

math612.0 A1: From numbers through algebra to calculus and linear algebra

Gunnar Stefansson (editor) with contributions from very many students

7. mars 2022

Copyright This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

Acknowledgements

This project has received direct funding from the EU H2020 project Minouw, to provide technical support for students who take tutorials on the EAFM in general and discard models in particular.

Most of the content has been developed as a part of giving courses at the University of Iceland and at GRÓ-FTP, with additions and developments in 2019-2021 funded in part by FarFish.

MareFrame is a EC-funded RTD project which seeks to remove the barriers preventing more widespread use of the ecosystem-based approach to fisheries management.

<http://mareframe-fp7.org>

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no.613571.

<http://mareframe-fp7.org>

The University of Iceland uses the tutor-web in many courses and funds content-development as a part of this use.

The University of Iceland Research Fund has funded many of the studies developing algorithms uses in tutor-web.

<http://www.hi.is/>

This project has received funding from the European Commission's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 634495 for the project Science, Technology, and Society Initiative to minimize Unwanted Catches in European Fisheries (MINOUW).

<http://minouw-project.eu/>

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 727891.

www.farfish.eu

Efnisyfirlit

1	Numbers, arithmetic and basic algebra	13
1.1	Natural Numbers	13
1.1.1	Details	13
1.1.2	Examples	14
1.2	Starting with \mathbb{R}	14
1.2.1	Details	14
1.2.2	Examples	14
1.3	The Integers	15
1.3.1	Details	15
1.3.2	Examples	15
1.4	Rational numbers	15
1.4.1	Details	16
1.4.2	Examples	16
1.5	The real line	21
1.5.1	Details	21
1.5.2	Examples	21
2	Data vectors	23
2.1	The plane	23
2.1.1	Details	23
2.1.2	Examples	23
2.2	Simple plots in \mathbb{R}	24
2.2.1	Examples	24
2.3	Data	25
2.3.1	Details	25
2.3.2	Examples	25
2.4	Indices for a data vector	25
2.4.1	Details	26
2.4.2	Examples	26
2.5	Summation	26
2.5.1	Examples	27
3	More on algebra	28
3.1	Some Squares	28
3.1.1	Details	28
3.2	Pascal's Triangle	28
3.2.1	Details	28
3.2.2	Examples	28
3.3	Factorials	29
3.3.1	Details	29
3.3.2	Examples	29
3.4	Combinations	30
3.4.1	Details	30
3.4.2	Examples	30
3.5	The binomial theorem	31
3.5.1	Details	31
3.5.2	Examples	31

4	Discrete random variables and the binomial distribution	32
4.1	Simple probabilities	32
4.1.1	Details	32
4.1.2	Examples	32
4.2	Random variables	33
4.2.1	Examples	33
4.2.2	Handout	34
4.3	Simple surveys with replacement	34
4.3.1	Examples	34
4.4	The binomial distribution	35
4.4.1	Examples	35
4.5	General discrete probability distributions	38
4.5.1	Details	38
4.5.2	Examples	39
4.6	The expected value or population mean	39
4.6.1	Details	39
4.6.2	Examples	40
4.7	The population variance	42
4.7.1	Details	42
4.7.2	Examples	42
5	Functions	43
5.1	Functions of a single variable	43
5.1.1	Details	43
5.1.2	Examples	43
5.2	Functions in \mathbb{R}	43
5.3	Ranges and plots in \mathbb{R}	44
5.3.1	Examples	44
5.4	Plotting functions	44
5.4.1	Examples	44
5.5	Functions of several variables	45
5.5.1	Examples	45
6	Polynomials	46
6.1	The general polynomial	46
6.1.1	Details	46
6.2	The quadratic	46
6.2.1	Details	46
6.3	The cubic	47
6.4	The Quartic	47
6.5	Solving the linear equation	47
6.5.1	Details	47
6.6	Roots of the quadratic equation	48
6.6.1	Details	48
6.6.2	Examples	48
7	Simple data analysis in \mathbb{R}	50
7.1	Entering data; dataframes	50
7.1.1	Details	50
7.1.2	Examples	50
7.2	Histograms	50

7.2.1	Examples	51
7.3	Bar Charts	51
7.3.1	Details	51
7.4	Mean, standard error, standard deviations	51
7.4.1	Details	51
7.5	Scatter plots and correlations	53
7.5.1	Details	53
7.5.2	Examples	53
8	Indices and the apply commands in R	54
8.1	Giving names to elements	54
8.1.1	Examples	54
8.2	Regular matrix indices and naming	54
8.2.1	Details	54
8.2.2	Examples	54
8.3	The apply command	55
8.4	The tapply command	55
8.4.1	Examples	55
8.5	Logical indexing	56
8.5.1	Examples	56
8.6	Lists, indexing lists	56
8.6.1	Examples	56
9	Functions of functions and the exponential function	58
9.1	Exponential growth and decline	58
9.1.1	Details	58
9.2	The exponential function	58
9.2.1	Details	58
9.3	Properties of the exponential function	59
9.4	Functions of functions	59
9.4.1	Details	59
9.4.2	Examples	59
9.5	Storing and using R code	59
9.5.1	Examples	59
9.6	Storing and calling functions in R	60
9.6.1	Examples	60
10	Inverse functions and the logarithm	61
10.1	Inverse Function	61
10.1.1	Details	61
10.1.2	Examples	61
10.2	When the inverse exists: The domain question	61
10.2.1	Examples	61
10.3	The base 10 logarithm	62
10.3.1	Details	62
10.3.2	Examples	62
10.4	The natural logarithm	63
10.5	Properties of logarithm(s)	63
10.5.1	Details	63
10.6	The exponential function and the logarithm	64
10.6.1	Details	64

10.6.2	Examples	64
11	Continuity and limits	66
11.1	The concept of continuity	66
11.1.1	Details	66
11.2	Discrete probabilities and cumulative distribution functions	66
11.2.1	Details	66
11.2.2	Examples	67
11.3	Notes on discontinuous function	67
11.3.1	Details	67
11.4	Continuity of polynomials	68
11.4.1	Details	68
11.5	Simple Limits	68
11.5.1	Details	68
11.5.2	Examples	69
11.6	More on limits	69
11.6.1	Examples	69
11.7	One-sided limits	71
11.7.1	Details	71
12	Sequences and series	72
12.1	Sequences	72
12.1.1	Details	72
12.1.2	Examples	72
12.2	Convergent sequences	72
12.2.1	Details	72
12.2.2	Examples	72
12.3	Infinite sums (series)	73
12.3.1	Details	73
12.3.2	Examples	73
12.4	The exponential function and the Poisson distribution	74
12.4.1	Details	74
12.5	Relation to expected values	74
12.5.1	Details	74
13	Slopes of lines and curves	76
13.1	The slope of a line	76
13.1.1	Details	76
13.2	Segment slopes	76
13.2.1	Details	76
13.3	The slope of $y = x^2$	77
13.3.1	Examples	77
13.4	The tangent to a curve	77
13.4.1	Details	77
13.4.2	Examples	78
13.5	The slope of a general curve	78
13.5.1	Details	78

14 Derivatives	79
14.1 The derivative as a limit	79
14.1.1 Details	79
14.2 The derivative of $f(x) = a + bx$	79
14.2.1 Details	79
14.3 The derivative of $f(x) = x^n$	79
14.3.1 Details	80
14.4 The derivative of \ln and \exp	80
14.4.1 Details	80
14.5 The derivative of a sum and linear combination	80
14.5.1 Details	80
14.5.2 Examples	81
14.6 The derivative of a polynomial	81
14.6.1 Details	81
14.6.2 Examples	81
14.7 The derivative of a product	81
14.7.1 Details	81
14.7.2 Examples	82
14.8 Derivatives of composite functions	82
14.8.1 Examples	82
15 Applications of differentiation	84
15.1 Tracking the sign of the derivative	84
15.1.1 Details	84
15.1.2 Examples	84
15.2 Describing extrema using f''	84
15.2.1 Details	84
15.3 The likelihood function	85
15.3.1 Details	85
15.3.2 Examples	86
15.4 Plotting the likelihood	86
15.4.1 Examples	86
15.5 Maximum likelihood estimation	86
15.5.1 Details	86
15.5.2 Examples	86
15.6 Least squares estimation	87
15.6.1 Details	87
15.6.2 Examples	87
16 Integrals and probability density functions	89
16.1 Area under a curve	89
16.1.1 Details	89
16.2 The antiderivative	89
16.2.1 Examples	89
16.3 The fundamental theorem of calculus	90
16.3.1 Detail	90
16.3.2 Examples	90
16.4 Density functions	91
16.4.1 Details	91
16.4.2 Examples	91
16.5 Probabilities in R: The normal distribution	92

16.5.1	Details	92
16.5.2	Examples	93
16.6	Some rules of integration	93
16.6.1	Examples	93
16.6.2	Handout	93
17	Principles of programming	95
17.1	Modularity	95
17.1.1	Details	95
17.1.2	Examples	95
17.2	Modularity and functions	95
17.2.1	Details	95
17.2.2	Examples	95
17.3	Modularity and files	96
17.3.1	Details	96
17.3.2	Examples	96
17.4	Structuring an R project	97
17.4.1	Details	97
17.4.2	Examples	98
17.5	Loops, for	98
17.5.1	Details	98
17.5.2	Examples	98
17.6	The if and ifelse commands	99
17.6.1	Examples	99
17.7	Indenting	100
17.7.1	Details	100
17.8	Comments	100
17.8.1	Examples	100
18	The Central Limit Theorem and related topics	102
18.1	The Central Limit Theorem	102
18.1.1	Details	102
18.1.2	Examples	103
18.2	Properties of the binomial and Poisson distributions	103
18.2.1	Details	103
18.2.2	Examples	104
18.3	Monte Carlo simulation	105
18.3.1	Examples	105
19	Miscellanea	107
19.1	Simple probabilities in R	107
19.1.1	Examples	107
19.2	Computing normal probabilities in R	108
19.2.1	Details	108
19.2.2	Examples	108
19.3	Introduction to hypothesis testing	109
19.3.1	Details	109

20	Multivariate probability distributions	111
20.1	Joint probability distribution	111
20.1.1	Details	111
20.1.2	Examples	111
20.2	The random sample	112
20.2.1	Details	112
20.2.2	Examples	112
20.3	The sum of discrete random variables	113
20.3.1	Details	113
20.3.2	Examples	113
20.4	The sum of two continuous random variables	114
20.4.1	Details	114
20.4.2	Examples	114
20.5	Means and variances of linear combinations of independent random variables	115
20.5.1	Details	115
20.5.2	Examples	115
20.6	Means and variances of linear combinations of measurements	116
20.6.1	Examples	116
20.7	The joint density of independent normal random variables	117
20.7.1	Details	117
20.8	More general multivariate probability density functions	117
20.8.1	Examples	117
20.8.2	Handout	118
21	Some distributions related to the normal	119
21.1	The normal and sums of normals	119
21.1.1	Details	119
21.1.2	Examples	119
21.2	The Chi-square distribution	121
21.2.1	Details	121
21.3	Sum of Chi square Distributions	121
21.3.1	Details	121
21.4	Sum of squared deviation	122
21.4.1	Details	122
21.5	The t-distribution	122
21.5.1	Details	122
22	Estimation, estimates and estimators	124
22.1	Ordinary least squares for a single mean	124
22.1.1	Examples	124
22.2	Maximum likelihood estimation	124
22.2.1	Examples	124
22.2.2	Detail	125
22.3	Ordinary least squares	125
22.3.1	Details	125
22.4	Random variables and outcomes	126
22.4.1	Details	126
22.4.2	Examples	126
22.5	Estimators and estimates	126
22.5.1	Details	126

23	Test of hypothesis, P values and related concepts	128
23.1	The principle of the hypothesis test	128
23.1.1	Examples	128
23.2	The one sided z test for normal mean	129
23.2.1	Examples	129
23.3	The two-sided z test for a normal mean	130
23.3.1	Details	130
23.3.2	Examples	130
23.4	The one-sided t-test for a single normal mean	130
23.4.1	Details	131
23.4.2	Examples	131
23.5	Comparing means from normal populations	131
23.5.1	Details	131
23.6	Comparing means from large samples <Ól.B.M.>	132
23.6.1	Details	132
23.7	The P-value	133
23.7.1	Examples	133
23.8	The concept of significance	133
23.8.1	Details	133
24	Power and sample sizes	135
24.1	The power of a test	135
24.1.1	Details	135
24.2	The power of tests for proportions	135
24.2.1	Examples	135
24.3	The Power of the one sided z test for the mean	138
24.3.1	Details	138
24.3.2	Examples	139
24.4	Power and sample size for the one-sided z-test for a single normal mean	140
24.4.1	Details	140
24.4.2	Examples	140
24.5	The non central t - distribution	141
24.5.1	Details	141
24.6	The power of t-test for a normal mean (warning: errors)	141
24.6.1	Details	141
24.7	Power and sample size for the one sided t-test for a mean	142
24.7.1	Details	142
24.7.2	Examples	142
24.8	The power of the 2-sided t-test	145
24.8.1	Details	145
24.8.2	Examples	146
24.9	The power of the 2-sample one and two-sided t-tests	147
24.9.1	Details	147
24.10	Sample sizes for two-sample one and two-sided t-tests	149
24.10.1	Details	150
24.11	A case study in power	150
24.11.1	Handout	150

25	Vectors and Matrix Operations	157
25.1	Numbers, vectors, matrices	157
25.1.1	Examples	157
25.2	Elementary Operations	158
25.2.1	Examples	158
25.3	The tranpose of a matrix	159
25.3.1	Details	159
25.3.2	Examples	159
25.4	Matrix multiplication	160
25.4.1	Details	160
25.4.2	Examples	160
25.5	More on matrix multiplication	161
25.6	Linear equations	161
25.6.1	Details	161
25.6.2	Examples	162
25.7	The unit matrix	162
25.8	The inverse of a matrix	163
25.8.1	Examples	163
26	Some notes on matrices and linear operators	164
26.1	The matrix as a linear operator	164
26.1.1	Examples	164
26.2	Inner products and norms	165
26.2.1	Details	165
26.2.2	Examples	166
26.3	Orthogonal vectors	166
26.3.1	Details	166
26.4	Linear combinations of i.i.d. random variables	167
26.4.1	Examples	167
26.5	Covariance between linear combinations of i.i.d random variables	168
26.5.1	Details	168
26.5.2	Examples	168
26.6	Random vectors	169
26.6.1	Details	169
26.6.2	Examples	169
26.7	Transforming random vectors	170
26.7.1	Details	170
27	Ranks and determinants	172
27.1	The rank of a matrix	172
27.1.1	Details	172
27.1.2	Examples	172
27.2	The determinant	174
27.2.1	Details	174
27.2.2	Examples	175
27.3	Ranks, inverses and determinants	176
27.3.1	Details	177

28	Multivariate calculus	178
28.1	Vector functions of several variables	178
28.1.1	Examples	178
28.2	The gradient	179
28.2.1	Details	180
28.2.2	Examples	180
28.3	The Jacobian	180
28.3.1	Details	181
28.3.2	Examples	181
28.4	Univariate integration by substitution	181
28.4.1	Details	182
28.5	Multivariate integration by substitution	182
28.5.1	Details	182
28.5.2	Examples	183
29	The multivariate normal distribution and related topics	185
29.1	Transformations of random variables	185
29.1.1	Details	185
29.2	The multivariate normal distribution	185
29.2.1	Details	185
29.3	Univariate normal transforms	187
29.3.1	Details	187
29.4	Transforms to lower dimensions	187
29.4.1	Details	187
29.5	The OLS estimator	188
29.5.1	Details	188
30	Independence, expectations and the moment generating function	190
30.1	Independent random variables	190
30.1.1	Details	190
30.2	Independence and expected values	191
30.2.1	Details	191
30.2.2	Examples	192
30.3	Independence and the covariance	193
30.4	The moment generating function	193
30.4.1	Examples	193
30.5	Moments and the moment generating function	194
30.5.1	Details	194
30.6	The moment generating function of a sum of random variables	194
30.6.1	Details	194
30.7	Uniqueness of the moment generating function	195
31	The gamma distribution	196
31.1	The gamma distribution	196
31.1.1	Details	196
31.2	The mean, variance and mgf of the gamma distribution	196
31.2.1	Details	196
31.3	Special cases of the gamma distribution: The exponential and chi-squared distributions	199
31.3.1	Details	199
31.4	The sum of gamma variables	201
31.4.1	Details	201

32	Notes and examples: The linear model	202
32.1	Simple linear regression in R	202
32.1.1	Details	202
32.2	Multiple linear regression	204
32.3	The one-way model	204
32.3.1	Details	204
32.3.2	Examples	205
32.4	Random effects in the one-way layout	205
32.4.1	Details	206
32.5	Linear mixed effects models (lmm)	207
32.5.1	Details	207
32.5.2	Examples	208
32.6	Maximum likelihood estimation in lmm	208
32.6.1	Details	208
33	Some regression topics	210
33.1	Poisson regression	210
33.2	The generalized linear model (GLM)	210
33.2.1	Details	210
34	Overview drills	212

1 Numbers, arithmetic and basic algebra

1.1 Natural Numbers

The positive integers are called natural numbers.

These numbers can be added, multiplied together and so forth.

Notation: $\mathbb{N} = \{1, 2, 3, 4, \dots\}$

Subtraction and division are not defined on these numbers.

An arbitrary element of \mathbb{N} is most commonly denoted by i , j , n , or m , but any symbol can be used.

1.1.1 Details

Definition 1.1. The set of positive integers is usually denoted by \mathbb{N} , i.e. $\mathbb{N} = \{1, 2, 3, 4, \dots\}$ and is called the set of **natural numbers**. In some cases the number zero is included as a natural number, but here we will use the symbol \mathbb{N}_0 to denote the integers 0, 1, 2 and up.

Within this set of numbers it is possible to add and multiply numbers together. Arithmetic operations are denoted by $+$ for addition and \cdot (or \times) for multiplication. A natural number can also be raised to the power of a natural number, e.g. $3^5 = 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3$ or in general $m^n = m \cdot m \cdot \dots \cdot m$ (n times).

When stating general properties of the natural numbers one needs to use symbols to indicate that the property holds for an arbitrary number. It is not enough to just write the property for a few numbers. For example, to declare that one can interchange numbers in a sum, it is not enough to say $4 + 3 = 3 + 4$ but one must explicitly state "the addition operator has the property that any two natural numbers, $n, m \in \mathbb{N}$ satisfy $n + m = m + n$ ".

An arbitrary element of \mathbb{N} is most commonly denoted by i , j , n , or m , but any symbol, a , b , c , \dots , can be used.

Several rules of arithmetic apply (some by definition, others can be derived) such as

$$\begin{aligned} ab &= ba \\ a + b &= b + a \\ a + bc &= a + (bc) \\ a(b + c) &= ab + ac \\ (a + b) + c &= a + (b + c) \\ (ab)c &= a(bc) \end{aligned}$$

Subtraction and division are not generally defined. In addition, we define one integer, n , to the power of another, m , to mean n multiplied by itself m times: $n^m = \underbrace{n \cdot n \cdot \dots \cdot n}_m$.

Definition 1.2. The power is an **operator** just like addition and multiplication, and is defined to have higher priority than the other two.

1.1.2 Examples

Example 1.1. If we have $x = 4$ and $y = 2$ and want to evaluate

$$x^y + y^x$$

then we replace the values of x and y in the expression, and evaluate it, taking care to observe the correct order of operations:

$$4^2 + 2^4 = 16 + 16 = 32.$$

1.2 Starting with R

Download R from the R website: <http://www.r-project.org/>
Look at on-line information on R, and take the tutor-web R tutorial:
<http://tutor-web.net/stats/stats240.1>

Simple R commands:

- Assignment: $x \leftarrow -2$
- Arithmetic: $2 * 5 + 4$

1.2.1 Details

To assign values to a variable in R one can use " \leftarrow " or "="; however, these are **NOT** equivalent. Using the equals sign is confusing and therefore not recommended.

1.2.2 Examples

Example 1.2. Assigning values to a variable:

```
x ← -2  
y ← -3  
z ← -x + y
```

Example 1.3. Viewing assigned values:

Type the name, i.e. "z", to view the assigned value.

```
z  
[1] 5
```

1.3 The Integers

The set of positive and negative integers:

$$\mathbb{Z} = \{\dots, \dots, -2, -1, 0, 1, 2, \dots\}$$

1.3.1 Details

Definition 1.3. The set of all integers is denoted by \mathbb{Z} , i.e.

$$\mathbb{Z} = \{\dots, \dots, -2, -1, 0, 1, 2, \dots\}.$$

Note 1.1. Note that within this set it is possible to subtract as well as add and multiply. Within this set we cannot, however, in general, perform division.

When performing multiple mathematical operations within the same equation, i.e. $79 - 8 \cdot 3$, there is a conventional order for which the operations must be performed.

Definition 1.4. The conventional order of operations for equations with multiple mathematical operations is referred to as an **operator precedence**.

1.3.2 Examples

Example 1.4. To compute $79 - 8 \cdot 3$ start by multiplying and then subtracting:

$$79 - 8 \cdot 3 = 79 - 24 = 55$$

Example 1.5. To compute $15 - (24 + 36)$ we first note that the parentheses (brackets) imply a precedence; anything inside brackets should be evaluated first.

Thus, we first add 36 to 24 and then we subtract that from 15.

$$15 - (24+36) = 15 - 60 = -45$$

Note that the answer is a negative number.

Example 1.6. Simple arithmetic in R is easily done at the command prompt.

```
79-8*3
```

```
[1] 55
```

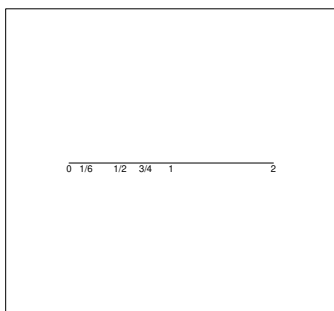
```
15-(24+36)
```

```
[1] -45
```

1.4 Rational numbers

Rational numbers are fractions denoted p/q , where p and q are integers. We can simplify fractions if the numerator and denominator contain common terms.

1.4.1 Details



Definition 1.5. Rational numbers are fractions denoted p/q , where p and q are integers. The set of all rational numbers is usually denoted \mathbb{Q} .

Note 1.2. Note that every integer is a rational number (obtained by taking $q = 1$).

We can simplify fractions if the numerator and denominator contain common terms.

When the rationals are ordered on to a line there are points missing, i.e. there are "gaps", for example there is no rational number p/q such that $(p/q)^2 = 2$.

1.4.2 Examples

Example 1.7. $\frac{2}{6} = \frac{2}{2 \cdot 3} = \frac{1}{3}$

The rational numbers can be put in order along a line as in the figure.

Example 1.8. As an elaborate example of a fraction, consider the evaluation of the quantity

$$\frac{\frac{2}{3} + \frac{2}{5}}{\frac{1}{3} + \frac{1}{2}}$$

Example 1.9. Evaluate

$$\frac{\frac{2}{3} + \frac{2}{5}}{\frac{1}{3} + \frac{1}{2}}$$

Solution: We can either start by calculating the numerator

$$\frac{2}{3} + \frac{2}{5}$$

or the denominator

$$\frac{1}{3} + \frac{1}{2}$$

Here we choose to start with the numerator. The first step is to make the two fractions in the numerator have a common denominator.

We can either find the least common denominator or use the product of the two denominators. Here they are the same number, 15.

So the first step is:

$$\frac{2}{3} \cdot \frac{5}{5} + \frac{2}{5} \cdot \frac{3}{3} = \frac{2 \cdot 5}{3 \cdot 5} + \frac{2 \cdot 3}{5 \cdot 3} = \frac{10}{15} + \frac{6}{15}$$

Now it is possible to add the two fractions, which is the second step:

$$\frac{10+6}{15} = \frac{16}{15}$$

Next, the same process has to be performed for the original denominator.

With the same method (LCM - least common multiple) we get:

$$\frac{1 \cdot 2}{3 \cdot 2} + \frac{1 \cdot 3}{2 \cdot 3} = \frac{2}{6} + \frac{3}{6} = \frac{5}{6}$$

Then the total answer is:

$$\frac{\frac{16}{15}}{\frac{5}{6}} = \frac{16}{15} \cdot \frac{6}{5} = \frac{96}{75} = \frac{96/3}{75/3} = \frac{32}{25}$$

We can see that in the last step of the equation, the factor has been simplified. To do this we use factoring. Here we obtain:

$$= \frac{96}{75} = \frac{3 \cdot 32}{3 \cdot 25}$$

We can now remove "3", or the multiplier, as it is on both sides of the fraction. So we have:

$$= \frac{32}{25} = \frac{25}{25} + \frac{7}{25} = 1 \frac{7}{25}$$

In step 1 above we used Cross-Multiplication.

Definition 1.6. Cross-Multiplication is when we multiple the numerator by the reciprocal of the denominator.

So in this case we rewrite

$$\frac{\frac{16}{15}}{\frac{5}{6}}$$

or

$$\frac{16}{15} \div \frac{5}{6}$$

as

$$\frac{16}{15} \cdot \frac{6}{5}$$

As you can see all we are doing is turning

$$\frac{5}{6}$$

upside down: and multiplying it with

$$\frac{16}{15}$$

This gives:

$$\frac{96}{75}$$

In some cases it is possible to draw a **square root** of a fraction $s = \frac{p}{q}$, i.e. find a number $r \in \mathbb{Q}$ such that $r^2 = s$. The square root is denoted \sqrt{r} .

Example 1.10. Consider the expression

$$\left(\sqrt{\frac{1}{9}} \times 2^4\right) + \left(\frac{1}{5} \times \sqrt{25}\right)$$

To evaluate this expression, first consider separately the two parts on each side of the plus symbol.

The first part is

$$\left(\sqrt{\frac{1}{9}} \times 2^4\right)$$

and the second part is

$$\left(\frac{1}{5} \times \sqrt{25}\right)$$

In addition, by definition of root,

$$\sqrt{\frac{1}{9}} = \frac{1}{3}$$

First part:

$$\left(\sqrt{\frac{1}{9}} \times 2^4\right) = \frac{1}{3} \times 16 = \frac{16}{3}$$

Second part:

$$\left(\frac{1}{5} \times \sqrt{25}\right) = \frac{1}{5} \times 5 = 1$$

Finally, add the first part and the second part:

$$\frac{16}{3} + 1 = \frac{19}{3}$$

Example 1.11. Consider the following fraction example, to be solved step by step:

$$\frac{\frac{4}{2} + \left(\frac{1}{4} \cdot \frac{5}{3}\right)}{\frac{2}{6} \div \frac{1}{5}}$$

First we need to be aware of operator precedence, sometimes called BODMAS (brackets, multiplication/division, then addition/subtraction).

$$\left(\frac{1}{4} \cdot \frac{5}{3}\right) = \frac{5}{12}$$

After solving the bracket we can proceed with adding

$$\frac{4}{2}$$

to

$$\frac{5}{12}$$

as there is no other action left for the nominator of the main fraction. So:

$$\frac{4}{2} + \frac{5}{12}$$

When adding fractions together we first have to find a common denominator, in this case 12 would work as

$$2 \cdot 6 = 12$$

So we multiply both the numerator and the denominator of that fraction by 6 and then add the two numerators of the fractions together, keeping the same denominator.

$$\frac{4}{2} + \frac{5}{12} = \frac{4 \cdot 6}{2 \cdot 6} + \frac{5}{12} = \frac{24}{12} + \frac{5}{12} = \frac{29}{12}$$

Now we have the top half of the fraction solved. We then proceed with dividing the two fractions of the bottom half. When dividing fractions we use the so called cross multiplication technique. This arithmetic trick is derived from the fact that if you divide a fraction by its duplicate you get 1. If you multiple a fraction by its reciprocal (it's reverse) you also get 1. Like so:

$$\frac{1}{2} \div \frac{1}{2} = 1$$

and

$$\frac{1}{2} \cdot \frac{2}{1} = 1$$

These functions always provide the same result and therefore we can turn the fraction we are dividing by upside down and multiply it to the other fraction as that is usually much easier.

We can therefore rewrite

$$\frac{2}{6} \div \frac{1}{5}$$

as

$$\frac{2}{6} \cdot \frac{5}{1} = \frac{10}{6}$$

We've now solved both halves of the original fraction and can therefore proceed to solve it, again with the cross multiplication technique as fractions are after all just divisions:

$$\frac{29}{12} \div \frac{10}{6} = \frac{29}{12} \cdot \frac{6}{10} = \frac{174}{120}$$

Now

$$\frac{174}{120}$$

is a pretty bad looking fraction and we'd preferably like to simplify it.

To do this we use factoring.

Definition 1.7. Factoring essentially means to break a number down into its smallest factors or multipliable prime numbers.

In this case we get

$$\frac{2 \cdot 3 \cdot 29}{2 \cdot 3 \cdot 20}$$

These are the smallest prime numbers that can multiply together into 174 and 120 respectively.

A way of doing this in your head is by first dividing both numbers (174,120) by two. Which gives us:

$$\frac{2 \cdot 87}{2 \cdot 60}$$

and then dividing those numbers (87,60) by 3, since they can't be divided by 2. Dividing by 3 gives you

$$\frac{3 \cdot 29}{3 \cdot 20} = \frac{29}{20}$$

which is a lot nicer than

$$\frac{174}{120}$$

The reasoning behind this factoring simplification is that we can remove multipliers if they are on both sides of a fraction. This is because the result of a fraction where the numerator and the denominator are the same is always 1. Like so:

$$\frac{1}{1} = 1$$

or

$$\frac{2}{2} = 1$$

or

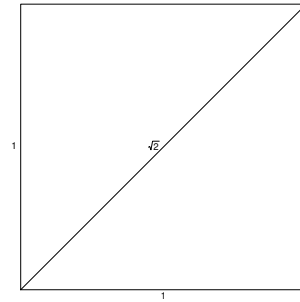
$$\frac{3}{3} = 1$$

The final answer therefore is

$$\frac{\frac{4}{2} + (\frac{1}{4} \cdot \frac{5}{3})}{\frac{2}{6} \div \frac{1}{5}} = \frac{29}{20}$$

1.5 The real line

Some obvious numbers are not fractions.
The set of numbers making up the real line is denoted by the symbol \mathbb{R} .



The diagonal of a rectangle with unit side lengths of $\sqrt{2}$.
Note that $\sqrt{2}$ is not a fraction.

1.5.1 Details

Some obvious numbers, which commonly occur, are not fractions. These are in between the rational numbers (fractions). Filling in the missing points to obtain a continuum results in the set of "real numbers".

Denoted by \mathbb{R} the entire set of "real numbers" which corresponds to "filling in" the "missing pieces of the line".

1.5.2 Examples

Example 1.12. If C is the circumference of a circle and D is the diameter and we define $\pi = \frac{C}{D}$ then π is not a fraction.

Example 1.13. One example of a non fraction is the number e (Euler's number) which can be defined by

$$e = \sum_{n=0}^{\infty} \frac{1}{n!}$$

Example 1.14. If you have a right triangle with unit side length, what is the length of its hypotenuse and what class of numbers does it belong to?

An isosceles triangle is defined as having adjacent and opposite sides of same length, connected by a 90° angle. Unit side length of these, refers to a side length of

$$1$$

.

As we have a 90° angle, we can use Pythagoras' theorem:

$$a^2 + b^2 = c^2$$

With

$$a = \textit{adjacent}$$

$$b = \textit{opposite}$$

$$c = \textit{hypotenuse}$$

So with

$$a, b = 1$$

:

$$c^2 = 1^2 + 1^2$$

$$c^2 = 1 + 1$$

$$c^2 = 2$$

We take the square root to get

$$c$$

$$c = \sqrt{2}$$

Now that we answered the first part of the question, it needs to be defined, which class of number

$$\sqrt{2}$$

belongs to.

$$\sqrt{2}$$

is an irrational number, and belongs thereby to the set of real numbers

$$\mathbb{R}$$

Real numbers can be imagined as points on an infinitely long line, which is also called the real line.

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students
This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

2 Data vectors

2.1 The plane

Pairs of numbers can be depicted as points on a plane.
The plane is normally denoted by \mathbb{R}^2 .

2.1.1 Details

Pairs of numbers can be depicted as points on a plane.

Definition 2.1. A **plane** is a perfectly flat surface with no thickness and no end, it can extend forever in all directions. It has two-dimensions, length and width. We need two values to find a point on the plane.

Normally we talk about "the plane" as the collection of all pairs of numbers and denoted it by

$$\mathbb{R}^2 = \{(x,y) : x,y \in \mathbb{R}\}$$

, giving coordinates to each point.

2.1.2 Examples

Example 2.1. Plotting the point (2,4) in the x-y plane using R.

```
plot(2,4,xlim=c(0,6),ylim=c(0,6),xlab="x",ylab="y",cex=2)  
text(2,4,"(2,4)",pos=4,cex=2)
```

Additional points can be added using the *points* function:

```
points(3,5, cex = 0.5) ## a point at (3,5)
```

If you have 2 sets of coordinates on a plane you can calculate the distance between the 2 points and graph the line connecting the points

Example 2.2. What is the distance between the 2 points (3,9) and (5,1)?

We will use the Pythagorean theorem:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

We insert our values into the formula:

$$d = \sqrt{(5 - 3)^2 + (1 - 9)^2}$$

When we combine inside the parenthesis we get:

$$d = \sqrt{(2)^2 + (-8)^2}$$

Squaring both terms:

$$d = \sqrt{4 + 64}$$

Then we take the square root:

$$d = \sqrt{68}$$

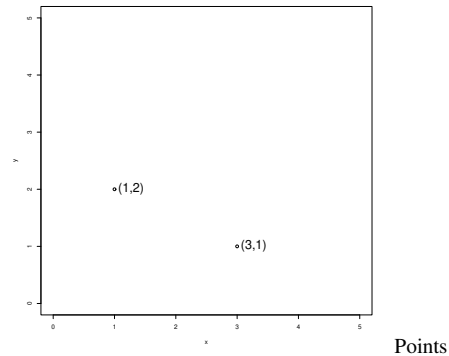
The result:

$$d = 8.2462$$

2.2 Simple plots in R

Graphing functions in R

- `plot` - plots a scatter plot (as a line plot)
- `points` - adds points to a plot
- `text` - adds text to a plot
- `lines` - adds lines to a plot



2.2.1 Examples

Example 2.3. `plot(2,3)`

gives a single plot and

```
plot(2,3, xlim=c(0,5), ylim=c(0,5))
```

gives a single plot but forces both axes to range from 0 to 5.

