# MEDB 5501, Module 13

2025-11-18

# Topics to be covered, 1

- What you will learn
  - The two by two crosstabulaton
  - Test of equality of two proportions
  - R code for the test of equality of two proportions
  - Sample size calculations
  - R code for sample size calculations

# Topics to be covered, 2

- What you will learn
  - Chi-squared test of independence
  - R code for the chi-squared test of independence
  - Odds ratios and relative risks
  - R code for odds ratios and relative risks
  - Your homework

# The crosstabulation of two binary variables

| | | Variable2 | |
|---|---|---|---|
| | | 0 | 1 |
| Variable1 | 0 | n00 | n01 |
| | 1 | n10 | n11 |

Speaker notes

One of the most common tables you will see in Statistics is the 2 by 2 crosstabulation. This table shows the counts associated with the combination of the levels of the two binary variables. There are only four numbers in this table, but there are numerous statistics that you can use to summarize what’s going on in this table.

# Example: Titanic data

- Crosstabulation

```
        survived
sex      yes  no
  female 308 154
  male   142 709
```

Speaker notes

This is an example of a crosstabulation.The number in the upper left corner, 308, represents the number of female passengers who survived (did not die). This includes Kate Winslet. The number in the lower right corner, 709, represents then number of male passengers who did not survive. This includes, sad to say, Leonardo diCaprio.

# Multiple ways to display, 1

- Crosstabulation

```
        survived
sex      yes  no
  female 308 154
  male   142 709
```

- Swap columns

```
        survived
sex       no yes
  female 154 308
  male   709 142
```

- Swap rows

```
        survived
sex      yes  no
  male   142 709
  female 308 154
```

- Swab both

```
        survived
sex       no yes
  male   709 142
  female 154 308
```

Speaker notes

The cross tabulation changes when you swap the rows, the columns, or both.

# Multiple ways to display, 2

- Transposed

```
        sex
survived female male
     yes    308  142
     no     154  709
```

- Transposed, swap columns

```
        sex
survived male female
     yes  142    308
     no   709    154
```

- Transposed, swap rows

```
        sex
survived female male
     no     154  709
     yes    308  142
```

- Transposed, swap both

```
        sex
survived male female
     no   709    154
     yes  142    308
```

Speaker notes

You can get four more tables by transposing the matrix. What was the rows becomes the columns and what was the columns becomes the rows.

# Row and column percents

- Crosstabulation with row totals

```
        survived
sex      yes  no Sum
  female 308 154 462
  male   142 709 851
```

- Row percents

```
        survived
sex            yes        no       Sum
  female 0.6666667 0.3333333 1.0000000
  male   0.1668625 0.8331375 1.0000000
```

- Cross tabulation with column totals

```
        survived
sex      yes  no
  female 308 154
  male   142 709
  Sum    450 863
```

- Column percents

```
        survived
sex            yes        no
  female 0.6844444 0.1784473
  male   0.3155556 0.8215527
  Sum    1.0000000 1.0000000
```

Speaker notes

The column percents are computed by dividing by the column totals. They add up to 100% within each column. The row percents are computed by dividing by the row totals. They add up to 100% within each row. The cell percents are computed by dividing by the overall total. They only add up to 100% when you sum across both the rows and the columns.

# Cell percents

- Cell totals

```
        survived
sex       yes   no  Sum
  female  308  154  462
  male    142  709  851
  Sum     450  863 1313
```

- Cell percents

```
        survived
sex            yes        no       Sum
  female 0.2345773 0.1172887 0.3518660
  male   0.1081493 0.5399848 0.6481340
  Sum    0.3427266 0.6572734 1.0000000
```

# My recommendation

- Outcome variable is the columns
- Intervention/exposure variable is the rows
- Calculate row percentages

```
        survived
sex            yes        no
  female 0.6666667 0.3333333
  male   0.1668625 0.8331375
```

Speaker notes

I have found that nine times out of ten, you want row percentages with the exposure/intervention variable as the rows and the outcome variable as the columns. This doesn’t always work, but it is usually what I try first. It shows how much the chances of a good outcomes (or sometimes the chances of a bad outcome) change when you switch levels of the exposure/intervention.

The orientation is also important. You want the percentages that are most interesting as close as possible. This is the proximity principle. The values within a column are nested above/below the other. The values within a row are farther apart.

# Always round, 1

```
  0.6666667
- 0.1668625
  ---------
  ?????????
```

Speaker notes

Often in a cross tabulation, you will do some mental math. You might, for example, want to subtract survival probabilities.

