Ronald Christensen
Department of Mathematics and Statistics
University of New Mexico

Statistical Inference

A Work in Progress

Springer

Seymour and Wes.

Preface

"But to us, probability is the very guide of life."
Joseph Butler (1736). *The Analogy of Religion, Natural and Revealed, to the Constitution and Course of Nature*, Introduction.
https://www.loc.gov/resource/dcmsiabooks.analogyofreligio00butl_1/?sp=41

"What is important is to spread confusion, not eliminate it." Salvidor Dali.
Quoted on the PUZZLES page of the *Albuquerque Journal*, October 9, 2024.

Normally, I wouldn't put anything this incomplete on the internet but I wanted to make parts of it available to my Advanced Inference Class, and once it is up, you have lost control.

Seymour Geisser was a mentor to Wes Johnson and me. He was Wes's Ph.D. advisor. Near the end of his life, Seymour was finishing his 2005 book *Modes of Parametric Statistical Inference* and needed some help. Seymour asked Wes and Wes asked me. I had quite a few ideas for the book but then I discovered that Seymour hated anyone changing his prose. That was the end of my direct involvement. The first idea for this book was to revisit Seymour's. (So far, that seems only to occur in Chapter 1.)

Thinking about what Seymour was doing was the inspiration for me to think about what I had to say about statistical inference. And much of what I have to say is inspired by Seymour's work as well as the work of my other professors at Minnesota, notably Christopher Bingham, R. Dennis Cook, Somesh Das Gupta, Morris L. Eaton, Stephen E. Fienberg, Narish Jain, F. Kinley Larntz, Frank B. Martin, Stephen Orey, William Suderth, and Sanford Weisberg. No one had a greater influence on my career than my advisor, Donald A. Berry. I simply would not be where I am today if Don had not taken me under his wing.

The material in the beginning of this book is what I (try to) cover in the first semester of a one year course on Advanced Statistical Inference. The other semester

I use Ferguson (1996) and here I have felt free to reference (but not prove) many results from his book. In addition to requiring a first course in Statistical Inference, such as Casella and Berger (?), the course is for students who have had at least Advanced Calculus and hopefully some Introduction to Analysis. The course is at least as much about introducing *some* mathematical rigor into their studies as it is about teaching statistical inference. (Rigor also occurs in our Linear Models class but there it is in linear algebra and here it is in analysis.) I try to introduce just enough measure theory for students to get an idea of its value (and to facility my presentations). But I get tired of doing analysis, so occasionally I like to teach some inference in the advanced inference course. The end of the book involves topics that are of interest to me.

Many years ago I went to a JSM (Joint Statistical Meeting) and heard Brian Joiner make the point that *everybody* learns from examples to generalities/theory. Ever since I have tried, with varying amounts of success, to incorporate this dictum into my books. (*Plane Answers*' first edition preceded that experience.) This book has four chapters of example based discussion before it begins the theory in Chapter 5. There are only three chapters of theory but they are tied to extensive appendices on technical material. The first appendix merely reviews basic (non-measure theoretic) ideas of multivariate distributions. The second briefly introduces ideas of measure theory, measure theoretic probability, and convergence. Appendix C introduces the measure theoretic approaches to conditional probability and conditional expectation. Appendix D adds a little depth (very little) to the discussion of measure theory and probability. Appendix E introduces the concept of identifiability. Appendix F merely reviews concepts of multivariate differentiation. Chapters 8 through 13 are me being self-indulgent and tossing in things of personal interest to me. (They don't actually get covered in the class.)

References to *PA* and *ALM* are to my books *Plane Answers* and *Advanced Linear Modeling*.

Recommended Additional Reading

Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, New York.
I would use this as a text if it were not out of print! I may have borrowed even more than I realized from this.
Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, Second Edition. John Wiley and Sons, New York.
Covers almost everything I want to cover. Not bad to read but I have found it impossible to teach out of.
Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton.
Excellent! All the analysis you need to do math stat? Limited because of age.
Lehmann, E. L. (1983) *Theory of Point Estimation*. John Wiley and Sons, New York. (Now Lehmann and Casella.)
Lehmann, E. L. (1986) *Testing Statistical Hypotheses*, Second Edition. John Wiley and Sons, New York. (Now Lehmann and Romano.)
Berger, J. O. (1993). *Statistical Decision Theory and Bayesian Analysis*. Revised Second Edition. Springer-Verlag, New York.
We used to teach Advanced Inference out of above three books.
Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.
More profound on ideas, less on math.
Manoukian, E. B. (1986), *Modern Concepts and Theorems of Mathematical Statistics*. Springer-Verlag, New York.
SHORT! Usually the first book off my shelf. Statements, not proofs or explanations.
Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, Third Edition. John Wiley and Sons, New York.
For our purposes, a source for multivariate normal only. An alternative to *PA* for this.
Wilks, *Mathematical Statistics*; Zacks, *Theory of Statistical Inference*.
Both old but thorough. Wilks is great for order statistics and distributions related to discrete data.
Wasserman, Larry (2004). *All of Statistics*. Springer, New York.

Statistical Inference

Cox, D.R. (2006). *Principles of Statistical Inference*. Cambridge University Press, Cambridge.
Fisher, R.A. (1956). *Statistical Methods and Scientific Inference*, Third Edition, 1973. Hafner Press, New York.
Geisser, S. (1993). *Modes of Parametric Statistical Inference*. Wiley, New York.

Bayesian Books

de Finetti, B. (1974, 1975). *Theory of Probability*, Vols. 1 and 2. John Wiley and Sons, New York.
Jeffreys, H. (1961). *Theory of Probability*, Third Edition. Oxford University Press, London.
Savage, L. J. (1954). *The Foundations of Statistics*. John Wiley and Sons, New York.
DeGroot, M. H. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.
First three are foundational. There are now TONS of other books, see mine for other references.

Large Sample Theory Books

Ferguson, Thomas S. (1996). *A Course in Large Sample Theory*. Chapman and Hall, New York. I teach out of this. (Need I say more?)
Lehmann, E. L. (1999) *Elements of Large-Sample Theory*. Springer.
Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*, Wiley, (paperback, 2001)
I haven't read either of the last two, but I hear they are good.

Probability and Measure Theory Books

Ash, Robert B. and Doleans-Dade, Catherine A. (2000). *Probability and Measure Theory*, Second Edition. Academic Press, San Diego.

I studied the first edition of this while in grad school and continue to use it as a reference.

Billingsley, Patrick (2012). *Probability and Measure*, Fourth Edition. Wiley, New York.

I haven't read this, but I hear it is good. There are lots of others. There are also lots of good books on probability alone.

Contents

Chapter 1
Overview

1.1 Early Examples

The 12th century theologian, physician, and philosopher Maimonides used probability to address a temple tax problem associated with women giving birth to boys when the birth order is unknown, see Geisser (2005) and Rabinovitch (1970).

One of the earliest uses of statistical testing was made by Arbuthnot (1710). He had available the births from London for 82 years. Every year there were more male births than females. Assuming that yearly births are independent and the probability of more males is 1/2, he calculated the chance of getting all 82 years with more males as $(0.5)^{82}$. This being a ridiculously small probability, he concluded that boys are born more often.

In the last half of the eighteenth century, Bayes, Price, and Laplace used what we now call Bayesian estimation and Daniel Bernoulli used the idea of maximum likelihood estimation, cf. Stigler (2007).

1.2 Testing

One of the famous controversies in statistics is the dispute between Fisher and Neyman-Pearson about the proper way to conduct a test. Hubbard and Bayarri (2003) give an excellent account of the issues involved in the controversy. Another famous controversy is between Fisher and almost all Bayesians. In fact, Fienberg (2006) argues that Fisher was responsible for giving Bayesians their name. Fisher (1956) discusses one side of these controversies. Berger's Fisher lecture attempted to create a consensus about testing, see Berger (2003).

The Fisherian approach is referred to as significance testing. The Neyman-Pearson approach is called hypothesis testing. The Bayesian approach to testing is an alternative to Neyman and Pearson's hypothesis testing. A quick review and comparison of these approaches is given in Christensen (2005). Here we cover much

of the same material but go into more depth with Chapter 2 examining significance testing, Chapter 3 discussing hypothesis testing, and Chapter 4 drawing comparisons between the methods. These three chapters try to introduce the material with a maximum of intuition and a minimum of theory.

1.3 Decision Theory

Chapter 5 introduces decision theory. Chapters 6 and 7 particularize decision theory into the subjects of estimation and hypothesis testing, respectively. My first course in statistical inference was out of Lindgren (1968). That edition of Lindgren's book took a decision theoretic approach to statistical inference and I have never been able to view statistical inference (other than significance testing) except through the lens of decision theory.

von Neumann and Morgenstern (1944) developed game theory and in particular the theory of two person zero sum games. (In a zero sum game whatever you win, I lose.) Wald (1950) recognized that Statistics involved playing a game with god. Blackwell and Girshick (1954) presented the early definitive work on decision theory. Raiffa and Schlaiffer (1961) systematized the Bayesian approach. Ferguson (1966), DeGroot (1970), and more recently Parmigiani and Inoue (2009) all make notable contributions.

The Likelihood Principal is that whenever two likelihoods are proportional, all statistical inference should be identical, cf. Barnard (1949). Berger and Wolpert (1984) have written a fascinating book on the subject. Royball (1997) has argued for basing evidentiary conclusions on relative values of the likelihood function. Hill (1987) questions the validity of the likelihood principal.

1.4 Some Ideas about Inference

Chapters 8 through 11 contain various ideas I have had about statistical inference.

1.5 The End

The last two chapters are easy going. The first of these contains edited reprints of my *JASA* reviews for two books on statistical inference by great statisticians: D. R Cox and Erich Lehmann. The last chapter is as a reprint. It is a short biography of Seymour Geisser.

1.6 The Bitter End

The absolute end of the book is a series of appendices that cover multivariate distributions, an introduction to measure theory and convergence, a discussion of how the Radon-Nikodym theorem provides the basis for measure theoretic conditional probability, some additional detail on measure theory, and finally a summary of multivariate differentiation.

Chapter 2
Significance Tests

In his seminal book *The Design of Experiments*, R.A. Fisher (1935) illustrated significance testing with the example of "The Lady Tasting Tea," cf. also Salsburg (2001). Briefly, a woman claimed that she could tell the difference between whether milk was added to her tea or if tea was added to her milk. Fisher set up an experiment that allowed him to test that she could not.

Fisher (1935, p.14) says, "In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result."

The fundamental idea of significance testing is to extend the idea of a proof by contradiction into probabilistic settings.

Fisher (1925, p.80) says, "The term Goodness of Fit has caused some to fall into the fallacy that the higher the value of P the more satisfactorily is the hypothesis verified. Values over .999 have sometimes been reported which, if the hypotheses were true, would only occur in only once in a thousand trials. ... In these cases the hypothesis considered is as definitely disproved as if P had been.001."

2.1 Generalities

The idea of a proof by contradiction is that you start with a collection of (antecedent) statements, you then work from those statements to a conclusion that cannot possibly be true, i.e., a contradiction, so that you can conclude that *something* must be wrong with the original statements. For example if I state that "all women have blue eyes" and that "Sharon has brown eyes" but then I observe the data that "Sharon is a woman," it follows that either the statement that "all women have blue eyes" and/or the statement that "Sharon has brown eyes" must be false. Ideally, I would know that all but one of the antecedent statements are correct, so a contradiction would tell me that the final statement must be wrong. Since I happen to know that

the antecedent statement "Sharon has brown eyes" is true, it must be the statement "all women have blue eyes" that is false. I have proven by contradiction that "not all women have blue eyes," but to do that I needed to know that both of the statements "Sharon has brown eyes" and "Sharon is a woman" are true.

In significance testing we collect data that we take to be true ("Sharon is a woman") but we rarely have the luxury of knowing that all but one of our antecedent statements are true. In practice, we do our best to validate all but one of the statements (we look at Sharon's eyes and see that they are brown) so that we can have some idea which antecedent statement is untrue.

In a significance test, we start with a probability model for some data, we then observe data that are supposed to be generated by the model, and if the data are impossible under the model, we have a contradiction, so something about the model must be wrong. The extension of proof by contradiction that is fundamental to significance testing is that, if the data are merely weird (unexpected) under the model, that gives us a philosophical basis for questioning the validity of the model. In the Lady Tasting Tea experiment, Fisher found weird data suggesting that something might be wrong with his model, but he designed his experiment so that the only thing that could possibly be wrong was the assumption that the lady was incapable of telling the difference.

EXAMPLE 2.1.1. *One observation from a known discrete distribution.*
Consider the probability model for a random variable y

r	1	2	3	4
$\Pr(y=r)$	0.980	0.005	0.005	0.010

If we take an observation, supposedly from this distribution, and see anything other than y equal to 1, 2, 3, or 4, we have an absolute contradiction of the model. The model must be wrong. More subtly, if we observe y equal to 2, 3, or 4, we have seen something pretty weird because 98% of the time we would expect to see $y = 1$. Seeing anything other than $y = 1$ makes us suspect the validity of this model for such datum. It isn't that 2, 3, or 4 cannot happen, it is merely that they are so unlikely to happen that seeing them makes us suspicious.

Note that we have used the probability distribution itself to determine which data values seem weird and which do not. Obviously, observations with low probabilities are those least likely to occur, so they are considered weird. In this example, the weirdest observations are $y = 2,3$ followed by $y = 4$. □

The crux of significance testing is that you need somehow to determine how weird the data are. Arbuthnot (1710) found it suspicious that for 82 years in a row, more boys were born in London than girls. Suspicious of what? The idea that male and female births are equally likely. Many of us would find it suspicious if males were more common for merely 10 years in a row. If birth sex has a probability model similar to coin flipping, each year the probability of more males should be 0.5 and outcomes should be independent. Under this model the probability of more males 10 years in a row is $(0.5)^{10} \doteq 0.001$. Pretty weird data, right? But no weirder

than seeing ten years with more boys the first year then alternating with girls, i.e. (B,G,B,G,B,G,B,G,B,G), and no weirder than any other specific sequence, say (B,B,G,B,B,G,B,G,G,G). What seems relevant here is the total number of years with more boys born, not the particular pattern of which years have more boys and which have more girls.

Therein lies the rub of significance testing. To test a probability model you need to summarize the observed data into a test statistic and then you need to determine the relative weirdness of the possible values of the test statistic as determined by the model. Typically, we would choose a test statistic that will be sensitive to the kinds of things that we think most likely to go wrong with the model (e.g, one sex might be born more often than the other). If the distribution of the test statistic is discrete, it seems pretty clear that a good measure of weirdness is having a small probability. If the distribution of the test statistic is continuous, it seems that a good measure of weirdness is having a small probability density, but we will see later that there are complications associated with using densities.

For our birth sex problem, the coin flipping model implies that the probability distribution for the number of times boys exceed girls in 10 years is binomial, specifically, $\text{Bin}(10,0.5)$. The natural measure of weirdness for the outcomes in this model is the probability of the outcome. The smaller the probability, the weirder the outcome.

Traditionally, a *P value* is employed to quantify the idea of weirdness. *The P value is defined as the probability of seeing something as weird or weirder than you actually saw*. It measures how consistent the data are with the hypothesized model. If there is little consistency with the model, that provides evidence that *something* is wrong with the model. We won't know what in particular is wrong with the model unless we can validate all of the assumptions in the model except one.

EXAMPLE 2.1.1 CONTINUED.
One observation from a known discrete distribution.
In this example, the weirdest observations are 2 and 3 and they are equally weird. The probability of seeing something as weird as seeing a 2 is the probability of seeing a 2 or 3, which is $0.005+0.005=0.01$. Similarly, the P value when observing 3 is also 0.01. If you see $y=4$, both 2 and 3 are weirder, so the probability of seeing something as weird or weirder than $y=4$ is $\Pr(y=2 \text{ or } 3 \text{ or } 4)=0.02$. The following table presents all of the possible P values for this model.

r	1	2	3	4
$\Pr(y=r)$	0.980	0.005	0.005	0.010
P value	1.00	0.01	0.01	0.02

What we expect to see is $y=1$ and, in this model, seeing it is fully consistent with the model as illustrated by having a P value of 1.

Actually, the weirdest things to see here are values of y other than 1, 2, 3, 4 because such values have (collectively) zero probability. Observing any such data gives a P value of 0, and a complete contradiction of the model. Moreover, their collective probability of 0 adds nothing to the P values for observing 1, 2, 3, 4. □

Anything with a P value of 0 is something that cannot happen and gives an absolute contradiction to the assumed probability model. In practice, one rarely obtains P values of 0 but often encounters P values being rounded off to 0.

EXAMPLE 2.1.2. *Birth sex P values.*
Our coin flipping model for birth sex implies the $\text{Bin}(10, 0.5)$ model for the number of years in which boys births exceed girl births. The probability distribution and possible P values are given below.

r	0	1	2	3	4	5	6	7	8	9	10
$\Pr(y = r)$	.001	.010	.044	.117	.205	.246	.205	.117	.044	.010	.001
P value	.002	.022	.110	.344	.754	1.000	.754	.344	.110	.022	.002

Note that seeing all girls is just as weird as seeing all boys, so the P for seeing 10 out of 10 boys is twice the probability of that outcome. The datum that is most consistent with the model is seeing 5 boys. If we do see 10 out of 10 boys, that suggests that the $\text{Bin}(10, 0.5)$ model is wrong but does not itself suggest why the model is wrong or what about the model is wrong. □

In a significance test there is only one probability model being tested so there is no real need to give that model a label. But historically, significance tests have been confounded with the hypothesis tests discussed in the next chapter, and it has become common to refer to the probability model being tested as the "null model" or as the "null hypothesis." In many situations it makes sense to think of there being an overall model for the data and a specific claim about that model (called the *null hypothesis*) which, together, define the *null model*. If the data are inconsistent with the null model, we do not know if (the data are just weird or if) the overall model is wrong or if the null hypothesis is wrong. In practice, we do our best to validate the overall model, so that it makes some sense to claim that the null hypothesis may be wrong (cf. the Duhem-Quine thesis).

In significance testing, the P value is the fundamental concept but it can be useful to discuss α level tests. Technically, an α level is simply a decision rule as to which P values will cause one to reject the null model. In other words, it is merely a decision point as to how weird the data must be before rejecting the null model. *In an α level significance test, if the P value is less than or equal to α, the null model is rejected.* Implicitly, an α level determines what data would cause one to reject the null model and what data will not cause rejection. The α level rejection region is defined as the set of all data points that have a P value less than or equal to α.

Note that in Example 2.1.1, an $\alpha = 0.01$ test is identical to an $\alpha = 0.0125$ test. Both reject when observing either $r = 2$ or 3. Moreover, the probability of rejecting an $\alpha = 0.0125$ test when the null model is true is *not* 0.0125, it is 0.01. However, significance testing is not interested in what the probability of rejecting the null hypothesis *will be*, it is interested in what the probability *was* of seeing weird data.

If an α level is chosen, for any semblance of a contradiction to occur, the α level must be small. On the other hand, making α too small will defeat the purpose of the test, making it extremely difficult to reject the test for any reason. A reasonable view would be that an α level should never be chosen; that a scientist should simply

evaluate the evidence embodied in the P value. (But that would not allow us to define confidence regions associated with significance tests.)

EXAMPLE 2.1.3. *Fisher's Exact Test for* 2×2 *Tables.*
Consider the following subset of the knee injury data of Christensen (1997, Example 2.3.1). The two factors are surgical outcome: either excellent (E) or good (G), and the source of injury to the knee: either the knee was subject to a direct blow (Direct) or the knee was also twisted (Both).

	Result		
Injury	E	G	Total
Direct	3	2	5
Both	7	1	8
Total	10	3	13

We want to test whether the probability of an excellent outcome (given that the outcome is E or G) is the same for a direct as for both. Fisher's Exact Test can be based on the conditional distribution of the number y_1 of E outcomes for Direct injuries, *given* that the total number of E outcomes was 10, the number of Direct injuries was 5, and the number of Both injuries was 8. We will show later that the distribution of y_1 and the P values are given by

r	0	1	2	3	4	5
$\Pr(y_1 = r)$	0	0	10/286	80/286	140/286	56/286
P value	0	0	$10/286 = 0.035$	$146/286 = 0.510$	1	$66/286 = 0.231$

Since there are 5 Direct injuries, the possible values for the number of E outcomes, say r, is 0 to 5. However, we are conditioning on having seen a total of 10 E outcomes, and it is impossible to get a total of 10 Es from samples of 5 Directs and 8 Boths without having at least 2 of the Direct injuries being Es. Given the numbers of Direct and Both injuries (both numbers treated as fixed) and the total number of excellent outcomes, the only outcome that would constitute any reasonable evidence that something is wrong with the null model (has a small P value) is seeing $y_1 = 2$, which we did not see. (I did all these computations by hand except for finding the decimal P values.) □

We now derive the conditional distribution. The assumed model for the data is that independently $y_i \sim \text{Bin}(N_i, p_i)$, $i = 1, 2$. The null hypothesis is $p_1 = p_2 \equiv p$, so under the null model y_1 and y_2 are independent $y_i \sim \text{Bin}(N_i, p)$. Write the 2×2 table as

	Success	Failure	Total
Group 1	y_1	$N_1 - y_1$	N_1
Group 2	y_2	$N_2 - y_2$	N_2
Total	t	$N_1 + N_2 - t$	$N_1 + N_2$

Note that under the null model $t \equiv y_1 + y_2 \sim \text{Bin}(N_1 + N_2, p)$, so

$$\Pr(t=s) = \binom{N_1+N_2}{s} p^s (1-p)^{N_1+N_2-s}.$$

Because y_1 and y_2 are independent,

$$\begin{aligned}
\Pr(y_1 = r \text{ and } t = s) &= \Pr(y_1 = r \text{ and } y_2 = s-r) \\
&= \Pr(y_1 = r)\Pr(y_2 = s-r) \\
&= \binom{N_1}{r} p^r (1-p)^{N_1-r} \times \binom{N_2}{s-r} p^{s-r} (1-p)^{N_2-s+r} \\
&= \binom{N_1}{r}\binom{N_2}{s-r} p^s (1-p)^{N_1+N_2-s}.
\end{aligned}$$

It follows that

$$\Pr(y_1 = r | t = s) = \frac{\Pr(y_1 = r \text{ and } t = s)}{\Pr(t=s)} = \binom{N_1}{r}\binom{N_2}{t-r} \Big/ \binom{N_1+N_2}{t},$$

for allowable r. This is known as the *hypergeometric distribution*. Remember that N_1 and N_2 are also being treated as fixed and known here. The same distribution would obtain if we had a multinomial sample of the four categories in the table and conditioned on one column total and one row total.

Unlike most tests for categorical data, this test does not depend on any large sample approximations. It gives exact probabilities for any sample sizes N_1 and N_2. However, the computations become difficult with large samples.

Exercise 2.1. Show that the same distribution results if the 4 values in the table are multinomial with Success/Failure independent of Groups when conditioning on one row total and one column total.

EXAMPLE 2.1.3. *Fisher's Exact Test for Twins.*
Fisher (1925, Section 20.1, Example 13.1) examined data from Lange on 30 people who had been convicted of a crime that were twins. The data are whether their fellow twin had also been convicted of a crime.

	Result		
Twins	Convicted	Not	Total
Identical	10	3	13
Fraternal	2	15	17
Total	12	18	30

As just demonstrated, the probability distribution for Fisher's Exact Test is

$$\Pr(y_1 = r) = \binom{13}{r}\binom{17}{12-r} \Big/ \binom{30}{12},$$

where we have suppressed in the notation that the probability is conditional on seeing 12 total convictions, 13 identical twins, and 17 fraternal twins.

The reason for addressing this example is because Fisher argued that the only more extreme tables are

	Result				Result		
Twins	Convicted	Not	Total	Twins	Convicted	Not	Total
Identical	11	2	13	Identical	12	1	13
Fraternal	1	16	17	Fraternal	0	15	17
Total	12	18	30	Total	12	18	30

so Fisher computed a P value as the sum of the probabilities of these three tables giving

$$\frac{619}{1330665} = 0.000465.$$

This looks like some kind of one-sided test rather than a significance test.

Indeed, it is not a significance test. The probability of seeing what we saw is

$$\Pr(y_1 = 10) = \frac{1}{\binom{30}{12}}\binom{13}{10}\binom{17}{2} = \frac{1}{\binom{30}{12}}\frac{13\cdot 12\cdot 11}{3\cdot 2}\frac{17\cdot 16}{2} = \frac{1}{\binom{30}{12}}17\cdot 13\cdot 16\cdot 11.$$

The extreme table in the other direction is

	Result		
Twins	Convicted	Not	Total
Identical	0	13	13
Fraternal	12	5	17
Total	12	18	30

which has probability

$$\Pr(y_1 = 0) = \frac{1}{\binom{30}{12}}\binom{13}{0}\binom{17}{12} = \frac{1}{\binom{30}{12}}\frac{17\cdot 16\cdot 15\cdot 14\cdot 13}{5\cdot 4\cdot 3\cdot 2} = \frac{1}{\binom{30}{12}}17\cdot 13\cdot 2\cdot 14.$$

Obviously, $16\cdot 11 > 2\cdot 14$, so $\Pr(y_1 = 10) > \Pr(y_1 = 0)$ and $\Pr(y_1 = 0)$ should be added in when computing the P value. It turns out that the second most extreme case in the other direction has $\Pr(y_1 = 1) > \Pr(y_1 = 10)$, so it does not need to be added to P. I used R to compute the distribution as shown below and obviously $\Pr(y_1 = 10) < \Pr(y_1 = r)$, $r = 1,\ldots,9$.

r	$\Pr(y_1 = r)$	r	$\Pr(y_1 = r)$
0	0.00007154318	7	0.1227681
1	0.001860123	8	0.03541387
2	0.01753830	9	0.005621250
3	0.08038387	10	0.0004497000
4	0.2009597	11	0.00001533068
5	0.2893819	12	0.0000001503008
6	0.2455362		

The significance test P value is 0.000472, not a whole lot different from Fisher's 0.000465. □

EXAMPLE 2.1.4. *Exact Tests for Contingency Tables.*
With modern computational tools it is possible to extend Fisher's exact test to bigger tables than 2×2. The complete knee injury data from Christensen (1997, Example 2.3.1) are

		Result			
	n_{ij}	E	G	F-P	Totals
Injury	Twist	21	11	4	36
	Direct	3	2	2	7
	Both	7	1	1	9
	Totals	31	14	7	52

where the surgical result F-P has collapsed together the Fair and Poor outcomes. The overall model for the data is that the results for each type of injury constitute independent multinomial distributions and the null hypothesis is that the probabilities are identical in each of the multinomials. The test statistic is taken as Pearson's chi-square or the generalized likelihood ratio statistic or some other appropriate statistic. One then finds the probability distribution of the test statistic given both the row totals and the column totals. The problem is that you have to enumerate every possible table that has the fixed row and column totals.

Somewhat ironically, testing for equal probabilities in independent multinomials is actually the most difficult problem as far as enumerating all of the possible tables. More complicated models for contingency tables, such as those used for three and higher dimensional tables, are conceptually easier to enumerate because, the more constraints you put on the model, the fewer possible models there are to enumerate. As Christensen (1997, Section 2.6) points out, even logistic regression models are just more highly constrained two-way contingency table models, so they are candidates for exact tests. Cyrus Mehta and Nitin Patel were in the forefront of developing this theory and associated software. See Christensen (2020c) for more discussion and references. □

To summarize, *as in any proof by contradiction, the results are skewed. If the data tend to contradict the model, we have evidence that the model is invalid. If the data are consistent with (do not tend to contradict) the model, we have an attempt at proof by contradiction in which we got no contradiction. If the model is not rejected, the best one can say is that the data are consistent with the model. Not rejecting certainly does not prove that the model is correct, whence comes the common exhortation that one can reject, but should never accept, a null hypothesis (model).*

To see that high consistency does not provide evidence that the model is correct, suppose your model for a population is that a distribution of heights is, in inches, $N(69,9)$. If your test statistic is the sample mean and the population was NBA basketball players, a sample of size 10 would surely convince you that the

model is wrong. If the population was male University of New Mexico (UNM) students, a sample of size 10 with sample mean of $\bar{y}_\cdot = 69.3$ is consistent with the model, having a P value of 0.92. But $\bar{y}_\cdot = 69.3$ is even more consistent with the model $N(69.0001, 9)$ and is most consistent with the model $N(69.3, 9)$. Of course $\bar{y}_\cdot = 69.3$ is also perfectly consistent with the model that *every* male at UNM has the height 69.3 inches, but (presumably) aspects of the data other than their mean could prove that model incorrect.

Seeing data that are consistent with the model does not make the model correct, anymore than making a bunch of assumptions and not being able to find a contradiction makes the assumptions correct. The logic is one directional. A contradiction means the assumptions are wrong, weird data suggest the model may be wrong. Admittedly, I do feel that collecting data that are consistent with a null model is a more worthwhile activity than merely thinking about assumptions and failing to find a contradiction.

A significance test can really be thought of as a model validation procedure. It makes no reference to any alternative model(s). We have the distribution of the null model and we examine whether the data look weird or not.

2.2 Continuous Distributions

For discrete distributions, i.e., distributions with a finite or countable number of outcomes, using the probabilities of those outcomes as a measure of their weirdness seems unexceptionable. I have never heard anyone object to it. For continuous distributions, no particular outcome has positive probability, so we use probability density functions to define probabilities via calculus. The probability of any set is the integral of the density over that set. It is then natural to rank the weirdness of continuous outcomes by how small the probability density is at that outcome. Frequently, this works well and we begin with illustrations of it working well. But probability densities are much more nebulous things than probabilities and we will also examine problems that arise from defining weirdness in terms of densities.

Earlier I said that "Anything with a P value of 0 is something that cannot happen and gives an absolute contradiction to the assumed probability model." With a continuous distribution, every observation is something that "cannot happen" yet things do happen. The truth is that we never see the outcome of a continuous distribution because we are incapable of making such observations. All measurement devices have a finite ability to measure, so all measurement devices only tell us that an observation has occurred within some interval. With good measuring devices we just ignore the interval and act like we saw the point. A potential method for dealing with the problems of densities is to go back to focusing on the observation intervals that have positive probability. (One difference between continuous and discrete distributions is that in a discrete distribution the total probability of all the outcomes with 0 probability is 0 but with continuous distributions all outcomes individually have 0 probability but collections of such outcomes can have positive probability.)

2.2.1 One Sample Normals

Assume a random sample of n observations,

Data	Distribution
$y_1, y_2, \ldots, y_n$ independent	$N(\mu, \sigma^2)$

The key assumptions are that the observations are independent and have the same distribution. In particular, we assume they have the same mean μ and the same variance σ^2. Assuming that the common distribution is normal facilitates working with the probability distributions. In particular, if we compute the sample mean $\bar{y}_\cdot$ and the (unbiased) sample variance s^2, it is well known that

$$\frac{\bar{y} - \mu}{\sqrt{s^2/n}} \sim t(n-1),$$

where $t(n-1)$ indicates the famous Student t distribution with $n-1$ degrees of freedom. Figure 2.1 illustrates three t distributions using the fact that $t(\infty) \equiv N(0,1)$. As the degrees of freedom df get larger, the $t(df)$ distributions get closer to a $N(0,1)$ distribution. For our purposes, the key facts are that all of these distributions have a maximum density at 0, are symmetric, and the density decreases as we get farther from 0.

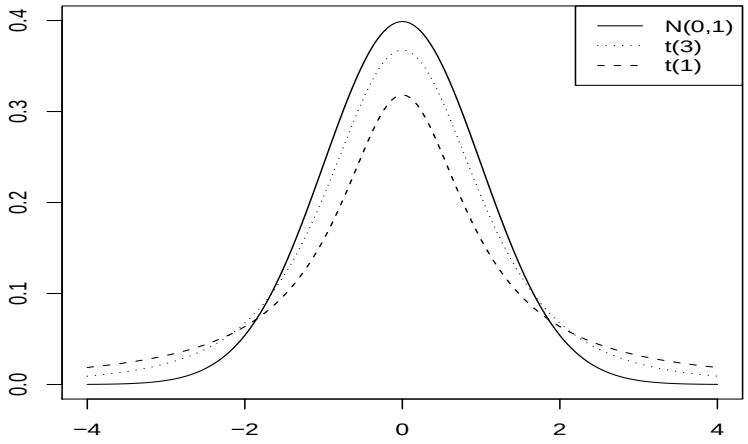

Fig. 2.1 Three distributions: solid, $N(0,1)$; long dashes, $t(1)$; short dashes, $t(3)$.

EXAMPLE 2.2.1. *One Sample t Test.*
We have specified the overall model for the data. Now specify a null hypothesis as, say, $H_0 : \mu = 3$. For a sample size of $n = 100$, the null model implies that

$$\frac{\bar{y} - 3}{\sqrt{s^2/100}} \sim t(99).$$

We are summarizing the data using the test statistic

$$t \equiv \frac{\bar{y} - 3}{\sqrt{s^2/100}}.$$

We want to collect data and then use them to determine whether or not the data give a test statistic that is consistent with the $t(99)$ distribution.

Taking weird observations to be those that have small probability densities, because of the shape of $t(df)$ distributions, weird observations are values of $(\bar{y} - 3)/\sqrt{s^2/100}$ that are far from 0. Because of symmetry about 0, the level of weirdness is determined by $|\bar{y} - 3|/\sqrt{s^2/100}$ with larger values more weird than smaller values.

If we happen to observe $\bar{y}_{obs} = 1$ and $s^2_{obs} = 4$, we get

$$t_{obs} \equiv \frac{\bar{y}_{obs} - 3}{\sqrt{s^2_{obs}/100}} = \frac{1 - 3}{\sqrt{4/100}} = -10,$$

which is a very strange thing to observe from a $t(99)$ distribution. (A $t(99)$ will be quite similar to a $N(0,1)$ distribution.) By symmetry about 0, seeing $(\bar{y} - 3)/\sqrt{s^2/100} = -10$ is exactly as weird as seeing $(\bar{y} - 3)/\sqrt{s^2/100} = 10$ and less weird than seeing anything with $|\bar{y} - 3|/\sqrt{s^2/100} < 10$. So the P value, being the probability of seeing anything as weird or weirder than the -10 that we actually saw, is the probability that a $t(99)$ distribution is (as far or) farther away from 0 than 10, i.e.,

$$P = \Pr[|t(99)| \geq |-10|] = \Pr[t(99) \leq -10] + \Pr[t(99) \geq 10],$$

which is *approximately* 0. This constitutes a great deal of evidence against the null model but it does not necessarily constitute evidence against the null hypothesis. Perhaps $\mu \neq 3$ but perhaps the data are not normal or perhaps the data are not independent or perhaps the observations have different variances or different means.

□

EXAMPLE 2.2.2. *One Sample F Test.*
It is a well known fact that the square of a $t(df)$ distribution is an $F(1, df)$ distribution. As in the previous example we have specified the overall model for the data and a null hypothesis $H_0 : \mu = 3$. For a sample size of $n = 100$, the null model implies that

$$\left[\frac{\bar{y}-3}{\sqrt{s^2/100}}\right]^2 \sim F(1,99).$$

We are summarizing the data using the test statistic

$$F \equiv \left[\frac{\bar{y}-3}{\sqrt{s^2/100}}\right]^2.$$

We want to collect data and then use them to determine whether or not the data give a test statistic that is consistent with the $F(1,99)$ distribution.

We again take weird observations to be those that have small probability densities but now the shape of the $F(1,df)$ distribution is as illustrated in Figure 2.2. Because of the shape of $F(1,df)$ distributions, weird observations are values of $(\bar{y}-3)^2/(s^2/100)$ that are above 0 with larger values more weird than smaller values.

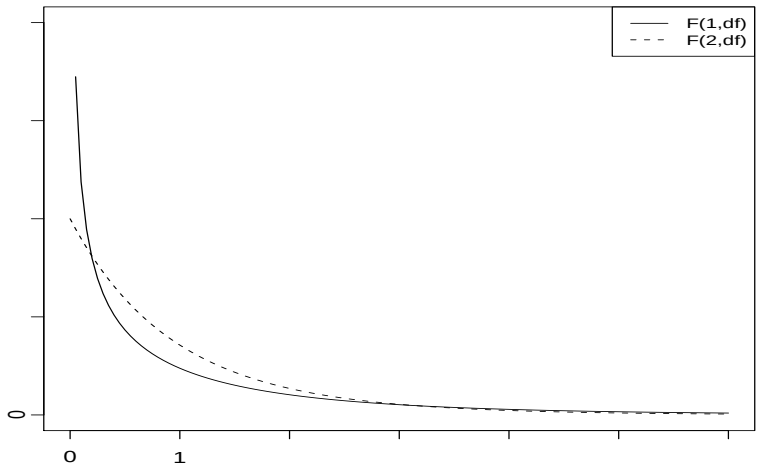

Fig. 2.2 $F(1,df)$ and $F(2,df)$ densities.

If we happen to observe $\bar{y}_{obs} = 1$ and $s^2_{obs} = 4$, we get $F_{obs} \equiv (\bar{y}_{obs}-3)^2/(s^2_{obs}/100) = 100$ which is a very strange thing to observe from a $F(1,df)$ distribution. The P value, being the probability of seeing anything as weird or weirder than the 100 that we actually saw, is the probability that a $F(1,99)$ distribution is (as far or) farther away from 0 than 100, i.e.,

$$P = \Pr[F(1,99) \geq 100] = \Pr[|t(99)| \geq 10],$$

which is *approximately* 0. In this case, the t and F significance tests correspond perfectly and give exactly the same interpretations. □

Exactly the same arguments apply to all the $t(df)$ tests and their corresponding $F(1,df)$ tests that arise in regression analysis, in analysis of variance (ANOVA), and in general linear models, cf. Christensen (1996, 2015, 2020a). Note that the equivalence was based entirely on the fact that $[t(df)]^2 \sim F(1,df)$ and on the shapes of the distributions.

Generally, to determine weirdness we have to know and evaluate the density of the test statistic under the null model. For an $F(d_1,d_2)$ distribution, that density is

$$f(x|d_1,d_2) \equiv \frac{1}{\mathrm{B}\left(\frac{d_1}{2},\frac{d_2}{2}\right)}\left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1}\left(1+\frac{d_1}{d_2}x\right)^{-\frac{d_1+d_2}{2}},$$

where $\mathrm{B}(\cdot,\cdot)$ is the Beta function, which is defined in terms of the Gamma function as

$$\mathrm{B}(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.$$

For $d_1 = 1,2$ the shape of the density was illustrated in Figure 2.2. For $d_1 \geq 3$, the shape is illustrated in Figure 2.3.

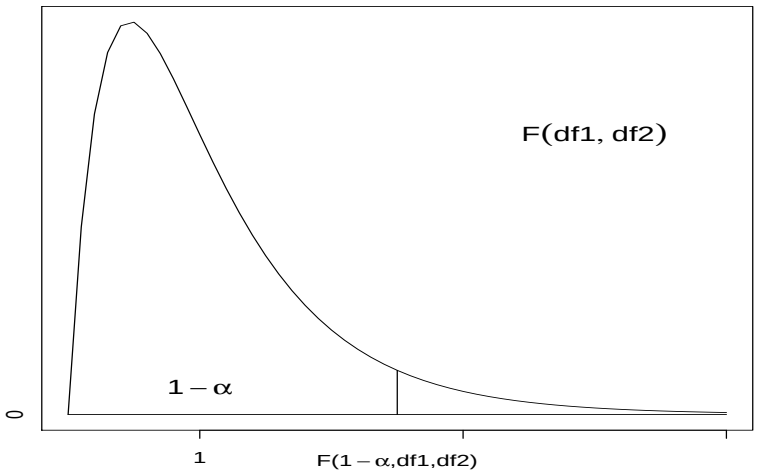

Fig. 2.3 Percentiles of $F(df1,df2)$ distributions; $df1 \geq 3$.

2.2.1.1 Linear Model F Tests

In regression analysis, analysis of variance (ANOVA), and in general linear models key statistics associated with a model are the sum of squares for error (SSE), degrees of freedom for error (dfE), and the mean squared error (MSE), wherein

$$MSE \equiv \frac{SSE}{dfE}.$$

As mentioned earlier, although significance tests themselves involve no alternative hypothesis (other than the trivial one that the null model is wrong), typically the test statistic is chosen with an eye towards those aspects of the null model that are least likely to be true. With linear models, test statistics are most often chosen with some "full" model in mind that ideally represents our overall model for the data, and we perform a test for the validity of a "reduced" model (one that is a special case of the full model), that represents our null model. (The null hypothesis consists of the restrictions on the full model needed to specify the reduced model.) As usual, if we reject the null model it means that either (we got weird data or) the null hypothesis is wrong or the full model is wrong. If we can validate the full model, we have reason to believe that the null hypothesis is what is wrong. Christensen (2015, Chapter 3) goes into details of significance tests for linear models. Christensen (2020a) goes into the details of linear model theory.

Using obvious notation, the general form of these tests are F tests where the test statistic is

$$F = \frac{[SSE(Red.) - SSE(Full)]/[dfE(Red.) - dfE(Full)]}{MSE(Full)}$$

and under the null model the distribution of F is

$$F \sim F\left[dfE(Red.) - dfE(Full), dfE(Full)\right].$$

The idea of the test is that under the null model both

$$\frac{SSE(Red.) - SSE(Full)}{dfE(Red.) - dfE(Full)} \quad \text{and} \quad MSE(Full)$$

should be unbiased estimates of the variance parameter σ^2, so under the null model F should take on a value close to 1, even though 1 is not typically the mean, median, or mode for the $F\left[dfE(Red.) - dfE(Full), dfE(Full)\right]$ distribution.

All software that I have seen computes $P \equiv \Pr[F > F_{obs}]$ where F_{obs} is the value of F computed from the observed values of $SSE(Red.)$, $dfE(Red.)$, $SSE(Full)$, and $dfE(Full)$. Indeed, if the full model is valid, only large values of F are inconsistent with the null model. But we do not know that the full model is valid. With weirdness defined by the density function, Figures 2.2 and 2.3 show that this $P \equiv \Pr[F > F_{obs}]$ computation only gives the significance test P value when $dfE(Red.) - dfE(Full) = 1, 2$. For $dfE(Red.) - dfE(Full) \geq 3$, finding the significance test P value requires

finding the density value $f(F_{obs}|dfE(Red.)-dfE(Full),dfE(Full))$ and a second value F_* such that $f(F_*) = f(F_{obs}|dfE(Red.)-dfE(Full),dfE(Full))$ and finding the probability that, if $F_* \leq F_{obs}$,

$$P = \Pr[F \leq F_*] + \Pr[F \geq F_{obs}]$$

or, if $F_* \geq F_{obs}$,

$$P = \Pr[F \geq F_*] + \Pr[F \leq F_{obs}].$$

EXAMPLE 2.2.3. *Numerical Examples*
Suppose we have 5 and 40 degrees of freedom with $F_{obs} = 2.75$. The one-sided P value associated with this is 0.031595. The density at F_{obs} is $f(2.75|5,40) = 0.0481$. It turns out that $f(0.0349|5,40) = 0.0481$ so $F_* = 0.0349$. The significance test P value is then

$$P = \Pr[F \leq 0.0349] + \Pr[F \geq 2.75] = 0.000691 + 0.031595 = 0.032286.$$

Not much difference.

What about small values? With the same degrees of freedom suppose $F_{obs} = 0.15$. The commonly computed one-sided P value is 0.978878 which is almost too good to be true and large enough to make many of us suspicious that something is wrong. The density at F_{obs} is $f(0.15|5,40) = 0.3113$ and it turns out that $f(1.504|5,40) = 0.3113$, so $F_* = 1.504$. The significance test P value is then

$$P = \Pr[F \leq 0.15] + \Pr[F \geq 1.504] = 0.021122 + 0.210298 = 0.232420,$$

which does not seem suspicious at all. □

We will revisit these numerical examples once we have developed another approach to these problems.

2.3 Testing Two Sample Variances.

Using the F density becomes more problematic when we seek to test the equality of the variances from two normal samples.

EXAMPLE 2.3.1. We examine Jolicoeur and Mosimann's log turtle height data, cf. Christensen (2015, Example 4.4.1), consisting of 24 female heights and 24 male heights. The sample variance of log female heights is $s_1^2 = 0.02493979$ and the sample variance of log male heights is $s_2^2 = 0.00677276$. The overall model is that the observations are independent with one normal distribution for females and another one for males. The null hypothesis is that the variances for males and females are the same, i.e., $\sigma_2^2 = \sigma_1^2$. The standard $\alpha = 0.01$ level hypothesis (not significance) test is rejected, i.e., we conclude that the null model with is wrong, if

$$F_{obs} = 0.2716 = \frac{0.00677276}{0.02493979} = \frac{s_2^2}{s_1^2} > F(0.995, 23, 23) = 3.04$$

or if

$$F_{obs} = 0.2716 < F(0.005, 23, 23) = \frac{1}{F(0.995, 23, 23)} = \frac{1}{3.04} = 0.33.$$

The second of these inequalities is true, so the null model with equal variances is rejected at the 0.01 level. This is a simple hypothesis test (by no means an optimal one).

The significance test is a bit more work to determine. Denote the density for the $F(23,23)$ distribution $f(z|23,23)$. Evaluating the density at F_{obs} gives $f(0.2716|23,23) = 0.03597$. It turns out that $f(2.50835|23,23) = 0.03597$, so $F_* = 2.50835$ and $F_{obs} = 0.2716$ are equally rare events. The P value, given the shape of $f(z|23,23)$, is

$$P = \Pr(F \leq 0.2716) + \Pr(F \geq 2.50835) = 0.00138 + 0.015974 = 0.017.$$

The significance test is not rejected at the 0.01 level.

What makes this problematic for significance testing is that because s_1^2 and s_2^2 differ only in their labels, there is no possible reason to prefer looking at s_2^2/s_1^2 rather than s_1^2/s_2^2. But the P value changes when you reverse the order.

If we base the test on

$$F_{obs} = \frac{s_1^2}{s_2^2} = 3.6824,$$

the significance test requires us to find $f(3.6824|23,23) = 0.002650$ and then a value F_*, different from F_{obs}, with $f(F_*|23,23) = 0.002650$. Since $f(0.17979|23,23) = 0.002650$, $F_* = 0.17979$ and the significance test P value becomes

$$P = \Pr(F \geq 3.6824) + \Pr(F \leq 0.17979) = 0.00138 + 0.0000573 = 0.001$$

which is far smaller than the other one. □

In the next section we will fix this particular problem. Incidentally, the standard (nonoptimal) α level hypothesis test illustrated at the beginning of the section does not change when you reverse the order of the sample variances, because of the mathematical fact that $F(\alpha, r, s) = F(1-\alpha, s, r)$, and that is true even when you have different numbers of degrees of freedom in the numerator and denominator.

Before proceeding we also need to look at a similar F significance test where the numerator and denominator degrees of freedom are not the same.

EXAMPLE 2.3.2. Consider the final point total data of Christensen (2015, Example 4.4.2.). For a sample of 15 females the sample variance was $s_1^2 = 487.28$ and for 22 males the sample variance was $s_2^2 = 979.29$. The test statistic can be $F = s_1^2/s_2^2$ with

$$F_{obs} = \frac{s_1^2}{s_2^2} = \frac{487.28}{979.29} = 0.498.$$

To find the significance test P value observe that $f(0.498|14,21) = 0.6207$, so $F_* = 1.2051$ because $f(1.2051|14,21) = 0.6207$. Using the $F(14,21)$ distribution

$$P = \Pr(F \geq 1.2051) + \Pr(F \leq 0.498) = 0.34028 + 0.09125 = 0.432$$

Alternatively, the test statistic can be $F = s_2^2/s_1^2$ with

$$F_{obs} = \frac{s_2^2}{s_1^2} = \frac{979.29}{487.28} = 2.010.$$

To find the significance test P value observe that $f(2.010|21,14) = 0.15339$, so $F_* = 0.31933$ because $f(0.31933|21,14) = 0.15339$. Using the $F(21,14)$ distribution

$$P = \Pr(F \geq 2.010) + \Pr(F \leq 0.31933) = 0.09125 + 0.00901 = 0.100.$$

Again the P values are unacceptably far apart. □

R code for finding the P values follows. A routine that can be adapted for finding F_* appears in the next section.

```
a=487.28/979.29
b=1/a

# Check densities at F_{obs} and F_*
df(a,14,21)
df(1.2051,14,21)
# Probabilities
pf(a,14,21)
1-pf(1.2051,14,21)

# Check densities at F_{obs} and F_*
df(b,21,14)
df(.31933,21,14)
# Probabilities
pf(.31933,21,14)
1-pf(b,21,14)
```

Do an example where we collect data 12.3, etc. compute F statistic, recompute with everything at 12.25 and recompute with 12.35, to see how much F changes. Discretize based on F changes.

2.4 Fisher's z distribution.

The F statistic was invented by George Snedecor in the 1930's at Iowa State University and labeled F in honor of R.A. Fisher. Many of Fisher's applications to which we now apply F tests were invented before the F distribution, so obviously Fisher did not originally use F tests. He used *Fisher's z distribution* which is

$$z \equiv \frac{1}{2}\log(F),$$

see Fisher (1924) and Aroian (1941). In particular, this has the density

$$\tilde{f}(x|d_1,d_2) = \frac{2d_1^{d_1/2}d_2^{d_2/2}}{B(d_1/2,d_2/2)} \frac{e^{d_1 x}}{(d_1 e^{2x}+d_2)^{(d_1+d_2)/2}},$$

which always has a mode of 0 and is symmetric for $d_1 = d_2$. Recall that F should be near 1 but typically 1 is not the mode of the F distribution. Here, $0 = (1/2)\log(1)$, so the density of Fisher's z distribution has its mode at the point that should be most consistent with the null model.

Frankly, given the current state of statistics, I cannot think of a single reason to use Fisher's z distribution *except* for using its density to define weirdness for F statistics We will continue to compute P values using software for F distributions but the sets we compute the probabilities for will be determined by the Fisher's z distribution.

EXAMPLE 2.4.1. We again examine the log turtle height data consisting of 24 female heights and 24 male heights. The sample variance of log female heights is $s_1^2 = 0.02493979$ and the sample variance of log male heights is $s_2^2 = 0.00677276$.

$$F_{obs} = 0.2716 = \frac{0.00677276}{0.02493979} = \frac{s_2^2}{s_1^2},$$

so

$$z_{obs} = \log(0.2716)/2 = -0.6517$$

Denote the density for the $z(23,23)$ distribution $\tilde{f}(z|23,23)$. From the symmetry of the distribution $\tilde{f}(-0.6517|23,23) = \tilde{f}(0.6517|23,23)$. The P value becomes

$$\begin{aligned} P &= \Pr[z \le -0.6517] + \Pr[z \ge 0.6517] \\ &= \Pr[F \le 0.2716] + \Pr[F \ge 1/0.2716] \\ &= \Pr[F \le 0.2716] + \Pr[F \ge 3.6824] \\ &= 0.0014 + 0.0014 = 0.0028. \end{aligned}$$

It turns out that if you perform the test with s_1^2/s_2^2 you will get the same P value. □

Even with unequal degrees of freedom, z gives the same P value either way you do the test.

EXAMPLE 2.4.2. Consider again the final point total data. For a sample of 15 females the sample variance was $s_1^2 = 487.28$ and for 22 males the sample variance was $s_2^2 = 979.29$. The test statistic can be $F = s_1^2/s_2^2$ with

$$F_{obs} = \frac{s_1^2}{s_2^2} = \frac{487.28}{979.29} = 0.497584985.$$

This leads to

$$z_{obs} = \log(0.498)/2 = -0.3489945$$

To find the significance test P value we first need to calculate $\tilde{f}(z_{obs}|14,21) = \tilde{f}(-0.3489945|14,21) = 0.6168776$, then find a value z_* on the other side of the mode from z_{obs} with $\tilde{f}(z_*|14,21) = \tilde{f}(z_{obs}|14,21)$, that is $\tilde{f}(0.3335880|14,21) = \tilde{f}(-0.3489945|14,21)$, and then we can find the probability.

$$\begin{aligned} P &= \Pr[z \le -0.3489945] + \Pr[z \ge 0.3335880] \\ &= \Pr[F \le 0.497584985] + \Pr[F \ge 1.948726] \\ &= 0.09125 + 0.081016 = 0.1723. \end{aligned}$$

Alternatively, the test statistic can be $F = s_2^2/s_1^2$ with

$$F_{obs} = \frac{s_2^2}{s_1^2} = \frac{979.29}{487.28} = 2.00970694.$$

This leads to

$$z_{obs} = \log(2.010)/2 = -\log(0.498)/2 = 0.3489945.$$

Although the test statistics are symmetric, with unequal degrees for freedom, Fisher's z distribution is not, so the P values will be slightly different. To find the significance test P value we first need to calculate $\tilde{f}(z_{obs}|21,14) = \tilde{f}(0.3489945|21,14) = 0.6168776$, then find a value z_* on the other side of the mode from z_{obs} with $\tilde{f}(z_*|21,14) = \tilde{f}(z_{obs}|21,14)$, that is $\tilde{f}(-0.3335880|21,14) = \tilde{f}(0.3489945|21,14)$, and then we can find the probability.

$$\begin{aligned} P &= \Pr[z \ge 0.3489945] + \Pr[z \le -0.3335880] \\ &= \Pr[F \ge 2.00970694] + \Pr[F \le 0.51315568] \\ &= 0.09125 + 0.081016 = 0.1723. \end{aligned}$$

□

The following R code illustrates the symmetry of the test.

```
fobs=0.497584985
xx=.3489945
```

```
x=c(log(fobs)/2,xx)
d1=14
d2=21
ftilde = (2*d1^(d1/2)*d2^(d2/2)/beta(d1/2,d2/2))*
exp(d1*x)/(d1*exp(2*x)+d2)^((d1+d2)/2)
matrix(c(x,ftilde),,2)

fobs=2.00970695
xx=-.3335880
x=c(log(fobs)/2,xx)
d1=21
d2=14
ftilde = (2*d1^(d1/2)*d2^(d2/2)/beta(d1/2,d2/2))*
exp(d1*x)/(d1*exp(2*x)+d2)^((d1+d2)/2)
matrix(c(x,ftilde),,2)
```

In the following R code by playing around with `a`, `b`, and `c` that determine `xx` you can figure out what z_* has to be. The first entry in `x` is z_{obs}, so the first entry in `ftilde` is what you are trying to reproduce. You can start with $a = -5$, $b = 5$, $1/10^c = .1$. Once you see what z_{obs} is, z_* should be somewhere near its negative. Pick `a` and `b` appropriately and then decrease the last term in `seq` by a factor of 10 as needed. Minor changes allow finding F_*. In particular, you would want to change to $a = 0$.

```
# Routine for finding z_*
# Adaptable for finding F_*
a=-5
b=5
c=1
fobs=0.497584985
xx=seq(a,b,1/10^c)
x=c(log(fobs)/2,xx)
# For F_* use
# x=c(fobs,xx)
d1=14
d2=21
ftilde = (2*d1^(d1/2)*d2^(d2/2)/beta(d1/2,d2/2))*
exp(d1*x)/(d1*exp(2*x)+d2)^((d1+d2)/2)
# For F_* use
# ftilde=df(x,d1,d2)
matrix(c(x,ftilde),,2)
```

Plots of these 3 densities, 36,36 and 14,21 and 21,14.

```
x=seq(-1,1,.01)
d1=23
d2=23
```

```
ftilde = (2*d1^(d1/2)*d2^(d2/2)/beta(d1/2,d2/2))*
exp(d1*x)/(d1*exp(2*x)+d2)^((d1+d2)/2)
plot(x,ftilde,type="l",ylim=c(0,2),ylab="",
xlab="",lty=3,labels=T)
d1=14
d2=21
ftilde1 = (2*d1^(d1/2)*d2^(d2/2)/beta(d1/2,d2/2))*
exp(d1*x)/(d1*exp(2*x)+d2)^((d1+d2)/2)
lines(x,ftilde1,type="l",lty=2)
d1=21
d2=14
ftilde2 = (2*d1^(d1/2)*d2^(d2/2)/beta(d1/2,d2/2))*
exp(d1*x)/(d1*exp(2*x)+d2)^((d1+d2)/2)
lines(x,ftilde2,type="l",lty=1)
legend("topright",c("z(23,23)","z(14,21)",
"z(21,14)"),lty=c(3,2,1))
```

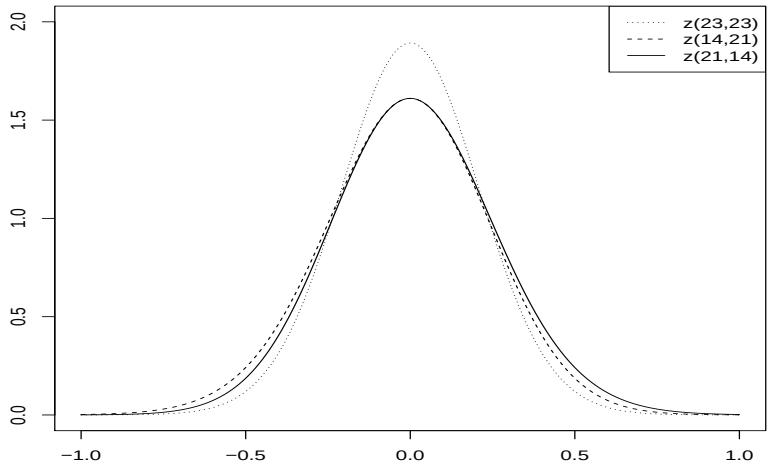

Fig. 2.4 Three Fisher $z(df1, df2)$ densities.

The R package `VGAM` has a command `dlogF` that gives the density for $2z = \log(F)$.

EXAMPLE 2.4.3. *Linear model F tests: Numerical Examples with $d_1 \geq 3$.*
As in Example 2.2.3., suppose we have 5 and 40 degrees of freedom with $F_{obs} = 2.75$. The one-sided P value associated with this is 0.031585. The significance test

P value from the F distribution is

$$P = \Pr[F \le 0.0349] + \Pr[F \ge 2.75] = 0.000691 + 0.031585 = 0.032.$$

$F_{obs} = 2.75$ leads to

$$z_{obs} = \log(2.75)/2 = 0.505800456$$

To find the significance test P value we first need to calculate $\tilde{f}(z_{obs}|5,40) = \tilde{f}(0.505800456|5,40) = 0.2647452$, then find a value z_* on the other side of the mode from z_{obs} with $\tilde{f}(z_*|5,40) = \tilde{f}(z_{obs}|5,40)$. In particular, $\tilde{f}(0.505800456|5,40) = \tilde{f}(-0.6823366|5,40)$, so we can find the probability.

$$\begin{aligned} P &= \Pr[z \le -0.6823366] + \Pr[z \ge 0.505800456|] \\ &= \Pr[F \le 0.2554641] + \Pr[F \ge 2.75] \\ &= 0.065451 + 0.031585 = 0.097. \end{aligned}$$

What about small values? With the same degrees of freedom suppose $F_{obs} = 0.15$. The commonly computed one-sided P value is 0.978878 which is almost too good to be true and large enough to make many of us suspicious that something is wrong. The significance test P value from an F test is

$$P = \Pr[F \le 0.15] + \Pr[F \ge 1.504] = 0.021122 + 0.210298 = 0.231.$$

$F_{obs} = 0.15$ leads to

$$z_{obs} = \log(0.15)/2 = -0.94856$$

To find the significance test P value from a z test, we first need to calculate $\tilde{f}(z_{obs}|5,40) = \tilde{f}(-0.94856|5,40) = 0.09340470$, then find a value z_* on the other side of the mode from z_{obs} with $\tilde{f}(z_*|5,40) = \tilde{f}(z_{obs}|5,40)$. With $\tilde{f}(-0.94856|5,40) = \tilde{f}(0.64079287|5,40)$, we can find the probability

$$\begin{aligned} P &= \Pr[z \le -0.94856] + \Pr[z \ge .64079287] \\ &= \Pr[F \le 0.15] + \Pr[F \ge 3.6023476] \\ &= 0.0211224 + 0.008772 = 0.0299. \end{aligned}$$

```
fobs=2.75
d1=5
d2=40
1-pf(fobs,d1,d2)
df(fobs,d1,d2)
df(.0349,d1,d2)
pf(.0349,d1,d2)
pf(.0349,d1,d2)+1-pf(fobs,d1,d2)

xx=-.6823366
```

```
x=c(log(fobs)/2,xx)
ftilde = (2*d1^(d1/2)*d2^(d2/2)/beta(d1/2,d2/2))*
exp(d1*x)/(d1*exp(2*x)+d2)^((d1+d2)/2)
matrix(c(x,ftilde),,2)
exp(2*x)
pf(exp(2*x),d1,d2)
```

□

EXAMPLE 2.4.4. *Linear model F tests, $d_1 = 1,2$.*
Earlier we showed that, because of the density shapes in Figure 2.2, F tests with $d_1 = 1,2$ are one-sided tests and that for $d_1 = 1$ an F test is equivalent to the corresponding t test.

We have specified the overall model for the data. Now specify a null hypothesis as, say, $H_0 : \mu = 3$. For a sample size of $n = 100$, the null model implies that

$$\frac{\bar{y} - 3}{\sqrt{s^2/100}} \sim t(99).$$

We are summarizing the data using the test statistic

$$t \equiv \frac{\bar{y} - 3}{\sqrt{s^2/100}}.$$

If we happen to observe $\bar{y}_{obs} = 2.6$ and $s^2_{obs} = 4$, we get

$$t_{obs} \equiv \frac{\bar{y}_{obs} - 3}{\sqrt{s^2_{obs}/100}} = -2,$$

and $F_{obs} = 4$.

$$z_{obs} = \log(4)/2 = 0.6931472$$

To find the significance test P value we first need to calculate $\tilde{f}(z_{obs}|5,40) = \tilde{f}(0.6931472|1,99) = 0.2196706$, then find a value z_* on the other side of the mode from z_{obs} with $\tilde{f}(z_*|1.99) = \tilde{f}(z_{obs}|1,99)$, that is $\tilde{f}(0.6931472|1,99) = \tilde{f}(-1.245495|1,99)$, and then we can find the probability.

$$\begin{aligned} P &= \Pr[z \le -1.245495] + \Pr[z \ge 0.6931472] \\ &= \Pr[F \le 0.0828279] + \Pr[F \ge 4] \\ &= 0.225897 + 0.048240 = 0.274. \end{aligned}$$

This oddity occurs because this z distribution is highly skewed to the left.

Personally, this is the only situation in which I would choose to use P values from the F distribution rather than Fisher's z. But I have not really investigated behavior with 2 degrees of freedom in the numerator. As a practical matter, although I do not choose to do it, I live with the one-sided P values imposed by standard software. And they are the same for both distributions. □

2.5 Final Notes

Rejecting a Significance test suggests that something is wrong with the (null) model. It does not specify what is wrong.

The example of a t test raises yet another question. Why should we summarize these data by looking at the t statistic,

$$\frac{\bar{y}-0}{s/\sqrt{n}}?$$

One reason is purely practical. In order to perform a test, one must have a known distribution to compare to the data. Without a known distribution there is no way to identify which values of the data are weird. With the normal data, even when assuming μ is known, we do not know σ^2 so we do not know the distribution of the data. By summarizing the data into the t statistic, we get a function of the data that has a known distribution, which allows us to perform a test. Another reason is essentially: why not look at the t statistic? If you have another statistic you want to base a test on, the Significance tester is happy to oblige. Fisher (1956, p. 49) indicates that the hypothesis should be rejected "if any relevant feature of the observational record can be shown to [be] sufficiently rare". After all, if the null model is correct, it should be able to withstand any challenge. Moreover, there is no hint in this passage of worrying about the effects of performing multiple tests. Inflating the probability of Type I error (rejecting the null when it is true) by performing multiple tests is not a concern in Significance testing because the probability of Type I error is not a concern in Significance testing.

The one place that possible alternative hypotheses arise in Significance testing is in the choice of test statistics. Again quoting Fisher (1956, p. 50), "In choosing the grounds upon which a general hypothesis should be rejected, personal judgement may and should properly be exercised. The experimenter will rightly consider all points on which, in the light of current knowledge, the hypothesis may be imperfectly accurate, and will select tests, so far as possible, sensitive to these possible faults, rather than to others." Nevertheless, the logic of Significance testing in no way depends on the source of the test statistic.

Although Fisher prefered his idea of fiducial inference, one can use Significance testing to arrive at "confidence regions" that do not involve either fiducial inference or repeated sampling. If you have a null model determined by an overall model and a null hypothesis about a parameter's value, a $(1-\alpha)$ confidence region can be defined simply as a collection of parameter values that would not be rejected by a α level significance test, that is, a collection of parameter values that are consistent with the data as judged by an α level test. This definition involves no long run frequency interpretation of "confidence." It makes no reference to what proportion of hypothetical confidence regions would include the true parameter. It does, however, require one to be willing to perform an infinite number of tests without worrying about their frequency interpretation. This approach also raises some curious ideas. For example, with the normal data discussed earlier, this leads to standard t confi-

dence intervals for μ and χ^2 confidence intervals for σ^2, but one could also form a joint 95% confidence region for μ and σ^2 by taking all the pairs of values that satisfy

$$\frac{|\bar{y} - \mu|}{\sigma/\sqrt{n}} < 1.96.$$

Certainly all such μ, σ^2 pairs are consistent with the data as summarized by $\bar{y}$.

