

# 431 Class 18

Thomas E. Love

2018-11-01

# Today's Agenda

- A little survey
- Finishing off power stuff for comparing 2 means
- Comparing Rates/Proportions
- The `taby1` function in the `janitor` package
- Analyzing a 2x2 Cross-Tabulation
- Power and Sample Size concerns

# 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.

# Today's R Setup

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

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

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

## Returning to the Satiety Example

# A Small Example: Studying Satiety

- I want to compare people eating meal A to people eating meal B in terms of impact on satiety, as measured on a 0-100 scale. I'm interested in a two-sided test, since I don't know which will be better in advance.
- I can afford to enroll 160 people (or, more specifically, prepare 160 meals) in the study.
- I expect that a difference that's important will be about 10 points on the satiety scale.
  - Perhaps this is because I saw a **17** point difference in a pilot study.
- I don't know the standard deviation, but the whole range (0-100) gets used, so I'll estimate the SD conservatively with  $(\text{range}/4)$  or 25.

## The Key Questions

- 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?

# Using power.t.test

Measure	Paired Samples	Independent Samples
type =	"paired"	"two.sample"
$n$	# of paired diffs	# in each sample
$\delta$	true mean of diffs	true diff in means
$s = sd$	true SD of diffs	true SD, either group <sup>1</sup>
$\alpha = \text{sig.level}$	max. Type I error rate	Same as paired.
$1 - \beta = \text{power}$	power to detect effect $\delta$	Same as paired.

# Satiety: Assumption Set 1

Let's assume:

- Independent Samples, so test type = "two.sample"
- Total sample size I can afford = 160, so a balanced design gives 80 in each group.
- $\delta$  = smallest difference I want to be sure I detect = 10 points.
- standard deviation of satiety scores is unknown, but we'll guess 25 (since range/4 = 25).
- we'll use alpha = .10 (90% confidence) and a two-sided test.



# Satiety: Assumption Set 1 (Results)

```
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

# Satiety: Assumption Set 2

What if we use a paired samples design instead?

- Paired samples, so test type = “paired”
- Total = 160 meals, so that's  $n = 80$  paired differences (each subject eats A and B, in a random order)
- standard deviation of *differences* in satiety will be smaller than 25, let's be very conservative and say 20.
- $\delta$  remains 10 points, two-sided test with 90% confidence.

## Satiety: Assumption Set 2 (Results)

```
power.t.test(n = 80, delta = 10, sd = 20,  
             sig.level = 0.10, alt = "two.sided",  
             type = "paired")
```

Paired t test power calculation

```
      n = 80  
delta = 10  
      sd = 20  
sig.level = 0.1  
      power = 0.9973528  
alternative = two.sided
```

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

**Note:** If sd = 25 with paired samples, power = 0.971.

# What happens if you change something else?

If we start with  $n = 80$ ,  $\delta = 10$ ,  $sd = 25$ ,  $\alpha = 0.10$ , that yields power 0.809.

To increase the power to 0.90, what can we do?

- ① Increase the sample size,  $n$  to 108 in each group
  - If  $n = 108$  instead, power is now 0.90
- ② or Increase the minimum detectable effect size  $\delta$  to 11.7
- ③ or Reduce the standard deviation to 21.5
- ④ or Increase  $\alpha$  (willingness to tolerate Type I error) to 0.215
  - So confidence level would 79.5% instead of 90%.
- ⑤ or Switch from independent to matched/paired samples.
  - Even if we leave  $sd$  alone, 55 matched differences yields 90% power

# 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) is `pwr.t2n.test` from the `pwr` package.

- 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  $n_1 = 40$ ,  $d = 10/25$  ( $\delta / sd$ ),  $\text{sig.level} = 0.1$  and  $\text{power} = 0.8$

- 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

```
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 while adjusting for a covariate?
- ⑤ Lots of arbitrary decisions here. How do I improve the decisions I make?

More to come on all of this, later.



