Hypothesis testing

Snehal M. Shekatkar

Department of Scientific Computing, Modeling, and Simulation,
Savitribai Phule Pune University, Pune, India 411007

1 Introduction

Hypothesis testing is the third type of problem in inferential statistics. A **hypothesis** is simply a statement that claims that something is true. A **statistical hypothesis** is a statement that claims that something is true in a *statistical sense*. Hypothesis testing refers to deciding whether a given statistical hypothesis could be accepted as true or not based on some observed data. At the heart of hypothesis testing are two statistical hypotheses called **Null hypothesis** and **Alternative hypothesis** which are always opposite of each other. As we will see, based on the observed data, we may decide to retain or reject the null hypothesis.

In the real-world, we often need to make a decision based on some data. For example, whether to accept the claim by a pharmaceutical company that a newly developed drug actually works for a particular disorder, or whether to regulate violent programs on television because they may be making children violent on average. Here we can think of the two hypotheses as given below:

1. The first one, called **Null hypothesis** (usually denoted by H_0) is the default position. It assumes that if we have no information about the system, then we can't claim that something special is happening. For example, H_0 would imply that the new drug has no effect on the disorder or that violent programs have no effect on the behavior of children. It is very important to realize that although H_0 is called 'hypothesis', it actually doesn't assume anything. That's why we use the word 'null' to describe it.

2. The **Alternative hypothesis** (usually denoted by H_1, and sometimes also referred to as **research hypothesis**) is the negation of H_0. It asserts that something special is happening. In the two examples above H_1 may say that the said drug is effective for the disorder or that the violent scenes make children more violent.

Given the data $X_1, X_2, \cdots, X_n$, our task is to decide whether there is **sufficient evidence** in the data to move away from our default position. If yes, we say that we **reject the null hypothesis**. If not, we say that we **could not reject** H_0 or just that we **retain** H_0. Another way of looking at this is to see whether the observed outcome would be rare if we assume H_0 to be true. If yes, probably H_0 is not true.

The topic of hypothesis testing is complex and conceptually tough to learn. Thus, here we will present its simple outline along with some procedures that can be used for solving practical problems. However, before actually starting to describe these procedures, let us try to understand important elements of the topic whose understanding is of utmost necessity for conceptual understanding.

- **Randomness forbids a deterministic decision:** Let us say that you want to decide whether the distance between cities A and B is greater than 100 kms or not. If you have no doubt about the measurement, and if you also assume that cities don't wander around, then you can be pretty sure about your conclusion in a single measurement. The point is that once you fix the points in the two cities, then the measurement of the distance would yield the same value even if you perform the measurement several times. This is *not* the case for the systems with randomness where we can meaningfully talk only about the **average** behavior. Think of the question of deciding whether the

Indian cricket team is better than the cricket team of Bangladesh. If you take the last 10 matches played between the two teams, it may have happened that the Indian team won more matches either because it is actually a better team or just by chance. Thus, we can never be sure about the such systems. In other words, rejecting H_0 based on a particular sample does not mean that H_1 is true or the other way around.

- **Multiple measurements / Groups are needed:** Continuing with the last example, why did we talk about last 10 matches instead of just the last match? The reason is same: the randomness. Even a good team may lose a match against a weaker team. That's why it is a bad idea to conclude anything based on a single measurement when you know that the system's behavior is noisy. On the same note, in a problem of deciding the effectiveness of a drug, one has to keep in mind that checking the effect on a single person is completely useless because whether to attribute the observed positive effect to the medicine or better immunity of *that particular person* is not clear. Hence whenever possible, we need as many measurements as possible either in the form of multiple measurements or multiple subjects. (Here the word *subject* means a human or an animal which is part of the experiment that is being carried out to gather relevant data).

