Week 7: Bivariate tests for one categorical and one continuous variable

Jenine Harris

October 18, 2018

Week 8 schedule

- t-test activity
- t-test workshop
- sign up for book club topic

Workshop outline

These slides review the basics of *hypothesis testing* and several specific hypothesis tests to compare continuous variables across groups:

- Review null hypothesis significance testing (NHIST)
- One-sample t-test
- Independent samples t-test
- Dependent or paired t-test
- Analysis of Variance (ANOVA)
- Testing assumptions
- Alternative tests after not meeting assumptions
- Extra stuff

Null hypothesis significance testing (NHIST)

Relationships among variables are often described as being *statistically significant* or *not statistically significant* as a result of NHIST.

The process of NHIST is:

- Write the null and alternate hypotheses
- Compute the appropriate test statistic
- Calculate the probability that your test statistic is as big as it is if there is no relationship (i.e., the null is true)
- If the probability that the null is true is very small–less than 5%–reject the null hypothesis (if the p is low, the null must go!)
- If the probability that the null is true is not small–5% or higher–retain the null hypothesis

One-sample t-test

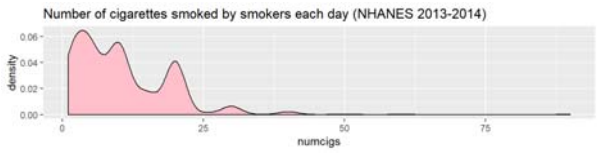
- The *one-sample t-test* compares a **sample mean** to a **hypothesized** or **population mean**.
- For example, we might hypothesize that the mean number of cigarettes a smoker smokes is 20, since this is the number of cigarettes in one pack.
- To test NHANES smokers against this hypothetical mean, we can compare the mean number of cigarettes smoked per day during the past 30 days by the NHANES participants to the hypothetical value of 20.

Write the null and alternate hypotheses

- Null (H0): Smokers smoke an average of 20 cigarettes per day (m = 20)
- Alternate (HA): Smokers do not smoke an average of 20 cigarettes per day (m is not equal to 20)

Examine the data

```
# Bring in the NHANES smoker data call it smokeNHANES
library(RNHANES)
smokeNHANES <- nhanes_load_data(file_name = "SMQ_H",
                                year = "2013-2014",
                                demographics = TRUE)

# clean the number of cigarettes variable
library(car)
smokeNHANES$numcigs <- car::recode(smokeNHANES$SMD650, "777 = NA; 999 = NA")

# explore the data with a graph
library(ggplot2)
ggplot(data = smokeNHANES, aes(x = numcigs)) +
  geom_density(fill = 'pink') +
  ggtitle("Number of cigarettes smoked by smokers each day (NHANES 2013-2014)")
```



Compute the appropriate test statistic

The one-sample t-test uses the t-statistic (sort of like a z-statistic) to test whether a sample came from a population with the hypothesized mean:

$$t = \frac{m - \mu}{\frac{s}{\sqrt{n}}}$$

- m is the sample mean
- μ is the population mean (or hypothetical mean)
- s is the sample standard deviation
- n is the sample size

Get mean, sd, n from R and compute

```
# mean and standard deviation for number of cigarettes
mean(smokeNHANES$numcigs, na.rm = TRUE)
```

```
## [1] 10.4053
```

```
sd(smokeNHANES$numcigs, na.rm = TRUE)
```

```
## [1] 8.428882
```

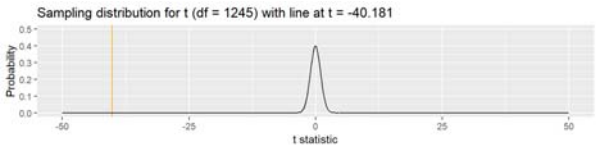
```
# use naniar library to get sample size treating NA as missing
library(naniar)
n_complete(smokeNHANES$numcigs)
```

```
## [1] 1246
```

$$t = \frac{10.405297 - 20}{\frac{8.4288816}{\sqrt{1246}}} = -40.1809877$$

Find the probability of your test statistic

- The area under the curve of the sampling distribution contains all possible values of the one-sample t-statistic for samples with 1245 d.f. that *came from a population where m = 20*
- The probability of getting a t-statistic of -40.181 (or larger) for a sample from *a population where m = 20* is the area under the curve to the left of the orange line
- This is a very tiny probability of getting a sample where the t-statistic is this big (or bigger) *if m = 20 in the population*
- So, *the mean (m) is probably not 20 in the population that this sample came from*



