An introduction to statistical testing

Guillaume Filion

January 21, 2013

All the code shown below has to be executed in `R`, which should be installed on your machines. To execute the code, you have to start the application `R` or `R64` and type the commands at the prompt inside the terminal.

1 Testing: a gentle introduction

1.1 The Gaussian distribution

The Gaussian distribution, sometimes referred to as the "bell curve" represents a random distribution. All we care for is that we can produce random samples from this distribution with the function `rnorm()` in `R`.

Exercise 1

Typing `rnorm(1)` in the `R` terminal produces 1 Gaussian random number. Try producing Gaussian samples of sizes 10 and 50.

Exercise 2

Produce several Gaussian samples of size 10. Are the values identical?

Exercise 3

Produce a large Gaussian random sample (between 10,000, and 10,000,000). In `R`, typing `hist(x)` plots a histogram of the sample `x`. Use the function `hist()` to plot the histogram of the sample. You can increase the resolution by increasing the `breaks` parameter like `hist(x, breaks=50)`.

In `R`, if `x` is a sample and `a` is a number, you can produce a new sample where the value in `a` is added to every value in `x` by typing `x + a`.

Exercise 4

Choose a number. Produce a large Gaussian random sample and then add that number to every value of the sample. Does the new sample look Gaussian?

1.2 The first magic number

The "magic number" is a secret number that I am the only one to know. The goal of this section is to gain information about the magic number.

The function `magic1()` generates a Gaussian random sample to which I add the magic number.

```
> # Equivalent to m + rnorm(15) where m is unknown.
> first_magic_sample <- magic1(15)
> first_magic_sample

 [1]  0.2395244  0.5698225  2.3587083  0.8705084  0.9292877  2.5150650
 [7]  1.2609162 -0.4650612  0.1131471  0.3543380  2.0240818  1.1598138
[13]  1.2007715  0.9106827  0.2441589
```

The function `magic1()` does not exist in `R`, I used it to mask the value of the magic number. You can copy the values show above in your `R` session by typing `first_magic_sample <- c(0.240, 0.570, 2.359, ...)`.

Exercise 5

How would you know whether the magic number is 0?

1.3 Testing

If the magic number is 0, the central value of my sample is 0. But with only 15 random numbers there is quite some uncertainty about this central value. Here we want to know if a true Gaussian random sample, *i.e.* one that is produced when the magic number is 0 could potentially produce the random sample shown above.

Exercise 6

In R, if `f(x)` is a function call, you can repeat it 10,000 times by typing `for (i in 1:10000) { f(x) }`. Generate a random sample of size 15 when the magic number is 0 (we will call this a standard Gaussian sample). How would you generate 10,000 standard Gaussian samples of size 15?

To assign a value to a variable in `R` you can use the following syntax: `x <- 2.3`. One of the specificities of `R` is that every variable is actually a *vector*, which means that it is an ordered collection of values. In the above, the variable `x` is a collection of one value, equal to 2.3. **If you have already defined** a variable `x`, you can create or change the value at position 1, 2, 3, *etc.* with the following syntax: `x[1] <- 1.1; x[2] <- 2.2; x[3] <- 1.3` *etc.*

Exercise 7

Generate 10,000 standard Gaussian samples of size 15. Use the function `mean()` to compute the 10,000 means and store them in a variable called `means` (the first value of `means` will be the first mean, the second value will be the second mean *etc*).

Exercise 8

Use the previous exercise to test whether the magic number is 0.

Exercise 9

Summarize this first approach. How sure are you that you gave the right answer? How could you be absolutely sure?

Even though it may not be obvious, in this section we have performed a statistical test in the strict sense. The critical steps of statistical testing are

1. Formulate a null hypothesis, usually denoted H_0.
2. Choose a score, called a *statistic* (no *s* at the end).
3. Generate the disribution of the statistic under H_0. This is called the *null* distribution.
4. See how likely the observed statistic is under H_0 and conclude.

Exercise 10

What were our null hypothesis, our statistic, our null distribution and our conclusion?

Answers and complements

Answer of Exercise 1

For the record, the density of the standard Gaussian distribution is given by the equation

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

The function `rnorm()` in `R` produces a random sample taken from this distribution.

```
> rnorm(10)

 [1] -0.17285279 -2.53433771 -0.09445901 -0.13067621 -0.34517178  0.02523371
 [7] -0.10775485  1.10556689  0.78663222 -0.90489091

> rnorm(50)

 [1]  1.70900675 -1.98691149  1.79645711 -0.44570452 -1.57051376 -2.30149971
 [7]  0.42844057 -1.38776897  0.26287738 -1.14012939 -0.07166047 -1.00201055
[13] -0.93702727 -1.56549507  0.28864150 -2.21452052 -0.03764876  0.19259853
[19] -0.53824154  0.83207881 -0.89963325 -0.72455874  0.53268874 -0.74390461
[25]  0.92397326  0.66496712 -0.26565063  0.06876814  0.06843948  1.07338214
[31]  0.61566000 -0.35916567  0.93633310  0.57798273 -0.42339765  2.05340410
[37]  0.09380175  0.48342642  0.01499521 -1.00481748 -2.12386886 -1.16213725
[43] -0.89316645 -1.78759001 -1.56616714  0.34832769 -1.13651936 -0.29687038
[49]  1.04155924  1.19948572
```

