Notes on Hypothesis Testing

Dr. Syring

February 7, 2020

Coverage

Here we will discuss testing from the point of view of chapter 4 in the Hogg textbook. Further study of tests using likelihood ratios will be covered later.

What is hypothesis testing?

Hypothesis testing can be thought of as a part of the scientific method. Roughly, scientists make hypotheses about how the universe works. Then, they design experiments that test those hypotheses, collect data, and try to gauge to what extent the observations tend to support or refute the original hypothesis.

Statistical hypothesis testing

Statisticians, with the help of the practitioners/subject matter experts, seek to translate those scientific hypotheses into statistical hypotheses in which the parameter of a probability model of a population is hypothesized to take its value in a certain set.

For example, the discovery of the Higgs Boson hinged, in part, on a statistical hypothesis test that could be viewed (in an oversimplification) as testing whether or not the mean of a Poisson distribution of particle counts was more or less than a given known value. We write this as

$$H_0 : \mu \leq \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0$$

where μ is the true unknown mean, μ_0 is a known value, and H_0 and H_1 are called the "null" and "alternative" hypotheses. You may think of the null hypothesis as the status quo, the current accepted theory, and the alternative as the negation (or complement) of the null.

Outcomes of hypothesis tests

In a hypothesis testing problem we make a decision to either "reject" or "retain" the null hypothesis. These words are actually rather important and we'll describe why below.

Because we do not actually know whether the null hypothesis is true or false, four outcomes are possible:
1. We reject the null when it is true. This is referred to as a Type 1 error.
2. We reject the null when it is false.
3. We do not reject the null (we retain it) when it is true.
4. We do not reject the null (we retain it) when it is false. A Type 2 error.

The data we collect is random and constitutes an incomplete picture of the entire population. In essentially all interesting problems we will never observe the whole population, so we can never know the value of the population parameter about which we are testing. Therefore, we will never know precisely which of the four outcomes happens; we can control the decision but not whether or not we make the correct decision. So, what we do in practice is try to understand the probability that we make each type of error and design testing procedures so as to avoid making the errors we deem most undesirable.

Illustration of a basic test

Consider a normal population with known vairance and unknown mean. We hypothesize

$$H_0 : \mu \leq 0 \quad \text{versus} \quad H_1 : \mu > 0.$$

Suppose we can collect data $X_1, \ldots, X_n$, a random sample from this population. How should we decide whether to reject or retain H_0 ?

We already know $\bar{X}$, the sample mean, is a reasonable point estimate of μ . Naturally, we would consider rejecting H_0 if the data seemed to suggest $\mu > 0$, and this happens when $\bar{X}$ is something larger than zero. So, let's say we reject H_0 when $\bar{X} > C$ for some $C > 0$ we have not yet specified. We want to avoid errors, so let's consider computing the chance of making an error. The probability of rejecting H_0 corresponds to

$$P(\bar{X} > C) = P(Z > \frac{C - \mu}{\sigma/\sqrt{n}}).$$

We do not know σ so we cannot actually compute this probability. However, it is clear that when H_0 is true

$$P(Z > \frac{C - \mu}{\sigma/\sqrt{n}}) \leq P(Z > \frac{C - \mu_0}{\sigma/\sqrt{n}})$$

since $\mu \leq \mu_0 = 0$ under H_0 . Therefore, if $C = \mu_0 + z_{1-\alpha}\sigma/\sqrt{n}$ where $P(Z < z_\alpha) = \alpha$ then the Type I error is bounded above by α .

