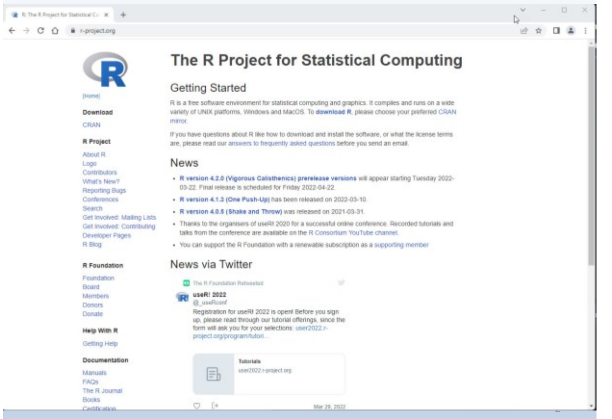
Research Methods

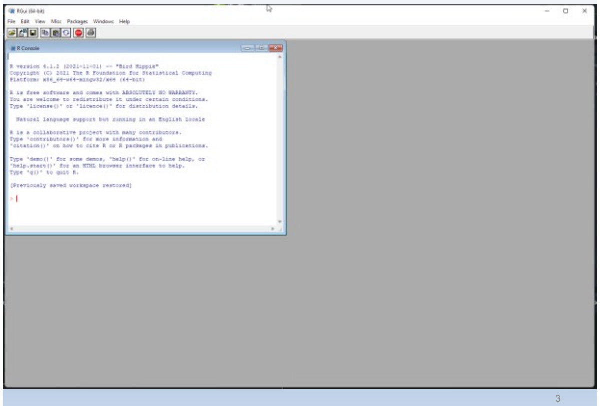
Introduction to Statistics with R

Andy Gillett

1

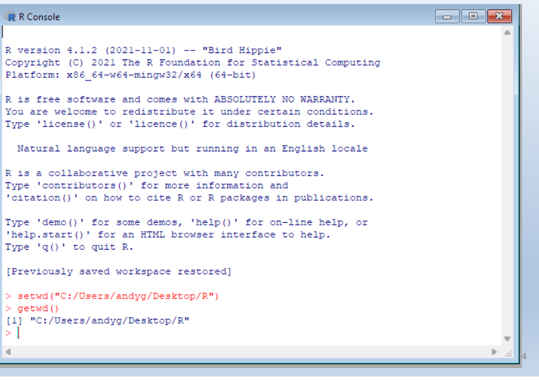


2



3

Working Directory



4

Describing Data using Statistics

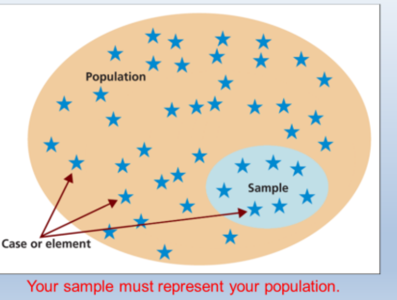
Descriptive	Inferential
Description	Inference
Sample	Population

"When an individual uses descriptive statistics, he talks about the data he has; but with inferential statistics, he talks about data he does not have."

(Popham & Sirotnik, 1973: 40)

5

Data Collection



Your sample must represent your population.

(Saunders et al, 2016: 275)

6

Describing Data using Statistics

Description

- Central Tendency
- Dispersion

Inference

- Exploring Relationships
  - Correlation
  - Regression
- Comparing Groups
  - t-Test
  - ANOVA
    - One-Way
    - Two-Way
    - ANCOVA

7

Describing Data using Statistics

Description:

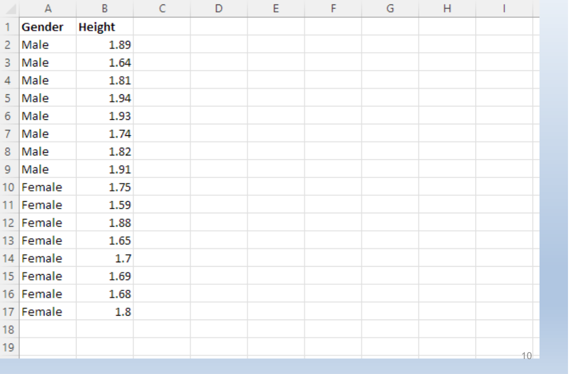
- The central tendency, which is the common, middle or average value.
- The dispersion, which is how the data values are spread, or dispersed, around the central tendency.

8

Gender	Height (m)
Male	1.89
Female	1.75
Male	1.64
Male	1.81
Female	1.59
Male	1.94
Female	1.88
Female	1.65
Male	1.93
Male	1.74
Female	1.70
Male	1.82
Male	1.91
Female	1.69
Female	1.68
Female	1.80

9

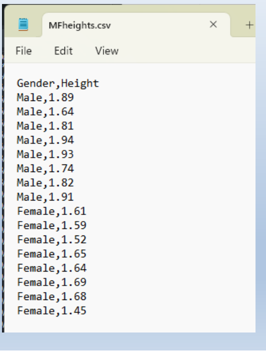
Data



	A	B
1	Gender	Height
2	Male	1.89
3	Male	1.64
4	Male	1.81
5	Male	1.94
6	Male	1.93
7	Male	1.74
8	Male	1.82
9	Male	1.91
10	Female	1.75
11	Female	1.59
12	Female	1.88
13	Female	1.65
14	Female	1.7
15	Female	1.69
16	Female	1.68
17	Female	1.8

10

MFheights.csv



```
Gender,Height
Male,1.89
Male,1.64
Male,1.81
Male,1.94
Male,1.93
Male,1.74
Male,1.82
Male,1.91
Female,1.61
Female,1.59
Female,1.52
Female,1.65
Female,1.64
Female,1.69
Female,1.68
Female,1.45
```

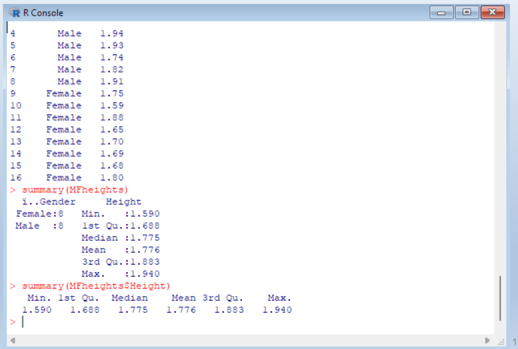
11

Import into R

```
> MFheights <- read.csv("MFheights.csv", header = TRUE, colClasses = c("factor", "numeric"))
> MFheights
   Gender Height
1    Male   1.89
2    Male   1.64
3    Male   1.81
4    Male   1.94
5    Male   1.93
6    Male   1.74
7    Male   1.82
8    Male   1.91
9  Female   1.61
10 Female   1.59
11 Female   1.52
12 Female   1.65
13 Female   1.64
14 Female   1.69
15 Female   1.68
16 Female   1.45
```

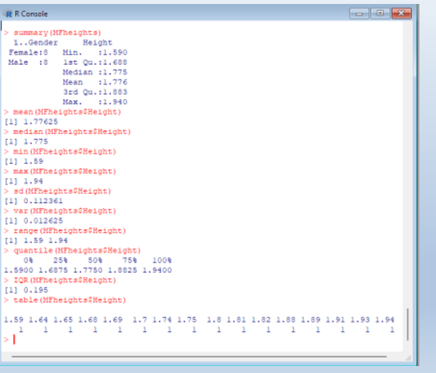
12

Summary



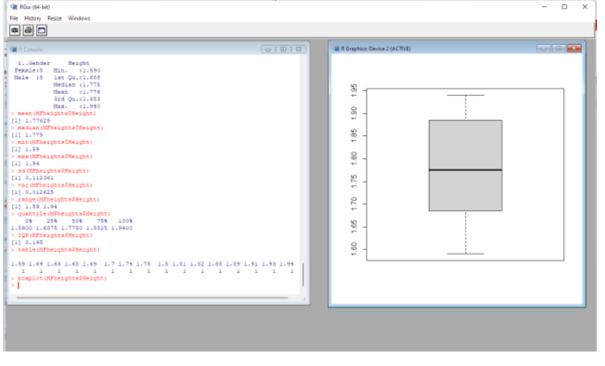
13

Various Descriptive Statistics



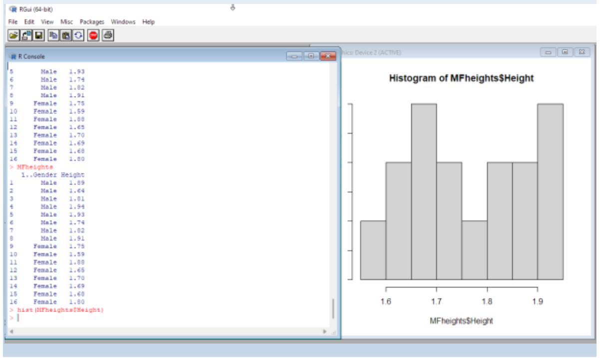
14

Boxplot



15

Histogram



16

Describing Data using Statistics

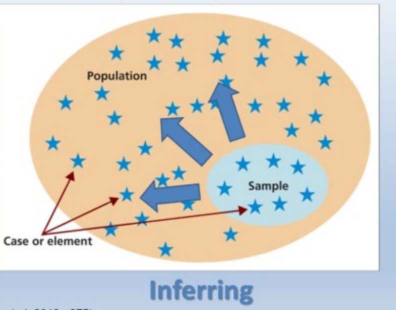
Descriptive	Inferential
Description	Inference
Sample	Population

"When an individual uses descriptive statistics, he talks about the data he has; but with inferential statistics, he talks about data he does not have."

