Name:

Stat 1040, Fall 1998
Final Test, Tuesday 15 December, 3:30–5:20

Show your work. The test is out of 100 points and you have 110 minutes, so budget your time accordingly.

1. In your town, there is a unique gumball machine. In the Holiday season, it contains red gumballs and green gumballs. It is constantly churning to mix the gumballs. There are a *very* large number of gumballs and 40% are red, the rest green.

   (a) (5 points) If you buy 2 gumballs, what is the chance that they are both the same color?

   RR $\quad .4 \times .4 = .16$

   \+

   GG $\quad .6 \times .6 = .36$

   so the chance is $.52$.

   (b) (5 points) If you buy 8 gumballs, what is the chance that 5 will be red and 3 will be green?

   $n = 8$

   $p = .4$ or $p = .6$

   $k = 5$ $\quad k = 3$

   $$\frac{8!}{5!3!}(.4)^5(.6)^3 = 56 \times .010 \times .216 = .1239$$

2. Suppose that 60% of all people who are eligible for jury duty in a large city are in favor of capital punishment. We are interested in how this fact might affect the composition of a jury in a murder trial. Suppose a jury of 12 is to be randomly selected from all who are eligible for jury duty in that city.

   (a) (5 points) What is the chance that none of the 12 jurors selected favors capital punishment? $(.4)^{12} = .000017$

   (b) (3 points) If the jury were selected at random, would you be surprised if none of the 12 jurors selected favored capital punishment? Explain.

   Yes - it's very unlikely that you'd get none. I'd suspect that they were <u>not</u> chosen at random.

3. (5 points) The correlation between two IQ tests is 0.8. Both tests have an average of 100 and an SD of 15. If someone scores 75 on the first test, which is the most likely:

   (a) the person will score 75 on the second test.

   (b) the person will score somewhat less than 75 on the second test.

   (c) the person will score somewhat more than 75 on the second test.

   (Choose one option and explain briefly). The regression effect says that a low score will tend to improve somewhat.

4. (5 points) This year, the toy "Furbies" are selling out everywhere. To estimate the percentage of people who want to buy one, a market research company plans to sample 500 people in Salt Lake City and 500 people in Logan. Other things being equal,

   (a) The accuracy in Logan will be about the same as the accuracy in Salt Lake City.

   (b) The accuracy in Logan will be quite a bit less than the accuracy in Salt Lake City.

   (c) The accuracy in Logan will be quite a bit more than the accuracy in Salt Lake City.

   (Choose one option and explain briefly).

   Only the sample size determines accuracy.

5. In a study, reading comprehension is tested for a large group of third grade students, once at the beginning of the school year and once at the end of the school year. During the school year, the students work on reading comprehension skills. The following results are obtained:

   x beginning of year: average score = 75, SD = 15
   y end of year: average score = 80, SD = 17, $r = 0.6$.

   Assume that the scatter plot of the data shows a football shaped cloud.

   (a) (7 points) Write the equation of the regression line for predicting the end-of-year score from the beginning-of-year score.

   $$\text{slope} = r\,\frac{SD_y}{SD_x} = .6 \times \frac{17}{15} = .68$$

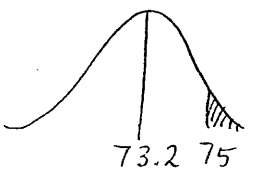
   $$\text{intercept} = \text{ave}_y - \text{slope} \cdot \text{ave}_x$$
   $$= 80 - .68(75)$$
   $$= 29$$

   So the equation is end-of-year = 29 + .68 · beginning-of-year

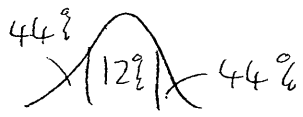
   (b) (8 points) For those students who scored 65 on the beginning-of-year score, what percentage scored 75 or higher on the end-of-year score?

   $29 + (.68 \times 65) = 73.2$ new average

   rms error $= \sqrt{1 - .6^2} \times 17 = 13.6$

   $$\frac{75 - 73.2}{13.6} = .13$$

   So the percentage should be around 44%.

6. (7 points) A night time cold medicine "Relief" contains acetaminophen. A random sample of 65 capsules from a large batch of "Relief" has an average acetaminophen content of 589 mg per capsule with an SD of 21 mg. Find a 95% confidence interval for the average acetaminophen content per capsule for the batch.

$SE_{sum} = \sqrt{65} \times 21 = 169$

$SE_{ave} = 2.6$

CI: $589 \pm 2 \times 2.6$ mg per capsule

i.e. 589 ± 5.2

7. (10 points) This question concerns a study of fulltime workers age 25–54 in Cache Valley. A simple random sample of 400 such people with high school degrees had an average income of \$22,500 and an SD of \$16,000. A simple random sample of 225 such people with college degrees had an average income of \$31,400 and an SD of \$17,500. For fulltime workers age 25–54 in Cache Valley, is there evidence that there is a difference between the average incomes of those who have a high school degree and those who gave a college degree? Set up a null and alternative hypotheses, perform the test, and clearly state your conclusions.

null: the average income of all fulltime workers age 25-54 in Cache Valley who have high school degrees is the same as for those who have college degrees.

alt: the two population averages are different.

high school: $SD_{box} \approx 16000$

$SE_{sum} = \sqrt{400} \times 16000 = 320000$

$SE_{ave} = 800$

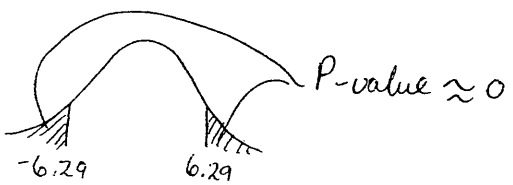
college: $SD_{box} \approx 17500$

$SE_{sum} = \sqrt{225} \times 17500 = 262500$

$SE_{ave} = 1166$

$SE_{diff} = \sqrt{800^2 + 1166^2} = 1414$

$z = \dfrac{22500 - 31400}{1414} = -6.29$



Since the P-value is small we reject the null hypothesis and conclude that the population averages are different.

8. (12 points) Many scientists believe that alcoholism is linked to social isolation. One measure of social isolation is marital status, i.e., whether a person is married or not. To test the notion that alcoholics are socially isolated, 280 adults were randomly selected and each was classified as a diagnosed alcoholic, undiagnosed alcoholic, or nonalcoholic and categorized according to his or her marital status. A summary of the response is shown in the table.

		Alcoholic Classification			Total
		Diagnosed	Undiagnosed	Nonalcoholic	
Marital Status	Married	21 *33*	37 *41*	58 *42*	116
	Not married	59 *47*	63 *59*	42 *58*	164
Total		80	100	100	280

Is there evidence that marital status and alcoholic classification are dependent? Test this hypothesis and state your conclusion.

116 out of 280 are married - that's 41%.

41% of 80 is 33 } which lead to the expected counts written
41% of 100 is 41 } in the table above.

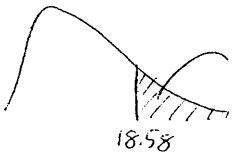
null: marital status & alcoholic status are independent.

alt: " " " " dependent.

χ^2:

obs	exp	$(o-e)^2/exp$
21	33	4.36
59	47	3.06
37	41	.39
63	59	.27
58	42	6.09
42	58	4.41
		18.58

$df = (2-1)(3-1) = 2$



P-value is less than 1%.

We reject the null & conclude that marital status & alcoholism are not independent.

9. (10 points) In a flower breeding trial, the results were as follows:

Type of Flower	Observed number	Expected number	$(Obs-Exp)^2/Exp$
tall stem yellow	278	256	1.89
short stem yellow	306	328	1.48
tall stem white	223	205	1.58
short stem white	96	114	2.84
			$\chi^2 = 7.79$

The expected counts were obtained using the breeder's theory. Do the observed counts support the breeder's theory? Test at the 1% level. Set up the null and alternative hypotheses, perform the test, and clearly state your conclusions.

null: the data are consistent with what we would expect from the model.

alt: the data are not consistent with the model.

$\chi^2 = 7.79$ (working beside table) on $4-1=3$ df.

The P-value is a little over 5%. It's borderline at the 5% level - at the 1% level we fail to reject the null and conclude that there is no evidence the model doesn't hold.

10. A high school teacher working at an inner city high school is concerned about the time students spend working at after school jobs. She randomly selects 15 of her students and finds the average time spent working at after school jobs is 13.1 hours, with an SD of 11.5 hours.

(a) (10 points) The national average is 10.7 hours. Assuming that the hours worked follow the normal curve, test to see whether this teacher's students work longer, on average, than those in the nation as a whole.

null: the average for her students is 10.7 hours.

alt: " " " " is more than 10.7 hours.

$SD^+ = \sqrt{\frac{15}{14}} \times 11.5 = 11.9$

$SE_{sum} = 46.1$

$SE_{ave} = 3.07$

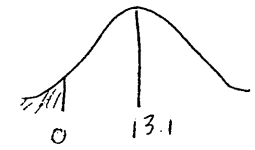
$t = \frac{13.1 - 10.7}{3.07} = .78$

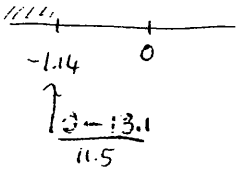
$df = 15 - 1 = 14$

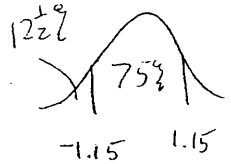
We do not reject the null. We conclude there is not enough evidence to say that her students work longer.

P-value is between 10% and 25%.

(b) (5 points) Assuming that the hours worked follow the normal curve, estimate the percentage of this teacher's students that work less than 0 hours.







The percentage is about 12½%.

(c) (3 points) From your result in part (b), do you think the hours worked really does follow the normal curve for these students? Explain.

No - if it did, we'd have 12½% negative - but you can't work negative hours! 12½% is a lot to be in an impossible region, so we conclude the numbers do not follow the normal curve.

Name:

Stat 1040, Spring 1999
Final Test, Monday May 3, 3:30–5:20

Show your work. The test is out of 100 points and you have 110 minutes, so budget your time accordingly.

1. In a school district with 1500 kindergarten children, the heights of 64 randomly chosen children are measured. The average height of these 64 children is 49.1 inches with an SD of 2.4 inches. Suppose that the heights of kindergarten children are known to follow the normal curve.

   (a) (7 points) Find a 95% confidence interval for the average height of all 1500 kindergarten children in that school district.

   SD ≈ 2.4"

   $SE_{sum} = \sqrt{64} \times 2.4 = 19.2$

   $SE_{ave} = \frac{19.2}{64} = .3$

   CI: $49.1 \pm 2 \times .3$ i.e. $49.1" \pm .6"$

   (b) (5 points) Approximately 95% of all the kindergarten children in that school district will have heights in the interval $49.1" \pm 4.8"$ ($2 \times 2.4"$).

   (c) Now suppose that you find out that the heights do **not** follow the normal curve.

   i. (3 points) Is your interval in part (a) still valid? Explain.

   Yes - it's a confidence interval and the sample size is large (at least, it's large for continuous data like heights).

   ii. (3 points) Is your interval in part (b) still valid? Explain.

   No - we need the normal curve to find an interval containing 95% of the individual heights.

   (d) (4 points) True or False (no explanation is required).

   What if it's an unusual sample?

   i. The average height of all these 1500 kindergarten students is in the interval found in part (a) 95% of the time. False

   ii. There is a 68% chance that the average height of all these 1500 kindergarten children is between 48.8 and 49.4 inches. False

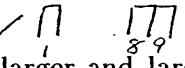
2. (5 points) In 1998 a researcher looked at a large, representative group of women. She found that, on average, the older these women were, the less meat they consumed. True or (false) and explain: "The data show that on average, women eat less meat as they get older".

