Quantitative Statistics

Dr Scott Oatley

Table of Contents

Quantitative Statistics 1

Introduction to Statistical Analysis 3

Randomisation and sampling 3

Data and Distributions 5

Probability 7

Hypothesis and P-value 18

Contingency tables and Chi2 25

Requirements 27

Correlation 30

T-test 30

ANOVA 31

Assumptions for statistical inference 34

Association and causality 35

Regression Diagnostics 37

Introduction to Statistical Analysis

A statistical analysis is commonly classified into two distinct typologies. The first as descriptive, and the second inferential. The analysis depends on whether its main purpose is to describe the data or make predictions respectively.

The values a variable can take form the measurement scale of that variable. For a quantitative/metric/continuous/scale variable, the possible numerical values are said to form an interval scale, because they have a numerical distance or interval between each pair of levels. For categorical variables that are unordered, the scale does not have a ''high'' or ''low'' end. In this instance the categories are said to form a nominal scale. A third type of scale falls, in a sense, between nominal and interval. It consists of categorical scales having a natural ordering of values. The categories form an ordinal scale.

A variable is said to be discrete if its possible values form a set of separate numbers, such as (0,1,2,3…,n) it is continuous if it can take an infinite continuum of possible real number values.

Randomisation and sampling

A simple random sample of *n* subjects from a population is one in which each possible sample of that size has the same probability of being selected. To select a random sample, we need a list of all subjects in the population. This list is called the sampling frame. To select a random sampling frame, you need to (1) number the subjects in the sampling frame, (2) generate a set of these numbers randomly, and (3) sample the subjects whose numbers were generated.

Sampling variability and potential bias

For conclusions based on statistical inference to be worthwhile, we should know the potential sampling error. The sampling error of a statistic is the error that occurs when we use a statistic based on a sample to predict the value of a population parameter.

Random sampling protects against bias, in the sense that the sampling error tends to fluctuate about 0, sometimes being positive, other times being negative. For sample sizes of about 1000, we'll see that the sampling error for estimating percentages is usually no greater than plus or minus 3%. This bound is called the margin of error.

Sampling Bias: Nonprobability Sampling

For simple random sampling, each possible sample *n* subjects have the same probability of selection. This is a type of probability sampling method, meaning that the probability any sample will be selected is known. Inferential statistical methods assume probability sampling.

The most common nonprobability sampling is volunteer sampling.

The sampling bias inherent in volunteer sampling is also called selection bias. It is problematic to evaluate policies and programs when individuals can choose whether to participate in them.

Even within random sampling, sampling bias can occur. One case is when the sampling frame suffers from under coverage- lacking representation.

Response bias is another example of bias that is often the result of confusing questions.

Nonresponse bias: missing data

Some subjects who are selected for the sample may refuse to participate, or it may not be possible to reach them. This results in what is known as nonresponse bias. This adds to a problem called missing data which can be difficult to solve.

Systemic random sampling

Denote the sample size by *n* and the population size by *N*. Let K=*N/n,* the population size be divided by the sample size. A systematic random sample (1) selects a subject at random from the first k names in the sampling frame, and (2) selects every kth subject after that one. The number k is called the skip number.

Stratified Random Sampling

A stratified random sample divides the population into desperate groups, called strata, and then selects a simple random sample from each stratum. Stratified

random sampling is called proportional if the sampled strata proportions are the same as those in the entire population.

Cluster Sampling

Divide the population into many clusters, such as city blocks. Select a simple random sample of the clusters. Use the subjects in those clusters as the sample.

Multistage Sampling

When conducting a survey for predicting elections, the Gallop organisations often identify election districts as clusters and take a simple random sample of them. But then it also takes a simple random sample of households within each selected selection district. This is more feasible than sampling every household in the chosen districts. This is an example of multistage sample, which uses combinations of sampling methods.

Data and Distributions

The center of the data is typically understood as a typical observation. The variability of the data is understood to be the spread around the center of the data. The mean describes the center, and the standard deviation describes the variability. The correlation describes the strength of the association, and regression analysis predicts the value of one variable from a value of the other variable.

Population Distribution

The sample state distribution is a blurry photo of the population distribution. As the sample size increases, the sample proportion in any interval gets closer to the true population proportion. Thus, the sample data distribution looks more like the population distribution.

The shape of the distribution

A U-shaped distribution indicates a polarisation on the variable between two sets of subjects. A bell-shaped distribution indicates that most subjects tend to fall near a central value. Often one tail is much longer than another. A distribution is said to be skewed to the right or left according to which tail is longer.

Notation for observations, mean, and summations.

The sample size is symbolised by *n*. For a variable denoted by y, its operations are denoted as y1,y1…in. The sample mean is denoted by $\bar{y}$.

The sample mean is defined as:

$$\bar{y} = y_1 + y_2 +, \ldots, y_n/n$$

The symbol Σ represents the process of summing. Using the summation symbol, we have a shortened expression for the sample of n observations:

$$\bar{y} = \sum y_i/n$$

Properties of the Mean

The overall sample mean for the combined set of $(n_1 + n_2)$ observations is the weighted average:

$$\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2}$$

The numerator is the sum of all observations, and the denominator is the total sample size.

Properties of the median

For skewed distributions the mean lies toward the longer tail relative to the median. The mean is larger than the median for distributions that are skewed to the right and smaller when the distribution is skewed to the left.

For the mean we need quantitative data, the median also applies for ordinal scales.

Probability

Standard Deviation

The deviation of an observation y_i from the sample mean $\bar{y}$ is $(y_i - \bar{y})$, the difference between them. The standard deviation *s* of *n* observations is:

$$s = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}}$$

Also known as the root of the squared deviations divided by the sample size minus one.

This is the positive square root of the variance s^2, which is:

$$s^2 = \frac{\sum(y_i - \bar{y})^2}{n-1} = \frac{(y_i - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n-1}$$

The expression $\sum(y_i - \bar{y})^2$ in these formulas is called a sum of squares. It represents squaring each deviation and then adding those squares. The larger the deviation, the larger the sum of squares and the larger *s* tends to be.

Interpreting the magnitude of s: the empirical rule

If you produce a histogram of the data and that data is approximately bell shaped, then: about 68 per cent of the observations fall between $\bar{y} - s$ and $\bar{y} + s$, about 96 per cent of the observations fall between $\bar{y} - 2_s \; and \; \bar{y} + 2_s$, and All or nearly all observations fall between $\bar{y} - 3_s \; and \; \bar{y} + 3_s$.

Quartiles and other percentiles

The *pth* percentile is the point such that p% of the observations fall below or at that point and (100-p)% fall above it. The 25% percentile is called the lower quartile. The 75% percentile is called the upper quartile. One quarter of the data fall below the lower quartile and one quarter falls above the upper quartile.

