CLT and Hypothesis Testing

Objectives

- Central Limit Theorem
- Hypothesis Formulation
- Null and Alternative Hypothesis
- Type I and Type II Errors
- Confidence Intervals
- Hypothesis Testing
- One tailed v/s two tailed test
- Test of mean

Central Limit Theorem

“Sampling Distribution of the mean of any independent random variable will be normal”

- The random variable should have a well defined mean and variance (standard deviation).
- Applicable even when the original random variable is not normally distributed.

<u>Assumptions</u>:

- The data must be randomly sampled.
- The samples values must be independent of each other.
- **The 10% condition:** When the sample is drawn without replacement, the sample size *n*, should be no more than 10% of the population.
  - In general, a sample size of 30 is considered sufficient.
- The sample size must be sufficiently large.
  - If the population is skewed, pretty large sample size is needed.
  - For a symmetric population, even small samples are acceptable.

Central Limit Theorem *(contd.)*

Assume a dice is rolled in sets of 4 trials and the faces are recorded. This is repeated for a month (30 days)

Sample	Throw 1	Throw 2	Throw 3	Throw 4	Mean
1	4	1	6	2	3.25
2	1	2	3	2	2
3	5	6	4	6	5.25
4	4	3	6	1	3.5
5	2	2	4	3	2.75
6	4	2	1	6	3.25
7	3	6	6	4	4.75
8	2	4	2	5	3.25
9	2	1	5	6	3.5
10	1	3	6	6	4
11	4	3	3	3	3.25
12	6	5	4	1	4
13	3	3	3	1	3.25
14	2	5	2	6	3.75
15	1	3	1	6	2.75

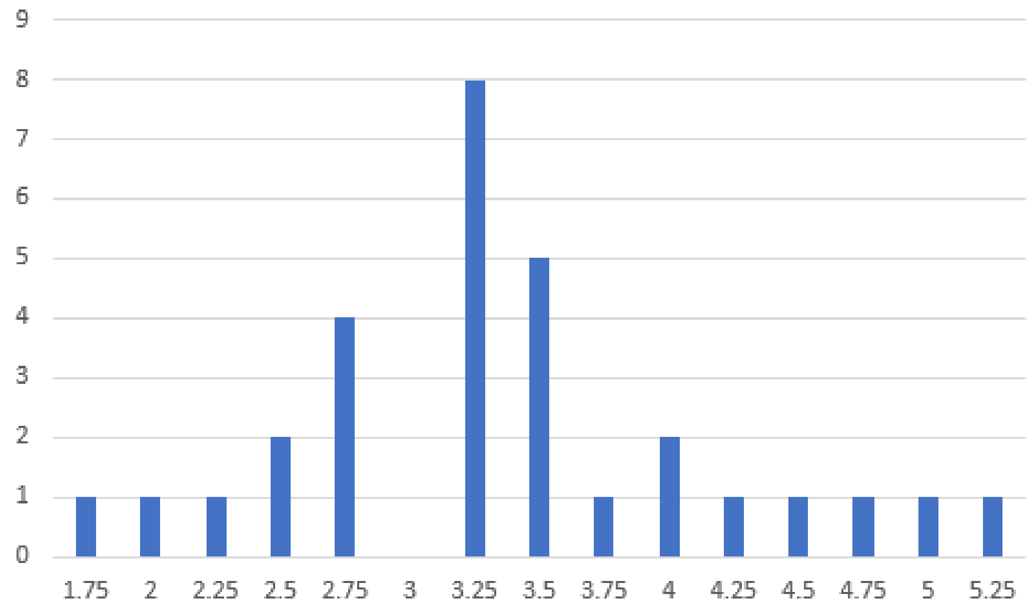
Sample	Throw 1	Throw 2	Throw 3	Throw 4	Mean
16	6	4	5	5	5
17	3	2	3	6	3.5
18	1	3	2	1	1.75
19	6	1	3	3	3.25
20	5	2	5	6	4.5
21	1	2	1	6	2.5
22	3	2	6	2	3.25
23	3	1	3	4	2.75
24	3	2	6	4	3.75
25	6	1	1	5	3.25
26	1	5	2	2	2.5
27	4	2	2	3	2.75
28	4	6	2	5	4.25
29	4	2	3	5	3.5
30	3	1	4	1	2.25

Central Limit Theorem *(contd.)*

The means of the 30 samples are obtained are recorded in a frequency distribution table:

Mean	1.75	2	2.25	2.5	2.75	3	3.25	3.5	3.75	4	4.25	4.5	4.75	5	5.25
Frequency	1	1	1	2	4	0	8	5	1	2	1	1	1	1	1

Plotting the sample distribution of the sample mean, the following curve is obtained:



Sampling Distribution: CLT

If *n* samples are drawn from a population that has a mean μ and standard deviation σ:

The sampling distribution follows a normal distribution with:

- Mean: μ
- Standard Deviation: $\sigma / \sqrt{n}$ (also known as Standard Error)

The corresponding z-score transformation is:

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

- If the population is normal, this holds true even for smaller sample sizes.
- However, if the population is not normal, this holds true for sufficiently large sample sizes.

Population parameters : Mu=Mean
Std Dev=Sigma

Sample Statistics : Mean=x-bar
Std Dev =s

Standard Error = sigma/sqrt(n)
Ex: 30 sample 1 : x-bar 1= 8 lakhs
30 samples 2 : x-bar2 = 8.5 lakhs

Refer: http://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/BS704_Probability12.html

Sample = mean =50; std dev=5

Sample 2 = mean = 50 ; std dev = 10

Why Hypothesis Tests?

In a practical scenario, we seldom know about the population (or population parameters so to speak). By doing certain Hypothesis Tests, we try to infer about the population parameters with the help of sample statistic. There are Hypothesis Tests designed estimating or inferring about various population parameters.

Hypothesis Formulation

Coca Cola’s most selling product is the 600 ml coke or Coca Cola. Since the 600 ml info is on the label, we assume it to be true. But, is it actually true ?

As a customer, we’re concerned that there is *at least* 600 ml in the bottle. If little more, we’re okay.

As a manufacturer, we would want the volume to be *exactly* 600 ml.

- Under-filling upsets the customers, while overfilling results in higher costs of production.

<u>As a customer</u>:

- On an average, is there at least 600 ml coke in every bottle?

$\Rightarrow$ Quantity >= 600 ml

<u>As a manufacturer</u>:

- On an average, is there exactly 600 ml coke in every bottle?

=> Quantity = 600 ml

Hypothesis Formulation *(contd.)*

To experiment the same:

We collect 100 bottles from all over the country, so that we have a *random* sample.

We then measure the volume of each bottle in the sample to find the mean of 100 bottles.

Using this sample mean, we test the assumption (of what is on the label) i.e. status quo

This is the key to inferential statistics: making inferences about the population from the sample.

Company claims to have discovered new ingredients, using which they can sell volumes > 600 ml for the same cost.

$\Rightarrow$ Company claim: Volume > 600 ml   (This may or may not be true)

