Collecting Data

1. Observations vs. Experiments
2. Simulations
3. Surveys
   - Census
   - Random
   - Cluster
   - Stratified
   - Convenience
   - Systematic
   - Multi-stage
4. Experiments
   - Components
   - Randomized
   - Blocked
   - Matched Pairs
5. Bias vs. Sampling Error/Sampling Variability
6. Bias
   - Voluntary Response Bias
   - Undercoverage Bias
   - Non-response Bias
   - Response Bias
   - Selection Bias

Displaying Data

1. Qualitative/Categorical vs. Quantitative
2. Qualitative
   - Dot Plots, Bar Charts, Segmented Bar Charts, Pie Charts-relative Frequencies
   - Mode
3. Quantitative
   - Histograms, Stem Plots, Dot Plots, box plots, cumulative frequency plots, pie charts
   - CUSS-Center Shape Spread Unusual Features
   - Center- Mean, median mode,
   - Shape-Skewed, uniform, symmetric, bi-modal, mound
   - Spread-range, IQR, variance/standard deviation
   - Unusual Features-outliers, gaps
4. Effects of changing units
   - Adding a Constant
   - Multiplying by a Constant
Collecting Data—Observations and Experiments

1. Two studies are run to compare the experiences of families living in high-rise public housing to those of families living in townhouse subsidized rentals. The first study interviews 25 families who have been in each government program for at least 1 year, while the second randomly assigns 25 families to each program and interviews them after 1 year. Which of the following is a true statement?
   A. Both studies are observational studies because of the time period involved.
   B. Both Studies are observational studies because there are no control groups
   C. The first study is an observational study, while the second is an experiment
   D. The first study is an experiment, while the second is an observational study
   E. Both studies are experiments.

2. When the estrogen-blocking drug tamoxifen was first introduced to treat breast cancer, there was concern that it would cause osteoporosis as a side effect. To test this concern, cancer subjects were randomly selected and given tamoxifen, and their bone density was measured before and after treatment. Which of the following is a true statement?
   A. This was an observational study.
   B. This study was a sample survey of randomly selected cancer patients.
   C. This study was an experiment in which the subjects were used as their own controls.
   D. With the given procedure, there cannot be a placebo effect.
   E. Causation cannot be concluded without knowing the survival rates.

3. Which of the following are true statements?
   I. Based on careful use of control groups, experiments can often indicate cause and effect relationships.
   II. While observational studies may suggest relationships, great care must be taken in concluding that there is cause and effect because of the lack of control over lurking variables.
   III. A complete census is the only way to establish a cause-and-effect relationship absolutely.
   A. I and II
   B. I and III
   C. II and III
   D. I, II, and III
   E. None of the above gives the complete set of true responses.

4. In one study subjects were randomly given either 500 or 1000 milligrams of vitamin C daily, and the number of colds they came down with during a winter season was noted. In a second study people responded to a questionnaire asking about the average number of hours they sleep per night and the number of colds they came down with during a winter season.
   A. The first study was an experiment without a control group, while the second was an observational study.
   B. The first study was an observational study, while the second was a controlled experiment.
   C. Both studies were controlled experiments.
   D. Both studies were observational studies.
   E. None of the above is a correct statement.

5. Which of the following are true statements?
   I. IN an experiment some treatment is intentionally forced on one group to not the response
   II. In an observational study information is gathered on an already existing situation.
   III. Sample surveys are observational studies, not experiments
   A. I and II
   B. I and III
   C. II and III
   D. I, II, and III
   E. None of the above gives the complete set of true responses.
Collecting Data—Observations and Experiments 1999 Question 3

The dentists in a dental clinic would like to determine if there is a difference between the number of new cavities in people who eat an apple a day and in people who eat less than one apple a week. They are going to conduct a study with 50 people in each group.

Fifty clinic patients who report that they routinely eat an apple a day and 50 clinic patients who report that they eat less than one apple a week will be identified. The dentists will examine the patients and their records to determine the number of new cavities the patients have had over the past two years. They will then compare the number of new cavities in the two groups.

(a) Why is this an observational study and not an experiment?

This is an observational study because:

- No treatments were imposed
- The data used was existing data
- There was no random assignment of subjects to treatment groups

(b) Explain the concept of confounding in the context of this study. Include an example of a possible confounding variable.

Confounding occurs when the effects of the variables cannot be distinguished from one another. In this case confounding would occur if it became impossible to measure the true effect on the number of cavities attributed to a group who eats more than one apple a week as compared to the number of cavities for a group that ate less than apple a week as the result of a related variable that is related to dental health. For instance, it is reasonable to assume that the group which consumes more than one apple a week has on an average a healthier overall lifestyle and that could be the cause for that groups having fewer cavities when compared to the group that eats less than one apple a week.

(c) If the mean number of new cavities for those who ate an apple a day was statistically significantly smaller than the mean number of new cavities for those who ate less than one apple a week, could one conclude that the lower number of new cavities can be attributed to eating an apple a day? Explain.

(d) Because this is not an experiment, it is not reasonable to draw a cause and effect relationship between eating apples and a lower number of new cavities.
Every Monday a local radio station gives coupons away to 50 people who correctly answer a question about a news fact from the previous day’s newspaper. The coupons given away are numbered from 1 to 50, with the first person receiving coupon 1, the second person receiving coupon 2, and so on, until all 50 coupons are given away. On the following Saturday, the radio station randomly draws numbers from 1 to 50 and awards cash prizes to the holders of the coupons with these numbers. Numbers continue to be drawn without replacement until the total amount awarded first equals or exceeds $300. If selected, coupons 1 through 5 each have a cash value of $200, coupons 6 through 20 each have a cash value of $100, and coupons 21 through 50 each have a cash value of $50.

(a) Explain how you would conduct a simulation using the random number table provided below to estimate the distribution of the number of prizewinners each week.

(b) Perform your simulation 3 times. (That is run 3 trials of your simulation.) Start at the leftmost digit in the first row of the table and move across. Make your procedure clear so that someone can follow what you did. You must do this by marking directly on or above the table. Report the number of winners in each of your 3 trials.

| 72749 | 13347 | 65030 | 26128 | 49067 | 02904 | 49953 | 74674 | 94617 | 73317 |
| 81638 | 36566 | 42709 | 33717 | 59943 | 12027 | 46547 | 61303 | 46690 | 76423 |
| 38449 | 46438 | 91579 | 01907 | 72146 | 05764 | 22400 | 94490 | 49890 | 09258 |

- Trial 1: $13\text{347} \quad \text{2 winners} \quad \text{\$300 prizes}$
- Trial 2: $4\text{2709} \quad \text{5 winners} \quad \text{\$350 prizes}$
- Trial 1: $4\text{2709} \quad \text{4 winners} \quad \text{\$300 prizes}$

(b) Trial 1 -- 3 winners $\text{\$300 prizes}$
Trial 2 -- 5 winners $\text{\$350 prizes}$
Trial 1 -- 4 winners $\text{\$300 prizes}$
Collecting Data—Surveys

6. Which of the following is a true statement about sampling error?
   A. Sampling error can be eliminated only if a survey is both extremely well designed and extremely well conducted.
   B. Sampling error concerns natural variation between samples, is always present, and can be described using probability.
   C. Sampling error is generally larger when the sample size is larger.
   D. Sampling error implies an error, possibly very small, but still an error on the part of the surveyor.
   E. Sampling error is highest when bias is present.

7. Which of the following is a true statement?
   A. If bias is present in a sampling procedure, it can be overcome by dramatically increasing the sample size.
   B. There is no such thing as a “bad sample.”
   C. Sampling techniques that use probability techniques effectively eliminate bias.
   D. Convenience samples often lead to undercoverage bias.
   E. Voluntary response samples often underrepresent people with strong opinions.

8. To conduct a survey of long-distance calling patterns, a researcher opens a telephone book to a random page, closes his eyes, puts his finger down on the page, and then reads off the next 50 names. Which of the following is incorrect?
   A. The survey design incorporates chance?
   B. Assuming the page and starting point on the page are randomly selected, each person in the phone book has an equal chance of being selected.
   C. The procedure could easily result in selection bias
   D. The procedure does not result in a simple random sample
   E. This is the typical methodology of a systematic sample.

9. A survey was administered to parents of high school students in a certain state to see if the parents thought the students’ academic needs were being met. To select the sample, the parents were divided into two groups—one group of parents who live in cities with populations of more than 100,000 and the other group of parents who live in cities with populations less than or equal to 100,000. A random sample of 100 parents from each group was taken. Which of the following statements about the sample of 200 parents is true?
   A. It is a convenience sample because the sample of parents was easily obtained.
   B. It is a stratified random sample because parents were randomly selected from each group.
   C. It is a random cluster sample because parents were randomly selected from each group.
   D. It is a random cluster sample because groups of high schools were randomly selected.
   E. It is a systematic sample because the parents were systematically divided into two groups.

10. The buyer for an electronics store wants to estimate the proportion of defective wireless game controllers in a shipment of 5,000 controllers from the stores’ primary supplier. The shipment consists of 200 boxes each containing 25 controllers. The buyer numbers the boxes 1 to 200 and randomly selects six numbers in that range. She opens the six boxes with the corresponding numbers, examines all 25 controllers in each of these boxes, and determines the proportion of the 150 controllers that are defective. What type of sample is this?
   (A) Biased random sample
   (B) Nonrandom sample
   (C) Simple random sample
   (D) Stratified random sample
   (E) Cluster random sample
Collecting Data—Surveys—2011 Question 3

An apartment building has nine floors and each floor has four apartments. The building owner wants to install new carpeting in eight apartments to see how well it wears before she decides whether to replace the carpet in the entire building.