- **Sample should be representative of the population:** If in our drug effectiveness experiment we tried drug only on the people who are on their deathbed, then we would see that almost everybody dies after taking the drug. On the other hand, if the drug is given only to very fit people, then we will see improvement in everybody. Since just like other problems in inferential statistics here also we are interested in the effect on the population, not a particular sample, it is important to choose a sample that is representative of the population. As we have seen already, the best way to do this is to choose the sample uniformly randomly from the population. Of course this may not always be feasible, and in that case we need to minimize the bias in the sample using other ways which we will discuss later. Later we will talk about **Randomized Control Trials** in which some bias in the sample may be acceptable. However, it is almost always true that any significant bias in the sample will lead to biased conclusions however careful we are with processing the data.

2 Specifying H_0 and H_1

The first and probably the most important step in hypothesis testing is to correctly specify the null hypothesis and the alternative hypothesis. Let us consider an example. A friend of yours claims that she knows a person who can predict coin tosses almost correctly, and she believes that this is probably because of some supernatural power. Of course she has to come up with some evidence to support her claim, so she shows you the data she obtained the last week. The data consists of asking the person supposedly possessing the powers the result of a coin toss exactly 100 times. She shows you the data, and turns out that the person could actually predict 78 out of 100 coin tosses correctly! What would you make of this? Should you accept your friend's claim? Or can you somehow refute it? The first step here is to correctly identify H_0 and H_1 as given below.

- H_0: The person cannot predict coin tosses better than random guess

- H_1: The person has supernatural powers to predict the toss of a coin

Observe how H_0 doesn't assume anything; the claim is always made by H_1 by negating what H_0 says. Depending upon how many coin tosses were actually predicted, we may want to retain or reject H_0.

In most cases, we are interested in some average over the *population*, and not about one particular element. In such cases, H_0 and H_1 are statements about the population means. For example, if the claim is that average height of people from Chennai is greater than the rest of the India, then we will have:

- H_0: $\mu_{\text{Chennai}} = \mu_{\text{India}}$

- H_1: $\mu_{\text{Chennai}} \neq \mu_{\text{India}}$

Here μ denotes the average height. Of course you may find a person in Chennai taller than say a person from say Delhi, or the other way around, but the statements are about the averages. This particular example, is an example of what is known as the **two-tailed test**. The "two" in "two-tailed" here signifies that we are only saying that the two averages are different; we are not claiming that one average is larger or smaller than other.

Now consider a slightly different claim: the average height of people from Chennai is greater than the average height of people from rest of the India. In this case the two hypotheses would be:

- H_0: $\mu_{\text{Chennai}} \leq \mu_{\text{India}}$

- H_1: $\mu_{\text{Chennai}} > \mu_{\text{India}}$

Observe how we had to use different H_0 than the one we used in the previous example. This is actually an example of a **right-tailed** test because $>$ sign appears in the alternative hypothesis. If instead H_1 had claimed that the mean height of people from Chennai is *less than* the people from the rest of the India, then that would have been an example of a **left-tailed** test. If a test is either left-tailed or right-tailed, then it is called a **single-tailed** test.

3 Retaining or rejecting H_0

Suppose that one starts guessing the result of a coin toss completely randomly. If the coin is fair, and you guess either *Heads* (H) or *Tails* (T) each with probability 0.5, then you would actually be able to predict around half of the tosses correctly! So you must decide what would be the number of coin tosses predicted beyond which you will accept the claim. It is not hard to see that many times just by chance you may even be able to predict more than 50 coin tosses out of 100. So what is the level of acceptability for supernatural powers? Is more than 65 correct predictions enough? Or must it be more than 90? You see, to some extent this is a subjective decision. You need to decide what qualifies as a rare outcome if you assume that the person does *not* possess those powers. To me, 60 looks rather feeble but 100 sounds too much (we must consider the possibility that even the supernaturals will fail sometimes!).

