Significance and Hypothesis Testing exercises. Part 1

Alex Sanchez-Pla

2024-10-17

Table of contents

1 Problem 1

The following figures (Cushny and Peebles' data), are quoted quote by R.A. Fisher [1] from a "Student's" paper, and show the result of an experiment with ten patients on the effect of two supposedly soporific drugs, A and B, in producing sleep.

The last column gives a controlled comparison of the efficacy of the two drugs as soporifics, (a) Propose and apply a test of significance to help decide if both drugs can be considered to have

the same soporific effect. (b) Answer the previous question assuming that the researchers had decide to record only the sign of the difference, but not the numerical value. (c) Answer the first question assuming that soporific A and B were not tested on the same subjects but, instead on two independent (not matched) groups of subjects.

Patient	A	B	Difference (B - A).
1	+0.7	+1.9	+1.2
2	-0.6	+0.8	+2.4
3	-0.2	+1.1	+1.3
4	-1.2	+0.1	+1.3
5	-0.1	-0.1	0.0
6	+3.4	+4.4	+1.0
7	+3.7	+5.5	+1.8
8	+0.8	+1.6	+0.8
9	0.0	+4.6	+4.6
10	+2.0	+3.4	+1.4
Mean ($\bar{x}$)	+0.75	+2.33	+1.58

Table 1: Additional Hours of Sleep galned by the Use OF TWO TESTED DRUGS

```
drugA <- c(0.7,-0.6,-0.2,-1.2,-0.1,3.4,3.7, 0.8,0  ,2.0)
drugB <- c(1.9,0.8 ,1.1, 0.1 ,-0.1,4.4,5.5, 1.6,4.6,3.4)
BvsA <- drugB-drugA
```

If we compute observed statistics for each test and then the associated p-values we obtain:

1.1 Paired samples with numerical values

Here we can state the null hypothesis as: $H_0 : \mu_D = 0$

We can rely on a student's T statistic, tha, as a corollary of Fisher's Theorem, is known to follow a $t_n - 1$ distribution.

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}, \text{where: } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

If we compute the observed value of the test statistic yields: $\tilde{t}_{obs} = 3.91$.

The p-value is defined as $P[T \geq \tilde{t}_{obs}|H_0] = P[t_{n-1}\tilde{t}_{obs}|H_0]$

This can be computed in r as:

```
pt(3.91,9,lower.tail=FALSE)
```

```
[1] 0.001782377
```

The p-value is very small so it is very unlikely that the observed difference is due to chance which leads us to decide H_0 is not acceptable, that is, *there is a significant difference between the drugs.*

1.2 Paired samples. Only signs

Assume we have paired samples and only the signs of the differences

In that case, although we still aim at determining if the drugs have the same effect we can only work with the "number of positive signs".

The statistic "N = # of positive signs" in a sample of n observations where the probability of obtaining a positive sign is p follows a binomial distribution:

$$N_{n,p} \sim B(n,p)$$

Now the null hypothesis can be stated as: $H_0 : p = 1/2$ which means that under this null hypothesis

$$N_{n.p} \sim \text{Bin}(n = 10, p = 0.5)$$

Now, given that we he observed 9 positive signs, the observed value of the statistic is $\tilde{n}_{\text{obs}} = 9$ and the p-value can be computed as:

$$P\left[N_{n,p} \geq \tilde{n}_{\text{obs}} \mid H_0\right] = P\left[N_{10,0.5} \geq \tilde{n}_{\text{obs}}\right] = p(N = 9) + p(N = 10)$$

This can be computed using R as:

```
p10<- dbinom( 10,10,0.5)
p9<- dbinom (9,10,0.5)
p9+p10
```

```
[1] 0.01074219
```

The p-value is still small (though not so much as in the previous case) so it is unlikely that the observed difference is due to chance which leads us to decide H_0 is not acceptable, that is, there is a significant difference between the drugs.

Notice that in this case, the strength of evidence against H_0 reflected by the p-value is smaller, which is reasonable given that we have less information about how the data deviate from the hypothesis.

