Name:

STAT 1040, SPRING 2002
Final Test, Tuesday April 30, 7:00–8:50 am

[100]

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

[20] 1. (5 points) To study the effectiveness of vitamin C in preventing colds, 200 volunteers were recruited. The researcher randomly assigned 100 of them to take vitamin C for 10 weeks and the remaining 100 to take nothing. The 200 participants recorded how many colds they had during the 10 weeks. The two groups were compared, and the researcher announced that taking vitamin C reduces the frequency of colds. Is there any problem with the design of the study? If your answer is yes, identify the problem and briefly describe how to properly design this particular study.

(8) Yes, this is not really a controlled experiment. There is no "control group" – people in the (missing) "control group" received "nothing, i.e., no placebo. (1) A placebo pill that looks similar but is ineffective should have been given to the "control group", resulting in a blind (or double-blind) experiment

explanation: (5)

[24] 2. (6 points) A local politician who has to run for office sees the following article from The Utah Statesman (March 27 2002):

Online poll

Do you think the state's current liquor laws need to be changed? (Current results in bold.)

- **Yes, the laws restrict the freedoms of the citizens of the state. (42%)**
- **No, the laws are not any tougher than those found in other states. (38%)**
- **I don't really care either way. (11%)**
- **I need more information on the subject. (2%)**
- **None of the above (7%)**

Visit us on the Web at *www.utahstatesman.com* to cast your vote and see results from past Utah Statesman online polls.

- Voluntary answers - only people that knew about this Web site and could access it before the outcome was published had a chance to indicate their opinion.
- The population of people that might answer the opinion poll is not identical with the population that may vote during the next election (e.g., out-of-state students, people who find the poll but do not live in Utah, etc.) and there → is certainly a bias towards students and younger people.
- People can vote more than once - so people who have a strong opinion and would like the liquor laws to be changed may vote many times on this opinion poll.
- We don't know the "sample" size – is 42% really significantly more than 38% ? And what about the 20% non-response ?

each explanation: (4)

She wonders whether changing the state's liquor laws would be a good thing to have as part of her campaign platform, hoping to get more votes by supporting what the majority of people think. What do *you* think? Should she support a change in the current liquor laws to boost her election chances? Answer yes or no and give 3 different reasons to support your answer, based on information in the Statesman poll, not on your personal opinion.

No!

(12)

[20] 3. (5 points) A large department store wants to know if consumers would be willing to pay slightly higher prices to have computers available throughout the store to help them locate items. The store posted an interviewer at the door and told her to collect a sample of 100 opinions by asking the next person who came in the door each time she had finished an interview. Is this a simple random sample, a cluster sample, or a sample of convenience? (10) Why? Comment on the reliability of the information she would obtain.

This is a sample of convenience (or "convenience sample"). The interviewer takes the next available person, i.e. only people that came to the store that time of the day. The information will not be very reliable because the interviewer most likely will only interview people that look approachable and seem to have time to answer questions. There will certainly be some bias (5) towards such customers.

explanation: (5)

[32] 4. A 10-sided die shows the numbers 1 through 10 with equal probability. A 6-sided die shows the numbers 1 through 6 with equal probability.

[16] (a) (4 points) Which is more likely, and why?

- Getting double-six when you roll two 6-sided dice
- Getting double-ten when you roll two 10-sided dice

(6)

$\frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$ (chance of double-6)

greater than (6)

$\frac{1}{10} \cdot \frac{1}{10} = \frac{1}{100}$ (chance of double-10)

(4)

Double-6 more likely than double-10.

[16] (b) (4 points) If I roll one 6-sided die and one 10-sided die, what is the chance that the total number of spots is 10?

6-sided: 1 2 3 4 5 6

10-sided: 1 2 3 4 5 6 7 8 9 10

- $6 \cdot 10 = 60$ possible combinations (6)
- 6 combinations result in "10": (6)
  1-9, 2-8, 3-7, 4-6, 5-5, 6-4
- chance of "10": $\frac{6}{60} = \frac{1}{10} = 10\%$ (4)

[40] 5. (10 points) There are approximately 20,000 students at USU. For a simple random sample of 500 USU students, we learn that 80% are satisfied with President Kermit Hall. If possible, construct a 95% confidence interval for the percentage of all USU students who are satisfied with President Hall. If you cannot do this, say why not.

