431 Class 14

thomaselove.github.io/431

2020-10-08

Today's Agenda

- p values and Researcher Degrees of Freedom
  - The "Garden of Forking Paths"
- Comparing Means using Two Independent Samples
  - Regression models to obtain pooled t comparisons
  - Welch's t test (not assuming equal population variances)
  - Bootstrap methods for comparing means in 2 samples
  - Rank-based alternatives (Wilcoxon-Mann-Whitney)
- Comparing More than Two Means with ANOVA
  - using regression to compare more than two population means

Today's R Packages and Data

```
library(broom)
library(janitor)
library(knitr)
library(magrittr)
library(patchwork)
library(tidyverse)

theme_set(theme_bw())

source("data/Love-boost.R") # new today!

dm431 <- readRDS("data/dm431_2020.Rds")
```

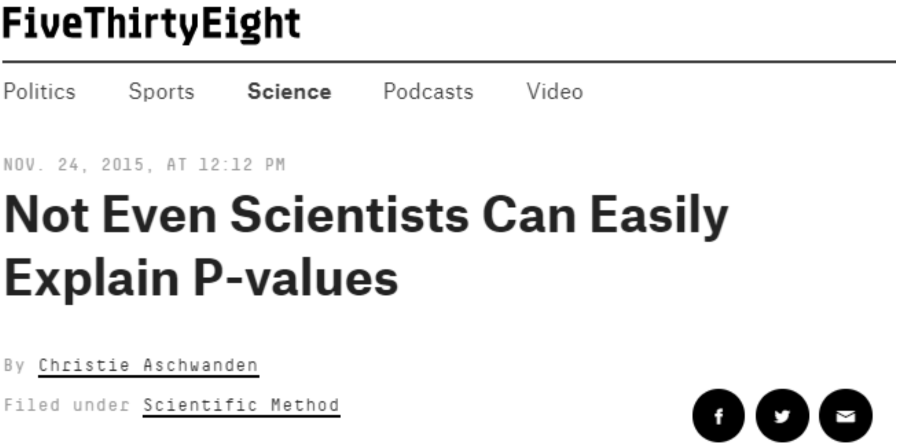
What is a p value?

From FiveThirtyEight

> *P-values have taken quite a beating lately. These widely used and commonly misapplied statistics have been blamed for giving a veneer of legitimacy to dodgy study results, encouraging bad research practices and promoting false-positive study results.*
>
> *Last week, I attended the inaugural METRICS conference at Stanford, which brought together some of the world's leading experts on meta-science, or the study of studies. I figured that if anyone could explain p-values in plain English, these folks could.*

(Christie Aschwanden, FiveThirtyEight, 2015-11-24)

Let's Go To The Videotape



FiveThirtyEight

Politics Sports **Science** Podcasts Video

NOV. 24, 2015, AT 12:12 PM

Not Even Scientists Can Easily Explain P-values

By Christie Aschwanden

Filed under Scientific Method

Link: https://fivethirtyeight.com/features/not-even-scientists-can-easily-explain-p-values/

Do Scientists Get This Right?

> *Scientists regularly get it wrong, and so do most textbooks, said Steven Goodman (co-director of the METRICS conference.) Even after spending his "entire career" thinking about p-values, he said he could tell me the definition, "but I cannot tell you what it means, and almost nobody can."*
>
> *When Goodman speaks to large audiences of scientists, he often presents correct and incorrect definitions of the p-value, and they "very confidently" raise their hand for the wrong answer. "Almost all of them think it gives some direct information about how likely they are to be wrong, and that's definitely not what a p-value does," Goodman said.*

(Christie Aschwanden, FiveThirtyEight, 2015-11-24)

Can we define a p value better?

> *I've come to think that the most fundamental problem with p-values is that no one can really say what they are.*
>
> *What I learned by asking all these very smart people to explain p-values is that I was on a fool's errand. Try to distill the p-value down to an intuitive concept and it loses all its nuances and complexity, said science journalist Regina Nuzzo, a statistics professor at Gallaudet University. "Then people get it wrong, and this is why statisticians are upset and scientists are confused."*
>
> *You can get it right, or you can make it intuitive, but it's all but impossible to do both.*

(Christie Aschwanden, FiveThirtyEight, 2015-11-24)

Last time, we built a 90% CI for μ = population mean SBP. . .

Use indicator variable regression to produce a t-based interval.

```
model1 <- lm(sbp ~ 1, data = dm431)

tidy(model1, conf.int = TRUE, conf.level = 0.90) %>%
  select(term, estimate, std.error,
         conf.low, conf.high, p.value) %>%
  knitr::kable(digits = 2)
```

term	estimate	std.error	conf.low	conf.high	p.value
(Intercept)	131.26	0.89	129.79	132.74	0

- Our point estimate for the population mean SBP (μ) is 131.26 mm Hg.
- Our 90% confidence interval is (129.79, 132.74) mm Hg for μ.

Assumptions of a t-based Confidence Interval

> *"Begin challenging your assumptions. Your assumptions are your windows on the world. Scrub them off every once in awhile or the light won't come in." (Alan Alda)*

1. Sample is drawn at random from the population or process.
2. Samples are drawn independently from each other from a population or process whose distribution is unchanged during the sampling process.
3. Population or process follows a Normal distribution.

Can we drop any of these assumptions?

Only if we're willing to consider alternative inference methods.

Confidence Intervals using Bootstrap Resampling

Resampling is A Big Idea

If we want our sample mean to accurately estimate the population mean, we would ideally like to take a very, very large sample, so as to get very precise estimates. But we can rarely draw enormous samples. So what can we do?

Oversimplifying, the idea is that if we sample (with replacement) from our current data, we can draw a new sample of the same size as our original.

- And if we repeat this many times, we can generate as many samples of "431 systolic blood pressures" as we like.
- Then we take these thousands of samples and calculate (for instance) the sample mean for each, and plot a histogram of those means.
- If we then cut off the top and bottom 5% of these sample means, we obtain a reasonable 90% confidence interval for the population mean.

Bootstrap 90% confidence interval

The bootstrap can be used to build a confidence interval for μ without the assumption that the population follows a Normal distribution.

```
set.seed(2020)
Hmisc::smean.cl.boot(dm431$sbp, conf.int = .90, B = 1000)
```

```
    Mean    Lower    Upper 
131.2645 129.8165 132.7594 
```

- I often use B = 1,000 (the default) or 10,000 bootstrap replications for building CIs, but it isn't usually critical.
- A bootstrap interval is often asymmetric, and for highly skewed data, the point estimate might not be near the center of the interval.

