431 Class 19

thomaselove.github.io/431

2020-10-29

Today's Agenda

Comparing Rates/Proportions/Percentages/Probabilities/Risks

- Point Estimates and Confidence Intervals for a Population Proportion
  - Five Methods to accomplish this task
- Comparing Two Proportions using Independent Samples
  - Standard Epidemiological Format
  - Working with 2x2 Tables

Today's Setup and Data

```
knitr::opts_chunk$set(comment = NA)
options(dplyr.summarise.inform = FALSE)

library(Epi) # new today
library(janitor)
library(knitr)
library(magrittr)
library(mosaic) # not usually something I load
library(broom)
library(tidyverse)

theme_set(theme_bw())

dm431 <- readRDS("data/dm431_2020.Rds")
source("data/Love-boost.R")
```

Confidence Intervals for a Population Proportion

Moving on from Means to Proportions

We've focused on creating statistical inferences about a population mean when we have a quantitative outcome. Now, we'll tackle a **categorical** outcome.

We'll estimate a confidence interval around an unknown population proportion, or rate, symbolized with π, on the basis of a random sample of n observations from the population of interest.

The sample proportion is called $\hat{p}$, which is sometimes, unfortunately, symbolized as p.

- This $\hat{p}$ is the sample proportion - not a p value.

Hemoglobin A1c < 8 rate?

The dm431 data yields these results on whether each subject's Hemoglobin A1c level (a measure of blood sugar control) is below 8%[1].

```
dm431 %$%
  tabyl(a1c < 8)
```

```
 a1c < 8   n     percent valid_percent
   FALSE 147 0.341067285     0.3434579
    TRUE 281 0.651972158     0.6565421
      NA   3 0.006960557            NA
```

What can we conclude about the true proportion of Northeast Ohio adults ages 31-70 who live with diabetes whose A1c is below 8%?

[1] Having an A1c < 8 is a good thing, generally, if you have diabetes.

Our Sample and Our Population

Sample: 431 adult patients living in Northeast Ohio between the ages of 31 and 70, who have a diagnosis of diabetes.

- 281 of our 431 adult patients, or 65.2% have A1c < 8.

Our population: **All** adult patients living in Northeast Ohio between the ages of 31 and 70, who have a diagnosis of diabetes.

Our first inferential goal will be to produce a **confidence interval for the true (population) proportion** with A1c < 8, across all adults with diabetes ages 31-70 living in NE Ohio, based on this sample.

A Confidence Interval for a Proportion

A 100(1-α)% confidence interval for the population proportion π 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, that confidence interval estimate 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 π in the A1c $<$ 8 data

- We'll build a 95% confidence interval for the true population proportion, so $\alpha = 0.05$
- We have n $= 431$ subjects
- Sample proportion is $\hat{p} = .652$, since $281/431 = 0.652$.

The standard error of that sample proportion will be

$$\text{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.652(1-0.652)}{431}} = 0.023$$

Confidence Interval for π = Pr(A1c < 8)

Our 95% confidence interval for the true population proportion, π, of people whose A1c is below 8 is:

$$\hat{p} \pm Z_{.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.652 \pm 1.96(0.023) = 0.652 \pm 0.045$$

or (0.607, 0.697).

To verify that $Z_{0.025} = 1.96...$

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

```
[1] 1.959964
```

Likely Accuracy of this Confidence Interval?

Since $n\hat{p} = (431)(0.652) = 281$ and $n(1-\hat{p}) = (431)(1-0.652) = 150$ are substantially greater than 5, the CI should be reasonably accurate.

What can we conclude from this analysis?

- Point estimate of the population proportion with A1c < 8 is 0.652
- 95% confidence interval for the population proportion is (0.607, 0.697)

What is the "margin of error" in this confidence interval?

- The entire confidence interval has width 0.09 (or 9 percentage points.)
- The margin of error (or half-width) is 0.045, or 4.5 percentage points.

Happily, that's our last "by hand" calculation.

Using R to estimate a CI for a Proportion

I'll discuss five procedures for estimating a confidence interval for a population proportion. Each can be obtained using the `binom.test` function from within the `mosaic` package. For a 95% CI, we use:

```
mosaic::binom.test(x = 281, n = 431,
                   p = 0.5, conf.level = 0.95, # defaults
                   ci.method = "XXX")
```

where the appropriate `ci.method` is obtained from the table below.

