4. Data, Statistics, and Probability

Data, statistics, and probability are mathematical topics of widespread utility. In this set of problems, we treat these subjects.

There are several separate, albeit related, topics of concern here. The first is data and the various methods of representing it. Representation of data may be graphical (bar charts, scatter plots, and the like) or by means of statistics that describe the data (mean, median, standard deviation, quartiles, etc.). It is important that students learn the most common of these methods and, for statistics, the calculations that give the statistics. Many state standards give the impression that good representation of data is rather straightforward and that the focus should be on interpretation. Actually, deciding on an appropriate representation itself requires good judgment, often in combination with arithmetical, algebraic, and/or geometrical understanding. *We believe that in state standards, one emphasis as regards Data and Analysis should be on the representation and statistical description of data*; see problems 1-5 below.

Another part of statistics treats the topic of drawing conclusions and making suitable predictions based on data. *We believe that the standards of some states as regards data analysis and experimental design should be significantly changed.* Instead of an emphasis on *making predictions*, we believe that the focus should be on *critical reasoning* in the context of data analysis. In particular, we recommend that the curriculum in this area include increased contrasting of valid and invalid arguments involving statistics. In the problems below, we give some examples.

A first reason for this recommendation is that it lays the foundation for the use of statistics in other courses, where subject-specific knowledge is often important, if not crucial, to a full understanding of the data and its implications. Indeed, we embrace including additional work with statistics in the science and social sciences curriculum. (We do, however, believe that the actual carrying out of all but the simplest experiments belongs in such classes rather than in the mathematics classroom.) Second, it is easy to misapply statistical methods (e.g. by treating dependent data as if it were independent), resulting in conclusions that are seriously flawed for non-obvious reasons. We believe that it is important for citizens to have a good understanding of such issues in order to critically analyze arguments using statistics. Third, in general we have great concern that students not be told in essence to plug in formulas that they may not genuinely understand and then to trust the answers. In the context of statistics we believe that this concern can be well-addressed by an increased emphasis on critical reasoning.

Let us also note that state examination questions in data analysis and in the related area of experimental design all too often, shockingly, reflect a real misunderstanding of this subject. Students are asked to draw a line of best fit when there is no indication that the problem is linear or of what "best fit" means, to design experiments without the proper mathematical background in this area (e.g. a discussion of correlation), to model data without being able to measure whether the model fits the data well, and to draw conclusions from experiments though those conclusions cannot be justified without additional, unstated, hypotheses. These problems suggest a certain overreaching. We believe that students would be far better served

by learning instead of mislearning. It may be wiser to cover less material about statistics but to cover it more thoroughly, and then to encourage students to continue to learn about statistics as their mathematical background increases. We also believe that, outside of descriptive statistics, this material is best taught in a focused way once students already have a solid mathematical foundation in arithmetic and middle-school mathematics (including algebra), rather than being presented in smaller, continuing, doses.

Problems related to probability are also given in this section, primarily to highlight the distinction between probability, which is a purely mathematical subject motivated by real-world connections, and concerns with data and statistics in the real world. We explain this relation further in the subsection on Probability below.

Data and Statistics

1: Make a bar chart showing the population in millions of the 10 provinces and 3 territories of Canada (counting Northwest Territories as a single territory).

Discussion: Typically such a problem would be accompanied by a table showing the population in each province and territory (omitted here). This problem, an exercise in presenting data nicely for the reader, is consistent with the standards of many states.

Bar charts are occasionally called bar graphs.

2: Convert the bar chart of Problem 1 to a bar chart representing percentages of the total Canadian population.

Discussion: This problem is for students who have learned about the connection between decimals and percentages. The bar chart is the same as for Problem 1 except for labels.

3: Construct a bar chart showing the average populations per square mile of the provinces and territories of Canada. Make sure that the bar chart fits on one piece of paper but is not minuscule.

