Name:

Stat 1040, Spring 2003
Final Test, Monday April 28, 9:30–11:20 am

400

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

12 → 48

1. The text below shows the headlines and an excerpt from an article that was published in the *CNN* "Health" section at `www.cnn.com`.

> **Study: Men have biological clocks, too**
> *Thursday, February 6, 2003 Posted: 8:51 AM EST (1351 GMT)*
>
> **Women aren't the only ones with a ticking biological clock. A new study adds to the evidence that men's fertility declines with age, too.**
> "If men are choosing to delay fatherhood, they may want to reconsider," said Dr. Brenda Eskenazi, a professor at the University of California at Berkeley and one of the lead researchers of the study. "There may be an impact on the probability that they will be able to father a child." [...]
> Eskenazi and her team looked at 97 men, aged 22 to 80 years, and found that, as men age, the quality of their sperm declines. [...]

More details on this study were published in an article entitled "The association of age and semen quality in healthy men" in the journal *Human Reproduction*, February 2003, where it is stated that a

> "sample of 97 non–smoking men (aged 22-80 years) without known fertility problems was recruited [...]. The men provided semen samples and additional information relating to lifestyle, diet, medical and occupational details."

6

8 (a) (2 points) Based on the information you have, is it an **observational study** or a **controlled experiment**? Circle your answer and explain briefly.

(2) no intervention took place – the men were not told to do anything special

6

8 (b) (2 points) Based on the information you have, is it a **cross–sectional study** or a **longitudinal study**? Circle your answer and explain briefly.

(2) there is no indication that the men had to provide more than just one sample, i.e., there is no clue that these men were followed over time

12 (c) (3 points) From a statistical point of view, why did the men need to provide the additional information relating to lifestyle, diet, medical and occupational details?

8

these factors may be confounding factors that could have an effect on the sperm quality

(4) for general explanation

20 (d) (5 points) Based on the information you have about this study, is the statement "as men age, the quality of their sperm declines" justified? Why or why not? Give two statistical reasons.

(10) No, not justified!

- "association" is not "causation" (but CNN headline suggests causation)
- since this is a cross-sectional study, we do not know whether the sperm quality really declines; older men may have never had a higher sperm quality even when they were younger (due to nutrition, etc.)

(5) for each valid reason

- this study only holds for non-smoking men (at the best); we cannot generalize to all men; smoking (independent from age) may immediately effect sperm quality

9 → 36 2. The regression line for estimating the value of a home from its size (in square feet) has an intercept of −$48,000 and a slope of $80 per square foot.

12 (a) (3 points) Explain why it is OK for the regression line to have a negative intercept.

there are no houses with just a few square feet in size (if we'd ever try to predict the value for a house that is 1 square foot in size, this would be extreme extrapolation); for houses that are of typical size (a few thousand square feet), the predicted value will be positive, i.e., OK

12 (b) (3 points) If possible, predict the value of a Cache Valley home that has 3,800 square feet. If this is not possible, clearly explain why not.

possible: predicted value = $-48{,}000 + 80 \cdot 3{,}800 = \$256{,}000$

12 (c) (3 points) If possible, predict the size (in square feet) of a Cache Valley home that is valued at $250,000. If this is not possible, clearly explain why not.

not possible: we only know the regression line for predicting value from size; we have no information how to predict size from value

$\left(\frac{250{,}000}{80}\right.$ does not work here$\left.\right)$

9 → 36 3. In one region of Cache Valley there are 41 homes for sale. The average listing price of these homes is $284,822, the standard deviation is $248,968 and the median is $209,000. The most expensive home is listed as $1,599,002. **Answers to the following questions are among the 5 choices A,B,C,D and E, and no choice can be used more than once.**

SD: A: $138,847
median: B: $199,450
avg: C: $251,968
D: $283,611
E: $293,466

(a) sum of all 41 homes: $41 \cdot 248{,}968 = 11{,}677{,}702$
sum of 40 homes (most expensive excluded): $11{,}677{,}702 - 1{,}599{,}002 = 10{,}078{,}700$
avg of these 40 homes: $\frac{10{,}078{,}700}{40} = 251{,}968$ → C

(b) the SD must be smaller; it will be considerably effected → A

(c) the median must be smaller; it will only be somewhat effected → B

12 (a) (3 points) If we excluded the most expensive home, the average of the other 40 homes would be A,B,(C),D,E (circle which one). C

12 (b) (3 points) If we excluded the most expensive home, the standard deviation of the other 40 homes would be (A),B,C,D,E (circle which one). A

