Statistics 1040 Fall 2010 Final Exam

Do not open the exam until you are instructed to do so.

Directions: You have 110 minutes to complete the exam, which has 200 total possible points. Be sure to answer every question. Where numeric answers are required, you must show your work to receive credit. The formulas and the tables on the last few pages of the exam may be useful for some of the questions, and you may detach the last page if you choose. (Don't detach the normal table page.) Where calculations are required, round to two decimal places.

Student Name: Solutions

Under your recitation leader's name, circle your recitation time:

Oleksandr Gromenko	Bryan Stephenson	Anthonie Nichols
10:30	8:30	8:30
11:30	12:30	10:30
12:30	1:30	11:30
		1:30

1. A psychology professor recently hypothesized that people who are reminded of their mortality (i.e., that they will someday die) would be more likely to want to namesake a child (i.e., name a child after themselves). To test this, she had the 116 students in her large introductory psychology class participate in a study. The professor assigned the 65 freshmen in the class to write an essay on their emotions regarding their own death, and the 47 non-freshmen in the class to write an essay on their emotions regarding possibly failing an exam. After submitting their essays, the students each answered a question about whether they would be likely to namesake a future child. Those who had thought about their own death were significantly more likely to want to namesake a future child.

[+2 if only say there was a control group or something about confounding]

   (a) (8 points) Was this study observational or controlled? Choose one and explain briefly.

   $\underbrace{\text{controlled}}_{+4}$ - $\underbrace{\text{the researcher assigned participants to groups}}_{+4}$

   (b) (8 points) Was this study randomized? Answer yes or no, and explain briefly.

   $\underbrace{\text{No}}_{+4}$ - $\underbrace{\text{assignment not made randomly}}_{+4}$

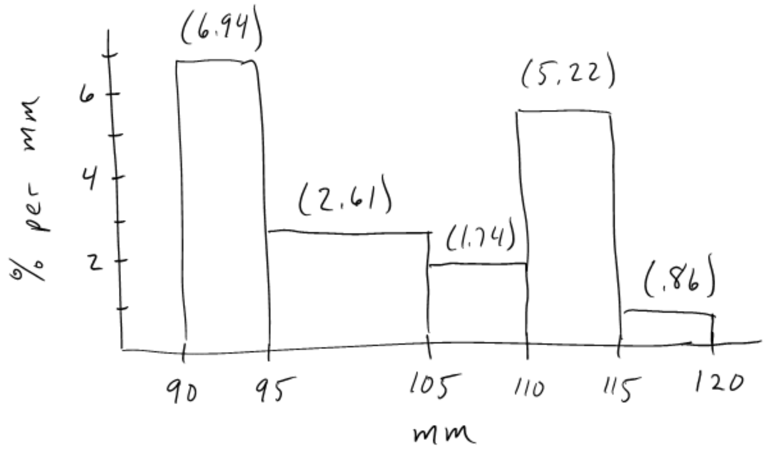
   [+2 if selection to participate instead of assignment to groups]

2. (18 points) The systolic blood pressure was recorded for 23 females, and is summarized in the table below. Class intervals include the left but not the right endpoints, and blood pressure is measured in millimeters (mm). Sketch a histogram to summarize these data, being sure to clearly indicate rectangle heights, and label both axes correctly.

Blood Pressure	Number of Females	%	h.u.	% per h.u.
90-95	8	34.7	5	6.94
95-105	6	26.1	10	2.61
105-110	2	8.7	5	1.74
110-115	6	26.1	5	5.22
115-120	1	4.3	5	.86
	23			

Partial credit:

+2 for horiz. units
+9 for correct heights (% ÷ h.u.)
+2 for rectangles
+3 for axis labels
+2 for heights written
(-6 for % × h.u.)



3. In a well-shuffled deck of 52 cards, there are 4 suits (clubs ♣, spades ♠, diamonds ♢, and hearts ♡), and each suit has 13 cards. Clubs and spades are black, while diamonds and hearts are red. Suppose that 2 cards are to be drawn at random, without replacement.

   (a) (10 points) What is the probability of drawing a club on the first and a red card on the second?