One “Downside” of the Bootstrap

We get (somewhat) different answers if we resample the data with a new seed.

```
set.seed(2020); Hmisc::smean.cl.boot(dm431$sbp, conf.int = .9)
```

```
    Mean    Lower    Upper 
131.2645 129.8165 132.7594 
```

```
set.seed(12); Hmisc::smean.cl.boot(dm431$sbp, conf.int = .90)
```

```
    Mean    Lower    Upper 
131.2645 129.7350 132.6775 
```

```
set.seed(431); Hmisc::smean.cl.boot(dm431$sbp, conf.int = .9)
```

```
    Mean    Lower    Upper 
131.2645 129.8046 132.7290 
```

What changes when we set the seed?

Bootstrap vs. t-Based Confidence Intervals

- `Hmisc`'s `smean.cl.boot` function (unlike most R functions) deletes missing data automatically, as does the `smean.cl.normal` function, which produces the t-based confidence interval.

```r
set.seed(431)
dm431 %$% Hmisc::smean.cl.boot(sbp, conf = 0.90)
```

```
    Mean    Lower    Upper 
131.2645 129.8046 132.7290 
```

```r
dm431 %$% Hmisc::smean.cl.normal(sbp, conf = 0.90)
```

```
    Mean    Lower    Upper 
131.2645 129.7940 132.7350 
```

Bootstrap: Estimating a confidence interval for μ

What the computer does:

1. Resample the data with replacement, until it obtains a new sample that is equal in size to the original data set.
2. Calculates the statistic of interest (here, a sample mean.)
3. Repeat the steps above many times (default is 1,000 with our approach) to obtain a set of 1,000 results (here: 1,000 sample means.)
4. Sort those 1,000 results in order, and estimate the 90% confidence interval for the population value based on the middle 90% of the 1,000 bootstrap samples.
5. Send us a result, containing the sample estimate, and the bootstrap 90% confidence interval estimate for the population value.

The bootstrap idea can be used to produce interval estimates for almost any population parameter, not just the mean.

What about p values?

```
dm431 %$% Hmisc::smean.cl.normal(sbp, conf = 0.90)
```

```
    Mean    Lower    Upper 
131.2645 129.7940 132.7350 
```

```
set.seed(431); dm431 %$% Hmisc::smean.cl.boot(sbp, conf = 0.9
```

```
    Mean    Lower    Upper 
131.2645 129.8046 132.7290 
```

1. What can we say about the p value for $H_0 : \mu = 0$ vs. $H_A : \mu \neq 0$ based on this bootstrap? How about based on the t-distribution CI?
2. What can we say about the p value for $H_0 : \mu = 130$ vs. $H_A : \mu \neq 130$ based on the bootstrap? How about based on the t-distribution CI?

When is a Bootstrap CI for μ Reasonable?

The interval will be reasonable as long as we are willing to believe that:

- the original sample was a random sample (or at least a completely representative sample) from a population,
- and that the samples are independent of each other (selecting one subject doesn't change the probability that another subject will also be selected)
- and that the samples are identically distributed (even though that distribution may not be Normal.)

It is still possible that the results can be both:

- **inaccurate** (i.e. they can, include the true value of the unknown population mean less often than the stated confidence probability) and
- **imprecise** (i.e., they can include more extraneous values of the unknown population mean than is desirable).

The Wilcoxon Signed Rank Approach (if the data come from a symmetric population)

The Wilcoxon Signed Rank Procedure for CIs

The Wilcoxon signed rank approach can be used as an alternative to build interval estimates for the population *pseudo-median* when the population cannot be assumed to follow a Normal distribution.

What is a Pseudo-Median?

- For any sample, the pseudomedian is defined as the median of all of the midpoints of pairs of observations in the sample.
- As it turns out, if you're willing to assume the population is **symmetric** (but not necessarily Normally distributed) then the pseudo-median is equal to the population median.

Wilcoxon rank sum based 90% confidence interval

```
wilcox.test(dm431$sbp, conf.int = TRUE, conf.level = 0.90)
```

```

    Wilcoxon signed rank test with continuity
    correction

data:  dm431$sbp
V = 93096, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 0
90 percent confidence interval:
 129.0 131.5
sample estimates:
(pseudo)median
           130
```

Interpreting the Wilcoxon Signed Rank CI

If we're willing to believe the sbp values come from a population with a symmetric distribution, the 90% Confidence Interval for the population median would be (129, 131.5)

For a non-symmetric population, this only applies to the *pseudo-median*.

Note that the pseudo-median is actually fairly close in this situation to the sample mean as well as to the sample median, as it usually will be if the population actually follows a symmetric distribution, as the Wilcoxon approach assumes.

```r
mosaic::favstats(~ sbp, data = dm431)
```

```
Registered S3 method overwritten by 'mosaic':
  method                           from
  fortify.SpatialPolygonsDataFrame ggplot2

 min  Q1 median  Q3 max     mean       sd   n missing
  90 120    130 141 208 131.2645 18.52038 431       0
```

Tidying the Wilcoxon Results

```r
w1 <- dm431 %$%
  wilcox.test(sbp, conf.int=TRUE, conf.level=0.90)

tidy(w1) %>% select(method, alternative) %>% kable()
```

method	alternative
Wilcoxon signed rank test with continuity correction	two.sided

```r
tidy(w1) %>% select(estimate, conf.low, conf.high, p.value) %>
```

estimate	conf.low	conf.high	p.value
130	129	131.5	0

Three Methods for Estimating a Single Population Mean

For estimating the population mean…

1. A **t-based** estimate and confidence interval, available from an intercept-only linear model, or (equivalently) from a t test.
    - This approach will require an assumption that the population comes from a Normal distribution.
2. A **bootstrap** confidence interval, which uses resampling to estimate the population mean.
    - This approach won't require the Normality assumption, but has some other constraints.
3. A **Wilcoxon signed rank** approach, but that won't describe the mean, only a pseudo-median.
    - This also doesn't require the Normality assumption, but no longer describes the population mean (or median) unless the population can be assumed symmetric. Instead it describes the *pseudo-median*.

Conclusions in our study of SBP in `dm431`

`dm431` sample mean was 131.26, and sample median was 130.

Approach	$H_A : \mu \neq 0$	90% CI for μ
t-test	p $<$ 0.0001	(129.79, 132.74)
bootstrap	p $<$ 0.10	(129.80, 132.73)

Approach	$H_A : psmed \neq 0$	90% CI for *psmed*
Wilcoxon signed rank	p $<$ 0.0001	(129, 131.5)

- *psmed* = population pseudo-median
- Bootstrap with `set.seed = 431`

p Hacking and “Researcher Degrees of Freedom”

p values?

