CHAPTER 6

Inference for Categorical Data

6.1 Inference for Single Proportion

Sampling Distribution of $\hat{p}$

The sampling distribution for $\hat{p}$ based on a sample size n from a population with a true proportion p is nearly normal when:

1. The sample's observations are independent (Simple Random Sample).
2. We expected to see at least 10 successes and 10 failures in the sample.

When these conditions are met, the sampling distribution of $\hat{p}$ is nearly normal with mean p and standard error $SE = \sqrt{\frac{p(1-p)}{n}}$

Confidence Interval for Population Proportion

If a point estimate closely follow a normal model with standard error SE, then a confidence interval for the population parameter is

$$\hat{p} \pm z * \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

where *z** correspond to the confidence level selected.

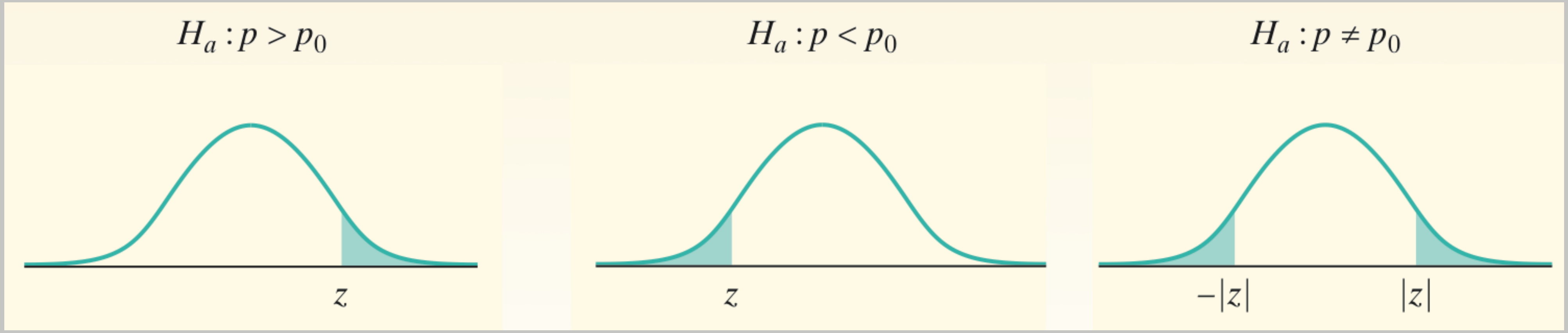
Hypothesis Testing for a Single Proportion

Use this test only when the expected numbers of successes and failures are both at least 10.

Choose an SRS of size *n* from a large population that contains an unknown proportion *p* of successes. To test the hypothesis H_0: $p = p_0$, compute the *z* statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Find the *P*-value by calculating the probability of getting a *z* statistic this large or larger in the direction specified by the alternative hypothesis H_a:

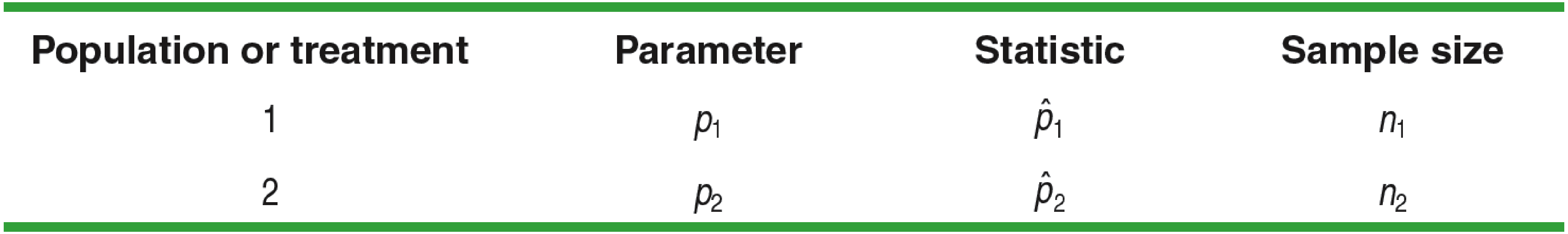


6.2 Difference of two Proportions

We would like to extend the method from section 6.1 to apply confidence intervals and hypothesis tests to difference in population proportions $p_1 - p_2$.

First, we will identify a reasonable point estimate of $p_1 - p_2$ based on the sample which is $\hat{p}_1 - \hat{p}_2$.

