431 Class 14

github.com/THOMASELOVE/2019-431

2019-10-10

Today's Agenda (Notes Chapters 19-20)

Statistical Inference and the `dm431` data

1. Paired vs. Independent Samples
2. Moving from Wide to Long and back again
3. 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"))
```

Section 1

Comparing Population Means

So far, we've been thinking about one population, and one sample

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).

Are these Samples Paired (Matched) or Not?

Now, suppose we want to compare two subpopulations of our bigger population, using the relevant subsamples of our `dm431` data.

- Deciding whether or not the samples are paired (matched) is something we do before we analyze the data.

The best way to establish whether a study uses paired or independent samples is to look for the **link** between the two measurements that creates paired differences.

dm431 Example 1.

Suppose we want to compare the mean ldl cholesterol level for subjects who are currently taking a statin medication to the mean ldl for subjects who are not currently taking a statin.

```
dm431 %>% select(subject, ldl, statin) %>% tail()
```

```
# A tibble: 6 x 3
  subject   ldl statin
  <chr>   <dbl>  <dbl>
1 S-426     100      1
2 S-427      86      1
3 S-428      88      1
4 S-429     166      1
5 S-430      34      0
6 S-431      77      0
```

`dm431` Example 1.

Suppose we want to compare the mean `ldl` cholesterol level for subjects who are currently taking a `statin` medication to the mean `ldl` for subjects who are not currently taking a statin.

```
mosaic::favstats(ldl ~ statin, data = dm431)
```

```
  statin min Q1 median    Q3 max     mean       sd   n
1      0  31 76   98.0 114.5 177 97.41667 29.22364  72
2      1  36 70   88.5 113.0 227 96.40683 35.33276 322
  missing
1      14
2      23
```

- What is the outcome of interest here?
- What are the two exposure groups we are comparing?
- Does this design create paired samples or independent samples?

dm431 Example 2.

Suppose we want to compare the mean ldl cholesterol level for a set of subjects this year to the mean ldl for the same subjects two years ago.

```
dm431 %>% select(subject, ldl, ldl_old) %>% head()
```

```
# A tibble: 6 x 3
  subject   ldl ldl_old
  <chr>   <dbl>   <dbl>
1 S-001     126      71
2 S-002     172     182
3 S-003     105     127
4 S-004     127      NA
5 S-005     100      86
6 S-006      65      90
```

dm431 Example 2.

Suppose we want to compare the mean ldl cholesterol level for a set of subjects this year to the mean ldl for the same subjects two years ago.

```
mosaic::favstats(~ ldl, data = dm431)
```

```
 min Q1 median  Q3 max     mean       sd   n missing
  31 72     90 113 227 96.59137 34.26558 394      37
```

```
mosaic::favstats(~ ldl_old, data = dm431)
```

```
 min Q1 median  Q3 max     mean       sd   n missing
  31 72     90 115 244 96.98744 34.73313 398      33
```

- What is the outcome of interest here?
- What are the two exposure groups we are comparing?
- Does this design create paired samples or independent samples?

dm431 Example 3.

Suppose we want to compare the mean systolic blood pressure for male subjects to the mean systolic blood pressure among female subjects?

```
dm431 %>% select(subject, sbp, sex) %>% head()
```

```
# A tibble: 6 x 3
  subject   sbp sex
  <chr>   <dbl> <fct>
1 S-001      98 F
2 S-002     162 F
3 S-003     154 F
4 S-004     138 M
5 S-005     118 F
6 S-006     124 F
```

dm431 Example 3.

Suppose we want to compare the mean systolic blood pressure for male subjects to the mean systolic blood pressure among female subjects?

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

```
  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
```

- What is the outcome of interest here?
- What are the two exposure groups we are comparing?
- Does this design create paired samples or independent samples?

Formatting the Data (Wide vs. Long)

Wide format (most appropriate for paired/matched samples)

subject	treatment1	treatment2
A	140	150
B	135	145
C	128	119

Long format (most appropriate for independent samples)

subject	sbp	group
A	140	treatment1
A	150	treatment2
B	135	treatment1
B	145	treatment2
C	128	treatment1
C	119	treatment2

Suppose you have a wide data set. . .

```r
tempdat_wide <- tibble(
  subject = c("A", "B", "C"),
  treatment_1 = c(140, 135, 128),
  treatment_2 = c(150, 145, 119)
)

tempdat_wide
```

```
# A tibble: 3 x 3
  subject treatment_1 treatment_2
  <chr>         <dbl>       <dbl>
1 A               140         150
2 B               135         145
3 C               128         119
```

Pivot Data to make it longer

We want more rows, fewer columns. Each subject*treatment combination will become a row.

