# STATS 2107
# Statistical Modelling and Inference II
# Solutions
# Workshop 9:
# $\chi^2$ test of association

Matt Ryan

Semester 2 2022

## Contents

# What and Why?

## The Titanic

```
titanic <- read_csv(here::here("data/titanic.csv"), col_types = cols())
```

The Titanic is a very famous ship that sank in 1912, and we have data on it! This data set contains the following information on 891 passengers:

```
tribble(
  ~"Variable", ~"Info",
  "survived", "Categorical, 1 for yes, 0 for no",
  "pclass", "Categorical, either 1, 2, or 3",
  "sex", "Categorical, either F or M"
) %>%
  knitr::kable()
```

| Variable | Info |
|---|---|
| survived | Categorical, 1 for yes, 0 for no |
| pclass | Categorical, either 1, 2, or 3 |
| sex | Categorical, either F or M |

## The question

Is there a relationship between a passengers class and their survival rate?

How can we go about answering this?

## The story so far

```
tribble(
   ~"", ~"Continuous", ~"Categorical",
  "Continuous", "Linear regression", "t-test",
  "Categorical", "Next year!", "$\\chi^2$ test"
) %>%
  knitr::kable()
```

| | Continuous | Categorical |
|---|---|---|
| Continuous | Linear regression | t-test |
| Categorical | Next year! | $\chi^2$ test |

## $\chi^2$ test

The $\chi^2$ test of association tests for independence between two categorical variables $X$ and $Y$. Suppose

- $X$ has $I$ levels $i = 1, 2, \ldots, I$
- $Y$ has $J$ levels $j = 1, 2, \ldots, J$

## Welcome to the cross tabs

| | $Y_1$ | $Y_2$ | $\cdots$ | $Y_J$ |
|---|---|---|---|---|
| $X_1$ | $N_{11}$ | $N_{12}$ | $\cdots$ | $N_{1J}$ |

| | $Y_1$ | $Y_2$ | $\cdots$ | $Y_J$ |
|---|---|---|---|---|
| $X_2$ | $N_{21}$ | $N_{22}$ | $\cdots$ | $N_{2J}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $X_I$ | $N_{I1}$ | $N_{I2}$ | $\cdots$ | $N_{IJ}$ |

### A null hypothesis

The $\chi^2$ test works under the following hypothesis:

$$H_0: \text{ There is no association between } X \text{ and } Y$$

vs

$$H_a: \text{ There is an association between } X \text{ and } Y$$

### A test statistics

$$\chi = \sum_{i,j} \frac{(N_{ij} - \mathrm{E}[N_{ij}])^2}{\mathrm{E}[N_{ij}]}.$$

Under $H_0$, $\chi \sim \chi^2_{(I-1)(J-1)}$.

## Back to our question

### Is there a relationship between a passengers class and their survival rate?

We test the hypothesis:

$$\begin{aligned} H_0 : &\text{ There is no association between the passenger class} \\ &\text{ and whether they survived.} \end{aligned}$$

vs

$$\begin{aligned} H_a : &\text{ There is an association between the passenger class} \\ &\text{ and whether they survived.} \end{aligned}$$

### What does the data say?

```r
titanic %>%
  count(pclass, survived) %>%
  pivot_wider(names_from = survived, values_from = n) %>%
  knitr::kable()
```

| pclass | 0 | 1 |
|---|---|---|
| 1 | 80 | 136 |
| 2 | 97 | 87 |
| 3 | 372 | 119 |

### Perform a test

First, we need this data in a nice format. We can do it `tidy`, but `base` is better here:

```
(survival_class_crosstabs <- table(titanic$pclass, titanic$survived))
```

```
##
##       0   1
##   1  80 136
##   2  97  87
##   3 372 119
```

### `chisq.test`

```
chisq.test(survival_class_crosstabs)
```

```
##
##  Pearson's Chi-squared test
##
## data:  survival_class_crosstabs
## X-squared = 102.89, df = 2, p-value < 2.2e-16
```

## Your turn

### What to do

1. We rejected the hypothesis test. By looking at the cross tabs, which class do you think is most related to survival outcome? Why do you think this is?

