

# Angewandte (Mathematische) Statistik (405.330)

## Elementary toy example hypothesis testing<sup>1</sup>

**Univ.-Prof. Dr. Wolfgang Trutschnig**

Research group for Stochastics/Statistics

Department for Mathematics

University Salzburg

[www.trutschnig.net](http://www.trutschnig.net)

Salzburg, May 2022

---

<sup>1</sup>Slides originally created for an introductory statistics course for non-mathematicians



## Example (Elementary 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

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

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

- ▶ 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

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

- ▶ If  $H_1$  is correct then we have

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

- ▶ 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  $\alpha$ :

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

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



## Example (Toy example hypothesis testing, cont.)

- ▶ @type II error  $\beta$ :

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

- ▶ 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  $\alpha = 1$ .
- ▶  $\alpha$  and  $\beta$  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  $\alpha$  corresponds to the situation that the defendant will be declared guilty although he is innocent.
- ▶ The afore-mentioned type II error  $\beta$  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 is the term 'Beyond reasonable doubt' underlining this fact.
- ▶ In other words: We want to keep the type I error  $\alpha$  (very) small.
- ▶ The same applies to hypothesis testing:  $\alpha$  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  $\alpha$  is fixed it is the statisticians' job to develop optimal tests, i.e. decision rules (criteria) with a probability of (at most)  $\alpha$  for a type I error and, at the same time, minimal type II error  $\beta$ .





## 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  $x_1, x_2, \dots, x_n$  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

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

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

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

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  $x_1, \dots, x_n$ :
- ▶ 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

```
1 #determine the threshold for the test  $H_0: p=2/3$  versus  $H_1: p=1/3$ 
2 plot(0:100, pbinom(0:100, size=100, prob=2/3), type="p")
3 abline(h=0.05)
4
5 t<-qbinom(p=0.05, size=100, prob=2/3)-1
6 t
7 [1] 58
8
9 pbinom(t, size = 100, prob=2/3)
10 [1] 0.04337149
11
12 #calculate beta
13 1-pbinom(t, size=100, prob=1/3)
14 [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).

```

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

1
2 # R=10000 runs under H0
3 R<-10000
4 reject<-rep(0,R)
5 for(i in 1:R){
6   x<-sample(c(1,0),size=n,replace = TRUE,prob=c(2/3,1/3))
7   if(length(x[x==1])<=58){reject[i]<-1}
8 }
9 mean(reject)
10 [1] 0.0445
11
12 barplot(table(reject))

```



## Example (Toy example hypothesis testing, cont.)

- ▶ Simulations for the type II error.

```
1 # R=10000 runs under H1
2 R<-100000
3 false<-rep(0,R)
4 for(i in 1:R){
5   x<-sample(c(1,0),size=n,replace = TRUE,prob=c(1/3,2/3))
6   if(length(x[x==1])>=58){false[i]<-1}
7 }
8 mean(false)
9 [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 41:

- ▶ 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  $\alpha$  and (more importantly) on  $\beta$ .



## 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  $x_1, \dots, x_n$  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  $x_1, \dots, x_n$  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  $\alpha$  and  $\beta$  can not be minimized simultaneously, so  $\alpha$  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

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

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  $\beta$  and set

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

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



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

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

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

```

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

```

- ▶ yields

- ▶ Exact **binomial** test

```

2
3 data: successes and n
4 number of successes = 61, number of trials = 100, p-value =
  0.9895
5 alternative hypothesis: true probability of success is less than
  0.5
6 95 percent confidence interval:
7 0.0000000 0.6918993
8 sample estimates:
9 probability of success
10 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.



- ▶ We analyze the performance of binom.test via simulations

```
1 #assume that H0 holds
2 #repeat the above procedure R=10000 times and calculate the
  portion of false decisions (type I error)
3 R <- 10000
4 error <- rep(0,R)
5 for(i in 1:R){
6   p <- 0.6
7   n <- 100
8   x <- sample(c(0,1), size=n, replace=TRUE, prob=c(1-p,p))
9   successes <- sum(x)
10  test <- binom.test(successes, n, p=0.5, alternative="less")
11  if(test$p.value < 0.05){error[i] <- 1}
12 }
13 mean(error)
```

- ▶ yields

```
1 [1] 0.0036
```



```
1 #worst case scenario (what is different to before?)
2 R <- 10000
3 error <- rep(0,R)
4 for(i in 1:R){
5   p <- 0.5
6   n <- 100
7   x <- sample(c(0,1), size=n, replace=TRUE, prob=c(1-p,p))
8   successes <- sum(x)
9   test <- binom.test(successes, n, p=0.5, alternative="less")
10  if(test$p.value < 0.05){error[i] <- 1}
11 }
12 mean(error)
```

► yields

