431 Class 17

Thomas E. Love

2017-10-26

Today's Agenda

- Power and Sample Size Considerations Comparing 2 Means
  - With power.t.test for balanced designs
  - With pwr for unbalanced designs
- A little in-class survey
- Comparing Two Population Means: Decision Support
- Project Task B - Building the Survey
- The Signal and The Noise
  - Chapter 7: Disease Outbreaks
  - Chapter 8: Bayes' Theorem

Today's R Setup

```
# devtools::install_github('jtleek/slipper')
# use above line if you haven't installed slipper

library(pwr); library(tidyverse)

angina <- read.csv("data/angina.csv") %>% tbl_df
implant <- read.csv("data/implant.csv") %>% tbl_df

source("Love-boost.R")
```

Error Types, Confidence, Power, α and β

- α is the probability of rejecting H_0 when H_0 is true.
  - So $1 - \alpha$, the confidence level, is the probability of retaining H_0 when that's the right thing to do.
- β is the probability of retaining H_0 when H_A is true.
  - So $1 - \beta$, the power, is the probability of rejecting H_0 when that's the right thing to do.

–	H_A is True	H_0 is True
Test Rejects H_0	Correct Decision (1 - β)	Type I Error (α)
Test Retains H_0	Type II Error (β)	Correct Decision (1 - α)

Most common approach: pre-specify $\alpha = 0.05$, and $\beta = 0.80$

Using `power.t.test`

Measure	Paired Samples	Independent Samples
`type =`	`"paired"`	`"two.sample"`
n	# of paired diffs	# in each sample
δ	true mean of diffs	true diff in means
s = sd	true SD of diffs	true SD, either group[1]
α = sig.level	max. Type I error rate	Same as paired.
$1 - \beta$ = power	power to detect effect δ	Same as paired.

Specify `alt = "greater"` or `alt = "less"` for a 1-sided comparison.

Sample Size & Power: Pooled t Test

For an independent-samples t test, with a balanced design (so that $n_1 = n_2$), R can estimate any one of the following elements, given the other four, using the `power.t.test` function, for a one-sided or two-sided t test.

- n = the sample size in each of the two groups being compared
- δ = delta = the true difference in means between the two groups
- s = sd = the true standard deviation of the individual values in each group (assumed to be constant, since we assume equal population variances)
- α = sig.level = the significance level for the comparison (maximum acceptable risk of Type I error)
- 1 - β = power = the power of the t test to detect the effect of size δ

If you want a two-sample power calculation for an unbalanced design, you will need to use a different library and function in R.

A Small Example: Studying Satiety

- I want to compare people eating this meal to people eating this meal in terms of impact on satiety.
- My satiety measure ranges from 0-100.
- People either eat meal A or meal B.
- I can afford to enroll 160 people in the study.
- I expect that a difference that's important will be about 10 points on the satiety scale.
- I don't know the standard deviation, but the whole range (0-100) gets used.
- I want to do a two-sided test.
- How many should eat meal A and how many meal B to maximize my power to detect such a difference? And how much power will I have if I use a 90% confidence level?

Satiety Example: Power

- n = the sample size in each of the two groups being compared
- δ = delta = the true difference in means between the two groups
- s = sd = the true standard deviation of the individual values in each group (assumed to be constant, since we assume equal population variances)
- α = sig.level = the significance level for the comparison (maximum acceptable risk of Type I error)
- 1 - β = power = the power of the t test to detect the effect of size δ

What do I know?

Satiety Example Calculation

```
power.t.test(n = 80, delta = 10, sd = 25,
             sig.level = 0.10, alt = "two.sided",
             type = "two.sample")
```

```
     Two-sample t test power calculation

              n = 80
          delta = 10
             sd = 25
      sig.level = 0.1
          power = 0.8089716
    alternative = two.sided

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

What if 32 people ate both meals (different times?)

Impact on standard deviation? Let's say $\sigma_d = 15$...

