Applied Statistics practice (use R)

* Installed R/Rstudio?
* Download R from: <https://www.r-project.org/>
* If you use cluster: <http://cluster2.inm.ras.ru/~hpc19>
  + Log in ssh [hpc19@dodo.inm.ras.ru](mailto:hpc19@dodo.inm.ras.ru)
  + makedir [yourname]
  + cd [yourname]
  + TYPE R then click ENTER
  + Check R plots in browser (<http://cluster2.inm.ras.ru/~hpc19>/yourname)

1. **Basic statistics (5 mins)**

**x = 1; x**

**class(x);**

**x <- 1; x**

R program: Calculate the minimum value of the sample (2,3,3,3,7,7,7,7,9,10,21)

**S<-c(2,3,3,3,7,7,7,7,9,10,21)**

**min(S)**

[1] 2

#minimum index

**which.min(S)**

1. 1

R program: Calculate the maximum value of the sample (2,3,3,3,7,7,7,7,9,10,21)

**S<-c(2,3,3,3,7,7,7,7,9,10,21)**

#maximum

**max(S)**

[1] 21

#maximum index

**which.max(S)**

[1] 11

R program: calculate the **median** of the sample (1,2,3,4,5)

**S<-c(1,2,3,4,5)**

**median(S)**

[1] 3

R program: Calculate the quartile of the sample (1,2,3,4,5,6,7,8,9)

**S<-c(1,2,3,4,5,6,7,8,9)**

**quantile(S)**

0% 25% 50% 75% 100%

1 3 5 7 9

**fivenum(S)**

[1] 1 3 5 7 9

Calculate the **mode** of the sample (1,2,3,3,3,7,7,7,7,9,10,21)

**S<-c(1,2,3,3,3,7,7,7,7,9,10,21)**

**names(which.max(table(S)))**

1. "7"
2. **Random variables (10 mins)**

R program: generate random variable in (0,1,2,3,4,5)

**S<-1:5**

**sample(S,1)**

1. 2

**sample(S,1)**

[1] 3

**sample(S,4)**

1. 3 5 4 1

**sample(x=x, size=5, replace=T)**, where size specifies the number of samplings, and "replace" means to repeat. That is, the elements can be repeatedly sampled, which is the so-called sampling with replacement.

R program: Generate 10 continuous random numbers of the sample in the space (0,1):

**runif(10, 0,1)**

0.3819569 0.7609549 0.6692581 0.6314708 0.5552201 0.8225527 0.7633086 0.4667188 0.1883553 0.3741653

The rules of **runif(n, min=0, max=1)** function: n represents the number of random numbers generated, min represents the lower limit of **uniform distribution**, max represents the upper limit of uniform distribution; if the parameters min and max are omitted, it will be generated by default [0 ,1] uniformly distributed random number.

R program: Calculate the mathematical expectation of the sample (1,2,3,7,21)

**S<-c(1,2,3,7,21)**

**mean(S)**

1. 6.8

R program: Calculate the variance of the sample (1,2,3,7,21)

**S<-c(1,2,3,7,21)**

**var(S)**

1. 68.2

R program: Calculate the standard deviation of the sample (1,2,3,7,21)

**S<-c(1,2,3,7,21)**

**sd(S)**

[1] 8.258329

R program: Calculate the covariance of (1,2,3,4) and (5,6,7,8)

**X<-c(1,2,3,4)**

**Y<-c(5,6,7,8)**

**cov(X,Y)**

[1] 1.666667

R program: Calculate the correlation coefficient of (1,2,3,4) and (5,7,8,9)

**X<-c(1,2,3,4)**

**Y<-c(5,7,8,9)**

**cor(X,Y)**

[1] 0.9827076

1. **Discrete distribution (15 mins)**

* Binomial distribution

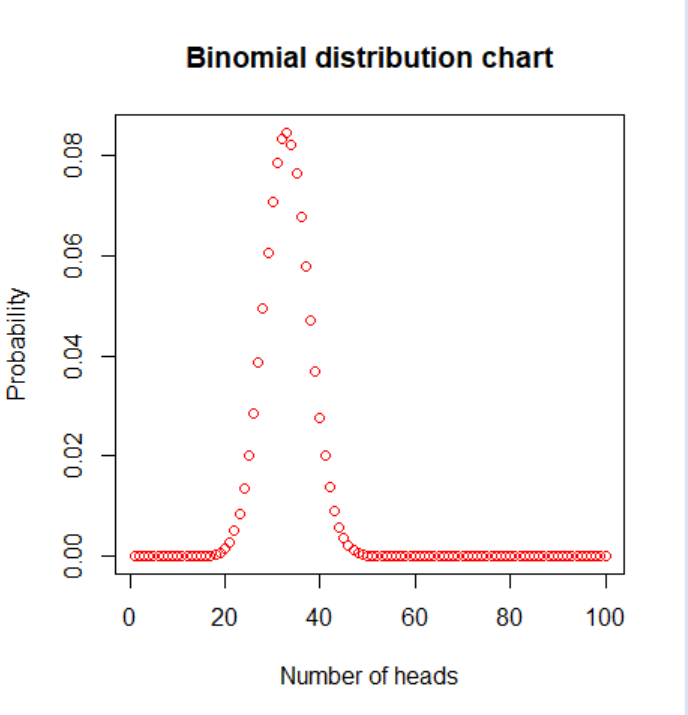
Assuming that an uneven coin is tossed, the probability that the result is heads each time is 0.33, and the probability of 20 heads in tossing 100 times is：

**dbinom(x=20, size=100, prob=0.33)**

Among them, the first parameter indicates the total number of occurrences of an event, the second parameter indicates the total number of trials, and the third parameter indicates the probability of an event occurring once.

**Plot density function:**

**plot(dbinom(1 : 100, 100, 0.33), col = "red", main = "Binomial distribution chart", xlab = "Number of heads", ylab = "Probability")**



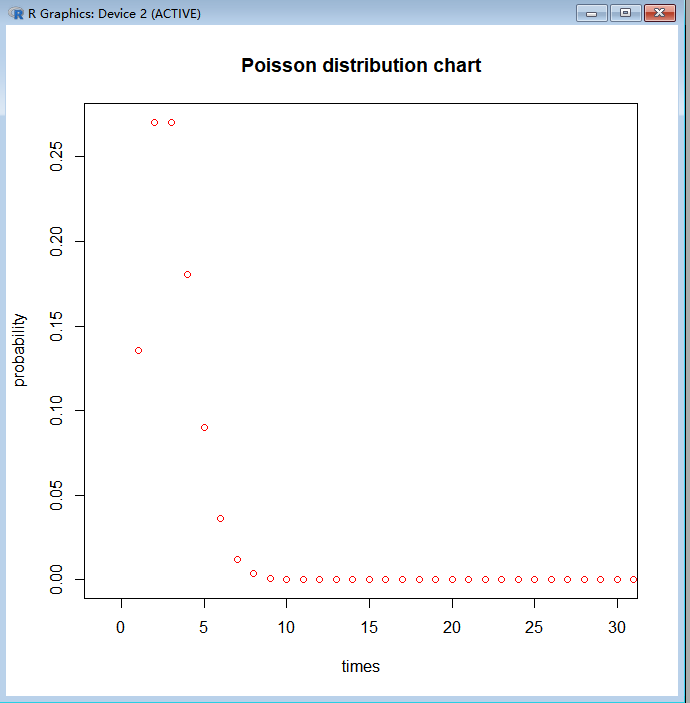
* Poisson distribution

Assuming that a hospital receives 2 patients per minute on average, assuming that the probability of receiving 4 patients per minute at a certain time is:

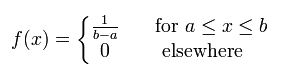
**dpois(4, lambda = 2)**

Graph of probability density distribution：

**plot(dpois(0:30, lambda = 2), col = "red", xlim = c(-1,30), xlab = "time", ylab = "probability", main = "Poisson distribution chart")**



1. **Continuous distribution (25 mins)**
2. **Uniform probability**
3. Probability density function



The probability of each point of x in the interval [a, b] is the same, and the probability of each point has nothing to do with the value of x.

