# **Producing data**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Key concepts | | | | | |
| Statistical significance  An observed effect is so large that it would rarely occur by random chance  Population vs. sample  Sample: a subset of the population that we actually want to examine (but cannot due to challenges)     * Parameter (number that describes population) vs statistic (number that describes sample, often to estimate unknown parameter) | | | | | |
| Design of studies | | | | | |
| Retrospective study | | Observational study | | Designed experiments | |
| * Uses historical data over a period of time. Any time period is okay, so long as justifiable to be representative of what you want to investigate. Rule of thumb is 1-3 years so that it is representative. | | * Observes a population of interest, usually carried out over a relatively short period of time and allows variables not usually measured to be collected * Variables of interest are measured but there is no attempt to influence the responses   \*in some ways, addresses flaws of retrospective studies, as data collected will be relevant (operationalised as you want, relevant factors also collected, more current) | | * Deliberate changes made to controlling variables, resulting change in dependent variable observed * Controlled environment, easier to identify other relevant factors   Representing relationships found in data collected through a mathematical expression.  Mathematical expression, if statistically valid, can be used to make predictions and thus support/disprove hypothesis. | |
| Design of experiments | | | | | |
| Key terms and concepts | | | | | |
| * Experimental units: individuals on which the experiment is done. When the units are humans, they are called subjects * Treatment: the specific experimental condition applied to the units * Factors: the explanatory variables in an experiment – many studies often examine the joint effects of several factors. In such experiments, each treatment is formed by combining a specific value (a level) of each of the factors.   The need for a control group  Without a control group, we cannot tell whether the observed response was due to the treatment or unobserved lurking variables. | | | | | |
| Randomised comparative designs | | | | | |
| * Systematic randomisation produces two groups of subjects that we expect to be similar in all respects before application of treatment * Comparative design helps ensure that influences other than factor operate equally on both groups * Differences in outcome must be due to treatment or chance assignment of subjects in control and experimental group  1. Compare two or more treatments – this will control the effects of lurking variables on response 2. Randomise – use impersonal chance to assign experimental units to treatment and control groups 3. Repeat with large samples – to reduce random variation in the results | | | | | |
| Completely randomised design | | Randomised block design | | Matched pair design | |
| **Working principle:** randomly assign subjects to control or treatment with equal probability of selection.  **Method:** draw with replacement (but without repetition)   1. Using a table of random digits, pick one randomly 2. Check if already picked, if yes, draw again until you get a value that has not been picked 3. Repeat until all assigned to treatment or control groups | | **Working principle:**   * Completely randomised designs still risk there being a difference in treatment and control groups by sheer nature of randomness, creating risk of lurking variables affecting study conclusions * A block is a group of units that are known before the experiments to be similar in some way *that is expected to affect the response to the treatments*. * Random assignment of units to treatments is carried out separately within each block to create treatment and control groups that are more assuredly similar.   **Method:**   1. Consider all x as a block, all y as a block… 2. Within each block, randomly allocate units to control and treatment groups of equal size n | | **Working principle:**   * A special case of randomised block designs, where block size = 2 * Subjects are matched in similar pairs – producing results more precise that simple randomisation as matched subjects are more similar and comparing their responses is more efficient than comparing the responses of groups of randomly assigned subjects * Which of the two subjects is control and treatment is randomly decided * Common variation: subject is same person, just before (control) and after (treatment) | |
| Sampling design | | | | | |
| We gather information about a subset of a population – a sample – to draw conclusions about the general population. How we proceed with collecting this information (sample designs) determines how far our conclusions can apply to the general population.  Cautions of sampling design   * Undercoverage: when some groups in the population are left out in process of choosing the sample. Leads to non-generalisable conclusions. * Non-response: leads to biased conclusions * Biased questions/data-collection | | | | | |
| Voluntary response sampling | Simple random sampling | | Stratified sampling | | Multistage sampling |
| * Sample consists of people who *chose* themselves to respond. * Tend to be biased because people with strong feelings, especially negative ones, are most likely to respond. | * Sample consist of *n* individuals from the population chosen in such a way that every set of *n* individuals has an equal chance to be the sample selected | | * Completely random sampling risks the sample not being representative of the population due to randomness, resulting in conclusions not being generalisable * Population is divided into *similar* groups (strata), then SRS in each stratum in done. Combine each SRS to form full sample. | | * At each stage, population is split into clusters (that make sense – e.g. geographical location). Probability sampling is done to choose a cluster, then probability sampling again to choose a sub-cluster, and so on, until desired sample size is reached. * More logistically efficient than SRS or stratified sampling as process iteratively reduces size to randomise from |
| Bias of sample statistics | | | Variability of sample statistics | | |
| Bias concerns the centre of the estimator distribution: a statistic used to estimate a parameter is unbiased if the mean of its sampling distribution is equal to the true value of the parameter being estimated.  An experimental design is biased if it systematically favours certain outcomes.  Reducing bias from experimental design   * **Randomised comparative design** * **Double blinding:** neither subjects nor experimenters know which treatment subjects received – avoiding unconscious bias   Reducing bias from sampling design   * Avoid undercoverage * Ensure data collection (e.g. surveys) are not biased * Use probability sampling methods | | | Variability of a statistic is described by the spread of its sampling distribution – determined by sampling design and sample size n.  Reducing variability from sampling   * Use a larger sample   Population size does not matter  Variability of a statistic does not depend on the population size if the population is >100 times larger than the sample | | |