# Always round, 2

```
  0.67
- 0.17
  ----
  ????
```

Speaker notes

Notice how much easier the subtraction becomes when you round to two significant figures.

# Break #1

- What you have learned
  - The two by two crosstabulaton
- What’s coming next
  - Test of equality of two proportions

# The binomial distribution

- You have n trials of an “experiment”
- Two outcomes, “success” or “failure”
- $\pi$ is probability of success
- Each trial is independent

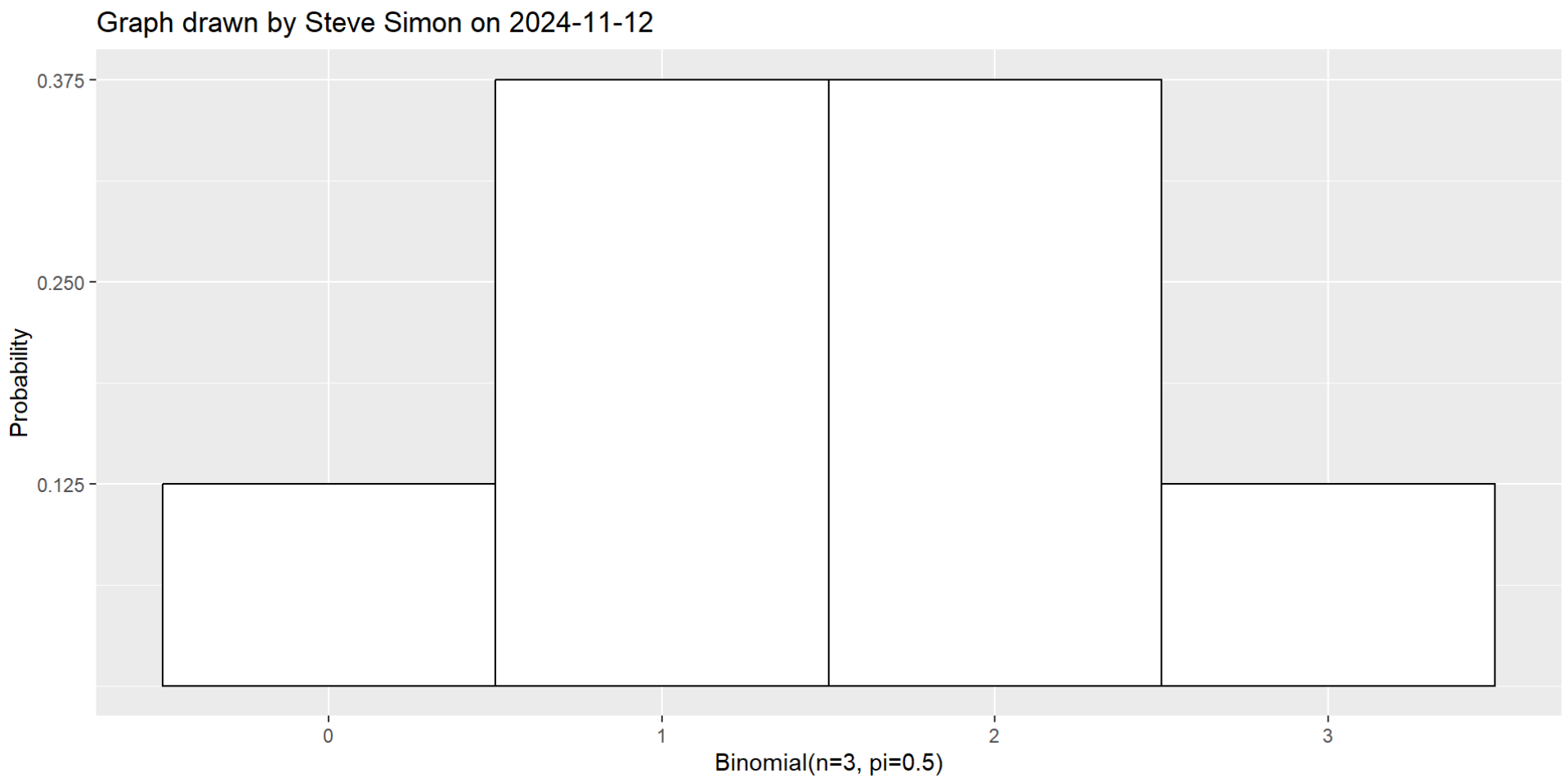
# Example, creating a family of three children, 1

