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. $H_A : \mu > \mu_0$, $n = 11$, $T = 1.91$

   $$H_A : \mu > \mu_o \quad n = 11 \quad T = 1.91 \quad df = 11 - 1 = 10 \quad 0.025 < \text{p-value} < 0.05 \quad \text{Reject } H_0$$

   b. $H_A : \mu < \mu_0$, $n = 17$, $T = -3.45$

   $$H_A : \mu < \mu_o \quad n = 17 \quad T = -3.45 \quad df = 17 - 1 = 16 \quad \text{p-value} < 0.005 \quad \text{Reject } H_0$$

   c. $H_A : \mu \neq \mu_0$, $n = 7$, $T = 0.83$

   $$H_A : \mu \neq \mu_o \quad n = 7 \quad T = 0.83 \quad df = 7 - 1 = 6 \quad \text{p-value} > 0.20 \quad \text{Fail to reject } H_0$$

   d. $H_A : \mu > \mu_0$, $n = 28$, $T = 2.13$

   $$H_A : \mu > \mu_o \quad n = 28 \quad T = 2.13 \quad df = 28 - 1 = 27 \quad 0.01 < \text{p-value} < 0.025 \quad \text{Reject } H_0$$

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.

   $H_0 : \mu = 8$ (New Yorkers sleep 8 hrs per night on average.) $H_A : \mu < 8$ (New Yorkers sleep less than 8 hrs per night on average.)

   b. Check conditions, then calculate the test statistic, T, and the associated degrees of freedom.

   Before calculating the test statistic we should check that the conditions are satisfied.

   1. Independence: The sample is random and 25 is less than 10% of all New Yorkers, so the observations are independent.
   2. Normality: All observations are within three standard deviations of the mean. While this is encouraging, it would be useful to see the raw data. However, for now we will proceed while acknowledging that we are assuming the skew is perhaps moderate or less (moderate skew would be acceptable for this sample size).

   The test statistic and degrees of freedom can be calculated as follows:

$$T = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{7.73 - 8}{\frac{0.77}{\sqrt{25}}} = \frac{-0.27}{0.154} = -1.75$$

$$df = 25 - 1 = 24$$

c. Find and interpret the p-value in this context. Drawing a picture may be helpful.

p-value $= P(T_{24} < -1.75) \rightarrow 0.025 <$ p-value < 0.05. If in fact the true population mean of the amount New Yorkers sleep per night was 8 hours, the probability of getting a random sample of 25 New Yorkers where the average amount of sleep is 7.73 hrs per night or less is between 0.025 and 0.05.

d. What is the conclusion of the hypothesis test?

Since p-value < 0.05, reject H_0. The data provide convincing evidence that New Yorkers sleep less than 8 hours per night on average.

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?

No, the hypothesis test suggests that the average amount of sleep New Yorkers get is significantly lower than 8 hours per night, therefore we wouldn't expect 8 hours to be in the interval. Note that the confidence level corresponding to this test is 90%, since the test is one-sided and uses a significance level of 5%.

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.

With a larger critical value, the confidence interval ends up being wider. This makes intuitive sense as when we have a small sample size and the population standard deviation is unknown, we should have a wider interval than if we knew the population standard deviation, or if we had a large enough sample size.

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.

$$H_0 : \mu = 35, \quad H_A : \mu > 35$$

b. Explicitly state and check all conditions necessary for inference on these data.

1. Independence: 52 police officers are less than 10% of all police officers and if we can assume that these 52 officers represent a random sample, we can assume that the blood lead concentration of one officer in the sample is representative of another.
2. Normality: We don't have a plot of the distribution that we can use to check this condition, however given that the sample average is more than three times as large as the standard deviation, it is conceivable that the distribution is approximately normal (the 68-95-99.7% rule could apply here). There is also no reason to suspect extreme skew in the distribution of blood lead concentration.

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.