# **Preliminary data analysis**

## **Measures of central tendency**

Calculation of all four measures simultaneously can give you some information on presence of outliers, symmetry and skew

|  |  |  |  |
| --- | --- | --- | --- |
| Sample arithmetic mean | Median | Mode | Geometric mean |
| Indicates where the centre of the distribution lies but is strongly influenced by extreme values or outliers.   * Centroid of graph | 1. Order data in ascending order, calculate rank of 50th percentile/2nd quartile value   Percentile rank = 1 +percentile(n+1)  = 1 + 0.5(n+1)   1. Interpolate if necessary, to get median   Central value in ordered set of data, or weighted average of two central values if even number od data points. Less sensitive to extreme values.   * Point that divides area under density curve in half | 1. Sort data into occurrences 2. Count number of each occurrence – sample mode is the most frequent one   Most commonly used with categorical rather than numerical data (eg. Good/bad vs rating from 1­­-5).  Note that can have multiple modes. | Often used in averaging values that represent a rate of change because variable follows an exponential (i.e. logarithmic law), hence often used for bacteria growth/population growth.  Geometric mean is always less than or equal to the arithmetic mean, and more resistant to extreme values. |

## **Measures of dispersion**

|  |  |  |  |
| --- | --- | --- | --- |
| Range | Interquartile range | Variance & standard deviation | Coefficient of variation |
| Range = max – min | 1. Calculate 75th and 25th percentile values (3rd/1st quartile) 2. IQR = Q3 – Q1   IQR is less affected by extreme values and outliers but may be less representative because it only examines dispersion in part of the dataset.   * *p*th percentile: value that has *p*% of observations that fall at or below it * Q1 is the median of the observations left of the median, Q3 for the right | =  Both measure average deviation from central value (mean). The use of is a bias correction for using sample mean to estimate population variance.  Be careful of units!   * s measures spread about *mean* and should only be used when mean is chosen as measure of centre * not resistant to outliers | The CoV normalises the amount of dispersion to allow for comparison of dispersion across studies with different units and contexts.  If CoV < 1, not very dispersed.  Generally, for a normal distribution, 68% of all data within 1 deviation from mean, 95% within 2 deviations, ~all data is contained within 3 deviations (99.7%). |

## **Outliers and shape**

|  |  |
| --- | --- |
| Outliers | Data points that do not conform to the distribution of values of the rest of the data. Possible reasons:   1. Assumptions of conditions no longer hold (e.g. halfway through collection conditions change notably) 2. Interfering processes can also contribute to outliers 3. Errors of faulty instrumentation, measurement, observation or recording of data   To exclude outliers, appropriate analysis and evaluation should be carried out, it is ultimately a judgement call that must be justified based on whether the outliers are caused by something relevant (or not) to your study. If not relevant, can exclude.  Outlier 🡪 >1.5 IQR from 25th/75th percentile  Extreme outlier 🡪 > 3 IQR from 25th/75th percentile |
| Symmetry and skew | Skew is a measure of symmetry, measuring extent to which a distribution has tails on one side or the other.  Symmetrical 🡪 skew = 0 (e.g. normal distribution)  Tail to the right 🡪 positive skew, majority of data on right of mean  Tail to the left 🡪 negative skew, majority of data on left of mean |

