431 Class 15

thomaselove.github.io/431

2021-10-11

Today's R Setup

```
knitr::opts_chunk$set(comment=NA) # as always
options(width = 55) # to fit things on the slides

library(broom)
library(Hmisc) # for smean.cl.boot(), mostly
library(janitor)
library(knitr)
library(magrittr)
library(naniar)
library(pwr) # for specialized power functions
library(tidyverse)

source("data/Love-boost.R") # for bootdif() function

theme_set(theme_bw())
```

Today's Agenda

- Comparing Two Population Means Using Paired Samples
  - T test and Bootstrap Approaches
- Power and Sample Size When Comparing Means

The Hypoxia MAP Data

These data also come from the Cleveland Clinic's Statistical Education Dataset Repository.

Source: Turan et al. "Relationship between Chronic Intermittent Hypoxia and Intraoperative Mean Arterial Pressure in Obstructive Sleep Apnea Patients Having Laparoscopic Bariatric Surgery" *Anesthesiology* 2015; 122: 64-71.

```
hypox_raw <- read_csv("data/HypoxiaMAP.csv") %>%
  clean_names() %>%
  mutate(subject = row_number())

dim(hypox_raw)
```

```
[1] 281  37
```

Background and Study Description

[The Hypoxia MAP study] retrospectively examined the intraoperative blood pressures in 281 patients who had laparoscopic bariatric surgery between June 2005 and December 2009 and had a diagnosis of OSA within two preoperative years.

Time-weighted average (TWA) intraoperative MAP was the main outcome in the study. MAP (or mean arterial pressure) is a term used to describe an average blood pressure in a subject.

MAP is normally between 65 and 110 mmHg, and it is believed that a MAP $>$ 70 mmHg is enough to sustain the organs of the average person. If the MAP falls below this number for an appreciable time, vital organs will not get enough oxygen perfusion, and will become hypoxic, a condition called ischemia.

Our Objective with these Data

We will focus today on two measurements of MAP for each subject (outside of some missing data).

- MAP1 = time-weighted average mean arterial pressure from ET intubation to trocar insertion, in mm Hg.
- MAP2 = time-weighted average mean arterial pressure from trocar insertion to the end of the surgery, in mm Hg.

We are interested in estimating the **difference** between the two MAP levels, across a population of subjects like those enrolled in this study.

Our Key Variables

- For each subject, we have two outcomes to compare: their MAP1 and their MAP2.

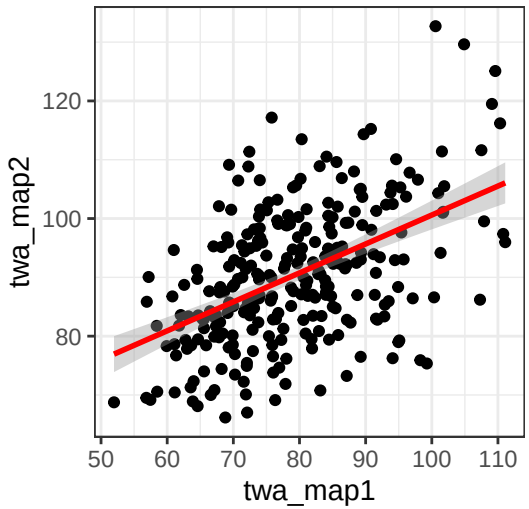
```
hypox <- hypox_raw %>%
  select(subject, twa_map1, twa_map2) %>%
  mutate(map_diff = twa_map2 - twa_map1)

hypox %>% head(., 4)
```

```
# A tibble: 4 x 4
  subject twa_map1 twa_map2 map_diff
    <int>    <dbl>    <dbl>    <dbl>
1       1     67.9     87.4     19.5
2       2     67.0     83.3     16.3
3       3     91.6     83.0     -8.59
4       4     67.1     79.9     12.8
```

We have Paired Samples in this setting

- Every MAP1 value is connected to the MAP2 value for the same subject. We say that the MAP1 and MAP2 are paired by subject.



