MEDB 5501, Module 14

2024-11-19

Topics to be covered

- What you will learn
  - Chi-squared test of independence
  - Odds ratio and relative risk
  - R code
  - Your homework

Listing of titanic.yaml, 1

```yaml
---
data_dictionary: titanic.txt

description: >
  The Titanic was a large cruise ship, the biggest of its kind in 1912. It was
  thought to be unsinkable, but when it set sail from England to America in
its
  maiden voyage, it struck an iceberg and sank, killing many of the passengers
  and crew. You can get fairly good data on the characteristics of passengers
  who died and compare them to those that survived. The data indicate a strong
  effect due to age and gender, representing a philosophy of "women and
children
  first" that held during the boarding of life boats.

additional_description:
  http://www.statsci.org/data/general/titanic.html
```

Listing of titanic.yaml, 2


```
download_url:
  - http://www.statsci.org/data/general/titanic.txt

format:
  tab-delimited

varnames:
  first row of data

missing_value_code:
  NA
```

Listing of titanic.yaml, 3


```
size:
  rows: 1313
  columns: 5

vars:
  Name:
    label: Passenger name

  PClass:
    label: Passenger class
    scale: ordinal
    values: 1st, 2nd, 3rd
```

Listing of titanic.yaml, 4


```
Age:
  unit: years
  scale: ratio
  range: positive real numbers
  missing: NA

Sex:
  scale: binary
  values: female, male
```

Listing of titanic.yaml, 5


```
  Survived:
    scale: binary
    values:
      1: yes
      0: no
---
```

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

- $P[A|B] = \frac{P[A \cap B]}{P[B]}$
  - 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 > α

Example with Titanic data

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

```
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 #1

- What you have learned
  - Chi-squared test of independence
- What's coming next
  - Odds ratio and relative risk

The crosstabulation of two binary variables, 1

```
              Good      Bad
             Outcome  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

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

2024 ELECTION ODDS

CANDIDATE	ELECTION ODDS	IMPLIED % CHANCE
Joe Biden	13/8	38.1%
Donald Trump	3/1	25%
Ron DeSantis	16/1	5.9%
Robert Kennedy Jr	16/1	5.9%
Kamala Harris	40/1	2.4%
Michelle Obama	40/1	2.4%

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

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

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.

Odds for Titanic

```
        survived
sex        yes   no  odds
  female  308  154  154/308 = 0.5
  male    142  709  709/142 = 4.993

Odds ratio = 4.993 / 0.5 = 9.986

Odds ratio = 308*709 / 154*142 = 9.986
```

Break #2

- What you have learned
  - Odds ratio and relative risk
- What's coming next
  - R code

Listing of simon-5501-14-titanic.qmd, 1

```
---
title: "Analysis of Titanic dataset"
format:
  html:
    embed-resources: true
---

This program reads data on survival of passengers on the Titanic. Find more
information in the [data dictionary][dd].

[dd]: https://github.com/pmean/data/blob/main/files/titanic.yaml

This code was written by Steve Simon on 2024-11-09 and is placed in the public
domain.
```

Listing of simon-5501-14-titanic.qmd, 2

````
## Load the tidyverse library

```{r}
#| label: setup
#| message: false
#| warning: false
library(epitools)
library(tidyverse)
```
````

Listing of simon-5501-14-titanic.qmd, 3

```
#### Comments on the code

For most of your programs, you should load the [tidyverse library][tid1]. The
messages and warnings are suppressed.

[tid1]: https://www.tidyverse.org/

In previous programs, I put a label for each chunk inside the curly braces
({}).
It is recommended instead to put the label on a separate line inside the
program
chunk. It is a bit more work to provide a unique label for each chunk, but it
helps quite a bit to isolate where to look when your code produces an error.
```

Listing of simon-5501-14-titanic.qmd, 4

````
## Read the data and view a brief summary

```{r}
#| label: read
ti <- read_tsv(
  file="../data/titanic.txt",
  col_names=TRUE,
  col_types="ccncn",
  na="NA")
names(ti) <- tolower(names(ti))
glimpse(ti)
```
````

