Lecture Notes:

Experimental Design

Maciej Szydłowski

October 2023

Owner

Name and Surname _____

Address _____

Telephone _____

Basic R

R is free software which is widely used in statistics. You can download R from http://www.r-project.org/. R Environment can be used for data management, matrix algebra, statistics, plots and graphs, and your own programmable computation.

Standard installation = R Environment + 8 standard packages. Other packages can be installed later.

Start R under Linux:

```
$ mkdir mydirectory
$ cd mydirectory
$ R
>…
>q( )
```

Start R in Windows:

Create your own directory
Edit shortcut and change „Start in" (eg. c:\mydirectory)

Set language:

```
Sys.setenv(LANG = "en")
```

Help in R:

```
help( table )   or   ?table
help.start( )
help.search
example( subject )
```

R language

V and v are two different variables
; separate 2 commands within a single line
this line is not executed
↑ use arrow up to repeat your last command

You can write all your commands in a text file and then

```
Linux:   source( "myRcommands.txt" )
Windows: menu File->Source
```

Objects

```
objects( )    #display all current objects
rm( x, y )    #removes objects x and y
```

Workspace

All objects create a workspace. Your workspace can be saved in your directory to continue calculation next day.

Arrays

```
# Multiple ways to create an array of values (object)
assign( x, c( 2.0,  3.0,  4.0 ) )
x  <-  c( 2.0,  3.0,  4.0 )
c( 2.0,  3.0,  4.0  )  ->  x
Y <- c( x, 0, x )
```

Algebra

```
+  -
*   /
^   sqrt()
log exp
sin   cos   tan
```

Array algebra

```
x                  #displays 2 3 4
max( x )           #gives 4
min( x )           #gives 2
length( x )        #gives 3
sum( x )           #gives 9
v  <-   2 * x  +  y  +  1
#        2  3  4    2  3  4    2   (x, x, 2)
#        2  3  4    2  3  4    2   (x, x, 2)
#        2  3  4    0  2  3    4   ( y=[x,0,x] )
#        1  1  1    1  1  1    1
#              ----------------------------
#        7  10 13  5  9  12 9     (v)
```

Statistical functions

```
mean( x )               #gives 3
sum( x ) / length( x )  #gives 3
var( x )                #gives 1
sd( x )
```

Regular sequences

```
z <- 1:7
z <- seq( -1.5,  1.0,  by=0.5 )
z <- seq( length=4, from=10, by=2 )
```

Arrays of strings

```
software <- c( "SAS",  "R" )
price <- c( "expensive.",  "free." )
paste( software, price, sep= " is " )
```

Select a value from array

```
a <- ( "dog", "cat", "mouse", "parrot")
a[2]              #cat
a[4]              #parrot
a[3:4]            #mouse parrot
a <- c( 6, 7, 8, 9 )
a[ -(2:3) ]                         #gives 6 and 9
a[ a >= 8 ]                         #gives 8 and 9
a[ a != 7 ]                         #gives 6, 8 and 9
a[ a>6 & a!=8 ]           #gives 7 and 9
a[ a<7 | a>8 ] -> b       #now b includes 6 and 9
```

Factor

```
music <- c( "jazz", "folk", "classic", "classic", "folk" )
style <- factor( music )
style
# Number of CDs in the collection
table( style )
price <- c( 40, 30, 61, 63, 40)
# average price of CD within each style of music
tapply ( price, style, mean )
```

Matrix

```
a <- 1:6  #1 2 3 4 5 6
dim( a ) <- c( 2, 3 )
      #1  3  5
      #2  4  6
b = a ;   b[ 2, 1 ] <- 0;  c <- cbind( a, b )
      #1    3    5    1    3    5
      #2    4    6    0    4    6
d<-rbind( c, c)
      #1    3    5    1    3    5
      #2    4    6    0    4    6
      #1    3    5    1    3    5
      #2    4    6    0    4    6
List - complex data structure
family <- list( father='Jan', mother='Maria', nkids=3, ages = c(5, 2, 1))
family[2]     #is  Maria
family[3]  #is 3
family[4] #is 5 2 1
family$father     #is  Jan
```

Data frame

```
#Example data for dogs
#ID Name Breed WeightA WeightB
#1023 Pirat spaniel 24 22
#1049 Aniel jamnik 18 61
#1219 Rabi spaniel 26 42.5
# read the data from the local file or internet
Dogs <- read.table( ”dogdata.txt" , header=TRUE )
Dogs <- read.table( ”http://…." , header=TRUE )
NewDogs <- edit( Dogs )       #data editing
```

Example data sets

```
data( )              # show data sets
data( ToothGrowth  ) # start working with …
library ( nls )      # use package nls and more data sets
data( )
data( Puromycin )
structure( Puromycin )
head( Puromycin )
```

Attach

```
data( )
data( chickwts )
chickwts$weight         # show all values of weight
attach( chickwts )      # now the chickwts is default data set
weight                  # show all values of weight
detach( )               # detach last attached data set
```

Select records from data set

```
subset( Dogs, Breed  == 'spaniel' & WeightA > 20) -> FatSpaniels
```

Merge data sets

```
# merge two data frames by ID and Country
total <- merge( data frameA,data frameB,by=c("ID","Country") )
```

Create a new variable

```
# create two categories
d$obesity <- ifelse( d$weight > 70, c("obese"), c("lean") )
```

Apply a function to each level of a factor

```
tapply(  WeightA, Breed, mean )
```

Create a function

```
addseven <- function ( x ) { x + 7 }
addseven( 10 )
```

Bar chart

```
## Example: chicken meat production: Africa 4.7 mln tons,
## Americas 39.4, Asia 31.0, Europe 14.5, Oceania 1.3.
count <- c(4.7, 39.4, 31.0, 14.5, 1.3)
lbls <- c( 'Africa', 'Americas', 'Asia', 'Europe', 'Oceania')
barplot( height=count, names.arg=lbls, ylab="Mln tons" )
```

Pie chart

```
## Example: Beef production: U.S. 10.8 mln tons Brazil 7.4 China 6.6
## Argentina 3.4 Australia 2.0 Russia 1.8 Others 25.6
slices <- c(10.8, 7.4, 6.6, 3.4, 2.0, 1.8, 25.6)
lbls <- c("US","Brazil","China","Argen.","Austr.","Rus.","Other")
pct <- round(slices/sum(slices)*100)
```

```
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=rainbow(length(lbls)), main="Beef")
```

Histogram

```
## Example: Chick weight from R example data set chickwts
hist(chickwts$weight)
```

Stem and leaf plot

A stem-and-leaf plot of a quantitative variable is a textual graph that classifies data items according to their most significant numeric digits. In addition, we often merge each alternating row with its next row in order to simplify the graph for readability.

```
stem(chickwts$weight)
```

Binomial distribution in R

Consider an infertility problem in a herd of 100 cows. A cow can be fertile or infertile (probability 0.03).

```
##calculate the probability for observing 10 infertile cows
##within a total 100 cows
dbinom( x=10, size=100, prob=0.03 )
##Calculate probabilities for all possible outcomes
##( 0, 1, 2, 3, … , 100 infertile cows)
dbinom( x=0:100, size=100, prob=0.03 )
##Calculate probabilities of observing
##maximum 2 infertile cows
sum ( dbinom( x=0:2, size=100, prob=0.03 ) )
pbinom( q=2, size=100, prob=0.03 )
##Generate seven random samples from the binomial ##distribution, each for a ##hundred cows
rbinom( n=7, size=100, prob=0.03 )
```

Multinomial distribution in R

```
##Calculate the probability for observing 43 blue, 33 yellow
##and 24 gold individuals,
##when color probabilities are 0.5, 0.3 and 0.2, respectively.
p = c( 0.5, 0.3, 0.2 )
x = c( 43, 33, 24 )
dmultinom( x=x, prob=p )
##Generate two random samples from the multinomial
##distribution for a hundred dogs
p = c( 0.5, 0.3, 0.2 )
rmultinom( n=2, size=100, prob=p )
```

Normal distribution in R

```
## example: x ~ N(10, 3)
mu = 10; s = 3;
curve( dnorm( x, mean=mu, sd=s), mu-3.5*s, mu+3.5*s, lwd=2 )
pnorm( 1, mu, s )
qnorm(0.95,0,1)
pnorm(1.644854,0,1)
rnorm(123, 10, 3)
```

[Tasks]

Task 1. Construct a bar chart describing swine meat production (in 1000 MT CWE): China 54700, EU-27 22450, US 10705, Brazil 3435, Russia 2300.

Task 2. Create a pie chart describing chicken meat production: Africa 4.7 ml tons, Americas 39.4, Asia 31.0, Europe 14.5, Oceania 1.3.

Task 3. Create a histogram of average body gain in pigs (column "gain") from a file swine-data.txt

Task 4. Make a stem and leaf plot for average daily body gain from a file, swine-data.txt

Two branches of statistics

- Descriptive statistics: Typically based on records for the entire population (e.g., sales of certain products, political votes). The objective is to summarize, organize, and describe the data.

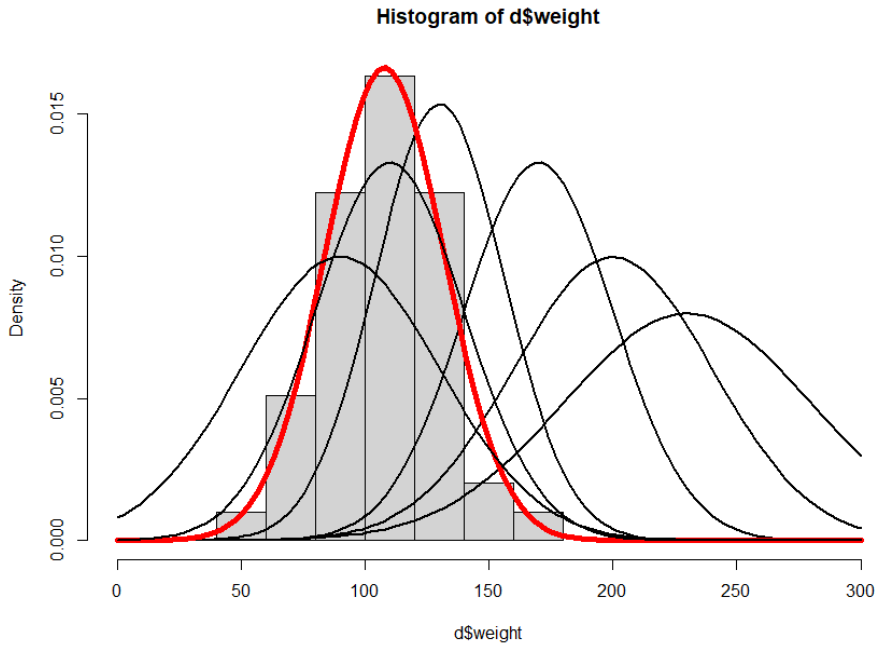
- Statistical inference: This is based on a limited sample of data. We utilize the information in the sample to draw conclusions about the population from which the sample was drawn. Data description is also essential in this context.

Probability distributions

We frequently observe that the data distribution exhibits a shape resembling one of the theoretical distributions. For instance, when we create a histogram using cow milk yield data, it often resembles a normal (Gaussian) distribution. Assuming that milk yield follows a normal distribution simplifies data analysis and facilitates predicting the level and range of yield in the future. Modeling the quantitative characteristics of animals using a normal distribution is a common statistical technique. There are infinitely many normal distributions, determined by two key parameters: the mean (μ) and variance (σ^2). In the modeling process, we select these parameters based on the data.

When we assume that milk yield follows a normal distribution, we can express it as $y \sim N(\mu, \sigma^2)$. From this point onwards, we operate under the belief that there exist true values for the mean and variance, which serve as the parameters of the normal distribution. Regrettably, these values remain unknown. Nevertheless, by analyzing performance data from a specific group of cows, we can attempt to approximate what these values might be. This process of determining the value of a parameter in a theoretical distribution is referred to as estimation.