```
power.t.test(delta = 10, sd = 15, sig.level = 0.10,
        n = 32, alt = "two.sided", type = "paired")
```

```
     Paired t test power calculation

              n = 32
          delta = 10
             sd = 15
      sig.level = 0.1
          power = 0.979437
    alternative = two.sided

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

Power for an unbalanced design

- If you have independent samples, the most powerful design, for a given total sample size, will always be a balanced design.
- If you must use an unbalanced design in setting up a sample size calculation, you typically have meaningful information about the cost of gathering samples in each group, and this may help you estimate the impact of Type I and Type II errors so you can trade them off appropriately.

The tool I use (and demonstrate in the Notes, Part B, section on Power for Independent Sample T tests with Unbalanced Designs) is from the `pwr` library and is called `pwr.t2n.test`.

- Must specify both n1 and n2
- Instead of specifying *delta* and sd separately, we specify their ratio, with *d*.

Satiety Example Again

If we can only get 40 people in the tougher group to fill, how many people would we need in the easier group to get at least 80% power to detect a difference of 10 points, assuming a standard deviation of 25, and using 90% confidence. (Remember that we met this standard with 80 people in each group using a balanced design)...

We have n1 = 40, d = 10/25 (δ / sd), sig.level = 0.1 and power = 0.8

- What's your guess, before I show you the answer, as to the number of people I'll need in the easier group?

Satiety Example, Unbalanced Design

```
library(pwr)
pwr.t2n.test(n1 = 40, d = 10/25, sig.level = .1,
             power = .80, alt="two.sided")
```

```
     t test power calculation

             n1 = 40
             n2 = 1174.101
              d = 0.4
      sig.level = 0.1
          power = 0.8
    alternative = two.sided
```

In Class Survey

Three Questions.

- Please fill it out, then raise your hand and we'll come get your response.
- Don't show your survey to others in the class, please.

Assessing Normality if Comparing Population Means

Paired Samples - compute paired differences, assess Normality

- Best tool: Histogram + Boxplot + Normal Q-Q plot
- Purpose: determine whether a t test is appropriate
- Big Deal? No.
  - If obviously non-Normal, avoid t test
  - If Normal seems like a reasonable approximation, but you're not certain, easy to run both t test and other approach (like bootstrap) - if they give similar answers, then not a problem. If they don't give similar answers, use the approach with fewer assumptions.

If you're having trouble getting calibrated, try this. Stop looking for a plot to scream "I am normal" and instead focus on making sure you identify plots that scream "I am NOT normal."

Assessing Normality when Comparing Population Means

Independent Samples - assess Normality in each of the two samples

- Best tool: Comparison boxplot + Faceted Histograms or Normal Q-Q plots, or Superimposed Density Functions
- Purpose: determine whether a t test is appropriate
- Big Deal? No.
  - If obviously non-Normal, avoid t test
  - If Normal seems like a reasonable approximation, but you're not certain, easy to run both t test and other approach (like bootstrap) - if they give similar answers, then not a problem. If they don't give similar answers, use the approach with fewer assumptions.

Equal Variances Assumed - a big deal?

Is the issue of which t test to use for independent samples (pooled t test or Welch's t test) an important one?

Practically, no.

- If the sample sizes are equal (or close), whether you use a pooled t test or a Welch t test will almost never yield an important difference in confidence intervals or p values.
- If the sample sizes are meaningfully different, then use a Welch test to be safe. If the sample variance in group 1 is between 2/3 and 3/2 the size of the variance in group 2, you'll rarely see a meaningful difference between the two approaches anyway.
- R defaults to Welch approximation.

