Intro Hypothesis Testing

1. For each of the research statements below, determine whether it represents a null hypothesis claim or an alternative hypothesis claim.

   (a) The number of hours that grade-school children spend doing homework predicts their future success on standardized tests. *Alternative*

   (b) King cheetahs on average run the same speed as standard spotted cheetahs. *Null*

   (c) For a particular student, the probability of correctly answer a 5-option multiple choice test is larger than 0.2 (i.e. better than guessing) *Alternative*

   (d) The probability of getting in a car accident is the same if using a cell phone then if not using a cell phone. *Null*

2. Write out the null and alternative hypotheses in words and also in statistical notation for each of the following situations. When writing in statistical notation, be sure to define quantities in context.

   (a) New York is known as "the city that never sleeps". A random sample of 25 New Yorkers were asked how much they sleep they get per night. Does these data providing convincing evidence that New Yorkers on average sleep less than 8 hours per night?

   $H_0 : \mu = 8$ *(On average New Yorkers sleep 8 hours a night) versus* $H_A : \mu < 8$ *(On average New Yorkers sleep less than 8 hours a night), where* μ *is the mean hours of sleep New Yorkers receive.*

   (b) A study suggests that 25% of 25 year-olds have gotten married. You believe that this is incorrect and decide to collect your own data to conduct a hypothesis test.

   $H_0 : p = 0.25$ *(True proportion of 25 year-olds who have gotten married is 25%) versus* $H_A : p \neq 0.25$ *(True proportion of 25 year-olds who have gotten married is not 25%)*

3. A Survey USA poll conducted in Seattle, WA in May 2021 reports that of the 650 respondents (adults living in this area), 159 support proposals to defund police departments.

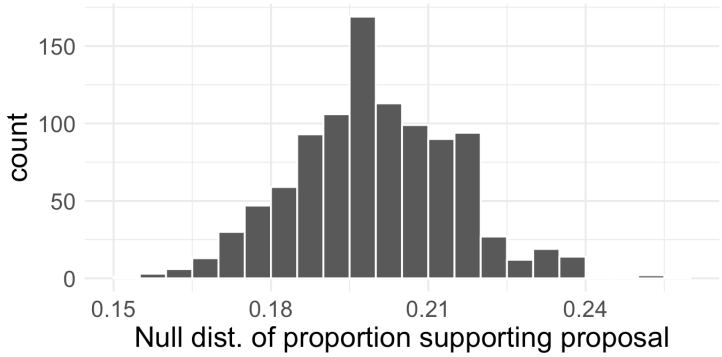
   (a) A journals writing a news story on the poll results wants to use the headline: "More than 1 in 5 adults living in Seattle support proposals to defund police departments". You caution the journalist that they should first conduct a hypothesis test to see if the poll data provide convincing evidence for this claim. Write the hypotheses for this test using proper notation, defining any necessary quantities.

   $H_0 : p = 0.20$ *versus* $H_A : p > 0.20$ *where* p *is the true proportion of Seattle adults who support proposals to defund.*

(b) Describe in words a simulation scheme that would be appropriate for this situation. Also describe how the p-value can be calculated using the simulation results.

Example solution: . Take 100 cards, 20 black cards representing those who support proposals to defund police departments and 80 red cards representing those who do not. Shuffle the cards and draw with replacement (shuffling each time in between draws to get our "infinite population") 650 cards representing the 650 respondents to the poll. After each iteration, calculation $\hat{p}_{sim}$, the proportion of black cards which represents the simulated proportion of adults in favor. The p-value will be the proportion of simulations where $\hat{p}_{sim} \geq 0.245$.

(c) The histogram below shows the distribution of 1000 simulated proportions under H_0. Estimate the p-value using the plot and use it to evaluate your hypotheses (i.e. make a conclusion). Assume a significance level of 0.05.



There is only one simulated proportion that is at least 0.245, therefore the approximate p-value is 0.001. Since $0.001 < 0.05$, reject H_0. The data provide convincing evidence that the proportion of Seattle adults who support proposals to defund police departments is greater than 0.20.

