STAT8007

Worksheet 7 effect size, sample size and statistical significance

Here we use the rnorm function in R to create datasets with a normal distribution and specified means and variances. We will then use two sample t-tests to determine whether there are statistically significant differences between the sample means.

The rnorm(n,μ ,σ) function generates a vector of n pseudo-random numbers from a normal distribution of mean μ and standard deviation σ.

The numbers are described as *pseudo* random because computers are deterministic and so cannot create truly random numbers. Any random number generated on a computer is calculated using an algorithm, if the starting conditions of the algorithm are the same, then the same *random* numbers will be generated. The starting conditions of the random number generator in R can be specified using set.seed()

Let’s try:

set.seed(1)

rnorm(10,0 ,1)

-0.6264538 0.1836433 -0.8356286 1.5952808 0.3295078 -0.8204684 0.4874291 0.7383247 0.5757814 -0.3053884

Here we have 10 numbers generated from a normal distribution with a mean of 0 and a standard deviation of 1.

In this worksheet, we will test for significant differences between two samples (sample A and sample B) for 4 different cases:

Case 1:

Sample A has a mean of 0 and a standard deviation of 4

Sample B has a mean of 1 and a standard deviation of 4

Case 2:

Sample A has a mean of 0 and a standard deviation of 4

Sample B has a mean of 0.1 and a standard deviation of 4

Case 3:

Sample A has a mean of 0 and a standard deviation of 1

Sample B has a mean of 0.1 and a standard deviation of 1

Case 4:

Sample A has a mean of 0 and a standard deviation of 10

Sample B has a mean of 0.1 and a standard deviation of 10

**Case 1**

Set the seed at random and generate two vectors of size 10, A\_10 and B\_10.

A\_10<-rnorm(10, 0, 4)

B\_10<-rnorm(10, 1, 4)

We wish to test the hypothesis:

H0 There is no difference between the population means, μA = μB

HA There is a difference between the population means, μA ≠ μB

Clearly we know that the alternative hypothesis is true since we defined the distributions the two samples are taken from, μA = 0and μB = 1. The aim of this exercise is to see how sample size affects the outcome of the hypothesis test. We can use the two sample t-test to test for differences between the sample means.

First calculate the sample means and standard deviations: remember these are random samples taken from the population.

Next look up the t.test function and test the two samples.

Are the sample means significantly different?

Repeat the process again using samples of size 20

A\_20<-rnorm(20, 0, 4)

B\_20<-rnorm(20, 1, 4)

Record your results in the table for Case 1, like this:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample Size | Sample A Mean | Sample B mean | P-Value | H0 or HA at 5% |
| 10 | -1.202080 | 1.431363 | 0.1827 | H0 |
| 20 |  |  |  |  |
| 30 |  |  |  |  |
| 50 |  |  |  |  |
| 100 |  |  |  |  |
| 1000 |  |  |  |  |
| 10,000 |  |  |  |  |

In addition to the table, we can demonstrate the effect of increasing sample size on the accuracy of the samples visually using hist(). Recreate the plots on the following page using your own data.

You can plot histograms in one plot using:

windows(10,20)

par(mfrow = c(5,2))

Look up ?hist to plot density rather than frequency.



**Case 2**

Next we reduce the effect size, so that we are now trying to detect a difference between samples taken from populations with distributions:

A\_10<-rnorm(10,0,4)

B\_10<-rnorm(10,0.1,4)

we have reduced the difference between the means from 1 to 0.1.

Repeat the process above and record your results in a table to see how large a sample size we need to detect the difference between the population means.

**Case 3**

This time let’s reduce the variation associated with the data:

A\_10<-rnorm(10,0,1)

B\_10<-rnorm(10,0.1,1)

We have reduced the standard deviation from 4 to 1.

Repeat the process again and record your results in a table to see how large a sample size we need to detect the difference between the population means now that there is less variation.

**Case 4**

Now, let’s increase the variation associated with the data:

A\_10<-rnorm(10,0,10)

B\_10<-rnorm(10, 0.2,10)

We have increased the standard deviation from 1 to 10.

Repeat the process again and record your results in a table to see how large a sample size we need to detect the difference between the population means now that there is more variation.

Case 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample Size | Sample A Mean | Sample B mean | P-Value | H0 or HA at 5% |
| 10 | 0.1641318 | 0.5688750 | 0.7436 | H0 |
| 20 | 0.0345187 | 1.13464643 | 0.4144 | H0 |
| 30 |  |  |  |  |
| 50 |  |  |  |  |
| 100 |  |  |  |  |
| 1000 |  |  |  |  |
| 10,000 |  |  |  |  |

Case 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample Size | Sample A Mean | Sample B mean | P-Value | H0 or HA at 5% |
| 10 | 0.1641318 | 0.5688750 | 0.7436 | H0 |
| 20 | 0.0345187 | 1.13464643 | 0.4144 | H0 |
| 30 |  |  |  |  |
| 50 |  |  |  |  |
| 100 |  |  |  |  |
| 1000 |  |  |  |  |
| 10,000 |  |  |  |  |

Case 3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample Size | Sample A Mean | Sample B mean | P-Value | H0 or HA at 5% |
| 10 | 0.1641318 | 0.5688750 | 0.7436 | H0 |
| 20 | 0.0345187 | 1.13464643 | 0.4144 | H0 |
| 30 |  |  |  |  |
| 50 |  |  |  |  |
| 100 |  |  |  |  |
| 1000 |  |  |  |  |
| 10,000 |  |  |  |  |

Case 4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample Size | Sample A Mean | Sample B mean | P-Value | H0 or HA at 5% |
| 10 | 0.1641318 | 0.5688750 | 0.7436 | H0 |
| 20 | 0.0345187 | 1.13464643 | 0.4144 | H0 |
| 30 |  |  |  |  |
| 50 |  |  |  |  |
| 100 |  |  |  |  |
| 1000 |  |  |  |  |
| 10,000 |  |  |  |  |