Stats Exam Review
1st Semester

(1) **Beanstalks.** Beanstalk Clubs are social clubs for very tall people. To join, a man must be over 6'2" tall, and a woman over 5'10". The National Health Survey suggests that heights of adults may be Normally distributed, with mean heights of 69.1" for men and 64.0" for women. The respective standard deviations are 2.8" and 2.5".

a) You are probably not surprised to learn that men are generally taller than women, but what does the greater standard deviation for men's heights indicate?

b) Who are more likely to qualify for Beanstalk membership, men or women? Explain.

(2) **Acid rain.** Based on long-term investigation, researchers have suggested that the acidity (pH) of rainfall in the Shenandoah Mountains can be described by the Normal model $N(4.9,0.6)$.

a) Draw and carefully label the model.

b) What percent of storms produce rain with pH over 6?

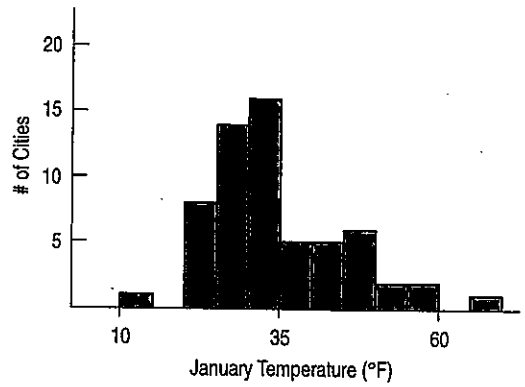
c) What percent of storms produce rainfall with pH under 4?

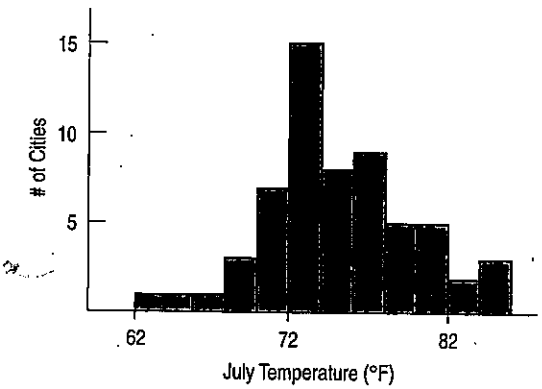
d) The lower the pH, the more acidic the rain. What is the pH level for the most acidic 20% of all storms?

e) What is the pH level for the least acidic 5% of all storms?

f) What is the IQR for the pH of rainfall?

(3) **Seasons.** Average daily temperatures in January and July for 60 large U.S. cities are graphed in the histograms below.





(4) **Music and memory.** Is it a good idea to listen to music when studying for a big test? In a study conducted by some Statistics students, 62 people were randomly assigned to listen to rap music, Mozart, or no music while attempting to memorize objects pictured on a page. They were then asked to list all the objects they could remember. Here are the 5-number summaries for each group:

	n	Min	Q1	Median	Q3	Max
Rap	29	5	8	10	12	25
Mozart	20	4	7	10	12	27
None	13	8	9.5	13	17	24

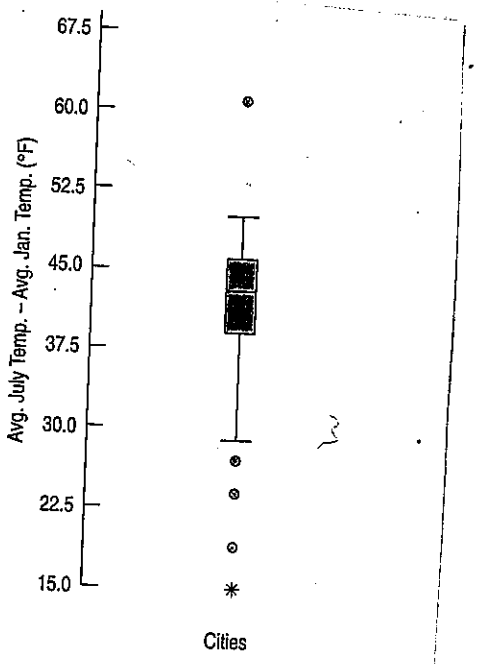
a) Describe the W's for these data: *Who, What, Where, Why, When, How.*

b) Name the variables and classify each as categorical or quantitative.

c) Create parallel boxplots as best you can from these summary statistics to display these results.

d) Write a few sentences comparing the performances of the three groups.

