Test 4 Review

Statistics for Computer Science – 201-H02-HR

1. You construct a survey to discover what proportion of Heritage College students would be in favour of a Fall reading week. Of the 100 random respondents, 43% were in favour of the proposal.

   (a) Use a hypothesis test to determine if exactly 50% of Heritage College students do favour a Fall reading week. Use that a majority of Heritage College students do not favour the proposal as your alternative hypothesis. Use a significance level of 5% to make your decision.

   (b) If you could re-start the entire survey so that you could guarantee that your confidence interval for the proportion would have a 90% confidence and a margin of error of 5 percentage points, how many random students would you have to survey? Use a conservative bound on the standard deviation for your work.

2. Many people have trouble setting up all the features of their smartphones, so a company has developed what it hopes will be easier instructions. The goal is to have at least 96% of customers succeed. The company tests the new system on 300 people, of whom 282 were successful. Is this strong evidence that the new system fails to meet the company's goal?

   (a) Write appropriate hypotheses.

   (b) Check the necessary assumptions.

   (c) Perform the mechanics of the test. What is the P-value?

   (d) Explain carefully what the P-value means in this context.

   (e) What is you conclusion?

3. You developed an app for iOS and for Android that has over 30,000 users. In a survey of randomly-selected users, 67% of 113 iOS users and 61% of 98 Android users reported satisfaction with the app.

   (a) Construct a confidence interval for the difference of proportions with 95% confidence. Be sure to verify all relevant conditions.

   (b) Test the hypothesis that there really is no true difference in these proportions. Use a significance level of 5%.

4. A study published in the *Archives of General Psychiatry* examined the impact of depression on a patient's ability to survive cardiac disease. Researchers identified 450 people with cardiac disease, evaluated them for depression, and followed the group for four years. Of the 361 patients with no depression, 67 died. Of the 89 patients with minor or major depression, 26 died. Among people who suffer from cardiac disease, are depressed patients more likely to die than nondepressed ones?

   (a) Write appropriate hypotheses.

   (b) Are the assumptions and conditions necessary for inference satisfied?

   (c) Test the hypothesis and state your conclusion.

   (d) Explain what your P-value means in this context.

   (e) Create a 95% confidence interval for the difference in survival rates.

   (f) Interpret your interval in this context.

   (g) Carefully explain what "95% confidence" means.

5. Hoping to lure more shoppers downtown, a city builds a new public parking garage in the central business district. The city plans to pay for the structure through parking fees. During a two-month period (44 weekdays), daily fees collected averaged $126, with a standard deviation of $15.

   (a) What assumptions must you make in order to use these statistics for inference?

   (b) Write a 90% confidence interval for the mean daily income this parking garage will generate.

   (c) Explain in context what this confidence interval means.

   (d) Explain what "90% confidence" means in this context.

   (e) The consultant who advised the city on this project predicted that parking revenues would average $130 per day. Based on your confidence interval, do you think the consultant was correct? Why?

   (f) Someone suggests that the city use its data to create a 95% confidence interval instead of the 90% interval first created. How would this interval be better for the city? (You need not actually create the new interval.)

   (g) How would the 95% interval be worse for the planners?

   (h) How many day's worth of data must they collect to have 95% confidence of estimating the true mean to within 3$

6. You are testing the running time a computer program that you wrote. You run your computer program on 16 instances of randomly-generated input data and measure the running time of each. Your histogram of the results is roughly unimodal and symmetric with a sample mean of 47 milliseconds and a standard deviation of 24 milliseconds.

   (a) What conditions must you check if you want to construct a confidence interval or hypothesis test for the distribution of sample means?

   (b) Construct a confidence interval for your sample mean with a confidence of 95%.

   (c) Test the hypothesis that the true mean is 60 milliseconds. Consider the true mean being less than 60 milliseconds as your alternative hypothesis and use a significance level of 5%.

7. In 1960, census results indicated that the age at which Canadian women first married had a mean of 22.6 years. It is widely suspected that young people today are waiting longer to get married. We want to find out if the mean age of first marriage has increased during the past 40 years.

   (a) Write appropriate hypotheses.

   (b) We plan to test our hypotheses by selecting a random sample of 40 women who married for the first time last year. Do you think the necessary assumptions for inference are satisfied? Explain.

   (c) Describe the approximate sampling distribution model for the mean age in such samples.

   (d) The women in our sample married at an average age of 27.2 years, with a standard deviation of 5.3 years. What is the P-value for this result?

   (e) Explain (in context) what this P-value means.

   (f) What is your conclusion?

Selected Answers:

1. (a) With $H_0 : p = 0.50$ and $H_A : p < 0.50$, $P(X < -1.4) = 0.0808$ so we fail to reject the null hypothesis at a significance level of 5%.

   (b) See notes for proof that we would want $n \geq 272.25$ (since we are counting students, $n \geq 273$). As a side-note, notice that this sample size would disobey the 10% Condition.

