431 Class 16

Thomas E. Love

2017-10-19

A Little Teaser on P Values

ASA Statement: "Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value."

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

. . . 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.**

Today's R Setup

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

library(broom); library(magrittr)
library(slipper); library(tidyverse)

dm192 <- read.csv("data/dm192.csv") %>% tbl_df
angina <- read.csv("data/angina.csv") %>% tbl_df
implant <- read.csv("data/implant.csv") %>% tbl_df

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

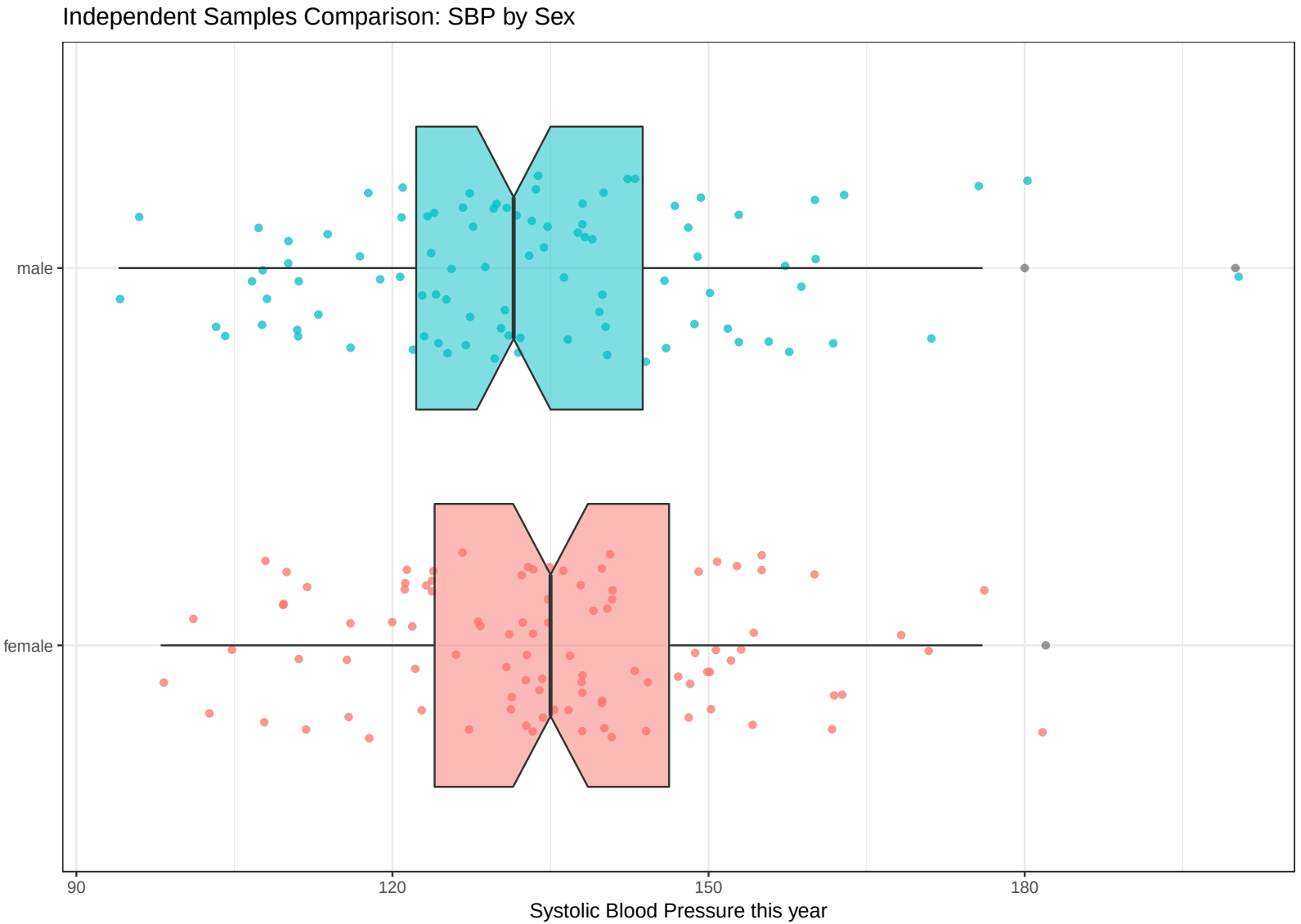
What if the Samples Aren't Paired?

Using the dm192 frame, is the average systolic blood pressure larger in **male** or in **female** adults in NE Ohio living with diabetes?

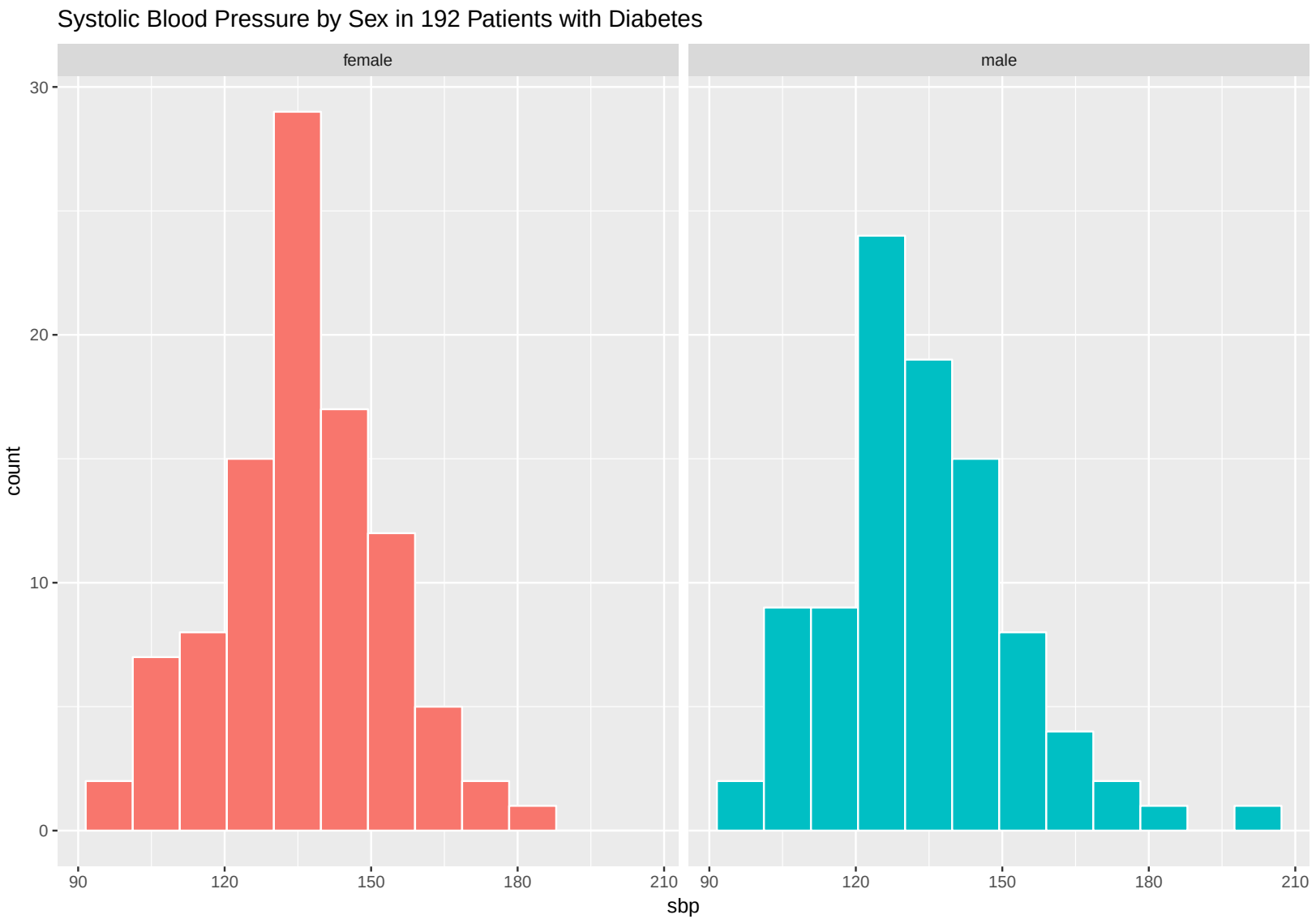
```
dm_second <- select(dm192, pt.id, sex, sbp)
summary(dm_second)
```

```
     pt.id            sex          sbp       
 Min.   :  1.00   female:98   Min.   : 94.0  
 1st Qu.: 48.75   male  :94   1st Qu.:123.0  
 Median : 96.50               Median :133.0  
 Mean   : 96.50               Mean   :134.2  
 3rd Qu.:144.25               3rd Qu.:144.5  
 Max.   :192.00               Max.   :200.0  
```