In the example below, the histogram illustrates the distribution of chicken body weight values. This particular phenotypic trait was assumed to adhere to a normal distribution. The best fit for the normal distribution was achieved when the distribution's parameters were estimated using the available data.




```
subset( ChickWeight, Time == 10 ) -> d
hist( d$weight, xlim=c(0,300), freq=F)
curve( dnorm(x, mean=230, sd=50), add=T, col="black", lwd=2)
curve( dnorm(x, mean=200, sd=40), add=T, col="black", lwd=2)
curve( dnorm(x, mean=170, sd=30), add=T, col="black", lwd=2)
curve( dnorm(x, mean=130, sd=26), add=T, col="black", lwd=2)
curve( dnorm(x, mean=110, sd=30), add=T, col="black", lwd=2)
curve( dnorm(x, mean=90,  sd=40), add=T, col="black", lwd=2)
mu<-mean( d$weight)
s<-sd(d$weight)
curve( dnorm(x, mean=mu, sd=s),  add=T, col="red",   lwd=5)
```

Population and sample

A population refers to a group of individuals who share common characteristics, such as a group of Holstein-Friesian cows in Poland. The traits exhibited by animals within a population are the outcomes of a specific random process. The objective of statistical research is to gain insight into this process. For example, we seek to understand the expected level of production from the Holstein-Friesian breed in Poland.

One useful conceptual approach when dealing with populations is to imagine an infinite number of individuals within that population. The study of such a hypothetical population would provide us with infinitely precise knowledge about the random process that shapes the population.

However, populations are constrained by their actual numbers. The population of Holstein-Friesian cows in Poland stands at approximately 1.8 million. The larger

the population, the more accurately it mirrors the underlying random process. The considerable size of the Holstein-Friesian population undoubtedly enables us to determine the breed's potential with a high degree of precision. Nevertheless, some populations are quite small; for instance, certain dog breeds are represented in Poland by only a handful of individuals.

Collecting animal data can often be a labor-intensive and costly endeavor, leading to only a fraction of the existing population being accessible. For instance, in Poland, only 39% of cows are included in the production data collection program. When dealing with a very limited sample size, there's a risk that the sample may not accurately represent the entire population. Ideally, when drawing a sample from the population, efforts should be made to ensure that the sample is indeed representative.

It's essential to keep in mind that, whether you are analyzing the entire available population or just a sample from it, your data always reflects the outcomes of a certain random process. Our ultimate goal in analysis is to comprehend this underlying process.

Parameters and statistics

Theoretical probability distributions are characterized by specific parameters. For instance, a normal distribution is defined by two parameters: the mean and variance. These parameters are typically unknown, and we refer to them as population parameters. This terminology can be somewhat misleading, as it implies that these parameters can be derived from the data if the entire population is observed.

From the data collected, you can calculate the mean and variance, which we distinguish as the sample mean and sample variance. It's important to note that the population mean and sample mean are distinct terms, just as population variance and sample variance differ. In general, values computed from data are termed statistics.

Statistics serve multiple purposes: they can provide a concise summary of data, and they can also offer approximations for the values of parameters in theoretical probability distributions.

Experimental design

Experimental design often aims to optimize the process of sampling and calculations to achieve the most accurate approximation of multiple parameters while minimizing the required sample size.

In the realm of experimental design, our primary concern revolves around the analysis of data generated from an experiment. It is prudent to invest time and effort in organizing the experiment meticulously to ensure that we have the right type of data, and a sufficient amount of it, to effectively address the questions of interest.

Before embarking on an experiment, it is crucial to clearly identify the specific questions it is intended to answer. Additionally, we should make an effort to pinpoint all known or expected sources of variability within the experimental units. One of the central objectives of a well-designed experiment is to minimize the impact of these sources of variability on the responses to the questions at hand. In essence, the purpose of experiment design is to enhance the precision of our answers.

Quantitative data

Milk yield in kg (eg. 5210, 6730, 6209, 7504, 5411)
Milk fat percentage (eg. 3.22, 3.03, 3.26, 3.09, 3.12)
Litter size (eg. 7, 8, 7, 9, 8, 10, 9)

Discrete and continuous quantitative variables

When observations can take on a countable number of distinct values, we refer to the variable as 'discrete.' For instance, this includes variables like the number of accidents per year at an intersection or the count of piglets in a litter. A special case of discrete variables is the binary variable, where observations can only be zero or one.

On the other hand, when observations can assume any value within a continuous range, we label the variable as a 'continuous random variable.' Examples of this type of variable include the daily maximum temperature in Poznań or the milk yield of a cow.

Histogram

```
## Example: Chick weight from R example data set chickwts
hist(chickwts$weight)
```

Stem and leaf plot

A stem-and-leaf plot for a quantitative variable is a textual representation that organizes data items based on their leading significant numeric digits. Additionally, it's common practice to merge every alternating row with the following row to enhance the plot's readability.

```
stem(chickwts$weight)
```

Numerical descriptive measures for describing data

The two most common numerical descriptive measures are central tendency and variability. These measures aim to describe the center of the distribution of measurements and the extent to which measurements deviate from this center.

Measures of central tendency

We can use mode, median or mean.

Mode

The mode is the measurement that occurs most often (with the highest frequency).

Example: Slaughter weights: 962, 1005, 1033, 980, 965, 1030, 975, 989, 955, 1015, 1000, 970, 1042, 1005, 995. The mode is 1005.

The mode is also the midpoint of the interval with the highest frequency. From the histogram of chickwt data set and using R:hist(chickwts$weight) we can find the mode to be 325g.

Mode is often used as a measure of popularity.

Example: What food do you like most? Answers: seafood, beef, potatoes, beef. The mode is beef.

However, some data have two modes (bimodal distribution) or more. Mode is applicable to qualitative and quantitative data.

Median

Median is the middle value when the measurements are arranged in order of magnitude.

Example: Ages: 5, 7, 9, 10, 12, 17, 19, 22, 29, 33, 40, 42, 49. The median is 19.

The median of the even number of measurements is the average of the two middle values.

Example: Ages: 5, 7, 9, 10, 12, 17, 19, 22, 29, 33, 40, 42, 49, 53. The median is (19+22) / 2 = 20.5

There is only one median for the data set. Median is applicable to quantitative data only. Its value is not influenced by extreme measurements.

Mean

The mean as parameter is denoted by μ, whereas the mean from a sample is $\bar{y}$. In statistical inference, we use $\bar{y}$ to estimate μ.

The arithmetic mean (or simply mean or average) is the sum of measurements divided by the total number of measurements. Mean is $\sum y$ / n.

```
ages<-c(5, 7, 9, 10, 12, 17, 19, 22, 29, 33, 40, 42, 49, 53)
mean(ages)
```

There is only one mean for a data set. Mean is applicable to quantitative data only. Its value is influenced by extreme measurements.

Measures of variability (dispersion)

We can calculate range, percentile, variance and standard deviation, coefficient of variation.

Range

The range is the difference between the largest and smallest measurements.

```
ages<-c(5, 7, 9, 10, 12, 17, 19, 22, 29, 33, 40, 42, 49, 53)
max(ages) - min(ages)
```

Percentile

The p-th percentile of a set of n measurements arranged in order of magnitude is that has p% of the measurements below and (100-p)% above it. Specific percentiles of interest are the 25th, 50th, and 75th percentiles, often called the lower quartile, the middle quartile (median), and the upper quartile, respectively.

```
quantile(ages)
0%   25%   50%   75%  100%
5.00 10.50 20.50 38.25 53.00
```

Variance

Variance is a popular and practical measure of variability. Variance as a parameter is denoted by σ^2, and for a sample by s^2.

Sample variance (s^2) is the sum of the squared deviations from the mean divided by n-1. Sample variance is $\sum(y-\bar{y}) / (n-1)$.

```
#Example age at slaughter (in days): 220, 230, 228.
#Mean is 226 days.
#Deviations from the mean: -6, 4, 2.
#Squared deviations: 36, 16, 4
#Sum of squared deviations: 56
#Variance 56 / 2 = 28
var( c(220, 230, 228) )
```

Standard deviation (SD)

Standard deviation (s.d. or SD) is the positive square root of variance.

```
V <- var( c(220, 230, 228) )
sqrt( v )
sd( c(220, 230, 228) )
```

The standard deviation is often more intuitive than variance. For example, we can say that the standard deviation of the age at slaughter is 5.29 days.

In notation, the standard deviation as a population parameter is denoted by σ, while the standard deviation for a sample is denoted by s.

Coefficient of variation (CV)

The coefficient of variation (CV) is defined as the ratio of the standard deviation σ to the mean μ, or for a sample, s to $\bar{y}$.

What makes the CV valuable is its unit-independence; it's a dimensionless number. When comparing datasets with different units or significantly different means, it's more meaningful to use the coefficient of variation instead of the standard deviation.

However, it's important to note that the CV becomes less useful when the mean value is close to zero. In such cases, the coefficient of variation can approach infinity and becomes sensitive to even minor changes in the mean.

```
## Example: which trait is more dispersed:
## back fat thickness over shoulder or over back?
## Back fat thickness over shoulder (in mm): 20, 18, 17, 19, 27
## Back fat thickness over back: 21, 18, 14, 19, 22
shoulder <- c( 20, 18, 17, 19, 27 )
back <-  c( 21, 18, 14, 19, 22 )
cv <- function( x ) sqrt(  var( x )  ) / mean( x )
cv( shoulder )
cv( back )
```

Box plot

In descriptive statistics, a box plot, or boxplot, is a useful graphical representation for displaying groups of numerical data by highlighting their quartiles. Box plots often feature vertical lines extending from the boxes, known as whiskers, to indicate variability beyond the upper and lower quartiles. This is why they are sometimes referred to as 'box-and-whisker plots' or 'box-and-whisker diagrams.' Additionally, any outliers in the data may be plotted as individual points.

```
boxplot(count ~ spray, data = InsectSprays, col = "lightgray" )
```

Transformation of data to achieve desired mean and SD

The transformation x = (y - $\bar{y}$) / s standardizes the dataset, resulting in a dataset with a mean of 0 and a standard deviation of 1.

If you wish to transform the original data to a dataset with a different mean and standard deviation, for instance, a mean of 50 and a standard deviation of 20, your transformation should take the form: x = 50 + 20(y - $\bar{y}$) / s.

[Tasks]

Task 1. A study was conducted to determine the urine flow of sheep (in milliliters per minute). The urine flows of 10 sheep are recorded here: 0.7, 0.5, 0.5, 0.6, 0.5, 0.4, 0.3, 0.9, 1.2, and 0.9. (a) Calculate the mean, median, and mode for this sample data. (b) Imagine that the largest measurement is 6.8 instead of 1.2. How would this change affect the mean, median, and mode?

Task 2. Generate a synthetic dataset using the following R code: milk <- rnorm(9999, 8000, 1000). Now, you have fabricated observations of milk yield from 9999 cows. Create a histogram. Does the distribution of observations around the central point appear approximately symmetrical?

Task 3. Compute the sample mode, mean, and median for milk yields. You can utilize the sort function to arrange all values.

Task 4. Calculate the range, variance, standard deviation, and coefficient of variation for milk yields.

Task 5. Determine the lower and upper quartiles of milk yields.

Task 6. Simulate a dataset using the following R code: milkfat <- rchisq(9999, 5)/2. Now, you have fabricated observations of fat percentage in milk from 9999 cows. Create a histogram. Does the distribution of observations around the central point appear approximately symmetrical?

Task 7. Determine basic statistics for central tendency for milk fat percentage. Are they consistent with each other?

Task 8. Calculate the lower and upper quartiles of the milk fat percentage.

Task 9. Provide basic descriptive statistics for abdominal fat weight from the file 'swine-data.txt' (column AbdominalFat).

Task 10. Transform the data on abdominal fat weight to have a mean of 100 and a standard deviation of 10.

Task 11. Construct a boxplot illustrating the average daily gains (column gain) for each breed. Utilize the dataset in the file 'swine-data.txt'. There are three breeds: PLW (Polish Large White), PL (Polish Landrace), and L990 (artificial line 990) (column Breed).

Normal distribution

The normal distribution is frequently employed to model natural events that result from the influence of numerous factors, each having a minor impact. Take, for instance, the milk yield of a cow, which is affected by a multitude of factors, including genetics, seasonal variations, and diet. Typically, each individual factor exerts only a modest influence on the final milk yield. Consequently, the distribution of milk yield is often assumed to follow a normal distribution.

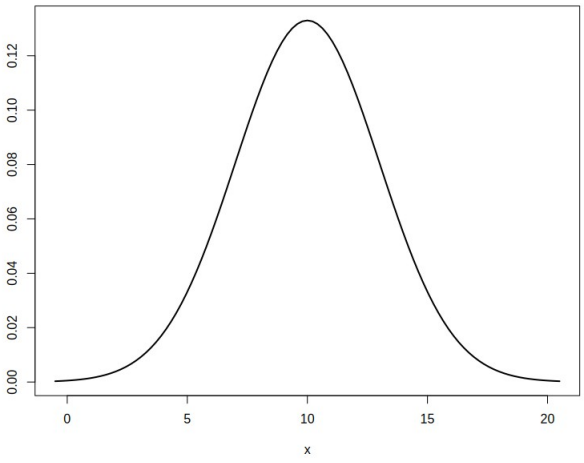
In cases where a variable is continuous, it becomes exceedingly challenging to assign a precise probability to each potential value. Instead, we express the probability of 'y' falling within a specific interval, which is referred to as probability density.

