Probability & Statistics Formulas

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 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 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 favorite 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: Categorial data can also be ordinal. This type of data can be ordered.

Interval scale of measurement: Data measured 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.
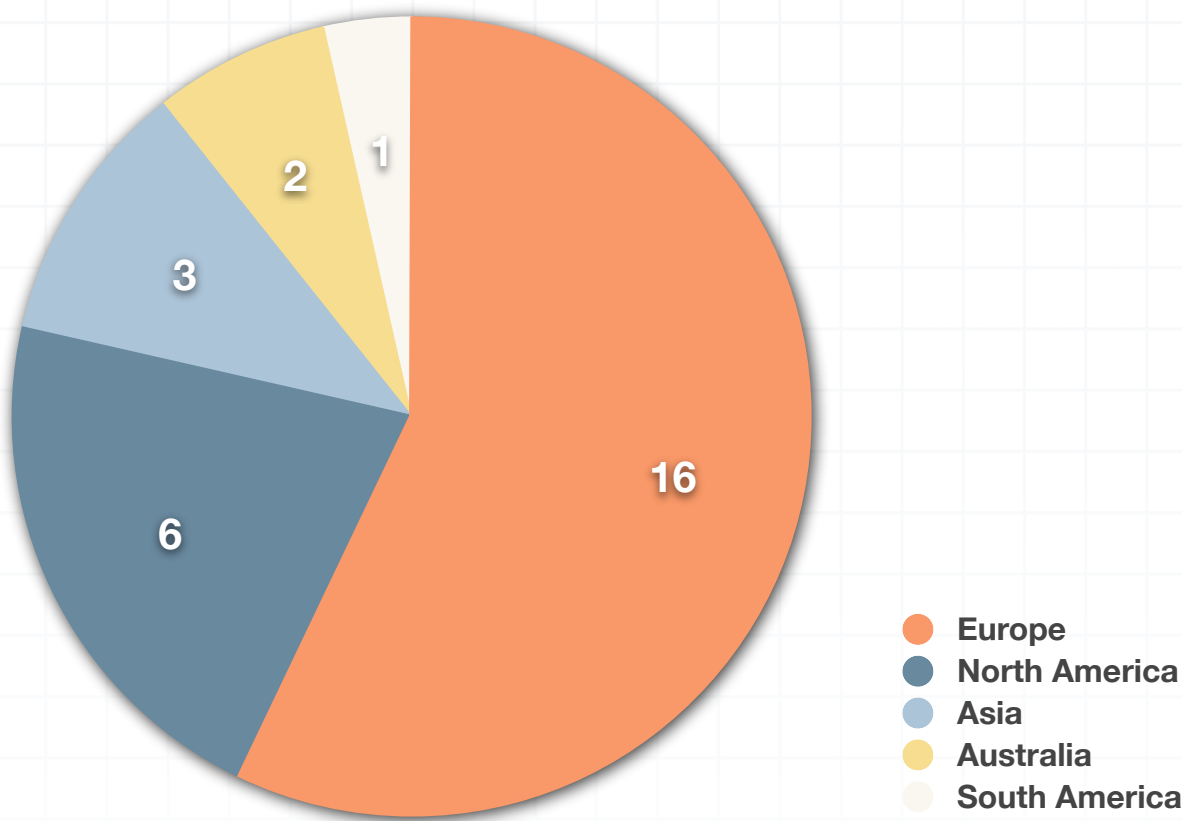
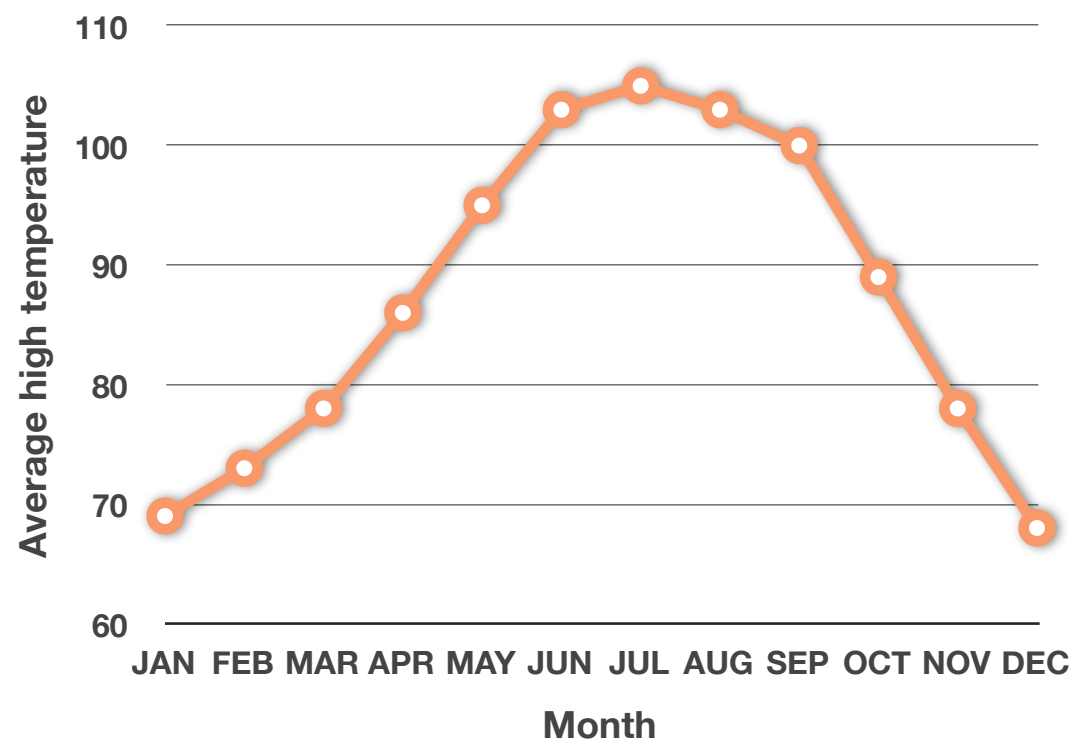## Bar graphs and pie charts

