Chapter 7

Hypothesis Testing

7.1 Introduction

Forming and testing hypotheses are the most important activities in the method of science. In Chapter **??** we have seen that, faced with a question, a scientist first guesses the possible answers to the question. When a guess satisfies certain minimum criteria, it can be called a scientific guess, or a 'hypothesis'. For any given question, scientists form a number of such hypotheses, and then set about testing the hypotheses. The basic objective of the test is to reject the wrong hypotheses.

Notice that proposing a 'successful' hypothesis (the one that ultimately survives) is not the sole purpose of doing good science. The enterprise of science depends on proposing a maximum number of hypotheses that are consistent with the clues on a given question. That is why a scientist who proposes a hypothesis that ultimately is proved to be wrong, is also doing good science. To show that a particular line of thought is wrong is good science, because it saves many scientists from following a wrong path. That is why, proposing and testing hypotheses are bonafide scientific activities even when the hypothesis in question is proved wrong.

7.2 The hypothetico-deductive model

The usual way of investigating a problem follows what is known as the hypothetico-deductive model. In this approach, faced with a problem, a scientist first looks up the literature to find if any possible explanation has already been proposed. If so, she checks if the explanation is satisfactory, i.e., it has been experimentally confirmed. If not, she proceeds to propose hypotheses that satisfy the conditions given in Chapter **??**.

Then one has to deduce a prediction from each hypothesis, i.e., if we assume that the hypothesis is true, what consequences follow? This is the 'deduction' stage. Then one has to use the result of this stage to eliminate the wrong hypotheses.

The next step is vital. One has to look for the *opposite* of each consequence in order to disprove a hypothesis. It is a logical error to seek evidence of the predicted event and cite it as proof of the hypothesis. This error is called affirming the consequent[1].

7.3 Planning experiments for hypothesis testing

The method of science expects that the demands of objectivity have to be strictly followed when planning and executing the test of a hypothesis. This is because, a scientist who has painstakingly worked out a hypothesis including its possible tests, may tend to believe in it. That belief may interfere with the experimental readings and measurements. History of science is replete with

[1] If P leads to Q, it does not imply that the occurrence of Q implies the occurrence of P. We can make the statement that "if the lamp were broken, then the room would be dark" but cannot infer its converse: "The room is dark, so the lamp is broken." This is because there may be other reasons for the room being dark. This is called the error of affirming the consequent. The cause of such a logic error is failure to realize that just because P is a possible condition for Q, P may not be the only condition for Q, i.e., Q may follow from another condition as well.

examples where scientists erred in recording information or ignored data when these contradicted one's 'pet' hypothesis.

Let us illustrate with an example how one can plan an experiment such that subjective biases cannot influence the results.

Suppose a scientist has hypothesized that a particular drug A is effective in curing a particular disease B. How would he test the hypothesis?

A person unexposed to the method of science may say that he would administer the drug on a few patients of the disease and if they get cured then the hypothesis is true and if they do not get cured the hypothesis is false. But that immediately raises a few questions. What if some people get cured and some do not? Can we still say that the drug cures the disease? For the people who get cured, can we say that the drug and not something else was the causative factor?

It is clear that the above procedure is too naïve to serve our purpose. One has to use the techniques of statistics to obtain meaningful answers to these questions. And for that, the experiment needs to be planned by satisfying the requirements of statistical tests.

7.3.1 Null and alternative hypothesis

We have seen earlier that hypotheses are always formulated in pairs: a null hypothesis and an alternative hypothesis. The null hypothesis is the statement that the proposed effect does not occur and the alternative is that the proposed effect occurs. These are denoted by the symbols H_0 and H_1 respectively.

For the present example, the null hypothesis would be that the substance has no effect on the disease. A scientific procedure would demand that the scientist believes in the null hypothesis unless he/she finds enough evidence against it. So one has to plan the experiment to test the correctness of the *null hypothesis*.

7.3.2 Sampling

The first requirement of statistical test is that a reasonably *large number* of patients of the disease have to be subjected to the test. These 'samples' should not be biased, i.e., should contain a mix of people of different categories, male and female (unless the disease is sex-specific) and people of all age-groups (unless the disease is age-specific).

