Data Analysis I

T QI

Contents

1 Study Design

Human randomness or systematic choices are inherently biased. Computer generated randomness are much better.

Aim to design a study to have meaningful data, in order to

- Draw conclusions that are valid (unbiased)
- Provide estimates of responses with high precision

Biased is a lack of accuracy.

1.1 Study design

Terms for study design

- Study units, cases, subjects are the units where data are obtained, one observation per study unit
- Response variable, dependent variable, the variable of interest which we expect may depend on other variables
- Explanatory variables, independent variables, are variables that are used to explain or predict the response variable
- Other explanatory variables, dependent variables that are not the primary interest to the study. We aim to control these variables or measure them to eliminate them.

Principles of good study design include both validity of the study (biases) and also the precision of the data.

1.2 Validity and biases

Validity is about the biasness of the study. A bias study may measure/include implicit other predictor variables that we are not interested at. To minimize the potential of biases

- Comparison: comparing with a control group, placebo, current best, natural groups.
- Control: through restrictions (discrimination), protocols (being systematic in experiments), blinding (unaware groups for placebos), attempting to eliminate/hold constant the confounding variables
- Randomization, randomized subjects that are representative of the population, randomized groups of subjects that are similar with the baseline (basically reducing effects of confounding variables)

The terms:

- Control groups are experimental units which do not receive the treatment, but are similar to other experimental units
- Placebos are special types of control groups which are fake treatments. The placebo effect are effects not produced by placebo itself, and attributed through the patient's belief
- The current best is the group given the current best treatment
- Natural groups refers to groups that occur naturally
- Restrictions are limiting the range of samples/population, but it limits the explanatory power of the study/dataset
- Protocols refers to the calibration and consistency of the experiment
- Blinding is to attempt to blind the experiment units about the group they are in. A study is blind if either the experimenter or the subject is blind, it is double-blind if neither knows.

Confounding factors affect both the dependent and independent variables (relation with treatment group and causation with the outcome). For example, difference between doctors administering each treatment may affect the outcome, and we are actually measuring the doctor's effectiveness.

Confounding factors may create a correlation between the explanatory and response variable, but does not necessitates a causal relationship.

A lurking variable is an unobserved, unmeasured variable that affects the study.

To deal with confounding factors, we aim to make the subject groups similar in respect to the possible confounding variable

- Randomly assign subjects to groups, as to ensure a fair distribution of all characteristics within a group
- Randomizing the order of treatments when subjects receive both treatments
- Restriction of some form, say on the number of the experimenter
- Blocking, have each experimenter perform both treatments

For randomization of the treatment/subject groups, we can compute measures of potential confounding variables for each treatment group before deciding to intervene.

1.3 Precision

A study or projection is precise if we get close estimates.

To maximize precision,

- Blocking is the act to divide the study units into blocks, blocks which contains similar characteristics within (say same gender), but different between blocks. This reduces the variability of the result due to the confounding variable. We then randomize the treatment for study units within a block. In observational studies, this is called stratification.
- Replication, increasing the number of representative subjects sampled — higher sample size, or to replicate the measurements on each study group and use the average data. This is not repeatability/reliability, which is how repeatable the study is in other independent similar studies.
- Balance, it is better to have equal sized groups.

An example of blocking is the matched pair design, where a similar pair of twins creates a block. Moreover, a block can be a single subject if the subject undertakes both treatments (of course, with randomized ordering of treatments).

1.4 Study types

Observational studies have the data collected through observing purely the responses. In observational studies, the subject decides the group they are in (implicitly through their decision making). It may produce correlations in the relationship between the explanatory and response variable, but cannot accurately generate a causal link.

In designed experiments, the experimenter deliberately imposes some treatment to study units to observe the response. The experimenter assigns the study units to have different treatments — the experimenters decide the group the study unit belongs to. Designed experiments can prove causation.

Observational studies follow the evidence by observation, where data is collected from observations. Designed experiments follow the evidence by design, where data is collected from a well-thought-out, deliberately designed experiment.

Problems with observational studies

- Selection bias, the act that the samples in observational studies (say surveys) may be selected in a biased manner, or biased through self selection
- Reporting bias, that specific groups may be biased in responding/reporting
- Confounders are not accounted for, because we cannot randomly assignment treatments.

In designed studies, the experimenters have better control over the randomization process (in assigning treatments, and in blocking) to reduce biases and uncontrolled variables (or even nuisance variable that decreases precision).

A design that uses randomization without any manual matching is called a completely randomized design. Randomization can be accomplished using mechanical devices or computers.

In general for a causal relationship

$$\text{data} = \text{signal} + \text{noise}$$

We can remove the noise by including the element that causes the noise in our signal.

2 Exploratory Data Analysis

EDA is exploratory data analysis.

Data analysis is used to

- Discover important features of the data
- Improve understandings of the underlying population
- Transform data into information

and uses

- Appropriate statistical techniques
- Obtain results
- Analyze results

To explore data, we aim to: examine the variability within the data, model the data, and to detect the story the data can tell.

The steps of exploratory data analysis are

- Display the sample data (graphs)
- Summarize the distribution of the sample data (statisics)
- Describe the statistics, what is revealed by the pictures and summary
- Conjecture what is happening to the population

2.1 Variable types

Data are either numerical or categorical. The types of variable determine what questions we can ask and how we can display it.

Categorical has: nominal and ordered variables. Numerical has: discrete and continuous variables. Nominal variables are simply groups, while ordered variables can have the groups to be ordered in some meaningful way.

There exists a hierarchy of information on the types of variables.

- Categorical nominal variables have the least information, for they only contains categories
- Categorical Ordinal variables have more information, as they also contain the information of order
- Discrete variables have even more information, as they provide a scale component

- Continuous variables have the most information

The type of questions we can ask is related to the type of variable we have measured.

For categorical variables we can ask: the category the data is in, the mode of the dataset, the association between two columns [1].

For numeric variables, we can ask: the mean, variance, min and max, outliers.

The distribution of a random variable refers to the way that possible values from the population/sample are located/spread across the range of possible values.

Comments and features of distributions

- Shape: symmetrical/skewed, unimodal/multimodal
- Location: mean
- Spread: variance
- Unusual: outliers or groupings
- Association: correlations between pairs of variables

2.2 Categorical variables

Bar charts are used to show the distribution of a qualitative (categorical ordinal) variable, as counts or percentages.

Simpson's paradox is a phenomenon where the trend exhibited within each group of the dataset changes when grouping them together. This is caused by balancing issues within the group size, and often generates misleading conclusions for the most frequently sampled value get overproportionally weighted.

2.3 Numeric variables

The same comments and features of distribution on a graph: shape, center, spread, outliers.

Plots to use include: dot plot, histogram, box plot. The type of plot used depends on the size of the dataset and the intent to summarize, showcase individual values, or to compare groups.

In general, dotplots for individual data, boxplot to summarize, comparative boxplot for comparisons, and dotplots for small dataset with histograms for large datasets.

[1] Association is often used for categorical trends, correlation for numeric trends, and relationships for either

If the dataset is stretched to the right, it is positively skewed. If the dataset is stretched to the left, it is negatively skewed.

Histograms Histograms are bar charts for numeric values. It is a bar chart representing frequencies in a range of values. To make a histogram

1. Divide the range of the data by k, this is the width of each bin
2. Round each bin to a sensible unit
3. Choose the first bin/class to contain the smallest observation, and keep adding bins adjacent, making sure the other bins/intervals don't overlap

A rough rule to choose k, the number of classes/bins, is $k = \log_2(n)$, where n is the size of the data.

Box plots A box plot is a visual display of the five number summary: the minimum, lower quartile, median, upper quartile, and maximum. The steps to construct a box plot is

1. Draw the box between the LQ and UQ, with a line at the median
2. Whiskers are drawn to the largest and smallest elements that are not outliers, outliers are marked with circle outlines explicitly
3. The quantitative variable maybe on either the horizontal or vertical axis, however horizontal axis makes most box plots easier to read

2.4 Numerical measurements — Location

Numerical data contains some interesting summary measurements.

Mean Mean is the average value of the dataset. For observations x_i, its mean is

$$\bar{x} = \frac{1}{n} \sum_i x_i$$

Order statistic The order statistics for a set of observations refers to the sorted series of observations, where

$$x_1 \leq x_2 \leq \dots$$

we can then compute the minimum and maximum values by looking at the first and last value.

Median The median is the middle entry in the order statistics of a dataset. It is

$$m = x_{\frac{n+1}{2}}$$

If n is even, we take the arithmetic average between the middle values and define that as the median. The median is robust against outliers.

