25

Testing hypotheses: essentials

The statistical methods that we have discussed until now have been developed to infer knowledge about certain features of the model distribution that represent our quantities of interest. These inferences often take the form of numerical estimates, as either single numbers or confidence intervals. However, sometimes the conclusion to be drawn is *not* expressed numerically, but is concerned with choosing between two conflicting theories, or *hypotheses*. For instance, one has to assess whether the lifetime of a certain type of ball bearing deviates or does not deviate from the lifetime guaranteed by the manufacturer of the bearings; an engineer wants to know whether dry drilling is faster or the same as wet drilling; a gynecologist wants to find out whether smoking affects or does not affect the probability of getting pregnant; the Allied Forces want to know whether the German war production is equal to or smaller than what Allied intelligence agencies reported. The process of formulating the possible conclusions one can draw from an experiment and choosing between two alternatives is known as *hypothesis testing*. In this chapter we start to explore this statistical methodology.

25.1 Null hypothesis and test statistic

We will introduce the basic concepts of hypothesis testing with an example. Let us return to the analysis of German war equipment. During World War II the Allied Forces received reports by the Allied intelligence agencies on German war production. The numbers of produced tires, tanks, and other equipment, as claimed in these reports, were a lot higher than indicated by the observed serial numbers. The objective was to decide whether the actual produced quantities were smaller than the ones reported.

For simplicity suppose that we have observed tanks with (recoded) serial numbers

$$61 \quad 19 \quad 56 \quad 24 \quad 16.$$

Furthermore, suppose that the Allied intelligence agencies report a production of 350 tanks.[1] This is a lot more than we would surmise from the observed data. We want to choose between the proposition that the total number of tanks is 350 and the proposition that the total number is smaller than 350. The two competing propositions are called *null hypothesis*, denoted by H_0, and *alternative hypothesis*, denoted by H_1. The way we go about choosing between H_0 and H_1 is conceptually similar to the way a jury deliberates in a court trial. The null hypothesis corresponds to the position of the defendant: just as he is presumed to be innocent until proven guilty, so is the null hypothesis presumed to be true until the data provide convincing evidence against it. The alternative hypothesis corresponds to the charges brought against the defendant.

To decide whether H_0 is false we use a statistical model. As argued in Chapter 20 the (recoded) serial numbers are modeled as a realization of random variables $X_1, X_2, \ldots, X_5$ representing five draws *without replacement* from the numbers $1, 2, \ldots, N$. The parameter N represents the total number of tanks. The two hypotheses in question are

$$H_0 : N = 350$$
$$H_1 : N < 350.$$

If we reject the null hypothesis we will accept H_1; we speak of *rejecting* H_0 *in favor of* H_1. Usually, the alternative hypothesis represents the theory or belief that we would like to accept if we do reject H_0. This means that we must carefully choose H_1 in relation with our interests in the problem at hand. In our example we are particularly interested in whether the number of tanks is *less* than 350; so we test the null hypothesis against $H_1 : N < 350$. If we would be interested in whether the number of tanks *differs* from 350, or is *greater* than 350, we would test against $H_1 : N \neq 350$ or $H_1 : N > 350$.

Quick exercise 25.1 In the drilling example from Sections 15.5 and 16.4 the data on drill times for dry drilling are modeled as a realization of a random sample from a distribution with expectation μ_1, and similarly the data for wet drilling correspond to a distribution with expectation μ_2. We want to know whether dry drilling is faster than wet drilling. To this end we test the null hypothesis $H_0 : \mu_1 = \mu_2$ (the drill time is the same for both methods). What would you choose for H_1?

The next step is to select a criterion based on $X_1, X_2, \ldots, X_5$ that provides an indication about whether H_0 is false. Such a criterion involves a test statistic.

[1] This may seem ridiculous. However, when after the war official German production statistics became available, the average monthly production of tanks during the period 1940–1943 was 342. During the war this number was estimated at 327, whereas Allied intelligence reported 1550! (see [27]).

TEST STATISTIC. Suppose the dataset is modeled as the realization of random variables $X_1, X_2, \ldots, X_n$. A *test statistic* is any sample statistic $T = h(X_1, X_2, \ldots, X_n)$, whose numerical value is used to decide whether we reject H_0.

In the tank example we use the test statistic

$$T = \max\{X_1, X_2, \ldots, X_5\}.$$

Having chosen a test statistic T, we investigate what sort of values T can attain. These values can be viewed on a credibility scale for H_0, and we must determine which of these values provide evidence in favor of H_0, and which provide evidence in favor of H_1. First of all note that if we find a value of T larger than 350, we immediately know that H_0 as well as H_1 is false. If this happens, we actually should be considering another testing problem, but for the current problem of testing $H_0 : N = 350$ against $H_1 : N < 350$ such values are irrelevant. Hence the possible values of T that are of interest to us are the integers from 5 to 350.