Example: uniform distribution on the interval [-1,1], probability density at x=0：

dunif(0,-1,1)

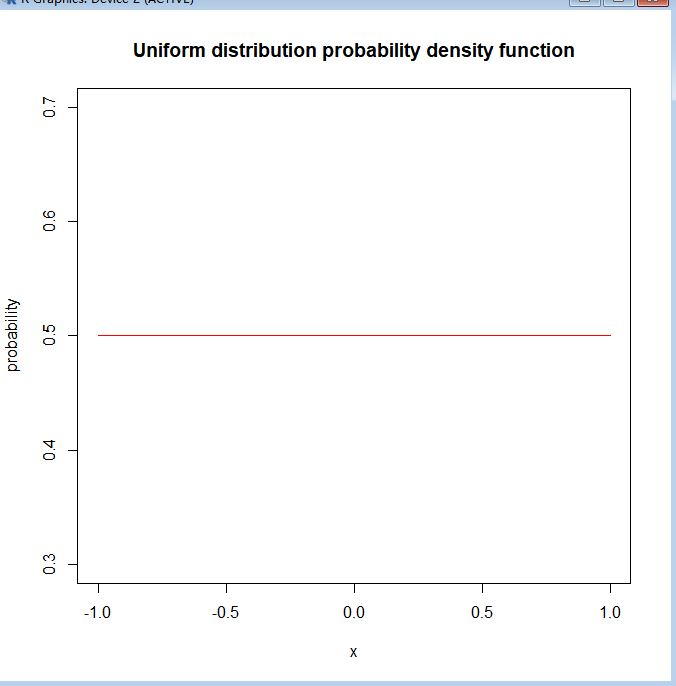
Among them, the second parameter is the interval minimum, and the third parameter is the interval maximum.

Uniform distribution probability density plot：

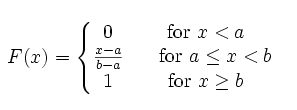
**x = seq(-1, 1, length.out = 100)**

**y = dunif(x, -1, 1)**

**plot(x, y, col = "red", type = "l", main = "Uniform distribution probability density function", ylab = "probability")**



1. Cumulative distribution function



**x<-seq(-5,5,length.out=100)**

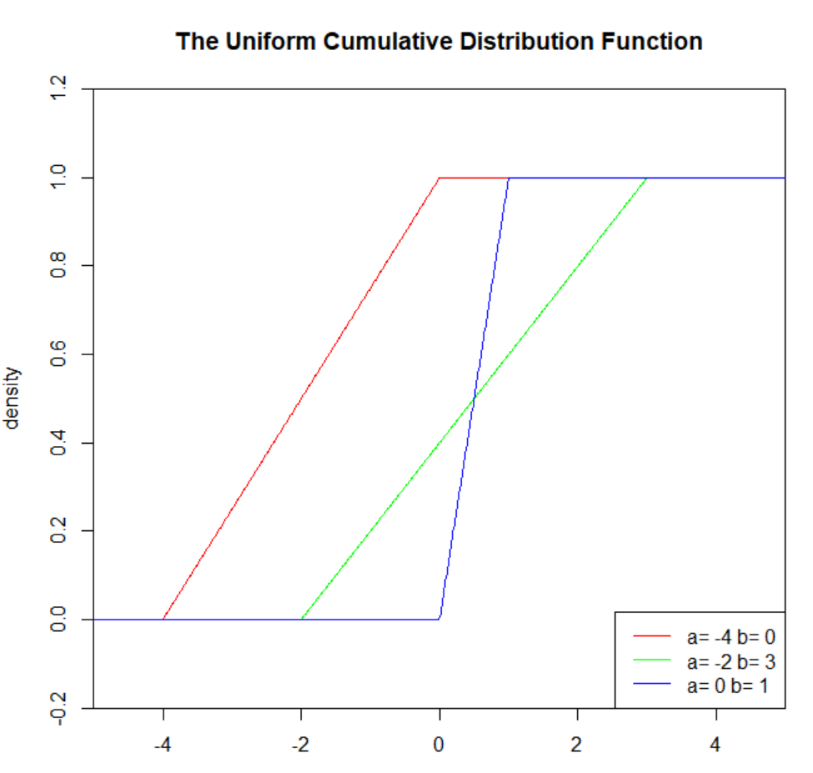
**y<-punif(x,-4,0)**

**plot(x,y,col="red",xlim=c(-5,5),ylim=c(-0.2,1.2),type='l', xaxs="i",yaxs="i",ylab='density',xlab='',main="The Uniform Cumulative Distribution Function")**

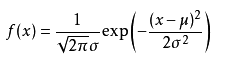
**lines(x,punif(x,-2,3),col="green")**

**lines(x,punif(x,0,1),col="blue")**

**legend("bottomright",legend=paste("a=",c(-4,-2,0),"b=",c(0,3,1)),lwd=1,col=c("red","green","blue"))**



1. **Normal distribution**
2. Probability density function



The normal distribution is the most common in life. The heights of men and women, test scores, life span, etc. all obey the normal distribution.

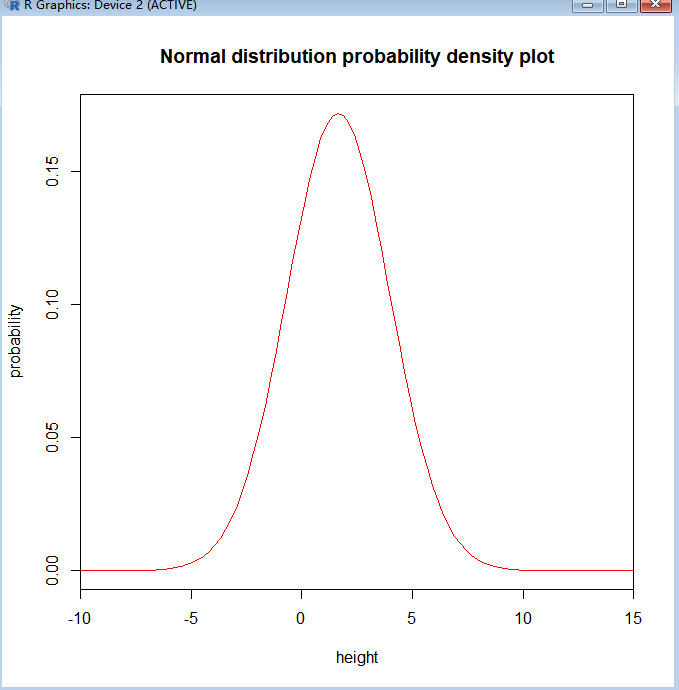
Assuming that the height of a classmate obeys a normal distribution, the average height of the class is 1.65m, and the variance is 2.32, then the probability of the height of 1.70m is:

**dnorm(x=1.70, mean=1.65, sd=2.32)**

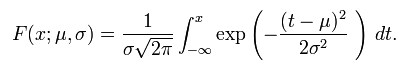
Among them, the second parameter is the average value, and the third parameter is the variance.

