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

21 April 2012

High School Statistics and Probability★

Overview

In high school, students build on knowledge and experience described in the 6-8 Statistics and Probability Progression. They develop a more formal and precise understanding of statistical inference, which requires a deeper understanding of probability. Students learn that formal inference procedures are designed for studies in which the sampling or assignment of treatments was random, and these procedures may not be informative when analyzing non-randomized studies, often called observational studies. For example, a random selection of 100 students from your school will allow you to draw some conclusion about all the students in the school, whereas taking your class as a sample will not allow that generalization.

Probability is still viewed as long-run relative frequency but the emphasis now shifts to conditional probability and independence, and basic rules for calculating probabilities of compound events. In the plus standards• are the Multiplication Rule, probability distributions and their expected values. Probability is presented as an essential tool for decision-making in a world of uncertainty.

In the high school Standards, individual modeling standards are indicated by a star symbol (★). Because of its strong connection with modeling, the domain of Statistics and Probability is starred, indicating that all of its standards are modeling standards.

• Additional mathematics that students should learn in order to take advanced courses such as calculus, advanced statistics, or discrete mathematics is indicated by (+).

Interpreting categorical and quantitative data

Summarize, represent, and interpret data on a single count or measurement variable Students build on the understanding of key ideas for describing distributions—shape, center, and spread—described in the Grades 6-8 Statistics and Probability Progression. This enhanced understanding allows them to give more precise answers to deeper questions, often involving comparisons of data sets. Students use shape and the question(s) to be answered to decide on the median or mean as the more appropriate measure of center and to justify their choice through statistical reasoning. They also add a key measure of variation to their toolkits.

In connection with the mean as a measure of center, the *standard deviation* is introduced as a measure of variation. The standard deviation is based on the squared deviations from the mean, but involves much the same principle as the mean absolute deviation (MAD) that students learned about in Grades 6-8. Students should see that the standard deviation is the appropriate measure of spread for data distributions that are approximately normal in shape, as the standard deviation then has a clear interpretation related to relative frequency.

The margin shows two ways of comparing height data for males and females in the 20-29 age group. Both involve plotting the data or data summaries (box plots or histograms) on the same scale, resulting in what are called *parallel* (or *side-by-side*) *box plots* and *parallel histograms*.[S-ID.1] The parallel box plots show an obvious difference in the medians and the IQRs for the two groups; the medians for males and females are, respectively, 71 inches and 65 inches, while the IQRs are 4 inches and 5 inches. Thus, male heights center at a higher value but are slightly more variable.

The parallel histograms show the distributions of heights to be mound shaped and fairly symmetrical (approximately normal) in shape. Therefore, the data can be succinctly described using the mean and standard deviation. Heights for males and females have means of 70.4 inches and 64.7 inches, respectively, and standard deviations of 3.0 inches and 2.6 inches. Students should be able to sketch each distribution and answer questions about it just from knowledge of these three facts (shape, center, and spread). For either group, about 68% of the data values will be within one standard deviation of the mean.[S-ID.2,S-ID.3] They should also observe that the two measures of center, median and mean, tend to be close to each other for symmetric distributions.

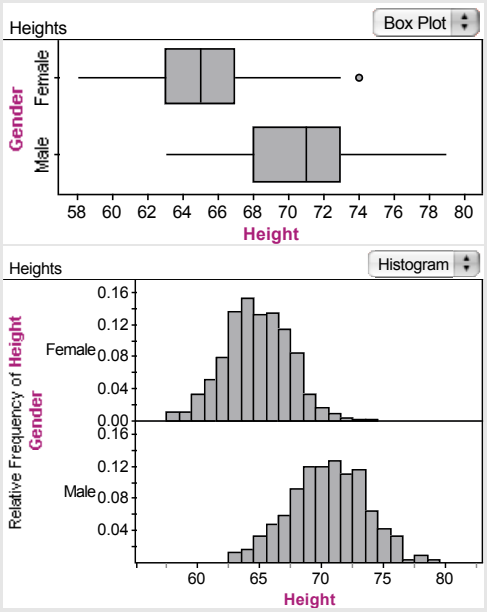
Data on heights of adults are available for anyone to look up. But how can we answer questions about standardized test scores when individual scores are not released and only a description of the distribution of scores is given? Students should now realize that we can do this only because such standardized scores generally have

Shapes of distributions

In Grade 6, students began to distinguish among distributions that were skewed or approximately symmetric. In high school, they distinguish among the latter, according to whether or not they are approximately normal, that is approximately bell-shaped.

S-ID.1 Represent data with plots on the real number line (dot plots, histograms, and box plots).

Comparing heights of males and females



Heights of U.S. males and females in the 20–29 age group. Source: *U.S. Census Bureau, Statistical Abstract of the United States: 2009*, Table 201.

S-ID.2 Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, standard deviation) of two or more different data sets.

S-ID.3 Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers).

a distribution that is mound-shaped and somewhat symmetric, i.e., approximately normal.• For example, SAT math scores for a recent year have a mean of 516 and a standard deviation of 116.• Thus, about 16% of the scores are above 632. In fact, students should be aware that technology now allows easy computation of any area under a normal curve. "If Alicia scored 680 on this SAT mathematics exam, what proportion of students taking the exam scored less than she scored?" (Answer: about 92%.)[S-ID.4]

Summarize, represent, and interpret data on two categorical and quantitative variables As with univariate data analysis, students now take a deeper look at bivariate data, using their knowledge of proportions to describe categorical associations and using their knowledge of functions to fit models to quantitative data.[MP7, MP4]

The table below shows statistics from the Center for Disease Control relating HIV risk to age groups. Students should be able to explain the meaning of a row or column total (marginal), a row or column percentage (conditional) or a "total" percentage (joint). They should realize that possible associations between age and HIV risk are best explained in terms of the row or column conditional percentages. Are the comparisons of percentages valid when the first age category is much smaller (in years) than the others?[S-ID.5]

HIV risk by age groups, in percent of population

	Age	18–24	25–44	45–64	Row Total
Not at risk	Row %	14.0	59.6	26.4	100.0
	Column %	35.0	51.7	27.2	
	Total %	5.6	23.6	10.5	39.6
At risk	Row %	17.1	36.5	46.4	100.0
	Column %	65.0	48.3	72.8	
	Total %	10.3	22.0	28.1	60.4
Column total	Row %	15.9	45.6	38.5	100.0
	Column %	100.0	100.0	100.0	100.0
	Total %	15.9	45.6	38.5	100.0

Source: Center for Disease Control,
`http://apps.nccd.cdc.gov/s_broker/WEATSQL.exe/weat/freq_year.hsql`

Students have seen scatter plots in Grade 8 and now extend that knowledge to fit mathematical models that capture key elements of the relationship between two variables and to explain what the model tells us about that relationship. Some of the data should come from science, as in the examples about cricket chirps and temperature, and tree growth and age, and some from other aspects of their everyday life, e.g., cost of pizza and calories per slice (p. 6).