Example 2.4. The following R commands can be used to generate a plot with two points:

```
plot(1,2,xlim=c(0,5),ylim=c(0,5),xlab="x",ylab="y")
```

```
points(3,1)
```

```
text(1,2,"(1,2)",pos=4, cex=2)
```

```
text(3,1,"(3,1)",pos=4, cex=2)
```

Example 2.5. In this example, we plot 3 points. The first two points are by including vectors with a length of 2 as the x and y arguments of the plot function. The third plot was added with the points function. The second and third points were labeled using the text function and a line was drawn between them using the lines function.

Note 2.1. Note that if you are unsure of what format the arguments of an R function needs to be, you can call a help file by typing "?" before the function name (e.g. "?lines")

```
plot(c(2,3),c(3,4),xlim=c(2,6),ylim=c(1,5),xlab="x",ylab="y")
points(4,2)
text(3,4,"(3,4)",pos=4, cex=2)
text(4,2,"(4,2)",pos=4, cex=2)
lines(c(3,4), c(4,2))
```

2.3 Data

Data are usually a sequence of numbers, typically called a vector.

2.3.1 Details

When we collect data these are one or more sequences of numbers, collected into data vectors. We commonly think of these data vectors as columns in a table.

2.3.2 Examples

Example 2.6. In R, if the command

```
x <- c(4,5,3,7)
```

is given, then x contains a vector of numbers.

Example 2.7. Create a function in R, give it a name "Myfunction" which takes the sum of x,y.

```
Myfunction<- function(x,y) {
  sum(x,y)
}
```

If you input the vectors 1:3 and 4:7 into the function it will calculate the sum of $x \leftarrow (1+2+3)$ and $y \leftarrow (4+5+6+7)$ as follows

```
> Myfunction(1:3,4:7)
28
```

2.4 Indices for a data vector

If data are in a vector x, then we use indices to refer to individual elements.

2.4.1 Details

If i is an integer then x_i denotes the i 'th element of x .

Note that although we do not distinguish (much) between row- and column vectors, usually a vector is thought of as a column. If we need to specify the type of vector, row or column, then for vector x , the column vector would be referred to as x' and the row vector as x^T (the **transpose** of the original).

2.4.2 Examples

Example 2.8. If $x = (4, 5, 3, 7)$ then $x_1 = 4$ and $x_4 = 7$

Example 2.9. How to remove all indices below a certain value in R

```
x <- c(1,5,8,9,4,16,12,7,11)
x
[1] 1 5 8 9 4 16 12 7 11
y <- x[x>10]
y
[1] 16 12 11
```

Example 2.10. Consider a function that takes to vectors

$$a \in \mathbb{R}^n, b \in \mathbb{N}^m$$

as arguments with

$$n \geq m$$

and

$$1 \leq b_1, \dots, b_m \leq n$$

. The function returns the sum

$$\sum_{i=1}^m a_{b_i} \tag{1}$$

Long version:

```
fN <- function(a,b)
result <- sum(a[b])
return(result)
```

Short version:

```
|fN <- function(a,b) sum(a[b])|
```

2.5 Summation

We use the symbol Σ to denote sums.
In R, the sum function adds numbers.

2.5.1 Examples

Example 2.11. If $x = (4, 5, 3, 7)$
then

$$\sum_{i=1}^4 x_i = x_1 + x_2 + x_3 + x_4 = 4 + 5 + 3 + 7 = 19$$

and

$$\sum_{i=2}^4 x_i = x_2 + x_3 + x_4 = 5 + 3 + 7 = 15.$$

Within R one can give the corresponding commands:

```
x<-c(4,5,3,7)
x
[1] 4 5 3 7
sum(x)
[1] 19
sum(x[2:4])
[1] 15
```

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students
This work is licensed under the Creative Commons Attribution-ShareAlike License. To
view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a
letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

3 More on algebra

3.1 Some Squares

If a and b are real numbers, then

$$(a + b)^2 = a^2 + 2ab + b^2$$

3.1.1 Details

If a, b are real numbers, then:

$$(a + b)^2 = a^2 + 2ab + b^2$$

This can be proven formally with the following argument:

$$\begin{aligned}(a + b)^2 &= (a + b)(a + b) \\ &= (a + b)a + (a + b)b \\ &= a^2 + ba + ba + b^2 \\ &= a^2 + 2ab + b^2\end{aligned}$$

3.2 Pascal's Triangle

Pascal's triangle is a geometric arrangement of the binomial coefficients in a triangle

$$\begin{array}{cccc} & & 1 & & \\ & & & 1 & & 1 \\ & 1 & & 2 & & 1 \\ \vdots & \vdots & & \vdots & & \vdots \end{array}$$

3.2.1 Details

$$\begin{array}{l} n = 0: \quad \quad \quad 1 \\ n = 1: \quad \quad 1 \quad 1 \\ n = 2: \quad \quad 1 \quad 2 \quad 1 \\ n = 3: \quad 1 \quad 3 \quad 3 \quad 1 \end{array}$$

To build Pascal's triangle, start with "1" at the top, and then continue placing numbers below it in a triangular pattern. Each number is just the two numbers above it added together (except for the edges, which are all "1").

3.2.2 Examples

Example 3.1. The following function in R gives you the Pascal's triangle for $n = 0$ to $n = 10$.

```
fN <- function(n) formatC(n, width=2)
for (n in 0:10) {
  cat(fN(n), ":", fN(choose(n, k = -2:max(3, n+2))))
}
```

```

    cat("\n")
}

0 :  0  0  1  0  0  0
1 :  0  0  1  1  0  0
2 :  0  0  1  2  1  0  0
3 :  0  0  1  3  3  1  0  0
4 :  0  0  1  4  6  4  1  0  0
5 :  0  0  1  5 10 10  5  1  0  0
6 :  0  0  1  6 15 20 15  6  1  0  0
7 :  0  0  1  7 21 35 35 21  7  1  0  0
8 :  0  0  1  8 28 56 70 56 28  8  1  0  0
9 :  0  0  1  9 36 84 126 126 84 36  9  1  0  0
10 : 0  0  1 10 45 120 210 252 210 120 45 10  1  0  0

```

Changing the numbers in the line `f or(n in 0:10)` will give different portions of the triangle.

3.3 Factorials

We define the factorial of an integer n as

$$n! = n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot 3 \cdot 2 \cdot 1$$

3.3.1 Details

Definition 3.1. We define the factorial of an integer n as

$$n! = n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot 3 \cdot 2 \cdot 1.$$

3.3.2 Examples

Example 3.2. Suppose you have 6 apples, $\{a, b, c, d, e, f\}$ and you want to put each one into a different apple basket, $\{1, 2, 3, 4, 5, 6\}$.

For the first basket you can choose from 6 apples $\{a, b, c, d, e, f\}$, and for the second basket you have then 5 apples to choose from and so it goes for the rest of the baskets, so for the last one you only have 1 apple to choose from.

The end result would then be: $6! = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 720$ possible allocations.

This could also be calculated in R with the factorial function:

```

factorial(6)
[1] 720

```

3.4 Combinations

The number of different ways one can choose a subset of size x from a set of n elements is determined using the following calculation:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

3.4.1 Details

Definition 3.2. A **combination** is an un-ordered collection of distinct elements

Suppose we want to toss a coin n times. In each toss we obtain head (H) or tail (T) resulting in a sequence of H,T,T,H, ... T.

How many of these possible sequences contain exactly x tails? There are n positions in the sequence, we can choose x of these in $\binom{n}{x}$ ways and put our "Ts" in those positions. If the probability of landing tails is p then each one of these sequences with exactly x tails has probability $p^x(1-p)^{n-x}$ so the total probability of landing exactly x tails in n independent tosses is

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

For convenience we define $0!$ to be 1.

3.4.2 Examples

Example 3.3. Consider tossing a coin four times.

(a) How many times will this experiment result in exactly two tails?

There are a total of 16 possible sequences of heads and tails from four tosses. These can simply all be written down to answer a question like this.

We get two tails in 6 of these tosses. We can explicitly write the corresponding combinations of two tails as follows

HHTT
HTHT
HTTH
THTH
TTHH
THHT

(b) How many times you will end up with 1 tail? The answer is 4 times and the output can be written as;

HHHT
HTHH
TTHH
HHTH

The case of a single tail is easy: The single tail can come up in any one of four positions.

3.5 The binomial theorem

$$(a + b)^n = \sum_{x=0}^n \binom{n}{x} a^x b^{n-x}$$

3.5.1 Details

If a and b are real numbers and n is an integer then the expression $(a + b)^n$ can be expanded as:

$$(a + b)^n = a^n + \binom{n}{1} a^{n-1} b + \binom{n}{2} a^{n-2} b^2 + \dots + \binom{n}{n-1} a b^{n-1} + b^n$$

$$(a + b)^n = \sum_{i=1}^n \binom{n}{x} a^x b^{n-x}$$

This can be seen by looking at $(a + b)^n$ as a product of n parentheses and multiply these by picking one item (a or b) from each. If we picked a from x parentheses and b from $(n - x)$, then the product is $a^x b^{n-x}$. We can choose the x a 's in a total of $\binom{n}{x}$ ways so the coefficient of $a^x b^{n-x}$ is $\binom{n}{x}$.

3.5.2 Examples

Example 3.4. Since

$$(a + b)^n = \sum_{x=0}^n \binom{n}{x} a^x b^{n-x},$$

it follows that

$$2^n = (1 + 1)^n = \sum_{x=0}^n \binom{n}{x}$$

i.e.

$$2^n = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \dots + \binom{n}{n}$$

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students
This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

4 Discrete random variables and the binomial distribution

4.1 Simple probabilities

4.1.1 Details

Of all the possible 3-digit strings, $\binom{3}{x}$ of them have x heads. So the probability of landing x heads is $\binom{3}{x}p^x(1-p)^{3-x}$.

4.1.2 Examples

Example 4.1. Consider a biased coin which has probability p of landing heads up. If we toss this coin 3 independent times the possible outcomes are:

<i>sequence</i>	<i>probability</i>	<i>Number of heads</i>
<i>HHH</i>	$p \cdot p \cdot p = p^3$	3
<i>HHT</i>	$p^2(1-p)$	2
<i>HTH</i>	$p^2(1-p)$	2
<i>HTT</i>	$p(1-p)^2$	1
<i>THH</i>	$p^2(1-p)$	2
<i>THT</i>	$p(1-p)^2$	1
<i>TTH</i>	$p(1-p)^2$	1
<i>TTT</i>	$(1-p)^3$	0

Example 4.2. It is also possible to aggregate these values into a table and describe only the number of heads obtained:

heads	probability $p(x)$
0	$(1-p)^3$
1	$3p(1-p)^2$
2	$3p^2(1-p)$
3	p^3

If we are only interested in the number of heads, then this table describes a **probability mass function** p , namely the probability $p(x)$ of every possible outcome x of the experiment.

Example 4.3. Given that a year is 365 days and each day has the same probability of being someone's birthday. What's the probability of at least 2 people sharing a birthday in a group of 25 people?

Now, calculating each of the possible outcomes could become very tedious. That is calculating the odds that 2 people share a birthday, 3 people, 4 people, etc. So instead we try to find out the odds that no one in the group shares a birthday and subtract those

odds from 1 (100%).

First, let's look at the odds of only two people having distinct birthdays.

$$\frac{365}{365} \cdot \frac{364}{365} = 0.9973$$

Person one can be born on any day and the odds of having a distinct birthday are therefore 1. The next person can be born on everyday but the 1 the other person was born on, so 364 days.

Now let's say we add the 3rd person and calculate his/her odds of having a distinct birthday.

$$\frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} = 0.9918$$

This can also be rewritten as

$$\frac{365 \cdot 364 \cdot 363}{365^3}$$

And we can do this on and on for all the 25 people we are interested in. But that may also become a bit tedious. So we use factorials instead. So instead of doing

$$\frac{365 \cdot 364 \cdot 363 \dots \cdot 341}{365^{25}}$$

we do

$$\frac{\frac{365!}{340!}}{365^{25}} = 0.4313$$

Essentially the division of factorials here removes all the values < 341 , leaving 340, 339, 338 ... 1

Now remember this is the probability that no one shares a birthday. So when we subtract this from 1 we get

$$1 - 0.4313 = 0.5687$$

or roughly 57% odds of at least 2 people in a group of 25 sharing the same birthday.

4.2 Random variables

A random variable is a concept used to denote the outcome of an experiment before it is conducted.

4.2.1 Examples

Example 4.4. Let X denote the number of heads in a coin tossing experiment. We can then talk about the probabilities of certain events such as obtaining two heads, i.e. $X = 2$. We write this as

$$P[X = 2] = \binom{n}{2} p^2 (1-p)^{n-2}$$

In general:

$$P[X = x] = \binom{n}{x} p^x (1 - p)^{n-x}$$

where $x = 0, 1, \dots, n$

4.2.2 Handout

Definition 4.1. A **random variable**, X , is a function defined on a sample space, with outcomes in the set of real numbers.

It is simpler to think of a random variable as a symbol used to denote the outcome of an experiment before it is conducted.

Note 4.1. Note that it is **essential** to distinguish between upper case and lower case letters when writing these probabilities - it makes no sense to write $P[x = x]$.

Note 4.2. Random variables are generally denoted by upper case letters such as X , Y and so on.

Note 4.3. To see how a random variable is a function, it is useful to consider the actual outcomes of two coin tosses. These outcomes can be denoted $\{HH, HT, TH, TT\}$. Now consider a random variable X which describes the number of heads obtained. This random variable attributed 2 to the outcome HH and 0 to TT , i.e. X is a function with $X(HH) = 2$, $X(HT) = X(TH) = 1$ and $X(TT) = 0$.

4.3 Simple surveys with replacement

If we randomly draw individuals (with replacement) and ask a question with two possible answers (positive or negative), then the number of positive answers will come from a binomial distribution.

4.3.1 Examples

Example 4.5. Suppose we are participating in a lottery. We pick a number from a lottery bowl (a simple random sample). We can put the number aside, or we can put it back into the bowl. If we put the number back in the bowl, it may be selected more than once; if we put it aside, it can be selected only one time.

Definition 4.2. When an element can be selected more than one time, we are sampling **with replacement**.

Definition 4.3. When an element can be selected only one time, we are sampling **without replacement**.

4.4 The binomial distribution

If we toss a biased coin n independent times, each with probability p of landing heads up, then the probability of obtaining x heads is

$$\binom{n}{x} p^x (1-p)^{n-x}$$

4.4.1 Examples

Example 4.6. Suppose we toss a coin, with probability p of landing on heads n times obtaining a sequence of Hs (when it lands heads) and Ts (when it lands tails). Any sequence,

$$HTH\dots HTHHH$$

which has x heads (H) and $n-x$ tails (T), has the probability $p^x(1-p)^{n-x}$. There are exactly $\binom{n}{x}$ such sequences, so the total probability of landing x heads in n tosses is

$$\binom{n}{x} p^x (1-p)^{n-x}.$$

Example 4.7. Let the probability that a certain football club wins a match be equal to 0.4. If the total number of matches played in the season is 30, what is the probability that the football club wins the match 10% of the time?

We first calculate the number of times a match was played and won by multiplying the percentage of wins by the number of matches played.

10% of 30 times = 3 times

We can now proceed to calculate the probability that they will win the match given that their probability of a winning is 0.4 if they play 3 times in a season. This can be computed as follows:

$$\begin{aligned} \binom{30}{3} \times (0.4)^3 \times (1-0.4)^{30-3} \\ = 0.000265 \end{aligned}$$

This can be calculated in R using the code below:

```
dbinom(3,30,0.4)
```

```
[1] 0.0002659437
```

This is equal to the manual calculation using the binomial theorem.

Example 4.8. Suppose a youngster puts his shirt on by himself every day for five days. The probability that he puts it on the right way each time is $p = 0.2$. We let X be a random variable that describes the number of times the youngster puts his shirt on the right way. The youngster can either put the shirt on the wrong or the right way so X follows the binomial distribution with the parameters $p = 0.2$ (the probability of a successful trial) and $n = 5$ (number of trials). We can now calculate for example the probability that the youngster will put it on the right way for at least 4 days.

Putting the shirt on the right way for at least 4 days means that the youngster will either put it on the right way for either four or five days (at least four or more days of five days total). We thus have to calculate the probability that the youngster will put his shirt on the right way for 4 and 5 days separately and then we add it together. We can write this process as follows:

$$\begin{aligned}
 P(X \geq 4) &= P(X = 4) + P(X = 5) \\
 &= \binom{5}{4} \times 0.2^4 \times (1 - 0.2)^{5-4} + \binom{5}{5} \times 0.2^5 \times (1 - 0.2)^{5-5} \\
 &= 5 \times 0.2^4 \times 0.8^1 + 1 \times 0.2^5 \times 0.8^0 \\
 &= 5 \times 0.2^4 \times 0.8 + 0.2^5 \times 1 \\
 &= 5 \times 0.8 \times 0.2^4 + 0.2^5 \\
 &= 4 \times 0.2^4 + 0.2^5 \\
 &= 4 \times 0.0016 + 0.00032 \\
 &= 0.00672
 \end{aligned}$$

The probability that the youngster will put his shirt on the right way for at least four out of five is thus 0,7%.

This is possible to calculate in R in a several ways, either using the command `dbinom` or `pbinom`. The command `dbinom` calculates

$$P(X = k)$$

and the command `pbinom` calculates

$$P(X \leq k)$$

where k is the number of successful trials. If n is the number of trials and p is the probability of a successful trials then the commands are used by writing: `dbinom(k,n,p)` and `pbinom(k,n,p)`.

To calculate the probability that the youngster will put his shirt on the right way for at least four days of five we thus write the command:

```
dbinom(4,5,0.2) + dbinom(5,5,0.2)
```

which gives 0.00672.

This is the same as writing:

```
dbinom(c(4,5),5,0.2)
```

or

```
dbinom(4:5,5,0.2)
```

which give two separate numbers: 0.00640 and 0.00032 which can be added together to get 0.00672.

There is also a command to add them together for us:

```
sum(dbinom(c(4,5),5,0.2))
```

or

```
sum(dbinom(4:5,5,0.2))
```

They give the answer 0.00672.

The fourth way of calculating this in R is to use pbinom. As said before pbinom calculates

$$P(X \leq k)$$

where k is the number of successful trials. Here we want to calculate the probability that the youngster will put his shirt on the right way in 4 or 5 times (of 5 total) so the number of successful trials is 4 or greater. That means we want to calculate

$$P(X \geq 4)$$

which equals

$$1 - P(X \leq 3)$$

. We thus put k as 3 and the R command will be:

```
1 - pbinom(3,5,0.2)
```

which also gives 0.00672.

Example 4.9. In a certain degree program, the chance of passing an examination is 20%. What is the chance of passing at most 2 exams if the student takes five exams?

Solution:

In this problem, we compute the chance of a student passing, 0, 1 or 2 exams. This is given by,

$$\begin{aligned} p(X = 0 \text{ or } 1 \text{ or } 2) &= \binom{5}{0} 0.2^0 0.8^5 + \binom{5}{1} 0.2^1 0.8^4 + \binom{5}{2} 0.2^2 0.8^3 \\ &= 1 \times 0.2^0 0.8^5 + 5 \times 0.2^1 0.8^4 + 10 \times 0.2^2 0.8^3 \end{aligned}$$

$$= 0.32768 + 0.4096 + 0.2048$$

$$= 0.94208$$

In the R console, we can use the command, `sum(dbinom(c(0:2), 5, 0.2))`, which also gives

$$0.94208.$$

The same answer is obtained with

`dbinom(0, 5, 0.2) + dbinom(1, 5, 0.2) + dbinom(2, 5, 0.2)`

and with

`pbinom(2, 5, 0.2)`

Example 4.10. Consider the probability of someone jumping off a cliff is 0.35. Suppose we randomly selected four individuals to participate in the cliff jumping activity. What is the chance that exactly one of them will jump off the cliff?

Consider a scenario where one person jumps:

$P(A = \text{jump}, B = \text{refuse}, C = \text{refuse}, D = \text{refuse})$

$= P(A = \text{jump}) P(B = \text{refuse}) P(C = \text{refuse}) P(D = \text{refuse})$

$= (0.35)(0.65)(0.65)(0.65) = (0.35)^1(0.65)^3 = 0.096$

But there are three other scenarios (B, C, or D) in which one only person decides to jump. In each of these cases, the probability is again 0.096. These four scenarios exhaust all the possible ways that exactly one of the four people jumps:

$$4 \cdot (0.35)^1(0.65)^3 = 0.38.$$

In the R console we can use the command: `dbinom(1, 4, 0.35)` which gives the answer as 0.384475.

4.5 General discrete probability distributions

A general discrete probability distribution can be described by a list of all possible outcomes and associated probabilities.

4.5.1 Details

A general discrete probability distribution is described by the possible outcomes

$$x_1, x_2, \dots$$

and associated probabilities, denoted by p_1, p_2, \dots or $p(x_1), p(x_2), \dots$

If a random variable X has this distribution, then we can write

$$P[X = x_i] = p(x_i) = p_i$$

or in general

$$P[X = x] = p(x)$$

where it is understood that $p(x) = 0$ if x is not one of these x_i .

4.5.2 Examples

Example 4.11. If X is the number of heads (H) before obtaining the first tail (T) when tossing an unbiased coin 4 independent times, then the possible basic outcomes are:

In binary	Toss				# H before T
	1	2	3	4	
0000	H	H	H	H	4
0001	H	H	H	T	3
0010	H	H	T	H	2
0011	H	H	T	T	2
0100	H	T	H	H	1
0101	H	T	H	T	1
0110	H	T	T	H	1
0111	H	T	T	T	1
1000	T	H	H	H	0
1001	T	H	H	T	0
1010	T	H	T	H	0
1011	T	H	T	T	0
1100	T	T	H	H	0
1101	T	T	H	T	0
1110	T	T	T	H	0
1111	T	T	T	T	0

Since the coin is unbiased, each of these has the same probability of occurring. We can now count sequences to find the number of possibilities of a particular number of heads, H , before a tail in 4 coin tosses and thus obtain the corresponding probabilities as:

Number of tosses before a heads	Probability
x	$p(x)$
0	$\frac{8}{16} = \frac{1}{2}$
1	$\frac{4}{16} = \frac{1}{4}$
2	$\frac{2}{16} = \frac{1}{8}$
3	$\frac{1}{16}$
4	$\frac{1}{16}$

4.6 The expected value or population mean

The expected value is the sum of the possible outcomes, weighted with the respective probabilities (discrete variable). Think of this in terms of an urn full of marbles, each labelled with number.

4.6.1 Details

If the possible outcomes are x_1, x_2, \dots with probabilities p_1, p_2, \dots then the expected value is

$$\mu = x_1 \cdot p_1 + x_2 \cdot p_2 + \dots$$

The fact that this is the only sensible definition of an expected value follows from considering random draws from a finite population where there are n_i possibilities of obtaining the value x_i . If we set $n = \sum x_i$ and $p_i = n_i/n$ then the expected value above is the simple average of all the numbers in the original population.

In the case of the **binomial distribution** with n trials and success probability p it turns out that

$$\mu = n \cdot p$$

If X is the corresponding random variable, we denote this quantity by $E[X]$.

4.6.2 Examples

Example 4.12. If we toss a fair coin 10 independent times, we expect on average $np = 10 \cdot \frac{1}{2} = 5$ heads.

Example 4.13. Toss a fair die and pay \$60 if a six comes up and nothing otherwise. The expected outcome is

$$\frac{5}{6} \cdot \$0 + \frac{1}{6} \cdot \$60 = \$10.$$

Example 4.14. In Las Vegas, a particular sports bet has about a 30% chance of winning. If the bet wins, the bettor will win 15 dollars. If the bet loses, the bettor will lose 10 dollars. The expected return of placing one of these bets is -2.50 dollars.

Detailed calculation:

$$\$15 \cdot 0.3 - \$10 \cdot 0.7 = -\$2.5$$

Example 4.15. Class starts at 8:00 and the last bus that will get you to class on time leaves at 7:30. The teacher has a policy that if you are late to class 6 of the 30 classes, then she drops your final grade by 1/10 points. You know that if you set your alarm for 7:15, you miss the 7:30 bus approximately every fourth time, but if you set it for 7:10, you'll only miss the bus approximately every eighth time. If you set it for 7:00, you'll only miss the bus every one hundredth time.

Part A: Assuming you try to go to class every time, can you expect to have your grade dropped in the following scenarios?

1 - You set your alarm for 7:15 throughout the duration of the class.

2 - You set your alarm for 7:15 until you reach 5 missed classes, then switch to 7:10.

3 - You set your alarm for 7:15 until you reach 5 missed classes, then switch to 7:00.

Part B: What is your expected grade in the course, assuming you would have had a 7/10 without the late penalty, and:

1 - You would never choose the first alarm-clock strategy and you would most likely choose scenario 2 (let's say 9/10 times), but there's a small chance you might choose the 3rd strategy (let's say 1/10 times).

2 - You would never choose the first alarm-clock strategy and you would most likely choose scenario 3 (let's say 9/10 times), but there's a small chance you might choose the 2nd strategy (let's say 1/10 times).

Answers:

A1 - Let's call X our random variable, which we want to be the number of times we make it to class on-time. With the alarm set to 7:15 we expect to make it to class on-time:

$$E[X] = 30 \times \left(1 - \frac{1}{4}\right) = 22\frac{1}{2}$$

You're grade would most likely be dropped.

A2 - First we need to see how many classes we go to before we reach the 5-late-classes threshold:

$$E[X] = n \times \left(1 - \frac{1}{4}\right) = n - 5$$

$$E[X] = n \left(\left(1 - \frac{1}{4}\right) - 1\right) = -5$$

$$E[X] = n = \frac{-5}{-\frac{1}{4}}$$

$$E[X] = n = \frac{20}{1} = 20$$

So, the night before our 21st class, you get worried and change alarm-clock strategies. If you set it at 7:15 for the rest of the course (10 classes), you will be on time:

$$E[X] = 15 + 10 \times \left(1 - \frac{1}{8}\right) = 23\frac{3}{4}$$

You're grade would most likely be dropped.

A3: If you instead start setting the alarm clock for 7:00 for the rest of the course, you will be on time:

$$E[X] = 15 + 10 \times \left(1 - \frac{1}{100}\right) = 24\frac{1}{9}$$

You're grade would most likely NOT be dropped.

Part B: **This seems to contain errors** In Part A, we calculated the mean of several binomial distributions that described the expected number of days that you will arrive on-time to class. Each distribution corresponded to a different alarm-setting scenario. In this part, we are describing a different binomial distribution. It describes your expected grade. Therefore, the grade is the outcome n , weighted by the probability of you choosing the particular alarm-clock setting procedure:

$$1 - E[X] = 0 \times 6 + 0.9 \times 6 + 0.1 \times 7 = 6.1$$

$$1 - E[X] = 0 \times 6 + 0.1 \times 6 + 0.9 \times 7 = 6.9$$

Note that the probabilities of these three choices ($0 + 0.9 + 0.1$) must equal 1, since these are the only three choices defined.

4.7 The population variance

The (population) variance, for a discrete distribution, is

$$\sigma^2 = E[(X - \mu)^2] = (x_1 - \mu)^2 p_1 + (x_2 - \mu)^2 p_2 + \dots$$

where it is understood that the random variable X has this distribution and μ is the expected value.

In the case of the binomial distribution, it turns out that:

$$\sigma^2 = np(1 - p)$$

4.7.1 Details

Definition 4.4. If μ is the expected value, then the **variance of a discrete distribution** is defined as

$$\sigma^2 = (x_1 - \mu)^2 p_1 + (x_2 - \mu)^2 p_2 + \dots$$

If a random variable X has associated probabilities, $p_i = P[X = x_i]$, then one can equivalently write

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

4.7.2 Examples

Example 4.16. In the case of the binomial distribution, it turns out that:

$$\sigma^2 = np(1 - p).$$

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students
This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

5 Functions

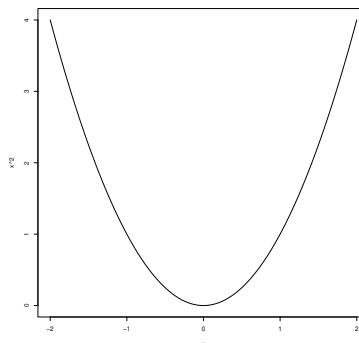
5.1 Functions of a single variable

A function describes the relationship between variables.

Examples:

$$f(x) = x^2$$

$$y = 2 + 3 \cdot x^4$$



5.1.1 Details

Functions are commonly used in statistical applications, to describe relationships.

Definition 5.1. A **function** describes the relationship between variables. A variable y is described as a function of a variable x by completely specifying how y can be computed for any given value of x .

An example could be the relationship between a dose level and the response to the dose.

The relationship is commonly expressed by writing either $f(x) = x^2$ or $y = x^2$.

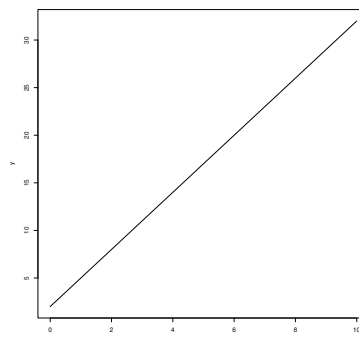
Usually names are given to functions, i.e. to the relationship itself. For example, f might be the function and $f(x)$ could be its value for a given number x . Typically $f(x)$ is a number but f is the function, but the sloppy phrase "the function $f(x) = 2x + 4$ " is also common.

5.1.2 Examples

Example 5.1. $f(x) = x^2$ or $y = x^2$ specifies that the computed value of y should always be x^2 , for any given value of x .

5.2 Functions in R

A function can be defined in R using the "function" command



5.3 Ranges and plots in R

Functions in R can commonly accept a range of values and will return a corresponding vector with the outcome.

5.3.1 Examples

```
ERROR: Unknown script MIME type text/x-tex
Traceback (most recent call last):
  File "/srv/sites/tutor-web-2/src/tutorweb.content/tutorweb/content/transforms/script
    raise ValueError("Unknown script MIME type %s" % kwargs['mimetype'])
ValueError: Unknown script MIME type text/x-tex
```

```
Example 5.2. f <- function(x) {return(x*12)}
x <- seq (-5,5,0,1)
y <- f(x)
plot {(x,y) type= 'l'}
```

5.4 Plotting functions

In statistics, the function of interest is commonly called the response function. If we write $Y=f(x)$, the outcome Y is usually called the response variable and x is the explanatory variable. Function values are plotted on vertical axis while x values are plotted on horizontal axis. This plots Y against x .

5.4.1 Examples

```
ERROR: Unknown script MIME type text/x-tex
Traceback (most recent call last):
  File "/srv/sites/tutor-web-2/src/tutorweb.content/tutorweb/content/transforms/script
    raise ValueError("Unknown script MIME type %s" % kwargs['mimetype'])
ValueError: Unknown script MIME type text/x-tex
```

```
Example 5.3. The following R commands can be used to generate a plot for function;
 $Y= 2+3x$ 
x<- seq(0:10)
g <- function(x){
+ yhat <- 2+3*x
+ return(yhat)
+ }
```

```
x<-seq(0,10,0.1)
y<- g(x)
plot(x,y,type="l", xlab="x",ylab="y")
```

5.5 Functions of several variables

5.5.1 Examples

Example 5.4.

$$z = 2x + 3y + 4 \quad (2)$$

$$v = t^2 + 3x \quad (3)$$

$$w = t^2 + 3b * x \quad (4)$$

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students
This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

6 Polynomials

6.1 The general polynomial

The general polynomial:

$$p(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n$$

The simplest: $p(x) = a$

6.1.1 Details

Definition 6.1. A **polynomial** describes a specific function consisting of linear combinations of positive integer powers of the explanatory variable.

The general form of a polynomial is:

$$p(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n$$

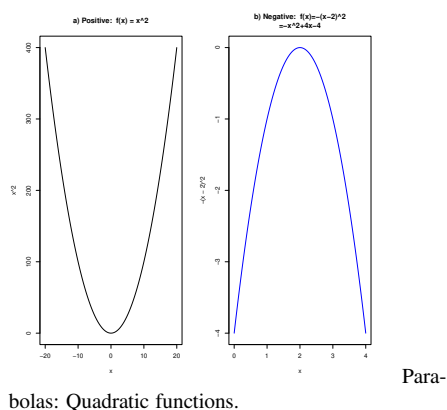
The simplest of these is the constant polynomial $p(x) = a$.

6.2 The quadratic

The general form of the quadratic (parabola) is

$$p(x) = ax^2 + bx + c.$$

The simplest quadratic is $p(x) = x^2$



6.2.1 Details

The quadratic polynomial of the form $p(x) = ax^2 + bx + c$ describes a parabola when points (x, y) with $y = p(x)$ are plotted.

The simplest parabola is $p(x) = x^2$ (Fig. a) which is always non-negative $p(x) \geq 0$ and $p(x) = 0$ only when $x = 0$.

Note 6.1. Note that $p(-x) = p(x)$ since $(-x)^2 = x^2$.

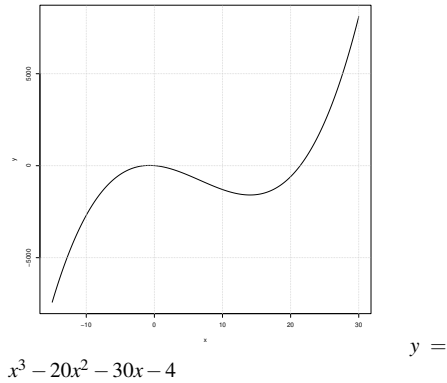
If the coefficient at the highest power is negative, then the parabola is "upside down"(Fig. b).

This is sometimes used to describe a response function.

6.3 The cubic

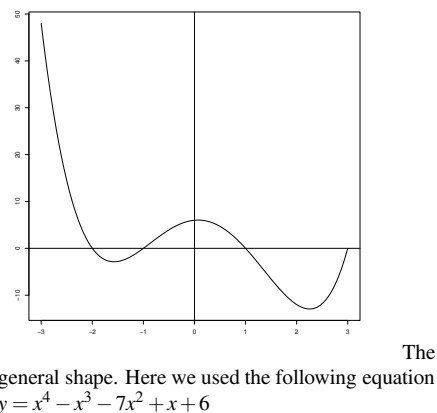
The general form of a cubic polynomial is:

$$p(x) = ax^3 + bx^2 + cx + d$$



6.4 The Quartic

The general form of the quartic polynomial is $p(x) = ax^4 + bx^3 + cx^2 + dx + e$



6.5 Solving the linear equation

If the value of y is given and we know that x and y are on a specific line so that $y = a + bx$, then we can find the value of x

6.5.1 Details

If a value of y is given and we know that x and y lie on a specific straight line so that $y = a + bx$, then we can find the value of x by considering $y = a + bx$ as an equation to be solved for x , since y , a and b are all known.