The idea is that we are seeking to derive meaningful conclusion regarding the 'population' of people affected by the disease (its response to the drug), and for that, we should draw *random* samples from that population free from any bias or judgement.

7.3.3 Experimental group and control group

Suppose the experimenter has collected 100 such patients afflicted by the disease. If the proposed drug is applied on all the patients, there will be no way of distinguishing between application and non-application of the drug.

So the usual procedure is to divide them into two groups: the 'experimental group' and the 'control group'. The drug would be applied on the experimental group and will not be applied on the control group. The scientist would then expect to infer the causal connection between the hypothesized drug and the cure of the disease from the data obtained from the two groups.

Note that, if the drug cannot cure the disease, still some people of the experimental group will be cured, and if it is effective, still some people of that group will not be cured. Therefore the decision has to rest on some statistical analysis of the results obtained from the individuals belonging to the two populations.

The first major requirement is to assign subjects to the experimental and control groups without any bias, i.e., completely randomly. But how to do the randomization properly? There are two common practices.

Completely Randomized Design: In this method, one labels each subject with a number, then uses a random number genera-

tor in a computer to select from the labelled subjects.

Randomized Block Design: There are situations where an experimenter may be aware of specific differences between the experimental subjects or objects, which may have some influence on the result. In such situations experimental subjects are first divided into homogeneous blocks depending on that characteristic before they are randomly assigned to a treatment group.

For example, if one is trying to find cure of a disease, the experimenter may have reasons to believe that people of different age groups may respond differently to the proposed curative procedure. In such a situation she should first divide the experimental subjects into 'blocks' of different age groups, such as under 30 years old, 30-60 years old, and over 60 years old. Then, individuals out of these blocks would be assigned to the experimental and control groups using a completely randomized design.

7.3.4 Recording the results

How does the scientist obtain information about who is getting cured and who is not? Should she go about asking the patients "How are you feeling today"? If she does that, she will get subective responses which may be affected by the mood of the patient that day. More importantly, in order to subject the results to statistical tests, she will have to obtain the results in terms of numbers. This could be the blood pressure of the patient, the body temperature, or the bacteria count in blood samples—anything that can be recorded as numbers.

7.3.5 Eliminating experimenter bias

If the experimenter *believes* in either the null or the alternative hypothesis, there is a possibility that the belief may inadvertently influence the results. For example, if the experimenter believes that the hypothesized drug is really a cure of the disease and she is testing it by counting the number of bacteria in the blood

samples of the patients, she may over-count the bacteria in the blood samples of the patients of the control group and under-count those in the patients of the experimental group.

In order to eliminate possibility of such experimenter bias, the standard procedure is to have the measurements done by somebody who does not know which patient belongs to the experimental group and which to the control group. Elaborate procedures are evolved to mask the information from the person doing the measurements. This is called *single blind test*.

7.3.6 Eliminating experimental subject bias

As we have said earlier, the patients of the control group are not given the proposed drug. Now consider the psychological state of the person: she is sick and knows that she is not receiving any treatment. This knowledge may affect her psychologically, and the person's psychological state may affect her physical health.

To avoid such effects, the standard procedure is to administer the substance A on the people of the experimental group, and to administer a *placebo* (a substance that looks and tastes like A but is not A) on the people of the control group. Thus, a condition is created where the experimental subjects do not know whether they belong to the experimental group or to the control group. This is called a *double blind test*.

7.4 The statistical test

Some time after administering the drug and the placebo, one has to measure the effect in terms of some objectively measurable quantity. As an example consider an infectious disease, caused by some bacteria, and suppose that the state of the disease is reflected in the bacteria count in blood samples. So the scientist would have to get the bacteria in blood samples counted, and through these measurements, the scientist gets a mass of data obtained from the two groups.

As regards the data, the two hypotheses can be stated as:

H_0 : The mean number of bacteria in the blood of people of the two groups is the same

H_1 : The mean number of bacteria in the blood of people of the experimental groups is less than that of the control group.

Notice that there are in principle two 'populations': those who have the disease and have been administered the drug, and those who have the disease and have not been administered the drug. The experimental arrangement is to draw 'samples' from these two populations.

For the sake of convenience, let us now introduce the following notations.