1.3 Independent samples

Assume samples are independent and only the signs of the differences

If for whatever reason the drugs had been tested on distinct patients, the question about if the drugs have the same effect may be re-stated, for instance as: $H_0 : \mu_A = \mu_B$.

In this case we may, again, rely on a statistic whose distribution is known as a corollary of Fisher's theorem.

Given two independent simple random samples:

$$X_1, X_2, \ldots, X_{n_1} \overset{iid}{\sim} N(\mu_1, \sigma_1) \quad Y_1, Y_2, \ldots, Y_{n_2} \overset{iid}{\sim} N(\mu_2, \sigma_2)$$

The statistic

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{(n_1 - 1) S_1^2/\sigma_1^2 + (n_2 - 1) S_2^2/\sigma_2^2}} \sqrt{\frac{n_1 + n_2 - 2}{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

is distributed as a Student's t with $n_1 + n_2 - 2$ degrees of freedom. Under the assumption that $\sigma_1^2 = \sigma_2^2$ and $n_1 = n_2, \sigma^2$ cancels out and the statistic can be computed for the sample.

Computing the value of the statistic from the sample yields: $\tilde{t}_{\text{obs}} = 1.80$ and using R the p-value can be computed as:

```
pt (q=1.80, df=18, lower.tail=FALSE)
```

```
[1] 0.04432216
```

Notice that this p-value can lead to the temptation to discuss significance with respect to a threshold, which should be avoided! Instead it is preferable to notice that there is not much evidence leading to accept there is a significant difference between the drugs.

1.4 A Hypothesis testing approach

1.4.1 Paired samples with values

$$H_0 : \mu_D = 0; \quad H_1 : \mu_D > 0$$

Student's T-test is the optimal test for this procedure:

```
t.test(BvsA, alternative = "greater")
```

```

    One Sample t-test

data:  BvsA
t = 3.9128, df = 9, p-value = 0.001775
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 0.7866405       Inf
sample estimates:
mean of x
     1.48
```

1.5 Paired samples with signs

$$H_0 : \#\text{positive signs} = \#\text{negative signs}$$

There is no optimal test for this problem but the signs tests or bintest provides a good approximation.

```
binom.test (9,10)
```

```
    Exact binomial test

data:  9 and 10
number of successes = 9, number of trials = 10, p-value = 0.02148
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5549839 0.9974714
sample estimates:
probability of success
                   0.9
```

1.6 Independent samples

Under the assumption that variances are equal the two-sample t-test provides an optimal solution for this problem

$$H_0 : \mu_A = \mu_B; \quad H_0 : \mu_A < \mu_B;$$

```
t.test(drugB,drugA, alternative = "greater", var.equal=TRUE)
```

```
    Two Sample t-test

data:  drugB and drugA
t = 1.7964, df = 18, p-value = 0.04461
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.05138695        Inf
sample estimates:
mean of x mean of y
     2.33      0.85
```

2 Problem 2

Assume a certain process that leads to a TRUE/FALSE decision is expected to be *fair*, wihch means that either TRUE or FALSE are expected to habppen with probability 0.5

In order to check this *fairness*, from time to time, the process is repeated 100 times and the number of TRUE results is recorded. If this number is above 60 or below 40 the process is declared *out-of-control* or *unfair* and it is re-adjusted.

a) Show how to turn this decision rule into a test with critical region $W = \{\tilde{x}\, s.t. \, |X-50| > 10\}$

b) Calculate the probability of the type I error for the test associated with W.

2.1 Critical region

A subset of the sampling space $W \in \Omega$ is the critical region of a test with null hypothesis H_0 if it is verified that, for any sample $\tilde{x}$

$$P[\tilde{x} \in W | H_0 \text{ true}] \leq \alpha,$$

Where α is called the size of the test. When we work with the equality $P[\tilde{x} \in W | H_0 \text{ true}] = \alpha$, α is said to be the *significance level* of the test and also the *Type I error probability.*

Now, if we consider the random variable X: "number of TRUE results in 100 repeats of the process", it happens that X has a Binomial distribution with parameters $n = 100$ and p unknown, and the 100 repeats of the process is equivalent to sampling one observation of X.

