ST 305: Final Exam Spring 2017

By handing in this completed exam, I state that I have neither given nor received assistance from another person during the exam period. I have used no resources other than the exam itself and the basic mathematical functions of a calculator (ie, no notes, electronic communication, notes stored in calculator memory, etc.) Using your calculator for values from probability distributions like the normal or t is OK; however, if you are doing a calculation from a normal, t, or F distribution show your work all the way to the point of calculating the z, t, or F statistic. I have not copied from another person's paper. I understand that the penalty if I am found guilty of any such cheating will include failure of the course and a report to the NCSU Office of Student Conduct. **I understand that I must show all work/calculations, even if they seem trivial, to get credit for my answers.**

Name: KEY

ID#: ____________

$\bar{x} = \frac{1}{n}\sum x_i$ $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$ $Z = \frac{X - \mu}{\sigma}$ $r = \frac{\sum \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)}{n-1}$ $b_1 = r\frac{s_y}{s_x}$ $b_0 = \bar{y} - b_1\bar{x}$ residual $= y - \hat{y}$ $P(A \text{ or } B) = P(A) + P(B)$ $P(A^C) = 1 - P(A)$ $P(A \text{ and } B) = P(A) \times P(B)$	$\mu_X = \sum x_i p_i$ $\mu_{a+bX} = a + b\mu_X$ $\mu_{X+Y} = \mu_X + \mu_Y$ $\sigma_X^2 = \sum (x_i - \mu_X)^2 p_i$ $\sigma_{a+bX}^2 = b^2\sigma_X^2$ $\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$ $\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2$ $\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y$ $\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y$ $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$ $P(A \text{ and } B) = P(A)P(B \mid A)$ $P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}$	$\mu_X = np$ $\sigma_X = \sqrt{np(1-p)}$ $\hat{p} = X / n$ $\mu_{\hat{p}} = p$ $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ $P(X = k) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$ $\mu_{\bar{X}} = \mu$ $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ $m = z^* \sigma/\sqrt{n}$ $\bar{x} \pm m$ $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ $n = \left(\frac{z^*\sigma}{m}\right)^2$

Simple Linear Regression

$$b_1 = r\frac{s_y}{s_x};\ \ b_0 = \bar{y} - b_1\bar{x}$$

$$e_i = y_i - \hat{y}_i;\ \ s^2 = \frac{\sum e_i^2}{n-2}$$

$$b_j \pm t^* SE_{b_j};\ \ t = \frac{b_j}{SE_{b_j}}\ \ \text{df} = n-2$$

$$\hat{\mu}_y \pm t^* SE_{\hat{\mu}};\ \ \hat{y} \pm t^* SE_{\hat{y}}$$

$$\text{SST} = \sum (y_i - \bar{y})^2$$

$$\text{SSM} = \sum (\hat{y}_i - \bar{y})^2$$

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2$$

$$SE_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$$

$$SE_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$$

$$SE_{\hat{\mu}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

$$SE_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

Multiple regression changes:

$$s^2 = \frac{\sum e_i^2}{n-p-1}$$

$$\text{df} = n - p - 1$$

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}},\ \text{df} = n-1$$

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},\ \text{df} = \min(n_1, n_2) - 1$$

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},\ \text{df} = n-1$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},\ \text{df} = \min(n_1, n_2) - 1$$

$$(\bar{x}_1 - \bar{x}_2) \pm t^* s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\ \text{df} = n_1 + n_2 - 2$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},\ \text{df} = n_1 + n_2 - 2$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$\psi = \sum a_i\mu_i,\ \ c = \sum a_i\bar{x}_i$$

$$SE_c = s_p\sqrt{\sum \frac{a_i^2}{n_i}}$$

Simple Linear Regression

$$b_1 = r\frac{s_y}{s_x}; \;\; b_0 = \bar{y} - b_1\bar{x}$$

$$e_i = y_i - \hat{y}_i; \;\; s^2 = \frac{\sum e_i^2}{n-2}$$

$$b_j \pm t^* SE_{b_j}; \;\; t = \frac{b_j}{SE_{b_j}} \text{ df} = n-2$$

$$\hat{\mu}_y \pm t^* SE_{\hat{\mu}}; \;\; \hat{y} \pm t^* SE_{\hat{y}}$$

$$\text{SST} = \sum\left(y_i - \bar{y}\right)^2$$

$$\text{SSM} = \sum\left(\hat{y}_i - \bar{y}\right)^2$$

$$\text{SSE} = \sum\left(y_i - \hat{y}_i\right)^2$$

$$SE_{b_1} = \frac{s}{\sqrt{\sum\left(x_i - \bar{x}\right)^2}}$$

$$SE_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum\left(x_i - \bar{x}\right)^2}}$$

$$SE_{\hat{\mu}} = s\sqrt{\frac{1}{n} + \frac{\left(x^* - \bar{x}\right)^2}{\sum\left(x_i - \bar{x}\right)^2}}$$

$$SE_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{\left(x^* - \bar{x}\right)^2}{\sum\left(x_i - \bar{x}\right)^2}}$$

