Experimental Statistics for Engineers I

Kamala Dadashova

Problem 1: Beginning with the probability statement $P(\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} > -t_\alpha) = 1-\alpha$, derive the expression for the exact upper bound of a one-sided confidence interval for a sample from a normal distribution.
Solution:

$$P(\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} > -t_\alpha) = 1-\alpha$$

$$P(\bar{X}-\mu > -t_\alpha \frac{S}{\sqrt{n}}) = 1-\alpha$$

$$P(-\mu > -\bar{X} - t_\alpha \frac{S}{\sqrt{n}}) = 1-\alpha$$

$$P(\mu < \bar{X} + t_\alpha \frac{S}{\sqrt{n}}) = 1-\alpha$$

After collecting data, $100(1-\alpha)\%$ upper confidence bound for μ is approximately $\bar{x} + t_\alpha \frac{S}{\sqrt{n}}$. CI is $(-\infty, \bar{x} + t_\alpha \frac{S}{\sqrt{n}})$.

Problem 2: Newly hatched trout are known as "fry." Suppose that the lengths of the trout fry in a pond at the fish hatchery have the population standard deviation of 0.8 inch. A random sample of 49 fry is netted and their lengths measured, and the sample mean is found to be $\bar{x} = 3.4$ inches.

(a) Construct and interpret a 95% confidence interval for the overall mean length of the trout fry in the pond.

Solution:

95% CI for μ is computed as following way,

$$(\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}})$$

We have $\bar{x} = 3.4$, $n = 49$, $\sigma = 0.8$. Plugging all values in the above expression, we get the confidence interval is $(3.176, 3.624)$. The interpretation is that we are doing sampling which

each sample consists of 49 fry and for each sample we find $\bar{X}$; we say that 95% of those samples along with mean values include true mean value. That is, we can be highly confident, at the 95% confidence level, that $3.176 < \mu < 3.624$).

(b) What is the minimum sample size required obtaining a 90% confidence interval with width 0.4 inches?

Solution:

Since the width is equal to 0.4 inches, we are interested in

$$\bar{X} + z_{0.1/2} \cdot \frac{\sigma}{\sqrt{n}} - (\bar{X} - z_{0.1/2} \cdot \frac{\sigma}{\sqrt{n}}) = 0.4$$

$$2 \cdot 1.645 \cdot \frac{0.8}{\sqrt{n}} = 0.4$$

$$\sqrt{n} = 6.58$$

$$n = 43.2964$$

Thus, minimum sample size is 43.

(c) Some of the researches only care that the mean length of trout fry in the pond is greater than 3 inches. Can the researchers say this with 95% confidence? Justify your answer.

Solution:

We first find

$$\bar{x} - z_{\alpha} \cdot \frac{\sigma}{\sqrt{n}} < \mu$$

$$3.4 - z_{0.05} \cdot \frac{0.8}{7} < \mu$$

$$3.4 - 1.645 \cdot \frac{0.8}{7} < \mu$$

$$3.212 < \mu$$

Thus, it is reasonable that the researchers say the mean length of trout fry in the pond is greater than 3 inches with 95% confidence.

Problem 3: A meteorologist who samples 13 thunderstorms found that the average speed at which they traveled across a certain state was 15 miles per hour. The standard deviation of the sample was 1.7 miles per hour. From past experience, the meteorologist knows the average storm speed is normally distributed.

(a) Find and interpret the 99% confidence interval of the mean.

Solution: CI is computed as below,

$$(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}})$$

We have $\bar{x} = 15$ $n = 13$ $\sigma = 1.7$ and $z_{0.01/2} = 2.580$. Plugging all values in above expression, we get (13.7835, 16.2165). That is, we can be highly confident, at the 99% confidence level, that $13.7835 < \mu < 16.2165$).

(b) Determine and interpret the 95% prediction interval

Solution:

The prediction interval is computed as the following way,

$$(\bar{x} - t_{\alpha/2,n-1} \cdot s\sqrt{1 + \frac{1}{n}},\ \bar{x} + t_{\alpha/2,n-1} \cdot s\sqrt{1 + \frac{1}{n}})$$

where $t_{\alpha/2,n-1} = t_{0.025,13-1} = qt(p = 0.025, df = 12) = 2.178813$. Plugging all values, we obatin

$$(15 - 2.178813 \cdot 1.7\sqrt{1 + \frac{1}{13}},\ 15 + 2.178813 \cdot 1.7\sqrt{1 + \frac{1}{13}}) = (11.1562, 18.8438).$$

The interpretation is that following to collecting a sample of observations and calculating a 95% prediction interval, there exists a 95% probability that a future observation will be in the prediction interval. Moreover, there is also a 5% probability that the subsequent observation will not be contained within the interval.

