Mathematical Statistics

2016

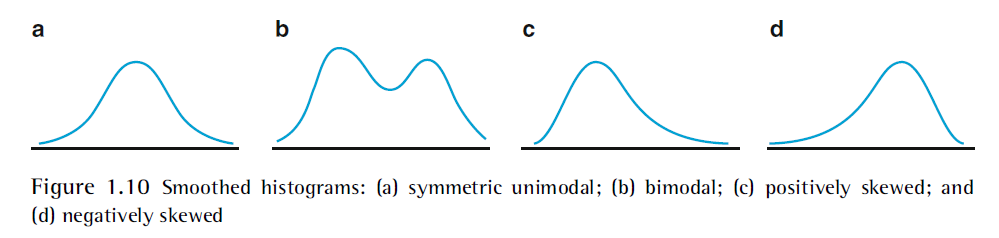
**Book:**

Modern Mathematical Statistics with Applications, 2nd Edition (2012), by Jay L. Devore, Kenneth N. Berk

# Descriptive Statistics

## Pictorial Methods

**Histogram shapes**



**Mode** – The data point that is the most frequent. Hence the terms **unimodal**, **bimodal**, **multimodal**.

**Positively skewed** = more data on the positive side of the central peak

**Multivariate Data** – For each object there are two or more measurements.

The sample is **bivariate** if each observation consists of two measurements, so that the data set can be represented as (x1, y1) ... (xn, yn).

## Measures of Location

A **trimmed mean** is the mean of the data with the outliers removed. Say the 10% of the outlying data is removed.

## Measures of Variability

**Sample Variance**

|  |
| --- |
|  |

**Population Variance**

σ = population standard deviation

**Why divide by (n-1) instead of divide by (n)**

If dividing by "n", the sample variance will tend to under-estimate the population variance. So divide by "n-1" instead, so to boost the sample variance, to be more in line with the population variance.

**Example:** The population of numbers is {1, 1, 2, 2, 3, 3}, and three numbers are sampled.

Some ways this population can be sampled:

* the sample {1, 1, 3} will have higher variance than the population
* the sample {1, 2, 2} will have lower variance than the population
* the sample {1, 2, 3} will exactly represent the population

There are total of 6C3 = 20 ways to sample the population

To produce a higher variance, choose 3 numbers out of {1, 1, 3, 3}. There are 4C3 = 4 ways to choose numbers that produce higher variance.

There are 2 \* 2 \* 2 = 8 ways to choose numbers to produce exact variance

There are 20 – 4 – 8 = 8 ways to choose numbers to produce lower variance

So most of the time divide by "n" will produce a sample variance that is lower than the population variance.

**s2 is based on (n-1) degree of freedom, instead of (n) degree of freedom:**

The s2 is dependent on "n" terms: x1, x2, ... xn. But it's known that:

So knowing (n-1) of the terms is sufficient to know the final term.

**A Computing Formula**

Proof:

Start with the , expand and the distribute the summation sign. Note the " is a constant that can be distributed outside of the summation sign.

Further simplification using and .

**Properties:**

Let x1, x2, ..., xn be a sample and c be a constant.

1. **If y1 = x1 + c, y2 = x2 + c, ..., yn = xn + c, then sy2=sx2**

This means if each data point is shifted by a constant amount, then variance remains the same – because the spread of the data remains the same. When each data point is shifted by the same amount, the mean will be shifted by the same amount, and the data to mean distance remains the same.

1. **If y1 = cx1, y2 = cx2, ..., yn=cxn, then sy2=c2sx2, sy=|c|sx**

This means if each data point is scaled by a constant, then the standard deviation will be scaled by the same constant. When each data point is scaled by a constant, the mean will be scaled by the same constant. The distance between the data point and the mean will be scaled by the same constant.

**Fourth Spread (fs)**

fs = upper fourth – lower fourth

**outlier** – Any observation more than 1.5 fs from the closest fourth

**extreme outlier** – Any observation more than 3 fs from the closest fourth. Otherwise, it's a **mild outlier**

**Interquartile range** – this is another way of saying "Fourth Spread".

**IQR = Q3 – Q1** – abbreviation for interquartile range

# Probability

## Sample Spaces and Events

**Sample space (S)** – the set of all possible outcomes of that experiment

**event** – any collection (subset) of outcomes contained in the sample space

**simple event** – event with one outcome; otherwise it's called a **compound event**

Example:

Three vehicles turning left (L) or right (R).

The sample space S = all eight outcomes = {LLL, RLL, LRL, LLR, LRR, RLR, RRL, RRR}.

A simple event is a single outcome, like E1 = {LLL}.

An example of a compound event is having one out of three cars turn right. A = {RLL, LRL, LLR}.

**Relations from Set Theory**

|  |  |
| --- | --- |
|  | **union** of two events A and B |
|  | **intersection** of two events A and B |
|  | **complement** of an event A |
|  | When events A and B have no outcomes in common, they are said to be **disjoint** or **mutually exclusive** events. |

## Axioms and Properties of Probability

**Union of Two Events**

|  |
| --- |
|  |

**Union of Three Events**

**Distributive Law from Set Theory**

## Counting Techniques

**Permutation**

Permutation of "k" objects taken from a set of "n" distinct objects:

|  |
| --- |
|  |

**Combination**

Given a set of "n" distinct objects, any unordered subset of size "k" of the objects is called a **combination**.

The number of combinations of size "k" that can be formed from "n" distinct objects:

|  |
| --- |
|  |

## Conditional Probability

For any two events A and B with P(B) > 0, **the conditional probability of A given that B has occurred** is defined by

|  |
| --- |
|  |

**The Multiplication Rule**

**Mutually Exclusive and Exhaustive Events**

Events are **mutually exclusive** if no two have any common outcomes. The events are **exhaustive** if one must occur, so that .

**The Law of Total Probability**

Let be mutually exclusive and exhaustive events. Then for any other event B,

|  |
| --- |
|  |

**Bayes' Theorem**

Let be mutually exclusive and exhaustive events. Then for any other event B,

|  |
| --- |
|  |

This theorem links to .

## Independence

Two events A and B are **independent** if and **dependent** otherwise.

. If two events are mutually exclusive, they cannot be independent.

Events A and B are independent if and only if

|  |
| --- |
|  |

**Independence of More than Two Events**

Events A1, ..., An are mutually independent if for every k (k=2,3, ..., n) and every subset of indices i1, i2, ..., ik,

# Discrete Random Variables

rv = random variable

**Bernoulli random variable** – a random variable whose only possible values are 0 and 1

## Probability Distributions

**PMF**

The **probability distribution** or **probability mass function (pmf)** of a discrete rv is defined for every number by .

**Heavy tail** – Any probability distribution having a large amount of probability far from the average µ. Such heavy tails make it difficult to make inferences about µ.

**CDF**

The **cumulative distribution function (cdf)** of a discrete rv X with pmf is defined for every number x by:

Note the way that this is defined, F(5) will include the probability of "x=5". For integer probabilities that means:

## Expected Values

Let X be a discrete rv with set of possible values and pmf . The **expected value** or **mean value** of X is

|  |
| --- |
|  |

**Expected Value of a function h(X)**

**Expected Value of the Linear Function**

|  |
| --- |
|  |

**Variance and Standard Deviation**

Let X have pmf and expected value . Then the **variance** of X is:

|  |
| --- |
|  |

The **standard deviation** (SD) of X is:

|  |
| --- |
|  |

**Shortcut Formula for**

|  |
| --- |
|  |

Proof

expands to:

The first term is E[X2].  
In the second term the "-2µ" can be factored out to become -2E(X)\*E(X).  
The third term is [E(X)]2

**Rules of Variance**

|  |
| --- |
|  |

Addition of the constant to a random variable does not affect the variance.

## Moments

**Definition of Moments**

First moment about 0:

Second moment about 0:

Second moment about the mean:

**skewness** – a measure of departure from symmetry. This is the third moment about the mean divided by σ3:

A negative skewness value means "X-µ" is negative, so the possible outcomes are skewed to the left of the mean.

**Moment Generating Function (mgf)**

|  |
| --- |
|  |

The mgf is unique for each probability distribution - If the mgf exists and is the same for two distributions, then the two distributions are the same.

**Producing Expected Values from mgf**

**MGF of**

Let X have the mgf MX(t), and let .

Proof:

## Binomial Probability Distribution

Binomial Experiment:

* Consists of a sequence of "n" smaller experiments called trials.
* Success and failure are the only two possible outcomes per trial.
* Each trial has a "p" probability of success.

**Replacements in Binomial Experiments** – the perfect binomial experiment is done with replacements, meaning for each trial, the probability of success is exactly the same. If the experiment is done without replacements, then it's not strictly binomial.

Example: Suppose a city has 50 licensed restaurants, of which 15 currently have at least one serious health code violation. Suppose five inspections are made. The first inspection will therefore have a 15/50 chance of finding the bad restaurant. If the first inspection failed to find the bad restaurant, the second inspection has a 15/49 chance of doing so. Note the probability for success is no longer 15/50. This is an experiment without replacement, so the probability distribution is not strictly binomial.

Example #2: Suppose a city has 5000 restaurants, and we are inspecting 10 of them. Again the probability for catching the bad restaurant is not the same per inspection, but this time, because it's 5000 instead of 50, the probability is nearly constant per inspection. So the probability distribution is not strictly binomial, but can be approximated by the binomial distribution.

**5% Rule:** When sampling without replacement, if the number of trials "n" is at most 5% of the population size, then the experiment can be analyzed using binomial distribution.

**Equations**

|  |
| --- |
| n = number of trials  p = probability of success per trial  rv X = the number of successes  PMF:  Mean:  Variance: |

**Moment Generating Function**

Proof:

This can be simplified using the binomial theorem:

## Hypergeometric and Negative Binomial

**Hypergeometric Probability Distribution PMF**

|  |
| --- |
|  |

Denominator: The population has a total size of "N", and we are choosing "n" out of "N"

Numerator: There are two outcomes.

* Outcome #1 has M elements total in the whole population.
* Outcome #2 has N elements.
* We are choosing "x" out of "M" outcome #1, and the rest "n - x" is coming from outcome #2.

**Hypergeometric Probability Distribution Mean and Variance**

Let .

, which is the same as the binomial.

This differs from the binomial by the factor, which is called the **finite population correction factor**. The hypergeometric rv has smaller variance than the binomial rv.

**Negative Binomial Probability Distribution PMF**

The "r" is the number of successes. The "x" is the number of failures that PRECEDES the "r-th" success. So the negative binomial random variable probability can be reworded as:

|  |
| --- |
|  |

**Negative Binomial MGF**

Proof:

We need to know:

to get the MGF. This expression is why it's called the "negative binomial" distribution.

**Expansion of**

Proof:

Note this expansion goes on forever.

You will need to know:

**Alternate expression for**

Proof:

Change the "n" to "-r":

Factor out the "-1". The numerator has "x" terms. Rearrange to get:

The way combination works, 10C3 is the same as 10C7, so the above can be re-written as:

## Poisson Probability Distribution

**PMF, Mean, and Variance**

|  |
| --- |
|  |

This distribution initially increases due to the , but then decreases rapidly due to the

**Proof that the PMF adds to 1**

The power series for is:

**The Poisson distribution as a limit of the binomial**

The binomial converges to the Poisson if .

**Rule of thumb - the Poisson approximation applies if n > 50 and np < 5.**

Proof:

The binomial pmf:

Rearrange into:

Apply

To evaluate the limit, use

**Poisson Moment Generating Function**

Proof:

Finish up using

**The Poisson Process**

Conditions for the Poisson Process:

1. For any short time interval of length , the probability that exactly one event occurred is .
2. The probability of more than one event occurring during is .

The is some function that follow the property . So needs to be something that goes to zero faster than

is a function that will satisfy this restriction, while will not satisfy this restriction.

The probability of no event during is .

1. The number of events during is independent of the number of events prior to .

The number of events from the above process will follow a Poisson distribution.

Let denote the probability of events during any particular time interval length .

Note the average of Poisson is , which is why . The is a rate, and is the duration. The expected number of events is .

**Proof that the Poisson Process results in the Poisson Distribution (Exercise #107)**

**Exercise #107 Part A ~ C**

Suppose there are no events in the interval .

Mathematically:

Rearrange:

Apply :

This crucial step is made possible due to the restriction that .

The solution to the differential equation is

**Exercise #107 Part D**

Parts A through C looks at the situation k=0. Part D generalizes "k".

To have "k" events in the interval there are two possibilities:

* "k" events in and zero event in
* "k-1" events in and one event in

Mathematically:

Rearrange:

Apply :

The solution to the differential equation is

# Continuous Random Variables

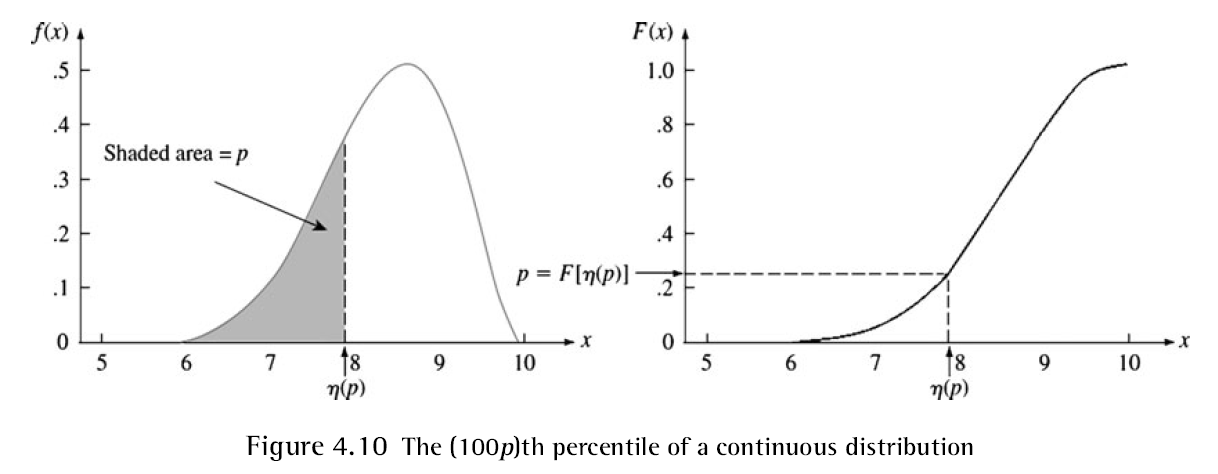
## PDF and CDF

**Probability Density Function (PDF) - f(x)**

**Cumulative Distribution Functions (CDF) - F(x)**

**η(p)** - location of (100p)th percentile.

For example η(0.25) is the value on the measurement axis that marks the 25th percentile. 25% of the area under the pdf graph lies to the left of η(25).



## Expected Value sand MGF

**Expected Value E(X)**

|  |
| --- |
|  |

**Variance V(X), Standard Deviation**

|  |
| --- |
|  |

Similar properties as the discrete case:

|  |
| --- |
|  |

**Approximating the Mean Value and Standard Deviation - The Delta Method**

Suppose Y = h(X), and the variance of X is small, such that the distribution of X is largely concentrated on an interval of values close to μ, then:

|  |
| --- |
|  |

Proof:

The key is the assumption that h(x) can be approximated using a first-order Taylor series expansion:

To prove , the setup is

This expands to three integrals:

To compute E[h(x)2], the setup is

This expands to the following integrals:

The variance is then computed by E[h(x)2] - E[h(x)]2

**Moment Generating Function (mgf)**

|  |
| --- |
|  |

The properties from the discrete case carries over.

As a check, MX(0) should always be 1.

The mgf uniquely identifies the distribution.

Moments are related to the mgf via:

|  |
| --- |
|  |

If random variables and are related by , then

|  |
| --- |
|  |

## Normal Distribution

**PDF**

|  |
| --- |
|  |

**Proof that**

We need to show:

First simplify the problem with a variable substitution.

Now the problem reduces to proving that

Now reframe the problem as a volume problem. Define:

Next, find the volume of a solid defined by z = f(x) \* f(y). This will be done two different ways.

One way is to use the shell method.

Another way is to use the double integral method.

Let , then V = A2. Previously V = 1, so A = 1.

**Standard Normal Distribution**

μ=0, σ=1

|  |
| --- |
|  |

**zα Notation**

Example: z0.01 = 2.326

This means the z value of 2.326 corresponds to the top 1%. There is 1% area to the right, and 99% area to the left.

**Standardizing X ~ N(μ, σ2)**

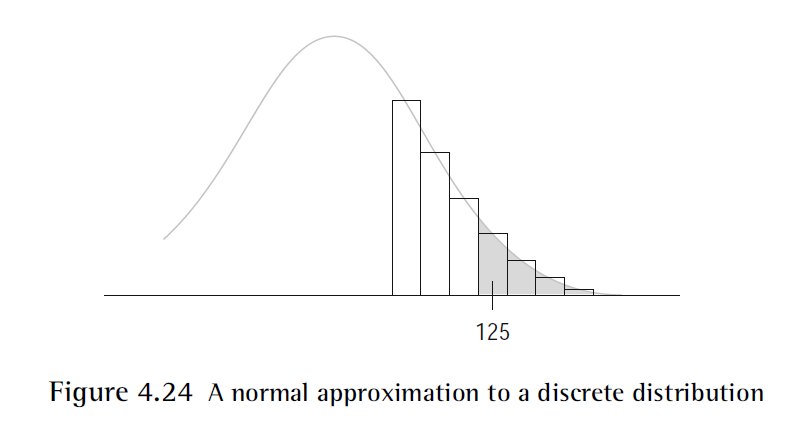
"Z" is the number of standard deviations from the mean.

Roughly 68% of the values are within 1 SD of the mean, 95% within 2 SDs, and 99.7% are within 3 SDs.

**Continuity Correction for Discrete Populations**

This is a minor correction to be made if the normal distribution is being applied to something that is discrete in nature.

Example: IQ (as measured by a standard test) is known to be approximately normally distributed with μ = 100 and σ = 15. What is the probability that a randomly selected individual has an IQ of at least 125?



The IQ scores are integers, so the 125 rectangle really begins at 124.5.

z = (124.5 - 100) / 15

percent = 1 - stats.norm.cdf(z) # 0.051

**Approximating the Binomial Distribution using the Normal Distribution**

As long as the binomial probability histogram is not too skewed, binomial probabilities can be approximated by the normal distribution.

For a binomial distribution:

Due to the continuity correction, you should add 0.5 to x.

The Φ() is a function that turns a "z" value into a probability.

The approximation is good if BOTH AND .

**Moment Generating Function for Normal Distribution**

|  |
| --- |
|  |

Proof

First get the moment generating function for the standard normal:

Apply a complete the square operation:

The term comes out:

The integral part is a normal distribution with mean of "t" and standard deviation of 1. So the integral part is 1.

Transform from the standard normal to the general normal:

The moment generating function for the general normal:

## Gamma Distribution

**The Gamma Function**

|  |
| --- |
|  |

The most important properties of the gamma function are:

1. For α > 1, Γ(α) = (α-1) \* Γ(α-1)
2. For positive integer n, Γ(n) = (n-1)!

Proof:

#1. Use integration by parts.

The "uv" term becomes (0 - 0) = 0.

#2. This follows from #1. The gamma function is like a factorial function that works for floating points.

#3. The 1/2 case:

This integral is a lot like the standard normal distribution integral:

Use the substitution z2/2 = x. The limit of integration will change - the "z" range is -∞ to +∞. The "x" range is 0 to ∞, and be sure to note that this happens two times. The "z" going covering maps to "x" covering one time, while the "z" covering maps to "x" covering a second time.

The variable substitution results in:

**The Family of Gamma Distributions**

|  |
| --- |
|  |

The **standard gamma distribution** has β = 1.

**Discussion:**

**Why the pdf integrates to "1":**

For the standard gamma distribution (β = 1) case, will produce the gamma function, which is in the denominator. That's why the expression is a valid pdf.

Starting with the gamma function integral:

If we change the term to

Do a change of variable:

leads to:

which is times the gamma function. That's why the factor appears on the denominator to keep the total probability 1.

**Parameter α:**

When (α < 1), the variable part of the pdf writes out to be:

So the f(x) will always be decreasing.

When (α > 1), the variable part of the pdf writes out to be:

So the f(x) will initially grow. Ultimately the polynomial function at the top will be overwhelmed by the exponential function at the bottom, and the function will decrease.

**Parameter β:**

The variable part of the pdf:

The β controls how fast the denominator will overwhelm the growth in the numerator. The larger the β, the slower the denominator grows, and the function becomes more stretched out. On the other hand, a smaller β compresses the f(x) function in the horizontal direction.

**Moment Generating Function**

|  |
| --- |
|  |

Proof:

Start with:

The strategy is to make the integral look like the gamma function pdf, which will result in the integral becoming a "1".

The substitution is:

After substitution:

The integral part is 1.

The new "b" is not the same quantity as the old "β". To get the final answer, express the new "b" in terms of the old "β".

**Mean and Variance of the Gamma Distribution**

|  |
| --- |
|  |

**CDF of the standard gamma random variable**

This is called the **incomplete gamma function**.

**CDF of Gamma Distribution with the β Parameter**

|  |
| --- |
|  |

Proof

Start with

Do a substitution to make the integral look like the "β=1" case.

Notice that the limit of integration changes from "0 🡪 x", to "0 🡪 x/β".

When the substitution is made, the only occurrence of "β" is in the limit of integration.

### Exponential Distribution

This is a special case of the general gamma pdf, with α = 1 and β = 1/λ.

|  |
| --- |
|  |

**Memoryless Property**

Discussion:

Suppose that component lifetime is modeled by the exponential distribution.

A

t

0

The probability that a component will last beyond time "t" is the area A.

Suppose it's already known that a component has lasted time "t0"

B

A

0

t

t0

the probability that this component will last another time "t" is "(B)/(A+B)".

So after time "t0" passed by, the component's failure probability in time "t" remain exactly the same. It's as if there is no wear and tear.

**MGF**

Note there is a requirement in order to evaluate the .

## Other Distributions

### Weibull Distribution

|  |
| --- |
|  |

**Discussion**

It's similar to the gamma distribution, but it's not exactly the same thing as it has no Γ(α).

Similar to the gamma distribution, the numerator grows faster than the denominator at first, but eventually the denominator will grow faster.

