STOR 151 SECTION 1 FINAL EXAM
DECEMBER 10 2019

This is an open book exam. Course text, personal notes and calculator are permitted. You have 3 hours to complete the test. Answers should be written in a blue book. Personal computers and cellphones are not allowed. If you have any queries about the meaning of a question, ask the instructor for advice.

SHOW ALL WORKING — even correct answers will not get full credit if it's not clear how they were obtained. Incorrect answers will gain substantial credit if the method of working is substantially correct.

Answer six of the eight questions. If you attempt more than six, all the answers will be graded but only the best six (complete questions) will count. Each question is worth a total of 20 points and the whole exam is worth 120 points (which will be rescaled to a maximum of 40 for grading purposes). Points for each individual part of a question are also given in square brackets.

Questions may be answered in any order.

1. In an election being held in a certain foreign country, there are three political parties, Conservative, Liberal and Socialist. An opinion poll company is trying to determine how support for the three parties varies with age. They sample 500 people who intend to vote from each of the three age groups: 18–29, 30–49 and 50+. The participants in the survey state their support in the following table (assume each participant supports one and only one party):

Age Group	18–29	30–49	50+	Total
Conservative	112	232	286	630
Liberal	188	130	117	435
Socialist	200	138	97	435
Total	500	500	500	1500

   (a) We are interested in testing whether the distribution of voters among the three parties varies with age group. Show how to formulate this as a hypothesis testing problem and state H_0 and H_A. Is it a test of equal column probabilities within each row, or of equal row probabilities within each column, or of independence? [**4 points**]

   (b) Carry out the test you proposed in (a). Compute the test statistic; state the distribution the test is based on; then state whether you accept or reject the null hypothesis. Assume a significance level of 0.05. [**8 points**]

   (c) The pollster notices that support for the Liberal and Socialist parties seems very similar across the three age groups. *Without* repeating the full calculations for the test of (b), but restricting to the Liberal and Socialist parties, say whether you think there is an age-dependent difference in the support for those two parties. [**4 points**]

   (d) In non-technical language, state your conclusions about how support for the three political parties varies among the three age groups. [**4 points**]

2. In a game played at the NC State Fair, players are invited to throw a baseball at a set of three blocks on a table; if they knock all three blocks of the table, they win a prize. Past results suggest that the probability of winning a prize on a single play of the game is about 0.12.

   (a) Jenny plays the game six times. What is the probability that she wins at least twice? **[6 points]**

   (b) Peter decides to keep playing until he wins once. What are the mean and standard deviation of the number of throws he requires? **[7 points]**

   (c) Over the course of a two-hour session, the game is played 150 times. What is the probability that at least 25 of those games result in winning a prize? **[7 points]**

3. According to data from the Census Bureau and the CDC (adjusted for the purpose of this question — some races are omitted), 63.2% of Americans are non-hispanic white, 12.6% are non-hispanic black, 17.4% are hispanic, 5.6% are Asian, and 1.2% are native American. Among non-hispanic white adults over the age of 20, the proportion with type 2 diabetes is 12.3%. For other races, the corresponding proportions are: non-hispanic black 13.4%, hispanic 13.9%, Asian 13.9%, native American 32.1%.

   (a) Represent this information in the form of a tree diagram, expanding the tree first by race and then by prevalence of type 2 diabetes. **[6 points]**

   (b) What proportion of the total population is non-hispanic white and has type 2 diabetes? **[4 points]**

   (c) What proportion of the total population is hispanic and does not have type 2 diabetes? **[4 points]**

   (d) For someone selected at random from the entire population, if this person has type 2 diabetes, what is the probability that he or she is native American? **[6 points]**

4. According to national statistics, the mean number of hours of TV watched per day by 16-year olds in the US is 2.58, with a standard deviation of 1.71. The superintendent of the Chapel-Hill Carrboro City Schools (CHCCS) is interested in knowing whether those figures apply in this district. She randomly selects 61 16-year-old students from the CHCCS system and asks them to record their viewing hours. The mean for those students is 2.11 hours per day. Assume that the standard deviation in the CHCCS school system is the same as the national value.

