(Elementary) Regression Methods & Computational Statistics (405.952)

Part III: Hypothesis Testing and Confidence Intervals

Assoz.Prof. Dr. Wolfgang Trutschnig

Arbeitsgruppe Stochastik/Statistik
Fachbereich Mathematik
Universität Salzburg
www.trutschnig.net

Salzburg, December 2018

- The alternative distribution is fully determined by one single parameter $p \in [0, 1]$.
- If X has an alternative distribution we will write $X \sim A(p)$ in the sequel.
- If $X \sim A(p)$ then X can only assume two values: 0 and 1; more precisely

$$\mathbb{P}(X = 1) = p, \quad \mathbb{P}(X = 0) = 1 - p.$$

- We all know examples of variables with alternative distribution:
- If we denote the result of flipping a coin by 0 (tails) and 1 (heads), then $p = \frac{1}{2}$ and $X \sim A(\frac{1}{2})$.
- If X denotes the result of rolling a dice and we write 1 if the result is either 5 or 6 and 0 otherwise then $p = \frac{1}{3}$ and $X \sim A(\frac{1}{3})$.

- In practice, we do not know the parameter p and have to estimate it based on a sample $x_1, x_2, \ldots, x_n$.

Example (Election forecasts simplified)

- Suppose that one week before the election 100 (randomly drawn) people are asked which of the two candidates '0' and '1' they will vote for.
- 42 answer '0' and 58 answer '1'.
- How would you estimate $p = \mathbb{P}(X = 1)$?
- Natural choice is $\overline{x}_{100} = 0.58 =: \hat{p}_{100}$.

- Suppose that $Z \sim A(p)$ and we repeat the 'experiment' n times.
- Let X denote the number of 1s observed in the n trials.
- We will write $X \sim Bin(n, p)$ and say that X has *binomial distribution* (with parameters n and p).
- X can attain all integer values between 0 and n.
- With a little bit of mathematics we get the following well-known formula:

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k \in \{0, 1, \ldots, n\}.$$

Example (Election again)

- Suppose that in the election exactly 50% voted for candidate '0' and candidate '1' each.
- We ask 100 randomly selected voters, which candidate they voted for.
- What is the probability that 42 answer '0' and 58 answer '1'?
- $\mathbb{P}(X = 58) = \binom{100}{58} 0.5^{58} 0.5^{42} \approx 0.022.$

Example (Toy example hypothesis testing)

- Suppose that somebody rolls a dice (that you can not see).
- You only know that the dice either has (i) a '1' on four sides and a '0' on the other two sides or (ii) a '1' on two sides and a '0' on the other four sides.
- If we let X denote the result of rolling this dice once, then we either have

$$(i)\ p := \mathbb{P}(X=1) \quad = \quad \frac{4}{6} = \frac{2}{3} \quad \text{and} \quad \mathbb{P}(X=0) = \frac{2}{6} = \frac{1}{3} = 1-p$$

or

$$(ii)\ p := \mathbb{P}(X=1) \quad = \quad \frac{2}{6} = \frac{1}{3} \quad \text{and} \quad \mathbb{P}(X=0) = \frac{4}{6} = \frac{2}{3} = 1-p.$$

- In other words, $X \sim A(p)$ and we know that $p \in \Theta = \{\frac{2}{3}, \frac{1}{3}\}$.
- We will call $H_0 : p = \frac{2}{3}$ the *null hypothesis* and $H_1 : p = \frac{1}{3}$ the *alternative hypothesis* (for whatever reason).

Example (Toy example hypothesis testing, cont.)

- For the moment we focus on $H_0 : p = \frac{2}{3}$.
- Suppose that the dice is rolled twice and the result is denoted by (X_1, X_2).
- Possibility 1: $(X_1, X_2) = (1, 1)$. Would you stick to H_0 or reject H_0 (i.e. change to H_1), and why?
- Possibility 2: $(X_1, X_2) = (1, 0)$. Would you stick to H_0 or reject H_0, and why?
- Possibility 3: $(X_1, X_2) = (0, 1)$. Would you stick to H_0 or reject H_0, and why?
- Possibility 4: $(X_1, X_2) = (0, 0)$. Would you stick to H_0 or reject H_0, and why?
- Which criterion is your decision based upon?
- For a given observation we check under which of the two hypotheses the observation has higher probability.

Example (Toy example hypothesis testing, cont.)

- If H_0 is correct then we have

$$\begin{aligned}\mathbb{P}_{H_0}(X_1 = 1, X_2 = 1) &= \frac{4}{9}, \quad \mathbb{P}_{H_0}(X_1 = 1, X_2 = 0) = \frac{2}{9}\\ \mathbb{P}_{H_0}(X_1 = 0, X_2 = 1) &= \frac{2}{9}, \quad \mathbb{P}_{H_0}(X_1 = 0, X_2 = 0) = \frac{1}{9}.\end{aligned}$$

- If H_1 is correct then we have

$$\begin{aligned}\mathbb{P}_{H_1}(X_1 = 1, X_2 = 1) &= \frac{1}{9}, \quad \mathbb{P}_{H_1}(X_1 = 1, X_2 = 0) = \frac{2}{9}\\ \mathbb{P}_{H_1}(X_1 = 0, X_2 = 1) &= \frac{2}{9}, \quad \mathbb{P}_{H_1}(X_1 = 0, X_2 = 0) = \frac{4}{9}.\end{aligned}$$

- In case of $(1, 1)$ we do not reject H_0.
- In case of $(1, 0)$ and in case of $(0, 1)$ we do not reject H_0 (the observation is equally probable under both hypotheses, so by changing from H_0 to H_1 we don't gain anything).
- In case of $(0, 0)$ we reject H_0.

Example (Toy example hypothesis testing, cont.)

- We intuitively reject H_0 if - under the assumption that H_0 is true - the observation we made is very unlikely (in the sense of having low probability).
- In our toy setting we can make two different mistakes:
- **Type I error**: We reject H_0 although it is correct.
- **Type II error**: We do not reject (accept) H_0 although it is wrong.
- Let us calculate the probability of a type I and the probability of a type II error in our toy setting:
- @type I error α:

$$\alpha := \mathbb{P}_{H_0}(\text{reject } H_0) = \mathbb{P}_{H_0}(X_1 = 0, X_2 = 0) = \frac{1}{9}$$

- We have a chance of more than 11% to make a type I error.

Example (Toy example hypothesis testing, cont.)

- @type II error β:

$$\begin{aligned}\beta := \mathbb{P}_{H_1}(\text{accept } H_0) &= \mathbb{P}_{H_1}(X_1 = 1, X_2 = 1) + \mathbb{P}_{H_1}(X_1 = 1, X_2 = 0) \\ &\quad + \mathbb{P}_{H_1}(X_1 = 0, X_2 = 1) \\ &= 1 - \mathbb{P}_{H_1}(X_1 = 0, X_2 = 0) = \frac{5}{9}\end{aligned}$$

- We have chance of more than 55% to make a type II error.
- Could we improve our decision criterion to reduce the type I and the type II error?
- Is there a perfect decision rule such that $\alpha = \beta = 0$?
- If we want $\alpha = 0$ then we can NEVER reject H_0, so we get $\beta = 1$.
- If we want $\beta = 0$ then we always have to reject H_0, so we get $\alpha = 1$.
- α and β are antagonists.
- Which one is more important? Think of a criminal trial...

