

Into the roots

Parikshit Sanyal – Version 1.0, 2006-20

Table of Contents

General principles

About learning and teaching

The nature of truth

 The scientist's dilemma

 Descriptive statistics

 Central tendency

 Dispersion

 Sampling

 Errors in sampling

 Calculation of sample size

Research design

 Descriptive studies

 Analytical studies

Inferential statistics

 Probability

 Probability distributions and Models

 Discrete probability distributions: X can have only a finite number of values

 Continuous probability distributions: X can have any real value within an interval

 Bayesian statistics

 Hypothesis testing

 Regression Models

 Relation between categorical and numeric variables

 Multiple regression

Index

General principles

About learning and teaching

“*Bene ascolta chi la nota.*

("He listens well who takes notes.")

— *Dante Inferno. 15:99*

“*"The time has come," the Walrus said,*

"To talk of many things:"

— *Lewis Carroll*

Through the Looking-Glass and What Alice Found There

The nature of truth

The scientist's dilemma

Science believes in generalising the special, to make general rules that apply to everybody, from a *limited* number of observations. This is the only way research is done, because it is not humanly possible to examine every human on this planet to verify if he has a set of femurs. When Vesalius cut up his very first corpse, he *instantly* made the decision that the human *species*, not just the specimen on his table, but the *entirety* of the human species, must have two femurs. Herein lies the scientist's dilemma. A *formal proof*, as iterated in pure mathematics, should not be based on induction (that is, no one specimen should be held representative of whole of the population). A proof must establish, by rigorous mathematical procedures, an identity between two sides of an equation. This kind of reasoning, however smart it may sound, is absurd in most kinds of research except pure mathematics. For the more mundane kind of research, we can not examine whole populations and we have to deal with samples. A few questions immediately spring up

- what should be the sample size to make a reasonable conclusion (i.e. can I make a statement "every human has two femurs" by examining only one corpse, or do I need more samples)
- what should be the benchmark of 'statistical significance' (i.e. how can I decide if between two events, one is being caused by another, or if its just by chance - there is no real relationship between them)

- what will be the amount of error (how valid will be our research) when we deviate from the mathematical 'formal proof' method (i.e. by doing research on samples)
- how much will the research be biased by a distorted sample (i.e. a study done on shoe size on a sample of acromegals)

"How many roads must a man walk down Before you call him a man? How many seas must a white dove sail Before she sleeps in the sand? How many times must the cannon balls fly Before they're forever banned? The answer, my friend, is blowin' in the wind The answer is blowin' in the wind"

(Bob Dylan, 1962)

Statistics is the only plausible answer to these questions.

Measurements

By definition, any set of rules for assigning numbers to attributes of objects is *measurement*. Not all measurement techniques are equally useful in dealing with the world, however, and it is the function of the scientist to select those that are more useful. The physical and biological scientists generally have well-established, standardized, systems of measurement, unlike social scientists.

The issue of measurements were discussed in great detail by S.S.Stevens in an article in 1951.

Properties of measurement scales

Magnitude

The property of *magnitude* exists when an object that has more of the attribute than another object, is assigned a larger number by the rule system, i.e. if A is heavier than B then weight of A is more than weight of B.

Intervals

The property of intervals is concerned with the relationship of differences between objects. If a measurement system possesses the property of intervals it means that the unit of measurement means the same thing throughout the scale of numbers. That is, an inch is an inch, no matter where it falls - immediately ahead or a mile down the road.

Rational Zero

A measurement system possesses a rational zero if an object that has none of the attribute in question is assigned the number zero by the system of rules. The object does not need to really exist in the "real world", as it is somewhat difficult to visualize a "man with no height". The requirement for a rational zero is this: if objects with none of the attribute did exist would they be given the value zero.

Scale types

In the same article in which he proposed the properties of measurement systems, S. S. Stevens (1951) proposed four scale types. These scale types were Nominal, Ordinal, Interval, and Ratio, and each possessed different properties of measurement systems.

Nominal Scales

Nominal scales are measurement systems that possess none of the three properties discussed earlier. Nominal renaming scales apply random numbers (or words) to objects (i.e. social security numbers, classification of diseases). Nominal categorical scales apply a different number to each category of objects (i.e. Belgians = 1, Indians = 2, Irish = 3).

Ordinal Scales

Ordinal Scales are measurement systems that possess the property of magnitude, but not the property of intervals. The property of rational zero is not important if the property of intervals is not satisfied. Any time ordering, ranking, or rank ordering is involved, the possibility of an ordinal scale should be examined. As with a nominal scale, computation of most of the statistics described in the rest of the book is not appropriate when the scale type is ordinal. Rank ordering people in a classroom according to height and assigning the shortest person the number "1", the next shortest person the number "2", etc. is an example of an ordinal scale.

Interval Scales

Interval scales are measurement systems that possess the properties of magnitude and intervals, but not the property of rational zero (i.e. the height of a person). It is appropriate to compute the statistics described in the rest of the book when the scale type is interval.

NOTE

Quantiles: How to remove the property of intervals

Arranging lists according to centiles (i.e. 'X belongs to 91st centile' means that 91% of values fall below X) or quartiles (the list divided in four quarters) or deciles (ten parts) removes the interval property, and converts an interval scale to an ordinal scale.

Ratio Scales

Ratio scales are measurement systems that possess all three properties: magnitude, intervals, and rational zero. The added power of a rational zero allows ratios of numbers to be meaningfully interpreted; i.e. the ratio of John's height to Mary's height is 1.32, whereas this is not possible with interval scales.

Descriptive statistics

Data are the figures you derive directly from the source. Primary data is obtained directly from the population (as in census), and secondary data from a record. Obviously, primary data is always more reliable than secondary.

Data has two parts, i.e. a variable (like age, gender, income, number of spouses etc. - denoted by x) and a value (i.e. 40 years, male, Rs. 3500, 3).

A *variable* may be qualitative (Boolean data, subjective data) or quantitative (numbers). Quantitative data may be further classified as *continuous* (i.e. height, weight - which may have any value) or *discrete* (size of shoes - which can have only a fixed set of values).

Universe is the extent of a statistical survey being undertaken, also called the population. The numerical size of the population is denoted by n . Subgroups in this universe are called samples.

An *event* is just that, an event that has or will occur.

Representation of data: Frequency distribution

This special variety of tables describe frequency of an event within *class intervals* of a universe. Mind that the class intervals (the domains) do not overlap and are of equal width.

Table 1. Frequency distribution table of height of students in a class

Height (cm)	Number of students
163	9
165	12
167	15
170	10
173	7
177	4
180	3
182	1

A *histogram* is the representation of an frequency distribution table, so the bars are adjacent (to emphasize the discrete nature of the variable being measured, in this case, shoe size, which can be any value between 7-11). By adding the midpoint of top of these bars – you can get a *frequency polygon*.

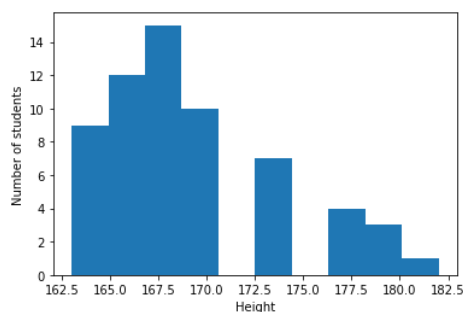


Figure 1. A histogram showing distribution of shoe sizes in a sample of students

Representation of data: box plots

A box plot is a five figure summary of dispersion.

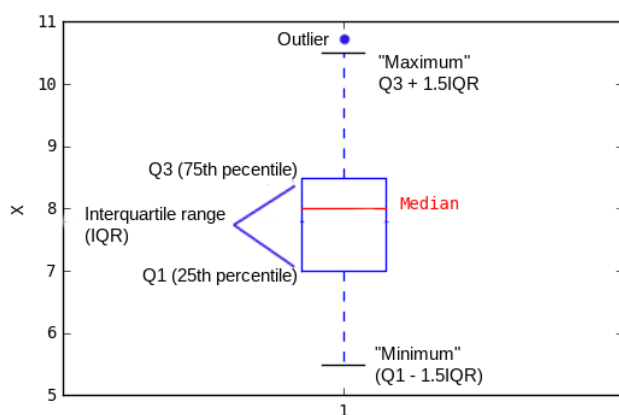


Figure 2. Box plot of data

Central tendency

1. The **mean** is the arithmetic average, denoted by an \bar{X} (for a sample) or μ (for the whole population). It is not, however, a very good indicator of the actual distribution of the variable. Suppose you admit a dinosaur (or any fellow with really big feet) in your class, then

the mean shoe size will be dramatically altered even though only one member has been added. Mean is affected severely by the values at the end.

2. The **median** is the midpoint value of a distribution arranged in ascending order, or the average of two midpoint values (if number of data is even). The median is not affected by terminal values.
3. The **mode** is the most commonly occurring value in a distribution.

Dispersion

Range

The set of values X can take.

Mean deviation

A particular value x of a variable is said to be deviating from the mean \bar{X} by an amount $x - \bar{X}$. This is the deviation of x from mean. The mean deviation is the average of all such deviations. In a population of size η

$$d = \frac{\sum |X - \mu|}{\eta}$$

The mean deviation gives an idea on how widely the data varies. Mind the absolute value sign '|' around the deviations. If we do not ignore the sign of deviations, positive and negative variations tend to cancel each other out.

Variance

For an entire population, variance

$$\sigma^2 = \frac{\sum (X - \bar{X})^2}{\eta}$$

For any sample of size n , the variance s^2 is

$$s^2 = \frac{\sum (X - \bar{X})^2}{n}$$

If the mean of the *population* is unknown, then we must use Bessel's correction (why? see later).

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

The variance of a sample shows how widely it is distributed. Suppose, in two different classes, the mean shoe size is same. In this case, the class with the more variance has a more varying set of students (i.e. there are more number of students who have a very small or very large shoe size) than the class with less variance. Variance is only a number, and its unit is the square of the unit of the thing we want to measure. The reason we use a square is that marginal values get even more marginal after squaring.

Standard deviation

Standard deviation is the positive square root of variance.

$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{\eta}}$$

when dealing with the whole population, or

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

when in a sample. If the mean of the population is unknown, Bessel's correction must be applied

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

Properties of standard deviation

For constant c and random variable X ,

- $\sigma(X + c) = \sigma(X)$; the SD is not changed if each value is incremented by same amount
- $\sigma(cX) = |c|\sigma(X)$; i.e. if each value of a population gets multiplied, the SD is also multiplied

Importance of standard deviation

The standard deviation serves as the 'unit' of variability. We speak 'the shoe size of this student is 3 standard deviations more than the mean, i.e. the shoe size is mean + 3 × standard deviation = 7.818 + 3 × 1.266 = 11.616

Example 1. Finding the mean and SD of a population

Say the shoe sizes in a class of 101 are distributed as follows.

Number of students	Shoe size (X)
5	5.5

6	6
9	6.5
13	7
17	7.5
13	8
13	8.5
10	9
8	9
5	10
2	10.5

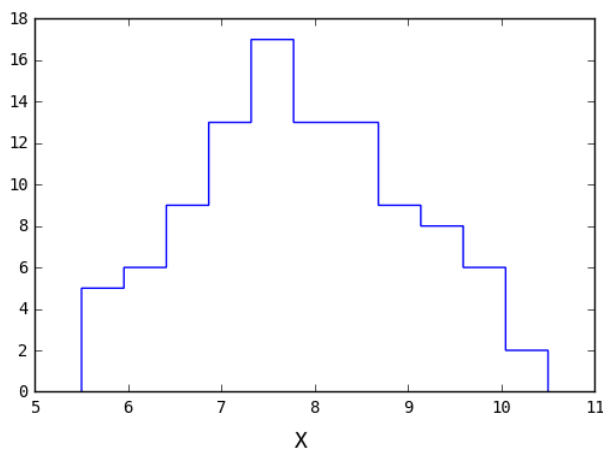


Figure 3. Histogram of the shoe sizes in this class

In this population

- size of population $\eta = 101$
- **mean** of the population $\mu = 7.85$
- **variance** of the population $\frac{(X - \mu)^2}{\eta} = 1.41$
- **standard deviation** $\sigma = \sqrt{\text{variance}} = 1.19$

The distribution does not have any particular shape, it's random. So no statistical test can not be applied to it directly. The most we can say, from the table, is that $p(\text{shoe size } 10 \text{ or more}) = (5 + 2) / 101 = 0.069$, and similar probabilities.

NOTE What is this degree of freedom thing? A fine example from David Lane,

Some estimates are better than others. An estimate based on 100 samples is better than that based on 5. The **degrees of freedom (df)** of an estimate is the number of independent pieces of information that has gone into it.

As an example, let's say that we know that the mean height of Martians is 6 and wish to estimate the variance of their heights. We randomly sample one Martian and find that its height is 8. so variance $(8-6)^2 = 4$, is an estimate of the mean squared deviation for all Martians. Therefore, based on this sample of one, we would estimate that the population variance is 4. This estimate is based on a single piece of information and therefore has 1 df.

If we sampled another Martian and obtained a height of 5, then we could compute a second estimate of the variance, $(5-6)^2 = 1$. We could then average our two estimates (4 and 1) to obtain an estimate of 2.5. Since this estimate is based on two independent pieces of information, it has two degrees of freedom.

However, mostly we are not aware of the population mean when we are estimating the variance. Instead, we have to first estimate the population mean (μ) with the sample mean (\bar{X}). The process of estimating the mean affects our degrees of freedom. We have sampled two Martians and found that their heights are 8 and 5. Therefore \bar{X} , our estimate of the population mean, is

$$\bar{X} = \frac{8 + 5}{2} = 6.5$$

We can now compute two estimates of variance:

1. Estimate 1 = $(8-6.5)^2 = 2.25$
2. Estimate 2 = $(5-6.5)^2 = 2.25$

Now for the key question: Are these two estimates independent? The answer is no because each height contributed to the calculation of \bar{X} . Since the first Martian's height of 8 influenced \bar{X} , it also influenced Estimate 2. Another way to think about the non-independence is to consider that if you knew the mean and one of the scores, you would know the other score. For example, if one score is 5 and the mean is 6.5, you can compute that the total of the two scores is 13 and therefore that the other score must be $13-5 = 8$.

In general, the degrees of freedom for an estimate is equal to the number of values minus the number of parameters estimated en route to the estimate in question. In the Martians example, there are two values (8 and 5) and we had to estimate one parameter (μ) on the way to estimating the parameter of interest (σ^2). Therefore, the estimate of variance has $2 - 1 = 1$ degree of freedom. If we had sampled 12 Martians, then our estimate of variance would have had 11 degrees of freedom. Therefore, the degrees of freedom of an estimate of variance is equal to $n - 1$, where n is the number of observations.

“

Bessel's correction

Lets pick a sample of eleven (11) random students from this population. For the sake of argument, we'll assume that we do not know the entire population yet. We just have the sample at hand.

Serial no	Shoe size (X)	Mean (\bar{X})	Deviation $X - \bar{X}$	$ X - \bar{X} $	Mean deviation ($\frac{\sum X - \bar{X} }{n}$)	$(X - \bar{X})^2$	Variance $s^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$	Standard deviation s_n
		7.818			12.182/11 = 1.107		17.63/ (11-1) = 1.763	$\sqrt{1.763} = 1.32$
1	6		-1.818	1.818		3.3		
2	6		-1.818	1.818		3.3		
3	7		-0.818	0.818		0.67		
4	7		-0.818	0.818		0.67		
5	7		-0.818	0.818		0.67		
6	8		0.182	0.182		0.03		
7	8		0.182	0.182		0.03		
8	9		1.182	1.182		1.39		
9	9		1.182	1.182		1.39		
10	9		1.182	1.182		1.39		
11	10		2.182	2.182		4.76		

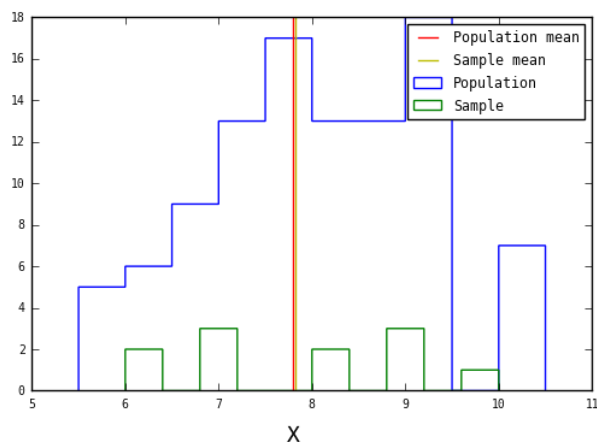


Figure 4. Histogram of population and sample of shoe sizes

We assume now, that the sample mean is a good approximate of population mean (given that the sample is sufficiently random), and thus try to calculate the variance of the sample, which by way of extension, should serve as variance of the entire *population*.

$$s^2 = \sum \frac{(X - \bar{X})^2}{n} = \frac{17.63}{11} = 1.6033$$

Now let's look back at the population, its mean is 7.8. If we do the same calculation with the population mean

$$s^2 = \sum \frac{(X - \mu)^2}{n} = 1.6035$$

This is bigger (and will always be bigger, as can be proved) from the variance which is derived from the population mean. Our purpose is to calculate sum of the squared distance from the population mean, $X - \mu$, which can be written as

$$\sum (X - \mu)^2 = \sum (X - \bar{X} + \bar{X} - \mu)^2 = \sum (X - \bar{X})^2 + \sum 2(X - \bar{X})(\bar{X} - \mu) + \sum (\bar{X} - \mu)^2$$

The sum of $(X - \bar{X})$, i.e. sum of distances from mean, will naturally be 0, so the entire middle term is 0; thus continuing

$$\sum (X - \mu)^2 = \sum (X - \bar{X})^2 + \sum (\bar{X} - \mu)^2$$

Which proves that calculating from the sample mean will always produce a result smaller than calculating from the population mean.