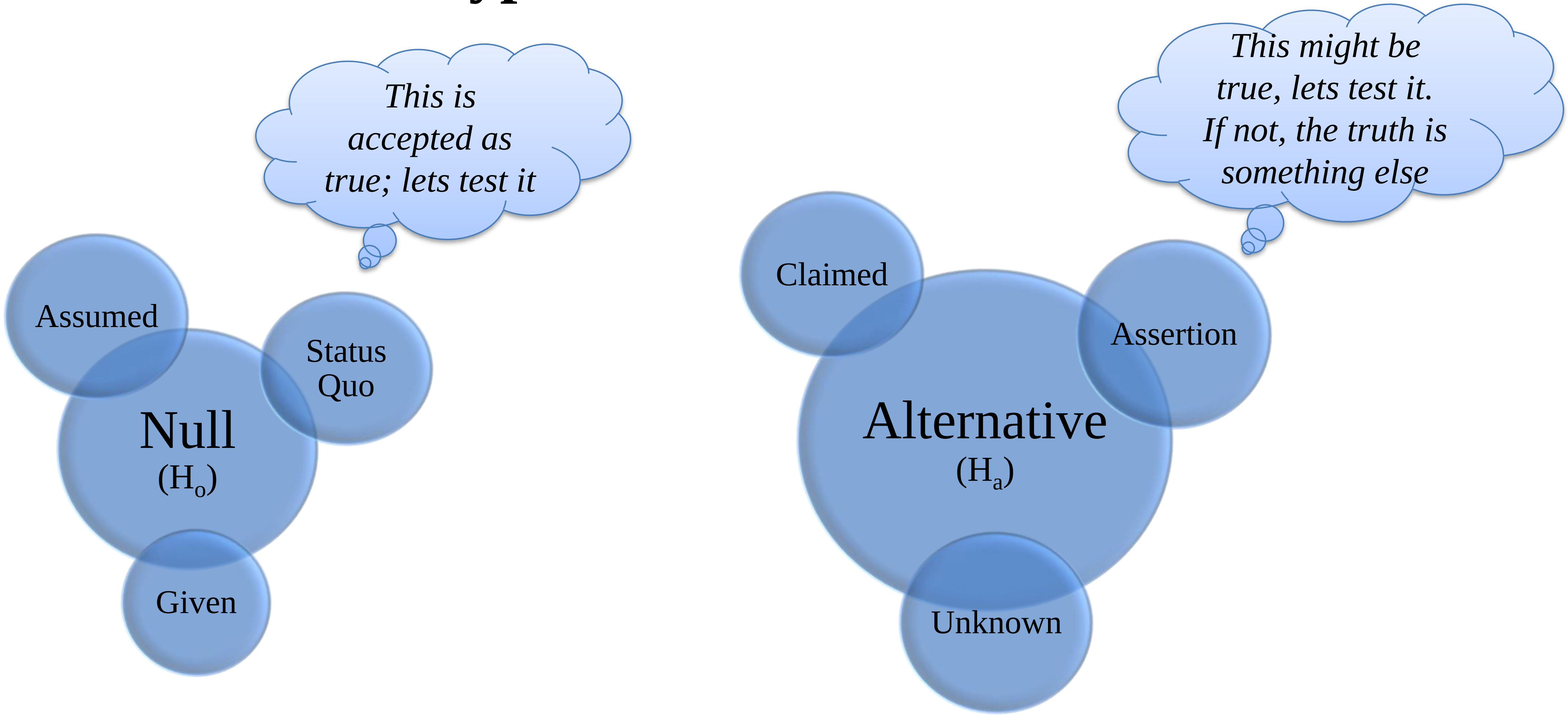
The company is making a claim that would be tested. It is not testing an already existing assumption (status quo).

When formulating hypothesis:

Am I testing an assumption or status quo, that already exists?

Am I testing a claim or an assertion, which is something beyond what I already know?

Null and Alternative Hypothesis



The null and alternative hypotheses are opposite; mutually exclusive.
We may begin by formulating either the null or the alternative hypothesis; the other is just a complement.

Null and Alternative Hypothesis

- Null: Assumption or the status quo holds, there’s no new observation.
- Alternative: Rejection of assumption or status quo.

- Null: Negation of the research question.
- Alternative: Research question to be proven.

- Null: Always contains equality (<=, >=, =)
- Alternative: Does not contain equality (<, >, !=)

All statistical conclusions are made in reference to the null hypothesis.

We either reject the null hypothesis or fail to reject the null hypothesis; we do not accept the null hypothesis.

From the start, we assume the null hypothesis to be true, later the assumption is rejected or we fail to reject it.

- When we reject the null hypothesis, we can conclude that the alternative hypothesis is supported.
- If we fail to reject the null hypothesis, it does not mean that we have proven the null hypothesis is true.
  - Failure to reject the null hypothesis does not equate to proving that it is true.
  - It just holds up our assumption or the status quo.

Null and Alternative Hypothesis: Example

Assumption: Volume = 600 ml

⇒ H_o: μ 600 ml H_a: μ != 600 ml

Case 1: Data indicates that bottles are filled properly.

For example, μ = 600.4 ml

⇒ We fail to reject the null hypothesis i.e. fail to reject the assumption.

We cannot say that the null is proved, we can just say that the assumption has held up.

Case 2: Data indicates that bottles are not filled properly

For example, μ = 645 ml

⇒ We reject the null hypothesis i.e. reject our assumption.

We have statistical evidence to say that alterative hypothesis is valid or supported.

Null and Alternate Hypothesis Formulation

A medical trial is conducted to test whether or not a new medicine reduces cholesterol by 25%. State the null and alternative hypotheses.

