Discussion 7

Hanying Jiang

T and Bootstrap CI for a Population Mean

Review: T CI (`F23 STAT 371 05 Estimation II CIs 2023.10.03.pdf`)

(Two-tailed) Recall that we have learned Z-CI $\bar{X} \pm (Z_{\frac{\alpha}{2}})(\sigma_{\bar{\bar{X}}})$ and t-CI $\bar{X} \pm (t_{\frac{\alpha}{2}})(\hat{S}_{\bar{\bar{X}}})$ for the population mean:

When to use which? (Hint: recall how to calculate $\sigma_{\bar{\bar{X}}}$ and $\hat{S}_{\bar{\bar{X}}}$.)

R function: `qt()`

To compute t critical value with N (thus $df = N - 1$), we can use `qt(1-0.5 * alpha, N - 1)`.

For example, when I want to find the 99% CI with $N = 36$, the t critical value would be $t_{\frac{\alpha}{2}} = t_{0.005}$ with $df = N - 1 = 35$:

```
qt(0.995, 35)
```

```
## [1] 2.723806
```

Try in your computer. When I want to find the 98% CI with $N = 31$, what is the t critical value? (Check whether you can get 2.457262.)

Exercise 1

Try to answer (a) and solve (b) by hand.

Consider the Volume measurements of 31 cherry trees stored in the trees data frame in R.

```
volumes <- trees$Volume
(x_bar <- mean(volumes))
```

```
## [1] 30.17097
```

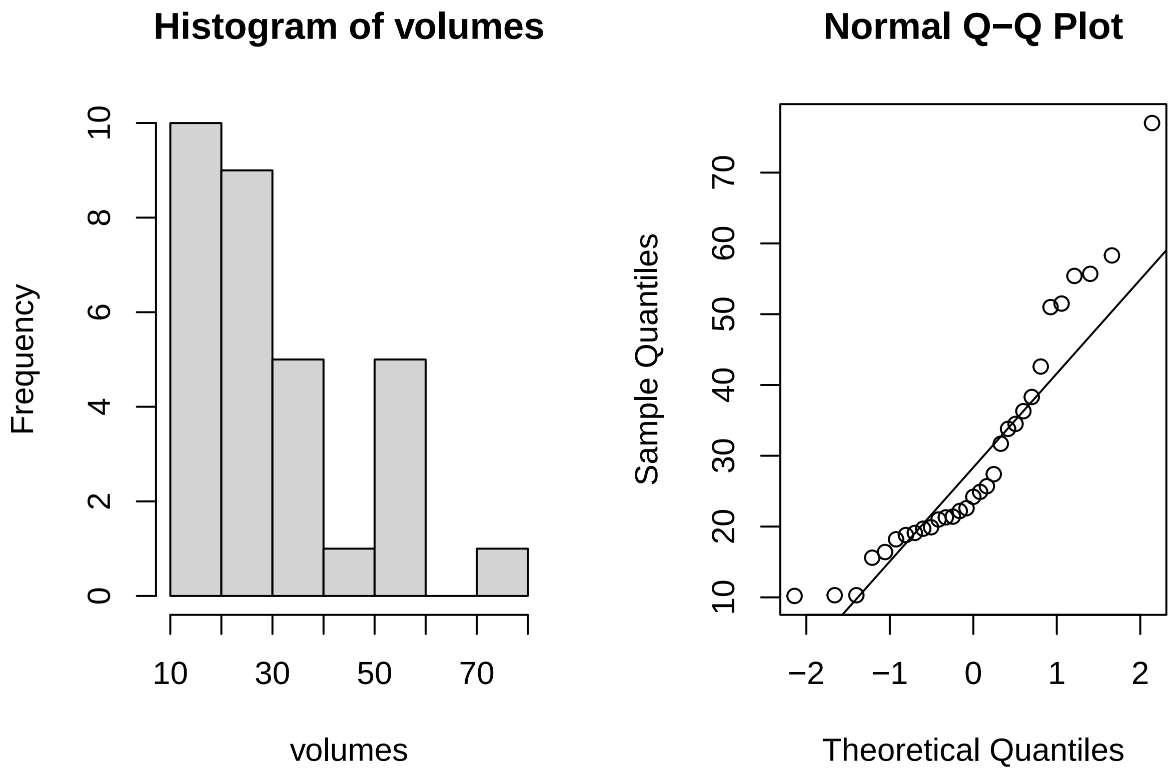
```
(s <- sd(volumes))
```

```
## [1] 16.43785
```

```
(n <- length(volumes))
```

```
## [1] 31
```

```
par(mfrow = c(1,2))
hist(volumes)
qqnorm(volumes); qqline(volumes)
```



```r
par(mfrow = c(1,1))
```

