GLOSSARY

A

absolute cell referencing A spreadsheet feature that blocks automatic adjustment of cell references when formulas are moved or copied. References preceded by a dollar sign—A1, for example—are left unchanged.

accidental relationship A correlation between two variables that happens by random chance.

action plan A logical sequence of specific steps for completing a project or testing a hypothesis.

addition rule for mutually exclusive events The principle relating the probabilities of events that cannot occur at the same time. For example, if events A and B are mutually exclusive, then $P(A \text{ or } B) = P(A) + P(B)$.

addition rule for non-mutually exclusive events The principle relating the probabilities of events that can occur at the same time. For example, if events A and B are not mutually exclusive, then $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$.

additive counting principle (rule of sum) The principle that, if one mutually exclusive action can occur in m ways, a second in n ways, a third in p ways, and so on, then there are $m + n + p \ldots$ ways in which one of these actions can occur.

adjacent vertices Vertices that are connected by an edge.

algorithm A procedure or set of rules for solving a problem.

arrow diagram A diagram that uses lines and arrows to illustrate the sequence of steps in a process.

B

balloting process A sampling process in which the instrument is an anonymous ballot.

bar graph A chart or diagram that represents quantities with horizontal or vertical bars whose lengths are proportional to the quantities.

Bernoulli trials Repeated independent trials measured in terms of successes and failures.

bias Systematic error or undue weighting in a statistical study.

bimodal Having two modes, or "humps." See ***mode***.

binomial distribution A distribution having independent trials whose outcomes are either success or failure. The probability of success is unchanged from one trial to the next, and the random variable is the number of successes in a given number of trials.

binomial theorem A theorem giving the expansion of powers of a binomial: For any natural number n, $(a+b)^n = {}_nC_0a^n + {}_nC_1a^{n-1}b + \ \ldots + {}_nC_ra^{n-r}b^r + \ldots + {}_nC_nb^n$

box-and-whisker plot A graph that summarizes a set of data by representing the first quartile, the median, and the third quartile with a box and, the lowest and highest data with the ends of lines extending from the box.

C

CANSIM and **CANSIM II** The Canadian Socio-economic Information Management System, an extensive database compiled by Statistics Canada. It profiles the Canadian people, economy, and industries. CANSIM II is the updated version of CANSIM.

categorical data Data that can be sorted or divided by type rather than by numerical values.

cause-and-effect relationship A relationship in which a change in an independent variable (X) produces a change in a dependent variable (Y).

cell references A letter and number that indicate the column and row of a cell in a spreadsheet. For example, B3 refers to the third row of column B.

census An official count of an entire population or class of things.

circle graph A graph that represents quantities with segments of a circle that are proportional to the quantities. Circle graphs are also called pie charts.

circuit A path in a network that begins and ends at the same vertex.

class A set of data whose values lie within a given range or interval.

classical probability The probability of an event deduced from analysis of the possible outcomes. Classical probability is also called theoretical or *a priori* probability.

cluster sampling A survey of selected groups within a population. This sampling technique can save time and expense, but may not give reliable results unless the clusters are representative of the population.

coefficient of determination (generalized correlation coefficient) A measure of how closely a curve fits a set of data. The coefficient of determination is denoted by r^2, and can be calculated using the formula $r^2 = \frac{\sum(y_{\text{est}} - \bar{y})^2}{\sum(y - \bar{y})^2}$, where $\bar{y}$ is the mean y value and y_{est} is the value estimated by the best-fit curve.

coding matrix A matrix used to encode a message.

column matrix A matrix having only one column.

column sum The sum of the entries in a column of a matrix.

combination A selection from a group of items without regard to order. The number of combinations of r items taken from a set of n items is denoted by ${}_nC_r$, $C(n, r)$, or $\binom{n}{r}$, and equal to $\frac{n!}{r!(n-r)!}$.

combinatorics The branch of mathematics dealing with ideas and methods for counting, especially in complex situations.

common-cause factor An external variable that causes two variables to change in the same way.

common element An element that is in two sets.

complement of an event The set of all outcomes that are not included in an event. The complement of an event A is the event that event A does *not* happen, and is denoted as A' or $\sim A$.

complete network A network that has an edge between every pair of vertices.

compound event An event consisting of two or more events.

conditional probability of an event, $P(B|A)$ The probability that event B occurs, given that event A has already occurred.

confidence interval A range of values that is centred on the sample mean and has a specified probability of including the population mean, μ. For example, there is a 0.9 probability that μ will lie between the upper and lower limits of the 90% confidence interval.

