431 Class 19

Thomas E. Love

2018-11-06

Today's Agenda

- Comparing Rates/Proportions
- The `tabyl` function in the `janitor` package
- Analyzing a 2x2 Cross-Tabulation
- Power and Sample Size When Comparing Proportions
- The In-Class Survey (from Class 18)

Today's R Setup

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

library(Hmisc); library(pwr); library(broom); library(Epi)
library(magrittr); library(janitor) # new and "exciting"
library(tidyverse) # always load tidyverse last

dm192 <- read.csv("data/dm192.csv") %>% tbl_df()
class18a <- read.csv("data/class18a.csv") %>% tbl_df
class18b <- read.csv("data/class18b.csv") %>% tbl_df
```

Comparing Rates/Proportions

Comparing Two Proportions

Quinnipiac U. poll December 16-20, 2015 of 1,140 registered U.S. voters

- **Would you support or oppose a law requiring background checks on people buying guns at gun shows or online?**
- **Do you personally own a gun or does someone else in your household own a gun?**

Reported summaries of that poll get me to the following table:

–	Support Law	Oppose Law	*Total*
No Gun	542	24	*566*
Gun Household	440	73	*513*
Total	*982*	*97*	*1,079*

- Links to sources: fivethirtyeight and pollingreport

2 x 2 Table of Guns and Support, Prob. Difference

–	Support	Oppose	*Total*
No Gun in HH	542	24	*566*
Gun Household	440	73	*513*
Total	*982*	*97*	*1,079*

- Of those living in a no gun household, $542/566 = 95.8\%$ support universal background checks.
- Of those living in a gun household, $440/513 = 85.8\%$ support universal background checks.
- So the sample shows a difference of 10 percentage points, or a difference of 0.10 in proportions

Can we build a confidence interval for the population difference in those two proportions?

2 x 2 Table of Guns and Support, Relative Risk

–	Support	Oppose	*Total*
No Gun in HH	542	24	*566*
Gun Household	440	73	*513*
Total	*982*	*97*	*1,079*

- Pr(support | no gun in HH) $= 542/566 = 0.958$
- Pr(support | gun in HH) $= 440/513 = 0.858$
- The ratio of those two probabilities (risks) is $.958/.858 = 1.12$

Can we build a confidence interval for the relative risk of support in the population given no gun as compared to gun?

2 x 2 Table of Guns and Support, Odds Ratio

–	Support	Oppose	*Total*
No Gun in HH	542	24	*566*
Gun Household	440	73	*513*
Total	*982*	*97*	*1,079*

- Odds = Probability / (1 - Probability)
- Odds of Support if No Gun in HH $= \frac{542/566}{1-(542/566)} = 22.583333$
- Odds of Support if Gun in HH $= \frac{440/513}{1-(440/513)} = 6.027397$
- Ratio of these two Odds are 3.75

In a 2x2 table, odds ratio = cross-product ratio.

- Here, the cross-product estimate $= \frac{542*73}{440*24} = 3.75$

Can we build a confidence interval for the odds ratio for support in the population given no gun as compared to gun?

2x2 Table Results in R

```
twobytwo(542, 24, 440, 73,
      "No Gun in HH", "Gun Household", "Support", "Oppose")
```

This `twobytwo` function is part of the `Love-boost.R` script we sourced in earlier. Without that, this will throw an error message.

Full Output

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : Support
Comparing : No Gun in HH vs. Gun Household

              Support Oppose  P(Support)  95% conf. int.
No Gun in HH      542     24     0.9576   0.9375   0.9714
Gun Household     440     73     0.8577   0.8247   0.8853

                                    95% conf. interval
             Relative Risk: 1.1165    1.0735   1.1612
         Sample Odds Ratio: 3.7468    2.3230   6.0431
Conditional MLE Odds Ratio: 3.7424    2.2867   6.3174
    Probability difference: 0.0999    0.0659   0.1355

Exact P-value: 0                Asymptotic P-value: 0
------------------------------------------------------
```

Bayesian Augmentation in a 2x2 Table?