μ_E: The mean bacteria count in the blood of the experimental population,

μ_C: The mean bacteria count in the blood of the control population,

σ_E: The standard deviation of the bacteria count in the blood of the experimental population,

σ_C: The standard deviation of the bacteria count in the blood of the control population.

Let $\bar{x}_E$, $\bar{x}_C$, s_E and s_C represent the corresponding values obtained from the samples. Suppose, further, that there were n_E data points obtained from the experimental group, and n_C from the control group (these need not be equal).

Now the two hypotheses can be expressed in compact form as

$$H_0 : \ \mu_E = \mu_C$$
$$H_1 : \ \mu_E \neq \mu_C$$

Notice that these statements concern the population means, and not sample means.

The test procedure

In the scientific method, one starts by believing the null hypothesis and tries to test if there is sufficient ground to reject the null hypothesis. If the null hypothesis is false, the sample means would differ greatly, i.e., the *difference* between the sample means can be taken as the test statistic with which one can try to decide if the population means also have sufficient difference to justify rejection of the null hypothesis. So we need to focus on the quantity $\bar{x}_E - \bar{x}_C$, and have to decide whether this quantity is 'large enough'.

From the data we can calculate the difference $\bar{x}_E - \bar{x}_C$. But this is only a sample. If we take another pair of samples from the two 'populations', the value obtained may be different. If all possible pairs of samples are drawn (one does not really have to do that), that would result in a *sampling distribution of the difference between two sample means*. The question is, can we estimate this distribution from the data taken?

The Central Limit Theorem tells us that, if the sample size is fairly large (at least 25), then the sampling distribution of the mean is approximately a normal distribution with mean μ and SD=$\sigma/\sqrt{n}$. There is a very important result in statistics which says that the above is true also for the sampling distribution of the differences. More accurately, the distribution of the difference between two (independent) random variables which are each normally distributed is also a normal distribution.

What will be the mean and the standard deviation of the difference statistic? The mean is simply the difference between the two population means $\mu_E - \mu_C$. The variance is given by

$$\frac{\sigma_E^2}{n_E} + \frac{\sigma_C^2}{n_C}$$

And hence the standard deviation, or the *standard error of the*

difference between two sample means is given by

$$\text{SE} = \sqrt{\frac{\sigma_E^2}{n_E} + \frac{\sigma_C^2}{n_C}} \tag{7.1}$$

It is beyond the scope of this text to prove these results; one may consult any standard textbook on statistics for that.

Thus, we are able to construct the distribution of the difference between two sample means. This allows us to answer the question regarding the correctness of the null hypothesis.

If the null hypothesis is true, then the two population means are equal, i.e., $\mu_E - \mu_C = 0$ and hence the sampling distribution will be approximately a normal distribution with mean 0 and standard deviation given by (7.1). If it is a normal distribution, 95% of the differences $\bar{x}_E - \bar{x}_C$ will be within 1.96 standard deviations of the mean $\mu_E - \mu_C = 0$. In other words, differences larger than 1.96 standard deviations is unlikely to occur if the null hypothesis is true. On the other hand, if differences of this size do occur, there will be basis of rejecting the null hypothesis.

Thus, the test involves a comparison of the difference between the sample means $\bar{x}_E - \bar{x}_C$ and 1.96 SE. This is normally done by obtaining the quantity $(\bar{x}_E - \bar{x}_C)/\text{SE}$, and comparing it with 1.96. We reject H_0 if

$$\text{either} \quad \frac{\bar{x}_E - \bar{x}_C}{\text{SE}} \leq -1.96 \quad \text{or} \quad \frac{\bar{x}_E - \bar{x}_C}{\text{SE}} \geq 1.96$$

However, we need σ_E and σ_C in order to calculate SE. We overcome the problem the way we have done earlier: by using s_E and s_C as estimates for σ_E and σ_C. Thus the estimated value of SE (or ESE) is given by

$$\text{ESE} = \sqrt{\frac{s_E^2}{n_E} + \frac{s_C^2}{n_C}}$$

We define a quantity z as

$$z = \frac{\bar{x}_E - \bar{x}_C}{\text{ESE}}.$$

We reject the null hypothesis if z is large. If

$$\text{either} \quad z \leq -1.96 \quad \text{or} \quad z \geq 1.96$$

then we say that the null hypothesis is rejected with a 95% confidence. The is the standard z-test.