12 (c) (3 points) If we excluded the most expensive home, the median of the other 40 homes would be A,(B),C,D,E (circle which one). B

−12 if flipped

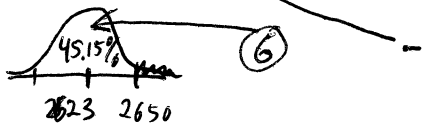
8 → 32 4. (8 points) Cache Valley homes have an average size of 2,623 square feet and an SD of 1,000 square feet. I plan to take a simple random sample of 500 Cache Valley homes. If possible, find the chance that the sample average will be more than 2,650 square feet. If this is not possible, clearly state why not.

possible: box avg: 2,623
box SD: 1,000
draws: 500

(6) $SE_{sum} = \sqrt{500} \cdot 1{,}000 = 22{,}361$

(6) $SE_{avg} = \frac{22{,}361}{500} = 44.7$

S.u.: $\frac{2650 - 2623}{44.7} = \frac{27}{44.7} = 0.60$ (8)



−4 if not SE avg

area above 0.60: $\frac{100\% - 45.15\%}{2} = 27.4\%$ (6)

10 → 40 5. A drawer of socks contains 24 socks of which 5 are black, 10 are blue, and 9 are green. In the dark, a child chooses two socks at random to wear to school.

-2 each calculation error (or no final result)

8 (a) (2 points) What is the chance that the first sock is green?

$\frac{9}{24} = 0.375 = 37.5\%$

8 (b) (2 points) What is the chance that the second sock is green?

$\frac{9}{24} = 0.375 = 37.5\%$ (we don't know anything about the first sock)

8 (c) (2 points) What is the chance that the first sock is blue or black?

(3) $\frac{10}{24} + \frac{5}{24}$ (3) $= \frac{15}{24} = 0.625 = 62.5\%$

(2)

8 (d) (2 points) What is the chance that both the first sock and the second sock are not green?

first not green: $\frac{15}{24}$

second not green, given first not green: $\frac{14}{23}$

first and second not green:

(3) $\frac{15}{24} \cdot \frac{14}{23}$ (3) $= \frac{210}{552} = 0.380 = 38.0\%$

8 (e) (2 points) What is the chance that at least one of the socks is green?

(2)

event (e) is opposite of event (d):

$1 - \frac{210}{552} = \frac{342}{552} = 0.620 = 62.0\%$

10 → 40 6. (10 points) On April 7, 2003, *Time*, page 72, reported the following numbers:

24,645 civilians have been immunized against smallpox

2 of these civilians have died from the vaccine

Suppose a Federal health agency plans to vaccinate all civilians against smallpox. If possible, find a 95% confidence interval for the percentage of all civilians who would die from the vaccine. If it is not possible to construct a valid confidence interval, clearly state all the reasons why not.

-35 if "possible" and CI calculated

(20) not possible: 2 main reasons:

- it is very unlikely that these 24,645 civilians that received the vaccine have been selected via a random sample; these people do not represent the entire population (i.e., all civilians); among these 24,645, there probably won't be too many children, elderly, people with other diseases; more likely, these are people working in a medical field (e.g., doctors, nurses, etc.)

(10) for each valid reason

- the sample % is $\frac{2}{24,645} = 0.000081 = 0.0081\%$

which is very close to 0% and does not allow the construction of a valid CI

10 → [40] 7. (10 points) A simple random sample of 10 people over age 85 take a memory test. The average score for these 10 people is 82.5, with a standard deviation of 8.6. For the general population, the average score is known to be 92.3. Test to see whether this is evidence that people over 85 score lower than the general population, and **clearly state any assumption(s) that you need to make in order to perform the test.**

t-test: • sample size < 30 ✓
• SD for box unknown ✓
• additional assumption: data follows normal curve (4)

-3 if null, alt swapped

1) null: people over 85 on avg have the same score, i.e., avg = 92.3 (2) (1)
alternative: people over 85 on avg have a lower score, i.e., avg < 92.3 (2) (1)

2) $SD^+ = \sqrt{\frac{10}{10-1}} \cdot 8.6 = 1.05 \cdot 8.6 = 9.1$ (4)

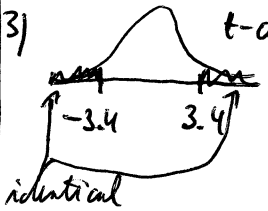
$SE_{sum} = \sqrt{10} \cdot 9.1 = 28.8$ (3)

$SE_{avg} = \frac{28.8}{10} = 2.88$ (3)

$t = \frac{82.5 - 92.3}{2.88} = \frac{-9.8}{2.88} = -3.4$ (4)

$df = 10 - 1 = 9$ (3)

3)



3.25 → 0.5%
3.4 (or -3.4) → less than 0.5% (4)

4) • reject the null (3)
• result is highly stat. significant (3) (p-value < 1%)
• people over 85 on avg have a lower score (3)

12 → [48] 8. (12 points) The following table describes types of collisions in rural and urban settings, for a random sample o f two-vehicle accidents that occurred in a given region last year. Carry out an appropriate test of significance to decide whether the type of collision is independent of whether the accident took place in a rural versus an urban setting. Is the result statistically significant? Is it highly statistically significant? What are your conclusions?

