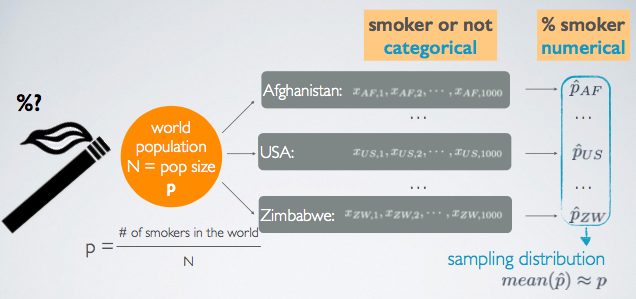
**Inference for Categorical Data**

YOUNGJIN LEE

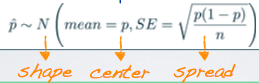
**Sampling Variability and CLT for Categorical Data**

**Sampling Distribution for Categorical Data**



* Remember a sampling distribution is different from a sample distribution.
* For categorical data, the sampling distribution is composed of the mean proportions from each sample.

**CLT for Proportions**



The Central Limit Theorem tells

1. the shape,
2. the center,
3. the spread of the distribution.

* The sampling distribution of sample proportions is nearly normal, centered at the population proportion, and with a standard error inversely proportional to the sample size.

**Conditions for the CLT**

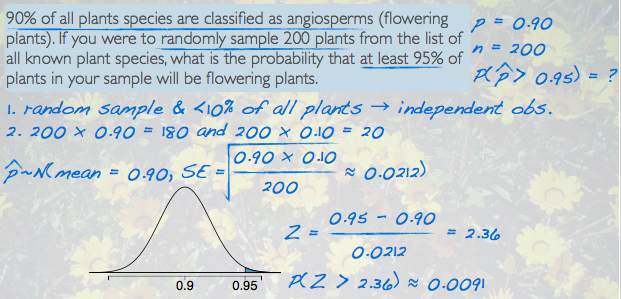
1. Independence: Sampled observations must be independent.

* Random sample / assignment
* If sampling without replacement, n < 10% of population

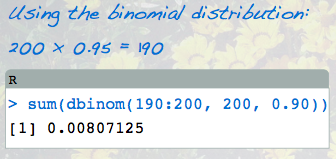
1. Sample size/skew: There should be at least 10 successes and 10 failures in the sample.

* Np >= 10 and n(1-p) >= 10
* The same idea of success and failure condition as in the normal approximation of a binomial distribution holds here as the sample proportion needs to be nearly normally distributed.
* When considering the sampling distribution of sample proportions, we don’t have a requirement of *n* ≥ 30. To determine if the sample size of categorical data is high enough, we instead check the success-failure condition.

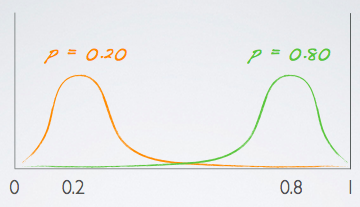
**Example)**



Since in this case the observation is a sample proportion (sampling distribution of the mean proportions), standard deviation of that is going to be measured by the standard error, and that gives us a Z score of 2.36.



**Shape of the Sampling Distribution**



If the success-failure condition is not met:

* The center of the sampling distribution will still be around the true population proportion
* The spread of the sampling distribution can still be approximated using the same formula for the standard error.
* The shape of the sampling distribution will depend on whether the true population proportion is closer to 0 or closer to 1.

**Back to example)**

What would you expect the shape of the sampling distribution of percentages of angiosperms in random samples of 50 plants to look like? (Remember, 90% of all plants species are classified as angiosperms.)

The success-failure condition is not met:

50 \* 0.9 = 45 >10 but 50 \* 0.1 = 5 < 10

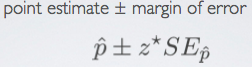
Therefore, the CLT doesn’t apply and the sampling distribution is not nearly normal. Since the true population proportion is close to 1, and the center of the sampling distribution will be at the true population proportion, we expect a shorter tail on the right side and longer tail on the left, yielding a left skewed distribution.

