Inference for numerical and categorical data

Computational Mathematics and Statistics Camp
University of Chicago
September 2018

1. An independent random sample is selected from an approximately normal population with an unknown standard deviation. Find the p-value for the given set of hypotheses and T test statistic. Also determine if the null hypothesis would be rejected at $\alpha = 0.05$.

   a. $H_A : \mu > \mu_0$, $n = 11$, $T = 1.91$
   b. $H_A : \mu < \mu_0$, $n = 17$, $T = -3.45$
   c. $H_A : \mu \neq \mu_0$, $n = 7$, $T = 0.83$
   d. $H_A : \mu > \mu_0$, $n = 28$, $T = 2.13$

2. 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 of these data are shown below. Do these data provide strong evidence that New Yorkers sleep less than 8 hours a night on average?

n	$\bar{x}$	s	min	max
25	7.73	0.77	6.17	9.78

   a. Write the hypotheses in symbols and in words.
   b. Check conditions, then calculate the test statistic, T, and the associated degrees of freedom.
   c. Find and interpret the p-value in this context. Drawing a picture may be helpful.
   d. What is the conclusion of the hypothesis test?
   e. If you were to construct a 90% confidence interval that corresponded to this hypothesis test, would you expect 8 hours to be in the interval?

3. For a given confidence level, $t^{\star}_{df}$ is larger than $z^{\star}$. Explain how t^{*}_{df} being slightly larger than z^{*} affects the width of the confidence interval.

4. Researchers interested in lead exposure due to car exhaust sampled the blood of 52 police officers subjected to constant inhalation of automobile exhaust fumes while working traffic enforcement in a primarily urban environment. The blood samples of these officers had an average lead concentration of 124.32 μg/l and a SD of 37.74 μg/l; a previous study of individuals from a nearby suburb, with no history of exposure, found an average blood level concentration of 35 μg/l.

   a. Write down the hypotheses that would be appropriate for testing if the police officers appear to have been exposed to a higher concentration of lead.
   b. Explicitly state and check all conditions necessary for inference on these data.
   c. Test the hypothesis that the downtown police officers have a higher lead exposure than the group in the previous study. Interpret your results in context.
   d. Based on your preceding result, without performing a calculation, would a 99% confidence interval for the average blood concentration level of police officers contain 35 μg/l?

5. Determine if the following statements are true or false. If false, explain.

   a. In a paired analysis we first take the difference of each pair of observations, and then we do inference on these differences.
   b. Two data sets of different sizes cannot be analyzed as paired data.
   c. Consider two data sets that form paired data. Each observation in one data set has a natural correspondence with exactly one observation from the other data set.
   d. Consider two data sets that form paired data. In the analysis, each observation in one data set is subtracted from the average of the other data set's observations.

6. In each of the following scenarios, determine if the data are paired.

a. We would like to know if Intel's stock and Southwest Airlines' stock have similar rates of return. To find out, we take a random sample of 50 days, and record Intel's and Southwest's stock on those same days.
b. We randomly sample 50 items from Target stores and note the price for each. Then we visit Walmart and collect the price for each of those same 50 items.
c. A school board would like to determine whether there is a difference in average SAT scores for students at one high school versus another high school in the district. To check, they take a simple random sample of 100 students from each high school.

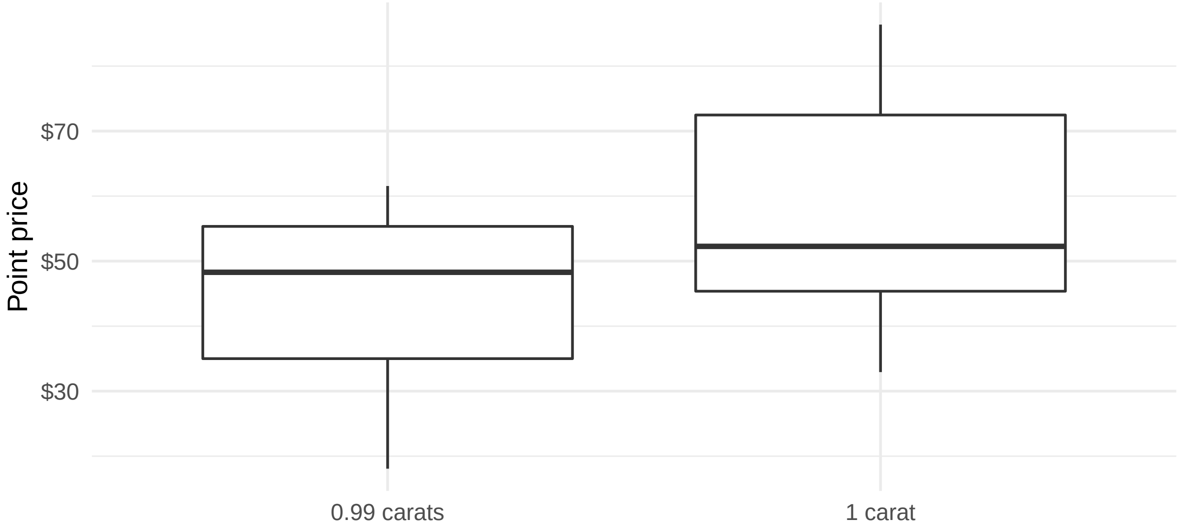
7. We measured the differences between the temperature readings in January 1 of 1968 and 2008 at 51 locations in the continental US. The mean and standard deviation of the reported differences are 1.1 degrees and 4.9 degrees.