Our comparison now is between females and males



Another Way to Picture Two Independent Samples



Numerical Summary for Two Independent Samples

```
by(dm_second$sbp, dm_second$sex, mosaic::favstats)
```

```
dm_second$sex: female
 min  Q1 median     Q3 max     mean       sd  n
  98 124    135 146.25 182 135.1327 16.75637 98
 missing
       0
-------------------------------------
dm_second$sex: male
 min     Q1 median     Q3 max     mean       sd  n
  94 122.25  131.5 143.75 200 133.2447 18.82785 94
 missing
       0
```

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.

Testing Options for Independent Samples

1. Pooled t test or Indicator Variable Regression Model (t test assuming equal population variances)
2. Welch t test (t test without assuming equal population variances)
3. Wilcoxon-Mann-Whitney Rank Sum Test (non-parametric test not assuming populations are Normal)
4. Bootstrap confidence interval for the difference in population means

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:

```
t.test(dm_second$sbp ~ dm_second$sex, var.equal=TRUE)
```

```
    Two Sample t-test

data:  dm_second$sbp by dm_second$sex
t = 0.73467, df = 190, p-value = 0.4634
alternative hypothesis: true difference in means is not equal
95 percent confidence interval:
 -3.181093  6.957037
sample estimates:
mean in group female   mean in group male
            135.1327             133.2447
```

We can use regression to obtain these results, too.

Indicator Variable Regression

Regression Equation: sbp = 135.13 - 1.89 (sex = male)

- where (sex = male) = 1 if male, 0 if female

```
modA <- lm(sbp ~ sex, data = dm192)
modA
```

```

Call:
lm(formula = sbp ~ sex, data = dm192)

Coefficients:
(Intercept)      sexmale
    135.133       -1.888
```

Indicator Variable Regression Summary

```
> summary(modA)

Call:
lm(formula = sbp ~ sex, data = dm192)

Residuals:
    Min      1Q  Median      3Q     Max
-39.245 -11.161  -1.133  11.033  66.755

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  135.133      1.798  75.152   <2e-16 ***
sexmale       -1.888      2.570  -0.735    0.463
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.8 on 190 degrees of freedom
Multiple R-squared:  0.002833,  Adjusted R-squared:  -0.002416
F-statistic: 0.5397 on 1 and 190 DF,  p-value: 0.4634
```

Male - Female difference point estimate: -1.89, $p = 0.463$

Confidence Interval via Regression

Our point estimate of the male - female difference is -1.89, with standard error 2.57, and two-sided $p = 0.463$

```
confint(modA)
```

```
                 2.5 %     97.5 %
(Intercept) 131.585817 138.679490
sexmale      -6.957037   3.181093
```

Tidying a Two-Sample (Pooled) t Test

```
dm_second %$%
  tidy(t.test(sbp ~ sex, var.equal = TRUE))
```

```
  estimate1 estimate2 statistic   p.value parameter
1  135.1327  133.2447 0.7346677 0.4634476       190
   conf.low conf.high             method alternative
1 -3.181093  6.957037  Two Sample t-test   two.sided
```

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 choice in R.

Welch t test without assuming equal population variances

```
t.test(dm_second$sbp ~ dm_second$sex)
```

```
    Welch Two Sample t-test

data:  dm_second$sbp by dm_second$sex
t = 0.73288, df = 185.39, p-value = 0.4646
alternative hypothesis: true difference in means is not equal
95 percent confidence interval:
 -3.194236  6.970180
sample estimates:
mean in group female   mean in group male
            135.1327             133.2447
```

Tidying a Two-Sample (Welch) Test

```
dm_second %$%
  tidy(t.test(sbp ~ sex))
```

```
  estimate estimate1 estimate2 statistic   p.value
1 1.887972  135.1327  133.2447 0.7328845 0.4645545
  parameter  conf.low conf.high
1  185.3938 -3.194236   6.97018
                     method alternative
1 Welch Two Sample t-test   two.sided
```

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.

Wilcoxon-Mann-Whitney Rank Sum Test

The rank sum test is a non-parametric test of whether the two samples were selected from populations having the same distribution.

```
wilcox.test(dm_second$sbp ~ dm_second$sex, conf.int = TRUE)
```

```
    Wilcoxon rank sum test with continuity
    correction

data:  dm_second$sbp by dm_second$sex
W = 5035.5, p-value = 0.2649
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -2.000061  7.999993
sample estimates:
difference in location
              2.999918
```

The Continuity Correction

The *p* value for the rank sum test is obtained via a Normal approximation, using the test statistic W.

- That approximation can be slightly improved through the use of a continuity correction (a small adjustment to account for the fact that we're using a continuous distribution, the Normal, to approximate a discretely valued test statistic, W.)
- The continuity correction is particularly important in the case where we have many tied ranks, and is applied by default in R.
- If you want (for some reason) to not use it, add `correct = FALSE` to your call to the `wilcox.test()` function.