Hypothesis testing vs. criminal trials

- Consider a criminal trial.
- Based on evidence the jury (or the judge) has to decide whether the defendant is guilty or not.
- Suppose that $H_0 = \{\text{innocent}\}$ and that $H_1 = \{\text{guilty}\}$.
- Right at the start the jury (or the judge) accepts H_0 and assumes that the defendant is innocent.
- Only if enough evidence is brought in, H_0 will be rejected and the defendant will be declared guilty.
- The afore-mentioned type I error α corresponds to the situation that the defendant will be declared guilty although he is innocent.
- The afore-mentioned type II error β corresponds to the situation that the defendant will be declared innocent although he is guilty.

- Which error has worse consequences for the defendant?
- Obviously the type I error.
- In the Anglo-Saxon jurisdiction system there there is the term 'Beyond reasonable doubt' underlining this fact.
- In other words: We want to keep the type I error α (very) small.
- The same applies to hypothesis testing: α should be small, standard *significance levels* are $\alpha = 0.05$ and $\alpha = 0.01$ (one error out of twenty or one out of hundred).
- As soon as α is fixed it is the statisticians' job to develop optimal tests, i.e. decision rules (criteria) with a probability of (at most) α for a type I error and, at the same time, minimal type II error β.

Example (Toy example hypothesis testing, cont.)

- Suppose we fix $\alpha = 0.05$ and want to develop a decision rule (i.e. a criterion when to reject H_0) such that the probability of a type I error is at most 0.05.
- Since, under $H_0 : p = \frac{2}{3}$ all four possible outcomes have at least a probability of $\frac{1}{9}$ the only choice we have is never to reject H_0, in which case $\beta = 1$.
- This looks pretty bad at first sight...keeping in mind, however, the criminal trial comparison it would mean that the jury should not declare the defendant guilty if there is not enough evidence against it (remember: 'Beyond reasonable doubt').
- If, instead of sample size two (two observations), we had sample size $n = 100$ the situation would improve - let's develop a simple test for this situation:
- As before we have $H_0 : p = \frac{2}{3}$ and $H_1 : p = \frac{1}{3}$ and we want the error of type I to be at most 0.05.
- A natural idea is the following: Reject H_0 if the sample $x_1, x_2, \ldots, x_n$ contains 0 too many times or, equivalently, 1 not often enough.

Example (Toy example hypothesis testing, cont.)

- How to determine the threshold t?
- Under H_0 the number k of 1s in the sample of size $n = 100$ has a Binomial distribution $Bin(n, p)$ with parameter $p = \frac{2}{3}$, i.e

$$\mathbb{P}_{H_0}(K = k) = \binom{100}{k}\left(\frac{2}{3}\right)^k\left(\frac{1}{3}\right)^{100-k}.$$

- The threshold t has to fulfill

$$\mathbb{P}_{H_0}(K \leq t) \stackrel{!}{=} 0.05. \tag{1}$$

- There is no exact solution t of equation (1) so we calculate the biggest t fulfilling

$$\mathbb{P}_{H_0}(K \leq t) \leq 0.05 \tag{2}$$

and get $t = 58$ (see R-Code).

Example (Toy example hypothesis testing, cont.)

- Altogether we have arrived at the following test for H_0 vs. H_1 given $n = 100$ observations $x_1, \ldots, x_n$:
- Reject H_0 if the number K of 1s in the sample fulfills $K \leq 58$.
- Do not reject H_0 if $K > 58$.
- It follows from the construction (again see R-Code) that

$$\alpha = \mathbb{P}_{H_0}(\text{reject } H_0) = \mathbb{P}_{H_0}(K \leq 58) = 0.04337149,$$

  i.e. in 4.3% of all cases we reject H_0 although it is correct.
- How big is the probability of a type II error?
- We calculate it as before and get

$$\beta = \mathbb{P}_{H_1}(\text{accept } H_0) = \mathbb{P}_{H_1}(K > 58) = 1 - \mathbb{P}_{H_1}(K \leq 58) = 0.00000012907.$$

- How can this be interpreted?

Example (Toy example hypothesis testing, cont.)

- A quick look at the R-Code

```
#determine the threshold for the test H0: p=2/3 versus H1: p=1/3
plot(0:100,pbinom(0:100,size=100,prob=2/3),type="p")
abline(h=0.05)

t<-qbinom(p=0.05,size=100,prob=2/3)-1
t
[1] 58

pbinom(t,size = 100,prob=2/3)
[1] 0.04337149

#calculate beta
1-pbinom(t,size=100,prob=1/3)
[1] 1.290734e-07
```

Example (Toy example hypothesis testing, cont.)

- Let us check if the just developed test really performs as it should - we run simulations (always important especially in the context of hypothesis testing).

```
#evaluate performance of the developed test
# one run under H0:
n<-100
p<-2/3
x<-sample(c(1,0),size=n,replace = TRUE,prob=c(2/3,1/3))
if(length(x[x==1])<=58){print("reject H0")}
```

```

# R=10000 runs under H0
R<-10000
reject<-rep(0,R)
for(i in 1:R){
  x<-sample(c(1,0),size=n,replace = TRUE,prob=c(2/3,1/3))
  if(length(x[x==1])<=58){reject[i]<-1}
}
mean(reject)
[1] 0.0445

barplot(table(reject))
```

Example (Toy example hypothesis testing, cont.)

- Simulations for the type II error.

```
# R=10000 runs under H1
R<-10000
reject<-rep(0,R)
for(i in 1:R){
  x<-sample(c(1,0),size=n,replace = TRUE,prob=c(1/3,2/3))
  if(length(x[x==1])<=58){reject[i]<-1}
}
1-mean(reject)
[1] 0
```

- The type II error is really (almost) zero, i.e. if $H_1 : p = \frac{1}{3}$ is true, the test detects it (almost) every time.

Exercise 27:

- Suppose that the toy example is slightly modified as follows:
- You only know that the dice either has (i) a 1 on three sides and a 0 on the other three sides or (ii) a 1 on two sides and a 0 on the other four sides.
- Develop a test with type I error of at most 0.05 for this situation, i.e. a test for $H_0 : p = \frac{1}{2}$ vs. $H_1 : p = \frac{1}{3}$.
- Evaluate the performance of this test by modifying the provided R-Code accordingly.
- Work with different sample sizes, e.g. $n = 10, n = 20, n = 50, n = 100, n = 500$, and describe the influence of the sample size on α and (more importantly) on β.

Quick reminder

- We had an experiment X with a binary output 1 and 0.
- We knew that the success probability $p = \mathbb{P}(X = 1)$ was either $p = \frac{2}{3}$ or $p = \frac{1}{3}$.
- We developed a hypothesis test for $H_0 : p = \frac{2}{3}$ versus $H_1 : p = \frac{1}{3}$ based on samples $x_1, \ldots, x_n$ of size $n = 100$.
- The test we developed at a significance level $\alpha = 0.05$ was to reject H_0 if the number K of ones in $x_1, \ldots, x_n$ fulfills $K \leq 58$.
- The probability of a type I error (what was that?) was $\alpha = \mathbb{P}_{H_0}(K \leq 58) = 0.04337149$.
- The probability of a type II error (what was that?) was $\beta = \mathbb{P}_{H_1}(K > 58) = 0.00000012907$.
- How can these two values be interpreted?

- Assume that H_0 is correct:
- Then out of $R = 10.000$ times we falsely reject H_0 approx. 434 times
- Assume that H_1 is correct:
- Then out of $R = 10.000$ times we do not reject H_0 approx. 0 times
- Remember that α and β can not be minimized simultaneously, so α comes first (criminal trial comparison).

- Suppose we now want to test $H_0 : p \geq \frac{1}{2}$ vs. $H_1 : p < \frac{1}{2}$ at significance level $\alpha = 0.05$.
- Why is this situation more complicated and what is the key difference to $H_0 : p = \frac{2}{3}$ versus $H_1 : p = \frac{1}{3}$?
- H_0 and H_1 are **composite**, i.e. they contain more than one value of the parameter.