   $P(\text{Club } \#1 \text{ and Red } \#2) = P(\text{Club}\#1) \times P(\text{Red } \#2 \mid \text{Club}\#1)$

   [+5 for mult. rule of correct events
   +5 for correct #'s (-2 each)] $= \frac{13}{52} \times \frac{26}{51}$

   (b) (6 points) Let event A be that the first card drawn is a club, and event B be that the second card drawn is red. Then events A and B are (circle one):

   i. independent
   ii. mutually exclusive
   iii. neither independent nor mutually exclusive [no partial credit]

4. In a large study of women aged 25-34, the relationship between height and income was roughly linear (and football-shaped) and can be summarized as follows:

   x : average height = 64 inches SD = 2.5 inches
   y : average income = \$21,000 SD = \$20,000 r = 0.25

   (a) (20 points) Write the equation of the line to predict income from height, in the form $y = mx + b$.

   $$m = \underbrace{r \cdot \frac{SD_y}{SD_x}}_{+3 \text{ formula}} = \underbrace{0.25 \times \frac{20{,}000}{2.5} = 2{,}000}_{+4 \text{ for } \#\text{'s}}$$

   $$\underbrace{21{,}000 = 2{,}000(64) + b}_{+4 \text{ for pt. of ave.}} \longrightarrow \underbrace{b = -107{,}000}_{+3 \text{ for solving}}$$

   $$\underbrace{\underline{y = 2{,}000x - 107{,}000}}_{+6 \text{ for equation in this form}}$$

   - OR -

   $$y - (21{,}000) = \underbrace{\left(.25 \times \frac{20{,}000}{2.5}\right)}_{+7}(x - 64) \quad \}+5$$

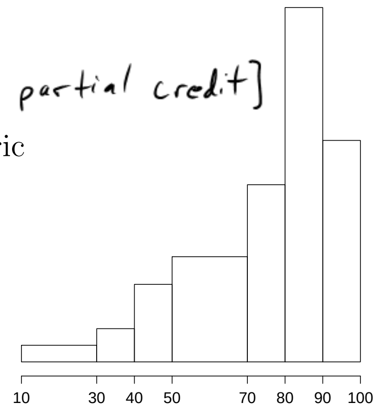
   [then +8 for simplifying to form $y = 2{,}000x - 107{,}000$]

   (b) (3 points) Calculate the predicted income for a woman 67 inches tall.

   $2{,}000(67) - 107{,}000 = \underline{27{,}000}$

   [full points if used equation (a) correctly]

5. The histogram below represents the distribution of the reading scores of 100 4th graders selected as a SRS in a certain school district. The median is 81 points, the average is 74, and the SD is 18.4.



   (a) (6 points) This distribution is (circle one): [no partial credit]

   (i) skewed left (ii) skewed right (iii) symmetric

   (b) (12 points) From this SRS, the 90% confidence interval for the average is calculated as 70.96 to 77.04. Seeing this, a district administrator says that "90% of 4th graders in this district have reading scores between 70.96 and 77.04." Is she correct? If yes, explain why. If no, give a correct interpretation of the confidence interval.

+6 { No

+3 { We can be 90% confident

+3 { that the true average [reading score for all 4th graders in the district] is in this interval.

6. (10 points) [NOTE: This question does not ask you to perform a test of significance. Instead, it asks you to explain a particular step in the χ^2 Goodness of Fit test.] A plant breeder uses a flower-breeding strategy that gives the observed counts below, while a certain geneticist's theory says the breeding strategy should give the expected counts below.

Flower type	# Observed	# Expected
tall yellow	278	256
short yellow	306	328
tall white	223	205
short white	96	114
total:	903	903

A χ^2 Goodness of Fit test results in a test statistic of $\chi^2 = 7.79$, with $df = 3$; the resulting P-value is between between 5% and 10%.

Referring to this example, explain briefly why the degrees of freedom in a χ^2 Goodness of Fit test are "# categories − 1."

If you know the total is 903,

key (+10) { you only need counts for 3 categories (4-1)

before you can get counts for the remaining categories.

7. A vitamin pill is advertised as containing (on average) 500 mg Vitamin C. A consumer protection organization wants to make sure the true average isn't much different than 500 mg. A random sample of 7 pills are assessed for actual Vitamin C content; the following amounts (in mg) are observed:

498 513 521 487 514 515 504

The pills' Vitamin C amounts have an approximately normal distribution. Perform an appropriate test of significance.

(a) (6 points) State the null and alternative hypotheses (in terms of some parameter).

+3 { Null: average is 500

+3 { Alt: average is not 500 (-2 if one-sided)

(b) (15 points) Compute the test statistic.

$$t = \frac{\text{obs} - \text{exp}}{SE} \qquad SE_{\text{ave}} = \frac{SD^+}{\sqrt{\text{size}}}$$

+4 (calc.) { ave = 507.43 , SD = 10.91

+4 { $SD^+ = \sqrt{\frac{7}{6}} \times 10.91 = 11.78$

+4 { $SE_{\text{ave}} = \frac{11.78}{\sqrt{7}} = 4.45$

+3 { $t = \frac{507.43 - 500}{4.45} = 1.67$ (d.f. = 7-1 = 6)

(c) (4 points) Report the P-value (a useful range is okay here).

between 5% & 10%

↳ 10% < P < 20%

(+2 if one-tailed)

(d) (6 points) Report your conclusion.

