Mu Sigma

Hypothesis Testing

Day 4

Do The Math

Chicago, IL
Bangalore, India
www.mu-sigma.com

2016

Contents

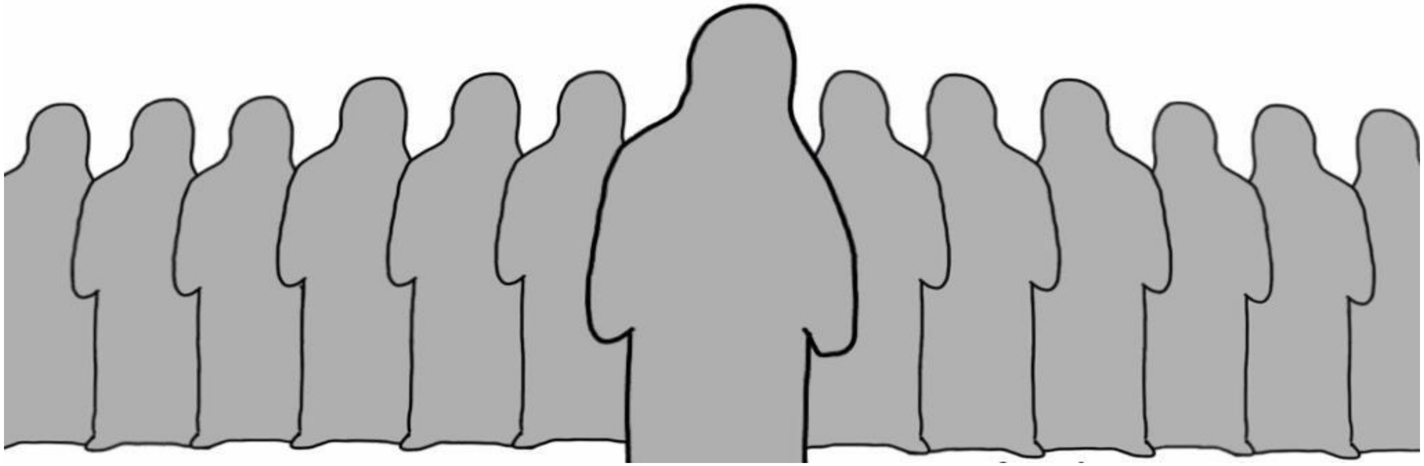
- Introduction to Hypothesis Testing
- Approach to Hypothesis Testing
- One tail and two tailed tests
- Z test
- t test

CATBERT: EVIL DIRECTOR OF HUMAN RESOURCES
I CAN'T DECIDE WHAT'S CHEAPER...
...AN EMPLOYEE WELLNESS PROGRAM TO REDUCE SICK DAYS OR INCENTIVIZING THE OLDER, HIGHLY PAID WORKERS TO DIE.
MAYBE YOU COULD USE MATH TO FIGURE IT OUT.
WHEN I SAID CHEAPER, I MEANT MORE FUN.
www.dilbert.com
scottadams@aol.com
9-27-04 © 2004 Scott Adams, Inc./Dist. by UFS, Inc.

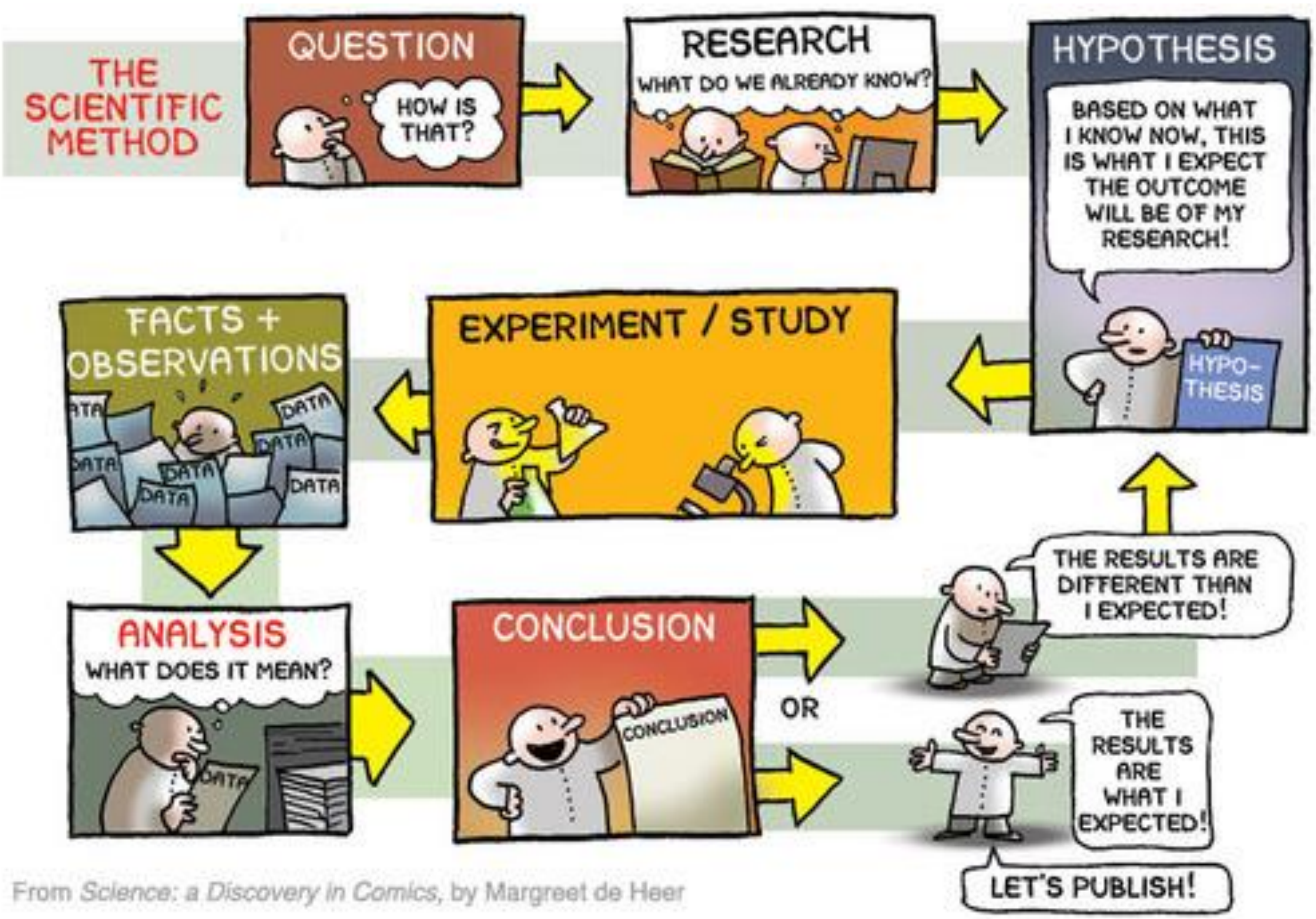
© UFS, Inc.

