431 Class 15

github.com/THOMASELOVE/2019-431

2019-10-15

Today's Agenda (Notes Chapters 19-20)

1. Discussion of Quiz 1
2. Discussion of Draft Survey for Project Study A
3. Statistical Inference and the `dm431` data: Comparing Population Means using Independent Samples
   - Pooled t / Indicator Variable Regression
   - Welch's t
   - Wilcoxon-Mann-Whitney rank sum
   - Bootstrap with `bootdif`

Today's Setup and Data

```r
library(magrittr); library(janitor)
library(patchwork); library(here);
library(boot); library(broom)
library(tidyverse)

source(here("R", "Love-boost.R"))

dm431 <- readRDS(here("data", "dm431.Rds"))
```

Comparing means using Independent Samples

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 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 systolic blood pressure among females in our population as compared to males in our population?

2. Can we estimate the difference in the population mean LDL level for those on a statin as compared to those not on a statin?

3. Can we estimate the difference in the population mean hemoglobin A1c for those with Medicaid vs. Medicare insurance?

SBP for females vs. males

Systolic Blood Pressure by Sex in 431 Adults with Diabetes
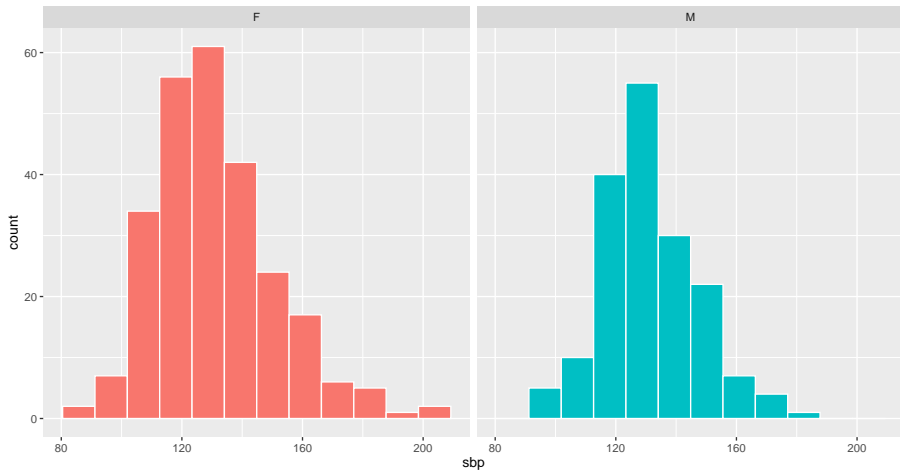
Numerical Summary for Two Independent Samples

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

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

  sex min  Q1 median  Q3 max     mean       sd   n
1   F  90 118    128 142 208 131.1673 20.14962 257
2   M  98 120    130 140 182 131.4080 15.86577 174
  missing
1       0
2       0
```

Confidence Interval Options for Independent Samples

1. Pooled t CI or Indicator Variable Regression Model (t approach 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)

Hypotheses Under Consideration

The hypotheses we are testing are:

- H_0: mean in population 1 = mean in population 2 + hypothesized difference Δ_0 vs.
- H_A: mean in population $1 \neq$ mean in population 2 + hypothesized difference Δ_0,

where Δ_0 is almost always zero. An equivalent way to write this is:

- $H_0 : \mu_1 = \mu_2 + \Delta_0$ vs.
- $H_A : \mu_1 \neq \mu_2 + \Delta_0$

Yet another equally valid way to write this is:

- $H_0 : \mu_1 - \mu_2 = \Delta_0$ vs.
- $H_A : \mu_1 - \mu_2 \neq \Delta_0$,

where, again, Δ_0 is almost always zero.

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] The population variances in the two groups being compared are the same, so we can obtain a pooled estimate of their joint variance.

The Pooled Variances t test in R

Also referred to as the t test assuming equal population variances:

```
tt <- t.test(sbp ~ sex, data = dm431, var.equal = TRUE)
tidy(tt) %>%
  select(estimate1, estimate2,
         conf.low, conf.high, method, alternative) %>%
  knitr::kable(digits = 2)
```

estimate1	estimate2	conf.low	conf.high	method	alternativ
131.17	131.41	-3.82	3.34	Two Sample t-test	two.sided

Full output from Pooled T test

```
    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
95 percent confidence interval:
 -3.818555  3.337094
sample estimates:
mean in group F mean in group M
       131.1673          131.4080
```

Indicator Variable Regression Approach

```
model1 <- lm(sbp ~ sex, data = dm431)
tidy(model1, conf.int = TRUE, conf.level = 0.95) %>%
  select(term, estimate, conf.low, conf.high) %>%
  knitr::kable(digits = 2)
```

term	estimate	conf.low	conf.high
(Intercept)	131.17	128.89	133.44
sexM	0.24	-3.34	3.82

`sexM` indicator shows the effect of being Male, so the displayed CI estimates $\mu_{male} - \mu_{female}$. Invert the signs to get the $\mu_{female} - \mu_{male}$ estimate.

summary(model1)

```
> summary(model1)

Call:
lm(formula = sbp ~ sex, data = dm431)

Residuals:
    Min      1Q  Median      3Q     Max
-41.167 -11.408  -1.408   9.833  76.833

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 131.1673     1.1566 113.409   <2e-16 ***
sexM          0.2407     1.8203   0.132    0.895
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.54 on 429 degrees of freedom
Multiple R-squared:  4.077e-05,  Adjusted R-squared:  -0.00229
F-statistic: 0.01749 on 1 and 429 DF,  p-value: 0.8949
```

Results for the SBP and Sex Study

Procedure	p for $H_0 : \mu_F = \mu_M$	95% CI for $\mu_M - \mu_F$
Pooled t	0.90	(-3.3, 3.8)

What conclusions should we draw, at $\alpha = 0.05$?

Assumptions of the Welch t test

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.

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

Welch t test not assuming equal population variances

```
t.test(sbp ~ sex, data = dm431)
```

```
    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
95 percent confidence interval:
 -3.660307  3.178845
sample estimates:
mean in group F mean in group M
       131.1673        131.4080
```

Results for the SBP and Sex Study

Procedure	p for $H_0 : \mu_F = \mu_M$	95% CI for $\mu_M - \mu_F$
Pooled t	0.90	(-3.3, 3.8)
Welch t	0.89	(-3.2, 3.7)

What conclusions should we draw, at $\alpha = 0.05$?

Assumptions of the Wilcoxon-Mann-Whitney Rank Sum Test

The Wilcoxon-Mann-Whitney Rank Sum 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.

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.

It also doesn't really compare population means. It compares pseudo-medians again.

Wilcoxon-Mann-Whitney Rank Sum Test

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

```
    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
95 percent confidence interval:
 -4.999979  2.000039
sample estimates:
difference in location
              -1.99992
```

Results for the SBP and Sex Study

Procedure	p for $H_0 : \mu_F = \mu_M$	95% CI for $\mu_M - \mu_F$
Pooled t	0.90	(-3.3, 3.8)
Welch t	0.89	(-3.2, 3.7)

Procedure	p for $H_0 : psmed_F = psmed_M$	95% CI for M - F shift
Rank Sum	0.42	(-2.0, 5.0)

What conclusions should we draw, at $\alpha = 0.05$?

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 SBP levels in the two independent samples (M vs. F).
- For each bootstrap sample, we are calculating a mean difference between the two groups (M vs. F).
- We then determine the 2.5th and 97.5th percentile of the resulting distribution of mean differences (for a 95% confidence interval).

Using the `bootdif` function to compare means based on independent samples

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...