a. Consider the summary measures, histogram, and qqnorm() plot of the sample of 31 volume measurements given above. What evidence do we have that the population of all cherry tree volume measurements is not normally distributed?

b. Construct a 98% t confidence interval "by hand" and using t.test() for the average volume μ_V. As part of your process identify the **point estimate for μ_V, approximate SE for $\bar{X}$, critical value, and margin of error**. Explain why the t confidence interval might be a reasonable tool, even though we stated in (a) that the population of volume is not normally distributed.

c. Construct a 98% bootstrap t confidence interval for μ_V using the code from lecture. Compare and contrast it to the student's t confidence interval computed in (b).

Solution

a. Consider the summary measures, histogram, and qqnorm() plot of the sample of 31 volume measurements given above. What evidence do we have that the population of all cherry tree volume measurements is not normally distributed?

The sample data looks right skewed in the histogram. The `qqnorm()` plot also shows strong non-linearity. We also see that the minimum possible volume (0) is within 2 standard deviations of the mean (so values on lower half more tightly clustered).

b. Construct a 98% t confidence interval "by hand" and using t.test() for the average volume μ_V. As part of your process identify the **point estimate for μ_V, approximate SE for $\bar{X}$, critical value, and margin of error**. Explain why the t confidence interval might be a reasonable tool, even though we stated in (a) that the population of volume is not normally distributed.

"By hand"

Recall that the t CI is $\bar{X} \pm (t_{\frac{\alpha}{2}})(\hat{S}_{\bar{\bar{X}}})$.

$\bar{X}$ is computed to be 30.17097 by `x_bar <- mean(volumes)`.

S_X is 16.43785 by `s <- sd(volumes)`.

$\hat{S}_{\bar{\bar{X}}} = \frac{S_X}{\sqrt{N}} = \frac{16.43785}{\sqrt{31}} = 2.952324$.

$\alpha = 1 - 98\% = 0.02$ Critical value:

```r
qt(1-0.02 / 2, 31 - 1)
```

```
## [1] 2.457262
```

Margin of error: $2.457262 * 2.952324 = 7.254634$.

98% t CI: $30.17097 \pm 7.254634 = (22.91634, 37.4256)$. We are 98% confident the true average volume of cherry trees from which this sample was take is between: 22.91634 and 37.4256.

With R

```r
t.test(volumes, conf.level = .98)
```

```
##
##  One Sample t-test
##
## data:  volumes
## t = 10.219, df = 30, p-value = 2.753e-11
## alternative hypothesis: true mean is not equal to 0
## 98 percent confidence interval:
##  22.91633 37.42560
## sample estimates:
## mean of x
##  30.17097
```

c. Construct a 98% bootstrap t confidence interval for μ_V using the code from lecture. Compare and contrast it to the student's t confidence interval computed in (b).

