Conducting and Interpreting t-Tests

Dependent samples t-test

Jenine Harris
Brown School

Import and clean the data

```r
# 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 related sample means with a dependent samples t-test

- Sometimes means to compare could be related.
- This usually happens in one of two ways; either the same people are measured twice or people in the sample are siblings or spouses or co-workers or have some other type of relationship.
- It may seem strange to measure the same people twice, but often people will be measured before and after some sort of intervention and the measures are compared to see if they changed.
- In the dependent samples t-test formula, the m_d is the *mean of the differences* between the related measures, the s_d^2 is the variance of the mean difference between the measures, and n_d is the sample size.

$$t = \frac{m_d - 0}{\sqrt{\frac{s_d^2}{n_d}}}$$

More about the dependent samples t-test

- Rather than the mean(s) from the groups of interest, the formula uses **the mean of the differences between the two related measures** (m_d).

- For example, if someone's systolic blood pressure were measured to be 110 before she went to the dentist and 112 after she went to the dentist, the difference between the two measures would be 2.

- If someone else were measured as having 115 before the dentist and 110 after, the difference between the two measures would be -5.

- In a study of blood pressure before and after going to the dentist, the numerator for the paired t-test would take the mean of those differences, 2 and -5, and subtract 0.

- The reason it would subtract zero is that zero is the mean difference if the measures of blood pressure were exactly the same before and after the dentist visit---this is the null hypothesis.

- In the case of the systolic blood pressure measurement for NHANES, the measure was taken up to four times for each person to ensure that it was accurate.

- The blood pressure numbers should be the same or very similar since they are from the same person and nothing in particular happened between the first and second measure to increase or decrease blood pressure.

EDA for dependent t-tests

- The codebook shows the variable names for the first and second measures of systolic blood pressure, `BPXSY1` and `BPXSY2`.
- To conduct the paired samples t-test, renaming these variables so they are easier to remember would be good.
- Create a variable for the differences between the first systolic blood pressure measure and the second one.
- Name the difference variable something easy to interpret like `diff.syst` for difference in systolic blood pressure.

```
# rename second systolic measure and create diff variable for
# difference between measure 1 and 2 for systolic BP
nhanes.2016.cleaned <- nhanes.2016 %>%
  mutate(RIAGENDR = recode_factor(.x = RIAGENDR,
                                  `1` = 'Male',
                                  `2` = 'Female')) %>%
  rename(sex = RIAGENDR) %>%
  rename(systolic = BPXSY1) %>%
  rename(systolic2 = BPXSY2) %>%
  mutate(diff.syst = systolic - systolic2)
```

Mean of difference in blood pressure measures

- Now that there is a variable, `diff.syst`, measuring the difference between the first and second systolic blood pressure measures, check the descriptive statistics and visualize the distribution of the new d variable, `diff.syst`.

```
# mean of the differences
nhanes.2016.cleaned %>%
  drop_na(diff.syst) %>%
  summarize(m.diff = mean(diff.syst))
```

```
##      m.diff
## 1 0.5449937
```

- The mean difference between the first and second systolic blood pressure measures was 0.54, which was not zero, but it was pretty small.
- On average, the systolic blood pressure measure was 0.54 different between the first measure and the second on the same person in the NHANES 2015-2016 data set.

Graphing the difference variable

```
# histogram of the differences between first and second
# blood pressure measures
nhanes.2016.cleaned %>%
  ggplot(aes(x = diff.syst)) +
  geom_histogram(fill = "#7463AC", color = "white") +
  theme_minimal(base_size = 18) +
  labs(x="Difference between SBP measure 1 and 2",
       y="Number of NHANES participants",
       title = "Difference between measure 1 and 2 for systolic blood pr
```

Interpreting the stats & histogram

- The distribution of differences looked close to normal and the center was near 0, but maybe not exactly 0.
- The mean difference was .54.
- If measures 1 and 2 were exactly the same for each person, there would just be one long bar at 0 in the histogram and the mean difference would be 0.
- Use the NHST process to see if the m_d of 0.54 was statistically significantly different from the 0 expected if the first and second measures of systolic blood pressure had been exactly the same for each person.

NHST Step 1: Write the null and alternate hypotheses

H0: There is no difference between measures 1 and 2 for systolic blood pressure.

HA: There is a difference between measures 1 and 2 for systolic blood pressure.

NHST Step 2: Compute the test statistic

- To substitute the mean, standard deviation, and sample size of `diff.syst` into the formula for the paired t-test statistic, add variance and sample size to the descriptive statistics code:

```
# mean, var, and sample size of the difference variable
nhanes.2016.cleaned %>%
  drop_na(diff.syst) %>%
  summarize(m.sbp = mean(diff.syst),
            var.sbp = var(diff.syst),
            n = n())
```

```
##       m.sbp  var.sbp    n
## 1 0.5449937 23.99083 7101
```

Computing the test statistic manually

- Notice that the sample size is even smaller this time; it looks like 7,101 people had data for both measures available.

$$t = \frac{.5449937 - 0}{\sqrt{\frac{23.99083}{7101}}} = 9.38$$

Computing the test statistic in R

- A t-statistic this large seems likely to be statistically significant, but rather than guess, use the `t.test()` function again, but this time with the `paired = TRUE` argument since the default for the command is an *independent* samples t-test and this is a *dependent* samples t-test.

```
# paired t-test for systolic measures 1 and 2
t.test(x = nhanes.2016.cleaned$systolic,
       y = nhanes.2016.cleaned$systolic2,
       paired = TRUE)
```

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

- The p-value is shown in scientific notation as `< 2.2 e-16` which is well below .05.
- The probability is very low of finding a mean difference between `systolic` and `systolic2` of 0.54 if there were no difference between the measures in the population that the sample came from.

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

- The t-statistic has a low probability, so there was sufficient evidence to reject the null hypothesis in favor of the alternate hypothesis.
- Even though the mean difference between the first and second measures was small, it was statistically significant.
- The probability of this sample coming from a population where the first and second measures of blood pressure were equal was very low.
- This sample is likely to be from a population where systolic blood pressure is not consistent over time.
- The confidence interval is the range where the difference between the first and second measures likely lies in the population.
- The difference statistic was calculated by subtracting the second measure of systolic blood pressure, `systolic2`, from the first measure, `systolic` and the mean difference is **positive**.
  - This indicated that the first measure of systolic blood pressure tended to be higher than the second measure in the sample.