Original command:

```
twobytwo(542, 24, 440, 73,
      "No Gun in HH", "Gun Household", "Support", "Oppose")
```

Bayesian augmentation approach (add a success and add a failure in each row):

```
twobytwo(542+1, 24+1, 440+1, 73+1,
      "No Gun in HH", "Gun Household", "Support", "Oppose")
```

Full Output with Bayesian augmentation

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : Support
Comparing : No Gun in HH vs. Gun Household

             Support Oppose  P(Support) 95% conf. int.
No Gun in HH     543     25      0.9560  0.9357  0.9701
Gun Household    441     74      0.8563  0.8233  0.8840

                                   95% conf. interval
             Relative Risk: 1.1164    1.0731   1.1614
         Sample Odds Ratio: 3.6446    2.2768   5.8342
Conditional MLE Odds Ratio: 3.6405    2.2413   6.0875
    Probability difference: 0.0997    0.0655   0.1355

Exact P-value: 0               Asymptotic P-value: 0
------------------------------------------------------
```

Using a data frame, rather than a 2x2 table

For example, in the `dm192` data, suppose we want to know whether statin prescriptions are more common among male patients than female patients. So, we want a two-way table with "Male", "Statin" in the top left.

```
dm192 %$% table(sex, statin)
```

```
        statin
sex       0  1
  female 24 74
  male   21 73
```

So we want male in the top row and statin yes in the left column...

Rebuilding the data frame

```
dm192 <- dm192 %>%
  mutate(sex_f = fct_relevel(sex, "male"),
         statin_f = fct_recode(factor(statin),
                               on_statin = "1", no_statin = "0"),
         statin_f = fct_relevel(statin_f, "on_statin"))

dm192 %$% table(sex_f, statin_f)
```

```
        statin_f
sex_f    on_statin no_statin
  male          73        21
  female        74        24
```

Using tabyl from janitor

```
t1 <- dm192 %>% tabyl(sex_f, statin_f)

t1
```

```
  sex_f on_statin no_statin
   male        73        21
 female        74        24
```

```
class(t1)
```

```
[1] "tabyl"      "data.frame"
```

"Adorning" the tabyl

```
dm192 %>% tabyl(sex_f, statin_f) %>%
  adorn_totals() %>%
  adorn_percentages("row") %>%
  adorn_pct_formatting(digits = 1) %>%
  adorn_ns(position = "front") %>%
  adorn_title(row = "Sex", col = "Statin Status") %>%
  knitr::kable(align = "rr", caption = "dm192 statin by sex")
```

Table 5: dm192 statin by sex

	Statin Status	
Sex	on_statin	no_statin
male	73 (77.7%)	21 (22.3%)
female	74 (75.5%)	24 (24.5%)
Total	147 (76.6%)	45 (23.4%)

Running twoby2 against a data set

The twoby2 function from the Epi package can operate with tables (but not, alas, tabyls) generated from data.

```
twoby2(dm192 %$% table(sex_f, statin_f))
```

(edited output on next slide)

With Bayesian Augmentation

```
twoby2(dm192 %$% table(sex_f, statin_f) + 1)
```

(edited output on the slide after that)

twoby2 output on Raw Data (No Augmentation)

```
2 by 2 table analysis:
------------------------------------------------------
Outcome: on_statin          Comparing: male vs. female

       on_statin no_statin P(on_statin) 95% conf. interval
male          73        21       0.7766    0.6815   0.8496
female        74        24       0.7551    0.6605   0.8301

                                    95% conf. interval
             Relative Risk: 1.0285    0.8795   1.2026
         Sample Odds Ratio: 1.1274    0.5775   2.2010
Conditional MLE Odds Ratio: 1.1267    0.5473   2.3330
    Probability difference: 0.0215   -0.0985   0.1399

P-values:        Exact: 0.7368      Asymptotic: 0.7253
------------------------------------------------------
```

twoby2 WITH Bayesian Augmentation

```
2 by 2 table analysis:
------------------------------------------------------
Outcome: on_statin          Comparing: male vs. female

       on_statin no_statin P(on_statin) 95% conf. interval
male          74        22       0.7708    0.6764   0.8441
female        75        25       0.7500    0.6561   0.8251

                                    95% conf. interval
             Relative Risk: 1.0278    0.8783   1.2027
         Sample Odds Ratio: 1.1212    0.5814   2.1624
Conditional MLE Odds Ratio: 1.1206    0.5520   2.2869
    Probability difference: 0.0208   -0.0988   0.1389

P-values:        Exact: 0.7414      Asymptotic: 0.7328
------------------------------------------------------
```

Power and Sample Size When Comparing Proportions

Relation of α and β to Error Types

Recall the meanings of 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)

Tuberculosis Prevalence Among IV Drug Users

Here, we investigate factors affecting tuberculosis prevalence among intravenous drug users.

Among 97 individuals who admit to sharing needles, 24 (24.7%) had a positive tuberculin skin test result; among 161 drug users who deny sharing needles, 28 (17.4%) had a positive test result.

What does the 2x2 table look like?

Tuberculosis Prevalence Among IV Drug Users

Here, we investigate factors affecting tuberculosis prevalence among intravenous drug users.

Among 97 individuals who admit to sharing needles, 24 (24.7%) had a positive tuberculin skin test result; among 161 drug users who deny sharing needles, 28 (17.4%) had a positive test result.

The 2x2 Table is...