## **Graphical methods**

|  |  |
| --- | --- |
| Line diagram/bar chart | Often used for occurrences of a discrete/categorical variable to show trends or relative amounts |
| Pie chart | Used to compare relative amounts of discrete/categorical variables |
| Dotplot | * Helpful mostly for showing a small pool of data (~<25) * Difficult to see pattern in dispersion, but unusual features (symmetry, outliers) can be seen because data points are there |
| Boxplot | **Workings:**    **Use case:**   * Helpful in highlighting dispersion features (range, IQR) and outliers, individual points not there * Good for comparing data measuring related or similar characteristics |
| Stemplot | **Workings:**   * To read: middle column is first digit(s), side numbers are last digit(s) * Gives a quick picture of shape of distribution while including actual numerical values * Work best for small number of observations all greater than 0 |
| Histogram | **How to plot:**   1. Number of bins ~ 2. Width ~ range/number of bins   **Use cases and problems:**   * Normally for large data set (n>25), because with small dataset histogram shape can change significantly with number of bins, and will not really highlight distribution/symmetry features * Good for compact summary of large dataset – shape of distribution, central tendency and dispersion * Approximates a continuous distribution |
| Time series | **Use cases:**   * Looking at trends across time periods |

## **Describing our data**

|  |  |
| --- | --- |
| Descriptive statistical summaries | * 5-number summary: Minimum Q1 M Q3 Maximum. Usually better than mean and st.dev for describing a skewed distribution or a distribution with significant outliers. * 9 number summary: Sample size, Mean, SE of Mean, Stdev + 5 number summary |
| Qualitative description | Within ranges or totality of data:   * Shape/trend/functional form * Symmetry and skew * Presence of outliers or influential points * Centrality and spread |

# **Random variables and probability**

## **Basics of probability**

|  |  |
| --- | --- |
| Interpretation and axioms | **Relative frequency interpretation:** that the proportion of occurrences of an outcome stabilises to one value in the very long run. This is known as probability and quantified by the relative frequency of occurrence.  **Total probability of all events in sample space = 1**  P(S) = 1  0 ≤ P(E) ≤ 1 |
| Sample space and events | **Sample space:** set of all possible outcomes S = { … }  **Event:** subset of sample space, specific occurrences  **Union:** the event that consists of all outcomes contained in any of the constituent events  **Intersection:** the event that consist of all outcomes that are contained in all constituent events  **Complement:** the set of outcomes in the sample space that is not in the event (i.e. E’)  **Mutual exclusivity:** two/more events that cannot occur simultaneously (i.e. intersection is a null set)  P(S) = 1  0 ≤ P(E) ≤ 1  P(A’) = 1- P(A)  P(A ⋃ B) = P(A) + P(B) – P(A ⋂ B) |
| Conditional probability | Probability that an event B has occurred, given that another event A has occurred. i.e. the relative frequency of (event B) occurring among trials that produce an outcome where (event A) has also occurred.  P(B|A) = |
| Statistical independence | Statistical independence means that by numbers/probability alone, an event occurring does not affect the probability of another event occurring. Ie. The knowledge that an event has occurred does not change what you know about another event.  Note that there still could be a causal/dependent relationship, but it does not manifest and cannot be identified by numbers of relative frequency (i.e. probability).  **Tests: independent if…**  P(B|A) = P(B)  P(A ⋂ B) = P(A) × P(B) |
|  |  |

## **Discrete and continuous random variables**

|  |  |
| --- | --- |
| Probability and QR | |
| By collecting data for QR, we are sampling – choosing a subset of elements from an entire population, with the aim of finding information about the population using the subset (sample). This is because it is often impossible to do a census of the entire population.  Data points can take on a range of values, each with an associated frequency – and each ‘equally representing’ the population. **Random variables**, which can take on a range of values, each with an associated probability (probability distribution) is thus the only way to represent the subject of interest in totality.  **Population:** totality of observations of interest (but infinite of them, cannot possibly observe all)  **Sample:** subset of observations obtained, probability and random variables used to find information that applies to entirety of population | |
| Expectation operator | |
| “Taking expectation” is finding the average value of a variable in the long run, equal to taking the probability-weighted summation of all outcomes in the sample space. By convention, 1 more d.p. than the original value. | |
| Properties: | Notable examples: |
|  |  |
| Discrete random variables | Continuous random variables |
| Random variables with a finite number of possible values.   * Probability distribution is described by **probability mass functions.** * sum of probabilities of all outcomes = 1 | Random variables with an infinite number of possible values.   * Probability distribution is described by **probability density functions.** * Area under PDF = 1 (because each value on x is infinitely small) |

