Statistical Learning - Hypothesis Testing

Agenda

- Sampling distribution
- Central Limit Theorem
- Confidence intervals
- Hypothesis Formulation
- Null and Alternative Hypothesis
- Type I and Type II Errors
- Hypothesis Testing
  - One tailed v/s two tailed test
  - Test of mean
  - Test of proportion
  - Test of variance
- Examples

Concepts of sampling distribution

- Why do we need sampling?
- Analyse the sample and make inferences about the population
- Sample statistic vs population parameter
- Sampling distribution – distribution of a particular sample statistic of all possible samples that can be drawn from a population – sampling distribution of the mean

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: σ / √n (also c/a Standard Error)
- The corresponding z-score transformation is:

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

- 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.

Central Limit Theorem

- "Sampling Distribution of the mean of any independent random variable will be normal"
- This applies to both discrete and continuous distributions.
- The random variable should have a well defined mean and variance (standard deviation).
- Applicable even when the original variable is not normally distributed.

- Assumptions:
- 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

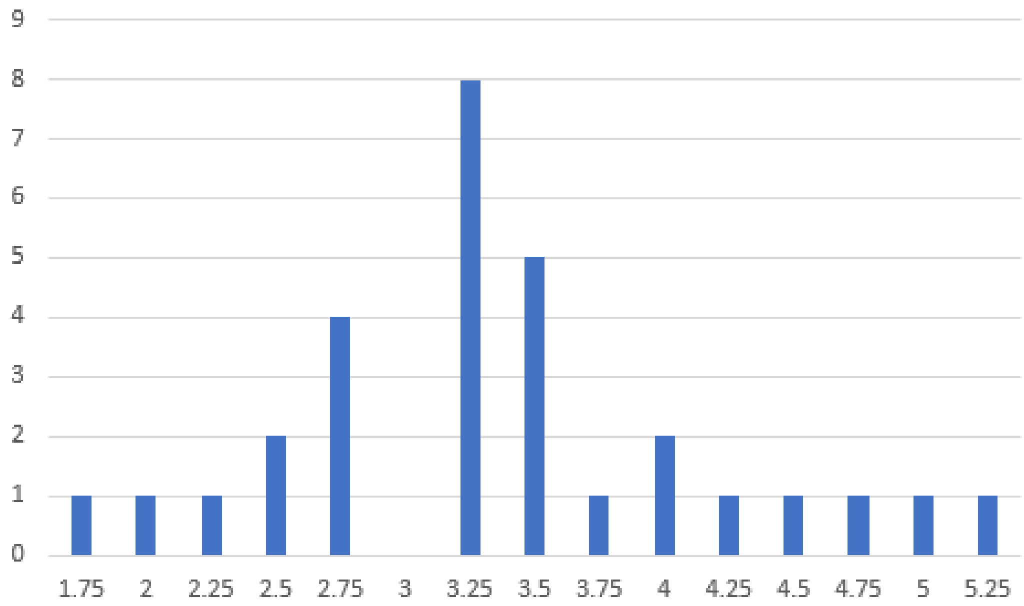
Sample	Throw 1	Throw 2	Throw 3	Throw 4	Mean
16	6	4	5	5	5
17	3	2	3	6	3.5
18	1	3	2	1	1.75
19	6	1	3	3	3.25
20	5	2	5	6	4.5
21	1	2	1	6	2.5
22	3	2	6	2	3.25
23	3	1	3	4	2.75
24	3	2	6	4	3.75
25	6	1	1	5	3.25
26	1	5	2	2	2.5
27	4	2	2	3	2.75
28	4	6	2	5	4.25
29	4	2	3	5	3.5
30	3	1	4	1	2.25

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:



Hypothesis

Hypothesis is an assumption

- Null hypothesis – is the status quo – rejection leads to desired conclusion
- Alternate hypothesis – what a person believes is true or wants to prove

Hypothesis Formulation

Coca Cola’s most selling product is the 600ml 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.