```
         TB+   TB-
share     24    73
don't     28   133
```

- rows describe needle sharing, columns describe TB test result
- row 1 people who share needles: 24 TB+, and 97-24 = 73 TB-
- row 2 people who don't share: 28 TB+ and 161-28 = 133 TB-

twobytwo (with Bayesian Augmentation)

To start, we'll test the null hypothesis that the population proportions of intravenous drug users who have a positive tuberculin skin test result are identical for those who share needles and those who do not.

$$H_0 : \pi_{share} = \pi_{donotshare}$$

$$H_A : \pi_{share} \neq \pi_{donotshare}$$

We'll use the Bayesian augmentation.

```
twobytwo(24+1, 73+1, 28+1, 133+1,
         "Sharing", "Not Sharing",
         "TB test+", "TB test-")
```

Two-by-Two Table Result

```
Outcome   : TB test+
Comparing : Sharing vs. Not Sharing

          TB test+ TB test- P(TB test+) 95% conf. int.
Sharing         25       74      0.2525  0.1767 0.3471
Not Sharing     29      134      0.1779  0.1265 0.2443

                                     95% conf. interval
             Relative Risk: 1.4194    0.8844   2.2779
         Sample Odds Ratio: 1.5610    0.8520   2.8603
Conditional MLE Odds Ratio: 1.5582    0.8105   2.9844
    Probability difference: 0.0746   -0.0254   0.1814

Exact P-value: 0.1588      Asymptotic P-value: 0.1495
```

What conclusions should we draw?

Designing a New TB Study

PI:

- OK. That's a nice pilot.
- We saw $p_{nonshare} = 0.18$ and $p_{share} = 0.25$ after your augmentation.
- Help me design a new study.
    - This time, let's have as many needle-sharers as non-sharers.
    - We should have 90% power to detect a difference as large as what we saw in the pilot, or larger, so a difference of 7 percentage points.
    - We'll use a two-sided test, and $\alpha = 0.05$, of course.

What sample size would be required to accomplish these aims?

How `power.prop.test` works

`power.prop.test` works much like the `power.t.test` we saw for means.

Again, we specify 4 of the following 5 elements of the comparison, and R calculates the fifth.

- The sample size (interpreted as the # in each group, so half the total sample size)
- The true probability in group 1
- The true probability in group 2
- The significance level (α)
- The power (1 - β)

The big weakness with the `power.prop.test` tool is that it doesn't allow you to work with unbalanced designs.

Using power.prop.test for Balanced Designs

To find the sample size for a two-sample comparison of proportions using a balanced design:

- we will use a two-sided test, with $\alpha = .05$, and power $= .90$,
- we estimate that non-sharers have probability .18 of positive tests,
- and we will try to detect a difference between this group and the needle sharers, who we estimate will have a probability of .25

R Command to find the required sample size

```
power.prop.test(p1 = .18, p2  = .25,
                alternative = "two.sided",
                sig.level = 0.05, power = 0.90)
```

Results: `power.prop.test` for Balanced Designs

```
power.prop.test(p1 = .18, p2  = .25,
                alternative = "two.sided",
                sig.level = 0.05, power = 0.90)
```

```
Two-sample comparison of proportions power calculation
n = 721.7534
p1 = 0.18, p2 = 0.25
sig.level = 0.05, power = 0.9, alternative = two.sided
NOTE: n is number in *each* group
```

So, we'd need at least 721 non-sharing subjects, and 721 more who share needles to accomplish the aims of the study, or a total of 1442 subjects.

Another Scenario

Suppose we can get 400 sharing and 400 non-sharing subjects. How much power would we have to detect a difference in the proportion of positive skin test results between the two groups that was identical to the data above or larger, using a *one-sided* test, with $\alpha = .10$?

```
power.prop.test(n=400, p1=.18, p2=.25, sig.level = 0.10,
                alternative="one.sided")
```

```
Two-sample comparison of proportions power calculation
n = 400, p1 = 0.18, p2 = 0.25
sig.level = 0.1, power = 0.8712338
alternative = one.sided
NOTE: n is number in *each* group
```

We would have just over 87% power to detect such an effect.

Using the pwr package to assess sample size for Unbalanced Designs

The `pwr.2p2n.test` function in the `pwr` package can help assess the power of a test to determine a particular effect size using an unbalanced design, where n_1 is not equal to n_2.

As before, we specify four of the following five elements of the comparison, and R calculates the fifth.

- `n1` = The sample size in group 1
- `n2` = The sample size in group 2
- `sig.level` = The significance level (α)
- `power` = The power (1 - β)
- `h` = the effect size h, which can be calculated separately in R based on the two proportions being compared: p_1 and p_2.

Calculating the Effect Size `h`

To calculate the effect size for a given set of proportions, use `ES.h(p1, p2)` which is available in the `pwr` package.

For instance, in our comparison, we have the following effect size.