Listing of simon-5501-14-titanic.qmd, 5

```
#### Comments on the code

Use read_tsv from the [readr package][rea1] to  read this file. Use
col_names=TRUE because the column names are included as the first row of the
file. The col_types="ccncn" specifies the first second and fourth columns as
strings and the third and fifth as numeric. There are missing values in this
dataset, designated by the letters "NA".

[rea1]: https://readr.tidyverse.org/
```

Listing of simon-5501-14-titanic.qmd, 6

````
## Replace numeric codes for survived

```{r}
#| label: replace-numbers
ti$survived <-
    factor(
        ti$survived,
        level=1:0,
        labels=c("yes", "no"))
```
````

Listing of simon-5501-14-titanic.qmd, 7

````
## Get counts of sex by survival

```{r}
#| label: counts
table1 <-xtabs(~sex+survived, data=ti)
table1
```
````

Listing of simon-5501-14-titanic.qmd, 8

```
#### Comments on the code

The [table function][tab1] or the [xtabs function][xta1] creates a matrix with
the number of observations in each combination of sex and survived. These
values are placed in a single column. The [xtabs function][xta1] or the
[count function][cou1] provide slightly different approaches.

[cou1]: https://dplyr.tidyverse.org/reference/count.html
[tab1]: https://stat.ethz.ch/R-manual/R-devel/library/base/html/table.html
[xta1]: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/xtabs.html
```

Listing of simon-5501-14-titanic.qmd, 9

````
#### Interpretation of the output

There are 154 female passengers who died and 308 who survived. There are 709
male passengers who died and 142 who survived.

## Chi-squared test, 1

```{r}
#| label: chi-squared-test-1
m1 <- chisq.test(table1, correct=FALSE)
m1
```
````

Listing of simon-5501-14-titanic.qmd, 10

```
#### Comments on the code

The [chisq.test function][chi1] calculates a chi-square test of independence.
It
takes input in a variety of forms. In this example, it uses a crosstabulation
computed by the xtabs command as input.

This function also will run a goodness-of-fit test, which is not discussed in
this lecture.

[chi1]: https://stat.ethz.ch/R-manual/R-
devel/library/stats/html/chisq.test.html
```

Listing of simon-5501-14-titanic.qmd, 11

````
#### Interpretation of the output

The chi-squared statistic is much larger than the degrees of freedom and the
p-value is small. You should reject the null hypothesis and conclude that sex
and survival are related (not independent)

## Chi-squared test, 2

```{r}
#| label: chi-squared-test-2
m1$observed
```
````

Listing of simon-5501-14-titanic.qmd, 12

````
## Chi-squared test, 3

```{r}
#| label: chi-squared-test-3
m1$expected
```

## Fisher's Exact test

```{r}
#| label: fishers-exact
fisher.test(table1)
```
````

Listing of simon-5501-14-titanic.qmd, 13

```
#### Comments on the code

The [fisher.test function][fis1] calculates the Fisher's exact test, which is
helpful for small sample sizes. The 1,313 passengers on the Titanic do not
constitute a small sample size by any means. This test is just shown as an
example of how to calculate this test.

[fis1]: https://stat.ethz.ch/R-manual/R-
devel/library/stats/html/fisher.test.html
```

Listing of simon-5501-14-titanic.qmd, 14

````
#### Interpretation of the output

The p-value is small. You should reject the null hypothesis and conclude that
sex and survival are related (not independent). The estimated odds ratio is
9.97. The confidence interval for the odds ratio excludes the value of 1
leading to the same conclusion. In fact, even after allowing for sampling
error
that odds of survival are at least 7.6 times greater for women than for me.

## Odds ratio calculation

```{r}
#| label: odds-ratio
oddsratio(table1)
```
````

Listing of simon-5501-14-titanic.qmd, 15

```
#### Comments on the code

The oddsratio function and riskratio function (see below) are part of the
[epitools library][epi1]. It produces an odds ratio and confidence interval
and p-values associated with the Fisher's Exact test and the Chi-squared test
of
independence.

[epi1]: https://cran.r-project.org/web/packages/epitools/epitools.pdf
```

Listing of simon-5501-14-titanic.qmd, 16

```
#### Interpretation of the output

We are 95% confident that the odds ratio of survival for women versus men is
at
least 7.7 and possibly as large as 13, after accounting for sampling error.
This interval excludes the value of 1, so you can conclude that the risk of
death is much higher for men than for women. Equivalently you could conclude
that the odds of survival are much higher for women than for men.
```