Answer of Exercise 2

No, the values are different.

```
> rnorm(10)

 [1]  1.80205966  0.03367561  0.67445665 -0.59114583 -0.03631942 -0.70623471
 [7] -0.23587878 -0.03445549 -1.36407018 -0.76019406

> rnorm(10)

 [1]  0.8900684 -1.3508408 -0.2012391 -0.3065655 -0.2028977 -0.4177020
 [7] -0.6651209  1.0460433  1.9556649 -0.6988360
```

In case the values need to be identical, we can set or reset the random number generator with the function `set.seed()`. With this you can make a reproducible random example.

```
> set.seed(123)
> rnorm(10)

 [1] -0.56047565 -0.23017749  1.55870831  0.07050839  0.12928774  1.71506499
 [7]  0.46091621 -1.26506123 -0.68685285 -0.44566197

> set.seed(123)
> rnorm(10)
```

```
 [1] -0.56047565 -0.23017749  1.55870831  0.07050839  0.12928774  1.71506499
 [7]  0.46091621 -1.26506123 -0.68685285 -0.44566197
```

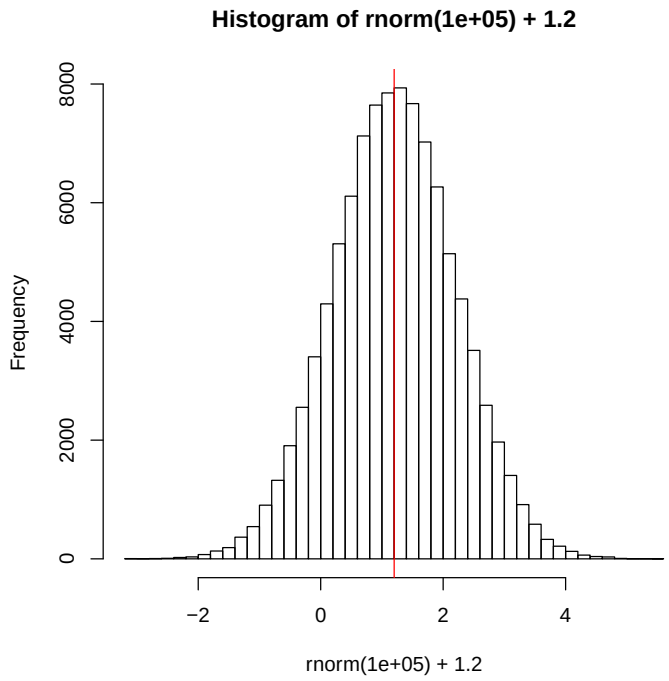
Answer of Exercise 3

```
> hist(rnorm(100000), breaks=50)
```

Answer of Exercise 11

The sample is Gaussian.

```
> hist(rnorm(100000) + 1.2, breaks=50)
> # Add a vertical bar at position 1.2.
> abline(v=1.2, col="red")
```



Answer of Exercise 6

```
> rnorm(15)

 [1]  1.7869131  0.4978505 -1.9666172  0.7013559 -0.4727914 -1.0678237
 [7] -0.2179749 -1.0260044 -0.7288912 -0.6250393 -1.6866933  0.8377870
[13]  0.1533731 -1.1381369  1.2538149

> for (i in 1:10000) { rnorm(15) }
```

Answer of Exercise 7

We use `i`, the index of the 'for' loop to assign a value to the `i`-th element of `means`.

```
> means <- NA
> for (i in 1:10000) { means[i] <- mean(rnorm(15)) }
```

Answer of Exercise 8

We can count how often the central value of a standard Gaussian sample is more distant from 0 than in my sample.

```
> observed_mean <- mean(first_magic_sample)
> observed_mean

[1] 0.9523843

> sum(abs(means) > abs(observed_mean))

[1] 3
```

Answer of Exercise 10

Our null hypothesis was that the first magic number is 0. Our statistic was the mean. Our null distribution was taken from the values of the random sample `means`. Our conclusion was that the magic number is different from 0.

In reality our null hypothesis is more complex. We also implicitly assumed that the sample is Gaussian and that sampling is independent (the generation of one number does not influence the values of the others). In addition, in this oversimplified case, we have assumed that the distribution of the random sample is *standard* Gaussian (the variance is equal to 1). All these elements were used to generate the null distribution.

2 Student's t test: opening the black box

2.1 The Gaussian distribution revisited

In `R` if `x` is a sample (or more accurately a vector) you can multiply every value in `x` by a constant `a` with the command `a * x`.

Exercise 11

Choose a number. Produce a large Gaussian random sample and then multiply that number with every value of the sample. Does the new sample look Gaussian? Try adding a number *and* multiplying by another one.

2.2 The second magic number

The function `magic2()` generates a Gaussian random sample which I multiply by a secret factor, and I then add the second magic number (different from the first magic number). Note that we are not interested in the secret factor, but only in the magic number. Here is a sample generated by this process.

```
> # Equivalent to m + rnorm(15) * a where a and m are unknown.
> second_magic_sample <- magic2(15)
> second_magic_sample

 [1]  2.02142105  1.96430625  0.21414665  2.48272986  0.67412698  0.25318027
 [7] -3.40171291  3.15525467 -0.06084991 -4.49286646  2.32304029 -2.48098098
[13]  1.30041911  0.07348790  2.61465595
```

Exercise 12

Is the distribution of the second magic sample Gaussian?

