STATS 477/577 – Data Analysis 1
Diasorin Case Study

Department of Mathematics & Statistics
University of New Mexico

15 February 2018

Background

- Diasorin is a commercial assay (test) which, its manufacturers claim, can differentiate between individuals with low and normal bone turnover.

Background

- Diasorin is a commercial assay (test) which, its manufacturers claim, can differentiate between individuals with low and normal bone turnover.
- Bone turnover refers to the continual process where old bone cells in the body are replaced with new bone cells.

Background

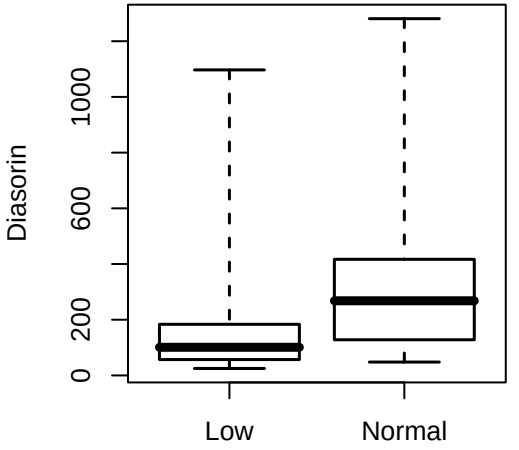
- Diasorin is a commercial assay (test) which, its manufacturers claim, can differentiate between individuals with low and normal bone turnover.
- Bone turnover refers to the continual process where old bone cells in the body are replaced with new bone cells.
- When kidneys fail to maintain proper levels of phosphorous and calcium in the blood, knowing a patient's rate of bone turnover is important for managing their health.

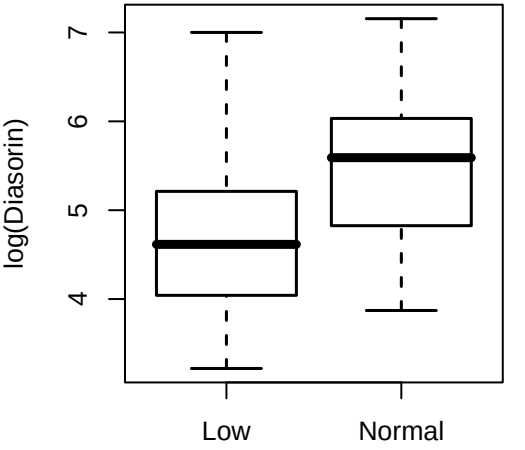
The Study

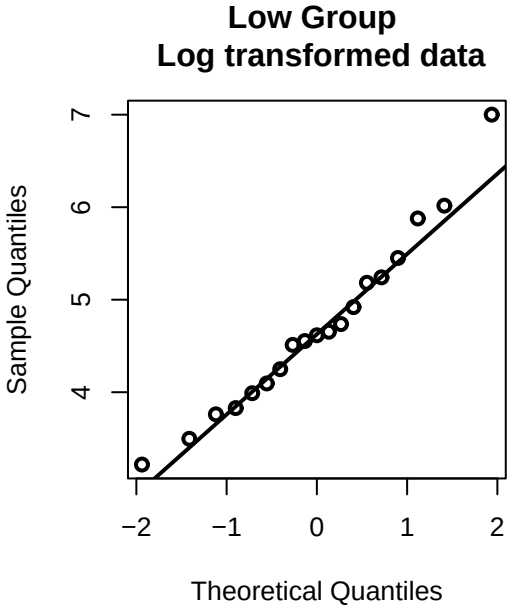
- 34 kidney patients from the bone registry at the University of Kentucky were identified as low or normal turnover by other means.
- These patients were then given the commercial assay to determine whether it could correctly identify them.