Discussion: It would be natural to use a calculator for this problem. Here, it would be a mistake to ask students to convert the bar chart to one representing percentages of the whole, since these data are themselves percentages that have been obtained by dividing by different numbers for different items.

Problems 1–3, Additional Discussion: There are many methods of presenting data. It is more important that students learn the advantages and disadvantages of the most common methods

rather than obtain an encyclopedic knowledge of less common methods. Though we have not included problems about pie charts, such charts are popular in newspapers and other data presentations and should be discussed.

In the next two problems we are concerned with the statistics that can be calculated from numerical data as an aide to describing the data: average (also called mean), median, quartiles, quantiles, standard deviations, and variances. Students should understand these descriptors, and also know which changes in data affect the mean mostly and which have more influence on the median.

4: At our company, the mean annual salary is \$100,000; the median annual salary is \$25,000. Our company has 1,000 employees.

(a) Can the total annual amount allocated to paying salaries be determined from just the median salary and the number of employees? If it can, do so.

(b) Can the total annual amount allocated to paying salaries be determined from just the mean salary and the number of employees? If it can, do so.

Discussion: The answer to (a) is "no". The answer to (b) is "yes".

Here is an inappropriate 'problem' based on the data in this problem: *Which better represents the salaries of employees at this company: the mean or the median?* Outside of a context, this question cannot be answered. If the purpose is to calculate the annual amount allocated to salaries or some related task, then the mean is better. On the other hand, a prospective employee might be more interested in the median or even the first quartile. *Here and in general, the choice of which statistic or statistics to use to best describe given data depends on context, that is, on how the information will be used.*

5: Sally uses a measuring tape to measure the heights of all the children in her class, rounding each measurement to the nearest whole number of inches. Her data are

$$50\,,\ 50\,,\ 50\,,\ 52\,,\ 53\,,\ 53\,,\ 57\,,\ 58\,,\ 58\,,\ 59\,,\ 60\,,\ 61\,,\ 61\,.$$

Find the mean, median, first quartile, and all modes of this collection of data.

Discussion: Notice that there is a unique mode—namely 50. Sometimes modes are mistakenly described as indicators of 'centers of data'. This problem shows that any such phrase is misleading as a description of modes. In fact, modes are far less important than means and medians; modes are very sensitive to small changes in data whereas means and medians are not.

Inherent in data organization is the issue of rounding numbers. It is a good exercise for students who have had algebra to prove that if Sally had re-measured to the nearest half-inch,

then—whatever the actual heights of the 13 students—the means of Sally's first data and her second would differ by no more than 1/2, and that the same is true for the medians and quartiles.

Good representations of data accompanied by calculations of important statistics seem to invite one to draw conclusions or make predictions about future events or even to make certain decisions. But in doing so it is easy to make subtle but important errors without having a hint that errors are being made. Throughout their lives, students will be confronted by others making claims on the basis of statistics, and it is thus important for them to learn the variety of subtle but deceiving mistakes that can arise. The problems below, which require the student to look critically at predictions and other uses of data, help promote this learning.

6: What is wrong with the following problem and proposed solution?

> The average snowfall in Minneapolis, Minnesota during the fall-winter-spring snow season is 56.3 inches. Suppose that at the end of February in some particular snow season, the total snowfall thus far is 15.2 inches. What is a good estimate of the amount of snow yet to come during that snow season?
>
> Proposed solution: 56.3-15.2 = 41.1 inches

Discussion: The major error: a possibility has been overlooked—that the 15.2 inches of snow thus far observed might indicate that the total snowfall for the entire snow season will be far below the average of 56.3 inches. (For instance, winter weather patterns are affected by global ocean currents.) We have no information about the average snowfall in years for which the snowfall through February is only 15.2 inches. Indeed, if the snowfall through the entire snow season less one day were 25 inches, would the best estimate for the last day be a 31.3 inch blizzard?!

7: What is wrong with the following problem and proposed solution?