Rather than the Type 2 error, we often analyze the Power, defined as one minus the probability of a type 2 error, or, in other words, the power is the probability of rejecting H_0 when it is false. Then, this test that power function

$$\gamma(\mu) = P(\bar{X} > \mu_0 + z_{1-\alpha}\sigma/\sqrt{n}).$$

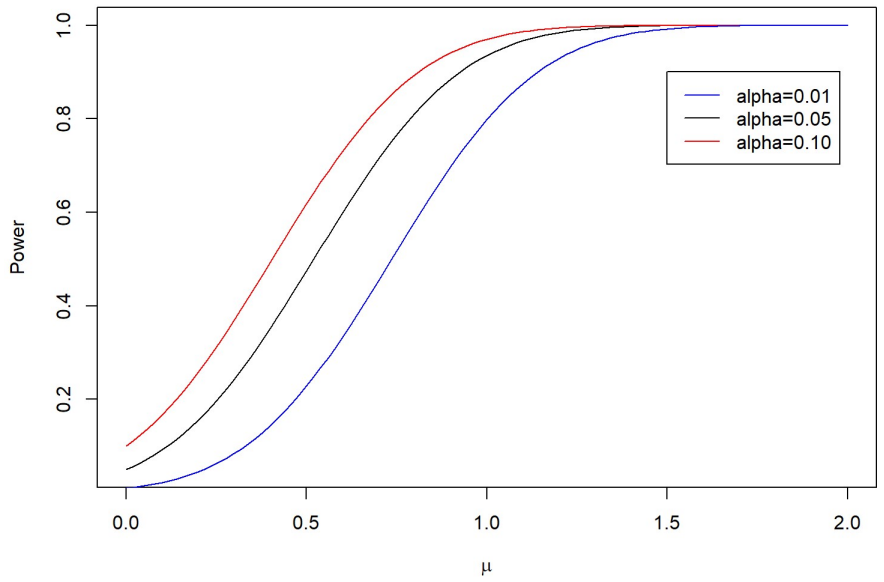
When H_0 is false, $\mu > \mu_0$ so the power function can be written

$$\gamma(\mu) = P(Z > z_{1-\alpha} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}) \geq \alpha$$

which increases in μ .

Power curves for the one-sample test of a normal mean, variance known

```r
alpha<-0.05
mu0<-0
sigma<-1
n<-10
pwr.mytest <- function(mu) 1-pnorm(qnorm(1-alpha)+((mu0-mu)/(sigma/sqrt(n))))
curve(pwr.mytest, from = 0, to = 2, xlab = expression(mu), ylab = "Power")
alpha<-0.10
pwr.mytest <- function(mu) 1-pnorm(qnorm(1-alpha)+((mu0-mu)/(sigma/sqrt(n))))
curve(pwr.mytest, from = 0, to = 2, xlab = expression(mu), ylab = "Power", col = 'red', add=TRUE)
alpha<-0.01
pwr.mytest <- function(mu) 1-pnorm(qnorm(1-alpha)+((mu0-mu)/(sigma/sqrt(n))))
curve(pwr.mytest, from = 0, to = 2, xlab = expression(mu), ylab = "Power", col = 'blue', add=TRUE)
legend(1.5, .9, legend=c("alpha=0.01", "alpha=0.05", "alpha=0.10"),col=c("blue","black", "red"), lty=1)
```

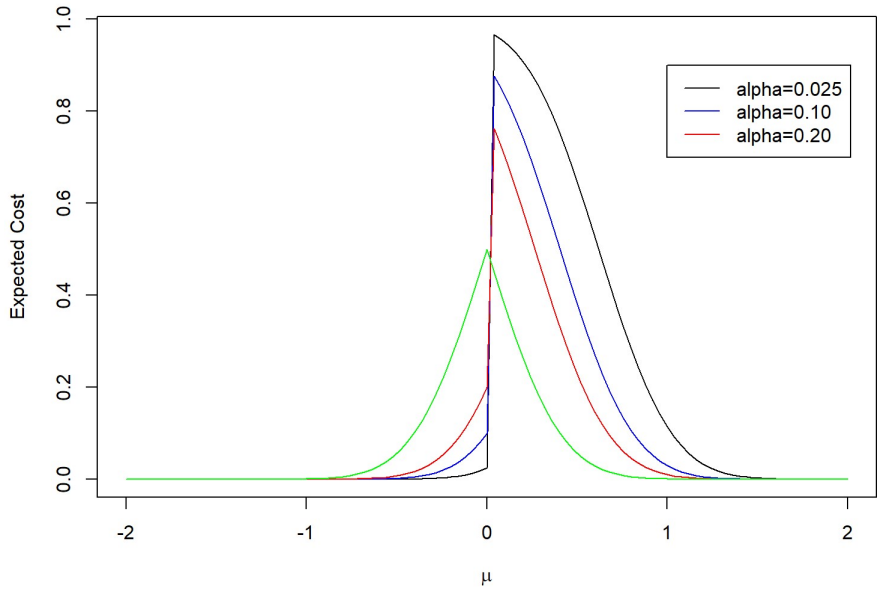


Cost and Expected Cost

A Z-test for a normal mean.

