MATH 338

FINAL EXAM

MON/THURS, MAY 15/18, 2017

Your name: ________________________

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

Problem 1:	____/10
Problem 2:	____/12
Problem 3:	____/17
Problem 4:	____/12
Problem 5:	____/13
Problem 6:	____/9
Problem 7:	____/9.5
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 independent samples 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. After accounting for petal length, there is a linear relationship between sepal length and sepal width.

B. When people “guess a number between 1 and 10,” every integer has equal chance of being picked.

C. There is no association between hand size and shoe size.

D. Freshmen, sophomores, juniors, and seniors spend the same amount of time on homework per week.

E. Once they enter, men and women are equally likely to graduate from college.

F. People take longer road trips in hybrid cars than they do in gasoline-powered cars.

G. A model predicting your college GPA from your high school GPA and SAT scores is, overall, significant.

H. Multiple-choice questions are more likely to be unanswered at the end of the test than the beginning.

I. People are willing to spend more on the same drug when they are told it is a brand-name (vs. generic).

J. The radar gun that caught you doing 55 in a 54, on average, overestimates speeds by at least 1 mph.

Problem 2. A 2016 study looked at the amount of added sugar in Canadian kids' meals. The table below summarizes their findings for the grams of added sugar in kids' beverages and kids' desserts:

Type	n	Mean	SD	Min	Q1	Median	Q3	Max
Beverage	33	16	20	0	0	11	28	73
Dessert	35	12	7	0	8	14	16	30

A. [1 pt] The distribution of grams of added sugar in kids' beverages is most likely (circle one):

skewed left symmetric skewed right

B. [1 pt] How many beverages in the sample contain 0 grams of added sugar (circle the correct range)?

8 or fewer 9-16 17-24 25 or more

C. [2 pts] Is the beverage with 73 grams of added sugar an outlier? Justify your answer mathematically.

D. [7 pts] Is there a difference in the mean amount of added sugar between beverages and desserts? Perform a statistical test to support your answer. Assume the assumptions of the test are met.

E. [1 pt] Based on the sample data, which has more grams of added sugar, an "average" beverage or an "average" dessert? Justify your answer mathematically.

Problem 3. In one of the largest and most famous public health experiments ever conducted, in 1954 a randomized controlled trial was run to see whether a vaccine developed Dr. Jonas Salk was effective in preventing paralytic polio. A total of 401,974 children, chosen to be representative of those who might be susceptible to the disease, were randomized to two groups: 200,745 children were injected with a harmless saline solution and the other 201,229 children were injected with Salk's vaccine.

A. [2 pts] What was the point of giving saline solution to the children who didn't get the vaccine?

B. [2 pts] Would it have been possible to run this experiment in a double-blind fashion? Would it have been a good idea to do so? Explain your answers briefly.

C. [7 pts] The results of the trial were as follows: 33 of the 201,229 children who got the vaccine later developed paralytic polio, whereas 115 of the other 200,745 children developed paralytic polio. Perform an appropriate statistical hypothesis test for these data to draw conclusions about the effectiveness of the Salk vaccine.

Problem 3 (Continued).

D. [3 pts] Describe in context what the Type I and Type II errors are in for this scenario (do not perform any computations). What would you argue is the more detrimental error to commit? Defend your answer.

E. [3 pts] Assess the assumptions needed to perform your statistical hypothesis test in part (C). Write a few sentences commenting on the validity of your conclusions in part (C) in terms of these assumptions and sample size.

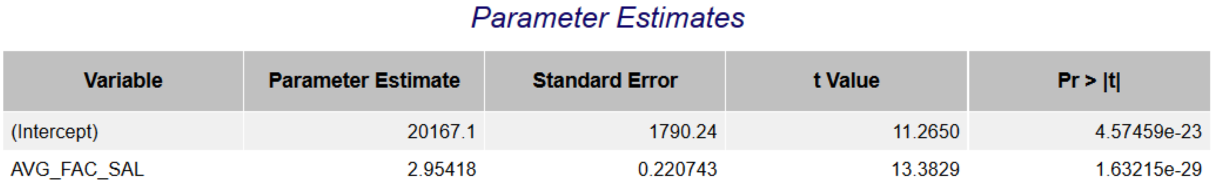
Problem 4. Acceptance sampling is a method by which manufacturers decide whether a lot of products (either produced or supplied) can be considered to conform to specifications.

