Stat 1040, Fall 2008 **Name:** ______________

Final Test, December 9, 1:30pm–3:20pm

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

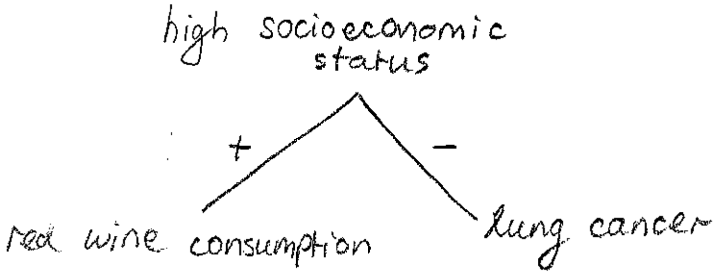
1. The following information came from **FOX NEWS,** October 10 2008:

> Drinking red wine not only reduces your risk for cardiovascular disease, but it may also reduce your risk for lung cancer especially if you are a current or ex-smoker, Reuters reported Thursday.
>
> People who do or have smoked and drink at least one glass of wine each day are 60 percent less likely to develop lung cancer than those who have smoked and don't drink red wine, said Dr. Chun Chao, of the Kaiser Permanente Southern California in Pasadena.
>
> Chao said it's the resveratrol and flavonoids in red wine that are protective – something white wine does not have.
>
> The reduction seen with red wine "lends support to a causal association for red wine and suggests that compounds that are present at high concentrations in red wine but not in white wine, beer or liquors may be protective against lung carcinogenesis," Chao wrote in her study.
>
> However, previous studies examining the correlation between alcohol consumption and lung cancer haven't always had the same results, Chao and her team noted in the journal Cancer Epidemiology, Biomarkers and Prevention.

(a) (1 point) Was the study a controlled experiment or an observational study?

This was an observational study because they obviously could not tell people how much to drink!

(b) (3 points) Clearly explain why socioeconomic status could be a confounding factor in this study and why this might make you doubt their conclusion.



People in lower socioeconomic groups may not drink red wine because it's expensive, but they may be more likely to smoke because there is more social pressure to smoke in lower socioeconomic groups. So this might make it appear that red wine prevents lung cancer, when really it's just that red wine drinkers are less likely to be smokers.

This is an example answer. They could also say lower s.e. groups more likely to drink beer than red wine because beer is more socially accepted for them. Also, higher s.e. groups more educated so less likely to smoke, more access to good healthcare, better diet, less likely to be working in toxic environments, etc.

2. Background: to participate in an online dating service, people are required to answer a number of questions, including the length of their index finger. Why do they ask this? Female participants often complain that men have lied about their height. Finger length is positively correlated with height, so perhaps the dating service collects finger length to check whether men are telling the truth about their height.

Heights and finger lengths for a group of male students are summarized by:

x	Finger length:	average = 7.83 cm	SD = .65 cm	r = 0.34
y	Height:	average = 179.0 cm	SD = 6.3 cm	

The scatter-diagram is football-shaped.

(a) (3 points) Find the equation of the regression line for predicting height from finger length.

$$\text{slope} = r\frac{SD_y}{SD_x} = .34\left(\frac{6.3}{.65}\right) = 3.295$$

$$\text{intercept} = \text{ave}_y - \text{slope}(\text{ave}_x) = 179 - 3.295(7.83) = 153.2$$

$$\underset{\text{height}}{y} = 153.2 + 3.295\underset{\text{finger length}}{x}$$

(b) (2 points) If a male student has a finger length of 8.56 cm, how tall do you predict him to be?

$$\text{height} = 153.2 + 3.295(8.56)$$
$$= 181.4$$

$\underline{\text{or}}$: 8.56 is $\frac{8.56-7.83}{.65} = 1.123$ SDs above average: $r(1.123) = .38$

.38 SDs above average in height is $(.38)(6.3) + 179 = 181.4$

(c) (1 points) Find the rms error for your answer in (b).