So let us assume that the probability of obtaining the correct predictions just by chance should be no more than 0.00001. If we assume that the probability of correctly guessing when a person does not possess any powers is p, the probability of correctly predicting k or more coin tosses out of n is:

$$P_{pred} = \sum_{i=k}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

Since without supernatural powers a person has 50% chance to correctly predict the result of a toss, let us set $p = 0.5$. Let us also set $n = 100$. Then we want the value of k which will make this probability less than our acceptability threshold 10^{-5}. It is not hard to verify that $P_{pred} < 10^{-5}$ when $k = 72$. Thus, if somebody can actually predict 72 or more tosses correctly out of 100, with this acceptability threshold, we should accept that the person does possess some kind of power! The whole point is that the chance of happening this just by fluke is extremely small. You can always question the threshold level. Why 10^{-5} and not 10^{-8} for example? That's your choice. This is decided by *your definition* of a rare event. This is the *level of significance* or the *critical value* that you have set. Once the level of significance is fixed, philosophy of hypothesis testing tells us that depending upon the data we either need to retain H_0 or reject it (If you are wondering, in this particular case, the person would actually pass the test even if you set the threshold to 10^{-8} because he could predict 78 tosses correctly!).

When we reject H_0, we sometimes say that we have obtained a **statistically significant result**, and you may see researchers using the phrase **test of significance** to refer to a hypothesis test.

Hypothesis testing problems are of varied type. For example, one may want to check whether a given data comes from the normal distribution, whether given two variables can be assumed to be independent, whether the mean of the underlying distribution is equal to 1, and so on. Usually, for a given type of testing, we use a particular **test statistic**, a function $g(X_1, X_2, \cdots, X_n)$ of the data. Here we will restrict ourselves only to hypothesis testing problems which use the mean of a distribution as the test statistic. **This necessarily involves sampling distribution of the mean** as we explain using an example in the next section.

4 Z-test

A company that manufactures cupcakes prints on each of its packets containing several cakes that the mean weight of the packet is 200 grams. You purchase two packets from one of their stores, and later out of curiosity you actually weigh them. You find that the two packets weigh 192 grams and 190 grams. You feel cheated, and you go to that store again and complain about it. A representative in the store tells you that making of cupcakes is a manual process, and such fluctuations are expected, although the mean weight of a packet is definitely 200 *grams or more*. He takes out another packet from the shelf, and weighs it in front of you. The electronic balance shows weight 204 grams. Moreover, he even offers you that if you want to verify his claim, you may purchase *all* the packets from his store, weigh them and the average of those numbers will come out 200 grams or more for sure!

Unless you want to spend several thousand rupees just to verify this claim, and also to go through tedious weighing activity, you will need to some other way to find whether the claim is true or not. A friend tells you about a government agency which is established to ensure that the marketed food products are up to the mark, and you think that they can bear such expense easily and complain to them about this particular company's product. While talking to the agency's representative you tell her that the agency can and should buy all the packets from the store, and if the average weight of a packet is less than 200 grams, you must fine them. However, the representative, having learned basic statistics, tells you that this is neither a necessary nor a correct way of approaching the problem.

A part of the problem with the approach is that the company has many manufacturing units in the country, and they supply these packets to stores in their geographical vicinity. Thus, the packets in a single store do not form a **sample representative of the population** i.e all the packets that the company sells. Because of this, even if the average weight for a packet from a particular store comes out to be more or less than 200 grams, we can't immediately fine the company. Moreover, even the agency won't buy *all* the packets from any store because certainly they don't want to open their own store to sell those packets after the testing is done!

The way out, your are told, is to choose a **random sample** from stores all over the country, and then use weights of those packets to test whether the company's claim of 200 grams is really correct or not. Of course now this won't be a deterministic decision because we are not using all the packets that the company produces (and the packets that it has produced in the past, and those which it will produce in future). However, if our sample is sufficiently representative of the population, we would be able to determine with high confidence whether the mean weight is actually 200 grams or more, or not.

In this case, the two hypotheses are:

- **H_0:** $\mu \geq 200$ grams

- **H_1:** $\mu < 200$ grams

Notice that it is *you* who is making the claim, not the company. The default position is that the mean weight is indeed 200 grams or more. The **burden of proof** is on you, the person making the claim, to prove that this default position is false. Certainly it would be silly to argue that the company has to prove that the mean weight is 200 grams or more to satisfay your doubt.

