math612.3 612.3 Some notes on statistics and probability

Gunnar Stefansson

19. desember 2016

Acknowledgements
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`
Háskóli Íslands
`http://www.hi.is/`

Efnisyfirlit

1 Multivariate probability distributions

1.1 Joint probability distribution

If
$X_1,\ldots,X_n$ are discrete random variables with
$P[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] = p(x_1,\ldots,x_n)$, where $x_1,\ldots,x_n$ are numbers, then the function p is the joint probability mass function (p.m.f.) for the random variables $X_1,\ldots,X_n$.

For continuous random variables $Y_1,\ldots,Y_n$, a function f is called the joint probability density function if,
$P[Y \in A] = \int\int\ldots\int f(y_1,\ldots y_n)dy_1dy_2\cdots dy_n$.

1.1.1 Details

Definition 1.1. If $X_1,\ldots,X_n$ are discrete random variables with $P[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] = p(x_1,\ldots,x_n)$ where $x_1\ldots x_n$ are numbers, then the function p is the joint **probability mass function (p.m.f.)** for the random variables $X_1,\ldots,X_n$.

Definition 1.2. For continuous random variables $Y_1,\ldots,Y_n$, a function f is called the joint probability density function if,
$P[Y \in A] = \underbrace{\int\int\ldots\int}_{A} f(y_1,\ldots y_n)dy_1dy_2\cdots dy_n$.

Note 1.1. Note that if $X_1,\ldots,X_n$ are independent and identically distributed, each with p.m.f. p, then $p(x_1,x_2,\ldots,x_n) = q(x_1)q(x_2)\ldots q(x_n)$, i.e, $P[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] = P[X_1 = x_1]P[X_2 = x_2]\ldots P[X_n = x_n]$.

Note 1.2. Note also that if A is a set of possible outcomes ($A \subseteq \mathbb{R}^n$), then we have

$$P[X \in A] = \sum_{(x_1,\ldots,x_n)\in A} p(x_1,\ldots,x_n).$$

1.1.2 Examples

Example 1.1. 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 $P[X = \text{“}b\text{“}, Y = \text{”}l\text{“}] = \frac{5}{20}$.

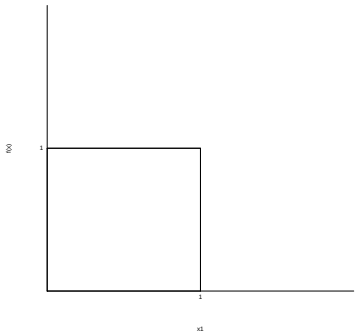
The joint p.m.f. is:

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

1.2 The random sample

A set of random variables $X_1, \ldots, X_n$ is a random sample if they are independent and identically distributed (i.i.d.).

A set of numbers $x_1, \ldots, x_n$ are called a random sample if they can be viewed as an outcome of such random variables.



1.2.1 Details

Samples from populations can be obtained in a number of ways. However, to draw valid conclusions about populations, the samples need to obtained randomly.

Definition 1.3. In **random sampling**, each item or element of the population has an equal and independent chance of being selected.

A set of random variables; $X_1 \ldots X_n$ is a random sample if they are independent and identically distributed (i.i.d.).

Definition 1.4. If a set of numbers $x_1 \ldots x_n$ can be viewed as an outcome of random variables, these are called a **random sample**.

1.2.2 Examples

Example 1.2. If $X_1, \ldots, X_n \sim U(0,1)$, 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 $x_1, x_2 \in \mathbb{R}$

$$f(x_1,x_2) = f_1(x_1)\,f_2(x_2) = \begin{cases} 1 & \text{if } 0 \le x_1, x_2 \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

Example 1.3. Toss two dice independently, and let X_1, X_2 denote the two (future) outcomes.

Then

$$P[X_1 = x_1, X_2 = x_2] = \begin{cases} \frac{1}{36} & \text{if } 1 \le x_1, x_2 \le 6 \\ 0 & \text{elsewhere} \end{cases}$$

is the joint p.m.f.