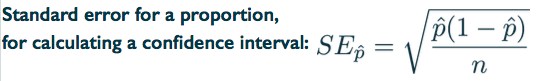
**Confidence Interval for a Proportion**

**Parameter of Interest: p**

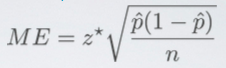
**Point Estimate: p̂**

**Estimating a Proportion**

****



**Calculating the Required Sample Size for Desired ME (Margin of Error)**



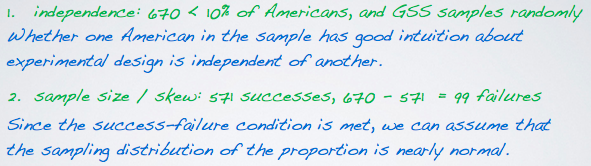
* If we don’t have any knowledge about the characteristics of the population and there is a previous study that we can rely on for the value of p̂, use that in the calculation of the required sample size
* If not, use p̂ = 0.5
  + If you don’t know any better .5 is a good guess
  + Gives the most conservative estimate (highest ME) – highest possible sample size

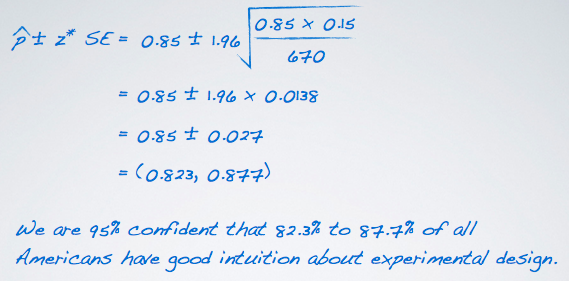
**INSTRUCTIONS**

1. Check the conditions
2. Construct a confidence interval by estimating the proportion

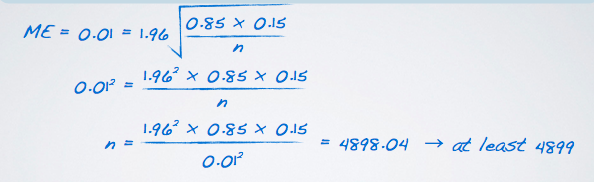
**Example)**

The GSS found that 571 out of 670 (~85%) of Americans answered the question on experimental design correctly. Estimate (using a 95% confidence interval) the proportion of all Americans who have good intuition about experimental design?



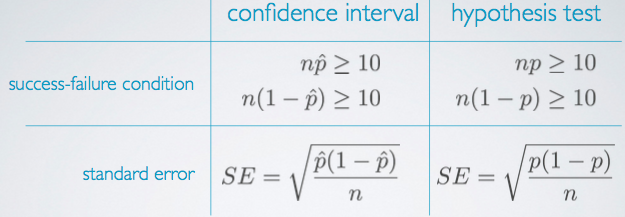


The margin of error for the previous confidence interval was 2.7%. If, for a new confidence interval based on a new sample, we wanted to reduce the margin of error to 1% while keeping the confidence level the same, at least how many respondents should we sample?



**Hypothesis Test for a Single Proportion**

**Working with One Proportion: p̂ vs p**



* Use the sample proportion p̂ for confidence interval when when you don’t know the population proportion.
* Use the population proportion p or at least the null hypothesized value of the population proportion when we’re doing a hypothesis test which dictates that we must assume that the null hypothesis is true

**INSTRUCTIONS**

1. Set the hypothesis

* H0 : p = null value
* H1 : p < or > or != null value

1. Check conditions
2. Calculate the point estimate p̂
3. Draw sampling distribution, calculate test statistic, shade p-value,

../../Screen%20Shot%202018-08-06%20at%2011.42.02%20PM.png

1. Make a decision, and interpret it in context of the research question:

* If p-value < alpha, reject H0; the data provide convincing evidence for H1
* If p-value > alpha, fail to reject H0 the data do not provide convincing evidence for H1
* Draw the curve and shape where the p value belongs to before calculating the p value
* Used p, instead of p hat, and that is because in a hypothesis test, we have to assume that the null hypothesis is true. If you think about the definition of a p value, it says, probability of observed or mare, more extreme outcome, if the null hypothesis is true. So, when going through the conditions, or any other portion of the hypothesis test, we must assume that the null is true, and therefore, wherever we see a p, we plug in whatever the null value for that p is, that's set forth in the null hypothesis. So, we could read this as not ten observed successes and ten observed failures, but instead as ten expected successes and ten expected failures.

