Crash Course on Basic Statistics

Marina Wahl, marina.w4hl@gmail.com
University of New York at Stony Brook

November 6, 2013

Contents

Chapter 1

Basic Probability

1.1 Basic Definitions

Trials

- ⋆ Probability is concerned with the outcome of **trials**.
- ⋆ **Trials** are also called *experiments* or *observations* (multiple trials).
- ⋆ **Trials** refers to an event whose outcome is unknown.

Sample Space (S)

- ⋆ Set of **all possible elementary outcomes** of a trial.
- ⋆ If the trial consists of flipping a coin twice, the sample space is $S = (h, h), (h, t), (t, h), (t, t)$.
- ⋆ The probability of the sample space is **always 1**.

Events (E)

- ⋆ An *event* is the **specification** of the outcome of a trial.
- ⋆ An *event* can consist of a **single** outcome or a **set** of outcomes.
- ⋆ The **complement** of an *event* is everything in the sample space that is not that event (not E or $\sim$ E).
- ⋆ The probability of an *event* is always **between 0 and 1**.
- ⋆ The probability of an *event* and its **complement** is always 1.

Several Events

- ⋆ The **union** of several simple events creates a compound event that **occurs if one or more of the events occur**.
- ⋆ The **intersection** of two or more simple events creates a compound event that occurs **only if all the simple events occurs**.
- ⋆ If events cannot occur together, they are **mutually exclusive**.
- ⋆ If two trials are **independent**, the outcome of one trial does not influence the outcome of another.

Permutations

- ⋆ **Permutations** are all the possible ways elements in a set can be arranged, where the **order is important**.
- ⋆ The number of permutations of subsets of size k drawn from a set of size n is given by:

$$nPk = \frac{n!}{(n-k)!}$$

Combinations

- ⋆ **Combinations** are similar to permutations with the difference that the **order of elements is not significant**.
- ⋆ The number of combinations of subsets of size k drawn from a set of size n is given by:

$$nPk = \frac{n!}{k!(n-k)!}$$

1.2 Probability of Events

- ⋆ If two events are **independents**, $P(E|F) = P(E)$. The probability of both E and F occurring is:

$$P(E \cap F) = P(E) \times P(F)$$

- ⋆ If two events are **mutually exclusive**, the probability of either E or F:

$$P(E \cup F) = P(E) + P(F)$$

- ⋆ If the events are **not mutually exclusive** (you need to correct the 'overlap'):

$$P(E \cup F) = P(E) + P(F) - P(E \cap F),$$

where

$$P(E \cap F) = P(E) \times P(F|E)$$

1.3 Bayes' Theorem

Bayes' theorem for any two events:

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

- ⋆ **Frequentist**:
  - – There are true, fixed parameters in a model (though they may be unknown at times).
  - – Data contain random errors which have a certain probability distribution (Gaussian for example).
  - – Mathematical routines analyse the probability of getting certain data, given a particular model.
- ⋆ **Bayesian**:
  - – There are no true model parameters. Instead all parameters are treated as random variables with probability distributions.
  - – Random errors in data have no probability distribution, but rather the model parameters are random with their own distributions.
  - – Mathematical routines analyze probability of a model, given some data. The statistician makes a guess (prior distribution) and then updates that guess with the data.

Chapter 2

Basic Definitions

2.1 Types of Data

There two types of measurements:

- **Quantitative**: **Discrete** data have finite values. **Continuous** data have an infinite number of steps.
- **Categorical (nominal)**: the possible responses consist of a set of categories rather than numbers that measure an amount of something on a continuous scale.

2.2 Errors

- **Random error**: due to chance, with no particular pattern and it is assumed to cancel itself out over repeated measurements.
- **Systematic errors**: has an observable pattern, and it is not due to chance, so its causes can be often identified.

2.3 Reliability

How consistent or repeatable measurements are:

- **Multiple-occasions reliability (test-retest, temporal)**: how similarly a test perform over repeated administration.
- **Multiple-forms reliability (parallel-forms)**: how similarly different versions of a test perform in measuring the same entity.
- **Internal consistency reliability**: how well the items that make up instrument (a test) reflect the same construct.

2.4 Validity

How well a test or rating scale measures what is supposed to measure:

- **Content validity**: how well the process of measurement reflects the important content of the domain of interests.
- **Concurrent validity**: how well inferences drawn from a measurement can be used to predict some other behaviour that is measured at approximately same time.
- **Predictive validity**: the ability to draw inferences about some event in the future.

2.5 Probability Distributions

- Statistical inference relies on making assumptions about the way data is distributed, transforming data to make it fit some known distribution better.
- A **theoretical probability distribution** is defined by a formula that specifies what values can be taken by data points within the distribution and how common each value (or range) will be.

2.6 Population and Samples

- We rarely have access to the entire population of users. Instead we rely on a subset of the population to use as a proxy for the population.
- **Sample statistics** estimate **unknown population parameters**.
- Ideally you should select your **sample randomly** from the parent population, but in practice this can be very **difficult** due to:
  - issues establishing a truly random selection scheme,
  - problems getting the selected users to participate.
- Representativeness is more important than randomness.

Nonprobability Sampling

- Subject to sampling bias. Conclusions are of limited usefulness in generalizing to a larger population:
  - **Volunteer** samples.
  - **Convenience samples**: collect information in the early stages of a study.
  - **Quota sampling**: the data collector is instructed to get response from a certain number of subjects within classifications.

Probability Sampling

- Every member of the population has a know probability to be selected for the sample.
- The simplest type is a **simple random sampling** (SRS).
- **Systematic sampling**: need a list of your population and you decide the size of the sample and then compute the number n, which dictates how you will select the sample:
  - Calculate n by dividing the size of the population by the number of subjects you want in the sample.
  - Useful when the population **accrues over time** and there is **no predetermined list** of population members.
  - One caution: making sure data is not cyclic.
- **Stratified sample**: the population of interest is divided into non overlapping groups or *strata* based on common characteristics.
- **Cluster sample**: population is sampled by using pre-existing groups. It can be combined with the technique of sampling proportional to size.

2.7 Bias

- Sample needs to be a good representation of the study population.
- If the sample is biased, it is not representative of the study population, conclusions draw from the study sample might not apply to the study population.
- A statistic used to estimate a parameter is **unbiased** if the expected value of its sampling distribution is equal to the value of the parameter being estimated.
- Bias is a source of systematic error and enter studies in two primary ways:
  - During the **selection and retention** of the subjects of study.
  - In the way **information is collected** about the subjects.

Sample Selection Bias

- **Selection bias**: if some potential subjects are more likely than others to be selected for the study sample. The sample is selected in a way that systematically excludes part of the population.

- ⋆ **Volunteer bias**: the fact that people who volunteer to be in the studies are usually not representative of the population as a whole.

- ⋆ **Nonresponse bias**: the other side of volunteer bias. Just as people who volunteer to take part in a study are likely to differ systematically from those who do not, so people who decline to participate in a study when invited to do so very likely differ from those who consent to participate.

- ⋆ **Informative censoring**: can create bias in any longitudinal study (a study in which subjects are followed over a period of time). Losing subjects during a long-term study is common, but the real problem comes when subjects do not drop out at random, but for reasons related to the study's purpose.

Information Bias

- ⋆ **Interviewer bias**: when bias is introduced intro the data collected because of the attitudes or behaviour of the interviewer.

