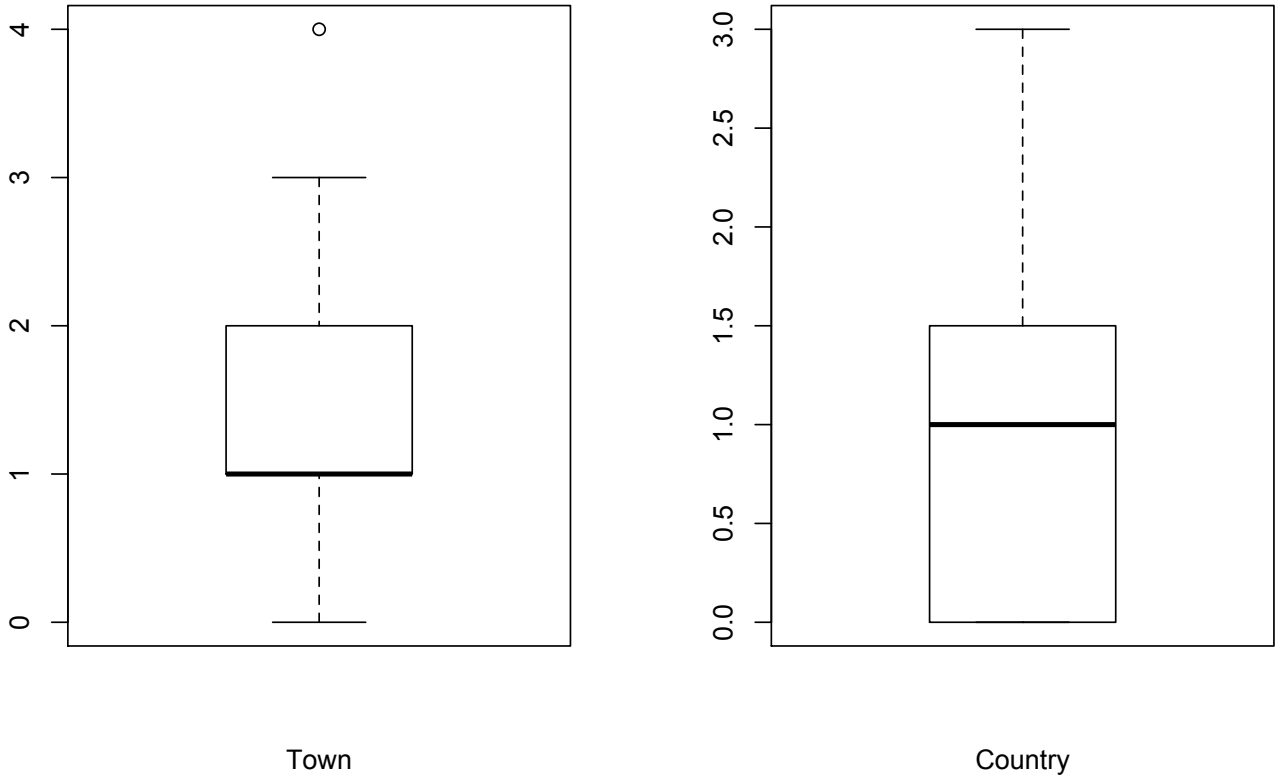
# Example from Table 2.1 data

- $Y_{jk}$ number of chronic medical conditions for women that use similar practitioner services.
- $j = 1$ 'town'; $j = 2$ 'country'
- $k = 1, 2, \ldots, k_j$ where $k_1 = 26$ and $k_2 = 23$.
- For 'town' , $\bar{Y}_{1.} = 1.423$ and $SD = 1.17$.
- For 'country', $\bar{Y}_{2.} = 0.913$ and $SD = 0.9$.
- We can try to model $Y'_{jk}s$ as independent $Poisson(\theta_j)$. $\theta_j$ rate for group $j$.
- Recall that if $Y_i \sim Poisson(\theta)$, the *log-likelihood function* is

$$l(\theta; Y) = \sum_i Y_i log(\theta) - \theta \sum_i Y_i + \sum_i log(y_i!)$$

- Do women have similar levels in the two groups?

# Boxplots of the data by group



- Hypothesis testing problem,

$$H_0 : \theta_1 = \theta_2 = \theta \text{ vs } H_1 : \theta_1 \neq \theta_2$$

- Two *models*: Under $H_0$, $Y_{jk} \sim Poisson(\theta)$.
- Second model, $Y_{jk} \sim Poisson(\theta_j)$ (nested models).
- Under $H_0$,

$$l_0 = l(\theta; Y) = \sum_{j=1}^{2} \sum_{k=1}^{k_j} \left[ y_{jk} log(\theta) - \theta - log(y_{jk}!) \right]$$