A. [4 pts] Suppose that the lengths of widgets are normally distributed with mean 10 cm and standard deviation 0.2 cm. A widget conforms to specifications if it has a length between 9.8 and 10.5 cm. What is the probability that a single randomly selected widget conforms to specifications?

B. [3 pts] Suppose that 4% of the widgets in a very large lot do not conform to specifications. What is the probability of obtaining at least 1 nonconforming widget, if you randomly sample 20 widgets?

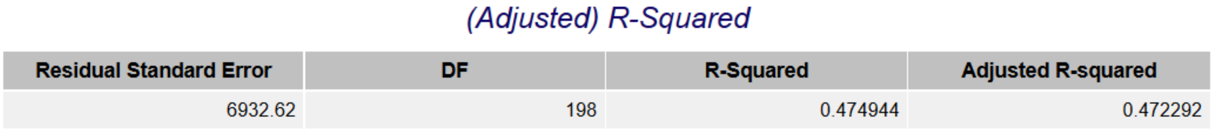
C. [5 pts] Suppose that your acceptance sampling plan will reject 5% of "good" lots, but accept 10% of "bad" lots. 2% of lots from your supplier are "bad." What is the probability that the lot is "bad," given that you reject the lot?

Problem 5. I took a simple random sample of 200 predominantly Bachelor's degree-awarding colleges and universities in the United States. The Rguroo output shown below is from an analysis predicting median earnings of a college's graduates (10 years after entering) from the average full-time faculty monthly salary (AVG_FAC_SAL).



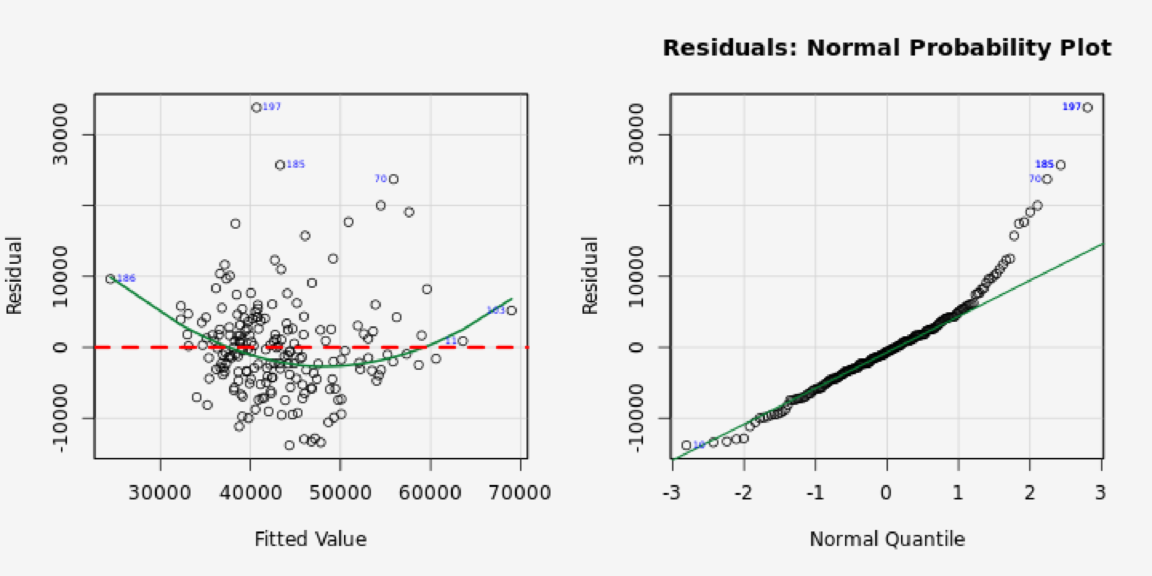
Parameter Estimates

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
(Intercept)	20167.1	1790.24	11.2650	4.57459e-23
AVG_FAC_SAL	2.95418	0.220743	13.3829	1.63215e-29



