2.4 Testing Independence

Example: Study of 159 depression patients

Depression	Marital Status			Total
	Single	*Married*	*Wid/Div*	
Severe	16	22	19	57
Moderate	29	33	14	76
Mild	9	14	3	26
Total	54	69	36	159

Expected Counts

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

H_0 means that for all (i, j)

$$\mathrm{P}(X = i, Y = j) = \mathrm{P}(X = i)\mathrm{P}(Y = j)$$
$$\pi_{ij} = \pi_{i+}\pi_{+j}$$

Expected frequency is

$$\begin{aligned} \mu_{ij} &= \text{mean of dist. of cell count } n_{ij} \\ &= n\pi_{ij} \\ f &= n\pi_{i+}\pi_{+j} \quad \text{under H}_0 \end{aligned}$$

MLEs under H_0 are

$$\begin{aligned} \widehat{\mu}_{ij} = n\widehat{\pi}_{i+}\widehat{\pi}_{+j} = n\left(\frac{n_{i+}}{n}\right)\left(\frac{n_{+j}}{n}\right) &= \frac{n_{i+}n_{+j}}{n} \\ &= \frac{\text{row total} \times \text{column total}}{\text{overall total}} \end{aligned}$$

$\widehat{\mu}_{ij}$'s are called estimated expected frequencies (or simply expected counts).

Expected Counts for the Depression Example

Depression	**Marital Status**			**Row**
	Single	*Married*	*Wid/Div*	Total
Severe	$\frac{57\times54}{159} = 19.37$	$\frac{57\times69}{159} = 24.74$	$\frac{57\times36}{159} = 12.91$	57
Moderate	$\frac{76\times54}{159} = 25.81$	$\frac{76\times69}{159} = 32.98$	$\frac{76\times36}{159} = 17.21$	76
Mild	$\frac{26\times54}{159} = 8.83$	$\frac{26\times69}{159} = 11.28$	$\frac{26\times36}{159} = 5.89$	26
Column Total	54	69	36	159

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

Pearson's Chi-Squared Test of Independence

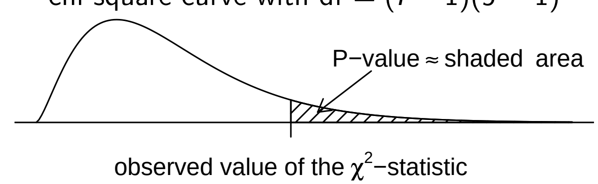
$$X^2 = \sum_{ij} \frac{(n_{ij} - \widehat{\mu}_{ij})^2}{\widehat{\mu}_{ij}} = \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

X^2 has a large-sample chi-squared dist. under H_0, with

$$df = (I - 1)(J - 1)$$

where I = number of rows, J = number of columns.

chi-square-curve with df $= (I-1)(J-1)$



(See p. 343 of text for the Chi-square Table)

Note: chi-squared dist. has mean $= df$, $\sigma = \sqrt{2 \times df}$, is right-skewed and becomes more bell-shaped as df. increases.

Back to the Depression Example

The observed counts and the expected counts (in parentheses):

Depression	Marital Status			
	Single	*Married*	*Wid/Div*	Total
Severe	16 (19.36)	22 (24.74)	19 (12.90)	57
Moderate	29 (25.81)	33 (32.98)	14 (17.21)	76
Mild	9 (8.83)	14 (11.28)	3 (5.89)	26
Total	54	69	36	159

The observed value of the χ^2 test statistic is

$$X^2 = \frac{(16-19.36)^2}{19.36} + \frac{(22-24.74)^2}{24.74} + \ldots + \frac{(3-5.89)^2}{5.89}$$
$$= 6.83$$

The table is 3×3, so

$$df = (I-1)(J-1) = 2 \times 2 = 4$$

p-value $= \mathrm{P}(X^2 > 6.83) = 0.145$

The evidence against H_0 is weak:
it is not strong enough to say the level of depression is associated (dependent) with marital status.

