431 Class 17

Thomas E. Love

2018-10-30

Today's Agenda

- Comparing Two Population Means
  - A New Example
  - Decision Support
  - Bootstrapping using the `slipper` package
- On P Values: A Little Taste
- Power and Sample Size Considerations Comparing 2 Means
  - With power.t.test for balanced designs
  - With pwr for unbalanced designs

Today's R Setup

```
source("Love-boost.R") # helps to load Hmisc explicitly

library(pwr); library(broom); library(Hmisc)
library(tidyverse) # always load tidyverse last
```

We'll also install and load the `slipper` package, from Jeff Leek - see https://github.com/jtleek/slipper

```
devtools::install_github('jtleek/slipper')
library(slipper)
```

The Research Question

Suppose we consider the population of adults with diabetes in Northeast Ohio, from whom we have sampled the data in `dm192`.

Suppose we want to compare the population mean of LDL cholesterol for the adults with **Medicare** insurance to the population mean of LDL cholesterol for the adults with **Medicaid** insurance.

This will involve filtering our sample to include only those subjects with:

- insurance of either "medicare" or "medicaid" but not "commercial" or "uninsured"
- complete data on LDL cholesterol, since we don't want to deal with missingness today

We'll create a new data set called `dm_third` to do this.

Creating the dm_third tibble we'll need

```r
dm192 <- read.csv("data/dm192.csv") %>% tbl_df

dm_third <- dm192 %>%
  filter(insurance %in% c("medicare", "medicaid")) %>%
  filter(complete.cases(ldl, insurance)) %>%
  select(pt.id, ldl, insurance) %>%
  droplevels() # drop unused levels from insurance factor

tail(dm_third, 3) # show last 3 rows
```

```
# A tibble: 3 x 3
  pt.id   ldl insurance
  <int> <int> <fct>
1   187   105 medicare
2   189    74 medicare
3   191   158 medicaid
```

2-Sample Study Design, Comparing Means

Suppose we want to compare the population mean of LDL cholesterol for the adults with *Medicare* insurance to the population mean of LDL cholesterol for the adults with *Medicaid* insurance.

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)?

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

Distribution of LDL by Insurance Group in `dm_third`

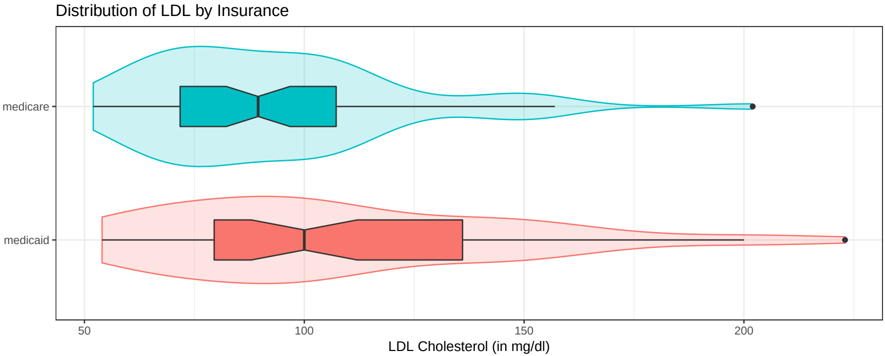


Table 1: LDL by Insurance in dm_third

insurance	n	mean	median	sd	min	max
medicaid	55	107.4	100.0	39.5	54	223
medicare	60	93.6	89.5	29.4	52	202

Code to Accomplish Previous Slide

```r
ggplot(dm_third, aes(x = insurance, y = ldl,
                     fill = insurance)) +
    geom_violin(aes(col = insurance), alpha = 0.2) +
    geom_boxplot(notch = TRUE, width = 0.3) +
    coord_flip() +
    guides(fill = FALSE, color = FALSE) +
    theme_bw() +
    labs(x = "", y = "LDL Cholesterol (in mg/dl)",
         title = "Distribution of LDL by Insurance")

dm_third %>% group_by(insurance) %>%
  summarize(n = n(), mean = round(mean(ldl),1),
            median = median(ldl), sd = round(sd(ldl),1),
            min = min(ldl), max = max(ldl)) %>%
  knitr::kable(caption = "LDL by Insurance in dm_third")
```

2-Sample Study Design, Comparing Means

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?