$\underbrace{\text{Insufficient evidence}}_{\text{key}}$ to say ave. isn't 500

8. (10 points) Complete the following sentence: If you perform a test of significance and obtain a P-value of 2%, this means there is a 2% probability of ...

+4 { observing a result [or test statistic, or event]

+3 { at least as extreme as what was seen

+3 { just by chance when the null hypothesis is true

[+3 if only say 2% < 5% so reject null]

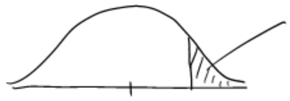
9. According to the U.S. Census Bureau, the average family size for U.S. citizens is 3.20. Researchers suspect that U.S. citizens of Vietnamese ancestry have larger family sizes than that. They take a SRS of 120 families of Vietnamese ancestry; the sample average family size is 3.38, with SD 1.1. Their one-sample z test statistic is $\underline{z = 1.8}$.

(a) (6 points) State the null and alternative hypotheses (in terms of some parameter).

+3 { Null: average is 3.20

+3 { Alt: average is > 3.20 (−2 if two-sided)

(b) (4 points) Report the P-value.



1.8

$\frac{100 - 92.81}{2} = 3.60\%$

(c) (6 points) Report your conclusion.

Average [family size of Vietnamese families] is > 3.20

10. In a randomized controlled double-blind study of 112 recent heart attack patients, 50 are randomly assigned to receive pill "A", while the other 62 receive pill "B." All 112 patients are very diligent in taking their pill each day, and during the next 2 years they are monitored for heart attack recurrence. The following table summarizes the results as category counts, with row and column totals given:

	recurrence	no recurrence	
pill A	10	40	50
pill B	18	44	62
	28	84	112

Perform an appropriate test of the null hypothesis "Heart attack recurrence and pill taken are independent" vs. the alternative hypothesis "Heart attack recurrence and pill taken are not independent." (Note that no conclusion is required below.)

(a) (12 points) Compute the test statistic.

+4 { expected:

12.5	37.5
15.5	46.5

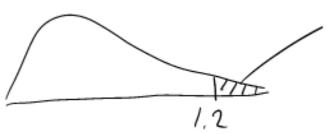
$\frac{(\text{obs} - \text{exp})^2}{\text{exp}}$:

.50	.17
.40	.13

+4

→ sum: $\chi^2 = 1.2$ [d.f. = 1] } +4

(b) (3 points) Report the P-value (a useful range is okay here).



1.2

10% < P < 30%

11. The data from the previous question can be used in another test of percentages. Notice that % recurrence for A is 10/50, or 20%, and % recurrence for B is 18/62, or 29%. Perform an appropriate test of the null hypothesis “Recurrence percentage is the same for both pills” vs. the alternative hypothesis “Recurrence percentage is not the same for both pills.”

(a) (15 points) Compute the test statistic.

+3 SD's, +4 SE's

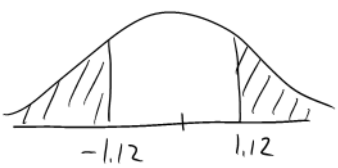
$$\left\{ \begin{array}{llll} A: & SD = \sqrt{.2 \times .8} = .4 & & SE_A = \dfrac{.4}{\sqrt{50}} \times 100\% = 5.66\% \\ B: & SD = \sqrt{.29 \times .71} = .45 & & SE_B = \dfrac{.45}{\sqrt{62}} \times 100\% = 5.72\% \end{array} \right.$$

$$+5 \left\{ \; SE_{A-B} = \sqrt{5.66^2 + 5.72^2} = 8.05 \right.$$

$$+3 \left\{ \; z = \frac{(20-29) - 0}{8.05} = \underline{-1.12} \right.$$

(positive okay)

(b) (4 points) Report the P-value.



$P = 100\% - 72.87\% = 27.13\%$

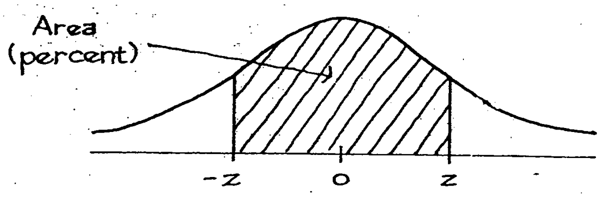
(c) (6 points) Report your conclusion.

Insufficient evidence to say the percentages aren't the same.