Approach	`ci.method` to be used
Wald	"Wald"
Clopper-Pearson	"Clopper-Pearson" or "binom.test"
Score	"Score" or "prop.test"
Agresti-Coull	"agresti-coull"
Plus4	"plus4"

These 5 Approaches (each is approximate)

1. **Wald** is the "basic biostatistics" method we just calculated, where we estimate the standard error using the sample proportion and then use the Normal distribution to set the endpoints. The Wald interval is always symmetric, and can dip below 0 or above 1.
2. **Clopper-Pearson** is used by `stats::binom.test()` in R as well. It guarantees coverage at least as large as the nominal coverage rate, but may produce wider intervals than the other methods.
3. **Score** is used by `stats::prop.test()` and creates CIs by inverting p-values from score tests. It can be applied with a continuity correction (use ci.method = "prop.test") or without.
4. **Agresti-Coull** is the Wald method after adding Z successes and Z failures to the data, where Z is the appropriate quantile for a standard Normal distribution (1.96 for a 95% CI)
5. **Plus4** is the Wald method after adding 2 successes and 2 failures (so 4 observations) to the data.

Formulas? See Wikipedia's entry: Binomial proportion confidence interval

Method 1: The Wald Procedure

```
method1 <- binom.test(x = 281, n = 431, conf.level = 0.95,
                      ci.method = "Wald")
```

```
method1
```

```
Exact binomial test (Wald CI)

data:  281 out of 431
number of successes = 281, number of trials = 431,
                                        p-value = 2.795e-10
alternative hypothesis: true probability of success
                                        is not equal to 0.5
95 percent confidence interval: 0.6070013 0.6969430
sample estimates: probability of success  0.6519722
```

Tidying up a binom.test result from mosaic

```
tidy1 <- tidy(method1)

tidy1 %>%
  select(estimate, conf.low, conf.high, statistic, parameter) %>%
  kable(dig = 4)
```

estimate	conf.low	conf.high	statistic	parameter
0.652	0.607	0.6969	281	431

Method 2: The Clopper-Pearson Procedure

```
method2 <- binom.test(x = 281, n = 431, conf.level = 0.95,
                      ci.method = "Clopper-Pearson")

tidy2 <- tidy(method2)

tidy2 %>%
  select(estimate, conf.low, conf.high, statistic, parameter) %>%
  kable(dig = 4)
```

estimate	conf.low	conf.high	statistic	parameter
0.652	0.6049	0.6969	281	431

Method 3: The Score Procedure

```
method3 <- binom.test(x = 281, n = 431, conf.level = 0.95,
                      ci.method = "Score")

tidy3 <- tidy(method3)

tidy3 %>%
  select(estimate, conf.low, conf.high, statistic, parameter)
  kable(dig = 4)
```

estimate	conf.low	conf.high	statistic	parameter
0.652	0.6058	0.6954	281	431

Method 4: The Agresti-Coull Procedure

```
method4 <- binom.test(x = 281, n = 431, conf.level = 0.95,
                      ci.method = "agresti-coull")

tidy4 <- tidy(method4)

tidy4 %>%
  select(estimate, conf.low, conf.high, statistic, parameter) %>%
  kable(dig = 4)
```

estimate	conf.low	conf.high	statistic	parameter
0.652	0.6058	0.6954	281	431

Method 5: The Plus4 Procedure

```
method5 <- binom.test(x = 281, n = 431, conf.level = 0.95,
                      ci.method = "plus4")

tidy5 <- tidy(method5)

tidy5 %>%
  select(estimate, conf.low, conf.high, statistic, parameter) %>%
  kable(dig = 4)
```

estimate	conf.low	conf.high	statistic	parameter
0.652	0.6058	0.6954	281	431

Comparison of Methods

```
res1 <- tidy1 %>% select(estimate, conf.low, conf.high)
res2 <- tidy2 %>% select(estimate, conf.low, conf.high)
res3 <- tidy3 %>% select(estimate, conf.low, conf.high)
res4 <- tidy4 %>% select(estimate, conf.low, conf.high)
res5 <- tidy5 %>% select(estimate, conf.low, conf.high)

res <- bind_rows(res1, res2, res3, res4, res5)
res <- res %>% mutate(
  approach = c("Wald", "Clopper-Pearson", "Score",
               "Agresti-Coull", "Plus4"))
```

