Appendix

Glossary

A

allocation An allocation is an arrangement for the values in a data set. For example, the data sets {1, 2, 3, 4, 5} and {3, 3, 3, 3, 3} each have a mean and a median equal to 3, but they are very different allocations. Allocation can also be used to describe the proximity of values to the mean; values may be closely distributed to or widely distributed from the mean, for example.

association An association between two variables exists when a change in the values for one variable produces a systematic change in the other. If an increase in one variable tends to result in an increase in the other, the association is positive. If an increase in one variable tends to result in a decrease in the other, the association is negative.

B

bias Bias, or systematic error, favors particular results. A measurement process is biased if it systematically overstates or understates the true value of a variable.

binomial experiment A binomial experiment consists of *n* trials, where each trial is like a coin toss—it has exactly two possible outcomes. In each trial, the probability for each outcome remains constant.

binomial probability model The binomial probability model specifies the probabilities for each of the two possible outcomes in a binomial experiment.

bivariate analysis Bivariate analysis is a kind of data analysis that explores the association between two variables.

box plot A box plot, also known as a box-and-whiskers plot, is a graphical representation of the Five-Number Summary of a data set. A box is drawn from the lower quartile (Q1) to the upper quartile (Q3); a horizontal line across the box indicates the median. Two whiskers are drawn, one from the lower quartile to the minimum and one from the upper quartile to the maximum. Box plots can be used to make graphical comparisons between data sets and to measure the variation within parts of a data set.

C

census A census is an attempt to include every individual in a given population in a sample.

comparative experimental study A comparative experimental study seeks to determine "cause and effect." In an experimental study, two groups are selected, and each group is given a different treatment. At the end of the experiment, the results for each group are compared to determine whether or not the treatment had an influence on the results. For example, an experimental study might indicate that people who were told to drink more milk daily had a decreased incidence of osteoporosis.

comparative observational study A comparative observational study seeks to determine differences in measured groups, where each group is selected based on a differentiating criterion. For example, an observational study might compare smokers to nonsmokers, or men to women. The difference between an observational study and an experimental study is that in an experimental study, participants are actively given different behaviors, while in an observational study, the different behaviors are predetermined and are used to place participants into groups.

comparative study A comparative study focuses on the relationship(s) between two or more sets of data. For example, a comparative study might demonstrate that, on average, the winners of a Best Actress award are younger than the winners of a Best Actor award. Comparative studies often use box plots and other statistical comparisons to prove that the distributions are different in a significant way.

contingency table A contingency table lists the number of values in each quadrant of a scatter plot.

continuous variable A continuous variable is a quantitative variable whose values can take on any value on a number line; it may contain a decimal or fractional value. For example, time is a continuous variable since its values can be any number zero or greater. Time can be measured on a number line, and any point on the number line is a possible point in time. This is in contrast to a discrete variable, which can only accept whole numbers as values (such as the number of raisins in a box).

co-variation Co-variation describes the way two variables simultaneously change together.

Glossary, cont'd.

cumulative frequency Cumulative frequency specifies how many data values are of a particular number or smaller. For example, in the data set {1, 1, 2, 2, 3, 3, 4, 4, 4, 5, 26}, the cumulative frequency for the value 4 is nine, since there are nine values in the set that are 4 or less. The cumulative frequency for the value 2 is four; the cumulative frequency for the value 26 is 11; and so on. The statement "You scored higher than 10 other students in this class" is a statement of cumulative frequency.

cumulative frequency table A cumulative frequency table is a representation of data that shows the cumulative frequency of each value in the data set.

D

data Data are a set of values for a measured variable.

design of a comparative study The design of a comparative study is the step-by-step description of how the study is conducted, including the selection process of participants and the process of data collection. Designs must be created in ways that reduce potential sources of bias.

deviation from the mean Deviation from the mean for a data value is the difference between the value and the mean. The deviation from the mean can be positive, negative, or zero. For example, in the data set {1, 2, 3, 4, 5}, the mean is 3, and the deviations from the mean for each data value are {-2, -1, 0, 1, 2}. Adding all the deviations from the mean, positive and negative, must result in zero, since the mean represents a balance point for these deviations—the point at which the excesses and deficits are perfectly balanced.

discrete data Discrete data are data whose measurements are obtained by counting and whose values must be whole numbers. The number of people living in a town, the number of times a person has been struck by lightning, the number of licks it takes to get to the center of a lollipop—these are all discrete data.

distribution The distribution of data describes the shape of a data set when displayed on a histogram. There are dozens of specific statistical distributions found in data, but two of the most common are uniform distribution (intervals with equal frequency) and normal distribution (a bell-shaped histogram).

