§9.1 Estimating a Population Proportion

Often, we are interested in estimating a population proportion rather than a population mean. For example, we may be interested in estimating the true proportion of people in a population who possess some characteristic of interest, as in the following:

- What is the true proportion of voters who support a certain political candidate?
- What proportion of people infected with the ebola virus will die?
- What is the true proportion of people who will suffer side effects from a certain drug?

To answer questions like these, we would collect a random sample of data, estimate the population proportion, p, using the **sample proportion**, $\hat{p}$ **,** and then use information about the **sampling distribution of the sample proportion** to calculate a confidence interval for the population proportion.

§9.1.1 Sampling Distribution of the Sample Proportion, $\hat{p}$

The binomial distribution forms the basis for the sampling distribution of $\hat{p}$ when small samples are drawn. However, we will consider only the case where our sample is "large"; in this case, the sample proportion ($\hat{p}$) approximately follows a normal distribution with mean p (p= population proportion), and variance p(1-p)/n. (This is an application of the normal approximation to the binomial.)

§9.1.2 Approximate (1-α)x100% Confidence Intervals for the Population Proportion, p

An approximate "large sample" (1-α)x100% confidence interval for the population proportion, p, is:

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

Where $\hat{p}$ is the sample proportion (i.e. the point estimate of the population proportion), and $Z_{\alpha/2}$ is the (1-α/2) x 100th percentile of the standard normal distribution. **This "large sample" approximate confidence interval is appropriate if nṕ and n(1-ṕ) are both ≥ 5.**

Note that, in the formula for the confidence interval, the variance of ṕ, p(1-p)/n, is being *estimated* by ṕ (1-ṕ)/n. This is similar to the situation where, in order to calculate a confidence interval for the mean of a population when the population variance, σ^2, is unknown, we use the sample variance, s^2 to estimate the population variance.

Example:
47% of a random sample of 1495 subjects who are taking a prescription pain killer report having a particular side effect. Find a 95% confidence interval for the true proportion of subjects who experience the side effect after taking the drug.

Example:

Georgia

Swing Toward The Democrats

Wednesday, Oct. 28, 2020

But most voters still expect Trump to win

West Long Branch, NJ – The race for Georgia's electoral votes remains very close, but Joe Biden has gained ground on Donald Trump in the latest ***Monmouth ("Mon-muth") University Poll***. Democrats have also improved their standing in the two U.S. Senate races, erasing a GOP lead in the regularly scheduled contest and leaving two Republicans fighting for a spot in the special election runoff.

Among all registered voters in Georgia, Biden is supported by 50% and Trump is supported by 45%. Another 2% say they will vote for Libertarian Jo Jorgensen and 2% are undecided. These results represent a small swing in the Democrat's favor, but the numbers are not statistically different from Trump's single point edge last month (47% to 46% for Biden) or the tied result in July (47% each).

METHODOLOGY

The *Monmouth University Poll* was sponsored and conducted by the Monmouth University Polling Institute from October 23 to 27, 2020 with a statewide random sample of 504 Georgia voters drawn from a list of registered voters. This includes 160 contacted by a live interviewer on a landline telephone and 344 contacted by a live interviewer on a cell phone, in English. Monmouth is responsible for all aspects of the survey design, data weighting and analysis. The full sample is weighted for party primary vote history, age, gender, race, education, and region based on state voter registration list information and U.S. Census information (CPS 2018 supplement). Data collection support provided by Braun Research (field) and Aristotle (voter sample). For results based on the full voter sample, one can say with 95% confidence that the error attributable to sampling has a maximum margin of plus or minus 4.4 percentage points (unadjusted for sample design). Sampling error can be larger for sub-groups (see table below). In addition to sampling error, one should bear in mind that question wording and practical difficulties in conducting surveys can introduce error or bias into the findings of opinion polls.

Note: Above, we have learned about computing confidence intervals for a single proportion. It is also possible to perform *hypothesis tests* for a population proportion; this will not be covered in class. However, you should read about this in the Weiss text (section 12.2).

§9.2 Hypothesis Tests for $p_1 - p_2$: The Chi-square Test of Homogeneity

We will consider only the case where we are testing:

H_0: $p_1 - p_2 = 0$
H_A: $p_1 - p_2 \neq 0$

To investigate the hypotheses shown above, we collect two independent random samples of data, one from each population, and then perform the chi-square test of homogeneity. In this test, the data are summarized in a 2 x 2 'contingency table'. The test statistic for this test compares the observed frequencies in the table against the frequencies one would expect if the null hypothesis was true. The test statistic follows a chi-square distribution with one degree of freedom.