The normal distribution is characterized by two essential parameters: the mean and the variance. When we assume that a variable follows a normal distribution with a mean of 10 and a variance of 5, we can represent it as y~N(10,5).

Example: What is the probability, or more precisely, the probability density function, of y being equal to 11 if we assume that y~N(10, 5)? We can calculate this using R code: dnorm(11, 10, sqrt(5)), which yields a result of approximately 0.1614342.

Normal distribution in R

```
## example: x ~ N(10, 3)
mu = 10; s = 3;
curve( dnorm( x, mean=mu, sd=s), mu-3.5*s, mu+3.5*s, lwd=2 )
```



Cumulative distribution function (CDF)

The cumulative distribution function (CDF), also referred to as the distribution function, provides information about the probability that a random variable X, following a specific probability distribution, will have a value less than or equal to x. When dealing with continuous variables, the CDF represents the area under the probability density function from negative infinity up to x.

CDF in case of normal distribution

CDF for normal distribution is not easy to calculate, but we can use R:pnorm function from R.

Example: Let's consider a scenario where milk yield follows a normal distribution represented as y~N(8000, 250000). We want to determine the probability of observing a cow that produces less than 7000 kg of milk per lactation. Using the R code pnorm(7000, 8000, sqrt(250000)), we can calculate that this probability is approximately 0.0227. Thus, we can expect that only around 2% of cows from the population will produce less than 7000 kg of milk per lactation.

To visually compare the shapes of the probability density distribution and cumulative distribution function, you can use the following R code:

```
curve(pnorm, from=-4,to=4)
curve(dnorm, from=-4,to=4, add=T)
```

Standard normal distribution

A standard normal variable, denoted as 'z,' follows a normal distribution with mean 0 and variance 1, represented as z~N(0, 1). Essentially, a standard normal distribution is a specific type of normal distribution with these particular mean and variance values. There is only one standard normal distribution.

In many statistics textbooks, cumulative distribution function (CDF) values for the standard normal distribution are tabulated, allowing us to obtain the desired probabilities from a table without the need for additional calculations. For example, we can consult such a table and find that for N(0, 1), the CDF at 1.65 is equal to 0.95.

A variable 'y' following a normal distribution with a given mean and standard deviation can be transformed into a 'z' variable that follows the standard normal distribution, z~N(0,1), using the transformation formula: z = (y - mean) / SD.

Example: Let's consider a scenario where milk yield follows a normal distribution, denoted as y~N(8000, 250000). Now, what is the minimum milk yield required for the top 5% of cows? We can achieve this by standardizing 'y' to a standard normal variable 'z' using the formula z = (y - 8000) / 500. This transforms 'y' into 'z,' which follows a standard normal distribution, and these values are often available in tables. From a standard normal distribution table, we find that the cumulative distribution function (CDF) at 1.65 is equal to 0.95. This indicates that only 5% of values will exceed 1.65 in the standard normal distribution. To obtain the threshold in the original scale, we reverse the transformation: 1.65 * 500 + 8000 = 8825. Therefore, we can conclude that each of the top 5% of cows should produce a minimum of 8825 kg of milk per lactation to qualify as the best performers.

[Tasks]

Task 1. Consider basic phenotypes of farm animals. Provide three examples of traits that can be regarded as discrete variables and three examples of traits that can be viewed as continuous variables.

Task 2. Simulate milk yields for 9999 cows using the R code rnorm(9999, 8000, 500). Create a histogram. Describe the shape of the distribution. Is it bell-shaped?

Task 3. Use the following R code to simulate milk yields for two herds (A and B) and display histograms for each herd:

```
herdA <- rnorm(9999, 8000, 500)
herdB <- rnorm(9999, 9000, 500)
p1 <- hist(herdA)
p2 <- hist(herdB)
plot(p1, col='blue', xlim=c(0000,15000))
plot(p2, col='red', xlim=c(0000,15000), add=T)
```

Task 4. Both distributions are normal, with the same standard deviation but different means. Use a similar R code to simulate milk yields for another two herds with the same milk production level (8000 kg) but different standard deviations (1500 and 500). Describe the differences between the corresponding histograms.

Task 5. Determine the mean, mode, and median of the standard variable 'z'.

Task 6. Using the table with cumulative distribution functions (CDF) for the standard normal distribution (http://en.wikipedia.org/wiki/Standard_normal_table), (a) find the probabilities for z<-3, z<-2, z<-1, z<1, z<2, and z<3, and (b) find the 'z' values for CDF=0.95 and CDF=0.99.

Task 7. For a production trait that follows a normal distribution, calculate the number of individuals within the range defined as mean±1SD, mean±1.96SD, and mean±3SD.

Task 8. Assume that in a population of 9000 cows, a trait, milk yield in lactation, follows a normal distribution y~N(6000, 1000^2). How many cows do you expect to produce milk in the range of 5000-7000 kg per lactation?

Task 9. Construct a histogram for the average daily gain (column gain) from the dataset in a file named 'swine-data.txt'. (a) Determine if the distribution is similar to a normal distribution. (b) Calculate the mean and variance of the trait. (c) Using the calculated statistics, simulate 10000 synthetic values from a normal distribution and create a second histogram for the simulated values on the same graph. Describe the differences between the empirical distribution for daily gain and the idealized distribution for the simulated values.

Estimation of parameter

We can estimate a parameter of the population, which is a characteristic assumed to follow a certain distribution, based on the samples available to us. Generally, the larger the sample size, the more accurate the estimation becomes. However, gathering a large sample can be costly. Therefore, it's highly valuable to assess the quality of estimates derived from an existing sample before deciding to collect more data.

There are two primary types of estimation procedures. One involves computing a single value, referred to as a point estimate, along with the corresponding standard error of the estimate. Alternatively, we can compute a range, represented by two values, known as an interval estimate.

When planning an experiment, it's essential to consider how large the sample size should be in order to meet our criteria regarding the standard error.

Central Limit Theorem

When a random sample of 'n' measurements is repeatedly drawn from a population with a finite mean μ and standard deviation 'σ,' the histogram of the sample means (calculated from the repeated samples) tends to approximate a normal distribution as 'n' becomes large. The mean of these sample means remains μ while the standard deviation decreases to 'σ/√n.'

You can experimentally verify the Central Limit Theorem through simulation. Begin by simulating an entire population comprising 10,000 measurements of a specific trait. Then, extract a small sample of 100 measurements from this population and calculate the mean for each sample. Repeat this sampling process 700 times.

Finally, create a histogram based on the 700 calculated means. Below is an R code for conducting this simulation:

```
## Simulate a population from non-normal distribution
population <- rchisq( 10000, 5)
## Calculate the true population mean
mean( population )
## Calculate expected sampling SD
sqrt( var( population ) ) / sqrt( 100 ) ;
## Allocate space for storing 700 means
means <- rep(NA, 700) ;
## Repeat the procedure 700 times
for( i in (1:700) ) {
  onesample <- sample( population, 100) ;
  onemean = mean( onesample ) ;
  means[ i ] = onemean ;
}
## Histogram will show a normal distribution
hist( means )
## The mean should be similar to population mean
mean( means )
## Sampling SD should similar to the expected sampling SD
sd( means )
```

In each sampling procedure from the population, we obtain a unique sample of 100 measurements. Consequently, each sample provides a different mean denoted as '$\bar{y}$.' Notably, these means tend to fluctuate around the true population mean (μ). It is worth noting that the sampling variance tends to be higher when the population variance is high, and conversely, it decreases as the sample size increases. As a general rule, the standard deviation (SD) of the calculated means will be '$\sigma/\sqrt{n}$.' This prediction aligns with the outcomes observed during the simulation, where the SD of the 700 stored values approximated '$\sigma/\sqrt{n}$.' Furthermore, it is apparent that these 700 values follow a normal distribution.

It's noteworthy that the original distribution of the entire population was non-normal. The population was initially generated from the chi-square distribution, which is asymmetrical in nature. To gain insight into the shape of this distribution, you can create a histogram based on all 10,000 measurements.

Point estimation of mean (μ)

According to the Central Limit Theorem, the mean of a sample provides insights into the actual mean of the entire population. In essence, we consider the sample mean as an estimate of the population mean. Nevertheless, it's crucial to recognize that this estimate is contingent upon the specific sample selected, and as such, estimates can differ when different samples are drawn. The extent of this variation in estimates is contingent upon both the population's variance and the sample size.

Typically, we only have access to a single sample, and it's not feasible to observe the full spectrum of variation among multiple estimates. However, we can gauge the reliability of our single estimate by considering factors such as sample size and population variance.

Standard error of the mean (SEM)

The Standard Error of the Mean (SEM) represents the standard deviation of the estimate of a population mean derived from sample means. Grounded in the Central Limit Theorem, we understand that the standard deviation of various estimates of population means is 'σ/√n,' where 'σ' stands for the true population variation, which is often unknown. However, we can replace this elusive 'σ' with the sample's standard deviation ('s'). Consequently, the standard deviation of different estimates of population means becomes 's/√n.' Calculating the SEM is a straightforward process, and it serves as a valuable measure of the accuracy of our estimation. A larger SEM indicates a lower quality of estimation.

It's important to acknowledge that when calculating SEM, we substitute the unknown 'σ' with the sample's 's.' In reality, SEM is an estimate of an unknown true SEM. However, for practical convenience, we often treat SEM as a reliable measure of estimation quality.

In summary, when estimating the population mean, calculating SEM is a critical step in assessing the precision of our estimate.

Example: Let's consider an example where the weights of 71 chicks were recorded in a sample:

179, 160, 136, 227, 217, 168, 108, 124, 143, 140, 309, 229, 181, 141, 260, 203, 148, 169, 213, 257, 244, 271, 243, 230, 248, 327, 329, 250, 193, 271, 316, 267, 199, 171, 158, 248, 423, 340, 392, 339, 341, 226, 320, 295, 334, 322, 297, 318, 325, 257, 303, 315, 380, 153, 263, 242, 206, 344, 258, 368, 390, 379, 260, 404, 318, 352, 359, 216, 222, 283, 332.

We treat this sample of 71 chicks as a representation of a larger population, and our aim is to determine the mean of this broader population. The calculated mean of the sample is 261.3, which we consider as an estimate of the true population mean. The sample's standard deviation is 78.1, yielding a Standard Error of the Mean (SEM) of 78.1/√71, which equals 9.3. In statistical terms, we can express the estimated mean weight of the chicks as 261.3 ± 9.3.

At this point, we must assess whether this SEM meets our precision requirements for estimation. To determine this, you need to consider the context and specific needs of your analysis.

If you wish to calculate SEM for a sample using R, you can utilize the following code:

```
SEM <- function(x) { sqrt( var(x) / length(x) ) }
mean( chickwts$weight )
SEM( chickwts$weight )
```

Assuming a population variance of $\sigma^2 = 500$, you can investigate how the Standard Error of the Mean (SEM) decreases as the sample size increases using the following R code:

```
SEM <- function(n) sqrt( 500 / n )
curve( SEM, from=10, to=3000 )
```

Indeed, the plot suggests that as the sample size increases, the Standard Error of the Mean (SEM) decreases. However, there comes a point where increasing the sample size further doesn't significantly reduce the SEM. This is a key concept in statistics known as the law of diminishing returns regarding sample size.

In practical terms, it means that once you have a sufficiently large sample size, collecting even more data may not substantially improve the precision of your mean estimate. Therefore, it's essential to strike a balance between the desired level of precision and the practicality of collecting more data, as collecting larger samples can be resource-intensive and costly.

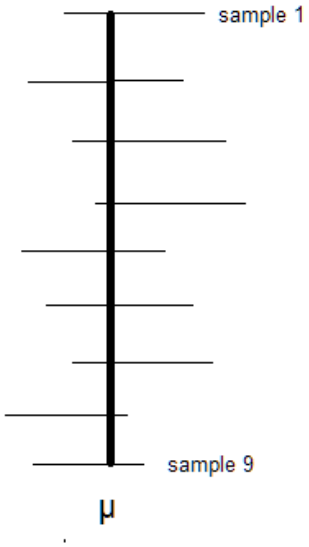
Interval estimate of mean

Interval estimation is an alternative approach to point estimation, first developed by the Polish mathematician and statistician Jerzy Neyman. This concept is visually represented in a diagram. Within this diagram, a vertical line symbolizes the unknown true population mean, while nine horizontal lines represent interval estimates derived from nine separate random samples. As these samples are random, the intervals they produce vary between each sample.

