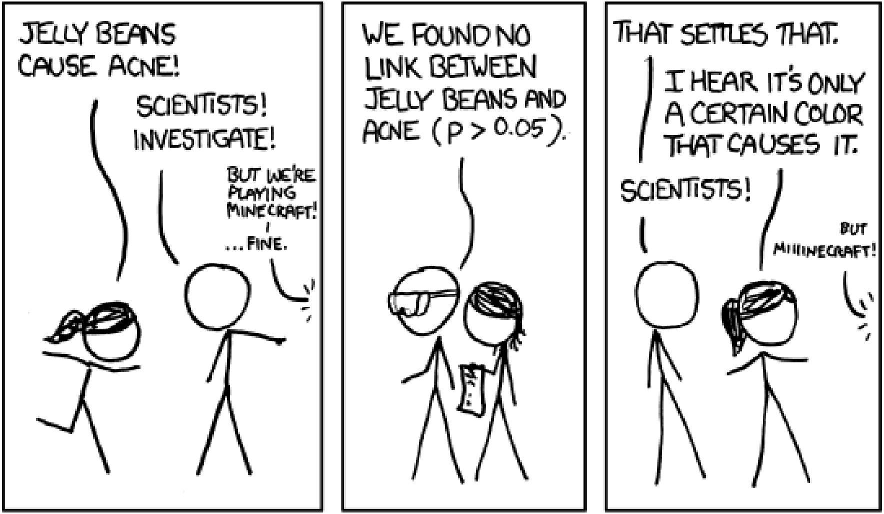
431 Class 17

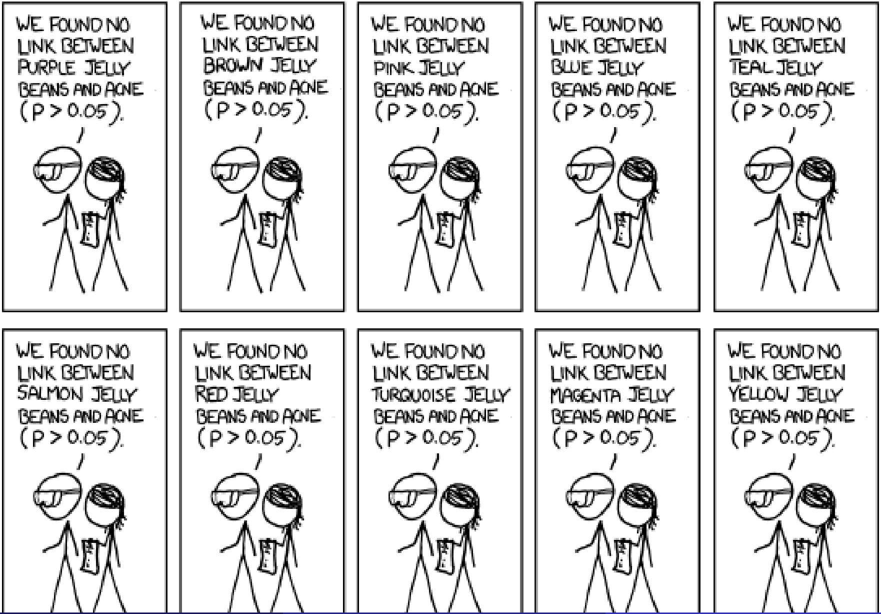
github.com/THOMASELOVE/2019-431

2019-10-29

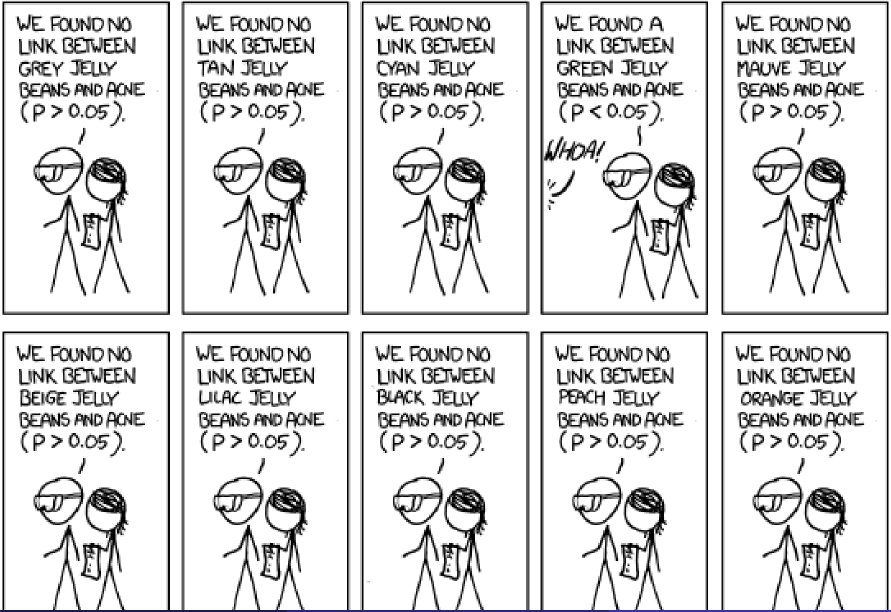
From XKCD (https://xkcd.com/882/)



