Statistical Inference

Non-Parametric Tests II

Contents

Kruskal Wallis test

- The Kruskal Wallis test is considered a non-parametric alternative to one way analysis of variance (ANOVA).
- The Kruskal Wallis test is used to compare differences between more than two independent groups when the dependent variable is either ordinal or continuous, but not normally distributed.
- H0: K samples come from the same population
  H1: Not H0.

Kruskal Wallis test procedure

- Combine all the observations from k samples into a single sample of size n and arrange them in ascending order .
- Assign ranks to them from smallest to largest as 1 to n. if there is a tie at two or more places, each observation is given the mean of the ranks for which it is tied.
- The ranks assigned to observations in each of the k groups are added separately to give k rank sums.
- The test statistic is

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_j^2}{n_j} - 3(n+1)$$

$n_j = \text{number of observations in } j^{th} \text{ sample}$

$n = \text{number of observations in the combined sample}$

$R_j = \text{sum of the ranks in the } j^{th} \text{ sample.}$

- H follows Chi Square Distribution with k-1 df

Case Study - 1

Background

The data consists of the aptitude scores of 3 groups of employees.

Objective

To check whether there is difference in scores among the three groups.

Sample Size

Sample size: 20
Variables: aptscore, Group

Data Snapshot

Kruskal Wallis Test

Variables

Observations

aptscore	Group
11.5	GroupI
12.1	GroupI
14.5	GroupI
16.1	GroupI
19.8	GroupI
23.6	GroupI
11	GroupII
12.2	GroupII
15	GroupII
15.7	GroupII
18.5	GroupII

Combined sample arranged from smallest to largest

Ordered sample	Group	Rank
11	GroupII	1
11.1	GroupIII	2
11.5	GroupI	3
12.1	GroupI	4
12.2	GroupII	5
14.5	GroupI	6
15	GroupII	7
15.7	GroupII	8
16.1	GroupI	9
18.5	GroupII	10
19.1	GroupIII	11
19.8	GroupI	12
21	GroupIII	13
22.2	GroupIII	14

Rank ordered sample

Columns	Description	Type	Measurement	Possible values
aptscore	Aptitude score	Numeric	-	Positive values
Group	Group of employees	character	GroupI, GroupII, GroupIII	3

Kruskal Wallis test

Testing distribution of more than two samples

Objective	To test the **null hypothesis** that all the samples came from same population

Null Hypothesis (H_0): The three samples are from the same population
Alternate Hypothesis (H_1): The three samples do not come from the same population

Test Statistic	$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(n+1)$ n_j = number of observations in j^{th} sample n = number of observations in the combined sample R_j = sum of the ranks in the j^{th} sample.
Decision Criteria	Reject the null hypothesis if the p-value < 0.05

Kruskal Wallis test example

Calculations :

	Value
Sample size	$n_1 = 6$ $n_2 = 7$ $n_3 = 7$
R_1	50
R_2	68
R_3	92
H	2.2309
p-value	0.3278

Kruskal Wallis test in R

```
# Import the CSV file
data<-read.csv("Kruskal Wallis Test.csv",header=TRUE)

# Kruskal walis test
kruskal.test(formula=aptscore~Group,data=data)
```

- *kruskal.test performs the Kruskal waliss test on the data.*
- *aptscore is the analysis variable.*
- *Group is the factor variable.*

Kruskal Walis test in R

Output:

```
        Kruskal-Wallis rank sum test

data:  aptscore by Group
Kruskal-Wallis chi-squared = 2.2309, df = 2, p-value = 0.3278
```

Interpretation :

- *Since the p-value is >0.05, do not reject H0. Aptitude score is the same for all three groups of employees.*

Chi-square test of Association

- The chi-square test for independence, also called as Pearson's chi-square test or the chi-square test of association, is used to test if there is a relationship between two categorical variables.

- The two categorical variables can be nominal or ordinal.

- H0: Two attributes are independent (not associated)
  H1: Not H0.

Chi-square test procedure

- Assume that there are 'r' categories of attribute A and 'c' categories of attribute B. Therefore, we have a cross table of r*c (r rows and c columns).

- Let Ri be the total of the ith row and Cj be the total of the jth column.

- Observed frequencies are calculated from the data.
  Oij: Observed frequency in ith row and jth column.

- Expected frequencies are given by Eij = (Ri * Cj)/ n where n is total sample size.
  Expected frequencies are computed under the null hypothesis.

- Test statistic

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

  where Oij are the observed frequencies in the ith row and jth column.
  Eij are the expected frequencies in the ith row and jth column.

- χ^2 follows a Chi-Square Distribution with (r-1)(c-1) degrees of freedom.

Case Study - 2

Background

The data consists of information regarding the Performance & Recruitment Source of employees.

Objective

To check whether Performance & Source of Recruitment are associated.

Sample Size

Sample size: 870
Variables: sn, performance, source

Data Snapshot

chi square test of association

Variables

Observations

sn	performan	source
1	Excellent	Internal
2	Excellent	Internal
3	Excellent	Internal
4	Excellent	Internal
101	Excellent	Campus
102	Excellent	Campus
251	Excellent	Jobportal
252	Excellent	Jobportal
253	Excellent	Jobportal
254	Excellent	Jobportal
291	Good	Internal
292	Good	Internal
293	Good	Internal
491	Good	Jobportal
492	Good	Jobportal
493	Good	Jobportal
591	Poor	Internal

Columns	Description	Type	Measurement	Possible values
sn	Serial number	Numeric	-	-
performance	Employee performance	character	Excellent, Good,Poor	3
source	Source of recruitment	Character	Campus, Internal, Jobportal	3

Get the observed frequency (count) table from this data.

Chi-square test of Association

Testing association between two categorical variables

Objective	To test the **null hypothesis** that two categorical variables are **independent**

Null Hypothesis (H_0): performance and source are not associated

Alternate Hypothesis (H_1): performance and source are associated

Test Statistic	$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$ **Oij = observed frequencies in the ith row and jth column. Eij = expected frequencies in the ith row and jth column.**
Decision Criteria	Reject the null hypothesis if the p-value < 0.05

Chi-square test example

Observed Frequency table

	Recruitment Source			
Performance	Campus	Internal	Jobportal	Total
Excellent	150	100	40	290
Good	100	100	100	300
Poor	80	50	150	280
Total	330	250	290	870

Expected Frequency table

	Recruitment Source			
Performance	Campus	Internal	Jobportal	Total
Excellent	=(330*290)/870	83	97	290
Good	114	=(250*300)/870	100	300
Poor	106	80	=(290*280)/870	280
Total	330	250	290	870

	Value
r	3
c	3
χ^2	107.3786

Chi-Square test in R

```
# Import the CSV file
data<-read.csv("chi square test of association.csv", header=TRUE)
```

```
# Install and use the package “gmodels”
install.packages("gmodels")
library(gmodels)
```

“gmodels” is needed for the contingency table. The table displays frequencies, relative frequencies of two categorical variables.

```
# Chi-square test of association
CrossTable(data$performance, data$source, chisq=TRUE)
```

CrossTable function performs Chi-square test of association when chisq=TRUE.

Chi-Square test in R

```
# Output:
```

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  870
```

```
                 | data$source
data$performance |    Campus |  Internal | Jobportal | Row Total |
-----------------|-----------|-----------|-----------|-----------|
       Excellent |       150 |       100 |        40 |       290 |
                 |    14.545 |     3.333 |    33.218 |           |
                 |     0.517 |     0.345 |     0.138 |     0.333 |
                 |     0.455 |     0.400 |     0.138 |           |
                 |     0.172 |     0.115 |     0.046 |           |
-----------------|-----------|-----------|-----------|-----------|
            Good |       100 |       100 |       100 |       300 |
                 |     1.672 |     2.207 |     0.000 |           |
                 |     0.333 |     0.333 |     0.333 |     0.345 |
                 |     0.303 |     0.400 |     0.345 |           |
                 |     0.115 |     0.115 |     0.115 |           |
-----------------|-----------|-----------|-----------|-----------|
            Poor |        80 |        50 |       150 |       280 |
                 |     6.467 |    11.531 |    34.405 |           |
                 |     0.286 |     0.179 |     0.536 |     0.322 |
                 |     0.242 |     0.200 |     0.517 |           |
                 |     0.092 |     0.057 |     0.172 |           |
-----------------|-----------|-----------|-----------|-----------|
    Column Total |       330 |       250 |       290 |       870 |
                 |     0.379 |     0.287 |     0.333 |           |
-----------------|-----------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  107.3786     d.f. =  4     p =  2.635987e-22
```

Interpretation :

- *Since the p-value is <0.05, reject H0. Recruitment source and employee performance are associated.*

Quick Recap

Kruskal Wallis test	• Nonparametric alternative to one way ANOVA.
Chi-Square test	• Also called Pearson's chi-square test or the chi-square test of association. It is used to test if there is a relationship between two categorical variables (nominal or ordinal).

Statistical Inference

Non Parametric Tests - 1

Contents

Non-parametric statistical test

- Tests based on t and F distributions assume that populations are normally distributed.
- A large body of statistical methods is available which do not make assumptions about the nature of the distribution(e.g. normality)
- These testing procedures are termed nonparametric tests or distribution-free tests.
- If the underlying assumptions of the parametric test are met, then a parametric test will be more powerful than a non-parametric test.

* Note : Always check for normality assumptions using the test explained earlier and then decide which hypothesis test is more accurate, depending upon the problem statement.

Mann-Whitney test

- The Mann-Whitney test is considered as a non-parametric alternative to t test for independent samples.
- The Mann-Whitney U test is used to compare differences between two independent groups when the dependent variable is either ordinal or continuous, but not normally distributed.
- The test is equivalent to Wilcoxon rank-sum test (WRS).
- The null hypothesis is that the distributions of both groups are identical, so that there is a 50% probability that an observation randomly selected from one population exceeds an observation randomly selected from another population.

Mann-Whitney test

- **Steps to follow :**
  - Combine the two samples.
  - Rank all the observations from smallest to largest.
  - Keep track of the group to which each observation belongs.

- Tied observations (observations with same value) are assigned a rank equal to the mean of the rank positions for which they are tied.

- The test statistic is

$$U = T - \frac{m(m+1)}{2}$$

Where T is the sum of the ranks of the first sample in the combined ordered sample, m and n are sample sizes.

$$E(U) = \frac{mn}{2} \qquad V(U) = \frac{mn(m+n+1)}{12}$$

- Standardized U is assumed to follow normal distribution.

- Compare the p-value with the level of significance & conclude.

Case Study - 1

Background

Data consists of the aptitude scores of 2 groups of employees.

Objective

To compare aptitude scores of the two groups and test if they come from the same population.

Sample Size

Sample size: 13
Variables: aptscore, Group

Data Snapshot

Mann Whitney test

Variables

Observations

aptscore	Group
11.5	G1
23.6	G1
12.1	G1
14.5	G1
16.1	G1
19.8	G1
11	G2
25	G2

Combine the two sample from smallest to largest

Ordered Sample	Group	Rank
11	GroupII	1
11.5	GroupI	2
12.1	GroupI	3
12.2	GroupII	4
14.5	GroupI	5
15	GroupII	6
15.7	GroupII	7
16.1	GroupI	8
18.5	GroupII	9

Rank ordered samples

Columns	Description	Type	Measurement	Possible values
aptscore	Aptitude score	Numeric	-	Positive values
Group	group of employees	character	G1,G2	2

- T is the sum of the ranks of the first sample in the combined ordered sample. m and n are sample sizes.
  T=39 , m=6, n= 7
  U=18 , E(U)=21 ,V(U)= 49

Mann-Whitney test

Testing distribution of two samples

Objective	To test the **null hypothesis** that **the median** of both samples is the same

Null Hypothesis (H_0): The two samples come from the same population

Alternate Hypothesis (H_1): The two samples do not come from the same population

Test Statistic	$U = T - \frac{m(m+1)}{2}$ **Where T is the sum of the ranks of first sample in the combined ordered sample, m and n are sample sizes**
Decision Criteria	Reject the null hypothesis if the p-value < 0.05

Mann-Whitney test in R

