Data and Chance Vocabulary

A

Addition rule – the probability of either of two disjoint events occurring is the sum of the probabilities of the two events. For non-disjoint events, the probability is the sum minus the probability of the intersection of the two events occurring.

B

Bar Graph – a graph that shows frequencies of categorical data or values of numeric data using rectangles or bars

Bias – in sampling, occurs when not all individuals are equally likely to be chosen in a sample. In estimation, occurs when the sample statistics differs from the population statistic.

Binomial distribution – the distribution of successes of a collection of independent events where each event has the same probability of success

Box plot – a graphical representation of a data set that shows the median, the inter-quartile range, and the minimum and maximum values

C

Categorical Variable – a variable whose value is not numeric or is a numbers that does not have a numeric meaning

Cause and effect – a relationship between variables where changes in one variable (the cause) result in changes in the other (the effect)

Census – an examination of all individuals in a population

Center – one number that characterizes the middle or average value in a data set. Mean, median, and mode are all measures of center.

Central Limit Theorem – as the size of the sample increases, the sampling distribution of the sample mean tends to a normal distribution with mean equal to the population mean and standard deviation equal to the population standard deviation divided by the square root of the sample size

Chance - the possibility of a particular outcome in an uncertain situation

Chartjunk – items on graphs that are not part of the basic graph type

Cluster – groups of data that are close together in value.

Combination – a subset of a set of objects

Conditional probability – the probability that an event will occur given that another event has already occurred. It is the probability that both events occur divided by the probability that the first event occurs.

Confounding- when there is more than one possible variable influencing a result but it is not possible to separate the variables to determine the cause

Convenience sampling – sampling a population by choosing individuals that are easy to include in the sample

Correlation --- the degree to which two variables are (linearly) related

D

Data – numbers with a context.

Dependent event – the probability of one event depends on whether the other event occurs or not.

Derangement – a derangement is a permutation of the numbers 1 to N where no number is in its correct position

Digital image – a collection of samples taken from a picture that are represented numerically and whose numeric value represents brightness of color information

Disjoint (mutually exclusive) events – two events are disjoint if the do not contain a common outcome.

Distribution – the values of a variable and how frequently each occurs

Dominate – if one strategy is superior to another regardless of what an opponent does.

Dot plot – a graph showing distribution using dots

E

Event – a subset of a sample space

Excel – computer program to help analyze collect, manage, and display data

Expected value – for a probability model whose outcomes have numeric values, the sum over all outcomes of the probability of the outcome times the outcome (value). Over the long run of experiments, it is the mean of the outcomes.

Experiment – applying treatments to individuals to observe and study responses

Experimental Design – the assignment of treatments to individuals

Explanatory Variable – a variable that is believed to explain or cause changes in the response variable

F

Factor – an explanatory variable in an experiment

Frequency – how often a value occurs in a distribution

Fundamental Counting Principle – the number of ways to choose two objects, one from a set of N1 objects and one from a set of N2 objects is N1xN2. This can be extended to selecting an object from any number of sets.

G

Gap – space between clusters of data

Geometric distribution – a distribution skewed to the right that decreases at a constant rate

Graph – pictorial representation of data

Gray scale – shades of gray ranging from black to white

H

Histogram – A graph of frequency distribution of a quantitative variable

I

Incorrect response bias – bias that occurs because of individuals in a sample responding incorrectly

Independent events – the probability of one event is does not depend on whether the other event occurs.

Intercept – the point at which a line meets an axis.

Interquartile range (IQR) – the difference between the lower and upper quartiles

J

K

L

Latin square – an N x N array of numbers from 1 to N where each row and column contain distinct numbers

Law of Large Numbers – as more and more samples are taken from a population, the mean of the samples gets closer to the mean of the population.

Line graph – A graphs that shows time on one axis and data values on the other, with data points connect by line segments.

Lower Quartile – the number that divides smallest quartile from the rest of the numbers in the data set. It is the median of the lower half of the numbers.

Lurking variable – a variable that has an effect on the response variable but is not one of the explanatory variables

M

Maximum – largest value in a data set

Mean – the sum of the data values divided by the number of data items

Measurement bias – bias that occurs because of the errors in the measurement of samples

Median – for a data set with an odd number of values, the one that has the same number of values that are greater than it as those that are smaller than it. In a set with an even

number of values, it is the mean of the two numbers in the set that split the set into the same number smaller and larger values than the two numbers.

Minimum – smallest value in a data set

Mode – the value (or values) that occur most often in a data set.

Modified boxplot – a boxplot that shows the outliers as separate dots

Multiplication rule -- the probability of both of two independent events occurring is the product of the probabilities of the two events

N

Nonresponse bias – bias that occurs because some individuals in a sample fail to respond

Normal distribution – a symmetric distribution shaped like a bell.