I am what is
The default, the status quo
I am already accepted, can only be rejected
The burden of proof is on the alternative

I am the null hypothesis



The scientific method



From *Science: a Discovery in Comics*, by Margreet de Heer

Motivation for testing hypothesis

- The intent of hypothesis testing is to formally examine two opposing conjectures (hypotheses) H_0 and H_1

- H_0 is referred to as the null hypothesis which can be your *status quo*

- H_1 is the alternative hypothesis – which has been put up against the null hypothesis

- These two hypotheses are mutually exclusive so that one is true to the exclusion of the other

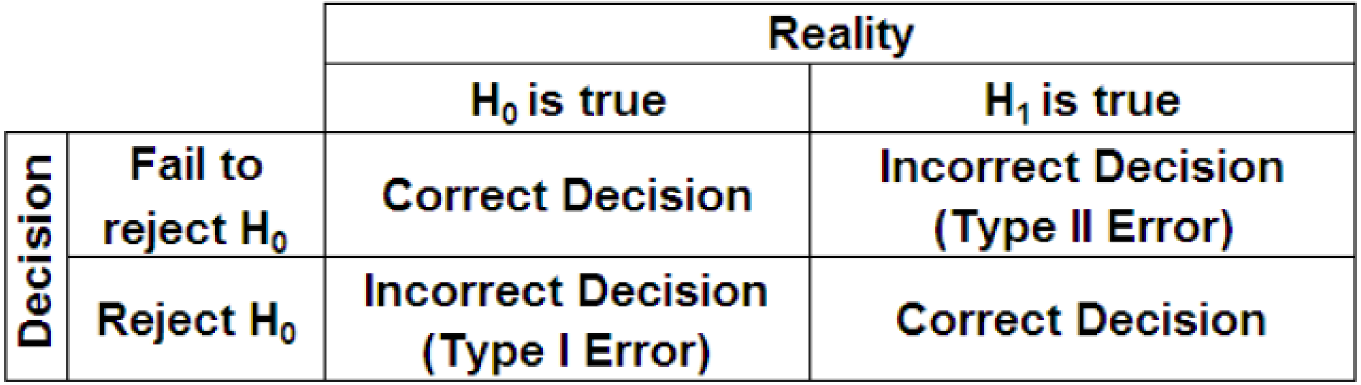
- It is important to remember that hypotheses are always statements about the population or distribution under study, not statements about the sample

- The hypotheses may result from past experiences or knowledge of the process or even from previous experiments

A test of hypothesis is a procedure leading to a decision

- Hypothesis testing procedures rely on using information in a random sample from population of interest
- If the information is consistent with the hypothesis, we will conclude that the hypothesis is true
- However, if this information is inconsistent, we will conclude that the hypothesis is false
- Truth or falsity of a particular hypothesis can never be known with certainty, unless we can examine the entire population

Any decision making process is in danger of errors



		Reality	
		H_0 is true	H_1 is true
Decision	Fail to reject H_0	Correct Decision	Incorrect Decision (Type II Error)
	Reject H_0	Incorrect Decision (Type I Error)	Correct Decision

- Two kinds of errors exist in any decision making process
- Rejecting the null hypothesis (H_0) when it is true is defined as **Type I** error (α)
- Failing to reject the null hypothesis when it is false, is defined as **Type II** error (β)

Case Study: What are the chances of putting an innocent prisoner behind bars?

So what does putting prisoners in trial have to do with hypothesis testing?

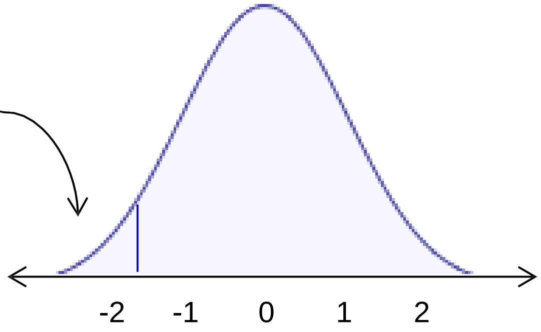
A Type I error is when you reject Ho when actually it is correct

Type 1 error is like putting an innocent person behind the bars.

Type 1 error occurs when null hypothesis is rejected but it is true

A Type II error is when you accept H0 When it is actually wrong

If you get a type I error , your test statistic must be here in the critical region



Type 2 error would result in letting a guilty person go innocent

Mr. X claims that he is able to distinguish between Diet Coke and Coke

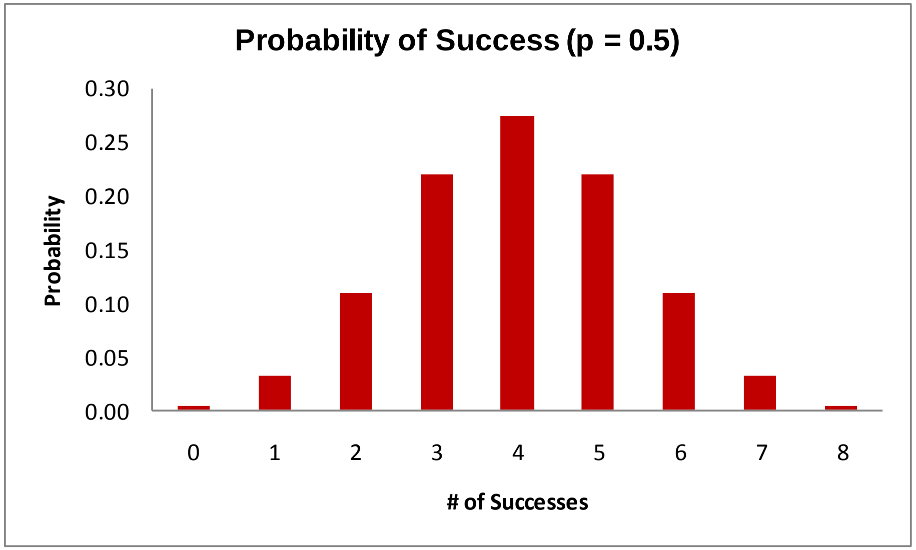
- To test this claim, 8 cups were offered to him to taste and make a decision
- In each cup, there would be either coke or diet coke - which Mr. X would have to guess after tasting
- If Mr. X is able to guess 6 out of 8 cups - can we say Mr. X has the ability to distinguish between the two types of drinks?
- What is the procedure to test the claim?