Question 1. If the *p* value is smaller than our pre-specified α level, then we can declare the results to be statistically significant and celebrate?

Well, no.

Question 2. What if the *p* value is greater than our α, say $p > 0.05$? Then what?

P > 0.05
GAME OVER, TRY AGAIN
imgflip.com

What is a *p* value?

> *The probability of getting results at least as extreme as the ones you observed, given that the null hypothesis is correct.*

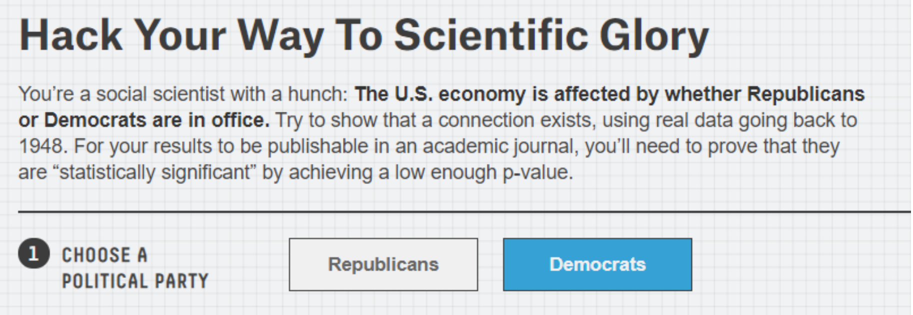
It's a conditional probability statement. That's all.

> *We want to know if results are right, but a p-value doesn't measure that. It can't tell you the magnitude of an effect, the strength of the evidence or the probability that the finding was the result of chance.*

Quotes from Christie Aschwanden "Not Even Scientists Can Easily Explain P-values", at FiveThirtyEight.com on 2015-11-24

Hack Your Way To Scientific Glory

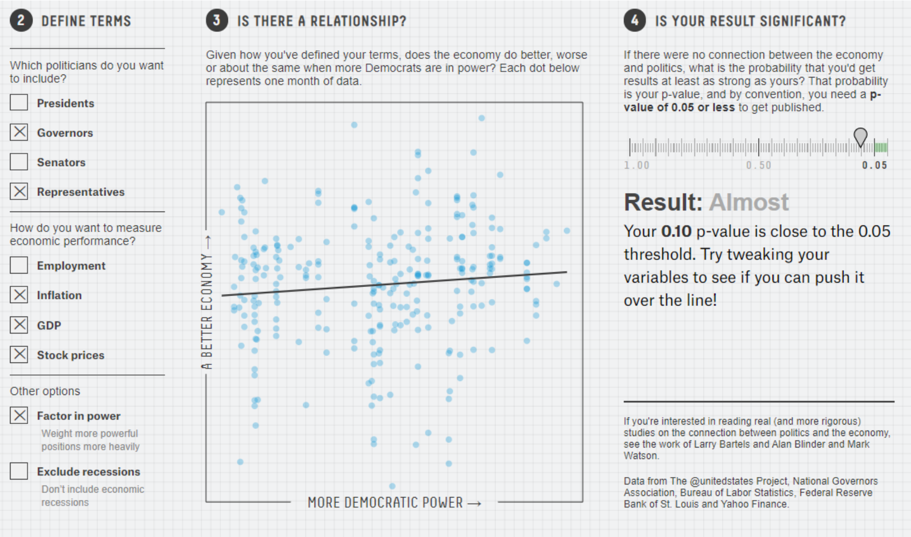
https://fivethirtyeight.com/features/science-isnt-broken



Hack Your Way To Scientific Glory

You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

What will you be doing?



2 DEFINE TERMS

Which politicians do you want to include?

☐ **Presidents**

☒ **Governors**

☐ **Senators**

☒ **Representatives**

How do you want to measure economic performance?

☐ **Employment**

☒ **Inflation**

☒ **GDP**

☒ **Stock prices**

Other options

☒ **Factor in power**

Weight more powerful positions more heavily

☐ **Exclude recessions**

Don't include economic recessions

3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Democrats are in power? Each dot below represents one month of data.

4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a **p-value of 0.05 or less** to get published.

Result: **Almost**

Your **0.10** p-value is close to the 0.05 threshold. Try tweaking your variables to see if you can push it over the line!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

Data from The @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

We'll give you a few minutes.

Again, the link is https://fivethirtyeight.com/features/science-isnt-broken

1. Choose a party to frame your argument (Democrats or Republicans)
2. Define your terms (as a group) in the way you feel is most interesting. What is the resulting p value and conclusion?
3. Switch things up by defining new terms, switching parties or whatever you like.

   - Can you obtain a p value that supports the notion that your party has a positive effect on the economy?
   - Can you obtain a p value that supports the notion that your party has a negative effect on the economy?

4. What is the range (minimum to maximum) of p values you obtained?

Go!

So, what happened?

What can you get?

In just a few minutes, I was able to get

- $p < 0.01$ (*positive* effect of Democrats on economy)
- $p < 0.01$ (*negative* effect of Democrats on economy)
- $p = 0.02$ (positive effect of Democrats)
- $p = 0.02$ (negative effect of Democrats)

but also . . .

- $p = 0.06$, 0.11, 0.21, 0.39, 0.47, 0.54, 0.65, 0.78 and even p $>$ 0.99

without even switching parties (I chose to frame this in terms of the Democrats), just by checking different boxes to define my terms (section 2 of the graphic.)

“Researcher Degrees of Freedom”, 1

> *[I]t is unacceptably easy to publish “statistically significant” evidence consistent with any hypothesis.*
>
> *The culprit is a construct we refer to as **researcher degrees of freedom**. In the course of collecting and analyzing data, researchers have many decisions to make: Should more data be collected? Should some observations be excluded? Which conditions should be combined and which ones compared? Which control variables should be considered? Should specific measures be combined or transformed or both?*

Simmons et al. *link*

"Researcher Degrees of Freedom", 2

> *. . . It is rare, and sometimes impractical, for researchers to make all these decisions beforehand. Rather, it is common (and accepted practice) for researchers to explore various analytic alternatives, to search for a combination that yields statistical significance, and to then report only what worked. The problem, of course, is that the likelihood of at least one (of many) analyses producing a falsely positive finding at the 5% level is necessarily greater than 5%.*

For more, see

- Gelman's blog 2012 – 11 – 01 "Researcher Degrees of Freedom",
- Paper by *Simmons* and others, defining the term.