Results with too many decimal places

95% confidence intervals based on x = 281 successes in n = 431 trials.

```
res %>% kable(dig = 5)
```

estimate	conf.low	conf.high	approach
0.65197	0.60700	0.69694	Wald
0.65197	0.60492	0.69692	Clopper-Pearson
0.65197	0.60584	0.69542	Score
0.65197	0.60582	0.69544	Agresti-Coull
0.65197	0.60577	0.69538	Plus4

This is way more precision than we can really justify, but I just want you to see that the five results are all (slightly) different.

Results after some rounding

95% confidence intervals based on x = 281 successes in n = 431 trials.

```
res %>% kable(dig = 3)
```

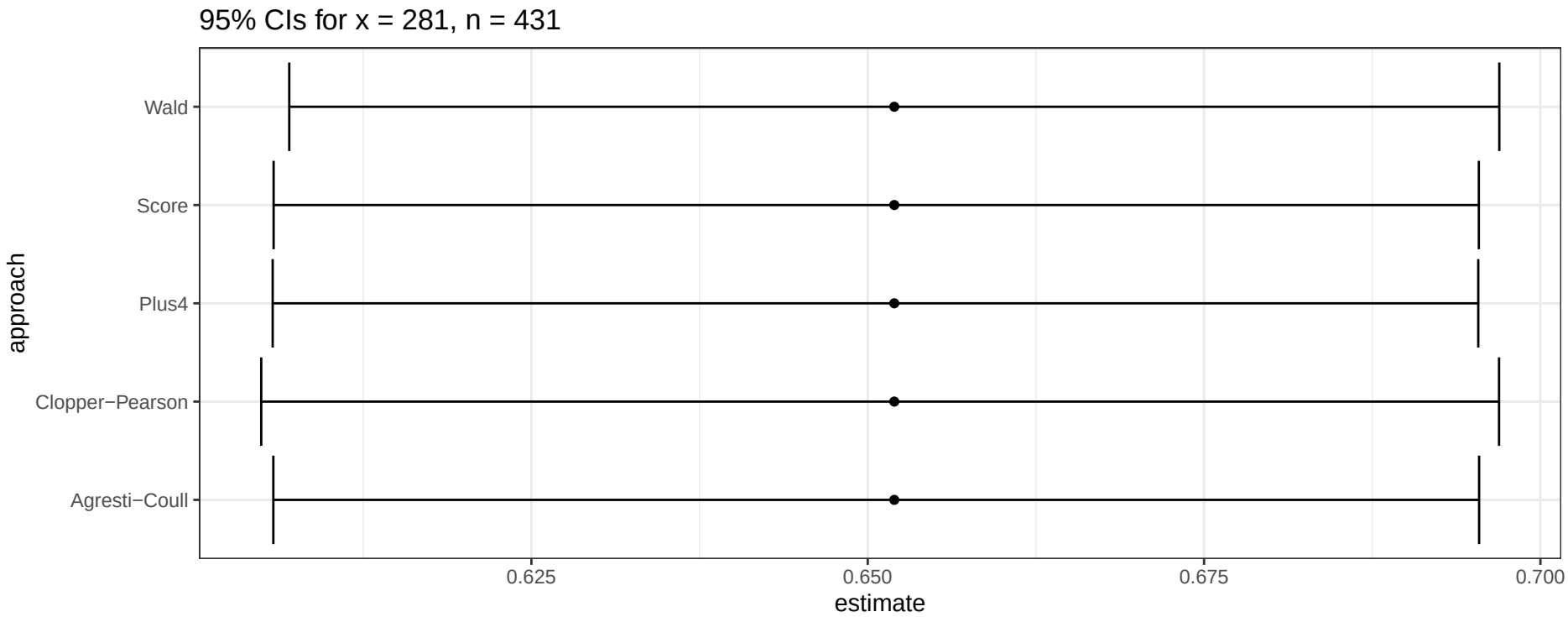
estimate	conf.low	conf.high	approach
0.652	0.607	0.697	Wald
0.652	0.605	0.697	Clopper-Pearson
0.652	0.606	0.695	Score
0.652	0.606	0.695	Agresti-Coull
0.652	0.606	0.695	Plus4

Here's a somewhat more plausible rounding approach.