This is a cross-sectional study. To understand what happens as people age, you have to watch them age, i.e. you need a longitudinal study. For example, perhaps these older women just always ate less meat, even when they were young. This is not association ≠ causation – we don't even know if there really is any association.

3. (5 points) For a statistics class of 50 students, the average score on Test 1 is 73.5 with an SD of 15. The average score on Test 2 is 76.5 with an SD of 18. Explain why it is **not** appropriate to perform a 2-sample z-test to decide whether the difference between the scores on Test 1 and Test 2 is significant.

We have the whole population. The averages are 73.5 and 76.5 – they differ. There is no sampling, no randomization, nothing to justify a box model.

4. A box contains three tickets: 1, 8, and 9. We draw from this box with replacement. Fill in the blanks and explain:



(a) (3 points) As the number of draws gets larger and larger, the data histogram of the draws will look more and more like the probability histogram for the box.

(b) (3 points) As the number of draws gets larger and larger, the probability histogram for the average of the draws (when put in standard units) will look more and more like the normal curve.

5. (7 points) In order to estimate the percentage of USU students who go to Salt Lake City more than three times a month, a simple random sample of 400 students was selected. Forty said that they go to Salt Lake at least four times a month. Construct a 95% confidence interval for the percentage of USU students who go to Salt Lake more than three times a month.

Sample percentage is $\frac{40}{400} \times 100\% = 10\%$.

$SD_{box} \approx \sqrt{.1 \times .9} = .3$ (bootstrap)

$SE_{sum} = \sqrt{400} \times .3 = 6$

$SE_{\%} = \frac{6}{400} \times 100\% = 1.5\%$

CI: $10\% \pm 3\%$

6. A new food additive is regularly fed to 300 white mice selected at random from a set of 600. The remaining 300 are fed a normal lab diet as controls. After two years, the 600 mice are dissected and examined for 20 different forms of cancer - lung, liver, bone, and so on. A two-sample z test is conducted for each of the 20 cancer rates, and the p-value for the test on stomach cancer is 0.03 (against an alternative that the additive increases the chance of stomach cancer). The other 19 tests all have p-values of greater than 0.05.

(a) (7 points) Is this strong evidence that the additive causes stomach cancer? Explain.

No - we expect 1 out of 20 to have a P-value of less than .05 when the null hypotheses are true. This could easily be due to chance.

(b) (3 points) What should be the next step in this investigation?

Repeat the experiment with just this type of cancer as the thing of interest.

7. **(6 points) The first year professional baseball allowed free agency (allowing players the right to negotiate contracts with any team they chose), a few of the wealthiest clubs paid enormous sums to obtain a few players who had had very high batting averages the previous year. During the course of the year, sports writers had lots of fun pointing out what bad judgement the owners had shown, for almost none of the high-priced players did as well that year as they had the year before. Is there a statistical explanation for this result, which does not assume the (now-rich) players had grown fat and lazy? Explain.**

The regression effect suggests that the players who did very well the first year wouldn't do quite as well the second year. It could just be because of natural variation.

8. (10 points) In the game of chess, the first few moves play a very important role in determining the final outcome. Five different opening strategies are highly favored by chess experts. To determine whether one or more of these strategies is most preferred by grand masters in international competition, a random sample of 100 grand masters is taken, and each is asked which of the strategies he or she would prefer to employ. A summary of their responses is given below:

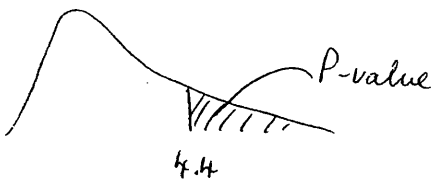
Strategy	A	B	C	D	E
Frequency	17	27	22	15	19

Make a χ^2-test of the null hypothesis that there is no preference between these strategies by grand masters in international competition. You should state the null and the alternative hypotheses and clearly state your conclusions.

null: no preference - it's as though they choose at random.
alt: some strategies are preferred, at least by some players.

obs	exp	$(obs-exp)^2/exp$
17	20	.45
27	20	2.45
22	20	.20
15	20	1.25
19	20	.05

$\chi^2 = 4.4$ $\quad df = 5-1 = 4$



The P-value is between 30% and 50% so we do not reject the null - we conclude that there is no evidence of any preference. They could just be picking at random.

9. According to USA Today (1997), the age of adults getting knee replacements is distributed according to the following table (the endpoint convention is that the class intervals include the left endpoint but not the right endpoint).

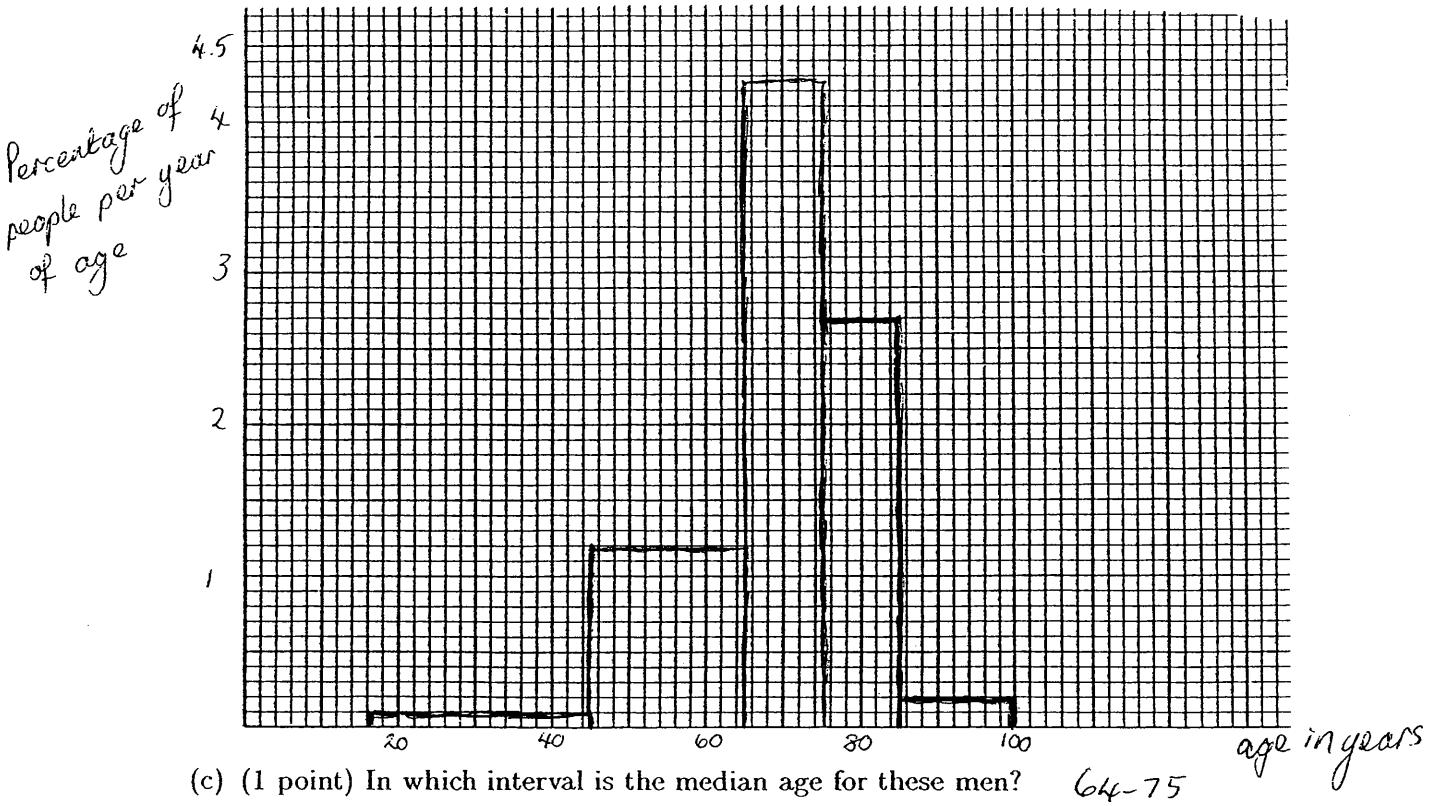
Age (in years)		Percentage of people	
18–45	width 27	2.8	.10
45–65	20	24.6	1.23
65–75	10	43.3	4.33
75–85	10	26.7	2.67
85–100	15	2.9	.19

Note: for the purposes of this question, we have assumed that only adults less than 100 years old were included.

(a) (3 points) The total percentage is $(2.8 + 24.6 + 43.3 + 26.7 + 2.9)\% = 100.3\%$. Does this imply that they made a mistake? What else could account for the fact that the percentage is slightly higher than 100%?

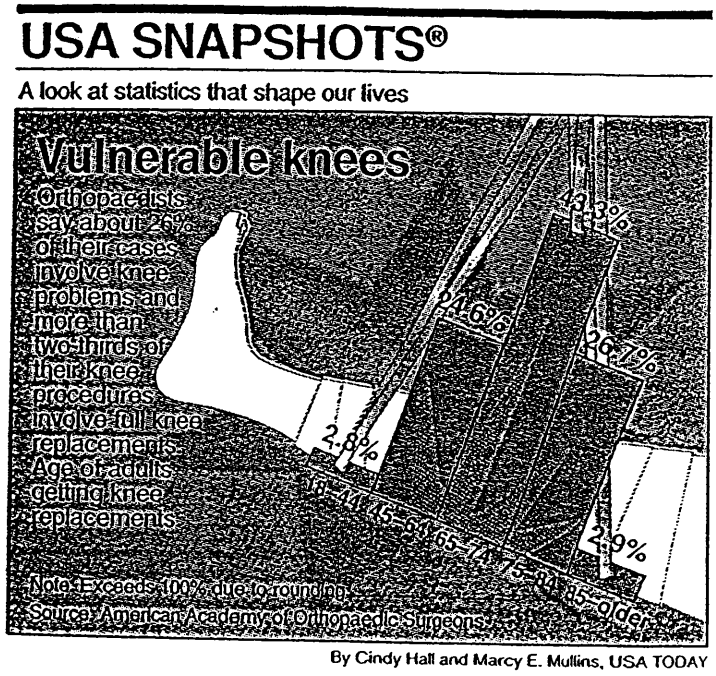
Roundoff

(b) (10 points) In the space provided, sketch a histogram for these data. The picture does not need to be artistically perfect but it must be drawn to scale and the axes must be properly labelled.



(c) (1 point) In which interval is the median age for these men? 64–75

(d) (5 points) Now consider the figure given in USA Today:



Does the USA Today figure correctly represent the data? Explain clearly. (Ignore the fact that they used a different class interval endpoint convention).

