Name:

Stat 1040, Fall 2001
Final Test, Thursday December 13, 9:30–11:20 am

Show your work. The test is out of 100 points and you have 110 minutes.

1. A recent study in Europe looked at a large group of women of childbearing age. The researchers asked each woman how much alcohol they had consumed over the past 12 months. The researchers found that women who drank moderate amounts of alcohol were somewhat less likely to have infertility than women who did not (November, 2001). The study said it "controlled for age, income and religion".

   (a) (3 points) Based on the information above, was this a controlled experiment or an observational study? Explain briefly.

   no intervention was used – nobody was told to drink / not to drink

   (b) (3 points) Why did they "control for" age, income and religion?

   these factors may be confounding factors

   (c) (4 points) Is this convincing evidence that infertility would decrease if women with infertility started to drink moderate amounts of alcohol? (Note: we are only asking about infertility. There may be other problems introduced by such behavior, but ignore these for answering this question).

   no! – we only know that there is association between drinking and fertility; drinking does not cause fertility

   (d) (4 points) Suggest a possible confounding factor (other than age, income, or religion) and clearly explain why you think it might be a confounding factor.

   general health (condition):
   someone who has some other medical problem may not drink and also be less fertile

2. A selection of 65 varieties of cereal were tested for calories and sodium (in milligrams) for a one-cup serving. The results may be summarized as follows:

y: Average sodium = 240 mg SD = 131 mg
x: Average calories = 149 calories SD = 62 calories r = 0.53

(a) Suppose we were to convert our 65 sodium measurements to grams, by dividing each measurement by 1000. Using this new set of measurements,

i. (4 points) What will the average and SD of sodium (in grams) be?

$$\text{avg} = \frac{240}{1000} = 0.24\ g$$

$$SD = \frac{131}{1000} = 0.131\ g$$

ii. (3 points) What will the correlation between calories and sodium be now?

$r = 0.53$, i.e., the same

(b) (5 points) Find the regression estimate for the number of mg of sodium in a one-cup serving of a cereal that has 200 calories per cup.

x: calories
y: sodium

$$\text{slope} = r \cdot \frac{SD_y}{SD_x} = 0.53 \cdot \frac{131}{62} = 1.12$$

intercept = $\text{avg}_y - \text{slope} \times \text{avg}_x = 240 - 1.12 \cdot 149 = 73.12$

regression equation: sodium = $73.12 + 1.12 \cdot$ calories

for 200 calories:
sodium = $73.12 + 1.12 \cdot 200 = 297.12$

(c) (4 points) Explain why it would not be a good idea to use the information in the question to estimate the amount of sodium for a cereal with 350 calories per cup.

350 calories: $\frac{350 - 149}{62} = 3.2$ s.u.

350 is more than 3 s.u. above the average (149 calories);
this seems to be extrapolation; the result most likely will be meaningless

3. (8 points) According to the U.S. Census Bureau, 68% of Utah residents are 18 years of age or older. What is the chance that in a simple random sample of 100 Utah residents, less than 50% will be 18 years of age or older?

$\boxed{1}$: 18 or older
$\boxed{0}$: below 18

box: $68 \times \boxed{1} \quad 32 \times \boxed{0}$

number of draws: 100

box avg = 0.68

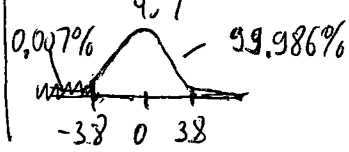
box SD = $\sqrt{0.68 \cdot 0.32} = 0.47$

$SE_{sum} = \sqrt{100} \cdot 0.47 = 4.7$

$EV_{\%} = 68\%$

$SE_{\%} = \frac{4.7}{100} \cdot 100\% = 4.7\%$

s.u.: $\frac{50 - 68}{4.7} = -3.8$



$$\frac{100\% - 99.986\%}{2} = 0.007\%$$

4. An elementary school in Logan employs 15 teachers; 11 are women and 4 are men. Two teachers are selected at random to meet the governor and attend a reception in SLC. Answer each part separately.