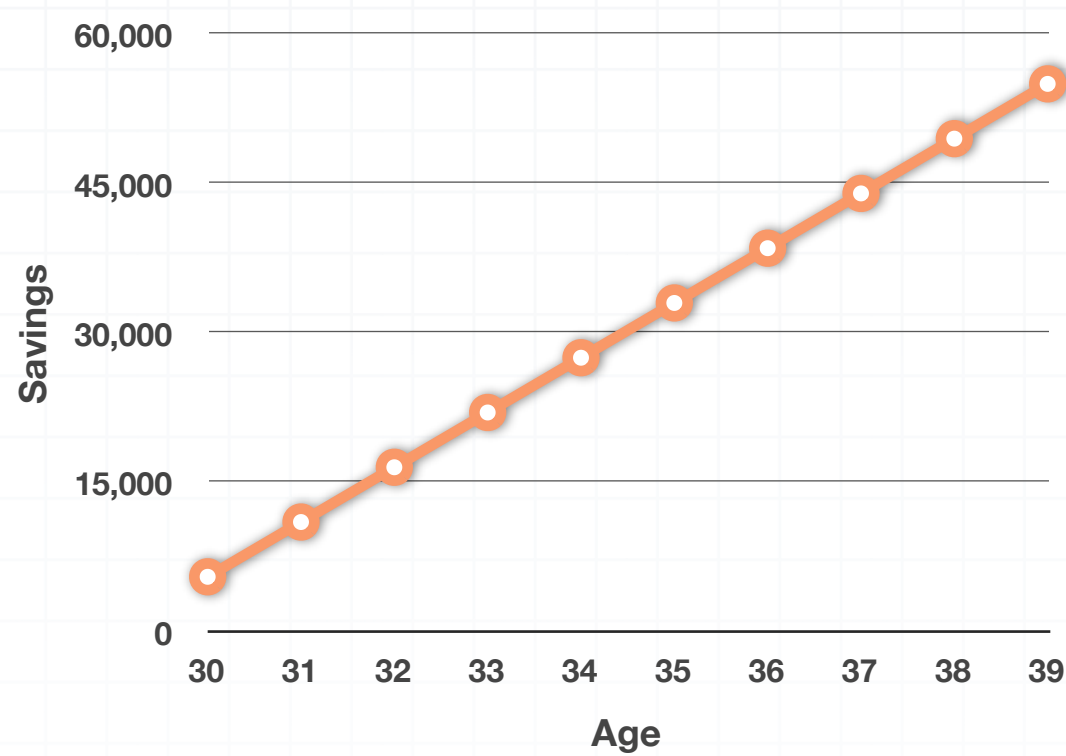Bar graph, bar chart:
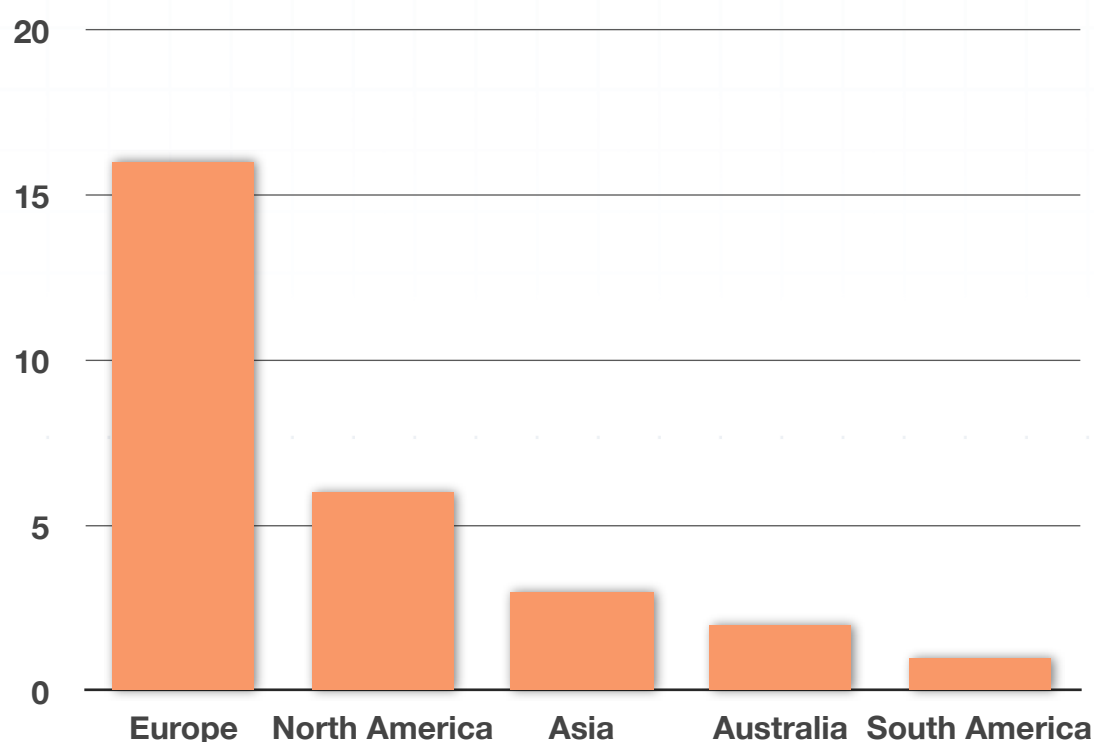
Pie chart:

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:

2014   2015   2016   2017
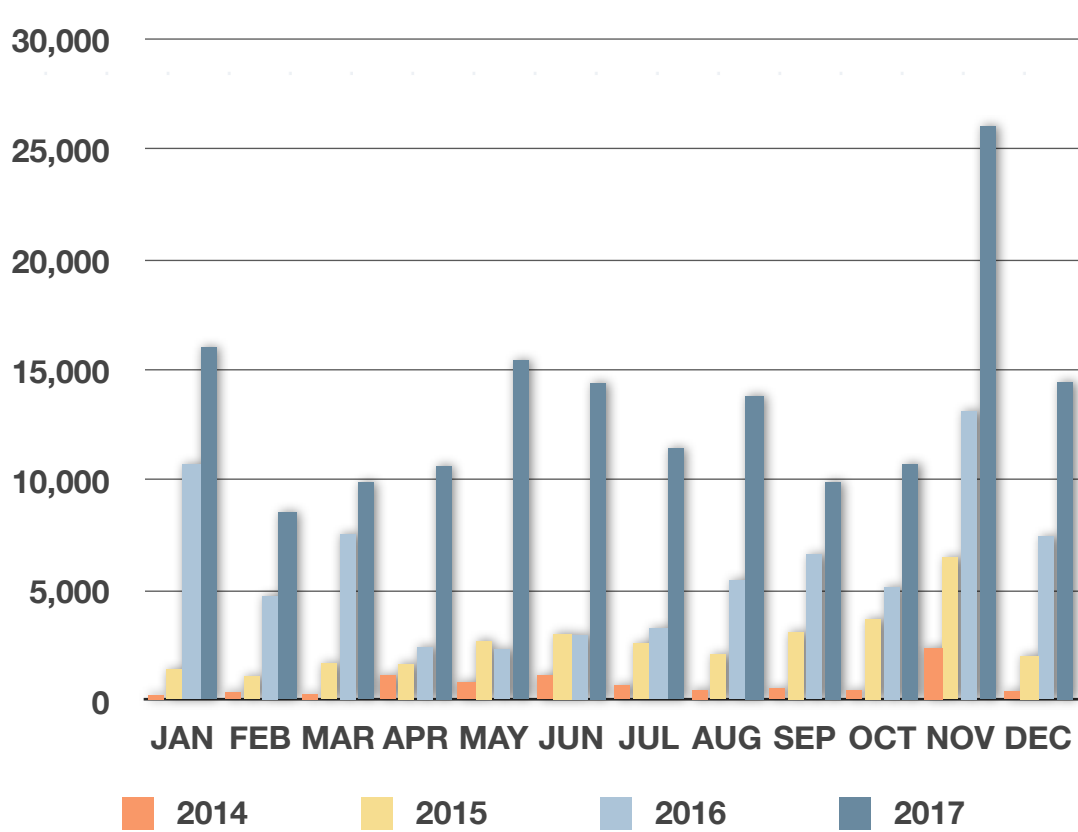
Comparison line graph:

| | 2014 | | 2015 | | 2016 | | 2017 |
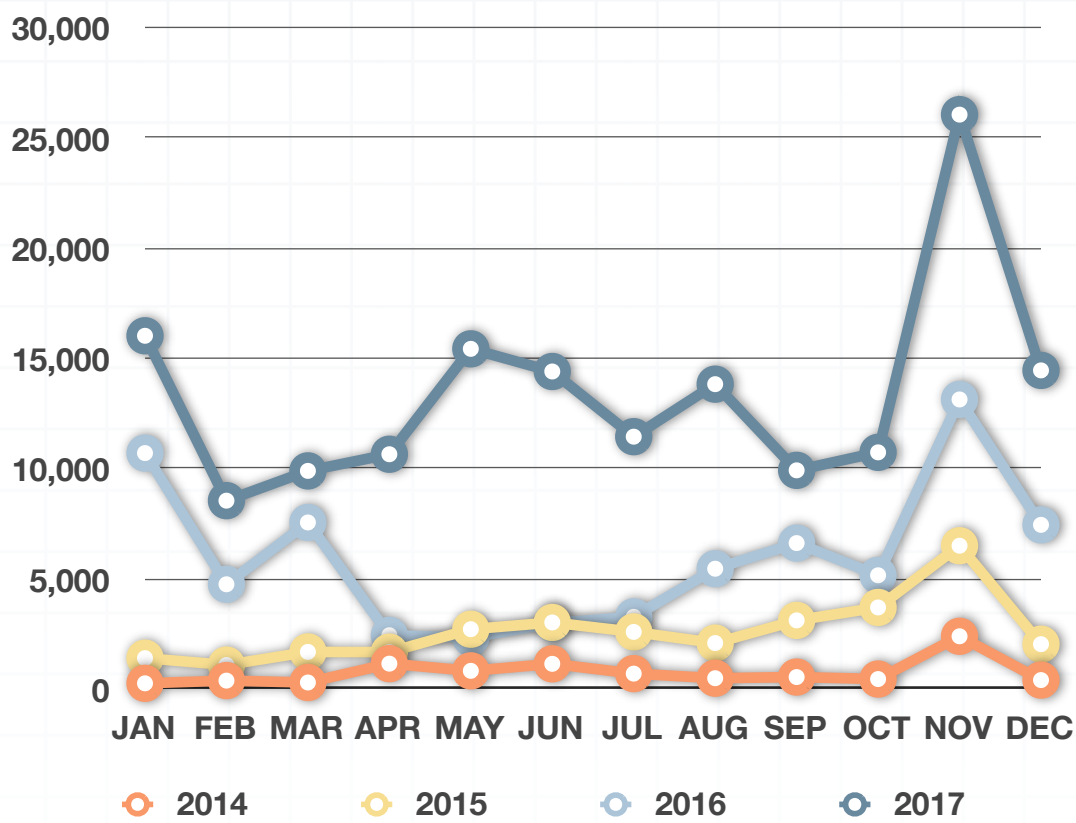
Venn diagrams