Use R to compute the actual p-value

```
# conduct the t-test with the hypothesized population mean (mu) as 20
t.test(smokeNHANES$numcigs, mu=20)
```

```
##
##  One Sample t-test
##
## data:  smokeNHANES$numcigs
## t = -40.181, df = 1245, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 20
## 95 percent confidence interval:
##   9.936827 10.873767
## sample estimates:
## mean of x
##   10.4053
```

- The t-statistic is -40.181 and the p-value for the t-statistic is < 2.2e-16, which is less than .05
- The probability of getting a t-statistic of -40.181 or even more extreme, *if the null hypothesis is true*, is < .00000000000000022
- That is a *very* low probability of getting such a large t-statistic, so the null is *probably not true*. We should *reject the null hypothesis*.

Interpret the results

- There is sufficient evidence to reject the null hypothesis
- The mean of 10.41 cigarettes smoked per day is statistically significantly different from the hypothesized mean of 20 cigarettes per day (t = -40.18; p < .05)
- The sample of smokers likely came from a population with a daily mean smoking rate of between 9.94 and 10.87 cigarettes per day (see confidence intervals from the t.test printout)

Altogether: A one-sample t-test comparing the mean number of cigarettes NHANES participants smoke per day to a hypothesized mean of 20 found that the mean of 10.41 cigarettes (sd = 8.43) smoked per day by NHANES participants is statistically significantly different from 20 cigarettes per day (t = -40.18; p < .05). The sample of smokers likely came from a population with a daily mean smoking rate of between 9.94 and 10.87 cigarettes per day (95% CI: 9.94-10.87).

Independent samples t-test

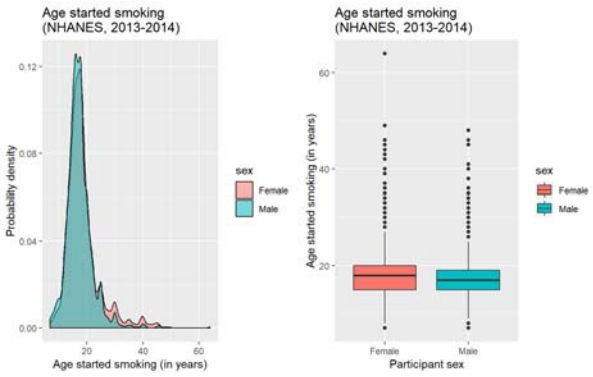
- Instead of comparing one mean to a hypothesized or population mean, the independent samples t-test compares the means of two groups to each other to see if they likely come from a population where the two groups had the same mean
- In the NHANES smoker data there are males and females
- Some research shows female smokers start at younger ages than male smokers
- We can use the independent samples t-test to find out if this is the case

Write the null and alternate hypotheses

- H0: There is no difference in mean age of starting smoking for male and female smokers
- HA: There is a difference in mean age of starting smoking for male and female smokers

```
# recode the variables
smokeNHANES$sex <- car::recode(smokeNHANES$RIAGENDR, "1 = 'Male'; 2 = 'Female'")
smokeNHANES$age.smoking <- car::recode(smokeNHANES$SMD030, "0 = NA;
                                777 = NA; 999 = NA")
```


Explore the data with graphs



Calculate the test statistic

By hand:

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

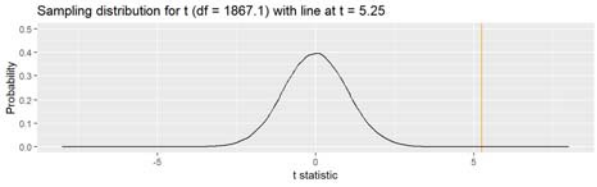
Or, with R:

```
# t-test for age smoking initiation by sex
age.smoke.sex.t <- t.test(smokeNHANES$age.smoking ~ smokeNHANES$sex)
age.smoke.sex.t
```

```
##
##  Welch Two Sample t-test
##
## data:  smokeNHANES$age.smoking by smokeNHANES$sex
## t = 5.2488, df = 1867.1, p-value = 1.705e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.7366393 1.6155312
## sample estimates:
## mean in group Female   mean in group Male
##             18.65580             17.47972
```