4. A study conducted in 2020 found that the U.S. adjusted divorce rate was 14 per 1000 married women. Joe is suspicious and disagrees with the stated divorce rate. Joe somehow collected data from 323 married or previously-married women, and asked them if they had a divorce in 2020. 55 of the women responded that they indeed had a divorce in 2020.

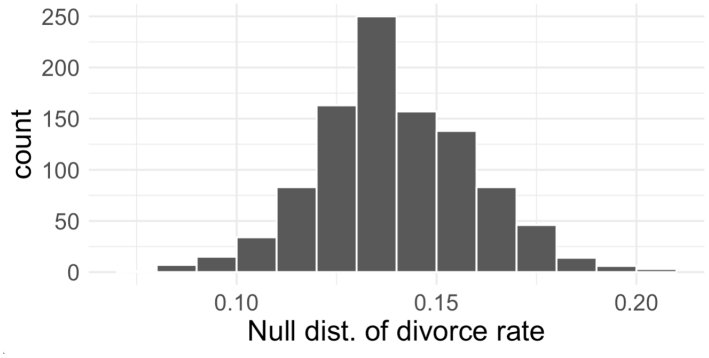
(a) Write out the hypotheses corresponding to this scenario.

$H_0 : p = 0.14$ versus $H_A : p \neq 0.14$ where p is the true divorce rate among married women.

(b) Describe in words a simulation scheme that would be appropriate for this situation. Also describe how the p-value can be calculated using the simulation results.

Similar to previous problem.

(c) The histogram below shows the distribution of 100 simulated proportions under H_0. Estimate the p-value using the plot and use it to evaluate Joe's hypotheses (i.e. make a conclusion). Assume a significance level of 0.05.



The observed proportion was $\hat{p}_{obs} = 55/323 = 0.17$. Since the alternative is two-sided, the p-value is approximately 0.13. Since this is larger than 0.05, fail to reject. The data do not provide convincing evidence that the divorce rate amount married women is different from 0.14.

(d) Joe has some free time and also created a 90% bootstrap confidence interval for the divorce rate.

He obtained the following interval: (0.136, 0.207). Interpret this interval in context.

Joe is 90% confidence that the true divorce rate among married women is between 0.136 and 0.207.

(e) Based on this interval, would it be appropriate for Joe to conclude that the study's reported rate was wrong? Explain your reasoning.

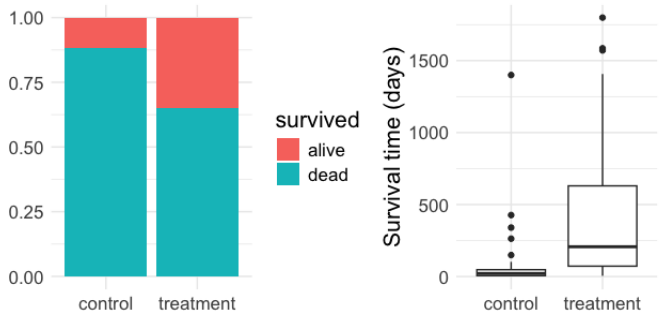
No! 0.14 is included in the interval, so it is a plausible value.

(f) How do your conclusions from (c) and (e) compare?

They agree!

Randomization

1. The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was designated an official heart transplant candidate, meaning that they were gravely ill and would most likely benefit from a new heart. Some patients got a transplant and some did not. The variable `transplant` indicates which group the patients were in: treatment (received transplant) or control (no transplant). The variable `survived` indicates whether the patient was alive at the end of the study or died. Of the 34 patients in the control group, 30 died. Of the 69 people in the treatment group, 45 died.



(a) What do the two plots above suggest about 1) if survival is independent of receiving a transplant and 2) the efficacy of heart transplants? Explain your reasoning.

Proportion of patients who are alive at the end of the study is higher in the treatment group than in the control group. These data suggest that survival is not independent of whether the patient got a transplant.

(b) What proportion of patients in the treatment group and the control group died?