The general solution is found through the following steps:

- Equation: $y = a + bx$
- Subtract a from both sides
 - $y - a = bx$
 - $bx = y - a$
- Divide by b on both sides if b is not equal to 0.
 - $x = \frac{1}{b}(y - a)$.

6.6 Roots of the quadratic equation

The general solution of $ax^2 + bx + c = 0$ is given by $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$.

6.6.1 Details

Suppose we want to solve $ax^2 + bx + c = 0$, where $a \neq 0$.

The general solution is given by the formula

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

if $b^2 - 4ac \geq 0$. On the other hand, if $b^2 - 4ac < 0$, the quadratic equation has no real solution.

6.6.2 Examples

Example 6.1. Solve $x^2 - 3x + 2 = 0$

Putting this into the context of the formulation $ax^2 + bx + c = 0$, the constants are;
 $a = 1, b = -3, c = 2$

Inserting this into the formula for the roots gives:

$$x = \frac{-(-3) \pm \sqrt{(-3)^2 - 4(1)(2)}}{2(1)}$$

$$x = \frac{3 \pm \sqrt{9 - 8}}{2}$$

$$x = \frac{3 \pm \sqrt{1}}{2}$$

$$x = \frac{3 + 1}{2}, \frac{3 - 1}{2}$$

$$x = \frac{4}{2}, \frac{2}{2}$$

$$x = 2, 1$$

Example 6.2. Find the roots of the following polynomial

$$3x^4 + 14x^2 + 15$$

We can use the quadratic equation to solve for the roots of this polynomial if we substitute a variable for

$$x^2$$

Let's use the letter

$$a$$

$$3a^2 + 14a + 15$$

We then plug the constants in to the quadratic equation.

$$x = \frac{-(14) \pm \sqrt{14^2 - (4)(3)(15)}}{(2)(3)}$$

which simplifies to

$$\frac{-(14) \pm \sqrt{196 - 180}}{6}$$

which equals

$$-1\frac{2}{3}$$

and

$$-3$$

.

Then, since we substituted a for

$$x^2$$

we need to take the square root of these values to get the roots of the polynomial.

So,

$$x_{1,2} = \pm \sqrt{-1\frac{2}{3}}$$

and

$$x_{3,4} = \pm \sqrt{3}$$

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students
This work is licensed under the Creative Commons Attribution-ShareAlike License. To
view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a
letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

7 Simple data analysis in R

7.1 Entering data; dataframes

Several methods exist to enter data into R:

1. Enter directly: `x<-c(4,3,6,7,8)`
2. Read in a single vector: `x<-scan("filename")`
3. Use: `x<-read.table("file address")`

7.1.1 Details

The most direct method will not work if there are a lot numbers; therefore, the second method is to read in a single vector by `x<-scan("filename")`, "filename"- text string, either a full path name or refers to a file in the working directory.

The `scan()` command returns a vector, but the `read.table()` command returns a dataframe, which is a rectangular table of data whose columns have names. A column can be extracted from a data frame, e.g., with `x<- dat$a` where "dat" is the name of the data frame and "a" is the name of a column.

Note 7.1. Note that for `read.table("file address")`, "file address" refers to the location of the file. Thus, it can be the URL or the complete file directory depending on where the table is stored.

7.1.2 Examples

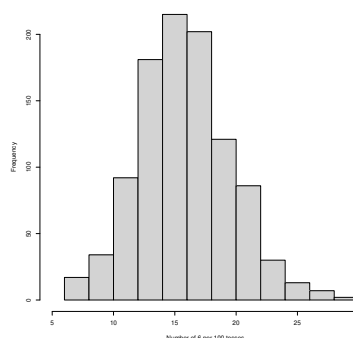
Example 7.1. Below are three examples using R code to enter data

1. `x<-c(4,3,6,7,8)`
2. `x<-scan("lecture 70.txt")`
3. `x<-read.table("http://notendur.hi.is/ gunnar/kennsla/alsm/data/set115.dat", header=T)`

7.2 Histograms

A histogram is a graphical display of tabulated frequencies, shown as bars.

In R use the command: `hist()`



7.2.1 Examples

A histogram is a graphical display of tabulated frequencies, shown as bars.

Example 7.2. If we toss a fair die 100 times and record the number of sixes, then we can view that as the outcome of a random variable X , which is binomial with $n = 100$ and $p = \frac{1}{6}$, i.e. $X \sim b(n = 100, p = \frac{1}{6})$

Now this can be done e.g. 1000 times to obtain numbers, x_1, \dots, x_{1000} . Within R this can be simulated using

```
x <- rbinom(1000,100,1/6)
```

We would typically plot these using a histogram, e.g.

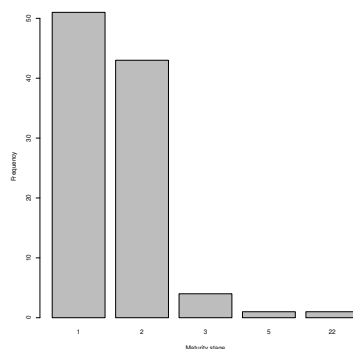
```
hist(x)
```

or

```
hist(x,nclass=50);l
```

7.3 Bar Charts

The bars in a bar chart usually correspond to frequencies in categories and are therefore kept apart.



7.3.1 Details

A bar chart is similar to the histogram but is used for categorical data.

7.4 Mean, standard error, standard deviations

7.4.1 Details

The most familiar measure of central tendency is the arithmetic mean.

Definition 7.1. An **arithmetic mean** is the sum of the values divided by the number values, typically expressed as:

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

Definition 7.2. The **sample variance** is a measure of the spread of a set of values from the mean value:

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

The sample standard deviation is more commonly used as a measure of the spread of a set of values from the mean value.

Definition 7.3. The **standard deviation** is the square root of the variance and may be expressed as:

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

Definition 7.4. The **standard error** is a method used to indicate the reliability of the sample mean:

$$SE_{\bar{y}} = \sqrt{\frac{s^2}{n}}$$

If a vector x in R contains an array of numbers then:

`mean(x)` returns the average, \bar{x}

`sd(x)` returns the standard deviation, s

`var(x)` returns the variance, s^2

We may also want to use several other related operations in R:

`median(x)`, the median value in vector x

`range(x)`, which list the range: `max(x) - \code{min(x)}`;

If the variable x contains discrete categories, `table(x)` returns counts of the frequency in each category.

7.5 Scatter plots and correlations

If we have paired explanatory and response data we are often interested in seeing if a relationship exists between them. To do this, we first plot the data in a scatter plot.

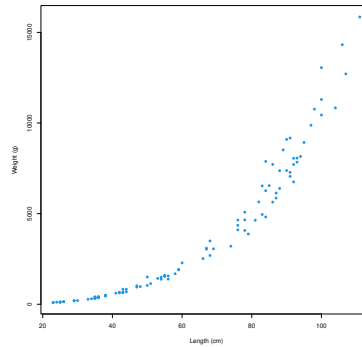


Figure: Scatter plot showing the length-weight relationship of fish species "X". Data source : Marine Resource Institution - Iceland.

7.5.1 Details

A first step in analyzing data is to prepare different plots. The type of variable will determine the type of plot. For example, when using a scatter plot both the explanatory and response data should be continuous variables.

The equation for the Pearson correlation coefficient is:

$$r_{x,y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

where \bar{x} and \bar{y} are the sample means of the x- and y-values.

The correlation is always between -1 and 1.

7.5.2 Examples

The following R commands can be used to generate a scatter plot for vectors x and y

Example 7.3. `plot(x, y)`

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students
This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

8 Indices and the apply commands in R

8.1 Giving names to elements

We can name elements of vectors and data frames in R using the "names" command.

8.1.1 Examples

```
Example 8.1. X<-c(41, 3, 73)
names(X)<-c("One", "Two", "Three")
```

View the results by simply typing "X" and the output of "X" is given as follows:

```
X
One Two Three
41 3 73
```

With this we can refer to the elements by name as well as locations using...

```
X[1]
One

X["Three"]
Three
73
```

8.2 Regular matrix indices and naming

A matrix is a table of numbers. Typical matrix indexing: `mat[i,j]`, `mat[1:2,]` etc

A matrix can have row and column names Indexing with row and column names:
`mat["a","B"]`

8.2.1 Details

Definition 8.1. A **matrix** is a (two-dimensional) table of numbers, indexed by row and column numbers.

Note 8.1. Note that a matrix can also have row and column names so that the matrix can be indexed by its names rather than numbers.

8.2.2 Examples

Example 8.2. Consider a matrix with 2 rows and 3 columns. Consider extracting first element (1,2), then all of line 2 and then columns 2-3 in an R session:

```
mat<-matrix(1:6,ncol=3)
mat
      [,1] [,2] [,3]
[1,]  1  3  5
[2,]  2  4  6
```

```
mat[1,2]
[1] 3
```

```
mat[2,]
[1] 2 4 6
```

```
mat[,2:3]
      [,1] [,2]
[1,]  3  5
[2,]  4  6
```

Next, consider the same matrix, but give names to the rows and columns. The rows will get the names "a" and "b" and the columns will be named "A", "B" and "C".

The entire R session could look like this:

```
mat<-matrix(1:6,ncol=3)
dimnames(mat)<-list(c("a","b"),c("A","B","C"))
mat
  A B C
a 1 3 5
b 2 4 6

mat["b",c("B","C")]
B C
4 6
```

8.3 The apply command

The apply command...

`apply(mat,2,sum)` – applies the sum function within each column

`apply(mat,1,mean)` – computes the mean within each row

8.4 The tapply command

Commonly one has a data vector and another vector of the same length giving categories for the measurements. In this case one often wants to compute the mean or variance (or median etc) within each category. To do this we use the tapply command in R.

8.4.1 Examples

Example 8.3. `z<-c(5,7,2,9,3,4,8)`
`i<-c("m","f","m","m","f","m","f")`

A. Find the sum within each group

```
tapply(z, i, sum)
  f m
18 20
```

B. Find the sample sizes

```
tapply(z, i, length)
  f m
 3 4
```

C. Store outputs and use names

```
n<-tapply(z, i, length)
n
  f m
 3 4
n["m"]
m
4
```

8.5 Logical indexing

A logical vector consists of *TRUE* (1) or *FALSE* (0) values. These can be used to index vectors or matrices.

8.5.1 Examples

```
Example 8.4. i<-c("m", "f", "m", "m", "f", "m", "f")
z<-c(5, 7, 2, 9, 3, 4, 8)

i=="m"
[1] TRUE FALSE TRUE TRUE FALSE TRUE FALSE

z[i=="m"]
[1] 5 2 9 4

z[c(T, F, T, T, F, T, F)]
[1] 5 2 9 4
```

8.6 Lists, indexing lists

A list is a collection of objects. Thus, data frames are lists.

8.6.1 Examples


```
Example 8.5. x<-list(y=2,z=c(2,3),w=c("a","b","c"))
x[["z"]]
[1] 2 3
names(x)
[1] "y" "z" "w"
x["w"]
$w
[1] "a" "b" "c"
x$w
[1] "a" "b" "c"
```

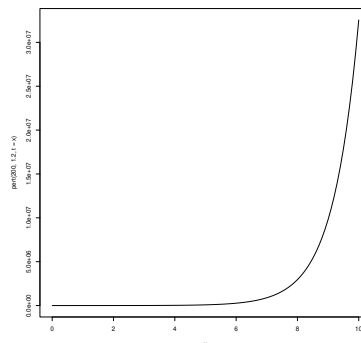
Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students
This work is licensed under the Creative Commons Attribution-ShareAlike License. To
view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a
letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

9 Functions of functions and the exponential function

9.1 Exponential growth and decline

Exponential growth is typically expressed as:

$$y(t) = Ae^{kt}$$



Exponential growth curve

9.1.1 Details

Definition 9.1. Exponential growth is the rate of population increase across time when a population is devoid of limiting factors (i.e. competition, resources, etc.) and experiences a constant growth rate.

Exponential growth is typically expressed as:

$$y(t) = Ae^{kt}$$

where

A (sometimes denoted P)=initial population size

k = growth rate

t =number of time intervals

Note 9.1. Note that exponential growth occurs when $k > 0$ and exponential decline occurs when $k < 0$.

9.2 The exponential function

An exponential function is a function with the form: $f(x) = b^x$

9.2.1 Details

For the exponential function $f(x) = b^x$, x is a positive integer and b is a fixed positive real number. The equation can be rewritten as:

$$f(x) = b^x = b \cdot b \cdot b \dots b$$

When the exponential function is written as $f(x) = e^x$ then, it has a growth rate at time x equivalent to the value of e^x for the function at x .

9.3 Properties of the exponential function

Recall that the methods of the basic arithmetic implies that:

$$e^{a+b} = e^a e^b$$

for any real numbers a and b .

9.4 Functions of functions

9.4.1 Details

Consider two functions, f and g , each defined for some set of real numbers. Where x can be solved in function f using $Y = f(x)$ when $g(Y)$ exists for all such resulting Y . If $Y = f(x)$ and $g(Y)$ exist then we can compute $g(f(x))$ for any x .

If

$$f(x) = x^2 \text{ and}$$

$$g(y) = e^y \text{ then}$$

$$g(f(x)) = e^{f(x)} = e^{x^2}$$

If we call the resulting function h ;

$$h(x) = g(f(x))$$

Then h is commonly written as

$$h = g \circ f$$

9.4.2 Examples

Example 9.1. If

$$g(x) = 3 + 2x \text{ and}$$

$$f(x) = 5x^2$$

Then

$$g(f(x)) = 3 + 2f(x)$$

$$g(f(x)) = 3 + 10x^2$$

$$f(g(x)) = 5(g(x))^2$$

$$f(g(x)) = 5(3 + 2x)^2$$

$$f(g(x)) = 45 + 60x + 20x^2$$

9.5 Storing and using R code

As R code gets more complex (more lines) it is usually stored in files. Functions are typically stored in separate files.

9.5.1 Examples

Example 9.2. Save the following file (test.r):

```
x=4  
y=8  
cat("x+y is", x+y, "\n")$
```

To read the file use:

```
source("test.r")
```

and the outcome of the equation is displayed in R

9.6 Storing and calling functions in R

To save a function in a separate file use a command of the form "function.r".

9.6.1 Examples

```
Example 9.3. f<-function(x) {  
  return (exp(sum(x)))  
}
```

can be stored in a file function.r and subsequently read using the source command.

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students
This work is licensed under the Creative Commons Attribution-ShareAlike License. To
view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a
letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

10 Inverse functions and the logarithm

10.1 Inverse Function

If f is a function, then the function g is the inverse function of f if

$$g(f(x)) = x$$

for all x in which $f(x)$ can be calculated

10.1.1 Details

The inverse of a function f is denoted by f^{-1} , i.e.

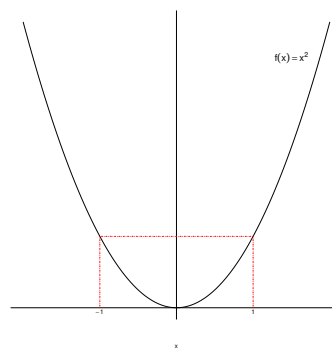
$$f^{-1}(f(x)) = x$$

10.1.2 Examples

Example 10.1. If $f(x) = x^2$ for $x < 0$ then the function g , defined as $g(y) = \sqrt{y}$ for $y > 0$, is not the inverse of f since $g(f(x)) = \sqrt{x^2} = |x| = -x$ for $x < 0$.

10.2 When the inverse exists: The domain question

Inverses do not always exist. For an inverse of f to exist, f must be one-to-one, i.e. for each x , $f(x)$ must be unique.



The function $f(x) = x^2$ does not have an inverse since $f(x)=1$ has two possible solutions -1 and 1.

10.2.1 Examples

Example 10.2. $f(x) = x^2$ does not have an inverse since $f(x) = 1$ has two possible solutions -1 and 1.

Note 10.1. Note that iff f is a function, then the function g is the inverse function of f , if $g(f(x)) = x$ for all calculated values of x in $f(x)$.

The inverse function of f is denoted by f^{-1} , i.e. $f^{-1}(f(x)) = x$.

Example 10.3. What is the inverse function, f^{-1} , of f if $f(x) = 5 + 4x$.

The simplest approach is to write $y = f(x)$ and solve for x :

With

$$f(x) = 5 + 4x$$

we write

$$y = 5 + 4x$$

which we can now rewrite as

$$y - 5 = 4x$$

and this implies

$$\frac{y - 5}{4} = x$$

And there we have it, very simple:

$$f^{-1}(f(x)) = \frac{y - 5}{4}$$

10.3 The base 10 logarithm

When x is a positive real number in $x = 10^y$, y is referred to as the base 10 logarithm of x and is written as:

$$y = \log_{10}(x)$$

or

$$y = \log(x)$$

10.3.1 Details

If $\log(x) = a$ and $\log(y) = b$, then $x = 10^a$ and $y = 10^b$, and

$$x \cdot y = 10^a \cdot 10^b = 10^{a+b}$$

so that

$$\log(xy) = a + b$$

10.3.2 Examples

Example 10.4.

$$\log(100) = 2$$

$$\log(1000) = 3$$

Example 10.5. If

$$\log(2) \approx 0.3$$

then

$$10^y = 2$$

Note 10.2. Note that

$$2^{10} = 1024 \approx 1000 = 10^3$$

therefore

$$2 \approx 10^{3/10}$$

so

$$\log(2) \approx 0.3$$

10.4 The natural logarithm

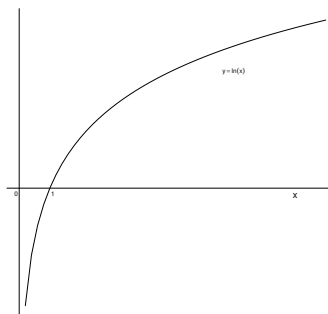
A logarithm with e as a base is referred to as the natural logarithm and is denoted as \ln :

$$y = \ln(x)$$

if

$$x = e^y = \exp(y)$$

Note that \ln is the inverse of \exp .



The curve depicts the function $y = \ln(x)$ and shows that \ln is the inverse of \exp . Note that $\ln(1) = 0$ and when $y = 0$ then $e^0 = 1$.

10.5 Properties of logarithm(s)

Logarithms transform multiplicative models into additive models, i.e.

$$\ln(a \cdot b) = \ln a + \ln b$$

10.5.1 Details

This implies that any statistical model, which is multiplicative becomes additive on a log scale, e.g.

$$y = a \cdot w^b \cdot x^c$$
$$\ln y = (\ln a) + \ln(w^b) + \ln(x^c)$$

Next, note that

$$\begin{aligned}\ln(x^2) &= \ln(x \cdot x) \\ &= \ln x + \ln x \\ &= 2 \cdot \ln x\end{aligned}$$

and similarly $\ln(x^n) = n \cdot \ln x$ for any integer n .

In general $\ln(x^c) = c \cdot \ln x$ for any real number c (for $x > 0$).

Thus the multiplicative model (from above)

$$y = a \cdot w^b \cdot x^c$$

becomes

$$y = (\ln a) + b \cdot \ln w + c \cdot \ln x$$

which is a linear model with parameters $(\ln a)$, b and c .

In addition, the log-transform is often variance-stabilizing.

10.6 The exponential function and the logarithm

The exponential function and the logarithms are inverses of each other

$$x = e^y \Leftrightarrow y = \ln x$$

10.6.1 Details

Note 10.3. Note the properties:

$$\ln(x \cdot y) = \ln(x) + \ln(y)$$

and

$$e^a \cdot e^b = e^{a+b}$$

10.6.2 Examples

Example 10.6. Solve the equation

$$10e^{1/3x} + 3 = 24$$

for x .

First, get the 3 out of the way.

$$10e^{1/3x} = 21$$

Then the 10.

$$e^{1/3x} = 2.1$$

Next, we can take the natural log of 2.1. Since \ln is an inverse function of e this would result in

$$\frac{1}{3}x = \ln(2.1)$$

This yields

$$x = \ln(2.1) \cdot 3$$

which is

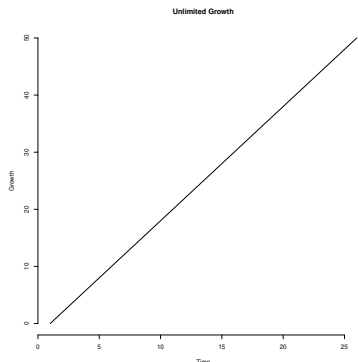
$$\approx 2.23$$

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students
This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

11 Continuity and limits

11.1 The concept of continuity

A function is continuous if it has no jumps. Thus, small changes in each x_0 , the input, correspond to small changes in the output, $f(x_0)$.



The above figure is an example of linear growth. Thomas Robert Malthus (1766-1834) warned about the dangers of uninhibited population growth.

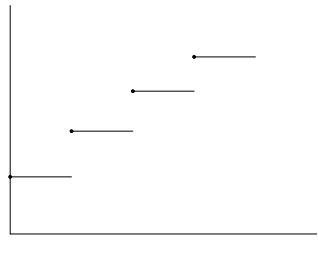
11.1.1 Details

A function is said to be discontinuous if it has jumps. The function is continuous if it has no jumps. Thus, for a continuous function, small changes in each x_0 , the input, correspond to small changes in the output, $f(x_0)$.

Note 11.1. Note that polynomials are continuous as are logarithms (for positive numbers).

11.2 Discrete probabilities and cumulative distribution functions

The cumulative distribution function for a discrete random variable is discontinuous.



11.2.1 Details

Definition 11.1. If X is a random variable with a discrete probability distribution and the probability mass function of

$$p(x) = P[X = x]$$

then the **cumulative distribution function**, defined by

$$F(X) = P[X \leq x]$$

is discontinuous, i.e. it jumps at points in which a positive probability occurs.

Note 11.2. When drawing discontinuous functions it is common practice to use a filled circle at $(x, f(x))$ to clarify what the function value is at a point x of discontinuity.

11.2.2 Examples

Example 11.1. If a coin is tossed 3 independent times and X denotes the number of heads, then X can only take on the values 0, 1, 2 and 3. The probability of landing exactly x heads, $P(X = x)$, is $p(x) = \binom{n}{x} p^n (1 - p)^{n-x}$. The probabilities are

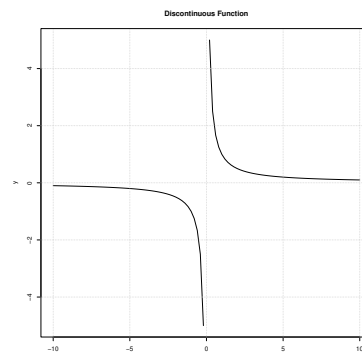
x	$p(x)$	$F(x)$
0	1/8	1/8
1	3/8	4/8
2	3/8	7/8
3	1/8	1

The cumulative distribution function, $F(x) = P[X \leq x] = \sum_{t \leq x} p(t)$ has jumps and is therefore discontinuous.

Note 11.3. Notice on the above figure how the circles are filled in, the solid circles indicate where the function value is.

11.3 Notes on discontinuous function

A function is discontinuous for values or ranges of the variable that do not vary continuously as the variable increases. In other words, breaks or jumps.



$$f(x) = \frac{1}{x}, \text{ where } x \neq 0$$

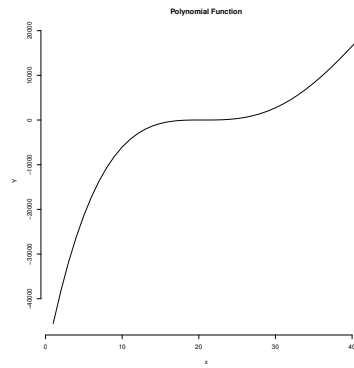
11.3.1 Details

A function can be discontinuous in a number of different ways. Most commonly, it may jump at certain points or increase without bound in certain places.

Consider the function f , defined by $f(x) = 1/x$ when $x \neq 0$. Naturally, $1/x$ is not defined for $x = 0$. This function increases towards $+\infty$ as x goes to zero from the right but decreases to $-\infty$ as x goes to zero from the left. Since the function does not have the same limit from the right and the left, it can not be made continuous at $x = 0$ even if one tries to define $f(0)$ as some number.

11.4 Continuity of polynomials

All polynomials, $p(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n$, are continuous.



11.4.1 Details

It is easy to show that simple polynomials such as $p(x) = x$, $p(x) = a + bx$, $p(x) = ax^2 + bx + c$ are continuous functions.

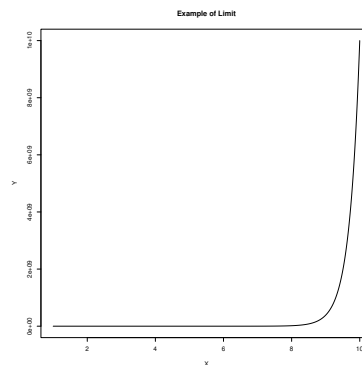
It is generally true that a polynomial of the form

$$p(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n$$

is a continuous function.

11.5 Simple Limits

A "limit" is used to describe the value that a function or sequence "approaches" as the input or index approaches some value. Limits are used to define continuity, derivatives and integrals.



$$f(x) = x^x, \text{ for } x > 0$$

11.5.1 Details

Definition 11.2. A **limit** describes the value that a function or sequence approaches as the input or index approaches some value.

Limits are essential to calculus (and mathematical analysis in general) and are used to define continuity, derivatives and integrals.

Consider a function and a point x_0 . If $f(x)$ gets steadily closer to some number c as x gets closer to a number x_0 , then c is called the limit of $f(x)$ as x goes to x_0 and is written as:

$$c = \lim_{x \rightarrow x_0} f(x)$$

If $c = f(x_0)$ then f is **continuous** at x_0 .

11.5.2 Examples

Example 11.2. A simple example of limits:

Evaluate the limit of $f(x) = \frac{x^2-16}{x-4}$ when $x \rightarrow 4$, or

$$\lim_{x \rightarrow 4} \frac{x^2 - 16}{x - 4}.$$

Notice that in principle we can not simply stick in the value $x = 4$ since we would then get $0/0$ which is not defined. However we can look at the numerator and try to factor it. This gives us:

$$\frac{x^2 - 16}{x - 4} = \frac{(x - 4)(x + 4)}{x - 4} = x + 4$$

and the result has the obvious limit of $4 + 4 = 8$ as $x \rightarrow 4$.

Example 11.3. Consider the function

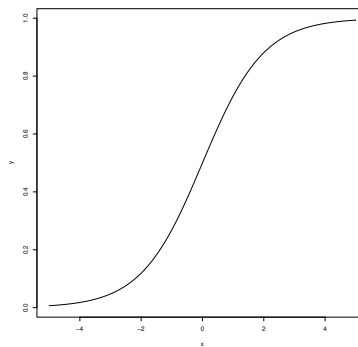
$$g(x) = \frac{1}{x}$$

where x is a positive real number. As x increases, $g(x)$ decreases, approaching 0 but never getting there since $\frac{1}{x} = 0$ has no solution. One can therefore say, “The limit of $g(x)$, as x approaches infinity, is 0,” and write

$$\lim_{x \rightarrow \infty} g(x) = 0.$$

11.6 More on limits

Limits impose a certain range of values that may be applied to the function.



The function $f(x) = \frac{1}{1+e^{-x}}$.

The

11.6.1 Examples

Example 11.4. The Beverton-Holt stock recruitment curve is given by:

$$R = \frac{\alpha S}{1 + \frac{S}{K}}$$

where $\alpha, K > 0$ are constants and $S =$ biomass and $R =$ recruitment.

The behavior of this curve as S increases $S \rightarrow \infty$ is

$$\lim_{S \rightarrow \infty} \frac{\alpha S}{1 + \frac{S}{K}} = \alpha K.$$

This is seen by rewriting the formula as follows:

$$\lim_{S \rightarrow \infty} \frac{\alpha S}{1 + \frac{S}{K}} = \lim_{S \rightarrow \infty} \frac{\alpha}{\frac{1}{S} + \frac{1}{K}} = \alpha K.$$

Example 11.5. A popular model for proportions is:

$$f(x) = \frac{1}{1 + e^{-x}}$$

As x increases, e^{-x} decreases which implies that the term $1 + e^{-x}$ decreases and hence $\frac{1}{1 + e^{-x}}$ increases, from which it follows that f is an increasing function.

Notice that $f(0) = \frac{1}{2}$ and further,

$$\lim_{x \rightarrow \infty} f(x) = 1.$$

This is seen from considering the components:

Since $e^{-x} = \frac{1}{e^x}$ and the exponential function goes to infinity as $x \rightarrow \infty$, e^{-x} goes to 0 and hence $f(x)$ goes to 1.

Through a similar analysis one finds that

$$\lim_{x \rightarrow -\infty} f(x) = 0,$$

since, as $x \rightarrow \infty$, first $-x \rightarrow \infty$ and second $e^{-x} \rightarrow \infty$.

Example 11.6. Evaluate the limit of

$$f(x) = \frac{\sqrt{x+4} - 2}{x}$$

as

$$x \rightarrow 0$$

$$\lim_{x \rightarrow 0} \frac{\sqrt{x+4} - 2}{x}$$

Since the square root is present we cannot just directly substitute the 0 as x . This will give us $\frac{0}{0}$, which is an indeterminate form. We must perform some algebra first. The way to get rid of the radical is to multiply the numerator by the conjugate.

$$\frac{\sqrt{x+4}-2}{x} \cdot \frac{\sqrt{x+4}+2}{\sqrt{x+4}+2}$$

This gives us

$$\frac{(\sqrt{x+4})^2 + 2(\sqrt{x+4}) - 2(\sqrt{x+4}) - 4}{x(\sqrt{x+4}+2)}$$

The numerator reduces to x , and the x s will cancel out leaving us with

$$\frac{1}{\sqrt{x+4}+2}$$

At this point we can direct substitute 0 for x , which will give us

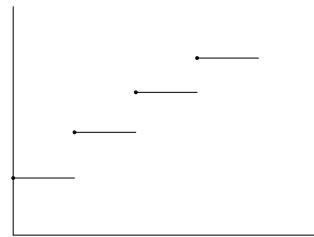
$$\frac{1}{\sqrt{0+4}+2}$$

Therefore,

$$\lim_{x \rightarrow 0} \frac{\sqrt{x+4}-2}{x} = \frac{1}{4}$$

11.7 One-sided limits

$f(x)$ may tend towards different numbers depending on whether $x \rightarrow x_0$:
 from the right ($x \rightarrow x_{0+}$)
 or from the left ($x \rightarrow x_{0-}$).



11.7.1 Details

Sometimes a function is such that $f(x)$ tends to different numbers depending on whether $x \rightarrow x_0$ from the right ($x \rightarrow x_{0+}$) or from the left ($x \rightarrow x_{0-}$).

If

$$\lim_{x \rightarrow x_{0+}} f(x) = f(x_0)$$

then we say that f is continuous from the right at x_0 .

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students
 This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

12 Sequences and series

12.1 Sequences

A **sequence** is a string of indexed numbers a_1, a_2, a_3, \dots . We denote this sequence with $(a_n)_{n \geq 1}$.

12.1.1 Details

In a sequence the same number can appear several times in different places.

12.1.2 Examples

Example 12.1. $(\frac{1}{n})_{n \geq 1}$ is the sequence $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots$

Example 12.2. $(n)_{n \geq 1}$ is the sequence $1, 2, 3, 4, 5, \dots$

Example 12.3. $(2^n n)_{n \geq 1}$ is the sequence $2, 8, 24, 64, \dots$

12.2 Convergent sequences

A sequence a_n is said to **converge** to the number b if for every $\varepsilon > 0$ we can find an $N \in \mathbb{N}$ such that $|a_n - b| < \varepsilon$ for all $n \geq N$. We denote this with $\lim_{n \rightarrow \infty} a_n = b$ or $a_n \rightarrow b$, as $n \rightarrow \infty$.

12.2.1 Details

A sequence a_n is said to **converge** to the number b if for every $\varepsilon > 0$ we can find an $N \in \mathbb{N}$ such that $|a_n - b| < \varepsilon$ for all $n \geq N$. We denote this with $\lim_{n \rightarrow \infty} a_n = b$ or $a_n \rightarrow b$, as $n \rightarrow \infty$. If x is a number then,

$$(1 + \frac{x}{n})^n \rightarrow e^x \text{ as } n \rightarrow \infty$$

12.2.2 Examples

Example 12.4. The sequence $(\frac{1}{n})_{n \geq \infty}$ converges to 0 as $n \rightarrow \infty$

Example 12.5. If x is a number then,
 $(1 + \frac{x}{n})^n \rightarrow e^x$ as $n \rightarrow \infty$

12.3 Infinite sums (series)

We are interested in, whether infinite sums of sequences can be defined.

12.3.1 Details

Consider a sequence of numbers, $(a_n)_{n \rightarrow \infty}$.

Now define another sequence $(s_n)_{n \rightarrow \infty}$, where

$$s_n = \sum_{k=1}^n a_k.$$

If $(s_n)_{n \rightarrow \infty}$ is convergent to $S = \lim_{n \rightarrow \infty} s_n$, then we write

$$S = \sum_{n=1}^{\infty} a_n.$$

12.3.2 Examples

Example 12.6. If

$$a_k = x^k, k = 0, 1, \dots$$

then

$$s_n = \sum_{k=0}^n x^k = x^0 + x^1 + \dots + x^n$$

Note also that

$$xs_n = x(x^0 + x^1 + \dots + x^n) = x + x^2 + \dots + x^{n+1}$$

We have

$$\begin{aligned} s_n &= 1 + x + x^2 + \dots + x^n \\ xs_n &= x + x^2 + \dots + x^n + x^{n+1} \\ s_n - xs_n &= 1 - x^{n+1} \end{aligned}$$

i.e.

$$s_n(1 - x) = 1 - x^{n+1}$$

and we have

$$s_n = \frac{1 - x^{n+1}}{1 - x}$$

if $x \neq 1$. If $0 < x < 1$ then $x^{n+1} \rightarrow 0$ as $n \rightarrow \infty$ and we obtain $s_n \rightarrow \frac{1}{1-x}$ so $\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$.