The "α" determines how fast the numerator grows. Increase it to increase the peak of the pdf.

The "β" determines how fast the denominator catches up to the numerator. It will stretch or compress the graph horizontally.

### Lognormal Distribution

If X is said to have a lognormal distribution, then that means Y = ln(X) has a normal distribution.

So we start with a normally distributed Y:

Note that the μ and σ are the mean and standard deviation of Y, not of X.

To change the random variable from Y to X, use the "differentiate first, then substitute" method mentioned in the future section "Transformations of a Random Variable."

The PDF of the lognormal distribution:

|  |
| --- |
|  |

### Beta Distribution

|  |
| --- |
|  |

The case A = 0, B = 1 gives the standard beta distribution:

**Discussion**

The shape of the graph is determined by the rational function .

If α and β are less than one, such as

then you have asymptotes at x = 0 and x = 1. The graph is going to be U-shaped bathtub curve.

If α and β are larger than one, such as

then you have zeros at x = 0 and x = 1. The graph is going to be dome shaped in the interval .

## Transformations of a Random Variable

**To obtain the PDF, differentiate first, then substitute**

Given:

What is the pdf of Y?

One solution is to get the CDF FX(x), substitute x=y/60, and then differentiate.

So the order of operation is (1) substitute, (2) dy. This requires the knowledge of the FX(x) expression.

Alternatively, invert the order, **differentiate first, and then substitute**.

|  |
| --- |
|  |

For the example given

The pdf must be positive. So the (dx/dy) term should be |dx/dy|.

**Non-Monotonic Transformations**

Example: Suppose that X is a standard normal random variable, and that the transform is Y = X2. What is the pdf of Y?

The book has a more formal solution, using absolute value function |X|.

Start with

The procedure

will give

This is not the pdf for Y, but very close. In the transformation Y = X2, two different sets of X value will map to the same Y value. X = +2, and X = -2, will both produce Y = 4. The pdf of X is symmetric, so both X near +2, and X near -2 will contribute the same amount of probability.

The final correct answer is double of the expression shown above.

Another way to think about this is that we are applying the "differentiate first, then substitute" procedure for half of fX(x), say the x > 0 half. Then double the result because the other half of fX(x) is the exact same thing.

**Breaking Down Non-Monotonic Transformations**

Example: fX(x)=(x+1)/8, -1<x<3, and Y=X2. What is the pdf of Y?

The X to Y mapping is graphed below.

Y

X

-1 0 1 3

The 1 < x < 3 case is the standard case.

The -1 < x < 1 case is a special case. The Y pdf has two contributing components.

For 1 < x < 3, the "differentiate first, then substitute" procedure gives

For 0 < x < 1, the pdf is the same thing.

For -1 < x < 0, the pdf uses a different substitution

This "pdf" has a problem, it's negative for 0 < y < 1.

The origin of the problem is that the (x+1)/8 pdf is running from x = -1 to x = 0. The Y pdf needs to run from y = 0 to y = 1. For the -1 < x < 0 interval, this require the x variable to run from x = 0 to x = -1.

So the pdf for the -1 < x < 0 region is:

The 0 < y < 1 pdf is the sum of the two pdfs

Final answer

# Joint Probability Distributions

## Jointly Distributed Random Variables

|  |  |
| --- | --- |
| **joint probability mass function** | **joint probability density function** |
| **marginal probability mass function** | **marginal probability density function** |

**Independent Random Variables**

Two random variables X and Y are said to be independent if:

This needs to hold for EVERY (x,y) entry.

For the continuous case, the region of positive density needs to be a rectangle - such that x1 and x2 have the same possible y values.

**Multinomial Experiment**

This is a generalization of the binomial. Instead of just heads and tails, there are "r" outcomes, with probabilities p1, p2, ... pr. There are "n" total trials, x1 trials with outcome #1, x2 trials with outcome #2, and so on, with n = x1 + x2 + ... + xr .

|  |
| --- |
|  |

**Discussion**

The is the probability that a particular combination of outcomes will happen in a SINGLE particular order.

The is the number of order this particular set of outcomes can happen.

Suppose n=10, and there are 4 possible outcomes, imagine:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Outcome #1 | | | Outcome #2 | | Outcome #3 | | | Outcome #4 | |
| Labels: | A | B | C | D | E | F | G | H | I | J |
| Slots: |  |  |  |  |  |  |  |  |  |  |

n! is the number of ways to position the "A" ... "J" labels into the 10 slots.

For outcome #1, the following positioning are all equivalent

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Slots: |  | A |  | B | C |  |  |  |  |  |
| Slots: |  | A |  | C | B |  |  |  |  |  |
| Slots: |  | B |  | A | C |  |  |  |  |  |

They all mean:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Slots: |  | outcome  #1 |  | outcome  #1 | outcome  #1 |  |  |  |  |  |

There are 3! = 6 ways to position the "A, B, C" labels for a single ordering of "outcome #1".

So 10! / 3! will REMOVE the excessive counting related to outcome #1.

Divide by 2! to REMOVE the excessive counting related to outcome #2, and so on.

## Expected Values, Covariance, and Correlation

**Expected values of two jointly distributed random variables**

|  |  |
| --- | --- |
|  |  |

For independent variables, the integrals and summations would be separable into products.

|  |
| --- |
|  |

**covariance**

|  |
| --- |
|  |

Discussion:

The covariance is most positive if large positive values of X occur with large positive values of Y. It is most negative if large positive values of X occur with large negative values of Y.

X and Y need not have the same number of outcomes.

Every X is paired with every Y. The p(x,y) determines which of these pairings are "strong".

**covariance computation formula**

|  |
| --- |
|  |

Proof:

Start with:

Expand to get:

The E[Y] is µY, and E[X] is µX .

|  |
| --- |
|  |

Useful:

**covariance linearity**

Proof

**Correlation**

|  |
| --- |
|  |

Discussion:

The defect of covariance is that its computed value depends on the units of measurement. By dividing the standard deviations, the correlation is "standardized".

**Correlation Properties**

1. Corr(X,Y) = 1 or -1 iff for some numbers a and b.
2. If X and Y are independent, then Corr(X,Y) = 0, but correlation being zero does not imply independence.

**Correlation Properties Discussion**

**#1.**

The statement means linear scaling and shifting does not affect correlation. This can be proven by expected value expansion.

In the numerator:

eventually reduces to:

So the covariance function is already "immune" to the effect of shifting by "b" and "d". It does get affected by the scale factors "a" and "c".

For the denominator:

This eventually reduces to:

The "a" and "c" in the numerator will cancel with the "a" and "c" in the denominator.

**#2 and #3.**

1. Corr(X,Y) = 1 or -1 iff for some numbers a and b.

These two statements are related.

The proof relies on a trick where the X and Y are "standardized" as ZX and ZY.

These redefined variables make the math much easier.

In discussion "#1" we already saw:

The Cov(ZX, ZY) can be expanded and further simplified

The expected values of "Z" are zero due to the standardization definition, so we have in total

The "ρ" has been introduced to make the expression more compact.

This is a huge simplification - from Corr(X,Y) to just E() of two redefined variables.

The next trick is to use the following inequality as a starting point:

The expected value expands to be

The keys to simplifying this expression is to use the from before, and to realize that:

So that

This leads to the

Referring to , in the special case that ρ = 1, we have

In order for this to be true, we need everywhere. The expected value is a sum, and if any single is not zero, then will be above zero, and the final total cannot be zero.

The requirement expands to become

When ρ=1, Y = AX + b.

This means the correlation coefficient is a measure of the degree of **linear relationship** between X and Y. Correlation is only 1 if there is a perfect linear relationship.

**#4.**

1. If X and Y are independent, then Corr(X,Y) = 0, but correlation being zero does not imply independence.

Start with the definition of Cov(X,Y)

When X and Y are independent, the pdf for X refers to only "x" and the pdf for Y refers to only "y". The integral or summation is then separable.

Separately each term is zero.

Here's an example where there is a non-linear relationship , yet the correlation is numerically zero if certain points are sampled.

The x\*y product sums to be zero because there is y-axis symmetry. For every (2,1.5), there is a (-2, 1.5) to balance it out.

The µX and µY are also zero, because of the way the points are selected. The µX is zero because the x values selected are -2, -1, 1, 2. The µY is zero because I shifted the parabola so that f(2) is balanced out by f(1).

So the covariance is zero, but that's a sampling hack. If more points are sampled, the covariance will not be zero.

, and the hack achieve a total zero by having all three terms be zero.

Having E(XY) = 0 require symmetry about the y-axis, meaning f(-x) = f(x). If the x-value are chosen symmetrically, meaning choosing x=-2 to balance out that x=2, then having E(Y) = 0 require symmetry about the origin, meaning f(-x) = -f(x). These two requirements, and are mutually exclusive, so cannot happen in general.

From: <http://people.math.gatech.edu/~ecroot/3225/rho_notes.pdf>

Another counter example is to define two random variables, A and B, that each take values of   
{-1, 1}. The probability for each value is 50%, so that the mean of A and B will be zero. This {head, tail} encoded as {-1, 1}.

Then define: X = A + B, Y = A - B. The Cov(X, Y) will be zero --- because the mean of X, mean of Y, and mean of (XY = A2-B2) is zero.

Again this example is contrived. It works by having lots of zeros in the various expected values.

Since it's so hard to come up with a "normal" example of a dependent relationship that leads to Cov(X,Y) being zero, seeing Cov(X,Y) being zero in the real world has a good chance of X and Y being relatively independent.

## Conditional Distributions

Conditional distribution is the distribution over a subset of the whole probability.

|  |  |
| --- | --- |
|  |  |

Other aspects of statistics can be restricted to a subset of the whole probability as well.

* Mean - condition mean, expected value of Y given that X = x
* Variance - conditional variance of Y given X = x

For something like E(Y) or V(Y), there is just mean or variance value for the whole probability.

In contrast, E(Y|X) or V(Y|X) are random variables. The E(Y|X) depends on what kind of restriction we put on X.

If X and Y are independent, then that means knowing X doesn't change the probabilities for Y, leading to:

**The mean and variance via the conditional mean and variance**

It's possible to go from the conditional mean to the mean:

|  |
| --- |
|  |

Note that the inner E(Y|X) is a random variable.

The f(x,y) numbers can be thought of as a table, with a particular "x" being a particular row of that table. The inner E(Y|X) summarizes a particular row of this table. The overall mean is a weighted average of all the E(Y|X) for each row of the table.

This is also called "**Law of total expectation**".

Knowing both the conditional mean and conditional variance can produce the variance:

|  |
| --- |
|  |

This is known as the "**Law of Total Variance**".

It's not obvious at first glance, but this law is the result of repeatedly applying just two other laws:

Definition of variance:

The law of total expectation just mentioned above:

Start with:

Apply

Apply

Distribute the very first "E()"

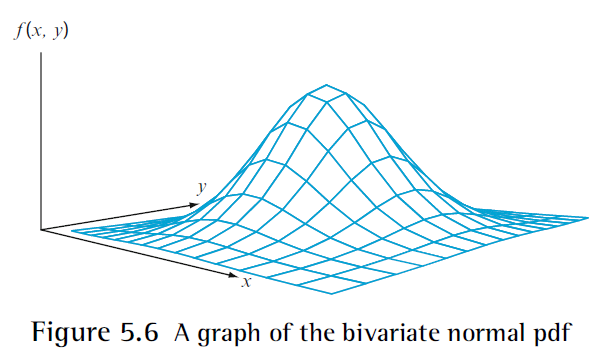
Recognize that the last two terms is actually a variance:

### Bivariate Normal Distribution

|  |
| --- |
|  |

Where:

(µX, σX) are the mean and standard deviation of X  
(µY, σY) are the mean and standard deviation of Y  
ρ is the correlation between X and Y



Discussion:

The f(x,y) is shaped like a hill, or a normal distribution rotated about its vertical axis.

The horizontal cross sections are ellipses centered at (x, y) = (µX, µY). This is because to have constant f(x,y), we would set the exponential expression to a constant.

The vertical cross sections are proportional to normal distribution densities.

If ρ=0, then f(x,y) can be separated into fX(x) \* fY(y). So for this particular distribution, ρ=0 does imply independence. It's still important to note that in general ρ=0 does not imply independence.

The marginal distribution is a normal distribution with mean µ1 and standard deviation σ1.

The conditional density is a normal density with mean and variance:

The conditional distribution would be one of the vertical cross sections of the f(x,y). The mean does depend on X - if the horizontal cross section is a rotated ellipse. The variance does not depend on X.

When ρ is close to 1, which means X and Y are close to a linear relationship, the σY|X=x will be nearly zero. In this case the V(Y|X=x) is very small, much smaller than V(Y), because X almost predicts the Y. Alternatively, if ρ is close to 0, then V(Y|X=x) will be close to V(Y), and knowledge of X does not help to predict Y.

**Regression to the Mean**

The conditional mean can be re-expressed as:

Suppose ρ=0.5. So if a particular X is 2 standard deviations above the mean, then on average Y is only 1 standard deviation above the mean. This is noticed by Francis Galton in biology, where tall parents tend to produce taller kids, those kids also tend to revert to the mean in terms of height.

## Transformations of Random Variables

Previously, the transformation from one random variable to another can be accomplished by what I call the "differentiate first, then substitute" procedure:

If there are multiple random variables, then there will be multiple derivatives.

Definitions:

Input random variables X1, X2  
Output random variables Y1, Y2  
Mapping functions from X to Y: and   
Mapping functions from Y to X: and   
The pdf of the two original variables: f(x1, x2)  
The pdf of the two new variables: g(y1,y2)  
Region S of positive density for f(x1, x2): S = {(x1, x2) : f(x1, x2) > 0}  
Region T of positive density for g(y1,y2): T = {(y1,y2) : g(y1,y2) > 0}

To find the pdf g(y1,y2) given the above definitions:

|  |
| --- |
|  |

Note that the target region "T" is not necessarily the same shape as the source region "S".

Sometimes the det(M) is written as:

If there are three random variables: X1, X2, and X3, then the matrix has dimensions 3x3. The entry in the i-th row and j-th column becomes .

## Order Statistics

**Definition of the order statistics Y1, Y2, ... Yn**

Suppose that X1, X2, ... Xn is a random sample from a continuous distribution with cumulative function F(x).

The **order statistics** are random variables Y1, Y2, ... Yn defined by:

Y1 = the smallest among X1, X2, ..., Xn  
Y2 = the second smallest among X1, X2, ..., Xn  
.  
.  
.  
Yn = the second smallest among X1, X2, ..., Xn

Examples:

When "n" is odd, the sample median is Y(n+1)/2

The sample range is Yn - Y1

For n=10, the 20% trimmed mean is

**The CDF of Yn**

F(y)

y

The probability of a single value being less than "y" is F(y), which is the shaded region in the graph. If this happens "n" times, then the largest value is "y".

**The CDF of Y1**

y

Simply [1-F(y)]n is the probability of getting value larger than "y" for "n" times.

The CDF is defined as the function of probability smaller than "y", as in CDF = P(Y1<y)

**Joint Distribution of the n Order Statistics**

Discussion

Suppose we are sampling three random numbers: X1, X2, and X3. The joint pdf of this random sample:

The total probability of picking this (x1, x2, x3) group of numbers is a product of individual probabilities.

The order statistics Y1, Y2, Y3 is the ordered version of X1, X2, and X3. In the Y world, the (y1, y2, y3) is ordered as in y1 < y2 < y3 , but in the X world they are not required to be ordered. There are 3! = 6 ways to order the three numbers in the X world. That is why:

**Distribution of a Single Order Statistic**

Discussion:

Suppose we sample a random variable 6 times. The pdf of the ordered statistics:

To get the statistics for just the 3rd value, Y3, you need to integrate out: y1, y2, y4, y5, and y6.

Note how the limits of integration are chosen. The dy2 and dy4 are the two outermost integrals that are done last. For each variable, the limit of integration is a constant (∞) on one end and the next integration variable on the other end. The integral can be restructure as:

The generalization of the above integral, and its simplification, leads to the formula shown at the beginning of the section. The key to this simplification is that after you completed the inner integral, you get:

Don't expand and multiply things out. Use the power law instead. The f(y2) is the derivative of the [F(y2)-C].

**Intuitive Derivation of Order Statistic PDF**

To get the g(yi), divide the probability graph into three zones:

For each variable before yi, the probability of occurrence is F(yi). There are (i-1) variables before yi, so the total probability of occurrence is [F(yi)]i-1.

The product of the three "p()" functions will give the sequential probability of getting (i-1) variables smaller than yi, 1 variable at yi, (n-i) variables larger than yi. The product is one way to get the right set of numbers, but there are multiple ways. The "X" world is unordered. Out of the whole (x1, x2, ... xn), you are not requiring x1 < x2, for example. You can have [F(yi)] type outcome, followed by [1-F(yi)] type outcome, then followed by [F(yi)] outcome, and so on. You don't need these outcomes to be in a certain order, only that the total number of [F(yi)] outcomes is (i-1).

We can give these outcomes labels:

[F(yi)] outcome labels: A1, A2, ... Ai-1

[f(y­i)] outcome label: B

[1-F(yi)] outcome label: C1, C2, ... Cn-i

You can spell n! words with these labels. This is the total number of ways to order all these labels.

The ordering of say (A1, A2) is considered the same as (A2, A1). Divide by "(i-1)!" to get rid of all the over counting. Similar division needed for the "C" labels.

This is a multinomial probability problem - review that section for clarification. There are "n!" ways to order all the X variables, but this over counts the "(i-1)!" ways to get (i-1) smaller numbers, and also over counts the "(n-i)!" ways to get the (n-i) larger numbers.

Final formula:

**Distribution of Two Order Statistic**

Suppose you sample six numbers, Y1 ... Y6, and would like g(y3, y5). The integral set up is:

Simplify either by using the intuitive ideas described above, or by actually evaluating the integral:

Note that the 6! approach uses purely PDFs and integrals. The approach uses a combination of PDFs and CDFs.

# Sampling Distributions

## Statistics and Their Distributions

**point estimate** - The value of the sample mean from any particular sample is called a point estimate of the population mean µ.

**statistic** - A statistic is a value calculated from sample data. Since the sample of a population is not fully predictable, a statistic is a random variable. Use upper case to represent random variables and lower case to represent calculated values.

- the sample mean random variable  
 - the calculated value of sample mean

- the sample standard deviation random variable  
s - the calculated value of sample standard deviation

**sampling distribution** - A statistic is a random variable, and therefore has a probability distribution called the sampling distribution. This distribution describes how the statistic varies in value across different samples.

**random samples** - This is one method of sampling. The random variables Xi are independent, and every Xi has the same probability distribution. Ideally this means sampling with replacement, or sampling from an infinite population.

If at most 5% of the population is sampled, meaning n/N < 0.5, then the samples are approximately random samples. If you sample too much, then the probability distribution has changed.

**simulation experiments** - Statistic distributions can be obtained by computer simulation. The characteristics of such experiment are:

1. The **statistic** of interest
2. The **population distribution**
3. The sample size "**n**"
4. The number of replications "**k**"

## The Distribution of the Sample Mean

|  |
| --- |
| Let X1, X2, ..., Xn be a random sample from a distribution with mean value µ and standard deviation σ.  Sample mean distribution characteristics:  The "sample total" TO is defined as: TO = X1 + X2 + ... + Xn  Sample total distribution characteristics: |

Discussion:

Variance of independent variables can be added up (see next section). That means .

The variance gets smaller for but larger for .

Averaging will move the probability towards the middle, whereas totaling spreads the probability out over a wider range of values.

**for Normal Population Distribution**

Let X1, X2, ..., Xn be a random sample from a normal distribution. Then for any n, is normally distributed.

The is closely related to TO, so TO is also normally distributed.

**Central Limit Theorem**

Let X1, X2, ..., Xn be a random sample from any distribution. Then and TO are asymptotically normal - they converge to the normal distribution as "n" gets larger.

For "sufficiently large n", and TO has an approximately normal distribution.

So when is "n" sufficiently large to apply the approximation? The accuracy of the approximation depends on the shape of the original underlying distribution being sampled. If the underlying distribution is symmetric, then the approximation will be good for even a small "n". If the original distribution is skewed, or if there is a lot of probability in the tails, then a larger "n" will be required. The extreme situation is that a distribution can have so much probability in the tail that the mean will fail to exist and the Central Limit Theorem does not apply, so no "n" is large enough.

Rule of thumb - If n > 30, the Central Limit Theorem can be used.

**The normal approximation to the binomial distribution**

The Central Limit Theorem justifies approximating the binomial distribution using the normal distribution.

The binomial experiment is defined as:

The binomial distribution is the number of successes. This is the sum

The sample total will converge to the normal distribution if the underlying distribution, the Xi, is symmetric. So the normal approximation is best when success and failure have equal probability, which is p = 0.5. If "p" is near 0 or 1, then "n" needs to be larger. That's why the criteria for applying the normal approximation is and . This way, both small and large "p" will be required to have a large "n" value.

**The origin of the log-normal distribution**

Start with random variables X and Y:

So ln(Y) is a sample total, and regardless of the distribution of X, ln(Y) will have a normal distribution. Y's distribution is therefore log-normal, no matter what distribution X has.

**The Law of Large Numbers**

When "n" gets large, the will remain at µ, but the will approach zero. So the whole distribution approaches µ, a single value.

In equations:

Variance is the mean of the squared difference between and µ, so we say that converges in mean square to µ.

Alternative way to state the law:

## The Mean, Variance, and MGF for Several Variables

**Linear Combination**

Let Y be the **linear combination** of Xi's:

The sample mean is a special case of linear combination where

The sample total TO is a special case of linear combination where

We are not requiring the Xi's to be independent or identically distributed. All the Xi's could have different distributions.

|  |
| --- |
| Whether or not the Xi's are independent:  If the Xi's are independent: |

Discussion:

When the Xi's are independent, cov(Xi, Xj) is zero.

, but . So the expected value is being subtracted, while the variance is being summed.

**Proofs for the n = 2 Case**

Start with:

The integrals can then be separated and then simplified:

The expected value's linear combination property is used to prove the variance linear combination property. To start, the variance can be expressed as an expected value:

Rearrange the expression inside the E{...} into (X1-µ1) and (X2-µ2) grouping:

Apply the square:

**Linear Combination of Normal Random Variables**

Linear combination of normally distributed random variables also has the normal distribution.

