Steve Sawin
Statistics 3352

Problem Set VI (Solutions)

Due April 20.

All write-ups should be individually done, complete, neat and precise. Please tell me anyone you worked with or got help from. You may use our book and your notes, and me. Other sources are not permitted without my permission.

1. Suppose α is your confidence level and for convenience let $\beta = 1 - \alpha$ (so for a 95% confidence interval β would be 5%. Suppose we know σ and are trying to estimate μ from a normal population, using a simple random sample of size n with a sample mean of $\overline{x}$..

   (a) Show that the interval
   $$\left[\overline{x} - z_{\beta/3}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{2\beta/3}\frac{\sigma}{\sqrt{n}}\right]$$
   is an α confidence interval for μ (that is, show that the proportion of samples for which this interval contains μ is α). Here $z_{\beta/3}$ means the z-score such that $\beta/3$ of the data in a standard normal distribution falls above it and the remaining $2\beta/3$ falls below. This is called an "asymmetrical confidence interval." (Hint: what is the probability a normal variable will fall between $-z_{\beta/3}$ and $z_{2\beta/3}$?)

   (b) Show that the standard α confidence interval
   $$\left[\overline{x} - z_{\beta/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\beta/2}\frac{\sigma}{\sqrt{n}}\right]$$
   is necessarily shorter than the α confidence interval
   $$\left[\overline{x} - z_{\beta/3}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{2\beta/3}\frac{\sigma}{\sqrt{n}}\right].$$
   Hint: You will use the fact that the pdf of z is decreasing for $z > 0$ and increasing for $z < 0$. More specifically you should draw a graph of the pdf, mark off $z_{\beta/3}$, $z_{\beta/2}$, and $z_{2\beta/3}$ on the graph and compare areas to see
   $$z_{2\beta/3} - z_{\beta/2} > z_{\beta/2} - z_{\beta/3}.$$
   From there it will be pretty easy. The length of the first is $2z_{\beta/2}\sigma/\sqrt{n}$ and of the second is $(z_{\beta/3} + z_{2\beta/3})\sigma/\sqrt{n}$ so we need only show
   $$z_{\beta/3} + z_{2\beta/3} > 2z_{\beta/2}$$
   or equivalently
   $$z_{2\beta/3} - z_{\beta/2} > z_{\beta/2} - z_{\beta/3}.$$
   Now between $z_{2/beta/3}$ and $z_{\beta/2}$ on the graph of the pdf of the standard normal curve we know the area under the curve is exactly $\beta/6$. Also, because $f(z)$ is decreasing where f is the pdf for the standard normal (which I do not need to

write out, so I won't), we know that the curve lies below $f(z_{\beta/2})$. So that means if we draw a rectangle containing the area between $z_{2/beta/3}$ and $z_{\beta/2}$ we see

$$f(z_{\beta/2}) * (z_{2\beta/3} - z_{\beta/2}) > \frac{\beta}{6}.$$

On the other hand the area under the curve between $z_{\beta/2}$ and $z_{\beta/3}$ is also $\beta/6$, and if we draw a rectangle *inside* that area we see

$$f(z_{\beta/2}) * (z_{\beta/3} - z_{\beta/3}) < \frac{\beta}{6}.$$

Putting these two together and dividing by $f(z_{\beta/2})$ tells you the result.

2. Suppose again that we are estimating μ from a normal population with σ known, using a simple random sample with sample mean $\overline{x}$. Suppose you want to pick your sample size n so that the margin of error (which is half the total width of the confidence interval) is equal to e. Give a formula for the minimum n that will accomplish this as a function of σ, e and z^*_α. Remember $z^*_\alpha = z_{(1-\alpha)/2}$ is the z-score such that α is the chance a standard normal random variable will fall between $-z^*_\alpha$ and z^*_α. Since the α confidence interval has a margin of error

$$z^*_\alpha \sigma/\sqrt{n},$$

we have a probability α that

$$|\overline{x} - \mu| < z^*_\alpha \sigma/\sqrt{n}$$

and therefore if we need α confidence that

$$|\overline{x} - \mu| < e$$

it is sufficient to make sure

$$e > z^*_\alpha \sigma/\sqrt{n}.$$

Solving for n in this

$$\begin{aligned} e &> z^*_\alpha \sigma/\sqrt{n} \\ \sqrt{n} &> z^*_\alpha \sigma/e \\ n &> (z^*_\alpha)^2 \frac{\sigma^2}{e^2}. \end{aligned}$$