```
ES.h(p1 = .18, p2 = .25)
```

```
[1] -0.1708995
```

Using `pwr.2p2n.test` in R

Suppose we can have 700 samples in group 1 (the not sharing group) but only 400 in group 2 (the group of users who share needles).

How much power would we have to detect this same difference (p1 = .18, p2 = .25) with a 5% significance level in a two-sided test?

R Command to find the resulting power

```
pwr.2p2n.test(h = ES.h(p1 = .18, p2 = .25),
              n1 = 700, n2 = 400, sig.level = 0.05)
```

Results of using `pwr.2p2n.test`

```
pwr.2p2n.test(h = ES.h(p1 = .18, p2 = .25),
              n1 = 700, n2 = 400, sig.level = 0.05)
```

```
difference of proportion power calculation
for binomial distribution (arcsine transformation)

h = 0.1708995, n1 = 700, n2 = 400
sig.level = 0.05, power = 0.7783562
alternative = two.sided
NOTE: different sample sizes
```

We will have about 78% power under these circumstances.

Comparison to Balanced Design

How does this compare to the results with a balanced design using 1100 drug users in total, i.e. with 550 patients in each group?

```
pwr.2p2n.test(h = ES.h(p1 = .18, p2 = .25),
              n1 = 550, n2 = 550, sig.level = 0.05)
```

which yields a power estimate of 0.809. Or we could instead have used...

```
power.prop.test(p1 = .18, p2 = .25, sig.level = 0.05,
                n = 550)
```

which yields an estimated power of 0.808.

Each approach uses approximations, and slightly different ones, so it's not surprising that the answers are similar, but not identical.

Exploring the In-Class Survey from Class 18

In-Class Survey

We chose (using a computer) a random number between 0 and 100.

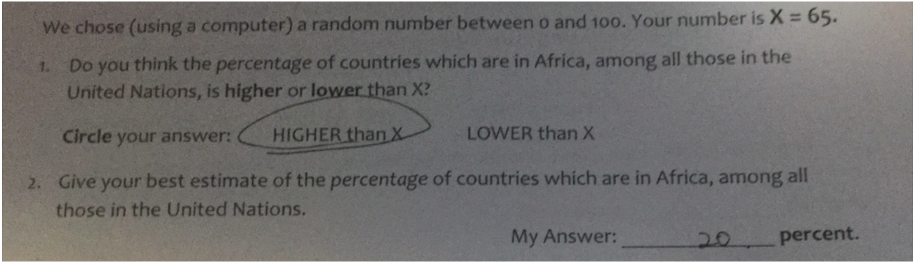
Your number is X = 10 (or 65).

1. Do you think the percentage of countries which are in Africa, among all those in the United Nations, is higher or lower than X?
2. Give your best estimate of the percentage of countries which are in Africa, among all those in the United Nations.

The facts

- There are 193 sovereign states that are members of the UN.
- The African regional group has 54 member states, so that's 28%.
- UN regions for countries found at this Wikipedia link.

A troubling situation



We chose (using a computer) a random number between 0 and 100. Your number is X = 65.

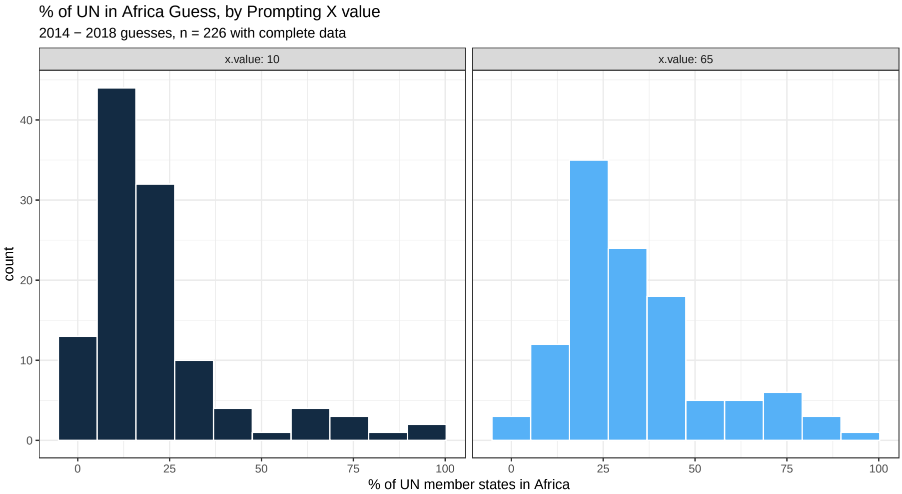
1. Do you think the *percentage* of countries which are in Africa, among all those in the United Nations, is **higher** or **lower** than X?

   Circle your answer: HIGHER than X LOWER than X

2. *Give your best estimate of the* *percentage* of countries which are in Africa, among all those in the United Nations.

   My Answer: ____20____ percent.

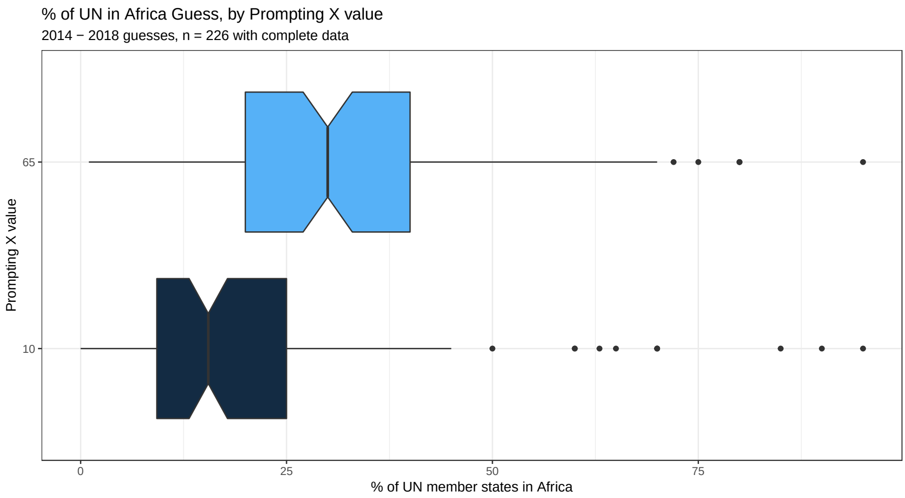
`class18a` Africa percentage guess by X = 10 or 65



`class18a` Analysis, Step-by-Step

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?

`class18a` Africa percentage guess by X = 10 or 65



class18a Descriptive Statistics

