STAT22000 Autumn 2013 Lecture 25

Yibi Huang

November 25-27, 2013

9.1 Inference For Two-Way Tables
9.2 Formulas and Models for Two-Way Tables

Two-Way Tables

Two-way tables, also known as **contingency tables**, often formed from counts of two **categorical** variables.

Example: Study of 159 depression patients categorized by

- level of depression (severe, moderate, mild), which is the *row variable* in the table
- marital status (single, married, widowed/divorced), which is the *column variable* in the table

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

- Each combination of values of the two combination defines a *cell*.

Marginal Distributions

$$\textbf{marginal distribution} \text{ of the row variable} = \frac{\text{row total}}{\text{overall total}}$$

$$\textbf{marginal distribution} \text{ of the column variable} = \frac{\text{column total}}{\text{overall total}}$$

Depression	**Marital Status**			**Row Total**
	Single	*Married*	*Wid/Div*	
Severe				$\frac{57}{159} = 0.358$
Moderate				$\frac{76}{159} = 0.478$
Mild				$\frac{26}{159} = 0.164$
Column Total	$\frac{54}{159} = 0.340$	$\frac{69}{159} = 0.434$	$\frac{36}{159} = 0.226$	$\frac{159}{159} = 1.000$

The table tells us, 34.0% of people in the sample are single, 43.4% married, and 22.6% widowed or divorced.

Also, 35.8% of people in the sample are severely depressed, 47.8% moderately, and 16.4% mildly.

Conditional Distribution

The observed **conditional distribution** distribution of the row variable given the column variable is the cell counts divided by the corresponding column totals.

Example. The conditional distributions of level of depression given marital status is

Depression	**Marital Status**			**Row Total**
	Single	*Married*	*Wid/Div*	
Severe	$\frac{16}{54} = 0.296$	$\frac{22}{69} = 0.319$	$\frac{19}{36} = 0.528$	$\frac{57}{159} = 0.358$
Moderate	$\frac{29}{54} = 0.537$	$\frac{33}{69} = 0.478$	$\frac{14}{36} = 0.389$	$\frac{76}{159} = 0.478$
Mild	$\frac{9}{54} = 0.167$	$\frac{14}{69} = 0.203$	$\frac{3}{36} = 0.083$	$\frac{26}{159} = 0.164$
Column Total	$\frac{54}{54} = 1.000$	$\frac{69}{69} = 1.000$	$\frac{36}{36} = 1.000$	$\frac{159}{159} = 1.000$

E.g., 29.6% of the single people in the sample are severely depressed, 53% moderately, and 16.7% mildly.

Conditional Distribution

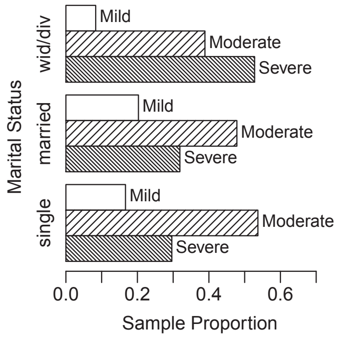
Likewise, the observed **conditional distribution** distribution of the column variable given the row variable is the cell counts divided by the corresponding row totals.

Example. The conditional distributions of level of depression given marital status is

Depression	**Marital Status**			**Row Total**
	Single	*Married*	*Wid/Div*	
Severe	$\frac{16}{57} = 0.281$	$\frac{22}{57} = 0.386$	$\frac{19}{57} = 0.333$	$\frac{57}{57} = 1$
Moderate	$\frac{29}{76} = 0.382$	$\frac{33}{76} = 0.434$	$\frac{14}{76} = 0.184$	$\frac{76}{76} = 1$
Mild	$\frac{9}{26} = 0.346$	$\frac{14}{26} = 0.538$	$\frac{3}{26} = 0.115$	$\frac{26}{26} = 1$
Column Total	$\frac{54}{159} = 0.340$	$\frac{69}{159} = 0.434$	$\frac{36}{159} = 0.226$	$\frac{159}{159} = 1$

E.g., 34.6% of the people with mild depression in the sample are single, 53.8% are married, and 11.5% are widowed or divorced.

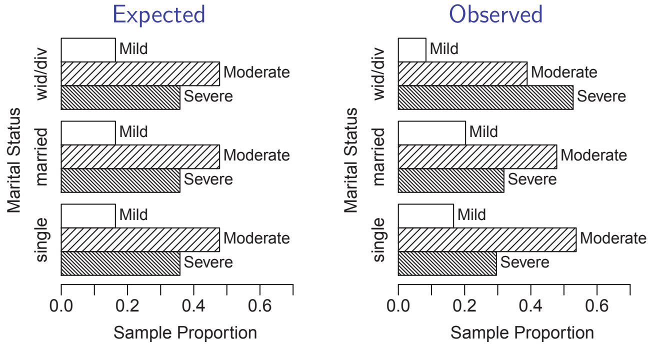
The barplots below show the **conditional distributions** of level of depression given marital status.



Does the level of depression depend on marital status?

- More of widowed or divorced people seems to have severe depression than single or married people

Is this simply chance variation, or the two variables (level of depression, marital status) are indeed associated?