The difference between the upper and lower quartile ranges is called the interquartile range. An observation is classified as an outlier if it falls more than 1.5(IQR) above the upper quartile or more than 1.5(IQR) below the lower quartile.

The z-score

Another way to measure position by the number of standard deviations that a value falls from the mean. This is known as the z-score. The number of standard deviations from the mean equals the z-score: z=observation – mean/standard deviation.

By the empirical rule, for a bell-shaped distribution it is very unusual for an observation to gall more than three standard deviations from the mean.

The z symbol in a normal table refers to the distance between a possible value y of a variables and the mean μ of its probability distribution, in terms of the number of standard deviations that y falls from μ

For a probability distribution with a mean μ and standard deviation σ, it equals:

$$z = \frac{Variable\ value - Mean}{Standard\ deviation} = \frac{y - \mu}{\sigma}$$

Using z-scores to find probabilities or y-values

If we have a value y and need to find a probability, convert y to a z-score using z=(y-μ)/σ, and then convert z to the probability of interest using a table of normal probabilities. Is we have a probability and need to find a value of y, convert the probability to a tail probability and find the z-score and then evaluate y=μ+zσ

Association between response and explanatory variables

An analysis of association between two variables is called a bivariate analysis because there are two variables. Usually, one is an outcome variable on which comparisons are made at levels of the other variable. The outcome variable is called the response variable. The variable that defines the groups is called the explanatory variable.

Sample Statistics and population parameters

We call μ and σ the population mean and population standard deviation, respectively. These parameters are constants.

Probability as a long run relative frequency

With a random sample or randomised experiment, the probability that an observation has a particular outcome is the proportion of times that outcome would occur in a very long sequence of like observations- probability.

The Basic Rules of Probability

Let P(A) denote the probability of a possible outcome denoted by the letter A.

$$P(not\ A) = 1 - P(A)$$

If you know the probability a particular outcome occurs, the probability it does not occur is 1 minus that probability. If A and B are distinct possible outcomes (with no overlap), then

$$P(A\ or\ B) = P(A) + P(B)$$

If A and B are possible outcomes, then

$$P(A\ and\ B) = P(A) * P(B\ given\ A)$$

The probability P(B given A) is called a conditional probability and is often denoted by

$$P(B\ I\ A)$$

If A and B are independent, then

$$P(A\ and\ B) = P(A) * P(B)$$

Probability distributions for discrete variables

The probability distribution of a discrete variable assigns a probability to each possible value of the variable. Each probability is a number between 0 and 1. The some of the probabilities of all possible values equals 1. Let P(y) denote the probability of a possible outcome for a variable y. Then:

$$0 \leq P(y) \leq 1\ and \sum_{all\ y} P(y) = 1$$

Probability distributions for continuous variables

Mean of a probability distribution (expected value). The mean of the probability distribution for a discrete variable y is:

$$\mu = \sum yP(y)$$

The sum is taken over all possible values of the variable. This parameter is also called the expected value of y and denoted by E(y).

The standard deviation of a probability distribution, denoted by σ, measures its variability. The more spread out the distribution, the larger the value of σ. If a probability distribution is bell shaped, about 68% of the probability falls between μ-σ and μ+σ, about 95% falls between μ-2σ and μ+2σ, and all or nearly all falls between μ-3σ and μ+3σ. The standard deviation is the square root of the variance of the probability distribution. The variance measures the average squared deviation of an observation from the mean. That is, it is the expected value of (y-μ)squared. In the discrete case, the formula is:

$$\sigma^2 = E(y-\mu)^2 = \sum (y-\mu)^2 P(y)$$

The Normal Probability Distribution

The normal distribution is symmetric, bell shaped, and characterised but its mean μ and standard deviation σ. The probability within any number of standard deviations of μ is the same for all normal distributions. This probability equals 0.68 within 1 standard deviation, 0.95 within 2 standard deviations, and 0.997 within 3 standard deviations

The standard normal distribution

The standard normal distribution is the normal distribution with mean μ=0 and standard deviation σ=1. If a variable has a normal distribution, and if its values are converted to z-scores by subtracting the mean and dividing by the standard deviation, then the z-scores have the standard normal distribution

Bivariate probability distributions: covariance and correlation

Each variable in a bivariate distribution has a mean and a standard deviation. Denote them by (μ_x, σ_x) for x and by (μ_y, σ_y) for y. The way that x and y vary together is described by their covariance, which is defined to be:

$$Covaraince(x,y) = E[(x-\mu_x)(y-\mu_y)]$$

Which represents the average of the cross products about the population means (weighted by their probabilities). If y tends to fall above its mean when x falls above its mean, the covariance is positive. If y tends to fall below its mean when x falls above it mean the covariance is negative.

The covariance can be any real number. For interpretation, it is simpler to use:

$$Correlation(x,y) = \frac{covariance(x,y)}{(\sigma_x)(\sigma_y)}$$

This equals:

$$\frac{E[(x-\mu_x)(y-\mu_y)]}{\sigma_x\sigma_y} = E\left[\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right] = E(z_x z_y),$$

Sampling distribution

A sampling distribution of a statistic is the probability distribution that specifies probabilities for the possible values the statistic can take.

The standard deviation of the sampling distribution of $\bar{y}$ is called the standard error of $\bar{y}$ and is denoted by $\sigma_{\bar{y}}$. For a random sample of size n, the standard error of $\bar{y}$ depends on n and the population standard deivation σ by:

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

For sampling a population, the sampling distribution of $\bar{y}$ states the probabilities for the possible values of $\bar{y}$. For a random sample of size n from a population having mean μ and standard error:

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

The effect of sample size on sampling distribution and precision of estimates

The standard error gets smaller as the sample size n gets larger. The reason for this is that the denominator ($\sqrt{n}$) of the standard error formula increases as n increases. In summary, error occurs when we estimate μ *by* $\bar{\bar{y}}$, because we sampled only part of the population. This error, which is the sampling error, tends to decrease as the sample size n increases. The standard error is fundamental to inferential procedures that predict the sampling error in using $\bar{\bar{y}}$ *to estimate* μ.

Central limit theorem

For random sampling with a larger sample size n, the sampling distribution of the same mean $\bar{\bar{y}}$ is approximately a normal distribution. This is what we call the central limit theorem.

Point and Interval Estimation

A point estimate is a single number that is the best guess for the parameter value. An interval estimate is an interval of numbers around the point estimate that we believe contains the parameter value. This interval is also called the confidence interval.

An estimator is unbiased if its sampling distribution centers around the parameter. Specifically, the parameter is the mean of the sampling distribution. By contrast, a biased estimator tends to underestimate the parameter, on the average, or it tends to overestimate the parameter. An estimator that has a standard error that is smaller than those of other estimators is said to be efficient.