Find the probability of t under the null

- The area under the curve of the sampling distribution contains all possible values of the t-statistic for samples with 1867.14 d.f. that *came from a population where the means are equal*
- The probability of getting a t-statistic of 5.25 (or larger) for a sample from *a population where the means are equal* is the area under the curve to the right of the orange line
- This is a very tiny probability of getting a sample where the t-statistic is this big (or bigger) *if the means are equal in the population*
- So, *the mean ages of smoking initiation are probably not equal for males and females in the population that this sample came from*



Write a conclusion/interpretation

- There is sufficient evidence to reject the null hypothesis ($t = 5.25$; $p < .05$).
- The mean age of smoking initiation for males is statistically significantly different the mean age of smoking initiation for females.
- Male smokers and female smokers in this sample likely come from a population with different mean ages of smoking initiation for males and females.
- Males start smoking earlier (m = 17.48, sd = 4.59) than females (m = 18.66, sd = 6.06).

Altogether:

An independent samples t-test comparing the mean age of smoking initiation by sex for NHANES participants found that there is a statistically significant difference between males and females ($t = 5.25$; $p < .05$). Males start smoking earlier (m = 17.48, sd = 4.59) than females (m = 18.66, sd = 6.06).

Dependent samples t-test

- Sometimes the means you want to compare will be related (not independent groups).
- For example, the mean number of cigarettes smoked or the mean BMI before and after an intervention.
- When the two groups being compared are related (same people before & after, siblings, spouses, or two otherwise matched groups) an adjustment to the t-test to account for the non-independence is used.
- Everything else about the test stays the same! *See the Dalgaard text for more information.*

Comparing 3 or more means with Analysis of Variance (ANOVA)

- When you have three or more group means to compare you need something a little fancier than a t-test.
- The statistical test for comparing means across 3 or more groups is ANOVA.
- For example, we could test the mean age of smoking initiation by race/hispanic origin.

Clean the data:

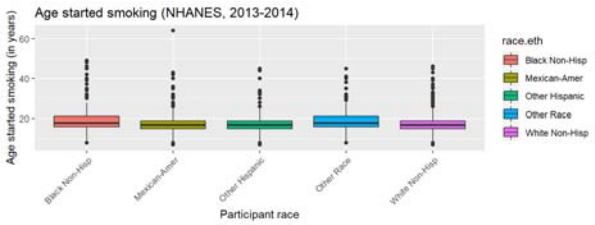
```
# recode the race-eth variable
smokeNHANES$race.eth <- car::recode(smokeNHANES$RIDRETH1, "1 = 'Mexican-Amer';
2 = 'Other Hispanic'; 3 = 'White Non-Hisp'; 4 = 'Black Non-Hisp';
5 = 'Other Race'")

# this is a factor variable or a category, so tell R to treat it as such!
smokeNHANES$race.eth <- factor(smokeNHANES$race.eth)
```

Write the null and alternate hypotheses

- H0: There is no difference in the mean age of smoking initation across race/ethnicity groups.
- HA: There is a difference in the means.

Explore with a graph:

```
# plot race-eth by groups
ggplot(data = smokeNHANES, aes(x = race.eth, y = age.smoking)) +
  geom_boxplot(aes(fill = race.eth)) +
  xlab("Participant race") +
  ylab("Age started smoking (in years)") +
  ggtitle("Age started smoking (NHANES, 2013-2014)") +
  theme(axis.text.x = element_text(angle=45, hjust=1))
```

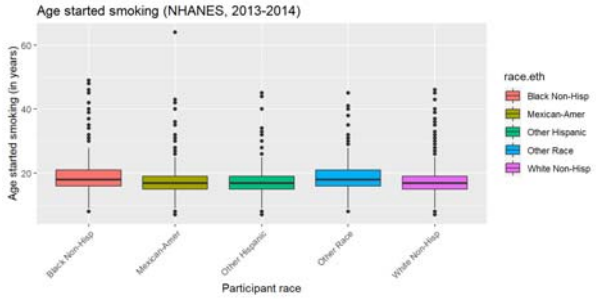


Calculate the test statistic

- In ANOVA, the test statistic is F
- F is the ratio of between-group variability to within-group variability

$$F = \frac{between - group - variability}{within - group - variability}$$

- Is there more difference *between* groups than there is *within* groups?



