Universitetet i Bergen
STAT200 - Applied Statistics

Summer 2019

Assignment 2 - Testing

In this assignment, we will deal with concepts that allow us to take decisions in situations requiring hypothesis testing. These occur regularly when verifying claims about a population by means of a sample, for example:

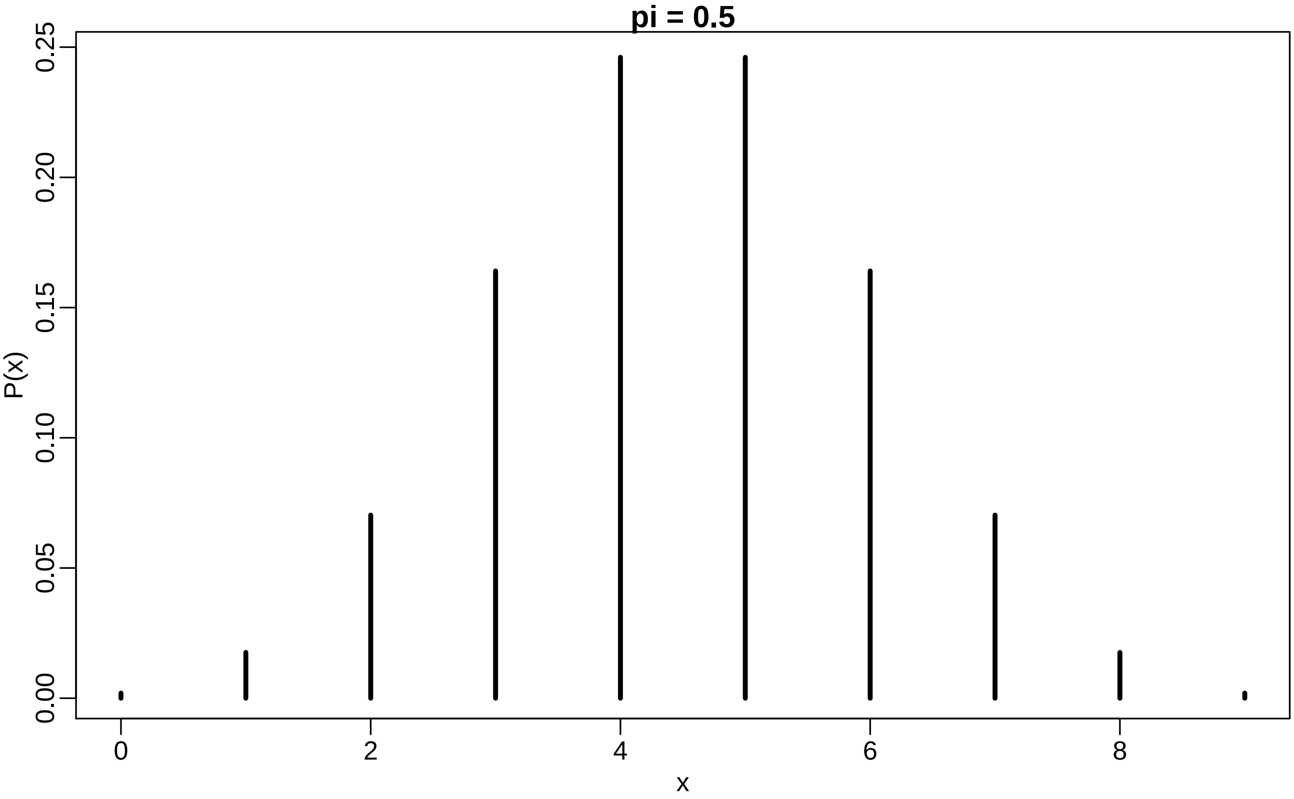
- A new drug reduces the risk of heart attack
- The climate has changed in the last 10 years
- The market share of a certain product has increased

Exercise 1

Main new R functions: `read.csv()`, `t.test()`, `shapiro.test()`, `read.table()`.