The root cause of the issue

When population mean is fixed, the n elements of a sample are not free to have any value. The first $n - 1$ elements might take on any value, but the last one has to conform to regress to the population mean. Thus, the *degree of freedom* is reduced to $(n - 1)$.

Sampling

The use of sampling has already been underlined. In most studies, we will be sampling **without replacement**, i.e. the same member will not be repeated in a sample.

Random sampling

Random sampling is left entirely to nature's own laws of entropy, and everyone has equal probability of being selected. It provides the greatest number of possible samples, but it is also most prone to produce a distorted sample. For example, out of fifty girls and fifty boys, a random sample of 20 has to be chosen. Now it is perfectly possible that the sample of 20 that we choose has exactly 10 girls and 10 boys. It is, however, much more probable for our sample to be distorted either way, i.e. we could select more girls than boys or vice versa. It is, by sheer chance, also possible that we select only girls, and that would be an embarrassingly distorted sample, not even close to representing the population.

Matched random sampling

If we really want to do a scientific study, let's say about the pattern recognition skill differences between boys and girls, we must select pairs of a boy and a girl, who are identical in all aspects (age, mental growth, family background etc) except that one is a boy and other a girl. We could make a comparison only if such matching has been done.

Systematic sampling

Suppose we pick every 10th person from our hundred (i.e. 1, 11, 21, 31 ... or 4, 14, 24 etc), which reduces the number of possible samples (in our particular case, only 10 samples can now be chosen, beginning from 1, 11, 21 to 9, 19, 29). However, if the boys and girls are so arranged that every 10th person is a girl (i.e. if there is periodicity in the population and we resonate with that periodicity), we would end up with 10 girls again. This is the drawback of systematic sampling.

Stratified sampling

We could divide the population into reasonable groups (strata) and then take samples from each group so that no group gets preselected (we could split our hundred into 'boys' stratum and 'girls' stratum, and then go on random/ systematic sampling within each stratum; this way ensure that we do not end up with only girls or only boys in our sample). Stratification can be done on the basis of age, sex, religion or any other attribute.

Cluster sampling

We could also set up groups of 10 among our hundred (each group may include both boys and girls) and select one from each group. This kind of sampling is used in immunisation survey among children.

Convenience sampling

This is waiting for samples to come to you, i.e. noting down the model of every car passing down the road. This is rife with bias and has very little use in statistics.

Errors in sampling

Sampling error

Each sample of a universe differs from another sample, and this is unavoidable, omnipresent whim of nature.

Non sampling errors

1. Overcoverage: Inclusion of data from outside of the population.
2. Undercoverage: Sampling frame does not include elements in the population.
3. Measurement error: The respondent misunderstands the question.

4. Processing error: Mistakes in data coding.
5. Non-response: People unwilling to take part in a survey may get included in the sample.

Calculation of sample size

Where the sample and population are identical in characteristic, statistical theory yields exact recommendations on sample size. However, where it is not straightforward to define a sample representative of the population, it is more important to understand the cause system of which the population are outcomes and to ensure that all sources of variation are embraced in the sample. Large number of observations are of no value if major sources of variation are neglected in the study.

In general, if we want a margin of error (d) $z\sigma_{\bar{X}}$, then it follows

$$d = z\sigma_{\bar{X}} = z \frac{\sigma}{\sqrt{n}}$$

or

$$n = z^2 \frac{\sigma^2}{d^2}$$

If σ is not known, we have to do with the sample SD (s).

Example 2. Calculation of sample size: The simplest scenario

Previous studies have shown that average shoe size of 10 year old children born to a particular community is 7.5in, with SD of 1.5in. We want to study effect of good childhood nutrition on shoe size. What is the sample size required, for a 95% confidence interval and a margin of error of of maximum 0.5 in?

$$\text{In this case, } n = \frac{z^2 \sigma^2}{d^2} = 1.96^2 \cdot 1.5^2 / 0.5^2 = 34.57.$$

Research design

Descriptive studies

The measurement of a variable over time/ place/ person, without any analysis of inference. Descriptive studies do not need a minimum sample size, and the usual outcome is

- a mean (if numeric variable)
- a proportion (if categorical variable)

Analytical studies

Usually some sort of inference is involved.

Cross sectional studies

Cross sectional studies measure snapshot of a variable in time/ place/ person.

Sample size for cross sectional studies

Categorical variables

You need

1. the Z score for an alpha limit; if the alpha limit is selected to be 0.05, then $Z = 1.96$ (two sided) or 1.64 (one sided)
2. prevalence p of the variable in population
3. margin or error d to be decided by you; depends on what exactly you are measuring

Suppose you want to estimate the number of acromegalics in the population. In previous pilot studies, the prevalence has come to be around 1% (0.01). For a binomial distribution such as this (disease present/ absent), the variance is $\sigma^2 = p(1 - p)$. Thus sample size

$$n = \frac{z^2 p(1 - p)}{d^2}$$

which comes around to be 15. However, if the same study were conducted to estimate malnutrition in children, which has an approximate prevalence of 10%, then the figure would turn out to be 138.

Numeric variables

$$n = \frac{z^2 \cdot \sigma^2}{d^2}$$

where σ is the known standard deviation of the variable (from previous pilot studies).

Case control studies

Case control studies measure previous 'exposure' to a factor which has caused the 'effect', in a retrospective manner.

Qualitative variables

Supposing

1. r = ratio of cases to controls
2. p_{exp} = (prevalence of the exposure in cases + prevalence of the variable in control) / 2 (from pilot studies)
3. z_{β} = the z score for the power of a test; for a power $(1-\beta)$ of 80%, $z_{\beta} = 0.84$; if power is increased to 90%, $z_{\beta} = 1.28$
4. z_{α} = the z score for any α ; for $\alpha = 0.05$, $z = 1.96$ (two sided)
5. p_c = prevalence of disease in cases (from pilot studies)
6. p_o = prevalence of disease in controls (from pilot studies)

Then, the sample size

$$n = \left(\frac{r+1}{r} \right) \cdot p_{\text{exp}} \cdot (1 - p_{\text{exp}}) \cdot \frac{(z_{\beta} + z_{\alpha})^2}{(p_c - p_o)^2}$$

Quantitative variables

If d is the mean difference of the variable between cases and controls and σ is the SD of the variable (both from previous studies)

$$n = \left(\frac{r+1}{r} \right) \cdot \sigma^2 \frac{(z_{\beta} + z_{\alpha})^2}{d^2}$$

Cohort studies

Cohort studies start with exposure and non-exposure groups, and follow them up to see the effect.

Given

1. r = ratio of controls to experimental group
2. p_e = prevalence of disease in experimental group (from previous studies)
3. p_c = prevalence of disease in control group (from previous studies)
4. $p = (p_e + r p_c) / (r+1)$

Then, sample size

$$n = \frac{\left(z_{\alpha} \cdot \sqrt{1 + \frac{1}{r}} \cdot p(1-p) + z_{\beta} \cdot \sqrt{p_c} \cdot \frac{1-p_c}{r} + p_e(1-p_e) \right)^2}{(p_c - p_e)^2}$$

Bayesian studies/ diagnostic test studies

Such studies compare a new test against a known 'gold standard' test. If the sensitivity of the gold standard test is sn and specificity sp , then sample size

$$n_{sn} = \frac{z^2 sn(1 - sn)}{d^2}$$

and

$$n_{sp} = \frac{z^2 sp(1 - sp)}{d^2}$$

Pick whichever is greater as sample size.

Inferential statistics

Probability

As much as we are fond of data and patterns of data, there is always a limit to how much data we can collect, and at some point of time, we have to stop collecting data and do some hypothesizing. It would have been very fortunate if we could measure all the data about every aspect of everybody, but until that happens, we have to indulge in speculation and forecasting. The beauty of inferential statistics is that, not only does it allow you to make a reasonable prediction, but also allows you to specify the possible amount of error (i.e. "I am 93% sure that there is 13% chance of rain today"). Probability is a theory of uncertainty which deals with these speculations. It is a necessary concept because the world according to the scientist[23] is unknowable in its entirety. However, prediction and decisions are obviously possible. Probability theory is a rational means of dealing with an uncertain world.

Probabilities are numbers associated with events that range from zero to one (0-1). A probability of zero means that the event is impossible. For example, if I were to flip a coin, the probability of a leg is zero, due to the fact that a coin may have a head or tail, but not a leg. Given a probability of one, however, the event is certain. For example, if I flip a coin the probability of heads, tails, or an edge is one, because the coin must take one of these possibilities.

In real life, most events have probabilities between these two extremes. For instance, the probability of rain tonight is 0.40; tomorrow night the probability is 0.10. Thus it can be said that rain is more likely tonight than tomorrow.

The 'odds' of an event

Probability of an event happening / the probability of it *not* happening. If the probability of rain tonight is 0.3, the odds of rain tonight are $0.3 / (1 - 0.3)$ or $0.3/0.7$.

“The meaning of the term probability depends upon one’s philosophical orientation. In the CLASSICAL approach, probabilities refer to the relative frequency of an event, given the experiment was repeated an infinite number of times. For example, the .40 probability of rain tonight means that if the exact conditions of this evening were repeated an infinite number of times, it would rain 40% of the time.

In the SUBJECTIVE approach, however, the term probability refers to a “degree of belief/ confidence.” That is, the individual assigning the number .40 to the probability of rain tonight believes that, on a scale from 0 to 1, the likelihood of rain is .40. This leads to a branch of statistics called “Bayesian statistics.” [1] This has led to the term **confidence intervals** for the zones in the normal curve (i.e. that a 95% of values lie within 2 standard deviation means that I am 95% confident any value in this interval belongs to the population)

— David W. Stockburger
Introductory Statistics: Concepts Models and Applications

The probability of an event is the expected number of events among all possible events.

Some basic rules

1. The probability of nothing happening is 0
2. The probability of *something* happening is 1.
3. The probability of two independent events $p(A \cap B) = p(A)p(B)$
4. The probability of *either* of two mutually exclusive events $p(A \cup B) = p(A) + p(B)$
 - a. If however, the events are *not* mutually exclusive, $p(A \cup B) = p(A) + p(B) - p(A \cap B)$
5. For opposite events, $p(X) = 1 - p(\neg X)$
6. If an event A implied that some other event B has already occurred, then $p(A) < p(B)$
 - a. $p(A | B) = \frac{p(A \cap B)}{p(B)}$

The probability mass function (PMF) for discrete variables (X can have only certain values)

For a variable X, the PMF takes one of the possible values 'x' and returns the probability $p(X = x)$. For a coin, $p(X=\text{'heads'})$ is 0.5; and thus for a die, $p(X=1)$ is 1/6.

1. $p(X)$ must always be 0
2. The sum of $p(X=x)$ for all possible x must be 1

The probability density function (PDF) for continuous variables

The function accepts a *range* of [a,b] rather than a single value, and returns a probability $p(X \text{ in range } [a,b])$

1. For all x, $p(X \text{ in } [a,b])$ must be more than 0
2. The area under the function must be 1

Probability distributions and Models

Models are the necessary abstractions to make any sense of an inherently uncertain world. Models are generated from the *belief* that a few algebraic equations underlie the endless variability observed in the real world. **Probability distributions** is the name given to a table of *all* possible values of a variable X and their probability, and not just the ones observed. Because probability distributions are all inclusive, the sum of all probabilities (and thus, the area under the graph), is 1.

Every probability distribution has two characteristics

Expectance

The **expectance** $E(X)$, also called the *theoretical mean* (μ), or the most expected value X should take; it is not always the commonest, or even possible, value of X; it is the $\sum x \cdot p(X = x)$, where x is any one value that X can have. It represents the center of mass of a physical object, if probabilities are thought of solid bar graphs.

In the general case of a discrete distribution,

$$E(f(X)) = \sum x \cdot p(X = x)$$

In a *continuous* distribution, where the probability $P(X=x)$ is given by a function $f(x)$

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

In the special case where all values of x are equally likely, and x can have only n definite values, i.e. the probability of any value is $1/n$:

$$E(X) = \sum_{x=1}^n x \cdot \frac{1}{n}$$

This is what we mean when we say 'mean', although it is actually the arithmetic average.

Variance

The variance of a random variable X is the __expected value of the squared deviation from the theoretical mean

$$\sigma^2 = E[(X - \mu)^2]$$

i.e. if i took all the possible values of X, and subtracted the theoretical mean, these differences will be both negative and positive, and might cancel each other out. However, if I square them, then they will form a distribution among themselves, and the variance is the theoretical mean of the *distribution of squared differences*. (Similar to moment of inertia in physics)

The variance (σ^2), in a *discrete* distribution is the square notation of the average deviation from the mean

$$\sigma^2 = \sum (X - \mu)^2 = \sum (x - \mu)^2 \cdot p(X = x)$$

which can be shown to be the same as

$$\sigma^2 = E(X^2) - [E(X)]^2 = E(X^2) - \mu^2$$

The variance of a *continuous* distribution of the function f(x) is

$$\sigma^2 = \int (x - \mu)^2 f(x) dx$$

Example 3. Variance of a die roll

For a 6 sided die, the expectance (theoretical mean) $E(X) = 1x\frac{1}{6} + 2x\frac{1}{6} \dots = 3.5$. Now, $E(X^2)$ will simply be

$$1^2x\frac{1}{6} + 2^2 + \frac{1}{6} \dots 6^2x\frac{1}{6} = 15.17$$

$$\sigma^2 = E(X^2) - E(X)^2 = 2.92$$

Example 4. Variance of a biased coin

If probability of heads is p

$$E(X) = 1 \cdot p + 0 \cdot (1 - p) = p$$

and

$$E(X^2) = 1^2p + 0^2(1 - p) = p$$

thus

$$\sigma^2 = E(X^2) - E(X)^2 = p - p^2 = p(1 - p)$$

Variance of a population

For a population of size of size η , Variance

$$\begin{aligned} \sigma^2 &= E(X^2) - E(X)^2 = \frac{\sum X^2}{\eta} - \mu^2 = \frac{\sum X^2}{\eta} - 2\mu^2 + \mu^2 \\ &= \frac{\sum X^2}{\eta} - 2\mu \frac{\sum X}{\eta} + \mu^2 = \frac{\sum X^2 - 2\mu \sum X + \mu^2}{\eta} = \frac{\sum (X - \mu)^2}{\eta} \end{aligned}$$

This is usually the form we write variance of a population.

Sample variance

For sample variance, use n - 1 as denominator (see Bessel's correction). Like sample means, *the sample variances also form a normal distribution around the population variance*.

Discrete probability distributions: X can have only a finite number of values

Bernoulli distribution

Let's say that a certain event has only *two* outcomes, such as tossing a coin. It has two possible outcomes head (H) and tail (T). Let θ be the probability of head, $1 - \theta$ be the probability of tails (for a *fair* coin, both are 1/2). If I toss the coin only **once**, then what is the probability of the event X which denotes heads? Obviously, its just p . For any out come x

$$p(X = x) = \theta^x (1 - \theta)^{1-x}$$

- The only 'expected' value of X is p, which is the E(X) or the theoretical mean
- The variance (σ^2) is $E(X^2) - [E(X)]^2 = p - p^2 = p \cdot (1 - p) = pq$.

Example 5. Calculating the variance of the Bernoulli distribution by hand

The variance, as we know, is defined as the sum of (all value of X - theoretical mean)² * p(X), i.e.

$$\sigma^2 = \sum (X - E(X))^2 \cdot p(X) \text{ i.e. } = (1 - p)^2 \cdot p + (0 - p)^2 \cdot (1 - p) = p(1 - p)$$

Binomial distribution

Now we do the tossing twice. The possible number of outcomes now become four (HH, HT, TH, TT). What is the probability of each of these occurrences? Remember that the probabilities of two independent events get multiplied if they are to occur together. The outcomes of the two tosses are independent (i.e. the results of the first toss does, in no way, affect the second), so probabilities must be multiplied.