E

equal-shares allocation See **fair allocation.**

experimental probability Experimental probability is the proportion of times a particular outcome actually occurs when a random experiment is repeated a large number of times.

F

fair allocation Fair allocation, or the equal-shares allocation, is an allocation in which each data value is equal to the mean. For example, if five people are to share 35 cookies, the fair allocation is for each person to have the mean of 7 cookies.

Five-Number Summary The Five-Number Summary of a data set is a five-item list comprising the minimum value, first quartile, median, third quartile, and maximum value of the set. It divides a data set into four sets, each of which contains 25% of the set.

frequency The frequency of a value in a data set is the number of times that that value appears in the set. For example, in the data set {1, 1, 2, 2, 3, 3, 4, 4, 4, 5, 26}, the frequency of the value 3 is two, the frequency of the value 26 is one, and the frequency of the value 6 is zero.

frequency bar graph A frequency bar graph is a graphical representation of data in which the values of the data are placed on the horizontal axis, and bars extend vertically above each value to indicate the frequency of that value. A bar graph indicating the population of a dozen cities is an example of a frequency bar graph.

frequency table A frequency table is a representation of data that shows the frequency of each value in the data set.

G

grouped frequency table A grouped frequency table is a representation of data in which the number (frequency) of data values that occurs within each interval (group) of a data set is listed.

Glossary, cont'd.

H

histogram A frequency histogram is a graphical representation of grouped continuous data. The groups of data values are placed on the horizontal axis, and bars are placed vertically above each value to indicate the frequency of the data for that interval.

I

interquartile range The interquartile range is the length of the interval between the lower quartile (Q1) and the upper quartile (Q3). This interval indicates the central, or middle, 50% of a data set.

interval An interval is a range of values for data. Some common intervals include the interval from the lowest data value to the highest data value and the interval that contains the middle 50% of data.

L

least squares line Also called the line of best fit, the least squares line is the line that most closely approximates a data set.

line of best fit See **least squares line.**

line plot A line plot is a graphical representation of data in which the values of the data are placed on the horizontal axis, and dots are placed vertically above each value to indicate the number of times that that value appears in the data. A line plot is sometimes called a dot plot.

M

mathematical probability Mathematical probability, or theoretical probability, is the proportion of times a particular outcome is expected to occur when a random experiment is repeated a large number of times.

mean The mean of a data set is the arithmetic average of the data set, which is obtained by adding all the values, then dividing by the number of values in the set. For example, in the data set {1, 1, 2, 2, 3, 3, 4, 4, 4, 5, 26}, the mean is 5; you find it by dividing the sum of the values in the set (55) by the number of values (11). The mean may or may not be an actual value in the set.

mean absolute deviation (MAD) The mean absolute deviation (MAD) of a data set is the average of the absolute values of all deviations from the mean in that set. For example, in the data set {1, 2, 3, 4, 5}, the mean is 3, the deviations from the mean are {-2, -1, 0, 1, 2}, the absolute deviations from the mean are {|-2|, |-1|, |0|, |1|, |2|}, and the MAD is $(2 + 1 + 0 + 1 + 2) / 5 = 1.2$. The MAD is a measure of, on average, how far the values in a data set are from the mean.

measure of central tendency A measure of central tendency is a value that represents the data set. The mean, median, and mode are examples of measures of central tendency. Although all measures of central tendency represent the data set, they are not necessarily the same value.

median The median of a data set is the value in the center of an ordered list of the data. It is also the value for which there are as many values above it as there are below it. For example, in the data set {1, 1, 2, 2, 3, 3, 4, 4, 4, 5, 26}, the sixth value has five above it and five below it. This value, 3, is the median. If a data set contains an even number of values, the median is found by taking the mean of the two values in the center of the ordered list.

midrange The midrange of a data set is the average of the minimum and maximum values.

mode The mode is the most frequently occurring value in a data set. For example, in the data set {1, 1, 2, 2, 3, 3, 4, 4, 4, 5, 26}, the mode is 4, which has a frequency of three. It is possible for a data set to have more than one mode if two or more values each have the highest frequency. It is also possible for a data set to have no mode if all of its values have the same frequency.