Exercise 13

Set up a statistical test to know whether the second magic number is 0. To do so, define the 4 critical steps and discuss in group whether they are valid in the present case.

Exercise 14

Is the mean a good choice for a statistic? Justify.

2.3 Standard deviation and effect size

The fact that we don't know the secret factor is more annoying than it seemed. Here we will try to find ways to know it, or to go around the difficulty.

Exercise 15

In `R` the function `sd()` computes the standard deviation of a sample. The standard deviation is the square root of the variance and quantifies the vari-

ability of the values in the sample. Choose a number, produce a large standard Gaussian sample and multiply all the values by that number and compute the standard deviation of the result. What do you observe?

Exercise 16

Could we estimate the secret factor by the standard deviation and use that number to generate the null distribution of the mean? Justify.

Exercise 17

The effect size of a sample is the mean of that sample divided by its standard deviation. Generate 10,000 standard Gaussian samples of size 15 that you multiply by the factor of your choice. Compute their effect sizes and store them in a variable called `esizes`.

Exercise 18

Plot the histogram of the effect sizes you obtained. Produce other samples with different secret factors, and compare the distribution of the effect sizes. What do you observe?

Exercise 19

Is the distribution of the effect size Gaussian?

Exercise 20

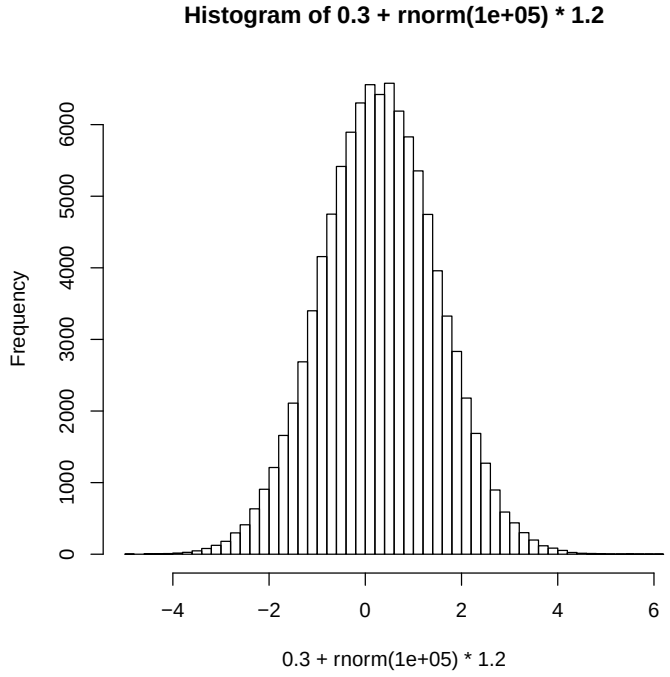
Use the effect size to finish the test and conclude whether the second magic number is 0.

Answers and complements

Answer of Exercise 11

The samples are Gaussian. The Gaussian family is stable by addition and multiplication by a constant. In other words, you always obtain a Gaussian random variable when you add and/or multiply contsants to another Gaussian random variable.

```
> hist(0.3 + rnorm(100000) * 1.2, breaks=50)
```



Answer of Exercise 12

The previous exercise shows that multiplying a Gaussian variable by a constant and adding another constant yields a variable with a Gaussian distribution, so the second magic sample has a Gaussian distribution.

Answer of Exercise 14

The mean is not a very good choice in the present case because we cannot find its null distribution. For this we would need to know the secret factor.

Answer of Exercise 15

For a large standard Gaussian sample, the standard deviation is close to the multiplication factor.

```
> sd(rnorm(100000)*2.54)

[1] 2.540203
```

Answer of Exercise 16

This sounds like a good idea, but in reality it is not. By using the standard deviation, we make an error that can be quite substantial. Here is an example. Assume that the secret factor is 2.54, but on a sample of size 15, the standard deviation can be far off target.

```
> sd(rnorm(15)*2.54)

[1] 1.998879
```

We could produce samples of the form `mean(rnorm(15) * 1.99)`, but this is not the null distribution, *i.e.* it is not the distribution of the mean under the null hypothesis H_0.

Answer of Exercise 17

First let us show what you should not do, the following code is a **mistake**.

```
> esizes <- NA
> for (i in 1:10000) { esizes[i] <- mean(rnorm(15)*2.54) /
+     sd(rnorm(15)*2.54) }
```

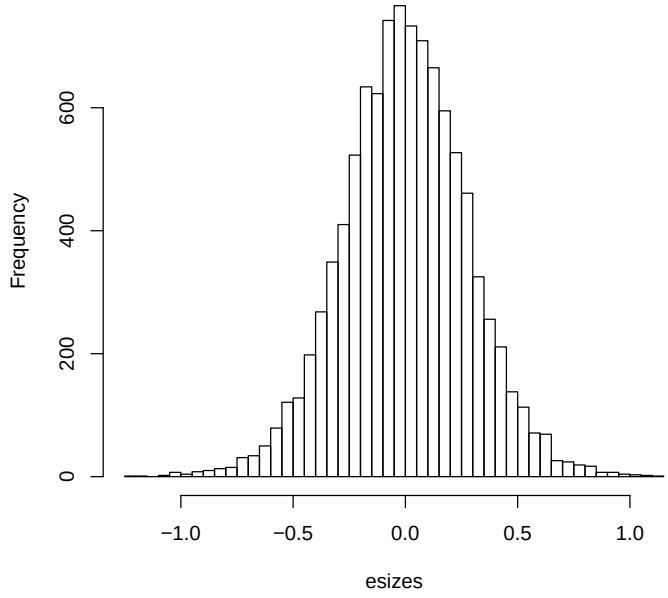
Because the function `rnorm()` is called two times, the samples passed to the functions `mean()` and `sd()` are different. To make sure that you pass the same sample to both functions, the easiest is to assign it to a temporary variable, `smpl` in the case below.

```
> esizes <- NA
> for (i in 1:10000) {
+    smpl <- rnorm(15) * 2.54
+    esizes[i] <- mean(smpl) / sd(smpl)
+ }
```

