JARED SCHULTZ

Statistics Concepts

An Overview of Upper-division Statistics with R

This is a brief overview of upper division statistics with some examples. This book does not cover anything as in-depth as textbooks but instead gives the main formulas and rules used in each chapter with some examples. To contact me please email me at jaredschultz@outlook.com.

CONTENTS

Chapter 1 : Basic Ideas

Definitions

Statistics : is the study of procedures for collecting, describing and drawing conclusions for information.

Population : is the entire collection of individuals about which information is sought.

Sample : is a subset of the population containing the individuals that are actually observed.

Simple Random Sample : of size n is a sample chosen by a method in which each collection of n populations items is equally likely to make up the sample ex. Lottery

Sample of Convenience : is a sample that is not drawn by a well defined method.

Statistic : is a number that describes a sample.

Parameter : is a number that describes a population.

Qualitative Variables : classify into categories

Quantitative Variables : tells us quantity either discrete or continuous.

Nominal Data : a type of qualitative data that is distinguished by having no order

Ordinal Data : a type of qualitative data that is distinguished by having order ex. US States

Chapter 2: Probability

Definitions

Trial - Each occasion we observe a random phenomenon.

Outcome - At each trail, we note the value of the random phenomenon.

Sample Space (s) - Collection of all possible outcomes.

Event - Combinations of outcomes (subset of the sample space).

Disjoint or mutually exclusive - A and B are disjoint events if they have not outcomes on common. ie: $P(A \cap B) = 0$

Dependent - The outcome of the first event affects the outcome of the second event

Independent - The outcome of the first event doesn't affect the outcome of the second event. ie: $P(A \cap B) = P(A) * P(B)$

Complement - The complement of an event A is denoted A' is the set outcomes not in an event A. ie: $P(A') = 1 - P(A)$

Probability

Probability of an event A :

$$P(A) = \frac{\text{Number of outcomes in A}}{\text{Number of possible outcomes}}$$

Example :

We want to roll a fair 6 sided die. What is the probability of rolling a 1

Solution:

s = {1,2,3,4,5,6}

P(1) = 1/6

Probability Rules

1. A probability is number within 0 and 1 for any event A.

2. The probability of the set of all possible outcomes of a trail must be 1.

3. Complement rule. $P(A) = 1 - P(A')$

Union and Intersections :

A **Union** ∪ is an “or” statement.

$P(A \cup B) = P(A) + P(B)$, where A & B are *disjoint*.

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$, where A & B are *dependent*.

An **Intersection** ∩ is a “and” statement.

$P(A \cap B) = P(A) * P(B)$, where A & B are *independent*.

$P(A \cap B) = P(A|B) * P(B)$ or $P(A \cap B) = P(B|A) * P(A)$, where A & B are *dependent*.

Conditional Probability

The conditional probability of an event **A given B** is : $P(A|B) = \frac{P(A \cap B)}{P(B)}, P(B) \neq 0$

The conditional probability for independent event A given B is : $P(A|B) = P(A)$

Example :

Suppose there is a very reliable test for a certain type of cancer. If I have this cancer, then the test will be positive with probability 0.96. If I do not have the test, then the test will be negative with probability 0.94. Also, it is assumed that 1/145 people in my age group have the cancer with out knowing. If I get the test and it comes out positive, what is the probability that I have cancer.

Solution:

S : sample space

C : event of cancer

N : event of no cancer

“+” : positive test

“-” : negative test

We wish to know *conditional probability that cancer is present given a positive result* : P(C|+)

By definition:

$P(C|+) = \frac{P(C\cap+)}{P(+)}$, where $P(C\cap+) = P(+|C) * P(C) = (0.96)(1/145)$, and $P(+) = P(C\cap+) + P(N\cap+) = (0.96)(1/145) + (0.06)(144/145)$.

thus, $P(C|+) = 0.1$

Counting Techniques

Product rule : If the first element or object of an ordered pair can be selected n ways the second m ways. The number of pairs n*m

Example :

Given 5 shirts each having 12 pants and 9 hats, how many possible outcomes exists?

Solution:

N = (5)(12)(9) = 540

Permutations and Combinations

Combinations are when order *is not* important. ("combinations don't care")

$$C_r^n = \frac{n!}{r!\,(n-r)!}$$

Permutations are when order *is* important.

$$P_r^n = \frac{n!}{(n-r)!}$$

Special Permutations when you have different sizes in your group.

$$S_{k_1\ldots m}^n = \frac{n!}{k_1 * k_2 * \ldots * k_m}$$

Example :

How many different ways can you arrange " Tennessee "

Solution:

$$S = \frac{9!}{1! * 4! * 2! * 2!} = 60480$$

Chapter 3: Discrete Random Variables and Their Probability Distributions

Definitions

Random Variable : is a real valued function defined over a sample space. they are either discrete or continuous

Discrete Random Variable : A random variable Y is said to be discrete if it can assume only a finite or countably infinite number of distinct values.

Probability Distribution of a Discrete Random Variable (pmf): is a list of probabilities associated with each of its possible values. Also called a Probability mass function.

P(Y = y) or p(y) :"The probability that Y takes on the values y" is defined as the sum of the probabilities of all the sample points in S that are assigned the value y.

Probability Distributions : For a discrete variable Y can be represented by a formula , table or graph that provides p(y) for all y.

Expected value E(x) or mew of a Discrete Rv : Calculated as the sum of all possible values each multiplied by their probability of occurrence. $E(x) = \sum\left(x * p(x)\right)$

Outcome	Payout	Probability	Deviation
Death	100,000	0.001	99,800
Disability	50,000	0.002	49,800
neither	0	0.997	-200

Example

Given the data table find the expected annual payout on the policy. Then find the SD(X).

Solution :

On R we can use the following command `sum(data$Payout*data$prob)`.

E(x)= 100,000(0.001) + 50,000(0.002) + 0(.997) = 200. We expect the insurance company will payout $200 a year.

On R to find variance we can use `sum(data$Prob*(data$Deviation)^2)`

99,800^2(.001) + 49,800^2(.002) + (-200)^2(.997) = 14,960,000

SD(X) = $\sqrt{14{,}960{,}000} \approx$ 3967.82

Variance and Expected Values

Variance of a Rv, with E(X) = k : $\sigma^2 = Var(X) = E(X-k)^2 = E(X^2) - E(X)^2$

Variance of a discrete Rv : $\sigma^2 = Var(X) = \sum(X-k)^2 p(X)$

Proof:

E(X-k)^2 = E(X^2) - E(X)^2

$$E(X-k)^2 = E(X^2 - 2Xk + k^2)$$

$$= E(X^2) - 2kE(X) + E(X)^2$$

$$= E(X^2) - 2E(X)^2 + E(X)^2$$

$$= E(X^2) - E(X)^2$$

Properties of Expected values and Variances

Adding a Constant C

$$E(X \pm C) = E(X) \pm C$$

$$Var(X \pm C) = Var(X)$$

Multiplying by a Constant C

$$E(CX) = C * E(X)$$

$$Var(CX) = C^2 * Var(X)$$

For SD we use take |C|

Bernoulli and Binomial Distributions

- Both distributions are discrete distributions.

3 assumptions of Bernoulli Distributions

1. Each trial has two outcomes: Success or Failure.
2. Trials are independent.
3. On each trial the probability of success is p and failure is 1-p with $p \in [0,1]$

The Bernoulli distribution pmf can be expressed as:

$$P(X) = \begin{cases} p & \text{if X = 1} \\ 1-p & \text{if X = 0} \end{cases}$$

Mean : p

Variance : p(1-p)

4 assumptions of Binomial Distributions

1. The experiment consists of a sequence of n smaller experiments called trials, where n is fixed in advanced.
2. Each trial has two possible outcomes: Success or Failure
3. Trials are independent.
4. The Probability of success P(S) is constant from trial to trial, denoted probability p.

Binomial rv : the number of Successes among n trials

X ~ Bin(n,p) : indicates the X is a binomial rv based on n trails and success probability p.

Mean : n*p

Variance : n*p(1-p)

The binomial distribution pmf can be expressed as:

$$b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

with, x = 0, 1, 2, …, n and $p \in [0,1]$

Example

suppose we are tossing a fair coin and we decided to toss a coin 3 times. We denote heads to be successful and tail to be a failure. In this case what is the probability of only having 1 success.

Solution :

First our p = .5 since its a fair coin. Our trial size or n = 3 and we want x = 1

using the formula we get $b(1; 3{,}0.5) = \binom{3}{1}(.5)^1\left(1 - (.5)\right)^{3-1} = .375$

In R we can use : `dbinom(1,3,0.5)`.

or we can write out: s = { SSS SSF SFS SFF FSS FSF FFS FFF } thus,

P(X =1) = 1/8 + 1/8 + 1/8 = .375

Poisson Probability Distribution

3 assumptions of Poisson Distributions

Allow X to be a random variable that represents the number of events that occur in a time interval of length t. Then X will have the probability distribution if :

1. The average rate at which events occur is the same at all times.
2. The number of events that occur in nonoverlapping time intervals are independent.
3. For a very short interval of t:

a. It is essentially impossible for more than one event to occur within the time interval

b. The probability that one event occurs in the interval is approximately equal to L*t where L is the average rate at which events occur.

The Poisson distribution pmf can be expressed as:

$$p(x; L, t) = e^{-Lt}\frac{(Lt)^x}{x!}$$

where X = 0,1,2,3,... and not that often we write M = L*t to simplify the formula.

Example

Let X denote the number of defects on a machine part and it has a Poisson distribution with L*t = M = 2 per day. Find the probability that there at most 3 defects

Solution :

$$P(X \leq 3) = \sum_{x=0}^{3}\frac{e^{-2}2^x}{x!} = 0.135 + 0.271 + 0.271 + 0.18 = .857$$

or using r : `ppois(3,2,lower.tail = T)`.

Note: given a binomial pmf bin(x;n,p) if we let n –> inf and p –> 0 where n*p approaches a value M then bin(x;n,p) ~ Pos(x;M). As a general rule of thumb the approximation can be applied if n > 50 and* np < 5

Mean: M = L*t

Variance: M = L*t

Geometric Probability Distribution

4 assumptions of Geometric Distributions

1. The experiment involves identical and independent trials
2. Only 2 outcomes: Success or Failure
3. The probability of success is equal to p and is constant from trial to trial.
4. The geometric Rv X is the number of trials on which the first success occurs. ie FFFFFFF….FFS

The Geometric distribution pmf can be expressed as:

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

where x = 1,2,3,…

Mean : 1/p

Variance : $\frac{1-p}{p^2}$

Example

Suppose that 30% of people get accepted to UC Davis. Applicants are reviewed sequentially from a random pool. Find the probability that the first accepted student will be on the 5th interview.

Solution :

$$P(X = 5) = .3(1 - .3)^{5-1} = .072$$

or in r : `dgeom(4,.3)`

Note the on the functions our pmf = dgeom and cdf = pgeom

Hypergeometric Probability Distribution

3 assumptions of Hypergeometric Distributions

1. The population of the set is finite and consists of N individuals, objects or elements.

2. Each individual can be characterized as S or F and there are M Successes in the population.

3. A sample size of n individuals are selected without replacement in such a way each subset of size n is equally likely to be chosen.

The Hypergeometric distribution pmf can be expressed as:

$$p(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

where x = 0,1,2,3,… and $x \leq M$ and $n - x \leq N - M$

note:

N : population size

M : number of successes

n : sample size

x : number of successes in the sample of n

Mean : $\frac{n*M}{N}$

Variance : $\frac{n*M}{N}\left(\frac{N-n}{N-1}\right)\left(1 - \frac{M}{N}\right)$

Example

During a particular period the UC Davis IT department received 20 repair orders on computers. Of the 20 8 are Apple OS and 12 are Windows OS. A random sample of 5 computers are selected to be repaired that day. What is the probability that exactly 2 of these are Windows OS.

Solution :

given:

N : population size = 20

M : number of successes = 12

n : sample size = 5

x : number of successes in the sample of n = 2

M-N : 12

$$p(X=2)=\frac{\binom{12}{2}\binom{20-12}{5-2}}{\binom{20}{5}}=\frac{77}{323}=.238$$

or using r: `dhyper(2,12,8,5)`

Moments and moment generating functions

Definitions

Kth moment of a Rv X (about origin): is defined to be $E(X^k)$ denoted μ_k'

Kth moment of a Rv X (about mean): is defined to be $E[(X-\mu)^k]$ denoted μ_k

The *First Moment* about the origin is $E(X)=\mu$

The *Second Moment* about the origin is $E(X^2)$

Variance : $E[(X-\mu)^2]$

Moment Generating Function : of a Rv X is defined to be $m(t)=E[e^{tx}]$. We say that a moment generating function for X exists if there is a positive constant b s.t m(t) is finite for $|t|\leq b$.

