Final Exam

STAT 305, Section 3
Fall 2019

Instructions

- The exam is scheduled for 120 minutes, from 09:45 AM to 11:45 AM. At 11:45 AM the exam will end.
- Total points for the exam is 100. Points for individual questions are given at the beginning of each problem. Show all your calculations clearly to get full credit. Put final answers in the box at the right (except for the diagrams!).
- A formula sheet is attached to the end of the exam. Feel free to tear it off.
- You may use a calculator during this exam.
- Answer the questions in the space provided. If you run out of room, continue on the back of the page.
- If you have any questions about, or need clarification on the meaning of an item on this exam, please ask your instructor. No other form of external help is permitted attempting to receive help or provide help to others will be considered cheating.
- **Do not cheat on this exam.** Academic integrity demands an honest and fair testing environment. Cheating will not be tolerated and will result in an immediate score of 0 on the exam and an incident report will be submitted to the office of the dean.

Name: _____

Student ID: _____

1. (2 points) A random sample of 1000 students' Statistics exam scores was drawn from the population of all possible comparable Stat exam scores (an unknown population/distribution). The sample mean, once computed, has the exact value of the distribution/population mean.

   A. True B. False

2. (2 points) While trying to figure out the probability that the sample mean for a data of size 10 would exceed a value, we can apply the central limit theorem.

   A. True B. False

3. An agriculturist is attempting to determine which of three species of corn (A, B, and C) yield the most grain per acre. Since the yield may depend on the fertilizer used, the researcher intends to use fertilizers with different concentrations of Nitrogen as well - low Nitrogen, medium-low Nitrogen, medium-high Nitrogen, and high Nitrogen. There are 8 fields (scattered around Iowa) available to perform this expiriment. Each field is divided into 24 single acre plots and the combinations of species and fertilizer are randomly assigned so that within each field every combination is used exactly twice. Since the size of the plants may impact their growth when placed close by each other, it was decided that all species would be planted in a grid with each plant exactly four feet from its nearest neighbor. The agriculturalist also decided not to use any pest control system during growth. At harvest time, the weight of grain each plot yields is recorded and the combination of corn species and fertilizer that gives the highest average yield is chosen.

   (a) (2 points) Explain why this is an experiment and not an observational study.

   The experimenter has an active role in manipulation of the experiment under study by intentionally changin fertilizer amount & species of corn.

   (b) Identify each of the following and describe them as numeric (in which case, identify whether it is continuous or discrete) or categorcial (in which case list the possible levels).

   i. (2 points) Identify the response variable(s):

   weight of grain yield

   ii. (2 points) Identify the experimental variable(s):

   Cois species (A, B, C) & fertilizer amount.

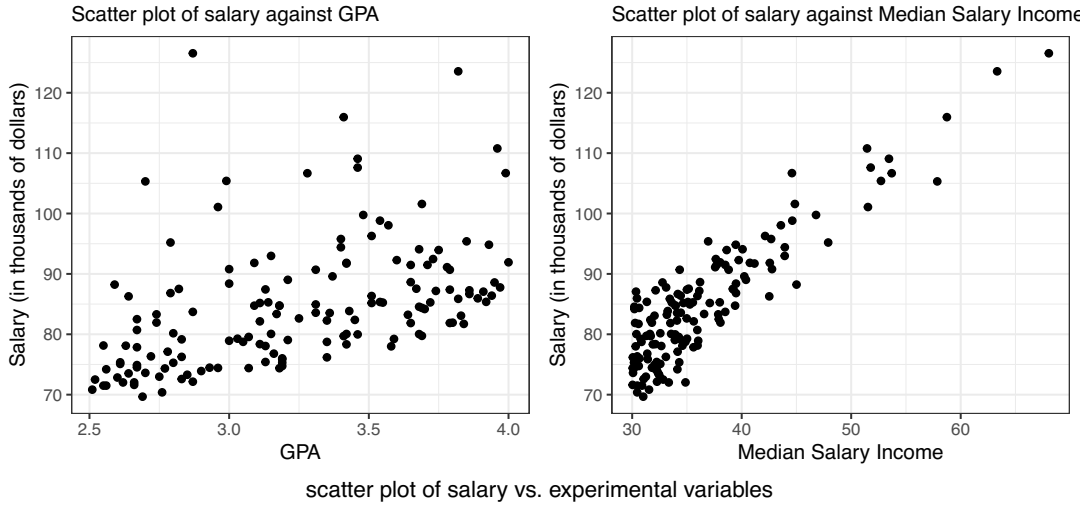
   iii. (2 points) Blocking variable(s):

   8 Fields around IA.

   (c) (2 points) Identify two controlled variables used in this process.