```
# Import the CSV file
data<-read.csv("Mann Whitney test.csv", header=TRUE)

# Mann-Whitney test
wilcox.test(formula=aptscore~Group,data=data)
```

- *The Mann-whitney test is also known as the Wilcoxon Rank Sum test.*
- *The wilcox.test function gives the value of U(as W) and p-value.*
- *aptscore is the analysis variable.*
- *group is the factor.*

Mann-Whitney test in R

```
# Output:
```

```
        Wilcoxon rank sum test

data:  aptscore by Group
W = 18, p-value = 0.7308
alternative hypothesis: true location shift is not equal to 0
```

Interpretation :

- *Since p-value is >0.05, do not reject H0. aptitude score is same for both the groups i.e. samples come from the same population.*

Wilcoxon Signed Rank Test for paired data

- The Wilcoxon Signed Rank test is considered as a nonparametric alternative to paired t test .
- The Wilcoxon Signed Rank test is used to compare differences between two related or paired groups when the variable is either ordinal or continuous, but not normally distributed.
- H0: The median of difference in the population is zero
  H1: Not H0.

Wilcoxon Signed Rank Test for paired data

- **Steps to follow :**
  - Define Di = Xi- Yi, which are the differences between two values for each pair.
  - Obtain |Di|, which are absolute values of differences.
  - Rank all |Di| from smallest to largest.
  - Define Ri = rank of |Di|.
  - Obtain 'W', which is the sum of the ranks associated with positive Di.

- The test statistic is W, which is the sum of the ranks associated with positive Di. n is the sample size.

$$E(W) = \frac{n(n+1)}{4}$$

$$V(W) = \frac{n(n+1)(2n+1)}{24}$$

- Standardized W is assumed to follow normal distribution.

- Compare the p-value with the level of significance & conclude.

Case Study - 2

Background

A company organized a training program and the scores before and after training were recorded.

Objective

To test whether the median of paired samples is same.

Sample Size

Sample size: 12
Variables: Before, After

Data Snapshot

- A company organized a training program and the scores before and after training were recorded.

Variables

Observations

Before	After
58	74
52	65
61	60
48	45
50	58
39	53

Combine the two sample to get Di

Di=Before-After	Abs(Di)	Rank (Di)
-16	16	12
-13	13	10
1	**1**	**1**
3	**3**	**3**
-8	8	5
-14	14	11
-9	9	6

Rank associated with positive Di

Columns	Description	Type	Measurement	Possible values
Before	Score before training	Numeric	-	Positive values
After	Score after training	Numeric	-	Positive values

- W is sum of the ranks associated with positive Di. n is sample size.
  W=4 , n= 12
  E(W)=39 , V(W)= 162.5

Wilcoxon Signed Rank Test for paired data

Testing distribution of paired samples

Objective	To test the **null hypothesis** that **median** of paired samples is same.

Null Hypothesis (H_0): The median of the difference in the population is zero
Alternate Hypothesis (H_1): The median of the difference in the population is less than zero.

Test Statistic	**w=sum of the ranks associated with positive Di.** **Di = Xi- Yi which are the differences between data and specified median value.**
Decision Criteria	Reject the null hypothesis if p-value < 0.05

Wilcoxon Signed Rank Test for paired data **in** R

```
# Import the CSV file
data<-read.csv("Wilcoxon Signed Rank test for paired  data.csv",
               header=TRUE)
# Wilcoxon Signed Rank test
wilcox.test(data$Before, data$After, paired=TRUE,
            alternative = "less")
```

- *wilcox.test function gives the value of W (as V) and p-value.*
- *wilcox.test function performs Wilcox signed rank test for paired data when paired=TRUE is specified.*
- *Before and After are the paired observations.*
- *alternative=less specifies one tail test .since, score will be more if training program is effective.*

Wilcoxon Signed Rank Test for paired data in **R**

Output:

```
        Wilcoxon signed rank test

data:  data$Before and data$After
V = 4, p-value = 0.001709
alternative hypothesis: true location shift is less than 0
```

Interpretation :

- *Since the p-value is <0.05, reject H0. The training program is effective as the score after training is more than before training.*

Quick Recap

Non Parametric Test	• Non parametric tests are performed if the normality assumption is not satisfied.
Mann-Whitney test	• Nonparametric alternative to the t test for independent samples.
Wilcoxon Signed Rank test	• Nonparametric alternative to the t test for paired samples.

Statistical Inference

Post Hoc analysis in ANOVA

Contents

Post Hoc testing in ANOVA

- Post-hoc (Latin, meaning “after this”) means analyzing results further after main analysis. They are often based on a **familywise error rate -** it is the probability of making at least one Type I Error.

- Post-hoc pairwise comparisons are commonly performed after significant effects have been found when there are three or more levels in a factor.

- After an ANOVA, you may know that the means of your response variable differ significantly across your factors, but you do not know which pairs of the factor levels are significantly different from each other. At this point, you can conduct pairwise comparisons.

- **Post hoc tests:**
  - Pairwise t tests (Not recommended)
  - Pairwise t tests using the Bonferroni adjustment
  - Tukey test

Case Study

To execute a Post Hoc analysis in ANOVA in R, we shall consider this case as an example.

Background

A company has recorded aptitude scores for three groups of employees.

Objective

To test whether there is no significant difference in the Mean Aptitude score of three groups of employees.

Sample Size

Sample size: 24
Variables: aptscore, group

Data Snapshot

Variables

Observations

aptscore	group
34	GrI
28	GrI
41	GrI
31	GrI
39	GrI
44	GrI
25	GrI
20	GrI
27	GrII
31	GrII
33	GrII

Columns	Description	Type	Measurement	Possible values
aptscore	Aptitude score	Numeric	-	Positive value
group	Group of employees	character	GrI,GrII,GrIII	3

ONE WAY ANOVA

Testing **equality** of more than two means

Objective	To test the **null hypothesis** that **means** scores **are same**

Null Hypothesis (H_0): **Mean Aptitude score of three groups are equal**

$$\mu_1 = \mu_2 = \mu_3$$

Alternate Hypothesis (H_1): At least one group mean is different than other

Test Statistic	**Test statistic is based on F distribution.**
Decision Criteria	Reject the null hypothesis **if p-value < 0.05**

ANOVA in R

```
# Import data
data<-read.csv("Post Hoc Tests-Anova.csv", header=TRUE)

# Anova table
anova<-aov(formula=aptscore~group,data=data)
summary(anova)
```

- *'aov' is the R function for ANOVA .*
- *anova is user defined object name created to store output.*
- *summary function displays the ANOVA table output.*

ANOVA in R

```
# Output:
```

```
              Df Sum Sq Mean Sq F value Pr(>F)  
group          2  501.7  250.87   4.832 0.0188 *
Residuals     21 1090.2   51.92                 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation :

- *P-value<0.05,reject h0. There is significant difference in mean aptitude scores of three groups of employees.*
- *Now ,we will do the pairwise testing to identify which pairs are having difference in aptitude score.*

Pairwise t tests

```
pairwise.t.test(data$aptscore,data$group,p.adj="none")
```

- *pairwise.t.test is the R function for pairwise comparison .*
- *p.adj =none specifies no adjustments are used.*

```
# Output:
```

```
        Pairwise comparisons using t tests with pooled SD

data:  data$aptscore and data$group

      GrI    GrII
GrII  0.6082 -
GrIII 0.0261 0.0083
```

Interpretation :

- *P-value<0.05 for group I and III and also for group II and III. Aptitude test score is significantly different between Group I and III and Group II and III.*

Why Bonferroni Adjustment is needed ?

- The Bonferroni correction is used to limit the **possibility of getting a statistically significant result** when testing multiple hypotheses. It's needed because the more tests you run, the more likely you are to get a significant result. The correction lowers the area where you can reject the null hypothesis. In other words, it makes your p-value smaller.

- **Example :** Imagine looking for the Ace of Clubs in a deck of cards: if you pull one card from the deck, the odds are pretty low (1/52) that you'll get the Ace of Clubs. Try again (and try perhaps 50 times), you'll probably end up getting the Ace. The same principal works with hypothesis testing: the more simultaneous tests you run, the more likely you'll get a "significant" result. Let's say you were running 50 tests simultaneously with an alpha level of 0.05. The probability of observing at least one significant event due to chance alone is:
- P (significant event) = 1 – P(no significant event)
- = 1 – (1-0.05)^50 = 0.92.
- That's almost certain (92%) that you'll get at least one significant result.