Note: IN general the nth derivative of m(t) evaluated at t=0 is $E[X^n]$ for n grater than 1.

Tchebusheff's Theorem : Let X be a Rv with mean μ and finite variance σ^2. Then for all k>0

$$P(|X-\mu|<k\sigma)\geq 1-\frac{1}{k^2} \text{ or } P(|X-\mu|<k\sigma)\leq\frac{1}{k^2}$$

This thm allows us to get boundaries.

Example

let $m(t) = \frac{1}{6}e^{t} + \frac{2}{6}e^{2t} + \frac{3}{6}e^{3t}$. Find E(Y), Var(Y).

Solution :

a. $E(Y) = m'(0) = \frac{1}{6}e^{0} + \frac{2*2}{6}e^{2*0} + \frac{3*3}{6}e^{3*0} = \frac{1}{6} + \frac{4}{6} + \frac{9}{6} = \frac{7}{3}$

b. $E(Y^2) = m''(0) = \frac{1}{6}e^{0} + \frac{2*2*2}{6}e^{2*0} + \frac{3*3*3}{6}e^{3*0} = \frac{1}{6} + \frac{8}{6} + \frac{27}{6} = 6$

$Var(Y) = E(Y^2) - E(Y)^2 = 6 - \frac{49}{9} = \frac{5}{9}$

Example

suppose we know the number of items produced in a factory during a week is a Rv X with mean 500 and variance of 100. What can be said about the probability that this weeks production will be between 400 and 600?

Solution :

we want $P(400 \leq X \leq 600)$ thus we will use Tchebysheffs thm.

$$P(400 \leq X \leq 600) = P(|X - \mu| < k\sigma) = P(mu - k\sigma < X < mu + k\sigma)$$

$$\mu + k\sigma = 600 => k = 10$$

$$P(400 \leq X \leq 600) = 1 - \frac{1}{100} = .99$$

Chapter 4: Continuous Random Variables and Their Probability Distributions

Cumulative Distribution Function

The Cumulative Distribution Function (CDF) : Let X denote any random variable. The distribution function of X, denoted by F(x), is such that

$$F(x) = P(X \leq x) \text{ for } -\infty < x < \infty$$

Example

x	p.x.
0	0.20
1	0.25
2	0.30
3	0.15
4	0.10

What is F(1) and find F(x) the cumulative distribution function ?

Solution :

$$F(1) = P(X \leq 1) = P(X = 0) + P(X = 1) = 0.20 + 0.25 = 0.45$$

$$F(x) = \begin{cases} 0 & \text{if x < 0} \\ .20 & \text{if } 0 \leq x < 1 \\ .45 & \text{if } 1 \leq x < 2 \\ .75 & \text{if } 2 \leq x < 3 \\ .90 & \text{if } 3 \leq x < 4 \\ 1 & \text{if } x \geq 4 \end{cases}$$

Properties of CDF

If F(y) is a CDF then,

1. $F(-\infty) = \lim_{y \to -\infty} F(y) = 0.$

2. $F(\infty) = \lim_{y \to \infty} F(y) = 1.$

3. F(y) is a *non-decreasing* function of y. $[y_1 < y_2 \Rightarrow f(y_1) \leq f(y_2)]$

Continuous random variable (CRV) : is one that takes an infinite number of possible values.

Density Function of a CRV

Allow F(x) be a CDF of a CRV X. Then f(x) the probability density function is given by:

$$f(x) = \frac{dF(y)}{dy} = F'(y)$$

For a function f(x) that represents the density curve we know

1. The function must have no negative values.

2. The total area under the density function is 1.

Properties of a CRV

Mean:

$$E(X) = \int x \cdot f(x)dx$$

$$E\big(g(x)\big) = \int g(x) \cdot f(x)dx$$

Variance:

$$Var(X) = E(x-\mu)^2 = \int (x-\mu)^2 \cdot f(x)dx \text{ or}$$

$$E(X^2) - E(X)^2 = \int x^2 \cdot f(x) - \left[\int x \cdot f(x)\right]^2$$

Example

The pdf of a CRV is given as

$$f(x) = \begin{cases} \frac{3}{2}(1-x^2) & \text{if } 0 \leq x \leq 1 \\ 0 & \text{Otherwise} \end{cases}$$

Find the E(x), Var(x) and F(x).

Solution :

a. E(x)

$$E(x) = \int x \cdot f(x)dx = \int_0^1 x \cdot \frac{3}{2}(1 - x^2)dx, \text{u = 1-x^2 du = -2x}$$

$$\frac{-3}{4}\int_1^0 u \cdot du = \frac{-3}{8}(0 - 1) = \frac{3}{8}$$

or in r :

```
integrand = function(x){x*(1.5*(1-x^2))}
```

```
integrate(integrand,lower = 0, upper = 1).
```

b. Var(X)

with r :

`integrand = function(x){x^2*(1.5*(1-x^2))}` = 1/5

$$V(x) = \frac{1}{5} - \left(\frac{3}{8}\right)^2 = 0.059$$

c. F(x)

integrating gives:

$$F(X) = \begin{cases} 0 & \text{x < 0} \\ \frac{3}{6}(3x - x^3) & \text{if } 0 \leq x \leq 1 \\ 1 & \text{x > 1} \end{cases}$$

Continuous Normal Distribution

The normal distribution is one of the most important probability and stats.

Normal distribution density function

Normal probability distribution: A Rv Y has a normal pmf iff σ >0 and $-\infty \leq \mu \leq \infty$.

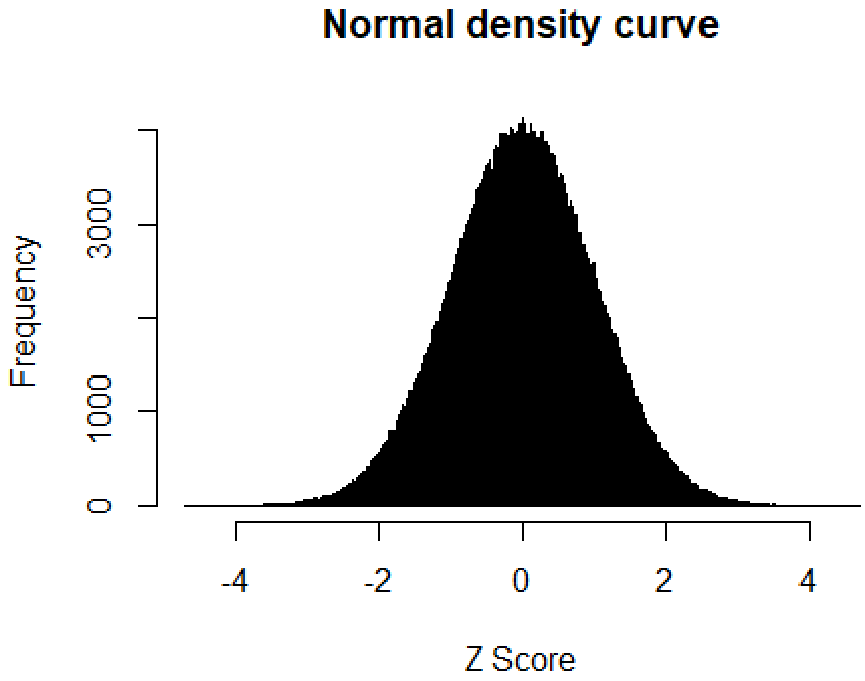
The density function of a pmf Y is

$$f(y) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\left(\frac{(y-\mu)^2}{2\sigma^2}\right)}$$

If we wish to find $P(a \leq x \leq b)$ then we integrate f(y) from a to b.

The mean, median and mode are all the same in a normal distribution.

Here is a plot of 1 million random normally distributed values plotted against their frequency.



Standard Normal Distribution

Standard Normal Distribution is a special case of a normal distribution with a mean of 0 and standard deviation of 1. Denoted by Z ~ $N(\mu = 0, \sigma^2 = 1)$

$$f(z) = \int \frac{1}{\sqrt{2\pi}} \cdot e^{-\left(\frac{z^2}{2}\right)}$$

Example

Calculating probabilities with z scores in R.

$P(Z \leq 1.25)$ = `pnorm(1.25)`

$P(Z > 1.25)$ = 1 - `pnorm(1.25)`

$P(-0.38 \leq Z \leq 1.25)$ = `pnorm(1.25)` - `pnorm(-0.38)`

Percentiles of Standard Normal Distribution

For critical values denoting percentiles we use $Z_\alpha = P(Z \geq Z_\alpha) = \alpha$

Thus $Z_{.10}$ captures the upper-tail area containing 10 percent of the area.

For a standard normal distribution Z_α is the $100(1-\alpha)th$ percentile.

Non-standard normal distribution

When X ~ $N(\mu, \sigma^2)$, probabilities involving X are computed by standardizing.

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

The Uniform Probability Distribution

If $\theta_1 < \theta_2$, a Rv Y is said to have a continuous uniform probability distribution on the interval (θ_1, θ_2) iff the density function of Y is :

$$f(y) = \begin{cases} \dfrac{1}{\theta_2 - \theta_1} & \text{if } \theta_1 \leq y \leq \theta_2 \\ 0 & \text{Otherwise} \end{cases}$$

A Poisson distribution on the interval (0,t) is approximately uniform given that one event has occurred in the interval.

Example

Suppose arrivals of customers at a check out counter follow a Poisson distribution and it is known that within the 30 min time period that one customer arrived at the check out counter. then find the probability that the customer arrived in the last 5 minutes of the 30 min interval.

Solution :

interval : (0,30)

In this example we can say that it is approximately uniform.

$$P(25 \le x \le 30) = \int_{25}^{30} \frac{1}{30} dx = \frac{1}{6}$$

note the arrival of any 5-minute interval is also $\frac{1}{6}$

Mean : $\mu = E(X) = \frac{\theta_1+\theta_2}{2}$

Variance : $\sigma^2 = Var(X) = \frac{(\theta_1-\theta_2)^2}{12}$

Gamma Distributions

A Rv Y is said to have a gamma distribution with parameters $\alpha > 0$ and $\beta > 0$ iff the density function of Y is :

$$f(y) = \begin{cases} \frac{y^{\alpha-1} \cdot e^{\frac{-y}{\beta}}}{\beta^{\alpha} \cdot \Gamma(\alpha)} & \text{if } 0 \le y \le \infty \\ 0 & \text{Otherwise} \end{cases}$$

Where,

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} \cdot e^{-y} dy.$$

A Gamma Rv Y is the waiting time until the α^{th} event occurs.

α : is the number of events for which you are waiting to occur. (Shape Parameter)

β : the rate of the events happening. (Scale Parameter)

Mean :$E(Y) = \alpha\beta$

Variance : $\alpha\beta^2$

Properties of the Gamma function

1. $\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} \cdot e^{-x} dx$
2. $\int_0^{\infty} x^{\alpha-1} \cdot e^{-\lambda x} dx = \frac{\Gamma(\alpha)}{\lambda^{\alpha}}$ for $\lambda > 0$
3. $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$
4. $\Gamma(n) = (n-1)!$ n $\in \mathbb{Z}$
5. $\Gamma\left(\frac{1}{2]}\right) = \sqrt{\pi}$

An application of the gamma distribution would be the time between arrivals at a supermarket check out.

Exponential Distribution

A Rv Y is said to have an exponential distribution with parameter $\beta > 0$ iff the density function of Y is :

$$f(y) = \begin{cases} \frac{1}{\beta} e^{\frac{-y}{\beta}} & \text{if } 0 \leq y \leq \infty \\ 0 & \text{Otherwise} \end{cases}$$

This distribution describes the arrival time of a randomly recurring independent event sequence.

β : The mean waiting time until the next occurrence.

Mean :$E(Y) = \beta$

Variance : $Var(Y) = \beta^2$

Decay parameter : $\frac{1}{\beta}$

Applications include the length of electronic comments, time until an earthquake occurs and more.

