STAT115: Introduction to Biostatistics

University of Otago

Ōtākou Whakaihu Waka

Lecture 29: Tests of Association for Contingency Tables

Outline

- Contingency table
- Looking at the relationship between two categorical variables
- Investigate approaches to test independence of two categorical variables
- Compare observed and expected counts
- Introduce χ^2 distribution

Data: Passengers on the Titanic

- Data from the adult passengers on the titanic. Two variables:
  - Class: 1st, 2nd, 3rd or crew
  - Survived: yes or no

		survived		
		no	yes	Total
Class	1st	122	197	319
	2nd	167	94	261
	3rd	476	151	627
	Crew	673	212	885
	Total	1438	654	2092

- Do survival probabilities depend on the class?

Big picture

- We have investigated when both variables have two levels (groups)
- Here one of the variables has four levels
  - 1st – 3rd class, crew
- If the survival probabilities vary by class
  - The two variables (class and survival) are related
- If the survival probabilities do not vary by class
  - The two variables (class and survival) are independent
  - Knowing the class of a passenger tells us nothing about their survival probability
  - Recall: Definition of independence when we looked at probability
- Idea: Compare the observed data to what we would expect if two variables were independent

Expected counts

- We can use the margin totals to find the expected counts under independence

$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{\text{table total}}$$

- Work through the Titanic table to understand this

Expected counts: Titanic

$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{\text{table total}} = \frac{319 \times 654}{2092} = 99.73$$

		survived		
		no	yes	Total
Class	1st		99.73	319
	2nd			261
	3rd			627
	Crew			885
	Total	1438	654	2092

- Proportion of passengers who are 1st class
  - $\frac{\text{row total}}{\text{table total}} = \frac{319}{2092} = 0.1525$
  - 15.25% of passengers are 1st class
- If survival and class are independent
  - Expected number is the total number of passengers who survive $\times$ the proportion of passengers who are 1st class
  - Or column total $\times \frac{\text{row total}}{\text{table total}}$

Expected counts: Titanic

$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{\text{table total}} = \frac{627 \times 1438}{2092} = 430.99$$

		survived		
		no	yes	Total
Class	1st		99.73	319
	2nd			261
	3rd	430.99		627
	Crew			885
	Total	1438	654	2092

- Proportion of passengers who are 3rd class
  - $\frac{\text{row total}}{\text{table total}} = \frac{627}{2092} = 0.2997$
  - 29.97% of passengers are 3rd class
- If survival and class are independent
  - Expected number is the total number of passengers who died $\times$ the proportion of passengers who are 3rd class
  - Or column total $\times \frac{\text{row total}}{\text{table total}}$

Expected counts: Titanic

- Put it all together to give observed (black) and expected (blue)

		survived		
		no	yes	Total
Class	1st	122 (219.27)	197 (99.73)	319
	2nd	167 (179.41)	94 (81.59)	261
	3rd	476 (430.99)	151 (196.01)	627
	Crew	673 (608.33)	212 (276.67)	885
	Total	1438	654	2092

- The observed and expected counts will vary: there is natural variation in the data
  - Do they vary more than we would expect if variables are truly independent?

Test for independence/association

- We can look at this with a hypothesis test
  - H_0 : the two variables are independent
  - H_A : the two variables are related (associated)
- The test statistic we will use is

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

  - For each cell we calculate $\frac{(\text{observed}-\text{expected})^2}{\text{expected}}$ and add them up

Test statistic

		survived		
		no	yes	Total
Class	1st	122 (219.27)	197 (99.73)	319
	2nd	167 (179.41)	94 (81.59)	261
	3rd	476 (430.99)	151 (196.01)	627
	Crew	673 (608.33)	212 (276.67)	885
	Total	1438	654	2092

$$X^2 = \frac{(122 - 219.27)^2}{219.27} + \frac{(197 - 99.73)^2}{99.73} + \ldots + \frac{(212 - 276.67)^2}{276.67}$$

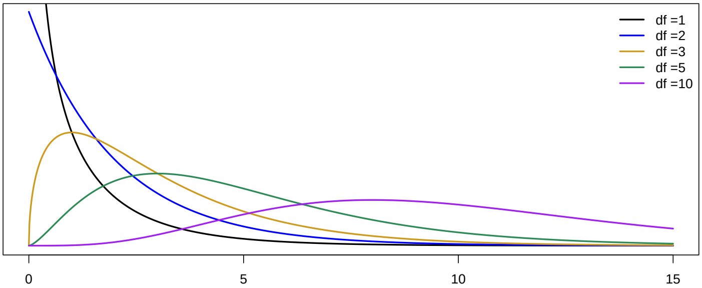
$$= 177.8$$

Test statistic

- If the null hypothesis is true
  - The test statistic, X^2, will be a realisation from a χ^2-distribution with $(R-1) \times (C-1)$ degrees of freedom
    - R is the number of rows; C is the number of columns
- Titanic data: $R = 4$, $C = 2$
  - $df = (4-1) \times (2-1) = 3$

