Section 1: VISUALIZING DATA

One-way tables

Individuals: The set of elements (whether people or otherwise) that are surveyed to form a set of data about those individuals

Variables: Each property we collect in our data about the individuals

Data: The collection of individuals and variables

Data table: A table that organizes the data, including the individuals and their variables

Categorical Variables: Non-numerical variables, also called "qualitative" variables. Their values aren't represented with numbers.

Quantitative variables: Numerical variables. Their values are numbers.

Discrete variables: Variables we can obtain by by counting. Therefore, they can take on only certain numerical values.

Continuous variables: Variables that can include data such as decimals, fractions, or irrational numbers.

Nominal scale of measurement: Things like favourite food, colors, names, and 'yes' or 'no' responses have a nominal scale of measurement. Only categorical data can be measured with a nominal scale.

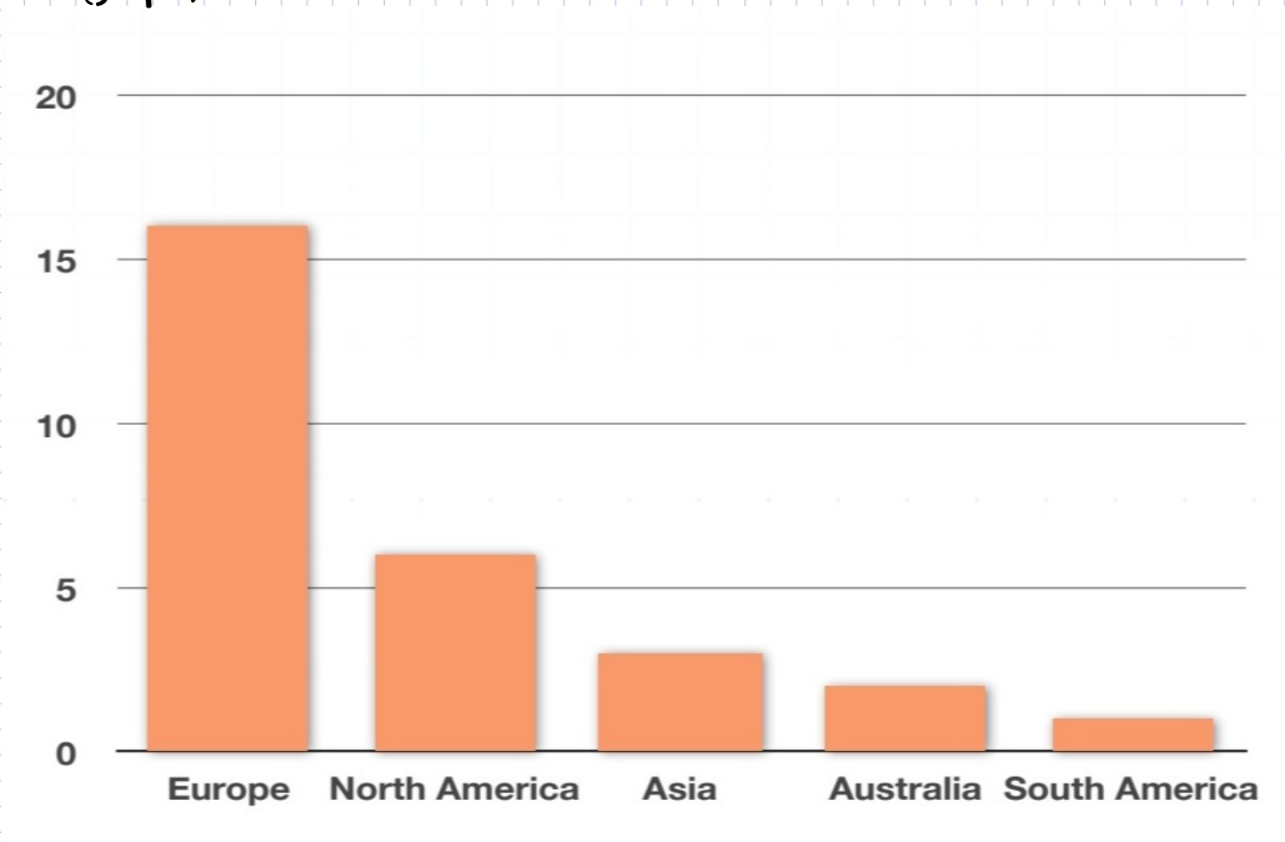
Ordinal scale of measurement: Categorical data can also be ordinal. This type of data can be ordered.

Interval scale of measurement: Data scaled using an interval scale can be ordered like ordinal data. But interval data also gives us a known interval between measurements.