---

**Solutions:**
Looking at the crosstabs table, we see that A LOT of third class passengers died. This is likely because

a. They were lower class citizens.
b. They were trapped at the bottom of the boat.

---

2. Obtain a crosstabs table relating passenger sex to survival rate.

---

**Solutions:**

---

```
(survival_sex_crosstabs <- table(titanic$sex, titanic$survived))
```

```
##
##            0   1
##   female  81 233
##   male   468 109
```

3. Test the hypothesis at the 5% level that there is no association between sex and whether a passenger survived or not.

---

**Solutions:**

```
chisq.test(survival_sex_crosstabs)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  survival_sex_crosstabs
## X-squared = 260.72, df = 1, p-value < 2.2e-16
```

**Solutions:**
Under the null hypothesis, our test statistics follows a $\chi^2_1$ distribution. Our observed statistic is 260.72 with a p-value less than $2.2 \times 10^{-16} < 0.05$. Hence, we reject the null hypothesis and conclude there is an association between passenger sex and whether they survived.

# Why does this all work?

## Consider the $2 \times 2$ case

| | $Y_1$ | $Y_2$ | Total |
|---|---|---|---|
| $X_1$ | $W_1$ | $N_1 - W_1$ | $N_1$ |
| $X_2$ | $W_2$ | $N_2 - W_2$ | $N_2$ |
| Total | $\mathbf{W}$ | $\mathbf{N} - \mathbf{W}$ | $\mathbf{N}$ |

## What happens under the null hypothesis?

Since there is no association between $Y$ and $X$, there is a probability $\pi$ of seeing $Y_1$ and $1 - \pi$ of seeing $Y_2$, no matter what $X$ is. Hence

$$W_i \sim \text{Bin}(N_i, \pi)$$

## Estimating $\pi$ and expected values

Our best guess of the probability of being in $Y_1$ is

$$\hat{\pi} = \frac{W}{N}\,.$$

Hence,

$$\text{E}\left[W_i\right] \approx N_i\hat{\pi} \quad \text{and} \quad \text{var}\left(W_i\right) \approx N_i\hat{\pi}(1 - \hat{\pi})\,.$$

## Table of expected values

### Truth

| | $Y_1$ | $Y_2$ | Total |
|---|---|---|---|
| $X_1$ | $W_1$ | $N_1 - W_1$ | $N_1$ |
| $X_2$ | $W_2$ | $N_2 - W_2$ | $N_2$ |
| Total | $\mathbf{W}$ | $\mathbf{N} - \mathbf{W}$ | $\mathbf{N}$ |

**Expected**

| | $Y_1$ | $Y_2$ |
|---|---|---|
| $X_1$ | $N_1\hat{\pi}$ | $N_1(1-\hat{\pi})$ |
| $X_2$ | $N_2\hat{\pi}$ | $N_2(1-\hat{\pi})$ |

## Our test statisitic

$$\begin{aligned}\chi = &\frac{(W_1 - \mathrm{E}[W_1])^2}{\mathrm{E}[W_1]} + \frac{(N_1 - W_1 - \mathrm{E}[N_1 - W_1])^2}{\mathrm{E}[N_1 - W_1]} \\ &+ \frac{(W_2 - \mathrm{E}[W_2])^2}{\mathrm{E}[W_2]} + \frac{(N_2 - W_2 - \mathrm{E}[N_2 - W_2])^2}{\mathrm{E}[N_2 - W_2]}\end{aligned}$$

