Name______________________

Stat 1040, Spring 2004
Final Test, Friday April 23rd, 7:30-9:20 am
100 → $\boxed{400}$

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

−2 for each calculation error

9 → $\boxed{36}$ 1. (9 pts) In an experiment, 216 stamped, addressed letters are lost, and the rate of return is recorded. Some of the letters were addressed to Mr. M. J. Davis; some to Dandee Davis, c/o Hooters Club; some to M.J. Davis, c/o Friends of the Communist Party. Test to see if there is a relationship between the addressee and the rate of return. Clearly state the null and alternative hypotheses, calculate the appropriate test statistic, find the p-value, and state your conclusion.

−27 for incorrect test

−3 if null, alt swapped

	Davis	Hooters	Communist	Total
Returned	32	17	29	78
Not Returned	40	55	43	138
Totals	72	72	72	216

expected:

26	26	26
46	46	46

6× 2 = 12

χ^2-test for independence

1/ null: addressee and rate of return are independent, 2
i.e., boxes are identical 1
alternative: addressee and rate of return are not independent (rate of return depends 2 on addressee), i.e., at least one box is different 1

2/ expected: $\frac{78 \cdot 72}{216} = 26$ etc. (see table "expected" above)

$$\chi^2 = \text{sum of } \frac{(\text{obs}-\text{exp})^2}{\text{exp}} = \frac{(32-26)^2}{26} + \frac{(17-26)^2}{26} + \frac{(29-26)^2}{26} + \frac{(40-46)^2}{46} + \frac{(55-46)^2}{46} + \frac{(43-46)^2}{46}$$

$df = (3-1)\cdot(2-1) = 2$ 2 $\quad = 7.59$ 5

3/ $\chi^2 = 7.59$ between 5.99 and 9.21
$\quad\downarrow$ 5% $\quad\downarrow$ 1%

⇝ P-value between 1% and 5% 5

4/ • reject the null (P-value < 5%) 2
• result is statistically significant 2
• rate of return depends on addressee (they are not independent) 2

13 → $\boxed{52}$

2.(6 pts) In a random sample of 200 homes in Jefferson, Illinois (population 25,000), it was found that the average number of televisions sets per home was 3.2 with a standard deviation of 0.9.

$\boxed{24}$ a. Construct a 95% confidence interval for the average number of television set per home in Jefferson, Illinois.

−2 for each calculation error

avg = 3.2

SD = 0.9

$SE_{sum} = \sqrt{200} \cdot 0.9 = 12.73$ (6)

$SE_{avg} = \frac{12.73}{200} = 0.06365 \approx 0.064$ (6)

95% CI: $3.2 \pm 2 \cdot 0.064 = 3.2 \pm 0.128 \approx 3.07$ to 3.33

(2) (2) (2) (2) (2) (2)

$\boxed{8}$ b. T (F) 95% of the homes in Jefferson have between (your lower limit) and (your upper limit) televisions. (2 pts) [False: We are 95% confident that the average falls between these two limits — but this does not mean that 95% of all homes fall between these limits]

$\boxed{8}$ c. T (F) Your interval in part (a) represents a 95% confidence interval for the average number of television set in the sample. (2 pts)

[False: This is a 95% CI for the average number of TVs per home in the entire population.]

$\boxed{12}$ d. (3 pts) Suppose that you find out that televisions sets in Jefferson, Illinois do not follow the normal curve, but have a long right tail. Can you still rely on your results in part (a)? Explain

(8) Yes, we can!

We are looking at the CI for the average. Since the sample size (200) is quite big, we can assume that the distribution of the average follows the normal curve (and thus our results are justified) even if the data do not follow the normal curve. (4)

---

3 → $\boxed{12}$ 3. Roger has tossed a fair coin six times and got heads every time. His friend tells him that his chance of getting a tail on the next toss is almost certain. Do you agree? (3 pts)

(12) No!

[The coin is fair - the chance to get a tail on the next toss remains at 50%.]