1: satisfied
0: not satisfied

box: [? × 0, ? × 1]

number of draws: 500

"SD box" = $\sqrt{0.8 \cdot 0.2} = \sqrt{0.16} = 0.4$ (10)

$SE_{sum} = \sqrt{500} \cdot 0.4 = 8.9$ (5)

$SE_{\%} = \frac{8.9}{500} \cdot 100\% = 1.78\% \approx 1.8\%$ (10)

95% CI: $80\% \pm 2 \cdot 1.8\% = 80\% \pm 3.6\%$: 76.4% to 83.6%

(5) (5) (5)

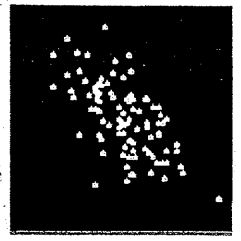
−1 if no final result

16 6. (4 points) Match the four scatterplots with their correlations from the list:

$-1.03, \ -0.98, \ -0.62, \ 0.05, \ 0.65, \ 0.80, \ 0.98, \ 1.03$

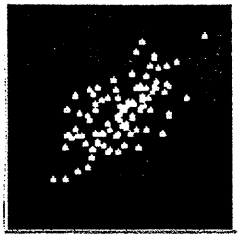
4: Correct
2: for / unit off
1: more than / unit off
0: -1.03 or 1.03

Plot A



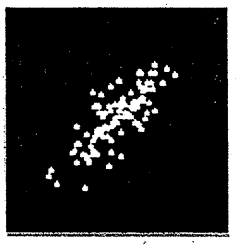
r = −0.62

Plot B



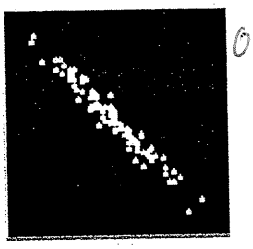
r = 0.65

Plot C



r = 0.80

Plot D



r = −0.98

68 7. For a random sample of 20 car models, the average weight (in pounds) was 3236, with an SD of 523. The average gas mileage (in miles per gallon) was 21.4 with an SD of 4.2. The correlation between weight and gas mileage was -0.87. The scatter diagram was football shaped.

16 (a) (4 points) Assuming the histogram for gas mileage follows the normal curve, would you be surprised if someone told you that one of these cars got 27 miles per gallon? Explain your reasoning.

(5) calculation

$\frac{27.0-21.4}{4.2} = \frac{5.6}{4.2} = 1.33$ s.u.

	avg	SD
x: weight	3236	523
y: mileage	21.4	4.2

r = −0.87

(6) No, 27 miles per gallon is just 1.33 s.u. above the average and not unusual at all. (5) explanation

12 (b) (3 points) Find the equation of the regression line for predicting gas mileage from weight.

(5) slope $= r \cdot \frac{SD_y}{SD_x} = -0.87 \cdot \frac{4.2}{523} = -0.00698 \approx -0.007$

(5) intercept: $21.4 - (-0.007) \cdot 3236 = 44.052 \approx 44.1$

(2) equation: mileage = 44.1 − 0.007 · weight

16 (c) (4 points) Predict the gas mileage of a car that weighs 3500 pounds.

mileage for 3500 pounds: $44.1 - 0.007 \cdot 3500 = 19.6$

(14) (2)

16 (d) (4 points) Would you be surprised if someone told you that one of these cars weighing 3500 pounds got 27 miles per gallon? Explain your reasoning, using the rms error.

r.m.s. error $= \sqrt{1-r^2} \cdot SD_y = \sqrt{1-(-0.87)^2} \cdot 4.2 = \sqrt{0.2431} \cdot 4.2 = 0.493 \cdot 4.2 = 2.07$ (10) (2)

$\frac{27-19.6}{2.07} = \frac{7.4}{2.07} = 3.57$ s.u.; 27 is about 3.6 r.m.s. errors above the predicted value of 19.6. So, yes, this is very unusual... (2) (2)

8 (e) (2 points) In one sentence, explain what the correlation coefficient tells you about the relationship between gas mileage and weight of cars like these.

(8)

There is a strong negative association (perhaps even causation?) between weight and mileage, i.e., the heavier a car, the less mileage it has.

or (8)

72

8. For a simple random sample of 225 patients from Logan Regional Hospital, the average cholesterol count was 230 with an SD of 30. For a simple random sample of 400 patients from the University of Utah Hospital, the average cholesterol count was 240 with an SD of 40. You may assume that there are 5000 patients at each hospital.

   The question we want to consider is whether or not the average cholesterol count of all 5000 patients at Logan Regional Hospital is the same as the average cholesterol count of all 5000 patients at the U of U Hospital.

16 (a) (4 points) Formulate appropriate box models - indicate the number of tickets in each box, what the numbers on the tickets represent, and say whether you do or do not know the average and SD of each box.

