# STATS 2107
# Statistical Modelling and Inference II

# Workshop 9:
# $\chi^2$ test of association


Matt Ryan

School of Mathematical Sciences, University of Adelaide


Semester 2 2022

# What and Why?

# The Titanic

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:

| 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

| | 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

| | $Y_1$ | $Y_2$ | $\cdots$ | $Y_J$ |
|---|---|---|---|---|
| $X_1$ | $N_{11}$ | $N_{12}$ | $\cdots$ | $N_{1J}$ |
| $X_2$ | $N_{21}$ | $N_{22}$ | $\cdots$ | $N_{2J}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $X_I$ | $N_{I1}$ | $N_{I2}$ | $\cdots$ | $N_{IJ}$ |

# 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:

$H_0$ : There is no association between the passenger class and whether they survived.

vs

$H_a$ : There is an association between the passenger class and whether they survived.

# What does the data say?

| 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?
2. Obtain a crosstabs table relating passenger sex to survival rate.
3. Test the hypothesis at the 5% level that there is no association between sex and whether a passenger survived or not.

# 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,

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

# 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

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

# 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 $W_1 \dot{\underset{\cdot}{\sim}} N(N_1\hat{\pi}, N_1\hat{\pi}(1-\hat{\pi}))$, hence

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

# 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 $\chi \sim \chi_1^2$?

# Your turn

# What to do

1. Consider the contingency table for sex and survival:

```
##
##              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?

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