## **Probability distributions**

The distribution of a variable tells us what values it takes and how often it takes these values.

|  |  |
| --- | --- |
| Normal (Gaussian) and standard normal distribution | **Normal distribution** X ~ N(µ, σ2)   * ‘Bell shaped’ PDF, fully described by µ and σ alone. * Used as basis to generate other important distributions * Used often due to Central Limit theorem     **Standard normal distribution** Z ~ N(0,1)   * Standardisation makes calculation easier and allows for use of (left-tailed) statistical tables. |
| Log-normal distribution | If ln(X) is normally distributed, then X is log-normally distributed.  Y = ln(X), Y ~ N(λ, ξ2)  A close up of a map  Description automatically generated |
| Chi-square distribution | X = Z2 follows chi-square (χ2) distribution.  Sum of k number of variables with χ2 distribution gives different shape for PDF. k is known as ‘degree of freedom’. Higher DoF, move from right tail, positive skew towards symmetric normal distribution.  A picture containing text  Description automatically generated  Application of this is in linear combination of variables with normal distribution that are squared (e.g. variance).  Like log-normal, it is a non-linear transformation of normal distribution, but not one-one. |
|  |  |

## **Transformation of data**

|  |  |
| --- | --- |
| Data transformations | Where appropriate, data transformation can be applied to aid visualisation and processing.    But when performing analysis and reporting, we often report in original, untransformed scale as it is more readily understandable. |
| Linear combinations of random variables | Linear combination of random variables gives you another random variable.  Y = X1 + X2 + X3 … +Xn + b (note that in this case, Xi may not be identical variables from the same population. They are any random variable – discrete OR continuous.)  **Distribution:** generally, not the same as X, also not the same as ‘sum of X’. because when combined, outcomes have more permutations, changing distribution – cannot assume. But:   * Will be normal if CLT applies (n ≥ 30) * Will be same distribution as constituent variables if all are same distribution and respectively   + Normal distributed   + Chi-square distributed   + Gamma distributed   **Mean:**  µy = E(Y) = a1 E(X1) + a2 E(X1) … an E(Xn) + b  **Variance:**  σ2 = E[ (Y- µy)2] = a12E(X1 -µ1)2 + … an2E(Xn -µn)2 + sum of covariance terms  covariance = anam E(Xn - µn)( Xm - µm) for all possible combinations. (e.g. 1-2, 1-3, 2-3 etc.)    If variables X1 to Xn are independent, covariance terms = 0. |

## **Normal distribution in practice**

|  |  |
| --- | --- |
| The 68-95-99.7 rule | In a normal distribution with mean μ and standard deviation σ:   * ~68% of observations fall within σ of μ * ~ 95% of observations fall within 2σ of μ * ~99.7% of observations fall within 3σ of μ |
| Central limit theorem | **Theorem**  The sum of a large number (≥ 30) of identical and independent random variables (i.e. n independent observations of same population) has an approximately normal distribution. Distribution tends to normal as n 🡪 ∞.  X1 + X2 + X3 … +Xn ~ N(nµ, nσ2) if n ≥ 30.  E(X1 + X2 + X3 … +Xn) = E(X1) + … E(Xn)  = µ + … µ  = nµ  Var(X1 + X2 + X3 … +Xn) = Var(X1) + … Var(Xn)  = σ2 + … σ2  = nσ2  **Application to sampling and QR**  Normally, population distribution is unknown. CLT allows us to approximate distribution and thus perform calculations and data processing, even if true population distribution is unknown. |
| Checking for normal distribution | **Normal quantile plots**   * Used when histogram and stemplot appear unimodal and symmetrical, but want to confirm normality of distribution * How to plot:  1. Arrange observed data from smallest to largest and record percentile of each data point. 2. Find standardised normal values of z corresponding to these percentiles (e.g. z = -1.645 is at 5% point of standard normal distribution) 3. Plot data point (raw value) against normal score. If plotted line is straight, distribution is close to a normal distribution.     **Histogram:** check to see if uni-modal and symmetrical  **% of standardised residual:** % of points between -1.96 and 1.96 should be close to 95%, as in a standardised normal distribution.  **Mean-median-mode comparison:** if approximately normal, these three values will be approximately equal. |
| Normal distribution calculations | * Probability values * X values given proportion/probability   **Useful manipulations:**  P( a < X < b) = P( X < b ) - P( X < a )  P( X > a) = 1 – P( X < a ) |