1.3 The sum of discrete random variables

1.3.1 Details

Suppose X and Y are discrete random values with a probability mass function p. Let $Z = X + Y$. Then

$$P(Z = z) \quad = \sum_{\{(x,y):x+y=z\}} p(x,y)$$

1.3.2 Examples

Example 1.4. $X, Y = outcomes$,

```
     [,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
```

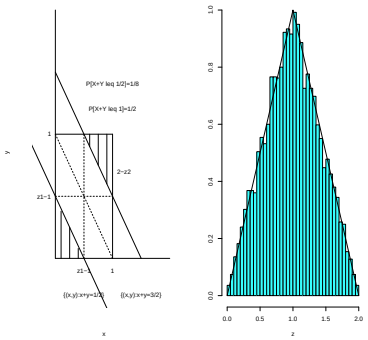
$$P[X + Y = 7] = \frac{6}{36} = \frac{1}{6}$$

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

$$P[X+Y=4]=\frac{3}{36}=\frac{1}{12}$$

1.4 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.



1.4.1 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

$$P[Z\leq z]=P[X+Y\leq z]=\int\int_{\{(x,y):x+y\leq z\}} f(x,y)dxdy.$$

1.4.2 Examples

Example 1.5. If X and Y $\sim$ U(0,1), independent and $Z=X+Y$ then

$$P[Z\leq z]=\begin{cases} 0 & \text{for} \quad z\leq 0 \\ \frac{z^2}{2} & \text{for} \quad 0<z<1 \\ 1 & \text{for} \quad z>2 \\ 1-\frac{(2-z)^2}{2} & \text{for} \quad 1<z<2 \end{cases}$$

the density of z becomes

$$g(z)=\begin{cases} z & \text{for} \quad 0<z\leq 1 \\ 2-z & \text{for} \quad 1<z\leq 2 \\ 0 & \text{for} \quad elsewhere \end{cases}$$

Example 1.6. 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.

1.5 Means and variances of linear combinations of independent random variables

If X and Y are random variables and $a, b \in \mathbb{R}$, then

$$E[aX + bY] = aE[X] + bE[Y].$$

1.5.1 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

$$E[\sum_{i=1}^{n} a_i X_i] = \sum_{i=1}^{n} a_i E[X_i]$$

when $X_i, \ldots, X_n$ are random variables and $a_1, \ldots, a_n$ 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

$$V[\sum_{i=1}^{n} a_i X_i] = \sum_{i=1}^{n} a_i^2 V[X_i]$$

1.5.2 Examples

Example 1.7. $X, Y \sim U(0,1)$, i.i.d. then

$$E[X+Y] = E[X] + E[Y] = \int_0^1 x \cdot 1 dx + \int_0^1 x \cdot 1 dx = [\frac{1}{2}x^2]_0^1 + [\frac{1}{2}x^2]_0^1 = 1.$$

Example 1.8. Let $X, Y \sim N(0,1)$. Then $E[X^2 + Y^2] = 1 + 1 = 2$.

1.6 Means and variances of linear combinations of measurements

If $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ are numbers, and we set

$$z_i = x_i + y_i$$

$$w_i = ax_i$$

where a>0, then

$$\overline{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \overline{x} + \overline{y}$$

$$\overline{w} = a\overline{x}$$

$$s_w^2 = \frac{1}{n-1}\sum_{i=1}^{n}(w_i - \overline{w})^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(ax_i - a\overline{x})^2$$

$$= a^2 s_x^2$$

and

$$s_w = a s_x$$

1.6.1 Examples

Example 1.9. We set:

```
a<-3
x<-c(1:5)
y<-c(6:10)
```

Then:

```
z<-x+y
w<-a*x
n<-length(x)
```

Then $\overline{z}$ is:

```
(sum(x)+sum(y))/n
[1] 11
mean(z)
[1] 11
```

and $\overline{w}$ becomes:

```
a*mean(x)
[1] 9
mean(w)
[1] 9
```

and s_w^2 equals:

```
sum((w-mean(w))^2))/(n-1)
[1] 22.5
sum((a*x - a*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
```

