Progressions for the Common Core State Standards in Mathematics (draft)

26 December 2011

6–8 Statistics and Probability

Overview

In Grade 6, students build on the knowledge and experiences in data analysis developed in earlier grades (see K-3 Categorical Data Progression and Grades 2-5 Measurement Progression). They develop a deeper understanding of variability and more precise descriptions of data distributions, using numerical measures of center and spread, and terms such as cluster, peak, gap, symmetry, skew, and outlier. They begin to use histograms and box plots to represent and analyze data distributions. As in earlier grades, students view statistical reasoning as a four-step investigative process:

- Formulate questions that can be answered with data
- Design and use a plan to collect relevant data
- Analyze the data with appropriate methods
- Interpret results and draw valid conclusions from the data that relate to the questions posed.

Such investigations involve making sense of practical problems by turning them into statistical investigations (MP1); moving from context to abstraction and back to context (MP2); repeating the process of statistical reasoning in a variety of contexts (MP8).

In Grade 7, students move from concentrating on analysis of data to production of data, understanding that good answers to statistical questions depend upon a good plan for collecting data relevant to the questions of interest. Because statistically sound data production is based on random sampling, a probabilistic concept, students must develop some knowledge of probability before launching into sampling. Their introduction to probability is based on seeing probabilities of chance events as long-run relative frequencies of their occurrence, and many opportunities to develop the connection between theoretical probability models and empirical probability approximations. This connection forms the basis of statistical inference.

With random sampling as the key to collecting good data, students begin to differentiate between the variability in a sample and the variability inherent in a statistic computed from a sample when samples are repeatedly selected from the same population. This understanding of variability allows them to make rational decisions, say, about how different a proportion of "successes" in a sample is likely to be from the proportion of "successes" in the population or whether medians of samples from two populations provide convincing evidence that the medians of the two populations also differ.

Until Grade 8, almost all of students' statistical topics and investigations have dealt with univariate data, e.g., collections of counts or measurements of one characteristic. Eighth graders apply their experience with the coordinate plane and linear functions in the study of association between two variables related to a question of interest. As in the univariate case, analysis of bivariate measurement data graphed on a scatterplot proceeds by describing shape, center, and spread. But now "shape" refers to a cloud of points on a plane, "center" refers to a line drawn through the cloud that captures the essence of its shape, and "spread" refers to how far the data points stray from this central line. Students extend their understanding of "cluster" and "outlier" from univariate data to bivariate data. They summarize bivariate categorical data using two-way tables of counts and/or proportions, and examine these for patterns of association.

Grade 6

Develop understanding of statistical variability Statistical investigations begin with a question, and students now see that answers to such questions always involve variability in the data collected to answer them.[6.SP.1] Variability may seem large, as in the selling prices of houses, or small, as in repeated measurements on the diameter of a tennis ball, but it is important to interpret variability in terms of the situation under study, the question being asked, and other aspects of the data distribution (MP2). A collection of test scores that vary only about three percentage points from 90% as compared to scores that vary ten points from 70% lead to quite different interpretations by the teacher. Test scores varying by only three points is often a good situation. But what about the same phenomenon in a different context: percentage of active ingredient in a prescription drug varying by three percentage points from order to order?

Working with counts or measurements, students display data with the dot plots (sometimes called line plots) that they used in earlier grades. New at Grade 6 is the use of histograms, which are especially appropriate for large data sets.

Students extend their knowledge of symmetric shapes,[4.G.3] to describe data displayed in dot plots and histograms in terms of symmetry. They identify clusters, peaks, and gaps, recognizing common shapes[6.SP.2] and patterns in these displays of data distributions (MP7).

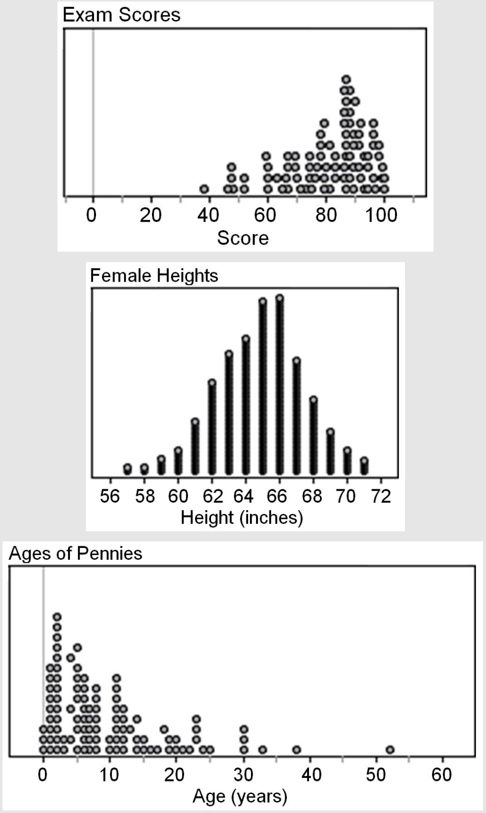
A major focus of Grade 6 is characterization of data distributions by measures of center and spread.[6.SP.2,6.SP.3] To be useful, center and spread must have well-defined numerical descriptions that are commonly understood by those using the results of a statistical investigation. The simpler ones to calculate and interpret are those based on counting. In that spirit, center is measured by the *median*, a number arrived at by counting to the middle of an ordered array of numerical data. When the number of data points is odd, the median is the middle value. When the number of data points is even, the median is the average of the two middle values. *Quartiles*, the medians of the lower and upper halves of the ordered data values, mark off the middle 50% of the data values and, thus, provide information on the spread of the data.[1] The distance between the first and third quartiles, the *interquartile range* (IQR), is a single number summary that serves as a very useful measure of variability.[6.SP.3]