This way to define the process allows us to re-state the null hypothesis ("The process is fair") as

$$H_0 : p = 0.5, \qquad H_1 : \, p \neq 0.5$$

So the form of the critical region is:

$$P[\tilde{x} \in W | H_0 \text{ true}] = P[\tilde{x} \in W | p = 0.5] = P[|X - 50| > 10 | p = 0.5]$$

This can be re-stated by saying that W is *the procedure that rejects* $H_0 : p = 0.5$ in favour of $H_A : p \neq 0.5$ when $|X - 50| > 10$_.

Note btw that X is a binomial distribution where n is big and p not small so the (Laplace-De Moivere) CLTheorem can be used to compute probabilities based on the Normal approximation to a Binomial.

Next section shows how to do it applying a *continuity correction* to compensat from the fact that, even if n is big values of X are discrete, while the distribution used for the approximation is continuous.

2.2 Type I error probability.

Calculate the probability of the type I error for the test associated with W.

$$\begin{aligned}
P(\text{ Type I error }) &= P(|X-50|>10 \mid p=0.5) = \\
&= P(X>60 \text{ o } X<40 \mid p=0.5) \\
&= 1-P(40\leq X\leq 60 \mid p=0.5) = [\text{continuity correction}] = \\
&= 1-P\left(\frac{39.5-100p}{\sqrt{100p(1-p)}}\leq Z\leq \frac{60.5-100p}{\sqrt{100p(1-p)}}\right) \\
&= 1-P\left(\frac{39.5-50}{\sqrt{25}}\leq Z\leq \frac{60.5-50}{\sqrt{25}}\right) \\
&= 1-P(-2.1\leq Z\leq 2.1) \\
&= 1-(\Phi(2.1)-\Phi(-2.1)) = 1-(0.9821-0.0179) = 0.0357
\end{aligned}$$

Using R this can be computed as:

```
n <- 100
p <- .5
Type.I.error.prob <-
  1 - (  pnorm((60.5 - n*p)/sqrt(n*p*(1-p)))
       - pnorm((39.5 - n*p)/sqrt(n*p*(1-p))))
print(Type.I.error.prob)
```

```
[1] 0.03572884
```

In R, we could have directly used the binomial distribution

```
Type.I.error.prob_2 = pbinom(q= 60, size= n, prob = p, lower.tail=TRUE) -
  pbinom(q= 40, size= n, prob = p, lower.tail=FALSE)
```

2.3 Power function

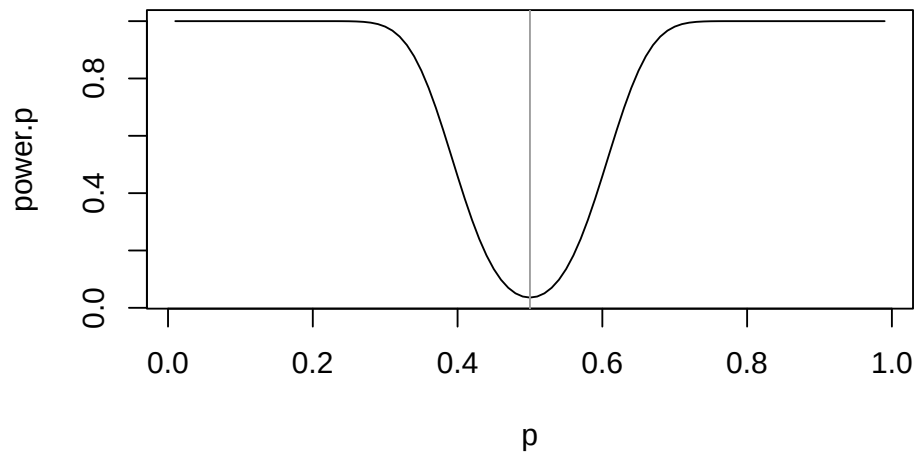
The power function is obtained by expressing the power of the test as a function of the parameter's values.

