

Goodness of Fit

PO7001: Quantitative Methods I
Kenneth Benoit

17 November 2010

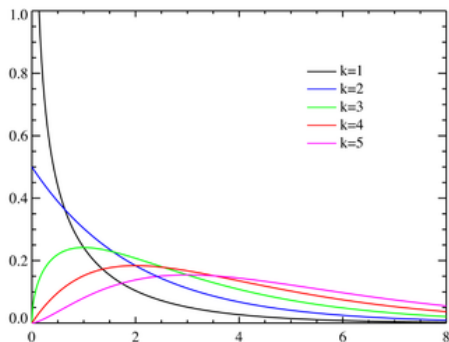
Goodness of fit

- ▶ The **goodness of fit** refers to the extent to which a particular observed set of data confirms to the expected data.
- ▶ Different ways to assess this, depending on the model
- ▶ This week we will assess the fit of categorical data to models of independence (no relationship) and assess statistical significance
- ▶ Later, we will assess the fit of linear models to bivariate data

The Chi-square distribution

- ▶ The sum of squares of normally distributed random variables has a chi-square distribution, also denoted as χ^2

$$f(x; k) = \begin{cases} \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2} & \text{for } x > 0, \\ 0 & \text{for } x \leq 0, \end{cases}$$



Chi-squared test: example

Party	Seats	Recall
Fianna Fail	45.8%	46.7%
Fine Gael	33.1%	20.1%
Labour	10.2%	8.0%
Other	10.9%	25.2%

Table: Recall of the 1997 vote in 2002 Irish National Election Study

Is the recall of vote choice in the 2002 survey significantly different from the outcome of the elections?

Pearson's chi-squared statistic

$$\chi^2 = \sum \frac{(\textit{observed} - \textit{expected})^2}{\textit{expected}} = \sum_{i=1}^k \frac{(Y_i - np_i)^2}{np_i}$$

$$d.f. = k - 1$$

The χ^2 distribution is a good fit if the *expected* cell counts are all ≥ 5

Chi-squared test: example

```
recall <- c(46.7, 20.1, 8, 25.2)
seats <- c(45.8, 33.1, 10.2, 10.9)

seats.prop <- seats / sum(seats)

expected <- 100 * seats.prop

chi2 <- sum((recall - expected)^2/expected)

1 - pchisq(chi2, df=4-1)
```

Chi-squared test: example

```
recall <- c(46.7, 20.1, 8, 25.2)
seats <- c(45.8, 33.1, 10.2, 10.9)

seats.prop <- seats / sum(seats)

chisq.test(recall, p=seats.prop)
```

Chi-squared test: exercise

Verzani 9.2: The table below “contains the results of a poll of 787 registered voters and the actual race results (in percentages of total votes) in the 2003 gubernatorial recall election in California. Is the sample data consistent with the actual results?”

Candidate	Party	Poll amount	Actual
Schwarzenegger	Republican	315	48.6%
Bustamante	Democrat	197	31.5%
McClintock	Republican	141	12.5%
Camejo	Green	39	2.8%
Huffington	Independent	16	0.6%
other		79	4.0%

Table: California gubernatorial recall election

Multivariate Chi-squared test of independence

	Democracy	Dictatorship
No war	40	74
War	3	11

Do dictatorships more often have war on their territory than democracies?

Chi-squared test of independence

We still need:

$$\chi^2 = \sum \frac{(\textit{observed} - \textit{expected})^2}{\textit{expected}}$$

So how to calculate expected values for a cross-table?

Chi-squared test of independence

Basic intuition: if the two variables were independent of each other, the relative proportions should be similar to the marginal distributions.

E.g. the proportion of democracies at war should be similar to the proportion of countries at war.

Since we have two margins, we need to calculate the proportion as:

$$\hat{p}_{democracy,war} = \hat{p}_{democracy} \times \hat{p}_{war}$$

Generally:

$$\text{Expected Frequency} = \frac{r}{n} \cdot \frac{c}{n} \cdot n = \frac{rc}{n}$$

Chi-squared test of independence: example

	Democracy	Dictatorship	
No war	40	74	114
War	3	11	14
	43	85	128

Next step: calculate expected proportions by multiplying marginal proportions.

Chi-squared test of independence: example

	Democracy	Dictatorship	
No war	$114/128 \times 43/128$	$114/128 \times 85/128$	$114/128$
War	$14/128 \times 43/128$	$14/128 \times 85/128$	$14/128$
	$43/128$	$85/128$	1

Next step: calculate this through.

Chi-squared test of independence: example

	Democracy	Dictatorship	
No war	.299	.591	.891
War	.037	.073	.109
	.336	.664	1

Next step: multiply expected proportions by $N = 128$ to get expected counts.

Chi-squared test of independence: example

	Democracy	Dictatorship	
No war	38.3	75.7	114
War	4.7	9.3	14
	43	85	128

Next step: compare expected to observed values.

Chi-squared test of independence: example

		Democracy	Dictatorship
No war	Observed	40	74
	Expected	38.3	75.7
War	Observed	3	11
	Expected	4.7	9.3

Next step: calculate χ^2 .

Chi-squared test of independence

$$\chi^2 = \sum_{i=1}^{n_r} \sum_{j=1}^{n_c} \frac{(Y_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}}$$

$$d.f. = (n_r - 1)(n_c - 1)$$

Yates correction for $2 \times 2 \chi^2$

When the expected frequency is small (< 10) then a $2 \times 2 \chi^2$ test may become inflated. Yates proposed a correction to reduce the inflated χ^2 value.