- How could we extend the definition of the type I error $\mathbb{P}_{H_0}(\text{reject } H_0)$ to this situation?
- If the true parameter is p then H_0 holds whenever $p \geq \frac{1}{2}$.
- What we want is

$$\mathbb{P}_p(\text{reject } H_0) \leq 0.05 \tag{3}$$

  for every $p \geq \frac{1}{2}$.
- Mathematically speaking we want

$$\max_{p \in H_0} \mathbb{P}_p(\text{reject } H_0) \leq 0.05$$

- Does it make sense to proceed analogously with the type II error β and set

$$\beta = \max_{p \in H_1} \mathbb{P}_p(\text{accept } H_0)?$$

- No, because we would get $\beta = 1$.

- As a consequence we calculate β for every value $p \in H_1$ and simply write $\beta(p)$, i.e.

$$\beta(p) = \mathbb{P}_p(\text{accept } H_0) \tag{4}$$

- In our situation we expect $\beta(p)$ to be small if p is very small (close to 0).
- And we expect $\beta(p)$ to be big if p is close to $\frac{1}{2}$.
- The function $\pi(p) = 1 - \beta(p)$ is called **power function** - the higher the value the better.

- Back to the original problem: How to construct a hypothesis test for $H_0 : p \geq \frac{1}{2}$ vs. $H_1 : p < \frac{1}{2}$?
- Why might such a test be of practical relevance?

The binom.test function in R

- The test we are looking for is already implemented in R.

```
#binom.test for testing H0: p>=0.5 versus H1: p<0.5
p <- 0.55
n <- 100
x <- sample(c(0,1),size=n,replace=TRUE,prob=c(1-p,p))
successes <- sum(x)
test <- binom.test(successes,n,p=0.5,alternative="less")
test
```

- yields

```
Exact binomial test

data:  successes and n
number of successes = 61, number of trials = 100, p-value =
     0.9895
alternative hypothesis: true probability of success is less than
      0.5
95 percent confidence interval:
0.0000000 0.6918993
sample estimates:
probability of success
0.61
```

- How can the output be interpreted? Is H_0 rejected or not?
- How is the p-value calculated and what does it tell us?
- We reject H_0 if the **p-value** returned by R is smaller than $\alpha = 0.05$.
- The smaller the p-value the more evidence against H_0.

- Loosely speaking, the p-value is the probability under H_0, to observe 'something at least as extreme as the current value'.
- What does 'something at least as extreme as 61' mean in our case?
- It means that the number of successes X is at most 61.
- In other words:

$$p = \max_{p \in H_0} \mathbb{P}_p(X \leq 61) = \mathbb{P}_{0.5}(X \leq 61) \approx 0.9895$$

- How can we check if binom.test really does what it should?
- We check by simulations if the type I error is at most 0.05.
- Afterwards we approximate the power function again via simulations.

Checking the performance of binom.test

- We analyze the performance of binom.test via simulations

```
#assume that H0 holds
#repeat the above procedure R=10000 times and calculate the
    portion of false decisions (type I error)
R <- 10000
error <- rep(0,R)
for(i in 1:R){
  p <- 0.6
  n <- 100
  x <- sample(c(0,1),size=n,replace=TRUE,prob=c(1-p,p))
  successes <- sum(x)
  test <- binom.test(successes,n,p=0.5,alternative="less")
  if(test$p.value<0.05){error[i] <- 1}
}
mean(error)
```

- yields

```
[1] 0.0036
```

Checking the performance of binom.test

```r
#worst case scenario (what is different to before?)
R <- 10000
error <- rep(0,R)
for(i in 1:R){
  p <- 0.5
  n <- 100
  x <- sample(c(0,1),size=n,replace=TRUE,prob=c(1-p,p))
  successes <- sum(x)
  test <- binom.test(successes,n,p=0.5,alternative="less")
  if(test$p.value<0.05){error[i] <- 1}
}
mean(error)
```

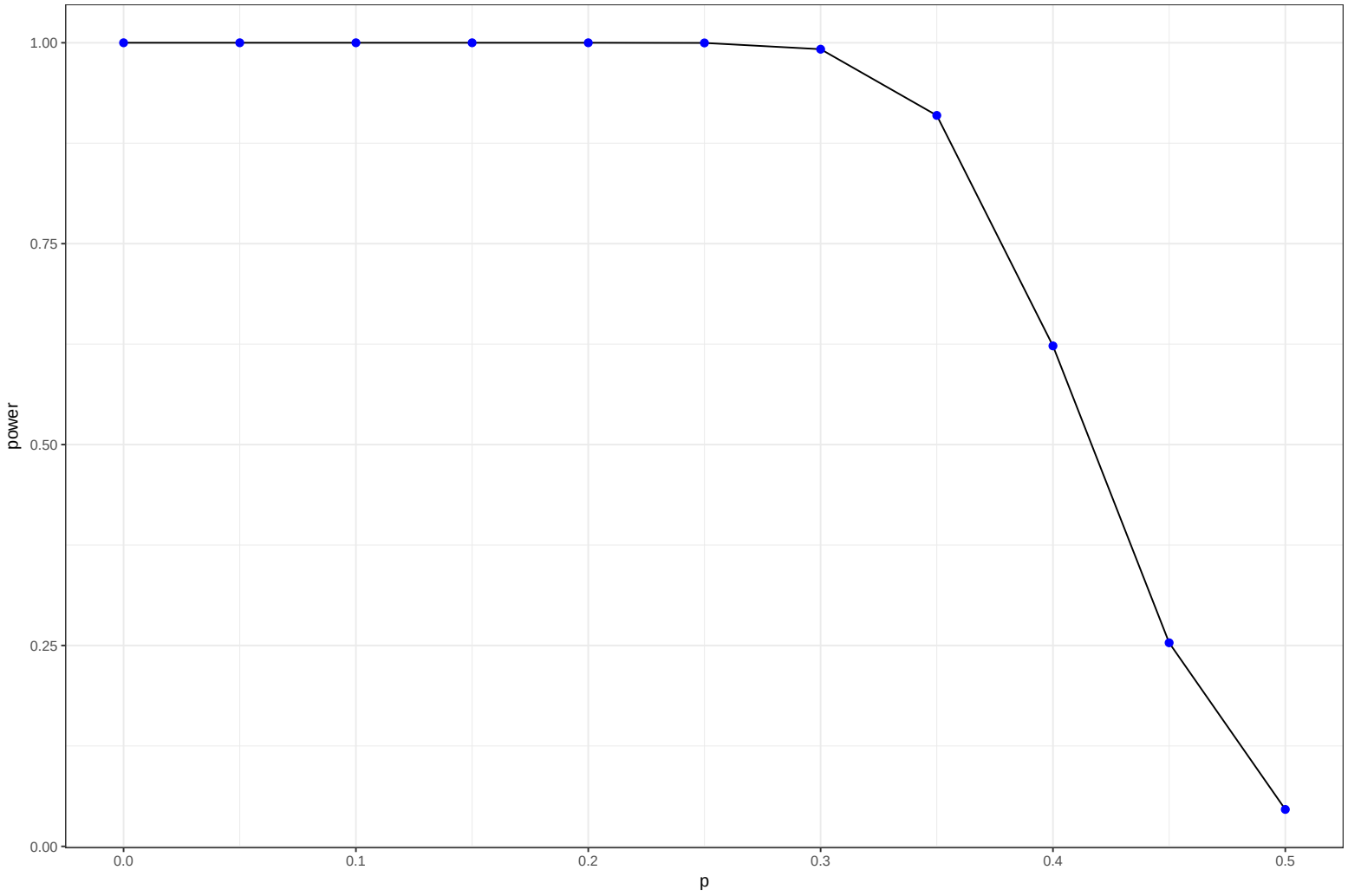
- yields

```
[1] 0.0441
```

Checking the performance of binom.test

```
#@power: choose different values for p in H1 and calculate the
    power
pgrid <- seq(0,0.5,by=0.05)
power <- rep(0,length(pgrid))
for(j in 1:length(pgrid)){
 print(j)
 R <- 5000
 error <- rep(0,R)
 for(i in 1:R){
  p <- pgrid[j]
  n <- 100
  x <- sample(c(0,1),size=n,replace=TRUE,prob=c(1-p,p))
  successes <- sum(x)
  test <- binom.test(successes,n,p=0.5,alternative="less")
  if(test$p.value>=0.05){error[i] <- 1}          #type II error
 }
 power[j]<-1-mean(error)
}
power
[1] 1.0000 1.0000 1.0000 1.0000 1.0000 0.9998 0.9920 0.9134
    0.6220 0.2532 0.0474
```