Memoryless Property of Exponential Distribution

If Y has an Exponential distribution then for a >0 and b > 0

$$P(x > a + b|x > a) = P(x > b)$$

Moment generating function

If Y is a continuous Rv, then the moment generating function of Y is :

$$m(t) = E(e^{ty}) = \int e^{ty} \cdot f(y)$$

Useful R commands

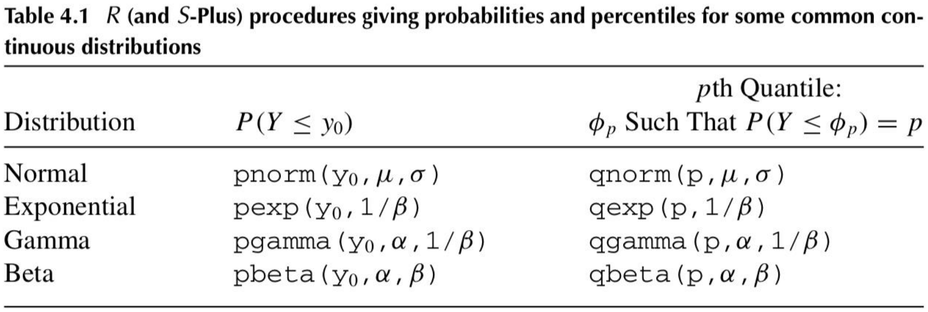


Table 4.1 *R* (and *S*-Plus) procedures giving probabilities and percentiles for some common continuous distributions

Distribution	$P(Y \leq y_0)$	pth Quantile: ϕ_p Such That $P(Y \leq \phi_p) = p$
Normal	pnorm(y_0,μ,σ)	qnorm(p,μ,σ)
Exponential	pexp(y_0,1/β)	qexp(p,1/β)
Gamma	pgamma(y_0,α,1/β)	qgamma(p,α,1/β)
Beta	pbeta(y_0,α,β)	qbeta(p,α,β)

Chapter 5: Multivariate Probability Distributions

Two Discrete Random Variables

The pmf of a single discrete Rv X specifies how much probability mass is placed on each possible x.

Similarly the joint pmf of *two* Rv's X and Y describe how much probability mass is placed on the values of (x,y).

Joint Probability Mass Function p(x,y) : Allow X and Y to be two discrete Rv's defined on a sample space. The joint pmf for each pair (x,y) is given by

$$P(x,y) = P(X = x, Y = y)$$

Where it must be the case that p(x,y) ≥ 0 and $\sum_x \sum_y p\,(x,y) = 1$.

Example

Consider having two die the first Rv X will represent the number of values on die 1. Rv Y will represent the same fr die 2. Then we have 36 outcomes in pairs of (x,y).

Graphically this could be represented in a 3D graph each with a height of 1/36.

Marginal pmf of X : The marginal pmf of X, denoted $P_x(x)$ is given by

$$P_x(x) = \sum_{y:p(x,y)>0} p\,(x,y) \text{ for all x}$$

Marginal pmf of Y : The marginal pmf of Y, denoted $P_y(y)$ is given by

$$P_y(y) = \sum_{x:p(x,y)>0} p\,(x,y) \text{ for all y}$$

Example cont.

Now using the example from before we want to find the marginal pmf of X. Thus

$$p_x(1) = p(1,1) + p(1,2) + p(1,3) + p(1,4) + p(1,5) + p(1,6) = \frac{6}{36} = \frac{1}{6}$$

From this we know that p_x(1,2,3,4,5,6) will all be the same so the marginal pmf will look like

$$P_x(x) = \begin{cases} \frac{1}{6} & \text{if x =1} \\ \frac{1}{6} & \text{if x =2,3,4,5} \\ \frac{1}{6} & \text{if x =6} \\ 0 & \text{Otherwise} \end{cases}$$

Two Continuous Random Variables

The probability that the observed value of a continuous Rv X lies in a one dimensional set A is obtained by integrating the pdf f(x) over the set A.

Similarly, the probability that the pair (x,y) of continuous rv's fall in a two dimensional set A (rectangle) is obtained by integrating the joint density function.

Let X and Y be continuous Rv's. A join probability density function f(x,y) for these two variables satisfy f(x,y)>0 and $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f\,(x,y)dxdy = 1$. Then for any two- dimensional set A

$$p[(x,y) \in A] = \int\int_A f\,(x,y)dxdy$$

where A = {(x,y)|$a \leq x \leq b\,, c \leq y \leq d$}.

See Examples in Books

Marginal Probability Density Functions

The marginal pmf of X and Y denoted $f_x(X)$ and $f_y(Y)$ are given by

$$f_x(X) = \int_{-\infty}^{\infty} f\,(x,y)dy \text{ for } -\infty \leq x \leq \infty$$

$$f_y(Y) = \int_{-\infty}^{\infty} f\,(x,y)dx \text{ for } -\infty \leq x \leq \infty$$

Note that we can extend to as much as n random variables.

Independent Random Variables

Two Rv's X and Y are said to be **Independent** if for every pair of x and y :

$p(x,y) = p_x(x) \cdot p_y(y)$, when X and Y are discrete.

$f(x,y) = f_x(x) \cdot f_y(y)$, when X and Y are continuous.

If these are not satisfied then X and Y are said to be **dependent**

Conditional Distributions

Let X and Y be two continuous random variables with joint pdf f(x,y) and marginal pdf $f_x(x)$. Then the *Conditional discrete probability function* of Y given X is

$$P(Y|X) = \frac{f(x,y)}{f_x(x)}$$

Multivariate Probability Distributions

Expected value of a function of random variables

Any function h(x) of a single Rv X is itself a RV.

Let X and Y be jointly distributed Rv's with pmf p(x,y) or pdf f(x,y) according to whether the variables are discrete or continuous. Then the expected value of a function h(x,y), denoted E[h(x,y)] or $\mu_{h(x,y)}$ is given by :

$E[h(x,y)] = \sum_x \sum_y h\,(x,y) \cdot p(x,y)$ if X and Y are discrete.

$E[h(x,y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h\,(x,y) \cdot p(x,y) dxdy$ if X and Y are continuous.

Properties of Expected Values

Let c be a constant and $g(Y_1, Y_2)$ be a function of the random variables Y_1 and Y_2. Then,

1. $E(c) = c$
2. $E[c \cdot g(Y_1, Y_2)] = c \cdot E[g(Y_1, Y_2)]$
3. $E[g_1(Y_1, Y_2) + g_2(Y_1, Y_2)] = E[g_1(Y_1, Y_2)] + E[g_2(Y_1, Y_2)]$

note : $E[Y_1 Y_2] \neq E[Y_1]E[Y_2]$

Covariance

When two random variables X and Y are not independent, it's frequently common to asses how strongly they are related to one another.

The **Covariance** between two random variables can is given by

$$Cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X \mu_Y \Rightarrow$$

$\Rightarrow \sum_x \sum_y (X - \mu_X)\,(Y - \mu_Y) \cdot p(x,y)$ if X and Y are discrete.

$\Rightarrow \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (X - \mu_X)\,(Y - \mu_Y)(x,y) dxdy$ if X and Y are continuous.

The covariance of an random variables X and Y will be a number between [-1,1] and the closer to 1 the higher the *positive* relation they have and the lower to -1 then lower the *negative* relation

note : Cov(X,X) = Var(X)

Expected Value and Variance of Linear Functions of Rv's

Given a collection of n random variables $Y_1, Y_2, \dots, Y_n$ and n numerical constraints $a_1, a_2, \dots, a_n$, the Rv

$$U_1 = a_1 Y_1 + a_2 Y_2 + \dots + a_n Y_n = \sum_{i=1}^{n} a_i \, Y_i$$

is called a linear combination of the $Y_i's$.

Properties

Allow $Y_1 Y_2, \dots, Y_n$ and $X_1, X_2, \dots, X_m$ to be random variables with $E(Y_i) = \mu_i$ and $E(X_j) = \mu_j$. Define U_1 from before and $U_2 = \sum_{j=1}^{m} b_j \, X_j$. Then ,

1. $E(U_i) = \sum_{i=1}^{n} a_i \, \mu_i$.
2. $Var(U_1) = \sum_{i=1}^{n} a_i^2 \, V(Y_i) + 2 \sum \sum_{(i,j)} a_i \, a_j Cov(Y_1, Y_2)$, where the double sum is over all pairs (i,j) for i>j.
3. $Cov(U_1, U_2) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i \, b_j Cov(Y_i, X_j)$.

Chapter 6 : Function of Random Variables

Chapter 6 is a mathematically dense chapter; this is a summary of the methods.

(1) Method of Distribution Functions

Let U be a function of random variables $Y_1, Y_2, \ldots, Y_n$. The first out of 3 methods of finding the probability distribution of U is called the method of distribution function

Procedure

1. Find the region U = u in the $Y_1, Y_2, \ldots, Y_n$ space
2. Find the region $U \leq u$
3. Find $F_U(u) = P(U \leq u)$ by integrating f($y_1, y_2, \ldots, y_n$) over the region $U \leq u$
4. Find the density function $f_U(u)$ by differentiating $F_U(u)$.

(2) Method of Transformations

The second method is called the method of transformations. Allow U = h(y), where h(y) is either an increasing or decreasing function of y for all y such that $f_Y(y) > 0$.

Procedure

1. Find the inverse function $y = f^{-1}(u)$
2. Take the derivative of y with respect to u $\frac{d}{du}[h^{-1}(u)]$.
3. Find $f_U(u) = f_Y[h^{-1}(u)]\left|\frac{d}{du}[h^{-1}(u)]\right|$

(3) Method of Moment Generating Functions

Let U be defined as before.

Procedure

1. Find the moment-generating function for U, $m_U(t)$
2. Compare $m_U(t)$ with other well known moment generating functions . If $m_U(t) = m_V(t)$ for all values of t then U and V have identical distributions.

See book for **Order Statistics chapter** and examples of these methods.

Chapter 7: Sampling Distributions and the Central Limit Theorem

Definitions

Statistic : is a function of the observable random variables in a sample and known constants.

Each observation is a random variable where we denote these samples by $X_1, X_2, \ldots, X_n$.

Before obtaining data, there is uncertainty about each x_i values and the values we calculate using them.

Sampling Distribution : Also known as the probability distribution of a statistic, it describes how the statistic varies in value across all samples that could be selected.

Simple Random Sample (IID) : The Rv's $X_1, X_2, \ldots, X_n$ form an IID of size n if each x_i is independent and has the same probability distribution.

For a IID with Rv's $X_1, X_2, \ldots, X_n$ and mean μ and standard deviation σ then:

Sample Mean : $E(\bar{x}) = \mu$

Sample Variance : $V(\bar{x}) = \frac{\sigma^2}{n}$

Standard error of the mean refers to the standard deviation of the sample and it describes the magnitude of a a typical deviation of the sample mean from the population mean.

Normal population Distribution

Allow $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean μ and SD σ. Then for any n, $\bar{X}$ is a normal distributed with mean μ and SD $\frac{\sigma}{\sqrt{n}}$.

Here $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

Example

The distribution of egg weights in grams is normal with a mean of 53 and SD of .3. Consider randomly selecting 4 of these eggs. What is the probability that the sample mean exceeds 53.5.

Solution :

$\bar{X} \sim N\left(53, \frac{0.3}{\sqrt{4}}\right)$

$$P(\bar{X} > 53.5) = P\left(Z > \frac{53.5 - 53}{.15}\right) = P(Z > 3.33)$$

`1-pnorm(3.33)` = 0.0004

Central Limit Theorem

Allow $X_1, X_2, \ldots, X_n$ to be a random sample from any distribution with mean μ and SD σ. Then if n is sufficiently large, $\bar{X}$ is approximately normally distributed with mean μ and SD $\frac{\sigma}{\sqrt{n}}$.

Rule of thumb : CLT can generally be used given n>30.

Example

Using the previous example, how many observations should be included in a sample if we wish to have the sample mean to be within 0.03 grams of μ with 95% probability.

Solution :

Want :

$$P(\mu - 0.03 < \bar{x} < \mu + 0.03) = 0.95 \Rightarrow P\left(\frac{\mu - 0.03 - \mu}{\frac{0.3}{\sqrt{n}}} < \bar{x} < \frac{\mu + 0.03 - \mu}{\frac{0.3}{\sqrt{n}}}\right) = 0.95$$

$$\Rightarrow P\left(-0.1\sqrt{n} < Z < 0.1\sqrt{n}\right) = 0.95$$

95 % confidence interval means $\alpha = 1.96$ thus $0.1\sqrt{n} = 1.96 \Rightarrow n = 385$

The Normal Approx. to the Binomial Distribution

If Y is a Binomial Rv with parameters n and p, and if n is sufficiently large then Y has approximately normal distribution with $\mu = n \cdot p$ and $\sigma = \sqrt{n \cdot p \cdot (1 - p)}$

$Bin(n, p) \sim N(\mu, \sigma)$

Rule of thumb: np > 10 and n(1-p) > 10

Example

Candidate A believes that she can win a city election if she can earn at least 55% of the votes in her precinct. She also believes that about 50% of the city voters are in favor of her. If n = 100 voters show up in her precinct what is the probability that candidate A will receive at least 55% of their votes.