$$\begin{aligned}
P(\text{¤} \notin W \mid p) &= P(|X-50|\leq 10 \mid p) = \\
&= P(40\leq X\leq 60 \mid p) = [\text{continuity correction}] = \\
&= P\left(\frac{39.5-100p}{\sqrt{100p(1-p)}}\leq Z\leq \frac{60.5-100p}{\sqrt{100p(1-p)}}\right) \\
&= \left(\Phi\left(\frac{60.5-100p}{\sqrt{100p(1-p)}}\right)-\Phi\left(\frac{39.5-100p}{\sqrt{100p(1-p)}}\right)\right)
\end{aligned}$$

p	li	ls	$\Phi(li)$	$\Phi(ls)$	Power
0.20	4.8750	10.125	1.0000	1.0000	1.0000
0.25	3.3486	8.1984	0.9996	1.0000	0.9996
0.30	2.0731	6.6556	0.9809	1.0000	0.9809
0.35	0.9435	5.3463	0.8273	1.0000	0.8273
0.40	-0.1021	4.1845	0.4594	1.0000	0.4594
0.45	-1.1055	3.1156	0.1345	0.9991	0.1354
0.50	-2.1000	2.1000	0.0179	0.9821	0.0357
0.55	-3.1156	1.1055	0.0009	0.8655	0.1354
0.60	-4.1845	0.1021	0.0000	0.5406	0.4594
0.65	-5.3463	-0.9435	0.0000	0.1727	0.8273
0.70	-6.6556	-2.0731	0.0000	0.0191	0.9809
0.75	-8.1984	-3.3486	0.0000	0.0004	0.9996
0.80	-10.125	-4.8750	0.0000	0.0000	1.0000

```r
n <- 100
p <- seq(.01,.99,by=.01)
power.p <-
  1 - (  pnorm((60.5 - n*p)/sqrt(n*p*(1-p)))
       - pnorm((39.5 - n*p)/sqrt(n*p*(1-p))))

plot(p,power.p,type="l")
abline(v=.5,col=8)
```



2.4 The continuity correction

See this intuitive eplanation

3 Problem 3

Let $X_1, \dots, X_n$ be a simple random sample from a Uniform distribution on$(0, \theta)$. We would like to test $H_0 : \theta \geq 2$ versus $H_1 : \theta < 2$. Let $Y_n = max(X_1, \dots, X_n)$ and consider the procedure that has as critical region all the results such that $Y_n \leq 1.5$.

a) Find the power function of such procedure.
b) Calculate the size of the procedure.

3.1 Power function

$$F_{Y_{(n)}}(y) = (F_X(y))^n = \left(\frac{y}{\theta}\right)^n$$

$$\Pi(\theta) = P\left(Y_{(n)} \leq \frac{\frac{3}{2}}{\theta}\right) = \left(\frac{1.5}{\theta}\right)^n$$

if $\theta \leq \frac{3}{2}$ then $Y_{(n)} \leq \frac{3}{2}$ and $P\left(Y_{(n)} \leq \frac{\frac{3}{2}}{\theta}\right) = 1 \rightarrow \Pi(\theta) = 1$

if $\theta > \frac{3}{2}$ then $P\left(Y_{(n)} \leq \frac{\frac{3}{2}}{\theta}\right) = F_{Y_{(n)}}(\frac{3}{2}) = \left(\frac{1.5}{\theta}\right)^n$

3.2 Size of the procedure

$$\text{size } = \alpha = \Pi(2) = \left(\frac{3}{4}\right)^n$$

4 Problem 4

A team of researchers plans a study to see if a certain drug can increase the speed at which mice move through a maze. An average decrease of 2 seconds through the maze would be considered effective, so the researchers would like to have a good chance of detecting a change this large or larger. Would 20 mice be a large enough sample? Assume the standard deviation is $\sigma = 3$sec. and that the researchers will use a significance level of $\alpha = 0.05$.