```
class18a %>%
  filter(complete.cases(africa.pct)) %>%
  group_by(x.value) %>%
  summarise(n = n(),
            mean = round(mean(africa.pct),2),
            sd = round(sd(africa.pct),2),
            median = median(africa.pct))
```

```
# A tibble: 2 x 5
  x.value     n  mean    sd median
    <int> <int> <dbl> <dbl>  <dbl>
1      10   114  22    18.8   15.5
2      65   112  33.6  19.0   30
```

class18a comparisons (results: next slide)

```
t.test(africa.pct ~ x.value,
       data = class18a) # Welch
t.test(africa.pct ~ x.value, data = class18a,
       var.equal = TRUE) # Pooled t
wilcox.test(africa.pct ~ x.value, conf.int = TRUE,
            data = class18a)
set.seed(43123)
bootdif(class18a$africa.pct, class18a$x.value)
```

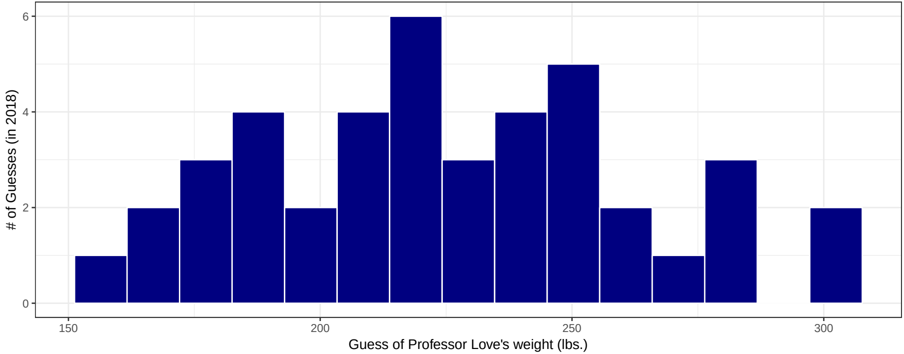
$$\Delta = \mu_{65} - \mu_{10}$$

Procedure	Est. Δ	95% CI for Δ	p
Welch t	11.6	(6.7, 16.6)	6.5e-06
Pooled t	11.6	(6.7, 16.6)	6.5e-06
Rank Sum	12.0	(8.0, 15.0)	2.9e-09
Bootstrap	11.6	(6.5, 16.4)	$<$ .05

Conclusions?

In-Class Survey (`class18b` data)

3. Provide a point estimate for Dr. Love's current weight (in pounds.)



- 2018 Weight Guesses: $n = 42$, $\bar{x} = 225.6$ lbs., $s = 36.6$ lbs.
- Five Number Summary: 154 200 220 250 300

50% and 90% "Intervals" from Group Estimates

4. Now estimate one interval, which you believe has a 50% chance of including Dr. Love's current weight (again, in pounds.) Then do the same for a 90% interval.

We have $n = 42$ independent guesses, with $\bar{x} = 225.6$ lbs., $s = 36.6$ lbs. Let's first obtain quantiles, and use the crowd's wisdom.