The fundamental idea behind interval estimation is to construct, based on a single sample, an interval that would encompass the true mean in the majority of cases, assuming multiple random samples were taken. This interval is known as a confidence interval. Typically, statisticians are interested in confidence intervals with a confidence level of 95% or 99%. This means that a 95% confidence interval will include the true mean in about 95% of similar cases.

It's crucial to note that when you construct a 95% confidence interval, it doesn't imply that there's a 95% probability that the true mean falls within that interval. The interpretation of confidence intervals isn't entirely intuitive and may not be easily explained.

To construct 95% confidence interval we can use the formula $\bar{y} \pm 1.96s/\sqrt{n}$. Here: $\bar{y}$ represents the sample mean, s denotes the sample standard deviation, n signifies the sample size. This formula is suitable for larger samples, typically when $n>30$. However, for smaller samples, an alternative formula or statistical approach becomes necessary.



The coefficient 1.96 is derived from the standard normal distribution, denoted as N(0, 1). In this distribution, it is established that 95% of the most probable values fall within the interval (-1.96, +1.96).

For constructing a 99% confidence interval, we employ a similar logic and use the formula: $\bar{y} \pm 2.58\ s/\sqrt{n}$. By using this formula, we create a wider interval that captures 99% of the most likely values.

Let's revisit the dataset of body weights recorded for 71 chicks. The sample statistics include a mean of $\bar{y}$ = 261.3g, a sample standard deviation of s=78.1, and a sample size of n=71. Consequently, we can calculate the 95% confidence interval for the mean as follows: $261.3 \pm 1.96 \times 78.1/\sqrt{71} = 261.3 \pm 18.2$. This interval, ranging from 243.1 to 279.5, constitutes a 95% confidence interval for μ, the true population mean. In simpler terms, we can express a 95% level of confidence that the average weight of chicks falls within the range of 243.1 to 279.5 grams.

Choosing the sample size for estimating population mean

Data collection indeed involves costs, and determining an appropriate sample size is a crucial aspect of research design. The choice of sample size should be based on several factors:

1. Desired Precision: Determine how precise you want your estimate to be. If you need a very precise estimate, you'll require a larger sample size. If rough estimates are acceptable, a smaller sample size may suffice.

2. Variability: The amount of variability or spread in your data affects the required sample size. Highly variable data requires a larger sample to achieve a specific level of precision.

3. Confidence Level: The level of confidence you desire in your estimate plays a role. Common confidence levels include 95% and 99%. Higher confidence levels require larger sample sizes.

4. Margin of Error: Decide on an acceptable margin of error. A smaller margin of error necessitates a larger sample size.

5. Population Size: In some cases, the size of the population you're sampling from can impact the required sample size. For very large populations, you may use an adjustment to the formula.

6. Resources: Consider practical constraints, such as time and budget, which may limit the sample size.

To determine the sample size, you can use sample size calculation formulas tailored to your study design, often involving standard error and critical values from the normal distribution or t-distribution. The formula will typically take into account the desired level of confidence, margin of error, and estimated population standard deviation.

In practice, software tools or statistical calculators can simplify this process. It's also common to conduct a pilot study to estimate the population standard deviation, which can then be used in the sample size calculation.

Remember that the choice of sample size is a trade-off between cost and precision. A larger sample provides more precise estimates but may be costlier and more time-consuming to collect. Balancing these factors is essential for efficient and effective data collection.

If a 100(1-α) confidence interval is to be $\bar{y} \pm E$, then the required sample size is $n = (z_{\alpha/2})^2 \sigma^2 / E^2$. Here, α is a small value (traditionally 0.05 or 0.01) called the level of significance, $z_{\alpha/2}$ is the value for which CDF of standard normal distribution is 1-α/2.

Note that determining the sample size to estimate μ requires knowledge of the population variance σ2. We can obtain an approximate sample size by estimating σ^2, using one of these two methods: Use information from a prior experiment to calculate a sample variance s^2, and then replace σ^2 with s^2. Or use the information on the range of the observations and replace σ^2 with range/4.

Example: A biologist aims to assess the impact of an antibiotic on the growth of a specific bacterium by measuring the mean bacterial quantity per culture plate. From prior experiments, it is known that the standard deviation (SD) of the bacterial

quantity is approximately 13 cm^2. To determine the number of culture plates needed for estimating the bacterial quantity with a 99% confidence interval and a half-width of 3 square centimeters, we require a 99% confidence interval in the form of $\bar{y}\pm3$. We can employ the following formula: $n = (z_{\alpha/2})^2\sigma^2 / E^2$, where we substitute the unknown variance σ^2 with 13^2. Here, the level of significance is $\alpha=0.01$, and we find the value of z for which the cumulative distribution function (CDF) is 0.995, which is approximately 2.58 based on tabulated values. Therefore, $n = (2.58^2)(13^2)/(3^2) \approx 2972$. This calculation suggests that around 2972 culture plates should be developed to estimate the bacterial quantity with a 99% confidence interval and a half-width of 3 square centimeters.

[Tasks]

Task 1. a. Utilize the data for the average daily gain (found in the 'gain' column) from the file 'swine-data.txt' to estimate μ (the population mean) and calculate the Standard Error (SE) for this estimated mean. b. Calculate the 95% and 99% confidence intervals for the mean.

Task 2. Create a plot illustrating how the true population variance (σ^2) impacts the Standard Error of the Mean (SEM) while keeping the sample size fixed at n=90. Explore whether n=90 is consistently appropriate.

Task 3. For an initial screening of 10 pigs, where breeders recorded 10 measurements of back fat thickness over the shoulder (in cm) - 2.5, 3.1, 2.6, 2.2, 3.3, 2.9, 2.0, 2.2, 2.0, 1.4, calculate the required sample size needed to estimate the population mean of this trait with a 95% confidence interval width of 1 mm.

A statistical test for μ

A statistical test aims to provide an answer to a fundamental question: 'Is the population mean equal to a specified value μ_0?' When we have access to measurements for the entire population, answering this question is straightforward. However, when we only have a sample from the population, statistical testing becomes essential to address this query.

Statistical testing operates on the principle of proof by contradiction and typically comprises five key components:

1. Null Hypothesis (H_0): This is the hypothesis we aim to challenge or test.

2. Research (Alternative) Hypothesis (H_1 or H_A): This is the hypothesis we seek to establish if we can refute the null hypothesis.

3. Test Statistic: A statistical value calculated from the sample data.

4. Rejection Region: A range of values for the test statistic that, if met or exceeded, leads to the rejection of the null hypothesis.

5. Conclusion: The decision based on whether the null hypothesis is rejected or not.

For instance, consider the scenario where we wish to investigate whether the average milk yield in the Polish population of cows exceeds 6000 kg per lactation. Our research hypothesis, denoted as $\mu > 6000$, reflects what we aim to prove. To achieve this, we employ proof by contradiction and challenge another hypothesis known as the null hypothesis, denoted as H_0: $\mu = 6000$. The null hypothesis posits that the population mean is exactly 6000 kg, which contradicts our research hypothesis. By demonstrating that H_0 is improbable or false, we infer that the alternative hypothesis is plausible. In essence, we conclude that it is unlikely for the population mean μ to be lower than 6000 kg.

Now, we have the ability to collect a sample from a population of cows and calculate the sample mean, denoted as $\bar{y}$. What can we anticipate regarding the potential values of $\bar{y}$ if the null hypothesis holds true?

When the null hypothesis is true ($\mu = 6000$), we should expect $\bar{y}$ to fluctuate around the actual population mean, which is $\mu = 6000$. In this scenario, it becomes highly probable to obtain $\bar{y}$ values that are in close proximity to the true mean of 6000 kg. Conversely, it is considerably less likely to obtain $\bar{y}$ values that deviate significantly from 6000.

In essence, if the null hypothesis holds true, and μ is indeed 6000, it becomes improbable to draw a sample that produces a $\bar{y}$ vastly different from 6000 kg. However, what if our calculated $\bar{y}$ turns out to be significantly greater than 6000 kg? One could argue that our single sampling was an unusual occurrence, although still possible. Alternatively, we might assert that our null hypothesis is unlikely to be valid. In other words, it becomes improbable that μ equals 6000 kg.

It is valuable to ascertain the probability of obtaining a specific value of $\bar{y}$ under the null hypothesis. In statistical testing, we employ certain test statistics that follow a known probability distribution when the null hypothesis holds true. In our example, we can utilize a test statistic based on the standard normal distribution. This statistic is represented as $z = (\bar{y} - \mu_0) / (s/\sqrt{n})$. The expectation is that this statistic adheres to a standard normal distribution, making it relatively straightforward to calculate the probability associated with a particular value of the statistic when the null hypothesis is valid.

Suppose we've collected a sample of n = 100 milk yield observations and computed a sample mean of $\bar{y}$ = 6073 kg, along with a sample standard deviation of s = 350 kg. The test statistic can be calculated as z = (6073 - 6000) / (350/√100) = 2.09. It's essential to note that the value z = 2.09 falls within the tail of the standard normal distribution and corresponds to a very low probability. In general, within the standard normal distribution, there is only a 5% probability of obtaining a z value greater than 1.65, and merely a 1% chance of acquiring a z value exceeding 2.33.

In statistical testing, it's crucial to establish a rejection region. It's reasonable to define the rejection region for the test statistic z as [1.65, ∞]. This rejection region encompasses the most unexpected values of z under the assumption of the null hypothesis. If the null hypothesis holds true, the likelihood of obtaining a sample that produces a z-value within this region is very low, approximately 5%. Consequently, if our calculated z falls within this region, it raises doubts about the validity of the null hypothesis. In cases where z > 1.65, we make the decision to reject the null hypothesis. This rejection leads us to accept the alternative hypothesis as true, employing a proof by contradiction.

In our specific scenario, we have calculated a z-value of 2.09, which falls within the defined rejection region. Therefore, based on the available sample, we can reasonably conclude that the true population mean is indeed greater than 6000 kg.

Now, let's consider a different scenario where we calculate the test statistic to be, for example, z = 1.1. In this case, this value does not fall within the rejection region. Consequently, we cannot reject the null hypothesis, nor can we affirm the alternative hypothesis. Does this imply that the null hypothesis is true? No, it simply means that based on the data at hand, we cannot draw any definitive conclusions regarding the validity of either hypothesis. It suggests that you might need to increase the sample size to provide stronger evidence in favor of the alternative hypothesis. However, if you have a significantly large sample size and still cannot reject the null hypothesis, it would strongly suggest that the null hypothesis aligns closely with reality.

The rejection region typically encompasses the 5% most unexpected values of the test statistic. Therefore, in statistical testing, it's common to use a 5% rejection region, equivalent to a significance level of α = 0.05. Another commonly used significance level is α = 0.01. It's essential to decide on the significance level before conducting the test. It's worth noting that if you choose to test at α = 0.05 and you reject the null hypothesis, there is a 5% risk of making a Type I error (incorrectly rejecting a true null hypothesis). Similarly, if you test at α = 0.01 and reject the null hypothesis, there is a 1% risk of committing a Type I error. These levels of risk, whether 5% or 1%, are typically considered acceptable in most cases.

Now we can summarize our statistical testing procedure in the 5 parts:

H0: μ > 6000 kg
HA: μ = 6000 kg
Test statistic: z = 2.09
Rejection region: $z_{0.05}$ > 1.65 (α=0.05)

Conclusion: Under the significance level α=0.05 the available data support the hypothesis that mean value of milk yield in the population is greater than 6000 kg.

Indeed, it's crucial to consider the level of risk associated with the decision made in statistical testing. As mentioned, when you choose a significance level of α = 0.05 (corresponding to a 5% risk), you are accepting that there is a 5% chance of making a Type I error, which means incorrectly rejecting a true null hypothesis. If you opt for a more stringent significance level like α = 0.01, you are reducing that risk to 1%.

However, reducing the risk of Type I error comes at a cost. To achieve a lower risk, you often need a larger sample size because you need more evidence to confidently reject the null hypothesis. So, if you are aiming to accept only a 1% risk of supporting the hypothesis that μ > 6000 when it's actually incorrect, you may need to collect more data to achieve the desired level of confidence.

The choice of significance level should be based on the specific context of your study, the consequences of making a Type I error, and the resources available for data collection. It's a balance between the level of confidence you require and the practical constraints of your research.

[Tasks]

Task 1. Considering the example mentioned earlier, what would be the rejection region if you choose a significance level of α = 0.01?