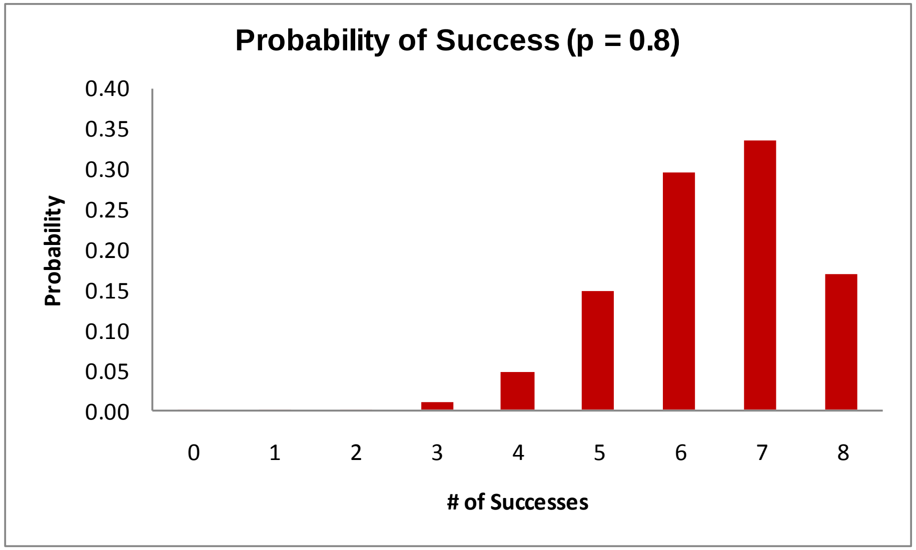
The underlying probability distribution in this example is Binomial with 8 trials and probability of success $\boldsymbol{\theta}$

- The probability of ***k*** successes in 8 trials is given by

$$P(X=k)=\binom{8}{k}\theta^{k}(1-\theta)^{8-k}$$

where $\boldsymbol{\theta}$ is the probability of correctly making a guess





We evaluate the chance of observing the data obtained under the null hypothesis

- Our null hypothesis is H_0: $\theta = 0.5$, i.e., Mr. X does not possess any ability to distinguish between Diet Coke and Coke
- Alternative hypothesis being, H_1: $\theta > 0.5$. We do not consider $\theta < 0.5$, since it is not meaningful in this context
- Moreover we would like our significance level to be at α = 0.05, i.e., we expect to make wrong decisions not more than 5% of the times
- With the above assumptions, we calculate the chance of observing the data as extreme as it were (at least 6 successes out of 8 trials)

$$\begin{aligned} P_{H_0}(X \geq 6) &= \sum_{k=6}^{8} \binom{8}{k} (0.5)^k (1-0.5)^{8-k} \\ &= 0.10938 + 0.03125 + 0.00391 \\ &= 0.14453\,(P - value) \end{aligned}$$

We reject our null hypothesis if our p-value is less than the significance level

- Our test of null hypothesis is completely defined with the critical region for the test – i.e., the region of rejection
- We would like to reject our null hypothesis at level α if $P_{H_0}(X \geq k) \leq \alpha$
- The critical point or region in this case is identified by the point *k* which satisfies the above equation
- In our example, Mr. X has been able to guess six out of eight cups correctly
- With a p-value of 0.14, the chance of getting six or more correct answers by simply guessing is too high to make us accept Mr. X’s claim
- Hence in the light of the given data, we are unable to reject our null hypothesis θ = 0.5, i.e., Mr. X is simply guessing

A manufacturer claims, burning rate of a particular type of propellant is 50 cm/sec

- If we wish to challenge such a claim, we frame our hypothesis as

$$H_0 : \mu = 50 \text{ cm/sec}$$
$$H_1 : \mu \neq 50 \text{ cm/sec}$$

- Suppose the standard deviation of burning rate is σ = 2.5 cm/sec

- To test the hypothesis, we draw a sample of n = 10 specimens and calculate the average burning rate $\overline{X}$

- What is the critical region for the above test?

We assume that the burning rate has normal distribution

- From the sampling distribution of the sample mean we obtain, the average burning rate

$$\bar{X} \sim N\left(\mu, \sigma/\sqrt{n}\right)$$

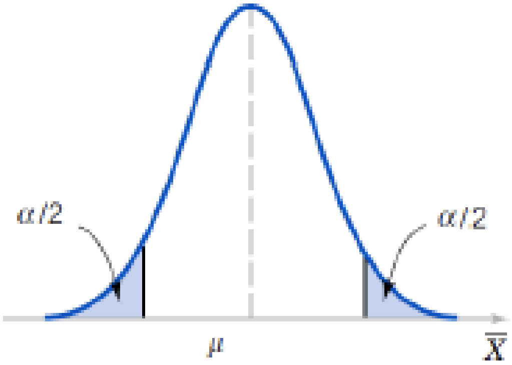
- Our critical region for the test is defined as at the point $Z_{\alpha/2}$ such that

$$P_{H_0}\left(\left|\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right| \geq Z_{\alpha/2}\right) = \alpha$$

  i.e., we find a point above which the probability is α/2

- From the standard normal probability distribution table (μ = 0, σ = 1) we find for α = 0.05, $Z_{\alpha/2}$ is 1.96

- We reject H_0 at 5% level if $\left|\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right| \geq 1.96$



Case Study: SnoreCull claims to cure snoring with a 90% success rate

Is your snoring getting you down
Then you need new Snorecull.
The ultimate remedy for snoring.