Let μ denote the true mean decrease in time through the maze.

Then the researchers are testing $H_0 : \mu = 0$ versus $H_A : \mu > 0$.

If the null hypothesis holds, then the sampling distribution of $\bar{X}$ is normal with mean 0 and standard error $3/\sqrt{20} = 0.6708$.

Using a one-sided test at $\alpha = 0.05$, the researchers will reject the null hypothesis if the z-score of the test statistic satisfies $Z \geq 1.645$.

This corresponds to $Z = (\bar{X} - 0)/0.6708 \geq 1.645$ or $\bar{X} \geq 1.1035$.

So, if the true mean decrease in time is 2 seconds , what is the probability of correctly rejecting the null hypothesis of $\mu = 0$?

$$\begin{aligned} 1 - \beta &= P(\text{ Reject } H_0 \mid H_A \text{ true }) \\ &= P(\bar{X} \geq 1.1035 \mid \mu = 2) \\ &= P\left(\frac{\bar{X} - 2}{0.6708} \geq \frac{1.1035 - 2}{0.6708}\right) \\ &= P(Z \geq -1.3365) \\ &= 0.9093 \end{aligned}$$

Thus, the researchers have a 91% chance of correctly concluding the drug is effective if the true average decrease in time is 2 s .

```
pow<- power.t.test ( n=20,
    delta=2,
    sd=3,
    sig.level=0.05,
    power=NULL,
    type="one.sample",
    alternative="one.sided")
show(pow)
```

```
     One-sample t test power calculation

              n = 20
          delta = 2
             sd = 3
      sig.level = 0.05
          power = 0.8902459
    alternative = one.sided
```

Notice that there is a small difference between the analytical solution and the one using the powerfunction because, in the first case we are working with the unrealistic assumption that σ is known. The second case (using R) assumes it is unknown and estimated.

5 Problem 5

Suppose the researchers in the previous example want a 95% chance of rejecting $H_0 : \mu = 0$ at $\alpha = 0.01$ if the true change is a 1.5 sec. decrease in time. What is the smallest number of mice that should be included in the study?

On the standard normal curve, $q = 2.3264$ is the cutoff value for the upper 0.01 tail (i.e. the 0.99 quantile). Thus, we need $(\bar{X} - 0)/(3/\sqrt{n}) \geq 2.3264$, or $\bar{X} \geq 6.9792/\sqrt{n}$.