Solution :

Y : number of voters that visit the precinct.

Here we can assume its approximately binomial given the conditions with n = 100

and p =.5

np = 50 = n(1-p) thus we can say that it is approximently normal. note that $\sigma = 5$

thus $P(Y \geq 55) = P\left(Z \geq \frac{55-50}{5}\right) = 1 - P(Z \leq 1) = 1 - 0.8413 = 0.1587$

NOTE THIS IS APPROXIMATELY CORRECT : we need to use continuity correction to get the exact value.

Continuity Correction : This is needed since binomial is a discrete distribution and Normal is a continuous distribution.

$$P(Y \leq x) \Rightarrow P(Y \leq x + 0.5)$$

$$P(Y \geq x) \Rightarrow P(Y \leq x - 0.5)$$

Example cont.

Using the continuity correction we have the following

$$P(Y \geq 55) \approx P(Y \geq 54.5) \Rightarrow P(Z \geq 0.9) = .1841$$

using r : `1- pbinom(54.5,100,.5)` , note we still used continuity correction

Chi, t and F distribution

Probability dist. for a function of the statistic S^2

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample of size n from a normal distribution with mean μ and SD σ. Then,

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

has a Ch-Squared (χ^2) distribution with (n-1) degrees of freedom here

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

t - distribution

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample of size n from a normal distribution with mean μ and SD σ.

Let $Z \frac{\bar{Y}-\mu}{\frac{\sigma}{\sqrt{n}}}$ be a standard normal random variable and W = $\frac{(n-1)S^2}{\sigma^2}$ be a Ch-Squared distribution with (n-1) degrees of freedom. Then, if Z and W are independent,

$$T = \frac{Z}{\sqrt{W/v}}$$

is said to have a t distribution with v degrees of freedom.

F - distribution

Let W_1 and W_2 be independent χ^2 distributed random variables with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom respectively. Then,

$$F = \frac{W_1/v_1}{W_2/v_2}$$

is said to have an F distribution with v_1 numerator degrees of freedom and v_2 numerator degrees of freedom.

Chapter 8: Estimation

The purpose of statistics is to use the information contained in a sample to make inferences about the population.

Definitions

Parameters : numerical descriptive measures that characterize a population.

Point Estimate : A single value or point is given as the estimate of μ.

Interval Estimate : The estimate of μ is given in the form of an interval.

Estimator : A rule, often expressed as a formula that tells how to calculate the value of an estimate based on the measurements in a sample.

An actual estimation is made by using an Estimator for the target parameter.

Example

Sample Mean is one possible point estimator of the population mean.

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

To see how we choose a estimator and more see the book or notes.

Bias, MSE and Error of Estimation

Unbiased Estimator : Let $\hat{\theta}$ be a point estimator for a parameter θ. Then $\hat{\theta}$ is an unbiased estimator if $E(\hat{\theta}) = \theta$. If $E(\hat{\theta}) \neq \theta$ then it is said to be biased given by $B(\hat{\theta}) = E(\hat{\theta}) - \theta$.

When choosing estimators we desire ones with smaller variance.

Mean Square Error (MSE) of a point estimator :

$$MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right] = Var(\hat{\theta}) + [B(\hat{\theta})]^2$$

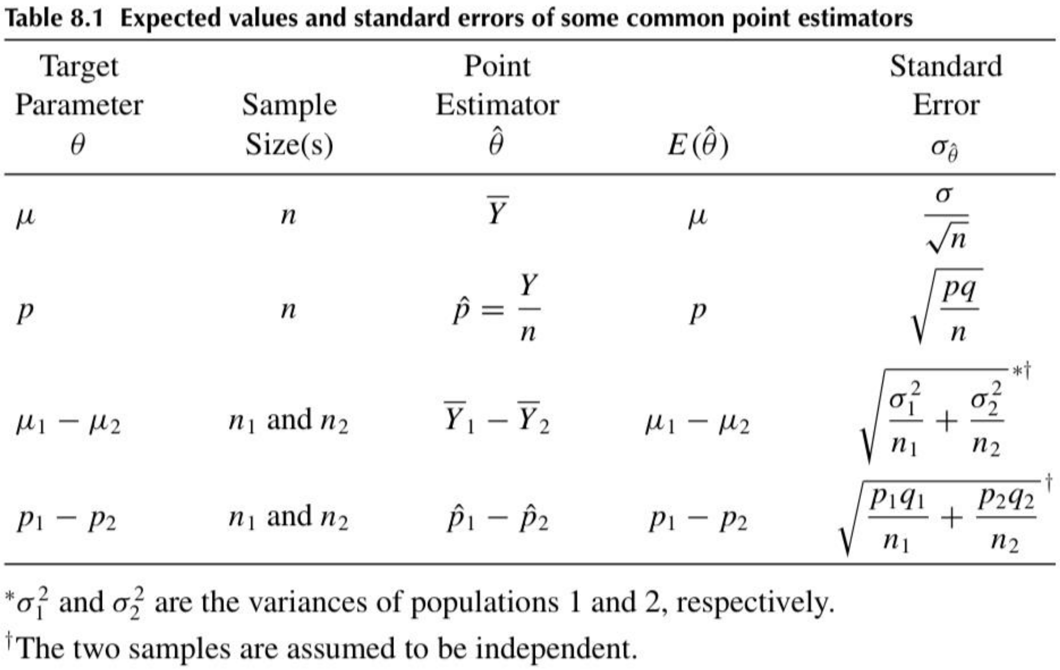


Table 8.1 Expected values and standard errors of some common point estimators

Target Parameter θ	Sample Size(s)	Point Estimator $\hat{\theta}$	$E(\hat{\theta})$	Standard Error $\sigma_{\hat{\theta}}$
μ	n	$\overline{Y}$	μ	$\frac{\sigma}{\sqrt{n}}$
p	n	$\hat{p} = \frac{Y}{n}$	p	$\sqrt{\frac{pq}{n}}$
$\mu_1 - \mu_2$	n_1 and n_2	$\overline{Y}_1 - \overline{Y}_2$	$\mu_1 - \mu_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ *†
$p_1 - p_2$	n_1 and n_2	$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$	$\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$ †

*σ_1^2 and σ_2^2 are the variances of populations 1 and 2, respectively.

†The two samples are assumed to be independent.

Error of Estimation : The distance between an estimator and its target parameter.

$$\epsilon = |\hat{\theta} - \theta|$$

Suppose we have $P(\epsilon < b)$ for some b that is picked to be a probabilistic bound. We know that $P(\epsilon < b) = P(\theta - b < \hat{\theta} < \theta + b)$ by expanding. Now we wish to find $P(\epsilon < b) = .90$ then we use $\int_{\theta-b}^{\theta+b} f(\hat{\theta}) d\hat{\theta} = .90$.

Here if we allow $k \geq 1$ and $b = k\sigma_{\hat{\theta}}$ then by tchebysheffs theorem we know $\epsilon < k\sigma_{\hat{\theta}}$ with probability 1 - $1/k^2$. A common one will be k = 2 and will often ask for a 2-standard error bound on the error of estimation.

Example

A sample of n = 1000 voters, randomly selected from a city, showed y = 560 in favor of candidate Jones. Estimate p, the fraction of voters in the population favoring Jones, and place a 2-standard-error bound on the error of estimation.

Solution

Since our target parameter is p then the estimator we will use is $\hat{p} = \frac{Y}{n}$ and Standard error is $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$. We want a 2-standard error bound on the error of estimation thus $b = 2\sigma_{\hat{p}}$

$$\hat{p} = \frac{560}{1000} = 0.56$$

$$b = 2\sigma_{\hat{p}} = 2\sqrt{\frac{(0.56)(0.44)}{1000}} \approx .03$$

The significance is that we can say that the probability that error of estimation is less than .03 is approximately .95. Also we can be reasonably confident that our estimate .56 is within .03 of the true value of p, the proportion that favors Jones.

Interval Estimation

Interval Estimators : is a rule specifying the method for using the sample measurements to calculate two numbers that form an endpoints of an interval. Also called Confidence intervals

Confidence Coefficient : Suppose that $\hat{\theta}_L$ and $\hat{\theta}_U$ are random lower and upper confidence limits for a parameter θ. Then if

$$P\left(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U\right) = 1 - \alpha$$

then the probability $(1-\alpha)$ is the Confidence Coefficient. The resulting interval $\left[\hat{\theta}_L, \hat{\theta}_U\right]$ is called the *Two-sided Confidence Interval*

One sided Confidence Intervals : It is also possible to form a one sided CI where

$$P\left(\hat{\theta}_L \leq \theta\right) = 1 - \alpha \text{ or } P\left(\hat{\theta}_U \geq \theta\right) = 1 - \alpha$$

Where the CI's are $[\hat{\theta}_L, \infty)$ or $(\infty, \hat{\theta}_U]$ respectively.

Distribution of Estimator : If the target parameter θ is $\mu, p, \mu_1 - \mu_2$ or $p_1 - p_2$ then for large samples

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$$

has approximately a standard normal distribution. Also, the confidence interval is of the form

$$\hat{\theta} \pm Z_{\alpha/2}\sigma_{\hat{\theta}}$$

Common Critical Values for two tails

Confidence_level	alpha	Crit_Value
90%	0.10	1.645
95%	0.05	1.960
99%	0.01	2.575

Example

Two brands of refrigerators, denoted A and B, are each guaranteed for 1 year. In a random sample of 50 refrigerators of brand A, 12 were observed to fail before the guarantee period ended. An independent random sample of 60 brand B refrigerators also revealed 12 failures during the guarantee period. Estimate the true difference (p1−p2) between proportions of failures during the guarantee period, with confidence coefficient approximately .98.

Solution

Our target parameter is $p_1 - p_2$ and we have a confidence interval of 0.98. Then our confidence interval will take the form of

$$(p_1 - p_2) \pm Z_{\alpha/2}\sqrt{\frac{p_1 q_2}{n_1} + \frac{p_2 q_2}{n_2}}$$

We have the following : $\hat{p}_1 = 0.24, n_1 = 50, \hat{p}_2 = 0.20, n_1 = 60, Z_{\alpha/2} = 2.33$

Plugging in and computing gives us: $0.04 \pm 0.1851 \Rightarrow (-0.1451, 0.2251)$

Example

An experimenter wishes to compare the effectiveness of two methods of training industrial employees to perform an assembly operation. The selected employees are to be divided into two groups of equal size, the first receiving training method 1 and the second receiving training method 2. After training, each employee will perform the assembly operation, and the length of assembly time will be recorded. The experimenter expects the measurements for both groups to have a range of approximately 8 minutes. If the estimate of the difference in mean assembly times is to be correct to within 1 minute with probability .95, how many workers must be included in each training group?

Solution

Confidence interval is 95% within 1 minute and thus our z value is Z = 1.96. Also, we are working with a difference of means with $n_1 = n_2$ and $\sigma_1^2 = \sigma_2^2$. Thus we have

$$1.96\sqrt{\frac{\sigma^2}{n}+\frac{\sigma^2}{n}} = 1$$

Note that $4\sigma = Range$ thus since ours is 8 we have $\sigma \approx 2$.

Small Sample CI

Small Sample Confidence Interval for Mean and Difference of means

Small-Sample Confidence Interval for the Mean : Here we have a quantity

$$T = \frac{\bar{Y}-\mu}{S/\sqrt{n}}$$

Where T has a t distribution with n-1 degrees of freedom.

Note : the t distribution has a density function very much like the standard normal density except the tails are thicker.

Confidence interval for the Mean : Has the form

$$\bar{y} \pm t_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right)$$

Small-Sample Confidence Interval for the difference of Means : Has the form

$$\bar{y}_1 - \bar{y}_2 \pm t_{\alpha/2}\cdot S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

where $t_{\alpha/2}$ is determined from the t -distribution with $n_1 + n_2 - 2$ degrees of freedom. And

Pooled Estimator :

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

Example

see book.

Confidence Intervals for Variance

see book.

Chapter 9: Properties of Point Estimators and Methods of Estimation

Estimators

Relative Efficiency : Given two unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ of a parameter θ, with variances $V(\hat{\theta}_1)$ and $V(\hat{\theta}_2)$, respectively, then the *efficiency* of $\hat{\theta}_1$ relative to $\hat{\theta}_2$, denoted, $eff(\hat{\theta}_1, \hat{\theta}_2)$, is defined by the ratio :

$$eff(\hat{\theta}_1, \hat{\theta}_2) = \frac{V(\hat{\theta}_2)}{V(\hat{\theta}_1)}$$

$eff(\hat{\theta}_1, \hat{\theta}_2) > 1$ iff $V(\hat{\theta}_2) > V(\hat{\theta}_1)$ where $\hat{\theta}_1$ is a better unbiased estimator. The opposite is also true.

