Analyzing Association Between Categorical Variables.

A. Skripnikov[1]

[1]New College of Florida

IDC 5205

Contingency Table: Example.

Table 11.1 Happiness and Family Income, from 2012 General Social Survey

	Happiness			
Income	**Not Too Happy**	**Pretty Happy**	**Very Happy**	**Total**
Above average	29	178	135	342
Average	83	494	277	854
Below average	104	314	119	537

Qs:

- Which variable makes more sense as a response? Why?
- Is there an association? Why?
- If there is an association, what is its nature?

Conditional Percentages

Example (cont'd). Having selected **happiness** as the response,

Q: How could we best compare happiness across income levels?

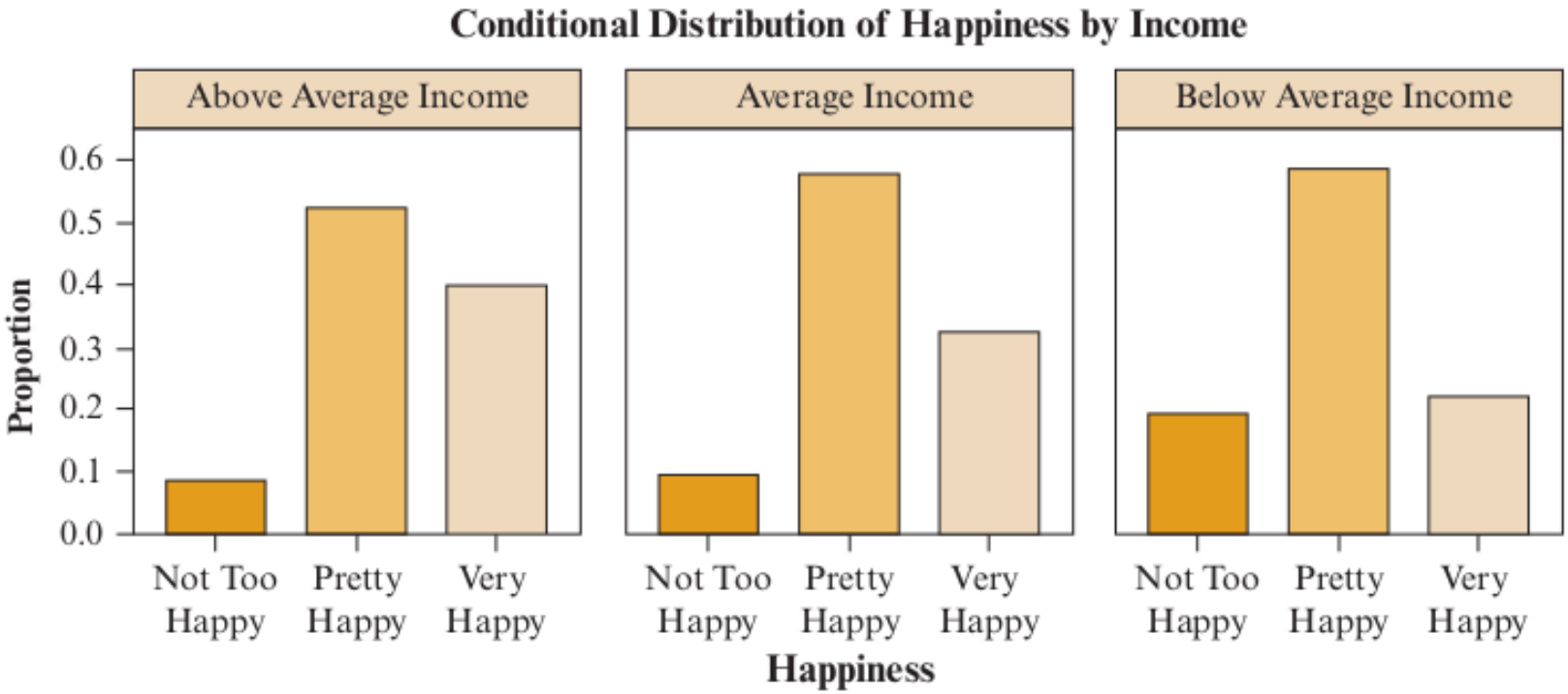
A: Conditional percentages.