No! The interval above 45-64 has width = 20
The interval above 75-84 has width = 10.
Both blocks are about the same height, which makes it look like there are about the same percentage per year in each - but in fact, it's only half as much for the first block as for the second (approx)

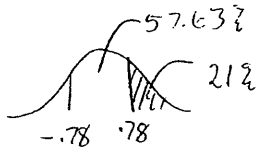
10. (10 points) Scores on the Verbal Graduate Record Examination test were recorded. For 68 randomly selected women from a given population, the average score was 538.82 with an SD of 114.16. For 86 randomly selected men from the same population, the average score was 525.23 with an SD of 97.23. Test to see if the population average of the women is higher than that of the men for this population. State the Null and Alternate Hypotheses and clearly state your conclusion.

null: average for men & women is the same for this pop"
alt: average for women is higher.

women	men
$SD \approx 114.16$	$SD \approx 97.23$
$SE_{sum} = \sqrt{68} \times 114.16 = 941$	$SE_{sum} = \sqrt{86} \times 97.23 = 902$
$SE_{ave} = 13.8$	$SE_{ave} = 10.5$

$SE_{diff} = \sqrt{13.8^2 + 10.5^2} = 17.34$

$z = \frac{538.82 - 525.23}{17.34} = .78$



P-value is approx 21% so we don't reject the null. We conclude there is no evidence the women's population average is higher than the men's.

Name:

Stat 1040, Fall 1999
Final Test, Friday 17 December, 7:00–8:50 am

Show your work. The test is out of 100 points and you have 110 minutes, so budget your time accordingly.

1. As part of a study on exercise and health, a group of 1,000 people was followed for 5 years. At the beginning of the study, the researchers asked each person whether they exercised regularly or not. At the end of the study, the researchers measured several health-related variables, and in doing so, they noticed that the death rate for the exercise group was lower than for the no-exercise group. They performed a statistical test and found that the difference was "highly statistically significant".

   (a) (6 points) Does this result necessarily imply that if people who do not exercise start to exercise regularly, they will live longer, on average, than if they do not? Explain clearly.

   No- this result merely shows that the type of people who choose to exercise differ from the type who do not. This is an observational study and there could be many confounding factors. For example, people who exercise might smoke less and it might be smoking that's harmful not exercise that's beneficial. Taking a smoker & making them exercise might have no impact.

   (b) (3 points) Is this an example of an observational study or a controlled experiment? (Explain).

   They were asked what they did- they weren't told what to do.

   (c) (3 points) What does it mean for a test statistic to be "highly statistically significant"? (Explain briefly).

   It means that the P-value is less than 1%. In other words, if the two groups are the same (in the population) there's only a 1% chance (or less) that the sample values would be as different as (or more different than) what we saw.

2. (12 points) In the January 27, 1999 edition of *The Journal of the American Medical Association* they describe a randomized clinical trial in which the 236 sedentary people were randomly allocated to either a lifestyle physical activity group (118 people) or a traditional structured exercise group (118 people). At the end of 24 months, the average drop in systolic blood pressure for the lifestyle group was 3.63 mm Hg, with an SD of 10.58 mm Hg. The average drop in systolic blood pressure for the traditional group was 3.26 mm Hg, with an SD of 11.08 mm Hg. Test to see whether there is a significant difference between the two treatments with respect to the drop in systolic blood pressure. Remember to clearly state the null and alternative hypotheses and your conclusion.

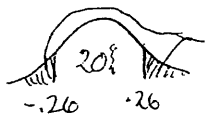
null: the two treatments are indistinguishable in their effect on BP

alt: " " " are different. " " " "

lifestyle	trad. exercise
$SD \approx 10.58$	$SD \approx 11.08$
$SE_{sum} = \sqrt{118} \times 10.58 = 115$	$SE_{sum} = \sqrt{118} \times 11.08 = 120$
$SE_{ave} = 115/118 = .97$	$SE_{ave} = 120/118 = 1.02$

$$SE_{diff} = \sqrt{.97^2 + 1.02^2} = 1.41$$

$$z = \frac{3.63 - 3.26}{1.41} = .26$$ so



80% = P-value

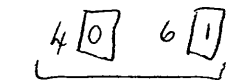
We do not reject the null. We conclude that there is no evidence of a difference wrt BP.

3. (12 points) A journal article claims that 60% of North American adults are too sedentary. Suppose you think that a lower percentage of Cache Valley adults are too sedentary. To test your belief, you take a simple random sample of 120 Cache Valley adults and find that only 68 are too sedentary. Test to see if your belief is correct. (You may assume that everybody is using the same definition of "too sedentary", although in practice this would be contentious).

null: 60% of all Cache Valley adults are "too sedentary"

alt: less than 60% " " " " " "

If the null is true, the sample percentage should be like the % of 1's in 120 draws from:



ave = .6

$SD_{box} = \sqrt{.6 \times .4} = .49$

observed percent = $\frac{68}{120} \times 100\% = 57\%$

$SE_{sum} = \sqrt{120} \times .49 = 5.37$

$SE_{\%} = \frac{5.37}{120} \times 100\% = 4.47\%$

$$z = \frac{57 - 60}{4.47} = -.67$$

The P-value is about 26% so we do not reject the null. There is no evidence that Cache Valley adults are less sedentary than the national figure.

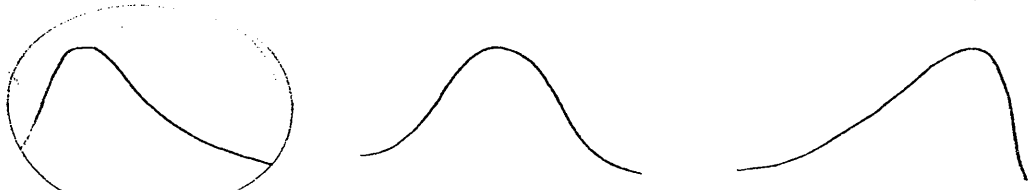
4. The manager of a commercial web site is interested in the average length of time a visitor spends at their site. She has the computer randomly sample 1000 visits from the past month and finds that the average length of time is 38 seconds, with an SD of 228 seconds. The median time is 18 seconds.

(a) (3 points) When she sees that the SD is so large, she thinks there has been some mistake. She remembers from her statistics class that 68% of the numbers should be within 1 SD of the average, but when she subtracts 228 from 38, she gets a negative number, and nobody can visit for a negative amount of time. Assuming the numbers are correct, what could account for such a large standard deviation?

Lots of short visits and a few really long ones (eg. people went to lunch, went home, went on vacation, etc.!)

(b) (3 points) The histogram for the times looked like: (choose the best one).

this one



(c) (6 points) If possible, find a 95% confidence interval for the average length of time for all the visits for the last month. If this is not possible, clearly state why not.

average = 38

SD = 228

$SE_{sum} = \sqrt{1000} \times 228 = 7210$ $\quad SE_{ave} = \frac{7210}{1000} = 7.21$

CI: $38 \pm 2 \times 7.21$

i.e. 38 ± 14

(d) (5 points) For a commercial web page, the manager wants to find a confidence interval for the average order total. She has a record for *all* order totals for the last 30 days. There were 596 orders and the average amount was \$57.56 with a standard deviation of \$22.39. If appropriate, find a 95% confidence interval for the average order total. If it is not appropriate, clearly explain why not.

This is not appropriate because she has the whole population for this 30 days and no information for any other time. There is no random sampling.

5. (12 points) An insurance company wants to know whether the occurrence of automobile accidents is independent of cellular phone use. They randomly sample 774 people and find:

	Accident in the last 12 months	
	Yes	No
Cellular Phone User	39	282
Not a Cellular Phone User	46	407

Is there evidence that cellular phone use and automobile accidents are dependent? Test the appropriate hypothesis and clearly state your conclusion.

null: cellular phone use and accidents are independent

alt: " " " " are dependent.

expected counts:

35	286	321
50	403	453
85	689	774

obs	exp	$(o-e)^2/e$
39	35	.457
46	50	.320
282	286	.056
407	403	.040
	$\chi^2 =$	.873

$df = (2-1)\times(2-1) = 1$

P-value ≈ between 30% and 50%

Since the P-value is large, we do not reject the null. We conclude that there is no evidence that cell phone use is related to having an accident in the last 12 months.

6. In a box of 12 chocolates, 3 are mint, 2 are orange, 5 are caramel and 2 are cherry. I choose 2 chocolates at random (without replacement!). **Answer each of the following questions separately.**

   (a) (3 points) What is the chance the first is mint?

   $\frac{3}{12}$

   (b) (4 points) What is the chance the first two are both mint?

   $\frac{3}{12} \times \frac{2}{11}$

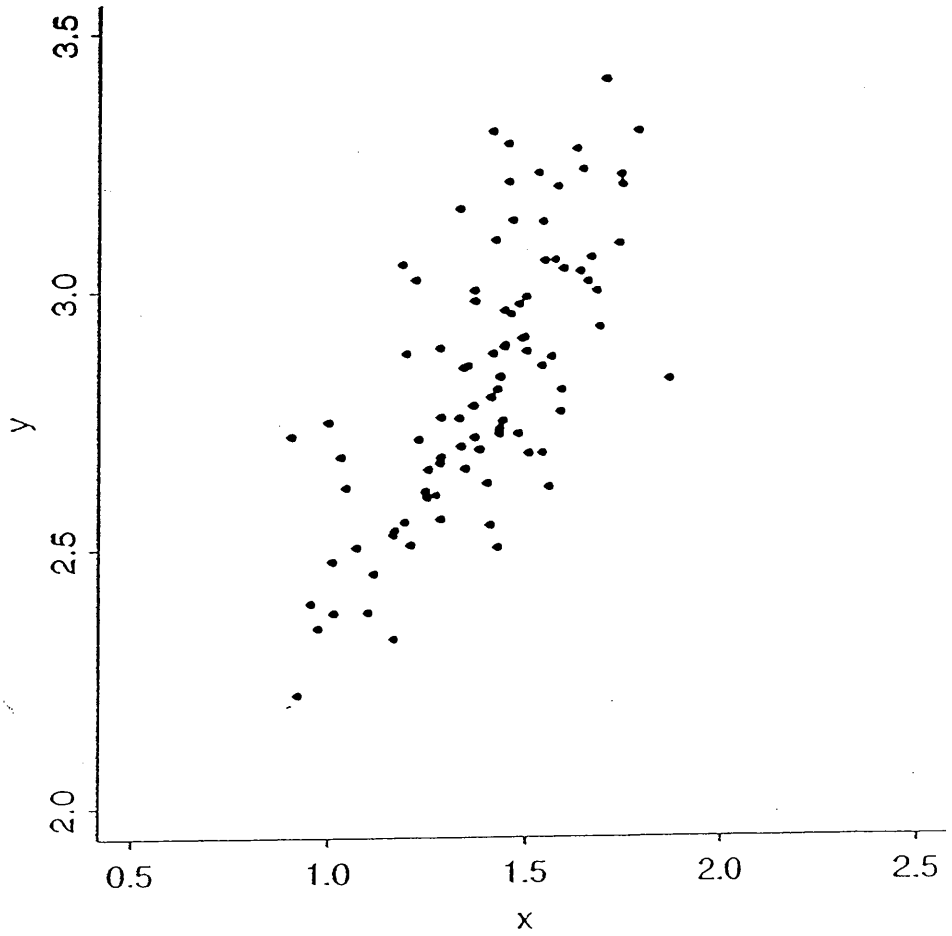
   (c) (4 points) What is the chance the first is mint and the second is orange?

   $\frac{3}{12} \times \frac{2}{11}$

   (d) (4 points) If I only like caramel, what is the chance that I like neither of the chocolates I choose?

   $\frac{7}{12} \times \frac{6}{11}$

7. Consider the scatter plot given below.



(a) (3 points) The average of x is closest to:

(i) 1.4

ii. 1.6

iii. 2.6

iv. 2.8

(b) (3 points) The standard deviation of x is closest to:

(i) 0.2

ii. 0.5

iii. 1.0

iv. 1.2

(c) (3 points) The correlation between x and y is closest to:

i. 0.0

ii. 0.35

(iii) 0.70

iv. 0.95

8. In web polls, anyone who views a certain web page is allowed to vote by clicking on their choice of button. In fact, there is nothing to stop someone voting as many times as they want. The results of one such poll suggest that almost 90% of the US population wants to ban firearm sales. The poll has a very large sample size (over 1 million) and the reported "margin of error" is less than 1%.