Listing of simon-5501-14-titanic.qmd, 17

````
## Risk ratio calculation, 1

```{r}
#| label: risk-ratio-1
table1 |>
  proportions("sex")
```
````

Listing of simon-5501-14-titanic.qmd, 18

````
#### Interpretation of the output

Before calculating the risk ratio, let's look at the row percentages one more
time. The probability of survival is around 2/3 for women and about 1/6 for
men.
This means that the risk ratio from a survival perspective is around 4 (2/3
divided by 1/6). The probability of death is 1/3 for females and about 5/6 for
males. The risk ratio from a mortality perspective is 0.4 (1/3 divided by
5/6).

## Risk ratio calculation, 2

```{r}
#| label: risk-ratio-2
riskratio(table1)
```
````

Listing of simon-5501-14-titanic.qmd, 19

```
#### Interpretation of the output

The risk ratio is comparing the probability of death between men and women.
Men
have 2.5 times higher probability of death compared to women. The confidence
interval excludes the value of 1, indicating a statistically significant
increase.

## Risk ratio calculation, 3

```{r}
#| label: risk-ratio-3
riskratio(table1, rev="columns")
```
```

Listing of simon-5501-14-titanic.qmd, 20

````
#### Interpretation of the output

The risk ratio is comparing the probability of survival between men and women.
Men has one-fourth the probability of survival compared to women. The
confidence
interval excludes the value of 1, indicating that men have a statistically
significantly lower probability of survival compared to women.

## Save data for later use

```{r save}
save(ti, file="../data/titanic.RData")
```
````

Listing of simon-5501-14-titanic.qmd, 21

```
#### Comments on the code

It is usually a good idea to [save][sav1] your data in an RData file to make
it
easier to retrieve this data later (with the [load function][loa1]).

[sav1]: https://stat.ethz.ch/R-manual/R-devel/library/base/html/save.html
[loa1]: https://stat.ethz.ch/R-manual/R-devel/library/base/html/save.html
```

Break #3

- What you have learned
  - R code
- What's coming next
  - Your homework

Listing of simon-5501-14-directions.md, 1

```
---
title: "Directions for 5501-14 programming assignment"
---

This programming assignment was written by Steve Simon on 2024-11-19 and is
placed in the public domain.
```

Listing of simon-5501-14-directions.md, 2

```
## Program

-   Download the [program][tem]
    -   Store it in your src folder
-   Modify the file name
    -   Use your last name instead of "simon"
-   Modify the documentation header
    -   Add your name to the author field
    -   Optional: change the copyright statement

[tem]: https://github.com/pmean/classes/blob/master/biostats-1/14/src/simon-
5501-14-titanic.qmd
```

Listing of simon-5501-14-directions.md, 3

```
## Data

-   Download the [data][dat] file
    -   Store it in your data folder
-   Refer to the [data dictionary][dic], if needed.

[dat]: https://github.com/pmean/data/blob/main/files/titanic.txt
[dic]: https://github.com/pmean/data/blob/main/files/titanic.yaml
```

Listing of simon-5501-14-directions.md, 4

```
## Question 1

Create a new variable, third_class that indicates whether a passenger is in
third class or not. What is the odds ratio comparing survival between third
class passengers and first/second class passengers? Interpret this odds ratio
and the associated confidence interval.

## Question 2

Calculate a chi-squared test of independence that examines the association
between passenger class (third versus first/second) and mortality. Interpret
the test result.
```

Listing of simon-5501-14-directions.md, 5

```
## Your submission

-   Save the output in html format
-   Convert it to pdf format.
-   Make sure that the pdf file includes
    -   Your last name
    -   The number of this course
    -   The number of this module
-   Upload the file

## If it doesn't work

Please review the [suggestions if you encounter an error page][sim3].

[sim3]: https://github.com/pmean/classes/blob/master/general/suggestions-if-
you-encounter-an-error.md
```

Summary

- What you have learned
  - Chi-squared test of independence
  - Odds ratio and relative risk
  - R code
  - Your homework