Venn diagram:
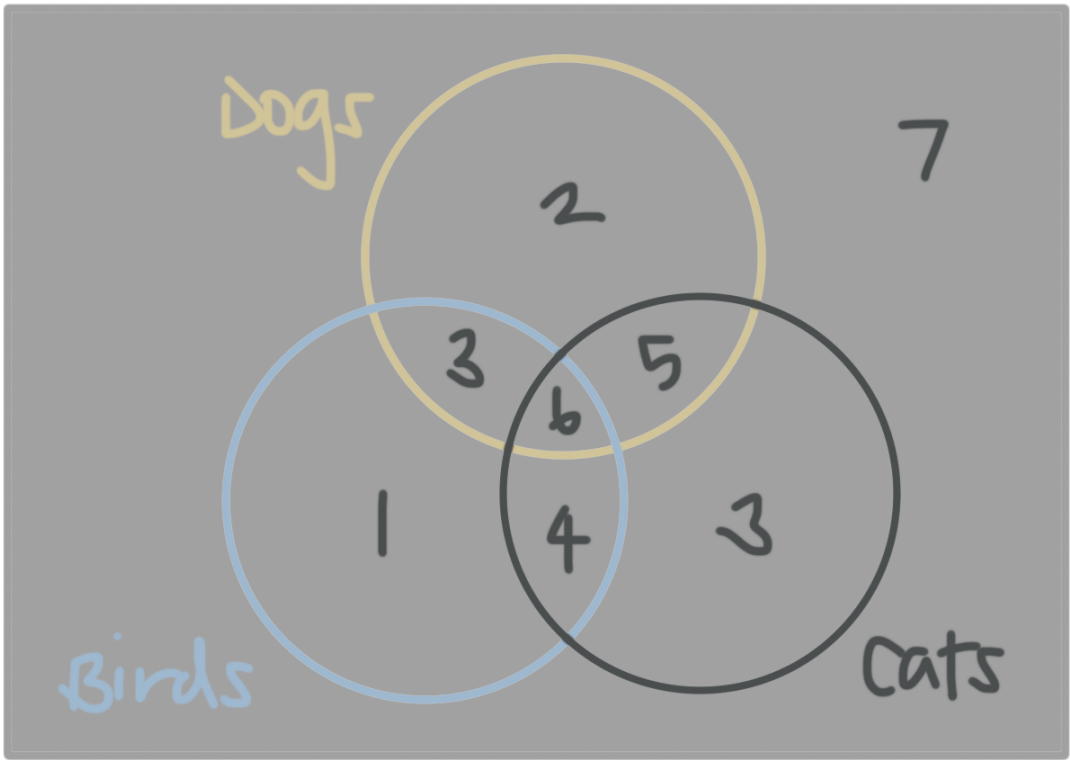
Frequency tables and dot plots

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
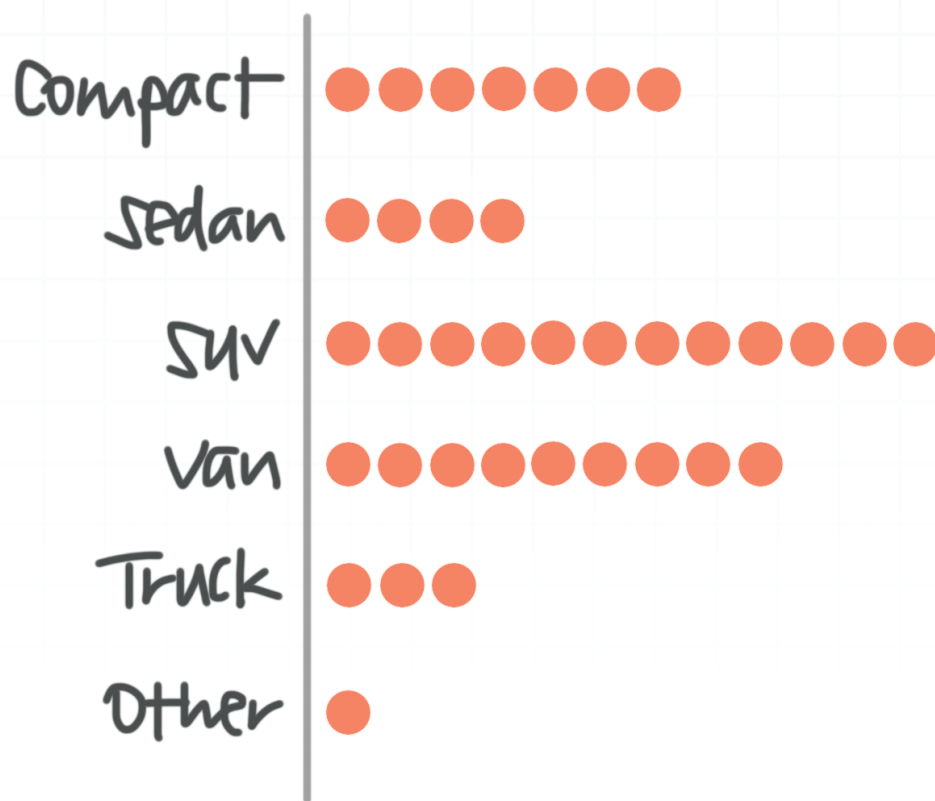
Relative frequency tables

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 data set.

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 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 we'll work with during our statistical studies. However, we might want to use up to 20 classes when we deal with larger data sets. It all depends on how large our data set is and the number of classes that would best represent the data.