In this case we look at the curves of Type 1 and 2 error probabilities for given tests (given α values). Note that a given test criteria implies a relative cost difference between type 1 and 2 errors. The relative costs are equal precisely when $\alpha = 0.5$.

```
alpha<-0.025  # change to customize
mu0<-0
sigma<-1
n<-10
C <- qnorm(1-alpha)
pwr.mytest <- function(mu) 1-pnorm(C+((mu0-mu)/(sigma/sqrt(n))))
expected.cost <- function(mu) {
  p.reject <- pwr.mytest(mu)
  p.retain <- 1-p.reject
  cost.type1 <- 1  # change to customize
  cost.type2 <- 1  # max(0,mu-mu0-0.1)  # change to customize
  exp.cost <- ifelse(mu<=mu0, p.reject*cost.type1, p.retain*cost.type2)
  return(exp.cost)
}
apply.exp.cost <- function(mu.vec) apply(matrix(mu.vec,length(mu.vec),1), 1, expected.cost)
curve(apply.exp.cost, from = -2, to = 2, xlab = expression(mu), ylab = 'Expected Cost')
alpha<-.10
C <- qnorm(1-alpha)
curve(apply.exp.cost, add=TRUE, col = 'blue')
alpha<-.20
C <- qnorm(1-alpha)
curve(apply.exp.cost, add=TRUE, col = 'red')
alpha<-.50
C <- qnorm(1-alpha)
curve(apply.exp.cost, add=TRUE, col = 'green')
legend(1.0, .9, legend=c("alpha=0.025", "alpha=0.10", "alpha=0.20"),col=c("black","blue", "red"), lty=1)
```


Expected Cost
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
μ
alpha=0.025
alpha=0.10
alpha=0.20

Choice of test criteria and error rates

It is clear that the testing rule determines the probabilities of Type 1 and 2 errors. It can also be seen that a more conservative choice of test that reduces type 1 error probability will increase type 2 error probability. So, how should the test be chosen?

It is partially a convention that small α values are targeted for tests, which implies a low type 1 error probability, and a high cost of type 1 error relative to type 2 error. This sort of makes sense practically. If we think of the null hypothesis as the best current theory of the universe (or whatever we are studying) then we should not want to reject this theory unless we have strong evidence against it. However, in many situations the null hypothesis is "weak" in the sense that we actually have very little expectation of its being true. So, in practice the statistician and practicioners/scientists should discuss the construction of the test to reflect the actual relative costs for the specific situation. This is often not easy to do.

Normal mean, unknown variance

In this case we know

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t(n-1)$$

so that, similarly to the above case with known variance, the test that rejects $H_0 : \mu \leq \mu_0$ in favor of $H_1 : \mu > \mu_0$ when $\bar{X} > \mu_0 + t_{1-\alpha}(n-1)s/\sqrt{n}$ has Type I error probability no more than α .

Power function:
The power can be written

$$P(barX > C) = P(\bar{X} > \mu_0 + t_{1-\alpha}(n-1))$$

$$= P\left(\frac{\bar{X}-\mu}{s/\sqrt{n}} + \frac{\mu-\mu_0}{s/\sqrt{n}} > t_{1-\alpha}(n-1)\right).$$

A random variable

$$T = \frac{Z+a}{\sqrt{V/k}}$$

that is a ratio of a standard normal r.v. plus a constant and the root of a chi-squared r.v. divided by its df (numerator and denominator independent) has a non-central Student t distribution with k df.
The power of a t-test can be written using a non-central t distribution since