Tool for Selecting a Comparison Procedure

If we want to compare the means of two populations,

1. Are these paired or independent samples?
2. If paired, then are the paired differences Normally distributed?
    1. Yes –> Use **paired t** test
    2. No – > are the differences reasonably symmetric?
        1. If symmetric, use **Wilcoxon signed rank** or **bootstrap** via `smean.cl.boot`
        2. If skewed, use **sign test** or **bootstrap** via `smean.cl.boot`
3. If independent, is each sample Normally distributed?
    1. No –> use **Wilcoxon-Mann-Whitney rank sum** test or **bootstrap**, via `bootdif`
    2. Yes –> are sample sizes equal?
        1. Balanced Design (equal sample sizes) - use **pooled t** test
        2. Unbalanced Design - use **Welch** test

2-Sample Study Design, Comparing Means

1. What is the outcome under study?
2. What are the (in this case, two) treatment/exposure groups?
3. Were the data collected using matched / paired samples or independent samples?
4. Are the data a random sample from the population(s) of interest? Or is there at least a reasonable argument for generalizing from the sample to the population(s)?
5. What is the significance level (or, the confidence level) we require here?
6. Are we doing one-sided or two-sided testing/confidence interval generation?
7. If we have paired samples, did pairing help reduce nuisance variation?
8. If we have paired samples, what does the distribution of sample paired differences tell us about which inferential procedure to use?
9. If we have independent samples, what does the distribution of each individual sample tell us about which inferential procedure to use?

The Cochlear Implant Study, 1

The `implant.csv` data are consonant recognition scores for 42 patients wearing second (S) or third (T) generation cochlear implants[1]. Third-generation devices incorporate changes and enhancements suggested by experience with second-generation devices.

```
knitr::kable(implant[1:4,])
```

patient	style	score
1	S	21
2	S	47
3	S	24
4	S	13

[1] mostly from Woodworth (2004) Biostatistics: A Bayesian Introduction, Table 4.5, from the Iowa Cochlear Implant Project, but with a change to one subject by TEL.

The Cochlear Implant Study, 2a

1. What is the outcome under study?
2. What are the (in this case, two) treatment/exposure groups?
3. Were the data collected using matched / paired samples or independent samples?

The Cochlear Implant Study, 2b

1. What is the outcome under study?
2. What are the (in this case, two) treatment/exposure groups?
3. Were the data collected using matched / paired samples or independent samples?

```
   patient       style       score      
Min.   : 1.00   S:21   Min.   : 8.00  
1st Qu.:11.25   T:21   1st Qu.:31.00  
Median :21.50          Median :45.00  
Mean   :21.50          Mean   :47.98  
3rd Qu.:31.75          3rd Qu.:70.00  
Max.   :42.00          Max.   :92.00  
```

Cochlear Implant Study, 3

4. Are the data a random sample from the population(s) of interest? Or is there at least a reasonable argument for generalizing from the sample to the population(s)?
5. What is the significance level (or, the confidence level) we require here?
6. Are we doing one-sided or two-sided testing/confidence interval generation?

What is H_0?

How about H_A?