confidence level The probability that a measurement or conclusion is correct. The confidence level is equal to $1 - \alpha$, where α is the ***significance level***.

connected network A network having at least one path connecting each pair of vertices.

consumer price index (CPI) A collective measure of the cost of items purchased by a typical family.

continuity correction Treating the values of a discrete variable as continuous intervals in order to use a normal approximation for a binomial distribution.

continuous Involving a variable or data that can have an infinite number of possible values in a given interval. A continuous function or distribution can be graphed as a smooth curve.

control group The group for which the independent variable is held constant in an experiment or statistical study.

convenience sample A sample selected simply because it is easily accessible. Such samples may not be random, so their results are not always reliable.

correlation coefficient A summary statistic that gives a quantified measure of the linear relationship between two variables. Sometimes referred to as the *Pearson product-moment coefficient of correlation*, this coefficient is denoted by r and can be calculated using the formula $r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$

correlational research The application of statistical methods to determine the relationship between two variables.

covariance The mean of the products of the deviations of two variables.

cubic function A polynomial function that can be written in the form $y = ax^3 + bx^2 + cx + d$, where a, b, c, and d are numerical coefficients and $a \neq 0$.

cumulative-frequency graph A graph that shows the running total of the frequencies for values of the variable starting from the lowest value.

cumulative probability The probability that a variable is less than a certain value.

curve of best fit The curve that fits closest to the data points in a scatter plot or best represents the relationship between two variables.

D

data Facts or pieces of information. A single fact is a *datum*.

database An organized store of records.

degree The number of edges that begin or end at a vertex in a network; also, the highest power of the variables in an equation or polynomial.

dependent event An event whose outcome depends directly on the outcome of another event.

dependent (or **response**) **variable** A variable whose value is affected by another variable.

deviation The difference in value between a datum and the mean.

dimensions The number of rows and columns in a matrix, usually expressed in the form $m \times n$ with the number of rows listed first.

direct linear correlation A relationship in which one variable increases at a constant rate as the other variable increases, a perfect positive linear correlation.

discrete variable A variable that can take on only certain values within a given range.

disjoint events Events that cannot occur at the same time, mutually exclusive events.

E

edges Line segments in a network.

element of a set An item or member in a set.

empirical probability The number of times that an event occurs in an experiment divided by the number of trials. The empirical probability is also known as the experimental or relative-frequency probability.

entry A number appearing in a matrix.

E-STAT An interactive educational web site hosted by Statistics Canada. This site enables students to access data from the CANSIM database.

expected value, $E(X)$ The predicted mean of all possible outcomes of a probability experiment.

experimental group The group for which the independent variable is changed in an experiment or statistical study.

exponential distribution A continuous probability distribution that predicts waiting times between consecutive events in a random sequence of events.

exponential function A function that can be written in the form $y = ab^x$, where a and b are numerical coefficients.

exponential regression An analytic technique for finding the equation with the form $y = ab^x$ or $y = ae^{kx}$ that best models the relationship between two variables.

extraneous variables Variables that affect or obscure the relationship between an independent and a dependent variable.

extrapolation Estimating variable values beyond the range of the data.

F

factorial, $n!$ A product of sequential natural numbers having the form
$n! = n \times (n - 1) \times (n - 2) \times \ldots \times 2 \times 1$.
The notation $n!$ is read "n factorial."

fair game A game in which the expectation is 0.

field A location in a database record where specific data are displayed or entered.

first-step probability vector, $S^{(1)}$ The row matrix in a Markov chain that gives the probabilities of being in any state after one trial.

fractal A geometric figure that is created using an iterative process and self-similar shapes.

frequency diagram A diagram, such as a histogram or frequency polygon, that shows the frequencies with which different values of a variable occur.

frequency polygon A plot of frequencies versus variable values with the resulting points joined by line segments.

function A built-in formula in a graphing calculator, spreadsheet, or other software.

fundamental (multiplicative) counting principle The principle that, if a task or process is made up of stages with separate choices, the total number of choices is $m \times n \times p \times \ldots$, where m is the number of choices for the first stage, n is the number of choices for the second stage, p is the number of choices for the third stage, and so on.

G

geometric distribution A distribution having independent trials whose outcomes are either success or failure. The probability of success is unchanged from one trial to the next, and the random variable is the number of trials before a success occurs.

graph In graph theory, a collection of line segments and vertices, which can represent interconnections between places, items, or people; a network.

graph theory A branch of mathematics in which graphs or networks are used to represent relationships and solve problems in many fields.

gross domestic product (GDP) A measure of a country's overall economic output, including all goods and services.