```
1 [1] 0.0441
```



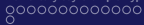
```

1  ##@power: choose different values for p in H1 and calculate the
   power
2  pgrid <- seq(0,0.5,by=0.05)
3  power <- rep(0,length(pgrid))
4  for(j in 1:length(pgrid)){
5    print(j)
6    R <- 5000
7    error <- rep(0,R)
8    for(i in 1:R){
9      p <- pgrid[j]
10     n <- 100
11     x <- sample(c(0,1),size=n,replace=TRUE,prob=c(1-p,p))
12     successes <- sum(x)
13     test <- binom.test(successes,n,p=0.5,alternative="less")
14     if(test$p.value >=0.05){error[i] <- 1}           #type II error
15   }
16   power[j] <- 1 - mean(error)
17 }
18 power
19 [1] 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 0.9998 0.9920 0.9134
    0.6220 0.2532 0.0474

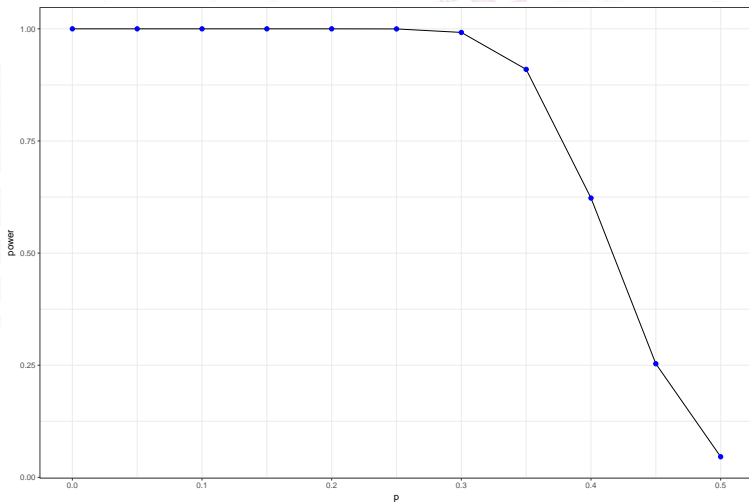
```







## Checking the performance of binom.test

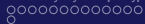


```

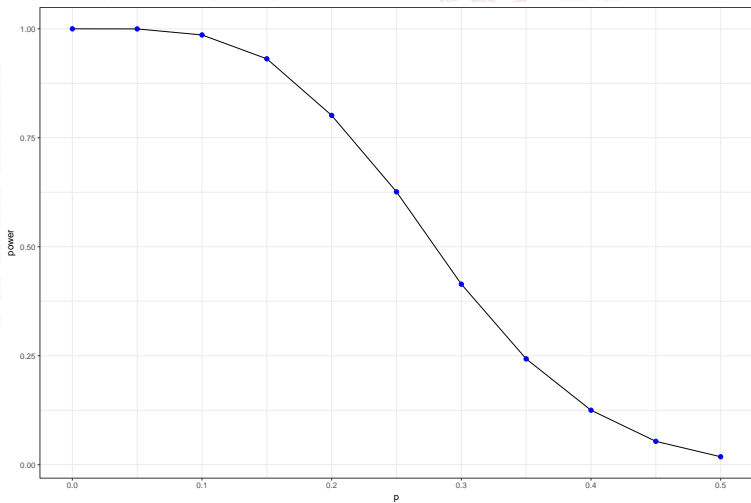
1 #@power: same for smaller sample size n
2 pgrid <- seq(0,0.5,by=0.05)
3 power <- rep(0,length(pgrid))
4 for(j in 1:length(pgrid)){
5   print(j)
6   R <- 5000
7   error <- rep(0,R)
8   for(i in 1:R){
9     p <- pgrid[j]
10    n <- 500
11    x <- sample(c(0,1), size=n, replace=TRUE, prob=c(1-p,p))
12    successes <- sum(x)
13    test <- binom.test(successes,n,p=0.5, alternative="less")
14    if(test$p.value >=0.05){error[i] <- 1}
15  }
16  power[j] <- 1 - mean(error)
17 }
18 power
19 [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 42:

- ▶ 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  $\pi$  in each case.
- ▶ How can the results be interpreted?



### Exercise 43:

- ▶ 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  $\pi$  in each case
- ▶ How can the results be interpreted?

