Name:

Stat 1040, Fall 2002
Final Test, Wednesday December 11, 9:30–11:20 am

400

For full credit, **show your work.** *The test is out of 100 points and you have 110 minutes.*

4 → 56 1. Here is the beginning of an article from the YAHOO Health News web site on January 7, 1998:

"Wednesday January 7, 1998 For Yahoo News by Reuters
Daily Two–Mile Walk Halves Death Risk
NEW YORK (Reuters) — Walking two miles or more per day can cut the overall risk of dying in half, according to a new study. It also reduces the risk of dying from cancer — and appears to cut the risk of death due to cardiovascular diseases, US researchers report. Between 1980 and 1982, multicenter researchers in the Honolulu Heart Program studied 707 nonsmoking, retired men, aged 61 to 81 years, and collected mortality data on these men over the following 12 years. During the study, 208 of the men died. The study results show that while 43.1% of men who walked less than one mile per day died, only half this figure – 21.5% – of the men who walked more than two miles per day died."

8 (a) (2 points) Is the research described in the article an observational study or an experiment? (Please circle your answer.)

8 (b) (2 points) Is the research described in the article a cross-sectional study or a longitudinal study? (Please circle your answer.)

24 (c) (6 points) List two different possible confounding factors that are likely to have an effect on the outcome of this study. Carefully explain exactly why you think these are confounding factors.

- age: people aged 61 to 65 probably walk more than people aged 77 to 81 (and are less likely to die – independent of walking)
- general health status at the beginning of the study: someone with a severe medical problem right in the beginning may not be able to walk (and is more likely to die)
- eating and drinking habits (people that walk may have healthier habits)
- additional exercise (people that walk may do additional exercise)
- risks in earlier part of life: job, environmental factors, smoking in youth, etc.

12 for each correct factor/explanation

16 (d) (4 points) Based on the description in the body of this article, is the headline ("Daily Two–Mile Walk Halves Death Risk") justified? Yes or No? Circle your answer and give a brief explanation.

12 0-4 for explanation

The headline suggests that walking two (or more) miles per day "causes" the reduction in the risk of death. This is clearly not the case. First of all, the headline generalises from "nonsmoking, retired men, aged 61 to 81 years" to the entire population. Then, as seen in (c), even for this specific group of people, there are many confounding factors. So, at best, "walking two miles or more per day" may be associated with fewer deaths, but there is certainly no causation.

8 → 32

2. From 100 subjects in a health study, the following data were collected:

-10 if x, y swapped

x Average height = 68 inches SD = 2.5 inches

y Average blood pressure = 120 mm SD = 15 mm r = -0.2

16 (a) (4 points) Find the equation of the regression line for predicting blood pressure from height.

$\text{slope} = r \cdot \frac{SD_y}{SD_x} = -0.2 \cdot \frac{15}{2.5} = -1.2$ (6)

$\text{intercept} = \text{ave}_y - \text{slope} \cdot \text{ave}_x = 120 - (-1.2) \cdot 68 = 201.6$ (6)

equation: blood pressure = 201.6 − 1.2 · height (4)

or $y = 201.6 - 1.2 \cdot x$

16 (b) (4 points) Predict the blood pressure of a subject who is 73 inches tall.

-8 for impossible value

blood pressure for 73 inches tall person:

$201.6 - 1.2 \cdot 73 = 114$ mm (16)

8 → 32

3. (8 points) A simple random sample of 400 Cache Valley homes reveals that the average value is \$100,000 with an SD of \$100,000. Do the home values follow the normal curve? (Use the normal curve to find the percentage of homes that would have a negative value - what do you conclude?)

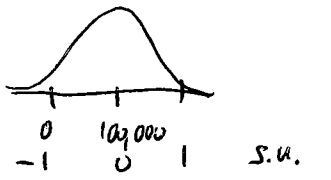
-24 for SE calculations (and not normal)

avg: 100,000

SD: 100,000

chance that house has a value of \$0 or below:

S.U.: $\frac{0 - 100{,}000}{100{,}000} = -1.0$ (12)



area from −1 to 1: 68.27%

area below −1: $\frac{100\% - 68.27\%}{2} = 15.865\% \approx 16\%$ (12)

There would be a chance of about 16% that a house has a negative value. This makes no sense! Home values clearly do not follow the normal curve. (8)

4 → 16

4. (4 points) A physician looks at all of her patients who have had a physical in the last 12 months. She selects the ones whose cholesterol levels were above the 90th percentile, and asks them to come in to the office and have a second cholesterol test. She finds that for these people, the average cholesterol level dropped between the first and second cholesterol test. This could be due to which of the following things? (Please circle the correct answer.)