```
quantile(class18b$love.lbs,
  probs = c(0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95))
```

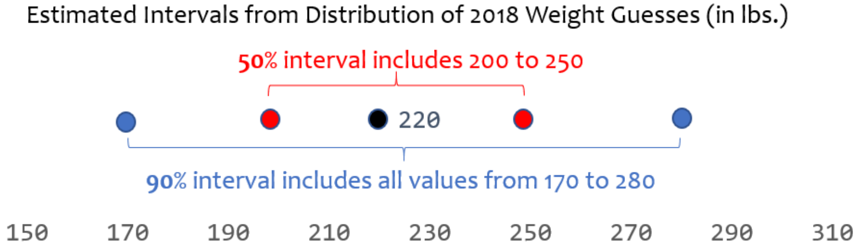
```
    5%    10%    25%    50%    75%    90%    95%
170.25 177.30 200.00 220.00 250.00 279.00 280.00
```

- What's a rational 50% interval for estimating my weight?
- How about a 90% interval?

One Possible, Rational, Set of Intervals

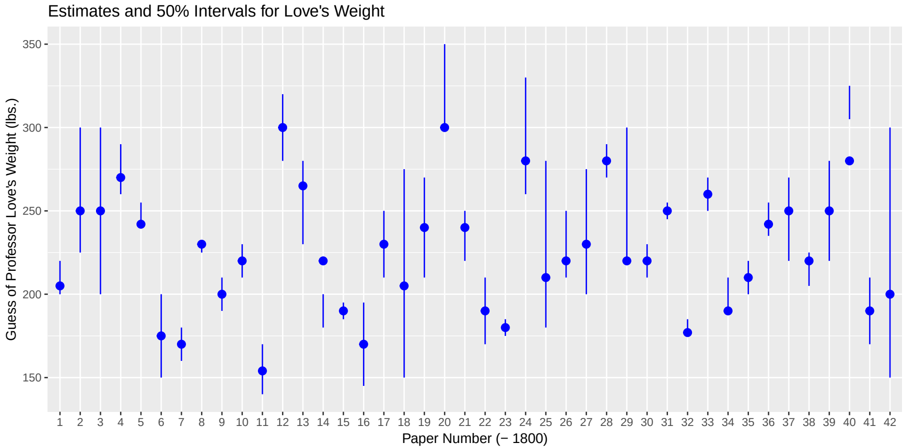
Suppose my estimate is 220 pounds.

- Then suppose I assign probability 0.50 to the interval (200, 250)
- And suppose I assign probability 0.90 to the interval (170, 280)



In-Class Survey (`class18b` data)

4a. Now estimate one interval, which you believe has a 50% chance of including Dr. Love's current weight (again, in pounds.)



In-Class Survey (`class18b` data)

4b. Now do the same, but for a 90% interval...

Some Troubling Selections (prior classes), 1

3. Provide a point estimate for Dr. Love's current weight (in pounds.) If you think in kilograms, multiply kg by 2.2 to get pounds.

My Answer: 260 pounds.

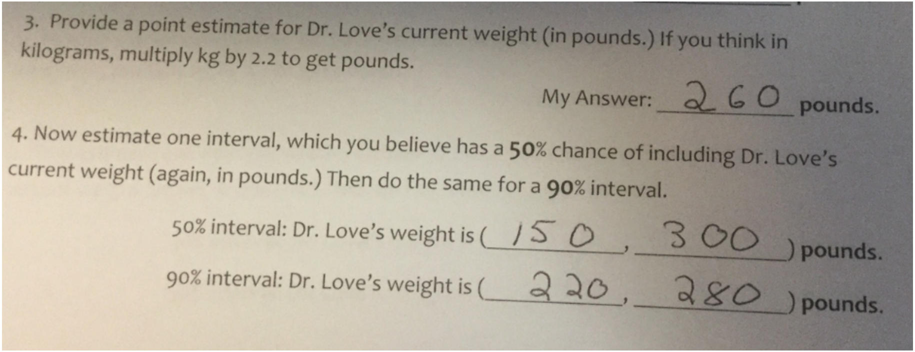
4. Now estimate one interval, which you believe has a **50%** chance of including Dr. Love's current weight (again, in pounds.) Then do the same for a **90%** interval.

50% interval: Dr. Love's weight is (150 , 300) pounds.

90% interval: Dr. Love's weight is (220 , 280) pounds.

- Why does this set of intervals not make sense?

Some Troubling Selections (prior classes), 1


3. Provide a point estimate for Dr. Love's current weight (in pounds.) If you think in kilograms, multiply kg by 2.2 to get pounds.

My Answer: 260 pounds.

