7. Predictive Modeling

7.1 Hypothesis testing

- **Example 1: The food label on a cookie bag states that there is at most 2 grams of saturated fat in a single cookie.**
- **In a sample of 35 cookies, it is found that the mean amount of saturated fat per cookie is 2.1 grams.**
- **Assume that the population standard deviation is 0.25 grams.**
- **At 0.05 significance level, can we reject the claim on the food label?**

7.1 Hypothesis testing - continued

- **The elements about a population that are more often tested are:**

1. **The population mean (e.g. average pizza delivery time is 30 minutes)**
2. **The population proportion (e.g. 80% of the voters support the view of the politician)**
3. **The difference in two proportions means or proportions (e.g. is it true that a greater proportion of males experience the side effects of a drug than females do)**

7.1 Hypothesis testing- continued

- A hypothesis test is a technique for using data to validate or invalidate a claim about a population.
- For example, a politician may claim that 80% of the people in her state agree with her views on a certain issue.
- A Pizza restaurant may claim that they deliver pizzas in 30 minutes or less.
- Medical researchers use hypothesis tests all the time to test whether or not a certain drug is effective, to compare a new drug to an existing drug in terms of its side effects.

7.1 Hypothesis testing- continued

- Every hypothesis test contains a set of two opposing statements, or hypotheses, about a population parameter.
- The first hypothesis is called the null hypothesis, denoted by H_0, The null hypothesis always states that the population parameter is equal to the claim value.
- The opposite of the null hypothesis is the alternative hypothesis, denoted by H_a or H_1.

Ref: http://www.dummies.com/how-to/content/how-to-set-up-a-hypothesis-test-null-versus-altern.html

7.1 Hypothesis testing - continued

Steps in hypothesis testing

Step 1 Frame null and alternate hypothesis

Step 2 Compute the test statistic, z

- **The z statistic will compare the observed sample mean to an expected population mean.**
- **To obtain a z-score for the entire population with parameters μ and σ, use** $z = \frac{(x - \mu)}{\sigma}$
- **If population mean and standard deviation are estimated from a sample parameters (x-bar and s)**

$$z = \frac{(x - \bar{x})}{s}$$

7.1 Hypothesis testing - continued

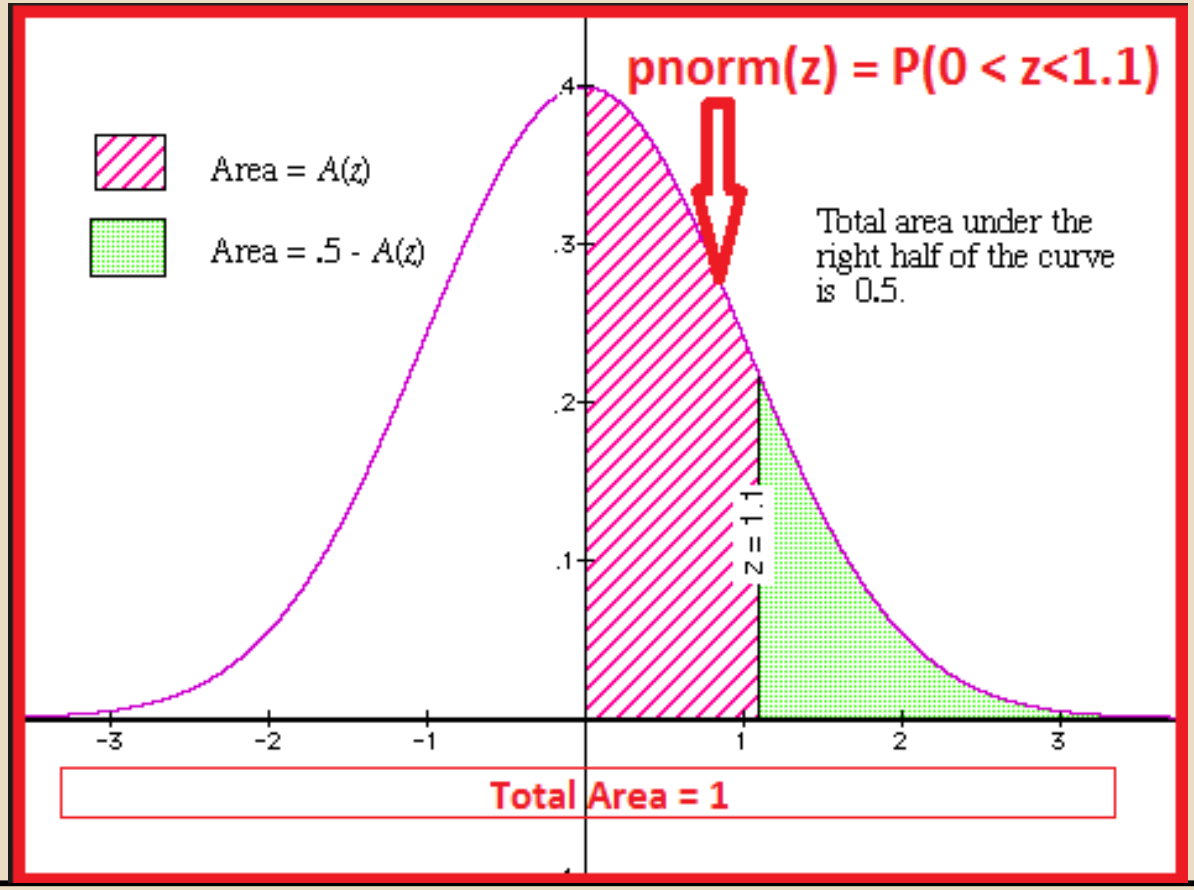
Steps in hypothesis testing – continued

Step 3 Compute p value and conclude

- **The test statistic is converted to a conditional probability, p-value. This answers the question, "if the null hypothesis were true, what is the probability of observing the current data or data is more extreme?"**
- **When p-value is =< 0.05, the observed difference is significant.**
- **When p-value is =< 0.01, the observed difference is highly significant.**

7.1 Hypothesis testing - continued

Hypothesis testing – continued



7.1 Hypothesis testing - continued

Solution:

Here the null hypothesis (H_0) is $\mu \leq 2$

N = 35; sample mean = 2.1 gm;

Population std. deviation = 2.1 gm;

Population mean = 2.0 gm

We use the pnorm function to compute the upper tail of the test statistic.

```
> pnorm(2.0,mean=2.1,sd=0.25/sqrt(35))
[1] 0.008980239
>
```

As it turns out to be less than the 0.05 significance level, we reject the null hypothesis that is H_0: $\mu \leq 2$

7.1 Hypothesis testing - continued

Example 2:

The customer accounts of a certain departmental store have an average balance of Rs. 12000 and a standard deviation of Rs. 4000. Assuming that the account balances are normally distributed,