*See example*

**Estimating the Difference Between Two Proportions**

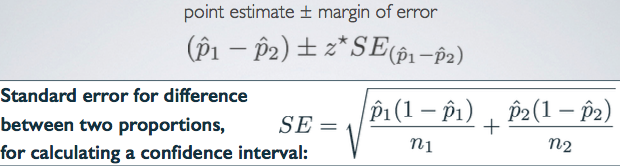
**Parameter of Interest: pA - pB**

Difference between the proportions of population A and population B

**Point Estimate: p̂A - p̂B**

Difference between the proportions of sampled A and sampled B

**Estimating the Difference Between Two Proportions**

****

**Conditions for Inference for Comparing Two Independent Proportions**

1. Independence:

* Within groups: sampled observations must be independent within each group
  + Random sample / assignment
  + If sampling without replacement, n < 10% of population
* Between groups: the two groups must be independent of each other (non-paired)

1. Sample size / skew: Each sample should meet the success-failure condition

* n1p1 ≥ 10 and n1(1-p1) ≥ 10
* n2p2 ≥ 10 and n2(1-p2) ≥ 10

**INSTRUCTIONS**

1. Check the conditions

* if we are not 100% sure about random sampling condition for one group, we need to be careful at making conclusions.

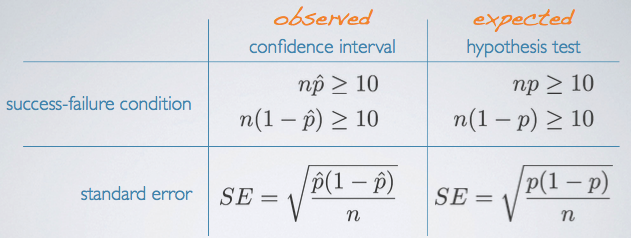
1. Construct a confidence interval by Estimating the Difference Between Two Proportions

* The order (subtracting one from another) does not matter. Both will give the same conclusion regarding the hypothesis.

*See example*

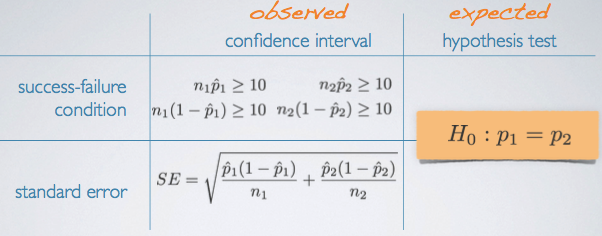
**Hypothesis Tests for Comparing Two Proportions**

**Working with One Proportion: p̂ vs p (Review)**

****

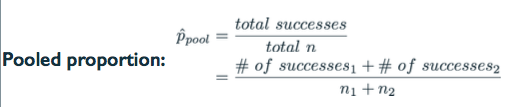
When working with one proportion, we had different formulae for success-failure condition and standard error in both building a confidence interval and conducting a hypothesis test.

**Working with Two Proportions**

****

However, it becomes again different when working with two proportions. We have H0: p1 = p2, but then we wonder what they should be equal to.

**Pooled Proportion**

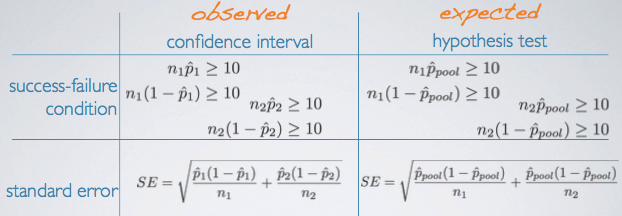
****

* Calculating the expected successes and failures or the expected proportion for the hypothesis test. For difference between two proportions is not as simple. Take a look at this null hypothesis. We simply say in the null hypothesis that the two population proportions should be equal to each other. Or that their difference should be equal to 0. But at no point do we define what this, these should be equal to. So, we don't have a readily available null value. Of the population proportion that we can use for the two groups to calculate expected successes and expected failures. What do we do? We make one up. This is called the pooled proportion. So the idea here is that even though your null hypothesis does not equate the two population proportions to something. Could we actually come up with a best guess for what these could be equal to under the assumption of the null hypothesis.

**When to Use Pooled Proportion**

When conducting a hypothesis test for comparing two proportions

**Working with Two Proportions**



**INSTRUCTIONS**

1. Calculate the pooled proportion
2. Set the hypothesis

* H0 : p1 = p2
* H1 : p1 < or > or != p2

1. Calculate the point estimate p̂1 - p̂2
2. Check conditions
3. Draw sampling distribution, calculate test statistic, shade p-value,
4. Make a decision, and interpret it in context of the research question:

*See example*

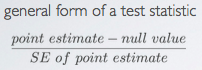
**~~Small Sample Proportion.pdf~~**

**~~Comparing Two Small Sample Proportions.pdf~~**

**Chi-Square GOF test**

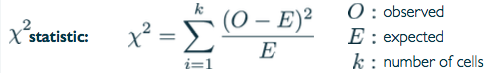
* Used to evaluate the distribution of one categorical variable with more than 2 levels.
* Evaluating by comparing the distribution of that categorical variable to a hypothetical distribution.
* Used to evaluate if the distribution of levels of a single categorical variable follows a hypothesized distribution.