Next, we will apply the same process we used in single-proportion .



Population or treatment	Parameter	Statistic	Sample size
1	p_1	$\hat{p}_1$	n_1
2	p_2	$\hat{p}_2$	n_2

Sampling Distribution of the difference of Two Proportions

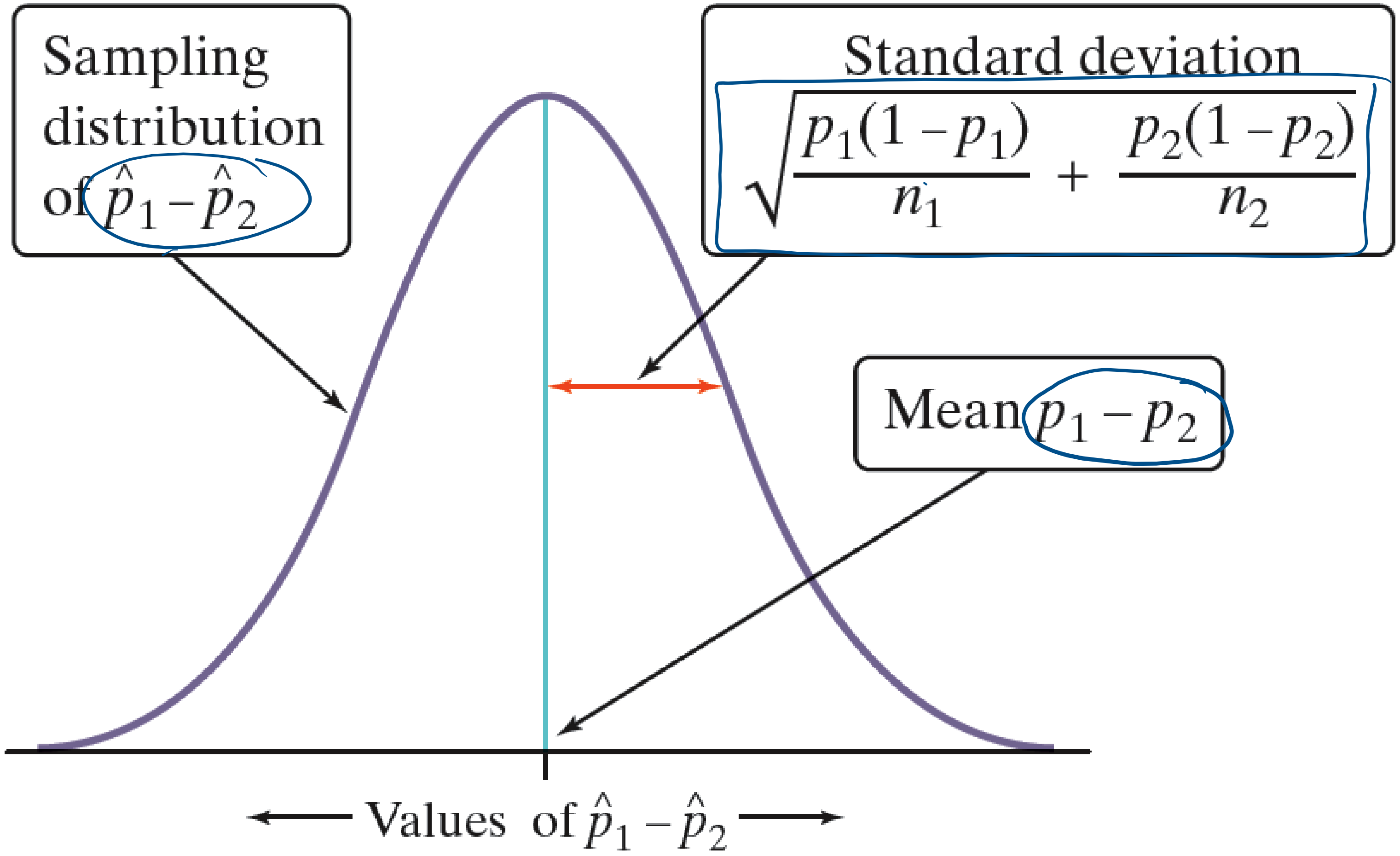
Conditions for the sampling distribution of $\hat{p}_1 - \hat{p}_2$to be Normal

- *Independence*, *extended*. The data are independent within and between the two groups. Generally this is satisfied if the data come from two independent random samples or if the data come from a randomized experiment.
- *Success-failure condition*. The success-failure condition holds for both groups, where we check successes and failures in each group separately.