(Adjusted) R-Squared

Residual Standard Error	DF	R-Squared	Adjusted R-squared
6932.62	198	0.474944	0.472292



A. [1 pt] Is there a positive relationship between the two variables (circle one)? Yes No

B. [1 pt] Is there a causal relationship between the two variables (circle one)? Yes No

C. [1 pt] Based on the output, the correlation between the variables is closest to (circle one):

-0.75 -0.5 -0.25 0 0.25 0.5 0.75

Problem 5 (Continued).

D. [1 pt] In an ANOVA table for this regression model, MSE would be closest to (circle one):

2000 7000 3 million 50 million 400 million

E. [2 pts] Predict the median earnings of graduates of a college that pays its full-time faculty an average of $6000 per month. (Just give a single value)

The summary of AVG_FAC_SAL below may be useful in Parts F-H:

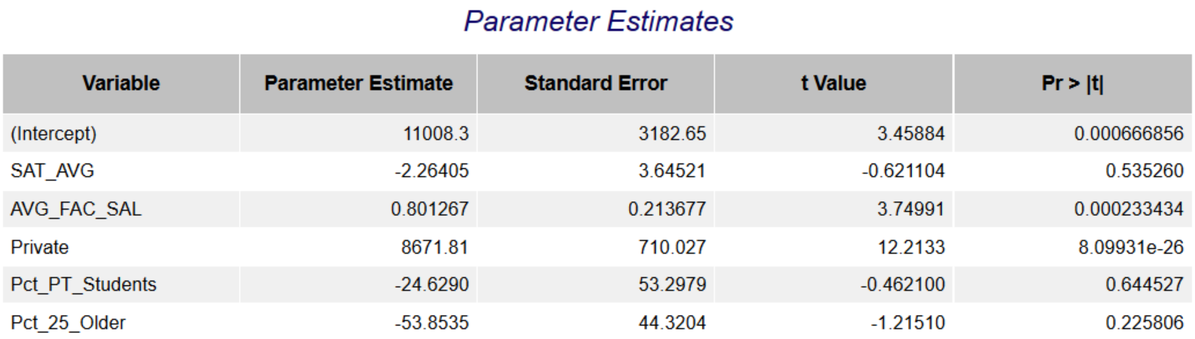
Min: 1451 Max: 16529 Mean: 7800.09 SD: 2226.30

F. [1 pt] The prediction in part (E) was an example of (circle one): interpolation extrapolation

G. [4 pts] Construct, **<u>but do not interpret</u>**, a 95% prediction interval for the median earnings of graduates of a college that pays its full-time faculty an average of $6000 per month.

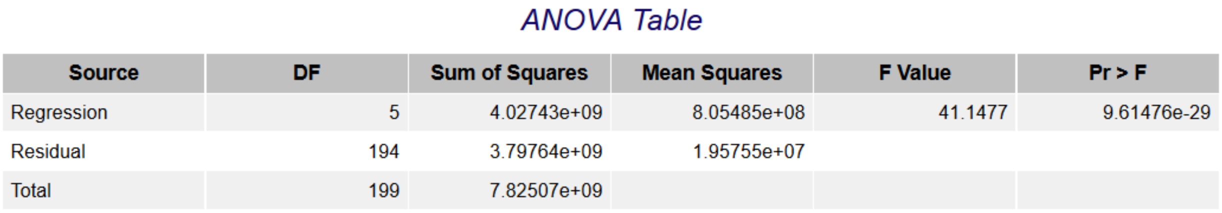
H. [2 pts] Does the 95% prediction interval you calculated in part (G) have valid real-world meaning? If so, interpret it. If not, explain why.

