STAT 220 Lecture Slides
Testing for Independence in Two-Way Tables

Yibi Huang
Department of Statistics
University of Chicago

Outline

This slide covers Section 6.4 in the text.

- Testing for independence of two categorical variables (two-way tables)

Example: Depression and Marital Status

Study of 159 depression patients categorized by level of depression (severe, moderate, mild), and marital status (single, married, widowed/divorced).

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

Does the conditional distribution of depression level change marital status?

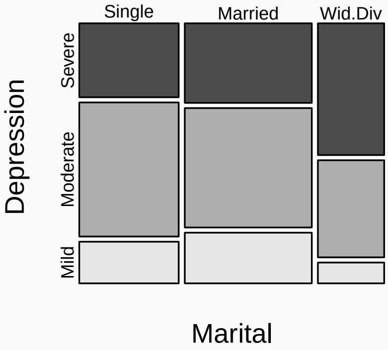
Example: Depression and Marital Status

Recall the conditional distributions of depression level, given marital status can be obtained by dividing cell counts by the corresponding column totals.

Depression	**Marital Status**			
	Single	*Married*	*Wid/Div*	Overall
Severe	$\frac{16}{54} \approx 0.30$	$\frac{22}{69} \approx 0.32$	$\frac{19}{36} \approx 0.53$	$\frac{57}{159} \approx 0.36$
Moderate	$\frac{29}{54} \approx 0.54$	$\frac{33}{69} \approx 0.48$	$\frac{14}{36} \approx 0.39$	$\frac{76}{159} \approx 0.48$
Mild	$\frac{9}{54} \approx 0.17$	$\frac{14}{69} \approx 0.20$	$\frac{3}{36} \approx 0.08$	$\frac{26}{159} \approx 0.16$
Column Total	1	1	1	1

Example: Depression and Marital Status

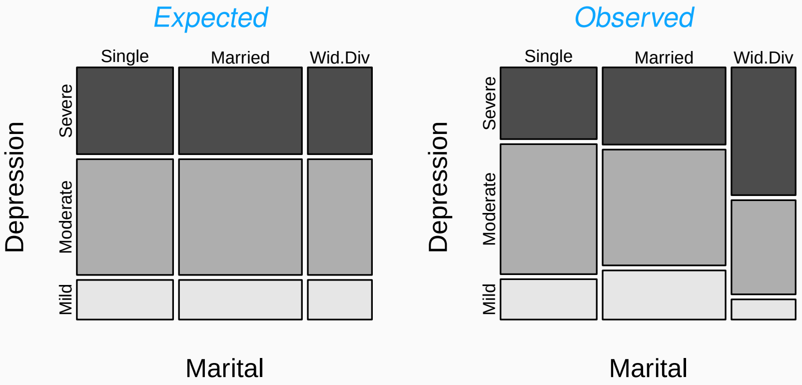
The mosaic plot below shows the **conditional distributions** of depression level, given marital status.



Does the conditional distribution of depression level change 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 (depression level, marital status) are indeed associated?



- Recall if depression level 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,

$$P(\text{column var.} \mid \text{row var.}) = \frac{\text{cell count}}{\text{row total}}$$

should be equal to the marginal distribution of the column variable,

$$P(\text{column var.}) = \frac{\text{column total}}{\text{overall total}}.$$

That is,

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

Thus the expected cell counts under the independence assumption are

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

Expected Counts

The expected counts for the depression and marital status data are

Depression	**Marital Status**			**Row Total**
	Single	*Married*	*Wid/Div*	
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 need NOT be **whole numbers**.

Test for Independence

1. 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

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

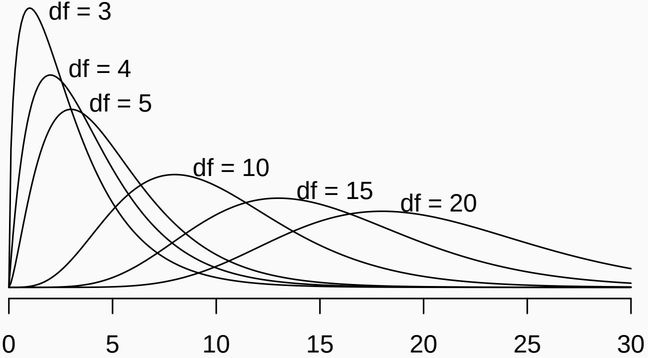
3. If H_0 is true, the observed counts and expected counts should be "close" Their differences are measured using a **chi-squared statistic**