- The MLE is

$$\hat{\theta} = \sum_{j=1}^{2} \sum_{k=1}^{k_j} y_{jk} / N; N = 26 + 23 = 49$$

- Notice the MLE is ”pooled estimate” of the town and country means.
- The value of $l_0$ at $\hat{\theta}$ is $\hat{l}_0 = -68.3868$.
- If $H_1$ is true,

$$l_1 = l(\theta_1, \theta_2; Y) = \sum_{j=1}^{2} \sum_{k=1}^{k_j} \left( y_{jk} log(\theta_j) - \theta_j - log(y_{jk}!) \right)$$

- Sum of terms with $\theta_1$ plus sum of terms with $\theta_2$.
- The MLEs are

$$\hat{\theta_1} = 1.423; \hat{\theta_2} = 0.913$$

- The value of $l_1$ at the MLEs is $\hat{l}_1 = -67.0230$

- Notice that regardless of the data $\hat{l}_1 \geq \hat{l}_0$. Why?
- How can we decide if difference in log-likelihood is significant?
- Need to know the *sampling distribution* of $l(\theta; Y)$.
- What is the distribution of $\hat{l}_1 - \hat{l}_0$?
- Could also rely on the *Akaike Information Criterion* AIC.

$$AIC = -2l(\hat{\theta}; Y) + 2p$$

where $p$ is the number of parameters in the statistical model.

- Select model with minimum *AIC*.

- For our example,

$$AIC_0 = 2(68.3868) + 2(1) = 138.7736$$

$$AIC_1 = 2(67.0230) + 2(2) = 138.046$$

- So the preferred model is $\theta_1 \neq \theta_2$ but barely.
- AIC rewards *goodness of fit* or models with large likelihood
- Includes *penalty* that increases with number of parameters.
- Attempts to avoid *overfitting*.
- Different penalties provide different criteria,

$$BIC = -2l(\hat{\theta}; Y) + p * log(N)$$

- Residuals based on,

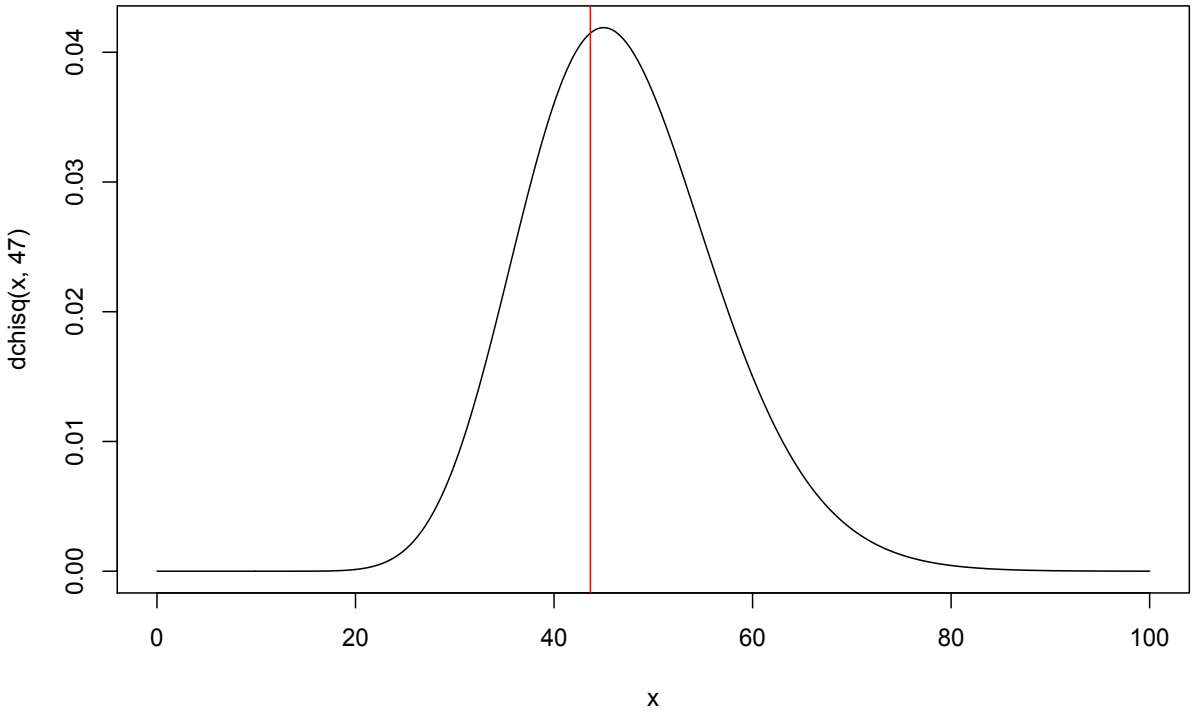
$$r_{ik} = \frac{(Y_{ik} - \hat{\theta}_k)}{\sqrt{\hat{\theta}_k}}$$

- For case where $\theta_0 = \theta_1$, $\hat{\theta}_k = 1.184, k = 1, 2$
- $\sum_k \sum_i r_{ik}^2 = 46.746$
- Compare to a $\chi^2(m)$ distribution, where

