STT 231 STATISTICS FOR SCIENTISTS

Chapter 3: Inference for Categorical Data

Lecture 3-1: Parametric procedures for a single proportion

We can use the normal distribution to model the sampling distribution of a sample proportion $\hat{p}$

KEY IDEA: If certain conditions are met, the sampling distribution of $\hat{p}$ will be

nearly normal .

Condition #1: observations are independent

Condition #2: Sample is sufficiently large

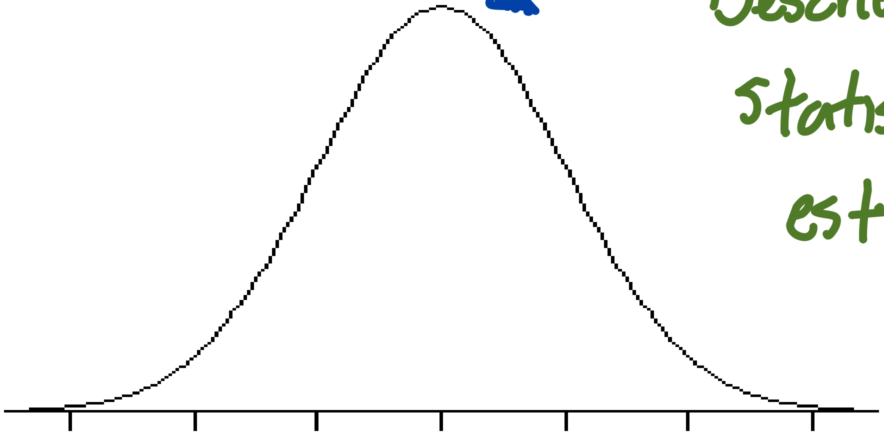
"Success-failure" requirement: $np \geq 10$ and $n(1-p) \geq 10$

Lecture 3-1: Parametric procedures for a single proportion

A quick sketch:

$N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

Describes possible values of statistic $\hat{p}$ as an estimator of p.



The mean of this normal distribution is p, true proportion/rate

The standard error is $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

The one-proportion z-test

The one-proportion z-test	
What is it?	A procedure for evaluating contradictory hypotheses about the true value of a single population proportion. The test statistic is: $z = \frac{\hat{p}-p_o}{SE_{\hat{p}}}$, where $SE_{\hat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}}$
What does it do?	It provides a formal way to quantify the likelihood of our observed data, or data more extreme, from a null distribution, which is a type of sampling distribution that is based on a particular value for a population proportion.
How does it do it?	The test statistic z computes the number of standard errors $\hat{p}$ fell from the value expected by the null model. Values that are very far from 0 tend to discredit the null hypothesis.
How is it used?	Independent observations must comprise the sample data and the sample size must be sufficiently large. If this is the case, the $N(0,1)$ distribution is used to find a p-value corresponding to the test statistic.

Page 66

Example 3.1: Xylitol & Ear Infections

p_0

- Roughly 38% of young children develop ear infections at various points before the age of five. Xylitol is a food sweetener that has recently gained scientific interest for its antibacterial properties.
- A Finnish experiment randomly divided children in a local daycare center into two groups. One group regularly chewed gum containing xylitol, and another regularly took xylitol lozenges.
- Over a three-month period, researchers recorded whether each child at any time developed an ear infection.

Page 66

Example 3.1: Xylitol & Ear Infections

Group	No. of children	Ear Infection? = 'Yes'
Xylitol gum	100	22
Xylitol lozenges	250	71

a. Conduct a hypothesis test to determine if regularly chewing Xylitol gum is associated with a *decreased* risk of ear infections.

Step 1: Determine the null and alternative hypotheses.

H_0: $p = 0.38$ H_a: $p < 0.38$

where the parameter p represents...

true rate of ear infections among all daycare-aged children who chew Xylitol gum.

Note: The direction of extreme is left-tailed

(determines how p-value is computed)

Page 66

Example 3.1: Xylitol & Ear Infections

Group	No. of children	Ear Infection? = 'Yes'
Xylitol gum	100	22
Xylitol lozenges	250	71

Step 2: Create a null distribution to see what's typical. In this case, we wish to use the __normal__ distribution for our null model. To do so, we much check the success-failure assumption for performing the test.

The data are assumed to be independent observations.

Check if $np_0 \geq 10$ and $n(1 - p_0) \geq 10$.

$$100(0.38) = \underline{38} \geq 10$$

No. of ear infections expected by null model

$$100(0.62) = 62 \geq 10$$

Page 66

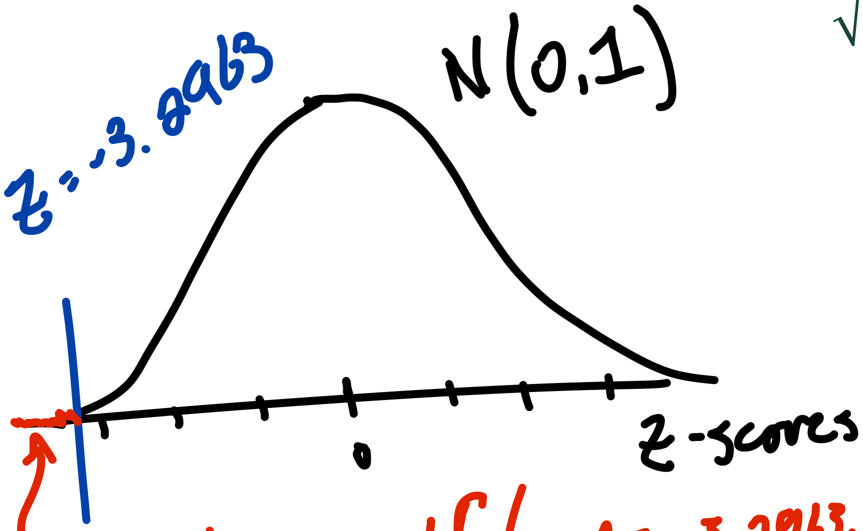
Example 3.1: Xylitol & Ear Infections $\hat{p} = \frac{22}{100}$

Group	No. of children	Ear Infection? = 'Yes'
Xylitol gum	100	22
Xylitol lozenges	250	71

Step 3: Evaluating evidence: calculate the test statistic and determine the *p*-value.