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{Observed count} - \text{Expected count})^2}{\text{Expected count}}$$

4. The larger the χ^2-statistic, the stronger is the evidence against H_0 (and the more likely to reject H_0)

5. How large is the χ^2-statistic usually under H_0?

The Chi-Square (χ^2) Distribution



- There is one curve with each number of **degree of freedom**
- All χ^2-curves are right-skewed
- As the degrees of freedom ↑, the curves flatten out and move off to the right, and become less skewed (more symmetric)
- Expected value $= df$, SD $= \sqrt{df}$

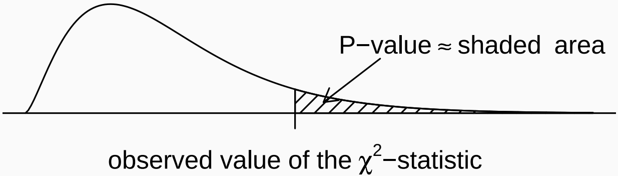
Distribution of the Chi-square Statistic

The χ^2 statistic has an approximate χ^2 distribution 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 $df = (3-1)(3-1) = 4$.

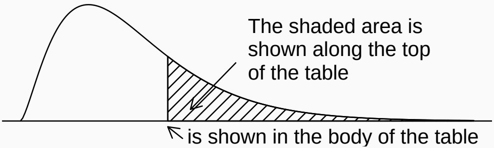
The *P*-value approximately is the area of the upper-tail under the χ^2-curve with $(R-1)(C-1)$ degrees of freedom beyond the chi-square statistic.

χ^2-curve with $(R-1)(C-1)$ degrees of freedom



Chi-Square Probability Table (p.432 in text)

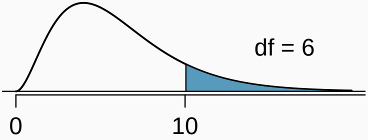
The χ^2-curve, with degrees of freedom shown along the left of the table.



Upper tail	0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df 1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32
8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12
9	10.66	12.24	14.68	16.92	19.68	21.67	23.59	27.88
10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	29.59
11	12.90	14.63	17.28	19.68	22.62	24.72	26.76	31.26
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Practice: Finding p-Value for a χ^2-Statistic

Suppose a χ^2-statistic is 10.3, with df = 6. Find the p-value.



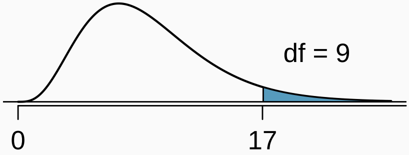
p-value = $P(\chi^2_{df=6} > 10.3)$
is between 0.1 and 0.2

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32

```
> pchisq(10.3, df = 6, lower.tail = FALSE)
[1] 0.1125737
```

Practice: Finding *p*-Value for a χ^2-Statistic

Suppose a χ^2-statistic is 17.56, with df = 9. Find the *p*-value.



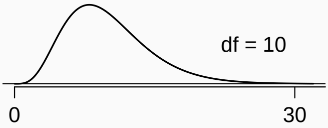
p-value = $P(\chi^2_{df=9} > 17.56)$
is between 0.02 and 0.05

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32
	8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12
	9	10.66	12.24	14.68	16.92	19.68	21.67	23.59	27.88
	10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	29.59

```
> pchisq(17.56, df = 9, lower.tail = FALSE)
[1] 0.04063539
```

Practice: Finding p-Value for a χ^2-Statistic

Suppose a χ^2-statistic is 30.9, with df = 10. Find the p-value



p-value = $P(\chi^2_{df=10} > 30.9)$

is less than 0.001

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001	→
df	8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12	
	9	10.66	12.24	14.68	16.92	19.68	21.67	23.59	27.88	
	10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	29.59	→
	11	12.90	14.63	17.28	19.68	22.62	24.72	26.76	31.26	

```
> pchisq(30.9, df = 10, lower.tail = FALSE)
[1] 0.0006094554
```

Back to the Depression Example

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

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

The observed value of the χ^2 test statistic is

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

Back to the Depression Example

The table is 3×3, so there are $(R-1)(C-1) = (3-1)(3-1) = 4$ degrees of freedom.