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} + \frac{\mu-\mu_0}{s/\sqrt{n}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} + \frac{\mu-\mu_0}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{(n-1)\sigma^2}}}.$$

In R, pwr.t.test

```
library(pwr)
```

```
## Warning: package 'pwr' was built under R version 3.4.4
```

```
d = 1 #(mu-mu0)/sigma, called the "effect size"
pwr.t.test(n=10, d = d, sig.level = 0.05, power = NULL, type = "one.sample", alternative = "greater")
```

```
##
##      One-sample t test power calculation
##
##               n = 10
##               d = 1
##       sig.level = 0.05
##           power = 0.897517
##     alternative = greater
```

Two-sided tests

A two-sided or "point null" test has hypotheses like

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0.$$

For the normal population problem with either known or unknown variance we will reject H_0 whenever $\bar{X} > u$ or $\bar{X} < l$ for some $l < \mu_0 < u$.

Make the choices $l = \mu_0 + z_{\alpha/2}\sigma/\sqrt{n}$ and $u = \mu_0 + z_{1-\alpha/2}\sigma/\sqrt{n}$ and find that

$$P(\text{type 1 error}) = P(\bar{X} < \mu_0 + z_{\alpha/2}\sigma/\sqrt{n} | \mu = \mu_0) + P(\bar{X} > \mu_0 + z_{1-\alpha/2}\sigma/\sqrt{n} | \mu = \mu$$

Note, there is no upper bounding necessary, the type I error probability is exactly α .

Power computations are slightly more complicated due to the fact that there are two criteria for rejection:

$$\gamma(\mu) = P(\bar{X} < \mu_0 + z_{\alpha/2}\sigma/\sqrt{n} | \mu \neq \mu_0) + P(\bar{X} > \mu_0 + z_{1-\alpha/2}\sigma/\sqrt{n} | \mu \neq \mu_0)$$

but if $\mu >> \mu_0$ then one would expect the first statment on the RHS to be nearly zero, and vice versa.

P-values

The way we have been describing the criterion used for deciding whther to reject or retain H_0 is often called the "critical-value" or "test-statistic value" method. We use the desired α level to determine a cutoff value involving a quantile of the distribution of a statistic.

The p-value method is essentially the reverse. We compute the Type I error probability implied by the value of the test statistic and compare this value (called the p-value) to the previously chosen α .

For example, consider a one-sample test for a normal mean with known variance $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$. Suppose $\alpha = 0.05$, $\sigma = 2$, $n = 10$, and suppose that we observe $\bar{x} = 1$. Then, the test statistic is $z = \frac{\bar{x}-0}{2/\sqrt{10}} = 1.58$. For $\alpha = 0.05$ we will not reject the null hypothesis because $z_{1-\alpha/2} = 1.96$ and $1.58 < 1.96$. But, if we had chosen $\alpha = 0.114$ then 1.58 would be the cutoff for the test. So, any α value of 0.114 or larger would result in us rejecting H_0 for this data.

The p-value is the smallest α value that would cause us to reject H_0 for the given data. In this case, the p-value is 0.114 Therefore, we can perform the test by computing the p-value corresponding to the observed data and comparing it to α . If pvalue $< \alpha$, reject H_0 , otherwise retain H_0 .

Caution! In practice we must always choose α before we do the test! If we choose α based on the data, then we are cheating and our test will not preserve Type I error probability at α .

P-values are uniform under the null hypothesis

This is pretty generally true. Here's an example for the simple test of a normal mean:

Based on the previous example it's not hard to see that the p-value is defined as

$$p = 2\left[1 - \Phi\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right)\right]$$

when $\bar{X} > \mu_0$ and as twice the lower tail:

$$p = 2\Phi\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right)$$

when $\bar{X} \leq \mu_0$. Then, compute $P(p < s)$

$$P(p < s) = P\left(2\Phi\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right) < s\right) + P\left(2\left[1 - \Phi\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right)\right] < s\right)$$

$$= P(Z < \Phi^{-1}(s/2)) + P(Z > \Phi^{-1}(1 - s/2)) = s/2 + 1 - (1 - s/2) = s.$$

Since $P(p < s) = s$ then p must be a uniform $(0, 1)$ random variable.