Outcome	Probability
HH	$p \times p = p^2$
HT	$p \times q = pq$
TH	$q \times p = pq$
TT	$q \times q = q^2$

Looking at the table, if we consider now, the possible values of the 'head' event as x after two tosses, x can have values 0, 1, 2 and 3. What are the probabilities of these values?

Table 2. If we toss a coin twice, what is the probability distribution of heads?

Value of X	Probability
0	q^2
1	$2pq$
2	p^2

Example 6. Finding the theoretical mean and variance of tossing a coin twice

The theoretical mean in this table (i.e. the expected number of 'heads' after two coin tosses) is $2p$. The variance is the *theoretical mean of squared differences from mean* of doing two coin tosses, i.e. $\sum (x - \mu)^2 \cdot p(X = x)$

$(0 - 2p)^2 \cdot q^2 + (1 - 2p)^2 \cdot 2p \cdot q + (2 - 2p)^2 \cdot p^2$ substituting $q = 1 - p$

$(0 - 2p)^2 \cdot (1 - p)^2 + (1 - 2p)^2 \cdot 2p \cdot (1 - p) + (2 - 2p)^2 \cdot p^2$ which turns out to be $= 2p(1 - p)$

In general, for n tosses, theoretical mean $= np$ and variance $= np(1 - p)$

The set of possibilities of all possible outcomes of an event (such as 'heads' in a coin toss), is the *binomial distribution*. When tossing only twice ($n = 2$), the set of probabilities of an event 'heads' is $[p^2, 2pq, q^2]$. For tossing n times, we can generalise the formula by mathematical induction; it would be the elements in the expansion of

$$(p + q)^n$$

which results in the set (following Newton's binomial theorem; nC_r denotes possible combinations of r things among n slots)

$$[p^n, {}^nC_1 \cdot p^{n-1} \cdot q, {}^nC_2 \cdot p^{n-2} \cdot q^2, \dots, {}^nC_r \cdot p^{n-r} \cdot q^r, \dots, q^n]$$

or, to put it concisely, the probability of X taking a value x ($x < n$) is

$$P(X = x) = {}^nC_x \times p^x \cdot q^{n-x}$$

where r is number of times we get a 'head' out of those n number of tosses.

Suppose the coin is biased, so that $p = 0.6$ and $q = 0.4$. Then, the probability distribution of variable X (incidence of 'heads') becomes

x	0	1	2
p(x)	$q^2 = 0.16$	$2pq = 0.48$	$p^2 = 0.36$

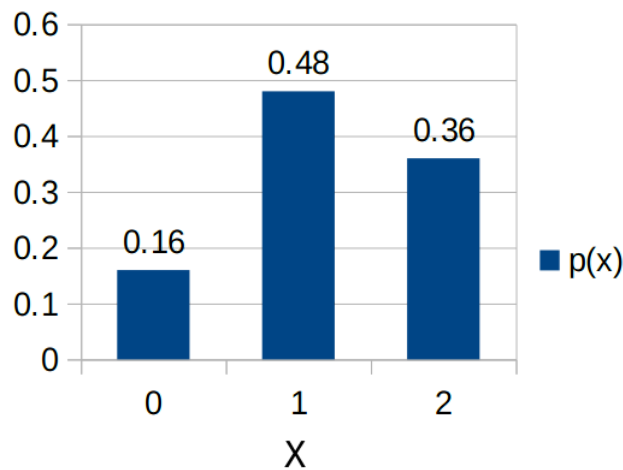


Figure 5. Plot of probability distribution of a biased coin

Expectance (theoretical mean) of the binomial distribution.

The expectance, in this case, is

$$\mu = E(X) = \sum x \cdot p(x) = 1.2$$

which, if you notice, is *not* a possible value of X (reason it is called the theoretical mean). However, as n increases, the average of X tends to E(X).

$$\sigma^2 = E(X^2) - [E(X)]^2 = 0.48$$

When n is large enough (> 25), and p is close to 0.5, then the **binomial distribution begins to look like normal distribution**, so that a z score might (and a subsequent p value) be calculated, with a mean np and SD of \sqrt{npq}

Confidence interval of a binomial estimate

Suppose a random trial of a drug of 100 (n) people show 56 (\hat{p}) cures . Is it evidence enough it is an effective drug than placebo (i.e. cures more than 50% of the time? We know for a binomial distribution (cure or no cure), the variance $\sigma^2 = p(1 - p)$, where p is the probability of cures. Thus the confidence interval for 'cure' is

$$\hat{p} \pm \sqrt{\frac{p(1-p)}{n}}$$

If p = 0.5, then this becomes $\hat{p} \pm \sqrt{0.5 \cdot \frac{0.5}{n}} = 0.56 \pm \frac{1}{\sqrt{100}} = [0.46, 0.56]$, which is the confidence of interval of cures. Note that this includes 0.50, thus the drug might not be very effective.

A binomial distribution can be simulated in most programming languages

```
#Python
numpy.random.binomial(10, 0.4) # Flip a (biased) coin 10 times whose probability of heads is 0.4
>>> 4 # i.e. number of times 'heads' turned out
```

```
#R
binom.test(10,20,0.4) # Gives probability of 10 'heads' in 20 flips, with a biased coin with 'head' probability 0.4
```

The sampling distribution of a proportion

Let's say we want to find the proportion of smokers in the world (ρ) (note that smoking is a categorical, binomial variable). The distribution of proportions of smokers in these samples is the sampling distribution.

Central limit theorem for proportions

If you take a whole lot of samples from the population and find the proportion of smokers (c) in each sample, you will find that these proportions ($p_1, p_2, p_3 \dots p_n$) form a normal distribution around the original ρ , i.e. the mean($p_1, p_2, p_3 \dots p_n$) $\sim \rho$.

The standard deviation of this normal distribution of sample proportions, is called **standard error of the proportion**, and will be equal to

$$s_p = \sqrt{\frac{\rho(1-\rho)}{n}}$$

For this theorem to work 1. the samples should be independent 2. each category should be represented in each sample (i.e. you can't have a sample of only smokers)

Example 7. Calculating estimates from a proportion

The proportion of smokers in a population is 0.9. What is the probability that in a sample of 200, >95% will be smokers?

The Z-score of the sample proportion, in this case

$$Z = \frac{p - \rho}{s_p} = \frac{p - \rho}{\sqrt{\frac{\rho(1-\rho)}{n}}} = \frac{0.95 - 0.9}{\sqrt{0.9 \cdot \frac{0.1}{200}}} = 2.35$$

Which translates to an area to the right of $Z = 0.009$.

But wait! This is a binomial distribution, so we can also calculate this probability as ($r = 95\%$ of $200 = 190$)

$$\sum_{i=r}^n \{C_i - n\} \rho^i (1 - \rho)^{n-i}$$

Which can be calculated in R language as `sum(dbinom(190:200, 200, 0.9))` = 0.008, and in python:

```
np.sum(stats.binom.pmf(k=np.arange(190,200), n=200, p=0.9))
```

Note that this won't match exactly with the Z-test, as the binomial becomes close to normal only at high sample sizes, and when both 'heads' and 'tails' are represented in the sample in numbers > 10 (i.e. at least 10 'heads' and at least 10 'tails').

Calculating confidence interval of a proportion

The confidence interval for ρ can be calculated similarly: find a Z-score for selected α , i.e. for one sided $\alpha = 0.05$, $Z = 1.64$. Then

$$\rho = p \pm Z s_p = 0.8 \pm 1.64 \left\{ \sqrt{0.8 \frac{1-0.8}{200}} \right\}$$

So if you get sample proportion 0.8 out of a sample of 200, the confidence interval will be [0.69, 0.90]

Sample size for assesment of proportion

The margin or error

$$d = Z s_p = Z \sqrt{p \frac{1-p}{n}}$$

thus

$$n = \frac{Z^2 p(1-p)}{d^2}$$

When n tends to infinite, and p is very small, the binomial distribution becomes the Poisson distribution

For rare events, the Binomial formula can be rewritten as

$$P(X = x) = {}^n C_x \cdot p^x \cdot q^{n-x} = \frac{n!}{x!(n-x)!} \cdot p^x \cdot (1-p)^{n-x}$$

Substituting $\lambda = np$ and limiting $n \rightarrow \infty$, this can be proved to be

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

where λ is the 'expectance' of the distribution, equal to np . You can then calculate a z score similar to binomial distribution.

Continuous probability distributions: X can have any real value within an interval

In such distributions, X can have any real value, so that $P(X=x)$ for any real x is not defined (which is just a vertical line with no area), but $P(a < X < b)$ is defined (which is an area), i.e. probabilities come only for an *area*. Usually, $p(X=x)$ is defined not by discrete values but a function $f(x)$.

The normal distribution

If we could go on forever, and collect shoe size data of an *infinite* number of students, and if shoe sizes would vary *continuously* (that is, the class intervals, would be adjacent - there was no discrete jumps from size 6 to 6.5 but 6.1, 6.11 and so on), we would produce a normal (Gaussian) curve. It is a curve which has been sketched after infinite number of observations and with no gaps in between class intervals. It is a bell shaped, symmetrical curve with absolute continuity.

The equation of the normal curve is

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

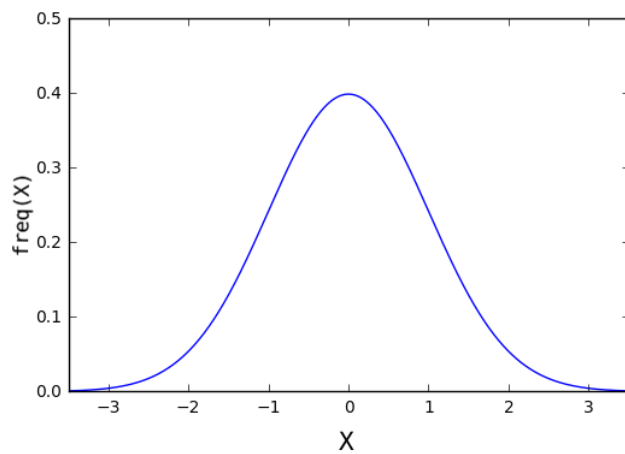


Figure 6. The normal curve

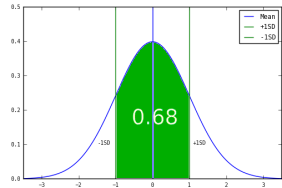
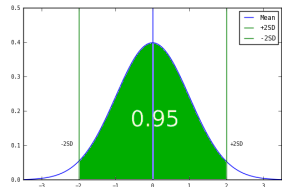
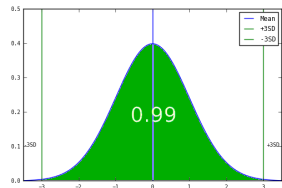
The *expectance* or **mean** of a normal distribution $E(X)$ is the same as its median and mode.

The standard normal curve

The standard normal curve is a member of the family of normal curves with $\mu = 0.0$ and $\sigma = 1.0$. The value of 0.0 was selected because the normal curve is symmetrical around μ and the number system is symmetrical around 0.0. The value of 1.0 for σ is simply a unit value. The X-axis on a standard normal curve is often re labelled with multiples of σ and called Z scores (see later).

There are three areas on a standard normal curve that should become second nature.

Table 3. Three areas in the normal curve, denoted by distance from mean in terms of standard deviation: $z = (x - \mu) / \sigma$

Z score	Area inside this Z score (one sided)	Area inside this Z score (two sided)	
1	0.34	0.68	
2 (approx) ^[2]	0.475	0.95	
3	0.495	0.99	

For the rest of this chapter, we will work with a population of children in a class of 101; we will assume a normal distribution of shoe sizea with mean $\mu = 7.8$ and $\sigma = 1.19$.

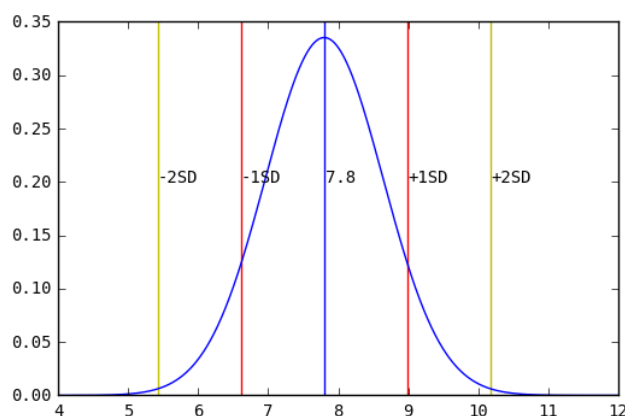


Figure 7. The normal distribution of shoe sizes

Conversion of a normal curve to a standard normal curve

Suppose a variable X has a mean μ and a standard deviation σ , and is normally distributed. It is easier, if we want to emphasise the variation rather than the actual values, to plot it in a standard normal curve by making the mean 0 and standard deviation 1.

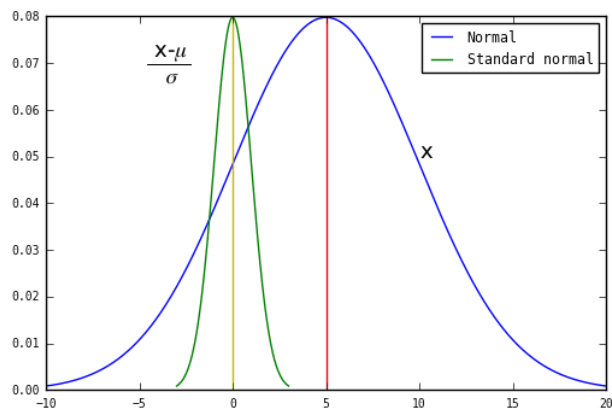


Figure 8. Making a standard normal curve from a normal curve

Think of a **relative deviate** Z which is the reduced form of X for a standard normal curve. For any value $X=x$

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

Now plot the curve of z . The curve of z is shifted parallel to x and is reduced in size, but the fractional areas under confidence intervals remain the same. In fact, we can prepare a table (by integrating the normal distribution function) to find the areas before and after a particular z -score. First, we find the area to the *left* of a particular score.

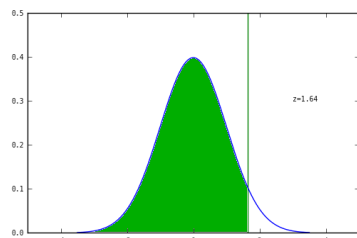


Table 4. The complete z table (area to the *left* of Z); to find a value $z = 1.64$, go to row 1.6 and column 0.04

z	+0.00	+0.01	+0.02	+0.03	+0.04	+0.05	+0.06	+0.07	+0.08	+0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55966	0.56360	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189

1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900

Next, the area to the *right* of a z score

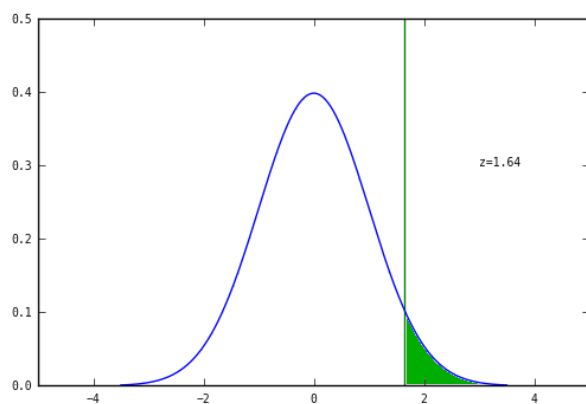


Table 5. The complete z table (area to the right of z)

z	+0.00	+0.01	+0.02	+0.03	+0.04	+0.05	+0.06	+0.07	+0.08	+0.09
0.0	0.50000	0.49601	0.49202	0.48803	0.48405	0.48006	0.47608	0.47210	0.46812	0.46414
0.1	0.46017	0.45620	0.45224	0.44828	0.44433	0.44034	0.43640	0.43251	0.42858	0.42465
0.2	0.42074	0.41683	0.41294	0.40905	0.40517	0.40129	0.39743	0.39358	0.38974	0.38591
0.3	0.38209	0.37828	0.37448	0.37070	0.36693	0.36317	0.35942	0.35569	0.35197	0.34827
0.4	0.34458	0.34090	0.33724	0.33360	0.32997	0.32636	0.32276	0.31918	0.31561	0.31207
0.5	0.30854	0.30503	0.30153	0.29806	0.29460	0.29116	0.28774	0.28434	0.28096	0.27760
0.6	0.27425	0.27093	0.26763	0.26435	0.26109	0.25785	0.25463	0.25143	0.24825	0.24510
0.7	0.24196	0.23885	0.23576	0.23270	0.22965	0.22663	0.22363	0.22065	0.21770	0.21476
0.8	0.21186	0.20897	0.20611	0.20327	0.20045	0.19766	0.19489	0.19215	0.18943	0.18673