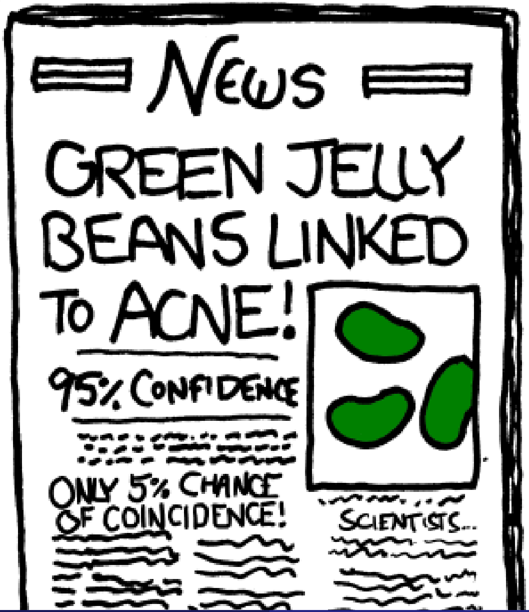
From XKCD (https://xkcd.com/882/)



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

Today's Setup and Data

```r
library(exact2x2); library(PropCIs) # new today

library(Epi)
library(magrittr); library(janitor)
library(here); library(tidyverse)

source(here("R", "Love-boost.R"))

dm431 <- readRDS(here("data", "dm431.Rds"))
```

Today: Comparing Population Proportions

We've focused on creating statistical inferences about differences between the means of populations, where we care about a quantitative outcome. Now, we'll tackle **categorical** outcomes, where we're interested in percentages, proportions or rates. We'll again do comparisons

- Using Independent Samples (see Chapter 21 of the Notes)
- Using Paired Samples (see Chapter 24)

We want to compare proportions π_1 and π_2 - comparisons across two populations, based on samples of size n_1 and n_2.

Comparing Proportions: Independent vs. Paired Samples

Goal: We want to use our data sampled from a population to make a comparison between the population proportion achieving our outcome who are [1] in exposure group 1 vs. [2] in exposure group 2.

1. The individual observations in exposure group 1 are not linked/matched to individual observations in exposure group 2. (Independent Samples)
2. Each individual observation in exposure group 1 is linked or matched to a specific observation in exposure group 2. (Paired Samples)

The determination as to whether the study design creates paired or independent samples can be determined without summarizing the data. It's a function of the design, not the responses.

A Polling Example

- 200 adult Ohio residents agreed to participate in a poll both two months ago and again today. Each of the 200 people met the polling organization's standards for a "likely voter in the 2020 Democratic presidential primary". 100 of those polled were under the age of 50 and the rest were 50 or older.
- In between the two polls, a major news event occurred which was relevant to Candidate X.

We asked them the same question at both times: "Are you considering voting for Candidate X?" We are interested in understanding what the data tell us about:

1. Were people under age 50 more likely to be considering Candidate X than people ages 50 and higher?
2. Were people more likely to be considering Candidate X after the news event than before?

Which of these uses *independent* samples, and which *paired* samples?

Comparing Proportions using Independent Samples (Course Notes Chapter 21)

`dm431` Example A.

