5. Hypothesis testing and goodness-of-fit test

Scientific statements

In chapter 1, I explained that science consists of theories and these comprise hypotheses. Scientists formulate these hypotheses as *universal statements* describing the world but they never know whether a hypothesis is true until it is rejected based on the empirical evidence. This makes science an infinite process of searching for true, to which we hopefully approach but never know whether we reach it or not.

Let's now return to the term *universal statement* I used in the previous paragraph and in chapter 1 as this is crucial to understand how empirical science works and hypothesis testing proceeds. Statements describing the world can be classified into two classes:

1. *Universal statements* apply generally on all objects concerned. E. g. "All (adult) swans are white" is a universal statement. This can be converted to a negative form: "Swans of other color than white do not exist." You can see that the universal statements prohibit certain patterns or events (e.g. observing a black swan here); therefore, they have the form of "natural laws". They can also be used to make predictions. If the white swan hypothesis is true, the next swan I will see will be white (and this is not dependent on how many white swans I saw before). A universal statement cannot be verified, i.e. confirmed to be true. We would need to inspect color all swans living on the Earth (and in the Universe) to do so (that is not very realistic) and even if we did so, we can never be sure that the next baby swan hatching from the egg would not be different from white at adulthood. By contrast, it is very easy to ***reject*** such universal statement on the basis of empirical evidence. Observing only a single swan of other color than white is sufficient for that.

2. *Singular statements* are asserted only on specific objects. E.g. "The swan I see is white." Such statement refers to a particular swan and does not predict anything about other swans. A specific class of singular statements are *existential* statements which can be derived from singular ones. The fact that I see a white swan (singular statement) can be used to infer that there is at least one swan which is white, i.e. white swans do *exist*. Based on the previous paragraph, you would probably not consider such statement any novel since it is in agreement with the universal statement on white swans. However, seeing a single black swan (Fig 5.1) completely changes the situation. It means, that at least one black swan exists and that the universal statement on white swans is not true. In general terms, this existential statement rejected the universal statement.

To sum up, scientific hypothesis must have a form of universal statements in order to have a predictive power, which we need to explain patterns in natural. They cannot be verified but can be rejected by empirical existential statements which are in conflict with the prediction the hypothesis makes.