(Popham & Sirotnik, 1973: 40)

17

Sample – Population



Inferring

(Saunders et al, 2016: 275)

18

Describing Data using Statistics

Description

- Central Tendency
  - Mean, Median, Mode
- Dispersion
  - Range, Standard Deviation, IQR

Inference

- Exploring Relationships
  - Correlation
  - Regression
- Comparing Groups
  - t-Test
  - ANOVA
    - One-Way
    - Two-Way

19

Correlation

There is much interest in whether there is a relationship between pupils' ability at sport and their maths ability. Accordingly a study was carried out in which both sport and maths ability were rated on each of 10 pupils sampled at random from a class of first year pupils in a secondary school. Sport ability was rated on a scale of 1 to 20. Maths ability was measured by a pencil and paper test on a scale from 1 to 30.

20

Correlation

Pupil No.	1	2	3	4	5	6	7	8	9	10
Sports score	15	19	12	14	9	16	20	14	15	19
Maths score	24	29	20	27	16	28	28	23	21	29

21

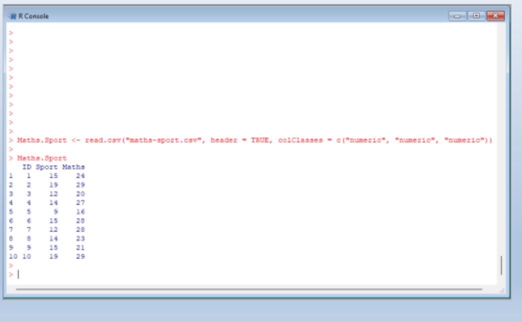
Data

A1 · fx · ID

	A	B	C	D	E	F
1	ID	Sport	Maths			
2	1	15	24			
3	2	19	29			
4	3	12	20			
5	4	14	27			
6	5	9	16			
7	6	15	28			
8	7	12	28			
9	8	14	23			
10	9	15	21			
11	10	19	29			
12						
13						
14						
15						

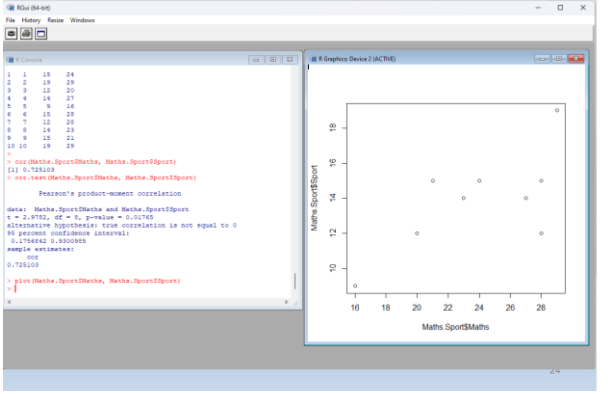
22

Import Data



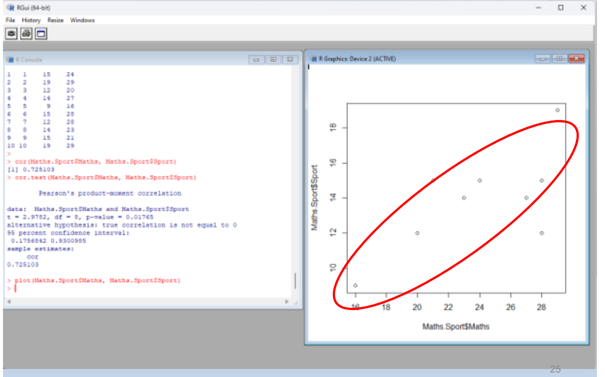
23

Plot the Relationship



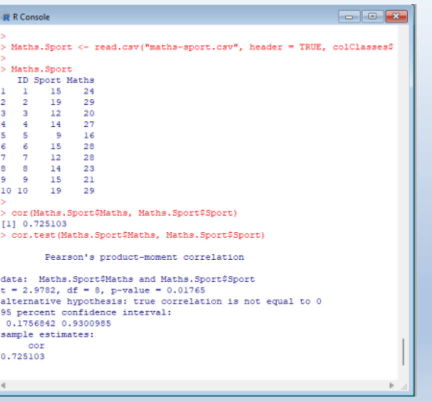
24

Correlation



25

Correlation



26

Correlation

The correlation, r, has the value 0.725.

The *p*-value is the probability that the correlation found between the two variables could be a chance result. As *p* here is less than 0.05 the correlation is unlikely to be a chance result.

The conclusion is that there is a relationship between SPORT scores and MATHS scores in the wider population of first year secondary school pupils. The results can be generalised.

27

27

Regression

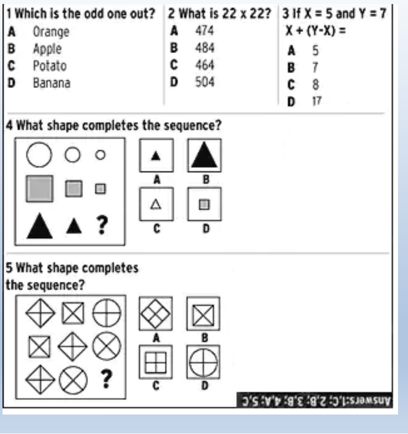
A study looked at the possibility of predicting students' future reading ability (READING) from their intelligence score (IQ).

Ten children were included in the investigation, selected at random from a primary school class. The children's IQ scores were obtained when they entered the school and the corresponding reading score was obtained at the end of the first year. The reading score was the result out of 100 on a test.

28

28

IQ



29

29

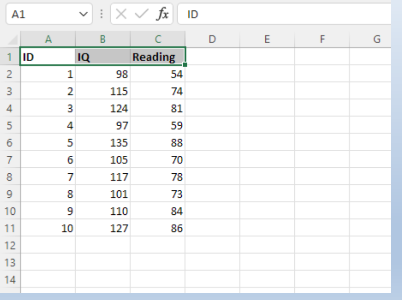
IQ : Reading

Pupil No.	1	2	3	4	5	6	7	8	9	10
IQ Score	98	115	124	97	135	105	117	101	110	127
Reading	54	74	81	59	88	70	78	73	84	86

30

30

IQ : Reading



ID	IQ	Reading
1	98	54
2	115	74
3	124	81
4	97	59
5	135	88
6	105	70
7	117	78
8	101	73
9	110	84
10	127	86

31

Hypotheses

H_0: READING score cannot be predicted from IQ score

H_1: READING score can be predicted from IQ score

H_0 = ***null hypothesis*** **-** no relationship between two measured phenomena.

Rejecting or disproving the **null hypothesis** is a central task in modern scientific practice.

Statistically significant means that the null hypothesis has been rejected.