Probability density plot of normal distribution:

**x = seq(-10,15, length.out = 100)  
y = dnorm(x, 1.65, 2.32)  
plot(x, y, xlim = c(-10, 15), type = "l", col = 'red', xaxs = "i", main = "Normal distribution probability density plot", xlab = "height", ylab = "probability")**



1. Cumulative distribution function



**x <- seq(-5,5,length.out=100)**

**y <- pnorm(x,0,1)**

**plot(x,y,col="red",xlim=c(-5,5),ylim=c(0,1),type='l',xaxs="i", yaxs="i",ylab='density',xlab='',main="The Normal Cumulative Distribution")**

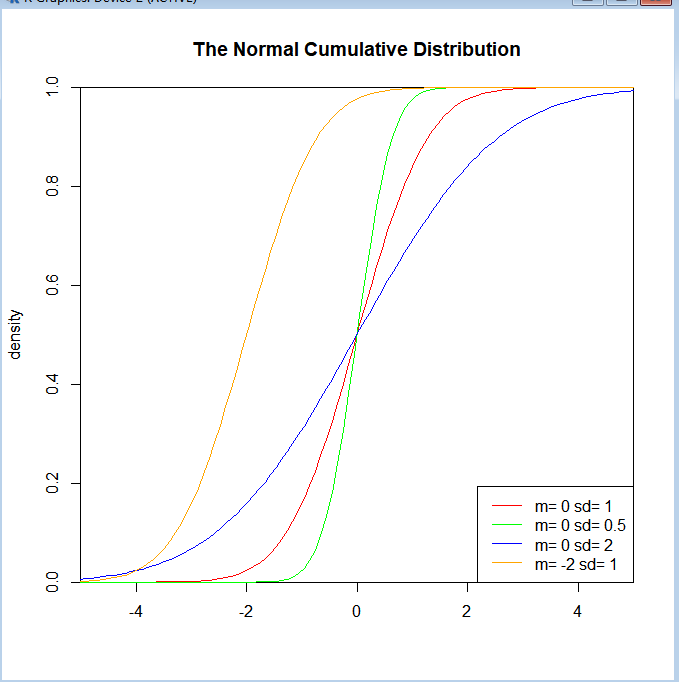
**lines(x,pnorm(x,0,0.5),col="green")**

**lines(x,pnorm(x,0,2),col="blue")**

**lines(x,pnorm(x,-2,1),col="orange")**

**legend("bottomright",legend=paste("m=",c(0,0,0,-2),"sd=",c(1,0.5,2,1)),lwd=1,**

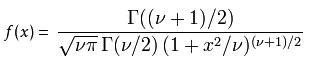
**col=c("red","green","blue","orange"))**



**(3)T distribution**

The t distribution is used to estimate the mean of a population with a normal distribution and unknown variance based on a small sample. It is the basis for the significance test of the difference between the two sample means.

1)Probability density function



*υ*=n is the degree of freedom (df), Г is the gamma function.

**x<-seq(-5,8,length.out=1000)**

**y<-dt(x,1,0)**

**plot(x,y,col="red",xlim=c(-5,8),ylim=c(0,0.5),type='l',xaxs="i",**

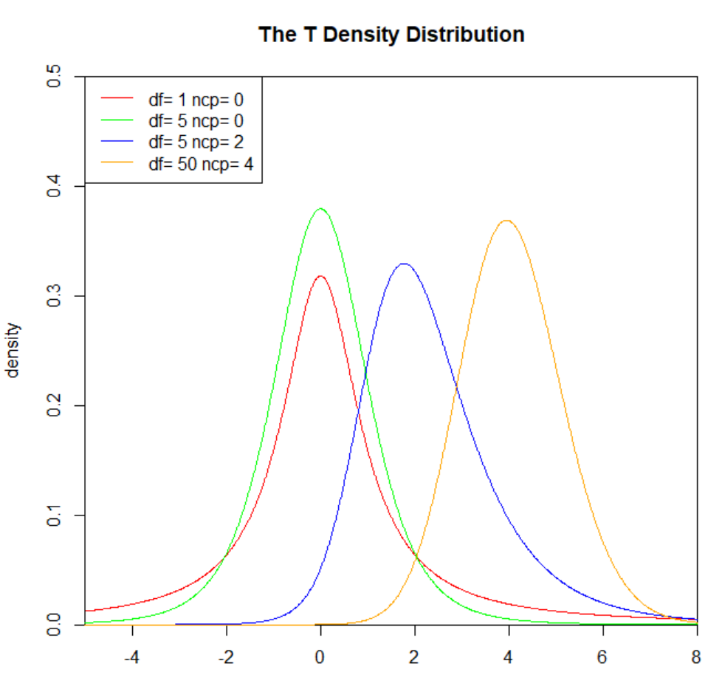
**yaxs="i",ylab='density',xlab='', main="The T Density Distribution")**