- You have 3 trials (pregnancies)
- Two outcomes, girl (success) or boy (failure)
- $\pi = 0.5$ is probability of success
- Each child is independent

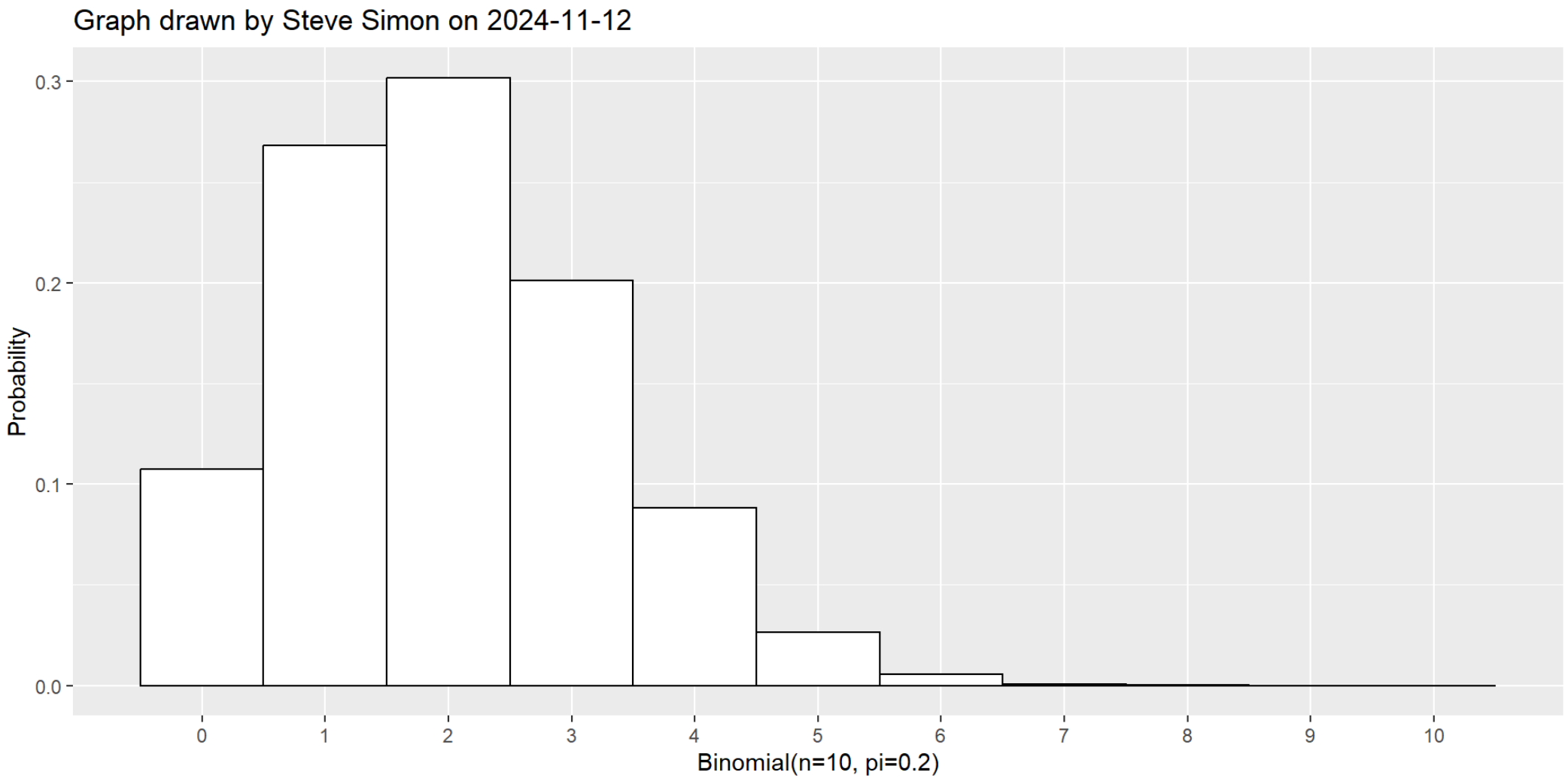
# Example, creating a family of three children, 2

- Eight possible outcomes
  - X=0, BBB
  - X=1, BBG, BGB, GBB
  - X=2, BGG, GBG, GGB
  - X=3, GGG

