Project 6 Solutions

Statistics 401: Fall 2016

Due Tuesday, November 1, 34 pts

1. (3 pts)

   (a) The center of the sampling distribution of an unbiased statistic is equal to the value of the parameter being estimated. Thus, unbiased statistics do not systematically over or under estimate a parameter and hence are preferred over biased statistics.

   (b) If the sampling variability of the unbiased statistic is large, then estimates will not tend to be "close" to the parameter value of interest, although the estimates will be centered on the parameter.

   (c) If the sampling variability of a biased statistic is a lot smaller than an unbiased statistic, and if the bias is small with respect to the sampling variability, then a biased statistic may be preferable to an unbiased statistic.

2. (5 pts) Regarding the lead data at the World Trade Center; see the Appendix for the R code that calculated the point estimates.

   (a) A point estimate for the population standard deviation σ is $1.37 \mu g/m^3$. The point estimator is the sample standard deviation S.

   (b) A point estimate for the population median $\tilde{\mu}$ is $0.705 \mu g/m^3$. The point estimator is the sample median $\tilde{x}$.

   (c) A point estimate for the population mean μ is $1.11 \mu g/m^3$. The point estimator is the sample mean $\bar{X}$.

   (d) If the lead data is normal, then $\bar{X} + 1.645\sigma$ is a point estimator for the parameter $\zeta = \mu + 1.645\sigma$. Thus, a point estimate for the 95th percentile is $1.11 + 1.645(1.37) = 3.36 \mu g/m^3$.

3. (2 pts) One advantage of using a 99% CI instead of a 90% CI is that you are more confident (more certain) that the parameter does lie in the CI. One disadvantage of using a 99% CI instead of a 90% CI is that the 99% CI will be wider than the 90% CI. Therefore, a 99% CI gives you a less precise interval estimate compared to a 90% CI.

4. (9 pts) Regarding the experiment in which thirty-two Berkeley undergrads volunteered, reported in March 2007 *Discover* article "Scents and Scents-Ability"

   (a) The individuals being measured are the Berkeley undergraduate students.

   (b) The variable being measured is whether or not the scent trail was followed successfully. The sample space is $S = \{\text{Yes, No}\}$.

   (c) A point estimate reported by *Discover* for the true proportion of all humans who could "sniff their way along a scent trail" is $p = \frac{2}{3}$.

   (d) The sample size is large because:

   - $n\hat{p} = 32\left(\frac{2}{3}\right) \approx 21 \geq 10$

- $n(1-\hat{p}) = 32\left(\frac{1}{3}\right) \approx 10 \geq 10$;

and the population is large because there are tens of thousands of Berkeley students which is much larger than 5% of the sample size of $n = 32$. So the Central Limit Theorem assures that the sample proportion $\hat{p}$ has an approximate normal distribution.

(e) The approximate normal distribution is $\hat{p} \dot{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$. Estimating p by $\hat{p}$, then $\hat{p} \dot{\sim} N\left(.\bar{6}, \sqrt{\frac{2/3(1-2/3)}{32}} \approx .08\bar{3}\right)$

(f) A 95% confidence interval for p is

$$\hat{p} \pm z_{.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{2}{3} \pm 1.96\sqrt{\frac{\frac{2}{3}(1-\frac{2}{3})}{32}} = \frac{2}{3} \pm 0.16\bar{3} = (0.5033, 0.8300).$$

.

(g) We are 95% confident that the true percentage of Berkeley undergraduates who can follow a scent trail is between 50% and 83%.

(h) Since the experiment was carried out on Berkeley undergraduates, then it may be safe to generalize the results to undergraduates at Berkeley only.

5. (1 pt) To estimate the true proportion of humans who would experience a "complete mystical experience" after taking psilocybin with a margin of error of 0.05 and 95% confidence, we need

$$n = .6(1-.6)\left(\frac{1.96}{.05}\right)^2 \approx 368.79.$$

Thus, recommend that your advisor recruit 369 students for an experiment at MSU.

6. (5 pts) Time to expire in an insect killing jar:

(a) The critical value to be used for a 75% CI is is 2.4142. This is because $\alpha = 25\%$, and so $1 - \alpha/2 = 0.875$, which means that the multiplier used in the CI is the 87.5th percentile from a t distribution with a measly $n-1 = 1$ degree of freedom, $t_{0.875,df=1} = 2.4142$.

