DATA2002

Testing for homogeneity

Garth Tarr

THE UNIVERSITY OF SYDNEY

Testing for homogeneity in 2×2 tables

Testing for homogeneity in general tables

Testing for homogeneity in 2×2 tables

COVID treatment

Liu, Lin, Baine, et al. (2020) performed a retrospective, propensity score-matched case-control study to assess the effectiveness of convalescent plasma therapy in 39 patients with severe or life-threatening COVID-19 at The Mount Sinai Hospital in New York City.

	Died	Discharged	Censored	
Plasma treatment	5	28	6	39
No plasma treatent	38	104	14	156
	43	132	20	195

? Is there any evidence that convalescent plasma is an effective treatment for severe COVID-19?

We will focus only on the patients who had an outcome that was able to be observed during the study (died or discharged).

Test of homogeneity

- Suppose that observations are sampled from two independent populations, each of which is categorised according to the same set of outcomes.
- We want to test whether the distribution (proportions) of the outcomes are the same across the different populations.

In our COVID-19 treatment example, we will consider the proportions of patients treated with plasma who died or were discharged and (separately) the proportion of patients who were not treated with plasma who died or were discharged.

	Outcome 1: Died	Outcome 2: Discharged	Row total (fixed)
Population 1: Plasma treatment	p_{11}	p_{12}	$p_{11} + p_{12} = 1$
Population 2: No plasma treatment	p_{21}	p_{22}	$p_{21} + p_{22} = 1$

Under the null hypothesis of **homogeneity** the proportion of patients who died is the same in both populations $p_{11} = p_{21}$, and the proportion of patients who were discharged is the same in both populations $p_{12} = p_{22}$.

Two way contigency table

	Died	Discharged	
Plasma treatment	5	28	33
No plasma treatent	38	104	142
	43	132	175

- A contingency table allows us to tabulate data from multiple categorical variables.
- Contingency tables are heavily used in health, survey research, business intelligence, engineering and scientific research.
- The above table is a two-way a contingency table, specifically a 2×2 contingency table.

Two way contigency table in R

```
library(tidyverse)
dat = read_csv("https://raw.githubusercontent.com/DATA2002/data/master/covidplasma.csv")
dplyr::glimpse(dat)
```

```
## Rows: 195
## Columns: 3
## $ subject   <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 1…
## $ treatment <chr> "Plasma", "Plasma", "Plasma", "Plasma", …
## $ outcome   <chr> "Died", "Died", "Died", "Died", "Died", …
```

```
dat = dat %>% filter(outcome != "Censored") %>%
  mutate(treatment = factor(treatment, levels = c("Plasma", "No plasma")),
         outcome = factor(outcome, levels = c("Died","Discharged")))
```

```
table(dat$treatment, dat$outcome)
```

```
##
##              Died Discharged
##   Plasma        5         28
##   No plasma    38        104
```

```
dat %>% janitor::tabyl(treatment, outcome)
```

```
##  treatment Died Discharged
##     Plasma    5         28
##  No plasma   38        104
```

Notation

In 2×2 contingency tables, for column c and row r let,

$$y_{\bullet c} = \sum_{j=1}^{2} y_{jc} \quad \text{and} \quad y_{r\bullet} = \sum_{j=1}^{2} y_{rj}$$

	Died	Discharged	Row total (fixed)
Plasma treatment	y_{11}	y_{12}	$y_{1\bullet} = n_1$
No plasma treatment	y_{21}	y_{22}	$y_{2\bullet} = n_2$
Column total	$y_{\bullet 1}$	$y_{\bullet 2}$	$n = n_1 + n_2$

Under the null hypothesis of homogeneity we have $p_{11} = p_{21}$ and $p_{12} = p_{22}$ so our best estimate of the proportion in each category is the Column total divided by the overall sample size,

$$\hat{p}_{ij} = \frac{y_{\bullet j}}{n}.$$

Under H_0, the expected counts are $e_{ij} = n_i \hat{p}_{ij} = y_{i\bullet} \frac{y_{\bullet j}}{n}$.