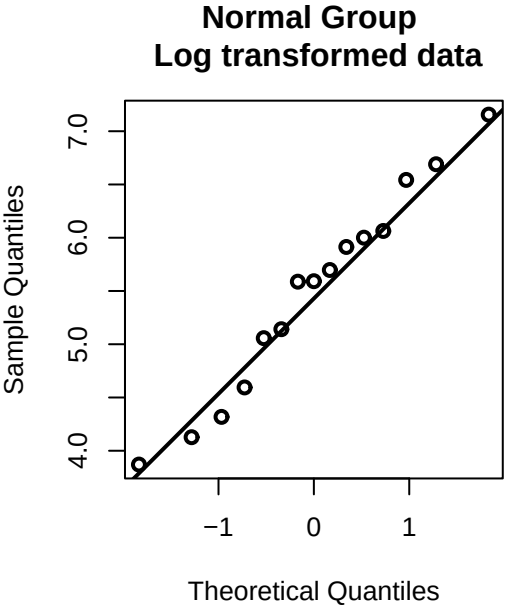
A Data Concern

- From boxplots a normal sampling model appears untenable due to marked skewness but boxplots and quantile plots of the log transformed data seem reasonably normal.


Diasorin
0
200
600
1000
Low
Normal


log(Diasorin)
4
5
6
7
Low
Normal


Low Group
Log transformed data
Sample Quantiles
4
5
6
7
−2
−1
0
1
2
Theoretical Quantiles


Normal Group
Log transformed data
Sample Quantiles
4.0
5.0
6.0
7.0
−1
0
1
Theoretical Quantiles

Data Structure

- In general we assume the sampling model

$$y_{11}, \ldots, y_{1n_1} | \mu_1, \tau_1 \overset{\text{iid}}{\sim} N(\mu_1, 1/\tau_1) \quad \perp\!\!\!\perp$$

$$y_{21}, \ldots, y_{2n_2} | \mu_2, \tau_2 \overset{\text{iid}}{\sim} N(\mu_2, 1/\tau_2).$$

Data Structure

- In general we assume the sampling model

$$y_{11}, \ldots, y_{1n_1} | \mu_1, \tau_1 \overset{\text{iid}}{\sim} N(\mu_1, 1/\tau_1) \quad \perp\!\!\!\perp$$

$$y_{21}, \ldots, y_{2n_2} | \mu_2, \tau_2 \overset{\text{iid}}{\sim} N(\mu_2, 1/\tau_2).$$

- We typically assume independent priors

$$\mu_1, \tau_1 \quad \perp\!\!\!\perp \quad \mu_2, \tau_2.$$

Data Structure

- In general we assume the sampling model

$$y_{11}, \ldots, y_{1n_1} | \mu_1, \tau_1 \overset{\text{iid}}{\sim} N(\mu_1, 1/\tau_1) \quad \perp\!\!\!\perp$$

$$y_{21}, \ldots, y_{2n_2} | \mu_2, \tau_2 \overset{\text{iid}}{\sim} N(\mu_2, 1/\tau_2).$$

- We typically assume independent priors

$$\mu_1, \tau_1 \quad \perp\!\!\!\perp \quad \mu_2, \tau_2.$$

- We can use any of the one-sample techniques: reference priors, conjugate priors, or independence priors, to determine the prior distributions.

Eliciting Priors on μ

- We need to elicit priors on both parameters of the log-normal distribution: μ and τ.
- We will elicit different priors for both our groups.
- We will also consider a reference prior for sensitivity analysis.

Eliciting Priors on μ

- Remember that we've log-transformed our data – so the distribution of the data, and by Jensen's Inequality its expectation, have changed.

Eliciting Priors on μ

- Remember that we've log-transformed our data – so the distribution of the data, and by Jensen's Inequality its expectation, have changed.
- The median is unchanged by the transformation, however, so we'll elicit prior information on it.

Eliciting Priors on μ

- Remember that we've log-transformed our data – so the distribution of the data, and by Jensen's Inequality its expectation, have changed.
- The median is unchanged by the transformation, however, so we'll elicit prior information on it.
- We specify a best guess, $\tilde{m}$, for the median, and a percentile, $\tilde{u}$, for which we are, say, 95% sure that the median is below (or above).