Proportion of patients who in the treatment group who are deceased: $45/69 = 0.652$. *Proportion of patients who in the control group who are deceased:* $30/34 = 0.882$.

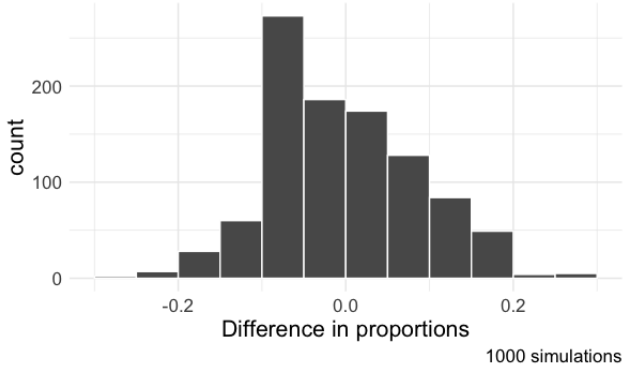
(c) Write out a null and alternative hypothesis for investigating whether there is statistically significant evidence that the treatment is effective.

H_0 : *group (treatment or control) and survival outcomes are independent (i.e. the treatment does not affect death rate.* H_A*: the group and survival outcome are dependent, and specifically, the treatment is effective (decreases death rate).*

(d) The paragraph below describes the set up for a randomization test if we did not have access to software. Fill in the blanks with a number or phrase using your answers to (b) and (c) for guidance:

We write the word *"alive" (or something similar)* on *28* cards representing patients who were alive at the end of the study, and *"dead" (or something similar)* on *75* cards representing the patients who were not. Then we shuffle these cards and split them into two groups: one group of size *69* representing treatment, and one group of size *34* representing *control*. We calculate the difference between the proportion of *"died"* cards in the treatment and control groups (treatment - control) and record this value. We repeat this 1000 times to build a distribution centered at *0*. This is called the *null* distribution. Lastly, we calculate the proportion of simulations where the simulated difference in proportions are *less than or equal to -0.23*. If this proportion is low, we conclude that that it is *unlikely* to have observed our data by chance assuming H_0 *is true*.

(e) What do the simulation results shown below suggest about the effectiveness of heart transplants?



The p-value is approximately 0, which suggest that the transplant program is effective.

(f) Suggest a more informative x-axis label for the plot above.

Include (treatment - control)

2. Understanding cultural differences in tobacco use across different demographic groups can lead to improved health care education and treatment. A recent study disaggregated tobacco use across Asian American ethnic groups, including Asian-Indian (n = 4373), Chinese (n = 4736), and Filipino (n = 4912), in comparison to non-Hispanic Whites (n = 275025). The number of current smokers in each group at the time of study was reported as:

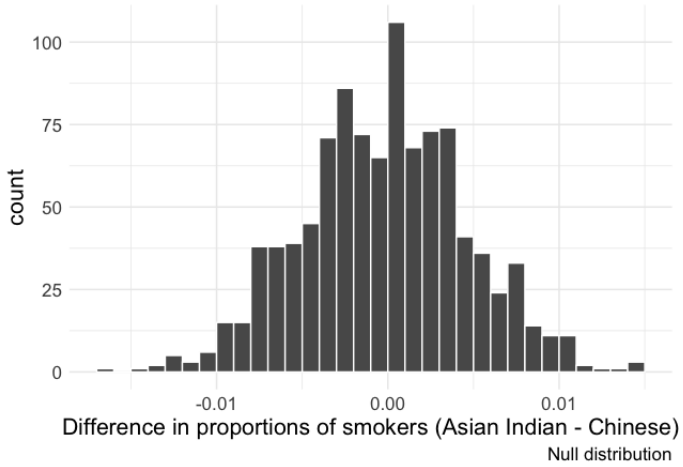
- Asian-Indian: 223
- Chinese: 279
- Filipino: 609
- non-Hispanic Whites: 50880

To determine whether the proportion of Asian-Indian Americans who are current smokers is different from the proportion of Chinese Americans who are smokers, a randomization simulation was performed.