   (a) Find a 99% confidence interval for μ, the mean number of hours of TV watched per day by 16-year olds in the CHCCS system. State any assumptions this involves. **[8 points]**

   (b) Test the null hypothesis that $\mu = 2.58$. Use a two-sided test with significance level 0.05. **[8 points]**

   (c) Would you say that the results of (a) and (b) are in any way discrepant? If so, what is the main reason for this discrepancy? **[4 points]**

5. A study is conducted of the exercise habits of men and women in a university. From each of ten dorms, labelled A through J, one man and one woman is randomly selected. They are followed to determine how many minutes of exercise they take, in minutes per day, for a period of one week. The results are as follows:

Dorm	A	B	C	D	E	F	G	H	I	J	Mean	SD
Man	82	61	44	76	96	70	70	57	59	86	70.1	15.5
Woman	59	61	73	60	35	58	46	57	40	75	56.4	12.9
Difference	23	0	−29	16	61	12	24	0	19	11	13.7	22.8

(a) Analyze the data as a paired comparison experiment. Is there a statistically significant difference between the exercise times of men and women? Use a two-sided test with a significance level of 0.05. [**8 points**]

(b) Analyze the data as a two-sample experiment. Is there a statistically significant difference between the exercise times of men and women? Use a two-sided test with a significance level of 0.05. [**8 points**]

(c) Which of the two tests do you feel is more appropriate in this case, and state your overall conclusions. [**4 points**]

6. A woman claims to be able to tell the difference between Coke and Pepsi in a blind tasting. To test this, 60 unlabelled cups are set out, 30 containing Coke and the other 30 containing Pepsi. Each one, she takes a sip and then states which of Coke or Pepsi she thinks it is. If she truly cannot tell the difference, she should get the right answer 50% of the time. In fact, she is successful in 37 guesses (61.7%).

(a) Construct a 99% confidence interval for p, the true probability of success in a single trial. What assumptions are involved here? [**7 points**]

(b) Carry out a formal test of the null hypothesis $H_0:\ p = 0.5$ against the alternative $H_A:\ p > 0.5$ What is the p-value in this case? What conclusion do you draw? [**7 points**]

(c) Suppose we require that the confidence interval in part (a) has margin of error less than 0.1. What sample size would be required for this? [**6 points**]

7. A researcher is interested in determining whether a flu shot is effective in preventing flu in adults over the age of 60. In November, she goes to Southpoint Mall and recruits 100 volunteers over the age of 60; 58 of them have had a flu shot and 42 have not. Six months later, she follows up and asks the same volunteers whether they have had flu in the intervening period. Among those who had the shot, 19 (33%) had flu; among those who did not have the shot, 22 (52%) report having had flu. Define p_1 to be the probability of getting flu for someone who had the flu shot, and p_2 to be the probability of getting flu for someone who did not have the flu shot.

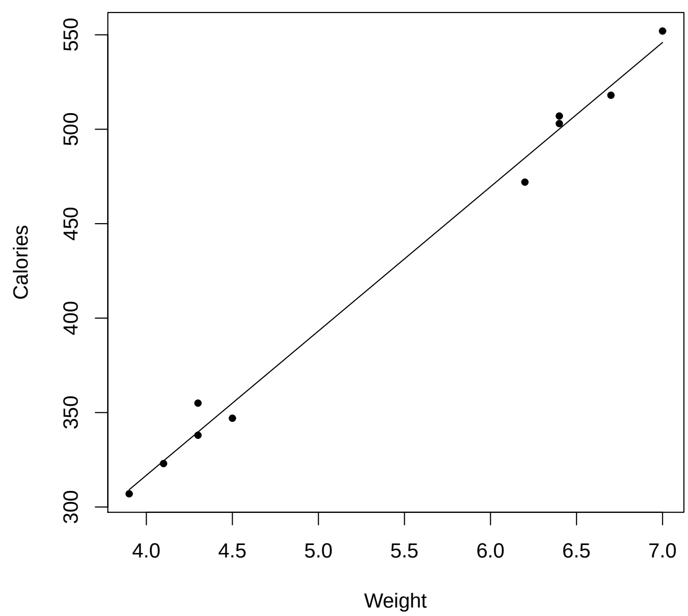
(a) Construct a 99% confidence interval for $p_2 - p_1$. [**8 points**]

(b) Suppose we are interested in testing the null hypothesis $H_0: \ p_1 = p_2$, against the alternative that $p_2 > p_1$. Do you accept or reject this hypothesis? Use the 0.05 significance level. [**8 points**]

(c) Do you think this is a good way to investigate whether flu shots are effective? If not, suggest a better way to do it. [**4 points**]

