STAT 226 Lecture 6

Section 2.4 Chi-Squared Tests of Independence

Yibi Huang

Example: Age & Source of News

A question asked in the 2008 General Social Survey is "*Where do you get most of your information about current news events?*" Possible answers included TV, Internet, and newspapers, as well as other possibilities such as radio, family, and friends. The table below summarizes the results by age group.

Age (X)	Source (Y)				Total
	TV	Internet	Newspapers	Other	
18-29	109	92	25	36	262
30-49	272	157	88	63	580
50+	345	59	165	63	632
Total	726	308	278	162	1474

Question: Did the way people get news change with age?

Getting Tablular Data into R

Age	TV	Internet	Newspapers	Other
18-29	109	92	25	36
30-49	272	157	88	63
50+	345	59	165	63

By default, R reads a matrix **by column**.

```
matrix(c(109,272,345,92,157,59,25,88,165,36,63,63),nrow=3)
     [,1] [,2] [,3] [,4]
[1,]  109   92   25   36
[2,]  272  157   88   63
[3,]  345   59  165   63
```

If one prefers entering the counts **by row**, just add `byrow=TRUE`.

```
matrix(c(109,92,25,36,272,157,88,63,345,59,165,63),nrow=3, byrow=TRUE)
     [,1] [,2] [,3] [,4]
[1,]  109   92   25   36
[2,]  272  157   88   63
[3,]  345   59  165   63
```

```
agenews = matrix(c(109,272,345,92,157,59,25,88,165,36,63,63),nrow=3)
dimnames(agenews) = list(
  Age=c("18-29", "30-49", "50+"),
  Source=c("TV", "Internet", "Newspapers", "Other")
)
agenews = as.table(agenews)
agenews
       Source
Age      TV Internet Newspapers Other
  18-29 109       92         25    36
  30-49 272      157         88    63
  50+   345       59        165    63
```

Marginal Totals

```
margin.table(agenews, 1)
Age
18-29 30-49   50+
  262   580   632
```

```
margin.table(agenews, 2)
Source
        TV    Internet Newspapers      Other
       726         308        278        162
```

```
addmargins(agenews)
       Source
Age          TV Internet Newspapers Other  Sum
  18-29   109       92         25    36  262
  30-49   272      157         88    63  580
  50+     345       59        165    63  632
  Sum     726      308        278   162 1474
```

Joint Distributions

Model/Population

Age	TV	Inter-net	News-papers	Other	Total
18-29	π_{11}	π_{12}	π_{13}	π_{14}	π_{1+}
30-49	π_{21}	π_{22}	π_{23}	π_{24}	π_{2+}
50+	π_{31}	π_{32}	π_{33}	π_{34}	π_{3+}
Total	π_{+1}	π_{+2}	π_{+3}	π_{+4}	π_{++}

Data/Sample

Age	TV	Inter-net	News-papers	Other	Total
18-29	n_{11}	n_{12}	n_{13}	n_{14}	n_{1+}
30-49	n_{21}	n_{22}	n_{23}	n_{24}	n_{2+}
50+	n_{31}	n_{32}	n_{33}	n_{34}	n_{3+}
Total	n_{+1}	n_{+2}	n_{+3}	n_{+4}	n_{++}

Estimated joint distribution: $\widehat{\pi}_{ij} = \dfrac{n_{ij}}{n_{++}}$

```
prop.table(agenews)
       Source
Age          TV Internet Newspapers   Other
  18-29 0.07395  0.06242    0.01696 0.02442
  30-49 0.18453  0.10651    0.05970 0.04274
  50+   0.23406  0.04003    0.11194 0.04274
```

Conditional Distributions

Conditional distribution $\mathrm{P}(\text{Source} = j \,|\, \text{Age} = i) = \dfrac{\pi_{ij}}{\pi_{i+}}$ is estimated by $\dfrac{n_{ij}}{n_{i+}}$

```
prop.table(agenews,1)
       Source
Age          TV Internet Newspapers   Other
  18-29 0.41603  0.35115    0.09542 0.13740
  30-49 0.46897  0.27069    0.15172 0.10862
  50+   0.54589  0.09335    0.26108 0.09968
```