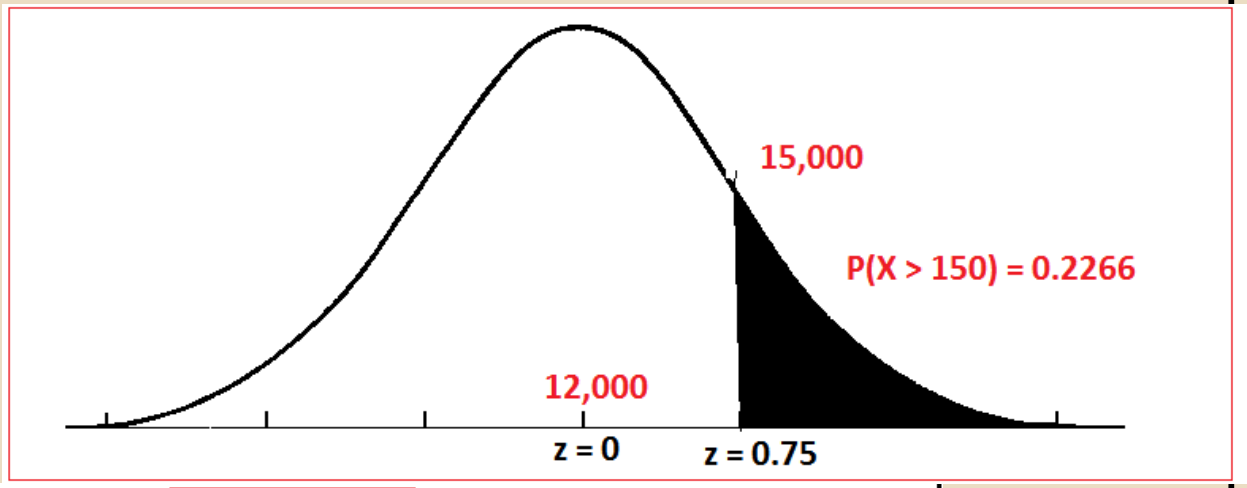
1. **What percentage of accounts has balance Rs.15000?**
2. **What proportion of accounts have balance ranging from Rs 10000 to Rs. 15000?**

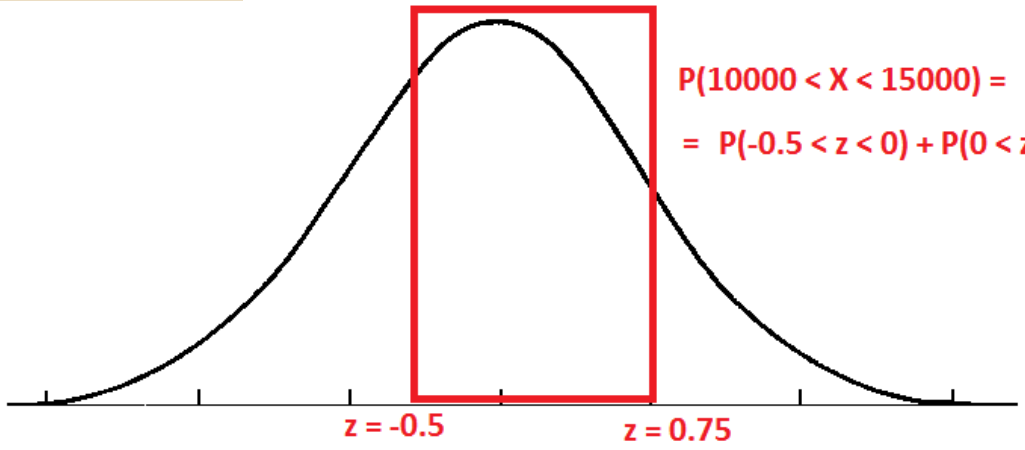
Solution: We first write a function to compute z score.

```
calculate_z<-function(meanx,popmean,sdx,n=1) {
z_calc<-
((meanx - popmean) /  (sdx/sqrt(n)))
return(z_calc)
}
```

7.1 Hypothesis testing - continued

Example 2 – continued:





7.1 Hypothesis testing - continued

1) Here μ = 12000; σ = 4000 and z = (X - μ)/ σ

- We calculate z score by calling the function calculate_z

```
> z <- calculate_z(15000,12000,4000)
> print(z)
[1] 0.75
```

- Then we find the P(z>0.75) by calling the function pnorm and setting the lower.tail = FALSE

```
> print(pnorm(z,lower.tail=FALSE))
[1] 0.2266274
```

- 22.66 % of the accounts will have balance over Rs.15000.

7.1 Hypothesis testing -continued

2) Here μ = 12000; σ = 4000 and z = (X - μ)/ σ
   We calculate the z score for X = 10000. We calculate z score by calling the function calculate_z

- Then we find the P(0.5 < z < 0) by calling the function pnorm and setting the lower.tail = TRUE

```
> # 2
> z <- calculate_z(10000,12000,4000)
> print(z)
[1] -0.5
> print(0.5 - pnorm(z,lower.tail=TRUE))
[1] 0.1914625
```

7.1 Hypothesis testing - continued

2) Here μ = 12000; σ = 4000 and z = (X - μ)/ σ
We calculate the z score for X = 15000. We calculate z score by calling the function calculate_z with lower .tail = FALSE

- Then we find the P(0 < z < 0.75) by calling the function pnorm and setting the lower.tail = FALSE

```
> z <- calculate_z(15000,12000,4000)
> print(z)
[1] 0.75
> print(0.5 - pnorm(z,lower.tail=FALSE))
[1] 0.2733726
```

- So, 46.49 % of the accounts will have balance over Rs.15000. (0.1915 + 0.2734 = 0.4649)

7.2 t Test

- A bottle filling machine is set to fill bottles with drinking water to a volume of 500 ml. The actual volume is known to follow a normal distribution.
- The manufacturer believes the machine is under filling bottles. A sample of 20 bottles is taken and the volume of liquid is measured.
- The results are given in the bottles dataset, which is available here.

```
bottles.volume <- c(484.11,  459.49, 471.38, 512.01, 494.48, 528.63,
493.64, 485.03, 473.88, 501.59, 502.85, 538.08, 465.68, 495.03,
475.32, 529.41, 518.13, 464.32, 449.08, 489.27)
```

7.2 t Test - continued

- Determine whether the bottles are being consistently under-filled.

```
> source("17_example1.R")

    In this example, the null hypothesis is
    The mean filling volume is not less than 500ml

    From the t test we had performed on the data, we observe

    1. Mean bottle volume for sample is: 491.5705 ml

    2. Mean filling volume is likely to be less than: 505.649487
       at 99% confidence level

    3. The probability of selecting a bottle with
       a Mean filling volume less than 500 ml is: 7.243113%


     Since the p-value is  less than the significance level of 0.01
     we accept the null hypothesis
>
```

7.2 t Test - continued

```
> bottles.volume
 [1] 484.11 459.49 471.38 512.01 494.48 528.63 493.64 485.03 473.88 501.59
[11] 502.85 538.08 465.68 495.03 475.32 529.41 518.13 464.32 449.08 489.27
> mean(bottles.volume)
[1] 491.5705
```

- Suppose you want to use a one-sample t-test to determine whether the bottles are being consistently under-filled, or whether the low mean volume for the sample is purely the result of random variation.
- A one-sided test is suitable because the manufacturer is specifically interested in knowing whether the volume is less than 500 ml.

7.2 t Test - continued

- The test has the null hypothesis that the mean filling volume is equal to 500 ml, and the alternative hypothesis that the mean filling volume is less than 500 ml.
- A significance level of 0.01 is to be used.
- A t-test is used to test hypotheses about the mean value of a population from which a sample is drawn.
- A t-test is suitable if the data is believed to be drawn from a normal distribution, or if the sample size is large.

7.2 t Test - continued

- The function t.test() can be used to perform both one and two sample t-tests on vectors of data.
- The function contains a variety of options and can be called as follows:
- t.test(x,y = NULL, alternative = c("two sided","less","greater"), mu= 0, paired= FALSE, var.equal= FALSE, conf.level = 0.95)
- Here, x is a numeric vector of data values and y is an optional numeric vector of data values.

