Data Types

Nominal Data: Unordered categories like gender, blood type.
Ordinal Data: Ordered categories like letter grades (A, B, C, D, F), preference ranking (love, like, indifferent, dislike).
Ranked Data: Quantitive observations arranged and ranked from highest to lowest.
Interval Data: No true, biologically meaningful zero: temperature measured on the Celsius scale, time of day.
Ratio Data: There is a true, biologically meaningful zero: height, eggs per nest.
Discrete Data: Variable can only take on integer values: eggs per nest.
Continuous Data: Values are approximated and depend on the precision of the instrument. height, age.

Graphs

Bar Chart: used to display a frequency distribution for nominal or ordinal data. The bars should be of equal width and separated from one another so as not to imply continuity.

Histogram: used to display a frequency distribution for discrete or continuous data. The frequency associated with each interval is represented not by the height of the bar, but by the bar's area.

Box Plot: a central box from the 25th to the 75th percentile with a midline at the 50th percentile. The lines projecting out from the box on either side extend to the most extreme observations that are no more than 1.5 times the height of the box beyond either quartile. In fairly symmetric data sets, the adjacent values should contain approximately 99% of the measurements. All points outside this range are outliers, represented by circles.

Numerical Summary Measures

The Empirical Rule When the data are symmetric and unimodal, 68% will fall within the first standard deviation, 95% within the first two standard deviations, and 99.7% will fall within the first three standard deviations of the mean.

Chebyshev's Inequality For any shape of data distribution, at least $1 - (1/k)^2$ of the measurements lie within k standard deviations of the mean.

Sampling

The *target population* is the ideal population we would like to describe. The *study population* is the group from which we can actually sample. The *sampling frame* is the list of items or people in the study population. *Selection bias* is a systematic tendency to exclude certain members of the target population.

Confounding Variable is an extraneous variable in a statistical model that correlates with both the dependent variable and the independent variable.

Simple Random Sample a subset of individuals (a sample) chosen from a larger set (a population). Each individual is chosen randomly and entirely by chance, such that each individual has the same probability of being chosen at any stage during the sampling process, and each subset of k individuals has the same probability of being chosen for the sample as any other subset of k individuals.
The sampling fraction of the population is n/N, where n is the size of the sample and N is the size of the underlying population.
When N has mean μ and standard deviation σ, a finite version of the central limit theorem states that the distribution of the sample mean $\overline{x}$ has mean μ and standard deviation $\sqrt{1-(n/N)}(\sigma/\sqrt{n})$. 1-(n/N) is the *finite population correction factor*.

Probability Sampling Not everybody has the same probability of selection. each unit's probability is known and weighted into the final statistic. In contrast, in *nonprobability sampling*, like convenience samples and samples made up of volunteers, the probability that an individual subject is included is not known. These types of samples are prone to bias and cannot be assumed to be representative of any target population.

Systematic Sampling aka *interval sampling* involves imposing a gap, or interval, between each selected unit in the sample. A sampling interval K is selected by dividing the number of units in the population by the desired sample size. A unit is selected every kth units, where k is a random number between one and K.

Stratified Random Sample The population is divided into homogeneous subgroups. Then simple random sampling or systematic sampling is applied within each stratum. This can produce a weighted mean that has less variability than the arithmetic mean of a simple random sample of the population.

Cluster Sampling A number of clusters are selected randomly to represent the total population, and then all units within selected clusters are included in the sample.

Multi-stage sampling The population is divided to clusters. Random units are selected from each cluster.

Multi-phase sampling Basic information is collected from a large sample of units and then, for a subsample of these units, more detailed information is collected.

Probability

For two disjoint events, P(A or B) = P(A union B) = P(A $\cup$ B) = P(A) + P(B)
For any two events, P(A or B) = P(A) + P(B) - P(A and B)
For two independent events, P(A and B) = P(A intersect B) = P(A $\cap$ B) = P(A)*P(B)
Conditional Probability: P(B|A) = P(A and B) / P(A)
P(A and B) = P(A)*P(B|A)
Two events are independent if P(B|A) = P(B)
The complement of event A is the event "not A", denoted $\overline{A}$, or A^c

Binomial Distribution

A random variable that can assume the values "failure" or "success" is called a *Bernoulli random variable*. The binomial distribution with parameters n and p is the discrete probability distribution of the number of

successes in a sequence of n independent yes/no experiments, each of which yields success with probability p.
$X \approx B\left(np, \sqrt{np(1-p)}\right)$ The variance is largest when p = 0.5