Rank Sum Test vs. Signed Rank Test

Each tests whether two samples were selected from populations having the same distribution.

1. The Rank Sum test (Wilcoxon - Mann - Whitney) is for **independent samples** comparisons.
    - Assign numerical ranks to all observations across the two groups.
        - 1 = smallest, n = largest. Use midpoint for any ties.
    - Add up the ranks from sample 1. Call that R_1.
        - R_2 is then known, since the sum of all ranks is $\frac{n(n+1)}{2}$
    - $U_1 = R_1 - \frac{n_1(n_1+1)}{2}$, where n_1 is the sample size for sample 1.
    - $U_1 + U_2$ is always just $n_1 n_2$, so it doesn't matter which sample you treat as sample 1.
    - The smaller of U_1 and U_2 is then called U, the test statistic.
    - Software converts U into a p value via a Normal approximation, given n_1 and n_2.

More details at the Wikipedia entry for Mann-Whitney U test.

Rank Sum Test vs. Signed Rank Test

Each tests whether two samples were selected from populations having the same distribution.

2. The Signed Rank test is for **paired samples** comparisons.
   - Calculate the paired difference for each pair, and drop those with difference = 0.
   - Let N be the number of pairs, so there are 2N data points.
   - Rank the pairs in order of smallest (rank = 1) to largest (rank = N) absolute difference.
   - Calculate W, the sum of the signed ranks by

$$W = \sum_{i=1}^{N} [sgn(x_{2,i} - x_{1,i})] \prod R_i)]$$

     - The sign function sgn(x) = -1 if x $<$ 0, 0 if x = 0, and +1 if x $>$ 0.
   - Statistical software will convert W into a p value, given N.

More details at the Wikipedia entry for Wilcoxon signed-rank test.

Tidying a Wilcoxon Rank Sum Test

```
dm_second %$%
  tidy(wilcox.test(sbp ~ sex, conf.int = TRUE))
```

```
  estimate statistic   p.value  conf.low conf.high
1 2.999918    5035.5 0.2648795 -2.000061  7.999993
                                                 method
1 Wilcoxon rank sum test with continuity correction
  alternative
1   two.sided
```

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 on the web site, and is also part of this Markdown file.

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 2.5^{th} and 97.5^{th} 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 the following. . .

```
set.seed(4314); bootdif(dm_second$sbp, dm_second$sex)
```

```
Mean Difference           0.025           0.975
      -1.887972      -6.977860        2.917249
```

Note that the two columns must be separated here with a comma rather than a tilde (~).

This CI describes the male - female difference (i.e. the negative of the F-M difference used earlier) – we can tell this by the listed sample mean difference.

Can we use slipper instead?

For differences in means between independent samples, we can use the tidy function in broom to obtain the point estimate, and then slipper on that result.

```
tidy(t.test(dm_second$sbp ~ dm_second$sex))
```

```
  estimate estimate1 estimate2 statistic   p.value
1 1.887972  135.1327  133.2447 0.7328845 0.4645545
  parameter  conf.low conf.high
1  185.3938 -3.194236   6.97018
                    method alternative
1 Welch Two Sample t-test   two.sided
```

Using `slipper` to run a bootstrap CI

For comparing the means of independent samples:

```
# requires library(slipper)
set.seed(4313)
dm_second %>%
  slipper((tidy(t.test(sbp ~ sex))$estimate),
          B = 500) %>%
  summarise(bootci_low = quantile(value, 0.025),
            bootci_high = quantile(value, 0.975))
```

```
  bootci_low bootci_high
1  -2.687109    6.970497
```

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 test	0.463	(-3.2, 7.0)
Welch t test	0.465	(-3.2, 7.0)
Rank Sum test	0.265	(-2.0, 8.0)
Bootstrap CI	$p > 0.05$	(-2.9, 7.0) via `bootdif`
Bootstrap CI	$p > 0.05$	(-2.7, 7.0) via `slipper`

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

On Reporting p Values

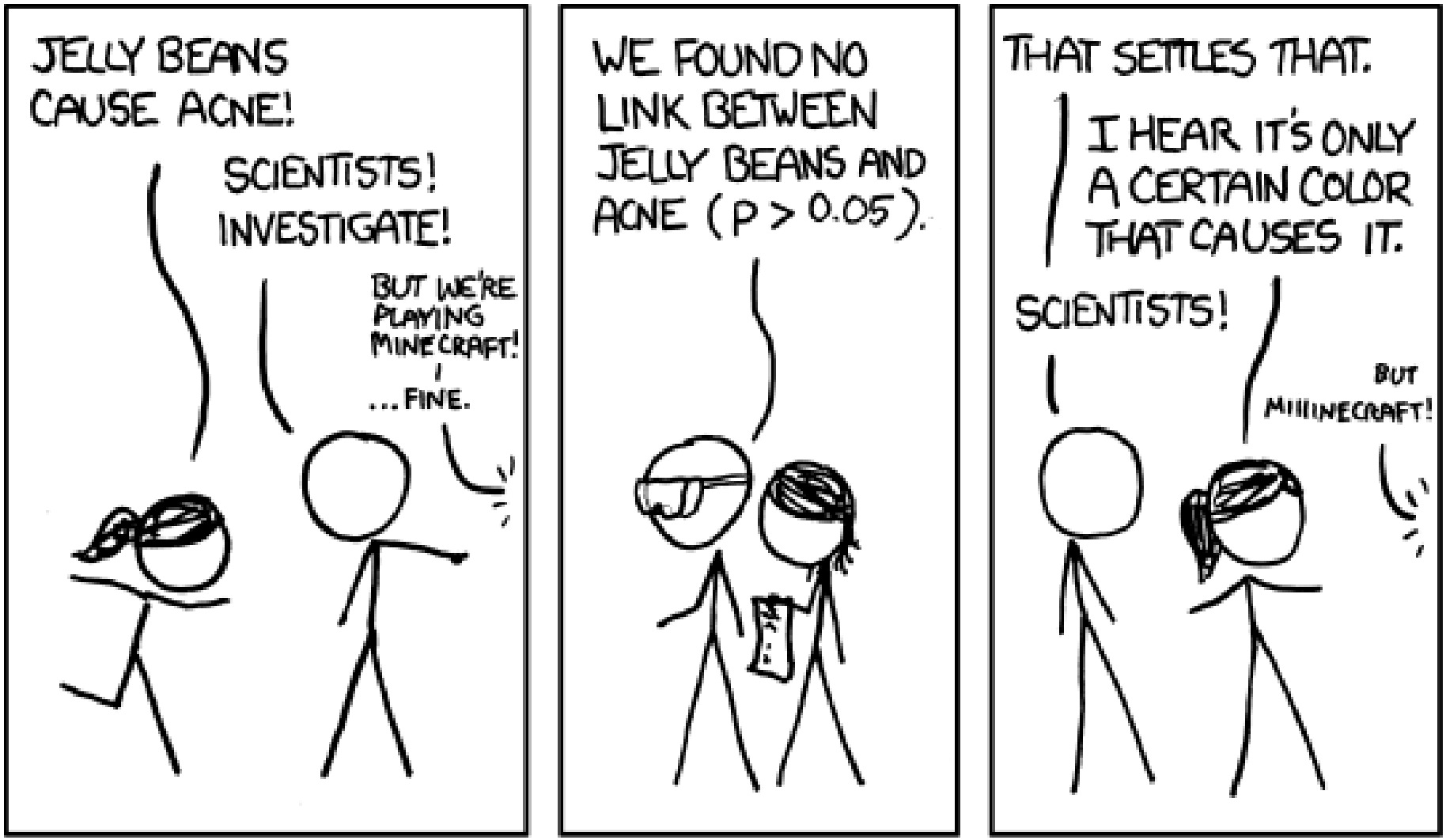
When reporting a p value and no rounding rules are in place from the lead author/journal/source for publication, follow these conventions. . .

