Public Policy 529
Association Between Categorical Variables
Part 1

Jonathan Hanson

Gerald R. Ford School of Public Policy
University of Michigan

November 1, 2023

Outline

1. Reminder from Probability Theory
2. Cross-Tabulation Analysis
3. The χ^2 (Chi-Squared) Test of Independence
4. Example

Outline

Example Frequency Distribution

In the GSS, respondents are asked if they own a gun and if they are afraid to walk in their neighborhood at night.

	Afraid to Walk?		
Own a Gun?	Yes	No	Total
Yes	133	395	528
No	413	707	1,120
Total	546	1,102	1,648

It appears that gun ownership is not independent of fear about walking in neighborhood at night, but what about statistical significance?

Reminder: Independence vs. Dependence

Two variables are independent when the conditional probability is the same as the unconditional probability.

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

Also, $P(B|A) = P(B)$ and $P(A \text{ and } B) = P(A)P(B)$.

In terms of our dependent variable y and our independent variable x, this means that the distribution of y does not depend upon x.

Conditional Means and Conditional Probabilities

We've also discussed expected values. By the CLT, the expected value of the sample mean is the population mean.

$$E[\bar{y}] = \mu_y$$

With independence, the value of one variable does not affect the expected mean or proportion of the other variable.

$$E[\bar{y}|x = 1] = E[\bar{y}|x = 2] = E[\bar{y}] = \mu_y$$

$$E[\hat{\pi}|x = 1] = E[\hat{\pi}|x = 2] = E[\hat{\pi}] = \pi$$

Example Frequency Distribution

The null hypothesis of independence implies a joint frequency distribution different from what we see here.

	Afraid to Walk?		
Own a Gun?	Yes	No	Total
Yes	133	395	528
No	413	707	1,120
Total	546	1,102	1,648

Keeping the same row/column totals, what numbers would appear in the four cells if the variables were independent?

Comparison with Scenario of Independence

In the overall sample, a proportion of .32 are gun owners. Find frequencies for the internal cells that reflect this overall proportion.

	Afraid to Walk?		
Own a Gun?	Yes	No	Total
Yes	133 vs. 174.9	395 vs. 353.1	528
No	413 vs. 371.1	707 vs. 748.9	1,120
Total	546	1102	1,648

We now need to test whether the observed internal cell frequencies deviate significantly from this scenario of independence.

Outline

Cross-Tabulation (Contingency Tables)

- Goal: Create groups that are as alike as possible except for the value of the independent variable.
- When all our variables are categorical, we can display their frequency distribution using cross-tabs (aka contingency tables).
- These tables illustrate the degree to which the frequency distribution of one variable depends upon the value of another.
- Typically, though not a requirement, our dependent variable makes the rows and the independent variable makes the columns.

Cross-Tabulation Format

	Category of x		
Category of y	Value 1	Value2	Total
Value 1	col. % (# cases)	col. % (# cases)	col. % (# cases)
Value 2	col. % (# cases)	col. % (# cases)	col. % (# cases)
Value 3	col. % (# cases)	col. % (# cases)	col. % (# cases)
Total	100% (# cases)	100% (# cases)	100% (# cases)

Example: Gun Ownership and Feelings of Fear

	Afraid to Walk?		
Own a Gun?	Yes	No	Total
Yes	24.36%	35.84%	32.04%
	(133)	(395)	(528)
No	75.64%	64.16%	67.96%
	(413)	(707)	(1,120)
Total	100%	100%	100%
	(546)	(1,102)	(1,648)

Compare along a row across values of the independent variable (x).

Imagine: If fear of walking at night made no difference, we would expect the rate gun ownership to be 32.04% for both sub-categories.

Is there a Statistical Relationship?

- Recall that, if the variables are independent, the conditional probabilities will equal the unconditional probabilities.
- In other words, does the distribution of y given a particular value of x match the overall distribution of y?
- In this case, does the expected value of the proportion π change when the value of x changes?

$$\mathrm{E}[\pi_{\text{gun}}|\text{afraid}] = \mathrm{E}[\pi_{\text{gun}}|\text{not afraid}] = \mathrm{E}[\pi_{\text{gun}}]?$$

Testing for Statistical Significance

- In cross-tabulation analysis, H_0 is that x and y are statistically independent.
  - the distribution of y within each category of x reflects the overall distribution.
- We need a way to measure the overall extent of deviations away from this null hypothesis.
- The statistic that measures these total deviations is a χ^2 statistic.