Quartiles The lower quartile is the 25th percentile, equal to $x_{0.25(n+1)}$. The upper quartile is the 75th percentile, equal to $x_{0.75(n+1)}$. Together with the median, they partition the dataset into roughly 4 groups of equal size.

We use linear interpolation to compute quartiles if the n is fractional (for example, $x_{2.25} = x_2 + 0.25(x_3 - x_2)$)

2.5 Numeric measurements — spread

Range The range is the difference between the largest value and the smallest value of the dataset. This is very sensitive to outliers. When used with the IQR, it can give a sense of spread in the dataset.

IQR The interquartile range is the range covered by the middle 50% of the data, it provides information on how close the data are to the median.

$$IQR = Q_3 - Q_1$$

The five number summary Q_i consists of the minimum Q_0, the lower quartile Q_1, the median Q_2, the upper quartile Q_3, and the maximum Q_4.

We should use the IQR when the dataset is skewed.

Standard Deviation The standard deviation measures the consistency/closeness that the observations are to its arithmetic mean. In other words, the expected deviation/difference of an observation with the sample mean.

For an unbiased estimator of the population standard deviation given the sample, we have

$$s = \sqrt{E[(X - \mu)^2]} = \sqrt{\frac{1}{n-1} \sum_i (x_i - \bar{x})^2}$$

where $\bar{x}$ is the sample mean. Note the division by $n - 1$, this is to account for the sampling and is an unbiased estimator for the population standard deviation (to get the variance/sd for the dataset, we still divide by n).

Standard deviations are nice for they have the same units as the data. Chebyshev's inequality states that for most distributions, no more than 25% of data points are outside 2 standard deviations from the mean (or that 75% of data points within 2sd)[2]. The 68-95-99.7 rule applies for the normal distribution. And that the square of the standard deviation is the variance.

Outliers To judge outliers, consider if: it is a legit data value, it is not an entry mistake, and the data belongs in the correct population group.

The quantitative rule is the IQR rule, is that if an observation is 1.5 IQR above or below Q_3 or Q_1, then it is a potential outlier.

In general, to describe a set of data points, consider its: shape, center, spread, and outliers.

2.6 Several numeric variables

To handle data pairs — measurements drawn from the same study units — we must use the appropriate graphics. We can either use scatterplots or dotplots on the variable differences.

Correlation Correlation is a measurement on the strength (degree of correlation) and direction (sign of correlation) of a linear relation between two random variables X and Y, defined as

$$r = \frac{1}{n-1} \sum_i \frac{x - \bar{x}}{s_x} \frac{y - \bar{y}}{s_y}$$

given the sample standard deviations s_x and s_y.

We can also normalize the samples (z distribution) using the sample standard deviations and use that.

The correlation is related to the covariance by

$$\mathrm{Cov}(X, Y) = \sigma_x \sigma_y r = \frac{1}{n} \sum_i (x - \bar{x})(y - \bar{y})$$

notice the use of the population standard deviation this time.

The correlation r has attributes:

- Range from -1 to 1

[2] In general, for most random variables X, Chebyshev's inequality states that the probability of having a value k standard deviations away from the mean is

$$Pr(|X - \mu| > k\sigma) \leq \frac{1}{k^2}$$

- Positive indicating positive association, negative implying negative association
- -1 and 1 indicating perfect correlation
- $r = 0$ implies no linear association between the variables (but not non-linear associations)
- Correlation is affected by outliers
- Correlation is unitless

Correlation must be used in conjunction with a scatterplot. Because pure numbers can lie about the shape of the plot.

In comparing two groups, we can compare their means (or medians), and the ratio between their variances (or IQRs).

2.7 One categorical, one numeric

Displays are: boxplots, histograms, and dotplots. The response variable is often numeric, with the categorical data the explanatory variable. These are called comparative plots (plots to compare between often the categorical explanatory variable).

For comparing between groups, we can also compare the difference between means, or ratios between standard deviations.

2.8 Summaries of graphical displays

Numeric Variables	Categorical Variables	Nice Displays
1	0	dotplot, histogram, boxplot
0	1	Table (with percentages) or barchart
2	0	Scatterplot, boxplot of differences
1	1	Comparative dotplot, boxplot
0	2	Continecy tables (with %) or comparative bar charts
3	0	Surface Plot
2	1	Grouped Scatterplot
1	2	Interaction plot
0	3	Cross tabulation or comparative bar chart

Note that interaction plots are discrete/categorical grouped scatterplots.

A graphical model for a set of data is used to create predictions. It represents an abstract formula (usually with a random error component) that quantitatively describes the relationship between the explanatory variable and response variable.

3 Modeling Randomness

Our goal is to use the sample data to infer the population distribution. While our sample data is random, we can aim to understand this random process by studying how it behaves.

3.1 Definitions

We look to model the hypothetical parent population to understand how sample statistics varies from sample to sample.

We use distributions to model two things

- The empirical distribution, to model the variable samples we take. This is the distribution we see
- The hypothetical distribution, which is a probability model, to model the generation (mimic) of the empirical distribution. This is the distribution we imagine

A random process/phenomena is an event where the outcome of a single trial is unpredictable. However, most random processes has the feature that, over many trials, the random process settles and becomes predictable — regression to the mean.

Probability is a way to quantitatively measure the chance of observing a particular outcome for a random process. Randomness in experimental data arises from

- The way we assign subjects to treatments
- The way we sample our subjects
- Measurement error

We define the probability using the frequentists' view, that the probability of an event is the relative frequency of its occurrence in an infinite sequence of trials. We denote the probability of the outcome x for a random process X as

$$Pr(X = x)$$

An event for a random process denotes one of the many outcomes the process could produce, with $Pr(\text{event})$ defined to be the probability of that event — so $X = x$ is the event where X measures to be x.

3.2 Random variables

A random variable is a numeric variable, with a value determined by the outcome of a random process.

The variable has an unknown value before the random process, and an observed value after the process. The former represents the population, while the latter is the sample. The observed numeric value is called a realization of the random variable.

The notation for random variables has the random variables be denoted as capital letters X, and observed/realized/sampled variables denoted as lowercase letters x.

A random variable is completely specified/defined by its probability distribution. A probability distribution summarized the probabilities associated with all the possible outcomes/observations of the random variable. It also qualitatively describes how the observed/realized values will be distributed among all the possible values.

The two types of random variables are: discrete and continuous random variables.

Discrete Random Variables Discrete random variables have countably many possible outcomes. Its value often represents counts. Continuous random variables can take any value in an interval. They are often measurements.

Discrete probability distributions are defined by: for a discrete random variable X, its probability distribution is characterized by the probability mass function pmf, $p_X(x)$, where

$$p_X(x) = Pr(X = x)$$

The pmf has some properties:

- $\forall x$, $p_X(x) \geq 0$, ie, probability cannot be negative
- $\sum_x p_X(x) = 1$, namely, the sum of probabilities for all the outcomes is 1.

Define the cumulative distribution function of any random variable X as

$$F(x) = Pr(X \leq x)$$

The probability for a discrete random variable to take a value between the range $[a, b]$ is

$$Pr(a \leq X \leq b) = \sum_{x=a}^{b} p_X(x) = F(b) = F(a-1)$$

Continuous Random Variables The probability that a continuous random variable to take a value between two values is the area under its PDF (probability density function) between the two values.

The area under the pdf is one and the pdf is always positive. The probability of getting a particular value is zero.

$$f_X(x) \geq 0$$
$$\int_{-\infty}^{\infty} f_X(x)\,dx = 1$$
$$Pr(a \leq X \leq b) = \int_{-a}^{b} f_X(x)\,dx$$

The cdf for a continuous random variable is the same as the discrete case, but with integrals.

3.3 Describing distributions

We want to think about the important features of the distribution, namely the location and the spread.

The expected value (mean) is defined as

$$E[X] = \sum_i x_i P(X = x_i) = \int_{-\infty}^{\infty} x f_X(x)\,dx$$

Properties of the mean

- Does not have to be an observable value
- It is the point of symmetry if the distribution is symmetrical
- $E[aX + b] = aE[X] + b$
- $E[X + Y] = E[X] + E[Y]$

The pth percentile of the distribution of X is the value x where p% of the population/probability falls below this value. This is denoted as the x where $Pr(x \leq X) = p$. The lower quartile is the 25th percentile, the upper quartile is the 75th percentile. The median is the 50th percentile. Percentiles provide information on both the location and spread.

Standard deviation is the measure of spread of a population (squared is the variance). It is the average distance of an observation from the mean (for a sample, the population standard deviation doesn't make too much sense because the sample is not the theoretical distribution). For variance

$$\sigma^2 = E[X^2] - E[X]^2 = \sum p(x)(x - E[X])^2 = \int_{-\infty}^{\infty} f(x)(x - E[X])^2\,dx$$

where $p(x)$ is the probability mass function.