Under-filling upsets the customers

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

Quantity >= 600 ml

Quantity < 600 ml

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

Overfilling results in higher costs of production.

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

Quantity = 600 ml

Quantity > 600 ml

Hypothesis Formulation (contd.)

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

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

Measure volume of each bottle in the sample to find the mean of 100 bottles.

Use sample mean to test assumption (status quo)

What is status quo in this scenario

Mean volume = 600ml

Hypothesis Formulation (contd.)

- Company claim: Volume > 600 ml (This may or may not be true)

What is the claim or assumption? → Mean volume > 600ml

When formulating hypothesis…

Am I testing a status quo, that already exists?

Am I testing an assumption or claim, that is something beyond what I know?

greatlearning
Learning for Life

When formulating hypothesis…

Am I testing a status quo, that already exists?

- Null hypothesis
- Negation of the research question
- Always Contains equality - =, <=, >=

Am I testing an assumption or claim, that is something beyond what I know?

- Alternate hypothesis
- Research question to be proven
- Does not contain equality - ≠, <, >

Null and Alternative Hypothesis

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.

Type 1 & Type 2 errors

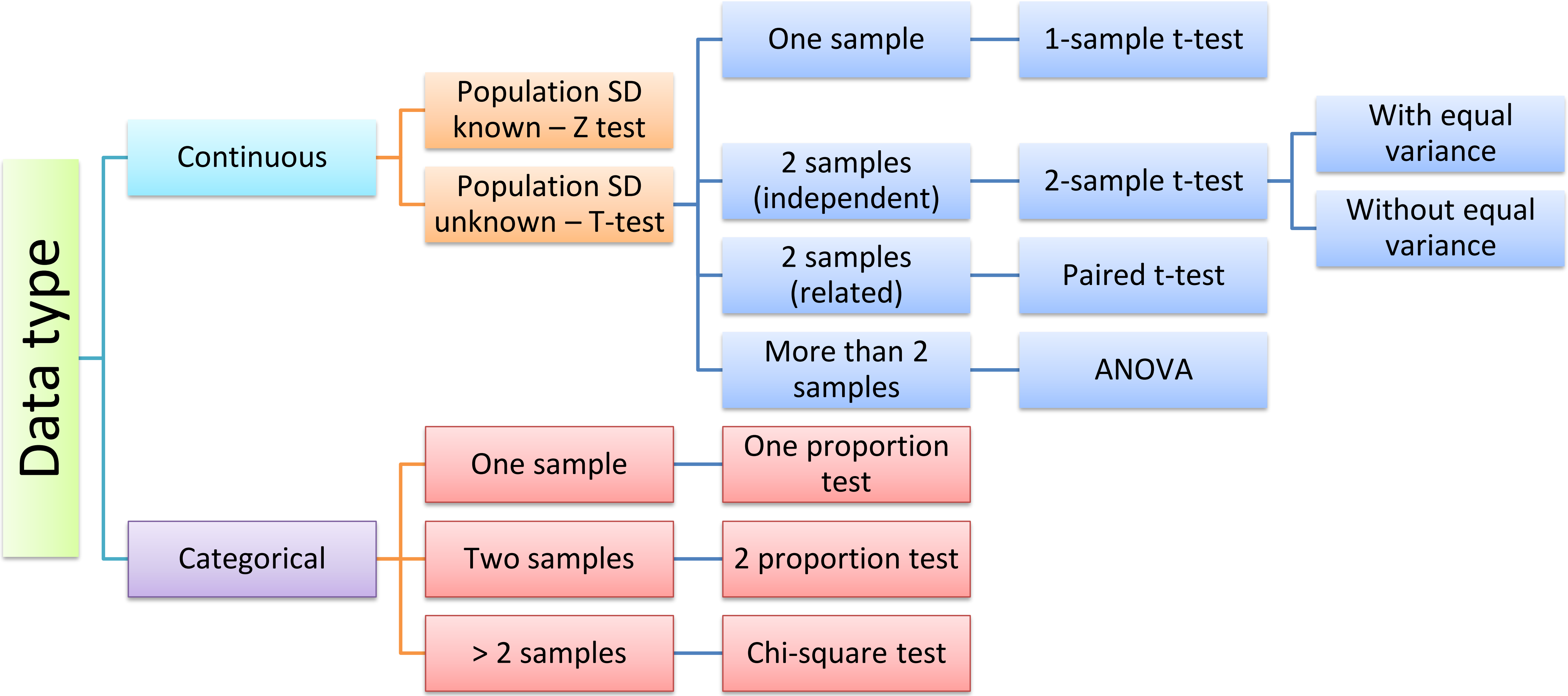
Confidence level = 1 - α	**H_o True (Should not reject)**	**H_o False (Should reject)**
Reject H_o	Type I Error (α)	Correct decision
Fail to Reject H_o	Correct decision	Type II Error (β)