(b) Assuming that the two insects are a SRS, a 75% confidence interval for the true average time, in seconds, that it'll take for the commute is

$$\bar{X} \pm t_{0.875,df=1}S/\sqrt{n} = 29 \pm 2.4142\frac{12.73}{\sqrt{2}} = (7.27,\ 50.73).$$

(c) Since the sample size is so small, we must assume that the kill times of the insects are from a normal distribution so that the 75% CI is valid.

(d) The 75% CI extremely wide (and uninformative) due to the tiny sample size.

(e) A SRS of size 18 insects 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:

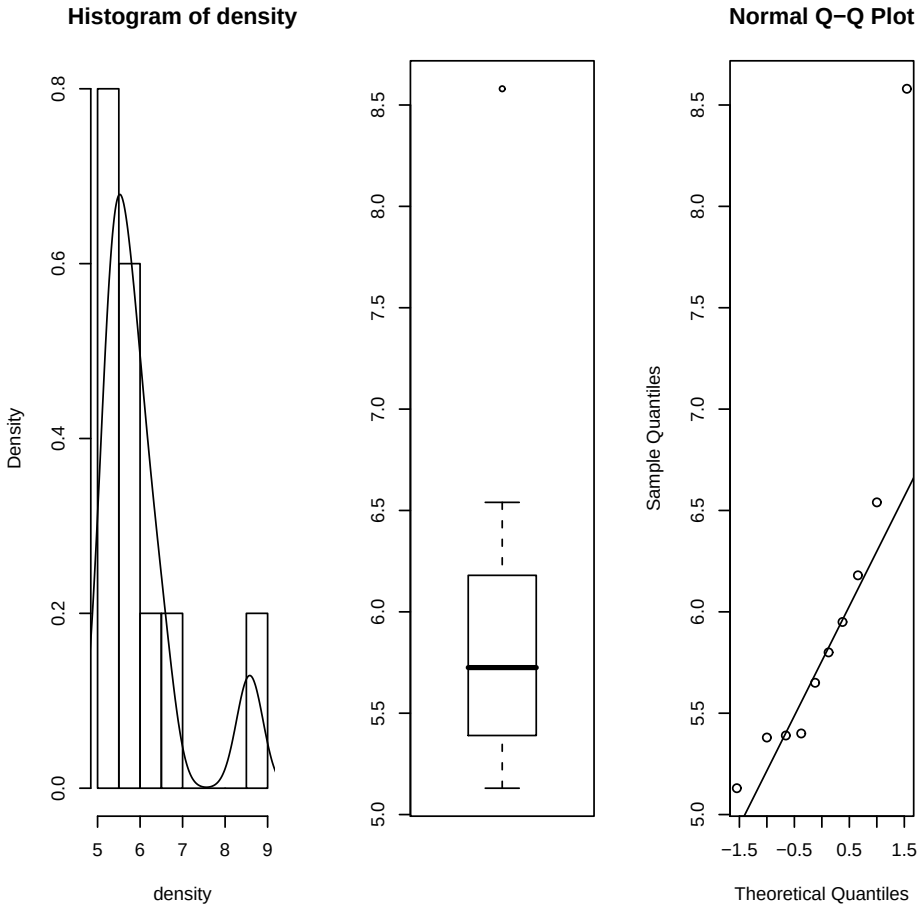
$$\left(\frac{1.645(12.89)}{5}\right)^2 \approx 17.54.$$

7. (9 pts) Biofilms:

(a) Since the sample size is small, $n = 10 < 15$, then the Central Limit Theorem does not apply. Thus, the 95% CI for μ_X is valid only if the data is normal.

