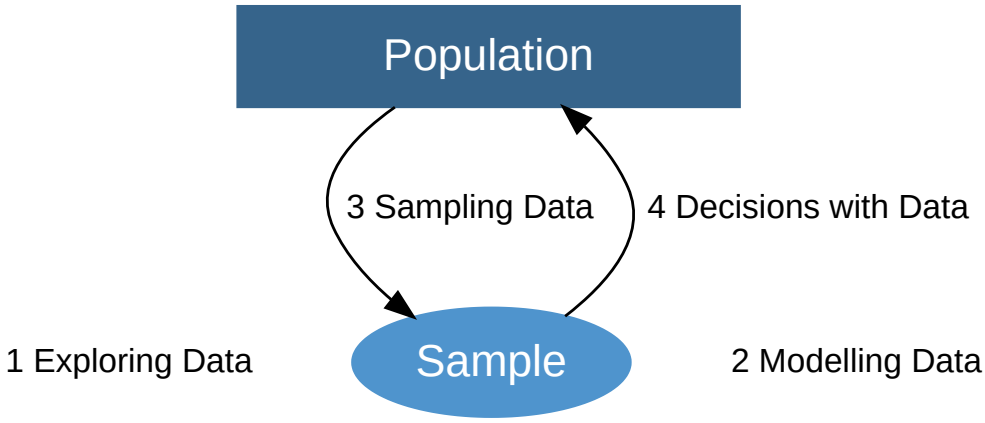
Chi-Square Tests

Decisions with Data | Tests for relationships

Unit Overview



Module4 Decisions with Data

Test for a Proportion

How can we make evidence based decisions? Is an observed result due to chance or something else? How can we test whether a population has a certain proportion?

Tests for a Mean

How can we test whether a population has a certain mean? Or whether 2 populations have the same mean?

Tests for a Relationship [DATA1001/MATH1115]

How can we test whether 2 variables are linearly related? How can we test whether a categorical variable is in certain proportions?

Topic 35 Chi-square tests

Data Stories

The innocent gambler and who drinks Redbull?

The innocent gambler?

Is it possible to test if a gambler has substituted a dice with a weighted dice? Did Mendel (the “father” of genetics) fudge his numbers?

Who drinks Redbull?

“Who buys Red Bull? Red Bull consumers are generally low income, Millennial males in the West Coast. Red Bull consumers are more likely to purchase Red Bull from Gas & Convenience stores with Marlboro cigarettes, Monster energy drinks and Starbucks Frappuccino products also in the basket.”
https://www.infoscout.co/brand/red_bull

Statistical Thinking

With the person next to you:

- What do you think the Australian profile of Red Bull drinkers might be? How would you collect your data, and what type of data would you collect?
- Do you think more Domestic or International students drink Redbull in DATA1001? What sort of data would you need to collect to test your hypothesis?

Tests for Qualitative Data

Tests for qualitative data

- So far, most of our tests have involved **quantitative** data:

Test	Data
1-Sample Z & T Tests	Cyclist endurance times
2-Sample T Test	Redbull effect on heart rate

- However, we did see that the Z Test (and T Tests) can be used for counting & classifying, by modelling a 0-1 box. This is equivalent to a qualitative variable with 2 categories.
- But what happens when we have 3+ categories?

Chi-square tests (χ^2)

- The chi-square test was invented in 1900 by Karl Pearson.
- We pronounce it like the “ki” in “kite”.
- We use the symbol χ^2.

Chi-square tests (χ^2)

The chi-square test is very versatile and is generally used in **3** ways.

- T1: Goodness of Fit: Test a hypothesis about the distribution (model) of a qualitative variable in a population.
- T2: Homogeneity: Test a hypothesis about the distribution of a qualitative variable in several populations.
- T3: Independence: Test a hypothesis about the relationship between two qualitative variables in a population.

The differences between the tests are subtle. We will discuss T1 and T3. For each test, the test statistic is

$$\chi^2 = \text{Sum of } [\frac{(\text{Observed frequency - Expected frequency})^2}{\text{Expected frequency}}]$$

Test for Goodness of Fit of Model (1 variable)

Is the gambler innocent?

A gambler is accused of using a loaded die, although she pleads innocent. The following data shows her last 60 throws.


```
##  [1] 3 4 5 1 3 1 1 5 5 2 3 3 6 4 6 4 6 1 4 6 1 3 5 2 3 4 2 4 1 4 4 5 4 3 6
## [36] 6 6 2 6 4 6 5 6 5 5 6 2 4 6 6 4 1 4 3 2 2 3 5 5 4
```

```
## throws1
##  1  2  3  4  5  6
##  7  7  9 14 10 13
```

Is she innocent?