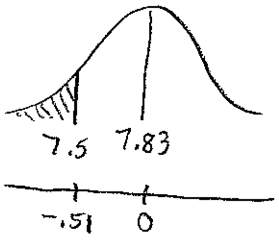
$$\text{rms} = \left(\sqrt{1-r^2}\right)(SD_y) = \left(\sqrt{1-.34^2}\right)(6.3) = (.94)(6.3) = 5.92$$

(d) (2 points) How useful is finger length for detecting whether or not men lie about their height? Use the numerical facts provided to support your answer.

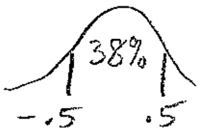
It is not very useful because the correlation of .34 is so weak and this means there is a lot of scatter around the line. Another way to see this is that the rms error is 5.92 and $SD_y = 6.3$ so there is almost as much error as there would be if we knew nothing about finger length.

3. From question 2, the average finger length for the male students is 7.83 cm with an SD of .65 cm. A histogram for the finger lengths is very close to the normal curve.

   (a) (5 points) What percentage of the students have fingers less than 7.5 cm long?



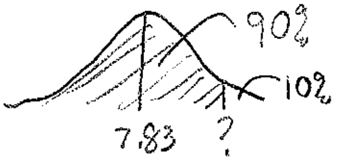
$z = \frac{7.5-7.83}{.65} = -.51$

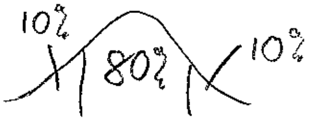


100%-38% = 62%

$\frac{62\%}{2}$ = 31% ANS

   (b) (5 points) If a student's finger length is at the 90th percentile, how long is it?





gives z = 1.3

1.3 SDs above average = (1.3)(.65) + 7.83
= 8.675

   (c) (2 points) If we were told that the histogram for the finger lengths did *not* follow the normal curve. Is your answer to (a) still valid (yes/[✓ no])? Is your answer to (b) still valid (yes/[✓ no])? Both of these calculations require the data to follow the normal curve.

4. (2 points) From question 3, the average finger length for a group of female students is 6.55 cm with an SD of .65 cm. If we combined the two groups, the SD would be (underline the correct answer)

   (a) equal to .65 cm.

   (b) smaller than .65 cm.

   [✓ (c)] larger than .65 cm.

5. (2 points) From question 3, suppose one of the students is an outlier because he has unusually short fingers. If we remove this student from the list, the average of the remaining male students will be (underline the correct answer)

   (a) equal to 7.83 cm.

   (b) smaller than 7.83 cm.

   [✓ (c)] larger than 7.83 cm.

6. (2 points) Anthropologists tell us humans tend to marry others similar to themselves. In one study, they recorded the heights of a group of students along with the students' estimates of the ideal height of their future spouse. They found r = -.33. However, when they looked more closely, they found that r = .60 for males and r=.56 for females. This is an example of