The figure below shows the floors of apartments in the building with their apartment numbers. Only the nine apartments indicated with an asterisk (*) have children in the apartment.

(a) For convenience, the apartment building owner wants to use a cluster sampling method, in which the floors are clusters, to select the eight apartments. Describe a process for randomly selecting eight different apartments using this method.

Using the floor numbers assign a number to each floor. 1 for 1st, 2 for 2nd all the way for 9 for 9th. Place the numbers in a bin and mix. Draw 2 numbers without replacement. The two numbers drawn represent the floors that are to be checked. Because this is a cluster sample all 4 apartments on both of the selected floors are to be analyzed for carpet wear.

(b) An alternative sampling method would be to select a stratified random sample of eight apartments, where the strata are apartments with children and apartments with no children. A stratified random sample of size eight might include two randomly selected apartments with children and six randomly selected apartments with no children. In the context of this situation, give one statistical advantage of selecting such a stratified sample as opposed to a cluster sample of eight apartments using the floors as clusters.

Because the amount of wear on the carpets in apartments with children could be different from the wear on the carpets in apartments without children, it would be advantageous to have apartments with children represented in the sample. The cluster sampling procedure in part (a) could produce a sample with no children in the selected apartments; for example, a cluster sample of the apartments on the third and sixth floors would consist entirely of apartments with no children. Stratified random sampling, where the two strata are apartments with children and apartments without children, guarantees a sample that includes apartments with and without children, which, in turn, would yield sample data that are representative of both types of apartments.
In response to nutrition concerns raised last year about food served in school cafeterias, the Smallville School District entered into a one-year contract with the Healthy Alternative Meals (HAM) company. Under this contract, the company plans and prepares meals for 2,500 elementary, middle, and high school students, with a focus on good nutrition. The school administration would like to survey the students in the district to estimate the proportion of students who are satisfied with the food under this contract.

Two sampling plans for selecting the students to be surveyed are under consideration by the administration. One plan is to take a simple random sample of students in the district and then survey those students. The other plan is to take a stratified random sample of students in the district and then survey those students.

a) Describe a simple random procedure that the administration could use to select 200 students from the 2,500 students in the district.

Assign each student a number 1-2500. Use a random number generator to randomly select 200 numbers without repeats. Survey the students who were assigned the numbers selected by the random number generator.

b) If a stratified random sampling procedure is used, give one example of an effective variable on which to stratify in this survey. Explain your reasoning.

Because nutritional requirements are likely to vary by age, I would stratify based on school type as that would control for differences in age. My strata would be elementary, middle and high school.

c) Describe one statistical advantage of using a stratified random sample over a simple random sample in the context of this study.

By stratifying by school type I am certain to have representation of the three different age groups. As a result I can control for the variation in nutritional needs that arises from the difference in age groups. I cannot be assured of representation of all 3 age groups if I use a simple random sample.
Collecting Data—Experiments 2003B Question 4

There have been many studies recently concerning coffee drinking and cholesterol level. While it is known that several coffee-bean components can elevate blood cholesterol level, it is thought that a new type of paper coffee filter may reduce the presence of some of these components in coffee.

The effect of the new filter on cholesterol level will be studies over a 10-week period using 300 nonsmokers who each drink 4 cups of caffeinated coffee per day. Each of these 300 participants will be assigned to one of two groups: the experimental group, who will only drink coffee that has been made with the new filter, or the control group, who will only drink coffee that has been made with the standard filter. Each participant’s cholesterol level will be measured at the beginning and at the end of the study.

(a) Describe an appropriate method for assigning the subjects to the two groups so that each group will have an equal number of subjects.

I would place all 300 names in a bin and mix well. I would then draw 150 names without replacement. The names that were drawn would be assigned the new coffee filter. The remaining names would be assigned the standard filter.

(b) In this study, the researchers chose to include a group who only drank coffee that was made with the standard filter. Why is it important to include a control group in this study even though cholesterol levels will be measured at the beginning and at the end of the study?

It is important to include a control group so that we can be certain that any changes in cholesterol level is due to the filter rather than dietary changes that might occur throughout the year.

(c) Which test would you conduct to determine whether the change in cholesterol level would be greater if people used the new filter rather than using the standard filter?

(d) Why would the researchers choose to use only nonsmokers in the study?

Because smoking might affect cholesterol levels and confound the results making it difficult to differentiate whether or not the filter or smoking habits were causing changes in cholesterol level, researchers might choose to only select non-smokers.
Collecting Data-Experiments Random Assignment & Matched Pairs 2005B Question 3

In search of a mosquito repellent that is safer than the ones that are currently on the market, scientists have developed a new compound that is rated as less toxic than the current compound, thus making a repellent that contains this new compound safer for human use. Scientists also believe that a repellent containing the new compound will be more effective than the ones that contain the current compound. To test the effectiveness of the new compound versus that of the current compound, scientists have randomly selected 100 people from a state.

Up to 100 bins, with an equal number of mosquitoes in each bin, are available for use in the study. After a compound is applied to a participant’s forearm, the participant will insert his or her forearm into a bin for 1 minute, and the number of mosquito bites on the arm at the end of that time will be determined.

(a) Suppose this study is to be conducted using a completely randomized design. Describe a randomization process and identify an inference procedure for the study.

Place all 100 names of the participants in a hat and mix. The first group of 50 names will be assigned to the current compound. The remaining group of 50 names will have the new compound applied to their arm. The bins of mosquitoes will be numbered 1-100. Each participant will draw a number from a hat containing the numbers 1-100. The participant will then place their arm in that particular bin for 1 minute. The mean number of bites per person for each group will then be compared.

(b) Suppose this study is to be conducted using a matched-pairs design. Describe a randomization process and identify an inference procedure for the study.

Each participant will be randomly assigned to a bin as above. First 50 names drawn from a hat will have the new repellant applied to their right arm and the current repellant applied to their left. The remaining 50 subjects will have the new repellant applied to their left arm and the current repellant applied to their right. Those who are assigned to an odd number will place their right arm in the bin first while those assigned to an even number bin will place their left arm in first. After one minute with the first arm, the second arm is placed in the bin. Then the different in bites in each arm will be noted.

(c) Which of the designs, the one in part (a) or the one in part (b), is better for testing the effectiveness of the new compound versus that of the current compound? Justify your answer.

The design in part b is better because it helps control potential sources of variation from person to person such as an individuals’ susceptibility to mosquito bites.
Collecting Data—Experiment 2007 Form B Question 3

The United States Department of Energy is conducting an experiment to compare the heat gain in houses using two different types of windows, A and B. Six windows of each type are available for the experiment. The Department has constructed a house with twelve windows as shown on the floor plan below.

In the interior of the house, each window is surrounded by a window box to capture and measure the amount of heat coming in through that window and to isolate the heat gain for each window.

(a) A randomized block experiment will be used to compare the heat gain for the two types (A and B) of windows. How would you group the window boxes into blocks? (Clearly indicate your blocks using the window box numbers.) Justify your choice of blocks.

The blocks are (1, 12); (2, 3); (4, 5); (6, 7); (8, 9); (10, 11).

Blocks were designed to be homogenous units so that each window type would be tested under the same conditions. The conditions of interest are exposure to the sun and proximity to openings. All blocks are on the same side of the building and should receive a similar amount of exposure to the sun.

(b) For the design in part (a), describe how you would assign windows types (A and B) to the numbered window boxes.

Window types A and B need to be assigned randomly, to do so I will flip a coin. If the coin is heads, the smaller numbered window box in the block will be assigned window type A and the larger numbered window box will be assigned window type B. If the coin is tails, the smaller numbered window box in the block will be assigned window type B and the larger numbered window Box will be assigned window type A.
Collecting Data—Bias

11. Anne Landers, who wrote a daily advice column appearing in newspapers across the country, once asked her readers, “If you had it to do over again, would you have children?” Of the more than 10,000 readers who responded, 70% said no. What does this show?

A. The survey is meaningless because of voluntary response bias.
B. No meaningful conclusion is possible without knowing something more about the characteristics of her readers.
C. The survey would have been more meaningful if she had picked a random sample of the 10,000 readers who responded.
D. The survey would have been less meaningful if she had used a control group.
E. This is a legitimate sample, randomly drawn from her readers and of sufficient size to allow the conclusion that most of her readers who are parents would have second thoughts about having children.

12. Which of the following is a true statement?

A. If bias is present in a sampling procedure, it can be overcome by dramatically increasing the sample size.
B. There is no such thing as a “bad sample.”
C. Sampling techniques that use probability techniques effectively eliminate bias.
D. Convenience samples often lead to undercoverage bias.
E. Voluntary response samples often underrepresent people with strong opinions.

13. Two possible wordings for a questionnaire on gun control are as follows:

I. The United States has the highest rate of murder by handguns among all countries. Most of these murders are known to be crimes of passion or crimes provoked by anger between acquaintances. Are you in favor of a 7-day cooling-off period between the filing of an application to purchase a handgun and the resulting sale?