Generalization of the Central Limit Theorem:

For , if "n" is sufficiently large, then Y has approximately a normal distribution.

**Example: Creating a bivariate normal distribution**

Let Z1 and Z2 be independent standard normal random variables.

Let and

The X and Y are linear combinations of normal random variables, so X and Y are normal distributions. The joint distribution of X and Y is bivariate normal.

X and Y have standard deviation 1. To verify this for Y:

The covariance Cov(X,Y) is:

µX = 0; µY doesn't matter but it's zero too

E(Z12) = E[(Z1-0)2] = variance of (Z1) = 1 by design

Because the Z1 and Z2 are independent, the integral is separable.

In conclusion:

**Moment Generating Functions**

Let X1, X2, ..., Xn be independent random variables with moment generating functions MX1(t), MX2(t), ..., M­Xn(t), respectively. Define Y = a1X1 + a2X2 + ... + anXn.

As long as the X random variables are independent, then the expected value integral can be broken up:

**Example: Adding two normal random variables**

Let X1 and X2 be independent normal random variables, and we want to compute Y = a1X1 + a2X2.

The MY(t) uniquely identifies a distribution. It's a normal distribution, with and .

**Example: Adding two Poisson random variables**

Let X be a Poisson random variable with mean λ, and Y be a Poisson random variable with mean v. We want to find X + Y.

"X + Y" results in a Poisson random variable with mean "λ+v".

## Distributions Based on a Normal Random Sample

Suppose we have five independent standard normal random variables: Z1, Z2, Z3, Z4, and Z5.

Chi-squared distribution:

t-distribution:

F distribution:

### Chi-Squared Distribution

The chi-squared distribution is a special case of the gamma distribution, with α=v/2, β=2. The "v" is a parameter called "number of degrees of freedom".

**If Z has a standard normal distribution, then Z2 is chi-squared with 1 df, .**

Proof:

The pdf for the standard normal distribution:

The mapping that will be applied:

The pdf for mapping the right half of the standard normal distribution:

The full pdf of X = Z2 includes the contribution from the left half of the standard normal distribution:

Finally,

**If X1 ~ χ2v1, X2 ~ χ2v2, and they are independent, then X1 + X2 ~ χ2v1+v2.**

Notation note: the χ2 denotes the Chi-squared distribution (to the power of 1). It's not really raised to the power of 2.

Proof:

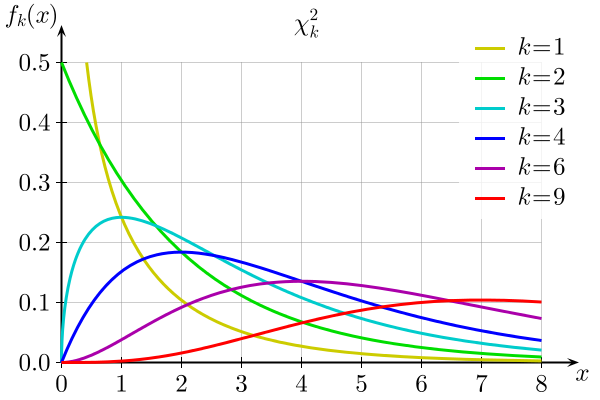
Use moment generating functions.

Let Z1, Z2, ..., Zn be "n" independent standard normal random variables:

where each standard normal square contributes one degree of freedom to the Chi-squared result.

So degree of freedom can be thought of as the number of standard normal random variables that it took to build up to a Chi-squared random variable.

**The peak of fχ2(x) is at**



Proof

For v = 2, the peak is at 0. It's a monotonically decreasing curve.

For v > 2, find the "x" where the derivative df(x)/dx = 0.

The solution will be x = v - 2.

Note the curve is being dragged to the right as v increases.

When v = 2, the peak is at 0, but the mean is clearly at some location above 0 - the distribution is actually exponential with mean 2. As more Chi-squared curves get added on, the curve gets dragged to the right - because that's where the data is - the mean is always to the right of the peak.

If you add enough random variables together, then it's going to converge to a normal distribution - the Chi-square distribution included.

**If X1, X2, ..., Xn are a random sample from a normal distribution, then and S2 are independent.**

Proof:

Start with: . The goal is show that is not correlated with the deviations .

This distributes out to be:

The second term is:

The first term simplifies to be: . So the covariance is zero. The simplification process is explained below:

Of all the covariance terms, only Cov(Xi, Xi) is non-zero because the X's are independent random variables.

For this proof I had help from: <http://www-ma4.upc.edu/~jfabrega/pipe/sample-mv-handout-4pp.pdf>

**If X1, X2, ..., Xn are a random sample from a normal distribution N(µ, σ2), then .**

Definitions:

Proof:

Start with:

Need to change this into:

The middle term is zero because is zero.

The last term is just the same thing happening for "n" times.

To match with the Chi-squared distribution, divide by σ2 on both sides:

The term on the left is sum of (standard\_normal\_distribution)2, so it's χ2n.

For the second term on the right, the has a mean of µ and a standard deviation of . So the second term on the left side is χ21.

For the first term on the right, refer to the definition to see that it is .

Altogether we have

which is

For this proof I had help from: <https://onlinecourses.science.psu.edu/stat414/node/174>

### t-Distribution

**The definition of t-distribution:**

Let be a standard normal random variable.  
Let be a Chi-squared random variable independent of "Z", with degree of freedom.

The random variable "T" will have a t-distribution.

**The t-distribution is standardized**

If X1, X2, ..., Xn is a random sample from a normal distribution , then

has the t-distribution with (n-1) degrees of freedom, tn-1.

This is part of a more general result:

**Proof**

Start with

The key is in the denominator. There is a prior result that says

So dividing a random variable by will get a random variable. The degree of freedom is the number of data points used to compute the minus one.

The quantity

is standard normal distribution, so we write

Dividing this by will produce a random variable.

There's another application of this principle in section 8.3.

**The PDF of the t-distribution**

Proof:

Start with

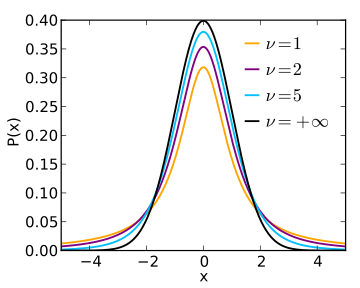
To figure out the limit of integration:

The f(x,z) is product of two pdf, the Chi-squared f(x) and the standard normal distribution f(z).

The fχ(x) is the Chi-squared pdf. To finish the proof, make the integral part look like a gamma distribution. Eventually it will look like:

The integral is the gamma distribution and it will evaluate to 1.

**Shape of the t-distribution PDF**



The PDF is similar in shape to the normal distribution PDF.

When v=1, the pdf is . This is called the Cauchy distribution and the tail is so fat that the mean does not exist.

As "v" approach ∞, the t-distribution approaches the standard normal distribution.

Proof using the definition :

The term on the right is the sample mean of the Z2 distribution.

When is large, the sample distribution will be bell shaped curve, with mean and variance . As gets really large, the variance goes to zero, and the distribution collapses to a single point, the E(Z2).

This E(Z2) = 1. One way to figure this out is to look up the expected value of the Chi-square distribution. Another way is that E(Z2) - E(Z)2 = variance of standard normal = 1. The E(Z) = 0, so E(Z2) = 1.

Since for large "v", "T" will equal "Z" for large "v".

**Mean of t-distribution**

This is because , and due to "Z" and "X" being independent. E(Z) is zero, so that means E(T) is zero.

**E(Xk) of chi-squared distribution**

You need to turn the integral into a gamma distribution

Final result:

**Variance of t-distribution**

We need E(T2).

Note it is NOT

Use the result with to get

You will need to use:

Final result:

The variance always exceeds 1 - since the t-distribution is fatter than the normal distribution. It will converge to 1 for large "v", as the t-distribution itself converges to the normal distribution.

### F Distribution

The F distribution describes ratio of chi-squared random variables.

X1 has v1 degree of freedom, and X2 has v2 degree of freedom.

The chi-squared distribution has the property:

dividing by the degree of freedom remove the (n-1) in the numerator

If "v" is large, then "X/v" will approach 1 as explained in the "Shape of the t-distribution PDF" section.

**Reciprocal of F**

The notation means .

Using this notation, the leftmost quantity from above is .

The rightmost quantity from above is , meaning .

So in general

**t-distribution squared is F distribution**

## Proof of the Central Limit Theorem

Start with:

and we need to prove that Y's distribution approach the standard normal distribution.

1. Expand
2. Use a substitution to replace the "X" with a standardized random variable W
3. Get the MGF for Y

The random variables are independent so the expression is separable:

So the should really be . It's called to be consistent with the textbook.

1. The limit that we need to show

where the is the MGF for the standard normal distribution.

Take logarithm of both sides:

1. Substitute to simplify the limit slightly
2. Understand what is M(0), M'(0), and M''(0), so to evaluate the limit.

, because of , , and .

1. Apply L'Hopital's rule, for the first time.
2. Apply L'Hopital's rule, for the second time.

# Point Estimation

## Concepts and Criteria

**Estimators**

θ - this is the statistics that we are estimating  
 - this is the estimator of θ, or the point estimate of θ

**Example**

is read as "the point estimator of the population mean is the sample mean ".  
 is read as "the point estimate of is 5.77".

For the population mean there are multiple estimators.

= First discard the smallest 10% and largest 10% of the sample, and then average.

**Error**

**Mean Squared Error:**

The mean square error can be computed using the formula for computing variance:

simplifies to due to the θ being a constant

simplifies to

|  |
| --- |
| variance of estimator  (bias)2 |

**Example - Estimating the from binomial distribution**

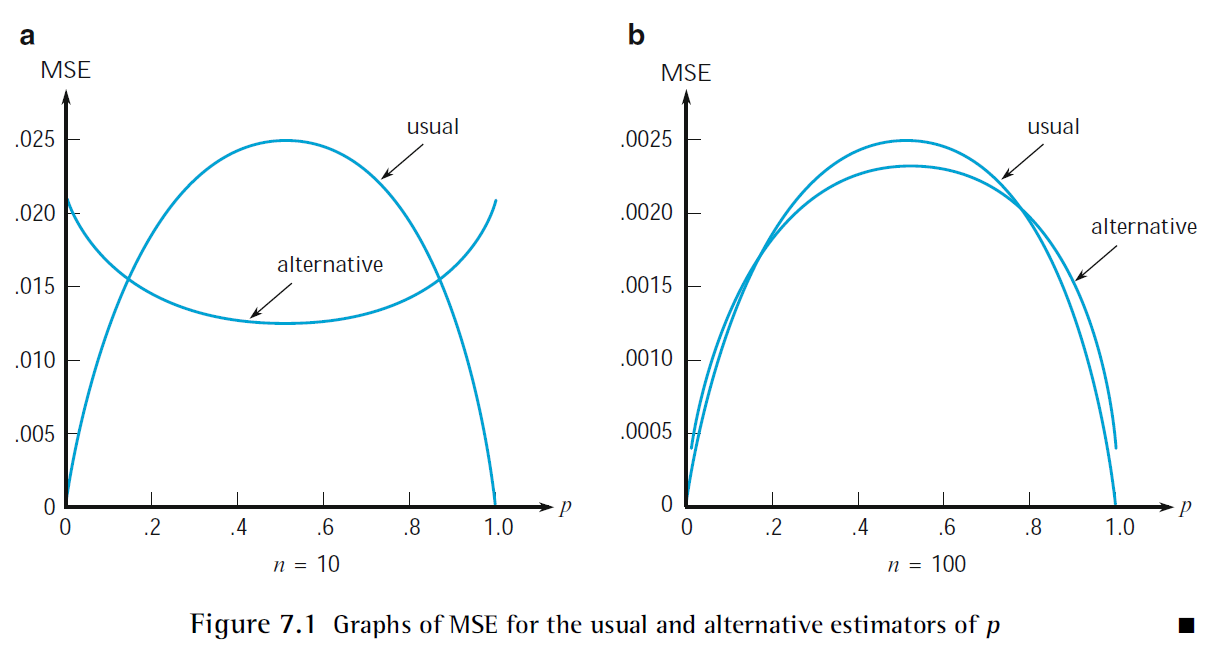
Suppose is the number of successes from a binomial distribution, the natural estimator of success probability is .

If is near 50%, then the estimator artificially pulls towards 50%, and so that estimator will achieve a lower MSE.

Mathematically:

For the binomial distribution, .

These two MSE expressions can then be graphed to show that when is small and is close to 50%, the estimator has a substantial advantage in terms of a lower MSE.

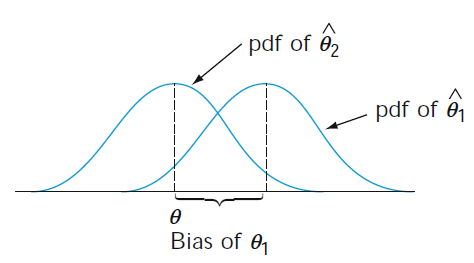


**Unbiased Estimator**

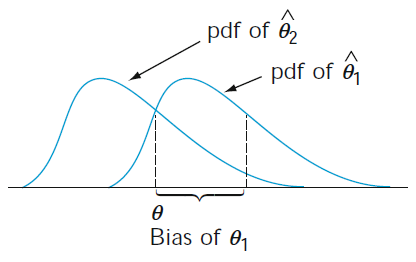
Un unbiased estimator has the property .

The amount of bias is .

Graphically, an unbiased estimator is "centered" at the true value. If θ = 100, then an unbiased estimator will have a distribution that is centered at 100. The distance between the estimator's expected value and the population's true value is the bias.



Being "centered" means only . The probability distribution need not be symmetrical.



**Example - Estimating the upper limit for the uniform distribution**

Suppose that is a random variable that has a uniform distribution from 0 to an unknown upper limit .

The estimator is a biased estimator because this estimator will never over estimate . An un-biased estimator will sometimes over estimate to balance out underestimating it at other times.

To compute , start with the CDF of .

An unbiased estimator would be

**Example - Unbiased estimator for the variance**

Suppose we are estimating the variance using the estimator

What value of will produce an unbiased estimator?

We want to compute

To simplify :

Use to get rid of the term.

To simplify :

Final answer:

To produce an unbiased estimator, use .

Note that even if , that doesn't mean .

So an unbiased estimator for the variance will be a biased estimator for the standard deviation.

If the random variable X comes from a normal distribution, then , where is the chi-squared distribution, and

The estimator is then

The variance of the estimator is

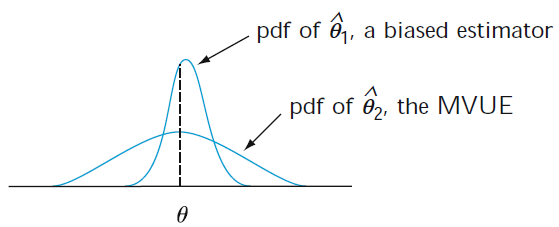
If is a random sample from a distribution with mean , then is an unbiased estimator of . If in addition the distribution is continuous and symmetric, then the median and any trimmed mean are also unbiased estimators of .

**Estimators with Minimum Variance**

**MVUE - minimum variance unbiased estimator** - Often there are multiple unbiased estimators. The MVUE is the estimator that is both unbiased and also have the minimum variance.

MSE = variance + (bias)2, meaning minimizing the variance will minimize MSE, for all existing unbiased estimators.

The MVUE does not necessarily lead to the lowest MSE - it's just the unbiased estimator that has the lowest MSE. A biased estimator can get an even lower MSE.



**Example - Variance of an estimator**

Suppose is a random variable from a uniform distribution .

In a previous example, the estimator is shown to have the PDF

and the expected value

An unbiased estimator is therefore

The variance of this unbiased estimator is

The since it's an unbiased estimator.

The is computed as

Note the PDF for is used, but it's NOT .

The final answer for the variance is

**How the PDF Affects the Estimator**

For a normally distributed random variable , the MVUE for is .

In general, the best estimator for µ will depend on the distribution that is being sampled.

The estimator is not good for distributions that have fat tails, such as the Cauchy distribution. In these cases, using the median as an estimator can be quite good.

The trimmed mean, with trimming proportions of 10% or 20%, is a **robust estimator** that works well across many different distributions.

**Example - estimating the of the exponential distribution using censoring**

Suppose a component has a lifetime distribution that is exponential with parameter λ. The estimator would be to run all "n" components until they fail and measure the life time.

**censoring** - this is an alternative procedure where we don't wait for "n" components to fail. Instead we just run until "r" components fail, and then average their lifetimes.

is an unbiased estimator for , which is for the exponential distribution.

Let Y1 be the shortest time, and Yr be the longest time. All "n" components will last for Y1 time, and "n-1" components will last between Y2-Y1 time, and so on. The Tr is then re-written as:

To compute we need , , and so on.

The random variables come from the exponential distribution. The CDF is

The is the minimum, meaning that components last longer than .

This is the CDF of an exponential distribution. .

Let . The event is that variables are of the quantity . The reasoning is exactly the same as and leads to

This pattern will hold, with .

Using just "r" components, instead of all "n" components will allow the experiment to be done faster. The estimated mean is unbiased. However, the variance of the "r" estimator will be higher than the "n" estimator.

**Standard Error**

The **standard error** of an estimator is its standard deviation .

If the standard error itself uses estimated parameters, then it's called **estimated standard error**, denoted by (with the "hat" over the σ) or by .

**Example**

Suppose follows the binomial distribution, and is being estimated by .

The variance of this estimated is

The "p" is an unknown parameter, so the notation becomes:

When "n" is large, the estimator will often be approximately normal. We can be confident that the true value of θ will lie within two standard errors.

When the estimator is not normal, but is unbiased, then the true value of θ can be as much as 4 standard errors away at most 6% of the time - again this is very conservative statement that applies to any unbiased .

**Bootstrap**

This is to obtain the by computer simulation.

First bootstrap sample: ; estimate =   
...  
Bth bootstrap sample: ; estimate =

The bootstrap estimate of 's standard error is the sample standard deviation:

If the f(x) is unknown, treat the existing data as the population. The existing data is sampled with replacement.

## Methods of Point Estimation

**The Method of Moments**

The idea is to express the population parameters in terms of , , and so on.

**population moment** - Each population distribution will have its own moments like E(X), E(X2).  
**sample moment** - Each set of data can be used to generate , , and so on.

**Example - Fitting data to the gamma distribution**

Let X be a random variable having the gamma distribution. The for the gamma distribution is

There are two parameters, so is needed. To get , use the MGF for the gamma distribution.

The and equations

together form a system of two equations with two unknowns: .

Expressing in terms of and gives

In terms of the sample data, this is

**Maximum Likelihood Estimation (mle)**

Given the joint pdf , this approach chooses the (θ1, ..., θm) to maximizes the pdf.

This function of is called **likelihood function**.

The mle is approximately the MVUE of θ. It is approximately unbiased , and has low variance. The mle is in general preferred over the moment estimators.

**Example: Estimating binomial distribution**

Suppose we observed 1, 0, 1, 0, 0, 0, 0, 0, 0, 1 - for a total of three failures out of 10 events. The joint PDF is

Maximizing "f" is the same as maximizing "ln(f)":

Take the derivative

This derivative is zero at .

**Example: Estimating exponential distribution**

The likelihood function for a random sample is

Maximizing "f" is the same as maximizing "ln(f)":

The derivative

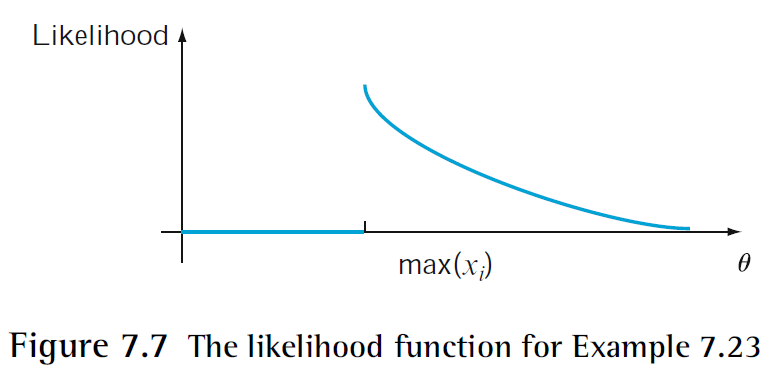
is zero at

The conclusion is

**Example: MLE of uniform distribution**

Suppose are sampled from a uniform distribution that range from 0 to . The likelihood function is just

At first glance, there is no mentioning of any of the sample values in this expression. Actually, the is restricted to .



So the mle .

**Example: Estimating the hypergeometric distribution**

Suppose given the hypergeometric pmf

where M, n, and x are all known - the is unknown.

To maximize the function, derivatives won't work here due to the discrete nature of the pmf. Instead, consider the ratio

When this ratio is > 1, the "p" function is growing. The ratio eventually simplifies to:

The numerator and denominator are the same except for that very last term. The ratio will be greater than one as long as .

The estimator for N:

## Sufficiency

**Overview**

A statistic will be useful for drawing conclusion about if the distribution of depends on

**Example: provides no information about the of**

Suppose and come from a normal distribution with mean and variance .

Let . The distribution of is a normal distribution with mean and variance .

Knowing is only useful for predicting and it's not useful for predicting , since the distribution of doesn't contain .

**Example: Predicting Poisson distribution's from data - the operational view**

Suppose we know is Poisson distribution, and has three pieces of data .

for Poisson, so we estimate Poisson by

It doesn't matter if the data is or - we always add it up and divide by 3 to get an estimate of . If we know the data is Poisson and , then only the sum really matters. This is the operational reason why the total is a sufficient statistic for .

This is not to say that the data doesn't matter. The data matters in that it determines the sum. For the purpose of estimating , the sum of the data already captured all the information.

**Example: Predicting Poisson distribution's from data - the probability view, which is used in the definition of sufficiency**

Suppose the data is . We would like to compute

is just a direct application of the Poisson PMF

For the sum of Poisson variables, use the MGF for Poisson

and the equation about linear combination of random variable MGFs

The is the same as except that became .

Each MGF corresponds to a unique PMF. The PMF of is

The final numerical answer about encountering given that is

This probability expression has no in it. That means that after knowing sum = 4, knowing the exact data value does not help in knowing more about .