Each subject provides a MAP1 and a MAP2

- The pairing is fairly strong here. The Pearson correlation of MAP1 and MAP2 across the subjects with complete data is 0.494.
- It makes sense to calculate the (paired) difference in MAP values for each subject, so long as there aren't any missing data.

Are there any missing values?

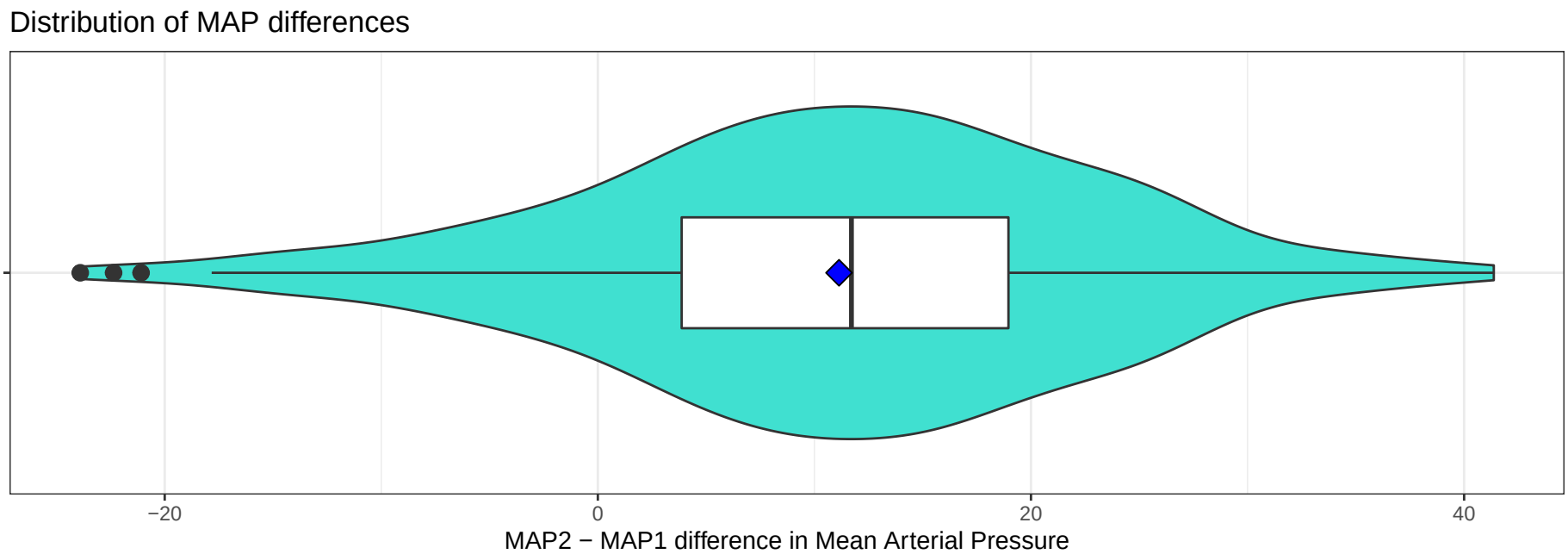
```
miss_var_summary(hypox)
```

```
# A tibble: 4 x 3
  variable n_miss pct_miss
  <chr>     <int>    <dbl>
1 twa_map1      4     1.42
2 map_diff      4     1.42
3 subject       0     0
4 twa_map2      0     0
```

```
hypox <- hypox %>% filter(complete.cases(map_diff))
```

Boxplot of the MAP differences

```
ggplot(data = hypox, aes(x = map_diff, y = "")) +
  geom_violin(fill = "turquoise") +
  geom_boxplot(width = 0.3, outlier.size = 3) +
  stat_summary(fun = "mean", geom = "point",
               shape = 23, size = 4, fill = "blue") +
  labs(x = "MAP2 - MAP1 difference in Mean Arterial Pressure",
       y = "", title = "Distribution of MAP differences")
```



Numerical Summaries

```
res1 <- as_tibble(bind_rows(
  mosaic::favstats(~ twa_map1, data = hypox),
  mosaic::favstats(~ twa_map2, data = hypox),
  mosaic::favstats(~ map_diff, data = hypox))) %>%
  mutate(item = c("map1", "map2", "map_diff")) %>%
  select(item, n, mean, sd, min, median, max)

res1 %>% kable()
```

item	n	mean	sd	min	median	max
map1	277	79.24274	11.73903	51.96	78.02	111.10
map2	277	90.37921	11.69104	66.17	89.74	132.71
map_diff	277	11.13646	11.78064	-23.90	11.71	41.37

- Is the mean of `map_diff` equal to the difference between the mean of `map2` and the mean of `map1`? Other summaries?