**lines(x,dt(x,5,0),col="green")**

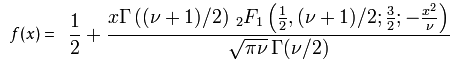
**lines(x,dt(x,5,2),col="blue")**

**lines(x,dt(x,50,4),col="orange")**

**legend("topleft",legend=paste("df=",c(1,5,5,50),"ncp=", c(0,0,2,4)), lwd=1, col=c("red", "green","blue","orange"))**



2) Cumulative distribution function



**x<-seq(-5,8,length.out=1000)**

**y<-pt(x,1,0)**

**plot(x,y,col="red",xlim=c(-5,9),ylim=c(0,1.1),type='l',**

**xaxs="i", yaxs="i",ylab='density',xlab='x',**

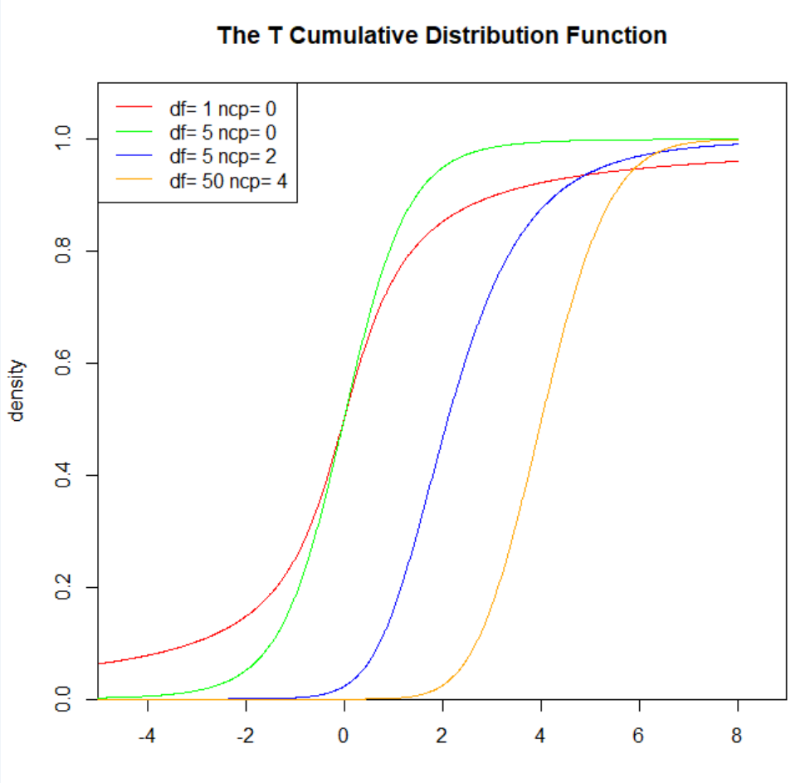
**main="The T Cumulative Distribution Function")**

**lines(x,pt(x,5,0),col="green")**

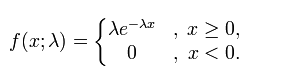
**lines(x,pt(x,5,2),col="blue")**

**lines(x,pt(x,50,4),col="orange")**

**legend("topleft",legend=paste("df=",c(1,5,5,50),"ncp=", c(0,0,2,4)), lwd=1, col=c("red", "green","blue","orange"))**

****

1. **Exponential probability distribution**
2. Probability density function



The exponential distribution can be derived from the Poisson distribution. The interval of the exponential distribution is [0, +∞), which has the characteristic of no memory. Exponential distribution can be used to represent the time interval of independent random events, such as the time interval for passengers to enter the airport, and the time when equipment fails.

Assuming that the probability of a bulb being damaged per unit time (for example, 1 hour) is 0.0168, the probability of failure within 72 hours is：

**pexp(72, rate=0.0168)**

The second parameter is the λ in the exponential density function

Exponential distribution probability density plot:

**x<-seq(-1,2,length.out=100)**

**y<-dexp(x,0.5)**

**plot(x,y,col="red",xlim=c(0,2),ylim=c(0,5),type='l',xaxs="i",yaxs="i",ylab='density',xlab='', main="The Exponential Density Distribution")**

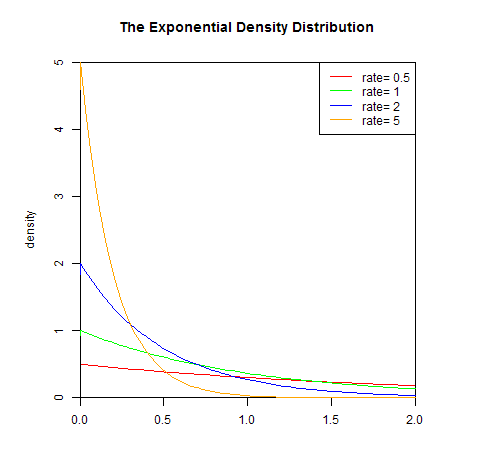
**lines(x,dexp(x,1),col="green")**

**lines(x,dexp(x,2),col="blue")**

**lines(x,dexp(x,5),col="orange")**

**legend("topright",legend=paste("rate=",c(.5, 1, 2,5)), lwd=1,col=c("red", "green","blue","orange"))**

#dexp is the probability density function of the exponential distribution



1. Cumulative distribution function



**x<-seq(-1,2,length.out=1000)**

**y<-pexp(x,0.5)**

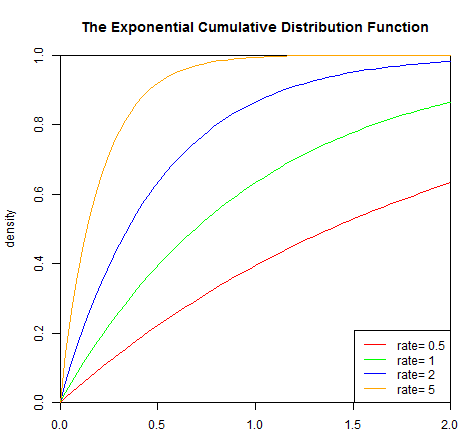
**plot(x,y,col="red",xlim=c(0,2),ylim=c(0,1),type='l',xaxs="i",yaxs="i",ylab='density',xlab='', main="The Exponential Cumulative Distribution Function")**

**lines(x,pexp(x,1),col="green")**

**lines(x,pexp(x,2),col="blue")**

**lines(x,pexp(x,5),col="orange")**

**legend("bottomright",legend=paste("rate=",c(.5, 1, 2,5)), lwd=1, col=c("red", "green","blue","orange"))**



1. **gamma distribution (optional)**

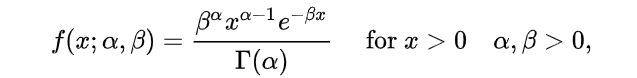
Gamma distribution is a two-[parameter](https://en.wikipedia.org/wiki/Statistical_parameter) family of continuous [probability distributions](https://en.wikipedia.org/wiki/Probability_distribution). The [exponential distribution](https://en.wikipedia.org/wiki/Exponential_distribution) and [chi-squared distribution](https://en.wikipedia.org/wiki/Chi-squared_distribution) are special cases of the gamma distribution. There are three different [parametrizations](https://en.wikipedia.org/wiki/Statistical_parameter) in common use:

1. With a [shape parameter](https://en.wikipedia.org/wiki/Shape_parameter) k and a [scale parameter](https://en.wikipedia.org/wiki/Scale_parameter) θ.
2. With a shape parameter α = k and an inverse scale parameter β = 1/θ, called a [rate parameter](https://en.wikipedia.org/wiki/Rate_parameter).
3. With a shape parameter k and a mean parameter μ = kθ = α/β.

The gamma distribution can be parameterized in terms of a [shape parameter](https://en.wikipedia.org/wiki/Shape_parameter) α = k and an inverse scale parameter β = 1/θ, called a [rate parameter](https://en.wikipedia.org/wiki/Rate_parameter). A random variable X that is gamma-distributed with shape α and rate β is denoted



1. Probability density function



**x<-seq(0,10,length.out=1000)**

**y<-dgamma(x,1,2)**

**plot(x,y,col="red",xlim=c(0,10),ylim=c(0,2),type='l',xaxs="i", yaxs="i",ylab='density',xlab='',main="The Gamma Density Distribution")**

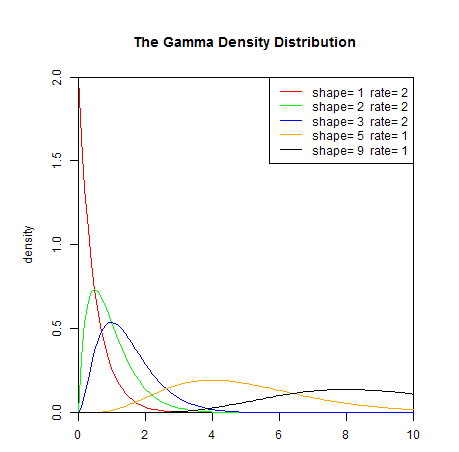
**lines(x,dgamma(x,2,2),col="green")**

**lines(x,dgamma(x,3,2),col="blue")**

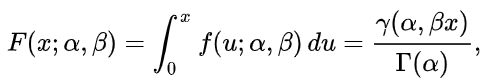
**lines(x,dgamma(x,5,1),col="orange")**

**lines(x,dgamma(x,9,1),col="black")**

**legend("topright",legend=paste("shape=",c(1,2,3,5,9),"rate=",c(2,2,2,1,1)),lwd=1,col=c("red","green","blue","orange","black"))**