0.9	0.18406	0.18141	0.17879	0.17619	0.17361	0.17106	0.16853	0.16602	0.16354	0.16109
1.0	0.15866	0.15625	0.15386	0.15151	0.14917	0.14686	0.14457	0.14231	0.14007	0.13786
1.1	0.13567	0.13350	0.13136	0.12924	0.12714	0.12507	0.12302	0.12100	0.11900	0.11702
1.2	0.11507	0.11314	0.11123	0.10935	0.10749	0.10565	0.10383	0.10204	0.10027	0.09853
1.3	0.09680	0.09510	0.09342	0.09176	0.09012	0.08851	0.08692	0.08534	0.08379	0.08226
1.4	0.08076	0.07927	0.07780	0.07636	0.07493	0.07353	0.07215	0.07078	0.06944	0.06811
1.5	0.06681	0.06552	0.06426	0.06301	0.06178	0.06057	0.05938	0.05821	0.05705	0.05592
1.6	0.05480	0.05370	0.05262	0.05155	0.05050	0.04947	0.04846	0.04746	0.04648	0.04551
1.7	0.04457	0.04363	0.04272	0.04182	0.04093	0.04006	0.03920	0.03836	0.03754	0.03673
1.8	0.03593	0.03515	0.03438	0.03362	0.03288	0.03216	0.03144	0.03074	0.03005	0.02938
1.9	0.02872	0.02807	0.02743	0.02680	0.02619	0.02559	0.02500	0.02442	0.02385	0.02330
2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101
2.3	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939	0.00914	0.00889	0.00866	0.00842
2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639
2.5	0.00621	0.00604	0.00587	0.00570	0.00554	0.00539	0.00523	0.00508	0.00494	0.00480
2.6	0.00466	0.00453	0.00440	0.00427	0.00415	0.00402	0.00391	0.00379	0.00368	0.00357
2.7	0.00347	0.00336	0.00326	0.00317	0.00307	0.00298	0.00289	0.00280	0.00272	0.00264
2.8	0.00256	0.00248	0.00240	0.00233	0.00226	0.00219	0.00212	0.00205	0.00199	0.00193
2.9	0.00187	0.00181	0.00175	0.00169	0.00164	0.00159	0.00154	0.00149	0.00144	0.00139
3.0	0.00135	0.00131	0.00126	0.00122	0.00118	0.00114	0.00111	0.00107	0.00104	0.00100

See the [Z score calculator](#)

```
#R
pnorm(1.96, lower.tail=FALSE) # prints 0.025, which is the area right of Z=1.96
pnorm(1.96, lower.tail=TRUE) # prints 0.975, which is the area left of Z=1.96
```

```
#Python
1 - stats.norm.cdf(1.96) # prints 0.025, which is the area to the right of Z=1.96
```

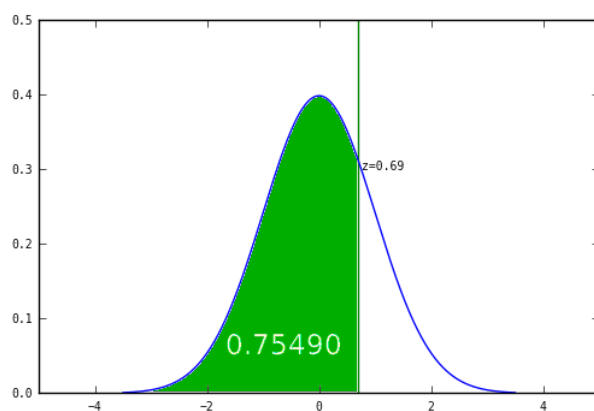


Figure 9. For example a Z score of 0.69 gives this area 0.75490 to the left of Z

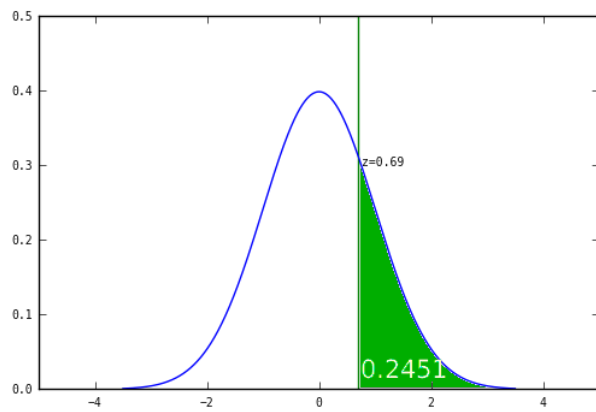


Figure 10. The same Z score gives an area 0.24510 to the right of Z

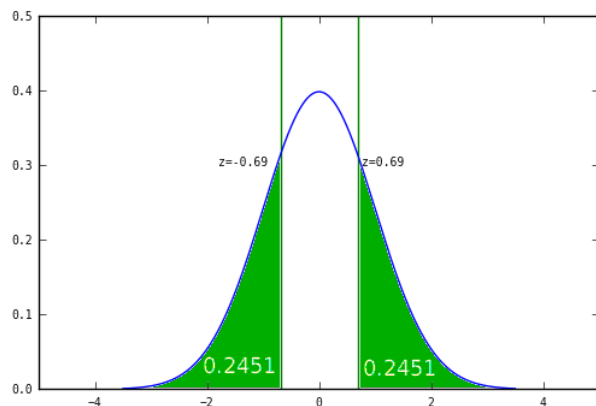


Figure 11. This is the same area to the left of z = -0.69

Skewed distributions

- positively skewed - mode < median < mean
- negatively skewed - mode > median > mean

Remember that *mean* is always dragged towards the tail of the distribution.

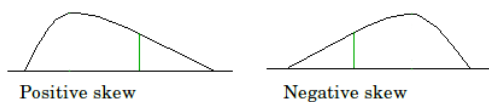


Figure 12. Skew

Probability and the normal curve

Each particular area bound between any two z-scores (any two vertical lines over the x axis) is associated with a certain degree of probability, and all such areas are called **confidence intervals**. In continuum with our example of shoe sizes, what of those students who have a size of 10.18 or more? Could they be considered acromegalics, or perfectly 'normal' people? If we introduce a child with a shoe size of 10.18, he would fall more than two standard deviations away from the mean (7.8) - look at the [normal distribution of shoe sizes](#) again. The area inside two standard deviations in a normal curve, as I have already illustrated, is 0.95. Thus the area outside 2 standard deviations is obviously, 0.05. This means that the person with a shoe size of 10.18 has a 0.05 probability (or 5%) chance of being 'normal' (or to put it more theoretically, a 5% chance of belonging to this class).

To be concise, the area of any interval of the normal curve, indicates the probability of a member from that interval being selected in a random sample. (And vice versa, i.e. if a random sample shows a member from this interval, the sample has only so much probability of actually belonging to the population).

$$\sigma^2 = \frac{\sum (X - \mu)^2}{n}$$

where n is the size of the population (see definition of variance earlier for deduction).

Standard deviation

The positive squareroot of variance (σ); the 'unit of confidence' in a normal distribution. It is a convenient method to designate distance of a member from the mean.

The central limit theorem

The central limit theorem is really at the heart of inferential statistics.

Suppose the mean of a population is μ . Each sample from this population has a different mean (\bar{X}) value. If adequate number of samples (in fact, an infinite number of them) are collected, each of a significant size, then the sample means \bar{X} - are found to be

normally distributed around the universal mean μ .

The distribution of sample means is called the **sampling distribution**. If we take any number of samples (each containing n members), then each sample will have its own mean \bar{X} . It is perfectly possible that we choose a sample from the lower end of the population, i.e. only the shoe sizes between 6-8, so that we get a sample mean = 7. It is also possible that we get, in our sample, only the largest shoe sizes, giving a sample mean = 11. But if we go on taking random samples, chances are that we would encounter every shoe size in most of our samples. So the means of our samples could be any or all of 6, 6.5, 8, 8.5, 11.5 or any plausible value. If we go on taking infinite number of samples, we would, obviously, get an infinite number of sample means. The mean of these infinite number of means, as we will find out, will be equal or very close to the universal mean μ , and the sample means themselves will form a normal curve. This *normal distribution of the sample means* will, again, have the universal mean μ as its mean. This is the central limit theorem.

$$\bar{x} \sim N(\mu, \sigma_{\bar{x}})$$

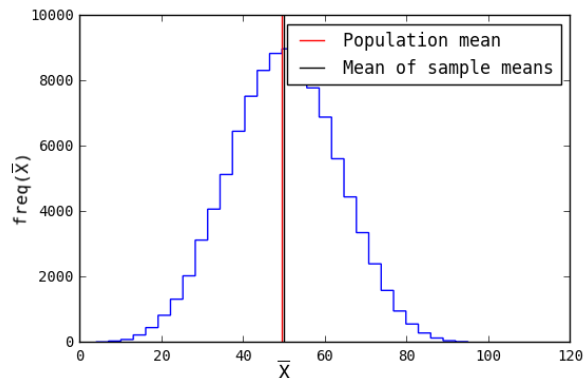


Figure 13. Central limit theorem; as number of sample increase, the mean of sample means form a normal distribution around population mean

For the central limit theorem to work

- the samples must be random/ independent
- if sampling without replacement, sample size must be $< 10\%$ of population
- the population should be normal; if *skewed*, ensure sample size is at least > 30

Standard error of the mean

The standard deviation of this new normal distribution is called **standard error of the mean** (SEM) and has the value

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

where n is the sample size. Note that the SEM no longer talks about the population variance; instead, it *shows the variability in the means of samples (of size n) drawn from the population*. It is just common sense, that the variance of the sampling distribution (sample means) will be lesser than the original variance of population. In fact, the variance will reduce proportionately with sample size (i.e. as you get better coverage of population with each sample, your samples begin to resemble each other more).

The standard error of the mean provides a method to determine how much a sample of size n , drawn from a population η , is representative of the population. The z score, of a sample then, becomes

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}}$$

Example 8. Estimating sample probabilities with Z distribution

To continue with our shoe size example, what is the probability that a sample of 25 students will have a mean 8.5 or more? Here,

- $\bar{x} = 8.5$
- $n = 25$
- $\sigma = 1.19$
- $\mu = 7.85$
- $\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{1.19}{5} = 0.238$

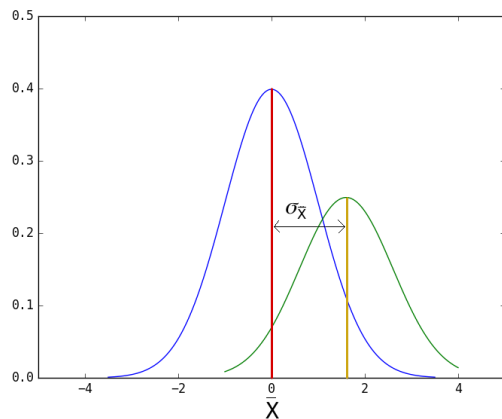
As per the central limit theorem, \bar{x} forms a normal distribution ; thus

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}} = \frac{8.5 - 7.85}{0.238} = 2.73$$

The Z score of 2.73 corresponds to an area 0.003 to the right of Z, which is the probability that a sample of 25 students will have a mean of 8.5 or more.

Finding the population mean μ from a sample mean \bar{X}

Now that we have established that sample means form a normal distribution around the population mean, let's ponder what that means. **95% of the time, the sample mean will be within 2 standard errors of the population mean.** The converse is also true, i.e. **the population mean, will be within 2 standard error from the sample mean, about 95% of the time**, i.e. in 95 out of 100 samples.^[3]



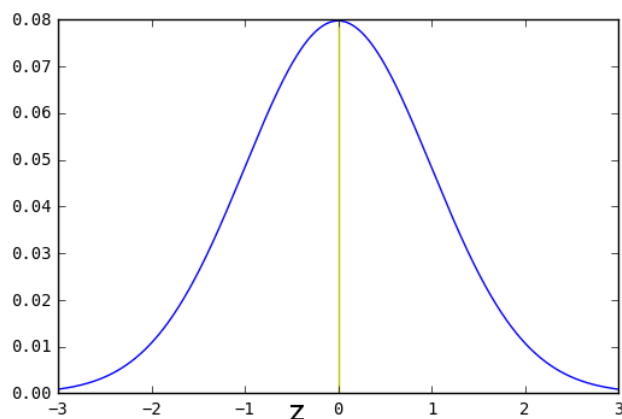
So, it can be said with 95% 'confidence' that the population mean lies within 2 standard errors of sample mean. In general, for any z score, the **confidence limit** of the population mean is $(\bar{X} \pm z \cdot \sigma_{\bar{X}})$, i.e. we can say that population mean lies within $\bar{X} - z \cdot \sigma_{\bar{X}}$ and $\bar{X} + z \cdot \sigma_{\bar{X}}$. The factor $z \cdot \sigma_{\bar{X}}$ is also called the **margin of error** (d).

To reduce the margin of error for a given z , (i.e. to narrow down the range of confidence limit), the only way is to reduce $\sigma_{\bar{X}}$, which, if you will recall, is just $\frac{\sigma}{\sqrt{n}}$. Since, σ is not accessible, the only way to bring this down is to increase n , the sample size. Because it is *square root of n* at work, to halve the standard error, the sample size must be increased *fourfold*.

The **precision** of a sample in determining the population mean, is proportional to sample size (i.e., it reflects the width of the confidence limit). The wider a confidence limit, the less precise it is. **Accuracy** of a sample denotes the closeness of the sample mean to the population mean.

If $Z_1, Z_2 \dots Z_n$ are all *standard normal* variables, then their sum of squares has a χ^2 distribution

Consider a standard normal distribution



What is the distribution of Z^2 then? Consider all values of Z , their frequency, and try to find how Z^2 matches up. Obviously, $Z=0$ is the most frequent, followed by $Z = -1$ and $+1$ (with everything in between) and so on. Thus the probability distribution of Z^2 becomes a chi square distribution with 1 degree of freedom (i.e. because only one variable has been used in construction)

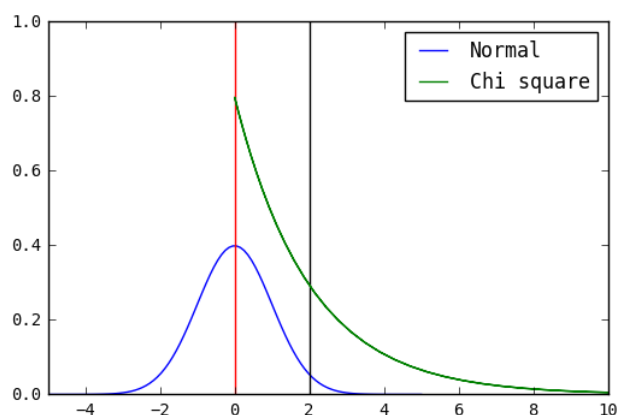


Figure 14. Chi square distribution

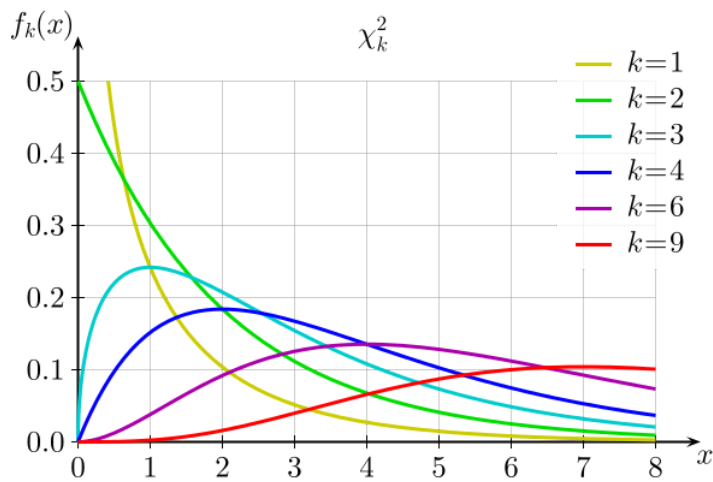


Figure 15. Chi square distribution over multiple degrees of freedom [source: Wikipedia]

In general, if k random variables $Z_1, Z_2, Z_3 \dots Z_k$ have *standard* normal distribution, then the sum $\sum_{i=1}^k Z_i^2$ has a χ^2 (pronounced 'chi squared') distribution with a single parameter, the **degree of freedom** = k . The distribution has a mean k .

```
np.random.chisquare(5, size=10000) # 5 degrees of freedom, 10000 Observations
```

As you might have noticed, when sample size is sufficiently large, the binomial distribution approximates the normal distribution, with p = probability of head ('success'), n = number of tosses, m = observed number of heads ('successes'), $q = 1-p$; Lancaster, DeMoivre and Laplace showed that the variance of binomial distribution was npq , so that it could a relative deviate 'chi' can be thought of (i.e trying to formally convert binomial distribution to normal)

$$\chi = \frac{m - np}{\sqrt{npq}}$$

thus

$$\chi^2 = \frac{(m - np)^2}{npq}$$

or

$$\chi^2 = \frac{(m - np)^2}{np(1 - p)} = \frac{(m - np)^2}{np - np^2} = \frac{(m - np)^2}{np}$$

(p being a fraction, p^2 can be safely ignored)

Of course, m is the observed incidence O of a 'success' (i.e. heads), np is the 'expected' incidence theoretical mean) E , thus

$$\chi^2 = \frac{(O - E)^2}{E}$$

If more than one such discrete variable is being studied (i.e. N variables)

$$\chi^2 = \sum_{i=1}^N \frac{(O_i - E_i)^2}{E_i}$$

This is the form that the chi square statistic is usually written. However, remember that the mean of this distribution is still np and SD is \sqrt{npq} . So if the test statistics χ^2 is determined, then χ is a normally distributed variable with mean np and SD \sqrt{npq} . If χ^2 comes out to be 3.841, then χ is 1.96 (if degree of freedom is 1), and the corresponding area right of this Z score in a normal distribution is $p = 0.05$. This is how the χ^2 table has been made.