# Comparing Rates/Proportions

# Moving on from Means to Proportions

We've focused on creating statistical inferences about a population mean, or difference between means, where we care about a quantitative outcome. Now, we'll tackle **categorical** outcomes.

- 1 We'll start by estimating a confidence interval around an unknown population proportion, or rate, which we'll symbolize with  $\pi$ , on the basis of a random sample of  $n$  observations from a sample which yields a proportion of  $\hat{p}$ , which is sometimes, unfortunately, symbolized as  $p$ . Note that this  $\hat{p}$  is the sample proportion - not a  $p$  value.
- 2 Then we'll look at comparing proportions  $\pi_1$  and  $\pi_2$  - comparisons across two populations, based on samples of size  $n_1$  and  $n_2$ .

# Firearm Laws should be more/less strict?

A Gallup Poll (2018-10-01 to 10-10) of 1,019 adults nationwide asked:

*In general, do you feel that the laws covering the sale of firearms should be made more strict, less strict, or kept as they are now?*

–	More Strict	Less Strict	Kept as now	Unsure
%	61	8	30	2

- 1 What can we conclude from this poll about the true percentage of registered U.S voters who would answer “More Strict”?
  - 2 The poll lists a “margin of error” of 4 percentage points. What does this mean?
- My source: <http://www.pollingreport.com/guns.htm>
  - Gallup’s graph, details: <https://news.gallup.com/poll/1645/guns.aspx>

# A Confidence Interval for a Proportion

A  $100(1-\alpha)\%$  confidence interval for the population proportion  $\pi$  can be created by using the standard normal distribution, the sample proportion,  $\hat{p}$ , and the standard error of a sample proportion, which is defined as the square root of  $\hat{p}$  multiplied by  $(1 - \hat{p})$  divided by the sample size,  $n$ .

Specifically, our confidence interval is  $\hat{p} \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

where  $Z_{\alpha/2}$  = the value from a standard Normal distribution cutting off the top  $\alpha/2$  of the distribution, obtained in R by substituting the desired  $\alpha/2$  value into: `qnorm(alpha/2, lower.tail=FALSE)`.

- *Note:* This interval is reasonably accurate so long as  $n\hat{p}$  and  $n(1 - \hat{p})$  are each at least 5.

# Estimating $\pi$ in “More Strict Firearms Laws” Poll

- We'll build a 95% confidence interval for the true population proportion, so  $\alpha = 0.05$
- We have  $n = 1,019$  subjects who responded
- Sample proportion saying “more guns” is  $\hat{p} = 0.61$ ; we'll assume that  $(1019)(0.61) = 622$  actually said this.

The standard error of that sample proportion will be

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.61(1-0.61)}{1019}} = 0.015$$

# Confidence Interval for $\pi$ in “More Guns”

Our 95% confidence interval for the true population proportion,  $\pi$ , of voters who would choose “more guns” is

$$\hat{p} \pm Z_{0.025} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

or  $0.61 \pm 1.96(0.015) = 0.61 \pm 0.029$ , or  $(0.581, 0.639)$

I simply recalled from our prior work that  $Z_{0.025} = 1.96$ , but we can verify this:

```
qnorm(0.025, lower.tail=FALSE)
```

```
[1] 1.959964
```

# Likely Accuracy of this Confidence Interval?

Since  $n\hat{p} = (1019)(0.61) = 622$  and  $n(1 - \hat{p}) = (1019)(1 - 0.61) = 397$  are substantially greater than 5, the CI should be reasonably accurate.

- ❶ What can we conclude from this poll about the true percentage of registered U.S voters who would answer “More Strict”?
  - Our best point estimate of the true population proportion who would say “more strict” is 0.61.
  - Our 95% confidence interval for the true population proportion is between 0.581 and 0.639.
- ❷ The poll lists a “margin of error” of 4 percentage points. What does this mean?
  - Our 95% confidence interval for  $\pi$  can also be expressed as  $0.61 \pm 0.029$ .
  - Why isn't it listed by Gallup as 3 points?

Happily, we don't have to do these calculations by hand ever again.