   positives   negative

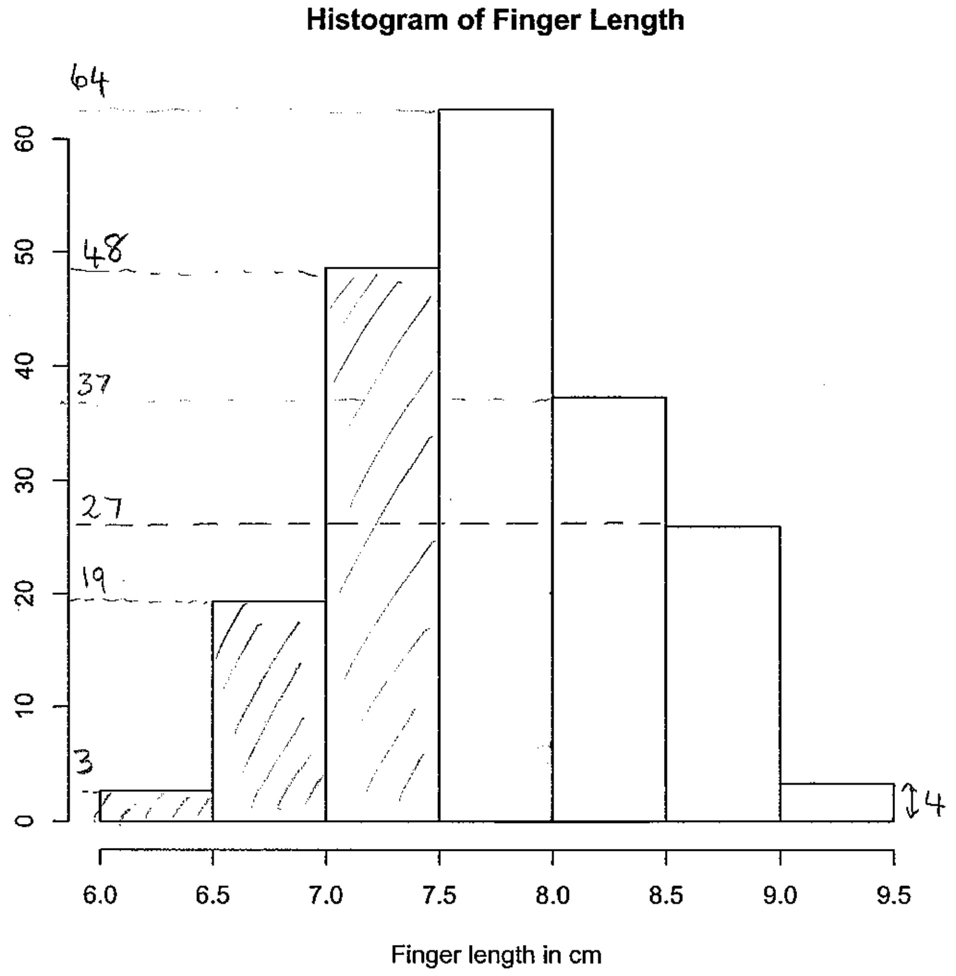
   [✓ (a)] Simpson's paradox.

   (b) ecological correlation.

   (c) correlation is not causation.

Simpson's paradox occurs when we get a different conclusion when we look at a big group than we do when we break it up into smaller groups.

7. The following histogram summarizes the finger lengths of 300 men. Note that these are **NOT** the same as the students in questions 2 through 6. Class intervals include the left endpoint but not the right.



(a) (1 point) Label the vertical axis. Percent per cm.

(b) (2 points) Using the histogram, what percentage of the men have fingers that are less than 7.5 cm long? Show your work.

See shading.

Area = .5(3) + .5(19) + .5(48) = 35%

(c) (2 points) Using the histogram, in which interval is the 70th percentile? Show your work.

7.5 to 8.0 has area .5(64) = 32% 35% + 32% = 67% not enough.

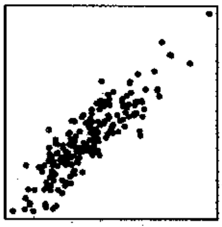
8.0 to 8.5 has area .5(37) = 18.5% 67% + 18.5% = 85.5% too much

It's in the 8.0 to 8.5 interval.

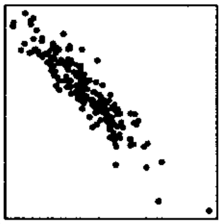
8. (6 points) Match each of the following scatterplots to their correlations from the list:

-.9, -.7, 0, .6, .85, .95

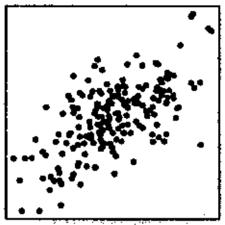
r = .85



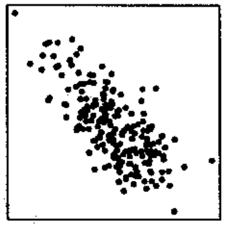
r = -.9



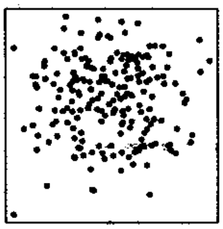
r = .6



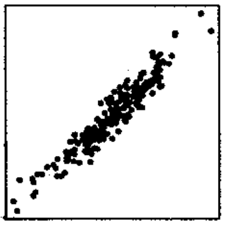
r = -.7



r = 0



r = .95



9. (7 points) A simple random sample of 500 Cache Valley voters shows that 125 of them voted for Obama in the 2008 Presidential election. Find a 95% confidence interval for the percentage of all Cache Valley voters who voted for Obama in the 2008 Presidential election.