SNORECULL cures 90%
of snores within 2 weeks

CULL THOSE SNORES WITH NEW SNORECULL

A doctor, well known in NYC for her skills to cure snoring is not convinced by the claim of the drug company and decides that she will conduct a test on 100 people

Here are the results

Cured	Yes	No
Frequency	80	20

I'm not sure if the claims are true. Had they been, more of my patients would be cured

To conclusively say that the claims are false, we would need to perform a hypothesis test

The claims of the drug company are false and people will continue to snore!!

It was a bad day for the doctor and the sample was biased. Those were the 100 best snorers NYC has !!

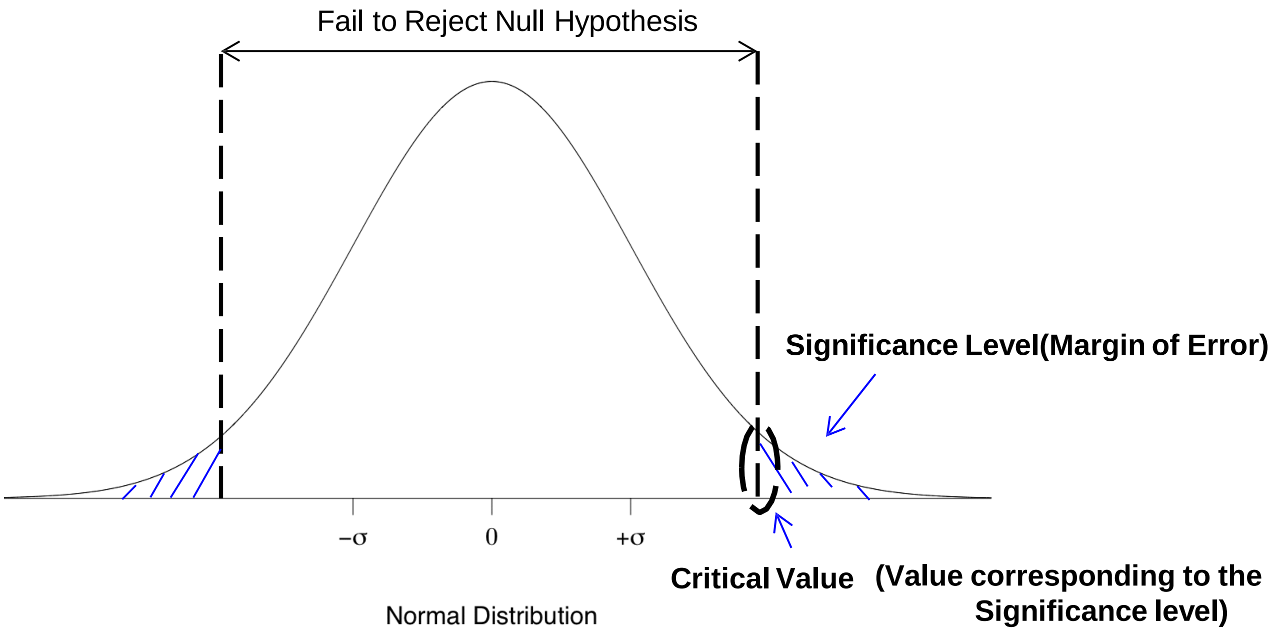
(-Also our Null Hypothesis)

Does that mean that the drug company is lying? Shouldn't the drug have cured more patients ?

But can we really be certain that the drug company is at fault ? Maybe the doctor was unlucky.

Guys,you need to do a "Test of Hypothesis".

The relation between significance, critical and p-value


Fail to Reject Null Hypothesis
Significance Level(Margin of Error)
−σ
0
+σ
Critical Value
(Value corresponding to the Significance level)
Normal Distribution

- p-value is the probability, that the observed values will be equal or more extreme than the critical value
- Lower the p-value, stronger is the evidence against p-value

Hypothesis testing is the first and most essential part of statistical inference

Steps in Hypothesis Testing

- **Decide on the hypothesis you are going to test**
  - The claim that we are putting on trial
- **Choose the test statistic**
  - A method that best suits to test the claims
- **Determine the critical region**
  - Level of certainty we are looking for, before we accept or reject the claim
- **Find the p-value of the test statistic**
  - What is the probability of the null hypothesis being actually true
- **See whether the test result is within the critical region**
- **Make decision**

One tailed tests help in making specific guess unlike two tailed tests

One Tailed Test

One tailed tests are what you use when you have a specific guess

- Earthquakes are more frequent in Japan then in America
- Kids like chocolates more than adults

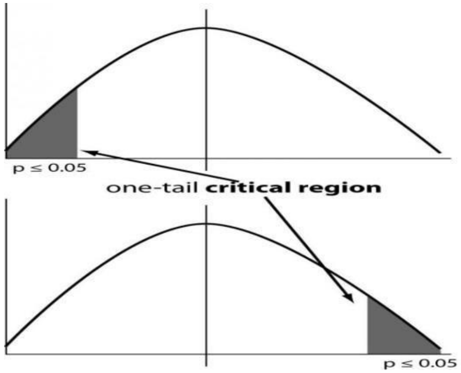
Vs.

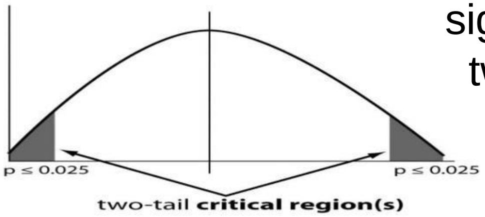
Two Tailed Test

Two tailed tests are used when you can't make a guess

- Number of deaths caused by Cancer and Aids are different
- Baseball and Soccer are not the same fun!

In a one tailed test, the significance level is tested in only one direction of the tail





In a two tailed test, the significance level is split into two parts & tested in both the directions of the tail

Z test: Known Variance-Small sample

The Concept

When to be used?

Is the the mean of the population known? ✓

Is the standard deviation of the population known? ✓

Does the population follow a normal distribution ✓

The *z*-test can be used to compare a sample mean to an accepted mean (one sample test) or to compare proportion of two groups

For a one sample test Z is defined as

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

$\bar{X}$ The average of the sample
μ The population mean
σ The standard deviation
n: The sample size

This follows a standard normal distribution i.e. a normal distribution with mean 0 and standard deviation 1, usually written as Z ~ N(0,1)

Hence it is governed by the properties of a normal distribution

t test: Unknown Variance-Small sample

The Concept

When to be used?