Chapter 3
Hypothesis Tests

Significance testing predates hypothesis testing. A theory of hypothesis testing was formally set out in a series of papers by Jersy Neyman and Egon Pearson, cf. Lehmann (2011). This theory set out to expand/improve on significance testing. Bayesian testing provides an alternative to the Neyman-Pearson theory.

Significance tests are about deciding whether a single probability model is reasonable. Hypothesis tests are about deciding which of two probability models (or sets of probability models) is better.

As in the previous chapter, we introduce the ideas using elementary examples. A more general discussion of hypothesis testing is introduced in Chapter 5 and expanded in Chapter 7.

3.1 Testing Two Simple Hypotheses

We begin by discussing how to decide which of two discrete probability distributions is more consistent with observed data. We typically identify the distributions using their densities (mass functions). We could call them f and g but we call them $f(r|\theta)$, specifying probabilities on outcomes r for two different values for θ. In particular, consider

r	1	2	3	4
$f(r\|0)$	0.980	0.005	0.005	0.010
$f(r\|2)$	0.098	0.001	0.001	0.900

The first of these is the same distribution that we used to illustrate significance testing in Section 2.1. But now we no longer consider the question of whether the data seem consistent with the simple null hypothesis $H_0 : \theta = 0$. Now we ask whether the observed data are more consistent with $H_0 : \theta = 0$ or with the the alternative hypothesis $H_A : \theta = 2$.

These hypotheses are *simple* in the sense that the distributions involved are completely specified. An hypothesis that data come from a family of two or more distributions is call a *composite hypothesis*.

This is a decision problem. We have two possible distributions and we are deciding between them. The reformulation of significance testing into a decision problem is a primary reason that Fisher objected to Neyman-Pearson testing, see Fisher (1956, Chp. 4).

Before examining formal testing procedures, look at the distributions. Intuitively, if we see $r = 4$ we are inclined to believe $\theta = 2$, if we see $r = 1$ we are quite inclined to believe that $\theta = 0$, and if we see either a 2 or a 3, it is still 5 times more likely that the data came from $\theta = 0$.

While significance testing does not use an explicit alternative, there is nothing to stop us from doing two significance tests: a test of $H_0 : \theta = 0$ and then another test of $H_0 : \theta = 2$. The significance tests both give perfectly reasonable results. The test for $H_0 : \theta = 0$ has small P values for any of $r = 2,3,4$. These are all strange values when $\theta = 0$. The test for $H_0 : \theta = 2$ has small P values when $r = 2,3$.

When $r = 4$, we do not reject $\theta = 2$; when $r = 1$, we do not reject $\theta = 0$; when $r = 2,3$, we reject both $\theta = 0$ and $\theta = 2$. The significance tests are not being forced to choose between the two distributions. Seeing either a 2 or a 3 is weird under both distributions. Hypothesis testing decides between the available choices, it does not allow one to reject both choices.

3.1.1 Neyman-Pearson tests

Neyman-Pearson testing involves the concepts of Type I and Type II error. *Type I error is rejecting the null hypothesis when it is true and Type II error is not rejecting the null hypothesis when it is false.* This very statement is a legacy of significance testing in that it focuses on the null hypothesis. In this problem it should be equivalent to describe Type I error as accepting the alternative hypothesis when it is false and Type II error as not accepting the alternative hypothesis when it is true. (In significance testing you may reject a null hypothesis (model) but you never accept it.)

Neyman-Pearson (N-P) tests treat the two hypotheses in fundamentally different ways. A test of $H_0 : \theta = 0$ versus $H_A : \theta = 2$ is typically different from a test of $H_0 : \theta = 2$ versus $H_A : \theta = 0$. We will examine the test of $H_0 : \theta = 0$ versus $H_A : \theta = 2$. In a simple versus simple test test like this, there is no good reason for the testing problems to be different, but – in a legacy from significance testing – they typically are treated as different.

The asymmetry stems from N-P theory seeking to find the best α *level test, where* α *is the probability of Type I error (rejecting* H_0 *when it is true).* The *rejection region* is the set of data values that cause one to reject the null hypothesis, so under H_0 the probability of the rejection region must be α. The *best* α level test is defined to be the *most powerful* one, the one with the highest power, that is, the one with the

highest probability of rejecting H_0 (observing data in the rejection region) when H_A is true. Equivalently, the best test minimizes the probability of Type II error.

Defining the α level as the probability of rejecting the null hypothesis when it is true places an emphasis on repeated sampling so that the Law of Large Numbers suggests that about α of the time you will make an incorrect decision, provided the null hypothesis is true in all of the samples. While this is obviously a reasonable definition prior to seeing the data, its relevance after seeing the data is questionable.

Consider again our example.

r	1	2	3	4
$f(r\|0)$	0.980	0.005	0.005	0.010
$f(r\|2)$	0.098	0.001	0.001	0.900
$f(r\|2)/f(r\|0)$	0.1	0.2	0.2	90

As demonstrated in the famous Neyman-Pearson lemma (see Chapter 7 or Lehmann, 1997, Chp. 3), optimal N-P tests are based on the likelihood ratio $f(r|2)/f(r|0)$. The best N-P tests reject for the largest values of the likelihood ratio. For example, our largest ratio occurs when $r = 4$. If we take $r = 4$ to be our rejection region we will get a most powerful test for its size. The size of the test is the probability of rejecting the null when the null is true, so here it is $\alpha = 0.01$.

If we reject whenever the likelihood ratio is 0.2 or larger, we will reject for $r = 2,3,4$. This will be most powerful for its size which is the probability of this rejection region under the null, i.e., $\alpha = 0.005 + 0.005 + 0.01 = 0.02$.

It is easy to pick rejection regions that will be most powerful for their size but how do we specify a size and find the most powerful rejection region? How would we find an optimal $\alpha = 0.005$ level test? We have to reject for the largest values of $f(r|2)/f(r|0)$. The largest value corresponds to $r = 4$ but, as we saw, rejecting for $r = 4$ yields an $\alpha = 0.01$ test. We can reduce the size to $\alpha = 0.005$ by flipping a coin and only rejecting $r = 4$ if the coin turns up heads. Similarly, we could reduce the size to $\alpha = 0.01/6$ by rolling a die and only rejecting $r = 4$ if the die turns up six. You say you don't like this idea? Tough! It if fundamental to N-P theory. You say you don't like such a theory? Well, neither do I.

To allow arbitrary α levels, one must consider randomized tests. A randomized test requires a randomized rejection region. How could one perform an $\alpha = 0.0125$ test in our example? Three distinct tests are: (a) reject whenever $r = 4$ and flip a coin, if it comes up heads, reject when $r = 2$, (b) reject whenever $r = 4$ and flip a coin, if it comes up heads, reject when $r = 3$, (c) reject whenever $r = 2$ or 3 and flip a coin twice, if both come up heads, reject when $r = 4$. It is difficult to convince anyone that these are reasonable practical procedures. Yet all have the correct size α and the first two are both most powerful tests. The third is not most powerful because it more readily rejects when $f(r|2)/f(r|0) = 0.2$ than it does when $f(r|2)/f(r|0) = 90$. Thus it violates the Neyman-Pearson lemma structure of rejecting for the highest likelihood ratios.

Note that the best N-P test of size $\alpha = 0.01$ (rejecting when $r = 4$) is completely different from the 0.01 significance test of $H_0 : \theta = 0$ that rejected when $r = 2,3$.

(On the other hand, the $\alpha = 0.02$ N-P test coincides with the significance test. Both reject when observing any of $r = 2, 3, 4$.) The power of the $\alpha = 0.01$ N-P test is 0.9 whereas the power of the significance $\alpha = 0.01$ test is only $0.001 + 0.001 = 0.002$. Clearly the significance test is not a good way to decide between these alternatives. But then the significance test was not designed to decide between two alternatives. It was designed to see whether the null model seemed reasonable and, on its own terms, it works well. Although the meaning of α differs between significance and N-P tests, we have chosen two examples, $\alpha = 0.01$ and $\alpha = 0.02$, in which the significance test rejection region also happens to define an N-P test with the same numerical value of α. Such a comparison would not be appropriate if we had examined, say, $\alpha = 0.0125$ significance and N-P tests because significance tests do not admit randomized decision rules.

In particular, the motivation for insisting on small α levels seems to be based entirely on the philosophical idea of proof by contradiction. In a significance test, using a large α level eliminates the suggestion that the data are unusual and thus tend to contradict H_0. However, N-P testing cannot appeal to the idea of proof by contradiction. Later we will examine situations in which most powerful N-P tests reject for those data values that are *most* consistent with the null hypothesis. In particular, such examples make it clear that significance test P values can have no role in N-P testing! See also Hubbard and Bayarri (2003) and discussion.

It seems that once you base the test on wanting a large probability of rejecting when the alternative hypothesis is true (high power), you have put yourself in the business of deciding between the two hypotheses. Even on this basis, the N-P test does not always perform very well. The rejection region for the $\alpha = 0.02$ optimal N-P test of $H_0 : \theta = 0$ versus $H_A : \theta = 2$ includes $r = 2, 3$, even though 2 and 3 are five times more likely under the null hypothesis than under the alternative. Admittedly, 2 and 3 are weird things to see under either hypothesis, but when deciding between these specific alternatives, rejecting $\theta = 0$ (accepting $\theta = 2$) for $r = 2$ or 3 does not seem reasonable. The Bayesian approach to testing, discussed in the next subsection, seems to handle this decision problem well.

Instead of arbitrarily deciding on a small value for α, good N-P testing needs to play off the relative probabilities of Type I and Type II error. If a small α causes too large a β (i.e., probability of Type II error), the N-P tester should pick a bigger α (which will make β smaller), even to the point where α may no longer be "small." Somewhat ironically, in our little example, picking a smaller α, 0.01 instead of 0.02, increases β to 0.1 from 0.098, but the change is β is much smaller than the change in α, so the smaller α may be preferred. The point is that good N-P testing requires consideration of both α and β (or α and the power), yet traditional N-P testing tends just to pick a small α and try to do the best with it.

3.1.2 Bayesian Tests

Bayesian analysis requires us to have prior probabilities on the values of θ. It then uses Bayes' Theorem to combine the prior probabilities with the information in the data to find "posterior" probabilities for θ given the data. All decisions about θ are based entirely upon these posterior probabilities. The information in the data is obtained from the likelihood function. For an observed data value, say $r = r^*$, the likelihood is the function of θ defined by $f(r^*|\theta)$.

In our simple versus simple testing example, let the prior probabilities on $\theta = 0, 2$ be $p(0)$ and $p(2)$. Applying Bayes' theorem to observed data r, we turn these prior probabilities into posterior probabilities for θ given r, say $p(0|r)$ and $p(2|r)$. To do this we need the likelihood function which here takes on only the two values $f(r|0)$ and $f(r|2)$. From Bayes' Theorem,

$$p(\theta|r) = \frac{f(r|\theta)p(\theta)}{f(r|0)p(0) + f(r|2)p(2)}, \qquad \theta = 0, 2.$$

Decisions are based on these posterior probabilities. Other things being equal, whichever value of θ has the larger posterior probability is the value of θ that we will accept. If both posterior probabilities are near 0.5, we might admit that we do not know which is right.

In practice, posterior probabilities are computed only for the value of r that was actually observed, but Table 2 gives posterior probabilities for all values of r and two sets of prior probabilities: (a) one in which each value of θ has the same probability, $1/2$, and (b) one set in which $\theta = 2$ is five times more probable than $\theta = 0$.

Prior	r	1	2	3	4
	$f(r\|0)$	0.980	0.005	0.005	0.010
	$f(r\|2)$	0.098	0.001	0.001	0.900
$p_a(0) = 1/2$	$p_a(0\|r)$	0.91	0.83	0.83	0.01
$p_a(2) = 1/2$	$p_a(2\|r)$	0.09	0.17	0.17	0.99
$p_b(0) = 1/6$	$p_b(0\|r)$	0.67	0.50	0.50	0.002
$p_b(2) = 5/6$	$p_b(2\|r)$	0.33	0.50	0.50	0.998

Table 3.1 Posterior probabilities of $\theta = 0, 2$ for two prior distributions a and b.

As is intuitively reasonable, regardless of the prior distribution, if you see $r = 4$ the posterior is heavily in favor of $\theta = 2$ and if you see $r = 1$ the posterior substantially favors $\theta = 0$.

The key point is what happens when r equals 2 or 3. With equal prior weight on the θs, the posterior heavily favors $\theta = 0$, that is, with $r = 2$, $p_a(0|2) = 0.83$, $p_a(2|2) = 0.17$ and with $r = 3$, $p_a(0|3) = 0.83$, $p_a(2|3) = 0.17$. It is not until our prior makes $\theta = 2$ five times more probable than $\theta = 0$ that we wash out the evidence from the data that $\theta = 0$ is more likely, that is, $p_b(0|2) = p_b(2|2) = 0.50$

and $p_b(0|3) = p_b(2|3) = 0.50$. Given the prior, the Bayesian procedure is always reasonable.

The Bayesian analysis gives no special role to the null hypothesis. It treats the two hypotheses on an equal footing. That N-P theory treats the hypotheses in different ways is something that many Bayesians find disturbing.

As discussed in Chapter 5, if actions have losses or utilities associated with them, the Bayesian can base a decision on maximizing expected posterior utility or minimizing expected posterior loss. Berry (2004) discussed the practical importance of developing approximate utilities for designing clinical trials.

The absence of a clear source for the prior probabilities seems to be the primary objection to the Bayesian procedure. Typically, if we have enough data, the prior probabilities are not going to matter because the posterior probabilities will be substantially the same for different priors. If we do not have enough data, the posteriors will not agree but why should we expect them to? The best we can ever hope to achieve is that reasonable people (with reasonable priors) will arrive at a consensus when enough data are collected. In the example, seeing one observation of $r = 1$ or 4 is already enough data to cause substantial consensus. One observation that turns out to be a 2 or a 3 leaves us wanting more data.

3.2 Simple Versus Composite Hypotheses

We now consider testing a single distribution against a pair of distributions. Recall that a composite hypothesis is any hypothesis that contains more than one distribution. Specifically, we consider the example

r	1	2	3	4
$f(r\|0)$	0.980	0.005	0.005	0.010
$f(r\|1)$	0.100	0.200	0.200	0.500
$f(r\|2)$	0.098	0.001	0.001	0.900
$f(r\|1)/f(r\|0)$	10/98	40	40	50
$f(r\|2)/f(r\|0)$	0.1	0.2	0.2	90

and test the simple null hypothesis $H_0 : \theta = 0$ versus the composite alternative hypothesis $H_A : \theta > 0$. Here we can write the set of all θs as $\Theta = \{0,1,2\}$ with the null being $\Theta_0 \equiv \{0\}$ and the alternative $\Theta_A \equiv \{1,2\}$. With this notation we can also write the testing problem quite generally as $H_0 : \theta \in \Theta_0$ versus $H_A : \theta \in \Theta_A$. Since the composite alternative has only two values, we could alternatively write $H_A : \theta = 1,2$ or $H_A : \theta \neq 0$. Looking at the distributions in Table 1, the intuitive conclusions are pretty clear. For $r = 1$, go with $\theta = 0$. For $r = 4$, go with $\theta = 2$. For $r = 2,3$, go with $\theta = 1$.

Significance testing has nothing new to add to this situation except the observation that when $\theta = 1$, none of the data are really weird. In this case, the strangest observation is $r = 1$ which has a P value of 0.1.

3.2.1 Neyman-Pearson Testing

The best thing that can happen in N-P testing of a composite alternative is to have a uniformly most powerful test. With $H_A : \theta > 0$, let θ^* be a particular value that is greater than 0. Test the simple null $H_0 : \theta = 0$ against the simple alternative $H_A : \theta = \theta^*$. If, for a given α, the most powerful test has the same rejection region regardless of the value of θ^*, then that test is the *uniformly most powerful (UMP) test*. It is a simple matter to see that the $\alpha = 0.01$ N-P most powerful test of $H_0 : \theta = 0$ versus $H_A : \theta = 1$ rejects when $r = 4$. We have already seen that that is also true when the alternative is $H_A : \theta = 2$. Since the most powerful tests of the alternatives $H_A : \theta = 1$ and $H_A : \theta = 2$ are identical, and these are the only permissible values of $\theta > 0$, this is the uniformly most powerful $\alpha = 0.01$ test. The test makes a "bad" decision when $r = 2, 3$ because with $\theta = 1$ as a consideration, you would intuitively like to reject the test.

The $\alpha = 0.02$ uniformly most powerful test rejects for $r = 2, 3, 4$, which is in line with our intuitive evaluation, but recall from the previous section that this is the test that (intuitively) should not have rejected for $r = 2, 3$ when testing only $H_A : \theta = 2$.

Theoretically, the key thing to note is that as r varies, the relative order of $f(r|1)/f(r|0)$ is identical to the relative order of $f(r|2)/f(r|0)$. The largest value of $f(r|1)/f(r|0)$ and $f(r|2)/f(r|0)$ occurs at $r = 4$. The smallest value of $f(r|1)/f(r|0)$ and $f(r|2)/f(r|0)$ occurs at $r = 1$. For $r = 2, 3$ the values of $f(r|1)/f(r|0)$ are the same and between the other two values. The same holds for $f(r|2)/f(r|0)$. This common ordering means that, regardless of size, any most powerful test for $H_0 : \theta = 0$ versus $H_A : \theta = 1$ will also be a most powerful test for $H_0 : \theta = 0$ versus $H_A : \theta = 2$. Recall that the size of the test depends on the rejection region but not on the alternative hypothesis. The most powerful test of size $\alpha = 0.01$ rejects when $r = 4$ for both alternatives. The most powerful test of size $\alpha = 0.02$ rejects when $r = 2, 3, 4$ for both alternatives. One most powerful test of size $\alpha = 0.0125$ for both alternatives rejects when $r = 4$ and rejects for $r = 2$ if a coin flip comes up heads.

This same idea works very generally. Suppose we are testing $H_0 : \theta = \theta_0$ versus $H_A : \theta \in \Theta_A$. If, as r varies, the relative order of $f(r|\theta)/f(r|\theta_0)$ remains the same for any $\theta \in \Theta_A$, the most powerful test will remain the same regardless of which $\theta \in \Theta_A$ we are considering, so we can find a uniformly most powerful test. In particular, it is fairly obvious that if, for any $\theta \in \Theta_A$, the function $f(r|\theta)/f(r|\theta_0)$ is always either increasing (or always decreasing) in r, the relative order of the likelihood ratios will remain the same regardless of the choice of $\theta \in \Theta_A$. This property is called having *monotone likelihood ratio* and it is sufficient (but not necessary) for the existence of uniformly most powerful tests. In fact, our little example displays monotone likelihood ratio.

If the likelihood ratios are always increasing, UMP tests *can be chosen* that reject for large values of r (possibly with random rejection for the smallest value of r in the rejection region), and if they are decreasing, UMP tests can reject for small values of r. For example, because the likelihood ratios are not strictly increasing, as in our simple versus composite example, both (a) reject whenever $r = 4$ and flip a coin, if it comes up heads, reject when $r = 2$, and (b) reject whenever $r = 4$ and flip a coin,

if it comes up heads, reject when $r = 3$, are uniformly most powerful tests of size $\alpha = 0.0125$. But the second test is a UMP test that rejects for large values of r.

3.2.2 Bayesian Testing

A (nontrivial) Bayesian approach requires positive prior probabilities on both Θ_0 and Θ_A. When Θ_A is composite, we need a prior distribution on θ given that $\theta \in \Theta_A$. The easiest way to achieve this is by simply putting a prior distribution on Θ for which $0 < \Pr(\Theta_0) < 1$. When Θ is finite or countable, this is easy to do.

For our simple versus composite example, an even handed Bayesian approach might take prior probabilities that are the same for the null hypothesis and the alternative, that is, $\Pr[\theta = 0] = 0.5$ and $\Pr[\theta > 0] = 0.5$. Moreover, we might then put the same prior weight on every possible θ value within the alternative, thus $\Pr[\theta = 1|\theta > 0] = 0.5$ and $\Pr[\theta = 2|\theta > 0] = 0.5$. Equivalently, $p(0) = 0.5$, $p(1) = 0.25$, and $p(2) = 0.25$. Using Bayes Theorem,

$$p(\theta|r) = \frac{f(r|\theta)p(\theta)}{f(r|0)p(0) + f(r|1)p(1) + f(r|2)p(2)},$$

the posterior probabilities become

r	1	2	3	4
$p(0\|r)$	0.908	0.047	0.047	0.014
$p(1\|r)$	0.046	0.948	0.948	0.352
$p(2\|r)$	0.045	0.005	0.005	0.634
$\Pr[\theta > 0\|r]$	0.091	0.953	0.953	0.986

These agree well with the intuitive conclusions that $r = 1$ suggests $\theta = 0$, $r = 4$ suggests $\theta = 2$, and $r = 2,3$ suggest $\theta = 1$. This is true even though the prior puts twice as much weight on $\theta = 0$ as on the other θs. For the specific decision problem, $\Pr[\theta > 0|r]$ gives our probability that the alternative hypothesis is true when observing r. (There is a little roundoff error for $r = 1$.)

The Bayesian approach to testing a simple null against a composite alternative can be recast as testing a simple null versus a simple alternative. Using the prior probability on the values of θ given that the alternative hypothesis is true, one can find the average distribution for the data under the alternative. With $\Pr[\theta = 1|\theta > 0] = 0.5$ and $\Pr[\theta = 2|\theta > 0] = 0.5$, the average distribution under the alternative is $0.5f(r|1) + 0.5f(r|2)$. The Bayesian test of the $\theta = 0$ density $f(r|0)$ against this average density for the data under the alternative yields the posterior probabilities $p(0|r)$ and $\Pr[\theta > 0|r]$, cf. Table 3.2.

It might also be reasonable to put equal probabilities on every θ value. In problems like this, where you know the data distributions, the only way to get unreasonable Bayesian answers is to use an unreasonable prior.

Prior	r	1	2	3	4
	$f(r\|0)$	0.980	0.005	0.005	0.010
	$f(r\|\theta>0)$	0.099	0.1005	0.1005	0.700
$\Pr(\theta=0)=1/2$	$\Pr(\theta=0\|r)$	0.908	0.047	0.047	0.014
$\Pr(\theta>0)=1/2$	$\Pr(\theta>0\|r)$	0.091	0.953	0.953	0.986

Table 3.2 Recasting a simple versus composite Bayes test as simple versus simple.

3.3 Composite versus Composite

Now we add a fourth distribution to our consideration, $f(r|-1)$, and test the composite null $H_0 : \theta \leq 0$ versus the composite alternative $H_A : \theta > 0$ or, more specifically, $H_0 : \theta \in \{-1,0\} \equiv \Theta_0$ versus $H_A : \theta \in \{1,2\} \equiv \Theta_A$. The distributions and likelihood ratios are given below.

r	1	2	3	4
$f(r\|-1)$	0.9803	0.0049	0.0049	0.0099
$f(r\|0)$	0.980	0.005	0.005	0.010
$f(r\|1)$	0.100	0.200	0.200	0.500
$f(r\|2)$	0.098	0.001	0.001	0.900
$f(r\|1)/f(r\|-1)$	0.102	40.82	40.82	50.51
$f(r\|2)/f(r\|-1)$	0.109	0.204	0.204	90.9
$f(r\|1)/f(r\|0)$	10/98	40	40	50
$f(r\|2)/f(r\|0)$	0.1	0.2	0.2	90

It is hard to make comparisons with significance testing because the composite hypotheses do not provide us with a model that determines a unique distribution for the data. One could make four different significance tests. For this example, $f(r|-1)$ was chosen to be a minor modification of $f(r|0)$.

3.3.1 Neyman-Pearson Testing

As always N-P theory (quite properly) focuses on likelihood ratios. (Problems arise with what N-P theory does with these ratios.) The example was chosen so that for every $\theta_0 \in \Theta_0$ and every $\theta_1 \in \Theta_A$ as r varies we have an identical ordering to the values of $f(r|\theta_1)/f(r|\theta_0)$. As we have seen earlier, this means we can find UMP tests. In particular, regardless of hypotheses, if $\theta_* < 0.5 < \theta_\#$, our likelihood ratios $f(r|\theta_\#)/f(r|\theta_*)$ are all monotone increasing, so UMP tests can be formed by rejecting for large values of r (perhaps with some randomization on the smallest r value). The trick is to find/define the size of these tests.

For a composite null $H_0 : \theta \in \Theta_0$, the size of a test is defined to be the largest probability for the rejection region among all $\theta \in \Theta_0$. Thus, if we reject when $r = 4$, the size for $\theta = 0$ is 0.01 and the size for $\theta = -1$ is 0.0099, so the overall size is the

maximum value, 0.01. For the rejection region $r = 2,3,4$, the size for $\theta = 0$ is 0.02 and the size for $\theta = -1$ is $0.0049 + 0.0049 + 0.99 = 0.197$, so the overall size is the maximum value, 0.02.

3.3.2 Bayesian Testing

Again a (nontrivial) Bayesian approach requires positive prior probabilities on both Θ_0 and Θ_A and specifying conditional distributions for the θs given either the null or the alternative. Again, the easiest way to achieve this is putting a prior distribution on Θ for which $0 < \Pr(\Theta_0) < 1$. This time we illustrate the prior $p(\theta) = 1/4$ for all θ, so $\Pr(\Theta_0) = \Pr(\Theta_A) = 1/2$ and $p(\theta|\Theta_i) = 1/2$ for allowable θs when $i = 0, A$.

Using Bayes Theorem,

$$p(\theta|r) = \frac{f(r|\theta)p(\theta)}{f(r|-1)p(-1) + f(r|0)p(0) + f(r|1)p(1) + f(r|2)p(2)},$$

the posterior probabilities become

r	1	2	3	4
$p(-1\|r)$	0.454	0.0235	0.0235	0.007
$p(0\|r)$	0.454	0.0235	0.0235	0.007
$p(1\|r)$	0.046	0.948	0.948	0.352
$p(2\|r)$	0.045	0.005	0.005	0.634
$\Pr[\theta > 0\|r]$	0.091	0.953	0.953	0.986

With $f(r|-1)$ so similar to $f(r|0)$ and equal probabilities $p(-1) = p(0) = 1/4$, the values of $\Pr(\Theta_0|r)$ from the last example have essentially been retained in this example but now $p(-1|r) \doteq p(0|r) \doteq \Pr(\Theta_0|r)/2$. **(I didn't actually do the arithmetic for the table but this "has to be" true.)**

As with the simple versus composite case, the composite versus composite Bayesian test can be recast as a simple versus simple test. As before, one finds the average data distribution under the alternative but now, rather than having a single data distribution under the null, the average alternative is tested against the average data distribution under the null. With the prior we chose, each of these is just a simple average of the probabilities. Table 3.3 illustrates the computations. The posterior probabilities in the table roundoff to the same values as in Table 3.2, even though they are slightly different, because I choose an $f(r|-1)$ extremely similar to $f(r|0)$.

3.4 More on Neyman-Pearson Tests

To handle more general testing situations, N-P theory has developed a variety of concepts such as unbiased tests, invariant tests, and α similar tests, see Chapter 7 or

Prior	r	1	2	3	4
	$f(r\|\theta \le 0)$	0.98015	0.00495	0.00495	0.00995
	$f(r\|\theta > 0)$	0.099	0.1005	0.1005	0.700
$\Pr(\theta \le 0) = 1/2$	$\Pr(\theta \le 0\|r)$	0.908	0.047	0.047	0.014
$\Pr(\theta > 0) = 1/2$	$\Pr(\theta > 0\|r)$	0.092	0.953	0.953	0.986

Table 3.3 Recasting a composite versus composite Bayes test as simple versus simple.

Lehmann (1997). For example, the one and two sample t tests are not a uniformly most powerful tests but are uniformly most powerful unbiased tests. Similarly, the standard F test in regression and analysis of variance is a uniformly most powerful invariant test.

Similar to significance testing, the N-P approach to finding confidence regions is also to find parameter values that would not be rejected by a α level test. However, just as N-P theory interprets the size α of a test as the long run frequency of rejecting a correct null hypothesis, N-P theory interprets the confidence $1-\alpha$ as the long run probability of these regions including the true parameter. The rub is that you only have one of the regions, not a long run of them, and you are trying to say something about this parameter based on these data. In practice, the long run frequency of α somehow gets turned into something called "confidence" that this parameter is within this particular region.

While I admit that the term "confidence," as commonly used, feels good, I have no idea what "confidence" really means as applied to the region at hand. Hubbard and Bayarri (2003) make a case, implicitly, that an N-P concept of confidence would have no meaning as applied to the region at hand, that it only applies to a long run of similar intervals. Students, almost invariably, interpret confidence as posterior probability. For example, if we were to flip a coin many times, about half of the time we would get heads. If I flip a coin and look at it but do not tell you the result, you may feel comfortable saying that the chance of heads is still 0.5 even though I know whether it is heads or tails. Somehow the probability of what is going to happen in the future is turning into confidence about what has already happened but is unobserved. Since I do not understand how this transition from probability to confidence is made (unless one is a Bayesian in which case confidence actually is probability), I do not understand "confidence."

3.5 More on Bayesian Testing

Bayesian tests can go seriously wrong only if you pick inappropriate prior distributions. This is the case in Lindley's famous paradox in which, for a seemingly simple and reasonable testing situation involving normal data, the null hypothesis is accepted no matter how weird the observed data are relative to the null hypothesis. The datum is $X|\mu \sim N(\mu,1)$. The test is $H_0 : \mu = 0$ versus $H_A : \mu > 0$. The priors on the hypotheses do not really matter, but take $\Pr[\mu = 0] = 0.5$ and $\Pr[\mu > 0] = 0.5$ In

an attempt to use a noninformative prior, take the density of μ given $\mu > 0$ to be flat on the half line. (This is an improper prior but similar proper priors lead to similar results.) The Bayesian test compares the density of the data X under $H_0 : \mu = 0$ to the average density of the data under $H_A : \mu > 0$. (The latter involves integrating the density of $X|\mu$ times the density of μ given $\mu > 0$.) The average density under the alternative makes any X you could possibly see, infinitely more probable to have come from the null distribution than from the alternative. Thus, anything you could possibly see will cause you to accept $\mu = 0$. Attempting to have a noninformative prior on the half line leads one to a nonsensical prior that effectively puts all the probability on unreasonably large values of μ so that, by comparison, $\mu = 0$ always looks more reasonable.

3.6 Hypothesis Test *P* Values

The definition of a P value for a significance test is straight forward, it is the probability of seeming something as weird or weirder than you actually saw. The probability is computed under the only distribution you have and that density of that distribution is used to define how weird an observation is. For an hypothesis test, you have at least two distributions and one again needs to define weird.

For a simple versus simple hypothesis test the standard definition of a P value is to compute the probability under the null distribution and to define weird in a relative sense as having a large value of the likelihood ratio (alternative density divided by null). This idea will also work for simple versus composite hypotheses for which UMP tests exist. However a key feature is that weird observations are not weird in any absolute sense, they are only weird relative to the alternative hypothesis. As such, and as we have seen, a small P value in this sense does not necessarily either contradict the null hypothesis, cf. Example 4.0.1, nor does it suggest that the alternative is more likely to be true, cf. Section 3.1 with $\alpha = .02$.

For more complicated problems, the idea of an hypothesis test P value is to find the α level for which the test would just barely reject. But that requires one to have a collection of tests indexed by α for which the rejections regions are getting smaller as α decreases in a continuous way so that the data always fall into some rejection region and a smallest rejection region exists.

Lehmann and Romano's (2005) most general definition of a P value is that if you have a collection of tests ϕ_α indexed by their size and if those tests have the property that $\phi_{\alpha_1}(x) \le \phi_{\alpha_2}(x)$ for any x and $\alpha_1 \le \alpha_2$, then the P value for this collection of tests is defined to be $P \equiv \inf\{\alpha|\phi_\alpha(X) = 1\}$. This makes P the smallest level of significance for which the null is rejected with probability 1. Flip an α coin has $P = 0$.

They also define for nonrandomized tests with nested rejection regions $R_{\alpha_1} \subset R_{\alpha_2}$ for any $\alpha_1 \le \alpha_2$, $P \equiv \inf\{\alpha|X \in R_\alpha\}$

3.7 Permutation Tests

Are they hypothesis tests rather than significance tests because they require an alternative? Probably an alternative based on stochastic inequality. The issue is that all outcomes would have equal probabilities so you need some outside definition of what constitutes “weird.”

Chapter 4
Comparing Testing Procedures

Significance tests and hypothesis tests are very different procedures. The clearest example of this, that I know of, is the following.

EXAMPLE 4.0.1. Consider a significance test of the null model $y \sim N(0,1)$ based on one observation. The density decreases as one gets further from the mean 0, so large $|y|$ values constitute evidence against the null model. In particular, an $\alpha = 0.05$ significance test rejects for $|y| \geq 1.96$.

Now consider testing $H_0: \ y \sim N(0,1)$ versus $H_A: \ y \sim N(0,\sigma^2), \ \sigma^2 < 1$. Figure 4.1 illustrates the densities $f(y|\sigma^2)$. With $\sigma^2 < 1$ the density of a $N(0,\sigma^2)$ is

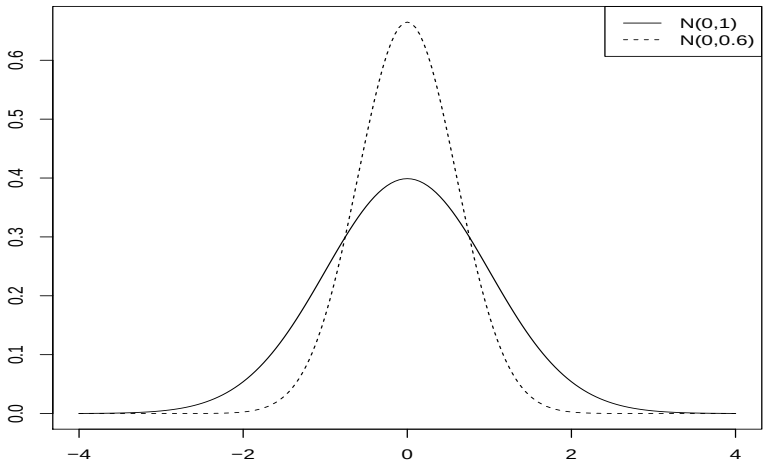

Fig. 4.1 $N(0,1)$ and $N(0,0.6)$ densities.

higher near 0 than the density of the $N(0,1)$ which means that the N-P hypothesis test for an $\alpha = 0.05$ test will reject for values close to 0, in particular it will reject for $|y| \leq 0.063$. Could two tests for the same null model be any more different? Seeing a small value of $|y|$ provides no evidence against the model $N(0,1)$ in any absolute sense, such values are merely more consistent with the alternative than they are with the null.

More technically, the density (likelihood) ratio is

$$\frac{f(y|\sigma^2)}{f(y|1)} = \frac{(1/2\pi\sigma)e^{-y^2/2\sigma^2}}{(1/2\pi\sigma)e^{-y^2/2}} = e^{-y^2/2\sigma^2+y^2/2} = \exp\left[-(y^2/2)\left(\frac{1}{\sigma^2}-1\right)\right],$$

which, for any value of $\sigma^2 < 1$, is maximized at 0 and decreases as $|y|$ gets further from 0. So according to the N-P lemma, the most powerful test for any particular $\sigma^2 < 1$, rejects for the smallest values of $|y|$. This holds regardless of the particular value of σ^2, hence the test is uniformly most powerful. In Chapter 7 it is an exercise to show that this is a UMP test.

The same moral can be learned from discrete distributions. Our simple versus composite hypothesis example of the previous chapter included the distributions

r	1	2	3	4
$f(r\|1)$	0.100	0.200	0.200	0.500
$f(r\|2)$	0.098	0.001	0.001	0.900
$f(r\|2)/f(r\|1)$	0.98	0.0005	0.0005	1.8

Now consider an N-P test of $H_0 : \theta = 1$ versus $H_A : \theta = 2$. The N-P Lemma indicates that we should most readily reject for the data point with the highest likelihood ratio, $r = 4$, which is precisely the data point that is most consistent with the null hypothesis. (For any N-P test with $\alpha < 0.5$, a randomized test is needed.)

4.1 Discussion

The basic elements of a significance test are: (1) There is a probability model for the data. (2) Multidimensional data are summarized into a test statistic that has a known distribution. (3) This known distribution provides a ranking of the "weirdness" of various observations. (4) The P value, which is the probability of observing something as weird or weirder than was actually observed, is used to quantify the evidence against the null hypothesis. (5) α level tests are defined by reference to the P value.

The basic elements of an N-P test are: (1) There are two (sets of) hypothesized models for the data: H_0 and H_A. (2) An α level is chosen which is to be the (maximum) probability of rejecting H_0 when H_0 is true. (3) A rejection region is chosen so that the probability of data falling into the rejection region is (at most) α when H_0 is true. With discrete data, this often requires the specification of a randomized rejection region in which certain data values are randomly assigned to be in or out of

the rejection region. (4) Various tests are evaluated based on their power properties. Ideally, one wants the most powerful test. (5) In complicated problems, properties such as unbiasedness or invariance are used to restrict the class of tests prior to choosing a test with good power properties.

Significance testing seems to be a reasonable approach to model validation. In fact, Box (1980) suggested significance tests, based on the marginal distribution of the data, as a method for validating Bayesian models. Significance testing is philosophically based on the idea of proof by contradiction in which the contradiction is not absolute.

Bayesian testing seems to be a reasonable approach to making a decision between alternative hypotheses. The results are influenced by the prior distributions, but one can examine a variety of prior distributions.

Neyman-Pearson testing seems to be neither fish nor fowl. It seems to mimic significance testing with its emphasis on the null hypothesis and small α levels, but it also employs an alternative hypothesis, so it is not based on proof by contradiction as is significance testing. Because N-P testing focuses on small α levels, it often leads to bad decisions between the two alternative hypotheses. Certainly, for simple versus simple hypotheses, any problems with N-P testing vanish if one is not philosophically tied down to small α values. For example, any reasonable test (as judged by frequentist criteria) must be within both the collection of all most powerful tests and the collection of all Bayesian tests, see Ferguson (1967, p. 204).

Although most problems with testing seem to stem from choosing too small an α at the expense of creating very large probabilities of type II error (β), we have seen an example where a decrease in α was appropriate because it barely increased β.

There is also the issue of whether α is merely a measure of how weird the data are, or whether is should be interpreted as the probability of making the wrong decision about the null. If α is the probability of making an incorrect decision about the null, then performing multiple tests to evaluate a composite null causes problems because it changes the overall probability of making the wrong decision. If α is merely a measure of how weird the data are, it is less clear that multiple testing inherently causes any problem. In particular, Fisher (1935, Chp. 24) did not worry about the experimentwise error rate when making multiple comparisons using his "least significant difference" method in analysis of variance. He did, however, worry about drawing inappropriate conclusions by using an invalid null distribution for tests determined by examining the data.

In significance testing the P value is well defined and an α level test is defined in terms of the P value. In hypothesis testing, an α level test is well defined and some people want to define P values for hypothesis tests. We have seen that significance tests and hypothesis tests are fundamentally different creatures, so any hypothesis testing P value, needs to have a different definition than a significance testing P value. Indeed, for all the restrictions one may need to place on N-P tests (composite α value, unbiasedness, invariance), ultimately N-P tests are trying to reject for values with large likelihood ratios. So a reasonable definition of an hypothesis testing P value will have to measure the weirdness of data by how much the likelihood ratio

favors the alternative over the null. But this will always lead to a choice between hypotheses and not a contradiction to the null model.

In the late 20-teens, it became fashionable to criticize *NHST (Null Hypothesis Significance Testing)*. Unfortunately, NHST is something of a straw man. Over the years many statisticians have conflated significance and hypothesis testing into NHST, which blurs the very important distinctions between the two methodologies. (Even the name NHST conflates the methodologies because a significance test is for a given model not a null hypothesis.) [Our use of the term "null model" is a capitulation to the prevalence of NHST.] The problem is exacerbated by the fact that the most commonly taught tests happen to be instances wherein the differences between significance and hypothesis testing are easily glossed over.

Exercise 4.1. Discuss the problems of significance testing $f(x) = .01, x = 1,\ldots,95, f(0) = .001, f(96) = .049, f(97) = 0$ and of hypothesis testing f against $g(x) = .001, x = 1,\ldots,95, g(0) = 0, g(96) = .1, g(97) = .805$.

4.2 Jeffreys' Critique

A test of significance examines not only the event that occurred, but discrepant events that did not occur. Jeffreys's (1961) criticized the procedure saying,

> "What the use of the P value implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred."

While that sounds good, its meaning is not clear. In the context of testing a null probability model, it seems hardly to apply. If a probability model were to predict different observable results, it would be a different probability model, and thus it cannot be true. From the point of view of testing a parameter, the quotation makes more sense. Suppose we observe $X = 2$ with $\mathrm{E}(X) = \theta$ and we are testing $H_0 : \theta = 0$. Whether we reject the hypothesis depends on assumptions we make about unobserved quantities. If we assume $X \sim N(\theta, 1)$, values of X greater than 2 units from 0 are not predicted to occur often, so we could reject the hypothesis with $P = .045$. If $X - \theta \sim t(2)$, $P = .18$ which is unlikely to make us reject the null. While this example fits the context of Jeffreys's statement, it does not seem a very damning criticism of significance tests.

Jeffreys's criticism is far more meaningful when applied to tests that involve the specification of an alternative hypothesis. In those cases it is appropriate to base conclusions on which alternative is more likely to have generated the observed data. In such a case, it seems ludicrous to incorporate into any conclusion the relative likelihoods of data that were not observed.

Discuss the stopping rule principal. Binomial versus negative binomial P values. Savage's quote (mentioned in Barnard conversation(?) and probably Rereading Fisher). Jessica's ESP experiment.

Chapter 5
Decision Theory

Decision theory is a very general theory that allows one to examine Bayesian estimation and hypothesis testing as well as Neyman-Pearson hypothesis testing and many aspects of frequentist estimation. I am not aware that it has anything to say about Fisherian significance testing.

In decision theory we start with *states of nature* $\theta \in \Theta$, potential *actions* $a \in \mathscr{A}$, and a *loss function* $L(\theta, a)$ that takes real values. We are interesting in taking actions that will reduce our losses. Some formulations of decision theory incorporate a utility function $U(\theta, a)$ and seek actions that increase utility. The formulations are interchangeable by simply taking $U(\theta, a) = -L(\theta, a)$.

Eventually, we will want to incorporate data in the form of a random vector X taking values in $\mathscr{X}$ and having density $f(x|\theta)$. The distribution of $X|\theta$ is called the *sampling distribution*.

We will focus on three special cases.

Estimation of a scalar state of nature involves scalar actions with $\Theta = \mathscr{A} = \mathbf{R}$. Three commonly used loss functions are

- Squared error, $L(\theta, a) = (\theta - a)^2$;
- Weighted squared error, $L(\theta, a) = w(\theta)(\theta - a)^2$, wherein $w(\theta)$ is a known weighting function taking positive values;
- Absolute error, $L(\theta, a) = |\theta - a|$.

Estimation of a vector involves $\Theta = \mathscr{A} = \mathbf{R}^d$. Three commonly used loss functions are

- $L(\theta, a) = (\theta - a)'(\theta - a) \equiv \|\theta - a\|^2$;
- $L(\theta, a) = w(\theta)\|\theta - a\|^2$, with known $w(\theta) > 0$;
- $L(\theta, a) = \sum_{j=1}^{d} |\theta_j - a_j|$.

Hypothesis testing involves two hypotheses, say $\Theta = \{\theta_0, \theta_1\}$, and two corresponding actions $\mathscr{A} = \{a_0, a_1\}$. What is key in this problem is that there are only two states of nature in Θ that we can think of as the null and alternative hypotheses, respectively, and two corresponding actions in $\mathscr{A}$ that we can think of as accepting

the null (rejecting the alternative) and accepting the alternative (rejecting the null). The *standard loss function* is

$L(\theta,a)$	a_0	a_1
θ_0	0	1
θ_1	1	0

A more general loss function is

$L(\theta,a)$	a_0	a_1
θ_0	c_{00}	c_{01}
θ_1	c_{10}	c_{11}

wherein, presumably, $c_{00} \leq c_{01}$ and $c_{10} \geq c_{11}$.

More generally in hypothesis testing we can partition a more general Θ into Θ_0 (the null hypothesis) and Θ_1 (the alternative hypothesis) with only two actions $\mathscr{A} = \{a_0, a_1\}$ and the standard loss function becomes

$L(\theta,a)$	a_0	a_1
$\theta \in \Theta_0$	0	1
$\theta \in \Theta_1$	1	0.

Again, a_1 is taken to mean rejecting the null hypothesis and a_0 is taken to mean accepting the null hypothesis. To reject is to 'not accept' and to 'not accept' is to reject. (Recall that in Significance Testing, not rejecting the null is different from accepting it and there is no formal alternative to accept.) The use of a_1 when $\theta \in \Theta_0$, i.e., rejecting the null hypothesis when it is true, is called a *Type I error*. Using a_0 when $\theta \in \Theta_1$, i.e., not rejecting/accepting the null hypothesis when it is false, is called a *Type II error*.

5.1 Optimal Prior Actions

If θ is random, i.e., if θ has a *prior distribution*, then the optimal action is defined to be the action that minimizes the expected loss, $\mathrm{E}[L(\theta,a)] \equiv \mathrm{E}_\theta[L(\theta,a)]$

Proposition 5.1.1. For $\Theta = \mathscr{A} = \mathbf{R}$ and $L(\theta,a) = (\theta - a)^2$, if θ is random, the optimal action is $\hat{a} = \mathrm{E}(\theta)$.

PROOF: It is enough to show that

$$\mathrm{E}[(\theta - a)^2] = \mathrm{E}[(\theta - \hat{a})^2] + (\hat{a} - a)^2$$

because then the minimizing value of a occurs when $\hat{a} = a$.

As is so often the case, the proof proceeds by subtracting and adding the correct answer.

$$\begin{aligned}
\mathrm{E}[(\theta-a)^2] &= \mathrm{E}[(\{\theta-\hat{a}\}+\{\hat{a}-a\})^2] \\
&= \mathrm{E}[(\theta-\hat{a})^2]+2\mathrm{E}[(\theta-\hat{a})(\hat{a}-a)]+\mathrm{E}[(\hat{a}-a)^2] \\
&= \mathrm{E}[(\theta-\hat{a})^2]+2(\hat{a}-a)\mathrm{E}[(\theta-\hat{a})]+(\hat{a}-a)^2 \\
&= \mathrm{E}[(\theta-\hat{a})^2]+(\hat{a}-a)^2
\end{aligned}$$

The third equality holds because $(\hat{a}-a)^2$ is a constant and the fourth holds because $\mathrm{E}[\theta-\mathrm{E}(\theta)]=0$. □

Proposition 5.1.2. For $\Theta=\mathscr{A}=\mathbf{R}$ and $L(\theta,a)=w(\theta)(\theta-a)^2$ with $w(\theta)>0$, if θ is random, the optimal action is $\hat{a}=\mathrm{E}[w(\theta)\theta]/\mathrm{E}[w(\theta)]$.

PROOF: The proof is an exercise. Write

$$\mathrm{E}[w(\theta)(\theta-a)^2]=\mathrm{E}[w(\theta)(\theta-\hat{a}+\hat{a}-a)^2].$$

□

Proposition 5.1.3. For $\Theta=\mathscr{A}=\mathbf{R}$ and $L(\theta,a)=|\theta-a|$, if θ is random, an optimal action is $\hat{a}=m\equiv\mathrm{Median}(\theta)$. Any median is optimal.

PROOF: Without loss of generality assume a is greater than the median m of θ under consideration so that $p_a\equiv P[\theta>a]\le 0.5$. By the definition of the median, $p_m\equiv P[\theta\le m]\ge 0.5$. (A median m also has $P[\theta\ge m]\ge 0.5$ with the inequalities used to cope with discrete distributions.) For discrete distributions we take $\int_c^d$ to mean $\int \mathscr{I}_{(c,d]}$ with $d=\infty$ irrelevant because $P[\theta=\infty]=0$. As with many proofs, ours involves adding and subtracting the correct item, but here we have to do it three times.

$$\begin{aligned}
\mathrm{E}[|\theta-a|] &= \int |\theta-a|\,dP(\theta) \\
&= \int_a^\infty (\theta-a)\,dP(\theta)+\int_m^a (a-\theta)\,dP(\theta)+\int_{-\infty}^m (a-\theta)\,dP(\theta) \\
&= \int_a^\infty (\theta-a)\,dP(\theta)+\int_a^\infty (a-m)\,dP(\theta)+\int_a^\infty (m-a)\,dP(\theta) \\
&\quad +\int_m^a (a-\theta)\,dP(\theta)+\int_m^a (\theta-m)\,dP(\theta)+\int_m^a (m-\theta)\,dP(\theta) \\
&\quad +\int_{-\infty}^m (a-\theta)\,dP(\theta)+\int_{-\infty}^m (m-a)\,dP(\theta)+\int_{-\infty}^m (a-m)\,dP(\theta) \\
&= \int_a^\infty (\theta-m)\,dP(\theta)+\int_a^\infty (m-a)\,dP(\theta) \\
&\quad +\int_m^a (\theta-m)\,dP(\theta)+\int_m^a (m+a-2\theta)\,dP(\theta) \\
&\quad +\int_{-\infty}^m (m-\theta)\,dP(\theta)+\int_{-\infty}^m (a-m)\,dP(\theta)
\end{aligned}$$

$$
\begin{aligned}
&= \int |\theta - m| \, dP(\theta) \\
&\quad + \int_a^{\infty} (m-a) \, dP(\theta) + \int_m^a (m+a-2\theta) \, dP(\theta) + \int_{-\infty}^m (a-m) \, dP(\theta) \\
&= \int |\theta - m| \, dP(\theta) + p_a(m-a) + \int_m^a (m+a-2\theta) \, dP(\theta) + p_m(a-m) \\
&= \int |\theta - m| \, dP(\theta) + (p_m - p_a)(a-m) + \int_m^a (m+a-2\theta) \, dP(\theta) \\
&\geq \int |\theta - m| \, dP(\theta) + (p_m - p_a)(a-m) + \int_m^a (m-a) \, dP(\theta) \\
&= \int |\theta - m| \, dP(\theta) + (p_m - p_a)(a-m) + (m-a)P[m < \theta \leq a] \\
&= \int |\theta - m| \, dP(\theta) + (p_m - p_a)(a-m) + [(1-p_a) - p_m](m-a) \\
&= \int |\theta - m| \, dP(\theta) + (p_m - p_a)(a-m) - [(1-p_a) - p_m](a-m) \\
&= \int |\theta - m| \, dP(\theta) + (2p_m - 1)(a-m) \\
&\geq \int |\theta - m| \, dP(\theta) \\
&= \mathrm{E}[|\theta - m|],
\end{aligned}
$$

where the first inequality holds because over the range of integration $m+a-2\theta \geq m+a-2a$ and the second inequality holds because, by definition, $p_m \geq 0.5$ and, by assumption, $a \geq m$.

The proof for $a < m$ is similar. □

Proposition 5.1.4. For $\Theta = \{\theta_0, \theta_1\}$, $\mathscr{A} = \{a_0, a_1\}$, and the standard loss function, the optimal action is

$$
\hat{a} = \begin{cases} a_0 & \text{if } \Pr(\theta = \theta_0) > 0.5, \\ a_1 & \text{if } \Pr(\theta = \theta_0) < 0.5. \end{cases}
$$

If $\Pr(\theta = \theta_0) = 0.5$, both actions are optimal.

PROOF: Note that

$$
\mathrm{E}[L(\theta, a_0)] = L(\theta_0, a_0)\Pr(\theta = \theta_0) + L(\theta_1, a_0)\Pr(\theta = \theta_1) = \Pr(\theta = \theta_1)
$$

and

$$
\mathrm{E}[L(\theta, a_1)] = L(\theta_0, a_1)\Pr(\theta = \theta_0) + L(\theta_1, a_1)\Pr(\theta = \theta_1) = \Pr(\theta = \theta_0).
$$

If $\Pr(\theta = \theta_1) < \Pr(\theta = \theta_0)$ the optimal action is a_0 and if $\Pr(\theta = \theta_1) > \Pr(\theta = \theta_0)$ the optimal action is a_1. However, $\Pr(\theta = \theta_0) + \Pr(\theta = \theta_1) = 1$, so $\Pr(\theta = \theta_1) < \Pr(\theta = \theta_0)$ if and only if $\Pr(\theta = \theta_0) > 0.5$ □

The final result is a generalization of Propostion 5.1.3 that establishes that quantiles/percentiles other than the median can be optimal actions for an appropriate loss function.

Proposition 5.1.5. For $\Theta = \mathscr{A} = \mathbf{R}$ and $L(\theta,a) = (1-\alpha)(a-\theta)_+ + \alpha(\theta-a)_+$ where $(x)_+$ is x when x is nonnegative and 0 when it is negative, and θ random, an optimal action $\hat{a}$ is any α quantile/percentile of the distribution of θ.

PROOF: The proof is similar to that for Propostion 5.1.3 but more involved. I have broken it into more manageable pieces. Of course $\alpha = 0.5$ is the special case addressed earlier. For notational simplicity and similarity with the proof for absolute error, we denote the α quantile as m rather than the more commonly used notation q_α.

First assume a is greater than the α quantile m of θ so that $F(m) \le F(a)$ where F is the cdf of θ. By the definition of the alpha quantile, $F(m) \equiv P[\theta \le m] \ge \alpha$ and $P[\theta \ge m] \ge 1-\alpha$ with the inequalities used to cope with discrete distributions. For discrete distributions we take $\int_c^d$ to mean $\int \mathscr{I}_{(c,d]}$ with $d = \infty$ irrelevant because $P[\theta = \infty] = 0$.

$$
\begin{aligned}
&\mathrm{E}[(1-\alpha)(a-\theta)_+ + \alpha(\theta-a)_+] \\
&= \int [(1-\alpha)(a-\theta)_+ + \alpha(\theta-a)_+]\, dP(\theta) \\
&= \alpha \int_a^\infty (\theta-a)\, dP(\theta) + (1-\alpha)\int_m^a (a-\theta)\, dP(\theta) + (1-\alpha)\int_{-\infty}^m (a-\theta)\, dP(\theta)
\end{aligned}
$$

We are going to look at these three terms separately and then put them back together.

The first term reduces to

$$
\begin{aligned}
&\alpha \int_a^\infty (\theta-a)\, dP(\theta) \\
&= \alpha \int_a^\infty (\theta-a)\, dP(\theta) + \alpha \int_a^\infty (a-m)\, dP(\theta) + \alpha \int_a^\infty (m-a)\, dP(\theta) \\
&= \alpha \int_a^\infty (\theta-m)\, dP(\theta) + \alpha \int_a^\infty (m-a)\, dP(\theta) \\
&= \alpha \int_a^\infty (\theta-m)\, dP(\theta) + \alpha[1-F(a)](m-a)
\end{aligned}
$$

The second term satisfies

$$
\begin{aligned}
&(1-\alpha)\int_m^a (a-\theta)\, dP(\theta) \\
&= (1-\alpha)\int_m^a (a-\theta)\, dP(\theta) \\
&\quad + (1-\alpha)\int_m^a (\theta-m)\, dP(\theta) + (1-\alpha)\int_m^a (m-\theta)\, dP(\theta)
\end{aligned}
$$

$$
\begin{aligned}
&= (1-\alpha)\int_m^a (\theta - m)\,dP(\theta) + (1-\alpha)\int_m^a (m + a - 2\theta)\,dP(\theta) \\
&= \alpha \int_m^a (\theta - m)\,dP(\theta) \\
&\quad + (1-2\alpha)\int_m^a (\theta - m)\,dP(\theta) + (1-\alpha)\int_m^a (m + a - 2\theta)\,dP(\theta) \\
&\geq \alpha \int_m^a (\theta - m)\,dP(\theta)
\end{aligned}
$$

where the last inequality holds because in the previous relation the sum of the last two terms is nonnegative. To see this,

$$
\begin{aligned}
&(1-2\alpha)\int_m^a (\theta - m)\,dP(\theta) + (1-\alpha)\int_m^a (m + a - 2\theta)\,dP(\theta) \\
&= \int_m^a (\theta - m + m + a - 2\theta)\,dP(\theta) - \int_m^a (2\alpha\theta - 2\alpha m + \alpha m + \alpha a - 2\alpha\theta)\,dP(\theta) \\
&= \int_m^a (a - \theta)\,dP(\theta) - \alpha \int_m^a (a - m)\,dP(\theta) \\
&\geq \int_m^a (a - m)\,dP(\theta) - \alpha \int_m^a (a - m)\,dP(\theta) \\
&= (1-\alpha)\int_m^a (a - m)\,dP(\theta) = (1-\alpha)[F(a) - F(m)](a - m) \geq 0.
\end{aligned}
$$

The third term has

$$
\begin{aligned}
&(1-\alpha)\int_{-\infty}^m (a - \theta)\,dP(\theta) \\
&= (1-\alpha)\int_{-\infty}^m (a - \theta)\,dP(\theta) + (1-\alpha)\int_{-\infty}^m (m - a)\,dP(\theta) + (1-\alpha)\int_{-\infty}^m (a - m)\,dP(\theta) \\
&= (1-\alpha)\int_{-\infty}^m (m - \theta)\,dP(\theta) + (1-\alpha)\int_{-\infty}^m (a - m)\,dP(\theta) \\
&= (1-\alpha)\int_{-\infty}^m (m - \theta)\,dP(\theta) + (1-\alpha)F(m)(a - m)\,dP(\theta)
\end{aligned}
$$

Putting these three together gives

$$
\begin{aligned}
&\mathrm{E}[(1-\alpha)(a-\theta)_+ + \alpha(\theta - a)_+] \\
&= \alpha \int_a^\infty (\theta - a)\,dP(\theta) + (1-\alpha)\int_m^a (a - \theta)\,dP(\theta) + (1-\alpha)\int_{-\infty}^m (a - \theta)\,dP(\theta) \\
&\geq \alpha \int_a^\infty (\theta - m)\,dP(\theta) + \alpha[1 - F(a)](m - a) \\
&\quad + \alpha \int_m^a (\theta - m)\,dP(\theta) \\
&\quad + (1-\alpha)\int_{-\infty}^m (m - \theta)\,dP(\theta) + (1-\alpha)F(m)(a - m)\,dP(\theta)
\end{aligned}
$$

$$= \mathrm{E}[(1-\alpha)(m-\theta)_+ + \alpha(\theta-m)_+] + \alpha[1-F(a)](m-a) + (1-\alpha)F(m)(a-m)$$
$$\geq \mathrm{E}[(1-\alpha)(m-\theta)_+ + \alpha(\theta-m)_+].$$

The last inequality holds because

$$\alpha[1-F(a)](m-a) + (1-\alpha)F(m)(a-m) \geq 0.$$

This is true because it is equivalent to having

$$(1-\alpha)F(m) \geq \alpha[1-F(a)],$$

which holds because $(1-\alpha) \geq [1-F(a)]$ and $F(m) \geq \alpha$.

To prove the case with $a < m$, you need to redefine $\int_c^d dP(\theta) \equiv \int \mathscr{I}_{[c,d)}\, dP(\theta)$, which leads to redefining F as $F(a) = \Pr[\theta < a]$ so that now the alpha quantile has $1-F(m) \geq 1-\alpha$.

$$\mathrm{E}[(1-\alpha)(a-\theta)_+ + \alpha(\theta-a)_+]$$
$$= \alpha \int_m^\infty (\theta-a)\, dP(\theta) + \alpha \int_a^m (\theta-a)\, dP(\theta) + (1-\alpha) \int_{-\infty}^a (a-\theta)\, dP(\theta)$$

Showing that

$$\alpha \int_a^m (\theta-a)\, dP(\theta) \geq (1-\alpha) \int_a^m (m-\theta)\, dP(\theta)$$

is left as an exercise. For the first and third terms, you also need to prove that

$$\alpha[1-F(m)](m-a) + (1-\alpha)F(a)(a-m) \geq 0,$$

which is equivalent to

$$\alpha[1-F(m)] \geq (1-\alpha)F(a),$$

which follows because $\alpha \geq F(a)$ and $[1-F(m)] \geq (1-\alpha)$. □

5.2 Optimal Posterior Actions

Suppose we have a data vector X with density $f(u|\theta)$. If θ is random, i.e., if θ has a prior density $p(\theta)$, a Bayesian updates the distribution of θ using the data and Bayes' Theorem to get the posterior density

$$p(\theta|X=u) = \frac{f(u|\theta)p(\theta)}{\int f(u|\theta)p(\theta)d\mu(\theta)}.$$

The *Bayes action* is defined to be the action that minimizes the expected loss,

$$\mathrm{E}[L(\theta,a)|X=u] \equiv \mathrm{E}_{\theta|X=u}[L(\theta,a)].$$

The Bayes action is just the optimal action when the distribution on θ is the posterior distribution given X. Recognizing this fact, the previous section immediately provides four results.

Proposition 5.2.1. For $\Theta = \mathscr{A} = \mathbf{R}$, data $X = u$, and $L(\theta,a) = (\theta - a)^2$, if θ is random, the Bayes action is $\hat{a} = \mathrm{E}_{\theta|X=u}(\theta) \equiv \mathrm{E}(\theta|X=u)$.

Proposition 5.2.2. For $\Theta = \mathscr{A} = \mathbf{R}$, data $X = u$, and $L(\theta,a) = w(\theta)(\theta - a)^2$ with $w(\theta) > 0$, if θ is random, the Bayes action is $\hat{a} = \mathrm{E}[w(\theta)\theta|X=u]/\mathrm{E}[w(\theta)|X=u]$.

Proposition 5.2.3. For $\Theta = \mathscr{A} = \mathbf{R}$, data $X = u$, and $L(\theta,a) = |\theta - a|$, if θ is random, a Bayes action is any $\hat{a} = m \equiv \mathrm{Median}(\theta|X=u)$.

Proposition 5.2.4. For $\Theta = \{\theta_0, \theta_1\}$, data $X = u$, $\mathscr{A} = \{a_0, a_1\}$, $L(\theta,a) = I(\theta \neq a)$, a Bayes action has

$$\hat{a} = \begin{cases} a_0 & \text{if } \Pr(\theta = \theta_0|X=u) > 0.5 \\ a_1 & \text{if } \Pr(\theta = \theta_0|X=u) < 0.5 \end{cases}.$$

A similar result also holds for quantile estimation. In Section 5 we will see that predictions problems have the same structure and the same results.

5.3 Traditional Decision Theory

With states of nature $\theta \in \Theta$, potential actions $a \in \mathscr{A}$, and a data vector X taking values in $\mathscr{X}$ and having density $f(x|\theta)$, a *decision function (rule)* δ is defined as a mapping of the data into the action space, i.e.,

$$\delta : \mathscr{X} \to \mathscr{A}.$$

With a loss function $L(\theta,a)$, the *risk function* is defined as

$$R(\theta,\delta) \equiv \mathrm{E}_{X|\theta}\{L[\theta,\delta(X)]\}.$$

To frequentists, the risk function is the soul of decision theory. They would like to pick a δ that minimizes $R(\theta,\delta)$ uniformly in θ. That is very hard to do.

Uniformly minimum variance unbiased (UMVU) estimators of $h(\theta)$ use squared error loss, minimize $R(\theta,\delta)$ uniformly in θ, but restrict δ to rules with $\mathrm{E}_{X|\theta}[\delta(X)] = h(\theta)$.

In testing problems with the standard loss function, we would love to minimize $R(\theta,\delta)$ uniformly in θ but we cannot. For $\theta \in \Theta_0$, $R(\theta,\delta)$ is the *probability of Type I error*. In particular, since $\delta(X)$ only takes two values,

$$\begin{aligned} R(\theta,\delta) &= L(\theta,a_0)P_{X|\theta}[\delta(X)=a_0] + L(\theta,a_1)P_{X|\theta}[\delta(X)=a_1] \\ &= 0 + P_{X|\theta}[\delta(X)=a_1]. \end{aligned}$$

For $\theta \in \Theta_1$, $R(\theta,\delta)$ becomes the *probability of Type II error*. If Θ_0 and Θ_1 are not single points, both probabilities are functions of θ. For $\theta \in \Theta_0$, $R(\theta,\delta)$ is sometimes called the *size function* of the test. (The actual size of a test is usually taken as $\sup_{\theta\in\Theta_0} R(\theta,\delta)$.)

The *power* of the test δ is the probability of rejecting the null hypothesis when it is false (picking a_1 when $\theta \in \Theta_1$), and equals $1 - R(\theta,\delta)$ when $\theta \in \Theta_1$. As a function of θ, $P_{X|\theta}[\delta(X)=a_1]$ gives the size function when $\theta \in \Theta_0$ and the power function when $\theta \in \Theta_1$, so it provides the *size-power function*. *Uniformly most powerful (UMP)* tests minimize $R(\theta,\delta)$ uniformly for $\theta \in \Theta_1$, but restrict δ to rules with $R(\theta,\delta) \le \alpha$ for all $\theta \in \Theta_0$. Uniformly most powerful unbiased (UMPU) and uniformly most powerful invariant (UMPI) tests place *additional* restrictions on the δ rules that are considered.