- ⋆ **Recall bias**: the fact that people with a life experience such as suffering from a serious disease or injury are more likely to remember events that they believe are related to that experience.

- ⋆ **Detection bias**: the fact that certain characteristics may be more likely to be detected or reported in some people than in others.

- ⋆ **Social desirability bias**: caused by people's desire to present themselves in a favorable light.

2.8 Questions on Samples

Representative Sampling

- ⋆ How was the sample selected?

- ⋆ Was it truly randomly selected?

- ⋆ Were there any biases in the selection process?

Bias

- ⋆ Response Bias: how were the questions worded and the response collected?

- ⋆ Concious Bias: are arguments presented in a disinterested, objective fashion?

- ⋆ Missing data and refusals: how is missing data treated in the analysis? How is attrition (loss of subjects after a study begins) handled?

Sample Size

- ⋆ Were the sample sizes selected large enough for a null hypothesis to be rejected?

- ⋆ Were the sample sizes so large that almost any null hypothesis would be rejected?

- ⋆ Was the sample size selected on the basis of a power calculation?

2.9 Central Tendency

Mean

- ⋆ Good if data set that is roughly symmetrical:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

- ⋆ **Outliers**: data error or they belong to other population.

Median

- ⋆ Middle value when the values are ranked in ascending or descending order.

- ⋆ When data is not symmetrical, mean can be heavily influenced by outliers, and median provides a better idea o most typical value.

- ⋆ For odd samples, the median is the central value $(n+1)/2$th. For even samples, it's the average of the two central values, $[n/2+(n+1)/2]/2$th.

- ⋆ In a **small sample** of data (less than 25 or so), the sample median tends to do a poor job of estimating the population median.
- ⋆ For **task-time data** the **geometric mean** tends to provide a better estimate of the population's middle value than the sample median.

Mode

- ⋆ The most frequently occurring value.
- ⋆ Is useful in describing ordinal or categorical data.

Dispersion

- ⋆ **Range**: simplest measure of dispersion, which is the difference between the highest and lowest values.
- ⋆ **Interquartile range**: less influenced by extreme values.
- ⋆ **Variance**:
  - – The most common way to do measure dispersion for continuous data.
  - – Provides an estimate of the average difference of each value from the mean.
  - – For a population:

    $$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

  - – For a sample:

    $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- ⋆ **Standard deviation**:
  - – For a population:

    $$\sigma = \sqrt{\sigma^2},$$

  - – For a sample:

    $$s = \sqrt{s^2}.$$

Chapter 3

The Normal Distribution

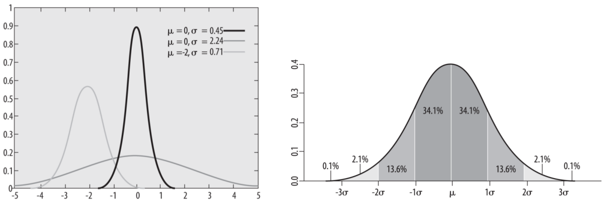


Figure 3.1: (left) All normal distributions have the same shape but differ to their μ and σ: they are shifted by μ and stretched by σ. (right) Percent of data failing into specified ranges of the normal distribution.

- ⋆ All **normal distributions** are **symmetric**, **unimodal** (a single most common value), and have a continuous range from negative infinity to positive infinity.
- ⋆ The total area under the curve adds up to 1, or 100%.
- ⋆ **Empirical rule**: For the population that follows a normal distribution, almost all the values will fall within three standard deviations above and bellow the mean (99.7%). Two standard deviations are about 95%. One standard deviation is 68%.

The Central Limit Theorem

- ⋆ As the sample size approaches infinity, the **distribution of sample means** will follow a normal distribution regardless of what the parent population's distribution (usually 30 or larger).
- ⋆ The **mean of this distribution of sample means** will also be equal to the mean of the parent population. If $X_1, X_2, ..., X_n$ all have mean μ and variance σ^2, sampling distribution of $\bar{X} = \frac{\sum X_i}{n}$:

  $$E(\bar{X}) = \mu, Var(\bar{X}, \frac{\sigma^2}{n}),$$

  using the central limit theorem, for large n, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.
- ⋆ If we are interested on p, a proportion or the probability of an event with 2 outcomes. We use the estimator $\hat{p}$, proportion of times we see the event in the data. The sampling distribution of $\hat{p}$ (unbiased estimator of p) has expected value p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$. By the central limit theorem, for large n the sampling distribution is $\hat{p} \sim N(p, \frac{p(1-p)}{n})$.

The Z-Statistic

- ⋆ A z-score is the distance of a data point from the mean, expressed in units of standard deviation. The z-score for a value from a population is:

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

- ⋆ Conversion to z-scores place distinct populations on the same metric.

- ⋆ We are interested in the probability of a particular sample mean. We can use the normal distribution even if we do not know the distribution of the population from which the sample was drawn, by calculating the **z-statistics**:

  $$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

  .

- ⋆ The **standard error of the mean** $\frac{\sigma}{\sqrt{n}}$ is the standard deviation of the sampling distribution of the sample mean:

  - It describes the **mean of multiple members of a population**.
  - It is **always smaller** than the standard deviation.
  - The larger our sample size is, the smaller is the standard error and less we would expect our sample mean to differ from the population mean.

- ⋆ If we know the mean but not the standard deviation, we can calculate the **t-statistic instead** (following sessions).

Chapter 4

The Binomial Distribution

- ⋆ An example of discrete distribution
- ⋆ Events in a **binomial distribution** are generated by a **Bernoulli** process. A single trial within a Bernoulli process is called a *Bernoulli trial*.
- ⋆ Data meets four requirements:
  - The outcome of each trial is one of two mutually exclusive outcomes.
  - Each trial is independent.
  - The probability of success, p, is constant for every trial.
  - There is a fixed number of trials, denoted as n.
- ⋆ The probability of a particular number of successes on a a particular number of trials is:

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

  where
  $$\binom{n}{k} = nCk = \frac{n!}{k!(n-k)!}$$
- ⋆ If both np and $n(1-p)$ are grater than 5, the binomial distribution can be approximated by the normal distribution.

Chapter 5

Confidence Intervals

Point Estimate

- ⋆ **Point estimate** is a single statistic, such as the mean, to describe a sample.
- ⋆ If we drew a different sample, the mean calculated from that sample would probably be different.
- ⋆ **Interval estimates** state how much a point estimate is likely to vary by chance.
- ⋆ One common interval estimate is the **confidence level**, calculated as $(1 - \alpha)$. where α is the **significance**.

Confidence Interval

- ⋆ The confidence interval is the range of values that we believe will have a specified chance of containing the unknown population parameter.
- ⋆ The confidence interval is a range of values around the mean for which if we drew an infinite number of samples of the same size from the same population, x% of the time the true population mean would be included in the confidence interval calculated from the samples. It gives us the information about the precision of a point estimate such as the sample mean.
- ⋆ CI will tell you the most likely range of the unknown population mean or proportion.
- ⋆ The confidence is in the method, not in any interval.
- ⋆ Any value inside the interval could be said to be a **plausible** value.
- ⋆ CI are twice the **margin of error** and provide both a measure of location and precision.
- ⋆ Three things affect the width of a confidence interval:
  - – **Confidence Level**: usually 95%.
  - – **Variability**: if there is more variation in a population, each sample taken will fluctuate more and wider the confidence interval. The variability of the population is estimated using the standard deviation from the sample.
  - – **Sample size**: without lowering the confidence level, the sample size can control the width of a confidence interval, having an inverse square root relationship to it.