Pairwise t tests using Bonferroni adjustment

```
pairwise.t.test(data$aptscore,data$group,p.adj="bonf")
```

- *pairwise.t.test is the R function for pairwise comparison .*
- *p.adj =bonf specifies Bonferroni adjustments are used.*

```
# Output:
        Pairwise comparisons using t tests with pooled SD

data:  data$aptscore and data$group

      GrI   GrII
GrII  1.000 -
GrIII 0.078 0.025

P value adjustment method: bonferroni
```

Interpretation :

- *P-value<0.05 for group II and III. Aptitude test score is significantly different for Group II and III.*

Get an Edge

```
#pairwise t tests with no adjustment
pairwise.t.test(data$aptscore, data$group, p.adj="none")
```

```
        Pairwise comparisons using t tests with pooled SD

data:  data$aptscore and data$group

      GrI    GrII
GrII  0.6082 -
GrIII 0.0261 0.0083
```

```
#Bonferroni adjustment multiples p value by number of
#comparisons k(here k=3).Note that maximum p value can be one.
pairwise.t.test(data$aptscore, data$group, p.adj="bonf")
```

```
        Pairwise comparisons using t tests with pooled SD

data:  data$aptscore and data$group

      GrI   GrII
GrII  1.000 -
GrIII 0.078 0.025

P value adjustment method: bonferroni
```

Tukey test

The purpose of the Tukey's test is to figure out which groups in your sample differ. It uses the "Honest Significant Difference", a number that represents the distance between groups, to compare every mean with every other mean.

```
TukeyHSD(anova,"group")
```

- *TukeyHSD test is used for pairwise comparison .*
- *anova is the object created in aov funvtion.*
- *group is the factor variable.*

Tukey test

Output:

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = aptscore ~ group, data = data)

$`group`
              diff       lwr        upr     p adj
GrII-GrI     1.875  -7.20576 10.9557603 0.8622638
GrIII-GrI   -8.625 -17.70576  0.4557603 0.0646166
GrIII-GrII -10.500 -19.58076 -1.4192397 0.0216622
```

Interpretation :

- *P-value<0.05 for group II and III. Aptitude test score is significantly different for Group II and III.*

Quick Recap

Post Hoc Analysis	• Post-hoc pairwise comparisons are commonly performed after significant effects have been found when there are three or more levels of a factor.
Post Hoc tests	• Pairwise t tests • Pairwise t test using Bonferroni adjustments • Tukey test

Statistical Inference

Multiway Factorial Analysis of Variance

Contents

Multiway Factorial Analysis of Variance

- Two Way Anova can be extended to assess the effects of simultaneous applications of three or more factors.

- **Example:** If there are three factors, say A,B and C then we can study

    - The Main effects of A,B and C
    - Two way interactions A*B, A*C and B*C
    - Three way interaction A*B*C

- The researcher may decide to exclude higher order interactions as they are difficult to interpret.

Case Study

Background

Two new marketing campaigns are tested along with traditional campaign(Control).
The campaigns are tested in mid size and large size stores of 3 regions (North ,west, east).

Objective

To test whether there is significant difference in growth among three campaigns, three regions & two sizes.

Sample Size

Sample size: 72
Variables: campaign, region, size, growth

Data Snapshot

Three Way Anova

Variables

Observations

campaign	region	size	growth
Test1	north	mid	11.9
Test1	north	mid	11.8
Test1	north	mid	11.6
Test1	north	mid	11.4
Test1	north	large	11.8
Test1	north	large	11.7
Test1	north	large	11.4
Test1	north	large	11.5
Test1	west	mid	12.3
Test1	west	mid	12.1
Test1	west	mid	12
Test1	west	mid	12.6

Columns	Description	Type	Measurement	Possible values
campaign	Campaign	Character	Control, Test1, Test2	3
region	Region	Character	east, north, west	3
size	Size of the store	Character	large, mid	2
growth	Growth in sales	Numeric	Percentage	+/- values

Three-way ANOVA in R

```
# Import data
data<-read.csv("Three Way Anova.csv", header=TRUE)
```

```
# Anova table
anovatable<-aov(formula=growth~campaign*region*size,data=data)
summary(anovatable)
```

- *'anovatable' is user defined object name created to store output.*
- *'aov' is the R function for ANOVA .*
- *formula specifies 'growth' as analysis (dependent) variable and 'campaign','region',' size' as factor (independent) variable.*
- *summary function displays the ANOVA table output.*

Three-way ANOVA in R

```
# Output:
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
campaign	2	1.818	0.909	24.475	2.71e-08	***
region	2	24.656	12.328	332.024	< 2e-16	***
size	1	0.009	0.009	0.239	0.6266	
campaign:region	4	1.102	0.275	7.418	7.75e-05	***
campaign:size	2	0.370	0.185	4.986	0.0103	*
region:size	2	0.175	0.088	2.360	0.1041	
campaign:region:size	4	0.221	0.055	1.485	0.2196	
Residuals	54	2.005	0.037			

Interpretation :

- *Since p-value is <0.05 for campaign, region, reject H0. there is significant difference in growth among three campaigns and three regions. Also, campaign*region and Campaign*size interaction is significant.*

Visualize main effects

```
# Box plots for main effects
```

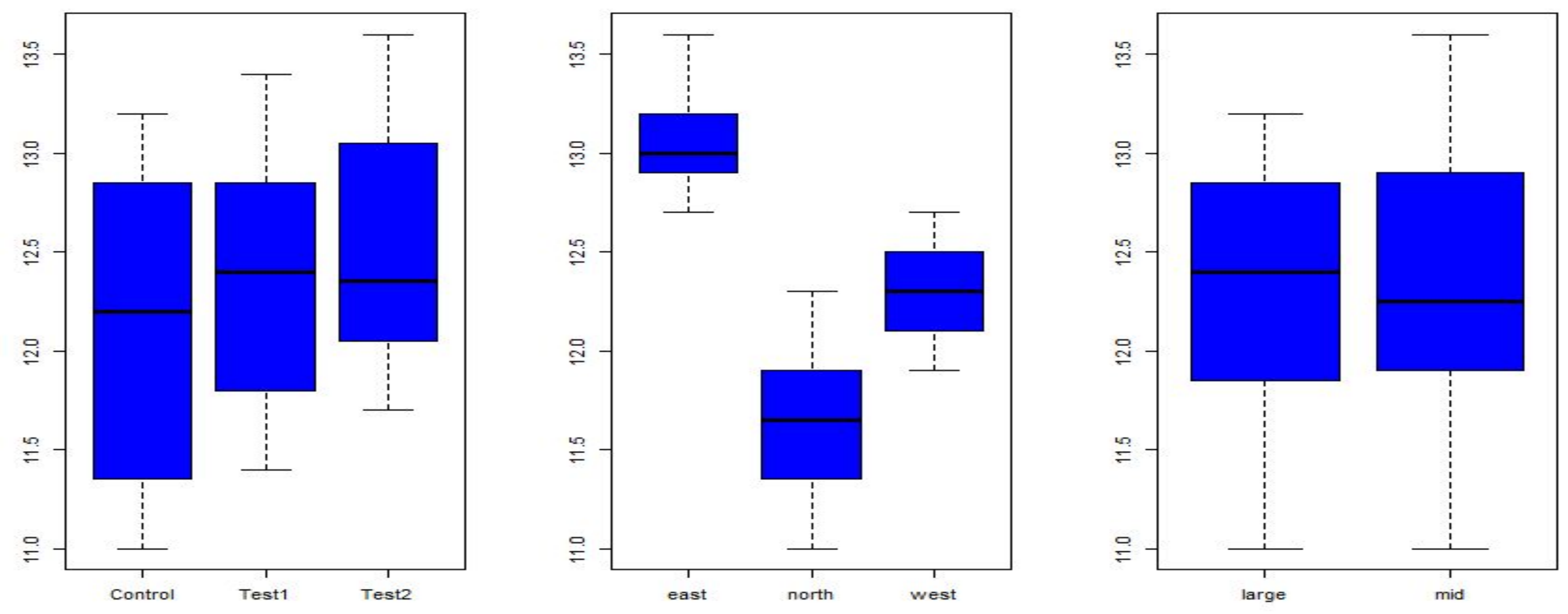
```
par(mfrow=c(1,3))

boxplot(growth~campaign,data=data,col="blue")
boxplot(growth~region,data=data,col="blue")
boxplot(growth~size,data=data,col="blue")
```

- *'par' function creates partition for graph output in 1 row and 3 columns.*
- *Boxplots are plotted for each factor with growth variable.*

Visualize main effects

```
# Output:
```



Interpretation :

- *There is significant difference in growth among three campaigns and three regions.*

Visualize interaction effects

```
# Box plots for interaction effects
```

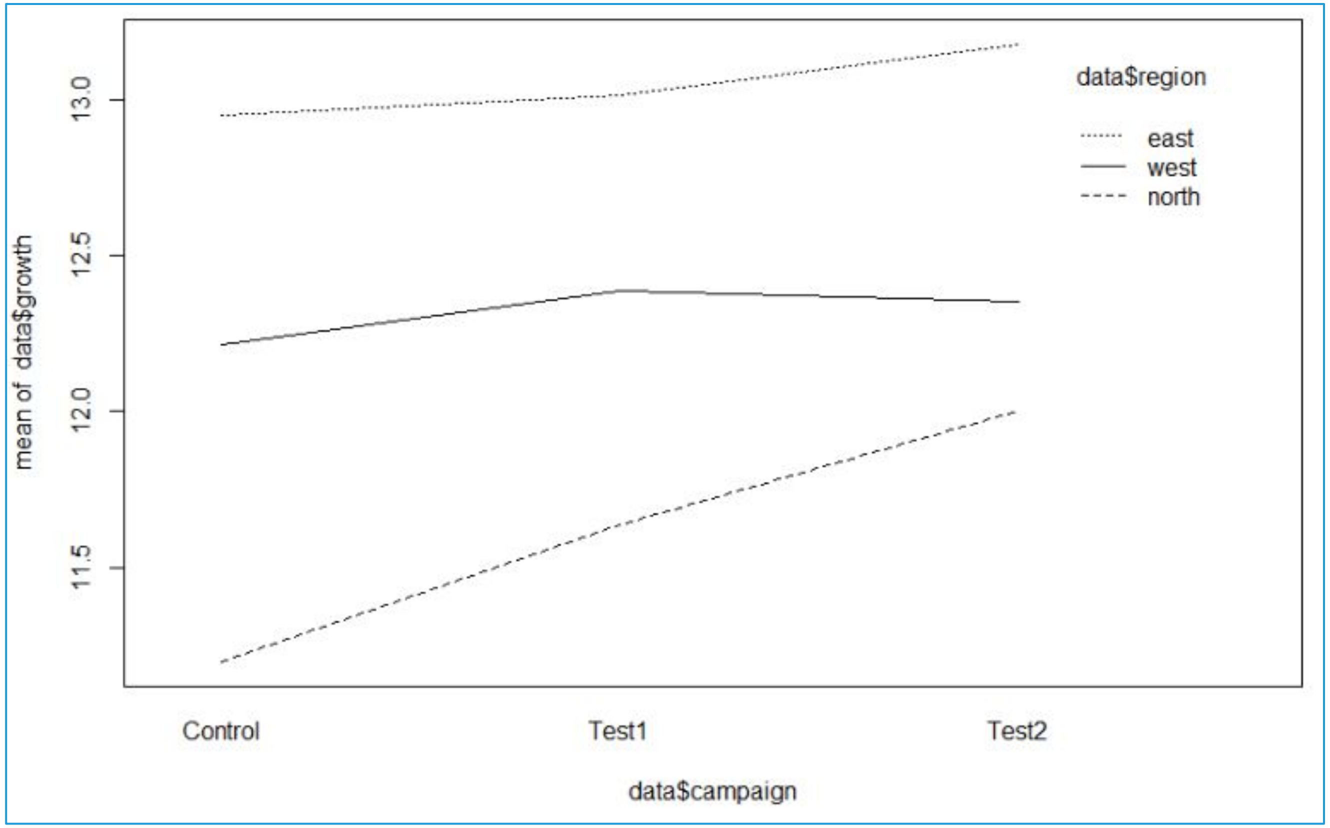
```
par(mfrow=c(1,1))

interaction.plot(data$campaign, data$region, data$growth)
```

