Glossary

A

accidental relationship A relationship that is based purely on coincidence.

aggregate data Data that are combined or summarized in such a way that the individual microdata can no longer be determined.

arrangement An ordered list of items.

attribute A quality or characteristic given to a person, group, or object.

B

bias Occurs when there is a prejudice for or against an idea or response. Biased samples can result from problems with either the sampling technique or the data collection method.

binomial probability distribution A distribution with independent trials whose outcomes are either success or failure.

binomial theorem A formula used to expand $(a+b)^n$.

$$(x+y)^n = {}_nC_0x^ny^0 + {}_nC_1x^{n-1}y^1 + {}_nC_2x^{n-2}y^2 + \ldots + {}_nC_{n-1}x^1y^{n-1} + {}_nC_nx^0y^n$$

The general term is ${}_nC_rx^{n-r}y^r$.

C

categorical (qualitative) data Data that can be sorted into distinct groups or categories.

cause and effect relationship The correlation between two variables in which a change in one directly causes a change in the other.

combination A selection from a group of objects without regard to order. The number of r object chosen from a set of n items is ${}_nC_r = \frac{n!}{(n-r)!r!}$.

common cause relationship The correlation between two variables in which both variables change as a result of a third common variable.

complement A set of possible outcomes not included in an event.

compound events Multiple events in a probability experiment which may or may not affect each other.

conditional probability The probability of a second event occurring, given that a first event occurred. The sample space for the second event is reduced from the first event.

confidence interval The range of possible values of a measured statistic at a particular confidence level.

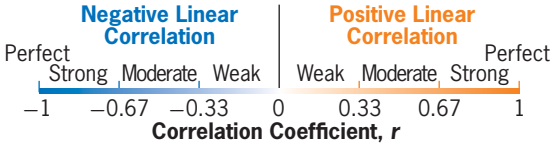
confidence level The probability that a particular statistic is within the range indicated by the margin of error.

continuity correction A correction applied when using the normal approximation to correct for the difference between a discrete and continuous distribution.

continuous random variable A variable that can have an infinite number of possible values in a given range.

control group The participants in an experiment who do not receive the specific treatment being measured.

correlation coefficient A measure of how well a linear model fits a two-variable set of data.



D

dependent events The occurrence or non-occurrence of one event influences the probability of the other event occurring.

discrete random variable A variable that can have only certain values within a given range.

E

event A set of outcomes that have a common characteristic.

expectation (expected value) The predicted average of all possible outcomes in a probability distribution.

$$E(X) = x_1 \cdot P(x_1) + x_2 \cdot P(x_2) + \ldots x_n \cdot P(x_n)$$
$$= \sum_{i=1}^{n} x_i \cdot P(x_i)$$

experimental probability Probability based on experimental trials.

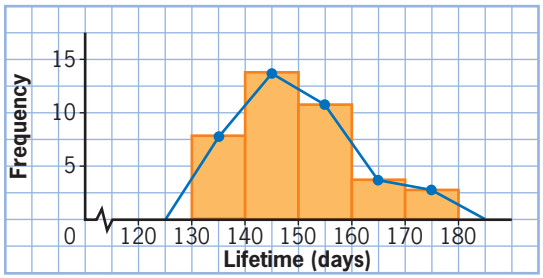
$P(A) = \frac{n(A)}{n(T)}$, where $P(A)$ is the probability that outcome A occurs, $n(A)$ is the number of times that outcome A occurred, and $n(T)$ is the total number of trials.

F

factorial A product of sequential natural numbers with the form $n! = n(n-1)(n-2) \ldots \times 2 \times 1$.

frequency histogram A graph with intervals on the horizontal axis and frequencies on the vertical axis.

frequency polygon A segmented line that joins the midpoints of the top of each column in a frequency histogram.



fundamental counting principle If one event can occur in m ways and a second event can occur in n ways, then together they can occur in $m \times n$ ways.

