Simple Overview of Hypothesis Testing

C.W.Cleghorn

Hypothesis Testing

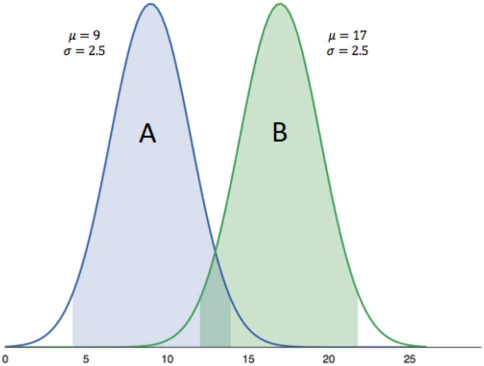
Often times we are not able to analytically prove a result. As such we generally

- Have an hypothesis
  - Algorithm A is faster than Algorithm B
- We then, via experimentation, try and determine if our hypothesis holds our not.

Hypothesis Testing

The question then is how much evidence is needed to ascertain if our hypothesis is correct?

- This is were hypothesis testing comes in.
- Imagine the below image represents the execution time of algorithms A and B
- How much overlap is tolerable, for our hypothesis to hold?



Hypothesis Testing: The Ingredients

In order to perform hypothesis testing we need the following components

The Ingredients

- **Null hypothesis**, h_0
- **Alternate hypothesis**, h_a (often labeled h_1)
  - Sometime called the **research hypothesis**
- **Test statistic**
- **Rejection region**

Hypothesis Testing: The Ingredients

The idea is:

- Use evidence that supports the **Alternate hypothesis**, h_a
  - To test if we can **Reject** the **Null hypothesis**, h_0
  - This setup **cannot** be used to **prove** the **Null hypothesis**, h_0

Conditions on h_0, and h_a

- $h_0 \cap h_a = \emptyset$
- Though sometimes $h_0 \cup h_a$ is all possibilities, this is **Not** actually a requirement

Hypothesis Testing: Error Types

There are two type of error that can occur during hypothesis testing

- **Type 1**
  - h_0 is **rejected** when h_0 was in fact **true**
  - We indicate the probability of this error is denoted as α
  - α is your level of significance, $1 - \alpha$ is you confidence level
- **Type 2**
  - h_0 is not **rejected** when the alternative hypothesis h_a is in fact true
  - We indicate the probability of this error is denoted as β
  - The power of a test is judged by $1 - \beta$.

Hypothesis Testing: Error Types

Table of error types		**Null hypothesis (H_0) is**	
		True	**False**
Decision about null hypothesis (H_0)	**Fail to reject**	Correct inference (True Negative) (Probability = 1 - α)	Type II error (False Negative) (Probability = β)
	Reject	Type I error (False Positive) (Probability = α)	Correct inference (True Positive) (Probability = 1 - β)

Hypothesis Testing: The Test Statistic

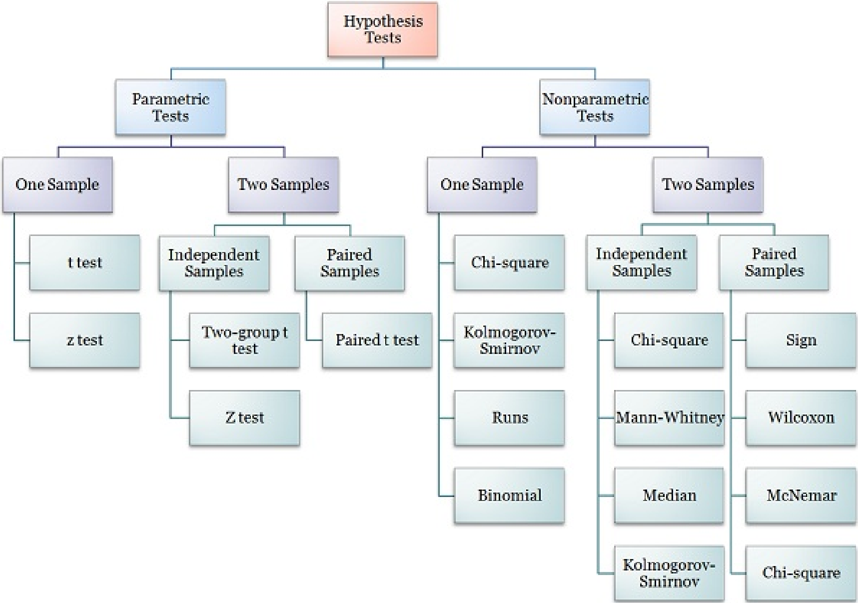
There are a number available test statistics. The two broad classes are