- *interaction.plot function plots the mean (or other summary) of the response for two-way combination of factors (campaign, region), thereby illustrating possible interactions.*

Visualize interaction effects

```
# Output:
```



Interpretation :

- *Campaign and region interaction is having impact on growth of sales.*

Quick Recap

Three Way ANOVA	Two Way Anova can be extended to assess the effects of simultaneous applications of 3 or more factors.
Main Effects and Interactions	There are 3 main effects- One for each factor Interaction effects are 2-way and 3 –way.
Data Visualization	Box-Plot for main effect Interaction effect plot

Statistical Inference

Two- Way Analysis of Variance

Contents

Two Way ANOVA

- Two Way Anova is used when there are 2 factors under study.

- Each factor can have 2 or more levels . Example: Gender and Age can be 2 factors.
  Gender with 2 levels as Male and Female
  Age with 3 levels as 18-30, 31-50 and >50

- Three hypothesis are tested.

  Factor A H0: All group means are equal
  H1: At least one mean is different from other means

  Factor B H0: All group means are equal
  H1: At least one mean is different from other means

  Interaction H0: The interaction is not significant
  H1: The interaction is significant

* For two-way ANOVA with interaction there has to be more than one observation per combination of the levels of factors.

Two Way ANOVA

Total variation is partitioned as below :

Total SS= Between Groups SS due to factor A (SSA)

+ Between Groups SS due to factor B (SSB)

+ Interaction SS due to factor A and B (SSAB)

+ Error SS (SSE)

where, SS stands for sum of squares

* SS formulae for two-way ANOVA with interaction are not specified due to their complexity.

Case Study

Background

A large company is assessing the difference in 'Satisfaction Index' of employees in Finance, Marketing and Client-Servicing departments. Experience level is also considered in the study.(<=5 years and >5 years)

Objective

To test the equality of the satisfaction index among employees of three departments (CS, Marketing, Finance) and among different experience bands.

Sample Size

Sample size: 36
Variables: satindex, dept, exp

Data Snapshot

Two Way Anova

Variables

Observations

satindex	dept	exp
75	FINANCE	lt5
56	FINANCE	lt5
62	FINANCE	gt5
66	FINANCE	gt5
58	FINANCE	gt5
58	MARKETIN	lt5
63	MARKETIN	lt5
53	MARKETIN	lt5
74	MARKETIN	lt5
77	MARKETIN	lt5
69	MARKETIN	lt5
57	MARKETIN	gt5
70	MARKETIN	gt5
68	MARKETIN	gt5
77	CS	lt5
71	CS	lt5

Columns	Description	Type	Measurement	Possible values
Satindex	Satisfaction Index	Numeric	-	Positive Values
Dept	Department	Character	MARKETING, CS, FINANCE	3
Exp	Years of Experience (grouped)	Character	lt5 = less than 5, gt5 = greater than 5	2

Two Way ANOVA

Testing equality of means in two factors.

Objective	To compare employee satisfaction index in three departments (CS, Marketing, Finance) and two experience level based groups.

Null Hypothesis

(H_{01}): Average satisfaction index is equal for 3 departments.

(H_{02}): Average satisfaction index is equal for 2 experience levels.

(H_{03}) Interaction effect(dept*exp) is not significant on satisfaction index.

The test statistic is computed for each of these null hypothesis.

Reject the null hypothesis if $p\text{-value} < 0.05$

Two Way ANOVA in R

```
# Import data
```

```
data<-read.csv("Two Way Anova.csv", header=TRUE)
```

```
# ANOVA Table
```

```
anovatable<-aov(formula=satindex~dept+exp+dept*exp,data=data)
summary(anovatable)
```

- *'aov' is the R function for ANOVA .*
- *formula specifies 'satindex' as analysis (dependent) variable and 'dept' and 'exp' as factor (independent) variables.*
- *dept*exp specifies the interaction effect.*
- *anovatable is user defined object name created to store output.*
- *summary function displays the ANOVA table output.*

Two Way ANOVA in R

```
# Output:
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
dept	2	164.2	82.11	1.679	0.204
exp	1	78.0	78.03	1.595	0.216
dept:exp	2	20.2	10.11	0.207	0.814
Residuals	30	1467.2	48.91		

Interpretation :

- *Since p-value is >0.05 for all three (dept, exp and dept*exp), do not reject H0 for all three tests. There is no significant difference in satisfaction index among 3 different departments and 2 experience levels.*
- *Also interaction effect is not significant.*

Quick Recap

Two Way Anova	• two way ANOVA is an extension of one way ANOVA when we have 2 factors in the study instead of one.
Null Hypothesis Drawing Inference	• Equality of means for levels in factor A • Equality of means for levels in factor B • No Interaction effect between 2 factors • Total sum of squares is split into 4 parts and each hypothesis is tested.

Knowledge check question

- A large retailer is testing a marketing campaign on 24 stores. 8 stores are selected randomly from each of 3 zones.
- The variable of interest is ' sales increment(%) during campaign month'. Objective is to test whether the campaign is equally effective in 3 regions. Data is given below.

NORTH	WEST	SOUTH
8	10.2	5.3
12.5	9.3	5.8
9.2	9.9	6
6.7	8.7	7.1
9.4	9.1	7
5.9	10.2	6.1
7.7	9.5	6.3
6.9	10	7.3

Is this One-way ANOVA problem or Two-way ANOVA problem?

ANSWER : One-way ANOVA

EXPLANATION : There is only one factor (zone) with 3 levels (North, West, South).

Statistical Inference

Analysis of Variance

Contents

Analysis of Variance (ANOVA)

Analysis of variance (ANOVA) is a collection of statistical models used to analyze the differences among more than two group means developed by statistician and evolutionary biologist **Ronald Fisher**.

Example: There are 20 plots of wheat and 5 fertilizers applied to four different plots. The yield of wheat is recorded for each of the 20 plots.

ANOVA can be used to find out whether the effect of these fertilisers on yields is equal or significantly different.

ANOVA

- Note that although the name is 'Analysis of Variance', the method is used to analyze the differences among group means.

- Variation in the variable is inherent in nature. In general, the observed variance in a particular variable is partitioned into components attributable to different sources of variation.

- The total variance in any variable is due to a number of causes which may be classified “assignable causes (which can be detected and measured)” and “chance causes (which are beyond human control and cannot be traced separately)”.

- Hence, ANOVA is the separation of variance ascribable to one group of causes from the variance ascribable to another.

ANOVA assumptions

The assumptions of ANOVA are listed below:

- The samples drawn are random samples.

-The populations from which samples are drawn have equal & unknown variances.

-The populations follow a normal distribution.

Testing Normality assumption

- An assessment of the normality of data is a prerequisite for many statistical tests because normal data is an underlying assumption in parametric testing.

- Normality can be assessed using two approaches: graphical and numerical.

  - **Graphical approach**

    - Box-Whisker plot (used to assess symmetry rather than normality)
    - Quantile-Quantile plot (Q-Q plot).

  - **Statistical approach**

    - Shapiro-Wilk test
    - Kolmogorov-Smirnov test

* **Normality tests are covered in the Parametric test session.**

One Way ANOVA

- One Way ANOVA can be considered as an extension of the t test for independent samples.
- One Way ANOVA is used to test the equality of K population means.
  (when K=2, t test can be used.)
- For two levels (K=2),the t test and One Way ANOVA provide identical results.
- The **Mathematical model** is :

  $X_{ij} = \mu_i + \varepsilon_{ij}$

  Where X_{ij} is the jth observation due to ith level of a factor. μ_i is the effect of ith level of a factor. ε_{ij} is the error term. $i=1,2,..,k$; $j=1,2,...,n_i$

- The null hypothesis is

$$H_0 : \mu_1 = \mu_2 = \dots\dots\dots\dots\dots\dots = \mu_K = \mu$$

Partitioning Total Variance

- Total variation is partitioned into two parts:
  Total SS= Between Groups SS + Within Groups SS
  where, SS stands for sum of squares

$$SS_{total} = \sum_i \sum_j \left(X_{ij} - \bar{X}_{..} \right)^2$$

Total variation
(Total SS)

$$SS_{between} = \sum_i n_i \left(\bar{X}_{i.} - \bar{X}_{..} \right)^2$$

Variation due to
Assignable causes
(Between SS)

$$SS_{error} = \sum_i \sum_j \left(X_{ij} - \bar{X}_{i.} \right)^2$$

Variation due to
Chance causes
(Within SS)

- Total SS is calculated using squared deviations of each value from overall mean.
- Between SS is calculated using squared deviation of each group mean from overall mean.
- Within Group SS can be obtained by subtracting Between SS from Total SS

Case Study - 2

Background

A large company is assessing the difference in the 'Satisfaction Index' of employees in it's Finance, Marketing and Client-Servicing departments.

Objective

To test whether the **mean satisfaction indices** for employees in three departments (CS, Marketing, Finance) are equal.

Sample Size

Sample size: 37
Variables: satindex, dept

Data Snapshot

One way anova

Variables

Observations

satindex	dept
75	FINANCE
56	FINANCE
72	FINANCE
59	FINANCE
66	FINANCE
58	FINANCE
58	MARKETING
63	MARKETING
51	MARKETING
64	MARKETING
55	MARKETING
72	CS
69	CS

Columns	Description	Type	Measurement	Possible values
satindex	Satisfaction Index	Numeric		Positive Values
dept	Department	Character	MARKETING, CS, FINANCE	3

One Way ANOVA

Testing equality of means in one factor with more than two levels.

Objective	**To test whether the mean satisfaction indices for employees in three departments (CS, Marketing, Finance) are equal.**

Null Hypothesis (H_0): Mean satisfaction index for 3 departments are equal i.e. $\mu 1 = \mu 2 = \mu 3$

Alternate Hypothesis (H_1): Mean satisfaction index for 3 departments are not equal

Test Statistic	**The test statistic is denoted as F and is based on F distribution.**
Decision Criteria	Reject the null hypothesis if $p\text{-value} < 0.05$

Calculation

Overall Mean	65.59	n=37
Mean for Finance	64.42	n1=12
Mean for Marketing	63.25	n2=12
Mean for CS	68.85	n3=13

Total SS = (75-65.59)^2+(56-65.59)^2+…………+(65-65.59)^2+(76-65.59)^2
= **1840.92**

Between Groups SS = 12*(64.42-65.59)^2+12*(63.25-65.59)^2+13*(68.85-65.59)^2
= **220.0599**

Within Groups SS = Total SS – Between SS = **1620.86**

One Way ANOVA table

Sources of variation	Degrees of freedom (df)	Sum of Squares (SS)	Mean Sum of Squares (MS=SS/df)	F-Value
Between groups	K-1=3-1 =2	SSA=220.0599	MSA=110.03	F=2.3080
Within groups (error)	n-k=37-3 =34	SSE=1620.86	MSE=47.6724	
TOTAL	n-1=37-1 =36	TSS=1840.92		

One Way ANOVA in R