The government agency now starts collecting samples from different stores around the country (You should keep in mind that this is a hypothetical agency; response of real governmental agencies to complaints by public is surely not this swift!). Suppose that they collect total 50 samples whose measured weights in grams are given below:

$$195, 194, 189, 196, 197, 195, 204, 192, 205, 202,$$
$$203, 202, 203, 204, 196, 205, 201, 202, 197, 195,$$
$$194, 209, 189, 200, 196, 191, 198, 198, 198, 190,$$
$$203, 200, 200, 204, 196, 190, 200, 203, 199, 191,$$
$$200, 204, 203, 194, 190, 193, 201, 199, 192, 199$$

The average of these 50 weights comes out to be 198.02 grams. Should you celebrate? Not just yet! The sample average for another sample of size 50 would likely produce a different value which might be greater than 200 grams or even less than 198 grams. How do you decide what to do? **We must decide how rare this outcome is under the assumption that H_0 is true**.

Let us assume that the population mean and population variance for the packet weights are μ and σ respectively. Then, as we already have studied, the sampling distribution of μ, i.e. the distribution of the sample mean $\overline{X}_n$ would be normal with mean μ and variance $\sigma/\sqrt{n}$ where n is the sample size (Recall again that the variance of the sampling distribution is the *standard error*). Of course since we don't know the sample variance, it needs to be estimated from the sample itself. Here we will use the unbiased estimator:

$$\widehat{\sigma}_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}$$

For our data, this gives $\widehat{\sigma} = 4.99$. Hence the estimated standard error is:

$$\widehat{se}_n = \frac{\widehat{\sigma}_n}{\sqrt{n}} = \frac{4.99}{\sqrt{50}} = 0.705$$

Let us define our *level of significance* to be $\alpha = 0.05$. Then any outcome for which the probability of occurrence of $\overline{X}_n$, **assuming that** $\mu \geq 200$, is less than or equal to 0.05 will be considered rare. In other words, if the observed value of $\overline{X}_n$ has indeed come from the sampling distribution with mean 200 and standard deviation 0.705, then we ask how rare that occurrence is. If the probability of that happening is less than the level of significance we have set (in this case $\alpha = 0.05$), then we will term that as a rare occurrence and reject the null hypothesis. Otherwise we will retain the null hypothesis. **Also notice that this is a left-tailed test**.

We know that the random variable $Z = (\overline{X}_n - \mu)/\widehat{se}_n$ has a standard normal distribution. Thus, we ask what is the value z_α such that the probability of getting a value z_α *or less* is equal to 0.05? Since this is a left-tailed test, we want z_α to satisfy:

$$\mathbb{P}(Z \leq z_\alpha) = 0.05$$

As you might have guessed, z_α can be easily found obtained in python as the following code shows. The code uses a special function 'trapz' from the 'scipy.integrate' module to integrate a function using Trapezoidal method of numerical integration.

```python
import numpy as np
from scipy.integrate import trapz

def normal(x, mu, sigma):
    """
    Return the values of the normal distribution computed on array x
    """
    dist = np.exp(-(x-mu)**2/(2*sigma**2))
    norm = 1./np.sqrt(2*np.pi*sigma**2)
    return norm * dist

# Here we assume that the value z_alpha is greater than -3.5. If alpha
# is smaller, you will need to reduce it to an even smaller value
for z_alpha in np.arange(-3.5, 0, 0.0001):
    x = np.arange(-100, z_alpha, 0.001)
    z = normal(x, 0, 1)
    if trapz(z, x) > 0.001:
        print(f"z_alpha = {z_alpha}")
        break
```

You should verify that the critical value $z_\alpha = -1.643$. Thus, the Z-score for our $\overline{X}_n$ is less than -1.643, and we should reject the null hypothesis. The Z-score is:

$$\frac{\overline{X}_n - \mu}{\widehat{se}_n} = \frac{198.02 - 200}{0.705} = -2.81 < -1.643$$

Thus, there is less than 5% chance that the true mean weight of the cake packet is 200 grams. We must reject H_0 and fine the company! The company at this point may object that our level of significance is too big. However, in this case, even if we set $\alpha = 0.01$, corresponding to 99% confidence, then we get $z_\alpha = -2.326$ which is greater than our Z-score. That means that the chance that company's claim is true, is in fact less than 1%. At this point, the hypothetical government agency, contrary to what the real government agencies usually do, fines the company and asks them to improve their packaging.