In other words, the joint PMF of is not a function of .

Alternatively, had we gotten some result like:

then we might wonder that even though the sum is 4, different data values has something to do with .

**Definition of Sufficiency**

A statistic is said to be **sufficient** for making inferences about a parameter - if given that , the joint distribution of does not depend upon .

**Neyman Factorization Theorem**

Let denote the joint pmf or pdf of . Then is a sufficient statistic for if and only if the joint pmf or pdf can be represented as a product of two factors

Note the function references the data only in the context of the statistics .

The function is not allowed to reference the statistic .

**Example: Factorizing the Poisson distribution with parameter**

Note that the first factor uses the data only in the context of the , and the second function does not use .

This shows that is a sufficient statistic, and it's much easier than a previous example which did the same thing using the probability based definition of sufficiency.

Any one-to-one function of a sufficient statistic is itself sufficient. So the mean is a sufficient statistic as well.

**Example: Factorizing the uniform distribution with upper limit**

Suppose we think that is a sufficient statistic for predicting . The problem is how do we introduce this expression into the equation.

Define an indicator function

This indicator function allows adding the expression.

The first factor uses the data only in the context of the statistics function , and the second factor does not use the parameter .

This proves that is a sufficient statistic for predicting .

**Why the factorization works - discrete case**

Returning to the Poisson statistic example. The example uses a total of four, but there can be other totals. For each total, say four, there are multiple data sets, say and . So the possibility of a particular data set can be viewed as a two-step event - first you choose the total, then you choose one of those data sets.

The is the possibility of a certain statistic value happening, and then is the possibility of a certain outcome happening, given that we are using the statistic .

The expression should reference the data in terms of the statistic only - that this portion of the total probability is not dependent on the actual value other than the statistic.

The expression should not reference the parameter that is being estimated, because at this stage of the total probability, the doesn't matter.

The above restrictions are in line with the definition of sufficient statistic.

**Jointly Sufficient Statistics**

This is an expanded notion of sufficient statistics, for statistical models that involve more than one parameter.

Suppose the joint pmf or pdf of the data involve unknown parameters . The statistics are .

The statistics are said to be **jointly sufficient** if the conditional distribution of the given that does not depend on any of the unknown parameters.

The factorization theorem extends to the jointly sufficient statistics. One expression should include the and refer to the data only through the statistics. The second expression does not involve the .

**Example - random sample from a normal distribution**

The data is only referred to in terms of and , so together they form a sufficient characteristic.

uniquely determines the sample mean .

, so and the sample mean uniquely determines sample variance.

The conclusion is that the sample mean and sample variances are sufficient statistics as well.

**Improving an Estimator**

Given:

A joint distribution of depends on some parameter   
 is a sufficient statistic for   
We are trying to estimate   
We already have an unbiased statistic that does not involve

Then:

The estimator is also unbiased for and has variance no greater than the original estimator .

**Reasoning**

The does not involve - because (U|T) by does not involve , by the definition of being a sufficient statistic.

has lower (or equal) variance than due to the Law of Total Variance

**Example - Poisson distribution estimator**

Given follows Poisson distribution.

We are trying to estimate , which is .

Start with the estimator . In other words

First to show that is an unbiased estimator

From a previous example, a sufficient statistic for the Poisson distribution is , so the new estimator is .

Note the summation indexing normally starts at 1. In the numerator case, must be zero, so the indexing on the summation starts at 2.

In an example in the start of this section, the sum of Poisson random variables have been shown to have a Poisson distribution with parameter .

So the new estimator is .

If , then

If using mle, then ,

So different estimators do give different results, because the principles of unbiasedness and maximum likelihood are in conflict. When is large, these different estimators should converge.

In general there will not be a unique sufficient statistic. So it's possible to get .

**Maximum Likelihood Estimator using only Sufficient Statistics**

The factorization theorem says that

The maximum likelihood estimate maximizes with respect to the parameters. This is the same as maximizing , since there are no variables to tweak in . Therefore, it is always possible to have a maximum likelihood estimator using only sufficient statistics.

## Information and Efficiency

**Fisher Information**

**Definition**

Differentiating the happens often in maximum likelihood estimation.

**Additional Properties**

An alternative formula for , when is hard to calculate, is

**Derivation of**

Use

**Derivation of the alternative equation for**

, and , so

The goal is to get .

Start with

Take derivative of both sides

Use to get

**Example: Computing for Bernoulli rv**

So the information is the reciprocal of .

The information is greatest when the variance is smallest.

**Information in a Random Sample**

The Fischer information in a random sample is just times the information in a single observation.

**Derivation**

When there are observations, the pmf or pdf is

The total information is then

**The Cramer-Rao Inequality**

Given:

Random sample with pmf or pdf

Statistic is an unbiased estimator for the parameter .

Then:

**Proof**

The basic idea is to consider the correlation between and , and use the inequality .

Start with being an unbiased estimator, meaning

Left side is 1

**Efficient Statistic**

The variance of must be at least .

is an **efficient** estimator if . The ratio of this lower bound to is the **efficiency** of .

An efficient estimator is a minimum variance unbiased estimator (MVUE).

**Example: Binomial distribution's MVUE**

From an example in the beginning of this section, for a Bernoulli random variable,

Let be an estimator of , where

The variance of has a lower bound:

The expected value of is

is binomial distribution, with

So T is unbiased.

The variance of is

Since is unbiased and has minimum variance, is a MVUE.

**Large Sample Properties of the MLE**

The maximum likelihood estimator is **consistent**, meaning it converges in probability to .

The mle is asymptotically normal with mean , meaning **asymptotic unbiasedness**, and variance equal to the Cramer-Rao lower bound, which is .

**Proof**

Start with the usual maximum likelihood score function:

An approximation for any derivative is

The score function is a derivative that should be zero at due to the maximum likelihood requirement.

Expand the to

The Central Limit Theorem can be used to say that will be a normal distribution as .

Due to

shows asymptotic unbiasedness .

shows the variance is at the Cramer-Rao lower bound when is large.

**Example: Large Sample MLE of**

The PDF is

The is

The maximum likelihood score function is

The maximum likelihood estimator for is

When is large, the value of is

To get , start with

Returning to the large behavior of

which is what the theorem predicted - that for large , will converge to .

Using the theorem, is

# Single Sample Statistical Intervals

## Basic Properties of Confidence Intervals

**95% Confidence Interval based on the Normal Distribution**

If we sample numbers, the mean of those numbers will follow the normal distribution, per Central Limit Theorem.

Given:

- number of data points  
 - the sample mean  
 - the population standard deviation

Equations for the 95% confidence interval:

Ways to express the interval:

Note the is actually constant, although its value is unknown. The is the random variable. If many experiments are done, the will stay in one place while the interval end points will move around.

**Other Levels of Confidence**

For other confidence intervals, replace the 1.96 with other Z scores.

For confidence interval:

Let be the interval width:

Given a certain width, the number of samples required would be

The **bound on the error of estimation** is defined as .

**Generalization**

Consider

The random variable is a function of , yet its probability distribution has a mean of zero - so the disappears in the probability distribution of . This is the key to enabling these confidence intervals.

More generally, the random variable satisfies two properties

1. The variable depends functionally on both and .
2. The probability distribution of the variable does not depend on or any other unknown parameters.

Satisfaction of these properties allow us to write

The key is to have values and that do not depend on . The above equation can then be re-arranged into

**Example - Confidence interval based on the chi-squared distribution**

Given:

Random variable has an exponential distribution, with parameter .

Let

In conclusion, has a chi-squared distribution, with

To get the chi-squared limit values, use

import scipy.stats as stats

stats.chi2.ppf(0.025, 20) # 9.591

stats.chi2.ppf(0.975, 20) # 34.170

The confidence interval is

Note the chi-squared distribution is not symmetrical. The 50% possibility is not in the middle.

stats.chi2.ppf(0.5, 20) # 19.337

In this example,

Interval for is

There are multiple ways to estimate the "center" value for .

Simply using gives 55.1

Using the chi-square 50% probability point gives 57.0

## Population Proportion

**Population Proportion**

Let the random variable be the total number of successes. This means is a binomial distribution, with the parameter .

An estimate for the proportion is

**Variance of Population Proportion**

The variance of the proportion is

Another approach to obtain the is to view the random variable as an individual outcome, so that each is a single Bernoulli trial.

In this view, the is , with .

**Distribution of Population Proportion**

When is large, the will be normally distributed.

**Solution of the Population Proportion Confidence Interval**

An approximation would be to use in the denominator, and get

The book recommends the exact solution, reasoning that the performance of the above approximation is not consistent.

For the exact solution:

The , , and are all specific values. So the equation of is a quadratic equation and this results in messy formulas. I have added a Python script for obtaining a numerical solution using the bisection method.

**Example 8.8 - Population Proportion Confidence Interval**

The above is solved numerically using the Python script

solve(lambda p: (1.96\*\*2) \* p \* (1-p) / 48,

lambda p: (0.3333-p)\*\*2,

0, 0.3333)

solve(lambda p: (1.96\*\*2) \* p \* (1-p) / 48,

lambda p: (0.3333-p)\*\*2,

1, 0.3333)

The answer is .

**Estimating the number of samples needed for a certain interval width**

The width of the confidence interval is

The quantity is maximum if

So a conservative estimate of the number of samples needed for interval width would be

## Sample Mean of a Normal Population

The sample mean follows the normal distribution only when the number of samples is large. The Central Limit Theorem cannot be invoked when is small.

**A Review of the t-Distribution**

If the population distribution is normal with mean , then the random variable

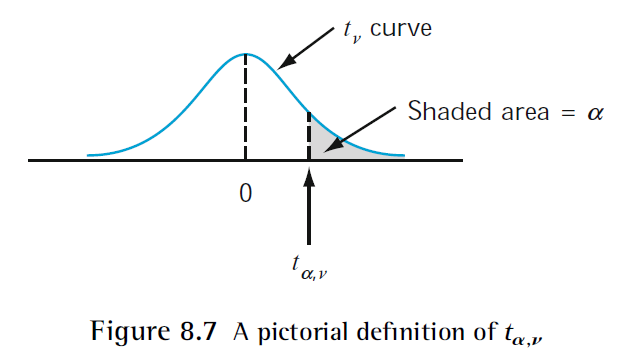
has the t-distribution with degrees of freedom (df).

A t-distribution is governed by only one parameter, the number of degrees of freedom.

Properties of the t-distribution:

1. Each curve is bell-shaped and centered at 0.
2. Each curve is more spread out than the standard normal curve.
3. As increases, the spread of the curve decreases.
4. As , the sequence of curves approaches the standard normal curve.

**t critical value:**



The confidence interval for a sample mean that follows the t-distribution:

**Prediction using**

Suppose that we have a random sample that came from a normal population distribution with parameters and .

To predict we use .

The prediction error is

is a linear combination of normally distributed random variables, so is normally distributed. It can be standardized into

A result from the t-distribution section says

The follows the t-distribution. The standard deviation is calculated from , totaling points. The degree of freedom is therefore .

The **prediction interval (PI)** is

Interpretation of the 95% PI - If the interval is calculated for sample after sample, then 95% of these intervals will include the corresponding future values of .

The is much wider than the . The reason is due to

With the sample mean, the eventually goes zero if is very large. With the prediction interval (PI), no matter how large is, the variance is still least .

**Tolerance Interval** - For example, an interval that contains at least 90% of the population values with a confidence level of 95%.

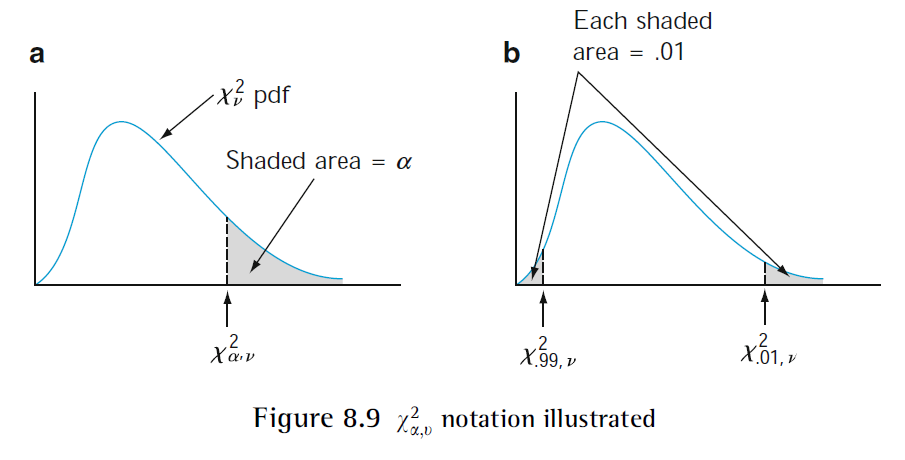
## Variance of a Normal Population

**Review of chi-squared distribution**

Let be a random sample from a normal distribution with parameters and . Then the random variable

has a chi-squared probability distribution with df.

**chi-squared critical value:**



The chi-squared distribution is not symmetric, so the left and right ends of the confidence interval needs to be tracked separately.

**Confidence interval for estimate of a normal population**

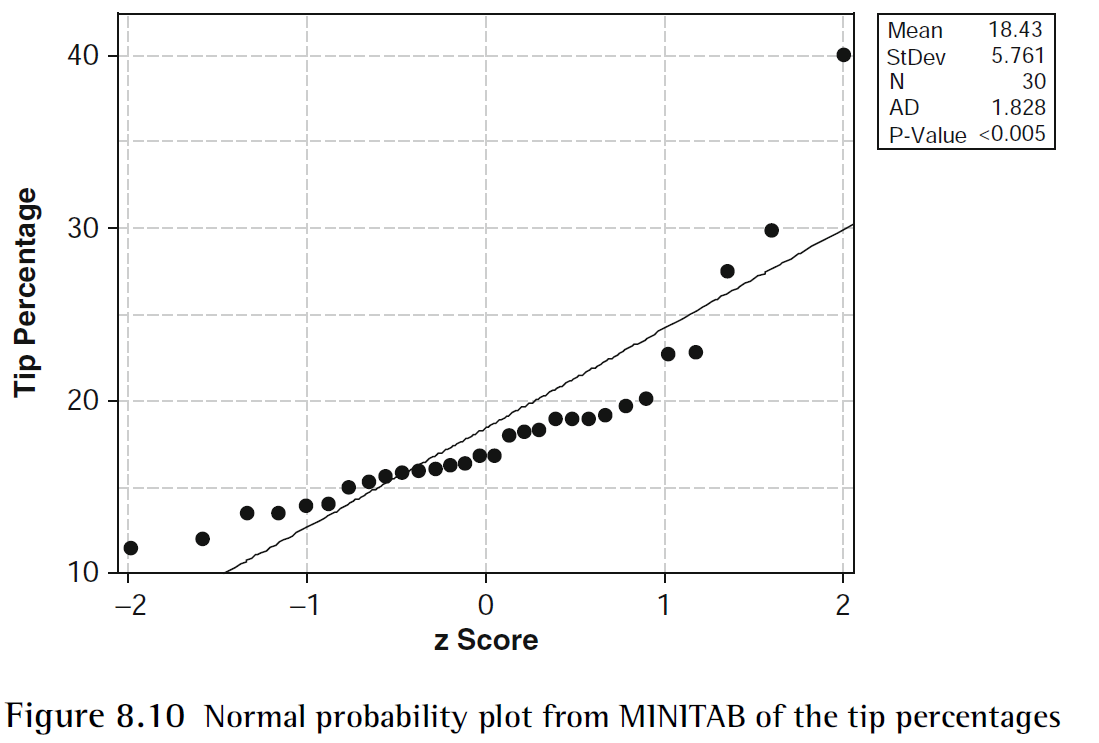
## Bootstrap Confidence Intervals

The strategy here is to get confidence intervals via simulation. This way the procedure will apply regardless of the underlying population.

The sampled data is treated as a population, and new "bootstrap samples" are created from this "population". New data are sampled with replacement - so the data in these "bootstrap samples" will contain repetitions.

**Example 8.15**

There's a random sample of 30 observed tip percentages. Normal probability plot of the data:



The goal is to estimate an interval for the sample mean. The sample size, 30, is small so it's not suitable for the central limit theorem.

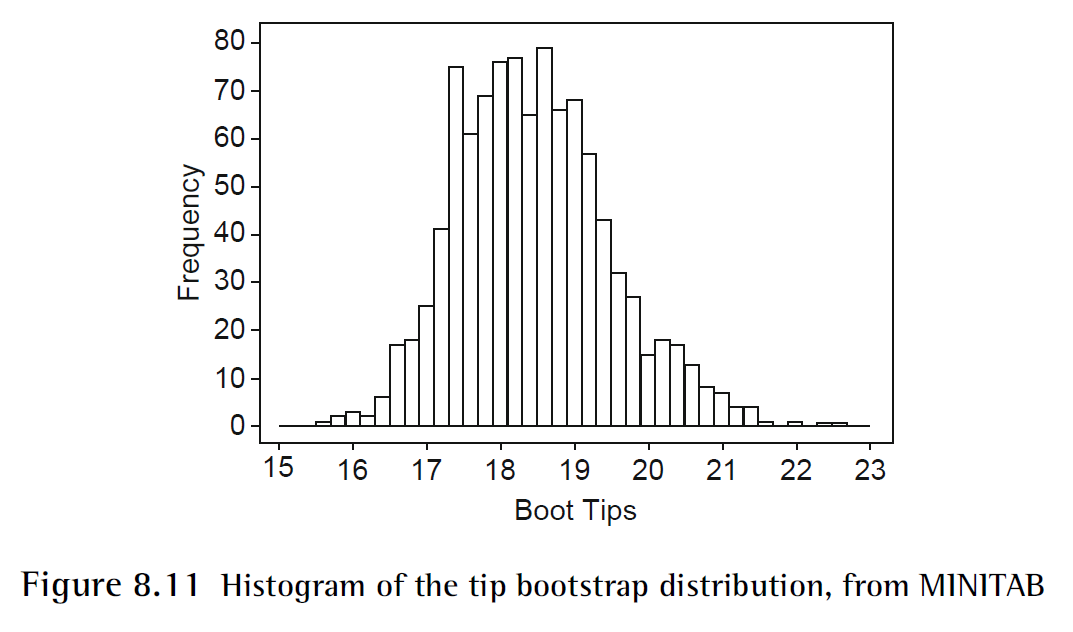
If the sample data follows the normal distribution, then the data points will fit onto a straight line. The sample data leads us to believe that the population is not normal, so the t-distribution should not be used.

The bootstrap approach takes a large number of random samples, with 999 being a common choice. In each of these 999 random samples, 30 numbers are selected from the sample of 30 numbers. This sampling is done with replacement, so repetitions are allowed.

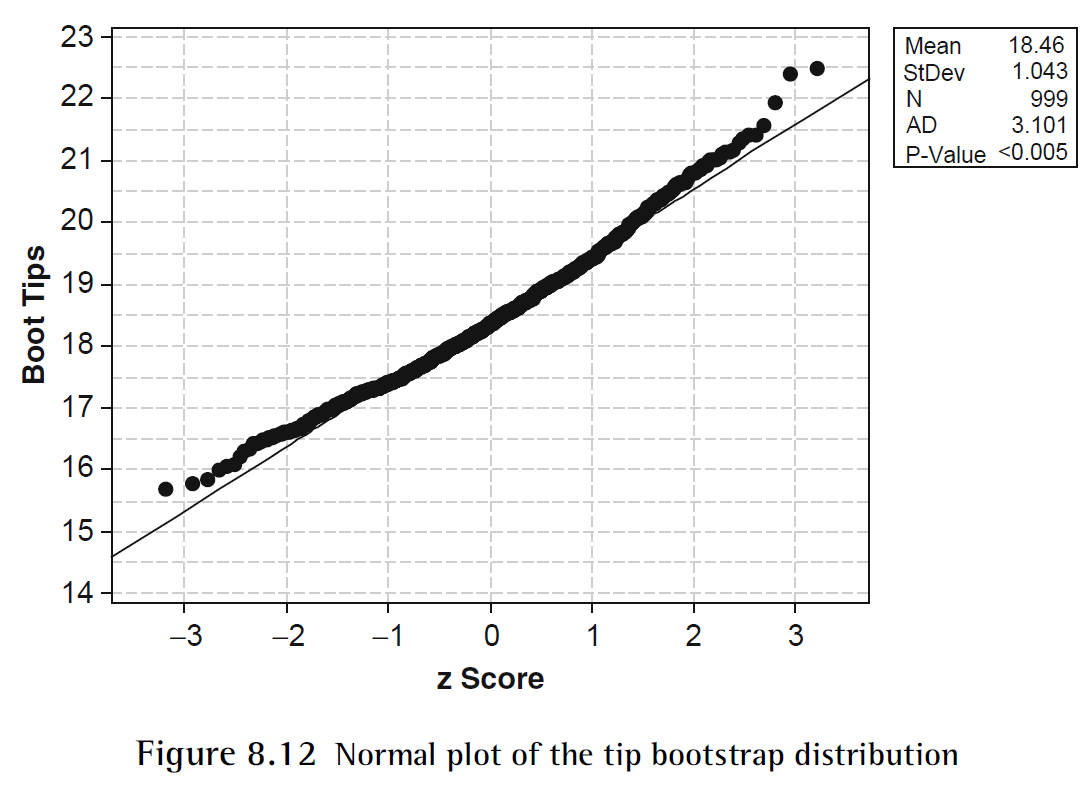
- This means the first of 999 samples have a mean of 18.41. The asterisk means this result comes from a bootstrap sample - not from the actual sample taken.

- a mean  
 - first of many means  
 - these means are bootstrap means

Histogram of the 999 bootstrap means



The histogram actually looks normal-like. Here's the normal plot



So the values are not really normally distributed.