Chi-squared test of homogeneity

With our observed counts and expected counts in each cell, we can construct a chi-squared test for homogeneity,

$$T = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(Y_{ij} - e_{ij})^2}{e_{ij}} \sim \chi_1^2$$

The expected cell counts are,

$$e_{ij} = n_i \hat{p}_{ij} = y_{i\bullet} \frac{y_{\bullet j}}{n} = \frac{\text{Row } i \text{ total} \times \text{Column } j \text{ total}}{\text{Overall total}}$$

? Why 1 degree of freedom for the chi-squared test?

Hypothesis testing workflow

The chi-squared test of homogeneity for a 2×2 contingency table is:

- **Hypothesis:** $H_0\colon\ p_{11} = p_{21}\ \&\ p_{12} = p_{22}$ vs $H_1\colon\ p_{11} \neq p_{21}\ \&\ p_{12} \neq p_{22}$
- **Assumptions:** observations randomly sampled from two independent populations and $e_{ij} = y_{i\bullet} y_{\bullet j}/n \geq 5$.
- **Test statistic:** $T = \displaystyle\sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(Y_{ij} - e_{ij})^2}{e_{ij}}$. Under H_0, $T \sim \chi^2_1$ approx.
- **Observed test statistic:** $t_0 = \displaystyle\sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(y_{ij} - e_{ij})^2}{e_{ij}}$.
- **P-value:** $P(T \geq t_0) \ = \ P(\chi^2_1 \geq t_0)$
- **Decision:** Reject H_0 if the p-value $< \alpha$

COVID treatment

Liu, Lin, Baine, et al. (2020) performed a retrospective, propensity score-matched case-control study to assess the effectiveness of convalescent plasma therapy in 39 patients with severe or life-threatening COVID-19 at The Mount Sinai Hospital in New York City.

	Died	Discharged	
Plasma treatment	5	28	33
No plasma treatent	38	104	142
	43	132	175

? Is there any evidence that convalescent plasma is an effective treatment for severe COVID-19?

y_{ij}	Died	Discharged	Row total
Plasma treatment	5	28	33
No plasma treatent	38	104	142
Column total	43	132	175

e_{ij}	Died	Discharged	Row total
Plasma treatment	$\frac{33 \times 43}{175} = 8.11$	$\frac{33 \times 132}{175} = 24.89$	33
No plasma treatent	$\frac{142 \times 43}{175} = 34.89$	$\frac{142 \times 132}{175} = 107.11$	142
Column total	43	132	175

Observed test statistic:

$$\begin{aligned} t_0 &= \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(y_{ij} - e_{ij})^2}{e_{ij}} \\ &= \frac{(5 - 8.11)^2}{8.11} + \frac{(28 - 24.89)^2}{24.89} + \frac{(38 - 34.89)^2}{34.89} + \frac{(104 - 107.11)^2}{107.11} \\ &= 1.9471 \end{aligned}$$