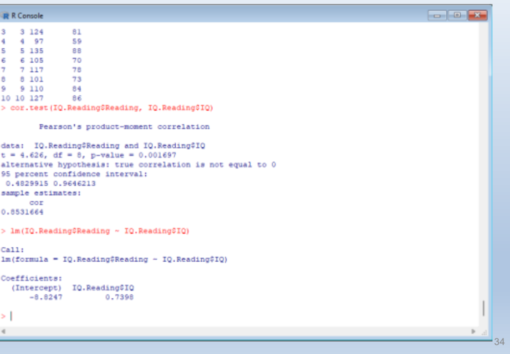
32

Import the Data

```
> IQ.Reading <- read.csv("IQ-Reading.csv", header = TRUE, colClasses = c("numeric", "numeric", "numeric"))
> IQ.Reading
   ID  IQ Reading
1   1  98      54
2   2 115      74
3   3 124      81
4   4  97      59
5   5 135      88
6   6 105      70
7   7 117      78
8   8 101      73
9   9 110      84
10 10 127      86
>
```


33

Regression Equation



```
3   3 124      81
4   4  97      59
5   5 135      88
6   6 105      70
7   7 117      78
8   8 101      73
9   9 110      84
10 10 127      86
> cor.test(IQ.Reading$Reading, IQ.Reading$IQ)

        Pearson's product-moment correlation

data:  IQ.Reading$Reading and IQ.Reading$IQ
t = 4.626, df = 8, p-value = 0.001697
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4829915 0.9646213
sample estimates:
      cor
0.8531664

> lm(IQ.Reading$Reading ~ IQ.Reading$IQ)

Call:
lm(formula = IQ.Reading$Reading ~ IQ.Reading$IQ)

Coefficients:
  (Intercept)  IQ.Reading$IQ
      -8.8247         0.7398

>
```


34

Regression Equation

$x = a + by$

$a = -8.825$
$b = .740$

35

Regression

The regression equation

"READING = −8.825 + 0.740 × IQ"

can be obtained by equating the dependent variable **Reading** to the *CONSTANT* value −8.825, plus the independent variable **IQ** multiplied by its coefficient 0.740.

36

IQ: Reading + Availability of Books

	A	B	C	D	E	F
1	ID	IQ	Books	Reading		
2	1	98	5	54		
3	2	115	25	74		
4	3	124	37	81		
5	4	97	10	59		
6	5	135	90	88		
7	6	105	80	70		
8	7	117	88	78		
9	8	101	73	73		
10	9	110	80	84		
11	10	127	100	86		
12						
13						

37

37

Hypotheses

H_0: READING score cannot be predicted from a combination of IQ score and number of Books Available

H1: READING score can be predicted from a combination of IQ score and number of Books Available

38

38

IQ/Books : Reading

	A	B	C	D	E	F
1	ID	IQ	Books	Reading		
2	1	98	5	54		
3	2	115	25	74		
4	3	124	37	81		
5	4	97	10	59		
6	5	135	90	88		
7	6	105	80	70		
8	7	117	88	78		
9	8	101	73	73		
10	9	110	80	84		
11	10	127	100	86		
12						
13						

IQ-Books-Reading.csv

39

39

Import the Data into R

```
R Console
>
> IQ.Books.Reading <- read.csv("IQ-Books-Reading.csv", header = TRUE, colClasses = c("numeric", "numeric", "numeric", "numeric"))
```

IQ.Books.Reading <- read.csv("IQ-Books-Reading.csv", header = TRUE, colClasses = c("numeric", "numeric", "numeric", "numeric"))

```
R Console
> IQ.Books.Reading
   ID  IQ Books Reading
1   1  98     5      54
2   2 115    25      74
3   3 124    37      81
4   4  97    10      59
5   5 135    90      88
6   6 105    80      70
7   7 117    88      78
8   8 101    73      73
9   9 110    80      84
10 10 127   100      86
> |
```

40

40

Regression Equation

```
R Console
1   1  98     5      54
2   2 115    25      74
3   3 124    37      81
4   4  97    10      59
5   5 135    90      88
6   6 105    80      70
7   7 117    88      78
8   8 101    73      73
9   9 110    80      84
10 10 127   100      86
> model <- lm(IQ.Books.Reading$Reading ~ IQ.Books.Reading$IQ + IQ.Books.Reading$Books)
> summary(model)

Call:
lm(formula = IQ.Books.Reading$Reading ~ IQ.Books.Reading$IQ +
    IQ.Books.Reading$Books)

Residuals:
   Min     1Q Median     3Q    Max
-5.230 -2.939 -1.188  2.856  7.862

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             6.47837   14.58373   0.444   0.6703
IQ.Books.Reading$IQ     0.53111    0.14093   3.769   0.0070 **
IQ.Books.Reading$Books  0.14046    0.05118   2.745   0.0287 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.624 on 7 degrees of freedom
Multiple R-squared:  0.8689,    Adjusted R-squared:  0.8315
F-statistic:  23.2 on 2 and 7 DF,  p-value: 0.0008151

> |
```

41

41

Multiple Regression Equation

$\boxed{x = a + by + cz}$

a = 6.478
b = .531
c = .140

42

42

Multiple Regression

The multiple regression equation

"READING = 6.478 + 0.531 × IQ + .140 x Books Available"

can be obtained by equating the dependent variable **Reading** to the *CONSTANT* value 6.478, plus the independent variables **IQ** multiplied by its coefficient 0.531 and the number of **Books Available** multiplied by .140.

43

43

IQ/Books/Sex : Reading

	A	B	C	D	E	F
1	ID	IQ	Books	Sex	Reading	
2	1	98	5	1	54	
3	2	115	25	2	74	
4	3	124	37	2	81	
5	4	97	10	1	59	
6	5	135	90	1	88	
7	6	105	80	1	70	
8	7	117	88	2	78	
9	8	101	73	2	73	
10	9	110	80	1	84	
11	10	127	100	2	86	
12						
13						

IQ-Books-Sex-Reading.csv

44

44

Import the Data into R

```
R Console
>
>
>
> IQ.Books.Sex.Reading <- read.csv("IQ-Books-Sex-Reading.csv", header = TRUE, colClasses = c("numeric", "numeric", "numeric", "numeric", "numeric"))
>
> |
```

IQ.Books.Sex.Reading <- read.csv("IQ-Books-Sex-Reading.csv", header = TRUE, colClasses = c("numeric", "numeric", "numeric", "numeric“, “numeric”))

```
R Console
>
>
>
>
> IQ.Books.Sex.Reading
   ID  IQ Books Sex Reading
1   1  98     5   1      54
2   2 115    25   2      74
3   3 124    37   2      81
4   4  97    10   1      59
5   5 135    90   1      88
6   6 105    80   1      70
7   7 117    88   2      78
8   8 101    73   2      73
9   9 110    80   1      84
10 10 127   100   2      86
> |
```

45

45

Regression Equation

model <- lm(IQ.Books.Sex.Reading$Reading ~ IQ.Books.Sex.Reading$IQ + IQ.Books.Sex.Reading$Sex + IQ.Books.Sex.Reading$Books)

```
R Console
> model <- lm(IQ.Books.Sex.Reading$Reading ~ IQ.Books.Sex.Reading$IQ + IQ.Books.Sex.Reading$Sex + IQ.Books.Sex.Reading$Books)
> summary(model)

Call:
lm(formula = IQ.Books.Sex.Reading$Reading ~ IQ.Books.Sex.Reading$IQ +
    IQ.Books.Sex.Reading$Sex + IQ.Books.Sex.Reading$Books)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6713 -2.7087 -0.6339  1.9184  8.6998

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                  6.3863    15.3608   0.416   0.6921
IQ.Books.Sex.Reading$IQ      0.5079     0.1542   3.294   0.0165 *
IQ.Books.Sex.Reading$Sex     1.8091     3.2472   0.557   0.5976
IQ.Books.Sex.Reading$Books   0.1405     0.0539   2.606   0.0403 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.87 on 6 degrees of freedom
Multiple R-squared:  0.8754,	Adjusted R-squared:  0.8131
F-statistic: 14.05 on 3 and 6 DF,  p-value: 0.004031

> |
```

46

46

Multiple Regression Equation

```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                  6.3863    15.3608   0.416   0.6921
IQ.Books.Sex.Reading$IQ      0.5079     0.1542   3.294   0.0165 *
IQ.Books.Sex.Reading$Sex     1.8091     3.2472   0.557   0.5976
IQ.Books.Sex.Reading$Books   0.1405     0.0539   2.606   0.0403 *
---
```

Reading = 6.39 + .5 x IQ + .14 x Books + 1.8 x Sex

Note that Sex is a categorical variable with two levels (Males = 1 & Females = 2).

In this case the regression weighting (1.8) is added to the predicted Reading value. So for females 1.8 (1.8 x 1) would be added and for males 3.6 (1.8 x 2) would be added.

47

47

Multiple Regression: Relative Importance

Which variables are important?

Install and load the "QuantPsyc" package, and use the function "lm.beta" on your model.

```
R Console
> lm.beta(model)
  IQ.Books.Sex.Reading$IQ   IQ.Books.Sex.Reading$Sex IQ.Books.Sex.Reading$Books
               0.58571381                 0.08464255                 0.44603647
>
```

You now have the standardised regression coefficients, which you can now compare.