(a) Using both symbols and words, provide the parameter and statistic of interest for this study. Do you know the numerical value of either the parameter or statistic of interest? If so, provide it.

$p_{AI} - p_C$ where p_{AI} is the proportion of Asian-Indian Americans who smoke, and p_C is the proportion of Chinese Americans who smoke. The parameter is the true difference (Asian-Indian - Chinese) in proportions of these groups, and the statistic is the observed difference. We don't know the parameter, by the statistic is $\hat{p}_{AI} - \hat{p}_C = 223/4373 - 279/4736 = -0.008$.

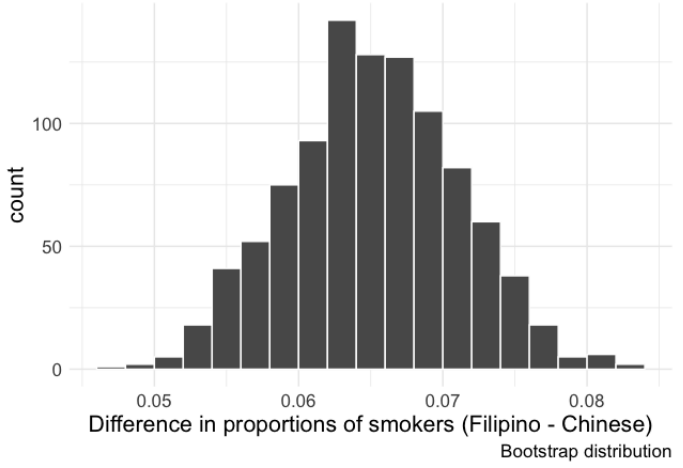
(b) The histogram below provides the simulated null distribution obtained from 1000 repetitions. Estimate the standard error.



(c) Consider the hypothesis test to determine if there is a difference in proportion of Asian-Indian Americans as compared to Chinese Americans who are current smokers. Write out the null and alternative hypotheses, and estimate a p-value using the randomization histogram from (b). If the significance level is $\alpha = 0.05$, what is your decision and conclusion in the context of the problem?

$H_0 : p_{AI} - p_C = 0$ versus $H_A : p_{AI} - p_C \neq 0$. I would estimate about 100 simulations resulted in simulated difference in proportions less than or equal to the observed -0.008 or greater than equal to 0.008. So i would fail to reject H_0. The data do not provide convince evidence that the true difference in proportion of current smokers is different across the two ethnic groups.

(d) Now consider the following bootstrap distribution of the difference in sample proportions of current smokers (Filipino Americans minus Chinese Americans) in 1000 repetitions. Find a 95% bootstrap confidence interval for the true difference in the proportion of current smokers in the population. Interpret the interval in the context of the problem, assuming our sample is representative.



A (symmetric) 95% bootstrap confidence interval is approximately $(0.054, 0.076)$. Thus, we are 95% confident that the proportion of Filipino Americans who smoke is between 0.054 and 0.076 higher than that of Chinese Americans.

HT Basics

1. For each of the statements (a) - (d), indicate if they are true or false interpretation of the following confidence interval. If false, provide or a reason or correction to the misinterpretation.

   "You collect a large sample and calculate a 95% confidence interval for the average number of cans of soda consumed annually per adult to be (440, 520), i.e. on average, adults in the US consume just under two cans of soda per day".

   (a) 95% of adults in the US consume between 440 and 520 cans of soda per year. *False.The interval is for the parameter (a number which describes the population), not for individual observational units.*

   (b) There is a 95% chance that the true population average per adult yearly soda consumption is between 440 and 520 cans. *False. Although unknown, the parameter is either in the interval or it is not (so either with probability zero or probability one).*

   (c) The true population average per adult soda consumption is between 440 and 520 cans, with 95% confidence. *True*

   (d) The average soda consumption of the people who were is sampled is between 440 and 520 cans of soda per year, with 95% confidence. *The sample mean is always inside the interval (it is the center!). We are 100% confident that the sample mean is in the interval*

2. A food safety inspector is called upon to investigate a restaurant with a few customer reports of poor sanitation practices. The food safety inspector uses a hypothesis testing framework to evaluate whether regulations are not being met. If the inspector determines the restaurant is in gross violation, its license to serve food will be revoked.