- $\sigma^2 \geq 0$
- The standard deviation is in the same units as the values
- 95% of data are within 2 std
- $V[aX + b] = a^2 V[X]$
- If X and Y are independent variables
$$V[aX + bY] = a^2 V[X] + b^2 V[Y]$$
- Otherwise
$$\begin{aligned} V[X + Y] &= V[X] + V[Y] + 2Cov[X, Y] \\ &= V[X] + V[Y] + 2(E[XY] - E[X]E[Y]) \end{aligned}$$
with $Cov[X, Y] = 0$ when X and Y are independent.

3.4 Standardization

We can scale the random variable with mean μ and sd σ to have a standardized mean and sd by

$$Z = \frac{X - \mu}{\sigma}$$

It is also possible to undo the standardization by performing the above in reverse.

3.5 IIDRVS

Stand for independent, identically distributed random variables. Consider a series of these variables X_i,

$$S_n = \sum_i^n X_i$$

Then

$$\begin{aligned} E[S_n] &= nE[X] \\ V[S_n] &= nV[X] \end{aligned}$$

The central limit theorem states that the sum of a large number of similarly distributed random variables is approximately normally distributed. Namely for a large n,

$$S_n \approx N(n\mu, \sqrt{n}\sigma)$$

This is why normal distributions are everywhere. Any variable that is considered to be the sum of many small random variables will be normally distributed.

Sampling Mean Distribution The sampling distribution of the sample means is

$$\bar{X} = \frac{1}{n} S_n$$

with

$$E[\bar{X}] = E[X]$$
$$V[\bar{X}] = \frac{1}{n} V[X]$$

This distribution is also approximately normally distributed, but the variance/sd will be smaller than the populations'.

3.6 Independent Events

Independent events are unrelated events: the weight of two different people are independent if the two people are randomly sampled.

In general, event A and B are independent when

$$Pr(A \cap B) = Pr(A)Pr(B)$$

For the random variables X and Y are independent if pair of its events are independent, namely

$$Pr(a \leq X \leq b \cap c \leq Y \leq d) = Pr(a \leq X \leq b)Pr(c \leq Y \leq d)$$

Moreover, the events A and B are independent implies

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

For dependent events, if

$$Pr(A|B) > Pr(A)$$

then the two events are positively associated. If

$$Pr(A|B) < Pr(A)$$

the two events are negatively associated.

4 Probability models

Statistical modeling assumes a theoretical population distribution. We will model the distribution with errors and make some assumptions about the errors.

The uniform distribution has each outcome to be equality likely.

The normal distribution approximates most real life distributions. We say a random variable has a normal distribution by

$$X \sim N(\mu, \sigma)$$

Its pdf is

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

- Is bell shaped
- Extends indef in both directions
- Symmetrical around the mean, with the median equal to the mean and mode
- Has points of inflection at $\pm\sigma$
- 95% of points are between two standard deviations
- Completely characterized by its mean and standard deviation

Let the standardized normal distribution be

$$Z \sim N(0, 1)$$

it denotes the number of standard deviations a value is away from the mean. We have that

$$Pr(X \leq x) = Pr(Z \leq \frac{x-\mu}{\sigma})$$

We use the standardized distribution to compute the probabilities for a generalized variable.

To use the z-table, draw a picture.

The sum of independent normal distributions are also normal. The expected value and variance formulas still apply.

4.1 Normality Tests

To assess normality of a distribution, use the 68-95-99.7 rule. In that we should check for the percentages of observations are within 1, 2, and 3 sd. We could also qualitatively verify.

QQ Plot A normal scores plot is a quantitative way to check for normality

- Order the raw data from smallest to largest
- Find the cdf for each of the sample data (inclusive of the current data point)
- Use the inverse z table to find the z values of each sample data using its probabilities (as if the distribution is normal) — this is to find the approximated normal scores.
- Plot the points of (z, x) with the estimated z values and the ordered sample statistics.

If the data is normally distributed, the QQ plot should be a line, with the intercept (expected z score is 0) being the mean, and slope being the standard deviation. Recognize that the normal scores are the scores as if the population is normal, and the estimated cdfs are empirical cdfs.

Note that to compute the empirical cdf, we first find the ordered statistics' orders k. Then by assuming a uniform distribution of the values, find the cdf for the ith ordered statistics by

$$F(X \leq x_i) = \frac{k}{n+1}$$

for there are a total of $n + 1$ steps (for an extra step below and an extra above step).

Normal Probability Plot A normal probability plot, plots the empirical cdf of the data values and scales the y-axis. The scaling used is one where the normally distributed cdf of the same mean and median would be graphed as a straight line. If the distribution is normal, the scaled empirical cdf should be a straight line. We compare with the empirical cdf with an actual normal distribution cdf to confirm its normality.

- The cdf must be monotonically increasing
- If the underlying distribution is normal, its normal probability plot will be a straight line.
- It is easier to judge if points are close to a line than checking whether a histogram looks normal

4.2 Binomial Random Variables

A Bernoulli trial is a random process with the only outcomes being success or failure. Its parameter p is the probability of success.

A binomial distribution is the sum of n Bernoulli trials. It must be that

- A fixed number of trials
- The outcome of all trials are independent

- The probability of success do not change between trials

The probability mass function for a binomial variable is some formula we will never need to know. Its parameters are

$$X \sim Bi(n,p)$$

The mean and sd for the distribution is

$$\mu = np, \sigma = \sqrt{np(1-p)}$$

For probabilities, round to 4 decimal places, or one more decimal place than the data.

We can add independent binomial variables with the same chance of success p to gain a new binomial variable with the number of trials equal to the sum of each trials.

4.3 Approximation to the binomial distribution

The probabilities for binomial distributions are hard to calculate for large n trials. We can use a normal distribution with the same mean and standard deviation to model a sufficiently large binomial variable.

A sufficiently large binomial variable is often defined when both the number of successes and number of failures are larger than 5.

The continuity correction accounts for the rounding from a continuous approximation to a discrete one. Essentially

$$Pr(X = x) = Pr(x - 0.5 < X^* < x + 0.5)$$

where we use a range (of width 1) around the point.

5 Statistical Modeling

Modeling data by its signals and noise.

Data consists of the signal, noise, and dirt

- The signals are the relationships, explaining trends
- The noise are the random variations
- The dirt are the mistakes we hope to remove

A model consists of some signal and some noise.

Every model is of the form: response = an equation plus a random error. The random error must have a probabilistic distribution. The error simply represents the part that can't be explained by the equation.

The deterministic part of the model is often pretty simply (a constant mean or a linear equation). The noise part is often just a normal distribution.

A statistical model is the combination of a deterministic function and some random (with distribution) error.

Stages of statistical modeling

- Formulate model
- Estimate parameters
- Check the assumptions of the model
- Use the model to estimate quantities of interest / inference.

5.1 Modeling process

The predicted value $\hat{y}$ is the value predicted from the deterministic equation.

The observed value y is the sampled value.

The estimated error is the residue between the predicted and observed value

$$\hat{e} = y - \hat{y}$$

(Error represents population concepts, Residuals represent sample concepts).

The residual standard deviation measures the variability of the residuals. We use the residuals to estimate the population errors

$$\hat{e}_i = e_i$$

For a population, its mean, standard deviation, correlation, etc are denoted to be parameters with Greek letters. For a sample, the sample mean, standard deviation, etc, are denoted to be sample statistics with English letters.

The formula on the residual standard deviation is

$$s_{\hat{e}} = \sqrt{\frac{\sum(e_i)^2}{n - df}}$$

The *df* is the number of degrees of freedom. A linear regression model has $df = 2$. The degree of freedom refers to the number of estimated parameters (not including the residual standard deviation).

No explanatory variables When we have no explanatory variables, when the error distribution is normal, the residual standard deviation is sample the sample standard deviation. We are assuming that the errors/samples are independent.

$$\begin{aligned} y_i &= \bar{x} + e_i \\ e &\sim N(0, \sigma) \end{aligned}$$

Numeric Explanatory variables Apply a polynomial fit

$$\begin{aligned} y_i &= \alpha + \beta x_i + \cdots + e_i \\ e &\sim N(0, \sigma) \end{aligned}$$

Categorical Explanatory variables Apply grouped fits, an index of each category fit.

$$\begin{aligned} y_{ij} &= f_j(x_i) + e_{ij} \\ e &\sim N(0, \sigma) \end{aligned}$$

When faced with a categorical grouping, we either model the collective dataset (without the groups), or account for the groups, but use the same error/residual distribution. To compute the grouping residual standard deviation, we sum the squared model's entire residuals and divide by the number of data points minus the degree of freedom (other sample parameters/statistics used).