(3)



a) What aspect of these histograms makes it difficult to compare the distributions?

b) What differences do you see between the distributions of January and July average temperatures?

c) Differences in temperatures (July–January) for each of the cities are displayed in the boxplot above. Write a few sentences describing what you see.

(5)

Mail. Here are the number of pieces of mail received at a school office for 36 days.

123	70	90
80	78	72
52	103	138
112	92	93
118	118	106
95	131	59
151	115	97
100	128	130
66	135	76
143	100	88
110	75	60
115	105	85

a) Plot these data.
b) Find appropriate summary statistics.
c) Write a brief description of the school's mail deliveries.
d) What percent of the days actually lie within one standard deviation of the mean? Comment.

(6)

Birth order. Are first-born children more likely to be interested in science, or perhaps younger siblings in, say, the humanities? A Statistics professor at a large university polled his students to find out what their majors were and what position they held in the family birth order. The results are summarized in the table.

a) What percent of these students are oldest or only children?
b) What percent of Humanities majors are oldest children?
c) What percent of oldest children are Humanities students?
d) What percent of the students are oldest children majoring in the Humanities?

Major	Birth Order*				
	1	2	3	4+	Total
Math/Science	34	14	6	3	57
Agriculture	52	27	5	9	93
Humanities	15	17	8	3	43
Other	12	11	1	6	30
Total	113	69	20	21	223

* 1 = oldest or only child

(7)

Age and party. The Gallup Poll conducted a representative telephone survey during the first quarter of 1999. Among their reported results was the following table concerning the preferred political party affiliation of respondents and their ages.

Age	Party			
	Rep.	Dem.	Ind.	Total
18–29	241	351	409	1001
30–49	299	330	370	999
50–64	282	341	375	998
65+	279	382	343	1004
Total	1101	1404	1497	4002

a) What percent of people surveyed were Republicans?
b) Do you think this might be a reasonable estimate of the percentage of all voters who are Republicans? Explain.
c) What percent of people surveyed were under 30 or over 65?
d) What percent of people were Independents under the age of 30?
e) What percent of Independents were under 30?
f) What percent of people under 30 were Independents?

(8)

Public opinion. For many years Martha Stewart was a popular expert in home decorating, an arbiter of good taste, and a very successful businesswoman. In June 2002 she first came under attack amidst rumors of insider stock trading. At that time a series of Gallup polls each contacted over 1000 people to ask about their overall opinion of her. Those polled could answer favorable, unfavorable, or that they did not know who she was. Results are summarized in the following table.

Opinion	Poll Results		
	Oct. 1999	June 2002	July 2002
Favorable	49%	46%	30%
Unfavorable	16%	27%	39%
Don't know	32%	23%	27%

a) Each poll should total 100%. Can you think of a reason why these do not?
b) How could the number of people who do not know who Martha Stewart is *increase* from June to July? Did people forget her? Or perhaps the poll is flawed? Explain how these results could be valid.
c) Display these results in a bar graph.
d) Display these results with pie charts.
e) Display these results with a timeplot.
f) Which display do you think best depicts these data? Why?
g) Write a few sentences describing Martha Stewart's public image at that time.

(9) **Bike safety.** The Massachusetts Governor's Highway Safety Bureau's report on bicycle injuries for the years 1991–2000 included the counts shown in the table.

Year	Bicycle Injuries Reported
1991	1763
1992	1522
1993	1452
1994	1370
1995	1380
1996	1343
1997	1312
1998	1275
1999	1030
2000	1118

a) What are the W's for these data?
b) Display the data in a stem-and-leaf display.
c) Display the data in a timeplot.
d) What is apparent in the stem-and-leaf display that is hard to see in the timeplot?
e) What is apparent in the timeplot that is hard to see in the stem-and-leaf display?
f) Write a few sentences about bicycle injuries in Massachusetts.

(10) **Profits.** Here is a stem-and-leaf display showing profits as a percent of sales for 29 of the *Forbes* 500 largest U.S. corporations. The stems are split; each stem represents a span of 5%, from a loss of 9% to a profit of 25%.

```
-0 | 99
-0 | 1234
 0 | 111123444
 0 | 5555679
 0 | 0013
   |
 2 | 2
 2 | 5
```

Profits (% of sales)
(-0 | 3 means a loss of 3%)

a) Find the 5-number summary.
b) Draw a boxplot for these data.
c) Find the mean and standard deviation.
d) Describe the distribution of profits for these corporations.