- Parametric
  - Assumes specific distribution on the data (generally the normal distinction)
- Non-parametric
  - Does **not** assume specific distribution on the data
  - Is however, **not** assumption free!

An important point is that parametric test are **stronger** at the cost of **more** assumptions.

  - If you can meet the assumptions (pre-conditions) parametric tests then use them.

Test Statistic: The Family



Test Statistic: One Sample versus Two Samples

One you have decided on parametric versus non-parametric you need to determine if you need a

- One sample test or a two samples test

The **one sample test**

- You have a stated population mean μ (Jeff states the average height is 180cm)
- You have drawn a set of samples of the data (polled n people), and you wish to check if
  - The proposed μ is actually correct

Test Statistic: One Sample versus Two Samples

The **two samples test**

- You have polled n_1 people from South Africa about their height
- You have polled n_2 people from Namibia about their height
- You now what to ask questions of the form (each would be a test)
  - Is the average height in South Africa the same as in Namibia?
  - Is the average height in South Africa greater than in Namibia?
  - Is the average height in South Africa less than in Namibia?

One Sample Parametric Test

The most common one sample parametric tests are

- z-test
  - Assumes normality
  - Assumes you know the **true** population variance
- t-test
  - Assumes normality
  - Uses the sample variance instead of the unknown population variance

One Sample T-Test Classic Example

Consider the following set up:

- An online store advertises that its average delivery time is less than six hours for local deliveries. A random sample of the amount of time taken to deliver packages:
  - 7 3 4 6 10 5 6 4 3 8

  Is there sufficient evidence to support the stores advertisement, at the 5% level of significance?
- A simple, but incomplete, approach would be to just check the average of the samples.
  - which gives as a $\bar{x} = 5.6$, but is their claim correct?
  - we need to use a more powerful approach.

One Sample T-Test Classic Example: The Set Up

We begin by stating the null and alternative hypothesis

- h_0: $\mu = 6$
- h_a: $\mu < 6$

We now use our **test statistic**, the t-score in this case

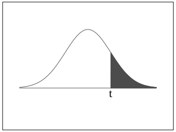
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{5.6 - 6}{2.3/(\sqrt{10})} = -0.557$$

where s^2 is just the sample variance. Now a library would be able to tell us how this maps to a p value and if $p \geq 0.95$ we would reject the null hypothesis. There is a way to calculate this, but the typical way is to use a t-table.

- We are doing a **lower** 1-tail test here
- So if $t < -t_{0.05}$ in the table we reject h_0.

One Sample T-Test Classic Example: T-Table

t-Distribution Table



The shaded area is equal to α for $t = t_\alpha$.

df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845

One Sample T-Test Classic Example: The Set Up

We have

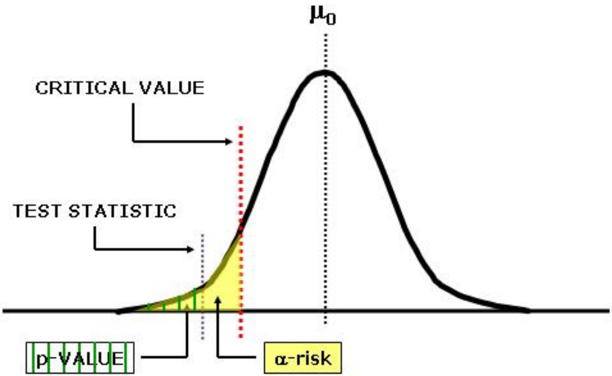
- $df = n - 1 = 9$
- Significance level of 0.05
- So the $t_{0.05}$ value is 1.833
- Our $t = -0.557$ is not less than $-t_{0.05}$
  - We don't have enough evidence to reject our null hypothesis