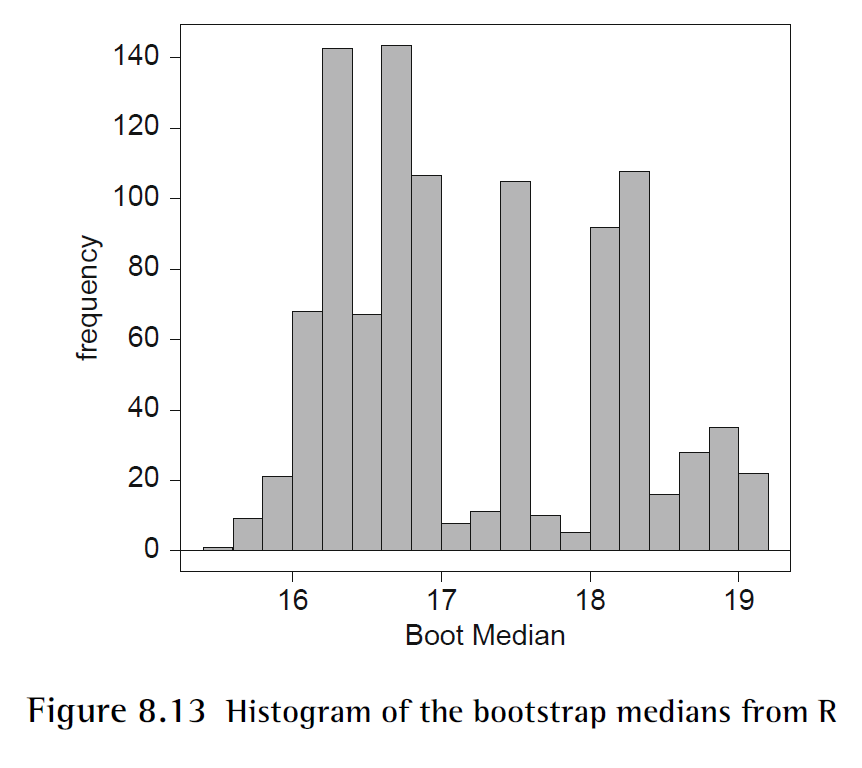
Had the data been more normal, a standard deviation can be calculated from the bootstrap samples:

and the 95% confidence interval would be:

Note the sample mean is still used as the center of the confidence interval, not the bootstrap sample means . This is of course assuming that the population is normal.

**Percentile Interval** - sort the 999 bootstrap sample means and use the 5th percentile value and the 95th percentile value to define the 95% interval. This procedure is the more appropriate one here because the population, as well as the distribution, are both non-normal.

**Bootstrapping the Median** - treat the 30 samples as population and create 999 samples. Compute the median from each of them. Histogram of these medians:



The data is definitely not normally distributed, so use the percentile approach to get a confidence interval.

The median is less affected by extreme observations, and the confidence interval will be smaller.

# Single Sample Test of Hypotheses

## Test Procedures

**Hypotheses**

**null hypothesis**  - the claim that is initially assumed to be true

**alternative hypothesis**

The two possible conclusions from hypothesis-testing analysis are to **reject**  or **fail to reject** . Note it's not to accept - it's just that the test doesn't reject . After the test, we will continue to believe in the plausibility of the null hypothesis.

null value - a number that appears in both and , separating the null from the alternative. Example: and .

**Test Procedures**

A test procedure is specified by:

1. A test statistic
2. A rejection region

**Type I Error** - rejecting the null hypothesis when it is true.

**Type II Error** - failing to reject the null hypothesis when it is false.

- Error probabilities for type I and type II errors.

**Example 9.1**

Random variable is binomial, with .

Rejection region:

α

β

8

20

5

0

The range of is 0 to 20.

The peak is at 20 \* 0.25 = 5.

The region is the rejection region. It has been marked as . The area of the rejection region is the probability of rejecting when it's correct.

The region is the fail to reject region. It has been marked as . The area of the fail to reject region is the probability of failing to reject when it's incorrect.

where is the CDF of the binomial distribution.

Note that there is no single value. Instead it is a function of . The is largest when is just above 0.25.

Any larger than 0.25 will shift the peak of the graph to the right. The rejection region will remain as . This means area of the graph will decrease.

There's a table in the book, and . When increases to 0.5, , which is much smaller.

If the null hypothesis is changed to ,

This time the is a function of .

is largest when . Anything smaller than 0.25 will pull the peak of the graph to the left, reducing the tail area that's called .

In general, focusing on the boundary value amounts to working explicitly with the "worst case" and will produce the largest and values.

**trade off**

The and mark complementary areas of a pdf or pmf. For any given , . When the rejection boundary is moved around, we are increasing one error at the expense of the other error.

The general practice is to specify and such that type I error is more serious than a type II error. In this view point, the rejection region is selected to specify the largest value of that can be tolerated.

The specification is the **significance level** of the test. Traditional levels of significance are 0.10, 0.05, and 0.01. The corresponding test procedure is called a **level test** (e.g., a level 0.05 test or a level 0.01 test).

## Population Mean Tests

**Case 1: A Normal Population with Known**

The null hypothesis is stated as .

The mean will be normally distributed. The will be the standard normal distribution.

The alternative hypothesis can be stated several different ways:

upper-tailed test

lower-tailed test

two-tailed test

**Example 9.7 - calculating**

The hypotheses are and .

Sample size , population standard deviation .

The test will be done with . What is the probability of a type II error when .

In the world:

α

30,000

In the world:

β

c

31,000

Think of and as two distinct situations. The common link between the two is the critical value between the reject region and the fail to reject region.

So with , , we get

What will produce ?

As gets larger, the value will get closer to the , widening the gap between and .

Also, as gets larger, the sample standard deviation will get smaller. The pdf becomes narrower, reducing the tail probabilities.

For the world:

For the world:

The two values has to match. Solve for

Final answer is .

**Case II: Large Sample Tests**

Due to the central limit theorem, the will be normally distributed.

The population is unknown, so use sample standard deviation to estimate . The standard normal curve is .

**Case III: A Normal Population with Unknown**

If the population is normal, then the mean will follow the t-distribution.

So the same thing as before, but replace the z-score with the t-score.

**Example 9.9 - Population mean inference using t-distribution**

The sample data is:

Does this data suggest the population mean exceeds 25? Test using .

|  |  |
| --- | --- |
| Get and from the sample. | import scipy.stats as stats  import numpy  x = numpy.array([25.8, 36.6, 26.3, 21.8, 27.2])  x\_bar = x.mean()  s = x.std(ddof=1) |
| The t-score for this sample: | (x\_bar - 25)/(s/(5\*\*0.5)) # 1.038 |
| The t-score for the limit: | stats.t.ppf(1-0.05, 4) # 2.1318 |

The conclusion is that the mean in the sample is not sufficiently far away from the to reject the hypothesis.

**Example 9.10 - Calculating for the t-distribution**

Test hypotheses: and

Population .

Sample size . Test at .

If the real , what is the type II error?

|  |  |
| --- | --- |
| In the world: | t = stats.t.ppf(1-0.05, 9) # 1.833  c = 2.5 + t \* 0.1 / 10\*\*0.5 # 2.558 |
| In the world: | t = (c-2.6)/(0.1/ 10\*\*0.5)  B = stats.t.cdf(t, 9) # 0.108 |

What sample size will give ?

The complication here is that the t-scores are not constant - as is varied, the t-scores will change as well.

The following loop prints out the different outcomes for the different values.

for n in range(10, 20):

# u = 2.5 case

t = stats.t.ppf(1-0.05, n-1)

c = 2.5 + t \* 0.1 / (n\*\*0.5)

# u = 2.6 case

t = (c-2.6)/(0.1/ 10\*\*0.5)

B = stats.t.cdf(t, n-1)

# results

print("n=" + str(n) + " ," + str(c) + ", B=" + str(B))

Output:

n=15 ,2.54547683216, B=0.0533380004325

n=16 ,2.54382625889, B=0.0479770386075

So final answer is .

Note the book obtained using a software package.

**Power**

When the null hypothesis is true, .

## Population Proportion Tests

**Example 9.11 - Large-Sample Test**

The test settings are

The data is , with sample size .

Let be the total number of successes. As the sample size is large, is modeled as a normal distribution.

In the null hypothesis world:

reject

fail to reject

The test data is 1276, which is greater than the 1272 cut off. So the test data is in the reject region. In terms of z-score, the test data is

Conclusion: H0 is rejected.

**Example 9.12 - Calculating and**

The test settings are

Initial sample size .

world:

fail to reject

reject

world:

reject

fail to reject

The probability of failing to reject scenario is .

Suppose the goal is to get to be 0.01 or lower by increasing .

In the world,

In the world,

Equal the two c's and get . Answer is .

**Example 9.13 - Small-Sample Tests**

When the sample size is small, model the possibilities using the binomial distribution rather than the normal distribution.

The test settings are

Sample size .

world:

fail to reject

reject

16

15

To find the rejection location, examine the CDF

import scipy.stats as stats  
  
# find rejection cutoff

stats.binom.ppf(0.05, 20, 0.9) # 16

stats.binom.cdf(16, 20, 0.9) # 0.13295

stats.binom.cdf(15, 20, 0.9) # 0.04317

15 is the in the rejection region because the CDF up to that point is less than 5%.

world:

fail to reject

reject

15

The probability of failing to reject scenario is

1 - stats.binom.cdf(15, 20, 0.8) # 0.6296

## P-Values

The **P-value** is the probability in the tail(s). The value is something that is fixed for the test, while the p-value is the actual probability of the tail, based on the location of the sample statistic on the pdf.

**Example**

. Experiment produced a mean of .

P-value = 25%. This means that if we run the experiment many times, we will get a mean that is larger than 25% of the time.

**P-value decision rule** - If , reject . Otherwise do not reject.

If we are free to choose the , the P-value is the lowest where we can reject the null hypothesis.

Another name for the P-value is **observed significance level (OSL)**.

For a two tailed test, the P-value is the sum of the probability in both tails.

**P-value Distribution**

The P-value is a random variable that has a distribution.

When the is true, the distribution of the P-value is uniform. The probability of getting the result in, say the upper 5% range, is actually 5% - as long as the model is correct.

When the is false, the P-value will not be uniform. It will be higher at whatever that is the most probable.

**Example**

Model for being true

P-value of 0%~5%

Real model

The P-value of 0%~5% normally happens 5% of the time, had the model been correct. In this case, the model is not correct, and the P-value of 0%~5% will happen 50% of the time.

## Likelihood Ratio Tests

**practical significance** versus **statistical significance** - a large sample size can create a sufficiently small P-value, which shows statistical significance that may favor rejecting in favor of , but you also have to look at the practical significance - is sufficiently different from .

**simple hypothesis** - if true, a simple hypothesis completely specifies the distribution of the sample 's. For example, if we are talking about an exponential distribution, is simple, but is not simple. If this later is true, there is a range of distributions, each with its own parameter.

**Neyman-Person Viewpoint**

This views the reject region, and the fail to reject region, as a ratio of two likelihood functions.

Let be the joint pmf or pdf of the 's.

The probability of the null hypothesis is

The probability of the alternate hypothesis is

The rejection region is

The rejection region is viewed as a place where the alternative hypothesis likelihood function is greater than the null hypothesis likelihood function scaled by some positive number .

Reject

Fail to Reject

The equations for type I error and type II error are

where is the complement of

**Example 9.20 - Poisson distribution statistics using likelihood functions**

The Poisson pmf:

The hypotheses are

The test parameters are:

The likelihood function:

Ratio of the likelihood functions:

The rejection region:

Take natural log of both sides

The k doesn't matter

The cutoff point:

For , has Poisson distribution with . This is something that has been worked out in an example in the "Sufficiency" section.

|  |  |
| --- | --- |
|  | import scipy.stats as stats  stats.poisson.cdf(8, 5) # 0.9319  stats.poisson.cdf(9, 5) # 0.96817 |

For , let {0~9} be the fail to reject region. The reject region starts at .

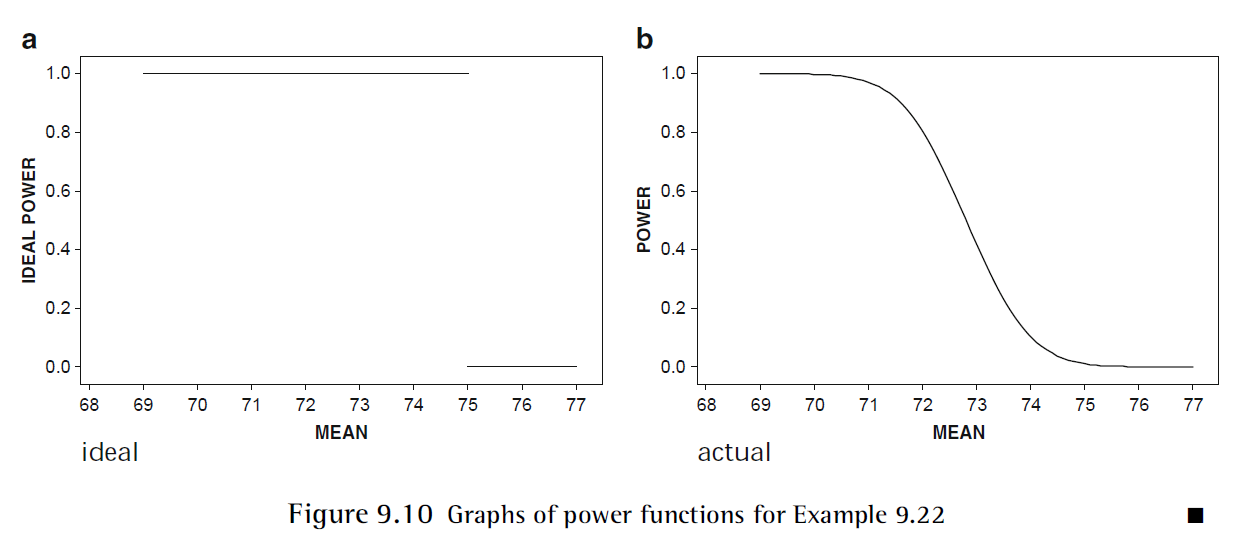
Type I and Type II errors:

**Power Function** - the probability of rejecting , considered as a function of

Formally:

**Example**

Suppose we want to test for versus



The left curve is the idea power function - the probability of rejecting the hypothesis when is 100%, and this probability instantly goes to zero if .

The curve on the right is the test done in example 9.22

Relationship of the power function to type I and type II errors:

where and are two disjoint sets of possible values of , such that and

Note the function is continuous. You just write an expression for under null hypothesis conditions.

**Example 9.22 - Power function for from a normal distribution**

Hypothesis:

Data comes from a normal distribution with parameters:

Test settings:

c

75

Cutoff point:

Power function:

|  |  |
| --- | --- |
|  | def power(theta):  z = (72.81 - theta) / (9.4 / 100\*\*0.5)  return stats.norm.cdf(z)  power(70) # 0.9986  power(75) # 0.01  power(76) # 0.00034488 |

Note that , the significance level

**Composite hypotheses** - this is when or (or both) consist of more than a single value

**uniformly most powerful (UMP) level test** - this term applies to composite hypotheses situations only. In this kind of test, the is maximized for any subjected to .

**Likelihood Ratio Tests** - given a ratio of likelihood functions

reject the null hypothesis when this ratio is too low. Again, the actual value of the ratio, say , doesn't matter. It's the form of the equation that matters.

A number of tests can be derived from the likelihood ratio principle.

**Example 9.24 - using the likelihood ratio principle to derive a test about the normal distribution**

Hypotheses

Population follows the normal distribution with unknown parameters and

The pdf for the normal distribution:

Likelihood function:

From section 7.2, the MLE for and are:

Likelihood ratio:

Manipulating the likelihood ratio expression into a test:

Simplification in the denominator:

# Two Sample Inferences

## Difference between Means

**Basic Assumptions**

is a random sample from a population with mean and variance .

is a random sample from a population with mean and variance .

The and samples are independent of each other.

**observational study** - the individuals or objects used in the comparison study are not assigned by investigators - they are "self-assigned".

**randomized controlled experiment** - the investigators do assign test subjects to two separate groups in a random fashion. The goal is that the difference in response is due solely to a difference in treatments.

**Large-Sample Tests** - Large sample assumption allows the use of the Central Limit Theorem, making and normal distributions regardless of the underlying population distributions. The parameters and can be approximated by sample standard deviations.

Rule of the thumb for characterizing sample sizes as large is and .

**Number of samples necessary for a certain confidence interval width**

Assume equal sample sizes, so both and has samples

## Two-Sample t-Test

**Difference of two means for normally distributed populations**

The random variable can only be assumed to be normally distributed if both sample sizes are large.

When the sample size is small, and and are both normally distributed, then will be t-distribution minus t-distribution, resulting in a third t-distribution.

The is a t distribution with degree of freedom

**Derivation of**

Start with

Standardize the numerator by dividing by the variance of

The numerator is , the standard normal distribution. The t-distribution is defined as . So for the whole thing to be a t-distribution, we need

where is a chi-squared random variable and is the degrees of freedom.

Properties from the chi-squared distribution section

Additional identities from the above properties

The key identities needed are

Returning to

Applying E( ) to both sides of the equation will only produce 1=1. Not very informative, but at least there's no contradiction.

Applying V( ) to both sides of the equation will give

Replace the population standard deviation values with sample standard deviation values to get an expression for computing the degree of freedom .

**Pooled t Procedure**

This procedure assumes equal variance in the two samples, meaning .

This procedure is not robust since it will perform poorly if it turns out that . The book recommends the standard two-sample t procedure, with separate standard deviations.

The benefit is that the degree of freedom is easily calculated as .

The procedure starts with standardizing , but this time with .

Replacing with

will produce a t-distribution, with degree of freedom .

Note the is a weighted average of the standard deviations.

**Derivation on why replacing with leads to t-distribution**

Start with

Property from the chi-squared distribution section

In this case

The ratio produce chi-squared distributions:

From the definition of the t-distribution:

So dividing by will produce a .

Dividing the standardized

by is effectively replacing by , and that will lead to a .

## Paired Data

Paired data is when one set of individuals or objects each have two observations.

So the data is and . The and is coming from the same individual or the same object, so these two data points are not independent. One of the assumptions in the previous sections is that and data points are independent.

**One sample t-test approach**

In this approach, define a new set of data, , that is the difference of and data. So . The and data points are from different objects or different individuals, so they can be assumed to be independent.

Then analyze the mean and standard deviation of the dataset as before.

As a review: If is large, will be normal due to the Central Limit Theorem. If is small, should be graphed to see if the data follows a normal distribution. If it does, then the t-distribution can be used.

**Two sample t-test approach**

In this approach, we try to compensate for the dependence between and using correlation .

As a review, if and are independent, we can write

When and are dependent,

Resulting in

Apply this more general property to results in

**Derivation**

## Difference between Proportions

**Distribution of the Difference between Proportions:**

Let and with and independent variables.

The proportions being estimated are

For large and , the Central Limit theorem will say that is normally distributed, with mean and variance:

**Null Hypothesis:**

In this case there is only one proportion, taken from the whole data set.

As seen above, this approach to calculating a single is equivalent to getting a weighted average of and .

The variance will then be

**Alternate Hypothesis:**

For calculating the type II error, the proportions are kept separate. The is a function of two variables.

**Example 10.12**

The hypotheses are .

The population's true proportions are .

The sample size is left as unknown , but the same for both X and Y group. We are trying to find the sample size that will produce when .

c

0

fail to reject

1.5e-4

def B(n):

p\_ave = (0.0003+0.00015) / 2

V\_null = 2 \* p\_ave \* (1- p\_ave) / (n)

c = 1.645 \* (V\_null) \*\* 0.5

V\_a = (3e-4\*(1-3e-4) + 1.5e-4\*(1-1.5e-4)) / n

z = (c - 1.5e-4) / V\_a \*\* 0.5

B = stats.norm.cdf(z)

return B

solve(B, lambda x: 0.1, 1, 1e6) # 1.713e+05

**Small-Sample Inference**

One option is the Fisher-Irwin test, which is based on the hypergeometric distribution.

## Ratio of Variances

**F-distribution**

The test for ratio of variances are based on the F-distribution.

Let be a random sample from a normal distribution with variance .

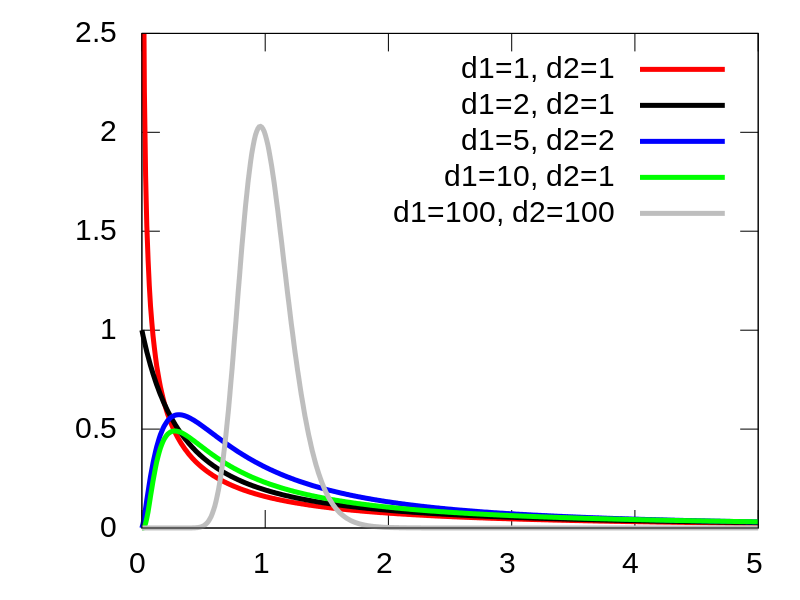
Let be another random sample, independent of the 's, from a normal distribution with variance .

Let and denote the two sample variances. Then the random variable

has an F-distribution with and .

Note the F-distribution parameters are NOT simply .

**Graph of the F-distribution**



When the degrees of freedom are approximately equal, the F-distribution is a bell shape curve that is centered at 1. The 50% probability is for to equal , which is consistent with

**Null Hypothesis Test**

The null hypothesis is .

From the definition of the F statistic

The assumption of means produce an F score that describes how likely is .

**Confidence Interval of**

Start with

and manipulate the equation into

The confidence interval is

Note that the F statistic is in the denominator. That is why produces the low end of the interval.