```
set.seed(4312019)
dm431 %$% bootdif(sbp, sex, conf.level = 0.95)
```

```
Mean Difference           0.025           0.975
     0.2407308       -3.1197510       3.5696286
```

- The two columns must be separated here with a comma rather than a tilde (~), and are specified using $ notation.
- This CI estimates $\mu_{male} - \mu_{female}$: observe the listed sample mean difference for the necessary context. Invert the signs, as before, to estimate $\mu_{female} - \mu_{male}$.

Results for the SBP and Sex Study

Procedure	p for $H_0 : \mu_F = \mu_M$	95% CI for $\mu_F - \mu_M$
Pooled t	0.90	(-3.3, 3.8)
Welch t	0.89	(-3.2, 3.7)
Bootstrap	$p > 0.05$	(-3.1, 3.6)

Procedure	p for $H_0 : psmed_F = psmed_M$	95% CI for M - F shift
Rank Sum	0.42	(-2.0, 5.0)

What conclusions should we draw, at $\alpha = 0.05$?

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 sbp in males and sbp in females) follows an approximately Normal distribution?

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

What did we see in our systolic BP data?
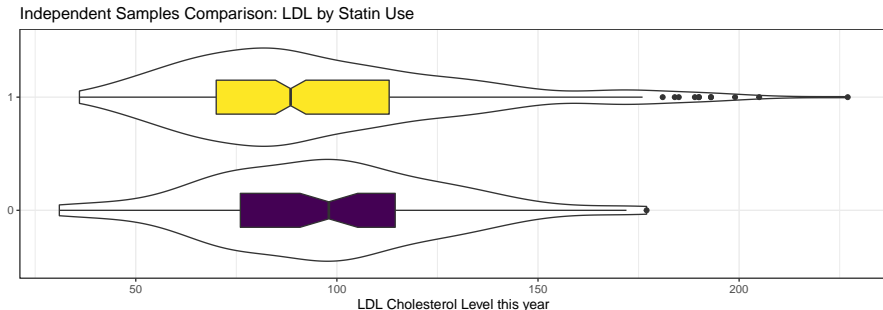
Systolic BP, within groups defined by sex

Independent Samples Comparison: SBP by Sex
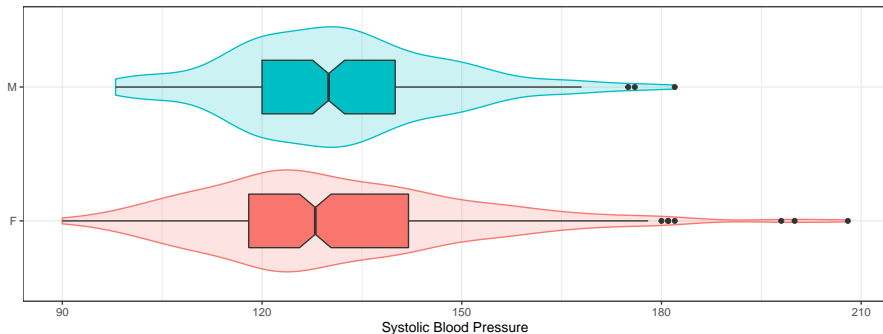
LDL of statin users and non-users

Warning: Removed 37 rows containing non-finite values
(stat_ydensity).

Warning: Removed 37 rows containing non-finite values
(stat_boxplot).

Independent Samples Comparison: LDL by Statin Use

What should we do about the missing values?

We could just look at the complete cases, and that's probably the best strategy when we're doing a two-sample t test. When we fit a more complicated regression (than just the simple indicator variable regression we need to get the pooled t test) where we're adjusting for multiple variables, then we'll consider imputation seriously.
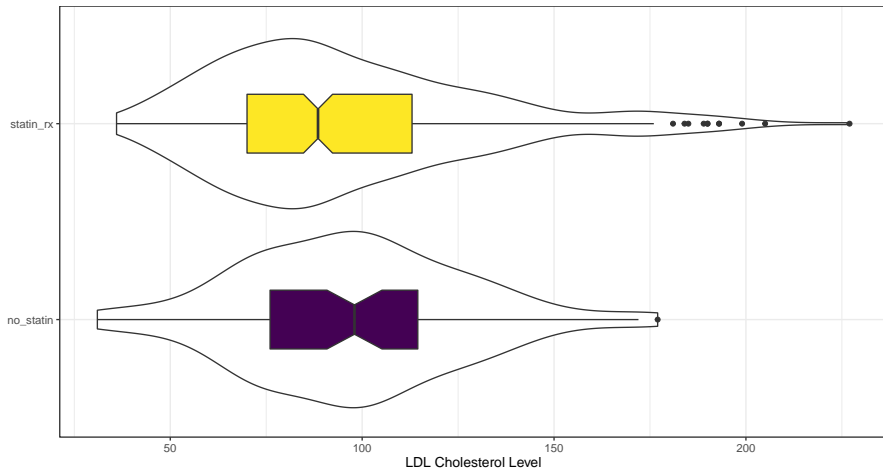
We also probably want to turn `statin` into a factor with meaningful names.