(11) **Grades.** A Statistics instructor created a linear regression equation to predict students' final exam scores from their midterm exam scores. The regression equation was $\widehat{fin} = 10 + 0.9\,mid$.

a) If Susan scored a 70 on the midterm, what did the instructor predict for her score on the final?
b) Susan got an 80 on the final. How big is her residual?
c) Suppose that the standard deviation of the final was 12 points and the standard deviation of the midterm was 10 points. What is the correlation between the two tests?
d) How many points would someone need to score on the midterm to have a predicted final score of 100?
e) Suppose someone scored 100 on the final. Explain why you can't estimate this student's midterm score from the information given.
f) One of the students in the class scored 100 on the midterm but got overconfident, slacked off, and scored only 15 on the final exam. What is the residual for this student?
g) No other student in the class "achieved" such a dramatic turnaround. If the instructor decides not to include this student's scores when constructing a new regression model, will the R^2 value of the regression increase, decrease, or remain the same? Explain briefly.
h) Will the slope of the new line increase or decrease?

(12) **Smoking and pregnancy.** The organization Kids Count monitors issues related to children. The table shows a 50-city average of the percent of expectant mothers who smoked cigarettes during their pregnancies.

Year	% Smoking While Pregnant	Year	% Smoking While Pregnant
1990	17.7	1995	12.7
1991	17.0	1996	11.9
1992	16.0	1997	11.2
1993	14.9	1998	10.8
1994	13.9	1999	10.4

a) Create a scatterplot and describe the trend you see.
b) Find the correlation.
c) How is the value of the correlation affected by the fact that the data are averages rather than percentages for each of the 50 cities?
d) Write a linear model and interpret the slope in this context.

No smoking? The downward trend in smoking you saw in the last exercise is good news for the health of babies, but will it ever stop?

a) Explain why you can't use the linear model you created in Exercise 22 to see when smoking during pregnancy will cease altogether.
b) Create a model that could estimate the year in which the level of smoking would be 0%.
c) Comment on the reliability of such a prediction.

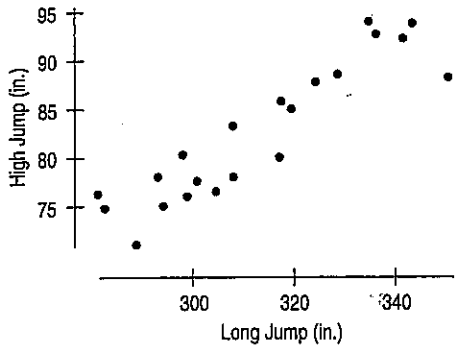
13 **Winter in the city.** Summary statistics for the data relating the latitude and average January temperature for 55 large U.S. cities are given below.

Variable	Mean	StdDev
Latitude	39.02	5.42
JanTemp	26.44	13.49

Correlation = −0.848

a) What percent of the variation in January *temperatures* can be explained by variation in *latitude*?
b) What is indicated by the fact that the correlation is negative?
c) Write the equation of the line of regression for predicting January *temperature* from *latitude*.
d) Explain what the slope of the line means.
e) Do you think the *y*-intercept of the line is meaningful? Explain.
f) The latitude of Denver is 40° N. Predict the mean January temperature there.
g) If the residual for a city is positive, what does that mean?

14 **Jumps.** How are Olympic performances in various events related? The plot shows winning long jump and high jump distances, in inches, for the 20th century Olympic Games.



a) Describe the association.
b) Do long jump performances somehow influence the high-jumpers? How do you account for the relationship you see?
c) The correlation for the given scatterplot is 0.92, but at the Olympics these jumps are actually measured in meters rather than inches. Does that make the actual correlation higher or lower?
d) What would you predict about the long jump in a year when the high-jumper jumped one standard deviation better than the average high jump?

15 **Modeling jumps.** Here are the summary statistics for the Olympic long jumps and high jumps displayed in the scatterplot above.

Event	Mean	StdDev
Long Jump	314.10	20.71
High Jump	83.04	7.26

Correlation = 0.917

a) Write the equation of the line of regression for estimating *high jump* from *long jump*.
b) Interpret the slope of the line.
c) In a year when the long jump is 340 inches, what high jump would you predict?
d) Why can't you use this line to estimate the long jump for a year when you know the high jump was 85 inches?
e) Write the equation of the line you need to make that prediction.