Detour: χ^2-distribution

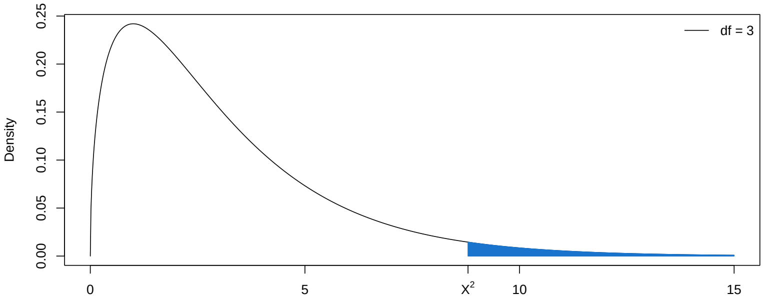
- The χ^2-distribution is a distribution for positive random variables



  - It is asymmetric (positively skewed)
  - It has one parameters: degrees of freedom

Finding a p-value

- An extreme X^2-value is one that is as large, or larger, than that observed
  - Indicative of increased divergence between observed and expected counts



- The p-value (blue area) is given by `1-pchisq(X2, df)`
  - `pchisq(X2, df)` gives probability of a value less than X^2

In R

- Data: each row is an observation
  - Titanic data: each row is a passenger
- Import into R

```r
titanic = read.csv('titanic.csv')
head(titanic)
```

```
##   Class Survived
## 1  Crew      Yes
## 2  Crew      Yes
## 3   2nd       No
## 4   1st      Yes
## 5  Crew      Yes
## 6   3rd       No
```

In R

- We use the `table` function to obtain contingency table

```
titan = table(titanic$Class, titanic$Survived)
```

  - First argument: variable 1 (class of passenger)
  - Second argument: variable 2 (survived: yes / no)

```
titan
```

```
##
##        No Yes
##   1st  122 197
##   2nd  167  94
##   3rd  476 151
##   Crew 673 212
```

```
addmargins(titan)
```

```
##
##          No  Yes  Sum
##   1st   122  197  319
##   2nd   167   94  261
##   3rd   476  151  627
##   Crew  673  212  885
##   Sum  1438  654 2092
```

- The function `addmargins` includes the margins on the table

In R

- The R function `chisq.test` evaluates the test

```
out1 = chisq.test(titan)
out1
##
##  Pearson's Chi-squared test
##
## data:  titan
## X-squared = 177.8, df = 3, p-value <2e-16
```

- The p-value $< \alpha = 0.05$. Observing a test statistic as large as we did is unusual if the two variables were independent
  - Evidence in support of H$_A$: that the variables are not independent

In R

- The `chisq.test` function can return the expected counts

```
out1$expected

##
##            No     Yes
##   1st  219.27  99.726
##   2nd  179.41  81.594
##   3rd  430.99 196.012
##   Crew 608.33 276.668
```

- Still important to know:
  - How to calculate them
  - What they represent (expected counts if variables are independent)

χ^2-test

- If $R = 2$ and $C = 2$: we have a 2×2 contingency table, e.g. smallpox in Boston
  - The χ^2 test is identical to test for difference in proportions
  - $\mathrm{H}_0 : p_1 - p_2 = 0$ and $\mathrm{H}_A : p_1 - p_2 \neq 0$
  - E.g. or smallpox data, following two concepts are the same:
    - Probability of death differs between those innoculated and those note;
    - There is an association between innoculation status (yes/no) and mortality (died/survived)
- The χ^2 test can also be used if both $R > 2$ and $C > 2$
- The χ^2 test is unreliable if any of the expected counts < 5
  - Options for resolving this problem are beyond the scope of course

Testing for Independence for Smallpox Data (in R)

```
SmallpoxTable = rbind(x,n-x)
SmallpoxTable # rbind combines rows to make a matrix (tabular array)

##   [,1] [,2]
## x    6  844
##    238 5136

out1 = chisq.test(SmallpoxTable)
out1

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  SmallpoxTable
## X-squared = 26, df = 1, p-value = 3e-07

out1$expected

##    [,1] [,2]
## x  33.3  817
##   210.7 5163
```

Testing for Independence for Smallpox Data

Results

- P-value is tiny: 3.37×10^{-7}
- Data highly inconsistent with H_0 (i.e. assumption of independence / non-association between variables)
- Conclude there is (very) strong evidence of an associated between mortality and innoculation status
- Test p-value and conclusion mirrors exactly that from last lecture, where we tested equality of mortality probabilities.

Summary

- χ^2 test for independence of contingency table
- Idea: compare observed counts with those expected under independence
- Assess evidence using χ^2-distribution
- Analysis 2×2 contingency table equivalent to comparing proportions