H

hidden variable An extraneous variable that is difficult to recognize.

histogram A bar graph in which the areas of the bars are proportional to the frequencies for various values of the variable.

hypergeometric distribution A distribution having dependent trials whose outcomes are either success or failure. The probability of success changes from one trial to the next, and the random variable is the number of successes.

hypothesis A proposition or thesis that is assumed to be true in order to investigate its validity.

hypothesis test A statistical procedure that uses a sample to determine the probability that a statement is correct.

I

identity matrix A matrix having entries of 1 along the main diagonal and zeros for all other entries, such as $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

independent event An event whose probability is not affected by the outcome of another event.

independent (or **explanatory**) **variable** A variable that affects the value of another variable.

index A number relating the value of a variable, or group of variables, to a base level, which is often the value on a particular date.

indirect method A problem-solving technique for finding a quantity by determining another quantity from which the first can be derived. Often the first quantity is difficult to calculate directly.

inflation An increase in the overall price of goods and services.

initial probability vector, $S^{(0)}$ A matrix that represents the probabilities of the initial state of a Markov chain.

initial value A value given for a term in the first step of a recursion formula.

instrument In statistics, any form of data-collection mechanism, such as questionnaires, personal interviews, telephone survey, or direct measurement.

interpolate Estimate a value between two known values.

interquartile range The range of the central half of a set of data when the data are arranged in numerical order.

intersection The set of elements common to two or more sets. The intersection of sets A and B is often written as $A \cap B$.

interval A set of all numbers between two given numbers.

inverse linear correlation A relationship in which one variable increases at a constant rate as the other decreases, a perfect negative linear correlation.

inverse matrix The matrix that produces the identity matrix when multiplied by a given matrix. The inverse of matrix A is written as A^{-1}. Only square matrices can have inverses.

iteration The process of repeating the same procedure over and over.

L

leading question A question which prompts a particular answer.

least-squares fit An analytic technique for determining the line of best fit by minimizing the sum of the squares of the deviations of the data from the line.

line of best fit The straight line that passes closest to the data points on a scatter plot and best represents the relationship between two variables.

linear correlation A relationship in which changes in one variable tend to be proportional to the changes in another.

linear regression An analytic technique for finding the equation with the form $y = ax + b$ that best models the relationship between two variables.

loaded question A question containing information or language intended to influence the respondents' answers.

lurking variable An extraneous variable that is difficult to recognize, a hidden variable.

M

margin of error The width of a confidence interval.

Markov chain A probability model in which the outcome of any trial depends directly on the outcome of the previous trial.

mathematical model A model that describes the relationship between variables in a quantitative fashion.

matrix A rectangular array of numbers used to manage and organize data.

mean The sum of the values in a set of data divided by the number of values.

mean absolute deviation The mean of the absolute values of the ***deviations*** of a set of data.

measurement bias Bias resulting from a data-collection method that consistently either under- or over-estimates a characteristic of the population.

measures of central tendency The values around which a set of the data tends to cluster.

measures of spread Quantities that indicate how closely a set of data clusters around its central values.

median The middle value of a set of data ranked from highest to lowest. If there is an odd number of data, the median is the midpoint between the two middle values.

member An item or element of a set.

midrange Half of the sum of the highest value and the lowest value in a set of data.

mind map A tool for organizing related topics and generating ideas about them and related sub-topics.

mind web A mind map in which related topics at the same level are joined with dotted lines.

modal interval For grouped data, the interval which contains more data than any other interval.

mode The value in a distribution or set of data that occurs most frequently.

modified box plot A box-and-whisker plot that shows outliers as separate points instead of including them in the whiskers.

multi-stage sampling A sampling technique that uses several levels of random sampling.

mutually exclusive events Events that cannot occur at the same time.

N

negative skew The pulling to the left of the tail in an asymmetric probability distribution.

neighbours In a network, vertices that are joined by an edge.

network In graph theory, a collection of line segments and vertices, which can represent interconnections between places, items, or people.

network matrix A matrix that represents a network.

node A point in a network at which edges end or meet, a vertex.

non-linear regression An analytical technique for finding a curve of best fit for a set of data.

non-mutually exclusive events Events that can occur simultaneously.

non-response bias Bias occurring when particular groups are under-represented in a survey because they choose not to participate.

normal distribution A common continuous probability distribution in which the data are distributed symmetrically and unimodally about the mean.

normal probability plot A graph of the data in a sample versus the z-scores of the corresponding quantiles for a normal distribution. If the plot is approximately linear, the underlying population can be assumed to be normally distributed.