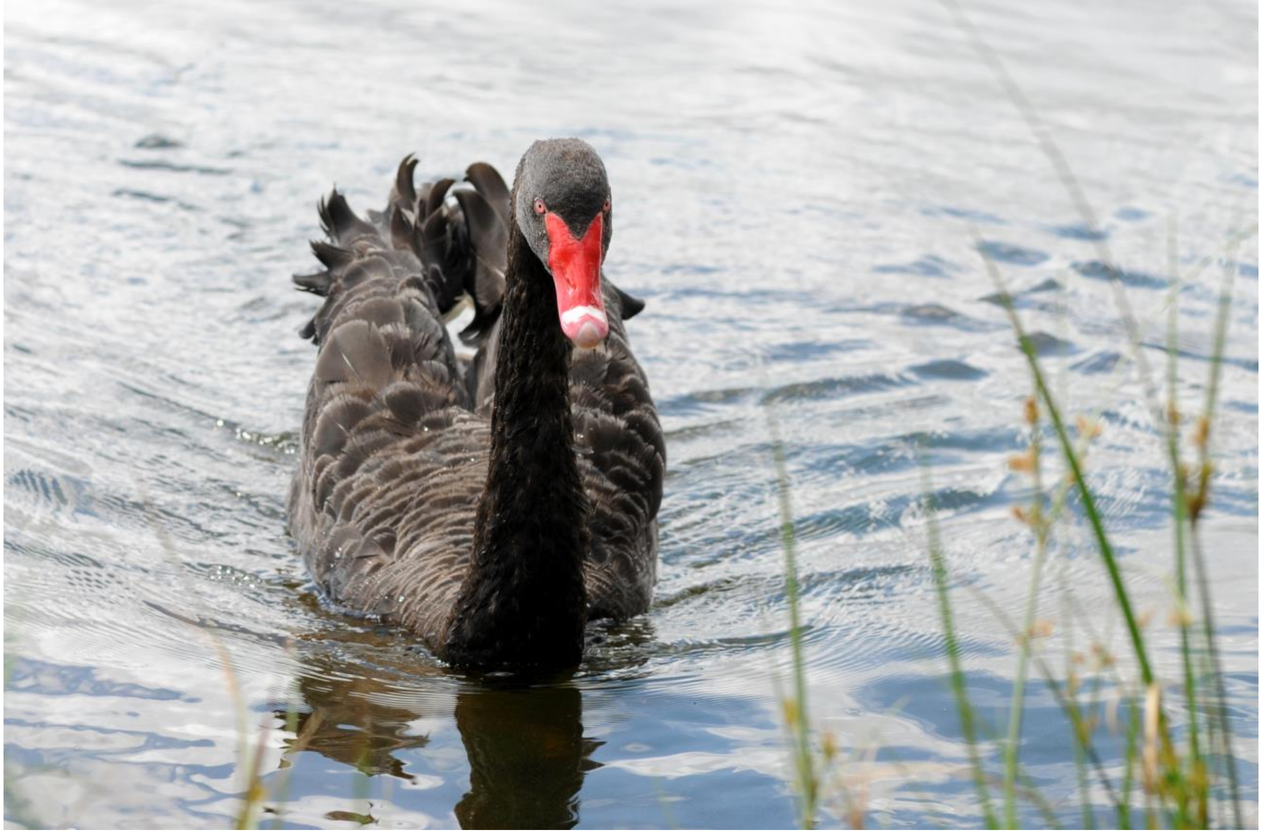

Fig. 5.1 A black swan in Perth (Western Australia).

Hypotheses and their testing

Empirical science is largely the process of hypothesis testing. This means searching for conflicts between predictions of hypotheses and collected/measured data. Once a hypothesis is rejected, a new hypothesis can be formulated to replace the old one. Note, here that there is no “objective” way how to formulate new hypotheses – they are rather genuine guesses.

An important implication from this is that for every scientific theory or hypothesis, it should be possible to define singular observations which if they exist would reject it. This means, that each scientific hypothesis must be *falsifiable*. Universal statements that are not falsifiable may be components of art, religion or pseudoscience but definitely not of science. Various conspiracy theories also belong to this class. These statements need not to be only dogmatic, they may also be tautological. Example of this is e.g. recently published theory of stability-based sorting in evolution (https://www.ncbi.nlm.nih.gov/pubmed/28899756), a “theory” which says that evolution operates with stability, i.e. organisms and traits which are more stable, persist for longer. The problem is that long persistence is a synonym for stability. So in fact the theory says “What is stable is stable”. Not very surprising. The authors declare the theory to explain everything (see the ending of the abstract), and this is indeed true, but the problem is that the theory neither produces any useful predictions nor can be tested by empirical data.

If we select only hypotheses which are falsifiable, and as such can be considered scientific statements, we may discover that there are multiple theories without any conflicts with the data. It is a natural question to ask, which one to choose over the others. Here, we should use the Occam's razor (https://en.wikipedia.org/wiki/Occam%27s_razor) principle and use the simplest (and also most universal and most easily falsifiable) hypothesis available. This is also termed "minimum adequate model" – i.e. choose the model with minimum number of parameters which fits adequately with the data.

Note on specifics of biology and ecology

Biological and ecological systems display high complexity arising from an interplay among complicated biochemical processes, evolutionary history and ecological interactions. As a result, quite large proportion of the research is exploratory aiming at discovering effects which were not anticipated yet. Therefore, no previous theory could have informed about them, or such information on absence of effect would be just redundant. In these cases, we test a universal statement, that the effect under investigation is zero.

Hypothesis testing with statistics

In statistics, we work with numbers and probabilities. Therefore, we do not record a clear-cut evidence to reject a hypothesis as in the example with swans. In other words, even improbable events do happen by chance and their observation may not be sufficient evidence to reject a hypothesis.