Independent Samples Results: Pooled t test

```r
t.test(ldl ~ insurance, data = dm_third, var.equal = TRUE)
```

```

    Two Sample t-test

data:  ldl by insurance
t = 2.1352, df = 113, p-value = 0.0349
alternative hypothesis: true difference in means is not equal
95 percent confidence interval:
  0.9956012 26.6043988
sample estimates:
mean in group medicaid mean in group medicare
                 107.4                   93.6
```

Based on these results, what can we conclude? Do the assumptions of this procedure match well to our data?

Independent Samples Results: Welch's t test

```r
t.test(ldl ~ insurance, data = dm_third)
```

```

    Welch Two Sample t-test

data:  ldl by insurance
t = 2.1084, df = 99.284, p-value = 0.03751
alternative hypothesis: true difference in means is not equal
95 percent confidence interval:
  0.8135641 26.7864359
sample estimates:
mean in group medicaid mean in group medicare
                 107.4                   93.6
```

Based on these results, what can we conclude? Do the assumptions of this procedure match well to our data?

Independent Samples Results: Wilcoxon-Mann-Whitney Rank Sum Test

```
wilcox.test(ldl ~ insurance, data = dm_third, conf.int = T)
```

```

    Wilcoxon rank sum test with continuity
    correction

data:  ldl by insurance
W = 1968.5, p-value = 0.07494
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -0.9999383 23.9999208
sample estimates:
difference in location
              10.00003
```

Based on these results, what can we conclude?

Using the bootdif function to compare means based on independent samples

So, to compare LDL cholesterol (our outcome) across the two levels of insurance (our grouping factor) for the subset of our original sample adult patients with diabetes in NE Ohio, run the following…

```
set.seed(20181030)
bootdif(dm_third$ldl, dm_third$insurance)
```

```
Mean Difference           0.025           0.975
    -13.8000000     -27.1150000      -0.7010227
```

Based on these results, what can we conclude? Do the assumptions of this procedure match well to our data?

Another Bootstrapping Approach - the slipper package

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

```
tidy(t.test(dm_third$ldl ~ dm_third$insurance))
```

```
# A tibble: 1 x 10
  estimate estimate1 estimate2 statistic p.value
     <dbl>     <dbl>     <dbl>     <dbl>   <dbl>
1     13.8      107.      93.6      2.11  0.0375
# ... with 5 more variables: parameter <dbl>,
#   conf.low <dbl>, conf.high <dbl>, method <chr>,
#   alternative <chr>
```

Using `slipper` to run a bootstrap CI

For comparing the means of independent samples:

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

```
  bootci_low bootci_high
1   2.130882    26.27912
```

Again: LDL by Insurance Group in `dm_third`

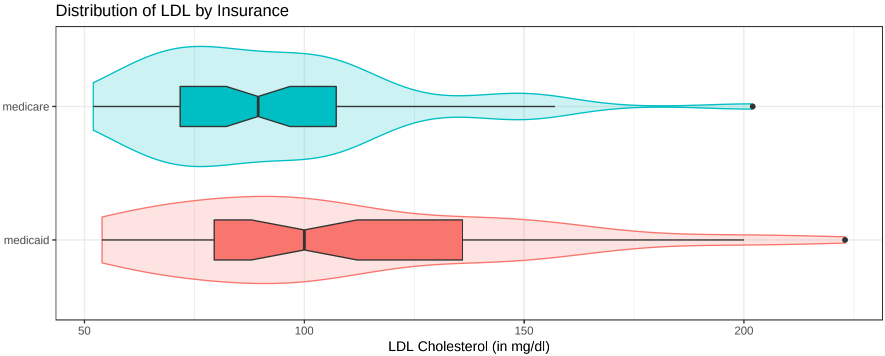


Table 2: LDL by Insurance in dm_third

insurance	n	mean	median	sd	min	max
medicaid	55	107.4	100.0	39.5	54	223
medicare	60	93.6	89.5	29.4	52	202