7.2 t Test - continued

- In this R code, the results of the t-test are stored in the variable t_test_val.
- We are interested in these attributes of t_*test_val: estimate, p.value and conf.int*

D:\Training\R_in_2_days\R_data\l7_example1.R - R Editor

```
#-----------------------+-----------------------------------------------------
# One tailed t - test
#-----------------------------------------------------------------------------
bottles.volume <- c(484.11,  459.49, 471.38, 512.01,
494.48, 528.63, 493.64, 485.03, 473.88, 501.59,
502.85, 538.08, 465.68, 495.03, 475.32, 529.41,
518.13, 464.32, 449.08, 489.27)
#
t_test_val        <-    t.test(bottles.volume, mu=500,alternative="less",conf.level = 0.99)
#
cat("\n    In this example, the null hypothesis is")
cat("\n    The mean filling volume is not less than 500ml \n")
cat("\n    From the t test we had performed on the data, we observe\n")
cat("\n    1. Mean bottle volume for sample is:",t_test_val$estimate,"ml",
"\n",sep = " ")
cat(sprintf("\n    2. Mean filling volume is likely to be less than: %f ",
t_test_val$conf.int[2]),sep="")
cat("\n       at 99% confidence level \n")
cat("\n    3. The probability of selecting a bottle with")
cat(sprintf("\n       a Mean filling volume less than 500 ml is: %f",(t_test_val$p.value*100)),
"%","\n\n",sep = "")
```

7.2 t Test - continued

```
concl_s1_ph1      <-    "    Since the p-value is "
less              <-    "less"
greater           <-    "not less"
concl_s1_ph2      <-    "than the significance level of 0.01 "
concl_s2_ph1      <-    "    we"
less              <-    "less"
greater           <-    "greater"
accept            <-    "accept"
reject            <-    "reject"
concl_s2_ph2      <-    "the null hypothesis"
accept_h0_1       <-    paste(concl_s1_ph1,less,concl_s1_ph2)
accept_h0_2       <-    paste(concl_s2_ph1,accept,concl_s2_ph2)
#
reject_h0_1       <-    paste(concl_s1_ph1,greater,concl_s1_ph2)
reject_h0_2       <-    paste(concl_s2_ph1,reject,concl_s2_ph2)
conclusion        <-    ifelse(t_test_val$p.value > 0.01,"accept","reject")
if (conclusion == "accept")
{
    cat("\n",accept_h0_1)
    cat("\n",accept_h0_2,"\n")
}else{
    cat("\n",reject_h0_1)
    cat("\n",reject_h0_2,"\n")
}
#
#---------------------------------------------------------------------
```

7.3 Non-Parametric tests

- **Test if the mean of goals suffered by two football teams over the years is the same.**

Team A	**6**	**8**	**2**	**4**	**4**	**5**
Team B	**7**	**10**	**4**	**3**	**5**	**6**

Solution

- **The Wilcoxon-Matt-Whitney test (or Wilcoxon rank sum test or Mann-Whitney U- test) is used when you are asked to compare the means of two groups that do not follow a normal distribution.**

7.3 Non-Parametric tests - continued

```
> source("l7_example2.R")

 Null hypothesis: Mean goals suffered by
 both teams are same


 P-value: 0.5174126

 Conclusion: p-value is greater than 0.05,
 Hence, we accept the null hypothesis

Warning message:
In wilcox.test.default(a, b, correct = FALSE) :
  cannot compute exact p-value with ties
```

R Programming Course

7.3 Non-Parametric tests - continued

D:\Training\R_in_2_days\R_data\I7_example2.R - R Editor

```
a                <- c(6,8,2,4,4,5)
b                <- c(7,10,4,3,5,6)
w                <- wilcox.test(a,b,correct=FALSE)
accept           <- "accept"
reject           <- "reject"
conclusion       <- ifelse (w$p.value >0.05,accept,reject)
cat("\n Null hypothesis: Mean goals suffered by")
cat("\n both teams are same","\n")
cat("\n\n P-value:",w$p.value)
line_accept1     <- "Conclusion: p-value is greater than 0.05,"
line_accept2     <- " Hence, we accept the null hypothesis"
#
line_reject1     <- "Conclusion: p-value is not greater than 0.05,"
line_reject1     <- " Hence, we reject the null hypothesis"
if (conclusion == "accept") {
cat("\n\n",line_accept1,"\n")
cat(line_accept2,"\n\n")
} else {
cat("\n\n",line_reject1,"\n")
cat(line_reject2,"\n\n")
}
```

7.3 Non-Parametric tests - continued

Parametric test

- **If the information about the population is completely known by means of its parameters then statistical test is called parametric test.**
- **Eg: t-test, f-test, z-test,ANOVA**
- **If there is no knowledge about the population or parameters, but still it is required to test the hypothesis of the population. Then it is called non-parametric test.**
- **Eg: Mann-Whitney, Rank sum test, Kruskal-Wallis test**

7.3 Non-Parametric tests - continued

Parametric test – continued

- **Non-parametric statistics is a statistical method wherein the data is not required to fit a normal distribution.**
- **Non-parametric statistics uses data that is often ordinal, meaning it does not rely on numbers, but rather a ranking or order of sorts.**

7.4 Chi- Square test

Example 3:

- **You have three groups based on age (under 45, between 45 and 59 and over 60) and you have nominal data for each group - the frequency of regular health check up (yearly, occasionally, never).**
- **Find out whether the outcomes for the two groups were statistically equal.**
- **Test the hypothesis whether the frequency of the medical check-up is independent of the age group at 0.05 significance level.**
- **You tally the frequency of medical check ups with the age group in the following table, contingency table of the two variables.**

7.4 Chi- Square test - continued

Age	Frequency of regular medical check ups		
	Yearly	Occasionally	Never
Under 45	91	90	51
45 – 59	150	200	155
60 and over	109	198	172

- H_0: Medical check-up is independent of the age group at 0.05 significance level
- H_1: Medical check-up is dependent of the age group at 0.05 significance level

7.4 Chi- Square test - continued

```
> source("17_example3.R")

 We frame hypotheses for this problem


 H0: Medical check-up is independent of the age group
     at 0.05 significance level
 H1: Medical check-up is dependent of the age group
     at 0.05 significance level

 P-value: 5.077788e-05

 Conclusion: p-value is not greater than 0.05,
 Hence, we accept the alternative hypothesis,H1

>
```

7.4 Chi- Square test - continued

The first part of the R script l7_example3.R

```
#l7_example3.R
#-----------------------------------------------------------------
# chi.square test
#-----------------------------------------------------------------
g1                <- c(91,91,51)
g2                <- c(150,200,155)
g3                <- c(109,198,172)
df                <- data.frame(rbind(g1,g2,g3))
#-----------------------------------------------------------------
ch                <- chisq.test(df)
accept            <- "accept"
reject            <- "reject"
conclusion        <- ifelse (ch$p.value >0.05,accept,reject)
cat("\n We frame hypotheses for this problem \n")
cat("\n H0: Medical check-up is independent of the age group")
cat("\n     at 0.05 significance level")
cat("\n H1: Medical check-up is dependent of the age group")
cat("\n     at 0.05 significance level")
```

7.4 Chi- Square test - continued

The second part of the R script l7_example3.R

```
cat("\n\n P-value:",ch$p.value)
line_accept1 <- "Conclusion: p-value is greater than 0.05,"
line_accept2 <- " Hence, we accept the null hypothesis,H0"
#
line_reject1 <- "Conclusion: p-value is not greater than 0.05,"
line_reject2 <- " Hence, we accept the alternative hypothesis,H1"
if (conclusion == "accept") {
cat("\n\n",line_accept1,"\n")
cat(line_accept2,"\n\n")
} else {
cat("\n\n",line_reject1,"\n")
cat(line_reject2,"\n\n")
}
```

7.4 Chi- Square test - continued

- Chi-square is a statistical test commonly used to compare observed data with data we would expect to obtain according to a specific hypothesis.
- Let the probabilities of various classes in a distribution be p_1, $p_{2, \ldots}$ p_k with observed frequencies m_1, $m_{2, \ldots}$ m_k.

$$\chi_s^2 = \sum_{i=1}^{k} \frac{(m_i - N\,p_i)^2}{N\,p_i}$$

- χ_s^2 This quantity is therefore a measure of the deviation of a sample from expectation, where N is the sample size.