Extrapolation is an act of faith.

Model performance is dependent on the random error standard deviation. The better the models, the smaller the residual/error standard deviation.

The best estimates of the model parameters are to minimize the sum of squared residuals. To compute these parameters (means, slope, intercept), we use computers. We also use computers to compute the residual sd.

Modeling Assumptions Assumptions we have to check for,

- That the errors are **independent**, **normally distributed**, and has zero mean and **constant sd**.
- The equational part is suitable; the **mean error** is zero. This is always guaranteed if the model is a least squared regression

We can check for normality using a QQ plot or probability plot. We can verify independence of errors by looking at the study design (data collection). Constant sd in errors implies that the residues are not affected by the explanatory variables — we plot the residues for each sampled explantory values (scatterplot) to check for constant spread and zero mean of residues.

The null model assumes no groups (constant deterministic part), and treats the whole dataset as originating from a single mean defined population.

Reasons for inference

- To give precision information of out estimates and predictions (confidence intervals)
- Decide which model gives the simplest and most appropriate model

6 Sampling Distributions

We are interested in how the sample estimates of the population parameters varies based on the samples. We want to know the distributions of the parameters from the sampling distribution.

The purpose is to

- Comment on how close the estimated parameters will be
- Whether the model is reasonable
- The range of plausible population parameters

Sampling distribution of sample means The distribution of all sample means has the properties

$$\begin{aligned} E[\overline{X}] &= \mu \\ SD[\overline{X}] &= \frac{\sigma}{\sqrt{n}} \\ \bar{X} &\sim N(\mu, \frac{\sigma}{\sqrt{n}}) \end{aligned}$$

Note that this distribution is normal only when n, the sample size of each sample, is large. This is because the mean of samples is only normally distributed when the sample size is high, from the CLT.

Law of large numbers The law of large numbers states that the sample mean is a better estimate to the population mean as the sample size increases. The more measurements we take in a specific sample, the closer to its population mean the sample mean will be.

Terminology

- Parameters are summary measures of the population data
- Statistics are summary measures of the sample data
- Estimators is a random variable that is the set of all estimates (for instance, $\bar{X}$ is an estimator for the population mean)
- Estimate is a realisation of an estimator. (for instance, the actual sample mean from the sampling distribution of sample means estimator)
- The sampling distribution of a mean is $\bar{X}$. It is the distribution of all possible sample means. Its mean is $\mu_{\bar{X}}$, and its sd is $\sigma_{\bar{X}}$

- The sampling variability is the spread of the sampling distribution of means, $SD[\overline{X}]$. It is the standard deviation of sampling distribution of means.
- The standard error is the estimated standard deviation of sampling distribution of means using the sample sd instead of the population sd, namely

  $$SE = \frac{s}{\sqrt{n}}$$

  It is an estimate for $SD[\overline{X}]$.

For proportions

- The estimator is $\hat{P}$, the random variable of all sample proportions of size n
- The mean of the sampling proportions is $E[\hat{P}] = \mu_{\hat{P}} = p$, the sd of the sampling proportions is $SD[\hat{P}] = \sigma_{\hat{P}} = \sqrt{\frac{p(1-p)}{n}}$.

Computing Probability Interval, Population SD To compute the probability interval of the sample mean using a sample with size n, we use the population standard deviation

$$\sigma_{\overline{X}} = SD[\overline{X}] = \frac{SD[X]}{\sqrt{n}}$$

If our sample mean is within 95% of the sampling distribution, then the true mean is also within 95% of our sample mean. Thus our population mean would be within

$$[\bar{x} - 2\sigma_{\overline{X}}, \bar{x} + 2\sigma_{\overline{X}}]$$

95% of the time, for we assume that the sampling distribution mean is normally distributed around the population mean.

For a binomial distribution, we will conduct the same calculations, but approximating the binomial distribution using a normal distribution, where $\overline{P}$ is the random variable for the sampling distribution proportion, and p is the known (often assumed) population proportion. We know that

$$\overline{P} \sim N(p, \sqrt{\frac{p(1-p)}{n}})$$

We can then compute the p-value given our sample proportion $\hat{p}$

$$Pr(\overline{P} \geq \hat{p})$$

Probability interval Interval A 95% probability interval for $\overline{X}$ is an interval where we expect 95% of the time, the sample mean to lie within the interval.

Probability Interval, Sample SD For the normal case, we can estimate the population sd σ using our sample sd s.

For the binomial case, we can estimate the population proportion p using our sample proportions $\hat{p}$.

Then we can use this estimate to estimate the standard deviation of sampling distribution $\hat{\sigma}_{\overline{X}}$ or $\hat{\sigma}_{\overline{P}}$ — the standard error.

Notice that the standard error depends on the sample proportions (higher proportions leads to higher SE).

I think this needs a slightly different distribution for confidence intervals.

Prediction interval The range where one sample from the population is likely to be observed. This is straight up using the population distribution.

The prediction interval is

$$\overline{x} \pm z \times \sqrt{\sigma_X^2 + \sigma_{\overline{X}}^2}$$
$$\overline{x} \pm z \times \sigma_X \sqrt{1 + \frac{1}{n}}$$

it represents an interval where we are likely to observe another sample, given the population sd and sample mean.

Now this is due to the variance in the population and sample mean, after we approximate the population mean with the sample mean. In other words, we need to account for

$$V(X + \overline{X})$$

When we also approximate the population sd with our sample sd, we must use a t distribution instead.

Of course, when we know the population parameters, we don't need any of this.

NOTICE THAT THE PREDICTION INTERVAL, THE PROBABILITY INTERVAL and CONFIDENCE INTERVAL ARE NOT THE SAME.

7 Inference

We were using the population parameters to understand the distribution of the samples (with certain probability). Inference is the use of samples to infer parameters of the populations, without probabilities.

Inference consists of

- Interval estimates (Confidence interval)
- Hypothesis testing

First focus on interval estimates. We are attempting to find a range of population means that are compatible with our observed sample mean (given the sd of sampling distribution).

To conduct inference on the population mean μ_X, we use our sample estimate $\overline{x}$ (a point estimate) and some error range (an interval estimate).

A point estimate is the best guess using a single point. The interval estimate is the best guess of the variable in a range.

This section is very fucking close to the last section, albeit more formally.

All confidence intervals have the form

$$\text{estimate} \pm \text{z/t value} \times \text{variability}$$
$$\overline{x} \pm z\sigma_{\overline{X}}$$

the z/t value depends on the distribution of sampling distribution as well as the degree of confidence.

- The CI provides a range of plausible values of the population parameter
- the distribution value refers to the certainty of our estimate range
- The variability refers to the error due to sampling variability (higher sample sizes reduces this number)
- The margin of error is the entire right-hand-side. It is the furthest a plausible the population parameter may be for the given level of confidence
- Increasing confidence level implies the increasing of the range of our CI, and vice versa

Known population SD We use the sample mean to estimate the population unknown mean, and we use the sd of sampling distribution $\sigma_{\overline{X}}$ as the precision of our estimate. We expect that 95% of the time,

$$|\mu_X - \overline{x}| \leq 1.96\sigma_{\overline{X}}$$

where the population mean is within 1.96 of our sd of sampling distribution. This is the margin of error, or known as the 95% confidence interval for our estimate.

More formally, define the random interval

$$(\overline{X} - 1.96\sigma_{\overline{X}}, \overline{X} + 1.96\sigma_{\overline{X}})$$

will contain the population mean μ_X 95% of the time. The realization of this interval $\overline{x} = \overline{X}$ using our sample mean, results in a 95% confidence interval.

Note, we cannot say that the mean is in the interval for the mean is not random. Instead, prefer that there is a 95% chance that the interval contains the population mean.

The CI provides a range of plausible values for μ_X (these are the values of population mean which are consistent with our sample mean).

Unknown population SD We estimate σ_X with the sample sd s_X, and compute the standard error. This results in the same confidence interval formula.

But we pay a price for estimating the population sd, and the z value must be larger and changed to a t value. Essentially, for random samples X_i where $X \sim N(\mu, \sigma)$, then

$$\frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

, where t_{n-1} is the t value of a t distribution with $n-1$ df.

Properties of t_ν:

- Symmetrical, mean zero, bell shape, but more spread out than a normal distribution
- As $\nu \to \infty$, the t distribution approaches normal

The 95% CI in this case is

$$\overline{x} \pm t_{n-1}(0.975) \times SE_n$$

where the t value is from the inverse cdf calculation.