Confidence Interval formed by point estimate margin of error.

A confidence interval for a parameter is an interval of numbers within which the parameter is believed to fall. The probability that this method produces an interval that contains the parameter is called the confidence level. This is a number chosen to be close to 1, such as 0.95 or 0.99. With probability about 0.95, the estimator falls within two standard errors. To construct a confidence interval, we add and substrate from the point estimate a z-score multiple of its standard error. This is the margin of error:

$$Form\ of\ confidence\ interval = Point\ estimate\ \pm Margin\ of\ Error$$

The sample proportion and its standard error

Let π denote a population proportion. Then π falls between 0 and 1. Its point estimator is the sample proportion. We denote the sample proportion by $\hat{\pi}$, since it estimates π. The population propotion π is the mean μ of the probability distirbution having probabilities:

$$P(1) = \pi \; and \; P(0) = 1 - \pi$$

The standard deviation of this probability distribution is $\sigma = \sqrt{\pi(1-\pi)}$. Since the formula for the standard error of a sample mean is $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$ the standard error of the sample proportion is:

$$\sigma_{\hat{\pi}} = \frac{\sigma}{\sqrt{\pi}} = \sqrt{\frac{\pi(1-\pi)}{n}}$$

As the sample size increases, the standard error gets smaller. The sample proportion then tends to fall closer to the population proportion.

Confidence Interval for a Proportion

For large random samples, the sampling distribution $\hat{\pi}$ is approximately normal about the parameter π it estimates. The interval $\hat{\pi} \pm 1.96\sigma_{\pi}$ is an interval estimate for π with confidence level 0.95. It is called a 95% confidence interval.

Calculating a Confidence Interval

First, we calculate the standard error of the sample proportion:

$$se = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

Then we can calculate the 95% confidence interval for π using:

$$\hat{\pi} \pm 1.96(se)$$

Controlling the confidence level by choice of z-score

With a confidence level of 0.95, that is 95% confidence, there is a 0.05 probability that the method produces a confidence interval that does not contain the parameter value. In some applications a 5% chance of an incorrect inference is unacceptable. To increase the change of a correct inference, we use a larger confidence level, such as 0.99. The general form for the confidence interval for a population proportion $\boldsymbol{\pi}$ is:

$$\hat{\pi} \pm z(se), with\ se = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$$

Larger sample sizes give narrower intervals.

We can estimate a population proportion $\boldsymbol{\pi}$ more precisely with a larger sample size. The margin of error is z(se). The larger the value of n, the smaller the margin of error and the narrower the interval. The width of the confidence interval increases as the confidence level increases and decreases as the sample size increases.

The Error Probability =1- confidence interval. The probability that an interval estimation method yields a confidence interval that does not contain the parameter is called the error probability. The Greek letter alpha denotes the error probability. For an error probability of a=0.005, the confidence level equals 1-a=0.95.

In practice the probability that the confidence interval contains π is approximately equal to the chosen confidence level. As n increases, the sampling distribution of $\hat{\pi}$ is more closely normal in form, by the Central Limit Theorem. This is what allows us to use z-scores from the normal distirbution in findings the margin of error.

The t distribution

To achieve generality, it has the disadvantage of assuming that the population distribution is normal. Suppose we knew the exact standard error of the sample mean. Then in addition with the assumption that the population is normal, for an n the appropriate confidence interval formula is:

$$\bar{\bar{y}} \pm z\sigma_{\bar{\bar{y}}}, which\ is\ \bar{\bar{y}} \pm z\sigma/\sqrt{n}$$

The t-score is like a z-score, but it comes from a bell-shaped distribution that is slightly more spread out than the standard distribution. This distribution is called the t-distribution.

Properties of the t-distribution

The t distribution is bell shaped and symmetric with a mean of 0. The standard deviation is a bit larger than 1. The precise value depends on what is called the degrees of freedom, denoted by df. The t-distribution has a slightly different spread for each distinct value of df, and different t-scores apply for each df value. For inference about a population mean the degrees of freedom equal df= n-1, one less than the sample size. The t distribution has thicker tails and is more spread out than the standard normal distribution. The larger the df value, however, the more closely it resembles the standard normal. When df is about 30 or more the two distributions are nearly identical. A t-score multiplied by the estimated standard error gives the merging of error for a confidence interval for the mean.

For a random sample from a normal population distribution, a 95% confidence interval for μ is:

$$\bar{\bar{y}} \pm t_{.025}(se), where\ se = s/\sqrt{n}$$

$$and\ df = n - 1\ for\ the\ t - score$$

Like the confidence interval for a proportion, this confidence interval has a margin of error that is a score multiplied by the estimated standard error.

Like the width of the confidence interval for a proportion, the width of a confidence interval for a mean also depends on the sample size n. Larger sample sizes result in narrower intervals

Robustness for violations of normal population assumption

The assumptions for the confidence interval for a mean are (1) randomisation for collecting the sample and (2) normal population distribution.

A statistical method is said to be robust with respect to an assumption if it performs adequately even when that assumption is violated. Statisticians have shown that the confidence interval for a mean using the t distribution is robust against violations of

the normal population assumption. Even if the population is not normal, confidence intervals based on the t distribution still work quite well, especially when n exceeds about 15. An important case when the method does not work well is when the data are extremely skewed or contain extreme outliers. The t confidence interval method is not robust to violations of the randomisation assumption.

Choice of Sample Size

The margin of error depends directly onto the standard error of the sampling distribution of the point estimator. The standard error itself depends on the sample size.

To determine the sample size, we must decide on the margin of error. We must also specify the probability with which the margin of error is achieved. This probability is the confidence level for the confidence interval.

The random sample size n having margin of error M in estimating π by the sample proportion $\hat{\pi}$ is:

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

The z score is the one for the chosen confidence level, such as z=1.96 for level 0.95. you need to guess π or take the sage approach of setting π=0.50.

Determining the sample size for estimating means

The random sample size n having margin of error M in estimating μ by the sample mean $\bar{y}$ is:

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

The z score is the one for the chosen confidence level, such as z=1.96 for level 0.95. you need to guess the population standard deviation σ.

Maximum likelihood method of estimation

Fisher proposed the maximum likelihood estimate. This estimate is the value of the parameter that is most consistent with the observed data, in the following sense: if the parameter equaled that number (the value of the estimate), the observed data would have had greater change of occurring than if the parameter wailed any other number

Fisher showed that, for large samples, maximum likelihood estimators have three desirable properties. They are efficient: other editors do not have smaller standard errors. They are consistent, in the sense that as n increases, they tend to get closer and closer to the unknown parameter value. They have little, if any, bias, with the bias diminishing to 0 as n increases. They have approximately normal sampling distributions.