8. A study of New York bagels listed the mean weights in ounces and calories of bagels from ten New York suppliers, as follows:
BAGEL ON THE SQUARE: 7 ounces, 552 calories.
PICK A BAGEL ON SECOND: 6.7 ounces, 518 calories.
DEAN & DELUCA (made by A & S Bagels): 6.4 ounces, 507 calories.
ESS-A-BAGEL: 6.4 ounces, 503 calories.
NOSH A BAGEL: 6.2 ounces, 472 calories.
H & H: 4.5 ounces, 347 calories.
ZARO'S: 4.3 ounces, 355 calories.
ZABAR'S (made by Columbia): 4.3 ounces, 338 calories.
BAGEL OASIS: 4.1 ounces, 323 calories.
E.A.T.: 3.9 ounces, 307 calories.

A scatterplot of the weights and calories with fitted straight line is as follows:



(a) You are given: $R = 0.9964$, $\bar{x} = 5.38$, $\bar{y} = 422.2$, $s_x = 1.25$, $s_y = 95.8$. Find the intercept and slope of the fitted straight line. [**8 points**]

(b) What is the expected number of calories for a bagel weighing 6 ounces? [**4 points**]

(c) Your dietician advises you not to eat a bagel containing over 400 calories. What is the largest weight of bagel you can eat? [**4 points**]

(d) Which of the bagels is the largest residual in this regression? Is the linear regression of calories on weight is a good fit? Explain why or why not. [**4 points**]

SOLUTIONS AND COMMENTS

Comments added after grading are in red

1. (a) The column totals are fixed; therefore, it is a test of equal column probabilities within each row. If $p_{i,j}$ is the proportion of voters in column j who support party i, the null hypothesis H_0 is that $p_{i,1} = p_{i,2} = p_{i,3}$ for each of $i = 1, 2, 3$, and the alternative H_A is anything that differs from H_0. It is not necessary to use this exact notation (or any mathematical notation) so long as you state the concept correctly.

   (b) The expected values under the null hypothesis are:

210	210	210
145	145	145
145	145	145

   The table of $\frac{(\text{Observed}-\text{Expected})^2}{\text{Expected}}$ (to one decimal place) is:

45.7	2.3	27.5
12.8	1.6	5.4
20.9	0.3	15.9

   The sum of those values is 132.4 whic is clearly significant under the χ^2_4 distribution. Formally: the rejection point under χ^2 at significance level 0.05 is 9.49, and 132.4 is greater than 9.49, so we (very clearly and emphatically) reject the null hypothesis.

   (c) If we repeated the test just for the last two rows we would find that the expected values under the null hypothesis are:

194	134	107
194	134	107

   and it's clear that the discrepancies between observed and expected values are *very* much smaller than in part (b) so we would expecte the null hypothesis to be accepted in this case. (If you do folllow through the χ^2 calculations you find $T = 2.4$, the 0.05-tail point of χ^2_2 is 5.99, so accept H_0.)

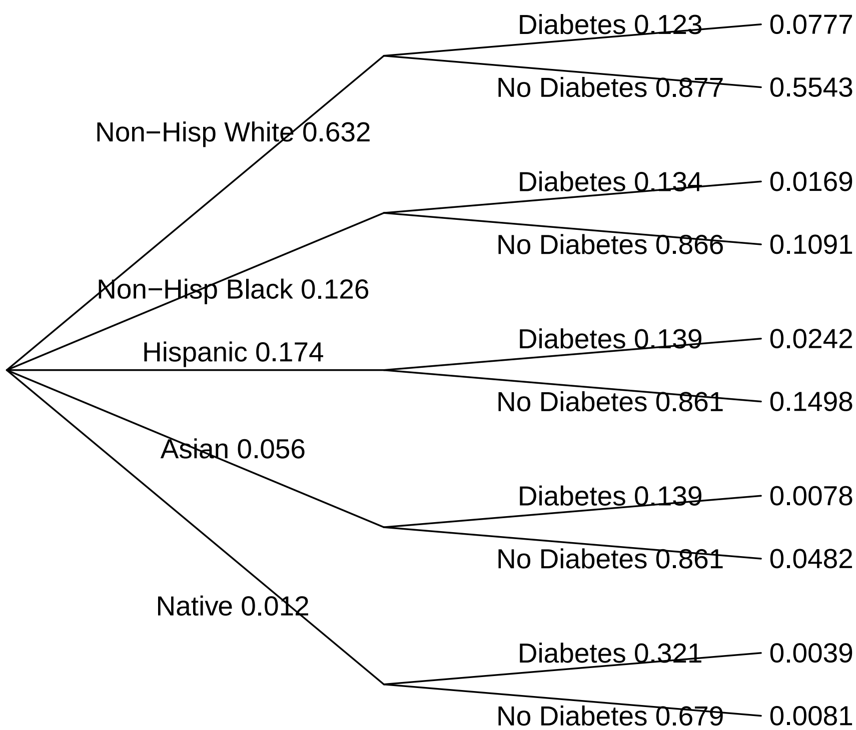
   (d) Support for the Conservative party increases with age. Support for the Liberal and Socialist parties decreases with age and is overall very similar among the three age groups.

