Testing Hypotheses

This is the problem of choosing between two theories, two models or two hypotheses.

(1) H_0 or the Null Hypothesis: This is the established hypothesis saying that the standard model is true or correct. If we are testing a new treatment against the current standard, H_0 denotes no difference between them.

Example. In a quality control setup, suppose that the procedure in use produces 2% defectives on average. Then new quality control measures are brought in and there is a claim of reduction in the proportion (p) of defectives. Then, we have

$$H_0 : p = 0.02 \text{ i.e., no change has occured.}$$

This effectively means that any observed changes can be explained as due to purely chance variations, and a new model is not needed.

(2) The hypothesis to be tested against H_0 is called H_1 or the Alternative Hypothesis.

In the example above,

$$H_1 : p < 0.02 \text{ i.e., a positive change has occured.}$$

This effectively means that the observed changes cannot be explained as due to chance variations under H_0 and a new model is needed.

Question. Is there enough evidence in the data to reject H_0 in favour of H_1?
We proceed as follows by considering the consequences of actions in a test procedure under the different possible states of nature.

	decision	
	accept H_0	reject H_0
H_0 is true	✓	type I error
H_1 is true	type II error	✓

Decisions (or actions, accept or reject H_0) are made using evidence from random samples or data. Therefore, we can only compute the probabilities of errors, and not know when they are committed. As shown above, *Type I Error* is the incorrect rejection of H_0, and
$P($ Type I Error $) = \alpha =$ level of significance.

Type II Error is the incorrect acceptance of H_0, and
$P($ Type II Error $) = \beta$; $1 - \beta =$ power of test.

Our approach is to fix α (at say, 0.05 or 0.01) and minimize β (or maximize the power, $1 - \beta$) to get the "most powerful" tests.

Let us consider some examples to intuitively see

- how to derive test statistics;
- how to derive test criteria.

Example. A pack of a certain brand of cigarettes displays the statement, "1.5 mg nicotine on average per cigarette". Let μ denote the actual average nicotine content per cigarette for all cigarettes of this brand. It is required to test if the actual average is higher than what is claimed. Suppose a sample of cigarettes is selected, and the nicotine content of each cigarette is determined. The observed contents are $X_1, \ldots, X_n$. Conduct the test at the level of significance of 5%. We assume that X_i are distributed as i.i.d $N(\mu, \sigma^2)$. Then, we desire to test

$$H_0 : \mu = 1.5 = \mu_0 \text{ versus } H_1 : \mu > 1.5 \text{ i.e., average is higher.}$$

How to we calculate evidence? Estimate μ from data. $\hat{\mu} = \bar{X}$. Compare $\hat{\mu}$ with μ_0. When do we say that H_0 is not true?

$\bar{X} - \mu_0 =$ observed difference between estimated mean and the hypothesized value.

However, $\bar{X}$ has variation from sample to sample. The expected variation in $\bar{X}$ values is s.e.$(\bar{X}) = \frac{s}{\sqrt{n}}$.
Now compare the observed difference with the expected, and enquire: Is

$$\frac{\bar{X} - \mu_0}{s/\sqrt{n}} >> 1?$$

i.e., is the observed difference much larger than the expected difference?

If so, reject H_0. If

$$\frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim 1 \text{ or } < 1$$

there is no evidence to reject H_0. Thus, we have the statistic for testing:

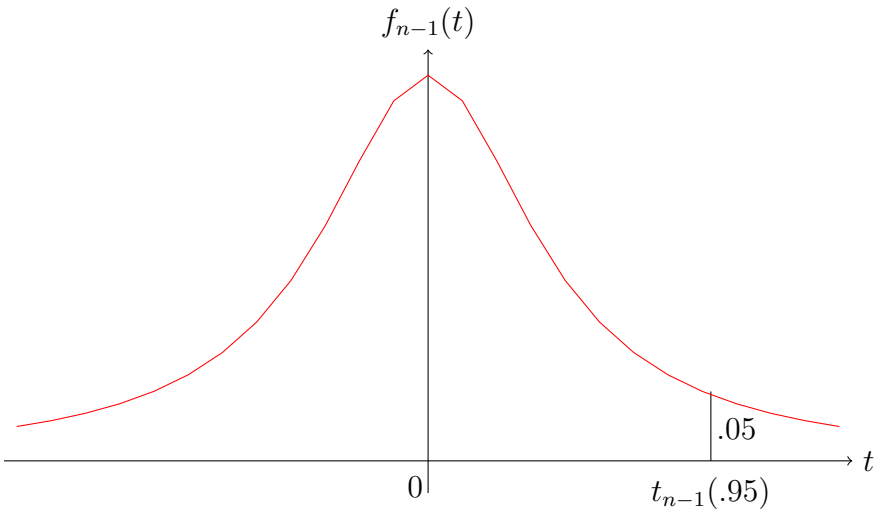
$$\text{Test statistic } = \frac{\text{observed departure from } H_0}{\text{expected departure}}.$$

In the present problem, the test statistic is

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1} \text{ if } H_0 \text{ is true.}$$

Reject H_0 if the observed value of the test statistic is large. But, how large?

$$\begin{aligned} 0.05 &= \alpha = P(\text{ Type I Error }) = P(\text{ Reject } H_0 \text{ when } H_0 \text{ is true }) \\ &= P\left(\frac{\bar{X} - \mu_0}{s/\sqrt{n}} > C\right) = P\left(t_{n-1} > C\right). \end{aligned}$$



Therefore, $C = t_{n-1}(1-\alpha) = t_{n-1}(.95)$ and,
we claim to have evidence against the null hypothesis at the 5% level of significance
if the observed value of $\frac{\bar{X}-\mu_0}{s/\sqrt{n}} > t_{n-1}(.95)$.

Example. $X \sim \text{Binomial}(n, p)$ and it is of interest to test

$$H_0 : p = p_0 \text{ versus } H_1 : p \neq p_0 \text{ a two-sided alternative.}$$

$\hat{p} = \frac{X}{n}$ and it would be natural to use the statistic:

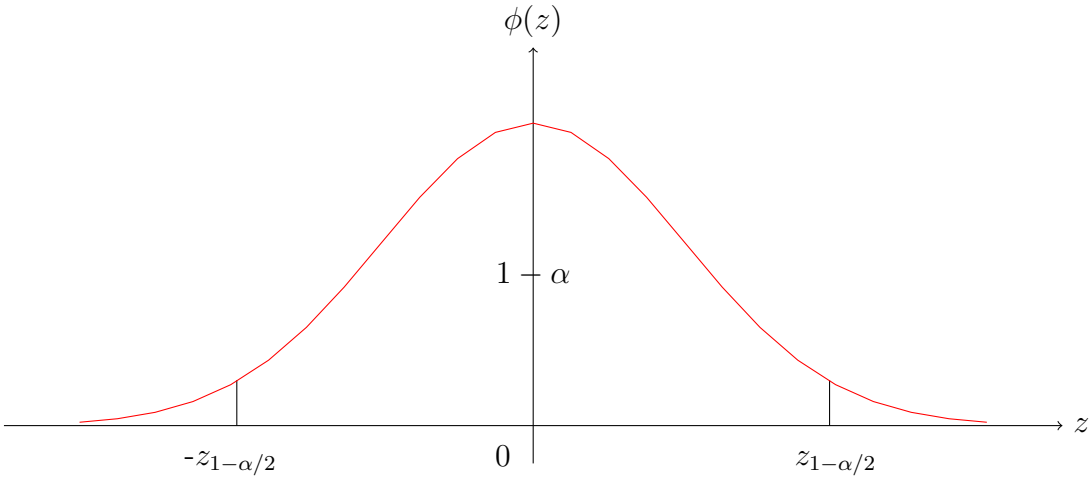
$$\frac{\text{estimated departure from } H_0}{\text{expected departure or s.e.}} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0,1),$$

approximately for large n, if H_0 is true. Since the alternative hypothesis is two-sided, the test statistic is $\left|\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}\right|$, large values of which provide

evidence against H_0 and in favour of H_1. To obtain the crtitical value (above which H_0 is to be rejected), note

$$\alpha = P_{H_0}\left(\left|\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}\right| > C\right) = P(|Z| > C),$$

implying that $C = z_{1-\alpha/2}$.