COVID treatment

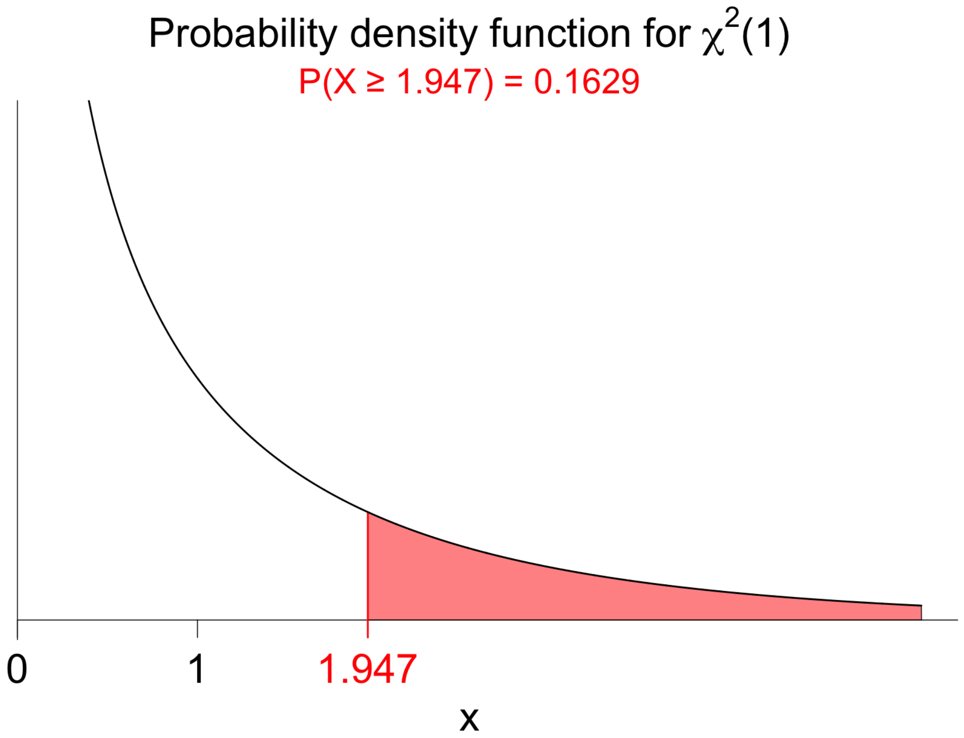
- **Hypothesis:** H_0: $p_{11} = p_{21}$ and $p_{12} = p_{22}$ vs H_1: $p_{11} \neq p_{21}$ and $p_{12} \neq p_{22}$ **or** death and discharge outcomes are homogenous across both the plasma and non-plasma populations.
- **Assumptions:** observations randomly sampled from two independent populations and $e_{ij} = y_{i\bullet} y_{\bullet j}/n \geq 5$.
- **Test statistic:** $T = \displaystyle\sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(Y_{ij} - e_{ij})^2}{e_{ij}}$. Under H_0, $T \sim \chi_1^2$ approx.
- **Observed test statistic:** $t_0 = 1.9471$
- **P-value:** $P(T \geq t_0) = P(\chi_1^2 \geq 1.9471) = 0.16$
- **Decision:** Do not reject H_0 as the p-value is quite large, i.e. there is no evidence to suggest there is a significant difference in the proportion of dead and discharged patients between the plasma and control groups.

```r
tab = table(dat$treatment, dat$outcome)
tab
```

```
##
##                Died Discharged
##    Plasma         5         28
##    No plasma     38        104
```

```r
chisq.test(tab, correct = FALSE)
```

```
##
##      Pearson's Chi-squared test
##
## data:  tab
## X-squared = 1.9471, df = 1, p-value = 0.1629
```



Probability density function for $\chi^2(1)$

$P(X \geq 1.947) = 0.1629$

```
n = sum(tab)
r = c = 2
(row_totals = apply(tab, 1, sum))   # rowSums(tab)
```

```
##    Plasma No plasma
##        33       142
```

```
(col_totals = apply(tab, 2, sum))   # colSums(tab)
```

```
##       Died Discharged
##         43        132
```

```
(rt = matrix(row_totals, nrow = r, ncol = c, byrow = FA
```

```
##      [,1] [,2]
## [1,]   33   33
## [2,]  142  142
```

```
(ct = matrix(col_totals, nrow = r, ncol = c, byrow = TR
```

```
##      [,1] [,2]
## [1,]   43  132
## [2,]   43  132
```

```
(etab = rt * ct / n)
# etab = row_totals%*%t(col_totals)/n
```

```
##              [,1]      [,2]
## [1,]  8.108571  24.89143
## [2,] 34.891429 107.10857
```

```
etab >= 5   # check Eij>=5
```

```
##      [,1] [,2]
## [1,] TRUE TRUE
## [2,] TRUE TRUE
```

```
(t0 = sum((tab - etab)^2/etab))
```

```
## [1] 1.947113
```

```
(p.value = 1 - pchisq(t0, 1))
```

```
## [1] 0.1628983
```

