

Project 6 Solutions

Statistics 401: Fall 2006

1. (Problem 9.2 on page 367) The center of an unbiased statistic is equal to the value of the parameter being estimated. Thus, unbiased statistics are preferred over an unbiased statistics. If the sampling variability of the unbiased statistic is large, then estimates will still not tend to be “close” to the parameter value of interest, although the estimates will be centered on the parameter. If the sampling variability of a biased statistic were a lot smaller than an unbiased statistic, and the bias was small with respect to the sampling variability, then a biased statistic may be preferable to an unbiased statistic.
2. Problem 9.8 on page 368:
 - (a) I would use the sample standard deviation to estimate σ . For this sample, $s = 1.886$.
 - (b) I would use the sample median to estimate $\tilde{\mu}$. For this sample, $\tilde{x} = 11.35$.
 - (c) The sample median can be used to estimate μ if the distribution is symmetric and heavy tailed. For this sample, $\tilde{x} = 11.35$. You could also used trimmed means. The 10% trimmed mean, which drops 10.1 and 16.2 from the data set and then computes the average, is $\bar{x}_{(10)} = 11.73$. The 20% trimmed mean, which drops 10.1, 10.5, 15.2 and 16.2 from the data set and then computes the average, is $\bar{x}_{(20)} = 11.45$.
 - (d) The sample mean is $\bar{x} = 11.967$. So the point estimate for the 90th percentile is $\bar{x} + 1.28s = 14.38$.
3. Problem 9.14 on page 379:
 - (a) Increasing the confidence level increases the width of the confidence interval.
 - (b) Increasing the sample size decreases the width of the confidence interval.
 - (c) The confidence interval width is largest for $p = 0.5$. As p moves away from 0.5, in either direction, the width of the confidence interval decreases.
4. 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.
5. Do problem 9.16 on page 379.
 - (a) This is an observational study since no treatments are being imposed.
 - (b) The individuals being measured are girls younger than 18 seeking services at Planed Parenthood in Wisconsin.
 - (c) The variable being measured is whether or not the girl will stop going to Planned Parenthood. The sample space of all possible outcomes is $S = \{\text{Stop}, \text{Contiinue}\}$.

- (d) A point estimate for the proportion of girls who would stop using Planned Parenthood if their parents were informed is $p = \frac{55}{118} \approx 0.4661$.
- (e) To determine if the sample size is large enough to assume that the sample proportion p has an approximate normal distribution, check:
- $n\pi = 55 \geq 10$
 - $n(1 - p) = 63 \geq 10$

Thus, the sample proportion p has an approximate normal distribution, $p \sim N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$. For $\pi = .4$, this becomes $p \sim N\left(.4, \sqrt{\frac{.4(1-.4)}{118}} \approx .0451\right)$

- (f) The 95% confidence interval for π is

$$p \pm 1.96\sqrt{\frac{p(1-p)}{n}} = 0.4661 \pm 0.09 = (0.3761, 0.5561).$$

- (g) We are 95% confident that the true proportion of girls that would stop using Planned Parenthood if their parents were informed is between 38% and 56%.
- (h) It would be reasonable to apply this estimate only to girls younger than 18 who are seeking services at Planned Parenthood clinics in Wisconsin.

6. (Problem 9.20 on page 379) The two calculations are

- $m = 0.05$, $z_{0.975} = 1.96$, $\pi \approx 0.27 \longrightarrow n \geq \frac{1.96^2(0.27)(1-0.27)}{0.05^2} = 302.87 \longrightarrow$ **Use n=303.**
- $m = 0.05$, $z_{0.975} = 1.96$, $\pi \approx 0.5 \longrightarrow n \geq \frac{1.96^2(0.5)(1-0.5)}{0.05^2} = 384.16 \longrightarrow$ **Use n=385.**

One should use the sample size of $n = 385$ because if we suspect that the true proportion in the area may be larger than the 0.27 estimate (but less than 0.73) and thus the $n = 303$ sample size would not be large enough to meet the margin of error criteria for a 95% CI.

7. Problem 9.42 on page 393:

- (a) Table 1 has 95% confidence intervals for each of the three triathlon events. Since the critical value is $t_{df=8, .975} = 2.31$, then the three CI's were computed using the formula: $\bar{X} \pm 2.31 \frac{s}{\sqrt{9}}$.