The number of *permutations* is the number of ways in which n objects can be ordered: n! = n(n-1)(n-2) ... (3)(2)(1)
The number of permutations of k out of n objects: $\frac{n!}{(n-k)!}$
The number of *combinations* is the number of ways in which k out of n objects can be selected: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Normal approximation for binomial distribution As n gets larger, the binomial distribution approaches a normal distribution. The normal approximation can be used when $np \geq 10$ and $n(1-p) \geq 10$. Better results can be obtained by adding 0.5 to k if we're interested in the probability that X is less than k, and substracting 0.5 if we are calculating the probability that X is greater than k. This *continuity correction* compensates for the fact that the discrete binomial distribution is being approximated by a continuous normal distribution.

The Poisson Distribution

The number of combinations in a binominal distribution is hard to evaluate for large n's. When n is very large and p is very small, the binomial distribution is well approximated by another theoretical probability distribution, called the Poisson distribution. The Poisson distribution is used to model discrete events that occur infrequently in time or space. The probability that a random variable X, that represents the number of occurences of some event over a given interval, assumes the value x is

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

λ is the average number of events in an interval.
$e \approx 2.71828$ is the base of the natural logarithms.
The underlying assumptions are, that the probability that a single event occurs within an interval is proportional to the length of the interval, and that the events occur independently.
$X \approx P(np, \sqrt{np})$ when p is very small, the variance $np(1-p) \approx np$.

The Normal Distribution

If the number of possible values for X approaches infinity, while the width of the intervals approach zero, the graph will increasingly resemble a smooth curve. A smooth curve is used to represent the probability distribution of a continuous random variable; the curve is called *a probability density*. The most common continuous distribution is the *normal distribution*, also known as the *Gaussian distribution* or the *bell-shaped curve*. Its probability density is given by the equation

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The z statistic

Confidence Interval for the mean of a normally distributed population:

$$\overline{x} \pm z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$$

To test a hypothesis H_0: $\mu = \mu_0$ we calculate the z statistic:

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

For example, the value for $z = 1.96$ is $P(z < 1.96) = 0.975$.
The z statistic has the standard normal distribution N(0,1).

The p-value does not measure the probability that the null hypothesis is true. It measures the probability of observing the sample data assuming the null hypothesis were true. The hypothesis is either true or false - the randomness is attached to the data, not to the hypothesis.

Sample size for a desired margin of error: $n = \left(\frac{z * \sigma}{\overline{x} - \mu_0}\right)^2$.
Choose a margin which is ecologically and economically important.

Sample Size Determination for Testing the Mean

$$n = \left[\frac{(z_\alpha + z_\beta)\sigma}{\mu_1 - \mu_0}\right]^2 \qquad \Leftarrow \qquad \begin{cases} \overline{x} = \mu_0 + z_\alpha \left(\frac{\sigma}{\sqrt{n}}\right) \\ \overline{x} = \mu_1 - z_\beta \left(\frac{\sigma}{\sqrt{n}}\right) \end{cases}$$

α - desired test significance level
$(1 - \beta)$ - desired test power

The t statistic

When σ is not known,

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

with n-1 degrees of freedom.
There is a different t distribution for each sample size. A particular t distribution is specified by its degrees of freedom. As the degrees of freedom increase, the t density curve approaches the N(0,1) distibution.
The t procedure is robust - it can be used even when the assumption of normality is violated, when the sample size $n \geq 40$.

Confidence Interval for the Mean

$$\overline{x} \pm t_{\alpha/2} * \frac{s}{\sqrt{n}}$$

Matched pairs t-test To compare two sample means in a matched pair design, apply t-test to the observed differences.

Two independent samples t-test

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \qquad t = \frac{\overline{x}_1 - \overline{x}_2}{SE} \qquad \text{with } n_1 + n_2 - 2 \text{ degrees of freedom.}$$

Confidence interval for the difference between two sample means:

$$\overline{x}_1 - \overline{x}_2 \pm t * \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Comparing two Standard Deviations When s_1^2 and s_2^2 are sample variances from independent standard random samples drawn from a normal population, the *F statistic* $F = \frac{s_1^2}{s_2^2}$ has the F distribution with $n_1 - 1$ *and* $n_2 - 1$ degrees of freedom when $H_0 : \sigma_1 = \sigma_2$ is true.