Eliciting Priors on μ

- An expert tells us that he thinks the median for the low bone turnover group will be 130. Further, he is 95% sure that the median of will be less than 142 in this patient population.

Eliciting Priors on μ

- An expert tells us that he thinks the median for the low bone turnover group will be 130. Further, he is 95% sure that the median of will be less than 142 in this patient population.
- For the normal bone turnover group, he believes the median will be 220, with 95% certainty that it is below 240.

Eliciting Priors on μ

- Our expert has given us priors on the real data scale, but we've log-transformed our data. So now we need to log-transform our priors.

Eliciting Priors on μ

- Our expert has given us priors on the real data scale, but we've log-transformed our data. So now we need to log-transform our priors.
- Why? The median of a Normal distribution is equal to the mean, so if we log-transform our expert's guesses at the median, we'll receive information on the mean of the Normal distribution for the log-transformed data.

Eliciting Priors on μ

- Our expert has given us priors on the real data scale, but we've log-transformed our data. So now we need to log-transform our priors.
- Why? The median of a Normal distribution is equal to the mean, so if we log-transform our expert's guesses at the median, we'll receive information on the mean of the Normal distribution for the log-transformed data.
- Then we obtain $\mu_L \equiv \mu_1 \sim N(4.87, 0.00288)$ and $\mu_N \equiv \mu_2 \sim N(5.39, 0.00280)$.

Eliciting Priors on τ

- τ is a much harder quantity to understand than μ. It measures the degree of variability in the data.

Eliciting Priors on τ

- τ is a much harder quantity to understand than μ. It measures the degree of variability in the data.
- Moreover, we want an idea of the variability of the data on its original scale, where we've got evidence of non-normality.

Eliciting Priors on τ

- τ is a much harder quantity to understand than μ. It measures the degree of variability in the data.
- Moreover, we want an idea of the variability of the data on its original scale, where we've got evidence of non-normality.
- The easiest way to elicit a prior on a scale/rate parameter like τ is to ask about percentiles of the underlying data.

Eliciting Priors on τ

- We elicit information from our expert about the 90th (or some other) percentile of the data distribution. The log of this is $\mu + 1.645\sqrt{1/\tau}$. The elicitation is now conditional on the best guess for μ being $\log(\tilde{m})$.

Eliciting Priors on τ

- We elicit information from our expert about the 90th (or some other) percentile of the data distribution. The log of this is $\mu + 1.645\sqrt{1/\tau}$. The elicitation is now conditional on the best guess for μ being $\log(\tilde{m})$.
- Then our best guess for τ, conditional on our best guess for μ, is obtained by solving $\log(u_{0.90}) = \log(\tilde{m}) + 1.645\sqrt{1/\tilde{\tau}}$

Eliciting Priors on τ

- We elicit information from our expert about the 90th (or some other) percentile of the data distribution. The log of this is $\mu + 1.645\sqrt{1/\tau}$. The elicitation is now conditional on the best guess for μ being $\log(\tilde{m})$.
- Then our best guess for τ, conditional on our best guess for μ, is obtained by solving $\log(u_{0.90}) = \log(\tilde{m}) + 1.645\sqrt{1/\tilde{\tau}}$
- We obtain $\log(u_{0.9}/\tilde{m}) = 1.645\sqrt{1/\tilde{\tau}}$ or

$$\tilde{\tau} = 1.645^2/\{\log(u_{0.9}/\tilde{m})\}^2.$$

Eliciting Priors on τ

- Since we assume a Gamma(c, d) prior for τ, set $(c-1)/d = \tilde{\tau}$ or equivalently $c = 1 + \tilde{\tau} d$.

Eliciting Priors on τ

- Since we assume a Gamma(c, d) prior for τ, set $(c-1)/d = \tilde{\tau}$ or equivalently $c = 1 + \tilde{\tau} d$.
- We can proceed with eliciting an upper limit on $u_{0.90}$ but often the expert wants to stop in which case we introduce the same large variability as in a proper Gamma reference prior by picking $d = 0.001$.