```
dm431_sub <- dm431 %>%
  filter(complete.cases(ldl, statin)) %>%
  mutate(statin_f =
           fct_recode(factor(statin),
                      statin_rx = "1", no_statin = "0"))
```

OK. Let's plot again.

LDL of statin users and non-users

Independent Samples Comparison: LDL by Statin Use

Summary Statistics for LDL by statin prescription

```
mosaic::favstats(ldl ~ statin_f, data = dm431_sub)

   statin_f min Q1 median    Q3 max     mean       sd
1 no_statin  31 76   98.0 114.5 177 97.41667 29.22364
2 statin_rx  36 70   88.5 113.0 227 96.40683 35.33276
    n missing
1 72       0
2 322       0
```

Pooled t-test from indicator variable regression

```
> model_sub <- lm(ldl ~ statin, data = dm431_sub)
> summary(model_sub)

Call:
lm(formula = ldl ~ statin, data = dm431_sub)

Residuals:
    Min      1Q  Median      3Q     Max
-66.417 -25.159  -6.412  16.593 130.593

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   97.417      4.043  24.094   <2e-16 ***
statin        -1.010      4.472  -0.226    0.821
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.31 on 392 degrees of freedom
Multiple R-squared:  0.00013,   Adjusted R-squared:  -0.002421
F-statistic: 0.05098 on 1 and 392 DF,  p-value: 0.8215
```

Conclusions at $\alpha = 0.05$?

Results for the LDL and Statin Study

Procedure	p for $H_0 : \mu_{statin} = \mu_{no}$	95% CI for $\mu_{statin} - \mu_{no}$
Pooled t	0.82	(-9.8, 7.8)
Welch t	0.80	(-8.9, 6.8)
Bootstrap	$p > 0.05$	(-8.7, 7.0)

Procedure	p for $H_0 : psmed_{st} = psmed_{no}$	95% CI for Statin - No shift
Rank Sum	0.27	(-12, 4)

What conclusions should we draw, at $\alpha = 0.05$?

Let's compare Hemoglobin A1c for Medicare vs. Medicaid

```
dm431_sub2 <- dm431 %>%
  filter(insurance %in% c("Medicaid", "Medicare")) %>%
  filter(complete.cases(a1c)) %>%
  droplevels()

mosaic::favstats(a1c ~ insurance, data = dm431_sub2)