Hypothesis Testing Comparing Paired Samples

- Null hypothesis H_0
  - H_0: population mean of paired differences (MAP2 - MAP1) $= 0$
- Alternative (research) hypothesis H_A or H_1
  - H_1: population mean of paired differences (MAP2 - MAP1) $\neq 0$

Two (related) next steps

1. Given the data, we can then calculate the paired differences, then an appropriate test statistic based on those differences, which we compare to an appropriate probability distribution to obtain a p value. Again, small p values favor H_1 over H_0.
2. More usefully, we can calculate the paired differences, and then use an appropriate probability distribution to help use the data to construct an appropriate **confidence interval** for the population of those differences.

Paired T test via Linear Model

```
m3 <- lm(map_diff ~ 1, data = hypox)

summary(m3)$coef
```

```
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 11.13646  0.7078298 15.73325 2.971357e-40
```

```
confint(m3, conf.level = 0.90)
```

```
               2.5 %   97.5 %
(Intercept) 9.743031 12.52989
```

```
summary(m3)$r.squared
```

```
[1] 0
```

Tidied Regression Model

```
tidy(m3, conf.int = TRUE, conf.level = 0.90) %>%
  select(term, estimate, conf.low, conf.high) %>%
  kable(digits = 3)
```

term	estimate	conf.low	conf.high
(Intercept)	11.136	9.968	12.305

```
tidy(m3, conf.int = TRUE, conf.level = 0.90) %>%
  select(term, estimate, std.error, statistic, p.value) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	11.136	0.708	15.733	0

Paired T test via t.test

```
hypox %$% t.test(map_diff, conf.level = 0.90)
```

```
    One Sample t-test

data:  map_diff
t = 15.733, df = 276, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
  9.968265 12.304660
sample estimates:
mean of x
 11.13646
```

Paired T Confidence Interval yet another way

```
hypox %$%
  smean.cl.normal(map_diff, conf = 0.90)
```

```
     Mean     Lower     Upper
11.136462  9.968265 12.304660
```

The function `smean.cl.normal` (and that's an L, not a 1 after C) comes from the `Hmisc` package.

So does the `smean.cl.boot` function we'll see on the next slide, which will let us avoid the key assumption of Normality for the population of paired differences.

Bootstrap for Comparing Paired Means

```
set.seed(20211006)
hypox %$%
  Hmisc::smean.cl.boot(map_diff, conf = 0.90, B = 1000)
```

```
     Mean      Lower      Upper
11.136462  9.932226 12.323684
```

```
set.seed(123431)
hypox %$%
  Hmisc::smean.cl.boot(map_diff, conf = 0.90, B = 5000)
```