CI for Completion Rate (Binary Data)

- ⋆ **Completion rate** is one of the most fundamental **usability metrics**, defining whether a user can complete a task, usually as a binary response.

- ⋆ The first method to estimate binary success rates is given by the **Wald interval**:

$$\hat{p} \pm z_{(1-\frac{\alpha}{2})}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

where $\hat{p}$ is the sample proportion, n is the sample size, $z_{(1-\frac{\alpha}{2})}$ is the critical value from the *normal distribution* for the level of confidence.

- ⋆ Very inaccurate for small sample sizes (less than 100) or for proportions close to 0 or 1.

- ⋆ For 95% ($z = 1.96$) confidence intervals, we can add two successes and two failures to the observed number of successes and failures:

$$\hat{p}_{\text{adj}} = \frac{x + \frac{z^2}{2}}{n + z^2} = \frac{x + \frac{1.96^2}{2}}{n + 1.96^2} \sim \frac{x+2}{n+4}, \quad (5.0.1)$$

where x is the number that successfully completed the task and n the number who attempted the task (sample size).

CI for Task-time Data

- ⋆ Measuring time on task is a good way to assess task performance and tend to be positively skewed (a non-symmetrical distribution, so the mean is not a good measure of the center of the distribution).

- ⋆ In this case, the *median* is a better measure of the center. However, there is two major drawbacks to the median: *variability* and *bias*.

- ⋆ The median does not use all the information available in sample, consequently, the medians of samples from a continuous distribution are more variable than their means.

- ⋆ The increased variability of the median relative to the mean is amplified when the sample sizes are small.

- ⋆ We also want our sample mean to be unbiased, where any sample mean is just as likely to overestimate or underestimate the population mean. The median does not share this property: at small samples, the sample median of completion times tends to consistently overestimated the population median.

- ⋆ For small-sample task-time data the geometric mean estimates the population median better than the sample median. As the sample sizes get larger (above 25) the median tends to be the best estimate of the middle value. We can use binomial distribution to estimate the confidence intervals: the following formula constructs a confidence interval around any percentile. The median (0.5) would be the most common:

$$np \pm z_{(1-\frac{\alpha}{2})}\sqrt{np(1-p)},$$

where n is the sample size, p is the percentile expressed as a proportion, and $\sqrt{np(1-p)}$ is the standard error. The confidence interval around the median is given by the values taken by the th integer values in this formula (in the ordered data set).

Chapter 6

Hypothesis Testing

- $\star$ H_0 is called the *null hypothesis* and H_1 is the *alternative hypothesis*. They are mutually exclusive and exhaustive. A null hypothesis cannot never be proven to be true, it can only be shown to be plausible.
- $\star$ The alternative hypothesis can be *single-tailed*: it must achieve some value to reject the null hypothesis,

  $$H_0 : \mu_1 \leq \mu_2, H_1 : \mu_1 > \mu_2,$$

  or can be *two-tailed*: it must be different from certain value to reject the null hypothesis,

  $$H_0 : \mu_1 = \mu_2, H_1 : \mu_1 \neq \mu_2.$$

- $\star$ **Statistically significant** is the probability that is not due to chance.
- $\star$ If we fail to reject the null hypothesis (find significance), this does not mean that the null hypothesis is true, only that our study did not find sufficient evidence to reject it.
- $\star$ Tests are not reliable if the statement of the hypothesis are suggested by the data: **data snooping**.

p-value

- $\star$ We choose the **probability level** or p-value that defines when sample results will be considered strong enough to support rejection of the null hypothesis.
- $\star$ Expresses the probability that extreme results obtained in an analysis of sample data are due to chance.
- $\star$ A **low p-value** (for example, less than 0.05) means that the null hypothesis is unlikely to be true.
- $\star$ With null hypothesis testing, all it takes is sufficient evidence (instead of definitive proof) that we can see as at least some difference. The size of the difference is given by the confidence interval around the difference.
- $\star$ A small p-value might occur:
  - by chance
  - because of problems related to data collection
  - because of violations of the conditions necessary for testing procedure
  - because H_0 is true
- $\star$ if multiple tests are carried out, some are likely to be significant by chance alone! For $\sigma = 0.05$, we expect that significant results will be 5% of the time.
- $\star$ Be suspicious when you see a few significant results when many tests have been carried out or significant results on a few subgroups of the data.

Errors in Statistics

- ⋆ If we **wrongly say that there is a difference**, we have a **Type I error**. If we **wrongly say there is no difference**, it is called **Type II error**.

- ⋆ Setting $\alpha = 0.05$ means that we accept a 5% probability of Type I error, *i.e.* we have a 5% chance of rejecting the null hypothesis when we should fail to reject it.

- ⋆ Levels of acceptability for Type II errors are usually $\beta = 0.1$, meaning that it has 10% probability of a Type II error, or 10% chance that the null hypothesis will be false but will fail to be rejected in the study.

- ⋆ The reciprocal of Type II error is *power*, defined as $1 - \beta$ and is the probability of rejecting the null hypothesis when you should reject it.

- ⋆ **Power of a test**:

  - The significance level α of a test shows how the testing methods performs if repeated sampling.
  - If H_0 is true and $\alpha = 0.01$, and you carry out a test repetitively, with different samples of the same size, you reject the H_0 (type I) 1 per cent of the time!
  - Choosing α to be very small means that you reject H_0 even if the true value is different form H_0.

- ⋆ The following four main **factors affect power**:

  - α level (higher probability of Type I error increases power).
  - Difference in outcome between populations (greater difference increases power).
  - Variability (reduced variability increases power).
  - Sample size (larger sample size increases power).

- ⋆ How to have a higher power:

  - The power is higher the further the alternative values away from H_0.
  - Higher significance level α gives higher power.
  - Less variability gives higher power.
  - The larger the samples size, the greater is the power.

Chapter 7

The t-Test

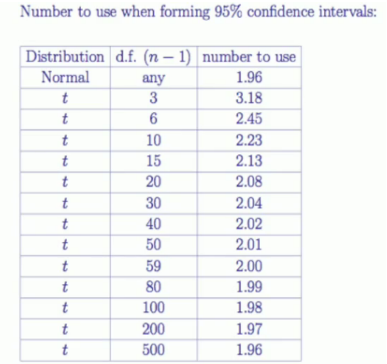
t-Distribution

- ⋆ The **t-distribution** adjusts for how good our estimative is by making the intervals wider as the sample sizes get smaller. It converges to the normal confidence intervals when the sample size increases (more than 30).
- ⋆ Two main reasons for using the t-distribution to test differences in means: when working with small samples from a population that is approximately normal, and when we do not know the standard deviation of a population and need to use the standard deviation of the sample as a substitute for the population deviation.
- ⋆ t-distributions are continuous and symmetrical, and a bit fatter in the tails.
- ⋆ Unlike the normal distribution, the shape of the t distribution depends on the degrees of freedom for a sample.
- ⋆ At smaller sample sizes, sample means fluctuate **more** around the population mean. For example, instead of 95% of values failing with 1.96 standard deviations of the mean, at a sample size of 15, they fall within 2.14 standard deviations.