1.7 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

$$f(z_1, z_2) = \frac{1}{(\sqrt{2\pi})^2} e^{\frac{-z_1^2}{2} - \frac{z_2^2}{2}}.$$

1.7.1 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 $X_1, \ldots, X_n \sim n(\mu, \sigma^2)$ are i.i.d., then

$$f(\underline{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} e^{-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{a\sigma^2}}$$

is the multivariate normal density in the case of i.i.d. variables.

1.8 More general multivariate probability density functions

1.8.1 Examples

Example 1.10. Suppose X and Y have the joint density

$$f(x,y) = \begin{cases} 2 & 0 \le y \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

First notice that $\int_{\mathbb{R}} \int_{\mathbb{R}} f(x,y)dxdy = \int_{x=0}^{1} \int_{y=0}^{x} 2dydx = \int_0^1 2xdx = 1$, 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 $P[X \le a] = 0$ but if $a \ge 0$, we obtain

$$F_X(a) = P[X \le a] = \int_{x_0}^{a} \int_{y=0}^{x} 2dydx = [x^2]_0^a = a^2.$$

The density of X is therefore

$$f_X(x) = \frac{dF(x)}{dx} \begin{cases} 2x & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}.$$

1.8.2 Handout

If
$f : \mathbb{R}^n \to \mathbb{R}$
is such that
$P[X \in A] = \int_A \dots \int f(x_1, \dots, x_n)dx_1 \cdots dx_n$
and $f(x) \ge 0$ 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.

2 Some distributions related to the normal

2.1 The normal and sums of normals

The sum of independent normally distributed random variables is also normally distributed.

2.1.1 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 $X_1, \ldots, X_n$ are normal and independent.

2.1.2 Examples

Example 2.1. 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 2.2. 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)
```

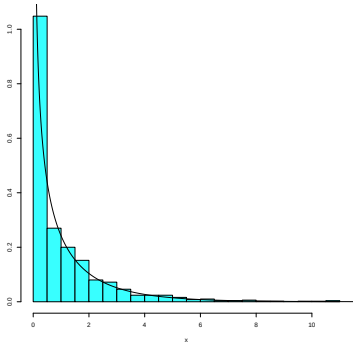
Example 2.3. Sum of nine normal distributions, all with $\mu = 42$ and $\sigma^2 = 2^2$