	Happiness			
Income	**Not Too Happy**	**Pretty Happy**	**Very Happy**	**Total**
Above average	8%	52%	39%	342 (100%)
Average	10%	58%	32%	854 (100%)
Below average	19%	58%	22%	537 (100%)

Q: Which variable do we condition on? Why?

Task. Comment on observed percentages.

Conditional Bar Charts.



These

sample conditional distributions

will be used to make **inferences** about the corresponding

population conditional distributions.

Independence Versus Dependence (Association).

Q: In *conditional proportion tables*, what would indicate

- **independence** of two categorical variables?
- their **dependence**?

Example.

	Happiness			
Gender	**Not Too Happy**	**Pretty Happy**	**Very Happy**	**Total**
Female	14%	56%	30%	100%
Male	14%	56%	30%	100%
Overall	**14%**	**56%**	**30%**	**100%**

Source: Data from CSM, UC Berkeley.

Qs:

1. Is this table indicative of **independence** or **dependence**? **Why**?

Independence Versus Dependence (Association).

Questions (cont'd):

1. Was the table on income & happiness indicative of **independence** or **dependence**? **Why**?

Independence and Dependence (Association)

Two categorical variables are **independent** if the population conditional distributions for one of them are identical at each category of the other. The variables are **dependent** (or **associated**) if the conditional distributions are not identical.

Independence Versus Dependence (Association).

Example (cont'd).

- **Q**: If income and happiness variables **were actually independent**, would we expect the **exact** same distribution across categories? Why?

- **Q**: If not, what kinds of *count distributions* across those categories would we anticipate if **variables were actually independent**?

The latter question is to be answered via **significance testing** for

H_0: {X and Y are independent} *vs* H_a: {X and Y are dependent},

where it becomes the classic

"If H_0 **were true**, then the test statistic is distributed as ..."

Significance Test for Independence.

In testing

H_0: {X and Y are independent} *vs* H_a: {X and Y are dependent},

the **main idea** is to compare the

- **observed** cell counts in the contingency table,

  with

- **expected** cell counts if H_0 were actually true

Expected Cell Counts under H_0.

Q: How do we calculate **expected** cell counts under H_0?

Main idea: Under H_0: $\{X$ and Y are independent$\}$, we expect to have

$$P(A \text{ and } B) = P(A) \times P(B)$$

where A - any event related to variable X; B - any event related to Y.

Example. Let $X = \{\text{Income}\}$, $Y = \{\text{Happiness}\}$, while the respective events are: $A = \{\text{Above-Average Income}\}$, $B = \{\text{Not Too Happy}\}$.

To calculate the **expected** ***count*** for the (A, B) cell:

1. Calculate $P(A)$, $P(B)$.
2. Under H_0, $P(A \text{ and } B) =$ - the **expected** ***probability***.
3. Then, the **expected** ***count*** for the $A \times B$ cell is

$$n \times \quad ,$$

where n - the total sample size.

Expected Cell Counts under H_0.

Task. Proceed to calculate the expected counts for

Table 11.1 Happiness and Family Income, from 2012 General Social Survey

	Happiness			
Income	**Not Too Happy**	**Pretty Happy**	**Very Happy**	**Total**
Above average	29	178	135	342
Average	83	494	277	854
Below average	104	314	119	537

Make sure to augment table with column totals & total sample size.

Expected Cell Counts under H_0.

Using the following notation:

- E_{ij} $\iff$ the expected count for cell (i, j),
- $X = i$ $\iff$ X taking on a value in i^{th} row,
- $Y = j$ $\iff$ Y taking on a value in j^{th} column,

we have:

$$E_{ij} = n \times P(X = i) \times P(Y = j) =$$

where

- $n_{i.}$ is the i^{th} row total,
- $n_{.j}$ is the j^{th} row total.

Expected Cell Count

For a particular cell, the **expected cell count** equals

$$\text{Expected cell count} = \frac{(\text{Row total}) \times (\text{Column total})}{\text{Total sample size}}.$$

Chi-Squared Test Statistic.

Chi-Squared Statistic

The **chi-squared statistic** is an overall measure of how far the observed cell counts in a contingency table fall from the expected cell counts for a null hypothesis. Its formula is

$$X^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}.$$

Qs:

- What is the range of possible values that X^2 can take on? Why?
- What values of X^2 indicate independence ($\Leftrightarrow$ H_0 being true) - **large or small? Why?**

See R code to find X^2 statistic for Happiness/Family Income.