3. Divide the range by the number of classes, then 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's 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 $3.999...$, and the interval $4 - 8$ contains data from 4 to $7.999...$ 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.

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 the data

Mean, arithmetic mean: The balancing point of the data

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

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 between the largest value and smallest value

Quartiles: The values that mark the 25th, 50th, 75th, and 100th 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 the 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 the data set. Scaling changes the mean, median, mode, range, and IQR.

Outlier: A number on the extreme upper end 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 to Q_3, and the whiskers extend beyond the box to the lower and upper ends of the range of the data.
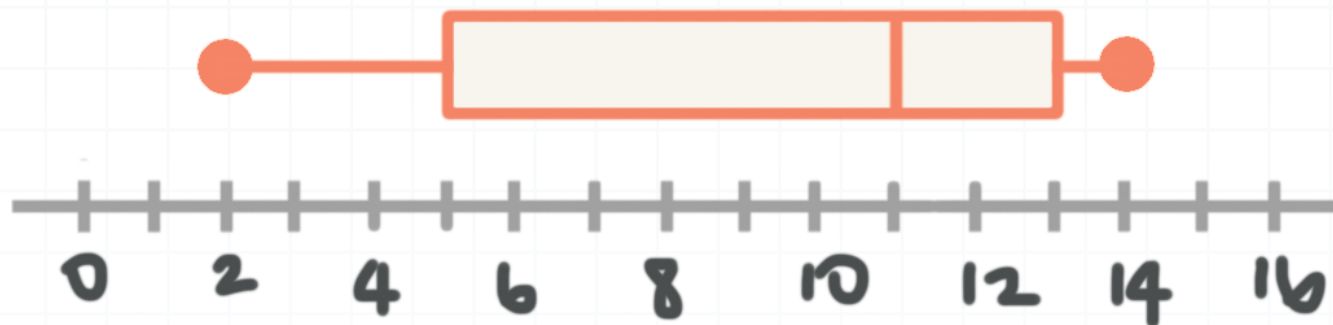
Five-number summary, five-figure summary: A summary table that includes the minimum and maximum values, the median, and Q_1 and Q_3 for the data set

Min	Q_1	Median	Q_3	Max
2	5	11	13	14

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

For a sample (biased): $s^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$

For a sample (unbiased): $s_{n-1}^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

Standard deviation:

For a population: $\sigma = \sqrt{\sigma^2} = \sqrt{\dfrac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$

For a sample: $s = \sqrt{s_{n-1}^2} = \sqrt{\dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$

Frequency histograms and polygons, and density curves

Histogram, 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 a 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 distribution: 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 or 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 $Q_1 - 1.5(\text{IQR})$, while high outliers are defined as values greater than $Q_3 + 1.5(\text{IQR})$.

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

68%

95%

99.7%
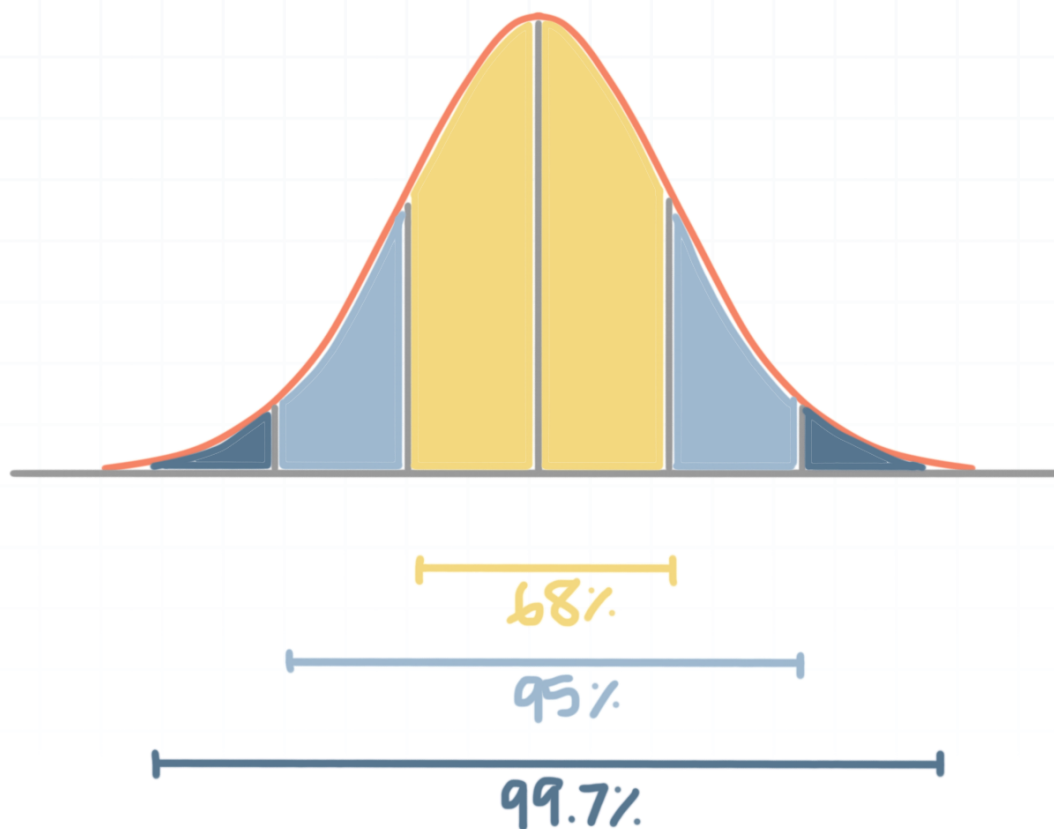
Percentile: n percent of the values in the data set lie below the nth percentile of the data set

z-score: The z-score for a data point x is the score that tells us the number of standard deviations between 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.