Plotting the extreme values, the quartiles, and the median (the *five-number summary*) on a number line diagram, leads to the *box plot*, a concise way of representing the main features of a data dis-

[1] Different methods for computing quartiles are in use. The Standards uses the method which excludes the median to create two halves when the number of data points is odd. See Langford, "Quartiles in Elementary Statistics," *Journal of Statistics Education*, 2006, for a description of the different methods used by statisticians and statistical software.

[6.SP.1] Recognize a statistical question as one that anticipates variability in the data related to the question and accounts for it in the answers.

Dot plots: Skewed left, symmetric, skewed right

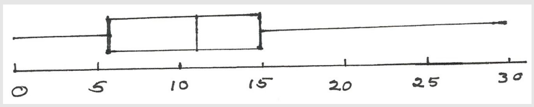


Students distinguish between dot plots showing distributions which are skewed left (skewed toward smaller values), approximately symmetric, and skewed right (skewed toward larger values). The plots show scores on a math exam, heights of 1,000 females with ages from 18 to 24, ages of 100 pennies in a sample collected from students.

[4.G.3] Recognize a line of symmetry for a two-dimensional figure as a line across the figure such that the figure can be folded along the line into matching parts. Identify line-symmetric figures and draw lines of symmetry.

[6.SP.2] Understand that a set of data collected to answer a statistical question has a distribution which can be described by its center, spread, and overall shape.

Box plot



For the data set {1, 3, 6, 7, 10, 12, 14, 15, 22, 30}, the median is 11 (from the average of the two middle values 10 and 12), the interquartile range is $15 - 6 = 9$, and the extreme values are 1 and 30.

tribution.• Box plots are particularly well suited for comparing two or more data sets, such as the lengths of mung bean sprouts for plants with no direct sunlight versus the lengths for plants with four hours of direct sunlight per day.[6.SP.4]

Students use their knowledge[6.NS.2,6.NS.3] of division, fractions, and decimals in computing a new measure of center—the *arithmetic mean*, often simply called the *mean*. They see the mean as a "leveling out" of the data in the sense of a unit rate (see Ratio and Proportion Progression). In this "leveling out" interpretation, the mean is often called the "average" and can be considered in terms of "fair share." For example, if it costs a total of $40 for five students to go to lunch together and they decide to pay equal shares of the cost, then each student's share is $8.00. Students recognize the mean as a convenient summary statistic that is used extensively in the world around them, such as average score on an exam, mean temperature for the day, average height and weight of a person of their age, and so on.

Students also learn some of the subtleties of working with the mean, such as its sensitivity to changes in data values and its tendency to be pulled toward an extreme value, much more so than the median. Students gain experience in deciding whether the mean or the median is the better measure of center in the context of the question posed. Which measure will tend to be closer to where the data on prices of a new pair of jeans actually cluster? Why does your teacher report the mean score on the last exam? Why does your science teacher say, "Take three measurements and report the average?"

For distributions in which the mean is the better measure of center, variation is commonly measured in terms of how far the data values deviate from the mean. Students calculate how far each value is above or below the mean, and these deviations from the mean are the first step in building a measure of variation based on spread to either side of center. The average of the deviations is always zero, but averaging the absolute values of the deviations leads to a measure of variation that is useful in characterizing the spread of a data distribution and in comparing distributions. This measure is called the *mean absolute deviation*, or MAD. Exploring variation with the MAD sets the stage for introducing the standard deviation in high school.

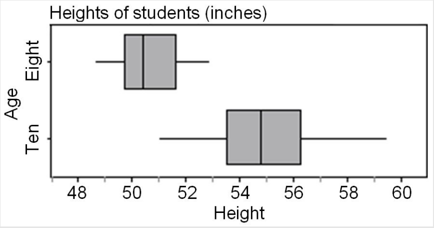
Summarize and describe distributions "How many text messages do middle school students send in a typical day?" Data obtained from a sample of students may have a distribution with a few very large values, showing a "long tail" in the direction of the larger values. Students realize that the mean may not represent the largest cluster of data points, and that the median is a more useful measure of center. In like fashion, the IQR is a more useful measure of spread, giving the spread of the middle 50% of the data points.