When the conditions are satisfied, the standard error $\hat{p}_1 - \hat{p}_2$ is

$$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Sampling Distribution of a Difference Between Proportions



Sampling distribution of $\hat{p}_1 - \hat{p}_2$

Standard deviation $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Mean $p_1 - p_2$

Confidence Intervals for $p_1 - p_2$

When the conditions are met, an approximate level C confidence interval for $\hat{p}_1 - \hat{p}_2$ is

$$\hat{p}_1 - \hat{p}_2 \pm Z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Where $z^*$is the critical value for the standard normal curve with area C between $-z^*$and z^*

Example 1

As part of the Pew Internet and American Life Project, researchers conducted two surveys in late 2009. The first survey asked a random sample of 800 U.S. teens about their use of social media and the Internet. A second survey posed similar questions to a random sample of 2253 U.S. adults. In these two studies, 73% of teens and 47% of adults said that they use social-networking sites. Use these results to construct and interpret a 95% confidence interval for the difference between the proportion of all U.S. teens and adults who use social-networking sites.

Significance Test for Comparing Proportions

An observed difference between two sample proportions can reflect an actual difference in the parameters, or it may just be due to chance variation in random sampling or random assignment. Significance tests help us decide which explanation makes more sense.

To do a test, standardize $\hat{p}_1 - \hat{p}_2$ to get a z statistic:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\text{standard deviation of statistic}}$$

If H_0: $p_1 = p_2$ is true, the two parameters are the same. We call their common value p. But now we need a way to estimate p, so it makes sense to combine the data from the two samples. This **pooled** (or **combined**) **sample proportion** is:

$$\hat{p} = \frac{\text{count of successes in both samples combined}}{\text{count of individuals in both samples combined}}$$

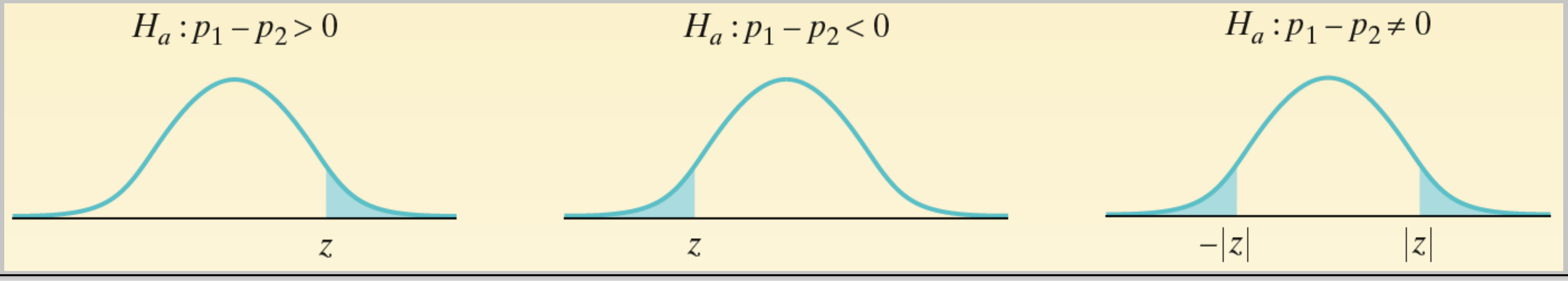
Significance Test for Comparing Proportions

Draw an SRS of size n_1 from a large population having proportion p_1 of successes, and draw an independent SRS of size n_2 from a large population having proportion p_2 of successes. To test the hypothesis $H_0 : p_1 - p_2 = 0$, first find the pooled proportion $\hat{p}$ of successes in both samples combined. Then compute the z statistic

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Use this test only when the expected numbers of successes and failures are both at least 10.

Find the P-value by calculating the probability of getting a z statistic this large or larger in the direction specified by the alternative hypothesis H_a :



Example

Researchers designed a survey to compare the proportions of children who come to school without eating breakfast in two low-income elementary schools. An SRS of 80 students from School 1 found that 19 had not eaten breakfast. At School 2, an SRS of 150 students included 26 who had not had breakfast. More than 1500 students attend each school. Do these data give convincing evidence of a difference in the population proportions? Carry out a significance test at the $\alpha = 0.05$ level to support your answer.

Example Cont.