   a. Calculate a 90% confidence interval for the average difference between the temperature measurements between 1968 and 2008.
   b. Interpret this interval in context.
   c. Does the confidence interval provide convincing evidence that the temperature was higher in 2008 than in 1968 in the continental US? Explain.

8. Prices of diamonds are determined by what is known as the 4 Cs: cut, clarity, color, and carat weight. The prices of diamonds go up as the carat weight increases, but the increase is not smooth. For example, the difference between the size of a 0.99 carat diamond and a 1 carat diamond is undetectable to the naked human eye, but the price of a 1 carat diamond tends to be much higher than the price of a 0.99 diamond. In this question we use two random samples of diamonds, 0.99 carats and 1 carat, each sample of size 23, and compare the average prices of the diamonds. In order to be able to compare equivalent units, we first divide the price for each diamond by 100 times its weight in carats. That is, for a 0.99 carat diamond, we divide the price by 99. For a 1 carat diamond, we divide the price by 100. The distributions and some sample statistics are shown below.

   Conduct a hypothesis test to evaluate if there is a difference between the average standardized prices of 0.99 and 1 carat diamonds. Make sure to state your hypotheses clearly, check relevant conditions, and interpret your results in context of the data.

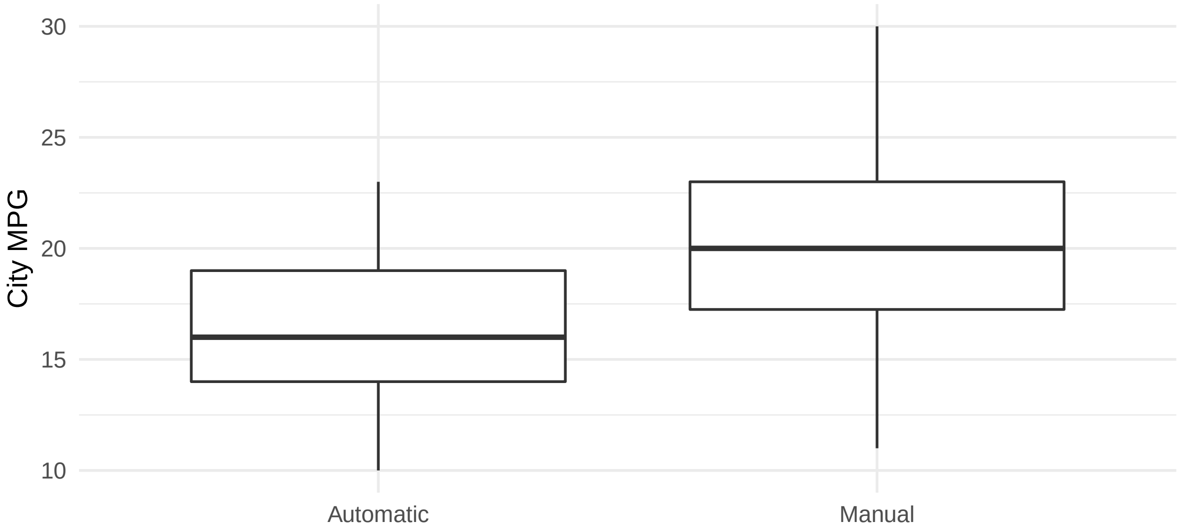
	0.99 carats	1 carat
Mean	$ 44.51	$ 56.81
SD	$ 13.32	$ 16.13
n	23	23



9. Each year the US Environmental Protection Agency (EPA) releases fuel economy data on cars manu-

factured in that year. Below are summary statistics on fuel efficiency (in miles/gallon) from random samples of cars with manual and automatic transmissions manufactured in 2012. Do these data provide strong evidence of a difference between the average fuel efficiency of cars with manual and automatic transmissions in terms of their average city mileage? Assume that conditions for inference are satisfied.