```
    Mean     Lower     Upper
11.13646 10.00769 12.30044
```

Gathered Estimates from our Paired Samples

Method	Estimate and 90% CI	Assumes Normality?
Paired t	11.14 (9.97, 12.30)	Yes
Bootstrap	11.14 (9.93, 12.32)	No

We estimate that the time-weighted average mean arterial pressure is 11.14 mm Hg higher (90% CIs shown above) after trocar insertion than it is during the period from ET intubation to trocar insertion, based on our sample of 277 subjects with complete data in this study.

- Does it matter much whether we assume Normality here?
- What can we say about the p values here?
- Is this a random sample of subjects?
- What population do these data represent?

Key Things to Remember in Hypothesis Testing

- Findings can be both statistically significant and practically significant or either or neither.
  - A statistically significant effect is not the same thing as a scientifically interesting effect.
  - A non-significant effect is not the same thing as "no difference".
- When we have large samples, we will regularly find small differences that have a small p value even though they have no practical importance.
- At the other extreme, with small samples, even large differences will often not be large enough to create a small p value.

Errors in Hypothesis Testing

In testing hypotheses, there are two potential decisions and each one brings with it the possibility that a mistake has been made.

–	H_A is true	H_0 is true
Test rejects H_0	Correct	Type I error (False positive)
Test retains H_0	Type II error (False negative)	Correct

- A Type I error can only be made if H_0 is actually true.
- A Type II error can only be made if H_A is actually true.

Specifying Error Probabilities (Type I)

If we say we are using 90% confidence, this means:

- we have a 10% significance level
- α, the probability of Type I error, is set to 0.10
- In general, confidence level = 100(1-α).
- The probability of correctly retaining H_0 is designed to be 0.90.

Specifying Error Probabilities (Type II)

- A Type II error is made if the alternative hypothesis is true, but you fail to choose it.
  - The probability depends on exactly which part of the alternative hypothesis is true, so that computing the probability of making a Type II error is not feasible.
- The **power** of a test is the probability of making the correct decision when the alternative hypothesis is true.
- β is defined as the probability of concluding that there was no difference, when in fact there was one (a Type II error).
- A more powerful test will have a lower Type II error rate, β.

Trading off significance (α) and power (β).

In many sample size decisions,

- we find that people set α, the tolerable rate of Type I error, to be 0.05.
- they then often try to set the sample size and other parameters so that the power (1 - β) is at least 0.80.

We'll advocate for thinking hard about the relative costs of Type I and Type II errors.

- The underlying framework that assumes a power of 80% with a significance level of 5% is sufficient for most studies is pretty silly.

Power and Sample Size Calculations

A power calculation is likely the most common element of an scientific grant proposal on which a statistician is consulted.

- The tests that have power calculations worked out in intensive detail using R are mostly those with more substantial assumptions.
  - t tests that assume population normality, common population variance and balanced designs in the independent samples setting
  - paired t tests that assume population normality
- These power calculations are also usually based on tests rather than confidence intervals. Simulation is your friend here.
- This process of doing power and related calculations is far more of an art than a science.

Paired vs. Independent Samples

If you can afford to obtain n = 400 observations to compare means under exposure A to means under exposure B, and you could either:

1. select a random sample from the population of interest containing 400 people and then randomly assign 200 people to receive exposure A and the remaining 200 people to receive exposure B (thus doing an independent samples study), or
2. select a random sample from the population of interest containing 200 people and then randomly assign 100 of them to get exposure A first, and then, a little later, when the effects have worn off, to then receive exposure B, while the other 100 people are assigned to receive B first, then A (thus doing a paired samples study)

Assuming the effect size is unchanged, which seems as though it would be the more powerful study design?

Power of an Independent Samples t test

```
power.t.test(n = 200, delta = 0.25, sd = 1,
             sig.level = 0.10)
```

```
     Two-sample t test power calculation

              n = 200
          delta = 0.25
             sd = 1
      sig.level = 0.1
          power = 0.8025858
    alternative = two.sided

NOTE: n is number in *each* group
```

Power of an Paired Samples t test

```
power.t.test(n = 200, delta = 0.25, sd = 1,
             sig.level = 0.10, type = "paired")
```

```
     Paired t test power calculation

              n = 200
          delta = 0.25
             sd = 1
      sig.level = 0.1
          power = 0.9698521
    alternative = two.sided

NOTE: n is number of *pairs*, sd is std.dev. of *differences*
```

What sample size do we need?

How many pairs of observations would we need to maintain 80% power?