The resulting confidence interval exactly works only when the population is normally distribution. But provided there are no strong skew in the population, t-procedures produce approximate and conservative estimates (overcounts CI). This robustness is due to

- CLT of sample means
- For large sample sizes, using the t dist is fine even if our population variable and sampling distribution is not normal. We know by the CLT it will be approximately normal.

- Law of large numbers state that the sample sd will approximate pop sd as n increases.

T procedure fails when there are outliers or skewness in the population distribution, especially when the sample size is small.

Assumptions

- Check normality of sample/population data
- If sample size is less than 15, use t procedure if data is close to normal. Otherwise dont
- If sample size is at least 15, use t unless outliers or major skewness.
- For even larger sample size, say 40, we can ues t for skewed populations. The more skew, the larger the sample size.

The assumptions for the confidence interval are

- The observations are random, independent samples from the population (how were the data collected)
- For small sample size, the population must be normally distributed (normal score plot)

Difference between the prediction interval, probability interval, confidence interval

- A confidence interval uses the sample statistics to predict the location of the population mean. It is always about a population parameter.
- A probability interval uses the population parameters to predict the location of the sample mean
- A prediction interval uses the parameters/statistics to predict the location of one sample from the population.

Sampling Distribution of Proportions For the proportions distribution, we often don't have the population mean. We then will simply estimate it using the sample proportions with

$$SE(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}$$

Importantly, we do not use a t distribution here because we have not used the sample sd. In reality, the sample proportions approximates the population proportions quite well for a z test.

In general for confidence intervals. When the 95% CI contains the expected population mean, we cannot be certain that the true population will be within the interval still.

Therefore, we will need more samples to reduce the CI or increase the strength by changing to a 99% CI in order to make the conclusions (or at least a more certain conclusion).

7.1 Sample size

We can compute the sample size required to get the 95% confidence interval under a specific range. We can just do algebra on the equation, and solving for n. If the population sd is unknown and we need to the t distribution, put it in a computer or trial and error.

For proportions, we cannot set a specific population sd, so we use $p = 0.5$ for a conservative confidence interval, and solve for n. This is due to the fact that in the worst case, the sd of sampling proportions will be maximized and we will overshoot the CI.

8 Inference, Hypothesis testing

We are trying to test if a sample mean belongs to a specific population. Obviously we can use a confidence interval on the sample mean, and check if the population mean is within the interval. Alternatively we can do a hypothesis test.

We will assume the null hypothesis, that the population parameters are suspected. We use the population parameter to test the z value of our sample mean, and compute the p value associated.

The process for a hypothesis test

- Define question/hypothesis, significance level
- Assume the null hypothesis, compute estimator
- Compute p value using our test statistic (sample)
- Decision between the two models
- Conclusion

More technically

- State hypothesis in terms of the parameter tested, and the significance level α
- Assume null hypothesis is true. Compute SE using sample or SD of sampling distribution.
- Compute test statistics under H_0, Determine p-value
- Reject or retain H_0
- Step conclusion in terms of the problem.

Statistical significance must be from a hypothesis test.

Terminologies

- Null hypothesis is the current best (boring) hypothesis, H_0
- Alternative hypothesis is the hypothesis we are testing, and hopes to be true, H_1
- One tail tests tests whether model that the mean is less than H_0, or greater, but not both. A two tail test tests both, that the mean is not equal to the null mean. Use a two tail test when uncertain
- Test statistic is just the standardized value of the sample stats. It measures the distance of the observation to the null mean.

- The p value is the level of extremeness observed, assuming that H_0 is true. The direction matters here.

One tail test We compute the p value by computing the cdf of getting the sample mean or more extreme. If the p value is less than our alpha, we can conclude our alternative hypothesis that our sample population has a pop parameter of less than (or greater than) the null hypothesis mean.

If our alternative hypothesis is that $\mu > x$, and our sample $\bar{x}$ is below x, then the p-value is still calculated as

$$P(X \geq \bar{x}) > 0.5$$

Two tail test Used for two tail p-values when our alternative hypothesis is that the population mean is not equal to a specific value (without a direction).

Small p-value is evidence that H_0 is wrong,

- If $p < 0.05$, there is a strong evidence, statistically significant
- If $p < 0.01$, very strong evidence, highly statistically significant
- If $p < 0.001$, overwhelming evidence

When the null hypothesis is indeed true, the distribution of the p-value would be uniform (think about how to generate the normal distribution using from the cdf); if the alternative hypothesis is true, the p-value distribution would be skewed towards zero with a higher proportion of significant p-values. Here, the proportion of significant p-values is known as the power of the test.

Errors and the cutoffs are decided prior to the hypothesis test. The test is a decision making process on whether to reject H_0.

- If H_0 is true in reality and we accept H_0, this is correct and a true negative
- If H_0 is false in reality and we rejected H_0, this is correct and a true positive. This is called the statistical power and has a value of $1 - \beta$.
- Type I error is when H_0 is true in reality but we rejected H_0. This is called a false positive. Its value is the significance value (alpha level)
- Type II error is when the H_0 is false in reality but we accepted H_0. This is a false negative. It is named beta.

Both Type I and Type II errors are bad. Reducing the probability of type I errors increases the probability of type II errors, and vice versa.

The level of significance is α. It is the cut-off point for the h test, and called the critical value. α defines how prepared we are to make a type I error, with the probability of a type I error (accepting alter when H_0 is actually true) equal to the alpha level.

Statistically significance does not equal to actual importance. The conclusion depends on the α value, and the alternative hypothesis true mean. Similarly, large p value does not imply that the alternative hypothesis must be false, and vice versa. Tests on low sample sizes tends to have low power.

The more difference between the means of two hypothesis, the higher the power of the test.

Tests are based on assumptions such as: normality, independence. If the assumptions don't hold, the inference may not be correct.

Notice that when we are computing the p-value, we use the respective variance (either population variance or standard error), as well as switching to a t-distribution when we are estimating using the standard error.

8.1 H test etiquette

Should publish all the results, instead of only the specific ones.

The conclusion should say either: "We have sufficient evidence, based on our sample, that . . . ", or "We have insufficient evidence, based on our sample, that . . . ". Always include the p-value or significance level, as well as a comment of the sample size and statistical power.

8.2 Paired data inference

We have some before and after data (single blocked designed study), where we are interested in the change of measurements for each study unit (block). We can apply the 1-data hypothesis testing (or confidence interval) using the differences as usual.

Usually, the null hypothesis would be that the population mean difference is zero, and the alternative to be non-zero differences.

To test whether there is a difference, we can also compute a confidence interval, and check if the population mean of zero is a plausible value (in the CI), or not, then using the fact that the mean difference is likely positive or negative to make conclusions. A 95% CI will make the same conclusion as a two tail 5% significance hypothesis test.

In our report, mention

- The type of study
- the sample size

- The significance
- Stats: P-value from h-test, significance and CI, rejection or retaining of null hypothesis, distribution value
- In context of the study

If we cannot use the mean due to outliers (or if the sample distribution is not normal), we can do inference on the median.

8.3 Proportions, Hypothesis testing

Similar approach, our null hypothesis would be

$$H_0 : p = 0.05, \qquad H_1 : p \neq 0.05$$

or greater or lesser.

We then approximate the sampling distribution of p to a normal distribution (requiring that n is large, use the hypothesis p for the normal approximation step). Then compute the sample statistic using our sampling distribution $\hat{P}$, using the z-distribution.

Notice that for the CI of proportions, we use the sample proportions and SE for the interval. For the hypothesis test, we use standard deviation of sampling distribution instead to compute the test statistic.

The critical value in general refers to the most extreme z-value we expect as our test-statistic under the null hypothesis.

Exact proportions hypothesis test We use the binomial distribution to test the probability (respecting the tail) to get our sample value or more extreme (so inclusive inequality), this will be our p-value.

We will NOT need to do a continuity correction on proportion hypothesis testing. We will need to know that the continuity correction may significantly change the p-value, and this is worse with more extreme p or smaller samples.

We must say that for proportions, the p-value is approximate.

Assumptions

- n is large
- the samples are independent
- p doesn't change

8.4 Inference on median

Applies when our sample is not normal and is very skewed. We use the sign test to judge whether the parameter is positive, equal, or negative.

Our hypothesis may be

$$H_0 : \text{median} = 0 \qquad H_1 : \text{median} > 0$$

Under the null hypothesis, we would expect *6=that there are equal numbers of positive and negative values. That is, the proportion of positive values are $p = 0.5$. We can then reframe this question to a proportions test on the proportions of positive values:

$$H_0 : p = 0.5 \qquad H_1 : p > 0.5$$

and continue with our sample dataset, with sample proportions as a statistic instead. We can use a binomial distribution here for the sample size is likely to be small, where the p-value is the probability of getting the observed number of events or more.