1. Cumulative distribution function



**x<-seq(0,10,length.out=1000)**

**y<-pgamma(x,1,2)**

**plot(x,y,col="red",xlim=c(0,10),ylim=c(0,1),type='l',xaxs="i",yaxs="i",ylab='density',xlab='', main="The Gamma Cumulative Distribution Function")**

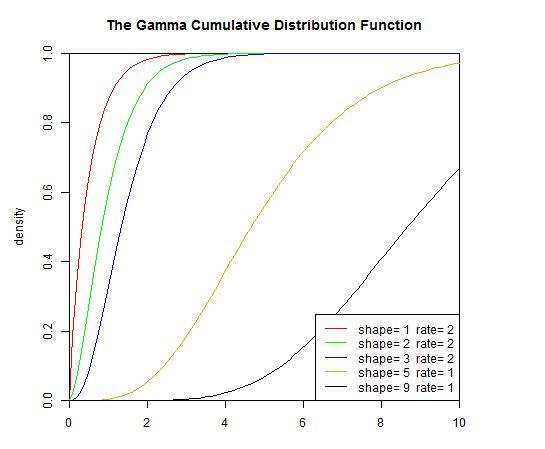
**lines(x,pgamma(x,2,2),col="green")**

**lines(x,pgamma(x,3,2),col="blue")**

**lines(x,pgamma(x,5,1),col="orange")**

**lines(x,pgamma(x,9,1),col="black")**

**legend("bottomright",legend=paste("shape=",c(1,2,3,5,9),"rate=",c(2,2,2,1,1)),lwd=1,col=c("red","green","blue","orange","black"))**



1. **χ² (chi-square) distribution (optional)**

If n mutually independent random variables ξ₁, ξ₂,..., ξn all obey the standard normal distribution, the sum of the squares of these n random variables obey the standard normal distribution constitutes a new random variable, and its distribution is called chi-square distribution. The parameter n is called the degree of freedom, and the difference in the degree of freedom is another χ² distribution, just as the difference in the mean or variance in the normal distribution is another normal distribution.

1. Probability density function



**x<-seq(0,10,length.out=1000)**

**y<-dchisq(x,1)**

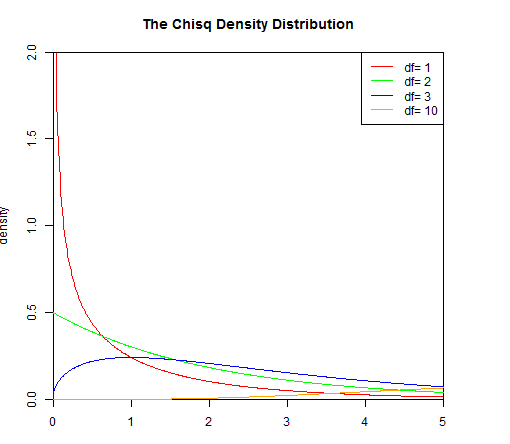
**plot(x,y,col="red",xlim=c(0,5),ylim=c(0,2),type='l',xaxs="i",yaxs="i",ylab='density',xlab='', main="The Chisq Density Distribution")**

**lines(x,dchisq(x,2),col="green")**

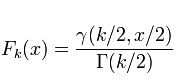
**lines(x,dchisq(x,3),col="blue")**

**lines(x,dchisq(x,10),col="orange")**

**legend("topright",legend=paste("df=",c(1,2,3,10)),lwd=1,col=c("red","green","blue","orange"))**



1. Cumulative distribution function



**x<-seq(0,10,length.out=1000)**

**y<-pchisq(x,1)**

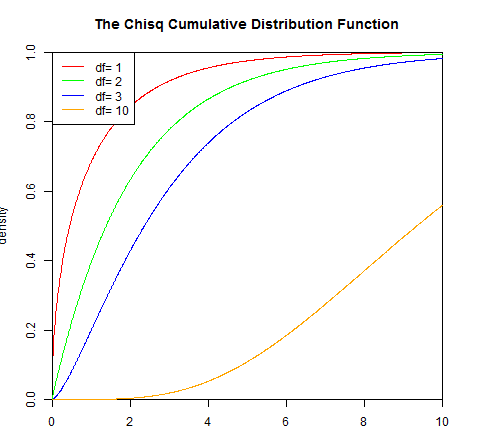
**plot(x,y,col="red",xlim=c(0,10),ylim=c(0,1),type='l',xaxs="i",yaxs="i",ylab='density',xlab='', main="The Chisq Cumulative Distribution Function")**

**lines(x,pchisq(x,2),col="green")**

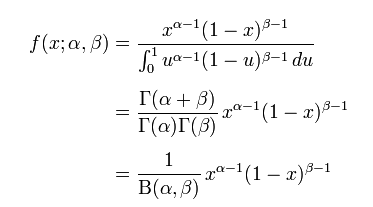
**lines(x,pchisq(x,3),col="blue")**

**lines(x,pchisq(x,10),col="orange")**

**legend("topleft",legend=paste("df=",c(1,2,3,10)),lwd=1,col=c("red","green","blue","orange"))**



1. **Beta distribution（optional）**
2. Probability density function



**x<-seq(-5,5,length.out=1000)**

**y<-dbeta(x,0.5,0.5)**

**plot(x,y,col="red",xlim=c(0,1),ylim=c(0,6),type='l',xaxs="i",yaxs="i",ylab='density',xlab='', main="The Beta Density Distribution")**

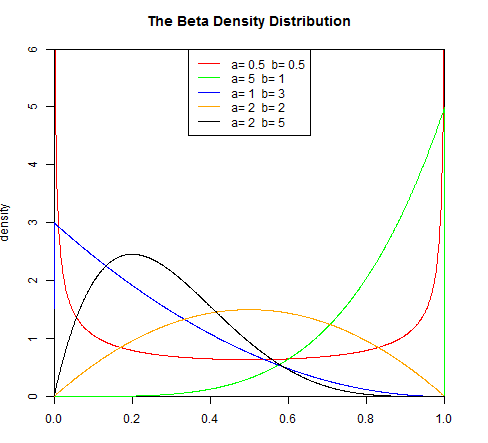
**lines(x,dbeta(x,5,1),col="green")**

**lines(x,dbeta(x,1,3),col="blue")**

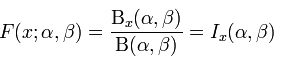
**lines(x,dbeta(x,2,2),col="orange")**

**lines(x,dbeta(x,2,5),col="black")**

**legend("top",legend=paste("a=",c(.5,5,1,2,2),"b=",c(.5,1,3,2,5)),lwd=1,col=c("red","green","blue","orange","black"))**



1. Cumulative distribution function



**x<-seq(-5,5,length.out=1000)**

**y<-pbeta(x,0.5,0.5)**

**plot(x,y,col="red",xlim=c(0,1),ylim=c(0,1),type='l',xaxs="i",yaxs="i",ylab='density',xlab='', main="The Beta Cumulative Distribution Function")**