Example: To study the effectiveness of a new flu vaccine, the numbers of flu cases in two independent random samples of people were observed. The 100 people in the first sample received the new flu vaccine, the 80 people in the other sample received a placebo vaccine. Suppose that 25 of the flu vaccine people and 31 of the placebo vaccine people had the flu. Test whether the proportion of flu vaccine people who get the flu is different from the proportion of placebo people who get the flu, at $\alpha = 0.01$.

H_0: $p_1 - p_2 = 0$
H_A: $p_1 - p_2 \neq 0$
$\alpha = 0.01$

where p_1 = true proportion of flu-vaccinated people who get the flu
p_2 = true proportion of placebo-treated people who get the flu

Let's summarize the information given in the example in a 2 x 2 contingency table.

		Populations		
		flu vaccine	placebo vaccine	**total**
outcomes	flu	25	31	**56**
	no flu	75	49	**124**
	total	**100**	**80**	**180**

[Note that two independent random samples of data were collected]

To test the hypotheses, we need to investigate whether or not the **observed** counts in the four cells of the table are 'similar' to the counts we would **expect** to see if the null hypothesis was true. For example, in the top left cell we would want to compare the observed frequency, 25, with the number of flu-vaccinated people who would be expected to get the flu if H_o were true. The latter expected value is calculated by multiplying the probability of getting the flu, assuming H_o is true, by the number of people who received the flu vaccine (100).

When H_0 is true, the probabilities of getting the flu in the two groups are the same; so $p_1 = p_2 = p_p$ where p_p represents the common probability of getting the flu. We can estimate this common probability by observing the overall proportion of people who got the flu in the study. [Note: p_p is called the 'pooled estimate of the probability of success', because it is determined by 'pooling' or combining the two samples and observing the overall proportion of successes.]

For the current example, we can estimate p_p by $\hat{p}_p = (25 + 31)/180 = 0.311111$

Thus, if H_0 is true, we would **expect** 0.311111 x 100 = 31.1111 of the flu-vaccinated people to catch the flu. Similarly, we would **expect** 0.311111 x 80 = 24.8889 of the 80 placebo-treated people to catch the flu. The remaining expected counts can either be determined using similar logic, or by subtracting the expected counts above from the relevant column totals. (In fact, if you think about it, once one expected count has been determined, the remaining three can be determined by subtraction from row and/or column observed totals).

In general, we can set up a 2 x 2 table and calculate the expected values as follows:

observed = O_{11}=25 expected = E_{11}=31.1111	observed = O_{12} = 31 expected = E_{12} = 24.8889
observed = O_{21}=75 expected = E_{21}=68.8889	observed = O_{22} = 49 expected = E_{22} =55.1111

To test the hypotheses of interest, we need a test statistic that measures how far away the observed counts in our sample are from the expected counts. The statistic that we use has the following form:

$\chi^2 = \Sigma$ (observed – expected)^2 / expected, (summation is across all of the cells)
or, in more compact notation: $\chi^2 = \Sigma [(O - E)^2 / E]$

Under H_0, this test statistic follows, approximately, a chi-square distribution with one degree of freedom (χ^2_1)…as long as at least 80% of the expected values are >= 5 (in effect, for a 2x2 table, all of the expected values must be >= 5). The degrees of freedom are equal to the number of expected frequencies that are free to vary, given that the row and column totals are known.

The larger the test statistic is, the less similar the observed and the expected values are. This indicates that the evidence is in favor of H_A.

In our example, $\chi^2 = \Sigma (O - E)^2 / E =$

P-value: $Pr(\chi^2_1 >$ test statistic value) =

Decision & Conclusion:

Which technique should we use to compare two proportions from independent populations when one or more of the expected values in the 2 x 2 contingency table is/are less than 5? We can use **Fisher's exact test.** Fisher's exact test makes use of the hypergeometric distribution (not covered in this class). You are not expected to calculate a Fisher's exact test by hand for this class; however, you are expected to be able to find and use the p-value for a two-sided Fisher's exact test on SAS output –to be covered in lab.

Note: The textbook discusses "chi-square goodness of fit" tests; these are extensions of the test of homogeneity. While we will not discuss goodness of fit tests in class (and you will not be quizzed on them), students should read about these tests.

§9.3 Comparing More Than Two Proportions: R x C Tables and Tests of Homogeneity

Using the contingency table approach, we can compare proportions across more than two populations. That is, the analysis need not be restricted to 2 rows and 2 columns. When the row and/or column variable(s) in a contingency table analysis has (have) more than 2 categories, the analysis for the resulting R x C table (R = number of rows, C = number of columns) is very similar to the 2x2 table analysis.