−2 for each calculation error

12→ $\boxed{48}$

M S B 50 books

4. A bookstore has a shelf with 15 math books, 10 stat books, and 25 biology books. (2 pts each). Suppose that three books are drawn at random **(without replacement)**:

(4) for multiplication

$\boxed{8}$ a. What is the probability that all three books are stat books?

1st S: $\frac{10}{50}$ (2) 2nd S, given 1st S: $\frac{9}{49}$ (1) 3rd S, given 1st & 2nd S: $\frac{8}{48}$ (1) chance all 3 S: $\frac{10}{50}\cdot\frac{9}{49}\cdot\frac{8}{48} = 0.00612 = \boxed{0.612\%}$

$\boxed{8}$ b. What is the probability the one of each type of book is selected? (2) for multiplication

6 possible orders of the draws: MSB, MBS, SMB, SBM, BMS, BSM (mutually exclusive) | for each order: (2) $\frac{15}{50}\cdot\frac{10}{49}$ (1) $\cdot\frac{25}{48}$ (1) $= 0.03188$ | final result: $6\cdot 0.03188 = 0.1913$ (2)

$\boxed{8}$ c. What is the probability that none of the books selected are stat books?

1st not S: $\frac{40}{50}$ (2) 2nd not S, given 1st not S: $\frac{39}{49}$ (1) 3rd not S, given 1st & 2nd not S: $\frac{38}{48}$ (1) chance all 3 not S: $\frac{40}{50}\cdot\frac{39}{49}\cdot\frac{38}{48} = 0.5041 = \boxed{50.41\%}$ (4) for multiplication

$= \boxed{19.13\%}$

$\boxed{8}$ d. What is the probability that at least one of the books selected is a stat book?

opposite of c.): $1 - 0.5041 = 0.4959 = \boxed{49.59\%}$ (8)

$\boxed{8}$ e. Given that the first book was a biology book, what is the probability that the second book is a stat book?

$\frac{10}{49} = 0.2041 = \boxed{20.41\%}$ (8)

$\boxed{8}$ f. What is the probability that the first book is math book or that the first book is a stat book?

1st M: $\frac{15}{50}$ (3) 1st S: $\frac{10}{50}$ (3) (mutually exclusive)

1st M or S: $\frac{15}{50} + \frac{10}{50} = \frac{25}{50} = \frac{1}{2} = \boxed{50.0\%}$

(2) for addition

6→ $\boxed{24}$ 5. (6 pts) A Logan radio station wishes to determine the reaction of students to President Hall's application to the University of Tennessee. They set up a booth on the quad in front of Widtsoe Hall and pass out surveys from 10 to 12 Thursday morning. Of 700 surveys that were distributed, 206 are returned and of these 169 are glad President Hall is returning. The radio station reports that 82% of USU students are in glad President Hall is returning. Do you agree with the radio station's conclusion? Clearly justify your conclusion.

(12) No! At least 2 major problems:

- non-response bias (only $\frac{206}{700} = 29.4\%$ response rate); so students with a strong opinion are more likely to respond (6)
- not a SRS; selection bias towards location (Widtsoe Hall) and time (students on campus 10–12am on Th) (6)

Need a SRS from all students to be able to make a similar statement.

1: lost weight

0: did not lose weight

13 → 52

−2 for each calculation error

6. A proponent of Dr. Atkins (A) claims that this diet is better than the South Beach (B) diet. In a simple random sample of 400 Atkins diet participants, 300 lost weight and in a simple random sample of 500 South Beach dieters, 360 lost weight. Can we conclude from these data that the Atkins diet is better than the South Beach diet?

−30 for incorrect test

−3 if null, alt swapped

40

a. (10 pts) Clearly state the null and alternative hypotheses, calculate the appropriate test statistic, find the p-value, and state your conclusion.