If H_0 is true, then what is a typical value for T and what is not? Remember from Section 20.1 that, because $n = 5$, the expectation of T is $\mathrm{E}[T] = \frac{5}{6}(N+1)$. This means that the distribution of T is centered around $\frac{5}{6}(N+1)$. Hence, if H_0 is true, then typical values of T are in the neighborhood of $\frac{5}{6} \cdot 351 = 292.5$. Values of T that deviate a lot from 292.5 are evidence *against* H_0. Values that are much greater than 292.5 are evidence against H_0 but provide even stronger evidence against H_1. For such values we will *not reject* H_0 in favor of H_1. Also values a little smaller than 292.5 are grounds *not to reject* H_0, because we are committed to giving H_0 the benefit of the doubt. On the other hand, values of T very close to 5 should be considered as strong evidence *against* the null hypothesis and are *in favor* of H_1, hence they lead to a decision to *reject* H_0. This is summarized in Figure 25.1.

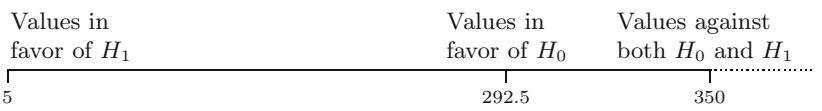

Fig. 25.1. Values of the test statistic T.

QUICK EXERCISE 25.2 Another possible test statistic would be $\bar{X}_5$. If we use its values as a credibility scale for H_0, then what are the possible values of $\bar{X}_5$, which values of $\bar{X}_5$ are in favor of $H_1 : N < 350$, and which values are in favor of $H_0 : N = 350$?

For the data we find

$$t = \max\{61, 19, 56, 24, 16\} = 61$$

as the realization of the test statistic. How do we use this to decide on H_0?

25.2 Tail probabilities

As we have just seen, if H_0 is true, then typical values of T are in the neighborhood of $\frac{5}{6} \cdot 351 = 292.5$. In view of Figure 25.1, the more a value of T is to the left, the stronger evidence it provides in favor of H_1. The value 61 is in the left region of Figure 25.1. Can we now reject H_0 and conclude that N is smaller than 350, or can the fact that we observe 61 as maximum be attributed to chance? In courtroom terminology: can we reach the conclusion that the null hypothesis is *false beyond reasonable doubt*? One way to investigate this is to examine how likely it is that one would observe a value of T that provides even stronger evidence against H_0 than 61, in the situation that $N = 350$. If this is very unlikely, then 61 already bears strong evidence against H_0.

Values of T that provide stronger evidence against H_0 than 61 are to the left of 61. Therefore we compute $\mathrm{P}(T \leq 61)$. In the situation that $N = 350$, the test statistic T is the maximum of 5 numbers drawn without replacement from $1, 2, \ldots, 350$. We find that

$$\begin{aligned} \mathrm{P}(T \leq 61) &= \mathrm{P}(\max\{X_1, X_2, \ldots, X_5\} \leq 61) \\ &= \frac{61}{350} \cdot \frac{60}{349} \cdots \frac{57}{346} = 0.00014. \end{aligned}$$

This probability is so small that we view the value 61 as strong evidence against the null hypothesis. Indeed, if the null hypothesis would be true, then values of T that would provide the same or even stronger evidence against H_0 than 61 are *very unlikely* to occur, i.e., they occur with probability 0.00014! In other words, the observed value 61 is *exceptionally small* in case H_0 is true.

At this point we can do two things: either we believe that H_0 is true and that something very unlikely has happened, or we believe that events with such a small probability do not happen in practice, so that $T \leq 61$ could only have occurred because H_0 is false. We choose to believe that things happening with probability 0.00014 are so exceptional that we *reject* the null hypothesis $H_0 : N = 350$ in favor of the alternative hypothesis $H_1 : N < 350$. In courtroom terminology: we say that a value of T smaller than or equal to 61 implies that the null hypothesis is false *beyond reasonable doubt*.

P-values

In our example, the more a value of T is to the left, the stronger evidence it provides against H_0. For this reason we computed the *left tail probability*

$\mathrm{P}(T \leq 61)$. In other situations, the direction in which values of T provide stronger evidence against H_0 may be to the right of the observed value t, in which case one would compute a *right tail probability* $\mathrm{P}(T \geq t)$. In both cases the tail probability expresses how likely it is to obtain a value of the test statistic T *at least as extreme as* the value t observed for the data. Such a probability is called a p-value. In a way, the size of the p-value reflects how much evidence the observed value t provides against H_0. The *smaller* the p-value, the *stronger evidence* the observed value t bears against H_0.