Cochlear Implant Study, 4

H_0: $\mu_S = \mu_T$ vs. H_A: $\mu_S \neq \mu_T$

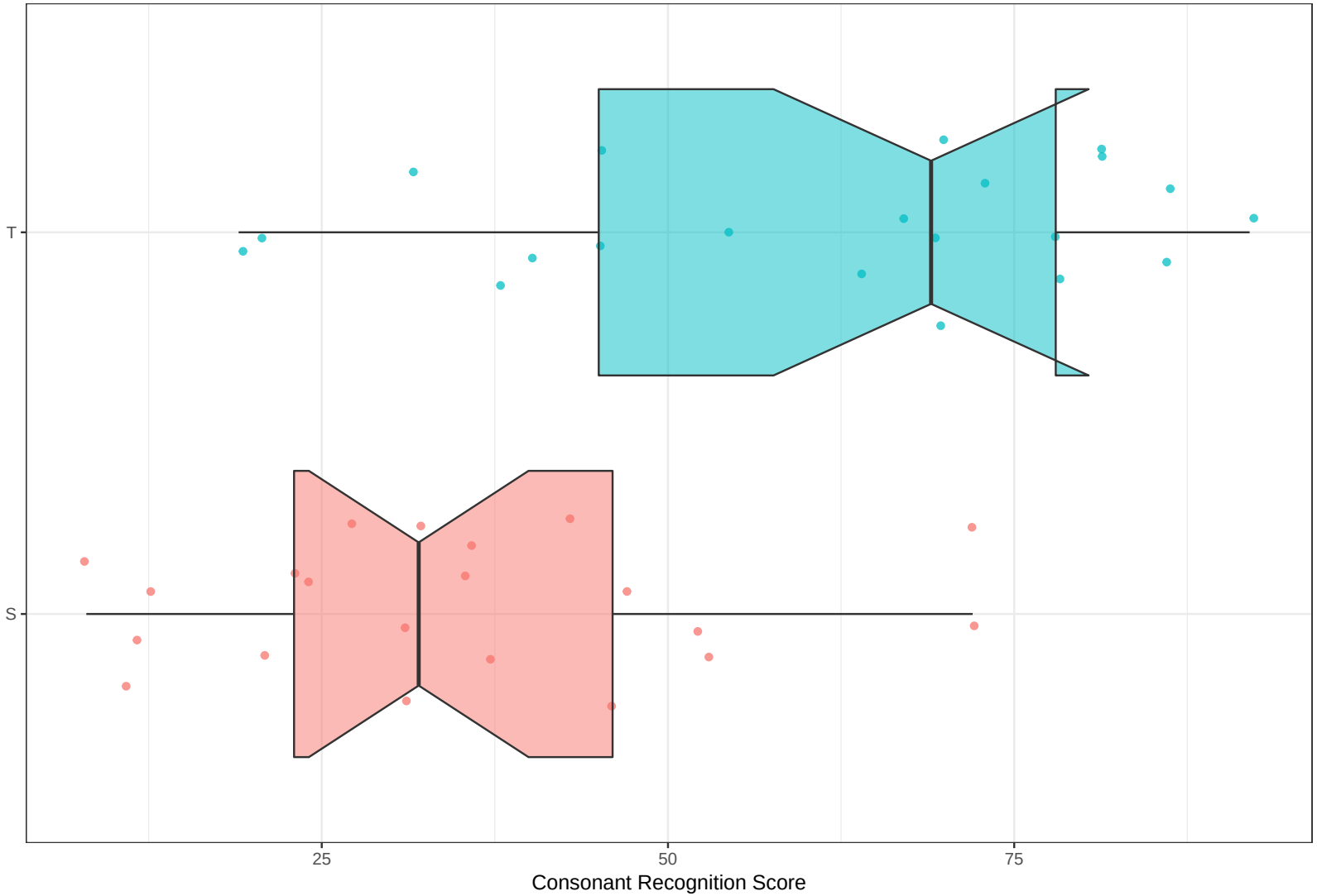
or, in our case. . .

H_0: $\mu_T - \mu_S = 0$ vs. H_A: $\mu_T - \mu_S \neq 0$

9. Since we have independent samples, what does the distribution of each individual sample tell us about which inferential procedure to use?

Cochlear Implant EDA: Best Choice of Test?

Independent Samples Comparison by Implant Type (S = 2nd Gen., T = 3rd Gen.)



Decision Tool for Two Independent Samples

H_0: $\mu_T - \mu_S = 0$ vs. H_A: $\mu_T - \mu_S \neq 0$

If independent, is each sample Normally distributed?

1. No –> use **Wilcoxon-Mann-Whitney rank sum** test or **bootstrap**, via `bootdif`
2. Yes –> are sample sizes equal?
    - Balanced Design (equal sample sizes) - use **pooled t** test
    - Unbalanced Design - use **Welch** test

Well?