Conditional distribution $\mathrm{P}(\text{Age} = i \,|\, \text{Source} = j) = \dfrac{\pi_{ij}}{\pi_{+j}}$ is estimated by $\dfrac{n_{ij}}{n_{+j}}$

```
prop.table(agenews,2)
       Source
Age          TV Internet Newspapers   Other
  18-29 0.15014  0.29870    0.08993 0.22222
  30-49 0.37466  0.50974    0.31655 0.38889
  50+   0.47521  0.19156    0.59353 0.38889
```

Estimation of Marginal Distributions

Marginal distribution of Age $\mathrm{P}(\text{Age} = i) = \pi_{i+}$ is estimated by

$$\widehat{\pi}_{i+} = \frac{n_{i+}}{n_{++}}.$$

```
prop.table(margin.table(agenews,1))
Age
 18-29  30-49    50+
0.1777 0.3935 0.4288
```

Marginal distribution of Source $\mathrm{P}(\text{Source} = j) = \pi_{+j}$ is estimated by

$$\widehat{\pi}_{+j} = \frac{n_{+j}}{n_{++}}.$$

```
prop.table(margin.table(agenews,2))
Source
        TV   Internet Newspapers      Other
    0.4925     0.2090     0.1886     0.1099
```

Independence of Variables (Review)

Recall two variables X & Y are said to be **independent** if the conditional distribution of Y given X doesn't change with X.

$$P(Y = j \mid X = i) = P(Y = j) \quad \text{for all } i, j.$$

As the conditional distribution $P(Y = j \mid X = i)$ is defined to be

$$P(Y = j \mid X = i) = \frac{P(X = i, Y = j)}{P(X = i)}$$

$P(Y = j \mid X = i) = P(Y = j)$ implies

$$P(X = i, Y = j) = P(X = i)P(Y = j).$$

which also implies

$$P(X = i \mid Y = j) = \frac{P(X = i, Y = j)}{P(Y = j)} = \frac{P(X = i)P(Y = j)}{P(Y = j)} = P(X = i)$$

Summary — Independence

Two variables X & Y are **independent** if any of the following is true

- the conditional distribution of Y given X doesn't change with X.

$$\mathrm{P}(Y = j \,|\, X = i) = \mathrm{P}(Y = j) \quad \text{for all } i, j.$$

- the conditional distribution of X given Y doesn't change with Y.

$$\mathrm{P}(X = i \,|\, Y = j) = \mathrm{P}(X = i) \quad \text{for all } i, j.$$

- The joint distribution is the product of the marginal distributions.

$$\mathrm{P}(X = i, Y = j) = \mathrm{P}(X = i)\mathrm{P}(Y = j) \quad \text{for all } i, j.$$

If any of the above is true, the remaining two are also true.

Expected Counts

If X and Y are independent, we know

$$\mathrm{P}(Y = j \mid X = i) = \mathrm{P}(Y = j)$$

We hence expect the estimates of the two sides to be about equal:

$$\frac{n_{ij}}{n_{i+}} \approx \frac{n_{+j}}{n_{++}} \quad \Rightarrow \quad n_{ij} \approx \frac{n_{i+} n_{+j}}{n_{++}}.$$

That is, the expected cell counts under the independence assumption are

$$\boxed{\widehat{\mu}_{ij} = \text{expected cell count} = \frac{\text{row total} \times \text{column total}}{\text{overall total}}}$$

We use the notion $\widehat{\mu}_{ij}$ to represent the expected count.