7.4 Chi- Square test - continued

Arguments for chisq.test()

X	a numeric vector or matrix x and y can also both be factors
y	a numeric vector; ignored if x is a matrix. If x is a factor, y should be a factor of same length
correct	a logical indicating whether to apply continuity correction when computing the test statistic for 2 by2 tables; one half is subtracted from all |O – E| differences; however, the correction will not be bigger than the differences themselves. No correction is done if simulate.p.value = TRUE.
p	a vector of probabilities of the same length of x.
rescale.p	a logical scalar; if TRUE then p is rescaled (if necessary) to sum to 1. If rescale.p is FALSE and p does not sum to 1, an error is given.
simulate.p.value	a logical indicating whether to compute p-values by Monte Carlo simulation.
B	an integer specifying the number of replicates used in the Monte Carlo test.

7.4 Chi- Square test - continued

- The function chisq.test is used for test of independence and goodness of fit.
- If "x" is a 2-D table, array, or matrix, then it is assumed to be a contingency table of frequencies, and a test of independence will be done.
- The correct = TRUE option applies the Yates continuity correction when "x" is a 2 X 2 table. Set this to FALSE, if the correction is not desired.
- For the goodness of fit test, set "p" equal to the null hypothesized proportions or probabilities for each of the categories represented in the vector "x".

7.4 Chi- Square test - continued

- A survey is conducted to study the student's smoking habits. There are four proper responses in the survey: "Heavy","Regularly","Occasionally" and "Never". The smoking data is multinominal. (Data taken from the built-in data set survey available in the library MASS)
- The frequency distribution is given below for the smoking data.
- As per the campus smoking statistics, we have

Heavy	Never	Occasionally	Regularly
11	189	19	17

7.4 Chi- Square test - continued

- As per the campus smoking statistics, we have

Heavy	Never	Occasionally	Regularly
4.5 %	79.5 %	8.5 %	7.5 %

- Determine whether the sample data in survey supports it at 0.05 significance level.

7.4 Chi- Square test - continued

```
> source("17_example4.R")

 We frame hypotheses for this problem

 H0: The sample data supports the campus-wide survey
     at 0.05 significance level
 H1: The sample data does not support the campus-wide survey
     at 0.05 significance level

 P-value: 0.9909295

 Conclusion: p-value is greater than 0.05,
 Hence, we accept the null hypothesis,H0

>
```

7.4 Chi- Square test - continued

```
#17_example4.R
#-----------------------------------------------------------------
# chi.square test
#-----------------------------------------------------------------
smoke.frequency     <- c(11,189,19,17)
smoke.prob          <- c(0.045,0.795,0.085,0.075)
#-----------------------------------------------------------------
chi_2               <- chisq.test(smoke.frequency,p=smoke.prob)
accept              <- "accept"
reject              <- "reject"
conclusion          <- ifelse (chi_2$p.value >0.05,accept,reject)
cat("\n We frame hypotheses for this problem \n")
cat("\n H0: The sample data supports the campus-wide survey")
cat("\n     at 0.05 significance level")
cat("\n H1: The sample data does not support the campus-wide survey")
cat("\n     at 0.05 significance level")
```

7.4 Chi- Square test - continued

```
cat("\n\n P-value:",chi_2$p.value)
line_accept1 <- "Conclusion: p-value is greater than 0.05,"
line_accept2 <- " Hence, we accept the null hypothesis,H0"
#
line_reject1 <- "Conclusion: p-value is not greater than 0.05,"
line_reject2 <- " Hence, we accept the alternative hypothesis,H1"
if (conclusion == "accept") {
cat("\n\n",line_accept1,"\n")
cat(line_accept2,"\n\n")
} else {
cat("\n\n",line_reject1,"\n")
cat(line_reject2,"\n\n")
}
```

7.4 Chi- Square test - continued

Multinomial Goodness of Fit

- A population is called multinomial if its data is categorical and belongs to a collection of discrete non-overlapping classes.
- The null hypothesis for goodness of fit test for multinomial distribution is that the observed frequency f_i is equal to an expected count e_i in each category.
- Null hypothesis is rejected, if the p-value of the Chi-Squared test statistic is less than a given significance level α.

7.5 ANOVA

- We have raised broods of flies on various flies on various sugars. We measure the size of the individual flies and record the diet for each. Our data file would consist of two columns; one for growth and one for sugar.

Growth	75	72	73	61	67	64	62	63	68
Sugar	C	C	C	F	F	F	S	S	S
Growth	72	77	78	82	83	78	59	61	63
Sugar	D	D	D	B	B	B	A	A	A

- Find out if there is significant effect of diet upon growth.

7.5 ANOVA - continued

- If so, which of these treatments are significantly different from other treatments.

Solution:

- We use aov() function to perform analysis of variance test.
- When we have a statistically significant effect in ANOVA and an independent variable of more than two levels, we typically want to make follow-up comparisons.
- We use Tukey Honest Significant Difference (HSD). for making pairwise comparisons.
- We use TukeyHSD(x) function, where x is a linear model object created using aov(formula,data) function.

7.5 ANOVA - continued

```
> source("l7_example5.R")

 We frame hypotheses for this problem

 H0: There is no significant effect of treatment (diet) upon growth
     at 0.05 significance level
 H1: There is a significant effect of treatment (diet) upon growth
     at 0.05 significance level

 P-value: 4.553398e-06

 Conclusion: p-value is not greater than 0.05
 Hence, we accept the alternative hypothesis,H1

 From TukeyHSD post-adhoc analysis, we find the following treatments
 are significantly different from other treatments and control:
  1 : B-A  2 : C-A  3 : D-A  4 : C-B  5 : F-B
  6 : S-B  7 : F-C  8 : S-C  9 : F-D  10 : S-D
>
```

7.5 ANOVA - continued

```
#l7_example5.R
#---------------------------------------------------------------------
# ANOVA test
#---------------------------------------------------------------------
growth                                          <- c(75,72,73,61,67,64,62,63,68,72,
77,78,82,83,78,59,61,63)
sugar                                           <- c("C","C","C","F","F","F","S","S","S",
"D","D","D","B","B","B","A","A","A")
#---------------------------------------------------------------------
my_aov                                          <- aov(growth ~ sugar)
p_value                                 <- summary(my_aov)[[1]][["Pr(>F)"]][1]
accept                                  <- "accept"
reject                                  <- "reject"
conclusion                              <- ifelse (p_value > 0.05,accept,reject)
cat("\n We frame hypotheses for this problem \n")
cat("\n H0: There is no significant effect of treatment (diet) upon growth")
cat("\n     at 0.05 significance level")
cat("\n H1: There is a significant effect of treatment (diet) upon growth")
cat("\n     at 0.05 significance level")
cat("\n\n P-value:",p_value)
#
```