	City MPG	
	Automatic	Manual
Mean	16.12	19.85
SD	3.58	4.51
n	26	26



10. Determine if the following statements are true or false, and explain your reasoning for statements you identify as false.

    a. When comparing means of two samples where $n_1 = 20$ and $n_2 = 40$, we can use the normal model for the difference in means since $n_2 \geq 30$.
    b. As the degrees of freedom increases, the t-distribution approaches normality.
    c. We use a pooled standard error for calculating the standard error of the difference between means when sample sizes of groups are equal to each other.

11. About 77% of young adults think they can achieve the American dream. Determine if the following statements are true or false, and explain your reasoning.

    a. The distribution of sample proportions of young Americans who think they can achieve the American dream in samples of size 20 is left skewed.
    b. The distribution of sample proportions of young Americans who think they can achieve the American dream in random samples of size 40 is approximately normal since $n \geq 30$.
    c. A random sample of 60 young Americans where 85% think they can achieve the American dream would be considered unusual.
    d. A random sample of 120 young Americans where 85% think they can achieve the American dream would be considered unusual.

12. On June 28, 2012 the U.S. Supreme Court upheld the much debated 2010 healthcare law, declaring it constitutional. A Gallup poll released the day after this decision indicates that 46% of 1,012 Americans agree with this decision. At a 95% confidence level, this sample has a 3% margin of error. Based on this information, determine if the following statements are true or false, and explain your reasoning.

    a. We are 95% confident that between 43% and 49% of Americans in this sample support the decision of the U.S. Supreme Court on the 2010 healthcare law.

b. We are 95% confident that between 43% and 49% of Americans support the decision of the U.S. Supreme Court on the 2010 healthcare law.

c. If we considered many random samples of 1,012 Americans, and we calculated the sample proportions of those who support the decision of the U.S. Supreme Court, 95% of those sample proportions will be between 43% and 49%.

d. The margin of error at a 90% confidence level would be higher than 3%.

13. The 2010 General Social Survey asked 1,259 US residents: "Do you think the use of marijuana should be made legal, or not?" 48% of the respondents said it should be made legal.

    a. Is 48% a sample statistic or a population parameter? Explain.

    b. Construct a 95% confidence interval for the proportion of US residents who think marijuana should be made legal, and interpret it in the context of the data.

    c. A critic points out that this 95% confidence interval is only accurate if the statistic follows a normal distribution, or if the normal model is a good approximation. Is this true for these data? Explain.

    d. A news piece on this survey's findings states, "Majority of Americans think marijuana should be legalized." Based on your confidence interval, is this news piece's statement justified?

14. The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was officially designated a heart transplant candidate, meaning that he was gravely ill and might benefit from a new heart. Patients were randomly assigned into treatment and control groups. Patients in the treatment group received a transplant, and those in the control group did not. The table below displays how many patients survived and died in each group.

	control	treatment
alive	4	24
dead	30	45

    A hypothesis test would reject the conclusion that the survival rate is the same in each group, and so we might like to calculate a confidence interval. Explain why we cannot construct such an interval using the normal approximation. What might go wrong if we constructed the confidence interval despite this problem?

15. According to a report on sleep deprivation by the Centers for Disease Control and Prevention, the proportion of California residents who reported insufficient rest or sleep during each of the preceding 30 days is 8.0%, while this proportion is 8.8% for Oregon residents. These data are based on simple random samples of 11,545 California and 4,691 Oregon residents. Calculate a 95% confidence interval for the difference between the proportions of Californians and Oregonians who are sleep deprived and interpret it in context of the data.

16. A professor using an open source introductory statistics book predicts that 60% of the students will purchase a hard copy of the book, 25% will print it out from the web, and 15% will read it online. At the end of the semester he asks his students to complete a survey where they indicate what format of the book they used. Of the 126 students, 71 said they bought a hard copy of the book, 30 said they printed it out from the web, and 25 said they read it online.

    a. State the hypotheses for testing if the professor's predictions were inaccurate.

    b. How many students did the professor expect to buy the book, print the book, and read the book exclusively online?

    c. This is an appropriate setting for a chi-square test. List the conditions required for a test and verify they are satisfied.

    d. Calculate the chi-squared statistic, the degrees of freedom associated with it, and the p-value.

    e. Based on the p-value calculated in part (d), what is the conclusion of the hypothesis test? Interpret your conclusion in this context.