Use R to calculate F

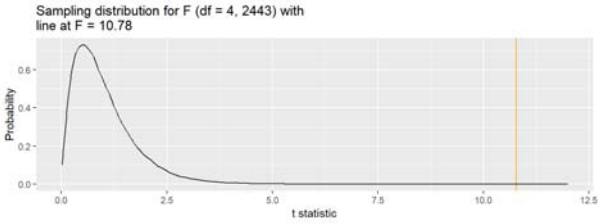
```
# use aov for ANOVA
race.smoking.age <- aov(age.smoking ~ race.eth, data = smokeNHANES)
summary(race.smoking.age)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## race.eth      4   1191  297.77   10.78 1.16e-08 ***
## Residuals  2443  67483   27.62
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 4720 observations deleted due to missingness
```

- F = 10.78
- p < .05

Find the probability of F under the null

- The area under the curve of the sampling distribution contains all possible values of the F-statistic for samples with 4 and 2443 d.f. that *came from a population where the means are equal*
- The probability of getting an F-statistic of 10.78 (or larger) for a sample from *a population where the means are equal* is the area under the curve to the right of the orange line
- This is a very tiny probability of getting a sample where the F-statistic is this big (or bigger) *if the means are equal in the population*
- So, *the mean ages of smoking initiation are probably not equal across race/ethnicity groups in the population that this sample came from*



Write a conclusion/interpretation

- There is sufficient evidence to reject the null hypothesis (F(4, 2443) = 10.78; p < .05)
- The mean age of smoking initiation is not equal across race/ethnicity groups
- The five groups likely have come from a population where mean age of smoking initiation differs by race/ethnicity

means:

```
## Black Non-Hisp   Mexican-Amer Other Hispanic     Other Race White Non-Hisp
##       18.98039       17.80859       18.23196       18.90400       17.37884
```

sd:

```
## Black Non-Hisp   Mexican-Amer Other Hispanic     Other Race White Non-Hisp
##       5.677966       5.953289       5.801805       5.170078       4.832132
```

Altogether:

An ANOVA analysis comparing the mean age of smoking initiation found a statistically significant difference (F(4, 2443) = 10.78; p < .05) in the means for Black non-Hispanic (m = 19; sd = 5.7), Mexican-American (m = 17.8, sd = 6), Other Hispanic (m = 18.2, sd = 5.8), Other Race (m = 18.9, sd = 5.2), and White non-Hispanic (m = 17.4, sd = 4.8) NHANES participants.

Which groups are different?

- ANOVA is an *omnibus test* so even though the graph and descriptive statistics show which group means seem different from one another, we need a statistical test to determine which means are *statistically significantly* different from one another.
- If the ANOVA is NOT SIGNFICANT, the post-hoc test is NOT NEEDED because there are no differences among the means

Conduct the post-hoc test

- There are several different post-hoc tests to choose from
- The **Bonferroni post-hoc test** is one possibility
- The Bonferroni post-hoc test uses the t-test to compare means, but adjusts the p-value so that the overall chance of being wrong across several tests (the familywise alpha) stays below 5%

```
# post-hoc test to determine which means are different
# x is continuous variable, g is groups (categorical variable)
pairwise.t.test(x = smokeNHANES$age.smoking, g = smokeNHANES$race.eth)
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  smokeNHANES$age.smoking and smokeNHANES$race.eth
##
##                Black Non-Hisp Mexican-Amer Other Hispanic Other Race
## Mexican-Amer   0.02910        -            -              -
## Other Hispanic 0.45753        0.79500      -              -
## Other Race     0.85069        0.13411      0.72616        -
## White Non-Hisp 7.9e-08        0.72616      0.21385        0.00027
##
## P value adjustment method: holm
```


Interpret the post-hoc test

- The values in the Bonferroni table are the p-values for t-tests of each pair of means.
- It appears that the mean age of smoking initiation for Non-Hispanic Blacks (m = 18.98) is significantly (p = .03) higher than for Mexican-Americans (m = 17.81).
- The age of initiation for Non-Hispanic Whites (m = 17.38) was statistically significantly (p < .05) lower than the age of initiation for Non-Hispanic Blacks (m = 18.98) and Other Race (m = 18.90) participants.