Power of the test is its ability to "reject" Ho when it is false

Power of the test = 1- β

Hypothesis testing roadmap

greatlearning
Learning for Life



Steps in hypothesis testing

Develop a clear research problem.

Establish both null and alternate hypothesis.

Determine appropriate statistical test.

Choose Type I error rate (α). This could be assumed to be 0.05, unless specified.

State the decision rule: when to reject / fail to reject null hypothesis.

Gather sample data.

Calculate test statistic.

IF p is low, NULL will GO

State statistical conclusion, based on the p-value

Make inference based on conclusion.

Type of hypothesis tests

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

greatlearning
Learning for Life

One tailed vs two tailed test

Case 1

- A customer complains that the mean volume is not equal to 600 ml
- What is Ho?
- What is Ha?
- Is this one-tailed or two tailed?

Case 2

- Coca Cola official claims that the mean volume in coke bottles is more than 600ml
- What is Ho?
- What is Ha?
- Is this one-tailed or two tailed?

One tailed vs two tailed test

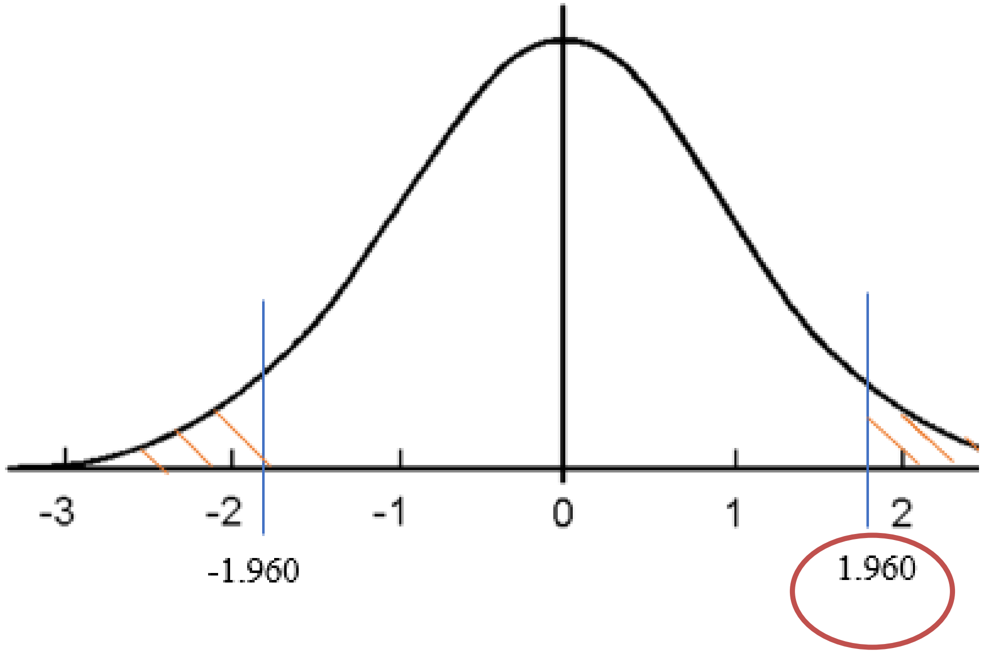
Case 1

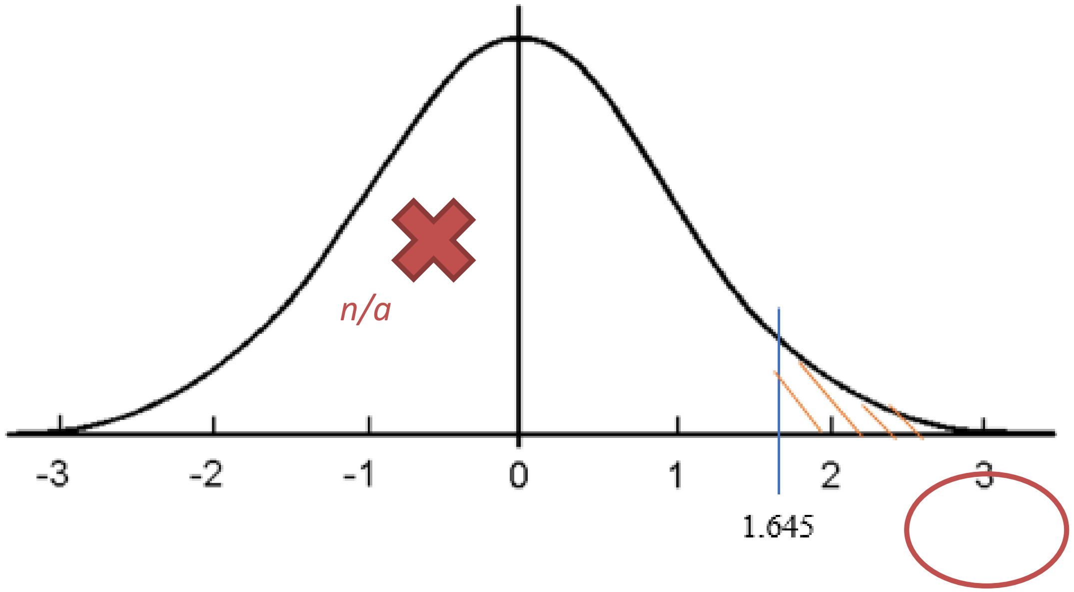
- Ho: $\mu = 600$ml
- Ha: $\mu \neq 600$ml
- Two-tailed test

Case 2

- Ho: $\mu <= 600$ml
- Ha: $\mu > 600$ml
- One-tailed test

One tailed vs two tailed test

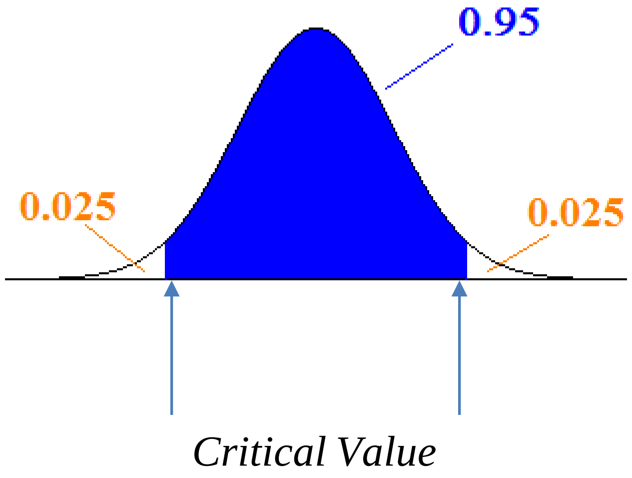




Confidence Intervals

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

⇒ This is called as 95% confidence interval.



