MATH 338

FINAL EXAM

THURSDAY, DECEMBER 15, 2016

Your name: ________________________

Your scores (to be filled in by Dr. Wynne):

Problem 1:	____/10
Problem 2:	____/19
Problem 3:	____/12
Problem 4:	____/9
Problem 5:	____/11
Problem 6:	____/12.5
Problem 7:	____/9
Total:	____/82.5

You have 110 minutes to complete this exam. This exam is closed book and closed notes with the exception of your two sheets of notes (front and back).

For full credit, show all work except for final numerical calculations (which can be done using a scientific calculator).

Problem 1. [1 pt each] Below are the names of a bunch of different hypothesis tests. For each claim in parts A-J, identify a correct test to use to test the claim (some claims may be tested using more than one test). Assume all assumptions for the tests are met. Tests may be used more than once or not at all.

a. One sample t-test
b. Two sample t-test
c. Matched pairs t-test
d. One sample proportion z-test
e. Two sample proportion z-test
f. Slope t-test for linear regression
g. ANOVA test for linear regression
h. One-way ANOVA
i. Chi-square test for independence
j. Chi-square test for goodness of fit

A. The distribution of Starburst colors is 25% red, 25% orange, 25% pink, and 25% yellow.

j

B. Gas costs more, per gallon, at Shell stations than at Mobil stations across the street from them.

c

C. There is no relationship between how much you eat (in calories) and your weight.

f is the best answer but I'd also accept g

D. When people experience pain, they report less pain when they swear than when they don't swear.

c

E. One-third of American adults can't identify the United States on a world map.

d is the best answer but I'd also accept j

F. Men are more likely to drive and text than women are.

e is the best answer but I'd also accept i

G. Undergraduates in every CSUF College, on average, have the same GPA.

h

H. On average, a "gallon of milk" contains less than 1 gallon.

a

I. Knowing your favorite burger chain tells me nothing about whether you prefer pancakes or waffles.

i

J. The more money you have, on average, the more problems you have.

f is the best answer but I'd also accept g, h, or i depending on how you define "money" and "problems"

Problem 2. In a famous meta-analysis, Linus Pauling examined an experiment in which a group of children at ski school were given a pill containing either 1 gram of Vitamin C or a placebo, and were followed to see if they caught a cold. The data is shown in the two-way table below.

	Cold	No Cold	Total
Vitamin C	17	122	139
Placebo	31	109	140
Total	48	231	279

A. [1 pt] Why did the study use a placebo (instead of just not giving some children any pill at all)?

To avoid the placebo effect: everyone thought they were getting Vitamin C, so any effect would be due to Vitamin C and not due to the psychological effect of thinking they're getting a supplement vs. not

B. [2 pt] The study was a <u>randomized</u> and <u>double-blind</u> experiment. Explain what the two underlined terms mean in the context of this study.

Randomized:

Children were randomly assigned to get either Vitamin C or the placebo. There was no systematic reason for some children to get Vitamin C and others to get placebo.

Double-blind:

Neither the children getting the pill, nor the people giving the children the pill, knew whether the pill was Vitamin C or placebo.

C. [7 pt] At the 5% significance level, is there a difference in the rate at which children catch colds when given Vitamin C compared to placebo?

1 pt work in a two-sample z HT framework

2 pts H_0: rate at which children catch colds is the same between Vitamin C and placebo ($p_1 = p_2$) and H_a: there is a difference in the rate at which children catch colds between Vitamin C and placebo ($p_1 \neq p_2$) OR Children catch fewer colds with Vitamin C

1 pt test statistic: $z = ((17/139) - (31/140))/\text{sqrt}((48/279)(231/279)(1/139 + 1/140)) = -2.193$

1 pt p-value between 0.0026 and 0.0456 OR Critical Region: $|z| > 1.96$ (alternatively: p-value between 0.013 and 0.0228, or Critical Region: $z < -1.645$)

1 pt Reject H_0 since $p < \alpha$lpha and/or z is in the rejection region

1 pt We conclude at the 5% level that there is a difference in the rate at which children catch colds

Alternately, get points for chi-square framework with appropriate H_0 and $\chi^2 = 4.81$, still reject H_0

Problem 2 (continued). The table is re-printed below for your convenience.