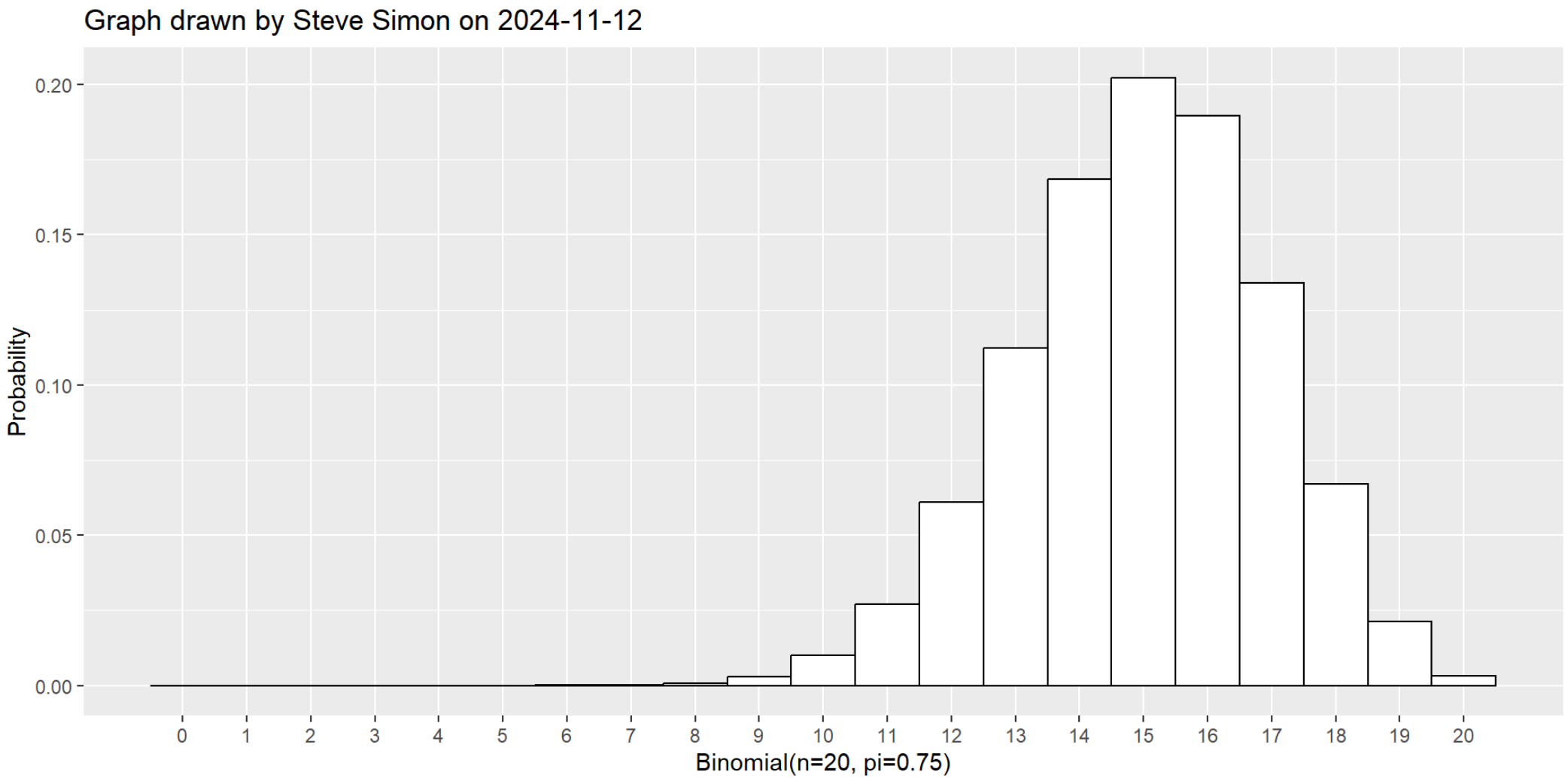
# Example, creating a family of three children, 3



# Example, Binomial(n=10, pi=0.2)



# Example, Binomial(n=20, pi=0.75)



# Two proportions model, 1

- Scenario 1
  - $X_1$ is a random binomial($n_1, \pi_1$)
  - $X_2$ is a random binomial($n_2, \pi_2$)
  - $X_1$ and $X_2$ are independent