Multiple regression changes:

$$s^2 = \frac{\sum e_i^2}{n - p - 1}$$

$$\text{df} = n - p - 1$$

Chapter 7 Stuff

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}, \text{ df} = n-1$$

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \text{ df} = \min(n_1, n_2) - 1$$

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \text{ df} = n-1$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \text{ df} = \min(n_1, n_2) - 1$$

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm t^* s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \text{ df} = n_1 + n_2 - 2$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \text{ df} = n_1 + n_2 - 2$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Definitions. (1 point each) **Clearly define each of the following terms.**

1. Distribution:

2. Correlation:

3. Statistic:

4. Type II Error:

5. Sample Space:

6. Central Limit Theorem:

7. Simple Random Sample:

8. Random Variable:

9. Residual:

10. p-value:

Pick the Procedure. (2 points each)

Select the procedure that would be the best approach to answer each of the following questions. You only need to give the letter of the answer (**please write it clearly to the left of each question**)

a. CI for a single mean
b. HT for a single mean
c. CI for a difference in two means
d. HT for a difference in two means
e. Matched Pairs CI
f. Matched Pairs HT
g. Simple linear regression
h. Multiple linear regression
i. 1-Way ANOVA
j. 2-Way ANOVA

A 11. How much do baby giraffes usually weigh?

(ch13) J 12. Do either sex (M, F) or blood type (A, B, or O) impact the density measurement of a patient's blood?

I 13. Are there differences in the average GPAs of seniors graduating from different universities in the Atlantic Coast Conference?

C 14. How much faster do sprinters run on synthetic tracks than on asphalt tracks?

(ch13) J 15. My manager wants to design an experiment to compare the effects of three marketing approaches, accounting for differences in effectiveness on children and adults.

F 16. Is there a difference between the average person's weight in the morning and at night?

G 17. If I spend 5 hours next month recruiting applicants for a job position, how many people will apply?

G 18. Is the average SAT score for a state related to the percentage of students in that state who apply to out-of-state colleges?

H 19. Do any of age, weight, or height help predict a person's IQ score?

I 20. After running a randomized experiment to look at profits from 8 investment strategies, I want to know if there are significant differences among them.

Multiple Choice. (2 points each)

21. A 95% confidence interval for mean response based on $n = 200$ observations will be
a. wider than a 95% prediction interval based on the same data
b. wider than a 99% confidence interval for mean response based on the same data
c. wider than a 90% confidence interval for mean response based on the same data

22. If X and Y are independent with standard normal distributions, Z = X - Y has variance
a. 0
b. 1
c. 2

23. A multiple regression analysis with a value of R^2 near 1 indicates
a. that at least one of the explanatory variables is statistically significant.
b. that none of the explanatory variables are statistically significant
c. that the explanatory variables explain most of the variation of y

24. 1-Way ANOVA is most useful when
a. Looking at the effects of several factors on a variable
b. Looking at the effects of several levels of one factor on a variable
c. Looking at the effects of several explanatory variables on a variable

25. Contrasts are likely to be part of
a. A multiple regression analysis
b. A 1-Way ANOVA analysis
c. A study with no pre-planned comparisons

26. If we have a large SRS of charities and record the annual income for each,
a. The histogram of incomes is likely to be bell-shaped
b. The difference between the mean and median incomes is probably small
c. The distribution of the sample mean income will probably be close to a normal distribution (But this is a poorly worded question)

27. Boxplots would be most useful in conjunction with
a. Simple linear regression analyses
b. 1-Way ANOVA analyses
c. Studies of interaction

School	Male	Fem.	School	Male	Fem.	School	Male	Fem.
Stanford*	144.4	129.9	Chicago*	122.5	110.2	UT Dallas	113.1	102.6
Columbia*	137.6	120.9	Duke*	120.6	97.7	S. Florida	111.8	74.1
Stevenson*	129.8	59.4	Yale*	120.4	113.3	UCLA	110.7	105.3
Penn*	127.1	107.8	Bryant*	119.6	107.6	Rutgers	109.1	113.9
MIT*	126.2	125.3	Penn State	122.9	113.7	Michigan	104.9	98.9
Harvard*	124.6	118.4	UC Berkeley	113.6	100.9	UNC	96.6	94.1
			Kansas	83.9	76.8	U Mass	101.8	93.7

The data in the table above are ***average*** salaries for male and female faculty at 20 US Universities (in thousands of dollars). The schools indicated with * are private, the rest are public. Here are some summary statistics for the data, with separate tables for Male and Female, separated by Public or Private within each table. Use the data and tables for questions 28-31.