	Cold	No Cold	Total
Vitamin C	17	122	139
Placebo	31	109	140
Total	48	231	279

D. [2 pt] Under the null hypothesis for the chi-square test of independence, fill in the expected counts in the table below, to two decimal places:

	Cold	No Cold
Vitamin C	23.91	115.09
Placebo	24.09	115.91

(139*48)/279 = 23.91, (139*231)/279 = 115.09

(140*48)/279 = 24.09, (140*231)/279 = 115.91

E. [2 pt] Using the table you constructed in Part D, compute the chi-square test statistic.

X^2 = (17 – 23.91)^2/23.91 + (122-115.09)^2/115.09 + (24.09 – 31)^2/24.09 + (109 – 115.91)^2/115.91

=4.81

In parts F-H, assume that one of the 279 children in the study is chosen at random.

F. [1 pt] What is the probability that the child was given a Vitamin C pill?

139/279 = 0.498

G. [2 pt] What is the probability that the child caught a cold, given that he or she got a Vitamin C pill?

17/139 = 0.122

For parts G and H, 1 pt setup, 1 pt answer

H. [2 pt] What is the probability that the child got a Vitamin C pill, given that he or she caught a cold?

17/48 = 0.354

Problem 3. Colquhoun (2014) claimed that, "if you use *p*=0.05 to suggest that you have made a discovery, you will be wrong at least 30% of the time." He uses the following example.

Suppose you pick a drug at random from 1000 candidate drugs and run a randomized controlled trial to determine whether the drug truly is effective. You perform a hypothesis test that has a significance level of 0.05 and a power of 0.8.

A. [2 pt] Write the null and alternative hypothesis for this hypothesis test, <u>using words only</u> (no math).

H_0: the drug has no effect (is not effective)

H_a: the drug has an effect (is effective)

B. [1 pt] What is the probability of committing a Type I Error on <u>this single</u> hypothesis test?

α = 0.05

C. [1 pt] What is the probability of committing a Type II Error on <u>this single</u> hypothesis test?

B = 1 – power = 0.2

D. [3 pt] If you independently run experiments and hypothesis tests for <u>four</u> drugs, what is the probability of committing <u>at least</u> one Type I Error when performing the four hypothesis tests?

1 pt use binomial formula
2 pts:

$P(X \geq 1) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = 0.1715 + 0.0135 + 0.0005 + 0.0000 = 0.1855$

OR

$P(X \geq 1) = 1 - P(X = 0) = 1 - 0.8145 = 0.1855$

E. [5 pt] 10% of the drugs truly have an effect, and the other 90% are no different from placebo. Find the probability that your randomly chosen drug has an effect, <u>given that you rejected the null hypothesis</u>.

Using Bayes's Rule:

$$P(H_a \text{ is true} | \text{reject } H_0) = \frac{P(\text{reject}|H_a \text{ true})\, P(H_a \text{ true})}{P(\text{reject}|H_a \text{ true})\, P(H_a \text{ true}) + P(\text{reject}|H_0 \text{ true})\, P(H_0 \text{ true})}$$

$$P(H_a \text{ is true} | \text{reject } H_0) = \frac{(0.8)(0.10)}{(0.8)(0.10) + (0.05)(0.9)} = 0.64$$

1 pt use Bayes's Rule and/or tree diagram, 3 pts plug in correctly, 1 pt final answer

Problem 4. In 2009 the LA Times Data Desk started its “Mapping LA” project, which included collecting information about all 114 distinct neighborhoods of the city of Los Angeles. The dataset we will work with in Problems 4-7 includes the 104 neighborhoods that, as of 2008, had at least one public school in the neighborhood. Here is some important information about the variables in the dataset:

- Neighborhood: the name of the neighborhood
- Income: the median income of the neighborhood’s residents, in thousands of dollars
- Schools: the median API score of schools in the neighborhood, measured on a scale from 200 (bad) to 1000 (good)
- Asian, Black, Latino, White: the percentage of residents with each of those ethnicities (i.e., 5% Asian corresponds to Asian = 5)
- Diversity: how likely neighborhood residents are to encounter a resident of a different ethnicity, measured on a scale from 1 (not diverse at all) to 10 (very diverse)
- Population: the neighborhood’s population, according to the 2000 U.S. Census, in thousands

A. [1 pt] How many cases are in this dataset?

104

B. [1 pt] Which variable is the label variable?

neighborhood

C. [1 pt] Is population a quantitative categorical variable (circle one)?

quantitative

D. [1 pt] Was this dataset collected from an experiment observational study (circle one)?

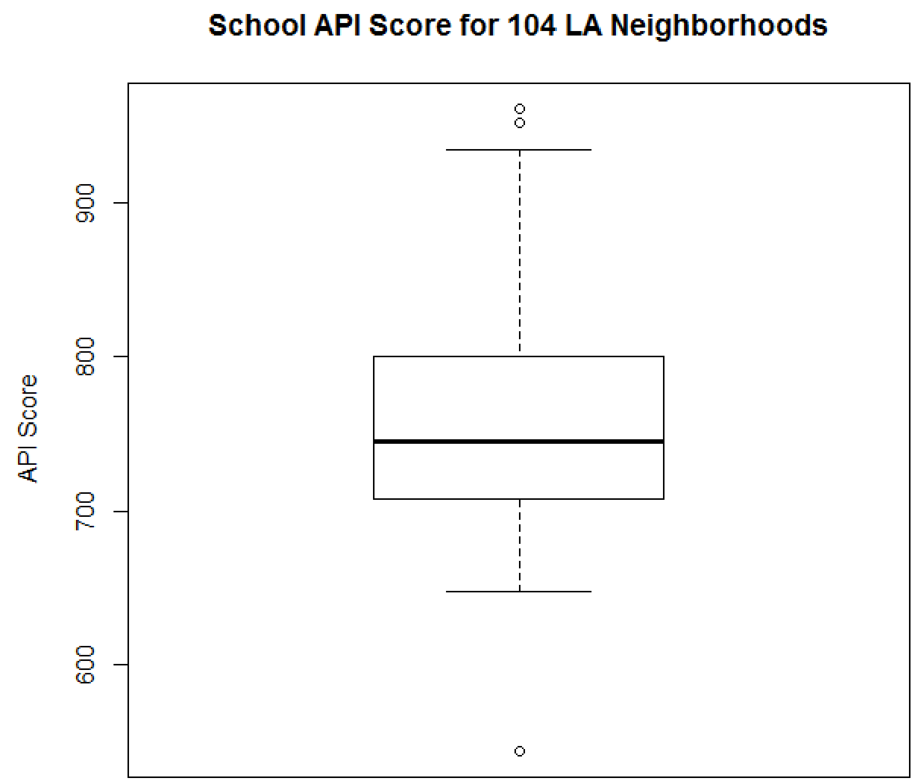
Observational study

E. [1 pt] Which of the following plots would be useful for identifying the distribution of Income:

histogram pie chart scatter plot (circle one)?

histogram

Problem 4 (continued). The boxplot below plots the variable Schools. Use it to answer parts F-I.



F. [1 pt] Which of these values is closest to the median of the variable Schools:

400 650 700 750 800 (circle one)?

750

G. [1 pt] Which of these values is closest to the interquartile range of the variable Schools:

50 100 300 400 750 (circle one)?

800 – 700 = 100

H. [1 pt] How many outliers are there for the variable Schools?

3 circles = 3 outliers

I. [1 pt] Is the distribution of the variable Schools most likely:

skewed left skewed right symmetric (circle one)?

Skewed right

In Problems 5, 6, and 7, we will make some attempts at predicting how good a neighborhood's schools are (median API) from the other variables in the dataset.

Problem 5. Our first attempt is a multiple linear regression model using all five of the variables related to ethnicity (Asian, Black, Latino, White, and Diversity). Use the ANOVA table for the model (below) to answer parts A-D.

