

Math 114 - Final
Tuesday, December 11th, 2007

Name: _____

Instructions:

1. On the multiple choice questions, you are required to circle the correct answer for each question. No partial credit will be given on those questions.
2. On the True/False questions, you are required to write T (True) or F (False) on top of the line provided. No partial credit will be given on those questions.
3. Answer all questions and **show all work** on the exam in the space provided (you may use the backs of pages if necessary). You will **not** receive credit if you do not justify your answers.
4. Please draw a **circle** or **box** around your final answers.
5. Unless otherwise specified, give exact answers.
6. This test is conducted under Tulane Honor Code: all work is your own.
7. Good luck!

1. (2 points) Circle the letter of the correct choice:

Given the sharp rise in insurance premiums after Hurricane Katrina, a pollster ran a survey among car owners in New Orleans with the question “Does your insurance policy cover damage to your auto?”. 750 selected users responded to this survey. In this case, the collection of responses from the 750 users is

- (a) the population
- (b) the sample
- (c) the variable
- (d) the unit

2. (2 points) Circle the letter of the correct choice:

In 1995, the Federal Highway Administration (FHWA) inspected each of the 470,515 highway bridges in the United States. All bridges were categorized as structurally deficient, functionally obsolete, or safe.

We say the data set obtained this way by the FHWA comes from

- (a) a published source
- (b) a designed experiment
- (c) an observational study
- (d) a survey

3. (2 points) Circle the letter of the correct choice:

Two events A and B are said to be independent if

- (a) $P(A \cup B) = P(A) + P(B)$
- (b) $P(B|A) = \frac{P(A \cap B)}{P(A)}$
- (c) $P(A \cap B) = P(A)P(B)$
- (d) $A \cap B = \emptyset$

4. (6 points) True or false:

_____ The formula of the sample standard deviation is $\frac{1}{n-1}(\sum x^2 - (1/n)(\sum x)^2)$

_____ The first quartile is the number such that at least 25% of the observations are equal to or fall below it, and at least 75% of the observations are equal to or fall above it.

_____ According to Chebyshev's Rule, there are at least 75% of the observations within 2 standard deviations of the mean.

_____ If H_0 was rejected at $\alpha = 0.025$, then it is also (always) rejected at $\alpha = 0.05$.

_____ Type I error is to accept H_0 when it is false.

_____ The so-called Least Squares method for fitting a multiple regression consists of minimizing the expression $\sum |y - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k)|$ in $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$, where $|\cdot|$ indicates the absolute value.

_____ The prediction interval is always wider than the associated confidence interval.

_____ To test for the overall usefulness of a multiple linear regression model, it suffices to check whether at least one t -statistic for the individual β_i s is significant.

_____ To test for the overall usefulness of a multiple linear regression model, it suffices to check whether at least one t -statistic for the individual β_i s is significant.

_____ If the Excel output gives a t -statistic of -2.5 for the coefficient β_1 with a p -value of 0.021 , then the one-sided p -value for testing $H_a : \beta_1 > 0$ is $0.021/2 = 0.0105$.

Continued...

5. (8 points) The table below contains the 2006 GDP growth rate (%) of 14 selected countries.

Country	GDP growth rate (%)
Argentina	8.5
Belgium	3.0
Brazil	3.7
Canada	2.8
Chile	4.0
Germany	2.8
Guatemala	4.6
Guyana	4.7
Japan	2.2
Madagascar	4.7
Singapore	7.9
Switzerland	2.7
United Kingdom	2.8
United States	2.9

(source: <https://www.cia.gov/library/publications/the-world-factbook/fields/2003.html>)

Remark: Remember that the GDP (Gross Domestic Product) is simply the income (in \$) generated in a country over the year.

- (a) Make a relative frequency histogram for the data set above using 2-3, 3-4, etc. as class intervals.
- (b) Is the distribution left-skewed, right-skewed, or symmetric? Why? Would you expect this shape for data on GDP growth rates? Why?

Continued...