**Anatomy of a Test Statistic**

****

1. Identifying the difference between a point estimate and an expected value if the null hypothesis were true.
2. Standardizing that difference using the standard error of the point estimate.

**Chi-Square Statistic**



When dealing with counts and investigating how far the observed counts are from the expected counts, we use a new statistic called the chi-square(X2) statistic.

* A cell is referred to a level of the categorical variable

**Why Square?**

* Want to get rid of negatives: positive standardized difference
* Not absolute, but square: highly unusual differences between observed and expected will appear even more unusual.

**Conditions for the Chi-square Test**

1. Independence: Sampled observations must be independent

* Random sample / assignment
* If sampling without replacement, n < 10% of population
* Each case only contributes to one cell in the table

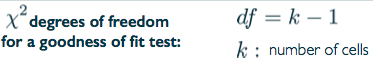
1. Sample size: Each particular scenario (i.e. cell) must have at least 5 expected counts

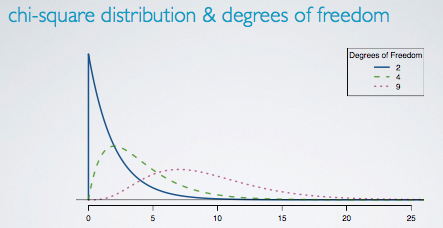
* We want to make sure that each case only contributes to one cell in the table. So we don't want to for example, identify a potential juror as both white and black. That is a possibility, but for the purposes of the chi-square test, we want to make sure each case can only go to one cell in the table. This is also another way of thinking about independence because if our cases showed up in multiple cells in the table, then the observations wouldn't exactly be independent of each other.

**Evaluating the Hypothesis**

* Quantify how different the observed counts are from the expected counts
* Large deviation from what would be expected based on sampling variation(chance) alone provide strong evidence for the alternative hypothesis
* Called a goodness of fit test since we’re evaluating how well the observed data fit the expected distribution

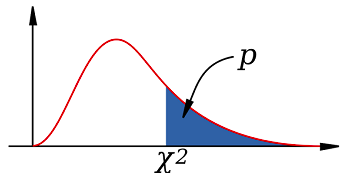
**Degrees of Freedom**





* To determine if the calculated X2 statistic is considered unusually high or not, we need to first describe its distribution.
* Chi-square distribution has only one parameter: degrees of freedom: influences the shape, center and spread.

**P – value**



* P-value for a chi-square test is defined as the tail area above the calculated test statistic
* Because the test statistic is always positive, and a higher test statistic means a higher deviation from the null hypothesis

**INSTRUCTIONS**

1. Set the hypothesis

* H0: Actual and expected distributions follow the same distribution
* H1: Actual and expected distributions do not follow the same distribution

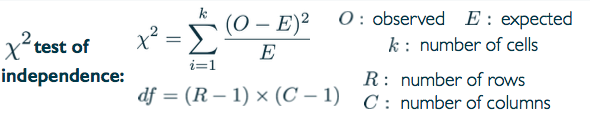
1. Calculate the expected number
2. Check conditions
3. Draw sampling distribution, calculate test statistic, shade p-value
4. Make a decision, and interpret it in context of the research question

*See example*

**The Chi-square Independence Test**

* Evaluating the relationship between two categorical variables, at least 1 with more than 2 levels

**Chi-Square Test of Independence**



**Expected Counts in Two-way Tables**

../../Screen%20Shot%202018-08-07%20at%201.07.56%20AM.png

**Evaluating the hypothesis**

* Quantify how different the observed counts are from the expected counts
* Large deviation from what would be expected based on sampling variation(chance) alone provide strong evidence for the alternative hypothesis
* Called an independence test since we’re evaluating the relationship between two categorical variables

**INSTRUCTIONS**

1. Define the query: Does there appear to be relationship between COL1 and COL2?
2. Set the hypothesis

* H0: nothing going on(independent)
* H1: something going on(dependent)

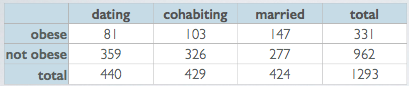
1. Check conditions
2. Calculate the expected counts
3. Calculate test statistic, find p-value
4. Make a decision, and interpret it in context of the research question

**Example)**

**Obesity and Marital Status**

A study reported in the medical journal Obesity in 2009 analyzed data from the National Longitudinal Study of Adolescent Health. Obesity was defined as having a BMI of 30 or more. The research subjects were followed from adolescence to adulthood, and all the people in the sample were categorized in terms of whether they were obese and whether they were dating, cohabiting, or married.