   - distance between the plants
   - No pest control system during growth.

4. A survey given to members of a national engineering society who graduated five years prior is attempting to determine the relationship between salary and undergraduate GPA. The graph below displays 150 responses.



The normal qq-plot of residual of the fit and the residual vs. predicted plot are as follows:

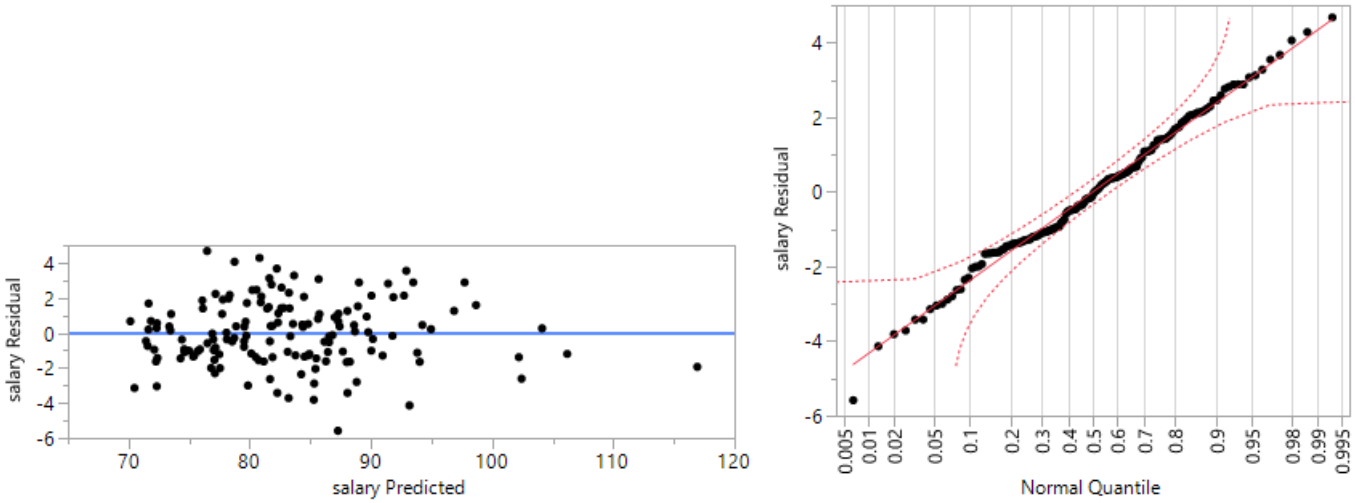


Figure 1: Residual vs. predicted variables (the left) and Normal quantile plot (the right)

Discouraged by the relationship between salary and GPA, the surveyors remember that they know the address of each respondant and are able to determine the median income of the area in which the respondant lives. The JMP output below comes from fitting a linear relationship using for annual salary of the respondant ("`salary`") using both the undergraduate GPA ("`GPA`") and the median income of the area in which the respondant lives ("`med_salary_loc`") (in thousands of dollars).

Response salary

Whole Model

Effect Summary

Summary of Fit

RSquare	0.945622
RSquare Adj	0.944882
Root Mean Square Error	1.866443
Mean of Response	83.03367
Observations (or Sum Wgts)	150

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	8905.0931	4452.55	1278.142
Error	147	512.0904	3.48	Prob > F
C. Total	149	9417.1835		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.4351638	1.590643	2.16	0.0324*
GPA	10.143862	0.349311	29.04	<.0001*
med_salary_loc	1.3328123	0.034538	38.59	<.0001*

(a) (2 points) Looking at the two plots on the **bottom of page 2** (Residual vs. predicted variables () and Normal quantile plot), are the assumptions of the regression model met?