Testing for homogeneity in general tables

Example: Voters

A survey of voter sentiment was conducted in Labor and Liberal to compare the fraction of voters favouring a new tax reform package. Random samples of 100 voters were polled in each of the two parties, with results as follows:

	Approve	Not approve	No comment	Row total
Labor	62	29	9	100
Liberal	47	46	7	100
Column total	109	75	16	200

? Do the data present sufficient evidence to indicate that the fractions of voters favouring the new tax reform package differ in Labor and Liberal?

A general two-way contigency table

	Category 1	Category 2	$\dots$	Category c	Row total (fixed)
Population 1	y_{11}	y_{12}	$\cdots$	y_{1c}	$y_{1\bullet}$
Population 2	y_{21}	y_{22}	$\cdots$	y_{2c}	$y_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
Population r	y_{r1}	y_{r2}	$\cdots$	y_{rc}	$y_{r\bullet}$
Column total	$y_{\bullet 1}$	$y_{\bullet 2}$	$\cdots$	$y_{\bullet c}$	$y_{\bullet\bullet} = n$

- A contingency table allows us to tabulate data from multiple categorical variables.
- We call the above table a two-way a contingency table, specifically a $r \times c$ contingency table.
- There are rc categories and either row or column totals are fixed (therefore, n is also fixed).

Test of homogeneity in general two-way tables

	Category 1	Category 2	$\dots$	Category c	Row total
Population 1	p_{11}	p_{12}	$\dots$	p_{1c}	$p_{1\bullet} = 1$
Population 2	p_{21}	p_{22}	$\dots$	p_{2c}	$p_{2\bullet} = 1$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
Population r	p_{r1}	p_{r2}	$\dots$	p_{rc}	$p_{r\bullet} = 1$

Under the null hypothesis of **homogeneity** $p_{11} = p_{21} = \ldots = p_{r1}$, $p_{12} = p_{22} = \ldots = p_{r2}$, $\ldots$, and $p_{1c} = p_{2c} = \ldots = p_{rc}$.

Test of homogeneity

	Approve	Not approve	No comment	Row total
Labor	y_{11}	y_{12}	y_{13}	$y_{1\bullet} = n_1$
Liberal	y_{21}	y_{22}	y_{23}	$y_{2\bullet} = n_2$
Column total	$y_{\bullet 1}$	$y_{\bullet 2}$	$y_{\bullet 3}$	n

Under the null hypothesis of homogeneity, $p_{11} = p_{21}$, $p_{12} = p_{22}$ and $p_{13} = p_{23}$.

As we don't know p_{ij}, we need to estimate it, $\hat{p}_{ij} = \dfrac{y_{\bullet j}}{n}$.

Under H_0, the expected counts are,

$$e_{ij} = n_i \hat{p}_{ij} = y_{i\bullet} \frac{y_{\bullet j}}{n} = \frac{\text{Row } i \text{ total} \times \text{Column } j \text{ total}}{\text{Overall total}},$$

and the test statistic is,

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(Y_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2_{(r-1)(c-1)}.$$

Degrees of freedom

	Approve	Not approve	No comment	Row total
Labor	y_{11}	y_{12}	y_{13}	$y_{1\bullet} = n_1$
Liberal	y_{21}	y_{22}	y_{23}	$y_{2\bullet} = n_2$
Column total	$y_{\bullet 1}$	$y_{\bullet 2}$	$y_{\bullet 3}$	n

- We need to estimate 3 parameters $\hat{p}_{\bullet 1}$, $\hat{p}_{\bullet 2}$ and $\hat{p}_{\bullet 3}$.
- The degrees of freedom for a 2×3 table is 2.
- More generally the degress of freedom for a $r \ \times \ c$ table is $(r-1)(c-1)$.