(c) Can the meteorologist be 95% confident that the mean thunderstorm speed is less than 16 miles per hour? Justify your answer.

Solution:

We will calculate

$$\mu < \bar{x} + z_{\alpha} \cdot \frac{\sigma}{\sqrt{n}}$$
$$\mu < 15 + z_{0.05} \cdot \frac{1.7}{\sqrt{13}}$$
$$\mu < 15 + 1.645 \cdot \frac{1.7}{\sqrt{13}}$$
$$\mu < 15.7756095494$$

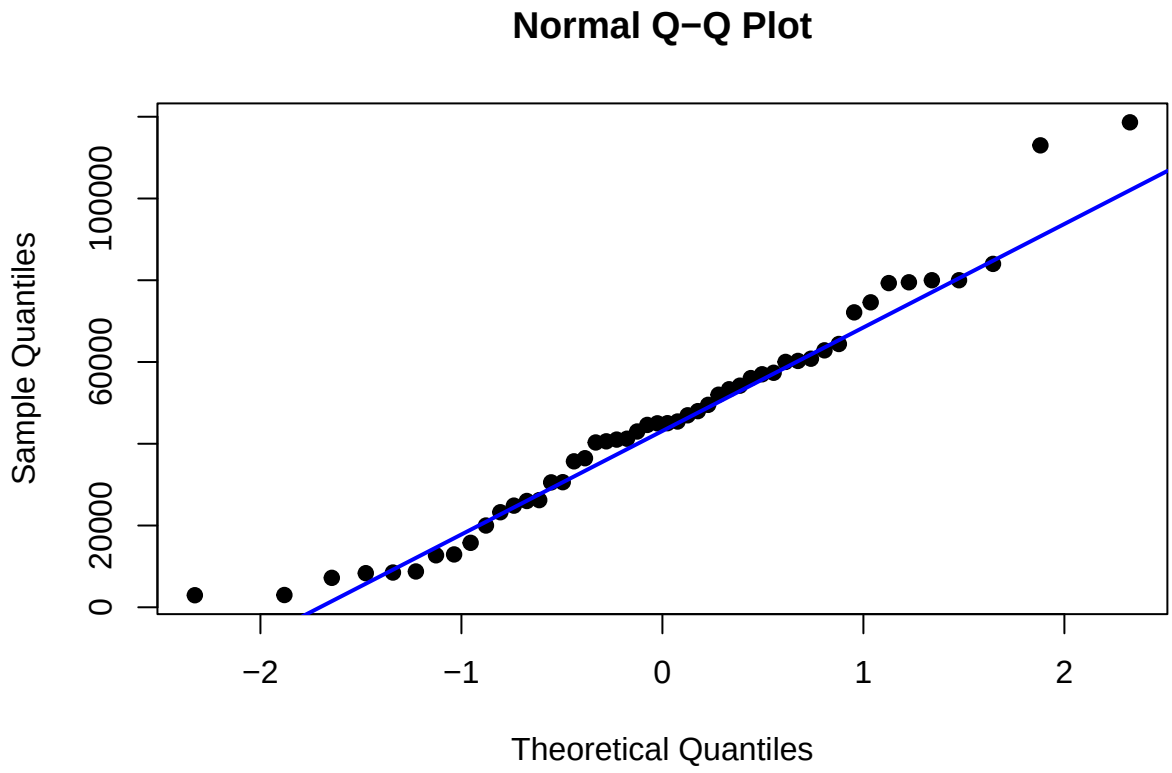
Thus, meteorologist' opinion about 95% confident that the mean thunderstorm speed is less than 16 miles per hour is true.

Codes

Problem 4: The mileage data sample from Example 7.6 in Devore is contained in a comma separated value file called mileage.csv. Use this data to do the following in R:

(a) Make a Q-Q plot of the data in the sample. Based on your observations, can we assume this data came from a normal distribution?

```
mileages <- read.csv("~/mileage.)
mile = mileages$mileage
n = length(mile)
qqnorm(mile, pch=19)
qqline(mile, lwd=2, col="blue")
```



As we observed, data has a normal distribution.