   (a) Write the hypotheses in words (no population parameters necessary).
   H_0: restaurant means food safety regulations. H_A: restaurant does not meet food safety regulations.

   (b) What is a Type I error in this context?
   The food safety inspector concludes that the restaurant does not meet food safety and sanitation regulations and shuts down the restaurant when the restaurant is actually safe.

   (c) What is a Type II error in this context?
   The food safety inspector concludes that the restaurant meets food safety and sanitation regulations and the restaurant stays open when the restaurant is actually not safe.

   (d) Which error is more problematic for the restaurant owner? For the diners? Why?
   A Type I error may be more problematic for the restaurant owner since his restaurant gets shut down even though it meets the food safety and sanitation regulations. A Type II error may be more problematic for diners since the restaurant deemed safe by the inspector is actually not.

(e) Do you think the diners would prefer a higher or lower significance level α compared to what the restaurant owner prefers? Explain.

A diner would probably prefer strong evidence as any indication of evidence might mean there may be an issue with the restaurant meeting food safety regulations, and diners would rather a restaurant that meet the regulations get shut down than a restaurant that doesn't meet the regulations not get shutdown

3. Consider the following simple random sample $x = (47, 4, 92, 47, 12, 8)$.

   Which of the following sets of values could be a possible bootstrap sample from the observe data above? If a set of values could not be a bootstrap sample, determine why not.

   (a) $(47, 47, 47, 47, 47, 47)$ *Yes*

   (b) $(92, 4, 13, 8, 47, 4)$ *No! 13 was not in original sample.*

   (c) $(4, 8, 12, 12, 47)$ *No; we need a resample of same size as original.*

   (d) $(12, 4, 8, 8, 92, 12)$ *Yes*

   (e) $(8, 47, 12, 12, 8, 4, 92)$ *No; we need a resample of same size as original.*

4. For each of the following statements (a)-(e), indicate if they are a true or false interpretation of the p-value. If false, provide a reason or correction to the misinterpretation.

   "You are wondering if the average amount of cereal in a 10 oz. cereal box is greater than 10 oz. You collect 50 boxes of cereal marketed as 10 oz, conduct simulation-based hypothesis test, and obtain a p-value of 0.23."

   (a) The probability that the average weight of all cereal boxes is 10 oz. is 0.23.

   False. The p-value describes a probability about data, not a probability about a parameter.

   (b) The probability that the average weight of all cereal boxes is something greater than 10 oz. is 0.23.

   False. The p-value describes a probability about data, not a probability about a parameter.

   (c) Because the p-value is 0.23, the average weight of all cereal boxes is 10 oz.

   False. The p-value describes a probability about data, not a probability about a parameter.

   (d) Because the p-value is small, the population average must be just barely about 10 oz.

   False; the p-value doesn't convey information about certainty and "just barely".

   (e) If H_0 is true, the probability of observing another sample with an average as or more extreme as the data is 0.23.

   True.

Normal

1. True or false? Briefly explain why.

   Among applicants to one law school, the average LSAT was about 169, the standard deviation about 9, and the highest score was 178. The distribution of the LSAT scores follows the normal curve.

   False: 68-96-99.7 rule.

2. In a law school class, the entering students averaged 160 on the LSAT. The variance was 64. The histogram of LSAT scores followed the normal curve reasonable well.

   (a) About what percentage of the class scores below 152?
   Since variance is 64, standard deviation is 8. Score of 152 is one standard deviation below mean. Using 68-95-99.7 rule and symmetry, this is approximately $0.5 - 0.34 = 0.16$.

   (b) One student was 0.5 standard deviations above average on the LSAT. About what percentage of the students had lower scores than he did?
   pnorm(164, 160, 8) = pnorm(0.5) = 0.691

1. Weights of 10-year-old girls are known to be Normally distributed with mean of 70 pounds and standard deviation of 13 pounds. Find the probability that a 10-year-old girl weighs between 60 and 85 pounds two ways:

   (a) Optional, but helpful: draw a sketch of the curve and shade in the region of interest.