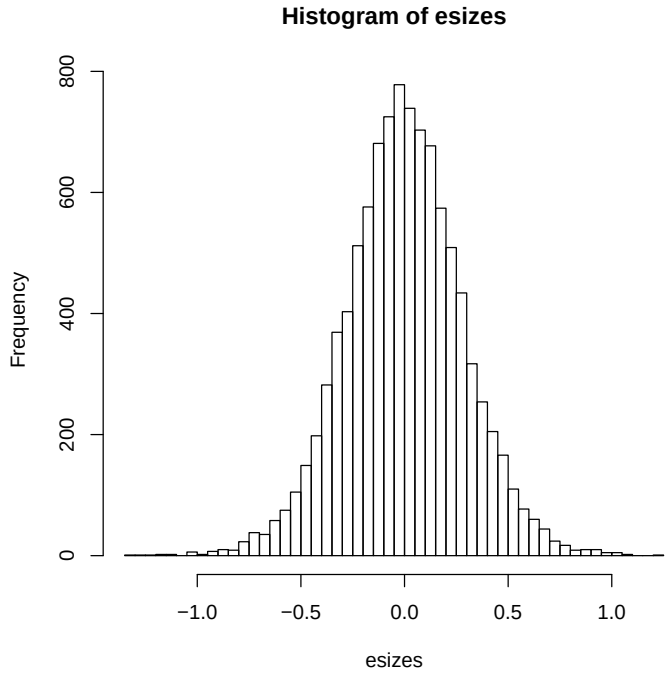
Answer of Exercise 18

The distributions are the same for every factor. This is the most fundamental property of the effect size.

```
> hist(esizes, breaks=50)
```




```
> esizes <- NA
> for (i in 1:10000) {
+    smpl <- rnorm(15) * 15456.97
+    esizes[i] <- mean(smpl) / sd(smpl)
+ }
> hist(esizes, breaks=50)
```



Note that there is some variation between the histograms, but this is negligible in comparison with the 6085-fold difference between the factors used to produce those plots.

Answer of Exercise 19

No. It has a bell shape, but it is distributed as a Student's t. The density of the distribution is

$$\frac{\Gamma(15/2)}{\sqrt{14\pi}\Gamma(7)}\left(1+\frac{x^2}{14}\right)^{-15/2}$$

where Γ is the function

$$\Gamma(\alpha)=\int_0^\infty e^{-t}t^{\alpha-1}dt.$$

By definition a Student's t variable is the ratio of a Gaussian variable and the square root of a χ^2 variable. If a sample is Gaussian, its variance follows a χ^2 distribution so its standard deviation will be the square root of a χ^2 variable.

The effect size is relatively easy to compute, but its distribution is more complicated than the Gaussian. The major difference with the Gaussian is the presence of "outliers" in random samples taken from Student's t distribution. They appear as points far away from the central value 0 in the histograms. In some samples, the standard deviation is very small, so dividing by this value yields a large effect size, which explains the origin of those outliers.

Answer of Exercise 20

```
> observed_esize <- mean(second_magic_sample) / sd(second_magic_sample)
> observed_esize

[1] 0.1931601

> sum(abs(esizes) > abs(observed_esize))

[1] 4699
```

In this section, we have performed a t test. The statistic of the t test is the effect size. Reasons for this have been shown above: because the standard deviation of the distribution is unknown, we cannot get a good null distribution for the mean. We have used a resampling approach to estimate the distribution of the effect size, but it can be computed in closed form. This was done by William Sealy Gosset as he was working for Guinness. He had to publish the result under the pseudonym "Student" to avoid a disclosure issue at the company.

Most of the statistical software have a function to perform Student's t test which uses this closed form distribution (this avoids the need to resample the effect size). In `R` this function is `t.test()`.

```
> t.test(second_magic_sample)

        One Sample t-test

data:  second_magic_sample
t = 0.7481, df = 14, p-value = 0.4668
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.8264836  1.7118648
sample estimates:
mean of x
0.4426906
```

3 Statistical decision theory

3.1 Non Gaussian distributions

The function `magic3()` generates a **non** Gaussian random sample. The third magic number is the expected value of this non Gaussian distribution. Here is a sample generated by this process.

```
> # Unknown process.
> third_magic_sample <- magic3(20)
> third_magic_sample

 [1]  0.14077222  0.68157532  1.29160423  0.57014027  0.62082276 -2.07083757
 [7]  2.64974356 -0.17183795 -0.32029763 -1.31236766 -0.94370525  0.75001101
[13]  0.08840992 -0.28609772  0.49218290  0.31438791 -0.09998969  2.16523635
[19] -0.51638700 -2.07356628
```

Exercise 21

Can we use the previous approach to test whether the third magic number is 0? Can we generate the null distribution of the effect size?