Observed test statistic: $$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.22 - 0.38}{\sqrt{\frac{0.38(0.62)}{100}}}$$

$$= \underline{-3.2963}$$



p-value = ncdf(-10^10, -3.2963, 0, 1)
= 0.00049

This means our observed $\hat{p}$ was 3.2963 standard errors <u>lower</u> than what was expected by H..

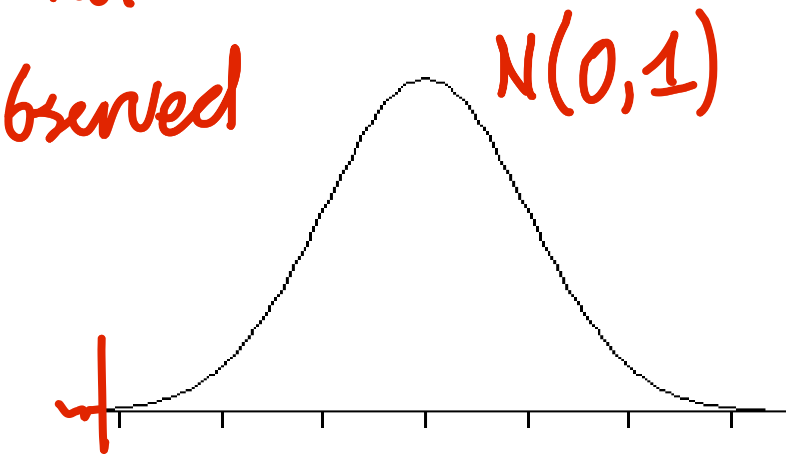
Page 66

Example 3.1: Xylitol & Ear Infections

Group	No. of children	Ear Infection? = 'Yes'
Xylitol gum	100	22
Xylitol lozenges	250	71

Step 4: Evaluate the *p*-value and report the conclusion in the context of the problem.

Our p-value of 0.00049 gives us very strong evidence the hypothesized ear infection rate of $p = 0.38$ does not adequately describe observed data.



Page 67

Example 3.1: Xylitol & Ear Infections

Group	No. of children	Ear Infection? = 'Yes'
Xylitol gum	100	22
Xylitol lozenges	250	71

b. Conduct a hypothesis test to determine if regularly taking Xylitol **lozenges** is associated with a *decrease* in risks of ear infection.

$H_0: p = 0.38$ vs. $H_a: p < 0.38$

$\hat{p} = 71/250 = 0.284$

$z = -3.1272$

p-value $= 0.00088$

Page 67

Example 3.1: Xylitol & Ear Infections

Group	No. of children	Ear Infection? = ‘Yes’
Xylitol gum	100	22
Xylitol lozenges	250	71

Step 1: Determine the null and alternative hypotheses.

H_0: $p = 0.38$ H_a: $p < 0.38$

where the parameter p represents...

true ear infection rate among daycare-aged children who take Xylitol lozenges.

Note: The direction of extreme is left tailed.

Page 67

Example 3.1: Xylitol & Ear Infections

Group	No. of children	Ear Infection? = 'Yes'
Xylitol gum	100	22
Xylitol lozenges	250	71

Step 2: The data are assumed to be independent observations.

Check if $np_0 \geq 10$ and $n(1 - p_0) \geq 10$.

$250(0.38) = 95 \geq 10$; $250(0.62) = 155 \geq 10$

Step 3: Evaluating evidence: calculate the test statistic and determine the *p*-value.

Observed test statistic: $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.284 - 0.38}{\sqrt{\frac{0.38(0.62)}{250}}}$

$\hat{p} = \frac{71}{250} = 0.284$

$= -3.1272$

Page 67

Example 3.1: Xylitol & Ear Infections

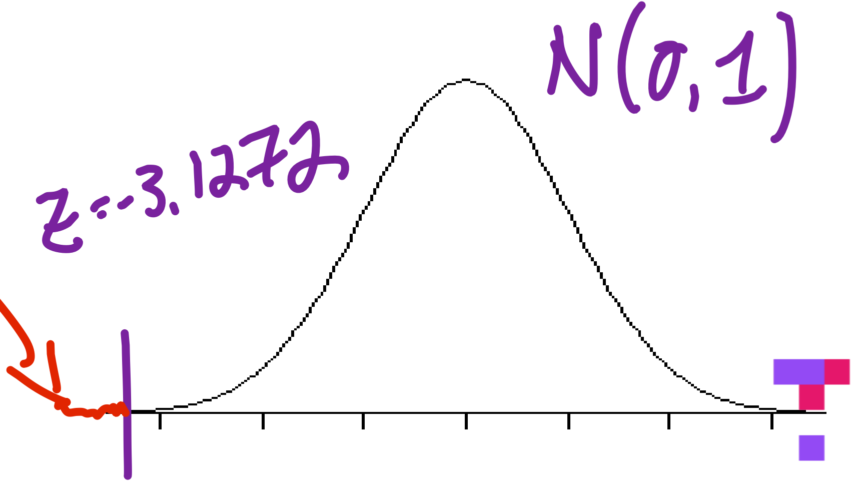
Group	No. of children	Ear Infection? = 'Yes'
Xylitol gum	100	22
Xylitol lozenges	250	71

Step 4: Evaluate the *p*-value and report the conclusion in the context of the problem.

Basically same conclusion as Xylitol gum...

p-value = ncdf(-10^10, -3.1272, 0, 1)

= 0.00088



Page 67

Example 3.1: Xylitol & Ear Infections

c. Consider the results of the two hypothesis tests above. We have good evidence that the null value of $p = 0.38$ is not a good fit for the ear infection risk of children regularly chewing gum or taking lozenges with Xylitol.

Would you recommend children take one over the other? Briefly justify your choice.

We have a discrepancy here: the p-values associated with the two tests are very similar, with only four-hundredths of a percentage point between them. That said, the observed sample proportions ARE meaningfully different, with xylitol gum having an observed ear infection rate of 0.22 and xylitol lozenges appearing about 6% less effective at reducing ear infection rates, with an observed rate of 0.284.

Is it worthwhile to recommend gum instead of lozenges, using the difference in observed sample proportions as evidence, even though the p-values are nearly identical?....