- If level of depression is **independent** of marital status, we expect the conditional distributions to be similar regardless of marital status.
- However, widowed/divorced patients seem to have a different conditional distribution from single or married patients.
- Is the difference statistically significant?

Expected Cell Counts

When the column variable and the row variable are independent, the conditional distribution and of the column given the row, which is

$$\frac{\text{count in a cell}}{\text{row total}}$$

will be the same as the marginal distribution of the column variable, which is

$$\frac{\text{column total}}{\text{overall total}}.$$

That is,

$$\frac{\text{count in a cell}}{\text{row total}} = \frac{\text{column total}}{\text{overall total}}$$

Thus the expected cell counts under the independence assumption are

$$\boxed{\text{expected count in a cell} = \frac{\text{row total} \times \text{column total}}{\text{overall total}}}$$

Expected Counts

The expected counts for the depression and marital status data are

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

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

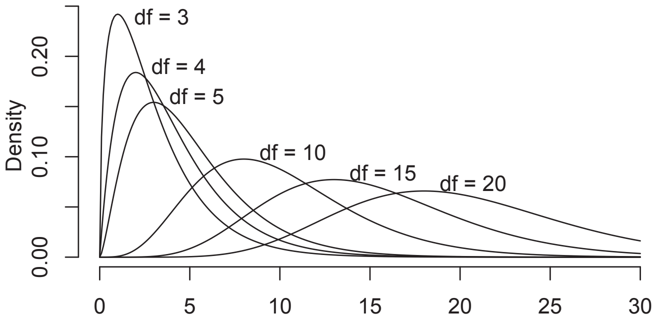
Test for Independence

1. First we state the null and alternative hypotheses:

   H_0: the row and column variables are independent

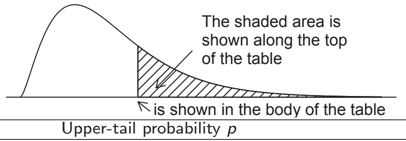
   H_a: the row and column variables are dependent
2. Construct table of expected counts using the formula
   $$\text{expected cell count} = \frac{\text{row total} \times \text{column total}}{\text{overall total}}$$
3. Compare expected with observed counts
4. If H_0 is true, the observed counts and expected counts should be "close"
5. Their differences are measured using a **chi-squared statistic**
   $$\chi^2 = \sum_{\text{all cells}} \frac{(\text{Observed count} - \text{Expected count})^2}{\text{Expected count}}$$
6. The larger the χ^2-statistic, the stronger is the evidence against H_0 (and the more likely to reject H_0)

Chi-Square (χ^2) Distribution



- Just like t-distribution, there is one χ^2 curve with each number of **degree of freedom**
- All χ^2-curves are right-skewed
- The larger the degrees of freedom, the more flatten out and the far to the right the χ^2 curves

The χ^2-curve, with degrees of freedom shown along the left of the table.


The shaded area is shown along the top of the table
is shown in the body of the table

	Upper-tail probability p											
df	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.52	22.11
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42
11	13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.72	26.76	28.73	31.26	33.14
12	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91	34.82
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
30	34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70	62.16
40	45.62	47.27	49.24	51.81	55.76	59.34	60.44	63.69	66.77	69.70	73.40	76.09
50	56.33	58.16	60.35	63.17	67.50	71.42	72.61	76.15	79.49	82.66	86.66	89.56
60	66.98	68.97	71.34	74.40	79.08	83.30	84.58	88.38	91.95	95.34	99.61	102.69
80	88.13	90.41	93.11	96.58	101.88	106.63	108.07	112.33	116.32	120.10	124.84	128.26
100	109.14	111.67	114.66	118.50	124.34	129.56	131.14	135.81	140.17	144.29	149.45	153.17

Exercise

	Upper-tail probability p											
df	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.52	22.11
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42

Q: Find the area under the χ^2-curve with 5 degrees of freedom to the right of (a) 6.63 (b) 9.24 (c) 15.09 .

A: (a) 0.25 (b) 0.1 (c) 0.01 .

Q: Find the area under the χ^2-curve with 10 degrees of freedom to the right of 18.

A: between 0.1 and 0.05.

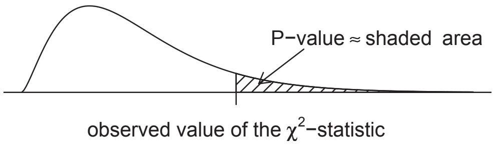
Distribution of the Chi-square Statistic

The χ^2 statistic defined on page 11 is approximately χ^2 distributed with $(r-1)(c-1)$ degrees of freedom, where $r = \#$ of rows, and $c = \#$ of columns in the table.

- e.g., the depression and marital status table has 3 rows and 3 columns, so $df = (3-1)(3-1) = 4$.

The P-value approximately is the area of the upper-tail under the χ^2-curve with $(r-1)(c-1)$ degrees of freedom beyond the chi-square statistic.

χ^2-curve with $(r-1)(c-1)$ degrees of freedom



Why $(r-1)(c-1)$ degrees of freedom?