Yes. The residual plot shows no pattern among the residual vs. predicted values & QQ plot shows that the residuals are normally distributed as they are around the line.

(b) (3 points) Using the JMP output, write the equation of the fitted multivariate linear relationship between GPA, median salary income and salary

$\hat{y} =$ 3.43 + 10.14(GPA) + 1.33(med. salary)

(c) (3 points) Using this fitted multivariate linear relationship, what do we suppose the salary would be for an engineer with a undergraduate GPA of 3.19 living in a location with a median income of 31.42 thousand dollars?

$\hat{y} =$ 3.43 + 12.14(3.19) + 1.33(31.42)

= 77, 5652

(d) (3 points) The actual salary of one engineer surveyed with a undergraduate GPA of 3.19 living in a location with a median income of \$31,420 was \$75,870. What is the residual for this specific engineer's actual salary using the fitted multivariate linear relationship?

e = -1.6952 or -16,952 $

$e = y - \hat{y}$

$= 75.870 - 77.5652 = -1.6952$

or $-\$16{,}952$

(e) (4 points) For the linear relationship, find r, the sample correlation coeffecient and R^2, the coeffecient of determination.

$r =$ 0.9724

$R^2 =$ 94.56%

(f) (3 points) For your response for R^2 in previous part, provide an interpretation.

94.56% of the variation among salary can be explained by the MLR between salary, GPA and med salary income.

(g) (3 points) Provide an estimate for σ^2.

$\hat{\sigma}^2 = s^2_{SF} =$ 3.48

(h) (3 points) Provide an estimate for the *variance* of the coefficient of *GPA*. $\text{Var}(b_1) =$ 0.1156

$0.34^2 = .1156$

(i) (5 points) Calculate and intepret the 95% two-sided confidence interval for the coefficient of *GPA*.
(9.4736 , 10.864)

$b_1 \pm t_{(n-p,\, 1-\alpha/2)}\, SE(b_1)$

$10.14 \pm t_{(147,\, 0.975)}(0.34)$

$10.14 \pm 1.96(0.34)$

$(9.4736,\ 10.8064)$

(j) (10 points) Conduct a formal hypothesis test at the $\alpha = 0.05$ significance level to determine if there is significance relationship between salary (y) and *median salary income* (x_1), holding GPA constant.
Note: Write down all six steps for full credit.

① $H_0: \beta_2 = 0$ vs. $H_a: \beta_2 \neq 0$

② $\alpha = 0.05$

③ I'll use $K = \dfrac{b_2 - 0}{SE(b_2)}$. If (i) The H_0 is true & (ii) The regression $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$ is valid,

$K \sim t_{(n-p)}$.

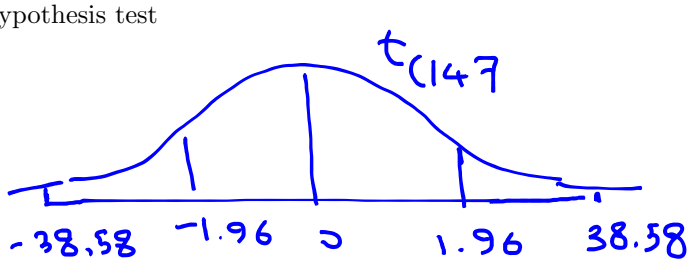
④ $K = \dfrac{1.33 - 0}{0.034} = 38.5891 \left(\sim 39.117\right)$

p-value $= P(|T| > K) = P(|T| > 38.5891)$

$T \sim t_{(147)}$

You may use this blank space for completing the steps of hypothesis test

$t_{(147\ 0.975)} \approx 1.96$



p-value $< t_{(149, 0.975)}$

(5) since p-value $< \alpha$, we reject H_0.

(6) There's enough evidence to reject H_0 concluding that there's significant linear relationship between salary and the median salary income holding GPA constant.