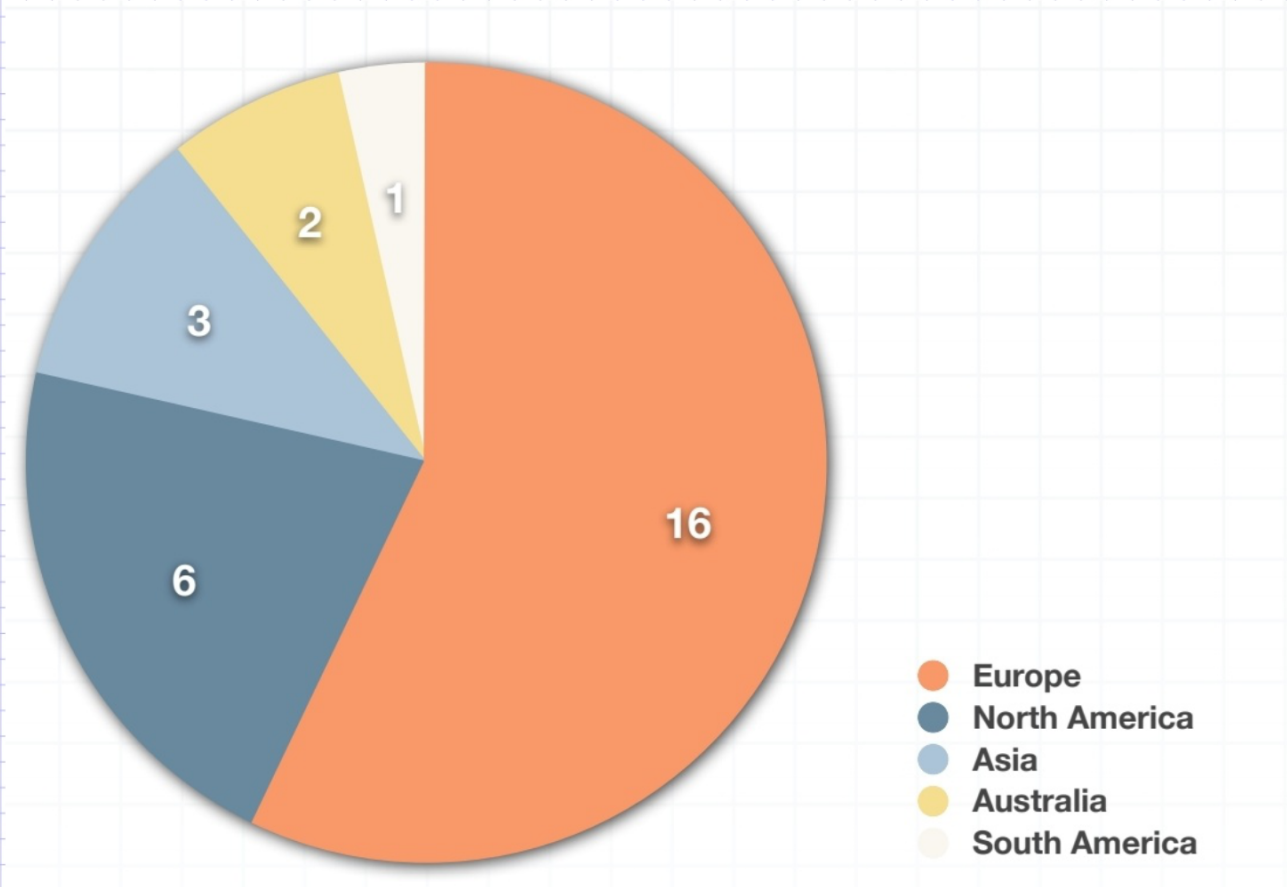
Ratio scale of measurement: Data measured using a ratio scale is just like interval scale data, except that ratio scale data has a starting point, or absolute zero.

Bar graphs and pie charts

Bar graph, bar chart:



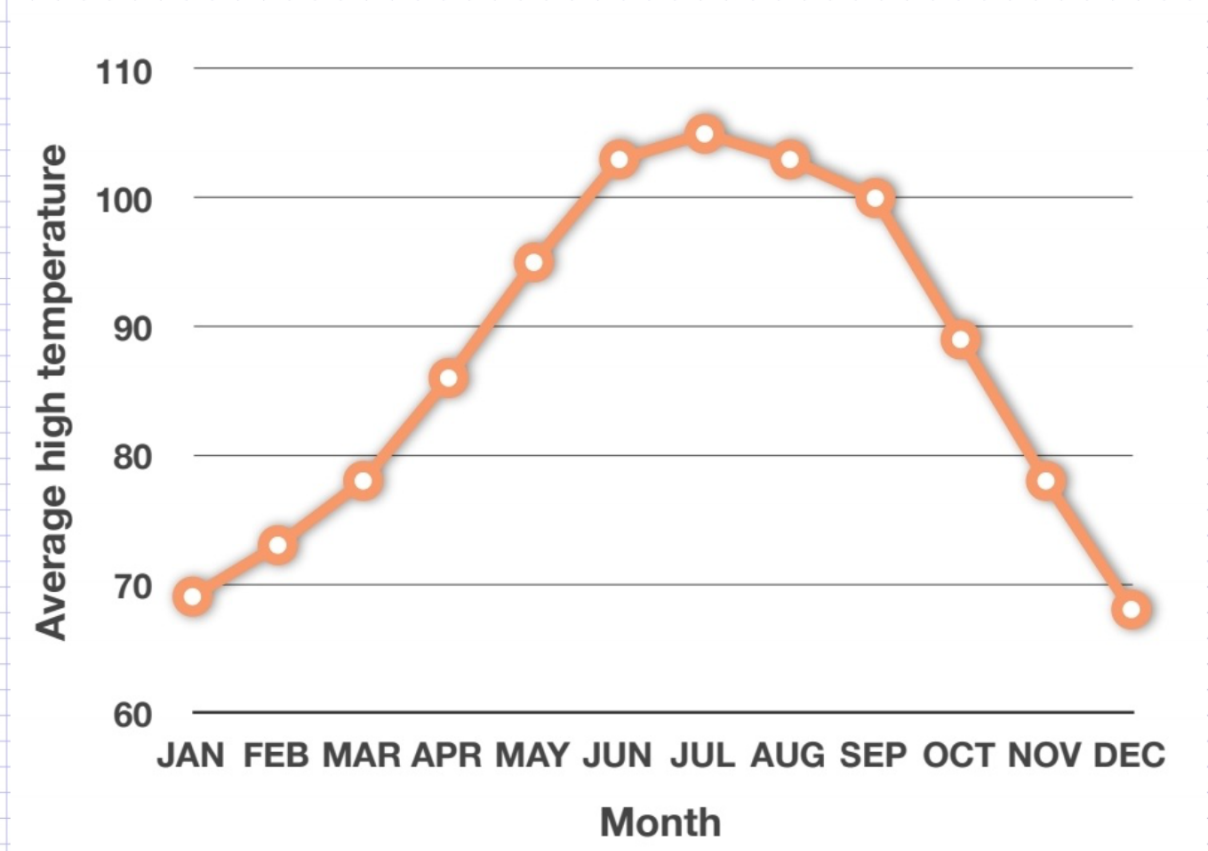
Pie chart



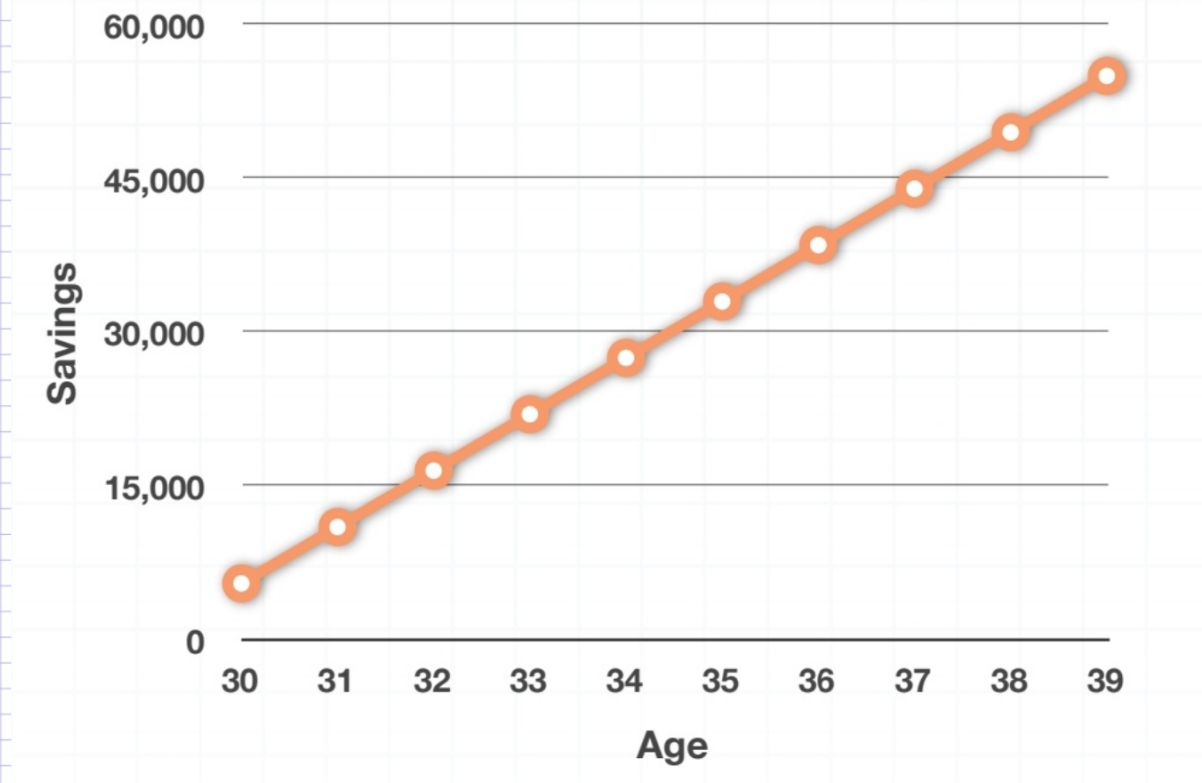
Frequency table: A summary table that shows the frequency or count of each categorical variable.

Line graphs and ogives

Line graph:

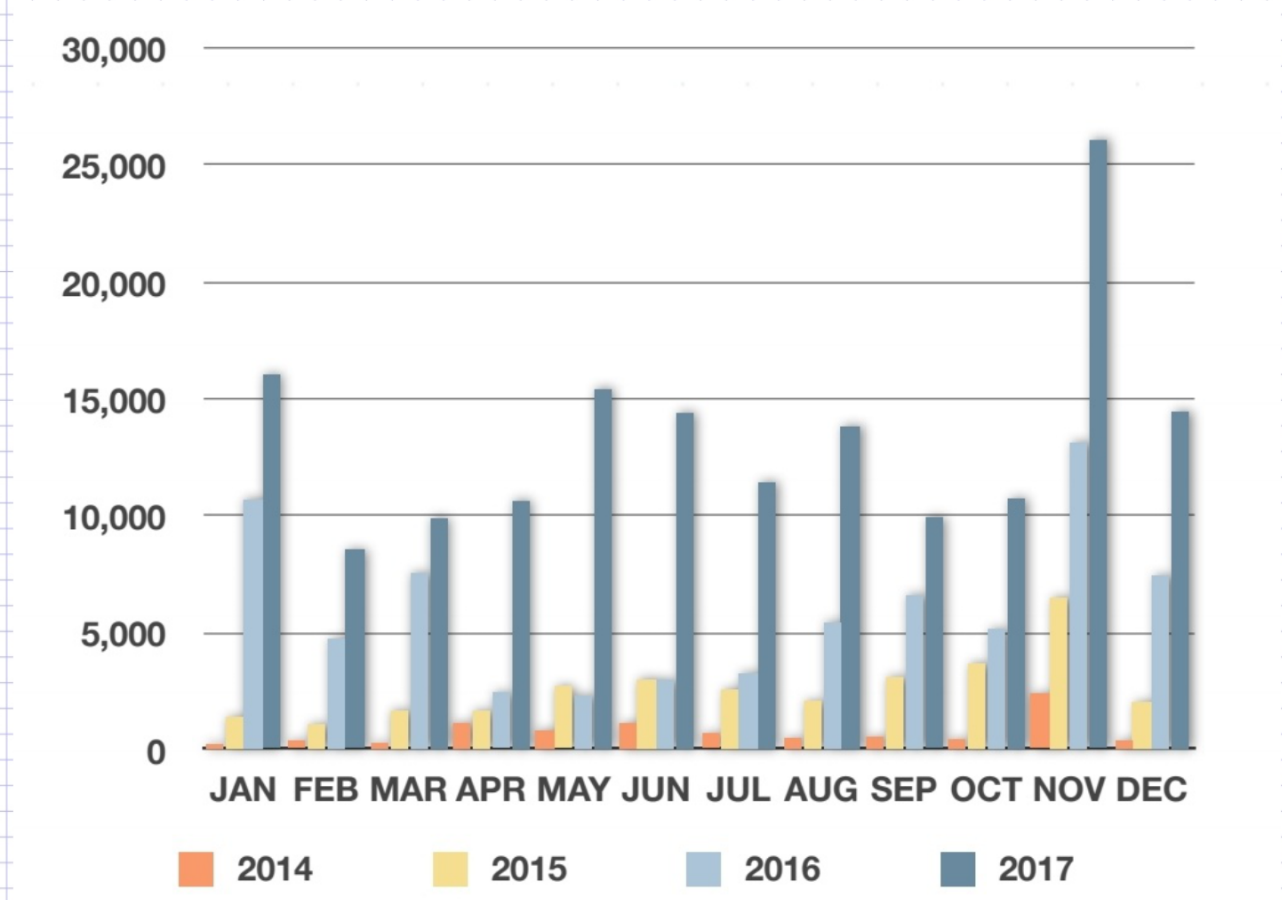


Ogive:

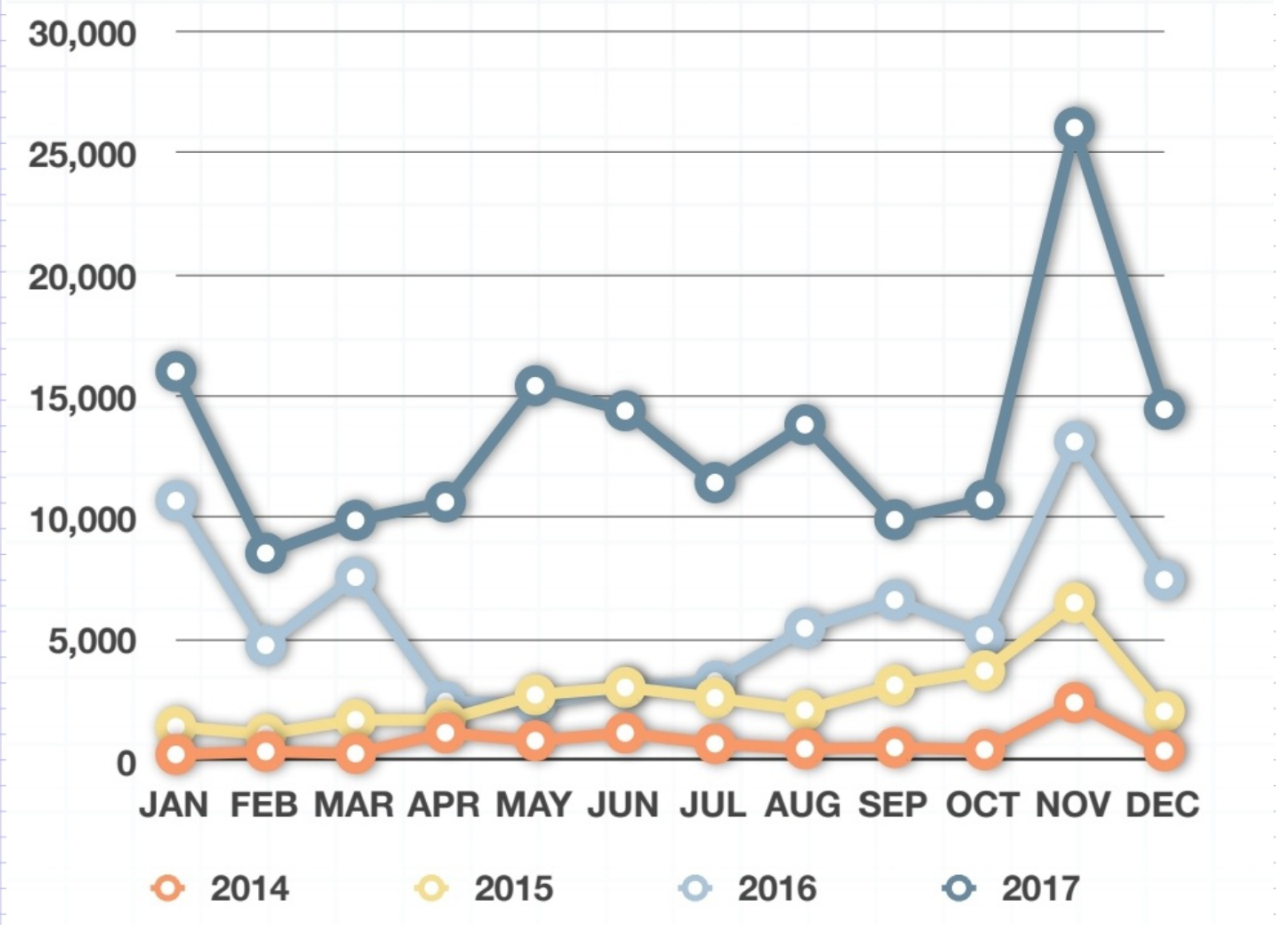


Two-way tables

Comparison bar graph:

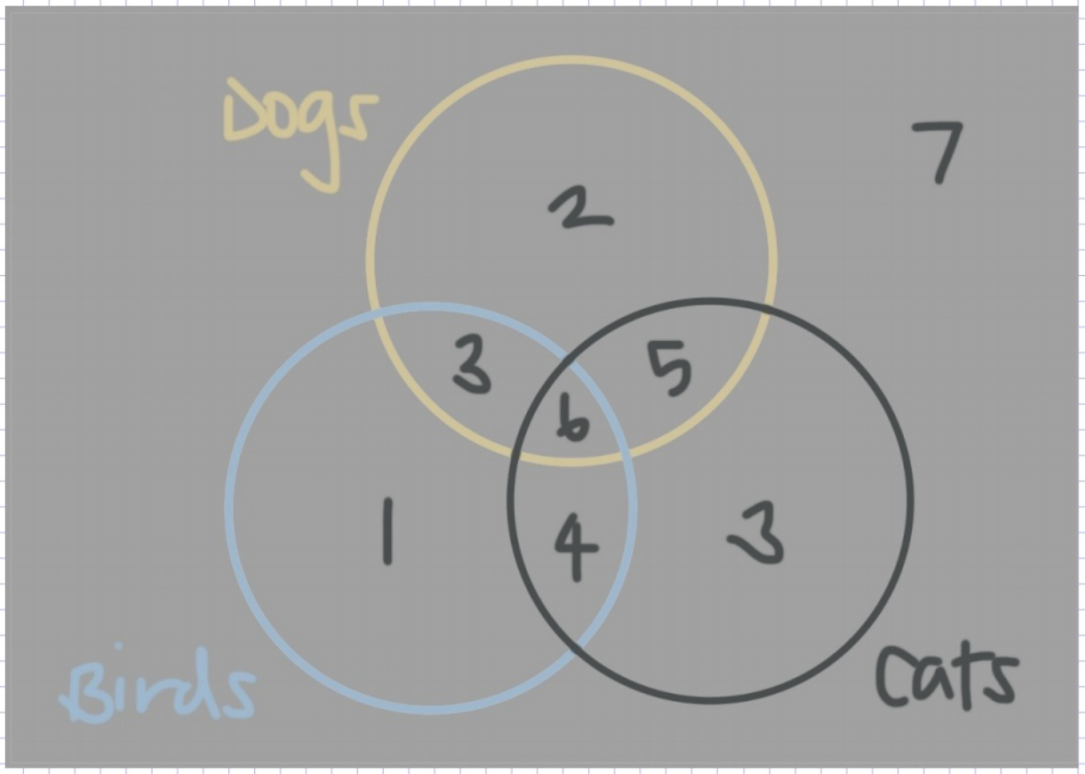


Comparison line graph:



Venn diagrams

Venn diagram:



Frequency tables and dot plots

Frequency table: A table that displays how frequently or infrequently something occurs

Dot plot: A plot that can be used to show the frequency of small data sets.

Relative frequency tables

Relative frequency table: A table that shows percentages instead of actual counts

Column-relative frequency table: A relative frequency table in which the columns sum to 1.00 but the rows do not

Row-relative frequency table: A relative frequency table in which the rows sum to 1.00 but the columns do not.

Total-relative frequency table: A relative frequency table in which both the rows and columns sum to 1.00. A grand total cell is also included, which is always 1.00.

Joint distributions

Joint distribution: A table of percentages similar to a relative frequency table. The difference is that, in a joint distribution we show the distribution of one set of data against the distribution of another set of data.

Marginal distribution: The total column or the total row in a joint distribution.

Conditional distribution: The distribution of one variable given a particular value of the other variable.

Histograms and stem-and-leaf plots

Histogram: Also called a frequency histogram, a histogram is like a bar graph, except that we collect the data into equally-sized buckets or bins, and then sketch a bar for each bucket. Histograms have no gaps between the bars, because they represent a continuous dataset.

Stem- and leaf plots: Also called a stem plot, a stem-and-leaf plot groups data together based on the first digit(s) in each number. Stems are the numbers on the left, and leaves are the numbers on the right.

