Conducting and Interpreting t-Tests

Descriptives & graphs for 1 categorical & 1 continuous

Jenine Harris
Brown School

Understanding the relationship between one categorical variable and one continuous variable s

- Import the NHANES 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.   :1.0
##  1st Qu.: 9.00   1st Qu.: 5.00   1st Qu.:2.00   1st Qu.:2.000   1st Qu.:1.0
##  Median :27.00   Median :10.00   Median :3.00   Median :3.000   Median :2.0
```

Examine the blood pressure variable

```r
# open tidyverse for graphing with ggplot2
library(package = "tidyverse")

# graph systolic blood pressure variable BPXSY1
sbp.histo <- nhanes.2016 %>%
  ggplot(aes(x = BPXSY1)) +
  geom_histogram(fill = "#7463AC", color = "white") +
  theme_minimal() +
  labs(x = "Systolic blood pressure (mmHg)",
       y = "NHANES participants",
       title = "Distribution of systolic blood pressure in mmHg for\n201
sbp.histo
```

Interpreting the histogram

- The distribution of sbp was close to normally distributed with a little right skew.
- The graph showed that most people have systolic blood pressure between 100 and 150.
- The CDC defines normal systolic blood pressure as below 120mmHg, at-risk between 120-139, and high as 140 and above.
- Viewing these ranges in the histogram might be useful.
- Add a logical statement to `fill =` to fills the histogram based on the statement.
- In this case add `BPXSY1 > 120` to fill the histogram with one color when R evaluated the statement and found that it was `FALSE` and another color when R evaluated the statement and found that it was `TRUE`.
- Add the two colors for `BPXSY1 > 120` is `TRUE` and `BPXSY1 > 120` is `FALSE` to the `scale_fill_manual()` layer along with labels corresponding to the two groups.
- This results in a histogram with purple representing normal systolic blood pressure and gray representing at-risk or high systolic blood pressure.

Histogram formatted to show normal and high bp

```r
# graph systolic bp BPXSY1
sbp.histo <- nhanes.2016 %>%
  ggplot(aes(x = BPXSY1, fill = BPXSY1 > 120)) +
  geom_histogram(color = "white") +
  theme_minimal() +
  scale_fill_manual(values = c("#7463AC","gray"),
                    labels=c("Normal range", "At-risk or high"),
                    name = "Systolic\nblood pressure") +
  labs(x = "Systolic blood pressure (mmHg)",
       y = "Number of NHANES participants",
       title = "Distribution of systolic blood pressure\nin mmHg for 201
sbp.histo
```

Histogram formatted to show normal and high bp

Diastolic blood pressure

- For diastolic blood pressure, the CDC defines normal as < 80 mmHG, at risk as 80 - 89 mmHG, and high as 90+ mmHg.

- Used the same code and change the variable name to `BPXDI1` and the threshold to 80mmHG.

```
# graph diastolic bp BPXDI1
nhanes.2016 %>%
  ggplot(aes(x = BPXDI1, fill = BPXDI1 > 80)) +
  geom_histogram(color="white") +
  theme_minimal() +
  scale_fill_manual(values = c("#7463AC","gray"),
                    labels=c("Normal range", "At-risk or high"),
                    name = "Blood pressure") +
  labs(x="Diastolic blood pressure (mmHg)",
       y="Number of NHANES participants",
       title = "Distribution of diastolic blood pressure\nin mmHg for 20
```

Diastolic blood pressure

- For diastolic blood pressure, the CDC defines normal as < 80 mmHG, at risk as 80 - 89 mmHG, and high as 90+ mmHg.
- Used the same code and change the variable name to `BPXDI1` and the threshold to 80mmHG.

Interpreting the dbp histogram

- The diastolic histogram had a tiny bar at 0, which seems like a terrible blood pressure.
- This is an indicator that it would be wise to check those observations later, they are probably a data entry problem or some missing value coding.
- More people were within the normal range for diastolic blood pressure than were in the normal range for systolic blood pressure.
- Looking at these two distributions, the mean systolic blood pressure in the sample was likely higher than the 120 threshold for healthy.

Descriptive statistics

- Based on observing the histograms, it appears the mean systolic blood pressure in the sample was higher than 120.

- In addition to the histogram, check this with the mean and standard deviation:

```
# mean and sd of systolic blood pressure
nhanes.2016 %>%
  drop_na(BPXSY1) %>%
  summarize(m.sbp = mean(BPXSY1),
            sd.sbp = sd(BPXSY1))
```

```
##      m.sbp   sd.sbp
## 1 120.5394 18.61692
```

- The observed mean was 120.54 which was just slightly higher than the threshold of 120.

- While it does not seem like a big difference, a t-test can determine whether the 120.54 is different enough from 120 to be statistically significantly different.