Conducting and Interpreting t-Tests

Independent samples t-test

Jenine Harris
Brown School

Import the data

```
# import nhanes 2015-2016
nhanes.2016 <- read.csv("/Users/harrisj/Box/teaching/Teaching/Fall2020/d

# check the data
summary(object = nhanes.2016)
```

```
##       SEQN          cycle               SDDSRVYR    RIDSTATR    RIAGENDR
##  Min.   :83732   Length:9544        Min.   :9   Min.   :2   Min.   :1.00
##  1st Qu.:86222   Class :character   1st Qu.:9   1st Qu.:2   1st Qu.:1.00
##  Median :88726   Mode  :character   Median :9   Median :2   Median :2.00
##  Mean   :88720                      Mean   :9   Mean   :2   Mean   :1.51
##  3rd Qu.:91210                      3rd Qu.:9   3rd Qu.:2   3rd Qu.:2.00
##  Max.   :93702                      Max.   :9   Max.   :2   Max.   :2.00
##
##     RIDAGEYR        RIDAGEMN        RIDRETH1        RIDRETH3        RIDEXMON
##  Min.   : 0.00   Min.   : 0.00   Min.   :1.00   Min.   :1.000   Min.   :1.00
##  1st Qu.: 9.00   1st Qu.: 5.00   1st Qu.:2.00   1st Qu.:2.000   1st Qu.:1.00
##  Median :27.00   Median :10.00   Median :3.00   Median :3.000   Median :2.00
##  Mean   :31.87   Mean   :10.76   Mean   :3.01   Mean   :3.216   Mean   :1.51
##  3rd Qu.:53.00   3rd Qu.:17.00   3rd Qu.:4.00   3rd Qu.:4.000   3rd Qu.:2.00
##  Max.   :80.00   Max.   :24.00   Max.   :5.00   Max.   :7.000   Max.   :2.00
##                  NA's   :8882
##     RIDEXAGM        DMQMILIZ        DMQADFC          DMDBORN4
##  Min.   :  0.0   Min.   :1.000   Min.   :1.000   Min.   : 1.000
##  1st Qu.: 41.0   1st Qu.:2.000   1st Qu.:1.000   1st Qu.: 1.000
##  Median :100.0   Median :2.000   Median :2.000   Median : 1.000
##  Mean   :104.5   Mean   :1.914   Mean   :1.531   Mean   : 1.244
```

Comparing two unrelated sample means with an independent samples t-test

- The one sample t-test is great for checking to see how well a sample represents a population for a single variable.
    - Is the sample mean statistically significantly different from some hypothesized mean? (population or something else)
    - For example, is the NHANES sample systolic blood pressure mean the same as 120?
- Instead of comparing one mean to a hypothesized or population mean, the **independent samples t-test** compares the means of two groups to each other.
- For example, the NHANES data set includes sex measured in two categories: males and females.
- You might be interested in whether the mean systolic blood pressure was the same for males and females in the population.
- That is, do males and females in the sample come from a population where males and females have the same mean systolic blood pressure?
- The independent samples t-test could be used to find out the answer.

EDA for independent samples t-test

- Comparing means across the groups of interest:

```
# compare means of BPXSY1 across groups
# sex variable is RIAGENDR
nhanes.2016 %>%
  drop_na(BPXSY1) %>%
  group_by(RIAGENDR) %>%
  summarize(m.sbp = mean(BPXSY1))
```

```
## # A tibble: 2 x 2
##   RIAGENDR m.sbp
##      <int> <dbl>
## 1        1  122.
## 2        2  119.
```

Data cleaning

- It certainly looks like there might be a difference between the two means, but it is unclear who has higher or lower blood pressure since the categories of sex are not labeled clearly.

- Use the codebook to find out how the `RIAGENDR` is coded and recode:

```
# add labels to sex and rename variables
nhanes.2016.cleaned <- nhanes.2016 %>%
  mutate(RIAGENDR = recode_factor(.x = RIAGENDR,
                                  `1` = 'Male',
                                  `2` = 'Female')) %>%
  rename(sex = RIAGENDR) %>%
  rename(systolic = BPXSY1)
```

Examine the means with recoded data

```r
# compare means of systolic by sex
nhanes.2016.cleaned %>%
  drop_na(systolic) %>%
  group_by(sex) %>%
  summarize(m.sbp = mean(x = systolic))
```

```
## # A tibble: 2 x 2
##   sex    m.sbp
##   <fct>  <dbl>
## 1 Male    122.
## 2 Female  119.
```

Examine the groups with a plot

```
# density plot of systolic by sex
dens.sex.bp <- nhanes.2016.cleaned %>%
  ggplot(aes(x = systolic,
             fill = sex)) +
  geom_density(alpha = .8) +
  theme_minimal(base_size = 18) +
  labs(x = "Systolic blood pressure", y = "Probability density",
       title = "Distribution of systolic blood pressure by sex in mmHg\n
  scale_fill_manual(values = c('gray', '#7463AC'),
                    name = "Sex")
dens.sex.bp
```

NHST Step 1: Write the null and alternate hypotheses

H0: There is no difference in mean systolic blood pressure for males and females in the US population.