5. A team of engineers is studying the differences in camera systems on self-driving cars. Their primary concern is in the car's ability to avoid obstacles. Each system was installed in a test car and on 10 consecutive days the car was sent on a 15 hour drive through a closed obstacle course where the number, timing, location, and type of obstacle the car encounters is randomly determined. The proportion of obstacles avoided during the 15 hour drive is independently recorded below (along with relevant summary statistics):

   - System 1: 0.87, 0.78, 0.79, 0.83, 0.86, 0.83, 0.86, 0.82, 0.82, 0.77 (with $\bar{x}_1 = 0.82, s_1^2 = 0.0012$)
   - System 2: 0.83, 0.77, 0.78, 0.77, 0.74, 0.77, 0.76, 0.73, 0.76, 0.78 (with $\bar{x}_2 = 0.77, s_2^2 = 0.0007$)

   Note: there are 10 sample data of each system. i.e $n_1 = n_2 = 10$

   (a) (5 points) Provide a 95% confidence interval for the mean proportion of obstacles avoided on the course using System 1. (0.7952 , 0.8447)

$\bar{x}_1 \pm t_{(n-1, 1-\alpha/2)}\ s_1/\sqrt{n_1}$

$0.82 \pm t_{(9, 0.975)} \sqrt{\frac{0.0012}{10}}$

$0.82 \pm 2.262(0.0109)$

   (b) (5 points) Provide a one-sided 95% lower bound confidence interval for the mean proportion of obstacles avoided on the course using System 2. (0.7546 , $+\infty$)

$(\bar{x}_2 - t_{(n-1, 1-\alpha)}\ s_2/\sqrt{n_2}\ , +\infty)$

$(0.77 - t_{(9, 0.95)} \sqrt{\frac{0.0007}{10}}, \infty)$

$(0.77 - 1.833(0.0083), +\infty)$

$(0.77 - 0.0153, +\infty)$

(c) (10 points) Assuming that the proportion of obstacles avoided is roughly normally distributed and that the both systems have the same variance in proportion of obstacles avoided, conduct a hypothesis test at the $\alpha = 0.05$ significance level for the claim that the true mean of the obstackes avoided of the two systems are significantly different.
Note: Write down all six steps for full credit.

① $H_0: \mu_1 - \mu_2 = 0$ vs. $H_a: \mu_1 - \mu_2 \neq 0$

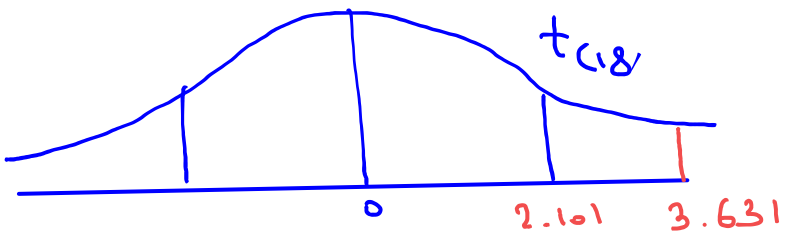
② $\alpha = 0.05$

③ I'll use $K = \dfrac{\bar{x}_1 - \bar{x}_2 - 0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ assuming that ① H_0 is true, ② $x_1, \dots x_{n_1}$ are $\overset{iid}{\sim} N(\mu_1, \sigma_1^2)$ ③ independent of sample ④ $\overset{iid}{\sim} N(\mu_2, \sigma_2^2)$ & ⑤ $\sigma_1^2 \approx \sigma_2^2$. Then $K \sim t_{(n_1+n_2-2)}$

④ $S_p = \sqrt{\dfrac{(9)(0.0012) + 9(0.0007)}{18}} = \sqrt{\dfrac{0.0171}{18}} = \sqrt{0.00095} = 0.0308$

$$K = \frac{0.82 - 0.77}{0.0308\sqrt{\frac{1}{10} + \frac{1}{10}}} = \frac{0.05}{0.01377} = 3.631$$

p-value $= P(|T| > 3.631)$, $t_{(n_1+n_2-2,\, 1-\alpha/2)} = t_{(18,\, 0.975)} = 2.101$



⑤ p-value $> \alpha$, So we reject the null.

⑥ There's enough evidence to reject H_0 concluding that the true mean obstacles avoided by the two systems are significantly different.