The estimated standard error and t distribution

The calculation of standard error of mean raises an obvious question: if we know the standard deviation of the population (σ) anyway, why bother about sampling at all? In fact, knowing σ is impossible; you can never survey the entire population. Hence calculation of standard error of mean ($\sigma_{\bar{X}}$) is also impossible. Rather, we have to content ourselves with **estimated** standard error, or $s_{\bar{X}}$, derived solely from the sample at hand.

$$s_{\bar{X}} = \frac{s}{\pm \sqrt{n - 1}}$$

where, s is the SD of the sample at hand, and n its size.

In calculating the population mean from a sample then, we must devise a new variable called t [4], which is the same as z score, but with *estimated* standard error

$$t = \frac{\bar{X} - \mu}{\pm s_{\bar{X}}} = \frac{\bar{X} - \mu}{\pm \frac{s}{\sqrt{n - 1}}}$$

With introduction of t , a variable derived from the sample at hand, we have committed ourselves to the mercy of the sample size.

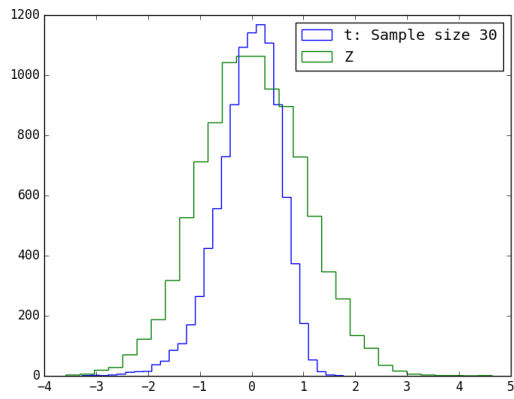


Figure 16. As sample size increases, t approaches z distribution

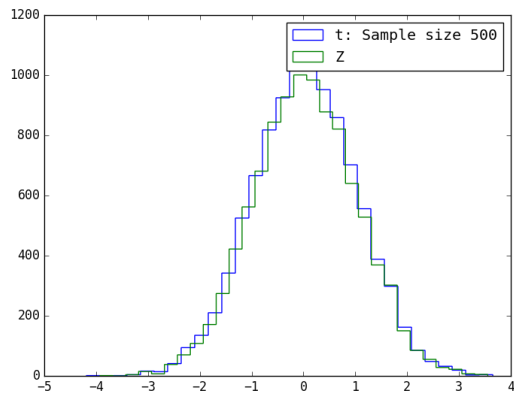


Table 6. Values of t for a given degree of freedom and α ($\alpha = 1 - \text{confidence interval}$); for t for a given degree of freedom, find a value which is just m

One-sided α												
	.25	.10	.05	.025	.01	.005	.0025	.001	.0005	.00025	.0001	.00005
Two-sided α												
	.50	.20	.10	.05	.02	.01	.005	.002	.001	.0005	.0002	.0001
df												
1	1.00	3.08	6.31	12.71	31.82	63.66	127.32	318.31	636.62	1273.24	3183.10	6366.20
2	.82	1.89	2.92	4.30	6.96	9.22	14.09	22.33	31.60	44.70	70.70	99.99
3	.76	1.64	2.35	3.18	4.54	5.84	7.45	10.21	12.92	16.33	22.20	28.00
4	.74	1.53	2.13	2.78	3.75	4.60	5.60	7.17	8.61	10.31	13.03	15.54
5	.73	1.48	2.02	2.57	3.37	4.03	4.77	5.89	6.87	7.98	9.68	11.18
6	.72	1.44	1.94	2.45	3.14	3.71	4.32	5.21	5.96	6.79	8.02	9.08
7	.71	1.42	1.90	2.37	3.00	3.50	4.03	4.79	5.41	6.08	7.06	7.88
8	.71	1.40	1.86	2.31	2.90	3.36	3.83	4.50	5.04	5.62	6.44	7.12
9	.70	1.38	1.83	2.26	2.82	3.25	3.69	4.30	4.78	5.29	6.01	6.59
10	.70	1.37	1.81	2.23	2.76	3.17	3.58	4.14	4.59	5.05	5.69	6.21
11	.70	1.36	1.80	2.20	2.72	3.11	3.50	4.03	4.44	4.86	5.45	5.92
12	.70	1.36	1.78	2.18	2.68	3.06	3.43	3.93	4.32	4.72	5.26	5.69
13	.69	1.35	1.77	2.16	2.65	3.01	3.37	3.85	4.22	4.60	5.11	5.51

14	.69	1.35	1.76	2.15	2.63	2.98	3.33	3.79	4.14	4.50	4.99	5.36
15	.69	1.34	1.75	2.13	2.60	2.95	3.29	3.73	4.07	4.42	4.88	5.24
16	.69	1.34	1.75	2.12	2.58	2.92	3.25	3.69	4.02	4.35	4.79	5.13
17	.69	1.33	1.74	2.11	2.57	2.90	3.22	3.65	3.97	4.29	4.71	5.04
18	.69	1.33	1.73	2.10	2.55	2.88	3.20	3.61	3.92	4.23	4.65	4.97
19	.69	1.33	1.73	2.09	2.54	2.86	3.17	3.58	3.88	4.19	4.59	4.90
20	.69	1.33	1.73	2.09	2.53	2.85	3.15	3.55	3.85	4.15	4.54	4.84
21	.69	1.32	1.72	2.08	2.52	2.83	3.14	3.53	3.82	4.11	4.49	4.78
22	.69	1.32	1.72	2.07	2.51	2.82	3.12	3.51	3.79	4.08	4.45	4.74
23	.68	1.32	1.71	2.07	2.50	2.81	3.10	3.49	3.77	4.05	4.42	4.69
24	.68	1.32	1.71	2.06	2.49	2.80	3.09	3.47	3.75	4.02	4.38	4.65
25	.68	1.32	1.71	2.06	2.49	2.79	3.08	3.45	3.73	4.00	4.35	4.62
26	.68	1.32	1.71	2.06	2.48	2.78	3.07	3.44	3.71	3.97	4.32	4.59
27	.68	1.31	1.70	2.05	2.47	2.77	3.06	3.42	3.69	3.95	4.30	4.56
28	.68	1.31	1.70	2.05	2.47	2.76	3.05	3.41	3.67	3.94	4.28	4.53
29	.68	1.31	1.70	2.05	2.46	2.76	3.04	3.40	3.66	3.92	4.25	4.51
30	.68	1.31	1.70	2.04	2.46	2.75	3.03	3.39	3.65	3.90	4.23	4.48
35	.68	1.31	1.69	2.03	2.44	2.72	3.00	3.34	3.59	3.84	4.15	4.39
40	.68	1.30	1.68	2.02	2.42	2.70	2.97	3.31	3.55	3.79	4.09	4.32
45	.68	1.30	1.68	2.01	2.41	2.69	2.95	3.28	3.52	3.75	4.05	4.27
50	.68	1.30	1.68	2.01	2.40	2.68	2.94	3.26	3.50	3.72	4.01	4.23
55	.68	1.30	1.67	2.00	2.40	2.67	2.93	3.25	3.48	3.70	3.99	4.20
60	.68	1.30	1.67	2.00	2.39	2.66	2.91	3.23	3.46	3.68	3.96	4.17
65	.68	1.29	1.67	2.00	2.39	2.65	2.91	3.22	3.45	3.66	3.94	4.15
70	.68	1.29	1.67	1.99	2.38	2.65	2.90	3.21	3.44	3.65	3.93	4.13
75	.68	1.29	1.67	1.99	2.38	2.64	2.89	3.20	3.43	3.64	3.91	4.11
80	.68	1.29	1.66	1.99	2.37	2.64	2.89	3.20	3.42	3.63	3.90	4.10
85	.68	1.29	1.66	1.99	2.37	2.64	2.88	3.19	3.41	3.62	3.89	4.08
90	.68	1.29	1.66	1.99	2.37	2.63	2.88	3.18	3.40	3.61	3.88	4.07
95	.68	1.29	1.66	1.99	2.37	2.63	2.87	3.18	3.40	3.60	3.87	4.06
100	.68	1.29	1.66	1.98	2.36	2.63	2.87	3.17	3.39	3.60	3.86	4.05
200	.68	1.29	1.65	1.97	2.35	2.60	2.84	3.13	3.34	3.54	3.79	3.97
500	.68	1.28	1.65	1.97	2.33	2.59	2.82	3.11	3.31	3.50	3.75	3.92
∞	.67	1.28	1.65	1.96	2.33	2.58	2.81	3.10	3.30	3.49	3.73	3.91

```
#R
pt(1.96, df=10, lower.tail = FALSE) #prints 0.03, which is area to the right of t=1.96
pt(1.96, df=500, lower.tail=FALSE) # prints 0.0252, which is almost the same area to the right of z=1.96; at high sample
```



```
sizes, t and z graphs look similar
qt(0.025,df=500) # prints -1.96, which is the t value, area to the left of which is 0.025
```

For example, if sample size is 21, degree of freedom is 20, and $t = -2.27$, then the closest value more than t (for degree of freedom=20) is 2.53, which corresponds to an $\alpha=0.02$ (two sided)

Once the sample size crosses about 100, the t and z distribution become almost the same.

An example : how to estimate population mean from a sample using t score

Suppose from 101 students of a class, we select 10 and take their shoe size. The known average shoe size of all children is 8.5. This time, we don't make the foolish assumption that we know the mean and SD of the entire class. Instead, we find the mean of the sample to be 7.6 and SD 1.11. What does it say about the sample?

- $H_0: \mu_h = 8.5$
- $H_A: \mu_h \neq 8.5$
- $\mu, \sigma: \text{unknown}$
- $\bar{x} = 7.6$
- $s_x = 1.11$
- $n = 10$

The estimated standard error

$$s_{\bar{X}} = \frac{s_x}{\sqrt{n-1}} = \frac{1.11}{3} = 0.37$$

The t score, is then

NOTE

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}}$$

Of course, we know neither σ nor μ . So we have to find t from the table, which shows areas of t distribution for a particular degree of freedom (in this case, 9). The corresponding t -value for an area of 95% is 2.22 (two-tailed). To rewrite the equations

$$\mu = \bar{X} \pm t s_{\bar{X}}$$

The 95% confidence interval will be $\bar{X} \pm t \cdot s_{\bar{X}} = 7.6 \pm 2.22 \cdot 0.37 = (6.77, 8.42)$. Thus, we are 95% sure that the mean shoe size of the population (the entire class) lies between 6.77 to 8.42.

Lets improve. We select a sample of 30, and mean 7.56, SD (s) 1.14, and estimated standard error $\frac{1.14}{\sqrt{29}} = 0.211$.

Calculating for a degree of freedom ($30-1$) = 29, t value for 95% confidence interval (two sided) is 2.05. Thus, the confidence interval

$$\bar{X} \pm t \cdot s_{\bar{X}} = 7.56 \pm 2.05 \cdot 0.211 = (7.12, 7.99)$$

With just an increase in sample size, our estimate has improved! Note that this does not include 8.5 (the hypothesised mean); so we have to reject null hypothesis.

Is this in agreement with hypothesis testing? The area outside of $t = 2.22$ ($df=9$) is 0.02, which is the p -value, and in this case, less than alpha limit. Thus it agrees with the confidence interval.

Bayesian statistics

Truth must ultimately be tested. The material sciences (physics, chemistry) usually provide us with the tools to conduct the tests, but it is statistics which tells how to interpret the test. Common to all tests is to find association between two events, i.e.

- to test whether the finding of bronchial sounds is diagnostic of pneumonia (a diagnostic test)
- to test whether smoking is associated with lung cancer (a study to discover risk factors)
- to test whether radiation is effective in lung cancer (a therapeutic test)

All tests must be

- **reproducible**, i.e. gives the same result over and over even if different samples are tested, by different observers with different skill levels
- **accurate**, i.e. yields result close to the GOLD STANDARD test, and reflects the true progression of the disease to that level which the test values suggests
- **valid**, i.e. can distinguish between a positive and a negative result (i.e. between diseased and non diseased) to a satisfactory degree.

Every diagnosis or decision is based on a test, be it the history, a symptom or sign, or some lab routines, or the history of an exposure to a risk factor. Such a diagnostic test must bear a few accreditations if it has to qualify for being used in clinical reasoning.

Reproducibility

It is the ability of a test to yield the same results over and over, irrespective of variations in basis of test, method or skill. No test is wholly reproducible. Suppose you have diagnosed a man to have pulmonary tuberculosis by examining sputum smears. A fellow physician may wholly disagree with you, because, when he did the examination, either

- patient became sputum negative (variation in the basis of the test)
- the physician used a different stain (variation of method)
- the physician made an error in spotting the AFB (variation of skill)

Because no test is wholly reproducible, the whole medical business continues to run on uncertainty.

Accuracy

A test is accurate when it yields results equal, on the average, to a GOLD STANDARD test

Suppose you devise method X to determine blood glucose. To qualify as accurate, this test has to yield results consistent with that of the Glucose Oxidase reaction.

Validity

This is the challenge: to distinguish truth from just another chance finding.

Put simply, let's say you attend some kids with complaint of cough, fever and mild dyspnea. You did some auscultation (the test) and find bronchial sounds in some of these kids, and label them as pneumonia. Next, your boss carries out a culture on their lung aspirate (the Gold Standard test) and disproves your findings. The final scenario is this.

	Diseased (Culture +ve)	Non diseased (Culture -ve)
+ve test (Bronchial sounds)	a	b
-ve test (Vesicular sound)	c	d

The **sensitivity** of the test is the probability of diseased people yielding positive result, $\frac{a}{a+c}$. The **specificity** is the reverse, the probability of healthy people to give negative result, $\frac{d}{b+d}$.

A test which is very non specific will yield too many false positive results; on the other hand, an insensitive test gives too many false negative results. The importance we attach to a positive or negative result is thus a function of the cut off value of the test; i.e. after which level we consider the results positive. Selecting a low cut off reduces the specificity of the test, and a high cut off dampens sensitivity [5]. In fact, the curve of sensitivity vs specificity looks like this

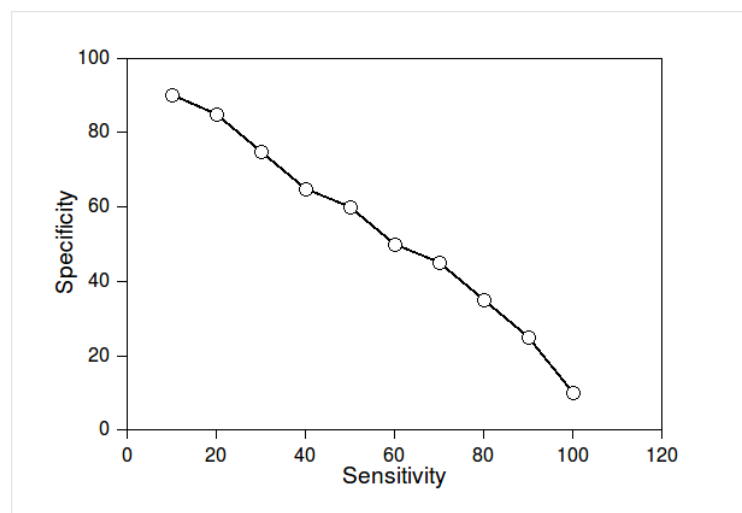


Figure 17. Sensitivity versus specificity curve

The perils of using non qualified tests are many. Too many false positive results are a burden on the health infrastructure, cause useless anxiety to the victim, and once labeled – it's hard to get rid of the stigma. Again, too many false negative results fail the entire purpose of screening.

Predictive value

This is the most important aspect of a test from a clinical viewpoint. A positive predictive value is the probability of someone testing positive actually having the disease, i.e. $\frac{a}{a+b}$. Similarly, a negative predictive value is the probability of *not* having the disease in

someone testing negative = $\frac{d}{c+d}$.