The *Bayes risk* is a frequentist idea of what a Bayesian should worry about. With a prior distribution, call it p, on θ, the Bayes risk is defined as

$$r(p,\delta) \equiv \mathrm{E}[R(\theta,\delta)].$$

Frequentists think that Bayesians should be concerned about finding the *Bayes decision rule* that minimizes the Bayes risk. Formally, for a prior p, the Bayes rule is a decision function δ_p with

$$r(p,\delta_p) = \inf_{\delta} r(p,\delta).$$

As discussed in the previous section, Bayesians think that they should be concerned with finding the Bayes action given the data. Fortunately, these amount to the same thing. To minimize the Bayes risk, you pick $\delta(x)$ to minimize

$$\begin{aligned} r(p,\delta) &= \mathrm{E}[R(\theta,\delta)] \\ &= \mathrm{E}_\theta\left(\mathrm{E}_{X|\theta}\{L[\theta,\delta(X)]\}\right) \\ &= \mathrm{E}_X\left(\mathrm{E}_{\theta|X}\{L[\theta,\delta(X)]\}\right). \end{aligned}$$

This can be minimized by picking $\delta(x)$ to be the Bayes action that minimizes

$$\mathrm{E}_{\theta|X=x}\{L[\theta,\delta(x)]\}$$

for every value of x.

One exception to Bayesians being concerned about the Bayes action rather than the Bayes decision rule is when a Bayesian is trying to design an experiment, hence is concerned with possible data rather than already observed data.

We now introduce other basic concepts from decision theory.

Definition 5.3.1 The rule δ is *inadmissible* if there exists δ_* such that, for any θ, $R(\theta,\delta_*) \leq R(\theta,\delta)$ and there exists θ_* such that $R(\theta_*,\delta_*) < R(\theta_*,\delta)$. In such a case we say that δ_* is *better than* δ. The rule δ is *admissible* if it is not inadmissible, i.e., if no rule is better than it. Two rules δ_1 and δ_2 are *equivalent* if $R(\theta,\delta_1) = R(\theta,\delta_2)$ for all θ. A rule δ_1 is *as good as* δ_2 if it is either better than or equivalent to δ_2.

For a discrete Θ and a prior that puts positive probability on each θ, the Bayes rule is admissible. Typically Bayes rules are admissible in decision problems unless something funky is going on. Suppose δ_p is Bayes and inadmissible with, say, δ being better so that $R(\theta,\delta) \leq R(\theta,\delta_p)$ with there existing θ_0 such that $R(\theta_0,\delta) < R(\theta_0,\delta_p)$. Obviously the prior p cannot put positive probability on θ_0 or else δ would have a strictly smaller Bayes risk. When Θ is not discrete, if the risk functions are continuous in θ in a neighborhood of θ_0, then there is a neighborhood of θ_0 on which $R(\theta,\delta) < R(\theta,\delta_p)$ and the difference is bounded above 0. The prior p cannot have positive probability on this neighborhood of θ_0 or else δ would have a strictly smaller Bayes risk.

Definition 5.3.2. A class of decision rules $\mathscr{C}$ is a *complete class* if for any $\delta \notin \mathscr{C}$ there exists $\delta_* \in \mathscr{C}$, with δ_* better than δ. A class of decision rules $\mathscr{C}^*$ is *essentially complete* if for any $\delta \notin \mathscr{C}^*$ there exists $\delta_* \in \mathscr{C}^*$, with δ_* as good as δ.

Every complete class contains all of the admissible rules. The *Complete Class Theorem* is that (under suitable conditions) the Bayes rules constitute a complete class. One rationale for being a Bayesian is that if all reasonable decision rules correspond to some prior distribution, before choosing something among the reasonable decision rules, you should investigate whether its corresponding prior seems reasonable.

Two generalizations of decision rules exist. You can randomly pick a decision rule or you can have a decision rule that yields randomized actions, i.e., if you see $X = x$ you randomly pick an action with the randomization allowed to depend on x. This second idea is called a *randomized decision rule*. For a randomized action A,

$$L(\theta,A) \equiv \mathrm{E}_A[L(\theta,A)].$$

For a randomized decision rule taking randomized actions $\delta(x)$,

$$R(\theta,\delta) \equiv \mathrm{E}_{X|\theta}\mathrm{E}_{\delta(X)|X}[L(\theta,\delta(X)].$$

Randomized decision rules are an integral part of Neyman-Pearson testing theory. When randomly picking a decision rule, say Δ,

$$R(\theta,\Delta) \equiv \mathrm{E}_{\Delta}\mathrm{E}_{X|\theta,\Delta=\delta}[L(\theta,\delta(X)].$$

For example, suppose Δ is that you flip a coin and use δ_1 if the coin is heads and δ_2 if the coin is tails, then

$$R(\theta,\Delta) = \frac{1}{2}R(\theta,\delta_1) + \frac{1}{2}R(\theta,\delta_2).$$

Neither of these ideas are very attractive to statisticians. You have certain evidence $X = x$ that global warming is true. Why would you flip a coin to decide whether that evidence is sufficient for you to act as if global warming is true. Nonetheless, we will see in Chapter 7 that randomized decision rules are a key feature of the theory of hypothesis testing. (Fortunately, they are not a key feature of its practice.)

Definition 5.3.3. The *value of a decision problem* is $\inf_{a\in\mathscr{A}}\sup_{\theta} L(\theta,a)$ or $\inf_{A}\sup_{\theta} L(\theta,A)$ for randomized actions.

EXAMPLE 5.3.4. In testing with the standard loss function, $\sup_{\theta} L(\theta,a_0) = 1$ and $\sup_{\theta} L(\theta,a_1) = 1$ so the value is 1. Either action will minimize your maximum loss and achieve the value of the problem.

With a randomized action,

$$A = \begin{cases} a_0 & \text{with probability } p \\ a_1 & \text{with probability } 1-p, \end{cases}$$

we get $L(\theta_0,A) = 1-p$ and $L(\theta_1,A) = p$ and $\sup_{\theta} L(\theta,A) = \max(p,1-p)$. With randomized actions, the value of the problem is $\inf_{A}\sup_{\theta} L(\theta,a) = \inf_{p}\max(p,1-p) = 0.5$, so it corresponds to the action of flipping a fair coin to decide which hypothesis to accept.

If you think of this as a game, you get to take actions, your opponent determines the states of nature, and whatever you lose your opponent wins. Your opponent acts to maximize your losses, so you can do better on average if you take randomized actions. With randomized actions, the worst thing that can happen to you in this example is half as bad and with fixed actions.

For the loss function

$L(\theta,a)$	a_0	a_1
θ_0	2	4
θ_1	3	2

action a_0 will minimize your maximum loss. This is the minimax pure action. □

Exercise 5.1 For the loss function immediately above, show that the randomized action that takes a_0 with probability 2/3 minimizes the maximum expected loss.

In the next section we extend these ideas to minimax decision rules.

5.4 Minimax Rules

Definition 5.4.1. A decision rule δ_0 is a *minimax rule* if

$$\sup_\theta R(\theta, \delta_0) = \inf_\delta \sup_\theta R(\theta, \delta).$$

In particular, δ_0 is a *minimax rule* if for any $\theta_* \in \Theta$ and any δ, $R(\theta_*, \delta_0) \leq \sup_\theta R(\theta, \delta)$

Definition 5.4.2. A prior distribution on θ, say g_*, is a least favorable distribution if

$$\inf_\delta r(g_*, \delta) = \sup_g \inf_\delta r(g, \delta) = \sup_g r(g, \delta_g),$$

where δ_g is a Bayes rule for g. If δ_* is a Bayes rule with respect to g_* then

$$r(g_*, \delta_*) = \inf_\delta r(g_*, \delta) = \sup_g \inf_\delta r(g, \delta) = \sup_g r(g, \delta_g).$$

In other words, g_* is the prior that is going to give a Bayesian the worst possible outcome (risk).

Exercise 5.2 Show that g_* is least favorable if and only if $\inf_\delta r(g_*, \delta_0) \geq \inf_\delta r(g, \delta)$ for any δ_0 and all g.

We present without proof the Minimax Theorem.

Theorem 5.4.3. *The Minimax Theorem.*

$$\inf_\delta \sup_g r(g, \delta) = \sup_g \inf_\delta r(g, \delta)$$

Note that on the right side, if $\sup_g \inf_\delta r(g, \delta) = \inf_\delta r(g_*, \delta)$ then g_* is a least favorable distribution.

Corollary 5.4.4. For any δ, $\sup_\theta R(\theta, \delta) = \sup_g r(g, \delta)$.

PROOF: Observe that

$$r(g, \delta) = \mathrm{E}[R(\theta, \delta)] \leq \mathrm{E}[\sup_\theta R(\theta, \delta)] = \sup_\theta R(\theta, \delta),$$

so

$$\sup_g r(g,\delta) \leq \sup_\theta R(\theta,\delta).$$

Conversely, by considering the subset of priors that take on the value θ with probability one, say g_θ, note that $r(g_\theta,\delta) = R(\theta,\delta)$ and

$$\sup_g r(g,\delta) \geq \sup_{g_\theta:\theta\in\Theta} r(g_\theta,\delta) = \sup_\theta R(\theta,\delta).$$

□

Proposition 5.4.5. If the Minimax Theorem holds, if δ_0 is a minimax rule, and if g_* is a least favorable distribution with corresponding Bayes rule δ_*, then δ_0 is also a Bayes rule with respect to the least favorable distribution. (If the Bayes rule happens to be unique, we must have $\delta_0 = \delta_*$.)

PROOF: Using Corollary 4, Definition 1, Corollary 4, the Minimax Theorem 3, and Definition 2,

$$\begin{aligned} r(g_*,\delta_0) &\leq \sup_g r(g,\delta_0) \\ &= \sup_\theta R(\theta,\delta_0) \\ &= \inf_\delta \sup_\theta R(\theta,\delta) \\ &= \inf_\delta \sup_g r(g,\delta) \\ &= \sup_g \inf_\delta r(g,\delta) \\ &= r(g_*,\delta_*) \end{aligned}$$

This must be an equality since we know by definition of the Bayes rule that

$$r(g_*,\delta_0) \geq r(g_*,\delta_*)$$

Since δ_0 and δ_* have the same Bayes risk, they must both be Bayes rules. □

The point is that a Bayes rule for a least favorable distribution isn't necessarily a minimax rule, but a minimax rule, if it exists, is necessarily a Bayes rule for a least favorable distribution.

We now introduce a method for finding minimax rules.

Definition 5.4.6. δ_0 is an equalizer rule if for some constant K, $R(\theta,\delta_0) = K$ for all θ.

Proposition 5.4.7. If the Minimax Theorem holds and if δ_0 is both an equalizer rule and the Bayes rule for some prior distribution g_0, then δ_0 is minimax.

PROOF:

$$\inf_{\delta}\sup_{\theta} R(\theta,\delta) \leq \sup_{\theta} R(\theta,\delta_0) = K = r(g_0,\delta_0) = \inf_{\delta} r(g_0,\delta) \leq \sup_{g}\inf_{\delta} r(g,\delta).$$

By the Minimax Theorem and Corollary 4, all of these are equal, so in particular

$$\inf_{\delta}\sup_{\theta} R(\theta,\delta) = \sup_{\theta} R(\theta,\delta_0).$$

□

Exercise 5.3a. Let $X_1,\ldots,X_n|\theta \sim N(\theta,\sigma^2)$. For squared error loss, show that the sample mean is an equalizer rule.

Exercise 5.3b. Let $X|\theta \sim \text{Bin}(n,\theta)$ and $\theta \sim \text{Beta}(\alpha,\beta)$. Assume that the Minimax Theorem holds! For squared error loss, find the Bayes rule, say, $\delta_{\alpha\beta}$. Find $R(\theta,\delta_{\alpha\beta})$. Pick α and β so that $\delta_{\alpha\beta}$ is an equalizer rule. Establish that $\delta_{\alpha\beta}$ a minimax rule.

5.5 Prediction Theory

In prediction theory one wishes to predict an unobserved random vector y based on an observed random vector x. Let's say that y has q dimensions and that x has $p-1$ dimensions. We assume that the joint distribution of x and y is known. Any predictor of y is some function of x, say $\tilde{y}(x)$. We define a predictive loss function, $L[y,\tilde{y}(x)]$ and seek to find a predictor $\hat{y}(x)$ that minimizes the expected prediction loss, $\text{E}\{L[y,\tilde{y}(x)]\}$, where the expectation is over both y and x. Note that

$$\text{E}_{x,y}\{L[y,\tilde{y}(x)]\} = \text{E}_x\left(\text{E}_{y|x}\{L[y,\tilde{y}(x)]\}\right)$$

or in alternative notation

$$\text{E}\{L[y,\tilde{y}(x)]\} = \text{E}\left(\text{E}\{L[y,\tilde{y}(x)]|x\}\right).$$

In particular, there is a one to one correspondence between prediction theory and the approach of traditional decision theory to Bayesian analysis. We associate y with θ and x with X. In prediction we assume a joint distribution for x and y whereas in Bayesian analysis we specify the sampling distribution and the prior that together determine the joint distribution of θ and X. A predictor $\tilde{y}(x)$ is analogous to a decision rule. The expected prediction error $\text{E}_{x,y}\{L[y,\tilde{y}(x)]\}$ is analogous to the Bayes risk. Just like in Bayesian analysis, the way to find the best predictor is, for each value of x, to find the value of $\tilde{y}(x)$ that minimizes $\text{E}\{L[y,\tilde{y}(x)]|x\}$.

The most common prediction problem is similar to linear regression in which y takes values in $\mathbf{R}$ and uses squared error loss,

$$L[y,\tilde{y}(x)] = [y-\tilde{y}(x)]^2.$$

We want to minimize the expected prediction error

$$\mathrm{E}\{L[y,\tilde{y}(x)]\} = \mathrm{E}\{[y-\tilde{y}(x)]^2\}$$

where the expectation is over both y and x. Identifying prediction with decision and conditioning on x, we see that Proposition 5.1.1 implies

Proposition 5.5.1. For data (x',y), $y \in \mathbf{R}$, and $L(y,\tilde{y}(x)) = [y-\tilde{y}(x)]^2$, the best predictor is $\hat{y} = \mathrm{E}(y|x)$.

Regression, both linear and nonparametric, is about estimating the optimal predictor $\mathrm{E}(y|x)$. Note that this result holds even when y is Bernoulli, in which case the best predictor under squared error loss is $\mathrm{E}(y|x) = \Pr[y=1|x]$. Using squared error loss with a Bernoulli variable y is essentially using *Brier Scores.*

Similarly we can get other best predictors.

Proposition 5.5.2. For data (x',y), $y \in \mathbf{R}$, and $L(y,\tilde{y}(x)) = w(y)[y-\tilde{y}(x)]^2$ with $w(\theta) > 0$, the best predictor is $\hat{y} = \mathrm{E}[w(y)y|x]/\mathrm{E}[w(y)|x]$.

Proposition 5.5.3. For data (x',y), $y \in \mathbf{R}$, and $L(y,\tilde{y}(x)) = |y-\tilde{y}(x)|$, a best predictor is any $\hat{y} = m \equiv \mathrm{Median}(y|x)$.

When y takes values in $\{0,1\}$, an alternative loss function is the so called *Hamming* loss,

$$L[y,\tilde{y}(x)] = \mathscr{I}[y \neq \tilde{y}(x)],$$

wherein $\mathscr{I}(\text{logical})$ is 0 if the logical statement is false and 1 if it is true and a predictor $\tilde{y}(x)$ also needs to take values in $\{0,1\}$. We want to minimize the expected prediction error

$$\mathrm{E}\{L[y,\tilde{y}(x)]\} = E\{\mathscr{I}[y \neq \tilde{y}(x)]\}$$

where the expectation is over both y and x. We see that Proposition 5.1.4 implies

Proposition 5.5.4. For data (x',y), $y \in \{0,1\}$ and $L(y,\tilde{y}(x)) = \mathscr{I}(y \neq \tilde{y}(x))$, a best predictor has

$$\hat{y}(x) \equiv \begin{cases} 0 & \text{if } \Pr(y=0|x) > 0.5 \\ 1 & \text{if } \Pr(y=0|x) < 0.5 \end{cases}.$$

In binary regression people tend to focus on the probability of getting a 1, rather than getting a 0 (which is analogous to a null hypothesis), so it is more common to

think of the optimal predictor as

$$\hat{y}(x) \equiv \begin{cases} 0 & \text{if } \Pr(y=1|x) < 0.5 \\ 1 & \text{if } \Pr(y=1|x) > 0.5 \end{cases}.$$

Binomial (e.g., logistic or probit) regression is about estimating the probability $\Pr(y=1|x)$. For squared error loss, this gives the estimated optimal predictor. For Hamming loss, the estimated optimal predictor is 0 or 1 depending on whether the estimated value of $\Pr(y=1|x)$ is less than 0.5

Fisher argued (similarly to Bayesians) that prediction problems should be considered entirely as conditional on the predictor vector x. However, there are some predictive measures, such as the squared multiple correlation coefficient (of which the coefficient of determination R^2 is an estimate), whose definition depends on the distribution of x. Measures that depend on the distribution of x are inappropriate to compare when the distribution of x changes. Thus it is common to argue that R^2 values for the same model on different data are not comparable. In fact, that is only true if the x data have been sampled from a different population – which is usually the case.

5.5.1 Prediction Reading List

Aitchison, J. and Dunsmore, I. R. (1975). *Statistical Prediction Analysis*. Cambridge University Press, Cambridge.

Billheimer, Dean (2019) Predictive Inference and Scientific Reproducibility, *The American Statistician*, 73, 291-295.

Briggs, W. (2016), *Uncertainty. The Soul of Modeling, Probability and Statistics*, Berlin, Germany: Springer International.

Clarke, B., and Clarke, J. (2012), "Prediction in Several Conventional Contexts," *Statistical Surveys*, 6, 1-73.

de Finetti, B. (1937), "La prevision: ses lois logiques, ses sources subjectives," *Annals Institute Henri Poincare*, 7, 1-68.

de Finetti, B. (2017), *Theory of Probability*, vol. I and II, New York:Wiley.

Geisser, S. (1988), "The Future of Statistics in Retrospect," in *Bayesian Statistics 3*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Oxford, UK: Oxford University Press, pp. 147-158. [292]

Geisser, S. (1993), *Predictive Inference: An Introduction*, London:Chapman and Hall/ CRC.

Nelder, J. A. (1999), "Statistics for the Millenium" *The Statistician*, 48, 257-269.

5.5.2 Linear Models

Do a prediction chapter. the prediction result from *PA* Exercise 2.1 can be used more generally when predicting $\tilde{y}$ and specifically some function of it $\tilde{\rho}'\tilde{y}$ like the difference in sample means between two predictive samples. When $\tilde{X}\beta$ is estimable,

$$\frac{\tilde{\rho}'\tilde{y} - \tilde{\rho}'\tilde{X}\hat{\beta}}{\sqrt{MSE\left(\tilde{\rho}'\tilde{\rho} + \tilde{\rho}'\tilde{X}(X'X)^{-}\tilde{X}\right)'\tilde{\rho}}}$$

$$\frac{(\bar{Y}_{M1} - \bar{Y}_{M2}) - (\bar{y}_{1\cdot}\bar{y}_{2\cdot})}{\sqrt{MSE\left[\frac{1}{M} + \frac{1}{M} + \frac{1}{N_1} + \frac{1}{N_2}\right]}}$$

Chapter 6
Estimation Theory

6.1 Basic Estimation Definitions and Results

Consider a parametric family of distributions for an observable random n-vector y defined by their densities (with respect to some dominating measure)

$$y|\theta \sim f(v|\theta); \qquad \theta \in \Theta.$$

Most often, $\Theta \subset \mathbf{R}^d$. Often we consider $y_1, \ldots, y_n$ iid with some common density $f_*(\cdot|\theta)$. In that case,

$$y = (y_1, \ldots, y_n)'; \qquad f(v|\theta) = \prod_{i=1}^{n} f_*(v_i|\theta).$$

Definition 6.1.1. Any function of y, say $T(y)$, is a *statistic*. Statistics can be real valued or vector valued.

Note that a statistic is not allowed to depend on θ but for fixed θ we can still treat functions $G(y,\theta)$ as random variables.

Definition 6.1.2. A statistic (estimator) $g(y)$ is *unbiased* for $h(\theta)$, if $\mathrm{E}_{y|\theta}[g(y)] = h(\theta)$ for all θ. The *bias* of $g(y)$ for estimating $h(\theta)$ is defined by $b_{gh}(\theta) \equiv \mathrm{E}_{y|\theta}[g(y) - h(\theta)]$. If $h(\theta) \equiv \theta$, we suppress the bias subscript h. These functions can be either real valued or vector valued.

For $g(y)$ and $h(\theta)$ real valued with the loss function

$$L(\theta, a) = [a - h(\theta)]^2,$$

the risk is

$$R[\theta, g(y)] = \mathrm{E}_{y|\theta}[g(y) - h(\theta)]^2 = \mathrm{Var}_{y|\theta}[g(y)] + [b_{gh}(\theta)]^2.$$

This is often referred to as the *mean squared error*. (In linear model theory the "mean squared error" is used to indicate the unbiased estimate of the variance, so to distinguish the concepts, this risk may be called the "expected squared error.")

Definition 6.1.3. The density of y, $f(v|\theta)$, is a function of the place holder variable v for known θ. We get to observe y but never know θ. For fixed y, the *likelihood function* is

$$L(\theta) \equiv f(y|\theta).$$

If we want to emphasize its dependence on y we may write $L(\theta|y)$.

6.1.1 Maximum Likelihood Estimation

Definition 6.1.4. A statistic $\hat{\theta}(y) \equiv \hat{\theta}$ is a *maximum likelihood estimate (MLE)* of θ if

$$L[\hat{\theta}(y)] = \sup_{\theta \in \Theta} L(\theta) \equiv \sup_{\theta \in \Theta} L(\theta|y).$$

For any function $h(\theta)$, the MLE is $h(\hat{\theta})$. cf. Cox and Hinkley (1974).

Under suitable regularity conditions, MLEs have excellent asymptotic properties, see Ferguson (1996). They converge in probability to what they are estimating and, when suitably normalized, they converge in distribution to a normal distribution.

A theorem exists that if a minimum variance unbiased estimate exists and if the maximum likelihood estimate is unbiased, it is the minimum variance unbiased estimate.

6.2 Sufficiency and Completeness

Calling a statistic $T(y)$ "sufficient" is intended to convey that all the information about θ is contained in $T(y)$.

Definition 6.2.1. $T(y)$ is a *sufficient statistic* (for θ) if the distribution of y given $T(y)$ does not depend on the parameter θ.

If $\Theta_0 \subset \Theta$, then any $T(y)$ that is sufficient for $\theta \in \Theta$ is automatically sufficient for $\theta \in \Theta_0$. The distribution of $y|T(y)$ can be used to examine whether our parametric family is appropriate, that is, it can be used for model checking.

The data themselves are always sufficient. If $y_1, \ldots, y_n$ are iid $f_*(\cdot|\theta)$, the order statistics $y_{(1)} \leq \cdots \leq y_{(n)}$ are always sufficient because the distribution of the data

given the order statistics is just a permutation. In other words,

$$P(y_1 = v_1, \ldots, y_n = v_n | y_{(1)} = u_1, \ldots, y_{(n)} = u_n) = \frac{1}{n!},$$

where $v_1, \ldots, v_n$ is any permutation of $u_1 \leq \cdots \leq u_n$. Clearly, this conditional distribution does not depend on θ.

We generally determine whether something is sufficient by using the following theorem.

Theorem 6.2.2. *The Factorization Criterion.*
$T(y)$ is a sufficient statistic for θ if an only if we can write

$$f(v|\theta) = h(v)g[T(v);\theta]$$

for some functions h and g.

First shown by Fisher (1922), this result was proven in great generality by Halmos and Savage (1949). When establishing properties of sufficient statistics, knowing the the factorization holds is very useful. However when finding sufficient statistics for a particular model, finding a factorization is how we typically establish sufficiency. So both directions in the if and only if statement are important. That notwithstanding, only a proof that the factorization implies sufficiency is given at the end of the section.

As a function of θ, the likelihood is now $L(\theta|y) \propto g[T(y);\theta]$, so, for example, the maximum of the likelihood function must occur at the maximum of $g[T(y);\theta]$, which only depends on y through the sufficient statistic $T(y)$. Similarly, if we have a prior density on θ, say $p_\theta(u)$, from Bayes Theorem the posterior density of θ given $y = v$ is

$$p_{\theta|y}(u|v) = \frac{f(v|u)p_\theta(u)}{\int f(v|u)p_\theta(u)\,du} = \frac{g[T(v);u]p_\theta(u)}{\int g[T(v);u]p_\theta(u)\,du},$$

where the last term shows that the posterior distribution depends on $y = v$ only through the fact that $T(y) = T(v)$. Thus the posterior depends on y only through the value of the sufficient statistic, any sufficient statistic.

EXAMPLE 6.2.3. Consider $y_1, \ldots, y_n$ iid $U(0,\theta)$, $\theta > 0$. For these data the largest order statistic is sufficient.

$$f(v|\theta) = \prod_{i=1}^{n} \frac{1}{\theta}\mathscr{I}_{(0,\theta)}(v_i) = \prod_{i=1}^{n} \frac{1}{\theta}\mathscr{I}_{(0,\theta)}(v_{(i)}) = \frac{1}{\theta^n}\mathscr{I}_{(0,\infty)}(v_{(1)})\mathscr{I}_{(0,\theta)}(v_{(n)}).$$

The second equality holds because we are merely reordering the terms in the product. The third equality follows from the definition of the order statistics. In the factorization criterion, $h(v) = \mathscr{I}_{(0,\infty)}(v_{(1)})$

EXAMPLE 6.2.4. Consider $y_1, \ldots, y_n$ iid $U(\theta_1, \theta_2)$, $\theta_2 > \theta_1$. For these data the smallest and largest order statistics are sufficient.

$$f(v|\theta_1,\theta_2) = \prod_{i=1}^{n} \frac{1}{(\theta_2-\theta_1)} \mathscr{I}_{(\theta_1,\theta_2)}(v_i) = \prod_{i=1}^{n} \frac{1}{(\theta_2-\theta_1)} \mathscr{I}_{(\theta_1,\theta_2)}(v_{(i)})$$
$$= \frac{1}{(\theta_2-\theta_1)^n} \mathscr{I}_{(\theta_1,\theta_2)}(v_{(1)}) \mathscr{I}_{(\theta_1,\theta_2)}(v_{(n)}).$$

If the smallest and largest order statistics are between θ_1 and θ_2, all of the order statistics must be between them.

When manipulating densities, if the density contains an indicator function, it is very important to retain the indicator throughout the manipulation.

Exercise 6.1. Let $y_1,\ldots,y_n$ iid $N(\mu,\sigma^2)$ with $\theta \equiv (\mu,\sigma^2)' \in \Theta = (\mathbf{R},\mathbf{R}^+)'$. Show that $(\bar{y}_\cdot, s^2)'$ is sufficient. Also show that $(\sum_i y_i, \sum_i y_i^2)'$ is sufficient.

Exercise 6.2. Let $y_1,\ldots,y_n$ iid $N(0,\sigma^2)$ with $\theta \equiv \sigma^2 \in \Theta = \mathbf{R}^+$. Show that $\sum_i y_i^2$ is sufficient.

Exercise 6.3. Let $y_1,\ldots,y_n$ iid $N(5,\sigma^2)$ with $\theta \equiv \sigma^2 \in \Theta = \mathbf{R}^+$. Show that $(\bar{y}_\cdot, s^2)'$ is sufficient. Show that $\sum_i (y_i-5)^2$ is sufficient.

Definition 6.2.5. A statistic $H(y)$ is a *complete statistic* if $\mathrm{E}_{y|\theta}\{Q[H(y)]\} = 0$ for all $\theta \in \Theta$ implies that $Q[H(y)] = 0$ a.e. in the dominating measure. A less restrictive version is that $P_{y|\theta}\{Q[H(y)] = 0\} = 1$ for every $\theta \in \Theta$. $H(y)$ is *boundedly complete* if the result holds for bounded functions Q.

Suppose you have two functions of a complete statistic, say $g_1[H(y)]$ and $g_2[H(y)]$, and both are unbiased for $h(\theta)$. Then $\mathrm{E}_{y|\theta}\{g_1[H(y)] - g_2[H(y)]\} = 0$ for all θ which means that $P_{y|\theta}[g_1[H(y)] - g_2[H(y)] = 0] = 1$. Basically, for any parameter $h(\theta)$, there can be only one unbiased estimate that is a function of $H(y)$. We will see in the next section that if a statistic $T(y)$ is both complete and sufficient, any function of it, say $g[T(y)]$ has to be a minimum variance unbiased estimate of its expectation, i.e., of $h(\theta) \equiv \mathrm{E}_{y|\theta}\{g[T(y)]\}$.

If $\Theta_0 \subset \Theta$, then any $T(y)$ that is complete for $\theta \in \Theta_0$ is automatically complete for $\theta \in \Theta$. Relative to subsets of Θ, sufficiency and completeness work in opposite ways. Think of Θ as indexing all distributions that are absolutely continuous *wrt* (with respect to) Lebesgue measure and for which the expected value exists. Think of Θ_0 as indexing $N(\mu,1)$ distributions. Consider $y_1,\ldots,y_n$ iid $f_*(\cdot|\theta)$. For Θ, the order statistics $y_{(1)} \leq \ldots \leq y_{(n)}$ are complete and sufficient, see Frasier (1957). For Θ_0, we will see later that the sample mean $\bar{y}_\cdot$ is complete and sufficient. For Θ_0, the order statistics are sufficient but not complete. For Θ, $\bar{y}_\cdot$ is complete but not sufficient.

Incidentally, $\bar{y}_\cdot$ is the minimum variance unbiased estimate of $\mathrm{E}(y_i)$ in both families. For $\theta \in \Theta$ it is relatively hard to find statistics that estimate the expected

value unbiasedly. Thus $\bar{y}_\cdot$, which is the mean of the order statistics, is the best of a relatively small group of unbiased estimators but is best for a wide array of distributions. For $\theta \in \Theta_0$ it is relatively easy to find statistics that estimate the expected value unbiasedly (partly because of symmetry). Thus $\bar{y}_\cdot$ is the best of a large group of estimators but is best for a relatively small collection of distributions.

6.2.1 Ancillary Statistics

Ancillary statistics are a somewhat controversial subject. If an ancillary statistic exists, some people (notably Fisher) argue that inference on θ should proceed by conditioning on the ancillary statistic.

Definition 6.2.8. Suppose $y \sim f(v|\theta)$. A statistic $A(y)$ is said to be *ancillary* if $P_{y|\theta}[A(y) \in B]$ for any (measurable) B does not depend on θ. A statistic $A(y)$ is said to be *first order ancillary* if $\mathrm{E}_{y|\theta}[A(y)]$ does not depend on θ.

Some people like to analyze experimental designs in which treatments are randomly assigned to experimental units based entirely on the random assignment. (For some simple applications see *PA*, Appendix G.) Obviously, the random assignment has nothing to do with any parameters related to the results of the experiment, so the result of the randomization is an ancillary statistic. If the randomization is the only thing random, conditioning on the ancillary statistic leaves nothing on which to base an analysis. Ironically, Fisher was big on both the idea of conditioning on ancillary statistics and on the idea of using only the random assignment of treatments as the basis for analyzing experiments.

When predicting y on the basis of x, Fisher argued that the only parameters of interest are associated with the conditional distribution given x. Pretty obviously, the distribution of x does not depend on any of those parameters, so is ancillary. (This is actually a somewhat more nuanced argument involving parameters of interest and nuisance parameters.)

In dealing with count data, Fisher's exact conditional test for 2×2 contingency tables conditions on ancillary statistics (row and column totals). In fact, all exact conditional tests for contingency tables involve conditioning on statistics that are ancillary for the parameters of interest. Whether these tests are more appropriate than unconditional tests is a source of controversy, cf. Agresti (1992).

The most famous result relating to ancillary statistics is due to Basu.

Basu's Theorem 6.2.8. If $T(y)$ is a boundedly complete sufficient statistic and if $A(y)$ is ancillary then $T(y)$ and $A(y)$ are independent.

PROOF: [Lehmann, 1983, p.46] Let $\eta_B(t) \equiv P_{y|\theta}[A(y) \in B|T(y) = t]$. By sufficiency this conditional probability does not depend on θ. Dropping the unnecessary subscript on the conditional probability,

$$\mathrm{E}_{y|\theta}[\eta_B(T(y))] = P[A(y) \in B] \equiv p_B,$$

which by ancillarity does not depend on θ. It follows that

$$\mathrm{E}_{y|\theta}[\eta_B(T(y)) - p_B] = 0,$$

for every θ. By (boundedly) completeness, $\eta_B(T(y)) - p_B = 0$ a.s. for all θ, so $P[A(y) \in B|T(y) = t] = P[A(y) \in B]$. If the conditional probability does not depend on what is being conditioned on, the objects are independent, so $A(y) \in B$ is independent of $T(y)$ for any B.

EXAMPLE 6.2.10. Consider random variables $x_1, \ldots, x_n$ iid each with a density $h(\cdot)$. Random variables $y_1, \ldots, y_n$ are iid from a location family if for some x_is they satisfy $y_i \sim x_i + \theta$ in which case they have density $f_{y|\theta}(v) = \prod_{i=1}^n h(v_i - \theta)$. The statistics $y_i - y_j$ are all ancillary because $y_i - y_j = x_i - x_j$ so the distribution does not depend on θ.

Consider $y_1, \ldots, y_n$ iid from a scale family defined by $y_i \sim x_i\theta$ in which case they have density $f_{y|\theta}(v) = \prod_{i=1}^n h(v_i/\theta)/\theta$. The statistics y_i/y_j are all ancillary because $y_i/y_j = x_i/x_j$ so the distribution does not depend on θ.

It would **seem** that Ancillary Statistics cannot be complete.

6.2.2 Proof of the Factorization Criterion

I think this if from Ferguson (1967).

Lemma 6.2.11. Let $r \geq 0$ and s be two functions and let $E[r(y)] \leq K_r < \infty$. Because it is nonnegative , r can act like a (not typically probability) density with regards to the dominating measure P, call this new measure μ_r, and so we can construct something akin to a conditional expectation relative to this new measure, call it $\mathrm{E}_r[s(y)|T(y)]$, then

$$\begin{aligned}
\int_{T^{-1}(B)} g[T(v)]r(v)s(v)dP(v) &= \int_{T^{-1}(B)} g[T(v)]\mathrm{E}[r(v)s(v)|T(v)]dP(v) \\
&= \int_{T^{-1}(B)} g[T(v)]\mathrm{E}_r[s(v)|T(v)]r(v)dP(v) \\
&= \int_{T^{-1}(B)} g[T(v)]\mathrm{E}_r[s(v)|T(v)]\mathrm{E}[r(v)|T(v)]dP(v).
\end{aligned}$$

PROOF: Let $g[T(v)]$ be an indicator and then do limits.

If $g[T(v)]$ be an indicator, combine it with the indicator for $T^{-1}(B)$, so it is enough to show the result without $g[T(v)]$ in the integral.

$$\int_{T^{-1}(B)} s(v)d\mu_r(v) = \int_{T^{-1}(B)} \mathrm{E}_r[s(y)|T(v)]d\mu_r(v)$$

so

$$\begin{aligned}\frac{1}{K_r}\int_{T^{-1}(B)} r(v)s(v)dP(v) &= \frac{1}{K_r}\int_{T^{-1}(B)} \mathrm{E}_r[s(y)|T(v)]r(v)dP(v)\\ &= \frac{1}{K_r}\int_{T^{-1}(B)} \mathrm{E}\left\{\mathrm{E}_r[s(y)|T(v)]r(y)\middle|T(v)\right\}dP(v)\\ &= \frac{1}{K_r}\int_{T^{-1}(B)} \mathrm{E}_r[s(y)|T(v)]\mathrm{E}\left\{r(y)\middle|T(v)\right\}dP(v)\end{aligned}$$

Lemma 6.2.12. Suppose ν is a σ-finite measure so there exists a partition A_i, $i = 1,2,\ldots$ such that for all i, $\nu(A_i) < \infty$. The

$$\nu_*(B) = \sum_{i=1}^{\infty} \frac{B \cap A_i}{2^i \nu(A_i)}$$

is a probability measure and

$$d\nu_*(v) = \sum_{i=1}^{\infty} \frac{\mathscr{I}_{A_i}(v)}{2^i \nu(A_i)}\, d\nu(v).$$

PROOF: Obvious.

PROOF OF THE FACTORIZATION CRITERION:
$\Leftarrow$ We must prove that, for all A, $\mathrm{E}[\mathscr{I}_A(y)|T(y)]$ does not depend on θ.
From Lemma 6.2.12 and implicitly defining the function h_1,

$$\begin{aligned}dP_\theta(v) = f(v|\theta)d\nu = f(v|\theta)\frac{1}{\sum_{i=1}^{\infty}\frac{\mathscr{I}_{A_i}(v)}{2^i\nu(A_i)}}\, d\nu_*(v)\\ = g[T(v);\theta]\frac{h(v)}{\sum_{i=1}^{\infty}\frac{\mathscr{I}_{A_i}(v)}{2^i\nu(A_i)}}\, d\nu_*(v)\\ \equiv g[T(v);\theta]h_1(v)\, d\nu_*(v).\end{aligned}$$

In Lemma 6.2.11 identify $g[T(v)] \to g[T(v);\theta]$, $r(v) \to h_1(v)$, $s(v) \to \mathscr{I}_A(v)\, dP(v) \to d\nu_*(v)$, so

$$\begin{aligned}\int_{T^{-1}(B)} g[T(v);\theta]h_1 s(v)dP(v) &= \int_{T^{-1}(B)} g[T(v)]\mathrm{E}[h_1(y)\mathscr{I}_A(y)|T(v)]\, d\nu_*(v)\\ &= \int_{T^{-1}(B)} g[T(v);\theta]\mathrm{E}_{h_1}[\mathscr{I}_A(y)|T(v)]h_1(v)\, d\nu_*(v)\\ &= \int_{T^{-1}(B)} g[T(v);\theta]\mathrm{E}_{h_1}[\mathscr{I}_A(y)|T(v)]\mathrm{E}[h_1(y)|T(v)]dP(v)\end{aligned}$$

$$\begin{aligned}\int_{T^{-1}(B)} \mathrm{E}[\mathscr{I}_A(y)|T(v)]dP_\theta(v) &= \int_{T^{-1}(B)} \mathscr{I}_A(v)g[T(v);\theta]h_1(v)\,d\nu_*(v)\\ &= \int_{T^{-1}(B)} g[T(v);\theta]\mathrm{E}_{h_1}[\mathscr{I}_A(y)|T(v)]h_1(v)\,d\nu_*(v)\\ &= \int_{T^{-1}(B)} g[T(v);\theta]\mathrm{E}_{h_1}[\mathscr{I}_A(y)|T(v)]\mathrm{E}_{\nu_*}[h_1(y)|T(v)]\,d\nu_*(v)\\ &= \int_{T^{-1}(B)} g[T(v);\theta]\mathrm{E}_{h_1}[\mathscr{I}_A(y)|T(v)]h_1(y)\,d\nu_*(v)\\ &= \int_{T^{-1}(B)} \mathrm{E}_{h_1}[\mathscr{I}_A(y)|T(v)]\,dP_\theta(v).\end{aligned}$$

By definition $\mathrm{E}_{h_1}[\mathscr{I}_A(y)|T(v)] = \mathrm{E}_{y|\theta}[\mathscr{I}_A(y)|T(v)]$ but $\mathrm{E}_{h_1}[\mathscr{I}_A(y)|T(v)]$ is defined with respect to the measure with density $h_1(v)\,d\nu_*(v)$ which does not depend on θ

$\Rightarrow$ We don't actually care that much about this direction.

6.3 Rao-Blackwell Theorem and Minimum Variance Unbiased Estimation

Suppose $T(y)$ is any statistic and $g(y)$ is unbiased for $h(\theta)$, both real valued. To simplify notation, for fixed θ write the conditional expectation and variance of $g(y)$ given $T(y)$ as both

$$\mathrm{E}_{y|\theta}[g(y)|T(y)] \equiv \mathrm{E}_{y|\theta,T(y)}[g(y)]$$

and write

$$\mathrm{Var}_{y|\theta}[g(y)|T(y)] \equiv \mathrm{Var}_{y|\theta,T(y)}[g(y)].$$

The key point in this section is that these numbers typically depend on θ but, when $T(Y)$ is sufficient, they do not.

Standard results, cf. Exercise A.1, on conditional probabilities provide that for any statistic $T(y)$,

$$h(\theta) = \mathrm{E}_{y|\theta}[g(y)] = \mathrm{E}_{y|\theta}\{\mathrm{E}_{y|\theta}[g(y)|T(y)]\} \tag{1}$$

and

$$\begin{aligned}\mathrm{Var}_{y|\theta}[g(y)] &= \mathrm{Var}_{y|\theta}\{\mathrm{E}_{y|\theta}[g(y)|T(y)]\} + \mathrm{E}_{y|\theta}\{\mathrm{Var}_{y|\theta}[g(y)|T(y)]\}\\ &\geq \mathrm{Var}_{y|\theta}\{\mathrm{E}_{y|\theta}[g(y)|T(y)]\}.\end{aligned} \tag{2}$$

Since we do not know θ, if $\mathrm{E}_{y|\theta}[g(y)|T(y)]$ depends on θ, it is not a statistic. It is not a number that we can actually find once we observe y, so it is not something we can use as an estimator. If $T(y)$ is a sufficient statistic, the distribution of $g(y)$ given

$T(y)$ does not depend on θ, so $\mathrm{E}_{y|\theta}[g(y)|T(y)] \equiv \mathrm{E}_{y|\theta,T(y)}[g(y)] = \mathrm{E}_{y|T(y)}[g(y)] \equiv \mathrm{E}[g(y)|T(y)]$ is a statistic, it is a function of y that does not depend on θ. It then follows from (1) that $\mathrm{E}[g(y)|T(y)]$ is an unbiased estimate of $h(\theta)$ and from (2) that $\mathrm{Var}_{y|\theta}\{\mathrm{E}[g(y)|T(y)]\} \leq \mathrm{Var}[g(y)]$, so the conditional expectation is at least as good an unbiased estimate as the original unbiased estimate. We have proven the following.

Theorem 6.3.1. *Rao-Blackwell Theorem.*
If $g(y)$ is an unbiased estimate of $h(\theta)$ and $T(y)$ is a sufficient statistic, then $\mathrm{E}[g(y)|T(y)]$ is also an unbiased estimate with variance no greater than that of $g(y)$.

If $T(Y)$ is both complete and sufficient, any function of it, say $\tilde{g}[T(y)]$, is unbiased for its expected value, say, $h(\theta) \equiv \mathrm{E}_{y|\theta}\{\tilde{g}[T(y)]\}$. If $g(y)$ is any other unbiased estimate of $h(\theta)$, then by sufficiency $\mathrm{E}[g(y)|T(y)]$ is also an unbiased statistic and a function of $T(Y)$, so $\mathrm{E}\{\tilde{g}[T(y)] - \mathrm{E}[g(y)|T(y)]\} = 0$ and by completeness of $T(y)$, $1 = \Pr_{y|\theta}\{\tilde{g}[T(y)] - \mathrm{E}[g(y)|T(y)] = 0\} = \Pr_{y|\theta}\{\tilde{g}[T(y)] = \mathrm{E}[g(y)|T(y)]\}$. It follows that

$$\mathrm{Var}_{y|\theta}\{\tilde{g}[T(y)]\} = \mathrm{Var}_{y|\theta}\{\mathrm{E}_{y|\theta}[g(y)|T(y)]\} \leq \mathrm{Var}_{y|\theta}[g(y)],$$

so the variance of $\tilde{g}[T(y)]$ is at least as small as the variance of any other unbiased estimate. This result is sometimes called the *Lehmann-Scheffé Theorem*.

The factorization criterion allows us to find sufficient statistics, the remaining question is how to find complete sufficient statistics. That will be addressed in the section on exponential families.

A similar result holds for any loss function $L(\theta, a)$ that is convex in a. Jensen's inequality applied to the conditional distribution of y given $T(y)$ implies

$$\mathrm{E}\{L[\theta, g(y)]|T(y)\} \geq L\{\theta, \mathrm{E}[g(y)|T(y)]\},$$

so

$$R[\theta, g(y)] = \mathrm{E}\left(\mathrm{E}\{L[\theta, g(y)]|T(y)\}\right) \geq \mathrm{E}\left(L\{\theta, \mathrm{E}[g(y)|T(y)]\}\right) = R\{\theta, \mathrm{E}[g(y)|T(y)]\}.$$

In particular, conditioning on a sufficient statistic can improve mean squared error even for biased estimates.

6.3.1 Minimal Sufficient Statistics

Definition 6.3.6 A sufficient statistic $T_0(y)$ is said to be *minimal sufficient* if for any other sufficient statistic $T(y)$ there exists a transformation q such that $T_0(y) = q[T(y)]$.

EXAMPLE 6.3.7. Suppose $y_1, \ldots, y_n$ are iid $N(\mu, \sigma^2)$, then the data and the order statistics are sufficient but $(\bar{y}_\cdot, s^2)$ and $(\sum_i y_i, \sum_i y_i^2)$ are minimal sufficient.

Suppose $T(y)$ is sufficient and $T_0(y)$ is minimal sufficient and suppose that $g[T(y)]$ and $g_0[T_0(y)]$ are unbiased for $h(\theta)$. By Rao-Blackwell, $\mathrm{E}\{g[T(y)] \,|\, T_0(y)]$ is at least as good as $g[T(y)]$ and may be better. However, by minimal sufficiency,

$$\mathrm{E}\{g_0[T_0(y)] \,|\, T(y)] = \mathrm{E}\left(g_0\{q[T(y)]\} \,|\, T(y)\right) = g_0\{q[T(y)]\} = g_0[T_0(y)],$$

so it cannot be an improvement. This establishes that a best unbiased estimator must be a function of the minimal sufficient statistic, however, without completeness there is no guarantee that merely conditioning an unbiased estimate on a minimal sufficient statistic will get you the best unbiased estimator. Galili and Meilijson (2016) provide a simple example with several unbiased functions of the minimal sufficient statistic giving different variances.

Lehmann and Scheffé (1950) show that if $T(y)$ is a boundedly complete sufficient statistic then it is minimal sufficient. If a minimal sufficient statistic and a UMVU estimate both exist, there has to be a UMVU that is a function of the minimal sufficient statistic.

6.3.2 *Unbiased Estimation: Additional Results from Rao (1973, Chapter 5)*

EXAMPLE 6.3.8. (This also appears in Cox and Hinkley.) Let $y_1, \ldots, y_n$ be iid $N(\mu, \sigma^2)$. Then $\mathrm{E}(s^2) = \sigma^2$, $\mathrm{Var}(s^2) = 2\sigma^4/(n-1)$, but $s^2(n-1)/n$ maximizes the likelihood function. Consider the risk under squared error loss from estimates of the form cs^2 for a constant c.

$$\mathrm{E}[cs^2 - \sigma^2]^2 = \left[\frac{2c^2}{n-1} + (1-c)^2\right]\sigma^4.$$

This is minimized for $c = (n-1)/(n+1)$. Squared error loss is kind of weird when there is a bound on the parameter space. You tend to get risk improvements by shrinking towards the bound. □

In normal theory statistical inference, what matters is not the point estimate of σ^2 but the fact that $\sum_i (y_i - \bar{y}_\cdot)^2/\sigma^2 \sim \chi^2(n-1)$. Using s^2 leads to the $t(n-1)$ and $F(1, n-1)$ distributions. If you use a different point estimate, you need to use different distributions, but they adjust is such a way that you get the same tests and confidence intervals.

EXAMPLE 6.3.9. *A case for unbiased estimation.* Consider k independent, possibly biased, estimators of θ, say $\hat{\theta}_1, \ldots, \hat{\theta}_k$. Take $\mathrm{E}\left[\hat{\theta}_i\right] = \theta + b$ and $\mathrm{Var}\left[\hat{\theta}_i\right] =$

$\sigma_i^2 \leq \sigma^2$. Now define $\bar{\theta} \equiv \sum_{i=1}^k \hat{\theta}_i/k$. We still get $\mathrm{E}\left[\bar{\theta}\right] = \theta + b$ but now $\mathrm{Var}\left[\bar{\theta}\right] = \sum_i \sigma_i^2/k^2 \leq \sigma^2/k$. So combining biased estimators reduces variance but does not help to reduce bias. □

The next result is very similar to a result in PA that something is the BP of y if and only if the prediction residuals are uncorrelated with any function of x.

Proposition 6.3.10. $T(y)$ is minimum variance unbiased for θ if and only if it is unbiased and for any function h with $\mathrm{E}[h(y)] = 0$, we have $\mathrm{Cov}[T(y), h(y)] = 0$.

PROOF: $\Leftarrow$ Suppose $\mathrm{E}[T(y)] = \theta = \mathrm{E}[h(y)]$, then $U(y) = T(y) - h(y)$ has $\mathrm{E}[U(y)] = 0$.

$$\mathrm{Var}[h(y)] = \mathrm{Var}[T(y) + U(y)] = \mathrm{Var}[T(y)] + \mathrm{Var}[U(y)] + 2\mathrm{Cov}[T(y), U(y)]$$

If $\mathrm{Cov}[T(y), U(y)] = 0$, then

$$\mathrm{Var}[h(y)] \geq \mathrm{Var}[T(y)]$$

$\Rightarrow$ Suppose$T(y)$ is minimum variance unbiased for θ and $\mathrm{E}[h(y)] = 0$. For any scalar λ,

$$\mathrm{Var}[T + \lambda h] = \mathrm{Var}[T] + 2\lambda\mathrm{Cov}[T, h] + \lambda^2\mathrm{Var}[h]$$

which is less than $\mathrm{Var}[T]$ if $2\lambda\mathrm{Cov}[T, h] + \lambda^2\mathrm{Var}[h] < 0$

If $\lambda > 0$,

$$0 < \lambda < \frac{-2\mathrm{Cov}[T, h]}{\mathrm{Var}[h]}$$

If $\lambda < 0$,

$$0 > \lambda > \frac{2\mathrm{Cov}[T, h]}{\mathrm{Var}[h]}$$

If $\mathrm{Cov}[T, f] \neq 0$ can find a λ so that T is not MVU

Corollary 6.3.11. $\mathrm{E}[T(y)] = \theta$ $\mathrm{E}[h(y)] = 0$ $T(y)$ is MVU if $\mathrm{Cov}[T(y), h(y)] \geq 0$.

PROOF: If $\mathrm{Cov}[T(y), T(y) - h(y)] \geq 0$, then $\mathrm{Var}[T(y)] = \mathrm{Cov}[T(y), h(y)]$

Corollary 6.3.12. $T(y)$ and $h(y)$ are minimum variance unbiased, then $\mathrm{Corr}[T(y), h(y)] = 1$.

PROOF: $\mathrm{Var}[T(y)] = \mathrm{Cov}[T(y), h(y)] = \mathrm{Var}[h(y)]$.

Corollary 6.3.13. $T(y) \in \mathscr{G}$ and unbiased. $T(y)$ is MVU within $\mathscr{G}$ if and only if $\mathrm{Cov}[T(y), h(y)] = 0$ for every $h \in \mathscr{G}$ with $E[h] = 0$.

PROOF: Same.

Often $\mathscr{G}$ is the class of linear estimators. In linear models $\lambda'\beta$ is a BLUE iff $\text{Cov}[\lambda'\hat{\beta}, \rho'Y] = 0$ when $\text{E}[\rho'Y] = 0$.

Corollary 6.3.14. Let $T(y) \in \mathscr{H}$. $\text{E}[T-\theta]^2$ is minimized within $\mathscr{H}$ iff $\text{E}[(T-\theta)(T-h)] = 0$ for $h \in \mathscr{H}$

PROOF: Similar

$$\text{E}[h-\theta]^2 = \text{E}[h-T]^2 + \text{E}[T-\theta]^2 + 2\text{Cov}[(h-T)(T-\theta)]$$

Corollary 6.3.15. If $T_1(y)$ and $T_2(y)$ are minimum variance unbiased for θ_1 and θ_2 then $b_1T_1(y) + b_2T_2(y)$ are minimum variance unbiased for $b_1\theta_1 + b_2\theta_2$.

PROOF: $\text{Cov}[b_1T_1(y) + b_2T_2(y), h] = 0$ if $\text{E}[h] = 0$. **look this up**

Corollary 6.3.16. Results hold for $y \sim f(v)$. If results hold for all η, with $y \sim f(v|\eta)$.

PROOF:

Theorem 6.3.17. If y is iid, any MVU is symmetric in the observations.

PROOF: iid implies that the order statistics are sufficient. Conditioning on the order statistics gives a symmetric function of the data.

$$\text{E}[h(y)|O] = \frac{\sum_{\text{permutation } p} h[p(y)}{n!}$$

Theorem 6.3.18. Generalization of Rao-Blackwell.

$$\text{E}[U-h(\theta)]^2 \geq \text{E}\{\text{E}[U|T] - h(\theta)\}^2$$

PROOF: Let $\text{E}[U] = g(\theta)$

$$\text{E}[U-h(\theta)]^2 = \text{E}[U-g(\theta)+g(\theta)-h(\theta)]^2 = \text{Var}[U-g(\theta)]^2 + [g(\theta)-h(\theta)]^2$$

$$\text{E}[\text{E}[U|T]-h(\theta)]^2 = \text{Var}[\text{E}[U|T]-g(\theta)]^2 + [g(\theta)-h(\theta)]^2$$

6.4 Scores, Information, and Cramér-Rao

We begin by introducing the the score and information functions. We then use these concepts to introduce the Cramér-Rao inequality.

For $\theta \in \Theta \subset \mathbf{R}^d$,

$$1 = \int f(v|\theta)\, dv.$$

Differentiating with respect to θ and assuming we can do it under the integral,

$$\begin{aligned} 0 &= \mathbf{d}_\theta 1 \\ &= \mathbf{d}_\theta \left[\int f(v|\theta)\, dv \right] \\ &= \int \mathbf{d}_\theta f(v|\theta)\, dv \\ &= \int \left[\frac{1}{f(v|\theta)} \mathbf{d}_\theta f(v|\theta) \right] f(v|\theta)\, dv \\ &= \mathrm{E}_{y|\theta} \left[\frac{1}{f(y|\theta)} \mathbf{d}_\theta f(y|\theta) \right]. \end{aligned}$$

Note that if $d > 1$, all of these quantities are d-dimensional row vectors. In the array of equations earlier, we divided by the density $f(v|\theta)$ but densities can be 0 over measurable sets. However, the derivative is assumed to exist, so it must also be 0, and whenever we divide by a 0 density we also multiply by a 0 density, so we can just exclude any regions of zero density from our computations.

Define the *score function* d-vector as

$$S(y;\theta) \equiv \frac{1}{f(y|\theta)} [\mathbf{d}_\theta f(y|\theta)]'.$$

Since it depends on θ, $S(y;\theta)$ is not a statistic. The score function can also be thought of as $\{\mathbf{d}_\theta \log[f(y|\theta)]\}'$.

We have just shown that $\mathrm{E}[S(y;\theta)] = 0$, so it follows that

$$\mathrm{Cov}_{y|\theta}[S(y;\theta)] = \mathrm{E}_{y|\theta}[S(y;\theta)S(y;\theta)'] \equiv \mathbf{I}(\theta),$$

where we use the equivalence to define $\mathbf{I}(\theta)$, the *information in y for* θ.

If the data are iid, the density is $f(v|\theta) = \prod_{i=1}^n f_*(v_i|\theta)$ and the score function for y becomes the sum of the score functions for the y_is, $S(y;\theta) = \sum_{i=1}^n S_*(y_i;\theta)$, because the multiplication rule Proposition F.2(a) lets us write

$$\begin{aligned} S(y;\theta) &= \frac{1}{f(y|\theta)} [\mathbf{d}_\theta f(y|\theta)]' \\ &= \frac{1}{f(y|\theta)} \left[\mathbf{d}_\theta \prod_{i=1}^n f_*(y_i|\theta) \right]' \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{f(y|\theta)} \left[\sum_{k=1}^{n} [\mathbf{d}_\theta f_*(y_k|\theta)] \prod_{i \neq k} f_*(y_i|\theta) \right]' \\
&= \frac{1}{f(y|\theta)} \left[\sum_{k=1}^{n} [\mathbf{d}_\theta f_*(y_k|\theta)] \left(\frac{f_*(y_k|\theta)}{f_*(y_k|\theta)} \right) \prod_{i \neq k} f_*(y_i|\theta) \right]' \\
&= \frac{1}{f(y|\theta)} \left[\sum_{k=1}^{n} \frac{[\mathbf{d}_\theta f_*(y_k|\theta)]}{f_*(y_k|\theta)} \prod_{i=1}^{n} f_*(y_i|\theta) \right]' \\
&= \frac{1}{f(y|\theta)} \left[\sum_{k=1}^{n} \frac{[\mathbf{d}_\theta f_*(y_k|\theta)]}{f_*(y_k|\theta)} f(y|\theta) \right]' \\
&= \frac{f(y|\theta)}{f(y|\theta)} \left[\sum_{k=1}^{n} \frac{[\mathbf{d}_\theta f_*(y_k|\theta)]}{f_*(y_k|\theta)} \right]' \\
&= \sum_{i=1}^{n} S_*(y_i;\theta).
\end{aligned}
$$

Similarly, the information matrix for y is the sum of the information matrices for the y_is, but those are all identical, so $\mathbf{I}(\theta) = n\mathbf{I}_*(\theta)$ where we use $\mathbf{I}_*(\theta)$ to denote the information in a single observation. All of this remains true if the y_is are iid r vectors so that y is actually an $rn \times 1$ vector.

Exercise 6.1.
(a) Show $I(\theta) = \mathrm{E}\left\{ [\sum_{i=1}^{n} S_*(y_i;\theta)] \left[\sum_{j=1}^{n} S_*(y_j;\theta) \right]' \right\} = n\mathbf{I}_*(\theta)$.
Hint: Independence allows you to get rid of cross-product terms.
(b) Show that $\mathbf{I}(\theta) = \mathrm{E}\left(-\{\mathbf{d}^2_{\theta\theta} \log[f(y|\theta)]\}\right)$.
Hint: Take the second derivative on each side of $1 = \int f(v|\theta)\, dv$.

The *Cramér-Rao Inequality* gives a lower bound for the variance of any estimate of a scalar parameter θ. Obviously, if you have an unbiased estimate that actually achieves the lower bound, it must be a minimum variance unbiased estimate.

The Cramér-Rao inequality involves applying the Cauchy-Schwarz inequality to a real valued estimator $g(y)$ of a real valued parameter θ and the real valued score function $S(y;\theta)$. If $g(y)$ is a possibly biased estimate of θ then

$$\theta + b_g(\theta) = \int g(v) f(v|\theta)\, dv$$

and differentiating (under the integral) wrt θ gives

$$1 + \dot{b}_g(\theta) = \int g(v) \dot{f}(v|\theta)\, dv = \int g(v) S(v;\theta) f(v|\theta)\, dv = \mathrm{E}_{y|\theta}\left[g(y) S(y;\theta) \right]. \quad (1)$$

The *Cauchy-Schwarz Inequality* states that for any random functions x and y,

$$[\mathrm{Cov}(x,y)]^2 \leq \mathrm{Var}(x)\mathrm{Var}(y).$$

Applying this to an estimate $g(y)$ of θ and the score, by Cauchy-Schwarz,

$$\{\mathrm{Cov}_{y|\theta}[g(y), S(y;\theta)]\}^2 \leq \mathrm{Var}_{y|\theta}[g(y)]\mathrm{Var}_{y|\theta}[S(y;\theta)].$$

It follows immediately that

$$\mathrm{Var}_{y|\theta}[g(y)] \geq \frac{\{\mathrm{Cov}_{y|\theta}[g(y), S(y;\theta)]\}^2}{\mathbf{I}(\theta)}.$$

Look at the covariance term. Using (1) and the fact that $0 = \mathrm{E}_{y|\theta}[S(y;\theta)]$,

$$\mathrm{Cov}_{y|\theta}[g(y), S(y;\theta)] = \mathrm{E}_{y|\theta}[g(y)S(y;\theta)] = 1 + \dot{b}_g(\theta).$$

Our final form for the *Cramér-Rao Inequality* is

$$\mathrm{Var}_{y|\theta}[g(y)] \geq \frac{[1 + \dot{b}_g(\theta)]^2}{\mathbf{I}(\theta)}.$$

If we restrict $g(y)$ to be an unbiased estimate of θ, then $b_g(\theta) = 0$, so the result reduces to

$$\mathrm{Var}_{y|\theta}[g(y)] \geq \frac{1}{\mathbf{I}(\theta)}.$$

Again, a sufficient (but not necessary) way to show that you have a minimum variance unbiased estimate of θ is to find an unbiased estimate that achieves equality in the Cramér-Rao lower bound.

Nayak (2002) discusses a similar inequality for prediction problems.

6.4.1 Information and Maximum Likelihood

The exact asymptotic distribution of the MLE depends on the information, cf. Ferguson (1996). Under suitable conditions with $\Theta \subset \mathbf{R}^d$,

$$[\mathbf{I}(\theta)]^{1/2}[\hat{\theta} - \theta] \xrightarrow{\mathscr{L}} N(0, I_d)$$

and

$$[\hat{\theta} - \theta]'[\mathbf{I}(\theta)][\hat{\theta} - \theta] \xrightarrow{\mathscr{L}} \chi^2(d).$$

Here we are using a Singular Value Decomposition to define $[\mathbf{I}(\theta)]^{1/2}$, cf. Christensen (2020). For iid data these reduce to

$$\sqrt{n}[\hat{\theta} - \theta] \xrightarrow{\mathscr{L}} N\left(0, \mathbf{I}_*(\theta)^{-1}\right)$$

and

$$n[\hat{\theta}-\theta]'[\mathbf{I}_*(\theta)][\hat{\theta}-\theta] \xrightarrow{\mathscr{L}} \chi^2(d).$$

Typically, $\hat{\theta} \xrightarrow{P} \theta$ and these relationships also hold when the information on θ is replaced with the information evaluated at $\hat{\theta}$.

6.4.2 Score Statistics

While the score function $S(y;\theta) \equiv [\mathbf{d}_\theta f(y|\theta)]'[1/f(y|\theta)]$ is not a statistic, if we replace θ with its MLE $\hat{\theta}$ we get the *score statistic*, $S(y;\hat{\theta})$. In the next chapter we will consider tests based on score statistics. Also, if we are only considering one value of θ, say $\theta = \theta_0$, we also refer to $S(y;\theta_0)$ as a (test) statistic.

6.5 Gauss-Markov Theorem

For a linear model

$$Y = X\beta + e; \quad \mathrm{E}(e) = 0; \quad \mathrm{Cov}(e) = \sigma^2 I,$$

the Gauss-Markov theorem is that for any estimable function, the least squares estimate is the best (minimum variance) linear unbiased estimate. See *Plane Answers* for details.

6.6 Exponential Families

A density function is in the *natural exponential family* if for functions c and h and a statistic $T(y)$ it can be written as

$$f(v|\theta) = c(\theta)h(v)\exp[\theta' T(v)] = h(v)\exp[\theta' T(v) - r(\theta)],$$

where $c(\theta) \equiv \exp[-r(\theta)]$ (provided that $c(\theta) > 0$). By the factorization criterion, $T(y)$ is a sufficient statistic.

The support of a distribution is the set of v values for which $f(v|\theta) > 0$. (The support is only defined up to sets of dominating measure 0.) The density of y is then equal to $f(v|\theta)$ times the indicator function of the support. If the support of the distribution depends on θ, the distribution is not in the exponential family. If the support depends on θ, we can write the support as a set $A(\theta) \neq \mathbf{R}^n$. The indicator for the support is $\mathscr{I}_{A(\theta)}(v)$ This cannot be part of $c(\theta)$ or $h(v)$ because it involves both θ and v. It also cannot be part of $\exp[\theta' T(v)]$ because $\exp[\theta' T(v)] > 0$ yet the indicator function of the support can be 0. For example, uniform distributions determined by a parametric endpoint are not members on an exponential family.

The mean of the statistic $T(y)$ can often be written in terms of $r(\theta)$.

$$1 = \int f(v|\theta)\,dv = \int h(v)\exp[\theta' T(v) - r(\theta)]\,dv.$$

If we can differentiate under the integral sign,

$$\begin{aligned} 0 = \mathbf{d}_\theta(1) &= \int h(v)\mathbf{d}_\theta\{\exp[\theta' T(v) - r(\theta)]\}\,dv \\ &= \int h(v)\exp[\theta' T(v) - r(\theta)]\left[T(v)' - \mathbf{d}_\theta r(\theta)\right]\,dv \\ &= \mathrm{E}\left[T(y)' - \mathbf{d}_\theta r(\theta)\right]. \end{aligned}$$

This leads to

$$\mathrm{E}[T(y)] = [\mathbf{d}_\theta r(\theta)]'.$$

Lehmann has given a couple of conditions under which $T(Y)$ is also complete. The following theorem gives what seems to be the simplest.

Theorem 6.6.1. [Lehmann, 1986, p.142] Suppose $y|\theta$ has a density in the natural exponential family. Then if neither θ nor $T(Y)$ is subject to a linear constraint, $T(y)$ is sufficient and complete.