Tests for difference of normal population means

Suppose we are studying two populations, e.g. individuals taking a medication and individuals taking a placebo, and we are interested in comparing the population means. Consider testing

$$H_0 : \mu_X - \mu_Y \leq D \quad \text{versus} \quad H_1 : \mu_X - \mu_Y > D$$

for some chosen constant D . There are two varieties of this test depending on whether or not we assume the two population variances are equal. If $X_i \overset{iid}{\sim} N(\mu_X, \sigma_X^2)$ and $Y_i \overset{iid}{\sim} N(\mu_X, \sigma_Y^2)$ for $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ then $V(\bar{X} - \bar{Y}) = \sigma_X^2/n + \sigma_Y^2/m$. And if $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ then $V(\bar{X} - \bar{Y}) = \sigma^2(1/n + 1/m)$.

In the first case of unequal variances

$$T = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{S_X^2/n + S_Y^2/M}}$$

does NOT have a t distribution. But, as showed by Welch (1947) T has a distribution that can be approximated by a t with a complicated df given by

$$\frac{(\frac{s_X^2}{n} + \frac{s_Y^2}{m})^2}{\frac{s_X^4}{n^2(n-1)} + \frac{s_Y^4}{m^2(m-1)}}$$

Pooled t-test:
If we assume $\sigma_X^2 = \simga_Y^2$ then

$$T = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_p\sqrt{1/n + 1/m}} \sim t(n + m - 2)$$

where $S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$.

Pooled versus unpooled? F-test?

If $X_i \overset{iid}{\sim} N(\mu_X, \sigma_X^2)$ and $Y_i \overset{iid}{\sim} N(\mu_X, \sigma_Y^2)$ for $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ then we have seen that

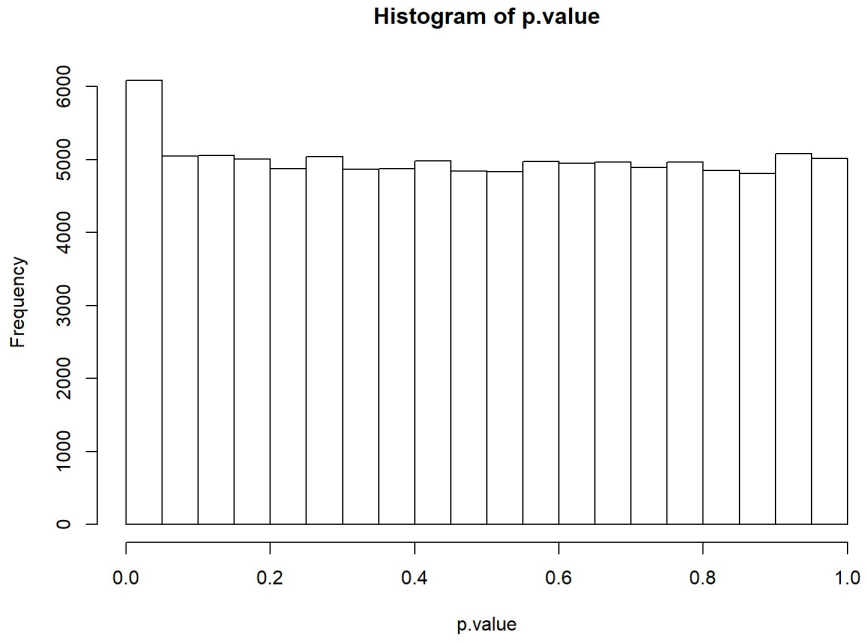
$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F(n-1, m-1).$$

One strategy:
1. Use F-test to decide if variances can be assumed equal
2. Use version of t-test suggested by above F-test.

