STAT 515 hw 6

CIs for mean with σ known, CIs for σ^2, large-sample CIs for mean/proportion

Attach a sheet with the R plots and R code printed on it. You may write out your other answers by hand if you want. Just try to make it easy for me grade!!

1. A beekeeper wishes to know the proportion of drones in a hive. She scoops a half-cup of bees from the hive and finds 40 drones among the 294 bees in the scoop. Assume each bee has the same probability of being in the scoop.

   (a) Give an estimate of the proportion of drones in the hive.

   > The estimated proportion drones in the hive is
   >
   > $$\hat{p}_n = \frac{40}{294} = 0.136.$$

   (b) Give a 95% Wald-type confidence interval for the proportion of drones in the hive.

   > The 95% Wald-type confidence interval for the proportion of drones in the hive is
   >
   > $$0.136 \pm 1.96\sqrt{0.136(1-0.136)/294} = (0.097, 0.175).$$

   (c) Are the conditions satisfied for using the Wald-type confidence interval?

   > We have $n\hat{p}_n = 40 \geq 15$ and $n(1-\hat{p}_n) = 294 - 40 = 254 \geq 15$, so the conditions are satisfied.

   (d) Give a 95% Agresti-Coull-type confidence interval for the proportion of drones in the hive.

   > The 95% Agresti-Coull-type confidence interval for the proportion of drones in the hive is
   >
   > $$42/298 \pm 1.96\sqrt{(42/298)(1-42/298)/298} = (0.101, 0.181).$$

   (e) The beekeeper would like the proportion of drones in the hive to be no greater than 15%. What does her data tell her?

   > The data are inconclusive; according to our confidence intervals, the percentage of drone bees might be as high as 18% or as low as 10%, if we round a bit.

   (f) If the true proportion of drones in the hive were equal to 0.15, with what probability would the beekeeper obtain 40 or more drones in her scoop of 294 bees?

      i. Compute this probability exactly, assuming that there are 30,000 bees in the hive.

If 15% of the bees are drones, then the number of drones is equal to $(0.15)30000 = 4500$. If we draw 294 bees without replacement, the probability that we will draw x drones is given by the hypergeometric probability mass function

$$\frac{\binom{4500}{x}\binom{30000-4500}{294-x}}{\binom{30000}{294}}, \qquad \text{for } x = 0, \ldots, 294.$$

If X is the number of drones in the scoop of 294 bees, we have

$$P(X \geq 40) = 1 - P(X \leq 39) = 1 - \sum_{x=0}^{39} \frac{\binom{4500}{x}\binom{30000-4500}{294-x}}{\binom{30000}{294}}.$$

We can obtain this using the `phyper()` function in R. We have

$$P(X \leq 40) = \texttt{1 - phyper(39, m=4500, n=30000-4500, k=294)} = 0.7724459.$$

ii. Compute this probability ignoring the fact that she is sampling without replacement.

If we ignore the fact that she is sampling without replacement, we can regard the drawings of the 294 bees as so many independent Bernoulli trials with probability of success $p = 0.15$. In this case we have

$$P(X \geq 40) = 1 - P(X \leq 39) = \texttt{1 - pbinom(39,294,0.15)} = 0.771256.$$

As expected, this answer is very close to our answer to the previous question, since the population (the hive) is large.

iii. Compute an approximation to this probability using the Normal distribution.

If X is the number of drones in the sample of 294 bees, then

$$P(X \geq 40) = P(X/294 \geq 40/294) = P(\hat{p}_n \geq 0.136).$$

From here, using the central limit theorem, we write

$$\begin{aligned} P(\hat{p}_n \geq 0.136) &= P((\hat{p}_n - 0.15)/\sqrt{0.15(1-0.15)/294} \geq (0.136 - 0.15)/\sqrt{0.15(1-0.15)/294}) \\ &\approx P(Z \geq -0.672), \quad Z \sim \text{Normal}(0,1) \\ &= \texttt{1 - pnorm(-0.672)} \\ &= 0.749. \end{aligned}$$