One may notice that the statistical procedure outlined above depends on the assumption that the distributions are normal. One may also come across situations where the distribution may not be assumed to be normal. In those cases other tests are recommended, for example the t-test, the χ^2 test, etc.

7.4.1 Types of error

In testing a hypothesis there are possibilities of making two types of error. When the null hypothesis is actually true but the test rejects it, that is called a 'Type-1 error'. On the other hand, when the null hypothesis is actually false but the test fails to reject it, that is called a type-2 error.

A type-1 error leads an investigator to conclude that a supposed effect or relationship exists when in fact it does not, i.e., to falsely infer the existence of something that is not there. This can happen when a scientist gets emotionally attached to a hypothesis and unwittingly tries to prove it right.

A type-2 error is to falsely infer the absence of something that is present. This can happen if a scientist unwittingly tries to stick to a common belief and bases himself or herself on false information.

Both these types of error can occur if a scientist carries a subjective bias or if the statistical procedure of hypothesis testing is not followed correctly.

Example 7.1: A scientist has the following questions regarding

the growth of a specific species of fish:

1. Are adult male fishes heavier than adult female fishes?
2. Are first-year male fishes heavier then first-year female fishes?
3. Are adult female fishes heavier than first-year female fishes?
4. Does the body-weight of male fishes increase after the first year?

She caught the fishes belonging to that species from a pond and collected the data on the weights (in grams) of 664 fishes. These are summarised in the table:

	Sample size	Sample mean	Sample SD
Adult males (AM)	95	113.4	11.92
Adult females (AF)	137	108.9	10.07
First-year males (FM)	152	110.2	9.68
First-year females (FF)	280	108.9	10.59

State the null and alternate hypotheses on each question, and check which one is supported by the data.

Solution: Let the mean weights of the four categories be denoted as $\mu_{\text{AM}}, \mu_{\text{AF}}, \mu_{\text{FM}}, \mu_{\text{FF}}$ respectively.

Question 1: The hypotheses may be stated as:

$$\begin{aligned} H_0 &: \mu_{\text{AM}} = \mu_{\text{AF}} \\ H_1 &: \mu_{\text{AM}} > \mu_{\text{AF}} \end{aligned}$$

The estimated standard error is:

$$\text{ESE} = \sqrt{\frac{s_{\text{AM}}^2}{n_{\text{AM}}} + \frac{s_{\text{AF}}^2}{n_{\text{AF}}}} = \sqrt{\frac{11.92^2}{95} + \frac{10.07^2}{137}} = 1.495$$

The test statistic is

$$z = \frac{\bar{x}_{\text{AM}} - \bar{x}_{\text{AM}}}{\text{ESE}} = \frac{113.4 - 108.9}{1.495} \approx 3.01$$

Since $z = 3.01 > 1.96$, we can reject the null hypothesis with 95% confidence.

We leave it as an exercise to answer the other three questions on the basis of the data provided. □

Exercise

1. In a hypothesis testing experiment, you have collected data from the experimental and control groups, and have obtained the sample mean and standard deviation as:

Control group:	$n_C = 52$	$\bar{x}_C = 82.3$	$s_C = 5.65$
Experimental group:	$n_E = 47$	$\bar{x}_E = 79.8$	$s_E = 4.95$

   If the hypotheses can be stated as $H_0 : \mu_C = \mu_E$, $H_1 : \mu_C > \mu_E$, is there enough ground for rejecting the null hypothesis?

2. You have conducted an experiment by creating a control group of 53 individuals and and an experimental group of 49 individuals, and have measured a quantity x in the two populations. From the data, you obtain the sample means as $\bar{x}_C = 3.56$ and $\bar{x}_E = 4.18$, and the standard deviations as $\sigma_C = 0.69$ and $\sigma_E = 0.82$.

   (a) If the null and alternate hypotheses demand $H_0 : \mu_E \leq \mu_C$, $H_1 : \mu_E > \mu_C$, do the data provide enough ground to reject the null hypothesis?

   (b) How would you state the values of x_C and x_E and in a paper?

   (c) What will you have to do if your are asked to halve the error bars of x_C and x_E?