12.4 The exponential function and the Poisson distribution

The exponential function can be written as a series (infinite sum):

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}.$$

The Poisson distribution is defined by the probabilities

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!} \text{ for } x = 0, 1, 2, \dots$$

12.4.1 Details

The exponential function can be written as a series (infinite sum):

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}.$$

Knowing this we can see why the Poisson probabilities

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

add to one:

$$\sum_{x=0}^{\infty} p(x) = \sum_{x=0}^{\infty} e^{-\lambda} \frac{\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1.$$

12.5 Relation to expected values

The expected value for the Poisson is given by

$$\begin{aligned} \sum_{x=0}^{\infty} xp(x) &= \sum_{x=0}^{\infty} xe^{-\lambda} \frac{\lambda^x}{x!} \\ &= \lambda \end{aligned}$$

12.5.1 Details

The expected value for the Poisson is given by

$$\begin{aligned} \sum_{x=0}^{\infty} xp(x) &= \sum_{x=0}^{\infty} xe^{-\lambda} \frac{\lambda^x}{x!} \\ &= e^{-\lambda} \sum_{x=1}^{\infty} \frac{x\lambda^x}{x!} \\ &= e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!} \\ &= e^{-\lambda} \lambda \sum_{x=1}^{\infty} \frac{\lambda^{(x-1)}}{(x-1)!} \\ &= e^{-\lambda} \lambda \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \\ &= e^{-\lambda} \lambda e^{\lambda} \\ &= \lambda \end{aligned}$$

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students
This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

13 Slopes of lines and curves

13.1 The slope of a line

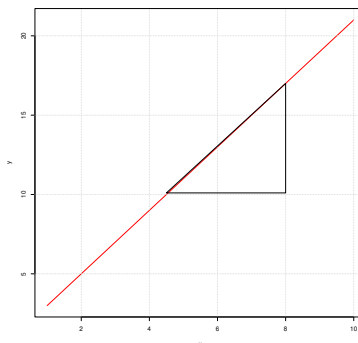
Linear functions produce straight-line graphs. In general, a straight line follows the following equation:

$$y = a + bx,$$

where a and b are fixed numbers.

The line on the graph is the set of points:

$$\{(x, y) : x, y \in \mathbb{R}, y = a + bx\}.$$



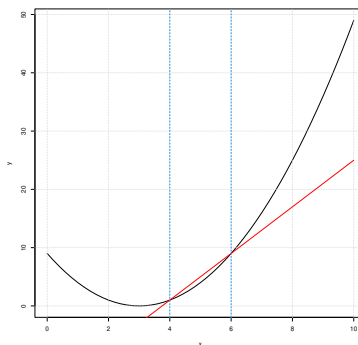
13.1.1 Details

The slope of a straight line represents the change in the y coordinate corresponding to a unit change in the x coordinate.

13.2 Segment slopes

Let's assume we have a more general function $y = f(x)$.

To find the slope of a line segment, consider 2 x -coordinates, x_0 and x_1 , and look at the slope between $(x_0, f(x_0))$ and $(x_1, f(x_1))$.



13.2.1 Details

Consider two points, (x_0, y_0) and (x_1, y_1) . The slope of the straight line that goes through these points is

$$\frac{y_1 - y_0}{x_1 - x_0}.$$

Thus, the slope of a line segment passing through the points $(x_0, f(x_0))$ and $(x_1, f(x_1))$, for some function, f , is

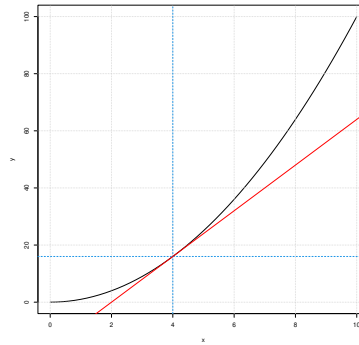
$$\frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

If we let $x_1 = x_0 + h$ then the slope of the segment is

$$\frac{f(x_0 + h) - f(x_0)}{h}.$$

13.3 The slope of $y = x^2$

Consider the task of computing the slope of the function $y = x^2$ at a given point.



13.3.1 Examples

Consider the function $y = f(x) = x^2$.

In order to find the slope at a given point (x_0) , we look at

$$y = \frac{f(x_0 + h) - f(x_0)}{h}$$

for small values of h .

For this particular function, $f(x) = x^2$, and hence

$$f(x_0 + h) = (x_0 + h)^2 = x^2 + 2hx_0 + h^2.$$

The slope of a line segment is therefore given by

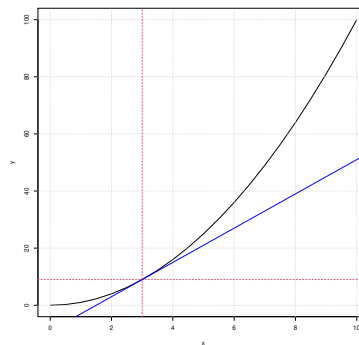
$$\frac{f(x_0 + h) - f(x_0)}{h} = \frac{2hx_0 + h^2}{h} = 2x_0 + h.$$

As we make h steadily smaller, the segment slope, $2x_0 + h$, tends towards $2x_0$. It follows that the slope, y' , of the curve at a general point x is given by $y' = 2x$.

13.4 The tangent to a curve

A **tangent** to a curve is a line that intersects the curve at exactly one point. The slope of a tangent for the function $y = f(x)$ at the point $(x_0, f(x_0))$ is

$$\lim_{h \rightarrow 0} \frac{f(x_0 + h) - f(x_0)}{h}.$$



13.4.1 Details

To find the slope of the tangent to a curve at a point, we look at the slope of a line segment between the points $(x_0, f(x_0))$ and $(x_0 + h, f(x_0 + h))$, which is

$$\frac{f(x_0 + h) - f(x_0)}{h}$$

and then we take h to be closer and closer to 0. Thus the slope is

$$\lim_{h \rightarrow 0} \frac{f(x_0 + h) - f(x_0)}{h}$$

when this limit exists.

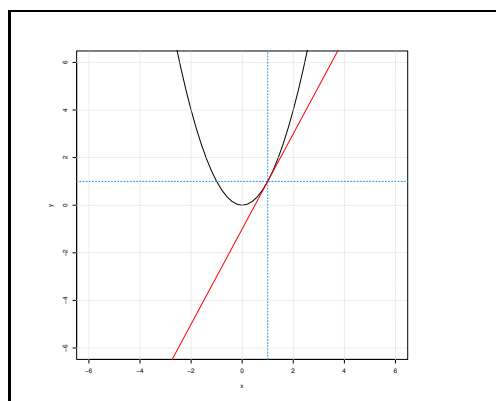
13.4.2 Examples

Example 13.1. We wish to find tangent line for the function $f(x) = x^2$ at the point $(1, 1)$. First we need to find the slope of this tangent, it is given as

$$\lim_{h \rightarrow 0} \frac{(1+h)^2 - 1^2}{h} = \lim_{h \rightarrow 0} \frac{2h + h^2}{h} = \lim_{h \rightarrow 0} (2 + h) = 2.$$

Then, since we know the tangent goes through the point $(1, 1)$ the line is $y = 2x - 1$.

13.5 The slope of a general curve



13.5.1 Details

Imagine a nonlinear function whose graph is a curve described by the equation, $y = f(x)$.

Here we want to find the slope of a line tangent to the curve at a specific point (x_0) .

The slope of the line segment is given by the equation $\frac{f(x_0+h) - f(x_0)}{h}$.

Reducing h towards zero, gives the slope of this curve if it exists.

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students
This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

14 Derivatives

14.1 The derivative as a limit

The derivative of the function f at the point x is defined as

$$\lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

if this limit exists.

14.1.1 Details

Definition 14.1. The derivative of the function f at the point x is defined as

$$\lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

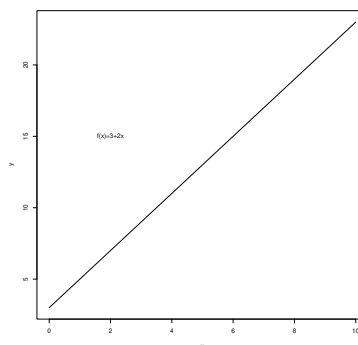
if this limit exists.

When we write $y = f(x)$, we commonly use the notation $\frac{dy}{dx}$ or $f'(x)$ for this limit.

14.2 The derivative of $f(x) = a + bx$

If $f(x) = a + bx$ then $f(x+h) = a + b(x+h) = a + bx + bh$ and thus

$$\lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} = \lim_{h \rightarrow 0} \frac{bh}{h} = b$$



14.2.1 Details

If $f(x) = a + bx$ then $f(x+h) = a + b(x+h) = a + bx + bh$ and thus

$$\lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} = \lim_{h \rightarrow 0} \frac{bh}{h} = b.$$

Thus $f'(x) = b$.

14.3 The derivative of $f(x) = x^n$

If $f(x) = x^n$, then $f'(x) = nx^{n-1}$.

14.3.1 Details

Let $f(x) = x^n$, where n is a positive integer. To calculate f' we use the binomial theorem in the third step:

$$\begin{aligned}\frac{f(x+h) - f(x)}{h} &= \frac{(x+h)^n - x^n}{h} \\ &= \frac{\sum_{q=0}^{n-1} \binom{n}{q} x^q h^{n-q}}{h} \\ &= \sum_{q=0}^{n-1} \binom{n}{q} x^q h^{n-q-1} \rightarrow \binom{n}{n-1} x^{n-1} = nx^{n-1}\end{aligned}$$

Thus, we obtain $f'(x) = nx^{n-1}$.

14.4 The derivative of ln and exp

If	$f(x) = e^x$
then	$f'(x) = e^x$
If	$g(x) = \ln(x)$
then	$g'(x) = \frac{1}{x}$

14.4.1 Details

The derivatives of the exponential function is the exponential function itself i.e. if

$$f(x) = e^x$$

then

$$f'(x) = e^x$$

The derivatives of the natural logarithm, $\ln(x)$, is $\frac{1}{x}$, i.e. if

$$g(x) = \ln(x)$$

then

$$g'(x) = \frac{1}{x}$$

14.5 The derivative of a sum and linear combination

If f and g are functions then the derivative of $f + g$ is given by $f' + g'$.

14.5.1 Details

Similarly, the derivative of a linear combination is the linear combination of the derivatives. If f and g are functions and $k(x) = af(x) + bg(x)$ then $k'(x) = af'(x) + bg'(x)$.

14.5.2 Examples

Example 14.1. If $f(x) = 2 + 3x$ and $g(x) = x^3$ then we know that $f'(x) = 3$, $g'(x) = 3x^2$ and if we write

$$h(x) = f(x) + g(x) = 2 + 3x + x^3$$

then

$$h'(x) = 3 + 3x^2$$

14.6 The derivative of a polynomial

The derivative of a polynomial is the sum of the derivatives of the terms of the polynomial.

14.6.1 Details

If

$$p(x) = a_0 + a_1x + \dots + a_nx^n$$

then

$$p'(x) = a_1 + 2a_2x + 3a_3x^2 + 4a_4x^3 + \dots + na_nx^{(n-1)}$$

14.6.2 Examples

Example 14.2. If

$$p(x) = 2x^4 + x^3$$

then

$$p'(x) = 2 \frac{dx^4}{dx} + \frac{dx^3}{dx} = 2 \cdot 4x^3 + 3x^2 = 8x^3 + 3x^2$$

14.7 The derivative of a product

If

$$h(x) = f(x) \cdot g(x)$$

then

$$h'(x) = f'(x) \cdot g(x) + f(x) \cdot g'(x)$$

14.7.1 Details

Consider two functions, f and g and their product, h :

$$h(x) = f(x) \cdot g(x).$$

The derivative of the product is given by

$$h'(x) = f'(x) \cdot g(x) + f(x) \cdot g'(x).$$

14.7.2 Examples

Example 14.3. Suppose the function f is given by

$$f(x) = xe^x + x^2 \ln x.$$

Then the derivative can be computed step by step as

$$\begin{aligned} f'(x) &= \frac{dx}{dx}e^x + x\frac{de^x}{dx} + \frac{dx^2}{dx}\ln x + x^2\frac{d\ln x}{dx} \\ &= 1 \cdot e^x + x \cdot e^x + 2x \cdot \ln x + x^2 \cdot \frac{1}{x} \\ &= e^x(1+x) + 2x\ln x + x \end{aligned}$$

14.8 Derivatives of composite functions

If f and g are functions and $h = f \circ g$ so that

$$h(x) = f(g(x)) \text{ then}$$

$$h'(x) = \frac{dh(x)}{dx} = f'(g(x))g'(x)$$

14.8.1 Examples

Example 14.4. For fixed x consider:

$$\begin{aligned} f(p) &= \ln(p^x(1-p)^{n-x}) \\ &= \ln p^x + \ln(1-p)^{n-x} \\ &= x \ln p + (n-x) \ln(1-p) \end{aligned}$$

$$\begin{aligned} f'(p) &= x\frac{1}{p} + \frac{n-x}{1-p}(-1) \\ &= \frac{x}{p} - \frac{n-x}{1-p} \end{aligned}$$

Example 14.5. $f(b) = (y - bx)^2$ (y, x fixed)

$$\begin{aligned}f'(b) &= 2(y - bx)(-x) \\ &= -2x(y - bx) \\ &= (-2xy) + (2x^2)b\end{aligned}$$

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students
This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

15 Applications of differentiation

15.1 Tracking the sign of the derivative

If f is a function, then the sign of its derivative, f' , indicates whether f is increasing ($f' > 0$), decreasing ($f' < 0$), or zero. f' can be zero at points where f has a maximum, minimum, or a saddle point.

15.1.1 Details

If f is a function, then the sign of its derivative, f' , indicates whether f is increasing ($f' > 0$), decreasing ($f' < 0$), or zero. f' can be zero at points where f has a maximum, minimum, or a saddle point.

If $f'(x) > 0$ for $x < x_0$, $f'(x_0) = 0$ and $f'(x) < 0$ for $x > x_0$ then f has a maximum at x_0

If $f'(x) < 0$ for $x < x_0$, $f'(x_0) = 0$ and $f'(x) > 0$ for $x > x_0$ then f has a minimum at x_0

If $f'(x) > 0$ for $x < x_0$, $f'(x_0) = 0$ and $f'(x) > 0$ for $x < x_0$ then f has a saddle point at x_0

If $f'(x) < 0$ for $x < x_0$, $f'(x_0) = 0$ and $f'(x) < 0$ for $x < x_0$ then f has a saddle point at x_0

15.1.2 Examples

Example 15.1. If f is a function such that its derivative is given by

$$f'(x) = (x-1)(x-2)(x-3)(x-4),$$

then applying the above criteria for maxima and minima, we see that f has maxima at 1 and 3 and f has minima at 2 and 4.

15.2 Describing extrema using f''

x_0 with $f'(x_0) = 0$ corresponds to a maximum if $f''(x_0) < 0$

x_0 with $f'(x_0) = 0$ corresponds to a minimum if $f''(x_0) > 0$

15.2.1 Details

If $f'(x_0) = 0$ corresponds to a maximum, then the derivative is decreasing and the second derivative can not be positive, (i.e. $f''(x_0) \leq 0$). In particular, if the second derivative is strictly negative, ($f''(x_0) < 0$), then we are assured that the point is indeed a maximum, and not a saddle point.

If $f'(x_0) = 0$ corresponds to a minimum, then the derivative is increasing and the second derivative can not be negative, (i.e. $f''(x_0) \geq 0$).

If the second derivative is zero, then the point may be a saddle point, as happens with $f(x) = x^3$ at $x = 0$.

15.3 The likelihood function

If p is the probability mass function (p.m.f.):

$$p(x) = P[X = x]$$

then the joint probability of obtaining a sequence of outcomes from independent sampling is

$$p(x_1) \cdot p(x_2) \cdot p(x_3) \dots p(x_n)$$

Suppose each probability includes some parameter θ , this is written,

$$p_{\theta}(x_1), \dots, p_{\theta}(x_n)$$

If the experiment gives x_1, x_2, \dots, x_n we can write the probability as a function of the parameters:

$$L_{\mathbf{x}}(\theta) = p_{\theta}(x_1), \dots, p_{\theta}(x_n).$$

This is the *likelihood function*.

15.3.1 Details

Definition 15.1. Recall that the **probability mass function (p.m.f)** is a function giving the probability of outcomes of an experiment.

We typically denote the p.m.f. by p so $p(x)$ gives the probability of a given outcome, x , of an experiment. The p.m.f. commonly depends on some parameter. We often write,

$$p(x) = P[X = x].$$

If we take a sample of independent measurements, from p , then the joint probability of a given set of numbers is,

$$p(x_1) \cdot p(x_2) \cdot p(x_3) \dots p(x_n)$$

Suppose each probability includes the same parameter θ , then this is typically written,

$$p_{\theta}(x_1), \dots, p_{\theta}(x_n)$$

Now consider the set of outcomes x_1, x_2, \dots, x_n from the experiment. We can now take the probability of this outcome as a function of the parameters.

Definition 15.2. $L_{\mathbf{x}}(\theta) = p_{\theta}(x_1), \dots, p_{\theta}(x_n)$

This is the **likelihood function** and we often seek to maximize it to estimate the unknown parameters.

15.3.2 Examples

Example 15.2. Suppose we toss a biased coin n independent times and obtain x heads, we know the probability of obtaining x heads is,

$$\binom{n}{x} p^x (1-p)^{n-x}$$

The parameter of interest is p and the likelihood function is,

$$L(p) = \binom{n}{x} p^x (1-p)^{n-x}$$

If p is unknown we sometimes wish to maximize this function with respect to p in order to estimate the *true* probability p .

15.4 Plotting the likelihood

missing slide – want to give a numeric example and plot L

15.4.1 Examples

missing example – want to give a numeric example and plot L

15.5 Maximum likelihood estimation

If L is a likelihood function for a p.m.f. p_θ , then the value $\hat{\theta}$ which gives the maximum of L :

$$L(\hat{\theta}) = \max_{\theta} (L_\theta)$$

is the maximum likelihood estimator (MLE) of θ

15.5.1 Details

Definition 15.3. If L is a likelihood function for a p.m.f. p_θ , then the value $\hat{\theta}$ which gives the maximum of L :

$$L(\hat{\theta}) = \max_{\theta} (L_\theta)$$

is the **maximum likelihood estimator** of θ

15.5.2 Examples

Example 15.3. If x is the number of heads from n independent tosses of a coin, the likelihood function is:

$$L_x(p) = \binom{n}{x} p^x (1-p)^{n-x}$$

Maximizing this is equivalent to maximizing the logarithm of the likelihood, since logarithmic functions are increasing. The log-likelihood can be written as:

$$\ln(L(p)) = \ln \binom{n}{x} + x \ln(p) + (n-x) \ln(1-p).$$

To find possible maxima, we need to differentiate this formula and set the derivative to zero

$$0 = \frac{dl(p)}{dp} = 0 + \frac{x}{p} + \frac{n-x}{1-p}(-1)$$

$$0 = p(1-p) \frac{x}{p} - p(1-p) \frac{n-x}{1-p}$$

$$0 = (1-p)x - p(n-x)$$

$$0 = x - px - pn + px = x - pn$$

So,

$$0 = x - pn$$

$$p = \frac{x}{n}$$

is the extreme and so we can write

$$\hat{p} = \frac{x}{n}$$

for the MLE

15.6 Least squares estimation

Least squares: Estimate the parameters θ by minimizing

$$\sum_{i=1}^n (y_i - g_i(\theta))^2$$

15.6.1 Details

Suppose we have a model linking data to parameters. In general we are predicting y_i as $g_i(\theta)$.

In this case it makes sense to estimate parameters θ by minimizing

$$\sum_{i=1}^n (y_i - g_i(\theta))^2.$$

15.6.2 Examples

Example 15.4. One may predict numbers, x_i , as a mean, μ , plus error. Consider the simple model $x_i = \mu + \varepsilon_i$, where μ is an unknown parameter (constant) and ε_i is the error in measurement when obtaining the i 'th observations, x_i , $i = 1, \dots, n$.

A natural method to estimate the parameter is to minimize the squared deviations

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

It is not hard to see that the $\hat{\mu}$ that minimizes this is the mean:

$$\hat{\mu} = \bar{x}.$$

Example 15.5. One also commonly predicts data y_1, \dots, y_n with values on a straight line, i.e. with $\alpha + \beta x_i$, where x_1, \dots, x_n are fixed numbers.

This leads to the *regression* problem of finding parameter values for $\hat{\alpha}$ and $\hat{\beta}$ which gives the best fitting straight line in relation to least squares:

$$\min_{\alpha, \beta} \sum (y_i - (\alpha + \beta x_i))^2$$

Example 15.6. As a general exercise in finding the extreme of a function, let's look at the function $f(\theta) = \sum_{i=1}^n (x_i \theta - 3)^2$ where x_i are some constants. We wish to find the θ that minimizes this sum. We simply differentiate θ to obtain $f'(\theta) = \sum_{i=1}^n 2(x_i \theta - 3)x_i = 2 \sum_{i=1}^n x_i^2 \theta - 2 \sum_{i=1}^n 3x_i$. Thus,

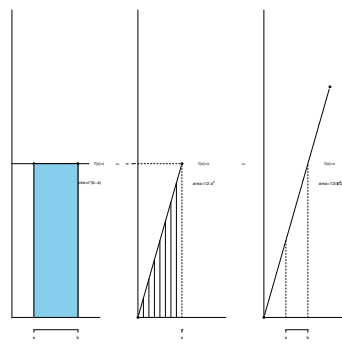
$$\begin{aligned} f'(\theta) &= 2\theta \sum_{i=1}^n x_i^2 - 2 \sum_{i=1}^n 3x_i = 0 \\ \Leftrightarrow \theta &= \frac{\sum_{i=1}^n 3x_i}{\sum_{i=1}^n x_i^2}. \end{aligned}$$

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students
This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

16 Integrals and probability density functions

16.1 Area under a curve

The area under a curve between $x=a$ and $x=b$ (for a positive function) is called the integral of the function.



Example 1, 2 and 3

16.1.1 Details

Definition 16.1. The area under a curve between $x=a$ and $x=b$ (for a positive function) is called the **integral of the function** and is denoted: $\int_a^b f(x)dx$ when it exists.

16.2 The antiderivative

Given a function f , if there is another function F such that $F' = f$, we say that F is the *antiderivative* of f . For a function f the antiderivative is denoted by $\int f dx$.

Note that if F is one antiderivative of f and C is a constant, then $G = F + C$ is also an antiderivative. It is therefore customary to always include the constant, e.g. $\int x dx = \frac{1}{2}x^2 + C$.

16.2.1 Examples

Example 16.1. The antiderivative of x to a power raises the power. $\int x^n dx = \frac{1}{n+1}x^{n+1} + C$.

Example 16.2. $\int e^x dx = e^x + C$.

Example 16.3. $\int \frac{1}{x} dx = \ln(x) + C$.

Example 16.4. $\int 2xe^{x^2} dx = e^{x^2} + C$.

16.3 The fundamental theorem of calculus

If f is a continuous function, and $F'(x) = f(x)$ for $x \in [a, b]$, then $\int_a^b f(x)dx = F(b) - F(a)$

16.3.1 Detail

It is not too hard to see that the area under the graph of a positive function f on the interval $[a, b]$ must be equal to the difference of the values of its antiderivative at a and b . This also holds for functions which take on negative values and is formally stated below.

Definition 16.2. Fundamental theorem of calculus: If F is the antiderivative of the continuous function f , i.e. $F' = f$ for $x \in [a, b]$, then $\int_a^b f(x)dx = F(b) - F(a)$. This difference is often written as $\int_a^b f dx$ or $[F(x)]_a^b$.

16.3.2 Examples

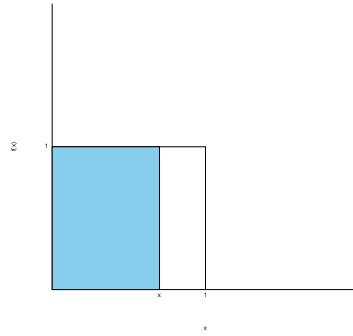
Example 16.5. The area under the graph of x^n between 0 and 3 is $\int_0^3 x^n dx = \left[\frac{1}{n+1}x^{n+1}\right]_0^3 = \frac{1}{n+1}3^{n+1} - \frac{1}{n+1}0^{n+1} = \frac{3^{n+1}}{n+1}$

Example 16.6. The area under the graph of e^x between 3 and 4 is $\int_3^4 e^x dx = [e^x]_3^4 = e^4 - e^3$

Example 16.7. The area under the graph of $\frac{1}{x}$ between 1 and a is $\int_1^a \frac{1}{x} dx = [\ln(x)]_1^a = \ln(a) - \ln(1) = \ln(a)$.

16.4 Density functions

The probability density function (p.d.f.) and the cumulative distribution function (c.d.f.).



16.4.1 Details

Definition 16.3. If X is a random variable such that

$$P(a \leq X \leq b) = \int_a^b f(x) dx,$$

for some function f which satisfies $f(x) \geq 0$ for all x and

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

then f is said to be a **probability density function (p.d.f.)** for X .

Definition 16.4. The function

$$F(x) = \int_{-\infty}^x f(t) dt$$

is the **cumulative distribution function (c.d.f.)**.

16.4.2 Examples

Example 16.8. Consider a random variable X from the uniform distribution, denoted by $X \sim U(0, 1)$. This distribution has density

$$f(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{e.w.} \end{cases}$$

The cumulative distribution function is given by

$$P[X \leq x] = \int_{-\infty}^x f(t) dt = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \end{cases}$$

Example 16.9. Suppose $X \sim P(\lambda)$, where X may denote the number of events per unit time. The p.m.f. of X is described by $p(x) = P[X = x] = e^{-\lambda} \frac{\lambda^x}{x!}$ for $x = 0, 1, 2, \dots$. Consider now the waiting time, T , between events, or simply until the first event. Consider the event $T > t$ for some number $t > 0$. If $X \sim p(\lambda)$ denotes the number of events per unit time, then let X_t denote the number of events during the time period for 0 through t . Then it is natural to assume

$X_t \sim P(\lambda t)$ and it follows that $T > t$ if and only if $X_t = 0$ and we obtain $P[T > t] = P[X_t = 0] = e^{-\lambda t}$. It follows that the c.d.f. of T is $F_T(t) = P[T \leq t] = 1 - P[T > t] = 1 - e^{-\lambda t}$ for $t > 0$.

The p.d.f. of T is therefore $f_T(t) = F_T'(t) = \frac{d}{dt} F_T(t) = \frac{d}{dt} (1 - e^{-\lambda t}) = 0 - e^{-\lambda t} * (-\lambda) = \lambda e^{-\lambda t}$ for $t \geq 0$ and $f_T(t) = 0$ for $t < 0$.

The resulting density

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{for } t \geq 0 \\ 0 & \text{for } t < 0 \end{cases}$$

describes the exponential distribution.

This distribution has the expected value

$$E[T] = \int_{-\infty}^{\infty} t f(t) dt = \int_0^{\infty} t \lambda e^{-\lambda t} dt.$$

the stuff below is all messed up...

We set $u = \lambda t$ and $du = \lambda dt$ to obtain

$$\begin{aligned} \int u e^{-u} du &= \frac{1}{\lambda} \int_0^{\infty} u e^{-u} du = \frac{1}{\lambda} \int_0^{\infty} 1 \cdot e^{-u} du \\ &= [-u e^{-u}]_0^{\infty} \\ &= \left[\frac{1}{\lambda} (-e^{-u}) \right]_0^{\infty} - 0 = \frac{1}{\lambda}. \end{aligned}$$

16.5 Probabilities in R: The normal distribution

R has functions to compute values of probability density functions (p.d.f.) and cumulative distribution functions (c.m.d.) for most common distributions.

16.5.1 Details

The p.d.f. for the normal distribution is

$$p(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

The c.d.f. for the normal distribution is

$$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

16.5.2 Examples

Example 16.10. `dnorm()` gives the value of the normal p.d.f.

Example 16.11. `pnorm()` gives the value of the normal c.d.f.

16.6 Some rules of integration

16.6.1 Examples

Example 16.12. Using integration by parts we obtain

$$\int \ln(x) x dx = \frac{1}{2} x^2 \ln(x) - \int \frac{1}{2} x^2 \cdot \frac{1}{x} dx = \frac{1}{2} x^2 \ln(x) - \int \frac{1}{2} x dx = \frac{1}{2} x^2 \ln(x) - \frac{1}{4} x^2.$$

Example 16.13. Consider $\int_1^2 2x e^{x^2} dx$. By setting $x = g(t) = \sqrt{t}$ we obtain

$$\int_1^2 2x e^{x^2} dx = \int_1^4 2\sqrt{t} e^t \frac{1}{2\sqrt{t}} dt = \int_1^4 e^t dt = e^4 - e.$$

16.6.2 Handout

The two most common "tricks" applied in integration are a) integration by parts and b) integration by substitution.

a) Integration by parts

$$(fg)' = f'g + fg'$$

by integrating both sides of the equation we obtain:

$$fg = \int f'g dx + \int fg' dx \Leftrightarrow \int fg' dx = fg - \int f'g dx$$

b) Integration by substitution

Consider the definite integral $\int_a^b f(x)dx$ and let g be a one-to-one differential function for the interval (c, d) to (a, b) . Then

$$\int_a^b f(x)dx = \int_c^d f(g(y))g'(y)dy$$

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students
This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

17 Principles of programming

17.1 Modularity

Modularity involves designing a system that is divided into a set of functional units (named modules) that can be composed into a larger application.

Any programming project should be split into logical module pieces of code which are combined into a complete program.

17.1.1 Details

Typically input, initialization, analysis, and output commands are grouped into separate parts.

17.1.2 Examples

Example 17.1. Input

```
dat<-read.table("http://notendur.hi.is/~gunnar/kennsla/alsm/data/
  set115.dat", header=T)
cols<- c("le", "osl")
```

Analysis

```
Mn<-mean(dat[, cols[1]])
```

Output

```
print (Mn)
```

17.2 Modularity and functions

In many cases groups of commands can be collected together into a function.

17.2.1 Details

Typically a project has several such functions.

17.2.2 Examples

Example 17.2. Suppose you want to plot the weight vs. length for several datasets in

```
http://hi.is/~gunnar/kennsla/alsm/data
```

A function can then be set up with the file number as an argument:

```
plotwtle<-function (fnum){
  fname<-paste(
    "http://hi.is/~gunnar/kennsla/alsm/data/set",fnum,".dat",sep="")
  cat("The URL is ", fname, "\n")
```

```
dat<-read.table(fname,header=T)
ttl<-paste("Data_from_file_number", fnum)
plot(dat$le,dat$osl,main=ttl)
}
```

Now call this with

```
plotwtle(105)
```

17.3 Modularity and files

It is advisable to split larger projects into several manageable files.

17.3.1 Details

Once a project reaches more than five lines of code, it should be stored in one or more separate files. In order to combine these files a single “source” command file can be created.

Typically function definitions are stored in separate files, so one may have several separate files like:

```
"input.r"
"function.r"
"analysis.r"
output.r"
```

While developing the analysis, the data would only be read once with

```
source("input.r")
```

The goal of this practice is to end up with a set of files which are completely self-contained, so one can start with an empty R session and give only the commands like:

```
source ("input.r")
source ("functions.r")
source ("analysis.r")
```

Furthermore, this ensures repeatability.

17.3.2 Examples

Example 17.3. For a given project “input”, “functions” “analysis” and “output” files can be created as below.

input.r

```
dat<-read.table("http://notendur.hi.is/~gunnar/kennsla/alsm/data/
set115.dat", header=T)
```


functions.r

```
plotwtle<-function(fnum){
  fname<-paste("http://notendur.hi.is/~gunnar/kennsla/alsm/data/set",
    fnum, ".dat", sep="")
  cat("The URL is", fname, "\n")
  dat<-read.table(fname,header=T)
  ttl<-paste("My data set was", fnum)
  plot(dat$le,dat$osl,main=ttl,xlab="Length(cm)",ylab="Live weight(g)")
}
```

output.r

```
source("functions.r")
for(i in 101:150){
  fnam<-paste("plot",i, ".pdf", sep="")
  pdf(fnam)
  plotwtle(i)
  dev.off()
}
```

These files can be executed with source commands as below:

```
source("input.r")
```

```
source("functions.r")
```

```
source("output.r")
```

17.4 Structuring an R project

17.4.1 Details

We already covered how to split code into different functions and linking them together with the help of one executable file that is "sourcing" the others. However, when you undertake a larger project, there will be a lot of different data and files and it is very advisable to have a consistent structure throughout the project.

A common project layout is to allocate all project files into a folder, something along the lines of:

```
/project
/data
/src
/doc
/figs (or /out)
```

Such a structure is quite normal in programming languages such as C, Matlab, and R.

Purpose of the different folders:

/data: Contains all important data to the project, which you will use. This folder should be read-only! No function is allowed to write anything into this folder.

/src: (abbreviation for "source(-code)") Here you will store all the functions that you programmed. You can decide to store the executable function here as well or, alternatively, have that one in the root project folder.

/doc: Contains further documentation material about your project. This could be, for example, readme files for other people who use your functions, or the paper you wrote about the project, or the latex files while you're writing.

/figs or /out: Here your functions are allowed to write and can produce the different results, like graphs, figures or anything else.

Finally, a large programming project should at some stage be split into packages and stored on dedicated servers such as github or CRAN.

17.4.2 Examples

Example 17.4. Consider first the issue of maintaining the code itself. It is common for R beginners to only work interactively within the command-line interface. However, it is essential that the code be kept in one or more files.

For large projects these will be several different files, each with its own purpose. To run a complete analysis one would typically set up one file to run all the tasks by reading in data through analyses to outputs.

For example, a file named "run.r" could contain the sequence of commands:

```
source("setup.r")  
  
source("analysis.r")  
  
source("plot.r")
```

17.5 Loops, for

If a piece of code is to be run repeatedly, the for-loop is normally used.

17.5.1 Details

If a piece of code is to be run repeatedly, the for-loop is normally used. The R code form is:

```
for(index in sequence){  
  commands  
}
```

17.5.2 Examples

Example 17.5. To add numbers we can use

```
tot <- 100
for(i in 1:100){
  tot <- tot + i
}
cat ("the sum is", tot, "\n")
```

Example 17.6. Define the plot function

```
plotwtle <- AS BEFORE
```

To plot several of these we can use a sequence:

```
plotwtle(101)
plotwtle(102)
.
.
.
```

or a loop

```
for (i in 101:150){
  fname<- paste("plot", i, ".pdf", sep="")
  pdf(fname)
  plotwtle(i)
  dev.off()
}
```

17.6 The if and ifelse commands

The "if" statement is used to conditionally execute statements.

The "ifelse" statement conditionally replaces elements of a structure.