Altogether:

An ANOVA analysis comparing the mean age of smoking initiation found a statistically significant difference (F(4, 2443) = 10.78; p < .05) in the means for Black non-Hispanic (m = 19; sd = 5.7), Mexican-American (m = 17.8, sd = 6), Other Hispanic (m = 18.2, sd = 5.8), Other Race (m = 18.9, sd = 5.2), and White non-Hispanic (m = 17.4, sd = 4.8) NHANES participants. A Bonferroni post-hoc analysis indicated that the mean age of smoking initiation for Non-Hispanic Blacks was significantly higher than for Mexican-Americans and the age of initiation for Non-Hispanic Whites was statistically significantly lower than for Non-Hispanic Blacks or Other Race participants.

Testing assumptions

All statistical tests make some underlying assumptions about the data being tested. Just like the mean is not a great indicator of central tendency when the distribution is skewed, statistical tests like t-tests and ANOVA are not great when the data fail to meet the underlying assumptions.

The one-sample t-test, independent samples t-test, and ANOVA rely on a few main assumptions:

- Independence of observations: Each observation in the data set is unrelated to the others in the data set. If your data includes siblings, spouses, or other related observations, it may not meet this assumption. *See Dependent Samples t-test*
- Normal distribution: The outcome variable is normally distributed
- Homogeneity of variance: The groups have the same or similar variance for the outcome **(not applicable in the one-sample test because there is only one group)**

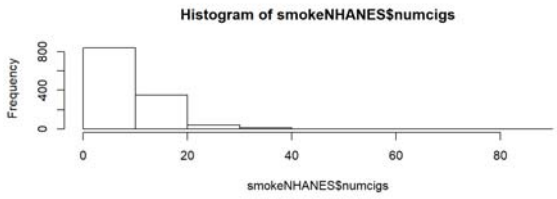
Testing the independence of observations

Independence of observations is not tested but instead is known based on how the data were collected.

Testing for normal distribution

Normal distribution is checked visually using a histogram or statistically using the Shapiro-Wilk test. First, check visually for the one-sample t-test:

```
# check the distribution of numcigs
hist(smokeNHANES$numcigs)
```



That is not normally distributed. The plot is right-skewed.

Statistical test of normality (Shapiro-Wilk)

The Shapiro-Wilk test tests the null hypothesis that the data are normally distributed for the variable of interest.

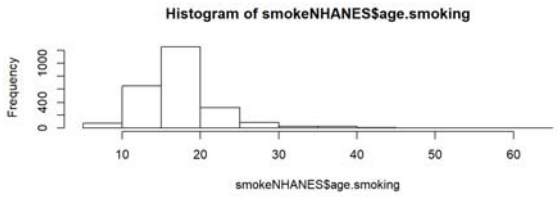
```
# check normality statistically with the Shapiro-Wilk test
shapiro.test(smokeNHANES$numcigs)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  smokeNHANES$numcigs
## W = 0.84839, p-value < 2.2e-16
```

The p-value less than .05, we reject the null hypothesis that numcigs is normally distributed. We conclude that numcigs is not normally distributed. This assumption is NOT MET.

Testing for normal distribution

Try the age.onset for the independent samples t-test and ANOVA:




```
##
##  Shapiro-Wilk normality test
##
## data:  smokeNHANES$age.smoking
## W = 0.84279, p-value < 2.2e-16
```

Doesn't look normally distributed by either the plot or the S-W test! Given the p-value for the S-W test is less than .05. This assumption is NOT MET.

Testing homogeneity of variance

- *Homogeneity of variance* is checked using the Levene Test.
- The Levene Test tests the null hypothesis that the variances are equal across groups.
- To meet this assumption we DO NOT want to reject the null hypothesis.
- Since this is about equal variance across groups, it is not relevant for the one-sample t-test because that test has only one group.

```
# Check the HoV of age.smoking from the independent samples t-test
leveneTest(age.smoking ~ sex, data = smokeNHANES)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value   Pr(>F)
## group    1  20.171 7.41e-06 ***
##       2446
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The small p-value indicates we *rejected* the null hypothesis that the variances are equal, so we do not meet this assumption. This assumption is NOT MET.

Testing the homogeneity of variance

```
# Check the HoV from the ANOVA
leveneTest(age.smoking ~ race.eth, data = smokeNHANES)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value  Pr(>F)
## group    4  2.9338 0.01963 *
##       2443
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The small p-value indicates we did reject the null, which means we do not meet this assumption. This assumption is NOT MET.