A general statistical testing procedure involves computation of *test statistic*. This statistic measures the discrepancy between the prediction of the *null hypothesis* and the data considering also strength of the evidence based on the number of observations. The test statistic is a random variable following certain theoretical distribution. As a result, probability of observing the actual data or data that differ even more from the null hypothesis expectation can be quantified. If this probability (called the *p-value*) is below certain threshold we can justify rejection of the null hypothesis.

The probability of observing certain data under null hypothesis can be very low but never zero. As a result, we are left with uncertainty concerning whether we did a right decision when rejecting or retaining the null hypothesis. We may take either right decision or make an error (Table 5.1).

Table 5.1. Possible outcomes of hypothesis testing by statistical tests. H_0 = null hypothesis

		Reality	
		H_0 is true	H_0 is false
Our Decision	Reject H_0	Type I Error	Ok
	Not reject H_0	Ok	Type II Error

Two types of error can be made, of which type I error is more harmful as it means rejection of a null hypothesis which is true. Such false positive evidence is misleading and may obscure the scientific research of given topic. By contrast, type II error (false negative) is typically invisible to anybody except to the researcher itself because results not rejecting the null hypothesis are not published. Statistical tools can quite precisely control the probability of making type I error, by setting an a-priori limit for the p-value. Typically, this limit called level of significance (α) is set to $\alpha = 0.05$ (5%). If the p-value resulting from the testing is higher than that, null hypothesis cannot be rejected. Note here, that such non-significant result does not mean that the null hypothesis is true. Non-significant results are indicative of absence of evidence, not of evidence of absence of an effect.

Concerning type II error (probability of which is denoted β), statistical inference is less informative. It can be quantified in some controlled experiments, but its precise value is not of particular interest. Instead, a useful concept is *power of the test*, which equals $1 - \beta$ and its relative rather than absolute size. Power of the test increases with sample size and with decreasing α, i.e. if the tester accepts an elevated risk of type I error.

Goodness-of-fit test

Let's have a look at an example of a statistical test. One of the most basic statistical tests is called goodness-of-fit tests (sometimes inappropriately chi-square test following the name of the test statistic). It is particularly suitable for testing frequencies (counts) of categorical data though the χ^2 distribution is quite universal and approximates e.g. very general likelihood ratio.

the formula is this: $\chi^2 = \sum \frac{(O-E)^2}{E}$

where O indicates observed and frequencies and E indicates frequencies expected under the null hypothesis. The sum is repeated for each of the categories under investigation.

The χ^2 value is subsequently compared with corresponding χ^2 distribution to determine the *p-value*. There are many χ^2 distributions which differ in the number of *degrees of freedom* (DF; Fig

5.2). The DF is a more general concept common to all statistical tests as it quantifies size of the data and/or complexity of the model. Here, it is important to know that for ordinary goodness-of-fit test:
DF = number of categories – 1.