7.5 ANOVA - continued

```
#
line_accept1                    <- "Conclusion: p-value is greater than 0.05"
line_accept2                    <- " Hence, we accept the null hypothesis,H0"
#
line_reject1                    <- "Conclusion: p-value is not greater than 0.05"
line_reject2                    <- " Hence, we accept the alternative hypothesis,H1"
#
if (conclusion == "accept") {
cat("\n\n",line_accept1,"\n")
cat(line_accept2,"\n\n")
} else {
        cat("\n\n",line_reject1,"\n")
        cat(line_reject2,"\n")
        tuk                             <- TukeyHSD(my_aov)
        line_3                  <- "From TukeyHSD post-adhoc analysis, we find the following treatments"
  line_4                        <-      "are significantly different from other treatments and control:"
        cat("\n",line_3)
        cat("\n",line_4,"\n")
        k <- length(tuk$sugar[,1])
        j <- 0
        for (i in 1:k) {
                conclusion_2 <- ifelse(tuk$sugar[i,"p adj"]> 0.05,accept,reject)
                if (conclusion_2 == "reject")  {
                 j      <- j +1; cat(" ",j,":",rownames(tuk$sugar)[[i]],sep=" ")
                 if (j %% 5 == 0) cat("\n")
                }
        }
        cat("\n")
}
```

7.5 ANOVA - continued

- Like the two-sample t-test, ANOVA lets us test hypotheses about the mean of a dependent variable across different groups.
- While the t-test is used to compare the means between two groups, ANOVA is used to compare means between 3 or more groups.

1 One-way ANOVA involves single independent variable

2. Two-way ANOVA involves two or more independent variables

- The null hypothesis for ANOVA is that the mean is the same for all groups. The alternative hypothesis is that the mean is not the same for all groups.

7.5 ANOVA - continued

- The ANOVA test procedure produces an F-statistic, which is used to calculate the p-value. If p <0.05, we reject the null hypothesis.
- We can then conclude that the mean of the dependent variable is not the same for all groups.
- With ANOVA, if the null hypothesis is rejected, then we only know that at least 2 groups are different from each other.
- In order to determine which of the two groups are different from which, post-adhoc t-tests (eg. Tukey HSD) are performed.

7.6 Monte Carlo Simulation

- There is an interesting article in the website "Frontline Solvers – Developers of the Excel Solver" about a Business Planning sample using Monte Carlo Simulation.

  (http://www.solver.com/monte-carlo-simulation-example)

- You, as a marketing manager of a firm are planning to introduce a new product. You need to estimate the first year profit from this product, which will depend on:

  1. Sales volume in units
  2. Price per unit
  3. Unit Cost
  4. Fixed Cost

7.6 Monte Carlo Simulation

- Net profit will be calculated as Net Profit = Sales Volume * (Selling Price – Unit Cost) – Fixed Cost.
- Fixed costs (for overhead, advertising, etc.) are known to be $120,000.
- But other factors all involve some uncertainty. Sales volume (in units) can over quite a range, and the selling price per unit will depend on competitor actions.
- Unit costs will also vary depending on vendor prices and production experience.

7.6 Monte Carlo Simulation - continued

Uncertain variables

- To build a risk analysis model, you must identify the uncertain variables – also called random variables. While there's some uncertainty in almost all variables in a business model, we want to focus on variables where the range of values is significant.

Sales and Price

- Based on the market research, you believe that there are equal chances that the market will be Slow, OK or Hot.
- In the "Slow market" scenario, you expect to sell 50,000 units at an average selling price of $11 per unit.

7.6 Monte Carlo Simulation - continued

- In the "OK market" scenario, you expect to sell 75,000 units at an average selling price of $10 per unit.
- In the "Hot market" scenario, you expect to sell 100,000 units at an average selling price of $8 per unit. In this scenario, your competitors will push your price down.

Unit Cost

- Another uncertain variable is unit cost. Your firm's production manager advises you that unit costs may vary anywhere between $5.50 to $7.50, with a most likely cost of $6.50. In this case, the most likely cost is also the average cost.

7.6 Monte Carlo Simulation - continued

Uncertain Functions

Net Profit

- **We now identify uncertain functions – also called functions of a random variable.**
- **Since Sales volume and Selling Price and Unit cost are all uncertain variables, the net profit calculated based on these variables is an uncertain function.**

7.6 Monte Carlo Simulation - continued

Solution:

- **Since there are equal chances that the market will be Slow, OK, or Hot, we want to create an uncertain variable that selects among these three possibilities by generating a random number – say 1 or 2 or 3 – with equal probability.**
- **We associate 1 with "Slow Market" state, 2 with "OK market" state and 3 with "Hot market" state.**
- **We generate this number easily by using the R function – "sample" and then base the Sales Volume and Selling Price of this uncertain variable.**

7.6 Monte Carlo Simulation - continued

Solution continued:

- **Unit cost may vary anywhere from $5.50 to $7.50, with a most likely cost of $6.50. We use triangular distribution to generate this variable.**

- ***In probability theory and statistics, the triangular distribution is a continuous probability distribution with a lower limit a, upper limit b and mode c, where $a < b$ and $a \leq c \leq b$.***

- **We install "triangle" package first before using the R to solve this problem.**

7.6 Monte Carlo Simulation - continued

Solution continued:

```
> install.packages("triangle")
--- Please select a CRAN mirror for use in this session ---
```

7.6 Monte Carlo Simulation - continued

Solution continued:

```
> # ------------------------------------------------------------------------------
> # Function to build a risk analysis model
> # We estimate the first year net profit which depends on
> # sales volume in units, price per unit, unit cost and fixed cost
> # Based on your market research, you believe that there are equal chances that the
> # market will be slow, ok  or hot
> # Slow market, sales volume is 50000 units selling price $ 11.00
> # OK market, sales volume is 75000 units selling price $ 10.00
> # Hot market, sales volume is 100000 units selling price $ 8.00
> # Unit cost varies from 5.50 to 7.50 with the most likely cost being 6.5
> # We simulate this model 10000 times
> # ------------------------------------------------------------------------------
> require(triangle)
> monteSimulation<-function(n=10000) {
+ fixed_cost=120000
+ result<-numeric(0)
+ for ( i in 1:n) {
+ r = sample(1:3,1)
+ sales_volume=(r * 25000) + 25000
+ sale_price=12- r
+ unit_cost=rltriangle(1,5.50, 7.50,6.50)
+ #
+ net_profit = sales_volume * (sale_price - unit_cost) - fixed_cost
+ result<-c(result,net_profit)
+ }
+ #print(class(result))
+ print(mean(result,na.rm=TRUE))
+ return(result)
+ #
+ }
```