HA: There is a difference in mean systolic blood pressure for males and females in the US population.

NHST Step 2: Compute the test statistic

- The test statistic for the independent samples t-test is a little more complicated to calculate since it now includes the means from both the groups in the numerator and the standard errors from the groups in the denominator.

- In the independent samples t-test formula, m_1 is the mean of one group and m_2 is the mean of the other group; the difference between the means makes up the numerator.

- The larger the difference between the group means, the larger the numerator will be and the larger the t-statistic will be!

- The denominator includes the variances for the first group, s_1^2, and the second group, s_2^2 and the sample sizes for each group, n_1 and n_2.

$$t = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Computing more descriptive stats

```r
# compare means of systolic by sex
nhanes.2016.cleaned %>%
  drop_na(systolic) %>%
  group_by(sex) %>%
  summarize(m.sbp = mean(systolic),
            var.sbp = var(systolic),
            samp.size = n())
```

```
## # A tibble: 2 x 4
##   sex    m.sbp var.sbp samp.size
##   <fct>  <dbl>   <dbl>     <int>
## 1 Male    122.    329.      3498
## 2 Female  119.    358.      3647
```

$$t = \frac{122.1767 - 118.9690}{\sqrt{\frac{329.2968}{3498} + \frac{358.2324}{3647}}} = 7.31$$

Compute the t-test with R

After watching Leslie substitute in the values and do the math, Nancy typed a line of code:

```
# compare systolic blood pressure for males and females
twosampt <- t.test(formula = nhanes.2016.cleaned$systolic ~ nhanes.2016.
twosampt
```

```
##
##      Welch Two Sample t-test
##
## data:  nhanes.2016.cleaned$systolic by nhanes.2016.cleaned$sex
## t = 7.3135, df = 7143, p-value = 2.886e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.347882 4.067432
## sample estimates:
##   mean in group Male mean in group Female
##             122.1767             118.9690
```

- In a lot of statistical tests, the object on the Left-Hand-Side (LHS) of the formula is the **outcome** or **dependent variable** while the object(s) on the Right-Hand-Side (RHS) of the formula are the **predictors** or **independent variables**.
- In this case, systolic blood pressure is the *outcome* being explained by the *predictor* of sex.

Results of the t-test

- The `t.test()` output shows a t-statistic of 7.3135.
- The degrees of freedom are 7142.9989031, which is the sample size of 7,145 minus two because there are two groups.
- In the case of the **independent samples t-test**, the degrees of freedom are computed as n - k, where n is the sample size and k is the number of groups.
- The 95% confidence interval is the interval around the **difference between the two groups**.
- In the sample, the difference between male systolic blood presure (m = 122.1766724) and female systolic blood pressure (m = 118.9690156) is 3.2076568.
- In the population this sample came from, the difference between the mean male and female systolic blood pressure is likely to be between 2.3478815 and 4.067432 (the 95% confidence interval).
- The confidence interval range does not contain zero, so in the population this sample came from, the difference between male and female blood pressure is not likely to be zero.
- Based on the difference in the sample and the other characteristics of the sample, there is likely some difference between male and female blood pressure in the sampled population.

NHST Step 3: Compute the probability for the test statistic (p-value).

- The p-value in this case was shown in *scientific notation* which can be converted to p = 0.0000000000002886278.
- In this case, use $p < .05$ instead since the longer version of the p-value was difficult to read and took up a lot of space.
- Interpret this as indicating that the value of this t-statistic would happen with a probability of much less than 5% **if the null hypothesis were true**.

NHST Steps 4 & 5: Interpret the probability and write a conclusion.

- In this case, the t-statistic was definitely in the rejection region, so there was sufficient evidence to reject the null hypothesis in favor of the alternate hypothesis.
- Even though the difference between the mean systolic blood pressure for males and females was small, it was statistically significant.
- The probability of this sample coming from a population where the means for males and females are equal is very low, it would happen about 0.00000000002886278% of the time.
- The sample was therefore likely to be from a population where males and females had different mean systolic blood pressure.
- Summarize the results:
    - There was a statistically significant difference [$t(7142.9989031) = 7.31$; $p < .05$] between the mean systolic blood pressure for males ($m = 122.18$) and females ($m = 118.97$) in the sample. The sample was taken from the US population indicating that males in the US likely have a different mean systolic blood pressure than females in the US. The difference between male and female mean systolic blood pressure was 3.21 in the sample; in the