One Sample T-Test

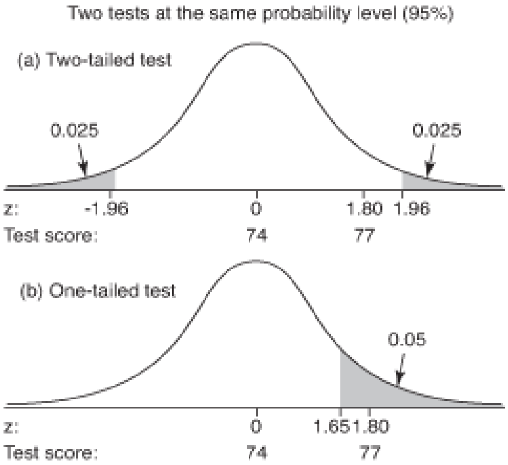
General Form

- $h_0 : \mu = \mu_0$
- $h_a : \begin{cases} \mu > \mu_0 \text{ (upper tail)} \\ \mu < \mu_0 \text{ (lower tail)} \\ \mu \neq \mu_0 \text{ (two tailed)} \end{cases}$
- Test statistic: $t = \frac{\bar{x} - \mu_o}{s/\sqrt{n}}$
- Rejection region: $\begin{cases} t > t_\alpha & \text{(upper tail RR)} \\ t < -t_\alpha & \text{(lower tail RR)} \\ |t| < t_{\alpha/2} & \text{(two-tail RR)} \end{cases}$

One Sample T-Test



One Sample T-Test



Two Sample T-Test (Independent Samples)

General Form

- Assumptions: Independent samples from normal distributions with $\sigma_1^2 = \sigma_2^2$
- $h_0 : \mu_1 - \mu_2 = D_0$
- $h_a : \begin{cases} \mu_1 - \mu_2 > D_0 \text{ (upper tail)} \\ \mu_1 - \mu_2 < D_0 \text{ (lower tail)} \\ \mu_1 - \mu_2 \neq D_0 \text{ (two tailed)} \end{cases}$
- Test statistic: $t = \frac{\bar{x}_1 - \bar{x}_2 - D_0}{s_p\sqrt{(1/n_1 + 1/n_2)}}$, where $s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$
- Rejection region: $\begin{cases} t > t_\alpha & \text{(upper tail RR)} \\ t < -t_\alpha & \text{(lower tail RR)} \\ |t| < t_{\alpha/2} & \text{(two-tail RR)} \end{cases}$

Two Sample T-Test (Paired Samples/ Dependent Samples)

Consider the following situation:

- You have a training program T for a company, call the set of staff S
- You want to demonstrate the effectiveness of T
- Now you cannot use average performance increase between S and $T(S)$, because the improvement induced by T is Dependant on the person who went through the training program T.
- We are actually interested in the sample pair $(s, T(s))$, where $s \in S$

We follow the same procedure as before, just with the **paired t-test** test statistic,

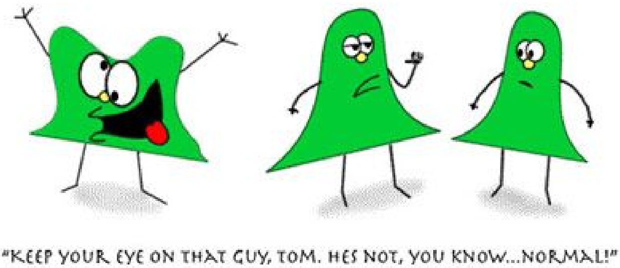
- You are basically using $T(s) - s$ as a single distribution about which you want to test hypotheses.

Can I use Parametric Statistics on My data?

As with all statistical tests you must

- Test if the needed assumptions hold!
- The first test you should perform is to confirm whether or not the data is normally distributed or not.
- Lots of option available, some of the very popular ones are
  - Anderson–Darling test
  - Shapiro–Wilk test
- Check what the standard in your field is!

So your data is not normal?



Just use the other branch of the tree!

Test Statistic: The Family