Population Proportion

$$\hat{p} = \frac{count\ of\ successes}{n} \qquad \hat{p} \approx N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

The population has to be at least 10 times as large as the sample.
z statistic for significance tests of a proportion can be used when $np \geq 10$ and $n(1-p) \geq 10$.
To improve the accuracy of the confidence intervals for a proportion, add four imaginary observations, two successes and two failures, to p and n.
Sample size for a desired margin of error: $n = \left(\frac{z}{m}\right)^2 p(1-p)$

Comparing two proportions When the samples are large,

$$\hat{p_1} - \hat{p_2} \approx N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

The formula for the SE can be used when $np \geq 10$ and $n(1-p) \geq 10$. Confidence intervals can be calculated only when the counts of successes and failures in both samples are 10 or greater.
To improve the accuracy of the confidence intervals for a proportion, add four imaginary observations, two successes and two failures, in each of the two samples.

The sample proportions can be pooled

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

Alternatively, we can use the Chi-Square test.

Contingency Tables

The Chi-Square statistic is a measure of how far the observed counts are from the expected counts in a two-way table.

	Liberal	Conservative	row total
Female	a	b	a + b
Male	c	d	c + d
column total	a + c	b + d	n

$$expected\ count = \frac{row\ total\ *\ column\ total}{table\ total} = row\ total * \frac{column\ total}{table\ total}$$

$$\chi^2 = \sum \frac{(observed\ count - expected\ count)^2}{expected\ count}$$

... with (r-1)(c-1) degrees of freedom.

Chi-square test can be used when no more than 20% of the expected counts are less than 5 and all individual expected counts ≥ 1. When sample sizes are small, *Fisher's exact test* can be used.

When df = 1 it is recommended to use the *Yates correction for continuity*, where the absolute value of each deviation of f_i (the observed frequency, or number of counts) from $\hat{f}_i$ is reduced by 0.5.

$$\chi_c^2 = \sum_{i=1}^{2} \frac{(|f_i - \hat{f}_i| - 0.5)^2}{\hat{f}_i}$$

The Log-Likelihood Ratio $G = 2 * \sum f_i ln(f_i/\hat{f}_i)$ may be used instead of chi-square.

McNemar's Test is a two-sample test for binomial proportions for matched-pair data. It is applied to 2 x 2 contingency tables with a dichotomous trait, to determine whether the row and column marginal frequencies are equal.
For example, suppose you have twins randomized to two treatment groups (Treatment and Control) then tested on a binary outcome (pass or fail).

	Treatment	
Control	Fail	Pass
Fail	a	b
Pass	c	d

In order to test if the treatment is helpful, we use only the numbers of discordant pairs of twins, b and c, since the other pairs of twins tell us nothing about whether the treatment is helpful or not. McNemar's test is $\chi^2 = \frac{(b-c-1)^2}{b+c}$. -1 is a continuity correction, since we're using discrete counts to estimate the continuous χ^2 distribution.

Odds If an event occurs with probability p, the *odds* in favor of the event are $p/(1-p)$ to 1. If an event occurs with probability 1/2, for instance, the odds in favor of the event are 1 to 1. Conversely, if the odds in favor of an event are a to b, the probability that the event occurs is a/(a + b).

Relative Risk or *risk ratio* (RR) is the ratio of the probability of an event occurring (for example, developing a disease) in an exposed group to the probability of the event occurring in a comparison, non-exposed group.

$$RR = \frac{P(disease|exposed)}{P(disease|unexposed)}$$

Odds Ratio or *relative odds* is the odds of disease among exposed individuals divided by the odds of disease among the unexposed.

$$OR = \frac{P(disease|exposed)/[1 - P(disease|exposed)]}{P(disease|unexposed)/[1 - P(disease|unexposed)]}$$

It can also be defined as the odds of exposure among diseased individuals divided by the odds of exposure among those who are not diseased:

$$OR = \frac{P(exposure|diseased)/[1 - P(exposure|diseased)]}{P(exposure|nondiseased)/[1 - P(exposure|nondiseased)]}$$

For rare diseases, the odds ratio is a close approximation of the relative risk. The sampling distribution of OR has better statistical properties than that of RR, hence it is generally preferable to work with.