```
power.t.test(delta = 0.25, sd = 1, sig.level = 0.10,
             power = 0.80, type = "paired")
```

```
     Paired t test power calculation

              n = 100.2877
          delta = 0.25
             sd = 1
      sig.level = 0.1
          power = 0.8
    alternative = two.sided

NOTE: n is number of *pairs*, sd is std.dev. of *differences*
```

Note that we'd need 101 pairs of measurements.

What happens when you change assumptions?

In our independent-samples test, we chose

- `n = 200` (per group)
- `delta = 0.25` (the minimum clinically important difference in means that we want to detect)
- `sd = 1` (assumed population standard deviation in each group)
- `sig.level = 0.10` (since we want 90% confidence)

Changing any of these will change the power from the calculated 0.802 to something else.

Which direction will power move in?

Original Setup yielded power = 0.802

- If we change n from 200 to 400, leaving everything else untouched, do you think the power will increase or decrease?

Which direction will power move in?

Original Setup yielded power = 0.802

- If we change n from 200 to 400, leaving everything else untouched, do you think the power will increase or decrease?
- If we change n from 200 to 400, power = 0.970

Which direction will power move in?

Original Setup yielded power = 0.802

- If we change n from 200 to 400, leaving everything else untouched, do you think the power will increase or decrease?
- If we change n from 200 to 400, power = 0.970
- What if we change n from 200 to 100?

Which direction will power move in?

Original Setup yielded power = 0.802

- If we change n from 200 to 400, leaving everything else untouched, do you think the power will increase or decrease?
- If we change n from 200 to 400, power = 0.970
- What if we change n from 200 to 100?
- If we change n from 200 to 100, power = 0.546

Changing the other parameters

Original Setup yielded power = 0.802

What do you think will happen?

New Setup	Resulting Power
a. δ from 0.25 to 0.5	Higher or Lower than 0.802?
b. δ from 0.25 to 0.1	?
c. `sd` from 1 to 2	?
d. `sd` from 1 to 0.5	?
e. α from 0.1 to 0.05	?
f. α from 0.1 to 0.2	?

- Which of these six setups will lead to **larger** power estimates than the original 0.802?

Results of Parameter Changes

Original Setup yielded power = 0.802

Change	Resulting power
δ from 0.25 to 0.5	0.9996
δ from 0.25 to 0.1	0.259
`sd` from 1 to 2	0.345
`sd` from 1 to 0.5	0.9996
α from 0.1 to 0.05	0.703
α from 0.1 to 0.2	0.888

- Setups **a** (larger δ), **d** (smaller `sd`) and **f** (smaller α) led to larger power estimates than the original setup.

What if you have an unbalanced design?

The most efficient design for an independent samples comparison will be balanced.

- What if we used our original setup for δ, sd and α, (which with n = 200 in each group, yielded power = 0.802) but instead we placed
  - 150 subjects into one exposure group, and
  - planned to recruit some number X larger than 150 into the other.
- How many people would we have to recruit into the second exposure group to yield the same power as our original 200 in each group result?

Using `pwr.t2n.test` from the `pwr` package

- Note the use here of d $= \delta/$sd.

```
pwr.t2n.test(n1 = 150, d = 0.25/1, sig.level = 0.10,
             power = 0.802)
```

```
     t test power calculation

             n1 = 150
             n2 = 298.1132
              d = 0.25
      sig.level = 0.1
          power = 0.802
    alternative = two.sided
```

So we can either have 200 and 200, or we can have 150 and 299 to maintain the same power.

Assessing Unbalanced Designs

The power is always stronger for a balanced design than for an unbalanced design with the same overall sample size.

See chapter 19 of the Course Notes for additional examples using the `pwr.t2n.test()` function within the `pwr` package.

One-Sided or Two-Sided

Note that I used a two-sided test to establish my power calculation - in general, this is the most conservative and defensible approach for any such calculation, unless there is a strong and specific reason to use a one-sided approach in building a power calculation, don't.

Coming Up

Class 16:

- Confidence Intervals for a Population Proportion
- Comparing Two Proportions with `twobytwo`
- Power Calculations for Studying Proportions
- Chi-Square Tests for Independence in Larger Contingency Tables

Classes 17-18:

- Analysis of Variance for Comparing $>$ 2 Population Means
- Methods of Dealing with Multiple Comparisons
- More on Statistical Significance and the trouble with p values
- Making Sure You Can Complete Project A's Analyses

and then we return to Multiple Regression