1. A large university has 27,000 undergraduate students and 3,000 graduate students. Suppose you select a S.R.S. of size 300 from the undergraduate student population and separately a S.R.S. of size 200 from the graduate student population. Because we have two S.R.S.s it is easy to estimate opinions in the two populations separately. Suppose estimates $p$-hat are calculated for the proportion of students favoring an issue in each of the two samples.

(a) Do a quick calculation of the margin of error for 95% confidence for each of the two sample statistics.

(b) The sample of undergraduates contains a smaller proportion of the population 1 out of 90, versus 1 out of 15 for graduates. Yet sampling 1 out 90 undergraduates gives a smaller margin of error than sampling 1 out of 15. Explain to someone who knows no statistics why this can happen.

2. According to a March 2007 *U.S.A. Today* /Gallup poll, 43% of Americans identify themselves as baseball fans. That is low by recent standards, as an average of 49% of Americans have said they were fans of the sport since Gallup started tracking this measure in 1993. The high point came in 1998, when Sammy Sosa and Mark McGwire pursued (and ultimately surpassed) Roger Maris’s single-season home run record, at which time 56% of Americans considered themselves baseball fans. The Gallup press release says:

For results based on this sample, one can say with 95% confidence that the maximum error attributable to sampling and other random effects is ± 3 percentage points.

(a) If you’ve been keeping abreast of baseball news, can you think of any reason why there might legitimately be fewer fans of professional baseball today compared with recent history?

(b) Give one example of an error in the poll result that is not included in the margin of error.

3. Comment on each of the following as a potential sample survey question. Are they slanted towards a desirable response? If yes, what kind of a error would it result in? Sampling or non-sampling error? Would the surveys announced margin of error take this source of error into account?

(a) In the light of escalating environmental degradation and incipient resource depletion, would you favor economic incentives for recycling of resource-intensive consumer goods?

(b) Which of the following represents your opinion on gun control?

(i) The government should take away our guns.
(ii) We have the right to bear arms.

c) Do you think that excessive restrictions placed on U.S. law enforcement hampered their ability to detect the 9/11 terrorists before their attack?

4. A survey organization wants to assess the proportion of Washingtonians who favor a sales tax hike. They use a S.R.S. of size 400 from the Seattle Metro directory and estimate the proportion to be 40%. What sources of error are likely present in this estimate? Are these errors examples of sampling error or non-sampling errors?

5. In the previous problem, suppose the survey organization decided to assess the proportion of Washingtonians who favor a sales tax hike by interviewing shoppers at a large mall. What sources of error are likely present in this estimate? Are these errors examples of sampling error or non-sampling errors?

6. In a study of the effectiveness of various punishments for drunk driving, three punishment options were considered: jail time, community service or a fine. Is the form of punishment a categorical or a quantitative variable? What values can this variable take?

7. Researchers in New Zealand interviewed 907 drivers at age 21. They had data on traffic accidents and they asked the drivers about marijuana use and whether the accidents were their fault. Here are data on the numbers of accidents caused by these drivers at age 19, broken down by marijuana use at the same time.

<table>
<thead>
<tr>
<th>Marijuana Use per Year</th>
<th>1-10 times</th>
<th>11-50 times</th>
<th>51+ times</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accidents</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At fault</td>
<td>59</td>
<td>36</td>
<td>15</td>
</tr>
<tr>
<td>Not at fault</td>
<td>393</td>
<td>193</td>
<td>55</td>
</tr>
<tr>
<td>Total</td>
<td>452</td>
<td>229</td>
<td>70</td>
</tr>
</tbody>
</table>

(a) Is this data from an observational study? Is it retrospectively or prospectively collected?

(b) Who are the individuals or observational units represented by this table? What two variables are measured on these individuals? Are they categorical or quantitative? What values can these variables take?

(c) Which of the two variables is considered explanatory and which is the response?

(d) Explain carefully why a useful graph must compare rates (at-fault accidents/total) rather than counts of accidents in the four marijuana classes.

(e) Compute the at-fault accident rates in the four marijuana use classes. After you have done this, make a graph that displays the accident rates for each class. This is called a bar graph. What do you conclude?

(f) Because the data comes from an observational study, you cannot conclude that marijuana use causes accidents. The reason is confounding variables. Can you think of a lurking variable that might potentially be confounding the relationship between marijuana use and at-fault accidents?
8. (Text, page 52, 8) The figure on the next page (adapted from the *San Francisco Chronicle*, May 18, 1992) shows the distribution of American families by income. Ranges include the left endpoint but not the right. For example, 3.7% of the families had incomes in the range $0-$4,999, 5.8% had incomes in the range $5,000-$9,999, and so forth. True or false, and explain.

(a) Although American families are not spread evenly over the whole income range, the families that earn between $10,000 and $35,000 are spread fairly evenly over that range.

![](image)

9. (Text, page 429, 5) The median salary of the 746 major league baseball players was about $500,000 in 1994. The lowest salary was $100,000 and the highest was over $5,000,000. Choose one option and explain.

(a) The total salary paid out to major league baseball players in 1994 was $373 million.

(b) The total salary paid out to major league baseball players in 1994 was substantially less than $373 million.

(c) The total salary paid out to major league baseball players in 1994 was substantially more than $373 million.

10. Household size in the U.S. has a mean of approximately 2.5 people. Which of these numbers would be a good guess for the standard deviation and why? 0.014, 0.14, 1.4 and 14. You may assume that the distribution of household size in the U.S. follows the 68-95 rule approximately.

11. Suppose that the proportion of all adult Americans who support an issue is $p=0.37$. If we took many S.R.S.s of size 3,500 from the population, the sample statistic $p$-hat would vary from sample to sample according to a Normal distribution with mean 0.37 and standard deviation 0.01. Use this fact to answer these questions.

(a) In many samples, what percentage of the values of $p$-hat fall above 0.37? Above 0.40?
(b) In a large number of samples, what range contains the central 95% of values of \( p \)-hat. (This range is called the “reasonable likely values” for \( p \)-hat with 95% confidence)