- 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, $\alpha = 0.05$

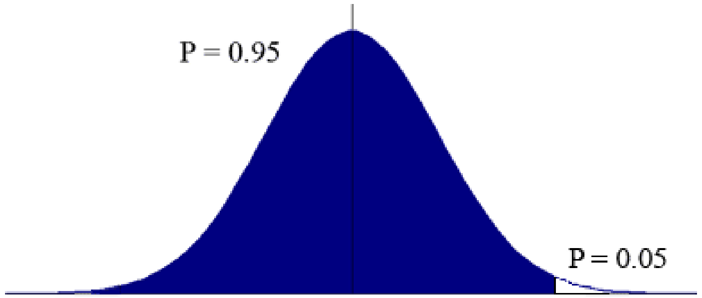
⇒ α is the <u>level of significance</u> or our tolerance level towards making a Type I error.

- 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)



Example – Confidence interval estimation

- 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.

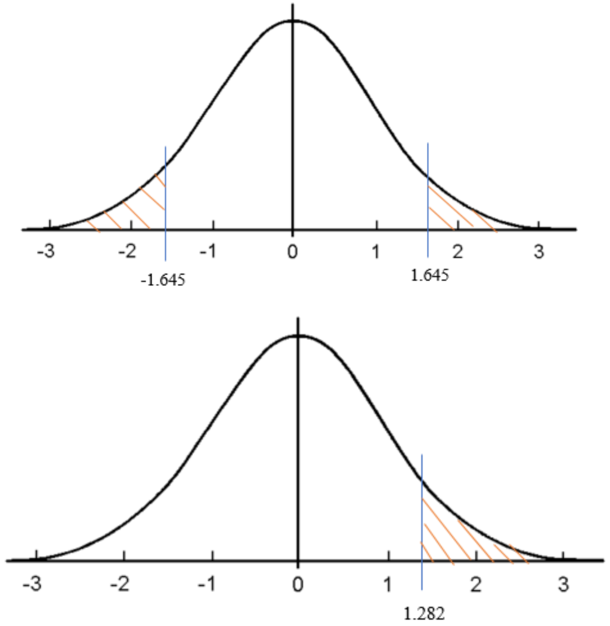
$$C.I. = \overline{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

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

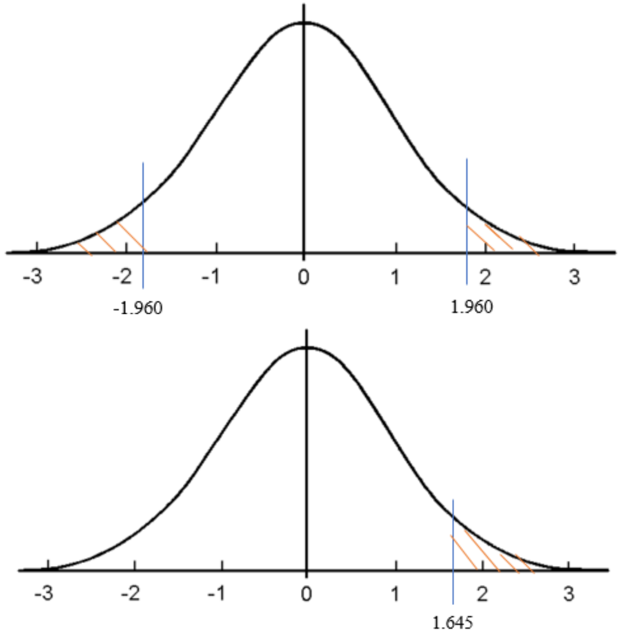
Test Statistic: $z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