(b) Calculate a large sample approximate confidence interval for the mean and an exact confidence interval for the mean based on the assumption the data is normally distributed. Compare the results. What does this say about your conclusion in (a). ?

```
# Large sample approximation.
approx <- c(mean(mile)-1.96*sd(mile)/sqrt(n),mean(mile)+1.96*sd(mile)/sqrt(n))
```

```
names(approx) = c("lower","upper")
approx
```

```
##    lower    upper
## 38294.66 53064.06
```

```
#Exact.
t.crit <- qt(0.025,n-1,lower.tail=F)
exact <- c(mean(mile)-t.crit*sd(mile)/sqrt(n),mean(mile)+t.crit*sd(mile)/sqrt(n))
names(exact) = c("lower","upper")
exact
```

```
##    lower    upper
## 38107.88 53250.84
```

In terms of comparison, we observe that the there is a small difference in lower bounds and in upper bounds for both large and exact sample approximation implying that our assumption in part (a) about normal distribution holds true.

(c) Generate a bootstrap percentile confidence interval for the mean from the sample data.

```
B <- 10000  # run 10000 bootstrap samples
means <- rep(0,B)  # create dummy matric for means
for(i in 1:B){
  rsample <- sample(mile,n,replace=T)
  means[i] <- mean(rsample)  # calculate mean and save in vector
}
x_bar_bar <- mean(means)  # calculate mean estimate


# t-CI with bootstrap SE
tboot <- c(mean(mile)-qt(0.025,n-1,lower.tail=F)*sd(means),
           mean(mile)+qt(0.025,n-1,lower.tail=F)*sd(means))
names(tboot) = c("lower", "upper")
tboot
```

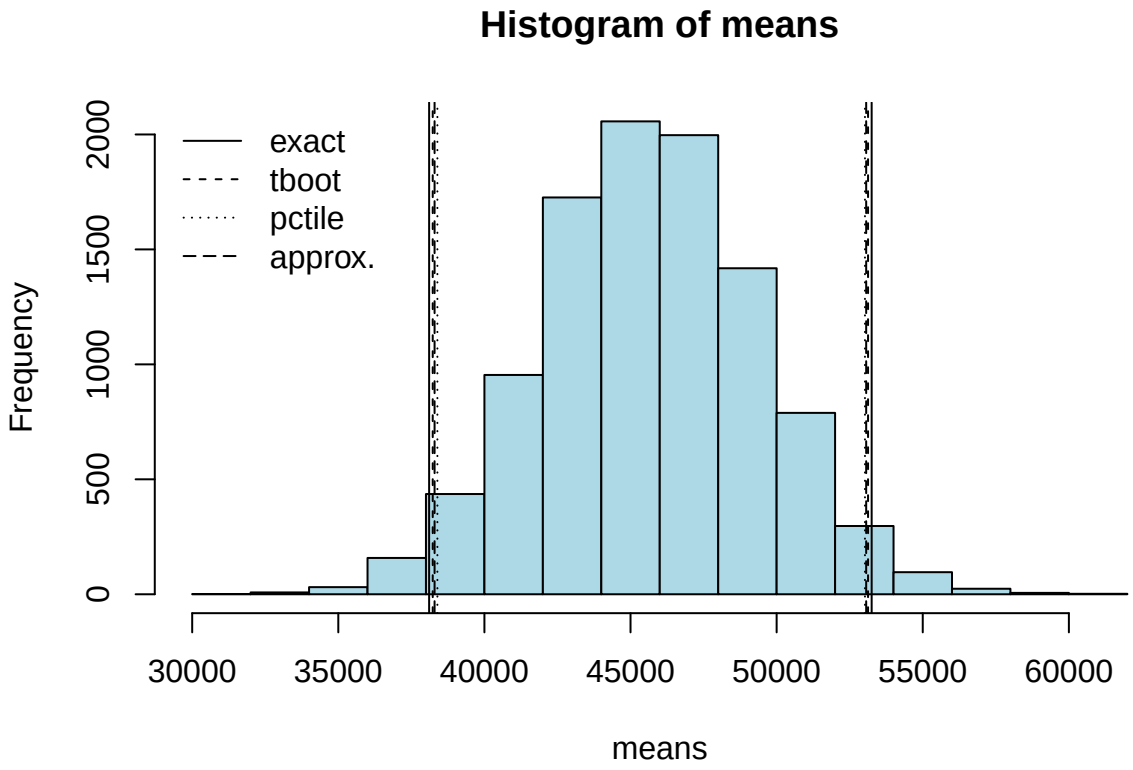
```
##    lower    upper
## 38231.06 53127.66
```

```
# bootstrap percentile confidence interval
pctile <- quantile(means,probs=c(0.025,0.975))
pctile
```

```
##     2.5%    97.5%
## 38391.44 53022.26
```