Controlling for Alternative Explanations

- To create comparisons that isolate the independent variable, we need to control for alternative/confounding factors.
- We add a control variable (z) to the analysis by examining the relationship between x and y inside categories of z.
- If the variables have a lot of categories, the resulting tables become large very quickly.
- We may want to divide them into separate tables, each for a different value of z.

Cross-Tabulation with Control Variable

	Value 1 of z			Value 2 of z		
y	Val. 1 x	Val. 2 x	Total	Val. 1 x	Val. 2 x	Total
Value 1	col. % (#)	col. % (#)	col. % (#)	col. % (#)	col. % (#)	col. % (#)
Value 2	col. % (#)	col. % (#)	col. % (#)	col. % (#)	col. % (#)	col. % (#)
Total	100.0% (#)	100.0% (#)	100.0% (#)	100.0% (#)	100.0% (#)	100.0% (#)

Example: Gun Ownership by Fear of Walking

	Does not Hunt			Hunts		
Gun?	Afraid	Not Afraid	Total	Afraid	Not Afraid	Total
Yes	19.1%	27.5%	24.6%	73.6%	74.7%	74.5%
	94	250	344	39	145	184
No	80.9%	72.5%	75.4%	26.4%	25.3%	25.5%
	399	658	1,057	14	49	63
Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	493	908	1,401	53	194	247

Here, we see that the gun ownership rate appears to be related to the level of fear among people who do not hunt but not among people who hunt.

Outline

χ^2 Test of Independence

- Assumption: we have a random sample with two categorical variables.
- Assumption: there is an expected frequency of at least 5 cases in every cell.
- H_0: the variables are independent.
- H_A: the variables are not independent

Testing for Independence

- The χ^2 test measures deviation of observed frequencies from expected frequencies.
- The expected frequencies are those that we should see if x and y are statistically independent.
- Label the observed frequencies f_o. Label the expected frequencies f_e.
- The χ^2 statistic captures the overall deviation of f_o from f_e in all the interior cells of the table (i.e. the categories of the variables).

Observed vs. Expected Frequencies

	Afraid to Walk?		
Own a Gun?	Yes	No	Total
Yes	24.36% (133)	35.84% (395)	32.04% (528)
No	75.64% (413)	64.16% (707)	67.96% (1,120)
Total	100% (546)	100% (1,102)	100% (1,648)

Under independence, we expect to observe that 32.04% of the 546 afraid people would own a gun. That would be 174.9.

We also would expect that 32.04% of the 1,102 non-afraid people would own a gun. That would be 353.1.

Calculating the Expected Frequency

Another way to find the expected frequency for a given cell is:

$$f_e = \frac{(\text{row total})(\text{column total})}{n}$$

Why does this work?

$$\begin{aligned} f_e &= \left(\frac{\text{row total}}{n}\right)(\text{column total}) \\ &= (\text{row proportion})(\text{column total}) \end{aligned}$$

This calculation gives us the expected cell frequency if the row proportions down the column are the same as overall row proportions.

Observed vs. Expected Frequencies

	Afraid to Walk?		
Own a Gun?	Yes	No	Total
Yes	24.36%	35.84%	32.04%
	(133)	(395)	(528)
No	75.64%	64.16%	67.96%
	(413)	(707)	(1,120)
Total	100%	100%	100%
	(546)	(1,102)	(1,648)

The expected frequency in the yes/yes cell is: $\frac{(528)(546)}{1{,}648} = 174.9$. We observed 133, fewer than would be expected.

The χ^2 Statistic

The χ^2 statistic is a way to sum of the deviations of f_e from f_o for all of the interior cells. The formula is:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

For each cell, we square the deviation between f_o and f_e and divide the result by f_e. We then add up the resulting numbers.