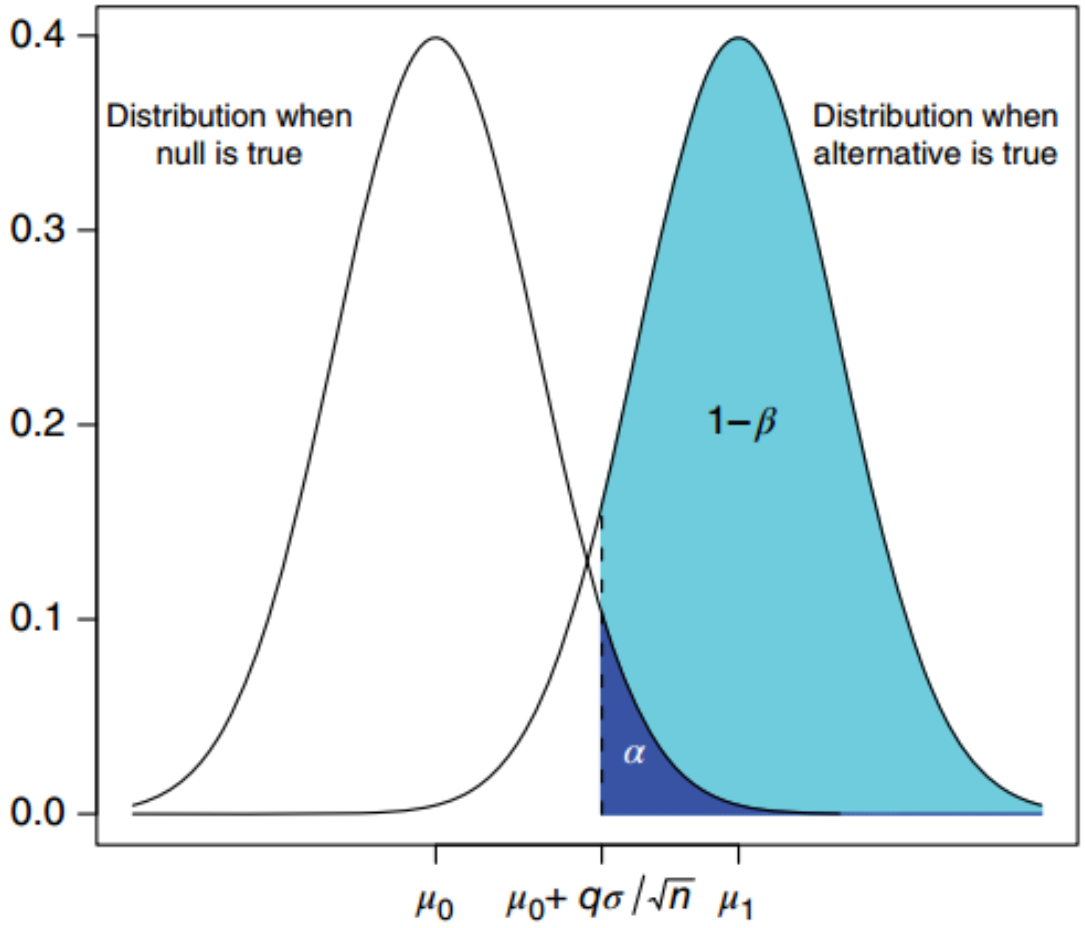


Figure 1: Distributions under the null and alternative hypotheses. Shaded regions represent power (1-beta) and significance level (alfa). Moving the critical value to the left increases power

See This link for an animation related to this plot

Thus,

$$\begin{aligned} 0.95 &= P\left(\bar{X} \geq \frac{6.9792}{\sqrt{n}} \middle| \mu = 1.5\right) \\ &= P\left(\frac{\bar{X} - 1.5}{3/\sqrt{n}} \geq \frac{6.9792/\sqrt{n} - 1.5}{3/\sqrt{n}}\right) \\ &= P\left(Z \geq 2.3264 - \frac{1.5}{3/\sqrt{n}}\right). \end{aligned}$$

Using the 0.05 quantile for the standard normal,

$$-1.645 = 2.3264 - \frac{1.5}{3/\sqrt{n}}$$

Thus, $n = 64$ is the smallest number of mice that the researchers should use. 1 Using R

```
power.t.test(n=NULL,
    delta=1.5,
    sd=3,
    sig.level=0.01,
    power=0.95,
    type="one.sample",
    alternative="one.sided")
```

```
     One-sample t test power calculation

              n = 65.82776
          delta = 1.5
             sd = 3
      sig.level = 0.01
          power = 0.95
    alternative = one.sided
```

As in the previous exercise there is adifference in results due to the unrealistic assumption that σ is known in the first case.

6 Problem 9 - Permutation tests with R

We wish to compare the mean survival time (in weeks) of two groups of mice which are used used to test a treatment for a liver disease. All mice are affected by the disease and half one group has been treated by a placebo (y) while the other has been given the drug being tested (z). Placebo and treatmen have been assigned at random to an otherwise homogeneous sample of mice. The resulting survival times are:

$$\begin{aligned} z = & \quad 94, 197, 16, 38, 99, 141, 23 \\ y = & \quad 52, 104, 146, 10, 51, 30, 40, 27, 46. \end{aligned}$$

a) Assuming the data are exponentially distributed build a likelihood ratio test and use it to test the null hypothesis of mean equality versus the alternative hypothesios that the mean survival times are different.

b) Implement a permutation test in R to compare the two group means. Use it with the data and a number of 1000 permutations to obtaina permutation p-value.

c) Compare the results of both tests and comment about the pros and cons of each method.