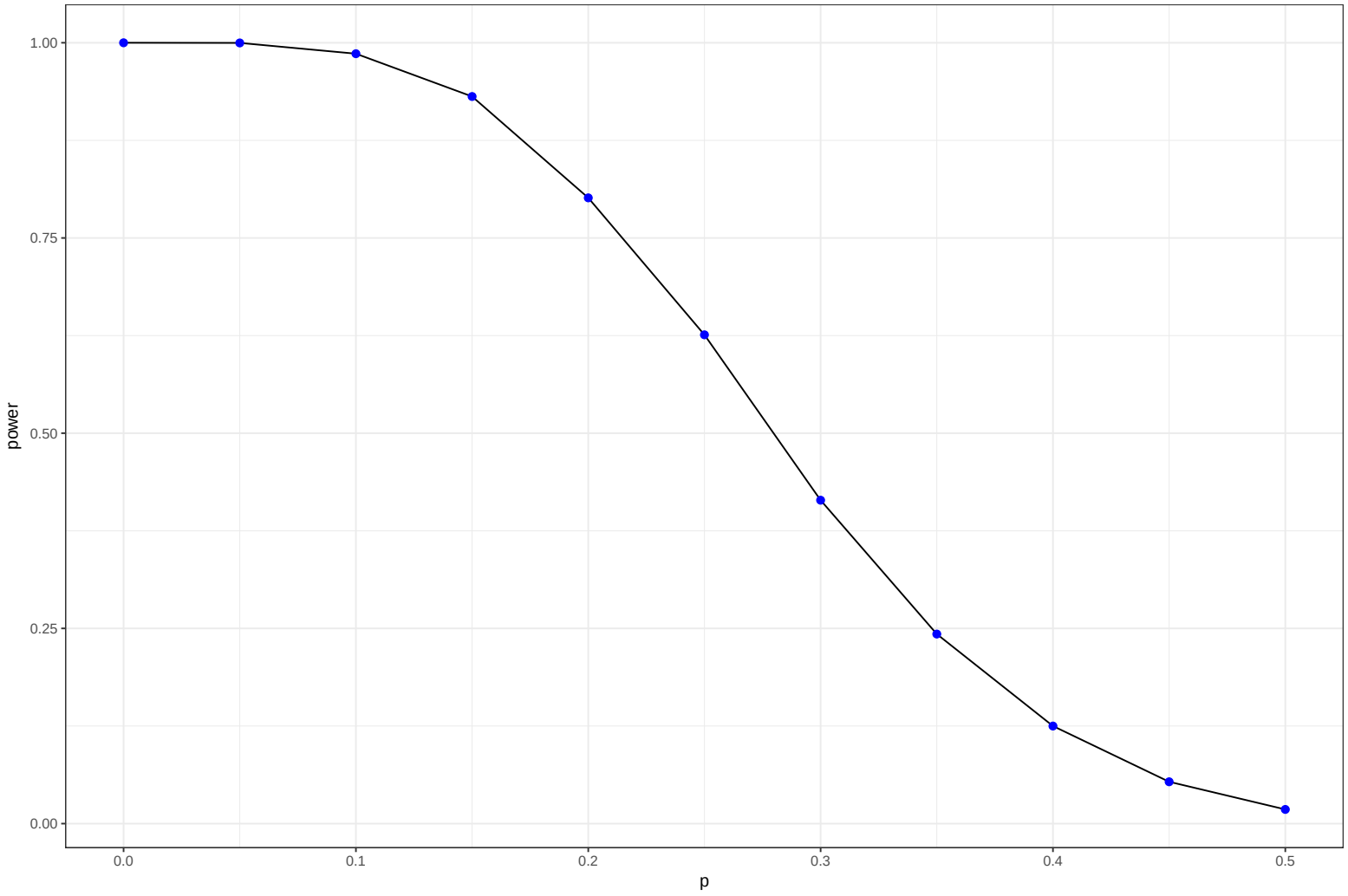
Checking the performance of binom.test



Checking the performance of binom.test

```
#@power: same for smaller sample size n
pgrid <- seq(0,0.5,by=0.05)
power <- rep(0,length(pgrid))
for(j in 1:length(pgrid)){
  print(j)
  R <- 5000
  error <- rep(0,R)
  for(i in 1:R){
    p <- pgrid[j]
    n <- 20
    x <- sample(c(0,1),size=n,replace=TRUE,prob=c(1-p,p))
    successes <- sum(x)
    test <- binom.test(successes,n,p=0.5,alternative="less")
    if(test$p.value>=0.05){error[i] <- 1}
  }
  power[j]<-1-mean(error)
}
power
[1] 1.0000 0.9996 0.9876 0.9284 0.8082 0.6130 0.4238 0.2458
    0.1306 0.0548 0.0210
```

Checking the performance of binom.test



Exercise 28:

- Use binom.test to test the hypothesis $H_0 : p \leq 0.7$ versus $H_1 : p > 0.7$.
- Check that the type I error is at most 0.05 for every $p \in H_0$.
- Calculate/approximate the power function $\pi(p)$ for sample size $n = 100$ via (sufficiently many) simulations.
- Work with different sample sizes, e.g. $n = 10$, $n = 20$, $n = 50$, $n = 100$, $n = 500$, $n = 1000$, and produce a plot of the power function π in each case.
- How can the results be interpreted?

Exercise 29:

- Use binom.test for testing the hypothesis $H_0 : p = 0.5$ versus $H_1 : p \neq 0.5$.
- Check that the type I error is at most 0.05.
- Calculate/approximate the power function $\pi(p)$ for sample size $n = 100$ via (sufficiently many) simulations.
- Work with different sample sizes, e.g. $n = 10$, $n = 20$, $n = 50$, $n = 100$, $n = 500$, $n = 1000$, and produce a plot of the power function π in each case
- How can the results be interpreted?

t-tests are possibly the most (mis)used tests in various disciplines; we start with the one-sample version:

One-sample t-tests

- Suppose that $X \sim \mathcal{N}(\mu, \sigma^2)$ but we do not know μ and σ^2.
- Given a sample $x_1, \ldots, x_n$ from X we are interesting in testing one of the following three hypotheses concerning μ:
- (i) $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$
- (ii) $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$
- (iii) $H_0 : \mu \geq \mu_0$ versus $H_1 : \mu < \mu_0$
- **NB: The test only does what it should if X has normal distribution!** (normality has to be checked in advance).
- All three tests are implemented in R via the function **t.test**, which works as follows in case of (i):

The t-test for one sample

```
#t-tests:
mu0 <- 0
sigma <- 1
n <- 1000
x <- rnorm(n,mean=mu0,sd=sigma)
hist(x)

test <- t.test(x,mu=mu0,alternative="two.sided")
test
```

yields

```
One Sample t-test

data:  x
t = -1.131, df = 999, p-value = 0.2583
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
-0.09744758  0.02618977
sample estimates:
mean of x
-0.03562891
```

- We reject H_0 if the **p-value** returned by R is smaller than $\alpha = 0.05$.
- Loosely speaking, the p-value is the probability under H_0, to observe something at least as extreme as the current sample.
- The smaller the p-value the more evidence against H_0.

