Topic 16
Interval Estimation
Additional Topics

Outline

Linear Regression

For ordinary linear regression, we have given least squares estimates for the slope β and the intercept α. For data $(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$, our model is

$$y_i = \alpha + \beta x_i + \epsilon_i$$

where ϵ_i are independent $N(0, \sigma)$ random variables. Recall that the estimator for the slope

$$\hat{\beta}(x, y) = \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x)}$$

is unbiased.

Exercise. Show that the variance of $\hat{\beta}$ equals

$$\frac{\sigma^2}{(n-1)\mathrm{var}(x)}.$$

Linear Regression

Generally, σ is unknown. However, the variance of the residuals,

$$s_u^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - (\hat{\alpha} - \hat{\beta}x_i))^2$$

is an unbiased estimator of σ^2 and s_u/σ has a t distribution with $n-2$ degrees of freedom. This gives the t-interval

$$\hat{\beta} \pm t_{n-2,(1-\gamma)/2}\frac{s_u}{s_x\sqrt{n-1}}.$$

Exercise. For the data on the humerus and femur of the five specimens of *Archeopteryx*, we have $\hat{\beta} = 1.197$. $s_u = 1.982$, $s_x = 13.2$, and $t_{3,0.025} = 3.1824$, Use this to find a 95% confidence interval for the slope.

Sample Proportions

For n Bernoulli trials with success parameter p, the sample proportion $\hat{p}$ has

$$\text{mean} \quad p \quad \text{and} \quad \text{variance} \quad \frac{p(1-p)}{n}.$$

The parameter p appears in the variance. Thus, we need to make a choice $\tilde{p}$ to replace p in the confidence interval

$$\hat{p} \pm z_{(1-\gamma)/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}.$$

One simple choice for $\tilde{p}$ is $\hat{p}$. Based on extensive numerical experimentation, one recent popular choice is

$$\tilde{p} = \frac{x+2}{n+4}$$

where x is the number of successes.

Sample Proportions

In order for a normal random variable to be a good approximation to the binomial, we ask that the mean number of successes np and the mean number of failures $n(1-p)$ each be at least 10.

For Mendel's data the F_2 generation consisted 428 for the dominant allele green pods and 152 for the recessive allele yellow pods. Thus, the sample proportion of green pod alleles is

$$\hat{p} = \frac{428}{428+152} = 0.7379.$$

The confidence interval, using $\tilde{p} = 0.7363$ is

$$0.7379 \pm z_{(1-\gamma)/2}\sqrt{\frac{0.7363 \cdot 0.2637}{580}} = 0.7379 \pm z_{(1-\gamma)/2} 0.0183 = 0.7379 \pm 0.0426$$

for $\gamma = 0.98$, $z_{0.01} = 2.326$. Note that this interval contains the predicted value of $3/4$.

Sample Proportions

For the difference in two proportions p_1 and p_2 based on n_1 and n_2 independent trials. We have, for the difference $p_1 - p_2$, the confidence interval

$$\hat{p}_1 - \hat{p}_2 \pm z_{(1-\gamma)/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

Exercise. Let p_{2001} be the fraction of the US adult population that *opposed* same sex marriage in 2001 and let p_{2013} be the corresponding number in 2013. We have the following data from the Pew Research Center.

year	$\hat{p}_{\text{year}}$	n_{year}
2001	0.59	3181
2013	0.43	3001

Find the 95% confidence interval for the difference $p_{2013} - p_{2001}$.

Interpretation of the Confidence Interval

- The confidence interval for a parameter θ is based on two statistics
  - $\hat{\theta}_\ell(\mathbf{x})$, the lower end of the confidence interval and
  - $\hat{\theta}_u(\mathbf{x})$, the upper end of the confidence interval.
- As with all statistics, these two statistics *cannot* be based on the value of the parameter.
  - Their formulas are determined *in advance* of having the actual data.
- Thus, the term confidence can be related to the *production* of confidence intervals.
  - If we produce independent confidence intervals repeatedly, then
  - each time, we may either succeed or fail to include the true parameter in the confidence interval.
  - The inclusion of the parameter value in the confidence interval is a Bernoulli trial with success probability γ.

Interpretation of the Confidence Interval

Exercise. Below are 100 confidence interval built from simulating independent normal random variables and constructing 95% confidence intervals. Which fail to include the mean value - 0?