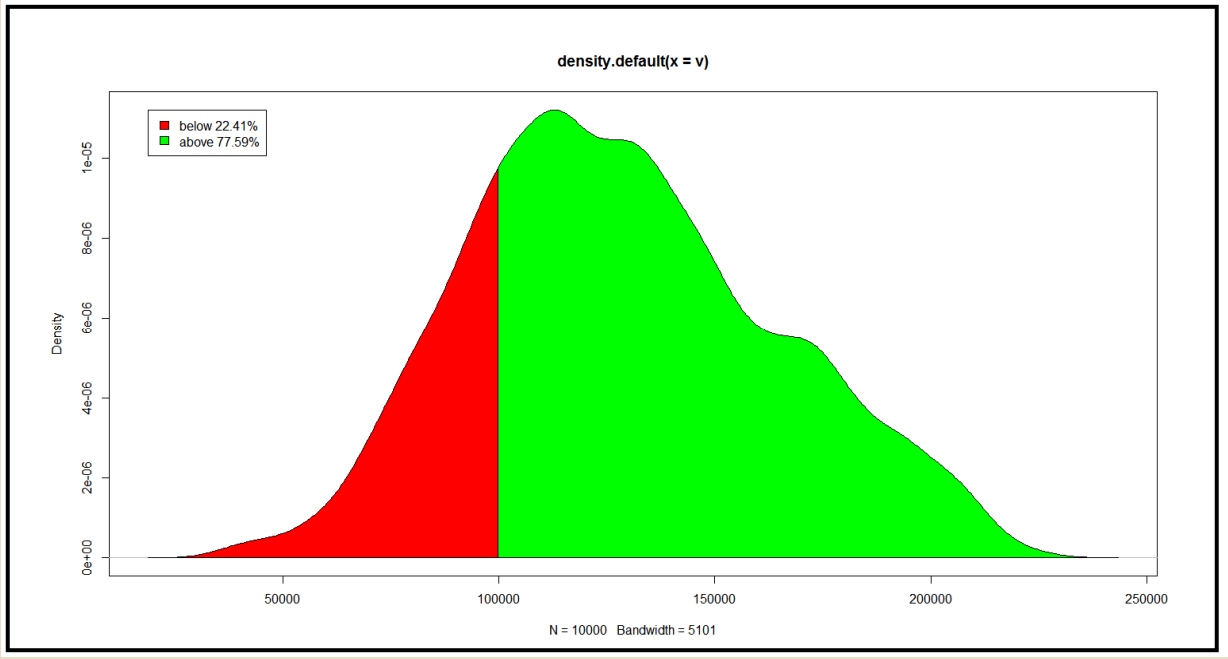
7.6 Monte Carlo Simulation - continued

Solution continued:

```
> threshold <- function(v, t, low.col="red", high.col="green") {
+ d = density(v)
+ l = length(d$x)
+ n = length(d$x[d$x<t])
+ plot(d)
+ x = c(d$x[1:n], d$x[n])
+ y = c(d$y[1:n], d$y[1])
+ polygon(x,y, col=low.col)
+ x = c(d$x[n], d$x[n:l])
+ y = c(d$y[1], d$y[n:l])
+ polygon(x, y, col = high.col)
+ pct = c(length(v[v<t])/length(v), 1 - length(v[v<t])/length(v))
+ pct = pct * 100
+ labels = c("below", "above")
+ labels = paste(labels, pct)
+ labels = paste(labels, "%", sep="")
+ legend(min(d$x) ,max(d$y), labels, fill=c(low.col, high.col))
+ }
> res<-monteSimulation(10000)
[1] 128605
> threshold(res, 100000)
```

7.6 Monte Carlo Simulation - continued

Solution continued:



We have observed that the average net profit after 10000 trials is $128,605. From this graph, we see that 77.59% of the trials resulted in the net profit above $100,00 and 22.41% of the trials resulted in the net profit below $100,000.

7.7 Correlation

- Correlation is a method of studying the relationship between two variables.
- In statistical analysis, we come across the study of two variables wherein the change in the value of one variable produces a change in the value of another variable. In that case, we say that the variables are correlated or there is a correlation between the two variables.
- Two variables may have a positive correlation, a negative correlation or they may be uncorrelated.

7.7 Correlation - continued

Example:

A tax consultant charges the following rates for his / her clients based on the annual income of the client. Using R draw a scatter plot.

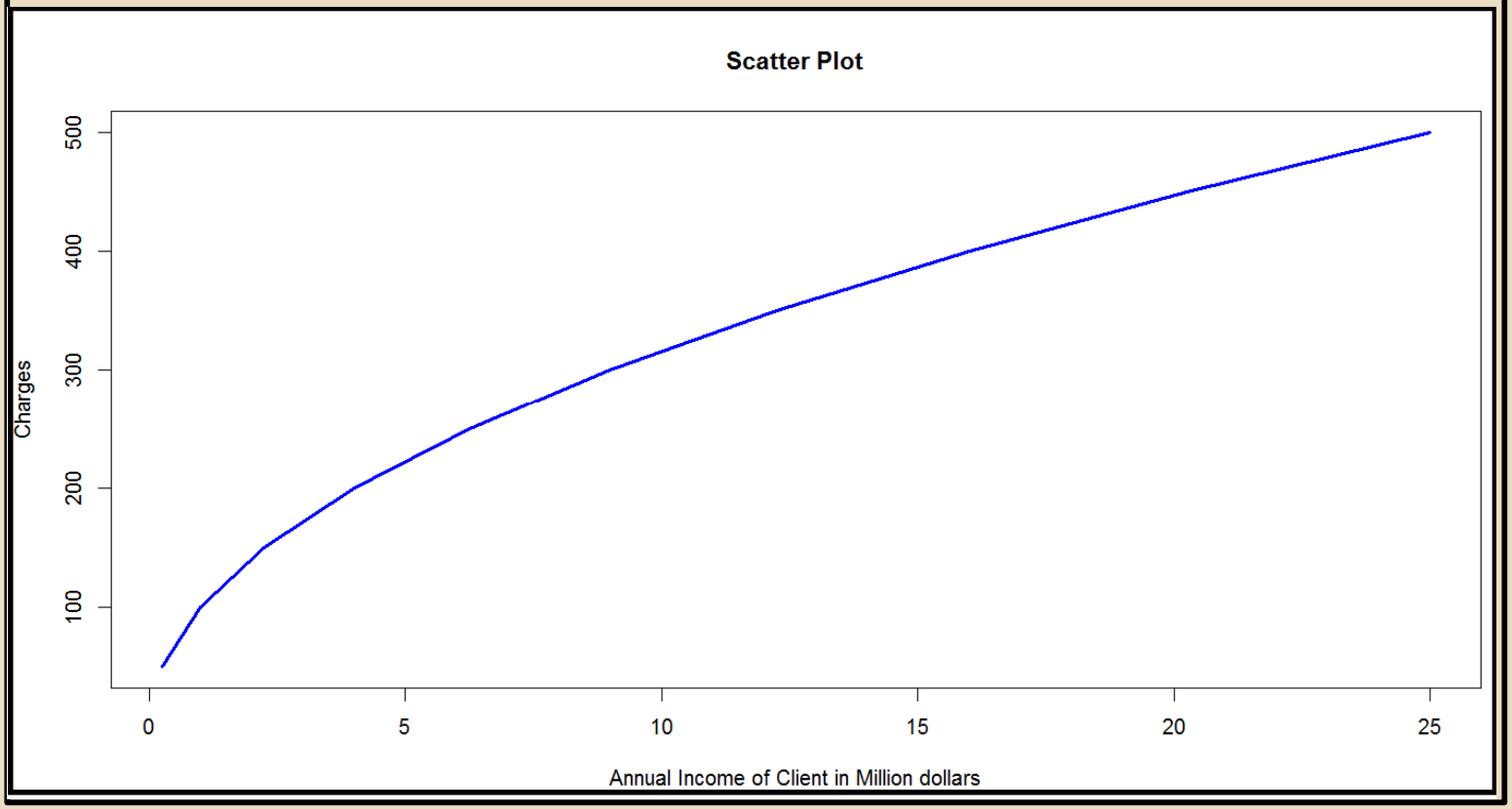
Charges	Annual Income of the client (in million dollars)
50	0.25
100	1.00
150	2.25
200	4.00
250	6.25
300	9.00
350	12.25
400	16.00
450	20.25
500	25.00

7.7 Correlation - continued

```
> setwd("D:/Training/R_in_2_days/R_data")
> source("17 example6.R")
```

```
#-----------------------------------------------
#17_example_6.R
#-----------------------------------------------
# Correlation
#-----------------------------------------------
y  <- seq(50,500,by=50)
x  <- c(0.25,1,2.25,4,6.25,9,12.25,16,20.25,25)
plot(x,y,main="Scatter Plot",
xlab="Annual Income of Client in Million dollars",
ylab="Charges",type="l",col="blue",lwd=3)
```

7.7 Correlation - continued



7.7 Correlation - continued

- The following table gives the age – distribution of the population and the number of unemployed in town.
- Find the coefficient of correlation between the mid-values of the age- groups and the percentage of unemployed in different age – constituents.