O

outcome An outcome is a possible result of a random experiment. Each outcome has a probability associated with it (between zero and one).

P

Pascal's Triangle Pascal's Triangle is a special triangular tabulation of numbers. Each row in the triangle corresponds to the frequencies in a binomial probability table for *n* trials.

population The population is the entire group that a study wants information about.

probability table A probability table shows each of the possible values for an outcome of an experiment, paired with its corresponding probability.

Glossary, cont'd.

Q

quadrants The four quadrants of a scatter plot are created when the graph is divided at the mean of each of the two variables. For example, the first quadrant consists of points that are above the mean for both variables.

qualitative data Qualitative data are the values of a measured qualitative variable.

qualitative variables Qualitative variables represent categories rather than numbers—for example, the colleges attended by the last 10 American presidents, or the five cars most likely to be stolen in the United States.

quantitative data Quantitative data are the values of a measured quantitative variable.

quantitative variables Quantitative variables represent numbers or quantities—for example, the number of lions in a box of animal crackers, or the height of each student in a classroom.

quartiles Quartiles are numbers that divide an ordered data set into four portions, each containing approximately one-fourth of the data. Twenty-five percent of the data values come before the first quartile (Q1). The median is the second quartile (Q2); 50% of the data values come before the median. Seventy-five percent of the data values come before the third quartile (Q3).

R

random assignment In a comparative experimental study, random assignment is frequently used to select the group in which participants are placed; this is done to reduce bias. For example, if an experiment attempted to study the effect of fear on people's ability to think clearly, such an experiment would be unreasonably biased if it were to ask for volunteers to make up its groups. Random assignment makes it equally likely that any participant will be placed in any group.

random error Random error is a nonsystematic measurement error that is beyond our control; its effects average out over a set of measurements.

random experiment A random experiment is an experiment whose outcomes are due to chance.

random sample A random sample is a sample that is selected completely by chance from the population.

relative frequency Relative frequency is frequency as a proportion of the whole set. For example, in the data set {1, 1, 2, 2, 3, 3, 4, 4, 4, 5, 26}, the relative frequency of the value 4 is 3/11, since the value 4 appears three times out of 11 total values. Relative frequencies can be expressed as fractions (3/11), decimals (.273), or percentages (27.3%). The total of all relative frequencies in a data set should be 1 (or 100%) but may instead be very close to 1, due to round-off error.

relative frequency bar graph A relative frequency bar graph is a graphical representation of data in which the values of the data are placed on the horizontal axis, and bars extend vertically above each value to indicate its relative frequency. A bar graph indicating the percentage of people who voted for each presidential candidate is an example of a relative frequency bar graph.

relative frequency histogram A relative frequency histogram is a histogram in which the relative frequency of each group appears on the vertical axis, rather than the actual frequency. Typically, the relative frequency is expressed as a percentage.

representative sample A representative sample is one in which the relevant characteristics of the sample members are generally the same as the characteristics of the population.

S

sample A sample is a segment of the population examined in a study to gain information about the entire population.

sample mean The sample mean is the mean of a sample. It can be used as an estimate of the mean of the population under study.

sample size The sample size is the number of observations taken from a population to form a sample. For example, when 500 people are polled regarding an upcoming election, the size of this sample is 500. Increasing the sample size generally leads to more accurate estimates.

sampling with replacement Sampling with replacement is a type of sampling in which it is possible for the same observation to be included more than once within a sample.