```
# Import data
data<-read.csv("One way anova.csv",header=TRUE)
```

```
# ANOVA table
anovatable<-aov(formula=satindex~dept, data=data)
summary(anovatable)
```

- *'aov' is the R function for ANOVA .*
- *formula specifies 'satindex' as analysis (dependent) variable and 'dept' as factor (independent) variable.*
- *anovatable is user defined object name created to store output.*
- *summary function displays the ANOVA table output.*

```
# Output:
          Df Sum Sq Mean Sq F value Pr(>F)
dept       2  220.1  110.03   2.308  0.115
Residuals 34 1620.9   47.67
```

Interpretation :

- *Since p-value is >0.05, do not reject H0. There is no significant difference in satisfaction index among 3 different departments.*

Quick Recap

ANOVA	• Analysis of variance (ANOVA) is a collection of statistical models used to analyze the difference among more than two group means developed by statistician and evolutionary biologist Ronald Fisher.
Partitioning the variance	• The total variance in any variable is due to a number of causes which may be classified as "assignable causes (which can be detected and measured)" and "chance causes (which is beyond human control and cannot be traced separately)".
One Way ANOVA	• Comparing several means of different levels of one factor. $H_0 : \mu_1 = \mu_2 = \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots = \mu_K = \mu$

Statistical Inference

Test for equality of variances

F-test for equality of variances

- The F test is used to test the equality of two population variances.
- Testing equality of variances is a prerequisite for many statistical tests (eg the Independent sample t-test).
- Under H0 $\sigma_1^2 = \sigma_2^2$
  Where σ_1^2 and σ_2^2 are the first and second population variances, respectively.

Assumptions for F-test

- The assumptions for the F-test are listed below:
  - Random sampling from a defined population
    (employees are selected at random from the company)

  - Population of the testing variable is normally distributed
    (The time taken to complete the MIS report should be normally distributed).

- Note: Generally the F test is used to the validate assumption of equal variance while performing the t test for equality of means. The parent population is assumed to follow a normal distribution.

Case Study - 1

Background

The company is analysing the time to complete an MIS report between two groups of employees.

Group I: Experience (0-1 years)

Group II: Experience(1-2 years)

Objective

To test the equality of the variances in time taken to complete MIS in two groups of employees.

Sample Size

Sample size: 14

Variables: time_g1, time_g2

Data Snapshot

F test for 2 variances

Variables

Observations

time_g1	time_g2
85	83
95	85
105	96
85	94
90	102

Columns	Description	Type	Measurement	Possible values
time_g1	Time to complete MIS report by group1	Numeric	Hours	Positive Values
time_g2	Time to complete MIS report by group2	Numeric	Hours	Positive Values

F-test

Testing equality of variances in two samples.

Objective	To test the **equality** of the variances in time taken to complete an MIS report in two groups of employees.

Null Hypothesis (H_0): Variances of time are equal in two groups. i.e. $\sigma_1^2 = \sigma_2^2$.
Alternate Hypothesis (H_1): Alternative Hypothesis H1: $\sigma_1^2 \neq \sigma_2^2$

Test Statistic	$F = \frac{s_1^2}{s_2^2} \sim F_{\alpha, n_1-1, n_2-1}$ **Where s_1^2 is the sample variance of first sample and, s_2^2 is the sample variance of second sample. n_1 and n_2 are sample sizes of the first and second sample respectively.**
Decision Criteria	Reject the null hypothesis if p-value < 0.05

Computation

	Group I	Group II
Sample Size	$n_1=12$	$n_2=14$
Mean	$\bar{x}_1 = 93.5833$	
Sample Variance	$s_1^2 = 41.9015$	$s_2^2 = 27.1484$
F Value	$F = \frac{s_1^2}{s_2^2}$	1.5434

F-test in R

```
# Import data
data<-read.csv("F test for 2 variances.csv",header=TRUE)

# Variance test
var.test(data$time_g1,data$time_g2,alternative = "two.sided")
```

- *time_g1,time_g2 are the variables under study.*
- *alternative="two.sided" , since under H1,variances are not equal.*

```
# Output :
```

```
        F test to compare two variances

data:  data$time_g1 and data$time_g2
F = 1.5434, num df = 11, denom df = 13, p-value = 0.4524
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4826988 5.2348866
sample estimates:
ratio of variances
          1.543428
```

Interpretation :

- *Since the p-value is >0.05, do not reject H0. There is no significant difference in variances of the two groups.*
- *Also, 95 percent confidence interval of ratio of variance contains 1,which means variances are same.*

Statistical Inference

Parametric Tests - II

Contents

Independent samples t-test

- The independent-samples t-test compares the means of two independent groups on the same continuous variable.

- The following hypotheses are tested in an independent samples t test
  - H_0: Two population means are equal
  - H_1: Two population means are not equal

Data Snapshot

INDEPENDENT SAMPLES t TEST

Variables

Observations

time_g1	time_g2
85	83
95	85
105	96
85	94
90	102

Columns	Description	Type	Measurement	Possible values
time_g1	Time to complete MIS report by group1	Numeric	Hours	Positive Values
time_g2	Time to complete MIS report by group2	Numeric	Hours	Positive Values

Case Study - 1

Background

The company is assessing the difference in time to complete an MIS report between two groups of employees :
Group I: Experience(0-1 years)
Group II: Experience(1-2 years)

Objective

To test whether the average time taken to complete the MIS by both the groups is same.

Sample Size

Sample size: 14
Variables: time_g1, time_g2

Assumptions for independent samples t-test

- The assumptions for the independent samples t-test are listed below :
  - The samples drawn are random samples.
    (Employees are selected at random from the company)
  - The populations from which samples are drawn have equal & unknown variances.
    (F-test is used to validate this assumption which will be covered in next presentation)
  - The populations follow normal distribution.
    (Time taken to complete MIS report should be normally distributed for both groups.
  - A normality assumption can be validated using a method explained earlier.

Independent sample t-test

Testing whether the means of the two groups are equal.

Null Hypothesis (H_0): $\mu 1 = \mu 2$

Alternate Hypothesis (H_1): $\mu 1 \neq \mu 2$

$\mu 1$= average time taken by group1 to complete MIS
$\mu 2$=average time taken by group2 to complete MIS .

Objective	To test the average time taken to complete the MIS by both the groups is same.

Test Statistic	$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$
Decision Criteria	Reject the null hypothesis if p-value < 0.05

Computation

	Group I	Group II
Sample Size	n1=12	n2=14
Mean		
Variance	S_1^2=41.9015	S_2^2=27.1483
Pooled Variance	Sp^2=33.9102	
Difference		
t	$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ =0.22345	

Independent samples t-test in R

```
# Import data
```

```
data<-read.csv("INDEPENDENT SAMPLES t TEST.csv", header=TRUE)
```

```
# t-test for independent samples
```

```
t.test(data$time_g1,data$time_g2, alternative="two.sided",
       var.equal=TRUE)
```

- *data$time_g1 and data$time_g2 are the variables to be compared.*
- *alternative="two.sided" since H1 is* $\mu 1 \neq \mu 2$
- *var.equal=TRUE assumes for equality of variance of two groups.*

(there is a test which validates this assumption which is explained in subsequent slides.)

* Before performing the t test, a normality test is done to ensure time variable is normally distributed in both the groups.

Independent samples t-test in R

```
# Output:
```

```
        Two Sample t-test

data:  data$time_g1 and data$time_g2
t = 0.22346, df = 24, p-value = 0.8251
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.216185  5.239994
sample estimates:
mean of x mean of y
 93.58333  93.07143
```

Interpretation :

- *Since p-value is >0.05, do not reject H0. There is no significant difference in average time taken to complete the MIS between both groups of employees.*
- *95% CI contains value 0 which is value under H0.($\mu1 = \mu2 \rightarrow \mu1-\mu2=0$). Hence, do not reject H0.*

Independent samples t-test when variances are not equal

- Welch's t test is used to test the equality of two means if the variances of two groups can not be assumed to be equal.
- Welch's t-test defines the statistic t by the following formula:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

- The denominator is not based on a pooled variance estimate.
- If two variances are not equal, the t test syntax in R below is used:

```
data<-read.csv("INDEPENDENT SAMPLES t TEST.csv", header=TRUE)

t.test(data$time_g1,data$time_g2,alternative="two.sided",
var.equal=FALSE)
```

Paired samples t-test

- The paired sample t-test is used to determine whether the mean difference between two sets of observations is zero ,where each subject or entity is measured twice resulting in a pair of observations.

- Commonly used when observations are recorded ‘before’ and ‘after’ training and the objective is to test whether the training is effective.

Case Study - 2

Background

The company organized a training program to improve efficiency. The time taken to complete an MIS report before and after training is recorded for 15 employees.

Objective

To test whether the average time taken to complete the MIS before and after training is not different.

Sample Size

Sample size: 15
Variables: time_before, time_after

Data Snapshot

PAIRED t TEST

Variables

Observations

time_before	time_after
85	74
95	91
92	80
102	91
95	88
88	81
90	78

Columns	Description	Type	Measurement	Possible values
time_before	Time to complete MIS report before training	Numeric	Hours	Positive values
time_after	Time to complete MIS report after training	Numeric	Hours	Positive values

Assumptions for paired sample t-test

- The assumptions of the paired-sample t-test are listed below:

  - Random sampling from a defined population
    (employees are selected at random from the company)

  - Population of the testing variable is normally distributed
    (Difference time taken to complete MIS report should be normally distributed).

- A Normality test can be performed by any of the methods explained earlier.

- The validity of the test is not seriously affected by moderate deviations from the 'Normality' assumption.

Paired sample t-test

Testing whether means of two dependent groups are equal.

Objective	To test the average time taken to complete MIS before and after training is not different.

Null Hypothesis (H_0): There is no difference in average time before and after the training. i.e. D=0

Alternate Hypothesis (H_1):Average time is less after the training. (Training is effective.) D>0

D= μBefore – μAfter

Test Statistic	$t = \frac{\bar{d}}{s_d/\sqrt{n}}$ **Where $\bar{d}$ is the sample mean of the difference before-after , s_d is the sample standard deviation of the difference, n is the sample size of difference. The quantity t follows a distribution called as 't distribution' with n-1 degrees of freedom.**
Decision Criteria	Reject the null hypothesis if p-value < 0.05

Computation

	Notation	Value
Sample Size	n	12
Mean difference (before-after)	$\bar{d}$	8.3333
Standard Deviation	s_d	3.9219
t	$t = \frac{\bar{d}}{s_d/\sqrt{n}}$	8.2295

Paired sample t-test in R

```
# Import data
data<-read.csv("PAIRED t TEST.csv",header=TRUE)

# t-test for paired samples
t.test(data$time_before,data$time_after,
       alternative="greater", paired=TRUE)
```

- *data$time_before and data$time_after are the variables under study.*
- *alternative="greater" ,Since under H1, difference of before –after will be large.*

* Before performing a t test, a normality test is carried out to ensure difference variable is normally distributed.

Paired sample t-test in R

```
# Output:
```

```
        Paired t-test

data:  data$time_before and data$time_after
t = 8.2295, df = 14, p-value = 4.919e-07
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 6.549798      Inf
sample estimates:
mean of the differences
               8.333333
```

Interpretation :

- *Since p-value is <0.05, reject H0. Average time taken to complete the MIS report after the training is less. Hence, training is effective.*
- *95% C.I does not contain value of D=0 (under H0),reject H0.*

t-test for Correlation

- The Correlation coefficient summarizes the strength of a linear relationship between two variables.
- A t-test is used to check if there is significant correlation between two variables.
- Sample correlation coefficient (r) is calculated using bivariate data.
- Null hypothesis of this test is
  H0: there is no correlation between 2 variables under study ($\rho=0$)

Case Study - 3

Background

A company with 25 employees has calculated a job proficiency score & an aptitude test score for its employees

Objective

To test if there is significant correlation between the job proficiency and aptitude test scores.

Sample Size

Sample size: 25
Variables: Empcode, Aptitude, Job_prof

Data Snapshot

Correlation test

Variables

Observations

Empcode	aptitude	job_prof
E101	86	88
E102	62	80
E103	110	96
E104	101	76
E105	100	80
E106	78	73
E107	120	58
E108	105	116
E109	112	104

Columns	Description	Type	Measurement	Possible values
Empcode	Employee code	Numeric	-	
Aptitude	Score of aptitude test	Numeric	-	Positive values
Job_prof	Job proficiency score	Numeric	-	Positive values

Correlation t-test

Testing for correlation coefficient value.

Objective	To test whether there exists significant correlation between job proficiency and aptitude score.

Null Hypothesis (H_0): There is no significant correlation between Job proficiency and Aptitude test ($\rho=0$).
Alternate Hypothesis (H_1): There is correlation between Job proficiency and Aptitude test ($\rho\neq0$)

Test Statistic	$t = \frac{r\sqrt{(n-2)}}{\sqrt{1-r^2}}$ where r is the sample correlation coefficient, the sample size. The quantity t follows a distribution called as 't distribution' with n-2 degrees of freedom.
Decision Criteria	Reject the null hypothesis if p-value < 0.05

Computation

	Notation	Value
Sample Size	n	25
Sample correlation coefficient	r	0.514411
t	$t = \frac{r\sqrt{(n-2)}}{\sqrt{1-r^2}}$	2.8769

Correlation **t-test in R**

```
# Import data
data<-read.csv("Correlation test.csv", header=TRUE)

# t-test for correlation
cor.test(data$aptitude, data$job_prof, alternative="two.sided",
         method="pearson")
```

- *data$aptitude and data$job_prof are the variables under study.*
- *alternative="two.sided". Since under H1, ($\rho \neq 0$) .*

Correlation t-test in R

```
# Output:
```

```
        Pearson's product-moment correlation

data:  data$aptitude and data$job_prof
t = 2.8769, df = 23, p-value = 0.008517
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1497097 0.7558981
sample estimates:
      cor
0.5144107
```

Interpretation :

- *Since p-value is <0.05, reject H0. There is correlation between aptitude test and job proficiency.*
- *95% C.I does not contain value ρ=0 (under H0), reject H0.*

Quick Recap

Independent sample t test	• It compares the means of two independent groups on the same continuous variable. • H0: $\mu1 = \mu2$
Paired sample t test	• Used to determine whether the mean difference between two sets of observations is zero ,where each subject or entity is measured twice resulting in pair of observations. • H0: $\mu1-\mu2=d=0$
t test for correlation	• Used to check if there is significant correlation between two variables. • H0: $\rho=0$

Statistical Inference

Parametric Tests - I

Contents

Normality test

- An assessment of the normality of data is a prerequisite for many statistical tests because normal data is an underlying assumption in parametric testing.

- Normality can be assessed using two approaches: graphical and numerical.
  - Graphical approach
    - Box-Whisker plot (used to assess symmetry rather than normality.)
    - Quantile-Quantile plot (Q-Q plot).

  - Numerical (Statistical) approach
    - Shapiro-Wilk test (used generally for **small samples**)
    - Kolmogorov-Smirnov test (used generally for **large samples**)

* A box-whisker plot is used to asses symmetry rather than normality. Hence, only the Q-Q plot method is explained.

Case Study - 1

Background

Data has 2 variables recorded for 80 guests in a large hotel.
Customer Satisfaction Index (csi) & Total Bill Amount in thousand Rs. (billamt)

Objective

To check the normality of the data

Sample Size

Sample size: 80
Variables: id, csi, billamt

Data Snapshot

Normality Testing Data

Variables

Observations

id	csi	billamt
1	38.35	34.85
2	47.02	10.99
3	36.96	24.73
4	43.07	7.9
5	38.77	9.38
6	63.04	9.49
7	43.17	19.58
8	35.14	6.15
9	38.33	13.29
10	38.7	9.62
11	31.44	8.51

Column	Description	Type	Measurement	Possible Values
id	Customer ID	Numeric		
csi	Customer Satisfaction Index	Numeric		Positive value
billamt	Total Bill Amount in thousand euros.	Numeric	Rs.	Positive value

Quantile-Quantile plot

- Very powerful graphical method of assessing normality.
- Quantiles are calculated using sample data and plotted against expected quantiles under normal distributions.
- If the normality assumption is valid then, a high correlation is expected between the sample quantiles and the expected (theoretical quantiles under normal distribution) quantiles.
- The Y axis plots the actual quantile values based on the sample. The X axis plots theoretical values.
- If the data is truly sampled from a normal distribution, the QQ plot will be linear.

Q-Q plot using R

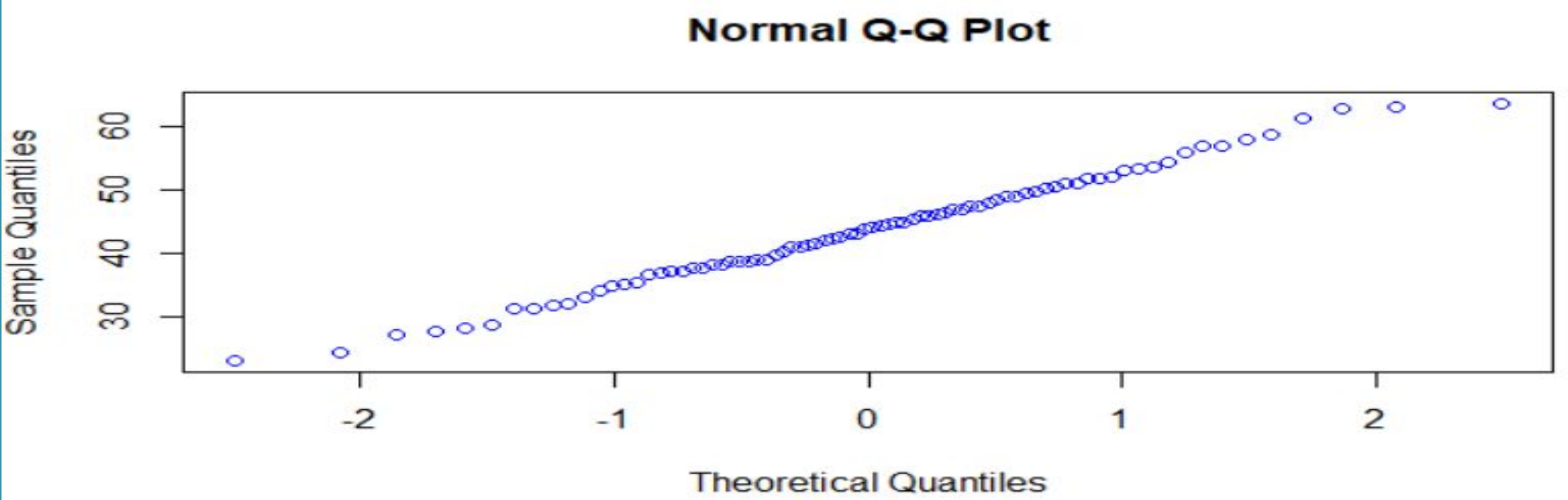
```
# Import data
data<-read.csv("Normality Testing Data.csv", header=TRUE)

# Q-Q plot for the variable csi
qqnorm(data$csi,col="blue")
```

- *data$csi is the variable for which normality is to be checked.*
- *Col=blue specifies the line color on graph.*

Q-Q plot using R