Consistent Estimators : The estimator $\hat{\theta}_n$ is said to be a *Consistent Estimator* of θ, if for any positive ϵ :

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| \leq \epsilon) = 1 \text{ or } \lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \epsilon) = 0$$

Theorem : An *Unbiased Estimator* $\hat{\theta}_n$ for θ is a consistent estimator of θ if

$$\lim_{n\to\infty} V(\hat{\theta}_n) = 0$$

Theorem : Suppose that $\hat{\theta}_n$ converges to probability θ and that $\hat{\theta}_n'$ converges to θ'. Then

a. $\lim_{n\to\infty}(\hat{\theta}_n + \hat{\theta}_n') = \theta + \theta'$.

b. $\lim_{n\to\infty}(\hat{\theta}_n \cdot \hat{\theta}_n') = \theta \cdot \theta'$.

c. $\lim_{n\to\infty}\left(\frac{\hat{\theta}_n}{\hat{\theta}_n'}\right) = \frac{\theta}{\theta'}$ if $\theta' \neq 0$.

d. $\lim_{n\to\infty} g(\hat{\theta}_n) = g(\theta)$.

Sufficient Statistic : Summarizes all the information in a sample about a chosen parameter.

Let $Y_1, Y_2, \dots, Y_n$ denote a random sample from a probability distribution with unknown parameters θ. Then the statistic $U = g(Y_1, Y_2, \dots, Y_n)$ is said to be *Sufficient* for θ if the conditional distribution of $Y_1, Y_2, \dots, Y_n$, given U does not depend on θ.

Theorem : Let U be a statistic based on the random sample $Y_1, Y_2, \dots, Y_n$. Then, U is a *Sufficient Statistic* for the estimation of a parameter θ iff the likelihood function $L(\theta) = L(y_1, y_2, \dots, y_n|\theta)$ can be factored into non-negative functions,

$$L(Y_1, Y_2, \dots, Y_n|\theta) = g(u, \theta) \times h(y_1, y_2, \dots, y_n)$$

where (u, θ) is a function only of u and θ and $h(y_1, y_2, \dots, y_n)$ is not a function of θ.

Example

see book and notes.

Rao-Blackwell Theorem : Let $\hat{\theta}$ be an unbiased estimator for θ such that $V(\hat{\theta}) < \infty$. If U is a sufficient statistics for θ, define $\hat{\theta}^* = E(\hat{\theta}|U)$. Then, for all θ,

$$E(\hat{\theta}^*) = 0 \text{ and } V(\hat{\theta}^*) \leq V(\hat{\theta})$$

Is also called *MVUE* : minimum value unbiased estimator.

Example

see book and notes

Method of Moments

This is used to estimate the parameter θ. We will equate $\mu_k' = m_k'$ and solve for our parameter.

Kth moment of a RV about the origin : $\mu_k' = E(Y^k)$

Kth sample moment is the average : $m_k' = \frac{1}{n}\sum_{i=1}^{n} Y_i^k$

Method of Maximum Likelihood (MLE)

Here we will use our likelihood function then take its derivative and equate to 0 to solve for the parameter.

Example

see book and notes

Chapter 10: Hypothesis Testing

Hypothesis Testing

Statistical Hypothesis : is a *claim* or *assertion* either the value of single or several parameters or the form of an entire probability distributions.

In a hypothesis test there are two contradicting hypothesis under considerations.

Ex : $\mu = 0.75$ vs $\mu \neq 0.75$

Null Hypothesis : Denoted by H_0 it is the claim initially assumed to be true. Has the form :

$$H_0 : \mu = \mu_0$$

Alternative Hypothesis : Denoted by H_a it is the assertion that is contradictory to H_0. This is the claim we wish to validate. Will have the forms :

a. $H_a : \mu > \mu_0$

b. $H_a : \mu < \mu_0$

c. $H_a : \mu \neq \mu_0$

The two outcomes of Hypothesis testing are to either :

1. Reject H_0 given the evidence suggests that H_0 is false, or
2. Fail to reject H_0 given the evidence does not contradict H_0.

The Elements of a Statistical Test

1. Null hypothesis H_0
2. Alternative hypothesis H_a
3. Test statistic
4. Rejection region

Examples

An Investor has developed a new energy efficient lawn mower engine. He claims that the engine will run continuously for 5 hours (300 min) on a single gallon of gas.

1. You suspect the mean run-time is not 300 min :
2. You suspect the mean run-time is greater than 300 min:

State the given hypothesis's.

Solution :

1. hyp:

$$H_0: \mu = 300$$

$$H_a: \mu \neq 300$$

2. hyp:

$$H_0: \mu = 300$$

$$H_a: \mu > 300$$

Errors in Hypothesis Testing

Type 1 Error : Consists of rejecting the null hypothesis H_0 when it is true. Also, P(Type 1 Error) = α.

Type 2 Error : Consists of not rejecting the null hypothesis H_0 when it is false. Also, P(Type 2 Error) = β.

Possible outcomes with Hypothesis Testing:

1. H_0 is true and reject H_0 ⇒ Type 1 Error.
2. H_0 is true and fail to reject H_0 ⇒ Correct decision.
3. H_0 is false and reject H_0 ⇒ Correct rejection.
4. H_0 is false and fail to reject H_0 ⇒ Type 2 Error.

Example

H_0: The defendant is innocent.

H_a The defendant is guilty.

Find all possible outcomes

Solution

1. Innocent and found guilty ⇒ Type 1 Error.
2. Innocent and found innocent ⇒ Correct.
3. guilty and found guilty ⇒ Correct.
4. guilty and found innocent ⇒ Type 2 Error.

Assumptions for Hypothesis Testing for One Population Mean with Sigma Known

1. Simple random testing.
2. Normal population or sample size is at least 30.
3. The population standard deviation σ is known.

Steps for Hypothesis Testing for one population mean

1. State the null and alternative hypothesis

a. Right tailed $H_a: \mu > \mu_0$

b. Left tailed $H_a: \mu < \mu_0$

c. Two tailed $H_a: \mu \neq \mu_0$

2. State the Level of Significance $\alpha = P(\text{Type 1 Error})$
3. Calculate the Statistic (denoted z_0):

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

4. Determine the p-value (Rejection region) : Area in curve (right tail, left or two)
5. Draw a conclusion: P-value $\leq \alpha$ then reject H_0 ; otherwise do not reject.
6. Interpret the decision

Example

Preforming a HT for a population mean (right-tailed)

We claim the mean age of newlywed women is 26.5 years based on rough data. A simple random sample of 213 newlywed women found the mean to be 27. Assume the population SD is 2.3 and we desire a 5% significance level. Determine if there is sufficient evidence to support the claim.

Solution

1. We have :

$H_0: \mu = 26.5$ $H_0: \mu > 26.5$

Thus we have a right tailed test

2. n = 213, $\bar{x} = 27$, $\sigma = 2.3$, $\alpha = 0.05$
3. $Z = (27 - 26.5)/\left[2.3/\sqrt{213}\right] = 3.17$
4. P = `pnorm(3.17,lower.tail = F)` = 0.0008

5. We desire P-value $\leq \alpha \Rightarrow 0.0008 < 0.05$. Thus we reject the null hypothesis.

6. there is sufficient evidence to support the claim that the mean is greater than 26.5

Sigma Unknown Hypothesis Testing for Population Means

Testing Conditions :

1. A single random sample used
2. The population standard deviation σ is unknown.
3. Either sample size $n \geq 30$ or approximately normal dist.

When these conditions are met we use a t-distributions for testing:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \text{ where df =1}$$

Example

Given $H_0: \mu = 100, H_a: \mu > 100, \alpha = 0.05, n = 27, s = 9.07, \bar{x} = 104.93, \mu_0 = 100$ and the distribution is approximately normal. Determine if there is sufficient evidence to support the claim.

Solution

3. $t = \frac{4.93}{9.07/\sqrt{27}} = 2.824, df = 26.$

4. P-Value = `pt(2.824,26,lower.tail = F)`

5. 0.0045 < 0.05 thus reject H_0

6. There is sufficient evidence.

Hypothesis Testing for one population mean using Rejection region.

When using the Reject region we look at the Z-score to see if our desired Z score is in our RR. If so then we can claim there is sufficient evidence.

Example cont.

Using this example : We claim the mean age of newlywed women is 26.5 years based on rough data. A simple random sample of 213 newlywed women found the mean to be 27. Assume the population SD is 2.3 and we desire a 5% significance level. Determine if there is sufficient evidence to support the claim.

Solution

1. We have :

$H_0 : \mu = 26.5$ $H_0 : \mu > 26.5$

Thus we have a right tailed test

2. n = 213, $\bar{x} = 27$, $\sigma = 2.3$, $\alpha = 0.05$

3. $Z = (27 - 26.5)/[2.3/\sqrt{213}] = 3.17$

4. Since our $\alpha = 0.05$ we will have a Z score of 1.96. Since 3.17 > 1.645 we reject H_0

5. there is sufficient evidence to support the claim that the mean is greater than 26.5

Hypothesis Testing for Population Proportion

Testing Conditions :

1. A single random sample used.
2. The population is at least 20 times as large as the sample,
3. The individuals in the population are divided into two categories.
4. The values np_0 and $n(1 - p_0)$ are at least 10.

Test Statistic for Z : $Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

Example

In a recent college survey 90% of female college students used TikTok. Assume this survey was based on a random sample of 500 students. Someone from TikTok wants to report that more than 85% of female college students use TikTok. Can you conclude the proportion of female college students is greater than 85% with $\alpha = 0.05$ level of significance.

Solution

First checking the assumptions we see that we have a random sample, two categories : TikTok users and non users, The size of the population is more than 20 times the sample of 500 and np_0 = 425 > 10 and $n(1 - p_0)$ = 75 >10.

1. Our hypothesis is $H_0 : p = 0.85$ and $H_1 : p > 0.85$

2. sig level of 5% $\Rightarrow Z_\alpha = 1.645$

3. $Z = \frac{0.90 - 0.85}{\sqrt{\frac{0.85(0.15)}{500}}} = 3.13$

4. Here we can that 1.645 < 3.13 thus we can reject the null hypothesis and claim that there is sufficient evidence to claim that more than 85% of female college students use TikTok.

Hypothesis Testing for Standard Deviation

For HT testing for standard deviation we use this test statistic :

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

Critical value will be found using chi dist. with df = n-1 .

The example would work the same way as previous except we are using a different testing statistic.

Example

If we wanted to find the critical value at a 5% confidence level for study that had 16 observations and was a left tailed test. then use `qchisq(.05,15)`.

Hypothesis Testing with Two Samples

For two sample hypothesis tests:

1. The Null Hypothesis is a statistical hypothesis about the difference between two parameters.

2. The alt. Hypothesis is a statistical hypothesis which is true when h_0 is false.

Conditions for Two Sample testing

1. Samples are randomly selected
2. Samples are independent
3. $n \geq 30$ or normal dist.

$$\mu_{\bar{x}-\bar{x}} = \mu_1 - \mu_2$$

$$\sigma_{\bar{x}-\bar{x}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$Z = \frac{(\bar{x}_2 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}-\bar{x}}}$$

Note that $\mu_1 - \mu_2$ go to zero since $H_0: \mu_1 = \mu_2$

Small Sample Hypothesis Testing with Two Samples

Conditions for Two sample t - Test :

1. Samples are randomly selected
2. Samples are independent
3. $n \leq 30$ or normal dist.

$$t = \frac{(\bar{x} - \bar{x}) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}-\bar{x}}}$$

IF $\sigma_1 = \sigma_2$ then,

$$\sigma_{\bar{x}-\bar{x}} = \hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\hat{\sigma} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}, \text{ df} = n_1 + n_2 - 2$$

IF $\sigma_1 \neq \sigma_2$ then,

$$\sigma_{\bar{x}-\bar{x}} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, \text{ df} = min(n_1 - 1, n_2 - 1)$$

Hypothesis Testing for Two Proportions

Conditions for Testing:

1. Two simple random sample used that are independent.
2. Each population is at least 20 times as large as the sample.
3. The individuals in each population are divided into two categories.
4. Both samples contain at least 10 individuals in each category.

Test statistic : $Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$

where we have $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$ is the pooled proportion.