The phrase "at least as extreme as the observed value t" refers to a particular direction, namely the direction in which values of T provide stronger evidence against H_0 and in favor of H_1. In our example, this was to the left of 61, and the p-value corresponding to 61 was $\mathrm{P}(T \leq 61) = 0.00014$. In this case it is clear what is meant by "at least as extreme as t" and which tail probability corresponds to the p-value. However, in some testing problems one can deviate from H_0 in *both* directions. In such cases it may not be clear what values of T are at least as extreme as the observed value, and it may be unclear how the p-value should be computed. One approach to a solution in this case is to simply compute the *one-tailed p-value* that corresponds to the direction in which t deviates from H_0.

Quick exercise 25.3 Suppose that the Allied intelligence agencies had reported a production of 80 tanks, so that we would test $H_0 : N = 80$ against $H_1 : N < 80$. Compute the p-value corresponding to 61. Would you conclude H_0 is false beyond reasonable doubt?

25.3 Type I and type II errors

Suppose that the maximum is 200 instead of 61. This is also to the left of the expected value 292.5 of T. Is it far enough to the left to reject the null hypothesis? In this case the p-value is equal to

$$\begin{aligned}\mathrm{P}(T \leq 200) &= \mathrm{P}(\max\{X_1, X_2, \ldots, X_5\} \leq 200) \\ &= \frac{200}{350} \cdot \frac{199}{349} \cdots \frac{196}{346} = 0.0596.\end{aligned}$$

This means that *if* the total number of produced tanks is 350, then in 5.96% of all cases we would observe a value of T that is at least as extreme as the value 200. Before we decide whether 0.0596 is small enough to reject the null hypothesis let us explore in more detail what the preceding probability stands for.

It is important to distinguish between (1) the *true state of nature*: H_0 is true or H_1 is true and (2) *our decision*: we reject or do not reject H_0 *on the basis of the data*. In our example the possibilities for the true state of nature are:

- H_0 is true, i.e., there are 350 tanks produced.
- H_1 is true, i.e., the number of tanks produced is less than 350.

We do not know in which situation we are. There are two possible decisions:

- We reject H_0 in favor of H_1.
- We do not reject H_0.

This leads to four possible situations, which are summarized in Figure 25.2.

		True state of nature	
		H_0 is true	H_1 is true
Our decision on the basis of the data	Reject H_0	*Type I error*	Correct decision
	Not reject H_0	Correct decision	*Type II error*

Fig. 25.2. Four situations when deciding about H_0.

There are two situations in which the decision made on the basis of the data is wrong. The null hypothesis H_0 may be true, whereas the data lead to rejection of H_0. On the other hand, the alternative hypothesis H_1 may be true, whereas we do not reject H_0 on the basis of the data. These wrong decisions are called type I and type II errors.

TYPE I AND II ERRORS. A *type I error* occurs if we falsely reject H_0. A *type II error* occurs if we falsely do not reject H_0.

In courtroom terminology, a type I error corresponds to convicting an innocent defendant, whereas a type II error corresponds to acquitting a criminal.

If $H_0 : N = 350$ is true, then the decision to reject H_0 is a type I error. We will never know whether we make a type I error. However, given a particular decision rule, we can say something about the probability of committing a type I error. Suppose the decision rule would be "reject $H_0 : N = 350$ whenever $T \leq 200$." With this decision rule the probability of committing a type I error is $\mathrm{P}(T \leq 200) = 0.0596$. If we are willing to run the risk of committing a type I error with probability 0.0596, we could adopt this decision rule. This would also mean that on the basis of an observed maximum of 200 we would reject H_0 in favor of $H_1 : N < 350$.

QUICK EXERCISE 25.4 Suppose we adopt the following decision rule about the null hypothesis: "reject $H_0 : N = 350$ whenever $T \leq 250$." Using this decision rule, what is the probability of committing a type I error?

The question remains what amount of risk one is willing to take to falsely reject H_0, or in courtroom terminology: how small should the p-value be to reach a conclusion that is "beyond reasonable doubt"? In many situations, as a rule of thumb 0.05 is used as the level where reasonable doubt begins. Something happening with probability less than or equal to 0.05 is then viewed as being too exceptional. However, there is no general rule that specifies how small the p-value must be to reject H_0. There is no way to argue that this probability *should be* below 0.10 or 0.18 or 0.009—or anything else.