- Is the distinction between methods important in this scenario?

Plotting the 95% CI Estimates

```
ggplot(res, aes(x = approach, y = estimate)) +
  geom_point() +
  geom_errorbar(aes(ymin = conf.low, ymax = conf.high)) +
  coord_flip() +
  labs(title = "95% CIs for x = 281, n = 431")
```



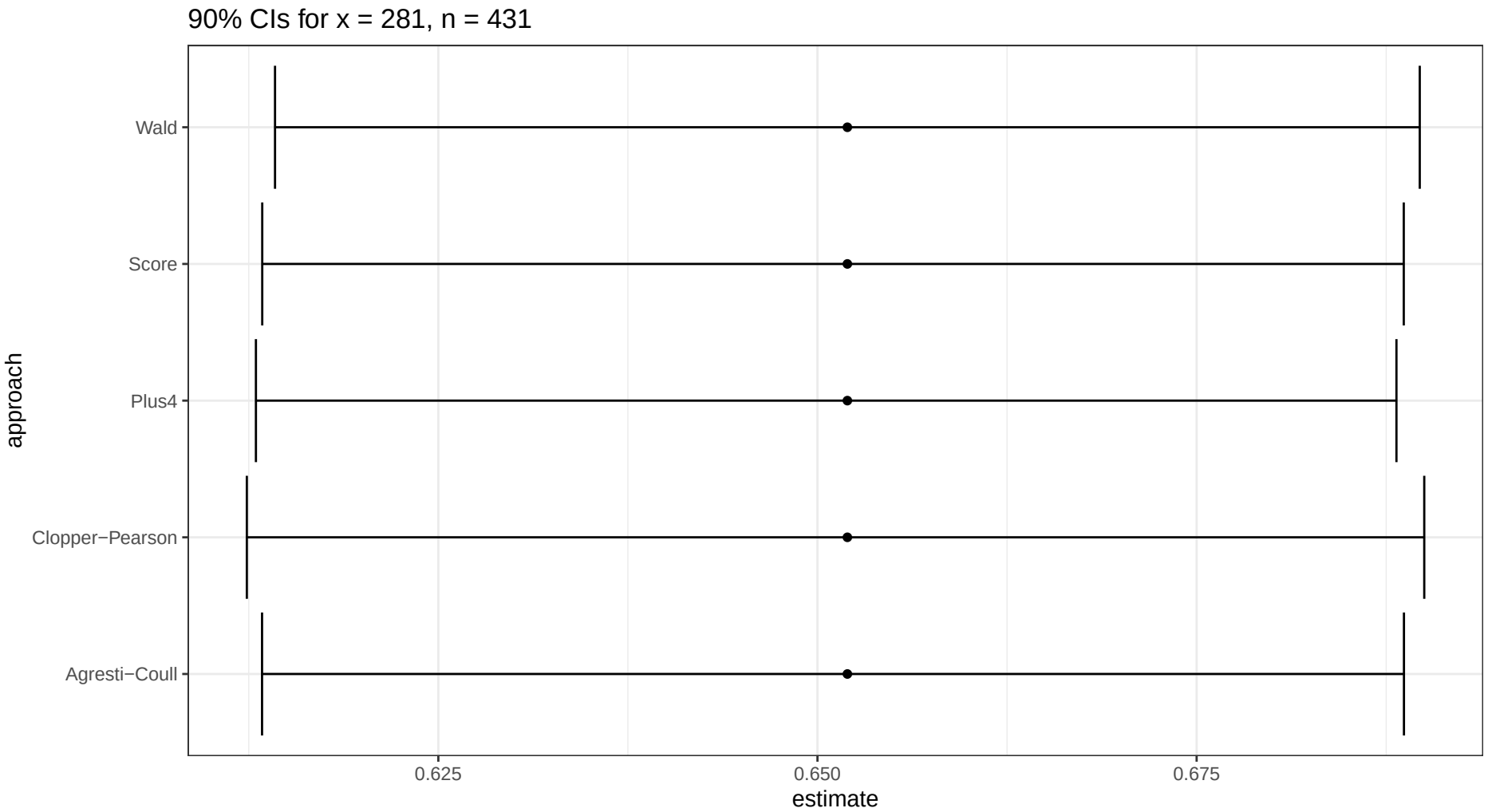
What if we ran 90% Confidence Intervals Instead?