What happens when assumptions are not met?

So, none of the tests met their assumptions! When data *do not* meet the assumptions for a specific test there are 3 options:

- Use an alternative test (the non-parametric version)
- Transform the continuous variable to meet the normality assumption and try again
- Report your results as not meeting assumptions and therefore not generalizable beyond the sample

Alternative to the one-sample t-test

The non-parametric version of the one-sample t-test is the Sign Test. The Sign Test computes whether the median is statistically significantly different than some population or hypothesized median. So, instead of comparing the sample mean to the mu, the Sign Test compares the sample median to a population or hypothesized median, like this:

- H0: The median number of cigarettes smoked per day by smokers is 20 (med = 20).
- HA: The median number of cigarettes smoked per day by smokers is not 20 (med does not equal 20).

Conduct the Sign Test

```
# conduct the sign with the hypothesized
# population median as 20
# be sure the BSDA package is installed beforehand
library(BSDA)
SIGN.test(smokeNHANES$numcigs, md=20)
```

```
##
##  One-sample Sign-Test
##
## data:  smokeNHANES$numcigs
## s = 58, p-value < 2.2e-16
## alternative hypothesis: true median is not equal to 20
## 95 percent confidence interval:
##   9 10
## sample estimates:
## median of x
##          10
##
## Achieved and Interpolated Confidence Intervals:
##
##                   Conf.Level L.E.pt U.E.pt
## Lower Achieved CI     0.9494      9     10
## Interpolated CI       0.9500      9     10
## Upper Achieved CI     0.9558      9     10
```


Interpret the Sign Test

The s statistic is 58 and the p-value for the s-statistic is < 2.2e-16 or < .00000000000000022, which is lower than the standard cutoff of .05.

Interpretation: The Sign Test indicates that the NHANES participants do not come from a population that smokes a median of 20 cigarettes per day (s = 58; p < .05). The median in the sample is 10 with a 95% confidence interval of 9 to 10 cigarettes per day. The sample likely comes from a population of smokers who smoke a median of 9 to 10 cigarettes per day.

Alternative to the independent samples t-test

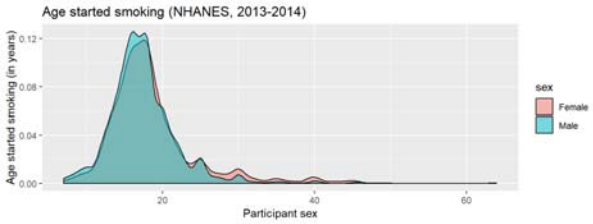
The non-parametric version of the independent samples t-test is the Mann-Whitney U test (also called the Wilcoxon test). Although sometimes interpreted as the difference in medians, the Mann-Whitney U test actually puts all the observations from both groups in order from lowest and highest and ranks them. The sums of the ranks for observations in the two groups are then compared. The null and alternate hypothesis would be:

- H0: Age of smoking initiation is equally distributed for males and females
- HA: Age of smoking initiation is not equally distributed for males and females

Exploring the data

Revisiting the distributions from earlier:

```
# compare male and female smoking age with density plot
ggplot(data = smokeNHANES, aes(fill = sex, x = age.smoking)) +
  geom_density(alpha=.5) +
  xlab("Participant sex") +
  ylab("Age started smoking (in years)") +
  ggtitle("Age started smoking (NHANES, 2013-2014)")
```



Conducting the Mann-Whitney U test

```
# testing for differences in distributions
wilcox.test(age.smoking ~ sex, data = smokeNHANES)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  age.smoking by sex
## W = 797420, p-value = 0.0001709
## alternative hypothesis: true location shift is not equal to 0
```

With a small p-value, the test appears to indicate that the null should be rejected. There is a statistically significant difference in the distribution of ages when males and females start smoking (W = 797420; p = .0002).

Examine the medians and make a conclusion

Since this is a non-parametric test for data that are not normally distributed, compute medians (instead of means) to add some context:

```
# tables of medians and IQR
library(tableone)
smoking.age.sex <- CreateTableOne(vars = "age.smoking",
                                  strata = "sex",
                                  data = smokeNHANES)
print(smoking.age.sex, nonnormal = "age.smoking")
```

```
##                              Stratified by sex
##                               Female               Male
##   n                            3714                 3454
##   age.smoking (median [IQR]) 18.00 [15.00, 20.00] 17.00 [15.00, 19.00]
##                              Stratified by sex
##                               p      test
##   n
##   age.smoking (median [IQR]) <0.001 nonnorm
```

We can conclude that there is a statistically significant difference (W = 797420; p = .0002) in the distribution of age of initation for males and females. Males likely start smoking at an earlier age (med = 17 years) than females (med = 18 years).

