Hypothesis Testing

In many situations of uncertainty, we have to make a choice between two alternatives. Given a coin, we might ask if it is fair. In order to help doctors we might have to decide whether a change to the components of a drug improve its performance. We could have to decide whether an additional ingredient in petrol improves cars performances.
The above decision processes usually come down to making a decision about the value of a parameter, or parameters that describe the outcome data both before and after a change. For example, with the coin we might want to test whether it is fair $p = 0.5$ or biased $p \neq 0.5$. Measurements can usually be approximated by a normal distribution and so we might want to test whether its mean has changed, i.e., $\mu = \mu_0$ or $\mu > \mu_0$ where μ_0 is the mean before a certain change to the process was made. This leads to the following definition
A **statistical hypothesis** is an assertion concerning the probability distribution of one or more random variables.

A **simple hypothesis** completely specifies the probability distribution.

A **composite hypothesis** does not completely specify the probability distribution involved. For example, the hypothesis might be about the mean of a normal distribution, when the variance is unknown.
When we are faced with two alternative hypotheses it is usual to call one of them **the null hypothesis**, denoted as H_0 and the other **the alternative hypothesis** denoted as H_1. Generally, the historical parameter value would be called the null hypothesis and a possible change from this the alternative hypothesis. There is, in general, no specific reason to call one hypothesis the null hypothesis and the other the alternative hypothesis.

Example:- In the coin tossing experiment, where we might be interested in fairness the hypotheses might be $H_0: p = 0.5$ with $H_1: p \neq 0.5$.

Having formulated the hypotheses, we need to derive a **statistical test** to decide between the two hypotheses. This consists of obtaining an appropriate random sample from the distribution involved and then choosing a **test statistic**, whose value can be used to choose between H_0 and H_1.

The critical region of a statistical test, denoted by C, is the set of values of the test statistic that lead to the acceptance of the alternative hypothesis H_1 or equivalently the rejection of the null hypothesis H_0.

The acceptance region of a statistical test, denoted by C', is the complement of C and leads to the acceptance of the null hypothesis H_0.

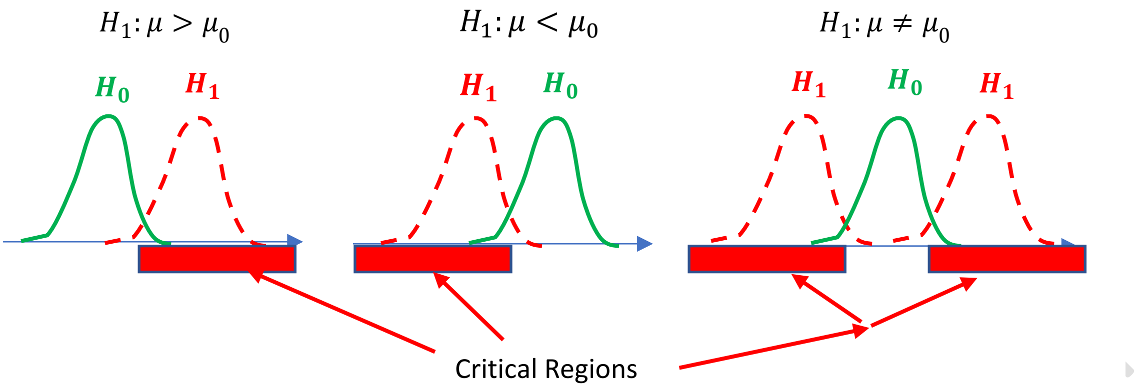
A one-sided test is one where the alternative hypothesis is directional, i.e.,

$$H_o: \mu = \mu_0 \text{ and } H_1: \mu > \mu_0 \quad \text{or} \quad H_o: \mu = \mu_0 \text{ and } H_1: \mu < \mu_0$$

A two-sided test is where the alternative hypothesis is not directional i.e.,

$$H_o: \mu = \mu_0 \text{ and } H_1: \mu \neq \mu_0$$

The critical regions are shaded in red in the figure below for these three tests.



In most cases the critical regions cover values of the test statistic that could have come from the distribution when H_0 or H_1 is true. Hence there is a potential error in the decision making process. **A Type I error** is said to have occurred if you accept H_1 when in fact H_0 is true, and **a Type II** error occurs if you accept H_0 when in fact H_1 is true. This situation is summarised in the table below.