Task 2. In the context of the hypothesis about the average milk yield, (a) what conclusion would you draw if the calculated test statistic were z = 0.56? (b) How about if it were z = 2.9?

The t-test for mean (μ)

In order to calculate the test statistic for the population mean, we often require the true population standard deviation, denoted as σ. However, σ is frequently unknown in practice and is replaced with the sample standard deviation, denoted as s. This substitution generally does not impact the testing procedure, except for

instances when the sample size is small. When dealing with small samples (e.g., n < 30), the statistical test based on the standard normal distribution may not provide sufficient precision. Therefore, an alternative test based on Student's t distribution has been proposed.

Student's t distribution closely resembles the standard normal distribution, but it varies based on a parameter known as degrees of freedom (df). Below, you will find an R code snippet that plots two t distributions with 1 and 30 degrees of freedom, along with a superimposed normal distribution for comparison.

```
curve( dnorm, from=-4, to=4 )
curve( dt( x, df=1), from=-4, to=4, col="blue", add=T )
curve( dt( x, df=30), from=-4, to=4, col="red", add=T )
```

You can observe that the Student's t distribution with df=30 is nearly indistinguishable from the standard normal distribution.

A statistical test that relies on the Student's t distribution is commonly referred to as a t-test. It performs well for both small and large sample sizes, making it more widely used than tests based solely on the standard normal distribution. In the case of the milk yield data, the test statistic is calculated as $t = (\bar{y} - \mu_0) / (s/\sqrt{n}) = 2.09$. The test employs n-1 degrees of freedom, in this case, df=100-1=99. We can consult a table containing percentage points of the t-distribution or utilize the R code qt(0.95, df=99), which yields a value of 1.66. The rejection region is $t_{0.05,\ 99}>1.66$, and it encompasses the test statistic value of 2.09. Consequently, we can reject the null hypothesis and conclude that the alternative hypothesis is valid.

The level of significance and p-value

In our previous example, we chose to perform a hypothesis test on the population mean at a significance level of $\alpha = 0.05$. This significance level establishes the rejection region as $z0.05 > 1.65$. Then, we can examine whether the calculated test statistic falls within this rejection region or not. A more convenient approach is to utilize the concept of a p-value. The p-value represents the probability of obtaining a test statistic that is at least as extreme as the one observed, assuming that the null hypothesis is accurate. Therefore, when the p-value is lower than α, we have grounds to reject the null hypothesis. Although computing the p-value can sometimes be complex, most statistical software packages are capable of performing this calculation. The p-value is a universal output and plays a crucial role in many statistical tests. Its convenience lies in the fact that it simplifies decision-making; all we need to do is compare the p-value to α (e.g., 0.05, 0.01, 0.001) to draw a final conclusion or make a decision.

In the example of testing the mean of milk yield, our test statistic was 2.09. This statistic is based on the standard normal distribution. The probability of

obtaining a test statistic as extreme (in this case, as high) as the observed one is represented by the p-value, which is calculated as follows: p-value = P(z ≥ 2.09) = 1 - CDF(2.09) = 0.0183. The p-value of 0.0183 is lower than α=0.05, but not as low as α=0.01.

Let's revisit the data on the weight of chicks. We can investigate whether the population mean is greater than 250g using the following R code.

```
t.test( chickwts$weight, mu=250, alternative="greater" )
```

```
One Sample t-test
data:  chickwts$weight
t = 1.2206, df = 70, p-value = 0.1132
alternative hypothesis: true mean is greater than 250
95 percent confidence interval:
 245.8648      Inf
sample estimates:
mean of x
 261.3099
```

The computed p-value is 0.1132; therefore, we can conclude that the sample does not yield sufficient statistical evidence to support the claim that the population mean is greater than 250g.

In some cases, the p-value is simply referred to as the 'level of significance' and is denoted by P or p. In many situations, we can report the results of a specific statistical test along with the p-value and leave the interpretation to the readers. It's important to note that the smaller the p-value, the stronger the statistical evidence in favor of the research hypothesis. For example, a statistical test with a level of significance of p=0.002 provides more compelling evidence for rejecting the null hypothesis than a test with p=0.020. Ultimately, it is the responsibility of the readers to assess whether the level of statistical evidence presented is sufficient to support a particular research hypothesis.

Type I error and Type II error

When conducting hypothesis testing, we face the possibility of committing two types of errors.

Type I Error (α): This occurs when we mistakenly reject the null hypothesis when it is actually true. In other words, we conclude that there is a significant effect or difference when there isn't one. The probability of making a Type I error is denoted by the symbol α (alpha).

Type II Error (β): This error takes place when we fail to reject the null hypothesis when it is, in fact, false, and the alternative hypothesis (research hypothesis) is true. Essentially, it means we miss a real effect or difference. The probability of making a Type II error is denoted by the symbol β (beta).

Balancing these two types of errors is crucial in hypothesis testing. Typically, as we decrease the risk of Type I errors (by choosing a smaller α level, like 0.01 instead of 0.05), we increase the risk of Type II errors. Finding an appropriate balance depends on the specific context and consequences of each type of error in a given study or experiment.

	Null hypothesis	
Decision	True	False
Reject H0	Type I error α	Correct 1-β
Accept H0	Correct 1-α	Type II error β

Indeed, there's a trade-off between Type I and Type II errors in hypothesis testing, and it's typically not possible to minimize both simultaneously. Statisticians and researchers must make a choice based on the specific goals of their study and the consequences of each type of error.

As you mentioned, statisticians often set a predefined level of significance (α) that represents the maximum acceptable probability of committing a Type I error. This level is determined by factors such as the importance of the decision being made and the potential consequences of incorrectly rejecting the null hypothesis.

The choice of α is somewhat arbitrary and depends on the field of study and the prevailing standards. Commonly used values for α include 0.1, 0.05, 0.01, and 0.001. A smaller α (e.g., 0.01 or 0.001) indicates a stricter criterion for rejecting the null hypothesis, which reduces the risk of Type I errors but increases the risk of Type II errors. Conversely, a larger α (e.g., 0.1) is more permissive regarding the rejection of the null hypothesis, which reduces the risk of Type II errors but increases the risk of Type I errors.

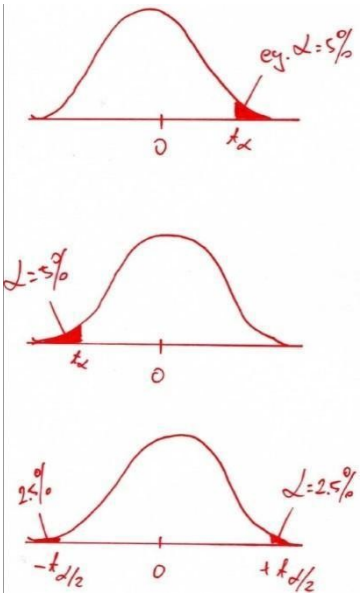
Choosing an appropriate α level involves considering the specific research question, practical implications, and the relative importance of avoiding each type of error.

One-tailed and two-tailed tests

In our initial test of the mean milk yield, we anticipated that data strongly in favor of the alternative hypothesis ($\mu > 6000$ kg) would yield extreme values of the test statistic on the right side of the distribution. Consequently, we positioned the 5% rejection region on the right side. However, let's contemplate another scenario with H_A: $\mu < 6000$. In such a case, conceivable sample observations that

robustly support this alternative hypothesis should generate extreme values of the test statistic on the left side of the distribution. Consequently, it becomes necessary to position the rejection region on the left side.

The placement of the rejection region in hypothesis testing depends on the directionality of the alternative hypothesis (H_A) and the specific research question.



When testing whether the population mean is greater than a specified value (H_A: $\mu > \mu_0$), you would typically place the rejection region on the right side of the distribution because extreme values in the right tail of the distribution would provide strong evidence against the null hypothesis.

Conversely, when testing whether the population mean is less than a specified value (H_A: $\mu < \mu_0$), you would place the rejection region on the left side of the distribution. In this case, extreme values in the left tail of the distribution would support the alternative hypothesis.

In two-tailed tests, where you are testing whether the population mean is not equal to a specified value (H_A: $\mu \neq \mu_0$), you would place rejection regions in both tails of the distribution to detect extreme values in either direction.

The choice of the rejection region's placement depends on the specific research question and the direction of the hypothesis being tested. This flexibility allows statisticians to tailor hypothesis tests to different scenarios and research goals.

In many cases, the alternative hypothesis is less specific and merely states H_A: $\mu \neq \mu_0$. In such instances, it is appropriate to allocate a 5% rejection region on both sides of the distribution. It is reasonable to position one half of the rejection region on the far left of the distribution and the other half on the far right.

Let's contemplate a test employing $\alpha = 0.05$ and the z statistic, which follows a standard normal distribution. If the alternative hypothesis indicates "greater than," the rejection region becomes [1.644854, infinity]. If the alternative hypothesis suggests "less than," the rejection region is [-infinity, -1.644854]. Lastly, if the alternative hypothesis implies "not equal to," the rejection regions are [-infinity, -1.959964] and [1.959964, infinity]. It's worth noting that for the "not equal to" alternative hypothesis, we can easily check if the absolute value $|z| > 1.959964$.

You can calculate these thresholds using the following R code:

```
alfa = 0.05
qnorm( alfa )
qnorm( 1 - alfa )
qnorm( ( 1 - alfa / 2 )  )
```

Now, we can summarize the statistical t test for population mean:

H_0: $\mu = \mu_0$
H_A: One of the following:
(a) $\mu > \mu_0$,
(b) $\mu < \mu_0$,
(c) $\mu \neq \mu_0$

Test statistic: $t = (\bar{y} - \mu_0) / (s/\sqrt{n})$

For a probability α of a Type I error and df=n-1,
(a) reject H0 if $t>t_\alpha$,
(b) reject H0 if $t<-t_\alpha$,
(c) reject H0 if $|t|>t_{\alpha/2}$,

[Tasks]

Task 1. Feed Conversion Ratio in Pigs: A group of breeders is considering investing in a newly developed feeding system for pigs, with the hope that it will result in a lower feed conversion ratio (the ratio of consumed fodder to body gain). Before making their final decision, they collected data from 33 pigs that were raised using this technology. Here are the feed conversion ratios for these pigs:
3.08 2.40 2.56 3.25 2.74 3.61 2.80 3.04 2.37 3.42 2.82
2.86 2.23 3.08 2.63 2.93 2.66 2.23 2.24 3.07 2.97 3.59

2.96 2.78 2.70 3.44 2.69 2.22 3.31 2.92 3.08 3.51 2.78
The breeders want to determine if they can be certain that the average feed conversion ratio in the entire pig population would drop below 3.0. What advice would you provide to the breeders based on this data?

Task 2. EU Law Impact on Pig Performance: A recent European Union law mandates that pigs "must have permanent access to a sufficient quantity of material to enable proper investigation and manipulation activities, such as straw, hay, wood, sawdust, mushroom compost, peat, or a mixture of such materials." This law has raised questions among pig breeders about its potential impact on pig performance. To investigate, the breeders collected data from pigs raised in this new environment. For each pig in the sample, they recorded the number of days it took to reach a weight of 100 kg, starting from 25 kg. The data can be found in the file 'swine-data.txt,' specifically in the 'Days100' column. Your task is to test the hypothesis that, under these new conditions, pigs require more or fewer days to reach 100 kg in body weight compared to the previously established average of 90.2 days.

Computing β

In past years, the average live weight of a farmer's steers prior to slaughter was 172.5 kg. This year, the farmer selected a random sample of 50 steers to be fed on a new diet. The data from this sample are as follows: n=50, sample mean ($\bar{y}$)=177.1 kg, sample standard deviation (s)=16.0 kg. The objective is to test the research hypothesis that the mean live weight for steers on the new diet is greater than 172.5 kg, using a significance level of α=0.01.

Null Hypothesis (H0): $\mu = 172.5$ kg

Alternative Hypothesis (H_A): $\mu > 172.5$ kg

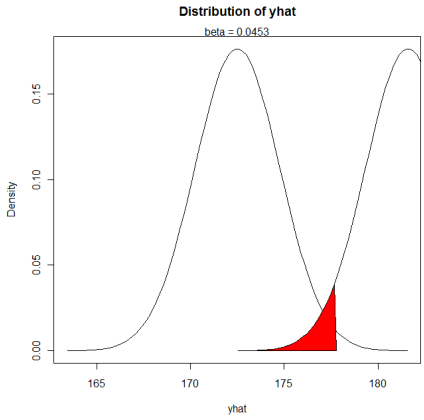
To calculate the test statistic (z), we use the formula: $z = (\bar{y} - \mu) / (s/\sqrt{n})$, which results in z = 2.03.