# **Samples, statistics and estimators**

|  |  |
| --- | --- |
| Statistics: point estimators of population parameters | To estimate population parameters, a function h(X1 … Xn) (statistic) is applied to observed data to condense them into a single number (point estimate), in which case the statistic is known as a point estimator.  **Notation:**   * Sample statistic, also a random variable (point estimator) * Population parameter **θ** * Output of statistic (point estimate) |
| Evaluating estimators | **Estimator bias (accuracy)**  Bias = E() – θ  Estimator is unbiased when Bias = 0 (i.e. E() = θ). If Bias is negative, underestimate; if positive, overestimate.  Note that a single outcome of a statistic (point estimate) may not necessarily produce the population parameter, but the average of all point estimates for an unbiased estimator will give the population parameter. i.e. biased estimator will give systematic difference from population parameter on average, unbiased will not.  There can be more than one unbiased estimator for population parameters.  **Estimator variance and standard error (precision)**  Estimator variance = Var() 🡪 manipulate by linear combinations to get expression  Because estimator is a random variable, it has both variance and standard deviation.  Standard error: standard deviation of estimator. This is often estimated because of unknown population parameters.  Rule of thumb: if > 1, estimator is too imprecise.  **Assessing estimator**  Mean-square error = bias2 + std error2 = bias2 + estimator variance  Terms are squared to account for negative/positive bias. Sometimes, biased estimators are used if mean-square error is smaller, or for convenience (e.g. standard deviation).  **Notable examples**   * Unbiased estimator of population variance   E(s2)= σ2, Var(s2) = |
| Sampling (estimator) distributions | The sampling distribution of a statistic is the distribution of values taken by the statistic in *all possible samples* of the same size from the same population.   * Normal distributions * Chi-square distributions (e.g. **sample** variance)      * t-distributions   When you want to use unbiased estimate of population variance s2 for sample mean, starting with sample variance, the substitution creates the form of a t distribution. tprobability, n-1 (degrees of freedom) |

|  |  |
| --- | --- |
| Confidence intervals | Reported with **confidence level** and **interval**.  **Confidence level:** a measure of the method used – in an infinite number of trials of this experiment, 95% of the calculated intervals will contain the true population parameter.  i.e. probability of any interval created using this method containing true population parameter = confidence level  We report with \_\_% confidence that the true population parameter \_\_ lies between (interval).  When population variance is known:    When unbiased estimate of population variance is used: |

# **Hypothesis testing**

|  |
| --- |
| Hypothesis testing |
| 1. Model distribution of variable (Z or t) 2. Set null and alternative hypotheses (μ inequality) 3. Find probability of test statistic occurring, if null hypothesis is true (p-value) 4. Depending on level of significance, **reject** or **fail to reject** H0. (**different from accepting**)   Concept: finding probability of event occurring by random chance if certain premises are true, from which you can falsify the premise (or not) – because of its low probability of occurrence. Hypothesis testing is a method of quantifying likelihood of an event to falsify something you want to falsify. It cannot prove something. |
| Errors |
| Type I: wrongly rejecting H0 when H0 is true. Hypothesis testing controls for this error through the level of significance – which is the probability of wrongly rejecting H0 when H0 is true, since the model is built assuming H0 is true.  Type II: wrongly accepting H0 when H0 is false. If this occurs, it means that the ‘true value’ is close to the null hypothesis, even if it is inaccurate.  Practically speaking, this is not likely to be a significant deviation – so you control for this by setting the maximum deviation of true mean from H0 that you can accept (counterpart to level of significance). You then test for it by finding an appropriate sample size that can allow you detect the deviation through hypothesis testing. |

# **Relationships in continuous bivariate data**

## **Relationships between variables**

|  |  |
| --- | --- |
| Variable types | Explanatory & Response: when there is a causal relationship between two variables |
| Association between variables | Positive: when above/below-average values tend to accompany above/below-average values of the other variable.  Negative: when above/below-average values tend to accompany below/above-average values of the other variable. |
| Visualising relationships | Bivariate data: scatterplots  Categorical variables in scatterplots: use a different plot colour for each category |