[6.SP.3] Recognize that a measure of center for a numerical data set summarizes all of its values with a single number, while a measure of variation describes how its values vary with a single number.

• "Box plot" is also sometimes written "boxplot." Because of the different methods for computing quartiles and other different conventions, there are different kinds of box plots in use. Box plots created from the five-number summary do not show points detached from the remainder of the diagram. However, box plots generated with statistical software may display these features.

[6.SP.4] Display numerical data in plots on a number line, including dot plots, histograms, and box plots.

Comparing distributions with box plots

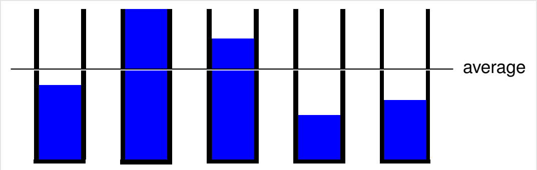


In Grade 6, box plots can be used to analyze the data from Example 2 of the Measurement Data Progression. Sixth graders can give more precise answers in terms of center and spread to questions asked at earlier grades. "Describe the key differences between the heights of these two age groups. What would you choose as the typical height of an eight-year-old? A ten-year-old? What would you say is the typical number of inches of growth from age eight to age ten?"

[6.NS.2] Fluently divide multi-digit numbers using the standard algorithm.

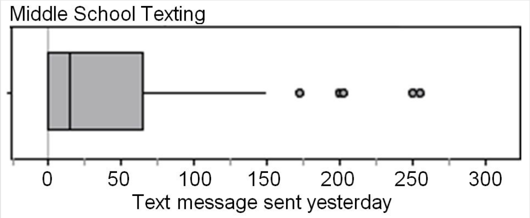
[6.NS.3] Fluently add, subtract, multiply, and divide multi-digit decimals using the standard algorithm for each operation.

Average as a "leveling out"



As mentioned in the Grades 2-5 Measurement Data Progression, students in Grade 5 might find the amount of liquid each cylinder would contain if the total amount in all the cylinders were redistributed equally. In Grade 6, students are able to view the amount in each cylinder after redistribution as equal to the mean of the five original amounts.

Middle School Texting

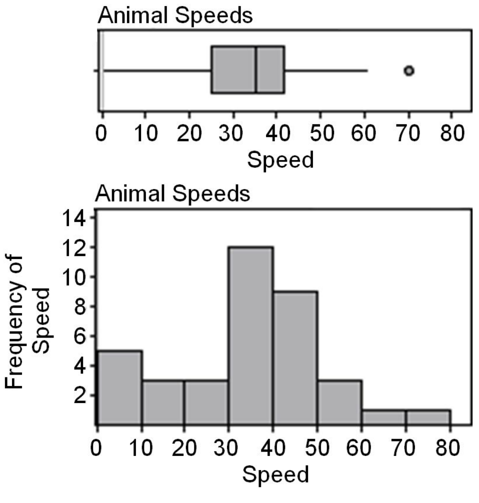


The 37 animal speeds shown in the margin can be used to illustrate summarizing a distribution.[6.SP.5a–c] According to the source, "Most of the following measurements are for maximum speeds over approximate quarter-mile distances. Exceptions—which are included to give a wide range of animals—are the lion and elephant, whose speeds were clocked in the act of charging; the whippet, which was timed over a 200-yard course; the cheetah over a 100-yard distance; humans for a 15-yard segment of a 100-yard run; and the black mamba snake, six-lined race runner, spider, giant tortoise, three-toed sloth, . . . , which were measured over various small distances." Understanding that it is difficult to measure speeds of wild animals, does this description raise any questions about whether or not this is a fair comparison of the speeds?

Moving ahead with the analysis, students will notice that the distribution is not symmetric, but the lack of symmetry is mild. It is most appropriate to measure center with the median of 35 mph and spread with the IQR of $42 - 25 = 17$. That makes the cheetah an outlier with respect to speed, but notice again the description of how this speed was measured. If the garden snail with a speed of 0.03 mph is added to the data set, then cheetah is no longer considered an outlier. Why is that?

Because the lack of symmetry is not severe, the mean (32.15 mph) is close to the median and the MAD (12.56 mph) is a reasonable measure of typical variation from the mean, as about 57% of the data values lie within one MAD of the mean, an interval from about 19.6 mph to 44.7 mph.

Box plot and histogram of 37 animal speeds



Note that the isolated point (the extreme value of 70 mph) has been generated by the software used to produce the box plot. The mild lack of symmetry can be seen in the box plot in the median (slightly off-center in the box) and in the slightly different lengths of the "whiskers." The geometric shape made by the histogram also shows mild lack of symmetry.