	Accept H_0 (C')	Accept H_1 (C)
H_0 true	Correct decision $(1-\alpha)$	Type I error (α)
H_1 true	Type II error (β)	Correct decision $(1-\beta)$

where $\alpha = P(C|H_0\ true) = 1 - P(C'|H_0\ true)$

and $\beta = P(C'|H_1\ true) = 1 - P(C|H_1\ true)$

Note C and C' are exclusive and exhaustive events and α and β are the probabilities of a Type I and Type II error respectively.

Example:- A coin was suspected of favouring Heads. If p is the probability of a Head, test the hypotheses $H_0: p = 0.5$ against $H_1: p > 0.5$. The coin was tossed 10 times and a critical region of the form $C = \{k, \dots, 10\}$ was used. Determine the value of k that would make the Type I error as close to 5%(0.05) as possible.

If H_0 is true the number of Heads obtained has the distribution B(10,0.5). The Type I error is the probability $\alpha = P(X \in C|H_0 \text{ is true}) = P(X \geq K|H_0 \text{ is true})$. The table below gives the values of α for different values of K.

α	0.0009	0.0107	0.0547	0.1719
K	10	9	8	7

It is clear that α is the closest to 0.05 when K = 8. The desired critical region is therefore $C = \{8,9,10\}$. **Note** the Type II error cannot be calculated because the alternative hypothesis is composite, i.e., no value of p specified.

Significance Tests

We have just seen in the last example that the Type II error probability cannot be calculated because the value of p wasn't specified, just that it was above 0.5. This is quite a common situation and so what can be done? In this situation, assuming the null hypothesis is a simple one, we often specify the value of α, usually the highest value we are prepared to tolerate, and then look for a test with this value of α. In this situation the value of α is called the significance level of the test, but it is still the probability of a Type I error.

Example:- The random variable X has a normal distribution with mean μ and variance 1. Given a random sample of size n, and the Hypotheses

$$H_0: \mu = 2 \quad \text{and} \quad H_1: \mu > 2,$$

which are to be tested using a critical region of the form $C = \{\bar{x} | \ \bar{x} > k\}$. Determine the value of k, in terms of n, to produce a significance level of 10%(0.1).

Under the null hypothesis $\bar{X} \sim N\left(2, \frac{1}{n}\right)$ or $Z = \frac{(\bar{X}-2)}{1/\sqrt{n}} \sim N(0,1)$. The significance level α is

$$\alpha = P(\bar{X} > k | H_0 \ true) = P\left(Z > \frac{(k-2)}{1/\sqrt{n}}\right) = 1 - \Phi\left(\frac{(k-2)}{1/\sqrt{n}}\right) = 0.1$$

$$\text{or} \qquad \Phi\left(\frac{(k-2)}{1/\sqrt{n}}\right) = 1 - 0.1 = 0.9 \quad i.e., \quad \frac{(k-2)}{1/\sqrt{n}} = 1.282 \quad \text{or} \quad k = 2 + \frac{1.282}{\sqrt{n}}$$

The critical region for this significance test is therefore $C = \left\{\bar{x} | \ \bar{x} > 2 + \frac{1.282}{\sqrt{n}}\right\}$.

If a 5% significance level was requested the critical region would be $C = \left\{\bar{x} | \ \bar{x} > 2 + \frac{1.645}{\sqrt{n}}\right\}$. Thus, if something was significant at the 10% level it wouldn't necessarily be significant at the 5% level, but if it was significant at the 5% it would be significant at the 10% level.

P – values

In the above we have identified the critical region to perform a hypothesis test. An equivalent method is to identify the probability of the test statistic being at least as extreme as the value observed, when the null hypothesis is true. Extreme here means in the direction favouring the alternative hypothesis.

Example:- Suppose you are testing the hypotheses $H_0: \mu = 0$ against $H_1: \mu > 0$, and your observed sample mean $\bar{x} = 0.5$, based on a sample of size 9 from the N(μ,1). The sample mean $\bar{X} \sim N\left(\mu, \frac{1}{9}\right)$. Extreme here means at least equal to 0.5 and so the p value is

$$P(\bar{X} \geq 0.5 | H_0 \text{ true}) = P(\bar{X} \geq 0.5) \text{ when } \bar{X} \sim N\left(0, \frac{1}{9}\right)$$

$$= 1 - \Phi\left(\frac{0.5-0}{\sqrt{\frac{1}{9}}}\right) = 1 - \Phi(1.5) = 1 - 0.933 = 0.067$$

If the significance level was 5% the critical region would be all those points for which $P(\bar{X} \geq 0.5 | H_0 \text{ true}) \leq 0.05$. The observed mean would not be in this critical region and so we accept the null hypothesis.

The statistical framework outlined so far will apply to any sampling distribution, we have restricted our examples to binomial and normal probability settings.