Confidence Interval for t-Distributions

To construct the interval,

$$\bar{x} \pm t_{(1-\frac{\alpha}{2})} \frac{s}{\sqrt{n}},$$



Number to use when forming 95% confidence intervals:

Distribution	d.f. $(n-1)$	number to use
Normal	any	1.96
t	3	3.18
t	6	2.45
t	10	2.23
t	15	2.13
t	20	2.08
t	30	2.04
t	40	2.02
t	50	2.01
t	59	2.00
t	80	1.99
t	100	1.98
t	200	1.97
t	500	1.96

where $\bar{x}$ is the sample mean, n is the sample size, s is the sample standard deviation, $t_{(1-\frac{\alpha}{2})}$ is the critical value from the t-distribution for $n-1$ *degrees of freedom* and the specified level of confidence, we need:

1. the mean and standard deviation of the mean,
2. the standard error,
3. the sample size,
4. the critical value from the t-distribution for the desired confidence level.

The *standard error* is the estimate of how much the average sample means will fluctuate around the true population mean:

$$se = \frac{s}{\sqrt{n}}.$$

The t-critical value is simply:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}},$$

where we need α (level of significance, usually 0.05, one minus the confidence level) and the degrees of freedom. The degrees of freedom (df) for this type of confidence interval is the sample size minus 1.

For example, a sample size of 12 has the expectation that 95% of sample means will fall within 2.2 standard deviations of the population mean. We can also express this as the margin of error:

$$me = 2.2\frac{s}{\sqrt{n}},$$

and the confidence interval is given as twice the mg.

t-test

Assumptions of t-tests:

- ⋆ The samples are unrelated/independent, otherwise the paired t-test should be used. You can test for linear independence.
- ⋆ t-tests assume that the underlying population variances of the two groups are equal (variances are pooled), so you should test for homogeneity.
- ⋆ Normality of the distributions of both variables are also assumed (unless the samples sizes are large enough to apply the central limit theorem).
- ⋆ Both samples are representative of their parent populations.
- ⋆ t-tests are robust even for small sample sizes, except when there is an extreme skew distribution or outliers.

1-Sample t-Test

One way t-test is used is to compare the mean of a sample to a population with a known mean. The null hypothesis is that there is no significant difference between the mean population from which your sample was drawn and the mean of the know population.

The standard deviation of the sample is:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

The degrees of freedom for the one sample t-test is $n - 1$.

CI for the One-Sample t-Test

The formula to compute a two-tailed confidence interval for the mean for the one-sample t-test:

$$CI_{1-\alpha} = \bar{x} \pm \left(t_{\alpha/2,df}\right)\left(\frac{s}{\sqrt{n}}\right)$$

If you want to calculate a one-sided CI, change the ± sign to either plus or minus and use the upper critical value and α rather than $\alpha/2$.

2-Sample t-Test (Independent Samples)

The t-test for independent samples determines whether the means of the populations from which the samples were drawn are the same. The subjects in the two samples are assumed to be unrelated and independently selected from their populations. We assume that the populations are approximately a normal distribution (or sample large enough to invoke the central limit theorem) and that the populations have approximately equal variance:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

where the pooled variance is

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

CI for the Independent Samples t-Test

$$CI_{1-\alpha} = (\bar{x}_1 - \bar{x}_2) \pm \Big(t_{\alpha/2,df}\Big)\left(\sqrt{s_p^2\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)}\right)$$

Repeated Measures t-Test

Also know as *related samples t-test* or the *dependent samples t-tests*, the samples are not independent. The measurements are considered as pairs so the two samples must be of the same size. The formula is based on the difference scores as calculated from each pairs of samples:

$$t = \frac{\bar{d} - (\mu_1 - \mu_2)}{\frac{s_d}{\sqrt{n}}},$$

where $\bar{d}$ is the mean of the different scores and n the number of pairs. The degrees of freedom is still $n-1$.

The null hypothesis for the repeated measures t-test is usually that the mean of the difference scores, $\bar{d}$, is 0.

CI for the Repeated Measures t-Test

$$CI_{1-\alpha} = \bar{d} \pm \Big(t_{\alpha/2,df}\Big)\Big(\frac{s_d}{\sqrt{n}}\Big)$$

Two Measurements

- $H_0 : \mu = \mu_0$, $H_A : \mu \neq \mu_0$
- observations: $n, \bar{x}, s, d = |\bar{x} - \mu_0|$
- Reject H_0 if

$$\begin{aligned} &P(|\bar{X} - \mu_0| \geq d) \leq \alpha \\ = &P(|\frac{\bar{X} - \mu_0}{\sqrt{s^2/n}}| \geq \frac{d}{\sqrt{s^2/n}}) \\ = &P(|t_{n-1}| \geq \frac{d}{\sqrt{s^2/n}}) \end{aligned}$$

Comparing Samples

The concept of the number of standard errors that the sample means differ from population means applies to both confidence intervals and significance tests.

The two-sample t-test is found by weighting the average of the variances:

$$t = \frac{\hat{x}_1 - \hat{x}_2}{\sqrt{\frac{s_1^2}{n_1 + \frac{s_2^2}{n_2}}}},$$

where $\hat{x}_1$ and $\hat{x}_2$ are the means from sample 1 and 2, s_1 and s_2 the standard deviations from sample 1 and 2, and n_1 and n_2 the sample sizes. The degree of freedom for t-values is approximately 2 less the smaller of the two sample sizes. A *p-value* is just a percentile rank or point in the t-distribution.

The null hypothesis for an independent samples t-test is that the difference between the population means is 0, in which case $\mu_1 - \mu_2)$ can be dropped. The degree of freedom is $n_1 + n_2 - 2$, fewer than the number of cases when both samples are combined.

Between-subjects Comparison (Two-sample t-test)

When a different set of users is tested on each product, there is variation both between users and between designs. Any difference between the means must be tested to see whether it is greater than the variation between the different users. To determine whether there is a significant difference between means of independent samples of users, we use the two-sample t-test:

$$t = \frac{\hat{x}_1 - \hat{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The degrees of freedom for one-sample t-test is $n-1$. For a two-sample t-test, a simple formula is $n_1 + 1n_2 - 2$. However the correct formula is given by *Welch-Satterthwaite procedure.* It provides accurate results even if the variances are unequal:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}$$

CI Around the Difference

There are several ways to report an effect size, but the most compelling and easiest to understand is the confidence interval. The following formula generates a confidence interval around the difference scores to understand the likely range of the true difference between products:

$$(\hat{x}_1 - \hat{x}_2) \pm t_a \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Chapter 8

Regression

Linear Regression

- ⋆ It calculates the equation that will produce the line that is as close as possible to all the data points considered together. This is described as minimizing the square deviations, where the squared deviations are the sum of the squared deviations between each data point and the regression line.

- ⋆ The difference between each point and the regression line is also called the **residual** because it represents the variability in the actual data points not accounted for by the equation generating the line.

- ⋆ Minimizing the squared deviations can be expressed as minimizing the errors of prediction or minimizing the residuals.

- ⋆ The most important assumption of the linear regression is the independence and normality of errors in the independent and dependent variables.

- ⋆ Other assumptions for simple linear regression are: data appropriateness (outcome variable continuous, unbounded), linearity, distribution (the continuous variables are approximately normally distributed and do not have extreme outliers), homoscedasticity (the errors of prediction are constant over the entire data range), independence.