The chi-square test allows us to determine whether an association exists between two independent nominal random variables, and McNemar's test does the same for paired dichotomous variables, but neither test provides a measure of the strength of the association. For a 2 x 2 table of two independent dichotomous variables, the *odds ratio* is one such measure.

Heterogeniety Chi-Square Analysis

Multiple 2 x 2 Contingency Tables When a relationship between a pair of dichotomous random variables is being investigated, it is sometimes examined in two or more populations. In some cases, the data originate from different studies; in other cases, the data are subclassified, or statified, by some factor that is believed to influence the outcome. For example, if the dichotomous variables are "disease" and "exposed", we may have two separate 2 x 2 contingency tables for males and females.
The **Mantel-Haenszel Method** include a test of homogeneity of the different strata. If the odds ratios for all strata tables can be shown to belong to the same population, the method provides a means of calculating both a point estimate and a confidence interval for the overall population relative odds. In addition, it allows us to test the null hypothesis of no association between exposure and disease.

For the more general case, individual chi-square test are performed and the chi-square values are summed. If the samples are homogeneous, then the total of the individual chi-square values should be close to the chi-square for the pooled frequencies. These two chi-square values can be compared using another chi-square test.

Kolmogorov-Smirnov Goodness of Fit for Discrete Data

For data on an ordinal scale, The cumulative observed frequencies (F_i) and cumulative expected frequencies ($\hat{F}_i$) are calculated. The test statistic is $d_{max} = max(|F_i - \hat{F}_i|)$.
This test is more powerful than the chi-square test when n is small or when $\hat{f}_i$ values are small, and often in other cases.

Kolmogorov-Smirnov Goodness of Fit for Continuous Data

From the cumulative observed frequencies we calculate the cumulative relative frequencies (rel $F_i = F_i/n$) and the cumulative relative expected frequencies (rel $\hat{F}_i$).
Next, we calculate $D_i = |rel\ F_i - rel\ \hat{F}_i|$ and $D'_i = |rel\ F_{i-1} - rel\ \hat{F}_i|$.
The test statistic is $D = max(maxD_i, maxD'_i)$.

Analysis of Variance

Analysis of variance (ANOVA) provides a statistical test of whether or not the means of several groups are equal, and therefore generalizes the t-test to more than two groups. A successful grouping will split the observations such that (a) each group has a low variance (meaning the group is relatively homogeneous) (b) and the mean of each group is distinct.

Total Sum of Squares (SS_{Total}) = Within Sum of Squares (SS_{Error}) + Between Sum of Squares ($SS_{Treatment}$):

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{\overline{y}})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{y}_i - \overline{\overline{y}})^2$$

ANOVA F statistic is the signal-to-noise ratio used to test the equality of several means:

$$F = \frac{MS_{Treatment}}{MS_{Error}} = \frac{SS_{Treatment}/(k-1)}{SS_{Error}/(N-k)}$$

The F distribution has k - 1 degress of freedom in the numerator and N - k degrees of freedom in the denominator, where k is the number of populations and N is the total number of observations combined. When F = 0, all treatments have equal effects. When F = 1, the between-groups variance equals the within-group variance, i.e., the different treatments have no effect on the variance in the experiment.

The ANOVA Assumptions

- Independence of observations.
  Independent errors. Achieved by random assignment of subjects to groups.

- Normality.
  The distributions of the residuals (and of data within populations) are normal.
  ANOVA is robust with respect to this assumption too, if the data don't deviate severly from it.

- Homoscedasticity, or Homogeneity of variance.
  Equal variance among populations.
  The ANOVA F test is robust; the test results are correct when the largest sample SD is no more than twice as large as the smallest sample SD. Bartlett's test for homogeneity might be used to determine whether this assumption is met. However, because Bartlett's test is not very efficient and is badly affected by non-normality, it is generally not worthwhile to use in conjunction with analysis of variance.

- Additivity of the effects of the factor levels.

Data Transforamtion

Logarithmic Transformation $X' = logX$, or preferrably, $X' = log(X+1)$ will correct data that are multiplicative rather than additive, or when there is heteroscedasticity and the standard deviations are proportional to the means. It may also convert a positively skewed distribution into a symmetrical one. If the distribution of X' is normal, the distribution of X is said to be *lognormal*.