(d) Which interval is most appropriate for this data? Justify your answer.

```
# plot and compare
par(mfrow=c(1,1))
hist(means,breaks=20,col="light blue")  #histogram of means
abline(v=exact,lty=1)   # overlay exact interval
abline(v=tboot,lty=2)   # overlay t-bootstrap interval
abline(v=pctile,lty=3)  # overlay percentile interval
abline(v=approx,lty=5)  # overlay percentile interval
legend("topleft",legend=c("exact","tboot","pctile",'approx.'),lty=c(1,2,3,5),bty="n")
```

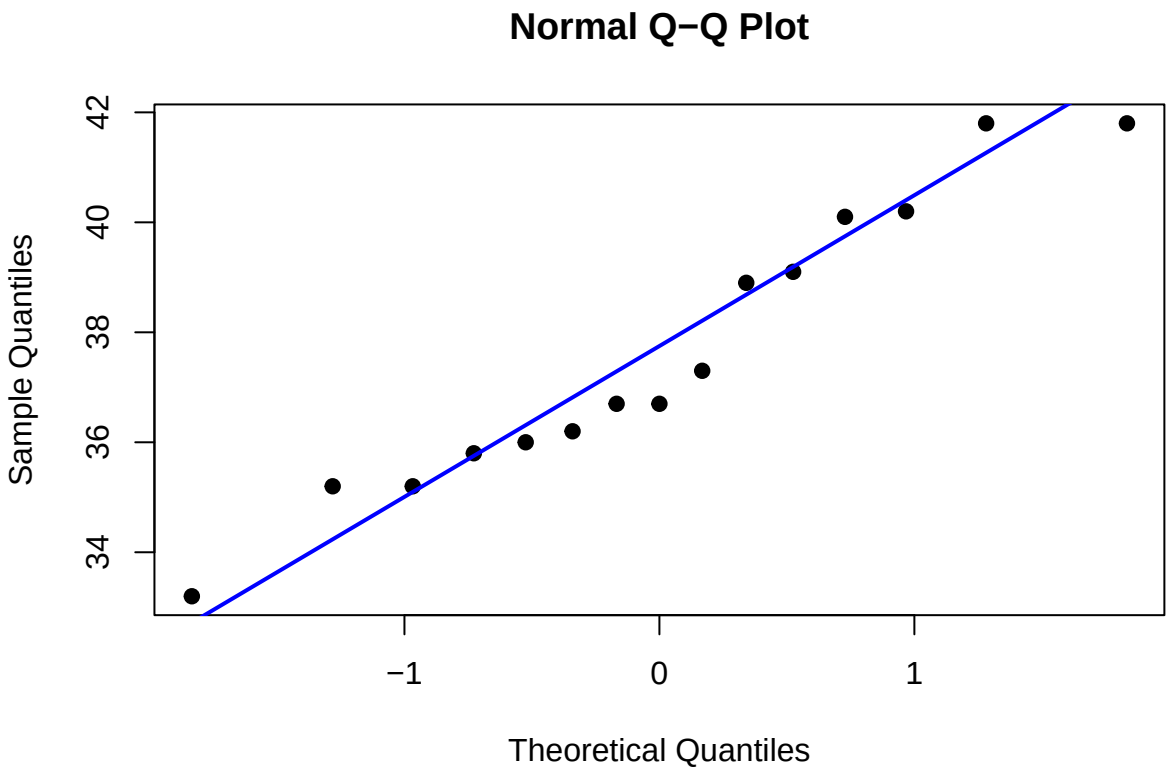


As we observe from histogram, tboot gives the best approximation to exact value.

Problem 5: Data from a sample of maximum pressures on concrete framework is contained in the file pressure.csv. Use this data to do the following in R:

(a) Make Q-Q plot of the data in the sample. Based on your observations, can we assume this data came from a normal distribution?

```
pressure <- read.csv("~/pressure.csv)
pressure = pressure[,1]
n = length(pressure)
qqnorm(pressure, pch=19)
qqline(pressure, lwd=2, col="blue")
```



As we observe, the data is normally distributed.

(b) Calculate an exact confidence interval for both the mean and the standard deviation based on the assumption the data is normally distributed.

```
#Exact.
t.crit <- qt(0.025,n-1,lower.tail=F)
exact <- c(mean(pressure)-t.crit*sd(pressure)/sqrt(n),mean(pressure)+t.crit*sd(pressure)/sqrt(n))
names(exact) = c("lower","upper")
exact
```

```
##    lower    upper
## 36.18928 39.03739
```

```
X2l.crit <- qchisq(0.025,n-1,lower.tail=T)
X2u.crit <- qchisq(0.975,n-1,lower.tail=T)
sd_interval <- c(sqrt((n-1)*var(pressure)/X2u.crit),sqrt((n-1)*var(pressure)/X2l.crit))
names(sd_interval)<-c("lower", "upper")
sd_interval
```

```
##    lower    upper
## 1.882671 4.055527
```