Problem 6. The multiple linear regression analysis output below predicts the average yearly net price students paid to attend college from five variables: average SAT score (SAT_Avg), average full-time faculty monthly salary (AVG_FAC_SAL), whether the school is Private (0 = Public, 1 = Private), percent of part-time students (Pct_PT_Students), and percent of students 25 or older (Pct_25_Older).



Parameter Estimates

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
(Intercept)	11008.3	3182.65	3.45884	0.000666856
SAT_AVG	-2.26405	3.64521	-0.621104	0.535260
AVG_FAC_SAL	0.801267	0.213677	3.74991	0.000233434
Private	8671.81	710.027	12.2133	8.09931e-26
Pct_PT_Students	-24.6290	53.2979	-0.462100	0.644527
Pct_25_Older	-53.8535	44.3204	-1.21510	0.225806



ANOVA Table

Source	DF	Sum of Squares	Mean Squares	F Value	Pr > F
Regression	5	4.02743e+09	8.05485e+08	41.1477	9.61476e-29
Residual	194	3.79764e+09	1.95755e+07		
Total	199	7.82507e+09			

A. [2 pts] How much of the variance in average yearly net price is explained by this model?

B. [2 pts] If this was our initial model and we used a backward selection algorithm, circle all explanatory variables that would be in our next model:

SAT_AVG AVG_FAC_SAL Private Pct_PT_Students Pct_25_Older

C. [5 pts] Construct and interpret a 95% confidence interval for the population slope corresponding to the variable Private. What can you say about the average yearly net price at public vs. private colleges?

Problem 7. In lecture, we explored part of a data set looking at wedding announcements in the *New York Times*. The graphs on the next page explore another aspect of the data set: what last name does the bride use after she marries? I divided that variable into four categories:

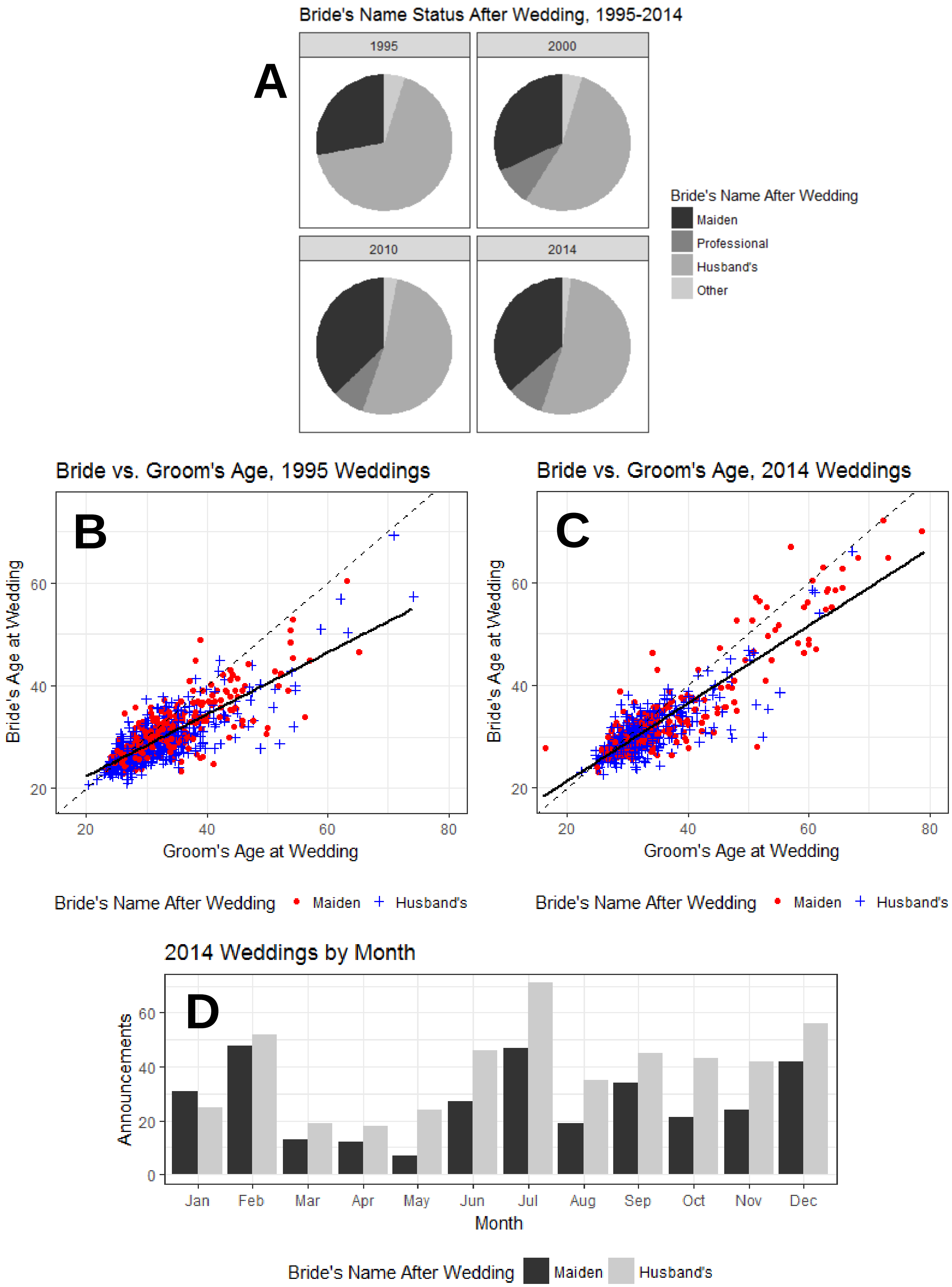
- Maiden: Bride uses her maiden (birth) name both professionally and socially
- Professional: Bride uses her maiden name professionally, but her husband's last name socially (for example, she goes by "Mrs. Smith" around town but still writes as "Ms. Jones")
- Husband's: Bride uses her husband's last name both professionally and socially
- Other: Something else, such as hyphenating her last name, combining both last names, etc.