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47

From the χ^2-table above, we see that 6.83 is between 5.99 and 7.78. Thus the *P*-value is between 0.2 and 0.1, not rejecting H_0 at level 0.05.

If H_0 is true, we have about 10% chance to get a χ^2-statistic larger than 7.78. Thus, a χ^2-statistic of 6.83 is not too surprising.

No strong evidence to say the level of depression is associated with marital status.

Chi-Square Test in 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
```

Chi-Square Test in R

```
> chisq.test(depr)

        Pearson's Chi-squared test

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

> chisq.test(depr)$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
```

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

We can safely use the chi-square test when:

- The samples are simple random samples (SRS)
- All individual expected counts are 5 or more (≥ 5)

Infant Malformation and Mother's Alcohol Consumption

The table below shows the result of a prospective study in 1987 about maternal drinking (measured as average number of drinks per day) and whether the child had congenital sex organ malformations[1].

Alcohol Consumption	Observed Malformation Absent	Observed Malformation Present	Expected Malformation Absent	Expected Malformation 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

For this table, the chi-square statistic does not have a χ^2 distribution for many cells have very small expected counts.

[1] Mills, J. L. and Graubard, B. I. (1987). Is moderate drinking during pregnancy associated with an increased risk for malformations? *Pediatrics* 80(3), 309¡V314.

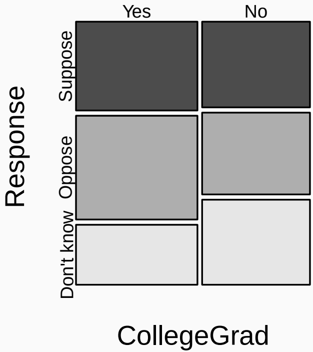
Exercise 6.47 Offshore Drilling

A 2010 survey asked 827 randomly sampled registered voters in California “Do you support? Or do you oppose? Drilling for oil and natural gas off the Coast of California? Or do you not know enough to say?” Below is the distribution of responses, separated based on whether or not the respondent is a college graduate.

	College Grad		
	Yes	No	Total
Support	154	132	286
Oppose	180	126	306
Do not know	104	131	235
Total	438	389	827

Complete a chi-square test for these data to check whether there is a statistically significant difference in responses from college graduates and non-graduates.

Exercise 6.47 Offshore Drilling



Conditional distribution of subjects' opinion on offshore drilling given whether they had a college degree.

	College Grad	
	Yes	No
Support	$\frac{154}{438} = 35.2\%$	$\frac{132}{389} = 33.9\%$
Oppose	$\frac{180}{438} = 41.1\%$	$\frac{126}{389} = 32.4\%$
Do not know	$\frac{104}{438} = 23.7\%$	$\frac{131}{389} = 33.7\%$
Total	1	1

H_0: College graduates and non-graduates did not differ in the distribution of opinion about offshore drilling (support, oppose, do not know)

H_a: The distribution of college graduates' opinion about offshore drilling (support, oppose, do not know) was different from the distribution of non-college graduates.

Exercise 6.47 Offshore Drilling

Expected counts:

	College Grad		
	Yes	No	Total
Support	$\frac{286\times438}{827} = 151.47$	$\frac{286\times389}{827} = 134.53$	286
Oppose	$\frac{306\times438}{827} = 162.07$	$\frac{306\times389}{827} = 143.93$	306
Do not know	$\frac{235\times438}{827} = 124.46$	$\frac{235\times389}{827} = 110.54$	235
Total	438	389	827

Observed counts:

	College Grad	
	Yes	No
Support	154	132
Oppose	180	126
Do not know	104	131

The chi-square statistic is

$$\chi^2 = \frac{(154 - 151.47)^2}{151.47} + \frac{(132 - 134.53)^2}{134.53} + \frac{(180 - 162.07)^2}{162.07} + \frac{(126 - 143.93)^2}{143.93} + \frac{(104 - 124.46)^2}{124.46} + \frac{(131 - 110.54)^2}{110.54}$$

$$\approx 11.46$$

Exercise 6.47 Offshore Drilling

The table is 3×2, so there are $(R-1)(C-1) = (3-1)(2-1) = 2$ degrees of freedom.

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	*0.005*	*0.001*
df	2	2.41	3.22	4.61	5.99	7.82	9.21	*10.60*	*13.82*

From the χ^2-table above, we see that 11.46 is between 10.60 and 13.82. Thus the P-value is between 0.005 and 0.001 (exact P-value ≈ 0.0032).