Alternative for ANOVA

Likewise, the non-parametric version of ANOVA is the *Kruskal-Wallis test*, which examines ranks across groups like the Mann-Whitney U test, but for more than two groups.

Try the Kruskal-Wallis test for age of smoking initiation by the race.eth variable:

- H0: The distribution of the age of smoking initiation is the same by race/ethnicity.
- HA: The distribution of the age of smoking initiation differs by race/ethnicity.

In R, the two arguments for this test are x, which is the continuous variable and g, which is the groups. The command can fail if the groups variable is not a factor data type, so you can add as.factor to ensure it will run.

```
# examine distributions across groups
kruskal.test(x = smokeNHANES$age.smoking, g = as.factor(smokeNHANES$race.eth))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  smokeNHANES$age.smoking and as.factor(smokeNHANES$race.eth)
## Kruskal-Wallis chi-squared = 61.586, df = 4, p-value = 1.346e-12
```

Interpreting Kruskal-Wallis

In this case, it appears to be significant. So, we can conclude that there is a statistically significant difference (K-W chi-squared = 61.59; p < .05) in the distribution of age of smoking initiation by race/ethnicity. A table can add some additional context:

```
##                            Stratified by race.eth
##                             Black Non-Hisp       Mexican-Amer
##   n                          1520                 1096
##   age.smoking (median [IQR]) 18.00 [16.00, 21.00] 17.00 [15.00, 19.00]
##                            Stratified by race.eth
##                             Other Hispanic       Other Race
##   n                           651                 1073
##   age.smoking (median [IQR]) 17.00 [15.00, 19.00] 18.00 [16.00, 21.00]
##                            Stratified by race.eth
##                             White Non-Hisp       p      test
##   n                          2828
##   age.smoking (median [IQR]) 17.00 [15.00, 19.00] <0.001 nonnorm
```

It appears that Mexican-Americans, Non-Hispanic white, and Other Hispanic all have median age of smoking initiation of 17 years old while Non-Hispanic Black and Other Race participants have a median of 18 years old for smoking initiation.

Post-hoc test for Mann-Whitney U

Like Bonferroni for ANOVA, use a Dunn's test to compare distributions for each pair of groups:

```
# be sure dunn.test is installed
# open the dunn.test package
# enter the continuous variable as x, the categorical as g (groups)
library(dunn.test)
dunn.test(x = smokeNHANES$age.smoking, g = smokeNHANES$race.eth)
```


Dunn test output

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 61.5856, df = 4, p-value = 0
##
##
##                            Comparison of x by group
##                                 (No adjustment)
## Col Mean-|
## Row Mean |   Black No   Mexican-   Other Hi   Other Ra
## ---------+--------------------------------------------
## Mexican- |   3.700592
##          |    0.0001*
##          |
## Other Hi |   2.370583  -0.877067
##          |    0.0089*     0.1902
##          |
## Other Ra |  -0.647326  -3.749900  -2.612287
##          |     0.2587    0.0001*    0.0045*
##          |
## White No |   6.455053   0.818430   1.808927   5.619179
##          |    0.0000*     0.2066     0.0352    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha/2
```


Interpret post-hoc test

The Dunn's post-hoc test indicates that the age of smoking initiation distribution is significantly (p < .05) different for Non-Hispanic Blacks (med = 18) compared to Mexican-Americans (med = 17), Other Hispanics (med = 17), and Non-Hispanic Whites (med = 17). Other Race participants (med = 18) are also statistically significantly different (p < .05) from Other Hispanic and Non-Hispanic Whites for age of smoking initiation. Non-Hispanic Blacks and Other Race participants start smoking later (med = 18) than the other groups (med = 17).

THE END

- The challenge is on GitHub
- The RMD used to create these slides is on GitHub and includes all the code used throughout and extra annotation for any new commands or new command options along with some fancy [optional] options for writing R Markdown documents for those of you who are interested!