2-sample z-test:

1) null: Atkins and South Beach diet work equally well, i.e., $lose_{A\%} - lose_{B\%} = 0\%$ 3 1

alternative: Atkins works better than South Beach diet, i.e., $lose_{A\%} - lose_{B\%} > 0\%$ 3 1

2)

A (Atkins)		B (South Beach)	
sample size: 400		sample size: 500	
$\text{sample}\%_A = \frac{300}{400} = 75\%$	1	$\text{sample}\%_B = \frac{360}{500} = 72\%$	1
$SD_A = \sqrt{.75 \cdot .25} = 0.433$	2	$SD_B = \sqrt{.72 \cdot .28} = 0.449$	2
$SE_{sum_A} = 0.433 \cdot \sqrt{400} = 8.66$	2	$SE_{sum_B} = 0.449 \cdot \sqrt{500} = 10.04$	2
$SE_{\%A} = \frac{8.66}{400} \cdot 100\% = 2.165\%$	2	$SE_{\%B} = \frac{10.04}{500} \cdot 100\% = 2.01\%$	2

$SE_{diff\%} = \sqrt{(2.165\%)^2 + (2.01\%)^2} = 2.55\%$ 4

$z = \frac{75\% - 72\%}{2.55\%} = 1.02$ 4

3) area between −1.00 and 1.00: 68.27%

→ P-value: $\frac{100\% - 68.27\%}{2} = 15.9\%$ 4

4) • do not reject the null (P-value > 5%) 3

• both diets work equally well 3

12

b. (3 pts) Is this an observational study or controlled experiment? Justify your choice. 9

No intervention took place − participants chose diet themselves. 3

---

4 → 16

7. (4 pts) A marketing firm wants to determine whether Barbie dolls are still a good market item. To estimate the percentage of people who want to buy one, the firm takes a simple random sample of 500 people in Utah and a simple random sample of 500 people in California. Other things being equal,

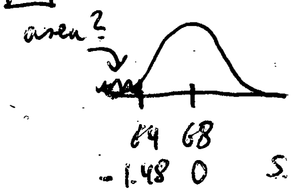
16 a. The accuracy in Utah will be about the same as the accuracy in California.
b. The accuracy in Utah will be less than the accuracy in California.
c. The accuracy in Utah will be more than the accuracy in California.
d. Cannot be determined from the information given.

[a.) The accuracy depends on the size of the sample and not on the size of the population.]

8→ $\boxed{32}$

8. In Pearson's data about fathers and sons, fathers had an average height of 68 inches with a standard deviation of 2.7 inches. (4 points each)

$\boxed{16}$ a. Suppose one father is 64 inches tall, express his height as a percentile.

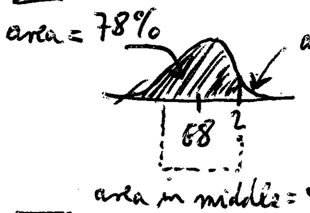


S.u.: $\frac{64-68}{2.7} = -1.48$ (6)

area from -1.5 to 1.5: 86.64% (4)

area below -1.5: $\frac{100\% - 86.64\%}{2} = 6.68\%$ (4), i.e., about 7th percentile (2)

$\boxed{16}$ b. Another father is in the 78[th] percentile, express his height in inches.



(4)

area from -0.75 to 0.75 (4): 54.67% (closest to 56%)

original units: $0.75 \cdot 2.7 + 68 = 70.0$ inches
(2) (2) (2) (2)

16→ $\boxed{64}$ 9. For men 18 – 24 in the HANES sample the following is recorded:

x average height = 70 inches SD = 2.7 inches

y average weight = 162 pounds SD = 30 pounds r = 0.6

−2 if x, y not stated

−4 if x, y swapped

$\boxed{24}$ a. (6 pts) Calculate the regression equation for predicting weight from height.

slope $= r \cdot \frac{SD_y}{SD_x} = 0.6 \cdot \frac{30}{2.7} = 6.67$ (4) (3) (3)

intercept = avg y − slope · avg x = $162 - 6.67 \cdot 70 = -304.9$ (2) (2) (2) (2) (2)

regression equation: $\boxed{y = -304.9 + 6.67 \cdot x}$ or $\boxed{\text{weight} = -304.9 + 6.67 \cdot \text{height}}$ (4)