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

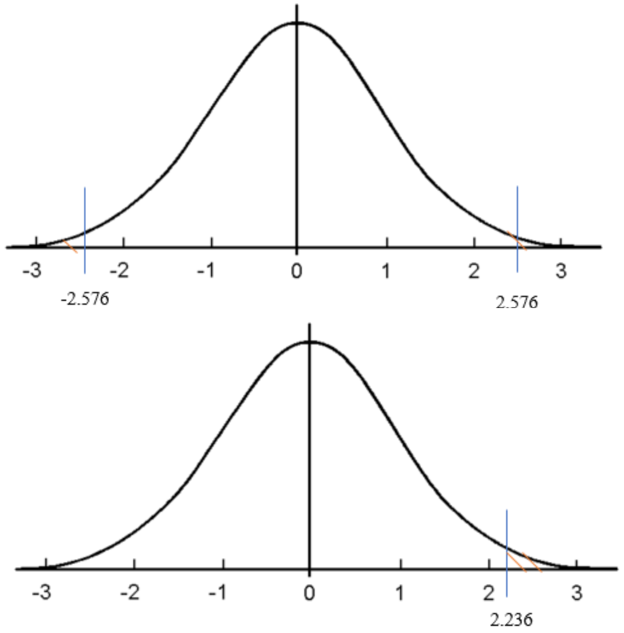
α = 0.10



α = 0.05



α = 0.01



Example problem - Single Sample z – test of mean

- 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.

Steps to solve the problem…

- One-tailed or two-tailed
- What is Ho and Ha
- Determine Z and Zstat
- Draw the normal curve
- Reject/Fail to reject Ho?

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

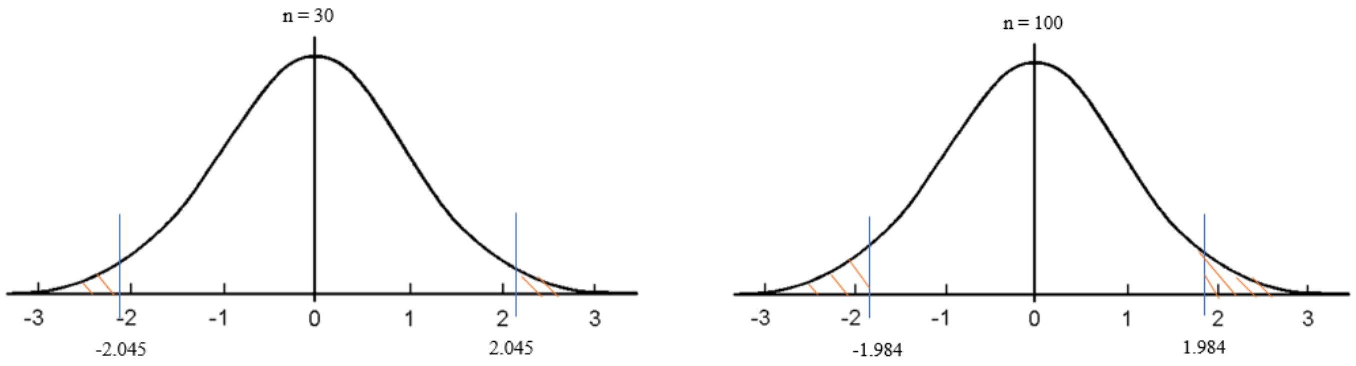
Test Statistic: $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

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

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 Wing A and Wing B of the hotel. Analyse whether is a difference in the average time taken by the 2 Wings?

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

greatlearning
Learning for Life

Example -Paired t-test of mean

- The data represents the compressive strength of 40 samples taken 2 days and 7 days after pouring.
  - AT 0.01 level of significance, is there evidence that the mean strength is lower at 2 days than at 7 days?

z-test of Proportion

In 2010, Coca Cola officials noted that 30% of the bottles were under-filled. They took corrective measures.
5 years later, they sampled 300 bottles and determined that 76 of them were under-filled.
At 5% significance level, is this evidence sufficient to show the impact of corrective measures?