```
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)
```

2.2 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:

$$Y \sim \chi^2$$



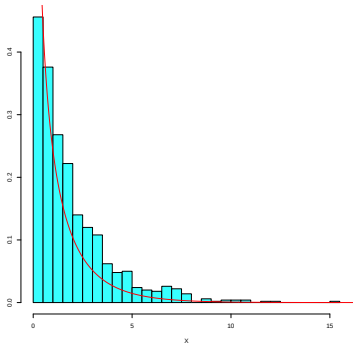
2.2.1 Details

Definition 2.1. If $X_1, X_2, \ldots, X_n$ are i.i.d. $N(0,1)$ then the distribution of $Y = X_1^2 + X_1^2 + \ldots + X_n^2$ has a **Chi square (χ^2)distribution**.

2.3 Sum of Chi square Distributions

Let Y_1 and Y_2 be independent variables. If $Y_1 = \chi^2_{\nu_1}$ and $Y_2 = \chi^2_{\nu_2}$,
then the sum of these two variables also follows a chi-squared (χ^2)distribution

$$Y_1 + Y_2 = \chi^2_{\nu_1+\nu_2}$$



2.3.1 Details

Note 2.1. Recall that if

$$X_1, \ldots, X_n \sim n(\mu, \sigma^2)$$

are i.i.d., then

$$\sum_{i=1}^{n} \left(\frac{\bar{X} - \mu}{\sigma} \right)^2 = \sum_{i=1}^{n} \frac{(\bar{X} - \mu)^2}{\sigma} \sim \chi^2$$

2.4 Sum of squared deviation

If $X_1, \cdots, X_n \sim n(\mu, \sigma^2)$ i.i.d, then

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2_n,$$

but we are often interested in

$$\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \sim \chi^2_{n-1}.$$

2.4.1 Details

Consider a random sample of Gaussian random variables, i.e. $X_1, \cdots, X_n \sim n(\mu, \sigma^2)$ i.i.d. Such a collection of random variables have properties which can be used in a number of ways.

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2_n,$$

but we are often interested in

$$\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \sim \chi^2_{n-1}.$$

Note 2.2. A degree of freedom is lost because of subtracting the estimator of the mean as opposed to the true mean.

The correct notation is:
$\underline{\mu}$ = population mean
$\underline{X}$ = sample mean (a random variable)
$\underline{x}$ = sample mean (a number)

2.5 The t-distribution

If $U \sim n(0,1)$ and $W \sim \chi^2_\nu$ are independent, then the random variable

$$T = \frac{U}{\sqrt{\frac{W}{\nu}}}$$

has a distribution which we call the t-distribution on ν degrees of freedom denoted $T \sim t_\nu$.

2.5.1 Details

Definition 2.2. If $U \sim n(0,1)$ and $W \sim \chi^2_\nu$ are independent, then the random variable

$$T := \frac{U}{\sqrt{\frac{W}{\nu}}}$$

has a distribution which we call the *t*-**distribution** on ν degrees of freedom, denoted $T \sim t_\nu$.

It turns out that if $X_1, \ldots, X_n \sim n(\mu, \sigma^2)$ and we set

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and

$$S = \sqrt{\frac{1}{1-n}\sum_{i=1}^{n}(X_i - X)^2}$$

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 $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim n(0,1)$, $\sum \frac{(X_i-\bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}$.

3 Estimation, estimates and estimators

3.1 Ordinary least squares for a single mean

If μ is unknown and $x_i, \ldots, x_n$ 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

$$\mu = \overline{x}$$

and can easily be derived by setting the derivative to zero.

3.1.1 Examples

Example 3.1. Consider the numbers $x_1, \ldots, x_5$ to be

$$13, 7, 4, 16 \text{ and } 9$$

We can plot $\sum (x_i - \mu)^2$ vs. μ and find the minimum.

3.2 Maximum likelihood estimation

If $(Y_1, \ldots, Y_n)'$ is a random vector from a density f_θ where θ is an unknown parameter, and **y** is a vector of observations then we define the **likelihood function** to be

$$L_{\mathbf{y}}(\theta) = f_\theta(y).$$

3.2.1 Examples

Example 3.2. If, $x_1, \ldots, x_n$ 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

$$\begin{aligned} & f(x_1) \cdot f(x_2) \cdot \ldots \cdot f(x_n) \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \cdot \ldots \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}} \\ &= \Pi_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \end{aligned}$$

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

$$\hat{\mu} = \overline{x}$$

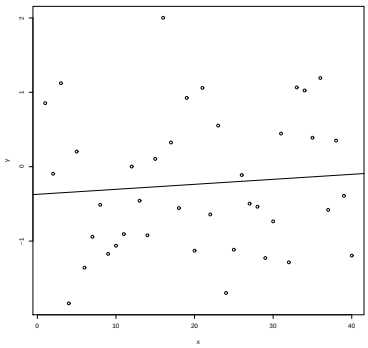
3.2.2 Detail

Definition 3.1. If $(Y_1,\ldots,Y_n)'$ 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

$$L_{\mathbf{y}}(\theta) = f_\theta(y).$$

3.3 Ordinary least squares

Consider the regression problem where we fit a line through (x_i, y_i) pairs with $x_1,\ldots,x_n$ fixed numbers but where y_i is measured with error.



Regression line through data pairs.

3.3.1 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

$$\sum_i (y_i - (\alpha + \beta x_i))^2$$

$$a = \overline{y} - b\overline{x}$$

$$b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

3.4 Random variables and outcomes

3.4.1 Details

Recall that $X_1, \ldots, X_n$ are random varibles (reflecting the population distribution) and $x_1, \ldots, x_n$ are numerical outcomes of these distributions. We use upper case letters to denote random variables and lower case letters to denote outcome or data.

3.4.2 Examples

Example 3.3. 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.

3.5 Estimators and estimates

In OLS regression, note that the values of a and b

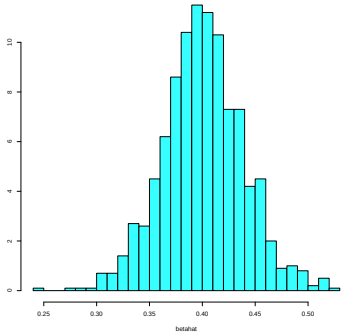
$$a = \bar{y} - b\bar{x}$$

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

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 $\hat{\beta}$

3.5.1 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)
```