- H_0 : The drug reduces cholesterol by 25%. $p = 0.25$
- H_a : The drug does not reduce cholesterol by 25%. $p \neq 0.25$

We want to test whether the mean GPA of students in American colleges is different from 2.0 (out of 4.0). The null and alternative hypotheses are:

- H0: $\mu = 2.0$
- Ha: $\mu \neq 2.0$

We want to test if college students take less than five years to graduate from college, on the average. The null and alternative hypotheses are:

- H0: $\mu \geq 5$
- Ha: $\mu < 5$

We want to test if it takes fewer than 45 minutes to teach a lesson plan. State the null and alternative hypotheses. Fill in the correct symbol (=, ≠, ≥, <, ≤, >) for the null and alternative hypotheses.
H0: μ __ 45 Ha:μ __ 45

- H0: $\mu \geq 45$
- Ha: $\mu < 45$

In an issue of U.S. News and World Report, an article on school standards stated that about half of all students in France, Germany, and Israel take advanced placement exams and a third pass. The same article stated that 6.6% of U.S. students take advanced placement exams and 4.4% pass. Test if the percentage of U.S. students who take advanced placement exams is more than 6.6%. State the null and alternative hypotheses.

- H0: $p \leq 0.066$
- Ha: $p > 0.066$

Type I and Type II Errors

<u>Type I Error</u>:

- Rejection of null hypothesis when it should not have been rejected.
- Incorrectly rejecting the null hypothesis.

<u>Type II Error</u>:

- Failure to reject the null hypothesis, when it should have been rejected.
- Incorrectly not rejecting the null hypothesis.

Decision/ Reality	H_o True (Should not reject)	H_o False (Should reject)
Reject H_o	Type I Error (α)	Correct Rejection (No error)
Fail to Reject H_o	Correct Decision (No error)	Type II Error (β)

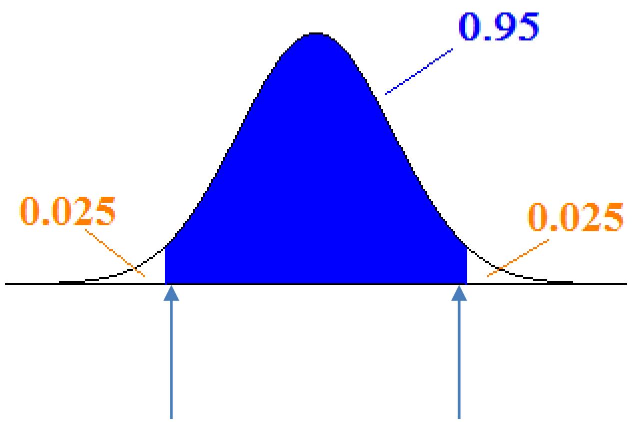
<u>Causes of Type I and Type II Errors</u>:

- By random chance, we may select a sample which is not representative of the population.
- Sampling techniques may be flawed.
- Assumptions in our null hypothesis may be flawed.

Confidence Intervals

- 95% of all sample means ($\bar{x}$) are hypothesized to be in this region.

⇒ This is called as 95% confidence interval.



Critical Value = +1.95
Actual = 2.5 hence will fall in the rejection area

- If sample mean is in the blue region, we fail to reject the null hypothesis
- If sample mean is in the white region, we reject the null hypothesis.

- Here, $\alpha = 0.05$

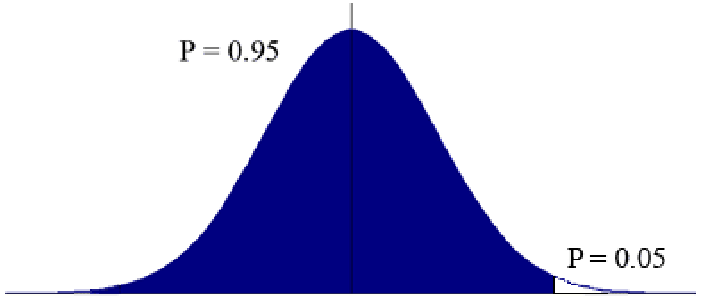
⇒ α is the <u>level of significance</u> or our tolerance level towards making a Type I error.

- If the null hypothesis is correct, (α * 100)% of the sample means should lie in the rejection region.

In case of one-tailed situation:

- All of α is in one tail or the other, depending on the alternative hypothesis.
- H_a points to the tail, where the critical value and the rejection region are.

(Case when observed mean > hypothesized mean)



Hypothesis Testing

greatlearning

<u>Steps</u>:

1. Develop a clear research problem.
2. Establish both null and alternate hypothesis.
3. Determine appropriate statistical test.
4. Choose Type I error rate (α). This could be assumed to be 0.05, unless specified.
5. State the decision rule: when to reject / fail to reject null hypothesis.
6. Gather sample data.
7. Calculate test statistic.
8. State statistical conclusion, based on the p-value
9. Make inference based on conclusion.