(a) (6 points) Web based polls such as this are notoriously susceptible to bias. Give 3 possible sources of bias for this poll.

1. People must own a computer - creates a bias towards younger, wealthier, more technologically-trained, etc. people.
2. People who feel strongly are more likely to vote + these people may vote multiple times.
3. The source will create a bias - what type of people read this web page? Maybe it's in a newspaper web page - only people who find it will have the chance to vote.

(b) (5 points) Are the sources of bias you listed in part (a) a problem even with a very large sample, or does the sample size imply that they can be ignored? Explain.

All a large sample does is repeat a mistake on a grand scale, if the data are subject to bias! So all we have is a large biased sample - we cannot ignore the problem.

Name:

Stat 1040, Spring 2000
Midterm 1, February 17

Show your work. The test is out of 100 points and you have 50 minutes, so budget your time accordingly.

1. Oregon has an experimental boot camp program to rehabilitate prisoners before their release. T he object is to reduce the "recidivism rate" - the percentage who will be back in prison within 3 years. Prisoners volunteer for the program, which lasts several months, but many prisoners drop out before completing the treatment.

   To evaluate the program, investigators compared prisoners who completed the program (the "graduates") with prisoners who dropped out (the "dropouts"). The recidivism rate for those who completed the program was 29%. For the dropouts, the recidivism rate was 74%.

   (a) (3 points) Was this an observational study or a controlled experiment?

   It was an observational study - they chose to be in boot camp.

   (b) (3 points) Was this study blind?

   No - they knew if they were in boot camp or not.

   (c) (7 points) Why did they compare the "graduates" and the "dropouts" instead of comparing the "graduates" and the prisoners who did not volunteer for the program? Explain.

   They wanted to compare groups that were as similar as possible except for boot camp. They suspected that the ones who volunteered might be more motivated, for example, than those who did not, so they wanted to eliminate the confounding effect of "volunteerism".

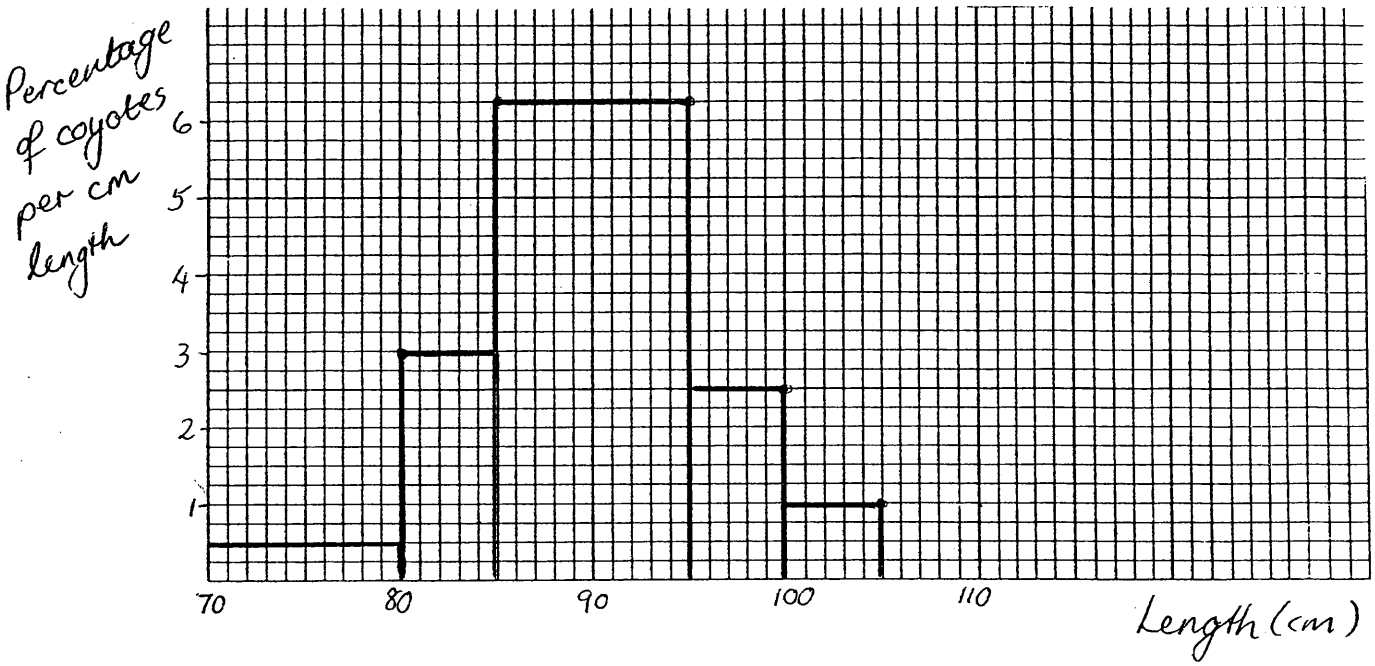
   (d) (8 points) Is the difference in the recidivism rates convincing evidence that Boot Camp works, or could there be another possible explanation for this difference? Explain clearly.

   It could be that the type of people who "drop out" differ from those who persevere, so perhaps it wasn't boot camp itself but just the "perseverence" effect that they were seeing.

2. Lengths of a group of 40 female coyotes are distributed as follows (intervals include the left endpoint but not the right endpoint):

Width	Length (cm)	Number of Coyotes	Percentage	Height
10	70-80	2	5	.5
5	80-85	6	15	3
10	85-95	25	62.5	6.25
5	95-100	5	12.5	2.5
5	100-105	2	5	1

(a) (11 points) Draw a histogram for the lengths. Be sure to label the axes.



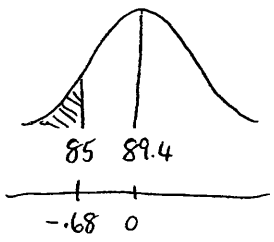
(b) (4 points) For these data, the average length is 89.4 cm with an SD of 6.5 cm. Find the average and the SD in inches. (Note: 2.54 cm = 1 inch).

average = 89.4/2.54
= 35.20"

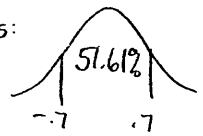
SD = 6.5/2.54
= 2.56"

Quite a number of people got an average of 89.4×2.54 = 227 inches - that's a rather large coyote!!!

(c) (10 points) Assuming that the data followed the normal curve, estimate the percentage of the coyotes that would be less than 85.0 cm long.



Tables:



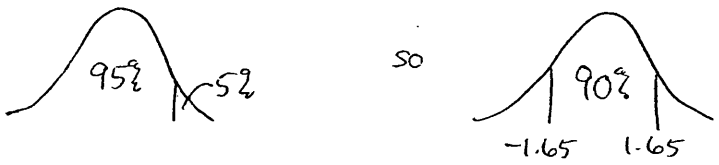
So each tail is about 24%.

$\frac{85-89.4}{6.5} = -.68$

(d) (3 points) Using either the histogram or the table, find the true percentage of the coyotes that are less than 85.0 cm long.

20% = 5% + 15%
(70-80) (80-85)

(e) (10 points) Assuming that the data followed the normal curve, estimate the 95th percentile (in cm).



So they are 1.65 SD's above average.
Average = 89.4
SD = 6.5
so they are $89.4 + (1.65 \times 6.5)$ cm $= 100.125$ cm.

(f) (3 points) Using either the histogram or the table, find the true 95th percentile (in cm).

100cm (only 5% are 100cm or longer)

Also acceptable: 99cm or 99.5 cm.

3. (8 points) In 1998 a researcher took a large representative sample of women in the US. She found that the older these women were, the less their daily average meat consumption. True or false and explain: "The data show that as women grow older, their daily average meat consumption drops".

False - this is a cross-sectional study. To find out what happens as people age, you need to watch them age. An alternative explanation could be that women who are young today eat more meat than those who were young in the past.
Not: "association is not causation" was not enough - even association is questionable here! (i.e. the point is, we don't even know that age & meat consumption are associated, and we are not trying to establish a cause).

4. (3 points each) True or false:

(a) If you add 10 to each entry on a list, the average and the SD also increase by 10.
F - the SD remains unchanged.

(b) The regression effect says that in test-retest situations, the people who score the highest on the first test will tend to be below average on the second test.
F - they will still do well, just not quite as well.

(c) Provided we take a large, representative sample, the area under the normal curve will give a good approximation to the percentage of cases in the given region.
F - consider a large representative sample of incomes!

5. In a study of a large number of California boys it was found that
Average height at age 6 = 46 inches, SD = 1.7 inches
Average height at age 18 = 70 inches, SD = 2.5 inches, r = 0.8
The scatter plot for the heights was football shaped.

(a) (7 points) Find the equation of the regression line for predicting height at age 18 from height at age 6.

$\text{slope} = .8 \times \frac{2.5}{1.7} = 1.176$

$\text{intercept} = 70 - 1.176 \times 46 = 15.88$

$y = 15.88 + 1.176x$

↑ height at 18 ↑ height at 6

(b) (4 points) Predict the height at age 18 for a boy who was 44 inches tall at age 6.

$y = 15.88 + (1.176 \times 44) = \underline{67.6''}$

(c) (4 points) Put a give-or-take number on your estimate.

$\sqrt{1-r^2}\ SD_y = \sqrt{1-.8^2} \times 2.5 = 1.5$

(d) (6 points) A child who is right at the 80th percentile for height at age 6 will likely be

(i) slightly below the 80th percentile for height at age 18.

ii. right at the 80th percentile for height at age 18.

iii. slightly above the 80th percentile for height at age 18.

Explain.

The regression effect says that they will still be tall, but probably not as tall (compared to the group).

Name:

Stat 1040, Spring 2000
Midterm 2, April 3

Show your work. The test is out of 100 points and you have 50 minutes, so budget your time accordingly.

1. A newspaper web site has a poll on gun control. In one hour, they had 1000 votes - 780 "for" gun control and the rest "against" gun control.

   (a) (15 points) Assuming that 1000 votes are a simple random sample from some very large population, find a 95% confidence interval for the percentage of people in the population who favor gun control.

   Sample percentage $= \frac{780}{1000} \times 100\% = 78\%$

   Bootstrap: 22 [0] 78 [1]

   $ave_{box} = .78$

   $SD_{box} = \sqrt{\frac{22}{100} \times \frac{78}{100}} = .414$

   $SE_{sum} = \sqrt{1000} \times .414 = 13$

   $SE_{\%} = \frac{13}{1000} \times 100\% = 1.3\%$

   so the confidence interval is $78.0\% \pm 2.6\%$.

   (b) (6 points) Explain why we would not expect these 1000 votes to be like a simple random sample.

   We would not expect them to be like a simple random sample because they have to be computer users, they are reading a newspaper web page, and they choose to vote. There is no way of working out the chance that anyone is in the sample because chance is not used.

   (c) (6 points) Give a plausible source of bias in the survey. (For example, an acceptable answer would be "people who feel strongly for (or against) gun control are more likely to respond to a survey on the subject and this would cause a bias for (or against) gun control").

   People who read newspapers on the web are more likely to be educated, white collar workers, and such people probably have different views on gun control than less educated, blue collar workers.