You can rip off the page of graphs to help you answer the following questions.

A. [2 pts] There are four panels, labeled A, B, C, and D. For each of the following plot types, identify which panel or panels contain that type of plot. If no panel contains that type of plot, write "None."

Bar plot: ______ Box plot: ______ Scatterplot: _____ Histogram: ______

B. [7.5 pts] Tell me some interesting things about these weddings. Make at least three claims about these weddings, support them with evidence from one or more of the panels, and suggest some "next steps" to take to further investigate each claim. Keep in mind that weddings announced in the *New York Times* may not be representative of all weddings around the world, or even in the United States.


Bride's Name Status After Wedding, 1995-2014
A
1995
2000
2010
2014
Bride's Name After Wedding
Maiden
Professional
Husband's
Other
Bride vs. Groom's Age, 1995 Weddings
B
Bride's Age at Wedding
60
40
20
20
40
60
80
Groom's Age at Wedding
Bride's Name After Wedding
Maiden
Husband's
Bride vs. Groom's Age, 2014 Weddings
C
Bride's Age at Wedding
60
40
20
20
40
60
80
Groom's Age at Wedding
Bride's Name After Wedding
Maiden
Husband's
2014 Weddings by Month
D
Announcements
60
40
20
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month
Bride's Name After Wedding
Maiden
Husband's

Extra Space. The tables below show a number of 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
-2.50	0.0062
-2.00	0.0228
-1.65	0.0495
-1.28	0.1003
-1.00	0.1587
-0.67	0.2514

z-score	Cumulative Proportion
0.67	0.7486
1.00	0.8413
1.28	0.8997
1.65	0.9505
2.00	0.9772
2.50	0.9938
3.00	0.9987

Refer to the following tables for t* and z* critical values for confidence 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
19	1.729	2.093	2.539	2.861
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
≈200	1.653	1.972	2.345	2.601
≈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