***n*th-step probability vector, $S^{(n)}$** The row matrix in a Markov chain that gives the probabilities of being in any state after n trials.

null set A set that has no elements.

O

odds against The ratio of the probability that the event will not occur to the probability that it will occur.

odds in favour The ratio of the probability that the event will occur to the probability that it will not occur.

ogive A graph of a cumulative-frequency distribution.

outcome A possible result, a component of an event.

outliers Points in a set of data that are significantly far from the majority of the other data.

P

Pascal's formula A formula relating the combinations in Pascal's triangle: ${}_nC_r = {}_{n-1}C_{r-1} + {}_{n-1}C_r$. This formula is also known as Pascal's theorem.

path A connected sequence of vertices in a network.

percentile A set of values that divides an ordered set of data into 100 intervals, each having the same number of data.

perfect negative linear correlation A relationship in which one variable increases at a constant rate as the other variable decreases. A graph of the one variable versus the other is a straight line with a negative slope.

perfect positive linear correlation A relationship in which one variable increases at a constant rate as the other variable increases. A graph of the one variable versus the other is a straight line with a positive slope.

permutation An arrangement of items in a definite order. The total number of permutations of distinct n items is denoted by ${}_nP_n$ or $P(n, n)$ and is equal to $n!$.

pictograph A chart or diagram that represents quantities with symbols.

planar network A network that can be drawn such that its edges do not cross anywhere except at vertices.

polynomial regression An analytic technique for finding the polynomial equation that best models the relationship between two variables.

population All individuals that belong to a group being studied.

positive skew The pulling to the right of the tail in an asymmetric probability distribution.

power regression An analytic technique for finding the equation with the form $y = ax^b$ that best models the relationship between two variables.

presumed relationship A correlation that does not seem to be accidental even though no cause-and-effect relationship or common-cause factor is apparent.

principle of inclusion and exclusion The principle that the total number of elements in either set A or set B is the number in A plus the number in B minus the number in both A and B:
$n(A \cup B) = n(A) + n(B) - n(A \cap B)$

probability theory A branch of mathematics that deals with chance, random variables, and the likelihood of outcomes.

probability density The probability per unit of a continuous variable.

probability-density function An equation that describes or defines the curve for a probability distribution.

probability distribution The probabilities for all possible outcomes of an experiment, often shown as a graph of probability versus the value of a random variable.

probability experiment A well-defined process consisting of a number of trials in which clearly distinguishable outcomes are observed.

probability of an event *A*, *P*(*A*) A quantified measure of the likelihood that event A will occur. The probability of an event is always a value between 0 and 1.

product rule for dependent events The principle that the probability of event B occurring after event A has occurred is $P(A \text{ and } B) = P(A) \times P(B|A)$, where $P(B|A)$ is the ***conditional probability*** of event B.

product rule for independent events The principle that the probability of independent events A and B both occurring is $P(A \text{ and } B) = P(A) \times P(B)$.

Q

quadratic function A function that can be written in the form $y = ax^2 + bx + c$, where a, b, and c are numerical coefficients and $a \neq 0$.

quantile One of a set of values that divide a set of data into groups with equal numbers of data; the variable value corresponding to a given cumulative probability. For example, the first quartile has the cumulative probability $P(X \leq x) = 0.25$.

R

random variable A variable that can have any of a set of different values. In statistics, a random variable is often denoted by a capital letter (commonly X or Y), while its individual values are denoted by the corresponding lowercase letter.

randomization A technique that ensures that all members of a population are equally likely to be selected for a sample. Such techniques reduce the likelihood that results will be inappropriately weighted in favour of one particular group within the population.

range The difference between the highest and lowest values in a set of data.

raw data Unprocessed information.

record A set of data that is treated as a unit in a database.

recursion formula A formula for calculating a series of terms, each of which is derived from the preceding terms.

regression analysis An analytic technique for determining the relationship between two variables.

regular Markov chain A Markov chain that always achieves a steady state.

relational database Databases in which different sets of records can be linked and sorted in complex ways based on the data contained in the records.

relative cell referencing A spreadsheet feature that automatically adjusts cell references in formulas when they are moved or copied.

relative frequency The frequency of a value or group of values expressed as a fraction or percent of the whole data set.

repeated sampling A sampling method that uses two or more independent samples from the same population.

residual The difference between the observed value of a variable and the corresponding value predicted by the regression equation.

residual plot A graph of the residuals of a set of data versus the independent variable.

response bias Bias that occurs when participants in a survey deliberately give false or misleading answers.

reverse cause-and-effect relationship A relationship in which the presumed dependent and independent variables are reversed in the process of establishing causality.

row matrix A matrix having only one row.

row sum The sum of the entries in a row of a matrix.