Logan

| individual tickets unknown | (4)

5000 tickets

U of U

| individual tickets unknown | (4)

5000 tickets

- tickets represent cholesterol count for each person in the population (4)
- box avg and box SD are unknown (4)

8 (b) (2 points) Clearly state the null and alternative hypotheses in terms of your box models.

Null: average cholesterol count the same at Logan and at U of U, i.e., $\text{avg}_{\text{Logan}} = \text{avg}_{\text{U of U}}$ (3) (1)

Alt: average cholesterol count different, i.e., $\text{avg}_{\text{Logan}} \neq \text{avg}_{\text{U of U}}$ (3) (1)

16 (c) (4 points) Compute an appropriate test statistic.

Logan

avg = 230

(1) $SE_{sum} = \sqrt{225} \cdot 30 = 15 \cdot 30 = 450$

(2) $SE_{avg} = \frac{450}{225} = 2$

U of U

avg = 240

(1) $SE_{sum} = \sqrt{400} \cdot 40 = 20 \cdot 40 = 800$

(2) $SE_{avg} = \frac{800}{400} = 2$

(5) $SE_{diff} = \sqrt{2^2 + 2^2} = \sqrt{8} = 2.83$

(5) $z = \frac{230 - 240}{2.83} = \frac{-10}{2.83} = -3.5$

8 (d) (2 points) Find the P-value.

-3 for 1-sided test

from z-table: -3.5 to 3.5: 99.953% (4)

p-value: 100% - 99.953% = 0.047% (4) < 1%

12 (e) (3 points) Using the 5% level, say whether or not you reject the null hypothesis and state your conclusion in terms of the original question.

(6) • reject the null hypothesis - the result is highly statistically significant (p-value < 1%)

(6) • the average cholesterol count is different at Logan and at U of U (it is smaller at Logan)

12 (f) (3 points) Did you assume that the histogram for the patients' cholesterol counts followed the normal curve? Why or why not?

(8) No - the normal curve assumption for the box is not required since we are working with averages of a large number of draws (sample sizes are 225 and 400).

(4) Yes - the probability distribution of the averages does follow the normal curve.

48 9. (12 points) The table below shows the number of computers per household for a simple random sample of 600 households from Cache Valley and Salt Lake City.

χ^2-test for independence

Number of Computers	0	1	2	Total
Cache Valley	30	50	20	100
Salt Lake City	160	300	40	500
Total	190	350	60	600

observed:

expected: | 32 | 58 | 10 | / | 158 | 292 | 50 | (12)

For Cache Valley and Salt Lake City, is the number of computers per household independent of location? 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.

(1) Null: location and #computers are independent, i.e. losses are identical (4)
Alt: location and #computers are not independent (#computers depends on location), i.e., losses different (4)

(2) $\frac{100\cdot 190}{600} = 32$ etc. (see table "expected" above)

χ^2 = sum of $\frac{(obs-exp)^2}{exp} = \frac{(30-32)^2}{32} + \frac{(50-58)^2}{58} + \frac{(20-10)^2}{10}$

$+ \frac{(160-158)^2}{158} + \frac{(300-292)^2}{292} + \frac{(40-50)^2}{50} = 13.47$ (8)

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

(3) from χ^2-table: 9.21 ⇝ 1%
13.47 ⇝ less than 1%; p-value < 1% (6)

(4)(5) reject the null hypothesis – the result is highly statistically significant (p-value < 1%);
(5) the number of computers depends on location

60 10. In the United States, the average age of men at the time of their first marriage is 24.8 years. To determine if Cache Valley men marry for the first time at an earlier average age than men in the country as a whole, a sociologist took a simple random sample of 24 male Cache Valley residents who were or had been married and found that their average age at first marriage was 23.5 years with a standard deviation of 3.2 years.

t-test

48 (a) (12 points) Did this survey show that Cache Valley men really marry for the first time at an earlier average age than the men in the country as a whole? Assume the histogram of mens' ages at first marriage closely follows the normal curve. Clearly state the null and the alternative hypothesis, conduct an appropriate test, and state your conclusions.

(1) Null: CV men marry at the same age, i.e, $avg_{CV} = 24.8$ (3) (1)
Alt: CV men marry at an earlier age, i.e., $avg_{CV} < 24.8$ (3) (1)

(2)(6) $SD^+ = \sqrt{\frac{24}{24-1}} \cdot 3.2 = 1.02 \cdot 3.2 = 3.26$

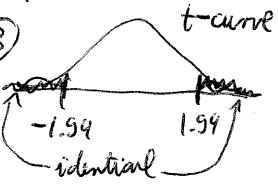
(2) $SE_{sum} = \sqrt{24} \cdot 3.26 = 15.97$

(4) $SE_{avg} = \frac{15.97}{24} = 0.67$

$t = \frac{23.5 - 24.8}{0.67} = -1.94$ (6)

$df = 24 - 1 = 23$ (4)

(3)



1.71 ⇝ 5%
2.07 ⇝ 2.5%
1.94 (or −1.94) ⇝ p-value between 2.5% and 5% (6)

12 (b) (3 points) Would your test in part (a) be valid even if the histogram of the mens' ages at first marriage had a very long right-hand tail? Why or why not? Explain briefly.