Calculating the χ^2 Statistic

	Afraid to Walk		
Own a Gun?	Yes	No	Total
Yes	$\frac{(133-174.9)^2}{174.9}$	$\frac{(395-353.1)^2}{353.1}$	528
No	$\frac{(413-371.1)^2}{371.1}$	$\frac{(707-748.9)^2}{748.9}$	1,120
Total	546	1,102	1,648

Calculating the χ^2 Statistic

	Afraid to Walk		
Own a Gun?	Yes	No	Total
Yes	10.1	5.0	528
No	4.7	2.4	1,120
Total	546	1,102	1,648

$$\chi^2 = 10.1 + 5.0 + 4.7 + 2.4 = 22.2$$

Interpreting the χ^2 Statistic

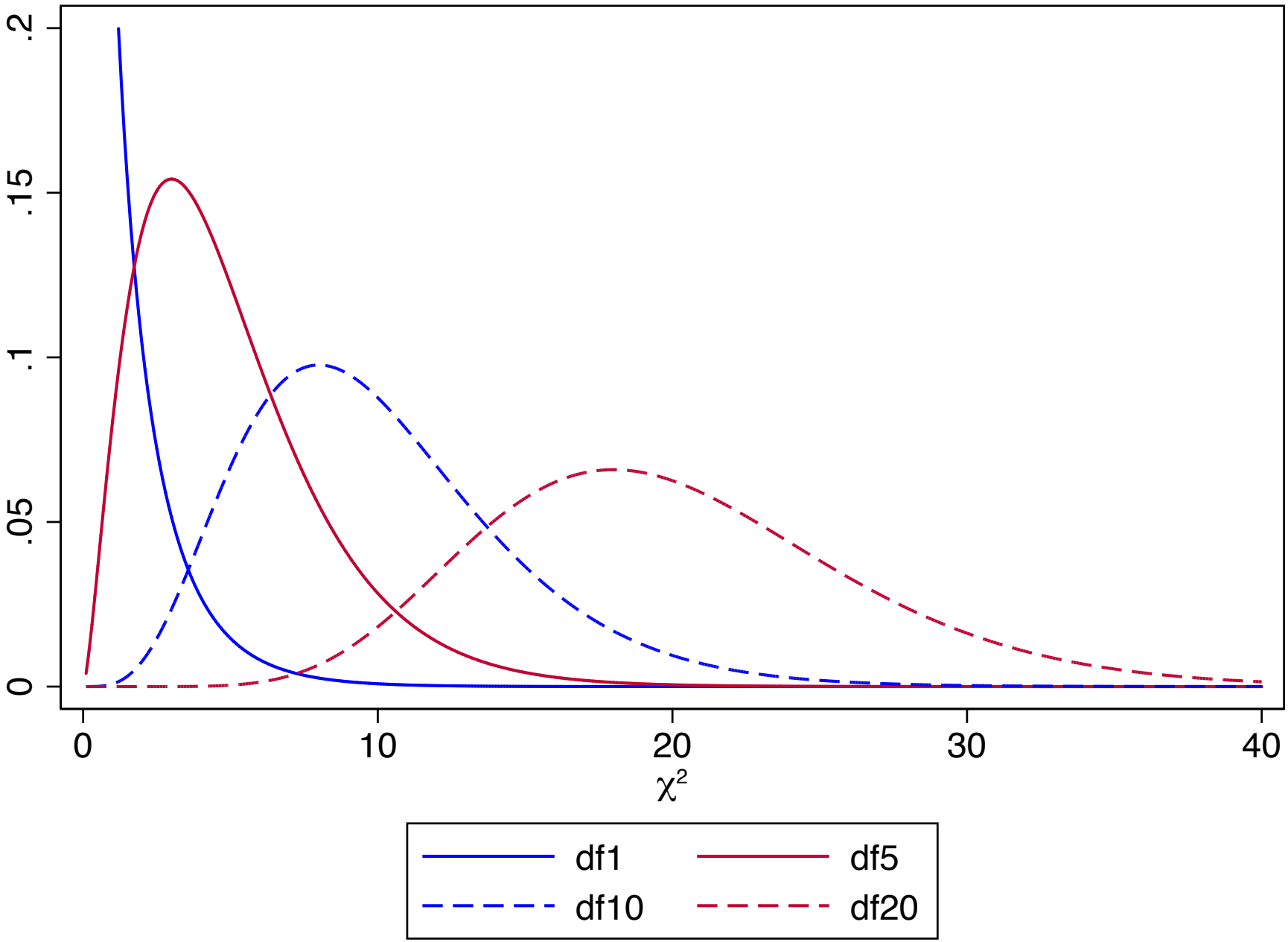
- Just like the other test statistics, the χ^2 refers to a distribution.
- Unlike the other test statistics, a χ^2 statistic is non-negative. The bigger the χ^2 statistic, the stronger the evidence for rejecting H_0. $\chi^2 = 0$ means all f_e and f_o are equal.
- Like the t distribution, the shape of the χ^2 distribution is a function of degrees of freedom. For this distribution:

$$\mu = df \text{ and } \sigma = \sqrt{2df}$$

- For the χ^2 test of independence, the degrees of freedom are based upon the number of rows (r) and columns (c) formed by the categories of x and y:

$$d.f. = (r-1)(c-1)$$

The χ^2 distribution at various degrees of freedom



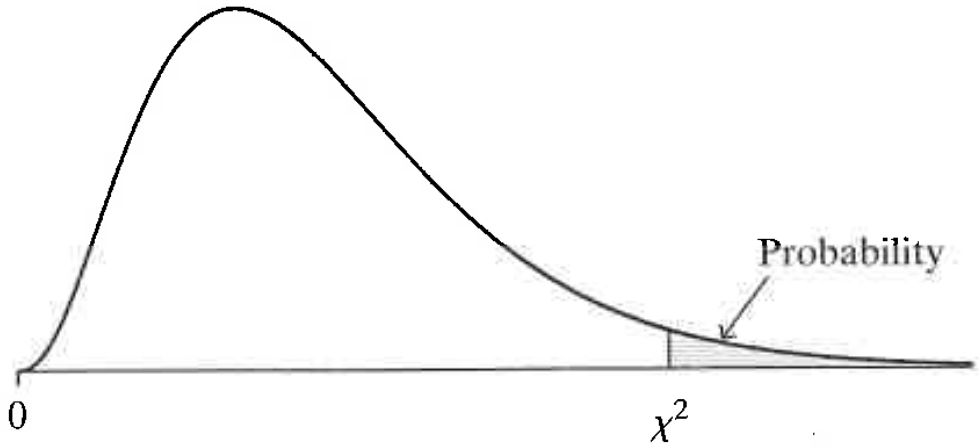
Statistical Significance

- As usual, we assess statistical significance at some level of α.
- The test is one-sided: all of α is in the right-hand tail.
- The decision rule is whether the χ^2 statistic is large enough that it crosses the critical χ^2 value associated with α.
- The *p*-value of this test refers to the area in the right-hand tail above our χ^2 statistic.
- Thus, if $p < \alpha$, we reject H_0 at that level of significance.