# Two proportions model, 2

- Scenario 2
  - Sample $X_{11}, X_{12}, \ldots, X_{1n_1}$
  - Sample $X_{21}, X_{22}, \ldots, X_{2n_2}$
  - Only possible values are 0, 1
  - $P[X_{1i} = 1] = \pi_1, P[X_{2i} = 1] = \pi_2$

# Two proportions model, 3

- $H_0: \ \pi_1 - \pi_2 = 0$
- $H_1: \ \pi_1 - \pi_2 \neq 0$

# Two proportions model, 4

- $T = \frac{p_1 - p_2}{se}$
  - $se = \sqrt{\bar{p}(1-\bar{p})(\frac{1}{n_1} + \frac{1}{n_2})}$
  - $\bar{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$
- Accept $H_0$ if T is close to zero.
  - If $Z(\alpha/2) < T < Z(1-\alpha/2)$

# Two proportions model, 5

- p-value = $2P[Z > |T|]$
- Accept $H_0$ if p-value > $\alpha$

Speaker notes

You can also compute a p-value which is the probability of observing the test statistic, T, or a value more extreme.

# Some variations of the test statistic, 1

- Yates continuity correction
  - $T = \frac{p_1 - p_2 + c}{se}$
  - Formula for c is messy
  - Net effect is to pull T closer to zero
  - Better to meet the normal approximation

# Some variations of the test statistic, 2

- Chi-squared test
  - $T = \left( \frac{p_1 - p_2}{se} \right)^2$
  - Accept $H_0$ if T < $\chi^2(1 - \alpha, df = 1)$
  - p-value = $P[\chi^2(df = 1) > T]$
  - Does not allow for easy test of one-sided hypothesis
- Chi-squared test with Yates continuity correction
  - $T = \left( \frac{p_1 - p_2 + c}{se} \right)^2$

# One-sided test, 1

- $H_0 : \ \pi_1 - \pi_2 = 0$
- $H_1 : \ \pi_1 - \pi_2 > 0$
  - Accept $H_0$ if T < $z(1-\alpha)$
  - p-value = $P[Z > T]$
  - Accept $H_0$ if p-value > $\alpha$

# One-sided test, 2

- $H_0: \ \pi_1 - \pi_2 = 0$
- $H_1: \ \pi_1 - \pi_2 < 0$
  - Accept $H_0$ if T > $z(\alpha)$
  - p-value = $P[Z < T]$
  - Accept $H_0$ if p-value > $\alpha$

# Two sample test of proportions with Titanic data, 1

```
        survived
sex      yes  no
  female 308 154
  male   142 709
```

# Two sample test of proportions with Titanic data, 2

```
        survived
sex            yes        no
  female 0.6666667 0.3333333
  male   0.1668625 0.8331375
```

# Two sample test of proportions with Titanic data, 3

```
    2-sample test for equality of proportions without continuity correction

data:  table1
X-squared = 332.06, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.4500519 0.5495564
sample estimates:
   prop 1    prop 2
0.6666667 0.1668625
```

# Break #2

- What you have learned
  - Test of equality of two proportions
- What’s coming next
  - R code for the test of equality of two proportions

# R code for the test of two proportions

Please refer to the program simon-5501-13-titanic.qmd

Speaker notes

The

# Break #3

- What you have learned
  - R code for the test of equality of two proportions
- What's coming next
  - Sample size calculations

# What you need for a continuous outcome

- Research hypothesis
- Standard deviation of your outcome measure
- Minimum clinically important difference
- Other details
  - Type I error rate (usually 0.05)
  - Power (usually 0.90)

Speaker notes

In an earlier lecture, you saw that a sample size calculation needed three major elements, a research hypothesis, a standard deviation, and the minimum clinically important difference.

That’s not all, but the Type I error rate or alpha is usually fixed at 0.05. Power is usually fixed at 0.9. You might go a bit lower, but don’t go much lower than 0.8 for power. It might be okay from your perspective, but power around 0.75 or less is considered a red flag by those who might examine your research (IRB members, granting agencies, journal peer reviewers).

# What you need for a categorical outcome

- Research hypothesis
- Expected proportion in the control group
- Minimum clinically important difference

Speaker notes

The elements that you need when you have a categorical outcome change a bit. Instead of a standard deviation, you need to specify the expected proportion of events in the control group.

