Conducting and Interpreting t-Tests

Assumptions for t-tests

Jenine Harris
Brown School

Import and clean the data

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

# 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) %>%
  rename(systolic2 = BPXSY2) %>%
  mutate(diff.syst = systolic - systolic2)

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

```
##       SEQN          cycle              SDDSRVYR    RIDSTATR      sex
##  Min.   :83732   Length:9544        Min.   :9   Min.   :2   Male  :4676
##  1st Qu.:86222   Class :character   1st Qu.:9   1st Qu.:2   Female:4868
##  Median :88726   Mode  :character   Median :9   Median :2
##  Mean   :88720                      Mean   :9   Mean   :2
##  3rd Qu.:91210                      3rd Qu.:9   3rd Qu.:2
##  Max.   :93702                      Max.   :9   Max.   :2
##
##     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
```

Underlying assumptions for using the t-test

- Just like chi-squared, t-tests have to meet a few assumptions before they can be used.
- The first assumption for the t-test is that the data are normally distributed.
- For the one-sample t-test, the single variable being examined should be normally distributed.
- For the independent samples t-test and the paired samples t-test the data within each of the two groups should be normally distributed.
- There are different ways to assess normality:
    - Visually, a histogram or a Q-Q plot is useful for identifying normal and non-normal data distribution.
    - Statistically, a Shapiro-Wilk test can be used.

Testing normality with a histogram

- For the one-sample t-test comparing systolic blood pressure to a hypothesized population mean of 120, the histogram to determine whether a t-test was appropriate would look like:

```
# graph systolic bp
nhanes.2016.cleaned %>%
  ggplot(aes(x = systolic)) +
  geom_histogram(fill = "#7463AC", col = "white") +
  theme_minimal(base_size = 16) +
  labs(x = "Systolic blood pressure (mmHg)",
       y = "NHANES participants",
       title = "Distribution of systolic blood pressure in mmHg\nfor 201
```

Q-Q plot for normality

- Another way to visually check normality is with a Q-Q plot, or quantile-quantile plot.
- This plot is made up of points below which a certain percentage of the observations fall.
- On the x-axis are normally distributed values with a mean of 0 and a standard deviation of 1. On the y-axis are the observations from the data.
- If the data are normally distributed, the values will form a diagonal line through the graph.
    - Consistent with the right-skewed histogram, the higher observed values at the top of the graph are further from the line representing normality.
    - The visual evidence is enough to state that the normality assumption is **not met**.
    - However, if the graphs showed the data were closer to normal, computing skewness or kurtosis, or using a statistical test for normality, would help to determine if the data were normally distributed.

Using stats to test normality

- Different statistical checks of normality are useful in different situations.
- The mean of a variable is sensitive to skew, so the measure of skewness is good to check when a statistical test relies on means (like t-tests).
- When the focus of a statistical test is on variance, it is a good idea to examine kurtosis because variance is sensitive to problems with kurtosis (e.g., a platykurtic or leptokurtic distribution).
- The **Shapiro-Wilk** test is an inferential test that tests the null hypothesis that the data are normally distributed.
    - The **Shapiro-Wilk** test is sensitive to even small deviations from normality and is not useful for sample sizes above 5,000 because it will *always* find non-normality.
    - Given these limitations, Shapiro-Wilk is useful for testing normality in smaller samples when it is important that small deviations from normality are identified.

Examining skewness

```
# skewness of systolic bp
semTools::skew(object = nhanes.2016.cleaned$systolic)
```

```
##   skew (g1)          se           z           p
##  1.07037232  0.02897841 36.93689298  0.00000000
```

- The cutoffs for skewness that are problematic for this sample size are z values outside the range -7 to 7.
- The z here is 36.94, so skew is definitely a problem! The data are not normal and this assumption is failed.