PROOF: Wlog (without loss of generality) assume that $I = [-aJ \le \theta \le aJ] \subset \Theta$ where J is a vector of as and $a > 0$ is a scalar. Consider $\theta \in I$.

Suppose ν is the dominating measure for the density of y. Define $\nu_T(B) = \nu[T(v) \in B]$ and further define ν_* via $d\nu_*(t) = h_T(t)\,d\nu_T(t)$. The distribution of $T \equiv T(y)$ is determined by a density with respect to ν_T of the form

$$f(t|\theta) = c(\theta)\exp[\theta' t]\,d\nu_*(t)$$

We need to show that if $\mathrm{E}_{T|\theta}[Q(T)] = 0$ for all θ, then $P_{T|\theta}[Q(T) = 0] = 1$ for all θ. As in Appendix D.2, write $Q = Q^+ - Q^-$. Note that since $0 = \mathrm{E}_{T|\theta}[Q(T)] = \mathrm{E}_{T|\theta}[Q^+(T)] - \mathrm{E}_{T|\theta}[Q^-(T)]$,

$$\int Q^+(t)c(\theta)\exp[\theta' t]\nu_*(t) = \int Q^-(t)c(\theta)\exp[\theta' t]\,d\nu_*(t), \tag{1}$$

hence for $\theta = 0$,

$$\int Q^+(t)\,d\nu_*(t) = \int Q^-(t)\,d\nu_*(t) \equiv K.$$

It follows that $Q^+(t)/K$ and $Q^-(t)/K$ can be viewed as densities wrt $d\nu_*(t)$ and equation (1) specifies that the densities have the same moment generating function, hence the densities determine the same distribution. If the distributions are the same, the densities must be the same a.e. (ν_*), i.e., $Q^+(t) = Q^-(t)$ a.e., hence $Q(t) = Q^+(t) - Q^-(t) = 0$ a.e. (ν_*) which is enough to ensure that the function is 0 a.e. (ν)

We can broaden the class of exponential families by considering "unnatural" ones. A density function is in the *exponential family* if for functions c, h, η, and a statistic $T(y)$ it can be written as

$$f(v|\theta) = c(\theta)h(v)\exp[\eta(\theta)'T(v)] = h(v)\exp[\eta(\theta)'T(v) - r(\theta)],$$

where $c(\theta) \equiv \exp[-r(\theta)]$ and $T(y)$ is sufficient. If η is a bijection (one-to-one and onto), we can reparameterize θ into a new parameter η that is in the natural exponential family.

See *PA*, Section 2.5 for a discussion of these ideas applied to linear models.

I seem to recall that complete sufficient statistics *only exist* for exponential families.

6.7 Asymptotic Properties

Consistency, Efficiency, etc. Leave the details of these procedures to other sources on asymptotic theory like Ferguson (1996) or Lehmann(1999).

Chapter 7
Hypothesis Test Theory

In Chapter 5 we introduced hypothesis testing as part of decision theory. This involves partitioning Θ into Θ_0 (the null hypothesis) and Θ_1 (the alternative hypothesis) and only two actions $\mathscr{A} = \{a_0, a_1\}$ with a_0 accepting the null hypothesis (rejecting the alternative) and a_1 rejecting the null hypothesis (accepting the alternative). The standard loss function is

$L(\theta, a)$	a_0	a_1
$\theta \in \Theta_0$	0	1
$\theta \in \Theta_1$	1	0.

Using a_1 when $\theta \in \Theta_0$ is called a *Type I error*. Using a_0 when $\theta \in \Theta_1$ is called a *Type II error*.

If either Θ_0 or Θ_1 contains only a single value, it is referred to as a simple hypothesis. If either contains more than one value, that is called a composite hypothesis. Simple nulls are tested against both simple and composite alternatives. Composite nulls are tested against composite alternatives but rarely against simple alternatives.

We consider a test $\phi(y)$ to be a randomized decision rule in the sense that for every v, $\phi(v)$ is the probability of rejecting the null hypothesis. The loss for a randomized action $\phi(v)$ is

$$L(\theta, \phi(v)) = L(\theta, a_1)\phi(v) + L(\theta, a_0)[1 - \phi(v)].$$

The risk associated with ϕ is

$$R(\theta, \phi) = \mathrm{E}_{y|\theta}\{L(\theta, a_1)\phi(y) + L(\theta, a_0)[1 - \phi(y)]\}.$$

With the standard loss function, the *size of a test* is defined to be

$$\alpha \equiv \sup_{\theta \in \Theta_0} R(\theta, \phi) = \sup_{\theta \in \Theta_0} \mathrm{E}_{y|\theta}\{L(\theta, a_1)\phi(y)\} = \sup_{\theta \in \Theta_0} \int \phi(v) f(v|\theta)\, dv.$$

This is often called the α *level* of the test and the (maximum) probability of Type I error.

For a simple null hypothesis $\Theta_0 = \{\theta_0\}$, the size is

$$\alpha \equiv R(\theta_0, \phi) = \int \phi(v) f(v|\theta_0)\, dv.$$

This is the probability of rejection under the null hypothesis, i.e., the *probability of Type I error*.

The *probability of Type II error* is $R(\theta, \phi)$ for $\theta \in \Theta_1$ and written as,

$$\beta_\phi(\theta) \equiv \int [1 - \phi(v)] f(v|\theta)\, dv.$$

The *power of a test* is the probability of rejecting the null hypothesis when it is false. For any $\theta \in \Theta$, the *power function* (more awkwardly but correctly called the *size-power function*) is

$$\pi_\phi(\theta) \equiv \int \phi(v) f(v|\theta)\, dv = \mathrm{E}[\phi(y)].$$

Despite the name, the power function only gives the power of the test when $\theta \in \Theta_1$, whence $\pi_\phi(\theta) = 1 - R(\theta, \phi) = 1 - \beta_\phi(\theta)$. Somewhat ironically, for $\theta \in \Theta_0$, the power function is actually the size function because $\pi_\phi(\theta) = R(\theta, \phi)$ is the probability of Type I error for $\theta \in \Theta_0$. The size of the test is $\sup_{\theta \in \Theta_0} \pi_\phi(\theta)$.

A test $\tilde{\phi}$ in a class of tests $\mathscr{C}$ is *uniformly most powerful (UMP) of size α in $\mathscr{C}$* if $\alpha = \sup_{\theta \in \Theta_0} R(\theta, \tilde{\phi})$ and if for any other test $\phi \in \mathscr{C}$ with $\alpha \geq \sup_{\theta \in \Theta_0} R(\theta, \phi)$,

$$\pi_{\tilde{\phi}}(\theta) \geq \pi_\phi(\theta) \qquad \text{for all } \theta \in \Theta_1.$$

If $\mathscr{C}$ is the collection of all tests, we merely say that $\tilde{\phi}$ is uniformly most powerful.

Two restricted classes of tests are often used.

A size α test ϕ is said to be *unbiased* if

$$\sup_{\theta \in \Theta_0} \pi_\phi(\theta) \leq \inf_{\theta \in \Theta_1} \pi_\phi(\theta).$$

A *uniformly most powerful unbiased (UMPU) test* is uniformly most powerful among ϕs that are unbiased (all with size α).

Consider a group of transformations $\mathscr{G}$ that map $\mathbf{R}^n$ into $\mathbf{R}^n$. A test ϕ is said to be invariant under $\mathscr{G}$ if

$$\phi(y) = \phi[G(y)],$$

for any $G \in \mathscr{G}$ and any y. A *uniformly most powerful invariant (UMPI) test* is uniformly most powerful among tests ϕ that are invariant (all with size α).

Throughout we assume that $y \sim f(v|\theta)$ for some family of densities subject to a dominating measure, i.e., if the set A has dominating measure 0, $P_{y|\theta}(y \in A) = 0$ for all $\theta \in \Theta$.

For a simple versus simple test, because we know the two θs involved, we can and will let tests depend on $f(y|\theta_1)/f(y|\theta_0)$. For composite hypotheses, we will

need a test statistic that may or may not depend on particular values of θ. Not infrequently, test statistics depend on a value $\theta_0 \in \Theta_0$.

7.1 Simple versus Simple Tests and the Neyman-Pearson Lemma

We begin by considering a simple null hypothesis $H_0 : \theta = \theta_0$ and a simple alternative $H_1 : \theta = \theta_1$. The key result in finding a most powerful test is the Neyman-Pearson lemma. With only two distributions to test $f(v|\theta_0)$ versus $f(v|\theta_1)$ the entire conceit of introducing a parametric family in unnecessary. We could discuss this as simply testing one distribution for y, say density f, against another distribution for y, say density g. [That Lehmann (1958) used such notation took some adjusting on my part back in the 1970s.]

The goal is to fix α and find the test that maximizes the power (minimizes β). The Neyman-Pearson lemma tells us how to do this.

For some $K \geq 0$ and function $\gamma(y)$ taking values in [0,1], consider a test $\tilde{\phi}$ (randomized decision rule) that rejects the null hypothesis with probabilities specified by

$$\tilde{\phi}(y) = \begin{cases} 1 & \text{if } f(y|\theta_1) > Kf(y|\theta_0), \\ \gamma(y) & \text{if } f(y|\theta_1) = Kf(y|\theta_0), \\ 0 & \text{if } f(y|\theta_1) < Kf(y|\theta_0). \end{cases}$$

One of the nice things about this test is that it is not a very randomized rule. Only when $f(y|\theta_1) = Kf(y|\theta_0)$ do you have to resort to randomization for picking an action. Nonetheless, when y has a discrete distribution, randomized actions are a vital part of the theory.

Notice that the test $\tilde{\phi}$ can be rewritten as

$$\tilde{\phi}(y) = \begin{cases} 1 & \text{if } f(y|\theta_1) - Kf(y|\theta_0) > 0, \\ \gamma(y) & \text{if } f(y|\theta_1) - Kf(y|\theta_0) = 0, \\ 0 & \text{if } f(y|\theta_1) - Kf(y|\theta_0) < 0. \end{cases}$$

This form of the test is actually more convenient for our next proof and three examples. The test can also be written in terms of the likelihood ratio $f(y|\theta_1)/f(y|\theta_0)$ being greater than, equal to, or less than K but that form requires us to worry about what happens when $f(y|\theta_0) = 0$.

Lemma 7.1.1. *The Neyman-Pearson Lemma*
Suppose the size of $\tilde{\phi}$ is α. If $\phi(y)$ taking values in [0,1] is any other randomized decision rule with no greater size, then $\tilde{\phi}$ is at least as powerful as ϕ.

PROOF: First suppose that

$$0 \leq \int [\tilde{\phi}(v) - \phi(v)][f(v|\theta_1) - Kf(v|\theta_0)]dv. \tag{1}$$

If that is true, then

$$0 \le \int [\tilde{\phi}(v) - \phi(v)] f(v|\theta_1) dv - K \int [\tilde{\phi}(v) - \phi(v)] f(v|\theta_0)] dv. \tag{2}$$

Looking at the second term,

$$\begin{aligned}\int [\tilde{\phi}(v) - \phi(v)] f(v|\theta_0)] dv &= \int \tilde{\phi}(v) f(v|\theta_0)] dv - \int \phi(v) f(v|\theta_0)] dv \\ &= R(\theta_0, \tilde{\phi}) - R(\theta_0, \phi) = \pi_{\tilde{\phi}}(\theta_0) - \pi_{\phi}(\theta_0).\end{aligned}$$

This is the difference in the sizes of the tests, so by assumption $[R(\theta_0, \tilde{\phi}) - R(\theta_0, \phi)] \ge 0$. Since $K \ge 0$, in the second term of (2) we are subtracting a nonnegative number, and since the difference is nonnegative, the first term must be nonnegative, thus

$$\begin{aligned}0 \le &\int [\tilde{\phi}(v) - \phi(v)] f(v|\theta_1) dv \\ &= \int \tilde{\phi}(v) f(v|\theta_1) dv - \int \phi(v) f(v|\theta_1) dv = \pi_{\tilde{\phi}}(\theta_1) - \pi_{\phi}(\theta_1).\end{aligned}$$

However, this is just the difference in the powers of the two tests, so $\tilde{\phi}$ must have at least as much power as ϕ.

To establish (1) it suffices to show that $[\tilde{\phi}(v) - \phi(v)][f(v|\theta_1) - Kf(v|\theta_0)] \ge 0$. We consider three cases: when the second term is positive, negative, and 0. When $[f(v|\theta_1) - Kf(v|\theta_0)] > 0$, we have $\tilde{\phi}(y) = 1$ and since $0 \le \phi(v) \le 1$, we have $[\tilde{\phi}(v) - \phi(v)] \ge 0$. Thus $[\tilde{\phi}(v) - \phi(v)][f(v|\theta_1) - Kf(v|\theta_0)] \ge 0$. Similarly, when $[f(v|\theta_1) - Kf(v|\theta_0)] < 0$, we have $\tilde{\phi}(y) = 0$, so $[\tilde{\phi}(v) - \phi(v)] \le 0$, and $[\tilde{\phi}(v) - \phi(v)][f(v|\theta_1) - Kf(v|\theta_0)] \ge 0$. Finally, when $[f(v|\theta_1) - Kf(v|\theta_0)] = 0$, we have $[\tilde{\phi}(v) - \phi(v)][f(v|\theta_1) - Kf(v|\theta_0)] = 0$. □

For values v with $[f(v|\theta_1) - Kf(v|\theta_0)] = 0$, there are many functions $\gamma(v)$ that can give a size α most powerful test, however there always exists a constant function $\gamma(v) \equiv \gamma_0$ that will give a most powerful test. In particular, to get an α level test, take $\tilde{K}$ to be the smallest K value with $1 - \alpha \le P_{y|\theta_0}[f(y|\theta_1) \le Kf(y|\theta_0)]$. Then define $\alpha_0 \equiv P_{y|\theta_0}[f(y|\theta_1) > \tilde{K} f(y|\theta_0)] \le \alpha$. As a function of K, the function $P_{y|\theta_0}[f(y|\theta_1) \le Kf(y|\theta_0)]$ is either continuous at $\tilde{K}$ or it is not. If it is continuous, $\alpha_0 = \alpha$, and taking $\gamma_0 = 0$ we are done. If it is discontinuous then $\alpha_0 < \alpha$ and we must have $0 < \eta_0 \equiv P_{y|\theta_0}[f(y|\theta_1) = \tilde{K} f(y|\theta_0)]$. In that case take $\gamma_0 = (\alpha - \alpha_0)/\eta_0$ and we are done.

EXAMPLE 7.1.0. Test $H_0 : y \sim f(r|0)$ versus $H_1 : y \sim f(r|2)$ where, as in Chapter 3,

r	1	2	3	4
$f(r\|0)$	0.980	0.005	0.005	0.010
$f(r\|2)$	0.098	0.001	0.001	0.900

Illustrate different α, K, and $\gamma(\cdot)$ choices.

EXAMPLE 7.1.1. Test $H_0 : y \sim U[0,1]$ versus $H_1 : y \sim U[2,3]$. It is intuitively clear that one should reject H_0 if $2 \le y \le 3$ and accept H_0 if $0 \le y \le 1$. We merely illustrate that the most powerful test behaves properly. Remember we reject H_0 for positive values of $f(y|\theta_1) - Kf(y|\theta_0)$, accept H_0 for negative values, and (if necessary) randomize when $0 = f(y|\theta_1) - Kf(y|\theta_0)$. If $P_{y|\theta}[0 = f(y|\theta_1) - Kf(y|\theta_0)] = 0$ for both θs, we don't need to worry about this possibility.)

In this example,

$$f(y|\theta_1) - Kf(y|\theta_0) = \mathscr{I}_{[2,3]}(y) - K\mathscr{I}_{[0,1]}(y) = \begin{cases} 0 & \text{if } y < 0 \\ -K & \text{if } 0 \le y \le 1 \\ 0 & \text{if } 1 < y < 2 \\ 1 & \text{if } 2 \le y \le 3 \\ 0 & \text{if } 3 < y \end{cases}.$$

We always reject H_0 if $2 \le y \le 3$ because then $\tilde{\phi}(y) = 1$, so the test has power 1 regardless of the value of K.

For any $K > 0$, if $0 \le y \le 1$, we never reject because $\tilde{\phi}(y) = 0$. The other values of y cannot occur under these two models, so it doesn't matter what we do when $\mathscr{I}_{[2,3]}(y) - K\mathscr{I}_{[0,1]}(y) = 0$. This test has size 0 (and power 1).

To construct a test with size greater than 0, we need to take $K = 0$ and randomly reject values $0 \le y \le 1$ with probability α. We could reject any such value with probability α (flip a coin that gives heads with probability α), or always reject when $0 \le y < \alpha$ but never when $\alpha < y \le 1$, or always reject when $1 - \alpha \le y < 1$ but never when $0 \le y \le 1 - \alpha$. All three (and an infinite variety of others) give most powerful tests with power 1. But it would be silly to do this when a size 0, power 1 test is readily available. □

EXAMPLE 7.1.2. Test $H_0 : y \sim U[-1,1]$ versus $H_1 : y \sim N(0,1)$. Write the standard normal density as $\varphi(y) \equiv \exp(-y^2/2)/\sqrt{2\pi}$. Then

$$f(y|\theta_1) - Kf(y|\theta_0) = \varphi(y) - K\mathscr{I}_{[-1,1]}(y)/2 = \begin{cases} \varphi(y) & \text{if } y < -1 \\ \varphi(y) - K/2 & \text{if } -1 \le y \le 1 \\ \varphi(y) & \text{if } 1 < y \end{cases}.$$

We always reject when $|y| > 1$ because then $\varphi(y) > 0$ and $f(y|\theta_1) - Kf(y|\theta_0) > 0$. For any $K \ge 0$, $P[\varphi(y) - K/2 = 0] = 0$ under both hypotheses, so there no need to worry about randomized tests. We accept H_0 if $\varphi(y) - K/2 < 0$ and that depends specifically on the value of K. To get a most powerful size α test pick $y_0 = \alpha$ so that under the null uniform distribution $P[|y| < y_0] = \alpha$. Take K so that $K/2 = \varphi(y_0)$, thus $\varphi(y) - K/2 > 0$ iff (if an only if) $\varphi(y) - \varphi(y_0) > 0$ iff $|y| < y_0$. So the most powerful size α test rejects when $|y| \le \alpha$ or $|y| > 1$. □

EXAMPLE 7.1.3. Now we reverse the roles of the distributions in the previous example and test $H_0 : y \sim N(0,1)$ versus $H_1 : y \sim U[-1,1]$. Again write the standard normal density as $\varphi(y) = \exp(-y^2/2)/\sqrt{2\pi}$. We get

$$f(y|\theta_1) - Kf(y|\theta_0) = \mathscr{I}_{[-1,1]}(y)/2 - K\varphi(y) = \begin{cases} -K\varphi(y) & \text{if } y < -1 \\ 0.5 - K\varphi(y) & \text{if } -1 \le y \le 1 \\ -K\varphi(y) & \text{if } 1 < y \end{cases}.$$

For $K > 0$, you never reject when $|y| > 1$ because then $\varphi(y) > 0$ and $f(y|\theta_1) - Kf(y|\theta_0) < 0$. To get a most powerful size α test pick y_0 so that under the null standard normal distribution $P[1 \ge |y| > y_0] = \alpha$. Take $K > 0$ so that $1/2K = \varphi(y_0)$, thus for $-1 \le y \le 1$, $0.5 - K\varphi(y) > 0$ iff $\varphi(y_0) - \varphi(y) > 0$ iff $1 \ge |y| > y_0$. So the most powerful size α test rejects when $|y| \ge 1 > 1 - \alpha_2$. □

In the case $\gamma(y) = \gamma_0$, when a sufficient statistic exists, an easy application of the factorization criterion shows that the test is a function of the sufficient statistic. (And that the important parts of the test, were $\tilde{\phi}(y) \ne \gamma_y$), always are.) In fact, we can respecify the Neyman-Pearson test structure as most powerful tests having the following structure: For for any sufficient statistic $T(y)$, for some $K \ge 0$, and, as discussed earlier, some value γ_0 taking values in [0,1], a most powerful test can be written as

$$\tilde{\phi}(y) = \begin{cases} 1 & \text{if } h(y)g[T(y);\theta_1] > Kh(y)g[T(y);\theta_0], \\ \gamma_0 & \text{if } h(y)g[T(y);\theta_1] = Kh(y)g[T(y);\theta_0], \\ 0 & \text{if } h(y)g[T(y);\theta_1] < Kh(y)g[T(y);\theta_0]. \end{cases}$$

But this can be rewritten in terms $T(y)$,

$$\tilde{\phi}[T(y)] = \begin{cases} 1 & \text{if } g[T(y);\theta_1] > Kg[T(y);\theta_0], \\ \gamma_0 & \text{if } g[T(y);\theta_1] = Kg[T(y);\theta_0], \\ 0 & \text{if } g[T(y);\theta_1] < Kg[T(y);\theta_0], \end{cases}$$

so a most powerful test exists that depends only on the sufficient statistic. More generally, we can consider $T \equiv T(y)$ with density $f_T(t|\theta) = h_T(t)g(t;\theta)$ and write a most powerful test $\tilde{\phi}$ (randomized decision rule) that rejects the null hypothesis with probabilities specified by

$$\tilde{\phi}(T) = \begin{cases} 1 & \text{if } f_T(T|\theta_1) > Kf_T(T|\theta_0), \\ \gamma(T) & \text{if } f_T(T|\theta_1) = Kf_T(T|\theta_0), \\ 0 & \text{if } f_T(T|\theta_1) < Kf_T(T|\theta_0), \end{cases}$$

Clearly, any size α test of this form will also have, when considered a function of y, the form of most powerful α level test.

7.2 One-sided Alternatives

With $\Theta = \mathbf{R}$, we consider a composite one-sided alternative $H_1 : \theta > \theta_0$. We begin with the simple null, $H_0 : \theta = \theta_0$ and then consider the composite null s $H_0 : \theta \le \theta_0$. In both cases the Neyman-Pearson Lemma allows us to find UMP tests provided that we have a property called *monotone likelihood ratio*.

The changes necessary for testing $H_0 : \theta = \theta_0$ or $H_0 : \theta \geq \theta_0$ versus $H_1 : \theta < \theta_0$ are minor.

7.2.1 Monotone Likelihood Ratio

Definition 7.2.1. The densities are said to have monotone likelihood ratio if for any $\theta_1 > \theta_0$, the ratio $f(v|\theta_1)/f(v|\theta_0)$ is a monotone function in v whenever both densities are nonzero.

For the general exponential family, it suffices to have $\eta(\theta)$ increasing and $T(v)$ monotone. We tend to think in terms of increasing likelihood ratios but the theory works as well for decreasing likelihood ratios.

More importantly, the placeholder variable v in the definition is implicitly real valued, which means that the definition applies for random variables y rather than random vectors. In practice, we will apply the definition to situations in which a real valued sufficient statistic $T(y)$ exists, and require monotone likelihood ratio in the densities of the sufficient statistic. .

Theorem 7.2.1. If T has nondecreasing likelihood ratio, then any test of the form

$$\tilde{\phi}(T) = \begin{cases} 1 & \text{if } T > t_0, \\ \tilde{\gamma} & \text{if } T = t_0, \\ 0 & \text{if } T < t_0, \end{cases}$$

has nondecreasing size-power function $\mathrm{E}_{T|\theta}[\tilde{\phi}(T)]$ and is uniformly most powerful of its size for testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ for any θ_0. Moreover, for any $0 \leq \alpha \leq 1$, there exist a t_0 and $\tilde{\gamma}$ that give a size α test.

The values t_0 and $\tilde{\gamma}$ are chosen to give a size α test at $\theta = \theta_0$ but the test does not depend on $\theta_1 > \theta_0$, so if it is most powerful for any θ_1 it is most powerful for all of them.

For $\theta < \theta_0$, if we think about θ_0 as the alternative, the power is α so the size must be be less than that. More specifically, if the size at θ is ξ, then we can make a test that always rejects with probability ξ, an the most powerful test at alternative θ_0 has to have power at least as great as ξ.

PROOF: For any $\theta_0 < \theta_1$, the most powerful test has the form

$$\tilde{\phi}(T) = \begin{cases} 1 & \text{if } f_T(T|\theta_1) > K f_T(T|\theta_0), \\ \gamma(T) & \text{if } f_T(T|\theta_1) = K f_T(T|\theta_0), \\ 0 & \text{if } f_T(T|\theta_1) < K f_T(T|\theta_0), \end{cases}$$

We need to show that the test in the theorem can be written in this form.

Write $g(t) \equiv f_T(t|\theta_1)/f_T(t|\theta_0)$ which is nondecreasing, so if $t > t_0$, we have $g(t) \geq g(t_0)$. The most powerful test now involves comparing $g(T)$ to K.

Let t_0 be the smallest t with $1-\alpha \leq P(T \leq t_0) \equiv 1-\alpha_0$. The function $P(T \leq t)$ is either continuous at t_0 or is not. If continuous, $\alpha_0 = \alpha$ and take $\tilde{\gamma} = 0$. If it is discontinuous then $\alpha_0 < \alpha$ and we must have $0 < \eta_0 \equiv P_{T|\theta_0}[T = t_0]$. In that case take $\tilde{\gamma} = (\alpha - \alpha_0)/\eta_0$. To see that the form of the theorem is the form of a most powerful test, observe that the test in the theorem is

$$\tilde{\phi}(T) = \begin{cases} 1 & \text{if } g(T) > g(t_0), \\ 1 & \text{if } g(T) = g(t_0); T > t_0, \\ \tilde{\gamma} & \text{if } g(T) = g(t_0); T = t_0, \\ 0 & \text{if } g(T) = g(t_0); T < t_0, \\ 0 & \text{if } g(T) < g(t_0), \end{cases}$$

in which $K \equiv g(t_0)$ and $\gamma(T)$ is defined by the three middle cases.

Example $t \sim N(\theta, 1)$ Algebra to show that $f(t|\theta_1) > kf(t|\theta_0)$ iff $t > (\theta_1 - \theta_0)^2/2 + log(K)$.

Exercise 7.1. Assume $y \sim N(0, \sigma^2)$. Show that the problem displays monotone likelihood ratio $0 < \sigma^2 \leq 1$. Find the UMP test for $H_0 : \sigma^2 = 1$ versus $H_1 : \sigma^2 < 1$.

The composite versus composite example in Section 3.3 had a uniformly most powerful test without having monotone likelihood ratio because it was monotone where it counted. The likelihood ratios were increasing when $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_A$ but that broke down when looking at two distributions from the same hypothesis. Monotone likelihood ratio (for θs both in Θ_0) also assures that the size of a test is the size associated with largest $\theta_0 \in \Theta_0$.

7.3 Two-sided Testing

I think it was Ed Bedrick who pointed out to me that the normal theory two-sided t is a clearly reasonable thing to do. So the fact that it is UMPU is less evidence that it a reasonable test and more the case that it gives credence to UMPU being a reasonable criterion on which to base a test.

Locally best tests.

7.4 Generalized Likelihood Ratio Tests

As I recall from Lehmann's wonderful (2011) book, some people, I believe Gossett and Eagon Pearson, were dissatisfied with the fact that significance tests did not tell

you what was wrong with the null model. So E. Pearson came up with the idea of specifying an alternative hypothesis and the generalized likelihood ratio test statistic – before he and Neyman began collaborating on the theory of hypothesis testing.

This involves partitioning Θ into Θ_0 (the null hypothesis) and Θ_1 (the alternative hypothesis). The generalized likelihood ratio test statistic is

$$T(y) \equiv \frac{\sup_{\theta\in\Theta_0} L(\theta|y)}{\sup_{\theta\in\Theta} L(\theta|y)} = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})}.$$

Reject H_0 if $T(y)$ is too small.

Asymptotics: Under H_0 typically $-2\log[T(y)] \stackrel{\mathscr{L}}{\to} \chi^2(d)$ where d requires some specification.

Linear model

$$Y = X\beta + e, \qquad e \sim N(0, \sigma^2 I),$$

or

$$Y \sim N(X\beta, \sigma^2 I).$$

Least squares estimates of β are maximum likelihood, i.e., any $\hat{\beta}$ satisfying $X\hat{\beta} = MY$. Exercise 3.1 in PA is to show that the usual one-sided F test of a reduced model is equivalent to the generalized likelihood ratio test.

7.5 A Worse than Useless Generalized Likelihood Ratio Test

Geisser (2005) contains a favorite example from Hacking (1965) illustrating foundational issues related to testing.

Consider the null model

$$\Pr[X = 0|\theta = 0] = 0.9$$

$$\Pr[X = i|\theta = 0] = 0.001, i = 1, \ldots, 100.$$

The Fisherian 0.1 test of significance for this distribution rejects $H_0 : \theta = 0$ for $X = i$, $i = 1, \ldots, 100$. Observing anything other than $X = 0$ is somewhat weird, so that tends to contradict the (null) hypothesis. The Fisherian size is determined by the P value rather than the probability of type I error. Also, significance tests do not involve an alternative, so power is not an issue.

Now consider the Neyman-Pearson (N-P) problem of testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ for $\theta = 0, 1, \ldots, 100$. The null distribution is as before and the alternative sampling distributions are

$$\Pr[X = 0|\theta = i] = 0.91$$

$$\Pr[X = i|\theta = i] = 0.09, i = 1, \ldots, 100.$$

The significance test also defines a Neyman-Pearson test, so we can explore it's N-P properties. In this example, the probability of type I error is 0.1. When used in N-P testing, significance tests can have very poor power for some alternatives since they are constructed without reference to any alternative. For these alternatives, the significance test has power 0.09 regardless of the alternative, so its power is less than its size. This is not surprising. Given any test, you can always construct an alternative that will have power less than the size.

The most powerful test for an alternative $\theta > 0$ depends on θ, so a uniformly most powerful test does not exist. The significance test is also the likelihood ratio test. The likelihood ratio examines the transformation

$$\begin{aligned} T(x) &= \frac{\Pr[X = x|\theta = 0]}{\max_{i=0,\ldots,100} \Pr[X = x|\theta = i]} \\ &= \begin{cases} 0.9/0.91 = 0.989 & \text{if } x = 0 \\ 0.001/0.09 = 1/90 & \text{if } x \neq 0 \end{cases} \end{aligned}$$

and rejects for small values of the test statistic $T(X)$. That the likelihood ratio test has power less than its size IS surprising.

The uniformly most powerful invariant (UMPI) test of size .1 is a randomized test. It rejects when $X = 0$ with probability $1/9$. The size is $.9(1/9) = 0.1$ and the power is $0.91(1/9) > 0.1$. Note, however, that observing $X = 0$ does not contradict the null hypothesis because $X = 0$ is the most probable outcome under the null hypothesis. Moreover, the test does not reject for any value $X \neq 0$, even though such data are 90 times more likely to come from the alternative $\theta = X$ than from the null.

7.6 Asymptotic Test Statistics

There are three commonly used asymptotic tests for a d dimensional parametric family of distributions: the generalized likelihood ratio test, Wald tests, and score tests. With the parameter vector $\theta \in \Theta$ and the null hypothesis $\theta \in \Theta_0$, we have previously discussed the asymptotic behavior of generalized likelihood ratio tests, namely that under H_0 typically

$$-2\log[L(\hat{\theta}_0) - L(\hat{\theta})] \xrightarrow{\mathscr{L}} \chi^2(r),$$

where $\hat{\theta}_0$ is the maximum of the likelihood function $L(\theta)$ when θ is restricted to values in $\Theta_0 \subset \Theta$ and r is the difference in the dimensions between Θ and Θ_0. It is well-known from linear model theory that for multivariate normal data the usual F tests rejecting for large values are equivalent to rejecting for small values of the generalized likelihood ratio statistic. But in normal theory linear models, we can find exact distributions for the test statistics and do not need asymptotic theory.

EXAMPLE 7.6.1 *One-Sample Normal Theory.*
Consider $y_1, \ldots, y_n$ iid $N(\mu, \phi)$, where $\phi \equiv \sigma^2$. Recall that the likelihood is

$$L(\theta) = f(y|\mu,\phi) = (2\pi)^{-n/2}\phi^{-n/2}\exp\left[\frac{-1}{2\phi}\sum_{i=1}^{n}(y_i-\mu)^2\right],$$

the log-likelihood is

$$\ell(\theta) = \log[f(y|\mu,\phi)] = \frac{-n}{2}\log(2\pi) - \frac{n}{2}\log(\phi) - \frac{1}{2\phi}\sum_{i=1}^{n}(y_i-\mu)^2,$$

and the score function is

$$S(y;\mu,\phi) = \begin{bmatrix}\mathbf{d}_\mu \log[f(y|\mu,\phi)]\\ \mathbf{d}_\phi \log[f(y|\mu,\phi)]\end{bmatrix} = \begin{bmatrix}\sum_{i=1}^{n}(y_i-\mu)/\phi\\ -\frac{n}{2\phi} + \frac{1}{2\phi^2}\sum_{i=1}^{n}(y_i-\mu)^2\end{bmatrix}.$$

The maximum likelihood estimates $\hat{\mu}$ and $\hat{\phi}$ will solve the vector equation $S(y;\hat{\mu},\hat{\phi}) = 0$. Taking $\hat{\mu} = \bar{y}_\cdot$ makes the first component of the equation 0 regardless of the value of ϕ. With this $\hat{\mu}$, the second component of the equation is solved by $\hat{\phi} = \sum_{i=1}^{n}(y_i-\bar{y}_\cdot)^2/n$. It is not hard to see that when restricting θ to Θ_0, we get $\hat{\theta}_0 = (\mu_0,\tilde{\phi})'$ where $\tilde{\phi} \equiv \sum_{i=1}^{n}(y_i-\mu_0)/n$.

Some algebra shows that $\tilde{\phi} = \hat{\phi} + (\bar{y}_\cdot - \mu_0)^2$, so the generalized likelihood ratio test looks at

$$\begin{aligned}
-2\log[L(\hat{\theta}_0) - L(\hat{\theta})] &= -2\left\{\log[f(y|\mu_0,\tilde{\phi})] - \log[f(y|\hat{\mu},\hat{\phi})]\right\}\\
&= -2\left[\frac{-n}{2}\log(2\pi) - \frac{n}{2}\log(\tilde{\phi}) - \frac{1}{2\tilde{\phi}}\sum_{i=1}^{n}(y_i-\mu_0)^2\right]\\
&\quad -(-2)\left[\frac{-n}{2}\log(2\pi) - \frac{n}{2}\log(\hat{\phi}) - \frac{1}{2\hat{\phi}}\sum_{i=1}^{n}(y_i-\hat{\mu})^2\right]\\
&= \left[n\log(\tilde{\phi}) + \frac{1}{\tilde{\phi}}\sum_{i=1}^{n}(y_i-\mu_0)^2\right] - \left[n\log(\hat{\phi}) - \frac{1}{\hat{\phi}}\sum_{i=1}^{n}(y_i-\hat{\mu})^2\right]\\
&= \left[n\log(\tilde{\phi}) + \frac{1}{\tilde{\phi}}n\tilde{\phi}\right] - \left[n\log(\hat{\phi}) - \frac{1}{\hat{\phi}}n\hat{\phi}\right]\\
&= n\log(\tilde{\phi}) - n\log(\hat{\phi})\\
&= n\log[\hat{\phi} + (\bar{y}_\cdot-\mu_0)^2] - n\log(\hat{\phi})\\
&= n\log\left[1 + \frac{(\bar{y}_\cdot-\mu_0)^2}{\hat{\phi}}\right].
\end{aligned}$$

Large values of this can be rejected if larger than the α percentile of the asymptotic $\chi^2(1)$ distribution, but we can find an exact small sample distribution for this test. Large values of this statistic correspond to large values of $(\bar{y}_\cdot-\mu_0)^2/(\hat{\phi}/n)$ which linear model theory tells us has an exact $\frac{n}{n-1}F(1,n-1)$ distribution and also to large values of $(\bar{y}_\cdot-\mu_0)^2/(s_y^2/n)$ which has an exact $F(1,n-1)$ distribution and also to large absolute values of $(\bar{y}_\cdot-\mu_0)\Big/\sqrt{s_y^2/n}$ which has an exact $t(n-1)$ distribution.

Exercise 7.1a. What is wrong with the following argument for showing that the likelihood ratio test is asymptotically $\chi^2(1)$. Let $W_n = (\bar{y}_\cdot - \mu_0)/\sqrt{\hat{\phi}/n}$. Then by the CLThm and LLN and Ferguson's version of Slutsky, $W_n \xrightarrow{\mathscr{L}} Z \sim N(0,1)$. Again by Ferguson's Slutsky, $W_n^2 \xrightarrow{\mathscr{L}} Z^2 \sim \chi^2(1)$. We know if $a_n \to a$, then $\left[1+\frac{a_n}{n}\right]^n \to e^a$, so again by Ferguson's Slutsky,

$$\left[1+\frac{W_n^2}{n}\right]^n \xrightarrow{\mathscr{L}} e^{Z^2}.$$

Again by Ferguson's Slutsky,

$$n\log\left[1+\frac{W_n^2}{n}\right] \xrightarrow{\mathscr{L}} Z^2,$$

so

$$n\log\left[1+\frac{(\bar{y}_\cdot - \mu_0)^2}{\hat{\phi}}\right] \xrightarrow{\mathscr{L}} Z^2 \sim \chi^2(1).$$

as was to be shown.

7.6.1 Wald tests

Wald tests are based directly on the asymptotic distribution of an estimate of the parameter being tested. Suppose we have some d dimensional parameter θ and that for some r dimensional differentiable function ξ we want to test $H_0 : \xi(\theta) = \xi_0$. Of course this makes no sense unless ξ_0 is in the range of the function ξ, i.e., for this to make sense there must exist θ_0 such that $\xi(\theta_0) = \xi_0$.

Wald tests can be based on any sequence of $d \times 1$ random estimates $\hat{\theta}_n$ that have

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathscr{L}} N\big[0, \Sigma(\theta)\big].$$

Frequently they are based on iid data and maximum likelihood estimates in which case, typically, $\Sigma(\theta) = \mathbf{I}_*(\theta)^{-1}$.

As discussed in Example B.2.7, this convergence in law implies that $\hat{\theta}_n \xrightarrow{P} \theta$, and for $\Sigma(\theta)$ continuous (in each entry), we get $\Sigma(\hat{\theta}_n) \xrightarrow{P} \Sigma(\theta)$. (More technically, we get $\mathrm{Vec}[\Sigma(\hat{\theta}_n)] \xrightarrow{P} \mathrm{Vec}[\Sigma(\theta)]$.) In general, we can consider an arbitrary positive definite estimate $\hat{\Sigma}_n$ with $\hat{\Sigma}_n \xrightarrow{P} \Sigma(\theta)$, for which, by continuity of the inverse function, $\hat{\Sigma}_n^{-1} \xrightarrow{P} [\Sigma(\theta)]^{-1}$.

Applying the delta method of Appendix Subsection B.2.6,

$$\sqrt{n}[\xi(\hat{\theta}_n) - \xi(\theta)] \xrightarrow{\mathscr{L}} Y_w \sim N\big\{0, [\mathbf{d}_\theta \xi(\theta)]\Sigma(\theta)[\mathbf{d}_\theta \xi(\theta)]'\big\}.$$

In particular, under the null hypothesis,

$$\sqrt{n}[\xi(\hat{\theta}_n) - \xi_0] \xrightarrow{\mathscr{L}} Y_w \sim N\{0, [\mathbf{d}_\theta \xi(\theta_0)]\Sigma(\theta_0)[\mathbf{d}_\theta \xi(\theta_0)]'\}.$$

A potential problem with this covariance matrix is that if the value θ_0 with $\xi(\theta_0) = \xi_0$ is not unique, the covariance matrix, and therefore the Wald test, may depend on the choice of θ_0 and not on ξ_0 alone.

Provided that the derivative $\mathbf{d}_\theta \xi(\theta)$ is continuous at θ_0, following Example B.2.7, Ferguson's Slutsky allows us to infer that

$$\begin{aligned}\left\{\sqrt{n}[\xi(\hat{\theta}_n) - \xi_0)]\right\}' \left\{[\mathbf{d}_\theta \xi(\hat{\theta})]\hat{\Sigma}_n[\mathbf{d}_\theta \xi(\hat{\theta})']\right\}^{-1} \left\{\sqrt{n}[\xi(\hat{\theta}_n) - \xi_0)]\right\} \\ \xrightarrow{\mathscr{L}} Y_w' \left[\mathbf{d}_\theta \xi(\theta_0)\Sigma(\theta_0)\mathbf{d}_\theta \xi(\theta_0)'\right]^{-1} Y_w \sim \chi^2(r).\end{aligned}$$

Evaluating the derivative at θ_0 rather than $\hat{\theta}$ in the test statistic gives the same asymptotic distribution but, as discussed in the example below, using $\hat{\theta}$ rather than θ_0 in the estimate of the covariance matrix is often advantageous for finite samples because under the alternative hypothesis, it often gives smaller variance estimates and thus larger test statistics and greater power.

Again as in Example B.2.7, if $r = 1$ we can base a Wald test on

$$\frac{\xi(\hat{\theta}_n) - \xi_0}{\sqrt{\left\{[\mathbf{d}_\theta \xi(\hat{\theta})]\hat{\Sigma}_n[\mathbf{d}_\theta \xi(\hat{\theta})']\right\}/n}} \xrightarrow{\mathscr{L}} \frac{1}{\sqrt{[\mathbf{d}_\theta \xi(\theta_0)]\,\Sigma(\theta_0)\,[\mathbf{d}_\theta \xi(\theta_0)']}} Y_w \sim N(0,1).$$

Exercise 7.2 Find the form of the Wald test statistic when $\xi(\theta) \equiv \Lambda'\theta$ for a fixed $r \times d$ matrix Λ of rank r. Show that for $r = 1$ and $\Lambda \equiv \lambda$, the Wald test can be based on

$$\frac{\lambda'\hat{\theta}_n - \lambda'\theta_0}{\sqrt{\lambda'\hat{\Sigma}_n\lambda/n}} \xrightarrow{\mathscr{L}} N(0,1).$$

Generalizing Wald χ^2 tests to deal with singular covariance matrices is merely awkward, rather than actually difficult.

EXAMPLE 7.6.1 CONTINUED. *One-Sample Normal Theory.*
We construct Wald tests based on the MLEs. To do that we first need to find the information matrix. Using the second derivative method of Exercise 6.1,

$$-\begin{bmatrix}\mathbf{d}_\mu S(y;\mu,\phi) \\ \mathbf{d}_\phi S(y;\mu,\phi)\end{bmatrix} = \begin{bmatrix} n/\phi & \sum_{i=1}^n (y_i - \mu)/\phi^2 \\ \sum_{i=1}^n (y_i - \mu)/\phi^2 & -\frac{n}{\phi^2} + \sum_{i=1}^n (y_i - \mu)^2/\phi^3 \end{bmatrix}.$$

From linear model theory it is well known that $\mathrm{E}_{y|\mu,\phi} \sum_{i=1}^n (y_i - \mu) = 0$ and $\sum_{i=1}^n (y_i - \mu)^2/\phi \sim \chi^2(n)$, so $\mathrm{E}_{y|\mu,\phi} \sum_{i=1}^n (y_i - \mu)^2 = \phi n$, thus

$$\mathbf{I}(\mu,\phi) = -\mathrm{E}_{y|\mu,\phi}\begin{bmatrix}\mathbf{d}_\mu S(y;\mu,\phi)\\ \mathbf{d}_\phi S(y;\mu,\phi)\end{bmatrix} = \begin{bmatrix} n/\phi & 0\\ 0 & -\frac{n}{2\phi^2}+\frac{n\phi}{\phi^3}\end{bmatrix} = \begin{bmatrix} n/\phi & 0\\ 0 & \frac{n}{2\phi^2}\end{bmatrix}$$

and

$$\mathbf{I}_*(\mu,\phi) = \begin{bmatrix} 1/\phi & 0\\ 0 & \frac{1}{2\phi^2}\end{bmatrix}$$

Asymptotically,

$$\mathrm{Cov}\begin{bmatrix}\hat\mu\\ \hat\phi\end{bmatrix} \doteq \mathbf{I}(\mu,\phi)^{-1} = \begin{bmatrix}\phi/n & 0\\ 0 & \frac{2\phi^2}{n}\end{bmatrix}.$$

This agrees closely with results in linear model theory from which we know, $\bar y_\cdot \sim N(\mu,\phi/n)$, $n\hat\phi/\phi \sim \chi^2(n-1)$, so that $\mathrm{E}(n\hat\phi/\phi) = n-1$ and $\mathrm{Var}(n\hat\phi/\phi) = 2(n-1)$, with $\bar y_\cdot \perp\!\!\!\perp \sum_{i=1}^n (y_i-\bar y_\cdot)^2$. It follows that actually

$$\mathrm{Cov}\begin{bmatrix}\hat\mu\\ \hat\phi\end{bmatrix} = \begin{bmatrix}\phi/n & 0\\ 0 & \frac{2\phi^2(n-1)}{n^2}\end{bmatrix}.$$

To test $H_0 : \mu = \mu_0$, define $\xi(\theta) = (1,0)\theta = \mu$, so $\xi(\hat\theta) = (1,0)\hat\theta = \hat\mu = \bar y_\cdot$. Take $\xi_0 = (1,0)\theta_0$ where $\theta_0 = (\mu_0,\phi_0)'$. Thankfully, this test will not depend on the choice of ϕ_0. Here we are using $\mathbf{I}(\theta)$ as the asymptotic covariance matrix of $\hat\theta$ and, since $\mathbf{d}_\theta\xi(\theta) = (1,0)$, the asymptotic variance of $\xi(\hat\theta)$ is

$$\begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix} n/\phi & 0\\ 0 & \frac{n}{2\phi^2}\end{bmatrix}^{-1}\begin{bmatrix}1\\0\end{bmatrix} = \phi/n.$$

We have two choices for estimating this, the standard choice using $\hat\theta$

$$\begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix} n/\hat\phi & 0\\ 0 & \frac{n}{2\hat\phi^2}\end{bmatrix}^{-1}\begin{bmatrix}1\\0\end{bmatrix} = \hat\phi/n,$$

and an alternative choice using $\hat\theta_0$, which is

$$\begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix} n/\tilde\phi & 0\\ 0 & \frac{n}{2\tilde\phi^2}\end{bmatrix}^{-1}\begin{bmatrix}1\\0\end{bmatrix} = \tilde\phi/n.$$

The standard Wald statistic becomes

$$\frac{\bar y_\cdot - \mu_0}{\sqrt{\hat\phi/n}}.$$

The alternative choice replaces $\hat\phi$ with $\tilde\phi$. Because $\tilde\phi = \hat\phi + (\bar y_\cdot - \mu_0)^2$, the test statistic with $\hat\phi$ is larger than with $\tilde\phi$, and since both are compared to a standard normal distribution, it provides a more powerful test in small samples. Of course under the null hypothesis, $(\bar y_\cdot - \mu_0)^2 \xrightarrow{P} 0$.

Exercise 7.3 Consider two independent samples of size n with the same variance but possibly different means. Write $\theta = (\mu_1, \mu_2)'$ and $\hat{\theta}_n = (\bar{y}_{1\cdot}, \bar{y}_{2\cdot})'$. Let ξ_0 be a fixed number (like 2). Using the pooled estimate of the variance s_p^2 (cf. Christensen, 1996 or 2015), find the Wald test for $H_0 : \mu_1/\mu_2 = \xi_0$. Now find the Wald test for $H_0 : \mu_1 = \xi_0\mu_2$, i.e., test $(1, -\xi_0)(\mu_1, \mu_2)' = 0$. Compare the two test statistics. Hint: For some number K, each test statistic can be written as

$$\frac{\bar{y}_{1\cdot} - \xi_0\bar{y}_{2\cdot}}{\sqrt{\frac{s_p^2}{n}K\left(1+\xi_0^2\right)}}.$$

Exercise 7.4 In standard linear model theory

$$Y = X\beta + e, \qquad e \sim N(0, \sigma^2 I).$$

To test an estimable function $\Lambda'\beta = d$, the standard F test is based on

$$\frac{[\Lambda'\hat{\beta} - d]'[\Lambda'(X'X)^-\Lambda]^{-1}[\Lambda'\hat{\beta} - d]/r(\Lambda)}{MSE} \sim F[r(\Lambda), d!f!E].$$

Using the least squares estimates and $\hat{\Sigma}_n \equiv (MSE)\, I_n$, show that the Wald test statistic has a known small sample distribution based on

$$[\Lambda'\hat{\beta} - d]'[(MSE)\,\Lambda'(X'X)^-\Lambda]^{-1}[\Lambda'\hat{\beta} - d] \sim r(\Lambda)F[r(\Lambda), d!f!E].$$

7.6.2 Score Tests

We now consider score tests. Recall from Section 6.4 that the score function is a d vector

$$S(y;\theta) \equiv \frac{1}{f(y|\theta)}[\mathbf{d}_\theta f(y|\theta)]' = [\mathbf{d}_\theta \log f(y|\theta)]',$$

with

$$\mathrm{E}[S(y;\theta)] = 0; \qquad \mathrm{Cov}[S(y;\theta)] \equiv \mathbf{I}(\theta).$$

For independent observations y_i with density $f_*(v_i|\theta)$, score function $S_*(y_i;\theta)$, and information $\mathbf{I}_*(\theta)$, we established that

$$S(y;\theta) = \sum_{i=1}^{n} S_*(y_i;\theta); \qquad \mathbf{I}(\theta) = n\mathbf{I}_*(\theta).$$

Now define

$$\bar{S}(\theta) = \frac{1}{n}\sum_{i=1}^{n} S_*(y_i;\theta) = \frac{1}{n}S(y;\theta),$$

then by the Central Limit Theorem

$$\sqrt{n}\left[\bar{S}(\theta)-0\right] \xrightarrow{\mathscr{L}} Y_s \sim N[0,\mathbf{I}_*(\theta)].$$

Again applying Example B.2.7, we get

$$\left\{\sqrt{n}\left[\bar{S}(\theta)\right]\right\}'\mathbf{I}_*(\theta)^{-1}\left\{\sqrt{n}\left[\bar{S}(\theta)\right]\right\} \xrightarrow{\mathscr{L}} Y_s'\mathbf{I}_*(\theta)^{-1}Y_s \sim \chi^2(d).$$

This result can be rewritten as

$$S(y;\theta)'\mathbf{I}(\theta)^{-1}S(y;\theta) \xrightarrow{\mathscr{L}} \chi^2(d)$$

and often holds even for non-iid data.

To test the hypothesis $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ the α level *score test* rejects when

$$S(y;\theta_0)'\mathbf{I}(\theta_0)^{-1}S(y;\theta_0) \geq \chi^2(1-\alpha,d).$$

If $d = 1$, the score test can be based on the approximation $S(y;\theta_0)/\sqrt{\mathbf{I}(\theta_0)} \sim N(0,1)$.

For testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ the α level score test rejects when

$$S(y;\hat{\theta}_0)'\mathbf{I}(\hat{\theta}_0)^{-1}S(y;\hat{\theta}_0) \geq \chi^2(1-\alpha,r),$$

where $\hat{\theta}_0$ is the value that maximizes the likelihood among values $\theta \in \Theta_0$ and r is the difference in the dimensions between the spaces Θ and Θ_0. Typically, $S(y;\hat{\theta}) = 0$ because the maximum of the log-likelihood would be in the interior of Θ, but typically $S(y;\hat{\theta}_0) \neq 0$ because the maximum of the log-likelihood would be on a boundary of Θ_0. (Remember $S(y;\theta)$ is defined for the full model. For a redefined null model score function, typically, $S_0(y;\hat{\theta}_0) = 0$.) Sometimes an exact small sample distribution can be found for the score test statistic.

An advantage of score tests is that you only have to find the MLE under the null model. If you want to do model selection among generalized linear models, doing score tests to compare various models with the smallest model of interest is an easy, but ill advised, way to proceed, cf. Christensen (2025, Section 4.4).

EXAMPLE 7.6.1 CONTINUED. *One-Sample Normal Theory.*
Now consider the score test of $H_0 : \mu = \mu_0$. In this case $\Theta = \{(\mu,\phi)|\mu \in \mathbf{R}, \phi > 0\}$ and $\Theta_0 = \{(\mu_0,\phi)|\phi > 0\}$. It is not hard to see that $S(y;\hat{\theta}_0) = S(y;\mu_0,\tilde{\phi})$ where $\tilde{\phi} \equiv \sum_{i=1}^n (y_i-\mu_0)/n$.

$$S(y;\mu_0,\tilde{\phi}) == \begin{bmatrix} \sum_{i=1}^n (y_i-\mu_0)/\tilde{\phi} \\ -\frac{n}{2\tilde{\phi}} + \frac{1}{2\tilde{\phi}^2}\sum_{i=1}^n (y_i-\mu_0)^2 \end{bmatrix} = \begin{bmatrix} (\bar{y}_\cdot-\mu_0)/(\tilde{\phi}/n) \\ -\frac{n}{2\tilde{\phi}} + \frac{n}{2\tilde{\phi}^2}\tilde{\phi} \end{bmatrix} = \begin{bmatrix} (\bar{y}_\cdot-\mu_0)/(\tilde{\phi}/n) \\ 0 \end{bmatrix}.$$

The information needed for the test is

$$\mathbf{I}(\hat{\theta}_0) = \begin{bmatrix} n/\tilde{\phi} & 0 \\ 0 & \frac{n}{2\tilde{\phi}^2} \end{bmatrix}$$

The asymptotic test statistic is

$$\begin{aligned} S(y;\hat{\theta}_0)'\mathbf{I}(\hat{\theta}_0)^{-1}S(y;\hat{\theta}_0) &= \begin{bmatrix} (\bar{y}_\cdot - \mu_0)/(\tilde{\phi}/n) \\ 0 \end{bmatrix}' \begin{bmatrix} n/\tilde{\phi} & 0 \\ 0 & \frac{1}{2\tilde{\phi}^2} \end{bmatrix}^{-1} \begin{bmatrix} (\bar{y}_\cdot - \mu_0)/(\tilde{\phi}/n) \\ 0 \end{bmatrix} \\ &= [(\bar{y}_\cdot - \mu_0)/(\tilde{\phi}/n)]^2[\tilde{\phi}/n] \\ &= \frac{(\bar{y}_\cdot - \mu_0)^2}{\tilde{\phi}/n} \end{aligned}$$

This gets compared to a $\chi^2(1)$ distribution because Θ is a two-dimensional space and Θ_0 is a one-dimensional space. The $\chi^2(1)$ test is the same as a two-sided test based on

$$\frac{\bar{y}_\cdot - \mu_0}{\sqrt{\tilde{\phi}/n}} \stackrel{\mathscr{L}}{\to} Z \sim N(0,1). \tag{1}$$

This is the same as the alternative Wald test, which we saw is less powerful than the standard Wald test for this problem.

7.6.3 Comparison of Tests

Exercise 6.1 on Fisher's information establishes that

$$\mathbf{I}(\theta) = \mathrm{E}\left(-\{\mathbf{d}^2_{\theta\theta}\log[f(y|\theta)]\}\right).$$

There is another concept known as the *observed information* defined as

$$\mathbf{J}(\theta|y) \equiv -\{\mathbf{d}^2_{\theta\theta}\log[f(y|\theta)].$$

For iid data with pdf $f_*(v|\theta)$, we have

$$\mathbf{J}_*(\theta|y_i) \equiv -\{\mathbf{d}^2_{\theta\theta}\log[f_*(y_i|\theta)]$$

and $\mathbf{J}(\theta|y) = \sum_{i=1}^n \mathbf{J}_*(\theta|y_i)$. Since $\mathbf{I}_*(\theta) = \mathrm{E}[\mathbf{J}_*(\theta|y_i)]$, by the LLN,

$$\frac{1}{n}\mathbf{J}(\theta|y) \stackrel{a.s.}{\to} \mathbf{I}_*(\theta)$$

and we can use the observed information $\mathbf{J}(\theta|y)$ as an approximation for the Fisher information $\mathbf{I}(\theta)$.

Figure 7.1 gives a plot of the log-likelihood function $\ell(\theta)$ that I stole from Cox (2006) and need to replace with one of my own construction. It illustrates the relationships between the generalized likelihood ratio, maximum likelihood Wald, and score test statistics for $d = r = 1$ and a simple null hypothesis $H_0 : \theta = \theta_0$. Because $r = 1$, we can be looking at two-sided normal tests. The plot illustrates the numerators of these tests. Each test needs to be standardized appropriately; the standardizations are different for each. The log-likelihood function is maxi-

mized at $\hat{\theta}$ and for large samples under the null hypothesis, $\hat{\theta}$ should be close to θ_0. Let $Z \sim N(0,1)$. The generalized likelihood ratio test statistic is based on $W_L = \ell(\hat{\theta}) - \ell(\theta_0)$ with $\sqrt{2W_L} \sim |Z|$. The maximum likelihood Wald statistic is based on $W_E = \hat{\theta} - \theta_0 \sim N(0, \mathbf{J}(\hat{\theta})^{-1})$ where the variance is approximated by the inverse of the negative second derivative of the $\ell(\theta)$ function at $\hat{\theta}$ but could be replaced by the negative inverse second derivative at θ_0 since the two points are asymptotically close to one another under the null hypothesis. Finally, the score test looks at the slope of the tangent line evaluated at θ_0, i.e. $W_S = S(y;\theta_0) \sim N(0, \mathbf{J}(\theta_0))$, and is standardized by the square root of the negative second derivative of $\ell(\theta)$ evaluated at θ_0. Note that with increased sample sizes the $\ell(\theta)$ curve should get more sharply peaked, so the second derivatives should get to be larger negative numbers, but also the slope of the tangent line should also get steeper.

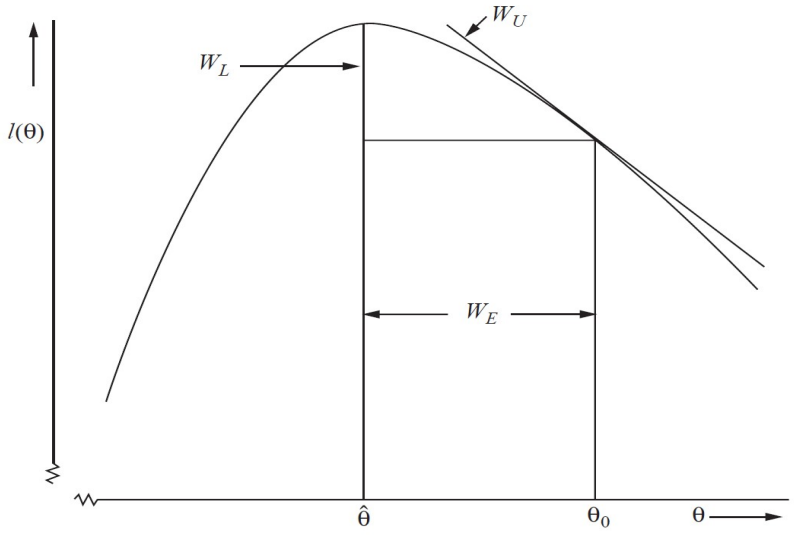

Fig. 7.1 Graph of a $d = 1$ log-likelihood function with numerators of the generalized likelihood ratio, maximum likelihood Wald, and score normal test statistics. **Taken from Cox (2006).** Need to make my own version.

Chapter 8
UMPI Tests for Linear Models

We examine the transformations necessary for establishing that the linear model F test is a uniformly most powerful invariant (UMPI) test. We also note that the Studentized range test for equality of groups means in a balanced one-way ANOVA is not invariant under all of these transformations so the UMPI result says nothing about the relative powers of the ANOVA F test and the Studentized range test. The discussion uses notation from Christensen (2020).

8.1 Introduction

It has been well-known for a long time that the linear model F test is a uniformly most powerful invariant (UMPI) test. Lehmann (1959) discussed the result in the first edition of his classic test and in all subsequent editions, e.g. Lehmann and Romano (2005). But the exact nature of this result is a bit convoluted and may be worth looking at with some simpler and more modern terminology.

Consider a (full) linear model

$$Y = X\beta + e, \qquad e \sim N(0, \sigma^2 I)$$

where Y is an n vector of observable random variables and consider a reduced (null) model

$$Y = X_0\gamma + e, \qquad C(X_0) \subset C(X),$$

where $C(X)$ denotes the column (range) space of X. Let M be the perpendicular projection operator (ppo) onto $C(X)$ and M_0 be the ppo onto $C(X_0)$. The usual F test statistic, which is equivalent to the generalized likelihood test statistic, is

$$F(Y) \equiv F \equiv \frac{Y'(M-M_0)Y/[r(X)-r(X_0)]}{Y'(I-M)Y/[n-r(X)]},$$

where $r(X)$ denotes the rank of X.

Consider a group of transformations $\mathscr{G}$ that map $\mathbf{R}^n$ into $\mathbf{R}^n$. A test statistic $T(Y)$ is invariant under $\mathscr{G}$ if

$$T(Y) = T[G(Y)],$$

for any $G \in \mathscr{G}$ and any Y. It is not too surprising that the F statistic is invariant under location and scale transformations. Specifically, if we define $G(Y) = a(Y + X_0\delta)$ for any positive real number a and any vector δ, it is easy to see using properties of ppos that $F(Y) = F[G(Y)]$. Unfortunately, this is not the complete set of transformations required to get the UMPI result. Note also that the location invariance is defined with respect to the reduced model, that is, it involves X_0. Given that the alternative hypothesis is the existence of a location $X\beta \neq X_0\gamma$ for any γ, one would not want a test statistic that is invariant to changes in the alternative, particularly changes that could turn the alternative into the null.

Before discussing the third group of transformations required for a UMPI F test, let's look at the best known alternative to the linear model F test. Consider a balanced one-way ANOVA model,

$$y_{ij} = \mu_i + \varepsilon_{ij}, \qquad \varepsilon_{ij} \text{ iid } N(0, \sigma^2),$$

$i = 1, \ldots, a$, $j = 1, \ldots, N$. The F statistic for $H_0 : \mu_1 = \cdots = \mu_a$ is

$$F(Y) \equiv F \equiv \frac{N\sum_{i=1}^{a}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2/[a-1]}{\sum_{i=1}^{a}\sum_{j=1}^{N}(\bar{y}_{ij} - \bar{y}_{i\cdot})^2/[a(N-1)]} \equiv \frac{MSGrps}{MSE}.$$

The best known competitor to an F test for H_0 is the test that rejects for large values of the studentized range,

$$Q(Y) \equiv Q \equiv \frac{\max_i \bar{y}_{i\cdot} - \min_i \bar{y}_{i\cdot}}{\sqrt{MSE/N}}.$$

We already know that F is location and scale invariant and it is easy to see that Q is too. In this case, location invariance means that the test statistic remains the same if we add a constant to every observation. Moreover, it is reasonably well known that neither of these tests is uniformly superior to the other, which means that Q must not be invariant under the full range of transformations that are required to make F a UMPI test.

We can decompose Y into three orthogonal pieces,

$$Y = M_0Y + (M - M_0)Y + (I - M)Y = X_0\hat{\gamma} + (X\hat{\beta} - X_0\hat{\gamma}) + (Y - X\hat{\beta}). \qquad (1)$$

The first term of the decomposition contains the fitted values for the reduced model. The second term is the difference between the fitted values of the full model and those of the reduced model. The last term is the residual vector from the full model. Intuitively we can think of the transformations that define the invariance as relating to the three parts of this decomposition. The residuals are used to estimate σ, the scale parameter of the linear model, so we can think of scale invariance as relating to $(I - M)Y$. The translation invariance of adding vectors $X_0\delta$ modifies M_0Y. To get

the UMPI result we need another group of transformations that relate to $(M-M_0)Y$. Specifically, we need to incorporate *rotations* of the vector $(M-M_0)Y$ that keep the vector *within* $C(M-M_0) = C(X_0)^{\perp}_{C(X)}$, the orthogonal complement of $C(X_0)$ with respect to $C(X)$. If we allow rotations of $(M-M_0)Y$ within $C(M-M_0)$, the end result can be any vector in $C(M-M_0)$ that has the same length as $(M-M_0)Y$. The length of a vector v is $\|v\| \equiv \sqrt{v'v}$. The end result of a rotation within $C(M-M_0)$ can be, for any n vector v with $\|(M-M_0)v\| \neq 0$,

$$\frac{\|(M-M_0)Y\|}{\|(M-M_0)v\|}(M-M_0)v.$$

Finally, the complete set of transformations to obtain the UMPI result is for any positive number a, any appropriate size vector δ, and any n vector v with $\|(M-M_0)v\| \neq 0$,

$$G(Y) = a\left[M_0Y + X_0\delta + (I-M)Y + \frac{\|(M-M_0)Y\|}{\|(M-M_0)v\|}(M-M_0)v\right].$$

Again, it is not difficult to see that $F(Y) = F[G(Y)]$.