Fisher found the maximum likelihood estimator is the sample mean, and that is preferred over the sample median.

To use the maximum likelihood, we need to assume about the shape of the population distribution. But sometimes we do not have enough information to make a sensible assumption. For such cases, a bootstrap is useful. This method treads the sample distribution as if it were the true population distribution and approximates by simulation the unknown sampling distribution. To do this, the method samples n observations, with replacement, from the sample distribution. That is, each of the original n data points has probability 1/n of selection for each "new" observation. This new sample of size n has its own point estimate of the parameter. The bootstrap method repeats this sampling process many times, for instance, selecting 1000 separate samples of size n and 1000 corresponding point estimate values.

Hypothesis and P-value

In statistics a hypothesis is a statement about a population. It takes the form of a predation that a parameter takes a numerical value or falls in a certain range of values. A statistical significance test uses data to summarize the evidence about a hypothesis. It does this by comparing point estimates of parameters to the values predicted by the hypothesis.

The null hypothesis, denoted by the symbol H0, is a statement that the parameter takes a value. The alternative hypothesis, denoted by Ha, states that the parameter

falls in some alternative range of values. Usually the value in H0 corresponds, in a certain sense, to no effect. The values in Ha then represent an effect of some type.

The p-value is the probability that the test statistic equals the observed value or a value even more extreme in the direction predicted by Ha. It is calculated by presuming that H0 is true. The p-value is denoted by p.

Significance test for a mean

The null hypothesis about a population mean μ has the form H0: μ=μ0. The alternative hypothesis contains alternative parameter values from the value in H0. The most common alternative hypothesis is Ha: μ=/= μ0, such as Ha: μ=/=0. This alternative hypothesis is called a two-side test, because it contains values both below and above the value listed in H0. Under the presumption that H0: μ=μ0 is true, the center of the sampling distribution of $\bar{y}$ is the value μ0. A value of $\bar{y}$ that falls far out in the tail provides strong evidence against H0, because it would be unusual if truly μ=μ0. We calculate the p-value under the presumption that H0 is true. The smaller p is, the stronger the evidence against H0 and in favour of Ha.

Correspondence between two-sided tests and confidence intervals

Conclusions using two-sided significate tests are consistent with conclusions using confidence intervals. If a test says that a value is believable for the parameter, then so does a confidence interval. Whenever the P>0.05 in a two-sided test about a mean μ, a 95%-confidence interval for μ necessarily contains the H0 value for μ. Whenever p<=0.05 in a two-sided test about a mean μ, a 95% confidence interval for μ does not contain the H0 for μ.

We can use a different hypothesis when a researcher predicts a deviation from H0 in a direction- also known as a one-sided significance test. It has one of the forms: Ha: μ>μ0 or Ha: μ0<μ0.

In practice two sided tests are more common than one sided test. Two sided tests can also detect an effect that falls in the opposite direction. In using two-sided P values, researchers avoid the suspicion that they chose Ha when they saw the direction in which the data occurred. That is not ethical.

Sometimes we need to decide whether the evidence against H0 is strong enough to reject it. We based the decision on whether the P-value falls below a prespecified cut

off point. The α-level is a number such that we reject H0 if the P-value is less than or equal to it. The α-level is also called the significance level. In practice, the most common α-levels are 0.05 and 0.01.

It is better to say "do not reject H0" than "accept H0" the population proportion has many plausible values besides the number in H0. When the P-value is larger than the α-level, saying "do not reject H0" instead of "accept H0" emphasis that that value is merely one of many believable values.

A given difference between an estimate and the H0 value has a smaller P-value as the sample size increases. The larger the sample size, the more certain we can be that sample deviations from H0 are indicative of true population deviations. Notice that even a small departure between $\hat{\pi}$and π_0 can yield a small P-value if the sample size is very large.

Type One and Type Two Errors

When H0 is true a Type 1 error occurs if H0 is rejected. When H0 is false, a type 2 error occurs if H0 is not rejected. With α=0.05 if H0 is true, the probability equals 0.05 of making a type 1 error and rejecting H0.

When we make α smaller in a significance test, we need a smaller P-value to reject H0. It then becomes harder to reject H0. But this means that it will also be harder even if H0 is false. The Stronger the evidence required to convict someone, the more likely we will fail to connive defendants who are guilty. The smaller P(Type 1 error) is, the larger P(Type 2 error) is.

Statistical versus practical significance

It is important to distinguish between statistical and practical significance. A small P-value, such as P=0.001, is highly statistically significant. It provides strong evidence against H0. It does not, however, imply an important finding in any practical sense.

Always inspect the different between the estimate and the H0 value to gauge the practical implications of a test result. Although significance tests can be useful, most statisticians believe they are overemphasised in social science research. It is preferable to construct confidence intervals for parameters instead of performing only significance tests. A test merely indicates whether the value in H0 is plausible.

It does not tell us which other potential values are plausible. The confidence interval, by contrast displays the entire set of believable values.

It is misleading to report results only if they are statistically significant. Some tests may be statistically significant just by chance. It is incorrect to interpret the P-value as the probability that H0 is true. True effects are often smaller than reported estimates.

Finding P(Type Two Error)

When H0 is false, a type 2 error results from not rejecting it. This probability has more than one value, because Ha contains a range of possible values. Each value in Ha has its own P(Type 2 error). P(Type 2 error) decreases as: The parameter value is farther from H0, The sample size increases, P(Type 1 error) increases.

The binomial distribution

For categorical data, often the following three conditions hold: Each observation falls into one of two categories, The probabilities for the two categories are the same for each observation. We denote the probabilities by $\boldsymbol{\pi}$ for category 1 and (1-$\boldsymbol{\pi}$) for category 2, The outcomes of successive observations are independent. That is, the outcome for one observation does not depend on the outcomes of other observations. Denote the probability of category 1, for each observation, by $\boldsymbol{\pi}$. For an independent observation, the probability that x of the n observations occurs in category 1 is:

$$P(x) = \frac{n!}{x!\,(n-x)!}\pi^{x}(1-\pi)^{n-x}, x = 0,1,2,\dots,n$$

The symbol n! Is called n factorial. It represents n!= 1*2*3*...*n for example, 1!=1, 2!=1*2=2, 3!=1*2*3=6. Also 0! Is defined to be 1.