And this is really hard to deal with. . .

The garden of forking paths: Why multiple comparisons can be a problem, even when there is no fishing expedition or p-hacking and the research hypothesis was posited ahead of time

> *Researcher degrees of freedom can lead to a multiple comparisons problem, even in settings where researchers perform only a single analysis on their data. The problem is there can be a large number of potential comparisons when the details of data analysis are highly contingent on data, without the researcher having to perform any conscious procedure of fishing or examining multiple p-values. We discuss in the context of several examples of published papers where data-analysis decisions were theoretically-motivated based on previous literature, but where the details of data selection and analysis were not pre-specified and, as a result, were contingent on data.*

- *Link* to the paper from Gelman and Loken

Comparing Population Means using Two Independent Samples

Two Examples Comparing Two Means with `dm431`

Our population: ALL adults ages 31-70 seen for care this year and two years ago who live in Northeast Ohio with a diabetes diagnosis.

Our `dm431` sample: 431 of those people, drawn in a way we hope is representative (but certainly isn't random).

1. Can we estimate the difference in the population mean LDL cholesterol for those who have a statin prescription as compared to those who do not?
2. Can we estimate the difference between females and males in terms of the population mean systolic blood pressure?

Today's Plan

1. We'll walk through example 1 (the harder example, as it turns out.)
2. Example 2 slides follow Example 1, for you to review on your own.

dm431 Example 1 (Comparing LDL by Statin usage)

Estimate the difference in the population mean LDL cholesterol for those who have a statin prescription as compared to those who do not.

```
mosaic::favstats(ldl ~ statin, data = dm431) %>%
  kable(digits = 2)
```

statin	min	Q1	median	Q3	max	mean	sd	n	missing
0	31	76	98.0	114.5	177	97.42	29.22	72	14
1	36	70	88.5	113.0	227	96.41	35.33	322	23

- What is the outcome here?

dm431 Example 1 (Comparing LDL by Statin usage)

Estimate the difference in the population mean LDL cholesterol for those who have a statin prescription as compared to those who do not.

```r
mosaic::favstats(ldl ~ statin, data = dm431) %>%
  kable(digits = 2)
```

statin	min	Q1	median	Q3	max	mean	sd	n	missing
0	31	76	98.0	114.5	177	97.42	29.22	72	14
1	36	70	88.5	113.0	227	96.41	35.33	322	23

- What is the outcome here?
- What are the two exposure groups we are comparing?

dm431 Example 1 (Comparing LDL by Statin usage)

Estimate the difference in the population mean LDL cholesterol for those who have a statin prescription as compared to those who do not.

```r
mosaic::favstats(ldl ~ statin, data = dm431) %>%
  kable(digits = 2)
```

statin	min	Q1	median	Q3	max	mean	sd	n	missing
0	31	76	98.0	114.5	177	97.42	29.22	72	14
1	36	70	88.5	113.0	227	96.41	35.33	322	23

- What is the outcome here?
- What are the two exposure groups we are comparing?
- What are the sample means, $\bar{x}_{Statin}$ and $\bar{x}_{NoStatin}$?

dm431 Example 1 (Comparing LDL by Statin usage)

Estimate the difference in the population mean LDL cholesterol for those who have a statin prescription as compared to those who do not.

```
mosaic::favstats(ldl ~ statin, data = dm431) %>%
  kable(digits = 2)
```

statin	min	Q1	median	Q3	max	mean	sd	n	missing
0	31	76	98.0	114.5	177	97.42	29.22	72	14
1	36	70	88.5	113.0	227	96.41	35.33	322	23

- What is the outcome here?
- What are the two exposure groups we are comparing?
- What are the sample means, $\bar{x}_{Statin}$ and $\bar{x}_{NoStatin}$?
- How might we estimate the difference in population means, $\mu_S - \mu_N$?

dm431 Example 1 (Comparing LDL by Statin usage)

Estimate the difference in the population mean LDL cholesterol for those who have a statin prescription as compared to those who do not.

```
mosaic::favstats(ldl ~ statin, data = dm431) %>%
  kable(digits = 2)
```

statin	min	Q1	median	Q3	max	mean	sd	n	missing
0	31	76	98.0	114.5	177	97.42	29.22	72	14
1	36	70	88.5	113.0	227	96.41	35.33	322	23

- What is the outcome here?
- What are the two exposure groups we are comparing?
- What are the sample means, $\bar{x}_{Statin}$ and $\bar{x}_{NoStatin}$?
- How might we estimate the difference in population means, $\mu_S - \mu_N$?
- Is there a problem in these data we need to deal with?

How much missing data do we have?

Do we have missing values in both columns, or just one?

```
dm431 %>% summarize(across(c(statin, ldl), ~ sum(is.na(.x))))
```

```
# A tibble: 1 x 2
  statin   ldl
   <int> <int>
1      0    37
```

So what shall we do?

- Drop the 37 cases, or
- Something else?

On Missing Data

Drop the Missing = A "Complete Case" analysis

- We could drop these 37, and do a **complete case analysis** on the other 431-37 = 394 subjects.
- We'll also create a factor (`statin_f`) with the statin information.

```
dm431_cc <- dm431 %>% filter(complete.cases(ldl, statin)) %>%
  mutate(statin_f = fct_recode(factor(statin),
                          "Statin" = "1", "No" = "0"))

mosaic::favstats(ldl ~ statin_f, data = dm431_cc) %>%
  kable(dig = 2)
```

statin_f	min	Q1	median	Q3	max	mean	sd	n	missing
No	31	76	98.0	114.5	177	97.42	29.22	72	0
Statin	36	70	88.5	113.0	227	96.41	35.33	322	0

- HUGE assumption: The 37 missing `ldl` are MCAR.

Missing Completely at Random (MCAR)

Our complete case analysis requires the HUGE assumption that these 37 observations are what Donald Rubin called "missing completely at random."

Missing Completely at Random (MCAR) means that there is no relationship between whether a data point is missing and any values in the data set, missing or observed. Thus, the missing values are just a random subset of the data.

- That is the huge assumption that is both impossible to prove and that is also tacitly made in many settings, more or less by default.
- The alternative is to consider other possible mechanisms (besides MCAR) for why data might be missing.

Assuming data are Missing at Random (MAR)?

Missing at Random (MAR): the reason a data point is missing is related to some observed data, but unrelated to the actual missing values.