1. As you have already seen in introductory courses of statistics, it is possible to construct a hypothesis test on the mean μ with unknown variance σ^2, assuming that the population follows a Gaussian law. What is the name of the test, what is the name of the function for carrying out this test in R, and what are the options?
2. Let's analyse the dataset 'Barley', which contains barley yields from a piece of land you intend to sell, and we assume that the sample follows a Gaussian distribution. In order to sell the land at a good price, you want to prove that the average yield is greater than 150 grams. Formulate the hypothesis and the alternative, and run the test in R. Is the hypothesis rejected at the level of $\alpha = 0.1$? And what happens when setting $\alpha = 0.05$?
3. Now a different assumption: let H_0 correspond to the case that the average yield equals 147. Test this hypothesis. Can you reject it for a given level of $\alpha = 10\%$?
4. Determine the confidence interval at 99% (99.9%) for the sample mean, it follows directly from the output of your test. The value 147, does it fall inside the confidence interval? What do you think is the relationship between the confidence interval and the level of significance of your test?
5. We change to a different data set. Suppose you a contacted by a friend from social sciences, who carried out a survey by handing out a questionnaire to 100 male and 100 female participants. Among other things, the participants have been asked to report the number of partners they have been in an intimate sexual relationship with. Your friend asks you for help with the statistical analysis, and has not particular hypothesis.
   The data set 'Partners.dat' contains the results. Inspect the data, import it into R using the `read.table()`-function, and use `attach()` for an easier access to the variables contained in the data set. Then, it is up to you do decide what to do. What are your conclusions?

Figure 1: Probability mass function for the number of votes



Exercise 2

Main new R functions: `dbinom()`, `par(mfrow = c(a, b))`, `plot()`, `lines()`, `legend()`.

In this exercise we consider presidential elections in New Jersey. Assume that one of the candidates, for example George Bush, said before the elections that he has the support of at least 50% of voters. This also corresponds to the experimental hypothesis. Therefore, in this case the unknown parameter of the population is the proportion of support π, and the statistical null hypothesis can be written as

$$H_0 : \pi < 0.5 \text{ (in general: } \pi < \pi_0)$$

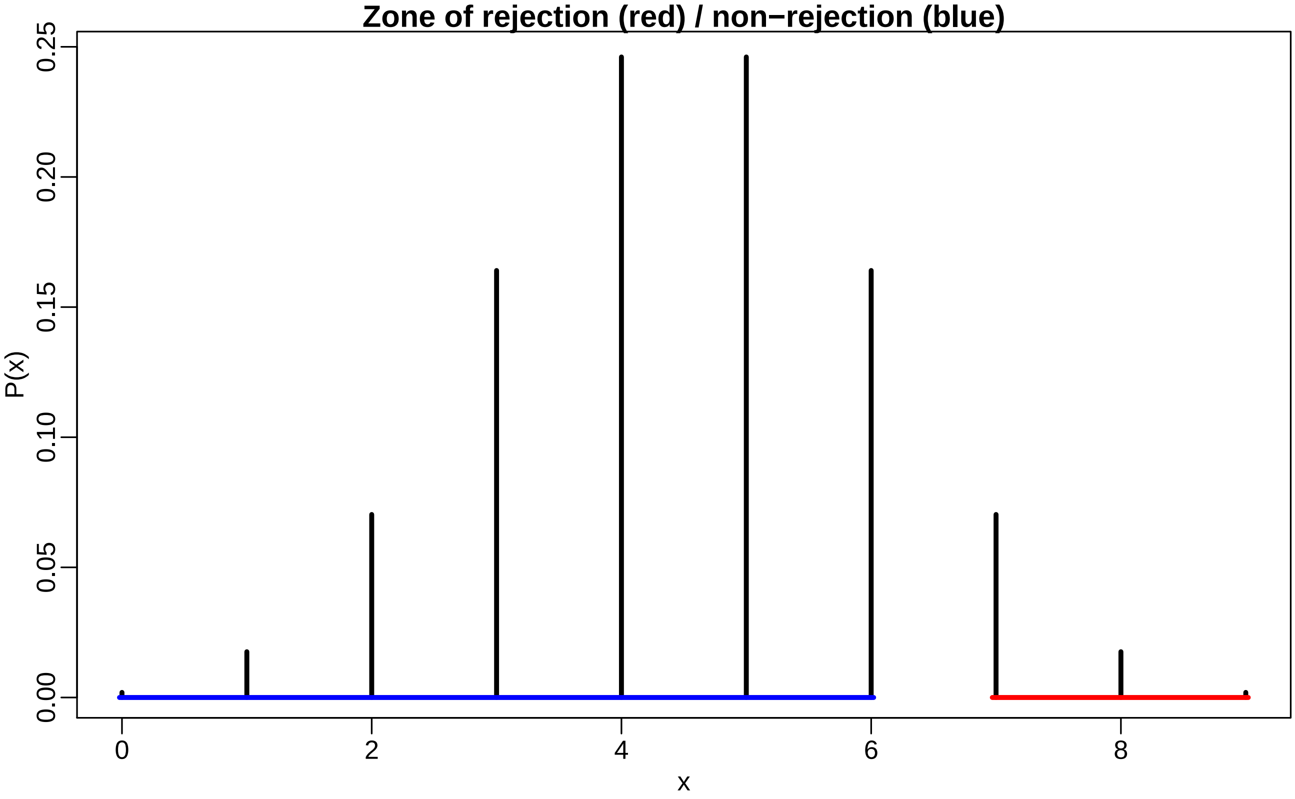
Consequently, the alternative hypothesis equals