Practice Problem 1

A recent study compared the proportions of young women and men who use Instagram. A total of 537 young women and 532 young men were surveyed. 328 of the women and 234 of the men stated that they use Instagram. construct and interpret a 95% confidence interval for the difference in the proportion of all young women and young men who use Instagram.

Practice Problem 2

Are young women and men equally likely to say they use Instagram? Recall the data from our earlier example: 328 out of 537 women (61.1%) and 234 out of 532 men (44.0%) use Instagram. Perform a significance test to see if the difference is large enough to lead us to believe that the population proportions are not equal.

6.3 Testing for Goodness of fit using Chi-square

The Chi-Square Test (χ^2)

- The Chi-square test can be used for a categorical variable with any number *k* of categories.
- The chi-square distribution has just one parameter called *degrees of freedom (df)*, which influences the shape, center, and spread of the distribution.
- The Chi-square statistic can be used to see whether a frequency distribution fits a specific pattern.
- Examples:
  - Suppose a market analyst whished to see whether consumers have any preference among five flavors of a new fruit soda.
  - A traffic engineer may wish to see whether accidents occur more often on some days than others, so that she can increase police patrols accordingly.

Example 1

A market analyst whished to see whether consumers have any preference among five flavors of a new fruit soda. A sample of 100 people provided these data. Is there enough evidence to reject that there is no preference in the selection of fruit soda flavors? Let $\alpha=.05$.

Frequency	Cherry	Strawberry	Orange	Lime	Grape
Observed	32	28	16	14	10
Expected					

Conditions for the Chi-Square Test

1. *Independence*: Each case that contributes a count to the table must be independent of all the other cases in the table.
2. *Sample size*: Each particular scenario (cell) must have at least 5 *expected* cases.
3. *df > 1*: Degrees of freedom must be greater than 1.

Testing Hypothesis about One Categorical Variable Goodness-of-Fit

Step 1: The null and the alternative Hypotheses

H_0: The probabilities for the k categories of a categorical variable are given by $\mathrm{p}_1, \mathrm{p}_2, \dots, p_k$

H_a: Not all probabilities specified in H_0 are correct.

Note: The probabilities specified in the null hypothesis must sum to 1.

Example 1

A market analyst whished to see whether consumers have any preference among five flavors of a new fruit soda. A sample of 100 people provided these data. Is there enough evidence to reject that there is no preference in the selection of fruit soda flavors? Let $\alpha = .05$.

Step 1: State the Hypotheses

Frequency	Cherry	Strawberry	Orange	Lime	Grape
Observed	32	28	16	14	10
Expected					

Step 2: The Chi-Square Statistic

Summaries the data into an appropriate test statistic after first verifying that necessary data conditions are met. The expected counts are greater than 5 and none are less than 1.

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

The **observed counts** are the counts in the cells of a two-way table of the sample data.

The **expected counts** for the $i - th$ category is computed by np_i

Example 1

A market analyst whished to see whether consumers have any preference among five flavors of a new fruit soda. A sample of 100 people provided these data. Is there enough evidence to reject that there is no preference in the selection of fruit soda flavors? Let $\alpha = .05$.

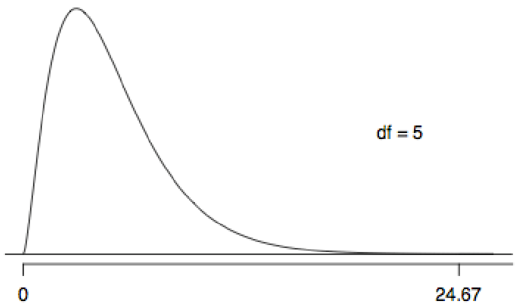
Step 2: Check the condition and

Compute the chi-square statistic.

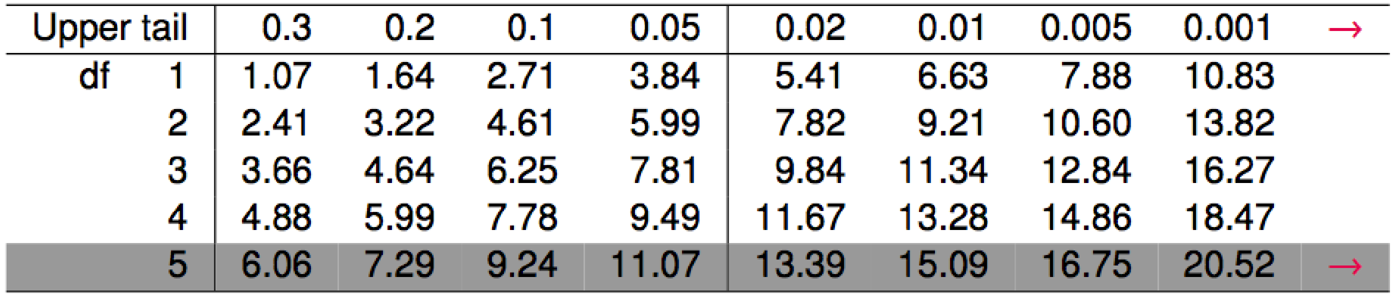
Frequency	Cherry	Strawberry	Orange	Lime	Grape
Observed	32	28	16	14	10
Expected					