The binominal distribution is perfectly symmetric only when $\boldsymbol{\pi}$=0.50. The sample proportion $\hat{\pi}$relates to the binominal variable x by: $\hat{\pi}$=x/n. The binominal distribution for x= how many of n observations fall in a category having probability $\boldsymbol{\pi}$ has mean and standard deviation:

$$\mu = n\pi \; and \; \sigma = \sqrt{n\pi(1-\pi)}$$

Standard Error of estimated difference between groups

To compare two populations, we estimate the difference between their parameters. To compare populations means μ1 and μ2, we treat μ2-μ1 as a parameter and estimate it by the difference of sample means, ȳ2-ȳ1. The sampling distribution of the estimator ȳ2-ȳ1 has expected value μ2-μ1. for larger random samples, by the Central Limit Theorem this sampling distribution has a normal shape. For two estimates from independent samples that have estimated standard errors se1 and se2, the sampling distribution of their difference has:

$$\sqrt{(se_1)^2 + (se_2)^2}$$

Confidence Interval for difference of proportions

For large, independent random samples, a confidence interval for the difference $\boldsymbol{\pi}2$-$\boldsymbol{\pi}1$ between two population proportions is:

$$(\widehat{\pi_2} - \hat{\pi}_1) \pm z(se), wher\ se = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

The z-score depends on the confidence interval, such as 1.96 for 95% confidence.

Contingency tables and conditional probabilities

Each row is a category of the explanatory variable which defines the two groups compared. Each column is a category of the response variable. The cells of the table contain frequencies for the four possible combinations of outcomes (*prtest*).

Confidence interval for μ2-μ1

For independent random samples from wo groups that have normal population distributions, a confidence interval for μ2-μ1 is:

$$(\bar{\bar{y}}_2 - \bar{\bar{y}}_1) \pm t(se), where\ se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The t-score is chosen to prove the desired confidence interval.

Interpreting a confidence comparing means

A confidence interval μ2-μ1 that contains only negative values suggests that μ2-μ1 is negative, meaning that we can infer that μ2 is less than μ1. When the confidence interval contains 0, insufficient evidence exists to conclude which of μ1 or μ2 is larger. It is then plausible that μ1=μ2.

To compare population means μ1 and μ2, we can also conduct a significance test of H0:μ1=μ2. For the difference of means parameter, this hypothesis is H0: μ2-μ1=0(no effect).

Dependent samples occur when each observation in sample 1 matches with an observation in sample 2. The data are often called matched pairs data, because of this matching. With matched-pairs data, for each pair we form: Difference= observation in sample 2- observation in sample 1.

Using dependent samples can have certain benefits. First sources of potential bias are controlled. Using the same subjects in each sample, for instance, keeps other factors fixed that could affect the analysis. Second the standard error of ȳ2-ȳ1 may be smaller with dependent samples.

When n12+n21 exceeds about 20, this statistic has approximately a standard normal distribution when H0 is true. This test is often referred to as McNemar's test. For smaller samples, use the binomial distribution to conduct the test.

A confidence interval for the difference of proportions is more informative than a significance test. For large samples, this is:

$$(\hat{\pi}_2 - \hat{\pi}_1) \pm z(se)$$

Fisher's exact test for comparing proportions.

For comparing proportions with independent samples are valid for relatively large samples. For small sample sizes, the sampling distribution of $\hat{\pi}_2 - \hat{\pi}_1$ may not be close to normality. You can then compare two proportions π_1 and π_2 using a method called Fishers exact test. The P-value for the two-sided alternative equals 0.0075. This is not double the one-sided P-value because, except in certain special cases, the sampling distribution (called the hyper-geometric distribution) is not symmetric. The test is called exact because it uses the actual (hypergeometric) sampling distribution rather than a normal approximation.

Nonparametric statistics for comparing groups.

A body of methods exists that makes no assumption about the shape of the population distribution. These methods are called nonparametric. Nonparametric methods still apply when the normality assumption for methods using the t distribution is badly violated. They are primarily useful for small samples, especially for on seeded tests, as parametric methods may then be invalid when the normal population assumption is badly violated.

Wilcoxon-Mann-Whitney Test

Most nonparametric comparisons of groups also assume identical shapes for the population distributions, but the shapes are not required to be normal. The model for the test is then: H0: Both y1 and y2 have the same distribution, Ha: the distributions for y1 and y2 have the same shape, but the one for y1 is shifted up or shifted down compared to the one for y2. Here, Ha is two sided. One sided Ha is also possible. The most popular test of this type is called the Wilcoxon test. This test is an ordinal-level method, in the sense that it uses only the rankings of the observations. Another nonparametric test is the Mann-Whitney test. It views all the pairs of observations, such that one observation is from one group and the other observations is from the other group. The test statistic is based on the number of pairs for which the observation from the first group was higher.

Effect size: proportion of better Reponses for a group

An effect size measure, $(\bar{y}1\text{-}\bar{y}2)/s$, for summarising the size of the difference between two groups. When the distributions are very skewed or have outliers, the means are less useful, and this effect size summary may be inappropriate. A nonparametric

effect size measure is the proportion of pairs of observations (one from each group) for which the observation from the first group was higher.

Treating ordinal variables as quantitative

Social scientists often uses parametrical statistical methods for quantitative data with variables that are only ordinal. They do this by assigning scores to the ordered categories. When the choice is unclear, such as with categories for happiness, it is a good idea to perform a sensitive study. Choose two or three reasonable sets of potential scores, such as (0.5.10), (0,6,10), (0,7,10) and check whether the ultimate conclusions are similar for each.

Contingency tables and Chi2

Data for the analysis of categorical variables are displayed in contingency tables. Two categorical variables are statistically independent if the population conditional distributions on one of them are identical at each category of the other. The variables are statistically dependent if the conditional distributions are not identical. The Chi-Square test of independence determines whether there is an association between categorical variables. It is a non-parametric test. The Chi-square test of indpeendence can only compare categorical variables. Additionally, the Chi-aqure test of indepdnence only assesses associations ebtween categorical variables, and cannot provide inferences about causation.

The definition of statistical independence refers to the population. Two variables are independent if the population conditional distributions on the response variable are identical. We address this with a significance test, by testing: H0: the variables are statistically independent, Ha: the variables are statistically dependent. Let f_o denote an observed frequency in a cell of the table. Let f_e denote an expected frequency. This is the count expected in a cell if the variables were independent. It quals the product of the row and column totals for that cell, divided by the total sample size.

The Pearson statistic for testing independence

The test statistic for H0: indolence summarises how close the expected frequencies fall to the observed frequencies. Symbolised by X squared:

$$X^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

With summation over all cells in the contingency table.