Table 1: 95% CI's for triathlon events

Event	\bar{x}	s	95% CI
Swimming	188	7.2	(185.6, 190.4)
Biking	186	8.5	(183.2, 188.8)
Running	194	7.8	(191.4, 196.6)

- (b) Even though the sample mean for maximum heart rate while running is larger than the others, all three CI's overlap, and so it is unclear whether the mean maximum heart during any one triathlon event is higher than the others. We will see later that we will need to perform an Analysis of Variance (ANOVA) to really answer this question.

- (c) Since we only have 9 male triathletes, then the Central Limit Theorem does not apply, and any conclusions based on these CI's are questionable.
- (d) We must assume that the data is normal so that the conclusions from these CI's are valid.
- (e) Assuming that the data is normal, then we are 95% confident that the average maximum heart rate of running male triathletes is between 191.4 and 196.6 beats per minute.
8. (Problem 9.46 on page 391) Estimating σ as $(700 - 50)/4 = 162.5$, we see that $n = \left(\frac{1.96(162.5)}{0.1}\right)^2 = 1014.423$. So $n = 1015$ wine specimens ought to be collected.
9. In the Discover article *Malaria Parasite Makes Humans Smell More Attractive to Mosquitoes* on the STAT401 website, researchers observe 100 mosquitos and find that 67 of them are attracted to kids carrying gametocytes - the transmissible, reproductive stage of the parasite *P. falciparum* - in their bloodstream while only 33 mosquitos were attracted to other kids.
- (a) Since $np = 67 \geq 10$ and $n(1 - p) = 33 \geq 10$ then the sample proportion p has an approximate normal distribution.
- (b) $.67 \pm 2.576\sqrt{\frac{.67(1-.67)}{100}} \approx .67 \pm 2.576(.047) = .67 \pm 0.1207 = (0.5493, 0.7907)$.
- (c) We are 99% confident that the true proportion of mosquitos to be attracted to kids with gametocytes versus kids without gametocytes is between 55% and 79%.
- (d) We are 99% confident that a majority of mosquitos prefer kids with gametocytes versus other kids since the entire 99%CI is larger than .5.
- (e) The margin of error is $2.576\sqrt{\frac{p(1-p)}{100}}$. Half of this is $2.576\sqrt{\frac{p(1-p)}{4(100)}} = 2.576\sqrt{\frac{p(1-p)}{4(100)}}$. Thus, if the researchers want to cut the margin of error in half, they should observe $n = 400$ mosquitos.
10. The Environmental Protection Agency has established an air quality standard for lead of $1.5 \mu\text{g}/\text{m}^3$. Listed below are measured amounts of lead (in micrograms per cubic meter or $\mu\text{g}/\text{m}^3$) 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) If the data is not normal, then the sample size of 12 is not large enough to assume that the sample mean \bar{X} is approximately normal since the Central Limit Theorem does not apply. Thus to proceed, we must assume that the data is normal.

- (b) The evidence suggests that the population distribution is not normal. Figure 1 displays the distribution plots (density plot, boxplot, and normal probability plot) of the lead concentrations. They all indicate a right skew and that the population distribution is non-normal. Furthermore, the correlation coefficient between the untransformed lead concentrations and the normal scores is 0.678519, which is less than $r_{\text{critical}} = .911$ for $n = 15$. Therefore, the untransformed lead concentrations do not appear to be normally distributed.
- (c) By the previous discussion, a transform of the data is necessary. Figure 2 displays the Box-Cox plot for the optimal power transformation. The chosen lambda is -1 (reciprocal of the data).
- (d) To be sure that the transform worked, the same diagnostic plots and correlation coefficient as in #10b are performed. Figure 3 displays the density plot, boxplot, and normal probability plot of the transformed lead concentrations. The transformed data appear to be normal. Furthermore, The correlation coefficient between the transformed lead concentrations and the normal scores is 0.9839, which is greater than $r_{\text{critical}} = .911$ for $n = 15$. Therefore, the evidence fails to suggest that the transformed data is non-normal. It looks like the transform worked.