Comparing observed and expected frequencies

If she is innocent, then we would expect equal numbers of each die face.

Face	Observed frequency	Expected frequency
1	7	10
2	7	10
3	9	10
4	14	10
5	10	10
6	13	10
Total	60	60

The gaps

If she is innocent, then the ‘gap’ between the observed and expected frequencies (O-E)should be small.

Face	Observed frequency (O)	Expected frequency (E)	O-E
1	7	10	-3
2	7	10	-3
3	9	10	-1
4	14	10	4
5	10	10	0
6	13	10	3
Total	60	60	

Looking at the gaps, do you think she’s innocent? What gaps would convince you she’s guilty?

Overall gaps

- We combine all this data by considering the weighted sum of the gaps.

$$\chi^2 = \text{Sum of } [\frac{(\text{Observed frequency - Expected frequency})^2}{\text{Expected frequency}}]$$

- This is the test statistic of the chi-square test.

Calculation of test statistic

Face	O	E	O-E	(O-E)^2/E
1	7	10	-3	0.9
2	7	10	-3	0.9
3	9	10	-1	0.1
4	14	10	4	1.6
5	10	10	0	0
6	13	10	3	0.9
Total	60	60	0	4.4

```
sum(((table(throws1) - 10)^2)/10)
```

```
## [1] 4.4
```

Aside: Note that the sum of the 4th column is (always) 0. Why?

H: Hypotheses

Null Hypothesis

- Assume H_0: The gambler is innocent.
- This is equivalent to drawing 60 times from a box, where the box has equal numbers of 1,2,3,4,5,6 (model).
- This can also be stated as H_0: proportions are equal (p1's= p2's= p3's= p4's= p5's= p6's = $\frac{1}{6}$).

Alternative Hypothesis

- Assume H_1: The gambler is guilty.
- This is equivalent to drawing 60 times from a box, where the box does not have equal numbers of 1,2,3,4,5,6.
- H_1: at least one proportion is different from the stated proportions.

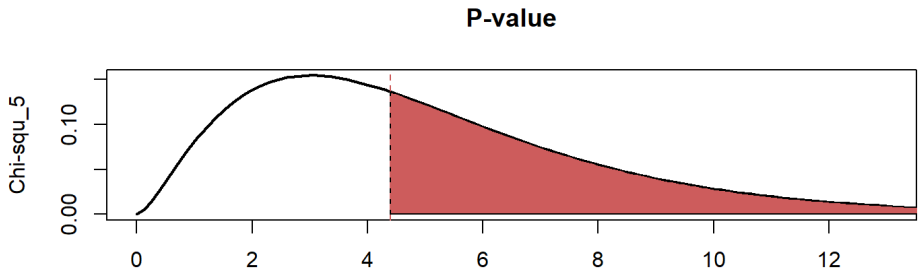
Weighing up the test statistic: T & P

Test Statistic

- As the test statistic is based on the ‘gaps’, **large** values of the test statistic will argue against H_0.
- This can occur if all or some of the gaps are big.

P-value

- So the p-value is the chance of observing 4.4 or more extreme on a chi-squared distribution (with $k-1$ degrees of freedom), where k=number of categories.
- Here: degrees of freedom = 6-1 = 5.




```
pchisq(4.4, 5, lower.tail = F)
```

```
## [1] 0.4933735
```

- As the p-value is so large, the data is consistent with H_0. So this suggests the gambler is innocent.

The reveal

Simulated data in R

The data did in fact come from a 1,2,3,4,5,6 box.

```
set.seed(1)
dice = c(1, 2, 3, 4, 5, 6)
fair = c(1/6, 1/6, 1/6, 1/6, 1/6, 1/6)
throws1 = sample(dice, 60, replace = T, prob = fair)
throws1a = table(throws1)
throws1a
```

```
## throws1
##  1  2  3  4  5  6
##  7  7  9 14 10 13
```

The speedy way to do the test!

```
chisq.test(throws1a, p = c(1/6, 1/6, 1/6, 1/6, 1/6, 1/6))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  throws1a
## X-squared = 4.4, df = 5, p-value = 0.4934
```

New Scenario: What if we had a loaded dice?

```
set.seed(1)
dice = c(1, 2, 3, 4, 5, 6)
loaded = c(1/6 - 1/7, 1/6, 1/6, 1/6, 1/6, 1/6 + 1/7)
throws2 = sample(dice, 60, replace = T, prob = loaded)
throws2a <- table(throws2)
throws2a
```

```
## throws2
##  1  2  3  4  5  6
##  1  8 10 11 15 15
```

```
chisq.test(throws2a, p = c(1/6, 1/6, 1/6, 1/6, 1/6, 1/6))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  throws2a
## X-squared = 13.6, df = 5, p-value = 0.01836
```

Note: The p-value is small so we would reject H_0.