Implant: Numerical Summary of the Samples

```
implant$style: S
 min Q1 median Q3 max     mean       sd  n missing
   8 23     32 46  72 34.57143 18.11787 21       0
------------------------------------------------------------
implant$style: T
 min Q1 median Q3 max     mean       sd  n missing
  19 45     69 78  92 61.38095 22.06462 21       0
```

Code to obtain results on next slide

```
## pooled t
t.test(implant$score ~ implant$style, var.equal = TRUE)
## Welch's t
t.test(implant$score ~ implant$style)
## rank sum
wilcox.test(implant$score ~ implant$style, exact = FALSE)
## bootstrap
set.seed(4313); bootdif(implant$score, implant$style)
```

Cochlear Implant Results

H_0: $\mu_T - \mu_S = 0$ vs. H_A: $\mu_T - \mu_S \neq 0$

Test	p value	Point Estimate	95% CI
Pooled t	.0001	26.8	(14.2, 39.4)
Welch's t	.0001	26.8	(14.2, 39.4)
Rank Sum	.0005	30.0	(14.0, 43.0)
Bootstrap	< 0.05	26.8	(14.9, 38.4)

What's our conclusion?

Angina Study

63 adult males with coronary artery disease were involved[2].

- On day A, they undergo an exercise test on a treadmill and we record the length of time from the start of the test until the patient experiences angina (pain or spasms in the chest).
  - They are then exposed to plain air for approximately one hour.
  - They then repeat the test and the time until the onset of angina is recorded again.
  - Outcome of interest is the percentage decrease in time to angina between the first and second tests.
- On day B, same tests and outcome, but during the interval between tests, they are exposed to a mixture of air and carbon monoxide, enough to increase the patient's carboxyhemoglobin level to 4% (lower than the level in smokers, but similar to that typically endured by a person in a poorly ventilated area in heavy traffic)

[2] Pagano and Gauvreau, 2000, Chapter 11.

The Angina Data ($n = 63$)

```
angina[1:5,]
```

```
# A tibble: 5 x 7
     id air.t1 air.t2    air co.t1 co.t2    co
  <int>  <int>  <int>  <dbl> <int> <int> <dbl>
1     1    983    957   2.65  1005   790 21.39
2     2    260    290 -11.54   256   239  6.64
3     3    709    489  31.03   593   400 32.55
4     4    655    490  25.19   710   520 26.76
5     5    505    490   2.97    NA    NA    NA
```

air = 100((air.t1 - air.t2)/air.t1) and co = 100((co.t1 - co.t2)/co.t1)