16 **Tobacco and alcohol.** Are people who use tobacco products more likely to consume alcohol? Here are data on household spending (in pounds) taken by the British Government on 11 regions in Great Britain. Do tobacco and alcohol spending appear to be related? What questions do you have about these data? What conclusions can you draw?

Region	Alcohol	Tobacco
North	6.47	4.03
Yorkshire	6.13	3.76
Northeast	6.19	3.77
East Midlands	4.89	3.34
West Midlands	5.63	3.47
East Anglia	4.52	2.92
Southeast	5.89	3.20
Southwest	4.79	2.71
Wales	5.27	3.53
Scotland	6.08	4.51
Northern Ireland	4.02	4.56

17 **Williams vs. Texas.** Here are the average weights of the football team for the University of Texas for various years in the 20th century.

Year	Weight (lb)
1905	164
1919	163
1932	181
1945	192
1955	195
1965	199

a) Fit a straight line to the relationship of *weight* by *year* for Texas football players.
b) According to these models, in what year will the predicted weight of the Williams College team from Exercise 36 first be more than the University of Texas team?
c) Do you believe this? Explain.

Vehicle weights. The Minnesota Department of Transportation hoped that they could measure the weights of big trucks without actually stopping the vehicles by using a newly developed "weigh-in-motion" scale. After installation of the scale, a study was conducted to find out whether the scale's readings correspond to the true weights of the trucks being monitored. In Exercise 16 of Chapter 7, you examined the scatterplot for the data they collected, finding the association to be approximately linear with $R^2 = 93\%$. Their regression equation is $\widehat{Wt} = 10.85 + 0.64\ scale$, where both the scale reading and the predicted weight of the truck are measured in thousands of pounds.

a) Estimate the weight of a truck if this scale read 31,200 pounds.
b) If that truck actually weighed 32,120 pounds, what was the residual?
c) If the scale reads 35,590 pounds, and the truck has a residual of −2440 pounds, how much does it actually weigh?
d) In general, do you expect estimates made using this equation to be reasonably accurate? Explain.
e) If the police plan to use this scale to issue tickets to trucks that appear to be overloaded, will negative or positive residuals be a greater problem? Explain.

Football weights. The Sears Cup was established in 1993 to honor institutions that maintain a broad-based athletic program, achieving success in many sports, both men's and women's. Since its Division III inception in 1995, the cup has been won by Williams College in every year except one. Their football team has a 85.3% winning record under their current coach. Why does the football team win so much? Is it because they're heavier than their opponents? The table shows the average team weights for selected years from 1973 to 1993.

Year	Weight (lb)	Year	Weight (lb)
1973	185.5	1983	192.0
1975	182.4	1987	196.9
1977	182.1	1989	202.9
1979	191.1	1991	206.0
1981	189.4	1993	198.7

a) Fit a straight line to the relationship between *weight* and *year*.
b. Does a straight line seem reasonable?
c. Predict the average weight of the team for the year 2003. Does this seem reasonable?
d. What about the prediction for the year 2103? Explain.
e. What about the prediction for the year 3003? Explain.

Models. Find the predicted value of y using each model for $x = 10$.

a) $\hat{y} = 2 + 0.8 \ln x$
b) $\log \hat{y} = 5 - 0.23x$
c) $\frac{1}{\sqrt{\hat{y}}} = 17.1 - 1.66x$

Cloning. In September 1998, *USA Weekend* magazine asked, "Should humans be cloned?" Readers were invited to register a "Yes" or "No" answer by calling one of two different 900 numbers. Based on 38,023 responses, the magazine reported that "9 out of 10 readers oppose cloning."

a) Explain why you think the conclusion is not justified. Describe the types of bias that may be present.
b) Reword the question in a way that you think might create a more positive response.

When to Stop? You play a game that involves rolling a die. You can roll as many times as you want, and your score is the total for all the rolls. But . . . if you roll a 6 your score is 0 and your turn is over. What might be a good strategy for a game like this?

a) One of your opponents decides to roll 4 times, then stop (hoping not to get the dreaded 6 before then). Use a simulation to estimate his average score.
b) Another opponent decides to roll until she gets at least 12 points, then stop. Use a simulation to estimate her average score.
c) Propose another strategy that you would use to play this game. Using your strategy, simulate several turns. Do you think you would beat the two opponents?