Options

Summary statistics for Male:
Group by: PubPri

PubPri	n	Mean	Variance	Std. dev.	Median	Range	Min	Max	Q1	Q3
Pri	10	127.28	65.346222	8.083701	125.4	24.8	119.6	144.4	120.6	129.8
Pub	10	106.84	116.40933	10.789316	109.9	39	83.9	122.9	101.8	113.1

Summary statistics for Female:
Group by: PubPri

PubPri	n	Mean	Variance	Std. dev.	Median	Range	Min	Max	Q1	Q3
Pri	10	109.05	393.73611	19.842785	111.75	70.5	59.4	129.9	107.6	120.9
Pub	10	97.4	181.52444	13.473101	99.9	39.8	74.1	113.9	93.7	105.3

28. The individuals in this data set are
    a. students
    b. universities
    c. faculty

29. The variable PubPri is
    a. categorical
    b. quantitative
    c. dependent

30. The *standard error* of the estimated mean for Females at Private universities is
    a. 19.84
    b. 6.27
    c. 7.05

$s/\sqrt{n}$

31. (Continue using the data and output from the previous page)

a) Carry out a test to see if the average salary of Male professors is higher than that of Female professors. Hint- pay attention to your answer to number 28. (5 points)

You will need a *standard error* that is not easily found from the table above- use the value 3.5.

Note that this is a paired sample - the M & F data come from the same individual (university). After computing the 20 differences, $\bar{x}_D = 13.8$

$H_0: \mu_D = 0$

$H_A: \mu_D > 0$

$\bar{x}_D = 13.8$

$t = \frac{13.8}{3.5} = 3.94 \quad df = 19$

p-val < .0005

$\Rightarrow$ reject H_0 & conclude that the mean M salary is greater than the mean F salary

b) Compute a 95% CI for the difference between the average salaries of Female professors at Public and Private universities. (4 points)

	$\bar{x}$	s
Pub	97.4	13.47
Pri	109.05	19.84

$\bar{X}_1 - \bar{X}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad df = 19$

$(97.4 - 109.05) \pm 2.093 \sqrt{\frac{13.47^2}{20} + \frac{19.84^2}{20}}$

c) Do you think your answers to parts (a) and (b) can be used to tell us about the average salaries of faculty at *all* US universities? Justify your answer. (2 points)

No. It is definitely not a SRS. "Elite" private universities & very large public universities make up most of the sample.

32. An experiment was conducted to investigate the effect of exercise (Yes, the subject exercised within 3 hours before going to bed or No, they did not exercise) on the quality of sleep (measured by a trained observer on a scale of 1-10, with 10 indicating high quality). The raw data from that experiment is in the following table; the entries in each cell are the sleep quality values for each of five subjects:

	Male	Female
Yes	10, 7, 9, 6, 8	5, 4, 6, 3, 2
No	5, 4, 7, 4, 5	3, 4, 5, 1, 2

a. Draw and label an appropriate graph to determine if there appears to be interaction between Gender and Exercise. (5 points)

CH 13

Not covered this semester

Much of this is Ch 13 - but expect a Ch 12 1-way ANOVA question like this

(32-continued) Here is the ANOVA Table for analyzing the data above.

Source	df	SS	MS	F	P-value
Exercise	1	20	20	8.89	0.0088
Gender	1	45	45	20.00	???
Gender x Exercise	1	5	5	2.22	0.1557
Error	(i)	36	2.25		
Total	(ii)	106			

b. What are the missing values labeled (i) and (ii) above? (1 point each)

i: $2.25 = 36/(i)$ ii: $16+1+1+1 = 19$

$(i) = 16$

c. What are the *factors* in this study? What are the *levels* of each factor? (4 points)