A possible solution is to solely report the p-value corresponding to the observed value of the test statistic. This is objective and does not have the arbitrariness of a preselected level such as 0.05. An investigator who reports the p-value conveys the maximum amount of information contained in the dataset and permits all decision makers to choose their own level and make their own decision about the null hypothesis. This is especially important when there is no justifiable reason for preselecting a particular value for such a level.

25.4 Solutions to the quick exercises

25.1 One is interested in whether dry drilling is *faster* than wet drilling. Hence if we reject $H_0 : \mu_1 = \mu_2$, we would like to conclude that the drill time is *smaller* for dry drilling than for wet drilling. Since μ_1 and μ_2 represent the drill time for dry and wet drilling, we should choose $H_1 : \mu_1 < \mu_2$.

25.2 The value of $\bar{X}_5$ is at least 3 and if we find a value of $\bar{X}_5$ that is larger than 348, then at least one of the five numbers must be greater than 350, so that we immediately know that H_0 as well as H_1 is false. Hence the possible values of $\bar{X}_5$ that are relevant for our testing problem are between 3 and 348. We know from Section 20.1 that $2\bar{X}_5 - 1$ is an unbiased estimator for N, no matter what the value of N is. This implies that values of $\bar{X}_5$ itself are centered around $(N + 1)/2$. Hence values close to 351/2=175.5 are in favor of H_0, whereas values close to 3 are in favor of H_1. Values close to 348 are against H_0, but also against H_1. See Figure 25.3.

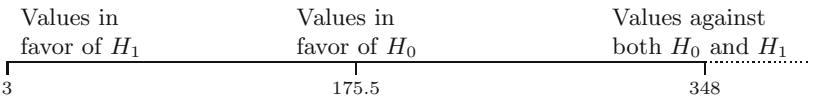

Fig. 25.3. Values of the test statistic $\bar{X}_5$.

25.3 The p-value corresponding to 61 is now equal to

$$\mathrm{P}(T \leq 61) = \tfrac{61}{80} \cdot \tfrac{60}{79} \cdots \tfrac{57}{76} = 0.2475.$$

If H_0 is true, then in 24.75% of the time one will observe a value T less than or equal to 61. Such values are not exceptionally small for T under H_0, and therefore the evidence that the value 61 bears against H_0 is pretty weak. We cannot reject H_0 beyond reasonable doubt.

25.4 The type I error associated with the decision rule occurs if $N = 350$ (H_0 is true) and $t \leq 250$ (reject H_0). The probability that this happens is $\mathrm{P}(T \leq 250) = \frac{250}{350} \cdot \frac{249}{349} \cdots \frac{246}{346} = 0.1838$.

25.5 Exercises

25.1 In a study about train delays in The Netherlands one was interested in whether arrival delays of trains exhibit more variation during rush hours than during quiet hours. The observed arrival delays during rush hours are modeled as realizations of a random sample from a distribution with variance σ_1^2, and similarly the observed arrival delays during quiet hours correspond to a distribution with variance σ_2^2. One tests the null hypothesis $H_0 : \sigma_1 = \sigma_2$. What do you choose as the alternative hypothesis?

25.2 ⊡ On average, the number of babies born in Cleveland, Ohio, in the month of September is 1472. On January 26, 1977, the city was immobilized by a blizzard. Nine months later, in September 1977, the recorded number of births was 1718. Can the increase of 246 be attributed to chance? To investigate this, the number of births in the month of September is modeled by a Poisson random variable with parameter μ, and we test $H_0 : \mu = 1472$. What would you choose as the alternative hypothesis?

25.3 Recall Exercise 17.9 about black cherry trees. The scatterplot of y (volume) versus $x = d^2h$ (squared diameter times height) seems to indicate that the regression line $y = \alpha + \beta x$ runs through the origin. One wants to investigate whether this is true by means of a testing problem. Formulate a null hypothesis and alternative hypothesis in terms of (one of) the parameters α and β.

25.4 ⊞ Consider the example from Section 4.4 about the number of cycles up to pregnancy of smoking and nonsmoking women. Suppose the observed number of cycles are modeled as realizations of random samples from geometric distributions. Let p_1 be the parameter of the geometric distribution corresponding to smoking women and p_2 be the parameter for the nonsmoking women. We are interested in whether p_1 is different from p_2, and we investigate this by testing $H_0 : p_1 = p_2$ against $H_1 : p_1 \neq p_2$.

a. If the data are as given in Exercise 17.5, what would you choose as a test statistic?

b. What would you choose as a test statistic, if you were given the extra knowledge as in Table 21.1?

c. Suppose we are interested in whether smoking women are less likely to get pregnant than nonsmoking women. What is the appropriate alternative hypothesis in this case?