Back to the Example

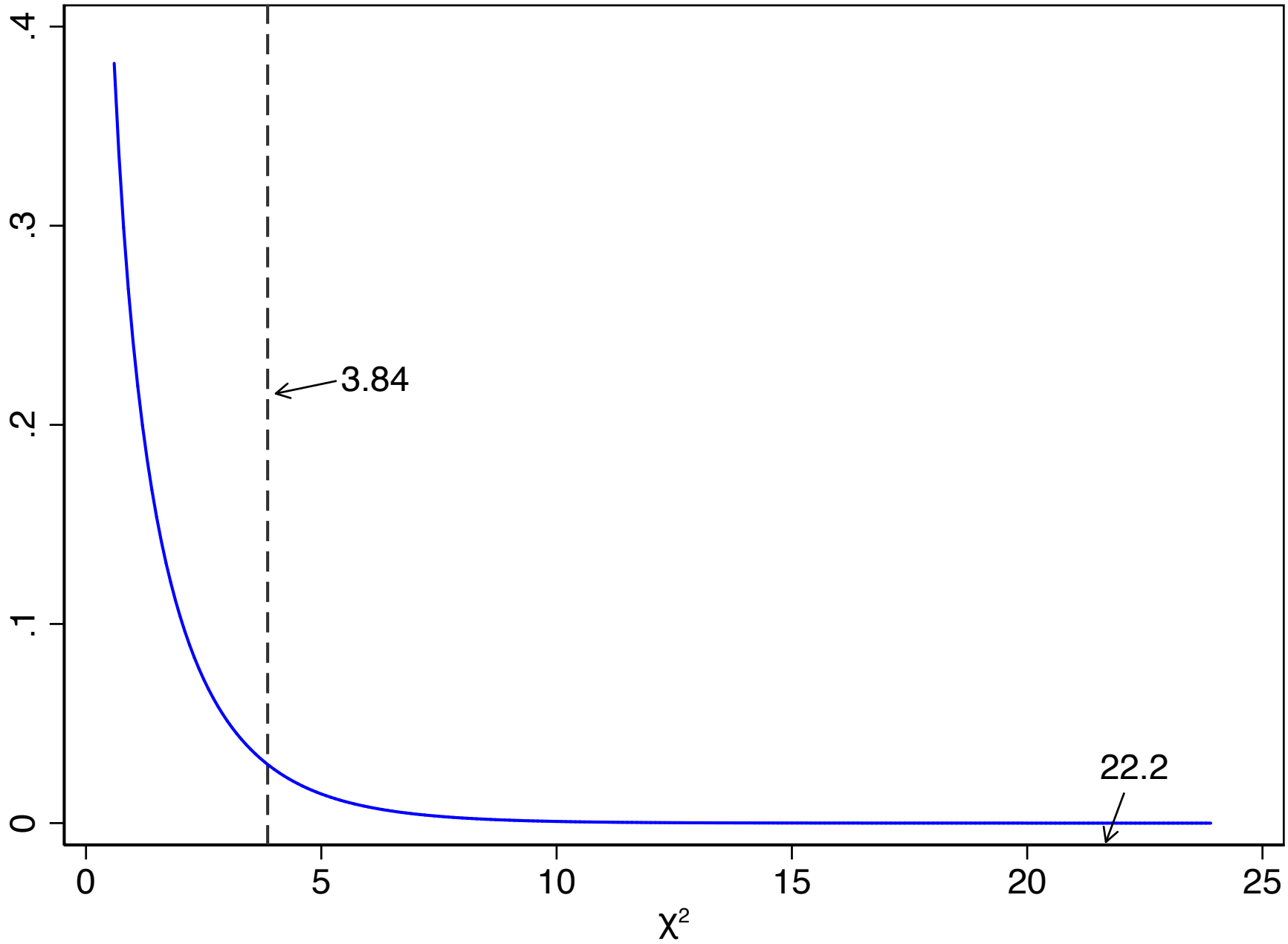
- The degrees of freedom in the example is 1.
- According to the χ^2 table, the critical value of χ^2 with 1 df that leaves an area of .05 in the right-hand tail is 3.84.
- We found a χ^2 statistic of 22.2, which is significantly greater than the critical value.
- Thus, we can reject H_0 with greater than 95% confidence.

TABLE C: Chi-Squared Distribution Values for Various Right-Tail Probabilities



	Right-Tail Probability						
df	0.250	0.100	0.050	0.025	0.010	0.005	0.001
1	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	2.77	4.61	5.99	7.38	9.21	10.60	13.82
3	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	10.22	13.36	15.51	17.53	20.09	21.96	26.12
9	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	12.55	15.99	18.31	20.48	23.21	25.19	29.59

Illustration of Significance Test



χ^2 Tests Using Stata

- When using the tabulate command, just add the option *chi* to get the χ^2 statistic reported with your results.

```
tabulate y x, col chi
```

- The command above creates cross-tabs with y making the rows and x making the columns.
- The options add column percentages and the χ^2 test.

```
. tab owngun fear, col chi
```