- ⋆ The sum of the squared deviations is the sum of squares of errors, or SSE:

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (ax_i + b))^2$$

- ⋆ In this formula, y_i is an observed data value and $\hat{y}_i$ is the predicted value according to the regression equation.

- ⋆ The variance of s is

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

and the covariance of x and y is

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}.$$

- ⋆ The slope of a simple regression equation is

$$a = \frac{S_{xy}}{S_{xx}},$$

and the intercept of a simple regression equation,

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

Independent and Dependent Variable

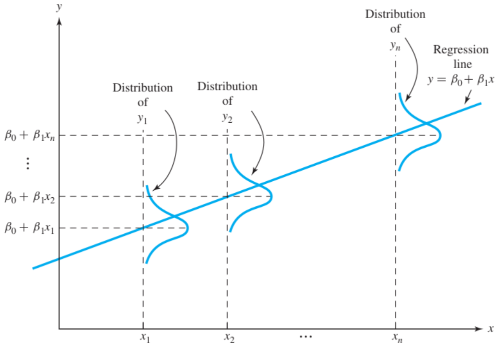
Variables can be either *dependent* if they represent an outcome of the study, or *independent* if they are presumed to influence the value of the dependent variables. There is also a third category, *control variables*, which might influence the dependent variable but are not the main focus of interest.

In a *standard linear model* such as an OLS (ordinary least squares) regression equation, the outcome or dependent variable is indicated by Y, and the independent variables by X:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \epsilon,$$

where ϵ means error and reflects the fact that we do not assume any regression equation will perfectly predict Y, and βs are the regression coefficients.

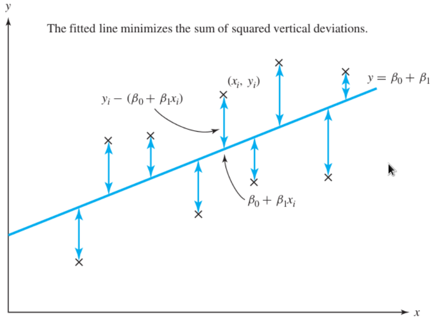
Other terms used for the dependent variable include the *outcome variable*, the *response variable*, and the *explained variable*. Other names for independent variables are *regressors, predictor, variables*, and *explanatory variables.*



Multiple Linear Regression

Multiple linear regression may be used to find the relationship between a single, continuous outcome variable and a set of predictor variables that might be continuous, dichotomous, or categorical; if categorical, the predictors must be recoded into a set of dichotomous dummy variables

Multiple linear regression, in which two or more independent variables (predictors) are related to a single dependent variables is much more common than simple linear regression. Two general principles apply to regression modelling. First, each variable included in the model should carry its own weight, it should explain unique variance in the outcome variable. Second, when you deal with multiple predictors, you have to expect that some of them will be correlated with one another as well as with the dependent variable.


y
The fitted line minimizes the sum of squared vertical deviations.
(x_i, y_i)
$y = \beta_0 + \beta_1$
$y_i - (\beta_0 + \beta_1 x_i)$
$\beta_0 + \beta_1 x_i$
x

$$Y = \beta_0 + \beta_1 X_1 - \beta_2 X_2 + .. + \beta_n X_n + \epsilon,$$

where Y is the dependent variable and β_0 is the intercept. No variable can be a linear combination of other variables.

Chapter 9

Logistic Regression

- We use **logistic regression** when the dependent variable is dichotomous rather than continuous.

- Outcome variables conventionally coded as 0-1, with 0 representing the absence of a characteristic and 1 its presence.

- Outcome variable in linear regression is a **logit**, which is a transformation of the probability of a case having the characteristic in question.

- The **logit** is also called the log odds. If p is the probability of a case having some characteristics, the logit is:

$$logit(p) = \log \frac{p}{1-p} = \log(p) - \log(1-p).$$

- The logistic regression equation with n predictors is:

$$logit(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon.$$

- The differences to multiple linear regression using a categorical outcome are:
  - The assumption of homoscedascitiy (common variance) **is not** met with categorical variables.
  - Multiple linear regression can return values outside the permissible range of 0-1.
  - Assume independence of cases, linearity (there is a linear relationship between the **logit** of the outcome variable and any continuous predictor), no multicollinearity, no complete separation (the value of one variable cannot be predicted by the values of another variable of set variables).

- As with linear regression, we have measures of model fit for the entire equation (evaluating it against the null model with no predictor variables) and tests for each coefficient (evaluating each against the null hypothesis that the coefficient is not significantly different from 0).

- The interpretation of the coefficients is different: instead of interpreting them in terms of linear changes in the outcome, we interpret them in terms of the odds ratios.

Ratio, Proportion, and Rate

Three related metrics:

- A *ratio* express the magnitude of one quantity in relation to the magnitude of another without making further assumptions about the two numbers or having them sharing same units.

- A *proportion* is a particular type of ratio in which all cases in the numerator are included in the denominator. They are often expressed as percents.

- A *rate* is a proportion in which the denominator includes a measure of time.

Prevalence and Incidence

- *Prevalence* describes the number of cases that exist in a population at a particular point in time. It is defined as the proportion of individuals in a population who have the characteristic at a particular point in time.

- *Incidence* requires three elements to be defined: new cases, population at risk, and time interval.

Odds Ratio

- The odds ratio was developed for use in case-control studies.

- The odds ratio is the ratio of the odds of exposure for the case group to the odds of exposure for the control group.

- The odds of an event is another way to express the likelihood of an event, similar to probability. The difference is that although probability is computed by dividing the number of events by the total number of trials, odds are calculated by dividing the number of events by the number of non-events:

$$odds = probability/(1 - probability)$$

or

$$probability = odds(1 + odds)$$

- The odds ratio is simply the ratio of two odds:

$$OR = \frac{odds_1}{odds_2} = \frac{p_1(1-p_1)}{p_2(1-p_2)}$$

Chapter 10

Other Topics

ANOVA

- **Analysis of variance** is a statistical procedure used to compare the mean values on some variable between two or more independent groups.

- It is called analysis of variance because the procedure involves partitioning variance, attributing the variance observed in a data set to different cause of factors, including group membership.

- ANOVA is a useful technique when analysing data from designed experiments.

- Assumptions of ANOVA:

  - Assumes independence, normality, and equality of variances.
  - It is more reliable when the study is balanced (sample sizes are proximately equal).

- The major test statistics for an ANOVA is the **F-ratio**, which can be used to determine whether statistically significant differences exist between groups.

- **One-Way ANOVA**: The simplest form of ANOVA, in which only one variable is used to form the groups to be compared. This variable is called *factor.* A one-way ANOVA with two levels is equivalent to performing a t-test. The null hypothesis in this type of design is usually that the two groups have the same mean.

- **Factorial ANOVA**: ANOVA with more than one grouping variable or factor. We are often interested in the influence of serveral factors, and how they interact. We might be interested in both main effects (of each factor alone) and interactions effects.

Factor Analysis

- Uses standardized variables to reduce data sets by using *principal component analysis* (PCA) (data reduction technique).

- It is based on an orthogonal decomposition of an input matrix to yield an output matrix that consists of a set of orthogonal components (or factors) that maximize the amount of variation in the variables from the input matrix.

- This process produces a smaller, more compact number of output components (produce a set of eigenvectors and eigenvalues).

- The components in the output matrix are linear combinations of the input variables, the components are created so the first component maximizes the variance captured, and each subsequent component captures as much of the residual variance as possible while taking on an uncorrelated direction in space (produce variables that are orthogonal).