90% confidence intervals based on x = 281 successes in n = 431 trials.

estimate	conf.low	conf.high	approach
0.652	0.614	0.690	Wald
0.652	0.612	0.690	Clopper-Pearson
0.652	0.613	0.689	Score
0.652	0.613	0.689	Agresti-Coull
0.652	0.613	0.688	Plus4

- I've hidden the code here, but it's available in the R Markdown.

Plotting the 90% CI Estimates



Estimating Rates 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 $\frac{x+2}{n+4}$ as an alternative[2]. We'll call this a *Bayesian augmentation*.

Estimates with and without the augmentation will be generally comparable, so long as. . .

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

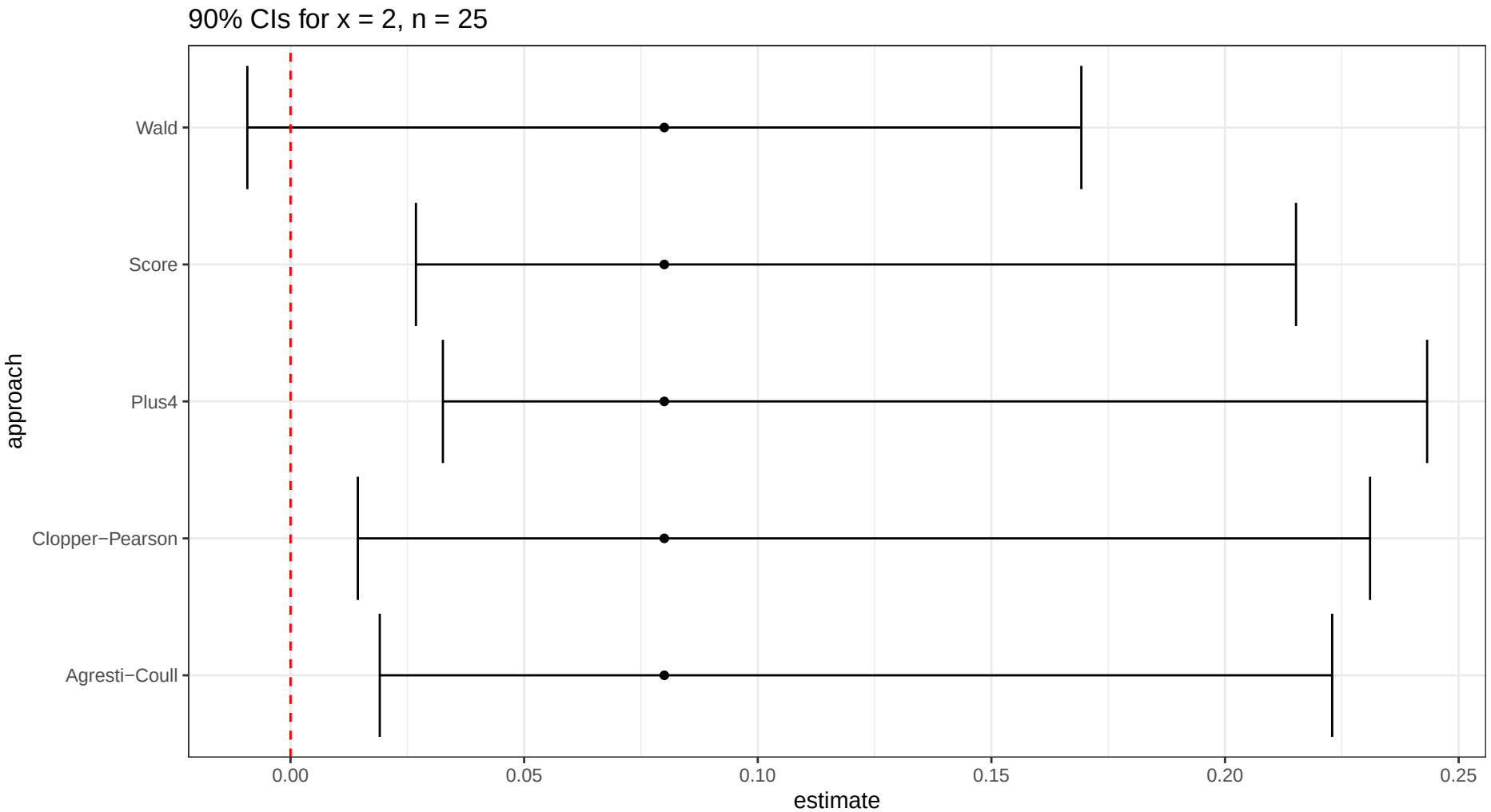
[2] See http://andrewgelman.com/2007/05/15 for instance.

