Stat 1040, Fall 2008 Name: _____

Final Test, December 9, 1:30pm–3:20pm

101 → 404

Show your work. The test is out of 101 points and you have 110 minutes to finish.

4 → 16 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.

4

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

[There was no intervention: The subjects decided themselves what to drink and whether to smoke or not.]

12 (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.

(12) for valid explanation

People in lower socioeconomic groups may not drink much red wine because it is 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 is just that red wine drinkers are less likely to be smokers.

Also, higher socioeconomic groups are more educated and so are less likely to smoke, and they have more access to good health care, better diet, and they are less likely to be working in toxic environments, etc.

8 → 32

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.

−2 for each calculation error
−2 if x/y swapped
−2 if x/y not specified

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

$$\text{slope} = r \cdot \frac{SD_Y}{SD_X} = 0.34 \cdot \frac{6.3}{0.65} = 3.295 \quad (5)$$

$$\text{intercept} = \text{avg}_Y - \text{slope} \cdot \text{avg}_X = 179 - 3.295 \cdot 7.83 = 153.2 \quad (5)$$

regression equation: $\boxed{Y = 153.2 + 3.295 \cdot X}$ (2)

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

height (for a man with a finger length of 8.56 cm)

$= 153.2 + 3.295 \cdot 8.56$ (8)

$= 181.4$ cm

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

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

$$= \sqrt{1-0.34^2} \cdot 6.3 = 5.92 \text{ cm} \quad (4)$$

8 (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.

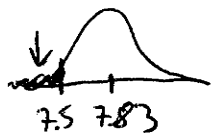
(6) Not very useful because the correlation of 0.34 is on the weaker side and this means there is a lot of scatter around the line. Also, note that the rms error is 5.92 cm, so the 2 rms error band becomes "predicted value ± 11.84 cm" and the 3 rms error band becomes "predicted value ± 17.76 cm". This is more than a head length!

(2) for valid explanation

12 → 48

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.

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



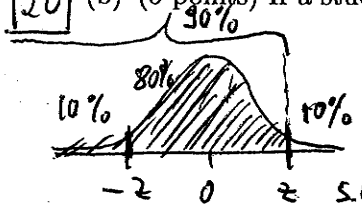
$s.u. = \frac{7.5 - 7.83}{0.65} = -0.51$ (7)

-2 for each calculation error

area between -0.50 and 0.50: 38.29% (7)

area below -0.50: $\frac{100\% - 38.29\%}{2}$ (6) = 30.86%

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



(5)

area between -1.30 and 1.30: 80.64% (closest to 80%)

original units: $1.30 \cdot 0.65 + 7.83 = 8.675$ cm

(5) (2) (3) (2) (3)

8 (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)?

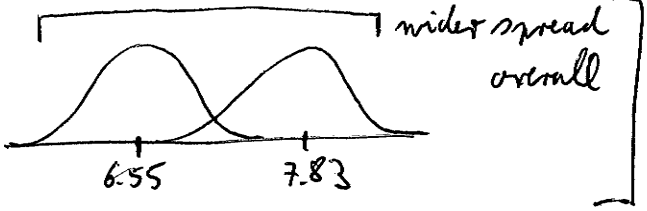
(4) [no! - both of these calculations require that the data follow the normal curve] (4)

2 → 8 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. (8)



2 → 8 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. (8)

[small outlier pulls average towards the smaller values; so average will be larger with this outlier being removed]

2 → 8 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

(a) Simpson's paradox. (8)