(a) the regression effect. (16)

(b) the correlation effect. (6)

(c) ecological correlation. (6)

(d) correlation is not causation. (6)

6 → 24

5. (6 points) For the following situations, state which of the following types of sample was used: simple random sample, cluster sample, or sample of convenience.

12 (a) To survey the opinions of its customers, an airline company made a list of all its flights and randomly selected 25 flights. All of the passengers on those flights were asked to fill out the survey.

cluster sample (12) everything else: (0)

12 (b) To survey USU students, a newspaper reporter stood outside the student center and interviewed people as they walked by.

sample of convenience (12) everything else: (0)

10 → 40

6. (10 points) In August 2000, the Gallup Poll asked 507 randomly sampled California adults the question "Do you think the possession of small amounts of marijuana should be treated as a criminal offense?" Of these, 238 responded "No." Find a 95% confidence interval for the percentage of all California adults who would respond "No".

1: no
0: yes

box: [2 x 10 | 2 x 1] number of draws: 507

sample% : $\frac{238}{507} = 0.469 = 46.9\%$

SD box = $\sqrt{\frac{238}{507} \cdot \frac{269}{507}} = \sqrt{\frac{64022}{257,049}} = \sqrt{0.2491} = 0.499 \approx 0.5$ (10)

$SE_{sam} = \sqrt{507} \cdot 0.5 = 22.52 \cdot 0.5 = 11.26$ (5)

$SE_{\%} = \frac{11.26}{507} \cdot 100\% = 2.22\%$ (10) -2 if no final result

95% CI: $46.9\% \pm 2 \cdot 2.22\% = 46.9\% \pm 4.44\% = 42.46\%$ to 51.34%

(5) (5) (5)

10 → 40

7. A gum-ball machine has 150 gum-balls of which 27 are red, 32 are blue, 17 are pink, 23 are white, 21 are green, and 30 are yellow. When the machine dispenses gum-balls, it is like selecting at random without replacement. Three children buy gum-balls from this machine. Answer each of the following questions separately.

8 (a) (2 points) What is the chance that the first child gets a pink gum-ball?

$\frac{17}{150} = 0.113 = 11.3\%$ (8)

8 (b) (2 points) Suppose the first child gets her gum-ball and it is pink. What is the chance the second child also gets a pink gum-ball?

$\frac{16}{149} = 0.107 = 10.7\%$ (8)

12 (c) (3 points) What is the chance that none of the 3 gum-balls are pink?

first not pink: $\frac{133}{150}$ (3), second not pink: $\frac{132}{149}$ (3), third not pink: $\frac{131}{148}$ (3)

chance all not pink: $\frac{133}{150} \cdot \frac{132}{149} \cdot \frac{131}{148} = \frac{2,299,836}{3,307,800} = 0.695 = 69.5\%$ (3)

12 (d) (3 points) What is the chance that at least one of the 3 gum-balls is pink?

multiplication rule

opposite rule (statement (d) is opposite of statement (c)):

$1 - 0.695 = 0.305 = 30.5\%$ 3 (12) -4 ... -6 if somewhat off -11 if far off

6 → 24

8. In a randomized controlled experiment, 600 mice are randomly assigned to a treatment group (which receives high levels of electromagnetic radiation) and a control group (which receives only normal levels of electromagnetic radiation). After a year, the 600 mice are dissected and examined for 20 different forms of cancer - lung, liver, stomach, bone, and so on.

8

(a) (2 points) A two-sample z test is conducted for each of the 20 cancer rates. For a particular type of cancer, what is the null hypothesis? (Please circle the correct answer.) (0)

i. electromagnetic radiation increases the risk of that particular cancer. (0)

ii. electromagnetic radiation decreases the risk of that particular cancer.

iii. electromagnetic radiation has no effect on the risk of that particular cancer. (8)

16

(b) (4 points) It turns out that the P-value for the test on stomach cancer is 3%. The other 19 tests all have P-values greater than 5%. Clearly explain why this is *not* convincing evidence that electromagnetic radiation causes stomach cancer. (Hint: how many would you expect to show up as "statistically significant" if electromagnetic radiation is harmless?) (0)...(16)

for explanation

If we do a test, even if the null hypothesis is true, we have a 5% chance of rejecting the null hypothesis just by chance. If we do 20 tests, we would expect to reject $20 \cdot 5\% = 20 \cdot 0.05 = 1$ null hypothesis just by chance. This has happened here! Since the P-value of 3% for stomach cancer is not even close to 0%, there is no convincing evidence that electromagnetic radiation causes stomach cancer.