17.6.1 Examples

Example 17.7. If we want to compute x^x for x -values in the range 0 through 5, we can use

```
xlist<-seq(0,5,0.01)
y<-NULL
for(x in xlist){
  if(x==0){
    y<-c(y,1)
```

```
}else{
  y<-c(y,x**x)
}
}
```

Example 17.8. `x<-seq(0,5,0.01)`
`y<-ifelse(x==0,1,x^x)`

Example 17.9. `dat<-read.table("file")`
`dat<-ifelse(dat==0,0.01,dat)`

Example 17.10. `x<-ifelse(is.na(x),0,x)`

17.7 Indenting

Code should be properly indented!

17.7.1 Details

fFunctions, for-loops, and if-statements should always be indented.

17.8 Comments

All code should contain informative comments. Comments are separated out from code using the pound symbol (#).

17.8.1 Examples

Example 17.11. #####
####SETUP DATA####

`dat<-read.table(filename)`
`x<-log(dat$le) #log-transformation of length`

```
y<-log(dat$wt) #log-transformation of weight

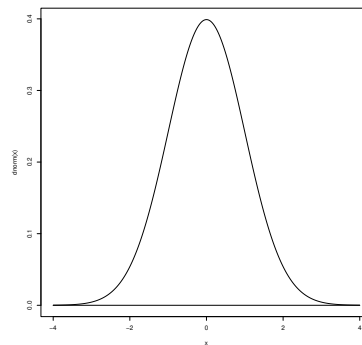
#####
####THE ANALYSIS####
#####
```

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students
This work is licensed under the Creative Commons Attribution-ShareAlike License. To
view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a
letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

18 The Central Limit Theorem and related topics

18.1 The Central Limit Theorem

If measurements are obtained independently and come from a process with finite variance, then the distribution of their mean tends towards a Gaussian (normal) distribution as the sample size increases.



standard normal density

The

18.1.1 Details

Theorem 18.1 The **Central Limit Theorem** states that if X_1, X_2, \dots are independent and identically distributed random variables with mean μ and (finite) variance σ^2 , then the distribution of $\bar{X}_n := \frac{X_1 + \dots + X_n}{n}$ tends towards a normal distribution.

It follows that for a large enough sample size n , the distribution random variable \bar{X}_n can be approximated by $n(\mu, \sigma^2/n)$.

The standard normal distribution is given by the p.d.f.

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

for $z \in \mathbb{R}$.

The standard normal distribution has an expected value of zero,

$$\mu = \int z\varphi(z)dz = 0$$

and a variance of

$$\sigma^2 = \int (z - \mu)^2 \varphi(z) dz = 1$$

If a random variable Z has the standard normal (or Gaussian) distribution, we write $Z \sim n(0, 1)$.

If we define a new random variable, Y , by writing $Y = \sigma Z + \mu$, then Y has an expected value of μ , a variance of σ^2 and a density (p.d.f.) given by the formula:

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

This is general normal (or Gaussian) density, with mean μ and variance σ^2 .

The Central Limit Theorem states that if you take the mean of several independent random variables, the distribution of that mean will look more and more like a Gaussian distribution (if the variance of the original random variables is finite).

More precisely, the cumulative distribution function of

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

converges to Φ , the $n(0, 1)$ cumulative distribution function.

18.1.2 Examples

Example 18.1. If we collect measurements on waiting times, these are typically assumed to come from an exponential distribution with density

$$f(t) = \lambda e^{-\lambda t}, \text{ for } t > 0$$

The Central Limit Theorem states that the mean of several such waiting times will tend to have a normal distribution.

Example 18.2. We are often interested in computing

$$w = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

which comes from a t-distribution (see below), if the x_i are independent outcomes from a normal distribution.

However, if n is large and σ^2 is finite then w values will look as though they came from a normal distribution. This is in part a consequence of the Central Limit Theorem, but also of the fact that s will become close to σ as n increases.

18.2 Properties of the binomial and Poisson distributions

The binomial distribution is really a sum of 0 and 1 values (counts of failures = 0 and successes = 1). So, a simple, single binomial outcome will correspond to coming from a normal distribution if the count is large enough.

18.2.1 Details

Consider the binomial probabilities:

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

for $x = 0, 1, 2, 3, \dots, n$ where n is a non-negative integer. Suppose p is a small positive number, specifically consider a sequence of decreasing p -values, specified with $p_n = \frac{\lambda}{n}$ and consider the behavior of the probability as $n \rightarrow \infty$. We obtain:

$$\binom{n}{x} p_n^x (1-p_n)^{n-x} = \frac{n!}{x!(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \quad (5)$$

$$= \frac{n(n-1)(n-2)\cdots(n-x+1)}{x!} \frac{\frac{\lambda^x}{n^x}}{\left(1 - \frac{\lambda}{n}\right)^x} \left(1 - \frac{\lambda}{n}\right)^n \quad (6)$$

$$= \frac{n(n-1)(n-2)\cdots(n-x+1)}{x!n^x} \frac{\lambda^x}{\left(1 - \frac{\lambda}{n}\right)^x} \left(1 - \frac{\lambda}{n}\right)^n \quad (7)$$

$$(8)$$

Note 18.1. Notice that $\frac{n(n-1)(n-2)\cdots(n-x+1)}{n^x} \rightarrow 1$ as $n \rightarrow \infty$. Also notice that $\left(1 - \frac{\lambda}{n}\right)^x \rightarrow 1$ as $n \rightarrow \infty$. Also

$$\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right) = e^{-\lambda}$$

and it follows that

$$\lim_{n \rightarrow \infty} \binom{n}{x} p_n^x (1-p_n)^{n-x} = \frac{e^{-\lambda} \lambda^x}{x!}, x = 0, 1, 2, \dots, n$$

and hence the binomial probabilities may be approximated with the corresponding Poisson.

18.2.2 Examples

Example 18.3. The mean of a binomial (n,p) variable is $\mu = n \cdot p$ and the variance is $\sigma^2 = np(1-p)$.

The R command `dbinom(q,n,p)` calculates the probability of q successes in n trials assuming that the probability of a success is p in each trial (binomial distribution), and the R command `pbinom(q,n,p)` calculates the probability of obtaining q or fewer successes in n trials.

The normal approximation of this distribution can be calculated with `pnorm($q,mu,sigma$)` which becomes `pnorm($q,n * p,sqrt(n * p(1-p))$)`. Three numerical examples (note that `pbinom` and `pnorm` give similar values for large n):

```
pbinom(3,10,0.2)
```

```
[1] 0.8791261
```

```
pnorm(3,10*0.2,sqrt(10*0.2*(1-0.2)))
```

```
[1] 0.7854023
```

```
pbinom(3,20,0.2)
```

```
[1] 0.4114489
```

```
pnorm(3,20*0.2,sqrt(20*0.2*(1-0.2)))
```

```
[1] 0.2880751
```

```
pbinom(30,200,0.2)
```

```
[1] 0.04302156
```

```
pnorm(30,200*0.2,sqrt(200*0.2*(1-0.2)))
```

```
[1] 0.03854994
```


Example 18.4. We are often interested in computing $w = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ which has a t-distribution if the x_i are independent outcomes from a normal distribution. If n is large and σ^2 is finite, this will look as if it comes from a normal distribution.

The numerical examples below demonstrate how the t-distribution approaches the normal distribution.

```

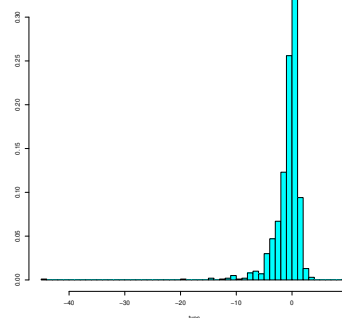
qnorm(0.7)
[1] 0.5244005
#This is the value which gives the cumulative probability of p=0.7
  for a  $n \sim (0,1)$ 
qt(0.7,2)
[1] 0.6172134
#The value, which gives the cumulative probability of p=0.7 with n=2
  for the t-distribution.
qt(0.7,5)
[1] 0.5594296
qt(0.7,10)
[1] 0.541528
qt(0.7,20)
[1] 0.5328628

qt(0.7,100)
[1] 0.5260763

```

18.3 Monte Carlo simulation

If we know an underlying process we can simulate data from the process and evaluate the distribution of any quantity based on such data.



A simulated set of t -values based on data from an exponential distribution.

18.3.1 Examples

Example 18.5. Suppose our measurements come from an exponential distribution and we want to compute

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

but we want to know the distribution of those when μ is the true mean.

For instance, $n = 5$ and $\mu = 1$, we can simulate (repeatedly) x_1, \dots, x_5 and compute a t-value for each. The following R commands can be used for this:

```
library(MASS)
n<-5
mu<-1
lambda<-1
tvec<-NULL
for(sim in 1:10000){
  x<-rexp(n,lambda)
  xbar<-mean(x)
  s<-sd(x)
  t<-(xbar-mu)/(s/sqrt(n))
  tvec<-c(tvec,t)
}

#then do...

truehist(tvec) #truehist gives a better histogram
sort(tvec)[9750]
sort(tvec)[250]
```

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students
This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

19 Miscellaneous

19.1 Simple probabilities in R

R has functions to compute probabilities based on most common distributions.

If X is a random variable with a known distribution, then R can typically compute values of the cumulative distribution function or:

$$F(x) = P[X \leq x]$$

19.1.1 Examples

Example 19.1. If $X \sim b(n, p)$ has binomial distribution, i.e.

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

then cumulative probabilities can be computed with `pbinom`, e.g.

```
pbinom(5, 10, 0.5)
```

gives

$$P[X \leq 5] = 0.623$$

where

$$X \sim b(n = 10, p = \frac{1}{2}).$$

This can also be computed by hand. Here we have $n = 10$, $p = 1/2$ and the probability $P[X \leq 5]$ is obtained by adding up the individual probabilities, $P[X = 0] + P[X = 1] + P[X = 2] + P[X = 3] + P[X = 4] + P[X = 5]$

$$P[X \leq 5] = \sum_{x=0}^5 \binom{10}{x} \frac{1}{2}^x \frac{1}{2}^{10-x}.$$

This becomes

$$P[X \leq 5] = \binom{10}{0} \frac{1}{2}^0 \frac{1}{2}^{10-0} + \binom{10}{1} \frac{1}{2}^1 \frac{1}{2}^{10-1} + \binom{10}{1} \frac{1}{2}^2 \frac{1}{2}^{10-2} + \binom{10}{3} \frac{1}{2}^3 \frac{1}{2}^{10-3} + \binom{10}{4} \frac{1}{2}^4 \frac{1}{2}^{10-4} + \binom{10}{5} \frac{1}{2}^5 \frac{1}{2}^{10-5}$$

or

$$P[X \leq 5] = \binom{10}{0} \frac{1}{2}^{10} + \binom{10}{1} \frac{1}{2}^{10} + \binom{10}{1} \frac{1}{2}^{10} + \binom{10}{3} \frac{1}{2}^{10} + \binom{10}{4} \frac{1}{2}^{10} + \binom{10}{5} \frac{1}{2}^{10} = \frac{1}{2}^{10} [1 + 10 + 45 + \dots].$$

Furthermore,

```
pbinom(10, 10, 0.5)
```

```
[1] 1
```

and

```
pbinom(0, 10, 0.5)
```

```
[1] 0.0009765625
```

It is sometimes of interest to compute $P[X = x]$ in this case, and this is given by the *dbinom* function, e.g.

```
dbinom(1, 10, 0.5)
[1] 0.009765625
```

or $\frac{10}{1024}$

Example 19.2. Suppose X has a uniform distribution between 0 and 1, i.e. $X \sim U(0, 1)$. Then the *punif* function will return probabilities of the form

$$P[X \leq x] = \int_{-\infty}^x f(t)dt = \int_0^x f(t)dt$$

where $f(t) = 1$ if $0 \leq t \leq 1$ and $f(t) = 0$. For example:

```
punif(0.75)
[1] 0.75
```

To obtain $P[a \leq X \leq b]$, we use *punif* twice, e.g.

```
punif(0.75) - punif(0.25)
[1] 0.5
```

19.2 Computing normal probabilities in R

To compute probabilities $X \sim n(\mu, \sigma^2)$ is usually transformed, since we know that

$$Z := \frac{X - \mu}{\sigma} \sim (0, 1)$$

The probabilities can then be computed for either X or Z with the *pnorm* function in R.

19.2.1 Details

Suppose X has a normal distribution with mean μ and variance

$$X \sim n(\mu, \sigma^2)$$

then to compute probabilities, X is usually transformed, since we know that

$$Z = \frac{X - \mu}{\sigma} \sim (0, 1)$$

and the probabilities can be computed for either X or Z with the *pnorm* function.

19.2.2 Examples

Example 19.3. If $Z \sim n(0, 1)$ then we can e.g. obtain $P[Z \leq 1.96]$ with

```
pnorm(1.96)
[1] 0.9750021
```

```
pnorm(0)
[1] 0.5
```

```
pnorm(1.96) - pnorm(-1.96)
[1] 0
```

```
pnorm(1.96) - pnorm(-1.96)
[1] 0.9500042
```

The last one gives the area between -1.96 and 1.96.

Example 19.4. If $X \sim n(42, 3^2)$ then we can compute probabilities either by transforming

$$\begin{aligned} P[X \leq x] &= P\left[\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right] \\ &= P\left[Z \leq \frac{x - \mu}{\sigma}\right] \end{aligned}$$

and calling *pnorm* with the computed value $z = \frac{x - \mu}{\sigma}$, or call *pnorm* with x and specify μ and σ .

To compute $P[X \leq 48]$, either set $z = (48 - 42)/3 = 2$ and obtain

```
pnorm(2)
[1] 0.9772499
```

or specify μ and σ

```
pnorm(48, 42, 3)
[1] 0.9772499
```

19.3 Introduction to hypothesis testing

19.3.1 Details

If we have a random sample x_1, \dots, x_n from a normal distribution, then we consider them to be outcomes of independent random variables X_1, \dots, X_n where $X_i \sim n(\mu, \sigma^2)$. Typically, μ and σ^2 are unknown but assume for now that σ^2 is known.

Consider the hypothesis:

$H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$

where μ_0 is a specified number.

Under the assumption of independence, the sample mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

is also an observation from a normal distribution, with mean μ but a smaller variance. Specifically, \bar{x} is the outcome of

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

and

$$X \sim n\left(\mu, \frac{\sigma^2}{n}\right)$$

so the standard deviation of X is $\frac{\sigma}{\sqrt{n}}$, so the appropriate error measure for \bar{x} is $\frac{\sigma}{\sqrt{n}}$, when σ is unknown.

If H_0 is true, then

$$z := \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

is an observation from an $n \sim n(0, 1)$ distribution, i.e. an outcome of

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

where $Z \sim n(0, 1)$ when H_0 is correct. It follows that e.g. $P[|Z| > 1.96] = 0.05$ and if we observe $|Z| > 1.96$ then we reject the null hypothesis.

Note that the value $z^* = 1.96$ is a quantile of the normal distribution and we can obtain other quantiles with the *pnorm* function, e.g. *pnorm*(0.975) gives 1.96.

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students
This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

20 Multivariate probability distributions

20.1 Joint probability distribution

If X_1, \dots, X_n are discrete random variables with $P[X_1 = x_1, X_2 = x_2, \dots, X_n = x_n] = p(x_1, \dots, x_n)$, where x_1, \dots, x_n are numbers, then the function p is the joint probability mass function (p.m.f.) for the random variables X_1, \dots, X_n .

For continuous random variables Y_1, \dots, Y_n , a function f is called the joint probability density function if,

$$P[Y \in A] = \int \int \dots \int f(y_1, \dots, y_n) dy_1 dy_2 \dots dy_n.$$

20.1.1 Details

Definition 20.1. If X_1, \dots, X_n are discrete random variables with $P[X_1 = x_1, X_2 = x_2, \dots, X_n = x_n] = p(x_1, \dots, x_n)$ where $x_1 \dots x_n$ are numbers, then the function p is the joint **probability mass function (p.m.f.)** for the random variables X_1, \dots, X_n .

Definition 20.2. For continuous random variables Y_1, \dots, Y_n , a function f is called the joint probability density function if,

$$P[Y \in A] = \underbrace{\int \int \dots \int}_A f(y_1, \dots, y_n) dy_1 dy_2 \dots dy_n.$$

Note 20.1. Note that if X_1, \dots, X_n are independent and identically distributed, each with p.m.f. p , then $p(x_1, x_2, \dots, x_n) = q(x_1)q(x_2) \dots q(x_n)$, i.e., $P[X_1 = x_1, X_2 = x_2, \dots, X_n = x_n] = P[X_1 = x_1]P[X_2 = x_2] \dots P[X_n = x_n]$.

Note 20.2. Note also that if A is a set of possible outcomes ($A \subseteq \mathbb{R}^n$), then we have

$$P[X \in A] = \sum_{(x_1, \dots, x_n) \in A} p(x_1, \dots, x_n).$$

20.1.2 Examples

Example 20.1. An urn contains blue and red marbles, which are either light or heavy. Let X denote the color and Y the weight of a marble, chosen at random

X/Y	L	H	TT
B	5	6	11
R	7	2	9
TT	12	8	20

We have $P[X = "b", Y = "l"] = \frac{5}{20}$.

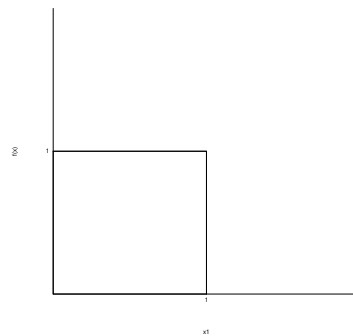
The joint p.m.f. is:

X/Y	L	H	TT
B	$\frac{5}{20}$	$\frac{6}{20}$	$\frac{11}{20}$
R	$\frac{7}{20}$	$\frac{2}{20}$	$\frac{9}{20}$
TT	$\frac{12}{20}$	$\frac{8}{20}$	1

20.2 The random sample

A set of random variables X_1, \dots, X_n is a random sample if they are independent and identically distributed (i.i.d.).

A set of numbers x_1, \dots, x_n are called a random sample if they can be viewed as an outcome of such random variables.



20.2.1 Details

Samples from populations can be obtained in a number of ways. However, to draw valid conclusions about populations, the samples need to be obtained randomly.

Definition 20.3. In **random sampling**, each item or element of the population has an equal and independent chance of being selected.

A set of random variables; $X_1 \dots X_n$ is a random sample if they are independent and identically distributed (i.i.d.).

Definition 20.4. If a set of numbers $x_1 \dots x_n$ can be viewed as an outcome of random variables, these are called a **random sample**.

20.2.2 Examples

Example 20.2. If $X_1, \dots, X_n \sim U(0, 1)$, i.i.d., i.e. X_1 and X_n are independent and each have a uniform distribution between 0 and 1. Then they have a joint density which is the product of the densities of X_1 and X_n .

Given the data in the above figure and if $x_1, x_2 \in \mathbb{R}$

$$f(x_1, x_2) = f_1(x_1)f_2(x_2) = \begin{cases} 1 & \text{if } 0 \leq x_1, x_2 \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

Example 20.3. Toss two dice independently, and let X_1, X_2 denote the two (future) outcomes.

Then

$$P[X_1 = x_1, X_2 = x_2] = \begin{cases} \frac{1}{36} & \text{if } 1 \leq x_1, x_2 \leq 6 \\ 0 & \text{elsewhere} \end{cases}$$

is the joint p.m.f.

20.3 The sum of discrete random variables

20.3.1 Details

Suppose X and Y are discrete random values with a probability mass function p . Let $Z = X + Y$. Then

$$P(Z = z) = \sum_{\{(x,y):x+y=z\}} p(x,y)$$

20.3.2 Examples

Example 20.4. $X, Y = \text{outcomes}$,

	[, 1]	[, 2]	[, 3]	[, 4]	[, 5]	[, 6]
[1,]	2	3	4	5	6	7
[2,]	3	4	5	6	7	8
[3,]	4	5	6	7	8	9
[4,]	5	6	7	8	9	10
[5,]	6	7	8	9	10	11
[6,]	7	8	9	10	11	12

$$P[X + Y = 7] = \frac{6}{36} = \frac{1}{6}$$

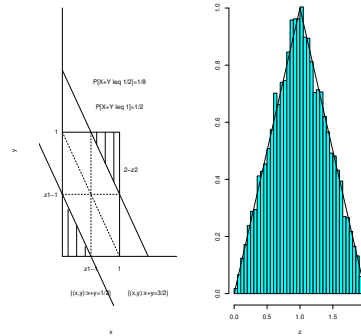
Because there are a total of 36 equally likely outcomes and 7 occurs six times this means that $P[X + Y = 7] = \frac{1}{6}$.

Also

$$P[X + Y = 4] = \frac{3}{36} = \frac{1}{12}$$

20.4 The sum of two continuous random variables

If X and Y are continuous random variables with joint p.d.f. f and $Z = X + Y$, then we can find the density of Z by calculating the cumulative distribution function.



20.4.1 Details

If X and Y are c.r.v. with joint p.d.f. f and $Z = X + Y$, then we can find the density of Z by first finding the cumulative distribution function

$$P[Z \leq z] = P[X + Y \leq z] = \iint_{\{(x,y):x+y \leq z\}} f(x,y) dx dy.$$

20.4.2 Examples

Example 20.5. If X and $Y \sim U(0,1)$, independent and $Z = X + Y$ then

$$P[Z \leq z] = \begin{cases} 0 & \text{for } z \leq 0 \\ \frac{z^2}{2} & \text{for } 0 < z < 1 \\ 1 & \text{for } z > 2 \\ 1 - \frac{(2-z)^2}{2} & \text{for } 1 < z < 2 \end{cases}$$

the density of z becomes

$$g(z) = \begin{cases} z & \text{for } 0 < z \leq 1 \\ 2 - z & \text{for } 1 < z \leq 2 \\ 0 & \text{for elsewhere} \end{cases}$$

Example 20.6. To approximate the distribution of $Z = X + Y$ where $X, Y \sim U(0, 1)$ i.i.d., we can use Monte Carlo simulation. So, generate 10.000 pairs, set them up in a matrix and compute the sum.

20.5 Means and variances of linear combinations of independent random variables

If X and Y are random variables and $a, b \in \mathbb{R}$, then

$$E[aX + bY] = aE[X] + bE[Y].$$

20.5.1 Details

If X and Y are random variables, then

$$E[X + Y] = E[X] + E[Y]$$

i.e. the expected value of the sum is just the sum of the expected values. The same applies to a finite sum, and more generally

$$E\left[\sum_{i=1}^n a_i X_i\right] = \sum_{i=1}^n a_i E[X_i]$$

when X_1, \dots, X_n are random variables and a_1, \dots, a_n are numbers (if the expectations exist). If the random variables are independent, then the variance also add

$$V[X + Y] = V[X] + V[Y]$$

and

$$V\left[\sum_{i=1}^n a_i X_i\right] = \sum_{i=1}^n a_i^2 V[X_i]$$

20.5.2 Examples

Example 20.7. $X, Y \sim U(0, 1)$, i.i.d. then

$$E[X + Y] = E[X] + E[Y] = \int_0^1 x \cdot 1 dx + \int_0^1 x \cdot 1 dx = \left[\frac{1}{2}x^2\right]_0^1 + \left[\frac{1}{2}x^2\right]_0^1 = 1.$$

Example 20.8. Let $X, Y \sim N(0, 1)$. Then $E[X^2 + Y^2] = 1 + 1 = 2$.

20.6 Means and variances of linear combinations of measurements

If x_1, \dots, x_n and y_1, \dots, y_n are numbers, and we set

$$z_i = x_i + y_i$$

$$w_i = ax_i$$

where $a > 0$, then

$$\bar{z} = \frac{1}{n} \sum_{i=1}^n z_i = \bar{x} + \bar{y}$$

$$\bar{w} = a\bar{x}$$

$$s_w^2 = \frac{1}{n-1} \sum_{i=1}^n (w_i - \bar{w})^2$$

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

$$= a^2 s_x^2$$

and

$$s_w = as_x$$

20.6.1 Examples

Example 20.9. We set:

$$a < -3$$

$$x < -c(1:5)$$

$$y < -c(6:10)$$

Then:

$$z < -x + y$$

$$w < -a * x$$

$$n < \text{length}(x)$$

Then \bar{z} is:

$$(\text{sum}(x) + \text{sum}(y)) / n$$

$$[1] \ 11$$

$$\text{mean}(z)$$

$$[1] \ 11$$

and \bar{w} becomes:

$$a * \text{mean}(x)$$

$$[1] \ 9$$

$$\text{mean}(w)$$

$$[1] \ 9$$

and s_w^2 equals:

$$\text{sum}((w - \text{mean}(w))^2) / (n - 1)$$

$$[1] \ 22.5$$

$$\text{sum}((a * x - a * \text{mean}(x))^2) / (n - 1)$$

$$[1] \ 22.5$$

```

a^2*var(x)
[1] 22.5
and s_w equals:
a*sd(x)
[1] 4.743416
sd(w)
[1] 4.743416

```

20.7 The joint density of independent normal random variables

If $Z_1, Z_2 \sim n(0, 1)$ are independent then they each have density

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, x \in \mathbb{R}$$

and the joint density is the product $f(z_1, z_2) = \phi(z_1)\phi(z_2)$ or

$$f(z_1, z_2) = \frac{1}{(\sqrt{2\pi})^2} e^{-\frac{z_1^2}{2} - \frac{z_2^2}{2}}.$$

20.7.1 Details

If $X \sim n(\mu_1, \sigma_1^2)$ and $Y \sim n(\mu_2, \sigma_2^2)$ are independent, then their densities are

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}$$

and

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}}$$

and the joint density becomes

$$\frac{1}{2\pi\sigma_1\sigma_2} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2} - \frac{(y-\mu_2)^2}{2\sigma_2^2}}$$

Now, suppose $X_1, \dots, X_n \sim n(\mu, \sigma^2)$ are i.i.d., then

$$f(\underline{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} e^{-\sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2}}$$

is the multivariate normal density in the case of i.i.d. variables.

20.8 More general multivariate probability density functions

20.8.1 Examples

Example 20.10. Suppose X and Y have the joint density

$$f(x,y) = \begin{cases} 2 & 0 \leq y \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

First notice that $\int_{\mathbb{R}} \int_{\mathbb{R}} f(x,y) dx dy = \int_{x=0}^1 \int_{y=0}^x 2 dy dx = \int_0^1 2x dx = 1$, so f is indeed a density function.

Now, to find the density of X we first find the c.d.f. of X , first note that for $a < 0$ we have $P[X \leq a] = 0$ but if $a \geq 0$, we obtain

$$F_X(a) = P[X \leq a] = \int_{x_0}^a \int_{y=0}^x 2 dy dx = [x^2]_0^a = a^2.$$

The density of X is therefore

$$f_X(x) = \frac{dF(x)}{dx} = \begin{cases} 2x & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}.$$

20.8.2 Handout

If

$$f : \mathbb{R}^n \rightarrow \mathbb{R}$$

is such that

$$P[X \in A] = \int_A \dots \int f(x_1, \dots, x_n) dx_1 \cdots dx_n$$

and $f(x) \geq 0$ for all $\underline{x} \in \mathbb{R}^n$

then f is the *joint density* of

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$$

If we have the joint density of some multidimensional random variable $X = (X_1, \dots, X_n)$ given in this manner, then we can find the individual density functions of the X_i 's by integrating the other variables.

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students
This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

21 Some distributions related to the normal

21.1 The normal and sums of normals

The sum of independent normally distributed random variables is also normally distributed.

21.1.1 Details

The sum of independent normally distributed random variables is also normally distributed. More specifically, if $X_1 \sim n(\mu_1, \sigma_1^2)$ and $X_2 \sim n(\mu_2, \sigma_2^2)$ are independent then $X_1 + X_2 \sim n(\mu, \sigma^2)$ since $\mu = E[X_1 + X_2] = \mu_1 + \mu_2$ and $\sigma^2 = V[X_1 + X_2]$ with $\sigma^2 = \sigma_1^2 + \sigma_2^2$ if X_1 and X_2 are independent.

Similarly

$$\sum_{i=1}^n X_i$$

is normal if X_1, \dots, X_n are normal and independent.

21.1.2 Examples

Example 21.1. Simulating and plotting a single normal distribution. $Y \sim n(0, 1)$

```
library(MASS) # for truehist
par(mfcol=c(2,2))
y<-rnorm(1000) # generating 1000 n(0,1)
mn<-mean(y)
vr<-var(y)
truehist(y,ymax=0.5) # plot the histogram
xvec<-seq(-4,4,0.01) # generate the x-axis
yvec<-dnorm(xvec) # theoretical n(0,1) density
lines(xvec,yvec,lwd=2,col="red")
ttl<-paste("Simulation and theory n(0,1)\n",
           "mean=",round(mn,2),
           "and variance=",round(vr,2))
title(ttl)
```

Example 21.2. Sum of two normal distributions.

$$Y_1 \sim n(2, 2^2)$$

and

$$Y_2 \sim n(3, 3^2)$$

```

y1<-rnorm(10000,2,2) # n(2,2^2)
y2<-rnorm(10000,3,3) # n(3, 3^2)
y<-y1+y2
truehist(y)
xvec<-seq(-10,20,0.01)
# check
mn<-mean(y)
vr<-var(y)
cat("The mean is",mn,"\n")
cat("The variance is",vr,"\n")
cat("The standard deviation is",sd(y),"\n")
yvec<-dnorm(xvec,mean=5,sd=sqrt(13)) # n() density
lines(xvec,yvec,lwd=2,col="red")
ttl<-paste("The sum of n(2,2^2) and n(3,3^2)\n",
           "mean=",round(mn,2),
           "and variance=",round(vr,2))
title(ttl)

```

Example 21.3. Sum of nine normal distributions, all with $\mu = 42$ and $\sigma^2 = 2^2$

```

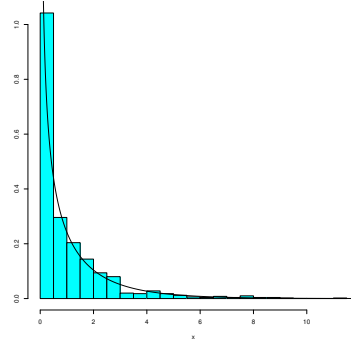
ymat<-matrix(rnorm(10000*9,42,2),ncol=9)
y<-apply(ymat,1,mean)
truehist(y)
# check
mn<-mean(y)
vr<-var(y)
cat("The mean is",mn,"\n")
cat("The variance is",vr,"\n")
cat("The standard deviation is",sd(y),"\n")
# plot the theoretical curve
xvec<-seq(39,45,0.01)
yvec<-dnorm(xvec,mean=5,sd=sqrt(13)) # n() density
lines(xvec,yvec,lwd=2,col="red")
ttl<-paste("The sum of nine n(42^2)\n",
           "mean=",round(mn,2),
           "and variance=",round(vr,2))
title(ttl)

```


21.2 The Chi-square distribution

If $X \sim n(0,1)$, then $Y = X^2$ has a distribution which is called the Chi - square distribution (χ^2) on one degree of freedom. This can be written as:

$$Y \sim \chi^2$$



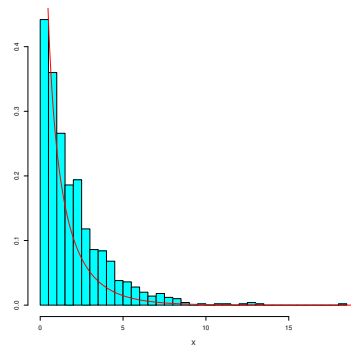
21.2.1 Details

Definition 21.1. If X_1, X_2, \dots, X_n are i.i.d. $N(0, 1)$ then the distribution of $Y = X_1^2 + X_2^2 + \dots + X_n^2$ has a **Chi square (χ^2) distribution**.

21.3 Sum of Chi square Distributions

Let Y_1 and Y_2 be independent variables. If $Y_1 = \chi_{v_1}^2$ and $Y_2 = \chi_{v_2}^2$, then the sum of these two variables also follows a chi-squared (χ^2) distribution

$$Y_1 + Y_2 = \chi_{v_1+v_2}^2$$



21.3.1 Details

Note 21.1. Recall that if

$$X_1, \dots, X_n \sim n(\mu, \sigma^2)$$

are i.i.d., then

$$\sum_{i=1}^n \left(\frac{\bar{X} - \mu}{\sigma} \right)^2 = \sum_{i=1}^n \frac{(\bar{X} - \mu)^2}{\sigma} \sim \chi^2$$

21.4 Sum of squared deviation

If $X_1, \dots, X_n \sim n(\mu, \sigma^2)$ i.i.d, then

$$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi_n^2,$$

but we are often interested in

$$\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi_{n-1}^2.$$

21.4.1 Details

Consider a random sample of Gaussian random variables, i.e. $X_1, \dots, X_n \sim n(\mu, \sigma^2)$ i.i.d. Such a collection of random variables have properties which can be used in a number of ways.

$$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi_n^2,$$

but we are often interested in

$$\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi_{n-1}^2.$$

Note 21.2. A degree of freedom is lost because of subtracting the estimator of the mean as opposed to the true mean.

The correct notation is:

μ = population mean

\bar{X} = sample mean (a random variable)

\bar{x} = sample mean (a number)

21.5 The t-distribution

If $U \sim n(0, 1)$ and $W \sim \chi_v^2$ are independent, then the random variable

$$T = \frac{U}{\sqrt{\frac{W}{v}}}$$

has a distribution which we call the t-distribution on v degrees of freedom denoted $T \sim t_v$.

21.5.1 Details

Definition 21.2. If $U \sim n(0, 1)$ and $W \sim \chi_v^2$ are independent, then the random variable

$$T := \frac{U}{\sqrt{\frac{W}{v}}}$$

has a distribution which we call the **t-distribution** on v degrees of freedom, denoted $T \sim t_v$.

It turns out that if $X_1, \dots, X_n \sim n(\mu, \sigma^2)$ and we set

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

and

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

This follows from \bar{X} and $\sum_{i=1}^n (X_i - \bar{X})^2$ being independent and $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim n(0, 1)$, $\sum \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2$.

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students
This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

22 Estimation, estimates and estimators

22.1 Ordinary least squares for a single mean

If μ is unknown and x_1, \dots, x_n are data, we can estimate μ by finding

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

In this case the resulting estimate is simply

$$\mu = \bar{x}$$

and can easily be derived by setting the derivative to zero.

22.1.1 Examples

Example 22.1. Consider the numbers x_1, \dots, x_5 to be

$$13, 7, 4, 16 \text{ and } 9$$

We can plot $\sum (x_i - \mu)^2$ vs. μ and find the minimum.