  insurance min  Q1 median  Q3  max     mean       sd
1  Medicaid 4.8 6.5   7.25 9.0 14.7 8.074000 2.302946
2  Medicare 4.3 6.5   7.15 7.9 16.7 7.635246 2.041225
    n missing
1 100       0
2 122       0
```

Hemoglobin A1c for Medicare vs. Medicaid

Independent Samples Comparison: A1c by Insurance
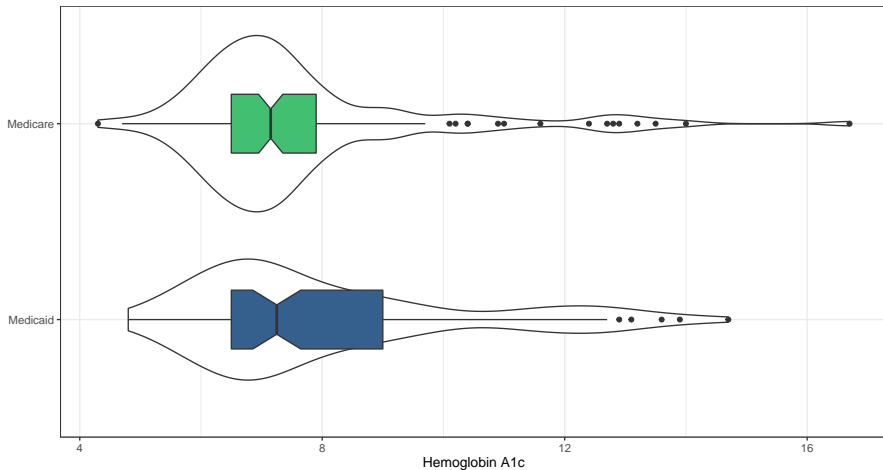
Four Approaches for comparing A1c by Insurance

Procedure	p for $H_0 : \mu_{Medicaid} = \mu_{Medicare}$	95% CI for $\mu_{Medicaid} - \mu_{Medicare}$
Pooled t	0.13	(-0.14, 1.01)
Welch t	0.14	(-0.14, 1.02)
Bootstrap CI	$p > 0.05$	(-0.17, 1.02)

Procedure	p for $H_0 : psmed_{Medicaid} = psmed_{Medicare}$	95% CI for Medicaid - Medicare shift
Rank Sum	0.24	(-0.2, 0.6)

What conclusions should we draw, at $\alpha = 0.05$?

A Few Reminders About Significance

- **A significant effect is not necessarily the same thing as an interesting effect.** For example, results calculated from large samples are nearly always "significant" even when the effects are quite small in magnitude. Before doing a test, always ask if the effect is large enough to be of any practical interest. If not, why do the test?

- **A non-significant effect is not necessarily the same thing as no difference.** A large effect of real practical interest may still produce a non-significant result simply because the sample is too small.

- **There are assumptions behind all statistical inferences.** Checking assumptions is crucial to validating the inference made by any test or confidence interval.

Compare the proportion of current Medicare subjects whose insurance status changed to Medicare in the last two years, to the proportion of current Medicare subjects who had another insurance status two years ago?

```
dm431 %>% filter(insurance == "Medicare") %>%
  count(insurance, insurance_old)
```

```
# A tibble: 4 x 3
  insurance insurance_old     n
  <fct>     <fct>         <int>
1 Medicare  Commercial       22
2 Medicare  Medicaid          4
3 Medicare  Medicare         93
4 Medicare  Uninsured         4
```

- Outcome? Exposure Groups? Paired or Independent Samples?

dm431 **Example 5.**

Among current Medicare subjects, compare the proportion with A1c below 8 to the proportion for the same patients two years ago.

```
dm431 %>% filter(insurance == "Medicare") %>%
  count(a1c < 8, a1c_old < 8)
```

```
# A tibble: 7 x 3
  `a1c < 8` `a1c_old < 8`     n
  <lgl>     <lgl>         <int>
1 FALSE     FALSE            16
2 FALSE     TRUE             12
3 FALSE     NA                1
4 TRUE      FALSE            16
5 TRUE      TRUE             74
6 TRUE      NA                3
7 NA        TRUE              1
```

- Outcome? Exposure Groups? Paired or Independent Samples?