H

hidden variable A variable that affects or obscures the relationship between two other variables.

hypergeometric probability distribution A distribution with dependent trials whose outcomes are either success or failure.

hypothesis A prediction about the relationship between variables or about the outcome of a research question.

I

independent events Situations in which the occurrence or non-occurrence of one event has no influence on the probability of the other event occurring.

indirect method A method of solving a permutation where you subtract the number of unwanted outcomes from the total number of outcomes without restrictions.

interquartile range (IQR) The difference between the first and third quartiles.

L

line of best fit A straight line that represents a trend in the scatter plot as long as the pattern is more or less linear.

linear correlation A relationship in which a change in one variable tends to correspond to a proportional change in another variable.

linear regression The formal process by which a line of best fit is mathematically determined.

M

margin of error The range of values that a particular measurement is said to be within.

mean The sum of the data entries divided by the number of entries.

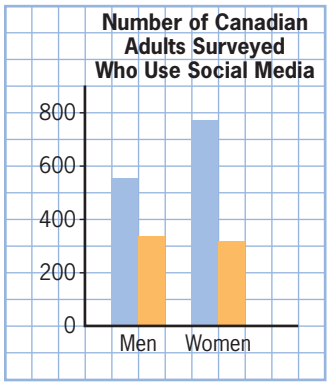
measurement bias When the survey collection method is such that the characteristics are consistently over- or under-represented.

median The middle value of all the data points when the data values are listed in order from least to greatest.

microdata An individual set of data about a single survey respondent.

mode The data value that occurs most often in a data set.

multiple bar graph A graph in which different quantities are represented by different colours and lengths of bars that are placed side-by-side.



mutually exclusive events Events that have different attributes and cannot occur simultaneously.

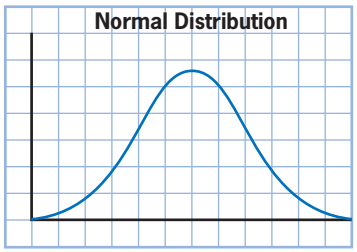
N

nominal data Qualitative data that cannot be ranked.

non-mutually exclusive events Different events that can happen at the same time.

non-response bias When the opinions of survey respondents differ in meaningful ways from those of non-respondents.

normal distribution A probability distribution around a central value, dropping off symmetrically to the right and left, forming a bell-like shape.



null set A set with no elements.

numerical (quantitative) data Data in the form of any number.

O

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

$$A = P(A) : P(A')$$

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

$$A = P(A') : P(A)$$

ordinal data Qualitative data that can be ranked.

outcome A possible result of an experiment.

outlier An element of the data set that is significantly different from the rest of the data points.

P

Pascal's triangle A triangular array of numbers in which each term is the sum of the two terms above it.

$$\begin{array}{ccccccccccc} & & & & 1 & & & & & & \\ & & & 1 & & 1 & & & & & \\ & & 1 & & 2 & & 1 & & & & \\ & 1 & & 3 & & 3 & & 1 & & & \\ 1 & & 4 & & 6 & & 4 & & 1 & & \end{array}$$

percentile The percent of all the data that are less than or equal to a specific data value.

$p = 100\dfrac{(L + 0.5E)}{n}$, where p is the percentile, L is the number of data less than the data point, E is the number of data equal to the data point, and n is the size of the population.

permutation An arrangement of n distinct items in a definite order. The total number of these permutations is written nPn or $P(n, n)$. The number of permutations of n items is ${}_nP_n = n!$.

population All the individuals in a group that is being studied.

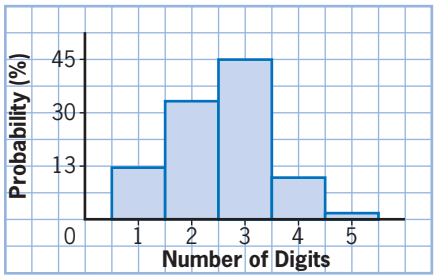
presumed relationship A relationship that makes sense but does not seem to have a causation factor.

primary source data Data that have been collected directly by the researcher and have not been manipulated or summarized.

probability The likelihood of something occurring.

probability distribution The probabilities for all possible outcomes of an experiment or sample space.

probability histogram A graph of a probability distribution in which equal intervals are marked on the horizontal axis and the probabilities associated with these intervals are indicated by the areas of the bars.