ANOVA Table

Source	DF	Sum of Squares	Mean Squares	F Value	Pr > F
Regression	5	404174	80834.7	41.9460	6.82533e-23
Residual	98	188857	1927.11		
Total	103	593031			

A. [2 pt] State the null and alternative hypothesis for the ANOVA F Test for this multiple linear regression model.

H_0: $\beta_{Asian} = \beta_{Black} = \beta_{Latino} = \beta_{White} = \beta_{Diversity} = 0$, none of the predictors have a linear relationship with median School API

H_a: at least one of the β_j's is not zero; at least one predictor has a linear relationship with School API

B. [1 pt] Find the F statistic and the p-value from the output above, and report their values below.

$F = 41.9460$, $p = 6.825 \times 10^{-23}$

C. [1 pt] Based on your answer to part B, is the overall model significant at the 1% level?

Yes, since $p < 0.01$

D. [2 pt] Find the squared multiple correlation coefficient (R^2) for this multiple linear regression model. Round your answer to 2 decimal places.

1 pt: $R^2 = SSM/SST$, 1 pt: $SSM/SST = 404174/593031 = 0.68$

Problem 5 (continued). Use the Parameter Estimates table for the model (below) to answer parts E-G.

Parameter Estimates

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
(Intercept)	486.570	396.456	1.22730	0.222651
Diversity	-3.94718	1.98950	-1.98401	0.0500517
Asian	6.05296	4.08813	1.48062	0.141916
Black	2.04206	4.01382	0.508757	0.612066
Latino	1.93060	3.97734	0.485401	0.628476
White	3.90155	4.12266	0.946368	0.346289

E. [1 pt] If we performed backward selection, which variable would we remove from this model?

The least significant predictor is Latino, so we would remove Latino

F. [2 pt] Interpret the value 6.05296 in the Parameter Estimates table.

1 pt: For every 1% increase in the Asian percentage of a neighborhood, the median school API is expected to increase by 6.05296 points…

1 pt: …holding the values of the other explanatory variables (Diversity, Black, Latino, and White) constant

G. [2 pt] Do you have any concerns about the choice of explanatory variables in this model? If so, what are they? If not, why not?

2 pts for anything discussing (1) collinearity among the chosen explanatory variables and/or (2) the choice not to put other reasonable explanatory variables (e.g. income) in the model

Problem 6. Our next attempt is a simple linear regression model using Income (in thousands of dollars) as the only predictor. Use the Parameter Estimates table (below) to answer all parts.

Parameter Estimates

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
(Intercept)	655.641	11.9676	54.7847	1.81476e-77
Income	1.81123	0.188148	9.62660	5.46340e-16

A. [1 pt] Write the equation used by this model to predict median school API from Income.

(predicted) Median API = 655.641 + 1.81123 (Income)

B. [1.5 pt] Name three plots that we should look at to determine whether model assumptions are met.

0.5 pts each for 3 of: scatter plot, residual plot, normal q-q plot, scale-location plot

C. [3 pt] Construct (don't interpret) a 95% confidence interval for the population slope of the model.

1 pt: CI = POINT ESTIMATE ± CRITICAL VALUE * STANDARD ERROR

2 pts: CI = 1.81123 ± (1.984)(0.188148) = (1.438, 2.185)

D. [2 pt] Predict the median school API for a neighborhood in which the median income is $40,000.

Predicted API = 655.641 + 1.81123 (40) = 728.0902 or about 728

1 pt for recognizing to plug in 40 (not 40,000) and 1 pt for plugging in correctly and getting answer

E. [5 pt] Income has a mean of 56.7312 and a variance of 835.483. The residual standard error (s) is 55.1934. Given these values and your answer to part D, construct <u>and interpret</u> a 90% prediction interval for the median school API in a neighborhood whose median income is $40,000.

$$PI = \hat{y} \pm t^*(s)\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)Var(x)}}$$

$$= 728.0902 \pm (1.660)(55.1934)\sqrt{1 + \frac{1}{104} + \frac{(40 - 56.7312)^2}{(103)(835.483)}}$$

$$= 728.0902 \pm 92.20867 = (635.88, 820.30)$$

We are 90% confident that a neighborhood with median income of $40,000 will have a median school API between about 636 and 820. 1 pt use PI formula, 2 pt plug in correctly, 1 pt final values, 1 pt interpretation