> The average snowfall in Minneapolis, Minnesota during the fall-winter-spring snow season is 56.3 inches, and the average for the last three months—March, April, and May—of the snow season is 13.7 inches. Suppose that at the end of February in some particular snow season, the total snowfall so far is 15.2 inches. What is a good estimate for the amount of snow yet to come during that snow season?
>
> Proposed solution: 13.7 inches, the mean for the remaining three months

Discussion: Error: In the proposed solution, the other information in the problem has been ignored, and this should only be done if the snowfall in the last three months is independent of the snowfall during the earlier part of the snow season. The problem gives no definite information about dependence or independence.

Information from previous years would be useful in assessing the question of dependence.

8: An advertisement says that, based on polling by an independent agency, 4 out of 5 dentists prefer Glistening White toothpaste to two other national brands. How might it be that the statement is factually correct but very misleading?

Discussion: There are many ways, so it might be a challenge to formulate a fair rubric for grading.

One way: Maybe the dentists were asked to rank eight different toothpastes and Glistening White was among the three worst but managed to be favored over two other national brands by 4 out of 5 dentists.

Second way: Maybe only 5 dentists were interviewed and the results are fully representative of the views of those 5 dentists. Of course, the person who hears this advertisement tends to imagine 500 dentists out of which 400 were enthusiastic about Glistening White.

Third way: Maybe several different independent agencies ran their own separate polls, say in different parts of the country, with widely varying results. And the Glistening White Company has chosen to only mention the one that is favorable to Glistening White toothpaste.

9: What is wrong with the following question?

> The students in Mr. Everson's biology classes are asked to vote for a biological symbol for the school year. The vote is among the following three symbols which have been previously nominated: frog, lizard, and gnu. Here are results of the vote:
>
> | | Class 1 | Class 2 | Class 3 |
> |---|---|---|---|
> | frog | 23 | 2 | 5 |
> | lizard | 7 | 6 | 16 |
> | gnu | 4 | 26 | 2 |
>
> What symbol should Mr. Everson choose and why?

Discussion: The appropriate question is: What mistake has Mr. Everson already made? Answer: He did not decide on the method of evaluating the vote and announce that decision before the vote was taken.

This issue is very closely related to an important issue in experimental design. Students should be taught that the plan for any experiment must include the questions to be addressed, a description of the data needed, how that data is to be collected, and how the data is to be analyzed, and this plan should be in place before the data is obtained. This should be included in the science curriculum where it can then be implemented in the science classroom as an illustration of the scientific method.

10: To carry out an experiment, a science class measures a quantity Q at 3 one-second

intervals. If at time $t = 0$, the quantity is given by $Q = 5$, at time $t = 1$ the quantity is $Q = 10$, and at time $t = 2$ the quantity is $Q = 15$, can you predict the quantity at time $t = 3$?

Discussion: Answer: there is no basis for making this prediction without additional information. Certainly, the quantity may be growing linearly, so that at $t = 3$ one would expect $Q = 20$, but the quantity may be better modeled by the equation $Q(t) = 5(t-1)^3 + 10$, or by $Q(t) = 10 - 5\cos(\pi t/2)$, or by infinitely many others.

Note that the above problems do not require advanced high-school mathematics. By contrast, algebra is useful for writing the formulas of statistics, as well as for understanding how changes of scale and accuracy of measurement affect statistical descriptions of data. Algebra, and in many cases Calculus, is needed for accurate modeling and predicting.

Probability

The relation of the mathematical subject of probability to the real world is similar to the relation of high-school geometry to the real world. No real-world triangles are ever exactly equilateral; similarly, no coins are ever exactly equally likely to come up heads as to come up tails. In geometry we do, nevertheless, spend considerable time talking about equilateral triangles and successfully apply the knowledge to the real world. And in probability we do discuss flipping fair coins and we apply the knowledge to many real-world coins as well as to situations where no coins are involved. (A fair coin is one that shows heads with probability 1/2 and tails with probability 1/2.)