For sign tests, completely ignore the group with zero differences (remove them from the sample size).

9 Comparative Inference

For two-tailed hypothesis testing, the confidence interval and h-test will give the same result for a given significance level.

We can also compare critical values with the test-statistics, which should give the same outcome as the p-value, significance level approach for both one and two-tails tests (a bit nuisanced).

For two independent populations, we can compare their parameters:

$$\mu_1 - \mu_2 \qquad p_1 - p_2$$

and note for any differences between the two populations.

This comparative inference applies for studies with only one categorical explanatory and one numeric response variable. With only two possible explanatory variables, we have two respective

$$X_i \qquad \mu_i \qquad \sigma_i \qquad \overline{X}_i \qquad \sigma_{\overline{X}_i}$$

Therefore when X_1 and X_2 are independent normal variables

$$\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

by the expected value and variance theorems on independent random variables.

This is the exact distribution when we know the population sd.

9.1 Equal variances, SE

There are other words for this equal variance assumptions.

When the std devs are equal, we have that

$$\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

We can of course estimate $\overline{X_1} - \overline{X_2}$ using $\bar{x}_1 - \bar{x}_2$. To estimate for the population standard deviation σ, we can combine the two sample standard deviations using

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

which is also the std dev of the residuals of the separate mean model (using two sample means). The pooled standard deviation should be between the two sample standard deviation.

Note that this is the residual std in statistical modeling for it provides an estimate to the average mean residuals of each population using the separate means model, a fair comparison.

Hence we can estimate

$$\sigma_{\overline{X}_1-\overline{X}_2} = SE = s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

and the confidence interval will be

$$(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2} \times SE$$

Assumptions for the equal variance method is that the two sample VARIANCES are not more than a multiple of 2 away

$$0.5 < \left(\frac{s_1}{s_2}\right)^2 < 2$$

To conduct a hypothesis test,

$$H_0 : \mu_1 - \mu_2 = 0 \qquad H_1 : \mu_1 - \mu_2 \neq 0$$

and assume H_0, with equal difference variance. Then continue as usual with SE or sd of estimator as the variability factor. Remember the different degrees of freedom of the t-dist test statistics.

When H_0 is retained, we are saying that there is no significant groupings and that a one population model provides a lower s_e. The alternative hypothesis suggests a separate means model.

9.2 Unequal variance, SE

An approximate standard error for the mean differences is

$$SE = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

with the test-statistic being an approximate t-distribution with degrees of freedom:

$$\nu = \min(n_1 - 1, n_2 - 1)$$

This provides a conservative estimate of the CI, and worse than the pooled std dev model for the low degrees of freedom.

Note that when the sample sizes are equal, the estimator standard error is the same whether we assume equal variance or not. The pooled method results in the same sd as the non-pooled, approximate method.

9.3 Two sample t assumptions

In general for the two sample t procedures

- The sample size can be small, for the t procedure is pretty robust. The sample sizes are relatively similar
- The population distributions are both approximately normal, checked using probability plot of samples
- The measurements within each group is random and independent, checked using study design
- The two sample groups are independently sampled (between each other), checked through study designs
- (Optional, for the pooled sample sd method), that the two population variances are equal $\sigma_1 = \sigma_2$, checked using the ratios.

We must not study matched pair, difference data using a two sample mean test, as they violate the group independence assumption. This is because the differences between the two populations are clouded by differences within the populations.

9.4 Two sample CIs

When the confidence intervals of both sample groups overlap, we cannot be sure that their difference is zero. We must use a two sample t hypothesis test to check the null hypothesis.

Assuming equal population std dev, and equal sample size, the CI of each population can overlap by a maximum of

$$1.96\frac{\sigma}{\sqrt{n}}(2-\sqrt{2}) \approx 1.15\frac{\sigma}{\sqrt{n}}$$

9.5 Two sample proportions test

To compare the proportions from two independent population, we can extend the one proportions method.

We will assume that the difference between the proportion estimators is approximately normal. The distribution of the difference of the proportion estimators is

$$\hat{P}_1 - \hat{P}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

We will estimate the difference of population proportions by the difference of sample proportions, and we will estimate the std dev of estimators by substituting p_1 and p_2 with $\hat{p}_1$ and $\hat{p}_2$. We will always use the z distribution for proportions.

The confidence interval

$$\hat{p}_1 - \hat{p}_2 \pm z \times SE$$

is approximate, and gives the plausible population proportion differences.

For hypothesis testing

$$H_0 : p_1 - p_2 = 0 \qquad H_1 : p_1 - p_2 \neq 0$$

we compute the estimator std dev by the null hypothesis, where both p are equal, and is

$$\sigma_{\hat{P}_1 - \hat{P}_2} = \sqrt{p_0(1-p_0)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

and $p_0 = p_1 = p_2$.

But we don't know p_0, but we can estimate the population proportions from the sample by merging the two groups (for their proportions are equal)

$$\hat{p}_0 = \hat{p}_p = \frac{x_1 + x_2}{n_1 + n_2}$$

and we can create a standard error of

$$SE = \sqrt{\hat{p}_0(1-\hat{p}_0)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Remember that p-values for a proportions test is always approximate.

Assumptions

- Distributions are assumed to be binomials
- Large enough sample sizes, higher than 5 expected successes and failures for both groups, relatively close sample sizes
- Independent samples between the groups and within the groups

10 Study design (B)

The statistical power of a test is the probability that we reject the null hypothesis when the null hypothesis is false. It is the chance that we detect an outcome that is not null.

To increase power

- Decreasing sd dev of estimator (decreasing population std dev, increase sample size)
- Increasing alpha, significance level
- Increasing the absolute difference in mean

One-sided tests are more powerful than two-sided tests on the same level of significance.

Too much power can have problems in: unnecessarily large sample sizes, or increasing Type I errors.

10.1 Power Curve

The power curve graphs the difference between two population means (or the population mean and the null hypothesis) with the power.

Power curve features

- the significance level is found at the minimum, the power when the difference between means is zero.
- the curve is symmetric when the 2-sample test is normal/t-distributed and two-tailed; the curve is single sided (no below zero difference) when the test is one-tailed.

10.2 Degrees of freedom

Replication improves precision: the variation of our estimates. It is either the increase of sample sizes or the increase of measurements.

Degrees of freedom is the number of things measured subtracting the things estimated. Adequate replication ensures that we have some degrees of freedom for our analysis; more replication allows us the flexibility in considering more sophisticated models via blocking/grouping.

11 Analysis of Variance

We wish to compare the means of multiple groups.

ANOVA targets numerical response variables with categorical explanatory variables. The parameters of interest include a set of means in different categories. The figures used are comparative box plots, dot plots, and histograms.

In an ANOVA

- We must consider the model used in the hypothesis test
- Use a test statistic that models the s_r of all candidate models, F-ratio
- We aim to check the significance of the model residuals to compare between models more systematically

For example, the null model can be the single mean model, and the alternative model is the separate means model. The ANOVA test will test if the population means for each categorical variable is really different (the alternative hypothesis) or are equal (the null hypothesis).

ANOVA idea

1. Formula the null and alternative model/hypothesis

   $$H_0 : \mu_i = \mu \qquad H_1 : \sigma_{\mu_i} > 0$$

   where H_1 is at least one mean differs. The alternative hypothesis is two-sided.

2. Compute mean of null model, compute the residuals and residual standard deviation s_r (using the formulas from statistical modeling), under

   $$y_{ij} = \mu + e_{ij} \qquad e_{ij} \sim N(0, \sigma_e)$$

3. Compute means of alternative model, compute residuals and res std dev s_r, under

   $$y_{ij} = \mu_i + e_{ij} \qquad e_{ij} \sim N(0, \sigma_e)$$

4. To test if there is a significant difference between the s_r of both models (does H_1 fit the data better than H_0 that is not due to chance?), we check if the variation is large between categorical variables (variance in sample means) compared to the variations within each group (variance within each sample group)

Notice that the alternative model residual std dev is the extended pooled sample standard deviations for all categorical groups

$$s_r = s_p = \sqrt{\frac{\sum_i (n_i - 1) s_i^2}{\sum_i n_i - n}}$$

11.1 H-Testing

We partition the total variability in the response into

- Explained variability, the variability between the means
- Unexplained variability, errors, the variability within each group (that is unexplained by the sample means)

A good model should have a lot of explained variability over the unexplained.

The ratio between the explained and unexplained variability is the f-ratio under the f-distribution[3]. The f-ratio test will always be one-tailed, where the H_1 is always such that one mean is not equal to the others.