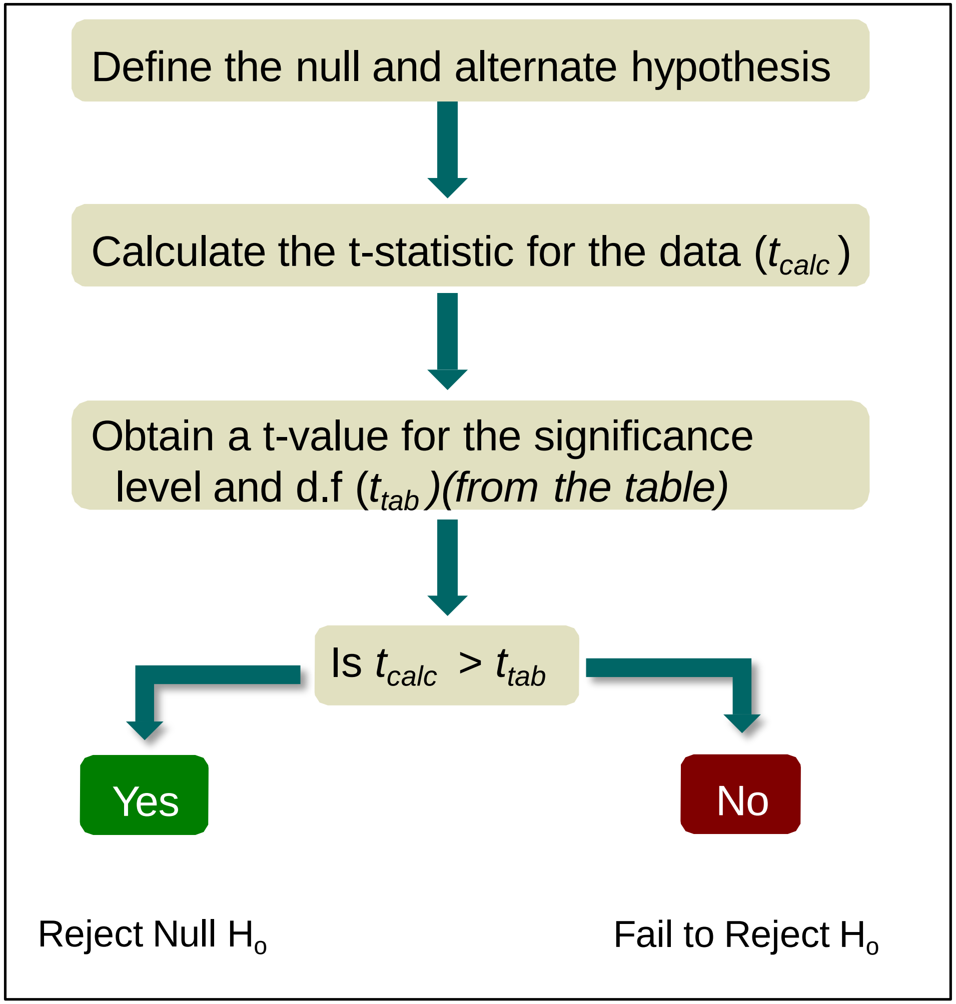
Is the mean of the population known? ✓

Is the standard deviation of the population known? ✗

Is the standard deviation of sample known? ✓

Does the population follow a normal distribution ✓

The *t*-test can be used to compare a sample mean to an accepted mean or the compare means of two groups



Define the null and alternate hypothesis

Calculate the t-statistic for the data (t_{calc})

Obtain a t-value for the significance level and d.f (t_{tab})*(from the table)*

Is $t_{calc} > t_{tab}$

Yes

No

Reject Null H_o

Fail to Reject H_o

Appendix

Z test : Illustration

Problem

- In a supermarket chain, in the year 2010 the average weekly sales was $485,000 with a standard deviation of 12,000 among stores with an area of about 40,000 sq. ft.
- In the first week of 2011, a random sample of 6 such stores is taken
- Examine if the average sales in all stores of this kind in the first week of 2011 has increase

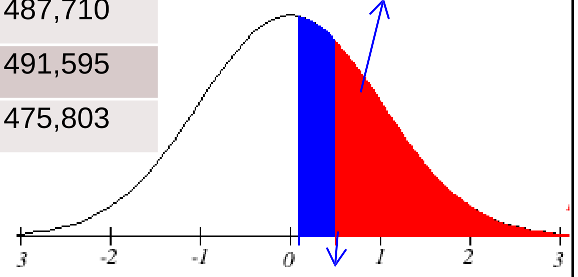
Solution

- Assumptions: Distribution of sales in first week of 2011 in the chain is normal with mean µ and s.d. 12,000
- H_0: µ = 485,000; H_A: µ > 485,000
- Under Ho: Z = (X-485,000)/(12,000/√6) ~ N(0,1)
- Observed value of Z is (when X=485,098) 0.003334
- p-value = P(Z > 0.003334) = 0.4987 = 49.87%
- This is quite large and so we do not reject H_0

  (For 5% Level of Significance the p-value has to be < 0.05)

  See the impact of std. dev. and sample size !!

Store	Sales
A	484,849
B	480,594
C	490,036
D	487,710
E	491,595
F	475,803



P = 0.4987

Z = 0.003334

Z test : lets take a different case

Problem

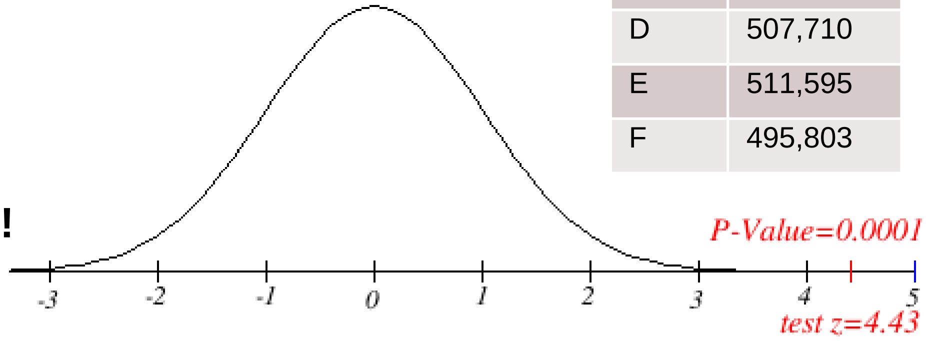
- In a supermarket chain, in the year 2010 the average weekly sales was $485,000 with a standard deviation of 12,000 among stores with an area of about 40,000 sq. ft.
- In the first week of 2011, a random sample of 6 such stores is taken
- Examine if the average sales in all stores of this kind in the first week of 2011 has increased

Solution

- Assumptions: Distribution of sales in first week of 2011 in the chain is normal with mean µ and s.d. 12,000
- H_0: µ = 485,000; H_A: µ > 485,000
- Under Ho: Z = ($\bar{X}$-485,000)/(12,000/√6) ~ N(0,1)
- Observed value of Z is (when $\bar{X}$=506,764) 4.43
- p-value = P(Z > 4.43) <0.0001
- This is very small and so we do reject H_0

See the impact of std. dev. and sample size !!