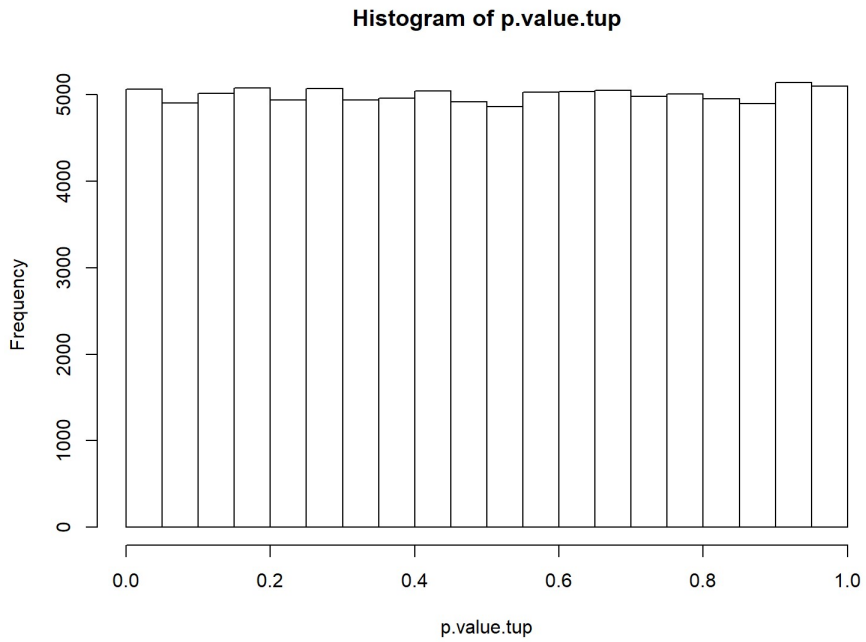
Is this actually a good idea?

Simulations

```r
mu_x = 0
mu_y = 0
sigma_x = 2
sigma_y = 1
n = 10
m = 20
p.value = rep(0,100000)
p.value.tup = rep(0,100000)
for( i in 1:100000){
  X <- rnorm(n,mu_x,sigma_x)
  Y <- rnorm(m,mu_y,sigma_y)
  vx <- var(X)
  vy <- var(Y)
  sp2 <- (vx*(n-1)+vy*(m-1))/(n+m-2)
  tp <- (mean(X)-mean(Y))/sqrt(sp2*(1/n+1/m))
  tup <- (mean(X)-mean(Y))/sqrt(var(X)/n+var(Y)/m)
  sw.df <- ((vx/n+vy/m)^2)/(((vx/n)^2)/(n-1)+((vy/m)^2)/(m-1))
  F <- ifelse((vx/vy >qf(.95,n-1,m-1)) || (vx/vy <qf(.05,n-1,m-1)), 1,0)
  t.pv <- ifelse(F==1,2*min(pt(tup,sw.df),1-pt(tup,sw.df)),2*min(pt(tp,n+m-2),1-pt(tp,n+m-2)))
  p.value[i] <- t.pv
  p.value.tup[i] <- 2*min(pt(tup,sw.df),1-pt(tup,sw.df))
}
hist(p.value)
```