However, in the balanced ANOVA problem, there exist such transformations G with $Q(Y) \neq Q[G(Y)]$, so Q is not invariant under these transformations and when we say that F is UMPI, it says nothing about the relative powers of F and Q. We know that Q is invariant to location and scale changes, so it must be the rotation that Q is not invariant to. Let J_m be an m dimensional vector of 1s. In a one-way ANOVA, write $Y = [y_{11}, y_{12}, \ldots, y_{aN}]'$ and

$$\begin{aligned}(M-M_0)Y &= X\hat{\beta} - X_0\hat{\gamma} \\ &= \begin{bmatrix} J_N & & & 0 \\ & J_N & & \\ & & \ddots & \\ 0 & & & J_n \end{bmatrix} \begin{bmatrix} \bar{y}_{1\cdot} \\ \bar{y}_{2\cdot} \\ \vdots \\ \bar{y}_{a\cdot} \end{bmatrix} - J_{aN}\bar{y}_{\cdot\cdot} = \begin{bmatrix} (\bar{y}_{1\cdot} - \bar{y}_{\cdot\cdot})J_N \\ (\bar{y}_{2\cdot} - \bar{y}_{\cdot\cdot})J_N \\ \vdots \\ (\bar{y}_{a\cdot} - \bar{y}_{\cdot\cdot})J_N \end{bmatrix}. \quad (2)\end{aligned}$$

Since Y is an arbitrary vector, $(M-M_0)v$ must display a similar structure. Also

$$\|(M-M_0)Y\|^2 = N\sum_{i=1}^{a}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 \equiv SSGrps. \quad (3)$$

Thinking about the decomposition in (1), if $Q(Y)$ were invariant we should get the same test statistic if we replace M_0Y with $M_0Y + X_0\delta$ (which we do) and if we replace $(M-M_0)Y$ with $[\|(M-M_0)Y\|/\|(M-M_0)v\|](M-M_0)v$ (which we do not). The numerator of Q is a function of $(M-M_0)Y$, namely, it takes the difference between the largest and smallest components of $(M-M_0)Y$. For $Q(Y)$ to be invariant, the max minus the min of $(M-M_0)Y$ would have to be the same as the max minus the min of $[\|(M-M_0)Y\|/\|(M-M_0)v\|](M-M_0)v$ for any Y and v. Alternatively, the max minus the min of $[1/\|(M-M_0)Y\|](M-M_0)Y$ would have to be the same

as the max minus the min of $[1/\|(M-M_0)v\|](M-M_0)v$ for any Y and v. In other words, given (2) and (3),

$$\frac{\max_i \bar{y}_{i\cdot} - \min_i \bar{y}_{i\cdot}}{\sqrt{SSGrps}}$$

would have to be a constant for any data vector Y, which it is not.

Chapter 9
Significance Testing for Composite Hypotheses

In their most natural form, significance tests typically lead to two-sided t and χ^2 tests. We review significance tests as probabilistic proofs by contradiction. We emphasize an appropriate definition of p values for significance testing and compare it to alternate definitions. We introduce the concept of composite significance tests, and illustrate that they can generate one-sided tests. We review interval estimation for both parameters and predictions based on significance testing and illustrate that one-sided interval estimates can be constructed from composite significance tests. Finally, we address the issue of multiple comparisons in the context of significance testing.

9.1 Introduction

Schervish, M.J. (1996), "P-Values: What They Are and What They Are Not," The American Statistician, 50, 203-206.

Fisher (1956) p. 94 seems to be saying the the significance of a composite hypothesis is the significance of each individual test. This can differ radically from N-P probability of type I error. $H_0 : (A \cap B)^c$ is true. Reject at level α if both A^c and B^c rejected at level α. Example really lends itself to an alternative $H_1 : A \cap B$ is true. Probability of type I error is much smaller than significance level.

Fisher was certainly interested in one side rejection regions!

Fisher 1925 p. 80-81 In preparing this table we have borne in mind that in practice we do not always want to know the exact value of P for any observed χ^2, but, in the first place, whether or not the observed value is open to suspicion. If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. Belief in the hypothesis as an accurate representation of the population sampled is confronted by the logical disjunction: *Either* the hypothesis is untrue *or* the value of χ^2 has attained by chance an exceptionally high value. The actual value of P obtainable from the table by interpolaton indicates the strength of the evidence

against the hypothesis. A value of χ^2 exceeding the 5 per cent point is seldom to be disregarded.

a paragraph later

The term Goodness of Fit has caused some to fall into the fallacy of believing that the higher the value of P the more satisfactorily is the hypothesis verified. Values over .999 have sometimes been reported which, if the hypothesis were true, would only occur once in a thousand trials. Generally such cases are demonstrably due to the use of inaccurate formulae, but occasionally small values of χ^2 beyond the expected range do occur, as in Ex. 4 with the colony numbers obtained in the plating method of bacterial counting. In these cases the hypothesis considered is as definitely disproved as if P had been .001.

Significance (Fisherian) tests are probabilistic versions of proof by contradiction. A probabilistic model is assumed and observed data are either deemed to be inconsistent with the model, which suggests that the model is wrong, or the data are consistent with the model, which suggests very little. Data consistent with the assumed model are almost always equally consistent with other models. The extent to which the data are consistent with the model is measured using a p value with small p values indicating data that are inconsistent with the model. Significance testing is distinct from the Neyman-Pearson theory of hypothesis testing, see Christensen (2005).

Significance tests for unimodal distributions typically yield two-sided tests. We begin with a discussion of simple significance tests and alternative definitions of p values. Section 3 discusses extensions of simple significance tests that include the possibility of one-sided tests. Section 4 discusses interval estimation with emphasis on the significance testing interpretation of prediction intervals and one-sided intervals. Finally, Section 5 briefly addresses the role of multiple comparison corrections in significance testing.

9.2 Simple Significance Tests

The standard form for significance testing assumes some model for a data vector $y = (y_1, \ldots, y_n)'$. A statistic $W \equiv W(y)$ with a known distribution is chosen to serve as a test statistic. Denote W's (possibly discrete) density $f(w)$. The model is called in question if the observed value of the test statistic looks weird relative to the density $f(w)$. Denote the observed data y_{obs} and let $W_{obs} = W(y_{obs})$.

To illustrate, assume $y_1, \ldots, y_n$ iid $N(0, \sigma^2)$, then a standard test statistic is

$$T \equiv T(y) \equiv \frac{\bar{y}_\cdot - 0}{s/\sqrt{n}} \sim t(n-1),$$

where $\bar{y}_\cdot$ and s are the sample mean and standard deviation, respectively. The weirdest data are those that correspond to small densities under the $t(n-1)$ distribution.

The $t(n-1)$ density decreases away from 0, so the weirdest observations are far from 0.

The p value is the probability of observing a test statistic as weird or weirder than we actually saw. In the illustration, because the $t(n-1)$ density is symmetric about 0, with $T_{obs} \equiv T(y_{obs})$ the p value is

$$p = \Pr[T \leq -|T_{obs}|] + \Pr[T \geq |T_{obs}|].$$

A small p value suggests that something is wrong with the model. Perhaps the mean is not 0 but perhaps the data are not normal, are not independent, or are heteroscedastic. Interestingly, this two-sided $t(n-1)$ test, when using the alternative test statistic T^2, corresponds to a one-sided $F(1,n-1)$ test, because the mode of an $F(1,n-1)$ distribution is at 0.

In general, with $W_{obs} \equiv W(y_{obs})$, the p value is

$$p = \Pr[f(W) \leq f(W_{obs})]. \tag{1}$$

For a less standard illustration, assume $y_1, \ldots, y_n$ iid $N(\mu, 4)$. With a test statistic

$$W = \frac{(n-1)s^2}{4} \sim \chi^2(n-1),$$

denote the density $\chi^2(w|n-1)$. For $n-1 > 2$, unless W_{obs} happens to be the mode, there are two values $w_1 < w_2$ that have

$$\chi^2(W_{obs}|n-1) = \chi^2(w_1|n-1) = \chi^2(w_2|n-1).$$

(One of w_1 and w_2 will be W_{obs}.) Here

$$p = \Pr[W \leq w_1] + \Pr[W \geq w_2].$$

For $n = 12$ and $s^2 = 6.0$, we get $W_{obs} = 16.50$ and

$$.0332 = \chi^2(4.21|11) = \chi^2(16.50|11),$$

so

$$p = \Pr[W \leq 4.21] + \Pr[W \geq 16.50] = .04 + .12 = .16.$$

While the machinations needed to find the p value may seem a bit complicated, they are simpler than those necessary to find a Neyman-Pearson theory two-sided uniformly most powerful unbiased test, see Lehmann (1997, pp.139, 194).

Although we are actually performing a test of the entire model that has been assumed for the data, under the influence of Neyman-Pearson theory, these illustrations are often called tests of the null hypotheses $H_0 : \mu = 0$ and $H_0 : \sigma^2 = 4$, respectively. In Neyman-Pearson theory, the hypotheses $H_0 : \mu = 0$ and $H_0 : \sigma^2 = 4$ are also referred to as composite hypotheses because the first test does not specify σ^2 and the second test does not specify μ. However, in our two examples the model and test statistic pairs provide simple significance tests.

The p value is universally accepted as being the probability of seeing data as weird or weirder than were actually observed. The problem is that there are alternative ideas of what it means to be weird. In (2.1) we used the value of the density from our assumed (null) model to determine what is weird. Cox (2005, p.32) indicates that a p value is the probability of seeing data as indicative or more indicative of "a departure from the null hypothesis of subject matter interest." The common method of applying p values to Neyman-Pearson theory posits an alternative hypothesis and weird data are those that have certain extreme values of the likelihood ratio.

The key idea of a test of significance and the related idea of a p value is that we are looking for data that contradict the null hypothesis. The following example, introduced briefly in Christensen (2008), shows that these alternative ideas of weird data do not lead to a probabilistic proof by contradiction for the null model.

Take our null model as

$$f(y) = \begin{cases} .001, & y = 0 \\ .01, & y = 1, \ldots, 95 \\ .049, & y = 96 \\ 0, & y = 97 \end{cases}.$$

With one observation take y as the test statistic, so from (2.1) the possible p values are

$$p(y_{obs}) = \begin{cases} .001, & y_{obs} = 0 \\ .951, & y_{obs} = 1, \ldots, 95 \\ 1, & y_{obs} = 96 \\ 0, & y_{obs} = 97 \end{cases}.$$

Among observations that have positive probability under f, $y = 0$ is the most inconsistent with the model and $y = 96$ is most consistent with the model. All the other observations $y = 1, \ldots, 95$ are equally weird as determined by $f(y)$, so the probability of seeing something as weird or weirder than, say, $y_{obs} = 2$ is just $\Pr[y \neq 96]$. We now illustrate that defining a p value relative to some formal or informal alternative is to abandon completely the idea of a proof by contradiction for the null model.

For a Neyman-Pearson most powerful test, we need an alternative. Take it to be

$$g(y) = \begin{cases} 0, & y = 0 \\ .001, & y = 1, \ldots, 95 \\ .1, & y = 96 \\ .805, & y = 97 \end{cases}.$$

A Neyman-Pearson test rejects the null model for the smallest values of the likelihood ratio

$$\frac{f(y)}{g(y)} = \begin{cases} \infty, & y = 0 \\ 10, & y = 1, \ldots, 95 \\ .49, & y = 96 \\ 0, & y = 97 \end{cases},$$

so y values are monotonically "weirder" (relative to the alternative) as they get larger. A Neyman-Pearson p value, call it $\tilde{p}$ to distinguish it from the significance

test p value, is the probability under the null model of seeing data with a likelihood ratio as small or smaller than that actually observed, hence

$$\tilde{p}(y_{obs}) = \begin{cases} 1, & y_{obs} = 0 \\ .999, & y_{obs} = 1, \ldots, 95 \\ .049, & y_{obs} = 96 \\ 0, & y_{obs} = 97 \end{cases}.$$

The point is that observing a 96 in no way contradicts the null model, it is the observation most likely to be observed, yet $\tilde{p}$ is small. On the other hand, observing 0 tends to contradict the null model, but $\tilde{p}$ is large.

Cox's (2005) approach seems to rely on the existence of an informal alternative suggesting that, say, larger values of the test statistic provide more evidence against the null model. In this case the p value, here called $\check{p}$, as a function of y_{obs} becomes

$$\check{p}(y_{obs}) = \begin{cases} 1, & y_{obs} = 0 \\ .049 + .01(96 - y_{obs}), & y_{obs} = 1, \ldots, 95 \\ .049, & y_{obs} = 96 \\ 0, & y_{obs} = 97 \end{cases}.$$

This is similar to $\tilde{p}$ in that $\check{p}$ again rejects the null model for the data most consistent with it ($y_{obs} = 96$) and fails to reject for data that are most inconsistent with the null model ($y_{obs} = 0$).

The significance test is designed as a probabilistic proof by contradiction of the the null model. In the parlance of philosophy of science, it is a probabilistic method for falsifying the null model. The evidence against the null is appropriately measured by p with small values required to conclude that the data are inconsistent with the null model. Neither $\tilde{p}$ nor $\check{p}$ provides the basis for such a proof by contradiction. The other "p" values provide appropriate, if not good, measures for evaluating the evidence between the null and some alternative. Since they do not provide a proof by contradiction of the null model, they seem to display an unnatural focus on the null hypothesis. With a formal alternative available, there seems to be little reason to focus on the null hypothesis as opposed to the alternative, hence little reason to restrict one's self to tests in which the p value or probability of Type I error is small. In a decision procedure, one should play off the probabilities of both Type I and Type II error so that both are reasonable. See Christensen (2005) for more discussion of testing null versus alternative hypotheses and in particular the virtues of Bayesian testing for evaluating the weight of evidence between the hypotheses.

The difficulty with significance testing is picking a test statistic with a known distribution. Fisher (1956, p.50) suggests that the choice of W should be made subject to the analyst's prior information. However, the requirement that W have a known distribution under the model can be quite restrictive. Composite significance testing discussed in the next section both relaxes this assumption and can generate familiar one-sided tests for common problems.

9.3 Composite Significance Tests

We now present a significance testing approach to expanded models that can generate familiar one-sided tests. We begin with two simple examples.

Suppose our model is that the data come from one of the distributions

$$y \sim N(\mu, 1); \qquad \mu < 0. \tag{1}$$

With one observation, take y as the test statistic. There are no distributions in this model that would make observing a y of 2 or anything larger than a 2 plausible. On the other hand, any observation less than 0 is completely plausible. The trick is deciding what values between 0 and 2 are plausible and how to quantify that idea. Clearly, for observations that are positive, the probability of seeing something as weird or weirder than we actually saw is appropriately measured by the probability that a standard normal is larger than y_{obs}. We take the position that all values for y_{obs} of 0 or less are perfectly consistent with the model, hence the p value as a function of y_{obs} is discontinuous at 0 jumping from 1 to .5. (A case could be made that that p should increase continuously to a value approaching 1 for huge negative y_{obs}.)

In significance testing one typically only specifies a null model. There is no alternative model. Although one could specify an alternative to model (3.1) that is simply "not model (3.1)," that alternative cannot be specified as $\mu \geq 0$ or even $y \sim N(\mu, 1)$; $\mu \geq 0$. We might observe $y \geq 2$ because the true distribution is a Cauchy centered at 0. There is about a 15% chance of seeing such data from a Cauchy. The point is that seeing $y \geq 2$ makes model (3.1) implausible. In the absence of other assumptions, it does not suggest what the true model might be. If you are willing to make such assumptions, you should do Bayesian or Neyman-Pearson testing.

Now suppose our model is

$$y \sim N(\mu, 1); \qquad -1 < \mu < 0.$$

Not only are y values of 2 and larger implausible for all allowable μ but values of -3 and lower are correspondingly implausible. The p value for observing 2 should be the same as the p value for observing -3. An appropriate quantification seems to be

$$p(\mu) = \Pr[y \leq -3] + \Pr[y \geq 2]$$

computed under either $\mu = -1$ or 0. The two values are both .024. Another reasonable choice for computing the p value could be $\mu = -.5$ but that minimizes $p(\mu)$. To ensure that the data contradict the model, the appropriate p value is the largest of the values $p(\mu)$, i.e.,

$$p = \sup_{-1<\mu<0} p(\mu).$$

If there is any value for μ that makes $p(\mu)$ large, the data do not contradict the model.

If an observation is consistent with any of the individual models in this hypothesis, the observation is consistent with the entire collection of models. If the observation is inconsistent with every individual model, it contradicts them all. This suggests that the relevant feature in determining which data are weird is the largest density that the observation can achieve under the model. The largest density for $y = 2$ is the same as the largest density for $y = -3$. To be weirder than another observation, the largest density must be smaller. The largest density for $y = 2.5$ occurs when $\mu = 0$. It is not only smaller than the largest density for 2, which also occurs when $\mu = 0$, but it is also smaller than the largest density for -3, which occurs when $\mu = -1$.

The largest density orders the values of y into observations that are less weird, as weird, or weirder than y_{obs}. The p value is the probability of seeing data as weird or weirder than we actually saw. To contradict the model, the p value must be small for every specific distribution in the model, so an appropriate quantification is to look at the largest of the p values.

General Definitions

We now present general terminology for composite significance tests and then return to the simple normal illustration.

A simple significance test, like those illustrated in Section 2, is based on a test statistics $W(y)$ that has a known distribution under the null model. In our illustrations of composite significance testing, we had a test statistic y but a variety of distributions for y that depend on a parameter. For the general discussion, we use an equivalent formulation in which the test is based on a function of the data and the parameter, a function that has a known distribution. Suppose we have a model for y with densities $q(y|\theta)$ for $\theta \in \Theta_0$. A composite significance test is based on a test function $W(y;\theta)$ such that when θ is the true parameter, the test function has a known density $f(w)$. Note that if, say, $1 \in \Theta_0$, then $W(y;1)$ is a random variable but $W(y;1) \sim f(w)$ only if $\theta = 1$.

The largest density for $W(y;\theta)$ is denoted

$$f_*(y) \equiv \sup_{\theta \in \Theta_0} f(W(y;\theta)).$$

The function f_* orders observations y by how weird they are relative to the null model. f_* is not a function of θ.

To compute a probability of obtaining data as weird or weirder than we actually saw, we pick a θ in Θ_0 and compute

$$p(\theta) = \Pr_{y|\theta}[f_*(y) \leq f_*(y_{obs})].$$

For the data to contradict the null model, this number must be small for every $\theta \in \Theta_0$, so the p value is defined as

$$p = \sup_{\theta \in \Theta_0} p(\theta) = \sup_{\theta \in \Theta_0} \Pr_{y|\theta}[f_*(y) \leq f_*(y_{obs})]. \tag{2}$$

Although p is a well defined quantity, it may not be easy to compute.

Returning to $y \sim N(\mu, 1)$ with $-1 < \mu < 0$, define the test function $Z(y;\mu) \equiv y - \mu$. Under the model, the test function has a $N(0,1)$ distribution, so the relevant density is $f(z) \equiv \phi(z) = \exp(-z^2/2)/\sqrt{2\pi}$.

The supremum of the densities will be identical for $-1 \leq y \leq 0$ with $f_*(y) = \phi(0)$. For $y \leq -1$, the supremum occurs when $\mu = -1$ and for $y \geq 0$ the supremum occurs when $\mu = 0$, thus the supremum of the densities is

$$f_*(y) = \begin{cases} \phi(y+1) & \text{for } y \leq -1, \\ \phi(0) & \text{for } -1 \leq y \leq 0, \\ \phi(y) & \text{for } y \geq 0. \end{cases}$$

Suppose $y_{obs} = 2$. We want to compute (3.2) with $f_*(y_{obs}) = f_*(2) = \phi(2)$. By symmetry, $y = -3$ is just as weird as $y = 2$ because $f_*(-3) = \phi(-3+1) = \phi(-2) = \phi(2) = f_*(2)$. It is not difficult to see that values $y < -3$ and $y > 2$ have $f_*(y) < f_*(2)$. It follows that $p = \sup_{-1<\mu<0} p(\mu)$ where

$$p(\mu) = \Pr_{y|\mu}[f_*(y) \leq f_*(2)] = \Pr_{y|\mu}[y \leq -3] + \Pr_{y|\mu}[y \geq 2].$$

The supremum occurs at $\mu = -1$ and 0 with

$$p = 0.001 + 0.023 = 0.024.$$

Composite significance tests provide a significance testing justification for one-sided t tests. Suppose our model is $y_1, \ldots, y_n$ iid $N(\mu, \sigma^2)$ with $\mu < 0$ using a test based on

$$T(y;\mu) \equiv \frac{\bar{y}_\cdot - \mu}{s/\sqrt{n}} \sim t(n-1). \tag{3}$$

The maximum $t(n-1)$ density is identical for every value with $\bar{y}_{obs} \leq 0$ and the p value will be 1. For $\bar{y}_\cdot > 0$, the supremum of the densities is the density for $\mu = 0$, and the supremum of the p values is the p value computed under $\mu = 0$. For $T(y_{obs};0) = 2$ and $n - 1 = 111$, we get the usual one-sided value $p = .0240$.

One-sided χ^2 tests are generated in a similar fashion.

Generalizing the test of $y \sim N(\mu, 1)$ with $-1 < \mu < 0$ to a corresponding test for a sample of normal observations with an unknown variance gets somewhat complicated and is relegated to the appendix. Although the development is complicated, it is far less complicated than the corresponding Neyman-Pearson theory.

9.4 Interval Estimation

Fisher was never comfortable with Neyman-Pearson confidence intervals, hence his development of fiducial intervals, see Fisher (1956). I think that interval estimates can be developed in a reasonable manner based on significance tests.

Significance tests are fundamentally based on p values. The standard procedure with a significance test is to report a p value, the evidence that the data are consistent with the model. To extend significance tests to interval estimates we first need the concept of an α level significance test. For $\alpha \in [0,1]$, define an α level significance test as a test that rejects the null model whenever $p \leq \alpha$. If the test is not rejected, we say that the data are consistent with the null model as determined by an α level test.

To get a two-sided t interval estimate, we consider a t test of the null model $y_1, \ldots, y_n$ iid $N(\mu_0, \sigma^2)$. The null model can be decomposed into the (base) model $y_1, \ldots, y_n$ iid $N(\mu, \sigma^2)$ and the parametric null hypothesis $H_0 : \mu = \mu_0$. Under this construction, the usual two-sided $(1-\alpha)100\%$ interval is precisely the set of all μ_0 values that are consistent with both the data and the model as determined by an α level significance test.

This idea applies whenever we can separate the null model into two parts: a (base) model and a parametric null hypothesis indexed by some parameter λ, for which we have available a simple significance test for every $\lambda = \lambda_0$. A $1-\alpha$ interval (actually, a "regional") estimate consists of all parameter values λ_0 that are consistent with the data and the model as determined by an α level test.

Technically, we are specifying a collection of models that are consistent with the data. In the normal example, there is a collection of models $y_1, \ldots, y_n$ iid $N(\mu_0, \sigma^2)$ that are consistent with the data. If we could find the distribution of the T statistic for Cauchy data with median μ, we could also discuss the collection of Cauchy models that are consistent with the data. The significance testing procedure is not telling us that we should believe the normal model, it is just telling us what the reasonable μ_0 values are, *if you believe the normal model*. Nonetheless, it is convenient to refer to the collection of models by specifying the parameter λ, hence the "interval estimate." (There is little reason to call this a $1-\alpha$ interval estimate rather than an α level estimate except that they often correspond to Neyman-Pearson $1-\alpha$ confidence intervals.)

Similarly, to construct a prediction interval for the normal model, we test whether an independent new observation $y_f \sim N(\mu, \sigma^2)$ is consistent with the data and the model. The standard α level significance test takes the form of rejecting if

$$\frac{|y_f - \bar{y}_\cdot|}{s\sqrt{1+\frac{1}{n}}} > t_{1-\alpha/2}(n-1)$$

where $t_\alpha(n-1)$ denotes the 100α percentile of the $t(n-1)$ distribution. The test would be executed upon observing all of $y_1, \ldots, y_n, y_f$. Treating y_f as the indexing parameter for the tests, the $1-\alpha$ prediction interval consists of all y_f values that are

consistent with both the model and the observed data $y_1, \ldots, y_n$ as determined by an α level test. The result is the standard prediction interval $\bar{y}_\cdot \pm t_{1-\alpha/2}(n-1)s\sqrt{1+\frac{1}{n}}$.

The interpretation of the significance testing interval as of a collection of parameters that are consistent with both the data and the model does not actually presume the model to be true. However, it is a small step to making that assumption, which in turn would allow a Bayesian or Neyman-Pearson analysis.

Finally, consider the collection of models

$$y_1, \ldots, y_n \text{ iid } N(\mu, \sigma^2); \qquad \mu < \mu_0$$

indexed by μ_0 with the associated t test. The model would not be rejected by an α level composite significance test for any μ_0 above the value that has

$$T(y_{obs}; \mu_0) = \frac{\bar{y}_{obs} - \mu_0}{s_{obs}/\sqrt{n}} = t_{1-\alpha}(n-1)$$

or $\mu_0 = \bar{y}_{obs} - t_{1-\alpha}(n-1)s_{obs}/\sqrt{n}$. This serves as the composite significance testing lower $1-\alpha$ bound for μ_0. It provides an infinite interval estimate for μ_0, not for μ. The one-sided interval tells us that μ_0, the upper bound on plausible μ values, must be at least $\mu_0 = \bar{y}_{obs} - t_{1-\alpha}(n-1)s_{obs}/\sqrt{n}$. This is a reasonable interpretation, but a far cry from the usual intuition of a one-sided interval.

A good Neyman-Peason-ite would correctly (if perhaps uselessly) interpret a one-sided confidence interval in terms of its long-run frequency of covering the true parameter μ. Nonetheless, a one-side confidence interval might be thought to contain a collection of parameter values μ that are reasonable. That is not the case! The data are never going to be consistent with any infinite interval of μ values. Suppose $\bar{y}_{obs} = 16$, $s_{obs} = 4$, and $n = 16$, so $t_{.95}(15) = 1.753$ and the .95 one-sided interval is $(14.25, \infty)$. This is not an interval of μ values that are consistent with the data because with these data $T(y; 116) = -100$. The data are clearly inconsistent with the normal model having $\mu = 116$ even though 116 is well within the one-sided interval. The large positive values contained in the one-sided interval can only be deemed consistent with the data as plausible values of μ_0, that is, as plausible upper bounds for μ.

9.5 Multiple Comparisons

It is by no means clear that significance testing has anything worthwhile to say about multiple comparisons problems. The interval estimates discussed in the previous section involve performing an infinite number of tests, but they involve no corrections for multiple testing.

Multiple comparison procedures are designed to control the (weak) experimentwise error rate, that is, to control the probability of rejecting any null hypothesis in a group of tests when all of the null hypotheses are true. Significance tests are de-

signed to measure how strange a set of data are relative to a null model. What does that have to do with the probability of errors in multiple tests? The principals of significance testing can be applied to multiple comparisons if we view the multiple tests as defining one overall test. If you want to be able to make statements about which individual hypotheses are correct or incorrect, you need to make stronger assumptions and use Neyman-Pearson or Bayesian procedures. But significance testing can perhaps help in identifying individual hypotheses that contribute to the evidence that the overall null model is wrong.

The very notion of evaluating the results of a collection of individual tests is contrary to the nature of significance testing. In significance testing, the collection of tests need to be combined into an overall measure of the evidence against some null model. This usually amounts to combining the individual tests into one test of a collective null model. (The overall null model may be nothing more than the collection of null models associated with the individual tests.)

Consider the common problem of significance testing for outliers in a normal linear model. For each of n data points, we get an associated t statistic, say $t_{i,obs}$, which is one observation on a random variable t_i that has a $t(dfE-1)$ distribution where dfE is the degrees of freedom for Error when fitting the model. The random variables t_i typically are correlated. If we came into the problem with a suspicion that case i' might be outlier, we could do a standard t test for that one case, comparing $t_{i',obs}$ to a $t(dfE-1)$ distribution to obtain a p value.

More commonly, we scan through the n different t_i statistics to see if any of them have large absolute values. In essence, we base our conclusion on the value of $\max_i |t_{i,obs}|$. Recall Fisher's dictum that one chooses the test statistic based on one's ideas about what may go wrong with the model. However, the test is still just a test of whether the data (as summarized by the test statistic) are consistent with the null model.

To compute a p value, we compare the number $\max_i |t_{i,obs}|$ to the distribution of the maximum of the random variables $|t_i|$. Finding the distribution of $\max_i |t_i|$ is difficult but clearly its density is maximized at 0 and decreases monotonically away from 0. (Clearly the density of $\max_i t_i$ is symmetric about 0 and decreases monotonically away from 0.) Therefore,

$$p = \Pr\left[\max_i |t_i| \geq \max_i |t_{i,obs}|\right].$$

The maximum of the $|t_i|$s is at least as large as $\max_i |t_{i,obs}|$ if (and only if) any one of the $|t_i|$s is as large as our observed value so

$$p = \Pr\left[\bigcup_i \left(|t_i| \geq \max_i |t_{i,obs}|\right)\right],$$

and by Bonferroni's inequality (finite subadditivity)

$$p \leq \sum_{i=1}^{n} \Pr\left[|t_i| \geq \max_i |t_{i,obs}|\right] = n\Pr\left[|t_i| \geq \max_i |t_{i,obs}|\right].$$

The naive individual p value computed from a t distribution can be as much as n times larger than the appropriate p value. Thus, to ensure that the overall p value is less than .05 when $n = 45$, we require, to quote Fisher (1935, p.60), the individual test p value "to be as small as 1 in 900, instead of 1 in 20, before attaching statistical significance" to the result.

Fisher was actually discussing the results of testing mean differences that were suggested by the data in analysis of variance, but the principles are identical. In significance testing, the issue is not how many tests you are making, the issue is using a distribution for an overall test statistic that is appropriate for the test procedure. For example, in balanced analysis of variance Tukey's method for multiple comparisons is based on knowing the distribution of the studentized range. The maximum difference between studentized means provides a significance test for the model that is rejected only if the largest and smallest observed means are too far apart. The issue of concluding which means are different and which are not is something significance testing does not formally address. Nonetheless, if one can safely assume that everything about the null model is true other than the equality of group population means, it is not difficult to infer using Tukey's method, with a level of evidence borrowed from Tukey's significance test, that some specific population means are different.

In significance testing, even interval estimation becomes a problem involving multiple testing. A significance testing interval estimate has a clear definition. It consists of the parameter values that would not be rejected by an α level test, thus it involves simultaneously testing a collection of null models for which no multiple comparison correction is needed or desirable. For example, a t interval involves testing $\mu = \mu_0$ for every value of μ_0, an infinite number of tests. Significance testing does not use multiple comparison corrections; it uses different test statistics and p values that are appropriate for those statistics. However, if one is willing to make the leap of assuming that only one aspect of the null model can be wrong, interval estimates can be interpreted as evidence in favor of certain parameter values, and tests like Tukey's can be interpreted as evidence that some population means are different. Both procedures require one to specify a fixed level of evidence, α. Finally, for these fixed α uses of significance tests, as for Neyman-Pearson and Bayesian procedures, we are well served by validating our model assumptions as far as practicable. Our conclusions are only ever as good as our model assumptions.

Appendix; More General Illustrations of Composite Significance Tests

While most of the ideas for defining a composite significance test are apparent from the previous simple normal illustration of Section 3, the notation was developed to

handle more complicated problems. Actually, the test function is $W(y;\eta)$ for $\eta \in \Theta_0$ with $W(y;\theta) \sim f(w)$. While this assumption is enough to define the test procedure, to actually compute the p value we need to think of $W(y;\eta)$ as a random variable for each fixed η. We know the distribution of $W(y;\theta)$ but we also need the distribution of $W(y;\eta)$ when the parameter is θ. The necessity of these requirements become clearer when dealing with t tests but we first introduce the ideas in the context of the simple normal example.

With $Z(y;\eta) \equiv y-\eta$, the general definition of f_* is $f_*(y) \equiv \sup_{-1<\mu<0} \phi(Z(y;\mu))$. The earlier analysis allows us to rewrite f_* as

$$f_*(y) = \begin{cases} \phi(Z(y;-1)) & \text{for } y \le -1, \\ \phi(0) & \text{for } -1 \le y \le 0, \\ \phi(Z(y;0)) & \text{for } y \ge 0. \end{cases}$$

The idea that $y=-3$ is just as weird as $y=2$, becomes

$$f_*(-3) = \phi(Z(-3;-1)) = \phi(-2) = \phi(2) = \phi(Z(2;0)) = f_*(2).$$

Because normal distributions with known variance are tractable, we were able to compute the p value earlier. Nonetheless, the p value can be rewritten in terms of $Z(y;\eta)$ as

$$\begin{aligned} p(\mu) &= \Pr_{y|\mu}[y \le -3] + \Pr_{y|\mu}[y \ge 2] \\ &= \Pr_{y|\mu}[y+1 \le -2] + \Pr_{y|\mu}[y \ge 2] \\ &= \Pr_{y|\mu}[Z(y;-1) \le -2] + \Pr_{y|\mu}[Z(y;0) \ge 2]. \end{aligned}$$

To compute this last expression we need to know the distributions of the random variables $Z(y;-1)$ and $Z(y;0)$ for all μ between -1 and 0. Again, because of the simple nature of this problem, the distributions of $Z(y;-1)$ and $Z(y;0)$ are readily available for all μ. In the next example, the equivalent random variables have noncentral t distributions.

For the model $y_1,\ldots,y_n$ iid $N(\mu,\sigma^2)$ with $-1<\mu<0$ we choose the t test function

$$T(y;\mu) \equiv \frac{\bar{y}_\cdot - \mu}{s/\sqrt{n}} \sim t(n-1).$$

To contradict the model, the data must be inconsistent with each parameter within the model, that is, $T(y_{obs};\mu)$ must be inconsistent with the $t(n-1)$ distribution for every value of μ allowed in the model. Denote the $t(n-1)$ density $t(\cdot|n-1)$. The supremum of the densities is

$$f_*(y) = \begin{cases} t\left(\frac{\bar{y}_\cdot+1}{s/\sqrt{n}}|n-1\right) \equiv t\left(T(y;-1)|n-1\right) & \bar{y}_\cdot \le -1 \\ t(0|n-1) & -1 \le \bar{y}_\cdot \le 0 \\ t\left(\frac{\bar{y}_\cdot}{s/\sqrt{n}}|n-1\right) \equiv t\left(T(y;0)|n-1\right) & \bar{y}_\cdot \ge 0 \end{cases}.$$

Suppose $\bar{y}_{obs}$ and s^2_{obs} are such that $T(y_{obs};0) = 2$. We must then have $\bar{y}_{obs} > 0$, so

$$f_*(y_{obs}) = t(2|n-1).$$

For any data with $T(y;0) > 2$, we again have $\bar{y}_\cdot > 0$, so $f_*(y) = t(T(y;0)|n-1) < t(2|n-1)$. Also, for any data with $T(y;-1) \le -2$, we must have $\bar{y}_\cdot < -1$ and by symmetry $f_*(y) = t(T(y;-1)|n-1) \le t(2|n-1)$.

To find the p value defined by (3.2) we need to maximize the probability that $T(y;-1) \le -2$ or $T(y;0) \ge 2$, that is, maximize the probability of

$$\left[\frac{\bar{y}_\cdot + 1}{s/\sqrt{n}} \le -2\right] \bigcup \left[\frac{\bar{y}_\cdot - 0}{s/\sqrt{n}} \ge 2\right],$$

over all parameter values in the model. These are disjoint sets, so we can compute the probabilities separately. More formally, define

$$\begin{aligned} p(\mu) &= \Pr_{y|\mu}[f_*(y) \le f_*(y_{obs})] \\ &= \Pr_{y|\mu}[t(T(y;-1)|n-1) \le t(2|n-1); \bar{y}_\cdot < -1] \\ &\quad + \Pr_{y|\mu}[t(T(y;0)|n-1) \le t(2|n-1); \bar{y}_\cdot > 0] \\ &= \Pr_{y|\mu}[T(y;-1) \le -2] + \Pr_{y|\mu}[T(y;0) \ge 2]. \end{aligned}$$

Again,

$$p = \sup_{-1<\mu<0} p(\mu).$$

To compute p we need the distributions of $T(y;-1)$ and $T(y;0)$ for μs between -1 and 0. For $\mu = 0$, $p(\mu)$ is the probability that a central $t(n-1)$ exceeds 2 and a noncentral t with parameter 1 is below -2. When $\mu = -1$, $p(\mu)$ is the probability that a central $t(n-1)$ is below -2 and a noncentral t with parameter -1 is above 2. Obviously, they are the same number. Other parameters μ give smaller values but involve using two noncentral $t(n-1)$ distributions. With $n-1 = 111$,

$$p = .0240 + .0015 = .0255.$$

Again, this may seem complicated, but the Neyman-Pearson theory for this test is considerably more complicated, see Hodges and Lehmann (1954, Sec.3).

If null distributions get stochastically larger as θ increases, Θ_0 and interval, do maximum densities occur at the ends of the interval? Relate to mode.

Replace Cox example with $X \sim N(0,\sigma^2)$ $H_0 : \sigma^2 = 1$ Fisherian test totally different from N-P $H_A : \sigma^2 < 1$.

Look at Fisher's exact test as a one-sided test, especially for extreme outcomes using negative binomial distribution. For extreme data, the sup is at the boundary. See Yung-Pin Chen (2011) Do the Chi-Square Test and Fisher's Exact Test Agree in Determining Extreme for 2×2 Tables?, The American Statistician, 65:4, 239-245,

Look at chi-square test assuming chi-squared distribution for test statistic. $f_*(y) = \sup_p \text{indep}\, \chi^2_3(X^2(y,p))$ I think this should be achieved, for any y at the minimum

chi-square value, $f_*(y) = \chi^2_3(X^2(y,\hat{p}(y)))$. Then show that $X^2(y,\hat{p}(y))$ is a monotone function of $f_*(y) = \chi^2_3(X^2(y,\hat{p}(y)))$ or maybe that the chi-squared 1 distribution is monotone for chi-square 3. Would work like a charm if we were only doing one-sided. Or that there is virtually no probability of a chi-square 3 density getting small when the distribution is actually chi-squared 1. that is, if $X^2(y,\hat{p}(y)) \sim \chi^2_1$ then

$$P(\chi^2_3(X^2(y,\hat{p}(y))) < \chi^2_3(X^2(y_{obs},\hat{p}(y_{obs}))) \doteq P(X^2(y,\hat{p}(y)) > X^2(y_{obs},\hat{p}(y_{obs}))$$

y values that get a chi-squared 3 density small have virtually no probability under a chi-squared 1. Or does it work backwards????

$\sup_{p \text{ indep}} \chi^2_3(X^2(y,p))$ occurs at the same place as $\sup_{p \text{ indep}} \log[\chi^2_3(X^2(y,p))]$ function is decreasing in $X^2(y,p)$ because derivative is

If you find the sup by minimizing the test statistic, then small values will become a problem. $\sup_{p \text{ indep}} \chi^2_3(X^2(y,p)) = \chi^2_3(\inf_{p \text{ indep}} X^2(y,p))$

seems like this needs to be comparing looking at noncentral chi-squares with lower df and central with higher df or vice versa. Simplest is

$$Y = X\beta + e, e \sim N(0,1)$$

so

$$\|Y - X\beta_0\|^2 \sim \chi^2_n \qquad \|Y - X\hat{\beta}\|^2 \sim \chi^2_{n-r(X)}$$

Relative to conclusions in The Fisher/Pearson Chi-Squared Controversy: A Turning Point for Inductive Inference
Author(s): Davis Baird
Source: The British Journal for the Philosophy of Science, Vol. 34, No. 2 (Jun., 1983), pp.105-118
who says testing results are different. For Baird, he should reject for the minimum of the test statistics, which would have a chi-squared 1 distribution. http://www.clarku.edu/faculty/facultybio.cfm?id=893

There are three ways to test $Y \sin N(X\beta, \sigma^2 I)$ with known variance. Let the density be $\psi(y|\beta)$.

$$W_1(Y,\beta) \equiv \psi(y|\beta) \qquad W_2(Y) \equiv W_2(Y,\beta) \equiv \|Y - X\hat{\beta}\|^2/\sigma^2 \sim \chi^2(n-r) \qquad W_3(Y,\beta) \equiv \|Y - X\beta\|^2/\sigma^2$$

The first leads to a one-sided $\chi^2(n-r)$ test, the second trivially leads to a two-sided $\chi^2(n-r)$ test because the test statistic does not depend on β, not sure what the third leads to. Have to find the density of W_3, maximize it relative to β (hopefully this is a function of W_2) and find which values of the maximized density are the smallest. Not that W_3 has a $\chi^2(n)$ distribution but that is only relevant for finding the value of β that maximizes the density and then determining the values of the maximized density that are smaller than the observed maximized density.

References

Box, George E. P. (1980). Sampling and Bayes' Inference in Scientific Modelling and Robustness. *Journal of the Royal Statistical Society*. Series A (General), 143(4), 383-430.

Christensen, Ronald (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician*, **59**, 121-126.

Christensen, R. (2008). Review of "Principles of Statistical Inference" by D. R. Cox, *Journal of the American Statistical Association*, **103**, 1719-1723.

Cox, D.R. (2006). *Principles of Statistical Inference*. Cambridge University Press, New York.

Hodges, J.L., Jr. and Lehmann, E.L. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society, Series B*, **16**, 261-268.

Fisher, Ronald A. (1925). *Statistical Methods for Research Workers*, Fourteenth Edition, 1970. Hafner Press, New York.

Fisher, R. A. (1935). *The Design of Experiments*, Ninth Edition, 1971. Hafner Press, New York.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*, Third Edition, 1973. Hafner Press, New York.

Hubbard, Raymond and Bayarri, M. J. (2003). Confusion over measures of evidence (ps) versus errors (αs) in classical statistical testing. *The American Statistician*, **57**, 171-177.

Lehmann, E.L. (1997) *Testing Statistical Hypotheses*, Second Edition. Springer, New York.

Chapter 10
Thoughts on prediction and cross-validation.

Suppose we have a random vector (y, x') where y is a scalar random variable and want to use x to predict y. We do this by defining some predictor function $f(x)$. We also have a prediction loss function $L[y, f(x)]$ that allows us to evaluate how well a predictor does. Want f that minimizes

$$\mathrm{E}_{y,x}\{L[y, f(x)]\}$$

which is called the *expected prediction loss* or the *expected prediction error*. Also, for whatever predictor we end up using, we want to be able to estimate the expected prediction error.

TWO EXAMPLES:

The loss function determines the best predictor. These problems are equivalent to Bayesian decision problems if we just think of y as θ, the marginal distribution of y as the prior of θ, and the prediction loss as the decision loss. In this context,

$$\mathrm{E}_{y,x}\{L[y, f(x)]\}$$

is the Bayes risk of the decision problem and the optimal predictor will be the Bayes decision rule.

The most common loss function for prediction is squared error, see PA Section 6.3,

$$L[y, f(x)] = [y - f(x)]^2$$

from which it follows that the optimal estimator is the "posterior" mean

$$m(x) \equiv E(y|x).$$

For the special case when $y \sim \mathrm{Bern}(p)$, write $p(x) \equiv E(y|x)$. The use of squared error loss leads to estimates of the expected prediction error called Brier scores. Another option called Hamming loss is

$$L[y, f(x)] = 1 - I_{\{0\}}[y - f(x)].$$

In other words, for Hamming loss if you predict y correctly there is no loss and if you predict it incorrectly the loss is 1. The expected prediction error is just the probability of mispredicting y, i.e.,

$$\mathrm{E}_{y,x}\{L[y,f(x)]\} = \mathrm{Pr}_{y,x}\{y \neq f(x)]\}$$

Note that with Hamming loss, it makes no sense to predict a value other than 0 or 1, so these will be referred to as valid predictions. Hamming loss is equivalent to the Bayes test procedure with $y = 1$ the alternative hypothesis and $y = 0$ the null. The optimal prediction is equivalent to rejecting when the posterior probability of the alternative is greater that 0.5, i.e., the optimal rule δ has

$$\delta(x) = \begin{cases} 1 & \text{if } p(x) > 0.5 \\ 0 & \text{if } p(x) < 0.5 \end{cases}.$$

We don't care which valid prediction we make (action we take) when $p(x) = 0.5$. This rule clearly minimizes the loss for each x but using Bayes Theorem one can also show that it has the form of the N-P Lemma so is a most powerful test. □

Note that

$$\mathrm{E}_{y,x}\{L[y,\delta(x)]\} = \mathrm{Pr}_{y,x}\{y \neq \delta(x)]\} = \int_{\{x|p(x)\geq 0.5\}} [1-p(x)]f(x)dx + \int_{\{x|p(x)<0.5\}} p(x)f(x)dx.$$

These rules depend on knowing the joint distribution of (y,x'), which is generally unknown in prediction problems. We want to use data to estimate both $E(y|x)$ and $\mathrm{E}_{y,x}\{L[y,m(x)]\}$. Suppose $(y,x'),(y_1,x_1'),\ldots,(y_n,x_n')$ are iid. Let Y be the vector of y_is and let X is the matrix with x_i' as its ith row.

Estimate $E(y|x)$.

A nonparametric approach to estimating $E(y|x)$ is to identify x_i values that are close to x and estimate $E(y|x)$ by taking a weighted mean of the y_is that correspond to close x_is. Obviously, the weights on the y_i might well depend on how far the x_is are from x. This is called a nearest-neighbor approach.

Quite generally, one can assume that $E(y|x)$ is a member of a parametric family, say $m(x;\theta)$ and use a maximum likelihood estimate of θ, say $\hat{\theta}$. In this set-up, the x_i are treated as fixed and the distributions of y_i given x_i are assumed independent and to be in a parametric family of distributions (largely) determined by its mean. This is already in the form of nonlinear regression but standard generalized linear models also fit this paradigm. Nonparametric regression techniques based on basis functions such as polynomials, wavelets, or sines and cosines also fit into the generalized linear model paradigm.

In general, we end up with an estimate

$$\hat{m}(x) \equiv m(x;\hat{\theta}).$$

.

Estimate $\mathrm{E}_{y,x}\{L[y,m(x)]\}$.

If we know m, an unbiased estimate is

$$\frac{1}{n}\sum_{i=1}^{n} L[y_i, m(x_i)]. \tag{1}$$

Generally, we have to estimate m, so we might use

$$\frac{1}{n}\sum_{i=1}^{n} L[y_i, \hat{m}(x_i)]. \tag{2}$$

Since $\hat{m}$ is a complex function of the data, the expected value of this function is hard to find. Conventional wisdom is that (2) underestimates the true expected prediction error, e.g.,

$$\mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n} L[y_i, \hat{m}(x_i)]\right\} \leq \mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n} L[y_i, m(x_i)]\right\}.$$

I wonder if Eaton's methods might be able to show this? (We will show not only that this is true for linear models but that cross-validation can be even more biased upwards.)

To "fix" this problem, people try Cross-Validation. Life is much easier if we have one set of (training) data from which to estimate $\mathrm{E}(y|x)$ and a different set of (test) data from which to estimate $\mathrm{E}_{y,x}\{L[y, m(x)]\}$. In such a case, $\hat{m}$ based on the training data is a fixed predictor with regard to the test data so equation (1) gives an unbiased estimate of expected prediction error for $\hat{m}$ given the training data. One might call this procedure, *validation*.

Cross-validation is based on using the validation idea repeatedly with the same data. For example, k-fold cross-validation randomly divides the data into k subsets of roughly equal size. First identify one subset as the test data and combine the other $k-1$ subsets into the training data. Estimate the best predictor from the training data and then use that estimate with the test data to estimate the expected prediction error. So far, this is just validation and the estimate of the expected prediction error should be conditionally unbiased.

However, in k-fold cross-validation there are k possible choices for the test data, so one goes through all k validation processes and averages the k estimates of the expected prediction error to give an overall estimate of the expected prediction error. With n data points, the largest choice for k is n, which is known as *leave one out cross-validation*.

Let's look at how all of this works in the most tractable case, linear models with squared prediction error loss. In linear models and more generally in nonparametric regression the model is typically taken as

$$y_i = m(x_i) + \varepsilon_i, \quad \mathrm{E}(y_i) = 0, \quad \mathrm{Var}(\varepsilon) = \sigma^2$$

with independent ε_is, or alternatively

$$y_i|X \quad \text{indep.} \quad \mathrm{E}(y_i|X) = m(x_i), \quad \mathrm{Var}(y_i|X) = \sigma^2.$$

This model will not work for $y \sim \text{Bern}(p)$ because the constant variance condition cannot hold except in degenerate cases. Under squared error loss

$$\text{E}_{y,x}\{L[y,m(x)]\} = \text{E}_x\text{E}_{y|x}\{L[y,m(x)]\} = \text{E}_x\sigma^2 = \sigma^2.$$

In a linear model

$$m(x) = x'\beta$$

.

It is not hard to see that (1) leads to

$$\frac{1}{n}\sum_{i=1}^{n} L[y_i, m(x_i)] = \frac{1}{n}\sum_{i=1}^{n}[y_i - x_i'\beta]^2$$

which is an unbiased estimate of σ^2. However, with least squares estimation and $\hat{m}(x) = x'\hat{\beta}$,

$$\text{E}_{Y|X}\left\{\frac{1}{n}\sum_{i=1}^{n}[y_i - x_i'\hat{\beta}]^2\right\} = \frac{n - r(X)}{n}\sigma^2,$$

which underestimates σ^2. Of course, what we really do in linear models is use the mean squared error, i.e.,

$$\text{E}_{Y|X}\left\{\frac{1}{n - r(X)}\sum_{i=1}^{n}[y_i - x_i'\hat{\beta}]^2\right\} = \sigma^2.$$

Finally, for leave one out cross-validation, the estimate uses the well known Press statistic, see PA Chapter 13. In the following, let $p \equiv r(X)$. With M the perpendicular projection operator onto the model matrix space, I believe

$$\begin{aligned}
\text{E}(Press/n) &= \frac{1}{n}\text{E}\left[Y'(I-M)D^2\left(\frac{1}{(1-m_{ii})}\right)(I-M)Y\right] \\
&= \frac{1}{n}\text{tr}\left[D^2\left(\frac{1}{(1-m_{ii})}\right)(I-M)\sigma^2 I(I-M)\right] \\
&= \frac{\sigma^2}{n}\text{tr}\left[D\left(\frac{1}{(1-m_{ii})}\right)(I-M)D\left(\frac{1}{(1-m_{ii})}\right)\right] \\
&= \frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{1}{(1-m_{ii})}.
\end{aligned}$$

For a one sample problem (intercept only model),

$$\text{E}(Press/n) = \frac{n}{n-1}\sigma^2$$

which is biased up. In fact, this is a lower bound for the expected value among models that include an intercept. Moreover, for $x = 0, 0.5, 1, 10, 19, 19.5, 20$ and fitting a cubic polynomial, I believe $\text{E}(Press/n) > 5\sigma^2$.

In fact, since

$$\sum_{i=1}^{n} \frac{(1-m_{ii})}{n} = \frac{n-p}{n}$$

Jensen's Inequality gives

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{(1-m_{ii})} \geq \frac{n}{n-p},$$

so it looks like Leave One Out CV is multiplicatively more biased upward than the naive estimator is biased downward.

I remember from talking to Rick Picard about his thesis years ago that he claimed Press really sucked. I wonder if this is why he said that.

Other Loss Functions

Using the formal equivalence between prediction and Bayesian decision theory, we can draw conclusions about other loss functions. For example, if for a positive weighting function $w(\cdot)$,

$$L[y, f(x)] = w(y)[y - f(x)]^2,$$

the BP is

$$\frac{\mathrm{E}[yw(y)|x]}{\mathrm{E}[w(y)|x]}.$$

If

$$L[y, f(x)] = |y - f(x)|,$$

the BP is

$$\mathrm{med}(y|x).$$

Moreover, if we use the absolute loss function with y Bernoulli, we get the same result as using Hamming loss, i.e., the BP is

$$\delta(x) = \begin{cases} 1 & \text{if } p(x) > .5 \\ 0 & \text{if } p(x) < .5. \end{cases}$$

Chapter 11
Notes on weak conditionality principle

We have two potential experiments to collect data y and learn about a parameter θ. Roughly, **the weak conditionality principle** says that if you flip a coin to decide to perform experiment $E = 1$ or $E = 2$ then the analysis should be conditional on the experiment you actually performed. What is unquestionably stupid would be to ignore which experiment was actually performed when you know it. But it is less clear that *conditioning* inferences on the observed experiment is actually necessary to get good results as opposed to using the joint distribution of the data and the experiment. (Inferences based on the marginal distribution of the data that ignore knowing the experiment are dumb.) Of course all these distributions are conditional on the parameter.

Note that the weak conditionality principle is a consequence of the ancillarity principle, since the outcome of the Experiment randomization is an ancillary statistic and should be conditioned on.

Fletch has an example

	E=1			E=2		
$f(y\|\theta, E=i)$	$y=0$	$y=1$	$y=2$	$y=0$	$y=1$	$y=2$
$\theta=0$	0.90	0.05	0.05	0.90	0.05	0.05
$\theta=1$	0.10	0.43	0.47	0.01	0.49	0.50
$\frac{f(y\|\theta=1,E=i)}{f(y\|\theta=0,E=i)}$	1/9	8.6	9.4	1/90	9.8	10

alternatively

	E=1			E=2		
$f(y, E=i\|\theta)$	$y=0$	$y=1$	$y=2$	$y=0$	$y=1$	$y=2$
$\theta=0$	0.45	0.025	0.025	0.45	0.025	0.025
$\theta=1$	0.05	0.215	0.235	0.005	0.245	0.25
$\frac{f(y,E=i\|\theta=1)}{f(y,E=i\|\theta=0)}$	1/9	8.6	9.4	1/90	9.8	10

From the second table, the unconditional MP $\alpha = .05$ test of $H_0 : \theta = 0$ versus $H_1 : \theta = 1$ rejects for $E = 2, y = 2$ or $E = 2, y = 1$. From the first table, the two MP $\alpha = .05$ tests of $H_0 : \theta = 0$ versus $H_1 : \theta = 1$, conditional on E, reject for $E = 1, y = 2$ and $E = 2, y = 2$, respectively. They are different results, but I'm not at all sure that

the conditional tests are better. As is the problem with so many N-P tests, the real problem is picking a stupid α level. In this case, by paying the small price of going from $\alpha = 0.05$ to $\alpha = 0.10$ you almost double the power.

I am not saying what procedure is better, only that it is not clear that one dominates the other. And I am all in favor of Bayes over NP. These are all simple versus simple tests, so the class of Bayes rules agrees with the class of NP rules. But, to me, Bayes is clearly a better way of choosing a test than arbitrarily picking a small level of α.

More generally, for two experiments $E = 1$ and $E = 2$ with $\Pr(E = 1) = p$ but getting to observe E, and with outcomes y from the experiments determined by $f(y|\theta, E = i)$, then the weak conditionality principal says that the analysis should be based on $f(y|\theta, E = i)$. The alternative to conditioning would be to base the analysis on the joint distribution of y, E,

$$f(y, E|\theta) = pf(y|\theta, E = 1)I(E = 1) + (1 - p)f(y|\theta, E = 2)I(E = 2)$$

which seems to lead to pretty reasonable results. However, if we look at the likelihood function

$$L(\theta|y, E) = f(y, E|\theta) \propto f(y|\theta, E)$$

so anything based on the likelihood is using the conditional distributions.

Another thing I looked at that seems to give good Fisherian inference from the joint distribution is a 50/50 mixture of

$$E = 1 : y \sim N(0, 1), \quad E = 2 : y \sim N(3, 1).$$

I come up with P values that agree with the conditional P values for $E = 1, y = 3$ and $E = 2, y = 3$, the first being small and the second being 1. Note that data points of equal weirdness to $E = 1, y = 3$ are $E = 1, y = -3$ $E = 2, y = 0$, and $E = 2, y = 6$. Similarly, the only point as weird as $E = 2, y = 3$ is $E = 1, y = 0$. I think, but did not check out, that things work reasonably for testing the alternative $H_1 : y \sim N(4, 1)$.

Chapter 12
Reviews of Two Inference Books

This chapter contains reviews of two excellent books on statistical inference. Both reviews were originally published in *JASA*: Christensen (2008, 2014). The first is a brief text on statistical inference by David Cox. The second is by Erich Lehmann on the historical development of statistical inference.

First my stories about D.R. Cox. I never met David Cox in the flesh but I saw him, and I talked to him on the phone, and I emailed with him. I talked to him on the phone in the mid 1980s when he edited *Biometrika* and I was an Assistant Professor at Montana State University. I had submitted a paper to *Biometrika* along with my friend and colleague Wes Johnson. David was famous for his fast editorial turnaround but our paper had been sitting there a long time. Wes was coming up for tenure at UC, Davis and he needed to hear about the paper. Calling England was not something you did lightly in the 1980s. I got up at 3:30AM and made the call. At first David was not very pleasant. He thought I was pressuring him to accept the paper. When I finally communicated that we just wanted an answer, he was sorry that it had taken so long, put down the phone, and went to another room to check on the paper. A few seconds later he came back on the phone to explain that he could not check right now because he had locked himself in his office and had to call his secretary to come let him out.

The paper did get published in *Biometrika*.

My next personal encounters with David were more than 20 years later, after I wrote a very detailed review of his wonderful little book, *Principles of Statistical Inference*. Out of the blue one day I got an email from him thanking me for the review. We chatted about a few things and I eventually told him that I had been telling the previous story for years. David's response was that, although the story as related could not actually be true, I should absolutely keep telling the story because it was too good not to tell.

12.1 "Principals of Statistical Inference" by D.R. Cox

I must admit that I write this review wondering why anyone would care what I have to say about a new book on statistical inference by D.R. Cox. Cox is, after all, arguably our greatest living statistician. He is the author of numerous books, one of which, *Planning of Experiments*, I consider to be one of the two best statistics books that I have ever read (the other being the small book by Shewhart (1939) edited by Deming). Interestingly, at about the same time Cox published this book on statistical inference, he also published a review article on applied statistics, Cox (2007), the first article ever published in a new IMS journal on the subject. But I am nothing if not opinionated, so I will persevere in my task. I should perhaps also add that, as with any review, this is not about what Cox said, but about what I thought he said.

This is a book on foundational issues in statistical inference. The mathematical level is aimed at university undergraduates in quantitative fields. In chapter one he states, "The object of the study of a theory of statistical inference is to provide a set of ideas that deal systematically with the above relatively simple situations and, more importantly still, enable us to deal with new models that arise in new applications." The book has nine chapters and two appendices. The nine chapters are: Preliminaries, Some concepts and simple applications, Significance tests, More complicated situations, Interpretations of uncertainty, Asymptotic theory, Further aspects of maximum likelihood, Additional objectives, Randomization-based analysis. At the ends of the chapters are Notes, which I think contain some of the most interesting material. In some ways, the chapter Notes are crucial. For example, the book contains interesting material on such things as linear rank statistics, Feiller's method, and ratio estimation, but the uninitiated would have no chance of relating those names to the material without the chapter Notes.

I'll say right up front that I think everyone should read the appendices. The first is "A brief history" of statistics and the second contains Cox's personal view on matters of inference. Although I do not agree with all his views, they are certainly worth encountering. Cox considers (p. 195) the main differences between Fisher and Neyman to be the nature of probability and the role of conditioning. Personally, I think the most important differences between them are the role of repeated sampling and the nature of testing. Neyman-Pearson theory seems to provide decision rules between a null and an alternative. I will discuss Fisherian testing later, but it is certainly a different approach. Neyman embraced the concept of repeated sampling, that is, the long run frequency (LRF) justification for statistical procedures, but my reading of Fisher is that he rejected LRF as a basis for testing. I think that in testing problems Fisher simply used the (null) probability model as a criterion to evaluate how weird the observed data were. Of course, that is intimately related to the different views that Fisher and Neyman had on the nature of probability. But as Cox says in one of my favorite sentences from the book, "Fisher had little sympathy for what he regarded as the pedanticism of precise mathematical formulation and, only partly for that reason, his papers are not always easy to understand." Amen!

Cox is clearly not a Bayesian. I am. He raises the issue of whether probability has the same meaning regardless of whether it is prior or posterior probability. While I

am sure I am missing the philosophical subtleties, as a practical matter it seems like the posterior does one of three things. In the best of cases, we obtain more specific knowledge from the posterior. (Reduced entropy?). If that is not happening, it suggests that we didn't know what we were talking about in the first place. A third scenario is where the data are inadequate to inform us about the parameterization, i.e., we have nonidentifiability. A simple example of this is diagnostic testing where there are three parameters: sensitivity, specificity, and prevalence, but the data are simply the number of individuals who test positive. The data only provide information on the apparent prevalence which is a function of the three underlying parameters. In this case, the conditional distribution of the three parameters given the apparent prevalence will be identical in the prior and the posterior.

Cox (p. 199) states that "Issues of model criticism, especially a search for ill-specified departures from the initial model, are somewhat less easily addressed within the Bayesian formulation." I think he is absolutely correct, but I see little reason to attempt such a search "within the Bayesian formulation." I see little reason not to use Fisherian (as distinct from Neyman-Pearson) tests to critique Bayesian models. In fact, I think it is essential to do so! I also think Cox is right to reject the axioms of personalistic probability as "being so compelling that methods not explicitly or implicitly using that approach are to be rejected." Bayesian statistics may be the only logically consistent form of inference, but it is not the only useful form of inference. Moreover, I think that Bayesian statistics is a wonderful medium for arriving at a consensus of thought. And I suggest that to believe we are doing more than arriving at a consensus is deluding ourselves.

Chapter 1 provides examples, notation, and overviews of what is to come. Throughout the book Cox uses a parameterization $\theta = (\psi, \lambda)$ where ψ is the parameter of interest, typically a scalar, and λ comprises the nuisance parameters. The book focuses primarily on confidence intervals for ψ and testing whether the data are consistent with a particular value ψ_0. To lesser extents, it examines prediction, decision theory, and formal model criticism. The book treats both frequentist and Bayesian approaches to inference, but to my mind, with much more emphasis on the frequentist. On page 8 Cox seems to restrict frequentists to those who accept LRF interpretations of procedures, so I am not sure where Fisher will fit into all of this (or me either since the two inferential procedures that make sense to me are Bayesian for fully specified models and a version of Fisherian testing for less specified models). Finally, chapter 1 contains a sentence I dearly wish I had written, "We follow the very convenient, although deplorable, practice of using the term *density* both for continuous random variables and for the probability functions of discrete random variables."

Chapter 2 starts with standard discussions of likelihood, sufficiency, exponential families and their conjugate priors. To me, it gets really interesting when sections 2.5 and 2.6 introduce machinery for constructing confidence intervals. I once accused the long-run frequency interpretation of confidence intervals of making Bayesian omelets without breaking Bayesian eggs. In sections 2.5 and 2.6 it seems like Cox is providing us with fiducial omelets (intervals) without breaking fiducial eggs. I do not pretend to be an expert on fiducial probability, but it looks to me like he has

set up the problems in these sections so that the fiducial argument is there for the making, yet Cox falls back on the LRF interpretation for the intervals he obtains. In the beginning of Chapter 3, Cox even says of his procedures, "This is close to but not the same as specifying a probability distribution for ψ; it avoids having to treat ψ as a random variable..." It took me a while to find an intellectual underpinning for Cox's arguments here because he does not seem to be using an approach that he endorses, that of reporting parameter values not rejected by a test. Cox explicitly eschews the Neyman-Pearson approach, yet here he is also not using any argument that I recognized as defining Fisherian tests. Moreover, Cox endorses the idea of tossing out probabilities of $\alpha/2$ on both ends of the intervals. That makes good sense to me from a fiducial viewpoint but I do not see any basis for it from inverting either Neyman-Pearson (N-P) or Fisherian tests.

Cox begins chapter 3 on significance tests with a list of six different situations in which one might wish to test a null hypothesis $H_0 : \psi = \psi_0$. Only the first and last of these seem to me appropriate for Fisherian testing. The others lend themselves to N-P testing. But Cox takes a broader view of significance testing than I do. To me the key idea in determining a p value is that the distribution of the test statistic under the (null) hypothesis is used to determine how weird an observed value is. Cox is prepared to use an alternative hypothesis to determine p values.

I spent a lot of time deciphering chapter 3, probably because I have very strong prior opinions about the subject matter and it was difficult to reconcile them with Cox's prose. Cox is clearly not addressing Neyman-Pearson theory, so I believe his goal is to construct p values that are measures of consistency with the null hypothesis. He considers two approaches. First, there are significance tests in which only a null hypothesis and a test statistic are specified (but this also requires a known distribution for the test statistic under the null). In the second, significance tests are considered when alternatives are specified. Initially, I thought that most of this chapter was about the second approach, but gradually I came to think that he is addressing the first approach in a way that is difficult for me to digest.

As I understand the first approach, the one I have been calling Fisherian, significance testing is a variation on proof by contradiction. The null hypothesis is assumed to be correct. If the p value, a measure of consistency with the null hypothesis, is small then the data are inconsistent with the model that incorporates the null hypothesis. This suggests that something is wrong with the null model. But what is wrong need not be the hypothesized parameter value! The p value is the probability of seeing a value of the test statistic that is as weird or weirder than we actually saw. Since there is no alternative, the null distribution has to determine how weird a particular observed value is. Weird values are those with a small probability (density) of occurring, thus the null density provides the ordering of how weird the observable values are. For example, in testing that the mean μ of a Poisson distributed random variable Y is 2, Table 1, column 3 gives the p values for $y = 0, \ldots, 6$. Column 2 gives the probabilities under the null, thus providing our measure of weirdness. For values $y > 6$, the values get increasingly weirder so the pattern is obvious without listing them.

y	$\Pr[Y=y]$	p	$\tilde{p}=\Pr[Y\geq y]$	$\tilde{p}=\Pr[Y\leq y]$
0	0.135335	0.27821	1.00000	0.135335
1	0.270671	1.00000	0.86466	0.406006
2	0.270671	1.00000	0.59399	0.676676
3	0.180447	0.45866	0.32332	0.857123
4	0.090224	0.14288	0.14288	0.947347
5	0.036089	0.05265	0.05265	0.983436
6	0.012030	0.01656	0.01656	0.995466

Table 12.1 Values of p and $\tilde{p}$ for testing a Poisson with mean 2.