If you have a keen ear and some crickets, can the cricket chirps help you predict the temperature? The margin shows data modeled in a scientific investigation of that phenomenon. In this situation, the variables have been identified as chirps per second and temperature in degrees Fahrenheit. The cloud of points in the scatter plot is essentially linear with a moderately strong positive relationship. It looks like there must be something other than random behavior in

• At this level, students are not expected to fit normal curves to data. (In fact, it is rather complicated to rescale data plots to be density plots and then find the best fitting curve.) Instead, the aim is to look for broad approximations, with application of the rather rough "empirical rule" (also called the 68%–95% Rule) for distributions that are somewhat bell-shaped. The better the bell, the better the approximation. Using such approximations is partial justification for the introduction of the standard deviation.

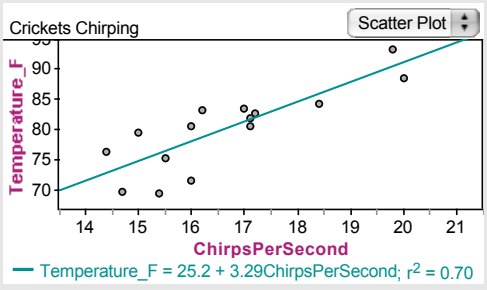
• See http://professionals.collegeboard.com/profdownload/2010-total-group-profile-report-cbs.pdf.

S-ID.4 Use the mean and standard deviation of a data set to fit it to a normal distribution and to estimate population percentages. Recognize that there are data sets for which such a procedure is not appropriate. Use calculators, spreadsheets, and tables to estimate areas under the normal curve.

MP7, MP4 Looking for patterns in tables and on scatter plots; modeling patterns in scatter plots with lines.

S-ID.5 Summarize categorical data for two categories in two-way frequency tables. Interpret relative frequencies in the context of the data (including joint, marginal, and conditional relative frequencies). Recognize possible associations and trends in the data.

Cricket chirps and temperature



Source: George W. Pierce, *The Songs of Insects*, Harvard University Press, 1949, pp. 12–21.

this association. A model has been formulated: The least squares regression line• has been fit by technology.[S-ID.6] The model is used to draw conclusions: The line estimates that, on average, each added chirp predicts an increase of about 3.29 degrees Fahrenheit.

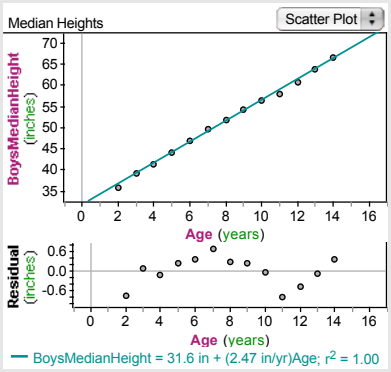
But, students must learn to take a careful look at scatter plots, as sometimes the "obvious" pattern does not tell the whole story, and can even be misleading. The margin shows the median heights of growing boys through the ages 2 to 14. The line (least squares regression line) with slope 2.47 inches per year of growth looks to be a perfect fit. [S-ID.6c] But, the *residuals*, the collection of differences between the corresponding coordinate on the least squares line and the actual data value for each age, reveal additional information. A plot of the residuals shows that growth does not proceed at a constant rate over those years.[S-ID.6b] What would be a better description of the growth pattern?

It is readily apparent to students, after a little experience with plotting bivariate data, that not all the world is linear. The figure below shows the diameters (in inches) of growing oak trees at various ages (in years). A careful look at the scatter plot reveals some curvature in the pattern,[S-ID.6a] which is more obvious in the residual plot, because the older and larger trees add to the diameter more slowly. Perhaps a curved model, such as a quadratic, will fit the data better than a line. The figure below shows that to be the case.

Would it be wise to extrapolate the quadratic model to 50-year-old trees? Perhaps a better (and simpler) model can be found by thinking in terms of cross-sectional area, rather than diameter, as the measure that might grow linearly with age.[S-ID.6a] Area is proportional to the square of the diameter, and the plot of diameter squared versus age in the margin does show remarkable linearity,[S-ID.6a] but there is always the possibility of a closer fit, that students familiar with cube root, exponential, and logarithmic functions[F-IF.7] could investigate. Students should be encouraged to think about the relationship between statistical models and the real world, and how knowledge of

• This term is used to identify the line in this Progression. Students will identify the line as the "line of best fit" obtained by technology and should not be required to use or learn "least squares regression line."

Median heights of boys



Source: National Health and Nutrition Examination Survey, 2002, `www.cdc.gov/nchs/about/major/nhanes/datatblelink.htm`.

S-ID.6 Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.

a Fit a function to the data; use functions fitted to data to solve problems in the context of the data.

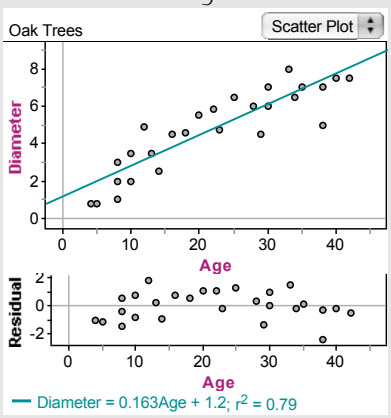
b Informally assess the fit of a function by plotting and analyzing residuals.

c Fit a linear function for a scatter plot that suggests a linear association.