## Consider $W_1$

$$\begin{aligned}&\frac{(W_1 - \mathrm{E}[W_1])^2}{\mathrm{E}[W_1]} + \frac{(N_1 - W_1 - \mathrm{E}[N_1 - W_1])^2}{\mathrm{E}[N_1 - W_1]} \\ &= \frac{(W_1 - N_1\hat{\pi})^2}{N_1\hat{\pi}} + \frac{(N_1 - W_1 - N_1(1-\hat{\pi}))^2}{N_1(1-\hat{\pi})} \\ &= \frac{(W_1 - N_1\hat{\pi})^2}{N_1\hat{\pi}} + \frac{(W_1 - \hat{\pi})^2}{N_1(1-\hat{\pi})} \\ &= \frac{(W_1 - N_1\hat{\pi})^2(1-\hat{\pi}) + (W_1 - N_1\hat{\pi})^2\hat{\pi}}{N_1\hat{\pi}(1-\hat{\pi})} \\ &= \frac{(W_1 - N_1\hat{\pi})^2}{N_1\hat{\pi}(1-\hat{\pi})}\end{aligned}$$

## What does this look like?

$$\left(\frac{\text{Random variable} - \text{mean}}{\text{SE}}\right)^2$$

For large $N_1$ CLT implies $W_1 \mathrel{\dot{\underset{\cdot}{\sim}}} N(N_1\hat{\pi}, N_1\hat{\pi}(1-\hat{\pi}))$, hence

$$\frac{(W_1 - N_1\hat{\pi})^2}{N_1\hat{\pi}(1-\hat{\pi})} \mathrel{\dot{\underset{\cdot}{\sim}}} \chi_1^2$$

## What are the degrees of freedom?

$$\chi = \frac{(W_1 - N_1\hat{\pi})^2}{N_1\hat{\pi}(1-\hat{\pi})} + \frac{(W_2 - N_2\hat{\pi})^2}{N_2\hat{\pi}(1-\hat{\pi})},$$

why is $\chi \sim \chi_1^2$?

---

**Solutions:**
Because we have $\hat{\pi}$ in each of the expressions, so this is not the sum of *independent* squared standard normals. This is analogous to the distribution of the sample variance!

---

# Your turn

## What to do

1. Consider the contingency table for sex and survival:

```
survival_sex_crosstabs
```

```
##
##            0   1
##   female  81 233
##   male   468 109
```

Calculate the table of expected counts. Under the null hypothesis, how many males would we expect to survive the sinking of the Titanic?

---

**Solutions:**
Our total $W$ is given by $81 + 468 = 549$. Our total $N$ is given by $81 + 233 + 468 + 109 = 891$. Hence, the probability of not surviving under the null hypothesis is $\hat{\pi} = \frac{W}{N} \approx 0.6162$. Finally, the number of females is $81 + 233 = 314$ and the number of males is $468 + 109 = 577$. Then we get the table of expected counts by:

```
W <- 81 + 468
N <- 81 + 233 + 468 + 109
pi.hat <- W/N
sex_counts <- rowSums(survival_sex_crosstabs)

nonsurvival_column <- sex_counts * pi.hat
survival_column <- sex_counts * (1-pi.hat)

(expect_tab <- cbind(nonsurvival_column, survival_column))
```

```
##        nonsurvival_column survival_column
## female           193.4747        120.5253
## male             355.5253        221.4747
```

---

**Solutions:**
Thus, we would expect approximately 221 males to survive the sinking of the Titanic under the null hypothesis.

---

2. Manually calculate the test statistics for this $\chi^2$ test. Does this agree with what you got before?

---

**Solutions:**
We calculate the test statistic as

```
(X2 <- sum((survival_sex_crosstabs - expect_tab)^2/(expect_tab)))
```

```
## [1] 263.0506
```

---

**Solutions:**
Notice that this does not agree with what we got before. However, if we look at the output from before, it says

"Pearson's Chi-squared test with Yates' continuity correction". If we do not use the continuity correction, we get

```
chisq.test(survival_sex_crosstabs, correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  survival_sex_crosstabs
## X-squared = 263.05, df = 1, p-value < 2.2e-16
```

**Solutions:**
and our result matches.