3. Question 11.12, p. 366: Show we can be at least α confident that $|\hat{\theta} - \theta|$ is less than an acceptable error e in estimating a population proprotion if we require

$$n > \frac{z^2_{(1-\alpha)/2}}{4e^2}.$$

Again we know that there is a probability α that

$$|\hat{\theta} - \theta| < z^*_\alpha \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

so to ensure it is less than e we need

$$e > z^*_\alpha \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}.$$

Solving for n gives

$$n > (z^*_\alpha)^2 \frac{\hat{\theta}(1-\hat{\theta})}{e^2}.$$

That is not satisfactory because it requires knowing what $\hat{\theta}$ is before we have even decided what n is going to be. This estimate is used when we have a good guess for $\hat{\theta}$ ahead of time. If we do not, we need replace $\hat{\theta}(1-\hat{\theta})$ with something bigger than it that does not depend on p. The biggest $\hat{\theta}(1-\hat{\theta})$ can be as we run through all $0 < \hat{\theta} < 1$ is found buy setting the derivative equal to 0, $2\hat{\theta} - 1 = 0 \implies p = 1/2$. So the function $\hat{\theta}(1-\hat{\theta})$ has a maximum at $\hat{\theta} = 1/2$, where it is 1/4. so $\hat{\theta}(1-\hat{\theta}) \leq 1/4$ and

$$n > \frac{(z^*_\alpha)^2}{4e^2}.$$

4. Question 11.24, p. 370: A study of the annual growth of certain cacti showed that 64 of them, selected at random in a desert region, grew an average of 52.80 mm with a standard deviation of 4.5 mm. Construct a 99% confidence interval for the true average annual growth of the given kind of cactus. Note we are given the *sample* standard deviation. so we have to use the t-procedure, not the z procedure appropriate when you know the *population* s.d. For 99% confidence with 63 degrees of freedom we get a critical t-value of 2.65614 (notice because n is so large this is nearly the same as the critical z value, which would have been 2.575, so those who used the σ know procedure got pretty close). The confidence interval is

$$\overline{x} \pm t\frac{s}{\sqrt{n}} = 52.80 \pm 1.49$$

or within the interval $[51.31, 54.29]$.

5. Question 11.38, p. 372: A sample survey at a supermarket showed that 204 of 300 shoppers regularly used cents-off coupons. Give a 95% confidence interval for the true proportion.

$$\frac{204}{300} \pm z^*_{.05}\sqrt{\frac{204/300(1-204/300)}{300}} = 68.00\% \pm 5.28\%.$$

6. Go to the file http://cs.fairfield.edu/šawin/MA217/Data/house_selling_prices.XLS for a sample of recently sold houses in some town. Give a 90% confidence interval for how much more houses with 2 or more baths sell for than those with fewer than 2. Number of Baths is column D, sale price is column G. Check all assumptions. Using the template to get an estimate without assuming same standard deviations we get

$$\$56{,}877 \pm \$14{,}565.$$

Since we do not know how this sample was take we will have to assume it was a simple random sample. The population of houses in the town would have to be bigger than $20 \cdot 100 = 2000$, which seems almost certainly true. The sample of large houses was plenty big enough for the 0:15:40 rule. The small houses had a sample of only 25, so we look at the histogram. This looks quite symmetric and gives no sign of outliers, so this assumption is met.

7. Go to the "High_Jump.XLS" file of the Data tab of my website cs.fairfield.edu/~sawin/217/data/High_Jump.XLS. The first two columns give the heights jumped by a sample of men and another sample of women athletes. Use the two sample mean confidence interval to give a 90% confidence interval for the amount by which the average height jumped by men exceeds the average height jumped by women. Tell me what you would need to know about the population sizes, variables and sampling process to know that the assumptions for using this procedure were met.
Again we use the two sample t procedure for the difference of means. Pasting the two columns in and setting the conf. level to 90% gives us that men jumped an averaqe of $0.25 \pm .09$ meters higher. Again we would need to know that the sample was a random sample of all athletes (I'm skeptical, but who knows), that the population had at least 500 men and 360 women (certainly if it is really all athletes, almost certainly even if it is much smaller population, say all American college high jumpers). Finally, since the samples are between 15 and 40 and highly skewed, it does not meet this assumption unless you knew that the population distributions were normal, which is pretty implausible.