125 out of 500 is $\frac{125}{500}\times 100\% = 25.0\%$.

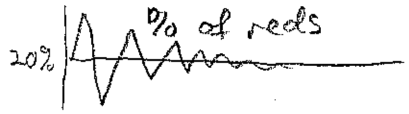
To approximate the SD we pretend the box looks like:

$\underbrace{75\,\boxed{0} \quad 25\,\boxed{1}}$ $SD_{box} \approx .433$

$SE_{sum} = \sqrt{500}\,(.433) = 9.68$

$SE_{\%} = \frac{9.68}{500}\times 100\% = 1.94\%$

CI: 25.0% ± 2(1.94%) i.e. 25.0% ± 3.9% <u>21.1% to 28.9%</u>



10. A box contains 2 red marbles and 8 blue marbles. I plan to sample **WITH** replacement from this box. In each of the following cases, circle the correct answer. No explanation is required; if you provide one it will not help your score.

    (a) (2 points) You win \$1 if red marbles are selected more than 15% of the time. Which is better for you: 100 draws or [✓ 500] draws?

    (b) (2 points) You win \$1 if red marbles are selected exactly $\frac{1}{5}$ of the time. Which is better for you: [✓ 100] draws or 500 draws?

    (c) (2 points) You win \$1 if red marbles are selected how between 15% and 25% of the time. Which is better for you: 100 draws or [✓ 500] draws?

2 marbles

11. (12 points) A box contains 2 red marbles and 8 blue marbles. For parts (a) through (c), assume we draw **WITH** replacement from the box. For parts (d) through (f), assume we draw **WITHOUT** replacement from the box.

with

    (a) What is the chance that both of the marbles are blue?

    $$\left(\frac{8}{10}\right)\left(\frac{8}{10}\right) = \underline{\underline{.64}}$$

    (b) What is the chance that one of the marbles is blue and the other is red?

    $$\underbrace{\left(\frac{8}{10}\right)\left(\frac{2}{10}\right)}_{B \quad R} + \underbrace{\left(\frac{2}{10}\right)\left(\frac{8}{10}\right)}_{R \quad B} = \underline{\underline{.32}}$$

    B R or R B

    (c) What is the chance that I get at least one red marble?

    1 – chance they are both blue $= 1 - \left(\frac{8}{10}\right)\left(\frac{8}{10}\right) = \underline{\underline{.36}}$

without

    (d) What is the chance that both of the marbles are blue?

    $$\left(\frac{8}{10}\right)\left(\frac{7}{9}\right) = \underline{\underline{.622}}$$

    (e) What is the chance that one of the marbles is blue and the other is red?

    $$\left(\frac{8}{10}\right)\left(\frac{2}{9}\right) + \left(\frac{2}{10}\right)\left(\frac{8}{9}\right) = \underline{\underline{.356}}$$

    B R or R B

    (f) What is the chance that I get at least one red marble?

    1 – chance they are both blue $= 1 - \left(\frac{8}{10}\right)\left(\frac{7}{9}\right) = \underline{\underline{.378}}$

12. (11 points) Rosuvastatin is a cholesterol-lowering medication that was recently tested using a randomized, controlled, double-blind experiment. The researchers wanted to know whether Rosuvastatin protected against cardiovascular death. Of the 8901 subjects in the Rosuvastatin group, 83 died from cardiovascular causes; of the 8901 subjects in the placebo group, 157 died from cardiovascular causes. *Source: The New England Journal of Medicine, November 2008.*

(a) Clearly state the null and alternative hypotheses.

2 null hypothesis: Rosuvastatin is the same as placebo for cardiovascular death

alt hypothesis: " prevents cardiovascular death

(b) Calculate the appropriate test statistic.