IQ: 0.58571381

Sex: 0.08464255

Books Available: 0.44603647

And you can see that the most important predictor is **IQ**, followed closely by **Books Available**. **Sex** is not contributing very much.

48

48

Describing Data using Statistics

Description

Inference

– Exploring Relationships
– Comparing Groups

"When an individual uses descriptive statistics, he talks about the data he has; but with inferential statistics, he talks about data he does not have."

(Popham & Sirotnik, 1973: 40)

Comparing Groups (2 groups)

An experiment is carried out to see if announcing the items on a conveyor belt as they pass helps children remember the items better. Twenty-two children, all aged 12 years, were randomly allocated; 10 to the announcement group, and the remaining 12 to the no-announcement group. The children were shown the items on the conveyor belt with or without announcing each item as they passed. The number of items recalled out of each of the ten letters in the list was recorded for each child:

Data

ID	Method	Recall
1	Announcement	5
2	Announcement	7
3	Announcement	8
4	Announcement	9
5	Announcement	5
6	Announcement	4
7	Announcement	9
8	Announcement	8
9	Announcement	7
10	Announcement	4
11	No Announcement	3
12	No Announcement	8
13	No Announcement	5
14	No Announcement	6
15	No Announcement	3
16	No Announcement	5
17	No Announcement	9
18	No Announcement	3
19	No Announcement	3
20	No Announcement	4
21	No Announcement	7
22	No Announcement	4

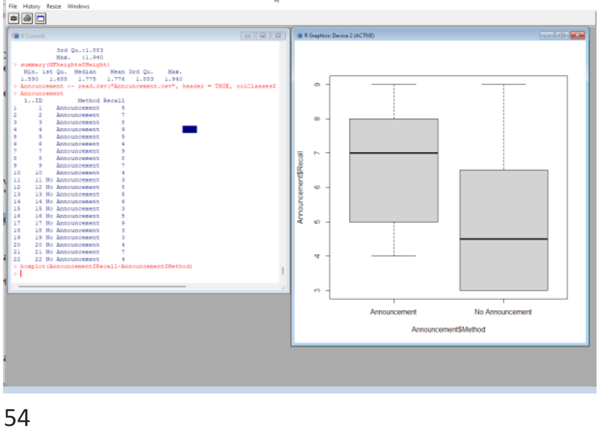
Import Data

```
lm(formula = IQ.Reading$IQ ~ IQ.Reading$Reading)

Coefficients:
      (Intercept)  IQ.Reading$Reading
          39.4035              0.9839

> Announcement <- read.csv("Announcement.csv", header = TRUE, colClasses = c("nu
> Announcement
   X..ID          Method Recall
1      1    Announcement      5
2      2    Announcement      7
3      3    Announcement      8
4      4    Announcement      9
5      5    Announcement      5
6      6    Announcement      4
7      7    Announcement      9
8      8    Announcement      8
9      9    Announcement      7
10    10    Announcement      4
11    11 No Announcement      3
12    12 No Announcement      8
13    13 No Announcement      5
14    14 No Announcement      6
15    15 No Announcement      3
16    16 No Announcement      5
17    17 No Announcement      9
18    18 No Announcement      3
19    19 No Announcement      3
20    20 No Announcement      4
21    21 No Announcement      7
22    22 No Announcement      4
>
```

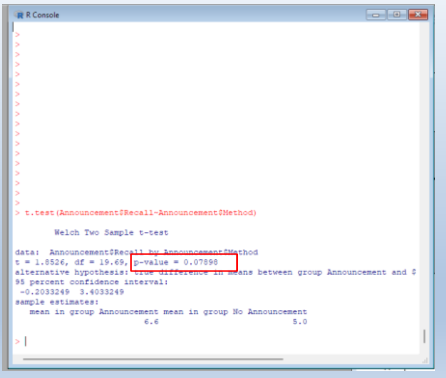

Summary

```
> summary(Announcement)
     X..ID                   Method       Recall
 Min.   : 1.00   Announcement   :10   Min.   :3.000
 1st Qu.: 6.25   No Announcement:12   1st Qu.:4.000
 Median :11.50                        Median :5.000
 Mean   :11.50                        Mean   :5.727
 3rd Qu.:16.75                        3rd Qu.:7.750
 Max.   :22.00                        Max.   :9.000
>
```


Draw a Boxplot



t-test



55

Analysis

P = 0.08 > 0.05 Therefore not significant

The conclusion is that there is no benefit to the wider population from the use of announcements.

56

Comparing Groups (3 groups or more)

An experiment was carried out in which a sample of adults was selected and randomly allocated to four groups. All were asked to consume amounts of alcohol as follows:

Group 1: NO ALCOHOL
Group 2: ONE UNIT OF ALCOHOL
Group 3: TWO UNITS OF ALCOHOL
Group 4: THREE UNITS OF ALCOHOL

after which they were asked drive a car simulator and perform an emergency stop at 40 miles per hour. The distance to stopping the simulator in metres by each group member is shown below.

57

Comparing Groups (3 groups or more)

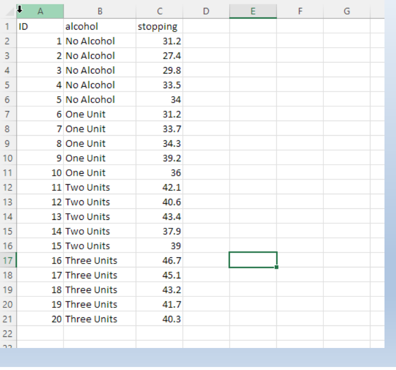
The results were as follows:

NO ALCOHOL	31.2	27.4	29.8	33.5	34.0
ONE UNIT	31.2	33.7	34.3	39.2	36.0
TWO UNITS	42.1	40.6	43.4	37.9	39.0
THREE UNITS	46.7	45.1	43.2	41.7	40.3

Stopping Distance in Metres

58

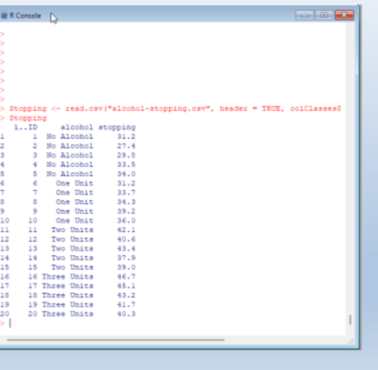
Data



	A	B	C	D	E	F	G
1	ID	alcohol	stopping				
2	1	No Alcohol	31.2				
3	2	No Alcohol	27.4				
4	3	No Alcohol	29.8				
5	4	No Alcohol	33.5				
6	5	No Alcohol	34				
7	6	One Unit	31.2				
8	7	One Unit	33.7				
9	8	One Unit	34.3				
10	9	One Unit	39.2				
11	10	One Unit	36				
12	11	Two Units	42.1				
13	12	Two Units	40.6				
14	13	Two Units	43.4				
15	14	Two Units	37.9				
16	15	Two Units	39				
17	16	Three Units	46.7				
18	17	Three Units	45.1				
19	18	Three Units	43.2				
20	19	Three Units	41.7				
21	20	Three Units	40.3				
22							

59

Import Data

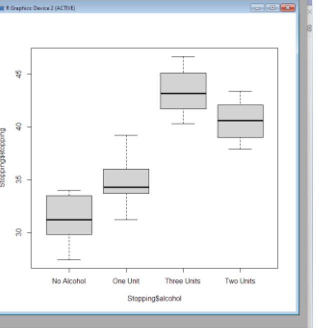


60

Compare Data

```
> Stopping <- read.csv("alcohol-stopping.csv", header=TRUE, colClass=c("numeric", "factor", "numeric"))
> Stopping
   ID units_of_alcohol stopping
1   1       No Alcohol     31.2
2   2       No Alcohol     27.4
3   3       No Alcohol     29.8
4   4       No Alcohol     33.5
5   5       No Alcohol     34.0
6   6         One Unit     31.2
7   7         One Unit     33.7
8   8         One Unit     34.3
9   9         One Unit     39.2
10 10         One Unit     36.0
11 11        Two Units     42.1
12 12        Two Units     40.6
13 13        Two Units     43.4
14 14        Two Units     37.9
15 15        Two Units     39.0
16 16      Three Units     46.7
17 17      Three Units     45.1
18 18      Three Units     43.2
19 19      Three Units     41.7
20 20      Three Units     40.3
>
>
>
>
>
>
>
>
>
> |
```