6. (4 points) A study conducted by the Natural Resources Defense Council revealed that 25% of bottled water is just tap water packaged in a bottle. Consider a sample of five bottles of water, and let X be the (random) number of bottles in the sample that contain just tap water.

(a) What is the probability that X is at most 2? (*hint*: what is the distribution of X ?)

(b) Find the expected value and standard deviation of X .

7. (4 points) If you wish to estimate a population mean within an error margin of 0.3 using a 95% confidence interval, and you know from prior sampling that σ^2 is approximately equal to 7.2, how many observations would have to be included in your sample? (*hint*: if you don't remember the formula, just write out the expression for the error margin and solve it for n)

8. (6 points) The distribution of a random variable X is as follows:

x	$f(x)$
1	0.9
3	0.1

This gives $E(X) = 1.2$ and $sd(X) = 0.6$ (you don't need to verify this).

(a) Give the sampling distribution of \bar{X} for a random sample of size two from the above population.

(b) Using the distribution found in (a), verify that $E(\bar{X}) = \mu$, where μ denotes the mean of X .

(c) Using the distribution found in (a), verify that $sd(\bar{X}) = \frac{\sigma}{\sqrt{2}}$, where σ denotes the standard deviation of X .

Continued...

9. (8 points) A pollster carried out a survey among 40 randomly selected adults in the South with the question: “Which of the four cities you think offers the most vibrant night life?: (1) New Orleans; (2) Houston; (3) Dallas; (4) Atlanta”. The data set thus obtained is shown below:

2 4 3 1 1 3 1 2 3 1 1 3 1 2 1 1 1 1 3 1
1 4 1 3 1 1 3 1 1 1 2 1 4 1 3 1 1 3 2 1

- (a) Find a point estimate of the proportion of adults who prefer New Orleans and the associated 90% confidence interval.
- (b) A past survey indicated that the proportion of adults who prefer New Orleans is 35%. Test the claim that such proportion has changed at 1% level of significance. Formulate the hypotheses, specify the rejection region, perform the test and calculate the p -value.

10. (6 points) Some college professors make bound lecture notes available to their classes in an effort to improve teaching effectiveness. At Helvetia State University, students from two classes were surveyed. One class required the purchase of notes, while the second did not offer lecture notes. At the end of the semester, the students were asked to respond to the statement “Having a copy of the lecture notes was [would be] helpful in understanding the material.” Responses were measured on a 9-point semantic difference scale, where 1 = “strongly disagree” and 9 = “strongly agree”. A summary of the results is reported in the table below.

class buying lecture notes	class <i>not</i> buying lecture notes
$n_1 = 86$	$n_2 = 35$
$\bar{x}_1 = 8.48$	$\bar{x}_2 = 7.80$
$s_1^2 = 0.94$	$s_2^2 = 2.99$

Construct a 99% confidence interval for the difference between mean responses. Interpret the result.

11. (6 points) Independent random samples selected from two normal populations produced the following results

sample 1	sample 2
$n_1 = 17$	$n_2 = 12$
$\bar{x}_1 = 5.4$	$\bar{x}_2 = 7.9$
$s_1^2 = 11.56$	$s_2^2 = 8.94$

Can we conclude that the two population means are different? Answer by performing a hypothesis test at $\alpha = 0.01$, assuming $\sigma_1 = \sigma_2$. State the hypotheses, specify the rejection region and perform the test.

12. (6 points) The data in the table represents typical salaries of technology professionals in 13 metropolitan areas for 2003 and 2005, which can be assumed to be normal. We want to determine whether the mean salary has increased from 2003 to 2005.