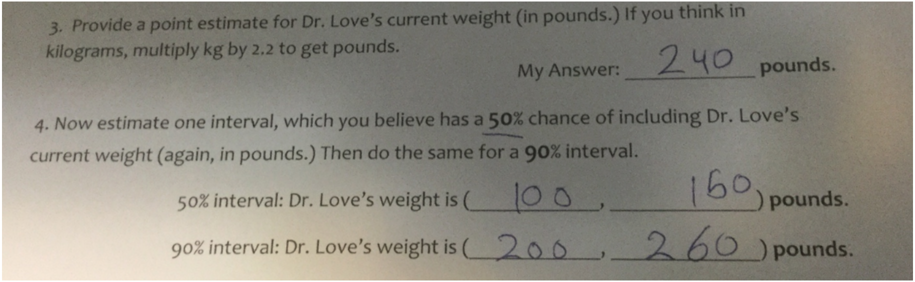
4. Now estimate one interval, which you believe has a **50%** chance of including Dr. Love's current weight (again, in pounds.) Then do the same for a **90%** interval.

50% interval: Dr. Love's weight is (150 , 300) pounds.

90% interval: Dr. Love's weight is (220 , 280) pounds.

- Why does this set of intervals not make sense?
- There were **8** students (out of 44) who had a wider 50% interval than 90% interval.

Some Troubling Selections (prior classes), 2



3. Provide a point estimate for Dr. Love's current weight (in pounds.) If you think in kilograms, multiply kg by 2.2 to get pounds.

My Answer: 240 pounds.

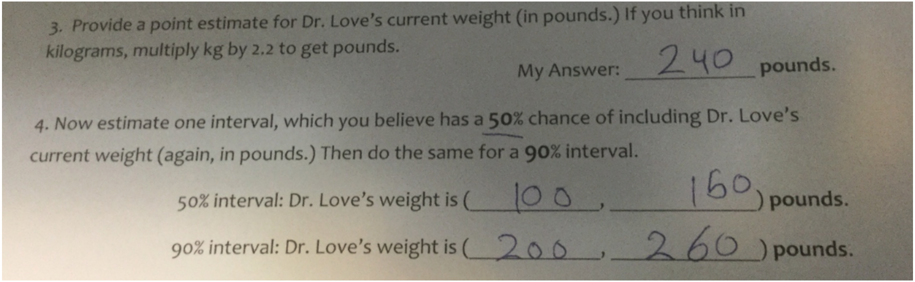
4. Now estimate one interval, which you believe has a **50%** chance of including Dr. Love's current weight (again, in pounds.) Then do the same for a **90%** interval.

50% interval: Dr. Love's weight is (100 , 160) **pounds.**

90% interval: Dr. Love's weight is (200 , 260) **pounds.**

- It wasn't clear enough that the interval estimate was meant to surround the point estimate.

Some Troubling Selections (prior classes), 2



3. Provide a point estimate for Dr. Love's current weight (in pounds.) If you think in kilograms, multiply kg by 2.2 to get pounds.

My Answer: 240 pounds.

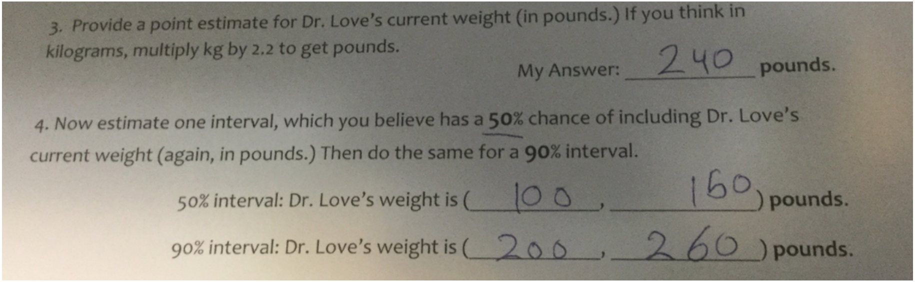
4. Now estimate one interval, which you believe has a **50%** chance of including Dr. Love's current weight (again, in pounds.) Then do the same for a **90%** interval.

50% interval: Dr. Love's weight is (100 , 160) **pounds.**

90% interval: Dr. Love's weight is (200 , 260) **pounds.**