Hypothesis Testing for Two Match Paired Samples

When we are talking about paired samples we are talking about data that is not independent. For example we collect data on 10 car transmission before tune up and after. Here our data results are not independent.

Conditions for Testing:

1. Simple random sample of matched pairs
2. Either large sample size (n>30) or there is no evidence of skewness or outliers.

Definitions

$\bar{d}$ is the sample mean of the differences between the values of matched pairs.

s_d is the sample deviation of the differences between the values of matched pairs.

μ_d is the population mean difference of the matched pairs.

Test Statistic : $t = \frac{\bar{d}-\mu_0}{s_d/\sqrt{n}}$

Note: μ_0 goes to zero typically.

Hypothesis Testing for Two population Standard deviations

Conditions for Testing:

1. We have independent random samples from two populations.
2. Both populations are normally distributed.

Test Statistic : $F = \frac{max(s_1^2,s_2^2)}{min(s_1^2,s_2^2)}$

For example assume we have two populations with sizes 10 and 6. We want to find a Critical point for a significance level of 5% then use : `qf(.05,5,9,lower.tail = F)`

Power and Probability of Type 2 Error for one tail

Power = $1-\beta$

Probability of a type 2 Error β

$\beta = P(\hat{\theta} \leq k$ when H_a is true)

Example

A vice president in charge of sales for a large corporation claims that salespeople are averaging no more than 15 sales contacts per week. (He would like to increase this figure.) As a check on his claim, n = 36 salespeople are selected at random, and the number of contacts made by each is recorded for a single randomly selected week. The mean and variance of the 36 measurements were 17 and 9, respectively. Test with level α = .05.

Suppose that the vice president wants to be able to detect a difference equal to one call in the mean number of customer calls per week. That is, he wishes to test H0 :μ = 15 against Ha :μ = 16. With the data as given in Example 10.5, find β for this test.

Solution

The rejection region for a $\alpha = 0.05$ is $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} > 1.645 \Rightarrow$

$$\bar{x} > \mu_0 + 1.645\frac{\sigma}{\sqrt{n}} \Rightarrow \bar{x} > 15 + 1.645\frac{3}{\sqrt{36}} \Rightarrow \bar{x} > 15.8225$$

Thus by definition $\beta = P(\bar{X} \leq 15.8225 \text{ when } H_a\!: \mu = 16)$

Then, $\beta = P\left(\frac{\bar{X}-\mu_a}{\sigma/\sqrt{n}} \leq \frac{15.8225-16}{3/\sqrt{36}}\right) = P(Z \leq -0.36) = 0.3594$

The Power = 0.6406

From our value of β we can see that our sample size of n = 36 will frequently fail to detect a difference of 1 unit from the hypothesized means. By increasing n we can reduce β

Sample size for Z test

$$Z_\alpha = \frac{k-\mu_0}{\sigma/\sqrt{n}}\ , Z_\beta = \frac{k-\mu_\alpha}{\sigma/\sqrt{n}}$$

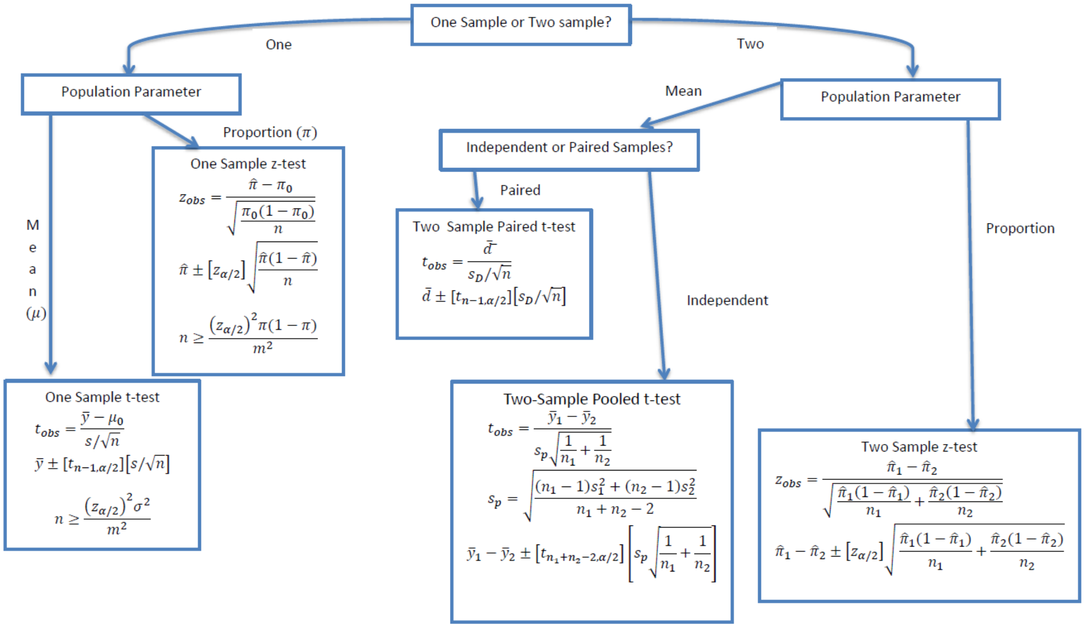
IF it is one sided:

$$n = \frac{\sigma^2\left(Z_\alpha + Z_\beta\right)^2}{(\mu_\alpha - \mu_0)^2}$$

If it is two sided:

$$n = \frac{\sigma^2\left(Z_{\alpha/2} + Z_\beta\right)^2}{(\mu_\alpha - \mu_0)^2}$$

A Nice Flow Chart


One Sample or Two sample?
One
Two
Population Parameter
Proportion (π)
Mean (μ)
One Sample z-test
$z_{obs} = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$
$\hat{\pi} \pm [z_{\alpha/2}]\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$
$n \geq \frac{(z_{\alpha/2})^2 \pi(1-\pi)}{m^2}$
One Sample t-test
$t_{obs} = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$
$\bar{y} \pm [t_{n-1,\alpha/2}][s/\sqrt{n}]$
$n \geq \frac{(z_{\alpha/2})^2 \sigma^2}{m^2}$
Population Parameter
Mean
Proportion
Independent or Paired Samples?
Paired
Independent
Two Sample Paired t-test
$t_{obs} = \frac{\bar{d}}{s_D/\sqrt{n}}$
$\bar{d} \pm [t_{n-1,\alpha/2}][s_D/\sqrt{n}]$
Two-Sample Pooled t-test
$t_{obs} = \frac{\bar{y}_1 - \bar{y}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$
$\bar{y}_1 - \bar{y}_2 \pm [t_{n_1+n_2-2,\alpha/2}]\left[s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$
Two Sample z-test
$z_{obs} = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}}$
$\hat{\pi}_1 - \hat{\pi}_2 \pm [z_{\alpha/2}]\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$

Chapter 11 : Linear Models and Estimation by Least Squares

Linear Statistical Model

A linear Statistical model relating a random response y to a set of independent variables $x_1, x_2, \ldots, x_n$ is of the form

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

The simplest case being :

$$Y = \beta_0 + \beta_1 x_1 + \epsilon$$

Where B_i, i=1,2,3,…,n are all unknown parameters. ϵ is a Rv representing randomness and the variable x_i are known. Assume $E(\epsilon) = 0$ then

$$E(Y) = \beta_0 + \beta_1 x_1$$

Method of Least Squares

The **least squares regression line** is the line that minimizes the sum of the squared error(residuals), We do this by minimizing the sum of residuals.

$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)^2\right).$

$\hat{y}_i = \hat{B}_0 + \hat{B}_1 x_i$ This line is our predicted observed value.

Residual : Observed y (y_i) - predicted y $(\hat{y}_i)$

$$\hat{\beta}_0 = \bar{y} - \bar{x}\beta_1$$

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

Example

Given the data set A find the least squares regression line. In data set A our x represents a drilling depth in feet. The y value represents the length of time it took to drill 5 feet at that depth.

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
x	35	50	75	95	120	130	145	155	160	175	185	190
y	5.88	5.99	6.54	6.37	7.07	6.93	6.78	7.57	7.88	7.62	6.99	7.9

Solution

Here we can either use our definitions above or use R to solve it.

Use the R code

```
lm.A = lm(y ~ x, data = A)
summary(lm.A)

##
## Call:
## lm(formula = y ~ x, data = A)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6567 -0.1152  0.0384  0.1860  0.5255
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.484235   0.261856  20.944 1.37e-09 ***
## x           0.011689   0.001928   6.062 0.000122 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.334 on 10 degrees of freedom
## Multiple R-squared:  0.7861, Adjusted R-squared:  0.7647
## F-statistic: 36.74 on 1 and 10 DF,  p-value: 0.0001217

lm.function = function(x){
  y = 5.4842 + 0.01169*x
  print(y)
}
```

From this we can see our $\beta_0 = 5.484$ and $\beta_1 = 0.0117$ and if we wish to predict an resulting y value we can use the linear model. For Example if we wish to know what the value is if x - 130 then we would use `lm.function(130)` = 7.035.

Interpreting the data:

slope : for each additional foot of depth the time to drill 5 feet increases by .0117 on average.

intercept : the time to drill 5 feet from the topsoil takes 5.5 seconds

Interpreting our slope and y-int

Interpretation of the Y Intercept : The intercept $\hat{\beta}_0$ is the expected mean value of y when x = 0. If x never equals 0 then the intercept has no intrinsic meaning. In general check if 0 is a reasonable value for x and if any values of x exist in the data set.

Interpretation of the Slope : $\hat{\beta}_1$ represents the estimated increase in Y per unit increase in x. Note: An increase may be negative when $\hat{\beta}_1$ is negative.

Warning

Do not use a least-squares regression line to make predictions outside the scope of the model. We can't be sure that a linear relationship exists outside of our sample.

Properties of the Least-Squares Estimators; Simple Linear Regression

1. The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased - that is , $E(\hat{\beta}_i) = \beta_i$, for i = 0,1.
2. V$(\hat{\beta}_0) = c_{00}\sigma^2$ where $c_{00} = \sum x_i^2/(nS_{xx})$.
3. V$(\hat{\beta}_1) = c_{11}\sigma^2$ where $c_{11} = \frac{1}{S_{xx}}$.
4. $Cov(\hat{\beta}_0, \hat{\beta}_1) = c_{01}\sigma^2$ where $c_{01} = \frac{-\bar{x}}{S_{xx}}$.
5. An unbiased estimator of σ^2 is $S^2 = SSE/(n-2)$ where $SSE = S_{yy} - \hat{\beta}_1 S_{xy}$ and $S_{yy} = \sum(y_i - \bar{y})^2$.

If, in addition the ϵ_i for i = 1,2,…,n are normally distributed,

6. Both $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed.
7. The random variable $\frac{(n-2)S^2}{\sigma^2}$ has a χ^2 distribution with n-2 df.
8. The statistic S^2 is independent of both $\hat{\beta}_0$ and $\hat{\beta}_1$.

Scatter plots and Linear Correlation Coefficient

Pearson's linear correlation coefficient : This is a measure of the strength and direction of the linear relation between two quantitative variables.

$$r = \frac{\sum\left(\frac{x_i = \bar{x}}{s_x}\right)\left(\frac{y_i = \bar{y}}{s_y}\right)}{n-1}$$

$\bar{x}$: Sample mean of the explanatory variable

S_x: Standard deviation of the explanatory variable

$\bar{y}$: Sample mean of the response variable

S_y: Standard deviation of response variable

n : number of individuals

Properties of the linear correlation coefficient

1. $r \in [-1,1]$
2. The closer r is to -1 and 1 the stronger the relationship is to it.
3. The closer r is to 0 there is little to no evidence of a linear relationship.
4. something about unit-less measure

Warning

Correlation is not causation

Example

Using the previous data set we can use Pearson's test on R and compute r

```
cor.test(A$x,A$y, method = "pearson")

## 
##  Pearson's product-moment correlation
## 
## data:  A$x and A$y
## t = 6.0618, df = 10, p-value = 0.0001217
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6366490 0.9679775
## sample estimates:
##       cor 
## 0.8866074

cor(A$x,A$y)

## [1] 0.8866074
```

Here, we can see a high correlation in our results.

The Coefficient of Determination

The **Coefficient of Determination** R^2 measures the proportion of total variation in the *response variable* that is explained by the least squares regression line.

Properties :

1. The Coefficient of Determination is between [0,1].
2. If $R^2 = 0$ then the line has no explanatory value.
3. If $R^2 = 1$ then all of the variation is explained in the response variable.

Example

Using the example before and using the our linear regression summary we see that $R^2 = .60$

Total deviation : $(y - \bar{y})$ Difference between observed value of the response variable and the mean value of the response variable.