# Hypothetical scenario, 1

- Weight loss study
  - Control is recommended diet and exercise routine
  - Treatment adds experimental weight loss drug
  - Binary event, losing at least 5 pounds after 6 months
- Hypothesis proportion is higher in treatment group than control
- Proportion in control group is 10%
- Need to see at least 5% more events in treatment group

# Hypothetical scenario, 2

- Set power to 90%
- Set Type I error rate (alpha) to 5%

```
power.prop.test(
    p1=0.10,
    p2=0.15,
    power=0.9,
    sig.level=0.05)
```

```
     Two-sample comparison of proportions power calculation

              n = 917.3206
             p1 = 0.1
             p2 = 0.15
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number in *each* group
```

# Rule of 50

- Want to detect a doubling or halving
- Need 25 to 50 events in each group
- Example
  - Control probability is 0.1
  - Want to see 20% in treatment group
  - 25/0.1 = 250; 50/0.1 = 500

# Break #4

- What you have learned
  - Sample size calculations
- What's coming next
  - R code for sample size calculations

# R code for the sample size calculations

Please refer to the program simon-5501-13-titanic.qmd.

# Break #5

- What you have learned
  - R code for sample size calculations
- What's coming next
  - Chi-squared test of independence

# Crosstabulation with row and column totals

- Counts

```
        survived
sex       yes   no  Sum
  female  308  154  462
  male    142  709  851
  Sum     450  863 1313
```

- Cell percents

```
        survived
sex            yes        no       Sum
  female 0.2345773 0.1172887 0.3518660
  male   0.1081493 0.5399848 0.6481340
  Sum    0.3427266 0.6572734 1.0000000
```

# Conditional probability

- $P[A|B] = \frac{P[A \cap B]}{P[B]}$
  - Read $P[A|B]$ as probability of A given B
  - Read $P[A \cap B]$ as probability of A and B
  - Note: $P[A|B] \neq P[B|A]$

# What does independence mean?

- $P[A|B] = P[A]$
- Equivalent definition of independence
  - $P[A \cap B] = P[A] \times P[B]$

# Positive association

- $P[A|B] > P[A]$
- $P[A \cap B] > P[A] \times P[B]$
  - Change direction for negative association

# Expected counts

| | Good Outcome | Bad Outcome | Total |
|---|---|---|---|
| Placebo | ? | ? | (a+b)/n |
| Treated | ? | ? | (c+d)/n |
| Total | (a+c)/n | (b+d)/n | 1 |

where n=a+b+c+d

- $E_{11} = n \times \frac{a+b}{n} \times \frac{a+c}{n}$
  - $E_{12}$, $E_{21}$, $E_{22}$ are defined similarly

# Expected counts for Titanic data

```
        survived
sex            yes        no       Sum
  female                    0.3518660
  male                      0.6481340
  Sum    0.3427266 0.6572734 1.0000000
```

- $E_{11}$ = 1313 $\times$ 0.3518660 $\times$ 0.3427266 = 158.3397091
- $E_{12}$ = 303.6603489
- $E_{21}$ = 291.6603167
- $E_{22}$ = 559.3396253

# Expected counts for Titanic data

- Observed counts

```
        survived
sex      yes  no
  female 308 154
  male   142 709
```

- Expected counts

```
        survived
sex           yes       no
  female 158.3397 303.6603
  male   291.6603 559.3397
```

# Test statistic

- $H_0: \ Independence$
- $H_1: \ Dependence$
  - $T = \Sigma_i \Sigma_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
  - p-value = $P[T > \chi^2(df = 1)]$
    - Accept $H_0$ if T < $\chi^2(1 - \alpha, df = 1)$
    - Accept $H_0$ if p-value > $\alpha$

# Example with Titanic data

- $H_0$ : Mortality is independent of sex
- $H_1$ : Mortality is related to sex

```r
m1 <- chisq.test(table1, correct=FALSE)
m1
```

```
    Pearson's Chi-squared test

data:  table1
X-squared = 332.06, df = 1, p-value < 2.2e-16
```

Speaker notes

Since the test statistic is a lot larger than the degrees of freedom and since the p-value is small, reject the null hypothesis and conclude that there is a relationship between sex and survival.

# Chi-squared test is an approximation

- Reasonable if all expected counts > 5
- Use Fisher's Exact test otherwise

# Fisher’s exact test for the Titanic data

```r
m2 <- fisher.test(table1)
m2
```

```
    Fisher's Exact Test for Count Data

data:  table1
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  7.601263 13.122462
sample estimates:
odds ratio
  9.965185
```

Speaker notes

Although the expected counts are much larger than 5, here is the code for running Fisher’s Exact test.