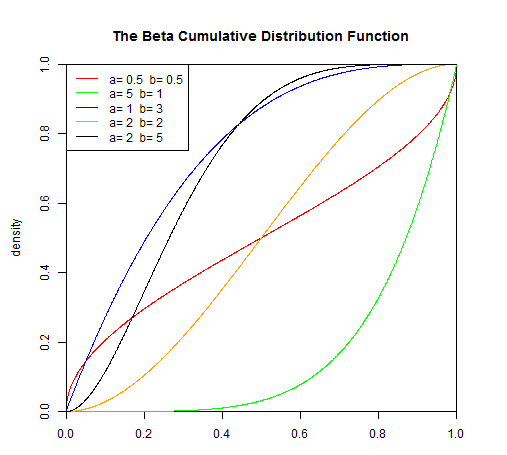
**lines(x,pbeta(x,5,1),col="green")**

**lines(x,pbeta(x,1,3),col="blue")**

**lines(x,pbeta(x,2,2),col="orange")**

**lines(x,pbeta(x,2,5),col="black")**

**legend("topleft",legend=paste("a=",c(.5,5,1,2,2),"b=",c(.5,1,3,2,5)),lwd=1,col=c("red","green","blue","orange","black"))**



**Hypothesis testing in R language （20 mins）**

一、Continuous distribution parameter test **(15 mins)**

**Example 1**. The lifetime (note as X, in hours) of a certain device obeys the normal distribution N (μ, σ), where both μ and σ2 are unknown. The life of 16 components is as follows：

                 159 280 101 212 224 379 179 264  
                222 362 168 250 149 260 485 170

   Is there any reason to believe that the average lifetime of the device is greater than 255 hours?

    Solution: According to the meaning of the question, need to check

**H0： μ ≤ 225     H1: μ > 225**

 This problem is a **one sample one-sided** test problem; **when the sample size is small, t-test is more preferable than z-test**. We can use the R language t.test

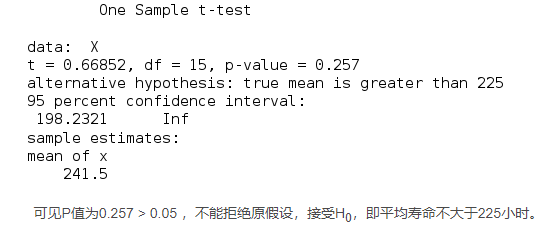
**t.test(x,y=NULL,alternative=c("two.sided","less","greater"),mu=0,paired=FALSE,var.equal=FALSE,conf.level=0.95)**

If only x is provided, the mean test of a single normal population will be performed. If x, y are provided, the mean test of two populations will be performed (Check about two sample t-test below). Alternative means to be assumed; here “two.sided” (default) means two-sided test (H1: μ≠μ0), “less” means one-sided test (H1:μ<μ0), “greater” means one-sided test (H1:μ>μ0), “mu” means the null hypothesis μ0, “conf.level” means the confidence level, that is 1-α, usually 0.95. Var.equal indicates a logical variable, var.equal=TRUE indicates that the variance of the two samples is the same, and var.equal=FALSE (default) indicates that the variance of the two samples is different.

 R language ：

**X<-c(159, 280, 101, 212, 224, 379, 179, 264, 222, 362, 168, 250, 149, 260, 485, 170)  
t.test(X,alternative = "greater",mu=225)**

result：



It can be seen that the P value is 0.257> 0.05, we don’t have enough evidence to reject null hypothesis.

**Example 2.** A test is carried out on an open hearth to determine whether changing the operation method will increase the yield of steel. The test is carried out on the same open hearth. The operation method for each furnace of steel should be the same as possible. The standard method makes one furnace, and then uses the new method to make one furnace, and then alternates, each with 10 furnaces, and the yield is respectively

  standard method    78.1 72.4 76.2 74.3 77.4 78.4 76.0 75.5 76.7 77.3  
    new method        79.1 81.0 77.3 79.1 80.0 79.1 79.1 77.3 80.2 82.1

Assume that these two samples are independent of each other and come from the normal population N(μ1, σ2) and N(μ2, σ2) respectively, where μ1, μ2 and σ2 are unknown. Can the new operation increase the yield? (Take α=0.05)

   Solution 1: According to the meaning of the question, suppose:

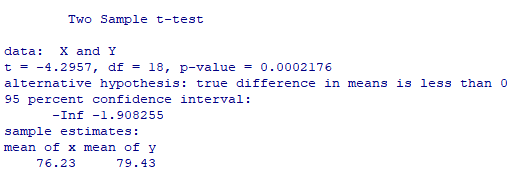
**H0： μ1 ≥  μ2     H1: μ1 <  μ2**

 It is assumed that σ12=σ22=σ2, so choose t.test, var.equal=TRUE

R code：

**X<-c(78.1,72.4,76.2,74.3,77.4,78.4,76.0,75.5,76.7,77.3)  
Y<-c(79.1,81.0,77.3,79.1,80.0,79.1,79.1,77.3,80.2,82.1)**

**t.test(X,Y,var.equal = TRUE,alternative = "less")**



 It can be seen that P value <0.05, we reject the null hypothesis, that is, the new operation can improve the yield.

**Binomial distribution parameter test (5 mins)**

Example 3. There is a batch of vegetable seeds with an average germination rate p0=0.85. Now 500 seeds are picked up and treated with water for seed soaking. As a result, 445 seeds have germinated. Test whether the water has any effect on the germination rate of seeds.

Solution: According to the meaning of the question, the question tested is

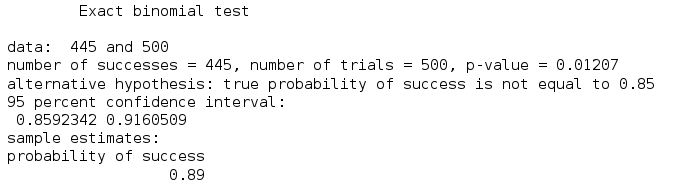
**H0：p=p0=0.85， H1：p≠p0**

 binom.test(x, n, p = 0.5, alternative = c("two.sided", "less", "greater"), conf.level = 0.95)

 Where x is the number of successes; or a two-dimensional vector composed of the number of successes and the number of failures. n is the total number of trials. When x is a two-dimensional vector, this value is invalid. P is the probability of the null hypothesis

 R language code：

binom.test(445,500,p=0.85)



It can be seen that the P value is 0.01207<0.05, reject the null hypothesis and indicate that water has a significant effect on the germination rate of seeds.