```
>  pchisq(11.46, df = 2, lower.tail = FALSE)
[1] 0.003247077
```

Conclusion: There is a significant difference in responses from college graduates and non-graduates.

From the data, we can see there are about same percentage

Example: Hormone Therapy for Menopause

The Women's Health Initiative conducted a randomized experiment to see if hormone therapy was helpful for postmenopausal women. The women were randomly assigned to receive the estrogen plus progestin hormone therapy or a placebo. After 5 years, 107 of the 8506 on the hormone therapy developed cancer and 88 of the 8102 in the placebo group developed cancer. Is this a significant difference?

	Cancer	No Cancer	Total
Hormone	107	8399	8506
Placebo	88	8014	8102
Total	195	16413	16608

Example: Hormone Therapy for Menopause

H_0: The two variables (Hormone or placebo and Cancer or not) are independent.
This implies that the hormone group and placebo group had the same rate of developing cancer.

$$p_{\text{hormone}} = p_{\text{placebo}}$$

H_a: The two variables are not independent.
This implies that the two group had the different rates of developing cancer.

$$p_{\text{hormone}} \neq p_{\text{placebo}}$$

Example: Hormone Therapy for Menopause

Expected counts:

	cancer	no cancer	total
hormone	$\frac{8506\times195}{16608}=99.87$	$\frac{8506\times16413}{16608}=8406.13$	8506
placebo	$\frac{8102\times195}{16608}=95.13$	$\frac{8102\times16413}{16608}=8006.87$	8102
total	195	16413	16608

Observed counts:

	cancer	no cancer
hormone	107	8399
placebo	88	8014

The chi-square statistic is

$$\chi^2 = \frac{(107-99.87)^2}{99.87} + \frac{(8399-8406.13)^2}{8406.13} + \frac{(88-95.13)^2}{95.13} + \frac{(8014-8006.87)^2}{8006.87}$$

$$\approx 1.0553 \qquad \text{with df } = (R-1)(C-1) = (2-1)(2-1) = 1.$$

Upper tail	0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df 1	*1.07*	1.64	2.71	3.84	5.41	6.63	7.88	10.83

The *P*-value is greater than 0.3 (exact *P*-value ≈ 0.304). The two groups didn't have significant different rates in developing cancer.

```
> pchisq(1.0553, df = 1, lower.tail = FALSE)
[1] 0.3042896
```

Example: Hormone Therapy for Menopause

Alternatively, one can perform a two sample z-test for proportions

$$H_0: \; p_{\text{hormone}} = p_{\text{placebo}} \quad \text{v.s.} \quad H_a: \; p_{\text{hormone}} \neq p_{\text{placebo}}$$

Under H_0, the pooled sample proportion is $\widehat{p} = \frac{107+88}{8506+8102} = \frac{195}{16608}$.

The test statistic is

$$z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\frac{107}{8506} - \frac{88}{8102}}{\sqrt{\frac{195}{16608}(1 - \frac{195}{16608})\left(\frac{1}{8506} + \frac{1}{8102}\right)}} \approx 1.02728$$

The 2-sided P-value is $\approx 2P(Z > 1.03) = 2(1 - 0.8485) = 0.303$.