12 → 48

9. (12 points) A survey is taken of college students, asking "What is the fastest (in miles per hour) you have ever driven a car?" For the 102 female students, the average was 88.4 mph and the SD was 14.4 mph. For the 87 male students, the average was 107.4 mph and the SD was 17.4 mph. Assuming that these are independent simple random samples of the very large populations of female and male students, perform a significance test of the hypothesis that the population averages are the same (versus the alternative that males have a higher average). Clearly state a null and an alternative hypothesis, find the P-value, and indicate what this means in terms of male and female college students.

2-sample z-test

A: female	B: male
avg_A: 88.4	avg_B: 107.4
SD_A: 14.4	SD_B: 17.4
sample size: 102	sample size: 87

1/ Null: avg max speed is the same for women and men, i.e., $avg_B - avg_A = 0$ (4)

Alt: avg max speed for men is higher than for women, i.e., $avg_B - avg_A > 0$ (4)

2/ $SE_{sum_A} = \sqrt{102} \cdot 14.4 = 145.4$ (3) $SE_{sum_B} = \sqrt{87} \cdot 17.4 = 162.3$ (3)

$SE\ avg_A = \frac{145.4}{102} = 1.43$ (3) $SE_{avg_B} = \frac{162.3}{87} = 1.87$ (3)

$SE_{diff} = \sqrt{1.43^2 + 1.87^2} = \sqrt{5.54} = 2.35$ (5)

$z = \frac{107.4 - 88.4}{2.35} = \frac{19}{2.35} = 8.08$ (5)

3/ 8.08 off the z-table $\Rightarrow$ P-value $\approx 0\%$ (4)

4/ reject null (6); result is highly statistically significant (P-value < 1%) (2);

avg max speed for men is higher than for women (6)

10 → [40]

10. (10 points) A university administrator claims that "80% of all students favor increased Tier II tuition". A journalism student thinks the percentage is much lower, so she decides to test the null hypothesis that the percentage is 80% against the alternative that the percentage is lower than 80%. She takes a simple random sample of 500 students and finds that 67% of the students in her sample favor increased Tier II tuition. Find a test statistic and a P-value for the appropriate test, identify whether or not you should reject the null hypothesis, and state your conclusions about whether or not the administrator's claim is correct.

z-test

1/ given: Null: pop. % that favor increased tuition is as claimed, i.e., pop % = 80%
Alt: pop % that favor increased tuition is less than claimed, i.e., pop % < 80%

2/ sample % = 67%
SD box = $\sqrt{0.67 \cdot 0.33} = \sqrt{0.2211} = 0.47$ (4)
$SE_{sum} = \sqrt{500} \cdot 0.47 = 10.5$ (2)
$SE_{\%} = \frac{10.5}{500} \cdot 100\% = 2.1\%$ (4)
$z = \frac{67\% - 80\%}{2.1\%} = \frac{-13\%}{2.1\%} = -6.2$ (6)

3/ −6.2 off the z-table ⇒ P-value ≈ 0% (6)

4/ reject null (8) result is highly statistically significant (P-value < 1%) (2); the administrator's claim is not correct – the percentage of students that favor increased tuition is less than 80% (8)

12 → [48]

11. (12 points) A researcher is interested in learning whether breast-feeding affects how well a mother "bonds" with her baby. The researcher takes a simple random sample of 100 mothers and records whether they breast-fed or not, and also measures the level of bonding. The results are given in the table below.

χ^2-test for independence

obs | exp	level of bonding					
	low		medium		high	
breast-fed	10	12	21	21	34	32	65
bottle-fed	8	6	11	11	16	18	35
	18		42		50		100

(12) for correct expected values

For this population, is the level of bonding independent of the way the babies were fed? Answer with an appropriate statistical test. State the null and alternative hypotheses, compute a test statistic, estimate the P-value, and clearly state your conclusion. −4 if Null & Alt swapped

1/ Null: level of bonding is independent of the way babies were fed, i.e., boxes are the same (4)
Alt: level of bonding is dependent on the way babies were fed, i.e., at least 1 box is different (4)

2/ $\frac{18 \cdot 65}{100} = 11.7$ ≈ 12 etc. (see table above)

$\chi^2 = \text{sum of } \frac{(obs - exp)^2}{exp} = \frac{(10-12)^2}{12} + \frac{(21-21)^2}{21} + \frac{(34-32)^2}{32}$
$+ \frac{(8-6)^2}{6} + \frac{(11-11)^2}{11} + \frac{(16-18)^2}{18} = 1.35$ (8)

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

3/ from χ^2-table: 1.35 between 0.71 and 1.39 (5)
↓ 70% ↓ 50%, i.e., P-value between 50% and 70% (6)

4/ do not reject null (P-value > 5%);
level of bonding is independent of the way babies were fed (5)

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

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

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

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

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

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

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

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

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

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