1. Use an italicized, lower-case p to specify the p value. Don't use p for anything else.
2. For p values above 0.10, round to two decimal places, at most.
3. For p values near α, include only enough decimal places to clarify the reject/retain decision.
4. For very small p values, always report either $p < 0.0001$ or even just $p < 0.001$, rather than specifying the result in scientific notation, or, worse, as $p = 0$ which is glaringly inappropriate.
5. Report p values above 0.99 as $p > 0.99$, rather than $p = 1$.

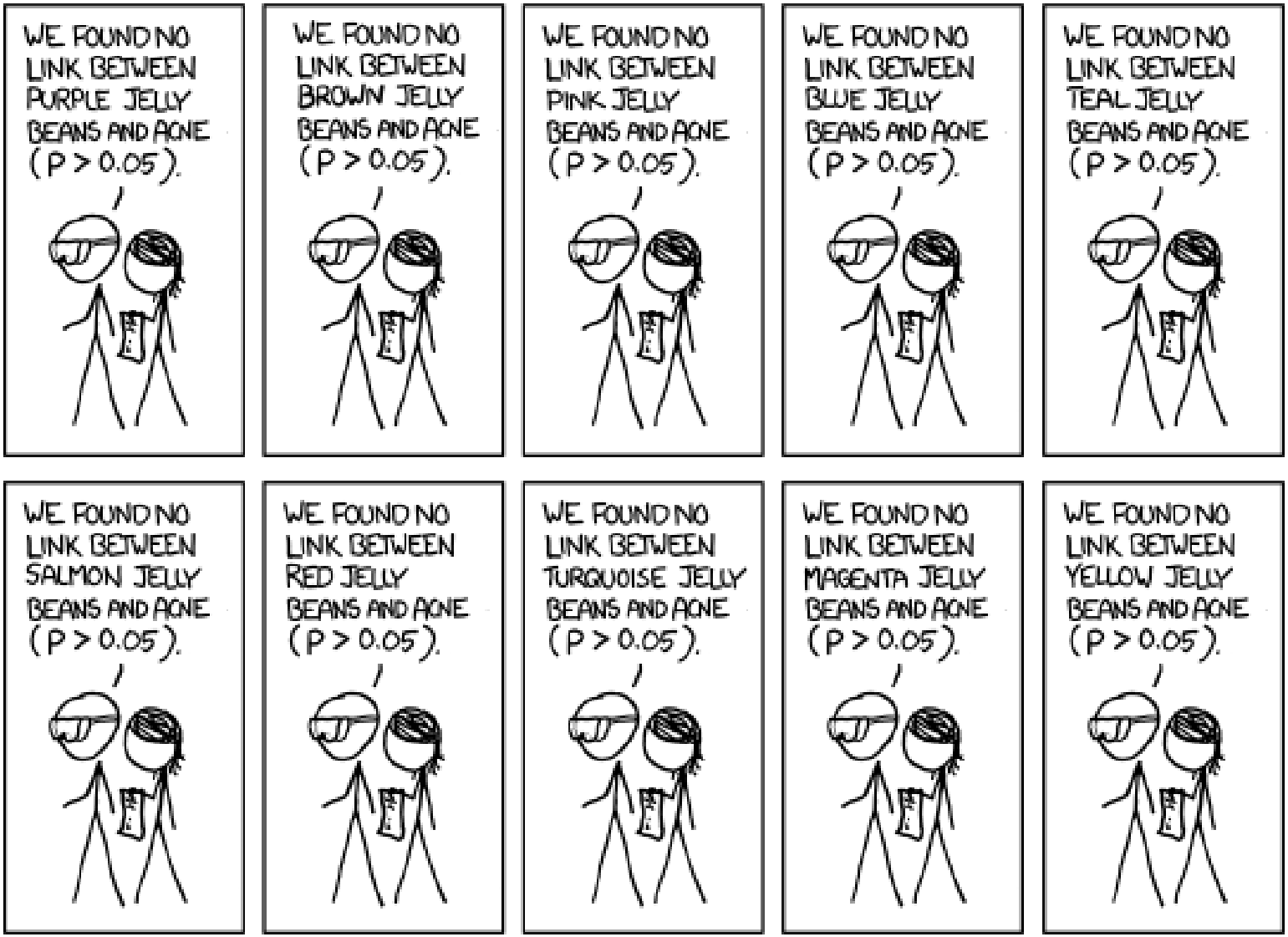
A Few Comments on Significance

- **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.