If this test were two-tailed instead of one-tailed, then we need to formulate the two hypotheses in a different fashion like this:

- H_0: $\mu = 200$ grams

- H_1:$\mu \neq 200$ grams

Here the aim is to see whether the actual mean weight is *different* from 200 grams. The company might be interested in conducting such study so that it is neither fined because of lower mean weight, nor it sells more quantity than the agreed-upon value. In this case, we define the value $z_{\alpha/2}$ similar to we did in constructing $1 - \alpha$ confidence interval:

$$\mathbb{P}(Z \leq -z_{\alpha/2}) + \mathbb{P}(Z \geq z_{\alpha/2}) = \alpha$$

If the Z-score of $\overline{X}_n$ is either less than $-z_{\alpha/2}$ or $z_{\alpha/2}$, it is an indication that the under the assumption that the true $\mu = 200$, the observed sample average is rare, and H_0 is rejected. Otherwise, H_0 is retained.

5 Student's t-test

Suppose that the problem similar to the cupcake problem above arose but now in this case the question is about the mileage of a car (i.e. the average distance a car can travel with per unit of fuel). A car manufacturing company claims that the average mileage is 40 kms per litre of fuel. You want to buy the car, but you would like to test the claim since cars are expensive!

You know that the company allows a reasonably long test drive during which you can actually check the total distance you drove and the amount of fuel consumed. However, the showroom in your city doesn't allow multiple test drives for the same person. Surely you would have liked to take around 50 test drives and use the Z-test but that looks difficult! Of course it is difficult to ask, say 50 different people, to do that since they will need to invest time and effort, and also need to have a driving license and so on. It should be clear that getting one sample point in this case is way too difficult than weighing cupcakes! You ask among your friends, and succeed to convince 4 of them to take the test drive for the same car model. Here are all five mileage values (including yours):

$$40.01,\ 35.99,\ 34.14,\ 41.18,\ 41.83$$

From these, we get $\overline{X}_n = 38.63$. Of course we shouldn't immediately say that this implies that the true mileage is less than 40 since another sample of 5 test drives would surely have yielded a different sample average. Thus, we must first clearly formulate the hypotheses for testing as stated below.

- H_0**:** $\mu \geq 40$
- H_1**:** $\mu < 40$

Can we now again apply the Z-test procedure and see whether the value $\overline{X}_n = 38.63$ is rare under the assumption that the population mean is 40? As you probably guessed correctly, this won't be a good idea since the sample size $n = 5$ is too small. The main issue is that since the standard error $se_n = \sigma/\sqrt{n}$ depends on unknown population standard deviation, we need to use its estimate $\widehat{se}_n = \widehat{\sigma}_n/\sqrt{n}$. But when n is small, the distribution of

$$T_n = \frac{\overline{X}_n - \mu}{\widehat{\sigma}_n/\sqrt{n}}$$

will not be normal. Hence, we cannot compute the critical t_α by assuming Gaussian distribution. Now if we make an assumption that the scores obtained from the test drives are Gaussian distributed, then we know that the variable T_n is distributed according to the Student's t-distribution with 4 degrees of freedom (since DOF $= n - 1$). What is the critical value t_α? This can be easily calculated in python (Go back to the notes on confidence intervals), and we get $t_\alpha = -2.132$.

In our case, the T-score is:

$$T_n = \frac{\overline{X}_n - \mu}{\widehat{\sigma}_n/\sqrt{n}} = \frac{38.63 - 40}{3.383/\sqrt{5}} = -0.906 > -2.132$$

Hence, at 0.05 significance level, we can't reject H_0. That is, we couldn't reject the manufacturer's claim that the mileage is 40 since our data does not contain evidence to do so. In fact since we couldn't reject H_0 at $\alpha = 0.05$, we also can't reject it at any lower α (Make sure you understand why). Only if it were the case that we rejected H_0 at certain level of significance, it might be possible to retain it at a lower significance level.