So we assume that we can predict the missing values effectively using other variables in the data, without causing any problems. That's a big assumption, but then we could *impute* (or fill in with predictions based on other variables) the missing data.

So to impute predicted `ldl` values for these 37 subjects, we'd need to:

- account for the fact that we're imputing in building estimates, and
- control for the variables which (together) predict why the data were missing, and
- remember that we are making a large and unverifiable assumption about why the data are missing.

If missing data aren't MCAR or MAR, then they are MNAR.

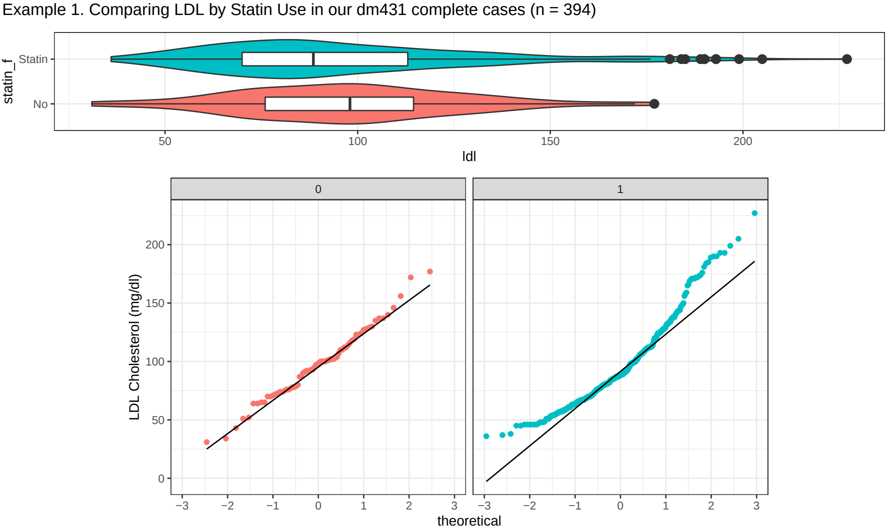
Three Types of Missingness

1. MCAR: Missing Completely At Random (ignorable nonresponse)
    - missing values are just a random subset of the data
    - unrealistically strong assumption in practice, although it's easy
    - makes a complete case analysis unbiased
2. MAR: Missing At Random
    - reason for missingness can be completely accounted for by variables where there is complete information
    - much more reasonable in many settings than MCAR, but impossible to verify statistically
    - imputing missing values here leads to a more robust conclusion
3. MNAR: Missing Not at Random (nonignorable nonresponse)
    - data are neither MCAR nor MAR
    - the reason the data is missing is related to its value, even after controlling for other variables.

These have different effects on the validity of the conclusions you build.

DTDP: Example 1. (Comparing LDL by Statin Use)

Assuming MCAR, we'll press on with a complete case analysis.



Linear Model for Example 1 (slide A)

Estimate the difference in population mean LDL cholesterol among people taking a statin as compared to those not taking a statin.

```
app1 <- lm(ldl ~ statin, data = dm431_cc)

tidy(app1, conf.int = T, conf.level = 0.90) %>% kable(dig = 2)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	97.42	4.04	24.09	0.00	90.75	104.08
statin	-1.01	4.47	-0.23	0.82	-8.38	6.36

- What can we learn from this output?
  - What is the sample mean `ldl` for those not on a statin?
  - What is the sample mean `ldl` for statin users?
  - The point estimate for $\mu_S - \mu_N$ is . . .

Linear Model for Example 2 (slide B)

Estimate the difference in population mean LDL cholesterol among people taking a statin as compared to those not taking a statin.

```
app1 <- lm(ldl ~ statin, data = dm431_cc)

tidy(app1, conf.int = T, conf.level = 0.90) %>% kable(dig = 2)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	97.42	4.04	24.09	0.00	90.75	104.08
statin	-1.01	4.47	-0.23	0.82	-8.38	6.36

- What can we learn from this output?
  - The point estimate for $\mu_S - \mu_N$ is **-1.01**
  - The 90% confidence interval for $\mu_S - \mu_N$ is . . .

Linear Model for Example 2 (slide C)

Estimate the difference in population mean LDL cholesterol among people taking a statin as compared to those not taking a statin.

```
app1 <- lm(ldl ~ statin, data = dm431_cc)

tidy(app1, conf.int = T, conf.level = 0.90) %>% kable(dig = 2)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	97.42	4.04	24.09	0.00	90.75	104.08
statin	-1.01	4.47	-0.23	0.82	-8.38	6.36

- What can we learn from this output?
  - The point estimate for $\mu_S - \mu_N$ is -1.01
  - The 90% confidence interval for $\mu_S - \mu_N$ is (-**8.38**, **6.36**)

Building Confidence Intervals for $\mu_1 - \mu_2$

The hypotheses we are testing are (Δ_0 is usually zero):

- H_0: $\mu_1 = \mu_2 +$ hypothesized difference Δ_0 vs.
- H_A: $\mu_1 \neq \mu_2 +$ hypothesized difference Δ_0.

Four Approaches

1. Indicator Variable Regression Model ("Pooled" t approach, or "t test" assuming equal population variances)
2. Welch t CI (t approach without assuming equal population variances)
3. Wilcoxon-Mann-Whitney Rank Sum Test (non-parametric test not assuming Normality but needing symmetry to be related to means)
4. Bootstrap confidence interval for the difference in population means (fewest assumptions of these options)

The Pooled t procedure (same as indicator variable regression)

Building a Pooled t CI

1. Best approach: use indicator variable regression
2. Also: direct call to t test with pooled variance estimate

```
t.test(ldl ~ statin, data = dm431_cc, alt = "two.sided", mu =
       var.equal = TRUE, conf.level = 0.90)
```

```

    Two Sample t-test

data:  ldl by statin
t = 0.22579, df = 392, p-value = 0.8215
alternative hypothesis: true difference in means is not equal
90 percent confidence interval:
 -6.363975  8.383644
sample estimates:
mean in group 0 mean in group 1
       97.41667        96.40683
```

`t` test can be tidied