# R Methods to get a CI for a Population Proportion

I am aware of at least three different procedures for estimating a confidence interval for a population proportion using R. All have minor weaknesses: none is importantly different from the others in many practical situations.

- 1 The `prop.test` approach (also called the Wald test)

```
prop.test(x = 622, n = 1019)
```

- 2 The `binom.test` approach (Clopper and Pearson “exact” test)

```
binom.test(x = 622, n = 1019)
```

- 3 Building a confidence interval via a SAIFS procedure

```
saifs.ci(x = 622, n = 1019)
```



# The prop.test approach (Wald test)

The prop.test function estimates a confidence interval for  $\pi$ :

```
prop.test(x = 622, n = 1019)
```

```
1-sample proportions test with continuity  
correction
```

```
data: 622 out of 1019, null probability 0.5  
X-squared = 49.24, df = 1, p-value = 2.264e-12  
alternative hypothesis: true p is not equal to 0.5  
95 percent confidence interval:  
 0.5796038 0.6403581  
sample estimates:  
      p  
0.6104024
```

# binom.test (Clopper-Pearson “exact” test)

```
binom.test(x = 622, n = 1019)
```

Exact binomial test

data: 622 and 1019

number of successes = 622, number of trials =  
1019, p-value = 1.85e-12

alternative hypothesis: true probability of success is not equal to 0.5

95 percent confidence interval:

0.579686 0.640479

sample estimates:

probability of success

0.6104024

# Estimating a Rate More Accurately

Suppose you have some data involving  $n$  independent tries, with  $x$  successes. The most natural estimate of the “success rate” in the data is  $x / n$ .

But, strangely enough, it turns out this isn't an entirely satisfying estimator. Alan Agresti provides substantial motivation for the  $(x + 1)/(n + 2)$  estimate as an alternative<sup>1</sup>. This is sometimes called a *Bayesian augmentation*.

---

<sup>1</sup>This note comes largely from a May 15 2007 entry in Andrew Gelman's blog at <http://andrewgelman.com/2007/05/15>

# Use $(x + 1)/(n + 2)$ rather than $x/n$

- The big problem with  $x / n$  is that it estimates  $p = 0$  or  $p = 1$  when  $x = 0$  or  $x = n$ .
- It's also tricky to compute confidence intervals at these extremes, since the usual standard error for a proportion,  $\sqrt{np(1 - p)}$ , gives zero, which isn't quite right.
- $(x + 1)/(n + 2)$  is much cleaner, especially when you build a confidence interval for the rate.
- The only place where  $(x + 1)/(n + 2)$  will go wrong (as in the SAIFS approach) is if  $n$  is small and the true probability is very close to 0 or 1.
  - For example, if  $n = 10$ , and  $p$  is 1 in a million, then  $x$  will almost certainly be zero, and an estimate of  $1/12$  is much worse than the simple  $0/10$ .
  - However, how big a deal is this? If  $p$  might be 1 in a million, are you going to estimate it with an experiment using  $n = 10$ ?

# Practical Impact of Bayesian Augmentation

It is likely that the augmented  $(x + 1) / (n + 2)$  version yields more accurate estimates for the odds ratio or relative risk or probability difference, but the two sets of estimates (with and without the augmentation) will be generally comparable, so long as...

- a. the sample size in each exposure group is more than, say, 30 subjects, and/or
- b. the sample probability of the outcome is between 0.1 and 0.9 in each exposure group.

# Bayesian Augmentation: Add a Success and a Failure

You'll get slightly better results if you use  $\frac{x+1}{n+2}$  rather than  $\frac{x}{n}$  as your point estimate, and to fuel your confidence interval using either the `binom.test` or `prop.test` approach.

- The results will be better in the sense that they'll be slightly more likely to meet the nominal coverage probability of the confidence intervals.
- This won't make a meaningful difference if  $\frac{x}{n}$  is near 0.5, or if the sample size  $n$  is large. Why?