Yates' correction: Whenever any $E < 10$, subtract .5 from the absolute size of all $|f_o - f_e|$ values (even if only 1).

So for the table

a	b
c	d

$$\text{Yates' } \chi^2 = \frac{N \cdot (|ad - bc| - \frac{N}{2})^2}{(a + b)(c + d)(a + c)(b + d)}$$

Exact tests (for 2×2 tables)

Exact tests offer non-parametric equivalents for tables that do not require any assumptions to be made. They can be used when

1. the overall total $N < 20$, or
2. $20 < N < 40$ and the smallest of the expected $a, b, c, d < 5$

a	b	e
c	d	f
g	h	N

The **exact probability** of any given 2×2 table is:

$$\frac{e!f!g!h!}{N!a!b!c!d!}$$

To test the probability of more extreme tables (less probable), we assess the probability of the observed table against a baseline standard (e.g. .05) of having occurred by chance.

Exact test: example

Gender	Don't Vote	Vote	Total
Female	1	3	4
Male	12	9	21
Total	13	12	25

The **exact probability** of this table, out of all other possible tables with the same margins, is:

$$\frac{4! 21! 13! 12!}{25! 1! 3! 12! 9!} = .2261$$

Exact test: example continued

(a)

0	4	4
13	8	21
13	12	25

$p = 0.0391$

(b)

1	3	4
12	9	21
13	12	25

$p = 0.2261$

(c)

2	2	4
11	10	21
13	12	25

$p = 0.4070$

(d)

3	1	4
10	11	21
13	12	25

$p = 0.2713$

(e)

4	0	4
9	12	21
13	12	25

$p = 0.0565$

Chi-squared test of independence: example

		Democracy	Dictatorship
No war	Observed	40	74
	Expected	38.3	75.7
War	Observed	3	11
	Expected	4.7	9.3

$$\chi^2 = \frac{(40 - 38.3)^2}{38.3} + \frac{(74 - 75.7)^2}{75.7} + \frac{(3 - 4.7)^2}{4.7} + \frac{(11 - 9.3)^2}{9.3} = 1.043$$

$$d.f. = (2 - 1)(2 - 1) = 1$$

Chi-squared test of independence: example

$$\chi^2 = \frac{(40 - 38.3)^2}{38.3} + \frac{(74 - 75.7)^2}{75.7} + \frac{(3 - 4.7)^2}{4.7} + \frac{(11 - 9.3)^2}{9.3} = 1.043$$

$$d.f. = (2 - 1)(2 - 1) = 1$$

Now you can look this up in a χ^2 -table, or you can just use R to calculate the probability that χ^2 has this value or higher:

```
1 - pchisq(1.043, df=1)
```

Chi-squared test of independence: example

	Democracy	Dictatorship
No war	40	74
War	3	11

Do dictatorships more often have war on their territory than democracies?

```
table <- rbind(c(40,74), c(3,11))  
chisq.test(table)
```

Chi-squared test of independence: example

	Democracy	Dictatorship
No war	40	74
War	3	11

Do dictatorships more often have war on their territory than democracies?

```
library(foreign)
aclp <- read.csv("~/Desktop/academic/data/AclpData.csv")
aclp <- aclp[aclp$YEAR == 1980,]
tbl <- table(aclp$WAR, aclp$REG)
chisq.test(tbl)
```

Chi-squared test of independence: exercise

	Most non-free	2	3	4	5	6	Most free
Country < 1945	17	8	6	6	11	9	3
Country > 1945	1	6	9	8	17	21	6

Is there an **association** between the age of the country and the level of freedom, according to the Freedom House scores?

Chi-squared test of independence: exercise

Verzani 9.12: The table below “contains data on the severity of injuries sustained during car crashes. The data is tabulated by whether or not the passenger wore a seat belt. Are the two variables independent?”

		Injury level			
		none	minimal	minor	major
Seat belt	yes	12,813	647	359	42
	no	65,963	4,000	2,642	303

Ordinal tests of association: Mann-Whitney U

To compute the Mann-Whitney U statistic, we only need R_1 , n_1 , and n_2 .

<i>Republicans</i>		<i>Democrats</i>	
X_1	R_1	X_2	R_2
40,000	8	16,000	21
41,000	7	17,000	20
43,000	5	20,000	19
42,000	6	21,000	18
190,000	1	39,000	9
44,000	4	38,000	10
55,000	3	36,000	11
60,000	2	35,000	12
31,000	14	34,000	13
30,000	15	29,000	16
	$\Sigma R_1 = 65$	28,000	17
$n_1 = 10$		$n_2 = 11$	

Ordinal tests of association: Mann-Whitney U

Steps:

1. Add the ranks for the 1st distribution, and solve for U :

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - \sum R_1$$

2. Use U , n_1 , and n_2 to solve for z_u :

$$z_u = \frac{U - n_1 n_2 / 2}{\sqrt{[n_1 n_2 (n_1 + n_2 + 1)] / 12}}$$

3. Compare the obtained value of z_u with the z score for the 95% confidence level, or $|z_{.05}| = 1.96$. If $|z_u| > |z_{.01}|$, then reject H_0 .

Testing whether a variable comes from a Normal distribution

- ▶ Kolmogorov-Smirnov test: `ks.test()`
- ▶ Shapiro-Wilk test: `shapiro.test()`
- ▶ and others exist