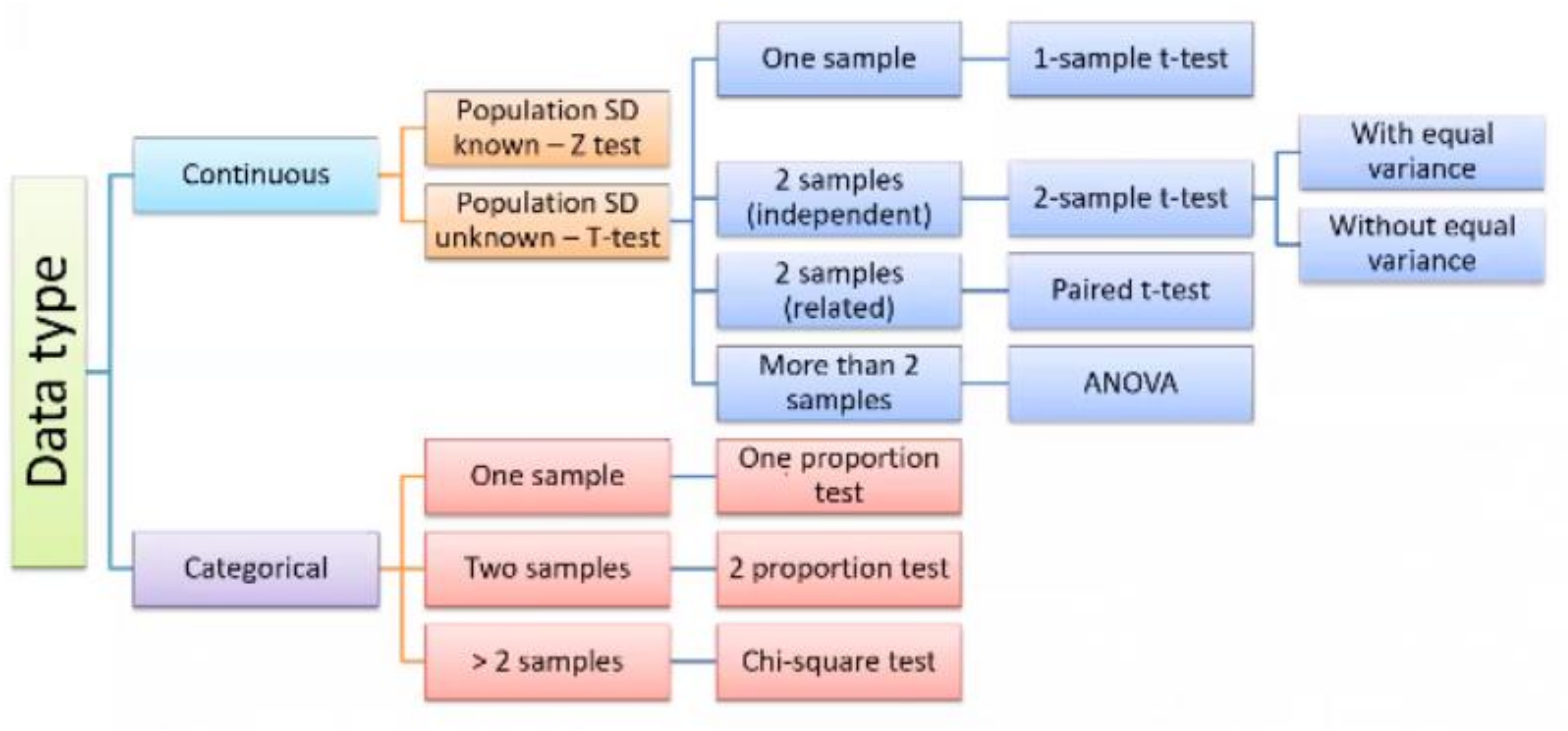
Given an alpha level , if:
p is low : Null will GO - <u>Reject</u>
p is high : Null will FLY- <u>Fail to Reject/(Accept)</u>

<u>Types of Hypothesis tests</u>:

- Single sample or two samples
- One tailed or two tailed
- Tests of mean, proportion or variance