II. The United States has one of the highest violent crime rates among all countries. Many people want to keep handguns in their homes for self-protection. Fortunately, U.S. citizens are guaranteed the right to bear arms by the Constitution. Are you in favor of a 7-day waiting period between the filing of an application to purchase a needed handgun and the resulting sale?

One of these questions showed that 25% of the population favored a 7-day waiting period between the application for purchase of a handgun and the resulting sale, while the other question showed that 70% of the population favored the waiting period. Which produced which result and why?

A. The first question probably showed 70% and the second question 25% because of the lack of randomization in the choice of pro-gun and anti-gun subjects as evidenced by the wording of the questions.
B. The first question probably showed 25% and the second question 70% because of a placebo effect due to the wording of the questions.
C. The first question probably showed 70% and the second question 25% because of the lack of a control group.
D. The first questions probably showed 25% and the second question 70% because of the response bias due to wording of the questions.
E. The first question probably showed 70% and the second question 25% because of response bias due to the wording of the questions.
Collecting Data-Bias 2005 Question 5
A survey will be conducted to examine the educational level of adult heads of households in the United States. Each respondent in the survey will be placed into one of the following two categories:
- Does not have a high school diploma
- Has a high school diploma

The survey will be conducted using a telephone interview. Random-digit dialing will be used to select the sample.

(a) For this survey, state one potential source of bias and describe how it might affect the estimate of the proportion of adult heads of households in the United States who do not have a high school diploma.
Response bias—people without a high school diploma may be reluctant to admit to their not having a diploma. As a result, the survey will likely underestimate the proportion of heads of households without a high school diploma and overestimate the proportion with a high school diploma.

(b) A pilot survey indicated that about 22 percent of the population of adult heads of households do not have a high school diploma. Using this information, how many respondents should be obtained if the goal of the survey is to estimate the proportion of the population who do not have a high school diploma to within 0.03 with 95 percent confidence? Justify your answer.

\[ \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \]
\[ \text{Margin of error} = .03 \quad \hat{p} = .22 \quad M.E = z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = .03 = 1.96 \sqrt{\frac{(0.22)(0.78)}{n}} \]
\[ n = 732.47 \]

The sample size necessary to estimate the proportion of adult heads of households without a high school diploma with the above conditions is 733.

(c) Since education is largely the responsibility of each state, the agency wants to be sure that estimates are available for each state as well as for the nation. Identify a sampling method that will achieve this additional goal and briefly describe a way to select the survey sample using this method.
I would stratify my sample using the individual states as the strata. From each state I would take a random sample of the heads of households. To perform the random sample within each state, I would obtain a list of all heads of households and assign each a number. I would use a random number generator to select the numbers and then contact those heads of households whose name corresponded to the numbers from the random number generator.
Collecting Data—Bias 2008 Question 2

A local school board plans to conduct a survey of parents’ opinions about year-round schooling in elementary schools. The school board obtains a list of all families in the district with at least one child in an elementary school and sends the survey to a random sample of 500 of the families. The survey question is provided below.

A proposal has been submitted that would require students in elementary schools to attend school on a year-round basis. Do you support this proposal? (Yes or No)

The school board received responses from 98 of the families, with 76 of the response indicating support for year-round schools. Based on this outcome, the local school board concludes that most of the families with at least one child in elementary school prefer year-round schooling.

(a) What is the possible consequence of nonresponse bias for interpreting the results of this survey?
Only 98 out of 500 families or 19.6% responded to the survey. It is likely that the opinions of parents who are most in favor of year-round schools were the ones who responded and there are over-represented in the survey.

(b) Someone advised the local school board to take an additional random sample of 500 families and to use the combined results to make their decision. Would this be a suitable solution to the issue raised in part (a)? Explain.
NO, this would not be reasonable. Even after increasing the sample size the non-response bias would still be present because there are still 402 of 500 families that did not respond from the first survey.

(c) Suggest a different follow-up step from the one suggested in part (b) that the local school board could take to address the issue raised in part (a).
The school board could contact the 402 families who did not respond and try to get their responses.
Or the school board could take a completely new survey using phone calls or person to person contact to try to eliminate the amount of non-response bias.
Probability

1. Set Theory
   - Complement
   - General Addition Rule
   - Mutually Exclusive
     - Addition Principle
   - Conditional Probability
     - 2-way Table
     - Tree Diagram
   - Independent
     - Multiplication Principle
   - Dependent
2. Law of Large Numbers
3. Expected Values & Standard Deviation
4. Combining Random Variables
5. Geometric
6. Binomial
7. Normal
Probability – Complement

1. The yearly mortality rate for American men from prostate cancer has been constant for decades at about 25 of every 100,000 men. (This rate has not changed in spite of new diagnostic techniques and new treatments.) In a group of 100 American men, what is the probability that at least 1 will die from prostate cancer in a given year.
   (A) .00025
   (B) .0247
   (C) .025
   (D) .9753
   (E) .99975

2. Let X represent a random variable whose distribution is normal, with a mean of 100 and a standard deviation of 10. Which of the following is equivalent to \( P(X > 115) \)?
   (A) \( P(X < 115) \)
   (B) \( P(X \leq 115) \)
   (C) \( P(X < 85) \)
   (D) \( P(85 < X < 115) \)
   (E) \( 1 - P(X < 85) \)

Probability – Independence & Mutually Exclusive

3. Which of the following statements is true for two events, each with probability greater than 0?
   (A) If the events are mutually exclusive, they must be independent.
   (B) If the events are independent, they must be mutually exclusive.
   (C) If the events are not mutually exclusive, they must be independent.
   (D) If the events are not independent, they must be mutually exclusive.
   (E) If the events are mutually exclusive, they cannot be independent.

4. Following are parts of the probability distributions for the random variables X and Y.

<table>
<thead>
<tr>
<th>x</th>
<th>P(x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>.2</td>
</tr>
<tr>
<td>3</td>
<td>.3</td>
</tr>
<tr>
<td>4</td>
<td>?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>y</th>
<th>P(y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.4</td>
</tr>
<tr>
<td>2</td>
<td>?</td>
</tr>
<tr>
<td>3</td>
<td>.1</td>
</tr>
</tbody>
</table>

If X and Y are independent and their joint probability \( P(X=1, Y=2) = .1 \) what is \( P(X=4) \)?
   (A) .1
   (B) .2
   (C) .3
   (D) .4
   (E) .5

5. In the November 27, 1994, issue of Parade magazine, the “Ask Marilyn” section contained this question: “Suppose a person was having two surgeries performed at the same time. If the chances of success for surgery B are 90%, what are the chances that both would fail?”
   What do you think of Marilyn’s solution: (.15)(.10) = .015 or 1.5%
   (A) Her solution is mathematically correct but not explained very well.
   (B) Her solution is both mathematically correct and intuitively obvious.
   (C) Her use of complementary events is incorrect.
   (D) Her use of the general addition formula is incorrect.
   (E) She assumed independence of events, which is most likely wrong.
6. Given that P(E) = .32, P(F) = .15 and P(E \cap F) = .048, which of the following is a correct conclusion?
   (A) The events E and F are both independent and mutually exclusive.
   (B) The events E and F are neither independent nor mutually exclusive.
   (C) The events E and F are mutually exclusive but not independent.
   (D) The events E and F are independent but not mutually exclusive.
   (E) The events E and F are independent, but there is insufficient information to determine whether or not they are mutually exclusive.

7. Suppose that, for any given year, the probabilities that the stock market declines, that women's hemlines are lower, and that both events occur are, respectively, .4, .35, and .3. Are the two events independent?
   (A) Yes, because (.4)(.35) ≠ .3
   (B) No, because (.4)(.35) ≠ .3
   (C) Yes, because .4 > .35 > .3
   (D) Yes, because (.5)(.3 + .4) ≠ .35
   (E) There is insufficient information to answer this question

8. A city water supply system involves three pumps, the failure of any one of which crashes the system. The probabilities of failure for each pump in a given year are .025, .034, and .02, respectively. Assuming the pumps operate independently of each other, what is the probability that the system does crash during the year?
   (A) Less than .05
   (B) .077
   (C) .079
   (D) .081
   (E) .923

9. Given that P(E) = .32, P(F) = .15 and P(E \cap F) = .048, which of the following is a correct conclusion?
   (A) The events E and F are both independent and mutually exclusive.
   (B) The events E and F are neither independent nor mutually exclusive.
   (C) The events E and F are mutually exclusive but not independent.
   (D) The events E and F are independent but not mutually exclusive.
   (E) The events E and F are independent, but there is insufficient information to determine whether or not they are mutually exclusive.

10. Given the probabilities P(A) = .4 and P(A \cup B) = .6 what is the probability P(B) if A and P are mutually exclusive? If A and B are independent?
    (A) .2, .4
    (B) .2, .33
    (C) .33, .2
    (D) .6, .33
    (E) .6, .4

11. An experiment has three mutually exclusive outcomes, A, B, and C. If P(A) = 0.12, P(B) = 0.61, and P(C) = 0.27, which of the following must be true?
    I. A and C are independent.
    II. P(A and B) = 0
    III. P(B or C) = P(B) + P(C)
    (A) I only
    (B) I and II only
    (C) I and III only
    (D) II and III only
    (E) I, II, and III
Probability—Contingency
Questions 12-13 refer to the following study: One thousand students at a city high school were classified both according to GPA and whether or not they consistently skipped classes.