(c) Generate a bootstrap percentile confidence interval for the mean and the standard deviation from the sample data.

```
B <- 10000  # run 10000 bootstrap samples
means <- rep(0,B)  # create dummy matric for means
for(i in 1:B){
  rsample <- sample(pressure,n,replace=T)
  means[i] <- mean(rsample)  # calculate mean and save in vector
}
x_bar_bar <- mean(means)  # calculate mean estimate
```

```r
se_est <- sd(means) #standard error estimates
se_est
```

```
## [1] 0.6381276
```

```r
# t-CI with bootstrap SE
tboot <- c(mean(pressure)-qt(0.025,n-1,lower.tail=F)*sd(means),
           mean(pressure)+qt(0.025,n-1,lower.tail=F)*sd(means))
names(tboot) = c("lower", "upper")
tboot
```

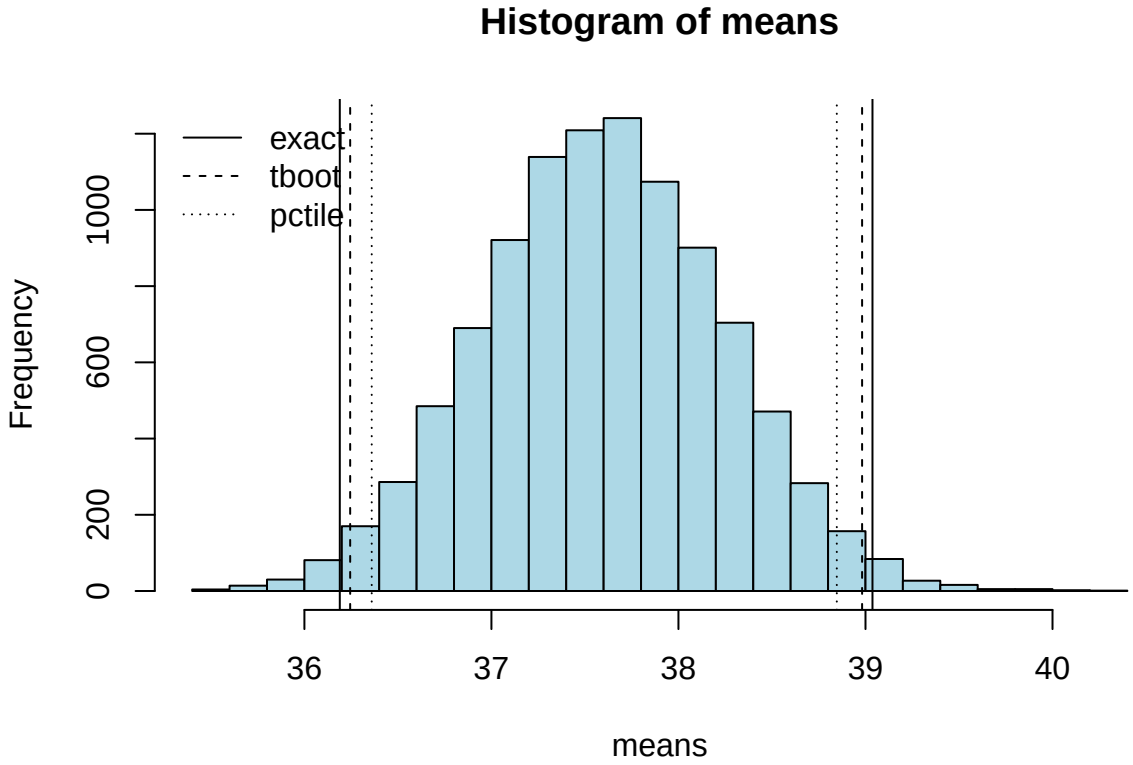
```
##    lower    upper
## 36.24469 38.98198
```

```r
# bootstrap percentile confidence interval
pctile <- quantile(means,probs=c(0.025,0.975))
pctile
```

```
##     2.5%    97.5%
## 36.36000 38.84667
```

(d) Which intervals are most appropriate for this data? Justify your answer

```r
# plot and compare
par(mfrow=c(1,1))
hist(means,breaks=20,col="light blue")  #histogram of means
abline(v=exact,lty=1)   # overlay exact interval
abline(v=tboot,lty=2)   # overlay t-bootstrap interval
abline(v=pctile,lty=3)  # overlay percentile interval
legend("topleft",legend=c("exact","tboot","pctile"),lty=c(1,2,3),bty="n")
```



As we observe from histogram, tboot gives the best approximation to exact value.