3.2 The Gaussian distribution re-revisited

In `R`, if `x` is a sample and `y` is another sample of the same size, you can do a vector addition (the pairwise sum) of `x` and `y` by typing `x + y`.

Exercise 22

Produce two large standard Gaussian random samples of the same size and compute their vector sum. Does the new sample look Gaussian?

The Gaussian distribution is stable by convolution, which means that when you add two or more Gaussian variables, you have a new Gaussian variable. But Gaussian distributions are more than just that.

Exercise 23

In `R` you can generate uniform random samples with the function `runif()` and exponential random samples with the function `rexp()`, which are used like `rnorm()`. Create a large (between 10,000 and 10,000,000) uniform sample and a large exponential sample and draw their histogram. Do they look Gaussian?

Exercise 24

Generate 10,000 means of a uniform random sample of size 20 and plot the histogram of the values. Do the same with the means of an exponential random sample of size 20. What do you observe?

Exercise 25

Assuming that the sample size is large enough, use the Central Limit Theorem to test whether the third magic number is 0.

3.3 Risk, test level and p-values

Testing is a branch of statistical decision theory, the purpose of which is to develop ways to take the right decision as often as possible. There are two main ways to conclude a statistical test. You can either work at fixed level, usually 0.05 or 0.01, or report the p-value of the test.

The level of a test is a maximum risk you are willing to take. This is the risk of false positive, *i.e.* the probability of rejecting the null hypothesis when it is actually correct.

The p-value of a test is a way to quantify the likelihood of the observed statistic. By definition, it is the probability that the test statistic is higher than the observed value (in absolute value) if the null hypothesis H_0 is true. This is the score we have used so far to decide whether the null hypothesis is true or false.

Exercise 26

If the null hypothesis H_0 is true, what is the probability that the p-value of a test is lower than 0.344?

Exercise 27

A statistician decides to work at level 0.05, what is the highest p-value for which she will accept the null hypothesis?

3.4 Power of a test

Let us come back to the first two magic numbers. Being wrong when you accept the null hypothesis is called a false negative, and being wrong when you reject it is called a false positive.

Exercise 28

What is the probability that we came to a false conclusion when we tested the nullity of the first magic number? What about the second magic number?

Exercise 29

If the first magic number were 0.000000000001, would the null hypothesis be true? What would be our chances of accepting it?

The false negative rate depends on the true value of the magic number. We can compute or estimate this probability and plot it as a function of the true value of the magic number. The result is a **power curve** and tells us the discriminatve power of the test.

Exercise 30

Supposing that we work at level 0.05, compute the power of the test when the first magic number is 0.1.

Exercise 31

Now suppose that we work at level 0.01. Does that change the power of the test if the first magic number is 0.1?

Exercise 32

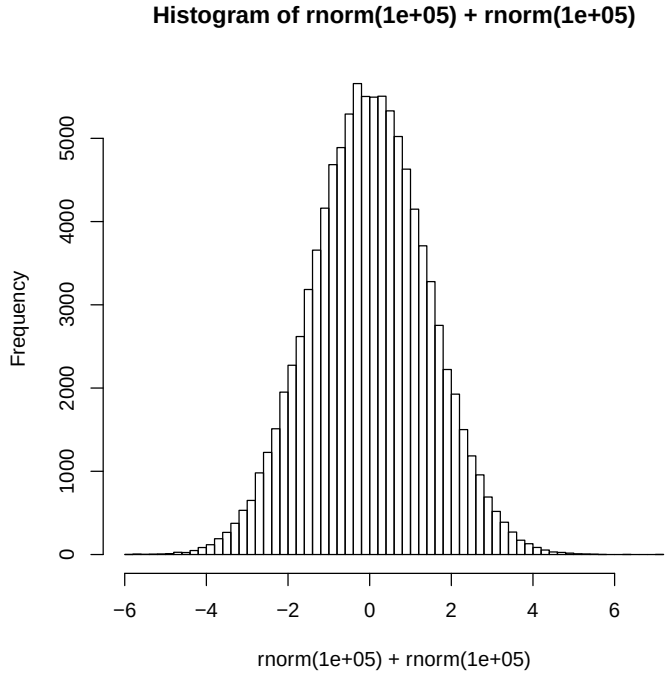
In the last case, decreasing the false positive rate by decreasing the level resulted in lower power (higher false negative rate). Is this always the case? What are the implications for statistical testing?

Answers and complements

Answer of Exercise 22

The sample is Gaussian.