22.2 Maximum likelihood estimation

If $(Y_1, \dots, Y_n)'$ is a random vector from a density f_{θ} where θ is an unknown parameter, and \mathbf{y} is a vector of observations then we define the **likelihood function** to be

$$L_{\mathbf{y}}(\theta) = f_{\theta}(\mathbf{y}).$$

22.2.1 Examples

Example 22.2. If, x_1, \dots, x_n are assumed to be observations of independent random variables with a normal distributions and mean of μ and variance of σ^2 , then the joint density is

$$\begin{aligned} & f(x_1) \cdot f(x_2) \cdot \dots \cdot f(x_n) \\ &= \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \cdot \dots \cdot \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}} \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\ &= \frac{1}{(2\pi)^{n/2} \sigma^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i-\mu)^2} \end{aligned}$$

and if we assume σ^2 is known then the likelihood function is

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

Maximizing this is done by maximizing the log, i.e. finding the μ for which:

$$\frac{d}{d\mu} \ln L(\mu) = 0,$$

which again results in the estimate

$$\hat{\mu} = \bar{x}$$

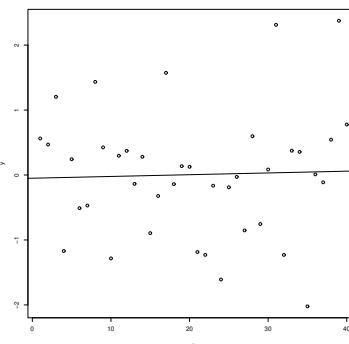
22.2.2 Detail

Definition 22.1. If $(Y_1, \dots, Y_n)'$ is a random vector from a density f_θ where θ is an unknown parameter, and \mathbf{y} is a vector of observations then we define the **likelihood function** to be

$$L_{\mathbf{y}}(\theta) = f_\theta(\mathbf{y}).$$

22.3 Ordinary least squares

Consider the regression problem where we fit a line through (x_i, y_i) pairs with x_1, \dots, x_n fixed numbers but where y_i is measured with error.



Regression line through data pairs.

22.3.1 Details

The ordinary least squares (OLS) estimates of the parameters α and β in the model $y_i = \alpha + \beta x_i + \varepsilon_i$ are obtained by minimizing the sum of squares

$$\sum_i (y_i - (\alpha + \beta x_i))^2$$

$$a = \bar{y} - b\bar{x}$$

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

22.4 Random variables and outcomes

22.4.1 Details

Recall that X_1, \dots, X_n are random variables (reflecting the population distribution) and x_1, \dots, x_n are numerical outcomes of these distributions. We use upper case letters to denote random variables and lower case letters to denote outcome or data.

22.4.2 Examples

Example 22.3. Let the mean of a population be zero and the $\sigma = 4$. Then draw three samples from this population with size, n , either 4, 16 or 64. The sample mean \bar{X} will have a distribution with mean zero and standard deviation of $\frac{\sigma}{\sqrt{n}}$ where $n=4, 16$ or 64 .

22.5 Estimators and estimates

In OLS regression, note that the values of a and b

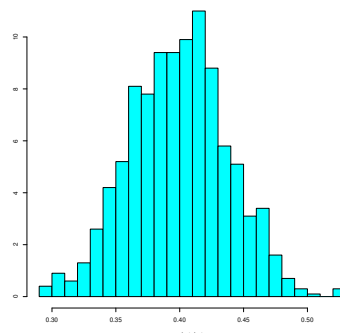
$$a = \bar{y} - b\bar{x}$$

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

are outcomes of random variables e.g. b is the outcome of

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

the estimator which has some distribution.



Shows an example of the distribution of the estimator $\hat{\beta}$

22.5.1 Details

The following R commands can be used to generate a distribution for the estimator $\hat{\beta}$

```
library(MASS)
nsim <- 1000 # replicates
betahat <- NULL
for (i in 1:nsim){
  n <- 20
  x <- seq(1:n) # Fixed x vector
  y <- 2 + 0.4*x + rnorm(n, 0, 1)
  xbar <- mean(x)
  ybar <- mean(y)
  b <- sum((x-xbar)*(y-ybar))/sum((x-xbar)^2)
  a <- ybar - b* xbar
  betahat <- c(betahat, b)
}
truehist(betahat)
```

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students

This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

23 Test of hypothesis, P values and related concepts

23.1 The principle of the hypothesis test

The principle is to formulate a hypothesis and an alternative hypothesis, H_0 and H_a respectively, and then select a statistic with a given distribution when H_0 is true and select a rejection region which has a specified probability (α) when H_0 is true. The rejection region is chosen to reflect H_a , i.e to ensure a high probability of rejection when H_a is true.

23.1.1 Examples

Example 23.1. Suppose we want to evaluate whether a coin is biased. We can plan an experiment for this. Suppose we toss the coin 5 times and count the number of heads. We can test the following hypothesis simply.

$H_0 : p = \frac{1}{2}$ where H_0 is the null hypothesis

$H_a ; p > \frac{1}{2}$ where H_a is an alternative hypothesis
and p is probability of having a head.

We reject H_0 if we get all heads. (Assuming the only interest is in a tendency towards larger probabilities). So the probability of rejecting the null hypothesis H_0 is:

$$P[\text{reject } H_0] = P[\text{all heads in 5 trials}] \equiv p^5$$

$$\text{If } H_0 \text{ is true, then } P[\text{reject } H_0] = \frac{1}{2}$$

Need to choose 5 trials to ensure $\frac{1}{2^5} = \frac{1}{32} < \frac{1}{32} < 0.05$

i.e. The probability of incorrectly rejecting H_0 is less than $\alpha = 0.05$

Example 23.2. Flip a coin to test

$$H_0 : P = \frac{1}{2} \text{ vs } H_a : P \neq \frac{1}{2}$$

Reject, if no heads or all heads are obtained in 6 trials, where the error rate is

$$P[\text{reject } H_0 \text{ when true}] = P[\text{all heads or all tails}]$$

$$= P[\text{all heads}] + P[\text{all tails}]$$

$$= \frac{1}{2^6} + \frac{1}{2^6} = 2 \frac{1}{64} = \frac{1}{32} < 0.05$$

A variation of this test is called the sign test, which is used to test hypothesis of the form, H_0 : true median = 0 using a count of the number of positive values.

23.2 The one sided z test for normal mean

Consider testing

$$H_0 : \mu = \mu_0$$

vs

$$H_a : \mu > \mu_0$$

Where data $x_1 \dots x_n$ are collected as independent observations of $X_1 \dots X_n \sim n(\mu, \sigma^2)$ and σ^2 is known. If H_0 is true, then

$$\bar{x} \sim n\left(\mu_0, \frac{\sigma^2}{n}\right)$$

So,

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim n(0, 1)$$

It follows that,

$$P[Z > z^*] = \alpha$$

Where

$$z^* = z_{1-\alpha}$$

So if the data $x_1 \dots x_n$ are such that,

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z^*$$

Then H_0 is rejected.

23.2.1 Examples

Example 23.3. Consider the following data set: 47, 42, 41, 45, 46.

Suppose we want to test the following hypothesis

$$H_0 : \mu = 42$$

vs

$$H_a : \mu > 42$$

$\sigma = 2$ is given

The mean of the given data set can be calculated as

$$\bar{x} = 44.2$$

we can calculate z by using following equation

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{44.2 - 42}{\frac{2}{\sqrt{5}}}$$

$$z = \frac{2.2}{0.8944} = 2.459$$

$$z^* = 1.645$$

Here

$$z > z^*$$

So H_0 is rejected with $\alpha = 0.05$

23.3 The two-sided z test for a normal mean

$$z := \frac{\bar{x} - \mu_0}{s\sqrt{n}} \sim n(0, 1)$$

23.3.1 Details

Consider testing $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$ based on observation from $\bar{X}_1, \dots, \bar{X} \sim n(\mu, \sigma^2)$ i.i.d. where σ^2 is known. If H_0 is true, then

$$Z := \frac{\bar{x} - \mu_0}{\sigma\sqrt{n}} \sim n(0, 1)$$

and

$$P[|z| > z^*] = \alpha$$

with

$$z^* = z_{1-\alpha/2}$$

We reject H_0 if $|z| > z^*$. If $|z| > z^*$ is not true, then we "Cannot reject the H_0 ".

23.3.2 Examples

Example 23.4. In R, you may generate values to calculate the z value. The command that is generally used is: `quantile`

To illustrate:

```
z<-rnorm(1000,0,1)
quantile(z,c(0.025,0.975))
  2.5% 97.5%
-1.995806 2.009849
```

So, the z value for a two-sided normal mean is $|-1.99|$.

23.4 The one-sided t-test for a single normal mean

Recall that if $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ i.i.d. then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

23.4.1 Details

Recall that if $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ i.i.d. then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

To test the hypothesis $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$ first note that if H_0 is true, then

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

so

$$P[T > t^*] = \alpha$$

if

$$t^* = t_{n-1, 1-\alpha}$$

Hence, we reject H_0 if the data x_1, \dots, x_n results in a value of $t := \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ such that $t > t^*$, otherwise H_0 can not be rejected.

23.4.2 Examples

Example 23.5. Suppose the following data set (12,19,17,23,15,27) comes independently from a normal distribution and we need to test $H_0 : \mu = \mu_0$ vs $H_a : \mu > \mu_0$. Here we have $n = 6, \bar{x} = 18.83, s = 5.46, \mu_0 = 18$ so we obtain

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = 0.37$$

so H_0 cannot be rejected.

In R, t^* is found using `qt(n-1,0.95)` but the entire hypothesis can be tested using

```
t.test(x, alternative="greater", mu=<$\mu_0$>)
```

23.5 Comparing means from normal populations

Suppose data are gathered independently from two normal populations resulting in x_1, \dots, x_n and y_1, \dots, y_m

23.5.1 Details

We know that if

$$X_1, \dots, X_n \sim n(\mu_1, \sigma)$$

$$Y_1, \dots, Y_m \sim n(\mu_2, \sigma)$$

are all independent then

$$\bar{X} - \bar{Y} \sim n(\mu_1 - \mu_2, \frac{\sigma^2}{n} + \frac{\sigma^2}{m})$$

Further,

$$\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2} \sim X_{n-1}^2$$

and

$$\sum_{j=1}^m \frac{(Y_j - \bar{Y})^2}{\sigma^2} \sim X_{m-1}^2$$

so

$$\frac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{j=1}^m (Y_j - \bar{Y})^2}{\sigma^2} \sim X_{n+m-2}^2$$

and it follows that

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S \sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim t_{n+m-2}$$

where

$$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{j=1}^m (Y_j - \bar{Y})^2}{n + m - 2}}$$

consider testing $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$. Hence, if H_0 is true then the observed value

$$t = \frac{\bar{x} - \bar{y}}{S \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

comes from a t-test with $n + m - 2$ df and we reject H_0 if $|t| > t^*$. Here,

$$S = \sqrt{\frac{\sum_i (x_i - \bar{x})^2 + \sum_j (y_j - \bar{y})^2}{n + m - 2}}$$

and $t^* = t_{n+m-2, 1-\alpha}$

23.6 Comparing means from large samples <Ó.L.B.M.>

If X_1, \dots, X_n and Y_1, \dots, Y_m , are all independent (with finite variance) with expected values of μ_1 and μ_2 respectively, and variances of σ_1^2 , and σ_2^2 respectively, then

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0, 1)$$

if the sample sizes are large enough.

This is the central limit theorem.

23.6.1 Details

Another theorem (Slutzky) states that replacing σ_1^2 and σ_2^2 with S_1^2 and S_2^2 will result in the same (limiting) distribution.

It follows that for large samples we can test

$$H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_a : \mu_1 > \mu_2$$

by computing

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}}$$

and reject H_0 if $z > z_{1-\alpha}$.

23.7 The P-value

The p-value of a test is an evaluation of the probability of obtaining results which are as extreme as those observed in the context of the hypothesis.

23.7.1 Examples

Example 23.6. Consider a dataset and the following hypotheses

$$H_0 : \mu = 42$$

vs.

$$H_a : \mu > 42$$

and suppose we obtain

$$z = 2.3$$

We reject H_0 since

$$2.3 > 1.645 + z_{0.95}$$

The p-value is

$$P[Z > 2.3] = 1 - \Phi(2.3)$$

obtained in R using

```
1-pnorm(2.3)
[1] 0.01072411
```

If this had been a two tailed test, then

$$\begin{aligned} P &= P[|Z| > 2.3] \\ &= P[Z < -2.3] + P[Z > 2.3] \\ &= 2 \cdot P[Z > 2.3] \end{aligned}$$

23.8 The concept of significance

23.8.1 Details

Two sample means are statistically *significantly different* if their null hypothesis ($\mu_1 = \mu_2$) can be *rejected*. In this case, one can make the following statements:

- The population means are different.
- The sample means are significantly different.
- $\mu_1 \neq \mu_2$
- \bar{x} is significantly different from \bar{y} .

But one does not say:

- The sample means are different.
- The population means are different with probability 0.95.

Similarly, if the hypothesis $H_0 : \mu_1 = \mu_2$ can not be rejected, we can say:

- There is no significant difference between the sample means.
- We can not reject the equality of population means.
- We can not rule out...

But we can not say:

- The sample means are equal.
- The population means are equal.
- The population means are equal with probability 0.95.

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students
This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

24 Power and sample sizes

24.1 The power of a test

Suppose we have a method to test a null hypothesis against an alternative hypothesis. The test would be "controlled" at some level α , i.e. $P[\text{reject } H_0] \leq \alpha$ whenever H_0 is true.

On the other hand, when H_0 is false one wants $P[\text{reject } H_0]$ to be as high as possible.

If the parameter to be tested is θ and θ_0 is a value within H_0 and θ_a is in H_a then we want $P_{\theta_0}[\text{reject } H_0] \leq \alpha$ and $P_{\theta_a}[\text{reject } H_0]$ as large as possible.

For a general θ we write

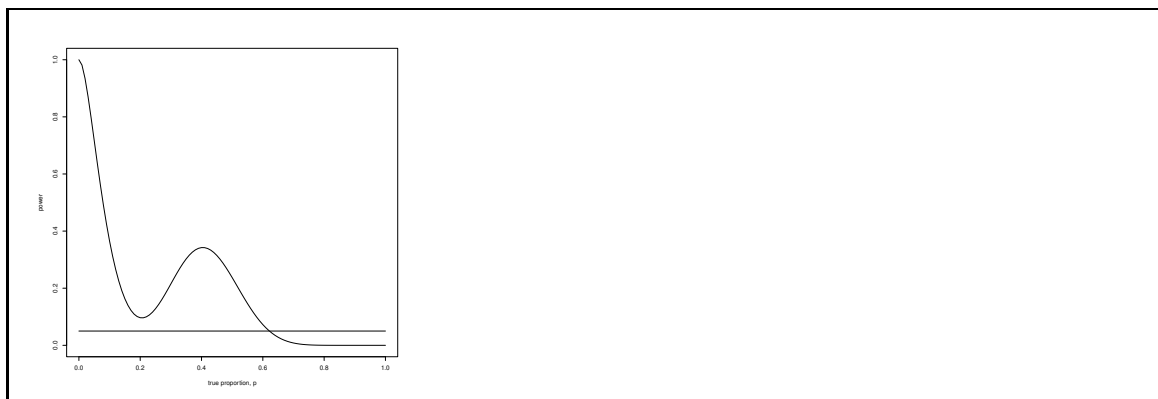
$$\beta(\theta) = P_{\theta}[\text{reject } H_0]$$

for the power of the test

24.1.1 Details

Do not use the phrase "accept".

24.2 The power of tests for proportions



24.2.1 Examples

Example 24.1. Suppose 7 students are involved in an experiment which is comprised of 7 trials and each trial consists of rolling a dice 9 times.

Experiment 1: A student records a 0 if they toss an even number (2,4,6), and records a 1 if they toss an odd number (1,3,5). After tossing the dice 9 times and recording a 0 or 1 the student tabulates the number of 1s. This process is repeated 6 more times.

Data and outcomes: x = number of successes in n trials $= \sum_{i=1}^n$. Thus, x = number of odd numbers

Question: Test whether $p = P[\text{odddnumber}] = \frac{1}{2}$ that is

$$H_0 : p = \frac{1}{2} \text{ vs. } H_a : p \neq \frac{1}{2}$$

Solution: Now, x is an outcome of $X \sim \text{Bin}(n, p)$. We know from the CLT that

$$\frac{X - np}{\sqrt{np(1-p)}} \sim \dot{N}(0, 1)$$

write $p_0 = \frac{1}{2}$ so if $H_0 : p = p_0$ is true then

$$Z := \frac{X - np_0}{\sqrt{np_0(1-p_0)}} \sim \dot{N}(0, 1)$$

so we reject H_0 if the observed value

$$z = \frac{x - np_0}{\sqrt{np_0(1-p_0)}}$$

is such that $|z| > z_{1-\frac{\alpha}{2}}$

Outcomes from 21 trials

7 4 4
 3 4 6
 5 3 4
 5 5 3
 6 4 5
 4 3 5
 3 6 7

$$z = \frac{7 - 9 \cdot \frac{1}{2}}{\sqrt{9 \cdot \frac{1}{2} \cdot \frac{1}{2}}} = \frac{7 - 4.5}{3 \cdot \frac{1}{2}} = \frac{14 - 9}{3} = \frac{5}{3} < 1.96$$

So we do not reject the null hypothesis!

Note 24.1. Note that we can rewrite the test statistics slightly

$$z = \frac{x - \frac{n}{2}}{\sqrt{n \cdot \frac{1}{2} \cdot \frac{1}{2}}} = \frac{x - \frac{9}{2}}{3 \cdot \frac{1}{2}} = \frac{2x - 9}{3}$$

Note 24.2. Note that we reject if $\frac{2x-9}{3} > 1.96$ i.e. if $2x > 9 + 3 \cdot 1.96 \approx 9 + 6 = 15$

$x > 7.5$ [for $x=8$ or 9] or $2x < 9 - 3 \cdot 1.96, x < 1.5$ [for $x=0$ or 1].

Example 24.2. Suppose 7 students are involved in an experiment which is comprised of 7 trials and each trial consists of rolling a dice 9 times.

Experiment 2: The procedure is the same as in experiment 1, but now the student records 0 for a 1 or 2 and a 1 for a 3,4,5, or 6.

Data and outcomes:

x = number of successes in n trials $= \sum_{i=1}^n$ Thus, x = number of 'b's

Solution: Outcomes from 21 experiments

5 4 3

8 5 7

5 7 3

7 6 5

7 8 8

5 6 4

2 5 7

This time our test is $H_0 : p = \frac{2}{3}$ vs $H_a : p = \frac{2}{3}$. Note that we reject H_0 if $\frac{6x-4n}{9} > 1,96$ [for $x=9$] or if $\frac{6x-4n}{9} < -1,96$ [for $x=0,1,2,3$].

We reject H_0 in 3 out of 21 trials.

Example 24.3. Suppose 7 students are involved in an experiment which is comprised of 7 trails and each trial consists of rolling a dice 9 times.

Experiment 3: Same as experiment 1 except 0 is recorded for 1,2,3,4,5 and a 1 is recorded for 6.

Data and outcomes:

x = number of successes in n trials $= \sum_{i=1}^n$ Thus, x = number of '1's

Solution: Outcomes from 21 experiments

0 1 2

1 2 1

1 4 2

1 1 1

1 3 1

1 1 2

0 2 0

With the same kind of calculations as above, we find that we reject the null hypothesis $H_0 : p = \frac{1}{6}$ in 14 out of 21 trials.

24.3 The Power of the one sided z test for the mean

The one sided z-test for the mean (μ) is based on a random sample where $X_1 \dots X_n \sim n(\mu, \sigma^2)$ are independent and σ^2 is known.

The power of the test for an arbitrary μ can be computed as:

$$\beta(\mu) = 1 - \Phi\left(\frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} + z_{1-\alpha}\right)$$

24.3.1 Details

The one sided z-test for the mean (μ) is based on a random sample where $X_1 \dots X_n \sim n(\mu, \sigma^2)$ are independent and σ^2 is known.

If the hypotheses are:

$H_0 : \mu = \mu_0$ vs

$H_a : \mu > \mu_0$

Then we know that, if H_0 is true

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim n(0, 1)$$

Given data x_1, \dots, x_n , the z-value is

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

We reject H_0 if $z > z_{1-\alpha}$

The level of this test is

$$\begin{aligned} P_{\mu_0}[\text{Reject } H_0] &= P_{\mu_0}\left[\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_{1-\alpha}\right] \\ &= P[z > z_{1-\alpha}] = \alpha \end{aligned}$$

since $Z \sim n(0, 1)$ when μ_0 is the true value.

The power of the test for an arbitrary μ can be computed as follows.

$$\begin{aligned} \beta(\mu) &= P_{\mu}[\text{reject } H_0] \\ &= P_{\mu}\left[\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_{1-\alpha}\right] \\ &= P_{\mu}\left[\bar{X} > \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}}\right] \\ &= P_{\mu}\left[\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} > \frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} + z_{1-\alpha}\right] \end{aligned}$$

$$= P\left[Z > \frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} + z_{1-\alpha}\right]$$

We obtain

$$\beta(\mu) = 1 - \Phi\left(\frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} + z_{1-\alpha}\right)$$

24.3.2 Examples

Example 24.4. Suppose we know $\sigma = 2$ and we will take a sample from $n(\mu, \sigma^2)$ intending to test the hypothesis $\mu = 3$ at level $\alpha = 0.05$. We want to know the power against a one-tailed alternative when the true mean is actually $\mu = 4$ when the sample size is $n = 25$.

We can set this up in R with:

```
alpha<-0.05
n<-25
sigma<-2
mu0<-3
mu<-4
zcrit<-qnorm(1-alpha)
```

Sticking the formula into R gives

```
1-pnorm((mu0-mu)/(sigma/sqrt(n))+zcrit)
[1] 0.803765
```

On the other hand, one can also use a simple simulation approach. First, decide how many samples are to be simulated (Nsim). Then, generate all of these samples, arrange them in a matrix and compute the mean of each sample. The z-value of each of these Nsim tests are then computed and a check is made whether it exceeds the critical point (1) or not (0).

```
Nsim<-10000
m<-matrix(rnorm(Nsim*n,mu,sigma),ncol=n)
mn<-apply(m,1,mean)
z<-(mn-mu0)/(sigma/sqrt(n))
i<-ifelse(z>zcrit,1,0)
sum(i/Nsim)
[1] 0.8081
```

24.4 Power and sample size for the one-sided z-test for a single normal mean

Suppose we want to test $H_0 : \mu = \mu_0$ vs $H_a : \mu > \mu_0$. We will reject H_0 if the observed value

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

is such that $z > z_{1-\alpha}$.

24.4.1 Details

Suppose we want to test $H_0 : \mu = \mu_0$ vs $H_a : \mu > \mu_0$. So based on $X_1, \dots, X_n \sim n(\mu, \sigma^2)$ i.i.d. with σ^2 known we will reject H_0 if the observed value

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

is such that $z > z_{1-\alpha}$. The power is given by:

$$\beta(\mu) = 1 - \Phi\left(\frac{\mu - \mu_0}{\sigma/\sqrt{n}} + z_{1-\alpha}\right)$$

and describes the probability of rejecting H_0 when μ is the correct value of the parameter. Suppose we want to reject H_0 with a prespecified probability β_1 , when μ_1 is the true value of μ . For this, we need to select the sample size so that

$$\beta(\mu_1) \geq \beta_1$$

i.e. find n which satisfies

$$1 - \Phi\left(\frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} + z_{1-\alpha}\right) \geq \beta_1$$

24.4.2 Examples

```
Example 24.5. mu0<-10
sigma<-2
mu1<-11
n<-50
d<-(mu1-mu0)

power.t.test(n=n,delta=d,sd=sigma,sig.level=0.05,type="one.sample",
  alternative="one.sided",strict
+ = TRUE)

One-sample t test power calculation

      n = 50
  delta = 1
     sd = 2
sig.level = 0.05
  power = 0.9672067
alternative = one.sided
```

24.5 The non central t - distribution

Recall that if $Z \sim n(0, 1)$ and $U \sim \chi^2_v$ are independent then

$$\frac{Z}{\sqrt{\frac{U}{v}}} \sim t_v$$

and it follows for a random sample $X_1 \dots X_n \sim n(\mu, \sigma^2)$ independent; that

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\sum (X_i - \bar{X})^2}{\frac{\sigma^2}{n-1}}}} \sim t_{n-1}$$

24.5.1 Details

On the other hand, if $W \sim n(\Delta, 1)$ and $U \sim \chi^2_v$ are independent, then $\frac{W}{\sqrt{\frac{U}{v}}}$ has a non central t-distribution with v degrees of freedom and non centrality parameter Δ . This distribution arises, if $X_1 \dots X_n \sim n(\mu, \sigma^2)$ independent and we want to consider the distribution of:

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} + \frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}}}{\frac{s}{\sqrt{n}}} = \frac{Z + \frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{U}{v}}}$$

Where $\mu \neq \mu_0$ which is a non central t with non centrality parameters

$$\Delta = \frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

with $n - 1$ df. Here $v = n - 1$ df since $Z \sim n(0, 1)$ and $U \sim \chi^2_{n-1}$ in this equation

24.6 The power of t-test for a normal mean (warning: errors)

24.6.1 Details

WARNING: This is all wrong and needs to be rewritten

Consider $X_1, \dots, X_n \sim n(\mu, \sigma^2)$ i.i.d. where σ^2 is unknown and we want to test $H_0 : \mu = \mu_0$ vs. $H_a : \mu > \mu_0$. We know that

$$T := \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

and we will reject H_0 if the computed value

$$t := \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

is such that

$$t > t^* = t_{n-1, 1-\alpha}.$$

The power of this test is:

$$\begin{aligned} B(\mu) &= P_{\mu}[\text{reject } H_0] = P_{\mu}\left[\frac{\bar{x} - \mu_0}{s/\sqrt{n}} > t^*\right] \\ &= P_{\mu}[\bar{x} - \mu_0 > t^* \cdot s/\sqrt{n}] \\ &= P_{\mu}\left[\frac{\bar{x} - \mu}{s/\sqrt{n}} > t^* + \frac{\mu_0 - \mu}{s/\sqrt{n}}\right]. \end{aligned}$$

Which is the probability that a $t_{n-1, 1-\alpha}$ -variable exceed $t^* + \frac{\mu_0 - \mu}{s/\sqrt{n}}$.

WARNING: This is all wrong and needs to be rewritten (the s in the above line is a random variable so this make no sense at all)

24.7 Power and sample size for the one sided t-test for a mean

Suppose we want to calculate the power of a one sided t-test for a single mean (one sample), this can easily be done in R with the `power.t.test` command.

24.7.1 Details

$$\Delta = \mu_1 - \mu_2$$

$$\delta = \frac{\mu_1 - \mu_2}{\sigma/\sqrt{n}}$$

24.7.2 Examples

Example 24.6. For a one sided power analysis we wish to test the following hypotheses:

For a one sample test:

$$H_0 : \mu = \mu_0 \text{ vs. } H_a : \mu > \mu_0$$

For a two sample test:

$$H_0 : \mu_1 = \mu_2 \text{ vs. } H_a : \mu_1 > \mu_2$$

In R, the `power.t.test` command is useful to calculate how many samples one needs to obtain a certain power of a test, but also to calculate the power when we have a given number of samples.

Example 24.7. How many samples do I need to get a power of .9?

```
power.t.test(power = .95, delta=1.5,sd=2, type="
  one.sample", alternative = "one.sided")
```

One-sample t test power calculation

```
      n = 20.67702
  delta = 1.5
     sd = 2
sig.level = 0.05
  power = 0.95
alternative = one.sided
```

We would thus need a sample size of $n = 31.15$ or ≈ 32 samples to obtain a power of 0.9 for our analysis.

Example 24.8. With a sample size of $n = 45$, what will the power of my test be?

```
power.t.test(n=45,delta=1.5,sd=2,sig.level=0.05,  
type="one.sample",alternative="one.sided")
```

One-sample t test power calculation

```
      n = 45  
    delta = 1.5  
      sd = 2  
sig.level = 0.05  
  power = 0.9995287  
alternative = one.sided
```

This is done the same way for two samples only by changing the alternative to "two.sample". For two sided power analysis, one only needs to change the alternative to "two.sided".

Example 24.9. If one is interested in doing a power analysis for an ANOVA test, this is done in a fairly similar way.

With a given sample size of n=20:

```
power.anova.test(groups=4, n=20, between.var=1,  
within.var=3)
```

Balanced one-way analysis of variance power calculation

```
  groups = 4  
    n = 20  
between.var = 1  
within.var = 3  
sig.level = 0.05  
  power = 0.9679022
```


To calculate the sample size needed to obtain a power of 0.90 for a test:

```
power.anova.test(groups=4, between.var=1, within  
.var=3, power=.9)
```

Balanced one-way analysis of variance power calculation

```
groups = 4  
n = 15.18834  
between.var = 1  
within.var = 3  
sig.level = 0.05  
power = 0.9
```

24.8 The power of the 2-sided t-test

A power analysis on a two-sided t-test can be done in R using the *power.t.test* command.

24.8.1 Details

For a one sample test:

$H_0 : \mu = \mu_0$ vs. $H_a : \mu \neq \mu_0$

The *power.t.test* command is useful to provide information for determining the minimum sample size one needs to obtain a certain power of a test:

```
power.t.test(n= ,delta= ,sd= ,sig.level= ,power=  
,type=c("two.sample","one.sample","paired"),  
alternative=c("two.sided"))
```

where:

n=sample size

d=effect size

sd=standard deviation

sig.level=significance level

power= normally 0.8, 0.9 or 0.95

type= two sample, one sample or paired (the type selected depends on the research)

alternative= either one sided or two sided

24.8.2 Examples

Example 24.10. How many samples do I need in my research to obtain a power of 0.8?

```
power.t.test(delta=1.5,sd=2,sig.level=0.05,power=0.8,type=c("two.sample"),alternative=c("two.sided"))
```

Two-sample t test power calculation

```
      n = 28.89962
  delta = 1.5
     sd = 2
sig.level = 0.05
  power = 0.8
alternative = two.sided
```

So, one needs 29 samples (n=29) to obtain a power level of 0.8 for this analysis.

24.9 The power of the 2-sample one and two-sided t-tests

The power of a two sample, one-sided t-test can be computed as follows:

$$\beta_{(\mu_1\mu_2)} = P_{\mu_1\mu_2} \left[\frac{Z + \Delta}{\sqrt{U/(n+m-2)}} > t_{1-\alpha, n+m-2}^* \right]$$

and the power of a two sample, two-sided t-test is give by:

$$\beta_{(\mu_1\mu_2)} = P_{\mu_1\mu_2} \left[\frac{Z + \Delta}{\sqrt{U/(n+m-2)}} > t_{1-\alpha, n+m-2}^* \right] + P_{\mu_1\mu_2} \left[\frac{Z + \Delta}{\sqrt{U/(n+m-2)}} < -t_{1-\alpha, n+m-2}^* \right]$$

where $\Delta = \frac{(\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}}$ and U is the SSE.

24.9.1 Details

Two Sample, One-sided t-Test:

Suppose data are gathered independently from two normal populations resulting in

$$X_1, \dots, X_n \sim n(\mu_1, \sigma^2)$$

$$Y_1, \dots, Y_m \sim n(\mu_2, \sigma^2)$$

where all data are independent then

$$\bar{X} - \bar{Y} \sim n\left(\mu_1 - \mu_2, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right)$$

The null hypothesis in question is $H_o : \mu_1 = \mu_2$ versus alternative $H_a : \mu_1 > \mu_2$. If H_o is true then the observed value

$$t = \frac{\bar{x} - \bar{y}}{s \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

comes from a t-distribution with $n + m - 2$ degrees of freedom and we reject H_o if $|t| > t_{1-\alpha, n+m-2}^*$

The power of the test can be computed as follows:

$$\begin{aligned}
\beta_{(\mu_1\mu_2)} &= P_{\mu_1\mu_2}[\text{reject } H_o] \\
&= P_{\mu_1\mu_2} \left[\frac{\bar{X} - \bar{Y}}{S\sqrt{\frac{1}{n} + \frac{1}{m}}} > t_{1-\alpha, n+m-2}^* \right] \\
&= P_{\mu_1\mu_2} \left[\frac{\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} + \frac{(\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}}{S/\sigma} > t_{1-\alpha, n+m-2}^* \right] \\
&= P_{\mu_1\mu_2} \left[\frac{Z + \frac{(\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}}{S/\sqrt{(n+m-2)}} > t_{1-\alpha, n+m-2}^* \right] \\
&= P_{\mu_1\mu_2} \left[\frac{Z + \Delta}{\sqrt{U/(n+m-2)}} > t_{1-\alpha, n+m-2}^* \right]
\end{aligned}$$

where $\Delta = \frac{(\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}$ and U is the SSE of the samples which is divided by the appropriate degrees of freedom to give a χ^2 distribution.

This is the probability that a non-central t -variable exceeds t^* .

Two Sample, Two-sided t-Test:

In this case the null hypothesis is defined as $H_o : \mu_1 = \mu_2$ versus alternative $H_a : \mu_1 \neq \mu_2$.

The power of the test can be computed as follows:

$$\begin{aligned}
\beta_{(\mu_1\mu_2)} &= P_{\mu_1\mu_2} [\text{reject } H_0] \\
&= P_{\mu_1\mu_2} \left[\left| \frac{\bar{X} - \bar{Y}}{S \sqrt{\frac{1}{n} + \frac{1}{m}}} \right| > t_{1-\alpha, n+m-2}^* \right] \\
&= P_{\mu_1\mu_2} \left[\frac{\bar{X} - \bar{Y}}{S \sqrt{\frac{1}{n} + \frac{1}{m}}} > t_{1-\alpha, n+m-2}^* \right] \\
&\quad + P_{\mu_1\mu_2} \left[\frac{\bar{X} - \bar{Y}}{S \sqrt{\frac{1}{n} + \frac{1}{m}}} < -t_{1-\alpha, n+m-2}^* \right] \\
&= P_{\mu_1\mu_2} \left[\frac{\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}} + \frac{(\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}}}{S / \sqrt{(n+m-2)}} > t_{1-\alpha, n+m-2}^* \right] \\
&\quad + P_{\mu_1\mu_2} \left[\frac{\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}} + \frac{(\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}}}{S / \sqrt{(n+m-2)}} < -t_{1-\alpha, n+m-2}^* \right] \\
&= P_{\mu_1\mu_2} \left[\frac{Z + \Delta}{\sqrt{U / (n+m-2)}} > t_{1-\alpha, n+m-2}^* \right] \\
&\quad + P_{\mu_1\mu_2} \left[\frac{Z + \Delta}{\sqrt{U / (n+m-2)}} < -t_{1-\alpha, n+m-2}^* \right]
\end{aligned}$$

where $\Delta = \frac{(\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}}$ and U is the SSE of the samples which is divided by the appropriate degrees of freedom to give a χ^2 distribution.