$$H_1 : \pi \geq 0.5 \text{ (in general: } \pi \geq \pi_0)$$

1. Suppose we interviewed only a (very) small sample of 9 voters at random before the elections in order to verify the claims of the candidate by means of the hypothesis and the alternative stated above. Naturally, in a real situation one would chose much bigger sample. Anyhow, in this experiment π corresponds to the true proportion of Bush's voters, which is unknown. Moreover, X is a random variable counting the number of voters for Bush in our sample. What distribution does X follow? Note the formula of the density. Then, represent the distribution graphically for $\pi = 0.1$ (0.5, 0.9).
   Remark: For $\pi = 0.5$, you should obtain a result looking approximately like Figure 1.

2. If the hypothesis is true, larger values of X are unlikely (why?). Thus, we should reject the hypothesis if X takes large values. For illustrative purposes, in this exercise we do not define

Figure 2: Zone of rejection and probability mass function for given $\pi = 0.5$



our level of significance (criterion) α first (which should usually be done), but simply define a zone of rejection directly. Suppose, for example, that the set $A = (7, 8, 9)$ constitutes the zone of rejection of H_0.

(a) Calculate the probability of (not) rejecting the hypothesis using the rejection zone A. What is the probability of an error of type I for this choice of A?

(b) Add a red horizontal line showing the zone of rejection to the graph shown in Figure 1, and a blue line for the non-rejection zone (example in Figure 2).