O

Odds – the probability of an event divided by the probability of the event not occurring.

Orthogonal array – an K x N^2 array of numbers from 1 to N where in any pair of rows, the two numbers in the N^2 columns are all distinct.

Outcome – the result of an experiment

Outlier – A value widely separated from most of the data (1.5xIQR from lower or upper quartile).

P

Pascal's Triangle – a triangle numbers where the K^{th} number in row N represents the number of ways to choose a subset of K objects from a set of N objects

Payoff – the amount returned to the player who has bet on a specific outcome

Permutation – a particular order of a set of objects

Personal (subjective) probability – an estimate of the probability of an event made by a person from whatever personal knowledge he or she has

Pie Chart – a circular graph showing relative frequencies of categorical data using sectors in a circle for each category

Pixel – short for picture element, one smallest part of a digital image

Population – A set of people or things (units) that is being studied

Population size – number of individuals in a population

Positive/negative trend – upward/downward movement of data values

Probability – a number from 0 to 1 that represents the chance that an event will occur where 0 indicates impossibility and 1 represents certainty.

Probability model – a sample space together with an assignment of probabilities to its outcomes.

Q

Quantitative Variable – a variable whose value is numeric

Quartile – the three numbers that divide a data set into 4 equally sized groups

Questionnaire bias – bias that occurs because of the way a question is worded

R

Random sample, Simple random sample (SRS) -- a sample in which all individuals of a population are equally likely to part of

Randomization –assigning individuals to different treatment groups using a random selection.

Range – difference between the minimum and maximum of a data set

Recentering – adding a number to all values in a data set. This doesn't change the shape or spread of the data set.

Regression – the study of the relationship between two quantitative variables

Regression line -- the line which best matches the relationship between two variables

Relative Frequency – how often a value occurs in a distribution expressed as a proportion or percentage of the all values

Rescaling – Multiplying all values in a data set by the same positive number. This doesn't change the shape but multiplies the averages, IQR, and standard deviation by same number.

Response Variable – a variable the measure the outcome of a study

S

Saddle point – the outcome in a game when both player have a dominant strategy and play it.

Sample – a set of individuals in a population

Sample mean – the mean of a sample.

Sample selection bias, Sampling bias – sampling that results in numerical summaries or characteristics for the sample that do not reflect those of the population

Sample space – a complete list or description of outcomes of a chance process

Sampling distribution of mean – the distribution of the means of a set of samples

Sampling distribution of sample proportion – the distribution of the proportion of successes of a set of samples

Sampling with replacement –sampling where samples are chosen and then replaced before choosing the next sample

Sampling without replacement –sampling where samples are chosen and not replaced before choosing the next sample

Scatterplot – a graph showing the relationship between to quantitative variables

Shape – the general outline of a histogram of data

Simulation – using a probability model to simulate a real-life situation.

Size bias – bias related to the size of the individual

Skewed left/right – distribution where data forms a long tail to the left/right and is bunched up in the opposite direction

Slope – a ratio of the rate a line rises. It can be calculated from two points by computing the change in y divided by the change in x.

Spread – the degree to which values in a distribution differ.

Stacked bar graphs – a graph where categorical frequencies are represented as bars stack on top of each other

Standard deviation of population/sample – a measure of spread that is the square root of the variance

Standard error – the standard deviation of a sample

Standard normal distribution -- the normal distribution with mean 0 and standard deviation 1

Statistic –a summary statistic of a sample

Statistics, Exploratory data analysis – Investigating data through mathematical techniques to find patterns using tables and graphs to display and summarize information

Stem-leaf plot – A graphical representation of data where each value is split into the stem, all digits except the last, and the leaf, the final digit. The stems are listed, in order, vertically and are separated by a vertical line from the leaves. Leaves are listed, in order, horizontally across from their stems.

Strategy -- a set of moves which a player plans to follow while playing a game

Summary statistic – a single number used to describe data, often a measure of center or spread of the distribution

Survey – gathering data about a population through samples

Symmetric – a distribution that has a vertical line of symmetry

T

Tree diagram – a graphic representation of the fundamental counting principle

Trend – the overall upward or downward movement of data values over time

U

Uniform distribution – a distribution where all values have the same frequency

Upper Quartile – the number that divides largest quartile from the rest of the numbers in the data set. It is the median of the upper half of the numbers.

V

Variable – the characteristic that is being measured or classified.

Variance – a measure of spread defined as the sum of the squares of the differences from the mean divided by one less than the number of data items (for a population it is divided by the number of data items)

Venn diagram – a graphical diagram representing relationships between sets

Voluntary sample response – sampling a population by have volunteers determine the sample

W

X

Y

Z

Zero-sum game - A game in which players make payments only to each other. One player's loss is the other player's gain, so the total amount of "money" available remains constant.