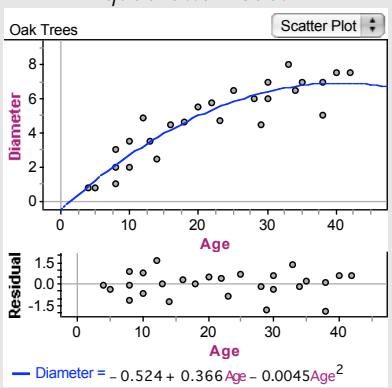
F-IF.7 Graph functions expressed symbolically and show key features of the graph, by hand in simple cases and using technology for more complicated cases.

Three iterations of the modeling cycle

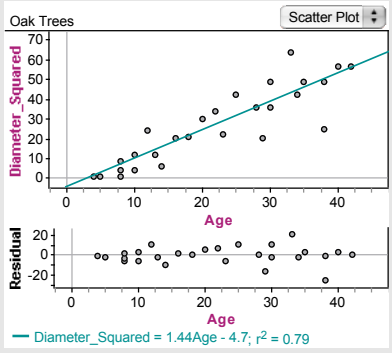
Linear model: Age vs diameter



A closer fit: Age vs diameter in a quadratic model



A simpler model: Age vs diameter squared



the context is essential to building good models.

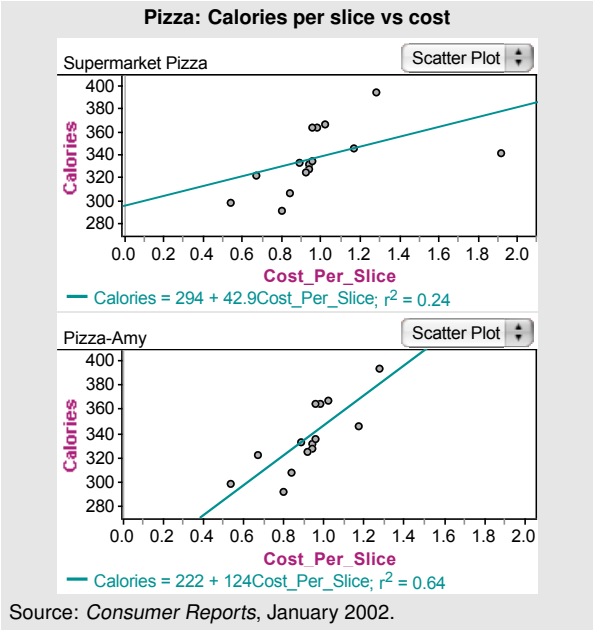
Interpret linear models Students understand that the process of fitting and interpreting models for discovering possible relationships between variables requires insight, good judgment and a careful look at a variety of options consistent with the questions being asked in the investigation.[MP6]

Suppose you want to see if there is a relationship between the cost per slice of supermarket pizzas and the calories per serving. The margin shows data for a sample of 15 such pizza brands, and a somewhat linear trend. A line fitted via technology might suggest that you should expect to see an increase of about 43 calories if you go from one brand to another that is one dollar more in price. But, the line does not appear to fit the data well and the correlation coefficient r (discussed below) is only about 0.5. Students will observe that there is one pizza that does not seem to fit the pattern of the others, the one with maximum cost. Why is it way out there? A check reveals that it is Amy's Organic Crust & Tomatoes, the only organic pizza in the sample. If the outlier (Amy's pizza) is removed and the discussion is narrowed to non-organic pizzas (as shown in the plot for pizzas other than Amy's), the relationship between calories and price is much stronger with an expected increase of 124 calories[S-ID.7] per extra dollar spent and a correlation coefficient of 0.8. Narrowing the question allows for a better interpretation of the slope of a line fitted to the data.[S-ID.8]

The *correlation coefficient* measures the "tightness" of the data points about a line fitted to data, with a limiting value of 1 (or -1) if all points lie precisely on a line of positive (or negative) slope. For the line fitted to cricket chirps and temperature (p. 4), the correlation is 0.84, and for the line fitted to boys' height (p. 5), it is about 1.0. However, the quadratic model for tree growth (p. 5) is non-linear, so the value of its correlation coefficient has no direct interpretation.[S-ID.8] (The square of the correlation coefficient, however, does have an interpretation for such models.)

In situations where the correlation coefficient of a line fitted to data is close to 1 or −1, the two variables in the situation are said to have a *high correlation*. Students must see that one of the most common misinterpretations of correlation is to think of it as a synonym for causation. A high correlation between two variables (suggesting a statistical association between the two) does *not* imply that one causes the other. It is not a cost increase that causes calories to increase in pizza, and it is not a calorie increase per se that causes cost to increase; the addition of other expensive ingredients cause both to increase simultaneously.[S-ID.9] Students should look for examples of correlation being interpreted as cause and sort out why that reasoning is incorrect (MP3). Examples may include medications versus disease symptoms and teacher pay or class size versus high school graduation rates. One good way of establishing cause

MP6 Reasoning abstractly but quantitatively in discovering possible associations between numerical variables.



Source: *Consumer Reports*, January 2002.

S-ID.7 Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.

S-ID.8 Compute (using technology) and interpret the correlation coefficient of a linear fit.

S-ID.9 Distinguish between correlation and causation.

is through the design and analysis of randomized experiments, and that subject comes up in the next section.

Making inferences and justifying conclusions

Understand and evaluate random processes underlying statistical experiments Students now move beyond analyzing data to making sound statistical decisions based on probability models. The reasoning process is as follows: develop a statistical question in the form of a hypothesis (supposition) about a population parameter; choose a probability model for collecting data relevant to that parameter; collect data; compare the results seen in the data with what is expected under the hypothesis. If the observed results are far away from what is expected and have a low probability of occurring under the hypothesis, then that hypothesis is called into question. In other words, the evidence against the hypothesis is weighed by probability.[S-IC.1]

S-IC.1 Understand statistics as a process for making inferences about population parameters based on a random sample from that population.

But, what is considered "low"? That determination is left to the investigator and the circumstances surrounding the decision to be made. Statistics and probability weigh the chances; the person in charge of the investigation makes the final choice. (This is much like other areas of life in which the teacher or physician weighs the evidence and provides your chances of passing a test or easing certain disease symptoms; you make the choice.)