Results for the LDL and Insurance Sub-Study

Procedure	p for $H_0 : \mu_{medicaid} = \mu_{medicare}$	95% CI for $\mu_{medicaid} - \mu_{medicare}$
Pooled t test	0.035	(1.0, 26.6)
Welch t test	0.038	(0.8, 26.8)
Rank Sum test	0.075	(-1, 24) [not means]
Bootstrap CI	$p < 0.050$	(0.7, 27.1) via `bootdif`
Bootstrap CI	$p < 0.050$	(2.1, 26.3) via `slipper`

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

On Reporting *p* Values

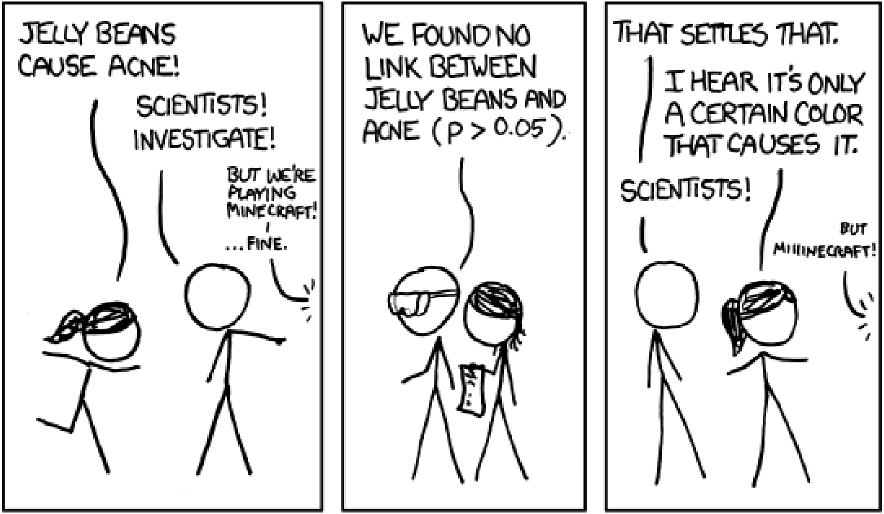
When reporting a *p* value and no rounding rules are in place from the lead author/journal/source for publication, follow these conventions...

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 $p = 1$.

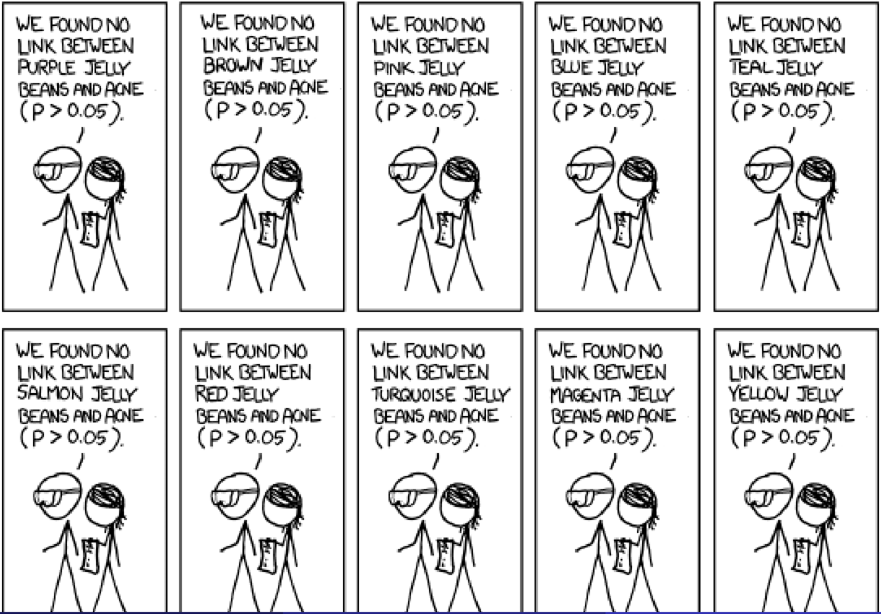
A Few Comments on Significance

- **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.