The Angina Study, 2a

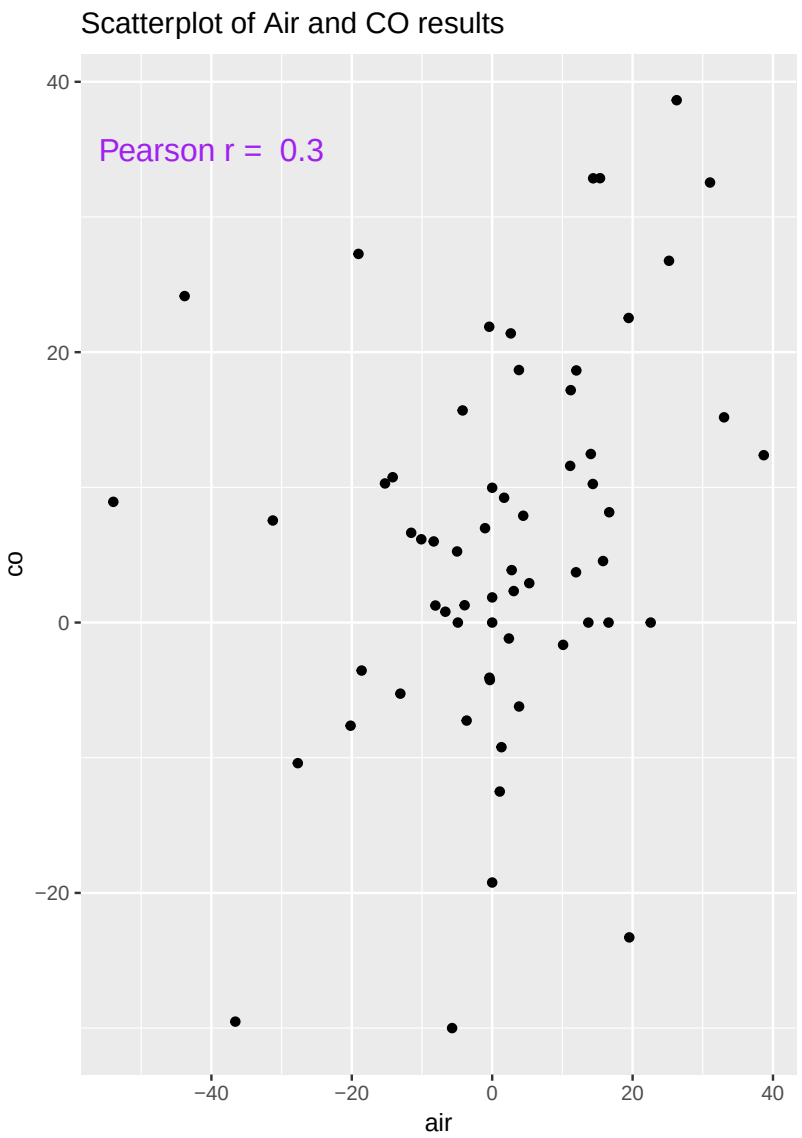
1. What is the outcome under study?
2. What are the (in this case, two) treatment/exposure groups?
3. Were the data collected using matched / paired samples or independent samples?