WE NEED A NEW WAY TO INTERPRET HYPOTHESIS TEST RESULTS!!

Introducing effect sizes

We are frequently interested in comparing a population parameter to a value specified by a null model or (in future lectures) to a parameter value of another population.

In many research situations, we want to estimate the

the size of the effect. (the size of the null model's inadequacy.)

The test statistic and p-value cannot help us here because they both

depend on sample size .

The ‘effect size’

Definition:

The effect size for a research study is a measure of how much the truth differs from the results expected by a statistical model.

A common formula for the effect size when conducting a hypothesis test regarding a single population proportion p is **Cohen’s *h*.** We estimate Cohen’s h with the following formula:

$$\hat{h} = 2 \cdot \arcsin\left(\sqrt{\hat{p}}\right) - 2\arcsin\left(\sqrt{p_0}\right)$$

★ In calculators, arcsin = $\sin^{-1}$ ★

(Radian mode, not degree mode)

Computing estimated effect sizes

Compute the effect size for each of the Xylitol hypothesis tests on the previous two pages.

a. **Gum:** In a z-test of $H_0 = 0.38$, we observed $\hat{p} = 0.22$. This resulted in a test statistic of $z_1 = -3.2963$ and p-value of 0.00049.

$$\hat{h} = 2\sin^{-1}(0.22)^{1/2} - 2\sin^{-1}(0.38)^{1/2} = -0.3520$$

b. **Lozenge:** In a z-test of $H_0 = 0.38$, we observed $\hat{p} = 0.284$. This resulted in a test statistic of $z_2 = -3.1272$ and p-value of 0.00088.

$$\hat{h} = 2\sin^{-1}(0.284)^{1/2} - 2\sin^{-1}(0.38)^{1/2} = -0.2043$$

Computing estimated effect sizes

Compute the effect size for each of the Xylitol hypothesis tests on the previous two pages.

c. Why are the z-scores in (a) and (b) approximately the same, if the effect sizes of the treatments are different?

$$z_a = \frac{0.22 - 0.38}{\sqrt{\frac{.38(.62)}{100}}} \qquad z_B = \frac{0.284 - 0.38}{\sqrt{\frac{0.38(0.62)}{250}}}$$

but z_B had a larger sample size.

Observed difference in $\hat{p} - p$ was larger in z_a

We'd consider the results in scenario (a) to have a Small-to-moderate effect size__________ and the results in scenario (b) to have a small effect size__________. For our purposes, Cohen's *h* is typically interpreted as one of four magnitudes.

Interpreting estimated effect sizes

$\hat{h}$	Effect size	Interpretation
$\|\hat{h}\| \leq 0.2$	small	We found an inadequacy in the null model so small we could only detect it with statistics.
$0.2 < \|\hat{h}\| \leq 0.5$	small-to-moderate	We found an inadequacy in the null model that someone with trained subject expertise would notice.
$0.5 < \|h\| \leq 0.8$	moderate-to-large	We found an inadequacy in the null model a careful observer would notice.
$\|\hat{h}\| > 0.8$	large	A problem with null model that is obvious to most observers.

Page 69

Lecture 3-1B: Confidence intervals

It is known that roughly two-thirds of humans have a dominant right foot or eye.

The article “Human Behavior: Adult Persistence of Head-Turning Asymmetry” reported that in a random sample of 125 kissing couples, both individuals in 80 of the couples leaned more to the right than the left.

Does this suggest that the two-thirds figure is implausible for kissing behavior?

two-tailed

Lecture 3-1B: Confidence intervals

Conduct the appropriate hypothesis test to answer the research question above. Include the procedure's test statistic, p-value, and estimate of Cohen's *h*.

$H_0: p = 0.6667$ $\quad H_a: p \underset{=}{\neq} 0.6667$

two-tailed

$\hat{p} = 80/125 = 0.64$

$$z = \frac{0.64 - 0.6667}{\sqrt{\frac{0.6667(.3333)}{125}}} = -0.6332$$

$$\text{p-value} = \underline{2}\ \underline{*}\ \text{normalcdf}(-10^{10}, -0.6332, 0, 1)$$

$$= 0.5266$$

$$\hat{h} = 2\sin^{-1}(0.64)^{1/2} - 2\sin^{-1}(.6667)^{1/2} = -0.0561$$

Analysis:

With p-val = 0.5266 and $\hat{h} = -0.0353$, we have little evidence to suggest $p = 2/3$ is a poor parameter to describe right-dominant kissing behavior.

Page 69-70

Lecture 3-1B: Confidence intervals

While establishing a null value after considering a sample statistic is *very* bad practice, it does show us that there as just as little evidence against using $p = 0.61$ as a rate for right-kissing behavior as $p = 0.67$, given our data.

In fact, there are many parameters that, if tested, would appear to be a plausible parameter for the process that created our observed data.

Rather than conduct numerous hypothesis tests, when statisticians want to create a range of plausible values for a population parameter, they create a

confidence interval.

The one-proportion confidence interval	
What is it?	A range of plausible values for a population parameter. For a single proportion, a confidence interval uses the formula below: $CI\colon\ \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
What does it do?	A confidence interval provides two pieces of information: (1) a range of plausible values for a population parameter and (2) a level for this range. The level corresponds to the reliability of the process used to create the interval.
How does it do it?	It uses the sample statistic $\hat{p}$ and its standard error to create an interval of plausible values for the parameter p. The way the confidence interval does this has a pre-specified success-rate of capturing the parameter, called the level of the interval. For instance, a 95% confidence interval is computed using a process that has a 95% success rate.
How is it used?	Independent observations must comprise the sample data and the sample size must be sufficiently large. If this is the case, a particular multiplier z^* is chosen by the statistician to dictate the width of the interval.

Example 3.2 – CI for right-kissing couples

Recall the earlier article “Human Behavior: Adult Persistence of Head-Turning Asymmetry,” which reported that in a random sample of 125 kissing couples, both individuals in 80 of the couples leaned more to the right than the left.

a. Use this information to create a 90% confidence interval, i.e., a range of plausible values modeling the true rate at which couples exhibit ‘right-kissing’ behavior.