12. (2 points) What topic(s) did you study most for this course that did not appear on this final exam?

(Anything)

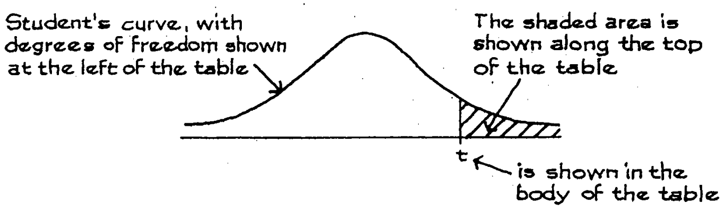
Tables



A NORMAL TABLE

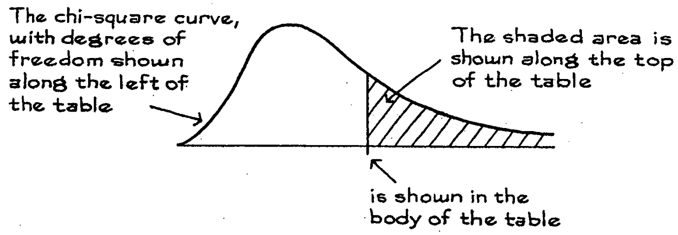
z	*Area*	*z*	*Area*	*z*	*Area*
0.00	0	1.50	86.64	3.00	99.730
0.05	3.99	1.55	87.89	3.05	99.771
0.10	7.97	1.60	89.04	3.10	99.806
0.15	11.92	1.65	90.11	3.15	99.837
0.20	15.85	1.70	91.09	3.20	99.863
0.25	19.74	1.75	91.99	3.25	99.885
0.30	23.58	1.80	92.81	3.30	99.903
0.35	27.37	1.85	93.57	3.35	99.919
0.40	31.08	1.90	94.26	3.40	99.933
0.45	34.73	1.95	94.88	3.45	99.944
0.50	38.29	2.00	95.45	3.50	99.953
0.55	41.77	2.05	95.96	3.55	99.961
0.60	45.15	2.10	96.43	3.60	99.968
0.65	48.43	2.15	96.84	3.65	99.974
0.70	51.61	2.20	97.22	3.70	99.978
0.75	54.67	2.25	97.56	3.75	99.982
0.80	57.63	2.30	97.86	3.80	99.986
0.85	60.47	2.35	98.12	3.85	99.988
0.90	63.19	2.40	98.36	3.90	99.990
0.95	65.79	2.45	98.57	3.95	99.992
1.00	68.27	2.50	98.76	4.00	99.9937
1.05	70.63	2.55	98.92	4.05	99.9949
1.10	72.87	2.60	99.07	4.10	99.9959
1.15	74.99	2.65	99.20	4.15	99.9967
1.20	76.99	2.70	99.31	4.20	99.9973
1.25	78.87	2.75	99.40	4.25	99.9979
1.30	80.64	2.80	99.49	4.30	99.9983
1.35	82.30	2.85	99.56	4.35	99.9986
1.40	83.85	2.90	99.63	4.40	99.9989
1.45	85.29	2.95	99.68	4.45	99.9991

A *t*-TABLE



Degrees of freedom	25%	10%	5%	2.5%	1%	0.5%
1	1.00	3.08	6.31	12.71	31.82	63.66
2	0.82	1.89	2.92	4.30	6.96	9.92
3	0.76	1.64	2.35	3.18	4.54	5.84
4	0.74	1.53	2.13	2.78	3.75	4.60
5	0.73	1.48	2.02	2.57	3.36	4.03
6	0.72	1.44	1.94	2.45	3.14	3.71
7	0.71	1.41	1.89	2.36	3.00	3.50
8	0.71	1.40	1.86	2.31	2.90	3.36
9	0.70	1.38	1.83	2.26	2.82	3.25
10	0.70	1.37	1.81	2.23	2.76	3.17
11	0.70	1.36	1.80	2.20	2.72	3.11
12	0.70	1.36	1.78	2.18	2.68	3.05
13	0.69	1.35	1.77	2.16	2.65	3.01
14	0.69	1.35	1.76	2.14	2.62	2.98
15	0.69	1.34	1.75	2.13	2.60	2.95
16	0.69	1.34	1.75	2.12	2.58	2.92
17	0.69	1.33	1.74	2.11	2.57	2.90
18	0.69	1.33	1.73	2.10	2.55	2.88
19	0.69	1.33	1.73	2.09	2.54	2.86
20	0.69	1.33	1.72	2.09	2.53	2.85
21	0.69	1.32	1.72	2.08	2.52	2.83
22	0.69	1.32	1.72	2.07	2.51	2.82
23	0.69	1.32	1.71	2.07	2.50	2.81
24	0.68	1.32	1.71	2.06	2.49	2.80
25	0.68	1.32	1.71	2.06	2.49	2.79