2. Shaquille O'Neal has a 50% chance of "making" a free-throw (i.e. getting the ball in the net). For parts a),b) and c), assume that this chance is constant and that all his free-throws are independent. If Shaquille takes 10 free-throws (i.e. shoots the ball 10 times):

(a) (6 points) what is the chance that he makes none of the 10 shots?

$n = 10$
$p = .5$
$k = 0$

$\frac{10!}{0!\,10!}(.5)^{10} = (.5)^{10}$ or $\left(\frac{1}{2}\right)^{10} = .00098$

(b) (6 points) what is the chance that he makes at least 1 of the 10 shots?

$1 - \left(\frac{1}{2}\right)^{10}$

(c) (6 points) would it be more likely for Shaquille to make exactly 5 shots out of 10, or exactly 10 shots out of 20? (No calculations are required).

He's more likely to make 5 shots out of 10 by the law of averages.

(d) (6 points) Do you think the 10 shots really would be independent? (Hint: think about what Shaquille might be feeling if he has missed 9 shots in a row and he steps up to take the tenth shot).

No - if he's failed for 9 shots he will perhaps feel as though he's having a "bad day" and give up. Or, he may get angry & try harder.

3. (16 points) A greenhouse has a large number of petunia seedlings of which 25% are purple, the rest white. If I buy 144 of these seedlings, chosen at random, find the chance that I get at least 30% purple ones.

The percentage of purples should be like the percentage of 1's in 144 draws from the box

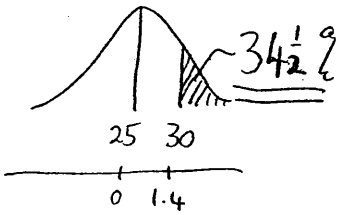
$ave_{box} = .25$

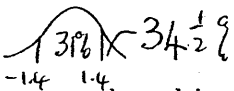
$SD_{box} = .433$

$EV_{\%} = 25\%$

$SE_{sum} = \sqrt{144} \times .433 = 5.2$

$SE_{\%} = \frac{5.2}{144} \times 100\% = 3.6\%$

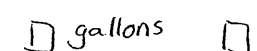




4. For 500 cars visiting a certain gas station, the average number of gallons of gas purchased is 9.84 with an SD of 3.92.

(a) (16 points) If I take a random sample of 100 of these cars, what is the chance that the total number of gallons purchased is more than 1000 gallons?

The total is like the sum of 100 draws from the box



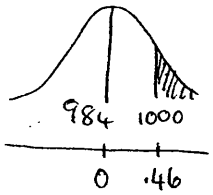
$ave_{box} = 9.84$

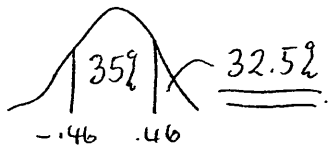
$SD_{box} = 3.92$

$EV_{sum} = 100 \times 9.84 = 984$

$SE_{sum} = \sqrt{100} \times 3.92 = 39.2$

$c.f. = \sqrt{\frac{500-100}{500-1}} = .895$ so the true SE_{sum} is $.895 \times 39.2 = 35$





(b) (6 points) When I do a histogram for the number of gallons of gas purchased by these 500 cars, I find that it does not follow the normal curve very closely. Does this mean that my answer in part a) is invalid? Explain briefly.

We have 100 draws, so the sum of the draws will follow the normal curve even if the tickets do not. The interval is still valid.

5. (6 points) True or false and explain briefly: When predicting Presidential election results, we can trust that a sample of size 10,000 will almost always give a correct prediction. For such a large sample, we don't need to know how the sample was collected.

This is false - a carelessly collected sample (even a large one) has no guarantee of giving a correct prediction.

6. (5 points) A local politician is interested in estimating the percentage of voters who are opposed to the marriage tax. Other things being equal, to get equal accuracy in Logan and Salt Lake City, she should sample:

(a) more people in Logan than in Salt Lake City.

(b) more people in Salt Lake City than in Logan.

(c) the same number of people in Logan and in Salt Lake City.

Note: you should assume that she can only afford to sample a small percentage of the population.

Name:

Stat 1040, Spring 2000
Final Test, Friday May 5, 7:00–8:50 am

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

1. In a recent study on SIDS (Sudden Infant Death Syndrome), one hospital collected data on 128 babies who died from SIDS in the last 12 months. They took a random sample of 500 babies (of similar ages) who did not die from SIDS (the "controls"), and they compared the two groups with respect to several variables of interest (e.g. whether the child slept on his or her stomach, birthweight, time of year, whether the mother smoked, whether she breast-fed, socio-economic status, etc).

   (a) (3 points) Is this a controlled experiment or an observational study? Explain.

   It is an observational study - there was no intervention.

   (b) One physician noticed that 63% of the SIDS babies had mothers who smoked during pregnancy, whereas only 26% of the control babies had mothers who smoked during pregnancy. Another physician claimed that low birthweight could be a "confounding factor".

      i. (7 points) Explain what it means for low birthweight to be a "confounding factor". Be specific.

      Perhaps smoking causes low birthweight and it is the low birthweight rather than smoking itself that is leading to higher rates of SIDS.

      ii. (3 points) If you had access to the data, what would you do to "control for" birthweight?

      Study babies with similar birthweights separately. eg. break up the comparison into groups of, say, babies 6-6.5 lb, 6.5-7 lb, 7-7.5 lb, etc.

2. (6 points) A market research company plans to take a simple random sample of people in Salt Lake City and an independent simple random sample of people in Logan to estimate the percentage who would purchase a new flavor of a gelatin dessert mix. The company can only afford to sample a few hundred people. All other things being similar, if they want to get similar accuracy in both locations they should:

   (a) sample more people in Salt Lake City than in Logan.
   (b) sample the same number of people in both locations.
   (c) sample the same percentage of people in both locations.
   (d) sample more people in Logan than in Salt Lake City.

   (Circle the best answer, no explanation is required.)

3. (10 points) I take a random sample of 500 Cache Valley drivers and find that 346 always wear a seatbelt. Find a 95% confidence interval for the percentage of *all* Cache Valley drivers who always wear a seatbelt. sample % = $\frac{346}{500} \times 100\% = 69.2\%$

?[0] ?[1], $SD_{box} \approx \sqrt{\frac{346}{500} \times \frac{154}{500}} = .46$

$SE_{sum} = \sqrt{500} \times .46 = 10.3$

$SE_{\%} = \frac{10.3}{500} \times 100\% = 2.06\%$

The 95% CI is $69.2\% \pm 2 \times 2.06\% \approx 69.2\% \pm 4.1\%$

i.e. 65.1% to 73.3%

4. (3 points) In question 3, did you need to assume that seatbelt usage followed the normal curve? Why/why not?

No because it's a confidence interval and the sample is large.

5. (12 points) Reading test scores are known to follow the normal curve. The average reading test score is supposed to be 100. I suspect that the children at one of the local schools have a higher average reading score, so I take a simple random sample of 10 of these children and find that their average reading score is 112, with an SD of 18. Is my suspicion correct? Perform the appropriate statistical test. You must clearly state a null and alternative hypothesis, compute a test statistic and a P-value and state your conclusions.

null: the average for <u>all</u> children at this school is 100.

alt: " " " " .. more than 100

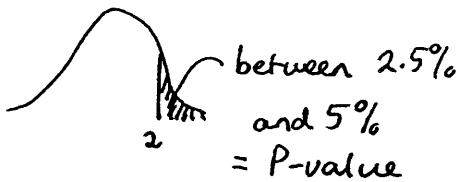
$SD^{+} = \sqrt{\frac{10}{9}} \times 18 = 19$

$SE_{sum} = \sqrt{10} \times 19 = 60$

$SE_{ave} = \frac{60}{10} = 6$

$t = \frac{112 - 100}{6} = 2$

$df = 10 - 1 = 9$



We reject the null because the P-value is small. We conclude that the students at this school do have a higher average.

6. (3 points) Would your test in question 5 still be valid if reading test scores were known to have a very skew histogram? Explain clearly.

No - the t-test assumes that the data are close to normal.

7. Suppose that 15% of all children in a very large population suffer from some form of learning disability. I choose 10 children at random from this population. *Answer the following questions separately:*

(a) (3 points) Find the chance that none of the 10 children suffer from some form of learning disability.

$(.85)^{10}$

(b) (4 points) Find the chance that exactly 2 of the 10 children suffer from some form of learning disability.

$n=10$, $p=.15$, $k=2$

$$\frac{10!}{2!8!}(.15)^2(.85)^8 = 45(.15)^2(.85)^8$$

(c) (4 points) Find the chance that at least one of the 10 children suffers from some form of learning disability.

$1-(.85)^{10}$

8. (12 points) In a study on snoring and nightmares, researchers found the following results:

	Frequency of nightmares				
	Never	Seldom	Occasionally	Frequently	Total
Nonsnorers	22 21.6	45 43.2	35 35.2	11 13.1	113
Snorers	16 16.4	31 32.8	27 26.8	12 9.9	86
Total	38	76	62	23	199

$\frac{113}{199} \times 100\% = 56.8\%$

56.8% of 38 is 21.6

56.8% of 76 is 43.2

56.8% of 62 is 35.2

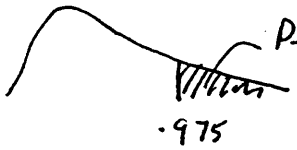
Assuming this is a random sample of people, is there evidence that the frequency of nightmares (never, seldom, occasionally, frequently) is different for snorers and nonsnorers? You must clearly state a null and alternative hypothesis, compute a test statistic and a P-value and state your conclusions.

null: nightmares & snoring are independent - the distribution of nightmare frequency is the same for snorers as nonsnorers

alt: nightmare frequency is different for snorers and nonsnorers

obs	exp	$(obs-exp)^2/exp$
22	21.6	.007
16	16.4	.010
45	43.2	.075
31	32.8	.099
35	35.2	.001
27	26.8	.001
11	13.1	.337
12	9.9	.445
		$.975 = \chi^2$

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



P-value between 70% and 90%

.975

We do not reject the null. We conclude that there is no evidence snorers have a different frequency of nightmares than nonsnorers.

9. One semester, a multi-section statistics class has 600 students. There are two versions of the final exam, version A and version B. The finals are graded and it is found that the average score for the 300 students who got version A is 67 with an SD of 21. The average score for the 300 students who got version B is 77 with an SD of 21.

(a) (3 points) If we were to calculate the average of all 600 exam scores, would it be smaller than 67, halfway between 67 and 77, or larger than 77? Or can this be determined from the information given?

halfway between 67 and 77

(b) (5 points) If we were to calculate the SD of all 600 exam scores, would it be smaller than 21, equal to 21, or larger than 21? Or can this be determined from the information given?

larger because the variability will be larger

(c) (12 points) Assuming that each student got version A or B at random, and that both versions were graded comparably, test the hypothesis that the two versions are equally difficult. You must clearly state a null and alternative hypothesis, compute a test statistic and a P-value and state your conclusions about the exams.

null: tests are equally difficult

alt: tests are not equally difficult

A: SD ≈ 21

$SE_{sum} = 363.7$

$SE_{ave} = 1.2$

B: SD ≈ 21

$SE_{sum} = 363.7$

$SE_{ave} = 1.2$

$$SE_{\substack{diff \\ ave}} = \sqrt{1.2^2 + 1.2^2} = 1.7$$

$$z = \frac{67-77}{1.7} = -5.9$$

P-value is tiny (off the chart)

We reject the null hypothesis & conclude that the tests are not equally difficult - version A seems to be harder.