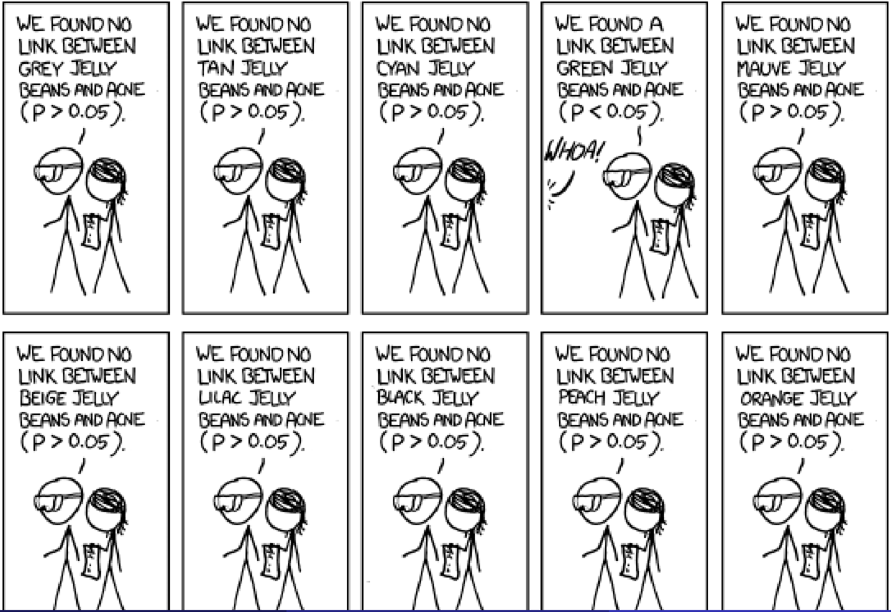
From XKCD (https://xkcd.com/882/)


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE ($P > 0.05$).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!

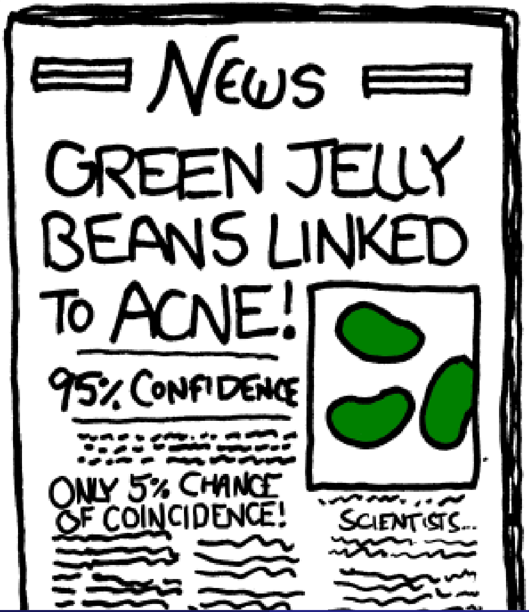
From XKCD (https://xkcd.com/882/)


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).

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


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

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



NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

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

Power and Sample Size Considerations: Getting Started

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?

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 $\beta = 0.20$

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]
$\alpha =$ 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

```r
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` package 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
```

What haven't I included here?

1. Some people will drop out.
2. What am I going to do about missingness?
3. And what if I want to compare something other than two means?
4. What if I want to do my comparison, adjusting for a covariate?

More to come.

Next Time

Comparing Rates and Proportions in 2x2 Tables