Note 24.3. Note that the power of a test can be obtained using the *power.t.test* function in R.

24.10 Sample sizes for two-sample one and two-sided t-tests

The sample size should always satisfy the desired power.

24.10.1 Details

Suppose we want to reject the H_0 with a pre-specified probability β_1 when μ_1 and μ_2 are true values of μ . For this, we need to select the sample size n and m so that $\beta(\mu_1\mu_2) \geq \beta_1$ i.e. find n and m which satisfies

$$P_{\mu_1\mu_2} \left[\frac{Z + \Delta}{\sqrt{U/(n+m-2)}} > t_{1-\alpha, n+m-2}^* \right]$$

for a two sample, one-sided t-test.

Similarly for a two sample, two-sided t-test we need to find n and m that satisfies

$$P_{\mu_1\mu_2} \left[\frac{Z + \Delta}{\sqrt{U/(n+m-2)}} > t_{1-\alpha, n+m-2}^* \right] + P_{\mu_1\mu_2} \left[\frac{Z + \Delta}{\sqrt{U/(n+m-2)}} < -t_{1-\alpha, n+m-2}^* \right]$$

24.11 A case study in power

Want to compute power in analysis of covariance:

$$y_{ij} = \mu_i + \beta x_{ij} + \varepsilon_{ij}, \quad i = 1, 2, \quad j = 1, \dots, J,$$

where $\varepsilon_{ij} \sim n(0, \sigma^2)$ are i.i.d.?

This can be done by simulation and can easily be expanded to other cases.

24.11.1 Handout

Example 24.11. If you want to compute a power analysis in analysis of covariance:

$$y_{ij} = \mu_i + \beta x_{ij} + \varepsilon_{ij}, \quad i = 1, 2, \quad j = 1, \dots, J,$$

where $\varepsilon_{ij} \sim n(0, \sigma^2)$ are i.i.d. then use simulation.

To do this one needs to first define the task in more detail, along with what exactly is known and what the assumptions are.

Note 24.4. Note that there are only two groups, with intercepts μ_1 and μ_2 . The "power" will refer to the power of a test for $\mu_1 = \mu_2$, i.e. we want to test whether the group means are equal, correcting for the effect of the continuous variable x .

In principle, the x -values will be either fixed a priori or they may be a random part of the experiment. Here we will assume that the x -values are randomly selected in the range 20-30 (could e.g. be the ages of patients).

Since this is in the planning stage of the experiment, we also have a choice of the sample size within each group. For convenience, the sample sizes are taken to be the same in each group, J so the total number of measurements will be $n = 2J$. We also need to decide at which levels of μ_1 and μ_2 the power is to be computed (but it is really only a function of the difference, $\mu_1 - \mu_2$).

The following pieces of R code can be saved into a file, "ancovapow.r" and then command

```
source("ancovapow.r")
```

can be used to run the whole thing.

The beginning of the command sequence merely consists of comments and definitions of parameter values. These need to be changed for each case separately.

```
#
```

```
# ancovapow.r - power computations for analysis  
  of covariance
```

```

# - one factor, two levels mu0, mu1
# - one covariate x, x0 stores possible values
  from which a random set is chosen
#
# first set values of parameters
#
alpha<-0.05
sigma<-7.5 # the common standard deviation
x0<-20:30 # the set of x values
delta<-10 # the difference in the means
mu0<-0 # the first mean
mu1<-mu0+delta # the second mean
slope<-2.5 # the slope in the ancova
J<-10 # the common sample size per factor level
n<-2*J # the total sample size
Nsim<- 40000 # the number of simulations for
  power computations

```

Rather than head straight for the ancova, start with a simpler case, namely ignoring the covariate (x) and merely doing a regular two-sample, two-tailed t -test. This should be reasonably similar to the ancova power computations anyway.

```

#
# Next do the power computations just for a
  regular two-sided, two-sample t-test
# and use simulation
#
Y1<-matrix(rnorm(J*Nsim,mu0,sigma),ncol=J) #
  Simulate Nsim samples of size J, ea n(mu1,
  sigma^2)
Y2<-matrix(rnorm(J*Nsim,mu1,sigma),ncol=J) #
  Simulate Nsim samples of size J, ea n(mu2,
  sigma^2)

```



```

y1mn<-apply(Y1,1,mean) # compute all the
  simulated y1-means
y2mn<-apply(Y2,1,mean) # compute all the
  simulated y2-means
sy1<-apply(Y1,1,sd) # compute all the simulated
  y1-std.devs
sy2<-apply(Y2,1,sd) # compute all the simulated
  y2-std.devs
s<-sqrt(((J-1)*sy1^2+(J-1)*sy2^2)/(n-2)) #
  compute all the pooled std.devs
t<-(y1mn-y2mn)/(s*sqrt(1/J+1/J)) # compute all
  the Nsim t-statistics
i<-ifelse(abs(t)>qt(1-alpha/2,n-2),1,0) # for ea
  t, compute 1=reject, 0=do not reject
powsim2<-sum(i)/Nsim # the simulated power
cat("The simulated power is ",powsim2,"\n")

```

The above gave the simulated power. In R there is a function to do the same computations and it is worth while to verify the code (and approach) by checking whether these give the same thing:

```

#
# Then compute the exact power for the t-test
#
pow2<-power.t.test(delta=delta,sd=sigma,sig.
  level=alpha,n=J ,type=c("two.sample"),
  alternative=c("two.sided"))
cat("The exact power:\n")
print(pow2)

```

Finally, start setting up the code to do the ancova simulations. Note that for this we need to generate the x-values. In this example it is assumed that the x-values are not under the control of the experimenter but arrive randomly, in the range

from 20 to 30 (could e.g. be the age of participants in an experiment).

```
#  
# Finally compute the power in the ancova - note  
# we already have simulated Y1, Y2-values but  
# have not added the x-part yet  
#  
x1<-matrix(sample(x0,Nsim*J,replace=T),ncol=J) #  
# simulate x-values for y1  
x2<-matrix(sample(x0,Nsim*J,replace=T),ncol=J) #  
# simulate x-values for y2  
Y1<-Y1+slope*x1  
Y2<-Y2+slope*x2  
fulldat<-cbind(Y1,Y2,x1,x2) # a row now contains  
# all y1, then all y2, then all x1, then all x2  
# ; Nsim rows
```

Rather than try to write code to do an ancova, it is natural to use the R function `lm` to do this. The “trick” below is to extract the P-value from the summary command. By defining a “wrapper” function which takes a single line as an argument, it will subsequently be possible to use the “apply” function to extract the P-values using a one-line R command.

```
ancova.pval<-function(onerow){ # extract the  
# ancova p-value for diff in means  
J<-length(onerow)/4  
n<-2*J  
y<-onerow[1:n] # get the y-data from the row  
x<-onerow[(n+1):(2*n)] # get the x-data from  
# the row  
grps<-factor(c(rep(1,J),rep(2,J))) # define the  
# groups  
sm<-summary(lm(y~x+grps)) # fit the ancova
```

```

    model
    pval<-sm$coefficients[3,4] # extract exactly
      the right thing from the summary command-the
      P-value for H0:mu1=mu2
    return(pval)
  }

```

Everything has now been defined so it is possible to compute all the P-values in a single command line:

```

pvec<-apply(fulldat,1,ancova.pval)
i2<-ifelse(pvec<alpha,1,0) # for ea test,
  compute 1=reject, 0=do not reject
ancovapow<-sum(i2)/Nsim # the simulated power
cat("The simulated ancova power is ",ancovapow,"
  \n")

```

When run, this script returns:

The simulated power is 0.803025

The exact power:

Two-sample t test power calculation

```

      n = 10
      delta = 10
      sd = 7.5
      sig.level = 0.05
      power = 0.8049123
      alternative = two.sided

```

NOTE: n is number in *each* group

The simulated ancova power is 0.775175

It is seen that when the x-values are not included in any way (in particular, $\beta = 0$), the power is 80.5%. However, this is

not the correct model in the present situation. Using the above value of β and taking this into account, the power is actually a bit lower or 77.5%.

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students

This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

25 Vectors and Matrix Operations

25.1 Numbers, vectors, matrices

Recall that the set of real numbers is \mathbb{R} and that a vector, $v \in \mathbb{R}^n$ is just an n-tuple of numbers.

Similarly, an $n \times m$ matrix is just a table of numbers, with n rows and m columns and we can write

$$A_{mn} \in \mathbb{R}^{mn}$$

Note that a vector is normally considered equivalent to a $n \times 1$ matrix i.e. we view these as column vectors.

25.1.1 Examples

Example 25.1. In R, a vector can be generated with:

```
X <- 3:6
```

```
X
```

```
[1] 3 4 5 6
```

A matrix can be generated in R as follows,

```
matrix(X)
```

```
 [,1]
```

```
[1,] 3
```

```
[2,] 4
```

```
[3,] 5
```

```
[4,] 6
```

Note 25.1. We note that R distinguishes between vector and matrices.

25.2 Elementary Operations

We can define multiplication of a real number k and a vector $v = (v_1, \dots, v_n)$ by $k \cdot v = (kv_1, \dots, kv_n)$. The sum of two vectors in \mathbb{R}^n , $v = (v_1, \dots, v_n)$ and $u = (u_1, \dots, u_n)$ as the vector $v + u = (v_1 + u_1, \dots, v_n + u_n)$. We can define multiplication of a number and a matrix and the sum of two matrices (of the same sizes) similarly.

25.2.1 Examples

Example 25.2. `A <- matrix(c(1,2,3,4), nr=2, nc=2)`

A

```
      [,1] [,2]
[1,]  1  3
[2,]  2  4
```

`B <- matrix(c(1,0,2,1), nr=2, nc=2)`

B

```
      [,1] [,2]
[1,]  1  2
[2,]  0  1
```

A+B

```
      [,1] [,2]
[1,]  2  5
[2,]  2  5
```

25.3 The tranpose of a matrix

In R, matrices may be constructed using the "matrix" function and the transpose of A , A' , may be obtained in R by using the "t" function:

```
A<-matrix(1:6, nrow=3)
t(A)
```

25.3.1 Details

If A is an $n \times m$ matrix with element a_{ij} in row i and column j , then A' or A^T is the $m \times n$ matrix with element a_{ij} in row j and column i .

25.3.2 Examples

Example 25.3. Consider a vector in R

```
x<-1:4
x
[1] 1 2 3 4
t(x)
      [,1] [,2] [,3] [,4]
[1,] 1 2 3 4
matrix(x)
      [,1]
[1,] 1
[2,] 2
[3,] 3
[4,] 4
t(matrix(x))
      [,1] [,2] [,3] [,4]
[1,] 1 2 3 4
```

Note 25.2. Note that the first solution gives a $1 \times n$ matrix and the second solution gives a $n \times 1$ matrix.

25.4 Matrix multiplication

Matrices A and B can be multiplied together if A is an $n \times p$ matrix and B is an $p \times m$ matrix. The general element c_{ij} of $n \times m$; $C = AB$ is found by pairing the i^{th} row of C with the j^{th} column of B, and computing the sum of products of the paired terms.

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}_{3 \times 2} \begin{bmatrix} 1 & 2 \\ 3 & 3 \end{bmatrix}_{2 \times 2} = \begin{bmatrix} 1 \cdot 1 + 2 \cdot 1 & 1 \cdot 2 + 2 \cdot 3 \\ 3 \cdot 1 + 4 \cdot 1 & 3 \cdot 2 + 4 \cdot 3 \\ 5 \cdot 1 + 6 \cdot 1 & 5 \cdot 2 + 6 \cdot 3 \end{bmatrix} = \begin{bmatrix} 3 & 8 \\ 7 & 18 \\ 11 & 28 \end{bmatrix}_{3 \times 2}$$

25.4.1 Details

Matrices A and B can be multiplied together if A is a $n \times p$ matrix and B is a $p \times m$ matrix. Given the general element c_{ij} of $n \times m$ matrix, $C = AB$ is found by pairing the i^{th} row of C with the j^{th} column of B, and computing the sum of products of the paired terms.

25.4.2 Examples

Example 25.4. Matrices in R

```
A<-matrix(c(1,3,5,2,4,6),3,2)
```

A

```
      [,1] [,2]
```

```
[1,]  1  2
```

```
[2,]  3  4
```

```
[3,]  5  6
```

```
B<-matrix(1,1,2,3)2,2)
```

```
B<-matrix(c(1,1,2,3),2,2)
```

B

```
      [,1] [,2]
```

```
[1,]  1  2
```



```

[2,] 1 3
A%*%B
      [,1] [,2]
[1,] 3  8
[2,] 7 18
[3,] 11 28

```

25.5 More on matrix multiplication

Let A , B , and C be $m \times n$, $n \times l$, and $l \times p$ matrices, respectively. Then we have

$$(AB)C = A(BC).$$

In general, matrix multiplication is not commutative, that is $AB \neq BA$.

We also have

$$(AB)' = B'A'.$$

In particular, $(Av)'(Av) = v'A'Av$, when v is a $n \times 1$ column vector.

More obvious are the rules

1. $A + (B + C) = (A + B) + C$
2. $k(A+B)=kA+kB$
3. $A(B+C)=AB+AC$,

where $k \in \mathbb{R}$ and when the dimensions of the matrices fit.

25.6 Linear equations

25.6.1 Details

Detail:

General linear equations can be written in the form $Ax = b$.

25.6.2 Examples

Example 25.5. The set of equations

$$2x + 3y = 4$$

$$3x + y = 2$$

can be written in matrix formulation as

$$\begin{bmatrix} 2 & 3 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$$

i.e. $A\underline{x} = \underline{b}$ for an appropriate choice of A , \underline{x} and \underline{b}

25.7 The unit matrix

The $n \times n$ matrix

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \dots & 0 & 1 \end{bmatrix}$$

is the identity matrix. This is because if a matrix \mathbf{A} is $n \times n$ then $\mathbf{AI} = \mathbf{A}$ and $\mathbf{IA} = \mathbf{A}$

25.8 The inverse of a matrix

If A is an $n \times n$ matrix and B is a matrix such that

$$BA = AB = I$$

Then B is said to be the inverse of A , written

$$B = A^{-1}$$

Note that if A is an $n \times n$ matrix for which an inverse exists, then the equation $Ax = b$ can be solved and the solution is $x = A^{-1}b$.

25.8.1 Examples

Example 25.6. If matrix A is:

$$\begin{bmatrix} 2 & 3 \\ 3 & 1 \end{bmatrix}$$

then A^{-1} is:

$$\begin{bmatrix} \frac{-1}{4} & \frac{3}{4} \\ \frac{3}{4} & \frac{1}{2} \end{bmatrix}$$

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students

This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

26 Some notes on matrices and linear operators

26.1 The matrix as a linear operator

Let A be an $m \times n$ matrix, the function

$$T_A : \mathbb{R}^n \rightarrow \mathbb{R}^m, T_A(\underline{x}) = A\underline{x},$$

is linear, that is

$$T_A(a\underline{x} + b\underline{y}) = aT_A(\underline{x}) + bT_A(\underline{y})$$

if $\underline{x}, \underline{y} \in \mathbb{R}^n$ and $a, b \in \mathbb{R}$.

26.1.1 Examples

Example 26.1. If $A = \begin{bmatrix} 1 & 2 \end{bmatrix}$ then $T_A(\underline{x}) = x + 2y$ where $\underline{x} = \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2$

Example 26.2. If $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ then $T_A \begin{pmatrix} x \\ y \end{pmatrix} = \begin{bmatrix} y \\ x \end{bmatrix}$

Example 26.3. If $A = \begin{bmatrix} 0 & 2 & 3 \\ 1 & 0 & 1 \end{bmatrix}$ then $T_A \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{bmatrix} 2y + 3z \\ x + z \end{bmatrix}$

Example 26.4. If $T \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x + y \\ 2x - 3y \end{pmatrix}$ then $T(\underline{x}) = A\underline{x}$ if we set $A = \begin{bmatrix} 1 & 1 \\ 2 & -3 \end{bmatrix}$

26.2 Inner products and norms

Assuming x and y are vectors, then we define their inner product by

$$x \cdot y = x_1y_1 + x_2y_2 + \cdots + x_ny_n$$

where $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ and $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$

26.2.1 Details

If $x, y \in \mathbb{R}^n$ are arbitrary (column) vectors, then we define their inner product by

$$x \cdot y = x_1y_1 + x_2y_2 + \cdots + x_ny_n$$

where $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ and $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$.

Note 26.1. Note that we can also view x and y as $n \times 1$ matrices and we see that $x \cdot y = x'y$.

Definition 26.1. The normal length of a vector is defined by $\|x\|^2 = x \cdot x$. It may also be expressed as $\|x\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$.

It is easy to see that for vectors a, b and c we have $(a + b) \cdot c = a \cdot c + b \cdot c$ and $a \cdot b = b \cdot a$.

26.2.2 Examples

Two vectors x and y are said to be orthogonal if $x \cdot y = 0$

Example 26.5. If $x = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$ and $y = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$, then

$$x \cdot y = 3 \cdot 2 + 4 \cdot 1 = 10,$$

and

$$\|x\|^2 = 3^2 + 4^2 = 25,$$

so

$$\|x\| = 5$$

26.3 Orthogonal vectors

Two vectors x and y are said to be orthogonal if $x \cdot y = 0$ denoted $x \perp y$

26.3.1 Details

Definition 26.2. Two vectors x and y are said to be **orthogonal** if $x \cdot y = 0$ denoted $x \perp y$

If $a, b \in \mathbb{R}^n$ then

$$\|a + b\|^2 = a \cdot a + 2a \cdot b + b \cdot b$$

so

$$\|a + b\|^2 = \|a\|^2 + \|b\|^2 + 2\underline{ab}.$$

Note 26.2. Note that if $a \perp b$ then $\|a + b\|^2 = \|a\|^2 + \|b\|^2$, which is Pythagoras' theorem in n dimensions.

26.4 Linear combinations of i.i.d. random variables

Suppose X_1, \dots, X_n are i.i.d. random variables and have mean μ_1, \dots, μ_n and variance σ^2 then the expected value of Y of the linear combination is

$$Y = \sum a_i X_i$$

and if a_1, \dots, a_n are real constants then the mean is:

$$\mu_Y = \sum a_i \mu_i$$

and the variance is:

$$\sigma^2 = \sum a_i^2 \sigma_i^2$$

26.4.1 Examples

Example 26.6. Consider two i.i.d. random variables, Y_1, Y_2 and a specific linear combination of the two, $W = Y_1 + 3Y_2$.

We first obtain

$$E[W] = E[Y_1 + 3Y_2] = E[Y_1] + 3E[Y_2] = 2 + 3 \cdot 2 = 2 + 6 = 8.$$

Similarly, we can first use independence to obtain

$$V[W] = V[Y_1 + 3Y_2] = V[Y_1] + V[3Y_2]$$

and then (recall that $V[aY] = a^2V[Y]$)

$$V[Y_1] + V[3Y_2] = V[Y_1] + 3^2V[Y_2] = 1^2 + 3^2 = 1(4) + 9(4) = 40$$

Normally, we just write this up in a simple sequence

$$V[W] = V[Y_1 + 3Y_2] = V[Y_1] + 3^2V[Y_2] = 1^2 + 3^2 = 1(4) + 9(4) = 40$$

26.5 Covariance between linear combinations of i.i.d random variables

Suppose Y_1, \dots, Y_n are i.i.d., each with mean μ and variance σ^2 and $a, b \in \mathbb{R}^n$. Writing $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$, consider the linear combination $a'Y$ and $b'Y$.

26.5.1 Details

The covariance between random variables U and W is defined by

$$\text{Cov}(U, W) = E[(U - \mu_u)(W - \mu_w)]$$

where

$$\mu_u = E[U], \mu_w = E[W]$$

Now, let $U = a'Y = \sum Y_i a_i$ and $W = b'Y = \sum Y_i b_i$, where Y_1, \dots, Y_n are i.i.d. with mean μ and variance σ^2 , then we get

$$\begin{aligned} \text{Cov}(U, W) &= E[(a'Y - \Sigma a_i \mu)(b'Y - \Sigma b_j \mu)] \\ &= E[(\Sigma a_i Y_i - \Sigma a_i \mu)(\Sigma b_j Y_j - \Sigma b_j \mu)] \end{aligned}$$

and after some tedious (but basic) calculations we obtain

$$\text{Cov}(U, W) = \sigma^2 a \cdot b$$

26.5.2 Examples

Example 26.7. If Y_1 and Y_2 are i.i.d., then

$$\begin{aligned} \text{Cov}(Y_1 + Y_2, Y_1 - Y_2) &= \text{Cov}\left((1, 1) \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}, (1, -1) \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}\right) \\ &= (1, 1) \begin{pmatrix} 1 \\ -1 \end{pmatrix} \sigma^2 \end{aligned}$$

$$= 0$$

and in general, $Cov(\underline{a}'\underline{Y}, \underline{b}'\underline{Y}) = 0$ if $\underline{a} \perp \underline{b}$ and Y_1, \dots, Y_n are independent.

26.6 Random vectors

$Y = (Y_1, \dots, Y_n)$ is a random vector if Y_1, \dots, Y_n are random variables.

26.6.1 Details

Definition 26.3. If $EY_i = \mu_i$ then we typically write

$$E(Y) = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = \mu$$

If $Cov(Y_i, Y_j) = \sigma_{ij}$ and $V[Y_i] = \sigma_{ii} = \sigma_i^2$, then we define the matrix

$$\Sigma = (\sigma_{ij})$$

containing the variances and covariances. We call this matrix the **covariance matrix** of Y , typically denoted $V[Y] = \Sigma$ or $Cov[Y] = \Sigma$.

26.6.2 Examples

Example 26.8. If Y_i, \dots, Y_n are i.i.d., $EY_i = \mu$, $VY_i = \sigma^2$, $a, b \in \mathbb{R}^n$ and $U = a'Y$, $W = b'Y$,

and $T = \begin{bmatrix} U \\ W \end{bmatrix}$

then

$$ET = \begin{bmatrix} \Sigma a_i \mu \\ \Sigma b_i \mu \end{bmatrix}$$

$$VT = \Sigma = \sigma^2 \begin{bmatrix} \Sigma a_i^2 & \Sigma a_i b_i \\ \Sigma a_i b_i & \Sigma b_i^2 \end{bmatrix}$$

Example 26.9. If \underline{Y} is a random vector with mean μ and variance-covariance matrix Σ , then

$$E[a'Y] = a'\mu$$

and

$$V[a'Y] = a'\Sigma a.$$

26.7 Transforming random vectors

Suppose

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

is a random vector with $E\mathbf{Y} = \mu$ and $V\mathbf{Y} = \Sigma$ where the variance-covariance matrix

$$\Sigma = \sigma^2 \mathbf{I}$$

26.7.1 Details

Note that if Y_1, \dots, Y_n are independent with common variance σ^2 then

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \dots & \sigma_{2n} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \dots & \sigma_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \sigma_{n3} & \dots & \sigma_n^2 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 & 0 & \dots & \dots & 0 \\ 0 & \sigma_2^2 & \ddots & 0 & \vdots \\ \vdots & \ddots & \sigma_3^2 & \ddots & \vdots \\ \vdots & 0 & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & \sigma_n^2 \end{bmatrix}$$

$$= \sigma^2 \begin{bmatrix} 1 & 0 & \dots & \dots & 0 \\ 0 & 1 & \ddots & 0 & \vdots \\ \vdots & \ddots & 1 & \ddots & \vdots \\ \vdots & 0 & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & 1 \end{bmatrix} = \sigma^2 \mathbf{I}$$

If A is an $m \times n$ matrix, then

$$E[\mathbf{AY}] = A\boldsymbol{\mu}$$

and

$$V[\mathbf{AY}] = A\boldsymbol{\Sigma}A'$$

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students

This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

27 Ranks and determinants

27.1 The rank of a matrix

The rank of an $n \times p$ matrix, A , is the largest number of columns of A , which are not linearly dependent (i.e. the number of linearly independent columns).

27.1.1 Details

Vectors a_1, a_2, \dots, a_n are said to be linearly dependent if the constant k_1, \dots, k_n exists and are not all zero, such that

$$k_1 \mathbf{a}_1 + k_2 \mathbf{a}_2 + \dots + k_n \mathbf{a}_n = \mathbf{0}$$

Note that if such constants exist, then we can write one of the a 's as a linear combination of the rest, e.g. if $k_1 \neq 0$ then

$$a_1 = \mathbf{c}_1 = -\frac{k_2}{k_1} a_2 - \dots - \frac{k_n}{k_1} a_n$$

It can be shown that the rank of A is the same as the rank of A' i.e. the maximum number of linearly independent rows of A .

Note 27.1. Note that if $\text{rank}(A) = p$, then the columns are linearly independent.

27.1.2 Examples

Example 27.1. If

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

the rank of $A = 2$, since

$$k_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + k_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

if and only if

$$\begin{pmatrix} k_1 \\ k_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

so the columns are linearly independent.

Example 27.2. If

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

the rank of $A = 2$.

Example 27.3. If

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

the rank of $A = 2$, since

$$1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + 0 \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + (-1) \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = 0$$

(and hence the rank can not be more than 2) but

$$k_1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + k_2 \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$

if and only if $k_1 = k_2 = 0$ (and hence the rank must be at least 2).

27.2 The determinant

Recall that for a 2x2 matrix,

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

the inverse of A is

$$A^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

27.2.1 Details

Definition 27.1. The number $ad - bc$ is called the **determinant** of the 2x2 matrix A .

Definition 27.2. Now suppose A is an $n \times n$ matrix. An **elementary product** from the matrix is a product of n terms based on taking exactly one term from each column of row x . Each such term can be written in the form $a_{1j_1} \cdot a_{2j_2} \cdot a_{3j_3} \cdot \dots \cdot a_{nj_n}$ where j_1, \dots, j_n is a permutation of the integers $1, 2, \dots, n$. Each permutation σ of the integers $1, 2, \dots, n$ can be performed by repeatedly interchanging two numbers.

Definition 27.3. A **signed elementary product** is an elementary product with a positive sign if the number of interchanges in the permutation is even but negative otherwise.

The determinant of A , $\det(A)$ or $|A|$ is the sum of all signed elementary products.

27.2.2 Examples

Example 27.4. $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$

then

$$|A| = a_{11}a_{22} - a_{12}a_{21}.$$

Example 27.5. $A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$

$$|A|$$

$= a_{11}a_{22}a_{33}$ This is the identity permutation and has positive sign

$-a_{11}a_{23}a_{32}$ This is the permutation that only interchanges 2 and 3

$-a_{12}a_{21}a_{33}$ Only one interchange

$+a_{12}a_{23}a_{31}$ Two interchanges

$+a_{13}a_{21}a_{32}$ Two interchanges

$-a_{13}a_{22}a_{31}$ Three interchanges

Example 27.6. $A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$

$$|A| = -1$$

Example 27.7. $A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$

$$|A| = 1 \cdot 2 \cdot 3 = 6$$

Example 27.8. $A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 3 & 0 \end{bmatrix}$

$$|A| = 0$$

Example 27.9. $A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 2 \\ 0 & 3 & 0 \end{bmatrix}$

$$|A| = -6$$

Example 27.10. $A = \begin{bmatrix} 2 & 1 \\ 2 & 1 \end{bmatrix}$

$$|A| = 0$$

Example 27.11. $A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{bmatrix}$

$$|A| = 0$$

27.3 Ranks, inverses and determinants

The following statements are true for an $n \times n$ matrix A :

- $\text{rank}(A) = n$
- $\det(A) \neq 0$
- A has an inverse

27.3.1 Details

Suppose A is an $n \times n$ matrix. Then the following are truths:

- $\text{rank}(A) = n$
- $\det(A) \neq 0$
- A has an inverse

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students

This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

28 Multivariate calculus

28.1 Vector functions of several variables

A vector-valued function of several variables is a function

$$f : \mathbb{R}^m \rightarrow \mathbb{R}^n$$

i.e. a function of m dimensional vectors, which returns n dimensional vectors.

28.1.1 Examples

Example 28.1. A real valued function of many variables: $f : \mathbb{R}^3 \rightarrow \mathbb{R}$, $f(x_1, x_2, x_3) = 2x_1 + 3x_2 + 4x_3$.

Note 28.1. Note that f is linear and $f(x) = Ax$ where $x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$ and $A = [2 \ 3 \ 4]$.

Example 28.2. Let

$$f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$$

where:

$$f(x_1, x_2) = \begin{pmatrix} x_1 + x_2 \\ x_1 - x_2 \end{pmatrix}$$

Note 28.2. Note that $f(x) = Ax$, where $A = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$.

Example 28.3. Let

$$f : \mathbb{R}^3 \rightarrow \mathbb{R}^4$$

be defined by

$$f(x) = \begin{pmatrix} x_1 + x_2 \\ x_1 - x_3 \\ y - z \\ x_1 + x_2 + x_3 \end{pmatrix}$$

Note 28.3. Note that:

$$f(x) = Ax$$

where

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \\ 1 & 1 & 1 \end{bmatrix}$$

Example 28.4. These multi-dimensional functions do not have to be linear, for example the function $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$

$$f(x) = \begin{pmatrix} x_1 x_2 \\ x_1^2 + x_2^2 \end{pmatrix},$$

is obviously not linear.

28.2 The gradient

Suppose the real valued function $f : \mathbb{R}^m \rightarrow \mathbb{R}$ is differentiable in each coordinate. Then the gradient of f , denoted ∇f is given by

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_m} \right).$$

28.2.1 Details

Definition 28.1. Suppose the real valued function $f : \mathbb{R}^m \rightarrow \mathbb{R}$ is differentiable in each coordinate. Then the **gradient** of f , denoted ∇f is given by

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_m} \right),$$

where each partial derivative $\frac{\partial f}{\partial x_i}$ is computed by differentiating f with respect to that variable, regarding the others as fixed.

28.2.2 Examples

Example 28.5.

$$f(\underline{x}) = x^2 + y^2 + 2xy; \quad \frac{\partial f}{\partial x} = 2x + 2y, \quad \frac{\partial f}{\partial y} = 2y + 2x, \quad \nabla f = (2x + 2y, 2y + 2x)$$

Example 28.6.

$$f(\underline{x}) = x_1 - x_2; \quad \nabla f = (1, -1)$$

28.3 The Jacobian

Now consider a function $f : \mathbb{R}^m \rightarrow \mathbb{R}^n$. Write f_i for the i^{th} coordinate of f , so we can write $f(x) = (f_1(x), f_2(x), \dots, f_n(x))$, where $x \in \mathbb{R}^m$. If each coordinate function f_i is differentiable in each variable we can form the *Jacobian matrix* of f :

$$\begin{pmatrix} \nabla f_1 \\ \vdots \\ \nabla f_n \end{pmatrix}.$$

28.3.1 Details

Now consider a function $f: \mathbb{R}^m \rightarrow \mathbb{R}^n$. Write f_i for the i^{th} coordinate of f , so we can write $f(x) = (f_1(x), f_2(x), \dots, f_n(x))$, where $x \in \mathbb{R}^m$. If each coordinate function f_i is differentiable in each variable we can form the *Jacobian matrix* of f :

$$\begin{pmatrix} \nabla f_1 \\ \vdots \\ \nabla f_n \end{pmatrix}.$$

In this matrix, the element in the i^{th} row and j^{th} column is $\frac{\partial f_i}{\partial x_j}$.

28.3.2 Examples

Example 28.7. For the function

$$f(x, y) = \begin{pmatrix} x^2 + y \\ xy \\ x \end{pmatrix} = \begin{pmatrix} f_1(x, y) \\ f_2(x, y) \\ f_3(x, y) \end{pmatrix},$$

the Jacobian matrix of f is the matrix

$$J = \begin{bmatrix} \nabla f_1 \\ \nabla f_2 \\ \nabla f_3 \end{bmatrix} = \begin{bmatrix} 2x & 2y \\ y & x \\ 1 & 0 \end{bmatrix}.$$

28.4 Univariate integration by substitution

If f is a continuous function and g is strictly increasing and differentiable then,

$$\int_{g(a)}^{g(b)} f(x) dx = \int_a^b f(g(t))g'(t) dt$$

28.4.1 Details

If f is a continuous function and g is strictly increasing and differentiable then,

$$\int_{g(a)}^{g(b)} f(x)dx = \int_a^b f(g(t))g'(t)dt$$

It follows that if X is a continuous random variable with density f and $Y = h(X)$ is a function of X that has the inverse $g = h^{-1}$, so $X = g(Y)$, then the density of Y is given by,

$$f_Y(y) = f(g(y))g'(y)$$

This is a consequence of

$$P[Y \leq b] = P[g(Y) \leq g(b)] = P[X \leq g(b)] = \int_{-\infty}^{g(b)} f(x)dx = \int_{-\infty}^b f(g(y))g'(y)dy$$

28.5 Multivariate integration by substitution

Suppose f is a continuous function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a one-to-one function with continuous partial derivatives. Then if $U \subseteq \mathbb{R}^n$ is a subset,

$$\int_{g(U)} f(\mathbf{x})d\mathbf{x} = \int_U (f(g(\mathbf{y}))|J|)d\mathbf{y}$$

where J is the Jacobian matrix and $|J|$ is the absolute value of its determinant.

$$J = \left| \begin{bmatrix} \frac{\partial g_1}{\partial y_1} & \frac{\partial g_1}{\partial y_2} & \cdots & \frac{\partial g_1}{\partial y_n} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial g_n}{\partial y_1} & \frac{\partial g_n}{\partial y_2} & \cdots & \frac{\partial g_n}{\partial y_n} \end{bmatrix} \right| = \left| \begin{bmatrix} \nabla g_1 \\ \vdots \\ \nabla g_n \end{bmatrix} \right|$$

28.5.1 Details

Suppose f is a continuous function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a one-to-one function with continuous partial derivatives. Then if $U \subseteq \mathbb{R}^n$ is a subset,

$$\int_{g(U)} f(\mathbf{x}) d\mathbf{x} = \int_U (g(\mathbf{y})) |J| d\mathbf{y}$$

where J is the Jacobian determinant and $|J|$ is its absolute value.

$$J = \left| \begin{bmatrix} \frac{\partial g_1}{\partial y_1} & \frac{\partial g_1}{\partial y_2} & \cdots & \frac{\partial g_1}{\partial y_n} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial g_n}{\partial y_1} & \frac{\partial g_n}{\partial y_2} & \cdots & \frac{\partial g_n}{\partial y_n} \end{bmatrix} \right| = \left| \begin{bmatrix} \nabla g_1 \\ \vdots \\ \nabla g_n \end{bmatrix} \right|$$

Similar calculations as in 28.4 give us that if \mathbf{X} is a continuous multivariate random variable, $\mathbf{X} = (X_1, \dots, X_n)'$ with density f and $\mathbf{Y} = \mathbf{h}(\mathbf{X})$, where \mathbf{h} is 1-1 with inverse $\mathbf{g} = \mathbf{h}^{-1}$. So, $\mathbf{X} = \mathbf{g}(\mathbf{Y})$, then the density of \mathbf{Y} is given by;

$$f_Y(\mathbf{y}) = f(\mathbf{g}(\mathbf{y})) |J|$$

28.5.2 Examples

Example 28.8. If $\mathbf{Y} = A\mathbf{X}$ where A is an $n \times n$ matrix with $\det(A) \neq 0$ and $X = (X_1, \dots, X_n)'$ are i.i.d. random variables, then we have the following results:

The joint density of $X_1 \cdots X_n$ is the product of the individual (marginal) densities,

$$f_X(\mathbf{x}) = f(x_1)f(x_2) \cdots f(x_n)$$

The matrix of partial derivatives corresponds to $\frac{\partial \mathbf{g}}{\partial \mathbf{y}}$ where $\mathbf{X} = \mathbf{g}(\mathbf{Y})$, i.e. these are the derivatives of the transformation: $\mathbf{X} = \mathbf{g}(\mathbf{Y}) = A^{-1}\mathbf{Y}$, or $\mathbf{X} = B\mathbf{Y}$ where $B = A^{-1}$.