**Example 10.14 and 10.15**

The two sets of data are

data\_B = [99.8, 105.4, 94.7, 107.8, 114.3, 106.3]

data\_W = [99.0, 98.2, 98.1, 102.1, 102.9, 104.1, 98.8, 99.5]

The null hypothesis:

Computing the likelihood of using the F-distribution:

m = len(data\_B)

n = len(data\_W)

V\_1 = numpy.var(data\_B, ddof=1) # 45.7657

V\_2 = numpy.var(data\_W, ddof=1) # 5.4655

f\_score = V\_1/V\_2

p = 1 - stats.f.cdf(f\_score, m-1, n-1) # 0.00723

The likelihood of getting an even larger ratio is 0.007. So the null hypothesis is rejected.

Calculating a 98% confidence interval for :

f\_score\_low = stats.f.ppf(0.01, m-1, n-1)

f\_score\_high = stats.f.ppf(0.99, m-1, n-1)

ci\_low = V\_1\*\*0.5 / (f\_score\_high\*\*0.5 \* V\_2\*\*0.5) # 1.0594

ci\_high = V\_1\*\*0.5 / (f\_score\_low\*\*0.5 \* V\_2\*\*0.5) # 9.3568

**Dependency on Normality**

The use of the F-distribution depends strongly on the underlying data following the normal distribution.

In contrast, the t procedures of this chapter will still work in the case of moderate departures from normality.

## Bootstrap and Permutation Methods

**Bootstrap for Two Samples**

Suppose data\_x = and data\_y = .

Create a random sample using data\_x as the population, and find the mean of this random sample. Call this mean , and do it 999 times. The set is called the **bootstrap sample**. Its distribution is called the **bootstrap distribution**.

If the bootstrap distribution is approximately normal, you can use its mean and standard deviation to get a confidence interval.

If the bootstrap distribution is not normal, use the percentile interval. If there are 999 data points in the bootstrap sample, then sort the data and use the 25th data point from the top, and the 25th data point from the bottom, as a confidence limit for a 95% confidence interval.

**Permutation Tests**

In summary, the permutation test tries out all possible arrangements and sees how the null hypothesis rank among these arrangements.

Suppose the null hypothesis is . When using the permutation test, the assumption is not only , but that the two data sets are in fact the same population. Then the data can be reordered in any way possible to generate the random samples.

For example, suppose the data is

B: 4.9, 5.5, 6.5  
G: 0, 1.3

If B and G are both from indeed the same population, then a valid random sample would be:

B\_random\_sample: 4.9, 5.5, 0  
G\_random\_sample: 6.5, 1.3

Note that the 6.5 that use to be in "B" went over to "G". Allowing data to change labels in thise way is what sets the permutation test apart - in contrast with previous section, where the random samples for B will only use the data from the original B dataset. This step assumes that not only , but that B and G are actually the same population.

The original data's mean is

The total number of arrangements possible is . Out of these 10 arrangements, the original data mean is the most extreme - because it uses the largest values for the B mean and the smallest values for the G mean. The probability of getting the original B mean, or higher, is 1/10.

Due to the factorial nature of the combination , the number of arrangements goes up very quickly. So the permutation test is generally approximated, with a few thousand randomly permutated samples, rather than an exhaustive search of all possibilities.

# ANOVA

**ANOVA = analysis of variance**

## Single-Factor ANOVA

**Hypothesis Setup**

Single-factor ANOVA compares means from multiple populations.

the number of treatments being compared

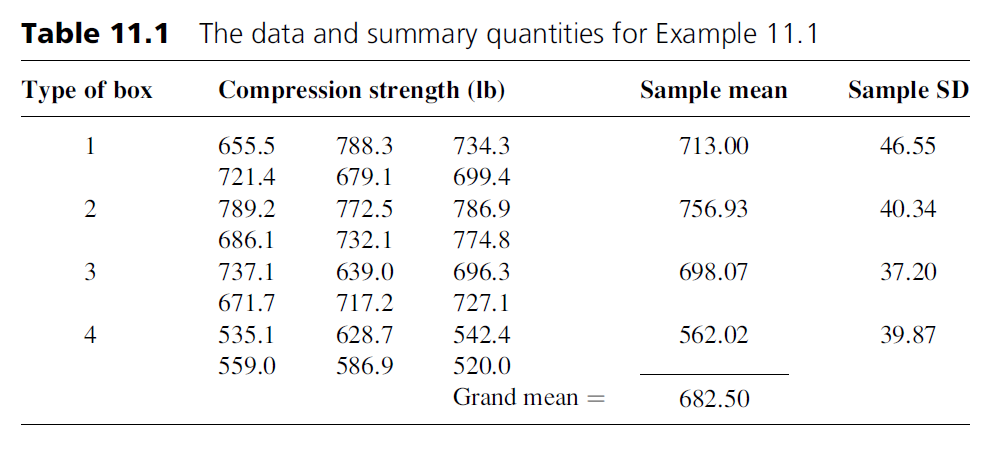
the mean of population 1  
 the mean of population I

Alternative names: is the response to **treatment 1**

Single-factor ANOVA as in only one parameter is being studied.

at least two of the 's are different

**Data Example**



There are four groups of data, each with its own mean. A box plot can visually show that box #4 is in a separate region.

**Notations**

the random variable (rv) denoting the i-th population's j-th measurement

the actual data value from the i-th population's j-th measurement

the number of observations in each population sample. For this section, all samples have the same number of measurements.

The population sample means are

The **grand mean** is

The sample variances are

**Assumptions**

independence - The variables within any particular population sample are independent. Different samples are independent of each other.

The population or **treatment distributions are all normal with the same variance** .

To check the normality assumption, there is usually not enough data within each treatment. Instead, subtract from the first sample, subtract from the second sample, and so on. Plot all this data on a single normal probability plot.

This data is a set of deviations, called **residuals**. The resulting plot is called the **normal plot of residuals**.

Rule of thumb - if the largest standard deviation is not much more than twice the smallest , then it's reasonable to assume equal variances.

**Sums of Squares**

The **treatment sum of squares SSTr** measures the spread of the treatments.

The **error sum of squares SSE** measures the spread of the data values within each treatment.

Note it's not minus the grand mean, but the mean of a particular treatment.

**Sums of Squares as Chi-squared Distributions**

**Review**

If , then and are independent.

Next, treat the as a data set, with its own variance.

The baseline assumption includes normal distribution and common variance .

If the null hypothesis is true, then the treatment means are also the same, so that effectively the all comes from the same population. In that case, we can say .

If the null hypothesis is false, then the numerator will have even greater variability - the values will be further spread out. The SSTr will be larger than the null case.

**SSE and SSTr are independent random variables**

SSTr can be expanded to show that it's a function of

So the SSE also has the term, but it can be manipulated into a function of treatment variances only. A previous result states that means are independent of variances.

**Mean Squares**

The **mean square for treatment**

It's sort of an "average" of the treatment sum of the square. A true average would divide by instead of .

**If is true, then**

If is true,

The expected value of a chi-squared variable is its degree of freedom.

If the null hypothesis is false, the numerator will grow and the will overestimate the variance.

The **mean square for error**

Again, a true average would have divided by .

Previously we have

**The F Test**

When is true,

The F random variable is F-distribution with parameters .

Its value should be close to 1.

**Explanation**

Previously we have

So

and F fits the definition of the F-distribution.

In terms of expected values:

if is true, and even larger if is false.

, whether is true or not.

So F should be close to 1 if is true.

The p-value of the F test is the upper tail of the F-distribution.

**Computational Formulas**

|  |  |  |
| --- | --- | --- |
|  | **Definition** | **Computing Formula** |
| **Sum of Squares Total** |  |  |
| **Sum of Squares Treatment** |  |  |
| **Sum of Squares Error** |  |  |

represent the sum of the for fixed

represent the sum of all IJ observations - the **grand total**

The is computed via

The F score is computed by formulas already discussed. These are reproduced below

**Fundamental ANOVA Identity**

The is the total variation in the data. This total variation is broken into two parts.

The is the variations that would be present whether is true or false. This is the variation within each group, without considering these group means are similar or not.

The is the variation due to differences between the group means and the grand mean. This is the variation that would be largest when is false.

When is large relative to the , then the is rejected.

**Derivations**

The connection between and is

For the first term, the connection between and is

For the second term, the connection between and is

Square both sides, apply summation

To show that the highlighted term is zero:

The highlighted area can be factored out because they don't contain any term.

**Levene Test**

The ANOVA assumes the populations have equal variances.

The Levene test is one way to test for equal variances. It is an ANOVA test on the absolute value of the residuals. So an ANOVA test is done on the quantities instead of . If the populations have equal variance, then the means should be equal.

## Multiple Comparisons in ANOVA

**Multiple Comparison Tests**

When the single-factor ANOVA is rejected, we would like to know which 's are different from each other.

The general idea is to calculate a confidence interval for each pairwise difference . If this interval does not include 0, conclude that and **differ significantly**.

**Studentized Range Distribution**

where are independent standard normal rv's

is a chi-squared rv, independent of the 's, with df

the critical value with upper-tail area

"Range" because the numerator of Q is the range of the 's.

"Studentized" because of the division by is similar to the definition of the t distribution.

**Mean Difference Confidence Intervals**

The confidence intervals for are

There are such intervals.

The confidence level that all of these intervals include the corresponding is .

**Derivation of following the Studentized Range Distribution**

There are means in the hypothesis.

For each mean follows the normal distribution

Due to the previous result

Due to

The highlighted terms all cancel out.

**Derivation of going from Q to a confidence interval**

You can get rid of the "max" if you say it's P(... for all i,j).

The inequality can be expanded into a confidence interval.

Note the Q subscript is , not .

**simultaneous confidence level**

The confidence level in this model is called a **simultaneous confidence level** because the same term is being applied to all intervals.

For instance, and , are two separate intervals, and the CI uses for both of them.

If the two intervals are independent, then 0.95 confidence CI on each interval will be confidence overall.

Yet these intervals are not all independent, for example and both share the same . So in this case, the overall probability is not a simple matter of multiplying together the individual probabilities.

**Tukey's Procedure**

Let

The confidence intervals are

If an interval does not include 0, then the corresponding and are significantly different.

Pictorially, the borderline case is

0

w

So if the difference between two means are smaller than , then the confidence interval will include zero, and these two means are said to be not significantly different.

The Tukey procedure is to sort the means and underline pairs that differ by less than .

**Example 11.5 (partial)**

**Confidence interval for**

The estimator is

Due to the ANOVA assumptions, the is normally distributed, with

Using the to estimate the variance will change the distribution from a normal distribution to a t-distribution. The confidence interval of is

## More on Single-Factor ANOVA

**Alternative Description of the ANOVA Model**

**Modeling as**

**Modeling as**

The claim that the 's are identical is equivalent to .

**Full equation for**

Previously we only had: .

The is a measure of the extent to which is false.

**Derivation**

Start with the computation formula for

Use

Use to get the final expression for .

**for the F Test**

The formula

allows the computation of probability when the null hypothesis is not true. The equation only depends on , so that acts as a sufficient statistic. The following situations all have

and so they will all have the same .

**Data Transformation**

The ANOVA methods assume the same variance for the various populations.

Data can sometimes be transformed into a population with the same variance, if there is a relationship between the variance and the mean.

**Example**

Suppose the data comes from Poisson, so , and therefore the variances are not equal if the null hypothesis is false.

Suppose there is a transformation function such that .

Through Taylor series, we can say

We want

When the is a function of , we can integrate to get a transformation function.

For the Poisson case

The transformation function to use is

**Generalization**

If , a known function of , then the transformation

will "stabilizes the variance" so that is approximately constant.

**A Random Effects Model**

**Single-Factor Fixed Effects Model**

's are random  
 and are fixed parameters whose values are unknown

The null hypothesis . If there is a difference , the difference is assumed to be a fixed number.

**Random Effects Model**

The becomes . The difference between means is now assumed to be a random variable. For example, the pH values of the soil that is being tested. There are many pH levels, and the experiment mere test a few of these levels from a large population of levels.

The null hypothesis is , which says that the different levels are actually the same.

So the model is different, but the two null hypotheses converge to the same thing, that the means are the same. The two models only diverge in the alternative hypothesis - that is they diverge on how the differ.

Since the null hypothesis works out to be the same thing, the same test procedure is used for the random effects model null hypothesis. Mathematically the equations are

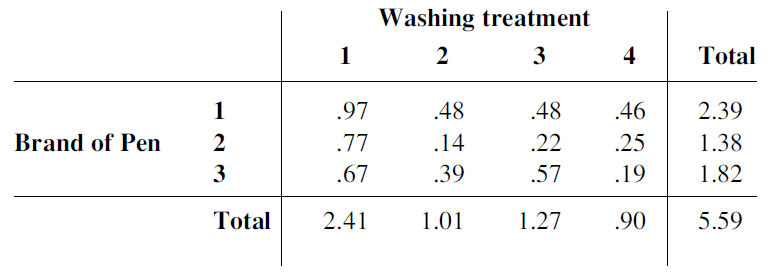
, which is the same as before

which is different from before

For the null hypothesis case, we have , which is the same as before.

## Two-Factor ANOVA with Kij = 1

**Example Data (from Example 11.11)**



The data is about applying pen marks to a fabric and then washing the fabric to remove the pen mark. There are two factors: the different brands of pen, and different wash treatments.

**Definitions**

Factor consist of levels and factor consists of levels. There are combination of levels and each of them is called a treatment.

Let

number of observations on the treatment consisting of factor at level and factor at level . For this section .

the random variable (rv) denoting the measurement when factor is held at level and factor is held at level

the observed value of

The convention of the book is to call the i-th row the level . Values in the i-th row are observed values of factor held constant at level .

**Means and Totals**

Data average when factor A is held at level

Data average when factor B is held at level

Grand mean

The totals are defined in a similar fashion, just without the bars. For example,

**An Incorrect Model**

The problem here is that there are too many parameters - the is parameters.

**The Model**

where and

The 's are independent, normally distributed, with mean 0, and common variance .

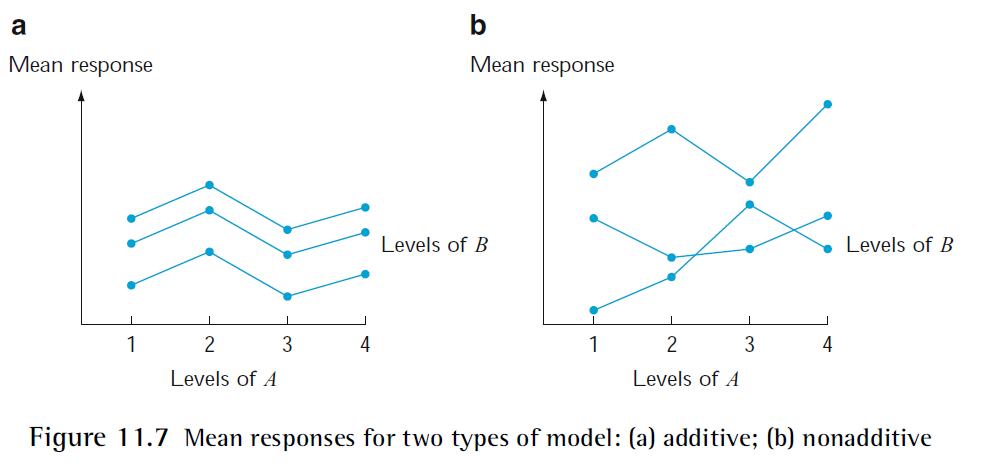
**Additive Model**

This is called an **additive model** where the mean for each is

If we vary factor A but hold factor B constant

which shows that factor B does not matter when it's held constant.

Graphically:

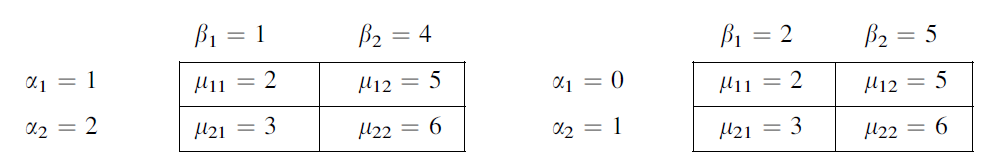


The data on average will go up and down in parallel. When factor A change from level 1 to level 2, the graph will go up by the same , regardless of the factor B level.

The actual data will not be exactly parallel due to the presence of variance . It's just that the means will go up and down in parallel, causing the data to be somewhat parallel.

**The and Restriction**s

Without these restrictions, it's possible to have different sets of and that produce the same .



In the situation above, going from left to right, the lost 1 and the gain 1.

**Parameters and their Estimators**

the true grand mean  
 the effect of factor A at level   
 the effect of factor B at level

**Hypotheses**

versus at least one

versus at least one

**Test Procedures**

The equations follow the same idea as the Single-Factor ANOVA. Instead of there is now and .

Note the now utilize the column and row means, in addition to the grand mean.

The fundamental identity is

The total variation is split into a part that is not explained by either factor A nor factor B, and two more parts that are related to factors A and B.

Test statistics are

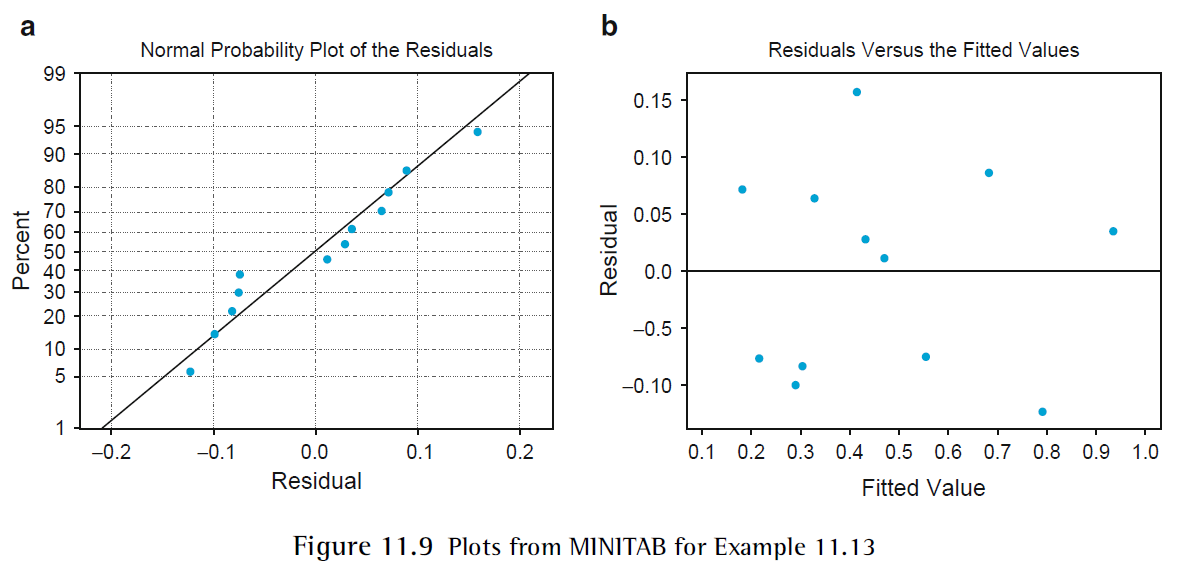
**Graphical Verification of Normality and Constant Variance Assumptions**

Define the **predicted values** as

These are also known as **fitted values**.

Define **residuals** as

The use of residuals aggregates all data onto a single graph. All data is reset relative to their group means, and supposedly all these groups should have the same variances.



For plot (a), the residuals should be normal.

For plot (b), the vertical spread of the residuals should not change as we move horizontally across the graph. If there is non-constant variance, then the residuals would be spread farther apart for some fitted values, and be clustered closer together for other fitted values.

If there is more variance for larger responses, then it might be fixed by taking logarithm of the data.

**Expected Mean Squares**

So like in the single-factor case, the expected values are all when the null hypothesis is true. The F-distribution is justified by saying that it models the ratio of .

## Two-Factor ANOVA with Kij > 1

**Constant**

For simplicity, this section assumes that . So the number of observations per group is constant.

**Additivity Assumption**

For , the additivity assumption was required. For though, this assumption is not required.

When additivity holds, a graph of the data means are parallel lines.

When additivity does not hold, there is **interaction** between different factor levels. The model for has an extra variable to model interaction.

**Fixed Effects Model, with Interaction**

the effect of factor A at level the effect of factor B at level

Restrictions:

The 's are normally distributed random variables, with mean 0 and variance . The other parameters are fixed values.

The 's are the **interaction parameters**.

The 's are the **main effects for factor A**.

The 's are the **main effects for factor B**.

**Hypotheses:**

for all

The needs to be tested first. If the is rejected, then there is no need to test for the other two.

**Analysis**

The total variation is divided into four parts

variance within group

variance between group means and the grand mean

|  |  |
| --- | --- |
| **Hypotheses** | **Test Statistics** |
| versus |  |
| versus |  |
| versus |  |

**Mixed Effects Model**

Suppose factor A is using the fixed factor model and factor B is using the random factor model:

The and 's are constants.

The 's, 's, and 's are random variables, with expected value 0 but different variances , , and , respectively.

# Regression and Correlation

## Linear and Logistic Models

**independent**, **explanatory**, or **predictor** variable  
 **dependent** or **response** variable

**Additive Model**

The f(x) is a deterministic function. The is a random error (random variable).

**The Simple Linear Regression Model**

The random deviation (random variable) is assumed to be normally distributed with .

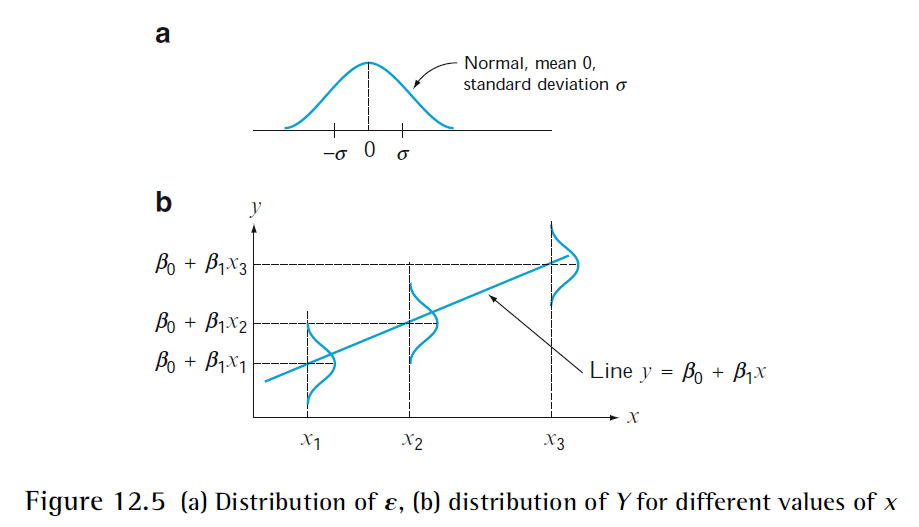
The data points are regarded as having been generated independently.

and are the coefficients of the **population** or **true regression line**

has a normal distribution, because of .

the expected value of for a given

the variance of for a given



For a rising line such as one shown x2 on average will produce a higher output than x1, but not always.