Note: this is a treatment/control example

10. A researcher is interested in the extent to which lead particulates emitted from automobiles are absorbed by competitive cyclists. For a group of 30 cyclists they find the following:

Hours of training:	average = 16.2,	SD = 5.9	
Blood lead (μmol/L):	average = .42,	SD = .19,	$r = 0.6$.

(a) (6 points) Find the regression line for predicting blood lead from training time.

$$\text{slope} = r\frac{SD_y}{SD_x} = \frac{.6 \times .17}{5.9} = .019$$

$$\text{int} = \text{ave}_y - \text{slope} \cdot \text{ave}_x$$
$$= .42 - (.019 \times 16.2) = .11$$

(b) (4 points) Predict the blood lead for a cyclist who trained for 21 hours.

$$.11 + (.019 \times 21) = .509$$

Name:

Stat 1040, Summer 2000
Test 1, June 26

Show all your work. The test is out of 100 points and you have 60 minutes, so budget your time accordingly.

1. Forty men have agreed to be subjects in an experiment on the effectiveness of a new throat spray that is supposed to reduce snoring. These people will be divided into a treatment group and a control group.

   (a) (6 points) Which of the following is the best way to form these two groups, and why?

   i. For each person, flip a coin to see which group he will be in.

   ii. Have a physician divide them into two groups of size 20 each such that the two groups are roughly the same with respect to general health.

   (i) is better because it avoids any unintentional (or intentional!) personal bias creeping in.

   (For the rest of the question, assume that the two groups were selected properly at this stage.)

   (b) (5 points) Once the two groups are chosen, each of the 40 men is given a spray bottle and told to use it every night for a week. Why are *all* the men given spray bottles? Are all the contents the same? Explain.

   They are all given spray bottles to make the experiment blind i.e. to minimize the effects of people reacting to the idea of treatment rather than the treatment itself. The placebo group (control group) get a spray that lacks the active ingredient

   (c) (6 points) To find out whether the spray works, at the end of the study the men are asked the following question:

   > *Do you think that with the spray you snore more than before, less than before, or about the same as before?*

   It turns out that on average the men in *both groups* thought that they snored less than before. One explanation why the treatment group might answer this way is that the spray works. But why did the men in the control group answer this way? Provide two plausible reasons.

   - they thought they were getting the treatment so they only thought they snored less
   - the placebo had some effect
   - they actually snored less because of psychosomatic effects

   (d) (3 points) Suggest a better way to evaluate the effectiveness of the spray.

   Record it or have an observer measure it_ self reporting is a bit unreliable for this!! They should record/observe before & after & compare.

2. (7 points) In a survey of 600 high school students, it was discovered that the GPA's of those from households with a PC were generally higher than those from households with no PC's. One PC company sees these data and claims that buying PC's improves students' GPA's, on average. Suggest a possible confounding factor that could invalidate the company's claim. Explain briefly but carefully why this is a confounding factor.

Wealth - wealthy families can afford a PC and perhaps these families can provide a better study environment so their kids would have higher GPA's.

Parents' education - more educated parents are more likely to own PC's and these parents might push/encourage their kids to do well in school

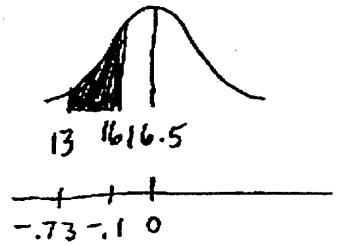
3. (6 points) The average of a list of numbers is 25 and the SD is 4. Suppose 5 is added to every number in the list. What is the average of the new list? What is the SD of the new list?

ave = 25 + 5 = 30

SD = 4 (unchanged)

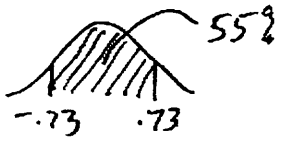
4. The lengths of 212 trout are measured. A histogram of their lengths follows the normal curve, and the average length is 16.5 cm, with an SD of 4.8 cm.

(a) (12 points) Estimate the percentage of these trout that are between 13.0 cm and 16.0 cm in length.



$\frac{16 - 16.5}{4.8} = -0.10$

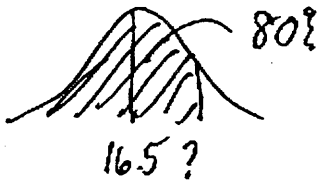
$\frac{13 - 16.5}{4.8} = -0.73$



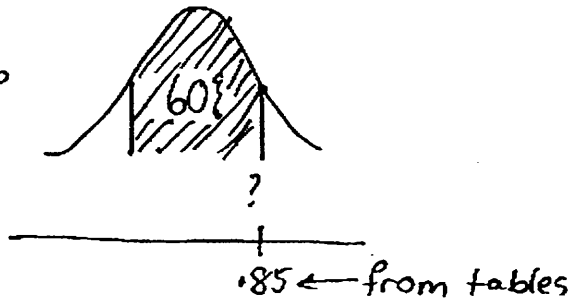
55% - 8% = 47%

so the area we want is $\frac{47\%}{2} = 23\frac{1}{2}\%$

(b) (12 points) Find a length such that about 80% of the trout are shorter than this length.



so



so length = ave + .85 × SD

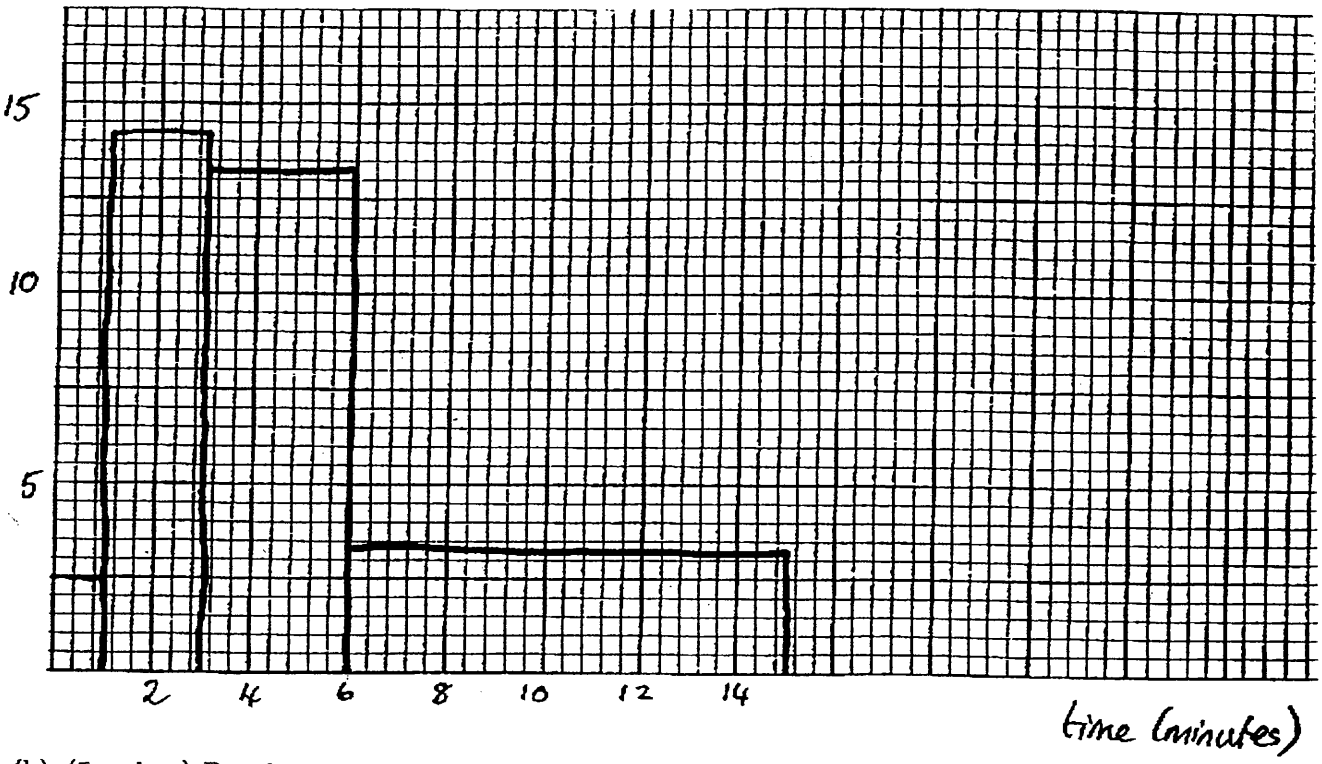
= 16.5 + .85 × 4.8

= 20.6 cm

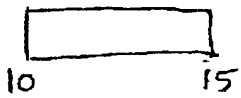
5. The waiting times, in minutes, of 80 customers in the checkout line of a supermarket are summarized below. The intervals include the left endpoint but not the right.

width	Waiting time (minutes)	Number of people	%	height = %/width
1	0-1	2	2.5	2.5
2	1-3	23	28.75	14.4
3	3-6	32	40	13.3
9	6-15	23	28.75	3.2
		80	100	

(a) (12 points) Draw a histogram for these data. Label both of the axes.



(b) (5 points) People who stand in line longer than 10 minutes get a candy bar. Approximately what percentage of people get candy bars? (Do *not* use the normal curve - use the histogram).



width = 5
height = 3.2
} area = $5 \times 3.2 = 16$

16% get candy!

(c) (5 points) Would it be appropriate to use the normal curve to approximate the percentage of people who get candy bars? Explain (no calculation is required).

No it would not be appropriate to use the normal curve because this histogram does not look like the normal curve - it's quite skew.

6. From the subjects in a health survey, the following data were collected:

Average height = 68 inches	SD 2.5 inches	x
Average blood pressure = 120 mm	SD 15 mm	y
correlation r = -0.2		

The scatter-diagram is football-shaped.

(a) (6 points) The correlation coefficient tells us that on average, taller men have (higher/lower) blood pressures than shorter men, and that the relationship between blood pressure and height is quite (strong/weak). *(cross out the wrong word in each pair).*

(b) (6 points) Find the equation of the regression line for predicting blood pressure from height.

$\text{slope} = r\frac{SD_y}{SD_x} = -.2 \times \frac{15}{2.5} = -1.2$

$\text{intercept} = \text{ave}_y - \text{slope} \cdot \text{ave}_x = 120 + 1.2 \times 68 = 201.6$

equation:

$\text{blood pressure} = 201.6 + -1.2 \cdot \text{height}$

$= 201.6 - 1.2 \cdot \text{height}$

(c) (6 points) Estimate the blood pressure of a man who is 73 inches tall.

$201.6 - 1.2 \times 73 = 114$ using (6)

or: 73" is 5" above average in height - that's 2 SD's.
Expect him to be $.2 \times 2 = .4$ SD's below ave in bp.
That's $.4 \times 15$ mm below 120 mm
i.e. 6 mm below 120 mm = 114 mm

(d) (3 points) If the heights of all the men were changed to centimeters, and the summary statistics were recalculated, would the correlation coefficient increase, decrease or stay the same?

It would stay the same.

Name:

Stat 1040, Summer 2000

Test 2, July 17

To ensure full credit, SHOW YOUR WORK. The test is out of 100 points.

1. In Harry Potter's school (Hogwarts), there are 4 houses: Gryffindor, Slytherin, Hufflepuff and Ravenclaw. First year students are assigned to one of these houses by the "sorting hat". Suppose the "sorting hat" just randomly picks a house for each student - so that the houses to which students get assigned are like independent draws (with replacement) from the box:

Gryffindor	Slytherin	Hufflepuff	Ravenclaw

Now consider 2 first year students.