Compare Data



ANOVA

```
> Stopping
   1..ID      alcohol stopping
1      1   No Alcohol     31.2
2      2   No Alcohol     27.4
3      3   No Alcohol     29.8
4      4   No Alcohol     33.5
5      5   No Alcohol     34.0
6      6     One Unit     31.2
7      7     One Unit     33.7
8      8     One Unit     34.3
9      9     One Unit     39.2
10    10     One Unit     36.0
11    11    Two Units     42.1
12    12    Two Units     40.6
13    13    Two Units     43.4
14    14    Two Units     37.9
15    15    Two Units     39.0
16    16  Three Units     46.7
17    17  Three Units     45.1
18    18  Three Units     43.2
19    19  Three Units     41.7
20    20  Three Units     40.3
>
> boxplot(Stopping$stopping~Stopping$alcohol)
> anova <- aov(Stopping$stopping ~ Stopping$alcohol)
> summary(anova)
                 Df Sum Sq Mean Sq F value   Pr(>F)
Stopping$alcohol  3  456.1  152.04   21.92 6.52e-06 ***
Residuals        16  111.0    6.94
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> |
```


Analysis

```
                 Df Sum Sq Mean Sq F value   Pr(>F)
Stopping$alcohol  3  456.1  152.04   21.92 6.52e-06 ***
Residuals        16  111.0    6.94
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p = < 0.05
Therefore significant

The conclusion is that there is a significant difference between the distance it takes to stop a car simulator according to the amount of alcohol consumed.

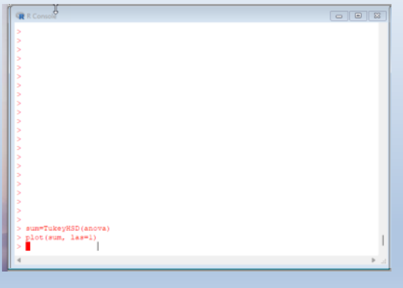
Multiple Comparisons

```
14    14   Two Units     37.9
15    15   Two Units     39.0
16    16 Three Units     46.7
17    17 Three Units     45.1
18    18 Three Units     43.2
19    19 Three Units     41.7
20    20 Three Units     40.3
>
> boxplot(Stopping$stopping~Stopping$alcohol)
> anova <- aov(Stopping$stopping ~ Stopping$alcohol)
> summary(anova)
                 Df Sum Sq Mean Sq F value   Pr(>F)
Stopping$alcohol  3  456.1  152.04   21.92 6.52e-06 ***
Residuals        16  111.0    6.94
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> TukeyHSD(anova)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Stopping$stopping ~ Stopping$alcohol)

$`Stopping$alcohol`
                         diff        lwr       upr     p adj
One Unit-No Alcohol      3.70 -1.0654663  8.465466 0.1596059
Three Units-No Alcohol  12.22  7.4545337 16.985466 0.0000091
Two Units-No Alcohol     9.42  4.6545337 14.185466 0.0001901
Three Units-One Unit     8.52  3.7545337 13.285466 0.0005430
Two Units-One Unit       5.72  0.9545337 10.485466 0.0161514
Two Units-Three Units   -2.80 -7.5654663  1.965466 0.3649771

> |
```

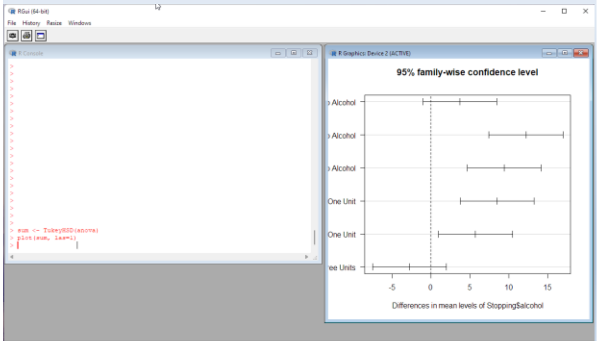

Multiple Comparisons

	diff	lwr	upr	p adj
One Unit-No Alcohol	3.70	-1.0654663	8.465466	0.1596059
Three Units-No Alcohol	12.22	7.4545337	16.985466	0.0000091
Two Units-No Alcohol	9.42	4.6545337	14.185466	0.0001901
Three Units-One Unit	8.52	3.7545337	13.285466	0.0005430
Two Units-One Unit	5.72	0.9545337	10.485466	0.0161514
Two Units-Three Units	-2.80	-7.5654663	1.965466	0.3649771

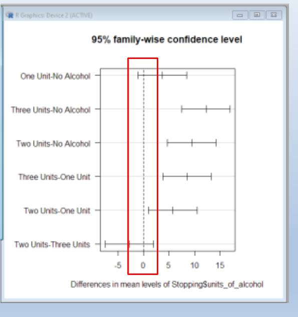
Tukey Plot



Tukey Plot



Tukey Plot



Comparing Groups

An experiment was carried out in which a sample of adults was selected and randomly allocated to four groups. All were asked to consume amounts of alcohol as follows:

Group 1: NO ALCOHOL
Group 2: ONE UNIT OF ALCOHOL
Group 3: TWO UNITS OF ALCOHOL
Group 4: THREE UNITS OF ALCOHOL

after which they were asked drive a car simulator and perform an emergency stop at 40 miles per hour. The distance to stopping the simulator in metres by each group member is shown below.

Is this the same for both males and females?

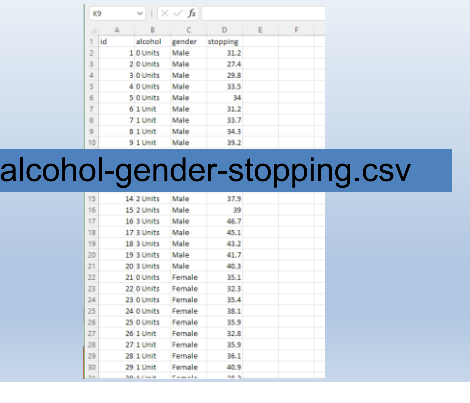
Comparing Groups (4 groups and M/F)

The results were as follows.

MALES	NO ALCOHOL	31.2	27.4	29.8	33.5	34.0
	ONE UNIT	31.2	33.7	34.3	39.2	36.0
	TWO UNITS	42.1	40.6	43.4	37.9	39.0
	THREE UNITS	46.7	45.1	43.2	41.7	40.3
FEMALES	NO ALCOHOL	35.1	32.3	35.4	38.1	35.9
	ONE UNIT	32.8	35.9	36.1	40.9	38.2
	TWO UNITS	38.1	35.9	40.2	39.8	37.6
	THREE UNITS	39.8	31.7	41.2	39.1	37.9

Stopping Distance in Metres

Comparing Groups (4 groups and M/F)



Three Questions

1. Is there a difference in the stopping distance according to the amount of alcohol consumed?
2. Is there a difference in the stopping distance between male and females?
3. Is the effect of amount of alcohol consumed on stopping distance different for males and females?

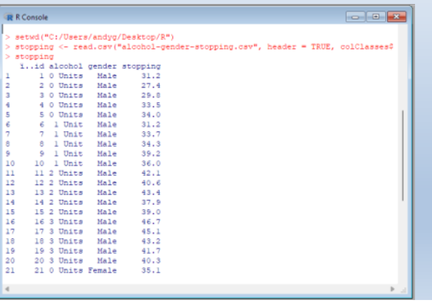
1 & 2 = main effect
3 = interaction

73

73

Import Data

> stopping <- read.csv("alcohol-gender-stopping.csv", header = TRUE, colClasses = c("numeric", "factor", "factor", "numeric"))

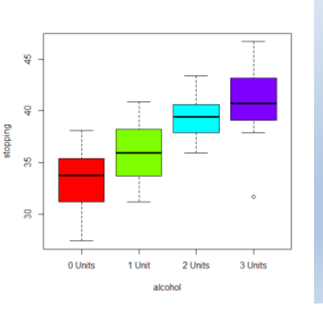


74

74

> boxplot(stopping~alcohol, data=stopping)

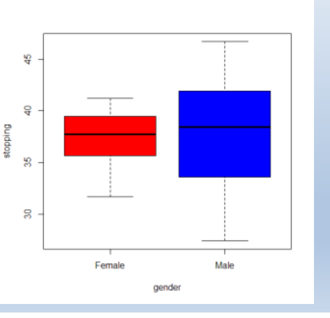
> boxplot(stopping~alcohol, data=stopping, col=rainbow(4))



75

75

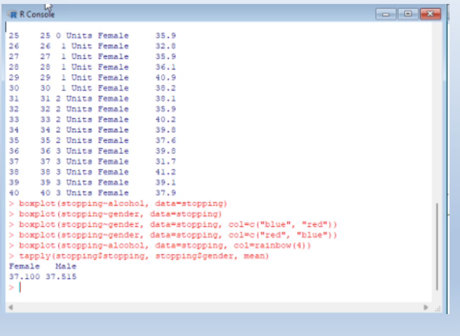
> boxplot(stopping~gender, data=stopping, col=c("red", "blue"))



76

76

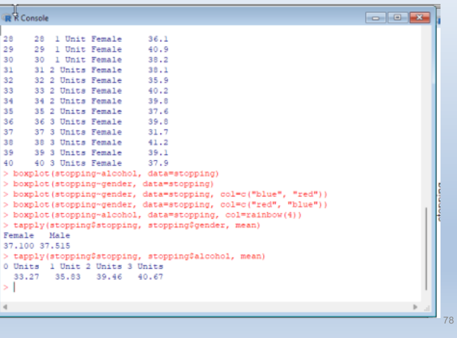
> tapply(stopping$stopping, stopping$gender, mean)



77

77

> tapply(stopping$stopping, stopping$alcohol, mean)



78

78

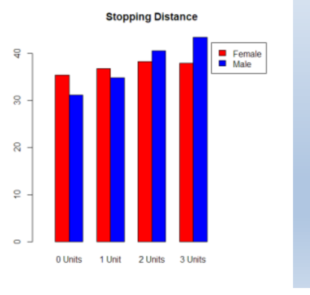
> tapply(stopping$stopping, list(stopping$gender, stopping$alcohol), mean)

