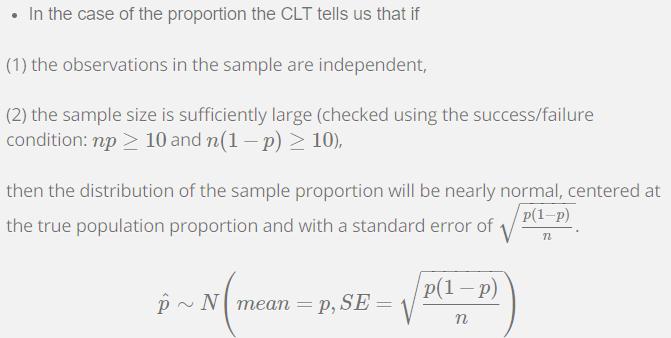
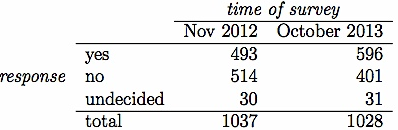
***Week 4 Quiz***

* Suppose you want to construct a CI for a population *proportion*. Which of the following, if it were true, would prevent you from being able to assume the distribution of the sample proportion is nearly normal w/ n = 104?
* ~~These observations are a simple random sample and make up less than 10% of the population.~~
* **Out of these 104 there are only a few failures (7), but relatively many successes (97).**



* In 2013, Edward Snowden leaked details of top-secret NSA spying activities to the media. A poll conducted by USA TODAY/Pew Research Center asked 1,504 people in U.S. whether Snowden’s leaks have helped or harmed the public interest. 53% of respondents answered “helped”. You want to test whether a majority of people in the U.S. believe he helped the public interest. Which of the following is the correct set of hypotheses?
* **H0: ρ=0.5; HA: ρ>0.5**
* In response to complaints from residents about too many (~15%) of the cars passing by the local school speeding, police started closely monitoring traffic. You want to check if police efforts had an effect on prevalence of speeding in this area. 1 day you observe 560 different cars pass by the school + find 70 of them speeding. You calculate a p-value of 0.0976. Assuming the cars are representative of all cars that drive by the school, which of the following is true?
* **If in fact the police’s efforts didn’t have an effect, the probability of getting a random sample of 560 cars where 70 or less or 98 or more cars are speeding is 0.0976.**
* p-value = P(observed or more extreme test statistic | H0 true)
* observed = 98, more extreme test statistics = 70
* When do we use the pooled proportion in calculation of the SE of the difference of 2 proportions?
* **When comparing p1 + p2 using a theoretical approach, + the null H0 = p1 − p2 = 0**
* Rock-paper-scissors is a hand game played by 2+ people where players choose to sign either ‘rock’, ‘paper’, or ‘scissors’ w/ their hands. We’d like to test if players choose between these 3 options randomly, or if certain options are favored above others. What hypothesis test should we conduct to answer this research question?
* **Chi square test of goodness of fit**
* evaluate if distribution of levels of a single categorical variable follow hypothesized distribution
* When doing a hypothesis test on a single proportion (i.e. for 1 categorical variable), we studied how to calculate the p-value for the hypothesis test, beginning w/ generating simulated samples. Which of the following is the best description for how you should generate simulated samples, and why?
* **Generate simulated samples based on the null b/c we need to see how extreme observed data looks if the null were really true.**
* generate simulated samples based on the null + then calculate the # of samples that are at least as extreme as the observed data.
* T/F: In calculation of required sample size for a given margin of error of the CI for a population proportion, we should use p^ = 0.5 if we don’t have any knowledge about the characteristics of the population.
* **True**
* Suppose in a population, 20% of people wear contact lenses. What is the expected shape of the sampling distribution of proportion of contact lens wearers in random samples of 1000 people from this population?
* **nearly normal**
* If the CLT doesn't apply + sample proportion is low (close to 0) the sampling distribution will likely be right=skewed + vice versa
* Here 🡺 random sample + 20%\*1000 = 50, so CLT applies