Eliciting Priors on τ

- Our expert provided his best guess for the 90th percentile of Diasorin values in the low ($u_{0.90,1} = 170$) and normal ($u_{0.90,2} = 280$) bone turnover groups.

Eliciting Priors on τ

- Our expert provided his best guess for the 90th percentile of Diasorin values in the low ($u_{0.90,1} = 170$) and normal ($u_{0.90,2} = 280$) bone turnover groups.
- We use gamma priors with modes 170 and 280, and with large variances: $\tau_1 \sim \text{Gamma}(1.0376, 0.001)$ and $\tau_2 \sim \text{Gamma}(1.04653, 0.001)$.

"This log-normal thing seems like a real pain. We can use whatever distribution we feel like for the priors, so why can't we pick an easier distribution for the data?"

"This log-normal thing seems like a real pain. We can use whatever distribution we feel like for the priors, so why can't we pick an easier distribution for the data?"

- At the end of the day, our models are only as good as their ability to represent reality. We want to use them to predict what will happen in the future, or to understand the probability associated with future events.

"This log-normal thing seems like a real pain. We can use whatever distribution we feel like for the priors, so why can't we pick an easier distribution for the data?"

- At the end of the day, our models are only as good as their ability to represent reality. We want to use them to predict what will happen in the future, or to understand the probability associated with future events.
- We need sensible distributions for any part of the model we're interested in interpreting after the fact.

"Eliciting expert information and turning them into prior distributions seems like a lot of work. Why can't we just use reference priors instead?"

"Eliciting expert information and turning them into prior distributions seems like a lot of work. Why can't we just use reference priors instead?"

- For me, there are two big reasons: intellectual honesty and understanding the problem.

"Eliciting expert information and turning them into prior distributions seems like a lot of work. Why can't we just use reference priors instead?"

- *Intellectual Honesty* – Science doesn't happen in a vacuum. The very fact that we're able to formulate scientific hypotheses says a lot about our level of knowledge about natural phenomena. Pretending that we don't have this information for the sake of "objectivity" is just papering over the truth. It's better to accept our prior beliefs and incorporate them into our work. Techniques like sensitivity analysis help us confront those beliefs head-on and see whether they're affecting our results.

"Eliciting expert information and turning them into prior distributions seems like a lot of work. Why can't we just use reference priors instead?"

- *Understanding the Problem* – Any high-level modeling work is going to involve a lot of parameters and data vectors, and it's easy to lose sight of what you really care about in a problem. Eliciting expert information and building informative priors is one way to explore your data and try to understand it better. If you have some sense of what data values you can expect, it's easier to spot output that doesn't make sense. More positively, it's easier to spot meaningful results when you see them.

"Normal priors aren't bad, but Gamma priors suck. Can't we use different priors that are easier to understand?"

"Normal priors aren't bad, but Gamma priors suck. Can't we use different priors that are easier to understand?"

- You can use whatever priors you want. But your models will work best – be most efficient, be free of errors – if you use priors that make sense in context.

"Normal priors aren't bad, but Gamma priors suck. Can't we use different priors that are easier to understand?"

- You can use whatever priors you want. But your models will work best – be most efficient, be free of errors – if you use priors that make sense in context.
- Uniform and (scaled) Beta priors only make sense when you know that there are hard limits for which values you can see, since they have finite support.

"Normal priors aren't bad, but Gamma priors suck. Can't we use different priors that are easier to understand?"

- You can use whatever priors you want. But your models will work best – be most efficient, be free of errors – if you use priors that make sense in context.
- Uniform and (scaled) Beta priors only make sense when you know that there are hard limits for which values you can see, since they have finite support.
- Normal priors are great for anything you think is symmetric.

"Normal priors aren't bad, but Gamma priors suck. Can't we use different priors that are easier to understand?"