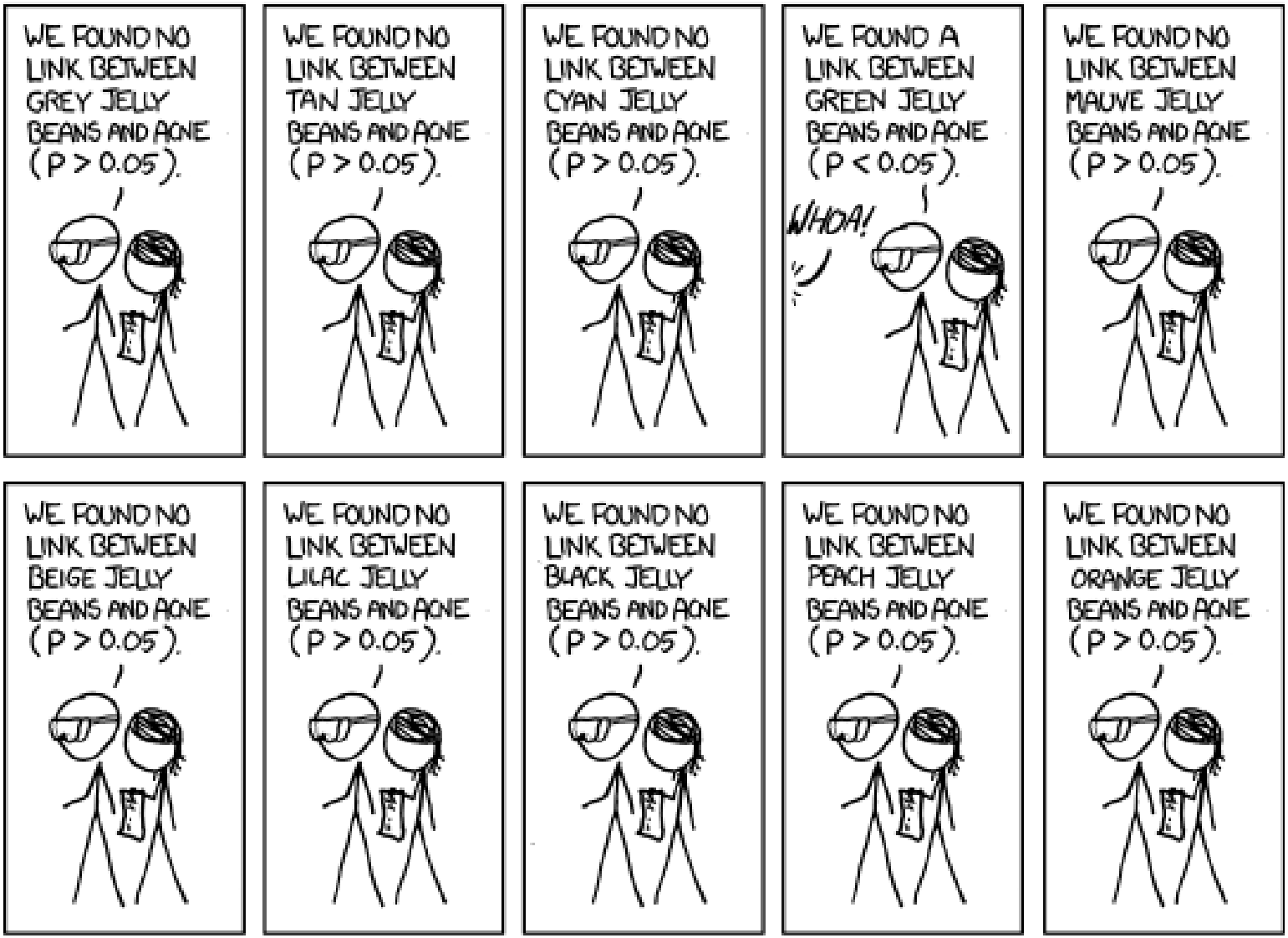
From XKCD (https://xkcd.com/882/)



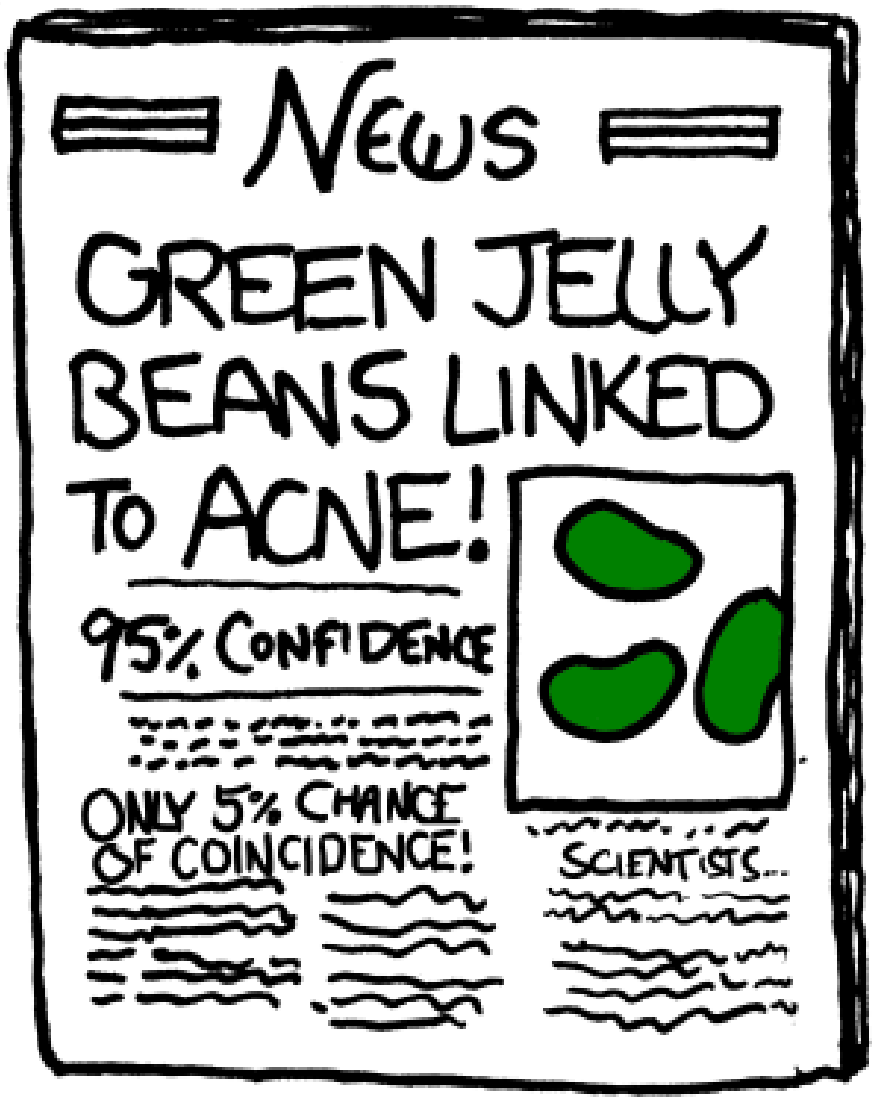
From XKCD (https://xkcd.com/882/)



From XKCD (https://xkcd.com/882/)



From XKCD (https://xkcd.com/882/)



From George Cobb - on why *p* values deserve to be re-evaluated

The **idea** of a p-value as one possible summary of evidence

morphed into a

- **rule** for authors: reject the null hypothesis if $p < .05$.

From George Cobb - on why *p* values deserve to be re-evaluated

The **idea** of a p-value as one possible summary of evidence

morphed into a

- **rule** for authors: reject the null hypothesis if $p < .05$,

which morphed into a

- **rule** for editors: reject the submitted article if $p > .05$.

From George Cobb - on why *p* values deserve to be re-evaluated

The **idea** of a p-value as one possible summary of evidence

morphed into a

- **rule** for authors: reject the null hypothesis if $p < .05$,

which morphed into a

- **rule** for editors: reject the submitted article if $p > .05$,

which morphed into a

- **rule** for journals: reject all articles that report p-values[1]

[1] http://www.nature.com/news/psychology-journal-bans-p-values-1.17001 describes the banning of null hypothesis significance testing by *Basic and Applied Psychology*.