Consider this example. You cannot seem to roll an even number with a certain number cube. The statistical question is, "Does this number cube favor odd numbers?" The hypothesis is, "This cube does not favor odd numbers," which is the same as saying that the proportion of odd numbers rolled, in the long run, is 0.5, or the probability of tossing an odd number with this cube is 0.5. Then, toss the cube and collect data on the observed number of odds. Suppose you get an odd number in each of the:

first two tosses, which has probability $\frac{1}{4} = 0.25$ under the hypothesis;

first three tosses, which has probability $\frac{1}{8} = 0.125$ under the hypothesis;

first four tosses, which has probability $\frac{1}{16} = 0.0625$ under the hypothesis;

first five tosses, which has probability $\frac{1}{32} = 0.03125$ under the hypothesis.

At what point will students begin to seriously doubt the hypothesis that the cube does not favor odd numbers? Students should experience a number of simple situations like this to gain an understanding of how decisions based on sample data are related to probability, and that this decision process does not guarantee a correct answer to the underlying statistical question.[S-IC.3]

S-IC.3 Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.

Make inferences and justify conclusions from sample surveys, experiments, and observational studies Once they see how probability intertwines with data collection and analysis, students use

this knowledge to make statistical inferences from data collected in sample surveys and in designed experiments, aided by simulation and the technology that affords it.[MP5, MP3]

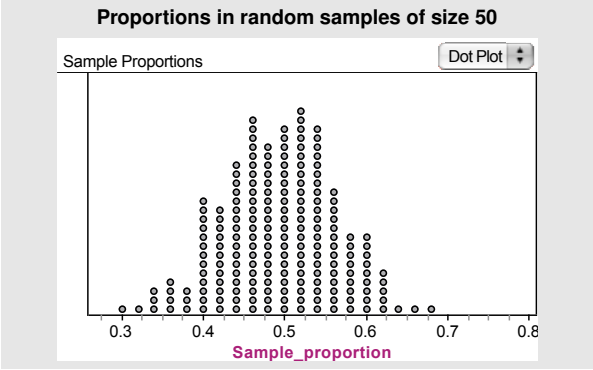
A *Time* magazine poll reported on the status of American women. One of the statements in the poll was "It is better for a family if the father works outside the home and the mother takes care of children." Fifty-one percent of the sampled women agreed with the statement while 57% of the sampled men agreed. A note on the polling methodology states that about 1600 men and 1800 women were randomly sampled in the poll and the margin of error was about two percentage points. What is the margin of error and how is it interpreted in this context? We'll come back to the *Time* poll after exploring this question further.

"Will 50% of the homeowners in your neighborhood agree to support a proposed new tax for schools?" A student attempts to answer this question by taking a random sample of 50 homeowners in her neighborhood and asking them if they support the tax. Twenty of the sampled homeowners say they will support the proposed tax, yielding a sample proportion of $\frac{20}{50} = 0.4$. That seems like bad news for the schools, but could the population proportion favoring the tax in this neighborhood still be 50%? The student knows that a second sample of 50 homeowners might produce a different sample proportion and wonders how much variation there might be among sample proportions for samples of size 50 if, in fact, 50% is the true population proportion. Having a graphing calculator available, she simulates this sampling situation by repeatedly drawing random samples of size 50 from a population of 50% ones and 50% zeros, calculating and plotting the proportion of ones observed in each sample. The result for 200 trials is displayed in the margin. The simulated values at or below the observed 0.4 number 25 out of 200, or $\frac{25}{200} = 0.125$. So, the chance of seeing a 40% or fewer favorable response in the sample even if the true proportion of such responses was 50% is not all that small, casting little doubt on 50% as a plausible population value.

Relating the components of this example to the statistical reasoning process, students see that the hypothesis is that the population parameter is 50% and the data are collected by a random sample. The observed sample proportion of 40% was found to be not so far from the 50% so as to cause serious doubt about the hypothesis. This lack of doubt was justified by simulating the sampling process many times and approximating the chance of a sample proportion being 40% or less under the hypothesis.[MP8]

Students now realize that there are other plausible values for the population proportion, besides 50%. The plot of the distribution of sample proportions in the margin is mound-shaped (approximately normal) and somewhat symmetric with a mean of about 0.49 (close to 0.50) and a standard deviation of about 0.07. From knowledge of the normal distribution,[S-ID.4] students knows that about 95% of the possible sample proportions that could be generated this way

[MP5, MP3] Using a variety of statistical tools to construct and defend logical arguments based on data.



[MP8] Observing regular patterns in distributions of sample statistics.

[S-ID.4] Use the mean and standard deviation of a data set to fit it to a normal distribution and to estimate population percentages. Recognize that there are data sets for which such a procedure is not appropriate. Use calculators, spreadsheets, and tables to estimate areas under the normal curve.

will fall within two standard deviations of the mean. This two-standard deviation distance is called the *margin of error* for the sample proportions. In this example with samples of size 50, the margin of error is $2 \cdot 0.07 = 0.14$.

Suppose the true population proportion is 0.60. The distribution of the sample proportions will still look much like the plot in the margin, but the center of the distribution will be at 0.60. In this case, the observed sample proportion 0.4 will not be within the margin of error. Reasoning this way leads the student to realize that any population proportion in the interval 0.40 ± 0.14 will result in the observed sample proportion of 0.40 being within the middle 95% of the distribution of sample proportions, for samples of size 50. Thus, the interval

observed sample proportion $\pm$ margin of error

includes the plausible values for the true population proportion in the sense that any of those populations would have produced the observed sample proportion within its middle 95% of possible outcomes. In other words, the student is confident that the proportion of homeowners in her neighborhood that will favor the tax is between 0.26 and 0.54.[S-IC.4] All of this depends on random sampling because the characteristics of distributions of sample statistics are predictable only if the sampling is random.