Normality in independent and dependent samples t-tests

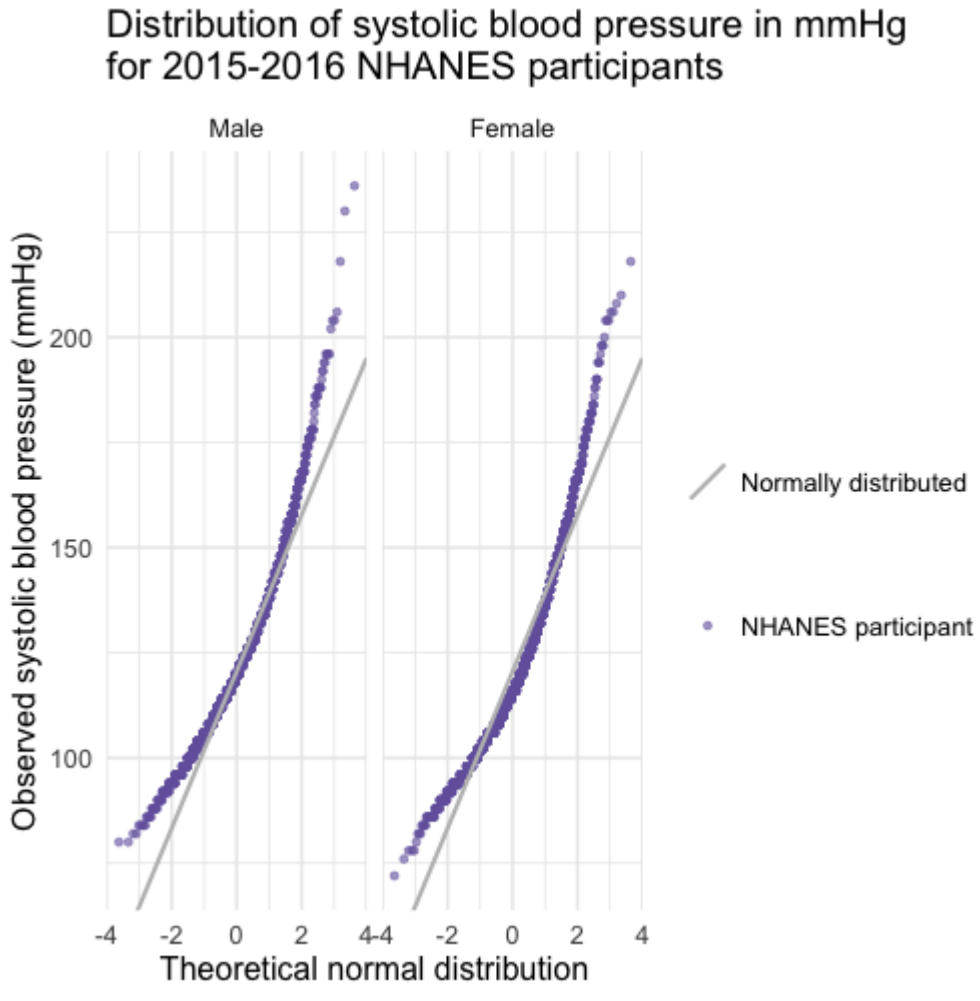
- Normality is checked for *each group* for the independent samples t-test and paired samples t-test.

```
#graph systolic bp by sex
nhanes.2016.cleaned %>%
  ggplot(aes(x = systolic)) +
  geom_histogram(fill = "#7463AC", col = "grey") +
  facet_grid(. ~ sex) +
  theme_minimal() +
  labs(x = "Systolic blood pressure (mmHg)",
       y = "NHANES participants",
       title = "Distribution of systolic blood pressure in mmHg\nfor 1,0
```

Normality in independent and dependent samples t-tests

- Normality is checked for *each group* for the independent samples t-test and paired samples t-test.

Normality with Q-Q plots



- The data within each group clearly *failed the assumption of normal distribution*.

Examining skew by group

- The skewness statistic could help to confirm this statistically for each of the two groups.
- Add the `semTools::skew()` code to the `summarize()` function to get the skew for each group.
- However, the `summarize()` function only prints a single number; print the z since that is the statistic used to determine how much skew is too much skew.
- The z is the third statistic printed in the `skew()` output, so Leslie added `[3]` to the end of the command to print the z.

```
# statistical test of normality for systolic bp by sex
nhanes.2016.cleaned %>%
  drop_na(systolic) %>%
  group_by(sex) %>%
  summarize(z.skew = semTools::skew(object = systolic)[3])
```

```
## # A tibble: 2 x 2
##   sex    z.skew
##   <fct>   <dbl>
## 1 Male     25.6
## 2 Female   27.6
```

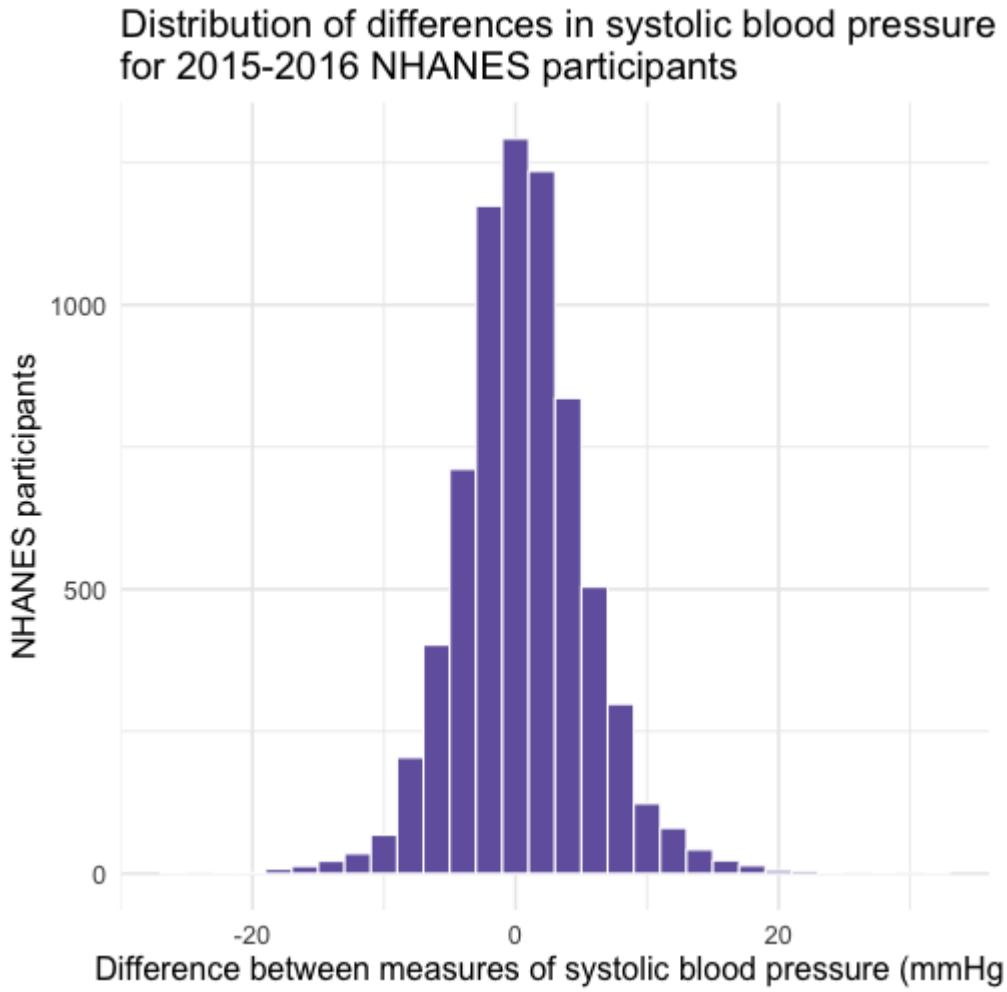
- The z values for skew of 25.64 for males and 27.59 for females were far above the acceptable range of -7 to 7 for this sample size, so both groups are skewed.