Equation wise, this is

$$\begin{aligned} \text{Total variability} &= \text{Var between groups} + \text{Var within groups} \\ SS_{\text{total}} &= SS_{\text{group}} + SS_{\text{error}} \\ \sum_{ij}(y_{ij} - \bar{y})^2 &= \sum_{i} n_i(\bar{y}_i - \bar{y})^2 + \sum_{ij}(y_{ij} - \bar{y}_i)^2 \end{aligned}$$

Where SS stands for the sum of squared residuals. Notice that n_i weighing of the SS of groups.

Essentially, the LHS is the s_r of the null hypothesis model, the RHS is the s_r of the alternative hypothesis model, plus the variance between the sample means.

The mean squared of residuals statistics are computed by dividing the sum of squared residuals by the degrees of freedom of that specific statistic. It represents the weighted s_r^2 for the respective model.

The f-ratio is computed by dividing the MS of group by the MS of errors. It divides the explained variance by the unexplained variance. The greater the ratio is, the more unlikely it is to be found under the null hypothesis, and the "better" the model is. (A better model implies that the improvement in s_r is not due to sampling variability).

Source	DF	SS	MS	F	P
Treatment (explained)	$k-1$	SS group	MS group	MS group / MS error	from F-ratio
Residual (unexplained)	$n-k$	SS error	MS error		
Total	$n-1$	SS total			

The f-distribution $F_{\nu_1\nu_2}$ has two degrees of freedom as parameter, with ν_1 being the df of the numerator, the explained df; and ν_2 being the df of the denom, the unexplained df.

[3] The f-distribution is the distribution of f-ratios under the null hypothesis, that there is no difference between the sample means.

Notice that MS error square rooted gives back the H_1 residual std dev, and SS total divided by $n-1$ gives the H_0 residual variance.

The alternative model's R^2 is the percentage of total variability in the response explained by the model, namely

$$R^2 = \frac{SS_{\text{group}}}{SS_{\text{total}}}$$

the higher the R^2, the better the model in explaining the total variability.

To state the conclusion of a f-test:

> Based on these samples, there is statistical evidence that the mean blah differs between blah treatments $(F_{\nu_1 \nu_2}, p)$ for the blah population. Also state the CI (or point estimates) of the means for each group.

Assumptions:

- The residuals of the separate means model are normally distributed, checked using a probability plot on the entire residuals of the separate means model.
- The standard deviation of the response variable for each treatment is equal. Checked by
$$0.5 < \frac{\max s}{\min s} < 2$$
without squaring when there are more than 2 means. We can also just plot the residuals against the categorical values (as in statistical modeling) to check this constant variance of residual assumption visually.
- Residuals are random and independent between and within groups (also that the samples within and between treatments are independent). This can be checked only by study design.

11.2 Difference between means

When there is statistical significance that the means are different. To confirm which means are different: if the CI between two sample means don't overlap, they are def different.

The key is to use the pooled std dev s_p, which is the square rooted MS_{error} from ANOVA.

To compute the CI for each sample mean (categorical group), we use

$$CI = \bar{x}_i \pm t_{df_{\text{error}}} SE$$

where the standard error is the pooled standard deviation (as we assumed that the standard deviations are equal for the population). This SE is

$$SE = \frac{s_p}{n_i}$$

with n_i being the samples within the ith group.

Another way is to construct an CI for the difference between two means, with the formula

$$CI = \bar{x}_1 - \bar{x}_2 \pm t_{df_{\text{error}}} SE$$

and

$$SE = s_p \sqrt{1/n_1 + 1/n_2}$$

with n being the sample sizes within the two categories.

11.3 Balanced design differences

Furthermore, if the samples in each group is equal, notice that the variability is always constant due to the pooled std dev, let

$$LSD = t_\nu SE = t_\nu s_p \sqrt{2/n}$$

we can then group the different sample means, grouping the means if they are within this LSD (least significant difference) and splitting it into another group otherwise (draw a horizontal line that connects the treatment means that are similar). We can therefore identify the groups of categories that are similar.

These confidence intervals of the difference between means are the Fisher Confidence Intervals. This grouping is called the Fisher Individual Test.

However, due to the high amount of CI tested, the probability that one CI fails increases, and the probability of type I error increases. The simultaneous significance level (confidence) will be lower (we don't need to compute the simultaneous confidence by hand).

To account for this difference in total significance level, we use Tukey's confidence Intervals, which is computed by Minitab. Tukey's CI fixes the simultaneous confidence level, with the result is that the individual difference confidence intervals will be higher.

Comparison between Fisher and Tukey's CI methods

- Fisher interval fixes the individual error rates, the simultaneous error rate will be higher. Use if we wish to identify all potential actual differences, reducing Type II errors
- Tukey interval fixes the simultaneous error rates, the individual error rate will be lower. Use if we wish to identify all significant differences, reducing Type I errors

Tukey intervals are more conservative and should be used when comparing more than two populations. When there is only two sample means, they give the same intervals because there is only one interval.

11.4 Unbalanced design differences

We cannot use the LSD method, but would have to recompute the standard errors of each Fisher interval between every category pair due to the different sample sizes within the groups.

12 Linear Regression

A categorical, separate means model for linear regression is a one-way ANOVA on the categorical explanatory variable. We are seeing if the explanatory model is better than the null model.

Numeric linear regression models the relationship between two continuous numeric variables. This is essentially fitting a separate mean to each numeric value.

12.1 Simple Linear Regression

The linear regression model for a numeric explanatory and a numeric response variable has the form

$$y_i = \alpha + \beta x_i + e_i \qquad e_i \sim N(0, \sigma)$$

The assumptions are that: normal errors, independent random residuals, equal variance of the residuals, and a linear relationship.

The regression line refers the deterministic model part

$$y_i = \alpha + \beta x_i$$

remember that the residuals are the differences between observed data $\hat{y}_i$ and the predicted data y_i

$$\hat{e}_i = \hat{y}_i - y_i$$

We assume that for a given x value, the observational y values are distributed normally around the regression line with equal variance (model error). Therefore, more of the data points are within 2sd about the line.

To estimate the parameters of the simple linear model, we estimate the coefficients

$$y_i = \hat{\alpha} + \hat{\beta} x_i + \hat{e}_i \qquad \hat{e}_i \sim N(0, s)$$

with the estimated (predictive) equation, or the fit of the dataset as

$$E(\hat{Y}|x) = \hat{\alpha} + \hat{\beta} x$$

where we are treating the expected value of the random variable Y as a function of the sampled x.

The line of best fit will always cross the sample x and y means. This means a list of samples with their regression lines will pivot around the population x and y means.

12.2 Method of Least Squares

To determine the coefficients for linear regression, we use the method of least squares.

The coefficients are determined where the sum of square residuals is minimized, that is

$$\sum (\hat{y}_i - y_i)^2 = SS_{\text{residuals}}$$

The parameters are

$$\begin{aligned} \hat{\beta} &= r\frac{s_Y}{s_X} \\ \alpha &= \bar{y} - \hat{\beta}\bar{x} \\ s &= \sqrt{\frac{\sum (\hat{y} - y(x))^2}{n-2}} \end{aligned}$$

where s_X is the sample standard deviations of the x values. The model (residual) sd is simply from statistical modeling.

The α is chosen so that the point $(\bar{x}, \bar{y})$ is always reached by the line.

12.3 Regression ANOVA

ANOVA in general refers to the splitting of variability into the model and errors.

The table contains the regression equation, hypothesis testing on the coefficients, model summary (residual sd), and ANOVA table on the variances of the explanatory and response variables.

The model residual standard deviation s_r is the standard deviations of the unexplained residuals of the simple regression model. It is the spread of the data points around the line.

The R^2 is the percentage of explained variability against the total variability

$$R^2 = \frac{SS_d}{SS_t}$$

Assumptions of the regression model: ANOVA and correct equations

- Normal samples within each explanatory group, with the same std dev. Only check for normality of the alternative, linear model residuals.
- Independent observations and residuals

- Correct line equation, where

  $$Y|x \sim N(\alpha + \beta x, \sigma)$$

  with

  $$e_i \sim N(0, \sigma)$$

  from the overall model. This implies that the observations are normal and centered around the line.

Stuff to check in the residuals vs fitted graphs

- Expecting no patterns: funnel shapes indicate unequal variance (need transformation), curved shape indicates wrong linear equation (use a better equation), unusual vertical points indicate outliers (clean data)
- Normal probability plot for residuals, p-value greater than 0.05 (so likely to be normal)

12.4 Parameter Inferences

Minitab will give out the regression parameters, intercept and slope estimates and SE. These parameters are distributed as a t-distribution with the df the total sample size subtract 2 from the two coefs. The 95% CI are computed then by

$$CI(\alpha) = \hat{\alpha} \pm t_{n-2}(0.975) SE(\alpha)$$
$$CI(\beta) = \hat{\beta} \pm t_{n-2}(0.975) SE(\beta)$$

We can conduct a hypothesis test on $H_0 : \alpha = 0$ or $H_0 : \beta = 0$ using the same SE. It is also possible to do H-test on specific slope or intercept values by modifying our H_0 (if we want to check the slope). The degrees of freedom and distribution of the test statistic is a t-distribution with $n - 2$ degrees of freedom.