Likelihood-Ratio Test of Independence

Test statistic

$$G^2 = -2\log\left(\frac{\text{maximized likelihood when } H_0 \text{ true}}{\text{maximized likelihood generally}}\right)$$
$$= 2\sum_{ij} n_{ij} \log\left(\frac{n_{ij}}{\widehat{\mu}_{ij}}\right)$$
$$= 2\sum_{\text{all cells}} \text{observed} \times \log\left(\frac{\text{observed}}{\text{expected}}\right)$$

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

Back to the Depression Example

Depression	Marital Status			
	Single	*Married*	*Wid/Div*	Total
Severe	16 (19.36)	22 (24.74)	19 (12.90)	57
Moderate	29 (25.81)	33 (32.98)	14 (17.21)	76
Mild	9 (8.83)	14 (11.28)	3 (5.89)	26
Total	54	69	36	159

The likelihood ratio chi-squared statistic is

$$G^2 = 2\left[16\log\left(\frac{16}{19.36}\right) + 22\log\left(\frac{22}{24.74}\right) + \ldots + 3\log\left(\frac{3}{5.89}\right)\right]$$
$$\approx 6.80$$

df $= 4$, P-value ≈ 0.147

Degrees of Freedom for Likelihood Ratio Test (LRT)

df for LRT = # parameters in general − # parameters under H_0

Example (Chi-squared test of independence)

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

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

- In general there are $IJ - 1$ free parameters $\{\pi_{ij}\}$: If we know $IJ - 1$ of the π_{ij}, then we know the last one because they must add to 1.
- Under H_0, there are $(I-1)+(J-1)$ free parameters: $(I-1)$ free π_{i+} and $(J-1)$ free π_{+j}. They determine the π_{ij} under H_0.

Thus

$$\begin{aligned} df &= (IJ-1) - [(I-1)+(J-1)] \\ &= (I-1)(J-1) \end{aligned}$$

Remarks About X^2 and G^2

- If all $n_{ij} = \widehat{\mu}_{ij}$, then $X^2 = 0$, $G^2 = 0$.
- 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 most $\widehat{\mu}_{ij} > 5$.
- These tests treat X and Y as **nominal**: reordering rows or columns leaves X^2, G^2 unchanged.

  Sec. 2.5 (we skip) presents tests of independence for ordinal variables. We'll introduce more powerful tests for ordinal variable in Ch. 6.

Definition of Standardized (or Adjusted) Residuals

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

Example:

Depression	Marital Status			Total
	Single	*Married*	*Wid/Div*	
Severe	16	22	19	57
Moderate	29	33	14	76
Mild	9	14	3	26
Total	54	69	36	159

$$n_{11} = 16, \quad \widehat{\mu}_{11} = \frac{57 \times 54}{159} \approx 19.36$$

$$r_{11} = \frac{16 - 19.36}{\sqrt{19.36(1 - \frac{57}{159})(1 - \frac{54}{159})}} \approx -1.17$$

Standardized Residuals for Depression Data

Depression	Marital Status		
	Single	*Married*	*Wid/Div*
Severe	−1.17	−0.91	2.41
Moderate	1.07	0.01	−1.22
Mild	0.08	1.18	−1.48

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

As all r_{ij}'s are < 2 or 3 in magnitude, none of the cells show very strong evidence of association.

Getting Tabled Data into R

Depression	Marital Status		
	Single	*Married*	*Wid/Div*
Severe	16	22	19
Moderate	29	33	14
Mild	9	14	3

- By default R reads a matrix **by columns**.

```
> depr = matrix(c(16,29,9,22,33,14,19,14,3), nrow=3)
> dimnames(depr) =
    list(Depression=c("Severe","Moderate","Mild"),
          Marital=c("Single","Married","Wid.Div"))
> depr = as.table(depr)
> depr
          Marital
Depression Single Married Wid.Div
  Severe       16      22      19
  Moderate     29      33      14
  Mild          9      14       3
```


Once the data are saved as a table as above, we can easily convert them to a data frame:

```
> depr.df = as.data.frame(depr)
> depr.df
  Depression Marital Freq
1     Severe  Single   16
2   Moderate  Single   29
3       Mild  Single    9
4     Severe Married   22
5   Moderate Married   33
6       Mild Married   14
7     Severe Wid.Div   19
8   Moderate Wid.Div   14
9       Mild Wid.Div    3
```


The data could also be read from the columns of a text file or a comma-separated (csv) file, which could be created with a text editor or a spreadsheet program. The text or csv file should have a separate row for each combination of factor levels.
Thus a text file `depr.txt` containing

```
Depression Marital Freq
    Severe  Single   16
  Moderate  Single   29
      Mild  Single    9
    Severe Married   22
  Moderate Married   33
      Mild Married   14
    Severe Wid.Div   19
  Moderate Wid.Div   14
      Mild Wid.Div    3
```

can be read into an R dataframe via

```
> depr.df = read.table("depr.txt", header=TRUE)
```


Ungrouped Data

Sometimes the data are ungrouped, like the data file `deprUG.dat`, in which, one row corresponds to the record of one patient.

```
Depression Marital
Moderate Married
Severe Wid.Div
Severe Single
Severe Married
Moderate Married
Mild Single
Severe Single
Severe Married
...
Mild Married
```