Normality for paired t-tests

- Testing normality for the paired-samples t-test is similar to the other t-tests.
- Use a graph and test for skewness of the `diff.syst` variable to see if the differences between the first and second measures are normally distributed.

```
#graph systolic difference between systolic and systolic2
nhanes.2016.cleaned %>%
  ggplot(aes(x = diff.syst)) +
  geom_histogram(fill = "#7463AC", col = "white") +
  theme_minimal(base_size = 16) +
  labs(x = "Difference between measures of systolic blood pressure (mmHg
       y = "NHANES participants",
       title = "Distribution of differences in systolic blood pressure r
```

Normality for paired t-tests



Q-Q plot for paired t-test normality test

```
#graph difference between systolic and systolic2
nhanes.2016.cleaned %>%
  drop_na(diff.syst) %>%
  ggplot(aes(sample = diff.syst)) +
  stat_qq(aes(color = "NHANES participant"), alpha = .6) +
  geom_abline(aes(intercept = mean(x = diff.syst),
                  slope = sd(x = diff.syst), linetype = "Normally distribute
                  color = "gray", size = 1) +
  theme_minimal(base_size = 16) +
  labs(x = "Theoretical normal distribution",
       y = "Observed differences between SBP measures",
       title = "Distribution of differences in systolic blood\npressure
    scale_color_manual(values = "#7463AC", name = "") +
    scale_linetype_manual(values = 1, name = "")
```

Q-Q plot for paired t-test normality test

Check the skew

```
# statistical test of normality for difference variable
semTools::skew(object = nhanes.2016.cleaned$diff.syst)
```

```
##    skew (g1)           se            z            p
## 2.351789e-01 2.906805e-02 8.090632e+00 6.661338e-16
```

- Despite the promising histogram, the Q-Q plot and z for skew of 8.09 suggest that the difference variable is not normally distributed. The `diff.syst` data failed this assumption.

Homogeneity of variances assumption

- While failing the normality assumption would be enough of a reason to choose another test, Kiara explained that there is one additional assumption to test for the independent samples t-test.
- The assumption is **homogeneity of variances** or equal variances across groups.
- Not only do the data need to be normally distributed, but the data should be equally spread out in each group.
- The histogram show these data might actually meet the assumption.
- **Levene's Test** is widely used to test the assumption of equal variances.
- The null hypothesis for Levene's Test is that **the variances are equal** while the alternate hypothesis is that the variances are not equal.
- A statistically significant Levene's Test would mean rejecting the null hypothesis of equal variances and failing the assumption.

Conducting the Levene's test

```
# equal variances for systolic by sex
car::leveneTest(y = systolic ~ sex, data = nhanes.2016.cleaned)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value  Pr(>F)
## group    1   3.552 0.05952 .
##       7143
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The p-value in the output is shown in the column with the heading `Pr(>F)`.
- This Levene's Test had a p-value of .06, which is not enough to reject the null hypothesis.
- Therefore, the assumption is met.
- The variances of systolic blood pressure for men and women are not statistically significantly different (p = .06) and the independent samples t-test *meets* the assumption of *homogeneity of variances*.
- Overall, none of the tests passed all assumptions. All of the tests failed the assumption of normal distribution.

Summary of assumptions

One-sample t-test assumptions

- continuous variable
- independent observations
- normal distribution

Independent-samples t-test assumptions

- continuous variable and two independent groups
- independent observations
- normal distribution in each group
- equal variances for each group

Dependent-samples t-test assumptions

- continuous variable and two dependent groups
- independent observations
- normal distribution of differences