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 and 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 occurring.

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

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 occurred 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.

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

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 a discrete random variable:

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

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 occurrences 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 occurring 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 success 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 nth attempt:

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

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

Mean: $\mu_X = E(X) = \dfrac{1}{p}$

Variance: $\sigma_X^2 = \dfrac{1-p}{p^2}$

Standard deviation: $\sigma_X = \sqrt{\dfrac{1-p}{p^2}}$

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 and then sample everyone in those clusters

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

Sampling distribution of the sample mean

Sampling with replacement: When we pick a random sample, and then "put that sample back" into the population before choosing another sample

Sampling distribution of the sample mean (SDSM): The probability distribution of all possible sample means for a certain sample size n

The Central Limit Theorem (CLT): Even if a population distribution is non-normal, as long as we use a large enough sample ($n \geq 30$), then we can make inferences using our sample statistics, because of the fact that the SDSM will be a normal distribution.

Mean, variance, and standard deviation of the SDSM:

Mean: $\mu_{\bar{x}} = \mu$

Variance: $\sigma_{\bar{x}}^2 = \dfrac{\sigma^2}{n} \approx \dfrac{s^2}{n}$

Standard deviation (standard error): $SE = \sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} \approx \dfrac{s}{\sqrt{n}}$

Finite population correction factor (FPC): If we're sampling without replacement or sampling from more than 5% of a finite population, we should multiply standard deviation by $\sqrt{(N-n)/(N-1)}$, and multiply variance by $(N-n)/(N-1)$.

Conditions for inference with the SDSM

Conditions for inference: In order to use the sampling distribution of the sample mean, we need to meet the random, normal, and independent conditions. We meet the normal condition if the original population is normal, and/or when $n \geq 30$. We meet the independent condition if we sample with replacement, and/or if we keep our sample size to 10% or less of the total population.

Sampling distribution of the sample proportion

Population proportion: The number of subjects in our population that meet a certain condition

Sample proportion: The proportion of subjects in our sample that meet a certain condition

$$\hat{p} = \frac{x}{n}$$

Sampling distribution of the sample proportion (SDSP): The probability distribution of all possible sample proportions for a certain sample size n

Mean, variance, and standard deviation of the SDSP:

Mean: $\mu_{\hat{p}} = p$

Variance: $\sigma_{\hat{p}}^2 = \dfrac{p(1-p)}{n}$

Standard deviation (standard error): $SE = \sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$

Conditions for inference with the SDSP

Conditions for inference: In order to use the sampling distribution of the sample proportion, we need to meet the random, normal, and independent conditions. We meet the normal condition when $np \geq 5$ and $n(1-p) \geq 5$. We meet the independent condition if we sample with replacement, and/or if we keep our sample size to 10% or less of the total population.

The student's t-distribution

Standard normal distribution, z-distribution: The normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$

Student's t-distribution: Similar to the standard normal distribution in the sense that it's symmetrical, bell-shaped, and centered around the mean $\mu = 0$, but flatter and wider. The exact shape depends on the number of degrees of freedom, df $= n - 1$.
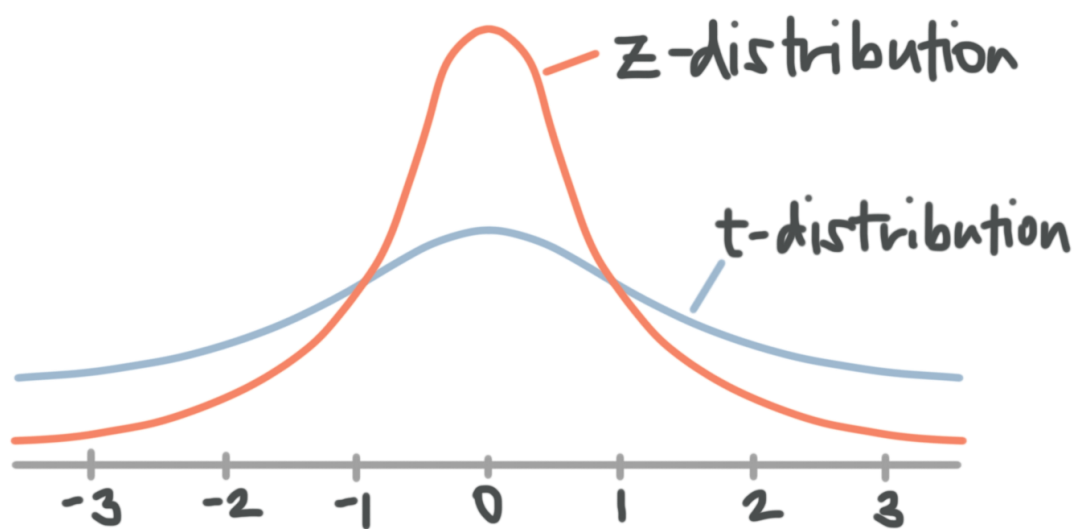
Confidence interval for the mean

Point estimate: An estimator for a singular value, like the sample mean for the population mean, or the sample standard deviation for the population standard deviation

Interval estimate: A range of values that estimate the interval in with some parameter may lie

Confidence level: The probability that an interval estimate will include the population parameter

Alpha value, α, level of significance, probability of making a Type I error:

$$\alpha = 1 - \text{confidence level}$$

Region of rejection: The region outside the confidence interval, in the tail(s) of the probability distribution. If the z-value we calculate falls in this region, we'll reject the null hypothesis.