## **Simple linear regression**

|  |  |
| --- | --- |
| Principles & quantitative implications of a linear model | Single variable: Y = µ + ε – Y has a mean value and is affected by random error.  Multi-variable: If there is a linear relationship between Y and another variable x, we can model it by having a mean value of Y for each specific value of x. ie. is the mean value of Y for each specific value of x. Used for predicting mean values of Y given values of x.  Assuming random error is normally and independently distributed with constant variance (ie. error is truly random, and Y is affected by x and random error only), then:  E(Y|x) = E() = =  Var(Y|x) = Var() = 0 + σ2  So σ2, and are unknown population parameters. |
| Estimators and the method of least squares | To find population parameters, we can use estimators from sample data. From which, we can quantify the probability of occurrence and use that in hypothesis testing – so that our random sample data can give us information about the population.  Estimator of variance    Regression coefficient estimators: using the method of least squares      Distributions of regression coefficient estimators    Estimator of error term (residual)      Acceptable residual plots. Left is ideal, right is ‘bowl shaped’. ‘Bowl shaped’ is acceptable only if there are more points in the middle, since more extreme values are rarer and it does not indicate the true behaviour of the population. |
| Hypothesis testing | **Setting the hypotheses**  Null hypothesis H0: no relationship (= 0)  Alternative hypothesis H1: some relationship </> 0  Rationale – hypothesis testing quantifies type I error to demonstrate there is a statistically significant relationship. So, you use the test to find the probability of obtaining your sample data (with the relationship) if there was truly no relationship, from which you make conclusions.  **Tests on slope**    **Making conclusions**  Check for assumptions of linear regression holding, and risks of type I (level of significance) and type II error. If all good, then there is a relationship. |
| Notes about least-square regression | * The slope and intercept of the line depend on the units of measurement – you cannot conclude anything from their magnitudes * There is a close relationship between **correlation and the slope of the least-squares line:**   β1 = r  Along the regression line, a change of one standard deviation of x corresponds to a change of r standard deviations in y.  Notice that when the variables are perfectly correlated (r = 1 or -1), the change in predicted y is the same as the change in x (in standardised units). So as the correlation becomes weaker, the prediction moves less in response to changes in x.   * Least-squares regression line always passes through the means of x and y. Relating that with the equation for β1, we can solve for the equation of the line by substituting in the mean values of x and y to find β0.   🡪 solve for   * The distinction between explanatory and response variables are essential in regression as it looks at distances from the line only in the y direction. If we reverse the variables, we get a different line. |

## **Coefficient of determination, correlation and covariance**

|  |  |
| --- | --- |
| Coefficient of determination | It indicates amount of variability in Y values that can accounted for by the regression model. Ranges from 0 to 1. Quantifies strength of relationship as it is the fraction of the variation in values of y that is explained by the least squares regression line (prediction) of y on x.        It is always possible to make R2 bigger by adding more x terms to the model (ie. perfect fit can be obtained for n data points with a polynomial of degree n-1).  Large value does not mean model will provide accurate predictions of future; small values need to be evaluated in context (eg. Human behaviour studies, you expect low values). Even if values are low, statistically significant coefficients will indicate at least the direction of relationship. |
| Covariance | Quantifies how two random variables vary together. |
| Correlation | **Working principle:**  Closer the value of correlation coefficient to -1 or 1, the more accurate the prediction based on regression. It indicates how the proportion of how much a variable varies with another as a fraction total variation, thus quantifying correlation.  Also the square root of coefficient of determination, though not directly connected, just mathematically linked. Does not necessarily indicate causation.  **Properties of correlation:**   * Only measures strength of *linear* relationship between variables * r does not change when we change units of measurement of x, y or both * makes no difference between explanatory or response variable – only about variability of one with the other * only applies to quantitative variables |