# Break #6

- What you have learned
  - Chi-squared test of independence
- What’s coming next
  - R code for the chi-squared test of independence

# R code for the Chi-squared test of independence

Please refer to the program simon-5501-13-titanic.qmd.

# Break #7

- What you have learned
  - R code for the chi-squared test of independence
- What's coming next
  - Odds ratios and relative risks

# The crosstabulation of two binary variables, 1

| | Good Outcome | Bad Outcome |
|---|---|---|
| Placebo | a | b |
| Treated | c | d |

- a, number of placebo patients with good outcome
- b, number of placebo patients with bad outcome
- c, number of treated patients with good outcome
- d, number of treated patients with bad outcome

Speaker notes

One of the most common tables you will see in Statistics is the 2 by 2 crosstabulation. This table shows the counts associated with the combination of the outcome and the treatment group. Is the risk of a bad outcome different between the two treatment groups?

# The crosstabulation of two binary variables, 2

- Note: rows could be
  - Exposed/Unexposed
  - Female/Male
  - Old/Young
  - Many other possibilities

Speaker notes

This table appears in many other contexts. You might want to compare people exposed to an environmental hazard to unexposed people. You might want to compare demographic groups: females to males, old to young, etc. There are many other possibilities.

# Example: Titanic data

- Crosstabulation

```
        survived
sex      yes  no
  female 308 154
  male   142 709
```

Speaker notes

This is an example of a crosstabulation. The number in the upper left corner, 308, represents the number of female passengers who survived (did not die). This includes Kate Winslet. The number in the lower right corner, 709, represents then number of male passengers who did not survive. This includes, sad to say, Leonardo diCaprio.

# Odds ratio

| | Good Outcome | Bad Outcome | Odds |
|---|---|---|---|
| Placebo | a | b | b/a |
| Treated | c | d | d/c |

- Odds ratio = $\frac{d/c}{b/a} = \frac{ad}{bc}$

Speaker notes

The odds are the number of bad outcomes divided by the number of good outcomes. The ratio of these odds is the odds ratio.

You may sometimes see the odds ratio computed as the product of the diagonal entries divided by the product of the off-diagonal entries.

# Relative risk (Risk ratio)

| | Good Outcome | Bad Outcome | Probabililty |
|---|---|---|---|
| Placebo | a | b | b/(a+b) |
| Treated | c | d | d/(c+d) |

- Relative risk = $\frac{\frac{b}{a+b}}{\frac{d}{c+d}}$

Speaker notes

The odds are the number of bad outcomes divided by the number of good outcomes. The ratio of these odds is the odds ratio.

You may sometimes see the odds ratio computed as the product of the diagonal entries divided by the product of the off-diagonal entries.

# Using odds

- Three to one in favor of victory
  - Expect three wins for every loss
- Four to one odds against victory
  - Expect four losses for every win
- Odds = Prob / (1- Prob)
- Prob = Odds / (Odds + 1)

*Speaker notes*

The relationship between odds and probability Another approach is to try to model the odds rather than the probability of BF. You see odds mentioned quite frequently in gambling contexts. If the odds are three to one in favor of your favorite football team, that means you would expect a win to occur about three times as often as a loss. If the odds are four to one against your team, you would expect a loss to occur about four times as often as a win.

You need to be careful with odds. Sometimes the odds represent the odds in favor of winning and sometimes they represent the odds against winning. Usually it is pretty clear from the context. When you are told that your odds of winning the lottery are a million to one, you know that this means that you would expect to having a losing ticket about a million times more often than you would expect to hit the jackpot.

It’s easy to convert odds into probabilities and vice versa. With odds of three to one in favor, you would expect to see roughly three wins and only one loss out of every four attempts. In other words, your probability for winning is 0.75.

If you expect the probability of winning to be 20%, you would expect to see roughly one win and four losses out of every five attempts. In other words, your odds are 4 to 1 against.

The formulas for conversion are

odds = prob / (1-prob)

and

prob = odds / (1+odds).

In medicine and epidemiology, when an event is less likely to happen and more likely not to happen, we represent the odds as a value less than one. So odds of four to one against an event would be represented by the fraction 1/5 or 0.2. When an event is more likely to happen than not, we represent the odds as a value greater than one. So odds of three to one in favor of an event would be represented simply as an odds of 3. With this convention, odds are bounded below by zero, but have no upper bound.