Although this seems like a logically sound way to proceed, there are several difficulties with it. The biggest problem is in choosing a test statistic with a known distribution. Fisher (1956, p.50) suggests that reasonable alternatives should inform the choice of test statistic. However, there is a potential lack of coherence in that, for example, a t test and a $t^2 = F$ test are fundamentally equivalent but provide different orderings of how weird observed values are. **[I cannot think why I said this unless I wrote it before I realized the shape of $F(1,df)$ and $F(2,df)$ distributions.]** (Although I suspect that the difference occurs because they provide different continuous approximations to our inherently discrete world.) This testing procedure, while having a sound logical basis, does not address all the issues one might like. Finally, I do not know of anybody who has consistently held to this procedure. Almost nobody applies this theory to χ^2 and F tests, although I suspect that is merely a matter of computational convenience. To "correctly" evaluate the p value in those cases, you need to find a second value of the statistic that gives the same density as your observed value and then compute the probability of being in either tail. These days, that would not be hard to program, but even today, it is not a computation that is commonly performed. Fisher (1925, Sec. 20) insisted that extremely large p values are as significant as extremely small ones. I view this as simply a convenient alternative to taking the trouble to make a correct p value computation. Box (1980) used this definition of p value for Bayesian model checking. This definition is also widely accepted in performing exact conditional tests on discrete data, e.g. Mehta and Patel (1983). Nonetheless, in Fisher's discussions of his exact test for 2×2 tables, he seems to have lent towards p values computed directionally. But again, that might have been for computational convenience.

I am not at all sure that Cox would agree with my description of the first approach because, what I consider a computational convenience, Cox seems to incorporate into the basic procedure. Even without explicitly defining alternatives, Cox p.32 indicates that the p value, here called a $\tilde{p}$ value to distinguish it from the other definition, is the probability of seeing data as indicative or more indicative of "a departure from the null hypothesis *of subject matter interest* [my italics]." For example, in testing that the mean μ of a Poisson distributed random variable Y is 2, Cox uses the idea that if $\mu > 2$, then only large values of Y are useful for detecting departures from the null, and conversely when $\mu < 2$. Cox provides a table similar to Table 1 of one sided $\tilde{p}$ values. Column 4 provides $\tilde{p}$ values when larger values of the test statistic are of interest or when the alternative mean is greater than 2 while

column 5 provides $\tilde{p}$ values when smaller values are of interest or the mean is less than 2. These are distinct from the p values computed by letting the null distribution determine which observations are most unusual.

I am confused about how these ideas make the transition into Cox's second approach, testing the null value of a parameter against some alternative value of the parameter, because I am not quite sure if these are merely examples of $\tilde{p}$ values with different departures of subject matter interest or if they are $\tilde{p}$ values based on the specification of alternative hypotheses. In this case, they seem to amount to the same thing. When alternatives are available, I suspect $\tilde{p}$ needs to be viewed as a measure of consistency with the null model *relative to the alternative*, in which case it could perhaps be formalized as the probability of seeing a likelihood ratio as extreme or more extreme than the one you actually saw. In fact, Cox presents this version of a $\tilde{p}$ value when discussing classification problems in section 5.17. Without a specific alternative, I am not sure how one could formalize $\tilde{p}$ values beyond what Cox has done. However, I am not convinced that this definition of $\tilde{p}$ can be reconciled with the idea of $\tilde{p}$ being a measure of consistency with the null hypothesis. Cox rightly points out that extremely small values of F statistics in linear models have large $\tilde{p}$ values under this paradigm but suggest inconsistency with the null model.

Defining a $\tilde{p}$ value relative to some formal or informal alternative is to abandon completely the idea of a proof by contradiction. A small $\tilde{p}$ does not assure us that the data contradict the null hypothesis nor does a large $\tilde{p}$ value assure us that the data are consistent with the null hypothesis. The following simple example using two discrete densities illustrates the point by improving on a similar example in Christensen (2005). The null is $f(x) = .01, x = 1, \ldots, 95, f(0) = .001, f(96) = .049, f(97) = 0$. Relative to the null, observing 0 is weird and has a p value of .001; observing 96 is consistent with the null and has a p value of one. If you have no alternative, this is extremely reasonable. Now consider a formal alternative $g(x) = .001, x = 1, \ldots, 95, g(0) = 0, g(96) = .1, g(97) = .805$ or just an informal alternative of rejecting for large values of x. Relative to the alternative the $\tilde{p}$ value for observing 0 is one and the $\tilde{p}$ value for observing 96 is .049. The point is that observing a 96 in no way contradicts the null hypothesis, yet $\tilde{p}$ is small. On the other hand, observing 0 contradicts the null, but $\tilde{p}$ is large. Since $\tilde{p}$ does not provide the basis for a proof by contradiction: 1) with no clearly specified alternative, I do not know what $\tilde{p}$ is giving us and 2) with a formal alternative available, it seems to me that there is no reason to focus on the null hypothesis, as opposed to the alternative, and we must be employing a procedure for deciding (or evaluating the evidence) between the null and alternative. This is the proper domain of Neyman-Pearson testing and Bayesian testing. Moreover, if Neyman-Pearson testing is a decision procedure and not a proof by contradiction, I see no reason to restrict one's self to tests in which the probability of Type I error is small. In a decision procedure, one should play off the probabilities of both Type I and Type II error so that both are reasonable.

Cox mentions in chapter 3, and returns in Section 6.5, to the idea of comparing two (nonnested) models by using each in turn as the null hypothesis. If p values are defined only by the null hypothesis, this seems like a very sensible approach.

The conclusion would be that the data are consistent with both, one, or neither of the models. With a formal alternative, this still seems like a sensible approach but one in which the conclusions need more explanation. As mentioned before, in a Neyman-Pearson test one should not merely choose a test with a small probability of Type I error but choose something with reasonable levels for both the Type I and Type II errors. With $\tilde{p}$ defined as the null probability of seeing the observed likelihood ratio or a more extreme one, $\tilde{p}$ provides a data driven choice for a level of Type I error. If we then compute the probability under the alternative of a more extreme likelihood ratio than that observed, we have the corresponding data driven choice for the level of Type II error. For continuous data, this is the $\tilde{p}$ value treating the alternative as the null. Here, a large $\tilde{p}$ value does not imply consistency with the null, it merely suggests that the null looks reasonable relative to the alternative. When doing both tests, one large $\tilde{p}$ value and one small $\tilde{p}$ value make the choice of distribution (weight of evidence determination) simple. Two large $\tilde{p}$ values suggest that it is difficult to distinguish between the alternatives with these data. Two small $\tilde{p}$ values suggest that the *pair* of alternatives is wrong, but I suspect that this does not *necessarily* imply that either one is wrong. Of course, composite hypotheses can make this procedure intractable, a problem that Cox gets around nicely in section 6.5 with large samples; see also page 128 and section 7.4.

Cox also discusses inverting tests to give confidence intervals. It is not clear to me how this relates to sections 2.5 and 2.6. Fisher's (1956) approach seems to have been that if his model passes the goodness of fit testing stage, even though that does not imply the model is correct, he goes on to estimation. Estimation takes the form of fiducial inference rather than confidence intervals obtained by inverting tests. When you develop tests without consideration of alternatives, it seems a bit shaky to then discuss the collection of all parameter values that would not be rejected by a test. Still, I am more comfortable with that than with the LRF interpretation of confidence intervals or fiducial intervals. Of course, what I am really comfortable with in this situation are posterior intervals.

Cox emphasizes a couple of interesting points relative to the p value approach to significance testing. Often the null hypothesis is actually a family of distributions for which a sufficient statistic may exist. By definition, the distribution of the data given the sufficient statistic does not depend on the unknown parameters of the model, so this conditional distribution can always be used as the basis for a test regardless of the choice of test statistic. He also points out that given just the test statistic and its null distribution, it is easy to construct an exponential family of (Neyman smooth) alternatives.

Chapter 4 goes into more complicated situations. Subsection 4.4.1 deals with conditioning on part of the data. Much of the discussion focuses on categorical data. Alas, we Americans only share second hand the cultural legacy of such luminaries as Shakespeare and Shaw, but as a great American sage might have put it: I pity the poor fool who reads this subsection without some previous knowledge of the models. Fortunately, I have such knowledge, so I found it very interesting. Cox's first example considers differences between Poisson, multinomial, and product-multinomial sampling. He emphasizes that, "Due care in interpretation is

of course crucial." Not only do I agree, but this is one reason I like to call generalized linear models a marvelous computing device in search of the theory. The last example in this subsection I found very disturbing. It is a case wherein the conditional procedure that has worked marvelously up to this point, breaks down. Cox takes this in stride indicating that insistence on the conditional approach may sometimes require paying too high a price. But I like my theories to stick together a little better than this. Perhaps that is why I am a Bayesian. In fact, as I was reading Example 4.1 in section 4.3 about conditional and unconditional inference for a random sample from the uniform $(\theta-1,\theta+1)$, I could not help but think that the issues would not be a problem if you were a Bayesian.

Chapter 5 is on interpretations of uncertainty and its beginning looked like it was headed way over my head. "There are two ways in which probability may be used in statistical discussions. The first is phenomenological, to describe in mathematical form the empirical regularities that characterize systems containing haphazard variation." So this first notion describes variability as opposed to the second role which "is in connection with uncertainty and is thus epistemological." I think the main question here is whether the philosophical nature of probability as it is used to describe random variation in observable data is the same as, or more to the point whether it can be combined with, the philosophical nature of probability used to describe our personal uncertainty about a parameter. The easy answer for a Bayesian is, "Yes." What we do not know about random data is the same stuff that we do not know about parameters. In fact, I have vague memories that Frank Lad may have made this more than an easy answer, based on the "God does not play dice." world view. In Example 5.3, Cox attempts to explain Fisher's notion of probability. My reaction was that it was so fraught with assumptions that I was glad not to be a frequentist. In section 5.2 some examples of the strong and weak likelihood principles would have been nice.

In section 5.3 Cox mentions Fisher's fiducial approach to inference but, and I am by no means sure of this, I think he is just dismissing it out of hand. Specifically, Cox considers $\bar{Y}\sim N(\mu,\sigma_0^2/n)$ so that $1-c=\Pr[\mu<\bar{Y}+k_c^*\sigma_0/\sqrt{n}]$. Fisher would argue that this defines a fiducial probability distribution on μ. Cox says that such a single set of limits "can in some respects be considered just like a probability statement for μ" but also that they cannot be combined or manipulated like probabilities, specifically that it is "clearly illegitimate" to give the probability that μ exceeds zero. I could be wrong, but it seems to me that Fisher would not have any problem discussing the probability that μ exceeds zero. Moreover, if you cannot do something as simple as give a probability for μ exceeding zero, in what sense can you treat these as probability statements about μ? Perhaps Cox is using this idea to justify a later statement that μ is more likely to be in the center of a [two-sided] confidence interval, and that if the model is appropriate and if μ "is outside the interval it is not likely to be far outside." It seems clear to me that nearly everyone, Cox and Fisher included, want Bayesian answers to problems, it is just that some people are not willing the bite the Bayesian bullet to get them. Cox also uses the Exchange Paradox to illustrate "in very simple form" the problems of switching probability statements about observables to probability statements about parameters. Alas, I know to my

personal sorrow that there is nothing simple about the exchange paradox except its statement, see Blachman, Christensen, and Utts (1996).

In Section 5.6 Cox returns to the subject of how to give a frequentist meaning to $\tilde{p}$ values. Back in Section 3.2, I thought he took the position that it should be viewed as a long run frequency related to deciding to reject the null hypothesis when it is true. The last thing he says in section 5.6 seems to suggest he is adopting the Fisherian view that it simply is what it is. The null model has probabilities associated with it, so there is probability associated with the p value computation. I would not want to have to bet on which interpretation Cox means to adopt.

Oops. I, a Bayesian, just said I didn't want to bet on something. Section 5.7 describes the basis for Bayesian personal probabilities in terms of making bets. It is quite a nice discussion given that it is just over a page long on a subject often treated at book length.

In section 5.8 Cox presents his Example 5.5 as showing that priors that work well in small dimensions can go astray in multiple dimensions. I think the example merely shows that infinite flat priors are stupid and, as with the Lindley-Jeffrey's paradox, we should not be surprised when they give stupid results. In reduced form, the example gives the conditional distribution of Y given θ so that $\mathrm{E}(Y|\theta) = \theta + 1$. The improper prior on θ causes $\mathrm{E}(\theta|Y) = Y + 1$. Computing the unconditional expectations, we get both $\mathrm{E}(Y) = \mathrm{E}(\theta) + 1$ and $\mathrm{E}(\theta) = \mathrm{E}(Y) + 1$. I do not see where multiple dimensions come into it. Moreover, I think the burden of proof should always be on showing that improper priors give reasonable answers. It should never be viewed as surprising when they do not.

Example 5.6 is to illustrate that "a prior that gives results that are reasonable from various viewpoints for a single parameter will have unappealing features if applied independently to many parameters." What he sees as a problem with the prior distribution, I see as a problem with the data and to some extent with biased estimation. Let me present an example similar to his: heteroscedastic one-way ANOVA. For $i = 1,\ldots,n,\ j = 1,\ldots,m$ let $y_{ij} = \mu_i + \varepsilon_{ij}$ with the ε_{ij}s independent $N(0,\sigma_i^2)$. Let s_i^2 be the sample variance from the ith group. To introduce some bias, we look at $[(m-1)/(m+1)]s_i^2$, which has better expected squared error properties than s_i^2. It seems like Cox's dissatisfaction in Example 5.6 should extend to the fact that as n gets large $\prod_{i=1}^n [(m-1)/(m+1)]s_i^2 / \prod_{i=1}^n \sigma_i^2$ does not become a good estimator of the number 1. In fact, it is an unbiased estimate of $[(m-1)/(m+1)]^n$, which approaches 0. Even without introducing the bias, $\prod_{i=1}^n s_i^2 / \prod_{i=1}^n \sigma_i^2$ is still not a really good estimate of the number 1. Introducing the bias turns a mediocre estimate into a bad one. Ultimately, although we are in an asymptotic framework, there are not enough data on any one parameter to expect good asymptotic behavior. Perhaps the quote from the beginning of this paragraph should be changed to: a procedure that gives results that are reasonable from various viewpoints for a single parameter may have unappealing features if applied to many parameters. In Section 8.3 Cox says something quite similar about biased point estimation while agreeing that a little bit of bias is not normally a bad thing.

Section 5.9 deals with reference priors but more specifically with the virtues and difficulties of developing reference priors through maximizing entropy. While

I think this is an interesting and useful theory, I think the most important role for reference priors is simply as priors we agree to use so that everyone has a common basis for comparison.

Section 5.11 discusses an approach to eliciting prior probabililities that should interest those of us who are not enamored with having betting as a key aspect to the foundations of Bayesianism. Cox's note on this section is quite amusing. Personally, rather than buying into coherent betting, the first thoughtful reason I had for being a Bayesian was based on a result in section 8.2 on Decision analysis, that is, that all reasonable procedures are Bayes procedures. If I have to act like a Bayesian anyway, why not use a prior that I think is reasonable. But to be honest, I had been indoctrinated before I ever heard of the complete class theorem.

Frankly, I was a little offended that in Section 5.12 on four ways to implement Bayesian procedures that the first two were not Bayesian. They are empirical Bayesian, which is a frequentist approach. In discussing the use of informative priors, Cox seems to want a consensus on what the prior distribution should be. Personally, I am far more interested in getting a consensus on the posterior distribution. It seems clear to me that if we cannot arrive at a practical consensus on the posterior distribution, we collectively do not yet know what is going on. I think the fact that reasonable people can obtain substantially different posteriors provides valuable information to the scientific community on our state of knowledge.

Chapter 6 treats asymptotic theory. I must say that Cox displays a facility with $O(\cdot)$, $o(\cdot)$, $O_p(\cdot)$, $o_p(\cdot)$ notation that boggles my mind. In fact, he displays amazing facility with the entire subject, although I am personally more comfortable with, say, Ferguson (1996). Figure 6.2 provides a fascinating illustration of the relationships between likelihood ratio, Wald, and score tests. It also quickly leads to Cox's conclusion that if these test statistics are substantially different, the asymptotic formulation is called in question.

Chapter 7 discusses some of the difficulties with asymptotic maximum likelihood theory as well as means for avoiding some of those difficulties. In particular, section 7.6 gives brief discussions of partial-, pseudo-, and quasi- likelihoods.

Chapter 8 is devoted to additional objectives. In section 8.1 Cox manages to treat the entire object of science, that is, prediction, in one page. But I'm being opinionated again. Technically, he presents (good) frequentist prediction in terms of testing equality of the parameters from the new observation and the observed data. That seems unnecessarily complicated to me. If the new observation is y^* and the old data are a vector y and if one can find a known distribution for some function of (y^*, y), then a prediction region consists of all values y^* such that (y^*, y_{obs}) is consistent with the known distribution at a level α. In other words, take y^* so that (y^*, y_{obs}) provides a p value greater than α. Again, values of (y^*, y) with the lowest density under the known distribution are the values that are most inconsistent. This difference in thinking about prediction regions is unlikely to engender much difference in practice. Bad frequentist prediction includes plugging estimates of the parameters into the sampling distribution of the new observation and then proceeding as if the sampling distribution was known. This systematically underestimates variability, a problem Bayesian prediction avoids.

Subsection 8.3.3 seems to have some key typographic errors or some logic I do not follow. It took me a while to figure out that subsection 8.4.2 is a generalization of the Grizzle, Starmer, Koch (1969) approach to categorical data.

Chapter 9 on randomization-based analysis, considers both sampling theory and designed experiments. The main idea is to contrast randomization-based analyses with the model based approach taken in the rest of the book. I think section 9.3 on design would be tough sledding for anyone who had not seen similar material before. One thing I particularly liked was his making a notational distinction between the variance appropriate for a completely randomized design versus that for a randomized complete block (paired) design. Too often we (I) just call them both σ^2.

Two final points about the book that I really like. First, Cox does not blithely assume independence. He repeated points out how crucial independence assumptions may be. Second, he stresses that all real data are discrete and that continuous models are just approximations that can go astray.

I would like to thank Prof. Cox for never having sat down with one of my books and picked it apart as I have done his. I hope he takes this as a sign of my high regard for his professional accomplishments. This is a great book by a great statistician. Buy it and read it.

References

Blachman, Nelson M., Christensen, Ronald and Utts, Jessica M. (1996). Comment on Christensen, R. and Utts, J. (1992), "Bayesian Resolution of the 'Exchange Paradox." *The American Statistician*, 50, 98-99.

Box, George E. P. (1980). Sampling and Bayes' Inference in Scientific Modelling and Robustness. *Journal of the Royal Statistical Society*. Series A (General), 143(4), 383-430.

Christensen, Ronald (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician*, **59**, 121-126.

Cox, D.R. (1958). *Planning of Experiments*. John Wiley and Sons, New York.

Cox, D.R. (2007). Applied Statistics: A Review. *The Annals of Applied Statistics* **1**, 1-17.

Ferguson, Thomas S. (1996). *A Course in Large Sample Theory*. Chapman and Hall, New York

Fisher, Ronald A. (1925). *Statistical Methods for Research Workers*, Fourteenth Edition, 1970. Hafner Press, New York.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*, Third Edition, 1973. Hafner Press, New York.

Grizzle, James E., Starmer, C. Frank, and Koch, Gary G. (1969). Analysis of categorical data by linear models. *Biometrics*, **25**, 489-504.

Mehta, C.R. and Patel, N.R. (1983). A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *Journal of the American Statistical Association*, **78**, 427-434.

Shewhart, W. A. (1939). *Statistical Method from the Viewpoint of Quality Control*. Graduate School of the Department of Agriculture, Washington. Reprint (1986), Dover, New York.

12.2 "Fisher, Neyman, and the Creation of Classical Statistics" by Erich L. Lehmann

Erich Lehmann was a class act and this short book *Fisher, Neyman and the Creation of Classical Statistics* is worthy of him. I found it great fun to read.

Lehmann was Neyman's Ph. D. student and spent most of his career in "Neyman's" department at Berkeley. He was a major contributor to Neyman-Pearson theory and is almost certainly the foremost expositor of that theory with classic books *Testing Statistical Hypotheses*, *Theory of Point Estimation*, and, my personal favorite, *Nonparametrics: Statistical Methods Based on Ranks*. He could be forgiven for being biased (I was on the look out) but for the most part the treatment is even-handed. When the issue of bias arises, I suspect it is less a matter of bias and more a matter of having an imperfect understanding of Fisher. (And who can be blamed for having an imperfect understanding of Fisher?) Having addressed the possible bias of the author, I should perhaps address the biases of the reviewer. While I have the utmost respect for both Fisher and Neyman, nobody would call me a fan of Neyman-Pearson theory and I am at core a Bayesian. In reviewing the book, which is itself a review of others' work, references that are not listed below come from Lehmann's book.

The first chapter starts out with some brief biographical information about the two protagonists as well as bios of supporting characters Karl and Egon Pearson and W.S. Gossett. To me, one highlight of this chapter was that Neyman learned from Karl Pearson that "scientific theories are no more than models of natural phenomena" which, not only do I agree with but, brought to mind Box's quote about all models being wrong. The chapter then spends a few pages on Fisher's classic 1922 paper, "On the mathematical foundations of theoretical statistics". Although Lehmann *seems* to credit Fisher for the invention of maximum likelihood estimates, he was aware of Stigler's (2007) work on their history. I myself remembered, from a previous life, that the "most probable number" used in serial dilution bioassays predated Fisher. "The estimate is 'most probable' only in the roundabout sense that it gives the highest probability to the observed results." This seems to me an admirable description of maximum likelihood estimation for discrete distributions. The previous quotation was taken from Cochran (1950) who later says, "Consequently the m.p.n. (most probable number) method is now generally used in a great variety of problems of statistical estimation, though it more frequently goes by the name of the 'method of maximum likelihood'." Cochran credits McCrady (1915) for originating

the m.p.n. in this application and the m.p.n. terminology lives on to this day. Stigler cites examples by Lagrange and Daniel Bernoulli of finding most probable values in the 18th century.

Like Fisher's 1922 paper, Chapter 2 of Lehmann's book transitions into testing, but now the focus shifts to Fisher's ground breaking (1925) book *Statistical Methods for Research Workers*. It seems that Fisher took quite a bit of grief for writing a practical manual directed at research workers and not including the underlying theory. Lehmann points out that after Fisher carefully derived Gossett's t distribution in 1912 (Fisher, 1915), Gossett urged on Fisher to derive more small sample distributions. Lehmann cites Gossett as pleading to Fisher, "But seriously, I want to know what is the frequency distribution of $r\sigma_x/\sigma_y$ [sic] for small samples, in my work I want that more than the r distribution now happily solved." In a section close to my own heart, Lehmann discusses the issue of testing two independent samples. This is the first time that I wondered if Lehmann was too firmly in Neyman's camp to fully understand Fisher. While Lehmann justifiably calls out Fisher for some technical sloppiness (citing Scheffé's admirable 1959 book), I think the bigger point goes wanting. The issue, of course, is whether to assume equal variances. Fisher's point, and I think it is well taken, is that the appropriate test is typically one of whether the two samples come from the same normal population. This is scientifically distinct from, although probabilistically equivalent to, testing whether two normal populations have the same mean given that they have the same variance. Quoting Lehmann, "Fisher concludes this later discussion by pointing out that one could of course ask the question: 'Might these samples be drawn from different normal populations having the same mean?' ... but that 'the question seems somewhat academic'." As Christensen et al. (2011, p. 123) point out, it is by no means clear that testing the equality of means when the variances are different is a worthwhile thing to do.

Chapter 3 moves on to Neyman-Pearson theory. Two things particularly struck me here. Apparently, the generalized likelihood ratio test statistic predates the theory of optimal testing and Neyman originally wanted to do the theory as a Bayesian. "This long two-part [1928] paper is a great achievement. It introduces the consideration of alternatives, the two kinds of error, and the distinction between simple and composite hypotheses. In addition, of course, it proposes the likelihood ratio test. This test is intuitively appealing, and Neyman and Pearson show that in a number of important cases it leads to very satisfactory solutions. It has become the standard approach to new testing problems." Optimal testing arrives in 1933 as Neyman and Pearson seek to solve a decision problem, "Without hoping to know whether each separate hypothesis is true of false, we may search for rules to govern our behavior with regard to them ..." Lehmann's summary of Neyman and Pearson's innovations follows: "The 1928 and 1933 papers of Neyman and Pearson discussed in the present chapter, exerted enormous influence. The latter initiated the behavioral [decision theoretic] point of view and the associated optimization approach. It brought the Fundamental Lemma and exhibited its central role, and it provided a justification for nearly all the tests that Fisher had proposed on intuitive grounds. On the other hand, the applicability of the Neyman-Pearson optimality theory was

severely limited. It turned out that optimum tests in their sense existed only if the underlying family of distributions was an exponential family (or, in later extensions, a transformation family). For more complex problems the earlier Neyman-Pearson proposal of the likelihood ratio test offered a convenient and plausible solution. It continues even today to be the most commonly used approach." The rub is whether this theory really does provide an appropriate justification for Fisher's tests. Fisher's dissent is the subject of Chapter 4.

It seems to me that Fisher's key objection to Neyman-Pearson testing is the introduction of alternative hypotheses. Ironically, it was correspondence between Gosset and Egon Pearson (discussed at the beginning of Chapter 3) that generated this idea. In a 1934 paper, Fisher very carefully stated, "The test of significance is termed uniformly most powerful *with regard to a class of alternative hypotheses* if this property [i.e. maximum power] holds with respect to all of them." [My italics. I am quoting Lehmann quoting Fisher. Presumably the "i.e." is Lehmann's.] Even at this early point, when Fisher's reaction to Neyman-Pearson theory was tepid and as yet involved no personal animosity, we have the hint that Fisher is not willing to accept this "class of alternative hypotheses" as the only possible alternatives. As I have argued elsewhere (Christensen, 2005), Fisherian testing is essentially subjecting the null hypothesis to a proof by contradiction in which the null is contradicted (rejected) or not contradicted. Unfortunately, the contradictions are almost never absolute and the strength of contradictory evidence is measured by a small P value. No alternatives are needed for a proof by contradiction. As indicated earlier, and as Fisher violently objected to, Neyman-Pearson theory is essentially a decision procedure (cf. p. 54), for which (unlike Neyman and Pearson's original thinking, cf. p. 35) there is no reason why having a small probability of type I error should be important if it leads to large probabilities of type II error, cf. p. 55.

It is in Chapter 4 that Lehmann seems, to me, to misunderstand Fisher most often. On page 48 he (technically correctly) describes an argument by Fisher as increasing the power of a test. Fisher's interest is in decreasing the P value. On page 57 Lehmann writes, "Fisher relied on his intuition, while Neyman strove for logical clarity." The word "while" seems to make this sentence inappropriately convey far more than the sum of its parts (neither of which I could disagree with). In another matter, I admit that Fisher is responsible for the silly dominance of the 0.05 level in testing, but I do not believe that he is to blame for an idea that others took to absurd lengths. On page 53 Lehmann presents 8 examples and states, "These examples, to which many others could be added, show that Fisher considered the purpose of testing to be to established [sic] whether or not the results were significant at the 5% level, and that he was not particularly interested in the p-values per se." I took the examples exactly opposite. Given that Fisher was reporting P values from tables of the t distribution, he seems to report them as accurately as the tables allow. Moreover, Fisher (1936) once pointed out that some of Gregor Mendel's data give P values that are suspiciously too high. On page 55 Lehmann suggests, I think unfairly, that the Neyman-Pearson attitude towards test sizes is more appropriate. In fact, I think it is equally fair to say that Neyman-Pearson are responsible, but not to blame, for their tests being used almost exclusively with small α levels.

Chapter 5 is entitled, "The Design of Experiments and Sample Surveys." While it is hard not to notice similarities in the ideas used in experimental design and sampling, I somehow felt that Dr. Lehmann was a little too cavalier (a word too often applied to me) about their differences. On page 64 Lehmann quotes a passage from Fisher's book *The Design of Experiments* that comes close to demonstrating that Fisher's concept of testing is essentially a proof by contradiction. On the same page I think Lehmann is quite right for chiding Fisher for not seeing the usefulness of power as a tool in determining sample size. Even in Fisher's concept of testing, it is worthwhile to consider the power of various alternatives for a fixed sized test. However, in Fisher's concept, one should take a wider view of what the various alternatives are. (I think Fisher occasionally used fixed sized tests but I am less sure that he would admit to it.) I have addressed the issue that Fisher found alternative hypotheses inappropriate (except to help in choosing a test statistic, cf. Fisher (1956, p. 50)), but on page 65 I found myself thinking about how Fisher and Neyman-Pearson would disagree even on what a null hypothesis was. In Neyman-Pearson theory a null hypothesis is an hypothesis about a parameter value within an underlying statistical model. In Fisherian testing the null "hypothesis" is better thought of as the null *model*. The correspondence is that the Neyman-Pearson model, together with their null hypothesis, is the null model in Fisherian testing.

In discussing randomized block designs, on page 67 Lehmann quotes Fisher as saying, "the discrepancies between the relative performances of different varieties in different blocks ... provide a basis for the estimation of error." I take this as a pretty clear statement that in a randomized complete block the treatment-block interaction is what you *want* to use as a measure of error. As I recently said in another context: If evidence for main effects is not so blatant that it overwhelms any block-treatment interaction we should not declare main effects.

Subsection 5.7.1 on randomization does not mention what I consider to be the most important reason for randomizing treatments. Randomization should (on average) alleviate the effects of any confounder variables, therefore randomization provides a philosophical basis for inferring that the effects we see are *caused* by the treatments.

Alas, page 73 closes with more comments reflecting the author's background. "These papers by Jack Kiefer [on optimal design] complemented and to some extent completed Fisher's work on experimental design as the Neyman-Pearson theory had done for Fisher's testing methodology." "In testing, the Neyman-Pearson theory provided justification for the normal-theory tests that Fisher had proposed on intuitive grounds." It has been pointed out that the t test is not reasonable because it is uniformly most powerful unbiased, the criterion of being uniformly most powerful unbiased may be reasonable because it gives the t test. (I got this from Ed Bedrick, who got it from Robinson (1991), who got it from Dawid (1976).)

Chapter 6 discusses estimation. Fiducial inference is one of the great mysteries of the statistical world. I have never personally met anyone who claimed to understand it. But Lehmann points out a passage from Fisher that I find crucial. In discussing Fisher's 1935 paper on "The foundations of inductive inference" Lehmann says a "new feature is the identification of fiducial limits with the set of parameters θ_0 for

which the hypothesis $\theta = \theta_0$ is accepted at the given level. This interpretation had already been suggested by Neyman in the appendix to his 1934 paper." Personally, I find this a much more reasonable basis for Fisherian interval estimation than fiducial inversion of probability distributions. This lets one state unambiguously that a Fisherian interval contains all the parameter values that are consistent with the data and the statistical model as determined by an α level test. (Remember that a Fisherian test requires a null model that is often a statistical model together with a null hypothesis for a parameter value and that if the test is not rejected at the α level we merely fail to contradict the null model so the data are merely consistent with the null model.) In the next section Lehmann repeats Neyman's important point about the long-run frequency interpretation of confidence intervals that the long run need not be on the same problem, but rather on all the confidence intervals that a statistician performs. (Not that that solves the problem of a confidence interval *really* saying nothing about the data at hand.) Section 6.4 seems to me to suggest that Fisher, like me, finds "confidence" to be nothing more than a backhanded way to get people to think of posterior probability, no matter how much one talks about long-run frequencies. Indeed, it seems that Fisher is identifying confidence with his own concept of fiducial probability. I find it rather comforting that these two concepts that I have never understood could be the same concept. It is fascinating to think of these renown anti-Bayesians trying desperately to make Bayesian omlettes without breaking eggs *a priori*. McGrayne (2011, p. 144) reveals that Abraham Wald, who appears five times in Lehmann's book as a key contributor to classical statistics, was a closet Bayesian. His reticence at coming out must ultimately be due to our protaganists.

The final chapter provides an epilogue that briefly summarizes the contributions of these giants to a variety of topics. In particular, the section "Hypothesis Testing", for good or ill, recapitulates many of the the points highlighted in this review. An appendix lists Fisher's works.

While I have gone to some lengths to point out what I think are biases in the book, let me reemphasize that given Lehmann's background, I think these are remarkably minor. And for all my disagreements, I found the book both fun and informative. Indeed, I found myself almost ashamed for having let Lehmann do all this hard work for me and definitely feel appreciative. If you find the title of Lehmann's book interesting, by all means buy it and read it. My hope is that this review will have whetted your appetite.

References

Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician*, 59, 121–126.

Christensen, R., Johnson, W., Branscum, A. and Hanson, T.E. (2011). *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*, CRC Press, Boca Raton.

Cochran, W.G. (1950). Estimation of bacterial densities by means of the "most probable number". *Biometrics*, 6, 105-116

Dawid, D. (1976). Discussion of the paper of O. Bandorff-Nielsen "Plausibility inference." *Journal of the Royal Statistical Society, Series B*, 38, 123-125.

Fisher, R.A. (1936). Has Mendel's work been rediscovered? *Annals of Science*, 1, 115-137.

Fisher, R.A. (1956), *Statistical Methods and Scientific Inference* (3rd. ed., 1973), Hafner Press, New York.

McCrady, M.H. (1915). The numerical interpretation of fermentation-tube results. *J. Infec. Dis.*, 17, 183-212.

McGrayne, S.B. (2011). *The Theory that Would Not Die: How Bayes Rule Cracked the Enigma Code, Hunted Down Russian Submarines & Emerged Triumphant from Two Centuries of Controversy*, Yale, New Haven.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, 6, 15-51.

Scheffé, H. (1959). *The Analysis of Variance*, John Wiley and Sons, New York.

Stigler, S.M. (2007). The epic story of maximum likelihood. *Statistical Science*, 22, 598-620.

Chapter 13
The Life and Times of Seymour Geisser.

This is a written version of a talk I gave at a 2005 Objective Bayes conference in Branson, MO. On my vita it is listed as published in the 2005 JSM Proceedings. (My talk was followed by one of Rob McCullagh's. I was so miserable giving mine and Rob was so joyously enthusiastic giving his, that it convinced me to stop giving talks except for special occasions.)

Seymour Geisser was a major figure in modern statistics, particularly in the development of Bayesian statistics. He was an active researcher, and able administrator, and an interesting man. I discuss his life, his impact on statistics, and recount some personal interactions.

13.1 Introduction

I am not quite sure how I came to be doing this. There are any number of people more qualified to discuss Seymour Geisser's personal and professional lives than me.

I was a student at the University of Minnesota before Seymour got there. And I am sure there were times when he thought that I would still be a student there when he left. Fortunately for us both, that turned out not to be true.

1. Seymour was not my advisor.
2. I never wrote a paper with him.
3. I never officially took a class from him.
4. I sat in on his prediction class one quarter where my unofficial status did not keep Seymour from making me do a class presentation.

Seymour was fond of observing that there are a lot of smart people in the world but that what matters is what you do with it. In graduate school, I may have been his poster boy for what to avoid.

After I finally left Minnesota for a position at Montana State University, I was told that my stock rose immeasurably when I published a little *American Statistician* article on "Bayesian point estimation using the predictive distribution," (Christensen and Huffman, 1985). Seymour was pretty sure that I would go to Montana and never be heard from again.

13.2 I Started Out as A Child

Seymour Geisser was born on October 5, 1929 in the Bronx, New York. He was the elder of two sons born to Polish immigrants who worked in the garment industry. His father left Poland after being discharged from the army having served in the Russo-Polish war of 1920.

At the age of two he moved to Brooklyn. By the age of 12, he was already widely recognized for being very clever (Zelen, 1996). He graduated from Lafayette High School. Seymour enjoyed high school and played some point guard with the basketball team. Seymour matriculated from the City College of New York in 1950 having spent much of his undergraduate years sleeping on the subway between City College and his home in Bensonhurst. His major was mathematics, in part because the math program was housed near the cafeteria where he played chess.

13.3 North Carolina

After graduation from City College, Seymour faced the same question many of us do and arrived at the same answer. The question is, "How am I going to support myself?" and the answer is, "There seem to be jobs in statistics." In Seymour's case, this came about through the intervention (or intercession) of Seymour's cousin Leon Gilford and his wife Dorothy Gilford. Both were statisticians. Dorothy had been a student of Harold Hotelling at Columbia. By then, Hotelling had moved to the University of North Carolina (UNC), so Seymour headed south.

At UNC Seymour hobnobbed with the likes of Sudish Ghurye, Ingram Olkin, Ram Gnanadesikan, Shanti Gupta, Marvin Kastenbaum, Marvin Zelen, and others. They drank beer, played cards, gambled, and did good statistics. To see how times have progressed, in graduate school we drank beer and played volleyball, softball, and basketball. But we had people like Bill Sudderth to teach us the evils, or at least the futility, of gambling.

The faculty at UNC included Hotelling, Gertrude Cox, Wassily Hoeffding, S.N. Roy, George Nicholson, R.C. Bose, and Herb Robbins. The students at UNC were apparently as intimidated by their professors as we were twenty five years later. I often wish that my own students were as intimidated by me.

Not surprisingly, Seymour worked on his Masters and Ph. D. with Harold Hotelling. His Masters thesis was on computing eigenvalues and eigenvectors. His

doctoral thesis was on mean square successive differences. This came about from spending summers working at the naval proving ground in Aberdeen, Maryland. The statistics group leader, Monroe Norden, had him following up work that John von Neumann had done during World War II. Contrary to the rumor (that I started), his thesis was not based on data obtained when the denizens of the proving ground went deer hunting with their cannon.

Seymour later described his interaction with Hotelling (in Christensen and Johnson, 2005). "He was very hard to get [to see] and every time I would find him and show him my work, he would always suggest something more to do. I got to be a little annoyed at this. I thought I had done enough. So the next time he asked me to do something, I went back and I did it and I thought, what would he ask next. I thought about it and said, probably this kind of thing, and I did it. Next time I came in, sure enough, he asked me to do exactly that and I said, 'Here, I've done it.' He said, 'Well then, I guess you're finished."'

13.4 Washington, DC

After graduating from UNC in 1955, Seymour took off to the National Bureau of Standards. It paid better than the University of Illinois and he liked living in Washington. He initially worked under Churchill Eisenhart in the Statistical Engineering Laboratory. He also worked with Marvin Zelen, Jack Youden, I. R. Savage, Bill Conner, and Bill Clatworthy.

Before long he joined the U.S. Public Health Service as a lieutenant j.g. The commission was necessary for joining the National Institutes of Health. Seymour spoke fondly of his lunchtime discussions with Sam Greenhouse, Max Halperin, Nate Mantel, Marvin Schneiderman, and Jerry Cornfield. His interactions with Jerry Cornfield changed his professional life.

Cornfield was interested in Bayesian ideas and the corresponding frequentist concepts. Seymour soon caught the Bayesian bug and, given his association with Hotelling, he not surprisingly began developing Bayesian approaches to multivariate problems such as discriminant analysis and profile analysis. The work on Bayesian discrimination lead naturally to looking at predictive probabilities of correct classification. Ultimately, that lead to Seymour's seminal work on prediction as the basis for statistical inference, first summarized in his 1971 paper "The inferential use of predictive distributions" and later compiled in his (1993) book *Predictive Inference: An Introduction*.

In 1959, Seymour published his infamous citation classic on the Greenhouse-Geisser correction to the F test. The adjective "infamous" is Seymour's. He was not overly taken with the work and opined, "There is no accounting for taste." Except Seymour said it in Latin.

In the early '60s, Seymour began his academic career teaching nights at George Washington University.

13.5 Buffalo

In 1965 Seymour moved to Buffalo, NY to be the founding Chair of the Department of Statistics at the State University of New York. Norman Severo and Bill Clatworthy were already there and together they brought in Marvin Zelen, Manny Parzen, Charles Mode, Jack Kalbfleisch, Peter Enis, and Jim Dickey.

13.6 Minnesota

Two years after I began at the University of Minnesota, Seymour became the first Director of the School of Statistics. That was 1971 and they forgot to consult me, perhaps because I was a sophomore in math education at the time. Seymour remained director for 30 years, helping develop a distinguished faculty. When I began graduate school the faculty included Don Berry, Kit Bingham, Bob Buehler, Dennis Cook, Joe Eaton (I even know why Morris L. Eaton is called Joe), Steve Fienberg, Cliff Hildreth, David Hinkley, Kinley Larntz, Bernie Lindgren, Frank Martin, Milton Sobel, Bill Sudderth, and Sandy Weisberg. Later additions included Katherine Chaloner, Jim Dickey, Charlie Geyer, Doug Hawkins, David Lane, Gary Oehlert, Glen Meeden, Luke Teirney and undoubtedly others that I am less familiar with.

Unintentionally, Seymour introduced a shibboleth by which Minnesota graduates recognize each other. Every year Seymour would tell the graduate students that seminar attendance was obligatory but not mandatory. To this day, if we hear anyone say that something is obligatory but not mandatory, we assume that person is from Minnesota. I doubt that Seymour was aware of this, but I am sure that he would have enjoyed me calling it a shibboleth.

I mentioned that we were intimidated by our faculty, and Seymour was certainly not an exception. A fellow graduate student, Dennis Jennings, got married and invited Seymour to the reception. I do not remember any other faculty being there. But I do remember Seymour and Anne sitting alone and the graduate students not having the nerve to go over and socialize. Perhaps one of the reasons I'm writing this is because I eventually worked up enough nerve.

13.7 Seymour's Professional Contributions

Seymour was always a very active researcher. He had over 175 publications. He had visiting professorships at 13 universities. He was a fellow of the Institute of Mathematical Statistics and the American Statistical Association. He was on numerous national committees.

In a two year period, late in life but before getting sick, Seymour published papers on:

interim analysis of lifetime data,
Bayesian method of moments,
order statistics in Bayesian analysis,
diagnostic tests,
and he edited a book on genetics.

In later years his research had focused on

Laplace approximations
Perturbation diagnostics and robustness
Curtailment of sampling
Paternity determination
Prior Distributions

Seymour was active as a expert witness in court cases particularly those involving DNA identification. In 2000 he published a fascinating and eye-opening account of his experiences as a witness for the defense, "Statistics, litigation, and conduct unbecoming."

Of course, Seymour was best known for his advocacy of the predictive approach to statistical inference. The essence of the predictive approach is that science, and statisticians, should be concerned about predicting future observables rather than estimating or testing parameters. While I never heard Seymour do it, this idea naturally calls in question the scientific validity of any discipline that displays a limited ability to provide verifiable predictions. Cosmology and the evolution of species come to my mind.

Predictive methodology largely boils down to, "Do unto the predictive distribution, that which you would do to the sampling distribution," (if you knew the value of the parameter). Thus, to estimate the mean θ of the sampling distribution, use the mean of the predictive distribution. Seymour noted that the mean of the predictive equals the posterior mean of θ.

In 1975, Seymour introduced predictive sample reuse as a "low structure" method for predictive inference. About the same time, Stone (1974) was independently introducing the same concept as cross-validation. By 1993, when he wrote his book, Seymour had given in somewhat and was using both terminologies for the concept. In any case, the concept has become a touchstone for evaluating predictive models, cf. Breiman (2000).

Seymour's interest in both foundational issues and prediction are illustrated in a (1985) reevaluation of Bayes' paper. Standard discussions of Bayes paper typically take one of two forms. The data $X_1, \ldots, X_n$ are iid Bernoulli with probability of success θ. The sampling distribution has

$$f(x_1, \ldots, x_n|\theta) = \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}.$$

The first version observes that

$$p(\theta|x_1, \ldots, x_n) \propto \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i},$$

so

$$p(\theta) = 1.$$

Stigler (1982) has Bayes assuming a marginal distribution for the data in which

$$\Pr\left[\sum X_i = r\right] = \frac{1}{N+1}.$$

A somewhat more complicated argument again leads to

$$p(\theta) = 1.$$

Seymour's version is entirely predictivist. He notes that the data are actually $Y_0, Y_1, \ldots, Y_n$ iid $U(0,1)$ but that all one observes is

$$X_i = \begin{cases} 1 & \text{if } Y_i \leq Y_0 \\ 0 & Y_i > Y_0 \end{cases}.$$

In other words,

$$\theta \equiv \Pr[Y_i \leq Y_0 | Y_0].$$

In Seymour's formulation,

$$f(x_1, \ldots, x_n | y_0) = y_0^{\sum x_i}(1-y_0)^{n-\sum x_i}$$

and by Bayes' theorem, the predictive distribution is

$$f(y_0 | x_1, \ldots, x_n) \propto y_0^{\sum x_i}(1-y_0)^{n-\sum x_i}.$$

There are no parameters. Everything is potentially observable.

The final major work of Seymour's career is a book on *Modes of Parametric Statistical Inference* (Geisser, 2005). Characteristically, the book begins with an example taken from the Book of Numbers. It also contains a favorite example from Hacking (1965) illustrating foundational issues related to testing.

Consider the null model

$$\Pr[X = 0 | \theta = 0] = .9$$

$$\Pr[X = i | \theta = 0] = .001, i = 1, \ldots, 100.$$

The Fisherian .1 test of significance for this distribution rejects $H_0 : \theta = 0$ for $X = i$, $i = 1, \ldots, 100$. Observing anything other than $X = 0$ is somewhat weird, so that tends to contradict the (null) hypothesis. The Fisherian size is determined by the P value rather than the probability of type I error. Also, Fisherian tests do not involve an alternative, so power is not an issue.

Now consider the Neyman-Pearson (N-P) problem of testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ for $\theta = 0, 1, \ldots, 100$. The null distribution is as before and the alternative sampling distributions are

$$\Pr[X = 0|\theta = i] = .91$$

$$\Pr[X = i|\theta = i] = .09, i = 1, \ldots, 100.$$

The Fisherian test also defines a Neyman-Pearson test, so we can explore it's N-P properties. In this example, the probability of type I error is .1. When used in N-P testing, Fisherian tests can have very poor power for some alternatives since they are constructed without reference to any alternative. For these alternatives, the Fisherian test has power .09 regardless of the alternative, so its power is less than its size. This is not surprising. Given any test, you can always construct an alternative that will have power less than the size.

The most powerful test for an alternative $\theta > 0$ depends on θ, so a uniformly most powerful test does not exist. The Fisherian test is also the likelihood ratio test. The likelihood ratio examines the transformation

$$\begin{aligned} T(x) &= \frac{\Pr[X = x|\theta = 0]}{\max_{i=0,\ldots,100} \Pr[X = x|\theta = i]} \\ &= \begin{cases} .9/.91 = .989 & \text{if } x = 0 \\ .001/.09 = 1/90 & \text{if } x \neq 0 \end{cases} \end{aligned}$$

and rejects for small values of the test statistic $T(X)$. That the likelihood ratio test has power less than its size IS surprising.

The uniformly most powerful invariant (UMPI) test of size .1 is a randomized test. It rejects when $X = 0$ with probability $1/9$. The size is $.9(1/9) = .1$ and the power is $.91(1/9) > .1$. Note, however, that observing $X = 0$ does not contradict the null hypothesis because $X = 0$ is the most probable outcome under the null hypothesis. Moreover, the test does not reject for any value $X \neq 0$, even though such data are 90 times more likely to come from the alternative $\theta = X$ than from the null.

In my humble opinion, Seymour's primary research contributions were to

Non-Bayesian Multivariate Analysis
Bayesian Multivariate Analysis
Predictive Sample Reuse
and Predictive Inference.

He was also proud of his role in building the program at Minnesota.

13.8 Family Life

Seymour had four children: Adam, Dan, Georgia, and Mindy. Mindy is a biostatistician in Minnesota. He had five grandchildren, including a set of triplets. When Seymour visited Harvard once, Marvin Zelen asked his administrator Anne to take care of him. She took the job so seriously that they married. Seymour's brother Martin taught high school and served as a counselor. Seymour enjoyed history, archeology,

religion and novels. More than once I was surprised at how well read he was. He seemed to know something about even my most obscure interests.

13.9 Conclusion

My personal favorite memories of Seymour come from professional meetings, particularly the Joint Statistical Meetings. Many's the time Wes Johnson would drag Seymour and I to some seedy or noisy bar. Seymour and I would get fed up with the place about the same time and the two of us would walk back to our hotels together, sometimes through rough parts of town. I always felt rather protective of Seymour on these occasions — rather like I would my father.

I learned relatively little from the lectures I attended during graduate school. I learned that material far better from reading the books recommended by my professors. Nonetheless, what I learned in graduate school was irreplaceable. I learned a sense of what is important in statistics. From Seymour I learned the importance of foundational issues and the importance of prediction. I try to view all of my work through that lens.

Seymour died March 11, 2004.

Acknowledgements

Aelise Houx, Anne Geisser, Jessica Utts, Marvin Zelen, Martin Geisser, and Wes Johnson helped accumulate this information. The late, great Larry Brown pointed out an error in an earlier draft. Actually, after my talk he privately pointed out a blunder and I have been forever grateful that he chose not to humiliate me. That was the only time I ever met him.

References

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53**, 370-418.

Breiman, Leo (2000). "Statistical Modeling: The Two Cultures," with discussion. *Statistical Science*, **16**, 199-231.

Christensen, Ronald and Johnson, Wesley (2005). "A Conversation with Seymour Geisser." Submitted to *Statistical Science*.

Christensen, Ronald and Huffman, Michael D. (1985). "Bayesian point estimation using the predictive distribution." *The American Statistician*, **39**, 319-321.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*, Third Edition, 1973. Hafner Press, New York.

Fisher, R. A. (1935). *The Design of Experiments*, Ninth Edition, 1971. Hafner Press, New York.

Geisser, Seymour (1971). The inferential use of predictive distributions. In *Foundations of Statistical Inference*, V.P. Godambe and D.A. Sprott (Eds.). Holt, Rinehart, and Winston, Toronto, 456-469.

Geisser, Seymour (1975). The predictive sample reuse method with applications. *Biometrika*, **70**, 320-328.

Geisser, Seymour (1985). On the predicting of observables: A selective update. In *Bayesian Statistics 2*, J.M. Bernardo et al. (Eds.). North Holland, 203-230.

Geisser, Seymour (1993). *Predictive Inference: An Introduction*, Chapman and Hall, New York.

Geisser, Seymour (2000). Statistics, litigation, and conduct unbecoming. In *Statistical Science in the Courtroom*, Joseph L. Gastwirth (Ed.). Springer-Verlag, New York, 71-85.

Geisser, Seymour (2005). *Modes of Parametric Statistical Inference*, John Wiley and Sons, New York.

Hacking, I. (1965). *Logic of Statistical Inference*. Cambridge University Press.

Lane, David (1996). "Story about Cosimo di Medici." In *Modelling and Prediction: honoring Seymour Geisser*, eds. Jack C. Lee, Wesley O. Johnson, Arnold Zellner. Springer- Verlag, New York.

Stigler, S.M. (1982). Thomas Bayes and Bayesian inference. *Journal of the Royal Statistical Society, A*, **145**(2), 250-258.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, B*, **36**, 44-47.

Zelen, Marvin (1996). "After dinner remarks: On the occasion of Seymour Geisser's 65th Birthday, Hsinchu, Taiwan, December 13, 1994." In *Modelling and Prediction: honoring Seymour Geisser*, eds. Jack C. Lee, Wesley O. Johnson, Arnold Zellner. Springer-Verlag, New York.

Appendix A
Multivariate Distributions

This appendix reviews properties of multivariate distributions. It contains a few additions to similar material in *PA*.

Let $y = (y_1, \ldots, y_n)'$ be a random vector. The joint cumulative distribution function (cdf) of $(y_1, \ldots, y_n)'$ is

$$F(v_1, \ldots, v_n) \equiv \Pr[y_1 \leq v_1, \ldots, y_n \leq v_n].$$

If $F(v_1, \ldots, v_n)$ is the cdf of a discrete random variable, we can define a (joint) probability mass function

$$f(v_1, \ldots, v_n) \equiv \Pr[y_1 = v_1, \ldots, y_n = v_n].$$

If $F(v_1, \ldots, v_n)$ admits the nth order mixed partial derivative, then we can define a (joint) density function

$$f(v_1, \ldots, v_n) \equiv \frac{\partial^n}{\partial v_1 \cdots \partial v_n} F(v_1, \ldots, v_n).$$

The cdf can be recovered from the density as

$$F(v_1, \ldots, v_n) = \int_{-\infty}^{v_1} \cdots \int_{-\infty}^{v_n} f(w_1, \ldots, w_n) dw_1 \cdots dw_n.$$

We (with D.R. Cox) will often adopt the "deplorable" habit of referring to probability mass functions as discrete densities, or if the context is clear, just densities.

For a function $g(\cdot)$ of $(y_1, \ldots, y_n)'$ into $\mathbf{R}$, the expected value is defined as

$$\mathrm{E}[g(y_1, \ldots, y_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(v_1, \ldots, v_n) f(v_1, \ldots, v_n) dv_1 \cdots dv_n.$$

We might also write this as $\mathrm{E}_y[g(y)]$.

The expected value of any random matrix is performed element wise, e.g.,

$$\mathrm{E}(y) = [\mathrm{E}(y_1), \ldots, \mathrm{E}(y_n)]'$$

and, if $\mathrm{E}(y) = \mu$,

$$\mathrm{Cov}(y) \equiv \mathrm{E}[(y-\mu)(y-\mu)'].$$

It is easily seen that for a conformable fixed matrix A and vector b,

$$\mathrm{E}(Ay+b) = A\mathrm{E}(y)+b; \qquad \mathrm{Cov}(Ay+b) = A\mathrm{Cov}(y)A'.$$

We now consider relationships between two random vectors, say $x = (x_1, \ldots, x_m)'$ and $y = (y_1, \ldots, y_n)'$. Assume that the joint vector $(x', y')' = (x_1, \ldots, x_m, y_1, \ldots, y_n)'$ has a density function

$$f_{x,y}(u,v) \equiv f_{x,y}(u_1, \ldots, u_m, v_1, \ldots, v_n),$$

where $u = (u_1, \ldots, u_m)'$ and $v = (v_1, \ldots, v_n)'$ are vectors of placehold variables corresponding to x and y respectively. Similar definitions and results hold if $(x', y')'$ has a probability mass function. Of particular note is that if $\mathrm{E}(x) = \mu_x$ and $\mathrm{E}(y) = \mu_y$,

$$\mathrm{Cov}(x,y) \equiv \mathrm{E}[(x-\mu_x)(y-\mu_y)'],$$

with, for conformable fixed matrices and vectors A_1, A_2, b_1, b_2:

$$\mathrm{Cov}(A_1x+b_1, A_2y+b_2) = A_1\mathrm{Cov}(x,y)A_2'.$$

The distribution of one random vector, say x, ignoring the other vector, y, is called the *marginal distribution* of x. The marginal cdf of x can be obtained by substituting the value $+\infty$ into the joint cdf for all of the y variables:

$$F_x(u) = F_{x,y}(u_1, \ldots, u_m, +\infty, \ldots, +\infty).$$

The marginal density can be obtained either by partial differentiation of $F_x(u)$ or by integrating the joint density over the y variables:

$$f_x(u) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{x,y}(u_1, \ldots, u_m, v_1, \ldots, v_n) dv_1 \cdots dv_n.$$

A.1 Conditional Distributions

The conditional density of a vector, say x, given the value of the other vector, say $y = v$, is obtained by dividing the density of $(x', y')'$ by the density of y evaluated at v, i.e.,

$$f_{x|y}(u|v) \equiv f_{x,y}(u,v) \big/ f_y(v).$$

The conditional density is a well-defined density, so expectations with respect to it are well defined. Let g be a function from $\mathbf{R}^m$ into $\mathbf{R}$,

$$\mathrm{E}[g(x)|y=v] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(u) f_{x|y}(u|v) du,$$

where $du \equiv du_1 du_2 \cdots du_m$. Sometimes we write

$$\mathrm{E}_{x|y=v}[g(x)] \equiv \mathrm{E}_x[g(x)|y=v] \equiv \mathrm{E}[g(x)|y=v]\,.$$

The standard properties of expectations hold for conditional expectations. For example, with a and b real,

$$\mathrm{E}[ag_1(x)+bg_2(x)|y=v] = a\mathrm{E}[g_1(x)|y=v]+b\mathrm{E}[g_2(x)|y=v]\,.$$

The conditional expectation of $\mathrm{E}[g(x)|y=v]$ is a function of the value v. Since y is random, we can consider $\mathrm{E}[g(x)|y=v]$ as a random variable. In this context we write $\mathrm{E}[g(x)|y]$ or $\mathrm{E}_x[g(x)|y]$ or $\mathrm{E}_{x|y}[g(x)]$. An important property of conditional expectations is

$$\mathrm{E}[g(x)] = \mathrm{E}[\,\mathrm{E}[g(x)|y]\,]\,.$$

To see this, note that $f_{x|y}(u|v)f_y(v) = f_{x,y}(u,v)$ and

$$\begin{aligned}
\mathrm{E}[\,\mathrm{E}[g(x)|y]\,] &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathrm{E}[g(x)|y=v]\, f_y(v) dv \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(u) f_{x|y}(u|v) du \right] f_y(v) dv \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(u) f_{x|y}(u|v) f_y(v) du\, dv \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(u) f_{x,y}(u,v) du\, dv \\
&= \mathrm{E}[g(x)]\,.
\end{aligned}$$

In fact, both the notion of conditional expectation and this result can be generalized. Consider a function $g(x,y)$ from $\mathbf{R}^{m+n}$ into $\mathbf{R}$. If $y=v$, we can define $\mathrm{E}[g(x,y)|y=v]$ in a natural manner. If we consider y as random, we write $\mathrm{E}[g(x,y)|y]$. It can be easily shown that

$$\mathrm{E}[g(x,y)] = \mathrm{E}[\mathrm{E}[g(x,y)|y]]\,.$$

A function of x or y alone can also be considered as a function from $\mathbf{R}^{m+n}$ into $\mathbf{R}$.

A second important property of conditional expectations is that if $h(y)$ is a function from $\mathbf{R}^n$ into $\mathbf{R}$, we have

$$\mathrm{E}[h(y)g(x,y)|y] = h(y)\mathrm{E}[g(x,y)|y]\,. \tag{1}$$

This follows because if $y=v$,

$$\mathrm{E}[h(y)g(x,y)|y=v] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(v) g(u,v) f_{x|y}(u|v) du$$

$$= h(v) \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(u,v) f_{x|y}(u|v) du$$
$$= h(v) \mathrm{E}[g(x,y)|y = v] .$$

This is true for all v, so (1) holds. In particular, if $g(x,y) \equiv 1$, we get

$$\mathrm{E}[h(y)|y] = h(y).$$

Finally, we can extend the idea of conditional expectation to a function $g(x,y)$ from $\mathbf{R}^{m+n}$ into $\mathbf{R}^s$. Write $g(x,y) = [g_1(x,y), \ldots, g_s(x,y)]'$. Then define

$$\mathrm{E}[g(x,y)|y] = (\mathrm{E}[g_1(x,y)|y], \ldots, \mathrm{E}[g_s(x,y)|y])' .$$

Exercise A.1 Show that

$$\mathrm{E}(x) = \mathrm{E}_y[\mathrm{E}_{x|y}(x)] \equiv \mathrm{E}_y[\mathrm{E}_x(x|y)]$$

and

$$\mathrm{Cov}(x) = \mathrm{Cov}_y[\mathrm{E}_{x|y}(x)] + \mathrm{E}_y\{\mathrm{Cov}_{x|y}(x)\} \equiv \mathrm{Cov}_y[\mathrm{E}(x|y)] + \mathrm{E}_y\{\mathrm{Cov}(x|y)\}.$$

(The equivalences in the displays are merely notation.)

A.2 Independence

If their densities exist, two random vectors are *independent* if and only if their joint density is equal to the product of their marginal densities, i.e., x and y are independent if and only if

$$f_{x,y}(u,v) = f_x(u) f_y(v).$$

Note that if x and y are independent,

$$f_{x|y}(u|v) = f_x(u).$$

If the random vectors x and y are independent, then any (reasonable) vector-valued functions of them, say $g(x)$ and $h(y)$, are also independent. This follows easily from a more general definition of the independence of two random vectors: The random vectors x and y are independent if for any two (reasonable) sets A and B,

$$\Pr[x \in A, y \in B] = \Pr[x \in A] \Pr[y \in B].$$

To prove that functions of random variables are independent, recall that the set inverse of a function $g(u)$ on a set A_0 is $g^{-1}(A_0) \equiv \{u|g(u) \in A_0\}$. That $g(x)$ and $h(y)$ are independent follows from the fact that for any (reasonable) sets A_0 and B_0,

$$\begin{aligned}\Pr[g(x) \in A_0, h(y) \in B_0] &= \Pr[x \in g^{-1}(A_0), y \in h^{-1}(B_0)] \\ &= \Pr[x \in g^{-1}(A_0)] \Pr[y \in h^{-1}(B_0)] \\ &= \Pr[g(x) \in A_0] \Pr[h(y) \in B_0].\end{aligned}$$

By "reasonable" I mean things that satisfy the mathematical definitions of being measurable, cf. Appendix B.

A.3 Characteristic Functions

The *characteristic function* of a random vector $y = (y_1, \ldots, y_n)'$ is a function from $\mathbf{R}^n$ to $\mathbf{C}$, the complex numbers. It is defined on a vector $t = (t_1, \ldots, t_n)'$ by

$$\varphi_y(t_1, \ldots, t_n) \equiv \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left[i \sum_{j=1}^{n} t_j v_j \right] f_y(v_1, \ldots, v_n) dv_1 \cdots dv_n = \mathrm{E}[e^{it'y}].$$

We are interested in characteristic functions because if $x = (x_1, \ldots, x_n)'$ and $y = (y_1, \ldots, y_n)'$ are random vectors and if

$$\varphi_x(t_1, \ldots, t_n) = \varphi_y(t_1, \ldots, t_n)$$

for all $(t_1, \ldots, t_n)$, then x and y have the same distribution.

The great advantage of characteristic functions is that $e^{it'y} \equiv \cos(t'y) + i \sin(t'y)$ so that the random variable is bounded, so its expectation always exists. The *moment generating function* gets rid of $i \equiv \sqrt{-1}$ and is

$$\psi_y(t_1, \ldots, t_n) \equiv \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left[\sum_{j=1}^{n} t_j v_j \right] f_y(v_1, \ldots, v_n) dv_1 \cdots dv_n = \mathrm{E}[e^{t'y}].$$

When it exists, it has similar properties to the characteristic function.

A.4 Inequalities

A.4.1 Chebyshev's Inequalities

Suppose g is a nondecreasing function on nonnegative real values with the properties that $g(0) \geq 0$ and $g(1) \geq 1$. In that case, the indicator function

$$\mathscr{I}_A(x, y) = \begin{cases} 1 & \text{if } (x, y) \in A \\ 0 & \text{if } (x, y) \notin A \end{cases},$$

has the property that

$$\mathscr{I}_{[1,\infty)}(v) \le g(v).$$

For example $g(v) = v^2$ works as does $g(v) = v$. In particular, for $\varepsilon > 0$,

$$\mathscr{I}_{[\varepsilon,\infty)}(|y-\mu_y|) = \mathscr{I}_{[1,\infty)}(|y-\mu_y|/\varepsilon) \le g(|y-\mu_y|/\varepsilon).$$

Taking expected values gives

$$P(|y-\mu_y| \ge \varepsilon) = \mathrm{E}\left[\mathscr{I}_{[1,\infty)}(|y-\mu_y|/\varepsilon)\right] \le \mathrm{E}\left[g(|y-\mu_y|/\varepsilon)\right].$$

In particular, for $g(v) = v^2$, we have $\mathrm{E}\left[|y-\mu_y|^2/\varepsilon^2\right] = \mathrm{Var}(y)/\varepsilon^2$, so we get

$$P(|y-\mu_y| \ge \varepsilon) \le \mathrm{Var}(y)/\varepsilon^2.$$

A.4.2 Jensen's Inequality

Jensen's inequality is that for any *convex function* $g(u)$ and random variable y, that

$$\mathrm{E}[g(y)] \ge g[\mathrm{E}(y)].$$

I find the easiest way to remember Jensen's inequality is to remember that it is an application of the fact that $0 \le \mathrm{Var}(y) = \mathrm{E}(y^2) - [\mathrm{E}(y)]^2$.

The function u^2 needs to be convex in the sense that for any points u_0 and u_1, and any $\alpha \in [0,1]$,

$$\alpha g(u_0) + (1-\alpha) g(u_1) \ge g[\alpha u_0 + (1-\alpha) u_1].$$

A.5 Change of Variables

Consider the random vectors $x = (x_1, \ldots, x_n)'$ and $y = (y_1, \ldots, y_n)'$. Suppose x has density $f_x(u)$ and that y is the result of a one-to-one transformation $y = T(x)$ so that $x = T^{-1}(y)$. We want to find $f_y(v)$ the density of y. This section involves derivatives the notation for which is set in Appendix F. The *change of variable formula* is that the density of y is

$$f_y(v) = f_x[T^{-1}(v)] \left|\det[\mathbf{d}_v T^{-1}(v)]\right|,$$

where $\mathbf{d}_v T^{-1}(v)$ is the derivative (matrix of partial derivatives) of T^{-1} evaluated at v.