```
# Output:
```



Interpretation :

- *Q-Q plot is Linear. Distribution of 'csi' can be assumed to be normal.*

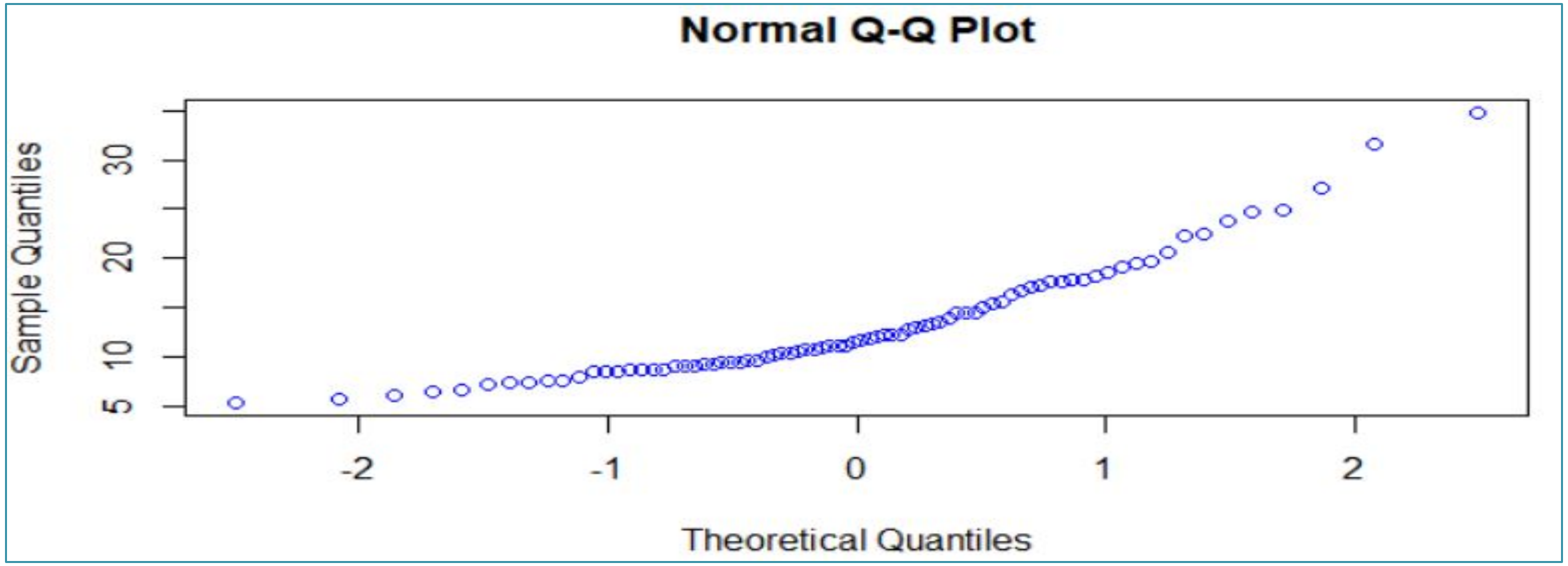
Q-Q plot using R

```
# Q-Q plot for the variable billamt
qqnorm(data$billamt,col="blue")
```

- *data$billamt is the variable for which normality is to be checked.*
- *Col=blue specifies the line color on graph.*

Q-Q plot using R

```
# Output:
```



Interpretation :

- *Q-Q plot is deviated from linearity. Distribution of 'billamt' appears to be non-normal.*

Shapiro-Wilk test

The Shapiro-Wilk test is widely used statistical test for assessing **normality**.

Objective	To test the **normality** of the data.

Null Hypothesis (H_0): Sample is drawn from Normal Population

Alternate Hypothesis (H_1): Sample is drawn from Non-Normal Population

The test is performed for the variables, 'csi' and 'billamt' separately.

Test Statistic	It correlates sample ordered values with expected Normal scores. (the actual calculation is very complex so we will avoid details)
Decision Criteria	Reject the null hypothesis if p-value < 0.05

* Here we continue to use previous data of CSI and Bill Amount for Shapiro-Wilk test.

Shapiro-Wilk test in R

```
# Shapiro Wilk test for the variable csi
```

```
shapiro.test(data$csi)
```

- *data$csi is the variable for which normality is to be checked.*

```
# Output:
```

```
        Shapiro-Wilk normality test

data:  data$csi
W = 0.99196, p-value = 0.9038
```

Interpretation :

- *Since p-value is >0.05,do not reject H0. Distribution of 'csi' can be assumed to be normal.*

Shapiro-Wilks test in R

```
# Shapiro Wilks test for the variable billamt
```

```r
shapiro.test(data$billamt)
```

- *data$billamt is the variable for which normality is to be checked.*

```
# Output:
        Shapiro-Wilk normality test

data:  data$billamt
W = 0.89031, p-value = 4.858e-06
```

Interpretation :

- *Since p-value is <0.05, reject H0. Distribution of 'billamt' appears to be non-normal.*

Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test is another widely used statistical test for assessing Normality.

Objective	To test the **normality** of the data.

Null Hypothesis (H_0): Sample is drawn from Normal Population

Alternate Hypothesis (H_1): Sample is drawn from Non-Normal Population

The test is performed for the variables, ‘csi’ and ‘billamt’ separately.

Test Statistic	Kolmogorov-Smirnov Test: It compares empirical (sample) cumulative distribution function (CDF) with Normal distribution CDF. The test statistic is the maximum difference between CDF's.
Decision Criteria	Reject the null hypothesis if p-value < 0.05

* Here we continue to use previous data of CSI and Bill Amount for Shapiro-Wilk test.

Kolmogorov-Smirnov test in R

```
# Install and use package ‘nortest’
```

```
install.packages("nortest")

library(nortest)
```

- *Package nortest contains the Kolmogorov smirnov test.*

```
# Kolmogorov Smirnov test
```

```
lillie.test(data$csi)
```

- *data$csi is the variable for which normality is to be checked.*

Kolmogorov-Smirnov test in R

Output:

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  data$csi
D = 0.042387, p-value = 0.9764
```

Interpretation :

- *Since p-value is >0.05,do not reject H0. Distribution of 'csi' can be assumed to be normal.*

Kolmogorov-Smirnov test in R

```
# Kolmogorov Smirnov test for the variable billamt
```

```
lillie.test(data$billamt)
```

- *data$billamt is the variable for which normality is to be checked.*

```
# Output:
```

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  data$billamt
D = 0.14244, p-value = 0.0003753
```

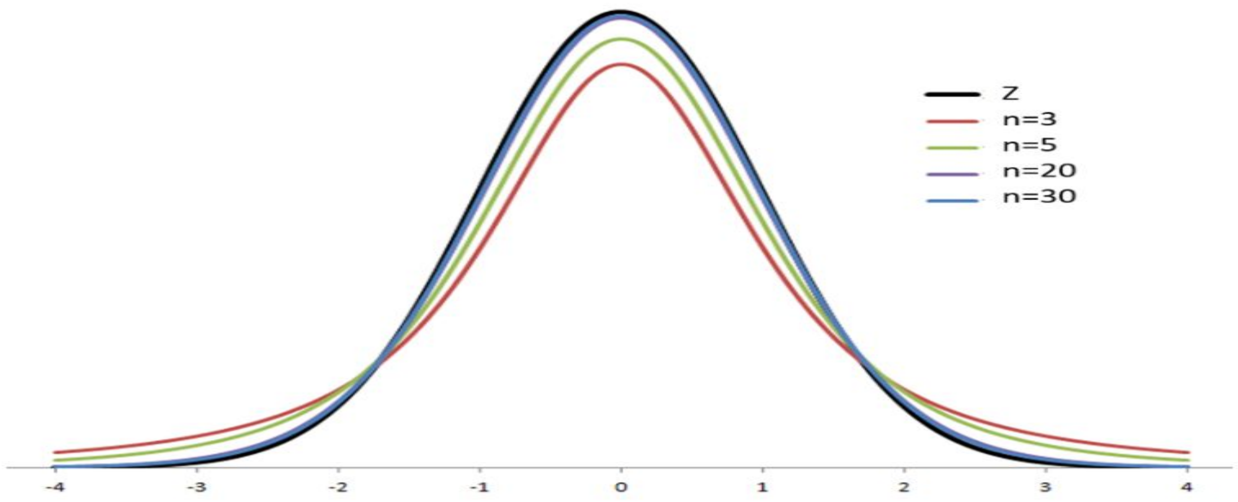
Interpretation :

- *Since p-value is <0.05, reject H0. Distribution of ‘billamt’ appears to be non-normal.*

t-distribution

- The t distribution is symmetric and its overall shape resembles the bell shape of a normally distributed variable with mean 0 and variance 1, except that it is a bit lower and wider.

- As the sample size increases so as the number of degrees of freedom grows, the t-distribution approaches the normal distribution with mean 0 and variance 1.



- In the above graph, z is a normal distribution with mean 0 and variance 1.

A note on Degrees of Freedom (DF)

- Degrees of freedom (df) is defined as the number of independent terms.
- "Sum of the squared deviations about mean of n values" has n-1 degrees of freedom. Knowing n-1 values, we can find last value since sum of deviations about mean is always zero.
- Sampling distributions like t, F and chi square have shapes based on degrees of freedom.
- Example , Give 5 numbers such that sum is 20. You can use 4 numbers freely but the fifth number should be such that sum is 20. Here df =4

One sample t-test

- The one sample t test is used to test a hypothesis about a single population mean.

- We use the one-sample t-test when we collect data on a single sample drawn from a defined population.

- For this design, we have one group of subjects, collect data on these subjects and compare a sample statistic to the hypothesized value of a population parameter.

- Subjects in the study can be patients, customers, retail stores etc.

Case Study - 2

Background

A large company is concerned about time taken by employees to complete the weekly MIS report.

Objective

To check if the average time taken to complete the MIS report is more than 90 minutes

Sample Size

Sample size: 12
Variables: Time

Data Snapshot

ONE SAMPLE t TEST

Variables

Observations

Time
85
95
105
85
90
97
104
95
88
90
94
95

Columns	Description	Type	Measurement	Possible values
Time	Time taken to complete MIS	Numeric	Minutes	Positive Values

Assumptions for one sample t-test

The assumptions of the one-sample t-test are listed below:

- Random sampling from a defined population
  (employees are selected at random from the company)

- The population is normally distributed
  (Time taken to complete MIS report should be normally distributed).

- The variable under study should be continuous.

A normality test can be performed by any of the methods explained earlier.

The validity of the test is not seriously affected by moderate deviations from 'normality' assumption.

One sample t-test

Testing whether mean is equal to a test value.

Objective	To test the average time taken to complete MIS is more than 90 minutes

Null Hypothesis (H_0): $\mu = 90$

Alternate Hypothesis (H_1): $\mu > 90$

Test Statistic	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
Decision Criteria	Reject the null hypothesis if p-value < 0.05

Computation

	Notation	Value
Sample Size	n	12
Mean		93.5833
Standard Deviation	S	6.4731
Standard Error	s/√n	1.8686
Difference		93.5833-90=3.5833
t	$\frac{\bar{x} - \mu_0}{S.E}$	1.9176

One sample t-test in R