   (b) Write the probability of interest in $P()$ form. Then write the `R` code necessary to find this probability, and actually execute the code to obtain the probability.
   $P(60 \leq X \leq 85)$ *where* X *the weight of a 10 year old girl. Code: pnorm(85, 70, 13) - pnorm(60, 70, 13) = 0.655.*

   (c) Confirm your solution in (b) by transforming to z-scores first, then using code again to obtain the probability.
   pnorm(15/13) - pnorm(-10/13)

2. Consider the same scenario as in 3. Without using any code than what is provided below, find the 60th percentile for the weight of 10-year-old girls.

   ```
   qnorm(0.6, mean = 0, sd = 1) = 0.2533471
   ```

   Working backwards: $0.253 = \frac{x-70}{13}$, *so* $x = 73.419$ *pounds.*

3. The length of human pregnancies from contraception to birth varies according to a distribution that is approximately normal with mean 266 days and standard deviation 16 days. Without using code, obtain the following:

(a) Between what values do the lengths of the middle 95% of all pregnancies fall?

68-95-99.7 rule: 234 and 298

(b) How short are the shortest 2.5% of all pregnancies? How long do the longest 2.5% last?

Shortest: qnorm(0.025, 266, 16) = 234.64 and Longest: qnorm(0.975, 266, 16) = 297.36 (agrees with our answers above)

CLT

1. A survey found that American families generate an average of 17.2 pounds of glass garbage each year. Assume that the standard deviation is 2.5 pounds.

   Suppose we randomly survey 40 families. Set up a calculation for (and if you have access to R, actually calculate) the probability that the mean of glass garbage of these 40 families is less than 18 pounds.

   Assuming independent garbage generation and no outliers, $\bar{X} \sim N(17.2, 2.5/\sqrt{(}40))$. *So our code is pnorm(18, 17.2, 0.40) = 0.977.*

2. Define what a sampling distribution of the sample proportion is. Describe how the shape, center, and spread of the sampling distribution change as the sample size increases when $p = 0.2$.

   The sampling distribution is the distribution of sample proportions from samples of the same size randomly sampled from the same population. As the same size increases, the shape of the sampling distribution (when $p = 0.2$*) will go from being right-skewed to being more symmetric and resembling the normal distribution. With larger sample sizes, the spread of the sampling distribution gets smaller. Regardless of the sample size, the center of the sampling distribution is equal to the true mean of that population, provided the sampling is independent.*

3. A survey of 1509 high school seniors who took the SAT and who completed an optional web survey shows that 55% of high school seniors are fairly certain that they will participate in a study abroad program in college.

   (a) Is this sample a representative sample from the population of all high school seniors in the US? Explain.

   No. The sample only represents students who took the SAT, and this was also an online survey.

   (b) Suppose the conditions for inference are met, regardless of your answer in (a). Using a mathematical model, construct a 90% confidence interval for the proportion of high school seniors who are fairly certain they will participate in a study abroad program in college. Interpret this interval in context.

   By CLT, $\hat{p} \sim N\left(p, \sqrt{p(1-p)/n}\right)$ *where* p *is the true proportion of high school seniors who want to study abroad.* $\hat{p}_{obs} = 0.55$. *Our critical value is* $z* = qnorm(0.94) \approx 1.64$ *and we need to estimate SE using observed proportion.. So*

our 90% CI for p is $0.55 \pm 1.65\left(\sqrt{\frac{0.55(0.45)}{1509}}\right) = (0.53, 0.57)$. *We are 90% confident the true proportion of high school seniors who took the SAT and are fairly certain that they will participate in a study abroad program in college is about 0.53 and 0.57.*

(c) Based on this interval, would it be appropriate to claim that the majority of high school seniors are fairly certain they will participate in a study abroad program in college?

Yes, since interval lies entirely above 50%.