Suppose you want to find a confidence interval when you have 2 successes in 10 trials. I'm suggesting that instead of `binom.test(x = 2, n = 10)` you might want to try `binom.test(x = 3, n = 12)`

# SAIFS confidence interval procedure

SAIFS = single augmentation with an imaginary failure or success<sup>2</sup>

- Uses a function I built in R for you (Part of Love-boost.R)

```
saifs.ci(x = 622, n = 1019)
```

Sample Proportion	0.025	0.975
0.610	0.580	0.641

`saifs.ci` already builds in a Bayesian augmentation, so we don't need to do that here.

---

<sup>2</sup>see Notes Part B for more details.

## Results for “More Guns” Rate ( $x = 622$ , $n = 1019$ )

Method	95% CI for $\pi$
prop.test	0.580, 0.640
binom.test	0.580, 0.640
saifs.ci	0.580, 0.641

Our “by hand” result, based on the Normal distribution, with no continuity correction, was (0.581, 0.639).

So in this case, it really doesn't matter which one you choose. With a smaller sample, we may not come to the same conclusion about the relative merits of these different approaches.



# Assumptions behind Inferences about $\pi$

We are making the following assumptions, when using these inferential approaches:

- 1 There are  $n$  identical trials.
- 2 There are exactly two possible outcomes (which may be designated as success and failure) for each trial.
- 3 The true probability of success,  $\pi$ , remains constant across trials.
- 4 Each trial is independent of all of the other trials.

## Accuracy of these Inferences about a Proportion

We'd like to see that both  $n\hat{p}$  = observed successes and  $n(1 - \hat{p})$  = observed failures exceed 5.

- If not, then the intervals may be both incorrect (in the sense of being shifted away from the true value of  $\pi$ ), and also less efficient (wider) than necessary.

# None of these approaches is always best

When we have a sample size below 100, or the sample proportion of success is either below 0.10 or above 0.90, caution is warranted<sup>3</sup>, although in many cases, the various methods give similar responses.

95% CI Approach	Wald	Clopper-Pearson	SAIFS
$X = 10, n = 30$	0.179, 0.529	0.173, 0.528	0.148, 0.534
$X = 10, n = 50$	0.105, 0.341	0.1, 0.337	0.083, 0.333
$X = 90, n = 100$	0.82, 0.948	0.824, 0.951	0.829, 0.96
$X = 95, n = 100$	0.882, 0.981	0.887, 0.984	0.894, 0.994

<sup>3</sup>These are great times for the Bayesian augmentation, for 'prop.test' or 'binom.test'

# Hypothesis Testing About a Population Proportion

To perform a hypothesis test about a population proportion, we'll usually use the `prop.test` or `binom.test` approaches in R<sup>4</sup>.

- The null hypothesis is that the population proportion is equal to some pre-specified value. Often, this is taken to be 0.5, but it can be any value, called  $\pi_0$ , that is between 0 and 1.
- The alternative hypothesis may be one-sided or two-sided. If it is two-sided, it will be that the population proportion is not equal to the value  $\pi_0$  specified by the null hypothesis.
- In the two-sided case, we have  $H_0 : \pi = \pi_0$  and  $H_A : \pi \neq \pi_0$
- In the one-sided “greater than” case, we have  $H_0 : \pi \leq \pi_0$  and  $H_A : \pi > \pi_0$

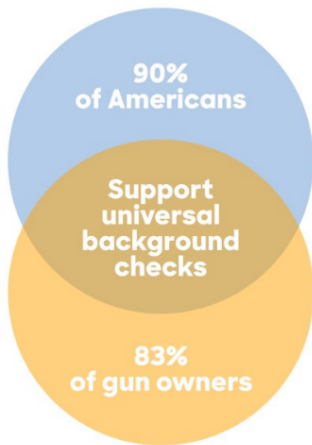
But, as usual, the focus is usually on the confidence intervals...

---

<sup>4</sup>Bayesian augmentation is helpful here, too.

# Comparing Proportions

# A Troublesome Tweet (2016-05-20)



**Hillary Clinton** ✓  
@HillaryClinton

 Follow

Dear Congress,

Let's get this done.