Not infrequently, the determinant of the derivative is replaced with an equivalent form. Using results from Appendix F, since $v = T[T^{-1}(v)]$, differentiating both sides

and using the chain rule on the right hand side gives $I = [\mathbf{d}_u T(u)|_{u=T^{-1}(v)}][\mathbf{d}_v T^{-1}(v)]$. Moreover, since $\det(I) = 1$ and, for conformable square matrices A and B, $\det(AB) = \det(A)\det(B)$, we get $\det[\mathbf{d}_v T^{-1}(v)] = 1/\det[\mathbf{d}_u T(u)|_{u=T^{-1}(v)}]$.

EXAMPLE A.5.1. *Location Families.*
Let the random variable x have density $h(u)$ and let $y = x + \theta$. We can find the the density of y from the cdf of y without resorting to the general formula. (Actually, the general formula is precisely the general statement of how you these simple change of variable problems.)

$$\int_{-\infty}^{a} f(v|\theta)dv = P[y \le a] = P[x+\theta \le a] = P[x \le a-\theta] = \int_{-\infty}^{a-\theta} h(u)du = \int_{-\infty}^{a} h(v-\theta)dv$$

The score function and the information are important concepts in statistical inference and are discussed in Chapter 6. The next equation implicitly defines the score function and then simplifies it for location families. Here $\dot{f}(y|\theta) \equiv \mathbf{d}_\theta f(v|\theta)$

$$S(y;\theta) \equiv \frac{\dot{f}(y|\theta)}{f(y|\theta)} = \frac{-\dot{h}(y-\theta)}{h(y-\theta)}.$$

Similarly the information is

$$\mathbf{I}(\theta) \equiv \mathrm{E}[S(y;\theta)]^2 = \int \frac{[\dot{h}(v-\theta)]^2}{h(v-\theta)}\,dv = \int \frac{[\dot{h}(v)]^2}{h(v)}\,dv.$$

This integral, and thus the information, does not depend on θ.

EXAMPLE A.5.2. *Scale Families.*
Let the random variable x have density $h(u)$ and let $y = \theta x$, $\theta > 0$. We can find the the density of y from the cdf of y without resorting to the general formula.

$$\int_{-\infty}^{a} f(v|\theta)dv = P[y \le a] = P[\theta x \le a] = P[x \le a/\theta] = \int_{-\infty}^{a/\theta} h(u)du = \int_{-\infty}^{a} \frac{1}{\theta} h(v/\theta)dv.$$

As in the previous example, the last equality is basically due to the general change of variable formula.

This paragraph was not in my notes and is currently wrong because all I did is swap $-$ for $/$. Using concepts defined in Chapter 5, the score function is

$$S(y;\theta) \equiv \frac{\dot{f}(y|\theta)}{f(y|\theta)} = \frac{-\dot{h}(y/\theta)}{h(y/\theta)}$$

and the information is

$$\mathbf{I}(\theta) \equiv \mathrm{E}[S(y;\theta)]^2 = \int \frac{[\dot{h}(v/\theta)]^2}{h(v/\theta)}\,dv = \int \frac{[\dot{h}(v)]^2}{h(v)}\,dv.$$

This integral, and thus the information, depend on θ. **Paragraph wrong**

EXAMPLE A.5.3. *Location-Scale Families.*
Let the random variable x have density $h(u)$ and let $y = \sigma x + \mu$, $\sigma > 0$.

$$f_y(v|\mu,\sigma) = \frac{1}{\sigma} h\left(\frac{v-\mu}{\sigma}\right).$$

A very common location transformation is between Celsius and Fahrenheit, $F = (9/5)C + 32$, although in Statistics we generally think of μ and σ as unknown parameters.

EXAMPLE A.5.4. *Generalized Linear Transformations.*
Generalized linear models involve an assumption that a random n vector Y has $\mathrm{E}(Y) = G(X\beta)$ where $X_{n\times p}$ is known and β is a parameter vector. The vector function G actually repeatedly applies a scalar function $G(u)$ by defining $G(v) \equiv [G(v_1),\cdots,G(v_n)]'$. Also define the scalar function $g(u) \equiv \mathbf{d}_u G(u)$ with its vector equivalent. It follows that $\mathbf{d}_v G(v) = D[g(v)]$ a diagonal matrix of derivatives and from the chain rule that $\mathbf{d}_\beta G(X\beta) = D[g(X'\beta)]X$.

For convenience, in this appendix we have assumed the existence of densities. Many of these concepts generalize easily with measure theory.

A.6 Exercises

Exercise A.6.1. Let x and y be independent. Show that
(a) $\mathrm{E}[g(x)|y] = \mathrm{E}[g(x)]$;
(b) $\mathrm{E}[g(x)h(y)] = \mathrm{E}[g(x)]\,\mathrm{E}[h(y)]$.

Exercise A.6.2. Using the methods of Subsection A.4.1, prove Markov's inequality: for $\varepsilon > 0$,

$$P(|y| \geq \varepsilon) \leq \mathrm{E}(|y|)/\varepsilon.$$

Exercise A.6.3. Let $U \sim U[0,1]$ with density $f_U(u) = \mathscr{I}_{(0,1)}$. Let $Y \equiv 2U$ Use the change of variable formula to find the density of Y. You should be able to guess the correct answer. The problem is show that it is correct.

Exercise A.6.4. Let the n-vector Z have independent standard normal components and for a fixed nonsingular matrix A and a fixed vector μ define $Y \equiv AZ + \mu$. Find the mean and covariance matrix of Y and use the change of variable formula to find the density of Y in terms of the mean and covariance matrix. Hints: Recall that determinants have the properties that $\det(A) = \det(A')$ and $\det(AB) = \det(A)\det(A)$.

Appendix B
Measure Theory and Convergence

This book does not require the reader to know measure theory or measure theoretic probability. But some of the ideas in measure theoretic probability are extremely useful and we seek to provide some intuition for them.

B.1 A Brief Introduction to Measure and Integration

Lebesgue measure generalizes the concepts of length, area, and volume to arbitrary sets in an arbitrary number of dimensions. Life being what it is, some people are smart enough to find sets of points for which even Lebesgue's theory is incapable of finding their length etc., but for most of us, any set we can dream up, Lebesgue's theory will measure.

Any reasonable measure of length has to satisfy certain properties. If A is any set, its length, say $\mu(A)$, has to be greater than or equal to 0. If you have any two sets A_1 and A_2, like (0.2,0.6] and (0.5,0.7], the total length of the set has to satisfy $\mu(A_1 \cup A_2) \leq \mu(A_1) + \mu(A_2)$. If the sets are disjoint the inequality becomes an equality. If this works for two sets, it works for any finite number of sets. We will assume that it also works for a countably infinite number of sets, although one can have philosophical debates about that. Incidentally, the finite version is enough to ensure that if $A_1 \subset A_2$, then $\mu(A_1) \leq \mu(A_2)$.

Let's get our hands dirty by showing that the length of the set of rational numbers in the unit interval is 0. Let $h = 1,2,3,\ldots$ and $i = 1,\ldots,h$. The numbers i/h are all of the rational numbers in $(0,1]$ (with many numbers – obviously 1 – repeated many times). We want to list all of these numbers with a single index so take $n = (h-1)h/2 + i$. If you know n, you can figure out what h and i have to be. If you only know i/h, rather than i and h there are lots of values n that correspond to it, but that is no problem.

Let $\mathbf{Q}$ denote the rationals in $(0,1]$. For any $\varepsilon > 0$, and any n we put a ball (interval) around i/h of length $\varepsilon/2^n$, call the ball A_n. Obviously $\{i/h\} \subset A_n$ and

$$\mu(\mathbf{Q}) = \mu\left(\bigcup_{h=1}^{\infty}\bigcup_{i=1}^{h}\{i/h\}\right) \leq \mu\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty}\mu(A_n) = \sum_{n=1}^{\infty}\frac{\varepsilon}{2^n} = \varepsilon.$$

So, proof by contradiction. If you claim that $\mu(\mathbf{Q}) = \delta > 0$, I can find $\varepsilon < \delta$ that contradicts your claim. The only length than can work for the rationals is $\mu(\mathbf{Q}) = 0$.

The same argument will establish that *any* countable set of points has Lebesgue measure 0, so in particular any finite set of points has Lesbegue measure 0.

By definition, the numbers in (0,1] that are not rational are irrational, say **Ir**. Since $(0,1] = \mathbf{Q} \cup \mathbf{Ir}$ and the sets are disjoint,

$$1 = \mu\{(0,1]\} = \mu(\mathbf{Q} \cup \mathbf{Ir}) = \mu(\mathbf{Q}) + \mu(\mathbf{Ir}) = 0 + \mu(\mathbf{Ir}) = \mu(\mathbf{Ir}).$$

Anything that occurs except on a set of measure 0 is said to occur *almost everywhere*. So irrational numbers occur in (0,1] almost everywhere.

An integral measures the area (volume, hypervolumne) under a curve. Suppose we have a function from (x,y) into z, say $f(x,y)$, and we want to measure the volume under the curve defined by $f(x,y)$.

The Reimann idea of an integral is to divide the (x,y) points into small regions and approximate the volume under the curve for that region as the area of the region times the height, where the height is just the value of $f(x,y)$ for any point (x,y) in the region. (There is a fair amount of slop here, but don't worry.) The approximate volume under the entire curve is just the sum of the approximate volumes for all the small regions. Does this always work? Of course not! If x and y are allowed to be any real numbers, you need an infinite number of small regions to sum over. If $f(x,y)$ is not well behaved, it can matter a great deal which (x,y) you pick in each region to use as the height for the approximate volume. We need a guarantee that the values of $f(x,y)$ cannot vary too much within our small (x,y) regions. But for lots of well-behaved functions, this idea works well.

The Lebesgue idea of an integral is different and works (by and large) for less well-behaved functions than you need for Reimann integration. There are Lebesgue integrable functions that are not Reimann integrable and Reimann integrable functions that are not Lebesgue integrable, but *typically when both integrals exist they give the same result.* One key point is that no function f is Lebesgue integrable unless its absolute value is also integrable. That is not a requirement for Reimann integrable functions.

In our 3-dimensional example, Reimann integration divides up the (x,y) plane into small regions whereas Lebesgue integration divides the z axis into small regions. For each small region of the z axis, there exist a set of points (x,y) that will give you a value of $f(x,y)$ in that small region of z. This set of (x,y) points can be very complicated, but as mentioned earlier, Lebesgue measure can be used to determine the area associated with this complicated set of points. Again, we will approximate the volume by the area of the set in the x,y plane times the height of the function but now the height is restricted to be in a very narrow region of the z axis and we are using Lebesgue measure theory to give us the area of the corre-

sponding set of (x,y) values. We have a good way of measuring areas, so this is subject to much less variability (in the z direction) than is Reimann integration.

Of course the theory is far more complicated than this. Lebesgue integrals are defined first for something called simple functions (a generalization of step functions) for which the integral is easy to compute and then simple functions are used to approximate more complicated functions, with the simple function integrals approximating the the more complicated function's integral.

Although it is rather redundant, the Lebesgue integral of a function that is 1 when $(x,y) \in A$ and 0 when $(x,y) \notin A$, is the area of the set A. In other words, define the indicator function

$$\mathscr{I}_A(x,y) = \begin{cases} 1 & \text{if } (x,y) \in A \\ 0 & \text{if } (x,y) \notin A \end{cases},$$

then

$$\mu(A) = \int \mathscr{I}_A(x,y)\, d\mu(x,y),$$

where μ is being used to denote Lebesgue measure and $d\mu(x,y)$ denotes that we are integrating with respect to Lebesgue measure. If $f(x,y) = g(x,y)$ almost everywhere, their integrals must be the same, i.e., $\int f(x,y)\, d\mu(x,y) = \int g(x,y)\, d\mu(x,y)$.

The whole idea of Lebesgue integration is based on Lebesgue measure which corresponds to our usual (Euclidean) conception of length, area, and volume. As systematized by Kolmogorov, probability is just an alternative measure that replaces length, area, or volume. Probabilities act much like areas. The area of any set has to be at least 0. Same for probabilities. The total area of your living quarters is the sum of the areas for each room. The probability of you being in your living quarters is the sum of the probabilities of you being in each room. The probability of the union of disjoint sets has to be the sum of the probabilities for each set. The probability of the union of any sets has to be less than or equal to the sum of the probabilities for each set. Again, that is pretty obvious if you have a finite number of (disjoint) sets but life is much, much easier if we also assume that it is true for an infinite number of disjoint sets. (Again, this is a matter of some controversy, especially among Bayesian statisticians.) The main difference between a probability measure and Lebesgue measure is that the biggest a probability can ever be is 1.

Suppose we have a function from (x,y) into z, say $f(x,y)$ and a probability measure on (x,y), say P. Using the same ideas as for Lebesgue measure we can define integrals with respect to the probability measure P, say

$$\int f(x,y)\, dP(x,y).$$

Now, anything that holds with probability one is said to hold *almost surely (a.s.)*.

Technically, a probability space is a triple $(\Omega, \mathscr{F}, P)$ where Ω is the set of possible outcomes, $\mathscr{F}$ is the collection of sets (the sigma field) of outcomes that we will work with (remember some sets are too weird for Lebesgue measure), and P is the probability measure which is defined for every set in $\mathscr{F}$. Our three main rules are

1. $P(\emptyset) = 0$; the probability that nothing happens (the empty set $\emptyset$ occurs) is 0

2. $P(\Omega) = 1$; the probability that something happens is 1
3. $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$ provided $A_i \cap A_j = \emptyset$ for all i, j.

Item 3 is referred to as *countable additivity*. It follows from item 3 that probabilities display *countable subadditivity*, namely,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i),$$

for any sets A_i.

A random variable y is a (measurable) mapping of Ω into the real line, $y : \Omega \to \mathbf{R}$. My favorite example of a random variable is that when you roll a die, some spots appear on the top of the die. Without even thinking, we say that we rolled some number that is 1, 2, 3, 4, 5, or 6. Now ω is the face of the die that comes up on top, and $y(\omega)$ is the number of spots on that face. In this simple case, $\mathscr{F}$ is literally all of the sets you can make up out of the six outcomes and you already know P.

The real valued function/random variable y is said to be measurable if it transforms $(\Omega, \mathscr{F})$ into $(\mathbf{R}, \mathscr{B})$ where $\mathscr{B}$ is another big set (sigma field) of allowable sets and if the transformation always relates sets in $\mathscr{B}$ to sets in $\mathscr{F}$. Specifically for any set B, define the inverse set

$$y^{-1}(B) \equiv \{\, \omega | y(\omega) \in B \,\}.$$

The random variable y is measurable if for any $B \in \mathscr{B}$, we always have $y^{-1}(B) \in \mathscr{F}$. For comparison, it is well known that if $(\Omega, \mathscr{F}) = (\mathbf{R}, \mathscr{B})$, so we can talk about continuous functions, y is continuous if and only if $y^{-1}(B)$ is an open set whenever B is an open set.

The expected value of y is defined by its integral with respect to the probability measure,

$$\mathrm{E}(y) \equiv \int y(\omega)\, dP(\omega).$$

When I first learned probability, one thing that stumped me is that at some point Ω went away. We stop really needing it because we focus exclusively on random variables. If we have $(\Omega, \mathscr{F}, P)$ and a random variable y, we can just as well work with a new probability space $(\mathbf{R}, \mathscr{B}, P_y)$ in which we define P_y by

$$P_y(B) \equiv P[y^{-1}(B)] \qquad \text{for any } B \in \mathscr{B},$$

so we can act like the probabilities were all defined on the real line to begin with. In this case, the random variable y defined on $(\mathbf{R}, \mathscr{B}, P_y)$, takes the value $y(u) = u$ for any $u \in \mathbf{R}$.

A random vector is a mapping $y : \Omega \to \mathbf{R}^n$ and we take the expected values elementwise, i.e., write $y(\omega) = [y_1(\omega), \ldots, y_n(\omega)]'$ and $\mathrm{E}(y) \equiv [\mathrm{E}(y_1), \ldots, \mathrm{E}(y_n)]'$. Everything discussed to this point extends is pretty obvious ways.

For a space $(\mathbf{R}^n, \mathscr{B}^n)$, a probability distribution P is said to be *absolutely continuous* with respect to Lebesgue measure μ if for any set $A \in \mathscr{B}^n$ with $\mu(A) = 0$, we

also have $P(A) = 0$. We will see in Appendix C that this is the property that allows us to find density functions as in Appendix A. Standard continuous distributions like the normal are specified by their density wrt Lebesgue measure and standard discrete distributions are specified by their density wrt counting measure (in which the measure of set is the number of integers in the set). Most of the standard continuous distributions used in Statistics are defined by their densities with respect to Lebesgue measure, so they are automatically absolutely continuous. Although our primary interest is in relating a probability measure to Lebesgue measure on $(\mathbf{R}^n, \mathscr{B}^n)$, the concept of absolute continuity works for any two measures defined on the same space.

For example, we might take (y_1, y_2) to be the probability of a randomly selected man and woman. It is actually hard to uniquely define a man or woman's height and all measurements are fundamentally discrete but lets pretend that heights are well-defined and continuous. What is the length of the set $\{65\}$? It is only one point, it has no length. Similarly, any probability (absolutely continuous with respect to Lebesgue measure) has the probability that a person is 65 inches tall is also 0. There is some positive probability that a person is between, 64.5 and 65.6 inches tall. But no chance that someone is exactly 65 inches. In reality, all of the measurements that we make in life (length, mass, time etc), although we think of them as measuring continuous variables, are really statements that the measurement is within some interval (centered at a rational number) determined by the accuracy of the measuring device. Approximating this as a continuous measurement, *rarely* causes problems.

B.2 A Brief Introduction to Convergence

We need to consider when a sequence of random variables (or vectors) y_n converges to another random variable (or vector), say, y. We will discuss four different types of convergence: *convergence in distribution (in law), convergence in probability, convergence with probability one (almost sure convergence), and convergence in $\mathscr{L}^2$ (mean square convergence)*. (Convergence in $\mathscr{L}^p$ for $p > 0$ is similar to $\mathscr{L}^2$.) There is a partial ordering on these. Almost sure convergence and convergence in mean square both imply convergence in probability but neither implies the other. Convergence in probability ensures convergence in distribution.

Because we presume previous exposure to these topics, *we illustrate them prior to defining them.* Our illustrative probability space is the unit interval $\Omega = [0,1]$ with the uniform distribution (and Borel sets). Thus, the probability of any set is just its length. We will define sequences of random variables y_n that converge to a random variable $y(\omega) = 0$ for all ω. (Remember $\omega \in [0,1]$.)

EXAMPLE B.2.1. Consider the random variable y_n defined as the indicator function of the set $[0, 1/n]$, i.e.,

$$y_n(\omega) = \mathscr{I}_{[0,1/n]}(\omega).$$

This random variable converges in all four ways.

It converges almost surely to 0 because for every $\omega \in (0,1]$, the sequence of numbers $y_n(\omega)$ converges to the number 0. (As soon as $1/n < \omega$, we get $y_n(\omega) = 0$ AND by assumption the probability that $\omega \in (0,1]$ is 1, i.e., $P\{(0,1]\} = 1$. The fact that $y_n(0) = 1$ for all n, so that $y_n(0) = 1 \not\to y(0) = 0$ does not matter because it occurs with zero probability.

To get convergence in mean square we need the limit of $\int [y_n - y]^2 \, dP$ to go to 0. Since $y = 0$ a.s.,

$$\int [y_n - y]^2 \, dP = \int [\mathscr{I}_{[0,1/n]}(\omega)]^2 \, d\omega = \int \mathscr{I}_{[0,1/n]}(\omega) \, d\omega = \int_0^{1/n} d\omega = \frac{1}{n} \to 0.$$

To get convergence in probability we need the limit of $P[|y_n - y| > \varepsilon]$ to go to 0 for any $\varepsilon > 0$. Since $y = 0$ a.s., for $0 < \varepsilon \leq 1$ (bigger εs are easy)

$$P[|y_n - y| > \varepsilon] = P[y_n = 1] = P(0 \leq \omega \leq 1/n) = \frac{1}{n} \to 0.$$

For convergence in distribution, we need the cdf of y_n, say $F_n(v) \equiv P[y_n \leq v]$ to converge to the cdf $F(v) \equiv P[y \leq v]$ for every point v at which $F(v)$ is continuous.

$$F(v) = \mathscr{I}_{[0,\infty)}(v),$$

so we need continuity everywhere except at $v = 0$. Since y_n is either 1 or 0

$$F_n(v) \equiv P[y_n \leq v] = \begin{cases} 0 & \text{if } v < 0, \\ 1 - \frac{1}{n} & \text{if } 0 \leq v < 1, \\ 1 & \text{if } v \geq 1. \end{cases}$$

For $v < 0$ or $v \geq 1$, $F_n(v) = F(v)$ and for $0 < v < 1$,

$$F_n(v) = 1 - \frac{1}{n} \to 1 = F(v).$$

In this case, we even get convergence at $v = 0$ which we do not need. □

EXAMPLE B.2.2. Everything is the same except we redefine y_n as

$$y_n(\omega) = \sqrt{n}\mathscr{I}_{[0,1/n]}(\omega).$$

The arguments for almost sure convergence, convergence in probability, and convergence in distribution continue to hold with very little change but this random variable does not converge to 0 in mean square. To get convergence in mean square we need the limit of $\int [y_n - y]^2 \, dP$ go to 0. Since $y = 0$ a.s.,

$$\int [y_n - y]^2 \, dP = \int [\sqrt{n}\mathscr{I}_{[0,1/n]}(\omega)]^2 \, d\omega$$
$$= n \int \mathscr{I}_{[0,1/n]}(\omega) \, d\omega = n \int_0^{1/n} d\omega = \frac{n}{n} \not\to 0. \quad \square$$

Exercise B.1. Show convergence a.s, in probability, and in distribution for Example B.2.2.

It is a little tricker to get something that converges in mean square but does not converge almost surely.

EXAMPLE B.2.3. Everything is the same except we redefine y_n. Let $h \to \infty$. For every h we are going to construct a collection of h different 0-1 random variables, y_{hi}, $i = 1, \ldots, h$, that change which ω values generate a 1 and which generate a 0,

$$y_{hi}(\omega) = \mathscr{I}_{[(i-1)/h, i/h]}(\omega).$$

To turn this into a single sequence of random variables, define $n = (h-1)h/2 + i$ and $y_n = y_{hi}$. The arguments for convergence in quadratic mean, probability, and distribution are minor modifications of Example B.2.1 but almost sure convergence now fails. For any ω, y_n has been constructed so that $y_n(\omega)$ is 0 infinitely often but is also 1 infinitely often, so it does not converge to anything. Since this occurs for every ω, y_n fails to converge on a set of ω values that has probability 1, as opposed to almost sure convergence which requires convergence on a set of ωs that have probability 1.

If we modify the example so that

$$y_{hi}(\omega) = \sqrt{h}\mathscr{I}_{[(i-1)/h, i/h]}(\omega)$$

we get a y_n that does not converge either almost surely or in quadratic mean but does converge in probability and in distribution. □

Exercise B.2. Establish convergence (or its lack) in quadratic mean, in probability, and in distribution for the sequences in Example B.2.3.

Exercise B.3. For the probability space in the Examples define $y_n(\omega)$ to be $(n-1)/n$ if ω is irrational and 0 if it is rational. To what and how does y_n converge?

It turns out that if y_n converges to y in distribution *and* y is constant with probability 1, then y_n converges to y in probability. So to produce an example of a sequence that converges in distribution but does not converge in probability, we need an example with a more sophisticated target than $y = 0$.

EXAMPLE B.2.4. Let $x \sim N(0,1)$. Define a sequence of random variables by $x = y = y_1 = y_3 = y_5 = y_7 = \cdots$ and $-x = y_2 = y_4 = y_6 = \cdots$. Quite clearly, $y_n \xrightarrow{\mathscr{L}} y$ because all of the distributions involved are identical. However, for n even,

$$P[|y_n - y| > \varepsilon] = P[|-2x| > \varepsilon]$$

which is the probability that a $N(0,4)$ is farther from 0 than ε. That number is certainly not getting close to 0, and the smaller ε gets the closer it is to 1. Clearly, $y_n \xrightarrow{P} y$.

We now give the general definitions of convergence for random vectors but first recall that the length of a vector is $\|y\| \equiv \sqrt{y'y}$.

Definition B.2.5. Consider a sequence of random d vectors y_n and a random vector y with respective cdfs F_n and F.

1. y_n converges to y in distribution (law), written $y_n \xrightarrow{\mathscr{L}} y$, if
$$F_n(v) \to F(v)$$
at every continuity point of F.
2. y_n converges to y in probability, written $y_n \xrightarrow{P} y$, if for any $\varepsilon > 0$,
$$P[\,\|y_n - y\| > \varepsilon\,] \to 0.$$
3. y_n converges to y almost surely, written $y_n \xrightarrow{a.s.} y$, if
$$P\left[\,\{\omega | \lim_{n\to\infty} \|y_n(\omega) - y(\omega)\| = 0\}\,\right] = 1.$$
4. y_n converges to y in quadratic mean, written $y_n \xrightarrow{q.m.} y$, if
$$\mathrm{E}\left[\,\|y_n - y\|^2\,\right] \to 0.$$

Note that all of these definitions involve checking whether certain sequences of numbers converge and all but number 3 reduce merely to checking convergence of numbers (as opposed to vectors). Moreover, from these definitions, the fact that $y_n \to y$ if and only if $y_n - y \to 0$ follows immediately for any form of convergence except distribution (law). In Example B.2.4 the even numbered y_ns converge in law to y but the distribution of $y_{2k} - y = -2x$ is not (nor does it approach) the 0 random variable.

When discussing convergence in distribution, say $y_n \xrightarrow{\mathscr{L}} y$, if we know the distribution of y, for example if y is multivariate normal with mean μ and covariance matrix Σ, we should write that information as $y_n \xrightarrow{\mathscr{L}} y \sim N(\mu, \Sigma)$. As a shorthand, we sometimes write $y_n \xrightarrow{\mathscr{L}} N(\mu, \Sigma)$ but when being careful, we need to establish the distribution of the limiting random vector y. Normal (Gaussian) distributions and χ^2 distributions are the two that most commonly appear in our limiting arguments.

It will also be useful to discuss convergence of random matrices.

Definition B.2.6. A sequence of random matrices W_n converges to a random

matrix W (possibly degenerate at some fixed value W_0) if and only if $\text{Vec}(W_n)$ converges to $\text{Vec}(W)$. [The Vec operator simply stacks the columns of an $r \times c$ matrix into an $rc \times 1$ vector.]

B.2.1 Characteristic Functions Are Not Magical

How can we tell when two probability measures P_1 and P_2, both defined on $(\Omega, \mathscr{F})$, are the same? Obviously they are the same if they give the same probability for every set in $\mathscr{F}$. In Appendix A we claimed that if two random variables had the same characteristic function, they had the same distribution. A *separating class* is a class of functions $\mathscr{S}$ such that if

$$\int f(\omega)\, dP_1(\omega) = \int f(\omega)\, dP_2(\omega)$$

for every $f \in \mathscr{S}$, it implies that $P_1 = P_2$. One separating class is pretty obviously all the indicator functions for sets in $\mathscr{F}$. It turns out that the class of all bounded continuous functions is a separating class. In fact, the class of functions e^{itx} for t in an interval around 0 is a separating class, which is why any two random variables having the same characteristic function must have the same distribution

It turns out (cf. Billingsley, 1999) that if

$$\text{E}[f(y_n)] \to \text{E}[f(y)] \qquad \text{for every } f \in \mathscr{S},$$

then

$$y_n \xrightarrow{\mathscr{L}} y.$$

In particular, if the characteristic functions of the y_ns converge for all t in an interval around 0 to the characteristic function of y, then $y_n \xrightarrow{\mathscr{L}} y$.

What if we do not know y so that all we know is that $\text{E}[f(y_n)]$ converges to something for every f? Consider a sequence of probability distributions $P_1, P_2, \ldots$ defined by a sequence of random variables $y_1, y_2, \ldots$. The sequence is said to be *tight* if for any $\varepsilon > 0$ there exists a closed bounded set B_ε such that $P_n(B_\varepsilon) \geq 1 - \varepsilon$ for every n. The sequence of distributions is tight in the sense that it does not have a lot of probability going off towards infinity. It turns out that if the sequence is tight and if $\text{E}[f(y_n)]$ converges to something for every f in a separating class, then there exists y such that $y_n \xrightarrow{\mathscr{L}} y$. In particular, if the characteristic functions of the y_ns converge for all t in an interval around 0 to some function $\varphi(t)$ that is continuous on a ball around 0, it turns out that the sequence has to be tight, so there exists a y with $y_n \xrightarrow{\mathscr{L}} y$ (and $\varphi(t)$ is the characteristic function of y).

One of my oldest (and therefore, *not necessarily accurate*) memories in statistics is hearing Joe (Morris L.) Eaton say that, if you need to use characteristic functions to prove something, you obviously don't know what you are doing. (Not withstanding, every edition of *PA* has used characteristic functions to prove that multivariate

normal distributions are uniquely defined by their mean vector and covariance matrix.)

B.2.2 Measure Theory Convergence Theorems

To use separating classes to show convergence in law, we need to be able to show that the expected values (integrals) of a sequence of (measurable) functions converge. We now present two of measure theory's most useful results along that line.

The Dominated Convergence Theorem. Let $(\Omega, \mathscr{F}, \mu)$ be a measure space and let f_n be a sequence of real valued functions that converge to f almost everywhere. If there exists a nonnegative measurable function g such that $\int g(\omega)\, d\mu(\omega) < \infty$, and if for all n, $|f_n(\omega)| \leq g(\omega)$ except on a set of measure 0, then

$$\int f_n(\omega)\, d\mu(\omega) \to \int f(\omega)\, d\mu(\omega).$$

The same result stated for (measurable) random variables is

The Probability Dominated Convergence Theorem. Let $(\Omega, \mathscr{F}, P)$ be a probability space and let y_n be a sequence of random variables that converge to y a.s. If there exists a nonnegative random variable x such that $\mathrm{E}(x) < \infty$, and if for all n, $|y_n(\omega)| \leq x(\omega)$ almost surely, then

$$\mathrm{E}(y_n) \to E(y).$$

The Monotone Convergence Theorem. Let $(\Omega, \mathscr{F}, \mu)$ be a measure space and let $f_n \geq 0$ be a sequence of real valued measurable functions that converge to f almost everywhere with $0 \leq f_n(\omega) \leq f_{n+1}(\omega)$ a.e., then

$$\int f_n(\omega)\, d\mu(\omega) \to \int f(\omega)\, d\mu(\omega).$$

B.2.3 Ferguson's Version of Slutsky's Theorem

Slutsky's Theorem has always been a very useful result in probability and statistics. The version presented in Ferguson (1996) is immensely useful.

Let $\mathscr{C}(f)$ be the set of all points at which the function f is continuous and let c be a fixed vector. Ferguson's (1996) random vector version of Slutsky's Theorem is

1. If $X_n \xrightarrow{\mathscr{L}} X$ and f has $\Pr[X \in \mathscr{C}(f)] = 1$, then $f(X_n) \xrightarrow{\mathscr{L}} f(X)$.
2. If $X_n \xrightarrow{\mathscr{L}} X$ and $X_n - Y_n \xrightarrow{P} 0$, then $Y_n \xrightarrow{\mathscr{L}} X$.
3. If $X_n \xrightarrow{\mathscr{L}} X$ and $Y_n \xrightarrow{P} c$, then

$$\begin{bmatrix} X_n \\ Y_n \end{bmatrix} \xrightarrow{\mathscr{L}} \begin{bmatrix} X \\ c \end{bmatrix}.$$

Ferguson also shows that $Y_n \xrightarrow{P} c$ if and only if $Y_n \xrightarrow{\mathscr{L}} c$. These three results also hold if convergence in law is replaced everywhere by convergence in probability but in part (3) convergence in probability still holds if Y_n converges to any random vector Y rather than a mere constant c. This stronger version of the theorem also holds for almost sure convergence.

A more traditional statement of Slutsky's Theorem for random variables is that if $X_n \xrightarrow{\mathscr{L}} X$ and $Y_n \xrightarrow{P} c$, then $X_n + Y_n \xrightarrow{\mathscr{L}} X + c$, $X_n Y_n \xrightarrow{\mathscr{L}} Xc$, and, for $c \neq 0$, $X_n/Y_n \xrightarrow{\mathscr{L}} X/c$. These results follow immediately from Ferguson's version.

Consider random matrices with $W_n \xrightarrow{P} W$. Based on Ferguson's Slutsky and our definition of matrix convergence, if the matrices are all nonsingular with probability 1, then by continuity, $W_n^{-1} \xrightarrow{P} W^{-1}$. Moreover, the eigenvalues will also converge in probability and at least one version of the eigenvectors will converge in probability. If all the matrices are positive definite, this allows us to find versions $W_n^{1/2} \xrightarrow{P} W^{1/2}$ and $W_n^{-1/2} \xrightarrow{P} W^{-1/2}$.

We now illustrate just how useful the Ferguson's Slutsky result is for manipulating random vectors.

EXAMPLE B.2.7. *Standardized statistics.*
Suppose for d vectors with a nonsingular covariance matrices

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathscr{L}} Y \sim N\big[0, \Sigma(\theta)\big].$$

This implies that $\hat{\theta}_n \xrightarrow{P} \theta$ but we need Ferguson's Slutsky to show it. In particular, $1/\sqrt{n} \to 0$, so degenerately $1/\sqrt{n} \xrightarrow{P} 0$, thus by Ferguson's Slutsky (3)

$$\begin{bmatrix} \sqrt{n}(\hat{\theta}_n - \theta) \\ 1/\sqrt{n} \end{bmatrix} \xrightarrow{\mathscr{L}} \begin{bmatrix} Y \\ 0 \end{bmatrix}$$

and since multiplication is a continuous function, by Ferguson's Slutsky (1)

$$(\hat{\theta}_n - \theta) = (1/\sqrt{n})\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathscr{L}} (0)Y = 0.$$

But as Ferguson shows, $(\hat{\theta}_n - \theta) \xrightarrow{\mathscr{L}} 0$ is equivalent to $(\hat{\theta}_n - \theta) \xrightarrow{P} 0$ and by definition $\hat{\theta}_n \xrightarrow{P} \theta$.

What we really want to show is that

$$n(\hat{\theta}_n - \theta)' \left[\Sigma(\hat{\theta})\right]^{-1} (\hat{\theta}_n - \theta) \xrightarrow{\mathscr{L}} \chi^2(d).$$

More generally, we would like to show that if $\hat{\Sigma} \xrightarrow{P} \Sigma(\theta)$, which is to say that if

$$\text{Vec}(\hat{\Sigma}) \xrightarrow{P} \text{Vec}[\Sigma(\theta)],$$

then

$$n(\hat{\theta}_n - \theta)'\hat{\Sigma}^{-1}(\hat{\theta}_n - \theta) \xrightarrow{\mathscr{L}} \chi^2(d).$$

We focus on this second result. Of course if $\Sigma(\theta)$ is continuous, Ferguson's Slutsky immediately gives that

$$\Sigma(\hat{\theta}) \xrightarrow{P} \Sigma(\theta).$$

(Actually this requires either the convergence in probability version of Slutsky or the equivalence of convergence in law and probability when converging to a constant.) By our assumptions and part (3) of Ferguson's Slutsky,

$$\begin{bmatrix} \sqrt{n}(\hat{\theta}_n - \theta) \\ \text{Vec}(\hat{\Sigma}) \end{bmatrix} \xrightarrow{\mathscr{L}} \begin{bmatrix} Y \\ \text{Vec}[\Sigma(\theta)] \end{bmatrix} \equiv Q.$$

Relying on our earlier comments about continuity of matrix inverses, a continuous function of the vector Q is $f(Q) \equiv Y'[\Sigma(\theta)]^{-1}Y$, so

$$f\left(\begin{bmatrix} \sqrt{n}(\hat{\theta}_n - \theta) \\ \text{Vec}(\hat{\Sigma}) \end{bmatrix}\right) = \sqrt{n}(\hat{\theta}_n - \theta)' \left[\hat{\Sigma}\right]^{-1} \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathscr{L}} Y'[\Sigma(\theta)]^{-1}Y.$$

It is a well-known fact from linear model theory (cf. Christensen, 2020, Section 1.3) that if a d vector has $Y \sim N\big[0, \Sigma(\theta)\big]$, then

$$Y'[\Sigma(\theta)]^{-1}Y \sim \chi^2(d).$$

Similarly, if $d = 1$ so that $\hat{\Sigma} \equiv \hat{\sigma}^2$ and $\Sigma(\theta) \equiv \sigma_{11}(\theta)$, then we get

$$\begin{bmatrix} \sqrt{n}(\hat{\theta}_n - \theta) \\ \hat{\sigma}^2 \end{bmatrix} \xrightarrow{\mathscr{L}} \begin{bmatrix} Y \\ \sigma_{11}(\theta) \end{bmatrix}.$$

By taking continuous functions,

$$\frac{\hat{\theta}_n - \theta}{\sqrt{\hat{\sigma}^2/n}} = \sqrt{\frac{\sigma_{11}(\theta)}{\hat{\sigma}^2}} \frac{1}{\sqrt{\sigma_{11}(\theta)}} \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathscr{L}} (1) \left(\frac{1}{\sqrt{\sigma_{11}(\theta)}}\right) Y \sim N(0,1).$$

Exercise B.4. Show that for any fixed d vector λ, that

$$\frac{\lambda'\hat{\theta}_n - \lambda'\theta}{\sqrt{\lambda'\hat{\Sigma}\lambda/n}} \xrightarrow{\mathscr{L}} N(0,1).$$

B.2.4 The Law of Large Numbers

The Strong Law of Large Numbers (LLN) states that if random vectors $x_1, \ldots, x_n$ are iid with $\mathrm{E}(x_i) = \mu$, then the sample mean $\bar{x}_\cdot$ has the property that

$$\bar{x}_\cdot \xrightarrow{a.s.} \mu.$$

For random variables, if you are willing to remove the identically distributed condition but willing to require the existence of variances that are bounded, you get the Weak Law of Large Numbers that

$$\bar{x}_\cdot \xrightarrow{P} \mu.$$

Of course convergence in probability also holds for iid random vectors.

EXAMPLE B.2.8. *Convergence of the sample covariance matrix.*
If $x_1, \ldots, x_n$ are iid with $\mathrm{E}(x_i) = \mu$ and $\mathrm{Cov}(x_i) = \Sigma$, the sample covariance matrix is defined as

$$\begin{aligned}
S_{xx} &\equiv \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_\cdot)(x_i - \bar{x}_\cdot)' \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i x_i' - \sum_{i=1}^{n} x_i \bar{x}_\cdot' - \sum_{i=1}^{n} \bar{x}_\cdot x_i' + n\bar{x}_\cdot \bar{x}_\cdot'\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i x_i' - n\bar{x}_\cdot \bar{x}_\cdot' - n\bar{x}_\cdot \bar{x}_\cdot' + n\bar{x}_\cdot \bar{x}_\cdot'\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i x_i' - n\bar{x}_\cdot \bar{x}_\cdot'\right] \\
&= \left(\frac{n}{n-1}\right)\frac{1}{n}\sum_{i=1}^{n} x_i x_i' - \frac{n}{n-1}\bar{x}_\cdot \bar{x}_\cdot'.
\end{aligned}$$

By the LLN

$$\bar{x}_\cdot \xrightarrow{P} \mu$$

so by Ferguson's Slutsky part (1),

$$\mathrm{Vec}(\bar{x}_\cdot \bar{x}_\cdot') \xrightarrow{P} \mathrm{Vec}(\mu\mu')$$

or

$$\bar{x}_\cdot \bar{x}_\cdot' \xrightarrow{P} \mu\mu'.$$

By our assumptions, $\mathrm{E}(x_i x_i') = \Sigma + \mu\mu'$, so again by the LLN,

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{Vec}(x_i x_i') \xrightarrow{P} \mathrm{Vec}(\Sigma + \mu\mu')$$

or

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i' \xrightarrow{P} \Sigma + \mu\mu'.$$

From these pieces we can apply Slutsky part (3) to see

$$\begin{bmatrix} \bar{x}. \\ \frac{1}{n}\sum_{i=1}^{n} \mathrm{Vec}(x_i x_i') \\ \frac{n}{n-1} \end{bmatrix} \xrightarrow{P} \begin{bmatrix} \mu \\ \mathrm{Vec}(\Sigma + \mu\mu') \\ 1 \end{bmatrix},$$

so by continuity and Slutsky part (1) we get

$$S_{xx} = \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n} x_i x_i' - \frac{n}{n-1}\bar{x}.\bar{x}.' \xrightarrow{P} (1)(\Sigma + \mu\mu') - (1)\mu\mu' = \Sigma.$$

The use of $n-1$ in the definition of $S_{xx} \equiv \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}.)(x_i - \bar{x}.)'$ is to get $\mathrm{E}(S_{xx}) = \Sigma$, it has little to do with convergence.

B.2.5 The Central Limit Theorem

The Central Limit Theorem states that if $x_1, \ldots, x_n$ are iid random variables with $\mathrm{E}(x_i) = \mu$ and $\mathrm{Var}(x_i) = \sigma^2$, then the sample mean $\bar{x}.$ has the property that

$$\frac{(\bar{x}. - \mu)}{\sqrt{\sigma^2/n}} \xrightarrow{\mathscr{L}} N(0,1).$$

Alternatively, $\sqrt{n}(\bar{x}. - \mu) \xrightarrow{\mathscr{L}} N(0, \sigma^2)$. A common way to prove this is to take a second-order Taylor's expansion of the characteristic function of $(\bar{x}. - \mu)/\sqrt{\sigma^2/n}$ and show that it converges to the characteristic function of a standard normal. We will not be doing that. We present without proof a more general result.

Lindeberg's Central Limit Theorem For each n, let y_{ni}, $i = 1, \ldots, n$, be independent with mean 0 and variance σ_{ni}^2. Let $z_n \equiv \sum_{i=1}^{n} y_{ni}$ and $B_n^2 \equiv \mathrm{Var}(z_n) = \sum_{i=1}^{n} \sigma_{ni}^2$. If for any $\varepsilon > 0$,

$$0 = \lim_{n\to\infty} \frac{1}{B_n^2}\sum_{i=1}^{n} \mathrm{E}\left[|y_{ni}|^2 \mathscr{I}_{[\varepsilon B_n, \infty)}(|y_{ni}|)\right], \tag{1}$$

then

$$\frac{z_n}{B_n} \xrightarrow{\mathscr{L}} N(0,1).$$

Lindeberg's result implies the usual Central Limit Theorem for iid random variables.

EXAMPLE B.2.9. Take $x_1,\ldots,x_n$ iid with $\mathrm{E}(x_i)=\mu$ and $\mathrm{Var}(x_i)=\sigma^2$. Set $y_{ni}\equiv x_i-\mu$, so $z_n=\sum_{i=1}^n(x_i-\mu)$ and $B_n^2=n\sigma^2$. If the Lindeberg condition holds

$$\frac{z_n}{B_n}=\frac{1/n}{1/n}\frac{\sum_{i=1}^n(x_i-\mu)}{\sqrt{n\sigma^2}}=\frac{(\bar{x}.-\mu)}{\sqrt{\sigma^2/n}}\xrightarrow{\mathscr{L}}N(0,1).$$

It remains to show that the Lindeberg condition (1) holds.

In this example the Lindeberg condition reduces to

$$\lim_{n\to\infty}\frac{1}{n\sigma^2}n\mathrm{E}\left[|x_i-\mu|^2\mathscr{I}_{[\varepsilon\sqrt{n}\sigma,\infty)}(|x_i-\mu|)\right]=0$$

so it suffices to show that

$$\mathrm{E}\left[|x_i-\mu|^2\mathscr{I}_{[\varepsilon\sqrt{n}\sigma,\infty)}(|x_i-\mu|)\right]\to 0.$$

However, $\lim_{a_n\to\infty}\mathscr{I}_{[a_n,\infty)}(u)=0$ and $\mathrm{E}[|x_i-\mu|^2]=\sigma^2$, so by probability dominated convergence with $|x_i-\mu|^2$ as the dominating function and the sequence $|x_i-\mu|^2\mathscr{I}_{[\varepsilon\sqrt{n}\sigma,\infty)}(|x_i-\mu|)$ converging to 0 as $n\to\infty$ a.s.,

$$\mathrm{E}\left[|x_i-\mu|^2\mathscr{I}_{[\varepsilon\sqrt{n}\sigma,\infty)}(|x_i-\mu|)\right]\to\mathrm{E}[0]=0.$$

EXAMPLE B.2.10. The x_ns are independent with $\Pr[x_n=\pm\sqrt{n-1}]=0.25$ and $\Pr[x_n=\pm 1]=0.25$. Note that x_n is an example of a sequence of random variables that is not tight. Define $y_{ni}\equiv x_i$ and observe that $\mathrm{Var}(x_i)=i/2$, so $B_n^2=\sum_{i=1}^n i/2=n(n+1)/4$. If the Lindeberg condition holds

$$\frac{\sum_{i=1}^n x_i}{\sqrt{n(n+1)/4}}=\frac{\bar{x}.}{\sqrt{1/4}}\frac{n}{\sqrt{n(n+1)}}\xrightarrow{\mathscr{L}}N(0,1).$$

Since $n/\sqrt{n(n+1)}\to 1$, according to Slutsky's theorem,

$$\frac{\bar{x}.}{\sqrt{1/4}}\xrightarrow{\mathscr{L}}N(0,1),$$

or $\bar{x}.\xrightarrow{\mathscr{L}}N(0,0.25)$.

It remains to show that the Lindeberg condition (1) holds. In this example the Lindeberg condition reduces to

$$\lim_{n\to\infty}\frac{1}{n(n+1)/4}\sum_{i=1}^n\mathrm{E}\left[|x_i|^2\mathscr{I}_{[\varepsilon\sqrt{n(n+1)/4},\infty)}(|x_i|)\right]=0.$$

However, for any $\varepsilon>0$, there exists an n large enough that

$$\sqrt{n-1}<\varepsilon\sqrt{n(n+1)/4}.$$

Once that happens, for all $i = 1, \ldots, n$

$$0 = \Pr\left[|x_i| \geq \varepsilon\sqrt{n(n+1)/4}\right] = \mathrm{E}\left[\mathscr{I}_{[\varepsilon\sqrt{n(n+1)/4},\infty)}(|x_i|)\right]$$

and, since the indicator functions are 0 a.s.,

$$0 = \frac{1}{n(n+1)/4}\sum_{i=1}^{n}\mathrm{E}\left[|x_i|^2\mathscr{I}_{[\varepsilon\sqrt{n(n+1)/4},\infty)}(|x_i|)\right].$$

There is a concept called entropy that measures the randomness in a random variable. It turns out that among distributions with a common expected value and variance, the normal distribution has the largest entropy. In the Central Limit Theorem we have fixed expected values and variances for the individual random variables, so the sample mean also has a fixed expected value and variance. The Central Limit Theorem is basically saying that the sample mean converges in distribution to the most random thing possible that retains the correct expected value and variance, cf. Barron (1986).

Finally we mention the vector version of the central limit theorem. If $x_1, \ldots, x_n$ are iid d vectors with with $\mathrm{E}(x_i) = \mu$ and $\mathrm{Var}(x_i) = \Sigma$, then the sample mean $\bar{x}_\cdot$ has the property that

$$\sqrt{n}(\bar{x}_\cdot - \mu) \xrightarrow{\mathscr{L}} N(0, \Sigma),$$

cf. Ferguson (1996)

B.2.6 The Delta Method

Let $\hat{\theta}_n$ be a sequence of $d \times 1$ random vectors and suppose that

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathscr{L}} N\big[0, \Sigma(\theta)\big].$$

Suppose that $g(\cdot)$ is a differentiable function taking d vectors into r vectors. Let $\mathbf{d}g$ denote the $r \times d$ matrix of partial derivatives of g evaluated at θ. Then

$$\sqrt{n}\big[g(\hat{\theta}_n) - g(\theta)\big] \xrightarrow{\mathscr{L}} N\big[0, \mathbf{d}g\,\Sigma(\theta)\,\mathbf{d}g'\big].$$

For technical reasons, it is advantageous to assume that $\mathbf{d}g$ and $\Sigma(\theta)$ are also continuous. For mathematical details, see Bishop, Fienberg, and Holland (1975, Section 14.6.3) or Ferguson (1996, Chapter 7). Somewhat surprisingly, this is not a direct consequence of Slutsky. Its proof involves a first order Taylor approximation.

Appendix C
Conditional Probability and Radon-Nikodym

The Radon-Nikodym Theorem basically tells us that for any absolutely continuous distribution, a density exists, and for any discrete distribution, a probability mass function (discrete density) exists. Typically, we define distributions in terms of their densities, so this is somewhat circular. A more important application of Radon-Nikodym is to defining conditional probabilities and expectations.

C.1 The Radon-Nikodym Theorem

Before stating the theorem we need a few additional ideas. A σ-finite measure μ may have $\mu(\Omega) = \infty$ but has sets $A_1, A_2, \ldots$ with $\mu(A_i) < \infty$ and $\Omega = \cup_{i=1}^{\infty} A_i$. Lebesgue measure is σ-finite. A signed measure λ simply allows negative measures. It must display countable additivity for sets of finite measure and $\lambda(\emptyset) = 0$. The set $\bar{\mathbf{R}}$ includes $\pm\infty$

Radon-Nikodym Theorem. (a) Let μ be a σ-finite measure and (b) λ a signed measure on the σ-field $\mathscr{F}$ of subsets of Ω. (c) Assume that λ is absolutely continuous with respect to μ. Then there is a Borel measurable function $g : \omega \to \bar{\mathbf{R}}$ such that

$$\lambda(A) = \int_A g d\mu \qquad \text{for all } A \in \mathscr{F}$$

If h is another such function, then $g = h$ a.e. $[\mu]$.

PROOF: See Ash and Doleans-Dade (2000). □

In the notation of Appendix A, (a) let μ be Lebesgue or counting product measure on $\mathbf{R}^m$ or some mixture of the two. (See Appendix D.3.) (b) Let x be a random vector on $\mathbf{R}^m$ that defines probabilities $P(x \in A)$ for Borel sets $A \subset \mathbf{R}^m$. (c) Assume that (i) if μ is product Lebesgue measure, x is continuous; (ii) if μ is product counting measure, x is discrete; (iii) if μ is a product measure parts of which are Lebesgue and

parts counting, x is an appropriate combination of continuous and discrete variables. Then there exists a (measurable) density function $f(u)$ such that

$$P(x \in A) = \int_A f(u) d\mu(u) = \begin{cases} \int_A f(u) du & \text{product Lesbegue} \\ \sum_{u \in A} f(u) & \text{product counting.} \end{cases}$$

If h is another such function, $h(u) = f(u)$ except on a set of product measure 0.

For probability mass functions like the binomial and Poisson, counting measure, say C, puts measure 1 on every natural number with everything else having collective measure 0. (You can put counting measure on the integers or natural numbers with the σ-field being all sets of integers (natural numbers) or you can put it on $(\mathbf{R}, \mathscr{B})$.) Probability mass functions that exist on the natural numbers are densities relative to counting measure. Integration with counting measure reduces to summation.

Probabilities are unitless. They are not measured in inches or centimeters or ergs. Counting measure is similarly unitless. But Lebesgue measure has units. Length is measured in meters or miles or some such thing. If we have a Lebesgue density for y, then $P_y(A) = \int \mathscr{I}_A(v) f_y(v)\, d\mu(v)$. The left side is unitless but $d\mu$ is in, say, meters. To get the meters to cancel out, the density has to have units that are $1/$meters. Because of this feature, sometimes using densities in statistical inference can get dicey and Cox (2006) rightly calls of the name "density" for a probability mass function "deplorable," but, like us, still uses it.

Another issue that there is no compelling reason why continuous densities should be defined relative to Lebesgue measure. They could just as well be defined relative to the $N(0,1)$ probability distribution.

Exercise C.1. Show that the density of a $U[0,1]$ relative to the $N(0,1)$ is just $\mathscr{I}_{[0,1]}(v)$ divided by the usual standard normal density. Why can you not find the density of a $N(0,1)$ relative to a $U[0,1]$?

C.2 Conditional Probability

Everyone agrees that the conditional probability of $x \in A$ given $y \in B$ is

$$P(x \in A | y \in B) \equiv \frac{P(x \in A, y \in B)}{P(y \in B)},$$

when $P(y \in B) > 0$. The problem is how to define conditional probability when $P(y \in B) = 0$. Specifically, we want to develop the ideas of $P(x \in A | y = v)$ when $P(y = v) = 0$, and $E(x|y = v)$, or more generally just $P(x \in A|y)$ and $E(x|y)$ as functions of y.

The key idea is to define $P(x \in A|y)$ in such a way that it is a function of y alone and that for any allowable B,

$$P(x \in A, y \in B) = \int_{y \in B} P(x \in A|y)\, dP.$$

More technically this is

$$P[\{\, \omega \,|\, x(\omega) \in A, y(\omega) \in B \,\}] = \int_{\{\, \omega \,|\, y(\omega) \in B \,\}} P[x \in A|y(\omega)]\, dP(\omega), \tag{1}$$

where in $P[x \in A|y(\omega)]$ the symbols $x \in A$ are only part of a name and do not actually depend on ω.

The vector $(x', y')'$ is mapping $(\Omega, \mathscr{F})$ into $(\mathbf{R}^{m+n}, \mathscr{B}^{m+n})$. ($\mathscr{B}^{m+n}$ can be thought of as the smallest σ-field generated by products of sets in $\mathscr{B}^m$ and sets in $\mathscr{B}^n$ hence also denoted $\mathscr{B}^m \times \mathscr{B}^n$.) For fixed A, $P(x \in A, y \in B)$ defines a new measure on $\mathbf{R}^n$ for $B \in \mathscr{B}^n$, typically not a probability measure, yet Radon-Nikodym assures us that some function of $y(\omega)$ exists that characterizes the new measure. We call this function $P(x \in A|y)$.

But we also want to avoid having to think about the ωs and just think about the random vectors. We can also write the definition of conditional probability using

$$P(x \in A, y \in B) \equiv P_{xy}(A \times B) = \int_{\mathbf{R}^m \times B} P(x \in A|y = v)\, dP_{xy}(u, v)$$

where the conditional probability $P(x \in A|y = v)$ has to be a function of v alone. If P_{xy} is absolutely continuous wrt some $m+n$ dimensional product measure (say Lebesgue, counting, or any combinations), $P(x \in A|y = v)$ is characterized by the following:

$$\begin{aligned}
P(x \in A, y \in B) &\equiv P_{xy}(A \times B) \\
&= \int_{\mathbf{R}^m \times B} P(x \in A|y = v) f_{xy}(u, v)\, d[\mu_x \times \mu_y](u, v) \\
&= \int_B \left[\int_{\mathbf{R}^m} P(x \in A|y = v) f_{xy}(u, v)\, d\mu_x(u) \right] d\mu_y(v) \\
&= \int_B P(x \in A|y = v) \left[\int_{\mathbf{R}^m} f_{xy}(u, v)\, d\mu_x(u) \right] d\mu_y(v) \\
&= \int_B P(x \in A|y = v) f_y(v)\, d\mu_y(v) \\
&= \int_B P(x \in A|y = v)\, dP_y(v)
\end{aligned}$$

where we perform iterated integration using the Fubini-Tonelli Theorem. We know from the definition of the joint density that

$$\begin{aligned}
P(x \in A, y \in B) &= \int_{A \times B} f_{xy}(u, v)\, d[\mu_x \times \mu_y](u, v) \\
&= \int_B \left[\int_A f_{xy}(u, v)\, d\mu_x(u) \right] d\mu_y(v)
\end{aligned}$$

$$= \int_B \left[\int_A f_{xy}(u,v)/f_y(v)\, d\mu_x(u) \right] f_y(v)\, d\mu_y(v)$$
$$= \int_B \left[\int_A f_{xy}(u,v)/f_y(v)\, d\mu_x(u) \right] dP_y(v),$$

so

$$P(x \in A | y = v) \equiv \int_A f_{xy}(u,v)/f_y(v)\, d\mu_x(u), \quad \text{a.s. } \mu_y,$$

is a function of y (it tells us what the function is for every $y = v$) such that when integrated over $y \in B$ gives $P(x \in A, y \in B)$ for any $B \in \mathscr{B}^n$. Radon-Nikodym tells us that any other such function must equal this one a.s. Note that $f_{xy}(u,v)/f_y(v)$ is undefined for $f_y(v) = 0$, but as a function of y, $f_y(v) = 0$ on a set of y probability 0, so we can define the ratio any way we desire on this set.

There is a slight catch. If we want to think about $P(x \in A | y = v)$ defining a conditional distribution on x given $y = v$, we need to think about v being fixed and varying the sets $A \in \mathscr{B}^m$. Although $P(x \in A | y = v)$ is unique up to sets of y probability 0, by changing A an uncountably infinite number of times, the uncountable accumulation of sets of y probability 0 might cause a problem. Fortunately, it can be shown that there is a version of $P(x \in A | y = v)$ that works fine. In fact, when we can do the iterated integrals, $P(x \in A | y = v) \equiv \int_A f_{xy}(u,v)/f_y(v)\, d\mu_x(u)$ is such a version because we can define the conditional probabilities as the result of the integral. In particular, $P(x \in A | y = v)$ admits a density wrt $\mu_x(u)$ which is

$$f_{x|y}(u|v) \equiv f_{xy}(u,v)/f_y(v).$$

Now we extend these ideas to conditional probabilities that are not product sets $A \times B$. For $D \in \mathscr{B}^m \times \mathscr{B}^n = \mathscr{B}^{m+n}$, to define $P[(x,y) \in D | y = v]$ we require, for every B in $\mathscr{B}^n$,

$$\begin{aligned} P[(x,y) \in D, y \in B] &= P\{(x,y) \in D, (x,y) \in [\mathbf{R}^m \times B]\} \\ &= \int_{\mathbf{R}^m \times B} P[(x,y) \in D | y = v]\, dP_{xy}(u.v) \qquad (2) \\ &= \int_B P[(x,y) \in D | y = v]\, dP_y(v). \end{aligned}$$

We need to define $D(v) = \{ u | (u,v) \in D \}$ so that

$$P[(x,y) \in D, y \in B] = \int_B \left[\int_{D(v)} f_{xy}(u,v)/f_y(v)\, d\mu_x(u) \right] f_y(v) d\mu_y(v),$$

which provides us with our working form,

$$P[(x,y) \in D | y = v] = \int_{D(v)} f_{xy}(u,v)/f_y(v)\, d\mu_x(u).$$

Measure theoretic discussions of conditional probability often refer to a probability space $(\Omega, \mathscr{F}, P)$ and a sub-σ-field $\mathscr{F}_0 \subset \mathscr{F}$. The conditional probability

$P(D|\mathscr{F}_0)$ is defined as a function, measurable wrt $\mathscr{F}_0$, satisfying, for every $B \in \mathscr{F}_0$,

$$P(D \cap B) = \int_B P(D|\mathscr{F}_0)\, dP.$$

The idea is that conditioning on $\mathscr{F}_0$ is conditioning on knowing whether every set in $\mathscr{F}_0$ either occurred or did not occur. This is pretty much equation (2) with the understandings that $\mathscr{F}_0 = \mathbf{R}^m \times \mathscr{B}^n$ and that knowing $y = v$ is equivalent to knowing whether every set B in $\mathscr{B}^n$ occurred, or equivalently whether every set $[\mathbf{R}^m \times B] \in \mathbf{R}^m \times \mathscr{B}^n$, occurred.

Finally, we examine conditional expected values. For a real valued measurable function of y, say, $g : \mathbf{R}^n \to \mathbf{R}$ and a measurable vector function $T : \mathbf{R}^n \to \mathbf{R}^d$. Denote the measure theoretic conditional expectation of $g(y)$ given $T(y)$ variously as,

$$\mathrm{E}_{y|T(y)}[g(y)] \equiv \mathrm{E}_y[g(y)|T(y)] \equiv \mathrm{E}[g(y)|T(y)].$$

Let $B \in \mathscr{B}^d$ so that $T^{-1}(B) \in \mathscr{B}^n$. The measure theoretic conditional expectation is defined as a function of $T(y)$ satisfying

$$\begin{aligned}\mathrm{E}\{g(y)\mathscr{I}_B[T(y)]\} &= \mathrm{E}\left[g(y)\mathscr{I}_{T^{-1}(B)}(y)\right] \\ &\equiv \int_{T^{-1}(B)} g(v)\, dP_y(v) = \int_{T^{-1}(B)} \mathrm{E}[g(y)|T(y)]\, dP_y(v)\end{aligned}$$

for all Borel sets B. As a notational matter, the collection of all sets $T^{-1}(B)$ for $B \in \mathscr{B}^d$ defines a sub-σ-field, say, $\mathscr{F}_0$ contained in $\mathscr{B}^n$ and sometimes $\mathrm{E}[g(y)|T(y)]$ is written as $\mathrm{E}[g(y)|\mathscr{F}_0]$.

In particular, we can apply this result replacing y with $(x', y')'$, $g(y)$ with $g(x,y)$, and $T(y)$ with y to see that $\mathrm{E}[g(x,y)|y]$ has the requirement that, for any $B \in \mathscr{B}^n$,

$$\begin{aligned}\mathrm{E}\left[g(x,y)\mathscr{I}_B(y)\right] &\equiv \int_{\mathbf{R}^m \times B} g(u,v)\, dP_{xy}(u,v) \\ &= \int_{\mathbf{R}^m \times B} \mathrm{E}[g(x,y)|v]\, dP_{xy}(u,v) = \int_B \mathrm{E}[g(x,y)|v]\, dP_y(v).\end{aligned}$$

If we have densities for a product measure $\mu_x \times \mu_y$,

$$\begin{aligned}\mathrm{E}\left[g(x,y)\mathscr{I}_B(y)\right] &= \mathrm{E}\left[g(x,y)\mathscr{I}_{\mathbf{R}^m \times B}(x,y)\right] \\ &\equiv \int_{\mathbf{R}^m \times B} g(u,v)\, dP_{xy}(u,v) \\ &= \int_{\mathbf{R}^m \times B} g(u,v) f_{xy}(u,v)\, d[\mu_x \times d\mu_y](u,v) \\ &= \int_B \left[\int g(u,v)[f_{xy}(u,v)/f_y(v)]\, d\mu_x(u)\right] f_y(v) d\mu_y(v) \\ &= \int_B \left[\int g(u,v)[f_{xy}(u,v)/f_y(v)]\, d\mu_x(u)\right] dP_y(v)\end{aligned}$$

so $\int g(u,v) f_{xy}(u,v)/f_y(v)\, d\mu_x(u)$ works as a version of $\mathrm{E}[g(x,y)|y]$.

Appendix D
Some Additional Measure Theory

D.1 Sigma fields

To deal with continuous distributions, probabilities are defined on sets rather than on outcomes. As discussed earlier, the probability that someone is 65 inches tall is zero but the probability that someone is in any neighborhood of 65 inches is generally positive. The sets on which we define probabilities (or any other measures) must constitute a σ-field.

Consider a set of outcomes Ω. For a set $F \subset \Omega$, define its *complement* to be

$$F^C \equiv \{\, \omega \in \Omega | \omega \notin F \,\}.$$

Definition D.1.1 A collection of subsets of Ω, say $\mathscr{F}$, is said to be a σ-*field* (σ-*algebra*) if it has three properties:

1. If $F \in \mathscr{F}$, then $F^C \in \mathscr{F}$.
2. If $F_1, F_2, \ldots \in \mathscr{F}$, then $\bigcup_{i=1}^{\infty} F_n \in \mathscr{F}$.

Items 1 and 2 imply that

$$\bigcap_{i=1}^{\infty} F_n = \left[\bigcup_{i=1}^{\infty} F_n^C \right]^C \in \mathscr{F}.$$

In particular, $F_1 \cap F_1^C = \emptyset \in \mathscr{F}$ and $\emptyset^C = \Omega \in \mathscr{F}$.

In the real line, the smallest σ-field that contains all the (open, closed, half-open – doesn't matter) finite intervals constitutes the *Borel sets*. As I said earlier, it takes a smarter person than me to dream up a set that is not Borel. In fact, we will see that the Borel sets in two and three dimensions are generated by finite rectangles and boxes, respectively.

A *measurable space* is a pair $(\Omega, \mathscr{F})$ on which we can define a measure.

D.2 Step and Simple Functions

Take numbers $-\infty \equiv a_0 < a_1 < a_2 < \cdots < a_{n-1} < \infty$ and define the sets $A_i = (a_{i-1}, a_i]$, $i = 1, \ldots, n-1$ and $A_n = (a_{n-1}, \infty)$ Also take numbers $f_1, \ldots, f_n$. The function

$$f(u) \equiv \sum_{i=1}^{n} f_i \mathscr{I}_{A_i}(u)$$

is a step function. Note that is extremely easy to compute the Reimann integral of a step function over any bounded interval. In particular,

$$\int_{a_1}^{a_{n-1}} f(u)\,du = \sum_{i=2}^{n-1} f_i(a_i - a_{i-1}).$$

Simple functions are generalization of step functions. Let $A_1, \ldots, A_n$ form a partition of the real numbers, i.e., $\mathbf{R} = \bigcup_{i=1}^{n} A_i$ and for $i \neq j$, $A_i \bigcap A_j = \emptyset$. (Obviously the A_i of the previous paragraph form a partition.) Also take numbers $f_1, \ldots, f_n$. The function

$$f(u) \equiv \sum_{i=1}^{n} f_i \mathscr{I}_{A_i}(u)$$

is a simple function. It is extremely easy to compute the Lebesgue integral of a simple function. In particular,

$$\int f(u)\,d\mu(u) = \sum_{i=1}^{n} f_i \mu(A_i).$$

Unfortunately, we need to worry about adding and subtracting infinity if we allow sets of infinite measure, which Lebesgue measure does. The Lebesgue integral over any closed bounded set is fine. Also, for any probability measure (or other measure for which the real line has bounded measure), the integral is always well-defined.