25.5 ⊡ Suppose a dataset is a realization of a random sample $X_1, X_2, \ldots, X_n$ from a uniform distribution on $[0, \theta]$, for some (unknown) $\theta > 0$. We test $H_0 : \theta = 5$ versus $H_1 : \theta \neq 5$.

a. We take $T_1 = \max\{X_1, X_2, \ldots, X_n\}$ as our test statistic. Specify what the (relevant) possible values are for T and which are in favor of H_0 and which are in favor of H_1. For instance, make a picture like Figure 25.1.

b. Same as **a**, but now for test statistic $T_2 = |2\bar{X}_n - 5|$.

25.6 ⊡ To test a certain null hypothesis H_0 one uses a test statistic T with a continuous sampling distribution. One agrees that H_0 is rejected if one observes a value t of the test statistic for which (under H_0) the right tail probability $\mathrm{P}(T \geq t)$ is smaller than or equal to 0.05. Given below are different values t and a corresponding left or right tail probability (under H_0). Specify for each case what the p-value is, if possible, and whether we should reject H_0.

a. $t = 2.34$ and $\mathrm{P}(T \geq 2.34) = 0.23$.

b. $t = 2.34$ and $\mathrm{P}(T \leq 2.34) = 0.23$.

c. $t = 0.03$ and $\mathrm{P}(T \geq 0.03) = 0.968$.

d. $t = 1.07$ and $\mathrm{P}(T \leq 1.07) = 0.981$.

e. $t = 1.07$ and $\mathrm{P}(T \leq 2.34) = 0.01$.

f. $t = 2.34$ and $\mathrm{P}(T \leq 1.07) = 0.981$.

g. $t = 2.34$ and $\mathrm{P}(T \leq 1.07) = 0.800$.

25.7 (Exercise 25.2 continued). The number of births in September is modeled by a Poisson random variable T with parameter μ, which represents the expected number of births. Suppose that one uses T to test the null hypothesis $H_0 : \mu = 1472$ and that one decides to reject H_0 on the basis of observing the value $t = 1718$.

a. In which direction do values of T provide evidence against H_0 (and in favor of H_1)?

b. Compute the p-value corresponding to $t = 1718$, where you may use the fact that the distribution of T can be approximated by an $N(\mu, \mu)$ distribution.

25.8 Suppose we want to test the null hypothesis that our dataset is a realization of a random sample from a standard normal distribution. As test statistic we use the Kolmogorov-Smirnov distance between the empirical distribution

function F_n of the data and the distribution function Φ of the standard normal:

$$T = \sup_{a \in \mathbb{R}} |F_n(a) - \Phi(a)|.$$

What are the possible values of T and in which direction do values of T deviate from the null hypothesis?

25.9 Recall the example from Section 18.3, where we investigated whether the software data are exponential by means of the Kolmogorov-Smirnov distance between the empirical distribution function F_n of the data and the estimated exponential distribution function:

$$T_{\text{ks}} = \sup_{a \in \mathbb{R}} |F_n(a) - (1 - \mathrm{e}^{-\hat{\lambda} a})|.$$

For the data we found $t_{\text{ks}} = 0.176$. By means of a new parametric bootstrap we simulated 100 000 realizations of T_{ks} and found that all of them are smaller than 0.176. What can you say about the p-value corresponding to 0.176?

25.10 ⊞ Consider the coal data from Table 23.1, where 23 gross calorific value measurements are listed for Osterfeld coal coded 262DE27. We modeled this dataset as a realization of a random sample from a normal distribution with expectation μ unknown and standard deviation 0.1 MJ/kg. We are planning to buy a shipment if the gross calorific value exceeds 23.75 MJ/kg. In order to decide whether this is sensible, we test the null hypothesis $H_0 : \mu = 23.75$ with test statistic $\bar{X}_n$.

a. What would you choose as the alternative hypothesis?

b. For the dataset $\bar{x}_n$ is 23.788. Compute the corresponding p-value, using that $\bar{X}_n$ has an $N(23.75, (0.1)^2/23)$ distribution under the null hypothesis.

25.11 ⊞ One is given a number t, which is the realization of a random variable T with an $N(\mu, 1)$ distribution. To test $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$, one uses T as the test statistic. One decides to reject H_0 in favor of H_1 if $|t| \geq 2$. Compute the probability of committing a type I error.