Expected Counts

The expected counts for the Age and News data are

Age (X)	Source (Y) TV	Internet	Newspapers	Other	Total
18-29	$\frac{262\times726}{1474} = 129.04$	$\frac{262\times308}{1474} = 54.75$	$\frac{262\times278}{1474} = 49.41$	$\frac{262\times162}{1474} = 28.8$	262
30-49	$\frac{580\times726}{1474} = 285.67$	$\frac{580\times308}{1474} = 121.19$	$\frac{580\times278}{1474} = 109.39$	$\frac{580\times162}{1474} = 63.74$	580
50+	$\frac{632\times726}{1474} = 311.28$	$\frac{632\times308}{1474} = 132.06$	$\frac{632\times278}{1474} = 119.2$	$\frac{632\times162}{1474} = 69.46$	632
Total	726	308	278	162	1474

Note the expected cell counts need NOT be **whole numbers**.

Do NOT round the expected counts to integers.

Pearson's Chi-Squared Test Statistic of Independence

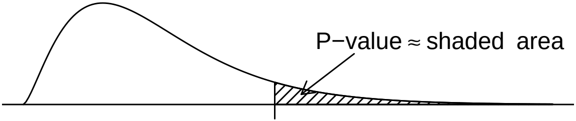
H_0: X and Y are independent vs H_a: X and Y are dependent

Pearson's chi-squared statistic is

$$X^2 = \sum_{ij} \frac{(n_{ij} - \widehat{\mu}_{ij})^2}{\widehat{\mu}_{ij}} = \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Under H_0, X^2 has a large-sample *chi-squared* distribution, with

$$df = (r-1)(c-1), \quad \text{where } \begin{cases} r = \text{number of rows} \\ c = \text{number of columns.} \end{cases}$$



Pearson's X^2-Statistic — Age News Example

The observed counts and the expected counts (in parentheses)

Age (X)	Source (Y): TV	Internet	Newspapers	Other	Total
18-29	109 (129.04)	92 (54.75)	25 (49.41)	36 (28.8)	262
30-49	272 (285.67)	157 (121.19)	88 (109.39)	63 (63.74)	580
50+	345 (311.28)	59 (132.06)	165 (119.2)	63 (69.46)	632
Total	726	308	278	162	1474

The observed value of Pearson's X^2 statistic is

$$X^2 = \frac{(109 - 129.04)^2}{129.04} + \frac{(92 - 54.75)^2}{54.75} + \cdots + \frac{(63 - 69.46)^2}{69.46}$$
$$= 120.0253$$

P-value of Pearson's X^2-Test — Age News Example

The table is 3×4, so

$$df = (r-1)(c-1) = (3-1)(4-1) = 6$$

P-value $= \mathrm{P}(X^2 > 120.0253) = 1.62 \times 10^{-23}$

```
pchisq(120.0253, df=6, lower.tail=FALSE)
[1] 1.61e-23
```

There is strong evidence against H_0:

people in different age groups clearly had different preference in ways of getting news.

Pearson's X^2-Test of Independence in R

```
chisq.test(agenews)


    Pearson's Chi-squared test


data:  agenews
X-squared = 120.025, df = 6, p-value < 2.22e-16
```

Likelihood-Ratio Test (LRT) of Independence

Test statistic

$$
\begin{aligned}
G^2 &= -2\log\left(\frac{\text{maximized likelihood when } \mathrm{H}_0 \text{ true}}{\text{maximized likelihood generally}}\right) \\
&= 2\sum_{ij} n_{ij}\log\left(\frac{n_{ij}}{\widehat{\mu}_{ij}}\right) \\
&= 2\sum_{\text{all cells}} \text{observed} \times \log\left(\frac{\text{observed}}{\text{expected}}\right)
\end{aligned}
$$

Large sample dist. of G^2 under H_0 is also approx. chi-squared $df = (r-1)(c-1)$.

