Stat 20: Discussion at section

B. M. Bolstad, bolstad@stat.berkeley.edu

Oct 13, 2003

Some confidence intervals

For the mean a level C confidence interval is given by

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

where σ is a known standard deviation and z^* is the value on the standard normal curve with area C between $-z^*$ and z^*. Note that is confidence interval is exact for normal populations and approximately correct if n is large for other cases.

A level C confidence interval for the population proportion is

$$\hat{p} \pm z^* \mathrm{SE}\left(\hat{p}\right)$$

where $\mathrm{SE}\left(\hat{p}\right) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ and z^* is as above.

A level C confidence interval for the difference between two proportions p_1 and p_2 is given by

$$\hat{p}_1 - \hat{p}_2 \pm z^* \mathrm{SE}\left(\hat{p}_1 - \hat{p}_2\right)$$

where z^* is as above. Note that $\mathrm{SE}\left(\hat{p}_1 - \hat{p}_2\right) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$.

Question 1

Interpret what is meant by a 95% confidence interval.

Answer: A 95% confidence interval means that 95% of the confidence intervals constructed using this method will contain the true population parameter. For example if we are making confidence intervals for the population parameter by taking a sample of size n then 95% of intervals of the form

$$\hat{p} \pm 1.96 \sqrt{\frac{p(1-p)}{n}}$$

would contain p the true population parameter.

Question 2

An entomologist samples a field for egg masses of a harmful insect by placing a yard square frame at random locations and carefullt examining the ground within the frame. A simple random sample of 75 pastoral locations in a particular county found egg masses in 13 locations. Give a 90% confidence interval for the proportion of all possible locations that are infested.

Answer: First note that the sample proportion is given by $\hat{p}`\frac{13}{75} = 0.173$ and $\text{SE}l\,(\hat{p}) = \sqrt{\frac{0.173(1-0.73)}{75}} = 0.044$. For a 90% confidence interval $z^* = 1.645$ and so out interval is given by $0.173 \pm 1.645(0.044)$ or more clearly (0.101,0.245).

Question 3

A study of 1711 cyclists 15 years of age or older who were fatally injured in bicycle accidents between 1987 and 1991 found that 386 had blood alcohol levels above 0.10%. Give a 99% confidence interval for the proportion of fatally injured cyclists who have blood alcohol level above 0.10%.

Answer: First note that the sample proportion is given by $\hat{p} = \frac{386}{1711} = 0.226$ and $\text{SE}l\,(\hat{p}) = \sqrt{\frac{0.226(1-0.226)}{1711}} = 0.010$. For a 99% confidence interval $z^* = 2.576$ and so out interval is given by$0.226 \pm 2.576(0.010)$ or more clearly (0.199,0.251).

Question 4

A researcher is interested in gender bias in introductory grammar textbooks. She samples 10 commonly used titles and counts the number of time the juvenile form is used when referring to each of the two genders eg boy or man, girl or woman. This table summarizes her results

Gender	n	X
F	60	48
M	132	52

Give a 90% confidence interval for the difference. Do you think there is a gender bias?

Answer: Let p_1 be the proportion of references to females using the juvenile form and p_2 be the proportion of references to males using the juvenile form. Our two sample estimates for the proportions are $\hat{p}_1 = \frac{48}{60} = 0.8$ and $\hat{p}_2 = \frac{52}{132} = 0.394$. So our point estimate of the difference is given by $\hat{p}_1 - \hat{p}_2 = 0.8 - 0.394 = 0.406$ and our estimate of the the standard error is given by

$$\text{SE}\,(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{0.8(1-0.8)}{60} + \frac{0.394(1-0.394)}{132}} = 0.067$$

For a 90% confidence interval $z^* = 1.645$. Remember that the confidence interval for the difference between two proportions is given by

$$\hat{p}_1 - \hat{p}_2 \pm z^*\text{SE}\,(\hat{p}_1 - \hat{p}_2)$$

and so in our case

$$.406 \pm 1.645(0.067)$$

leading to an interval $(0.296, 0.516)$. Since 0 is not inside this confidence interval we conclude that there is a difference in the proportion of references in the juvenile form between genders.

Question 5

Returning to the cyclists in Question 3 a further inspection of the data reveals a gender breakdown of the number of fatal bicycle accidents involving blood alcohol levels above 0.10%. The following table gives the number of each gender.

Gender	n	X
F	191	36
M	1520	350

Give a 95% confidence interval for the difference between males and females. Interpret your result.

Answer: Let p_1 be the proportion of fatally injured female cyclists having high blood alcohol levels and p_2 be the proportion of male cyclists (fatally injured) having a high blood alcohol level. Our two sample estimates for the proportions are $\hat{p}_1 = \frac{36}{191}$ and $\hat{p}_2 = \frac{350}{1520}$. So our point estimate of the difference is given by $\hat{p}_1 - \hat{p}_2 = -0.042$ and our estimate of the the standard error is given by $\text{SE}\,(\hat{p}_1 - \hat{p}_2) = 0.030$. For a 95% confidence interval $z^* = 1.96$. Remember that the confidence interval for the difference between two proportions is given by

$$\hat{p}_1 - \hat{p}_2 \pm z^* \text{SE}\,(\hat{p}_1 - \hat{p}_2)$$

and so in our case

$$-0.042 \pm 1.96(0.030)$$

leading to an interval $(-0.1013, 0.017)$. Since 0 is included in this interval we conclude that there is no difference in the proportion of cyclists who are fatally injured and have a high blood alcohol level between the two genders.