Age	Number of persons in '000	Number of unemployed
20 – 30	40	400
30 – 40	55	1100
40 – 50	32	960
50 - 60	20	1600
60 – 70	8	1600

7.7 Correlation - continued

```
> age_mid <-   c(25,35,45,55,65)
> population<-   c(40,55,32,20,8)
> unemployed<-   c(400,1100,960,1600,1600)
> df<-   data.frame(age=age_mid,perc=(unemployed * 100)/(population * 1000))
> cor(df$age,df$perc,method="pearson")
[1] 0.8856867
```

The Pearson coefficient of correlation between the mid-values of the age-groups and the percentage of unemployed in different age – constituents is approximately, 0.886.

7.8 Regression

- Correlation denotes the association between two quantitative variables.
- Regression involves estimating the best equation to summarize the association.
- The degree of association is measured by a correlation coefficient, denoted by r. It is sometimes called Pearson's correlation coefficient after its originator and is a measure of linear association.
- If a curved line is needed to express the relationship, other and more complicated measures of correlation must be used between the two variables.

7.8 Regression - continued

- Correlation describes the strength of an association between two variables, and is completely symmetrical, the correlation between A and B is the same as the correlation between B and A.
- However, if the two variables are related, it means that when one changes by a certain amount the other changes on an average by a certain amount.
- If y represents the dependent variable and x the independent variable, this relationship is described as the regression of y on x.

7.8 Regression - continued

- The regression equation representing how much y changes with any given change of x can be used to construct a regression line on a scatter diagram, and in the simplest case this is assumed to be a straight line.
- The direction in which the line slopes depends on whether the correlation is positive or negative.
- When the two sets of observations increase or decrease together (positive) the line slopes upwards them left to right; when one set decreases as the other increases the line slopes downwards from left to right.

7.8 Regression - continued

- As the line must be straight, it will probably pass through few, if any, of the dots.
- Given that the association is well described by a straight line we have to define two features of the line if we are to place it correctly on the diagram.
- First one is its distance from the baseline and the other being its slope.
- This is explained in the following regression equation, $y = a + bx$
- Regression equation enables us to predict y from x and gives us a better summary of the relationship between the two variables.

7.8 Regression - continued

- R includes a variety of tools for complex modeling, among them:
  1. *glm() for generalized linear models*
  2. *gam() for generalized additive models*
  3. *lme() and lmer() for linear mixed-effects models*
  4. *nls() and nlme() for nonlinear models*
- R functions such as aov(), lm() use a formula interface to specify the variables to be included in the analysis.
- The formula determines the model that will be built and tested by the R procedure. The basic format of such a formula is

  Response variable ~ explanatory variables

7.8 Regression - continued

- The tilde should be read “is modeled by” or “is modeled as a function of”.
- A basic regression analysis would be formulated this way:
  y ~ x
- … where “x” is the explanatory variable or IV and “y” is the response variable or DV. Additional variables would be added in as follows:
- y ~ x + z which would make this z multiple regression with two predictors.
- Meaning of the symbols used in the formula is given below:

7.8 Regression - continued

Example - Find the regression equation between Y1 and X1,X2,X3,X4

- Y1 is a measure of success in graduate school
- X1 is a measure of intellectual ability
- X2 is a measure of work ethic
- X3 is a second measure of intellectual ability
- X4 is a measure of spatial ability

Y1	125	158	207	182	196	175	145	144	160	175
	151	161	200	173	175	162	155	230	162	153
X1	13	39	52	29	50	64	11	22	30	51
	27	41	51	37	23	43	38	62	28	30
X2	18	18	50	43	37	19	27	23	18	11
	15	22	52	36	48	15	19	56	30	25

7.8 Regression - continued

X3	25	59	62	50	65	79	17	31	34	58
	29	53	75	44	27	65	62	75	36	41
X4	11	30	53	29	56	49	14	17	22	40
	31	39	36	27	20	36	37	50	20	33

```
> source("l7_example7.R")
```

Regression line is Y1 = 102.7439 + (1.254015 * X1) + (1.064291 * X2) + (-0.3713815 * X3) + (0.233896 * X4)

The model explains 94.85367 % of the variability of the response data around its mean

7.8 Regression - continued

```
#-----------------------------------------------------------------------------------------------------------------------
# l7_example7.R
#-----------------------------------------------------------------------------------------------------------------------
# Muliple Regression
#-----------------------------------------------------------------------------------------------------------------------
Y1          <- c(125,158,207,182,196,175,145,144,160,175,151,161,200,173,175,162,155,230,162,153)
X1          <- c(13,39,52,29,50,64,11,22,30,51,27,41,51,37,23,43,38,62,28,30)
X2          <- c(18,18,50,43,37,19,27,23,18,11,15,22,52,36,48,15,19,56,30,25)
#
X3          <- c(25,59,62,50,65,79,17,31,34,58,29,53,75,44,27,65,62,75,36,41)
#
X4          <- c(11,30,53,29,56,49,14,17,22,40,31,39,36,27,20,36,37,50,20,33)
mydf        <- data.frame(Y1,X1,X2,X3,X4)
attach(mydf)
my.lm       <- lm(Y1 ~ X1+X2+X3+X4)
summary(my.lm)
cat("\n Regression line is Y1 = ",summary(my.lm)$coefficients[1],
"+ (",summary(my.lm)$coefficients[2],"*","X1",
") + (",summary(my.lm)$coefficients[3],"*","X2",
") + (",summary(my.lm)$coefficients[4],"*","X3",
") + (",summary(my.lm)$coefficients[5],"*","X4",")\n",sep=" ")
cat("\n\n","The model explains ",summary(my.lm)$r.squared*100,"%"," of the variability of the response data around its
mean","\n",sep = " ")
detach(mydf)
```

7.8 Regression - continued

- By use of the logistic regression equation of vehicle transmission in the data set mtcars, estimate the probability of a vehicle being fitted with a manual transmission if it has a 120 hp engine and weights 2800 lbs.

Solution

- We apply the function glm to a formula that describes the transmission type (am) by the horsepower (hp) and weight (wt). This creates a generalized linear model (GLM) in the binomial family.

7.8 Regression - continued

- Now we apply the function predict to the generalized linear model am.glm along with mydf. We will have to select *response* prediction type in order to obtain the predicted probability.

```
#-------------------------------------------------------------------------------
#17_example8.R
#-------------------------------------------------------------------------------
# Logistic Regression
#-------------------------------------------------------------------------------
am.glm   <- glm(formula=am ~ hp + wt,data = mtcars, family= binomial)
mydf = data.frame(hp=120,wt=2.8)
cat("\n\n The probability of a vehicle with 120 hp engine and weights 2800 lbs
 being fitted with a manual transmission is ",
predict(am.glm,mydf,type="response"),"\n")
#-------------------------------------------------------------------------------
```

```
> source("17_example8.R")


 The probability of a vehicle with 120 hp engine and weights 2800 lbs
 being fitted with a manual transmission is  0.6418125
>
```

7.8 Regression - continued

- We use the logistic regression equation to predict the probability of a dependent variable taking the dichotomy values 0 or 1.
- Suppose $x_1, x_2, \ldots x_p$ are independent variables, α and β_k $(k=1,2,\ldots p)$ are parameters, and E(y) is the expected value of the dependent variable y, then the logistic regression equation is:

  $$E(y) = 1 / (1 + e^{-(\alpha + \sum \beta_k x_k)})$$

- For example, in the built-in data set mtcars, the data column am represents the transmission type of the automobile model (0 = automatic, 1 = manual).