Problem 7. On our final attempt, we divide the neighborhoods into two categories: "mostly white" (White >= 50) and "mostly minority" (White < 50). Here is a summary of the data:

Group	Number of Neighborhoods	Mean(Schools)	sd(Schools)
Mostly Minority	70	725.53	54.60
Mostly White	34	826.06	68.77

A. [7 pt] Using a two-sample t test, test the claim that there is no difference in school performance between mostly minority and mostly white neighborhoods. Use a significance level of 1%.

2 pt H_0: no difference in mean school performance between the two types of neighborhoods ($\mu_1 = \mu_2$), H_a: difference in school performance between the two types of neighborhoods ($\mu_1 \neq \mu_2$)

1 pt test statistic t = (725.53 – 826.06)/sqrt(54.60^2/70 + 68.77^2/34) = -7.46

OR use pooled variance, s_p^2 = ((69)(54.60^2) + (33)(68.77)^2)/(70 + 34 – 2) = 3546.74 so s_p = 59.55 and

t = (725.53 – 826.06)/(59.55*sqrt(1/70 + 1/34)) = -8.08

1 pt min(70 – 1, 34 – 1) = 33 degrees of freedom; 1 pt t* is about 2.75

1 pt critical region is $|t| > 2.75$; 1 pt since $|t_{obs}| > 2.75$, we reject H_0

1 pt At the 1% significance level, we have significant evidence that school performance differs between mostly white neighborhoods and mostly minority neighborhoods.

B. [2 pt] Name one assumption necessary for the two-sample t test that is violated in this analysis. Justify why the assumption does not hold.

The assumption that is violated most clearly is the assumption of simple random samples. You could argue that either we have the entire populations here and do not need to do inference, or that restricting the sample to Los Angeles makes it a non-random sample of neighborhoods in some larger population.

You could also argue that the assumption of independence of the samples is violated because some students in majority white neighborhoods attend school in mostly minority neighborhoods (and vice versa), or because the neighborhoods share boundaries, or any number of other plausible reasons.

Finally, you could argue that we should be working with the population of <u>schools</u> and not <u>neighborhoods</u>, so our analysis of median school API doesn't make any sense in the first place

Extra Space. The tables below show a number of critical values *z* for the standard normal variable $Z \sim N(0,1)$ and the corresponding cumulative proportions, corresponding to P(Z ≤ z).

z-score	Cumulative Proportion
-3.00	0.0013
-2.00	0.0228
-1.65	0.0495
-1.28	0.1003
-1.00	0.1587
-0.43	0.3336

z-score	Cumulative Proportion
0.43	0.6664
1.00	0.8413
1.28	0.8997
1.65	0.9505
2.00	0.9772
3.00	0.9987

Refer to the following tables for t* and z* critical values for confidence and prediction intervals:

Degrees of freedom	C = 0.90 (90%)	C = 0.95 (95%)	C = 0.98 (98%)	C = 0.99 (99%)
1	6.314	12.71	31.82	63.66
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
≈20	1.725	2.086	2.528	2.845
≈30	1.697	2.042	2.457	2.750
≈50	1.676	2.009	2.403	2.678
≈70	1.667	1.994	2.381	2.648
≈100	1.660	1.984	2.364	2.626
≈1000	1.646	1.962	2.330	2.581

	C = 0.90 (90%)	C = 0.95 (95%)	C = 0.98 (98%)	C = 0.99 (99%)
z* values	1.645	1.960	2.326	2.576

For a two-sided hypothesis test, use the column corresponding to C = 1 – α

For a one-sided hypothesis test, use the column corresponding to C = 1 - 2α

Refer to the following table for χ^2 critical values:

Degrees of freedom	α = 0.05	α = 0.01
1	3.84	6.63
2	5.99	9.21
3	7.81	11.34
4	9.49	13.28