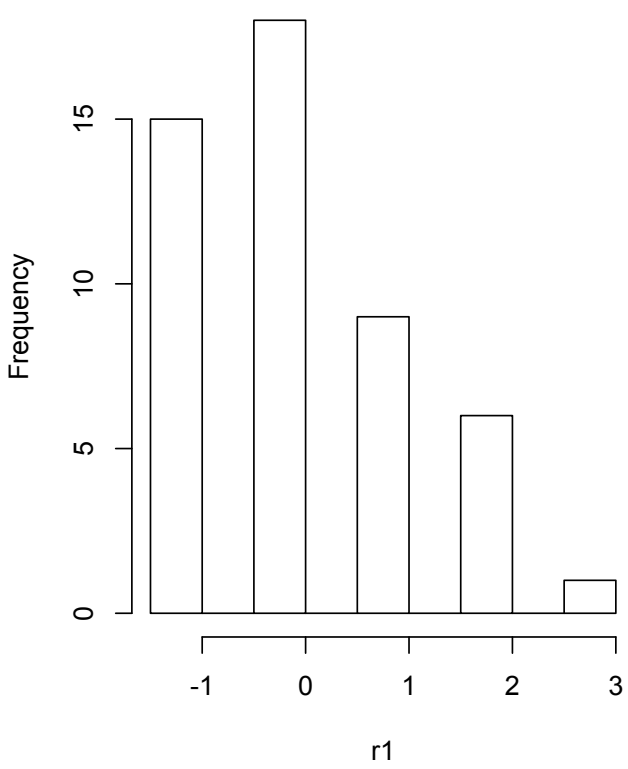
$$m = \textit{no. observations} - \textit{no. of estimated parameters}$$

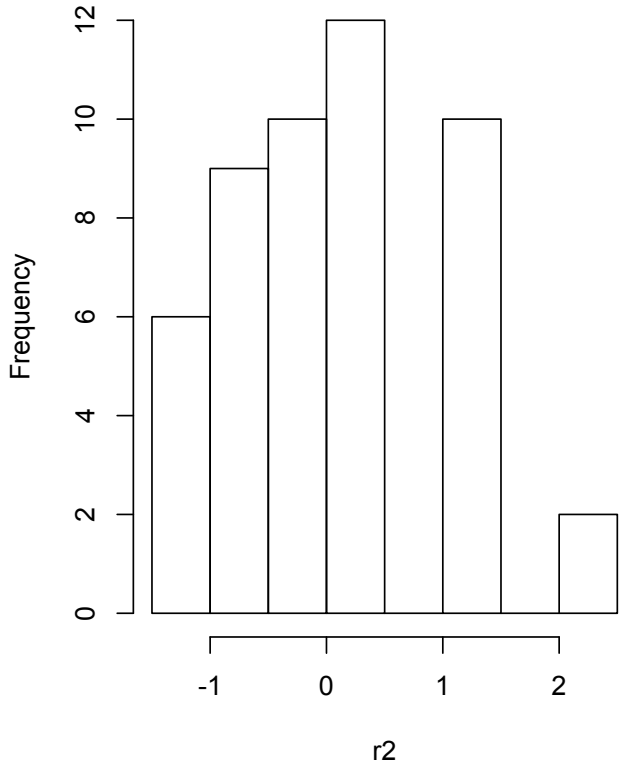
- $m = 26 + 23 - 1 = 48$ degrees of freedom.
- $Pr[\chi^2_{(48)} \geq 46.75] \approx 0.52$

- For Model 2: $\theta_0 \neq \theta_1$, $\sum_k \sum_i r_{ik}^2 = 43.659$
- $m = 26 + 23 - 2 = 47$ degrees of freedom.
- $Pr[\chi^2_{(47)} \geq 0.6117] \approx 0.52$



# Histograms of residuals for Model1 and Model 2.



# General principles (sec 2.3)

- Response variable $Y$ and predictors $X_1, X_2, \ldots, X_p$.
- Model building,
  - Specify (parametric) probability distribution of $Y$ (Normal, Poisson, etc.)
  - "Link" $E(Y)$ to predictors $X_1, X_2, \ldots, X_p$.

$$g(E(Y)) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

  - a function of the mean of $Y$ is a "linear component".
  - Parameter estimation: MLE, least squares, Bayes.
  - Model checking: consider model residuals.

- In *Linear regression* we use standardized residuals

$$r_i = \frac{(Y_i - \hat{Y}_i)}{\hat{\sigma}}.$$

  where $\hat{Y}_i$ is a fitted value and $\hat{\sigma}$ estimates the error SD.
- $Y_i \sim Poisson(\theta); i = 1, 2, \ldots, n$

$$r_i = \frac{(Y_i - \hat{\theta})}{\sqrt{\hat{\theta}}}$$

- Square root contribution to a *Pearson* goodness of fit statistic:

$$\sum_i (O_i - e_i)^2 / e_i$$

  where $O_i$ represents an observed value and $e_i$ an expected value.
- Exponential family of distributions.