Project 6 - Estimation

Statistics 401: Fall 2016

Due 1:40pm on Tuesday, November 1

Justify your answers. Feel free to use R for computations. As always, properly label figures and reference them from the body of your report.

1. (a) Why is an unbiased statistic or point estimator usually preferred over a biased statistic for estimating a population parameter?

   (b) Does unbiasedness alone guarantee that an estimate will be close to the true value of a parameter? Explain.

   (c) Under what circumstances might you choose a biased statistic over an unbiased statistic for estimating a population parameter?

2. The Environmental Protection Agency has established an air quality standard for lead of 1.5 μg/m^3. Listed below are measured amounts of lead (in micrograms per cubic meter or μg/m^3) in the air recorded at Building 5 of the World Trade Center site on different days immediately following the destruction caused by the terrorist attacks of September 11, 2001. After the collapse of the two World Trade Center buildings, there was considerable concern about the quality of the air. The data file "lead.txt" can be found on the Stat 401 website.

   5.40 1.10 0.42 0.73 0.51 1.10 0.66 1.02 0.45 0.69 0.72 0.55

   (a) Give a point estimate for σ. What point estimator did you use to obtain your estimate?

   (b) Give a point estimate of $\tilde{\mu}$. What point estimator did you use to obtain your estimate?

   (c) Give a point estimate of μ. What point estimator did you use to obtain your estimate?

   (d) Suppose that the lead data is normal. Then the true 95th percentile of the lead distribution is the parameter $\zeta = \mu + 1.645\sigma$ (i.e. 95% of the lead measurements are less than this value). Give (a) a point estimator for ζ, and (b) calculate a point estimate based on these data.

3. Give one advantage and one disadvantage of using a 99% confidence interval instead of a 90% confidence interval.

4. Read the March 2007 *Discover* article "Scents and Scents-Ability" available at the STAT401 web site. The following questions pertain to the first experiment in which thirty-two Berkeley undergrads volunteered.

   (a) Give the individuals being measured.

   (b) Give the variable being measured on each individual, and give the sample space of all possible outcomes.

(c) Give a point estimate for the true proportion of all humans who could "sniff their way along a scent trail" in 10 minutes.

(d) Are the population size and the sample size large enough to assume that the sample proportion $\hat{p}$ has an approximate normal distribution? Why or why not? Show all calculations and be sure to mention which theorem assures that your answer is correct.

(e) Using the estimate $p \approx \hat{p}$, give the sampling distribution of $\hat{p}$, being sure to give $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$.

(f) Construct a 95% CI for p.

(g) Interpret the CI in terms of the problem.

(h) To what population would it be reasonable to generalize the CI estimate?

5. The hallucinogenic effects of mushrooms that contain psilocybin have been studied for years. A study in 2007 found that such consuming mushrooms caused a "complete mystical experience" in 60% of subjects. Give the sample size required to estimate the true proportion of subjects who would experience a "complete mystical experience" after taking psilocybin with a margin of error of 5% and with 95% confidence.

6. Entomologists sometimes capture insects from the field to take them back to the lab. One means of doing this is putting the insects into a "kill jar" that contains an absorbant material (like cotton bars) soaked with finger nail polish (ethyl acetate). One insect is put into the jar - it takes 20 seconds to expire. A second insect is placed in the jar - it takes 38 seconds to expire.

   (a) Assuming that the two insects are a SRS, you will construct a 75% confidence interval for the true average time it'll take for the commute. Give the critical value that must be used to construct this 75% CI. *Hint*: Use R's **qt** function as outlined in the Chapter 4-6 notes on Estimation.

   (b) Assuming that these two guesses are a SRS, give a 75% confidence interval for the true average time it'll take for insects placed in the jar to expire.

   (c) What (besides being a SRS) must we assume about the data so that the 75% CI is valid?

   (d) Why is the 75% CI so wide?

   (e) How large of a SRS must be collected in order to construct a 90% confidence interval for the true average kill time with a margin of error of 5 seconds?

7. In the Center for Biofilm Engineering on MSU's campus, the thin layer (or slime) of bacteria that form on various surfaces, such as pipes and catheters, are studied. In experiments, microbiologists estimate the density (per mm^3) of bacteria that have formed to create a biofilm. The data file "bacteria.txt" can be found on the STAT 401 website, where the bacteria densities are in millions per mm^3.

   (a) We wish to construct a 95% CI for the center (i.e., either μ_X or $\tilde{\mu}_X$) of the distribution of the density of bacteria in a biofilm. What must be assumed about the data so that a CI for μ_X is valid? Why?

(b) Does the evidence suggest that the data are not normal? Use the techniques, including appropriate graphs and the correlation test from the Chapter 3 course notes to answer this question.

(c) Transform the data. Use Box-Cox to determine the appropriate transform. As in the Chapter 3 course notes, the R-command

```
boxcox(density ~ 1,plotit=TRUE,lambda=seq(-1.5,1.5,.01))
```

looks for the optimal λ value between -1.5 and 1.5. For this problem, you will need to use a larger range of λ's. Include the Box-cox plot in your report and specify which value of λ are you using for the transform.

(d) Use plots and the correlation test to make sure that your transform worked.

(e) Let X denote the original, untransformed data and let $Y = X^\lambda$ be the transformed data. Create a 95% CI for μ_Y

(f) Let X denote the original, untransformed data and let $Y = X^\lambda$ be the transformed data. Create a 95% CI for $\tilde{\mu}_X$, the median density of bacteria in the biofilm.

Start with the CI for μ_Y (the transformed mean from #7e) then back-transform the endpoints of the CI. For example, if you use a lambda of $\frac{1}{2}$ (square root) as the optimal power to transform the data to a normal distribution and the CI calculated based on the transformed data is (2.05, 4.26), then you can back-transform the end points by raising each end point to $\frac{1}{\lambda}$, or in this example, squaring each end point, so the CI on the original scale is (4.20, 18.15).

(g) Interpret the 95% CI for $\tilde{\mu}_X$ in the context of this problem.