Explained deviation : $(\hat{y} - \bar{y})$ Difference between the predicted value of the response variable and the mean value of the response variable.

Unexplained deviation : $(y - \hat{y})$ Difference between the observed value of the response variable of y and the predicted value of the response variable.

Total variation : $\sum(y - \bar{y})^2 = \sum(y - \hat{y})^2 + \sum(\hat{y} - \bar{y})^2$

Confidence and Prediction Intervals

Confidence Interval for $E(Y) = \beta_0 + \beta_1 x^*$

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

where $t_{\alpha/2}$ is based on n-2 df

$$S = \sqrt{\frac{\sum(y - \hat{y})^2}{n - 2}}$$

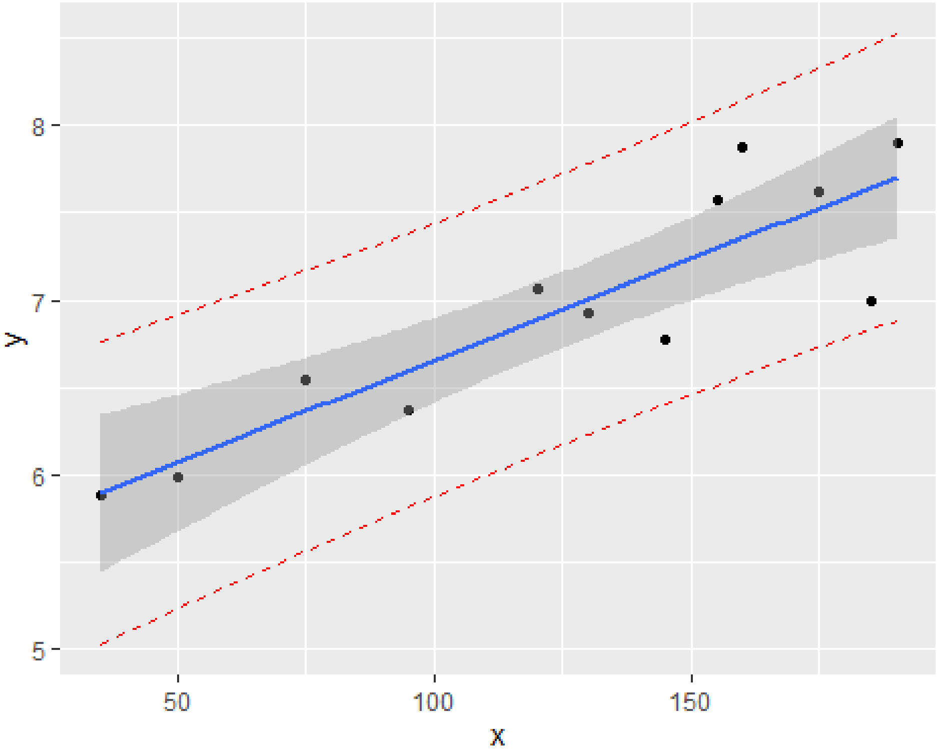
Prediction Interval for Y when $x = x^*$

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

Example

Make a 95% confidence interval and a 95% prediction interval using Data set A.

```
library(ggplot2)

# Here we are making our prediction interval
prediction = predict(lm.A, interval = "predict",level = .95)

# Combining the data
data.A = cbind(A,prediction)

# Plotting everything
ggplot(data.A,aes(x,y))+
  geom_point() +stat_smooth(method = lm) +
  geom_line(aes(y = lwr),col = "red",linetype = "dashed")+
  geom_line(aes(y = upr),col = "red",linetype = "dashed")
```



Fitting the Linear Model by Using Matrices

Here we will be working with *multiple linear regression models*

Suppose we have the linear model $Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon$ and we make n independent observations, $y_1, y_2, \ldots, y_n$ on Y. We can write the observation y_i as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \beta_k x_{ik} + \epsilon_i$$

For a simple linear model we of the form $Y = \beta_0 + \beta_1 x + \epsilon$, we have

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_2 \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n \end{bmatrix}, \widehat{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Least Squares Equations and solutions for a General Linear Model

Solution : $\boldsymbol{\beta} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y}$

$SSE \;= \mathbf{Y}^{\mathbf{T}}\mathbf{Y} - \widehat{\boldsymbol{\beta}}\mathbf{X}^{\mathbf{T}}\mathbf{Y}$

Example

Consider the data set B below with a linear fitted line of y = 1 + .7x

Here is an overview of this example

x	y
-2	0
-1	0
0	1
1	1
2	3

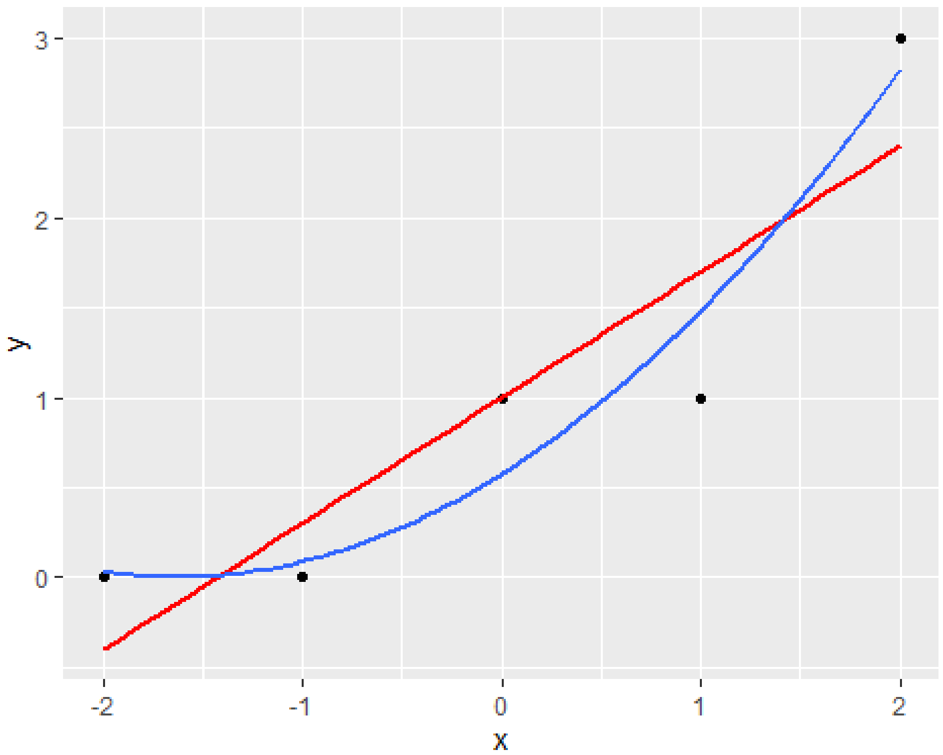
```
# To find the SSE

SSE = sum((B$y -predict(lm.b))^2)
SSE

## [1] 1.1
```

```
# To find the Variance

Variance = SSE/(length(B$x) -2)
Variance

## [1] 0.3666667

# To find the variance of the estimators

Var_B.0 = vcov(lm.b)[[1]]
Var_B.0

## [1] 0.07333333

Var_B.1 = vcov(lm.b)[[4]]
Var_B.1

## [1] 0.03666667

# To check covariance look at entries 2 and 3.

# To check Coefficient of determination

COD = 1- (SSE/sum((B$y - mean(B$y))^2))
COD

## [1] 0.8166667
```

Now lets see how our data fits on a non linear curve

```
# First lets do it how the book does it

x.0 = matrix(c(1,1,1,1,1))
x.1 = matrix(B$x)
x.2 = matrix(B$x^2)
X = cbind(x.0,x.1,x.2)
Y = matrix(B$y)

Beta = solve(t(X)%*%X,link = 'inverse')%*%t(X)%*%Y
Beta

##           [,1]
## [1,] 0.5714286
## [2,] 0.7000000
## [3,] 0.2142857

# or using r
```

```
new.B = cbind(B,x.2)

quad.model = lm(y ~ x + x.2, data = new.B)

summary(quad.model)$coefficients[1:3]

## [1] 0.5714286 0.7000000 0.2142857

# Here we have our COD

summary(quad.model)$r.squared

## [1] 0.9238095

# Lets make a plot with both fits

ggplot(new.B, aes(x,y)) +
  geom_point() +
 stat_smooth(method = lm, color = "red",level = 0) +
  stat_smooth(method = lm, formula = y ~ x + I(x^2), size = 1,level = 0)
```



As we can see a Quadratic model has a better fit than the linear fit. We can see this numerically by looking at the Coefficient of determination on each fitted model.

Properties of Least-Squares Estimators : Multiple Linear Regression

1. $E(\widehat{\beta_\iota}) = \beta_i$, i = 0,1,…,k.
2. $V(\widehat{\beta_\iota}) = c_{ii}\sigma^2$, where c_{ii} is the element in row i and column i of $(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}$
3. $Cov(\widehat{\beta_\iota}, \widehat{\beta_J}) = c_{ij}\sigma^2$
4. An unbiased estimator of σ^2 is $S^2 = SSE/[n-(k+1)]$, where $SSE = \mathbf{Y}^\mathbf{T}\mathbf{Y} - \widehat{\boldsymbol{\beta}}\mathbf{X}^\mathbf{T}\mathbf{Y}$

If, in addition the ϵ_i for i = 1,2,…,n are normally distributed,

5. Each $\widehat{\beta_\iota}$ is normally distributed.
6. The RV $\frac{[n-(k+1)]S^2}{\sigma^2}$ has χ^2 distribution with n - (k+1) df.
7. The statistic S^2 and $\widehat{\beta_\iota}$ are independent.

Notes

Be wary of outliers, high-influence points, lack of fit and multicollinearity in models.

Chapters 12 : Tests with Qualitative Data

Tests for qualitative or categorical data use the χ^2 distribution.

Preforming a Goodness of fit test

1. State null and alt. hypothesis. Here the null hypothesis specifies a probability for each category. The alt. says that some or all differ.

2. Compute the expected frequencies, make sure each has 5 or more.

3. Find a χ^2 score related to your desired level of significance.

4. Compute the Test statistic:

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

where O is each expected frequency and E is the expected value of that category.

5. Decided to reject or not and state conclusion

Example

Assume we have a fair die and we want to test to see if it is unfair or not. We roll the die 60 times and record our results in the data set A along with our expected values. Test to see if the dice is unfair.

Solution

1. We have $H_0: p_1 = p_2 = \ldots = p_6 = 1/6$ and H_1:some or all differ from 1/6

2. Our expected frequency is 10 per each p_i

3. Our significance level is 5% thus χ^2_α = `qchisq(.95,df = 5)` = 11.07 . Note : that our df = 5 and not 59 since it is by category.

4. Test statistic : χ^2 = `sum((O-E)^2/E)` = 9.4 . Note : Use vectors for easy math.

5. 9.4 < 11.07 thus we do not reject the null hypothesis. Not enough evidence to conclude the data is unfair.

Note : Our expected value is not always the same so we might have to compute multiple expected probabilities in our calculations.

Tests for Independence and Homogeneity

Please note that a test for Independence is identical for a test of homogeneity.

Def

Independence : One set does not affect the probability of the other.

Homogeneity : A data set is homogeneous if it is made up of things that are similar to each other.

Test of Independence

1. The null hypothesis states that the row and column variables are independent. The alt. states they are not independent

2. Compute the row and column totals and calculate a table of expected frequency : $E = \frac{\text{Row total} \cdot \text{Column total}}{\text{Grand total}}$

3.Find a χ^2 score related to your desired level of significance and use (r-1)(c-1) degrees of freedom.

4.Compute the Test statistic:

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

5. Decide a result and state your conclusion.

Example

We have a Data table B below and we wish to know if our data is independent. Our data involves college majors and hours studying per week. Use a significance level of 1%.

```
##              Humanities Social Science Business Engineering Row Total
## 0-10                 68            106      131          40       345
## 11-20               119            103      127          81       430
## 20+                  70             52       51          52       225
## Column Total        257            261      309         173      1000
```

Solution

1. We have H_0: Major and hours studying are independent, H_1: Major and hours studying are not independent.

2. To find the expected frequency we have to use our formula above. We will denote this table B.ex

```
##       Humanities Social Science Business Engineering
## 0-10      88.665         90.045  106.605      59.685
## 11-20    110.510        112.230  132.870      74.390
## 20+       57.825         58.725   69.525      38.925
```

3. Our $\chi_\alpha^2 =$ `qchisq(.99,6)` = 16.812

4. Our test statistic is

```
sum((B[1:3,1:4]-B.ex)^2/B.ex)

## [1] 34.63775
```

5. Thus, 16.812 < 34.638 so we reject our null hypothesis and we can state that numbers of hours studying and major is indeed not independent. The number of hours a student studies varies among majors.