Glossary, cont'd.

sampling without replacement Sampling without replacement is a type of sampling in which the same observation cannot be included more than once within a sample. If the same unit is randomly selected a second time, it is ignored.

scatter plot A scatter plot is a graph that allows you to visualize the simultaneous changes taking place in two variables. Each of the paired values of the two variables is plotted as a point on a graph in two dimensions.

standard deviation The standard deviation of a data set is the square root of the variance of that set. For example, in a data set whose variance is 2, the standard deviation is the square root of 2, which is approximately 1.414. Like the MAD, the standard deviation is a measure of the typical amount that the values in a data set vary from the mean.

stem and leaf plot A stem and leaf plot is a representation of data in which each data value is separated into two parts—a stem and a leaf. For example, if the data are two-digit numbers, then the stems are commonly the tens digits, and the leaves would be the units digits. The stems are listed vertically (from smallest to largest), and the corresponding leaves for the data values are listed horizontally beside the appropriate stem. On the final version of the stem and leaf plot, the leaves are usually ordered within each stem. Note that the stems on a stem and leaf plot provide a mechanism for grouping numeric data.

sum of squared errors The sum of squared errors, or SSE, is the sum of the squares of the vertical distances from the values in a data set to the corresponding points on a trend line. The line of best fit, or the least squares line, is the line with the smallest SSE.

summary measures Summary measures are numbers that describe some significant characteristics of your data. Summary measures include the mean, the median, the mode, the maximum, the minimum, and the quartiles of a data set.

T

Three-Number Summary The Three-Number Summary of a data set is a three-item list comprising the minimum, median, and maximum values of the set. It divides a data set into two sets, each of which contains 50% of the set.

treatment The treatment in a comparative study is the defining difference between the groups. In an experimental study, the treatment might be a new drug being clinically tested. An observational study does not impose a treatment on individual objects; it observes the objects as they are.

tree diagram A tree diagram is a schematic diagram that can be used to describe the possible outcomes of a random experiment.

Two-Number Summary The Two-Number Summary of a data set is a two-item list comprising the minimum and maximum values of the set.

V

variable A variable is a characteristic that may change (i.e., vary) from one observation to another.

variance The variance of a data set is the average of the squares of all the deviations from the mean in that set. For example, in the data set {1, 2, 3, 4, 5}, the deviations from the mean are {-2, -1, 0, 1, 2}, and the variance is $([-2]^2 + [-1]^2 + 0^2 + 1^2 + 2^2) / 5 = 2$.

variation Variation is any difference in measured data. Variation can occur for many reasons, including random error and bias.

Credits

Web Site Production Credits

Learning Math: Data Analysis, Statistics, and Probability is a production of WGBH Interactive and WGBH Educational Programming and Outreach for Annenberg/CPB.

The contents of this module were developed in part under a grant to PBS from the Department of Education, Award Number R286A950001-99. However, the contents do not necessarily represent the policy of the Department of Education, and you should not assume endorsement by the Federal Government.

Senior Producer
Ted Sicker

Curriculum Director
Denise Blumenthal

Content Developers
Gary D. Kader, Appalachian State University, NC
L. Mike Perry, Appalachian State University, NC

Classroom Case Studies Developer
DeAnn Huinker, University of Wisconsin, Milwaukee

Curriculum Developer
Anna Brooks

Curriculum Project Coordinators
Laura O'Neill
Sanda Zdjelar

Core Advisors
Bowen Kerins
Faye Nisonoff Ruopp

Additional Advisors
Suzanne Chapin
Susan Friel

Designers
Plum Crane
Lisa Rosenthal
Christian Wise

Web Developers
Joe Brandt
Kirsten Connelly

Online Video Segment Coordinator
Mary Susan Blout

Business Managers
Walter Gadecki
Joe Karaman

Unit Managers
Maria Constantinides
Adriana Sacchi

With the assistance of
Tim Barney
Jennifer Davis-Kay
Nina Farouk
Yasmin Madan
Jessica Rueter
Kate Smyres
Julie Wolf

Credits, cont'd.

Video Production Credits

Learning Math: Data Analysis, Statistics, and Probability is a production of WGBH Educational Foundation for Annenberg/CPB.