From George Cobb - on why *p* values deserve to be re-evaluated

The **idea** of a p-value as one possible summary of evidence

morphed into a

- **rule** for authors: reject the null hypothesis if $p < .05$, which morphed into a
- **rule** for editors: reject the submitted article if $p > .05$, which morphed into a
- **rule** for journals: reject all articles that report p-values.

Bottom line: **Reject rules. Ideas matter.**

Posted to an American Statistical Association message board Oct 14 2015

How Big A Sample Size Do I need?

1. What is the budget?
2. What are you trying to compare?
3. What is the study design?
4. How big an effect size do you expect (hope) to see?
5. What was that budget again?
6. OK, tell me the maximum allowable rates of Type I and Type II error that you want to control for. Or, if you like, tell me the confidence level and power you want to have.
7. And what sort of statistical inference do you want to plan for?

A Formula for Decoding Health News

Health Headlines are Advertising

Think about this headline: "Hospital checklist cut infections, saved lives."

- Suppose you are a little surprised that a checklist could really save lives. If you think say the odds of this being true are 1 in 4, you would set your initial gut feeling to 1/4. Because this number is less than one, it means initially you're less likely to believe the study.

Bayes' Rule

Final opinion = (initial gut feeling) * (study support for headline)
Source: Jeff Leek, *fivethirtyeight.com*

Assessing Study Support for a Headline

1. Was the study a clinical study in humans?
2. Was the outcome of the study something directly related to human health like longer life or less disease? Was the outcome something you care about, such as living longer or feeling better?
3. Was the study a randomized, controlled trial (RCT)?
4. Was it a large study (at least hundreds of patients)?
5. Did the treatment have a major impact on the outcome?
6. Did predictions hold up in at least two separate groups of people?

Assessing Study Support

Support for Headline: Multiply by 2 for every yes, and 1/2 for every no.

Evaluating A Research Article

Intensive care units (ICUs) at Michigan hospitals implemented a new strategy for reducing infections through training, a daily goals sheet and a program to improve the culture of safety in the ICUs. The doctors measured the rate of infection before and after implementing this safety program.

1. Was the study a clinical study in humans?
    - The study was done in humans in ICUs (+)
2. Was the outcome of the study something directly related to human health like longer life or less disease? Was the outcome something you care about, such as living longer or feeling better?
    - The outcome was the rate of infections after surgery; according to the article, these infections cost U.S. hospitals up to $2.3 billion annually. (+)

Evaluating a Research Article

3. Was the study a randomized, controlled trial (RCT)?
    - The study compared the same hospitals before and after a change in ICU policy. This is an example of a crossover study, which is not as strong as a randomized trial but does account for some of the differences among hospitals because the same ICUs were measured before and after using the checklist. (-)
4. Was it a large study (at least hundreds of patients)?
    - The study looked at more than 100 ICUs over 1,981 months. In total, it followed patients for 375,757 catheter-days. (A catheter-day means watching one patient for one day while she is on a catheter.) This is a huge number of days to monitor patients for potential infections. (+)
5. Did the treatment have a major impact on the outcome?
    - The study showed a sustained drop of up to 66 percent in infections. (+)
6. Did predictions hold up in at least two separate groups of people?
    - The study looked at 103 hospitals in Michigan. (+)

So we have 5 + and 1 - in our evaluation of this study.

Final Opinion?

- So, a large study showed a major drop in infections, and that is directly medically important. For the sake of the exercise, let's multiply by two every time we see a `yes` answer and by 1/2 every time we see a `no` answer. I would say this study's result is about 16 times more likely (five out of six `yes` answers or 2 x 2 x 2 x 2 x 2 x (1/2) = 16) if checklists really do reduce infections than if they don't. I set study support for headline = 16.
- Multiply to get final opinion on headline = 1/4*16 = 4, also expressed as 4/1. I would say that my updated odds are 4 to 1 that the headline is true. The strength of the study won over my initially skeptical gut feeling.

Bayes' Rule

Final opinion = (initial gut feeling) * (study support for headline)
Source: Jeff Leek, *fivethirtyeight.com*

Evaluating Health News: For Consumers

1. Watch out for single source stories. They're usually based on a press release, which will have a hidden agenda.
2. Beware of stories that don't mention cost. It's crucial information. (If the cost of the great, new treatment is out of reach – it's not that great, is it?)
3. Headline percentages are misleading. If something "reduces your risk of X by 50%" chances are that number doesn't mean what you think it means.
4. If it sounds too good to be true, it probably is. If a report presents only or primarily the benefits of a new treatment, it's a bad report. ALL healthcare interventions have trade-offs.
5. Patient anecdotes are not data. Beware of stories that rely on them. Anecdotes are used to compensate for data that are unavailable or flawed.

Source: *NPR*

Evaluating Health News: For Consumers

6. A "simple screening test" is never simple. The decision to take one is one of the most complex and difficult decisions a health consumer can make.
7. Watch out for hyperbolic language. "Breakthrough", "first-of-its-kind", and "game-changer" are red flags. When you read "it may become..." substitute "it may not become..."
8. Newer isn't always better. Often the latest test, treatment or procedure is no better than what already exists, just pricier.
9. Beware of disease-mongering. Risk factors, symptoms for diseases, or data can be exaggerated in a way that causes needless worry, and expense.
10. The latest treatment may not exist yet, or ever. "Awaiting FDA approval" or "in pre-clinical trial phase" means it's still a pipe dream.
11. There is a leap from mice to men. Getting from rodent trials to human use is a very, very long road, that may in fact lead nowhere.

Source: *NPR*

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)
- $1 - \beta$ = 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.