(a) (4 points) What is the chance they both get assigned to Gryffindor?

$\frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$

(b) (4 points) What is the chance the first gets assigned to Gryffindor and the second to Slytherin?

$\frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$

(c) (4 points) What is the chance they both get assigned to the same house?

$\frac{1}{16} + \frac{1}{16} + \frac{1}{16} + \frac{1}{16} = \frac{4}{16} = \frac{1}{4}$

(d) (4 points) What is the chance they get assigned to two different houses?

$1 - \frac{1}{4} = \frac{3}{4}$ (opposite of (c))

2. (12 points) Suppose a company wants to estimate the percentage of Utah parents who would be interested in a new device to prevent their children watching certain TV channels. They take a simple random sample of 500 Utah parents and find that 324 are interested in the device. Find a 95% confidence interval for the percentage of all Utah parents who would be interested in the device.

f: -3

sample percent = $\frac{324}{500} \times 100\% = 64.8\%$

wrong center -5

SD: -5

324 [1] 176 [0] $SD_{box} = \sqrt{\frac{324}{500} \times \frac{176}{500}} = .478$

$SE_{sum} = \sqrt{500} \times .478 = 10.7$

$SE_{\%} = \frac{10.7}{500} \times 100\% = 2.1\%$ CI is $64.8 \pm 4.2\%$

3. At a local supermarket, they have a "club card" which keeps track of customer spending. The computer shows that 2762 people used the card last week. For these 2762 customers, the average amount spent last week was \$47.63 with an SD of \$32.12.

(a) (14 points) If we take a random sample of 100 of these 2762 customers, what is the chance that the average spending of the 100 customers in the sample was more than \$50.00?

2762 tickets

[$] [$] $ave_{box} = 47.63$, $SD_{box} = 32.12$

D: 48% out of 8

$EV_{ave} = 47.63$

$SE_{sum} = \sqrt{100} \times 32.12 = 321.2$

$SE_{ave} = \frac{321.2}{100} = 3.212$

55%

22½%

47.63 50

0 .74

$\frac{50 - 47.63}{3.212} = .74$

(b) (5 points) Is your answer in a) valid if the spending does not follow the normal curve? Explain.

It's valid - we were looking at the average of a large no of draws (100).

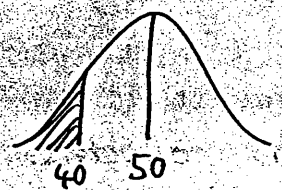
4. (14 points) Suppose it is known that 10% of all people in Utah have a specific blood type. If I take a simple random sample of 500 Utahns, what is the chance that fewer than 40 have that blood type?

1 9 0 $\quad ave_{box} = .1$

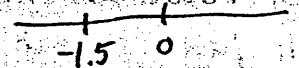
$SD_{box} = \sqrt{\frac{1}{10} \times \frac{9}{10}} = .3$

$EV_{sum} = 500 \times .1 = 50$

$SE_{sum} = \sqrt{500} \times .3 = 6.7$



$\frac{40-50}{6.7} = -1.5$



5. When the latest Harry Potter book was released, a local bookstore sold 500 copies in one day. They randomly sampled 300 of these people and found that the average age for these 300 people was 13.3 with an SD of 8.2.

(a) (15 points) Find a 95% confidence interval for the average age of all 500 customers.

$SD_{box} \approx 8.2$

$SE_{sum} = \sqrt{300} \times 8.2 = 142$

$SE_{ave} = \frac{142}{300} = .473$ $\qquad cf = \sqrt{\frac{500-300}{499}} = .633$

$SE_{ave} = .633 \times .473 = .3$

CI: $13.3 \pm 2 \times .3$

$13.3 \pm .6$

c.f. 4

$n = 500 \Rightarrow -3$.

wrong SE: -4

wrong center: -4

(b) True or false:

F i. (3 points) 95% of the 300 sampled customers had ages in the interval from part a).

T ii. (3 points) We cannot approximate the percentage of customers who had ages in the interval from part a) because the ages do not follow the normal curve.

F iii. (3 points) The interval from part a) is invalid because the ages do not follow the normal curve.

6. (3 points) I'm interested in estimating the percentage of USU summer school students who have read at least one of the Harry Potter books. I get a list of summer classes and randomly choose 10 of them. I send a representative to each of these 10 classes to find out how many students have read at least one of the Harry Potter books. Assuming all students are present that day, would this be:

(a) a simple random sample.

(b) a cluster sample.

(c) a sample of convenience (not a probability method).

7. (4 points) True or false and explain: a very large sample (say a million people or more) will usually give accurate results no matter how the sample was obtained.

False - it depends how the sample is collected. If there is a bias in the sampling procedure (eg towards the wealthy) a large sample won't help - it will just repeat the mistake on a grander scale!!

8. (4 points) True or false and explain: the law of averages says that when tossing a fair coin, as the number of tosses increases, the number of heads is more and more likely to be exactly equal to half the number of tosses.

False - it's less & less likely to be exactly half the number of tosses but more & more likely to be close in percentage terms.

9. (4 points) A local politician is interested in estimating the percentage of voters who are opposed to the marriage tax. She can only afford to sample 1000 people. Other things being equal, to get equal accuracy in Logan and Salt Lake City, she should sample:

(a) 500 people in Logan and 500 people in Salt Lake City.

(b) more people in Salt Lake City than in Logan.

(c) more people in Logan than in Salt Lake City.

Name:

Stat 1040, Summer 2000
Test 3, Friday Aug 4, 7:30–9:50 am

For full partial credit, show your work.

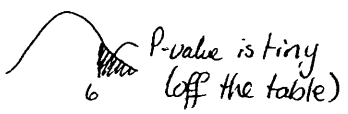
1. (15 points) In a study to evaluate the effectiveness of a promising treatment for malignant melanoma, researchers carefully conducted a randomized, controlled, double-blind experiment. At the end of 5 years, they found that of the 355 patients in the treatment group, 96 were still alive. Of the 371 patients in the control group, 37 were still alive. Is this treatment effective in terms of the 5-year survival rate of such patients? State the null and alternative hypotheses, perform the appropriate statistical test, and clearly state your conclusions.

null: treatment is not effective (in terms of 5-year survival)

alt: " is " " " " "

treatment	control
sample % $= \frac{96}{355} \times 100\% = 27\%$	sample % $= \frac{37}{371} \times 100\% = 10\%$
$SD_{box} \approx \sqrt{\frac{96}{355} \times \frac{259}{355}} = .444$	$SD_{box} \approx \sqrt{\frac{37}{371} \times \frac{334}{371}} = .300$
$SE_{sum} = \sqrt{355} \times .444 = 8.36$	$SE_{sum} = \sqrt{371} \times .300 = 5.78$
$SE_{\%} = 2.36\%$	$SE_{\%} = 1.56\%$

$SE_{diff} = \sqrt{2.36^2 + 1.56^2} = 2.8$ $\quad z = \frac{27-10}{2.8} = 6$



P-value is tiny (off the table)

So we reject the null & conclude that the treatment works!

2. (10 points) Suppose I teach a class of 200 students, 100 men and 100 women. I give a comprehensive examination, and I find that the average score for the men is 75.4 with an SD of 10.2 and the average score for the women is 78.5 with and SD of 10.8. (Both sets of scores follow the normal curve closely.) Clearly explain why it is not apporopriate to perform a statistical test to decide whether the men and women have different average exam scores.

It is not appropriate because we have the entire population - we know the average for the women is higher by $78.5 - 75.4 = 3.1$ points. We don't have random samples or a randomized experiment so we can't generalize.

3. True or false: (5 points each)

F (a) For the same sample, the P-value for a 1-tailed test will be twice as big as the P-value for a 2-tailed test.

T (b) For a very large sample, a "statistically significant" result may not be very important in a practical sense.

T (c) For a large enough sample, a z-test will be valid even if the sample values are somewhat skew.

T (d) We reject the null hypothesis when the P-value is small.

T (e) If we perform many statistical tests, we expect to find some "statistically significant" results just by chance.

4. Soil is said to be "neutral" if the pH is 7.0 and "acidic" if the pH is less than 7.0. A soil scientist is concerned that the soil in a certain field might be somewhat acidic. She takes 5 randomly selected samples of soil from the field and finds that the pH levels are 5.8, 6.3, 6.9, 6.2, and 5.5. average = 6.14 SD = .53

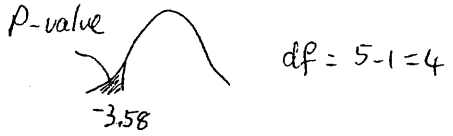
(a) (15 points) Suppose it is known that in this sort of situation, pH values follow the normal curve quite closely. Test to see whether her suspicion is correct. You must state the null and alternative hypotheses, perform the appropriate statistical test, and clearly state your conclusions.

null: soil is neutral (average is 7.0)
alt: soil is acidic (average is less than 7.0)

$t = \frac{6.14 - 7.0}{.24}$ $\quad SE_{sum} = \sqrt{5} \times .53 = 1.19$

$SE_{ave} = \frac{1.19}{5} = .24$

$= -3.58$



df = 5-1 = 4

The P-value is between 1% and 2.5% so we reject the null & conclude that she is right - it's acidic.

(b) (5 points) If she knew that in this sort of situation, pH values did not follow the normal curve at all well, would her test still be valid? Explain.

No - we need the values to be normal for the t-test to be valid.

What *should* she do (assuming she has lots of time and money!)?

Get more data.

5. (15 points) The Highway Patrol Department wants to assess if the cause of an accident is related to the outcome of the accident. The department takes a simple random sample of 250 accidents and finds the following results:

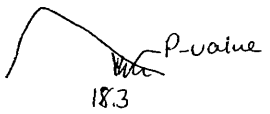
		Outcome: Death	Outcome: No death		expected: Death	expected: No death
Cause	Speeding	21	41	62	21.3	40.7
	Recklessness	13	30	43	14.8	28.2
	Fatigue	28	22	50	17.2	32.8
	Alcohol	19	38	57	19.6	37.4
	Other	5	33	38	13.1	24.9
		86	164	250		

Perform the appropriate statistical test. You must clearly state a null and alternative hypothesis, compute a test statistic and a P-value and clearly state your conclusions.

obs	exp	$(obs-exp)^2/exp$
21	21.3	.004
13	14.8	.219
28	17.2	6.78
19	19.6	.018
5	13.1	5.00
41	40.7	.002
30	28.2	.115
22	32.8	3.56
38	37.4	.0096
33	24.9	2.6
		18.3

null: outcome and cause are not related (indep)
alt: " " " are related (dependent)

$\chi^2 = 18.3$ on $(5-1)(2-1) = 4$ d.f.



The P-value is tiny (off the table) so we reject the null and conclude that outcome and cause are related.

6. (15 points) A magazine claims that "50% of young mothers say that finding child care is a problem". A child care referral group thinks the percentage is actually lower than 50% in their community. They take a simple random sample of 450 young mothers and find that 185 say that finding child care is a problem. Test to see whether the child care referral group is correct. You must clearly state a null and alternative hypothesis, compute a test statistic and a P-value and clearly state your conclusions.

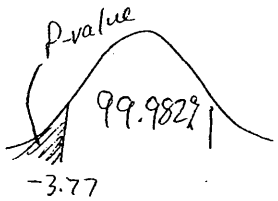
null: 50% say finding child care is a problem.
alt: less than 50% say that finding child care is a problem.