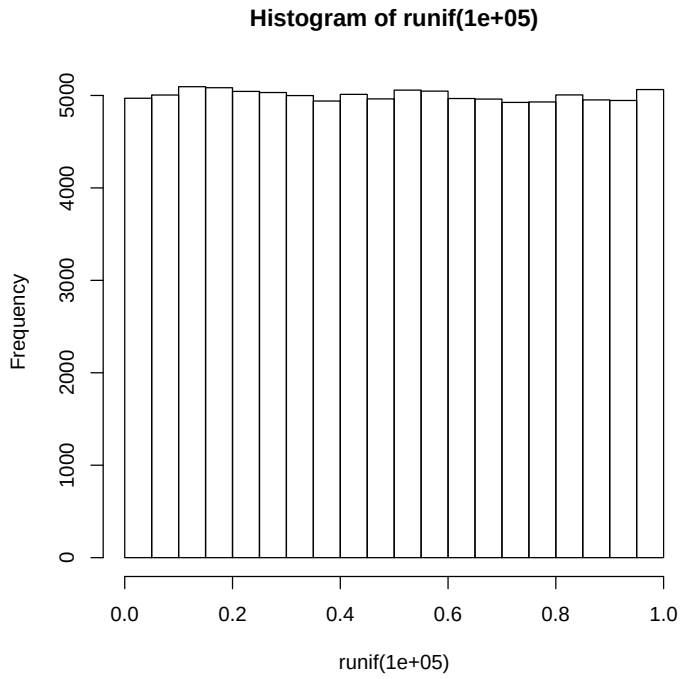
```
> hist(rnorm(100000) + rnorm(100000), breaks=50)
```



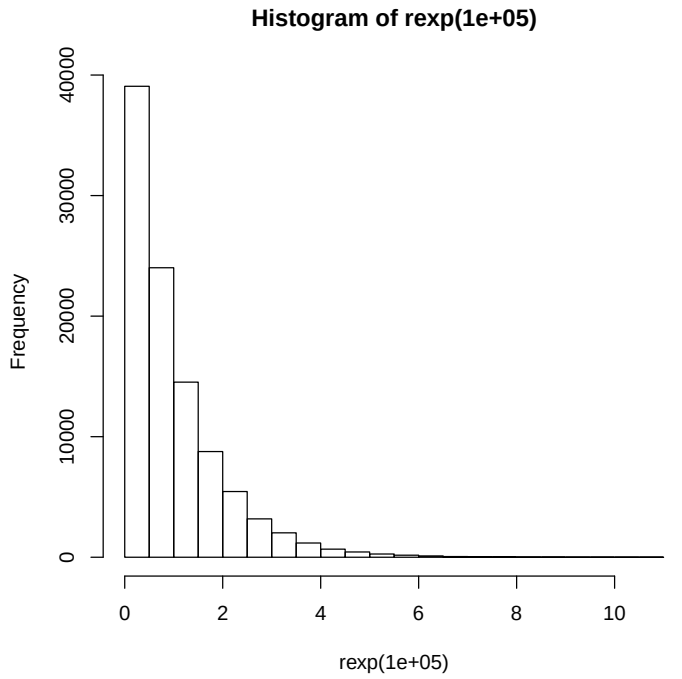
Answer of Exercise 23

The uniform and the exponential distributions do not look Gaussian.

```
> hist(runif(100000))
```

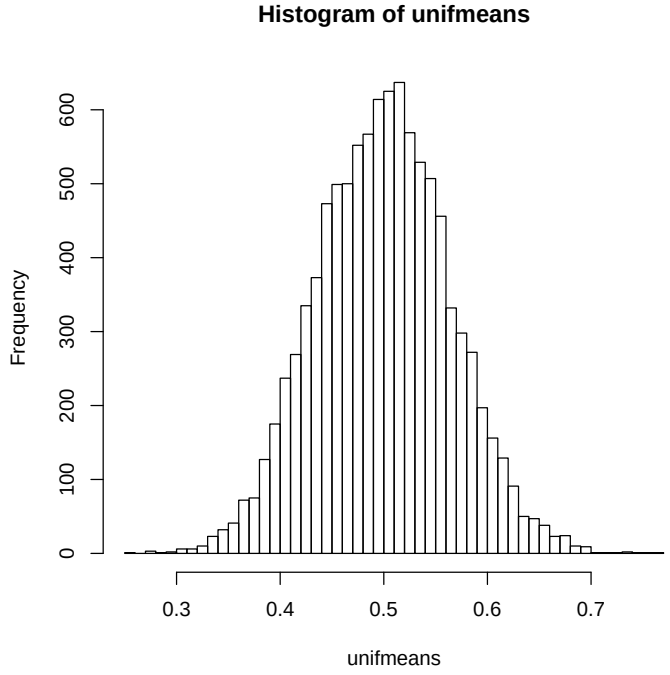



```
> hist(rexp(100000))
```

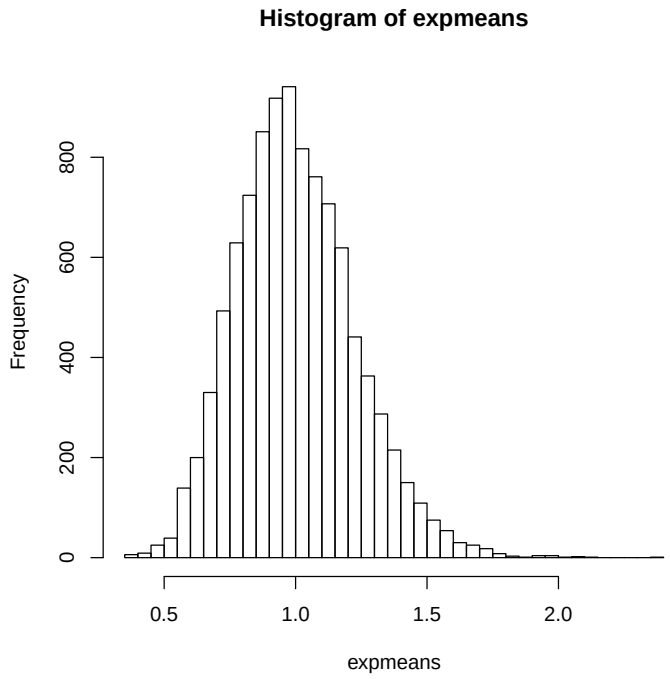


Answer of Exercise 24

```
> unifmeans <- NA
> for (i in 1:10000) { unifmeans[i] <- mean(runif(20)) }
> hist(unifmeans, breaks=50)
```




```
> expmeans <- NA
> for (i in 1:10000) { expmeans[i] <- mean(rexp(20)) }
> hist(expmeans, breaks=50)
```