```
R Console
32    32 2 Units Female     35.9
33    33 2 Units Female     40.2
34    34 2 Units Female     39.8
35    35 2 Units Female     37.6
36    36 3 Units Female     39.8
37    37 3 Units Female     31.7
38    38 3 Units Female     41.2
39    39 3 Units Female     39.1
40    40 3 Units Female     37.9
> boxplot(stopping~alcohol, data=stopping)
> boxplot(stopping~gender, data=stopping)
> boxplot(stopping~gender, data=stopping, col=c("blue", "red"))
> boxplot(stopping~gender, data=stopping, col=c("red", "blue"))
> boxplot(stopping~alcohol, data=stopping, col=rainbow(4))
> tapply(stopping$stopping, stopping$gender, mean)
Female   Male
37.100 37.515
> tapply(stopping$stopping, stopping$alcohol, mean)
0 Units  1 Unit 2 Units 3 Units
  33.27   35.83   39.46   40.67
> tapply(stopping$stopping, list(stopping$gender, stopping$alcohol), mean)
       0 Units 1 Unit 2 Units 3 Units
Female   35.36  36.78   38.32   37.94
Male     31.18  34.88   40.60   43.40
>
```

79

79

> count <- tapply(stopping$stopping, list(stopping$gender, stopping$alcohol), mean)
> barplot(count, main="Stopping Distance", beside=TRUE, xlim=c(0,16), legend=TRUE, col=c("red","blue"))



80

80

ANOVA

One-Way Between-Groups ANOVA
Single independent variable.

Two-Way Between-Groups ANOVA
Factorial/Multifactorial ANOVA

2 independent variables - two factors

81

81

2-Way ANOVA

> anova2 = aov(stopping ~ alcohol + gender + alcohol:gender, data = stopping)
> summary(anova2)

82

2-Way ANOVA

> anova2 = aov(stopping ~ alcohol + gender + alcohol:gender, data = stopping)
> summary(anova2)

```
R Console
[Previously saved workspace restored]

> setwd("C:/Users/andyg/Desktop/RData")
> stopping <- read.csv("alcohol-gender-stopping.csv", header = TRUE, colClasses$
> aov(stopping ~ alcohol + gender + alcohol:gender, data = stopping)
Call:
   aov(formula = stopping ~ alcohol + gender + alcohol:gender, data = stopping)

Terms:
                 alcohol   gender alcohol:gender Residuals
Sum of Squares  344.2407   1.7223       138.5087  230.7960
Deg. of Freedom        3        1              3        32

Residual standard error: 2.685587
Estimated effects may be unbalanced
> anova2 <- aov(stopping ~ alcohol + gender + alcohol:gender, data = stopping)
> summary(anova2)
               Df Sum Sq Mean Sq F value   Pr(>F)
alcohol         3  344.2  114.75  15.910 1.65e-06 ***
gender          1    1.7    1.72   0.239   0.6284
alcohol:gender  3  138.5   46.17   6.401   0.0016 **
Residuals      32  230.8    7.21
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

83

2-Way ANOVA

```
R Console
[Previously saved workspace restored]

> setwd("C:/Users/andyg/Desktop/RData")
> stopping <- read.csv("alcohol-gender-stopping.csv", header = TRUE, colClasses$
> aov(stopping ~ alcohol + gender + alcohol:gender, data = stopping)
Call:
   aov(formula = stopping ~ alcohol + gender + alcohol:gender, data = stopping)

Terms:
                 alcohol   gender alcohol:gender Residuals
Sum of Squares  344.2407   1.7223       138.5087  230.7960
Deg. of Freedom        3        1              3        32

Residual standard error: 2.685587
Estimated effects may be unbalanced
> anova2 <- aov(stopping ~ alcohol + gender + alcohol:gender, data = stopping)
> summary(anova2)
               Df Sum Sq Mean Sq F value   Pr(>F)
alcohol         3  344.2  114.75  15.910 1.65e-06 ***
gender          1    1.7    1.72   0.239   0.6284
alcohol:gender  3  138.5   46.17   6.401   0.0016 **
Residuals      32  230.8    7.21
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

84

2-Way ANOVA

R Console

```
37.100 37.515
> tapply(stopping$stopping, stopping$alcohol, mean)
0 Units  1 Unit 2 Units 3 Units
  33.27   35.83   39.46   40.67
> tapply(stopping$stopping, list(stopping$gender, stopping$alcohol), mean)
       0 Units 1 Unit 2 Units 3 Units
Female   35.36  36.78   38.32   37.94
Male     31.18  34.88   40.60   43.40
> count <- tapply(stopping$stopping, list(stopping$gender, stopping$alcohol), m$
>
> barplot(count, main="Stopping Distance", beside=TRUE, , xlim=c(0,15), legend=$
> barplot(count, main="Stopping Distance", beside=TRUE, , xlim=c(0,16), legend=$
> anova2 <- aov(stopping ~ alcohol + gender + alcohol:gender, data = stopping)
> summary(anova2)
Analysis of Variance Table

Response: stopping
               Df Sum Sq Mean Sq F value    Pr(>F)
alcohol         3 344.24 114.747 15.9097 1.652e-06 ***
gender          1   1.72   1.722  0.2388  0.628413
alcohol:gender  3 138.51  46.170  6.4014  0.001603 **
Residuals      32 230.80   7.212
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> |
```

85

R Console

```
37.100 37.515
> tapply(stopping$stopping, stopping$alcohol, mean)
0 Units  1 Unit 2 Units 3 Units
  33.27   35.83   39.46   40.67
> tapply(stopping$stopping, list(stopping$gender, stopping$alcohol), mean)
       0 Units 1 Unit 2 Units 3 Units
Female   35.36  36.78   38.32   37.94
Male     31.18  34.88   40.60   43.40
> count <- tapply(stopping$stopping, list(stopping$gender, stopping$alcohol), m$
>
> barplot(count, main="Stopping Distance", beside=TRUE, , xlim=c(0,15), legend=$
> barplot(count, main="Stopping Distance", beside=TRUE, , xlim=c(0,16), legend=$
> anova2 <- aov(stopping ~ alcohol + gender + alcohol:gender, data = stopping)
> summary(anova2)
Analysis of Variance Table

Response: stopping
               Df Sum Sq Mean Sq F value    Pr(>F)
alcohol         3 344.24 114.747 15.9097 1.652e-06 ***
gender          1   1.72   1.722  0.2388  0.628413
alcohol:gender  3 138.51  46.170  6.4014  0.001603 **
Residuals      32 230.80   7.212
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> |
```


alcohol = Amount of Alcohol Consumed – significant (p=0.0000 <0.05)

86

R Console

```
37.100 37.515
> tapply(stopping$stopping, stopping$alcohol, mean)
0 Units  1 Unit 2 Units 3 Units
  33.27   35.83   39.46   40.67
> tapply(stopping$stopping, list(stopping$gender, stopping$alcohol), mean)
       0 Units 1 Unit 2 Units 3 Units
Female   35.36  36.78   38.32   37.94
Male     31.18  34.88   40.60   43.40
> count <- tapply(stopping$stopping, list(stopping$gender, stopping$alcohol), m$
>
> barplot(count, main="Stopping Distance", beside=TRUE, , xlim=c(0,15), legend=$
> barplot(count, main="Stopping Distance", beside=TRUE, , xlim=c(0,16), legend=$
> anova2 <- aov(stopping ~ alcohol + gender + alcohol:gender, data = stopping)
> summary(anova2)
Analysis of Variance Table