6.SP.5a Summarize numerical data sets in relation to their context, such as by:

a Reporting the number of observations.

6.SP.5b Summarize numerical data sets in relation to their context, such as by:

b Describing the nature of the attribute under investigation, including how it was measured and its units of measurement.

6.SP.5c Summarize numerical data sets in relation to their context, such as by:

c Giving quantitative measures of center (median and/or mean) and variability (interquartile range and/or mean absolute deviation), as well as describing any overall pattern and any striking deviations from the overall pattern with reference to the context in which the data were gathered.

Table of 37 animal speeds

Animal	Speed (mph)
Cheetah	70.00
Pronghorn antelope	61.00
Lion	50.00
Thomson's gazelle	50.00
Wildebeest	50.00
Quarter horse	47.50
Cape hunting dog	45.00
Elk	45.00
Coyote	43.00
Gray fox	42.00
Hyena	40.00
Ostrich	40.00
Zebra	40.00
Mongolian wild ass	40.00
Greyhound	39.35
Whippet	35.50
Jackal	35.00
Mule deer	35.00
Rabbit (domestic)	35.00
Giraffe	32.00
Reindeer	32.00
Cat (domestic)	30.00
Kangaroo	30.00
Grizzly bear	30.00
Wart hog	30.00
White-tailed deer	30.00
Human	27.89
Elephant	25.00
Black mamba snake	20.00
Six-lined race runner	18.00
Squirrel	12.00
Pig (domestic)	11.00
Chicken	9.00
House mouse	8.00
Spider (Tegenearia atrica)	1.17
Giant tortoise	0.17
Three-toed sloth	0.15

Source: `factmonster.com/ipka/A0004737.html`

Grade 7

Chance processes and probability models In Grade 7, students build their understanding of probability on a relative frequency view of the subject, examining the proportion of "successes" in a chance process—one involving repeated observations of random outcomes of a given event, such as a series of coin tosses. "What is my chance of getting the correct answer to the next multiple choice question?" is not a probability question in the relative frequency sense. "What is my chance of getting the correct answer to the next multiple choice question *if I make a random guess among the four choices?*" is a probability question because the student could set up an experiment of multiple trials to approximate the relative frequency of the outcome.• And two students doing the same experiment will get nearly the same approximation. These important points are often overlooked in discussions of probability.[7.SP.5]

• Note the connection with MP6. Including the stipulation "if I make a random guess among the four choices" makes the question precise enough to be answered with the methods discussed for this grade.

[7.SP.5] Understand that the probability of a chance event is a number between 0 and 1 that expresses the likelihood of the event occurring. Larger numbers indicate greater likelihood. A probability near 0 indicates an unlikely event, a probability around 1/2 indicates an event that is neither unlikely nor likely, and a probability near 1 indicates a likely event.

Students begin by relating probability to the long-run (more than five or ten trials) relative frequency of a chance event, using coins, number cubes, cards, spinners, bead bags, and so on. Hands-on activities with students collecting the data on probability experiments are critically important, but once the connection between observed relative frequency and theoretical probability is clear, they can move to simulating probability experiments via technology (graphing calculators or computers).

It must be understood that the connection between relative frequency and probability goes two ways. If you know the structure of the generating mechanism (e.g., a bag with known numbers of red and white chips), you can anticipate the relative frequencies of a series of random selections (with replacement) from the bag. If you do not know the structure (e.g., the bag has unknown numbers of red and white chips), you can approximate it by making a series of random selections and recording the relative frequencies.[7.SP.6] This simple idea, obvious to the experienced, is essential and not obvious at all to the novice.• The first type of situation, in which the structure is known, leads to "probability"; the second, in which the structure is unknown, leads to "statistics."

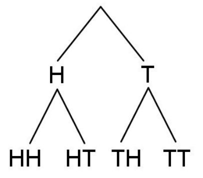
[7.SP.6] Approximate the probability of a chance event by collecting data on the chance process that produces it and observing its long-run relative frequency, and predict the approximate relative frequency given the probability.

• Examples of student strategies for generalizing from the relative frequency in the simplest case (one sample) to the relative frequency in the whole population are given in the Ratio and Proportional Relationship Progression, p. 11.

Different representations of a sample space

HH
HT
TH
TT

	H	T
H	HH	HT
T	TH	TT



All the possible outcomes of the toss of two coins can be represented as an organized list, table, or tree diagram. The sample space becomes a probability model when a probability for each simple event is specified.