<table>
<thead>
<tr>
<th>GPA</th>
<th>Many Skipped Classes</th>
<th>Few Skipped Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;2.0</td>
<td>80</td>
<td>175</td>
</tr>
<tr>
<td>2.0-3.0</td>
<td>25</td>
<td>450</td>
</tr>
<tr>
<td>&gt;3.0</td>
<td>5</td>
<td>265</td>
</tr>
</tbody>
</table>

12. What is the probability that a student has a GPA between 2.0 and 3.0?
   (A) 0.25
   (B) 0.227
   (C) 0.450
   (D) 0.475
   (E) 0.506

13. What is the probability that a student has a GPA under 2.0 and has skipped many classes?
   (A) 0.080
   (B) 0.281
   (C) 0.285
   (D) 0.314
   (E) 0.727

14. What is the probability that a student has a GPA under 2.0 or has skipped many classes?
   (A) 0.80
   (B) 0.281
   (C) 0.285
   (D) 0.314
   (E) 0.727

15. What is the probability that a student has a GPA under 2.0 given that he has skipped many classes?
   (A) 0.080
   (B) 0.281
   (C) 0.285
   (D) 0.314
   (E) 0.727

16. Are “GPA between 2.0 and 3.0” and “skipped few classes” independent?
   (A) No, because .475 ≠ .506.
   (B) No, because .475 ≠ .890.
   (C) No, because .450 ≠ .475.
   (D) Yes, because of conditional probabilities.
   (E) Yes, because of the product rule.

17. A computer technician notes that 40% of computers fail because of the hard drive, 25% because of the monitor, 20% because of a disk drive, and 15% because of the microprocessor. If the problem is not the monitor, what is the probability that it is in the hard drive?
   (A) 0.150
   (B) 0.400
   (C) 0.417
   (D) 0.533
   (E) 0.650
2003 Form B Question 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.

<table>
<thead>
<tr>
<th>Age Category</th>
<th>Annual Income</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>21-30</td>
<td>$25,000-$35,000</td>
<td>8</td>
</tr>
<tr>
<td>31-45</td>
<td>$35,001-$50,000</td>
<td>15</td>
</tr>
<tr>
<td>45-60</td>
<td>Over $50,000</td>
<td>27</td>
</tr>
<tr>
<td>Over 60</td>
<td></td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>50</td>
</tr>
</tbody>
</table>

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

\[ P(31-45) = \frac{89}{207} = .43 \]

(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.

\[ P(\text{over 50,000} \cap 31-45) \mid P(\text{over 50,000}) = \frac{P(\text{over 50,000} \cap 31-45)}{P(\text{over 50,000})} = \frac{35}{96} = .365 \]

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

Let A be the age 31-45 category
Let B be the income category 50,000 and over

\[ P(A \mid B) = \frac{P(A \cap B)}{P(B)} \]

and if independent, \[ P(A \mid B) = \frac{P(A)}{P(B)} = P(A) \]

\[ \therefore P(A \mid B) = P(A) \]

If independent then \[ \frac{P(31-45 \cap \text{over 50,000})}{P(\text{over 50,000})} = P(31-45) \]

However .365 does not equal .43 therefore the events are not independent. The age category 31-45 and the income category over 51,000 are not independent.
Problem 3 A laboratory test for the detection of a certain disease gives a positive result 5 percent of the time for people who do not have the disease. The test gives a negative result 0.3 percent of the time for people who have the disease. Large-scale studies have shown that the disease occurs in about 2 percent of the population.

(a) What is the probability that a person selected at random would test positive for this disease? Show your work.

<table>
<thead>
<tr>
<th></th>
<th>Positive</th>
<th>Not Positive</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diseased</td>
<td>.01994</td>
<td>.00006</td>
<td>.02</td>
</tr>
<tr>
<td></td>
<td>.02-.0006</td>
<td>.003x.02</td>
<td></td>
</tr>
<tr>
<td>Not diseased</td>
<td>.049</td>
<td>.931</td>
<td>.98</td>
</tr>
<tr>
<td></td>
<td>.98x.05</td>
<td>.98-.0049</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>.06894</td>
<td>.93106</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>.01994+.049</td>
<td>.00006+.931</td>
<td></td>
</tr>
</tbody>
</table>

The probability of testing positive for the disease is .06894

(b) What is the probability that a person selected at random who tests positive for the disease does not have the disease? Show your work

The probability that a person will test positive for the disease given that they do not have the disease is \( \frac{.049}{.06894} = .711 \).

71.1% of the people who test positive for the disease do not have the disease.
Probability—Law of Large Numbers/Expected Values & Standard Deviation

18. There are two games involving flipping a coin. In the first game you win a prize if you can throw between 40% and 60% heads. In the second game you win if you can throw more than 75% heads. For each game would you rather flip the coin 50 times or 500 times?
(A) 50 times for each game
(B) 500 times for each game
(C) 50 times for the first game, and 500 for the second.
(D) 500 times for the first game, and 50 for the second.
(E) The outcomes of the games do not depend on the number of flips.

19. Mathematically speaking, casinos and life insurance companies make a profit because of
(A) Their understanding of sampling error and sources of bias
(B) Their use of well-designed, well conducted surveys and experiments
(C) Their use of simulation of probability distributions
(D) The central limit theorem
(E) The law of large numbers

20. A magazine has 1,620,000 subscribers, of whom 640,000 are women and 980,000 are men. Thirty percent of the women read the advertisements in the magazine and 50 percent of the men read the advertisements in the magazine. A random sample of 100 subscribers is selected. What is the expected number of subscribers in the sample who read the advertisements?
(A) 30
(B) 40
(C) 42
(D) 50
(E) 80

21. A manufacturer makes light bulbs and claims that their reliability is 98 percent. Reliability is defined to be the proportion of non-defective items that are produced over the long term. If the company’s claim is correct, what is the expected number of non-defective light bulbs in a random sample of 1,000 bulbs?
(A) 20
(B) 200
(C) 960
(D) 980
(E) 1,000

22. The number of sweatshirts a vendor sells daily has the following probability distribution.

<table>
<thead>
<tr>
<th>Number of Sweatshirts $x$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(x)$</td>
<td>0.3</td>
<td>0.2</td>
<td>0.3</td>
<td>0.1</td>
<td>0.08</td>
<td>0.02</td>
</tr>
</tbody>
</table>

If each sweatshirt sells for $25, what is the expected daily total dollar amount taken in by the vendor from the sale of sweatshirts?

(A) $5.00
(B) $7.60
(C) $35.50
(D) $38.00
(E) $75.00
23. For an advertising promotion, an auto dealer hands out 1,000 lottery tickets with a prize of a new car worth $25,000. For someone with a single ticket, what is the standard deviation for the amount won?
   (A) $7.07
   (B) $25.00
   (C) $49.95
   (D) $790.17
   (E) $624,375

24. An insurance company charges $800 annually for car insurance. The policy specifies that the company will pay $1000 for a minor accident and $5000 for a major accident. If the probability of a motorist having a minor accident during the year is .2, and of having a major accident, .05 how much can the insurance company expect to make on a policy?
   (A) $200
   (B) $250
   (C) $300
   (D) $350
   (E) $450

25. A weighted die come up spots with the following probabilities:

<table>
<thead>
<tr>
<th>Spots</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.10</td>
</tr>
<tr>
<td>2</td>
<td>.15</td>
</tr>
<tr>
<td>3</td>
<td>.20</td>
</tr>
<tr>
<td>4</td>
<td>.25</td>
</tr>
<tr>
<td>5</td>
<td>.20</td>
</tr>
<tr>
<td>6</td>
<td>.10</td>
</tr>
</tbody>
</table>

   If two of these dice are thrown, what is the probability the sum is 10?
   (A) (.25)(.10)+(.20)^2
   (B) 2(.25)(.10)+(.20)^2
   (C) 2(.25)(.10)+2(.20)^2
   (D) (.10)(.20)(.10)+(.20)^2
   (E) (.10)(.20)(.10)+(.15)(.20)^2+ (.15)(.20)(.20)+ (.25)(.10)
2005 Form B Question 2  For an upcoming concert, each customer may purchase up to 3 child tickets and 3 adult tickets. Let \( C \) be the number of child tickets purchased by a single customer. The probability distribution of the number of child tickets purchased by a single customer is given in the table below.

<table>
<thead>
<tr>
<th>( c )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p(c) )</td>
<td>0.4</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
</tr>
</tbody>
</table>

(a) Compute the mean and the standard deviation of \( C \).

\[
E(X) = \mu_c = (0)(.4)+(1)(.3)+(2)(.2)+3(.1) = 1
\]

The number of child tickets purchased by a single customer is 1.

\[
\sigma_c = \sqrt{\sum p(x - \mu)^2} \quad \sigma_c = \sqrt{(0 - 1)^2(.4) + (1 - 1)^2(.3) + (2 - 1)^2(.2) + (3 - 1)^2(.1)} = 1
\]

The standard deviation for the number of child tickets purchased by a single customer is 1.