Square Root Transformation $X' = \sqrt{X}$, or preferrably, $X' = \sqrt{X+0.5}$. This transformation is applicable when the group variances are proportional to the means. This often occurs when samples are taken from a Poisson distribution (i.e. when the data consist of counts of randomly occurring objects or events).

Arcsine Transformation $p' = arcsin\sqrt{p}$. Percentages from 0 to 100% or proportions from 0 to 1 form a binomial, rather than a normal, distribution. This transformation is applicable to proportion or percentage data which are derived from counts. When nearly all values in the data lie between 0.3 and 0.7, there is no need for such transformation.

Multiple comparisons Procedure

If $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ is rejected, we can compare any two individual treatments using $t_{12} = \dfrac{\overline{y_1} - \overline{y_2}}{SE}$ where $SE = \sqrt{\dfrac{s^2}{n}}$ when group size are equal, or $SE = \sqrt{\dfrac{s^2}{2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}$ otherwise.

The Bonferroni Correction To set the overall probability of making a type I error at 0.05, we use $\alpha^* = *0.05/\binom{k}{2}$ as the significance level for an individual comparison.

The confidence interval for any treatment group is $\hat{\mu}_i \pm t * SE$.

Tukey test uses the studentized range distribution. $q = \dfrac{\overline{y_1} - \overline{y_2}}{SE}$
The critical value of q, based on three factors: α (the Type I error rate), k (the number of populations), and df (the number of degrees of freedom). The proper procedure is to compare first the largest mean against the smallest, then against the next smallest, until no differnce is found.

Newman Keuls Test is like the Tukey's test, except that the critical value p is the number of means in the range of means being tested.

Fix vs. Random Effects

When the levels of a factor are specifically chosen, the design is called a fixed-effects model, or a Model I, ANOVA. When the levels of a factor are chosen at random, the design is a random-effects model, or Model II, ANOVA. If a factorial design has both a fixed effect and a random effect factor, it is said to be a mixed-model, or Model III ANOVA.

ANOVA Designs

Completely Randomized Design (CRD) Each condition (specific combination of factor levels) is randomly assigned to an experimental unit. This randomization ensures that the effects of all possible other variables that might affect the response are, on average, equal in all treatment groups. A simultaneous analysis of the effect of more than one factor is termed **factorial analysis of variance**.

Two-Factor Analysis of Variance with equal replications Taking factor A and factor B, let us have a represent the number of levels in factor A, b the number of levels in factor B, and n the number of replicates.

$$SS_{Total} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{l=1}^{n}(X_{ijl} - \overline{X})^2 \ with \ df = N - 1$$

The variability within cells is:

$$SS_{Error} = \sum_{i=1}^{a}\sum_{j=1}^{b}\left[\sum_{l=1}^{n}(X_{ijl} - \overline{X}_{ij})^2\right] \ with \ df = ab(n-1)$$

The variability among cells (each cell being a combination of a level of factor A and a level of factor B) is:

$$SS_{Cells} = \sum_{i=1}^{a}\sum_{j=1}^{b} n(\overline{X}_{ij} - \overline{X})^2 \ with \ df = ab - 1$$

By considering factor A to be the sole factor in a single-factor ANOVA we calculate:

$$SS_A = bn \sum_{i=1}^{a} (\overline{X}_i - \overline{X})^2 \ with \ df = a - 1$$

The variability among cells include an interaction effect between factors A and B:

$$SS_{A\times B} = SS_{Cells} - SS_A - SS_B \ with \ df = (a-1)(b-1)$$

Testing H_0: Factor A has no effect: $F = \dfrac{MS_A}{MS_{Error}}$
Testing H_0: The differences in the response among levels of one factor are the same at all levels of the second factor (there is no interaction effect): $F = \dfrac{MS_{A\times B}}{MS_{Error}}$
Testing H_0: Factor A has no effect when both factors are random (Model II ANOVA): $F = \dfrac{MS_A}{MS_{A\times B}}$
when factor A is fixed and factor B is random, testing factor A: $F = \dfrac{MS_A}{MS_{A\times B}}$ and factor B: $F = \dfrac{MS_B}{MS_{Error}}$
It is generally not meaningful to speak of a factor effect if there is a significant interaction effect.