The test statistic and the p-value can be calculated as follows:

$$T = \frac{124.32 - 35}{\frac{37.74}{\sqrt{52}}} \approx 17.07$$
$$df = 52 - 1 = 51$$

$$\text{p-value} = P(T_{51} > 17.07) < 0.005$$

The hypothesis test yields a very small p-value, so we reject H_0. This indicates that the data provide very convincing evidence that the police officers have been exposed to a higher concentration of lead than individuals living in a suburban area.

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?

   Given that the one-sided p-value is less than 0.005, a two-sided hypothesis test at $\alpha = 0.01$ would also be rejected with these data. Since such a test is equivalent to a 99% confidence interval, we would not expect this interval to include the null value of $35\mu g/l$.

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.

      True

   b. Two data sets of different sizes cannot be analyzed as paired data.

      True

   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.

      True

   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.

      False. We find the difference of each pair of observations, and then we do inference on these differences.

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.

      Paired, on the same day the stock prices may be dependent on external factors that affect the price of both stocks.

   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.

      Paired, the prices are for the same 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.

      Not paired, these are two independent random samples, individual students are not matched.

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.

      A 90% confidence interval can be calculated as follows:

$$
\begin{aligned}
\bar{x}_{diff} \pm t^*_{df} \frac{s_{diff}}{\sqrt{n}} &= 1.1 \pm 1.68 \times \frac{4.9}{\sqrt{51}} \\
&= 1.1 \pm 1.68 \times 0.6861 \\
&= 1.1 \pm 1.15 \\
&= (-0.05, 2.25)
\end{aligned}
$$

b. Interpret this interval in context.

We are 90% confident that the average daily high in January 1, 2008 in the continental US was 0.05 degrees lower to 2.25 degrees higher than the average daily high in January 1, 1968.

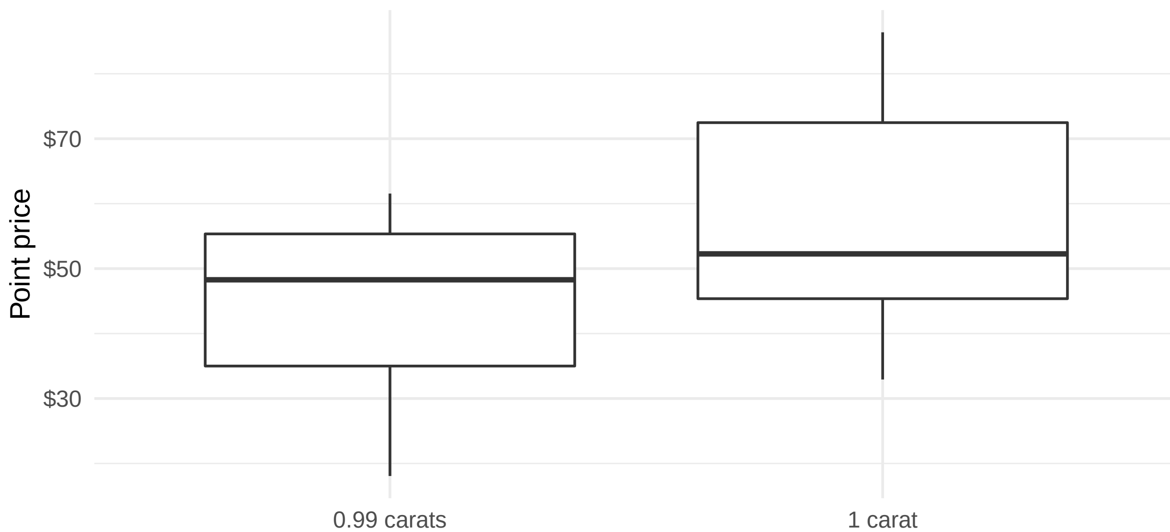
c. Does the confidence interval provide convincing evidence that the temperature was higher in 2008 than in 1968 in the continental US? Explain.