Step 3: P-value for the Chi-Square Goodness-of-Fit Test

- A Chi-square distribution has a parameter called degree of freedom; it is given by $df = k - 1$ (k = Number of categories)
- Using the χ^2 distribution with df, The p-value is the area to the right of the calculated test statistic χ^2 which we can use Chi-square distribution table or R-Studio to find.



p-value = $P(\chi^2_{df=5} > 24.67)$
is less than 0.001



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	→

Example 1 Cont.

A market analyst whished to see whether consumers have any preference among five flavors of a new fruit soda. A sample of 100 people provided these data. Is there enough evidence to reject that there is no preference in the selection of fruit soda flavors? Let $\alpha = .05$.

Step 3: Find the degree of freedom

and the p-value.

Frequency	Cherry	Strawberry	Orange	Lime	Grape
Observed	32	28	16	14	10
Expected					

Step 4: Making a Decision and Conclusion

Decide whether the result is statistically significant based on the p-value .choose a significant level α*; the standard is* α*=.05*.

P-value $< \alpha \rightarrow$ reject $H_0 \rightarrow$ conclude H_a (**Statistically Significant**)

P-value $\geq \alpha \rightarrow$ fail to reject $H_0 \rightarrow$ cannot conclude H_a

Report the conclusion in the context of the situation

Example 1 Cont.

A market analyst whished to see whether consumers have any preference among five flavors of a new fruit soda. A sample of 100 people provided these data. Is there enough evidence to reject that there is no preference in the selection of fruit soda flavors? Let $\alpha = .05$.

Step 4: Make a decision and state your conclusion

Example 2

The adviser of an ecology club at a large college believes that the group consists of 10% freshmen, 20% sophomores, 40% juniors, and 30% seniors. The membership for the club this year consisted of 14 freshman, 19 sophomores, 51 juniors, and 16 seniors. At α=*.10*, test the adviser claim

Frequency	Freshmen	Sophomores	Junior	Senior
Observed				
Expected				

Example 2 Cont.

The adviser of an ecology club at a large college believes that the group consists of 10% freshmen, 20% sophomores, 40% juniors, and 30% seniors. The membership for the club this year consisted of 14 freshman, 19 sophomores, 51 juniors, and 16 seniors. At $\alpha = .10$, test the adviser claim

Practice Problem

An ABC News poll asked adults whether they felt genetically modified food was safe to eat. Thirty-five percent felt it was safe, 52% felt it was not safe, and 13% had no opinion. A random sample of 120 adults was asked the same question at a local fair. Forty people felt that genetically modified food was safe, 60 felt that it was not safe, and 20 had no opinion. At the 0.01 level of significance, is there sufficient evidence to conclude that the proportions differ from those reported in the survey?

Practice Problem Cont.

An ABC News poll asked adults whether they felt genetically modified food was safe to eat. Thirty-five percent felt it was safe, 52% felt it was not safe, and 13% had no opinion. A random sample of 120 adults was asked the same question at a local fair. Forty people felt that genetically modified food was safe, 60 felt that it was not safe, and 20 had no opinion. At the 0.01 level of significance, is there sufficient evidence to conclude that the proportions differ from those reported in the survey?

6.4 Testing for Independence in Two-Way Tables

The Chi-Square Test For Two Way Table

When data can be tabulated in table form in terms of frequencies, Several Types of hypotheses can be tested by using a Chi-square test. One such test is:

Test of independence of variable: can be used to test whether a relationship that is observed in a sample can be used to infer that there is a relationship in the population from which the sample was drawn.

Condition for the Chi-Square Test

1. *Independence*: Each case that contributes a count to the table must be independent of all the other cases in the table.
2. *Sample size*: Each particular scenario (cell) must have at least 5 *expected* cases.
3. *df* > *1*: Degrees of freedom must be greater than 1.