6.1 Permutation test

```
z <- c(94, 197, 16, 38, 99, 141, 23)
y <- c(52, 104, 146, 10, 51, 30, 40, 27, 46)
```

We start by computing the *oberved difference in means*, that is, we calculate the mean and measure the difference

```
mean(z) ; length(z)
```

```
[1] 86.85714
```

```
[1] 7
```

```
mean(y) ; length(y)
```

```
[1] 56.22222
```

```
[1] 9
```

```
(diffMeans0 <- mean(z) - mean(y))
```

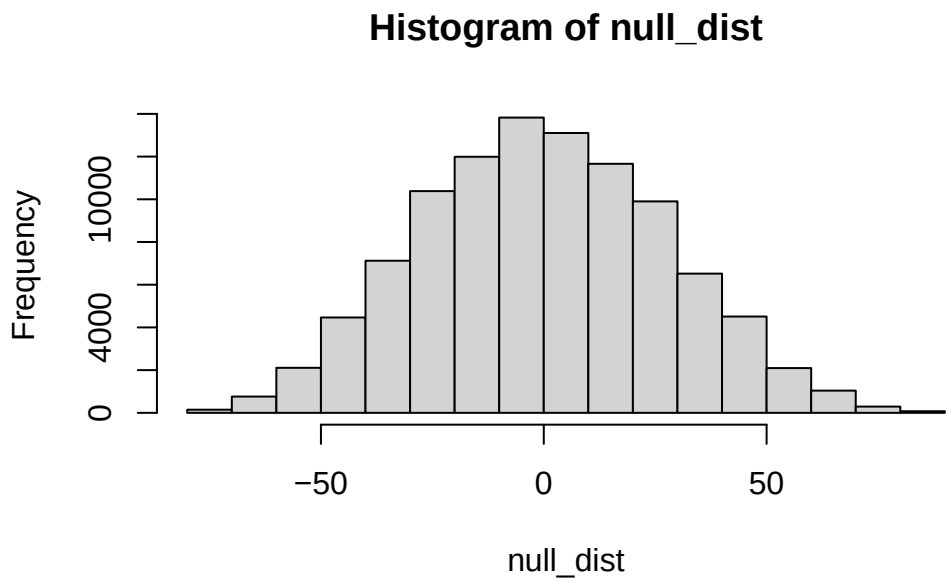
```
[1] 30.63492
```

Next we perform a two-sided permutation test using the following steps:

1. Let us combine the two datasets into a single dataset.
2. Randomly assign each data point into either z or y, although we need to maintain the original sample size (n=7) for Z and (n=9) for y.
3. After randomization, calculate the relevant statistic by taking the difference between mean(Zi) and mean(Yi).
4. Repeat the steps above until we have 10000 statistics.

```
combined_data <- c(z, y) # combines the data
set.seed(123) # set seed for reproducibility
null_dist <- c() # declaring a vector to contain the null distributions

# performs randomization at least 100000
for (i in 1:100000) {
  shuffled_data <- sample(combined_data)  # randomly shuffles the data
  shuffled_z <- shuffled_data[1:7] # assigns the first seven points to Z
  shuffled_y <- shuffled_data[8:16]  # assigns the last nine points to y
  null_dist[i] <- mean(shuffled_z) - mean(shuffled_y)
}
hist(null_dist)
```



5. Add the numbers of statistics that are equal to or greater the previously computed difference in means: `diffMeans0`.
6. Calculate the p-value of the permutation test by dividing the sum from step 5 by 10000 (the number of randomization performed).

```
(p_value <- (sum(null_dist >= diffMeans0) + sum(null_dist <= -diffMeans0))/length(null_dist))
```

```
[1] 0.28004
```