- You can use whatever priors you want. But your models will work best – be most efficient, be free of errors – if you use priors that make sense in context.
- Uniform and (scaled) Beta priors only make sense when you know that there are hard limits for which values you can see, since they have finite support.
- Normal priors are great for anything you think is symmetric.
- If you don't have hard limits or symmetry, i.e. if you have a skewed distribution without finite support, the Gamma distribution is usually the easiest choice.

We'll start by looking at the posterior distribution for these data under the informative priors, and those priors themselves.

(a) Posterior from Informative Priors

Parameter	mean	sd	2.50%	median	97.50%
μ_L	4.86	0.05212	4.757	4.86	4.961
μ_N	5.395	0.05143	5.294	5.395	5.496
$\mu_N - \mu_L$	0.5356	0.07315	0.3922	0.5355	0.6793
τ_L	1.275	0.3944	0.6245	1.231	2.161
τ_N	1.285	0.4389	0.5757	1.236	2.274
τ_N/τ_L	1.114	0.5547	0.3873	1.003	2.496

(b) Informative Priors

Parameter	mean	sd	2.50%	median	97.50%
μ_L	4.87	0.053	4.765	4.87	4.975
μ_N	5.39	0.053	5.286	5.39	5.494
$\mu_N - \mu_L$	0.52	0.07546	0.3724	0.5201	0.668
τ_L	1037	1019	29.45	729.59	3767.5
τ_N	1046	1023	30.48	738.26	3786.03
τ_N/τ_L	10.11	638.7	0.02897	1.01	35.67

Next, let's look at what would happen if we used reference priors (proper or improper).

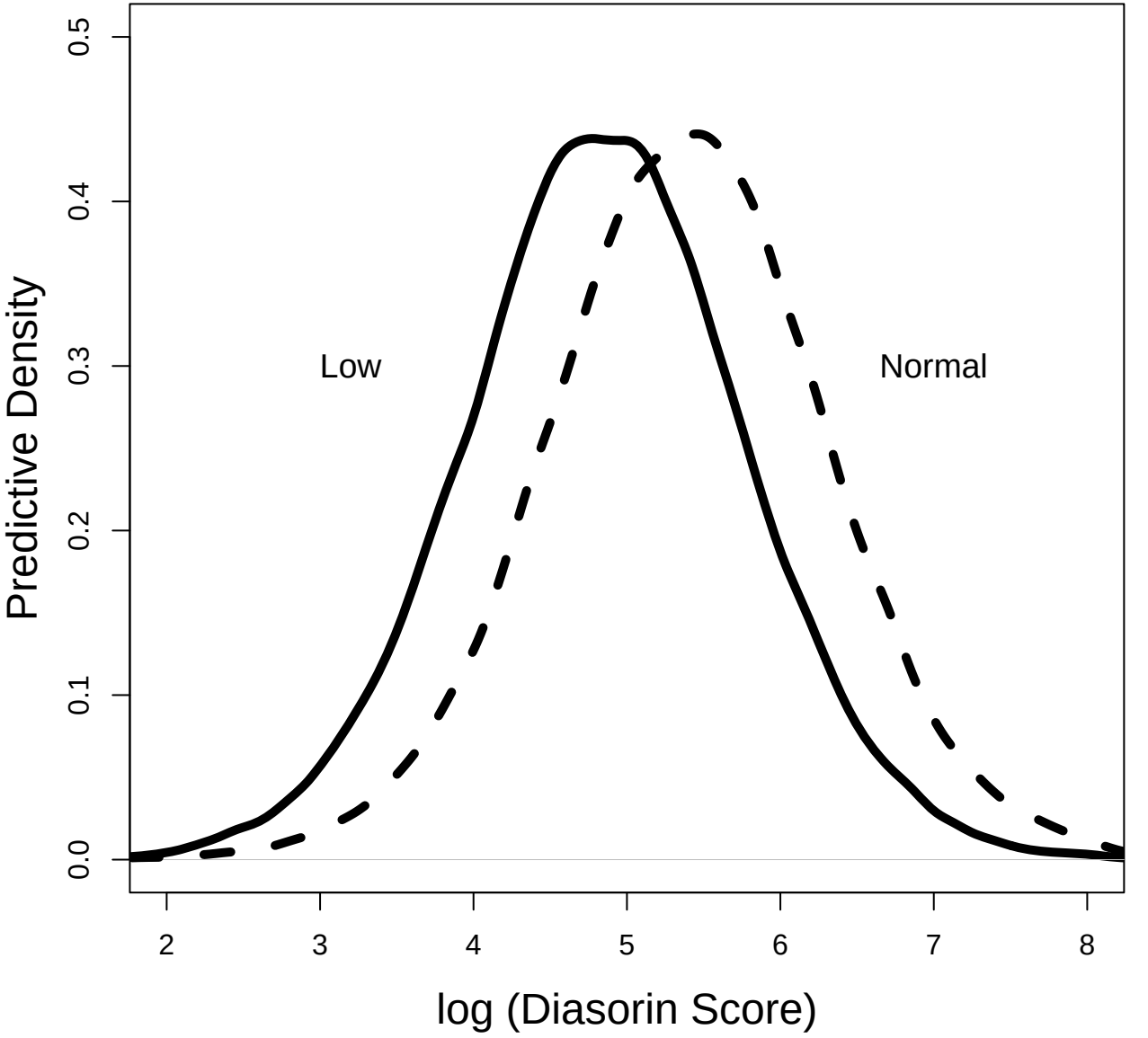
(c) Posterior from Proper Reference Prior for Means