Again, we first load it into R dataframe via the command `read.table()`

```
> depr.ug = read.table("deprUG.dat", header=TRUE)
```

Data in a dataframe can be converted to a table using the `xtabs()` or the `table()` function.

```
> xtabs(Freq ~ Depression + Marital, data=depr.df) # Grouped Data
          Marital
Depression Married Single Wid.Div
  Mild          14      9       3
  Moderate      33     29      14
  Severe        22     16      19

> xtabs( ~ Depression + Marital, data=depr.ug)     # Ungrouped Data
          Marital
Depression Married Single Wid.Div
  Mild          14      9       3
  Moderate      33     29      14
  Severe        22     16      19

> table(depr.ug)                              # Ungrouped Data Only
          Marital
Depression Married Single Wid.Div
  Mild          14      9       3
  Moderate      33     29      14
  Severe        22     16      19

> depr = xtabs( ~ Depression + Marital, data=depr.ug)
```

Note the rows and columns might be reordered.

Computations on Tables — Marginal Totals

```
> margin.table(depr, 1)
Depression
  Severe Moderate     Mild
      57       76       26

> margin.table(depr, 2)
Marital
 Single Married Wid.Div
     54      69      36

> addmargins(depr)
          Marital
Depression Single Married Wid.Div Sum
  Severe       16      22      19  57
  Moderate     29      33      14  76
  Mild          9      14       3  26
  Sum          54      69      36 159
```


Computations on Tables — Conditional Distributions

```
> prop.table(depr,1)
          Marital
Depression    Single   Married   Wid.Div
  Severe   0.2807018 0.3859649 0.3333333
  Moderate 0.3815789 0.4342105 0.1842105
  Mild     0.3461538 0.5384615 0.1153846

> prop.table(depr,2)
          Marital
Depression     Single    Married    Wid.Div
  Severe   0.29629630 0.31884058 0.52777778
  Moderate 0.53703704 0.47826087 0.38888889
  Mild     0.16666667 0.20289855 0.08333333

> round(prop.table(depr,2),3)
          Marital
Depression Single Married Wid.Div
  Severe    0.296   0.319   0.528
  Moderate  0.537   0.478   0.389
  Mild      0.167   0.203   0.083
```


Computations on Tables — Chi-Square Test for Indep.

```
> chisq.test(depr)

        Pearson's Chi-squared test

data:  depr
X-squared = 6.8281, df = 4, p-value = 0.1453
```

```
> depr.chisq = chisq.test(depr)

> names(depr.chisq)
[1] "statistic" "parameter" "p.value"   "method"
[5] "data.name" "observed"  "expected"  "residuals"
[9] "stdres"

> depr.chisq$statistic
X-squared
 6.828129

> depr.chisq$parameter
df
 4

> depr.chisq$p.value
[1] 0.1452544
```



```
> depr.chisq$observed
          Marital
Depression Single Married Wid.Div
  Severe       16      22      19
  Moderate     29      33      14
  Mild          9      14       3

> depr.chisq$expected
          Marital
Depression    Single  Married   Wid.Div
  Severe   19.358491 24.73585 12.905660
  Moderate 25.811321 32.98113 17.207547
  Mild      8.830189 11.28302  5.886792

> with(depr.chisq, sum((observed - expected)^2/expected))
[1] 6.828129
```

Likelihood Ratio Test Statistic G^2:

```
> G2 = with(depr.chisq, 2*sum(observed*log(observed/expected)))
> G2
[1] 6.799838
```


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.

```
> depr.chisq$residuals
          Marital
Depression       Single      Married      Wid.Div
  Severe   -0.763323068 -0.550083631  1.696432315
  Moderate  0.627632929  0.003285423 -0.773238674
  Mild      0.057145449  0.808861129 -1.189806121

> with(depr.chisq, (observed - expected)/sqrt(expected))
          Marital
Depression       Single      Married      Wid.Div
  Severe   -0.763323068 -0.550083631  1.696432315
  Moderate  0.627632929  0.003285423 -0.773238674
  Mild      0.057145449  0.808861129 -1.189806121
```


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

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

```
> depr.chisq$stdres
          Marital
Depression       Single      Married      Wid.Div
  Severe   -1.172764405 -0.912860689  2.408136051
  Moderate  1.068978941  0.006044052 -1.216799688
  Mild      0.076887954  1.175503738 -1.479091179
```

2.4.6 Partitioning Chi-squared

Diagnosis	*Drugs*	*No Drug*
Schizophrenia (S)	105	8
Affective disorder (A)	12	2
Neurosis (N)	18	19
Personality disorder (P)	47	52