2. (a)
   - H_0 : The percentage of successful customers with the new system is 96%. ($p = 0.96$)
   - H_A : H_0 : The percentage of successful customers with the new system is less than 96%. ($p < 0.96$)

   (b)
   - **Randomization Condition:** This sample is not random, so hopefully the customers you check with are representative of all the company's customers.
   - **10% Condition:** The 300 customers sampled may be considered less than 10% of all possible customers.
   - **Independence Assumption:** (Since Randomization wasn't met we must make a case for this assumption). There is no reason to think that the successful setting up of a smartphone for one customer will affect the probability that another customer successfully setting up their smartphone.
   - **Success/Failure Condition:** $np_0 = (300)(0.96) = 288$ and $nq_0 = (300)(0.04) = 12$ are both greater than 10, so the sample is large enough.

   (c) The sample of customers may not be representative of all customers, so we will proceed cautiously. A Normal model can be used to model the sampling distribution of the proportion, with $\mu_{\hat{p}} = p_0 = 0.96$ and $\sigma(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}} = \sqrt{\frac{(0.96)(0.04)}{300}} \approx 0.0113$.
   We can perform a one-proportion z-test. The observed proportion of successful customers is $\hat{p} = \frac{282}{300} = 0.94$. Thus $z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 n}} = \frac{0.94 - 0.96}{0.0113} \approx -1.77$
   The P-value is 0.0384 , since $P(z < -1.77) = .0384$

   (d) If the new system has the same rate as the old one, there is more than a 3.8% chance of seeing results as good or worse than our sample by natural sampling.

(e) With a P-value of 0.0384, we reject the null hypothesis. There is some evidence to suggest that the success rate of the customers with the new system is below the desired 96%.

3. (a) Verify independence, randomness, the 10% Condition, the 10 Successes and 10 Failures Condition, and the independence between groups condition. We estimate with 95% confidence that the difference of proportions (proportion of iOS users minus proportion of Android users) is between -7% and 19%.

   (b) $\hat{p}_{pooled} = 0.6421$ and $SE_{pooled}(\hat{p}_1 - \hat{p}_2) = 0.06617$ so with a two-tailed hypothesis test, $P(Z < -0.91) + P(Z > 0.91) = 0.3628$ so we definitely do not reject the null hypothesis at a significance level of 5%.

4. (a)
   - H_0: The proportion of cardiac patients without depression who died within the four years is the same as the proportion of cardiac patients with depression who died during the same time period. ($p_{None} = p_{Dep}$ or $p_{None} - p_{Dep} = 0$)
   - H_A: The proportion of cardiac patients without depression who died within the four years is less than the proportion of cardiac patients with depression who died during the same time period. ($p_{None} < p_{Dep}$ or $p_{None} - p_{Dep} < 0$)

   (b)
   - **Randomization Condition:** Assume that the cardiac patients followed by the study are representative of all cardiac patients.
   - **10% Condition:** 361 and 89 are both less than 10% of all patients.
   - Independent Samples Condition: The groups are not associated.
   - **Success/Failure Condition:** $n_{None}\hat{p_{None}} = 67$, $n_{None}\hat{q_{None}} = 294$, $n_{Dep}\hat{p_{Dep}} = 26$, and $n_{Dep}\hat{q_{Dep}} = 63$ are all greater than 10, sp the samples are both large enough.

   Since the conditions have been satisfied, we will model the sampling distribution of the difference in proportion with a Normal model with mean 0 and standard deviation estimated by:

$$SE_{pooled}(\hat{p}_{None} - \hat{p}_{Dep}) = \sqrt{\frac{\hat{p}_{pooled}\hat{q}_{pooled}}{n_{None}} + \frac{\hat{p}_{pooled}\hat{q}_{pooled}}{n_{Dep}}}$$

$$= \sqrt{\frac{\left(\frac{93}{450}\right)\left(\frac{357}{450}\right)}{361} + \frac{\left(\frac{93}{450}\right)\left(\frac{357}{450}\right)}{89}} \approx 0.0479.$$

(c) The observed difference between the proportions is $0.1856 - 0.2921 = -.1065$. Note that

$$z = \frac{-0.1065 - 0}{0.0479} \approx -2.22 \text{ and } P(z < -2.22) + 0.0131$$

Since the P-value = 0.0131 is low, we reject the null hypothesis. There is strong evidence to suggest that the proportion of non-depressed cardiac patients who die within four years is less than the proportion of depressed cardiac patients who die within four years.

(d) If there is no difference in the proportions, we will see an observed difference this large or larger only about 1.3% of the time by natural sampling variation.

(e) Since the conditions have already been satisfied above, we will find a two-proportion z-interval.

$$\begin{aligned} & (\hat{p}_{None} - \hat{p}_{Dep}) \pm z^* \sqrt{\frac{\hat{p}_{None}\hat{q}_{None}}{n_{None}} + \frac{\hat{p}_{Dep}\hat{q}_{Dep}}{n_{Dep}}} \\ = & \left(\frac{67}{361} - \frac{26}{89}\right) \pm z^* \sqrt{\frac{\left(\frac{67}{361}\right)\left(\frac{294}{361}\right)}{361} + \frac{\left(\frac{26}{89}\right)\left(\frac{63}{89}\right)}{89}} \\ = & \;]-0.209, -0.004[\end{aligned}$$

(f) We are 95% confident that the proportions of cardiac disease patients who die within four years is between 0.4% and 20.9% lower for non-depressed patients than for depressed patients.