Figure 1: Distribution Plots of Lead Concentration Data

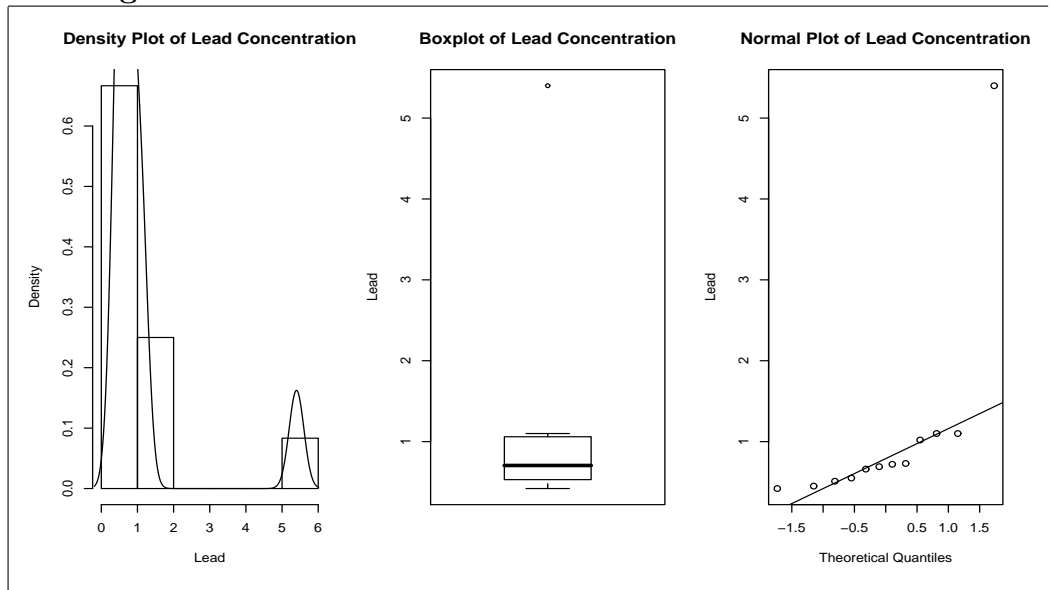


Figure 2: Likelihood Function for Various Values of λ

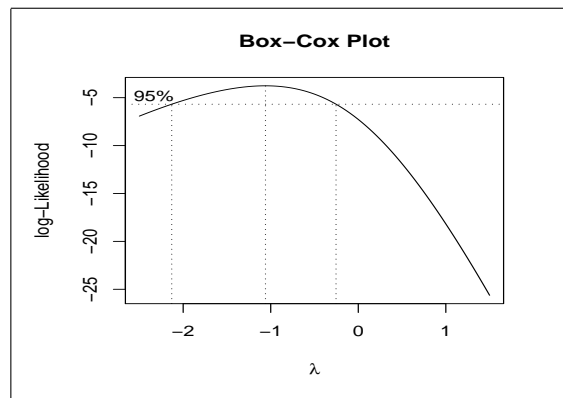
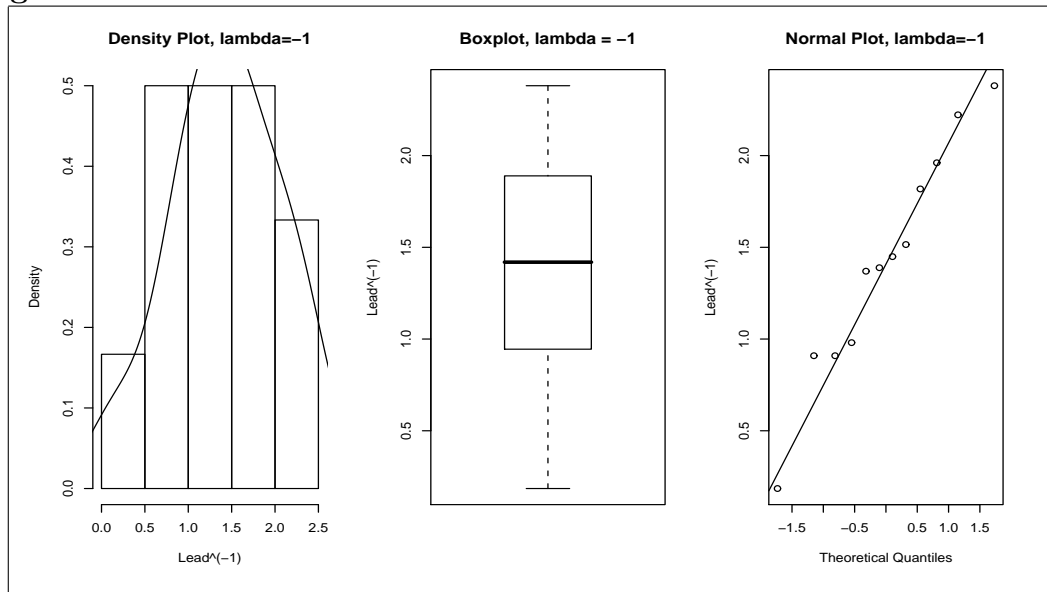