H_o: $p_o = 0.30$
H_a: $p_o < 0.30$

$\hat{p} = 76/300$

For one sample,

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

For two samples,

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

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

Example – One tailed test of proportions

There are 155 banks involved in certain international transactions. A federal agency claims that at least 35% of these banks have total assets of over $10 billion (In U.S. dollars). An independent agency wants to test this claim. It gets a random sample of 50 out of the 155 banks and finds that 15 of them have total assets of over $10 billion. Can the claim be rejected?

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

<u>Case 1</u>:

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 ?

<u>Case 2</u>:

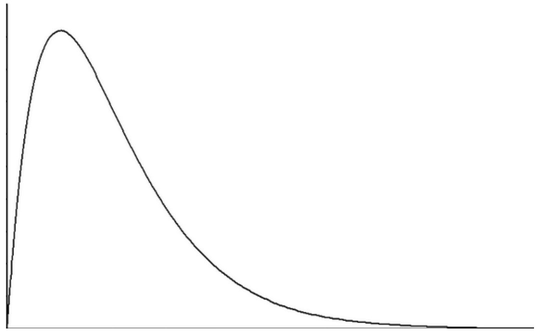
The audit teams wants to ensure that the production patterns should same remain equivalent across all manufactures.

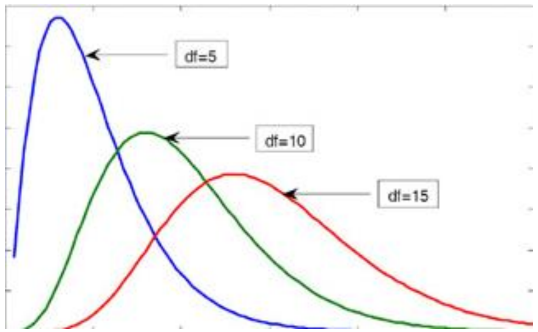
Do you think the two manufactures qualify this constraint ?

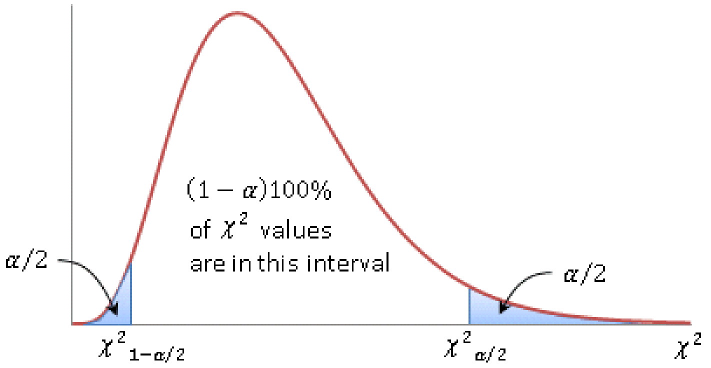
Chi square test of variance

When we take many samples of the same size from a normal population and find the sample means, they follow a normal distribution.

When we take many samples of the same size from a normal population and find the sample variances, they DO NOT follow a normal distribution; instead they follow a **chi-square (χ^2) distribution** , which is dependent on the degrees of freedom.







- Area under the curve is always 1.
- Cumulative Probability runs from right to left; 1 is towards the left end, while 0 is towards the right.