A *probability model* provides a probability for each possible non-overlapping outcome for a chance process so that the total probability over all such outcomes is unity. The collection of all possible individual outcomes is known as the *sample space* for the model. For example, the sample space for the toss of two coins (fair or not) is often written as {TT, HT, TH, HH}. The probabilities of the model can be either *theoretical* (based on the structure of the process and its outcomes) or *empirical* (based on observed data generated by the process). In the toss of two balanced coins, the four outcomes of the sample space are given equal theoretical probabilities of $\frac{1}{4}$ because of the symmetry of the process—because the coins are balanced, an outcome of heads is just as likely as an outcome of tails. Randomly selecting a name from a list of ten students also leads to equally likely outcomes with probability 0.10 that a given student's name will be selected.[7.SP.7a] If there are exactly four seventh graders on the list, the chance of selecting a seventh grader's name is 0.40. On the other hand, the probability of a tossed thumbtack landing point up is not necessarily $\frac{1}{2}$ just because there are two possible outcomes; these outcomes may not be equally likely and an empirical answer must be found be tossing the tack and collecting data.[7.SP.7b]

The product rule for counting outcomes for chance events should be used in finite situations like tossing two or three coins or rolling two number cubes. There is no need to go to more formal rules for permutations and combinations at this level. Students should gain experience in the use of diagrams, especially trees and tables, as the basis for organized counting of possible outcomes from chance processes.[7.SP.8] For example, the 36 equally likely (theoretical probability) outcomes from the toss of a pair of number cubes are most easily listed on a two-way table. An archived table of census data can be used to approximate the (empirical) probability that a randomly selected Florida resident will be Hispanic.

After the basics of probability are understood, students should experience setting up a model and using simulation (by hand or with technology) to collect data and estimate probabilities for a real situation that is sufficiently complex that the theoretical probabilities are not obvious. For example, suppose, over many years of records, a river generates a spring flood about 40% of the time. Based on these records, what is the chance that it will flood for at least three years in a row sometime during the next five years?[7.SP.8c]

Random sampling In earlier grades students have been using data, both categorical and measurement, to answer simple statistical questions, but have paid little attention to how the data were selected. A primary focus for Grade 7 is the process of selecting a random sample, and the value of doing so. If three students are to be selected from the class for a special project, students recognize that a fair way to make the selection is to put all the student names in a box, mix them up, and draw out three names "at random." Individual students realize that they may not get selected, but that each student has the same chance of being selected. In other words, random sampling is a fair way to select a subset (a sample) of the set of interest (the population). A statistic computed from a random sample, such as the mean of the sample, can be used as an estimate of that same characteristic of the population from which the sample was selected. This estimate must be viewed with some degree of caution because of the variability in both the population and sample data. A basic tenet of statistical reasoning, then, is that random sampling allows results from a sample to be generalized to a much larger body of data, namely, the population from which the sample was selected.[7.SP.1]

"What proportion of students in the seventh grade of your school

[7.SP.7a] Develop a probability model and use it to find probabilities of events. Compare probabilities from a model to observed frequencies; if the agreement is not good, explain possible sources of the discrepancy.

a Develop a uniform probability model by assigning equal probability to all outcomes, and use the model to determine probabilities of events.

[7.SP.7b] Develop a probability model and use it to find probabilities of events. Compare probabilities from a model to observed frequencies; if the agreement is not good, explain possible sources of the discrepancy.

b Develop a probability model (which may not be uniform) by observing frequencies in data generated from a chance process.

[7.SP.8] Find probabilities of compound events using organized lists, tables, tree diagrams, and simulation.

[7.SP.8c] Find probabilities of compound events using organized lists, tables, tree diagrams, and simulation.

c Design and use a simulation to generate frequencies for compound events.

[7.SP.1] Understand that statistics can be used to gain information about a population by examining a sample of the population; generalizations about a population from a sample are valid only if the sample is representative of that population. Understand that random sampling tends to produce representative samples and support valid inferences.

choose football as their favorite sport?" Students realize that they do not have the time and energy to interview all seventh graders, so the next best way to get an answer is to select a random sample of seventh graders and interview them on this issue. The sample proportion is the best estimate of the population proportion, but students realize that the two are not the same and a different sample will give a slightly different estimate. In short, students realize that conclusions drawn from random samples generalize beyond the sample to the population from which the sample was selected, but a sample *statistic* is only an estimate of a corresponding population *parameter* and there will be some discrepancy between the two. Understanding variability in sampling allows the investigator to gauge the expected size of that discrepancy.

The variability in samples can be studied by means of simulation.[7.SP.2] Students are to take a random sample of 50 seventh graders from a large population of seventh graders to estimate the proportion having football as their favorite sport. Suppose, for the moment, that the true proportion is 60%, or 0.60. How much variation can be expected among the sample proportions? The scenario of selecting samples from this population can be simulated by constructing a "population" that has 60% red chips and 40% blue chips, taking a sample of 50 chips from that population, recording the number of red chips, replacing the sample in the population, and repeating the sampling process. (This can be done by hand or with the aid of technology, or by a combination of the two.) Record the proportion of red chips in each sample and plot the results.