* T/F: When success-failure condition is not met, we should use a T test to compare 2 proportions.
* **False**
* Use simulation methods when sample size conditions aren't met for inference for categorical variables.
* When sample size isn't sufficiently large + parameter of interest = proportion/difference between 2 proportions, use simulation.
* *t*-distribution = only appropriate to for *means*.
* In hypothesis testing
* 1 categorical variable = generate simulated samples based on the null + then calculate # of samples that are at least as extreme as the observed data.
* 2 categorical variables = use a randomization test.
* When performing a *hypothesis test* on proportions (either where H0: p = p0 or where H0: p1 = p2) you should use the *observed* # of successes + failures when checking conditions.
* **FALSE =** use *observed* # of successes + failures for calculating a CI for a proportion + use n\*p0 + n\*(1−p0) successes + failures (*expected* # based on the null proportion)
* For CI’s, use p^ (*observed* sample proportion) when calculating SE + checking success/failure condition.
* For hypothesis tests, use p0 (null value) when calculating SE + checking success/failure condition.
* You are tasked w/ conducting a hypothesis test evaluating whether a majority or minority of Americans think it was a bad decision to hold the 2014 winter games in Russia. Using data from a 2014 Pew Research poll for 1,003 Americans asked, 44% responded yes. Which of the following is the correct set of hypotheses?
* **H0: p = 0.5; HA: p ≠ 0.5**
* The campaign manager for a congressional candidate claims the candidate has > 50% support from the district’s electorate. A newspaper collects a simple random sample of 500 likely voters in this district + estimates support for this candidate to be 52%. The p-value for the hypothesis test evaluating the campaign manager’s claim is 0.19. Which of the below is correct?
* ~~The success-failure condition is not met, so this p-value is not reliable.~~
* ~~95% of random samples of size 500 will estimate the support for this candidate to be 52%.~~
* **If in fact 50% of likely voters support this candidate, the probability of obtaining a random sample of 500 likely voters where 52% or more support the candidate is 0.19.**
* ~~The data provide convincing evidence for the campaign manager’s claim.~~
* Gallup conducts an annual poll of U.S. residents. Approximately 1k residents across all states + Washington D.C. are asked “Do you believe use of marijuana should be made legal?” Imagine a hypothesis test evaluating whether there is a difference from 2012 to 2013 between proportions of “yes” responses. Using the table, calculate the SE for this hypothesis test (closest answer)



> yes.2012 <- 493

> no.2012 <- 514

> undec.2012 <- 30

> total.2012 <- sum(yes.2012,no.2012,undec.2012)

> prop.yes.2012 <- yes.2012 / total.2012

>

> yes.2013 <- 596

> no.2013 <- 401

> undec.2013 <- 31

> total.2013 <- sum(yes.2013,no.2013,undec.2013)

> prop.yes.2013 <- yes.2013 / total.2013

>

> pooled.prop.point.estimate <- (yes.2012 + yes.2013) / (total.2012 + total.2013)

> > (se.prop.ht <- sqrt(((pooled.prop.point.estimate\*(1-pooled.prop.point.estimate))/total.2012) + ((pooled.prop.point.estimate\*(1-pooled.prop.point.estimate))/total.2013)))

[1] 0.02197318

* **0.022**
* “In statistical inference for proportions, SE is calculated differently for HT’s + CI’s confidence intervals.” Which of the following is the best justification for this statement?
* **Because in hypothesis testing, we assume the null is true, hence we calculate SE using the null value of the parameter. In CI’s, there is no null value, hence we use the sample proportion(s).**
* At the beginning of a semester an anonymous survey was conducted on students in a statistics class. 2 questions on the survey were about gender + whether or not students have equal, more, or less energy in the afternoon compared to the morning. What test should we perform to see if gender and energy level are associated?
* **Chi-square test of independence (**evaluating independence/relationship of 2 categorical variables where at least one has > 2 levels)
* A variety of studies suggest 10% of the world population is left-handed + also claims artists are more likely to be left-handed. To test this claim, take a random sample of 40 art students at a college + find that 6 (15%) left handed. Which of the following is the correct set-up for calculating the p-value for this test?
* **Roll a 10-sided die 40 times, record proportion of times you get a 1 (10% = null value), repeat this many times, + calculate proportion of simulations where sample proportion >= 15%**
* In HT for 1 categorical variable, generate simulated samples based on the null + calculate the # of samples that are at least as extreme as the observed data.

* True or false: The χ2 statistic is always non-negative.
* **True**
* 80% of Americans start the day w/ a cereal breakfast. Based on this, is the following statement T/F? “The sampling distribution of the proportions of Americans who start the day w/ a cereal breakfast in random samples of size 40 is right skewed.”
* **False** 🡪 40\*.8 = 35, 40\*.2 = 5, CLT does not apply + sample proportion is high (close to 1) = *left-skewed*
* At a stop sign, some drivers come to a full stop, some come to a ‘rolling stop’, + some do not stop at all. We’d like to test if there is an association between gender + type of stop + collect data by standing a few feet from a stop sign + taking note of type of stop + gender of driver. What are the hypotheses for testing for an association between gender and type of stop?
* **H0: Gender and type of stop are independent.**

**HA: Gender and type of stop are associated.**

* Does Weight Watchers work? Researchers randomly divided 500 people into 2 equal-sized groups. 1 group spent 6 months on the program, the other received a pamphlet about controlling portion sizes. At the end of the study 35% of subjects in pamphlet group + 55% of subjects in Weight Watchers group had lost at least 10 pounds. To test whether Weight Watchers is more effective for weight loss than pamphlets, a statistician used an index card to represent each subject in the study + wrote whether or not the subject lost at least 10 pounds on the index card. He then shuffled these cards together, + dealt them into 2 equal-sized groups. Which of the following best describes the expected result?
* **If Weight Watchers was effective, the difference between the proportions of cards indicating whether or not the subject lost at least 10 pounds will be more than 20%.**