This exercise illustrates the Central Limit Theorem. In essence, the theorem says that the mean of a sample has an approximate Gaussian distribution if the mean is taken over samples of a large enough size. This theorem has been known for a very long time in statistics. As soon as 1735, Abraham de Moivre gave a version of the theorem for 0-1 variables, and later Aleksandr Lyapunov, Paul Lévy, Jarl Waldemar Lindeberg and William Feller refined the conditions of application.

The example above shows that 15-20 is large enough if the underlying distribution is uniform, but for the exponential distribution, the distribution of the mean is a little bit skewed. The sample size that is large enough so that the approximation holds depends on the problem.

Answer of Exercise 25

Since we assume that the Central Limit Theorem holds, the mean of the third magic sample is supposed to be Gaussian. The standard deviation of the sample is the standard deviation of the mean, up to a factor $\sqrt{n}$, so it can be considered to be the standard deviation of a Gaussian variable. the computed effect size is the ratio of a Gaussian variable and a χ^2 variable, so it is approximately distributed as a Student's t.

```
> esizes <- NA
> for (i in 1:10000) {
+   smpl <- rnorm(20)
+   esizes[i] <- mean(smpl) / sd(smpl)
+ }
> observed_esize <- mean(third_magic_sample) / sd(third_magic_sample)
> sum(abs(esizes) > abs(observed_esize))
```

```
[1] 7236
```

```
> t.test(third_magic_sample)

        One Sample t-test

data:  third_magic_sample
t = 0.3696, df = 19, p-value = 0.7158
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.4592285  0.6562084
sample estimates:
 mean of x
0.09848999
```

Answer of Exercise 26

Let us call T the statistic of the test and $T_{0.344}$ the value of the statistic such that the p-value is 0.344, *i.e.* $T_{0.344}$ is the value such that $P(T \geq T_{0.344}) = 0.344$. The p-value of the test will be lower than 0.344 if and only if the observed test statistic if greater than $T_{0.344}$, which happens with probability 0.344 as we have just seen. So if the null hypothesis H_0 is true, the probability that the p-vlaue is lower than 0.344 is 0.344.

Answer of Exercise 27

The solution of the previous exercise shows that the risk of false positive is less than 0.05 when the p-value is lower than 0.05 and more than 0.05 when the p-value is more than 0.05. So the maximum p-value for which the null hypothesis is accepted is 0.05.

Answer of Exercise 28

We have rejected the null hypothesis for the first magic number. The p-value of the test was 3/1000, so this is our probability of a false positive.

We have accepted the null hypothesis for the second magic number. We have **no idea** of the probability of false negative. This is an important feature of statistical testing.

Answer of Exercise 29

The null hypothesis is that the second magic number is 0, so if it is equal to 0.000000000001 the null hypothesis is false. The noise associated with statistical analyses sometimes makes this confusing but it is important to realize that the null hypothesis is **not** that the second magic number is ≈ 0, but strictly 0.

Yet, the null distributions of the effect size when the second magic number is 0 or 0.000000000001 are practically indistinguishable. Since the probability of accepting the null hypothesis when the second magic number is 0 is 0.95, the probability of accepting it when the second magic number is 0.000000000001 is ≈ 0.95. So in this case we have more chances of being wrong than being right.

Answer of Exercise 30

First we need to find the threshold value of the mean for level 0.05. For that we regenerate the null distribution and take the value of the mean such that 95% of the values are smaller in absolute value (this is known as the quantile of order 0.95).

```
> means <- NA
> for (i in 1:10000) { means[i] <- mean(rnorm(15)) }
> threshold_value <- quantile(abs(means), 0.95)
> threshold_value

      95%
0.5060791
```

Then we need to resample the null distribution in the case the first magic number is 0.1 and count how often it is higher than the threshold value.

```
> means_0.1 <- NA
> for (i in 1:10000) { means_0.1[i] <- mean(rnorm(15)+.1) }
> sum(abs(means_0.1) > threshold_value)

[1] 649
```

Answer of Exercise 31

Yes, the power changes. The new threshold value is the quantile of order 0.99. We need to count how often the null distribution of the mean when the magic number is 0.1 is larger than this new threshold value.

```
> threshold_value <- quantile(abs(means), 0.99)
> sum(abs(means_0.1) > threshold_value)

[1] 170
```

Answer of Exercise 32

This is always the case. There is a trade-off between false positive rate and false negative rate. To decrease the false positive rate, you need to decrease the level, wich means that you have to increase the threshold value of the statistic. As a consequence, the probability that the observed statistic will be higher than the threshold decreases, whether the null hypothesis is true or not.

This is extremely important for statistical testing, and yet not very often emphasized. You often read claims like "we used a very stringent threshold ($\alpha = 10^{-6}$), therefore our conclusions are robust".