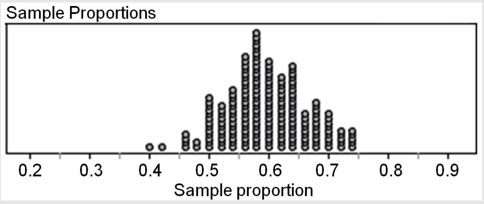
The dot plots in the margin shows results for 200 such random samples of size 50 each. Note that the sample proportions pile up around 0.60, but it is not too rare to see a sample proportion down around 0.45 or up around .0.75. Thus, we might expect a variation of close to 15 percentage points in either direction. Interestingly, about that same amount of variation persists for true proportions of 50% and 40%, as shown in the dot plots.

Students can now reason that random samples of size 50 are likely to produce sample proportions that are within about 15 percentage points of the true population value. They should now conjecture as to what will happen of the sample size is doubled or halved, and then check out the conjectures with further simulations. Why are sample sizes in public opinion polls generally around 1000 or more, rather than as small as 50?

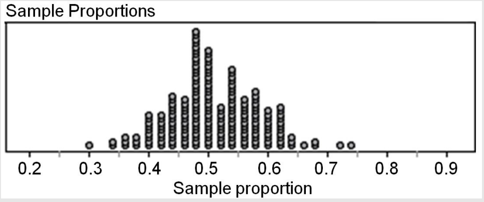
Informal comparative inference To estimate a population mean or median, the best practice is to select a random sample from that population and use the sample mean or median as the estimate, just as with proportions. But, many of the practical problems dealing with measures of center are comparative in nature, as in comparing average scores on the first and second exam or comparing average salaries between female and male employees of a firm. Such

[7.SP.2] Use data from a random sample to draw inferences about a population with an unknown characteristic of interest. Generate multiple samples (or simulated samples) of the same size to gauge the variation in estimates or predictions.

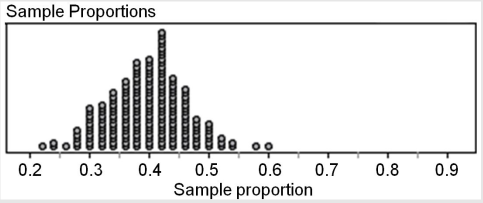
Results of simulations



Proportions of red chips in 200 random samples of size 50 from a population in which 60% of the chips are red.



Proportions of red chips in 200 random samples of size 50 from a population in which 50% of the chips are red.



Proportions of red chips in 200 random samples of size 50 from a population in which 40% of the chips are red.

comparisons may involve making conjectures about population parameters and constructing arguments based on data to support the conjectures (MP3).

If all measurements in a population are known, no sampling is necessary and data comparisons involve the calculated measures of center. Even then, students should consider variability.[7.SP.3] The figures in the margin show the female life expectancies for countries of Africa and Europe. It is clear that Europe tends to have the higher life expectancies and a much higher median, but some African countries are comparable to some of those in Europe. The mean and MAD for Africa are 53.6 and 9.5 years, respectively, whereas those for Europe are 79.3 and 2.8 years. In Africa, it would not be rare to see a country in which female life expectancy is about ten years away from the mean for the continent, but in Europe the life expectancy in most countries is within three years of the mean.

For random samples, students should understand that medians and means computed from samples will vary from sample to sample and that making informed decisions based on such sample statistics requires some knowledge of the amount of variation to expect. Just as for proportions, a good way to gain this knowledge is through simulation, beginning with a population of known structure.

The following examples are based on data compiled from nearly 200 middle school students in the Washington, DC area participating in the Census at Schools Project. Responses to the question, "How many hours per week do you usually spend on homework?," from a random sample of 10 female students and another of 10 male students from this population gave the results plotted in the margin.

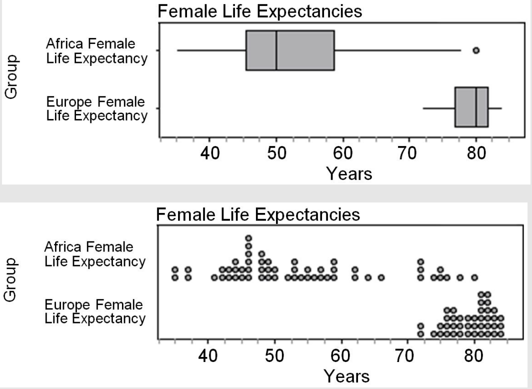
Females have a slightly higher median, but students should realize that there is too much variation in the sample data to conclude that, in this population, females have a higher median homework time. An idea of how much variation to expect in samples of size 10 is needed.

Simulation to the rescue! Students can take multiple samples of size 10 from the Census of Schools data to see how much the sample *medians* themselves tend to vary.[7.SP.4] The sample medians for 100 random samples of size 10 each, with 100 samples of males and 100 samples of females, is shown in the margin. This plot shows that the sample medians vary much less than the homework hours themselves and provides more convincing evidence that the female median homework hours is larger than that for males. Half of the female sample medians are within one hour of 4 while half of the male sample medians are within half hour of 3, although there is still overlap between the two groups.