What would this example say about homogeneity?

First our Null hypothesis would be that H_0: the hours spent studying for any major are the same vs H_1: The hours spent studying differ. Since we already did our test we can conclude that the numbers of hours studied varies across majors.

Note : Homogeneity of Variance is called *Homoscedasticity* and is used to describe a set of data that has the same variance. More on this later.

Chapters 13 : Inference in Linear Models

This chapter overlaps a lot of Chapter 11 from 115 so in here I will only include things not included in that document.

Residual plots

A residual plot is a where the residuals $(y - \hat{y})$ are plotted against the explanatory variable x.

Conditions for Residual Plots :

1. The residual plot must not exhibit any obvious patterns
2. The vertical spread of residuals must remain roughly within some bounds and neither increase nor decrease
3. There must not be any outliers

If a plot fails any of these conditions then we can say that the linear model is not valid.

Adjusted Coefficient of Determination

The C.O.D R^2 has a disadvantage as a measure of goodness of fit, that is if more explanatory variables are added to the model then the value of R^2 will not decrease and will increase even if the explanatory variables added are unrelated to the outcome variable. To compensate for this we use the adjusted R^2.

Adjusted C.O.D$R^2 = R^2 - \frac{p}{n-p-1}(1 - R^2)$

where n is the number of observations and p is the number of explanatory variables.

F-test goodness of fit

Use `summary(lm)` and look at the p value given after the F-score. Honestly more info is need and you can probably google it.

Chapter 14 : Analysis of Variance

Suppose we have the situation where we have n > 2 populations, and we wish to compare their means. To do so, if the assumptions are satisfied, we will do this will a method called Analysis of Variance (ie : ANOVA).

Definitions :

Factor : The explanatory variable which is qualitative.

Treatments : The different values for a factor. (ie: other sample variable)

One-Factor Experiment : Experiment with only one factor.

One-Way Table : Table containing the treatments, sample values, sample mean and sample standard deviation.

One-Way ANOVA : method of determining whether the population means differ.

Assumptions for One-Way ANOVA

1. We have independent simple random samples from 3 or more populations
2. Each population is approximately normal
3. The populations must all have the same variance, which we will denote by σ^2

In general we want to use a *Balanced design* in our data. To check that we do have a balanced design we must check our 3rd assumption with the condition :

$$\frac{max(\sigma^2)}{min(\sigma^2)} < 2$$

If it holds then we can assume the 3rd condition is satisfied.

Format for One-Way ANOVA Hypothesis Testing.

We have the format: $H_0\colon \mu_1 = \ldots = \mu_n$ and H_1: two or more of the μ_i are different.

Definitions :

Treatment Sum of Squares :(ie : SSTr) is a quantity used to measure the spread of the sample means around the sample grand mean.

$$SSTr = n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + \ldots + n_k(\bar{x}_k - \bar{\bar{x}})^2$$

Error Sum of Squares :(ie : SSE)

$$SSE = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots + (n_k - 1)s_k^2$$

Error Mean Square : (ie : MSE)

$$MSE = \frac{SSE}{N - k}$$

Where N = $(n_1 + \ldots + n_k)$

Treatment Mean Square : (ie : MSTr)

$$MSTr = \frac{SSTr}{k - 1}$$

F - Test for One-Way ANOVA

To test : $H_0 : \mu_1 = \ldots = \mu_n$ and H_1: two or more of the μ_i are different.

1. Check our assumptions are correct.
2. Compute SSTr and SSE.
3. Compute MSTr and SSE.
4. Compute the test statistic: F = $\frac{MSTr}{MSE}$
5. Find the critical value with degrees of freedom k - 1 and N-k
6. Decide whether to reject and state conclusion

Example

Given the tabular data set C test the null hypothesis that all the means are equal with a significance level of 5%.

```
##   Flux       Sample_Values Sample_Mean Sample_Standard_Deviation
## 1    A 250 264 256 260 239       253.8                     9.757
## 2    B 263 254 267 265 267       263.2                     5.404
## 3    C 257 279 269 273 277       271.0                     8.718
## 4    D 253 258 262 264 273       262.0                     7.450
```

Solution

1. Each population is approximately normal and independent. We can see from the data that if we divide the largest and smallest SD then we have a value less than 3 so our data is well balanced.
2. We will compute SSTr and SSE. Each sample has a sample size of 5 so n = 5

```
# sample size
n = 5

# Compute Grand Mean
grand.mean = mean(c(250, 264, 256, 260, 239,263, 254, 267, 265, 267, 257,
279, 269, 273, 277, 253, 258, 262, 264, 273))
grand.mean

## [1] 262.5

# Compute SSTr
SSTr = sum(n*(C$Sample_Mean - grand.mean)^2)
SSTr

## [1] 743.4

# Compute SSE

SSE = sum((n-1)*(C$Sample_Standard_Deviation)^2)
SSE

## [1] 1023.633
```

3 & 4. Now we will compute MSTr, MSE, Test statistic and Critical value.

```
# Number of samples
k = 4

# Compute grand sample size
N = k*n

# Compute MSTr
MSTr = SSTr/(k-1)
MSTr

## [1] 247.8

# Compute MSE
MSE = SSE/(N-k)
MSE

## [1] 63.97707

# Compute Test Statisitic
F.Test = MSTr/MSE
F.Test

## [1] 3.873263

# Find critical value
```

```
CV = qf(.95,3,16)
CV

## [1] 3.238872
```

5. Since we find that Our test statistic is greater than our critical value we will reject our null hypothesis. Thus the mean value is not the same for ewach sample and at least 2 samples have different means.

Tukey-Kramer Test of Pairwise Comparisons

In our previous example we concluded that all means are not equal. The F test does not tell usa which are different from the rest. We preform the Tukey-Kramer test on each pair to determine where two means are significantly different. This is a test is a *pairwise comparison*.

The Tukey-Kramer Test Statistic : Allow μ_i and μ_j represent two population means. The test statistic for the Tukey-Kramer Test of $H_0\colon \mu_i = \mu_j$ versus $H_0\colon \mu_i \neq \mu_j$ is

$$q = \frac{\left|\bar{x}_i - \bar{x}_j\right|}{\sqrt{\frac{MSE}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

When H_0 is true, the test statistic q has a distribution called the **Studentized range distribution**.

To find our critical value we use `qtukey()`.

Tukey-Kramer Test Layout

To test $H_0\colon \mu_i = \mu_j$ versus $H_0\colon \mu_i \neq \mu_j$

1. Check our assumptions stated before.
2. Compute sample means.
3. Compute MSE and critical value.

Now Compute steps 4-5 for each pair of sample means

4. Compute Test Statistic q.
5. Decided whether to reject.

Example

Using our previous example figure out which two samples differ.

Solution

1,2 & 3. Our assumptions have already been checked, and sample means were given. MSE was already previously computed. Test statistic at 5% Sig. level is `qtukey(.95,4,16)` = 4.046

4 & 5. Compute Test statistic for each pair

```
# first create a vector that will be easy to compute the difference of each mean pair.
# here xi will be a vector with our starting sample A and xj = will be the required match to make every possible pair.
A = 253.8
B = 263.2
C = 271.0
D = 262.0

xi = c(A,A,A,B,B,C)
xj = c(B,C,D,C,D,D)
q = abs(xi-xj)/(sqrt((MSE/2)*(1/5 + 1/5)))


# If q is greater than our critical value then we know that those samples had different means
q > 4.046

## [1] FALSE  TRUE FALSE FALSE FALSE FALSE

# Our results indicate that sample A and C's mean differ.There is not enough evidence to conclude that any other means differ.
```

Two-Way Analysis of Variance

Two-way ANOVA involves *two-factor experiments* where there are two factors. An example containing tabular data D is below where the two factors are teaching style and the size of the class with our output being final exam scores.

```
##             Lecture       Interactive   Combination
## Large Class "62 74 68 69" "87 75 69 84" "66 66 81 75"
## Small Class "75 77 82 74" "75 87 89 94" "76 84 90 93"
```

Assumptions for Two-Way ANOVA

1. Populations are approximately normal

2. Sample sizes are the same

3. Populations must have the same variance (aka: balanced data.)

Test procedure

1. Check assumptions and test the hypothesis that there is no interaction between the factors. Here use fairly large sig. level such as 10%.

2. If we reject that our null hypothesis from (1.) then stop since results might be misleading. Otherwise continue and test if there are differences among the means of factors one an two.

Example

In this example I am going tot format the data the way it would be needed to do a two way ANOVA test. We will make a data frame ANOVA2 with classes numeric and factors. Note: normally you would input this data then have to convert to factors later.

```
##   Final_Score Class_Type Class_Size
## 1          62        Lec         Sm
## 2          74        Lec         Sm
## 3          68        Lec         Sm
## 4          69        Lec         Sm
## 5          87        Int         Sm
## 6          75        Int         Sm

# Now we will test our data now that it is formatted
Test = aov(Final_Score ~ Class_Type + Class_Size +
             Class_Type:Class_Size, data = ANOVA2)
summary(Test)

##                       Df Sum Sq Mean Sq F value  Pr(>F)
## Class_Type             2  399.2   199.6   4.273 0.03031 *
## Class_Size             1  600.0   600.0  12.842 0.00212 **
## Class_Type:Class_Size  2   43.8    21.9   0.468 0.63355
## Residuals             18  841.0    46.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here we can see from `Class_Type:Class_Size` our p value is .643 which is greater than .10 so we can conclude that there is no interaction between factors. From the other results we see .03 and .02 implying that compared to a significance level of 5% our results indicate that there are differences in the mean of Class size and Class type. Thus we could conclude that the style of teaching and class size impact the results of finals scores.

Interaction Plots : we can compute the sample means for each category and plot it to see the magnitude of the interactions in a two way ANOVA.

Chapters 15 : Nonparametric Statistics

Definitions

Parametric Tests : Make assumptions about the distribution of a population.

Nonparametric Tests : Do NOT make an assumption about any specific distribution.

The 3 Tests that are Nonparametric are the **Sign Test**, **Wilcoxon Rank-Sum Test** and the **Wilcoxon Signed Rank-Sum Test**.

The Sign Test

The Sign Test is used to test hypothesis about a population median m.

Advantage and Disadvantage

Advantage : The sign test is valid for any continuous population. It does not require the assumption of normality.

Disadvantage : When the population is approximately normal, the sign test is less likely to reject a false null hypothesis than the t-test.

Test Statistic of the Sign Test

All x to be the minimum number of data points on either side of the supposed median m_0. Let n be the total number of plus and minus signs

If n $\leq$ 25 then the test statistic is x

If n > 25 then the test statistic is :

$$z = \frac{x + 0.5 - n/2}{\sqrt{n}/2}$$

To find the critical value we use a table or `binom.test`

Preforming the Sign Test

1. State null and alt. Hypothesis. The null hypothesis specifies a value for the population median m. We call this value m_0. The Null hypothesis is of the form $H_0\!: m = m_0$. The alt. hypothesis is stated one of three ways : $H_1\!: m < m_0$, $H_1\!: m > m_0$ or $H_1\!: m \neq m_0$.

2. Determine the data points above and below the median value m_0.

3. Compute test statistic and critical point.

4. State conclusion.

Example Consider our Data set E which is a set of the times it takes to complete a surgery in minutes. Test if the median time is less than 170 minutes. Use an alpha level of 1 %.

```
##  [1] 149 144 218 153 134 152 148 144 178 107 199 135 171 110 160 119  86
127 106
## [20] 153 169 153 153 173 156 145 205 132 169 174 130 175
```

Solution

1. We have $H_0: m = 170$ and $H_1: m < 170$

2. Assigning values

```
# Here we assign values

Values =  E < 170

table(Values)

## Values
## FALSE  TRUE
##     8    24
```

3. From here we can see that our table that we have 32 data points and an x value of 8.

From here we could follow the text and use the test statistic but instead lets use R

```
# we use this code since what were doing is basically the binomial dist.

# we use 24 here since we are testing less than.
binom.test(24,32)

##
##  Exact binomial test
##
## data:  24 and 32
## number of successes = 24, number of trials = 32, p-value = 0.007
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.5659506 0.8853840
## sample estimates:
## probability of success
##                   0.75
```

Here our out put shows us how certain we are at estimating the median value and gives us a p-value lower than .01 thus we would reject our null hypothesis.

LATER >>> The Rank-Sum Test & The Signed-Rank Test