Sampling Distribution of X^2: Chi-Squared (χ^2).

Larger X^2 - the greater the evidence against H_0: {independence}.

Q: How large is "large *enough*" to reject H_0?

A: We need to find the probability (p-value) of observing such value of X^2 test statistic (or more extreme) **given** H_0 were actually true.

For that, we need **sampling distribution of** X^2 under H_0.

χ^2 - sampling distribution of X^2-statistic.

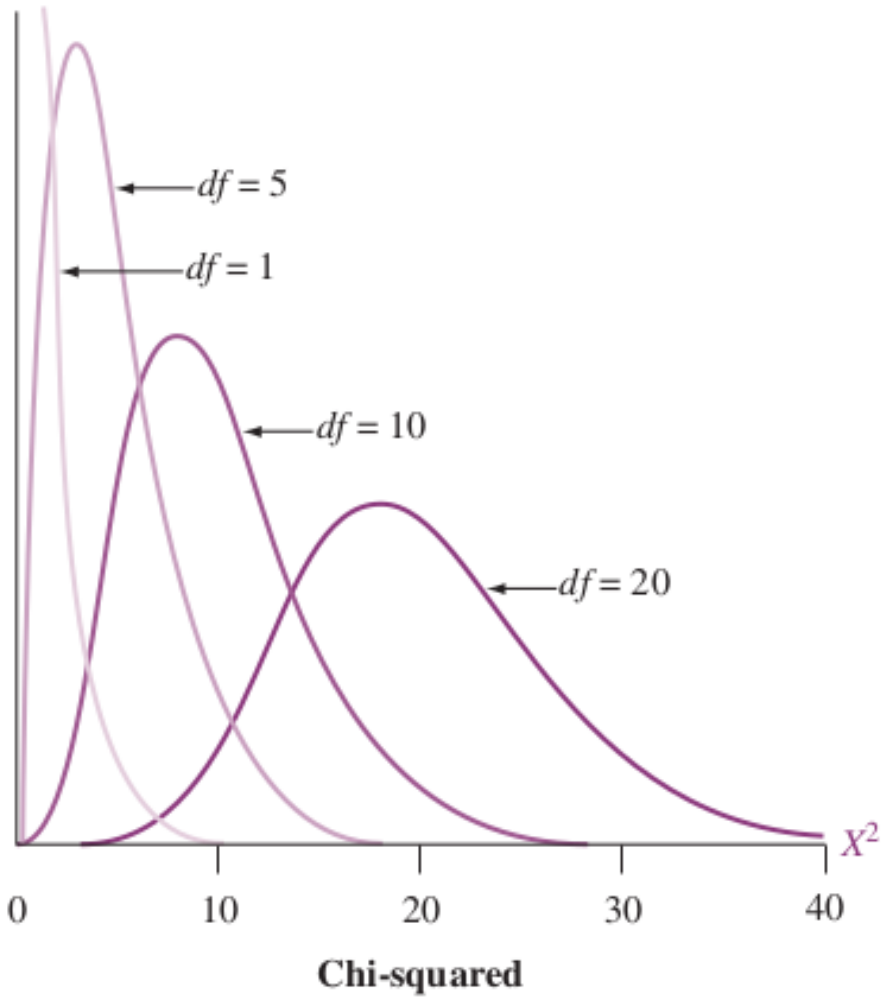
If the following assumptions are satisfied:

1. your data constitutes a **random** sample,
2. your sample size is **large enough** (all **expected** cell counts are ≥ 5),

then X^2-statistic has **chi-squared (χ^2) probability distribution**:

$$X^2 \mid H_0 \sim \chi^2_{df}$$

Properties of χ^2-distribution.