S

sample A group of items or people selected from a population.

sample space, *S* The set of all possible outcomes in a probability experiment.

sampling A process of selecting a group from a population in order to estimate the characteristics of the entire population.

sampling bias Bias resulting from a sampling frame that does not reflect the characteristics of the population.

sampling frame The members of a population that actually have a chance of being selected for a sample.

scalar A quantity having only magnitude (as opposed to a vector, which also has a direction).

scatter plot A graph in which data are plotted with one variable on the x-axis and the other on the y-axis. The pattern of the resulting points can show the relationship between the two variables.

self-similar shape A shape containing components that have the same geometrical characteristics.

semi-interquartile range One half of the ***interquartile range***.

seed value A value given for a term in the first step of a recursion formula, an initial value.

set A group of items.

significance level, α The probability that the result of a hypothesis test will be incorrect.

simple random sample A sample in which every member of a population has an equal and independent chance of being selected.

simulation An experiment, model, or activity that imitates real or hypothetical conditions.

square matrix A matrix with the same number of rows as columns.

standard deviation The square root of the mean of the squares of the deviations of a set of data. The standard deviation is given by the formulas $\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$ for a population and $s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$ for a sample.

standard normal distribution A normal distribution in which the mean is equal to 0 and the standard deviation is equal to 1.

statistical bias Any factor that favours certain outcomes or responses and hence systematically skews the survey results.

statistical fluctuation The difference between characteristics measured from a sample and those of the entire population. Such differences can be substantial when the sample size is small.

statistical inference A process used to determine whether causality exists between the variables in a set of data.

statistics The gathering, organization, analysis, and presentation of numerical information.

Statistics Canada A federal government department that collects, summarizes, and analyses a broad range of Canadian statistics.

steady-state vector A probability vector in a Markov chain that remains unchanged when multiplied by the transition matrix.

stratified sample A sample in which each stratum or group is represented in the same proportion as it appears in the population.

stratum A group whose members share common characteristics, which may differ from the rest of the population.

subjective probability An estimate of the likelihood of an event based on intuition and experience—an educated guess.

subset A set whose elements are all also elements of another set.

systematic sample A sample selected by listing a population sequentially and choosing members at regular intervals.

T

theoretical probability The probability of an event deduced from analysis of the possible outcomes. Theoretical probability is also called classical or *a priori* probability.

time-series graph A plot of variable values versus time with the adjacent data points joined by line segments.

total variation The sum of the squares of the deviations for a set of data, $\sum(y-\bar{y})^2$.

traceable network A network whose vertices are all connected to at least one other vertex and whose edges can all be travelled exactly once in a continuous path.

transition matrix, *P* A matrix representing the probabilities of moving from any initial state to any new state in a given trial.

transpose matrix A matrix in which the rows and columns have been interchanged so that a_{ij} becomes a_{ji}.

trial A step in a probability experiment in which an outcome is produced and tallied.

triangular numbers The sum of the first n natural numbers: $1 + 2 + \ldots + n$. Triangular numbers correspond to the number of items stacked in a triangular array.

U

uniform probability distribution A probability distribution in which each outcome is equally likely in any single trial.

unimodal Having only one mode, or "hump." See ***mode***.

union The set of all elements contained in two or more sets.The union of sets A and B is often written as $A \cup B$.

universal set, *S* The set containing all elements involved in a particular situation.

V

variable A quantity that can have any of a set of values.

variance The mean of the squares of the deviations for a set of data. Variance is denoted by σ^2 for a population and s^2 for a sample.

Venn diagram A pictorial representation of one or more sets, in which each set is represented by a closed curve.

vertex A point in a network at which edges end or meet. Also called a node.

voluntary-response sample A sampling technique in which participation is at the discretion or initiative of the respondent.

W

waiting time or **period** The number of unsuccessful trials or the elapsed time before success occurs. Waiting time is the random variable in geometric and exponential probability distributions.

weighted mean A measure of central tendency that reflects the greater significance of certain data.

Z

zero matrix A matrix in which all entries are zero.

***z*-score** The number of standard deviations from a datum to the mean. The z-score of a datum is given by the formula $z = \frac{x-\bar{x}}{s}$.