The positive predictive value is affected by the

- *prevalence* of the disease - in areas of high prevalence of some disease, tests for the disease are more valid

- *specificity* of the test

The usual challenge: determine post test probability from pre test knowledge

Bayesian statistics states that the positive predictive value is much higher if the disease in question is very prevalent. In another words, the usefulness of a test, apart from an inherent quality of itself (the sensitivity and specificity), *is also dependant upon how common the disease is*. Suppose the prevalence of a disease in a certain area is p in a population of η ; this means, the chance of any individual of that area of having the disease is p (the pretest probability). Now we do the test and get a positive result. Given the specificity and sensitivity of the test, what is the probability that the individual is truly diseased?

Table 7. Contingency table for a diagnostic test

	Diseased	Non diseased	Total
+ve test	a	b	a+b
-ve test	c	d	c+d
Total	a+c	b+d	

Now, obviously, the number of total diseased people $a + c = p\eta$ and non diseased people $b + d = (1 - p)\eta$. Again, the positive predictive value is (number of diseased people who tested +ve/ total number of people who tested +ve) = $\frac{a}{a + b}$.

Given the sensitivity of the test is sn and specificity is sp , we know that

$$sn = \frac{a}{a + c} \text{ and } sp = \frac{d}{b + d}$$

From these equations, calculate your heart out for the value of poitive preditive value $\frac{a}{a + b}$; you will find it to be

$$ppv = \frac{p \times sn}{(p \times sn) + (1 - sp)(1 - p)}$$

This is the positive predictive value of a test if the sensitivity, specificity and prevalence is given. There is, however, a more subtle way to achieve the same result. The **odds** of an individual having the disease before the test is, obviously, $\frac{p}{1 - p}$ (see definition of Odds).

Now, Bayesian statistics states that the odds (r) of having the disease after a positive test is

$$r = \frac{p}{1 - p} \times \frac{sn}{1 - sp}$$

Suppose post test probability (the same as positive predictive value) is ppv ; then by definition of odds

$$\frac{ppv}{1 - ppv} = r \text{ or } , ppv = \frac{r}{1 + r}$$

This ppv is the post test probability or the positive predictive value (do the actual calculation on a real problem and you will find the result from the two methods to be identical). The factor $\frac{sn}{1 - sp}$ is called the *likelihood ratio* of the test.

Hypothesis testing

Inferential statistics allows you to predict both the probability of an event and the amount of error of that prediction. This section is to determine that amount of error.

The GOLD STANDARD of hypothesis testing is of course, census; i.e. study the entire population.

Hypothesise first

The null hypothesis (H_0)

There are several ways to state the null hypothesis

- two events are unrelated
- two samples have similar distribution and mean
- ... and so on...

The alternate hypothesis (H_a).

It states that the two events are related, or that two populations differ. Obviously, both the hypotheses can not be true simultaneously.

Errors

Type I error (α)

The error that happens when null hypothesis is *rejected* in spite of it being true (i.e. you **wrongly diagnose an association** between reading comics and shoe size, or something equally bizzare). The probability of a type I error happening in a test is called **α limit** of the test.

The probability of finding something consistent with the H_a , given the H_0 was true, is the **p-value** (i.e. the probability of getting false positive results). If p-value is very low (specifially, less than α limit - then, in view of experimental findings agreeing with H_a , we will have

to reject H_0 .

Type II error (β)

The error that happens when the null hypothesis is accepted in spite of it being false (you miss a true association)

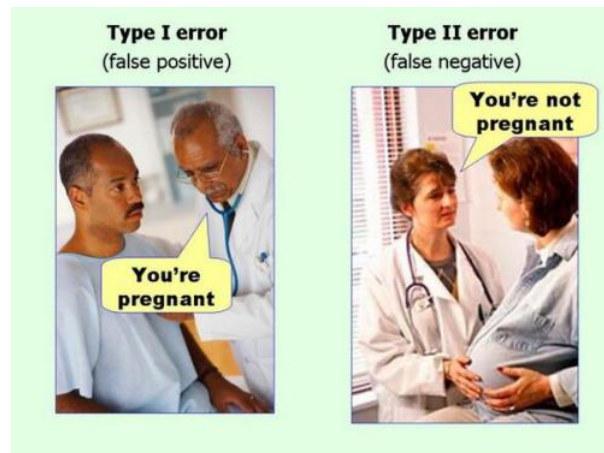


Figure 18. The two types of error (Image Courtesy: Wikipedia)

The two types of errors are inversely proportional, i.e. you can not reduce them simultaneously. *However*, it is usually accepted (both from a statistical and philosophical perspective), that α should be minimised, even if some β remains (i.e. 'innocent until proven guilty'). However, it is a matter of judgement and should be tailored to each case (i.e. while judging restaurant hygiene it is better to commit false positive errors (which will harm the restaurant), but not false negatives (which will affect the diners)).

Power of a test

The power of a test, i.e. the probability that a test detects any difference that actually exist (avoids false negatives), is $1-\beta$. It must be atleast 0.8 to qualify.

The power of a test ($1-\beta$) increases with

1. increasing α ; setting a stricter (lower) α will reduce false positives, but increases false negatives (i.e. β)
2. if the difference between sample mean and population mean increases (i.e. the sample is more extreme)
3. reducing estimated standard error (thus increasing t , pushing it more to extremes towards rejection regions); the way to reduce standard error is, of course to increase sample size

Table 8. Relation between α and β

		Reality	
		H_0 true	H_0 false
Test result	H_0 accepted	CORRECT, $1-\alpha$	β
	H_0 rejected	α	CORRECT, $1-\beta$

Example 9. How to determine power of a test

Let's exemplify these relations. Continuing with our example of shoe sizes

- mean shoe size of the class $\mu = 7.85$
- $\sigma = 1.19$.

We hypothesise (wrongly), that $\mu_H = 8.5$. Thus the alternate hypothesis H_a is that $\mu < 8.5$.

We have chosen $\alpha=0.05$, which means that area to the left of z (one sided only) is 0.05; so critical value of z becomes -1.64 (one tailed, see the Z table).

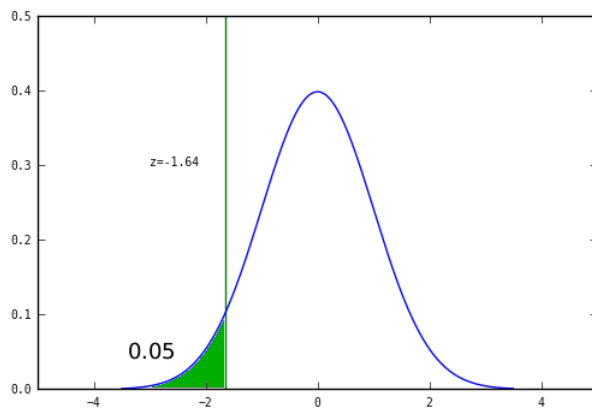


Figure 19. The area left of $z = -1.64$ is 0.05

Using central limit theorem, we find the relative deviate z of the distribution of \bar{X} (i.e. the distribution of different sample means from the same population also makes a normal distribution). Now we sample 21 students. This distribution of sample means \bar{X} , has an standard error of mean $= \frac{\sigma}{\sqrt{n}}$, where n is sample size, and its mean is going to be approximately the population mean μ_H .

$$z = \frac{\bar{X} - \mu_H}{\frac{\sigma}{\sqrt{n}}}, \therefore \bar{X} = z \cdot \frac{\sigma}{\sqrt{n}} + \mu_H = -1.64 \cdot \frac{1.19}{\sqrt{21}} + 8.5 = 8.07$$

Translating to values of \bar{X} , if the mean of a representative sample is less than 8.07, the null hypothesis is to be rejected, as it falls left to a z score of -1.64.

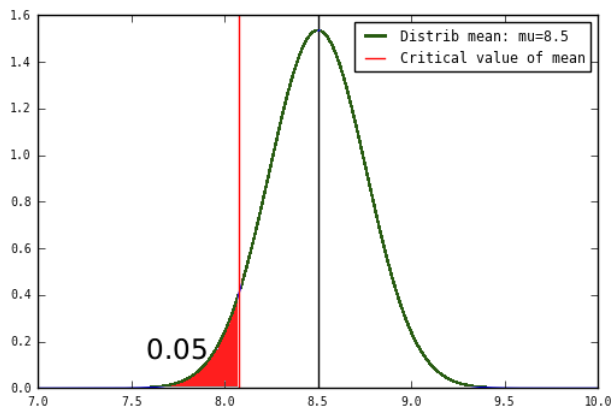


Figure 20. The probability of α (i.e. null hypothesis was rejected in spite of being true) is the area to the left of $\bar{X} = 8.07$

However, the population mean $\mu = 7.8$ (so the null hypothesis is actually false). If we are rejecting the null hypothesis only if $\bar{X} < 8.07$, the interval 7.8 - 8.07 is in the rejection zone. But in the zone $\bar{X} > 8.07$, **we will accept the null hypothesis in spite of it being false**.

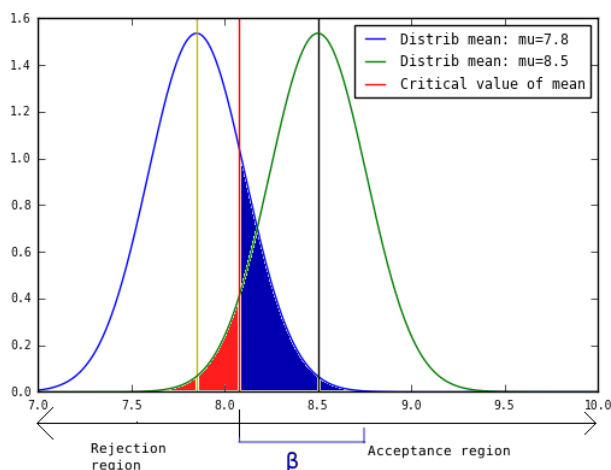


Figure 21. In the zone shaded in blue, the null hypothesis is *accepted* in spite of being false; thus its area is β

The critical value lies at $z = 0.049$ of the blue curve, giving an area 0.48 to its right, which is the value of β .

Now increasing just the sample size, we can reduce β .

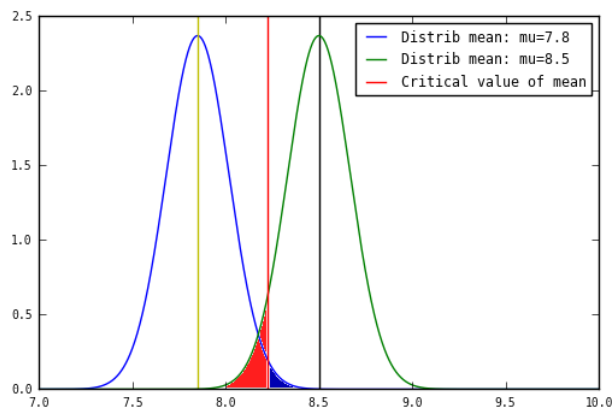


Figure 22. The plot with sample size = 50

Increasing α (i.e. making the test less stringent) will also reduce β .

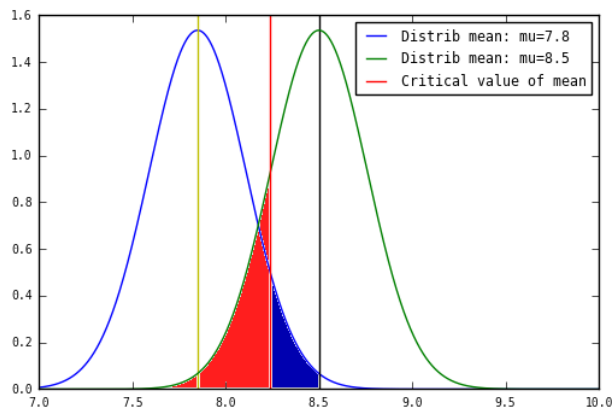


Figure 23. The same plot with $\alpha=0.15$, and corresponding $z=1.0$, reduces β

β will also reduce if the difference in mean of the population and the sample mean is greater. This difference between population and sample mean is called the **effect size**.

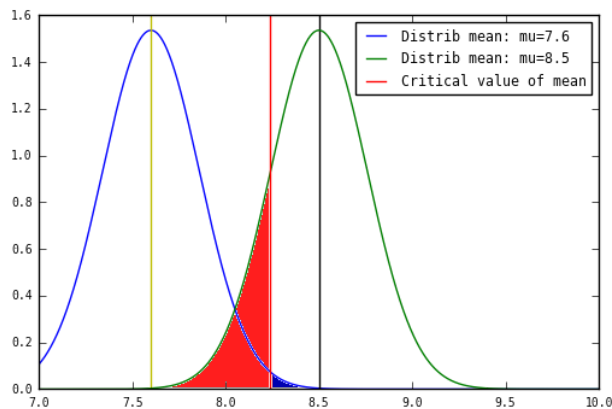


Figure 24. Plotting with population mean = 7.6

It is imperative that power be calculated before finalising the sample size.

CAUTION

Why not choose a very low α ?

A very low α makes it very difficult to reduce the null hypothesis, and β . Usually, in the fields of social sciences, we look forward to reduce α ('innocent until guilty'), even at the cost of rising β (acquitting criminals for lack of evidence).

Step two: Select significance level α

If the mean of our sample (or atleast, their *theoretical mean* or expectance) is far off from the population mean, we begin to suspect that the null hypothesis (i.e. mean of sample means = population mean, as per central limit theorem) might be false. If the probability that the sample was drawn from the population is less than a certain value α (usually 0.05) we reject the null hypothesis.

Recall that a sample which has a mean more than 2 standard deviations away from the population mean, has only 0.05 probability of belonging to the population. Thus is the the mean of our sample falls in this region, we begin to get uneasy, and tend to think that the population mean might be something else than the mean of sample means. However, that 5% chance lingers, and we can make a statement that **null hypothesis is false with 0.05 probability of α** , or simply, **null hypothesis rejected at $p<0.05$**

Note that significance level is complimentary to confidence interval, i.e. a significance level 0.05 is the same as a 95% confidence interval (two sided), and the same as 90% confidence interval (one sided) .

Directional hypothesis and rejection region

We can always for a null hypothesis that $\mu = \mu_h$, in which case, we can disprove H_0 in either direction. Lets make it a little more tangible. Suppose, out of this population, we draw a sample and the sample mean turns out to be \bar{X} .

Assuming that this sample is representative of the population, we know that sample means form a normal distribution around the population mean. Thus, this particular sample mean is a member of a normal distribution, which is located at a certain distance $\bar{X} - \mu_h$ away from the mean μ_h . But this happens if and only if the population mean is actually the hypothesised μ_h . If the population mean is anything else (supposing $\mu_h + \epsilon$), won't be normally distributed around it, but around $\mu_h + \epsilon$.

In fact, the probability that this sample actually belongs to the population is the height of the curve at \bar{X} . Samples with a mean right to this line have lesser chance (specifically, the red area beyond this line a) of belonging to this population; or to say the same in a roundabout manner, if a sample mean turns out to be at \bar{X} or right, the null hypothesis (that $\mu = \mu_h$) has a probability a of being true.

But what if our sample turned out to have a mean $-\bar{X}$? We could argue that this time, the population mean is actually $\mu_h - \epsilon$, and thus the sample with a mean $-\bar{X}$ is normally distributed around $\mu_h - \epsilon$. Thus, we could reject the null hypothesis with a probability a , where a is the area to the left of $-\bar{X}$. (Of course, the rejection regions of both side are equal, since the normal distribution is symmetrical)

To summarise, in a **two sided test**, a sample mean \bar{X} gives us two areas of rejection, to the right of \bar{X} , and to the left of $-\bar{X}$. If we have already selected an α , then the null hypothesis can be rejected only if $2a < \alpha$.

NOTE

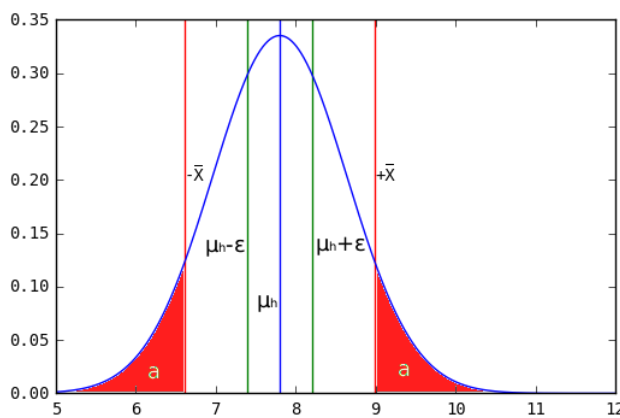


Figure 25. Two tailed rejection region: the combined area in both sides is $a + a = 2a$

Now, what if the null hypothesis was one tailed, i.e. $\mu \geq \mu_h$? In that case, $H_a: \mu < \mu_h$. Now the sample mean $-\bar{X}$ matters, because it is evidence that $\mu < \mu_h$, and can reject the null hypothesis provided it falls in the selected rejection region (defined by α). But the mean at the other end, \bar{X} , does not matter now? Why? A sample with mean \bar{X} will only prove that the population mean might be $\mu > \mu_h$, which is part of the null hypothesis. Thus it is no use in disproving the null hypothesis.