Building histograms from data sets

Class interval, class, bin: An individual grouping bucket within a histogram

Class width: The interval over which the class is defined

Class midpoint: The value halfway between the lower and upper edges of the class

How to build a histogram:

1. Put the data set in ascending order, then find the range as the difference between the largest and the smallest values.
2. Determine the number of bins, or classes, that we want to have in our histogram. As a rule of thumb, it's best to use 5-6 classes for most of the data. However, we might want to use up to 20 classes when we deal with larger data sets. It all depends on how large our dataset is and the number of classes would best represent data.
3. Divide the range by the number of classes, the round up the result to get the class width.
4. Build a table, putting each class in a separate row.
5. Find the frequency for each class by counting the data points that fall into each one. It is worth mentioning that we can choose either overlapping intervals or non-overlapping intervals in the previous step. However, for overlapping intervals like 0-4 and 4-8 the data point 4 should be included in the second interval. In other words, the interval 0-4 actually contains data from 0 to $3.\overline{9}$, and the interval 4-8 contains data from 4 to $7.\overline{9}$. Overlapping intervals are particularly useful when the data set contains decimals.
6. Graph the histogram by placing the classes along the horizontal axis and their frequencies along the vertical axis, such that the height of each bar is the frequency of each class.

Section 2: ANALYZING DATA

Measures of central tendency

Measures of central tendency: Different ways we've come up with to describe the 'middle', 'center', or most typical value of data

Mean, arithmetic mean: The balancing point of the data

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\text{the sum of all data points}}{\text{the number of data points}}$$

Median: The value at the middle of the data set when we line up all the data points in order from least to greatest

Mode: The value in the data set that occurs most often

Measures of Spread

Spread, dispersion, scatter: How, and by how much, our data set is spread out around its center

Range: The difference the largest value and smallest value

Quartiles: The values that mark the 25^{th}, 50^{th}, 75^{th}, and 100^{th} percentiles of the data, which are Q_1, Q_2, Q_3, and Q_4, respectively

Interquartile range (IQR): The difference between the first and third quartiles, $Q_3 - Q_1$

Changing the data and outliers

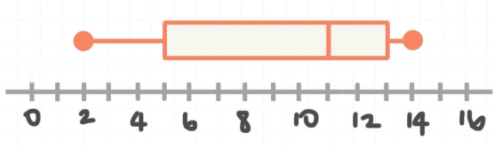
Shifting: Adding or subtracting a value from every point in data set. Shifting changes the mean, median and mode, but not the range or IQR.

Scaling: Multiplying or dividing a value from every point in data set. Scaling changes the mean, median, mode, range and IQR.

Outlier: A number on the extreme upper or extreme lower end of a data set.

Box-and-whisker plots

Box-and-whisker plots, box plots: A useful plot for representing the median and spread of the data at the same time. The box plotted in center extends from Q_1 and Q_3, and the whiskers extend beyond the box to the lower and upper ends of the range of the data.



Five-number summary, five-figure summary: A summary table that includes the minimum and maximum values, the median, and Q_1 and Q_3 for the data set.

Min	Q_1	Median	Q_3	Max
2	5	11	13	14

Section 3: DATA DISTRIBUTIONS

Mean, variance and standard deviation

Population: The entire group of subjects that we're interested in

Sample: A sub-section of the population

Population mean:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Sample mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Variance:

- for a population:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

- for a sample (biased):

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

- for a sample (unbiased):

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Standard deviation:

- for a population:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

- for a sample:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Frequency histograms and polygons, and density curves

Histograms, frequency histogram: Shows the frequency at which each category occurs

Relative frequency histogram: The same as a regular histogram, except that we display the frequency of each category as a percentage of the total of the data

Frequency polygon: The shape created by connecting the top of each bar in a frequency histogram

Density curve: The shape created by smoothing out the frequency polygon. The area under the density curve will always represent %100 of the data or 1.0

Symmetric and skewed distributions and outliers

Distribution: The distribution of data across a range of values.

Symmetric distributions: The distribution's mean and median are at the very center of the distribution, with an equal about of data to the left and right.

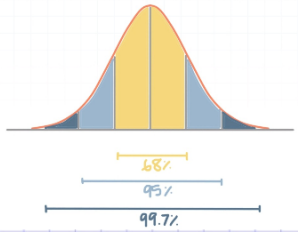
Normal distribution: A symmetric, bell-shaped distribution

Skewed distribution: A non-symmetric distribution that leans right and left. Negatively / left-skewed / left-tailed distributions have their tail on the left, and, from left-to-right, they have their mean, then median, then mode. Positively / right-skewed / right-tailed distributions have their tail on the right, and, from left-to-right, they have their mode, then median, then mean.

1.5-IQR rule: Low outliers are defined as values less than $Q_1 - 1.5(IQR)$, while high outliers are defined as values greater than $Q_3 + 1.5(IQR)$.

Normal distributions and z-scores

Empirical rule, 68-95-99.7 rule: For any normal distribution, there's a 68% chance, 95% chance, and 99.7% chance a data point falls within 1, 2, and 3 standard deviations of the mean, respectively



Percentile: n percent of the values in the data set lie below the n^{th} percentile of the data set

z-score: The z-score for a data point x is the score that tells us the number of standard deviations between x and the mean of μ. A data point is generally considered unusual if its z-score is $z = \pm 3$.

$$z = \frac{x - \mu}{\sigma}$$

Threshold: Some pre-defined cutoff point in the data set

Chebyshev's Theorem

Chebyshev's Theorem: At least $(1 - 1/k^2)\%$ of the data must fall within k standard deviations of the mean, for $k > 1$, regardless of the shape of the data's distribution. Because this theorem applies to distributions of all shapes, it's more conservative than the Empirical Rule.

- At least 75% of the data must be within $k = 2$ standard deviations of the mean.
- At least 89% of the data must be within $k = 3$ standard deviations of the mean.
- At least 94% of the data must be within $k = 4$ standard deviations of the mean.