Key
frequency *column percentage*

have gun in home	afraid to walk at night in neighborhood: yes	no	Total
1	**133** **23.54**	**395** **34.77**	**528** **31.04**
no	**413** **73.10**	**707** **62.24**	**1,120** **65.84**
refused	**19** **3.36**	**34** **2.99**	**53** **3.12**
Total	**565** **100.00**	**1,136** **100.00**	**1,701** **100.00**

Pearson chi2(**2**) = **22.2602** Pr = **0.000**

χ^2 Tests Using R

- First make a table object with y making the rows and x making the columns.

```
mytable <- table(data$y, data$x)
```

- Then use the `chisq.test()` function.

```
chisq.test(mytable)
```

- You can embed the `table()` function inside the `chisq.test()` function if you wish.

```
> mytable <- table(gss2014$owngun, gss2014$fear)
> addmargins(mytable)

           Not Afraid Afraid  Sum
  Has gun         395    133  528
  No gun          707    413 1120
  Refused          34     19   53
  Sum            1136    565 1701
> chisq.test(mytable, correct = F)

        Pearson's Chi-squared test

data:  mytable
X-squared = 22.26, df = 2, p-value = 1.466e-05
```

Shortcomings of χ^2 Test

1. The test statistic and p-value do not give information about the nature or the strength of association.
   - e.g. the statistic itself does not tell us whether gun owners are more or less likely to be afraid to walk in their neighborhoods at night.
   - e.g. it does not tell us how much more likely gun owners are more/less likely to be afraid.
   - The statistic treats ordinal categories the same as nominal categories, yet our interpretation of the relationship would change if categories were ordinal.
   - e.g. two Likert scales

Shortcomings of χ^2 Test

2. How we define the categories matters.
   - e.g. we could sub-divide democracies into presidential and parliamentary systems.
   - e.g. the p-values would change depending on the classification used.
   - On the other hand, this is also true with difference of means tests.

Some Words of Caution

- Sample size matters in this test. Use the raw frequencies rather than proportions from a contingency table.
- It is a large-sample test (at least 5 expected observations per cell).

Outline

A Previous Example: Education Program

Suppose we are examining the results of a program to boost the rate of high-school graduation. The data:

	Participant?			
Graduated?	Yes	No	Total	Overall Prop.
Yes	83	102	185	.864
No	10	19	29	.136
Total	93	121	214	1.00

Previously, we used a difference of proportions test, and we could not reject H_0. The z-statistic was 1.049. Let's use a χ^2 test.

Example: Education Program

Internal cells contain expected frequencies (f_e)

	Participant?			
Graduated?	Yes	No	Total	Overall Prop.
Yes	80.4	104.6	185	.864
No	12.6	16.4	29	.136
Total	93	121	214	1.00

Example: Education Program

Each cells contains: $\frac{(f_o - f_e)^2}{f_e}$

	Participant?	
Graduated?	Yes	No
Yes	$\frac{(83-80.4)^2}{80.4}$	$\frac{(102-104.6)^2}{104.6}$
No	$\frac{(10-12.6)^2}{12.6}$	$\frac{(19-16.4)^2}{16.4}$

Example: Education Program

Each cells contains: $\frac{(f_o - f_e)^2}{f_e}$

	Participant?	
Graduated?	Yes	No
Yes	.084	.065
No	.538	.413

$$\chi^2 = .084 + .065 + .538 + .413 = 1.1$$

Since the critical value of χ^2 with 1 degree of freedom is 3.84, we cannot reject the null hypothesis.