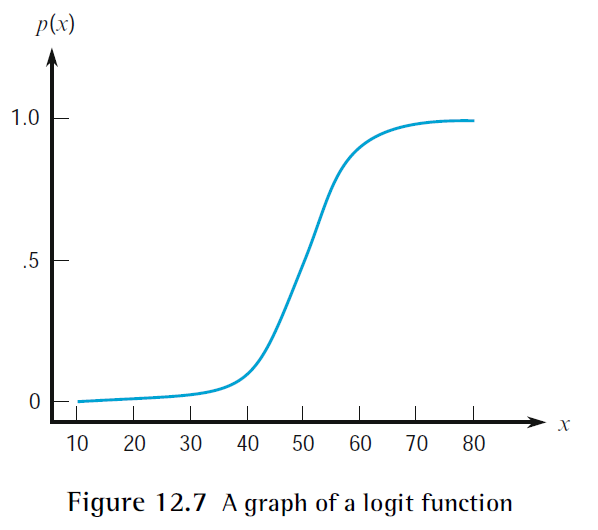
**Logistic Regression Model**

The **logit function**

This function will range from 0 to 1

so it's a good model for probabilities that depend on .

A graph of the function for positive and values:



If is negative, the probability would be decreasing.

Alternative form:

The expression on the left side is called the **odds ratio** because it's a ratio of success to failure.

If x increases by 1 unit, then the odds ratio increases by a factor of .

## Estimating Model Parameters

**Least Squares Model Parameters**

The vertical deviations, squared and summed:

The best and fits will satisfy

The partial derivative equations produce the **normal equations**:

where is the total number of data points

The solutions are

The and can be simplified

**Derivation**

The can leave the summation sign because it's not indexed. The second term is zero.

**Danger of extrapolation** - using the best fit line to fit x value beyond the range of the data

**Estimating**

The **fitted** (or **predicted**) **values** are . They come from utilizing the best fit line .

The **residuals** are the vertical deviations .

The **error sum of squares**

The (n-2) is the number of degree of freedom. There is a total of data points. For any given and , you can choose data points. The remaining 2 data points can be chosen by solving the normal equations using the data points, , and .

is an unbiased estimator for , but it is biased for . The mle of divides by rather than .

**The Coefficient of Determination**

**Total sum of squares**

It's like of like variance except that you don't divide by (n-1).

This is like SSE, if the estimator is simply using the mean .

SSE = the variations not explained by the linear model

SST = the total variations

Proportion of variations not explained by the linear model

**Coefficient of determination**

This is the proportion of variations that can be explained by the linear model

**Regression sum of squares (SSR)**

SSR is the amount of variation that is explained by the linear model.

see derivation section for proof

**Regression to the Mean**

Francis Galton found that taller fathers have sons that are above average in height, but not as tall as the father.

Due to the regression effect, things might naturally improve or get worse simply due to probability, and not due to whatever experimental changes.

For example, if special instructions are given to poorly performing students, their test scores might improve even if those instructions are useless.

**Derivations (Exercise #24)**

Denote the residuals as , where

**a1)**

One of the normal equations is

**a2)**

This matches with one of the normal equations

**b)**

**c)**

Now the goal is to prove the middle term is zero

Use results from parts (a) and (b)

The two summations are zero, according to part (a).

**d) prove the computational formula for SSE**

The hint for this problem says to use parts (b) and (c). However, I find the advice useless. The following reference the normal equations.

What we need is

|  |  |
| --- | --- |
| The first normal equation | The second normal equation |

Adding the two conclusions together gives the identity that we need.

## Inferences about the Regression Coefficient β1

**The Estimator for β1**

The simple linear regression model is

This β1 is a constant. The ε and Y are the random variables.

From actual data, the formula for estimating β1 is

The can leave the summation since it has no index attached to it.

The is . This happens multiple times in these derivations.

**is normally distributed**

The Y is a normally distributed random variable, due to the ε being normally distributed.

The is a linear combination of normally distributed random variables, so it too is normally distributed.

**Derivation:**

The expected value of the two highlighted terms are zero.

It is wrong to say

because is not a constant, so you cannot do a term by term division using this expression.

Both numerator and denominator becomes

The highlighted terms are constants, so they have a variance of zero.

. The ε is the source of variance in this model.

**The as t-Distribution**

The standardized variable

has a t-distribution with (n-2) df.

The S comes from

**Reasoning:**

Replace the population variance with an estimated variance will change the distribution to a t-distribution.

The degree of freedom comes from the having a degree of freedom of (n-2).

**The CI of**

**Fitting the Logistic Regression Model**

In the logistic regression model, the output is binary. is either 1 or 0. The probability is

The β­0 and β1 are obtained via maximum likelihood.

Suppose and the results are

|  |  |
| --- | --- |
| Successes: | x2, x4, x5 |
| Failures: | x1, x3 |

The likelihood function is

The β0 and β1 that maximize this function needs to be found numerically. There is no formula.

## Inferences Concerning Y

**Distribution of**

and are random variables. The is a fixed value.

The best fit equations for and are

**is normally distributed**

The previous section showed that is normally distributed. The is a function of , and that too is normally distributed.

**Hypothesis Testing using t-Distribution**

Replacing the σ2 with sample variance estimated from the data results in a t-distribution.

The variable

has a t-distribution with (n-2) df.

**Combined Confidence Intervals**

Suppose we are estimating two values using

and each value has a 95% confidence interval.

If the values and were independent, then the joint confidence would be .

But and appear in the estimate for both and , so the two values are not independent.

A result called the Bonferroni inequality says the rejection probability is at most , and so the confidence interval is at least 90%.

The key is to view the rejection probabilities as two events such that

**Prediction Interval for**

The confidence interval refers to the past data. It's about error that is based on

The prediction interval refers to the future. The error equation is

The difference is the . The confidence interval is working with existing data, which and there is no term. If the number of samples is very large, then the and will very closely approximate their actual population value. So the confidence interval width will approach zero.

In the prediction interval, even with perfect and knowledge, the future data still has . As a result, the prediction interval will not go to zero even if .

The past variance is

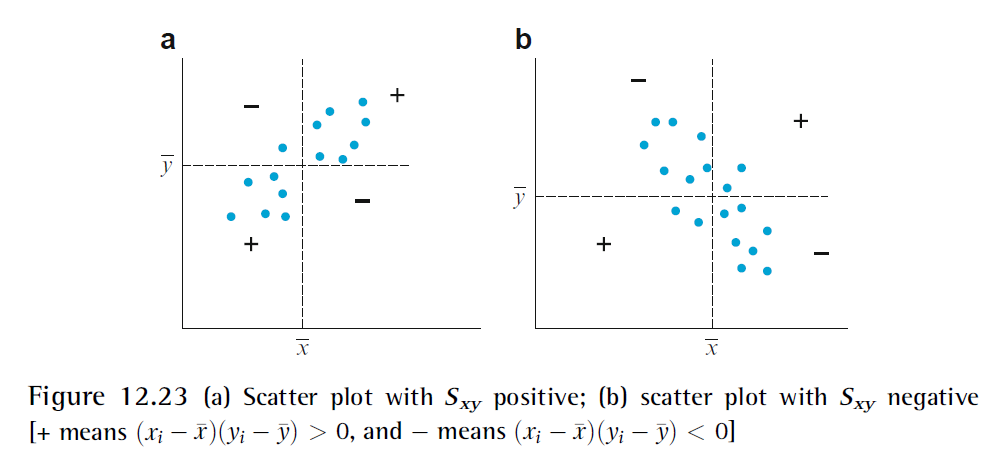
The future variance has the extra :

## Correlation

**Sxy**

If the relationship is strongly positive, then above will tend to be paired with above , so that , and that leads to being a large positive number.

If the relationship is strongly negative, then should be a negative number.



The defect of Sxy is that changing the unit of measurement for x or y can make Sxy arbitrarily large or small.

**The Sample Correlation Coefficient r**

The r is a standardized version of Sxy - it is independent of the units of measurement used.

If is replaced by , where is a positive number, the r will not change.

If is replaced by , the r will also not change.

**Properties of r**

There's another "r" introduced earlier, the coefficient of determination

These r's are the same "r".

**Derivation**

We are trying to show the green r

is the same "r" as the blue r

Start with SSR and proceed to express that in terms of the blue r

There's an identity from a previous section "Estimating Model Parameters", Derivations (Exercise #24), part b:

Now to make SSR in terms of blue r

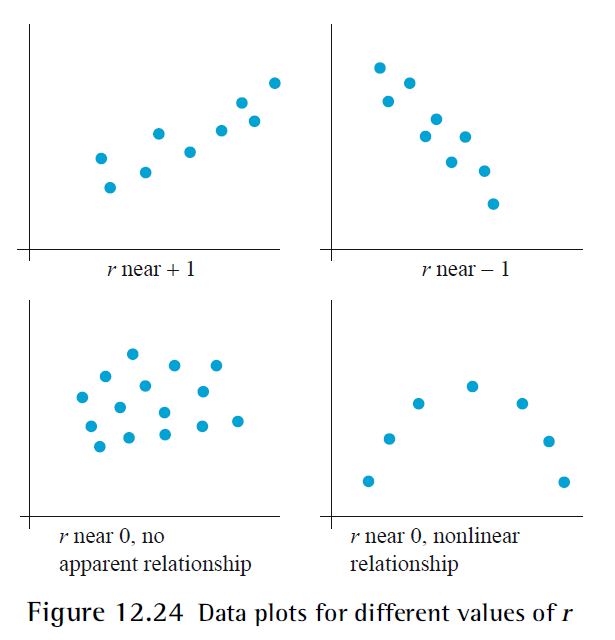
Now show the green r is the same as the blue r

Eliminate the SSE to get a relationship between and

From

we see that r2 is between 0 and 1. For r2 to be one, the SSE needs to be zero, meaning all the points are exactly on the best-fit line.

r near zero is only lack of a linear relationship. There could be a non-linear relationship lurking.



Rule of thumb:

Correlation is weak if

Correlation is strong if

**Relationship to the Population Correlation Coefficient**

In a previous chapter, the population correlation coefficient is defined as

The sample correlation coefficient r is a point estimate for the population correlation coefficient :

## Model Adequacy

To access model adequacy, you could try plotting the data on a graph. But then the scale used on the vertical axis could affect how good the fit looks.

**Properties of Residuals**

The residual

**Derivation for - Exercise #76**

You cannot say because and are not independent variables.

turn into using

In this context, all the x's are given values, and all the y's are random variables.

Put in terms of x's and y's.

The term goes to zero

Also use the index "j" instead of "i" to avoid confusing with the "i" in the problem. The "i" is the specific residue that we are talking about in terms of . The "j" will iterate across all values, from 1 through n.

For the random variables , the variable is special in that it has that extra coefficient "1" at the very start of the equation. The other (n-1) Y's only appear in the summation. Rewrite the equation to separate from the other Y's.

The different Y's are independent variables, so you can now say .

The is a common term between the two. Expand the first square by keeping this common term.

The summation now includes all of "j". There is no longer .

For now focus on just the summation portion of the equation.

Distribute the summation sign

Simplifications: and

**Standardized Residuals**

For large n

**Diagnostic Plots**

versus

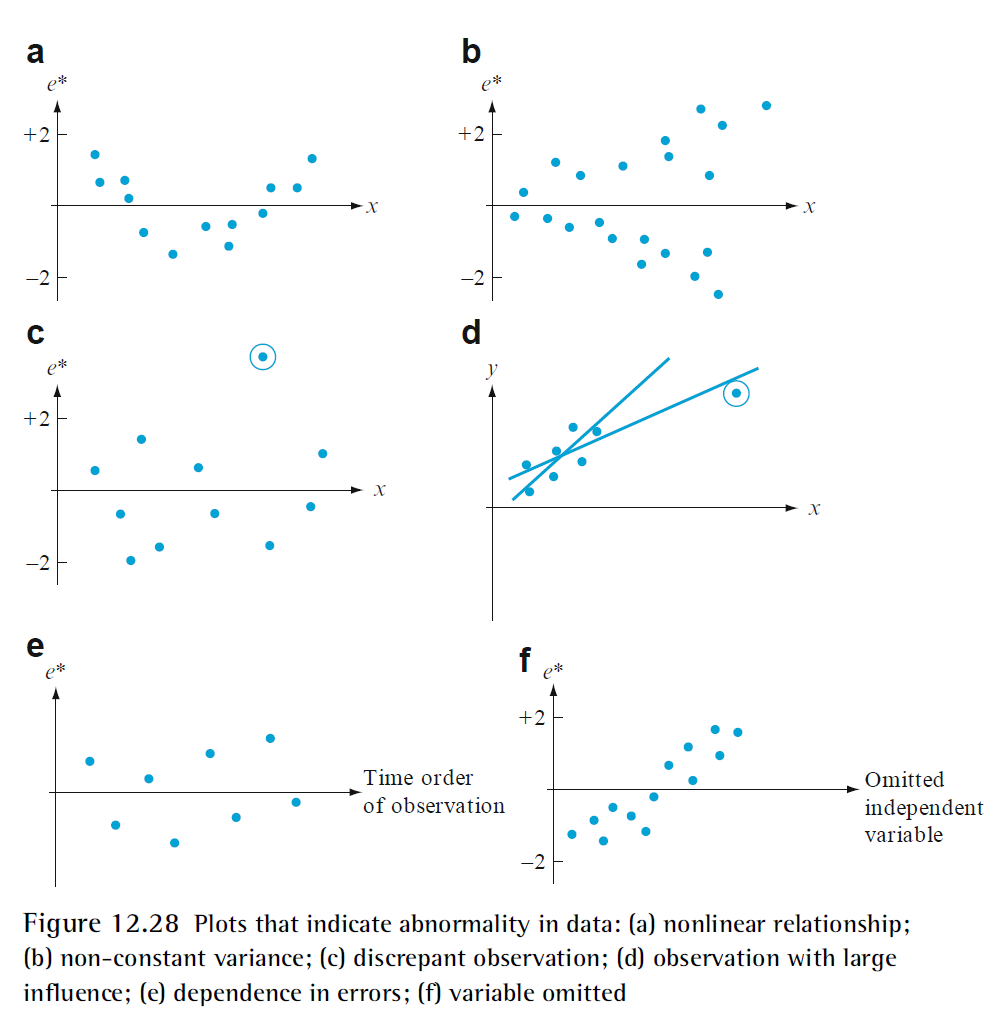
versus

versus

versus

A normal probability plot of

**Potential Difficulties**



## Multiple Regression Analysis

**General Additive Multiple Regression Model**

is normally distributed. The 's and 's are independent.

The 's are the **true** (or **population**) **regression coefficients**.

Let the value in the i-th observation and j-th predictor

The least square criterion function

produces (k+1) partial derivatives with respect to , which then all get set to zero to produce the normal equations.

**(Estimate of )**

Let be the **predicted** or **fitted values**.

The **residuals** are .

The is the degree of freedom. There are (k+1) normal equations that produce the model parameters. So for "n" data points you can only "freely" choose data points. The rest can be determined by solving the normal equations.

**Coefficient of Multiple Determination**

The is the proportion of total error that is not explained by the model.

The R2 is the proportion of the total error that is explained by the model.

Larger R2 is better but this metric can be inflated by adding predictors (more x's). What we want instead is just a few important predictors whose R2 value is high.

The **adjusted coefficient of multiple determination** is

Due to the fraction, the is always smaller than R2.

The fraction makes the much smaller than R2 when k is large. This is the adjustment to account for spamming predictors variables.

**Quadratic Regression Model**

**Interaction**

Model without interaction:

Model with interaction:

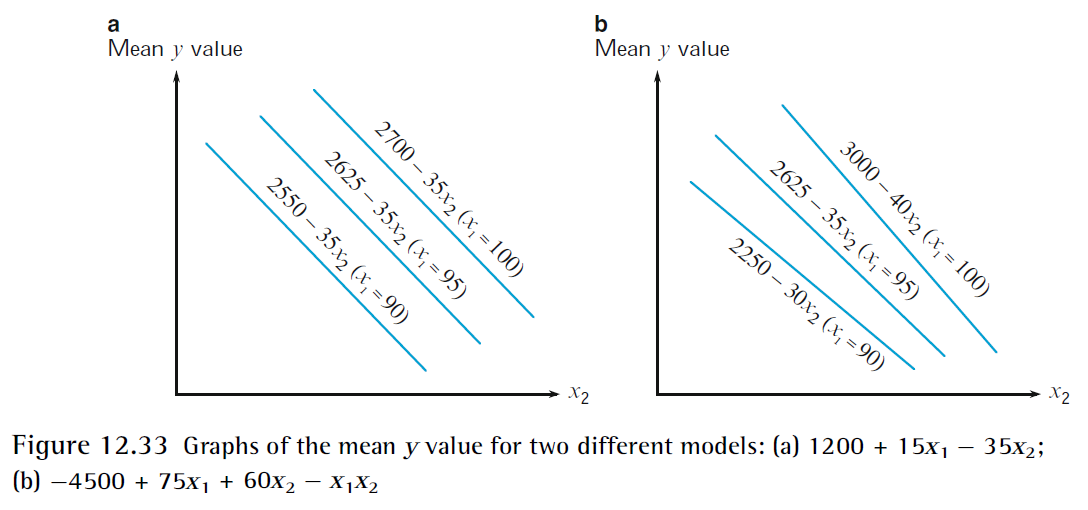
In the model without interaction, if x1 is held constant, then Y on average will be

The slope of the (x2, Y) plot does not depend on the value of x1.

In the model with interaction, if x1 is held constant, then Y on average will be

The slope of The slope of the (x2, Y) plot does depend on the value of x1.

When graphed, the model without interaction will produce parallel lines for different values of x1.



The **full quadratic** (or **complete second-order**) **model**:

**Categorical Data Modeling**

To model possible categories, you need variables.

For example, to model three teaching methods:

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Teaching method #1 | 1 | 0 |
| Teaching method #2 | 0 | 1 |
| Teaching method #3 | 0 | 0 |

You don't want to be using a single variable, with levels 0, 1, and 2. That kind of assignment might be implying that the teaching method associated with the 2 is twice as good as the teaching method associated with the 1. It's best to have only 0 and 1 in the encoding.

**Multiple Exponential Model**

Take ln on both sides to simplify to the additive model.

## Regression with Matrices

**Multiple Regression in Matrix Form**

There are "k" predictors and "n" data points.

There is a column of 1's, followed by a matrix of x-values.

**Dot Product in Matrix Form** - very common properties:

**Normal Equations**

**Derivation**

This is the least square metric.

For

Divide both sides by -2, and distribute the Σ.

System of equations

m=j

m=1

m=0

n=j

n=1

n=0

Fill in the blanks: First, second, third, forth

Previously ,

H is called the **hat matrix**.

**Residual**

Note it is not

**Implications**

Each column of X, including the columns of 1's is perpendicular to the residual vector .

The sum of the residuals is zero.

**Derivation**

This is the normal equation

is .

**ANOVA Definitions**

, where is .

SSR = sum of squares, regression

**The ANOVA Relationship**

**Derivation**

Property of transpose:

The two middle terms are the same because they are both .

It is sufficient to show

Property of transpose:

The two terms both zeros. Both cases using the identity .

For the second term, the in has a column of ones. That same identity is saying that this column of ones, when dot product with , will produce zero.

**Covariance Matrix**

a column vector of random variables with means

vector of the means

**Derivation**

The covariance matrix explores all possible combinations.

The different element for each row is .

Likewise, the different element for each row is .

These "all possible combination" matrices can be decomposed into two column vectors.

**Property:**

a matrix with constant entries

This is extending the linearity of the expectation operator to the matrix world.

**Property:**

**Derivation**

To move the "A" outside of E[ ], you will need , and .

Use

**Covariance of :**

Use

Let , which is a matrix consist of constants.

, meaning the transpose and the inverse operations can be interchanged.

Simplify further using

**Expected Value of :**

Use

The is a constant that can be factored out.

**Variance of :**

From previous:

Let

The variances are entries from the main diagonal, so that's why they are

**Distribution of**

is a linear combination of , so it will follow the normal distribution.

Replacing with will change from the normal distribution to the t-distribution.

**The Hat Matrix is Symmetric:**

Derivation:

You can derive this using and

**Property:**

Derivation:

The is a constant matrix.

**Property:**

Derivation:

The is a constant matrix.

, and both matrices are symmetric.

**Leverage**

The term measures the influence of the i-th observation on its predicted value. The diagonal elements of are called the **leverages**.

An observation with high leverage will pull the regression toward it, and its residual will be small.

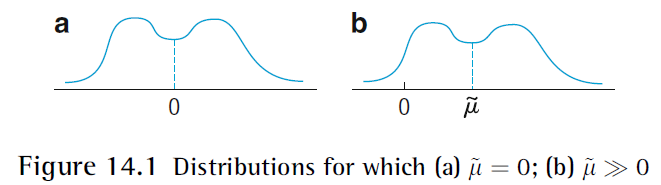
# Alternative Approaches to Inference

## The Wilcoxon Signed-Rank Test

**Assumption**

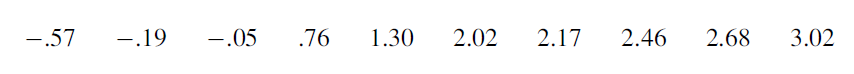
This test does not assume a particular distribution. But it does assume the distribution is symmetric, so there should be equal data on two sides of the true median . It's then further assumed that .

The null hypothesis is .