```r
t1 <- tidy(t.test(ldl ~ statin, data = dm431_cc,
                  var.equal = TRUE, conf.level = 0.90))
```

- `conf.level` must be specified to `t.test`. Otherwise, it uses 0.95.

Elements of `t1` are printed below (after rearrangement)

method	alternative	estimate1	estimate2
Two Sample t-test	two.sided	97.42	96.41

estimate	conf.low	conf.high	statistic	parameter	p.value
1.01	-6.36	8.38	0.23	392	0.82

- This estimates $\mu_{NoStatin} - \mu_{Statin}$. Invert the signs of the estimate and the endpoints of the CI to estimate $\mu_{Statin} - \mu_{NoStatin}$.

Assumptions of the Pooled T test

The standard method for comparing population means based on two independent samples is based on the t distribution, and requires the following assumptions:

1. [Independence] The samples for the two groups are drawn independently.
2. [Random Samples] The samples for each of the groups are drawn at random from the populations of interest.
3. [Normal Population] The two populations are each Normally distributed
4. [Equal Variances or Balanced Design] We must assume:

- Either the population variances in the two groups are the same, so a pooled estimate of their joint variance makes sense,
- OR the two samples are the same size (a balanced design.)

The Welch t procedure (t approach, not assuming equal population variances)

Assumptions of the Welch t approach

The Welch test still requires:

1. [Independence] The samples for the two groups are drawn independently.
2. [Random Samples] The samples for each of the groups are drawn at random from the populations of interest.
3. [Normal Population] The two populations are each Normally distributed

But it doesn't require:

4. [Equal Variances] The population variances in the two groups being compared are the same. (for instance, Welch's test still works if the larger variance σ_1^2 is more than 1.5 times as large as σ_2^2).

- If the design is balanced $(n_1 = n_2)$ or nearly so, the impact of assuming equal variances is minimal.

Welch's t test is the default `t.test` in R.

Building the Welch t CI

- Sensible approach when assuming Normal populations is OK, but we don't want to assume the two populations have the same variance (as pooled t requires)

```
t.test(ldl ~ statin, data = dm431_cc, alt = "two.sided", mu =
       conf.level = 0.90)
```

```

    Welch Two Sample t-test

data:  ldl by statin
t = 0.25455, df = 122.11, p-value = 0.7995
alternative hypothesis: true difference in means is not equal
90 percent confidence interval:
 -5.565462  7.585131
sample estimates:
mean in group 0 mean in group 1
       97.41667        96.40683
```

Welch `t` test can also be tidied

```
t2 <- tidy(t.test(ldl ~ statin, data = dm431_cc,
                  conf.level = 0.90))
```

- We must specify `conf.level` in the `t.test` unless we want 0.95.

Elements of `t2` are printed below (after rearrangement)

method	alternative	estimate1	estimate2
Welch Two Sample t-test	two.sided	97.42	96.41

estimate	conf.low	conf.high	statistic	parameter	p.value
1.01	-5.57	7.59	0.25	122.11	0.8

- Invert signs of estimate and CI limits to get $\mu_{Statin} - \mu_{No}$.

The Wilcoxon-Mann-Whitney Rank Sum procedure

Wilcoxon-Mann-Whitney Rank Sum Approach

The Wilcoxon-Mann-Whitney Rank Sum procedure requires:

1. [Independence] The samples for the two groups are drawn independently.
2. [Random Samples] The samples for each of the groups are drawn at random from the populations of interest.
3. [Symmetry] The two populations are each symmetrically distributed, and as a result, we're comfortable estimating the shift in location (measured by the pseudo-medians) rather than a shift in means.

But it doesn't require:

3. [Normal Population] The two populations are each Normally distributed
4. [Equal Variances] The population variances in the two groups being compared are the same.

As mentioned, it doesn't really compare population means, but instead pseudo-medians.

Wilcoxon-Mann-Whitney Rank Sum Approach

```r
wilcox.test(ldl ~ statin, data = dm431_cc,
            conf.int = TRUE, conf.level = 0.90)
```

```
    Wilcoxon rank sum test with continuity
    correction

data:  ldl by statin
W = 12560, p-value = 0.2683
alternative hypothesis: true location shift is not equal to 0
90 percent confidence interval:
 -2.000009 11.000062
sample estimates:
difference in location
              4.629345
```

Rank Sum test can also be tidied

```
w3 <- tidy(wilcox.test(ldl ~ statin, data = dm431_cc,
            conf.int = TRUE, conf.level = 0.90))
```

- Specify `conf.int` and `conf.level` in the `wilcox.test`.

Elements of `w3` are printed below (after rearrangement)

method	alternative	statistic
Wilcoxon rank sum test with continuity correction	two.sided	12559.5

estimate	conf.low	conf.high	p.value
4.63	-2	11	0.27

- Invert signs of estimate and CI to describe shift from No to Statin.

The Bootstrap

This bootstrap approach to comparing population means using two independent samples still requires:

1. [Independence] The samples for the two groups are drawn independently.
2. [Random Samples] The samples for each of the groups are drawn at random from the populations of interest.

but does not require either of the other two assumptions:

3. [Normal Population] The two populations are each Normally distributed
4. [Equal Variances] The population variances in the two groups being compared are the same.

The bootstrap procedure I use in R was adapted from Frank Harrell and colleagues. http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/BootstrapMeansSoftware

The bootdif function

The procedure requires the definition of a function, which I have adapted a bit, called bootdif, which is part of the Love-boost.R script we loaded earlier.

As in our previous bootstrap procedures, we are sampling (with replacement) a series of many data sets (default: 2000).

- Here, we are building bootstrap samples based on the LDL levels in the two independent samples (statin users vs. non-users.)
- For each bootstrap sample, we are calculating a mean difference between the two groups (statin vs. no statin.)
- We then determine the 2.5th and 97.5th percentile of the resulting distribution of mean differences (for a 95% confidence interval).

Using `bootdif` to compare mean(LDL) by statin

So, to compare LDL (our outcome) across the two levels of statin (our grouping factor) for the adult patients with diabetes in NE Ohio, run…

```r
set.seed(20201008)
boot4 <- dm431_cc %$% bootdif(ldl, statin, conf.level = 0.90)
boot4
```