Age (X)	Source (Y) TV	Internet	Newspapers	Other	Total
18-29	109 (129.04)	92 (54.75)	25 (49.41)	36 (28.8)	262
30-49	272 (285.67)	157 (121.19)	88 (109.39)	63 (63.74)	580
50+	345 (311.28)	59 (132.06)	165 (119.2)	63 (69.46)	632
Total	726	308	278	162	1474

The likelihood ratio G^2 statistic is

$$G^2 = 2\left[109\log\left(\frac{109}{129.04}\right) + 92\log\left(\frac{92}{54.75}\right) + \cdots + 63\log\left(\frac{63}{69.46}\right)\right]$$
$$\approx 126.44708951$$

df = $(3-1)(4-1) = 6$, P-value is approx. 7.192×10^{-25}.

```
pchisq(126.4471, df=6, lower.tail=FALSE)
[1] 7.1915915e-25
```

Degrees of Freedom for Likelihood Ratio Test (LRT) in General

df for LRT = # parameters under H_1 – # parameters under H_0

Example (LR test of independence)

H_0: $\pi_{ij} = \pi_{i+}\pi_{+j}$ (Independence)

$$\sum_{ij} \pi_{ij} = 1, \quad \sum_{i} \pi_{i+} = 1, \quad \sum_{j} \pi_{+j} = 1$$

- Under H_1: there are $rc - 1$ free parameters $\{\pi_{ij}\}$
  - If we know $rc - 1$ of the π_{ij}, the last one is known since they must add up to 1.
- Under H_0, the joint probabilities $\{\pi_{ij}\}$ are completely determined by the marginal probabilities $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ since $\pi_{ij} = \pi_{i+}\pi_{+j}$. there are $(r-1)+(c-1)$ free parameters: $(r-1)$ free π_{i+} and $(c-1)$ free π_{+j}.

Thus $df = (rc-1) - [(r-1)+(c-1)] = (r-1)(c-1)$.

```
agenews.chisq = chisq.test(agenews)

names(agenews.chisq)
[1] "statistic" "parameter" "p.value"   "method"    "data.name" "observ
[7] "expected"  "residuals" "stdres"
agenews.chisq$statistic
X-squared
  120.025
agenews.chisq$parameter
df
 6
agenews.chisq$p.value
[1] 1.60979e-23
```

```
agenews.chisq$observed
       Source
Age      TV Internet Newspapers Other
  18-29 109       92         25    36
  30-49 272      157         88    63
  50+   345       59        165    63
agenews.chisq$expected
       Source
Age          TV Internet Newspapers   Other
  18-29 129.045  54.7463    49.4138 28.7951
  30-49 285.672 121.1940   109.3894 63.7449
  50+   311.284 132.0597   119.1967 69.4600
with(agenews.chisq, sum((observed - expected)^2/expected))
[1] 120.025
```

Likelihood Ratio Test Statistic G^2:

```
G2 = with(agenews.chisq, 2*sum(observed*log(observed/expected)))
G2
[1] 126.439
```

Remarks About X^2 and G^2

- If observed count = expected count in all cells, then

$$X^2 = 0, \quad G^2 = 0.$$

- The larger the value of X^2 or G^2, the stronger the evidence against H_0
- The sampling distribution of X^2 converges to χ^2 faster than that of G^2, but X^2 and G^2 are usually similar if the expected counts are (almost) all ≥ 5
- These tests treat X and Y as **nominal**:
  - the order of categories are ignored
  - reordering rows or columns leaves X^2 & G^2 unchanged.
  - More powerful tests for ordinal variables in Ch 6.

Definition of Standardized (or Adjusted) Residuals

$$r_{ij} = \frac{n_{ij} - \widehat{\mu}_{ij}}{\sqrt{\widehat{\mu}_{ij}(1 - \widehat{\pi}_{i+})(1 - \widehat{\pi}_{+j})}}$$

	Source (Y)				
Age (X)	TV	Internet	Newspapers	Other	Total
18-29	109	92	25	36	262
30-49	272	157	88	63	580
50+	345	59	165	63	632
Total	726	308	278	162	1474

$$n_{11} = 109, \quad \widehat{\mu}_{11} = \frac{262 \times 726}{1474} \approx 129.04$$

$$r_{11} = \frac{109 - 129.04}{\sqrt{129.04(1 - \frac{262}{1474`})(1 - \frac{726}{1474})}} \approx -2.732$$