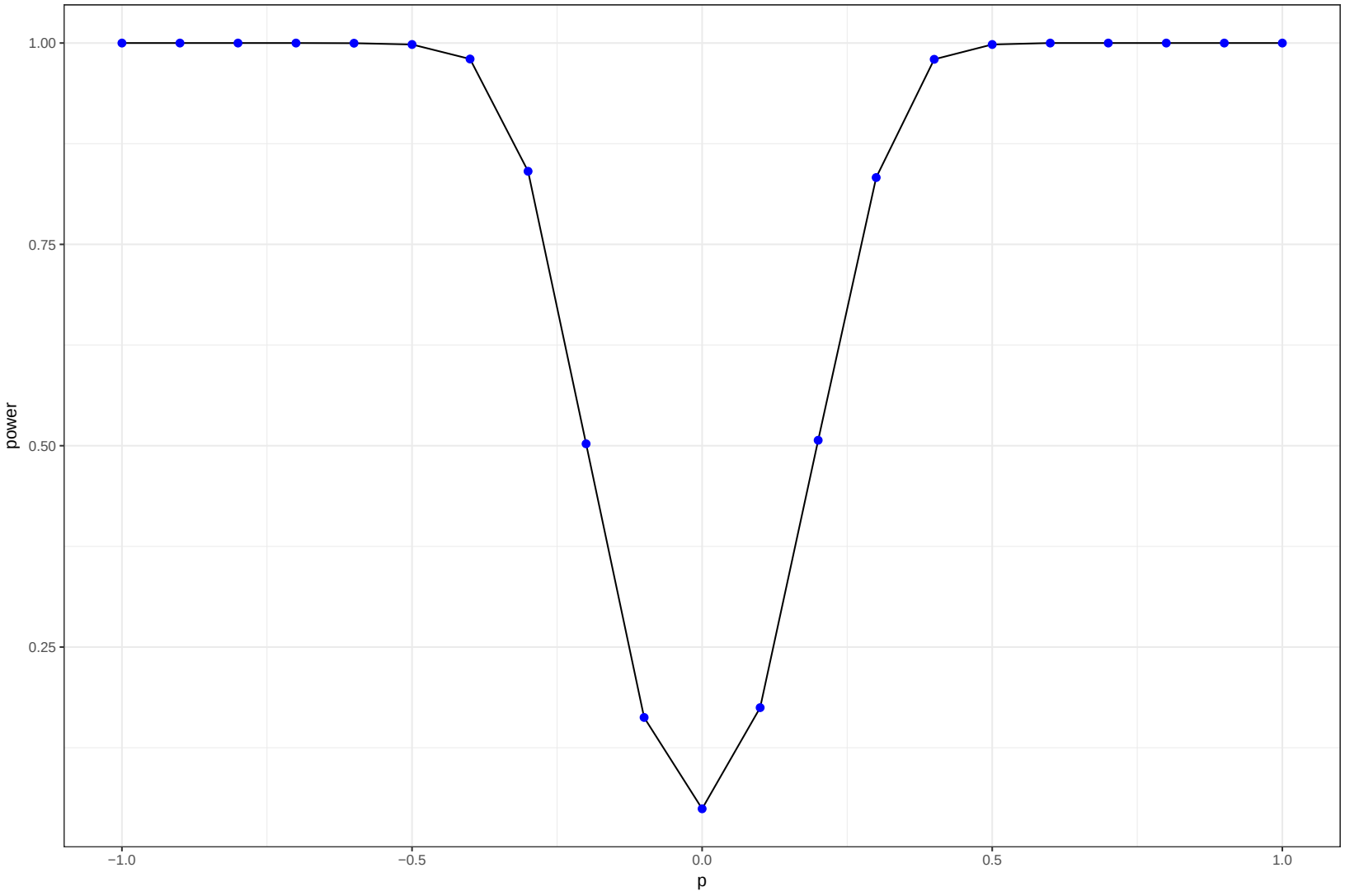
- How can we check if t.test really does what it should?
- We proceed analogously as with binom.test.
- We check by simulations if the type I error is at most 0.05.
- Afterwards we approximate the power function π again via simulations.

The t-test for one sample

```
#assume that H0: mu = mu0 holds
#repeat the above procedure R=10000 times and calculate the
    portion of false decisions (type I error)
R <- 10000
error <- rep(0,R)
for(i in 1:R){
  mu0 <- 0
  sigma <- 1
  n <- 100
  x <- rnorm(n,mean=mu0,sd=sigma)
  test <- t.test(x,mu=mu0,alternative="two.sided")
  if(test$p.value<0.05){error[i] <- 1}
}
mean(error)
[1] 0.0511
```

```
#@power: choose different values for mu in H1 and calculate the
    power
mugrid <- seq(-1,1,by=0.1)
power <- rep(0,length(mugrid))
for(j in 1:length(mugrid)){
  print(j)
  R <- 5000
  error <- rep(0,R)
  for(i in 1:R){
    mu0 <- mugrid[j]
    sigma <- 1
    n <- 100
    x <- rnorm(n,mean=mu0,sd=sigma)
    test <- t.test(x,mu=0,alternative="two.sided")
    if(test$p.value>=0.05){error[i] <- 1}
  }
  power[j]<-1-mean(error)
}
power
[1] 1.0000 1.0000 1.0000 1.0000 0.9998 0.9982 0.9802 0.8408
    0.5024 0.1626 0.0492 0.1748 0.5068 0.8330 0.9798 0.9982
    1.0000 1.0000 1.0000 1.0000 1.0000
```

The t-test for one sample



Exercise 30:

- Use t.test to test the hypothesis $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$.
- Check that the type I error is at most 0.05 whenever H_0 holds.
- Calculate/approximate the power function $\pi(p)$ for sample size $n = 100$ via (sufficiently many) simulations .
- Work with different sample sizes, e.g. $n = 10$, $n = 20$, $n = 50$, $n = 100$, $n = 500$, $n = 1000$, and produce a plot of the power function π in each case.
- How can the results be interpreted?

Exercise 31:

- Find out what the function 'power.t.test' does.
- How can it be used to solve Exercise 30?