$\hat{p} = 80/125 = 0.64$

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

$0.64 \pm 1.645 \sqrt{\frac{0.64(0.36)}{125}}$

Confidence interval: (0.5694 , 0.7106)

Example 3.2 – CI for right-kissing couples

b. Provide an interpretation of this confidence interval, along with its *level*.

With 90% confidence, we think the true rate of right dominant kissing is between 56.94% and 71.06%.

By "90% confidence," we mean to say that the method used to create these bounds has a 90% success rate at capturing the true parameter p.

Example 3.2 – CI for right-kissing couples

c. As a follow-up, create a 95% confidence for the true rate at which couples exhibit ‘right-kissing’ behavior. How does it compare to the interval in (a)?

$$0.64 \pm 1.96\sqrt{\frac{0.64\,(0.36)}{125}} \approx (0.5559,\ 0.7241)$$

KEY IDEA: Holding all else the same, confidence intervals that have a higher confidence level are wider than those with lower confidence levels.

Example 3.3 – Vaccine Efficiency

Annual iterations of the common flu vaccine are developed by growing cells in fertilized chicken eggs. 3900 subjects were administered vaccines developed by this method and 24 of them contracted the flu during the 28-week observational period that followed.

a. Estimate the true proportion of individuals who will develop the flu despite being given this vaccination. Give the standard error for the sample proportion interpret it in terms of an average distance.

$\hat{p} = 24/3900 = 0.0062$

$SE_{\hat{p}} = \sqrt{\frac{.0062(.9938)}{3900}}$

$= 0.0013$

We would estimate the average distance between the possible $\hat{p}$ values (from repeated samples) and true parameter p to be about 0.13%.

Page 71

Example 3.3 – Vaccine Efficiency

b. Create a 99% confidence interval for the true proportion of individuals who will develop the flu despite being given this vaccination, if it is produced en masse.

$$\hat{p} \pm z^* SE_{\hat{p}} = 0.0062 \pm 2.576(0.0013)$$

Confidence interval: (0.0029 , 0.0095)

Page 71

Example 3.3 – Vaccine Efficiency

c. A 95% confidence interval produced from the same survey results would be

[✓ *Narrower*] *wider* *the same width as*

the interval computed in (b).

Page 71-72

Example 3.3 – Vaccine Efficiency

d. Fill in the blanks for the typical interpretation of the confidence interval in part b:

"Based on this sample, with 99% confidence, we would estimate that somewhere between 0.29% and 0.95% of *all* individuals receiving this vaccine will later develop the flu."

e. Can we say the probability that the above (already observed) interval contains the population proportion *p* is 0.99?

No!

f. Can we say that 99% of the time the population proportion *p* will be in above (already observed) interval you computed in part (b)?

No!

Example 3.3 – Vaccine Efficiency

A common mistake laymen make is to turn a confidence interval into some

quantification of belief.

For example, if asked to interpret a 99% interval of $(0.003, 0.009)$, many people will incorrectly state…

> *"There is a 99% chance that the population proportion is between 0.3% and 0.9%"*

g. Why is this wrong?

The "chance" is not 99%. It is either 0% or 100% and we cannot verify it one way or another. Similarly, a flipped coin no longer has a "chance" of heads after it has landed.

Page 72

Example 3.4: Misinterpretations of confidence intervals

Suppose a student is working on a statistics project using data on pulse rates collected from a random sample of 100 students from her college. She finds a 95% confidence interval for the mean pulse rate of (65.5, 71.8). Each of the interpretations below are *incorrect* interpretations of this interval. Can you determine why?

population

a. I am 95% sure that the mean pulse rate for this sample of students will fall between 65.5 and 71.8 BPM.

95% the mean of all students

b. I am sure that 95% of all students at this college will have pulse rates between 65.5 and 71.8 BPM.

c. The mean pulse rates for students at this college will fall between 65.5 and 71.8 BPM 95% of the time.

nope!

Page 72

Lecture 3-2: Normal procedures for $\boldsymbol{p_1 - p_2}$

There are many cases in which we would like to make conclusions about the difference in two population proportions $p_1 - p_2$ using the normal model in a manner similar to the examples explored in Lecture 3-1.

To draw inferences about a difference in population proportions $p_1 - p_2$, we will use the sampling distribution of two sample proportions, $\hat{p}_1 - \hat{p}_2$.

Page 73

Lecture 3-2: Normal procedures for $\boldsymbol{p_1 - p_2}$

KEY IDEA: If certain conditions are met, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ can be modeled by a normal distribution.

Condition #1:

Observations are independent within and between samples.

(i.e., success-failure assumption)

Condition #2:

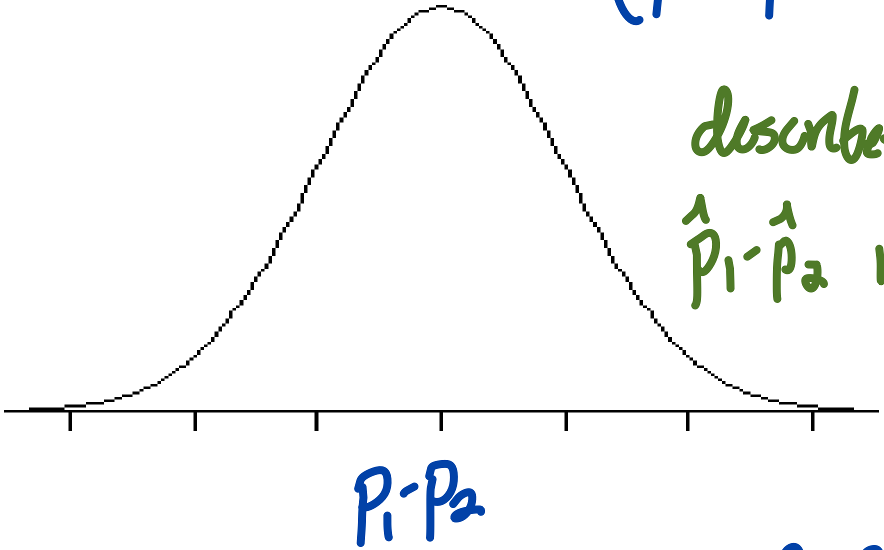
Both samples meet the success-failure condition.