But if $\mathbf{X} = B\mathbf{Y}$, then

$$X_i = b_{i1}y_1 + b_{i2}y_2 + \cdots + b_{ij}y_j + \cdots + b_{in}y_n$$

So, $\frac{\partial x_i}{\partial y_j} = b_{ij}$ and thus,

$$J = \left| \frac{\partial d\mathbf{x}}{\partial d\mathbf{y}} \right| = |B| = |A^{-1}| = \frac{1}{|A|}$$

The density of \mathbf{Y} is therefore;

$$f_Y(\mathbf{y}) = f_X(g(\mathbf{y}))|J| = f_X(A^{-1}\mathbf{y})|A^{-1}|$$

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students

This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

29 The multivariate normal distribution and related topics

29.1 Transformations of random variables

Recall that if X is a vector of continuous random variables with a joint probability density function and if $Y = h(X)$ such that h is a 1-1 function and continuously differentiable with inverse g so $X = g(Y)$, then the density of Y is given by

$$f_Y(y) = f(g(y))|J|$$

29.1.1 Details

J is the Jacobian determinant of g . In particular if $Y = AX$ then

$$f_Y(y) = f(A^{-1}y)|\det(A^{-1})|$$

if A has an inverse.

29.2 The multivariate normal distribution

29.2.1 Details

Consider i.i.d. random variables, $Z_1, \dots, Z_n \sim (0, 1)$, written

$\underline{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix}$ and let $\underline{Y} = A\underline{Z} + \underline{\mu}$ where A is an invertible $n \times n$

matrix and $\underline{\mu} \in \mathbb{R}^n$ is a vector, so $\underline{Z} = A^{-1}(\underline{Y} - \underline{\mu})$.

Then the p.d.f. of Y is given by

$$f_{\underline{Y}}(\underline{y}) = f_{\underline{Z}}(A^{-1}(\underline{y} - \underline{\mu}))|\det(A^{-1})|$$

But the joint p.d.f. of \underline{Z} is the product of the p.d.f.'s of Z_1, \dots, Z_n , so $f_{\underline{Z}}(\underline{z}) = f(z_1) \cdot f(z_2) \cdot \dots \cdot f(z_n)$ where

$$f(z_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z_i^2}{2}}$$

and hence

$$\begin{aligned}
f_{\underline{Z}}(\underline{z}) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{z_i^2}{2}} \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n z_i^2} \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\underline{z}'\underline{z}}
\end{aligned}$$

since

$$\sum_{i=1}^n z_i^2 = \|\underline{z}\|^2 = \underline{z} \cdot \underline{z} = \underline{z}'\underline{z}$$

The joint p.d.f. of \underline{Y} is therefore

$$\begin{aligned}
f_{\underline{Y}}(\underline{y}) &= f_{\underline{Z}}(A^{-1}(\underline{y} - \underline{\mu})) |det(A^{-1})| \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}(A^{-1}(\underline{y} - \underline{\mu}))'(A^{-1}(\underline{y} - \underline{\mu}))} \frac{1}{|det(A)|}
\end{aligned}$$

We can write $det(AA') = det(A)^2$ so $|det(A)| = \sqrt{det(AA')}$ and if we write $\Sigma = AA'$, then

$$|det(A)| = |\Sigma|^{\frac{1}{2}}$$

Also, note that

$$(A^{-1}(\underline{y} - \underline{\mu}))'(A^{-1}(\underline{y} - \underline{\mu})) = (\underline{y} - \underline{\mu})'(A^{-1})'A^{-1}(\underline{y} - \underline{\mu}) = (\underline{y} - \underline{\mu})'\Sigma^{-1}(\underline{y} - \underline{\mu})$$

We can now write

$$f_{\underline{Y}}(\underline{y}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\underline{y} - \underline{\mu})\Sigma^{-1}(\underline{y} - \underline{\mu})}$$

This is the density of the multivariate normal distribution.

Note that

$$E[\underline{Y}] = \underline{\mu}$$

$$V[\underline{Y}] = V[A\underline{Z}] = AV[\underline{Z}]A' = AIA' = \Sigma$$

Notation: $\underline{Y} \sim n(\underline{\mu}, \Sigma)$

29.3 Univariate normal transforms

The general univariate normal distribution with density

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

is a special case of the multivariate version.

29.3.1 Details

Further, if $Z \sim n(0, 1)$, then clearly $X = aZ + \mu \sim n(\mu, \sigma^2)$ where $\sigma^2 = a^2$

29.4 Transforms to lower dimensions

If $Y \sim n(\mu, \Sigma)$ is a random vector of length n and A is an $m \times n$ matrix of rank $m \leq n$, then $AY \sim n(A\mu, A\Sigma A')$.

29.4.1 Details

If $Y \sim n(\mu, \Sigma)$ is a random vector of length n and A is an $m \times n$ matrix of rank $m \leq n$, then $AY \sim n(A\mu, A\Sigma A')$.

To prove this, set up an $(n - m) \times n$ matrix, B , so that the $n \times n$ matrix, C , formed from combining the rows of A and B is of full rank n . Then it is easy to derive the density of CY which also factors nicely into a product, only one of which contains AY , which gives the density for AY .

29.5 The OLS estimator

Suppose $Y \sim n(X\beta, \sigma^2 I)$. The ordinary least squares estimator, when the $n \times p$ matrix is of full rank, p , where $p \leq n$, is:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

The random variable which describes the process giving the data and estimate is:

$$b = (X'X)^{-1}X'Y$$

It follows that

$$\hat{\beta} \sim n(\beta, \sigma^2(X'X)^{-1})$$

29.5.1 Details

Suppose $Y \sim n(X\beta, \sigma^2 I)$. The ordinary least squares estimator, when the $n \times p$ matrix is of full rank, p , is:

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

The equation below is the random variable which describes the process giving the data and estimate:

$$b = (X'X)^{-1}X'Y$$

If $B = (X'X)^{-1}X'$, then we know that

$$BY \sim n(BX\beta, B(\sigma^2 I)B')$$

Note that

$$BX\beta = (X'X)^{-1}X'X\beta = \beta$$

and

$$\begin{aligned} B(\sigma^2 I)B' &= \sigma(X'X)^{-1}X'[(X'X)^{-1}X']' \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned}$$

It follows that

$$\hat{\beta} \sim n(\beta, \sigma^2(X'X)^{-1})$$

Note 29.1. The earlier results regarding the multivariate Gaussian distribution also show that the vector of parameter estimates will be Gaussian even if the original Y -variables are not independent.

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students

This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

30 Independence, expectations and the moment generating function

30.1 Independent random variables

Recall that two events, A and B , are independent if,

$$P[A \cap B] = P[A]P[B]$$

Since the conditional probability of A given B is defined by:

$$P[A|B] = \frac{P[A \cap B]}{P[B]}$$

We see that A and B are independent if and only if

$$P[A|B] = P[A] \text{ (when } P[B] > 0 \text{)}$$

Two continuous random variables, X and Y , are similarly independent if,

$$P[X \in A, Y \in B] = P[X \in A]P[Y \in B]$$

30.1.1 Details

Two continuous random variables, X and Y , are similarly independent if,

$$P[X \in A, Y \in B] = P[X \in A]P[Y \in B]$$

Now suppose X has p.d.f. f_X and Y has p.d.f. f_Y . Then,

$$P[X \in A] = \int_A f_X(x) dx$$

$$P[Y \in B] = \int_B f_Y(y) dy$$

So X and Y are independent if:

$$P[X \in A, Y \in B] = \int_A f_X(x) dx \int_B f_Y(y) dy$$

$$\begin{aligned}
&= \int_A f_X(x) \left(\int_B f_Y(y) dy \right) dx \\
&= \int_A \int_B f_X(x) f_Y(y) dy dx
\end{aligned}$$

But, if f is the joint density of X and Y then we know that

$$P[X \in A, Y \in B]$$

$$\int_A \int_B f(x, y) dy dx$$

Hence X and Y are independent if and only if we can write the joint density in the form of,

$$f(x, y) = f_X(x) f_Y(y)$$

30.2 Independence and expected values

If X and Y are independent random variables then $E[XY] = E[X]E[Y]$.

Further, if X and Y are independent random variables then $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$ is true if g and h are functions in which expectations exist.

30.2.1 Details

If X and Y are random variables with a joint distribution function $f(x, y)$, then it is true that for $h : \mathbb{R}^2 \rightarrow \mathbb{R}$ we have

$$E[h(X, Y)] = \int \int h(x, y) f(x, y) dx dy$$

for those h such that the integral on the right exists.

Suppose X and Y are independent continuous r.v., then

$$f(x,y) = f_X(x)f_Y(y)$$

Thus,

$$\begin{aligned} E[XY] &= \int \int xyf(x,y)dxdy \\ &= \int \int xyf_X(x)f_Y(y)dxdy \\ &= \int xf_X(x)dx \int yf_Y(y)dy \\ &= E[X]E[Y] \end{aligned}$$

Note 30.1. Note that if X and Y are independent then $E[h(X)g(Y)] = E[h(X)]E[g(Y)]$ is true whenever the functions h and g have expected values.

30.2.2 Examples

Example 30.1. Suppose $X, Y \in U(0, 2)$ are i.i.d then,

$$f_X(x) = \begin{cases} \frac{1}{2} & \text{if } 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

and similarly for f_Y .

Next, note that,

$$f(x,y) = f_X(x)f_Y(y) = \begin{cases} \frac{1}{4} & \text{if } 0 \leq x,y \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

Also note that $f(x,y) \geq 0$ for all $(x,y) \in \mathbb{R}^2$ and

$$\int \int f(x,y)dxdy = \int_0^2 \int_0^2 \frac{1}{4}dxdy = \frac{1}{4} \cdot 4 = 1$$

It follows that,

$$E[XY] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x,y)dxdy$$

$$\begin{aligned}
&= \int_{y=0}^2 \int_{x=0}^2 xy \cdot \frac{1}{4} dx dy \\
&= \int_{y=0}^2 \left(\int_{x=0}^2 xy \frac{1}{4} dx \right) dy \\
&= \int_{y=0}^2 \left[\frac{1}{4}y \cdot \frac{1}{2}x^2 \right]_{x=0}^2 dy \\
&= \int_{y=0}^2 \frac{1}{4}y \left(\frac{1}{2} \cdot 2^2 - \frac{1}{2} \cdot 0 \right) dy \\
\int_0^2 \frac{2}{4}y dy &= \int_0^2 \frac{1}{2}y dy = \frac{1}{2} \cdot \frac{1}{2}y^2 \Big|_0^2 = \frac{1}{4} \cdot 2^2 = 1
\end{aligned}$$

But

$$E[X] = E[Y] = \int_{y=0}^2 x \cdot \frac{1}{2} dx = 1$$

So

$$E[XY] = E[X]E[Y]$$

30.3 Independence and the covariance

If X and Y are independent then $Cov(X, Y) = 0$.

In fact, if X and Y are independent then $Cov(h(X), g(Y)) = 0$ for any functions g and h in which expected values exist.

30.4 The moment generating function

If X is a random variable we define the moment generating function when t exists as: $M(t) := E(e^{tX})$.

30.4.1 Examples

Example 30.2. If $X \sim b(n, p)$ then $M(t) = \sum_{x=0}^n e^{tx} p(x) = \sum_{x=0}^n e^{tx} \binom{n}{x} p \cdot (1-p)^{n-x}$

30.5 Moments and the moment generating function

If $M_X(t)$ is the moment generating function (mgf) of X , then $M_X^{(n)}(0) = E[X^n]$.

30.5.1 Details

Observe that $M(t) = E[e^{tX}] = E[1 + X + \frac{(tX)^2}{2!} + \frac{(tX)^3}{3!} + \dots]$ since $e^a = 1 + a + \frac{a^2}{2!} + \frac{a^3}{3!} + \dots$. If the random variable $e^{|tX|}$ has a finite expected value then we can switch the sum and the expected value to obtain:

$$M(t) = E\left[\sum_{n=0}^{\infty} \frac{(tX)^n}{n!}\right] = \sum_{n=0}^{\infty} \frac{E[(tX)^n]}{n!} = \sum_{n=0}^{\infty} t^n \frac{E[X^n]}{n!}$$

This implies that the n^{th} derivative of $M(t)$ evaluated at $t = 0$ is exactly $E[X^n]$

30.6 The moment generating function of a sum of random variables

$M_{X+Y}(t) = M_X(t) \cdot M_Y(t)$ if X and Y are independent.

30.6.1 Details

Let X and Y be independent random variables, then

$$M_{X+Y}(t) = E[e^{Xt+Yt}] = E[e^{Xt} e^{Yt}] = E[e^{Xt}]E[e^{Yt}] = M_X(t)M_Y(t)$$

30.7 Uniqueness of the moment generating function

Moment generating functions (m.g.f.) uniquely determine the probability distribution function for random variables. Thus, if two random variables have the same m.g.f, then they must also have the same distribution.

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students

This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

31 The gamma distribution

31.1 The gamma distribution

If a random variable X has the density

$$f(x) = \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\beta^\alpha}$$

where $x > 0$ for some constants $\alpha, \beta > 0$, then X is said to have a gamma distribution.

31.1.1 Details

The function Γ is basically chosen so that f integrates to one, i.e.

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$$

It is not too hard to see that $\Gamma(n) = (n-1)!$ if $n \in \mathbb{N}$. Also, $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ for all $\alpha > 0$.

31.2 The mean, variance and mgf of the gamma distribution

Suppose $X \sim G(\alpha, \beta)$ i.e. X has density

$$f(x) = \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha}, x > 0$$

Then,

$$E[X] = \alpha\beta$$

$$M(t) = (1 - \beta t)^{-\alpha}$$

$$V[X] = \alpha\beta^2$$

31.2.1 Details

The expected value of X can be computed as follows:

$$\begin{aligned}
E[X] &= \int_{-\infty}^{\infty} xf(x)dx \\
&= \int_0^{\infty} x \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}} dx \\
&= \frac{\Gamma(\alpha+1)\beta^{\alpha+1}}{\Gamma(\alpha)\beta^{\alpha}} \int_0^{\infty} \frac{x^{(\alpha+1)-1} e^{-x/\beta}}{\Gamma(\alpha+1)\beta^{\alpha+1}} dx \\
&= \frac{\alpha\Gamma(\alpha)\beta^{\alpha+1}}{\Gamma(\alpha)\beta^{\alpha}}
\end{aligned}$$

so $E[X] = \alpha\beta$.

Next, the m.g.f.is given by

$$\begin{aligned}
E[e^{tX}] &= \int_0^{\infty} e^{tx} \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}} dx \\
&= \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \int_0^{\infty} x^{\alpha-1} e^{tx-x/\beta} dx \\
&= \frac{\Gamma(\alpha)\phi^{\alpha}}{\Gamma(\alpha)\beta^{\alpha}} \int_0^{\infty} \frac{x^{(\alpha-1)} e^{-x/\phi}}{\Gamma(\alpha)\phi^{\alpha}} dx
\end{aligned}$$

if we choose ϕ so that $\frac{-x}{\phi} = tx - x/\beta$ i.e. $\frac{-1}{\phi} = t - \frac{1}{\beta}$ i.e. $\phi = -\frac{1}{t-1/\beta} = \frac{\beta}{1-\beta t}$ then we have

$$\begin{aligned}
M(t) &= \left(\frac{\phi}{\beta}\right)^{\alpha} \\
&= \left(\frac{\beta/(1-\beta t)}{\beta}\right)^{\alpha} \\
&= \frac{1}{(1-\beta t)^{\alpha}}
\end{aligned}$$

or $M(t) = (1 - \beta t)^{-\alpha}$. It follows that

$$M'(t) = (-\alpha)(1 - \beta t)^{-\alpha-1}(-\beta) = \alpha\beta(1 - \beta t)^{-\alpha-1}$$

so $M'(0) = \alpha\beta$. Further,

$$\begin{aligned}M''(t) &= \alpha\beta(-\alpha-1)(1-\beta t)^{-\alpha-2}(-\beta) \\ &= \alpha\beta^2(\alpha+1)(1-\beta t)^{-\alpha-2}\end{aligned}$$

$$\begin{aligned}E[X^2] &= M''(0) \\ &= \alpha\beta^2(\alpha+1) \\ &= \alpha^2\beta^2 + \alpha\beta^2\end{aligned}$$

Hence,

$$\begin{aligned}V[X] &= E[X]^2 - E[X]^2 \\ &= \alpha^2\beta^2 + \alpha\beta^2 - (\alpha\beta)^2 \\ &= \alpha\beta^2\end{aligned}$$

31.3 Special cases of the gamma distribution: The exponential and chi-squared distributions

Consider the gamma density,

$$f(x) = \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\beta^\alpha}, x > 0$$

For parameters $\alpha, \beta > 0$.

If $\alpha = 1$ then

$$f(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}}, x > 0$$

and this is the density of exponential distribution.

Consider next the case $\alpha = \frac{\nu}{2}$ and $\beta = 2$ where ν is an integer, so the density becomes,

$$f(x) = \frac{x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}}{\Gamma(\frac{\nu}{2})2^{\frac{\nu}{2}}}, x > 0$$

This is the density of a chi-squared random variable with ν degrees of freedom.

31.3.1 Details

Consider, $\alpha = \frac{\nu}{2}$ and $\beta = 2$ where ν is an integer, so the density becomes,

$$f(x) = \frac{x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}}{\Gamma(\frac{\nu}{2})2^{\frac{\nu}{2}}}, x > 0$$

This is the density of a chi - squared random variable with ν degrees of freedom.

This is easy to see by starting with $Z \sim n(0, 1)$ and defining $W = Z^2$ so that the c.d.f. is:

$$\begin{aligned}
H(w) &= P[W \leq w] = P[Z^2 \leq w] \\
&= P[-\sqrt{w} \leq Z \leq \sqrt{w}] \\
&= 1 - P[|Z| > \sqrt{w}] \\
&= 1 - 2p[Z < -\sqrt{w}] \\
&= 1 - 2 \int_{-\alpha}^{\sqrt{w}} \frac{e^{-t^2}}{\sqrt{2w}} dt = 1 - 2\phi(\sqrt{w})
\end{aligned}$$

The p.d.f. of w is therefore,

$$\begin{aligned}
h(w) &= H'(w) \\
&= 0 - 2\phi'(\sqrt{w}) \frac{1}{2} w^{\frac{1}{2}-1}
\end{aligned}$$

but

$$\phi(x) = \int_{-\alpha}^x \frac{e^{-t^2}}{2\Pi} dt; \phi'(x) = \frac{d}{dx} \int_{-\alpha}^x \frac{e^{-t^2}}{2\Pi} dt = \frac{e^{-x^2}}{2\Pi}$$

So

$$\begin{aligned}
h[w] &= -2 \frac{e^{-w}}{2\Pi} \cdot \frac{1}{2} \cdot w^{\frac{1}{2}-1} \\
h[w] &= \frac{w^{\frac{1}{2}-1} e^{-w}}{2\Pi}, w > 0
\end{aligned}$$

We see that we must have $h = f$ with $\nu = 1$. We have also shown $\Gamma(\frac{1}{2})2^{\frac{1}{2}} = \sqrt{2\Pi}$, i.e $\Gamma(\frac{1}{2}) = \sqrt{\Pi}$. Hence we have shown the χ^2 distribution on 1 df to be $G(\alpha = \frac{\nu}{2}, \beta = 2)$ when $\nu = 1$.

31.4 The sum of gamma variables

In the general case if $X_1 \dots X_n \sim G(\alpha, \beta)$ are i.i.d. then $X_1 + X_2 + \dots X_n \sim G(n\alpha, \beta)$.

In particular, if $X_1, X_2, \dots, X_v \sim \chi^2$ i.i.d. then $\sum_{i=1}^v X_i \sim \chi_v^2$.

31.4.1 Details

If X and Y are i.i.d. $G(\alpha, \beta)$, then

$$M_X(t) = M_Y(t) = \frac{1}{(1 - \beta t)^\alpha}$$

and

$$M_{X+Y}(t) = M_X(t)M_Y(t) = \frac{1}{(1 - \beta t)^{2\alpha}}$$

So

$$X + Y \sim G(2\alpha, \beta)$$

In the general case if $X_1 \dots X_n \sim G(\alpha, \beta)$ are i.i.d. then $X_1 + X_2 + \dots X_n \sim G(n\alpha, \beta)$. In particular, if $X_1, X_2, \dots, X_v \sim \chi^2$ i.i.d., then $\sum_{i=1}^v X_i \sim \chi_v^2$

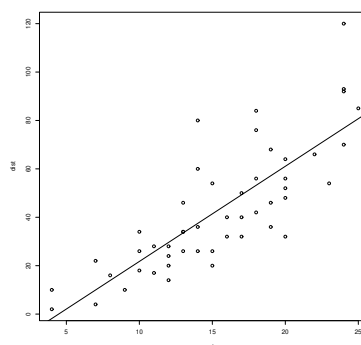
Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students

This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

32 Notes and examples: The linear model

32.1 Simple linear regression in R

To test the effect of one variable on another, simple linear regression may be applied. The fitted model may be expressed as $y = \alpha + \hat{\beta}x$, where α is a constant, $\hat{\beta}$ is the estimated coefficient, and x is the explanatory variable.



Example taken from R of a fitted model using linear regression.

32.1.1 Details

Below is the linear regression output using the R's data set "car". Notice that the output from the model may be divided into two main categories:

1. output that assesses the model as a whole, and
2. output that relates to the estimated coefficients for the model

Call:

```
lm(formula = dist ~ speed, data = cars)
```

Residuals:

```
      Min 1Q Median 3Q Max
-29.069 -9.525 -2.272  9.215 43.201
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791    6.7584   -2.601  0.0123 *
speed        3.9324    0.4155    9.464 1.49e-12 ***
```

Residual standard error: 15.38 on 48 degrees of freedom

Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438

F-statistic: 89.57 on 1 and 48 DF, p-value: 1.490e-12

Notice that there are four different sets of output (Call, Residuals, Coefficients, and Results) for both the constant α and the estimated coefficient $\hat{\beta}$ speed variable.

The estimated coefficients describe the change in the dependent variable when there is a single unit increase in the explanatory variable given that everything else is held constant.

The standard error is a measure of accuracy and is used to construct the confidence interval. Confidence intervals provide a range of values for which there is a set level of confidence that the true population mean will be within the given range. For example, if the CI is set at 95% percent then the probability of observing a value outside the given CI range is less than 0.05.

The p-value is represented as a percentage. Specifically, the p-value indicates the percentage of time, given that your null hypothesis is true, that you would find an outcome at least as extreme as the observed value. If your calculated p-value is 0.02 then 2

In the overall model assessment the R-squared is the explained variance over the total variance. Generally, a higher R^2 is better but data with very little variance makes it easy to achieve a higher R^2 , which is why the adjusted R^2 is presented.

Lastly, the F-statistic is given. Since the t-Statistic is not appropriate to compare two or more coefficients, the F-statistic

must be applied. The basic methodology is that it compares a restricted model where the coefficients have been set to a certain fixed level to a model which is unrestricted. The most common is the sum of squared residuals F-test.

32.2 Multiple linear regression

Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Formally, the model for multiple linear regression, given n observations, is

$$y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_p x_{i,p} + \varepsilon_i \text{ for } i = 1, 2, \dots, n$$

As always, we view the data, y_i as observations of random variables, so another way to describe the same model is

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_p x_{i,p} + \varepsilon_i \text{ for } i = 1, 2, \dots, n,$$

and we note that the x -values are just numbers and are usually assumed to be without any measurement error.

32.3 The one-way model

The one-way ANOVA model is of the form:

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

or

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

32.3.1 Details

The one-way ANOVA model is of the form:

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

where Y_{ij} is observation j in treatment group i and μ_i are the parameters of the model and are means of treatment group i . The ε_{ij} are independent and follow a normal distribution with

mean zero and constant variance σ^2 often written as $\varepsilon \sim N(0, \sigma^2)$.

The ANOVA model can also be written in the form:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

where μ is the overall mean of all treatment groups and α_i is the deviation of mean of treatment group i from the overall mean. The ε_{ij} follow a normal distribution as before.

The expected value of Y_{ij} is μ_i as the expected value of the errors is zero, often written as $E[Y_{ij}] = \mu_i$.

32.3.2 Examples

Example 32.1. In the rat diet experiment the model would be of the form:

$$y_{ij} = \mu_i + \varepsilon_{ij}$$

where y_{ij} is the weight gain for rat j in diet group i , μ_i would be the mean weight gain in diet group i and ε_{ij} would be the deviation of rat j from the mean of its diet group.

32.4 Random effects in the one-way layout

The simplest random effects model is the one-way layout, commonly written in the form

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

where $j = 1, \dots, J$ and $i = 1, \dots, I$.

Normally one also assumes $\varepsilon_{ij} \sim n(0, \sigma_A^2)$, $\alpha_i \sim n(0, \sigma_A^2)$, and that all these random variables are independent.

Note that we have stopped making a distinction in notation between random variables and measurements (the y -values are just random variables when distributions occur).

32.4.1 Details

Note that this is considerably different from the fixed effect model.

Since the factor has changed to a random variable with an expected value of zero, the expected value of all the y is the same:

$$E y_{ij} = \mu.$$

The variance of y now has two components:

$$V y_{ij} = \sigma_A^2 + \sigma^2.$$

In addition we have a covariance structure between the measurements and this needs to be looked at in some detail. First, the general case of a covariance between two general y_{ij} and $y_{i'j'}$, where the indices may or may not be the same:

$$\begin{aligned} \text{cov}(y_{ij}, y_{i'j'}) &= \text{cov}(\alpha_i + \varepsilon_{ij}, \alpha_{i'} + \varepsilon_{i'j'}) \\ &= E[(\alpha_i + \varepsilon_{ij})(\alpha_{i'} + \varepsilon_{i'j'})] \\ &= E[\alpha_i \alpha_{i'}] + E[\varepsilon_{ij} \alpha_{i'}] + E[\alpha_i \varepsilon_{i'j'}] + E[\varepsilon_{ij} \varepsilon_{i'j'}] \end{aligned}$$

Note 32.1. Recall that $E[UW] = E[U]E[W]$ if U, W are independent

So,

$$E[\varepsilon_{ij} \alpha_{i'}] = E[\alpha_i \varepsilon_{i'j'}] = E \alpha_i E \varepsilon_{i'j'} = 0.$$

Further,

$$E[\varepsilon_{ij} \varepsilon_{i'j'}] = \begin{cases} \sigma^2 & \text{if } i = i', j = j' \\ 0 & \text{otherwise} \end{cases}$$

and

$$E[\alpha_i \alpha_{i'}] = \begin{cases} \sigma_A^2 & \text{if } i = i' \\ 0 & \text{if } i \neq i' \end{cases}$$

so

$$\text{Cov}(y_{ij}, y_{i'j'}) = \begin{cases} \sigma_A^2 + \sigma^2 & \text{if } i = i', j = j' \\ \sigma_A^2 & \text{if } i = i', j \neq j' \\ 0 & \text{otherwise} \end{cases}$$

It follows that the correlation between measurements y_{ij} and $y_{ij'}$ (within the same group) are

$$\begin{aligned} \text{cor}(y_{ij}, y_{ij'}) &= \frac{\text{Cov}(y_{ij}, y_{ij'})}{\sqrt{v[y_{ij}]v[y_{ij'}]}} \\ &= \frac{\sigma_A^2}{\sqrt{(\sigma_A^2 + \sigma^2)^2}} \end{aligned}$$

$$\Rightarrow \text{Cor}(y_{ij}, y_{ij'}) = \frac{\sigma_A^2}{\sigma_A^2 + \sigma^2}$$

This is the intra-class correlation.

32.5 Linear mixed effects models (lmm)

The simplest mixed effects model is

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

where $\mu, \alpha_1, \alpha_2, \dots, \alpha_i$ are unknown constants,

$$\beta_j \sim n(0, \sigma_\beta^2)$$

$$\varepsilon_{ij} \sim n(0, \sigma^2)$$

(β_j and ε_{ij} independent).

32.5.1 Details

The μ and α_i are the fixed effects and β_j is the random effects.

Recall that in the simple one-way layout with $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, we can write the model in matrix form $\underline{y} = X\underline{\beta} + \underline{\varepsilon}$ where $\underline{\beta} = (\mu, \alpha_1, \dots, \alpha_I)'$ and X is appropriately chosen.

The same applies to the simplest random effects model $y_{ij} = \mu + \beta_j + \varepsilon_{ij}$ where we can write $\underline{y} = \mu \cdot \underline{1} + Z\underline{U} + \underline{\varepsilon}$ where $\underline{1} = (1, 1, \dots, 1)'$, $\underline{U} = (\beta_1, \dots, \beta_J)'$.

In general, we write the mixed effects models in matrix form with $\underline{y} = X\underline{\beta} + Z\underline{U} + \underline{\varepsilon}$, where $\underline{\beta}$ contains the fixed effects and \underline{U} contains the random effects.

32.5.2 Examples

- Example 32.2.** 1. $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ (SLR)
2. $y_{ij} = \mu + \alpha_i + \beta_j x_{ij} + \varepsilon_{ij}$ only fixed effects (ANCOVA)
3. $y_{ijk} = \mu + \alpha_i + b_j + \varepsilon_{ijk}$ where α_i are fixed but b_j are random.
4. $y_{ijk} = \mu + \alpha_i + b_j x_{ij} + \varepsilon_{ijk}$ where α_i are fixed but b_j are random slopes.

32.6 Maximum likelihood estimation in lmm

The likelihood function for the unknown parameters $L(\beta, \sigma_A^2, \sigma^2)$ is

$$\frac{1}{(2\pi)^{n/2} |\Sigma_y|^{n/2}} e^{-1/2(\mathbf{y}-X\beta)'\Sigma_y^{-1}(\mathbf{y}-X\beta)}$$

where $\Sigma_y = \sigma_A^2 Z Z' + \sigma^2 I$.

Maximising L over $\beta, \sigma_A^2, \sigma^2$ gives the variance components and the fixed effects. May also need $\hat{\mathbf{u}}$, this is normally done using BLUP.

32.6.1 Details

Recall that if W is a random variable vector with $EW = \mu$ and $VW = \Sigma$ then

$$E[AW] = A\mu$$

$$V[AW] = A\Sigma A'$$

In particular, if $W \sim n(\mu, \Sigma)$ then $AW \sim n(A\mu, A\Sigma A')$.

Now consider the lmm with

$$y = X\beta + Zu + \varepsilon$$

where

$$u = (u_1, \dots, u_m)'$$

$$\varepsilon = (\varepsilon_1, \dots, \varepsilon_m)'$$

and the random variables $U_i \sim n(0, \sigma_A^2)$, $\varepsilon_i \sim n(0, \sigma^2)$ are all independent so that $u \sim n(0, \sigma_A^2 I)$ and $\varepsilon \sim n(\mathbf{0}, \sigma^2 I)$.

Then $Ey = X\beta$ and

$$\begin{aligned} Vy &= \Sigma_y \\ &= V[Zu + V[\varepsilon]] \\ &= Z(\sigma_A^2 I)Z' + \sigma^2 I \\ &= \sigma_A^2 ZZ' + \sigma^2 I \end{aligned}$$

and hence $y \sim n(X\beta, \sigma_A^2 ZZ' + \sigma^2 I)$.

Therefore the likelihood function for the unknown parameters $L(\beta, \sigma_A^2, \sigma^2)$ is

$$= \frac{1}{(2\pi)^{n/2} |\Sigma_y|^{n/2}} e^{-1/2(y-X\beta)'\Sigma_y^{-1}(y-X\beta)}$$

where $\Sigma_y = \sigma_A^2 ZZ' + \sigma^2 I$. Maximizing L over $\beta, \sigma_A^2, \sigma^2$ gives the variance components and the fixed effects. May also need \hat{u} , which is normally done using BLUP.

Copyright 2022, Gunnar Stefansson (editor) with contributions from very many students

This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

33 Some regression topics

33.1 Poisson regression

Data y_i are from a Poisson distribution with mean μ_i and $\ln \mu_i = \beta_1 + \beta_2 x_i$. A likelihood function can be written and the parameters can be estimated using maximum likelihood.

33.2 The generalized linear model (GLM)

Data y_i are from a distribution within the exponential family, with mean μ_i and $g(\mu_i) = \mathbf{x}_i' \boldsymbol{\beta}$ for some link function, g . A likelihood function can now be written and the parameters can be estimated using maximum likelihood.

33.2.1 Details

Data y_i are from a distribution within the exponential family, with mean μ_i and $g(\mu_i) = \mathbf{x}_i' \boldsymbol{\beta}$ for some link function, g .

The exponential family includes distributions such as the Gaussian, binomial, Poisson, and gamma (and thus exponential and chi-squared).

The link functions are typically

- identity (with the Gaussian)
- log (with the Poisson and the gamma)
- logistic (with the binomial)

A likelihood function can be set up for each of these models and the parameters can be estimated using maximum likelihood.

The `glm` package in R has options to estimate parameters in these models.

Copyright 2022, Gunnar Stefansson (editor) with contributions from very many students

This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

34 Overview drills

Copyright 2021, Gunnar Stefansson (editor) with contributions from very many students

This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.