4 Test of hypothesis, P values and related concepts

4.1 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.

4.1.1 Examples

Example 4.1. 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[reject H_0]= P [all heads in 5 trials] $\equiv p^5$
If H_0 is true, then P [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 4.2. Flip a coin to test
$H_0 : P = \frac{1}{2}$ 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 [reject H_0 when true] = P [all heads or all tails]
= P[all heads] + P [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.

4.2 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

$$\bar{x} \sim n(\mu_0, \frac{\sigma^2}{n})$$

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.

4.2.1 Examples

Example 4.3. 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 $\alpha = 0.05$

4.3 The two-sided z test for a normal mean

$$z := \frac{\overline{x} - \mu_0}{s\sqrt{n}} \sim n(0,1)$$

4.3.1 Details

Consider testing $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$ based on observation from $\overline{X_1}, ..., \overline{X} \sim n(\mu, \sigma^2)$ i.i.d. where σ^2 is known. If H_0 is true, then

$$Z := \frac{\overline{x} - \mu_0}{\sigma\sqrt{n}} \sim n(0,1)$$

and

$$P[|z| > z^\star] = \alpha$$

with

$$z^\star = z_1$$

We reject H_0 if $|z| > z^\star$. If $|z| > z^\star$ is not true, then we "Cannot reject the H_0".

4.3.2 Examples

Example 4.4. 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 $|-1.99|$.

4.4 The one-sided t-test for a single normal mean

Recall that if $X_1, ..., X_n \sim N(\mu, \sigma^2)$ i.i.d. then

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

4.4.1 Details

Recall that if $X_1,\ldots,X_n \sim N(\mu,\sigma^2)$ i.i.d. then

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

To test the hypothesis $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$ first note that if H_0 is true, then

$$T = \frac{\overline{X}-\mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

so

$$P[T > t*] = \alpha$$

if

$$t* = t_{n-1,1-\alpha}$$

Hence, we reject H_0 if the data $x_1,\ldots,x_n$ results in a a value of $t := \frac{\overline{x}-\mu_0}{S/\sqrt{n}}$ such that t>t*, otherwise H_0 can not be rejected.