Summary: Chi-square Test for Goodness of Fit

Chi-square Test

Frequency table: Calculate table with observed (O) and expected (E) values.

H: H_0: Model fits data vs H_1: Model doesn't fit data.

A: Expected categories: none are empty, and no more than 20% are < 5. [Cochran's Rule]

T: $\chi^2 = \text{Sum of } [\frac{(\text{Observed frequency - Expected frequency})^2}{\text{Expected frequency}}$

P: Use χ^2_{n-1} curve to find upper tail area. n = number of categories.

C: Retain or Reject H_0.

Test for independence (2 variables)

Test for independence

We now expand our use of the χ^2 test to test for independence.

Statistical Thinking

In a class of 25 students, 15 are Domestic and 10 are International, of whom 9 Domestics smoke and 4 Internationals smoke. Is there a connection between background and smoking?

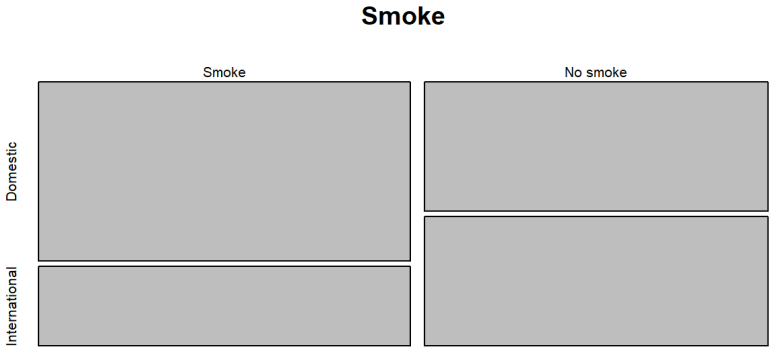
Summarise data in a contingency table

Category	Domestic	International	Total
Smoke	9	4	13
No smoke	6	6	12
Total	15	10	25

Visualising 2 qualitative variables

Visualise data using a mosaic plot

```
Smoke = matrix(c(9, 4, 6, 6), nrow = 2, ncol = 2, byrow = TRUE, dimnames = list(c("Smoke",
    "No smoke"), c("Domestic", "International")))
mosaicplot(Smoke)
```



- It looks like more Domestics seem to smoke compared to Interntationals. But is this significantly more than we would expect?

Test

H: We want to test the hypothesis of independence.

- H_0: Smoking is independent of background (Or: There is no association between smoking and background).
- H_1: Smoking is NOT independent of background (Or: There is an association between smoking and background).

```
chisq.test(Smoke)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  Smoke
## X-squared = 0.32719, df = 1, p-value = 0.5673
```

- T, & C: Given the large p-value, background and smoking preference appear to be independent suggesting that there is no background bias in smokers.

Yates Continuity Correction

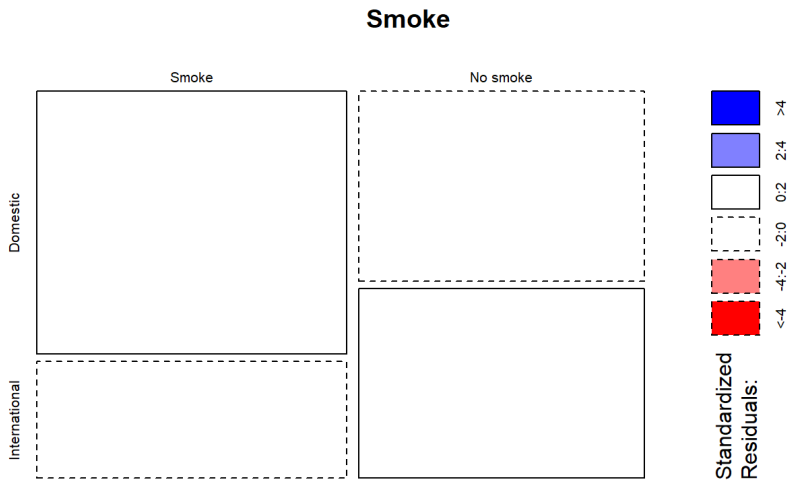
- Note that here the R Test has automatically adopted a Yate's Continuity Correction.
- We can perform the test without it.