```
Mean Difference             0.05             0.95
      -1.009834        -7.580275         5.326272
```

- The two columns must be separated here with a comma rather than a tilde (~), and are specified using $ notation.
- This CI estimates $\mu_{Statin} - \mu_{NoStatin}$. Observe the listed sample mean difference for the necessary context.
- If we change the `set.seed`, we'll get different endpoints for our CI.
- Note that we can infer the *p* value is above 0.10 from the CI. Why?

Results for the LDL and Statin Study

Procedure	p for $H_0 : \mu_S = \mu_N$	90% CI for $\mu_S - \mu_N$
Pooled t	0.82	(-8.4, 6.4)
Welch t	0.8	(-7.6, 5.6)
Bootstrap	$p > 0.100$	(-7.6, 5.3)

Procedure	p for $H_0 : psmed_S = psmed_N$	90% CI for S - N shift
Rank Sum	0.27	(-11, 2)

Which method should we use?

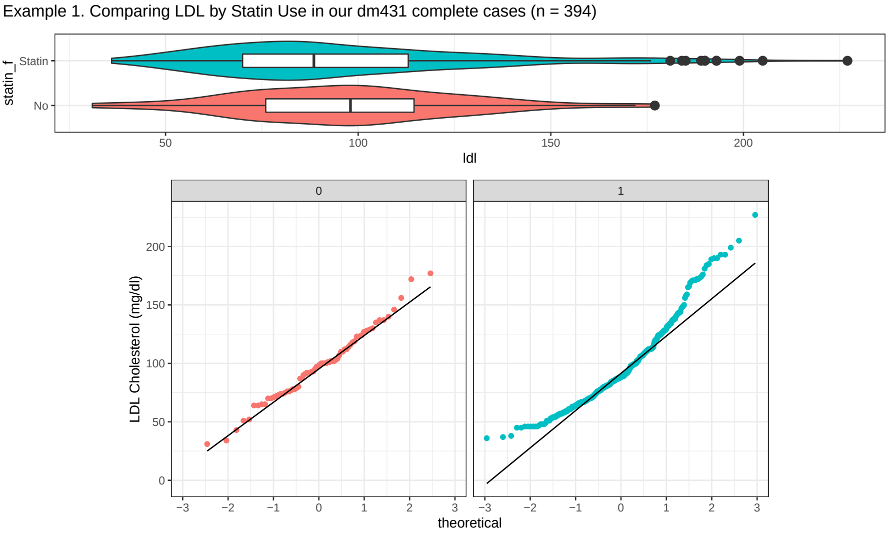
Which Method Should We Use?

1. Plot the distributions of the two independent samples.
2. Does it seem reasonable to assume that **each** distribution (here, both `ldl` in statin users and `ldl` in non-users) follows an approximately Normal distribution?

- If Yes, Normal models seem fairly appropriate, then
    - use the indicator variable regression (pooled `t` test) if the sample sizes are nearly the same, or if the sample variances are reasonably similar
    - use the Welch's `t` test, otherwise (default `t.test` in R)
- If No, Normal models don't seem appropriate at all, then
    - compare means using the bootstrap via `bootdif`, or
    - compare pseudo-medians using the WMW rank sum test

What did we see in our `ldl` data?

LDL, within groups defined by `statin`



Results for the LDL and Statin Study

Procedure	p for $H_0 : \mu_S = \mu_N$	90% CI for $\mu_S - \mu_N$
Pooled t	0.82	(-8.4, 6.4)
Welch t	0.8	(-7.6, 5.6)
Bootstrap	$p > 0.100$	(-7.6, 5.3)

Procedure	p for $H_0 : psmed_S = psmed_N$	90% CI for S - N shift
Rank Sum	0.27	(-11, 2)

What conclusions should we draw, at $\alpha = 0.10$?

Example 2 (Comparing SBP by Sex) slides follow, for you to review on your own. The main difference is that we have no missing values in SBP or Sex in the `dm431` data.

dm431 Example 2. (Comparing SBP by Sex)

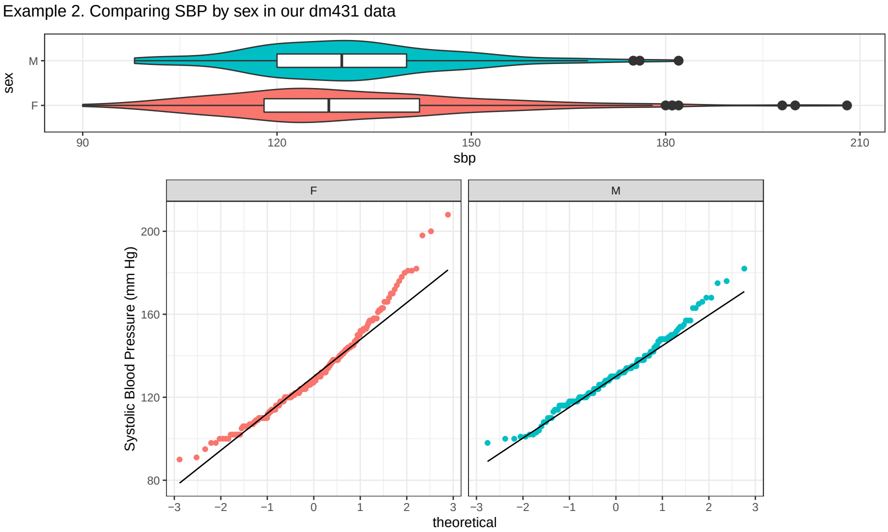
Estimate the difference in population mean systolic blood pressure among females as compared to males.

```
mosaic::favstats(sbp ~ sex, data = dm431) %>% kable(dig = 2)
```

sex	min	Q1	median	Q3	max	mean	sd	n	missing
F	90	118	128	142	208	131.17	20.15	257	0
M	98	120	130	140	182	131.41	15.87	174	0

- What is the outcome here?
- What are the two exposure groups we are comparing?
- What are the sample means, $\bar{x}_F$ and $\bar{x}_M$?
- Point estimate of the difference in population means, $\mu_F - \mu_M$?

DTDP for Example 2. (Comparing SBP by Sex)


Example 2. Comparing SBP by sex in our dm431 data

Linear Model for Example 2 (slide A)

Estimate the difference in population mean systolic blood pressure among females as compared to males.

```
m1 <- lm(sbp ~ sex, data = dm431)

tidy(m1, conf.int = T, conf.level = 0.90) %>% kable(dig = 2)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	131.17	1.16	113.41	0.00	129.26	133.07
sexM	0.24	1.82	0.13	0.89	-2.76	3.24

- What can we learn from this output?
  - What is the sample mean `sbp` for females?
  - What is the sample mean `sbp` for males?
  - The point estimate for $\mu_F - \mu_M$ is ...