Nonparametric Statistics

- ⋆ **Distribution-free statistics**: make few or no assumptions about the underlying distribution of the data.

- ⋆ **Sign Test**: analogue to the one-sample t-test and is used to test **whether a sample has a hypothesized median**.

  - Data values in the sample are classified as above (+) or below (-) the hypothesized median.
  - Under the null hypothesis that the sample is drawn from a population with the specified median, these classifications have a binomial distribution with $\pi = 0.5$ (probability of the population).

    $$Z = \frac{(X \pm 0.5) - np}{\sqrt{np(1-p)}},$$

    where X is the number of observed values greater than median, 0.5 is the continuity correction (negative in this case for a hypothesis $\pi > 0.5$), np is the mean of the binomial distribution (the expected value for X if the null hypothesis is true), and the denominator is the standard deviation of the binomial distribution.

Chapter 11

Some Related Questions

Design an experiment to see whether a new feature is better?

- ⋆ Questions:
  - How often Y occurs?
  - Do X and Y co-vary? (requires measuring X and Y)
  - Does X cause Y? (requires measuring X and Y and measuring X, and accounting for other independent variables)
- ⋆ We have:
  - goals for visitor (sign up, buy).
  - metrics to improve (visit time, revenue, return).
- ⋆ We are manipulating independent variables (independent of what the user does) to measure dependent variables (what the user does).
- ⋆ We can test 2 versions with a A/B test, or many full versions of a single page.
- ⋆ How different web pages perform using a random sample of visitor: define what percentage of the visitor are included in the experiment, chose which object to test.

Mean standard deviation of two blended datasets for which you know their means and standard deviations:

The difference between the means of two samples, A and B, both randomly drawn from the same normally distributed source population, belongs to a normally distributed sampling distribution whose overall mean is equal to zero and whose standard deviation (”standard error”) is equal to

$$\sqrt{(sd^2/n_a) + (sd^2/n_b)},$$

where sd^2 is the variance of the source population.

If each of the two coefficient estimates in a regression model is statistically significant, do you expect the test of both together is still significant?

- ⋆ The primary result of a regression analysis is a set of estimates of the regression coefficients.
- ⋆ These estimates are made by finding values for the coefficients that make the average residual 0, and the standard deviation of the residual term as small as possible.
- ⋆ In order to see if there is evidence supporting the inclusion of the variable in the model, we start by hypothesizing that it does not belong, i.e., that its true regression coefficient is 0.
- ⋆ Dividing the estimated coefficient by the standard error of the coefficient yields the t-ratio of the variable, which simply shows how many standard-deviations-worth of sampling error would have to have occurred in order to yield

an estimated coefficient so different from the hypothesized true value of 0.

- ⋆ What the relative importance of variation in the explanatory variables is in explaining observed variation in the dependent variable? The beta-weights of the explanatory variables can be compared.
- ⋆ Beta weights are the regression coefficients for standardized data. Beta is the average amount by which the dependent variable increases when the independent variable increases one standard deviation and other independent variables are held constant. The ratio of the beta weights is the ratio of the predictive importance of the independent variables.

How to analyse non-normal distributions?

- ⋆ When data is not normally distributed, the cause for non-normality should be determined and appropriate remedial actions should be taken.
- ⋆ Causes:
  - **Extreme values**: too many extreme values in a data set will result in a skewed distribution. Find measurement errors and remove outliers with valid reasons.
  - **Overlap of two processes**: see if it looks bimodal and stratify the data.
  - **Insufficient data discrimination**: round-off errors or measurements devices with poor resolution makes data look discrete and not normal. Use more accurate measurement systems or collecting more data.
  - **Sorted data**: if it represents simply a subset of the total output a process produced. It can happen if the data is collected and analysed after sorting.
  - **Values close to zero or a natural limit**: the data will skew. All data can be raised or transformed to a certain exponent (but we need to be sure that a normal distribution can be assumed).
  - **Data Follows a Different Distribution**: Poison distribution (rare events), Exponential distribution (bacterial growth), log-normal distribution (to length data), binomial distribution (proportion data such as percent defectives).
  - Equivalente tools for non-normally distributed data:
    - ∗ t-test, ANOVA: median test, kruskal-wallis test.
    - ∗ paired t-test: one-sample sign test.

How do you test a data sample to see whether it is normally distributed?

- ⋆ I would first plot the frequencies, comparing the histogram of the sample to a normal probability curve.
- ⋆ This might be hard to see if the sample is small, in this case we can regress the data against the quantities of a normal distribution with the same mean and variance as the sample, the lack of fit to the regression line suggests a departure from normality.
- ⋆ Another thing you can do, a back-of-the-envelope calculation, is taking the sample's maximum and minimum and compute the z-score (or more properly the t-statistics), which is the number of sample standard deviations that a sample is above or below the mean:

$$z = \frac{x - \bar{x}}{s/\sqrt{N}},$$

  and then compares it to the 68-95-99.7 rule.
- ⋆ If you have a 3 sigma event and fewer than 300 samples, or a 4 sigma event and fewer than 15k samples, then a normal distribution understates the maximum magnitude of deviations in the data.

- ⋆ For example: 6 sigmas events don't happen in normal distributions!

What's a Hermitian matrix? What important property does a Hermitian matrix's eigenvalues possess?

- ⋆ It is equal to its conjugate transpose, for example the Paul matrices. A real matrix is Hermitian if it is symmetric.
- ⋆ The entries in the main diagonal are real.
- ⋆ Any Hermitian matrix can be diagonalized by a unitary matrix (U*U=I)and all the eigenvalues are real and the matrix has n linearly independent eigenvectors.

Random variable X is distributed as N(a, b), and random variable Y is distributed as N(c, d). What is the distribution of (1) X+Y, (2) X-Y, (3) X*Y, (4) X/Y?

- ⋆
  $$N(\mu, \sigma) = \frac{1}{\sigma\sqrt{2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

  or the *standard normal distribution*, described by the probability density function:

  $$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{=\frac{1}{2}x^2}.$$

- ⋆
  $$Z \sim N(\mu_x, \sigma_x^2), Y \sim N(\mu_y, \sigma_y^2)$$

  them by convolution, we see that the Fourier transform is a Gaussian PDF with:

  $$Z = X + Y \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$

  .

- ⋆ Requires the assumption of independence: if the random variables are correlated, the joint random variables are still normally distributed however their variances are not additive, and we need to calculate de correlation.
- ⋆ Any linear combination of independent normal deviates is a normal deviate!
- ⋆ The product of two Gaussian PDFs is proportional to a Gaussian PDF with mean that is half of the coefficient of x and se that is the square root of half of the denominator:

  $$\sigma_{XY} = \sqrt{\frac{\sigma_x^2 \sigma_y^2}{\sigma_x^2 + \sigma_y^2}},$$

  and

  $$\mu_{XY} = \frac{\mu_x \sigma_y^2 + \mu_y \sigma_x^2}{\sigma_x^2 + \sigma_y^2}.$$

How to Write a Competitive Analysis?