# Ambiguity in odds

- “In favor” versus “Against”
- “Good” outcome versus “Bad” outcome
- Get clues from the context
  - Example: chances of winning the lottery (million to one)
    - One million winners for every loser?
    - One million losers for every winner?

# Example of odds and probability

Odds for winning election to U.S. president in 2024

- Biden: $\frac{8/13}{1+8/13} = \frac{8}{21} = 0.381$
- Trump: $\frac{1/3}{1+1/3} = \frac{1}{4} = 0.25$
- DeSantis: $\frac{1/16}{1+1/16} = \frac{1}{17} = 0.059$

*Speaker notes*

To convert from odds to probability, use the formula odds/(1+odds). You have to flip these around because 40 to 1 odds does not mean that Michelle Obama has 40 chances to win for every one chance of a loss.

Table downloaded from oddschecker.com

# Probability of winning 2022 World Cup

```
Brazil: 30.8%
Argentina: 18.2%
France: 16.7%
Spain: 13.3%
England: `10%
Portugal: 7.7%
Netherlands: 5.3%
Croatia: 2.8%
```

```
Switzerland: 1.5%
Japan: 1.5%
Morocco: 1.2%
USA: 1.1%
Senegal: 1%
South Korea: 0.67%
Poland: 0.55%
Australia: 0.5%
```

Argentina:

$$\frac{0.182}{1-0.182} = 0.2225 \approx 2/9$$

France:

$$\frac{0.167}{1-0.167} = 0.2004 \approx 1/5$$

*Speaker notes*

These probabilities were computed from a table of odds posted at the beginning of the round of 16 for the football world cup. Convert these back to odds.

These odds were taken from a December 2, 2022 blog post on the DraftKings website.

# Odds against winning 2022 football World Cup

Brazil: 9 to 4
Argentina: 9 to 2
France: 5 to 1
Spain: 13 to 2
England: 9 to 1
Portugal: 12 to 1
Netherlands: 18 to 1
Croatia: 35 to 1
Switzerland: 65 to 1
Japan: 65 to 1
Morocco: 80 to 1
USA: 90 to 1
Senegal: 100 to 1
South Korea: 150 to 1
Poland: 180 to 1
Australia: 200 to 1

*Speaker notes*

Here are all the odds. Notice that the United States was rightfully given almost no chance of winning. But wait until the women’s football World Cup.

# Odd ratio for Titanic data

```
        survived
sex       yes   no  odds
  female  308  154  154/308 = 0.5
  male    142  709  709/142 = 4.993 (round to 5.0)

Odds ratio = 5.0 / 0.5 = 10

        survived
sex       yes   no  odds
  male    142  709  709/142 = 4.993 (round to 5.0)
  female  308  154  154/308 = 0.5

Odds ratio = 0.5 / 5.0 = 0.1
```

# Relative risk for Titanic data, 1

```
        survived
sex       yes   no  total probability
  female  308  154  462   154/462 = 0.3333
  male    142  709  851   709/851 = 0.8331

Relative risk = 0.8331 / 0.3333 = 2.500

        survived
sex       yes   no  total probability
  male    142  709  851   709/851 = 0.8331
  female  308  154  462   154/462 = 0.3333

Relative risk = 0.3333 / 0.8331 = 0.4
```

# Relative risk for Titanic data, 2

```
        survived
sex       no   yes  total probability
  female  154  308  462   308/462 = 0.6667
  male    709  142  851   142/851 = 0.1669

Relative risk = 0.1669 / 0.6667 = 0.25

        survived
sex       no   yes  total probability
  male    709  142  851   142/851 = 0.1669
  female  154  308  462   308/462 = 0.6667

Relative risk = 0.6667 / 0.1669 = 4.0
```

# Break #8

- What you have learned
  - Odds ratios and relative risks
- What's coming next
  - R code for odds ratios and relative risks

# R code for odds ratios and relative risks

Please refer to the program simon-5501-13-titanic.qmd.

# Break #9

- What you have learned
  - R code for odds ratios and relative risks
- What's coming next
  - Your homework

# Your homework

Please refer to the file simon-5501-13-directions.qmd.

# Summary, 1

- What you have learned
  - The two by two crosstabulaton
  - Test of equality of two proportions
  - R code for the test of equality of two proportions
  - Sample size calculations
  - R code for sample size calculations

# Summary, 2

- What you have learned
  - Chi-squared test of independence
  - R code for the chi-squared test of independence
  - Odds ratios and relative risks
  - R code for odds ratios and relative risks
  - Your homework