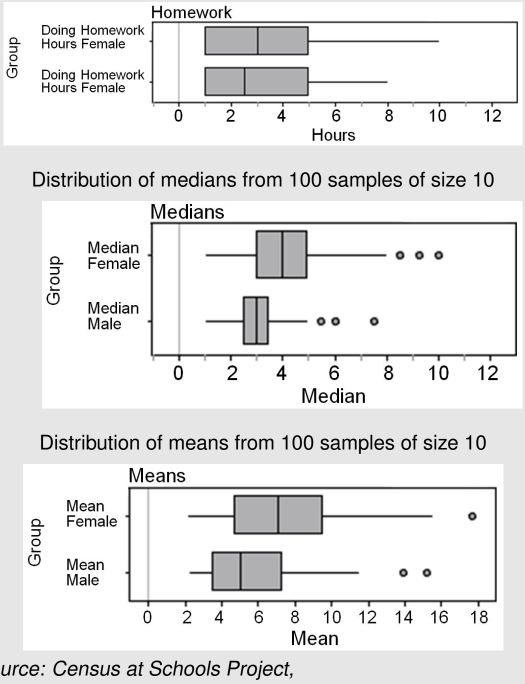
A similar analysis based on sample means gave the results seen in the margin. Here, the overlap of the two distributions is more severe and the evidence weaker for declaring that the females have higher mean study hours than males.

[7.SP.3] Informally assess the degree of visual overlap of two numerical data distributions with similar variabilities, measuring the difference between the centers by expressing it as a multiple of a measure of variability.

Female life expectancies in African and European countries



Hours spent on homework per week



Two random samples of size 10

Distribution of medians from 100 samples of size 10

Distribution of means from 100 samples of size 10

Source: Census at Schools Project, `amstat.org/censusatschool/`

[7.SP.4] Use measures of center and measures of variability for numerical data from random samples to draw informal comparative inferences about two populations.

Grade 8

Investigating patterns of association in bivariate data Students now have enough experience with coordinate geometry and linear functions[8.F.3,8.F.4,8.F.5] to plot bivariate data as points on a plane and to make use of the equation of a line in analyzing the relationship between two paired variables. They build statistical models to explore the relationship between two variables (MP4); looking for and making use of structure to describe possible association in bivariate data (MP7).

Working with paired measurement variables that might be associated linearly or in a more subtle fashion, students construct a scatter plot, describing the pattern in terms of clusters, gaps, and unusual data points (much as in the univariate situation). Then, they look for an overall positive or negative trend in the cloud of points, a linear or nonlinear (curved) pattern, and strong or weak association between the two variables, using these terms in describing the nature of the observed association between the variables.[8.SP.1]

For a data showing a linear pattern, students sketch a line through the "center" of the cloud of points that captures the essential nature of the trend, at first by use of an informal fitting procedure, perhaps as informal as laying a stick of spaghetti on the plot. How well the line "fits" the cloud of points is judged by how closely the points are packed around the line, considering that one or more outliers might have tremendous influence on the positioning of the line.[8.SP.2]

After a line is fit through the data, the slope of the line is approximated and interpreted as a rate of change, in the context of the problem.[8.F.4] The slope has important practical interpretations for most statistical investigations of this type (MP2). On the Exam 1 versus Exam 2 plot, what does the slope of 0.6 tell you about the relationship between these two sets of scores? Which students tend to do better on the second exam and which tend to do worse?[8.SP.3] Note that the negative linear trend in mammal life spans versus speed is due entirely to three long-lived, slow animals (hippo, elephant, and grizzly bear) and one short-lived, fast one (cheetah). Students with good geometry skills might explain why it would be unreasonable to expect that alligator lengths and weights would be linearly related.

Building on experience with decimals and percent, and the ideas of association between measurement variables, students now take a more careful look at possible association between categorical variables.[8.SP.4] "Is there a difference between sixth graders and eighth graders with regard to their preference for rock, rap, or country music?" Data from a random sample of sixth graders and another random sample of eighth graders are summarized by frequency counts in each cell in a two-way table of preferred music type by grade. The proportions of favored music type for the sixth graders are then compared to the proportions for eighth graders. If the two proportions for each music type are about the same, there is little or no

[8.F.3] Interpret the equation $y = mx + b$ as defining a linear function, whose graph is a straight line; give examples of functions that are not linear.

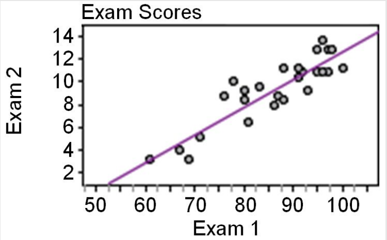
[8.F.4] Construct a function to model a linear relationship between two quantities. Determine the rate of change and initial value of the function from a description of a relationship or from two (x, y) values, including reading these from a table or from a graph. Interpret the rate of change and initial value of a linear function in terms of the situation it models, and in terms of its graph or a table of values.

