Name:

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

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

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."

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

(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.

(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?

(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.

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.

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

   (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.

   (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.

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.**

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

   (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).

   (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).

   (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).

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.

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.

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

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

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

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

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

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.

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.**

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

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.

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"

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

(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.

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

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

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

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

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

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

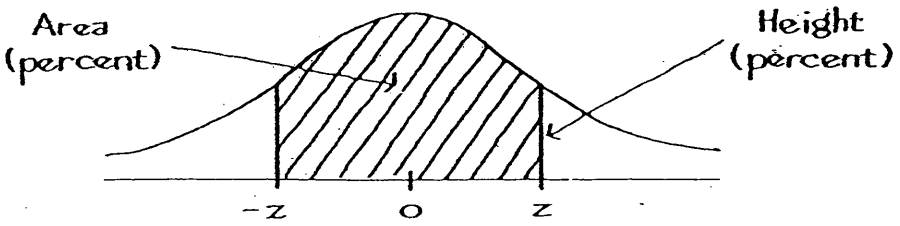
$$EV_{\text{sum}} = \text{number of draws} \times \text{ave}_{\text{box}}$$
$$SE_{\text{sum}} = \sqrt{\text{number of draws}} \times SD_{\text{box}}$$

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

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

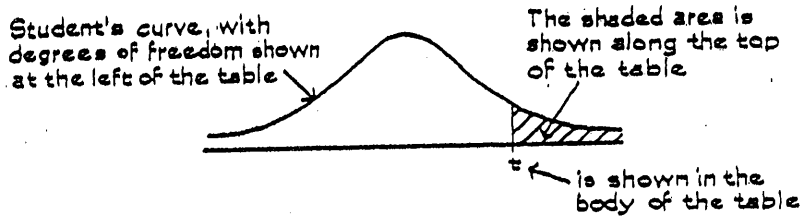
$SE_{\text{diff}} = \sqrt{a^2 + b^2}$ where a is the SE for the first quantity,
b is the SE for the second quantity, and the two quantities are independent

A NORMAL TABLE



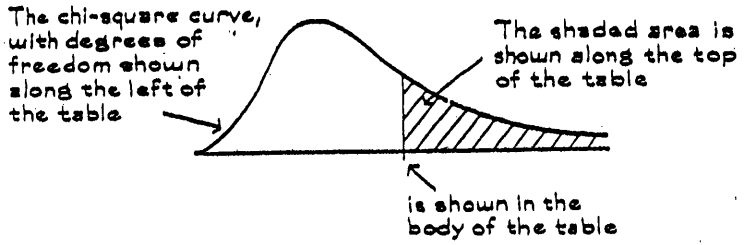
z	*Area*	*z*	*Area*	*z*	*Area*
0.00	0	1.50	86.64	3.00	99.730
0.05	3.99	1.55	87.89	3.05	99.771
0.10	7.97	1.60	89.04	3.10	99.806
0.15	11.92	1.65	90.11	3.15	99.837
0.20	15.85	1.70	91.09	3.20	99.863
0.25	19.74	1.75	91.99	3.25	99.885
0.30	23.58	1.80	92.81	3.30	99.903
0.35	27.37	1.85	93.57	3.35	99.919
0.40	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).