No, since 0 is included in the interval.

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



The hypotheses are: $H_0 : \mu_{0.99} = \mu_1$ and $H_A : \mu_{0.99} \neq \mu_1$. The conditions that need to be satisfied for the sampling distribution of $(\bar{x}_{0.99} - \bar{x}_1)$ to be nearly normal and the estimate of the standard error to be sufficiently accurate are:

1. Independence: Both samples are random and represent less than 10% of their respective populations. Also, we have no reason to think that the 0.99 carats are not independent of the 1 carat diamonds since they are both sampled randomly.
2. Normality: The distributions are not extremely skewed, hence we can assume that the distribution of the average differences will be nearly normal as well.

The test statistic and the p-value can be calculated as follows:

$$T = \frac{(\bar{x}_{0.99} - \bar{x}_1) - (\mu_{0.99} - \mu_1)}{\sqrt{\frac{s_{0.99}^2}{n_{0.99}} + \frac{s_1^2}{n_1}}}$$

$$= \frac{(44.51 - 56.81) - 0}{\sqrt{\frac{13.32^2}{23} + \frac{16.13^2}{23}}} = -\frac{12.3}{4.36} = -2.82$$

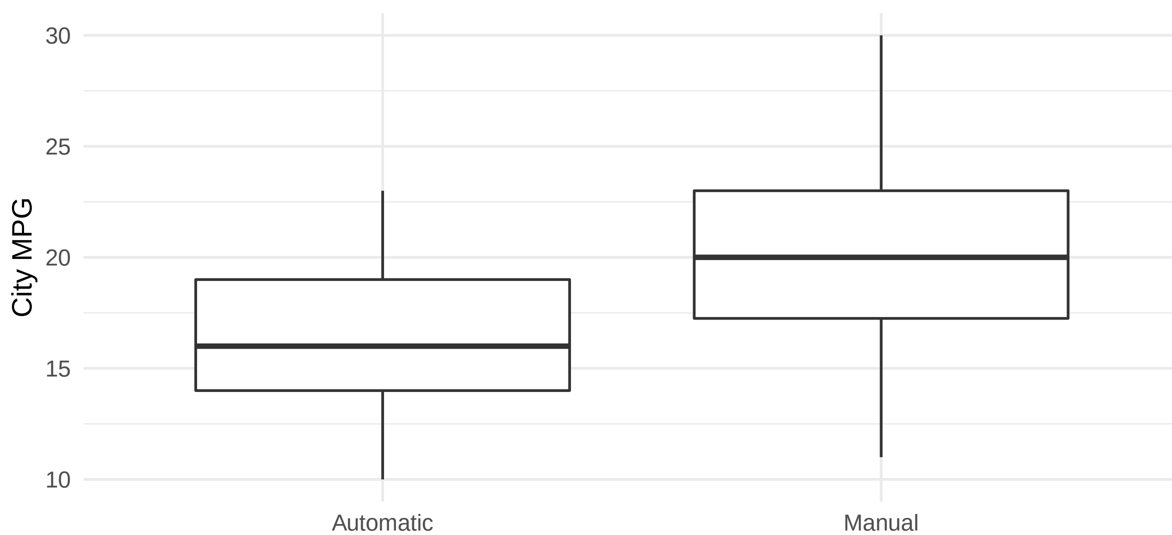
$$df = 23 - 1 = 22$$

$$\text{p-value} = P(|T_{22}| > 2.82) = 0.01$$

Since p-value < 0.05, reject H_0. The data provide convincing evidence that the average standardized price of 0.99 carats and 1 carat diamonds are different.