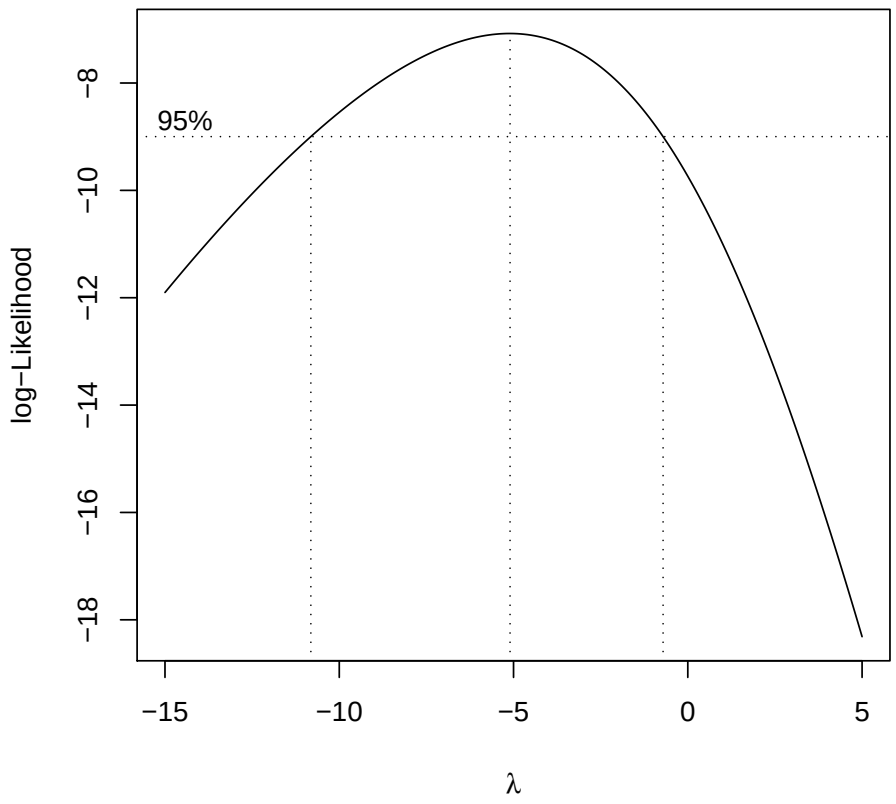
(b) Yes, the evidence DOES suggest that the data is not normal. Figure 1 displays the distribution plots (density plot, boxplot, and normal probability plot) of the bacteria densities (in millions per mm^3). They all indicate a right skew and that the population distribution is non-normal. Furthermore, the correlation coefficient between the untransformed bacteria densities and the normal scores is 0.85, which is less than $r_{\text{critical}} = 0.88$ for $n = 10$. Therefore, the untransformed bacteria densities do not appear to be normally distributed.

Figure 1: Diagnostic plots checking the bacteria densities for normality.



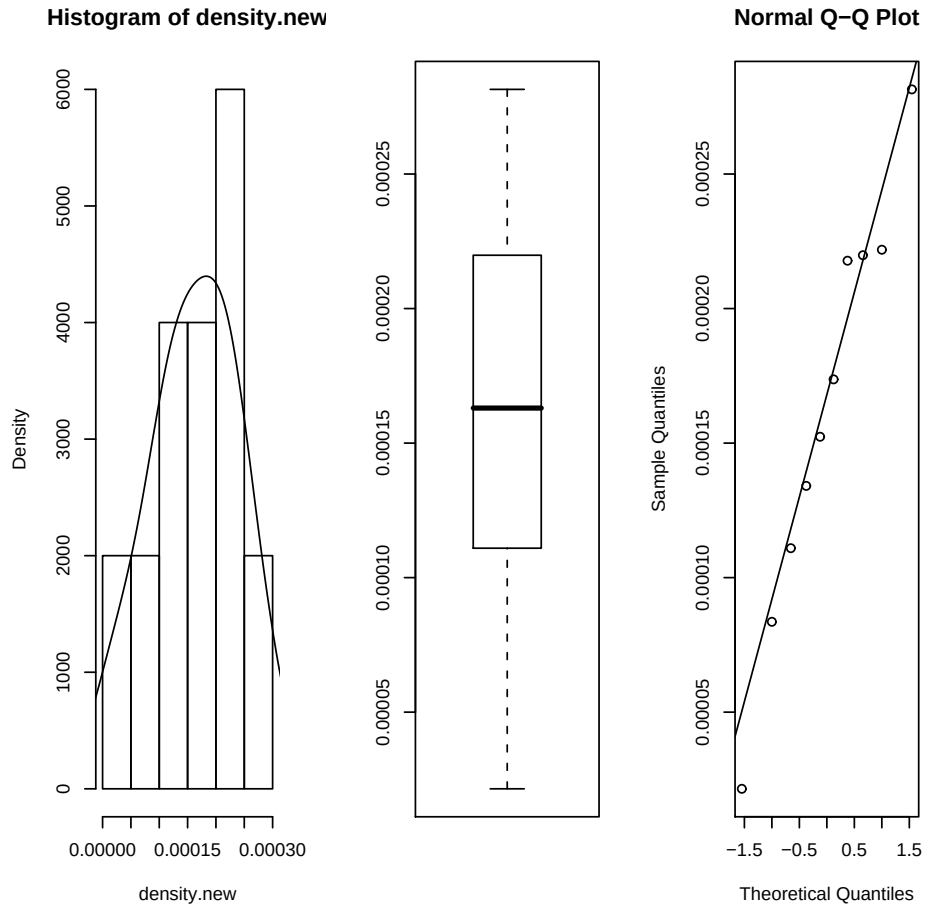
(c) According to the Box-Cox method (see Figure 2), the optimal transform corresponds to $\lambda = -5$, so the transformed data is $Y = \frac{1}{X^5}$.