Lecture 3-2: Normal procedures for $p_1 - p_2$

A quick sketch:

$$N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

describes possible values of $\hat{p}_1 - \hat{p}_2$ when sampled from population.



The mean of this normal distribution is $p_1 - p_2$, true difference in group rates

The standard error is $SE_{\hat{p}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Summary of two-proportion formulas

Note that the way to compute $SE_{\hat{p}_1-\hat{p}_2}$ depends on the procedure!	
Hypothesis test	$z = \frac{\hat{p}_1 - \hat{p}_2}{SE_{\hat{p}_1-\hat{p}_2}}$, where $SE_{\hat{p}_1-\hat{p}_2} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ and $\hat{p} = \frac{x_1+x_2}{n_1+n_2}$
Effect size	$\hat{h} = 2\arcsin\left(\sqrt{\hat{p}_1}\right) - 2\arcsin\left(\sqrt{\hat{p}_2}\right)$
Confidence interval	$\hat{p}_1 - \hat{p}_2 \pm z^* SE_{\hat{p}_1-\hat{p}_2}$, where $SE_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

Page 73

Two different ways to calculate $SE_{\hat{p}_1-\hat{p}_2}$

- The approximate normal model for the sampling distribution of the difference in sample proportions requires that the quantities n_1p_1, $n_1(1-p_1)$, n_2p_2, and $n_2(1-p_2)$ are at least 10. Since the population proportions are unknown, we will check if this assumption is reasonable using observed sample data.

- For **confidence interval estimation** we will need that the quantities $n_1\hat{p}_1$, $n_1(1-\hat{p}_1)$, $n_2\hat{p}_2$, and $n_2(1-\hat{p}_2)$ are all at least 10. If any of these four quantities fall below 10, the normal approximation to the sampling distribution of $\hat{p}_1-\hat{p}_2$ becomes unreliable. For **hypothesis testing**, we will replace the population proportions with an appropriate estimate assuming the null hypothesis is true, so watch for that subtle difference in checking the large sample sizes assumption.

Page 74

Example 3.5: Smoking and lung obstruction

The National Center for Health Statistics looked at the association between lung obstruction and smoking status in adults ages 40 to 79. In a random sample of 6297 adults without any lung obstruction, 54.1% never smoked. In a random sample of 1146 adults with lung obstruction (such as asthma or COPD) 23.8% never smoked.

a. Find and interpret a point estimate of the true difference between the proportion of adults without and with lung obstruction who never smoked. (That is, find your best estimate of $p_1 - p_2$).

$$\hat{p}_1 - \hat{p}_2 = 0.541 - 0.238 = 0.303$$

Page 74

Example 3.5: Smoking and lung obstruction

The National Center for Health Statistics looked at the association between lung obstruction and smoking status in adults ages 40 to 79. In a random sample of 6297 adults without any lung obstruction, 54.1% never smoked. In a random sample of 1146 adults with lung obstruction (such as asthma or COPD) 23.8% never smoked.

b. Compute the 95% confidence interval.

$$SE = \sqrt{\frac{.541(.459)}{6297} + \frac{.238(.762)}{1146}} = 0.0141$$

$$CI: \hat{p}_1 - \hat{p}_2 \pm z^* SE = 0.303 \pm 1.96(0.0141)$$

$$= (0.2754, 0.3306)$$

Page 74

Example 3.5: Smoking and lung obstruction

c. The statement below interprets the *level* of this interval. Comment on the manner in which it is incorrect.

> *We have a 95% interval for $p_1 - p_2$. The 95% level of this interval means there is a 95% chance that $p_1 - p_2$ is actually contained between the lower and upper bounds of the interval.*

Statement incorrectly!! applies a probability to the interval capturing the parameter.

Page 74

Example 3.5: Smoking and lung obstruction

d. In order for the interval in (c) to be valid, certain conditions must be met. State these conditions below and check to see whether they're safe to make.

The observations must be independent within and between samples, which is safe to assume given description & data collection in prompt.

Additionally, $n_1\hat{p}_1$, $n_1(1-\hat{p}_1)$, $n_2\hat{p}_2$, and $n_2(1-\hat{p}_2)$ should all be at least 10.

Page 75

Example 3.5: Smoking and lung obstruction

e. Briefly describe the consequences of computing a confidence interval as we did in (b) when the conditions in (d) are not met.

See AHR 3.2.

Page 75

Example 3.6: Ionizing your groceries

Ionizing radiation is gaining increased attention as a technique for preserving groceries found in the produce section. A recent study reported that 153 out of 180 irradiated garlic bulbs were of sellable status 240 days after treatment, compared to only 119 out of 180 untreated garlic bulbs.

a. Use this information to create a 99% confidence interval for the percentage increase we could expect in sellable status attributable to ionizing radiation.

Testing $H_0: p_1 = p_2$

We want to investigate whether there is an increased risk of cancer in dogs that are exposed to herbicide 2,4-dichlorophenoxyacetic acid (2,4-D). Study examined 491 dogs that had developed cancer and 945 dogs as a control group. Of these two groups, researchers identified which dogs had been exposed to 2,4-D in their owner’s yards. The results are shown below:

	Cancer	No cancer	Total
2,4-D	191	304	495
No 2,4-D	300	641	941
Total	491	945	1436

Testing $H_0: p_1 = p_2$

	Cancer	No cancer	Total
1 2,4-D	191	304	495
2 No 2,4-D	300	641	941
Total	491	945	1436

b. **STEP 1:** Set up the appropriate hypotheses to test for an investigation as to whether there is an *increased* risk of cancer in dogs exposed to 2,4-D.

H_0: $p_1 = p_2$ vs H_a: $p_1 > p_2$

Creating the null distribution

Step 2: Create a null distribution to see what's typical. In this case, we wish to summarize our sample data with a z-score and use the N(0,1) distribution for our null model. To do so, we much check the success-failure assumption for performing the test.

We're assuming (under H_0) that $p_1 = p_2$, so best estimate of common cancer incidence rate across all dogs (regardless of 2-4D exposure) is their weighted average:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Page 76

Recap: $SE_{\hat{p}_1-\hat{p}_2}$ for hypothesis tests

When the null hypothesis is that $p_1 = p_2$, it is useful to find the pooled estimate of the shared proportion between the two populations 1 & 2.