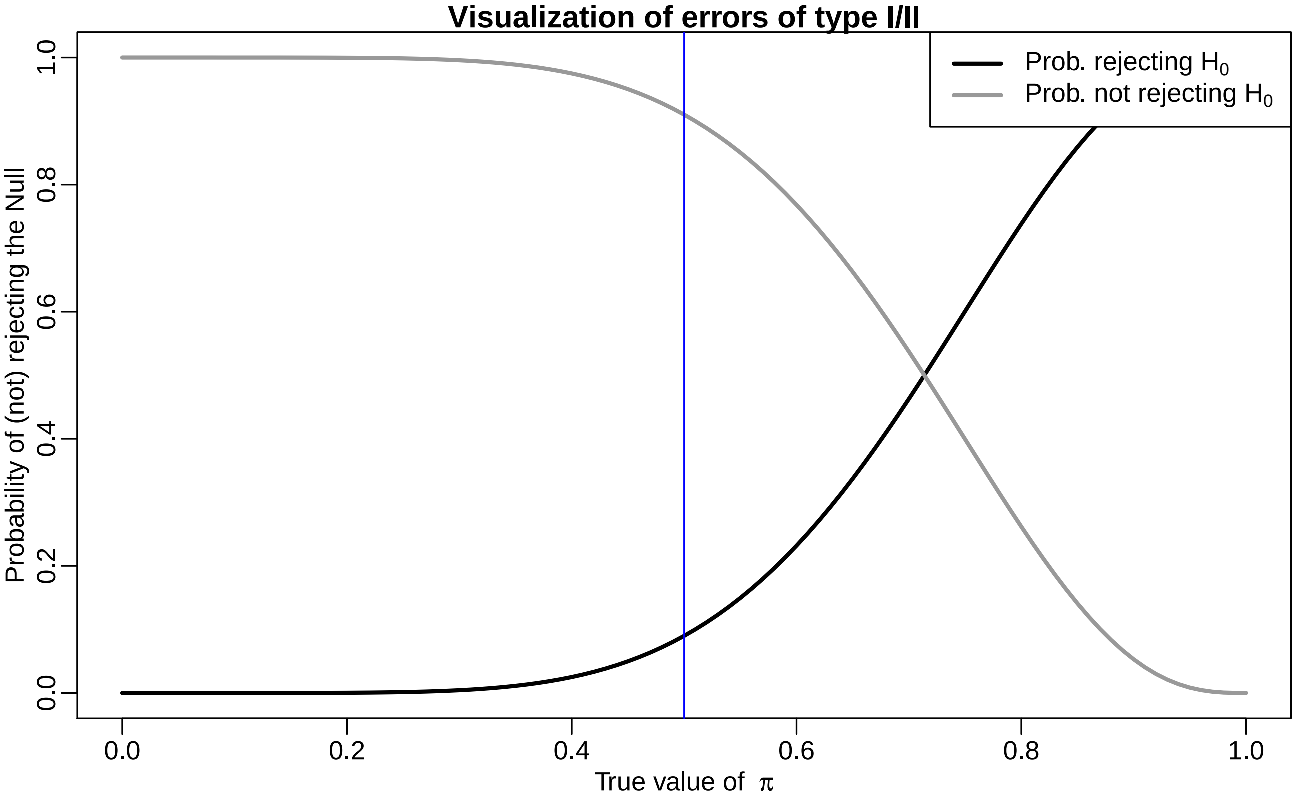
(c) Suppose you observe numbers of votes equalling $x \in 0, 1, ..., 9$. Calculate the corresponding p-value for each each value of x.

3. Lets now consider a different error, the error type II. Recall the definition of this type of error and calculate the probability of such an error if the true parameter π equals 50%. Interpret the result.

4. The values of 49.99% and 50% correspond to only two specific cases for the value of the parameter π. What is happening for other values? Complete the table below and specify the probabilities of error of type I and II, respectively.

π	H_0 true/false	P(reject H_0)	P(not reject H_0)
0.1	.	.	.
0.2	.	.	.
$\vdots$	$\vdots$	$\vdots$	$\vdots$
0.9	.	.	.

Figure 3: Errors of type I and II, given rejection zone A



Note that it is possible to use a loop here to shorten the program (but this is not necessary). A loop in R works as follows:

```
x = c()
for (i in 1 : 10) {
  x[i] = 2 * sqrt(i)
}
```

5. To finish the study of errors of type I and II, draw a graph corresponding to the one shown in Figure 3 and interpret it.

Exercise 3

Main new R functions: `qbinom()`, `pbinom()`.

In this exercise a more realistic setting is considered. Before the presidential elections in the United States in 2000, the Eagleton Institute of Politics of the Rutgers State University of New Jersey conducted a survey in New Jersey to determine the proportion of voters of George Bush and Al Gore. For this survey, 356 randomly selected individuals were interviewed and asked for which candidate they would vote. The result was 201 votes for Bush and 155 for Al Gore.

1. Now we proceed the usual way, and start by determining our criterion (level) α. For which level(s) of α would you reject the hypothesis H_0 ($\pi < 0.5$) in this context? Recall that the most common values for alpha are 5%, 1% and 0.1%.

2. How many votes should Bush receive at least in order (not) to reject H_0 a level of 0.1%, 1%, and 5%, respectively?

3. Furthermore, redraw the graph shown in Figure 3 to analyse the errors of type I/II, supposing that you fix α at 5%. What do you notice when comparing the two graphs to each other?