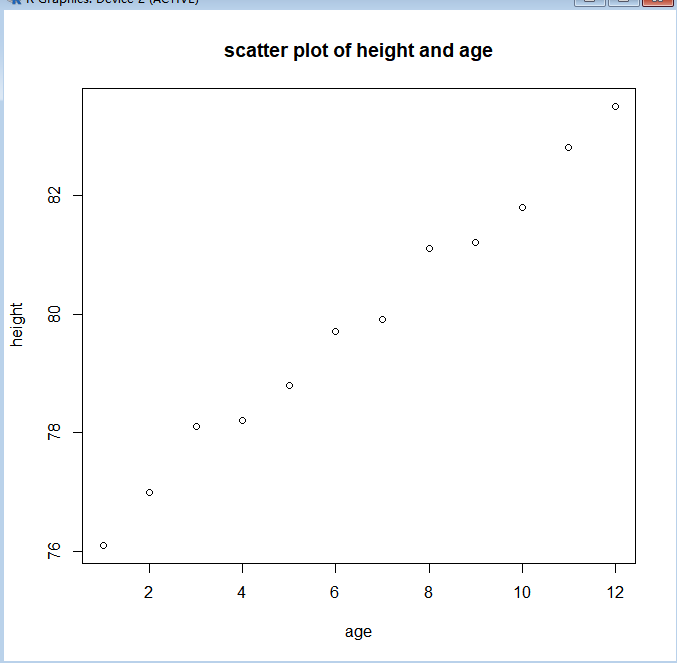
**Regression (30 mins)**

Example 1: Our data comes from a set of real recorded data about children's height and age. First of all, we first visually display the relationship between age and height, and draw a scatter plot with age as the abscissa and height as the ordinate. The code for R is as follows:

age=1:12 #Age from 1 to 12

height=c(76.1,77,78.1,78.2,78.8,79.7,79.9,81.1,81.2,81.8,82.8,83.5)

plot(age,height,main = "scatter plot of height and age")

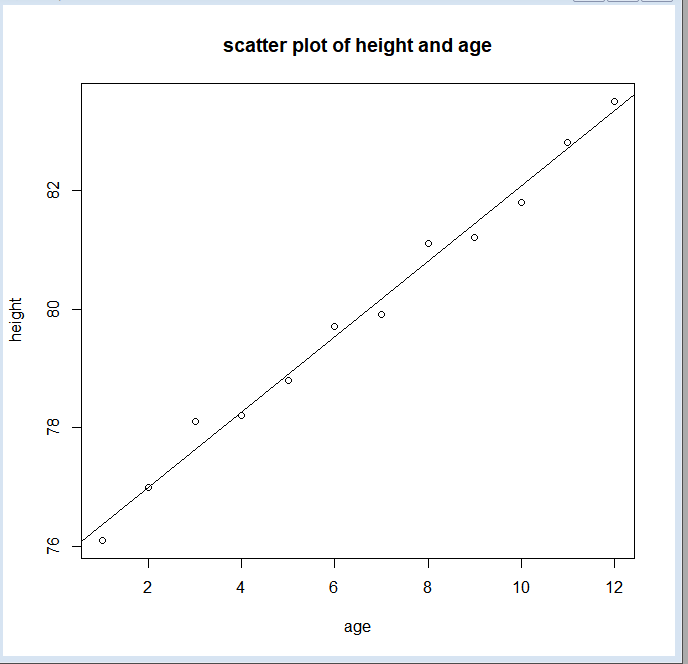


It can be observed from the figure that age and height are basically close to a straight line. It can be considered that the two have a linear relationship. Next, establish a regression model. The R code is as follows:

lm.reg <- lm(height~age)

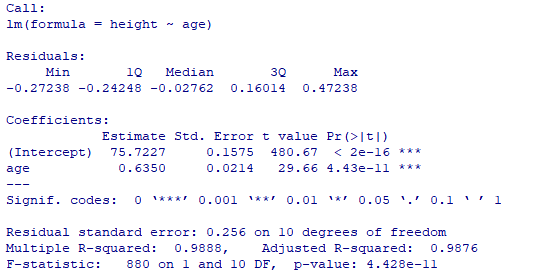
lm.reg

abline(lm.reg)



For the regression equation in the above example, we test the linearity.

summary(lm.reg) #Regression coefficient t test



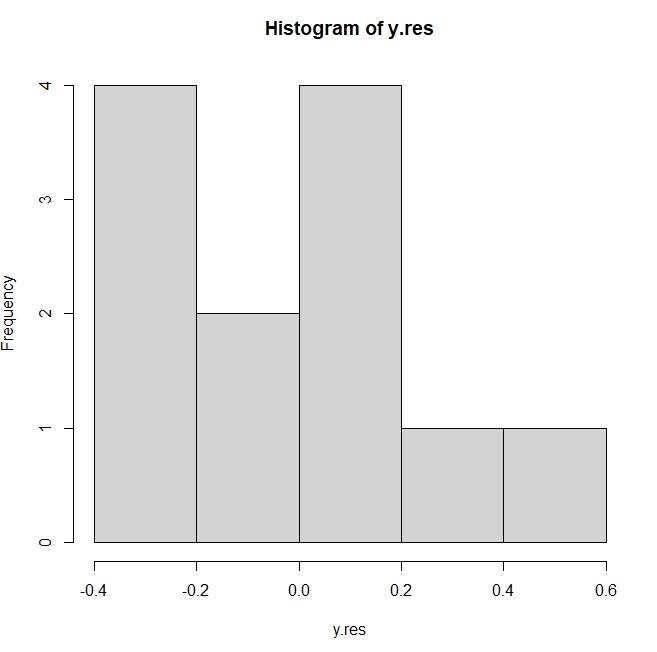
Since P<0.05, at the level of α=0.05, the regression coefficient of this case is statistically significant, and there is a regression relationship between height and age.

# residuals analysis

Histogram of residuals:

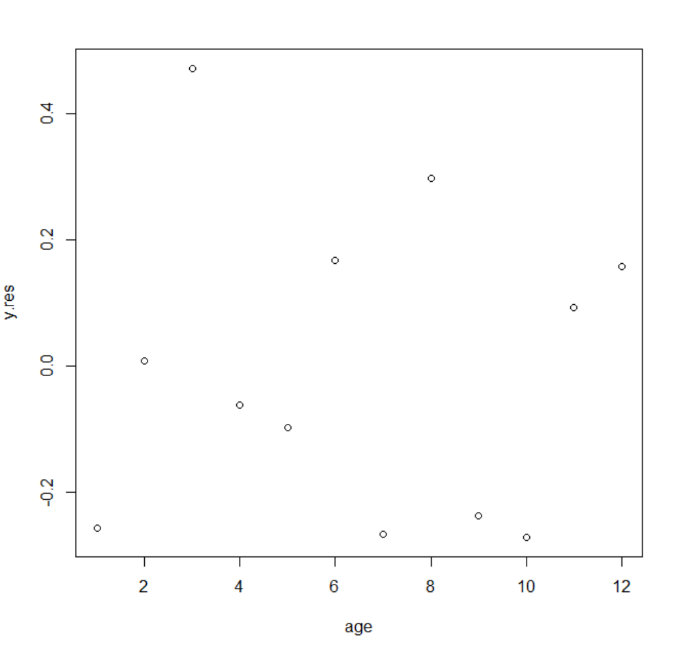
y.res = lm.reg$residuals

hist(y.res)



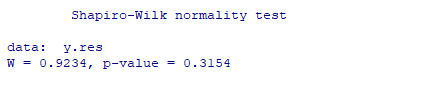
Plot residuals vs age:

plot(age, y.res)



# test of normality; use shapiro.test or qqplot

Shapiro.test(y.res)



**Example 2: A simulation example**

set.seed(1)

n = 100;

episilon = runif(n,-10,10);

x = rnorm(n,1,1);

alpha = 1; beta =5;

y = alpha + beta \* x + episilon;

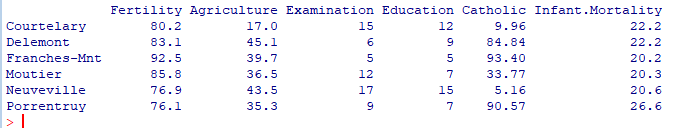
What is the estimation of alpha, beta? Is the estimation of beta equal to 5 (hypothesis test)?