	Cancer	No cancer	Total
2,4-D	191	304	495
No 2,4-D	300	641	941
Total	491	945	1436

$$\hat{p} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2} = \frac{191+300}{495+941} = \frac{491}{1436} = 0.3419$$

Note $\hat{p}_1 = \frac{191}{495} = 0.3859$ and $\hat{p}_2 = \frac{300}{941} = 0.3188$

Recap: $SE_{\hat{p}_1-\hat{p}_2}$ for hypothesis tests

- Step 2: Check success-failure assumption for performing the test.

$n_1\hat{p} = 495(0.3419) = 155.88$

$n_2\hat{p} = 941(0.3419) = 321.73$

Recap: $SE_{\hat{p}_1-\hat{p}_2}$ for hypothesis tests

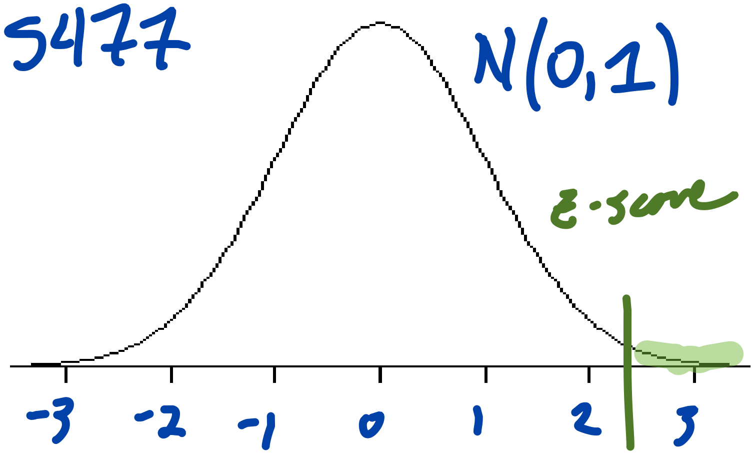
Step 3: Calculate the test statistic and determine the *p*-value.

Observed test statistic:

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

$$= \frac{0.3859-0.3188}{\sqrt{0.3419(0.6581)\left(1/495+1/441\right)}}$$

$$= 2.5477$$



p-value: 0.0054

Cohen's $\hat{h}$= 0.1406

Recap: $SE_{\hat{p}_1 - \hat{p}_2}$ for hypothesis tests

Step 4: Evaluate the *p*-value & effect size and report the conclusion in the context of the problem.

We have very strong evidence (p-value $= 0.0054$) that the cancer incidence rate is higher among dogs exposed to 2-4D then those who aren't. That said, $\hat{h} \approx 0.1406$, a small effect size, and the observed difference in incidence rates was only $\hat{p}_1 - \hat{p}_2 = 0.0671$.

Page 77

Prenatal Vitamins and Autism

Researchers studying the link between maternal use of prenatal vitamins and autism surveyed the mothers of a random sample of children aged 24 – 60 months with autism and conducted another sample of children with typical development.

Step 1: State the hypotheses appropriate for testing the *independence* of prenatal vitamin use and autism diagnosis:

H_0: $p_1 = p_2$ versus H_a: $p_1 \neq p_2$

- p_1 represents the population proportion of children whose mothers used prenatal vitamins later diagnosed with autism.
- p_2 represents the population proportion of children whose mothers did not use prenatal vitamins later diagnosed with autism.

Prenatal Vitamins and Autism

Group	Autism	Typical Development	Total
Vitamin	111	70	181
No Vitamin	143	159	302
Total	254	229	483

Step 2: Assume these samples are independent random samples. Verify the remaining assumption necessary to conduct the Z test.

$$\hat{p} = \frac{111+70}{483} = 0.3747$$

$$n_1\hat{p} = 229(0.3747) = 85.8063 \quad \checkmark$$

Prenatal Vitamins and Autism

Group	Autism	Typical Development	Total
Vitamin	111	70	181
No Vitamin	143	159	302
Total	254	229	483

Step 3: Conduct the test.

2 Prop Z test

$x_1 = 111$

$n_1 = 181$

$x_2 = 143$

$n_2 = 302$

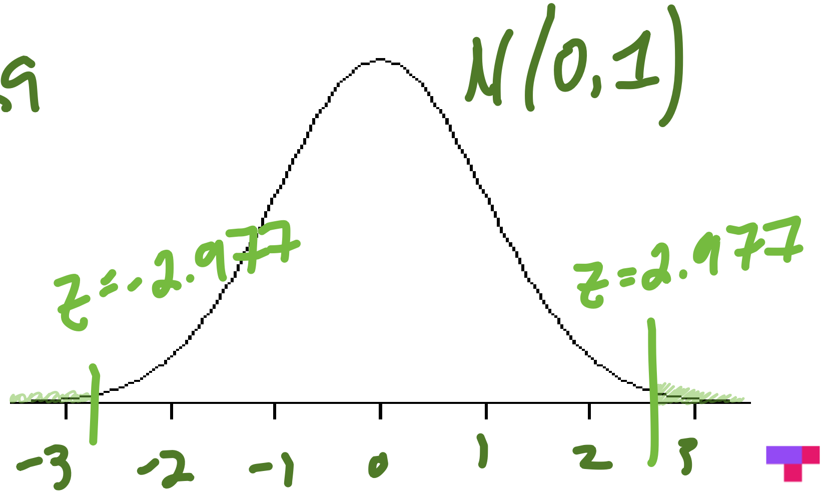
$H_a: \neq$

z-score = 2.977

p-val = 0.0029

$\uparrow$

$h = 0.2728$



Page 77

Prenatal Vitamins and Autism

Step 4: What is the appropriate conclusion?

We have very strong evidence (p-val $= 0.0029$) that prenatal vitamin use is related to autism development. The model that assumes $p_1 = p_2$ (i.e., independence) does not adequately fit the observed data. The size of the association is estimated as $\hat{h} = 0.2728$, which we'd consider small-to-moderate.

χ^2 procedures for categorical data

So far in Chapter 3, we've looked at inferential procedures for binomial characteristics that can be summarized with a proportion or with a difference between two proportions.

We now explore methods for assessing categorical data with **more than two outcomes**.