At the significance level of α, evidence to reject H_0 exists
if $|\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}| > z_{1-\alpha/2}$.

How does one derive 'most powerful' tests?

Power of a test. As noted previously, power of a test is P_{H_0}(reject H_0). Suppose $X_1, \ldots, X_n$ are i.i.d $N(\mu, \sigma^2)$ where σ^2 is known and we desire to test

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu > \mu_0$$

at the level of significance α. We use the test statistic

$$Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \text{ if } H_0 \text{ is true.}$$

Reject H_0 if the observed value of $Z > z_{1-\alpha}$ since

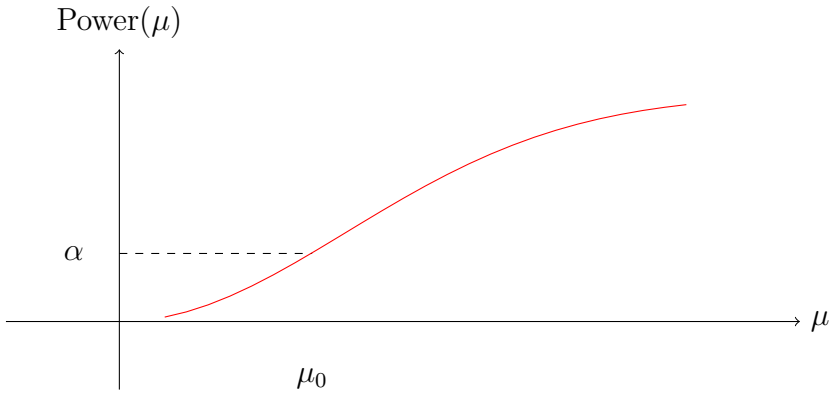
$$\alpha \;=\; P_{\mu=\mu_0}\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha}\right).$$

To compute the power of this test at any $\mu > \mu_0$ (under H_1), we have,

$$\begin{aligned}\text{Power}(\mu) &= P_\mu\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha}\right) = P_\mu\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} + \frac{\mu-\mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha}\right) \\ &= P_\mu\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > z_{1-\alpha} - \frac{\mu-\mu_0}{\sigma/\sqrt{n}}\right) = P\left(Z > z_{1-\alpha} - \frac{\mu-\mu_0}{\sigma/\sqrt{n}}\right),\end{aligned}$$

which is an increasing function of μ. i.e., if μ is much larger than μ_0, the test rejects H_0 easily.

Power(μ) as a function of μ is called the power curve.



P-values. The error probabilities of a test (significance level α and the power $1-\beta$ which are predetermined) do not provide a measure of the strength of evidence in a particular data set against H_0. The P-values defined below try to capture that.
Suppose $H_0 : \theta = \theta_0$ and your test is to reject H_0 for large values of a test statistic $W = W(\mathbf{X})$. Then, when $\mathbf{X} = \mathbf{x}$ is observed, the P-value is defined as

$$p(\mathbf{x}) = P_{\theta_0}\left(W > W(\mathbf{x})\right),$$

Any data point $\mathbf{x}$ for which $p(\mathbf{x}) \leq \alpha$ may be considered strong enough eveidence to reject H_0 at the significance level of α.

Example. Let $X_1, X_2, \ldots, X_9$ be i.i.d $N(\mu, 1)$. It is of interest to test $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$. Compute P-value for the Z-test if $\bar{x} = 0.9$ is observed. The Z-test rejects H_0 when

$$|Z| = \left|\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right| = 3|\bar{X}|$$

is large. $Z \sim N(0, 1$ when H_0 is true. Therefore the P-value when $\mathbf{x}$ is observed is $p(\mathbf{x}) = P(|Z| > 3\bar{x})$. If $\bar{x} = 0.9$, then the P-value is $2[1 - \Phi(2.7)]$ $= 0.007$.

How do we know that these are the 'best' tests avaiable to us? Let us discuss the Neyman-Pearson theory of deriving optimal tests for this.