Unit 6 Day 3: Exercise 22

Stat140-02

Confidence interval summary

Step 1: Identify the population parameter of interest

- Proportion
- Mean
- Difference in proportions
- Difference in Means

Step 2: Check the conditions for Central Limit Theorem

1. **Independence**: Sampled observations must be independent. This is difficult to verify, but is more likely if the **Randomization Condition** is met.

   - random sample or randomized experiment is used
   - or if sampling without replacement, $n < 10\%$ of the population.

2. **Sample size/skew**: do we have a large enough sample size? We should check the **Nearly normal condition**

   - if $n < 30$ the population distribution is normal or
   - $n > 30$ and the population dist. is not extremely skewed, or
   - n is much larger than 30 (approx. gets better as n increases).

Step 3: Compute point estimate and margin of error

Parameter	**Distribution**	**Standard Error (CI)**
Proportion	Normal	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Mean	$t, df = n - 1$	$\frac{s}{\sqrt{n}}$
Difference in Proportions	Normal	$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
Difference in Means	$t, df = \min(n_1, n_2) - 1$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Step 4: Write and interpret the confidence interval in the proper context

I am ____% confident that the true <u>[population parameter of interest]</u> is within the <u>[confidence interval]</u>.

Remember: Our uncertainty is about the interval, not the true mean. The interval varies randomly. The true mean sleep is neither variable nor random —just unknown.

Today: compute confidence interval with R

Example 1: College student sleep

I have data on the number of hours that 25 students slept and a histogram of the 25 observed amounts that students slept.

6	6	6
7	6	7
6	6	8
7	8	5
6	8	8
7	8	7
7	5	6
7	4	7
8		

I have loaded the dataset called `sleep`.

```
glimpse(sleep)
```

```
## Rows: 25
## Columns: 1
## $ hours <dbl> 6, 6, 6, 7, 6, 7, 6, 6, 8, 7, 8, 5, 6, 8, 8, 7, 8, 7, 7,...
```

Suppose we want to compute a 95% confidence interval for the mean of hours of sleep of all college students. To make the confidence interval, use the command:

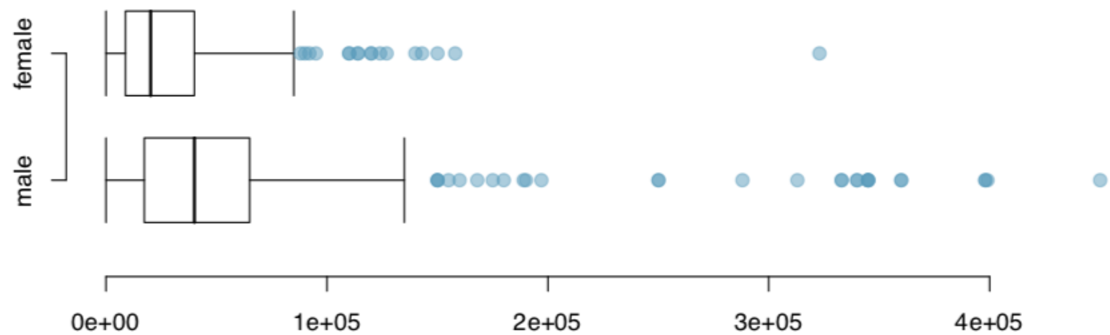
```
t.test(sleep$hours, conf.level=0.95)
```

```
##
##  One Sample t-test
##
## data:  sleep$hours
## t = 30.87, df = 24, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  6.196062 7.083938
## sample estimates:
## mean of x
##      6.64
```

Example 2: Gender gap in salaries

Since 2005, the American Community Survey polls approximately 3.5 million households yearly. The following summarizes distribution of salaries of males and females from a random sample of individuals who responded to the 2012 ACS:

	$\bar{x}$	s	n
male	55,890	68,767.88	470
female	29,240	32,025.98	373



I have loaded the dataset called `acs`.

```
glimpse(acs)
```

```
## Rows: 2,000
## Columns: 13
## $ income       <int> 60000, 0, NA, 0, 0, 1700, NA, NA, NA, 45000, NA, ...
## $ employment   <fct> not in labor force, not in labor force, NA, not i...
## $ hrs_work     <int> 40, NA, NA, NA, NA, 40, NA, NA, NA, 84, NA, 23, N...
## $ race         <fct> white, white, white, white, white, other, white, ...
## $ age          <int> 68, 88, 12, 17, 77, 35, 11, 7, 6, 27, 8, 69, 69, ...
## $ gender       <fct> female, male, female, male, female, female, male,...
## $ citizen      <fct> yes, yes, yes, yes, yes, yes, yes, yes, yes, yes,...
## $ time_to_work <int> NA, NA, NA, NA, NA, 15, NA, NA, NA, 40, NA, 5, NA...
## $ lang         <fct> english, english, english, other, other, other, e...
## $ married      <fct> no, no, no, no, no, yes, no, no, no, yes, no, no,...
## $ edu          <fct> college, hs or lower, hs or lower, hs or lower, h...
## $ disability   <fct> no, yes, no, no, yes, yes, no, yes, no, no, no, n...
## $ birth_qrtr   <fct> jul thru sep, jan thru mar, oct thru dec, oct thr...
```

Suppose we want to compute a 99% confidence interval for the average difference between the average salaries of males and females in the U.S.

```r
# Save the data in two different vector
female <- acs %>%
  filter(gender == "female") %>%
  pull(income)
male <- acs %>%
  filter(gender == "male") %>%
  pull(income)
# Compute t-test
t.test(male, female, conf.level=0.99)
```

```
##
##  Welch Two Sample t-test
##
## data:  male and female
## t = 8.1362, df = 1142.1, p-value = 1.056e-15
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  12490.81 24091.80
## sample estimates:
## mean of x mean of y
##  32627.30  14335.99
```

CLT based hypothesis testing

Example 1: Sleep versus Caffeine

Students were given words to memorize, then randomly assigned to take either a 90 min nap, or a caffeine pill. 2 and 1/2 hours later, they were tested on their recall ability. Is sleep or caffeine better for memory?

I have loaded the dataset `sleep.coffee`.

```r
glimpse(sleep.coffee)
```

```
## Rows: 24
## Columns: 2
## $ label <chr> "Sleep", "Sleep", "Sleep", "Sleep", "Sleep", "Sleep", "S...
## $ value <dbl> 14, 18, 11, 13, 18, 17, 21, 9, 16, 17, 14, 15, 12, 12, 1...
```

Let μ_s and μ_c be the mean number of words recalled after sleeping and after caffeine.

Step 1: Identify population parameter of interest

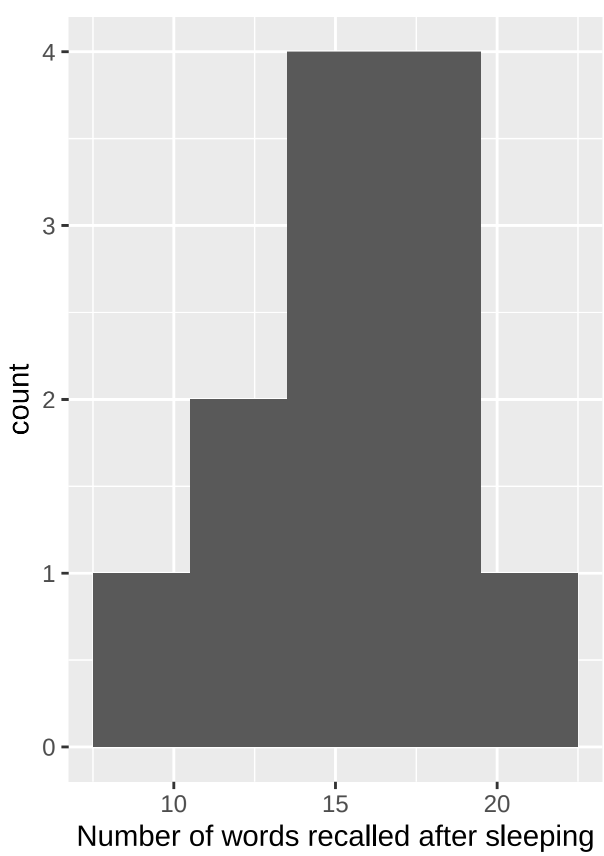
- **Population parameter**: $\mu_s - \mu_c$ the mean difference between the mean number of words recalled after sleeping and after caffeine