The residuals computed by `chisq.test()` are the unadjusted (raw) Pearson residuals:

$$\frac{n_{ij} - \widehat{\mu}_{ij}}{\sqrt{\widehat{\mu}_{ij}}}$$

not the standardized residuals we defined before.

```
agenews.chisq$residuals
       Source
Age             TV   Internet Newspapers      Other
  18-29 -1.7645379  5.0349193 -3.4730560  1.3426649
  30-49 -0.8088856  3.2524815 -2.0450846 -0.0933001
  50+    1.9110116 -6.3575932  4.1953110 -0.7751101
with(agenews.chisq, (observed - expected)/sqrt(expected))
       Source
Age             TV   Internet Newspapers      Other
  18-29 -1.7645379  5.0349193 -3.4730560  1.3426649
  30-49 -0.8088856  3.2524815 -2.0450846 -0.0933001
  50+    1.9110116 -6.3575932  4.1953110 -0.7751101
```

Standardized Residuals in R

The `stdres` given by `chisq.test()` are
the **standardized residuals** we defined above

$$r_{ij} = \frac{n_{ij} - \widehat{\mu}_{ij}}{\sqrt{\widehat{\mu}_{ij}(1 - \widehat{\pi}_{i+})(1 - \widehat{\pi}_{+j})}}.$$

```
agenews.chisq$stdres
       Source
Age            TV  Internet Newspapers     Other
  18-29 -2.731658  6.242944  -4.251991  1.569448
  30-49 -1.458025  4.695634  -2.915241 -0.126983
  50+    3.549392 -9.457686   6.162260 -1.087021
```

$\boxed{\text{Under } H_0\text{: independence, } r_{ij} \text{ is approx. } N(0,1)}$.

```
agenews.chisq$stdres
       Source
Age            TV  Internet Newspapers     Other
  18-29 -2.731658  6.242944  -4.251991  1.569448
  30-49 -1.458025  4.695634  -2.915241 -0.126983
  50+    3.549392 -9.457686   6.162260 -1.087021
```

- $r_{12} = 6.243$, $r_{22} = 4.696$ both above 3, $r_{32} = -9.458 < -3$
  - Younger (18-29 & 30-49) people were more likely to get news from the Internet than older (50+) people
- $r_{13} = -4.252 < -3$ and $r_{33} = 6.162 > 6$
  - Older (50+) people were more likely to get news from Newspaper than younger people (18-29)

Meaning of Independence

Which of the following means the primary way they get news is associated with their age? **Circle all that apply.**

a. There was a higher percentage of people getting news primarily from the Internet among younger people than among older people.
b. Most people in the 18-29 age group got news primarily from TV and the Internet.
c. Among those who got news primarily from the Newspaper, a majority of them were 50 or older.
d. Those who got news primarily from the internet were generally younger than those who got news primarily by reading a newspaper.
e. The percentage of 18-29 year-olds that got news primarily from the Internet was higher than the percentage that got news primarily from the internet in all age groups

a. P(Source = Internet | Age = young) $\neq$ P(Source = Internet | Age = old), which implies dependence.
b. If the same thing is observed in other age groups, the two variables could be independent
c. If 50+ year-olds were a majority for all sources of news, the two variables could be independent
d. P(Age | Source = Internet) $\neq$ P(Age | Source = Newspaper), which implies dependence.
e. P(Source = Internet | Age = 18-29) $\neq$ P(Source = Internet), which implies dependence.

Pearson's X^2-Test on 2×2 Tables

Pearson's X^2-Test on 2×2 tables is equivalent to the two-sample test for testing H$_0$: $\pi_1 = \pi_2$ with the z-statistic

	$Y = 1$	$Y = 2$	sum
$X = 1$	n_{11}	n_{12}	n_{1+}
$X = 2$	n_{21}	n_{22}	n_{2+}
sum	n_{+1}	n_{+2}	$n = n_{++}$

$$z = \frac{\widehat{\pi}_1 - \widehat{\pi}_2}{\sqrt{\widehat{\pi}(1 - \widehat{\pi})\left(\frac{1}{n_{1+}} + \frac{1}{n_{2+}}\right)}}, \quad \text{where } \widehat{\pi} = \frac{n_{11} + n_{21}}{n_{1+} + n_{2+}} = \frac{n_{+1}}{n}$$

Under H$_0$, z is approx. $N(0, 1)$.