2. (a) Binomial distribution: the probability Jenny wins either 0 times or 1 times are respectively $0.88^6 = 0.4644$ and $6 \times 0.88^5 \times 0.12 = 0.3800$. Therefore, the probability she wins at least twice is $1 - 0.4644 - 0.3800 = 0.1556$.

   (b) This is a geometric distribution with $p = 0.12$; the mean is $\mu = \frac{1}{p} = 8.3333$ and the standard deviation is $\sigma = \sqrt{\frac{1-p}{p^2}} = 7.817$,

   (c) With $n = 150,\ p = 0.12$, we have $np = 18 > 10,\ n(1-p) = 132 > 10$ so the normal approximation to the binomial distribution is applicable. The mean number of success is $n \times p = 18$ and the standard deviation is $\sqrt{n \times p \times (1-p)} = 3.9799$. Using the continuity correction, we calculate $z = \frac{24.5-18}{3.9799} = 1.63$; the left-tail probability associated with that is 0.9484. Therefore, the answer is (approximately) $1 - 0.9484 = 0.0516$. Two comments here. Some students wrongly interpreted the questions to mean the probability of exactly 25 successes. In that case, the answer would be $\binom{150}{25}0.12^{25}0.88^{125}$ which comes to 0.02145. Some students did give that exact answer, but since it's answering the wrong question, I only gave partial credit. The other common error was that students used the normal approximation but forgot about the continuity correction. Given the number of times we have mentioned this in the course, I felt it was appropriate to require that for full credit. Partial credit if you used the normal approximation but not the continuty correction.

3. (a) See figure:



(b) $0.632 \times 0.123 = 0.0777$.

(c) $0.174 \times (1 - 0.139) = 0.15$.

(d) The five "diabetes" branches of the tree have joint probabilities 0.0777, 0.0169, 0.0242, 0.0078, 0.0039 (total 0.1305 — this is the overall rate of diabetes). The conditional probability that a randomly selected person is native American, given they have diabetes, is 0.0039/0.1305=0.0299 or 3%.

4. (a) $2.11 \pm 2.58 \times \frac{1.71}{\sqrt{61}} = 2.11 \pm 0.5649 = (1.55, 2.67)$. This assumes independent sampling; an approximately normal distribution (no extreme skewness or extreme outliers) and that

the assumption about the standard deviation being the same as the general population is correct. The t distribution does not apply here because the standard deviation is assumed equal to the population standard deviation — we didn't use the sample standard deviation.

(b) The standard error is $\frac{1.71}{\sqrt{61}} = 0.2189$ so the z statistic is $\frac{2.11-2.58}{0.2189} = -2.15$. For a two-sided test with significance level 0.05, the critical z^* is $1.96 < 2.15$. Therefore, we reject the null hypothesis.

(c) If we interpret the answer to (a) as a hypothesis test, we would accept the null hypothesis (because 2.58 is *inside* the confidence interval). However this would correspond to a two-sided significance level of 0.01 — the main reason for the discrepant results is the difference in significance levels.

5. (a) For a paired comparison test, the standard error is $\frac{22.8}{\sqrt{10}} = 7.21$ and therefore the t statistic is $t = \frac{13.7}{7.21} = 1.90$. From the t table with 9 degrees of freedom, the critical $t^* = 2.26$. Therefore, in this case you accept the null hypothesis. This assumes that the pairs of observations are independently sampled from an approximately normal distribution. In this case, the standard deviations are sample standard deviations, therefore, use t distribution.