The Four Step in a Chi-Square Test of a Relationship Between Two Variables

Step 1: The null and the alternative Hypotheses

H_0: The variables are not related in the population

H_a: The variable are related in the population

Example

A researcher wishes to determine whether there is a relationship between the gender of an individual and the amount of alcohol consumed. A sample of 68 people is selected and the following data are obtained. At $\alpha = .10$, can the researcher conclude that alcohol consumption is related to gender?

Step 1: State the hypotheses.

Alcohol Consumption				
Gender	**Low**	**Moderate**	**High**	**Total**
Male	10	9	8	27
Female	13	16	12	41
Total	23	25	20	68

Step 2: The Chi-Square Statistic for Two-Way Tables

The **chi-square statistic** measures the difference between the observed counts and corresponding expected counts. It can be calculated as

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

Note: A large chi-square value occurs when there is a large difference between the observed and expected counts.

Expected Counts

The Basics of computing the chi-square statistic used for two-way tables are as follow:

- The **observed counts** are the counts in the cells of a two-way table of the sample data.
- The **expected counts** are hypotheticals counts that would occur in the cells of the table if the null hypothesis were true. For each cell in the table the expected count can be calculated as

$$Expected = \frac{Row\ total\ \times Column\ total}{Total\ n}$$

Example Cont.

Step 2: compute the Chi-Square Statistic.

Alcohol Consumption				
Gender	**Low**	**Moderate**	**High**	**Total**
Male	10	9	8	27
Female	13	16	12	41
Total	23	25	20	68

Step 3: P-value for the Chi-Square Test

- P-value is the probability that the calculated chi-square statistic, χ^2, could be as large as it is or larger than it is if the null hypothesis is true.

- We use R-Studio to find the p-value.

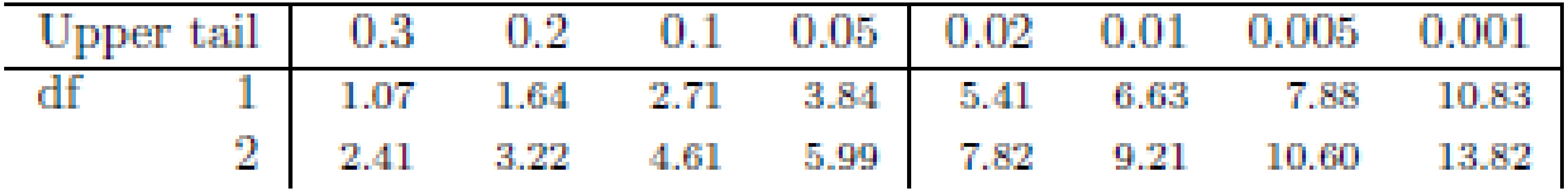
- A chi-square distribution has a parameter called degree of freedom.

  Degree of freedom = $df \ = (R-1)(C-1)$

  - R = number of row variable categories
  - C = Number of column variable categories.

Example Cont.

Step 3: Find the degree of freedom and the p-value.



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

Step 4: Making a Decision

Decide whether the result is statistically significant based on the p-value .choose a significant level α*; the standard is* α*=.05.*

P-value < $\alpha \longrightarrow$ reject $H_0 \longrightarrow$ conclude H_a (**Statistically Significant**)

P-value $\geq \alpha \longrightarrow$ fail to reject $H_0 \longrightarrow$ cannot conclude H_a

Example Cont.

A researcher wishes to determine whether there is a relationship between the gender of an individual and the amount of alcohol consumed. A sample of 68 people is selected and the following data are obtained. At $\alpha = .10$, can the researcher conclude that alcohol consumption is related to gender?

Step 4: Make a decision, and state your conclusion

Practice Problem

A sociologist whishes to see whether the number of years of college a person has completed is related to her or his place of residence. a sample of 88 people is selected and classified as shown.

Location	No College	Four-Year degree	Advanced degree	Total
Urban	15	12	8	35
Suburban	8	15	9	32
Rural	6	8	7	21
Total	29	35	24	88

Practice Problem Cont.

Practice Problem

A study was conducted to see if there was a relationship between the gender of an attorney and the type of practice he or she is engaged in. A sample of 240 attorney is selected, and the results are shown. At $\alpha = 0.05$, can it be assumed that gender and employment are independent?

Gender	Private Practice	Law firm	Government
Male	112	16	12
Female	64	18	18

Practice Problem Cont.