- It wasn't clear enough that the interval estimate was meant to surround the point estimate.
- There were **2** students out of 42 with this problem in their 50% interval, **0** in their 90% interval.

Some Troubling Selections (prior classes), 2


3. Provide a point estimate for Dr. Love's current weight (in pounds.) If you think in kilograms, multiply kg by 2.2 to get pounds.

My Answer: 240 pounds.

4. Now estimate one interval, which you believe has a **50%** chance of including Dr. Love's current weight (again, in pounds.) Then do the same for a **90%** interval.

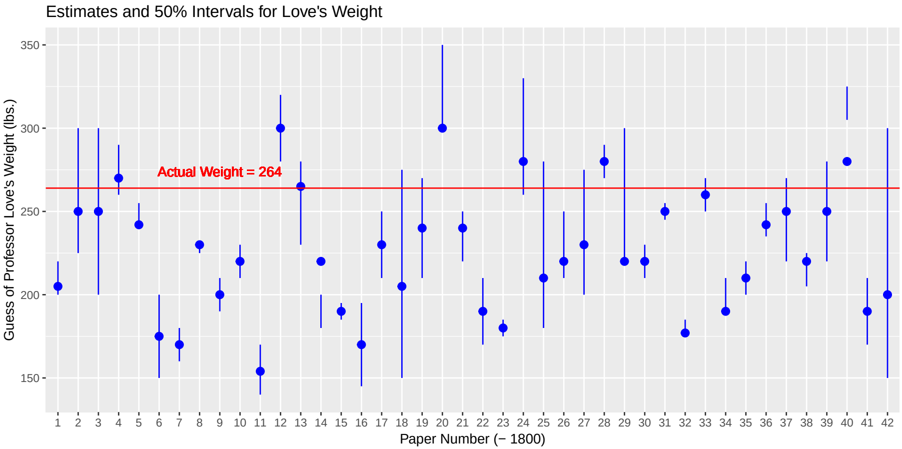
50% interval: Dr. Love's weight is (100 , 160) pounds.

90% interval: Dr. Love's weight is (200 , 260) pounds.

- It wasn't clear enough that the interval estimate was meant to surround the point estimate.
- There were **2** students out of 42 with this problem in their 50% interval, **0** in their 90% interval.
- For **8** students, the 90% interval did not contain the 50% interval.

The facts (with 50% intervals)

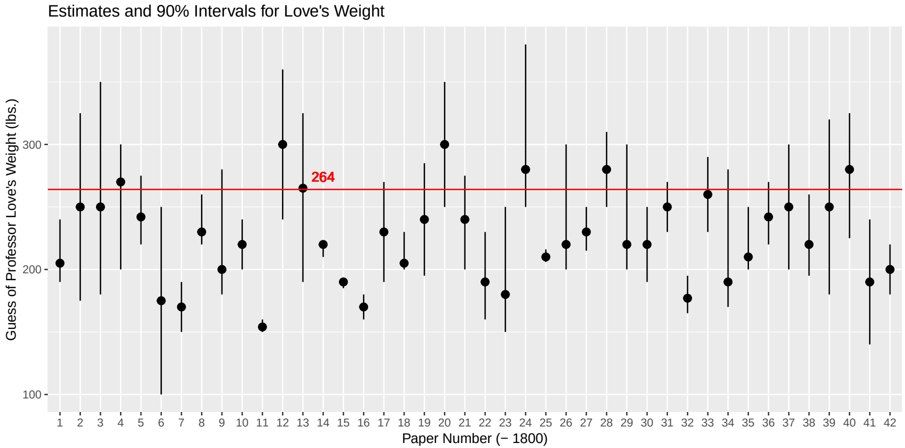
On 2018-11-01, Dr. Love actually weighed 264 lbs. or 119.7 kg or 18 stone and 12 pounds, dressed but without shoes.



- **14** of the 42 50% intervals estimated by students included 264 lbs.

The facts (with 90% intervals)

On 2018-11-01, Dr. Love actually weighed 264 lbs.



- **22** of the 42 90% intervals (or 52.3% of them) included 264 lbs.

Coming Up . . .

- Larger Cross-Tabulations
- Three-Way Tables
- Comparing Three or More Population Means with ANOVA

Please don't forget to do the Minute Paper after Class 19.