χ^2 procedures for categorical data

Below is data from a random sample of 275 jurors in a small county, sorted by their reported racial identity. We are interested in whether jurors are racially representative of the population.

If the jury *is* representative of the population, the proportions in the sample should resemble the population of eligible jurors, i.e. registered voters.

Race	White	Black	Hispanic	Other	Total
Representation in juries	205	26	25	19	275
Registered voters	0.72	0.07	0.12	0.09	1.00
Expected	0.72(275) 198	0.07(275) 19.25	0.12(275) 33	0.09(275) 24.75	

χ^2 procedures for categorical data

The hypotheses:

H_0: The jurors are a random sample, i.e., there is no racial bias in who serves on a jury, and the observed counts reflect natural sampling fluctuation.

H_a: The jurors are not randomly sampled, i.e. there is a racial bias in juror selection.

The χ^2 test statistic

Race	White	Black	Hispanic	Other	Total
Observed data	205	26	25	19	275
Expected counts	198	19.25	33	24.75	275

In previous hypothesis tests, we constructed a test statistic using the following form:

$$test\ statistic = \frac{observed - expected}{(null)\ SE\ of\ observed}$$

What would this test statistic be for individuals in the *White* category?

$$\chi_1 = \frac{205-198}{\sqrt{198}} = 0.4975$$

The χ^2 test statistic

Race	White	Black	Hispanic	Other	Total
Observed data	205	26	25	19	275
Expected counts	198	19.25	33	24.75	275

What would this test statistic be for the ‘Other’ Group?

$$\chi_1 = \frac{205-198}{\sqrt{198}} = 0.4975 \qquad \chi_3 = -1.3925$$

$$\chi_2 = \frac{26-19.25}{\sqrt{19.25}} = 1.5385 \qquad \chi_4 = -1.1558$$

$$\sum \chi_i = 0.4975 + 1.5385 + (-1.3925) + (-1.1558)$$

$$= -0.5123 \text{ ?!?}$$

The χ^2 test statistic

Summing all these test statistics gives a value that summarizing how far the actual counts are from what was expected. As it turns out, it is more common to add the squared values.

$$\sum \chi_i^2 = 0.4975^2 + 1.5385^2 + (-1.3925)^2 + (-1.1558)^2$$

$$= \underbrace{5.89}_{\text{summarizes total deviations from expected counts across entire table.}}$$

The χ^2 test statistic

Summing the squared test statistics has two consequences:

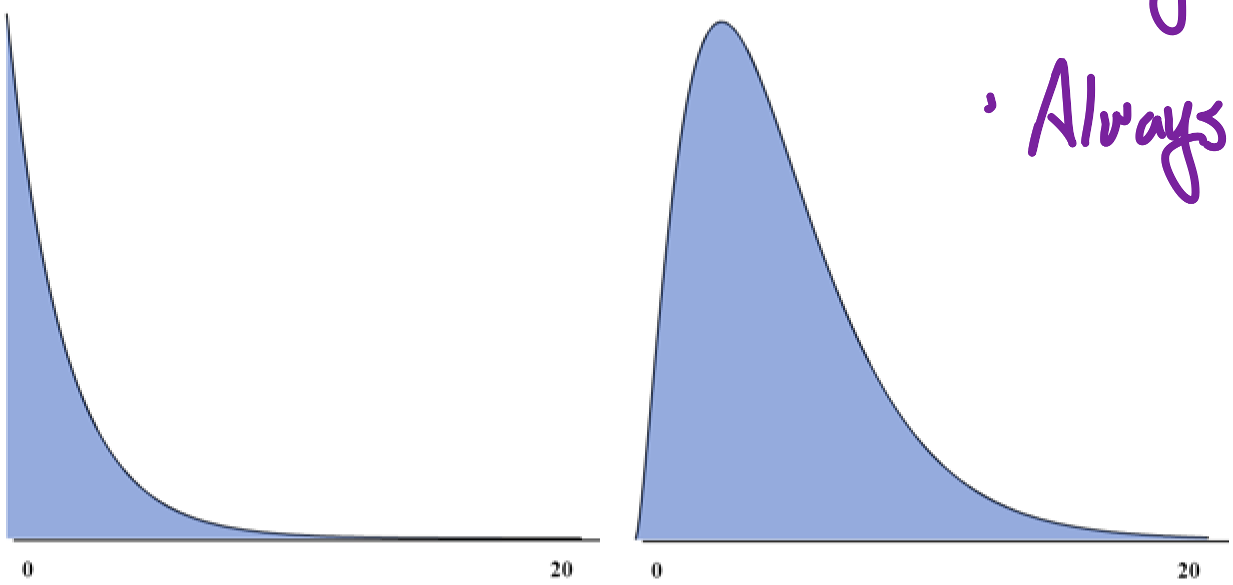
1. All χ^2 test statistics now positive!

2. Cells already unusual are further emphasized.

This final test statistic (in this case, the value 5.89) is called the χ^2 test statistic. We can use this statistic to compute a *p*-value and, thereafter, evaluate the two hypotheses of our test.

The χ^2 test

- Always right-skewed
- Always ≥ 0


The Chi-Square Distribution

- Skewed to the right
- Only take on positive values

Mean of a χ^2 distribution is...

$E(\chi^2) =$ df

Standard deviation of a χ^2 distribution is...

$\text{sd}(\chi^2) = \sqrt{2df}$

Using the χ^2 function

Consider the $\chi^2(\mathbf{4})$ distribution.

a. What is the mean?

df = 4

b. What is the standard deviation for this distribution?

$\sigma = \sqrt{2(4)} = 2\sqrt{2}$

c. How likely would it be to witness $\chi^2 \geq \mathbf{4}$?

0.4060

d. How likely would it be to witness $\chi^2 \geq \mathbf{10.3}$?

0.0357

Page 81

Example: Are jurors representative of their county?

Race	White	Black	Hispanic	Other	Total
Observed data	205	26	25	19	275
Expected counts	198	19.25	33	24.75	275

H_0: The jurors are a random sample, i.e., there is no racial bias in who serves on a jury, and the observed counts reflect natural sampling fluctuation.

H_a: The jurors are not randomly sampled, i.e. there is a racial bias in juror selection.