Unpaired (independent) two-sample t-test

- Suppose that $X \sim \mathcal{N}(\mu_x, \sigma^2)$, we do not know μ_x and σ^2.
- Suppose that $Y \sim \mathcal{N}(\mu_y, \sigma^2)$, we do not know μ_y and σ^2.
- Notice that the variance of X and Y is the same (and unknown).
- Given a sample $X_1, \ldots, X_n$ from X and a sample $Y_1, \ldots, Y_m$ from Y with $m, n \geq 2$ we want to test one of the following three hypotheses concerning $\mu_D := \mu_x - \mu_y$:
- (i) $H_0 : \mu_D = 0$ versus $H_1 : \mu_D \neq 0$
- (ii) $H_0 : \mu_D \leq 0$ versus $H_1 : \mu_D > 0$
- (iii) $H_0 : \mu_D \geq 0$ versus $H_1 : \mu_D < 0$
- For all three cases the test statistic $T_{n,m}$ is given by

$$T_{n,m} = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{S_p^2}\sqrt{\frac{1}{n} + \frac{1}{m}}},$$

where S_p^2 is the pooled sample variance and given by

$$S_p^2 = \frac{(n-1)S_n^2 + (m-1)S_m^2}{n+m-2}$$

Unpaired (independent) two-sample t-test

- Consider (i) $H_0 : \mu_D = 0$ versus $H_1 : \mu_D \neq 0$
- Under H_0 the test statistic $T_{n,m}$ has t_{n+m-2}-distribution
- Given concrete observations we will there reject H_0 at level α if and only if $|T_{n,m}| > t_{n+m-2,1-\frac{\alpha}{2}}$, where $t_{n+m-2,1-\frac{\alpha}{2}}$ denotes the $1-\frac{\alpha}{2}$-quantile the t-distribution with $n+m-2$ degrees of freedom.
- The p-value of this test is given by

$$p.value = \mathbb{P}(|T| \geq |T_{n,m}|)$$

  where $T_{n,m}$ is the value of the test statistic and $T \sim t_{n+m-2}$.
- **NB: The test only does what it should if X and Y have normal distribution!** (normality has to be checked in advance).
- All three tests are implemented in R via the function **t.test** and work as follows in the case of (i):

Unpaired (independent) two-sample t-test

```
mux <- muy <- 0
sigmax <- 1; sigmay <- 2
n <- 1000
x <- rnorm(n,mean=muy,sd=sigmax)
y <- rnorm(n,mean=muy,sd=sigmay)

test <- t.test(x,y,paired=FALSE,alternative="two.sided")
test
```

yields

```
Welch Two Sample t-test

data:  x and y
t = 0.32697, df = 1515.4, p-value = 0.7437
alternative hypothesis: true difference in means is not equal to
      0
95 percent confidence interval:
  -0.1125703  0.1576068
sample estimates:
  mean of x    mean of y
0.009047213 -0.013471049
```

Unpaired (independent) two-sample t-test

```
#repeat the above procedure R=10000 times and calculate the
    portion of false decisions (type I error)
R <- 10000
error <- rep(0,R)
for(i in 1:R){
  mux <- muy <- 0
  sigmax <- 1; sigmay <- 2
  n <- 1000
  x <- rnorm(n,mean=muy,sd=sigmax)
  y <- rnorm(n,mean=muy,sd=sigmay)
  test <- t.test(x,y,paired=FALSE,alternative="two.sided")
  if(test$p.value<0.05){error[i] <- 1}
}
mean(error)
```

- yields

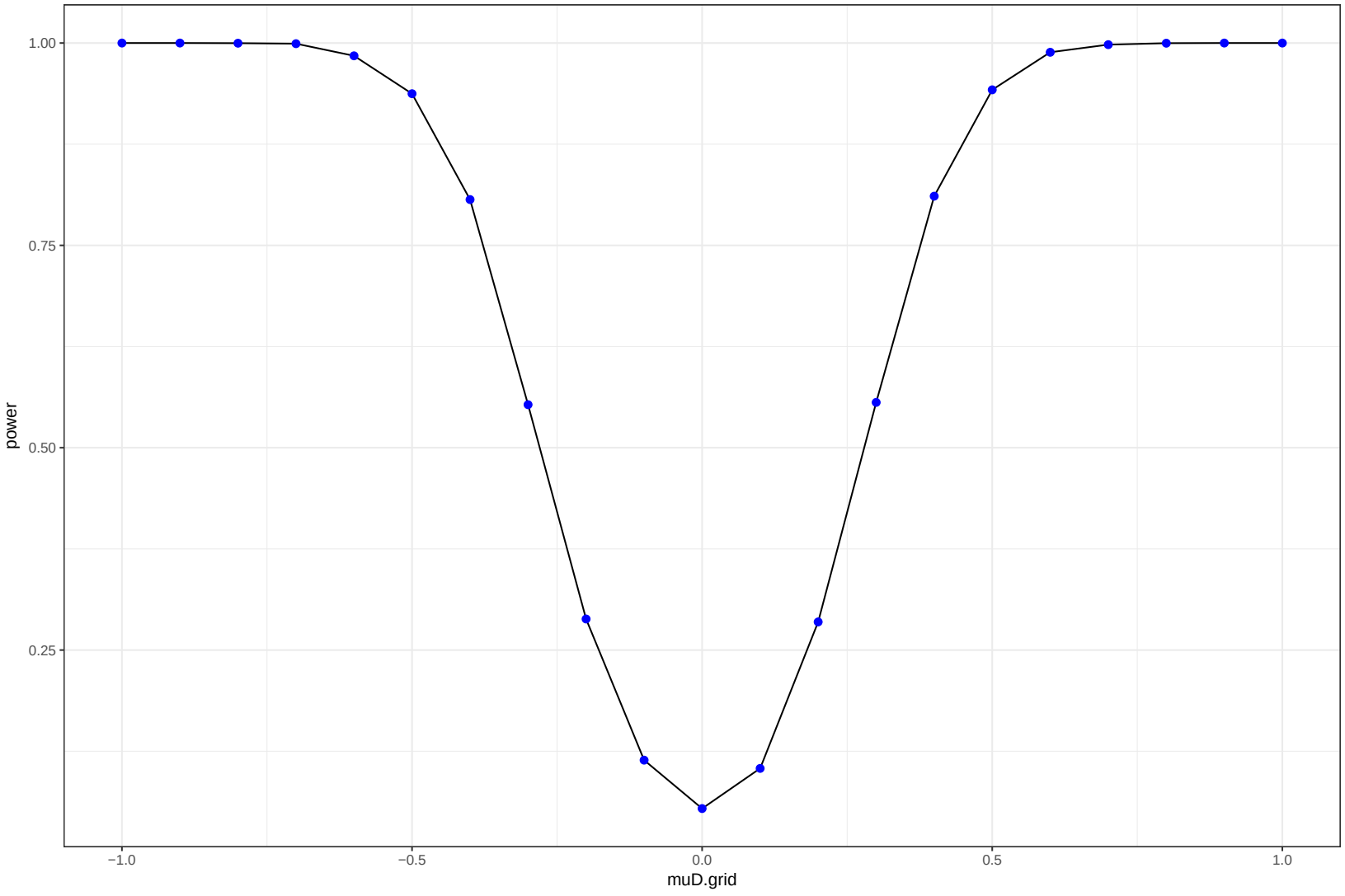
```
[1] 0.0499
```

Unpaired (independent) two-sample t-test

```
#estimate the power of the test
muD.grid <- seq(-1,1,by=0.1)
power <- rep(0,length(muD.grid))
for(j in 1:length(muD.grid)){
  print(j)
  R <- 5000
  error <- rep(0,R)
  for(i in 1:R){
    mux <- 0
    muy <- mux + muD.grid[j]
    sigma <- 1
    n <- 100
    x <- rnorm(n,mean=mux,sd=sigma)
    y <- rnorm(n,mean=muy,sd=sigma)
    test <- t.test(x,y,paired = FALSE,alternative="two.sided")
    if(test$p.value>0.05){error[i] <- 1}
  }
  power[j]<-1-mean(error)
}
power
```

- yields

Unpaired (independent) two-sample t-test



Exercise 32:

- Use t.test to test the hypothesis $H_0 : \mu_x \leq \mu_y$ versus $H_1 : \mu_x > \mu_y$.
- Check that the type I error is at most 0.05 whenever H_0 holds.
- Calculate/approximate the power function $\pi(\mu_D)$ for sample size $n = 1000$ via (sufficiently many) simulations and different values for μ_y (for instance on a grid from -1 to 1).
- How does the power function change if the sample size is decreased/increased?