- ⋆ **Your company's competitors**:
  - list of your company's competitors.
  - companies that indirectly compete with yours, ones that offer products or services that are aiming for the same customer capital.
- ⋆ **Competitor product summaries**:
  - Analyse the competition's products and services in terms of features, value, and targets.
  - How do your competitor's sell their wares? How do they market them?
  - Customer satisfaction surveys conducted. How do customers see your competition?
- ⋆ **Competitor strengths and weaknesses**:
  - be objective, no bias.
  - What makes their products so great?
  - If they are growing rapidly, what is it about their product or service that's promoting that growth?
  -
- ⋆ **The market outlook**:

  - What is the market for your company's product like now?
  - Is it growing? If so, then there are likely quite a few customers left to go around.
  - If it is flat, then the competition for customers is likely to be fierce.
  - Is the market splintering, is it breaking up into niches?

What is the yearly standard deviation of a stock given the monthly standard deviation?

$$SD_y = SD_m\sqrt{12}$$

Given a dataset, how do you determine its sample distribution? Please provide at least two methods.

- ⋆ Plot them using histogram: it should give you and idea as to what parametric family(exponential, normal, log-normal...) might provide you with a reasonable fit. If there is one, you can proceed with estimating the parameters.
- ⋆ If you use a large enough statistical sample size, you can apply the Central Limit Theorem to a sample proportion for categorical data to find its sampling distribution.
- ⋆ You may have 0/1 responses, so the distribution is binomial, maybe conditional on some other covariates, that's a logistic regression.
- ⋆ You may have counts, so the distribution is Poisson.
- ⋆ Regression to fit, test chi-square to the goodness of the fit.
- ⋆ Also a small sample (say under 100) will be compatible with many possible distributions, there is simply no way distinguishing between them on the basis of the data only. On the other hand, large samples (say more than 10000) won't be compatible with anything.

What's the expectation of a uniform(a, b) distribution? What's its variance?

- ⋆ The PDF:
$$f(x) = \frac{1}{b-a}$$
for $a \leq x \leq b$ and 0 otherwise.
- ⋆ Mean: $E(X) = \frac{a+b}{2}$
- ⋆ Variance: $V(X) = \frac{(b-a)^2}{12}$

What's the difference between the t-stat and R2 in a regression? What does each measure? When you get a very large value in one but a very small value in the other, what does that tell you about the regression?

- ⋆ Correlation is a measure of association between two variables, the variables are not designated as dependent or independent. The significance (probability) of the correlation coefficient is determined from the t-statistic. The test indicates whether the observed correlation coefficient occurred by chance if the true correlation is zero.
- ⋆ Regression is used to examine the relationship between one dependent variable and one independent variable. The regression statistics can be used to predict the dependent variable when the independent variables is known. Regression goes beyond correlation by adding prediction capabilities.
- ⋆ The significance of the slope of the regression line is determined from the t-statistic. It is the probability that the observed correlation coefficient occurred by chance if the true correlation is zero. Some researchers prefer to report the F-ratio instead of the t-statistic. The F-ratio is equal to the t-statistic squared.
- ⋆ The t-statistic for the significance of the slope is essentially a test to determine if the regression

model (equation) is usable. If the slope is significantly different than zero, then we can use the regression model to predict the dependent variable for any value of the independent variable.

- The coefficient of determination (r-squared) is the square of the correlation coefficient. Its value may vary from zero to one. It has the advantage over the correlation coefficient in that it may be interpreted directly as the proportion of variance in the dependent variable that can be accounted for by the regression equation. For example, an r-squared value of .49 means that 49% of the variance in the dependent variable can be explained by the regression equation.
- The R^2 statistics is the amount of variance in the dependent variable explained by your model. The F statistics is whether your model is significant. The t statistics (if it's a linear model of more than one predictor) is a measure of whether or not the individual predictors are significant predictors of the dependent variable. The rejection of H_0 depends on what it is. If it's about your overall model(s) you may look at the F and/or R^2 . If it is about the predictors you look at the t statistics (but only if your overall model is significant).
- The t-values and R2 are used to judge very different things. The t-values are used to judge the accuracy of your estimate of the β_i's, but R2 measures the amount of variation in your response variable explained by your covariates.
- If you have a large enough dataset, you will always have statistically significant (large) t-values. This does not mean necessarily mean your covariates explain much of the variation in the response variable.

What is bigger, the mean or the median?

If the distribution shows a positive skew, the mean is larger than the median. If it shows a negative skew, the mean is smaller than the median.

Pretend 1% of the population has a disease. You have a test that determines if you have that disease, but it's only 80% accurate and 20% of the time you get a false positive, how likely is it you have the disease.

- Fact: 0.01 of the population has the disease (given).
- Data: Test is only 0.8 accurate (given):
- How likely is it that you have the disease?
- To identify that you have the disease you have to test +ve and actually have the disease. Using Bayes:

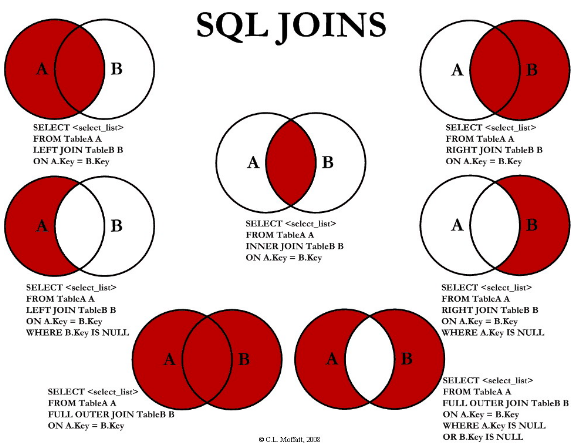
$$P(B|A) = \frac{0.8 * 0.01}{0.01 * 0.8 + 0.99 * 0.2} = 0.04$$

SQL, what are the different types of table joins? What's the difference between a left join and a right join?

Inner Join, Left Outer Join, Right Outer Join, Full Outer Join, Cross Join.

- Simple Example: Lets say you have a Students table, and a Lockers table.
- INNER JOIN is equivalent to "show me all students with lockers".
- LEFT OUTER JOIN would be "show me all students, with their corresponding locker if they have one".
- RIGHT OUTER JOIN would be "show me all lockers, and the students assigned to them if there are any".
- FULL OUTER JOIN would be silly and probably not much use. Something like "show me all students and all lockers, and match them up where you can"

- ⋆ CROSS JOIN is also fairly silly in this scenario. It doesn't use the linked "locker number" field in the students table, so you basically end up with a big giant list of every possible student-to-locker pairing, whether or not it actually exists.



```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
```

```
SELECT <select_list>
FROM TableA A
INNER JOIN TableB B
ON A.Key = B.Key
```

```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
```

```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
WHERE B.Key IS NULL
```

```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
WHERE A.Key IS NULL
```

```
SELECT <select_list>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.Key = B.Key
```

```
SELECT <select_list>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.Key = B.Key
WHERE A.Key IS NULL
OR B.Key IS NULL
```

© C.L. Moffatt, 2008

How many gigabytes would you need to run Google mail. How much do you think GMail costs to Google?

- ⋆ 16 000 gb is 800 dol/month, so 1 GB is 0.5 and each user have 15 GB, so 7.5 dollar.
- ⋆ Gmail has around 400 million user, so 3 billions month.
- ⋆ Marginal cost is the additional cost for adding a user.

How much revenue does Youtube make in a day?

- ⋆ Youtube's revenues at 4 billions for 2013 and their operating income at 711 millions. This puts their daily revenues at 11 millions and their daily income at 1.9 millions.
- ⋆ Average of 133 page views per month per user.