Store	Sales
A	514,849
B	510,594
C	500,036
D	507,710
E	511,595
F	495,803



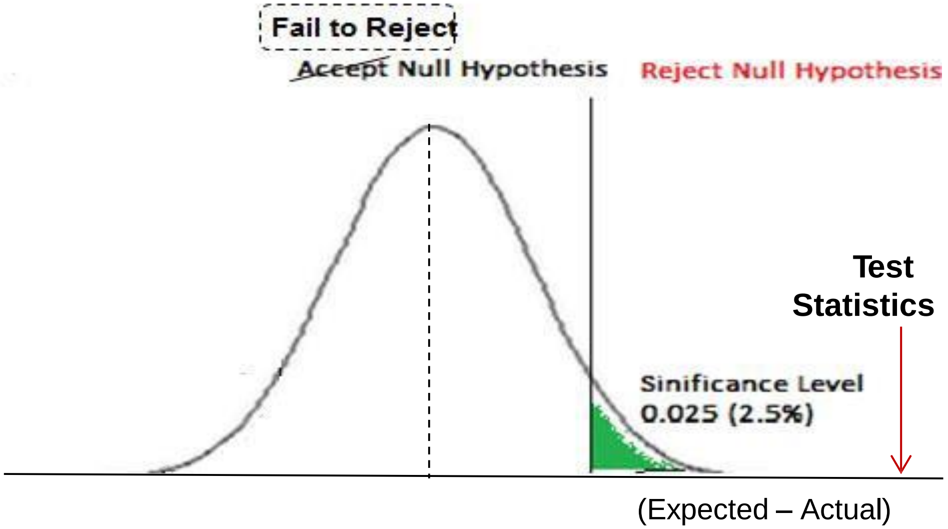
Building Blocks: Test statistic, Significance level....

Test Statistic

- A test statistic is a measure calculated from a sample of data. E.g. (Actual Cure – Expected Cure)
- The choice of a test statistic will depend on the hypothesis under question, population characteristic etc. and will be assumed to follow a pre-defined distribution
- Its value is used to decide whether or not the null hypothesis should be rejected in a hypothesis test.

Significance Level

- Measure of how unlikely you want the results of the sample to be before you reject the null hypothesis

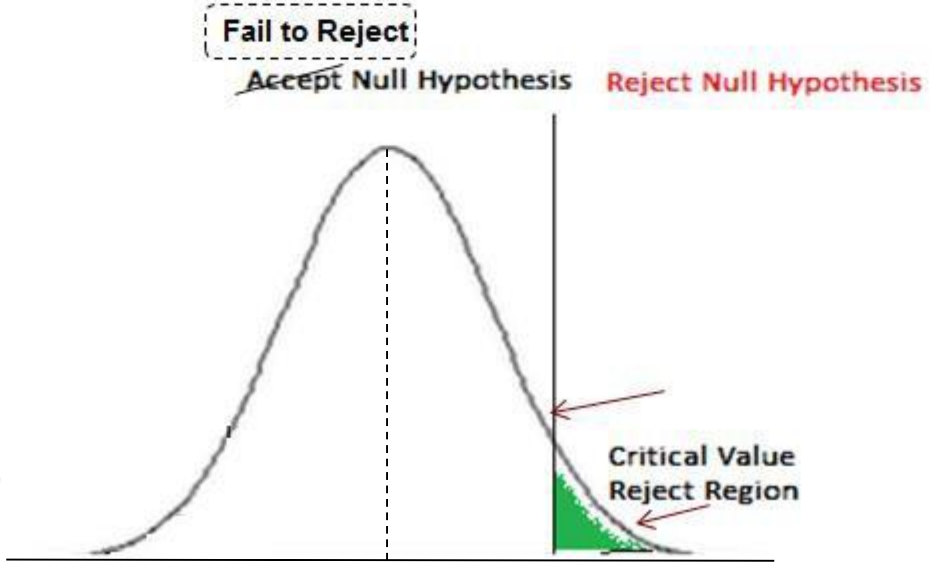


Building Blocks (contd.): ….Critical value and region, p-value

Critical Region

Set of values that present the most extreme evidence against the null hypothesis

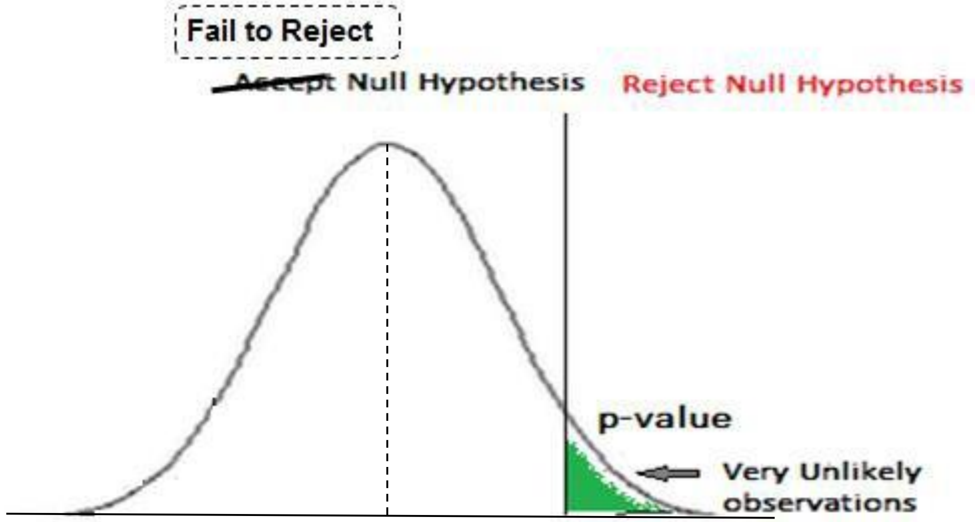
The fewer people in the sample who are cured, the stronger the evidence there is against the claims



p-value

p-value, is the probability given the null hypothesis is true, that the observed value is as extreme or more extreme than the actual

Smaller p-value, stronger evidence to reject null hypothesis



Chi square test: Test of frequency

The Concept

When to be used?