Comparison of Results

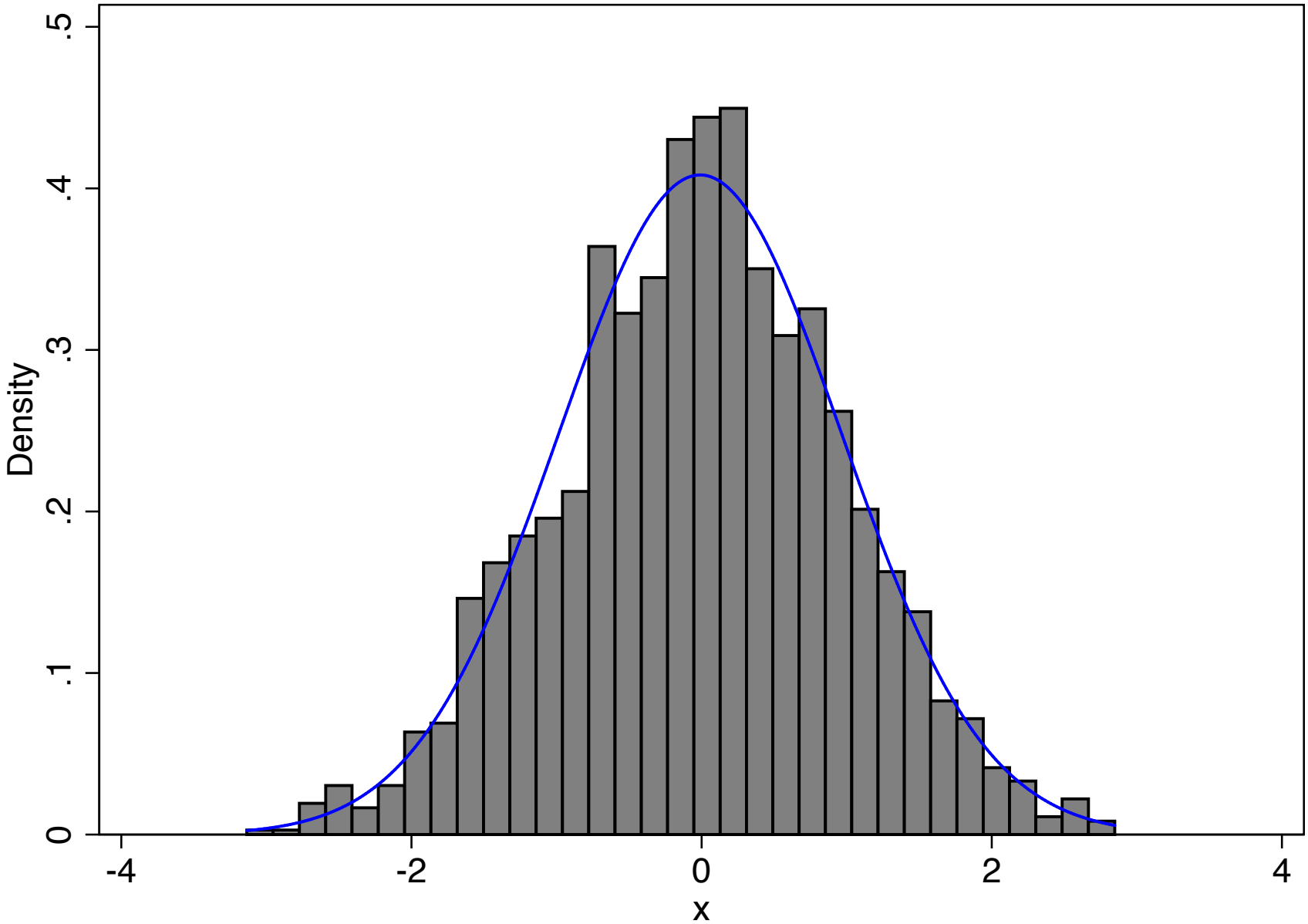
- In the difference of proportions test, the z-statistic was 1.049. We could not reject the null hypothesis of no difference.
- With the χ^2 test, the statistic was 1.1. We could not reject the null hypothesis that the variables are independent.
- These tests are related. Note that $1.049^2 = 1.1$. In other words, $z^2 = \chi^2$.
- Likewise, the area in tails of the two distributions, χ^2 (one-tailed) and z (two-tailed), is the same for these tests (p=.294).
- The χ^2 test, however, generalizes to variables that consist of more than two categories.