What if we'd had 2 successes in 25 trials instead?

90% confidence intervals based on x = 2 successes in n = 25 trials.

estimate	conf.low	conf.high	approach
0.08	-0.009	0.169	Wald
0.08	0.014	0.231	Clopper-Pearson
0.08	0.027	0.215	Score
0.08	0.019	0.223	Agresti-Coull
0.08	0.033	0.243	Plus4

90% CI Estimates for x = 2, n = 25



What if x = 0 or x = n?

The **Rule of Three** approach is often used.

- An approximate 95% CI for the proportion in a setting where x = 0 in n trials is $(0, \frac{3}{n})$
- An approximate 95% CI for the proportion where x = n in n trials is $(1 - \frac{3}{n}, 1)$

Comparing Proportions

Comparing Population Proportions

Suppose we want to compare the population proportions π_1 and π_2, based on samples of sizes n_1 and n_2. We can do this using independent samples or using paired samples.

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 presidential election". 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

Example A: dm431

Let's compare the proportion who have an A1c below 8 between:

- Group 1: Medicaid insured subjects who have never smoked
- Group 2: Medicaid insured subjects who are current smokers

```
dm431 %>%
  filter(insurance == "Medicaid") %>%
  filter(tobacco %in%
           c("Never", "Current")) %>%
    count(a1c < 8, tobacco) %>% kable()
```

a1c < 8	tobacco	n
FALSE	Current	13
FALSE	Never	12
TRUE	Current	20
TRUE	Never	22

The Table We'd Like To Get To

Let's compare the proportion who have an A1c below 8 between:

- Group 1: Medicaid insured subjects who have never smoked
- Group 2: Medicaid insured subjects who are current smokers

Standard Epidemiological Format

- rows are the exposure
- columns are the outcome

What do we want in our setting?

Our Goal: Standard Epidemiological Format

- exposure is *tobacco status* (Never or Current)
- columns are *hemoglobin A1c* category ($<$ 8, 8 or more)

	A1c < 8	A1c >= 8
Never used	a	b
Current user	c	d

dm431 Example A, rearranged

```
dm431 %>%
  filter(insurance == "Medicaid") %>%
  filter(tobacco %in%
           c("Current", "Never")) %>%
  mutate(a1c_cat = ifelse(a1c < 8,
                          "below_8", "8_or_more")) %>%
  tabyl(a1c_cat, tobacco)
```

```
  a1c_cat Current Former Never
8_or_more      13      0    12
  below_8      20      0    22
```

- 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(tobacco %in%
           c("Current", "Never")) %>%
  droplevels() %>%
  mutate(a1c_cat = ifelse(a1c < 8,
                          "A1c<8", "A1c>=8")) %>%
  tabyl(a1c_cat, tobacco)
```

```
 a1c_cat Current Never
   A1c<8      20    22
  A1c>=8      13    12
```

- 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(tobacco %in% c("Current", "Never")) %>%
  droplevels() %>%
  mutate(tobacco = fct_relevel(tobacco, "Never")) %>%
  mutate(a1c_cat = ifelse(a1c < 8, "A1c<8", "A1c>=8")) %>%
  tabyl(tobacco, a1c_cat)

tableA
```

```
 tobacco A1c<8 A1c>=8
   Never    22      12
 Current    20      13
```

- Does tableA match what we want now?
- Exposure = rows, Outcome = columns, correct top left count?

dm431 Example A

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

```
 tobacco A1c<8 A1c>=8 Total
   Never    22      12    34
 Current    20      13    33
   Total    42      25    67
```