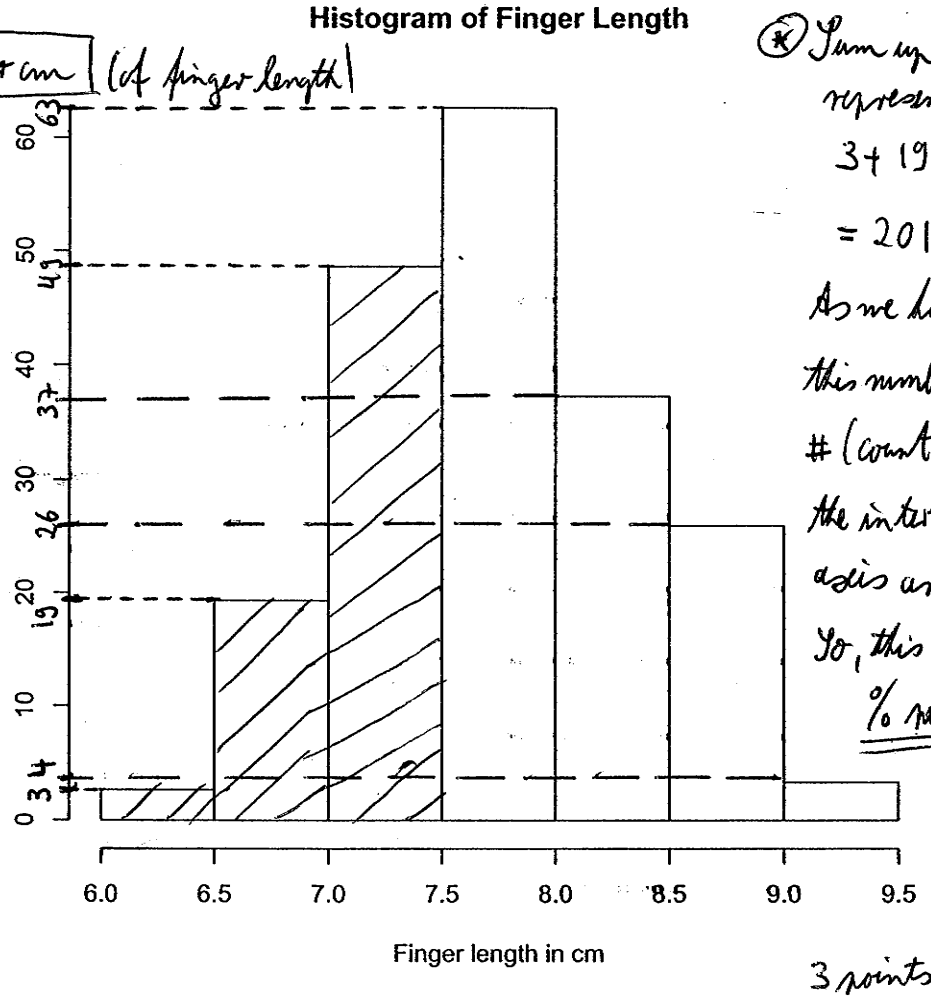
(b) ecological correlation.

(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 this group into smaller groups.]

5 → 20

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.



⊛ Sum up values that represent bar heights:
3 + 19 + 49 + 63 + 37 + 26 + 4
= 201
As we have data for 300 men, this number cannot represent the # (count) of men. Notice that the intervals on the horizontal axis are 0.5 cm wide. So, this must be % per cm (of finger length).

3 points for correct guess in b) & c) [with no work shown]

4 (a) (1 point) Label the vertical axis. ⊛ for explanation

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

Shaded area ≈ 3% · 0.5 + 19% · 0.5 + 49% · 0.5
= 35.5%

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

6.0 to 6.5: 3% · 0.5
6.5 to 7.0: 19% · 0.5
7.0 to 7.5: 49% · 0.5
= 35.5%; not enough
7.5 to 8.0: 63% · 0.5 = 31.5%
= 67%; not enough
8.0 to 8.5: 37% · 0.5 = 18.5%
= 85.5%; too much