Randomized Complete Block design (RCB) The experimental units are grouped into blocks according to known or suspected variation which is isolated by the blocks (e.g., age, litter of animals, field fertility, test site). Therefore, within each block, the conditions are as homogeneous as possible, but between blocks, large differences may exist.
Within each block, experimental units are completely randomized (independently within each block) to treatments such that every treatment occurs once and only once in each block and all treatments occur in each block.
The analysis of randomized block data is performed as a mixed model two-way ANOVA without replications. In this case the $SS_{Cells} = SS_{Total}$. Consequently, $SS_{Error} = 0$. The part of the total variability not accounted for by the effect of the two factors is $SS_{Remainder} = SS_{Total} - SS_A - SS_B$.
Testing H_0: Factor A has no effect: $F = \dfrac{MS_A}{MS_{Remainder}}$
This design relies on the assumption of additivitly between treatment effects and block effects upon the response variable (no interaction between treatments and blocks).

Latin Square Design A 3×3 square is a three-factor ANOVA with one fixed and two random factors (blocks), with no replications. Higher dimension squares are also possible.

II	I	III
III	II	I
I	III	II

Repeated-Measures Experimental Design also called a within-subject design, in one in which multiple measurements on the same experimental subject comprise the replicate data. This design is disadvantageous if there are effects of the sequence in which the treatments are administered to the subjects. Another disadvantage arises if insufficient time is allowed between the administration of different treatments to avoid *carryover* effects of the previous treatment. Carryover effect may often be counteracted by *counterbalancing*, whereby, to the extent possible, each subject receives the treatments in a different suquence.

Split-plot design The split-plot design involves two experimental factors, A and B. Levels of A are randomly assigned to whole plots (main plots), and levels of B are randomly assigned to split plots (subplots) within each whole plot.

Nested (Hierarchical) ANOVA Nested designs are used when levels of one factor are not represented within all levels of another factor. Each group (factor A) is divided into subgroups (factor B). These subgroups may be chosen randomly from a larger set of possible subgroups. Alternatively, the random-effects nested factor is included in order to account for some within-group variability.

Nonparametric Methods

Nonparametric methods do not rely on assumptions on the underlying distribution of the data (e.g. normality).

The Sign Test is used to test the null hypothesis that the median M of a distribution is equal to some value. It can be used in place of a one-sample t-test, in place of a paired t-test, or with ordinal data.
Suppose that r^+ of the observations are greater than M and r^- are smaller than M (in the case where the sign test is being used in place of a paired t-test, we count how many observations in group A are larger than the matched group B observations). Values which are exactly equal to M are ignored.
Under the null hypothesis we would expect half the x's to be above the median and half below. Therefore, under the null hypothesis both r^+ and r^- follow a binomial distribution with $p = \frac{1}{2}$.
We can use the normal approximation to the binomial distribution, or, if n is small, the exact method, which entails computing the binomial probability.

The Wilcoxon Signed-Rank Test is a nonparametric alternative to the paired t-test which takes into account the magnitude of the differences as well as their sign.
The absolute value of the differences between observations are ranked from smallest to largest, with the smallest difference getting a rank of 1, then next larger difference getting a rank of 2, etc. Ties get average ranks. The ranks of the positive and negative differences are then summed separately. The smaller of these two sums is the test statistic, T.
Under the null hypothesis, the median of the underlying population of differences should be equal to 0, with random distribution of positive and negative ranks. We evaluate this hypothesis with the statistic

$$z_T = \frac{T - \mu_T}{\sigma_T} \quad \text{where} \quad \mu_T = \frac{n(n+1)}{4}\,, \qquad \sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

The Wilcoxon Rank Sum Test aka **the Mann-Whitney U test**, is a nonparametric alternative to the two-sample t-test. It evaluates the hypothesis that the medians of the two populations are identical.
All observations are ranked from smallest to largest. Ties get average ranks. The ranks are then summed for each group. The smaller of the two sums is denoted by W.

$$z_W = \frac{W - \mu_W}{\sigma_W} \quad \text{with} \quad \mu_W = \frac{n_S(n_S + n_L + 1)}{2}\,, \qquad \sigma_W = \sqrt{\frac{n_S n_L(n_S + s_L + 1)}{12}}$$

n_S is the number of observations in the sample with smaller sum of ranks
n_L is the number of observations in the sample with larger sum of ranks

The Kruskal-Wallis test often called an "Analysis of variance by rank", can be used when the k samples do not come from normal populations, or when the population variances are somewhat heterogeneous. If k = 2, then the Kruskal-Wallis test is identical to the Mann-Whitney test.

