The Super Six

2005 #1: Urban and Rural Calories

1. The goal of a nutritional study was to compare the caloric intake of adolescents living in rural areas of the United States with the caloric intake of adolescents living in urban areas of the United States. A random sample of ninth-grade students from one high school in a rural area was selected. Another random sample of ninth graders from one high school in an urban area was also selected. Each student in each sample kept records of all the food he or she consumed in one day.

   The back-to-back stemplot below displays the number of calories of food consumed per kilogram of body weight for each student on that day.

Urban		Rural
9 9 9 9 8 8 7 6	2	
4 4 3 1 0	3	2 3 3 4
9 7 6 6 5	3	5 6 6 6 7
2 0	4	0 2 2 2 4
	4	5 6 8 8 9
	5	1

Stem: tens
Leaf: ones

(a) Write a few sentences comparing the distribution of the daily caloric intake of ninth-grade students in the rural high school with the distribution of the daily caloric intake of ninth-grade students in the urban high school.

(b) Is it reasonable to generalize the findings of this study to all rural and urban ninth-grade students in the United States? Explain.

(c) Researchers who want to conduct a similar study are debating which of the following two plans to use.

Plan I: Have each student in the study record all the food he or she consumed in one day. Then researchers would compute the number of calories of food consumed per kilogram of body weight for each student for that day.

Plan II: Have each student in the study record all the food he or she consumed over the same 7-day period. Then researchers would compute the average daily number of calories of food consumed per kilogram of body weight for each student during that 7-day period.

Assuming that the students keep accurate records, which plan, I or II, would better meet the goal of the study? Justify your answer.

d) Construct parallel boxplots to compare the calorie distribution of the rural vs. the urban students.
e) Verify whether or not there are outliers in either data set.
f) Describe how a researcher might use schools as clusters to gather data in a given county.

g) One researcher observed that rural students ate more home cooked meals than urban students. A journalist wrote an article stating that home cooked meals caused an increase in calorie intake. Describe a confounding variable that may be the cause of the higher calorie intake in rural students.
h) A researcher notes that rural students have less access to organic/vegan supermarkets than urban students and that this could explain the increase in calories in rural areas. However, another researcher noted that rural students eat more meals at home (that are meat and carb heavy, thus high in calories) than urban students. Explain how these variables (lack organic stores and home-cooked meals) are confounded.
i) Describe how you would use your calculator and a list of 9^{th} grade students from your school to conduct a simple random sample. Include a description of how you would implement your procedure.
j) Describe one variable that might be important to create strata and why you chose that variable.
k) What inference procedure would you use to compare the two groups?

2005 #2 Telephone Lines

2. Let the random variable X represent the number of telephone lines in use by the technical support center of a software manufacturer at noon each day. The probability distribution of X is shown in the table below.

x	0	1	2	3	4	5
$p(x)$	0.35	0.20	0.15	0.15	0.10	0.05

(a) Calculate the expected value (the mean) of X.

(b) Using past records, the staff at the technical support center randomly selected 20 days and found that an average of 1.25 telephone lines were in use at noon on those days. The staff proposes to select another random sample of 1,000 days and compute the average number of telephone lines that were in use at noon on those days. How do you expect the average from this new sample to compare to that of the first sample? Justify your response.

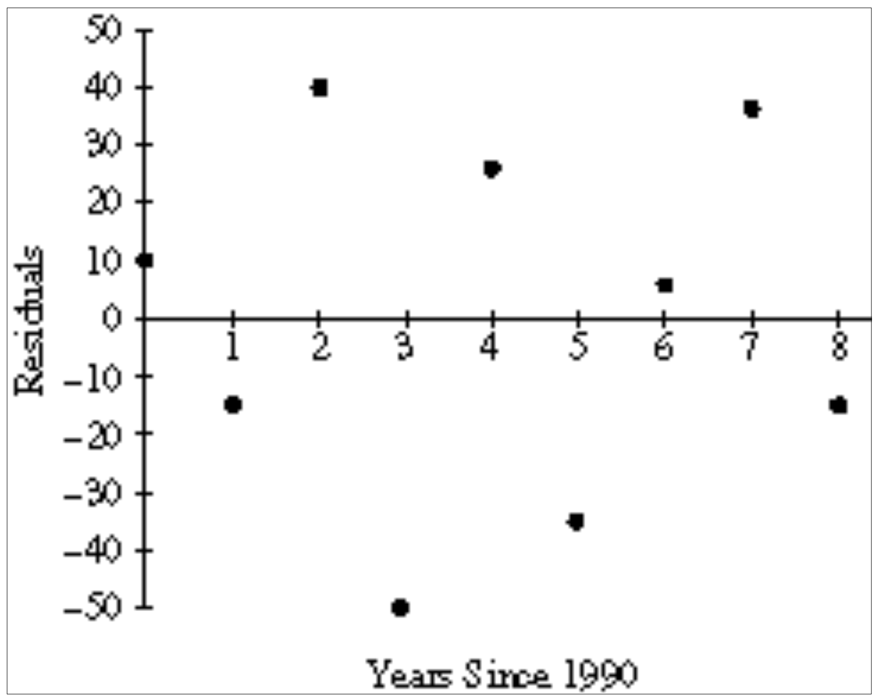
(c) The median of a random variable is defined as any value x such that $P(X \le x) \ge 0.5$ and $P(X \ge x) \ge 0.5$. For the probability distribution shown in the table above, determine the median of X.