With regard to the *Time* poll on the status of women, the student now sees that the plausible proportions of men who agree with the statement lie between 55% and 59% while the plausible proportions of women who agree lie between 49% and 53%. What interesting conclusions might be drawn from this?[S-IC.6]

Students' understanding of random sampling as the key that allows the computation of margins of error in estimating a population quantity can now be extended to the random assignment of treatments to available units in an experiment. A clinical trial in medical research, for example, may have only 50 patients available for comparing two treatments for a disease. These 50 are the population, so to speak, and randomly assigning the treatments to the patients is the "fair" way to judge possible treatment differences, just as random sampling is a fair way to select a sample for estimating a population proportion.

There is little doubt that caffeine stimulates bodily activity, but how much does it take to produce a significant effect? This is a question that involves measuring the effect of two or more treatments and deciding if the different interventions have differing effects. To obtain a partial answer to the question on caffeine, it was decided to compare a treatment consisting of 200 mg of caffeine with a control of no caffeine in an experiment involving a finger tapping exercise.

Twenty male students were randomly assigned to one of two treatment groups of 10 students each, one group receiving 200 milligrams of caffeine and the other group no caffeine. Two hours later

S-IC.4 Use data from a sample survey to estimate a population mean or proportion; develop a margin of error through the use of simulation models for random sampling.

S-IC.6 Evaluate reports based on data.

Finger taps per minute in a caffeine experiment

	0 mg caffeine	200 mg caffeine
	242	246
	245	248
	244	250
	248	252
	247	248
	248	250
	242	246
	244	248
	246	245
	242	250
Mean	244.8	248.3

Source: Draper and Smith, *Applied Regression Analysis*, John Wiley and Sons, 1981

the students were given a finger tapping exercise. The response is the number of taps per minute, as shown in the table.

The plot of the finger tapping data shows that the two data sets tend to be somewhat symmetric and have no extreme data points (outliers) that would have undue influence on the analysis. The sample mean for each data set, then, is a suitable measure of center, and will be used as the statistic for comparing treatments.

The mean for the 200 mg data is 3.5 taps larger than that for the 0 mg data. In light of the variation in the data, is that enough to be confident that the 200 mg treatment truly results in more tapping activity than the 0 mg treatment? In other words, could this difference of 3.5 taps be explained simply by the randomization (the luck of the draw, so to speak) rather than any real difference in the treatments? An empirical answer to this question can be found by "re-randomizing" the two groups many times and studying the distribution of differences in sample means. If the observed difference of 3.5 occurs quite frequently, then we can safely say the difference could simply be due to the randomization process. If it does not occur frequently, then we have evidence to support the conclusion that the 200 mg treatment has increased mean finger tapping count.

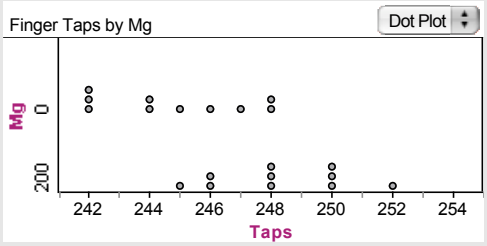
The re-randomizing can be accomplished by combining the data in the two columns, randomly splitting them into two different groups of ten, each representing 0 and 200 mg, and then calculating the difference between the sample means. This can be expedited with the use of technology.

The margin shows the differences produced in 400 re-randomizations of the data for 200 and 0 mg. The observed difference of 3.5 taps is equaled or exceeded only once out of 400 times. Because the observed difference is reproduced only 1 time in 400 trials, the data provide strong evidence that the control and the 200 mg treatment do, indeed, differ with respect to their mean finger tapping counts. In fact, we can conclude with little doubt that the caffeine is the *cause* of the increase in tapping because other possible factors should have been balanced out by the randomization.[S-IC.5] Students should be able to explain the reasoning in this decision and the nature of the error that may have been made.

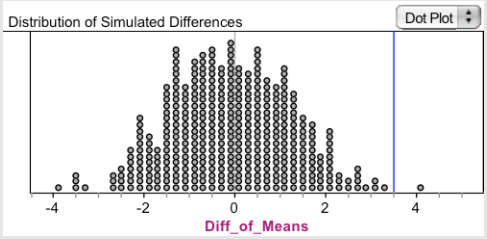
It must be emphasized repeatedly that the probabilistic reasoning underlying statistical inference is introduced into the study by way of random sampling in sample surveys and random assignment of treatments in experiments. No randomization, no such reasoning! Students will know, however, that randomization is not possible in many types of statistical investigations. Society will not condone the assigning of known harmful "treatments" (smoking, for example) to patients, so studies of the effects of smoking on health cannot be randomized experiments. Such studies must come from *observing* people who choose to smoke, as compared to those who do not, and are, therefore, called *observational studies*. The oak tree study (p. 5) and the pizza study (p. 6) are both observational studies.

Surveys of samples to estimate population parameters, random-

Plot of finger tapping data



Differences in re-randomized means for finger tapping data



S-IC.5 Use data from a randomized experiment to compare two treatments; use simulations to decide if differences between parameters are significant.

ized experiments to compare treatments and show cause, and observational studies to indicate possible associations among variables are the three main methods of data production in statistical studies. Students should understand the distinctions among these three and practice perceiving them in studies that are reported in the media, deciding if appropriate inferences seem to have been drawn.[S-IC.3]

S-IC.3 Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.

Conditional probability and the rules of probability

In Grades 7 and 8, students encountered the development of basic probability, including chance processes, probability models, and sample spaces. In high school, the relative frequency approach to probability is extended to conditional probability and independence, rules of probability and their use in finding probabilities of compound events, and the use of probability distributions to solve problems involving expected value. As seen in the making inferences section above, there is a strong connection between statistics and probability. This will be seen again in this section with the use of data in selecting values for probability models.

Understand independence and conditional probability and use them to interpret data In developing their understanding of conditional probability and independence, students should see two types of problems, one in which the uniform probabilities attached to outcomes leads to independence and one in which it does not. For example, suppose a student is randomly guessing the answers to all four true–false questions on a quiz. The outcomes in the sample space can be arranged as shown in the margin.[S-CP.1] Probabilities assigned to these outcomes should be equal because random guessing implies that no one outcome should be any more likely than another.