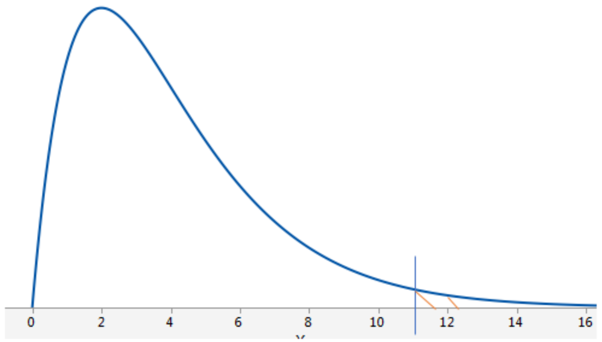
Chi square test of variance

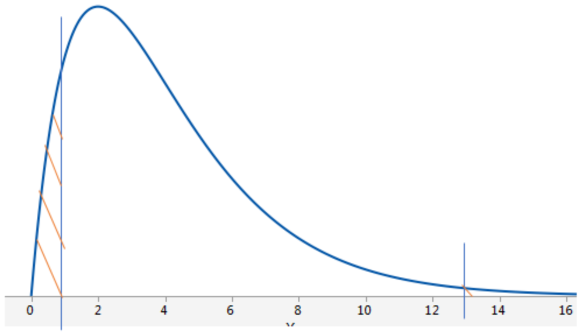
Chi-square (χ^2) test compares the population variance, with the hypothesized variance.

$\chi^2 = \frac{(n-1)\, s^2}{\sigma^2}$ where, n = sample size

s^2 = sample variance and σ^2 = population variance (which we wish to test)

At $\alpha = 0.05$ and n = 5 (df = 4)





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

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 given. Numbers are in thousands of copies sold over a three-month period.

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:

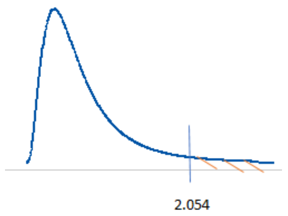
$F = \dfrac{s^2_x}{s^2_y}$ where, s^2_x is the larger sample variance while s^2_y is the smaller sample variance

(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



<u>p-value</u>: 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 **Use F.DIST function*

Example - F-test

An important measure of the risk associated with a stock is the standard deviation, or variance, of the stock's price movements. A financial analyst wants to test the one-tailed hypothesis that stock A has a greater risk (larger variance of price) than stock B. A random sample of 25 daily prices of stock A gives s2A=6.52, and a random sample of 22 daily prices of stock B gives a sample variance of s2B=3.47. Carry out the test at α=0.01.

Hypothesis Tests using Python

z-test

statsmodels.stats.weightstats.**ztest**(*x1*, *x2=None*, *value=0*, *alternative='two-sided'*)
Link to refer -
https://www.statsmodels.org/stable/generated/statsmodels.stats.weightstats.ztest.html

t-test

scipy.stats.ttest_ind(*a*, *b*, *axis=0*, *equal_var=True*, *nan_policy='propagate'***)**

Link to refer -

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html

Chi-square (χ^2) test

scipy.stats.chisquare(*f_obs*, *f_exp=None***)**
Link to refer -
https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chisquare.html

F-test

```
alpha = 0.05 #Or whatever you want your alpha to be.
 p_value = scipy.stats.f.cdf(F, df1, df2)
if p_value > alpha: # Reject the null hypothesis that Var(X) == Var(Y)
```

Link to refer - https://stackoverflow.com/questions/21494141/how-do-i-do-a-f-test-in-python

Hypothesis Testing Using Python

One Sample Testing

Some important functions:

1. *t_statistic, p_value = ttest_1samp(array, n)*

Here n= sample number , daily_intake= array

2. *z_statistic, p_value = wilcoxon(array – n)*

The Wilcoxon test is used if the data is demostrably not normally distributed.

greatlearning
Learning for Life

Hypothesis Testing Using Python

Two Sample Testing

Some important functions:

1. *t_statistic, p_value = ttest_ind(group1, group2)*
2. *u, p_value = mannwhitneyu(group1, group2)*
3. *t_statistic, p_value = ttest_1samp(post-pre, 0)*
4. *z_statistic, p_value = wilcoxon(post-pre)*
5. *levene(pre,post)*
6. *shapiro(post)*

ANOVA- One Way Classification

- The samples drawn from different populations are independent and random.
- The response variables of all the populations are normally distributed.
- The variances of all the populations are equal.

Hypothesis of One-Way ANOVA

- $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \ldots = \mu_k$
  - All population means are equal
- H_1 : Not all of the population means are equal
  - For at least one pair, the population means are unequal.

One-Way ANOVA