Rosuvastatin

sample % is $\frac{83}{8901} \times 100\% = .93\%$

$\underbrace{8818 \boxed{0} \quad 83 \boxed{1}}$, $SD_{box} \approx .10$

$SE_{sum} = \sqrt{8901}\,(.10) = 9.43$

$SE_{\%} = \frac{9.43}{8901} \times 100\% = .106\%$

Placebo

sample % is $\frac{157}{8901} \times 100\% = 1.76\%$

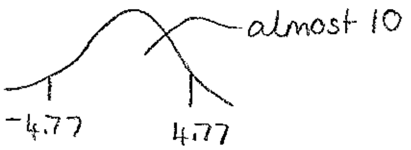
$\underbrace{8744 \boxed{0} \quad 157 \boxed{1}}$, $SD_{box} = .13$

$SE_{sum} = \sqrt{8901}\,(.13) = 12.26$

$SE_{\%} = \frac{12.26}{8901} \times 100\% = .138\%$

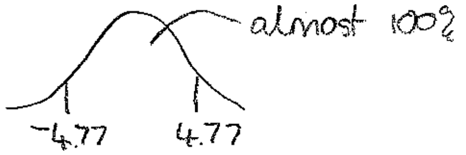
3 $SE_{diff} = \sqrt{.106\%^2 + .138\%^2} = .174\%$

2 $z = \frac{.93 - 1.76}{.174} = -4.77$



(c) Find the P–value.

1



so the P-value is almost 0%

(d) Do you reject the null hypothesis? Explain why or why not.

1 Yes! We reject the null hypothesis because the P-value is small (much less than 5%).

(e) State your conclusions.

2 We conclude that Rosuvastatin prevents cardiovascular death.

13. (12 points) In the 2008 Presidential election, a simple random sample of 500 people from each of Box Elder, Cache, and Weber Counties gave the following results:

	Obama	McCain	*Total*
Box Elder	134 91	366 409	500
Cache	134 120	366 380	500
Weber	134 192	366 308	500
Total	403	1097	1500

We are interested in whether or not voting behavior and County are independent in this population.

(a) Clearly state the null and alternative hypotheses.

2 null: voting behavior is independent of County
alt: " " " not " " "

(b) Calculate the appropriate test statistic.

5

obs	exp	$(obs-exp)^2/exp$
91	134	13.8
120	134	1.5
192	134	25.1
409	366	5.1
380	366	.5
308	366	9.2

$\chi^2 = 55.2$

1 (c) Find the degrees of freedom. $df = (3-1)(2-1) = 2$

1 (d) Find the P–value. off the chart

The P-value is way less than 1%.

1 (e) Do you reject the null hypothesis? Explain why or why not.

Yes! We reject the null because the P-value is small.

(f) State your conclusions.

2 We conclude that voting behavior is dependent on County for this population.

14. (3 points) In the 2008 Presidential election, the final voting results for Box Elder, Cache, and Weber Counties were as follows:

	Obama	McCain	*Total*	% Obama
Box Elder	3080	14340	17420	17.7%
Cache	9806	27799	37605	26.1%
Weber	24028	43250	67278	35.7%.
Total	36914	85389	122303	

Explain why it is *not* correct to perform a statistical hypothesis test with these data.

We have the whole population, so a statistical test is not necessary. We can see the percentage who voted for Obama is different in the 3 counties.

15. (9 points) A website claims that the average height of Utah men is 180 cm. A student thinks the average is lower than 180 cm. She takes a simple random sample of 10 men and finds that the average is 179.2 cm with an SD of 6.5 cm. Could this difference be due to chance error? Clearly state the null and alternative hypotheses, calculate the appropriate test statistic, find the P-value, and state your conclusion.

null: the average is 180 cm

alt: the average is less than 180 cm

$$SD^{+} = \sqrt{\frac{10}{9}}\ 6.5 = 6.85$$

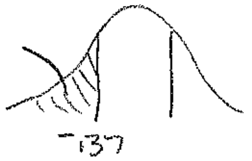
$$SE_{sum} = \sqrt{10}\,(6.85) = 21.67$$

$$SE_{ave} = \frac{21.67}{10} = 2.167$$

$$t = \frac{179.2 - 180}{2.167} = -.369 \approx -.37$$

$$df = 10 - 1 = 9$$

P-value



The P-value is larger than 25% so we fail to reject the null & conclude that we don't haveevidence the average is less than 180cm.