```
> 2*pnorm(1.02728, lower.tail = FALSE)
[1] 0.3042886
```

Observe the chi-square test and the two sample z-test for proportions give identical P-values, and hence will reach identical conclusions. This is not an coincidence.

Chi-square Test for 2×2 Tables

In fact, chi-square test for a 2×2 table is equivalent to a two-sided two-sample z-test for proportions

$$H_0 : p_1 = p_2 \quad \text{v.s.} \quad H_a : p_1 \neq p_2$$

	observed			expected	
	success	failure	total	success	failure
sample 1	X_1	$n_1 - X_1$	n_1	$n_1 \hat{p}$	$n_1(1 - \hat{p})$
sample 2	X_2	$n_2 - X_2$	n_2	$n_2 \hat{p}$	$n_2(1 - \hat{p})$
total	$X_1 + X_2$	$n_1 + n_2 - X_1 - X_2$	$n_1 + n_2$	where $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$	

One can show that

$$\chi^2\text{-statistic} = \sum \frac{(O - E)^2}{E} = \left(\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{p}(1 - \hat{p})}} \right)^2 = (z\text{-statistic})^2$$

where $\hat{p}_1 = X_1 / n_1$ and $\hat{p}_2 = X_2 / n_2$.

And the two tests give identical p-values.

Chi-square Test for 2×2 Tables — Proof (May Skip)

	Observed		Expected	
	success	failure	success	failure
sample 1	X_1	$n_1 - X_1$	$n_1\hat{p}$	$n_1(1-\hat{p})$
sample 2	X_2	$n_2 - X_2$	$n_2\hat{p}$	$n_2(1-\hat{p})$

$$
\begin{aligned}
\chi^2 &= \sum \frac{(O-E)^2}{E} \\
&= \frac{(X_1 - n_1\hat{p})^2}{n_1\hat{p}} + \frac{(n_1 - X_1 - n_1(1-\hat{p}))^2}{n_1(1-\hat{p})} + \frac{(X_2 - n_2\hat{p})^2}{n_2\hat{p}} + \frac{(n_2 - X_2 - n_2(1-\hat{p}))^2}{n_2(1-\hat{p})} \\
&= \frac{(X_1 - n_1\hat{p})^2}{n_1\hat{p}} + \frac{(X_1 - n_1\hat{p})^2}{n_1(1-\hat{p})} + \frac{(X_2 - n_2\hat{p})^2}{n_2\hat{p}} + \frac{(X_2 - n_2\hat{p})^2}{n_2(1-\hat{p})} \\
&= \frac{(X_1 - n_1\hat{p})^2}{n_1}\left(\frac{1}{\hat{p}} + \frac{1}{1-\hat{p}}\right) + \frac{(X_2 - n_2\hat{p})^2}{n_2}\left(\frac{1}{\hat{p}} + \frac{1}{1-\hat{p}}\right) \\
&= \frac{(X_1 - n_1\hat{p})^2}{n_1\hat{p}(1-\hat{p})} + \frac{(X_2 - n_2\hat{p})^2}{n_2\hat{p}(1-\hat{p})} \quad \text{since } \frac{1}{\hat{p}} + \frac{1}{1-\hat{p}} = \frac{(1-\hat{p}) + \hat{p}}{\hat{p}(1-\hat{p})} = \frac{1}{\hat{p}(1-\hat{p})}.
\end{aligned}
$$

Chi-square Test for 2×2 Tables — Proof (May Skip)

Observe that

$$X_1 - n_1\hat{p} = X_1 - n_1\left(\frac{X_1 + X_2}{n_1 + n_2}\right) = \frac{n_2 X_1 - n_1 X_2}{n_1 + n_2} = \frac{X_1/n_1 - X_2/n_2}{\frac{1}{n_1} + \frac{1}{n_2}} = \frac{\hat{p}_1 - \hat{p}_2}{\frac{1}{n_1} + \frac{1}{n_2}}$$

where $\hat{p}_1 = X_1/n_1$ and $\hat{p}_2 = X_2/n_2$.

Similarly, one can show that $X_2 - n_2\hat{p} = \frac{\hat{p}_2 - \hat{p}_1}{\frac{1}{n_1} + \frac{1}{n_2}} = -(X_1 - n_1\hat{p})$. So

$$\begin{aligned}
\chi^2 &= \frac{(X_1 - n_1\hat{p})^2}{n_1\hat{p}(1-\hat{p})} + \frac{(X_2 - n_2\hat{p})^2}{n_2\hat{p}(1-\hat{p})} \\
&= (X_1 - n_1\hat{p})^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{1}{\hat{p}(1-\hat{p})} \quad \text{since } X_2 - n_2\hat{p} = -(X_1 - n_1\hat{p}) \\
&= \left(\frac{\hat{p}_1 - \hat{p}_2}{\frac{1}{n_1} + \frac{1}{n_2}}\right)^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{1}{\hat{p}(1-\hat{p})} \quad \text{since } X_1 - n_1\hat{p} = \frac{\hat{p}_1 - \hat{p}_2}{\frac{1}{n_1} + \frac{1}{n_2}} \\
&= \frac{(\hat{p}_1 - \hat{p}_2)^2}{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{p}(1-\hat{p})} = \left(\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{p}(1-\hat{p})}}\right)^2 = (z\text{-statistic})^2
\end{aligned}$$