Are the variables categorical? ✓

Is the sample size adequate and random? ✓

Are the degrees of freedom known? ✓

The chi-square test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories

- Chi square statistics uses the degree of freedom which is the number of values that are free to vary after restriction has been placed on the data

- For instance, if the sum of four numbers is required to be 50, then three variables can assume any value but the fourth variable should have a value so that the sum comes to 50 hence d.o.f is 3

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

$O = \textit{the frequencies observed}$

$E = \textit{the frequencies expected}$

$\sum = \textit{the 'sum of'}$

Chi square test: Illustration

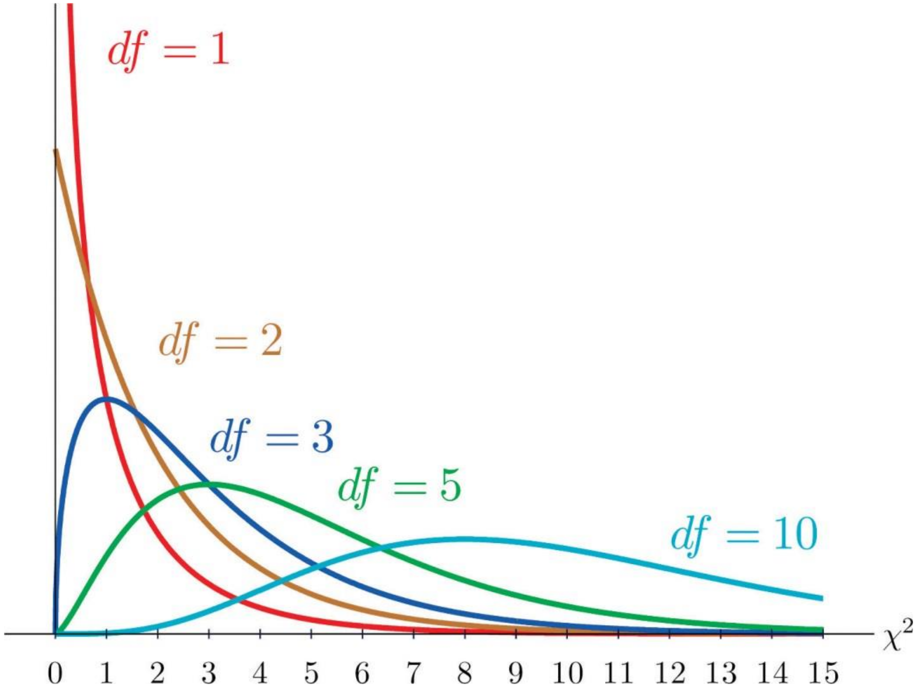
- A dice is rolled 36 times. The frequencies of the observed outputs are mentioned in the table
- The finding that the frequencies differ does not mean that the die is not fair
- To test whether the die is fair one has to conduct a significance test - the null hypothesis is that the die is fair
- If the probability of this observed frequency distribution is sufficiently low (compared to significance level), then the null hypothesis that the die is fair can be rejected

Number on Dice	Observed Frequency	Expected Frequency	$(O-E)^2/E$
1	8	6	0.667
2	5	6	0.167
3	9	6	1.500
4	2	6	2.667
5	7	6	0.167
6	5	6	0.167

$$\sum (O-E)^2/E = 5.33$$

Chi square test: Illustration

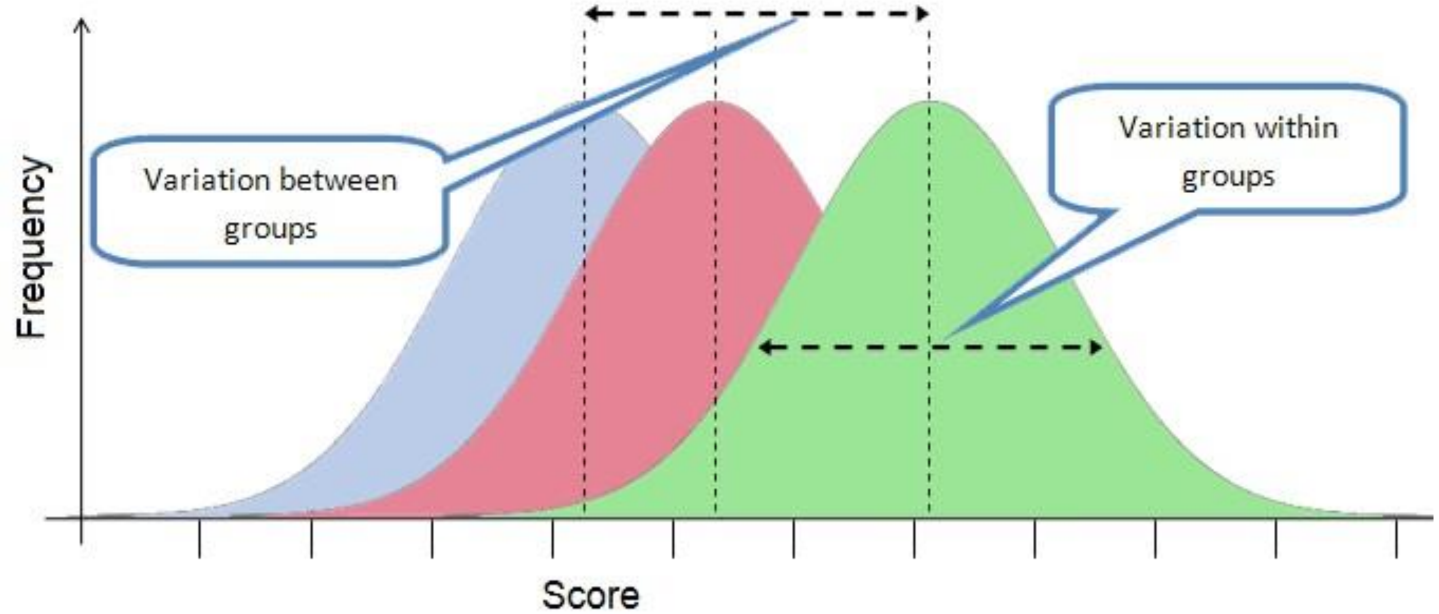
- Now that the degree of freedom is known (df = 6 – 1 = 5) and the value of chi-square has been calculated, the p-value has to be found out to get the significance level
- For this purpose, we use a chi-square table (or a calculator)
- In our case, when the $X2$ = 5.33 (df=5) the p-value associated with it is 0.377 (which is more than the level of significance) hence we fail to reject the null hypothesis
- As the df increases, the chi square distribution takes the shape of normal distribution