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



The hypotheses are as follows:

$$H_0 : \mu_{A,c} - \mu_{M,c}$$

$$H_A : \mu_{A,c} \neq \mu_{M,c}$$

We are told to assume that conditions for inference are satisfied. Then, the test statistic and the p-value can be calculated as follows:

$$T = \frac{(\bar{x}_{A,c} - \bar{x}_{M,c}) - (\mu_{A,c} - \mu_{M,c})}{\sqrt{\frac{s^2_{A,c}}{n_{A,c}} + \frac{s^2_{M,c}}{n_{M,c}}}} = \frac{(16.12 - 19.85) - 0}{\sqrt{\frac{3.58^2}{26} + \frac{4.51^2}{26}}} = \frac{-3.73}{1.13} = -3.3$$

$$df = \min(n_{M,c} - 1, n_{A,c} - 1) = \min(26 - 1, 26 - 1) = 25$$

$$\text{p-value} = P(|T_{25}| > 3.3) < 0.01$$

Since p-value < 0.05, reject H_0. The data provide strong evidence that there is a difference in the average city mileage between cars with automatic and manual transmissions.

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 $n_2 \geq 30$.

       False, in order to be able to use a Z test both sample sizes need to be above 30.

    b. As the degrees of freedom increases, the t-distribution approaches normality.

       True.

    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.

       False, we use the pooled standard deviation when the variability in groups is constant.

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.

       True. The success-failure condition is not satisfied:

       $$np = 20 \times 0.77 = 15.4 \text{ and } n(1 - p) = 20 \times 0.23 = 4.6$$

       Therefore we know that the distribution of $\hat{p}$ is not approximately normal. In most samples we would expect $\hat{p}$ to be close to 0.77, the true population proportion. While $\hat{p}$ can be as low as 0 (though we would expect this to happen very rarely), it can only go as high as 1. Therefore, since 0.77 is closer to 1, the distribution would probably take on a left skewed shape. Plotting the sampling distribution would confirm this suspicion.

    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 $n \geq 30$.

       False. Unlike with means, for the sampling distribution of proportions to be approximately normal, we need to have at least 10 successes and 10 failures in our sample. We do not use $n \geq 30$ as a condition to check for the normality of the distribution of $\hat{p}$.

    c. A random sample of 60 young Americans where 85% think they can achieve the American dream would be considered unusual.

       False. Standard error of $\hat{p}$ in samples with $n = 60$ can be calculated as:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.77 \times 0.23}{60}} = 0.0543$$

A $\hat{p}$ of 0.85 is only $Z = \frac{0.85-0.77}{0.0543} = 1.47$ standard errors away from the mean, which would not be considered unusual.

d. A random sample of 120 young Americans where 85% think they can achieve the American dream would be considered unusual.

True. Standard error of $\hat{p}$ in samples with $n = 120$ can be calcuated as:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.77 \times 0.23}{120}} = 0.0384$$

A $\hat{p}$ of 0.85 is $Z = \frac{0.85-0.77}{0.046} = 2.08$ standard errors away from the mean, which 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.

   False. A confidence interval is constructed to estimate the population proportion, not the sample proportion.

   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.

   True. This is the correct interpretation of the confidence interval, which can be calculated as $0.46 \pm 0.03 = (0.43, 0.49)$.

   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%.

   False. The confidence interval does not tell us what we might expect to see in another random sample.

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

   False. As the confidence level decreases, the margin of error decreases as well.

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.

   48% is a sample statistic, it's the observed sample proportion.

   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.

   A 95% confidence interval can be calculated as follows:

$$
\begin{aligned}
\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= 0.48 \pm 1.96 \sqrt{\frac{0.48(1-0.48)}{1259}} \\
&= 0.48 \pm 1.96 \times 0.014 \\
&= 0.48 \pm 0.0274 \\
&= (0.4526, 0.5074)
\end{aligned}
$$

We are 95% confident that approximately 45% to 51% of Americans think marijuana should be legalized.

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.