Figure 2: Box-Cox log Likelihood for Values of λ between -15 and 5.



(d) To be sure that the transform worked, the same diagnostic plots and correlation coefficient as in #7b are performed. Figure 3 displays the density plot, boxplot, and normal probability plot of the transformed bacteria densities. The transformed data appear to be normal. Furthermore, The correlation coefficient between the transformed bacteria densities and the normal scores is 0.9869, which is greater than $r_{\text{critical}} = .88$ for $n = 10$. Therefore, the evidence fails to suggest that the transformed data is non-normal. It looks like the transform worked.

Figure 3: Diagnostic plots checking the transformed bacteria densities for normality.



(e) Let X denote the original, untransformed data and let $Y = X^{-5}$ be the transformed data. A 95% confidence interval for μ_y is $\bar{y} \pm 2.26\frac{s_y}{\sqrt{10}} = 1.6171 \times 10^{-4} \pm 2.26\frac{7.7325\times10^{-5}}{\sqrt{10}} = (1.06 \times 10^{-4}, 2.17 \times 10^{-4})$.

(f) Back-transforming the 95% CI for μ_y gives a 95% confidence interval for the median $\tilde{\mu}_x$ by calculating $\left((2.1702 \times 10^{-4})^{\frac{1}{\lambda}}, (1.0639 \times 10^{-4})^{\frac{1}{\lambda}}\right) = \left((2.1702 \times 10^{-4})^{\frac{1}{-5}}, (1.0639 \times 10^{-4})^{\frac{1}{-5}}\right) = (5.40, 6.23)$.

(g) I am 95% confident that the true <u>median</u> density of bacteria in the biofilm is between 5.4 and 6.23 million per mm^3.

Appendix A

```
# PROBLEM 2

> lead
 [1] 5.40 1.10 0.42 0.73 0.51 1.10 0.66 1.02 0.45 0.69 0.72 0.55

> sd(lead)
[1] 1.370900

> median(lead)
[1] 0.705

> mean(lead)
[1] 1.1125

> mean(lead)+1.96*sd(lead)
[1] 3.799463



# PROBLEM 6

> qt(p=.75 + .125,df=1)
[1] 2.414214

> guess=c(20,38)

# CI by hand
> mean(guess) + c(-1,1)*qt(p=.875,df=1)*sd(guess)/sqrt(2)
[1]  7.272078 50.727922

# CI using R's t.test function
> t.test(guess,conf.level=.75)

        One Sample t-test

data:  guess
t = 3.2222, df = 1, p-value = 0.1916
alternative hypothesis: true mean is not equal to 0
75 percent confidence interval:
  7.272078 50.727922
sample estimates:
mean of x
       29

# Sample size calculation
> (1.645*sd(guess)/5)^2
```

```
[1] 17.53504


# PROBLEM 7
# Read in the data
> bac=read.table("project6bacteria.txt",header=TRUE)
> bac
   density
1     5.13
2     5.38
3     5.39
4     5.40
5     5.65
6     5.80
7     5.95
8     6.18
9     6.54
10    8.58
> attach(bac)

# Check for normality
> par(mfrow=c(1,3))
> hist(density,freq=FALSE)
> lines(density(density))
> boxplot(density)
> qqnorm(density)
> qqline(density)

> xy=qqnorm(density)
> cor(xy$x,xy$y)
[1] 0.8545355


 # Box-Cox transform
> boxcox(density ~ 1,plotit=TRUE,lambda=seq(-15,5,.1))
> density.new=density^(-5)


# Check the transformed data for normality
> par(mfrow=c(1,3))
> hist(density.new,freq=FALSE)
> lines(density(density.new))
> boxplot(density.new)
> qqnorm(density.new)
> qqline(density.new)
> xy.new=qqnorm(density.new)
> cor(xy.new$x,xy.new$y)
```

```
[1] 0.9869124

# 95% CI for mu_y
> mean(density.new) + c(-1,1)*qt(.975,df=9)*sd(density.new)/sqrt(10)
[1] 0.0001063932 0.0002170225

# 95% CI for mu_x
> (mean(density.new) + c(-1,1)*qt(.975,df=9)*sd(density.new)/sqrt(10))^(-1/5)
[1] 6.231853 5.403797

# Confirms that the transform is appropriate
> var(density)/mean(density)^2
[1] 0.0278358
```