The probability distribution of cell counts in conditional distributions or marginal distributions is then the multi-nominal distribution, which generalises the binominal distribution from two categories to several categories. For large sample sizes under randomisation, the sampling distribution of Squared is called the chi-squared probability distribution. The symbol X squared for the chi-squared distribution is the Greek analogy of the symbol X squared for the test statistic. That statistic is often called the Pearson chi-squared statistic. It is concentrated on the posited part of the real line. The X squared test statistic cannot be negative, since it sums squared differences divided by positive expected frequencies. The minimum possible value, X squared=0, would occur if f_o=$f_e$in each cell, it is skewed to the right. The precise shape of the distribution depends on the degrees of freedom (df). The mean μ=df and the standard deviation σ=$\sqrt{2df}$. Thus, the distribution tends to shift to the right and becomes more spread out for larger df values. In addition, as df increases, the skew lessens, and the chi-squared curve becomes more bell shaped. For testing H0: independent with a table having r rows and c columns, Df= (r-1)(c-1). Larger numbers of rows and columns produce larger df values. The larger the X squared value for a df, the stronger the evidence against H0:independence. The P-value equals the right tail probability above the observed X squared value. It measures the probability, presuming H0 is true, that X squared is at least as large as the observed value.

Chi-squared and difference of proportions for 2*2 tables

Let $\pi_1$represent the proportion of successes in population π_1 and let $\pi_2$represent the proportion of successes in population 2. Then (1-π_1) and (1-π_2) are the proportions of failures. If the response variable is statistically indent of the populations considered, then π_1=π_2. The null hypothesis of independence corresponds to the homogeneity hypothesis, H0: π_1=π_2. in fact, the chi-squared test of independence is equivalent to a test for equality of two population proportions. The chi-squared statistic relates to

the z statistic by Squared=z squared. The chi-squared right tail probability is the same as the two-tail standard normal probability for z. For instance, z=1.96 is the z score with a two-tail probability of 0.05. The square of this, (1.96)squared= 3.84, is the chi-squared score for df=1 with a right tail probability of 0.05.

Chi-squared needed for larger tables than 2*2

An advantage of the z test is that it also applies with one sided alternative hypothesis, such as Ha: $\pi_1 > \pi_2$. The direction of the effect is lost in squaring z and using X^2. When a table is larger than 2*2 and this df >1, we need more than one difference parameter to describe the association. We could use a z statistic for each comparison, but not a single z statistic for the overall test of independent.

Requirements

The chi-squared test, like one and two sample z tests for proportions, is a large sample test. For 2*2 contingency tables, a small sample test of independence is Fishers exact test.

In the chi-squared test, the value of X squared does not depend on which is the response variable, and which is the explanator variable. When a response variable is identified and the population conditional distributions are identical, they are said to be homogenous. The chi-squared test of independence is then often referred to as a test of homogeneity.

The chi-squared test tells us nothing about the nature or strength of association.

Residual Analysis

The difference (f_o-f_e) between an observed and an expected cell frequency is called a residual. A residual is positive when the observed frequency fo exceeds the value fe that independence predicts. The residual is negative when the observed frequency is smaller than independence predicts. The standardised residual for a cell is:

$$z = \frac{f_o - f_e}{se} = \frac{f_o - f_e}{\sqrt{f_e(1 - row\ proportion)(1 - column\ proportion)}}$$

Here, se denotes the standard error of $f_o - f_e$, presuming H0 is true. The standardised residual is the number of standard errors that $(f_o - f_e)$ falls from the value of 0 that we expect when H0 is true.

A large value for X squared in the chi-squared test of independence suggests that the variables are associated. It does not imply, however, that the variables have a strong association.

The Odds Ratio

The difference of proportions is easily interpretable. Several other measures are also being taken. The most important one for categorical data is the odds ratio. For a binary response variable, the odds of success are defined to be Odds= probability of success/ probability of failure. The estimated odds for a binary response equal the number of successes divided by the number of failures. The odds ratio is a measure of association for 2*2 contingency tables that equals the odds in row 1 divided by the odds in row 2.

The odds ratio takes the same value regardless of the choice of response variable. The odds ratio θ equals the ratio of the products of cell counts from diagonally opposite cells because of this property, the odds ratio is also called the cross-product ratio. The odds ratio can equal any nonnegative number. When the success probabilities are identical in the two rows of a 2*2 table then θ=1. When θ>1 the odds of success are higher in row 1 than in row 2. When θ<1 the odds of success are lower in row 1 than in row 2. Values of θ farther from 1.0 in each direction represent stronger associations. Two values for θ represent the same strength of association, but in opposite directions, when one value is the reciprocal of the other.

For contingency tables with more than two rows or more than two columns, the odds ration describes patterns in any 2*2 suitable.

A pair of observations is concordant if the subject who is higher on one variable also is higher on the other variable. A pair of observations is discordant if the subject who is higher on one variable is lower on the other.

Gamma

A positive difference for C-D occurs when C>D. This indicates a positive association. A negative difference for C-D reflects a negative association. Larger sample sizes

have larger numbers of pairs with, typically, larger absolute differences in C-D. Therefore, we standardize this difference to make it easier to interpret. To do this we device C-D by the total number of pairs that are either concordant or discordant, C+D. This gives the measure of association called gamma.

$$\hat{Y} = \frac{C - D}{C + D}$$

Properties of gamma: The value of gamma falls between -1 and +1, the sign of gamma indicates whether the association is positive or negative, the larger the absolute value of gamma, the stronger the association. A table for which gamma equals 0.6 or -0.6 exhibits a stronger association than one for which gamma equals 0.3 or -0.3 for example.

Common properties of ordinal measures

Gamma is one of several ordinal measures of association. Others are Kendalls tau-b, spearman's rho-b, and somers' d. Ordinal measures of association take values between -1 and +1. The sign tells us whether the association is positive or negative. If the variables are statistically independent then the population values of ordinal measures of association equal 0. The stronger the association, the larger the absolute value of the measure. Values of 1.0 and -1.0 represent the strongest associations- except for Somers' d, the ordinal measures of association names above do not distinguish between response and explanatory variables.

Confidence intervals help us gauge the strength of the association in the population. Let y denote the population value of gamma. For sample gamma, $\hat{y}$, its sampling distribution is approximately normal about y. Its standard error se describes the variation in $\hat{y}$ values around y among samples of the given size. The formula for se is complicated but it is reported by most software. Assuming random sampling, a confidence interval for y has the form:

$$\hat{y} \pm z(se)$$

Ordinal tests versus Pearson chi-squared test

A test of independence based on an ordinal measure is usually preferred to the chi-squared test when both variables are ordinal. The X^2 statistic ignores the ordering of the categories, taking the same value no matter how the levels are ordered. If a positive or negative trend exists, ordinal measures are usually more powerful for detecting it. Unfortunately, the situation is not clear cut. It is possible for the chi-squared test to be more powerful even if the data are ordinal. The null hypothesis of independence is not equivalent to a value of 0 for population gamma. although independence implies y=0, the converse is not true. An ordinal measure of association may equal 0 when the variables are statistically dependent, but the dependence does not have an overall positive or overall negative trend. The chi-squared test can perform better than the ordinal test when the relationship does not have a single trend.