Metro area	2003 salary	2005 salary	difference (2003 - 2005)
Silicon Valley	87.7	85.9	1.8
New York	78.6	80.3	-1.7
Washington, D.C.	71.4	77.4	-6.0
Los Angeles	70.8	77.1	-6.3
Denver	73.0	77.1	-4.1
Boston	76.3	80.1	-3.8
Atlanta	73.6	73.2	0.4
Chicago	71.1	73.0	-1.9
Philadelphia	69.5	69.8	-0.3
San Diego	69.0	77.1	-8.1
Seattle	71.0	66.9	4.1
Dallas-Ft. Worth	73.0	71.0	2.0
Detroit	62.3	64.1	-1.8

- (a) Why is a regular two-sample t -test *inappropriate* for comparing the mean salaries in 2003 and 2005? (*hint*: What assumption of the two-sample t -test is violated in the data set above?)
- (b) Let d denote the difference “2003 salary - 2005 salary”. For the data set above, $\sum d = -25.7$, $\sum d^2 = 206.59$ (you do NOT need to verify this). Through a hypothesis test at $\alpha = 0.10$, determine whether the mean salary in the US metropolitan areas has increased between 2003 and 2005. State the hypotheses, specify the rejection region and perform the test.

-
13. (6 points) A marketing professor assessed young children's abilities to recognize cigarette brand advertising symbols. She found that 15 out of 30 children under the age of 6, and 65 out of 92 children age 6 and over recognized Joe Camel, the brand symbol of Camel cigarettes.
- (a) Construct a 99% confidence interval for the difference between the proportions of children in the two age groups that recognize Joe Camel. Indicate the point estimate and the error margin.
- (b) Do the data indicate that older children are more prone to recognize Joe Camel? Answer by performing a hypothesis test at $\alpha = 0.05$. State the hypotheses, specify the rejection region, perform the test and calculate the p -value.

14. (2 points) Circle the letter of the correct choice.

Consider the multiple linear regression

$$y = \beta_0 + \beta_1x_1 + \dots + \beta_5x_5 + \varepsilon,$$

with a data set of size 35. The rejection region for the F test at $\alpha = 0.01$ is of the form $F > \text{"number"}$, where "number" is

- (a) $3.73 = F_{0.01}$ (at 5 and 29 degrees of freedom)
(b) $3.70 = F_{0.01}$ (at 5 and 30 degrees of freedom)
(c) $4.26 = F_{0.005}$ (at 5 and 29 degrees of freedom)
(d) $3.81 = F_{0.005}$ (at 6 and 35 degrees of freedom)

Continued...

15. (8 points) A company wants to examine whether advertising expenditure is in fact related to sales revenue. The following table lists the company's sales revenue in \$10,000 (y) and the associated advertising expenditure in \$1,000 (x).

x	y
1	1
2	1
3	2
4	2
5	4

- (a) Find the least squares estimates of the intercept β_0 and the slope β_1 (show your calculations). Interpret the estimate of the slope.

- (b) Find s (show your calculations).

- (c) Find the coefficient of determination and interpret its value (show your calculations).

- (d) Is there a *positive* linear relation between sales revenue and advertising expenditure? Answer by performing a test at $\alpha = 0.01$. State the hypotheses, specify the rejection region, and perform the test.

16. (6 points) Economic theory suggests that wages and quit rates are related. The next table lists quit rates (quits per 100 employees) and the average hourly wage in a sample of 15 manufacturing industries, where each row represents an industry. Consider the simple linear regression of quit rate y on average wage x .

average wage (x)	quit rate (y)
8.20	1.40
10.35	0.70
6.18	2.60
5.37	3.40
9.94	1.70
9.11	1.70
10.59	1.00
13.29	0.50
7.99	2.00
5.54	3.80
7.50	2.30
6.43	1.90
8.83	1.40
10.93	1.80
8.80	2.00

For the data set above, we have

$$SS_{xx} = 68.70, \quad SS_{yy} = 11.32, \quad SS_{xy} = -23.81, \quad \bar{x} = 8.60, \quad \bar{y} = 1.88$$

(you do NOT have to verify this).

- (a) Find a 95% prediction interval for the quit rate in an industry with an average hourly wage of \$9.00. Interpret the result.

- (b) Find the correlation coefficient (show your calculations). Interpret its value.

Continued...