(8) Not valid: t-test requires:
1) sample size < 30
2) "normal" use
3) SD unknown
In case of a long right-hand tail, 2) does not hold. (4)

(4) reject the null hypothesis – the result is statistically significant (p-value < 5%); (6)
CV men marry at an earlier age (6)

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

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

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

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

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

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

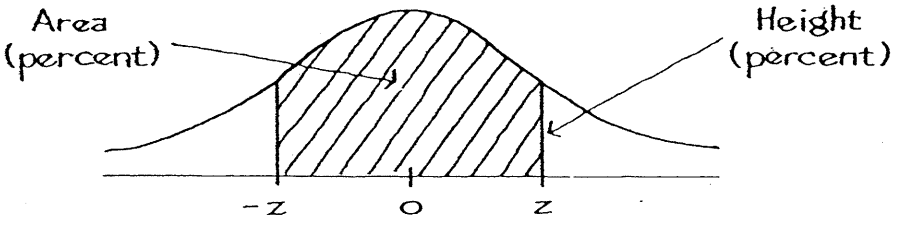
$$EV_{\text{sum}} = \text{number of draws} \times \text{ave}_{\text{box}}$$
$$SE_{\text{sum}} = \sqrt{\text{number of draws}} \times SD_{\text{box}}$$

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

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

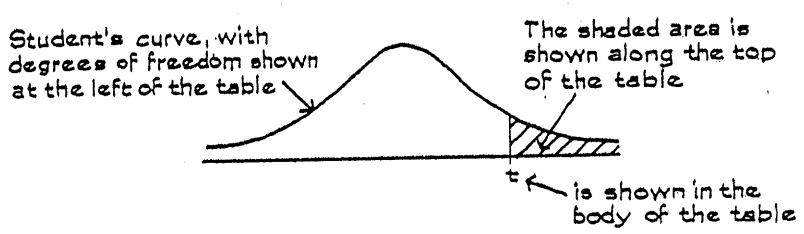
$SE_{\text{diff}} = \sqrt{a^2 + b^2}$ where a is the SE for the first quantity,
b is the SE for the second quantity, and the two quantities are independent

A NORMAL TABLE



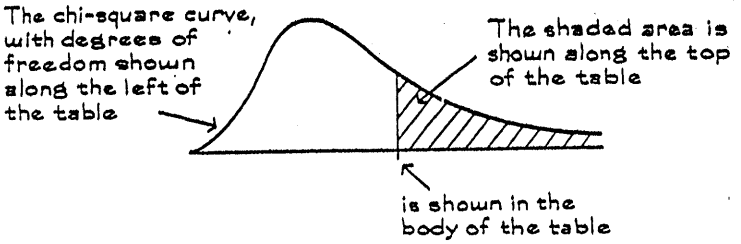
z	*Area*	*z*	*Area*	*z*	*Area*
0.00	0	1.50	86.64	3.00	99.730
0.05	3.99	1.55	87.89	3.05	99.771
0.10	7.97	1.60	89.04	3.10	99.806
0.15	11.92	1.65	90.11	3.15	99.837
0.20	15.85	1.70	91.09	3.20	99.863
0.25	19.74	1.75	91.99	3.25	99.885
0.30	23.58	1.80	92.81	3.30	99.903
0.35	27.37	1.85	93.57	3.35	99.919
0.40	31.08	1.90	94.26	3.40	99.933
0.45	34.73	1.95	94.88	3.45	99.944
0.50	38.29	2.00	95.45	3.50	99.953
0.55	41.77	2.05	95.96	3.55	99.961
0.60	45.15	2.10	96.43	3.60	99.968
0.65	48.43	2.15	96.84	3.65	99.974
0.70	51.61	2.20	97.22	3.70	99.978
0.75	54.67	2.25	97.56	3.75	99.982
0.80	57.63	2.30	97.86	3.80	99.986
0.85	60.47	2.35	98.12	3.85	99.988
0.90	63.19	2.40	98.36	3.90	99.990
0.95	65.79	2.45	98.57	3.95	99.992
1.00	68.27	2.50	98.76	4.00	99.9937
1.05	70.63	2.55	98.92	4.05	99.9949
1.10	72.87	2.60	99.07	4.10	99.9959
1.15	74.99	2.65	99.20	4.15	99.9967
1.20	76.99	2.70	99.31	4.20	99.9973
1.25	78.87	2.75	99.40	4.25	99.9979
1.30	80.64	2.80	99.49	4.30	99.9983
1.35	82.30	2.85	99.56	4.35	99.9986
1.40	83.85	2.90	99.63	4.40	99.9989
1.45	85.29	2.95	99.68	4.45	99.9991