The χ^2 test statistic is: $0.5^2 + 1.54^2 + (-1.39)^2 + (-1.16)^2 = 5.89$

Example: Are jurors representative of their county?

To calculate the *p*-value associated with $\chi^2 = 5.89$, we need to determine the appropriate degrees of freedom.

When evaluating a one-way table such as the one above, we use K-1 degrees of freedom.

Because there were 4 racial categories, we should calculate the *p*-value using the $\chi^2_{(3)}$ distribution.

Example: Are jurors representative of their county?

What conclusion should we make regarding the juror representation?

We have little evidence (p-value = 0.1171) against the claim that the jury selection process is unbiased & representative of its community.

The χ^2 independence test

- A 2010 study was conducted to determine whether the drug Nuvigil, a drug commonly prescribed for narcolepsy and sleep apnea, was effective at helping east-bound jet passengers adjust to jet lag.
- Subjects were randomly assigned either to one of three different doses of Nuvigil (low, medium, high) or to a placebo, flown to France in a plane in which they could not drink alcohol or coffee or take sleeping pills, and then examined in a lab where their state of wakefulness was classified as one of three categories (low, normal, alert).

Page 82

The χ^2 independence test

Treatment	Low	Normal	Alert	Total
High	42	36	25	103
Medium	47	42	16	105
Low	49	32	17	98
Placebo	73	27	21	121
Total	211	13~~2~~7	~~84~~ 79	427

H_0: The treatment a patient was given has no effect on their state of wakefulness OR, alternatively,

"wakefulness" and treatment are independent

H_a: The treatment a patient was given has effect on their state of wakefulness OR, alternatively,

"wakefulness" and treatment are associated

The χ^2 independence test

Treatment	Low	Normal	Alert	Total
High	42	36	25	**103**
Medium	47	42	16	**105**
Low	49	32	17	**98**
Placebo	73	27	21	**121**
Total	**211**	**13~~2~~7**	~~**84**~~ 79	**427**

The data available are like those explored in Part 3. But the different combinations of the two variables [treatment and wakefulness] are binned into a *two-way* table.

Because the data are provided in a two-way table, the expected counts and degrees of freedom for our χ^2 test will be computed differently than before.

Computing expected counts for a two-way table

Treatment	Low	Normal	Alert	Total
High	42	36	25	103
Medium	47	42	16	105
Low	49	32	17	98
Placebo	73	27	21	121
Total	211	13~~2~~7	~~84~~ 79	427

To compute the expected counts for a two-way table,

we compute $Expected = \dfrac{(\text{Column total})(\text{Row total})}{\text{overall total}}$

Computing expected counts for a two-way table

Treatment	Low	Normal	Alert	Total
High	42 (50.9)	36 (33.0)	25 (19.1)	103
Medium	47 (51.9)	42 (33.7)	16 (19.4)	105
Low	49 (48.4)	32 (31.4)	17 (18.1)	98
Placebo	73 (59.8)	27 (38.8)	21 (22.4)	121
Total	211	137	~~84~~ 79	427

$$\chi^2 = \frac{(42-50.9)^2}{50.9} + \dots + \frac{(21-22.4)^2}{22.4}$$

Adding the computed values for each cell gives the overall test statistic. In this case it is χ^2= 13.4788 .

Computing expected counts for a two-way table

Just as before, this test statistic follows a χ^2 distribution. To use this distribution to calculate a p-value, we need to know its degrees of freedom. For a two-way table,

$df =$ (# of rows − 1)(# of colums − 1)

In the case of the Nuvigil experiment, $df =$ 6.

Compute the p-value for this test. p-val = 0.0360 .

Page 83-84

Effect size for χ^2 tests of independence

Additionally, we can compute an estimated effect size for the level of association between two categorical variables using a statistic called *Cramer's V*.

The way this statistic should be interpreted is (as always the case with effect sizes) dependent on the context being studied.

$$Cramer's\, V = \sqrt{\frac{\chi^2}{n(k-1)}}$$

where
k-1 is the
smallest dimension
of the two-way table

b. Estimate the size of the effect for the Nuvigil experiment. Combine it with your p-value from (a) to draw a conclusion in the context of the study.

$$V = \sqrt{\frac{13.4788}{427\ (2)}} = 0.1256$$

Example: Painkillers & Pregnancy

- A recent study examined the relationship between miscarriage and the use of painkillers during pregnancy. 1009 pregnant women interviewed about use of painkillers. The researchers then recorded which of the pregnancies were successfully carried to term.

Treatment	Miscarriage	No miscarriage	Total
NSAIDS	18	57	75
Acetaminophen	24	148	172
No painkiller	103	659	762
Total	145	864	1009

Example: Painkillers & Pregnancy

Treatment	Miscarriage	No miscarriage	Total
NSAIDS	18 10.78	57 64.22	75
Acetaminophen	24 24.72	148 147.3	172
No painkiller	103 104.5	659 652.5	762
Total	145	864	1009

a. Does there appear to be an association between having a miscarriage and the use of painkillers?

$\sum \chi_i^2 = 6.1269$ ~ $df = (3-1)(2-1) = 2$

$\text{p-value} = \chi^2\text{cdf}(6.1269, 10^{\wedge}10, 2) = 0.0467$

$$V = \sqrt{\frac{6.1269}{1009\,(1)}} = 0.0779$$

Page 84

Example: Painkillers & Pregnancy

Treatment	Miscarriage	No miscarriage	Total
NSAIDS	18	57	75
Acetaminophen	24	148	172
No painkiller	103	659	762
Total	145	864	1009

b. If there is an association, can we conclude that the use of painkillers increases the chance of having a miscarriage?

Nope! Observational studies can rarely be used to justify causal claims.

Example: Painkillers & Pregnancy

Treatment	Miscarriage	No miscarriage	Total
NSAIDS	18 A	57 D	75
Acetaminophen	24 B	148 E	172
No painkiller	103 C	659 F	762
Total	145	864	1009

c. Suppose, in fact, 1019 women were interviewed in the study and the table above is missing the responses and pregnancy outcomes of 10 of these women. In what cell would these women have the largest impact on the value of the χ^2 test statistic? The least?

Largest → A

Smallest → F