4.4.2 Examples

Example 4.5. 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 $n = 6, \overline{x} = 18.83, s = 5.46, \mu_0 = 18$ so we obtain

$$t = \frac{\overline{x}-\mu_0}{s/\sqrt{n}} = 0.37$$

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=<$\mu_0$>)
```

4.5 Comparing means from normal populations

Suppose data are gathered independently from two normal populations resulting in $x_1,\ldots.,x_n$ and $y_1,\ldots y_m$

4.5.1 Details

We know that if

$$X_1,\ldots.,X_n \sim n(\mu_1,\sigma)$$
$$Y_1,\ldots.,Y_m \sim n(\mu_2,\sigma)$$

are all independent then

$$\bar{X}-\bar{Y} \sim n(\mu_1-\mu_2, \frac{\sigma^2}{n}+\frac{\sigma^2}{m})$$

Further,

$$\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \sim X^2_{n-1}$$

and

$$\sum_{j=1}^{m} \frac{(Y_j - \bar{Y})^2}{\sigma^2} \sim X^2_{m-1}$$

so

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}(Y_j - \bar{Y})^2}{\sigma^2} \sim X^2_{n+m-2}$$

and it follows that

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S\sqrt{(\frac{1}{n} + \frac{1}{m})}} \sim t_{n+m-2}$$

where

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_1 - \bar{X})^2 + \sum_{j=1}^{m}(Y_j - \bar{Y})^2}{n+m-2}}$$

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 $t^* = t_{n+m-2,1-\alpha}$

4.6 Comparing means from large samples <Ól.B.M.>

If $X_1, ...X_n$ and $Y_1, ...Y_m$, 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

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \dot{\sim} n(0,1)$$

if the sample sizes are large enough.

This is the central limit theorem.

4.6.1 Details

Another theorem (Slutzky) stakes 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

$$H_0 : \mu_1 = \mu_2 \qquad vs. \qquad H_a : \mu_1 > \mu_2$$

by computing

$$z = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}}$$

and reject H_0 if $z > z_{1-\alpha}$.

4.7 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.

4.7.1 Examples

Example 4.6. 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

$$\begin{aligned} P &= P[|Z| > 2.3] \\ &= P[Z < -2.3] + P[Z > 2.3] \\ &= 2 \cdot P[Z > 2.3] \end{aligned}$$

4.8 The concept of significance

4.8.1 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.

5 Power and sample sizes

5.1 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. $P[reject\ H_0] \leq \alpha$ whenever H_0 is true.

On the other hand, when H_0 is false one wants $P[reject\ H_0]$ 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 $P_{\theta_0}[reject\ H_0] \leq \alpha$ and $P_{\theta_a}[reject\ H_0]$ as large as possible.

For a general θ we write

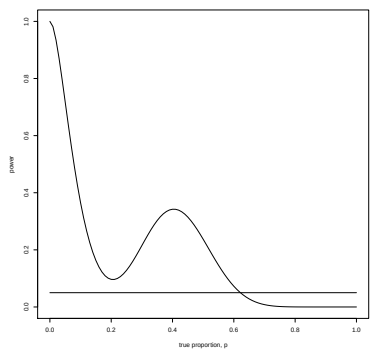
$$\beta(\theta) = P_\theta[reject\ H_0]$$

for the power of the test

5.1.1 Details

Do not use the phrase "accept".

5.2 The power of tests for proportions



5.2.1 Examples

Example 5.1. 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 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 $p = P[oddnumber] = \frac{1}{2}$ that is
$H_0 : p = \frac{1}{2}$ vs. $H_a : p \neq \frac{1}{2}$

Solution: Now, x is an outcome of $X \sim Bin(n, p)$. 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 5.1. 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 5.2. 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 5.2. 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 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 5.3. 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.

5.3 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 $X_1 \ldots X_n \sim n(\mu, \sigma^2)$ are independent and σ^2 is known.

The power of the test for an arbitrary μ can be computed as:

$$\beta(\mu) = 1 - \Phi\left(\frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} + z_{1-\alpha}\right)$$

5.3.1 Details

The one sided z-test for the mean (μ) is based on a random sample where $X_1 \ldots X_n \sim n(\mu, \sigma^2)$ 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 $x_1, \ldots x_n$, 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

$$P_{\mu_0}[RejectH_0] = P_{\mu_0}[\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_{1-\alpha}]$$

$$= P[z > z_{1-\alpha}] = \alpha$$

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.

$$\beta(\mu) = P_\mu[rejectH_0]$$

$$= P_\mu[\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_{1-\alpha}]$$

$$= P_\mu[\bar{X} > \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}]$$

$$= P_\mu[\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} > \frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} + z_{1-\alpha}]$$

$$= P[Z > \frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} + z_{1-\alpha}]$$

We obtain

$$\beta(\mu) = 1 - \Phi\left(\frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} + z_{1-\alpha}\right)$$

5.3.2 Examples

Example 5.4. 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
```