Mixed ordinal-nominal contingency tables

When the nominal variable has more than two categories, it is inappropriate to use an ordinal measure such as gamma. There are specialised methods for mixed nominal-ordinal tables, but it is usually simplest to treat the ordinal variable as quantitative by assigning scores to its levels.

Correlation

Correlation is a method for examining the magnitude of any assoication between two metric (continuous) variables. Prior to conducting a strict correlation test statistic, it is always beneficial to run a scatterplot graph to visually determine if any obvious relationship between two metric variables exist.

T-test

The T-test is a method for examining whether two populations (groups like men and women) vary on a metric variable (like income). We use this method if we have a binary categorical variable and a metric variable, or perhaps where we would want to know if there was a difference in score between two categories of a larger categorical variable. Error bar charts are often used to visualize the confidence of this association using confidence intervals. Confidence intervals are the range of population parameters for which the observed statistic is a plausible consequence.

ANOVA

ANOVA is a method we can use to assess the association between a metric variable (like income) and a categorical variable with two or more categories (like educational attainment) – where a variable has only two categories, we use a t-test. The one-way ANOVA (analysis of variance) compares the means of two or more independent groups to determine whether there is statistical evidence that the associated population means are significantly different. The measure of association reported with ANOVA is Eta and this varies between 0 and 1.

Linear functions and Regressions:

The formula y=a+Bx expresses observations on y as a linear function of observations on x. The formula has a straight-line graph with slope B(beta) and y-intercept a (alpha).

An observation is called influential if removing it results in a larger change in the prediction equation. Unless the sample size is large, an observation can have a strong influence on the slope if its x-value is low or high compared to the rest of the data and if it is a regression outlier

For an observation, the difference between an observed value and the predicted value of the response variable, y-ŷ, is called the residual

We summarise the size of the residuals by the sum of their squared values. This quantity, denotes by SSE is:

$$SSE = \sum (y - \hat{y})^2$$

The least squares estimate a and b are the values that provide the perdition equation ŷ=a+bx for which the residual sum of squares, is a minimum.

For the linear regression model, each value of x corresponds to a single value of y. Such a model is said to be deterministic. It is unrealistic in social science research because we do not expect all subjects who have the same x-value to have the same y-

value. Instead, the y-values vary. A probabilistic model for the relationship allows for variability in y at each vale of x.

Let E(y) denote the mean of a conditional distribution of y. The symbol E represents expected value.

$$E(y) = a + \beta x$$

A regression function is a mathematical function that describes how the mean of the response variable changes according to the value of an explanatory variable.

The linear regression model has an addition parameter σ describing the standard deviation of each conditional distribution. That is, σ measures the variability of the y-values for all subjects having the same x-value. We refer to σ as the conditional standard deviation. The most common assumption is that the conditional distribution of y is normal at each fixed value of x, with unknown standard deviation σ.

The sum of squares in the numerator of sy is called the total sum of squares.

The correlation between variables x and y, denoted by r, is:

$$r = \frac{\sum(x - \bar{\bar{x}})(y - \bar{\bar{y}})}{\sqrt{[\sum(x - \bar{\bar{x}})^2][\sum(y - \bar{\bar{x}})^2}}$$

The correlation is a standardised version of the slope. The standardisation adjusts the slope b for the fact that the standard deviations of x and y depend on their units of measurement. The correlation relates to the slope b of the prediction equation by:

$$r = \left(\frac{s_x}{s_y}\right) b$$

Because of the relationship between r and b, the correlation is also called the standardised regression coefficient for hot model E(y)=a+Bx

The correlation is valid only when a straight-line model is sensible for the relationship between x and y. Since r is proportional to the slope of a linear prediction equation, it measures the strength of the linear association: -1<=r<=1. the correlation, unlike the slope b, must fall between -1 and +1. R has the same sign as the slope b. This holds because their formulas have the same numerator, relating to the covariation of x and y, and positive denominators. Thus r>0 when the variables are positively related, and r<0 when the variables are negatively related. R=0 for those lines having b=0. When r=0, there is not a linear increasing or linear decreasing trend in the relationship. R= ±1 when all the sample points fall exactly on the prediction line. These correspond to perfect positive and negative linear associations. There is then no prediction error when we use $\hat{Y}$=a+bx to predict y. The larger the absolute value or r, the stronger the linear association. Variables with a correlation of -0.8 are more strongly linearly associated than variables with a correlation of 0.4. The correlation, unlike the slope b, treats x and y symmetrically. The predication equation using y to predict x has the same correlation as the one using x to predict y. The value of r does not depend on the variable's units.

The correlation implies regression toward the mean. We predict that y is closer to the mean, in standard deviation units. This is called regression toward the mean. The larger the absolute value of r, the stronger the association, in the sense that a standard deviation change in x corresponds to a greater proportion of a standard deviation change in y

A related measure of association summarise how well x can predict y. If we can predict y much better by substituting x-values into the prediction equation than without knowing the x-values, the variables are judged to be strongly associated. Rule1: (predicting y without using x) the best predictor is the sample mean. Rule2: (predicting y using x) when the relationship between x and y is linear, the prediction equation provides the best predictor of y. Prediction errors: the prediction error for each subject is the difference between the observed and predicted values of y. The prediction error using rule 1 is y-yflathat, and the prediction error using rule 2 is y-$\hat{Y}$. R2 is denoted by:

$$r^2 = \frac{E_1 - E_2}{E_1} = \frac{TSS - SSE}{TSS} = \frac{\sum(y - \hat{y})^2 - \sum(y - \hat{y})^2}{\sum(y - \hat{y})^2}$$

It is called r-squared, or sometimes the coefficient of determination. The notation squared is used for this measure because, in fact, the proportional reduction in error equals the square of the correlation r.

Since -1<=r<=1, r-squared falls between 0 and 1. The minimum possible value for SSE is 0, in which case r-squared=TSS/TSS=1. For SSE=0, all sample points must fall exactly on the prediction line. In that case, there is no error using x to predict y with the prediction equation. This condition corresponds to r=+-1. When the least squares slope b=0, the y intercept a equals $\bar{y}$. Then $\hat{y}= \bar{y}$ for all x. The two prediction rules are then identical, so SSE=TSS and r-squared=0. Like the correlation, r-squared measures the strength of linear association. The closer r-squared is to 1, the stronger the linear association. R-squared does not depend on the units of measurement, and it takes the same value when x predicts y as when y predicts x.

The sums of squares describe the conditional of marginal variability. A one standard deviation change in x corresponds to a predicted change of r standard deviations in y. The square of the correlation has a proportional reduction in error interpretation related to predicting y using $\hat{y}$=a+bx rather than $\hat{y}$. The total sum of squares summarises the variability of the observations on y, since this quantity divided by n-1 is the sample variance s2y of the y-values. Similarly, SSE summarises the variability around the prediction equation, which refers to variability for the conditional distributions.