17. (5 points) Obelix works for weather.com, and he was asked to calculate the predicted annual precipitation in California for a city at altitude 150 ft, latitude 40° and lying at 98 miles from the coast. However, after he ran the model in Excel, his hard drive crashed, and he lost not only the data set, but also an entry in the output:

Confidence and Prediction Estimate Intervals

Data	
Confidence Level	99%
Altitude given value	150
Latitude given value	40
Distance given value	98
t Statistic	2.778714523
Predicted Y (YHat)	?
For Average Predicted Y (YHat)	
Interval Half Width	10.64306588
Confidence Interval Lower Limit	11.6562109
Confidence Interval Upper Limit	32.94234265
For Individual Response Y	
Interval Half Width	32.62308099
Prediction Interval Lower Limit	-10.32380422
Prediction Interval Upper Limit	54.92235777

- (a) Indicate the 99% confidence and prediction intervals for average and predicted annual precipitation, respectively, for a city at altitude 150 ft, latitude 40° and lying at 98 miles from the coast.
- (b) Find the predicted annual precipitation for a city at altitude 150 ft, latitude 40° and lying at 98 miles from the coast.

18. (5 points) Suppose we are willing to model the price y (in dollars) of a case of Bordeaux wine as a linear regression on all or some of the following explanatory variables: x_1 = vintage year, x_2 = growing season temperature ($^{\circ}\text{C}$), x_3 = Sep./Aug. rainfall (cm), x_4 = rainfall in months preceding vintage (cm), x_5 = average Sep. temperature ($^{\circ}\text{C}$). Consider the following computer outputs for three different models (the symbol “-” indicates that the related explanatory variable is not included).

	x	$\hat{\beta}$	SE	t	p-value	
Model 1	x_1	0.035	0.014	2.584	0.015	$r^2 = 0.212, \quad s = 0.575$
	x_2	-	-	-	-	
	x_3	-	-	-	-	
	x_4	-	-	-	-	
	x_5	-	-	-	-	

	x	$\hat{\beta}$	SE	t	p-value	
Model 2	x_1	0.024	0.007	3.319	0.003	$r^2 = 0.828, \quad s = 0.287$
	x_2	0.616	0.095	6.471	0.000	
	x_3	-0.004	0.001	-4.765	0.000	
	x_4	0.000	0.000	0.243	0.810	
	x_5	-	-	-	-	

	x	$\hat{\beta}$	SE	t	p-value	
Model 3	x_1	0.024	0.007	3.213	0.004	$r^2 = 0.828, \quad s = 0.293$
	x_2	0.608	0.116	5.241	0.000	
	x_3	-0.004	0.001	-4.000	0.001	
	x_4	0.001	0.001	2.277	0.032	
	x_5	0.008	0.565	0.014	0.989	

Based on the output above, which of the three models would you recommend? (choose one and explain why it is better than the other two)

19. (8 points) A firm's shipping department wants to determine the relationship between the number of labor hours (y) and the explanatory variables: x_1 = thousands of pounds shipped, x_2 = percentage of units shipped by truck, and x_3 = average shipment weight in pounds. The output of the multiple linear regression model is displayed below.

Regression Statistics

Multiple R	?
R Square	?
Adjusted R Square	0.72699907
Standard Error	9.810345853
Observations	20

	df	SS	MS	F	Significance F
Regression	3	5158.313828	1719.437943	17.86561083	2.32332E-05
Residual	16	1539.886172	96.24288576		
Total	19	6698.2			

	Coefficients	Standard Error	t Stat	P-value
Intercept	131.9242521	25.69321439	5.134595076	9.98597E-05
Pounds (x1)	2.72608977	2.275004884	1.198278645	0.24825743
PctShip (x2)	0.047218412	0.093348559	0.505829045	0.6198742
AveWt (x3)	-2.587443905	0.642818185	-4.025156669	0.000978875

- (a) Construct a 95% confidence interval for β_3 (show your calculations).

- (b) Calculate the coefficient of determination (show your calculations).

- (c) Indicate r_a^2 and interpret its value.

- (d) What can we say about the overall usefulness of the model? Justify your answer.
- (e) If shipping department employees are paid \$7.50 per hour, how much less, on average, will it cost the company if the average shipment weight (x_3) increases from a level of 20 to 21? (assume x_1 and x_2 remain unchanged) Justify your answer.