5.4 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 $X_1 \ldots X_n \sim n(\mu, \sigma^2)$ independent; that

$$\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} = \frac{\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\sum_i (X_i - X)^2}{\frac{\sigma^2}{n-1}}}} \sim t_{n-1}$$

5.4.1 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 $X_1 \ldots X_n \sim n(\mu, \sigma^2)$ 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

5.5 The power of t-test for a normal mean

5.5.1 Details

Consider $X_1, \ldots, X_n \sim n(\mu, \sigma^2)$ 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

$$T := \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim_{t_n - 1}$$

and we will reject H_0 if the computed value

$$t := \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

is such that

$$t > t^\star = t_{n-1,1-\alpha}.$$

The power of this test is:

$$\begin{aligned} B(\mu) = P_\mu[reject\ H_0] &= P_\mu[\frac{\overline{x} - \mu_0}{s/\sqrt{n}} > t^\star] \\ &= P_\mu[\overline{x} - \mu_0 > t^\star \cdot s/\sqrt{n}] \\ &= P_\mu[\frac{\overline{x} - \mu}{s/\sqrt{n}} > t^\star + \frac{\mu_0 - \mu}{s/\sqrt{n}}]. \end{aligned}$$

Which is the probability that a $t_{n-1,1-\alpha}$-variable exceed $t^\star + \frac{\mu_0-\mu}{s/\sqrt{n}}$.

5.6 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

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

is such that $z > z_{1-\alpha}$.

5.6.1 Details

Suppose we want to test $H_0 : \mu = \mu_0$ vs $H_a : \mu > \mu_0$. So based on $X_1, ..., X_n \sim n(\mu, \sigma^2)$ i.i.d. with σ^2 known we will reject H_0 if the observed value

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

is such that $z > z_{1-\alpha}$. The power is given by:

$$\beta(\mu) = 1 - \Phi(\frac{\mu - \mu_0}{\sigma/\sqrt{n}} + z_{1-\alpha})$$

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

$$1 - \Phi(\frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} + z_{1-\alpha}) \geq \beta_1$$

5.6.2 Examples

Example 5.5.

```
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
```

5.7 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.

5.7.1 Details

$\triangle = \mu_1 - \mu_2$

$\delta = \frac{\mu_1 - \mu_2}{\sigma/\sqrt{n}}$

5.7.2 Examples

Example 5.6. For a one sided power analysis we wish to test the following hypotheses:

For a one sample test:
$H_0 : \mu = \mu_0$ vs. $H_a : \mu > \mu_0$

For a two sample test:
$H_0 : \mu_1 = \mu_2$ 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 5.7. 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 $\approx$ 32 samples to obtain a power of 0.9 for our analysis.