Hypothesis Testing Roadmap

greatlearning



Example – Confidence Interval

- A paper manufacturer has a production process that operates continuously. The paper is expected to have a mean length of 11 inches and a standard deviation of 0.02 inches. At periodic intervals, a sample is selected to determine whether the paper length is still equal to 11 inches. You select a random sample of 100 sheets and the mean paper length is 10.998 inches.
- Construct a 95% confidence interval.
- Construct a 99% confidence interval.

$$C.I. = \bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Mu=11
Xbar=10.998
Sigma=0..02
N=100
Alpha1=0.05
Alpha2=0.01

Example – Single Sample –z test of mean

greatlearning

- You are the manager of a fast food restaurant. You want to determine if the population mean waiting time has changed from the 4.5 minutes. You can assume that the population standard deviation is 1.2 minutes. You select a sample of 25 orders in an hour. Sample mean is 5.1 minutes. Use the relevant hypothesis test to determine if the population mean has changed from the past value of 4.5.

H0: Mu=4.5
H1: Mu <> 4.5

n=25
Sigma = 1.2
Xbar=5.1
Mu=4.5
Alpha = 0.05

Example – 2 Sample – t test of mean

- A hotel manager is concerned with increasing the return rate of customers. One aspect that affects this is the time taken to deliver to luggage to the guest's room after check-in. A random sample of 20 deliveries were selected from WIngA and WingB of the hotel. Analyse whether is a difference in the average time taken by the 2 Wings?

H0 :MuA=MuB
H1 : MuA<>MuB

Example – Paired – t test of mean

The data in the file Concrete.csv represents the compressive strength of 20 samples taken 2 days and 7 days after pouring.
-At 0.05 level of significance, is there evidence that the mean strength is lower at 2 days than 7 days

Ho: Mu2 = Mu7
H1: Mu2< Mu7

Example – Chi-Squared Test

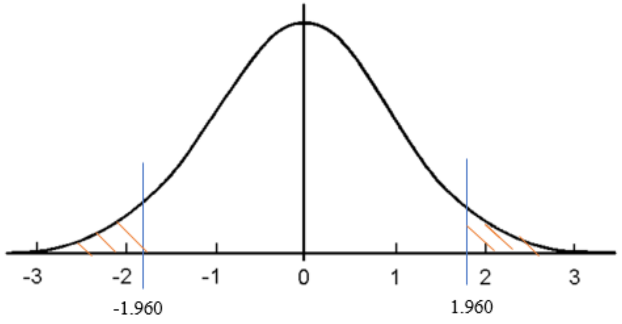
- When new paperback novels are promoted at bookstores, a display is often arranged with copies of the same book with differently colored covers. A publishing house wanted to find out whether there is a dependence between the place where the book is sold and the color of its cover. For one of its latest novels, the publisher sent displays and a supply of copies of the novels to large bookstores in five major cities. The resulting sales of the novel for each city-color combination are as follows. Numbers are in thousands of copies sold over a three-month period.

H0: Color and Location are independent
H1: Color and Location are dependent

One tailed vs two tailed test

Case 1: Coca Cola official claims that the mean volume in coke bottles is 600ml.

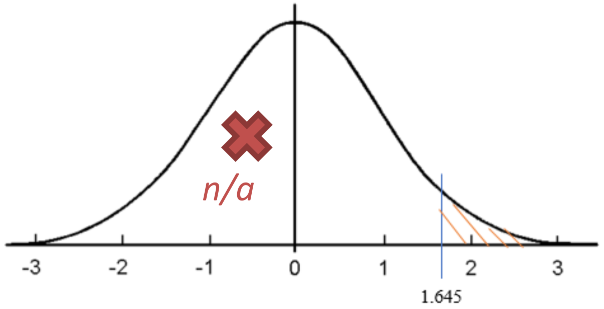
$\mu = 600$ml



(Two tailed test)

Case 2: Coca Cola official claims that the mean volume in coke bottles is more than 600ml.