- 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")
```

tobacco	A1c<8	A1c>=8	Total
Never	22 (64.7%)	12 (35.3%)	34 (100.0%)
Current	20 (60.6%)	13 (39.4%)	33 (100.0%)
Total	42 (62.7%)	25 (37.3%)	67 (100.0%)

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

  Among current Medicaid subjects, compare the proportion of never smokers with A1c below 8 to the proportion of current smokers 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 $>=$ 8	*Total*
Never	22	12	*34*
Current	20	13	*33*
Total	*42*	*25*	*67*

- Pr(A1c $<$ 8 | Never) = 22/34 = 0.647
- Pr(A1c $<$ 8 | Current) = 20/33 = 0.606
- The ratio of those two probabilities (risks) is 0.647/0.606 = 1.068.

Can we build a confidence interval for the relative risk of A1c $<$ 8 now in the never smokers as compared to the current smokers?

- The difference in those risks is 0.647 - 0.606 = 0.041.

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

2 x 2 Table for Example A, Odds Ratio

–	A1c $<$ 8	A1c $>=$ 8	*Total*
Never	22	12	*34*
Current	20	13	*33*
Total	*42*	*25*	*67*

- Odds = Probability / (1 - Probability)
- Sample Odds of A1c $<$ 8 now if Never smoked = $\frac{22/34}{1-(22/34)} = 1.833$
- Sample Odds of A1c $<$ 8 now if Current smoker = $\frac{20/33}{1-(20/33)} = 1.538$
- Ratio of these two Odds are 1.192.

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

- Here, the cross-product estimate = $\frac{22*13}{12*20} = 1.192$.

Can we build a confidence interval for the population odds ratio for A1c $<$ 8 now given "Never" as compared to "Current"?

Using `twobytwo` from the `Love-boost.R` script

–	A1c < 8	A1c >= 8	*Total*
Never	22	12	*34*
Current	20	13	*33*
Total	*42*	*25*	*67*

Code we need is:

```
twobytwo(22, 12, 20, 13,   # note order of counts
         "Never", "Current", # names of the rows
         "A1c<8", "A1c>=8",  # names of the columns
         conf.level = 0.90)  # default is 95% confidence
```

Complete Output shown on the next slide

```
2 by 2 table analysis:
------------------------------------------------
Outcome    : A1c<8
Comparing : Never vs. Current

        A1c<8 A1c>=8    P(A1c<8) 90% conf. interval
Never      22     12      0.6471    0.5040   0.7679
Current    20     13      0.6061    0.4613   0.7343

                                    90% conf. interval
            Relative Risk: 1.0676    0.7823   1.4571
        Sample Odds Ratio: 1.1917    0.5187   2.7377
Conditional MLE Odds Ratio: 1.1885    0.4625   3.0712
    Probability difference: 0.0410   -0.1486   0.2271

             Exact P-value: 0.8032
        Asymptotic P-value: 0.7288
------------------------------------------------
```

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 | Never) = Pr(A1c $<$ 8 | Current) vs. H_A: Pr(A1c $<$ 8 | Never) $\neq$ Pr(A1c $<$ 8 | Current).
- H_0: rows and columns of the table are *independent*, in that the probability of A1c $<$ 8 in each row is the same vs. H_A: the rows and columns of the table are *associated*.

```
    Exact P-value: 0.8032
Asymptotic P-value: 0.7288
```

- 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(22, 12, 20, 13,
      "Never", "Current",
      "A1c<8", "A1c>=8", conf.level = 0.90)
```

Bayesian augmentation approach: Add two successes and add two failures in each row. . .

```
twobytwo(22+2, 12+2, 20+2, 13+2,
      "Never", "Current",
      "A1c<8", "A1c>=8", conf.level = 0.90)
```

Output shown on next slide.

```
2 by 2 table analysis:
------------------------------------------------
Outcome    : A1c<8
Comparing : Never vs. Current

        A1c<8 A1c>=8    P(A1c<8) 90% conf. interval
Never      24     14      0.6316    0.4965   0.7488
Current    22     15      0.5946    0.4582   0.7178

                                    90% conf. interval
            Relative Risk: 1.0622    0.7851   1.4371
        Sample Odds Ratio: 1.1688    0.5355   2.5513
Conditional MLE Odds Ratio: 1.1664    0.4837   2.8243
    Probability difference: 0.0370   -0.1437   0.2148

             Exact P-value: 0.8147
         Asymptotic P-value: 0.7424
------------------------------------------------
```

Example B: Statin use in Medicaid vs. Uninsured

In the dm431 data, suppose we want to know whether statin prescriptions are more common among Medicaid patients than Uninsured subjects. So, we want a two-way table with "Medicaid", "Statin" in the top left.

```
dm431 %>%
  filter(insurance %in% c("Medicaid", "Uninsured")) %>%
  tabyl(insurance, statin)
```

```
  insurance  0  1
 Commercial  0  0
   Medicaid 17 83
   Medicare  0  0
  Uninsured 15 29
```

But we want the tabyl just to show the levels of insurance we're studying. . .