Another 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, at 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*
```

Assessing Normality when 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[2]. 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

[2] 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

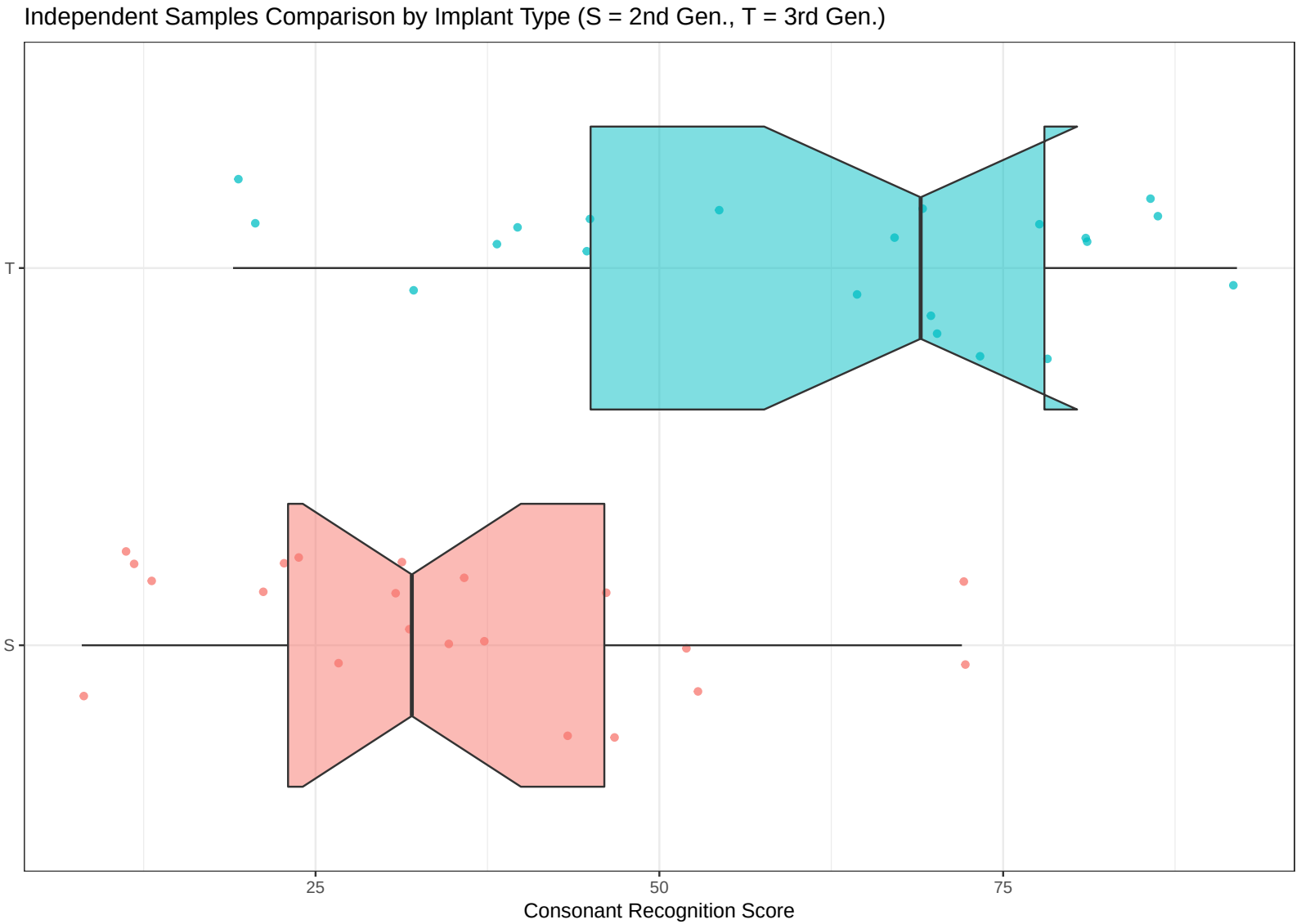
H_0: $\mu_S = \mu_T$ vs. H_A: $\mu_S \neq \mu_T$

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?



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
```

Results of all four tests are available in the Markdown file

```
## 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)
```

```
    Two Sample t-test

data:  implant$score by implant$style
t = -4.3032, df = 40, p-value = 0.0001055
alternative hypothesis: true difference in means is not equal
```

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[3].

- 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)

[3] 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

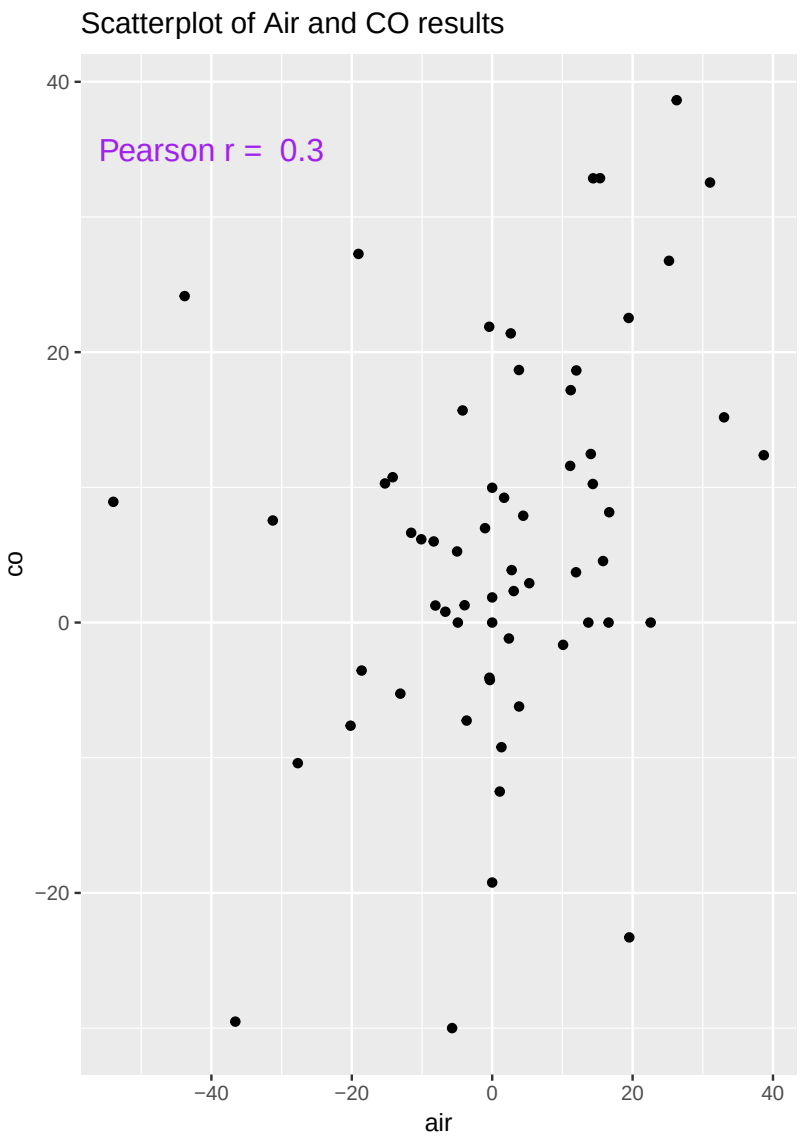
1. What is the outcome under study?
2. What are the (in this case, two) treatment/exposure groups?
3. Were the data collected using matched / paired samples or independent samples?