TYPE OF COLLISION

SETTING	angle	rear-end	other	
Urban	40	30	72	142
Rural	6	12	15	33
	46	42	87	175

6 × (2) = (12)

expected:

37	34	71	
9	8	16	

-4 if null, alt swapped

χ^2-test for independence:

1) null: setting and type of collision are independent (3), i.e., boxes are identical (1)
alternative: setting and type of collision are not (3) independent (type of collision depends on setting), i.e., at least one box is different (1)

2) $\frac{142.46}{175} = 37$ etc. (see table "expected" above)

$\chi^2 = \text{sum of } \frac{|obs - exp|^2}{exp} = \frac{|40-37|^2}{37} + \frac{|30-34|^2}{34} + \frac{|72-71|^2}{71}$

$+ \frac{|6-9|^2}{9} + \frac{|12-8|^2}{8} + \frac{|15-16|^2}{16} = 3.79$ (8)

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

3) from χ^2-table: 3.79 between 2.41 and 4.60
p-value between 10% and 30% (6)

4) • do not reject the null (5)
• setting and type of collision are independent (5)

12 → 48

9. (12 points) Four hundred volunteers agree to participate in clinical trial involving a dietary intervention. The investigators want to check how representative this sample is of the general population. One interesting finding is that 40 of the 400 volunteers are current cigarette smokers. Assume that 30% of the general population are current smokers. State the hypotheses needed to test whether the volunteer group is representative of the general population regarding cigarette smoking, compute a test statistic and a P-value and clearly state your conclusion. 1 = smoker, 0 = non-smoker

see FPP, p. 485/486 for a similar example

incorrect test: −30

z-test:

box: [30 × 1, 70 × 0] # draws = 400

box avg = 0.3

box SD = $\sqrt{0.3 \cdot 0.7} = \sqrt{0.21} = 0.46$ (6)

−4 if null, alt swapped

1/ null: volunteer group is representative for entire pop., i.e., vol% = 30% (3) (1)

alternative: volunteer group is not representative for entire pop., i.e., vol% ≠ 30% (3) (1)

2/ $SE_{sum} = \sqrt{400} \cdot 0.46 = 9.2$ (5)

$SE_{\%} = \frac{9.2}{400} \cdot 100\% = 2.3\%$ (5)

$obs\% = \frac{40}{400} = 0.1 = 10\%$

$exp\% = 0.3 = 30\%$

$z = \frac{10\% - 30\%}{2.3\%} = \frac{-20\%}{2.3\%} = -8.7$ (6)

3/ −8.7 off the z-table

⇝ P-value ≈ 0% (6)

4/ • reject the null (4)

• result is highly stat. significant (4)

(P-value < 1%)

• the volunteer group is not representative for the entire population (4)

8 → 32

10. In an article published in The New England Journal of Medicine, April 24, 2003, researchers looked at over 900,000 men and women. The article concludes:

> Increased body weight was associated with increased death rates for all cancers combined and for cancers at multiple specific sites.

On the CNN web site it claims: "Researchers say fat causes 90,000 U.S. cancer deaths a year"

12

(a) (3 points) What mistake did the CNN journalist make? Briefly explain using statistical terminology and concepts.

the CNN journalist mixed up "association" and "causation": association is not the same as causation; perhaps people with an increased body weight exercise less or they eat unhealthier food and these (or other) confounding factors also contribute to a higher cancer rate (8) (4)

20

(b) (5 points) Suppose the researchers looked at 50 different kinds of cancer, and found a P-value for testing the null hypothesis that the cancer is unrelated to body weight versus the alternative that the cancer is associated with body weight. They found 9 statistically significant P-values. Explain why we should be cautious in concluding that all 9 of these are true associations.

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 50 tests, we would expect to reject $50 \cdot 5\% = 50 \cdot 0.05 = 2.5$ (5)

of our null hypotheses just by chance.

Here we found 9 stat. significant P-values; so some of these might indeed represent true associations while others are based on chance.