Parameter	mean	sd	2.50%	median	97.50%
μ_L	4.706	0.2167	4.278	4.706	5.131
μ_N	5.49	0.2486	4.997	5.491	5.984
$\mu_N - \mu_L$	0.784	0.3296	0.135	0.7826	1.435
τ_L	1.252	0.3961	0.5991	1.208	2.147
τ_N	1.227	0.4319	0.5332	1.178	2.205
τ_N/τ_L	1.088	0.5596	0.3643	0.9723	2.49

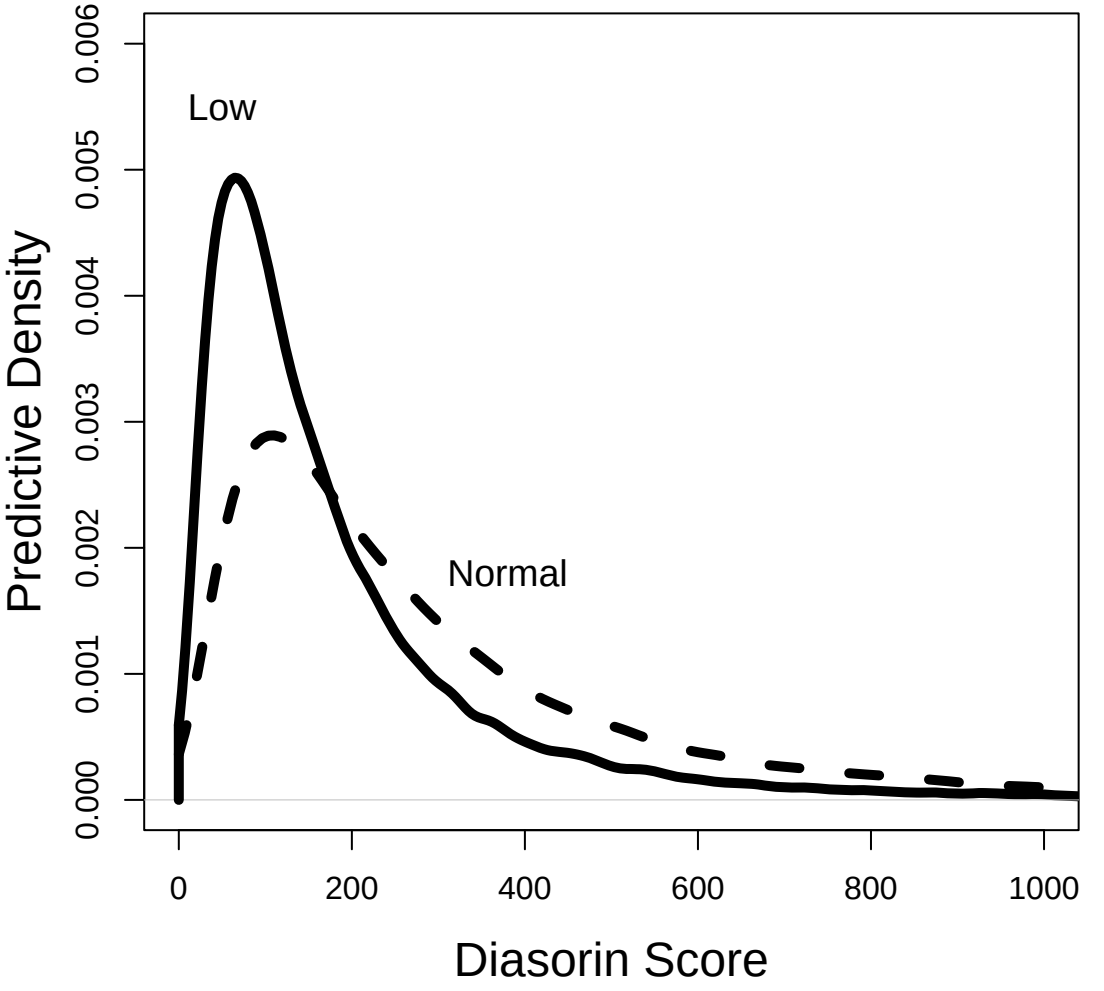
(d) Posterior from Improper Reference Prior