Hypothesis testing workflow

The chi-squared test of homogeneity for a $r \times c$ contingency table is:

- **Hypothesis:** H_0: $p_{1j} = p_{2j} = \ldots = p_{rj} \quad j = 1, 2, \ldots, c$ vs H_1: Not all equalities hold.
- **Assumptions:** $e_{ij} = y_{i\bullet} y_{\bullet j}/n \geq 5$ and independent observations sampled from the r populations.
- **Test statistic:** $T = \displaystyle\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(Y_{ij} - e_{ij})^2}{e_{ij}}$. Under H_0, $T \sim \chi^2_{(r-1)(c-1)}$ approx.
- **Observed test statistic:** $t_0 = \displaystyle\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(y_{ij} - e_{ij})^2}{e_{ij}}$
- **P-value:** $P(T \geq t_0) = P(\chi^2_{(r-1)(c-1)} \geq t_0)$
- **Decision:** Reject H_0 if the p-value $< \alpha$

Example: Voters

A survey of voter sentiment was conducted in Labor and Liberal to compare the fraction of voters favouring a new tax reform package. Random samples of 100 voters were polled in each of the two parties, with results as follows:

	Approve	Not approve	No comment	Row total
Labor	62	29	9	100
Liberal	47	46	7	100
Column total	109	75	16	200

? Do the data present sufficient evidence to indicate that the fractions of voters favouring the new tax reform package differ in Labor and Liberal?

Example: Voters

- **Hypothesis:** $H_0\colon\ p_{1j} = p_{2j}$ for $j = 1, 2, 3$ vs H_1: Not all equalities hold.
- **Assumptions:** $e_{ij} = y_{i\bullet} y_{\bullet j}/n \geq 5$.
- **Test statistic:** $T = \displaystyle\sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(Y_{ij} - e_{ij})^2}{e_{ij}}$. Under H_0, $T \sim \chi^2_2$ approx.
- **Test statistic:** $t_0 = \displaystyle\sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(y_{ij} - e_{ij})^2}{e_{ij}} = 6.1676$
- **P-value:** $P(T \geq t_0) = P(\chi^2_2 \geq 6.1676) = 0.046$
- **Decision:** The p-value is less than 0.05, therefore we reject the null hypothesis and conclude that voter preferences about the new tax reform package are not homogenous across Liberal and Labour voters.

```r
y = c(62, 47, 29, 46, 9, 7)
n = sum(y)
c = 3
r = 2
tab = matrix(y, nrow = r, ncol = c)  # default is to fill by column
colnames(tab) = c("Approve", "Not approve", "No comment")
rownames(tab) = c("Labor", "Liberal")
tab
```

```
##         Approve Not approve No comment
## Labor        62          29          9
## Liberal      47          46          7
```

```r
chisq.test(tab, correct = FALSE)
```

```
## 
##     Pearson's Chi-squared test
## 
## data:  tab
## X-squared = 6.1676, df = 2, p-value = 0.04579
```

```
# MARGIN = 1 means apply the sum FUNction
# down rows. Alternative: rowSums(tab)
(yr = apply(tab, MARGIN = 1, FUN = sum))
```

```
##   Labor Liberal
##     100     100
```

```
# MARGIN = 2 means apply the sum FUNction
# across columns. Alternative: colSums(tab)
(yc = apply(tab, MARGIN = 2,FUN = sum))
```

```
##     Approve Not approve  No comment
##         109          75          16
```

```
(yr.mat = matrix(yr, nrow = r, ncol = c,
                 byrow = FALSE))
```

```
##      [,1] [,2] [,3]
## [1,]  100  100  100
## [2,]  100  100  100
```

```
(yc.mat = matrix(yc, nrow = r, ncol = c,
                 byrow = TRUE))
```

```
##      [,1] [,2] [,3]
## [1,]  109   75   16
## [2,]  109   75   16
```

```
# elementwise multiplication and division
(etab = yr.mat * yc.mat / n)
```

```
##      [,1] [,2] [,3]
## [1,] 54.5 37.5    8
## [2,] 54.5 37.5    8
```

```
# could also do matrix multiplication %*%
(etab = yr %*% t(yc) / n)
```

```
##      Approve Not approve No comment
## [1,]    54.5        37.5          8
## [2,]    54.5        37.5          8
```

```
etab >= 5  # check e_ij >= 5
```