Response: stopping
               Df Sum Sq Mean Sq F value    Pr(>F)
alcohol         3 344.24 114.747 15.9097 1.652e-06 ***
gender          1   1.72   1.722  0.2388  0.628413
alcohol:gender  3 138.51  46.170  6.4014  0.001603 **
Residuals      32 230.80   7.212
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> |
```

alcohol = Amount of Alcohol Consumed – significant (p=0.0000 <0.05)

gender = M/F – not significant (p=0.63)

77

87

R Console

```
37.100 37.515
> tapply(stopping$stopping, stopping$alcohol, mean)
0 Units  1 Unit 2 Units 3 Units
  33.27   35.83   39.46   40.67
> tapply(stopping$stopping, list(stopping$gender, stopping$alcohol), mean)
       0 Units 1 Unit 2 Units 3 Units
Female   35.36  36.78   38.32   37.94
Male     31.18  34.88   40.60   43.40
> count <- tapply(stopping$stopping, list(stopping$gender, stopping$alcohol), m$
>
> barplot(count, main="Stopping Distance", beside=TRUE, , xlim=c(0,15), legend=$
> barplot(count, main="Stopping Distance", beside=TRUE, , xlim=c(0,16), legend=$
> anova2 <- aov(stopping ~ alcohol + gender + alcohol:gender, data = stopping)
> summary(anova2)
Analysis of Variance Table

Response: stopping
               Df Sum Sq Mean Sq F value    Pr(>F)
alcohol         3 344.24 114.747 15.9097 1.652e-06 ***
gender          1   1.72   1.722  0.2388  0.628413
alcohol:gender  3 138.51  46.170  6.4014  0.001603 **
Residuals      32 230.80   7.212
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> |
```


88

R Console

```
37.100 37.515
> tapply(stopping$stopping, stopping$alcohol, mean)
0 Units  1 Unit 2 Units 3 Units
  33.27   35.83   39.46   40.67
> tapply(stopping$stopping, list(stopping$gender, stopping$alcohol), mean)
       0 Units 1 Unit 2 Units 3 Units
Female   35.36  36.78   38.32   37.94
Male     31.18  34.88   40.60   43.40
> count <- tapply(stopping$stopping, list(stopping$gender, stopping$alcohol), m$
>
> barplot(count, main="Stopping Distance", beside=TRUE, , xlim=c(0,15), legend=$
> barplot(count, main="Stopping Distance", beside=TRUE, , xlim=c(0,16), legend=$
> model = lm(stopping ~ alcohol + gender + alcohol:gender, data = stopping)
> anova(model)
Analysis of Variance Table

Response: stopping
               Df Sum Sq Mean Sq F value    Pr(>F)
alcohol         3 344.24 114.747 15.9097 1.652e-06 ***
gender          1   1.72   1.722  0.2388  0.628413
alcohol:gender  3 138.51  46.170  6.4014  0.001603 **
Residuals      32 230.80   7.212
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> |
```

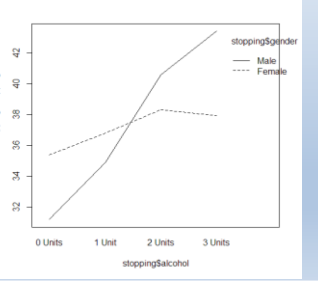
alcohol:gender - Interaction effect = whether the influence of alcohol on stopping distance depends on whether you are male or female – significant (p=0.0016)

89

89

Interaction Plot

> interaction.plot(stopping$alcohol, stopping$gender, stopping$stopping)

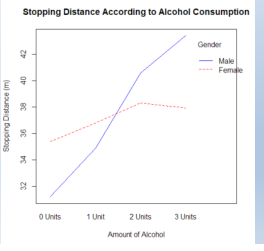


90

90

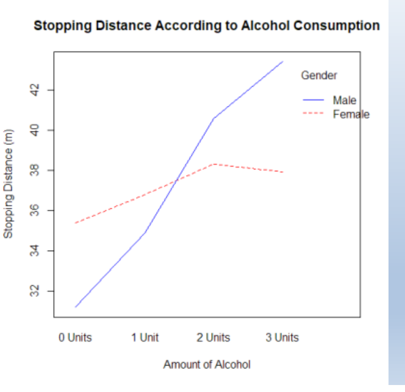
Interaction Plot

> interaction.plot(stopping$alcohol, stopping$gender, stopping$stopping, col=c("red", "blue"), trace.label="Gender", xlab="Amount of Alcohol", ylab="Stopping Distance (m)", main= "Stopping Distance According to Alcohol Consumption")



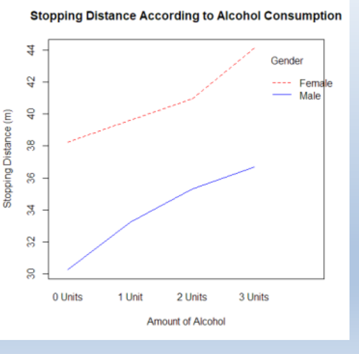
91

Interaction Plot



92

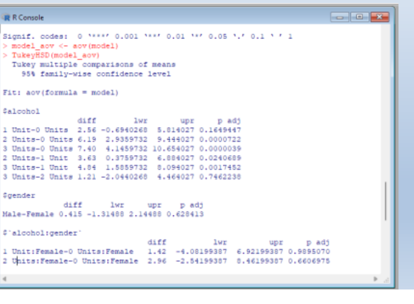
No Interaction



93

Tukey HSD

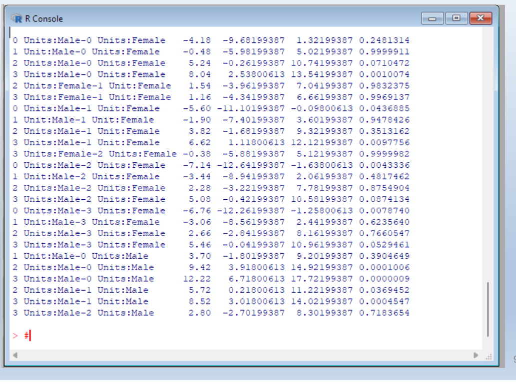
> model_aov <- aov(model)
> TukeyHSD(model_aov)



```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> model_aov <- aov(model)
> TukeyHSD(model_aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = model)

$alcohol
                diff        lwr       upr     p adj
1 Unit-0 Units  2.56 -0.6940268  5.814027 0.1649447
2 Units-0 Units 6.19  2.9359732  9.444027 0.0000722
3 Units-0 Units 7.40  4.1459732 10.654027 0.0000039
2 Units-1 Unit  3.63  0.3759732  6.884027 0.0240689
3 Units-1 Unit  4.84  1.5859732  8.094027 0.0017452
3 Units-2 Units 1.21 -2.0440268  4.464027 0.7462238

$gender
             diff      lwr     upr    p adj
Male-Female 0.415 -1.31488 2.14488 0.628413

$`alcohol:gender`
                              diff         lwr         upr     p adj
1 Unit:Female-0 Units:Female   1.42  -4.08199387  6.92199387 0.9895070
2 Units:Female-0 Units:Female  2.96  -2.54199387  8.46199387 0.6606975
```


94

Tukey HSD