For a one-tailed test with α=0.01, the rejection region is z>2.33. Thus, based on our data, we cannot reject the null hypothesis. To reach a conclusion about accepting H0, the experimenter would need to compute β. If β is found to be small for reasonable alternative values of μ, then H0 can be accepted.

Next, let's consider the sampling distribution of $\bar{y}$ under two scenarios: one where H_0 is true (μ=172.5 kg) and another where the true mean is 181.6 kg. The red area on the accompanying figure represents the probability β. It's essential to evaluate β under different situations. Additionally, what if the true mean is 177 kg? And what if you have a larger sample, say n=100?

You can explore the variations in β under these different conditions and sample sizes by utilizing and manipulating the provided R code.

```
mean0 = 172.5;  tmean = 181.6; sd=16; n=50 ;
## sampling SD
yhatsd = sd / sqrt( n ) ;
## some possible values of yhat and their density under H0
x <- seq( -4, 4,length=100 )* yhatsd + mean0 ;
density <- dnorm( x, mean0, yhatsd )
plot(x, density, type="l", xlab="yhat", ylab="Density") ;
## some possible values of yhat and their density under tmean
x <- seq( -4, 4, length=100 ) * yhatsd + tmean ;
density <- dnorm( x, tmean, yhatsd ) ;
lines( x, density ) ;
## add red area under curve
lb = tmean-4*yhatsd ; ub = mean0+2.33*yhatsd ;
i <- x >= lb & x <= ub ;
polygon(c(lb,x[i],ub), c(0,density[i],0), col="red") ;
## calculate beta
beta <- pnorm( ub, tmean, yhatsd) - pnorm(lb, tmean, yhatsd ) ;
result <- paste( "beta =", signif( beta, digits=3 ) ) ;
mtext( result,3 ) ;
```



The figure illustrates sampling distributions of $\bar{y}$ in two distinct scenarios. The left distribution represents the situation where the true mean is 172.5 kg, while the right distribution depicts the scenario with a true mean of 181.6 kg.

The shaded region under the curve in the right-hand graph represents the probability of committing a Type II error (β).

In many cases, accepting H_0 can be quite challenging unless we have an exceptionally large sample size. Frequently, our objective is not to affirm H0 but rather to state that we lack sufficient evidence to reject it.

It's important to note that at the outset of testing, we assume the validity of H_0. Subsequently, based on the available data, we may choose to either reject H_0 or not. Failure to reject H_0 doesn't imply that we are actively accepting it.

The formula for β, considering a null hypothesis of H_0: $\mu = \mu_0$ and an actual population mean of μ_T, is as follows:

- One-tailed test: $\beta = P[\ z < z_\alpha - |\mu_0-\mu_T|/\sigma_{\bar{y}}\ ]$
- Two-tailed test: $\beta = P[\ z < z_{\alpha/2} - |\mu_0-\mu_T|/\sigma_{\bar{y}}\ ]$

Power of test

The power of a statistical test represents the probability of correctly rejecting H_0 (the null hypothesis) when H_0 is indeed false. This probability is denoted as $1-\beta$, where β is the probability of making a Type II error.

The power of an experiment tends to increase with sample size. However, it's worth noting that when the sample size becomes sufficiently large, the additional gains in test power from collecting more observations become marginal. Consequently, in the realm of experimental design, it becomes crucial to pinpoint the smallest sample size that yields a satisfactory level of test power.

Choosing the sample size for testing μ

The amount of information available for a statistical test concerning μ (the population mean) is quantified through the probabilities of Type I and Type II errors, denoted as α and β, respectively.

Let's consider a scenario where we aim to test H_0: $\mu = \mu_0$ against H_A: $\mu > \mu_0$. Additionally, suppose we desire the probability of committing a Type I error (α) and the probability of making a Type II error (β) to be no greater than β when the true value of μ is Δ or more above μ_0.

In such cases, the formula to calculate the necessary sample size to meet these criteria is:

- One-tailed test: $n = \sigma^2(z_\alpha+z_\beta)^2/\Delta^2$

If population variance is unknown, we can substitute an estimated value (s^2) to get an approximate sample size.

For research hypothesis H_A: $\mu \neq \mu_0$, we have $\Delta=|\mu-\mu_0|$, and then

- Two-tailed test: $n = \sigma^2(z_{\alpha/2}+z_\beta)^2/\Delta^2$

Example: Determining Sample Size for Cat Food Packaging The producer of dry (extruded) cat food is concerned that one of his machines may be filling packages with a mean weight exceeding the labeled net weight of 16 kg. This discrepancy could potentially incur significant costs. Previous data indicates that the standard deviation (SD) of package fill weights is approximately 0.225 kg.

To address this concern, the producer needs to calculate the sample size required to establish that μ (the population mean) is greater than 16 kg, with a Type I error rate (α) of 0.05 and a Type II error rate (β) of 0.01 or less, assuming the actual mean is 16.1 kg.

Using the appropriate z-values, $z_{0.05}$ (1.645) and $z_{0.01}$ (2.33), the formula for sample size estimation becomes:

$n = (SD^2 * (z_{0.05} + z_{0.01})^2) / (0.1^2)$

Plugging in the values, we get:

$n \approx (0.225^2 * (1.645 + 2.33)^2) / (0.1^2) \approx 80$

Therefore, the producer needs to collect a random sample of n=80 cartons to conduct this test. If, after obtaining the sample, the computed test statistic does not fall in the rejection region, knowing that β is 0.01 or less when the true μ is 16.1 kg or more, the producer can confidently conclude to accept μ=16 kg.

[Tasks]

Task 1. Revisiting the Dry Cat Food Producer's Dilemma: Now, let's reconsider the issue faced by the dry cat food producer and calculate the necessary sample size for testing H0 against HA: μ≠16, when the actual μ deviates by more than 0.1 kg from the established 16 kg.

Task 2. Assessing the Efficiency of the New Milking Robot: In another scenario, our current knowledge regarding milking duration for a cow indicates that it typically takes about 40 seconds to collect 1 kg of milk from a cow, with a population standard deviation of approximately 5 seconds. A new milking robot has been introduced on a farm with the aim of speeding up the milking process. Now, we need to determine the sample size required to test the hypothesis that milking efficiency has indeed improved. We want to confirm this when the actual mean time required to collect 1 kg of milk using the new robot is 39 seconds or less. The farmer desires a 90% power of the test and sets the significance level α at 0.05.

Inferences about $\mu_1 - \mu_2$

Quite often, we find the need to compare two populations. For instance, we might want to assess the milk yield of two different breeds, compare the back fat thickness of pigs fed with two distinct types of fodder, or evaluate the growth of broilers from two different breeding systems. In each of these scenarios, we typically assume that we are drawing samples from two normal distributions: $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$. The associated sample sizes, means, and standard deviations are denoted as n_1, $\bar{y}_1$, and s_1 for the first distribution, and n_2, $\bar{y}_2$, and s_2 for the second distribution.

The statistical test for comparing μ1 and μ2 using independent samples is as follows:

H_0: $\mu_1 - \mu_2 = D_0$
H_A: One of the following:
(a) $\mu_1 - \mu_2 > D_0$,
(b) $\mu_1 - \mu_2 < D_0$,
(c) $\mu_1 - \mu_2 \neq D_0$

Test statistic: $t = (\bar{y}_1 - \bar{y}_2 - D_0) / [\ s_p\ /\ \sqrt{(1/n_1+1/n_2)}\ ]$

The term s_p is the estimate of the common standard deviation σ:
$s_p = \sqrt{[\ (\ (n_1-1)s^2_1\ +\ (n_2-1)s^2_2\ )\ /\ (n_1+n_2-2)\ ]}$

For a probability α of a Type I error and $df=n_1+n_2-2$,
(a) reject H_0 if $t>t_\alpha$
(b) reject H_0 if $t<-t_\alpha$
(c) reject H_0 if $|t|>t_{\alpha/2}$

Example: An experiment was conducted to compare the mean number of tapeworms in the stomachs of sheep that had been treated for worms against the mean number in those that were untreated:

Drug treated sheep: 18, 43, 28, 50, 16, 32, 13 worms.

Untreated sheep: 40, 54, 26, 63, 21, 37, 39 worms

We may wish to test an hypothesis that there is no difference in the mean number of worms between treated and untreated lambs (D0=0):

H0: $\mu_1 - \mu_2 = 0$ (or simply $\mu_1 = \mu_2$)

HA: $\mu_1 - \mu_2 < 0$ (or simply $\mu_1 < \mu_2$)

The sample data for treated sheep are: $n_1=7$, $\bar{y}_1=28.57$, $s^2_1=198.62$, and the sample

data for untreated sheep are n_2=7, $\bar{y}_2$=40.0, s^2_2=215.33. The estimate of the common population variance is s_p=14.39. The test statistic is t=-1.49. For α=0.05, the critical t-value for a one-tailed test with df=7+7-2=12 is t=1.782. We will reject the null hypothesis if t<-1.782. Since, the observed value t=-1.49 does not fall in the rejection region, we have insufficient evidence to say that the drug really works.

```
                    wor
ms     sample
18    treated
43    treated
28    treated
40  untreated
54  untreated
26  untreated
50    treated
63  untreated
…
```

```
## The computation in R:
sample1 <- c( 18, 43, 28, 50, 16, 32, 13);
sample2 <- c(40, 54, 26, 63, 21, 37, 39);
t.test(sample1, sample2, alt="less", var.equal=TRUE)
```

If the data are organized in a text file in two columns, the t.test function can be called as:

```
read.table( "myfile.txt", header=T ) -> d ;
attach( d ) ;
t.test( worms ~ sample, alt="less", var.equal=TRUE ) ;
```

In many instances, the true variance of the population is unknown, and we must rely on sample variance. It's beneficial if we can assume that both populations have similar variances (var.equal=TRUE). However, when we have knowledge that the two populations possess significantly different variances, we can employ a modified version of the t-test (var.equal=FALSE).

Inferences about $\mu_1 - \mu_2$ from paired data

The t-test discussed in the previous section is suitable for situations where independent random samples are drawn from two populations. However, in many experiments, each measurement in one sample is paired with a measurement in the other sample. For instance, we might record the number of eggs from a hen during two different laying periods. All measurements taken in the first period constitute the first sample, while those recorded in the second period form the second sample. Nonetheless, each observation in the first sample is linked to an observation in the second sample because they are both recorded from the same hen. Thus, the samples are no longer independent. When analyzing paired data, it's necessary to

compute the differences (d) between corresponding values from the two samples. It's important that both sample 1 and sample 2 have the same size, denoted as n. Then, we can analyze the mean difference between the samples.

The statistical test for $\mu_1 - \mu_2$ in a case of paired data is as follows:

H_0: $\mu_d = D_0$
H_A: One of the following:
(a) $\mu_d > D_0$,
(b) $\mu_d < D_0$,
(c) $\mu_d \neq D_0$

Test statistic: $t = (m_d - D_0) / (s_d / \sqrt{n})$

The terms m_d and s_d are the mean and standard deviation of the n differences.

For a probability α of a Type I error and df=n-1,
(a) reject H_0 if $t>t_\alpha$
(b) reject H_0 if $t<-t_\alpha$
(c) reject H_0 if $|t|>t_{\alpha/2}$

Example: Consider an experiment on egg production where each dam was recorded twice in two periods. We are interested whether the two periods are different with respect to the level of egg production:

```
## Dam:              1   2   3   4   5   6   7
## Laying period A: 240, 237, 266, 215, 228, 237, 245
## Laying period B: 250, 240, 267, 210, 224, 243, 265
periodA <- c( 240, 237, 266, 215, 228, 237, 245 ) ;
periodB <- c( 250, 240, 267, 210, 224, 243, 265 ) ;
t.test( periodA, periodB, paired=TRUE )
```

```
Paired t-test
data:  periodA and periodB
t = -1.3534, df = 6, p-value = 0.2247
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.435427   3.578285
sample estimates:
mean of the differences
             -4.428571
```

The test shows that there is no statistical evidence in the data to state that the two periods are different.

Choosing sample sizes for inferences about $\mu_1 - \mu_2$

- Independent samples, $n_1=n_2=n$, $|\mu_1 - \mu_2| > \Delta$

- One-sided test: $n = 2\sigma^2(z_\alpha+z_\beta)^2/\Delta^2$
- Two-sided test: $n = 2\sigma^2(z_{\alpha/2}+z_\beta)^2/\Delta^2$

- Paired data
  - One-sided test: $n = \sigma_d^2(z_\alpha+z_\beta)^2/\Delta^2$
  - Two-sided test: $n = \sigma_d^2(z_{\alpha/2}+z_\beta)^2/\Delta^2$

If variances are unknown, we must substitute estimated values.