HT for proportion

1. A recent poll found that 11% of US adults say they have smoked cigarettes in the past week, a historical low. In a random sample of 730 randomly selected students at four-year colleges, it was found that 66 students have smoked cigarettes in the past week. Test that claim that the smoking rate of students at four-year colleges is the same the national US adult average at the 0.05 significance level.

   *$H_0 : p = 0.11$ versus $H_A : p \neq 0.11$ where p is the smoking rate of students at four year colleges. To use CLT, we verify that we have independence via random sampling and the success-failure condition is met: $np_0 = 730(0.11) = 80.3 \geq 10$ and $n(1-p_0) = 649.7 \geq 10$. So the CLT tells us that our test statistic is $z = \frac{66/730 - 0.11}{\sqrt{0.11(0.89)/730}} = -1.69$. The p-value is $2 * $ `pnorm(-1.69)` $= 0.093$. Since the p-value is greater than 0.05, we fail to reject H_0. The data do not suggest that the smoking rate of students at four-year colleges is different from the US adult average rate.*

2. An apple farmer has historically lost an average of 4% of his trees each year. He believes that he has been losing more trees lately.

   a. In a sample of 300 trees, 20 have died. Test the farmer's claim at the 0.01 level.

      $H_0 : p = 0.04$ versus $H_A : p > 0.04$ where p is the loss rate of the farmer's trees. We probably have independence across trees (unless there's a blight going around). For success-failure: $np_0 = 300 * 0.04 = 12 \geq 10$ and $n(1-p_0) = 288 \geq 10$. So CLT tells us that our test statistic is $z = \frac{20/300 - 0.04}{\sqrt{0.04(0.96)/300}} = 2.357$. The p-value is $1 -$ `pnorm(2.357)` $= 0.009$. Since this p-value is below 0.01, we reject H_0. We have convincing evidence that suggests the loss rate of the farmer's trees has increased.

   b. How would the situation change if the farmer's sample size had been 200 instead of 300?

      The success-failure condition would not be satisfied: $np_0 = 200(0.04) = 8$. So we would not feel good about using CLT.

More CLT-based HTs

1. New York is known as "the city that never sleeps". A random sample of 25 New Yorkers were asked how much sleep they get per night. Statistical summaries includes a sample mean hours of 7.73 and standard deviation of 0.77. The point estimate suggests New Yorkers sleep less than 8 hours a night on average. Is the result statistically significant at the 0.05 level? Make a conclusion based on the your decision.

   $H_0 : \mu = 8$ versus $H_A : \mu < 8$ where μ is the average hours of sleep of a New Yorker. We have independence via random sample. We don't have a histogram of the data and the sample size $n = 25$ is small. We will cautiously proceed, but note that we are assuming there are no clear outliers in the data. Conducting a t-test, we have $t = \frac{7.73-8}{0.77/\sqrt{25}} = -1.75$. Our p-value is `pt(-1.75, df = 24)` $= 0.046$. *Since this is less than 0.05, we reject H_0. The data provide some convincing evidence that the true average hours of sleep New Yorkers receive is less than 8 hours,*

2. The population of all verbal GRE scores are known to have a standard deviation of 8.5. A certain graduate department hopes to receive applicants with a verbal GRE scores over 210. This year, the mean verbal GRE scores for the 42 applicants was 212.79. Using a significance level of 0.05, is this new mean significantly greater than the desired mean of 210?

   $H_0 : \mu = 210$ and $H_A : \mu > 210$ where μ is the mean verbal GRE score of this year's applicants to this graduate program. We can assume that the samples are independent as one applicant's score won't tell us about another's. The sample size $n = 42$ is greater than 30, so as long as there are not particularly extreme outliers, we can proceed with CLT. Since the population standard deviation is known, we conduct a z-test: $z = \frac{212.79-210}{8.5/\sqrt{42}} = 2.13$. Our p-value is `1 - pnorm(2.13)` $= 0.0166$. *Since this is less than 0.05, we reject H_0. The data provide convincing evidence that that the mean verbal GRE of this year's applicants is above 210.*