$\boxed{12}$ b. (3 pts) Predict the weight of a man who is 66 inches tall.

for height = 66:

weight $= -304.9 + 6.67 \cdot 66 = 135.32$ pounds (12)

$\boxed{12}$ c. (3 pts) About how far off do you expect your prediction to be?

r.m.s. error $= \sqrt{1 - r^2} \cdot SD_y = \sqrt{1 - 0.6^2} \cdot 30 = \sqrt{0.64} \cdot 30$
(3) (3) (3) (3)
$= 0.8 \cdot 30 = 24$ pounds

$\boxed{16}$ d. (4 pts) Men who were 73 inches tall averaged about 182 pounds. True or false and explain, the ones who weighed 182 pounds averaged about 73 inches tall. (12)

We cannot simply flip over the regression equation but we have to calculate a new regression equation for predicting height from weight! (4)

-2 for each calculation error
-24 for incorrect test
-3 if null, alt swapped

8→ $\boxed{32}$

10. (8 pts) Suppose that the ages at which children first walk alone are normally distributed with mean 11.5 months. A simple of sample of 12 children is selected from children having low birth weights. The average age these children first walked alone is 12.2 months with a standard deviation of 1.1. Can we conclude from these data that children with low birth weights walk later than the average? Clearly state the null and alternative hypotheses, calculate the appropriate test statistic, find the p-value, and state your conclusion.

t-test: • sample size < 30 ✓
• SD for box unknown ✓
• data follows normal curve ✓

1) null: children with low birthweights walk at same age, (2)
i.e., $box_{avg} = 11.5$ (1)
alternative: children with low birthweights walk later, (2)
i.e., $box_{avg} > 11.5$ (1)

2) $SD^+ = \sqrt{\frac{12}{11}} \cdot 1.1 = 1.044 \cdot 1.1 = 1.15$ (3)
$SE_{sum} = \sqrt{12} \cdot 1.15 = 3.46 \cdot 1.15 = 3.98$ (2)
$SE_{avg} = \frac{3.98}{12} = 0.33$ (2)

$t = \frac{12.2 - 11.5}{0.33} = 2.12$ (3)
$df = 12 - 1 = 11$ (3)

3) $t = 2.12$ between 1.80 and 2.20
$\downarrow$ 5% $\quad \downarrow$ 2.5%
⇝ P-value between 2.5% and 5% (4)

4) • reject the null (P-value < 5%) (3)
• result is statistically significant (3)
• children with low birthweights walk later (3)

8→ $\boxed{32}$

11. (8 pts) Many companies are experimenting with "flex-time," which is supposed to reduce absenteeism. One company employees have averaged 6.3 days off work in the past. The company introduces "flex-time" and a year later a simple random sample of 100 employees is selected. They average 5.5 days off work with a standard deviation of 2.9. Test to determine if flex-time reduces absenteeism. Clearly state the null and alternative hypotheses, calculate the appropriate test statistic, find the p-value, and state your conclusion.

-2 for each calculation error
-24 for incorrect test
-3 if null, alt swapped

z-test: • sample size ≥ 30

1) null: flex-time has no effect on absenteeism, (3) i.e., $box_{avg} = 6.3$ (1)
alternative: flex-time reduces absenteeism, (3) i.e., $box_{avg} < 6.3$ (1)

2) $SE_{sum} = \sqrt{100} \cdot 2.9 = 10 \cdot 2.9 = 29$ (3)
$SE_{avg} = \frac{29}{100} = 0.29$ (3)
$z = \frac{5.5 - 6.3}{0.29} = -2.76$ (4)

3) area between -2.75 and 2.75: 99.40%
⇝ P-value: $\frac{100\% - 99.40\%}{2} = 0.3\%$ (4)

4) • reject the null (P-value < 5%) (4)
• result is highly statistically significant (3)
• flex-time reduces absenteeism (3)