With a data set from normal distribution, suppose you can't get any observation greater than 5, how to estimate the mean?

- ⋆ Statistical analysis with small samples is like making astronomical observations with binoculars. You are limited to seeing big things: planets, stars, moons and the occasional comet. But just because you don't have access to a high-powered telescope doesn't mean you cannot conduct astronomy.
- ⋆ Comparing Means: If your data is generally continuous (not binary), such as task time or rating scales, use the two sample t-test. It's been shown to be accurate for small sample sizes.
- ⋆ Comparing Two Proportions: If your data is binary (pass/fail, yes/no), then use the N-1 Two Proportion Test. This is a variation on the better known Chi-Square test.
- ⋆ Confidence Intervals: While the confidence interval width will be rather wide (usually 20 to 30 percentage points), the upper or lower boundary of the intervals can be very helpful in establishing how often something will occur in the total user population.
- ⋆ Confidence interval around a mean: If your data is generally continuous (not binary) such as rating scales, order amounts in dollars, or the number of page views, the confidence interval is based on the t-distribution (which takes into account sample size).
- ⋆ Confidence interval around task-time: Task time data is positively skewed. There is a lower boundary of 0 seconds. It's not uncommon for some users to take 10 to 20 times longer than other users to complete the same task. To handle this skew, the time data needs to be log-transformed and the confidence interval is computed on the log-data, then transformed back when reporting.

- ⋆ Confidence interval around a binary measure: For an accurate confidence interval around binary measures like completion rate or yes/no questions, the Adjusted Wald interval performs well for all sample sizes.

- ⋆ Completion Rate: For small-sample completion rates, there are only a few possible values for each task. For example, with five users attempting a task, the only possible outcomes are $0\%, 20\%, 40\%, 60\%, 80\%$ and 100% success. It's not uncommon to have 100% completion rates with five users.

Derive the formula for the variance of OLS from scratch.

- ⋆ linear least squares is a method for estimating the unknown parameters in a linear regression model.

- ⋆ this method minimizes the sum of squared vertical distances between the observed responses in the dataset and the responses predicted by the linear approximation.

- ⋆ The resulting estimator can be expressed by a simple formula, especially in the case of a single regressor on the right-hand side.

- ⋆ Given a data set of n statistical units, a linear regression model assumes that the relationship between the dependent variable yi and the p-vector of regressors xi is linear. This relationship is modelled through a disturbance term or error variable ϵ_i an unobserved random variable that adds noise to the linear relationship between the dependent variable and regressors.

- ⋆ Thus the model takes the form

$$y_i = \beta_1 x_{i1} + .. + \beta_p x_{ip} + \epsilon_i = x_i^T \beta + \epsilon_1$$

where T denotes the transpose, so that xiT is the inner product between vectors xi and .

Bayesian analysis of drug testing (determining false positive rate).

- ⋆ Let say 0.5% of people are drug users and our test is 99% accurate (it correctly identifies 99% . What's the probability of being a drug user if you've tested positive?

- ⋆ $P(B|A)$ is 0.99, $P(B)$= 0.01*0.995+0.99*0.005, $P(A) = 0.005$.

I have a coin, I tossed 10 times and I want to test if the coin is fair.

- ⋆ A fair coin is an idealized randomizing device with two states (usually named "heads" and "tails") which are equally likely to occur.

- ⋆ In more rigorous terminology, the problem is of determining the parameters of a Bernoulli process, given only a limited sample of Bernoulli trials.

- ⋆ If a maximum error of 0.01 is desired, how many times should the coin be tossed? At 68.27%, level of confidence (Z=1), at 95.4% level of confidence (Z=2), at 99.90% level of confidence (Z=3.3).

- ⋆ Hypothesis testing lets you to decide, with a certain level of significance, whether you have sufficient evidence to reject the underlying (Null) hypothesis or you have do not sufficient evidence against the Null Hypothesis and hence you accept the Null Hypothesis.

- ⋆ I am explaining the Hypothesis testing below assuming that you want to determine if a coin comes up heads more often than tails. If you want to determine, if the coin is biased or unbiased, the same procedure holds good. Just that you need to do a two-sided hypothesis testing as opposed to one-sided hypothesis testing.

- ⋆ your Null hypothesis is $p \leq 0.5$ while your Alternate hypothesis is $p > 0.5$, where p is the probability that the coin shows up a head. Say now you want to perform your hypothesis testing at 10% level of significance. What you do now is to do as follows:
- ⋆ Let nH be the number of heads observed out of a total of n tosses of the coin.
- ⋆ Take p=0.5 (the extreme case of the Null Hypothesis). Let x∼B(n,0.5).
- ⋆ $P(x \geq ncH) = 0.1$, ncH gives you the critical value beyond which you have sufficient evidence to reject the Null Hypothesis at 10% level of significance. i.e. if you find nH≥ncH, then you have sufficient evidence to reject the Null Hypothesis at 10% level of significance and conclude that the coin comes up heads more often than tails.
- ⋆ If you want to determine if the coin is unbiased, you need to do a two-sided hypothesis testing as follows.
- ⋆ Your Null hypothesis is p=0.5 while your Alternate hypothesis is p≠ %0.5, where p is the probability that the coin shows up a head. Say now you want to perform your hypothesis testing at 10% level of significance. What you do now is to do as follows:
- ⋆ Let nH be the number of heads observed out of a total of n tosses of the coin.
- ⋆ Let x∼B(n,0.5).
- ⋆ Compute nc1H and nc2H as follows. $P(x \leq nc1H) + P(x \geq nc2H) = 0.1$ (nc1H and nc2H are symmetric about n2 i.e. nc1H+nc2H=n) nc1H gives you the left critical value and nc2H gives you the right critical value.
- ⋆ If you find nH(nc1H,nc2H), then you have do not have sufficient evidence against Null Hypothesis and hence you accept the Null Hypothesis at 10% level of significance. Hence, you accept that the coin is fair at 10% level of significance.

Why we can not use linear regression for dependent variable with binary outcomes?

- ⋆ the linear regression can produce predictions that are not binary, and hence ”nonsense”: interactions that are not accounted for in the regression and non-linear relationships between a predictor and the outcome
- ⋆ inference based on the linear regression coefficients will be incorrect. The problem with a binary dependent variable is that the homoscedasticity assumption (similar variation on the dependent variable for units with different values on the independent variable) is not satisfied. I will add that another concern is that the normality assumption is violated: the residuals from a regression model on a binary outcome will not look very bell-shaped... Again, with a sufficiently large sample, the distribution does not make much difference, since the standard errors are so small anyway.

How to estimating sample size required for experiment?

Sample sizes may be chosen in several different ways:

- ⋆ expedience - For example, include those items readily available or convenient to collect. A choice of small sample sizes, though sometimes necessary, can result in wide confidence intervals or risks of errors in statistical hypothesis testing.
- ⋆ using a target variance for an estimate to be derived from the sample eventually obtained
- ⋆ using a target for the power of a statistical test to be applied once the sample is collected.
- ⋆ How to determine the sample size?
  - – for a study for which the goal is to get a significant result from at test:

- set α to the residual power
- set H_A, estimate σ

Bibliography

[1] Quantifying the User Experience, Jeff Sauro and James R. Lewis, 2012

[2] OReilly Statistics in a Nutshell, 2012

[3] https://class.coursera.org/introstats-001/class/index