Here is a counter-example of the above. Imagine you have the medical files of many patients, but you don't know their sex. You decide that every patient that has breast cancer is a woman. The false positive rate is lower than 0.01 because the incidence of breast cancer in men is about 100 times lower. However, if you used this set as a "trusted set of patients of the female sex" you would most likely conclude that women have a lower life expectancy than the rest of the population. Here the error is not in the females you have included, but in the many females that should have been part of the set, but that you missed by being too "stringent".

4 The Wilcoxon-Mann-Whitney U test

4.1 Two-sample tests

Let us now ask a new question. Here are two samples generated by a sampling process of which nothing is known.

```
> another_magic_sample <- magic3(5)
> yet_another_magic_sample <- magic4(5)
> another_magic_sample
```

```
[1] -0.3065630 -0.9885932  0.1342170  0.3621566 -1.6284401
```

```
> yet_another_magic_sample
```

```
[1]  0.90841437 -0.08769287  1.26350761  1.46762361  0.18828323
```

Even if the distributions of these samples are unknown, I can tell you that the function `magic4()` creates a sample with the same distribution as `magic3()` and then adds a magic number. If the magic number is 0, the distributions are identical, otherwise they are shifted relative to the other.

Exercise 33

What would be the effect size for two samples?

Exercise 34

Does the Central Limit Theorem apply?

Exercise 35

Try to understand what the `rank()` function does in `R`. Try it on different samples of 1-10 values.

Exercise 36

If the null hypothesis is true, what is the expected distribution of the ranks when the samples are merged?

Exercise 37

You can use the function `sample(n)` to produce a random permutation of the first `n` numbers. Choose a statistic and use this function to derive its null distribution. Finish the test and conclude.

5 Multiple testing

Note: You will find additional information on multiple testing on my blog: http://blog.thegrandlocus.com/2012/09/The-most-dangerous-number and http://blog.thegrandlocus.com/2012/09/Focus-on-multiple-testing.

5.1 Hello multiple testing world

Null hypotheses are by definition not interesting, because they represent a knowledge that we already have (the phenomenon behaves as you expect). What is usually interesting for the researchers is the departure from the null, because it suggests the existence of new phenomena.

For this reason, scientific publications tend to report the rejected null hypotheses, and ignore the accepted ones.

Exercise 38

Suppose 7 statisticians test the same null hypothesis. They all work at level 0.05 and the null hypothesis is true. What are the chances that they all come to the right conclusion? What does that mean for scientific publications?

Exercise 39

Make an estimation of the number of statistical tests you perform in a year, and extrapolate to the end of your career. Estimate how many of them will be false positives at level 0.05. Assuming each false positive gives rise to a paper, how many papers should you retract?

Exercise 40

What can you do to fight this trend?

Answers and complements

Answer of Exercise 33

The difference of the sample means divided by their standard deviations. In practice, Student's t test uses another statistic, and the real effect size for two samples is computed somewhat differently.

Answer of Exercise 34

Here the sample sizes are very small. There is some uncertainty about the minimal size for which the theorem applies because this depends on the underlying distribution. As a rule of thumb, a sample size around 30 is usually sufficient for most distributions.

Answer of Exercise 35

It returns the rank of an observation in a sample. As you might expect the rank is an integer number between 1 and n, where n is the total number of values in the sample, which indicates the position of this value when the sample is sorted.

```
> rank(c(1.3,0.9))

[1] 2 1

> rank(c(2.2,0.9,1.7))

[1] 3 1 2
```

Answer of Exercise 36

If the null hypothesis is true, the distributions are identical. In that case, the ranks are a random permutation of $(1, ..., n)$.

Answer of Exercise 37

We can use the difference of mean rank.

```
> diffmeanranks <- NA
> for (i in 1:10000) {
+   smpl <- sample(10)
+   diffmeanranks[i] <- mean(smpl[1:5]) - mean(smpl[6:10])
+ }
> ranks <- rank(c(another_magic_sample, yet_another_magic_sample))
> observed_diffmeanrank <- mean(ranks[1:5]) - mean(ranks[6:10])
> sum(diffmeanranks > observed_diffmeanrank)

[1] 9686
```

Answer of Exercise 38

With the information available, it is impossible to tell. But if we assume that they work with independent datasets and that their work is unrelated, the probability that they all accept the null hypothesis is $0.95^7 = 0.75$.

This means that any hypothesis is more likely to be rejected in the scientific literature if a lot of researchers test it.

Answer of Exercise 39

An average experimenter does about 1 test per week, so about 50 tests per year. A career is (hopefully) 30 or more years, which totals 1500 tests. Assuming that all the null hypotheses are true, 5% of this total will be rejected, coming out to about 75, which is comparable to the number of papers for an experimental career.

Bioinformaticians can easily perform 100 tests per week (BLAST is a statistical test, which is enough to account for that number). The total amount of false positive is about 7,500.

In both cases, you should retract most of your papers.

Answer of Exercise 40

There are many solutions, but two of them are very common. First, you can make sure that your results are reproducible and reproduced, which reduces the false positive rate. If your result is bunk, there is only a 5% chance that it will be reproduced. Using the answer to the previous exercise, it means that if all your 75 papers are reproduced, you should retract only 3-4 of them.

The second solution is to correct for multiple testing. In essence, this decreases the level of the test to compensate for the fact that you perform 50-5000 tests per year and make sure that overall, only 5% of your production is bunk.