Figure 3: Distribution Plots of Transformed Lead Concentration Data



- (e) Let X denote the original, untransformed data and let $Y = X^{-1}$ be the transformed data. A 95% confidence interval for μ_y is $\bar{y} \pm 2.2 \frac{s_y}{\sqrt{12}} \approx (1.03, 1.82)$. Back-transform this CI to get a 95% confidence interval for μ_x by calculating $(1.03, 1.82)^{\frac{1}{\lambda}} = (1.03, 1.82)^{-1} = (.55, .97)$. Table 2 contains the sample mean of the untransformed data (\bar{x}), the sample standard deviation of the untransformed data (s_x) and the CI estimate for μ_x .

Table 2: Estimating the average lead content

\bar{x}	s_x	95% CI
1.1125	1.3709	(0.55, 0.97)

- (f) Back-transforming the CI yields a dubious CI for μ_x since $\frac{\sigma_x^2}{\mu_x^2} \approx \frac{s_x^2}{\bar{x}} = \frac{1.3709^2}{1.1125^2} \approx 1.5184$ is not less than 1.
- (g) By (f), it appears that our CI is not appropriate. If the confidence interval were legitimate, then we could say that we are 95% confident the true mean amount of lead in the air at the World Trade Center on the days immediately following 9/11 is between 0.55 and 0.97 $\mu\text{g}/\text{m}^3$.
- (h) Although we can not use the CI to get a statement of significance, only the single outlier of 5.4 $\mu\text{g}/\text{m}^3$ is larger than 1.5 $\mu\text{g}/\text{m}^3$. The rest of the data set is well below 1.5. This suggests that that the EPA standard for the amount of lead in the air may be met.

Appendix A

```
# Problem 2
> pine<-c(11.3,10.7,12.4,15.2,10.1,12.1,16.2,10.5,11.4,11,10.7,12)
> pine
 [1] 11.3 10.7 12.4 15.2 10.1 12.1 16.2 10.5 11.4 11.0 10.7 12.0
> sd(pine)
 [1] 1.885993
> median(pine)
 [1] 11.35
> mean(pine)
 [1] 11.96667
> mean(pine) + c(-1,1)*sd(pine)
 [1] 10.08067 13.85266

# Problem 10
D = read.table("lead.txt",header=TRUE)
attach(D)
library(lattice)
library(pastecs)
library(MASS)

# 10b
par(mfrow = c(1,3))
hist(lead,freq=FALSE,main="Density Plot of Lead Concentration",xlab="Lead")
lines(density(lead))
boxplot(lead,main="Boxplot of Lead Concentration",ylab="Lead")
qqnorm(lead,main="Normal Plot of Lead Concentration",ylab="Lead")
qqline(lead)

xy=qqnorm(lead)
cor(xy$x,xy$y)

# 10c
par(mfrow = c(1,1))
boxcox(lead~1,plotit=TRUE,lambda=seq(-2.5,1.5,0.01))
title(main="Box-Cox Plot")

lead.transformed=lead^(-1)

# 10d
par(mfrow = c(1,3))
hist(lead.transformed,freq=FALSE,main="Density Plot, lambda=-1",xlab="Lead^(-1)")
lines(density(lead.transformed))
boxplot(lead.transformed,main="Boxplot, lambda = -1",ylab="Lead^(-1)")
qqnorm(lead.transformed,main="Normal Plot, lambda=-1",ylab="Lead^(-1)")
```

```

qqline(lead.transformed)

xy.new=qqnorm(lead.transformed)
cor(xy.new$x,xy.new$y)

# 10e
xbar.trans = mean(lead.transformed)
xbar.trans
[1] 1.4240899
s.trans = sd(lead.transformed)
s.trans
[1] 0.6231341
n = length(lead)
n
[1] 12
lower.transformed = xbar.trans - qt(0.975,11)*s.trans/sqrt(n)
lower.transformed
[1] 1.028378

upper.transformed = xbar.trans + qt(0.975,11)*s.trans/sqrt(n)
upper.transformed
[1] 1.819822

lower.transformed^(-1)
[1] 0.9724051

upper.transformed^(-1)
[1] 0.5495043

# 10f
> sd(lead)^2/mean(lead)^2
[1] 1.518488

```