Possible outcomes: Guessing on four true–false questions

Number correct	Outcomes	Number correct	Outcomes	Number correct	Outcomes
4	CCCC	2	CCII	1	CIII
3	ICCC	2	CICI	1	ICII
3	CICC	2	CIIC	1	IICI
3	CCIC	2	ICCI	1	IIIC
3	CCCI	2	ICIC	0	IIII
		2	IICC		

C indicates a correct answer; I indicates an incorrect answer.

S-CP.1 Describe events as subsets of a sample space (the set of outcomes) using characteristics (or categories) of the outcomes, or as unions, intersections, or complements of other events ("or," "and," "not").

MP6 Attend to precision. "Two correct answers" may be interpreted as "at least two" or as "exactly two."

By simply counting equally likely outcomes,

$$\text{P(exactly}^{\text{MP6}}\text{ two correct answers)} = \frac{6}{16}$$

and

$$\begin{aligned}\text{P(at least one correct answer)} &= \frac{15}{16} \\ &= 1 - \text{P(no correct answers)}.\end{aligned}$$

Likewise,

$$\begin{aligned}\text{P(C on first question)} &= \frac{1}{2} \\ &= \text{P(C on second question)}\end{aligned}$$

as should seem intuitively reasonable. Now,

$$\begin{aligned}P[\text{(C on first question) and (C on second question)}] &= \frac{4}{16} \\ &= \frac{1}{4} \\ &= \frac{1}{2} \cdot \frac{1}{2},\end{aligned}$$

which shows that the two events (C on first question) and (C on second question) are independent, by the definition of independence.• This, too, should seem intuitively reasonable to students because the random guess on the second question should not have been influenced by the random guess on the first.

Students may contrast the quiz scenario above with the scenario of choosing at random two students to be leaders of a five-person working group consisting of three girls (*A*pril, *B*riana, and *C*yndi) and two boys (*D*aniel and *E*rnesto). The first name chosen indicates the discussion leader and the second the recorder, so order of selection is important. The 20 outcomes are displayed in the margin.

Here, the probability of selecting two girls is:

$$\begin{aligned} \text{P(two girls selected)} &= \frac{6}{20} \\ &= \frac{3}{10} \end{aligned}$$

whereas

$$\begin{aligned} \text{P(girl selected on first draw)} &= \frac{12}{20} \\ &= \frac{3}{5} \\ &= \text{P(girl selected on second draw)}. \end{aligned}$$

Because $\frac{3}{5} \cdot \frac{3}{5} \neq \frac{3}{10}$, these two events are not independent. The selection of the second person does depend on the selection of the first when the same person cannot be selected twice.

Another way of viewing independence is to consider the conditional probability of an event A given an event B, P(A|B), as the probability of A in the sample space restricted to just those outcomes that constitute B. In the table of outcomes for guessing on the true-false questions,

$$\begin{aligned} \text{P(C on second question} \mid \text{C on first question)} &= \frac{4}{8} \\ &= \frac{1}{2} \\ &= \text{P(C on second)} \end{aligned}$$

and students see that knowledge of what happened on the first question does not alter the probability of the outcome on the second; the two events are independent.

In the selecting students scenario, the conditional probability of a girl on the second selection, given that a girl was selected on the first is

$$\begin{aligned} \text{P(girl on second} \mid \text{girl on first)} &= \frac{6}{12} \\ &= \frac{1}{2} \end{aligned}$$

• Two events A and B are said to be independent if $\text{P(A)} \cdot \text{P(B)} = \text{P(A and B)}$.

Selecting two students from three girls and two boys

Number of girls	Outcomes	
2	AB	BA
2	AC	CA
2	BC	CB
1	AD	DA
1	AE	EA
1	BD	DB
1	BE	EB
1	CD	DC
1	CE	EC
0	DE	ED

and

$$\text{P(girl on second)} = \frac{3}{5}.$$

So, these two events are again seen to be dependent. The outcome of the second draw does depend on what happened at the first draw.[S-CP.3]

Students understand that in real world applications the probabilities of events are often approximated by data about those events. For example, the percentages in the table for HIV risk by age group (p. 4) can be used to approximate probabilities of HIV risk with respect to age or age with respect to HIV risk for a randomly selected adult from the U.S. population of adults. Emphasizing the conditional nature of the row and column percentages:

$$\text{P(adult is age 18 to 24 | adult is at risk)} = 0.171$$

whereas

$$\text{P(adult is at risk | adult is age 18 to 24)} = 0.650.$$

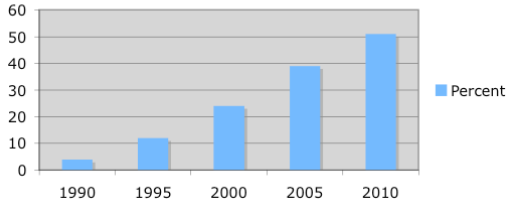
Comparing the latter to

$$\text{P(adult is at risk | adult is age 25 to 44)} = 0.483$$

shows that the conditional distributions change from column to column, reflecting dependence and an association between age category and HIV risk.[S-CP.4, S-CP.5]

Students can gain practice in interpreting percentages and using them as approximate probabilities from study data presented in the popular press. Quite often the presentations are a little confusing and can be interpreted in more than one way. For example, two data summaries from *USA Today* are shown below. What might these percentages represent and how might they be used as approximate probabilities?[S-CP.5]

Grandparents who are Baby Boomers



Top age groups for DUI	
21–25	29%
26–29	24%
18–20	20%
30–34	19%

Use the rules of probability to compute probabilities of compound events in a uniform probability model The two-way table for HIV risk by age group (p. 4) gives percentages from a data analysis that can be used to approximate probabilities, but students realize that such tables can be developed from theoretical probability models. Suppose, for example, two fair six-sided number cubes are rolled, giving rise to 36 equally likely outcomes.

S-CP.3 Understand the conditional probability of A given B as $P(A \text{ and } B)/P(B)$, and interpret independence of A and B as saying that the conditional probability of A given B is the same as the probability of A, and the conditional probability of B given A is the same as the probability of B.

S-CP.4 Construct and interpret two-way frequency tables of data when two categories are associated with each object being classified. Use the two-way table as a sample space to decide if events are independent and to approximate conditional probabilities.

S-CP.5 Recognize and explain the concepts of conditional probability and independence in everyday language and everyday situations.