Example 5.8. 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 5.9. 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
```

5.8 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.

5.8.1 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

5.8.2 Examples

Example 5.10. 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.

5.9 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.

5.9.1 Details

*T*wo Sample, One-sided t-Test:
Suppose data are gathered independently from two normal populations resulting in

$$X_1,\ldots,X_n \sim n(\mu_1,\sigma^2)$$
$$Y_1,\ldots,Y_m \sim n(\mu_2,\sigma^2)$$

where all data are independent then

$$\overline{X}-\overline{Y} \sim n(\mu_1-\mu_2, \frac{\sigma^2}{n} + \frac{\sigma^2}{m})$$

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

$$t = \frac{\overline{x}-\overline{y}}{s\sqrt{\frac{1}{n}+\frac{1}{m}}}$$

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:

$$\begin{aligned}
\beta_{(\mu_1\mu_2)} &= P_{\mu_1\mu_2}[reject\ H_o] \\
&= P_{\mu_1\mu_2}\left[\frac{\overline{X}-\overline{Y}}{S\sqrt{\frac{1}{n}+\frac{1}{m}}} > t^*_{1-\alpha,n+m-2}\right] \\
&= P_{\mu_1\mu_2}\left[\frac{\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} + \frac{(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}}{S/\sigma} > t^*_{1-\alpha,n+m-2}\right] \\
&= P_{\mu_1\mu_2}\left[\frac{Z+\frac{(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}}{S/\sqrt{(n+m-2)}} > t^*_{1-\alpha,n+m-2}\right] \\
&= P_{\mu_1\mu_2}\left[\frac{Z+\Delta}{\sqrt{U/(n+m-2)}} > t^*_{1-\alpha,n+m-2}\right]
\end{aligned}$$

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^*.

*T*wo 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:

$$\begin{aligned}
\beta_{(\mu_1\mu_2)} &= P_{\mu_1\mu_2}[reject\ H_o] \\
&= P_{\mu_1\mu_2}\left[\left|\frac{\overline{X}-\overline{Y}}{S\sqrt{\frac{1}{n}+\frac{1}{m}}}\right| > t^*_{1-\alpha,n+m-2}\right] \\
&= P_{\mu_1\mu_2}\left[\frac{\overline{X}-\overline{Y}}{S\sqrt{\frac{1}{n}+\frac{1}{m}}} > t^*_{1-\alpha,n+m-2}\right] \\
&\quad + P_{\mu_1\mu_2}\left[\frac{\overline{X}-\overline{Y}}{S\sqrt{\frac{1}{n}+\frac{1}{m}}} < -t^*_{1-\alpha,n+m-2}\right] \\
&= P_{\mu_1\mu_2}\left[\frac{\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} + \frac{(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}}{S/\sqrt{(n+m-2)}} > t^*_{1-\alpha,n+m-2}\right] \\
&\quad + P_{\mu_1\mu_2}\left[\frac{\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} + \frac{(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}}{S/\sqrt{(n+m-2)}} < -t^*_{1-\alpha,n+m-2}\right] \\
&= P_{\mu_1\mu_2}\left[\frac{Z+\Delta}{\sqrt{U/(n+m-2)}} > t^*_{1-\alpha,n+m-2}\right] \\
&\quad + P_{\mu_1\mu_2}\left[\frac{Z+\Delta}{\sqrt{U/(n+m-2)}} < -t^*_{1-\alpha,n+m-2}\right]
\end{aligned}$$

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 5.3. Note that the power of a test can be obtained using the *power.t.test* function in R.

5.10 Sample sizes for two-sample one and two-sided t-tests

The sample size should always satisfy the desired power.

5.10.1 Details

Suppose we want to reject the H_o 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
$P_{\mu_1\mu_2}\left[\frac{Z+\Delta}{\sqrt{U/(n+m-2)}} > t^*_{1-\alpha,n+m-2}\right] + P_{\mu_1\mu_2}\left[\frac{Z+\Delta}{\sqrt{U/(n+m-2)}} < -t^*_{1-\alpha,n+m-2}\right]$

5.11 A case study in power

Want to compute power in analysis of covariance:

$$y_{ij} = \mu_i + \beta x_{ij} + \varepsilon_{ij},\ i = 1,2,\ j = 1,\ldots J,$$

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.

5.11.1 Handout

Example 5.11. If you want to compute a power analysis in analysis of covariance:

$$y_{ij} = \mu_i + \beta x_{ij} + \varepsilon_{ij},\ i = 1,2,\ j = 1,\ldots J,$$

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 5.4. 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 covarariance
# - 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%.