35. Age and party. The Gallup Poll conducted a representative telephone survey during the first quarter of 1999. Among its reported results was the following table concerning the preferred political party affiliation of respondents and their ages.

Age	Party: Republican	Democratic	Independent	Total
18–29	241	351	409	1001
30–49	299	330	370	999
50–64	282	341	375	998
65+	279	382	343	1004
Total	1101	1404	1497	4002

a) What sampling strategy do you think the pollsters used? Explain.
b) What percentage of the people surveyed were Democrats?
c) Do you think this is a good estimate of the percentage of voters in the United States who are registered Democrats? Why or why not?
d) In creating this sample design, what question do you think the pollsters were trying to answer?

e) conditional distribution of democrats among 65+

f) marginal distribution of Independents

24 **Smoking and Alzheimer's.** Medical studies indicate that smokers are less likely to develop Alzheimer's disease than people who never smoked.

a) Does this prove that smoking may offer some protection against Alzheimer's? Explain.

b) Offer an alternative explanation for this association.

23 **Save the grapes.** Vineyard owners have problems with birds that like to eat the ripening grapes. Grapes damaged by birds cannot be used for winemaking (or much of anything else). Some vineyards use scarecrows to try to keep birds away. Others use netting that covers the plants. Owners really would like to know if either method works and, if so, which one is better. One owner has offered to let you use his vineyard this year for an experiment. Propose a design. Carefully indicate how you would set up the experiment, specifying the factor(s) and response variable.

26 **Knees.** Research reported in the spring of 2002 cast doubt on the effectiveness of arthroscopic knee surgery for patients with arthritis. Patients suffering from arthritis pain who volunteered to participate in the study were randomly divided into groups. One group received arthroscopic knee surgery. The other group underwent "placebo surgery" during which incisions were made in their knees, but no surgery was actually performed. Follow-up evaluations over a period of 2 years found that differences in the amount of pain relief experienced by the two groups were not statistically significant.

a) Why did the researchers feel it was necessary to have some of the patients undergo "placebo surgery"?

b) Because patients had to consent to participate in this experiment, the subjects were essentially self-selected—a kind of voluntary response group. Explain why that does not invalidate the findings of the experiment.

c) What does "statistically significant" mean in this context?

27 **NBA draft lottery.** Professional basketball teams hold a "draft" each year in which they get to pick the best available college and high-school players. In an effort to promote competition, teams with the worst records get to pick first, theoretically allowing them to add better players. To combat the fear that teams with no chance to make the playoffs might try to get better draft picks by intentionally losing late-season games, the NBA's Board of Governors adopted a weighted lottery system in 1990. Under this system the 11 teams that did not make the playoffs were eligible for the lottery. The NBA prepared 66 cards, each naming one of the teams. The team with the worst win-loss record was named on 11 of the cards, the second-worst team on 10 cards, and so on, with the team having the best record among the nonplayoff clubs getting only one chance at having the first pick. The cards were mixed, then drawn randomly to determine the order in which the teams could draft players. (Since 1995, 13 teams have been involved in the lottery, using a complicated system with 14 numbered Ping-Pong balls drawn in groups of four.) Suppose there are two exceptional players available in this year's draft and your favorite team had the third-worst record. Use a simulation to find out how likely it is that your team gets to pick first or second. Describe your simulation carefully.

28 **Security.** There are 20 first-class passengers and 120 coach passengers scheduled on a flight. In addition to the usual security screening, 10% of the passengers will be subjected to a more complete search.

a) Describe a sampling strategy to randomly select those to be searched.

b) Here is the first-class passenger list and a set of random digits. Select two passengers to be searched, carefully demonstrating your process.

65436 71127 04879 41516 20451 02227 94769 23593

Bergman	Cox	Fontana	Perl
Bowman	DeLara	Forester	Rabkin
Burkhauser	Delli-Bovi	Frongillo	Roufaiel
Castillo	Dugan	Furnas	Swafford
Clancy	Febo	LePage	Testut

c) Explain how you would use a random number table to select the coach passengers to be searched.

29 **Profiling?** Among the 20 first-class passengers on the flight described in Exercise 28 there were four businessmen from the Middle East. Two of them were the two passengers selected to be searched. They complained of profiling, but the airline claims that the selection was random. What do you think? Support your conclusion with a simulation.