[Tasks]

Task 1. Based on a sample of students in the classroom, test the hypothesis that the mean height of males is 15 cm greater than that of females.

Task 2. Determine the sample size for a two-sided test to detect a difference in the mean carcass length between two pig breeds of 0.5 cm or more with a desired power of 99.9%. Utilize the data in the file 'swine-data.txt' to estimate the variance of the trait (column CarcassL).

Task 3. An experiment involving 14 dogs examined the effect of Benzedrine on their heart rate (in beats per minute). After two weeks, each dog received a placebo to serve as its own control. Use this data to test the research hypothesis that Benzedrine increases heart rate.
Dog: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14
Placebo: 250, 271, 243, 252, 266, 272, 293, 296, 301, 298, 310, 286, 306, 309
Benzedrine: 258, 285, 245, 250, 268, 278, 280, 305, 319, 308, 320, 293, 305, 313

Task 4. Utilize the data in the file 'swine-data.txt' to test the research hypothesis that the mean back fat thickness at sacrum point I is different from the mean back fat thickness at sacrum point II (columns BFT3 and BFT4, respectively).

The Wilcoxon's rank sum tests (Mann-Whitney' test)

The t-test is applicable when two compared populations exhibit distributions similar to a normal distribution. However, there are situations where the distribution significantly deviates from a normal distribution. In such cases, we can compare the two populations using the Wilcoxon rank sum test, which demands fewer stringent assumptions. The shape of the distribution of the sample becomes

less crucial since the test operates on ranks rather than the original measurements. For instance, the original data on egg production can be transformed into ranks:

```
For Dam:              1    2    3    4    5    6    7
Laying period A:    240, 237, 266, 215, 228, 237, 245
Laying period B:    250, 240, 267, 210, 224, 243, 265
All ranks:
210(1) 215(2) 224(3) 228(4) 237(5.5) 237(5.5) 240(7.7) 240(7.5) 243(9) 245(10) 250(11) 265(12) 266(13)
267(14)
Ranks for period A:  7.5, 5.5, 13, 2, 4, 5.5, 10
Ranks for period B:   11, 7.5, 14, 1, 3, 9, 12
```

Under the null hypothesis of identical populations, the sum of the ranks for a sample will be directly proportional to the sample size.

There are two variants of Wilcoxon's test: one for independent samples and the other for paired observations, known as the Wilcoxon signed-rank test.

The Wilcoxon test is among the most powerful non-parametric tests for comparing two populations. It is employed to assess the null hypothesis suggesting that two populations share identical distribution functions, as opposed to the alternative hypothesis proposing differences in distribution functions, potentially related to location (such as median).

In numerous applications, the Wilcoxon test serves as a substitute for the two-sample t-test when the assumption of normality is questionable. Additionally, this test can be used when the data in a sample consist of ranks, representing ordinal data rather than direct measurements.

Example: The Wilcoxon rank sum test can be employed to evaluate whether the medians of egg production in two time periods differ.

```
periodA <- c( 240, 237, 266, 215, 228, 237, 245 ) ;
periodB <- c( 250, 240, 267, 210, 224, 243, 265 ) ;
wilcox.test( periodA, periodB, paired = TRUE )
Wilcoxon signed rank test
data:  periodA and periodB
V = 7, p-value = 0.2969
alternative hypothesis: true location shift is not equal to 0
```

The test indicates that there is no statistical evidence in the data to support the claim that the two periods are different.

The t-test is preferred over the Wilcoxon test in situations where we can confidently assume that two populations follow a normal distribution. In such cases, the t-test offers greater statistical power. However, when this assumption is not met, shifting to a non-parametric test is advisable. Alternatively, data transformation can be employed to attain normality.

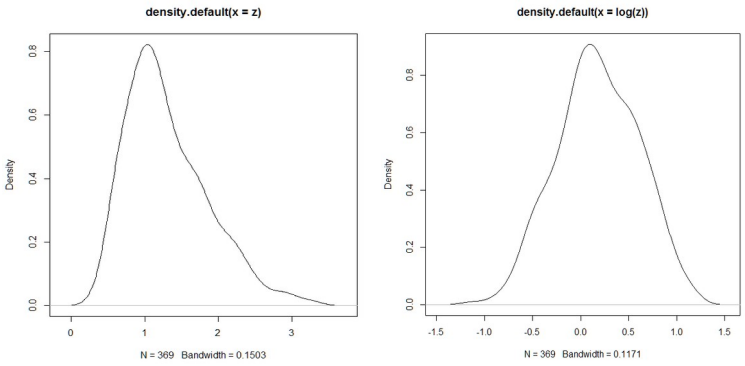
Parametric and non-parametric tests

The t-test is used to make inferences about the difference between means, which represent the parameters of two normal distributions. Thus, the t-test is classified as a parametric test. Non-parametric tests, on the other hand, do not revolve around parameters of theoretical distributions. The Wilcoxon test serves as an example of a non-parametric test, and numerous non-parametric tests rely on ranking data.

Transformation of data

Often, it's possible to attempt data transformation to attain normality or homogenize variance across samples. If data can be successfully transformed to achieve normality, then we can confidently employ statistical tests that necessitate normality. Various transformation techniques are available. For instance, log transformation can be employed to mitigate skewness in data. The figures illustrate the empirical distributions of intramuscular fat percentage (IMF) in pigs, both before and after log transformation. This applied transformation effectively reduced the skewness of the distribution.

```
read.table( "swine-data.txt", header=T ) -> d
attach( d )
IMF[ IMF>0 ] -> z #remove values 0, missing observations
plot( density( z ) )
plot( density( log( z ) ) )
```



[Tasks]

Task 1. Revisit the Benzedrine and dogs' heart rate data. Utilize a non-parametric method to assess the impact of Benzedrine.

Task 2. The SREBF1 gene could be a pivotal gene for fat accumulation. Evaluate its expression (transcript level) in two pig tissues using the data provided in the 'SREBF1.txt' file. Attempt to visualize the distribution shape of the SREBF1 gene's transcript level. Investigate whether data transformation can alleviate the existing skewness.

Task 3. Employ a non-parametric method to examine the distinction in SREBF1 mRNA levels between muscle and fat tissues.

Estimation of population variance σ^2

The sample variance, denoted as $s^2 = \sum(y-\bar{y})^2 / (n-1)$, offers a point estimate for the population variance σ^2 when n measurements are drawn from a normal distribution. Estimating variance accurately is more challenging than estimating population mean because it demands a larger dataset.

A more concise formula for sample variance is $s^2 = [\sum y^2 - (\sum y)^2 / n] / (n-1)$.

Testing for population variance σ^2

There are situations where the population's variability is of greater significance than the mean. For instance, in a slaughterhouse, it's crucial to receive pigs that exhibit similar sizes and proportions of body parts. In pig reproduction, it's vital that the number of piglets born from different dams is reasonably consistent. Populations with low variability among individuals may be favored for technical reasons. Conversely, higher variability among animals is necessary if the goal is to select the best animals for reproduction.

The statistical test for variance relies on the chi-square distribution. You can explore the distribution's shape based on the number of degrees of freedom (DF) using the following R code. Upper-tail and lower-tail values of the chi-square distribution are highlighted in red.

```
DF=9 #check the effect of various degrees of freedom
alfa = 0.05
x <- seq( 0, max(10,3*DF),length=1000 )
density <- dchisq( x, df=DF )
plot(x, density, type="l", xlab="Chi-square", ylab="Density")
# add red area under curve
lb = qchisq( alfa/2, df=DF )
ub = qchisq( 1-alfa/2, df=DF )
i <- x <= lb
polygon(c(0,x[i],lb), c(0,density[i],0), col="red")
i <- x >= ub
polygon(c(ub,x[i], max(x)), c(0,density[i],0), col="red")
```

The test for population variance:

H_0: $\sigma^2 = \sigma^2_0$
H_A: One of the following:
(a) $\sigma^2 > \sigma^2_0$,
(b) $\sigma^2 < \sigma^2_0$,
(c) $\sigma^2 \neq \sigma^2_0$

Test statistic: $\chi^2 = (n - 1)s^2 / \sigma^2_0$

For a specified value of α,
(a) reject H_0 if $\chi^2 > \chi^2_U$, the upper-tail value for α and df=n-1
(b) reject H_0 if $\chi^2 < \chi^2_L$, the lower-tail value for α and df=n-1
(c) reject H_0 if $\chi^2 > \chi^2_U$, the upper-tail value for $\alpha/2$ or and df=n-1
or $\chi^2 < \chi^2_L$, the lower-tail value for $\alpha/2$ and df=n-1.

Note, the test is known to be very sensitive to non-normality.

```
## The test is included in the R package TeachingDemos.
## Install and load the package.
x <- rnorm(20, mean = 15, sd = 7)
sigma.test(x, sigma = 6)

One sample Chi-squared test for variance
data:  x
X-squared = 37.8871, df = 19, p-value = 0.01227
alternative hypothesis: true variance is not equal to 36
95 percent confidence interval:
  41.51722 153.13926
sample estimates: var of x
71.78617
```

Test for comparing two population variances

At times, we need to assess whether two population variances are equal or not. A test for comparing two population variances is frequently employed to verify the assumption of equal variances before conducting a two-sample t-test (used for comparing two means). This variance comparison test relies on the F distribution. The provided R code illustrates the F distribution's shape, contingent on two parameters: two degrees of freedom values, DF1 and DF2.

```
#check the effect of various degrees of freedom
DF1=6; DF2=17; alpha= 0.05;
x <- seq( 0, 10,length=1000 ) ;
density <- df( x, df1=DF1, df2=DF2 )
plot( x, density, type="l", xlab="F", ylab="Density" ) ;
# add red area under curve
ub = qf( 1-alpha, df1=DF1, df2=DF2 ) ;
i <- x >= ub ;
polygon(c(ub,x[i],max(x)), c(0,density[i],0), col="red")
```

The test for population variance:

H_0: $\sigma^2_1 = \sigma^2_2$
H_A: One of the following:
(a) $\sigma^2_1 > \sigma^2_2$,
(b) $\sigma^2_1 \neq \sigma^2_2$,
Test statistic: $F = s^2_1 / s^2_2$

For a specified value of α,
(a) reject H_0 if $F>F_{\alpha,df1,\ df2}$
(b) reject H_0 if $F>F_{\alpha/2,df1,\ df2}$
where $df_1=n_1-1$, $df_2=n_2-1$

This test is highly sensitive to deviations from the normality of the underlying distributions. It's advisable to create separate plots for the data from each sample. If there are indications that one or both of the populations may not follow a normal distribution, exercise caution when drawing inferences, as the p-value may significantly differ from the theoretical value.

```
x <- rnorm(50, mean = 0, sd = 2)
y <- rnorm(30, mean = 1, sd = 1)
## Do x and y have the same variance?
var.test(x, y)
```

```
F test to compare two variances
data:  x and y
F = 5.7347, num df = 49, denom df = 29, p-value = 3.503e-06
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  2.881262 10.789419
sample estimates:
ratio of variances
          5.734732
```

Note, the var.test also estimates confidence interval for population variance.

[Tasks]

Task 1. Variability in milk production for a 305-day lactation period was observed in a random sample of 15 Holstein cows. (a) Use the milk-yield data to estimate a 95% confidence interval for population variance. (b) Determine the confidence interval for population standard deviation. (c) Compare the 95% confidence interval for population SD with the width of the interval for population mean. Which

interval is wider? (d) What sample size would you need to estimate the population mean with a similar interval width as for the SD? Milk-yield data for Holstein cows in kg: 12928, 12120, 14972, 14044, 14788, 13812, 14358, 8998, 10620, 14744, 11036, 9248, 9980, 11990, 14786

Task 2. Revisit the data on Holstein cows' milk-yield and test the research hypothesis that the true population variance of milk yield is less than 5000000.

Task 3. Utilize the data on pig carcass traits from the 'swine-data.txt' file. Test the hypothesis that the population variance of back fat thickness over the shoulder (BFT1) in the Polish Large White (PLW) is lower than in the synthetic line L990.

Categorical data

Many experiments yield data measured on a quantitative scale. However, there are also situations where we are interested in categorical or count data. For example, a pig's carcass may be classified into one of a few classes, and we may want to determine the probabilities associated with these classes.