Thanks

# 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	<i>Total</i>
No Gun	542	24	566
Gun Household	440	73	513
<i>Total</i>	982	97	1,079

- Links to sources: [fivethirtyeight](#) and [pollingreport](#)
- Source Images on next two slides

# Polling Topline Images, from fivethirtyeight.com

67. Would you support or oppose a law requiring background checks on people buying guns at gun shows or online?

							COLLEGE DEG	
	Tot	Rep	Dem	Ind	Men	Wom	Yes	No
Support	89%	87%	95%	86%	84%	94%	92%	88%
Oppose	9	12	5	12	14	5	7	10
DK/NA	2	1	-	2	2	1	1	2
	AGE IN YRS.....				Gun	DENSITY.....		
	18-34	35-49	50-64	65+	HsHld	Urban	Suburb	Rural
Support	98%	87%	87%	90%	84%	89%	91%	88%
Oppose	2	12	12	8	14	9	9	11
DK/NA	-	1	1	2	2	2	1	2

# Polling Topline Images, from [pollingreport.com](http://pollingreport.com)

**Quinnipiac University.** Dec. 16-20, 2015. N=1,140 registered voters nationwide. Margin of error  $\pm$  2.9.

**"Would you support or oppose a law requiring background checks on people buying guns at gun shows or online?"**

	<b>Support</b>	<b>Oppose</b>	<b>Unsure/ No answer</b>
	%	%	%
12/16-20/15	89	9	2
Republicans	87	12	1
Democrats	95	5	-
Independents	86	12	2
4/25-29/13	83	13	3

**"Do you personally own a gun or does someone else in your household own a gun?"**

	<b>Yes: Personally</b>	<b>Yes: Someone else</b>	<b>Yes: Both</b>	<b>No gun</b>	<b>Unsure/ No answer</b>
	%	%	%	%	%
12/16-20/15	25	12	9	50	4
Republicans	37	12	15	29	7
Democrats	13	11	1	74	1
Independents	28	11	12	46	4



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

—	Support	Oppose	<i>Total</i>
No Gun in HH	542	24	566
Gun Household	440	73	513
<i>Total</i>	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	<i>Total</i>
No Gun in HH	542	24	566
Gun Household	440	73	513
<i>Total</i>	982	97	1,079

- $\Pr(\text{support} \mid \text{no gun in HH}) = 542/566 = 0.958$
- $\Pr(\text{support} \mid \text{gun in HH}) = 440/513 = 0.858$
- 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")
```

2 by 2 table analysis:

-----  
Outcome : Support

Comparing : No Gun in HH vs. Gun Household

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

	interval
No Gun in HH	0.9714
Gun Household	0.8853

	95% conf. interval
Relative Risk: 1.1165	1.0735 1.1612

# 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 with % using row as denominator

```
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 9:** 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, `taby1s`) 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  
-----

## Returning (Briefly) to the Survey

# 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](#).



# Power and Sample Size When Comparing Proportions

# Relation of $\alpha$ and $\beta$ to Error Types

Recall the meanings of  $\alpha$  and  $\beta$ :

- $\alpha$  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.
- $\beta$  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 ( $1 - \beta$ )	Type I Error ( $\alpha$ )
Test Retains $H_0$	Type II Error ( $\beta$ )	Correct Decision ( $1 - \alpha$ )

# 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  $n$  in each group, so half the total sample size)
- The true probability in group 1
- The true probability in group 2
- The significance level ( $\alpha$ )
- The power ( $1 - \beta$ )

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

$n = 721.7534$

$p1 = 0.18, p2 = 0.25$

$\text{sig.level} = 0.05, \text{power} = 0.9, \text{alternative} = \text{two.sided}$

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 ( $\alpha$ )
- `power` = The power ( $1 - \beta$ )
- `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 ( $p_1 = .18$ ,  $p_2 = .25$ ) 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.



# Next Time

- Exploring the In-Class Survey Results
- Larger Cross-Tabulations
- Three-Way Tables