Before giving problems on probability, we first consider one matter of terminology. Consider the experiment of flipping a coin—an actual physical coin. One might be interested in the probability that it comes up heads. To estimate this probability one might flip the coin a large number, say 1000, times. Alternatively, one might observe that the coin is almost symmetrical and therefore conclude that the probability is close to 1/2. These are two methods of obtaining an estimate; it is misleading to call the result of one method 'experimental probability' and the other result 'theoretical probability'. It would be useful if state standards were not to employ these two terms. Rather, there is probability and the various methods of estimating probability. The proper term for describing the proportion of times some particular event happens when the same experiment is repeated many times is 'relative frequency'.

11: Suppose that a fair coin is flipped three times and that the flips are independent. Calculate the probability that exactly two heads are obtained.

Discussion: One thing that can help simplify the preceding problem is to explicitly write down the sample space. We are pleased that many state standards introduce sample spaces in middle school.

12: There are 2 red markers in a bag and 3 green markers, which are otherwise indistinguishable. Lillian draws two markers at random, one after the other, leaving 3 markers in the bag. What is the probability that Lillian draws both red markers? What is the probability that she draws a red marker on her first draw or a green marker on her second draw? What is the probability that she draws a red marker on her first draw and a green marker on her second draw?

Discussion: The numbers in the preceding problem are small enough for one to write down an explicit sample space. The questions are about subsets of that sample space and some of the questions involve intersections and unions of some easily identifiable subsets. Handling the language of sets in this manner along with related words such as 'or' and 'and' uses an important mathematical skill. As this question and discussion illustrate, some aspects of probability reinforce important aspects of other areas of mathematics.

13: There are $r \geq 2$ red markers in a bag and $g \geq 1$ green markers, which are otherwise indistinguishable. Lillian draws two markers at random, one after the other, leaving $r + g - 2$ markers in the bag. What is the probability that Lillian draws two red markers? What is the probability that she draws a red marker on her first draw or a green marker on her second draw? What is the probability that she draws a red marker on her first draw and a green marker on her second draw?

Discussion: This problem is a higher-level version of the preceding problem. It involves algebra, and requires the student who is taking a sample-space point of view to imagine the details of the sample space rather than writing them all down.

14: For the setting of Problem 12, represent the sample space as a tree. Then use the tree and the probabilities associated with its branches to calculate the probability that Lillian's first draw was green, given that her second draw was red.

Discussion: Conditional probability, which shows how to use partial information in a systematic manner, and calculations based on trees do belong in high school standards. They have the feature of highlighting the subtlety of the subject, and therefore can help promote critical reasoning about data analysis.

15: A furniture manufacturer produces 20 chairs to identical specifications, but it happens that two of them have subtle defects. Three chairs are chosen at random. What is the probability that none of the three are defective? Also, what is the expected number of defectives in this experiment?

Discussion: Means, medians, standard deviations, and variances are calculated for probability distributions as well as for collections of numerical data. This use of the same terms in two

different settings can create confusion. With respect to this double usage, the synonyms for 'mean' can play a useful role. One uses 'average' for collections of data but never for probability distributions. The terms 'expectation', 'expected value', and 'expected number' are synonyms for 'mean' in the context of probability distributions but not for collections of data.

16: Suppose that 75% of the workers in a certain company drive to their job. If 3 workers selected at random are asked to stay late to help move furniture, is it more likely or less likely that all 3 have driven to work?

Discussion: This is a problem that every student should be able to solve. The answer is less likely: the probability is 27/64, which is less than 1/2.

In writing problems such as this, one must be careful concerning the issue of independence. For example, if the problem were changed slightly to read *"Suppose that 75% of the workers in a certain company drive to their job. If 3 workers meet for coffee at break, is it more likely or less likely that all 3 have driven to work?"*, then the problem is flawed. There is no reason to think that the workers meeting over coffee are a random sample; for instance, maybe people who carpool are more likely to have coffee together.