(b) For a two-sample test, the standard error is $\sqrt{\frac{15.5^2+12.9^2}{10}} = 6.38$ and therefore the t statistic is $t = \frac{13.7}{6.38} = 2.15$. From the t table with 9 degrees of freedom, the critical $t^* = 2.26$. Therefore, in this case you accept the null hypothesis. This assumes that the male and female samples are randomly selected, independent and have an approximately normal distribution with means μ_M and μ_F. For both this and part (a), the test is for $H_0:\ \mu_M = \mu_F$ against $H_A:\ \mu_M \neq \mu_F$.

Further comment here. Precisely one student in the whole exam used the Satterthwaite approximation, but if you do that, it turns out that the df is 17.4, and the two-sided p-value in that case is 0.046, just significant at significance level 0.05. So it is possible to argue that the data do, in fact, support a difference in exercise levels between men and women. Full credit to the student who gave that answer, but I was *not* expecting you to actually do it that way.

(c) The two tests lead to very similar answers so one could argue it does not matter. There might be some dependence between the individuals because of living in the same dorm (e.g. if one of them was an athletic dorm, presumably both men and women in that dorm would have high exercise levels). However, the more logical and natural assumption is that the men and the women are sampled independently, in which case (b) would be the more appropriate test.

6. (a) The standard error is $\sqrt{\frac{0.617\times0.383}{60}} = 0.0628$ so the 99% confidence interval is $0.617 \pm 2.58 \times 0.0628 = 0.617 \pm 0.162 = (0.455, 0.779)$.

(b) For a hypothesis test we assume $p = 0.5$, so the standard error is $\sqrt{\frac{0.5\times0.5}{60}} = 0.0645$ and the z score for the test statistic is $\frac{0.117}{0.0645} = 1.814$. The one-sided p-value is $1 - 0.9649 = 0.0351$ which is "strong but not overwhelming" evidence that the null hypothesis is false. So the data do, to some extent, support the woman's claim. Here with $n = 60,\ p = 0.5,$

both $np = n(1-p) = 30 > 10$, so the normal approximation to the binomial distribution is satisfied.

(c) In this case we require that $\sqrt{\frac{0.5\times 0.5}{n}} \times 2.58 < 0.1$: this leads to an n of at least 167. Note that we assume the default $p = 0.5$ in questions of this nature because we do not know what the true value of p will be in the future experiment.

7. (a) $\hat{p}_1 = \frac{19}{58} = 0.3276$ and $\hat{p}_2 = \frac{22}{42} = 0.5238$. The standard error of the difference is $\sqrt{\frac{\hat{p}_1\times(1-\hat{p}_1)}{58} + \frac{\hat{p}_2\times(1-\hat{p}_2)}{42}} = 0.0987$. The conditions $n_1p_1 > 10$, $n_1(1-p_1) > 10$, $n_2p_2 > 10$, $n_2(1-p_2) > 10$ are satisfied so we can use the normal approximation. The 99% confidence interval is therefore $0.5238 - 0.3276 \pm 2.58 \times 0.0987 = (-0.0584, 0.4508)$.

(b) For a hypothesis test we use a pooled estimate $\hat{p} = \frac{19+22}{58+42} = \frac{41}{100} = 0.41$ and the standard error formula $\sqrt{\hat{p} \times (1-\hat{p}) \times \left(\frac{1}{58} + \frac{1}{42}\right)} = 0.0997$. The test statistic is $z = \frac{0.5238-0.3276}{0.0997} = 1.9679$ which is greater that the 1.645 critical value for a one-sided test at significance level 0.05. Therefore, we reject H_0 in this case.

(c) There are various points you could make (e.g. people who visit the mall are not necessarily a representative sample, but I think the biggest point is this: it wasn't a randomized test (participants weren't randomized to either receive or not receive the shot) and only a randomized test could definitively prove the hypothesis. A better way to do it would be a randomized test.

8. (a) $b_1 = \frac{Rs_y}{s_x} = 76.3641$ and $b_0 = \bar{y} - b_1\bar{x} = 11.3611$.

(b) $b_0 + 6b_1 = 469.5$.

(c) Solve $b_0 + b_1x = 400$ to get $x = \frac{400-b_0}{b_1} = 5.09$ — in words, do not eat a bagel over 5.09 ounces in weight.

(d) The largest residual (from the figure) is the one associated with the fourth point from the left, which is Zaro's ($x = 4.3$, $y = 355$, predicted value $b_0 + 4.3b_1 = 339.7$, residual 15.3). There is no evidence that the plot differs from a straight line though it could be noted that there seem to be two distinct types of bagels, one with weights in the 3.9–4.5 range and the other in the 6.2–7.0 range.