$ave_{box} = .5$
$SD_{box} = .5$

$SE_{sum} = \sqrt{450} \times .5 = 10.6$

$EV_{sum} = 450 \times .5 = 225$

$Z = \frac{185 - 225}{10.6} = -3.77$



The P-value is $(100-99.982)/2 = .009\%$ which is much smaller than 5% so we reject the null & conclude that in their community the percentage of young mothers who say that finding child care is a problem is less than 50%.

Name:

Stat 1040, Fall 2000
Final Test, Thursday December 14, 12:00–1:50 pm

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

1. The U.S. Bureau of Labor Statistics regularly collects information on the labor market. According to the bureau, workers employed in manufacturing industries earned an average \$577 per week in May 1999. Assume that this average is based on a simple random sample of 1600 workers selected from manufacturing industries and that the standard deviation of weekly earnings in this sample is \$100.

   (a) (10 points) Find a 99% confidence interval for the average weekly earnings of all U.S. workers employed in manufacturing industries in May 1999.

   $\$577 \pm (2.6 \times 2.5) = \$577 \pm 6.5$ i.e. \$570.50 to \$583.50

   $SD_{box} \approx 100$

   $SE_{sum} = \sqrt{1600} \times 100 = 4000$

   $SE_{ave} = \frac{4000}{1600} = 2.5$

   (b) (4 points) We know that income does not follow the normal curve. Since this is the case, does the confidence interval you found in part a) make much sense? Explain.

   Yes - we have a very large number of draws, so the probability histogram for the average will follow the normal curve even if the tickets in the box do not.

lence the
I is
valid

2. (6 points) After Florida Secretary of State Katherine Harris certified the Florida results in the 2000 presidential election as a 537-vote Bush win, a Salt Lake City TV station held a call-in poll. The station asked its viewers to call in with their answers to the question:

   "Do you believe Al Gore should stop trying to overturn legally certified votes in Florida and acknowledge that he has lost?"

   4,237 people phoned in, and 3,482 said "yes". The station then announced that "82% of Americans believe Al Gore should concede defeat."

   Identify at least three problems with this survey and the announced results.

   It's a phone-in survey - no random sampling, viewers only (people who feel strongly tend to respond)

   It's conducted in Salt Lake ⇒ conclusions are for "Americans"

   Question is emotionally loaded, conclusions are not.
   i.e. they didn't ask "Do you believe Al Gore should concede defe

3. (4 points) A political poll takes a simple random sample of 5000 registered voters from Palm Beach (around 400,000 registered voters) and an independent simple random sample of 1000 registered voters from Collier (around 80,000 registered voters). They plan to estimate the percentage of registered voters who think they voted for Al Gore. If everything else is similar for these two counties,

   (a) the accuracy in Palm Beach should be quite a bit lower than the accuracy in Collier.

   (b) the accuracy in Palm Beach should be about the same as the accuracy in Collier.

   (c) the accuracy in Palm Beach should be quite a bit higher than the accuracy in Collier.

   (Say which is true, no explanation is required.)

4. (4 points) Suppose that in a certain large Florida county, there are exactly 50% Bush voters and 50% Gore voters.

   A: getting exactly 50 Bush voters in a simple random sample of 100 voters from that county.

   B: getting exactly 100 Bush voters in a simple random sample of 200 voters from that county.

   Which of the following is true (no explanation is required):

   (a) A is more likely than B.

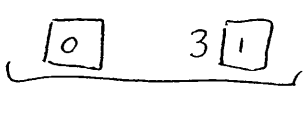
   (b) A and B are equally likely.

   (c) B is more likely than A.

5. (12 points) According to genetic theory, if we look at parents who are both heterozygous RH-positive, there is a 75% chance for their child to be RH-positive. I take a random sample of 400 children of such parents and I find that that 312 are RH-positive. Does the genetic theory seem to be appropriate for the population from which I sampled? Perform a statistical test. You must clearly state a null and alternative hypothesis, compute a test statistic and a P-value and state your conclusions.

null: the genetic theory is OK

alt: " " " " not OK



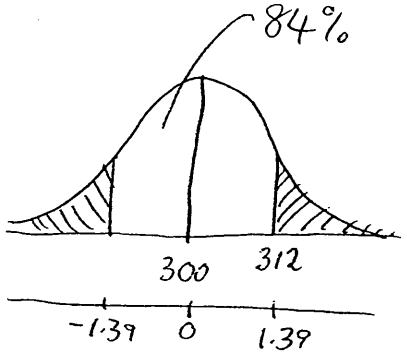
$ave_{box} = .75$

$SD_{box} = .433$

$EV_{sum} = 400 \times .75 = 300$

$SE_{sum} = \sqrt{400} \times .433 = 8.66$

$z = \frac{312 - 300}{8.66} = 1.39$



The P-value is 16% so we fail to reject the null. We conclude there is no evidence that the theory does not hold for this population.

6. "Do angry people have more heart disease?" In a 4-year study of a random sample of 8,474 people with normal blood pressure, their anger was measured using a specially designed test which measures how prone a person is to sudden anger. At the end of the study, physicians (who did not know the anger test results) determined whether the people had heart disease or not. The following results were obtained:

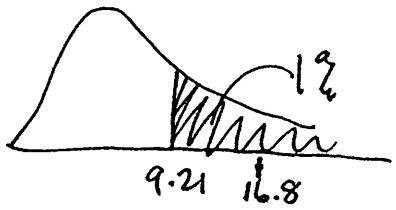
190 is 2.24% of 8474
2.24% of 3110 is 70
2.24% of 4731 is 106

	Anger score			
	Low anger	Moderate anger	High anger	Total
Heart disease	53 70	110 106	27 14	190
No Heart disease	3057 3040	4621 4625	606 619	8284
Total	3110	4731	633	8474

(a) (12 points) Is there evidence that anger and heart disease are related? You must clearly state a null and alternative hypothesis, compute a test statistic and a P-value and state your conclusions.

null: heart disease & anger are unrelated
alt: " " " " related

obs	exp	$(obs-exp)^2/exp$
53	70	4.1
3057	3040	.1
110	106	.2
4621	4625	.0
27	14	12.1
606	619	.3
		16.8



P-value is less than 1% so we reject the null & conclude that heart disease & anger are related.

$df = (3-1)\times(2-1) = 2$

(b) (3 points) This study is an example of

(i) an observational study.
ii. a controlled, blind experiment (not randomized).
iii. a randomized, controlled, blind experiment.

(Just say which one, no explanation is required).

(c) (5 points) Suggest a possible confounding factor for this study and clearly explain why it might be a confounding factor.

One example might be stress - stress might cause anger & stress might also be either causing high blood pressure or caused by high blood pressure (& hence heart disease).

(d) (3 points) Assuming the data show that anger and heart disease *are* related, does this show that anger causes heart disease? Explain clearly.

No - as the confounding factor shows, since this is an observational study we cannot conclude causation.

7. **For a group of 19 countries, researchers looked at the average wine alcohol consumption (average number of liters of wine alcohol per person per year) and the heart disease death rate (number of deaths from heart disease per 100,000 people per year) for each country. They found the following:**

alcohol
Wine Consumption (liters per year) : average = 3.03, SD = 2.4
Heart disease death rate (per 100,000) : average = 191.0, SD = 66.6, $r = -0.84$.

(a) (6 points) Find the equation of the regression line for predicting a country's heart disease death rate from its average alcohol consumption.

$\text{slope} = -.84 \times \frac{66.6}{2.4} = -23.31$

$\text{intercept} = 191.0 + 23.31 \times 3.03 = 261.63$

death rate = 261.63 − 23.31 · wine alcohol consumption

(b) (4 points) The correlation coefficient of $r = -0.84$ shows which (if any) of the following:

i. (T)/F countries where people drink more wine alcohol tend to have lower heart disease death rates, and vice versa.

ii. T/(F) the more wine alcohol a person drinks, the lower their chances of developing heart disease.

iii. T/(F) there is a strong negative association between the amount of wine alcohol a person drinks and their chances of dying from heart disease.

iv. (T)/F this is an example of an ecological correlation so the correlation coefficient is artificially close to -1.

For EACH ONE say whether it is true or false, no explanation is required. As always, we are asking which statements are correct interpretations of the correlation coefficient, not about what you believe to be true or untrue about alcohol consumption and heart disease.

8. I have 20 lightbulbs a large box. Unknown to me, 5 of these 20 bulbs are broken. I select 6 bulbs at random from these 20 bulbs to put in a chandelier. *Answer each of the following questions separately.*

(a) (1 points) What is the chance that the first bulb works?

$\frac{15}{20}$

(b) (3 points) What is the chance that the second bulb works?

$\frac{15}{20}$

(c) (3 points) What is the chance that all 6 of the bulbs work?

$\frac{15}{20} \times \frac{14}{19} \times \frac{13}{18} \times \frac{12}{17} \times \frac{11}{16} = .1937$

9. (12 points) The spermicide nonoxynol-9 kills HIV in the test tube, so researchers hypothesized that it might be useful in protecting high-risk women from HIV. Other researchers argued that nonoxynol-9 might increase the risk because it is an irritant. In a study of 990 prostitutes, participants were randomly divided into two groups. The treatment group were given a nonoxynol-9 gel. The control group were given a similar-looking but inactive gel. When the study ended in May 2000, 67 of the 495 women in the treatment group were HIV-positive, and 44 of the 495 women in the control group were HIV-positive. Perform a 2-tailed test to decide whether the treatment and control groups were significantly different with respect to HIV. Clearly state your conclusions.

null: treatment makes no difference to HIV

alt: " does make a difference

Treatment: $SD \approx \sqrt{\frac{67}{495} \times \frac{428}{495}} = .342$

% is $\frac{67}{495} \times 100\% = 13.5\%$

$SE_{sum} = \sqrt{495} \times .342 = 7.6$

$SE_{\%} = \frac{7.6}{495} \times 100\% = 1.5\%$

Control: $SD = \sqrt{\frac{44}{495} \times \frac{451}{495}} = .285$

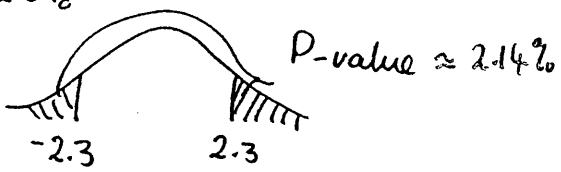
% is $\frac{44}{495} \times 100\% = 8.9\%$

$SE_{sum} = \sqrt{495} \times .285 = 6.3$

$SE_{\%} = \frac{6.3}{495} \times 100\% = 1.3\%$

$SE_{diff} = \sqrt{1.5^2 + 1.3^2} = 2.0\%$

$z = \frac{13.5\% - 8.9\%}{2.0\%} = 2.3$



The P-value is smaller than 5% so we reject the null & conclude that the nonoxynol-9 has increased the HIV-positive rate.

10. (3 points) The study in problem 9 is an example of

(a) an observational study.

(b) a controlled, blind experiment (not randomized).

(c) a randomized, controlled, blind experiment.

(d) a simple random sample.

(Just say which one, no explanation is required).

11. (5 points) The National Collegiate Athletic Association requires colleges to report the graduation rates of their athletes. At USU, 62% of all 157 student athletes who entered between Fall 1990 and Fall 1993 graduated within 6 years of starting. True or false, and explain:

An approximate 95% confidence interval for the percentage of all USU athletes who graduate within 6 years of starting runs from 54.3% to 69.7%.

False. What's the population? If all USU athletes ever, then we don't have a random sample. If all USU athletes entering Fall '90 to Fall '93, we have the entire population.