```
tempdat_long <- tempdat_wide %>%
  pivot_longer( -subject,
    names_to = "group", values_to = "sbp")
tempdat_long
```

```
# A tibble: 6 x 3
  subject group           sbp
  <chr>   <chr>         <dbl>
1 A       treatment_1     140
2 A       treatment_2     150
3 B       treatment_1     135
4 B       treatment_2     145
5 C       treatment_1     128
6 C       treatment_2     119
```

Spread the Data from Long to Wide

```r
tempdat_wide2 <- tempdat_long %>%
  pivot_wider(names_from = group, values_from = sbp)

tempdat_wide2
```

```
# A tibble: 3 x 3
  subject treatment_1 treatment_2
  <chr>         <dbl>       <dbl>
1 A               140         150
2 B               135         145
3 C               128         119
```

So, an independent samples design?

- Independent samples designs do not impose a matching, but instead sample two unrelated sets of subjects, where each group receives one of the two exposures.
- The two groups of subjects are drawn independently from their separate populations of interest.
- One obvious way to tell if we have an independent samples design is that this design does not require the sizes of the two exposure groups to be equal.

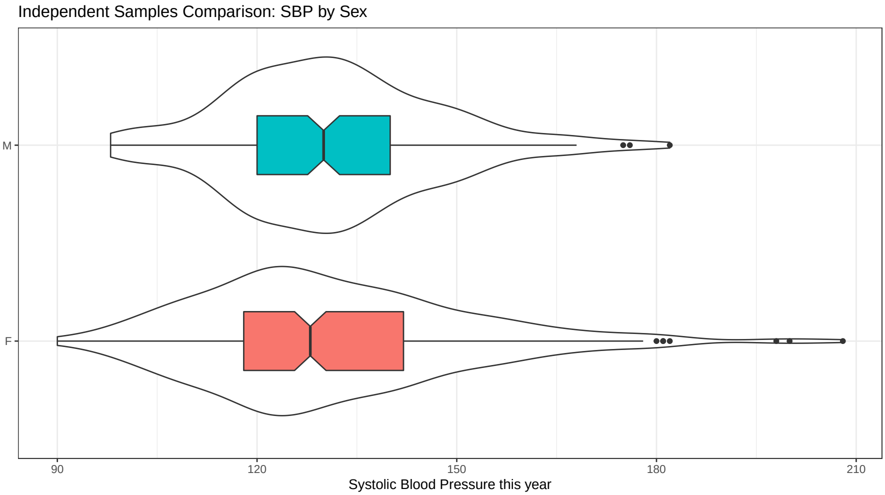
Three Small Analyses using Independent Samples

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.

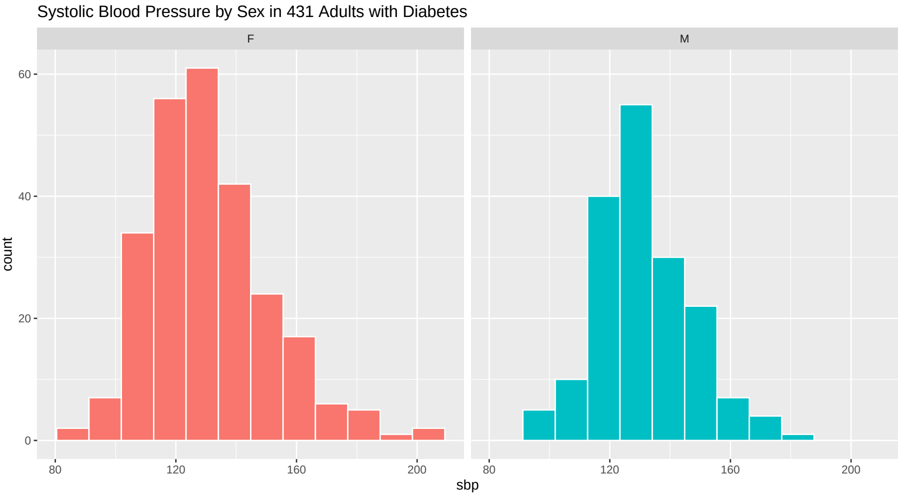
Our sample: 431 of those people, drawn in a way we hope is representative (but certainly isn't random).

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, in a boxplot



Another Picture for Two Independent Samples



Numerical Summary for Two Independent Samples

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

```
  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
```

Independent Samples: Confidence Intervals for $\mu_1 - \mu_2$

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)

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 $\alpha = 0.05$?