The codes below is mostly copied from the codes posted in Oct 5.

```
#function to build bootstrap CI
boot_ci <- function(cot, n_boot, alpha){
  #get summaries from data#
  n <- length(cot)
  x_bar <-mean(cot)
  s <- sd(cot)
  se <- s/sqrt(n)
  #vector to store bootstrap samples#
  t_hat <- numeric(1000)
  #bootstrap loop#
  for(i in 1:1000){
  #draw a SRS of size n#
  x_star <- sample(cot, size = n, replace = T)
  #calc resampled mean and sd#
  x_bar_star <-mean(x_star)
  s_star <-sd(x_star)
  #calc t-hat and store as vector#
  t_hat[i] <- (x_bar_star - x_bar) / (s_star/sqrt(n))
  }
  #Find upper and lower critical values of approximate distribution#
  t_lower <- quantile(t_hat, probs = alpha/2, names = F)
  t_upper <-quantile(t_hat, probs = 1-(alpha/2), names = F)
  #Build final CI#
  boot_ci_upper <- x_bar - t_lower*se
  boot_ci_lower <- x_bar - t_upper*se
  ci <- c(boot_ci_lower, boot_ci_upper)
  return(ci)
}
boot_ci(volumes, 1000, 0.02)
```

```
## [1] 24.12328 40.15484
```

We see that the total width of the bootstrap t CI is about the same as the standard t given above. However, the bootstrap t CI is shifted a bit higher (is estimating higher population mean volume). I would tend to use the bootstrap CI because it is better reflecting our uncertainty about how large n should be to rely on the CLT. Especially since it resulted in different estimates than the standard t tool.

Z CI for a Population Proportion

Review

When interested in the population proportion p, we usually use the sample proportion $\hat{p}$ as an estimator.

$E(\hat{p}) = p$.

$VAR(\hat{p}) = \frac{p(1-p)}{n}$.

The Z CI for p is $\hat{p} \pm (Z_{\frac{\alpha}{2}})(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$.

Assumptions to check before constructing CIs for pop proportions:

- Randomly selected sample.
- Sufficient sample size - “>5” rule of thumb.

Exercise 2.

Sleep apnea is a disorder in which there are pauses in breathing during sleep. People with this condition must wake up frequently to breathe. In 427 medical records in a health care system for people aged 65 and over, 104 of them had a diagnosis of sleep apnea.

a. The health care system CEO is interested in knowing the porportion of people aged 65 and over who have diagnoised sleep apnea in their health care system. Provide them this information and explain why a confidence interval is not necessary.

Hint: What is the population? What is the parameter? Is the parameter known?

b. Check that the assumptions for constructing a confidence interval for π are reasonable.

c. Construct a 95% Z confidence interval for π, the population proportion of those aged 65 and over in similar health care systems who have sleep apnea. As part of your process identify the **point estimate, approximate SE for $\hat{p}$, critical value, and margin of error**.

Solution

a. The health care system CEO is interested in knowing the porportion of people aged 65 and over who have diagnoised sleep apnea in their health care system. Provide them this information and explain why a confidence interval is not necessary.

If we have all of the medical records in the health care system for people aged 65 and over (427) and the count of those that have diagnosed sleep apnea (104), we can exactly compute $\pi = \frac{104}{427} = 0.2436$ where π is the proportion of people aged 65 and over who have diagnosed sleep apnea in their health care system.

b. Check that the assumptions for constructing a confidence interval for π are reasonable.

(1) Do we have a simple random sample from our population of interest?

Not exactly, but a SRS from our population of interest is probably not attainable. This sample has its limitations, but may be the best we can get.

(2) Are the observations like iid draws?

Sampling without replacement so we need to check that the population is very large compared to our sample size, which is true.

(3) Is our sample size large enough so that CLT is a good approximation?

Are $n\pi$ and $n(1-\pi)$ both at least 5? Both 104 and $427 - 104$ are larger than 5.

c. Construct a 95% Z confidence interval for π, the population proportion of those aged 65 and over in similar health care systems who have sleep apnea. As part of your process identify the **point estimate, approximate SE for $\hat{p}$, critical value, and margin of error**.

$\hat{p} = \frac{104}{427} = 0.2435597$.

$S\hat{E}(P) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.02077189$.

Critical value:

```
qnorm(0.975)
```

```
## [1] 1.959964
```

Margin error: $1.959964 * 0.02077189 = 0.04071216$.

The Z CI is $0.2435597 \pm 0.04071216 = (0.2028476, 0.2842719)$.