(g) We expect 95% of random samples of this size to produce intervals that contain the true difference between the proportions.

5. (a)
- **Randomization Condition:** The weekdays were not randomly selected. We will assume that the weekdays in our sample are representative of all weekdays.
- **Nearly Normal Condition:** We don't have the actual data, but since the sample of 44 weekdays is fairly large, it is okay to proceed.

The weekdays in the sample had a mean revenue of \$126 and a standard deviation in revenue of \$15. The sampling distribution of the mean can be modeled by a Student's t model, with 44-1=43 degrees of freedom. We will use a one-sample t-interval with 90% confidence for the mean daily income of the parking garage. (By hand, use $t^*_{40} \approx 1.684$)

(b) $\bar{y} \pm t^*_{n-1}\left(\frac{s}{\sqrt{n}}\right) = 126 \pm t^*_{43}\left(\frac{15}{\sqrt{44}}\right) \approx]122.2, 129.8[$

(c) We are 90% confident that the interval \$122.20 to \$129.80 contains the true mean daily income of the parking garage. (If you calculated the interval by hand, using $t^*_{40} \approx 1.684$ from the table, your interval will be $]122.19, 129.81[$, ever so slightly wider from the interval calculated using technology. This is not a big deal.)

(d) 90% of all random samples of size 44 will produce intervals that contain the true mean daily income of the parking garage.

(e) Since the interval is completely below the \$130 predicted by the consultant, there is evidence that the average daily parking revenue is lower than \$130.

(f) The 95% confidence interval would be wider than the 90% confidence interval. We can be more confident that our interval contains the mean parking revenue when we are less precise. This would be better for the city because the 95% confidence interval is more likely to contain the true mean parking revenue.

(g) The 95% confidence interval is wider than the 90% confidence interval, and therefore less precise. It would be difficult for budget planners to use this wider interval, since they need precise figures for the budget.

(h) The confidence interval that was calculated above won't help us to estimate the sample size. That interval was for 90% confidence. Now we want 95% confidence. A quick estimate with a critical value of $z^* = 2$ (from the 68-95-99.7 Rule) gives us a sample size of 100, which will probably work fine. Let's be a bit more precise, just for fun! Conservatively, let's choose t^* with fewer degrees of freedom, which will give us a wider interval. From the table, the next available number of degrees of freedom is $t^*_{80} \approx 1.990$, not much different than the estimate of 2 that was used before. If we

substitute 1.990 for t^*, we can estimate a sample size of about 99. Why not play it a bit safe? Use $n = 100$.

6. (a) We check independence, randomness, the 10% Condition, and that the data comes from a Nearly Normal distribution.

   (b) We estimate with 95% confidence that the true mean is between 34.214 ms and 59.786 ms.

   (c) $H_0 : \mu = 60$ and $H_A : \mu < 60$ we have $P(t < -2.17) < 0.025$ with 15 degrees of freedom so we reject the null hypothesis.

7. (a)
   - H_0: The mean age at which Canadian women first marry is 22.6 years ($\mu = 22.6$)
   - H_A: The mean age at which Canadian women first marry is greater than 22.6 years ($\mu > 22.6$)

   (b)
   - **Randomization Condition:** The 40 women were selected randomly.
   - **Nearly Normal Condition:** The population of ages of women at first marriage is likely to be skewed to the right. It is much more likely that there are women who marry for the first time at an older age than at a very young age. We should examine the distribution of the sample to check for serious skewness and outliers, but with a large sample of 40 women, it should be safe to proceed.

   (c) Since the conditions for inference are satisfied, we can model the sampling distribution of the mean age of women at first marriage with $N\left(22.6, \frac{\sigma}{\sqrt{n}}\right)$. Since we do not know σ, the standard deviation of the population, $\sigma(\bar{y})$ will be estimated by $SE(\bar{y}) = \frac{s}{\sqrt{n}}$, and we will use a Student's t model, with 40-1=39 degrees of freedom.

   (d) The mean age at first marriage in the sample was 27.2 years, with a standard deviation in age of 5.3 years. Use a one-sample t-test, modelling the sampling distribution of $\bar{y}$ with a t-distribution.

$$\begin{aligned} t &= \frac{\bar{y} - \mu_0}{SE(\bar{y})} \\ &= \frac{27.2 - 22.6}{\left(\frac{5.3}{\sqrt{40}}\right)} \end{aligned}$$

$$\approx \quad 5.49$$

The P-value is < 0.00001, since $P(t_{39} > 5.49) < 0.00001$.

(e) If the mean age at first marriage is still 22.6 years, there is a near-zero chance of getting a sample of mean 27.2 years or older simply from natural sampling variation.

(f) Since the P-value is low, we reject the null hypothesis. We have very strong evidence to suggest that the mean age of women at first marriage has increased from 22.6 years, the mean in 1960.