Multivariate Analysis of Variance (MANOVA)

There are experimental designs where more than one variable is measured on each experimental subject. We want to compare two or more variables' means among two or more groups. Doing so with multiple ANOVAs would result in an inflated Type I error. MANOVA also considers the correlation among multiple variables.
In MANOVA with two variables and k groups, the null hypothesis is $H_0 : \mu_{11} = \mu_{12} = ... = \mu_{1k}$ and $\mu_{21} = \mu_{22} = ... = \mu_{2k}$, where μ_{ij} denotes the population mean of variable i in group j.

There are several methods for comparing means to test MANOVA hypotheses. Computer programs may transform the resulting statistic into a value of F or chi-square, with associated probability.

Wilks' lambda Wilks' Λ, or Wilks' likelihood ratio, is the most commonly encountered MANOVA statistic, dating from the original formulation of the MANOVA procedure. Λ, ranging from 0 to 1, is a measure of the amount of variability among the data that is not explained by the effect of the levels of the factor. Thus, a measure of the proportion of the variability among the data that is explained by the experimental factor is $\eta^2 = 1 - \Lambda$.

Pillai's trace is a good choice when the variables are correlated.

Hotelling-Lawley Trace

Roy's Maximum Root is a better choice when the variables are not correlated.

Pearson's Correlation Coefficient

The strength of the linear relationship between two continuous variables X and Y can be quantified as

$$\rho = E\left(\frac{(X - \mu_x)}{\sigma_x}\frac{(Y - \mu_y)}{\sigma_y}\right)$$

The estimator of the population correlation is known as *Pearson's coefficient of correlation*, or simply the *correlation coefficient*.

$$r = \frac{1}{n-1}\frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_x s_y}$$

$-1 \leq r \leq 1$. If y tends to increase in magnitude as x increases, $r > 0$ and x and y are said to be positively correlated. if r = 1 or r = -1, there is a perfect correlation between x and y.

$$t_{n-2} = r\sqrt{\frac{n-2}{1-r^2}}$$

Spearman's Rank Correlation Coefficient is a nonparametric statistic, obtained by ranking the two sets of variables x and y separately, and caluculating the correlation coefficient of the ranked data. This method is not as sensitive to outliers as Pearson's correlation coefficient.

Regression

Like correlation analysis, simple linear regression is used to explore the nature of the relationship between two contiuous random variables. The primary difference between these two analytical methods is that regression enables us to investigate the change in one variable, called the *response*, which corresponds to a given change in the other, known as the *explanatory variable*. Correlation analysis makes no such distinction; the two variables involved are treated symmetrically. Rather than just quantifying the strength of the association, regression enables us to predict, or estimate, the value of the response that is associated with a fixed value of the explanatory variable.

$$\sigma_{y|x} = (1 - \rho^2)\sigma_y^2$$

Survival Analysis

In some studies, the response variable of interest in the amount of time from an initial observation until the occurrence of a subsequent event. Examples include the time from birth until death, the time from transplant surgery until the new organ fails, etc. This time interval between a starting point and a subsequent event, often called *failure*, is known as the *survival time*. The distributions of survival time measurements tend to be skewed to the right. We are usually interested in estimating the probability that an individual will survive for a given length of time. One common circumstance in working with survival data is that not all the individuals in a sample are observed until their respective times of failure. Incomplete observation of a time to failure is called *censoring*.
The survival function is represented by $S(t) = P(T > t)$. The graph of S(t) versus t is called a *survival curve*. In a life table, survival times are grouped within intervals of fixed length t to t+n. $S(t) = l_t/t_0$, where l_t is the number still alive at time t. q_t is the proportion of failure at interval t to t + n.

The product-limit method of estimating a survival function, also called the *Kaplan-Meier method*, is a nonparametric technique that uses the exact survival time for each individual in a sample instead of grouping the times into intervals.

Comparing the distribution of survival times for two different poplations If there are no censored observations in either group, the Wilcoxon rank sum test could be used to compare mediam survival times. Otherwise, the *Log-Rank Test* can be used: For each time t at which a death occurs, a 2x2 contingency table of group versus survival status is constructed. Next, the Mantel-Haenszel statistic compares the observed number of failures at each time to the expected number of failures given that the distributions of survival times for the two groups are identical. If the null hypothesis is true, the test statistic has an approximate chi-square disturbution with 1 degree of freedom.