Obtaining a 2x2 Table from a data frame

We want to know whether statin prescriptions are more common among Medicaid patients than Uninsured subjects.. So, we want a two-way table with "Medicaid", "Uninsured" in the top left.

```
dm431 %>%
  filter(insurance %in% c("Medicaid", "Uninsured")) %>%
  droplevels() %>%
  tabyl(insurance, statin)
```

```
 insurance  0  1
  Medicaid 17 83
 Uninsured 15 29
```

But we want Medicaid in the top row (ok) and "statin = yes" in the left column (must fix)...

Building and Releveling Factors in the data frame

```
exampleB <- dm431 %>%
  filter(insurance %in% c("Medicaid", "Uninsured")) %>%
  droplevels() %>%
  mutate(insur_f = fct_relevel(insurance, "Medicaid"),
         statin_f = fct_recode(factor(statin),
                               on_statin = "1", no_statin = "0"),
         statin_f = fct_relevel(statin_f, "on_statin"))

exampleB %>% tabyl(insur_f, statin_f)
```

insur_f	on_statin	no_statin
Medicaid	83	17
Uninsured	29	15

Since Medicaid was already on top, we didn't *have to* set `insur_f`.

Adorning the `tabyl` with % using row as denominator

```
exampleB %>% tabyl(insur_f, statin_f) %>%
  adorn_totals(where = c("row", "col")) %>%
  adorn_percentages(denom = "row") %>%
  adorn_pct_formatting(digits = 1) %>%
  adorn_ns(position = "front") %>%
  adorn_title(row = "Insurance", col = "Statin Status")
```

	Statin Status		
Insurance	on_statin	no_statin	Total
Medicaid	83 (83.0%)	17 (17.0%)	100 (100.0%)
Uninsured	29 (65.9%)	15 (34.1%)	44 (100.0%)
Total	112 (77.8%)	32 (22.2%)	144 (100.0%)

Running twoby2 against a data set

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

Original Data

```
twoby2(exampleB %$% table(insur_f, statin_f))
```

(output on next slide)

With Bayesian Augmentation

```
twoby2(exampleB %$% table(insur_f, statin_f) + 2)
```

(output on the slide after that)

```
2 by 2 table analysis:
------------------------------------------------------
Outcome    : on_statin
Comparing : Medicaid vs. Uninsured

           on_statin no_statin P(on_statin) 95% conf. interval
Medicaid          83        17      0.8300    0.7434   0.8916
Uninsured         29        15      0.6591    0.5090   0.7829

                                    95% conf. interval
              Relative Risk: 1.2593    1.0003   1.5854
          Sample Odds Ratio: 2.5254    1.1202   5.6933
Conditional MLE Odds Ratio: 2.5074    1.0252   6.1298
    Probability difference: 0.1709    0.0218   0.3307

              Exact P-value: 0.0299
          Asymptotic P-value: 0.0255
------------------------------------------------------
```

```
2 by 2 table analysis:
------------------------------------------------
Outcome    : on_statin
Comparing : Medicaid vs. Uninsured

           on_statin no_statin P(on_statin) 95% conf. interval
Medicaid          85        19      0.8173    0.7312   0.8803
Uninsured         31        17      0.6458    0.5023   0.7671

                                    95% conf. interval
             Relative Risk: 1.2655    1.0071   1.5901
         Sample Odds Ratio: 2.4533    1.1327   5.3136
Conditional MLE Odds Ratio: 2.4375    1.0464   5.6838
    Probability difference: 0.1715    0.0245   0.3261

              Exact P-value: 0.0251
          Asymptotic P-value: 0.0228
------------------------------------------------
```

Coming Up Next . . .

- Comparing two proportions with paired samples
- Power and Sample Size considerations for comparing means and proportions