Confidence interval:

When σ is known: $(a, b) = \bar{x} \pm z^* \dfrac{\sigma}{\sqrt{n}}$

With the FPC applied: $(a, b) = \bar{x} \pm z^* \dfrac{\sigma}{\sqrt{n}} \sqrt{\dfrac{N - n}{N - 1}}$

When σ is unknown and/or the sample size is small: $(a, b) = \bar{x} \pm t^* \dfrac{s}{\sqrt{n}}$

Required sample size for a fixed margin of error:

$$n = \left(\frac{z^*\sigma}{ME} \right)^2$$

Confidence interval for the proportion

Confidence interval for the proportion:

$$(a, b) = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Hypothesis testing

Inferential statistics and hypotheses

Hypothesis testing process:

1. State the null and alternative hypotheses.

2. Determine the level of significance.

3. Calculate the test statistic.

4. Find critical value(s) and determine the regions of acceptance and rejection.

5. State the conclusion.

Inferential statistics: Using information we have about the sample to make inferences about the population

Hypothesis: A statement of expectation about a population parameter that we develop for the purpose of testing it

Alternative hypothesis: The abnormality we're looking for in the data; it's the significance we're hoping to find

Null hypothesis: The opposite claim of the alternative hypothesis

Significance level and type I and II errors

Type I error: When we mistakenly reject a null hypothesis that's actually true. The probability of making a Type I error is given by α, which is the same as the level of significance for the hypothesis test.

Type II error: When we mistakenly accept a null hypothesis that's actually false. The probability of making a Type II error is given by β.

Power: The probability that we'll reject the null hypothesis when it's false (which is the correct thing to do). This is what we want, so we want our test to have a high power.

Test statistics for one- and two-tailed tests

Two-tailed test, two-sided test, non-directional test: The alternative hypothesis states that one value is unequal to another, while the null hypothesis states that one value is equal to the other

One-tailed test, one-sided test, directional test: Either an upper-tailed test or lower-tailed test

Upper-tailed test, right-tailed test: The alternative hypothesis states that one value is greater than another, while the null hypothesis states that one value is less than or equal to the other

Lower-tailed test, left-tailed test: The alternative hypothesis states that one value is less than another, while the null hypothesis states that one value is greater than or equal to the other

Test statistics:

$$\text{test statistic} = \frac{\text{observed} - \text{expected}}{\text{standard deviation}}$$

When σ is known: $z = \dfrac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} = \dfrac{\bar{x} - \mu_0}{\dfrac{\sigma}{\sqrt{n}}}$

When σ is unknown and/or we have small samples: $t = \dfrac{\bar{x} - \mu_0}{s_{\bar{x}}} = \dfrac{\bar{x} - \mu_0}{\dfrac{s}{\sqrt{n}}}$

For the proportion: $z = \dfrac{\hat{p} - p_0}{\sigma_{\hat{p}}} = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}}$

The p-value and rejecting the null

p-value, observed level of significance: The smallest level of significance at which we can reject the null hypothesis, assuming the null hypothesis is true, or the total area of the region of rejection

If $p \leq \alpha$, reject the null hypothesis

If $p > \alpha$, do not reject the null hypothesis

p-value approach when σ is known:

Lower-tailed test	Reject H_0 when $p \leq \alpha$
Upper-tailed test	Reject H_0 when $p \leq \alpha$
Two-tailed test	Reject H_0 when $p \leq \alpha$

p-value approach when σ is unknown and/or sample size is small:

Lower-tailed test	Reject H_0 when $p \leq \alpha$
Upper-tailed test	Reject H_0 when $p \leq \alpha$
Two-tailed test	Reject H_0 when $p \leq \alpha$

Critical value approach:

Lower-tailed test	Reject H_0 when $z \leq -z_\alpha$
Upper-tailed test	Reject H_0 when $z \geq z_\alpha$
Two-tailed test	Reject H_0 when $z \leq -z_{\frac{\alpha}{2}}$ or $z \geq z_{\frac{\alpha}{2}}$

Significance, statistical significance: The probability of obtaining the result by chance

Confidence interval for the difference of means

Sampling distribution of the difference of means (SDDM): The probability distribution of every possible difference of means for certain sample sizes n_1 and n_2 from populations 1 and 2, respectively

Mean and standard deviation of the SDDM:

Mean: $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2}$

Standard deviation (standard error): $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$

Confidence interval around the difference of means:

With known σ_1 and σ_2: $(a, b) = (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$

With unknown σ_1 and σ_2 and/or small s_1 and s_2:

Unequal population variances: $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$

with df $= \dfrac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{1}{n_1 - 1}\left(\dfrac{s_1^2}{n_1}\right)^2 + \dfrac{1}{n_2 - 1}\left(\dfrac{s_2^2}{n_2}\right)^2}$ (round down)

Equal population variances: $(a, b) = (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \times s_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}$

with df $= n_1 + n_2 - 2$

$$\text{with pooled standard deviation } s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Hypothesis testing for the difference of means

Test statistic with large samples and unequal population variances:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Test statistic with large samples and equal population variances:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Test statistic with small samples and unequal population variances:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \text{ with df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

Test statistic with small samples and equal population variances:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \text{ with df} = n_1 + n_2 - 2$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Matched-pair hypothesis testing

Independent samples: Samples for which the observations from one sample are not related to the observations from the other sample

Dependent samples: Samples for which the observations from one sample are related to an the observations from the other sample

Matched-pair test: A hypothesis test with dependent samples

Confidence interval for a matched-pair test:

With σ_d known: $(a, b) = \bar{d} \pm z_{\alpha/2}\dfrac{\sigma_d}{\sqrt{n}}$

With σ_d unknown and/or $n < 30$: $(a, b) = \bar{d} \pm t_{\alpha/2}\dfrac{s_d}{\sqrt{n}}$ with df $= n - 1$

Confidence interval for the difference of proportions

Confidence interval around the difference of proportions:

$$(a, b) = (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

Negative values in the confidence interval: When the confidence interval contains 0, it means there's likely no difference in proportions. On the other hand, when the confidence interval doesn't contain 0, it means there's likely a difference in proportions.