Exercise 33:

- Find out how the function 'power.t.test' can be used to solve the previous exercise.

- The t-test only does what it should if X and Y have normal distribution! (normality has to be checked in advance).
- What to do if the normality assumption is not met?
- Suppose that we are in again in the setting (i) from before, i.e. we want to test $\mu_D = 0$ given observations $X_1, \ldots, X_n$ from X and $Y_1, \ldots, Y_m$ from Y.
- Naive idea: If H_0 holds, then 'mixing' the observations should not change much, i.e. the absolute difference of the resulting means $|\overline{X}'_n - \overline{Y}'_m|$ should not be much bigger than the original absolute difference $|\overline{X}_n - \overline{Y}_m|$.
- Main idea of the permutation test is to 'mix' the observations many times and then check in how many cases $|\overline{X}'_n - \overline{Y}'_m| > |\overline{X}_n - \overline{Y}_m|$ holds.
- If the last inequality holds in less than 5% of the case H_0 is rejected.
- Let's check the details directly in R.

Permutation test

```
#permutation test for the case n=m: H0: mu_D=0
mux <- muy <- 0
n <- 100
sigma <- 1
x.orig <- rnorm(n,mean=mux,sd=sigma)
y.orig <- rnorm(n,mean=muy,sd=sigma)
dist.orig <- abs(mean(x.orig)-mean(y.orig))
all.orig <- c(x.orig,y.orig)

R <- 1000
dist.perm <- rep(0,R)
for(i in 1:R){
  all.perm <- sample(all.orig,size=2*n,replace = FALSE)
  x.perm <- all.perm[1:n]
  y.perm <- all.perm[(n+1):(2*n)]
  dist.perm[i] <- abs(mean(x.perm)-mean(y.perm))
}

greater <- ifelse(dist.perm>=dist.orig,1,0)
p.value <- mean(greater)
p.value
```

- yields

```
[1] 0.49
```

- In this case the test does not reject H_0 (as it should).
- Running the whole procedure many times yields a wrong decision in about 5% of the cases (see R-Code), i.e. the type I error of the permutation test is close to the one of the t-test, who was designed exclusively for this setting.
- The permutation test itself did not use any assumptions on the underlying distributions - it also works in the general setting.
- The next exercises aim at verifying this assertion.

Exercise 34:

- Develop a permutation test for testing (ii) $H_0 : \mu_x \leq \mu_y$ versus $H_1 : \mu_x > \mu_y$.
- Compare its performance with the standard t-test.

- We solve Exercise 34 together.
- Suppose that $X \sim \mathcal{N}(\mu_x, \sigma^2)$, we do not know μ_x and σ^2.
- Suppose that $Y \sim \mathcal{N}(\mu_y, \sigma^2)$, we do not know μ_y and σ^2.
- Notice that the variance of X and Y is the same (and unknown).
- Given a sample $X_1, \ldots, X_n$ from X and a sample $Y_1, \ldots, Y_m$ from Y with $m, n \geq 2$ we want to test the following hypothesis concerning $\mu_D := \mu_x - \mu_y$:
- (ii) $H_0 : \mu_D \leq 0$ versus $H_1 : \mu_D > 0$
- Remember that the test statistic $T_{n,m}$ is given by

$$T_{n,m} = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{S_p^2}\sqrt{\frac{1}{n} + \frac{1}{m}}},$$

where S_p^2 is the pooled sample variance and given by

$$S_p^2 = \frac{(n-1)S_n^2 + (m-1)S_m^2}{n+m-2}$$

- Last time we stated 'Under H_0 the test statistic $T_{n,m}$ has t_{n+m-2}-distribution'.
- Is this still true in the current setting?
- Given concrete observations we will reject H_0 at level α if and only if $T_{n,m} > t_{n+m-2,1-\alpha}$, where $t_{n+m-2,1-\alpha}$ denotes the $1-\alpha$-quantile of the t-distribution with $n+m-2$ degrees of freedom.
- Where does this come from?
- The p-value of this test is given by

$$p.value = \mathbb{P}(T \geq T_{n,m})$$

  where $T_{n,m}$ is the value of the test statistic and $T \sim t_{n+m-2}$.
- **NB: The test only does what it should if X and Y have normal distribution!** (normality has to be checked in advance).
- Let's compare our manually calculated values with the output of **t.test**:

One-sided unpaired two-sample t-test

```r
# t-test for H0: muD=mux-muy<=0
mux <- 0
muy <- 0.2
sigmax <- sigmay <- 1
n <- 50
x <- rnorm(n,mean=mux,sd=sigmax)
y <- rnorm(n,mean=muy,sd=sigmay)
test <- t.test(x,y,paired=FALSE,alternative="greater")
test

#t.test zu fuss
sp <- ((n-1)*var(x)+(n-1)*var(y))/(2*n-2)
test.stat <- (mean(x)-mean(y))/(sqrt(sp *(1/n+1/n)))
test.stat
1-pt(test.stat,df=2*n-2)
```

One-sided unpaired two-sample t-test

- Check the type I error via simulations.
-
```
#systematic:
R <- 10000
error <- rep(0,R)
for(i in 1:R){
  mux <- 0
  muy <- 0.2
  sigmax <- sigmay <- 1
  n <- 50
  x <- rnorm(n,mean=mux,sd=sigmax)
  y <- rnorm(n,mean=muy,sd=sigmay)
  test <- t.test(x,y,paired=FALSE,alternative="greater")
  test
  if(test$p.value<0.05){error[i] <- 1}
}
mean(error)
```
- yields
-
```
[1] 0.0029
```
- Why is this value not close to 0.05?
- What happens if we keep μ_x fixed and consider other values of μ_y?

One-sided unpaired two-sample t-test

- The bigger $\mu_y - \mu_x$ the smaller the type I error.
- Which role does the significance level α play?
- $H_0 : \mu_x \leq \mu_y$ is composite (i.e. contains many parameter constellations) - for each constellation $\mu_x \leq \mu_y$ the type I error is **at most** α.
- The type I error is exactly α if (and only if) $\mu_x = \mu_y$ (worst case scenario).
- Verify this with simulations.

Permutation test for the one-sided situation $H_0 : \mu_D = \mu_x - \mu_y \leq 0$

- As in the two-sided case let's develop a permutation test that does not build upon normality assumptions.
- We want to test $\mu_D = \mu_x - \mu_y \leq 0$ given observations $X_1, \ldots, X_n$ from X and $Y_1, \ldots, Y_m$ from Y.
- Naive idea: If $H_0 : \mu_D \leq 0$ holds, then 'mixing' the observations should bring the means closer together, the difference of the resulting means $\overline{X}'_n - \overline{Y}'_m$ should not be smaller than the original difference $\overline{X}_n - \overline{Y}_m$.
- Main idea of the permutation test in this situation is to 'mix' the observations many times and then check in how many cases $\overline{X}'_n - \overline{Y}'_m > \overline{X}_n - \overline{Y}_m$ holds.
- If the last inequality holds in less than 5% of the cases reject H_0.
- Let's check the details directly in R.