$df = 3$
$X^2 = 60.88$
$G^2 = 67.27$

Parameters:

Diagnosis	*Drugs*	*No Drug*
S	π_S	$1 - \pi_S$
A	π_A	$1 - \pi_A$
N	π_N	$1 - \pi_N$
P	π_P	$1 - \pi_P$

Test of Independence:

$H_0 : \pi_S = \pi_A = \pi_N = \pi_P$

$H_a : \pi_S, \pi_A, \pi_N, \pi_P$ not all equal

Estimates:

$$\widehat{\pi}_S = 105/(105 + 8) \approx 0.93$$
$$\widehat{\pi}_A = 12/(12 + 2) \approx 0.86$$
$$\widehat{\pi}_N = 18/(18 + 19) \approx 0.49$$
$$\widehat{\pi}_P = 47/(47 + 52) \approx 0.47$$

Partitioning Chi-squared

Testing $\pi_S = \pi_A$:

Diagnosis	*Drugs*	*No Drug*
S	105	8
A	12	2

$df = 1$
$X^2 = 0.89$
$G^2 = 0.75$

Testing $\pi_N = \pi_P$:

Diagnosis	*Drugs*	*No Drug*
N	18	19
P	47	52

$df = 1$
$X^2 = 0.0149$
$G^2 = 0.0149$

Testing $\pi_{S+A} = \pi_{N+P}$:

Diagnosis	*Drugs*	*No Drug*
S + A	117	10
N + P	65	71

$df = 1$
$X^2 = 60.56$
$G^2 = 66.50$

Conclusion: $\pi_S \approx \pi_A$, $\pi_N \approx \pi_P$,
but π_S and π_A are significantly different from π_N and π_P.

Partitioning Chi-squared

Sub-Table	X^2	G^2	df
S v.s. A	0.892	0.753	1
N v.s. P	0.015	0.015	1
(S + A) v.s. (N + P)	60.558	66.500	1
Sum over 3 sub-tables	61.465	67.268	3
Full Table	60.879	67.267	3

- The G^2 for the 3 sub-tables add up to the G^2 for the full table.
- The sum of X^2's for the 3 sub-tables is close but NOT equal to the X^2 for the full table.

Partitioning Chi-squared

- If $X \sim \chi^2_a$ is independent of $Y \sim \chi^2_b$, then $X + Y \sim \chi^2_{a+b}$
- G^2 statistic for testing independence can be partitioned exactly into components representing certain aspects of the association.
- Partition of G^2 is neither unique nor arbitrary.
  - (S v.s. A), (N v.s. P), and (S+A v.s. N+P) is a partition
  - Another partition: (S v.s. A), (S+A v.s. N), and (S+A+N v.s. P)
  - (S v.s. A), (N v.s. P), and (S v.s. N) is NOT a partition
  - Sub-tables in a partition must be **independent** of each other.
  - The general rule of partitioning Chi-squared is beyond the scope of STAT222.
- Partition of X^2 is NOT exact.

Another Partition of Chi-squared

S v.s A:

Diagnosis	*Drugs*	*No Drug*	$df = 1$
S	105	8	$X^2 = 0.89$
A	12	2	$G^2 = 0.75$

(S + A) v.s. N

Diagnosis	*Drugs*	*No Drug*	$df = 1$
S + A	117	10	$X^2 = 37.21$
N	18	19	$G^2 = 31.74$

(S + A + N) v.s. P

Diagnosis	*Drugs*	*No Drug*	$df = 1$
S + A + N	135	29	$X^2 = 35.16$
P	47	52	$G^2 = 34.77$

- For X^2, $0.89 + 37.21 + 35.16 = 73.26 \neq 60.88$
- For G^2, $0.75 + 31.74 + 34.77 = 67.26 = G^2$ for the full table

Not A Partition of Chi-squared

S v.s A:

Diagnosis	*Drugs*	*No Drug*	$df = 1$
S	105	8	$X^2 = 0.89$
A	12	2	$G^2 = 0.75$

N v.s. P

Diagnosis	*Drugs*	*No Drug*	$df = 1$
N	18	19	$X^2 = 0.0149$
P	47	52	$G^2 = 0.0149$

S v.s. N

Diagnosis	*Drugs*	*No Drug*	$df = 1$
S	105	8	$X^2 = 37.01$
N	18	19	$G^2 = 32.37$

For G^2, $0.75 + 0.015 + 32.37 = 33.135 \neq 67.26 = G^2$ for the full table, since the 3 sub-tables are NOT independent of each other.

What's Wrong? (Problem 2.21 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	17

C

Gender	Yes	No
Men	75	25
Women	86	14