A CHI-SQUARE TABLE



Degrees of freedom	99%	95%	90%	70%	50%	30%	10%	5%	1%
1	0.00016	0.0039	0.016	0.15	0.46	1.07	2.71	3.84	6.64
2	0.020	0.10	0.21	0.71	1.39	2.41	4.60	5.99	9.21
3	0.12	0.35	0.58	1.42	2.37	3.67	6.25	7.82	11.34
4	0.30	0.71	1.06	2.20	3.36	4.88	7.78	9.49	13.28
5	0.55	1.14	1.61	3.00	4.35	6.06	9.24	11.07	15.09
6	0.87	1.64	2.20	3.83	5.35	7.23	10.65	12.59	16.81
7	1.24	2.17	2.83	4.67	6.35	8.38	12.02	14.07	18.48
8	1.65	2.73	3.49	5.53	7.34	9.52	13.36	15.51	20.09
9	2.09	3.33	4.17	6.39	8.34	10.66	14.68	16.92	21.67
10	2.56	3.94	4.86	7.27	9.34	11.78	15.99	18.31	23.21
11	3.05	4.58	5.58	8.15	10.34	12.90	17.28	19.68	24.73
12	3.57	5.23	6.30	9.03	11.34	14.01	18.55	21.03	26.22
13	4.11	5.89	7.04	9.93	12.34	15.12	19.81	22.36	27.69
14	4.66	6.57	7.79	10.82	13.34	16.22	21.06	23.69	29.14
15	5.23	7.26	8.55	11.72	14.34	17.32	22.31	25.00	30.58
16	5.81	7.96	9.31	12.62	15.34	18.42	23.54	26.30	32.00
17	6.41	8.67	10.09	13.53	16.34	19.51	24.77	27.59	33.41
18	7.00	9.39	10.87	14.44	17.34	20.60	25.99	28.87	34.81
19	7.63	10.12	11.65	15.35	18.34	21.69	27.20	30.14	36.19
20	8.26	10.85	12.44	16.27	19.34	22.78	28.41	31.41	37.57

Source: Adapted from p. 112 of Sir R. A. Fisher, *Statistical Methods for Research Workers* (Edinburgh: Oliver & Boyd. 1958).

Possibly useful formulas for this exam

These are provided for your convenience, but it is your responsibility to know how and when to use them.

- calculator average: $\bar{x}$
- $SD = \sqrt{[\text{ave. of (value}^2)] - (\text{ave. of value})^2}$
- calculator SD: σ_n
- $r = \dfrac{(\text{ave. of x}\cdot\text{y}) - (\text{ave. of x})(\text{ave. of y})}{(\text{SD for x})(\text{SD for y})}$
- $slope = r \cdot \dfrac{\text{SD for y}}{\text{SD for x}}$
- rms error $= (\sqrt{1 - r^2}) \cdot (\text{SD for y})$
- P(A and B) = P(A)×P(B|A)
- P(A or B) = P(A) + P(B) − P(A and B)
- $EV_{sum} = (\#\text{ of draws}) \times ave_{box}$
- $SE_{sum} = \sqrt{\#\text{ of draws}} \times SD_{box}$
- conf. int.: estimate $\pm\, z_\alpha \times (SE)$
- $SD_{box} = \sqrt{(\text{fraction of 0's}) \times (\text{fraction of 1's})}$
- $EV_{\%} = \%$ of 1's in box
- $SE_{\%} = \dfrac{SE_{sum}}{\#\text{ of draws}} \times 100\% = \dfrac{SD_{box}}{\sqrt{\#\text{ of draws}}} \times 100\%$
- $EV_{ave} = ave_{box}$
- $SE_{ave} = \dfrac{SE_{sum}}{\#\text{ of draws}} = \dfrac{SD_{box}}{\sqrt{\#\text{ of draws}}}$
- $z = \dfrac{\text{obs} - \text{exp}}{\text{SE}}$
- $SD^{+} = \sqrt{\dfrac{\text{number of draws}}{\text{number of draws - 1}}} \times SD$
- $SE_{A-B} = \sqrt{SE_A^2 + SE_B^2}$
- $\chi^2 = \text{sum of } ((\text{obs} - \text{exp})^2/\text{exp})$
- $\text{exp} = \dfrac{(\text{row tot})(\text{col tot})}{\text{table tot}}$
- $y - ave_y = slope \cdot (x - ave_x)$