

## **A) Statistics & Probability » Introduction (excerpted from Common Core Standards)**

Decisions or predictions are often based on data—numbers in context. These decisions or predictions would be easy if the data always sent a clear message, but the message is often obscured by variability. Statistics provides tools for describing variability in data and for making informed decisions that take it into account.

Data are gathered, displayed, summarized, examined, and interpreted to discover patterns and deviations from patterns. Quantitative data can be described in terms of key characteristics: measures of shape, center, and spread. The shape of a data distribution might be described as symmetric, skewed, flat, or bell shaped, and it might be summarized by a statistic measuring center (such as mean or median) and a statistic measuring spread (such as standard deviation or interquartile range). Different distributions can be compared numerically using these statistics or compared visually using plots. Knowledge of center and spread are not enough to describe a distribution. Which statistics to compare, which plots to use, and what the results of a comparison might mean, depend on the question to be investigated and the real-life actions to be taken.

Randomization has two important uses in drawing statistical conclusions. First, collecting data from a random sample of a population makes it possible to draw valid conclusions about the whole population, taking variability into account. Second, randomly assigning individuals to different treatments allows a fair comparison of the effectiveness of those treatments. A statistically significant outcome is one that is unlikely to be due to chance alone, and this can be evaluated only under the condition of randomness. The conditions under which data are collected are important in drawing conclusions from the data; in critically reviewing uses of statistics in public media and other reports, it is important to consider the study design, how the data were gathered, and the analyses employed as well as the data summaries and the conclusions drawn.

Random processes can be described mathematically by using a probability model: a list or description of the possible outcomes (the sample space), each of which is assigned a probability. In situations such as flipping a coin, rolling a number cube, or drawing a card, it might be reasonable to assume various outcomes are equally likely. In a probability model, sample points represent outcomes and combine to make up events; probabilities of events can be computed by applying the Addition and Multiplication Rules. Interpreting these probabilities relies on an understanding of independence and conditional probability, which can be approached through the analysis of two-way tables.

Technology plays an important role in statistics and probability by making it possible to generate plots, regression functions, and correlation coefficients, and to simulate many possible outcomes in a short amount of time.

### **Connections to Functions and Modeling**

Functions may be used to describe data; if the data suggest a linear relationship, the relationship can be modeled with a regression line, and its strength and direction can be expressed through a correlation coefficient.

## B) Statistics & Probability » Overview

- **Interpreting Categorical and Quantitative Data**
  - Summarize, represent, and interpret data on a single count or measurement variable
  - Summarize, represent, and interpret data on two categorical and quantitative variables
  - Interpret linear models
- **Making Inferences and Justifying Conclusions**
  - Understand and evaluate random processes underlying statistical experiments
  - Make inferences and justify conclusions from sample surveys, experiments and observational studies
- **Conditional Probability and the Rules of Probability**
  - Understand independence and conditional probability and use them to interpret data
  - Use the rules of probability to compute probabilities of compound events in a uniform probability model
- **Using Probability to Make Decisions**
  - Calculate expected values and use them to solve problems
  - Use probability to evaluate outcomes of decisions
- **Mathematical Practices**
  1. Make sense of problems and persevere in solving them.
  2. Reason abstractly and quantitatively.
  3. Construct viable arguments and critique the reasoning of others.
  4. Model with mathematics.
  5. Use appropriate tools strategically.
  6. Attend to precision.
  7. Look for and make use of structure.
  8. Look for and express regularity in repeated reasoning.

### **Summarize, represent, and interpret data on a single count or measurement variable**

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

## **Summarize, represent, and interpret data on two categorical and quantitative variables**

- 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.
- 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. Use given functions or choose a function suggested by the context. Emphasize linear, quadratic, and exponential models.
  - 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.

## **Interpret linear models**

- 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.

## **C) Statistics & Probability » Interpreting Categorical & Quantitative Data**

### **Summarize, represent, and interpret data on a single count or measurement variable**

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

## **Summarize, represent, and interpret data on two categorical and quantitative variables**

- 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.
- 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. Use given functions or choose a function suggested by the context. Emphasize linear, quadratic, and exponential models.
  - 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.

## **Interpret linear models**

- 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.

## **D) Statistics & Probability » Making Inferences & Justifying Conclusions**

### **Understand and evaluate random processes underlying statistical experiments**

- S-IC.1. Understand statistics as a process for making inferences about population parameters based on a random sample from that population.
- S-IC.2. Decide if a specified model is consistent with results from a given data-generating process, e.g., using simulation. *For example, a model says a spinning coin falls heads up with probability 0.5. Would a result of 5 tails in a row cause you to question the model?*

### **Make inferences and justify conclusions from sample surveys, experiments, and observational studies**

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

- 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.5. Use data from a randomized experiment to compare two treatments; use simulations to decide if differences between parameters are significant.
- S-IC.6. Evaluate reports based on data.

## **E) Statistics & Probability » Conditional Probability & the Rules of Probability**

### **Understand independence and conditional probability and use them to interpret data**

- 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”).
- S-CP.2. Understand that two events  $A$  and  $B$  are independent if the probability of  $A$  and  $B$  occurring together is the product of their probabilities, and use this characterization to determine if they are independent.
- 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. *For example, collect data from a random sample of students in your school on their favorite subject among math, science, and English. Estimate the probability that a randomly selected student from your school will favor science given that the student is in tenth grade. Do the same for other subjects and compare the results.*
- S-CP.5. Recognize and explain the concepts of conditional probability and independence in everyday language and everyday situations. *For example, compare the chance of having lung cancer if you are a smoker with the chance of being a smoker if you have lung cancer.*

### **Use the rules of probability to compute probabilities of compound events in a uniform probability model**

- 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.
- 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.
- 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.