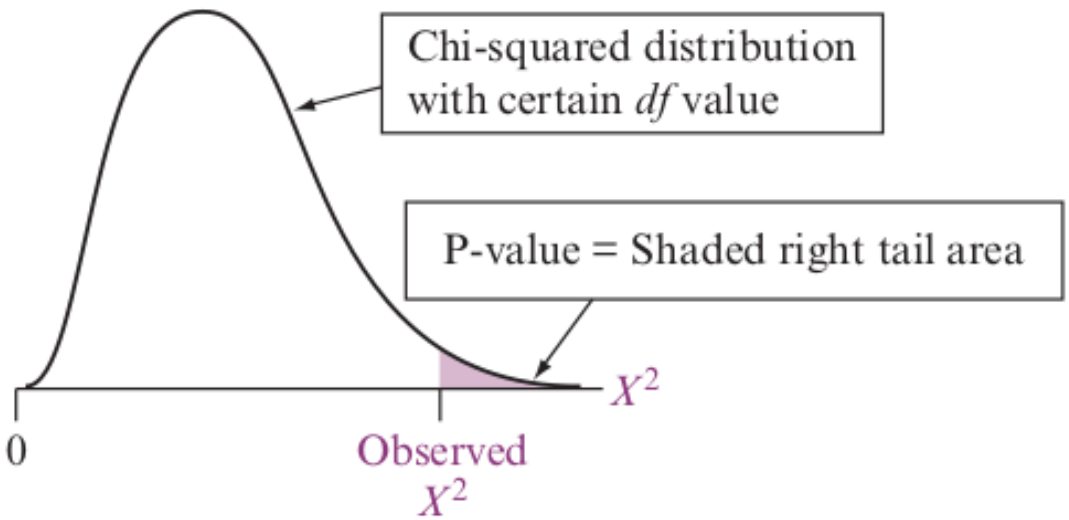
See `https://istats.shinyapps.io/ChisqDist/` for demo.

1. **Always positive**. (Why?)
2. **Degrees of freedom**: Like t-distribution, precise shape of χ^2 depends on degrees of freedom (df) parameter:
$$df = (r-1) \times (c-1),$$
for a $r \times c$ table.
3. $E[\chi^2] = df$, $V[\chi^2] = \sqrt{2df}$
"As # of rows & columns increases, X^2 tends to get larger and more varied."
4. **As $df \Uparrow$, distribution goes to bell shaped**.

Properties of χ^2-distribution.

Large X^2 provides **evidence against independence**,

$$\text{p-value} = P(\chi^2_{df} \geq X^2)$$



▲ **Figure 11.4 The P-value for the Chi-Squared Test of Independence.** This is the right-tail probability, above the observed value of the X^2 test statistic. **Question** Why do we not also use the left tail in finding the P-value?

Properties of χ^2-distribution.

Example. Proceed to conduct χ^2-test for independence between family income and happiness (**see *R* code for results**).

Qs: What are H_0 and H_a? What's the test statistic? Its distribution (are assumptions satisfied)? How is p-value calculated? What's the conclusion?

Steps of χ^2-test.

SUMMARY: The Five Steps of the Chi-Squared Test of Independence

1. **Assumptions:** Two categorical variables
   Randomization, such as random sampling or a randomized experiment
   Expected count ≥ 5 in all cells (otherwise, use small-sample test in Section 11.5)
2. **Hypotheses:**
   H_0: The two variables are independent.
   H_a: The two variables are dependent (associated).
3. **Test statistic:**

$$X^2 = \sum \frac{(\textit{observed count} - \textit{expected count})^2}{\textit{expected count}},$$

   where expected count $=$ (row total $\times$ column total) /total sample size

4. **P-value:** Right-tail probability above observed X^2 value, for the chi-squared distribution with $\boldsymbol{df = (r-1) \times (c-1)}$
5. **Conclusion:** Report P-value and interpret in context. If a decision is needed, reject H_0 when P-value $\leq$ significance level (such as 0.05).

"Fail to Reject H_0 (Independence)" $\neq$ "Accept H_0".

Caution

As with any hypotheses test, failing to reject the null hypothesis doesn't mean the variables are definitely independent. All that can be said is that independence is still plausible and can't be ruled out. ◀

Example. Presume for income & happiness example our X^2 were 1.3.

Task. Proceed to

- Calculate the p-value.
- Formulate the conclusion about H_0.

Limitations of the χ^2-test.

1. Presume you get a **low p-value**, then

   - while you have **statistically significant** evidence of association,

   - we know nothing about the nature or practical strength of that association.

2. χ^2 test is not appropriate whenever
   - Some expected cell counts are **too small** (e.g. $< \mathbf{5}$).

   OR

   - Rows (or columns) represent **dependent** samples (McNemar's test is appropriate here, but we won't cover it in this class).