Permutation test for the one-sided situation $H_0 : \mu_D = \mu_x - \mu_y \leq 0$

```
#permutation test for mu_D<=0
mux <- 0; muy <- 0.2
n <- 50; sigma <- 1
x.orig <- rnorm(n,mean=mux,sd=sigma)
y.orig <- rnorm(n,mean=muy,sd=sigma)
diff.orig <- mean(x.orig)-mean(y.orig)
all.orig <- c(x.orig,y.orig)

test <- t.test(x.orig,y.orig,paired=FALSE,alternative="greater")

R <- 1000
diff.perm <- rep(0,R)
for(i in 1:R){
  all.perm <- sample(all.orig,size=2*n,replace = FALSE)
  x.perm <- all.perm[1:n]
  y.perm <- all.perm[(n+1):(2*n)]
  diff.perm[i] <- mean(x.perm)-mean(y.perm)
}

greater <- ifelse(diff.perm>=diff.orig,1,0)
p.value <- mean(greater)
p.value
[1] 0.885
test$p.value
[1] 0.8996981
```

Permutation test for the one-sided situation $H_0 : \mu_D = \mu_x - \mu_y \leq 0$

```
outer.R <- 500
outer.tp <- rep(0,outer.R)
outer.p <- rep(0,outer.R)
for(j in 1:outer.R){
  mux <- 0; muy <- 0.2
  n <- 50; sigma <- 1
  x.orig <- rnorm(n,mean=mux,sd=sigma)
  y.orig <- rnorm(n,mean=muy,sd=sigma)
  diff.orig <- mean(x.orig)-mean(y.orig)
  all.orig <- c(x.orig,y.orig)
  outer.tp[j] <- t.test(x.orig,y.orig,paired=FALSE,alternative="
      greater")$p.value

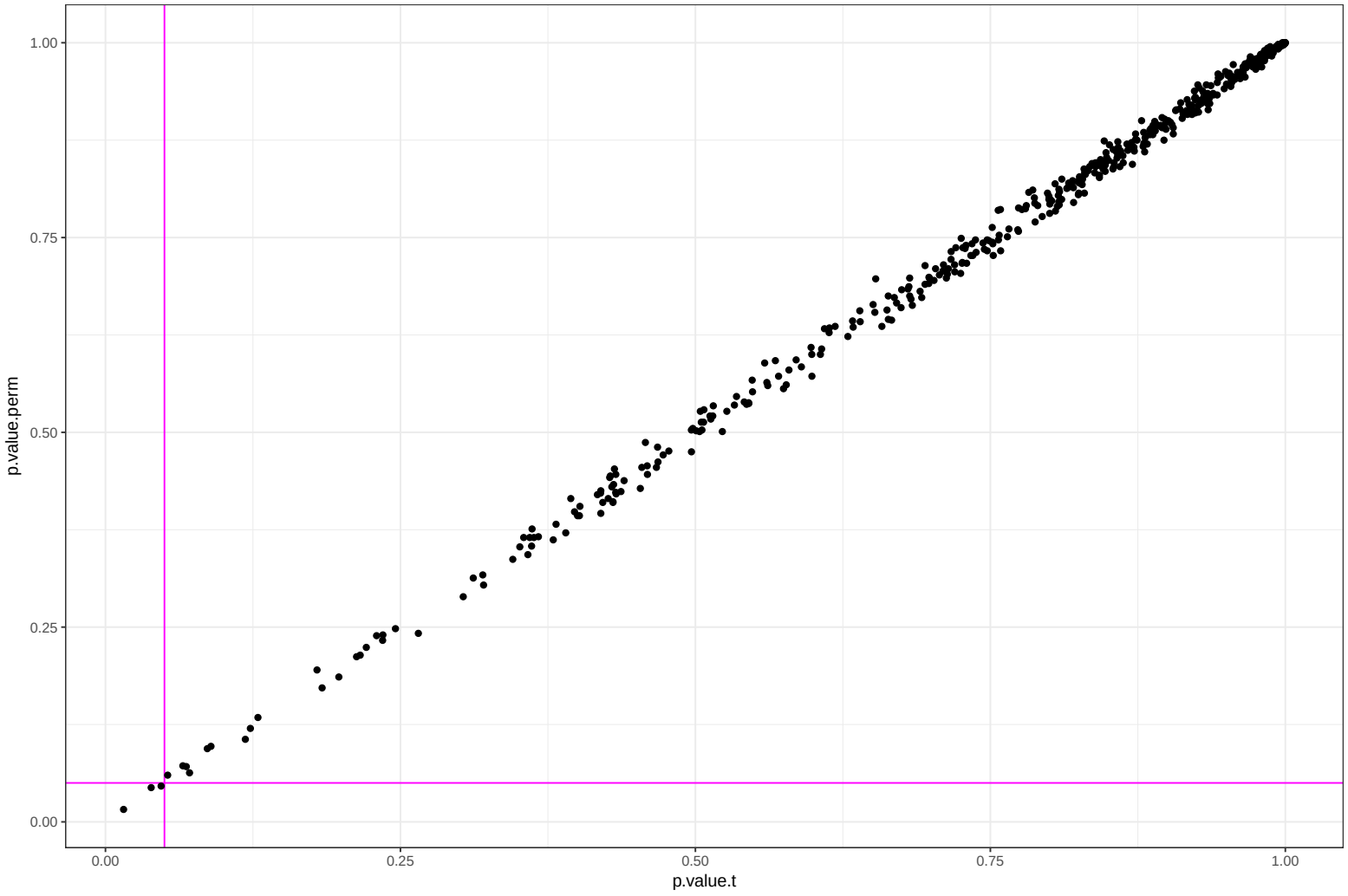
  R <- 1000
  dist.perm <- rep(0,R)
  for(i in 1:R){
    all.perm <- sample(all.orig,size=2*n,replace = FALSE)
    x.perm <- all.perm[1:n]
    y.perm <- all.perm[(n+1):(2*n)]
    diff.perm[i] <- mean(x.perm)-mean(y.perm)
  }
  greater <- ifelse(diff.perm>=diff.orig,1,0)
  p.value <- mean(greater)
  outer.p[j] <- p.value
}
```

Permutation test for the one-sided situation $H_0 : \mu_D = \mu_x - \mu_y \leq 0$

- Yields
- ```
  reject.rate.t <- mean(ifelse(outer.tp<0.05,1,0))
  reject.rate.t
  [1] 0.006

  reject.rate <- mean(ifelse(outer.p<0.05,1,0))
  reject.rate
  [1] 0.006
  ```
- Lucky coincidence?
- Let's compare the p-values of the t.test and the permutation test directly in a graphic.
- What can be seen?
- What happens if we change the value of μ_y?
  ```

Permutation test for the one-sided situation $H_0 : \mu_D = \mu_x - \mu_y \leq 0$



Exercise 35:

- Develop a permutation test for testing (iii) $H_0 : \mu_x \geq \mu_y$ versus $H_1 : \mu_x < \mu_y$.
- Compare its type I error with the type I error of the standard t-test.

Exercise 36:

- We have seen that the permutation test performs well w.r.t.t. type I error when applied to situations where the t-test can be applied.
- Return to the two-sided situation (i) $H_0 : \mu_x = \mu_y$ and compare the power of the t-test and the power of the corresponding permutation test.

Exercise 37:

- Return to the two-sided situation (i) $H_0 : \mu_x = \mu_y$.
- The permutation test was based on the comparison of $|\overline{X}'_n - \overline{Y}'_m|$ and $|\overline{X}_n - \overline{Y}_m|$.
- Another natural choice would be to compare the absolute difference of the medians.
- Find out with simulations if the resulting permutation test performs equally well.

Exercise 38 (for mathematicians):

- Suppose that $X \sim Ex(\theta_x)$ and $Y \sim Ex(\theta_y)$.
- Suppose that $X_1, \ldots, X_n$ are samples from X and $Y_1, \ldots, Y_m$ samples from Y.
- Develop a permutation test for testing $H_0 : \theta_x = \theta_y$ versus $H_1 : \theta_x \neq \theta_y$.
- Compare its type I error with the standard t-test (ignoring that it is not applicable in this situation).