- S-CP.9. (+) Use permutations and combinations to compute probabilities of compound events and solve problems.

## F) Statistics & Probability » Using Probability to Make Decisions

### Calculate expected values and use them to solve problems

- S-MD.1. (+) Define a random variable for a quantity of interest by assigning a numerical value to each event in a sample space; graph the corresponding probability distribution using the same graphical displays as for data distributions.
- S-MD.2. (+) Calculate the expected value of a random variable; interpret it as the mean of the probability distribution.
- S-MD.3. (+) Develop a probability distribution for a random variable defined for a sample space in which theoretical probabilities can be calculated; find the expected value. *For example, find the theoretical probability distribution for the number of correct answers obtained by guessing on all five questions of a multiple-choice test where each question has four choices, and find the expected grade under various grading schemes.*
- S-MD.4. (+) Develop a probability distribution for a random variable defined for a sample space in which probabilities are assigned empirically; find the expected value. *For example, find a current data distribution on the number of TV sets per household in the United States, and calculate the expected number of sets per household. How many TV sets would you expect to find in 100 randomly selected households?*

### Use probability to evaluate outcomes of decisions

- S-MD.5. (+) Weigh the possible outcomes of a decision by assigning probabilities to payoff values and finding expected values.
  - a. Find the expected payoff for a game of chance. *For example, find the expected winnings from a state lottery ticket or a game at a fast-food restaurant.*
  - b. Evaluate and compare strategies on the basis of expected values. *For example, compare a high-deductible versus a low-deductible automobile insurance policy using various, but reasonable, chances of having a minor or a major accident.*
- 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).

## (An Incomplete) Glossary Relative to Statistics

**Average.** See Mean

**Bivariate data.** Pairs of linked numerical observations. Example: a list of heights and weights for each player on a volleyball team.

**Box plot.** A method of visually displaying a distribution of data values by using the median, quartiles, and extremes of the data set. A box shows the middle 50% of the data.<sup>1</sup>

**Dot plot.** See: line plot.

**Expected value.** For a random variable, the weighted average of its possible values, with weights given by their respective probabilities.

**First quartile.** For a data set with median  $M$ , the first quartile is the median of the data values less than  $M$ . Example: For the data set {1, 3, 6, 7, 10, 12, 14, 15, 22, 120}, the first quartile is 6. Many different methods for computing quartiles are in use. See *also*: median, third quartile, interquartile range.

**Independently combined probability models.** Two probability models are said to be combined independently if the probability of each ordered pair in the combined model equals the product of the original probabilities of the two individual outcomes in the ordered pair.

**Interquartile Range.** A measure of variation in a set of numerical data, the interquartile range is the distance between the first and third quartiles of the data set. Example: For the data set {1, 3, 6, 7, 10, 12, 14, 15, 22, 120}, the interquartile range is  $15 - 6 = 9$ . See *also*: first quartile, third quartile.

**Line plot.** A method of visually displaying a distribution of data values where each data value is shown as a dot or mark above a number line. Also known as a dot plot.

**Mean.** A measure of center in a set of numerical data, computed by adding the values in a list and then dividing by the number of values in the list. To be more precise, this defines the arithmetic mean. Example: For the data set {1, 3, 6, 7, 10, 12, 14, 15, 22, 120}, the mean is 21.

**Mean absolute deviation.** A measure of variation in a set of numerical data, computed by adding the distances between each data value and the mean, then dividing by the number of data values. Example: For the data set {2, 3, 6, 7, 10, 12, 14, 15, 22, 120}, the mean absolute deviation is 20.

**Median.** A measure of center in a set of numerical data. The median of a list of values is the value appearing at the center of a sorted version of the list—or the mean of the two central values, if the list contains an even number of values. Example: For the data set {2, 3, 6, 7, 10, 12, 14, 15, 22, 90}, the median is 11.

**Number line diagram.** A diagram of the number line used to represent numbers and support reasoning about them. In a number line diagram for measurement quantities, the interval from 0 to 1 on the diagram represents the unit of measure for the quantity.

**Percent rate of change.** A rate of change expressed as a percent. Example: if a population grows from 50 to 55 in a year, it grows by  $5/50 = 10\%$  per year. (“amount of change” divided by “the original amount”)

**Probability distribution.** The set of possible values of a random variable with a probability assigned to each.

**Probability.** A number between 0 and 1 used to quantify likelihood for processes that have uncertain outcomes (such as tossing a coin, selecting a person at random from a group of people, tossing a ball at a target, or testing for a medical condition).

**Probability model.** A probability model is used to assign probabilities to outcomes of a chance process by examining the nature of the process. The set of all outcomes is called the sample space, and their probabilities sum to 1. See *also*: uniform probability model.

**Random variable.** An assignment of a numerical value to each outcome in a sample space. Rational expression. A quotient of two polynomials with a non-zero denominator.

**Repeating decimal.** The decimal form of a rational number. See *also*: terminating decimal.

**Sample space.** In a probability model for a random process, a list of the individual outcomes that are to be considered.

**Scatter plot.** A graph in the coordinate plane representing a set of bivariate data. For example, the heights and weights of a group of people could be displayed on a scatter plot.

**Standard deviation.** In statistics and probability theory the **standard deviation (SD)** (represented by the Greek letter sigma,  $\sigma$ ) measures the amount of variation or dispersion from the average.

**Terminating decimal.** A decimal is called terminating if its repeating digit is 0.

**Third quartile.** For a data set with median  $M$ , the third quartile is the median of the data values greater than  $M$ . Example: For the data set  $\{2, 3, 6, 7, 10, 12, 14, 15, 22, 120\}$ , the third quartile is 15. See *also*: median, first quartile, interquartile range.

**Uniform probability model.** A probability model which assigns equal probability to all outcomes. See *also*: probability model.