Imagine a table with r rows and c columns, and fixed row and column totals. (This table has rc cells.)

After we fill in $c-1$ values for a row, we can deduce the c-th value.

Likewise, we only need $r-1$ values to specify a column.

In total, there are $(r-1)(c-1)$ freely varying values.

*	*	*		N_{1+}
*	*	*		N_{2+}
				N_{3+}
N_{+1}	N_{+2}	N_{+3}	N_{+4}	N_{++}

Back to the Depression Example

The table below shows the observed counts and the expected counts (in parentheses)

Depression	Marital Status			Row
	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
Column Total	54	69	36	159

The observed value of the χ^2 test statistic is

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

Back to the Depression Example

The table is 3 × 3, so there are $(r-1)(c-1) = 2 \times 2 = 4$ degrees of freedom.

	Upper-tail probability p											
df	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	.0005
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

From the χ^2-table above, we see that 6.83 is between 6.74 and 7.78. Thus the P-value is between 0.15 and 0.10, not rejecting H_0 at level 0.05.

The evidence is not strong enough to say the level of depression is associated with marital status.

Example: Do Fighter Pilots Father More Daughters? (1)

It is conventional wisdom in military squadrons that pilots tend to father more girls than boys. Snyder (1961) gathered data for military fighter pilots. The gender of the pilot's offspring were tabulate for 3 kinds of flight duty during the month of conception.

Father's Activity	Female offspring	Male offspring	Row Total
Flying Fighters	51	38	89
Flying Transports	14	16	30
Not Flying	38	46	84
Column Total	103	100	203

Example: Do Fighter Pilots Father More Daughters? (2)

Now we have 3 populations of pilot fathers, those flying fighters, those flying transporters, and those not flying.

Let p_f, p_t, and p_n be respectively the proportion who father girls rather than boys in the 3 populations.

If the gender of their children is **independent** of their duty during the month of conception, then we expect

$$p_f = p_t = p_n.$$

So for this example, the hypotheses of the chi-square test of independence can be rephrased as

H_0: $p_f = p_t = p_n$.

H_a: p_f, p_t, p_n are not all equal

Example: Do Fighter Pilots Father More Daughters? (3)

Father's Duty		Gender of Offspring Girl	Boy	Row Total
Flying	obs'd	51	38	89
Fighters	exp'd	$\frac{89\times103}{203} = 45.16$	$\frac{89\times100}{203} = 43.84$	
Flying	obs'd	14	16	30
Transports	exp'd	$\frac{30\times103}{203} = 15.22$	$\frac{30\times100}{203} = 14.78$	
Not	obs'd	38	46	84
Flying	exp'd	$\frac{84\times103}{203} = 42.62$	$\frac{84\times100}{203} = 41.38$	
Column Total		103	100	203

$$\chi^2 = \frac{(51-45.16)^2}{45.16} + \frac{(38-43.84)^2}{43.84} + \frac{(14-15.22)^2}{15.22} + \frac{(16-14.78)^2}{14.78} + \frac{(38-42.62)^2}{42.62} + \frac{(46-41.38)^2}{41.38} = 2.75$$

which has $(3-1)\times(2-1) = 2$ degrees of freedom.
The approximate P-value is greater than 0.25, not rejecting H_0.
The difference can be simply due to chance variation.

Example: Jane Austen & Her Imitator

People have applied statistical techniques to distinguish the literary style of authors and imitators. The following is an example. When Jane Austen died, she left the novel *Sanditon* partially completed, and an admirer finished it. Morton (1978) compared the style of the imitator's work, and two other novels by 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
Total	375	440	202	196

Let's first check Austen's consistency from one work to another. The table below gives the obs'd count and the exp'd count (in paranthesis) in each cell of the table.

Word	Sense and Sensibility	Emma	Sanditon I
a	147 (160.0)	186 (187.8)	101 (86.2)
an	25 (22.9)	26 (26.8)	11 (12.3)
this	32 (31.7)	39 (37.2)	15 (17.1)
that	94 (87.0)	105 (102.1)	37 (46.9)
with	59 (59.4)	74 (69.7)	28 (32.0)
without	18 (14.0)	10 (16.4)	10 (7.5)

$\chi^2 = 12.27$ with $(6-1)\times(3-1) = 10$ degrees of freedom. P-value is greater than 0.25.

The relative frequencies with which Austen used these words did not change from work to work.

	Upper-tail probability p								
df	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	. . .
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	. . .

To compare Austen and her imitator, we can pool all Austen's work together as her style did not change.

Word	Imitator	Austen
a	83 (83.5)	434 (433.5)
an	29 (14.7)	62 (76.3)
this	15 (16.3)	86 (84.7)
that	22 (41.7)	236 (216.3)
with	43 (33.0)	161 (171.0)
without	4 (6.8)	38 (35.2)

$\chi^2 = 32.81$ with df $= (6-1)\times(2-1) = 5$.
P-value is less than 0.0005.

The imitator was not successful in imitating this aspect of Austen's style.

	Upper-tail probability p								
df	. . .	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	.0005
5	. . .	11.07	12.83	13.39	15.09	16.75	18.39	20.52	22.11