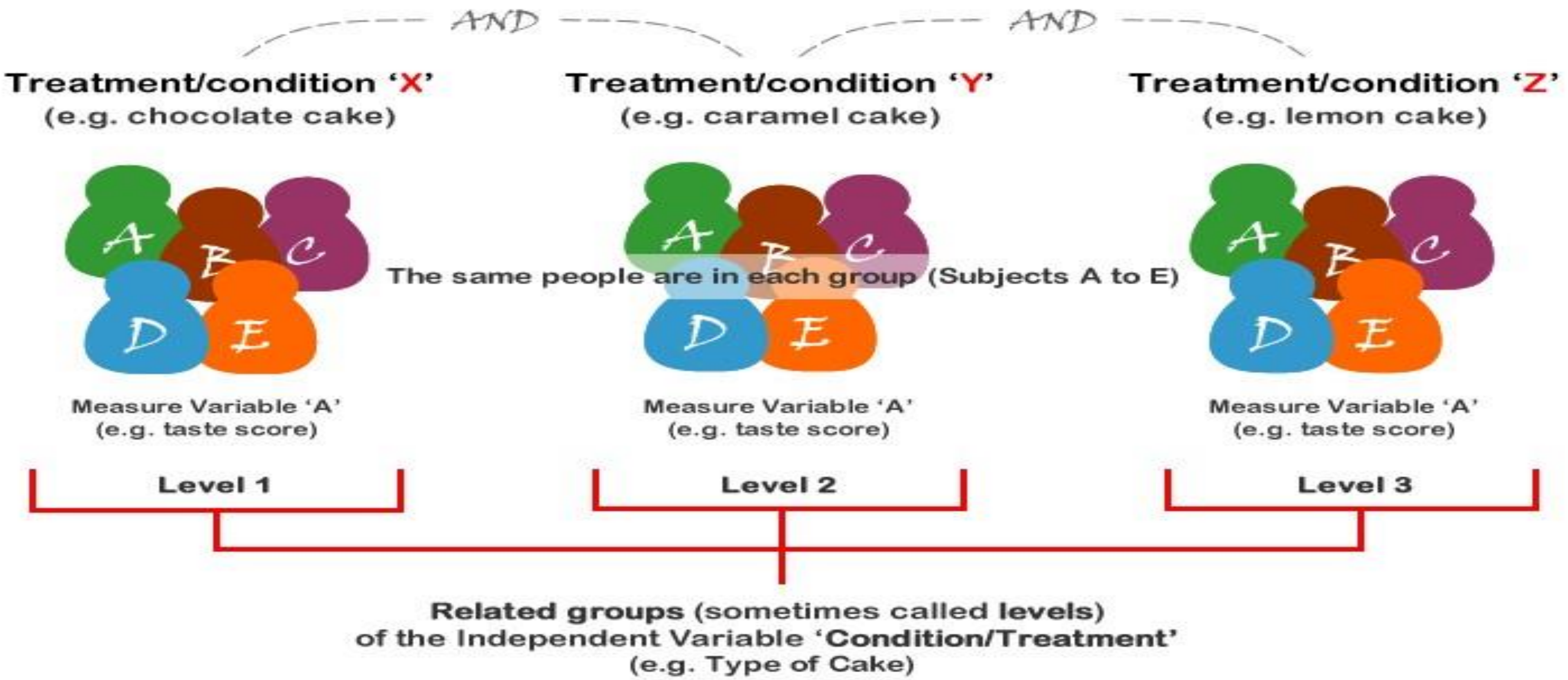
Analysis of Variance : ANOVA

- It is a way to test the hypothesis that there is not much difference between groups on some variable / treatment
- In the typical application of ANOVA, the null hypothesis is that all groups are simply random samples of the same population hence all treatments have the same effect (perhaps none). Rejecting the null hypothesis implies that different treatments result in altered effects

Consider the three graphs. They almost are the same and it is hard to tell the difference in their means just by looking



Visualizing the variability and variance within a group and amongst groups



To distinguish between these groups, the variability amongst the groups must be greater than that within the groups

Calculation of ANOVA depends on the number of groups being considered

Source	Degree of	Variance freedom	F -statistic
Between groups	K-1 (K is the number of groups)	SS_{bet} (Sum of squares between samples) ↓ (SS_{bet} / k-1) = MS_{bet} MS_{bet} (Variance between samples)	$F=MS_{bet}$ / MS_{w}
Within groups	N-K (N is the total number of observations across all the groups)	SS_{W} (Sum of squares within samples) ↓ (SS_{W} / N-k) = MS_{w} MS_{w} (Variance within samples)	

The F statistics generated gives the p-value based on the characteristics of F distribution

- This F-statistic can be compared with the tabulated values of F –distribution with (k-1) and (N-k) degrees of freedom
- The significance level tells whether to accept or reject the null hypothesis
- ***When we have only two groups to consider, a t-test can be used to compare the difference***
- ***However for more than two groups, conducting multiple t-tests (taking two at time) is not recommended as it can severely inflate Type 1 errors***
- Two important terms while doing ANOVA are:
  - Factor: A characteristic under consideration e.g. flavor, thought to influence the measured observation
  - Levels: A value of the factor e.g. chocolate, lemon etc.
- One way ANOVA is a test to study effects of more than one levels of a single factor so the test of hypotheses H_0: $\mu_i = \mu$ where i = 1,2,3,4,….k (μ_i is the mean of population for the different levels of i)