**Study results**

****

1. Define the query

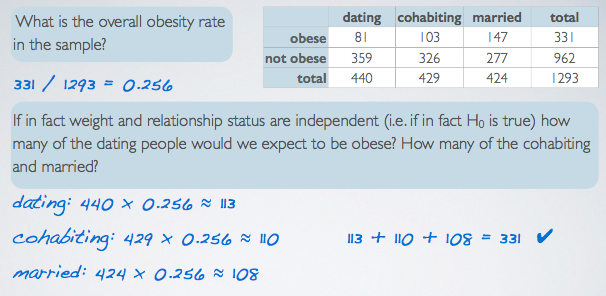
Does there appear to be a relationship between weight and relationship status?

1. Set the hypothesis

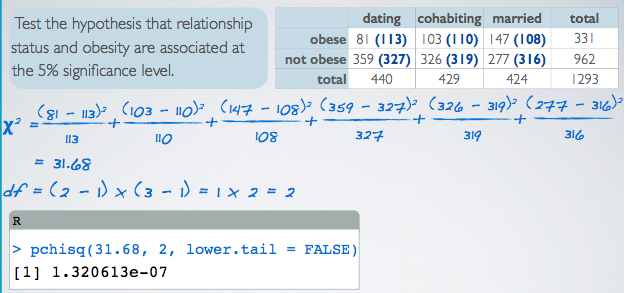
H0: Weight and relationship status are independent. Obesity rates do not vary by relationship status.

H1: Weight and relationship status are dependent. Obesity rates do vary by relationship status.

1. Check conditions
2. Calculate the expected counts

****

1. Calculate test statistic, find p-value

****

1. Make a decision, and interpret it in context of the research question

Can we conclude from these data that living with someone is making some people obese, and that marrying someone is making people even more obese? NO!

**Test yourself:**

1. Suppose 10% of Coursera students smoke. You collect many random samples of 100 Coursera students at a time, and calculate a sample proportion (p̂) for each sample, indicating the pro- portion of students in that sample who smoke. What would you expect the distribution of these p̂s to be? Describe its shape, center, and spread.
2. Suppose you want to construct a confidence interval with a margin of error no more than 4% for the proportion of Coursera students who smoke. How would your calculation of the required sample size change if you don’t know anything about the smoking habits of Coursera students vs. if you have a reliable previous study estimating that about 10% of Coursera students smoke?
3. Suppose a 95% confidence interval for the difference between male and female Coursera students who smoke (male - female) is (-0.08,0.11). Interpret this interval, making sure to incorporate into your interpretation a comparative statement about the two sexes of Coursera students.
4. Does the above interval suggest a significant difference between the true proportions of smokers in the two groups?
5. Suppose you had a sample of 100 male Coursera students where 11 of them smoke, and a sample of 80 female Coursera students where 10 of them smoke. Calculate p^pool.
6. When and why do we use p^pool in calculation of the standard error for the difference between two sample proportions?
7. Explain the different hypothesis tests one could use when assessing the distribution of a categorical variable (e.g. smoking status) with only two levels (e.g. levels: smoker and non-smoker) vs. more than two levels (e.g. levels: heavy smoker, moderate smoker, occasional smoker, non-smoker).
8. Why is the p-value for chi-square tests always “one sided”?
9. What are the null and alternative hypotheses in the chi-square test of independence?
10. Suppose a chi-square test of independence between two categorical variables (one with 5, the other with 3 levels) yields a test statistic of χ2 =14. What’s the conclusion of the hypothesis test at 5% significance level?
11. Suppose you want to estimate the proportion of Coursera students who smoke. You collect a random sample of 100 students, where only 8 of them smoke. Can you use theoretical methods (Z) to construct a confidence interval based on these data? If not, describe how you could calculate a 95% bootstrap confidence interval.