```
# Import data
data<-read.csv("ONE SAMPLE t TEST.csv",header=TRUE)

# t-test for one sample
t.test(data$time, alternative="greater", mu=90)
```

- *data$time is the variable under study.*
- *alternative="greater" ,Since under H1, value is tested for greater than 90.*
- *mu=90 is the value to be tested.*

* Before performing t test, normality test is done to ensure time variable is normally distributed.

One sample t-test in R

```
# Output:
```

```
        One Sample t-test

data:  data$time
t = 1.9176, df = 11, p-value = 0.04074
alternative hypothesis: true mean is greater than 90
95 percent confidence interval:
 90.22748      Inf
sample estimates:
mean of x
 93.58333
```

Interpretation :

- *Since the p-value is <0.05, reject H0. The average time taken to complete the MIS report is more than 90 minutes ’ .*

Quick Recap

Normality Test

- Normal data is an underlying assumption in parametric testing.
- Two approaches to test normality:
- Graphical (Box-Whisker plot,Quantile-Quantile plot
- Statistical (Shapiro-Wilks test,Kolmogorov-Smirnov test

One sample t test

- Used to test the hypothesis about a single population mean.
- H0: μ = μ0

Statistical Inference

An Introduction

Contents

Basic Terms as Prerequisite

- **Variable** (under study) – What you measure (ex. monthly salary of employees)
- **Population**- Set of all units in the study (all employees in the organization)
- **Sample**- Subset of units selected from population (ex. monthly salary of few selected employees in the organization)
- **Distribution**-How values of variable are distributed in the population (ex. normal distribution)
- **Factor**- Defines subgroups in the study.(ex. Gender, where gender wise salary distribution can be studied.)
- **Descriptive Statistics**- mean, median, standard deviation etc of the variable under study.. (ex. Average salary)

What is Statistical Inference ?

- Statistical inference is the process of drawing conclusion about unknown population properties, using a sample drawn from the population.

- These unknown population properties can be:
  - Mean
  - Proportion
  - Variance etc.

- Such unknown population properties are called as ‘Parameters’.

What is Statistical Inference ?

Statistical Inference

Estimation

Point estimation

- Summarize the sample by a single value as an estimate of the population parameter.

- Ex. Average salary of junior data scientists is. 55,000 euros.

Interval estimation

- A range of values within which, we believe,
- the true population parameter lies with high probability.

- Ex. Average salary of junior data scientists is in the range of (52,000,55,000)
- With 95% confidence level.

Testing of Hypothesis

- To decide whether a statement regarding population parameter is true or false, based on sample data.

- Ex. Claim: Average salary of junior data scientists is greater than.50,000 euros annually.

Parameter, Estimator, Estimate

- **Parameter**: Unknown property or characteristic of population
  - (population mean (μ), variance (σ^2), proportion (P))

- **Estimator**: A rule or function based on sample observations which is used to estimate the parameter
  - (sample mean, sample variance, sample proportion)

- **Estimate:** A particular value computed by substituting the sample observations into an Estimator.

Parameter, Estimator, Estimate

- Research Question: What is the average salary of junior data scientists in Europe?
  - Average salary of junior data scientists in Europe is Population **Parameter.**
  - Sample of 250 junior data scientists is observed and Sample mean is computed.
  - Sample mean is used as **Estimator** of Population Mean.
  - Sample mean "55,000" which is calculated from sample of 250 is the **Estimate.**

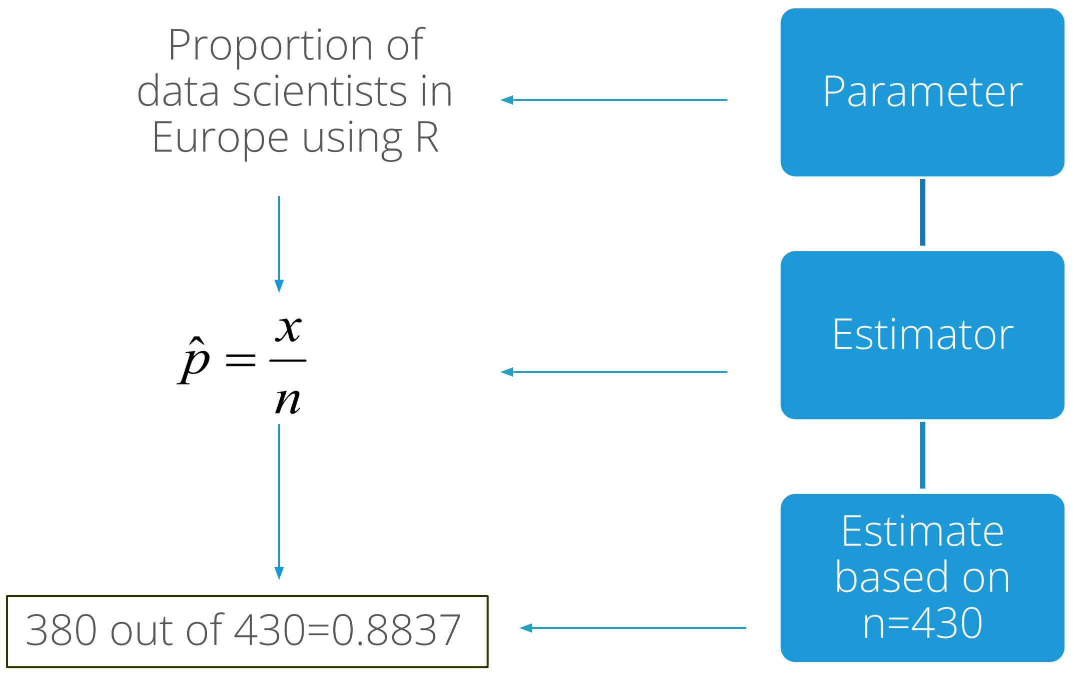
Average salary of junior data scientists (μ) ← Unknown population Parameter

$\hat{\mu} = \overline{x}$ ← Estimator

euro 55,000 ← Estimate based on n=250

Parameter, Estimator, Estimate

- Research Question: What is the proportion of data scientists in Europe who use R for data analysis?
  - Proportion of data scientists in Europe who use R for data analysis is population parameter.
  - Sample of 430 data scientists observed and proportion (or percentage) is calculated.
  - Sample proportion is used as an estimator of population proportion.
  - 380 out of 430 which is calculated from sample is **Estimate.**



Point Estimation vs. Interval Estimation

- In both the previous examples, (estimation of average salary of junior data scientists and proportion of data scientists using R) estimator is a single value estimating unknown population parameter.
- A confidence interval gives an estimated range of values which is likely to include an unknown population parameter with some probability, the estimated range being calculated from a given set of sample data.
- Generally, 95% or 90% Confidence Intervals are used.
- 95% confidence interval is a range estimate within which the true value of the parameter lies with probability 0.95.

Sampling distribution and Sampling error

- Research Question: What is the average salary of junior data scientists in Europe?

  - 50 samples, each of size 250 junior data scientists are observed and sample mean for each of theses 50 samples are computed. Here, sample mean will vary based on sample values.

- A probability distribution of all these means of the sample is called the **sampling distribution** of mean.

- **Standard error** is standard deviation of the these mean values.

Hypothesis Testing

- **Hypothesis**: An assertion about the distribution / parameter of the distribution of one or more random variables.
- **Null Hypothesis (H_0)**: An assertion which is generally believed to be true until researcher rejects it with evidence.
- **Alternative Hypothesis (H_1)**: A researcher's claim which contradicts null hypothesis.
- In simple words, testing of hypothesis is to decide whether a statement regarding population parameter is true or false, based on sample data.
- **Test Statistic**: The statistic on which decision rule of rejection of null hypothesis is defined.
- **Critical region or Rejection region**: the region, in which, if the value of test statistic falls, the null hypothesis is rejected.

Hypothesis Testing : Example

Objective	A consumer protection agency wants to test a Paint Manufacturer's claim, that average drying time of their new paint is less than 20 minutes.

- Sample: n=36 boards were painted from 36 different cans and the drying time was observed.
- Estimator of mean drying time is sample mean $\hat{\mu} = \bar{x}$

Null Hypothesis (H_0): $\mu = 20$

Alternate Hypothesis (H_1): $\mu < 20$

Test Statistic	In this case the test statistic is based on $\bar{x}$
Decision Criteria	Reject null hypothesis if test statistic based on sample mean is less than critical value.

Two types of error

- While testing the hypothesis using any decision rule, one of the following scenario might occur.

Decision	Reality	
	Ho is true	Ho is false
Reject Ho	Type I error	Correct
Do Not Reject Ho	Correct	Type II error

- **For example**, in legal system,
  H_0: person is not guilty H_1: person is guilty

Decision	Reality	
	Not Guilty	Guilty
Guilty	Type I Error -- Innocent person goes to jail	Correct
Not Guilty	Correct	Type II error Guilty person is set free

Two Types of error

- **Level of significance (LOS):** Probability of Type I error is called as 'Level of Significance (α)'
  generally set as 5% (los=0.05) and null hypothesis is rejected if observed risk(p value) is less
  than 0.05

- **p-value:** is the smallest level of significance that would lead to rejection of the null hypothesis (generally if $p < 0.05$, we reject the null hypothesis).

- **α =** Probability [Type I Error] = Probability [Reject H_0 | H_0 is True]

- **β =** Probability [Type II Error] = Probability [Do not reject H_0 | H_0 is not True]

- **Power of the test** is given by: $(1 - \beta)$

One tailed and two tailed tests

- Hypothesis test where the alternative hypothesis is one-tailed (right-tailed or left-tailed), is called a **one-tailed test**.

  $H0: \mu = \mu0$
  $H1: \mu > \mu0$ (Right-tailed) or $H1: \mu < \mu0$ (left-tailed)

- Hypothesis test where the alternative hypothesis is two-tailed is called **two-tailed test**.

  $H0: \mu = \mu0$
  $H1: \mu \neq \mu0$

Quick Recap

Statistical Inference	• It is the process of drawing conclusion about unknown population properties, using a sample drawn from the population.
Point Estimation	• Summarize the sample by a single value as an estimate of the population parameter.
Interval Estimation	• A range of values within which, we believe, the true population parameter lies with high probability.
Testing of Hypothesis	• To decide whether a statement regarding population parameter is true or false
Type I error	• α = Probability [Type I Error] = Probability [Reject H0 | H0 is True]
Type II error	• β = Probability [Type II Error] = Probability [Do not reject H0 | H0 is not True] • Power of the test is given by: $(1 – \beta)$