7.8 Regression - continued

- With the logistic regression equation, we can model the probability of a manual transmission in a vehicle based on its engine horsepower and weight data.

P(Manual Transmission)

$= 1 / (1 + e^{-(\alpha + \beta_1 * \text{horse power} + \beta_2 * \text{weight})})$

7.9 Clustering

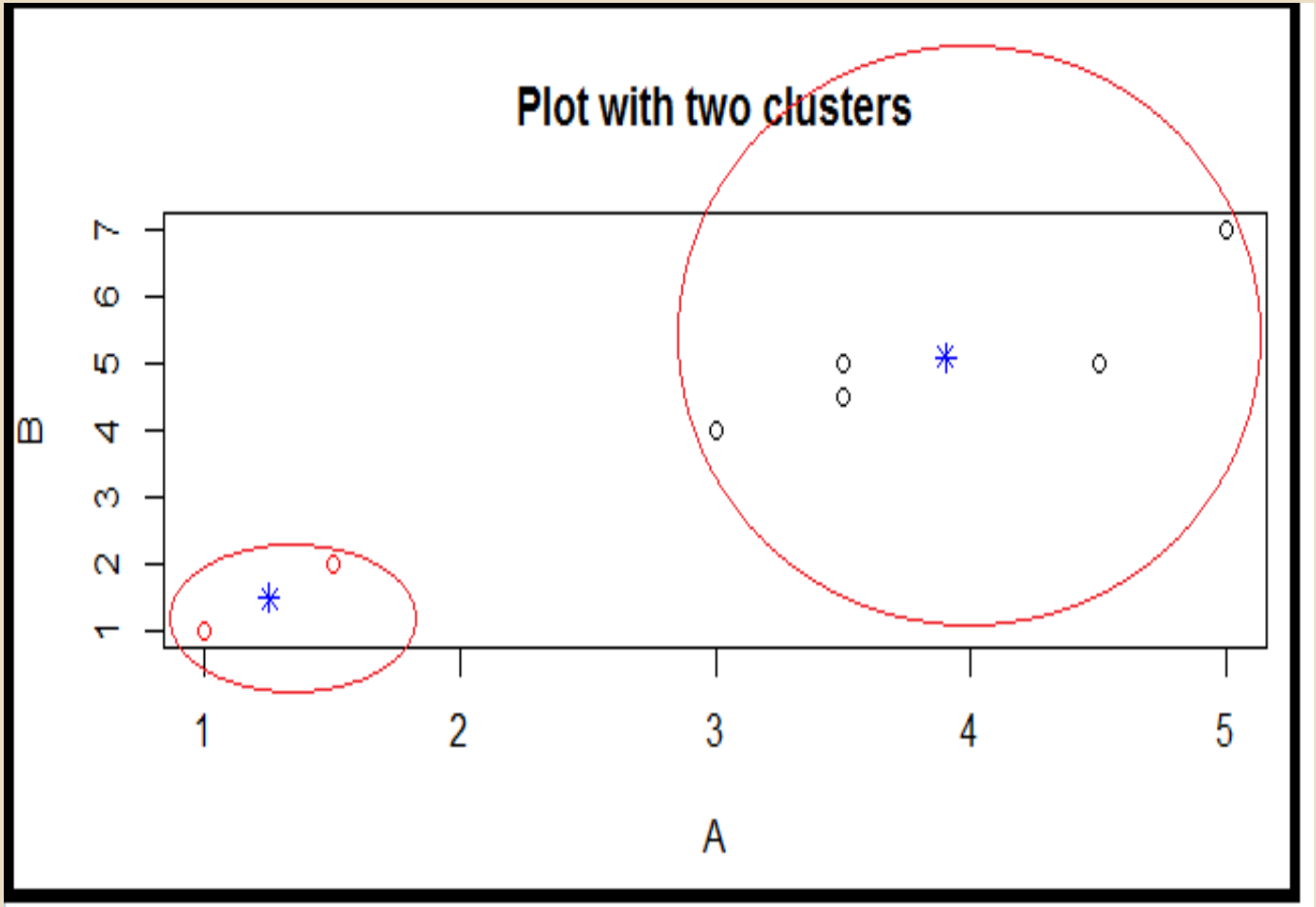
- **Consider the following data set consisting of the scores of two variables on each of seven individuals:**

Subject	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

7.9 Clustering- continued

```
> df <- data.frame(A= c(1.0,1.5,3.0,5.0,3.5,4.5,3.5), B =c(1.0,2.0,4.0,7.0,5.0,5.0,4.5))
> cluster1<- kmeans(df,2)
> par(mfrow = c(2,1))
> plot(df, col = cluster1$cluster, main = "Plot with two clusters")
> points(cluster1$centers, col = "blue", pch = 8)
> print(cluster1)
K-means clustering with 2 clusters of sizes 5, 2

Cluster means:
     A   B
1 3.90 5.1
2 1.25 1.5

Clustering vector:
[1] 2 2 1 1 1 1 1

Within cluster sum of squares by cluster:
[1] 7.900 0.625
 (between_SS / total_SS =  77.0 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
>
```

7.9 Clustering- continued



7.9 Clustering- continued

- This is a prototype-based, partitional clustering technique that attempts find a number of specified clusters (k), which are presented by their centroids (mean).
- K-means algorithm proceeds in such a way that the elements are assigned randomly to k clusters and the centroid (mean) is calculated for each cluster.
- In the next step, the elements are reassigned in such a manner that it belongs to the cluster with closest centroid.
- This process is iterated until two consecutive steps end up in the same assignment of elements.

7.9 Clustering- continued

- **In R package, k-means clustering is done using the function kmeans().**

  kmeans(x, centers, iter.max = 10, nstart = 1, algorithm = c("Hartigan-Wong","Lloyd","Forgy","MacQueen"), trace= FALSE)

Arguments

x	Numeric matrix of data, or an object that can be coerced to such a matrix
centers	Either the number of clusters, say k or a set of initial (distinct) cluster centers. If a number, a random set of (distinct) rows in x is chosen as the initial centers.
iter.max	The maximum number of iterations allowed
nstart	If centers is a number, how many random set should be chosen
Algothithm	Determines the algorithm to be used. Default is Hartigan-Wong
Trace	A logical value which product the tracing information on the progress of the algorithm, if TRUE

7.9 Clustering - continued

- Clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters.
- A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.
- Clustering algorithms
  - Exclusive clustering
  - Overlapping clustering
  - Hierarchical clustering
  - Probabilistic clustering

7.9 Clustering - continued

- **The four most used clustering algorithms:**
  - ***K-means***
  - ***Fuzzy C-means***
  - ***Hierarchical clustering***
  - ***Mixture of Gaussians***
- **Each of these algorithms belongs to one of the clustering types listed above.**
- **K-means is an exclusive clustering algorithm. Fuzzy C-means is an overlapping clustering algorithm.**
- **Hierarchical is obvious and lastly Mixture of Gaussian is a probabilistic clustering algorithm.**

R Programming Course

Activity 1:

Read the file "U08_R_Exercises_v1.pdf"