Illustrate Relationship between Normal and χ^2

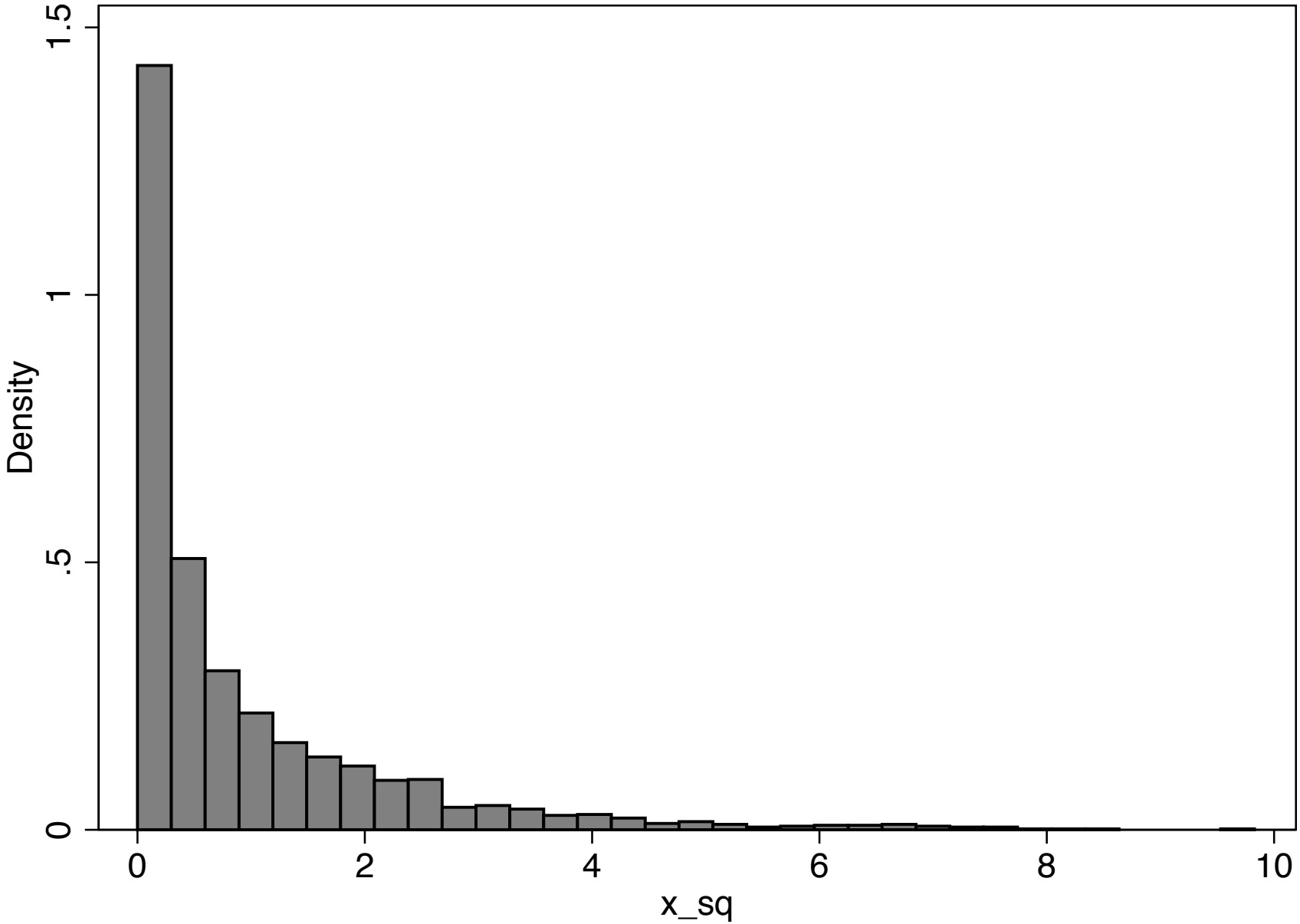
- I used Stata to generate a random sample of a variable (x) that comes from a population with a standard normal distribution.
- There are 2,000 randomly-generated data points. The sample mean is -.01, and the sample standard deviation is .98.
- I then created another variable (x_sq) consisting of the values of the first variable squared. The sample mean is .95 and the standard deviation is 1.32.
- Let's examine these distributions.

Distribution of x

Distribution of x with Normal Curve



Distribution of *x_sq*



Distribution of *x_sq* with $\chi^2(1df)$