Assumptions for statistical inference

Randomisation, such as a simple random sample in a survey. The mean of y is related to x by the linear equation E(y)=a+Bx. The conditional standard deviation is identical at each x-value. The conditional distribution of y at each value of x is normal.

A small P-value for H0:B=0 suggests that the regression line as a nonzero slope. We should be more concerned with the size of the slope than in knowing merely that it is not 0. A confidence interval for B has the formula: B ±t(se). Constructing a

confidence interval for the correlation is more complicated than for the slope B. The reason is that the sampling distribution of r is not symmetric expect when p=0.

In summary B=0 does not correspond to independence if the assumption of a linear regression model is violated. For this reason, you should always construct a scatterplot to check this fundamental assumption.

Observations

A disadvantage of least squares is that individual observations can unduly influence the results. When scatterplot shows a sever regression outlier, you should investigate the reasons for it. Observations that have a large influence on the model parameter estimates can also have a large impact on the correlation. Besides being influenced by outliers, the correlation depends on the rand of x-values sampled. When a sample has a much narrower range of variation in x than the population, the sample correlation tends to underestimate drastically the population correlation.

It is dangerous to apply a predication equality ion to values of x outside the range of observed values. The relationship might be far from linear outside that range.

Regression with error terms

An alternative formulation for the model expresses each observation on y, rather than the mean E(y) of the values, in terms of x. To allow variability, we include a term for the deviation of y from the mean, a+Bx. Each observation has its own error value ε. The ε term is called the error term, since it represents the error that results from using the mean value (a+Bx) of y at a certain value of x to predict the individual observation. We can interpret ε as a population residual. Thus, ε is the difference between the observation y and the mean a+Bx of all possible observations on y at the value of x.

Association and causality

A relationship must satisfy three criteria to be considered a causal one: Association between variables, an appropriate time order, the elimination of alternative explanations.

Causality requires an appropriate time order. For instance, race, age, and gender exist prior to current attitudes or achievements, so any causal association must treat

them as causes rather than effects. It is difficult to study cause and effect when two variables do not have a time order but are measured together over time.

Many alternative explanative involve an additional variable z or a set of variables. For example, there Amy be a variable z that causes both x and y. With observational studies we can never prove that one variable is a cause of another. We can disprove causal hypotheses, however, by showing that empirical evidence contradicts at least one of these criteria.

A randomised experiment is the ideal way to compare two groups. The outcome for a subject is observed after the group is assigned, so the time order is certain. Because of these factors it is easier to assess causality with randomised experiments than with observational studies.

A variable is said to be controlled when its influence is removed. To study the association between two categorical variables, while controlling for a third variable, we form contingency tables relating those variables separately for subjects at each level of that control variable. The separate tables that display the relationships within the fixed levels of the control variable are called partial tables.

A potential pitfall of almost all social science is the possibility that the study did not include an important variable. If you fail to control for a variable that strongly influences the association between the variables of primary interest, you will obtain misleading results. A variable that is not measured in a study but that influences the association under study is called a lurking variable.

An association between y and x1 is said to be spurious if both variables are dependent on a third variable x2 and their association disappears when x2 is controlled.

A chain causation in which x1 affects x2 which in turn affects y. Here x1 is an indirect rather than direct cause of y. Variable x2 is called an intervening or mediator variable.

Occasionally two variables show no association until a third variable is controlled. That control variable is called a suppressor variable.

When two explanatory variables both have effects on a response variable but are also associated with each other, there is said to be confounding. If our study neglects to

observe a confounding variable that explains a major part of that effect, our results and conclusions will be biased. Such bias is called omitted variable bias.

Backward Elimination

Backward elimination begins by placing all the explanatorily variables under consideration in the model. It deletes one at a time until reaching a point where the remaining variables all make significant partial contributions to predicting y. The variable deleted at each stage is the one that is the least significant, having the larger P-value in the significance test for tis effect.

Whereas backward elimination begins with all the potential explanatory variables in the mode, forward selection begins with none of them. Stepwise regression is a modification of forward selection that drops variables from the model if they lose their significance as other variables are added.

Indices for selecting a model: adjusted R-squared, press, and AIC

Recall that maximizing R-squared is not a sensible criterion because the most complicated model will have the largest R-squared value. In comparing predictive power of different models, it is more helpful to use adjusted R-squared:

$$R^2_{adj} = \frac{s_y^2 - s^2}{s_y^2} = 1 - \frac{s^2}{s_y^2}$$

Regression Diagnostics

Inference about parameters in regression model have these assumptions: The true function has the form used in the model (linear), the conditional distribution of y is normal, the conditional distribution of y has constant standard deviation throughout the range of values of the explanatory variables (This condition is called homoscedasticity), the sample is randomly selected.

Several checks of assumptions use the residuals y-ŷ. One check concerned the normality assumption. A standardised version of the residual equals the residual divided by its standard error, which describes how much residuals vary because of ordinary sampling variability. In regression, this is called a studentized residual. A

histogram of these residuals should have the appearance of a standard normal distributing (bell curve).

If the model assumes a linear effect but the effect is strongly nonlinear, some conclusions may be faulty. For multiple regression, it is also useful to construct a scatterplot of each explanatory variable against the response variable. For multiple regression models, plots of the residuals against the predicted values or against each explanatory variable also help is check for potential problems. In practice residual patterns are rarely as neat. Don't let a few outliers influence too strongly your interpretation of a plot.

VIF: multicollinearity causes variance inflation. Multicollinearity causes inflated standard errors for estimates of regression parameters. The standard error of the estimator of the coefficient Bj of xj in the multiple regression model can be expressed as:

$$se = \frac{1}{\sqrt{1 - R_j^2}} \left[\frac{s}{\sqrt{n - 1 s_{x_j}}} \right]$$

Where s is the square root of the residual mean square and six denotes the sample standard deviation of xj values. Let R-squared denote R-squared from the regression of xj on the other explanatory variables from the model. So, when xj overlaps a lot with the other explanatory variables, in the sense that R-squared is large for predicting xj using the other explanatory variables, this se is relatively large.

A warning sign occurs when the estimated coefficient for a predictor already in the model changes substantially when another variable is introduced. Perhaps the estimated coefficient of x1 is 2.4 for the bivariate model, but when x2 is added to the model, the coefficient of x1 changes to 25.9. When multicollinearity exists, it is rather artificial to interpret a regression coefficient as the effect of an explanatory variable when other variables are held constant.

Two approaches are common to deal with nonlinearity: Polynomial regression, Generalized linear model with a link function such as the logarithm. The data for the model fit consists of the y-values for the subjects in the sample, the x-values (called x1), and an artificial variable (x) consisting of the squares of the x-values.