6 Two important points

We should note two important points about this whole procedure of retaining or rejecting the null hypothesis.

1. **Rejecting H_0 is a stronger decision than retaining it**

   Why so? The reason is that when we retain H_0, we don't exactly mean that H_0 is true, but rather it *could* be true given the observed data. This is especially true when H_0 talks about a quantity like the *mean.* In our car mileage example, we failed to reject manufacturer's claim that the car mileage is 40 or more. But suppose that the true mileage is actually 39, and not 40, then the observed data is consistent with this μ too, and hence retaining H_0 that says $\mu \geq 40$ does not prove that it is really true. For this reason, the most logical choice of stating this outcome is to say that we *failed* to reject H_0.

   On the other hand, rejecting H_0 means that H_1 is true given the observed data, not that it could be true. Hence rejecting H_0 is a stronger decision.

2. **Rejecting H_0 refers to population while statistically significant result refers to the sample**

   This is probably confusing since we reject H_0 when we get a statistically significant result! This is indeed true, however, the way we should look at these is that whether the obtained result is statistically significant or not would depend on the sample. But we use this outcome to reject H_0 which is actually a statement about the population.

7 Types of Errors

Since the decision to retain or reject H_0 is probabilistic, it can go wrong sometimes. Accordingly, we define the following errors:

- **Type I error:** This is the error of rejecting H_0 when it is in fact true.

- **Type II error:** This is the error of retaining H_0 when in fact H_1 is true.

Type I error is considered more serious than type II error. To appreciate this, consider a legal trial in which a person is accussed of a serious crime. In this example, Type I error corresponds to punishing the defendant when he is actually innocent. On the other hand, Type II error corresponds to setting him free when he is actually guilty. Most of us would agree that it is better to set a criminal free than to punish an innocent guy.

What is the probability that we make Type I error? We reject H_0 whenever the probability of observing a sample assuming H_0 is true is less than the level of significance α. Thus, α is the probability that we make Type I error, and is called as **size** of the test. Can we reduce the size without any issue? When the matters are serious (for example, the punishment leads to death), we must reduce this probability. However, observe that in that case the chances that we would detect the true effect, if it is present, would also reduce because we will be retaining H_0 with high probability. In other words, reducing α actually increases the chance that we commit Type II error! The hypothesis test with $\alpha = 0$ will thus never detect any effect, however strong.

Similar to the probability of committing Type I error, we can talk about committing Type II error, and traditionally it is denoted by β (probably you see why). That is, β is the probability that we reject H_0 even when H_1 is false. Then, $1 - \beta$ is the probability that we reject H_0 when H_1 is really true, and is called the **power of a statistical test**. Stated differently, *statistical power* is the probability of correctly detecting the **effect** when it is present. Here I will not go into the discussion related to calculation of power, but will tell you that traditionally statistical tests with power at least 0.8 are preferred.

8 p-values

An alternative to the kind of hypothesis testing that we have been describing so far, is the usage of what are known as p-values. Here, no level of significance is specified before the collection of data. Instead, when the observation is made and the corresponding test statistic is calculated, we ask what is the probability of observing this outcome or an outcome that is more extreme **assuming the default position** H_0. This probability is called the p-value. Note that this is *different* from the level of significance which is something that we specify *before* analyzing the data.

Let's go back to our cupcake example, but now let us assume that we are performing two-tailed test. That is, we want to see whether the actual mean weight of a cupcake packet deviates (in any direction) from the presumed 200 grams. The Z-score that we get from our data is -2.81. Now notice an important point: *since the test is two-tailed, the extreme scores here correspond to Z-scores below* -2.81 *and those above* 2.81. Hence the p-value in this case is:

$$p = \mathbb{P}(Z \leq -2.81) + \mathbb{P}(Z \geq 2.81) = 2\mathbb{P}(Z \geq 2.81) = 0.005$$

This probability is very small. However, while reporting the results in terms of p-values we don't say that we reject H_0, only that it is unlikely to be true. Of course, we also need to report the corresponding p-value.