Executive Producer
Michele Korf

Senior Project Director
Amy Tonkonogy

Content Developers
Gary D. Kader, Appalachian State University, NC
L. Mike Perry, Appalachian State University, NC

Facilitator
Gary D. Kader

Producer
Christine Dietlin

Editor
Glenn Hunsberger

Advisors
Nicholas Branca, San Diego State University, CA

Suzanne Chapin, Boston University, MA

Mary Eich, Newton Public Schools, MA

Hollee Freeman, TERC, MA

Susan Friel, University of North Carolina, Chapel Hill

DeAnn Huinker, University of Wisconsin, Milwaukee

Susan Lamon, Marquette University, WI

Miriam Leiva, University of North Carolina, Charlotte

Joan Lukas, University of Massachusetts, Boston

Carol Malloy, University of North Carolina, Chapel Hill

Michelle Manes, Mathematics Teacher and Educational Consultant, MA

Bill Masalski, University of Massachusetts, Amherst

Arthur Powell, Rutgers-Newark College of Arts and Sciences, NJ

Sid Rachlin, East Carolina University, NC

Faye Nisonoff Ruopp, Educational Consultant, MA

Marty Simon, Pennsylvania State University, State College

Myriam Steinback, TERC, MA

Content Editor
Srdjan Divac, Harvard University, MA

Associate Producers
Irena Fayngold
Jessica Rueter

Project Coordinator
Sanda Zdjelar

Production Manager
Mary Ellen Gardiner

Credits, cont'd.

Post Production Associate Producer
Peter Villa

Production Coordinator
Lisa Eure

Location Coordinator
Nathan Gunner

Director
Fred Barzyk

Camera
Sam Ameen
Bill Charette
Bart Childs
Lance Douglas
Larry LeCain
Steve McCarthy

Audio
Steve Bores
Chris Bresnahan
Mario Cardenas
Charlie Collias
David King
Keith McManus
Giles Morin
Andy Turrett

Grip
Brian Snider

Intern
David Siegel

Design
Gaye Korbet
Daryl Myers
Alisa Placas

On-line Editors
David Eells
John O'Brien

Sound Mix
John Jenkins
Dan Lesiw

Original Music
Tom Martin

Narrator
Judy Richardson

Business Manager
Joe Karaman

Unit Manager
Maria Constantinides

Office Coordinators
Justin Brown
Ivy Moylan
Laurie Wolf

Credits, cont'd.

Special Thanks:

Session 1. Measurement Error: This Old Tape Measure

Norm Abram, Tom Silva, *This Old House*

Session 2. Data Organization and Representation: Weather Forecasting

Kim Martucci, Fox Studio 25, WFXT Boston

Session 3. Describing Distributions: New Balance

New Balance Inc.:

Jim Tompkins, President and C.O.O.

Stan Mescon, Sales and Product Planning Manager

Dave Elder, Manager of Production Planning

Session 4. The Median: Salary.com

Bill Coleman, Vice President of Compensation, Salary.com

Session 5. Variation About the Mean: The Boston Harbor Project

Massachusetts Water Resources Authority:

Dr. Andrea Rex, Lisa Wong, Nicole O'Neill

Session 6. Designing Experiments: Physicians' Health Study

Brigham and Women's Hospital:

Howard Sesso, Division of Preventive Medicine

Dr. Michelle Holmes, The Channing Laboratory

Session 7. Bivariate Data and Analysis: Anthropological Studies

Mark E. Mack, Professor, Biological Anthropology, Howard University

Archival photographs courtesy of The Library of Congress

Session 8. Probability: Risk Analysis

Doug McCrum, President, Global Specialty Risk

Archival footage courtesy of Sekani, Inc., and KD&E, Inc.

Session 9. Random Sampling and Estimation: Lake Victoria

New England Aquarium

William Oweke Ojwang, Les Kaufman, Biology Dept., Boston University

Archival footage courtesy of Nile Ziemba, North River Media, and New England Aquarium

Session 10. Classroom Case Studies

Suzanne L'Esperance, Northwest Elementary School, NH

Ellen Sabanosh, Gerald M. Parmenter School, MA

Paul Sowden, Eleanor Johnson Middle School, MA

Site Location

Oak Hill School, Newton, MA