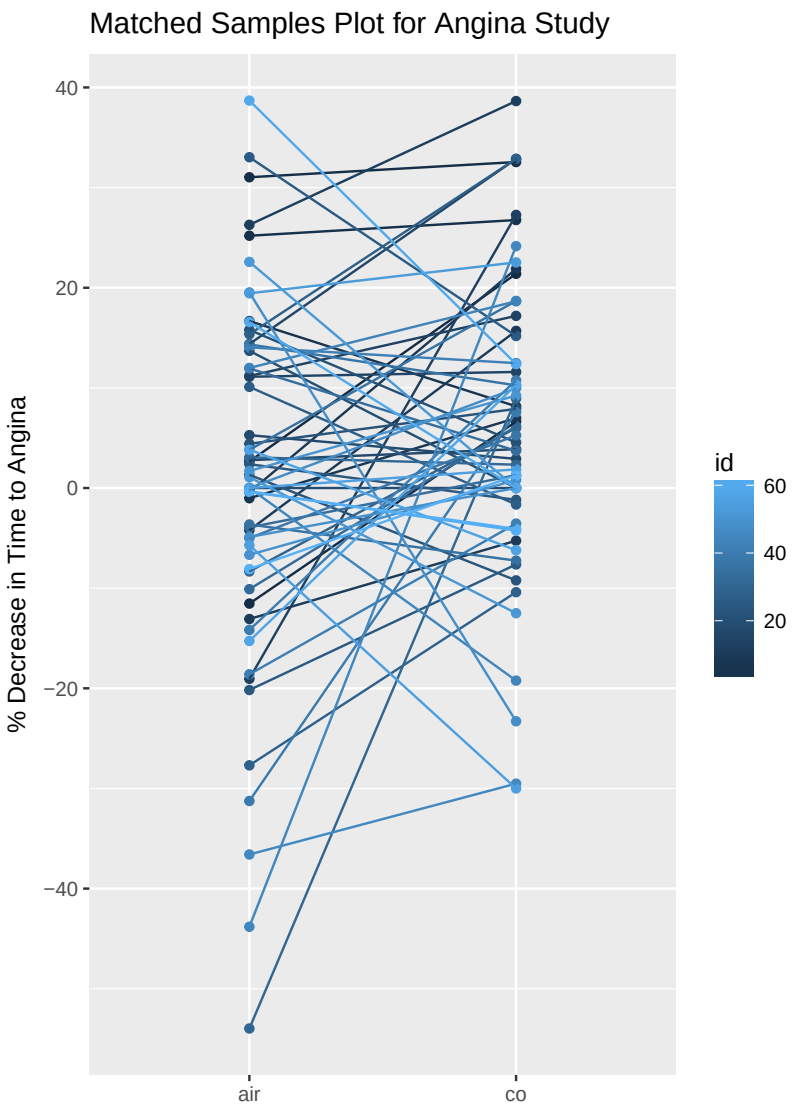
The Angina Study, 2b

1. What is the outcome under study?
2. What are the (in this case, two) treatment/exposure groups?
3. Were the data collected using matched / paired samples or independent samples?

```
# A tibble: 6 x 3
     id    air    co
  <int>  <dbl> <dbl>
1     1   2.65 21.39
2     2 -11.54  6.64
3     3  31.03 32.55
4     4  25.19 26.76
5     6  16.67  8.16
6     7  -1.02  6.98
```

Paired Samples? [1 sample removed]





The Angina Study, 3

4. Are the data a random sample from the population(s) of interest? Or is there at least a reasonable argument for generalizing from the sample to the population(s)?
5. What is the significance level (or, the confidence level) we require here?
6. Are we doing one-sided or two-sided testing/confidence interval generation?

What is H_0?

How about H_A?

The Angina Study, 4

H_0: $\mu_d = 0$ vs. H_A: $\mu_d \neq 0$

where μ_d = population mean of the CO - air differences.

```
ang2$diffs <- ang2$co - ang2$air
mosaic::favstats(ang2$diffs)
```

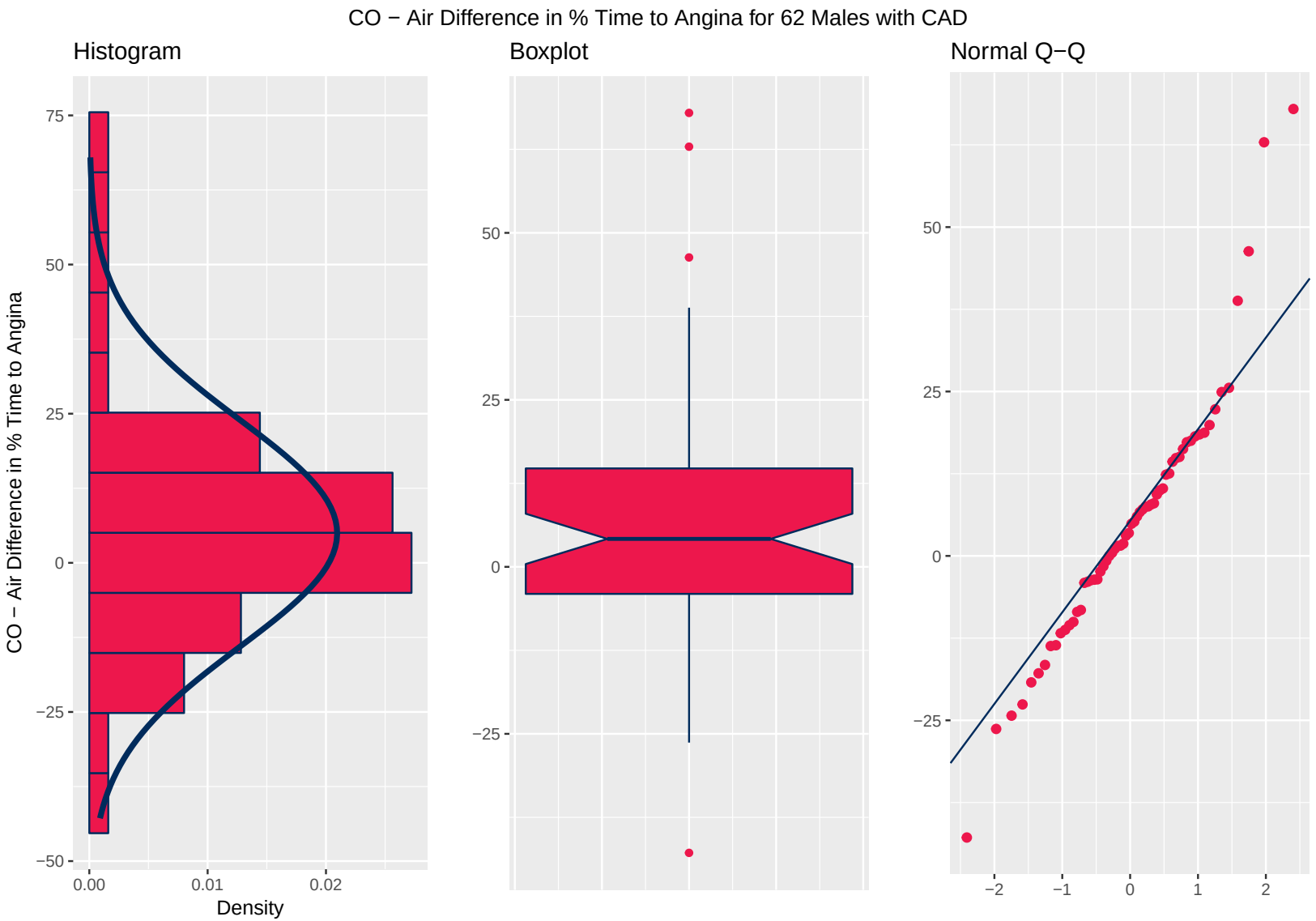
```
   min       Q1 median      Q3   max      mean        sd
-42.82 -4.0475  4.185 14.735 67.96 4.948387 19.04758
 n missing
62       0
```