Covariance

Covariance: How much two random variables vary together. Covariance reflects the directional relationship between two random variables, but not the magnitude of the relationship.

Population covariance

$$\sigma_{xy} = \frac{\sum_i (x_i - \mu_x)(y_i - \mu_y)}{N}$$

Sample covariance

$$s_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Direction of covariance:

- **Positive covariance:** Positive linear relationship between the variables
- **Approximately 0 covariance:** No linear relationship between variables
- **Negative covariance:** Negative linear relationship between the variables

Correlation coefficient

Correlation: The degree of the relationship between variables. If two variables are correlated, then a change in one variable results in a change in the other variable.

- **Perfectly correlated:** A specific change in one variable results in an exact and predictable change in the other variable
- **Somewhat correlated:** A change in one variable results in a predictable change in some general direction in the other variable
- **Uncorrelated:** A change in one variable doesn't result in any kind of predictable change in the other variable

Pearson correlation coefficient:

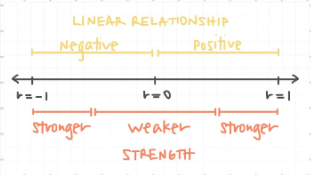
Population

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\left(\frac{\sum_i (x_i - \mu_x)(y_i - \mu_y)}{N}\right)}{\sigma_x \sigma_y}$$

Sample

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\left(\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n-1}\right)}{s_x s_y}, \quad \text{where } s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} \text{ and } s_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}$$

Value of the coefficient: The correlation coefficient indicates both the strength and direction of the relationship between two variables.



- Any value of the correlation coefficient between -0.7 and -1 (or 0.7 and 1) indicates a strong negative (or positive) correlation.
- Any value of the correlation coefficient between -0.3 and -0.7 (or 0.3 and 0.7) indicates a moderate negative (or positive) correlation.
- Any value of the correlation coefficient between 0 and -0.3 (or 0 and 0.3) indicates a weak negative (or positive) correlation.

Weighted means and grouped data

Weighted mean

- Weighted population mean:

$$\mu = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}$$

- Weighted sample mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Grouped mean

- Grouped population mean:

$$\mu = \frac{\sum_{i=1}^{N} f_i M_i}{N}$$

- Grouped sample mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i M_i}{n}$$

Grouped variance

- Grouped data population variance:

$$\sigma^2 = \frac{\sum_{i=1}^{N} f_i (M_i - \mu)^2}{N}$$

- Grouped data sample variance:

$$s^2 = \frac{\sum_{i=1}^{n} f_i (M_i - \bar{x})^2}{n-1}$$

Section 4: PROBABILITY

Simple probability

Probability: How likely it is that some event will occur. All probabilities are numbers equal to or between 0 and 1. This formula applies when all possible outcomes are equally likely.

$$P(\text{event}) = \frac{\text{outcomes that meet our criteria}}{\text{all possible equally likely outcomes}}$$

Sample space: The collection of 'all possible outcomes' from the denominator of the simple probability formula

Experiment: One event in the sample space

Experimental / empirical probability: The probability we find when we run experiments. Experimental probability changes as we run experiments over time. If the experiment is a good one, the experimental probability should get very close to the theoretical probability as we run more and more experiments

Theoretical / classical probability: The probability that an event will occur, based on an infinite number of experiments. This is the probability we get from the simple probability formula.

Law of large numbers: This law tells us that if we could run an infinite number of experiments, the experimental probability would eventually equal the theoretical probability.

The addition rule, and union vs intersection

Event: A specific collection of outcomes from the sample space

Addition rule, sum rule:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Mutually exclusive, disjoint events: Events that can't both occur. In this case, $P(A \text{ and } B) = P(A \cap B) = 0$. Therefore, for mutually exclusive events, the addition rule simplifies to $P(A \text{ or } B) = P(A) + P(B)$, or $P(A \cup B) = P(A) + P(B)$.

Union of events: $P(A \cup B)$ is the union of A and B, and it means the probability of either A or B or both occuring.

Intersection of events: $P(A \cap B)$ is the intersection of A and B, and it means the probability of A and B both occuring.

Independent and dependent events and conditional probability

Independent events: Events that don't affect one another, like two separate coin flips

Multiplication rule:

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

Dependent events: Events that affect one another, like pulling two cards from a deck without replacing the first card before pulling the second

Conditional probability: The probability that multiple dependent events occur

Bayes' Theorem

Bayes' Theorem / Law / Rule: Tells us the probability of an event, given prior knowledge of related events that occured earlier. To solve problems with Bayes' Theorem, write out what you know, build a tree diagram that includes all possibilities, and then 'trim the branches' of your tree that aren't relevant to the question being asked.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Section 5: DISCRETE RANDOM VARIABLES

Discrete Probability

Discrete random variable: A variable that can only take on discrete values

Continuous random variable: A variable that can take on any value in a certain interval

Expected value: The mean of a discrete random variable

Variance of discrete random variable:

$$\sigma_x^2 = \sum_{i=1}^{n} (X_i - M)^2 P(X_i)$$

Transforming random variables

- Shifting the data set doesn't affect standard deviation
- Scaling the data set by k scales standard deviation by k

Combinations of random variables

Mean and variance of the combination of normally distributed variables:

	Combination	Mean	Variance
Sum	$S = X + Y$	$\mu_S = \mu_X + \mu_Y$	$\sigma_S^2 = \sigma_X^2 + \sigma_Y^2$
Difference	$D = X - Y$	$\mu_D = \mu_X - \mu_Y$	$\sigma_D^2 = \sigma_X^2 - \sigma_Y^2$

Permutations and Combinations

Permutation: The number of ways we can arrange a set of things, and the order of the arrangement matters.

$${}_nP_k = \frac{n!}{(n-k)!}$$

Combination: The number of ways we can arrange a set of things, but the order of the arrangement doesn't matter

$${}_nC_k = \frac{n!}{k!(n-k)!}$$

Binomial random variables

Binomial variable: A variable that can take on exactly two values, like a coin flip. In order for a variable X to be a binomial random variable,

- each trial must be independent,
- each trial can be called a "success" or "failure",
- there are a fixed number of trials, and
- the probability of success on each trial is constant.