$\mu > 600$ml



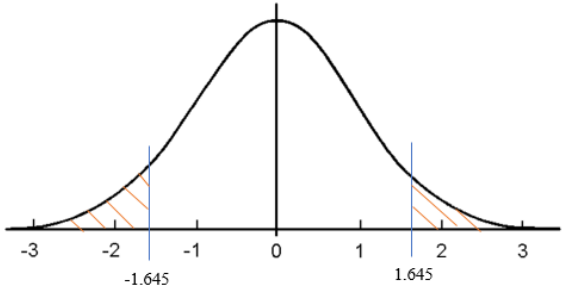
(One tailed test)

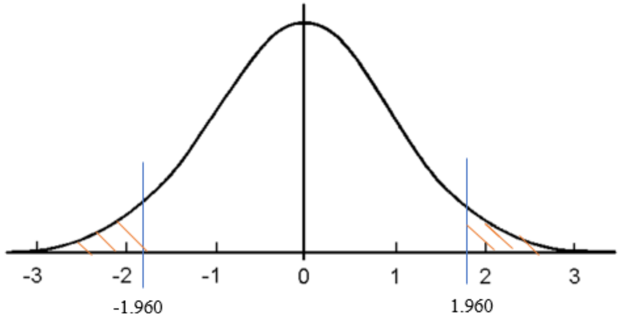
Single Sample z – test of mean *(known σ)*

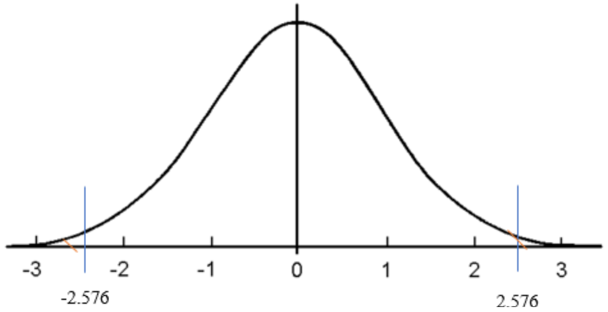
Test Statistic: $z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

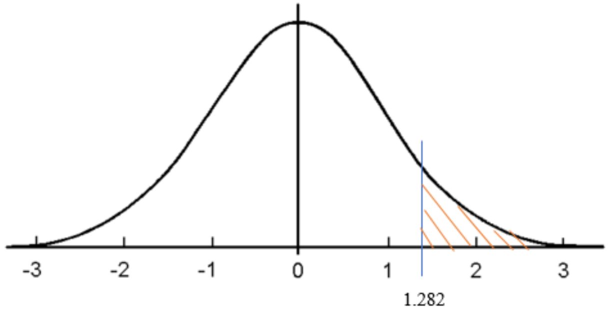
p-value(level of significance): How much of the area is above the test-statistic? *(Does test statistic fall in the rejection region?)*

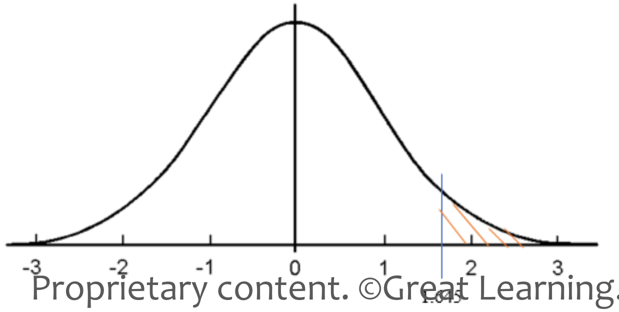
If it is less than the specific α, we reject the null hypothesis

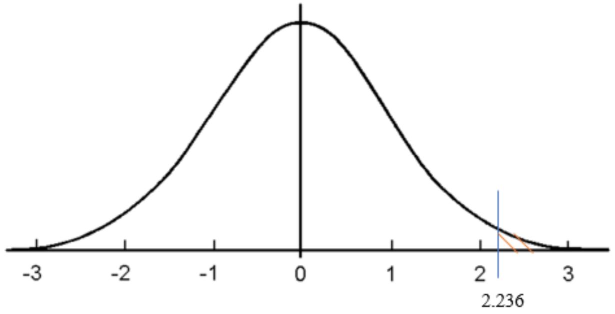












Single Sample t – test of mean *(unknown σ)*

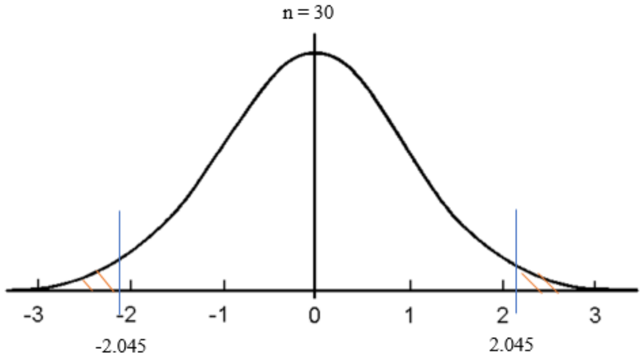
Test Statistic: $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$