The Angina Study, 5

7. If we have paired samples, did pairing help reduce nuisance variation?
8. If we have paired samples, what does the distribution of paired differences in the sample tell us about which inferential procedure to use?

- Recall that r = 0.3 for CO and Air, from our scatterplot earlier.

The Angina Study: EDA of Paired Differences



Decision Tool for Paired Samples

If paired, then are the paired differences Normally distributed?

1. Yes –> Use **paired t** test
2. No – > are the differences reasonably symmetric?
    1. If symmetric, use **Wilcoxon signed rank** or **bootstrap** via `smean.cl.boot`
    2. If skewed, use **sign test** or **bootstrap** via `smean.cl.boot`

What should we use here?

Code to obtain results on next slide

```
## paired t
t.test(ang2$diffs)
## rank sum
wilcox.test(ang2$diffs, conf.int = TRUE)
## bootstrap
set.seed(4314); Hmisc::smean.cl.boot(ang2$diffs)
```

Angina Study Results

H_0: $\mu_d = 0$ vs. H_A: $\mu_d \neq 0$

Test	p value	Point Estimate	95% CI
Paired t	.045	4.94	(0.1, 9.8)
Signed Rank	.062	4.15	(-0.2, 8.3)
Bootstrap	< 0.05	4.94	(0.2, 10.1)

What's our conclusion?

Building the Class Survey

To be developed.

The Signal and The Noise: Chapters 7 and 8

Predictions can be

- **self-fulfilling** (e.g. in election primary races) or
- **self-canceling** (e.g. when disease outbreaks are predicted, measures can be taken to prevent them, which can nullify the prediction)

When gauging the **strength** of a prediction, it's important to view the *inside* view in the context of the *outside* view.

- For example, many, if not most medical studies that claim 95% confidence aren't replicable.
- Should we take then 95% confidence figures at face value?

From Jonah Sinick at this link