(b) Suppose the mean and the standard deviation of the number of adult tickets purchased by a single customer are 2 and 1.2, respectively. Assume that the numbers of child tickets and adult tickets purchased are independent random variables. Compute the mean and the standard deviation of the total number of adult and child tickets purchased by a single customer.

\[
\mu_{c+a} = 1 + 2 = 3
\]

\[
\sigma_{c+a} = \sqrt{(1)^2 + (1.2)^2} = \sqrt{2.44} = 1.562
\]

(c) Suppose each child ticket costs $15 and each adult ticket costs $25. Compute the mean and the standard deviation of the total amount spent per purchase.

\[
\mu = 1(15) + 2(25) = 65. \quad \text{The expected price is 65 dollars}
\]

\[
\sigma = \sqrt{(1)^2(15)^2 + (1.2)^2(25)^2} = \sqrt{225 + 900} = \sqrt{1125} = 33.54. \quad \text{The standard deviation is 33.54 dollars}
\]

2002 Form B Problem 2  Airlines routinely overbook flights because they expect a certain number of no-shows. An airline runs a 5 p.m. commuter flight from Washington, D.C., to New York City on a plane that holds 38 passengers. Past experience has shown that if 41 tickets are sold for the flight, then the probability distribution for the number who actually show up for the flight is as shown in the table below.

<table>
<thead>
<tr>
<th>Number who actually show up</th>
<th>36</th>
<th>37</th>
<th>38</th>
<th>39</th>
<th>40</th>
<th>41</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>0.46</td>
<td>0.30</td>
<td>0.16</td>
<td>0.05</td>
<td>0.02</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Assume that 41 tickets are sold for each flight.

a) There are 38 passenger seats on the flight. What is the probability that all passengers who show up for this flight will get a seat?

\[
.46 + .30 + .16 = .92
\]

b) What is the expected number of no-shows for this flight?

\[
(5)(.46) + (4)(.30) + (3)(.16) + (2)(.05) + (1)(.02) + (0)(.01) = 4.1 \text{ people are expected to no-show for any flight}
\]

c) Given that not all passenger seats are filled on a flight, what is the probability that only 36 passengers showed up for the flight?

\[
P(36) = .46 \quad \text{Probability that not all seats are filled is } .46 + .30 = .76 = \frac{.46}{.76} = .605. \quad \text{There is a 60.5% chance that only 36 people will board the plane given that not all of the seats are filled.}
\]
Probability—Combining Random Variables

26. Suppose X and Y are random variables with $\mu_x=10$, $\sigma_x=3$, $\mu_y=15$, and $\sigma_y=4$. Given that X and Y are independent, what are the mean and standard deviation of the random variable $X + Y$.

(A) $\mu_{x+y}=25$, $\sigma_{x+y}=3.5$
(B) $\mu_{x+y}=25$, $\sigma_{x+y}=5$
(C) $\mu_{x+y}=25$, $\sigma_{x+y}=7$
(D) $\mu_{x+y}=12.5$, $\sigma_{x+y}=7$
(E) There is insufficient information to answer this question.

27. Suppose X and Y are random variables with $E(X)=500$, $Var(X)=50$, $E(Y)=400$, and $Var(Y)=30$. Given that X and Y are independent, what are the expected value and variance of the random sample $X - Y$?

(A) $E(X-Y)=100$, $Var(X-Y)=20$
(B) $E(X-Y)=100$, $Var(X-Y)=80$
(C) $E(X-Y)=900$, $Var(X-Y)=20$
(D) $E(X-Y)=900$, $Var(X-Y)=80$
(E) There is insufficient information to answer this question.

26. Suppose the average height of policemen is 71 inches with a standard deviation of 4 inches, while the average for policewomen is 66 inches with a standard deviation of 3 inches. If a committee looks at all ways of pairing up one male with one female officer, what will be the mean and standard deviation for the difference in heights for the set of possible partners?

(A) Mean of 5 inches with a standard deviation of 1 inches.
(B) Mean of 5 inches with a standard deviation of 3.5 inches.
(C) Mean of 5 inches with a standard deviation of 5 inches.
(D) Mean of 68.5 inches with a standard deviation of 1 inch.
(E) Mean of 68.5 inches with a standard deviation of 3.5 inches.
Probability—Combining Random Variables

2008  Problem 3  A local arcade is hosting a tournament in which contestants play an arcade game with possible scores ranging from 0 to 20. The arcade has set up multiple game tables so that all contestants can play the game at the same time; thus contestant scores are independent. Each contestant’s score will be recorded as he or she finishes, and the contestant with the highest score is the winner.

After practicing the game many times, Josephine, one of the contestants, has established the probability distribution of her scores, shown in the table below.

<table>
<thead>
<tr>
<th>Score</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>0.10</td>
</tr>
<tr>
<td>17</td>
<td>0.30</td>
</tr>
<tr>
<td>18</td>
<td>0.40</td>
</tr>
<tr>
<td>19</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Crystal, another contestant, has also practiced many times. The probability distribution for her scores is shown in the table below.

<table>
<thead>
<tr>
<th>Score</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>0.45</td>
</tr>
<tr>
<td>18</td>
<td>0.40</td>
</tr>
<tr>
<td>19</td>
<td>0.15</td>
</tr>
</tbody>
</table>

(a) Calculate the expected score for each player.

\[ E(X) = \mu_j = (16)(0.10) + (17)(0.30) + (18)(0.40) + (19)(0.20) \quad \mu_j = 17.7 \]

\[ E(X) = \mu_c = (17)(0.45) + (18)(0.40) + (19)(0.15) \quad \mu_c = 17.7 \]

(b) Suppose that Josephine scores 16 and Crystal scores 17. The difference (Josephine minus Crystal) of their scores is -1. List all combinations of possible scores for Josephine and Crystal that will produce a difference (Josephine minus Crystal) of -1, and calculate the probability of each combination.

<table>
<thead>
<tr>
<th>Scores J-C</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>16-17</td>
<td>(0.1)(0.45) = 0.045</td>
</tr>
<tr>
<td>17-18</td>
<td>(0.3)(0.40) = 0.12</td>
</tr>
<tr>
<td>18-19</td>
<td>(0.40)(0.15) = 0.06</td>
</tr>
<tr>
<td>Total</td>
<td>0.225</td>
</tr>
</tbody>
</table>

(c) Find the probability that the difference (Josephine minus Crystal) in their scores is -1. The probability is 0.045 + 0.12 + 0.06 = 0.225

(d) The table below lists all the possible differences in the scores between Josephine and Crystal and some associated probabilities.

<table>
<thead>
<tr>
<th>Difference</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3</td>
<td>0.015</td>
</tr>
<tr>
<td>-2</td>
<td>0.085</td>
</tr>
<tr>
<td>-1</td>
<td>0.225</td>
</tr>
<tr>
<td>0</td>
<td>0.325</td>
</tr>
<tr>
<td>1</td>
<td>0.260</td>
</tr>
<tr>
<td>2</td>
<td>0.090</td>
</tr>
</tbody>
</table>

Complete the table and calculate the probability that Crystal’s score will be higher than Josephine’s score.

\[ (0.015 + 0.085 + 0.225 + 0.325 + 0.260 + 0.090) = 1 \]

\[ P(-2) = 0.085 \]

The probability that Crystal beats Josephine is 0.015 + 0.085 + 0.225 = 0.325
27. The distribution of the diameters of a particular variety of oranges is approximately normal with a standard deviation of 0.3 inch. How does the diameter of an orange at the 67th percentile compare with the mean diameter?
   (A) 0.201 inch below the mean
   (B) 0.132 inch below the mean
   (C) 0.132 inch above the mean
   (D) 0.201 inch above the mean
   (E) 0.440 inch above the mean

28. The weights of a population of adult male gray whales are approximately normally distributed with a mean weight of 18,000 kilograms and a standard deviation of 4,000 kilograms. The weights of a population of adult male humpback whales are approximately normally distributed with a mean weight of 30,000 kilograms and a standard deviation of 6,000 kilograms. A certain adult male gray whale weighs 24,000 kilograms. This whale would have the same standardized weight (z-score) as an adult male humpback whale whose weight, in kilograms, is equal to which of the following?
   (A) 21,000
   (B) 24,000
   (C) 30,000
   (D) 36,000
   (E) 39,000

29. The lengths of individual shellfish in a population of 10,000 shellfish are approximately normally distributed with mean 10 centimeters and standard deviation 0.2 centimeter. Which of the following is the shortest interval that contains approximately 4,000 shellfish lengths?
   (A) 0 cm to 9.949 cm
   (B) 9.744 cm to 10 cm
   (C) 9.744 cm to 10.256 cm
   (D) 9.895 cm to 10.105 cm
   (E) 9.9280 cm to 10.080 cm

30. The distribution of the weights of loaves of bread from a certain bakery follows approximately a normal distribution. Based on a very large sample, it was found that 10 percent of the loaves weighed less than 15.34 ounces, and 20 percent of the loaves weighed more than 16.31 ounces. What are the mean and standard deviation of the distribution of the weights of the loaves of bread?
   