Among our subjects with Medicaid insurance, compare the proportion with A1c below 8 among those who identify their race/ethnicity as Hispanic or Latinx with those who identify as non-Hispanic White.

```r
dm431 %>%
  filter(insurance == "Medicaid") %>%
  filter(race_eth %in% c("Hispanic or Latinx",
                          "Non-Hispanic White")) %>%
    count(a1c < 8, race_eth)
```

```
# A tibble: 4 x 3
  `a1c < 8` race_eth               n
  <lgl>     <fct>              <int>
1 FALSE     Hispanic or Latinx     7
2 FALSE     Non-Hispanic White    16
3 TRUE      Hispanic or Latinx    15
4 TRUE      Non-Hispanic White    24
```

dm431 Example A, rearranged

```
dm431 %>%
  filter(insurance == "Medicaid") %>%
  filter(race_eth %in% c("Hispanic or Latinx",
                         "Non-Hispanic White")) %>%
  mutate(a1c_cat = ifelse(a1c < 8,
                          "below_8", "8_or_higher")) %>%
  tabyl(a1c_cat, race_eth)
```

```
    a1c_cat Black or African-American
 8_or_higher                         0
     below_8                         0
 Hispanic or Latinx Non-Hispanic White
                  7                 16
                 15                 24
```

- What should we do to remove the column with no data?
- Do we have the outcome/exposure combination we want at the top left?

dm431 Example A, after droplevels()

```
dm431 %>%
  filter(insurance == "Medicaid") %>%
  filter(race_eth %in% c("Hispanic or Latinx",
                         "Non-Hispanic White")) %>%
  droplevels() %>%
  mutate(a1c_cat = ifelse(a1c < 8,
                          "below_8", "higher")) %>%
  tabyl(a1c_cat, race_eth)
```

```
a1c_cat Hispanic or Latinx Non-Hispanic White
below_8                 15                 24
 higher                  7                 16
```

- Is this in standard epidemiological format, with the rows indicating the exposure, and the columns indicating the outcome?
- What did I do to flip the rows?

dm431 Example A, standard epidemiological format

```
tableA <- dm431 %>%
  filter(insurance == "Medicaid") %>%
  filter(race_eth %in% c("Hispanic or Latinx",
                         "Non-Hispanic White")) %>%
  droplevels() %>%
  mutate(a1c_cat = ifelse(a1c < 8,
                          "below_8", "higher")) %>%
  tabyl(race_eth, a1c_cat)

tableA
```

```
           race_eth below_8 higher
 Hispanic or Latinx      15      7
 Non-Hispanic White      24     16
```

- `tableA` has the exposure categories in the rows and the outcome categories in the columns.
- Do we have the cell we want in the top left now?

`dm431` Example A

```r
tableA %>% adorn_totals(where = c("row", "col"))
```

```
          race_eth below_8 higher Total
 Hispanic or Latinx      15      7    22
Non-Hispanic White      24     16    40
             Total      39     23    62
```

- How many subjects do we have in each exposure group?
- How many subjects fall into each outcome group?

Can we augment the table to help us understand:

- What is the probability of achieving each of the two possible outcomes?
- How do the outcome probabilities differ by exposure group?

dm431 Example A

```
tableA %>% adorn_totals(where = c("row", "col")) %>%
    adorn_percentages(denom = "row") %>%
    adorn_pct_formatting(digits = 1) %>%
    adorn_ns(position = "front")
```

```
           race_eth     below_8      higher         Total
 Hispanic or Latinx 15 (68.2%)  7 (31.8%) 22 (100.0%)
 Non-Hispanic White 24 (60.0%) 16 (40.0%) 40 (100.0%)
              Total 39 (62.9%) 23 (37.1%) 62 (100.0%)
```

- Why am I using `denom = "row"` here?

  Among current Medicaid subjects, compare the proportion of Hispanic/Latinx subjects with A1c below 8 to the proportion of Non-Hispanic White subjects with A1c below 8.

- What are the sample estimates for the two rates I am comparing?

2 x 2 Table for Example A: Comparing Probabilities

–	A1c < 8	A1c higher	*Total*
Hispanic/Latinx	15	7	*22*
Non-Hisp. White	24	16	*40*
Total	*39*	*23*	*62*

- Pr(A1c < 8 $|$ Hispanic/Latinx) $= 15/22 = 0.682$
- Pr(A1c < 8 $|$ Non-Hispanic Wh.) $= 24/40 = 0.6$
- The ratio of those two probabilities (risks) is $0.682/0.6 = 1.14$.

Can we build a confidence interval for the relative risk of A1c < 8 now in the Hispanic/Latinx population as compared to the Non-Hispanic White population?

- The difference in those risks is 0.682 - 0.6 $= 0.082$.

How about a confidence interval for the risk difference, too?