Try n =10.

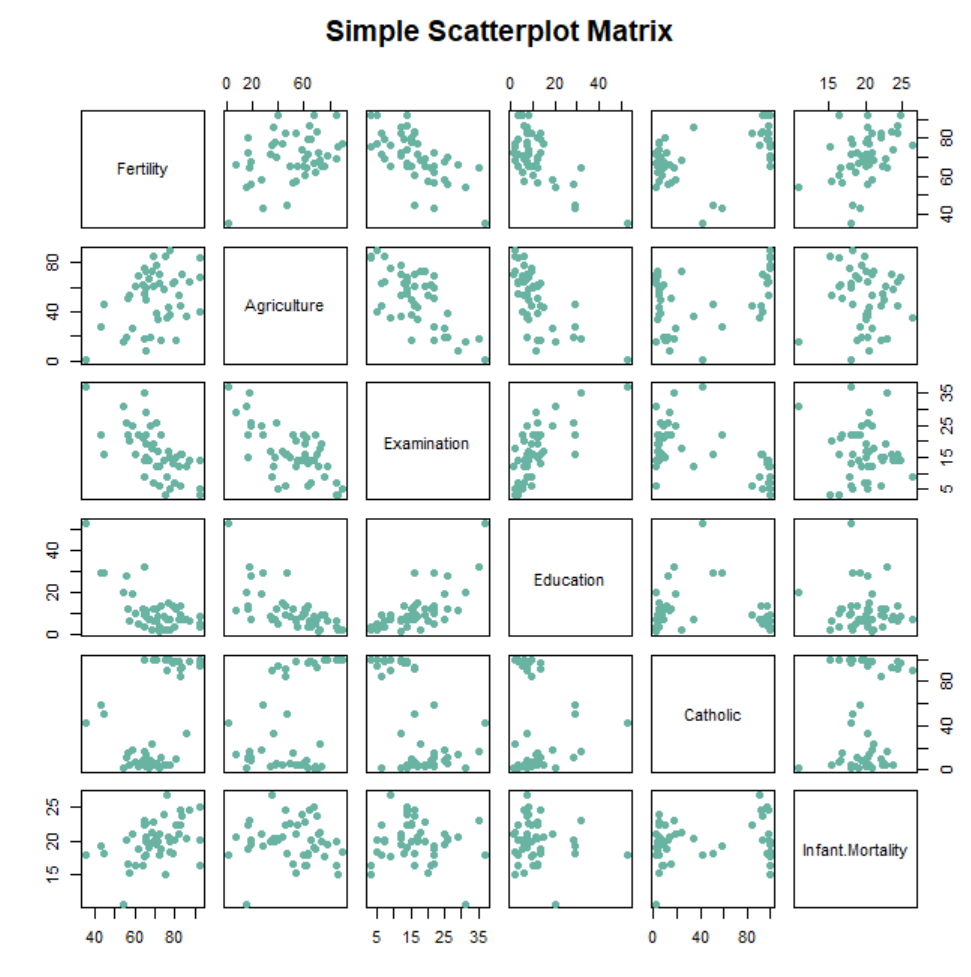
**Multiple linear regression (optional)**

Let's take the swiss data package as an example to explain multiple linear regression. This data package mainly talks about the fertility rate and socio-economic indicators of various cities in Switzerland.

head(swiss)



# Basic Scatterplot Matrix  
pairs(~Fertility + Agriculture + Examination + Education +Catholic + Infant.Mortality,data=swiss,main="Simple Scatterplot Matrix" ,pch=20 , cex=1.5 , col="#69b3a2")

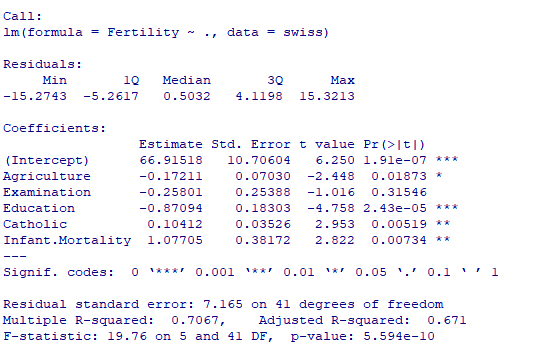


The model we intend to fit is:



lm2=lm(Fertility~.,data=swiss)

summary(lm2)



As can be seen from the output results above, the coefficients of Agriculture and Education are -0.17211 and -0.87094, respectively, and their p values are 0.01873 and 2.43e-05<0.05, so it can be considered that agriculture and education have a negative effect on Switzerland's fertility rate. Significantly!

The coefficient values of Catholic and Infant.Mortality are 0.10412 and 1.07705, respectively, and their P values are 0.00519 and 0.00734 respectively, which are less than 0.01. It can be concluded that Catholic and Infant.Mortality have a significant positive effect on Swiss fertility.

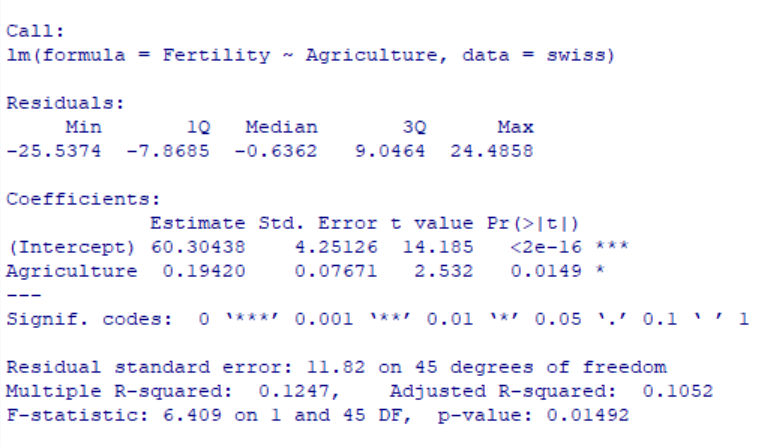
However, the P value of the coefficient of Examination is 0.31, which is greater than 0.05, so it cannot be considered that it has a significant impact on the Swiss fertility rate. So, what should we do next? You can try to delete the examination variable in the model.

The R2 value of the model is 0.671, indicating how much variance of Fertility can be explained by this model.

Now, we perform regression of Agriculture on Fertility:

lm3=lm(Fertility ~Agriculture, data=swiss)

summary(lm3)



What do you find? Why?

Check about marginal association, conditional association, joint association.