   Hint: Algebra

   (A) $\mu = 15.82$, $\sigma = 0.48$
   (B) $\mu = 15.82$, $\sigma = 0.69$
   (C) $\mu = 15.87$, $\sigma = 0.50$
   (D) $\mu = 15.93$, $\sigma = 0.46$
   (E) $\mu = 16.00$, $\sigma = 0.50$
Probability—Normal, Binomial & Geometric Distributions

2002 Problem 3 There are 4 runners on the New High School team. The team is planning to participate in a race in which each runner runs a mile. The team time is the sum of the individual times for the 4 runners. Assume that the individual times of the 4 runners are all independent of each other. The individual times, in minutes, of the runners in similar races are approximately normally distributed with the following means and standard deviations.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Runner 1</td>
<td>4.9</td>
<td>0.15</td>
</tr>
<tr>
<td>Runner 2</td>
<td>4.7</td>
<td>0.16</td>
</tr>
<tr>
<td>Runner 3</td>
<td>4.5</td>
<td>0.14</td>
</tr>
<tr>
<td>Runner 4</td>
<td>4.8</td>
<td>0.15</td>
</tr>
</tbody>
</table>

a) Runner 3 thinks that he can run a mile in less than 4.2 minutes in the next race. Is this likely to happen? Explain.

\[ P(\text{Runner 3} < 4.2) = P\left(\frac{4.2 - 4.5}{0.14}\right) = p(z < -2.14) = 0.0162. \] There is a 1.6% chance that Runner 3 will run a mile in less than 4.2 minutes.

b) The distribution of possible team times is approximately normal. What are the mean and standard deviation of this distribution?

The mean of the distribution is \( \mu_1 + \mu_2 + \mu_3 + \mu_4 = (4.9 + 4.7 + 4.5 + 4.8) = 18.9 \). The runners times are independent therefore the \( \sigma = \sqrt{\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2} = \sqrt{0.15^2 + 0.16^2 + 0.14^2 + 0.15^2} = 0.3003 \)

c) Suppose the team’s best time to date is 18.4 minutes. What is the probability that the team will beat its own best time in the next race?

\[ P(\text{best time} < 18.4) = P\left(\frac{18.4 - 18.9}{0.3003}\right) = p(z < -1.665) = 0.048. \] There is about a 4.8% chance that the team will beat its best time.
2011 Form B Problem 3 An airline claims that there is a 0.10 probability that a coach-class ticket holder who flies frequently will be upgraded to first class on any flight. This outcome is independent from flight to flight. Sam is a frequent flier who always purchases coach-class tickets.

(a) What is the probability that Sam's first upgrade will occur after the third flight?

Geometric distribution with $p = 0.10$

Let $x$ represent the number of upgrades

$P(x > 3) = P($Sam's flight did not occur during the 1st 3 flights$)$

$= 0.9 \times 0.9 \times 0.9 = 0.729$ There is a 72.9% chance that Sam's flight was not upgraded during the 1st 3 flights.

(b) What is the probability that Sam will be upgraded exactly 2 times in his next 20 flights?

Let $x$ represent the number of upgrades

Binomial distribution with $p = 0.1$ trials = 20 $x = 2$

$P(x=2) = \binom{20}{2}(0.1)^2(0.9)^2 = 0.285$

There is a 28.5% chance that Sam will have exactly 2 upgrades in 20 flights.

(c) Sam will take 104 flights next year. Would you be surprised if Sam receives more than 20 upgrades to first class during the year? Justify your answer.

Let $x$ represent the number of upgrades

Binomial distribution with $p = 0.1$ trials = 104

$P(x > 20) = 1 - P(x \leq 20)$

$= 1 - \left[ \left( \binom{104}{0}(0.1)^0(0.9)^{104} \right) + \ldots + \left( \binom{104}{20}(0.1)^{20}(0.9)^{84} \right) \right] = 0.0014$

It is unlikely that Sam will have more than 20 upgrades during the next 104 flights. The probability of more than 20 upgrades is 0.0014.
2006 Form B Problem 3  Golf balls must meet a set of five standards in order to be used in professional tournaments. One of these standards is distance traveled. When a ball is hit by a mechanical device, Iron Byron, with a 10-degree angle of launch, a backspin of 42 revolutions per second, and a ball velocity of 235 feet per second, the distance the ball travels may not exceed 291.2 yards. Manufacturers want to develop balls that will travel as close to the 291.2 yards as possible without exceeding that distance. A particular manufacturer has determined that the distances traveled for the balls it produces are normally distributed with a standard deviation of 2.8 yards. This manufacturer has a new process that allows it to set the mean distance the ball will travel.

(a) If the manufacturer sets the mean distance traveled to be equal to 288 yards, what is the probability that a ball that is randomly selected for testing will travel too far?

\[ Z = \frac{x-\mu}{\sigma} \]

\[ P(x > 291.2) = P(z > \frac{291.2-288}{2.8}) \Rightarrow P(z > 1.143) = .127 \]

The probability is .127 that the golf ball will travel farther than 291.2 yards.

(b) Assume the mean distance traveled is 288 yards and that five balls are independently tested. What is the probability that at least one of the five balls will exceed the maximum distance of 291.2 yards?

Binomial \[ P(x = 1 + \ldots + x = 5) = \binom{5}{1}(.127)^1(.873)^4 + \binom{5}{2}(.127)^2(.873)^3 + \binom{5}{3}(.127)^3(.873)^2 + \binom{5}{4}(.127)^4(.873)^1 + \binom{5}{5}(.127)^5(.873)^0 \]

\[ P = .127 = 1 - \left[ \binom{5}{0}(.127)^0(.873)^5 \right] \]

The probability that at least 1 of 5 balls travels farther than 291.2 is .493.

(c) If the manufacturer wants to be 99 percent certain that a randomly selected ball will not exceed the maximum distance of 291.2 yards, what is the largest mean that can be used in the manufacturing process?

\[ Z = \frac{x-\mu}{\sigma} \]

\[ Z = \frac{x-\mu}{\sigma} \quad 2.326 = \frac{291.2-\mu}{2.8} \quad \mu = 284.686 \]

The largest mean that can be used for the golf ball flight is 284.6 yards.
2003 Problem 3 Men's shirt sizes are determined by their neck sizes. Suppose that men's neck sizes are approximately normally distributed with mean 15.7 inches and standard deviation 0.7 inch. A retailer sells men's shirts in sizes S, M, L, and XL, where the shirt sizes are defined in the table below.

<table>
<thead>
<tr>
<th>Shirt Size</th>
<th>Neck size</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>14 ≤ neck size &lt; 15</td>
</tr>
<tr>
<td>M</td>
<td>15 ≤ neck size &lt; 16</td>
</tr>
<tr>
<td>L</td>
<td>16 ≤ neck size &lt; 17</td>
</tr>
<tr>
<td>XL</td>
<td>17 ≤ neck size &lt; 18</td>
</tr>
</tbody>
</table>

(a) Because the retailer only stocks the sizes listed above, what proportion of customers will find that the retailer does not carry any shirts in their sizes? Show you work.

\[ P(X < 14) + P(X > 18) = 1 - P(14 < x < 18) \]

\[ P(X < 14) + P(X > 18) = 1 - [P(Z < \frac{18 - 15.7}{0.7}) - P(Z < \frac{14 - 15.7}{0.7})] = 0.0081 \]

.81% of the population will not find their shirt size.

(b) Using a sketch of a normal curve, illustrate the proportion of men whose shirt size is M. Calculate this proportion.

\[ P(15 < x < 16) = [P(Z < \frac{16 - 15.7}{0.7}) - P(Z < \frac{15 - 15.7}{0.7})] = 0.507 \]

50.7% of the population wear size medium shirts.

(c) Of 12 randomly selected customers, what is the probability that exactly 4 will request size M? Show you work.

Binomial p = .507
Trials = 12
\[ X = 4 \]

\[ \binom{12}{4} \times (0.507)^4 \times (0.493)^8 = 0.1141 \]

There is a 0.1141 probability that exactly 4 out of 12 customers will choose a medium shirt.
Displaying Data

1. Qualitative/Categorical
   - Dot Plots, Bar Charts, Segmented Bar Charts, Pie Charts-relative Frequencies
   - Mode

2. Quantitative
   - Histograms, Stem Plots, Dot Plots, box plots, cumulative frequency plots, pie charts
   - CUSS-Center Shape Spread Unusual Features
   - Center- Mean, median mode,
   - Shape-Skewed, uniform, symmetric, bi-modal, mound
   - Spread-range, IQR, variance/standard deviation
   - Unusual Features-outliers, gaps

3. Effects of changing units
   - Adding a Constant
   - Multiplying by a Constant
When a tractor pulls a plow through an agricultural field, the energy needed to pull that plow is called the draft. The draft is affected by environmental conditions such as soil type, terrain, and moisture.

A study was conducted to determine whether a newly developed hitch would be able to reduce draft compared to the standard hitch. (A hitch is used to connect the plow to the tractor.) Two large plots of land were used in this study. It was randomly determined which plot was to be plowed using the standard hitch. As the tractor plowed that plot, a measurement device on the tractor automatically recorded the draft at 25 randomly selected points in the plot.

After the plot was plowed, the hitch was changed from the standard one to the new one, a process that takes a substantial amount of time. Then the second plot was plowed using the new hitch. Twenty-five measurements of draft were also recorded at randomly selected points in this plot.

(a) What was the response variable in this study? Identify the treatments. What were the experimental units? The response variable is the draft/energy needed to pull the plow. The treatments are the type of hitch that was used—the standard hitch and the new hitch. The experimental units are the 2 different fields.