A *t*-TABLE



Degrees of freedom	*25%*	*10%*	*5%*	*2.5%*	*1%*	*0.5%*
1	1.00	3.08	6.31	12.71	31.82	63.66
2	0.82	1.89	2.92	4.30	6.96	9.92
3	0.76	1.64	2.35	3.18	4.54	5.84
4	0.74	1.53	2.13	2.78	3.75	4.60
5	0.73	1.48	2.02	2.57	3.36	4.03
6	0.72	1.44	1.94	2.45	3.14	3.71
7	0.71	1.41	1.89	2.36	3.00	3.50
8	0.71	1.40	1.86	2.31	2.90	3.36
9	0.70	1.38	1.83	2.26	2.82	3.25
10	0.70	1.37	1.81	2.23	2.76	3.17
11	0.70	1.36	1.80	2.20	2.72	3.11
12	0.70	1.36	1.78	2.18	2.68	3.05
13	0.69	1.35	1.77	2.16	2.65	3.01
14	0.69	1.35	1.76	2.14	2.62	2.98
15	0.69	1.34	1.75	2.13	2.60	2.95
16	0.69	1.34	1.75	2.12	2.58	2.92
17	0.69	1.33	1.74	2.11	2.57	2.90
18	0.69	1.33	1.73	2.10	2.55	2.88
19	0.69	1.33	1.73	2.09	2.54	2.86
20	0.69	1.33	1.72	2.09	2.53	2.85
21	0.69	1.32	1.72	2.08	2.52	2.83
22	0.69	1.32	1.72	2.07	2.51	2.82
→ 23	0.69	1.32	1.71	2.07	2.50	2.81
24	0.68	1.32	1.71	2.06	2.49	2.80
25	0.68	1.32	1.71	2.06	2.49	2.79

A CHI-SQUARE TABLE



Degrees of freedom	*99%*	*95%*	*90%*	*70%*	*50%*	*30%*	*10%*	*5%*	*1%*
1	0.00016	0.0039	0.016	0.15	0.46	1.07	2.71	3.84	6.64
→ 2	0.020	0.10	0.21	0.71	1.39	2.41	4.60	5.99	9.21 ←
3	0.12	0.35	0.58	1.42	2.37	3.67	6.25	7.82	11.34
4	0.30	0.71	1.06	2.20	3.36	4.88	7.78	9.49	13.28
5	0.55	1.14	1.61	3.00	4.35	6.06	9.24	11.07	15.09
6	0.87	1.64	2.20	3.83	5.35	7.23	10.65	12.59	16.81
7	1.24	2.17	2.83	4.67	6.35	8.38	12.02	14.07	18.48
8	1.65	2.73	3.49	5.53	7.34	9.52	13.36	15.51	20.09
9	2.09	3.33	4.17	6.39	8.34	10.66	14.68	16.92	21.67
10	2.56	3.94	4.86	7.27	9.34	11.78	15.99	18.31	23.21
11	3.05	4.58	5.58	8.15	10.34	12.90	17.28	19.68	24.73
12	3.57	5.23	6.30	9.03	11.34	14.01	18.55	21.03	26.22
13	4.11	5.89	7.04	9.93	12.34	15.12	19.81	22.36	27.69
14	4.66	6.57	7.79	10.82	13.34	16.22	21.06	23.69	29.14
15	5.23	7.26	8.55	11.72	14.34	17.32	22.31	25.00	30.58
16	5.81	7.96	9.31	12.62	15.34	18.42	23.54	26.30	32.00
17	6.41	8.67	10.09	13.53	16.34	19.51	24.77	27.59	33.41
18	7.00	9.39	10.87	14.44	17.34	20.60	25.99	28.87	34.81
19	7.63	10.12	11.65	15.35	18.34	21.69	27.20	30.14	36.19
20	8.26	10.85	12.44	16.27	19.34	22.78	28.41	31.41	37.57

Source: Adapted from p. 112 of Sir R. A. Fisher, *Statistical Methods for Research Workers* (Edinburgh: Oliver & Boyd, 1958).