Example: Suppose we want to study the relationship between age at first birth and the development of breast cancer. In particular, we would like to know if the following proportions of women are the same:
the proportion aged < 20 at first birth who develop breast cancer
the proportion aged 20-24 at first birth who develop breast cancer

the proportion aged 25-29 at first birth who develop breast cancer
the proportion aged 30-34 at first birth who develop breast cancer
the proportion aged ≥ 35 at first birth who develop breast cancer

In order to study this question, we collect independent random samples of size 1700 from each of the above age at first birth categories and observe the number of breast cancer cases in each group:

Breast Cancer?	**Age at first birth** < 20	20-24	25-29	≥30	≥ 35	**Total**
Yes	313	364	440	507	1107	**2731**
No	1387	1336	1260	1193	593	**5769**
Total	**1700**	**1700**	**1700**	**1700**	**1700**	**8500**

The relevant test statistic is calculated using the same formula as in the 2 x 2 table analysis, except that here we will use all 10 cells of our 2 x 5 table. The test statistic for our 2 x 5 table will follow, approximately, a chi square distribution with (R-1) x (C-1) = (2-1) x (5-1) =4 degrees of freedom under the null hypothesis, as long as at least 80% of the expected values are ≥ 5.

Hypotheses:

Justification for the test:

Test statistic:

P-value:

Decision and conclusion:

§9.4 Chi-Square Test of Independence

Many different tests are based on the contingency-table approach, and the logic behind many of these tests is similar to the logic behind the test of homogeneity. The chi-square test of independence is no exception. In this test, the null hypothesis is that the row variable and column variable are independent, the alternative is that these variables are associated (i.e. not independent). Only one sample is collected in this test; once this sample is drawn from the population, each object/person is classified according to the two variables under study. An example where the test of independence is appropriate is shown below:

Example: We wish to investigate whether smoking (yes or no) is independent of education level (high

school completed, not completed). We randomly select 356 people and classify them with respect to smoking status and education level. The sample data are:

		Education: High school completed?	
		No	Yes
Smoke?	Yes	43	73
	No	50	190

In this example, one sample of 356 people was collected; the people sampled were **classified** into one of the four possible categories. This is in contrast to the test of homogeneity encountered in previous examples, where independent random samples were collected from each population of interest.

In this example, the hypotheses are written as:

H_0: Education and smoking are independent
H_A: Education and smoking are not independent

The test statistic is formula is exactly the same as for the test of homogeneity! Why? Read the optional section below…

The test statistic follows, approximately, a chi-square distribution with (R-1)(C-1) degrees of freedom, as long as at least 80% of the expected values are >= 5, just as in the test of homogeneity. Therefore, the computations are identical to those performed in the test of homogeneity. If the null hypothesis is rejected, we conclude that the row and column variables are not independent (i.e., they are associated); if we fail to reject, we do not have evidence to show that they are associated, and will continue to assume that they are independent.

Example: Complete the smoking and education example below:

Hypotheses: H_0: Education and smoking are independent
H_A: Education and smoking are not independent

Justification for the test:

Test statistic:

P-value:

Decision and conclusion:

§9.4.1 Optional: Rationale Behind the Construction of the Test Statistic for the Chi-Square Test of Independence

The test statistic is formula is exactly the same as for the test of homogeneity! Here's the rationale. Once again, we can investigate the hypotheses by comparing the **observed** counts in the cells of the contingency table with the counts we would **expect** to get if in fact the two variables were independent (i.e. if H_0 is true). If the observed and expected counts are 'very different' that would lead us to believe that the two variables are not independent (i.e. that the null hypothesis is false). If the counts are 'fairly similar' then the data do not suggest that the variables are dependent.

How do we calculate the expected counts for a test of independence?

Let's consider the top left cell consisting of people who dropped out of high school and smoked. The average number of people one would expect in this cell is equal to the probability of a randomly chosen person being a high school dropout **and** a smoker multiplied by the total number of people in our sample.

If the null hypothesis is true, then the probability of a randomly chosen person being a smoker **and** a high school dropout is equal to the probability of a randomly chosen person being a smoker multiplied by the probability of a randomly chosen person being a dropout (this is an application of the multiplication rule for independent events!) Therefore, the average number of people we would see in the top left cell of the table, assuming that H_o is true, is:

Average # of people in top left cell = P(smoker) x P(HS dropout) x 356

Of course, we do not know the true probability of someone being a smoker or the true probability of someone being a HS dropout; therefore, the expected count for this cell will estimate the average. We will estimate this average by substituting the sample probabilities in the formula above. That is, the expected count for the top left cell is:

E_{11} = Expected # of people who are smokers and dropouts
= Estimated P(smoker) x Estimated P(HS dropout) x 356
= (43+73)/356 x (43+50)/356 x 356
= (43+73) x (43+50)/356
= 30.30

The rationale behind the expected value calculations for each of the other cells is similar. Note that, in both the test of homogeneity and the test of independence, each expected value works out to be: the row total times the column total divided by the grand total.

Once the expected values have been calculated, the chi-square test statistic that we have already seen is used in the test of the hypotheses.