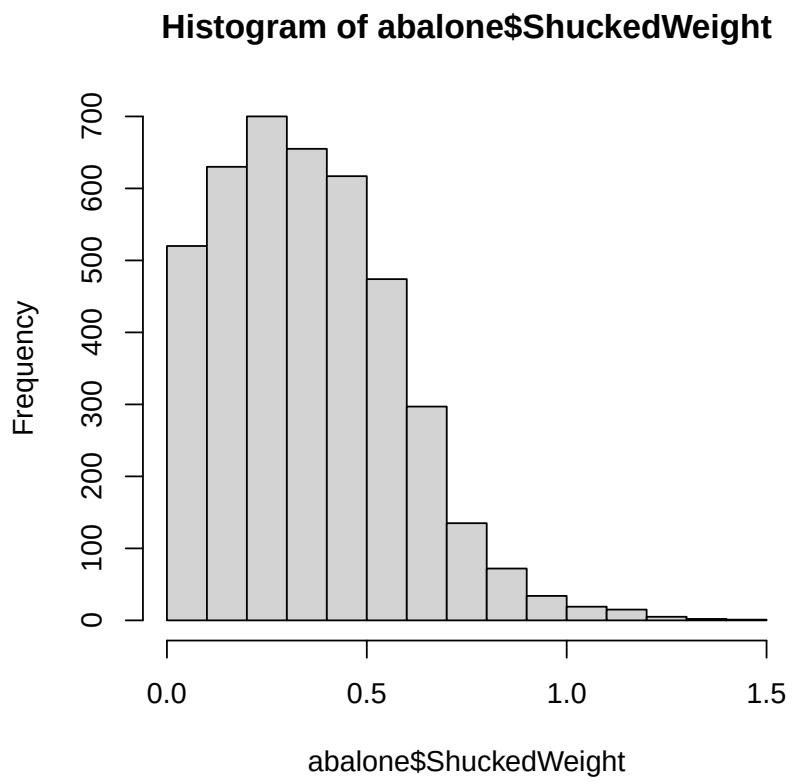
2. Import the abalone data set into R with this command:

```
load(url("https://people.stat.sc.edu/gregorkb/data/abalone.Rdata"))
```

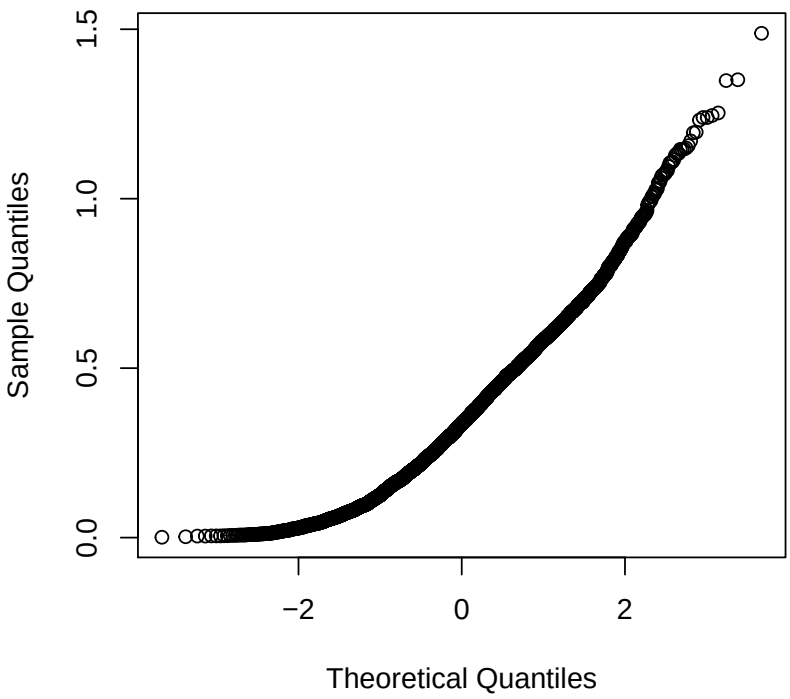
Regard the abalones in this data set as the entire population of abalones in the world. We are interested in the shucked weight of the abalones.

(a) Make a histogram of the shucked weights of the "world population" of abalones.

The histogram looks like



(b) Make a Normal quantile-quantile plot of the shucked weights using the `qqnorm()` function.



(c) Do the shucked weights of abalones appear to follow a Normal distribution?

Decidedly not. The histogram shows a right-skewed distribution.

(d) Give the mean μ and the standard deviation σ of the shucked weights of the "world population" of abalones.

We have $\mu = 0.3594$ and $\sigma = 0.222$.

(e) Take the first 40 abalones in the data set and regard them as a random sample of $n = 40$ abalones. Based on the shucked weights of these 40 abalones, give a confidence interval for the mean shucked weight of all abalones at the confidence level

i. 80%.

We have $\bar{X}_n = 0.2935$ and $z_{0.20/2} = z_{0.10} = \texttt{qnorm(0.90)} = 1.282$. So a 90% confidence interval for the mean shucked weight of abalones based on this sample is given by

$$0.2935 \pm 1.282 \cdot 0.222/\sqrt{40} = (0.249, 0.338).$$

ii. 90%.

The interval is
$$0.2935 \pm 1.645 \cdot 0.222/\sqrt{40} = (0.236, 0.351).$$

iii. 95%.