Hypothesis testing for the difference of proportions

Test statistic for the difference of proportions:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\text{with } \hat{p} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

Regression

Scatterplots and regression

Scatterplot, scattergraph, scatter diagram: A plot of the data points in a set that compares two variables about the data at the same time
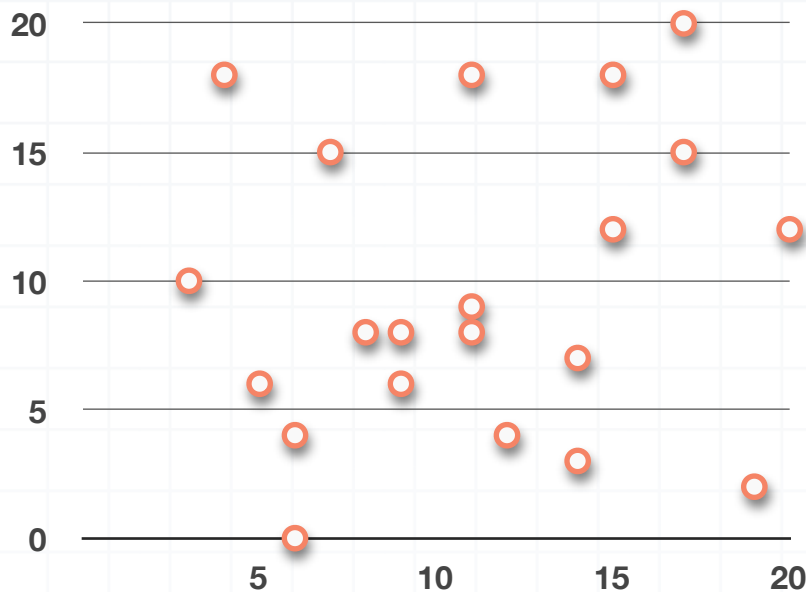
Approximating curve: The curve that approximates the shape of the points in the scatterplot

Curve fitting: The process of finding the equation of the approximating curve

Regression line, best-fit line, line of best fit, least-squares line: The line that best shows the trend in the data given in a scatterplot

$$\hat{y} = a + bx$$

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$a = \frac{\sum y - b \sum x}{n}$$

Positive linear relationship: When the regression line has a positive slope

Negative linear relationship: When the regression line has a negative slope

Strong linear relationship: When the data is tightly clustered around the regression line

Weak linear relationship: When the data is loosely clustered around the regression line

Regression: The process of estimating the value of the dependent variable from a given value of the independent variable

Correlation coefficient and the residual

Correlation coefficient: Tells us the strength of the relationship between two variables. The correlation coefficient will have a value on the interval $[-1,1]$; a value close to $r = -1$ indicates a strong negative relationship, a value close to $r = 1$ indicates a strong positive relationship, and a value close to $r = 0$ indicates no linear relationship.

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Residual: The residual e is the actual and predicted (based on the regression line) values for the same data point

Coefficient of determination and RMSE

Coefficient of determination, r^2: Measures the percentage of error we eliminated by using least-squares regression instead of just $\bar{y}$

Root-mean-square error (RMSE), root-mean-square deviation (RMSD): The standard deviation of the residuals

Chi-square tests

Types of χ^2-tests: A χ^2-test for homogeneity, a χ^2-test for association/independence, and a χ^2 goodness-of-fit test

Conditions for inference for a χ^2-test: Random, large counts, independent, categorical

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

t	Upper-tail probability p									
df	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.765	0.987	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									

x^2	Upper-tail probability p											
df	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.32	1.64	2.07	2.71	3.81	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.52	22.11
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42
11	13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.72	26.76	28.73	31.26	33.14
12	14.85	15.81	16.99	18.55	21.03	23.24	24.05	26.22	28.30	30.32	32.91	34.82
13	15.98	16.98	18.20	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53	36.48
14	17.12	18.15	19.41	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12	38.11
15	18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70	39.72
16	19.37	20.47	21.79	23.54	26.30	28.85	29.63	32.00	34.27	36.46	39.25	41.31
17	20.49	21.61	22.98	24.77	27.59	30.19	31.00	33.41	35.72	37.95	40.79	42.88
18	21.60	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31	44.43
19	22.72	23.90	25.33	27.20	30.14	32.85	33.69	36.19	38.58	40.88	43.82	45.97
20	23.83	25.04	26.50	28.41	31.41	34.17	35.02	37.57	40.00	42.34	45.31	47.50
21	24.93	26.17	27.66	29.62	32.67	35.48	36.34	38.93	41.40	43.78	46.80	49.01
22	26.04	27.30	28.82	30.81	33.92	36.78	37.66	40.29	42.80	45.20	48.27	50.51
23	27.14	28.43	29.98	32.01	35.17	38.08	38.97	41.64	44.18	46.62	49.73	52.00
24	28.24	29.55	31.13	33.20	36.42	39.36	40.27	42.98	45.56	48.03	51.18	53.48
25	29.34	30.68	32.28	34.38	37.65	40.65	41.57	44.31	46.93	49.44	52.62	54.95
26	30.43	31.79	33.43	35.56	38.89	41.92	42.86	45.64	48.29	50.83	54.05	56.41
27	31.53	32.91	34.57	36.74	40.11	43.19	44.14	46.96	49.64	52.22	55.48	57.86
28	32.62	34.03	35.71	37.92	41.34	44.46	45.42	48.28	50.99	53.59	56.89	59.30
29	33.71	35.14	36.85	39.09	42.56	45.72	46.69	49.59	52.34	54.97	58.30	60.73
30	34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70	62.16