So, the 70th percentile is in the 8.0 to 8.5 interval

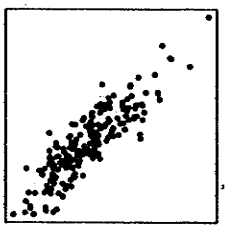
6 → 24

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

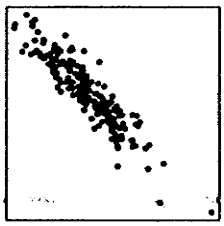
-.9, -.7, 0, .6, .85, .95

(4) each

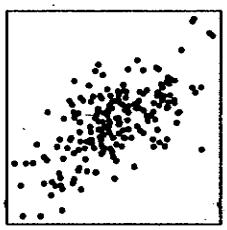
r = .85



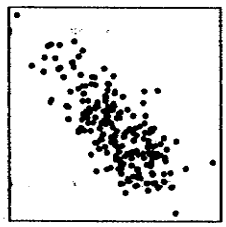
r = -.9



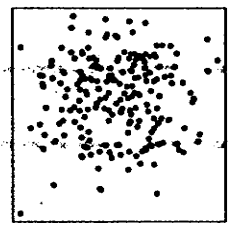
r = .6



r = -.7



r = 0



r = .95

7 → 28

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.

-2 for each calculation error

sample % = $\frac{125}{500}$ = 0.25 = 25% (3)

SD = $\sqrt{0.25 \cdot 0.75}$ = 0.433 (3)

$SE_{sum} = \sqrt{500} \cdot 0.433 = 9.68$ (3)

$SE_{\%} = \frac{9.68}{500} \cdot 100\% = 1.94\%$ (3)

95% CI : 25% ± 2 · 1.94% = 21.12% to 28.88%

(2) (2) (4) (2) (2) (2) (2)

6 → 24

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. $\frac{2}{10}$ = 20% red in box; sample % will get close to 20% if # of draws increases, but it is unlikely to be exactly 20%

[8] (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? (8)

[8] (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? (8)

[8] (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? (8)

12 → 48

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. We draw 2 marbles.

with

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

-2 for each calculation error (or no final result)

1st blue and 2nd blue

$\frac{8}{10} \cdot \frac{8}{10} = \frac{64}{100} = 0.64 = 64\%$ (3) (2) (3)

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

(1st blue and 2nd red) or (1st red and 2nd blue)

$\left(\frac{8}{10} \cdot \frac{2}{10}\right) + \left(\frac{2}{10} \cdot \frac{8}{10}\right) = \frac{32}{100} = 0.32 = 32\%$ (1)(1)(1) (2) (1)(1)(1)

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

1 − both blue from (a) $= 1 - 0.64 = 0.36 = 36\%$ (3)(2)(3)

without

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

1st blue and 2nd blue

$\frac{8}{10} \cdot \frac{7}{9} = \frac{56}{90} = 0.622 = 62.2\%$ (3) (2) (3)

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

(1st blue and 2nd red) or (1st red and 2nd blue)

$\left(\frac{8}{10} \cdot \frac{2}{9}\right) + \left(\frac{2}{10} \cdot \frac{8}{9}\right) = \frac{32}{90} = 0.356 = 35.6\%$ (1)(1)(1) (2) (1)(1)(1)

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

1 − both blue from (d) $= 1 - 0.622 = 0.378 = 37.8\%$ (3)(2)(3)

11 → 44

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.*

2-sample z-test:

(a) Clearly state the null and alternative hypotheses.

1/ null: Rosuvastatin (R) and Placebo (P) have the same effect on cardiovascular death, (3)
i.e., box% R − box% P = 0% (1)
alternative: R prevents cardiovascular death, (3)
i.e., box% R − box% P < 0% (1)

−33 for incorrect test
−4 if null, alt swapped
−2 for each calculation error

(b) Calculate the appropriate test statistic.

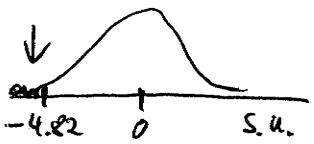
R	P
2/ sample size R: 8901	sample size P: 8901
sample % R: $\frac{83}{8901} = 0.0093 = 0.93\%$ (2)	sample % P: $\frac{157}{8901} = 0.0176 = 1.76\%$ (2)
$SD_R = \sqrt{0.0093 \cdot 0.9907} = 0.096$ (2)	$SD_P = \sqrt{0.0176 \cdot 0.9824} = 0.131$ (2)
$SE_{sum\,R} = \sqrt{8901} \cdot 0.096 = 9.06$ (2)	$SE_{sum\,P} = \sqrt{8901} \cdot 0.131 = 12.36$ (2)
$SE_{\%\,R} = \frac{9.06}{8901} \cdot 100\% = 0.101\%$ (2)	$SE_{\%\,P} = \frac{12.36}{8901} \cdot 100\% = 0.139\%$ (2)

$SE_{\%\,diff} = \sqrt{(0.101\%)^2 + (0.139\%)^2} = 0.172\%$ (4)

$z = \frac{0.93\% - 1.76\%}{0.172\%} = -4.82$ (4)

(c) Find the P–value.