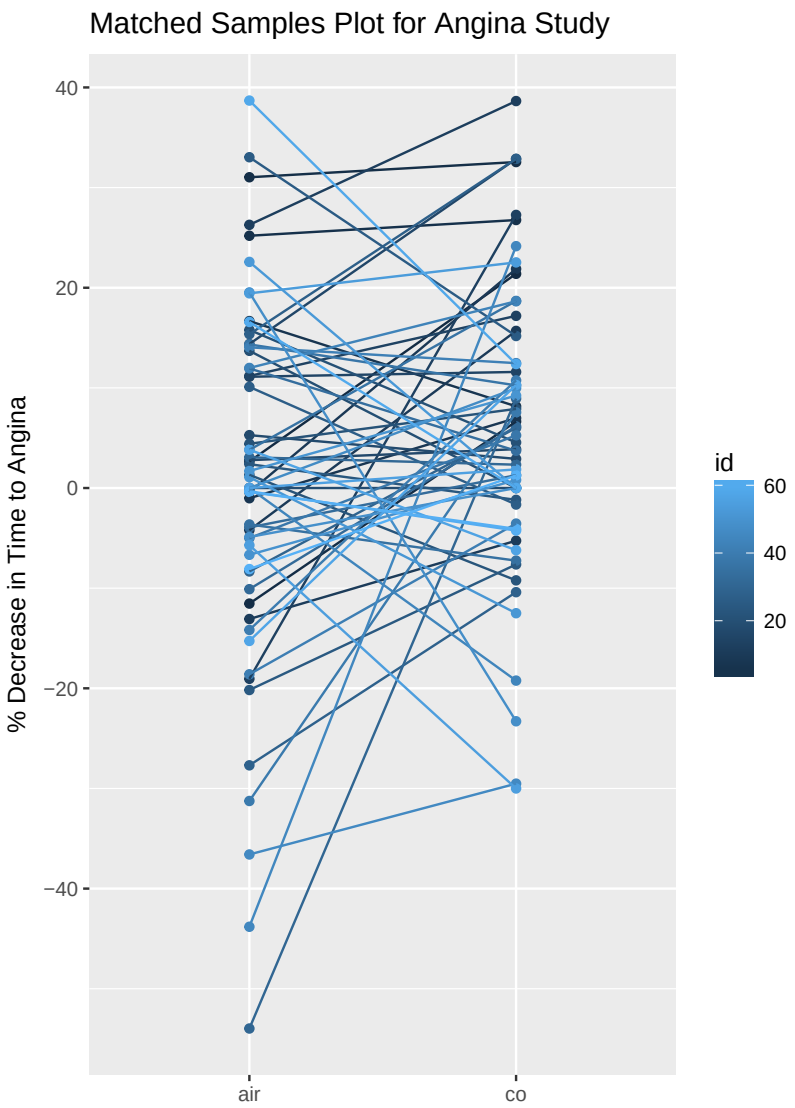
The Angina Study, 2b

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)
```

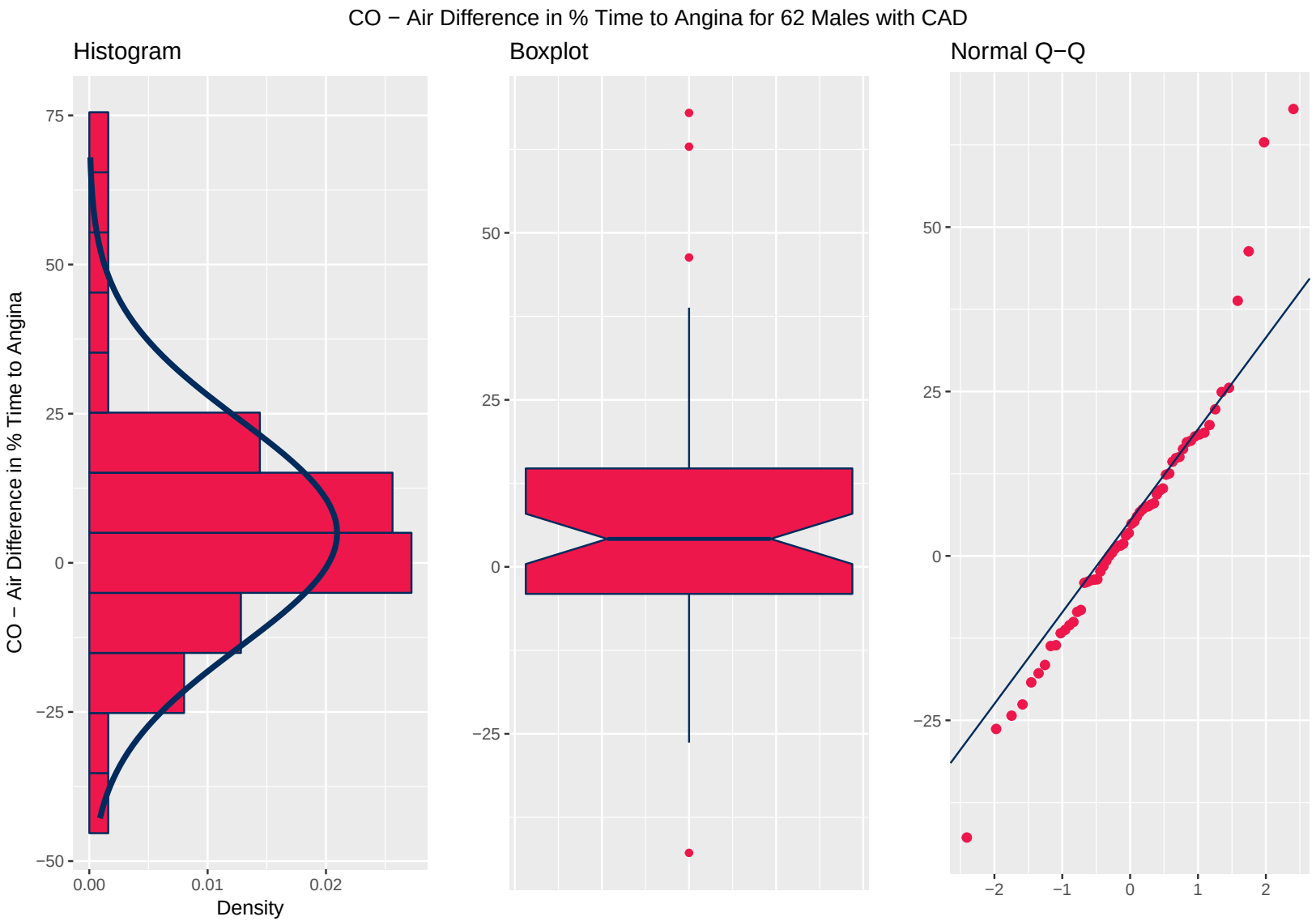
```
   min       Q1 median      Q3   max      mean        sd
-42.82 -4.0475  4.185 14.735 67.96 4.948387 19.04758
 n missing
62       0
```

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?

Results for 3 Analytic Approaches

```
## 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)
```

```
    One Sample t-test

data:  ang2$diffs
t = 2.0456, df = 61, p-value = 0.04512
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.1112082 9.7855660
sample estimates:
mean of x
```

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?