   (a) (3 points) What is the probability that both are women?

   $\frac{11}{15} \cdot \frac{10}{14} = \frac{110}{210} = 0.52 = 52\%$

   (b) (3 points) What is the probability that at least one is a woman?

   complement rule

   probability that both are men : $\frac{4}{15} \cdot \frac{3}{14} = \frac{12}{210}$

   probability that at least 1 is a woman : $1 - \frac{12}{210} = \frac{198}{210} = 0.94 = 94\%$

   (c) (3 points) What is the probability that both are the same gender?

   2 women or 2 men — mutually exclusive :

   addition rule $\frac{110}{210} + \frac{12}{210} = \frac{122}{210} = 0.58 = 58\%$

5. (12 points) Two types of water filters are to be compared in terms of the average reduction in impurities measured in parts per million (ppm). Fifty filters of each type were tested and the following data showing the amount of impurities removed were obtained:

	Type I	Type II
Average	8.0	7.1
SD	4.5	3.2

2-sample z-test

Is there statistical evidence that filters of type I perform better (remove more particles)? State null and alternative hypotheses. Compute the *P*-value and state your conclusions.

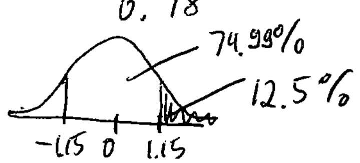
(1) Null : both filters perform equally well, i.e., $avg_{I} = avg_{II}$

Alternative: filters of type I perform better than filters of type II, i.e., $avg_{I} > avg_{II}$

(2) $SE_{avg_I} = \frac{\sqrt{50} \cdot 4.5}{50} = 0.64$ $SE_{avg_{II}} = \frac{\sqrt{50} \cdot 3.2}{50} = 0.45$

$SE_{diff} = \sqrt{0.64^2 + 0.45^2} = 0.78$

$z = \frac{8.0 - 7.1}{0.78} = 1.15$

(3) 

P-value = 12.5%

(4) P-value > 5% ; do not reject null hypothesis

→ both filters perform equally well

6. A survey was conducted to evaluate the effectiveness of a new flu vaccine that had been administered in a community of 20,000 people. The vaccine was provided free of charge in a two-shot sequence over a period of two weeks. Some people received the two-shot sequence, some appeared only for the first shot, and others received neither.

A simple random sample of 1,000 local inhabitants were surveyed the following spring and the results are shown in the table below.

obs exp · obs exp obs exp

	No vaccine	One shot	Two shots	Total
Flu	24 14	9 5	13 27	46
No Flu	289 299	100 104	565 551	954
Total	313	109	578	1000

χ^2-test for independence

(a) (10 points) Is there evidence that the vaccine classification and the occurrence or nonoccurrence of the flu are related? State the null and alternative hypotheses, compute the test statistic and the P-value and state your conclusion.

(1) Null: number of shots & occurrence of flu are independent, i.e., boxes are the same
Alternative: number of shots & occurrence of flu are dependent, i.e., at least one box is different

(2) e.g. $\frac{313 \cdot 46}{1000} = 14.4 \approx 14$ etc.

$d.f. = (2-1)\cdot(3-1) = 2$

$$\chi^2 = \frac{(24-14)^2}{14} + \frac{(9-5)^2}{5} + \frac{(13-27)^2}{27} + \frac{(289-299)^2}{299} + \frac{(100-104)^2}{104} + \frac{(565-551)^2}{551} = 18.45$$

(3) in table: 1% ~> 9.21, so 18.45 ~> less than 1%
P-value less than 1%

(4) reject null hypothesis (P-value < 1%); result highly statistically significant
-> number of shots has impact on occurrence of flu (shots & flu are dependent)

(b) (10 points) Find a 95% confidence interval for the percent of people in this community who received at least one shot of the vaccine.