S-CP.5 Recognize and explain the concepts of conditional probability and independence in everyday language and everyday situations.

Outcomes for specified events can be diagramed as sections of the table, and probabilities calculated by simply counting outcomes. This type of example is one way to review information on conditional probability and introduce the addition and multiplication rules. For example, defining events:

A is "you roll numbers summing to 8 or more"

B is "you roll doubles"

and counting outcomes leads to

$$\begin{aligned} \text{P(A)} &= \frac{15}{36} \\ \text{P(B)} &= \frac{6}{36} \\ \text{P(A and B)} &= \frac{3}{36}, \quad \text{and} \\ \text{P(B|A)} &= \frac{3}{15}, \quad \text{the fraction of A's 15 outcomes that also fall in B.} \end{aligned}$$

S-CP.6

Possible outcomes: Rolling two number cubes

	1	2	3	4	5	6
1	1, 1	1, 2	1, 3	1, 4	1, 5	1, 6
2	2, 1	2, 2	2, 3	2, 4	2, 5	2, 6
3	3, 1	3, 2	3, 3	3, 4	3, 5	3, 6
4	4, 1	4, 2	4, 3	4, 4	4, 5	4, 6
5	5, 1	5, 2	5, 3	5, 4	5, 5	5, 6
6	6, 1	6, 2	6, 3	6, 4	6, 5	6, 6

S-CP.6 Find the conditional probability of A given B as the fraction of B's outcomes that also belong to A, and interpret the answer in terms of the model.

Now, by counting outcomes

$$\text{P(A or B)} = \frac{18}{36}$$

or by using the Addition Rule[S-CP.7]

$$\begin{aligned} \text{P(A or B)} &= \text{P(A)} + \text{P(B)} - \text{P(A and B)} \\ &= \frac{15}{36} + \frac{6}{36} - \frac{3}{36} \\ &= \frac{18}{36}. \end{aligned}$$

S-CP.7 Apply the Addition Rule, $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$, and interpret the answer in terms of the model.

\+ By the Multiplication Rule[S-CP.8]

$$\begin{aligned} \text{P(A and B)} &= \text{P(A)P(B|A)} \\ &= \frac{15}{36} \cdot \frac{3}{15} \\ &= \frac{3}{36}. \end{aligned}$$

S-CP.8 (+) Apply the general Multiplication Rule in a uniform probability model, $P(A \text{ and } B) = P(A)P(B|A) = P(B)P(A|B)$, and interpret the answer in terms of the model.

The assumption that all outcomes of rolling each cube once are equally likely results in the outcome of rolling one cube being independent of the outcome of rolling the other.[S-CP.5] Students should understand that independence is often used as a simplifying assumption in constructing theoretical probability models that approximate real situations. Suppose a school laboratory has two smoke alarms as a built in redundancy for safety. One has probability 0.4 of going off when steam (not smoke) is produced by running hot water and the other has probability 0.3 for the same event. The probability

S-CP.5 Recognize and explain the concepts of conditional probability and independence in everyday language and everyday situations.

that they both go off the next time someone runs hot water in the sink can be reasonably approximated as the product $0.4 \cdot 0.3 = 0.12$, even though there may be some dependence between two systems operating in the same room. Modeling independence is much easier than modeling dependence, but models that assume independence are still quite useful.

Using probability to make decisions

Calculate expected values and use them to solve problems As students gain experience with probability problems that deal with listing and counting outcomes, they will come to realize that, most often, applied problems concern some numerical quantity of interest rather than a description of the outcomes themselves.[MP1] [MP2] Advertisers want to know how many customers will purchased their product, not the order in which they came into the store. A political pollster wants to know how many people are likely to vote for a particular candidate and a student wants to know how many questions he is likely to get right by guessing on a true-false quiz.

In such situations, the outcomes can be seen as numerical values of a *random variable*.• Reconfiguring the tables of outcomes for the true-false test (p. 13) and student selection (p. 14) in a way that emphasizes these numerical values and their probabilities gives rise to the probability distributions shown below.

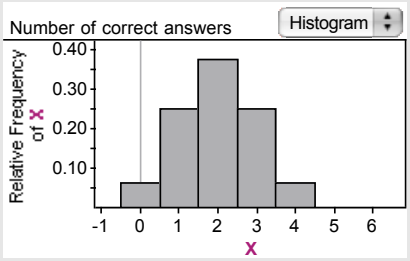
[MP1] Make sense of a problem, analyzing givens, constraints, relationships, and goals.

[MP2] Formulate a probability model for a practical problem that reflects constraints and relationships, and reason abstractly to solve the problem.

• Students should realize that random variables are different from the variables used in other high school domains; random variables are functions of the outcomes of a random process and thus have probabilities attached to their possible values.

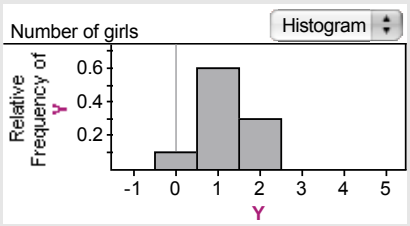
True–false test

Number correct, X	Probability
0	$\frac{1}{16}$
1	$\frac{4}{16}$
2	$\frac{6}{16}$
3	$\frac{4}{16}$
4	$\frac{1}{16}$



Selection of students

Number of girls, Y	Probability
0	$\frac{1}{10}$
1	$\frac{6}{10}$
2	$\frac{3}{10}$



Because probability is viewed as a long-run relative frequency, probability distributions can be treated as theoretical data distributions. If 1600 students all guessed at all four questions on the true-false test, about 400 of them would get three answers correct, about 100 four answers correct, and so on. These scores could then be averaged to come up with a mean score of:

$$0 \cdot \frac{1}{16} + 1 \cdot \frac{4}{16} + 2 \cdot \frac{6}{16} + 3 \cdot \frac{4}{16} + 4 \cdot \frac{1}{16} = 2.$$

With the *number correct* labeled as X, this value is called the *expected value* of X, usually expressed as E(X). Anyone guessing at all four true-false questions on a test can expect, over the long run, to get two correct answers per test, which is intuitively reasonable.