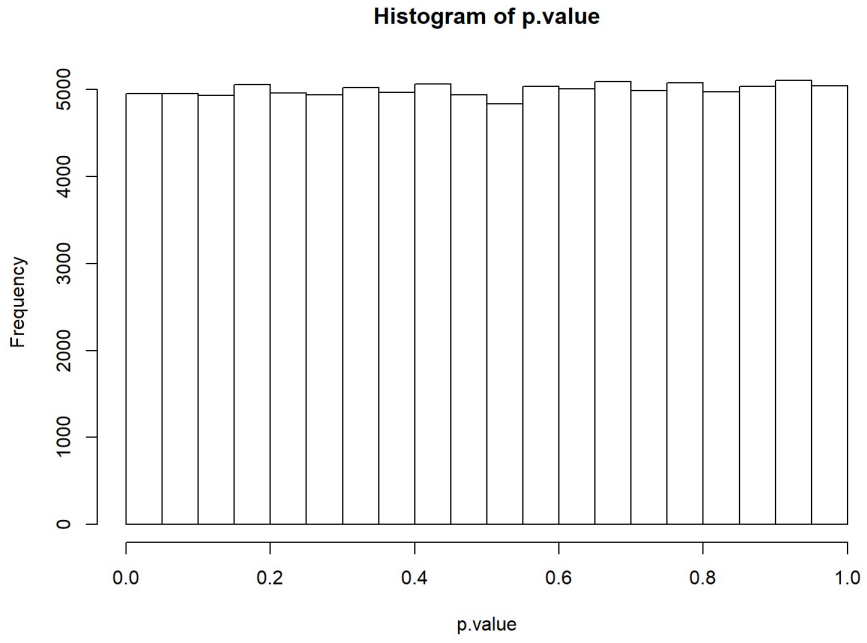



```r
hist(p.value.tup)
```

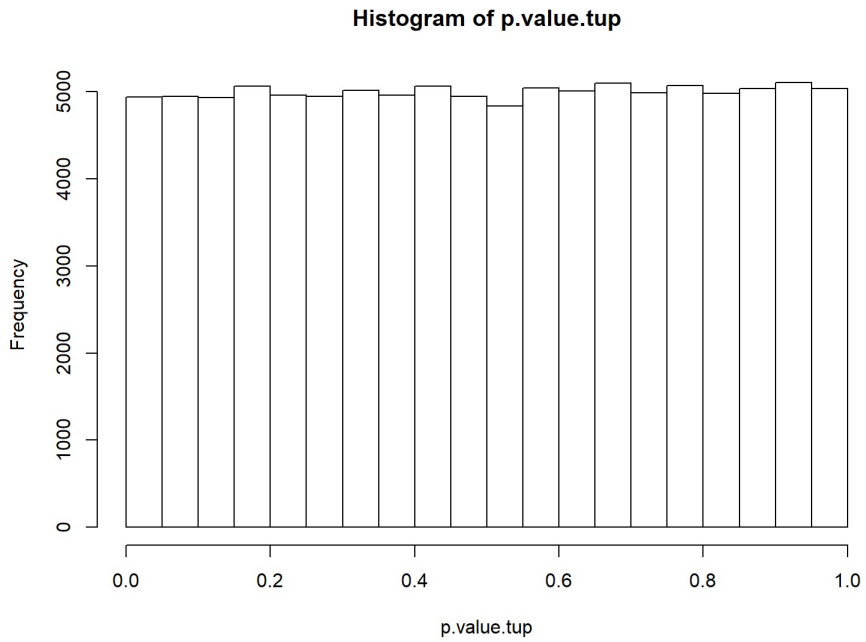

Histogram of p.value.tup
Frequency
0
1000
2000
3000
4000
5000
0.0
0.2
0.4
0.6
0.8
1.0
p.value.tup

Simulations

```r
mu_x = 0
mu_y = 0
sigma_x = 1
sigma_y = 1
n = 20
m = 20
p.value = rep(0,100000)
p.value.tup = rep(0,100000)
for( i in 1:100000){
  X <- rnorm(n,mu_x,sigma_x)
  Y <- rnorm(m,mu_y,sigma_y)
  vx <- var(X)
  vy <- var(Y)
  sp2 <- (vx*(n-1)+vy*(m-1))/(n+m-2)
  tp <- (mean(X)-mean(Y))/sqrt(sp2*(1/n+1/m))
  tup <- (mean(X)-mean(Y))/sqrt(var(X)/n+var(Y)/m)
  sw.df <- ((vx/n+vy/m)^2)/(((vx/n)^2)/(n-1)+((vy/m)^2)/(m-1))
  F <- ifelse((vx/vy >qf(.95,n-1,m-1)) || (vx/vy <qf(.05,n-1,m-1)), 1,0)
  t.pv <- ifelse(F==1,2*min(pt(tup,sw.df),1-pt(tup,sw.df)),2*min(pt(tp,n+m-2),1-pt(tp,n+m-2)))
  p.value[i] <- t.pv
  p.value.tup[i] <- 2*min(pt(tup,sw.df),1-pt(tup,sw.df))
}
hist(p.value)
```




```r
hist(p.value.tup)
```


Histogram of p.value.tup
Frequency
0
1000
2000
3000
4000
5000
0.0
0.2
0.4
0.6
0.8
1.0
p.value.tup