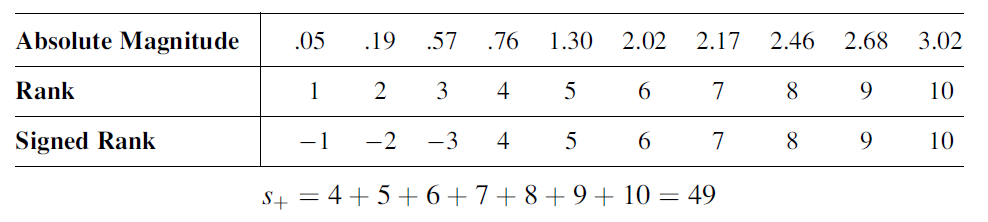


**Computation of S+**

The data



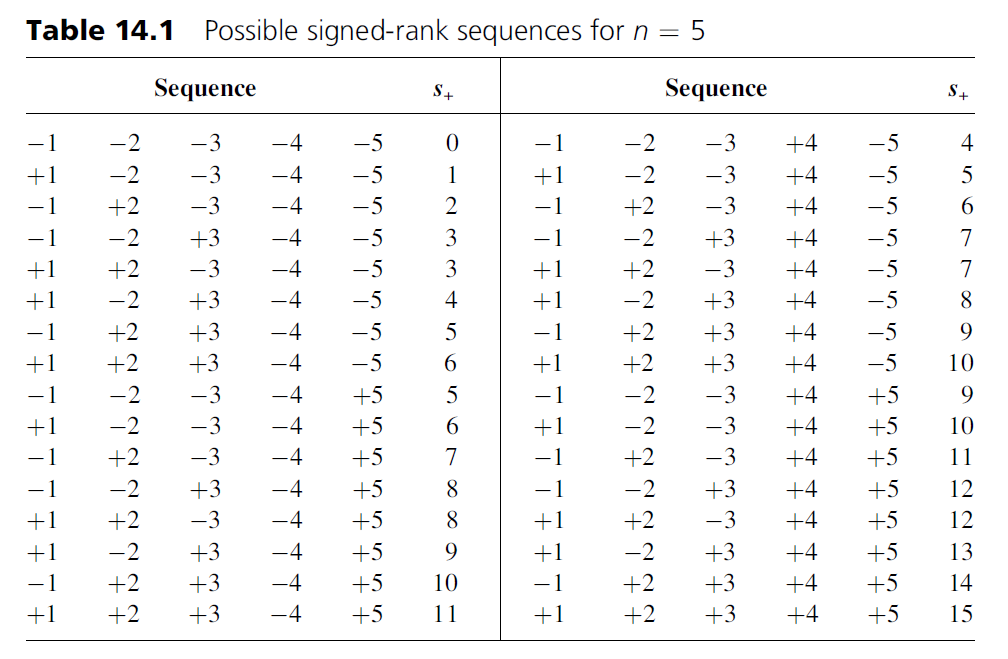
is sorted in terms of absolute value, and assigned a rank of 1 to 10.



The ranks are given a sign based on the value of the data. The S+ is a sum of the positive signed ranks.

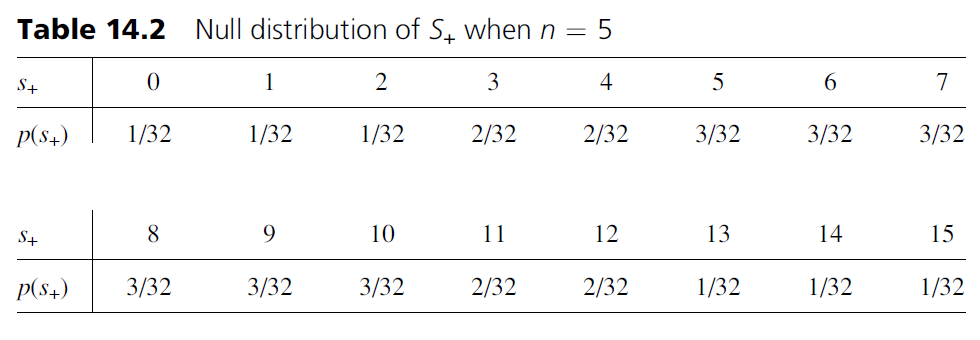
**The Distribution of S+**

The distribution can then be generated by listing all possible situations, and their S+.



The key argument is that when is true, the number of positive and negative values should be approximately equal. So each of these 32 situations are equally likely. On the other hand, if is true, then we cannot say the negative values are as likely as positive ones.

Assuming that each of these 32 situations are equally likely, counting the S+ will produce the probability distribution.



If the rejection region is set as , the probability is 3/32, meaning an alpha level of ≈ 0.1.

**Handling**

If the null hypothesis is not , but some finite instead, then the data needs to be transformed into . The signed ranks are computed from .

**Handling Ties**

If two data values, +10 and -10, are at rank 2 and rank 3 because they have the same absolute value, then the rank assigned to them will be +2.5 and -2.5.

## The Wilcoxon Rank-Sum Test

**Assumptions**

* Two independent random variables and .
* The and distributions have the same shape (and spread). They differ only in their means and .

Suppose there are three values, and four values.

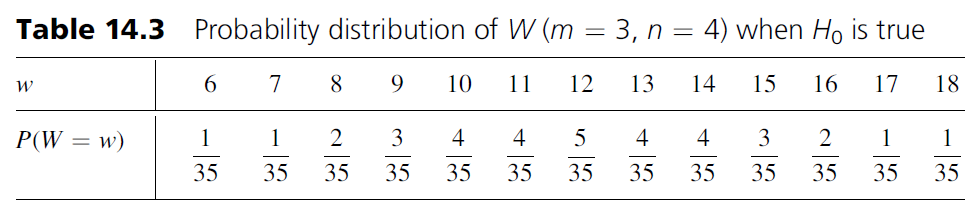
The values are sorted, and given a rank. The smallest of these 7 numbers is given rank 1, the largest number is given rank 7.

The test statistic is , the sum of the values' ranks.

The smallest possible . The largest possible value is 18.

There are possible rank triples.

Due to the assumption, the 35 possible rank triples are all equally likely.



The only way to get rank #6 is .

To get rank #9, we have .

If the rejection region is set as {17, 18},

**Handling instead of**

The data should be transformed into . Then we can assume the 35 ranks are equally likely.

## Bayesian Methods

**Bayes' Theorem**

the prior distribution about the statistic , based on knowledge up to the point of the experiment

the distribution of the experiment data

the posterior distribution about the statistic , based on both the prior knowledge and the experiment's data

**Example 14.7**

The experiment is a survey where 803 out of 1574 people said yes.

Model for the proportion prior to the experiment is a standardized beta distribution:

This distribution runs from to 1.

The "a" and "b" parameters need to be decided based on whatever is known before the experiment.

Distribution of the data:

This is the binomial distribution with x = 803 and n = 1574

**Example 14.8**

The experiment is a collection of IQ test scores for a group of first graders.

The population is .

Prior to the experiment, a mini survey with four entries was done. The result is . Since there are four entries, we can estimate that for this group of first graders, .

The model for prior to the experiment is .

The distribution of the data from the experiment is

# Formula Reference

## Descriptive Statistics

**Sample Variance**

## Probability

**Union of Two Events**

**Permutation**

**Combination**

**Conditional Probability**

**The Law of Total Probability**

**Bayes' Theorem**

**Independent Events**

## Discrete Random Variables

**Expected Values**

**Expected Value of the Linear Function**

**Variance and Standard Deviation**

**Rules of Variance**

**Moments**

**MGF of**

**Binomial Probability Distribution**

**Hypergeometric Probability Distribution**

**Negative Binomial Probability Distribution**

**Poisson Probability Distribution**

The Poisson is an approximation of the Binomial if n > 50 and np < 5.

## Continuous Random Variables

**Expected Value**

**Variance V(X), Standard Deviation**

**Approximating the Mean Value and Standard Deviation for**

**Moment Generating Function**

**Normal Distribution**

**Standard Normal Distribution**

**Normal Approximation for the Binomial Distribution**

The approximation is good if BOTH AND .

**The Gamma Function**

The most important properties of the gamma function are:

1. For α > 1, Γ(α) = (α-1) \* Γ(α-1)
2. For positive integer n, Γ(n) = (n-1)!

**Gamma Distributions**

The **standard gamma distribution** has β = 1.

**Exponential Distribution**

Memoryless Property:

**Weibull Distribution**

**Lognormal Distribution**

**Beta Distribution**

**Transformations of a Random Variable**

## Joint Probability Distributions

**Joint PMF and PDF**

**Marginal PMF and PDF**

**Multinomial Experiment**

**Expected Values**

**Expected Values for Independent Variables**

**Covariance**

**Correlation**

1. Corr(X,Y) = 1 or -1 iff for some numbers a and b.
2. If X and Y are independent, then Corr(X,Y) = 0, but correlation being zero does not imply independence.

**Conditional PMF and PDF**

**Law of total expectation**

**Law of Total Variance**

**Bivariate Normal Distribution**

**Transformations of Random Variables**

Input random variables X1, X2  
Output random variables Y1, Y2  
Mapping functions from Y to X: and   
The pdf of the two original variables: f(x1, x2)  
The pdf of the two new variables: g(y1,y2)  
Region T of positive density for g(y1,y2): T = {(y1,y2) : g(y1,y2) > 0}

**Order Statistics**

The **order statistics** are random variables Y1, Y2, ... Yn defined by:

Y1 = the smallest among X1, X2, ..., Xn  
Y2 = the second smallest among X1, X2, ..., Xn  
.  
.  
.  
Yn = the second smallest among X1, X2, ..., Xn

**The CDF of Yn**

**The CDF of Y1**

**Joint Distribution of the n Order Statistics**

**Distribution of a Single Order Statistic**

## Sampling Distributions

**Sample Mean Characteristics**

Let X1, X2, ..., Xn be a random sample from a distribution with mean value µ and standard deviation σ.

**Sample Total Characteristics**

**Law of Large Numbers**

**Linear Combination of Random Variables**

Whether or not the Xi's are independent:

If the Xi's are independent:

Generalization of the Central Limit Theorem: If "n" is sufficiently large, then Y has approximately a normal distribution.

**Distributions Based on a Normal Random Sample**

Suppose we have five independent standard normal random variables: Z1, Z2, Z3, Z4, and Z5.

Chi-squared distribution:

t-distribution:

F distribution:

**Chi-Squared Distribution**

The chi-squared distribution is a special case of the gamma distribution, with α=v/2, β=2.

If Z has a standard normal distribution, then Z2 is chi-squared with 1 df, .

If X1 ~ χ2v1, X2 ~ χ2v2, and they are independent, then X1 + X2 ~ χ2v1+v2.

Let Z1, Z2, ..., Zn be "n" independent standard normal random variables, then .

The peak of fχ2(x) is at

If X1, X2, ..., Xn are a random sample from a normal distribution, then and S2 are independent, where

If X1, X2, ..., Xn are a random sample from a normal distribution N(µ, σ2), then .

**t-Distribution**

Let be a standard normal random variable.  
Let be a Chi-squared random variable independent of "Z", with degree of freedom.

The t-distribution is standardized .

Let X1, X2, ..., Xn be a random sample from a normal distribution , then

has the t-distribution with (n-1) degrees of freedom, tn-1.

PDF of the t-distribution

When v=1, the pdf is . This is called the Cauchy distribution and the tail is so fat that the mean does not exist.

As approach ∞, the t-distribution approaches the standard normal distribution.

E(Xk) of chi-squared distribution

**F-Distribution**

The F distribution describes ratio of chi-squared random variables.

Reciprocal properties:

t-distribution squared is F distribution:

## Point Estimation

**Mean Squared Error of an Estimator**

**Neyman Factorization Theorem**

**Fisher Information**

**Information from a Random Sample**

**Cramer-Rao Inequality**

Given:

Random sample with pmf or pdf

Statistic is an unbiased estimator for the parameter .

**Large Sample Properties of the MLE**

The mle is asymptotically normal with mean and variance equal to the Cramer-Rao lower bound.

## Single Sample Intervals

**95% Confidence Interval based on the Normal Distribution**

**Population Proportion Confidence Interval**

The quantity is maximum if

**Prediction using**

Suppose that we have a random sample that came from a normal population distribution with parameters and . To predict we use .

The prediction interval (PI) is

**Confidence interval for estimate of a normal population**

## Single Sample Test of Hypotheses

**Type I Error** - rejecting the null hypothesis when it is true.

**Type II Error** - failing to reject the null hypothesis when it is false.

- Error probabilities for type I and type II errors.

**Power**

When the null hypothesis is true, .

The **P-value** is the probability in the tail(s).

Another name for the P-value is **observed significance level (OSL)**.

When the is true, the distribution of the P-value is uniform.

**Power Function**

## Two Sample Inferences

**Difference between Means**

**For normally distributed , width of the confidence interval**

**Two-Sample t-Test**

For small sample and normally distributed and

**Pooled t Procedure**

Assume the two samples have the same variance, so that .

The standardized is

Replacing with

will produce a t-distribution, with degree of freedom .

**Dependence between and**

**Difference between Proportions**

Let and with and independent variables.

**Ratio of Variances**

If the underlying population is normal, then the random variable

has an F-distribution with and .

**Confidence Interval of**

## ANOVA

**Single-Factor ANOVA**

the data value from the i-th population's j-th measurement

**Means**

**Sample Variances**

**Assumption:** The population or treatment distributions are all normal with the same variance .

**treatment sum of squares** - variation between treatments

**error sum of squares** - variation within each treatment

**Distributions**

**mean square for treatment**

If is true, then

**mean square for error**

**The F Test**

When is true, the F random variable is F-distribution with parameters . Its value should be close to 1.

**Computational Formulas**

|  |  |  |
| --- | --- | --- |
|  | **Definition** | **Computing Formula** |
| **Sum of Squares Total** |  |  |
| **Sum of Squares Treatment** |  |  |
| **Sum of Squares Error** |  |  |

represent the sum of the for fixed   
 represent the sum of all IJ observations - the **grand total**

**Fundamental ANOVA Identity**

**Multiple Comparisons in ANOVA**

The general idea is to calculate a confidence interval for each pairwise difference . If this interval does not include 0, conclude that and **differ significantly**.

**Studentized Range Distribution**

**Mean Difference Confidence Intervals**

The confidence intervals for are

There are such intervals.

**Tukey's Procedure**

The confidence intervals are

If two means differ by less than , then they do not differ significantly.

Sort the means and underline pairs that differ by less than .

**Confidence interval for**

**More on Single-Factor ANOVA**

**Alternative Description of the ANOVA Model**

**Data Transformation**

If , a known function of , then the transformation

will "stabilizes the variance" so that is approximately constant.

**A Random Effects Model**

**Two-Factor ANOVA with**

number of observations on the treatment consisting of factor at level and factor at level . For this section .

the random variable (rv) denoting the measurement when factor is held at level and factor is held at level

**Means**

**The Model**

where and

The 's are independent, normally distributed, with mean 0, and common variance .

**Estimators**

**Hypotheses**

versus at least one

versus at least one

**Test Procedures**

**predicted values**, also called **fitted values**

**Expected Mean Squares**

**Two-Factor ANOVA with**

**Fixed Effects Model, with Interaction**

the effect of factor A at level the effect of factor B at level

Restrictions:

The 's are normally distributed random variables, with mean 0 and variance . The other parameters are fixed values.

**Hypotheses**

for all

**Analysis**

variance within group

variance between group means and the grand mean

|  |  |
| --- | --- |
| **Hypotheses** | **Test Statistics** |
| versus |  |
| versus |  |
| versus |  |

## Regression and Correlation

**Linear and Logistic Models**

**The Simple Linear Regression Model**

**Logistic Regression Model**

The **logit function**

**Estimating Model Parameters**

**Least Squares Model Parameters**

the **normal equations**:

where is the total number of data points

The solutions are:

**Estimating**

The **error sum of squares**

is an unbiased estimator for , but it is biased for . The mle of divides by rather than .

**The Coefficient of Determination**

**Total sum of squares**

SSE = the variations not explained by the linear model

SST = the total variations

**Coefficient of determination**

**Regression sum of squares (SSR)**

**Additional Identities**

Denote the residuals as , where

**Inferences about the Regression Coefficient β1**

**The Estimator for β1**

The simple linear regression model is

This β1 is a constant. The ε and Y are the random variables.

**Distribution of**

is normally distributed

**in hypothesis testing as a t-Distribution**

The standardized variable

has a t-distribution with (n-2) df.

**Inferences Concerning Y**

**Distribution of**

is normally distributed

**Hypothesis Testing of**

**Combined Confidence Intervals**

Suppose we are estimating two values and each value has a 95% confidence interval.

A result called the Bonferroni inequality says the rejection probability is at most , and so the confidence interval is at least 90%.

**Prediction Interval for**

**Correlation**

**Sample Correlation Coefficient r**

**Population Correlation Coefficient**

**Properties of Residuals**

The residual

**Standardized Residuals**

For large n

**General Additive Multiple Regression Model**

is normally distributed. The 's and 's are independent.

**Coefficient of Multiple Determination**

**Regression with Matrices**

**Multiple Regression in Matrix Form**

**Normal Equations**

**Residual**

**ANOVA Definitions**

, where is .

**The ANOVA Relationship**

**Covariance Matrix**

a column vector of random variables with means

vector of the means

a matrix with constant entries

**Characteristics**

**The Hat Matrix**

**Leverage:**

# Python Reference

## Mean, Variance, Median

|  |  |
| --- | --- |
| Mean, variance, and standard deviation of | import numpy  x = numpy.array([1, 2, 3])  x.mean() #2  numpy.var(x) # 2/3 (population variance)  numpy.var(x, ddof=1) # 2/2 (sample variance)  numpy.std(x) # population standard deviation  numpy.std(x, ddof=1) # sample standard deviation |
| Mean of | x = numpy.array([[1, 2, 3],  [9, 10, 11]])  x.mean(0) # [5, 6, 7] - column mean  x.mean(1) # [2, 10] - row mean  x.mean() # 6 - overall mean |
| Trimmed mean | x = numpy.array([3, 3, 3, 3, 3, 3, 3, -10000])  x\_sorted = numpy.sort(x) # [-10000, 3, 3, ...]  # remove 10% of highest and lowest data  data\_start = len(x) // 10 + 1  x\_trimmed = x[data\_start:-1\*data\_start] # [3, 3, ...]  trimmed\_mean = np.mean(x\_trimmed) # 3 |
| Median of | x = numpy.array([[1, 2, 5],  [9, 10, 13],  [20, 21, 24]])  numpy.median(x, 0) # [9, 10, 13] - column median  numpy.median(x, 1) # [2, 10, 21] - row median  numpy.median(x) # 10 - overall median |

## Permutation, Combination

|  |  |
| --- | --- |
|  | n = 5  k = 3  p = 1  for i in range(n, n-k, -1):  p \*= i  print(p) # 60 |
|  | import scipy.misc as misc  misc.comb(5,3) # 10  import math  n = 5  k = 3  c = 1  for i in range(n, n-k, -1):  c \*= i  c /= math.factorial(k)  print(c) # 10 |

## Matrix Products

|  |  |
| --- | --- |
|  | x = numpy.array([1, 2, 3, 4])  p = numpy.array([0.25, 0.25, 0.25, 0.25])  x.dot(p) # 2.5  # same as x.mean() |
|  | v1 = numpy.array([[1, 2]])  v2 = numpy.array([[1],  [2]])  v1.dot(v2) # array([[5]]) |
|  | v2.dot(v1)  # array([[1, 2],  # [2, 4]]) |

## Distributions

|  |  |
| --- | --- |
| Binomial Distribution | import scipy.stats as stats  stats.binom.ppf(0.125, 3, 0.5) # 0  stats.binom.ppf(0.126, 3, 0.5) # 1  stats.binom.cdf(0, 3, 0.5) # 0.125  stats.binom.cdf(1, 3, 0.5) # 0.5 |
| Normal Distribution | import scipy.stats as stats  # percent point function  z = stats.norm.ppf(0.99) # 2.326  percent = stats.norm.cdf(2.326) # 0.99 |
| Gamma Distribution | from scipy.stats import gamma  x = 5  alpha = 2  gamma.cdf(x, alpha) # 0.95957  gamma.cdf(5, a=2) |
| Chi-Squared Distribution | import scipy.stats as stats  stats.chi2.cdf(9.21, 2) # 0.99  stats.chi2.ppf(0.99, 2) # 9.21 |
| t-Distribution  t0.048,2 = 3.0  P(T > 3.0) = 0.048 | import scipy.stats as stats  stats.t.cdf(3, 2) # 0.952267  # tail probability is then 0.048  stats.t.ppf(0.952267, 2) # 3 |
| F Distribution  F0.1,1,2 = 8.53  where v1 = 1, v2 = 2 | import scipy.stats as stats  stats.f.cdf(8.53, 1, 2) # 0.90  # tail probability is then 0.10  stats.f.ppf(0.90, 1, 2) # 8.526 |

## Solving Equations

Script for solving equations using the bisection method.

To solve , use

solve(lambda x: x, lambda x: x\*\*2, 0.5, 0.5)

import math

def solve(left\_equ, right\_equ, x1, x2):

# check orientation

y1 = left\_equ(x1) - right\_equ(x1)

y2 = left\_equ(x2) - right\_equ(x2)

# exit if there is no sign difference

if (y1 > 0 and y2 > 0) or (y1 <0 and y2 < 0):

print("No intersection found.")

return

# make x1 produce negative result, x2 positive result

if y1 > y2:

x1, x2 = x2, x1

y1, y2 = y2, y1

# keep running bisection method until error is small

error = y2 - y1

while error > 10\*\*-10:

# try new\_x value halfway between x1, x2

new\_x = (x1 + x2) / 2

new\_y = left\_equ(new\_x) - right\_equ(new\_x)

# the new\_x will replace either x1 or x2

if new\_y > 0:

x2 = new\_x; y2 = new\_y

else:

x1 = new\_x; y1 = new\_y

# update the error value

error = y2 - y1

# check for the double solution situation

if y1 == 0.0 and y2 == 0.0:

print("Two solutions found: " + str(x1) + ", " + str(x2))

return

# standard ending - print result

final\_answer = (x1 + x2) / 2

print("{0:.4g}".format(final\_answer))

# test cases:

# solve(lambda x: x, lambda x: x\*\*2, 0.5, 0.5)

# 0, 1 - two solutions found

# -1, 1 - two solutions found

# -0.5, 0.5

# 0.5, 1.5

# 1.5, 2 - no intersection found

# 0.5, 0.5 - no intersection found