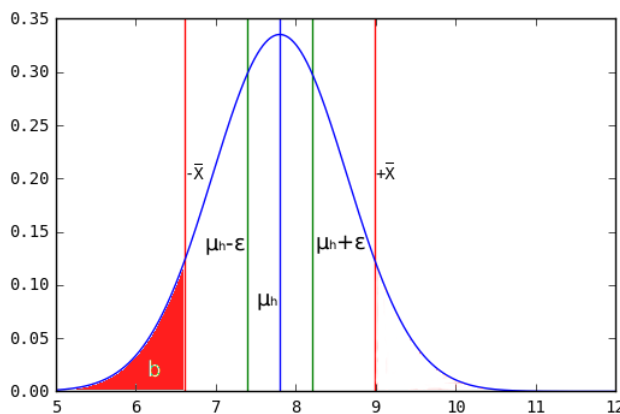


Figure 26. One tailed rejection region: the one tailed area b

If we have already selected an α , the area b alone must be less than α to reject the null hypothesis. Stated in another way, to prove that a sample mean differs significantly from the actual in one direction, it needs to be really far off. Using the same $\alpha=0.05$, the area 0.05 in one side (to the left) corresponds to a z score of -1.5, while in two tailed score it is -2. Notice that the z-value is less extreme than two tailed score, thus **one tailed tests are more powerful than two tailed ones**. But on the converse, using one tailed tests, we would completely miss any action on the other side. If we are studying a prospective antihypertensive drug, using a one sided test, we would miss any effect of the drug that might increase blood pressure.

It is imperative that after your test, you recheck with constructing a confidence interval based on your sample. **If you have rejected H_0 , but the confidence interval includes the null value (i.e. the hypothesised μ), then you might have made a false positive error.**

Step three: Select a test for statistical significance

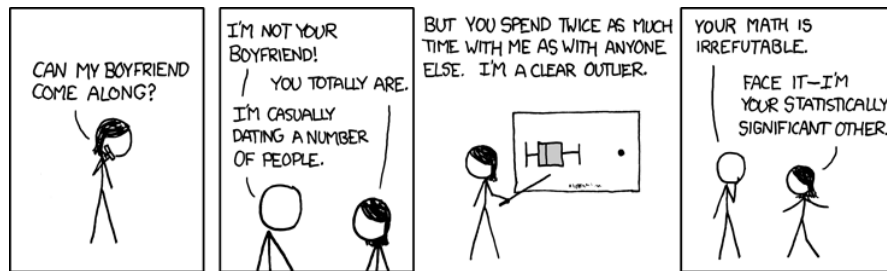


Figure 27. You can't deny mathematics[25]

The Z test

The Z test is useful only when

- the population is normally distributed
- the population mean and SD is known.
- your hypothesis is about a single mean of a population

The crux of the matter is this: there exists a population with a hypothesised mean μ (null hypothesis) and σ (known), which is normally distributed. We pick a sample of size n with a mean \bar{X} from this population, and define

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Then, if and only if null hypothesis is true (i.e. the population mean is actually μ)

1. z will have a *standard* normal distribution
2. values of z outside 2 standard errors ($z < -2\sigma_{\bar{X}}$ or $z > 2\sigma_{\bar{X}}$) will be rare; specifically, probability of such a value is 0.05

If \bar{X} translates to a z score in this **rejection range**, we reject the null hypothesis.

Example 10. Testing null hypothesis with z test

Suppose from the class of students (mean 7.8, SD 1.19), we pick a sample of 20 with mean shoe size 8.0 and sample SD = 1.03. Is this proof enough ($\alpha=0.05$) that the population mean is something else than 7.8?

Here

- $H_0 : \mu = 7.8$
- $H_a : \mu \neq 7.8$
- $\mu = 7.8$
- $\sigma = 1.19$
- $n = 20$
- $s = 1.03$
- $\bar{X} = 8.0$
- $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = 0.266$
- $\alpha = 0.05$

Thus, calculating z for the sample

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = 0.75$$

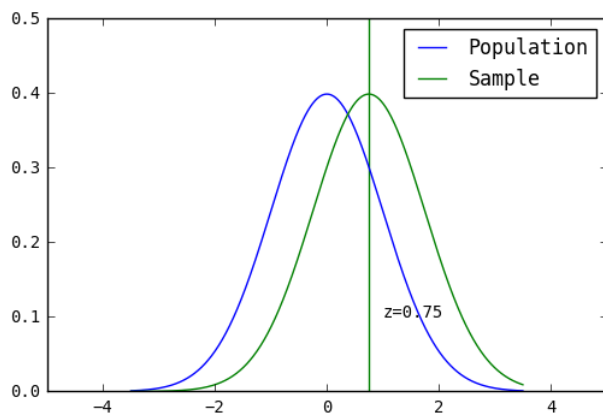


Figure 28. The sample is located at a Z score +0.75 from the population mean

The area beyond z-score 0.75 is 0.226. However, the alternate hypothesis is that population mean is *anything else* than 7.8. This is a two sided test, so the rejection region is in *both sides* of the mean = $0.226 * 2 = 0.452$, which is much greater than α . Thus **null hypothesis can not be rejected at $\alpha=0.05$** , i.e. this sample does not prove that the population mean is anything else than 7.8.

The perils of p-value

Supposing, you wrongly hypothesise that reading comics increases shoe size in children. The average shoe size in 5 year old children nationwide is 4.95. You pick a sample of 49 children and find the mean in this sample to be 5.0, with a standard deviation 0.2.

- $H_0: \mu_h = 4.95$
- $H_A: \mu_h > 4.95$
- $\bar{x} = 5$
- $s = 0.2$
- $n = 49$

CAUTION

Thus,

$$z = \frac{\bar{x} - \mu_h}{\frac{s}{\sqrt{n}}} = \frac{5 - 4.95}{\frac{0.2}{\sqrt{49}}} = 1.75$$

which corresponds to a $p = 0.04$ (one sided), not significant at 0.05 significance level. This is evident, because the effect size is very small (5-4.95 is 0.05, hardly a difference).

You now decide to take 100 children; you won't get much difference in **effect size** (because, let us admit that reading comics has nothing to do with shoe size). Let's say this time you got the same sample mean, thus

$$z = \frac{\bar{x} - \mu_h}{\frac{s}{\sqrt{n}}} = \frac{5 - 4.95}{\frac{0.2}{\sqrt{100}}} = 2.5$$

Which produces a p value 0.006, and behold, you have got a significant p-value just by increasing sample size!

The binomial test for discrete variables

Hypothesis testing for a proportion

Let

- the proportion (prevalence) of a disease in a population be assumed $H_0: \rho = 0.6$
- then $H_A: \rho > 0.6$
- proportion of disease in sample $p = 0.7$ ($n = 100$).

Then

$$Z = \frac{p - \rho}{s_r} = \frac{p - \rho}{\sqrt{\frac{\rho(1-\rho)}{n}}} = \frac{0.7 - 0.6}{\sqrt{0.6 \cdot \frac{0.4}{100}}} = 2.04$$

Which corresponds to a one sided p-value 0.02 (significant at $\alpha = 0.05$). However, the Z test can be done only when the binomial distribution approaches the normal. We can directly find the p-value from the binomial distribution. Here, observed number of heads = np = 70. Thus, the probability for 70 'heads' out of 100 tosses with a coin having inherent probability of heads = 0.6,

$$\sum_{i=70}^{100} {}^{100}C_i 0.6^i (1-0.6)^{100-i} = 0.01$$

```
sum(dbinom(70:100,100,0.6))
> 0.024
```

Which is the p-value; thus we can reject the null hypothesis and confirm the alternate at 95% confidence level.

The t test for independent mean

The t test is used when

- the population is normally distributed
 - if it is *not*, then the results will be distorted, but can be countered by increasing sample size
- mean and SD of the population are unknown

Hypothesis

- H_0 : The population mean is $\mu = 8.0$ (continuing with the shoe size example)

If this is true, then the variable t

$$t = \frac{\bar{X} - \mu_h}{\frac{s}{\sqrt{n}}}$$

(s = SD of a sample of size n)

will have the t distribution with degree of freedom ($n-1$). Lets say we choose 21 students from the class and get their shoe size.

CAUTION

Sample size for a t-test

The t test is based on sample SD, thus does not work very well with sample sizes < 40 (Bessel's correction might come into play). Over a sample size of 100, it becomes similar to z test.

Select critical levels for selected α

After you select α (say 0.05); select the t scores for the sample size you will be using, for a one tailed/ two tailed area (refer to t-table).

For an $\alpha=0.05$ and degree of freedom $(21-1) = 20$, the t value (two tailed) is 2.09.

Find the hypothetical population mean μ_h

If the null hypothesis is that the population mean is a certain μ_h , accept it for the time being. Continuing with our shoe size example, let us hypothesise that the mean shoe size of the class is 8.

Sampling

Draw a random sample; calculate its SD s and estimated standard error $s_{\bar{X}}$. We draw a sample of 21 students, and find the mean $\bar{X} = 7.68$, SD $s = 1.22$, estimated standard error $s_{\bar{X}} = 0.28$.

Calculate value of t

Let us find out where this sample lies among this population

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}} = \frac{7.38 - 8}{0.28} = -2.27$$

So the sample mean (7.38) lies 2.27 t -scores below the hypothesised population mean.

Compare the calculated and critical values of t

In the t distribution, an area outside -2.27 t -scores is 0.02 (two sided), which more than the selected $\alpha=0.05$. So the null hypothesis must be rejected with a 0.02 *probability of error* (i.e. there is still a 2% chance that the null hypothesis population mean might actually be 8.0).

The paired t test

This is exactly like the t -test for single mean, except that we test for mean *difference* between members of a group before and after some intervention/ experiment.

Example 11. Paired t -test

Suppose, for a year, we make all the students in a class read comic books and then measure their shoe sizes again. Like all statistical tests, we assume that there is, actually, no difference and the children grew up due to their natural growth spurt, not our comic books. To test this hypothesis, we must find the difference in shoe size of each individual student before and after injection, which we call d . Obviously, for 10 students, we will get 10 differences ($d_1, d_2, d_3 \dots d_{10}$). The summation of these differences ($\sum d$) divided by number of students (n), is the average difference or \bar{d} . The null hypothesis, states that the value of $\bar{d} = 0$ (i.e. there should no difference).

If the heights of 10 students before and after are [10,12,14,11,13,12,15,11,10,9] and [14,13,16,15,12,10,17,15,16,15], then, their differences are [4,1,2,4,-1,-2,2,4,6,6], and mean difference 2.6.

Now comes a crucial step. We *assume* if these differences d_1, d_2 etc are plotted, they will form a distribution among themselves (this is the key assumption behind the test). Some students will have grown much more than others (high difference) and some not much (low difference). We could, in theory, deduce the *standard deviation of differences*, which is

$$s_d = \sqrt{\sum \frac{(d - \bar{d})^2}{n - 1}}$$

Which in this case is 2.57. The t -value, in the paired t test is (note that we use $n-1$ because sample size is less than 30)

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n-1}}}$$

Which comes out to be 3.02. This t-value is checked, in the standard t table, for a degree of freedom $n - 1$, and a p-value could be deduced. Note the central theme in paired t testing is that we use the mean and standard error of differences before and after intervention, rather than using actual values.

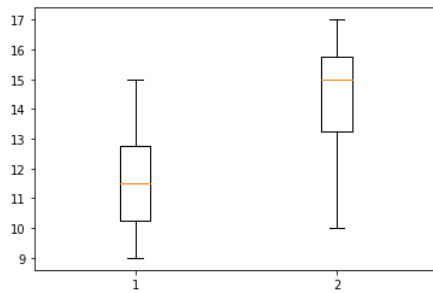


Figure 29. Box plot of pre and post shoe sizes

```
from scipy import stats
x1 = np.asarray([10,12,14,11,13,12,15,11,10,9])
x2 = np.asarray([14,13,16,15,12,10,17,15,16,15])
d = x2-x1
d_bar = d.mean()
sd = d.std()
n = len(x1)
mu_h = 0
se = (sd/math.sqrt(n - 1)) # standard error of differences
t = (d_bar - mu_h)/ se
area_to_right_of_t = 1 - stats.t.cdf(t,df=n-1) #prints area to the right of t at one tail
2*area_to_right_of_t # gives area on both tails, i.e. two sided p-value = 0.014
stats.ttest_1samp(d,0) # more direct way to do the same, returns (t,p) (double sided)
```

The p in this case is significant; let us also find a **critical t** for $df=9$ and $p=0.05$ (two sided, i.e. 0.025 on each side), which is 2.26. Building a confidence interval then

$$\bar{d} \pm t \cdot \sigma_d$$

which comes out to be $2.6 \pm 2.26(2.57/3) = (0.66, 4.57)$. Since this does not contain the null value (0), the results are significant.

The unpaired t-test for comparing independent samples with different variance

While comparing between two groups which are completely different,

1. H^0 : We assume that there is no difference at all between the two groups, i.e. $\mu_A = \mu_B$, or to put it succinctly, $\mu_A - \mu_B = d_\mu = 0$
2. Degree of freedom : a little complicated, but approximates to $n_1 - 1 + n_2 - 1$
3. Select alpha limit (usually 0.05)
4. Calculate the standard error of differences

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

Select value of **critical t** t_{df} for the degree of freedom; now, the confidence interval of the actual difference

$$(\bar{x}_1 - \bar{x}_2) \pm t_{df} \sigma_d$$

Now carry out the hypothesis test by calculating the **actual t**, which is $t = (\text{observed difference in means} - \text{hypothesised difference in means}) / \text{standard error of differences}$, i.e.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_\mu}{\sigma_d}$$

Find a p-value for this t and degree of freedom.

Example 12. Checking difference in shoe size between two groups

In a study for comics-reading and shoe size

	Read comics	Don't read comics
Number	26	30
Mean	5.4	6.1
SD	1.1	1.0

In this case, $df = 26 - 1 + 30 - 1 = 54$ and t_{df} (for two sided alpha limit 0.05) is 2, SE 0.281, thus confidence interval for mean difference is $(5.4 - 6.1) \pm 2 \cdot 0.281 = (-1.262, -0.138)$. This confidence interval does not include the null value, so the mean difference might be significant.

The actual $t = |(5.4 - 6.1) / 0.281| = 2.49$, which corresponds to double sided $p = 0.01$, significant.

The unpaired t-test for comparing independent samples with similar variance ^[1]

While comparing between two groups

1. H_0 : We assume that there is no difference at all between the two groups, i.e. $\mu_A = \mu_B$, or to put it succinctly, $\mu_A - \mu_B = d_\mu = 0$
2. Degree of freedom: $n_1 - 1 + n_2 - 1$
3. Select alpha limit (usually 0.05)
4. Calculate the standard error of differences

$$\sigma_d = \sqrt{\frac{\sum_{i=1}^{n_1} (x_i - \bar{x}_1)^2 + \sum_{i=1}^{n_2} (x_i - \bar{x}_2)^2}{n_1 - 1 + n_2 - 1}}$$

Because $s_1^2 = \frac{\sum_{i=1}^{n_1} (x_i - \bar{x}_1)^2}{n_1}$, it can be rewritten as

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

where n_1, n_2 are the number of members in two groups, respectively, and σ_1, σ_2 are the standard deviations of in those groups.

The t-value, in the unpaired t test, is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_d \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

```
#Python
scipy.stats.ttest_ind(a,b,equal_var=False) #returns t and p-value, by comparing two lists a and b
```

Analysis of variance

The question in an ANOVA is how much of variability is due to

- the individual
- and the *group* that he belongs to (and this is what we are interested in, while doing ANOVA)

and specifically, **whether there is a difference between at least a pair of groups**. ANOVA *won't* tell us which two groups show significant difference.

Let

1. Number of groups = k
2. Individual observations = X
3. Observations in each group = n_i
4. Overall observations = n
5. Individuals observations belonging to group $i = X_i$
6. Mean of a particular group = \bar{X}_i
7. SD of a group = s_i
8. Grand mean (across all groups) of is \bar{X}

Nutritional category	Shoe sizes
Bad	6,6.7,8,5,6,6,7,7
Moderate	7.2,7,6,8,7.1,6,8,7.1
Good	7.4,7,6,6,9,9,7.1,7.2,7.5

The null hypothesis H_0 is that the population mean of all groups is same, i.e. $\mu_1 = \mu_2 \dots = \mu_k$. Now, we calculate the sum of square total (SST) which is the **total variability** of the variable, which is similar to variance except it is not scaled down by sample size.

$$SST = \sum (X - \bar{X})^2$$

Then, the sum of squares groups (SSG) is calculated from mean of individual groups and grand mean

$$SSG = \sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2$$

The sum of square error (SSE), which is a measure of individual variability (which still remains after we eliminate the group variability) , is simply SST - SSG.