(b) Given that the goal of the study is to determine whether a newly developed hitch reduces draft compared to the standard hitch, was randomization used properly in this study? Justify your answer. Randomization was properly applied because the two hitches, the treatments, were randomly assigned to the fields, the experimental units, which were plowed.

(c) Given that the goal of the study is to determine whether a newly developed hitch reduces draft compared to the standard hitch, was replication used properly in this study? Justify your answer. Replication was not properly employed. Each treatment was applied to only one experimental unit. For replication to have occurred, the treatments/hitches would need to be used on multiple experimental units/plots of land.

(d) Plot of land is a confounding variable in this experiment. Explain why. The plot of land is a confounding variable in this experiment because differing soil conditions between plots such as rockiness, soil compaction, root growth, and wetness will impact the draft. Thus, we are not able to determine whether the hitch or the plot is creating the differences in the draft. The plot is confounded with the hitch.
A preliminary study conducted at a medical center in St. Louis has shown that treatment with small, low-intensity magnets reduces the self-reported level of pain in polio patients. During each session, a patient rested on an examining table in the doctor’s office while the magnets, embedded in soft pads, were strapped to the body at the site of pain. Sessions continued for several weeks, after which pain reduction was measured.

A new study is being designed to investigate whether magnets also reduce pain in patients suffering from herniated disks in the lower back. One hundred male patients are available for the new study.

a) Describe an appropriate design for the new study. Your discussion should briefly address treatments used, methods of treatment assignment, and what variables would be measured. Do not describe how the data would be analyzed.

I would place the names of all the male patients in a bin and mix thoroughly. The first 50 names draw would receive the magnet treatment. The remaining 50 would serve as the control and receive a placebo. I would double blind the experiment. To do so both the patients receiving treatment and those serving as a control would have soft pads strapped to their bodies at the pain site. Those receiving treatment would have magnets placed in the pouches while those not receiving treatment would have ferrous metal of similar size, shape and color placed in the pouches. In so doing those receiving the treatment and those administering the treatment would not be able to distinguish the differences. After an appropriate amount of treatment time had passed the average pain levels of those with the magnet treatments and those would be compared to determine the effectiveness of the treatment.

b) Would you modify the design above if, instead of 100 male patients, there were 50 male and 50 female patients available for the study? If so, how would you modify your design? If not, why not?

If there were both male and female patients I would block based on gender because treatment results might vary by gender. I would place the 50 male names into a bin and mix. The first 25 names drawn would receive the magnet treatment the remaining 25 names would serve as the control and would be administered a placebo as mentioned above. I would also place the names of the 50 females in a separate bin and mix. The first 25 names drawn would receive the magnet treatment the remaining 25 names would serve as the control and would be administered a placebo as mentioned above. The experiment could be double blind as mentioned above.
1. The boxplots summarize two data sets, A and B. Which of the following must be true?
   I. Set A contains more data than Set B.
   II. The box of Set A contains more data than the box of Set B.
   III. The data in Set A have a larger range than the data in Set B.
   (A) I only
   (B) III only
   (C) I and II only
   (D) II and III only
   (E) I, II, and III

2. The heights of adult women are approximately normally distributed about a mean of 65 inches with a standard deviation of 2 inches. If Rachael is at the 99th percentile in height for adult women, then her height, in inches, is closest to
   (A) 60
   (B) 62
   (C) 68
   (D) 70
   (E) 74

3. A survey of 57 students was conducted to determine whether or not they held jobs outside of school. The two-way table above shows the numbers of students by employment status (job, no job) and class (juniors, seniors). Which of the follow best describes the relationship between employment status and class?
   (A) There appears to be no association, since the same number of juniors and seniors have jobs.
   (B) There appears to be no association, since close to half of the students have jobs.
   (C) There appears to be an association, since there are more seniors than juniors in the survey.
   (D) There appears to be an association, since the proportion of juniors having jobs is much larger than the proportion of seniors having jobs.
   (E) A measure of association cannot be determined from these data.

4. Gina’s doctor told her that the standardized score (z-score) for her systolic blood pressure, as compared to the blood pressure of other women her age, is 1.50. Which of the following is the best interpretation of this standardized score?
   (A) Gina’s systolic blood pressure is 150.
   (B) Gina’s systolic blood pressure is 1.50 standard deviations above the average systolic blood pressure of women her age.
   (C) Gina’s systolic blood pressure is 1.50 above the average systolic blood pressure of women her age.
   (D) Gina’s systolic blood pressure is 1.50 times the average systolic blood pressure of women her age.
   (E) Only 1.5% of women Gina’s age have a higher systolic blood pressure than she does.
5. A company wanted to determine the health care costs of its employees. A sample of 25 employees were interviewed and their medical expenses for the previous year were determined. Later the company discovered that the highest medical expense in the sample was mistakenly recorded as 10 times the actual amount. However, after correcting the error, the correct amount was still greater than or equal to any other medical expense in the sample. Which of the following sample statistics must have remained the same after the correction was made?
(A) Mean  
(B) Median  
(C) Mode  
(D) Range  
(E) Variance

6. The back-to-back stem-and-leaf plot below gives the percentage of students who dropped out of school at each of the 49 high schools in a large metropolitan school district.

<table>
<thead>
<tr>
<th>School Year</th>
<th>School Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 4</td>
<td>0 4</td>
</tr>
<tr>
<td>9 9 9 9 9 8 8 7</td>
<td>0 5 6 6 6 7 7 7 8 8 8 9 9</td>
</tr>
<tr>
<td>4 4 4 4 4 3 3 2 2 2 2 1 1 1 1</td>
<td>1 0 0 0 0 1 1 1 1 2 2 2 3 3 4 4 4 4</td>
</tr>
<tr>
<td>9 9 9 7 7 6 6 6 6 5</td>
<td>1 5 5 6 6 6 6 7 7 7 7 8</td>
</tr>
<tr>
<td>4 2 2 2 1 0 0</td>
<td>2 1 3</td>
</tr>
<tr>
<td>8 8 8 7 6 2</td>
<td></td>
</tr>
<tr>
<td>2 2 3</td>
<td>3 0 1 1 2</td>
</tr>
<tr>
<td>7 6 6</td>
<td>3 5</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

For 1992-1993, 1 | 2 represents 12%

Which of the following statements is NOT justified by these data?

(A) The drop-out rate decreased in each of the 49 high schools between the 1989-1990 and 1992-1993 school years.
(B) For the school years shown, most students in the 49 high schools did not drop out of high school.
(C) In general, drop-out rates decreased between the 1989-1990 and 1992-1993 school years.
(D) The median drop-out rate of the 49 high schools decreased between the 1989-1990 and 1992-1993 school years.
(E) The spread between the schools with the lowest drop-out rates and those with the highest drop-out rates did not change much between the 1989-1990 and 1992-1993 school years.

7. For which of the following distributions is the mean greater than the median?

(A)  
(B)  
(C)  
(D)  
(E)  

For each graph, the x-axis ranges from 0 to 10, and the y-axis ranges from 0 to 10.
8. The distribution of the weights of loaves of bread from a certain bakery follows approximately a normal distribution. Based on a very large sample, it was found that 10 percent of the loaves weighed less than 15.34 ounces, and 20 percent of the loaves weighed more than 16.31 ounces. What are the mean and standard deviation of the distribution of the weights of the loaves of bread?

Hint: Algebra

(A) \( \mu = 15.82, \sigma = 0.48 \)
(B) \( \mu = 15.82, \sigma = 0.69 \)
(C) \( \mu = 15.87, \sigma = 0.50 \)
(D) \( \mu = 15.93, \sigma = 0.46 \)
(E) \( \mu = 16.00, \sigma = 0.50 \)

9. Suppose that the distribution of a set of scores has a mean of 47 and a standard deviation of 14. If 4 is added to each score, what will be the mean and the standard deviation of the distribution of new scores?

<table>
<thead>
<tr>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td>51</td>
</tr>
<tr>
<td>(B)</td>
<td>51</td>
</tr>
<tr>
<td>(C)</td>
<td>47</td>
</tr>
<tr>
<td>(D)</td>
<td>47</td>
</tr>
<tr>
<td>(E)</td>
<td>47</td>
</tr>
</tbody>
</table>

10. The lengths of individual shellfish in a population of 10,000 shellfish are approximately normally distributed with mean 10 centimeters and standard deviation 0.2 centimeter. Which of the following is the shortest interval that contains approximately 4,000 shellfish lengths?

(A) 0 cm to 9.949 cm
(B) 9.744 cm to 10 cm
(C) 9.744 cm to 10.256 cm
(D) 9.895 cm to 10.105 cm
(E) 9.9280 cm to 10.080 cm

11. The boxplots shown above summarize two data sets, I and II. Based on the boxplots, which of the following statements about these two data sets CANNOT be justified?

(A) The range of data set I is equal to the range of data set II.
(B) The interquartile range of data set I is equal to the interquartile range of data set II.
(C) The median of data set I is less than the median of data set II.
(D) Data set I and data set II have the same number of data points.
(E) About 75% of the values in data set II are greater than or equal to about 50% of the values in data set I.
12. The figure above shows a cumulative relative frequency histogram of 40 scores on a test given in an AP Statistics class. Which of the following conclusions can be made from the graph?
(A) There is greater variability in the lower 20 test scores than in the higher 20 test scores.
(B) The median test score is less than 50.
(C) Sixty percent of the students had test scores above 80.
(D) If the passing score is 70, most students did not pass the test.
(E) The horizontal nature of the graph for the test scores of 60 and below indicates that those scores occurred most frequently.