Parameter	mean	sd	2.50%	median	97.50%
μ_L	4.71	0.94	4.257	4.71	5.163
μ_N	5.49	0.97	4.953	5.49	6.027
τ_L	1.13	0.377	0.517	1.089	1.979
τ_N	1.06	0.401	0.427	1.013	1.983

- The next slide gives predictive densities for a future log Diasorin value from the low and normal groups. Note the similarity of the distributional shapes, which is due to the similarity of the precisions. With similar precisions, it becomes clear that the "normal" group has higher scores and that the means characterize the differences between the two distributions.


Predictive Density
0.0
0.1
0.2
0.3
0.4
0.5
Low
Normal
2
3
4
5
6
7
8
log (Diasorin Score)

- If the variances were not the same, the difference in means would not be nearly so meaningful. To illustrate this, the next slide gives the predictive distributions of the Diasorin scores. These densities are much more difficult to interpret relative to one another. In particular, it is not obvious that the difference in the means of the predictive distributions would be a good measure of how the two distributions differ.


Low
Normal
Predictive Density
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0
200
400
600
800
1000
Diasorin Score

The Heteroscedasticity Problem

- When we've got unequal variances for two populations we want to compare, comparing the means is not necessarily a good idea. We need to stop and think about what sort of comparisons we're actually interested in.

The Heteroscedasticity Problem

- When we've got unequal variances for two populations we want to compare, comparing the means is not necessarily a good idea. We need to stop and think about what sort of comparisons we're actually interested in.
- If we're interested in long-term averages, the mean may still be a reasonable point to consider. But if we're interested in uncommon events (which we sometimes call *tail behavior*), knowing which population has a higher density over a given region of the support becomes important.

The Heteroscedasticity Problem

- When we've got unequal variances for two populations we want to compare, comparing the means is not necessarily a good idea. We need to stop and think about what sort of comparisons we're actually interested in.
- If we're interested in long-term averages, the mean may still be a reasonable point to consider. But if we're interested in uncommon events (which we sometimes call *tail behavior*), knowing which population has a higher density over a given region of the support becomes important.
- Also, if we consider unbalanced loss functions (where an incorrect decision in one direction is more damaging than an incorrect decision in the other direction – think Type I and Type II error), we will be more concerned about the entire density of the two populations than just the mean of those populations.