- Pearson's $X^2 = \sum_{\text{all cells}} \frac{(\text{obs} - \text{exp})^2}{\text{exp}}$ is just $(z\text{-statistic})^2$
- The two tests give identical P-values.

Example: Aspirin & Heart Attack (p. 30)

Group	MI: Yes	MI: No	Total	
Placebo	189	10845	11034	$\Rightarrow \widehat{\pi}_1 = 189/11034 \approx 0.0171$
Aspirin	104	10933	11037	$\Rightarrow \widehat{\pi}_2 = 104/11037 \approx 0.0094$

For testing H_0: $\pi_1 = \pi_2$, $\quad \widehat{\pi} = \frac{189+104}{11034+11037} \approx 0.0132$

$$z = \frac{0.0171 - 0.0094}{\sqrt{0.0132(1-0.0132)\left(\frac{1}{11034} + \frac{1}{11037}\right)}} \approx \frac{0.0077}{0.00154} \approx 5.0014$$

2-sided P-value $= 0.00000057$, strong evidence against H_0.

```
chisq.test(matrix(c(189,104,10845,10933),nrow=2), correct=F)


    Pearson's Chi-squared test


data:  matrix(c(189, 104, 10845, 10933), nrow = 2)
X-squared = 25.0139, df = 1, p-value = 0.00000056919
```

$X^2 = 25.0139 \approx (5.0014)^2 = (z\text{-stat})^2$

Tests of Independence Can be Applied to Both Prospective & Retrospective Data

In a retrospective study, we can estimate

$$P(\text{exposed} \mid \text{disease}) \quad \text{and} \quad P(\text{exposed} \mid \text{no disease})$$

but not

$$P(\text{disease} \mid \text{exposed}) \quad \text{and} \quad P(\text{disease} \mid \text{unexposed})$$

which are only estimable in a prospective study.

Nonetheless, the former two are equal if and only if the latter two are equal by the property of independence.

We can thus test the equality of the latter two though neither of them can be estimated.

When is it Safe To Use a Chi-Square Test of Independence?

Both Pearson's X^2 and Likelihood Ratio G^2 require a *large sample* to be applicable.

We can safely use the chi-square test when:

- The observations are independent
- All individual expected counts are 5 or more (≥ 5)

Small-sample tests of independence are available in Section 2.6 (Fisher's exact test).

Infant Malformation and Alcohol Use (Table 2.6 on p.44)

A prospective study in 1987 about maternal drinking (mean number of drinks per day) and whether the baby had congenital sex organ malformations.

Alcohol	Observed Malformation		Expected Malformation	
Consumption	Absent	Present	Absent	Present
0	17,066	48	17,065.14	48.86
< 1	14,464	38	14,460.60	41.40
1-2	788	5	790.74	2.26
3-5	126	1	126.64	0.36
≥ 6	37	1	37.89	0.11

The result of Pearson's X^2-test ($X^2 = 12.1$, P-value $= 0.02$) is NOT consistent with LR test ($G^2 = 6.2$, P-value $= 0.19$).

For this table, neither X^2 nor G^2 has a χ^2 distribution for many cells have very small expected counts.

What's Wrong? (Problem 2.18 on p.60)

Each subject in a sample of 100 men and 100 women is asked to indicate which of the following factors (*one or more*) are responsible for increases in teenage crime:

- A. the increasing gap in income between the rich and poor;
- B. the increase in the percentage of single-parent families;
- C. insufficient time spent by parents with their children.

A cross classification of the responses by gender is

Gender	A	B	C
Men	60	81	75
Women	75	87	86

Can we do the chi-squared test of independence to this 2×3 table?