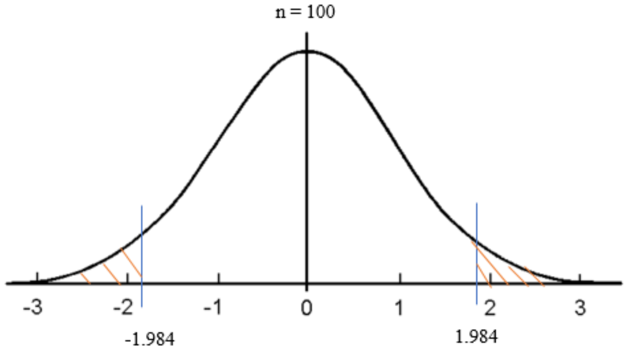
p-value: How much of the area is above the test-statistic? *(Does test statistic fall in the rejection region?)*

If it is less than the specific α, we reject the null hypothesis

* *t-statistic depends on the sample size*

$\alpha = 0.05$





Two sample tests of mean

To understand if the mean volume in coke bottle is 600ml, we decide to take two samples from two manufacturing centers.

The assumption would be that the mean difference between the two samples would be zero:

i.e. $\mu_1 = \mu_2$ => $\mu_1 - \mu_2 = 0$ *(Null hypothesis)*

- When σ is known, use z-distribution

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\left(\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}}\right)}$$

D_o : hypothesized mean b/w two means *(zero in the above example)*

- When σ is not known, use t-distribution

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where, *df* is calculated as:

$$df = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\left[\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}\right]}$$

Matched Sample / Paired t-test of mean

It is reported that the caffeine in coke had increased the permissible limit because of manufacturing issues.
A sample of 100 bottles taken reports the average caffeine to be 10.5 µg (Permissible level is 10 µg)
Coca Cola technicians derive a technique using which they would correct caffeine levels in the coke bottles, rather than having to throw them away.
The 100 bottles are made to undergo this technique and caffeine levels are measured in the same bottles.

The t-statistic here is calculated as:

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

$\bar{d}$ = Mean difference, μ_d = hypothesized difference (*usually 0*)
s_d = Standard deviation of the difference

Bottle	Caffeine before	Caffeine after	Difference
1	10.4	10.2	0.2
2	10.8	10.5	0.3
3	9.8	10	-0.2
…			

Calculate $\bar{d}$ and s_d

Test of variance

5 bottles from Manufacturer 1 show the following quantities:

- 607ml, 602ml, 590ml, 603ml, 598ml

5 bottles from Manufacturer 2 show the following quantities:

- 602ml, 597ml, 600ml, 603ml, 598ml

Case 1:

One of the two manufactures contract should be renewed at the end of the year.

Which one do you think should be renewed ? First, second or both ?

Case 2:

The audit teams wants to ensure that the production patterns should same remain equivalent across all manufactures.

Do you think the two manufactures qualify this constraint ?

F-ratio test of variance

When two independent random samples are taken from normal population(s) with equal variances, the sampling distribution of the ratio of those sample variances follows an F distribution.

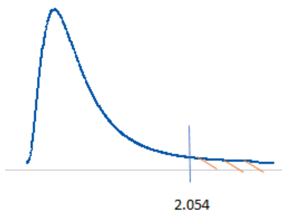
- Test of equality of variances: comparison of two sample variances
- The variances are compared using a ratio:

$F = \dfrac{s^2_x}{s^2_y}$ where, s^2_x is the larger sample variance while s^2_y is the smaller sample variance (Both numerator and denominator have their individual dfs)

F- distribution is only right-tailed:

H_o: $\sigma^2_x = \sigma^2_y$ H_a: $\sigma^2_x \neq \sigma^2_y$

For $\alpha = 0.05$ and df1 = 24, df2 = 21



p-value: How much of the area is above the test-statistic? *(Does test statistic fall in the rejection region?)*

If it is less than the specific α, we reject the null hypothesis

Note: Applications of the F-statistic or the F-test can be seen in more detail in ANOVA.

HappyLearning