Linear Model for Example 2 (slide B)

Estimate the difference in population mean systolic blood pressure among females as compared to males.

```r
m1 <- lm(sbp ~ sex, data = dm431)

tidy(m1, conf.int = T, conf.level = 0.90) %>% kable(dig = 2)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	131.17	1.16	113.41	0.00	129.26	133.07
sexM	0.24	1.82	0.13	0.89	-2.76	3.24

- What can we learn from this output?
    - The point estimate for $\mu_F - \mu_M$ is -0.24
    - The 90% confidence interval for $\mu_F - \mu_M$ is (-**3.24**, **2.76**)

Building a Pooled t CI: Example 2

1. Best approach: use indicator variable regression
2. Also: direct call to t test with pooled variance estimate

```r
t.test(sbp ~ sex, data = dm431, alt = "two.sided", mu = 0,
       var.equal = TRUE, conf.level = 0.90)
```

```

    Two Sample t-test

data:  sbp by sex
t = -0.13225, df = 429, p-value = 0.8949
alternative hypothesis: true difference in means is not equal
90 percent confidence interval:
 -3.241344  2.759883
sample estimates:
mean in group F mean in group M
       131.1673        131.4080
```

t test can be tidied

```
t1 <- tidy(t.test(sbp ~ sex, data = dm431,
                  var.equal = TRUE, conf.level = 0.90))
```

- conf.level must be specified to t.test. Otherwise, it uses 0.95.

Elements of t1 are printed below (after rearrangement)

method	alternative	estimate1	estimate2
Two Sample t-test	two.sided	131.17	131.41

estimate	conf.low	conf.high	statistic	parameter	p.value
-0.24	-3.24	2.76	-0.13	429	0.89

Building the Welch t CI: Example 2

- Sensible approach when assuming Normal populations is OK, but we don't want to assume the two populations have the same variance (as pooled t requires)

```
t.test(sbp ~ sex, data = dm431, alt = "two.sided", mu = 0,
       conf.level = 0.90)
```

```

    Welch Two Sample t-test

data:  sbp by sex
t = -0.13838, df = 419.27, p-value = 0.89
alternative hypothesis: true difference in means is not equal
90 percent confidence interval:
 -3.108582  2.627120
sample estimates:
mean in group F mean in group M
       131.1673        131.4080
```

Welch `t` test can also be tidied

```r
t2 <- tidy(t.test(sbp ~ sex, data = dm431,
                  conf.level = 0.90))
```

- We must specify `conf.level` in the `t.test` unless we want 0.95.

Elements of `t2` are printed below (after rearrangement)

method	alternative	estimate1	estimate2
Welch Two Sample t-test	two.sided	131.17	131.41

estimate	conf.low	conf.high	statistic	parameter	p.value
-0.24	-3.11	2.63	-0.14	419.27	0.89

Wilcoxon-Mann-Whitney Rank Sum: Example 2

```r
wilcox.test(sbp ~ sex, data = dm431,
            conf.int = TRUE, conf.level = 0.90)
```

```

    Wilcoxon rank sum test with continuity
    correction

data:  sbp by sex
W = 21329, p-value = 0.4167
alternative hypothesis: true location shift is not equal to 0
90 percent confidence interval:
 -4.000050  1.999972
sample estimates:
difference in location
              -1.99992
```

Rank Sum test can also be tidied

```
w3 <- tidy(wilcox.test(sbp ~ sex, data = dm431,
            conf.int = TRUE, conf.level = 0.90))
```

- Specify `conf.int` and `conf.level` in the `wilcox.test`.

Elements of `w3` are printed below (after rearrangement)

method	alternative	statistic
Wilcoxon rank sum test with continuity correction	two.sided	21328.5

estimate	conf.low	conf.high	p.value
-2	-4	2	0.42

Using bootdif to compare mean(SBP) by Sex

So, to compare systolic BP (our outcome) across the two levels of sex (our grouping factor) for the adult patients with diabetes in NE Ohio, run…

```r
set.seed(431431)
boot4 <- dm431 %$% bootdif(sbp, sex, conf.level = 0.90)
boot4
```

```
Mean Difference            0.05            0.95
      0.2407308      -2.6226195       3.1868901
```

- This CI estimates $\mu_M - \mu_F$: observe the listed sample mean difference for the necessary context.
- Invert the signs to estimate $\mu_F - \mu_M$.
- Again the p value must be larger than 0.10 since 0 is in the 90% CI.

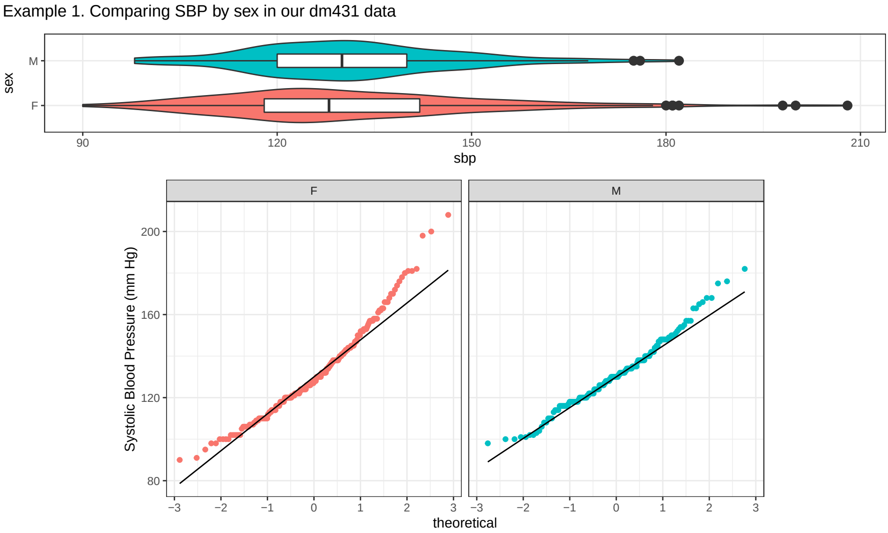
Results for the SBP and Sex Study

Procedure	p for $H_0 : \mu_F = \mu_M$	90% CI for $\mu_F - \mu_M$
Pooled t	0.8949	(-3.24, 2.76)
Welch t	0.89	(-3.11, 2.63)
Bootstrap	$p > 0.100$	(-3.19, 2.62)

Procedure	p for $H_0 : psmed_F = psmed_M$	90% CI for F - M shift
Rank Sum	0.4167	(-4, 2)

Which method should we use?

Systolic BP, within groups defined by sex



Results for the SBP and Sex Study

Procedure	p for $H_0 : \mu_F = \mu_M$	90% CI for $\mu_F - \mu_M$
Pooled t	0.8949	(-3.24, 2.76)
Welch t	0.89	(-3.11, 2.63)
Bootstrap	$p > 0.100$	(-3.19, 2.62)

Procedure	p for $H_0 : psmed_F = psmed_M$	90% CI for F - M shift
Rank Sum	0.4167	(-4, 2)

What conclusions should we draw, at $\alpha = 0.10$?