3/



area between −4.82 and 4.82: almost 100% (2)
area below −4.82: about 0% (2) = P-value

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

4/ • yes, reject the null (2) (P-value < 5%) (2)

(e) State your conclusions.

- result is highly stat. significant (2) (P-value < 1%)
- Rosuvastatin prevents cardiovascular death (2)

12 → 48

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:

obs. count

	Obama	McCain	*Total*
Box Elder	91 (1)	409 (2)	500
Cache	120 (3)	380 (4)	500
Weber	192 (5)	308 (6)	500
Total	403	1097	1500

exp count

134	366	500
134	366	500
134	366	500
402	1098	1500

χ^2-test for independence

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

- 36 for incorrect test
- -4 if null, alt swapped
- -2 for each calculation error

(a) Clearly state the null and alternative hypotheses.

1, null: voting behavior and County are independent, (3)
i.e., boxes are identical (1)
alternative: voting behavior and County are not independent, (3)
i.e., at least one box is different (1)

(b) Calculate the appropriate test statistic.

2, expected count:

$$(1) = (3) = (5) = \frac{403 \cdot 500}{1500} = 134$$

$$(2) = (4) = (6) = 500 - 134 = 366$$

(1) each × 6 ⇒ (6)

$$\chi^2 = \text{sum of } \frac{(\text{obs}-\text{exp})^2}{\text{exp}}$$

$$= \frac{(91-134)^2}{134} + \frac{(409-366)^2}{366} + \frac{(120-134)^2}{134} + \frac{(380-366)^2}{366} + \frac{(192-134)^2}{134} + \frac{(308-366)^2}{366} = 55.14 \quad (6)$$

(c) Find the degrees of freedom.

3, $df = (3-1)\cdot(2-1) = 2$ (4)

(d) Find the P–value.

$\chi^2 = 55.14$ is to the right of 9.21 (4)
→ P-value is below 1% (4)

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

4, • yes, reject the null (4) (P-value < 5%) (4)

(f) State your conclusions.

- result is highly stat. significant (4) (P-value < 1%)
- voting behavior and County are not independent (4)

3→ 12

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	17.68%
Cache	9806	27799	37605	26.08%
Weber	24028	43250	67278	35.71%
Total	36914	85389	122303	

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

We have data from the whole population (12) (i.e., this is not a sample)! No statistical test is needed. We can immediately indicate that the percentage of voters who voted for Obama is different in these 3 counties.

9→ 36

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.

-27 for incorrect test
-4 if null, alt swapped
-2 for each calculation error

t-test:
- sample size < 30
- box SD unknown
- assume data follow normal curve

1) null: Utah men are 180 cm tall on average, (3)
i.e., $box_{avg} = 180$ cm (1)
alternative: Utah men are less tall on average, (3)
i.e., $box_{avg} < 180$ cm (1)

2) observed (avg) = 179.2
expected (avg) = 180
SD = 6.5
$SD^t = \sqrt{\frac{10}{9}} \cdot 6.5 = 6.85$ (4)
$SE_{sum} = \sqrt{10} \cdot 6.85 = 21.66$ (3)
$SE_{avg} = \frac{21.66}{10} = 2.166$ (3)
$t = \frac{179.2 - 180}{2.166} = -0.369$ (3)

3) $df = 10 - 1 = 9$ (3)
$t = 0.369$ is to the left of 0.7 (3)
→ p-value is above 25% (3)

4) • do not reject the null (3)
(p-value > 5%)
• Utah men are 180 cm tall on average (3)