13. A small town employs 34 salaried, nonunion employees. Each employee receives an annual salary increase of between $500 and $2000 based on a performance review by the mayor’s staff. Some employees are members of the mayor’s political party, and the rest are not. Students at the local high school form two lists, A and B, one for the raises granted to employees who are in the mayor’s party, and the other for raises granted to employees who are not. They want to display a graph (or graphs) of the salary increases in the student newspaper that readers can use to judge whether the two groups of employees have been treated in a reasonably equitable manner.

Which of the following displays is least likely to be useful to readers for this purpose?
(A) Back-to-back stemplots of A and B
(B) Scatterplot of B versus A
(C) Parallel boxplots of A and B
(D) Histograms of A and B that are drawn to the same scale
(E) Dotplots of A and B that are drawn to the same scale

14. The statistics below provide a summary of the distribution of heights, in inches, for a simple random sample of 200 young children.

Mean: 46 inches
Median: 45 inches
Standard Deviation: 3 inches
First Quartile: 43 inches
Third Quartile: 48 inches

About 100 children in the sample have heights that are
(A) less than 43 inches
(B) less than 48 inches
(C) between 43 and 48 inches
(D) between 40 and 52 inches
(E) more than 46 inches

15. The distribution of the diameters of a particular variety of oranges is approximately normal with a standard deviation of 0.3 inch. How does the diameter of an orange at the 67th percentile compare with the mean diameter?
(A) 0.201 inch below the mean
(B) 0.132 inch below the mean
(C) 0.132 inch above the mean
(D) 0.201 inch above the mean
(E) 0.440 inch above the mean
16. The stemplot below shows the yearly earnings per share of stock for two different companies over a sixteen-year period.

<table>
<thead>
<tr>
<th>Company A</th>
<th>Company B</th>
</tr>
</thead>
<tbody>
<tr>
<td>92, 91, 90, 82, 78, 43, 38, 26</td>
<td>0 58, 75, 96, 98</td>
</tr>
<tr>
<td>49, 47, 44, 00</td>
<td>1 01, 10, 17, 21, 43, 43, 53, 65, 73</td>
</tr>
<tr>
<td>73, 27, 05, 02</td>
<td>2 09, 27, 29</td>
</tr>
<tr>
<td></td>
<td>3</td>
</tr>
</tbody>
</table>

Which of the following statements is true?
(A) The median of the earnings of Company A is less than the median of the earnings of the Company B.
(B) The range of the earnings of Company A is less than the range of the earnings of Company B.
(C) The third quartile of Company A is smaller than the third quartile of Company B.
(D) The mean of the earnings of Company A is greater than the mean of the earnings of Company B.
(E) The interquartile range of Company A is twice the interquartile range of Company B.

17. The histograms below represent the distribution of five different data sets, each containing 28 integers, from 1 through 7, inclusive. The horizontal and vertical scales are the same for all graphs. Which graph represents the data set with the largest standard deviation.

18. A botanist is studying the petal lengths, measured in millimeters, of two species of lilies. The boxplots above illustrate the distribution of petal lengths from two samples of equal size, one from species A and the other from species B. Based on these boxplots, which of the following is a correct conclusion about the data collected in this study?
(A) The interquartile ranges are the same for both samples.
(B) The range for species B is greater than the range for species A.
(C) There are more petal lengths that are greater than 70 mm for species A than there are for species B.
(D) There are more petal lengths that are greater than 40 mm for species B than there are for species A.
(E) There are more petal lengths that are less than 30 mm for species B than there are for species A.
19. The histogram below displays the times, in minutes, needed for each chimpanzee in a sample of 26 to complete a simple navigational task.

It was determined that the largest observation, 93, is an outlier since $Q_3 + 1.5(Q_3 - Q_1) = 87.125$. Which of the following boxplots could represent the information in the histogram?

(A) ![Boxplot A]

(B) ![Boxplot B]

(C) ![Boxplot C]

(D) ![Boxplot D]

(E) ![Boxplot E]
2010 Form B Question 1 As a part of the United States Department of agriculture's Super Dump cleanup efforts in the early 1990s, various sites in the country were targeted for cleanup. Three of the targeted sites—River X, River Y, and River Z—had become contaminated with pesticides because they were located near abandoned pesticide dump sites. Measurements of the concentration of aldrin (a commonly used pesticide) were taken at twenty randomly selected locations in each river near the dump sites.

The boxplots shown below display the five-number summaries for the concentrations, in parts per million (ppm) of aldrin, for the twenty locations that were sampled in each of the three rivers.

a) Compare the distributions of the concentration of aldrin among the three rivers. River X is skewed right and has the highest range, interquartile range and variability. River X also has the highest median with 50% of the readings being greater than 5.

River Y is reasonably symmetric and has median of 5 and its range is the second largest.

River Z is skewed left and has the smallest range, IQR and Variance. It has a median of 4.3. All of its readings are below the upper 50% of river X.

b) The twenty concentrations of aldrin for River X are given below.

3.4 4.0 5.6 3.7 8.0 5.5 5.3 4.2 4.3 7.3
8.6 5.1 8.7 4.6 7.5 5.3 8.2 4.7 4.8 4.6

Construct a stemplot that displays the concentrations of aldrin for River X.

3 4 7 4 0 2 3 6 6 6 7 8
5 1 3 3 5 6
6
7 3 5
8 0 2 6 7

713 = 7.3 ppm of aldrin

The gap at 6 cannot be seen in the box plot but is visible in the stemplot.

The gap at 6 cannot be seen in the box plot but is visible in the stemplot.
As gasoline prices have increased in recent years, many drivers have expressed concern about the taxes they pay on gasoline for their cars. In the United States, gasoline taxes are imposed by both the federal government and by individual states. The boxplot below shows the distribution of the state gasoline taxes, in cents per gallon, for all 50 states on January 1, 2006.

(a) Based on the boxplot, what are the approximate values of the median and the interquartile range of the distribution of state gasoline taxes, in cents per gallon? Mark and label the boxplot to indicate how you found the approximated values.

(b) The federal tax imposed on gasoline was 18.4 cents per gallon at the time the state taxes were in effect. The federal gasoline tax was added to the state gasoline tax for each state to create a new distribution of combined gasoline taxes. What are approximate values, in cents per gallon, of the median and interquartile range of the new distribution of combined gasoline taxes? Justify your answer.

When federal gas taxes are combined with the state gas taxes 18.4 cents per gallon is added to the tax charged by each state which results in the total tax distribution being an 18.4 shift increase relative to the state tax distribution. The median of the total tax distribution would be 21 + 18.4 or 39.4 cents per gallon and $Q_1 = 18 + 18.4$ or 36.4 and $Q_3 = 25 + 18.4$ or 43.4 cents per gallon. Thus the Interquartile range would for total gas tax would be 43.4 - 36.4 or 7 cents. Measures of spread do not change when a constant is added to every value, while measures of center shift by the value of the constant.
1998 Question 2 A plot of the number of defective items produced during 20 consecutive days at a factory is shown below.

(a) Draw a histogram that shows the frequencies of the number of defective items.

(b) Give one fact that is obvious from the histogram but is not obvious from the scatterplot.
   The data is mound shaped and symmetric with a center at 3 there are no outliers

(c) Give one fact that is obvious from the scatterplot but is not obvious from the histogram.
   The number of defects is declining over time
**2012 Question 3.** Independent random samples of 500 households were taken from a large metropolitan area in the United States for the years 1950 and 2000. Histograms of household size (number of people in a household) for the years are shown below.

a) Compare the distributions of household size in the metropolitan area for the years 1950 and 2000.

Both distributions are skewed right. The median household size in 1950 is about 5 whereas it is about 3 in 2000. The spread of the data is greater in 1950 than in 2000 as the range in 1950 was 1-14 and the range in 2000 was 1-12.
A large regional real estate company keeps records of home sales for each of its sales agents. Each month, the company publishes the sales volume for each agent. Monthly sales volume is defined as the total sales price of all homes sold by the agent during a month. The figure below displays the cumulative relative frequency plot of the most recent monthly sales volume (in hundreds of thousands of dollars) for these agents.

(a) In the context of this question, explain what information is conveyed by the circled point.

40% of the agents had monthly sales volumes of 300,000 dollars or less.

(b) What proportion of sales agents achieved monthly sales volume between $700,000 and $800,000?

.8-.7 = .1 10% of the agents achieved sales volumes between 700 and 800 thousand dollars.

(c) For values between 10 and 11 on the horizontal axis, the cumulative relative frequency plot is flat. In the context of this question, explain what this means.

The horizontal line indicates that no agent had monthly sales of 1-1.1 million dollars.

(d) A bonus is to be given to 20 percent of the sales agents. Those who achieved the highest monthly sales volume during the preceding month will receive a bonus. What is the minimum monthly sales volume an agent must have achieved to qualify for the bonus?

80% of the agents had monthly sales of less than 800,000 dollars. Thus the top 20 percent of agents had minimum sales of 800,000 dollars or more.