## **SLR in practice: assumptions, checks & remedies**

|  |  |  |  |
| --- | --- | --- | --- |
| Cautions with least squares regression | | | |
| Extrapolation  Do not extrapolate – model is learnt based on data within range, and predictions should thus only be made from interpolation  Presence of lurking variables  Take note of **lurking variables** – variables not among explanatory or response variables that may cause us to get a misrepresentative set of data (confounders).  Beware of correlations based on averaged data  Risk of ecological fallacy: a correlation based on averages over many individuals is usually higher than the correlation between the same variables based on data on individuals. Ecological fallacy occurs when this aggregate data is used to draw conclusions about individual level phenomena.  The restricted range problem  When data suffers from restricted range (not the full picture), r and r2 tend to be lower than if the full range could be observed, causing an underestimation.  Model assumptions  Check that assumptions of model are valid | | | |
| Assumption | Check | Causes & Implications | Remedies |
| Constant variance | Residual plots of regressor against fitted values are not ideal. | Incorrect functional form/model form, changes in data collection causing changes, presence of outliers.  Estimators will remain unbiased, but standard errors of estimators are biased, causing inflation/deflation of test statistics and rendering hypothesis test unreliable. | Adjusting model by including seasonal terms, data transformations on Y etc. |
| Independent distribution | Residual plots against time are not ideal. | Incorrect functional form, omitted variables, seasonal effects, measurement errors.  Estimators will remain unbiased, but standard errors of estimators are biased small, causing inflation of test statistics and rendering hypothesis test unreliable, higher risk of type I error. | Consider missing regressors or adjusting for seasonal effects, finding missing values or performing data transformations. |
| Independent distribution from x | Residual plot vs regressor x is not ideal. | Could be due to omitted variables/other factors or measurement errors.  Bias in estimators of coefficients, and bias in estimator for σ2. | Review selection of variables and factors in model. |
| Normally distributed | % of standardised residuals between 1.96 and -1.96, histogram of residuals. Note size of sample in this – larger samples are less likely to have an issue. | Incorrect functional form or model.  Estimators are unbiased, but not normally distributed. This renders hypothesis tests useless, though large samples should be fine due to CLT. | Transformation of variables to normal distribution. |
| Linear functional form | Residual plot against fitted values, correlation and coefficient of determination values, scatter plot. | Could be due to inclusion of irrelevant variables, exclusion of important ones, non-linear shape, etc.  This causes autocorrelation (dependence on time), non-constant variance. (even if they are in reality, a wrong model causes these to appear). | Review variables to be included in model. |
| Outliers and influential observations in regression | | **Outliers:** an observation that lies outside the overall pattern of other observations- outliers in the y direction often have large regression residuals, but those in x may not.  **Influence:** an observation (data point) is influential for a statistical calculation if removing it would markedly change the result of the calculation. Points that are outliers in the **x direction** are often influential for the least squars regression line.  To check, calculate values before and after removal of data points and compare. | |
| Applications of SLR | | * Investigating statistical significance of relationship between variables * Often used for prediction within range of data used to establish relationship * Used to substantiate, but never to prove causation | |
| Multi-variable regression | | Multi-variable regression – same principles. Least square estimators using partial derivatives, hypotheses testing (but now in plural), so conclusions can now be partially supported (eg. Only one of two variables have statistically significant relationship). n dimensional plot. | |

# **Relationships in categorical variables**

|  |  |  |
| --- | --- | --- |
| Describing relationships with categorical variables | | |
| The two-way table: summarising counts of observations for each combination of categorical variables    Joint distribution: a two-way table of **proportions relative to the total**    Marginal distribution: examining the **proportions of total of one variable relative to total** – for a two-way table, there are two marginal distributions, one for each categorical variable    Conditional distributions: looking at the ‘marginal distribution’ relative to total in one other variable (rather than full total). | | |
| Simpson’s paradox | | |
| An association or comparison that holds for all of several groups (at the group level) reversing when the data are combined to form a single group.    Reason: arises due to disproportionate numbers of the groups – causing skewed numbers. In this example, there are 1500 patients from A in poor condition, relative to 200 from B, thus skewing the numbers in favour of B.  Implications: be wary of conclusions drawn from aggregate data – look at individual conditional distributions. | | |
| Establishing causation | | |
| Explaining association: causation | Explaining association: common response | Explaining association: confounding |
| Direct causal mechanism present, no other variables that can explain phenomenon | No causal mechanism between two variables – both can be explained by their relationship to a single other common response. | Two variables are *confounded* when their effects on a response variable cannot be distinguished from each other – because z relates to both x *and* y.    Both the explanatory x and lurking z may influence the response variable y. Because x is confounded with z, we cannot distinguish the influence of x from the influence of z – and thus cannot conclude anything about the direct effect of x on y. |
| To establish causal links between categorical variables, the best method is thus to use carefully designed experiments to control the effects of possible lurking variables. | | |