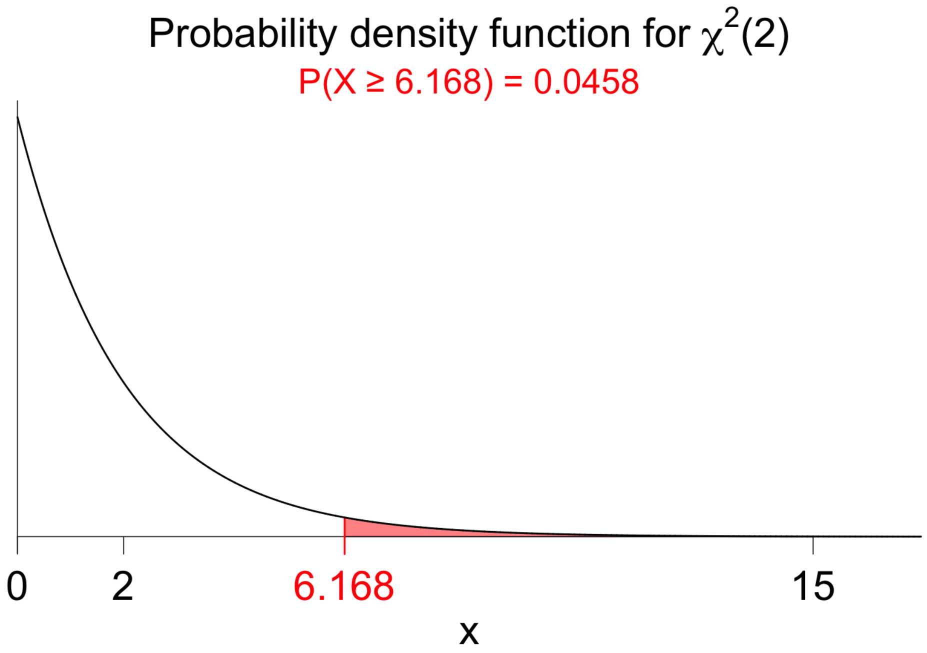
```
##      Approve Not approve No comment
## [1,]    TRUE        TRUE       TRUE
## [2,]    TRUE        TRUE       TRUE
```

```
(t0 = sum((tab - etab)^2/etab))
```

```
## [1] 6.167554
```

```
(p.value = 1 - pchisq(t0, (r - 1) * (c - 1)))
```

```
## [1] 0.04578601
```



References

Franke, T. M., T. Ho, and C. A. Christie (2012). "The Chi-Square Test: Often Used and More Often Misinterpreted". In: *American Journal of Evaluation* 33.3, pp. 448-458. DOI: 10.1177/1098214011426594. URL: https://journals-sagepub-com.ezproxy.library.sydney.edu.au/doi/10.1177/1098214011426594.

Liu, S. T. H., H. Lin, I. Baine, A. Wajnberg, J. P. Gumprecht, F. Rahman, D. Rodriguez, P. Tandon, A. Bassily-Marcus, J. Bander, et al. (2020). "Convalescent plasma treatment of severe COVID-19: a propensity score-matched control study". In: *Nature Medicine* 26.11, pp. 1708-1713. DOI: 10.1038/s41591-020-1088-9.