Students then develop the general rule that, for any discrete• random variable X,

$$E(X) = \sum (\text{value of X})(\text{probability of that value})$$

where the sum extends over all values of X.[S-MD.2]

• Students need not learn the term "discrete random variable." All of the random variables treated in this Progression are discrete random variables, that is, they concern only sample spaces which are collections of discrete objects.

[S-MD.2] (+) Calculate the expected value of a random variable; interpret it as the mean of the probability distribution.

For the random variable *number of girls*, Y, E(Y) = 1.2. Of course, 1.2 girls cannot be selected in any one group, but if the group selects leaders at random each day for ten days, they would be expected to choose about 12 girls as compared to 8 boys over the period.

The probability distributions considered above arise from theoretical probability models, but they can also come from empirical approximations. The margin displays the distribution of family sizes in the U.S., according to the Census Bureau. (Very few families have more than seven members.) These proportions calculated from census counts can serve as to approximate probabilities that families of given sizes will be selected in a random sample. If an advertiser randomly samples 1000 families for a special trial of a new product to be used by all members of the family, she would expect to have the product used by about 3.49 people per family, or about 3,490 people over all.

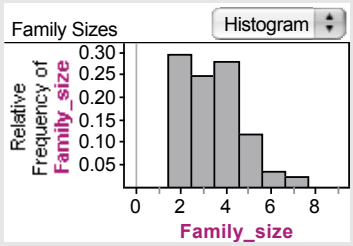
Use probability to evaluate outcomes of decisions Students should understand that probabilities and expected values must be thought of as long-term relative frequencies and means, and consider the implications of that view in decision making. Consider the following real-life example. The Wisconsin lottery had a game called "Hot Potato" that cost a dollar to play and had payoff probabilities as shown in the margin. The sum of these probabilities is not 1, but there is a key payoff value missing from the table. Students can include that key value and its probability to make this a true probability distribution and find that the expected payoff per game is about \$0.55.[S-MD.5] Losing a dollar to play the game may not mean much to an individual player, but expecting to take in \$450 for every \$1000 spent on the game means a great deal to the Wisconsin Lottery Commission!

Studying the behavior of games of chance is fun, but students must see more serious examples such as this one, based on empirical data. In screening for HIV by use of both the ELISA and Western Blot tests, HIV-positive males will test positive in 99.9% of the cases and HIV-negative males will test negative in 99.99% of the cases. Among men with low-risk behavior, the rate of HIV is about 1 in 10,000. What is the probability that a low-risk male who tests positive actually is HIV positive?

Having students turn the given rates into expected counts and placing the counts in an appropriate table is a good way for them to construct a meaningful picture of what is going on here. There are two variables, whether or not a tested person is HIV positive and whether or not the test is positive. Starting with a cohort of 10,000 low-risk males, the table might look like the one in the margin. The conditional probability of a randomly selected male being HIV positive, given that he tested positive is about 0.5! Students should discuss the implications of this in relation to decisions concerning mass screening for HIV.[S-MD.6, S-MD.7]

Empirical distribution of family size

Family size	Proportion
2	0.28
3	0.25
4	0.27
5	0.13
6	0.04
7	0.03



Source: U.S. Census Bureau, `http://www.census.gov/population/www/socdemo/hh-fam/cps2010.html`; Table F1

"Hot Potato" payoffs and probabilities

Payoff (\$)	Probability
1	$\frac{1}{9}$
2	$\frac{1}{13}$
3	$\frac{1}{43}$
6	$\frac{1}{94}$
9	$\frac{1}{150}$
18	$\frac{1}{300}$
50	$\frac{1}{2050}$
100	$\frac{1}{144000}$
300	$\frac{1}{180000}$
900	$\frac{1}{270000}$

For details about Hot Potato and other lotteries, see `www.wilottery.com/scratchgames/historical.aspx`.

S-MD.5 (+) Weigh the possible outcomes of a decision by assigning probabilities to payoff values and finding expected values.

HIV testing expected frequencies

	HIV+ male	HIV- male	Totals
HIV+ test result	0.999	1	1.999
HIV- test result	0.001	9,998	9,998.001
Totals	1	9,999	10,000

S-MD.6 (+) Use probabilities to make fair decisions (e.g., drawing by lots, using a random number generator).

S-MD.7 (+) Analyze decisions and strategies using probability concepts (e.g., product testing, medical testing, pulling a hockey goalie at the end of a game).

Where the Statistics and Probability Progression might lead

Careers A few examples of careers that draw on the knowledge discussed in this Progression are actuary, manufacturing technician, industrial engineer or statistician, industrial engineer and production manager. The level of education required for these careers and sources of further information and examples of workplace tasks are summarized in the table below. Information about careers for statisticians in health and medicine, business and industry, and government appears on the web site of the American Statistical Association (`www.amstat.org/careers/index.cfm`).

	Education	Location of information, workplace task
Actuary	bachelors	*Ready or Not*, p. 79; `http://beanactuary.org/how/highschool/`
Manufacturing technician	associate	*Ready or Not*, p. 81
Industrial engineer or statistician	bachelors	`http://www.achieve.org/node/205`
Industrial engineer; production manager	bachelors	`http://www.achieve.org/node/620`

Source: *Ready or Not: Creating a High School Diploma That Counts*, 2004, `www.achieve.org/ReadyorNot`

College Most college majors in the sciences (including health sciences), social sciences, biological sciences (including agriculture), business, and engineering require some knowledge of statistics. Typically, this exposure begins with a non-calculus-based introductory course that would expand the empirical view of statistical inference found in this high school progression to a more general view based on mathematical formulations of inference procedures. (The Advanced Placement Statistics course is at this level.) After that general introduction, those in more applied areas would take courses in statistical modeling (regression analysis) and the design and analysis of experiments and/or sample surveys. Those heading to degrees in mathematics, statistics, economics, and more mathematical areas of engineering would study the mathematical theory of statistics and probability at a deeper level, perhaps along with more specialized courses in, say, time series analysis or categorical data analysis. Whatever their future holds, most students will encounter data in their chosen field—and lots of it. So, gaining some knowledge of both applied and theoretical statistics, along with basic skills in computing, will be a most valuable asset indeed!