[8.F.5] Describe qualitatively the functional relationship between two quantities by analyzing a graph (e.g., where the function is increasing or decreasing, linear or nonlinear). Sketch a graph that exhibits the qualitative features of a function that has been described verbally.

[8.SP.1] Construct and interpret scatter plots for bivariate measurement data to investigate patterns of association between two quantities. Describe patterns such as clustering, outliers, positive or negative association, linear association, and nonlinear association.

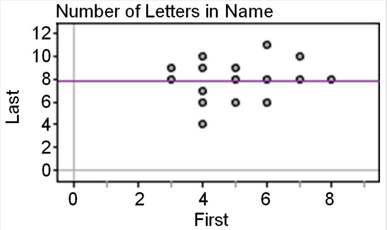
[8.SP.2] Know that straight lines are widely used to model relationships between two quantitative variables. For scatter plots that suggest a linear association, informally fit a straight line, and informally assess the model fit by judging the closeness of the data points to the line.

Scores on Exam 1 and Exam 2



The least squares line fitted to the points has a positive slope and the points are closely clustered about the line, thus, the scores said to show strong positive association. Students with high scores on one exam tend to have high scores on the other. Students with low scores on one exam tend to have low scores on the other.

Letters in first and last names of students



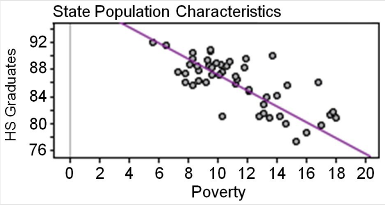
The line fitted to the points is horizontal. The number of letters in a student's first name shows no association with the number of letters in a student's last name.

[8.SP.3] Use the equation of a linear model to solve problems in the context of bivariate measurement data, interpreting the slope and intercept.

association between the grade and music preference because both grades have about the same preferences. If the two proportions differ, there is some evidence of association because grade level seems to make a difference in music preferences. The nature of the association should then be described in more detail.

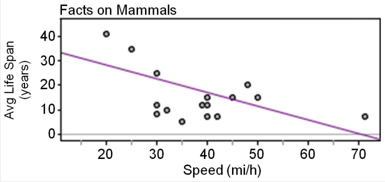
The table in the margin shows percentages of U.S. residents who have health risks due to obesity, by age category. Students should be able to explain what the cell percentages represent and provide a clear description of the nature of the association between the variables *obesity risk* and *age*. Can you tell, from this table alone, what percentage of those over the age of 18 are at risk from obesity? Such questions provide a practical mechanism for reinforcing the need for clear understanding of proportions and percentages.

High school graduation and poverty percentages for states



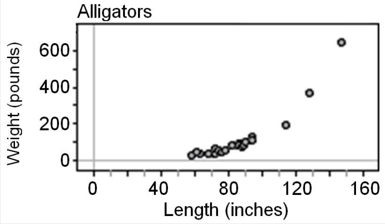
The line fitted to the data has a negative slope and data points are not all tightly clustered about the line. The percentage of a state's population in poverty shows a moderate negative association with the percentage of a state's high school graduates.

Average life span and speeds of mammals



The negative trend is due to a few outliers. This as can be seen by examining the effect of removing those points.

Weight versus length of Florida alligators



Source: http://www.factmonster.com/ipka/A0004737.html

A nonlinear association.

Table schemes for comparing frequencies and row proportions

	Rock	Rap	Country	Total
6th graders	a	b	c	d
8th graders	e	f	g	h

	Rock	Rap	Country	Total
6th graders	a/d	b/d	c/d	d
8th graders	e/h	f/h	g/h	h

Each letter represents a frequency count.

Obesity risk percentages

Age Category	Obesity		
	Not At Risk	At Risk	Row Total
Age 18 to 24	57.3	42.7	100
Age 25 to 44	38.6	61.4	100

Source: Behavioral Risk Factor Surveillance System of the Center for Disease Control

Where the Statistics and Probability Progression is heading

In high school, students build on their experience from the middle grades with data exploration and summarization, randomization as the basis of statistical inference, and simulation as a tool to understand statistical methods.

Just as Grade 6 students deepen the understanding of univariate data initially developed in elementary school, high school students deepen their understanding of bivariate data, initially developed in middle school. Strong and weak association is expressed more precisely in terms of correlation coefficients, and students become familiar with an expanded array of functions in high school that they use in modeling association between two variables.

They gain further familiarity with probability distributions generated by theory or data, and use these distributions to build an empirical understanding of the normal distribution, which is the main distribution used in measuring sampling error. For statistical methods related to the normal distribution, variation from the mean is measured by standard deviation.

Students extend their knowledge of probability, learning about conditional probability, and using probability distributions to solve problems involving expected value.