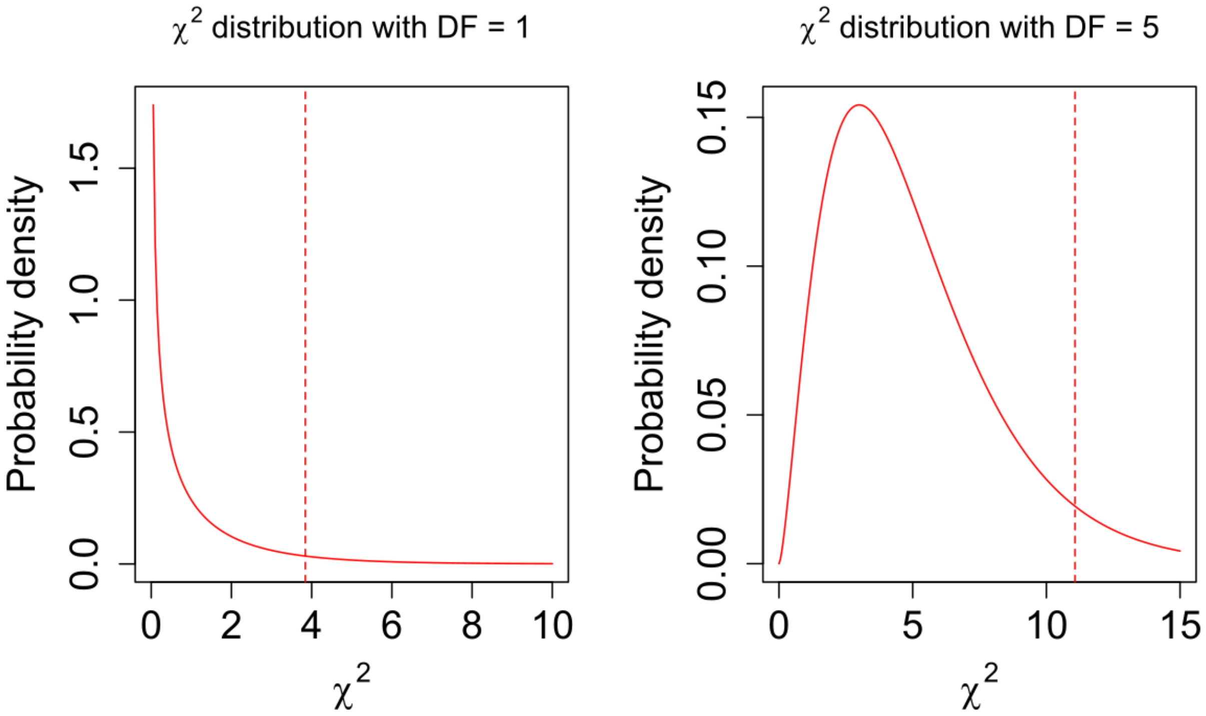

Fig. 5.2 Probability densities of two χ^2 distributions differing in the number of degrees of freedom. Dashed line indicates cut-off values for 0.05 probabilities on the upper tail.

Goodness-of-fit test example

A typical application of the goodness-of-fit test is in genetics as demonstrated in the following example:

You are a geneticist interested in testing the Mendelian rules. To do so, you cross red and white flowering bean plants. Red is dominant and white recessive, so in the F1 generation you only get red flowering individuals. You cross these and get 44 red flowering and 4 white flowering individuals in the F_2 generation. What can you say about the universal validity of the second Mendelian rule (which predicts 3:1 ratio between dominant and recessive phenotypes) at the level of significance α = 0.05?

First, you need to calculate the expected frequencies. These are:

E_{red} = 48 x 3 / 4 = 36

E_{white} = 48 x 1 / 4 = 12

then, computation of test statistic follows:

$\chi^2 = (44-36)^2/36+(4-12)^2/4 = \mathbf{7.11}$

DF = 1

$p(\chi^2 = 7.11, \text{DF} = 1) = 0.0077$

Conclusion (to be written in the text): Heredity in our bean-crossing system is significantly different from the second Mendelian rule (χ^2= 7.11, DF = 1, *p* = 0.0077). As a result, the second Mendelian rule is not universally true.

Here you can see that our experiment produced a singular statement on the number of bean plants. This was translated by the statistics into an existential statement that at least one (the our) genetic system exists which does not follow the Mendelian rule. This was then used to reject the universal statement.

How to do in R

Goodness-of-fit test: **chisq.test**

Parameter x is used for inputting the observed frequencies
Parameter p is used for inputting the null hypothesis-derived probabilities

Example with output:

```
chisq.test(x=c(44,4), p=c(3/4,1/4))
```

```
	Chi-squared test for given probabilities
data:  c(44, 4)
X-squared = 7.1111, df = 1, p-value = 0.007661
```

Probabilities of χ^2 distribution can be computed by **pchisq** (do not forget to set lower.tail=F to get the p=value).

```
pchisq(7.11, df=1, lower.tail = F)
```

```
[1] 0.007665511
```