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

```
##
##  Pearson's Chi-squared test
##
## data:  Smoke
## X-squared = 0.96154, df = 1, p-value = 0.3268
```

Mosaic plot with standardised residuals

```
mosaicplot(Smoke, shade = T)
```



- The standardised residuals indicate the “gaps” (size and direction) for each individual combination in the 2x2 matrix.
- Here the lack of shading (red and blue) indicates that these “gaps” are not large.
- The dotted and solid lines indicate that slightly more Domestics smoke than expected and slightly less Internationals, but this is not significant.

For an example with more dramatic residuals, try this:

```
mosaicplot(margin.table(Titanic, c(2, 4)), shade = TRUE)
```

Summary: Chi-square Test for Independence

Chi-square Test

Contingency table: Write down table with observed (O) values. R will work out the expected (E) values in `chisq.test()`.

H: H_0: Model fits data vs H_1: Model doesn't fit data.

A: Expected categories: none are empty, and no more than 20% are < 5.

T: $\chi^2 = \text{Sum of}\left[\frac{(\text{Observed frequency - Expected frequency})^2}{\text{Expected frequency}}\right.$

P: Use $\chi^2_{(m-1)\times(n-1)}$ curve to find upper tail area. m and n = number of categories.

C: Retain or Reject H_0.

Note:

- The only difference between the Goodness of Fit (GoF) test and the test for independence is the hypotheses and the way the degrees of freedom are calculated.
- The assumptions and form of the test statistic are the same.

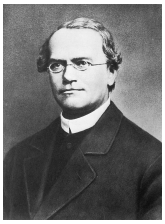
Test	Degrees of freedom	
Goodness of Fit	$n-1$	n = number of categories
Independence	$(m-1)\times(n-1)$	m = number of rows and n = number of columns.

An Example from Genetics

Fisher vs Mendel

- Sir Ronald A Fisher was a British statistician and geneticist.

- Fisher showed that Gregor Mendel's data was fudged.



Fisher vs Mendel χ^2 - test

- Mendel's experiments were all independent, for they involved different sets of plants.
- With independent experiments, the results can be pooled by adding up the separate χ^2 test statistics, and the separate degrees of freedom.
- So for each experiment, Fisher computed the χ^2-statistic, and then he pooled the results.

Hypothesis

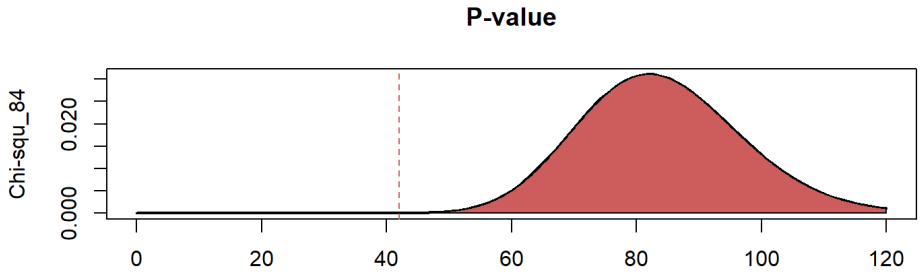
- The null hypothesis: Mendel's data was gathered honestly.
- The alternative hypothesis: Mendel's data was fudged to make the reported frequencies closer to the expected ones.

Test Statistic

- Fisher got a pooled χ^2-value under 42, with 84 degrees of freedom.

P-value

- The upper tail area is almost 1, which seems too good to be true.




```
pchisq(42, 84, lower.tail = F)
```

```
## [1] 0.9999646
```

Conclusion

- Fisher questioned whether Mendel (or Mendel's assistant) had fudged his results, to make the observed frequencies seem closer to the expected ones than chance variation would allow.

Summary

Key Words

Chi-square test; degrees of freedom; categorical & qualitative data; contingencey table; proportions; goodness of fit; homogeneity; independence.

Further Thinking

- If you have a small sample size (cells < 5), and you cannot increase the sample size by, for example, combining categories, then you can use Fisher's exact test `fishers.test` in R.
- In our smoking example, one cell has the value 4 (25%), and we cannot combine cells as there are only two categories. Therefore, we may wish to use Fisher's exact test.

```
fisher.test(Smoke)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  Smoke
## p-value = 0.4283
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   0.3377932 15.7643754
## sample estimates:
## odds ratio
##   2.176249
```

As with the conclusion with the chi-squared test, clearly gender and smoking preference are independent indicating that there is no significant gender bias in smokers.