X^2 Does **NOT** Measure **Strength** of Association.

While X^2 statistic quantifies the **statistical evidence** for association, it **does NOT** quantify the **practical strength** of association.

Example. Below are three hypothetical tables relating gender (expl.) and whether one attends religious services weekly (resp).

	Case A			Case B			Case C		
	Yes	**No**	***n***	**Yes**	**No**	***n***	**Yes**	**No**	***n***
Female	51%	49%	100	51%	49%	200	51%	49%	10,000
Male	49%	51%	100	49%	51%	200	49%	51%	10,000
Chi-squared = 0.08				Chi-squared = 0.16			Chi-squared = 8.0		
P-value = 0.78				P-value = 0.69			P-value = 0.005		

X^2 Does **NOT** Measure **Strength** of Association.

Example (cont'd). As sample size n increases ($A \to B \to C$),

- What happens to the **practical** difference between males & females? **Why?**
- What happens to the X^2 value? **Why?**

Measuring **Strength** of Association for 2×2 Tables.

Q: If not X^2, then how to measure the **strength** of association?

For a 2×2 contingency table, like the ones below:

Group	Yes	No
Placebo	p_1	$1 - p_1$
Regular dose	p_2	$1 - p_2$

	Yes	No	*n*
Female	51%	49%	100
Male	49%	51%	100

the classical options are:

1. **Difference of proportions**:

$$p_1 - p_2$$

2. **Ratio of proportions (AKA Relative Risk)**:

$$\frac{p_1}{p_2}$$

Measuring **Strength** of Association for 2×2 Tables.

Example. Below are the results of UCLA's large-scale survey on stress and depression among college freshmen:

	Stress				Depression		
Gender	**Yes**	**No**	**Total**	**Gender**	**Yes**	**No**	**Total**
Female	44%	56%	100%	Female	11%	89%	100%
Male	20%	80%	100%	Male	6%	94%	100%

Task: Proceed to

1. **Find** and **interpret** all two measures of association strength between gender & stress, gender & depression.
2. Which association is stronger: between gender and stress, or between gender and depression?

Measuring **Strength** of Association for 2×2 Tables.

Task (cont'd).

Measuring **Strength** of Association for 2×2 Tables.

Qs:

1. What is the range of all possible values for
   - difference of proportions $(p_1 - p_2)$,
   - relative risk $(\frac{p_1}{p_2})$
2. What type of values **for difference of proportions** represents
   - no association?
   - strong association?
3. What type of values for **relative risk** represents
   - no association?
   - strong association?

Strength of Association in $r \times c$ Tables.

While well-defined in 2×2 tables, how can we measure the association in larger tables?

One option: Pick out a particular response category and compare it across two rows.

Example. We found a large X^2 for the following table:

	Happiness		
Income	**Not Too**	**Pretty**	**Very**
Above	29 (8%)	178 (52%)	135 (39%)
Average	83 (10%)	494 (58%)	277 (32%)
Below	104 (19%)	314 (58%)	119 (22%)

Task: Interpret the strength of relationship.

Strength of Association in $r \times c$ Tables.

Task (cont'd): Interpret the strength of relationship.

Small Samples: Permutation Test.

Issue: χ^2-test relies on the distributional assumption

$$X^2 \mid H_0 \sim \chi^2_{df},$$

which only applies in **large enough samples**.

Q: What to do in case of small samples?

A: **Permutation** testing.

Small Samples: Permutation Test.

Example. "Edward Snowden - hero or criminal?" This question was posed to a set of US and International students at Williams college:

	Opinion on Edward Snowden			
Student Status	**Hero**	**Criminal**	**Neither**	**Total**
U.S.	1	9	2	12
International	5	2	1	8
Total	**6**	**11**	**3**	**20**

Qs:

1. What are the H_0 and H_a hypotheses here?
2. Can we use χ^2 test? Why?
3. How can we construct the permutation distribution of X^2 statistic **under the H_0 hypothesis being true**?
4. How can we judge whether the observed value of 6.9 is extreme and find the **permutation p-value**?

Small Samples: Permutation Test.

Slide for your notes. Please see R code and refer to in-class discussion.

Small Samples: Permutation Test.

Slide for your notes. Please see R code and refer to in-class discussion.

Small Samples: Permutation Test.