The basic idea is that a function f is Reimann integrable if it can be approximated well with step functions and Lebesgue integrable if it can be approximated well with simple functions.

To Reimann integrate an arbitrary function f we create approximating step functions

$$\tilde{f}(u) \equiv \sum_{i=1}^{n} f(x_i) \mathscr{I}_{A_i}(u),$$

where x_i is any point having $a_{i-1} \leq x_i \leq a_i$, which gives

$$\int_{a_1}^{a_{n-1}} \tilde{f}(u)\,du = \sum_{i=2}^{n-1} f(x_i)(a_i - a_{i-1}).$$

If $\int_{a_1}^{a_{n-1}} \tilde{f}(u)\,du$ converges to some number as you let $n \to \infty$, $a_i - a_{i-1} \to 0$ for all i, and regardless of how you pick x_i, but keep a_1 and a_n fixed, you have the Reimann

integral $\int_{a_1}^{a_{n-1}} f(u)\,du$. If you let a_1 and a_n go to $\mp\infty$, and the integrals converge to some finite number, you get $\int f(u)\,du$.

To Lebesgue integrate a bounded nonnegative (measureable) function f, instead of breaking up the x axis we partition the y axis with $0 \equiv b_0 < b_1 < b_2 < \cdots < b_n < \infty$ and sets $B_i = (b_{i-1}, b_i]$ with b_n equal to our bound on f. We create an approximating step function

$$\tilde{f}(u) \equiv \sum_{i=1}^{n} f_i \mathscr{I}_{f^{-1}(B_i)}(u),$$

where f_i is any point with $b_{i-1} \le f_i \le b_i$. This gives

$$\int \tilde{f}(u)\,d\mu(u) = \sum_{i=1}^{n-1} f_i \mu[f^{-1}(B_i)].$$

If $\int \tilde{f}(u)\,d\mu(u)$ converges to some number as you let $n \to \infty$, $b_i - b_{i-1} \to 0$ for all i, and regardless of how you pick f_i, you have a Lebesgue integral.

More precisely, in measure theory, we integrate nonnegative functions by getting a sequence of monotone simple functions that increase to them. We integrate regular functions by dividing them into their positive and negative parts, say

$$f^+(\omega) = \begin{cases} f(\omega) & \text{if } f(\omega) > 0 \\ 0 & f(\omega) \le 0; \end{cases} \qquad f^-(\omega) = \begin{cases} 0 & f(\omega) \ge 0 \\ -f(\omega) & \text{if } f(\omega) < 0\,. \end{cases}$$

Note that $f = f^+ - f^-$ and $|f| = f^+ + f^-$. Then $\int f d\mu \equiv \int f^+ d\mu - \int f^- d\mu$ provided the two integrals on the right both exist and are finite. If they both exist, $\int |f| d\mu \equiv \int f^+ d\mu + \int f^- d\mu$, so the integral of f only exists if the integral of $|f|$ is finite (unlike Reimann integration).

A classic example of a function that is Reimann integrable but not Lebesgue integrable is $\sin(v)/v$ on the real line. The situation is analogous to the fact that $\sum_n -1^n/n$ converges, rather like the Reimann integral of $\sin(v)/v$ existing. However $\sum_n |-1^n|/n$ does not converge, rather like the Lebesgue integral of $\sin(v)/v$ not existing because the Lebesgue integral of $|\sin(v)|/v$ does not exist.

D.3 Product Spaces and Measures

Consider the measurable space $(\mathbf{R}, \mathscr{B})$. You shouldn't be reading this if you don't know what $\mathbf{R}^2$ is, but you may not know the appropriate σ-field for $\mathbf{R}^2$. In one dimension, the Borel σ-field is generated by intervals. In two and three dimensions, it is generated by rectangles and boxes, respectively. For n dimensions let $A_1, \ldots, A_n$ be sets in $\mathbf{R}$ and $v = (v_1, \ldots, v_n)'$. Define *product sets*

$$A_1 \times \cdots \times A_n \equiv \{\, v | v_1 \in A_1, \ldots, v_n \in A_n \,\}.$$

The σ-field $\mathscr{B}^n$ is the smallest σ-field generated by all the product sets with A_is defined by finite intervals. In two dimensions, products of intervals are rectangles. In three dimensions, they are boxes.

It turns out that if you know the measure on a collection of sets that generate the σ-field, it is enough to determine the entire measure. For Lebesgue measure in n dimensions define $\mu_n(A_1 \times \cdots \times A_n) \equiv \prod_{i=1}^n \mu(A_i)$. In two or three dimensions we might write $\mu_2 = \mu \times \mu$ or $\mu_3 = \mu \times \mu \times \mu$. When getting lazy, we write $\mu_n \equiv \mu$ and let you figure out the dimension, like in Appendix B.

More generally we can have different measures on different parts of the space. For example we can have Lebesgue measure on some parts and counting measure on other parts. That would be useful if we have $\text{Bin}(N,\theta)$ data (having a density wrt counting measure) and a $\text{Beta}(\alpha,\beta)$ distribution on θ (having a density wrt Lebesgue measure). Together we would have a joint with respect to the product measure obtained by crossing counting measure with Lebesgue measure.

Consider a probability space $(\Omega,\mathscr{F},P)$ and a random vector $y : (\Omega,\mathscr{F}) \to (\mathbf{R}^n,\mathscr{B}^n)$. Specifically, write the random vector y in terms of random variables, $y = (y_1,\ldots,y_n)'$. We define P_y on $(\mathbf{R}^n,\mathscr{B}^n)$ by defining

$$P_y(A_1 \times \cdots \times A_n) \equiv P(y_1 \in A_1,\ldots,y_n \in A_n) \equiv P(y^{-1}(A_1 \times \cdots \times A_n),$$

for any sets $A_1,\ldots,A_n$ in $\mathscr{B}$. Knowing these probabilities is enough to determine the probability for any set $A \in \mathscr{B}^n$.

The random variables in y are defined to be *independent* if

$$P(y_1 \in A_1,\ldots,y_n \in A_n) = \prod_{i=1}^n P(y_i \in A_i),$$

for any sets $A_1,\ldots,A_n$ in $\mathscr{B}$. This constitutes a specification of $(\mathbf{R}^n,\mathscr{B}^n,P_y)$ in which P_y is a product measure.

The random variables are *identically distributed* if $P(y_i \in A) = P(y_j \in A)$ for any i,j and $A \in \mathscr{B}$. Random variables that are independent and identically distributed are called *iid* random variables.

If the n vector Y is multivariate normally distributed, i.e., $Y \sim N(\mu,\Sigma)$, the density is written

$$f(v|\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\det(\Sigma)^{1/2}} \exp[(v-\mu)'\Sigma^{-1}(v-\mu)/2].$$

This is the density wrt n dimensional Lebesgue product measure (which here we will call m_n) so that for any $A \in \mathscr{B}^n$

$$P(y \in A) = P_y(A) = \int \mathscr{I}_A(v) f(v|\mu,\Sigma)\, dm_n(v) \equiv \int_A f(v|\mu,\Sigma)\, dm_n(v),$$

where the last equivalence defines a shorthand notation. Similarly, the n dimensional multinomial distribution $y \sim \text{Mult}(N,p)$ has a density with respect to n dimensional product counting measure of

$$f(v|N,p) = \frac{N!}{\prod_{i=1}^{n} v_i!} \prod_{i=1}^{n} p_i^{v_i},$$

where v is any vector of nonnegative integers having $\sum_i v_i = N$.

D.4 Families of Distributions

Consider a family of probabilities on $(\Omega, \mathscr{F})$, say P_θ for $\theta \in \Theta$. Then

$$P_\theta(A) = \int_A dP_\theta = \int \mathscr{I}_A(\omega)\, dP_\theta(\omega).$$

If all of these are absolutely continuous with respect to a single (dominating) measure ν, then by Radon-Nikodym densities exist. Write the densities as

$$f(\omega|\theta), \qquad \theta \in \Theta$$

so that

$$P_\theta(A) = \int_A f(\omega|\theta) d\nu(\omega) = \int \mathscr{I}_A(\omega) f(\omega|\theta)\, d\nu(\omega).$$

Usually we think of ν as Lebesgue measure but if we let it be counting measure

$$P_\theta(A) = \sum_{\omega \in A} f(\omega|\theta),$$

where now A must be a set of integers.

For x to be a random variable, it need only be a measurable function from $(\Omega, \mathscr{F})$ to $(\mathbf{R}, \mathscr{B})$. All that requires is for $x^{-1}(B) \in \mathscr{F}$ whenever $B \in \mathscr{B}$. The random variable does not depend on θ, even though its distribution does. For $A \in \mathscr{B}$, write

$$P_\theta(x \in A) \equiv P_{x|\theta}(A).$$

The first of these is defined on $(\Omega, \mathscr{F})$ and the second is defined on $(\mathbf{R}, \mathscr{B})$. We have a dominating measure ν on $(\Omega, \mathscr{F})$ that corresponds to a dominating measure μ on $(\mathbf{R}, \mathscr{B})$ via

$$\mu(A) = \nu(x \in A).$$

The density $f(\omega|\theta)$ for $\theta \in \Theta$ relative to $(\Omega, \mathscr{F}, \nu)$ corresponds to the density $f_{x|\theta}(u) \equiv f(u|\theta)$, $\theta \in \Theta$ relative to $(\mathbf{R}, \mathscr{B}, \mu)$. [Pretty much the only way to tell $f(\omega|\theta)$ apart from $f(u|\theta)$ is context, but we almost always use the latter.]

Typically we write

$$x|\theta \sim f(u|\theta)$$

then

$$\mathrm{E}_{x|\theta}[g(x)] = \int g(u)\, dP_{x|\theta}(u) = \int g(u) f(u|\theta)\, d\mu(u)$$

$$= \int g[x(\omega)]\, dP_\theta(\omega) = \int g(\omega) f(\omega|\theta)\, d\nu(\omega).$$

Typically we would be doing that with Lebesgue measure as the domination measure μ but in the counting measure case it reduces to

$$\mathrm{E}_{x|\theta}[g(x)] = \int g[u]\, dP_{x|\theta}(u) = \sum_{\text{all } u} g(u) f(u|\theta)$$
$$= \int g[x(\omega)]\, dP_\theta(\omega) = \sum_{\text{all } \omega} g[x(\omega)] f(\omega|\theta).$$

Again, for rolling dice, think of ω as the top side of a die, $u = x(\omega)$ as the number of dots on the top side, P_θ could have θ indicating various ways to weight the die that affect which face comes up. $P_{x|\theta}$ is then the probability of the number that comes up from the weighted die (rather than the probability of what face comes up). $f(\omega|\theta)$ gives the probabilities for all the faces that may come up whereas $f_{x|\theta}(u) \equiv f(u|\theta)$ gives the probabilities for the numbers that come up. In particular,

$$P_\theta(x \in A) \equiv P_{x|\theta}(A) = \mathrm{E}_{x|\theta}\left[\mathscr{I}_A(x)\right].$$

Appendix E
Identifiability

This appendix examines the concept of identifiable parameters.

For better or worse (usually worse) much of statistical practice focuses on estimating and testing parameters. Identifiability is a property that ensures that this process is a sensible one.

Consider a collection of probability distributions $Y \sim P_\theta$, $\theta \in \Theta$. The parameter θ merely provides the name (index) for each distribution in the collection. Identifiability ensures that each distribution has a unique name/index.

Definition D.1 The parameterization $\theta \in \Theta$ is *identifiable* if $Y_1 \sim P_{\theta_1}$, $Y_2 \sim P_{\theta_2}$, and $Y_1 \sim Y_2$ imply that $\theta_1 = \theta_2$.

Being identifiable is easily confused with the concept of being well-defined.

Definition D.2 The parameterization $\theta \in \Theta$ is *well-defined* if $Y_1 \sim P_{\theta_1}$, $Y_2 \sim P_{\theta_2}$, and $\theta_1 = \theta_2$ imply that $Y_1 \sim Y_2$.

The problem with not being identifiable is that some distributions have more than one name. Observed data give you information about the correct distribution and thus about the correct name. Typically, the more data you have, the more information you have about the correct name. Estimation is about getting close to the correct name and testing hypotheses is about deciding which of two lists contains the correct name. If a distribution has more than one name, it could be in both lists. (Significance testing is about whether it seems plausible that a name is on a list, so identifiability seems less of an issue.) If a distribution has more than one name, does getting close to one of those names really help? In applications to linear models, typically distributions have only one name or they have an infinite number of names.

The ideas are roughly this. If the distributions are well-defined and I know that Wesley O. Johnson (θ_1) and O. Wesley Johnson (θ_2) are the same person ($\theta_1 = \theta_2$),

then, say, any collection of blood pressure readings on Wesley O. should look pretty much the same as comparable readings on O. Wesley. They would be two samples from the same distribution. Identifiability is the following: if all the samples I have taken or ever could take on Wesley O. look pretty much the same as samples on O. Wesley, then Wesley O. would have to be the same person as O. Wesley. (The reader might consider whether personhood is actually an identifiable parameter for blood pressure.)

For multivariate normal distributions, being well-defined is the requirement that if $Y_1 \sim N(\mu_1, V_1)$, $Y_2 \sim N(\mu_2, V_2)$, and $\mu_1 = \mu_2$ and $V_1 = V_2$, then $Y_1 \sim Y_2$. Being identifiable is that if $Y_1 \sim N(\mu_1, V_1)$, $Y_2 \sim N(\mu_2, V_2)$, and $Y_1 \sim Y_2$, then $\mu_1 = \mu_2$ and $V_1 = V_2$. Obviously, two random vectors with the same distribution have to have the same mean vector and covariance matrix. But life gets more complicated.

The more interesting problem for multivariate normality is a model

$$Y \sim N[F(\beta), V(\phi)]$$

where F and V are known functions of parameter vectors β and ϕ. To show that β and ϕ are identifiable we need to consider

$$Y_1 \sim N[F(\beta_1), V(\phi_1)], \qquad Y_2 \sim N[F(\beta_2), V(\phi_2)]$$

and show that if $Y_1 \sim Y_2$ then $\beta_1 = \beta_2$ and $\phi_1 = \phi_2$. From our earlier discussion, if if $Y_1 \sim Y_2$ then $F(\beta_1) = F(\beta_2)$ and $V(\phi_1) = V(\phi_2)$. We need to check that $F(\beta_1) = F(\beta_2)$ implies $\beta_1 = \beta_2$ and that $V(\phi_1) = V(\phi_2)$ implies $\phi_1 = \phi_2$. (Traditional analysis of variance parameterizations for the mean vector do not provide identifiability, c.f., Christensen, 2020.)

Appendix F
Multivariate Differentiation

F.1 Differentiation

If F is a function from $\mathbf{R}^s$ into $\mathbf{R}^t$ with $F(x) = [f_1(x), \ldots, f_t(x)]'$, then the derivative of F at c is the $t \times s$ matrix of partial derivatives,

$$\mathbf{d}_x F(x)|_{x=c} \equiv [\partial f_i(x)/\partial x_j|_{x=c}].$$

When the context is clear, we often use simpler notations such as

$$\mathbf{d}_x F(x)|_{x=c} \equiv \mathbf{d}_x F(c) \equiv \mathbf{d}F(c).$$

Critical points are points c where $\mathbf{d}F(c) = 0$. If $s = t = 1$, we occasionally write

$$\dot{F}(c) \equiv \mathbf{d}F(c).$$

A first order Taylor's expansion of F around the point c is

$$F(x) \doteq F(c) + [\mathbf{d}F(c)](x-c)$$

or, to be more mathematically precise,

$$F(x) = F(c) + [\mathbf{d}F(c)](x-c) + o(\|x-c\|),$$

where $\|x-c\|^2 = (x-c)'(x-c)$ and for scalars $a_n \to 0$, $o(a_n)$ has the property that $o(a_n)/a_n \to 0$. (This is a vector divided by a scalar converging to the 0 vector.)

In fact, the first order Taylor's expansion is essentially the mathematical definition of a derivative. The technical definition of a derivative, if it exists, is that it is some $t \times s$ matrix $dF(c)$ such that for any $\varepsilon > 0$, there exists a $\delta > 0$ for which any x with $\|x-c\| < \delta$ has

$$\|F(x) - F(c) - [\mathbf{d}F(c)](x-c)\| < \varepsilon\|x-c\|.$$

In other words, the linear (actually affine) function of x defined by $[\mathbf{d}F(c)](x-c)$ is a good approximation to the curved function $F(x)-F(c)$ in neighborhoods of c. Under suitable conditions, the matrix $\mathbf{d}F(c)$ is unique and equals the matrix of partial derivatives of F evaluated at the vector c.

In particular, if g maps $\mathbf{R}^s$ into $\mathbf{R}$, then $\mathbf{d}g(x)$ is a $1\times s$ *row* vector. In this case we can also define the second derivative matrix of g at c as

$$\mathbf{d}^2_{xx}g(c) \equiv \mathbf{d}^2_{xx}g(x)|_{x=c} \equiv \mathbf{d}_x\left\{[\mathbf{d}_x g(x)]'\right\}|_{x=c} = \left[\partial^2 g(x)/\partial x_i \partial x_j|_{x=c}\right],$$

which is an $s\times s$ matrix. Taylor's second-order expansion can be written

$$g(x) \doteq g(c) + [\mathbf{d}g(c)](x-c) + (x-c)'\left[\mathbf{d}^2 g(c)\right](x-c)/2$$

or, to be more mathematically precise,

$$g(x) = g(c) + [\mathbf{d}g(c)](x-c) + (x-c)'\left[\mathbf{d}^2 g(c)\right](x-c)/2 + o\left(\|x-c\|^2\right).$$

First and second order Taylor's expansions are fundamental to the models used in response surface methodology, cf. http://www.stat.unm.edu/~fletcher/TopicsInDesign or *ALM-II*.[1]

The chain rule can be written as a matrix product. If $f:\mathbf{R}^s\to\mathbf{R}^t$ and $g:\mathbf{R}^t\to\mathbf{R}^n$, then the composite function is defined by

$$(g\circ f)(x) \equiv g[f(x)]$$

and its derivative is an $n\times s$ matrix that satisfies

$$\mathbf{d}(g\circ f)(c) = [\mathbf{d}_v g(v)|_{v=f(c)}][\mathbf{d}_x f(x)|_{x=c}] \equiv \mathbf{d}g[f(c)]\mathbf{d}f(c).$$

EXAMPLE F.0. *Generalized Linear Transformations.*
Generalized linear models involve an assumption that a random n vector Y has $\mathrm{E}(Y) = G(X\beta)$ where $X_{n\times p}$ is known and β is a parameter vector. The vector function G actually repeatedly applies a scalar function $G(u)$ by defining $G(v)\equiv$

[1] Taylor's Theorem is fundamentally about mapping vectors into the real line. There are several ways to characterize the remainder involved. We have used the second-order Peano characterization $o(||x-c||^2)$. Ferguson (1996) uses an integral characterization for first-order Taylor's expansions (which are comparable to second-order Peano expansions) $g(x) = g(c) + [\mathbf{d}g(c)](x-c) + (x-c)'\left[\int_0^1\int_0^1 v\mathbf{d}^2 g(c+uv(x-c))dudv\right](x-c)/2$. Another characterization is due to Lagrange. A zero-order Taylor's expansion with the Lagrange characterization is known as the Mean Value Theorem. This Mean Value Theorem cannot be extended to the case of mapping vectors into other (nondegenerate) vectors but similar extensions can be obtained by generalizing the Taylor's Theorem integral and Peano characterizations of the remainder. Ferguson (1996) refers to the vector valued zero-order Taylor's theorem generalization with integral remainder as the Mean Value Theorem, $F(x) = F(c) + \left[\int_0^1 \mathbf{d}F(c+u(x-c))du\right](x-c)$. The comparable first-order Peano characterization extension is essentially just the definition of the derivative.

$[G(v_1),\cdots,G(v_n)]'$. Also define the scalar function $g(u) \equiv \mathbf{d}_u G(u)$ with its vector equivalent. It follows that $\mathbf{d}_v G(v) = D[g(v)]$ a diagonal matrix of derivatives and from the chain rule that $\mathbf{d}_\beta G(X\beta) = D[g(X'\beta)]X$.

I pretty much ripped this out of *ALM-III*. Not yet clear whether the following material is needed for this work.

Since we are examining linear models, we will be particularly interested in the derivatives of linear functions and quadratic functions.

Proposition F.1. Let A be a fixed $t \times s$ matrix with $t = s$ in part (b).
(a) $\mathbf{d}_x[Ax] = A$.
(b) $\mathbf{d}_x[x'Ax] = 2x'A$.

PROOF. (a) is proven by writing each element of Ax as a sum and taking partial derivatives. (b) is proven by writing $x'Ax$ as a double sum and taking partial derivatives. □

Let $A(u) = [a_{ij}(u)]$ be an $t \times s$ matrix that is a function of a scalar u. We define the derivative of $A(u)$ with respect to u as the matrix of derivatives for its individual entries, i.e.,

$$\mathbf{d}_u A(u) \equiv [\mathbf{d}_u a_{ij}(u)].$$

Functions like $A(u)$, from $\mathbf{R}$ into a matrix, do not fit into our definition of derivatives, but $\mathrm{Vec}[A(u)]$ can be thought of as a function from $\mathbf{R}$ into $\mathbf{R}^{ts}$ and all we are really doing un-Vec-ing the transpose of the derivative; the derivative being $1 \times ts$.

On the other hand, $A(u)x$ and $x'A(u)x$ are functions of u from $\mathbf{R}$ into $\mathbf{R}^t$ and $\mathbf{R}$, respectively, and it is easy to see that

$$\mathbf{d}_u[A(u)x] = [\mathbf{d}_u A(u)]x \quad \text{and that} \quad \mathbf{d}_u[x'A(u)x] = x'[\mathbf{d}_u A(u)]x.$$

We now present some useful rules for matrix derivatives. While these are specified for a scalar u, if A is a function of a vector θ, by thinking of $u = \theta_r$, we can obtain partial derivatives with respect to θ_r. (To find critical points we set all of the partial derivatives equal to 0.) The last three results in Proposition F.2 are particularly useful when dealing with likelihood functions associated with multivariate normal distributions.

Proposition F.2 For (c) and (d), $t = s$.
(a) A form of the product rule holds for conformable matrices $A(u)$ and $B(u)$,

$$\mathbf{d}_u[A(u)B(u)] = [\mathbf{d}_u A(u)]B(u) + A(u)[\mathbf{d}_u B(u)].$$

(b) When B and C are fixed matrices of conformable sizes,

$$\mathbf{d}_u[CA(u)B] = C[\mathbf{d}_u A(u)]B.$$

(c) The derivative of an inverse is

$$\mathbf{d}_u A^{-1}(u) = -A^{-1}(u)[\mathbf{d}_u A(u)]A^{-1}(u).$$

(d) The derivative of a trace is

$$\mathbf{d}_u\{\mathrm{tr}[A(u)]\} = \mathrm{tr}[\mathbf{d}_u A(u)].$$

(e) If $V(u)$ is positive definite for all u,

$$\mathbf{d}_u \log\{\det[V(u)]\} = \mathrm{tr}\left\{V^{-1}(u)[\mathbf{d}_u V(u)]\right\}.$$

The notations $\det[V]$ and $|V|$ are used interchangeably to indicate the determinant.

PROOF. See Exercise F.1 for a proof of the proposition. □

Exercise F.1. Prove Proposition F.2. Hints: For (a), consider $A(u)B(u)$ elementwise. For (b), use (a) twice. For (c), use (a) and the fact that $0 = \mathbf{d}_u I = \mathbf{d}\left[A(u)A^{-1}(u)\right]$. For (d) use the fact that the trace is a linear function of the diagonal elements. For (e), write $V = P\,\mathrm{Diag}(\phi_i)\,P'$ and show that both sides equal

$$\sum_{i=1}^{q} \mathbf{d}_u \phi_i(u)\,\frac{1}{\phi_i(u)}.$$

For the right-hand side, use (a) and the fact that $0 = \mathbf{d}_u I = \mathbf{d}_u PP'$.

Exercise F.2. *Standard Linear Models.*
The standard linear model is

$$Y = X\beta + e, \quad \mathrm{E}(e) = 0, \quad \mathrm{Cov}(e) = \sigma^2 I.$$

The least squares criterion is to choose an estimate of β that minimizes the squared Euclidean distance between Y and $X\beta$, namely

$$\|Y - X\beta\|^2 \equiv (Y - X\beta)'(Y - X\beta).$$

1. Using the chain rule, show that the first derivative of the squared error loss function is
$$\mathbf{d}_\beta (Y - X\beta)'(Y - X\beta) = -2(Y - X\beta)'X.$$
Note that setting the derivative equal to 0 leads to the well known *normal equations* $X'X\beta = X'Y$ for finding least squares estimates.
2. Show that the second derivative of $\|Y - X\beta\|^2$ is
$$\mathbf{d}^2_{\beta\beta}(Y - X\beta)'(Y - X\beta) = 2X'X.$$

Because the second derivative matrix is non-negative definite regardless of the value of β, all critical points are global minima but there may be an infinite number of them. When the second derivative matrix is positive definite, there will be a unique minimum.

Exercise F.3. *Log-Linear Models.*
Log-linear models involve observing a random q vector of counts, say, n and modeling their expectation, say, $\mathrm{E}(n) \equiv m$ using $\log(m) \equiv \mu = X\beta$ where the log function is applied to the vector m elementwise and it is merely convenient to have the name μ for $\log(m)$. The real modeling is in $X\beta$ which, just like standard linear models, has X as a known model matrix and β as a vector of unknown parameters. The parameter vector β determines both μ and m. Sometimes m and μ uniquely determine β.

As discussed in Chapters 10 and 12 of Christensen (2024), the most commonly used probabilistic models for n are independent Poisson sampling, multinomial sampling, and product-multinomial sampling. Letting J denote a vector of 1s, all of these sampling schemes lead to a log-likelihood function that is some fixed additive constant plus

$$\ell(n,\mu) \equiv n'\mu - \sum_{i=1}^{q} e^{\mu_i} = n'\mu - J'm = n'X\beta - J'\exp(X\beta), \tag{1}$$

so we seek to maximize this as a function of β. Often there are many β vectors that give the same maximization.

1. Show that the first derivative with respect to μ is

$$\mathbf{d}_\mu \ell(n,\mu) = \mathbf{d}_\mu \left(n'\mu - \sum_{i=1}^{q} e^{\mu_i} \right) == n' - m'.$$

2. To get the first derivative with respect to β use the chain rule to show

$$\mathbf{d}_\beta \ell[n,\mu(\beta)] = \left[\mathbf{d}_\mu \ell(n,\mu)\right]\left[\mathbf{d}_\beta \mu(\beta)\right] = [n - m(\beta)]'X.$$

   Setting this equal to 0 defines the *likelihood equations*.
3. Write

$$X = \begin{bmatrix} x_1' \\ \vdots \\ x_q' \end{bmatrix} = [x_{ij}],$$

   then

$$m(\beta) = (m_1, \ldots, m_q)' = \left(e^{x_1'\beta}, \ldots, e^{x_q'\beta}\right)'.$$

   and show

$$\mathbf{d}_\beta m(\beta) = \begin{bmatrix} x_{11}e^{x_1'\beta} & \cdots & x_{1p}e^{x_1'\beta} \\ \vdots & & \vdots \\ x_{q1}e^{x_q'\beta} & \cdots & x_{qp}e^{x_q'\beta} \end{bmatrix} = D[m(\beta)]X.$$

4. Use the previous result to show

$$\mathbf{d}^2_{\beta\beta}\ell[n,\mu(\beta)] = -X'D[m(\beta)]X.$$

It is implicit in our models that every element of $m(\beta)$ is positive, so the negative of the second derivative is nonnegative definite and critical points $\hat{\beta}$ will maximize of the likelihood function. When the negative of the second derivative is positive definite, the maxima will be unique.

F.2 Iterative Methods for Finding Extreme Values

Suppose we want to maximize or minimize a differentiable real valued function $L(\beta)$ of the vector β. To do this we find *critical points* where the derivative $\mathbf{d}_\beta L(\beta) \equiv \dot{L}(\beta)$ equals 0. In particular we find a sequence of points $\beta_0, \beta_1, \beta_2, \ldots$ that converge to a critical point.

F.2.1 Gradient (Steepest) Descent

This is the simplest method to find where $\mathbf{d}_\beta L(\beta) \equiv \dot{L}(\beta)$ equals 0. For some $\eta > 0$ set

$$\beta_{t+1} = \beta_t - \eta[\dot{L}(\beta_t)]'.$$

To maximize $L(\beta)$, use *steepest ascent* which replaces the minus sign with a plus sign. At convergence, $\beta_{t+1} = \beta_t$ so $[\dot{L}(\beta_t)]' = 0$ and β_t is a critical point.

Gradient descent is easy to program and easy to compute but can be inefficient. It is often used when the dimension of β is very large.

Exercise F.4. *Programming Steepest Descent.*
Consider the simple linear regression data in Table 7.1 of ANREG with data files available from `www.stat.unm.edu/~fletcher/avdr_data.zip`. Write a program for finding the least squares estimate using gradient descent. Submit a plot illustrating the sequence of values leading to the least squares estimate. Include in your plot at least three isobars of the least squares criterion function. These will be centered at the least squares estimate. For those of you programming in R, you may find it useful to modify the following program that I wrote to illustrate isobars for a normal density.

```
#install.packages("ellipse")
```

```
#Do this only once on your computer
library(ellipse)

b1=1
b2=2
A = matrix(c(1,.9,.9,2),2,2, dimnames = list(NULL, c("b1", "b2")))
A
E <-  ellipse(A,centre = c(b1, b2),t=.95,npoints = 100)
E1 <- ellipse(A,centre = c(b1, b2),t=.5 ,npoints = 100)
E2 <- ellipse(A,centre = c(b1, b2),t=.75,npoints = 100)

plot(E,type = 'l',ylim=c(.5,3.5),xlim=c(0,2),
     xlab=expression(y[1]),
 ylab=expression(y[2]),main="Normal Density")
text((b1+.01),(b2-.1),expression(mu),lwd=1,cex=1)
lines(E1,type="l",lty=1)
lines(E2,type="l",lty=1)
lines(b1,b2,type="p",lty=3)
```

F.2.2 Newton-Raphson

Newton-Raphson finds critical points of $\mathbf{d}_\beta L(\beta) \equiv \dot{L}(\beta)$ by using the second derivative $\mathbf{d}^2_{\beta\beta} L(\beta) \equiv \ddot{L}(\beta)$ in a Taylor's approximation centered on β_t of the first derivative function,

$$[\dot{L}(\beta)]' \doteq [\dot{L}(\beta_t)]' + [\ddot{L}(\beta_t)](\beta - \beta_t).$$

Setting $0 = [\dot{L}(\beta)]'$ we find β_{t+1} to be the solution for β in

$$0 = [\dot{L}(\beta_t)]' + [\ddot{L}(\beta_t)](\beta - \beta_t)$$

which gives

$$\beta_{t+1} = \beta_t - [\ddot{L}(\beta_t)]^{-1}[\dot{L}(\beta_t)]'.$$

At convergence we have

$$0 = \beta_{t+1} - \beta_t = -[\ddot{L}(\beta_t)]^{-1}[\dot{L}(\beta_t)]'$$

but by the (implicit) assumption that $[\ddot{L}(\beta_t)]$ is nonsingular, the only vector v with $-[\ddot{L}(\beta_t)]^{-1}v = 0$ is $v = 0$, so we have $\dot{L}(\beta_t) = 0$.

When β is a very high dimensional vector, finding the inverse matrix of $\ddot{L}(\beta_t)$ can be very difficult in which case the method is impractical. When this method is applicable, it tends to be very efficient. For example, it finds least squares estimates for a linear model in just one iteration regardless of the starting value β_0. Many computer programs for fitting generalized linear models employ Newton-Raphson under the name *iteratively reweighted least squares*.

See also *LLM-III*, Section 12.4a.

Exercise F.5. Use the results of Exercise F.2 to show that Newton-Raphson gives least squares estimates in only one iteration.

F.2.3 Gauss-Newton

Gauss-Newton applies only to minimizing $L(\beta) = [Y - F(\beta)]'[Y - F(\beta)]$. To minimize this, note that $[\dot{L}(\beta)] = 2[Y - F(\beta)]'[\dot{F}(\beta)]$. This method is particularly useful in nonlinear regression of which neural networks are a special case. Like Newton-Raphson it involves the inverse of a square matrix that has the same dimensions as β so, while efficient when applicable, it is not readily applied to high dimensional problems.

Use the approximation

$$F(\beta) \doteq F(\beta_t) + [\dot{F}(\beta_t)](\beta - \beta_t)$$

This approximation constitutes a linear model $Y = X\gamma + d + e$ with an offset. The offset is $d = F(\beta_t)$. The model matrix is $X = [\dot{F}(\beta_t)]$. The parameter vector is $\gamma = (\beta - \beta_t)$. From linear model theory we know the least squares solution is

$$\widehat{(\beta - \beta_t)} = \hat{\gamma} = (X'X)^{-1}X'(Y - d) = \left\{[\dot{F}(\beta_t)]'[\dot{F}(\beta_t)]\right\}^{-1}[\dot{F}(\beta_t)]'[Y - F(\beta_t)]$$

and set β_{t+1} to be the solution for β which makes

$$\beta_{t+1} = \beta_t + \hat{\gamma}.$$

At convergence we have

$$0 = \beta_{t+1} - \beta_t = \hat{\gamma} = \left\{[\dot{F}(\beta_t)]'[\dot{F}(\beta_t)]\right\}^{-1}[\dot{F}(\beta_t)]'[Y - F(\beta_t)]$$

Similar to the argument used with Newton-Raphson, this is only 0 when $0 = [\dot{F}(\beta_t)]'[Y - F(\beta_t)] = [\dot{L}(\beta_t)]'$ as we desired.

See also *ALM-III*, Subsection 7.4.1.

F.2.4 E-M (Expectation-Maximization)

We will see if I ever get around to writing this. I have had little interest in it but I gather it is extremely useful.

The EM algorithm seems to be particularly useful for missing data problems, for estimating the parameters of Gaussion (normal) mixture models, and for hidden variable models (like Factor Analysis).

References

Agresti, Alan (1992). A Survey of Exact Inference for Contingency Tables. *Statistical Science*, Vol. **7**, 131-153.

Aitchison, J. and Dunsmore, I. R. (1975). *Statistical Prediction Analysis*. Cambridge University Press, Cambridge.

Akaike, Hirotugu (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information*, edited by B.N. Petrov and F. Czaki. Akademiai Kiado, Budapest.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, Third Edition. John Wiley and Sons, New York.

Andrews, D. F. (1974). A robust method for multiple regression. *Technometrics*, **16**, 523-531.

Arbuthnot. (1710).

Arnold, S. F. (1981). *The Theory of Linear Models and Multivariate Analysis*. John Wiley and Sons, New York.

Aroian, Leo A. (1941). A Study of R. A. Fisher's z Distribution and the Related F Distribution. *The Annals of Mathematical Statistics*, **12**, 429-448.

Ash, Robert B. and Doleans-Dade, Catherine A. (2000). *Probability and Measure Theory*, Second Edition. Academic Press, San Diego.

Atkinson, A. C. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika*, **68**, 13-20.

Atkinson, A. C. (1982). Regression diagnostics, transformations and constructed variables (with discussion). *Journal of the Royal Statistical Society, Series B*, **44**, 1-36.

Atkinson, A. C. (1985). *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. Oxford University Press, Oxford.

Atwood, C. L. and Ryan, T. A., Jr. (1977). A class of tests for lack of fit to a regression model. Unpublished manuscript.

Bailey, D. W. (1953). *The Inheritance of Maternal Influences on the Growth of the Rat*. Ph.D. Thesis, University of California.

Barnard, G.A. (1949). Statistical Inference. *Journal of the Royal Statistical Society, Series B*, **11**, 115-149.

Barron, A. R. (1986). Entropy and the Central Limit Theorem. *The Annals of Probability*, **14**, 336-342.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53**, 370-418.

Bedrick, E. J., Christensen, R., and Johnson, W. (1996). A new perspective on priors for generalized linear models. *Journal of the American Statistical Association*, **91**, 1450-1460.

Bedrick, E. J. and Tsai, C.-L. (1994). Model selection for multivariate regression in small samples. *Biometrics*, **50**, 226-231.

Belsley, D. A. (1984). Demeaning conditioning diagnostics through centering (with discussion). *The American Statistician*, **38**, 73-77.

Belsley, D. A. (1991). *Collinearity Diagnostics: Collinearity and Weak Data in Regression*. John Wiley and Sons, New York.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley and Sons, New York.

Benedetti, J. K. and Brown, M. B. (1978). Strategies for the selection of log-linear models. *Biometrics*, **34**, 680-686.

Berger, J. O. (1993). *Statistical Decision Theory and Bayesian Analysis*. Revised Second Edition. Springer-Verlag, New York.

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, **18**, 1-32.

Berger, James O. and Wolpert, Robert (1984). *The Likelihood Principle*. Institute of Mathematical Statistics Monograph Series, Hayward, CA.

Berry, Donald A. (1997). *The American Statistician*, **51**, .

Berry, Donald A. (2004). Bayesian statistics and the efficiency and ethics of clinical trials. *Statistical Science*, **19**, 175-187.

Berger, J. O. (1993). *Statistical Decision Theory and Bayesian Analysis*, Revised Second Edition. Springer-Verlag, New York.

Berry, D. A. (1996). *Statistics: A Bayesian Perspective*. Duxbery, Belmont, CA.

Billingsley, Patrick (1999). *Convergence of Probability Measures*, Second Edition. Wiley, New York.

Billingsley, Patrick (2012). *Probability and Measure*, Fourth Edition. Wiley, New York.

Blachman, Nelson M., Christensen, Ronald and Utts, Jessica M. (1996). Comment on Christensen, R. and Utts, J. (1992), "Bayesian Resolution of the 'Exchange Paradox." *The American Statistician*, 50, 98-99.

Blackwell, David and Girshick, M.A. (1954). *Theory of games and statistical decisions*. John Wiley and Sons, New York. (1979, Dover Edition.)

Blom, G. (1958). *Statistical Estimates and Transformed Beta Variates*. John Wiley and Sons, New York.

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika*, **40**, 318-335.

Box, G.E.P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A*, **143**, 383-404.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, **26**, 211-246.

Box, G. E. P. and Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. John Wiley and Sons, New York.

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters*. John Wiley and Sons, New York.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. John Wiley and Sons, New York.

Breiman, Leo (2000). "Statistical Modeling: The Two Cultures," with discussion. *Statistical Science*, **16**, 199-231.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**, 1-3.

Bretz, F., Hothorn, T., and Westfall, P. (2011). *Multiple Comparisons Using R*. Chapman and Hall/CRC, Boca Raton, FL.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*, Second Edition. Springer-Verlag, New York.

Brownlee, K. A. (1965). *Statistical Theory and Methodology in Science and Engineering*, Second Edition. John Wiley and Sons, New York.

Casella, G. (2008). *Statistical Design.* Springer-Verlag, New York.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*, Second Edition. Duxbury Press, Pacific Grove, CA.

Cavanaugh, J. E. (1997). Unifying the derivations of the Akaike and corrected Akaike information criteria. *Statistics and Probability Letters*, **31**, 201-208.

Christensen, R. (1984). A note on ordinary least squares methods for two-stage sampling. *Journal of the American Statistical Association*, **79**, 720-721.

Christensen, R. (1987). The analysis of two-stage sampling data by ordinary least squares. *Journal of the American Statistical Association*, **82**, 492-498.

Christensen, R. (1989). Lack of fit tests based on near or exact replicates. *The Annals of Statistics*, **17**, 673-683.

Christensen, R. (1991). Small sample characterizations of near replicate lack of fit tests. *Journal of the American Statistical Association*, **86**, 752-756.

Christensen, R. (1993). Quadratic covariance estimation and equivalence of predictions. *Mathematical Geology*, **25**, 541-558.

Christensen, R. (1995). Comment on Inman (1994). *The American Statistician*, **49**, 400.

Christensen, R. (1996). *Analysis of Variance, Design, and Regression: Applied Statistical Methods*. Chapman and Hall, London.

Christensen, R. (1997). *Log-Linear Models and Logistic Regression*, Second Edition. Springer-Verlag, New York.

Christensen, R. (2001). *Advanced Linear Modeling: Multivariate, Time Series, and Spatial Data; Nonparametric Regression, and Response Surface Maximization*, Second Edition. Springer-Verlag, New York.

Christensen, R. (2003). Significantly insignificant F tests. *The American Statistician*, **57**, 27-32.

Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician*, **59**, 121-126.

Christensen, R. (2008). Review of *Principals of Statistical Inference* by D. R. Cox. *Journal of the American Statistical Association*, **103**, 1719-1723.

Christensen, Ronald (2014). "Review of *Fisher, Neyman, and the Creation of Classical Statistics* by Erich L. Lehmann." *Journal of the American Statistical Association*, **109**, 866-868.

Christensen, R. (2015). *Analysis of Variance, Design, and Regression: Linear Modeling for Unbalanced Data*, Second Edition. Chapman and Hall/CRC Pres, Boca Raton, FL.

Christensen, R. (2018). Comment on "A note on collinearity diagnostics and centering" by Velilla (2018). *The American Statistician*, **72**, 114-117.

Christensen, Ronald (2019). *Advanced Linear Modeling: Statistical Learning and Dependent Data*, Third Edition. Springer-Verlag, New York.

Christensen, Ronald (2020a). *Plane Answers to Complex Questions: The Theory of Linear Models*, Fifth Edition. Springer-Verlag, New York.

Christensen, Ronald (2020b). "Comment on 'Test for Trend With a Multinomial Outcome' by Szabo (2019)" *The American Statistician*, accepted.

Christensen, R. (2020c). *Log-Linear Models and Logistic Regression*, Third Edition. **Not yet published. Contact author.** Hopefully, Springer-Verlag, New York.

Christensen, R. (2019d). Another Look at Linear Hypothesis Testing in Dense High-Dimensional Linear Models. http://www.stat.unm.edu/~fletcher/AnotherLook.pdf

Christensen, R. and Bedrick, E. J. (1997). Testing the independence assumption in linear models. *Journal of the American Statistical Association*, **92**, 1006-1016.

Christensen, Ronald and Huffman, Michael D. (1985). "Bayesian point estimation using the predictive distribution." *The American Statistician*, **39**, 319-321.

Christensen, Ronald and Johnson, Wesley (2005). A Conversation with Seymour Geisser. *Statistical Science*, **22**, 621-636.

Christensen, R., Johnson, W., Branscum, A., and Hanson, T. E. (2010). *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*. Chapman and Hall/CRC Press, Boca Raton, FL.

Christensen, R., Johnson, W., and Pearson, L. M. (1992). Prediction diagnostics for spatial linear models. *Biometrika*, **79**, 583-591.

Christensen, R., Johnson, W., and Pearson, L. M. (1993). Covariance function diagnostics for spatial linear models. *Mathematical Geology*, **25**, 145-160.

Christensen, R. and Lin, Y. (2010). Linear models that allow perfect estimation. *Statistical Papers*, **54**, 695-708.

Christensen, R. and Lin, Y. (2015). Lack-of-fit tests based on partial sums of residuals. *Communications in Statistics, Theory and Methods*, **44**, 2862-2880.

Christensen, R., Pearson, L. M., and Johnson, W. (1992). Case deletion diagnostics for mixed models. *Technometrics*, **34**, 38-45.

Christensen, R. and Utts, J. (1992). Testing for nonadditivity in log-linear and logit models. *Journal of Statistical Planning and Inference*, **33**, 333-343.

Cochran, W. G. and Cox, G. M. (1957). *Experimental Designs*, Second Edition. John Wiley and Sons, New York.

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, **19**, 15-18.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*. John Wiley and Sons, New York.

Cook, R. D., Forzani, L., and Rothman, A. J. (2013). Prediction in abundant high-dimensional linear regression. *Electronic Journal of Statistics*, **7,** 3059-3088.

Cook, R. D., Forzani, L., and Rothman, A. J. (2015). Letter to the editor. *The American Statistician*, **69**, 253-254.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, New York.

Cook, R. D. and Weisberg, S. (1994). *An Introduction to Regression Graphics*. John Wiley and Sons, New York.

Cook, R. D. and Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*. John Wiley and Sons, New York.

Cornell, J. A. (1988). Analyzing mixture experiments containing process variables. A split plot approach. *Journal of Quality Technology*, **20**, 2-23.

Cox, D. R. (1958). *Planning of Experiments*. John Wiley and Sons, New York.

Cox, D. R. (2006). *Principals of Statistical Inference*. Cambridge University Press, Cambridge.

Cox, D. R. (2007). Applied Statistics: A Review. *The Annals of Applied Statistics* **1**, 1-17.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

Cox, D. R. and Reid, N. (2000). *The Theory of the Design of Experiments*. Chapman and Hall/CRC, Boca Raton, FL.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

Cressie, N. (1993). *Statistics for Spatial Data*, Revised Edition. John Wiley and Sons, New York.

Cressie, N. A. C. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. John Wiley and Sons, New York.

Daniel, C. (1959). Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics*, **1**, 311-341.

Daniel, C. (1976). *Applications of Statistics to Industrial Experimentation*. John Wiley and Sons, New York.

Daniel, C. and Wood, F. S. (1980). *Fitting Equations to Data*, Second Edition. John Wiley and Sons, New York.

Davies, R. B. (1980). The distribution of linear combinations of χ^2 random variables. *Applied Statistics*, **29**, 323-333.

de Finetti, B. (1974, 1975). *Theory of Probability*, Vols. 1 and 2. John Wiley and Sons, New York.

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.

deLaubenfels, R. (2006). The victory of least squares and orthogonality in statistics. *The American Statistician*, **60**, 315-321.

Doob, J. L. (1953). *Stochastic Processes*. John Wiley and Sons, New York.

Draper, N. and Smith, H. (1998). *Applied Regression Analysis*, Third Edition. John Wiley and Sons, New York.

Draper, N. R. and van Nostrand, R. C. (1979). Ridge regression and James-Stein estimation: Review and comments *Technometrics*, **21**, 451-466.

Duan, N. (1981). Consistency of residual distribution functions. Working Draft No. 801-1-HHS (106B-80010), Rand Corporation, Santa Monica, CA.

Durbin, J. and Watson, G. S. (1951). Testing for serial correlation in least squares regression II. *Biometrika*, **38**, 159-179.

Eaton, M. L. (1983). *Multivariate Statistics: A Vector Space Approach*. John Wiley and Sons, New York. Reprinted in 2007 by IMS Lecture Notes – Monograph Series.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science.* Cambridge University Press, Cambridge.

Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, New York.

Ferguson, Thomas S. (1996). *A Course in Large Sample Theory*. Chapman and Hall, New York.

Fienberg, S. E. (1980). *The Analysis of Cross-Classified Categorical Data*, Second Edition. MIT Press, Cambridge, MA.

Fienberg, S. E. (2006). When did Bayesian inference become "Bayesian"? *Bayesian Analysis*, **1**, 1–40

Fisher, R. A. (1922a). The goodness of fit of regression formulae, and the distribution of regression coefficients. *Journal of the Royal Statistical Society*, **85**, 597-612.

Fisher, Ronald A. (1922b). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A* , **222**, 309-368.

Fisher, R. A. (1924). "On a distribution yielding the error functions of several well known statistics," Proc. *International Math. Cong., Toronto*, **2**, 805-813.

Fisher, Ronald A. (1925). *Statistical Methods for Research Workers*, Fourteenth Edition, 1970. Hafner Press, New York.

Fisher, R. A. (1935). *The Design of Experiments*, Ninth Edition, 1971. Hafner Press, New York.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*, Third Edition, 1973. Hafner Press, New York.

Fraser, D. A. S. (1957). *Nonparametric methods in statistics*. John Wiley and Sons, New York.

Freedman, D. A. (2006). On the so-called "Huber sandwich estimator" and "robust standard errors". *The American Statistician*, **60**, 299-302.

Furnival, G. M. and Wilson, R. W. (1974). Regression by leaps and bounds. *Technometrics*, **16**, 499-511.

Galili, Tal and Meilijson, Isaac (2016). An example of an improvable Rao-Blackwell improvement, inefficient maximum likelihood estimator, and unbiased generalized Bayes estimator. *The American Statistician*, **70**, 108-113.

Geisser, Seymour (1971). The inferential use of predictive distributions. In *Foundations of Statistical Inference*, V.P. Godambe and D.A. Sprott (Eds.). Holt, Rinehart, and Winston, Toronto, 456-469.

Geisser, Seymour (1975). The predictive sample reuse method with applications. *Biometrika*, **70**, 320-328.

Geisser, Seymour (1985). On the predicting of observables: A selective update. In *Bayesian Statistics 2*, J.M. Bernardo et al. (Eds.). North Holland, 203-230.

Geisser, Seymour (1993). *Predictive Inference: An Introduction*, Chapman and Hall, New York.

Geisser, Seymour (2000). Statistics, litigation, and conduct unbecoming. In *Statistical Science in the Courtroom*, Joseph L. Gastwirth (Ed.). Springer-Verlag, New York, 71-85.

Geisser, Seymour (2005). *Modes of Parametric Statistical Inference*, John Wiley and Sons, New York.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*, Third Edition. Chapman and Hall/CRC, Boca Raton, FL.

Gnanadesikan, R. (1977). *Methods for Statistical Analysis of Multivariate Observations*. John Wiley and Sons, New York.

Goldstein, M. and Smith, A. F. M. (1974). Ridge-type estimators for regression analysis. *Journal of the Royal Statistical Society, Series B*, **26**, 284-291.

Graybill, F. A. (1976). *Theory and Application of the Linear Model*. Duxbury Press, North Scituate, MA.

Grizzle, J. E., Starmer, C. F., and Koch, G. G. (1969). Analysis of categorical data by linear models. *Biometrics*, **25**, 489-504.

Groß, J. (2004). The general Gauss–Markov model with possibly singular dispersion matrix. *Statistical Papers*, **25**, 311-336.

Guttman, I. (1970). *Statistical Tolerance Regions*. Hafner Press, New York.

Haberman, S. J. (1974). *The Analysis of Frequency Data*. University of Chicago Press, Chicago.

Hacking, I. (1965). *Logic of Statistical Inference*. Cambridge University Press.

Halmos, P. R. and Savage, L. J. (1949). Application of the Radon-Nikodym theorem to the theory of sufficient statistics. *Annals of Mathematical Statistics*, **20**, 225-241.

Hartigan, J. (1969). Linear Bayesian methods. *Journal of the Royal Statistical Society, Series B*, **31**, 446-454.

Harville, D. A. (2018). *Linear Models and the Relevant Distributions and Matrix Algebra.* CRC Press, Boca Raton, FL.

Haslett, J. (1999). A simple derivation of deletion diagnostic results for the general linear model with correlated errors. *Journal of the Royal Statistical Society, Series B*, **61**, 603-609.

Haslett, J. and Hayes, K. (1998). Residuals for the linear model with general covariance structure. *Journal of the Royal Statistical Society, Series B*, **60**, 201-215.

Hastie, T., Tibshirani, R. and Friedman, J. (2016). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition. Springer, New York.

Hill, Bruce M. (1987). The validity of the likelihood principle. *The American Statistician*, **43**, 95-100.

Hill, Joe R. (1990). A general framework for model-based statistics. *Biometrika*, **77**, 115-126.

Hinkelmann, K. and Kempthorne, O. (2005). *Design and Analysis of Experiments: Volume 2, Advanced Experimental Design*. John Wiley and Sons, Hoboken, NJ.

Hinkelmann, K. and Kempthorne, O. (2008). *Design and Analysis of Experiments: Volume 1, Introduction to Experimental Design*, Second Edition. John Wiley and Sons, Hoboken, NJ.

Hinkley, D. V. (1969). Inference about the intersection in two-phase regression. *Biometrika*, **56**, 495-504.

Hochberg, Y. and Tamhane, A. (1987). *Multiple Comparison Procedures*. John Wiley and Sons, New York.

Hodges, J. S. (2013). *Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models Using Random Effects*. Chapman and Hall/CRC, Boca Raton, FL.

Hoerl, A. E. and Kennard, R. (1970). Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics*, **12**, 55-67.

Högfeldt, P. (1979). On low F-test values in linear models. *Scandinavian Journal of Statistics*, **6**, 175-178.

Hsu, J. C. (1996). *Multiple Comparisons: Theory and Methods*. Chapman and Hall, London.

Hubbard, Raymond and Bayarri, M. J. (2003). Confusion over measures of evidence (ps) versus errors (αs) in classical statistical testing. *The American Statistician*, **57**, 171-177.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*, Second Edition. John Wiley and Sons, New York.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297-307.

Huynh, H. and Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact F-distributions. *Journal of the American Statistical Association*, **65**, 1582-1589.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer, New York.

Jeffreys, H. (1961). *Theory of Probability*, Third Edition. Oxford University Press, London.

John, P. W. M. (1971). *Statistical Design and Analysis of Experiments*. Macmillan, New York.

Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*, Sixth Edition. Prentice–Hall, Englewood Cliffs, NJ.

Kempthorne, O. (1952). *Design and Analysis of Experiments*. Krieger, Huntington, NY.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005). *Applied Linear Statistical Models*, Fifth Edition. McGraw-Hill Irwin, New York.

LaMotte, Lynn Roy (2014). The Gram-Schmidt Construction as a Basis for Linear Models, *The American Statistician*, **68**, 52-55.

Lane, David (1996). Story about Cosimo di Medici. In *Modelling and Prediction: honoring Seymour Geisser*, eds. Jack C. Lee, Wesley O. Johnson, Arnold Zellner. Springer- Verlag, New York.

Lehmann, E.L. (1959). *Testing Statistical Hypotheses*. John Wiley and Sons, New York.

Lehmann, E. L. (1983) *Theory of Point Estimation*. John Wiley and Sons, New York.

Lehmann, E. L. (1986) *Testing Statistical Hypotheses*, Second Edition. John Wiley and Sons, New York.

Lehmann, E.L. (1997) *Testing Statistical Hypotheses*, Second Edition. Springer, New York.

Lehmann, E. L. (1999) *Elements of Large-Sample Theory*. Springer, New York.

Lehmann, E. L. (2011). *Fisher, Neyman, and the Creation of Classical Statistics*. Springer, New York.

Lehmann, E.L. and Casella, George (1998). *Theory of Point Estimation*, 2nd Edition. Springer, New York

Lehmann, E.L. and Romano, J.P. (2005). *Testing Statistical Hypotheses*, Third Edition. Springer, New York.

Lehmann, E. L. and Scheffé, H. (1950). Completeness, similar regions and unbiased estimation, part I. *Sankhya*, 10, 305-340.

Lenth, R. V. (2015). The case against normal plots of effects (with discussion). *Journal of Quality Technology*, **47**, 91-97.

Lindgren, Bernard W. (1968). Statistical Theory, Second Edition. Macmillan, London.

Lindley, D. V. (1971). *Bayesian Statistics: A Review*. SIAM, Philadelphia.

McCullagh, P. (2000). Invariance and factorial models, with discussion. *Journal of the Royal Statistical Society, Series B*, **62**, 209-238.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, Second Edition. Chapman and Hall, London.

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models*, 2nd Edition. John Wiley and Sones, New York.

Mandansky, A. (1988). *Prescriptions for Working Statisticians*. Springer-Verlag, New York.

Mandel, J. (1961). Nonadditivity in two-way analysis of variance. *Journal of the American Statistical Association*, **56**, 878-888.

Mandel, J. (1971). A new analysis of variance model for nonadditive data. *Technometrics*, **13**, 1-18.

Manoukian, E. B. (1986), *Modern Concepts and Theorems of Mathematical Statistics*. Springer-Verlag, New York.

Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, **12**, 591-612.

Martin, R. J. (1992). Leverage, influence and residuals in regression models when observations are correlated. *Communications in Statistics – Theory and Methods*, **21**, 1183-1212.

Mathew, T. and Sinha, B. K. (1992). Exact and optimum tests in unbalanced split-plot designs under mixed and random models. *Journal of the American Statistical Association*, **87**, 192-200.

Mehta, C.R. and Patel, N.R. (1983). A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *Journal of the American Statistical Association*, **78**, 427-434.

Miller, F. R., Neill, J. W., and Sherfey, B. W. (1998). Maximin clusters for near replicate regression lack of fit tests. *The Annals of Statistics*, **26**, 1411-1433.

Miller, F. R., Neill, J. W., and Sherfey, B. W. (1999). Implementation of maximin power clustering criterion to select near replicates for regression lack-of-fit tests. *Journal of the American Statistical Association*, **94**, 610-620.

Miller, R. G., Jr. (1981). *Simultaneous Statistical Inference*, Second Edition. Springer-Verlag, New York.

Milliken, G. A. and Graybill, F. A. (1970). Extensions of the general linear hypothesis model. *Journal of the American Statistical Association*, **65**, 797-807.
Moguerza, J. M. and Muñoz, A. (2006). Support vector machines with applications. *Statistical Science*, **21**, 322-336.
Monlezun, C. J. and Blouin, D. C. (1988). A general nested split-plot analysis of covariance. *Journal of the American Statistical Association*, **83**, 818-823.
Morrison, D. F. (2004). *Multivariate Statistical Methods*, Fourth Edition. Duxbury Press, Pacific Grove, CA.
Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression*. Addison-Wesley, Reading, MA.
Nayak, T. K. (2002). Rao-Cramer type inequalities for mean squared error of prediction. *The American Statistician*, **56**, 102-106.
Neill, J. W. and Johnson, D. E. (1984). Testing for lack of fit in regression – a review. *Communications in Statistics, Part A – Theory and Methods*, **13**, 485-511.
Oehlert, G. W. (2010). *A First Course in Design and Analysis of Experiments.* http://users.stat.umn.edu/~gary/book/fcdae.pdf
Parmigiani, Giovanni and Inoue, Lurdes (2009). *Decision Theory : Principles and Approaches.* John Wiley and Sons, New York.
Peixoto, J. L. (1993). Four equivalent definitions of reparameterizations and restrictions in linear models. *Communications in Statistics, A*, **22**, 283-299.
Picard, R. R. and Berk, K. N. (1990). Data splitting. *The American Statistician*, **44**, 140-147.
Picard, R. R. and Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, **79**, 575-583.
Puri, M. L. and Sen, P. K. (1971). *Nonparametric Methods in Multivariate Analysis*. John Wiley and Sons, New York.
Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Administration, Harvard University, Boston.
Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, Second Edition. John Wiley and Sons, New York.
Rao, C. R. and Mitra, S. K. (1971). *Generalized Inverse of Matrices and Its Applications*. John Wiley and Sons, New York.
Ravishanker, N. and Dey, D. (2002). *A First Course in Linear Model Theory*. Chapman and Hall/CRC Press, Boca Raton, FL.
Rencher, A. C. and Schaalje, G. B. (2008). *Linear Models in Statistics*, Second Edition. John Wiley and Sons, New York.
Ripley, B. D. (1981). *Spatial Statistics*. John Wiley and Sons, New York.
Robert, C. P. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, Second Edition. Springer, New York.
Royall, Richard (1997). *Statistical Evidence : A Likelihood Paradigm*. Chapman & Hall, London.
St. Laurent, R. T. (1990). The equivalence of the Milliken-Graybill procedure and the score test. *The American Statistician*, **44**, 36-37.
Salsburg, David (2001). *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century.* Holt and Company, New York.
Savage, L. J. (1954). *The Foundations of Statistics*. John Wiley and Sons, New York.
Schafer, D. W. (1987). Measurement error diagnostics and the sex discrimination problem. *Journal of Business and Economic Statistics*, **5**, 529-537.
Schatzoff, M., Tsao, R., and Fienberg, S. (1968). Efficient calculations of all possible regressions. *Technometrics*, **10**, 768-779.
Scheffé, H. (1959). *The Analysis of Variance*. John Wiley and Sons, New York.
Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.
Searle, S. R. (1971). *Linear Models*. John Wiley and Sons, New York.
Searle, S. R. (1988). Parallel lines in residual plots. *The American Statistician*, **42**, 211.
Searle, S. R. and Pukelsheim, F. (1987). Estimation of the mean vector in linear models, Technical Report BU-912-M, Biometrics Unit, Cornell University, Ithaca, NY.

Seber, G. A. F. (1966). *The Linear Hypothesis: A General Theory*. Griffin, London.
Seber, G. A. F. (1977). *Linear Regression Analysis*. John Wiley and Sons, New York.
Seber, G. A. F. (2015). *The Linear Model and Hypothesis: A General Theory*. Springer, New York.
Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*, Wiley, (paperback, 2001)
Shapiro, S. S. and Francia, R. S. (1972). An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, **67**, 215-216.
Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, **52**, 591-611.
Shewhart, W. A. (1931). *Economic Control of Quality*. Van Nostrand, New York.
Shewhart, W. A. (1939). *Statistical Method from the Viewpoint of Quality Control*. Graduate School of the Department of Agriculture, Washington. Reprint (1986), Dover, New York.
Shi, L. and Chen, G. (2009). Influence measures for general linear models with correlated errors. *The American Statistician*, **63**, 40-42.
Shillington, E. R. (1979). Testing lack of fit in regression without replication. *Canadian Journal of Statistics*, **7**, 137-146.
Shumway, R. H. and Stoffer, D. S. (2011). *Time Series Analysis and Its Applications: With R Examples,* Third Edition. Springer, New York.
Skinner, C. J., Holt, D., and Smith, T. M. F. (1989). *Analysis of Complex Surveys*. John Wiley and Sons, New York.
Smith, A. F. M. (1986). Comment on an article by B. Efron. *The American Statistician*, **40**, 10.
Snedecor, G. W. and Cochran, W. G. (1980). *Statistical Methods*, Seventh Edition. Iowa State University Press, Ames.
Stefanski, L. A. (2007). Residual (sur)realism. *The American Statistician*, **61**, 163-177.
Stigler, S.M. (1982). Thomas Bayes and Bayesian inference. *Journal of the Royal Statistical Society, A*, **145**(2), 250-258.
Stigler, S.M. (2007). The epic story of maximum likelihood. *Statistical Science*, **22**, 598-620.
Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, B*, **36**, 44-47.
Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics, Part A, Theory and Methods*, **7**, 13-26.
Sulzberger, P. H. (1953). The effects of temperature on the strength of wood, plywood and glued joints. Department of Supply, Report ACA-46, Aeronautical Research Consultative Committee, Australia.
Tarpey, T., Ogden, R., Petkova, E., and Christensen, R. (2015). Reply. *The American Statistician*, **69**, 254-255.
Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B*, **58**, 267-288.
Tukey, J. W. (1949). One degree of freedom for nonadditivity. *Biometrics*, **5**, 232-242.
Utts, J. (1982). The rainbow test for lack of fit in regression. *Communications in Statistics—Theory and Methods*, **11**, 2801-2815.
Velilla, S. (2018). A note on collinearity diagnostics and centering. *The American Statistician*, **72**, 140-146.
von Neumann, John and Morgenstern, Oskar (1944). *Theory of Games and Economic Behavior*. (Third Edition, 1945; Reprinted, 2007.) Princeton University Press, Princeton.
Wald, Abraham (1950). *Statistical Decision Functions*. John Wiley and Sons, New York.
Wasserman, Larry (2004). *All of Statistics*. Springer, New York.
Weisberg, S. (2014). *Applied Linear Regression*, Fourth Edition. John Wiley and Sons, New York.
Wermuth, N. (1976). Model search among multiplicative models. *Biometrics*, **32**, 253-264.
Wichura, M. J. (2006). *The Coordinate-Free Approach to Linear Models*. Cambridge University Press, New York.
Wilks, S. S. (1962). *Mathematical Statistics*. John Wiley and Sons, New York.
Williams, E. J. (1959). *Regression Analysis*. John Wiley and Sons, New York.

Wu, C. F. J. and Hamada, M. S. (2009). *Experiments: Planning, Analysis, and Optimization*, 2nd Edition. John Wiley and Sons, New York.

Zacks, S. (197).1 *The Theory of Statistical Inference*. John Wiley and Sons, New York.

Zelen, Marvin (1996). After dinner remarks: On the occasion of Seymour Geisser's 65th Birthday, Hsinchu, Taiwan, December 13, 1994. In *Modelling and Prediction: honoring Seymour Geisser*, eds. Jack C. Lee, Wesley O. Johnson, Arnold Zellner. Springer-Verlag, New York.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. John Wiley and Sons, New York.

Zhu, M. (2008). Kernels and ensembles: Perspectives on statistical learning. *The American Statistician*, **62**, 97-109.

Albert, James (1997). *The American Statistician*, **51**, .

Moore, David (1997). *The American Statistician*, **51**, .

Index