Next, we enlist the degrees of freedom

- total $df_T = n - 1$
- group $df_G = k - 1$
- error $df_E = df_T - df_G = n - k$

The mean squared groups, MSG is SSG scaled down by degree of freedom, i.e. $MSG = \frac{SSG}{df_G}$; similarly, mean squared error

$MSE = \frac{SSE}{df_E}$. Then, the Fischer Snedecor ratio

$$F = \frac{MSG}{MSE}$$

This can be expressed better as the ratio of two two chi square variables divided by their degrees of freedom. Effectively, the underlying assumption is that the two chi-square variables, inter-group and intra-group variability, produce a variable 'F' which is also got its own, continuous, slightly skewed distribution called the F- distribution. Note that F is always positive, and it makes no sense to calculate the area left of F.

```
data = pd.DataFrame({'Group':['B','B','B','B','B','B','M','M','M','M','M','M','G','G','G','G','G','G'],\
                    'Shoe_size':[6,6.7,8,5.6,6.7,7,7.2,7,6.8,7.1,6.8,7.1,7.4,7.6,6.9,7.1,7.2,7.5]})
mod = ols('Shoe_size ~ Group', data=data).fit()
anova_res = anova_lm(mod, typ=2)
anova_res

>      sum_sq      df  F      PR(>F)
Group      1.143333    2.0  2.164493    0.149319
Residual    3.961667   15.0      NaN      NaN
```

The inter group variability (1.14) + residual (individual) variability (3.96) = make up the total variability 5.1; the inter group variability/ total variability (effect size) is just 1.14/(1.14+3.96) = 0.22, and thus F=2.16, which gives an area 0.14 to the right of F (p value = 0.14, not significant).

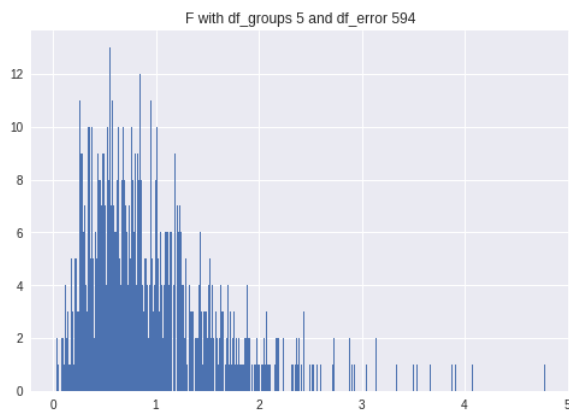


Figure 30. F distribution with df_groups 5 and df_error 594 (i.e. 6 groups, 100 observations in each group)

Conditions for ANOVA

1. individual members, and groups, need to be independent
2. the groups should be approximately normal and having roughly the same variance

The chi square test for independence

For a contingency table like this

	Acromegalics	Normal
Shoe size > 11	a	b
Shoe size < 11	c	d

The null hypothesis: We assume that shoe size is no indicator of acromegaly. Then the ones with small shoes are expected to have the same rate of acromegaly as those with large shoes (and that should be the prevalence of acromegaly in whole population in general).

The incidence of acromegaly in whole table is

$$\frac{a + c}{a + b + c + d}$$

Thus *expected* numbers in the four cells are calculated likewise, i.e. expected number in cell 'a' is total number of students with shoe size > 11 × incidence of acromegaly in whole population

$$(a + b) \times \frac{a + c}{a + b + c + d}$$

Similarly, expected number in cell 'b' = Total number of students with shoe size < 11 × incidence of 'no acromegaly' (normal children)

$$(a + b) \times \frac{b + d}{a + b + c + d}$$

χ^2 for each cell = (observed number - expected number)² / expected number. The summation of values of χ^2 is calculated. Now the degree of freedom of a table is (rows-1)(columns-1). The area to the right of this particular value of chi-square is

```
1 - stats.chi2.cdf(14.9,4) # Area above chi square > 14.9 with df = 4
```

Regression Models

Pearson's Correlation

For a dataset of two variables X and Y, the standard deviations

$$S_x = \sqrt{\frac{1}{n-1} \left(\sum_{i=1}^n (X_i - \bar{X}) \right)^2} \quad S_y = \sqrt{\frac{1}{n-1} \left(\sum_{i=1}^n (Y_i - \bar{Y}) \right)^2}$$

The covariance

$$Cov(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

and Pearson's correlation

$$R(X, Y) = \frac{Cov(X, Y)}{S_x S_y}$$

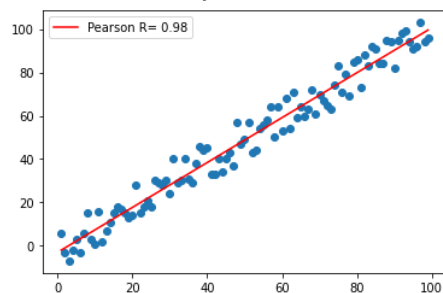


Figure 31. Positive correlation

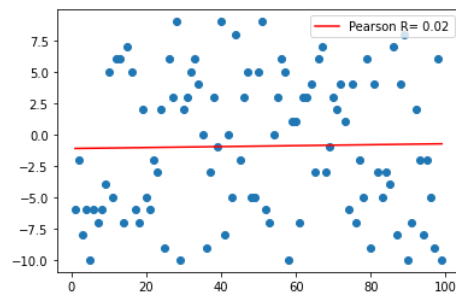


Figure 32. No correlation

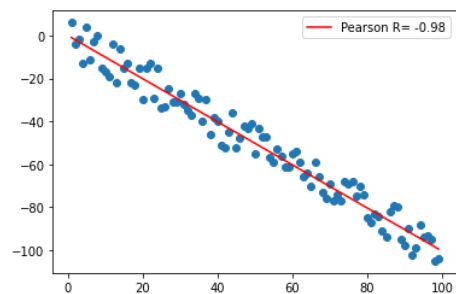
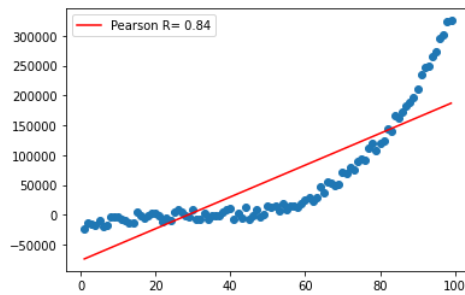


Figure 33. Negative correlation

No correlation, non linear relationship



```
import numpy as np
from scipy.stats import pearsonr

x = np.arange(1,100)
y = np.random.randint(1,100)
m, c = np.polyfit(x, y, 1) # find y = mx + c of regression line
R = pearsonr(x, y)
```

This last quantity is a unitless variable between -1 and 1, denoting strength of *linear* relationship between X and Y. Now comes the fun part: for a line $Y = mX + C$ that best fits through the XY plot (i.e. minimises the sum of square distances $\sum_{i=1}^n (Y_i - (m \cdot X_i + C))^2$, the values of m and C are

$$m = r(Y, X) \frac{S_y}{S_x}$$

and

$$C = \bar{Y} - m\bar{X}$$

Properties of R

- because the data has been *normalized* by subtracting the means, this line passes through \bar{Y} and \bar{X}

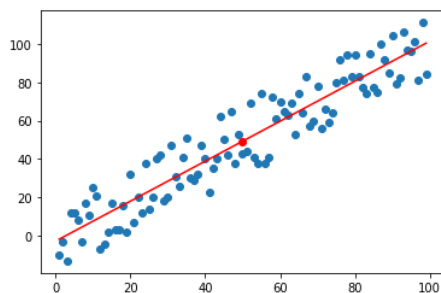


Figure 34. The least squares line will always pass through (\bar{X}, \bar{Y})

- if you reverse axis (i.e. X versus Y), you will have a new line; however, R stays the same ($R(X,Y) = R(Y,X)$)
- R is independent of unit of measurements
- R is sensitive to stray outliers

The square of the Correlation coefficient (R^2), is the proportion of variability explained by the model.

Total variability = R squared + variability of residuals

Which is to say, a two dimensional dataset is really variable in *two* directions.

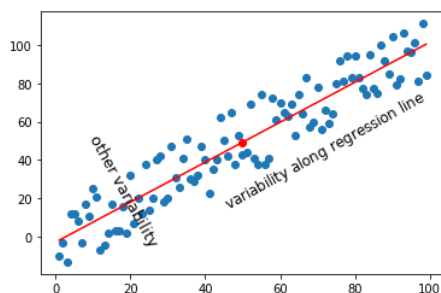


Figure 35. Partitioning variability

Residuals

The differences between $y_{\text{obs}} - y_{\text{pred}}$ ($y - \hat{y}$) are called **residuals**.

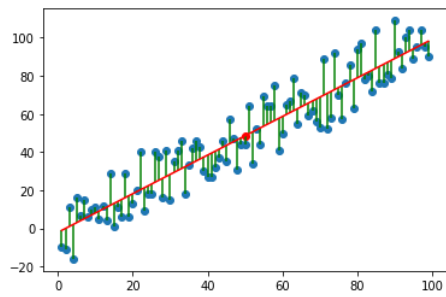
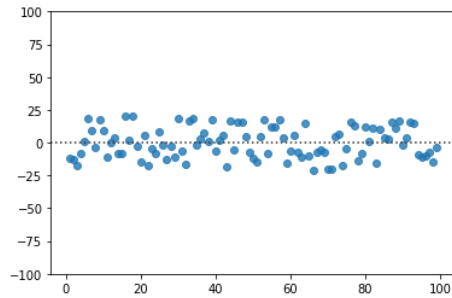


Figure 36. Residuals



The plot of the residuals, ideally, should be random noise around zero. There should be no more pattern remaining in residuals (otherwise, it's not true linear regression).

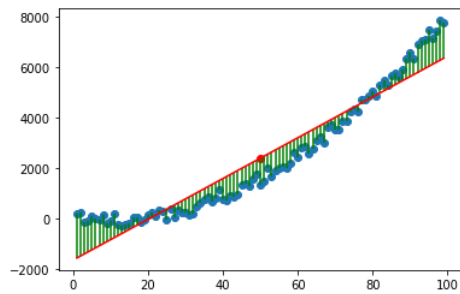


Figure 37. When the two variables have a non linear pattern, it also shows up in Residuals

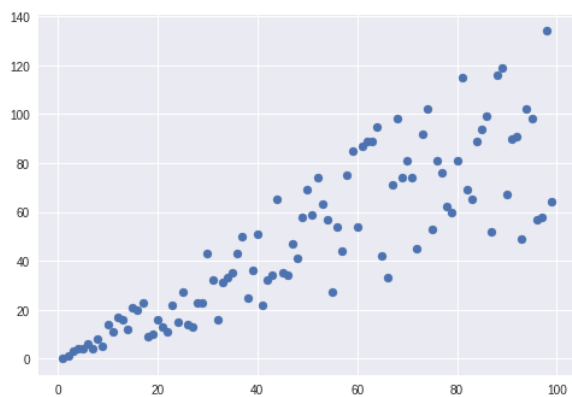
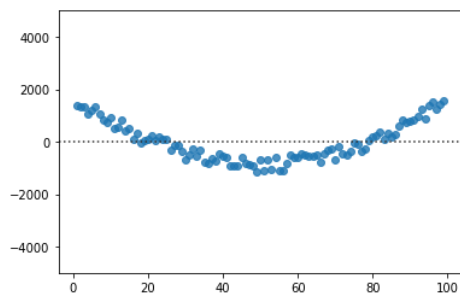
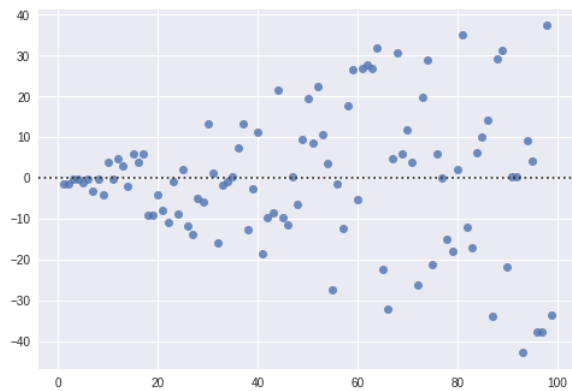


Figure 38. Analysis of residuals is useful to see 'variability in variance', in which case, linear regression doesn't work (i.e. it violates 'Homoscedasticity')



Relation between categorical and numeric variables

Regression also works between categorical and numeric variables; categorical variables can be represented in two way

Ordinal numbering

To assign a number to each category, i.e. education of parent {10th = 1; 12th = 2; Graduate = 3; Post graduate = 4}

Table 9. Table of shoe sizes with education level of parents

Shoe size	Education of parent
8	1
9	2
7.5	1
9	3
...	

One hot numbering

This is the more convenient method for regression analysis:

Table 10. Table of shoe sizes with education level of parents, one hot encoding

Shoe size	12th	Graduate	Post graduate
8	0	0	0
9	1	0	0
7.5	0	1	0
9	0	1	0
10	0	0	1

Note that the category '10th' is missing; this is because we assume a model such as

```
shoe size ~ c0 + c1*12th + c2*Graduate + c3*Postgraduate
```

Here, c_0 is the intercept, corresponding to the value of y when x is 0 (which in, this case is '10th', also called the *reference level*). In this model, $c_0 = 8$. Geometrically, this represents the equivalent of a line in a 5 dimensional coordinate system. In general, having n explanatory variables produces a 'hyperplane' in $n + 1$ dimensional space.

```
#Modelling the table above in R
model <- lm(Shoe.size ~ X12th + Graduate + Post.graduate, df) # df is the dataframe
summary(model)
>
      Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.000     1.061   7.542  0.0839 .
X12th         1.000     1.500   0.667  0.6257
Graduate      0.250     1.299   0.192  0.8790
Post.graduate  2.000     1.500   1.333  0.4097
```

Which indicates $c_0 = 8$, $c_1 = 1.0$, $c_2 = 0.25$, $c_3 = 2$. Now, if we try to partition the variability between different components

```
anova(model)
> Analysis of Variance Table

Response: Shoe.size
      Df Sum Sq Mean Sq F value Pr(>F)
X12th   1  0.1125   0.1125   0.1000  0.8050
Graduate 1  0.5625   0.5625   0.5000  0.6082
Post.graduate 1  2.0000   2.0000   1.7778  0.4097
Residuals 1  1.1250   1.1250
```

This indicates, total variability $(0.1125 + 0.5625 + 2 + 1.125) = 3.8$, thus R^2 for '12th' = $0.1125 / 3.8 = 0.029$; this is the same result that we arrive at if we calculate the correlation coefficient first and then square it.

```
cor(df$Shoe.size,df$X12th)**2
> 0.029
```

We have thus two ways of calculating R^2 : from correlation coefficient R , and from ANOVA table.

Multiple regression

So we can mix and match numeric and categorical variables and produce a 'line' in n -dimensional space, using similar principles.

$$y = c_0 + c_1x_1 + c_2x_2 \dots c_nx_n$$

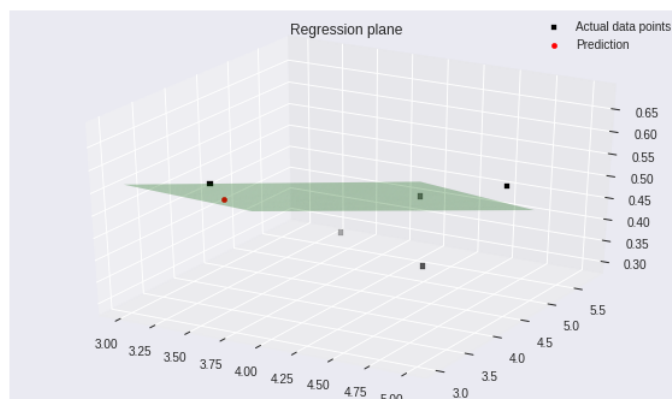


Figure 39. Regression plane with two explanatory variables

Adjusted R^2

For k predictor variables, the composite or 'adjusted' R_a^2 of the model, is defined as

$$R_a^2 = 1 - \left(\frac{SSE}{SST} \cdot \frac{n-1}{n-k-1} \right)$$

where SSE = sum square error of residuals, and SST = sum squared total variance, n = sample size. Often, the addition of new predictor variable will reduce R_a^2 , because of existing *collinearity* between predictors (i.e. they are not independent of each other). In such scenarios, it is better to plot pair of variables (screplot) to see existing collinearity.

Index

1. For variants of this test, see a specialised stats book; there are tests for groups with unequal variance, unequal sample sizes and so on
1. After Thomas Bayes
2. Actually, 1.96
3. Think about it. Possibly, this is the most important statement of inferential statistics
4. A term coined by 'Student', a pen-name taken by W A Gossett
5. If while screening for diabetes, you select a fasting sugar cut off of 80, of course you will detect all diabetics, but in addition, so many normal people who range over 80 mg/dL will be caught delinquent (loss of specificity). Think yourself what may happen if you select a cut off 160 mg/dL.