The Correct Analysis

A

Gender	Yes	No
Men	60	40
Women	75	25

B

Gender	Yes	No
Men	81	19
Women	87	13

C

Gender	Yes	No
Men	75	25
Women	86	14

Example: Jane Austen

People have used relative frequencies of words used as indices of literary style, and statistical techniques applied to word counts have been used in controversies about disputed authorship. An interesting account is given by Morton (1978). When Jane Austen died, she left the novel *Sanditon* unfinished, but she left a summary of the remainder. A highly literate admirer finished the novel, attempting to emulate Austen' style, and the hybrid was published. Morton counted the occurrences of various words in several works: Chapters 1 and 3 of *Sense and Sensibility*, Chapters 1, 2, and 3 of *Emma*, Chapters 1 and 6 of *Sanditon* (all three above were written by Austen); and Chapters 12 and 24 of *Sanditon* (written by her imitator). The counts Morton obtained for 6 words are given in the table on the next page

Example: Jane Austen

Word	Sense and Sensibility	Emma	Sanditon I (by Austen)	Sanditon II (by Imitators)
a	147	186	101	83
an	25	26	11	29
this	32	39	15	15
that	94	105	37	22
with	59	74	28	43
without	18	10	10	4

```
x = matrix(c(147,25,32,94,59,18,186,26,39,105,74,10,
               101,11,15,37,28,10,83,29,15,22,43,4), nrow=6)
dimnames(x) = list(
  word=c("a","an","this","that","with","without"),
  work=c("Sense and Sensitivity","Emma","Sandition I","Sandition II"))
x
         work
word       Sense and Sensitivity Emma Sandition I Sandition II
  a                          147  186         101           83
  an                          25   26          11           29
  this                        32   39          15           15
  that                        94  105          37           22
  with                        59   74          28           43
  without                     18   10          10            4
```

Pearson's X^2 test using the `chisq.test()` command:

```
x.chisq = chisq.test(x)
x.chisq


    Pearson's Chi-squared test


data:  x
X-squared = 45.5775, df = 15, p-value = 0.00006205
```

- What are the null and alternative hypotheses of the test above?
- What do we conclude from the test above?

The hypotheses stated in terms of literary style are as follows.

- H_0: The relative frequencies of the six words (a, an, this, that, with, without) do not differ between the four books
- H_a: Some the four books differs from each other in the relative frequencies of usage of the six words (a, an, this, that, with, without).

Example: Jane Austen (Residuals)

```
round(x.chisq$stdres,2)
         work
word      Sense and Sensitivity  Emma Sandition I Sandition II
  a                       -1.61 -0.19        2.32        -0.08
  an                      -0.74 -1.59       -1.22         4.23
  this                     0.17  0.51       -0.51        -0.37
  that                     2.16  1.67       -1.12        -3.75
  with                    -0.68  0.00       -1.23         2.09
  without                  1.70 -1.71        1.27        -1.19
```

The two largest residuals (4.23 and -3.75) both appear in the work by the imitator, *Sandition II*. We see the imitator used "*an*" a lot more often and "*that*" a lot less often than Austen did in her three works. Most of the residuals for the 3 works by Jane Austen are < 2 and all of them are < 2.5, which indicates little inconsistency in word usage in Austen's work.

Pearson's X^2-test on the first 3 columns (Austen's work) of the table only.

```
x[,1:3]
         work
word      Sense and Sensitivity Emma Sandition I
  a                         147  186         101
  an                         25   26          11
  this                       32   39          15
  that                       94  105          37
  with                       59   74          28
  without                    18   10          10
chisq.test(x[,1:3])

    Pearson's Chi-squared test

data:  x[, 1:3]
X-squared = 12.2714, df = 10, p-value = 0.2673
```

What does the test above tell us?

Pearson's X^2-test of word counts of Austen's 3 novels combined v.s. imitator's

Word	Austen	Imitator
a	434	83
an	62	29
this	86	15
that	236	22
with	161	43
without	38	4

```
x3 = cbind(rowSums(x[,1:3]),x[,4])
chisq.test(x3)


    Pearson's Chi-squared test


data:  x3
X-squared = 32.8096, df = 5, p-value = 0.0000041057
```

What does the test above tell us?