From the ANOVA point-of-view, the total variability is partitioned into the explained (model) and unexplained (residual) variability. The explained variability is the distance from the regression line to the overall mean, the unexplained is from the observed to the regression line.

In the ANOVA table for linear regression, the df is: $n - 1$ overall, $n - 2$ for the residuals, 1 for the regression (explained). Everything else is computed exactly as an ANOVA.

In general for regression, the df is: $n - 1$ overall, $n - k$ for the residuals, and k for the regression. k is the number of parameters minus 1 in the deterministic model.

Using the f test statistic, its distribution, and p-value, we can conduct a hypothesis test on H_0 being no linear relationship, and H_1 being some linear relationship. This allows us to check whether our regression model is significant enough and if the population actually follows a linear trend.

12.5 Model Correctness

The number s_r shows how variable the data is around the model. The lower the better, approximately 95% of data points are within 2 s_r around the regression line.

The number R^2 shows the variation accounted by the model. It measures how well the data fits the model, the higher the better. Only for linear regression, does $R^2 = r^2$, with r being the correlation coefficient.

The strength of the linear relationship is computed by the correlation coefficient r. We can do a hypothesis test on $H_0 : r = 0$. This is equivalent to testing for $H_0 : \beta = 0$ with the same assumptions and same data.

When the points are outliers or skewed (both in the x and y direction), they have high leverage (in the x direction) and have usually large standardized residual (in the y direction). They can influence the regression a lot.

Points with high leverage can change both R^2 and the coefficients. Points with high std residuals can do the same.

We don't usually want to remove data points from the analysis. In these cases, report the analysis including and excluding the outliers. If the two analysis produce similar results, use the removed one.

12.6 Response Inference

We can conduct a confidence and prediction interval for the response given an explanatory variable.

The confidence interval is about the distributions of the line (of means) on $\mu_{Y|x}$. The 95% confidence interval of the mean y-value is:

$$CI = \bar{y} \pm t_{n-2} \frac{s_r}{\sqrt{n}}$$

this is because we use the pooled variance to estimate the SE around the y-mean.

When knowing the explanatory value, the confidence interval of the mean at that x-value is:

$$CI = \hat{\mu}_{Y|x} \pm t_{n-2} \sqrt{\frac{s_r^2}{n} + (x - \bar{x})^2 SE(\beta)^2}$$

with the sampling variability also called the standard error of the fit

$$SE(\text{fit}) = \sqrt{\frac{s_r^2}{n} + (x - \bar{x})^2 SE(\beta)^2}$$

This variability contains both the y-mean variability and the slope variability.

The prediction interval is a range of plausible values for an observation $Y|x$. The prediction interval is

$$\begin{aligned} PI &= \hat{\mu}_{Y|x} \pm t_{n-2}\sqrt{s_r^2 + \frac{s_r^2}{n} + (x-\bar{x})^2 SE(\beta)^2} \\ &= \hat{\mu}_{Y|x} \pm t_{n-2}\sqrt{s_r^2 + SE(\text{fit})^2} \end{aligned}$$

The variability here contains the mean variability from the CI, and the variability of point around the mean.

The PI will always be wider than the confidence interval at every point.

In general for regression inferences, we can conclude

- Slope and intercept CI
- Response CI and PI
- Proportion of variability explained by the model, R^2
- The linear correlation between the explanatory and response variable, r
- Hypothesis tests on the slope, and on the model using f-ratio tests, model error std dev

12.7 Method Comparisons

For two sample groups of different numeric explanatory variables, the simple linear regression model, one way ANOVA on the two sample means, and the two-sample t-test produces the same results in

- Model residuals or pooled standard deviations
- The same p-value conclusion on the model hypothesis tests
- The same error/model degrees of freedom
- The CI around the predicted population means (for ANOVA, t-test) and the confidence interval using the regression (on the coefficients) are exactly the same

13 Comparing Several Proportions

The relationship between two categorical variables is called associations. This is the testing of independence between two categorical variables.

To conduct association tests on association tables, we compute the cell probabilities by dividing the observed frequencies by the total samples. The probability table is called a contingency table.

The hypothesis tests on association has the null hypothesis H_0 being that the two factors are independent, the alternative hypothesis H_1 is that the null hypothesis is incorrect.

Assume the null mode H_0 is true, then the probability in each cell of the contingency table is the product between the column and row probabilities. This creates the expected probabilities and frequencies of the cells.

The expected frequency in each cell given the expected probabilities p is and sample size n is binomially distributed with

$$X \sim B(n,p) \qquad E(X) = np \qquad p = \frac{r}{n}\frac{c}{n}$$

The expected count in each cell is computed by

$$x_{ij} = E(X) = n\frac{r}{n}\frac{c}{n}$$

By comparing the observed and expected frequencies, we get a test statistic u^2

$$U^2 = \sum \frac{(\hat{x}_{ij} - x_{ij})^2}{x_{ij}}$$

where $\hat{x}_{ij}$ are the observed frequencies, and x_{ij} are the expected frequencies. The test statistic is distributed by

$$U^2 \sim \chi^2$$

The degrees of freedom is the number of rows minus 1 multiplied by the number of columns minus 1

$$\nu = (n-1)(m-1)$$

It is the number of cells that can be altered to maintain the exact same expected cell frequencies.

A cell's contribution to the test-statistic is simply

$$\frac{(\hat{x}_{ij} - x_{ij})^2}{x_ij}$$

Notice that: expected counts should add up to the row and column sums, the pattern of row and column proportions are identical for all rows and columns.

The p-value is the probability to get a test statistic u^2 or more extreme under the null, chi-square distribution

$$P(U^2 \geq u^2)$$

with $U^2 \sim \chi^2_\nu$. This will always be a one-tailed test with a right tail.

13.1 Deducing Association

To identify the association, select the cells with large contributions to the chi-squared test statistic. These cells have a relative large difference between the observed and expected counts.

Within these large contribution cells, we can compare the expected and observed counts in order to identify the sign of the difference (greater than expected or fewer than expected) depending on the condition (the two categorical variable).

Cells with χ^2 contributions less than 1 are not significant.

Conclusions have the format of:

> With the given data, there is a significant association ($\chi^2_\nu = 40$, p-value < 0.05) between A and B. Namely that the group b_i is associated with higher a_i on average, and group b_j is associated with lower b_i on average. The group b_k has no association with a_k on average.

Assumptions are

- All expected frequencies in cells must be greater or equal to 1.
- All cells have expected frequencies distributed in a binomial way. For a normal approximation, this corresponds to at least 80% of the cells have expected frequencies greater or equal to 5

If we fail to make the expected frequencies assumption, we must combine similarly grouped rows or columns such that the assumption holds.

Technically, the chi-squared distribution with ν degrees of freedom is the sum of ν squared z-distributions

$$\chi^2 = \sum^n Z^2$$

hence our cell expected counts are approximated to be normal.

13.2 Goodness of Fit

Testing if a categorical data fits a pre-existing assumption/distribution.

Commonly, this will test if some event had changed the historical distribution of a categorical data. The null hypothesis H_0 is that the proportions are unchanged (the proportions and probabilities match the assumption), and that the event is ineffective in changing the distribution. The alternative hypothesis H_1 is that the proportions changed, and that the event is effective.

The test-statistic is the same normalized squared difference between the expected and observed counts, namely

$$u^2 = \sum \frac{(\hat{x}_i - x_i)^2}{x_i}, \qquad U^2 \sim \chi^2_{n-1}$$

with the degree of freedom $n - 1$, one less than the number of total samples. The p-value is computed the same way.

To make association inferences, we can compare the expected and observed frequencies once the difference is deemed to be significant.

For conclusions, state

- Effective or not
- Test statistics and p-value
- If significant, differences between observed and expected counts for each categorical value, and the precise increase and decrease percentages
- The event or campaign and its success
- Always in context

13.3 Summary

The varieties of tests for each variable type pairs are

Response	Explanatory	Graphs	Tests
numeric		dotplot, boxplot, histogram	1 sample t, 1 sample z
categorical		table, bar chart	1 proportion, χ^2 goodness of fit
numerical	categorical	dot box hist	2 sample t, 1 way ANOVA
numerical	numerical	scatterplot	linear regression
categorical	categorical	contingency table	χ^2 association