(d) In a sentence or two, comment on the relationship between the mean and the median relative to the shape of this distribution.

e) What is the probability that 3 or more lines are in use at noon?
f) What is the probability that at least 1 line is in use at noon?
g) Given that 3 or more lines are in use at noon, what is the probability that all 5 are in use?
h) Assuming that each day is independent of the next, what is the probability that on exactly 2 out of the next 5 days there are no lines in use at noon?
i) Suppose you come by every day at noon to see how many lines are in use. What are the chances that you don't find all 5 in use until your 7th visit?
j) Find the standard deviation of the number of lines in use this support center expects to have at noon.
k) Each call lasts an average of 3.75 minutes. What is the mean and standard deviation of the number of minutes at noon?
l) What are the mean and the standard deviation of the total number of lines in use this support center expects to have at noon over a 7-day week?
m) If another support center has a mean of 2.1 calls and a standard deviation of 1.8 calls, what is the mean and standard deviation of the total of number of calls of both centers at noon?

1999 #1 Aircraft in the 90's

1. Lydia and Bob were searching the Internet to find information on air travel in the United States. They found data on the number of commercial aircraft flying in the United States during the years 1990-1998. The dates were recorded as years since 1990. Thus, the year 1990 was recorded as year 0. They fit a least squares regression line to the data. The graph of the residuals and part of the computer output for their regression are given below.



Predictor	Coef	Stdev	t-ratio	p
Constant	2939.93	20.55	143.09	0.000
Years	233.517	4.316	54.11	0.000
s = 33.43				

a. Is a line an appropriate model to use for these data? What information tells you this?

b. What is the value of the slope of the least squares regression line?
Interpret the slope in the context of this situation.

c. What is the value of the intercept of the least squares regression line?
Interpret the intercept in the context of this situation.

d. What is the predicted number of commercial aircraft flying in 1992 ?

e. What was the actual number of commercial aircraft flying in 1992 ?

f) Interpret *s* in the context of this problem.
g) Create and interpret a 95% confidence interval for the slope.
h) R^2= 87.4%. Interpret this value in context.
i) Find and interpret the correlation coefficient.
j) If each new aircraft costs the FAA an additional $1000 in regulatory costs, how much are the costs increasing each year, on average?
k) Is there statistically convincing evidence that the number of aircraft is related to year? Explain.

2003B #2 Income & Age

2. A simple random sample of adults living in a suburb of a large city was selected. The age and annual income of each adult in the sample were recorded. The resulting data are summarized in the table below.

	Annual Income			
Age Category	\$25,000-\$35,000	\$35,001-\$50,000	Over \$50,000	Total
21-30	8	15	27	50
31-45	22	32	35	89
46-60	12	14	27	53
Over 60	5	3	7	15
Total	47	64	96	207

(a) What is the probability that a person chosen at random from those in this sample will be in the 31-45 age category?

(b) What is the probability that a person chosen at random from those in this sample whose incomes are over \$50,000 will be in the 31-45 age category? Show your work.

(c) Based on your answers to parts (a) and (b), is annual income independent of age category for those in this sample? Explain.

d) Make a graphical display to examine the relationship between Age and Income. Describe this graph.

e) Name an inference procedure that could be carried out to answer the independence question on part (c).

f) If Age and Income were completely independent, find the number of 46-60 year olds you would expect to have an income of over \$50,000.

The Super Six

2004B #3 Bauxite Ore Cars

3. Trains carry bauxite ore from a mine in Canada to an aluminum processing plant in northern New York state in hopper cars. Filling equipment is used to load ore into the hopper cars. When functioning properly, the actual weights of ore loaded into each car by the filling equipment at the mine are approximately normally distributed with a mean of 70 tons and a standard deviation of 0.9 ton. If the mean is greater than 70 tons, the loading mechanism is overfilling.

   (a) If the filling equipment is functioning properly, what is the probability that the weight of the ore in a randomly selected car will be 70.7 tons or more? Show your work.

   (b) Suppose that the weight of ore in a randomly selected car is 70.7 tons. Would that fact make you suspect that the loading mechanism is overfilling the cars? Justify your answer.

   (c) If the filling equipment is functioning properly, what is the probability that a random sample of 10 cars will have a mean ore weight of 70.7 tons or more? Show your work.

   (d) Based on your answer in part (c), if a random sample of 10 cars had a mean ore weight of 70.7 tons, would you suspect that the loading mechanism was overfilling the cars? Justify your answer.

e) Draw a careful sketch to show your answer to part (a)
f) Given the initial mean and standard deviation, how full are the most heavy 10% of the cars?
g) Describe the sampling distribution of the sample mean, if a sample of size 10 were chosen.
h) If we took a random sample of 40 cars instead of 10, how would that change your answer to part (g)?

2000 #5 Cholesterol and Exercise

5. High cholesterol level in people can be reduced by exercise or by drug treatment. A pharmaceutical company has developed a new cholesterol-reducing drug. Researchers would like to compare its effects to the effects of the cholesterol-reducing drug that is currently available on the market. Volunteers who have a history of high cholesterol and who are currently not on medication will be recruited to participate in a study.

(a) At this time, the company only has resources to test the effects of the drug treatment. Explain how you would carry out a completely randomized experiment for the study.

(b) Describe an experimental design that would improve the design in (a) by incorporating blocking.

(c) Can the experimental design in (b) be carried out in a double blind manner? Explain.

d) Describe a method for implementing your design in part (c).

e) What inference procedure would you use to compare the results obtained by method (a)?

f) After the method in part (a) was carried out, researchers found a difference with a p-value of 0.003 (in favor of the new medication over the old drug). Does this mean that the researchers can conclude that drug caused a reduction in cholesterol?

g) Identify the subjects, the treatment(s), the factor(s), the level(s), and the response variable in this experiment.

h) After an increasing in funding, the company is able to run an experiment where they test **both** the drug treatment and the effects of exercise. They decide to ask volunteers to exercise at a high level, a low level, or not all. Describe the factors, their levels, and the treatments of this expanded experiment.