The interval is
$$0.2935 \pm 1.96 \cdot 0.222/\sqrt{40} = (0.225, 0.362).$$

(f) Which of the intervals from the previous part contained the true mean?

Only the 95% confidence interval contained the true mean.

(g) Run a simulation: For the sample sizes $n = 5$ and $n = 40$, draw 1,000 random samples from the "world population" of abalones. With each random sample, construct a 90%, a 95%, and a 99% confidence interval for the mean shucked weight of abalones. Record for each confidence interval whether it contained the true value of the population mean. In the end, record in a table like the one below the proportion of times (out of the 1,000 random samples) the confidence interval contained the true mean:

n	90%	95%	99%
5	0.898	0.956	0.992
40	0.899	0.954	0.993

Here is some sample code to get you started:

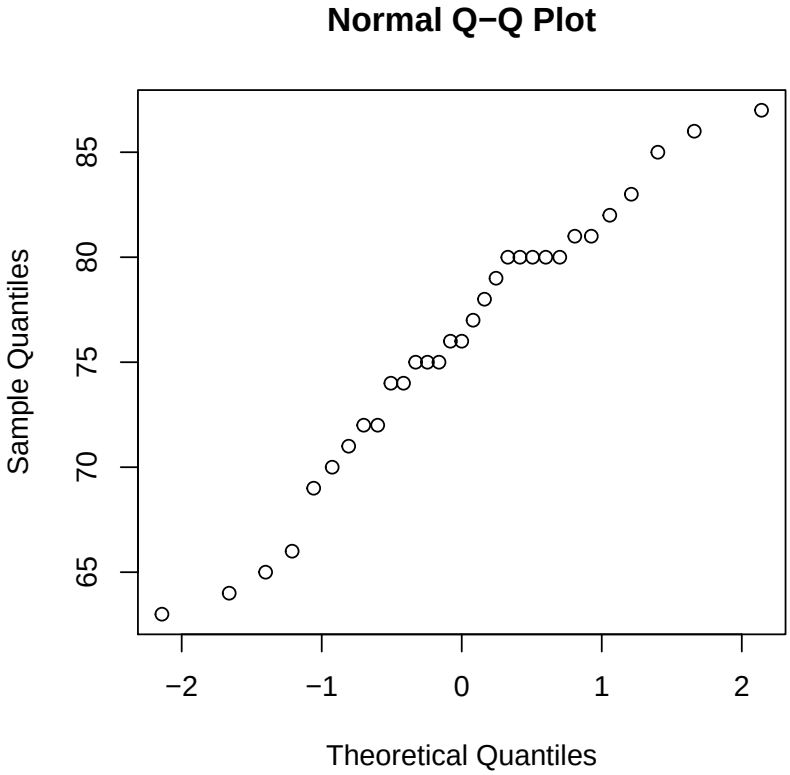
```
n <- 5
S <- 1000
cov90 <- numeric(S)
for(s in 1:S){

  X <- sample(population,n, replace = FALSE)
  X.bar <- mean(X)

  lo90 <- X.bar - 1.645 * 0.222 / sqrt(n)
  up90 <- X.bar + 1.645 * 0.222 / sqrt(n)

  cov90[s] <- (lo90 < mu) & (up90 > mu)

}
mean(cov90)
```

(h) Is there anything surprising about your results?

The central limit theorem appears to have kicked in very quickly—that is for very small n. The population distribution is not Normal; however, the confidence intervals at $n = 5$ still contained the true population mean with probability very close to the nominal, or stated, levels.

3. Bring the `trees` dataset (included in R) into the workspace with the command `data(trees)`. We will consider the heights of the trees.

   (a) Make a Normal quantile-quantile plot of the tree heights.



   (b) State whether you think the tree heights can be regarded as coming from a Normal population.

The points in the Normal Q-Q plot appear to fall fairly close to a straight line, so it is probably safe to proceed under the assumption that the population distribution is Normal.

   (c) Give a 95% confidence interval for the variance σ^2 of the population of tree heights.

We have $S_n^2 = 40.6$ with $n = 31$. So a 95% confidence interval for σ^2 is given by

$$\left(\frac{(31-1)(40.6)}{\chi^2_{31-1,0.975}}, \frac{(31-1)(40.6)}{\chi^2_{31-1,0.025}} \right) = (25.93, 72.54),$$

where $\chi^2_{31-1,0.975} =$ `qchisq(0.975,30)` $= 46.97924$ and $\chi^2_{31-1,0.025} =$ `qchisq(0.025,30)` $=$ 16.79077.

Optional (do not turn in) problems for additional study from McClave, J.T. and Sincich T. (2017) *Statistics*, 13th Edition: 7.8, 7.12, 7.60, 7.62