Q

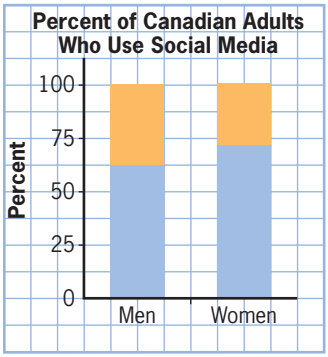
quartiles Three points that divide the data set into four equal groups. The first quartile (Q1) is the middle number between the smallest number and the median. The second quartile (Q2) is the median of the data set. The third quartile (Q3) is the middle number between the median and the largest number in a data set.

R

random variable A quantity that can have a range of values. Designated by X, with individual values designated by x.

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

relative split bar graph A graph in which different percents, totalling 100, are represented by different colours and lengths of bars that are placed one above the other.



reliable data Results of a study that can be duplicated in another study.

research question A question about a topic or problem that can be investigated or solved.

residual The difference between a data point's actual dependent value and the dependent value predicted by the line of best fit.

residual plot A graph which shows the value of each residual graphically as the vertical distance from a horizontal axis.

response bias When survey respondents change their answers to influence the results, to avoid embarrassment, or to give the answer they think the questioner wants.

reverse cause and effect relationship A relationship in which the independent and dependent variable are reversed.

S

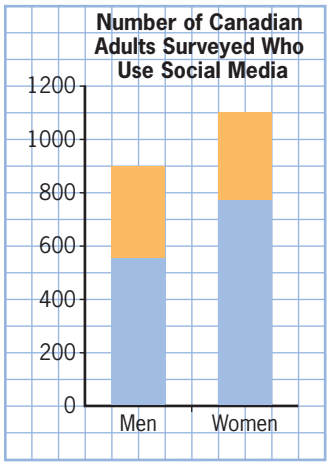
sample A group of items or people selected from the population.

sample space A collection of all possible outcomes. Sometimes called a sample set.

sampling bias When the sample does not closely represent the population.

secondary source data Data used by someone other than those who actually collected them.

split bar graph A graph in which different quantities are represented by different colours and lengths of bars that are placed one above the other.



standard deviation The average distance of the scores from the mean.

population standard deviation:

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

sample standard deviation:

$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$, where the population deviation is represented by $(x-\mu)$ and the sample deviation is represented by $(x-\bar{x})$.

subjective probability A probability estimate based on intuition.

T

theoretical probability A probability based on analysis of all possible outcomes.

$P(A) = \frac{n(A)}{n(S)}$, where $P(A)$ is the probability that event A can occur, $n(A)$ is the number of ways it can occur, and $n(S)$ is the total number of possible outcomes in the sample space.

treatment group The participants in an experiment who receive the specific treatment being measured.

U

uniform distribution Occurs when, in a single trial, all outcomes are equally likely.

$P(x) = \frac{1}{n}$, where n is the number of possible outcomes in the experiment.

V

valid data Results that accurately represent the entire population.

variability (in samples) Shows how samples are different from each other. The more similar the samples are, the lower the variability and the more accurately they represent the population.

variance The average squared difference of the scores from the mean.

population variance: $\sigma^2 = \frac{\sum(x-\mu)^2}{N}$

sample variance: $s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$,

where the population deviation is represented by $(x-\mu)$ and the sample deviation is represented by $(x-\bar{x})$.

W

weighted mean The mean of a set of numbers that are given weightings based on their frequency.

Z

***z*-score** The number of standard deviations an observation is from the mean.

population z-score: $z = \frac{x-\mu}{\sigma}$

sample z-score: $z = \frac{x-\bar{x}}{s}$

Index

D

E

T

U

V

W

Z