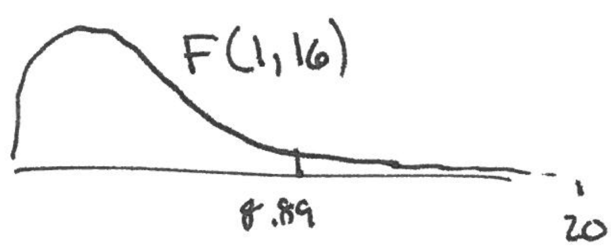
Exercise [Yes/No] Gender [Male/Female]

d. Is there statistically significant evidence of interaction? Justify your answer. (2 points)

No- pval=.1557 (CH 13)

e. Is the effect of Gender significant at the 0.05 level? Justify your answer (no, you don't need an F table) (2 points)

Yes. F=20, w/ same df (1,16) as Exercise. Since Exercise F was 8.89 & had p = .0088, the p-val for Gender (F=20) MUST be smaller.



33. The following table lists the 75th percentiles for the SAT math and Reading scores of incoming freshmen at the 12 ACC schools.

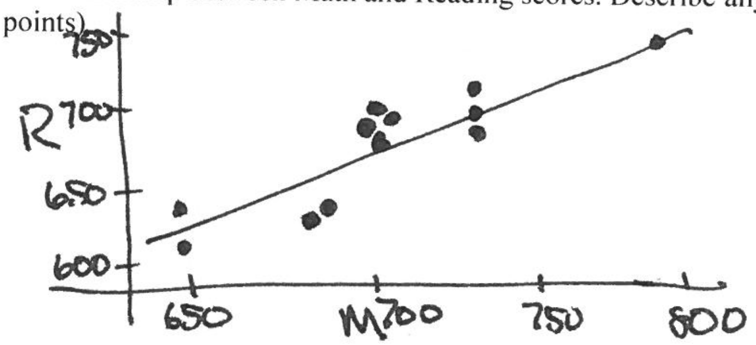
School	Math	Reading	School	Math	Reading	School	Math	Reading
Boston College	730	700	Georgia Tech	730	690	NC State	650	610
Clemson	680	640	Maryland	700	680	Virginia	730	710
Duke	790	750	Miami	700	680	Virginia Tech	670	630
Florida State	650	640	North Carolina	700	700	Wake Forest	710	690

a. Display the distribution of the Math scores for the 12 universities (3 points)

7 | 9
7 | 0001333
6 | 5578
6 |

b. Make a plot to look for a relationship between Math and Reading scores. Describe any relationship you find. (3 points)

There seems to be a fairly strong, positive, linear rel. w/ no outliers.



d. Provide a numerical summary of the distribution of Math score. The standard deviation of the levels above is calculated to be s = 39.4. (2 points)

Since it is not apparently skewed, & no outliers:

$\bar{x} = 700.3$

$s = 39.4$

e. Compute a 95% CI for the mean Math score. (4 points)

$700.3 \pm t^* \, 39.4/\sqrt{12}$

↑

df=11 2.201

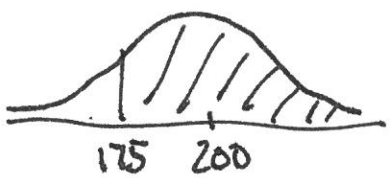
34. A mobile coffee vendor sells nothing but coffee: \$2 a cup. 80% of customers by one cup of coffee, 20% buy 2 cups. The number of customers per day follows a normal distribution with mean 200 and standard deviation 50. The vendor's daily costs are \$100/day for fixed expenses (gas, licenses, power, etc) plus \$0.50 to make each cup of coffee.

a. Find the probability that there will be more than 175 customers in a single day. (3 points)

C = # customers in 1day $C \sim N(200, 50)$

$$P(C > 175) = P\left(Z > \frac{175-200}{50}\right) = P(Z > -0.5)$$

$$= .6915$$



b. Find the mean and standard deviation of the amount spent by an individual customer. (4 points)

1 cup	\$2	.8
2 cups	\$4	.2

Mean: $2(.8) + 4(.2) = \boxed{\$2.40}$

Var: $(2-2.4)^2(.8) + (4-2.4)^2(.2) = .768$

$\Rightarrow$ std dev $= \sqrt{.768} = \boxed{.876}$

c. If the vendor has 240 customers one day, what is the probability that at least 25% of them buy 2 cups of coffee? (5 points)

$240 \times .25 = 60$

Let # who buy 2 cups = X

$X \sim Bin(240, .2)$

$X \dot{\sim} N(48, \sqrt{38.4})$ $\rightarrow \sqrt{np(1-p)}$

6.2

$$P(X \geq 60)$$

$$= P\left(Z > \frac{60-48}{6.2}\right)$$

$$= P(Z > 1.94) = .0262$$