Binomial probability:

$$P(k \text{ successes in } n \text{ attempts}) = \binom{n}{k} p^k (1-p)^{n-k}$$

Poisson distributions

Poisson process: Calculates the number of times an event occurs in a period of time, or in a particular area, or over some distance, or within any other kind of measurement.

1. The experiment counts the number of occurences of an event over some other measurement,
2. The mean is the same for each interval,
3. The count of events in each interval is independent of the other intervals,

4. The intervals don't overlap, and
5. The probability of the event occuring is proportional to the period of time.

Poisson probability:

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Poisson probability for a binomial random variable:

$$P(x) = \frac{(np)^x e^{-np}}{x!}$$

"At least" and "at most", and mean, variance and standard deviation

Probability of at least one succes or failure:

$$P(\text{at least 1 success}) = 1 - P(\text{all failures})$$

$$P(\text{at least 1 failure}) = 1 - P(\text{all successes})$$

Mean, variance, and standard deviation of a binomial random variable:

Mean: $\mu_x = E(X) = np$

Variance: $\sigma_x^2 = np(1-p)$

Standard deviation: $\sigma_x = \sqrt{np(1-p)}$

Bernoulli random variables

Bernoulli random variable: A special category of binomial random variables, with exactly one trial, in which "success" is defined as a 1 and "failure" is defined as a 0

Mean, variance, and standard deviation of a Bernoulli random variable:

Mean: $\mu = p$

Variance: $\sigma^2 = p(1-p)$

Standard deviation: $\sigma = \sqrt{p(1-p)}$

Geometric random variables

Geometric random variable: We run an infinite number of trials until we get some defined "success".

- Each trial must be independent,
- Each trial can be called a "success" or "failure", and
- The probability of success on each trial is constant.

Probability of success on the n^{th} attempt:

$$P(S=n) = p(1-p)^{n-1}$$

Mean, variance, and standard deviation of a geometric random variable:

Mean: $\mu_x = E(X) = 1/p$

Variance: $\sigma_x^2 = \frac{1-p}{p^2}$

Standard deviation: $\sigma_x = \sqrt{\frac{1-p}{p^2}}$

Section 6: SAMPLING

Types of studies

Statistic: Mean, standard deviation, proportion, etc., for a sample

Parameter: Mean, standard deviation, proportion, etc., for a population. We use statistics to estimate corresponding parameter values.

Observational studies: In an observational study, we're just looking at the information that's already there, or measuring it in some way, but we're adding nothing to the population that will change it in any way.

Treatment: Something that changes a population

Correlation: Two variables are correlated when they move together predictably. The variables are positively correlated when they increase together or decrease together. Variables are negatively correlated when they increase and decrease in opposite directions: one goes down while the other goes up, or one goes up while the other goes down.

Causation: One variable causes another variable to change. Showing correlation doesn't prove causation.

Confounding variable: A third variable that leads to both of the variables that were correlated

Control group: The group that does nothing, receives nothing, or isn't manipulated

Treatment / experimental group: The group that does something, receives something, or is treated in some way

Explanatory and response variables: In an experiment, we're looking to see whether one or more explanatory variables (the treatment) has an effect on the response variable (whatever is expected to be effected).

Blind experiment: When the participants don't know whether they're in the control group or the treatment group

Double-blind experiment: When neither the participants nor the people administering the experiment know which group anyone is in

Blocking: When researchers separate participants into like groups

Matched-pairs experiment: A more specific kind of blocking where we make sure that the participants in our experimental group and control group are matched based on similar characteristics

Sampling and bias

Representative / unbiased sample: When the sample data "scales up" to the population, or when the sample does a good job representing the population

Bias: When something skews our results and makes them inaccurate

Measurement bias: When there's something wrong with the tool we're using to collect the data, so our method of collecting observations or responses from the sample results in false values

Social desirability bias: When our survey asks something in a way that discourages people from responding truthfully

Leading questions: Questions that are framed in a way that push respondents toward a particular response

Selection bias, undercoverage: When we don't collect data from an entire group of subjects that should have been included in our data

Voluntary response sampling: When people voluntarily respond to or participate in the study

Convenience Sampling: When we choose a sample simply because it's convenient, instead of prioritizing getting a good, random representative sample

Non-response bias: When we get a large number of people who don't respond to our survey

Simple random sample: When we assign subjects to groups in a totally random way

Stratified random sample: When we put some parameter on the sample where we require an even number of subjects from different groups

Clustered random sample: Where we break our population into clusters, and then either

1) take a random sample within each cluster to be our total sample, or
2) randomly pick some clusters an then sample everyone in those clusters

Systematic sampling: When we assign numbers to individuals in a population and choose them at some specified interval