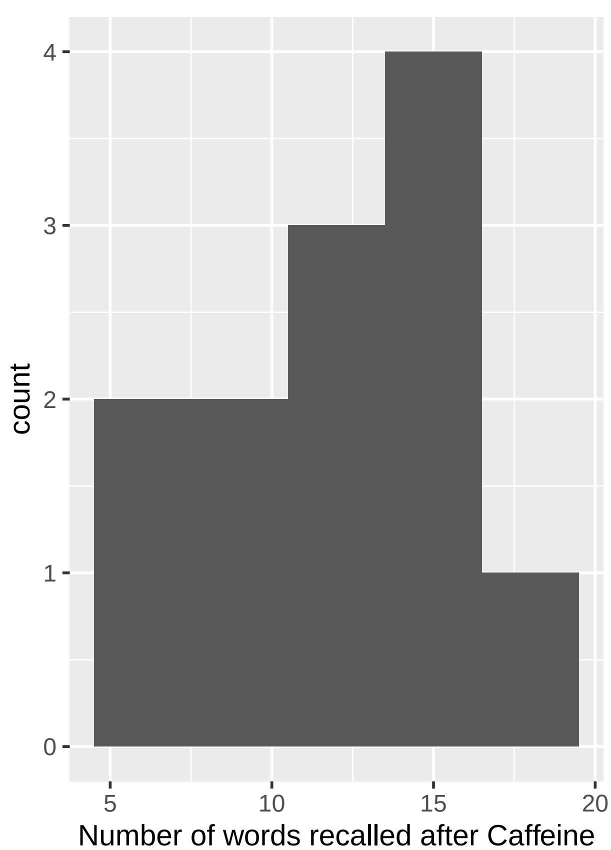
Step 2: State hypotheses

- **Null Hypothesis** (H_0): $\mu_s - \mu_c = 0$
- **Alternative Hypothesis** (H_A): $\mu_s - \mu_c \neq 0$

Step 3: Check conditions to apply the CLT

- **Randomization Condition**: Yes, the condition is met due to the random assignment in the experiment
- **Nearly normal condition**: Yes, the histogram of both groups is relatively normal
- **Independent groups?**: Yes, the two groups are independent.





Step 4: Calculate the p-value for the test

```r
# Save the data in two different vector
sleep <- sleep.coffee %>%
  filter(label == "Sleep") %>%
  pull(value)
coffee <- sleep.coffee %>%
  filter(label == "Caffeine") %>%
  pull(value)
# Compute t-test
t.test(sleep, coffee, mu = 0)
```

```
##
##  Welch Two Sample t-test
##
## data:  sleep and coffee
## t = 2.1438, df = 21.894, p-value = 0.04342
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.09699633 5.90300367
## sample estimates:
## mean of x mean of y
##     15.25     12.25
```

The `mu` argument gives the value with which you want to compare the sample mean. It is optional and has a default value of zero. By default, R performs a two-tailed test. To perform a one-tailed test, set the alternative argument to "greater" or "less", as shown below.

```r
t.test(sleep, coffee, mu = 0, conf.level=0.95, alternative="greater")
t.test(sleep, coffee, mu = 0, conf.level=0.95, alternative="less")
```

Step 5: Draw a conclusion, using a significance level of $\alpha = 0.05$.

Since the p-value is 0.04342, which is less than $\alpha = 0.05$, we reject the null hypothesis.

Example 2: College student sleep

Recall the college student sleep example, is the average hours of sleep per night for college students less than 7?

Step 1: Identify population parameter of interest

- **Population parameter**:

Step 2: State hypotheses

- **Null Hypothesis** (H_0):

- **Alternative Hypothesis** (H_A):

Step 3: Check conditions to apply the CLT

- **Randomization Condition**:

- **Nearly normal condition**:

- **Independent groups?**:

Step 4: Calculate the p-value for the test

Type in your R console the following code to define the `sleep` data frame. Then use the t.test to compute the p-value.

```
sleep <- data.frame(hours = c(6,6,6,7,6,7,6,6,8,7,
            8,5,6,8,8,7,8,7,7,5,6,7,4,7,8))
```

Step 5: Draw a conclusion, using a significance level of $\alpha = 0.05$.