   1. Independence: The sample is random, and comprises less than 10% of the American population, therefore we can assume that the individuals in this sample are independent of each other
   2. Success-failure: The number of successes (people who said marijuana should be legalized: $1259 \times 0.48 = 604.32$) and failures (people who said it shouldn't be: $1259 \times 0.52 = 654.68$) are both greater than 10, therefore the success-failure condition is met as well.

   Therefore the distribution of the sample proportion is expected to be approximately normal.

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?

   No, the interval contains 50%, suggesting that the true population proportion could be 50%, or even lower. Using this interval we wouldn't reject a null hypothesis where $p = 0.50$.

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?

Before we can calculate a confidence interval, we must first check that the conditions are met.

1. Independence: If patients are randomly assigned into the two groups, whether or not one patient in the treatment group survives is independent of another, and whether or not one patient in the control group survives is independent of another as well.
2. Success-failure: There are only 4 deaths in the control group.

Since the success-failure condition is not met, $(\hat{p}_C - \hat{p}_T)$ is not expected to be approximately normal and therefore cannot calculate a confidence interval for the difference between the proportion of patients who survived in the treatment and control groups using large sample techniques and a critical Z score.

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.

a. 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.

Before calculating the confidence interval we should check that the conditions are satisfied.

1. Independence: Both samples are random, and $11,545 < 10\%$ of all Californians and $4,691 < 10\%$ of all Oregonians, therefore how much one Californian sleeps is independent of how much another Californian sleeps and how much one Oregonian sleeps is independent of how much another Oregonian sleeps. In addition, the two samples are independent of each other.
2. Success-failure:

$$11,545\ times 0.08 = 923.6 > 10 \quad 11,545 \times 0.92 = 10621.4 > 10$$

$$4,691 \times 0.088 = 412.8 > 10 \quad 4,691 \times 0.912 = 4278.2 > 10$$

Since the observations are independent and the success-failure condition is met, $\hat{p}_{CA} - \hat{p}_{OR}$ is expected to be approximately normal. A 95% confidence interval for the difference between the population proportions can be calculated as follows:

$$\begin{aligned}(\hat{p}_{CA} - \hat{p}_{OR}) \pm z^* \sqrt{\frac{\hat{p}_{CA}(1-\hat{p}_{CA})}{n_{CA}} + \frac{\hat{p}_{OR}(1-\hat{p}_{OR})}{n_{OR}}} &= (0.08 - 0.088) \pm 1.96 \sqrt{\frac{0.08 \times 0.92}{11545} + \frac{0.088 \times 0.912}{4691}} \\ &= -0.008 \pm 0.009 \\ &= (-0.017, 0.001)\end{aligned}$$

We are 95% confident that the difference between the proportions of Californians and Oregonians who are sleep deprived is between -1.7% and 0.1%. In other words, we are 95% confident that 1.7% less to 0.1% more Californians than Oregonians are sleep deprived.

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.

The hypotheses are as follows:

- H_0: The distribution of the format of the book used by the students follows the professor's predictions.
- H_A: The distribution of the format of the book used by the students does not follow the professor's predictions.

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

- $E_{hardcopy} = 126 \times 0.60 = 75.6$
- $E_{print} = 126 \times 0.25 = 31.5$
- $E_{online} = 126 \times 0.15 = 18.9$

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

1. Independence: The sample is not random. However, if the professor has reason to believe that the proportions are stable from one term to the next and students are not affecting each other's study habits, independence is probably reasonable.
2. Sample size: All expected counts are at least 5.

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

The χ^2 statistic, the degrees of freedom associated with it, and the p-value can be calculated as follows:

$$\chi^2 = \sum \frac{O - E)^2}{E} = \frac{(71 - 75.6)^2}{75.6} + \frac{(30 - 31.5)^2}{31.5} + \frac{(25 - 18.9)^2}{18.9} = 2.32$$

$$df = 2$$

$$\text{p-value} > 0.3$$

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

Since the p-value is large, we fail to reject H_0. The data do not provide strong evidence indicating the professor's predictions were statistically inaccurate.