2 x 2 Table for Example A, Odds Ratio

–	A1c < 8	A1c higher	*Total*
Hispanic/Latinx	15	7	*22*
Non-Hisp. White	24	16	*40*
Total	*39*	*23*	*62*

- Odds = Probability / (1 - Probability)
- Sample Odds of A1c < 8 now if Hispanic or Latinx $= \frac{15/22}{1-(15/22)} =$ 2.143
- Sample Odds of A1c < 8 now if non-Hispanic White $= \frac{24/40}{1-(24/40)} = 1.5$
- Ratio of these two Odds are 1.43.

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

- Here, the cross-product estimate $= \frac{15*16}{24*7} = 1.43$.

Can we build a confidence interval for the odds ratio for A1c < 8 now in the population given "old A1c < 8" as compared to "old A1c high"?

Using twobytwo from the Love-boost.R script

```
twobytwo(15, 7, 24, 16,
      "Hispanic/Latinx", "Non-Hisp. White",
      "A1c < 8", "A1c higher")
```

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : A1c < 8
Comparing : Hispanic/Latinx vs. Non-Hisp. White

                A1c < 8 A1c higher    P(A1c < 8)
Hispanic/Latinx      15          7        0.6818
Non-Hisp. White      24         16        0.6000
                95% conf. interval
Hispanic/Latinx    0.4663   0.8401
Non-Hisp. White    0.4435   0.7385

                                     95% conf. interval
```

The Complete `twobytwo` Output

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : A1c < 8
Comparing : Hispanic/Latinx vs. Non-Hisp. White

                A1c < 8 A1c higher    P(A1c < 8) 95% conf. interval
Hispanic/Latinx      15          7        0.6818    0.4663   0.8401
Non-Hisp. White      24         16        0.6000    0.4435   0.7385

                                    95% conf. interval
             Relative Risk: 1.1364    0.7760   1.6641
         Sample Odds Ratio: 1.4286    0.4766   4.2820
Conditional MLE Odds Ratio: 1.4205    0.4241   5.0936
    Probability difference: 0.0818   -0.1675   0.3000

             Exact P-value: 0.5910
        Asymptotic P-value: 0.5242
------------------------------------------------------
```

Hypothesis Testing?

The hypotheses being compared can be thought of in several ways...

- H_0: $\pi_1 = \pi_2$, vs. H_A: $\pi_1 \neq \pi_2$.
- H_0: Pr(A1c < 8 | Hispanic or Latinx) $=$ Pr(A1c < 8 | non-Hispanic White) vs. H_A: Pr(A1c < 8 | Hispanic or Latinx) $\neq$ Pr(A1c < 8 | non-Hispanic White).
- H_0: rows and columns of the table are *independent*, in that the probability of a good outcome in each row is the same vs. H_A: the rows and columns of the table are *associated*.

```
Exact P-value: 0.5910
Asymptotic P-value: 0.5242
```

- The `Exact P-value` comes from Fisher's exact test, and is technically exact only if we treat the row and column totals as being fixed.
- The `Asymptotic P-value` comes from a Pearson χ^2 test.
- Neither approach is helpful if we don't have sufficient data to justify inference in the first place.

Bayesian Augmentation in a 2x2 Table?

Original command:

```
twobytwo(15, 7, 24, 16,
      "Hispanic/Latinx", "Non-Hisp. White",
      "A1c < 8", "A1c higher")
```

Bayesian augmentation approach: Add a success and add a failure in each row…

```
twobytwo(15+1, 7+1, 24+1, 16+1,
      "Hispanic/Latinx", "Non-Hisp. White",
      "A1c < 8", "A1c higher")
```

`twobytwo` Output with Bayesian Augmentation

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : A1c < 8
Comparing : Hispanic/Latinx vs. Non-Hisp. White

                A1c < 8 A1c higher    P(A1c < 8) 95% conf. interval
Hispanic/Latinx      16          8        0.6667    0.4612   0.8237
Non-Hisp. White      25         17        0.5952    0.4426   0.7314

                                   95% conf. interval
             Relative Risk: 1.1200    0.7681   1.6331
         Sample Odds Ratio: 1.3600    0.4765   3.8816
Conditional MLE Odds Ratio: 1.3537    0.4268   4.5252
    Probability difference: 0.0714   -0.1692   0.2863

             Exact P-value: 0.6080
        Asymptotic P-value: 0.5655
------------------------------------------------------
```

Coming Soon

- Another Independent Samples Example, plus Paired Samples Comparison of Proportions
- Comparing More than 2 Means with Independent Samples: Analysis of Variance
- Power and Sample Size Ideas
- Working with Larger Contingency Tables (Chi-Square Tests of Independence)
- Mantel-Haenszel Procedures for Three-Way Tables