Answer: yes, no, yes, yes, no, yes (or 1, 0, 1, 1, 0, 1)

Color: green, red, red, blue, green

Probability distributions of discrete random variables

One very useful probability distribution is the binomial distribution, while another useful distribution is the multinomial distribution.

Binomial experiment

We can consider some experiments as binomial experiments. A binomial experiment has the following properties: The experiment consists of n repeated trials. Each trial can result in just two possible outcomes. We refer to one of these outcomes as a success and the other as a failure. The probability of success, denoted by p, remains the same for every trial. The trials are independent; that is, the outcome of one trial does not affect the outcome of other trials.

Example: Here is an example of a binomial experiment. You flip a coin two times and count the number of times the coin lands on heads. This is a binomial experiment because: The experiment consists of repeated trials. We flip a coin 2 times. Each trial can result in just two possible outcomes - heads or tails. The probability of success is constant - 0.5 for every trial. The trials are independent; that is, getting heads on one trial does not affect whether we get heads on other trials.

Binomial distribution

The probability of observing y successes in n trials of a binomial experiment is: $P(y) = n! \times \pi^{y} \times (1-\pi)^{n-y} / [\ y! \times (n-y)!\ ]$

where n = number of trials, π = probability of success in a single trial, y = number of successes in n trials, n!=n(n-1)(n-2)(n-3)…(3)(2)(1).

Example: A drug company states that a new drug can cause undesirable side effects in 10% of horses treated. If a vet has 4 unrelated horses to treat with the new drug, what is the probability that all 4 will experience the side effect?

$P(3) = 4! \times 0.1^3 \times (1-0.1)^{4-3} / [\ 3! \times (4-3)!\ ] = 0.0036$

```
dbinom(3, size=4, prob=0.1)
```

Estimation of the binomial parameter

The estimate of the binomial parameter π is p=y/n. The confidence interval is $p \pm z_{\alpha/2} \sqrt{[p(1-p)/n]}$.

Test for binomial parameter π

The test is based on the standard normal distribution

H_0: $\pi = \pi_0$
H_A: One of the following:
(a) $\pi > \pi_0$,
(b) $\pi < \pi_0$,
(c) $\pi \neq \pi_0$

Test statistic: $z = (p - \pi_0) / \sqrt{[\ \pi_0(1- \pi_0)/n]}$

For a specified value of α,
(a) reject H0 if $z>z_\alpha$
(b) reject H0 if $z<-z_\alpha$
(c) reject H0 if $|z| >z_{\alpha/2}$

Comparing two binomial parameters

The test is based on the standard normal distribution.

H_0: $\pi_1 = \pi_2$
H_A: One of the following:

(a) $\pi_1 > \pi_2$,
(b) $\pi_1 < \pi_2$,
(c) $\pi_1 \neq \pi_2$

Test statistic: $z = (p_1 - p_2) / \sqrt{[p (1- p) / (1/n_1 + 1/n_2)]}$, where $p=(y_1 + y_2)/(n_1 + n_2)$

For a specified value of α,
(a) reject H_0 if $z>z_\alpha$
(b) reject H_0 if $z<-z_\alpha$
(c) reject H_0 if $|z| >z_{\alpha/2}$

[Tasks]

Task 1. A drug company claims that a new drug causes undesirable side effects in only 10% of treated horses. If a veterinarian treats 4 unrelated horses with the new drug, what are the probabilities that 0, 1, 2, 3, and all 4 will experience side effects? (Calculate the probability for each possible outcome.)

Task 2. All 4 horses treated with the new drug experienced side effects. What are your thoughts regarding the drug company's claim?

Task 3. Last year, a group of farmers culled 54 cows out of a total of 20,490 cows. Calculate the confidence interval for the probability of a cow being culled due to fertility problems.

Task 4. It was observed that in a herd of cows, female calves were more frequently born than male calves. Over five years, they produced 525 female calves and 475 male calves. (a) Calculate the interval estimate of π (proportion of female calves). (b) Test the hypothesis that the chances of obtaining male/female calves are not 50%/50%.

Task 5. Calving problems (dystocia) are a major cause of calf deaths. It was observed that in a herd of 145 cows, 23 had some calving problems. After cattlemen were trained on minimizing calving problems in the herd, it was observed that within another group of 79 cows, 10 had calving problems. Was the training successful?

Multinomial experiment

Count data are often understood in the context of the multinomial experiment. The experiment consists of n identical trials. Each trial results in one of k outcomes. The probability that a single trial will result in outcome i is π_i , i=1, 2, 3 …, k, $\sum\pi_i=1$, and remains constant from trial to trial. The trials are independent.

The multinomial distribution

When data comes from a multinomial experiment, the probability of a particular outcome can be modelled using the multinomial distribution
$P(n1, n2, \ldots, n_k) = n! \times \pi_1^{n1} \times \pi_2^{n2} \ldots \pi_k^{nk} / [\, n_1! \times n_2! \times \ldots \times n_k! \,]$
where n!=n(n-1)…1, and 0!=1.

Example 1: Consider a dog breed with only three possible coats: blue (probability 0.5), yellow (0.3) and gold (0.2).

```
##Calculate the probability for observing 43 blue,
##33 yellow and 24 gold individuals
p = c( 0.5, 0.3, 0.2 )
dogs = c( 43, 33, 24 )
dmultinom( x=dogs, prob=p )

##Generate two random samples from the multinomial
##distribution for a hundred dogs
p = c( 0.5, 0.3, 0.2 )
rmultinom( n=2, size=100, prob=p )
```

Example 2: It is known that the probability of obtaining a healthy calf from a mating is 0.83, while the probabilities of obtaining 0 and 2 healthy calves are 0.15 and 0.02, respectively. If a farmer breeds 3 dams from the herd, find the probability of obtaining exactly 3 healthy calves. The possible outcomes for a single mating are as follows:

Outcome 1: Number of progeny = 0, Probability = 0.15
Outcome 2: Number of progeny = 1, Probability = 0.83
Outcome 3: Number of progeny = 2, Probability = 0.02

We can have a situation where each of the 3 dams produces a single healthy calf (three times outcome 2). The probability of this is P(2,1,0) = 0.572. The other possibility is that one dam gives no calf (outcome 1), one dam gives one calf (outcome 2), and one dam gives 2 healthy calves (outcome 3). The probability of this is P(1,1,1) = 0.015. Finally, we need to sum up the probabilities of these two possible situations: 0.572 + 0.015 = 0.587.

```
dmultinom( c(0,3,0), prob=c(0.15,0.83,0.02) ) -> p1
dmultinom( c(1,1,1), prob=c(0.15,0.83,0.02) ) -> p2
p1 + p2
```

Chi-square goodness-of-fit test

The test allows the inference about parameters of multinomial experiment: π_1, π_2, π_3, … The test is as follows:

H_0: Each of the k probabilities is specified
H_A: At least one of the probability differs from the hypothesized value
Test statistic: $\chi^2 = \sum [(n_i - E_i)^2 / E_i]$
Reject H_0 if χ^2 exceeds the tabulated critical value for a specified value of α and df=k-1.

Example: A farmer knows from the literature that the probabilities of obtaining 0, 1, and 2 healthy calves from mating are 0.15, 0.83, and 0.02, respectively. Among the total 100 dams in his herd, 20 dams gave no healthy calves, 75 dams gave birth to a single calf each, and 5 dams gave birth to 2 healthy calves each. Test the hypothesis that the probabilities of obtaining 0, 1, and 2 calves from a mating are different from those known from the literature.

Solution: The expected values under the hypothesized probabilities in a herd of 100 dams are E_1=15 dams, E_2=83 dams, E_3=2 dams, whereas the observed values are n_1=20 dams, n_2=75 dams, n_3=5 dams. The test statistic is $\chi^2 = \sum [(n_i - E_i)^2 / Ei] = (20-15)^2/15 + (75-83)^2/83 + (5-2)^2/2 = 6.94$. For α=0.05 and df=3-1=2, the critical value is 5.99. We reject the null hypothesis and conclude that the true probabilities of obtaining 0, 1, and 2 calves from a mating must be different from those known from the literature.

```
null.probs <- c( 0.15, 0.83, 0.02 )
observed <- c( 20, 75, 5 )
chisq.test( observed, p=null.probs )
```

```
Chi-squared test for given probabilities
data:  observed
X-squared = 6.9378, df = 2, p-value = 0.03115
Warning message:
In chisq.test(observed, p = null.probs) :
  Chi-squared approximation may be incorrect
```

Test of independence

This non-parametric test examines the relationship between two distinct random variables that categorize a population based on two different criteria. A common question addressed by this test is whether there is an association between variables, such as dog breed and blindness. In other words, it helps determine if

there is a contingency or dependence between categorical variables. For instance, it explores whether specific dog breeds are more or less likely to have blind individuals compared to others.

It's crucial to note the significant difference between the chi-square goodness-of-fit test and the chi-square test of independence. The former investigates whether the probability of success (in this case, blindness) is consistent or varies across the breeds under examination. In the null hypothesis, we assume that each breed has the same likelihood of blindness and then assess whether our data align with this assumption. However, in this scenario, breed is not a random variable; instead, specific breeds are included in the study (e.g., 1000 dogs of each breed).

Conversely, in the test of independence, breed is a random variable. We collect information from veterinarians regarding both the blindness status and the breed of each dog. We don't know in advance how many dogs will fall into each breed group.

The test compares the observed values to the values expected under the null hypothesis. For instance, it can be applied to data on the diagnosis of progressive retinal atrophy across different breeds.

```
Breed: cocker spaniels, collies, Irish setters, Norwegian elkhounds, schnauzers, poodles
Blind:  79, 40, 20, 25, 90, 50
Sighted: 1231, 1322, 1420, 2421, 733, 2301
```

From this data, we can calculate the probability (frequency) of encountering a Cocker Spaniel dog. We observed 1310 Cocker Spaniels among a total of 11,863 dogs, resulting in a probability of encountering a Cocker Spaniel of 0.1104. Similarly, we can compute the probability of encountering a blind dog. Among the 11,863 dogs in the sample, there are 304 blind dogs, giving us a probability of blindness of 0.026.

If breed and blindness are independent, the probability of encountering a blind Cocker Spaniel is calculated as the product of these probabilities, resulting in 0.1104 * 0.026 = 0.00287. It's important to note that the multiplication of probabilities to calculate joint probability is only valid under the assumption that blindness and breed are independent (the null hypothesis). In a sample of 11,863 dogs, we would then expect to find approximately 0.00287 * 11,863 = 34 dogs.

However, the observed value is 79, indicating a deviation from the expected value. Expected values under the null hypothesis can be calculated for each individual cell, forming a table with r rows and c columns. In this experiment, there are 12 cells in total: 6 breeds (r) and 2 vision statuses (c). The deviations of each observed value from its corresponding expected value are then accumulated in the chi-square test statistic. Under the null hypothesis of independence, the test statistic follows the chi-square distribution with (r-1) * (c-1) degrees of freedom, which in this case equals df=5.

```
blind <- c( 79, 40, 20, 25, 90, 50 )
sighted <- c( 1231, 1322, 1420, 2421, 733, 2301 )
```

```
tab <- rbind(blind, sighted)
chisq.test( tab )

Pearson's Chi-squared test
data:  tab
X-squared = 260.4469, df = 5, p-value < 2.2e-16
```

[Tasks]

Task 1. For a biallelic locus and allelic frequencies p and q, the Hardy-Weinberg proportions of genotypes are p^2, 2pq and q^2. It is known that the frequency of the mutated allele T within the RYR1 locus is q=0.12. In a sample of 1000 pigs 1000, 20 pigs were TT and 720 were CC. Check whether this population has the Hardy-Weinberg proportions under q=0.12.

Task 2. It is suggested that some variant of the DIO2 gene increases the risk of blood hypertension in horses. A sample of 200 horses was collected. The blood pressure of each horse was measured and the genotype at DIO2 locus was determined. What this data suggests about the possible effect of the DIO2 gene.

Genotype	Very high	Elevated	Normal
AA	30	15	15
AT	40	10	50
TT	10	5	25

Task 3. Farmers around the world have been asked about their opinion on GMO. Is there any dependency between opinion and geography?

Location	Favour	Do not favour	Undecided
Americas	24	27	9
Europe	40	45	15
Asia	16	18	6

Task 4. In some local consumer tests, it was shown that the optimal percentage of intramuscular fat (IMF) is between 1-2%. Analyse the data on IMF from the 'swine-data.txt' file. Classify pigs into two groups according to optimal and nonoptimal IMF. Is there any contingency between this classification and breed? Note, missing observation on IMF is denoted by '0'.