6. Let X be a normal random variable with a mean of -2 and a varaince of 16 (i.e., $X \sim N(-2, 16)$) and let Z be a random variable following a standard normal distribution. Find the following probabilities (note: the attached standard normal probability table may be helpful):

(a) (2 points) $P(Z \leq 1.0)$

$= \Phi(1) = 0.8413$

(b) (2 points) $P(|Z| \leq 2.5)$ $= \Phi(2.5) - \Phi(-2.5)$

$= 0.9938 - 0.0062$

$= 0.9876$

(c) (2 points) $P(-8 \leq X < 1)$ $= P\left(\frac{-8-(-2)}{4} < Z < \frac{1-(-2)}{4}\right)$

$= P\left(-\frac{6}{4} < Z < \frac{3}{4}\right)$

$= \Phi(3/4) - \Phi(-3/2) = 0.7733 - 0.0668$

$= 0.7065$

(d) (3 points) $P(|X| \geq 4)$

$P(X>4) + P(X<-4) = P\left(Z > \frac{4+2}{4}\right) + P\left(Z < \frac{-2}{4}\right)$

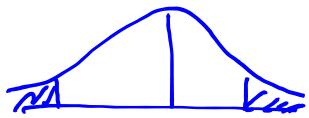
$= P(Z > 3/2) + P(Z < -1/2)$

$= 1 - \Phi(3/2) + \Phi(-1/2) = 1 - 0.9331 + 0.3085$

$= 0.3754$

(e) (5 points) find c such that $P(|X| \leq c) = 0.95$

$0.95 = P(-c < X < c) = P\left(\frac{-c+2}{4} < Z < \frac{c+2}{4}\right)$



$0.95 = 1 - 2\Phi\left(\frac{-c+2}{4}\right)$

$\Phi\left(\frac{-c+2}{4}\right) = 0.025 \rightarrow \frac{-c+2}{4} = -1.96$

$\Rightarrow -c+2 = -7.48$

$c = 9.48$

7. Seventy independent messages are sent from an electronic transmission center. Messages are processed sequentially, one after another. Transmission time of each message is Exponential with parameter $\alpha = 10$min.

(a) (3 points) what are the expected value and variance of the sample mean of all 70 messages?

$X \sim Exp(\alpha = 10)$

$E\bar{X} = 10$

$Var(\bar{X}) = \frac{10^2}{70} = \frac{100}{70} = \frac{10}{7}$

$E(\overline{X}) = 10$

$Var(\overline{X}) = \frac{10}{7} = 1.4285$

(b) (4 points) Find the probability that the average of all 70 messages are transmitted in less than 8 minutes.

$p = 0.04745$

$n > 25 \longrightarrow CLT$

$$P(\bar{X} < 8) = P\left(Z < \frac{8-10}{\sqrt{1.4285}}\right) = \Phi\left(\frac{-2}{1.1951}\right)$$

$$= \Phi(-1.6735)$$

$$= 0.04721$$

8. Two independent discrete random variab1e X and Y can be described using the following probability functions:

x	-1	0	1	2
$f_X(x)$	0.2	0.3	0.3	0.2

y	1	2	3	4
$f_Y(y)$	0.3	0.2	0.2	0.3

%

(a) (4 points) Find the means and standard deviations for X and Y respectively.

$EX = (-1)(0.2) + 0 + (0.3) + 0.4 = 0.5$

$EX^2 = 0.2 + 0.3 + 0.8 = 1.3$

$Var(X) = 1.3 - 0.5^2 = 1.05 \longrightarrow SD(X) = 1.024$

$EY = 0.3 + 0.4 + 0.6 + 1.2 = 2.5$

$EY^2 = 0.3 + 0.8 + 1.8 + 4.8 = 7.7$

$Var(Y) = 7.7 - 2.5^2 = 1.45 \longrightarrow SD(X) = 1.2041$

$\mu_X =$	0.5	$\sigma_X =$	1.024
$\mu_Y =$	2.5	$\sigma_Y =$	1.2041

(b) (2 points) Find the mean and standard deviation for the random variable $2X - 3Y + 5$.

$E(2X - 3Y + 5) = 2EX - 3EY + 5$

$= 2(0.5) - 3(2.5) + 5 = -1.5$

mean =	-1.5	s.d=	4.153

$V(2X - 3Y + 5) = 4V(X) + 9V(Y)$

$= 4(1.05) + 9(1.45)$

$= 17.25$

$\Rightarrow S.d = \sqrt{17.25} = 4.153$