sample % $= \frac{109+578}{1000} = 0.687 = 68.7\%$

"box SD" $= \sqrt{0.687 \cdot 0.313} = 0.46$

$SE_{sum} = \sqrt{1000} \cdot 0.46 = 14.5$

$SE_{\%} = \frac{14.5}{1000} \cdot 100\% = 1.45\%$

95% CI:

$68.7\% \pm 2 \cdot 1.45\%$

$= 68.7\% \pm 2.9\%$

$= 65.8\%$ to 71.6%

7. (10 points) A major manufacturer wants to test a new engine to determine whether it meets new air-pollution standards. The average emission of all engines of this type must be no more than 20 parts per million of carbon. Ten engines are manufactured for testing purposes, and the average and SD for this sample of engines are determined to be 24.1 and 3.0 parts per million, respectively. Assuming that these engines are like a simple random sample from a very large population and that the emissions follow the normal curve, is there evidence that this type of engine fails to meet the pollution standard? Set up a null and alternative hypothesis, perform the test, and clearly state your conclusion.

sample size = 10 (< 30), "normal" & SD box unknown → t-test

(1) Null: emission standard met (avg = 20 ppm)

Alternative: emission standard not met (avg > 20 ppm)

(2) $SD^+ = \sqrt{\frac{10}{10-1}} \cdot 3.0 = 3.16$

$SE_{sum} = \sqrt{10} \cdot 3.16 = 9.99 \approx 10$

$SE_{avg} = \frac{10}{10} = 1.0$

$t = \frac{24.1-20}{1.0} = 4.1$ $df = 10-1 = 9$

(3) in table: 0.5% → 3.25, so 4.1 less than 0.5%

p-value less than 0.5%

(4) reject null hypothesis (p-value < 0.5%), result highly statistically significant

→ emission standard not met

8. (5 points) The Salt Lake City metropolitan area has about 1.3 million people; the New York City metropolitan area has about 21.2 million — about 16.3 times as many as Salt Lake. Suppose we wish to take a survey to compare attitudes toward environmental policies in these two areas. We are happy with the accuracy (SE) of a survey of 1,200 Salt Lake residents. To get equivalent accuracy in New York, how many New York residents should we survey? Briefly explain.

1,200, i.e., about the same

accuracy (SE) only depends on sample size (as long as population size is at least 10 times larger than sample size)

9. (5 points) The average age of all 43 presidents when they entered office is 55.3 years, and the SD is 6.2 years. Explain why it would be inappropriate to use these numbers to conduct a significance test on the hypothesis that the average age of entering presidents is 50 years.

we <u>know</u> that the average is 55.3 years since this is the average for the whole "population of US presidents" – no need to perform any test

Memory Aids

Please note that these are provided for your convenience, but it is your responsibility to know how and when to use them.

$$\text{rms error } = \sqrt{1-r^2} \times SD_Y$$

$$\text{slope} = r \times \frac{SD_Y}{SD_X}$$

$$\text{intercept} = \text{ave}_Y - \text{slope} \times \text{ave}_X$$

$$SD^+ = \sqrt{\frac{\text{number of draws}}{\text{number of draws} - 1}} \times SD$$

$$SD_{\text{box}} = \sqrt{\text{fraction of 0's} \times \text{fraction of 1's}}$$

$$EV_{\text{sum}} = \text{number of draws} \times \text{ave}_{\text{box}}$$
$$SE_{\text{sum}} = \sqrt{\text{number of draws}} \times SD_{\text{box}}$$

$$EV_{\text{ave}} = \text{ave}_{\text{box}}$$
$$SE_{\text{ave}} = \frac{SE_{\text{sum}}}{\text{number of draws}}$$

$$EV_{\%} = \text{\% of 1's in the box}$$
$$SE_{\%} = \left(\frac{SE_{\text{sum}}}{\text{number of draws}}\right) \times 100\%$$

$SE_{\text{diff}} = \sqrt{a^2+b^2}$ where a is the SE for the first quantity,
b is the SE for the second quantity, and the two quantities are independent