```
0 Units:Male-0 Units:Female   -4.18  -9.68199387  1.32199387 0.2481314
1 Unit:Male-0 Units:Female    -0.48  -5.98199387  5.02199387 0.9999911
2 Units:Male-0 Units:Female    5.24  -0.26199387 10.74199387 0.0710472
3 Units:Male-0 Units:Female    8.04   2.53800613 13.54199387 0.0010074
2 Units:Female-1 Unit:Female   1.54  -3.96199387  7.04199387 0.9832375
3 Units:Female-1 Unit:Female   1.16  -4.34199387  6.66199387 0.9969137
0 Units:Male-1 Unit:Female    -5.60 -11.10199387 -0.09800613 0.0436885
1 Unit:Male-1 Unit:Female     -1.90  -7.40199387  3.60199387 0.9478426
2 Units:Male-1 Unit:Female     3.82  -1.68199387  9.32199387 0.3513162
3 Units:Male-1 Unit:Female     6.62   1.11800613 12.12199387 0.0097756
3 Units:Female-2 Units:Female -0.38  -5.88199387  5.12199387 0.9999982
0 Units:Male-2 Units:Female   -7.14 -12.64199387 -1.63800613 0.0043336
1 Unit:Male-2 Units:Female    -3.44  -8.94199387  2.06199387 0.4817462
2 Units:Male-2 Units:Female    2.28  -3.22199387  7.78199387 0.8754904
3 Units:Male-2 Units:Female    5.08  -0.42199387 10.58199387 0.0874134
0 Units:Male-3 Units:Female   -6.76 -12.26199387 -1.25800613 0.0078740
1 Unit:Male-3 Units:Female    -3.06  -8.56199387  2.44199387 0.6235640
2 Units:Male-3 Units:Female    2.66  -2.84199387  8.16199387 0.7660547
3 Units:Male-3 Units:Female    5.46  -0.04199387 10.96199387 0.0529461
1 Unit:Male-0 Units:Male       3.70  -1.80199387  9.20199387 0.3904649
2 Units:Male-0 Units:Male      9.42   3.91800613 14.92199387 0.0001006
3 Units:Male-0 Units:Male     12.22   6.71800613 17.72199387 0.0000009
2 Units:Male-1 Unit:Male       5.72   0.21800613 11.22199387 0.0369452
3 Units:Male-1 Unit:Male       8.52   3.01800613 14.02199387 0.0004547
3 Units:Male-2 Units:Male      2.80  -2.70199387  8.30199387 0.7183654

> #
```


95

Tukey HSD

3 Units:Male-1 Unit:Female	6.62	0.0097756
0 Units:Male-2 Units:Female	-7.14	0.0043336
0 Units:Male-3 Units:Female	-6.76	0.0078740
2 Units:Male-0 Units:Male	9.42	0.0001006
3 Units:Male-0 Units:Male	12.22	0.0000009
2 Units:Male-1 Unit:Male	5.72	0.0369452
3 Units:Male-1 Unit:Male	8.52	0.0004547

96

Interaction

The conclusion is that the effect of alcohol on car stopping distance was different for male subjects than for females.

A two-way between-groups analysis of variance was conducted to explore the impact of gender and amount of alcohol consumed on a car stopping distance. There was a significant interaction effect between the amount of alcohol consumed and the gender of the person driving the car on the distance needed to stop the car $F(3, 32) = 6.4$, $p<0.05$).

Comparing Groups

An experiment was carried out in which a sample of adults was selected and randomly allocated to four groups. All were asked to consume amounts of alcohol as follows:

Group 1: NO ALCOHOL
Group 2: ONE UNIT OF ALCOHOL
Group 3: TWO UNITS OF ALCOHOL
Group 4: THREE UNITS OF ALCOHOL

after which they were asked drive a car simulator and perform an emergency stop at 40 miles per hour. The distance to stopping the simulator in metres by each group member is shown below.

Does age make a difference?

Analysis of Covariance

If three or more distinct groups – need to control for 3rd variable:

Use one-way analysis of covariance or ANCOVA.

Data for Analysis of Covariance

An experiment was carried out in which a sample of adults was selected and randomly allocated to four groups. All were asked to consume amounts of alcohol as follows:
Group 1: NO ALCOHOL
Group 2: ONE UNIT OF ALCOHOL
Group 3: TWO UNITS OF ALCOHOL
Group 4: THREE UNITS OF ALCOHOL
after which they were asked drive a car simulator and perform an emergency stop at 40 miles per hour.

However, we need to control for age.

Three Questions

1. Is there a difference in the stopping distance according to the amount of alcohol drunk?
2. Is there a difference in the stopping distance depending on age?
3. Does age affect the stopping distance according to the amount of alcohol drunk?

Data for Analysis of Covariance

The stopping distances (m) obtained by each group member is shown below.
The results were as follows:

No Alcohol	31.2	27.4	29.8	33.5	34.0
One Unit	31.2	33.7	34.3	39.2	36.0
Two Units	42.1	40.6	43.4	37.9	39.0
Three Units	46.7	45.1	43.2	41.7	40.3

	Age	Stopping Distance (m)
No Alcohol	25.00	31.20
No Alcohol	24.00	27.40
No Alcohol	23.50	29.80
No Alcohol	24.70	33.50
No Alcohol	24.20	34.00
One Unit	24.50	31.20
One Unit	26.00	33.70
One Unit	27.00	34.30
One Unit	22.00	39.20
One Unit	25.00	36.00
Two Units	24.30	42.10
Two Units	24.60	40.60
Two Units	28.00	43.40
Two Units	27.00	37.90
Three Units	26.00	46.70
Three Units	25.00	45.10
Three Units	23.00	43.20
Three Units	28.00	41.70
Three Units	31.00	40.30

103

Data for Analysis of Covariance

	id	alcohol	age	stopping
1	id	alcohol	age	stopping
2	1	0 Units	25	31.2
3	2	0 Units	24	27.4
4	3	0 Units	23.5	29.8
5	4	0 Units	24.7	33.5
6	5	0 Units	24.2	34
7	6	1 Unit	24.5	31.2
8	7	1 Unit	26	33.7
9	8	1 Unit	27	34.3
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
13	12	2 Units	24.6	40.6
14	13	2 Units	28	43.4
15	14	2 Units	27	37.9
16	15	2 Units	26	39
17	16	3 Units	23	46.7
18	17	3 Units	25	45.1
19	18	3 Units	23	43.2
20	19	3 Units	28	41.7
21	20	3 Units	31	40.3

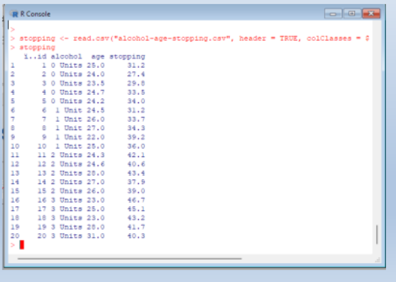
alcohol-age-stopping.csv

104

104

Import Data for Analysis of Covariance

> stopping <- read.csv("alcohol-age-stopping.csv", header = TRUE, colClasses = c("numeric","factor","numeric", "numeric"))

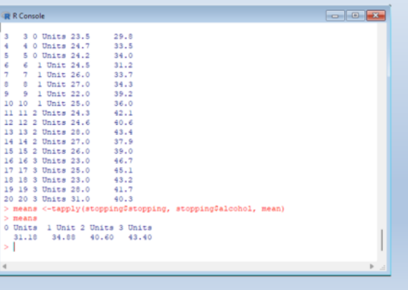


105

105

Data: Means

> means <-tapply(stopping$stopping, stopping$alcohol, mean)
> means

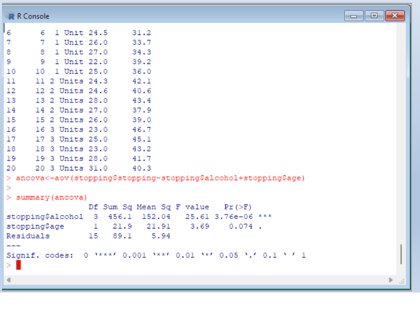


88

106

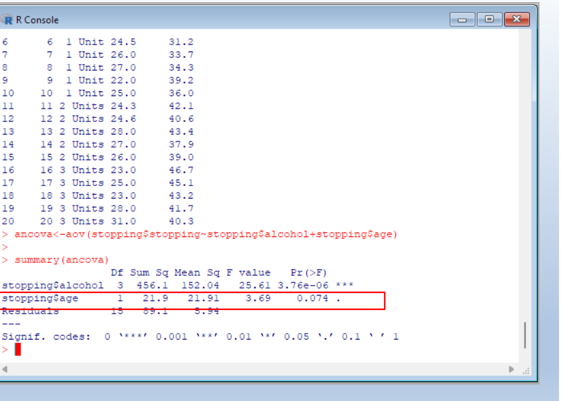
Analysis of Covariance

> ancova<-aov(stopping$stopping ~ stopping$alcohol+stopping$age)
> summary(ancova)



107

107



No Significant Effect For Age (p=.074)

108

108

```
R Console
6       6  1 Unit 24.5      31.2
7       7  1 Unit 26.0      33.7
8       8  1 Unit 27.0      34.3
9       9  1 Unit 22.0      39.2
10     10  1 Unit 25.0      36.0
11     11 2 Units 24.3      42.1
12     12 2 Units 24.6      40.6
13     13 2 Units 28.0      43.4
14     14 2 Units 27.0      37.9
15     15 2 Units 26.0      39.0
16     16 3 Units 23.0      46.7
17     17 3 Units 25.0      45.1
18     18 3 Units 23.0      43.2
19     19 3 Units 28.0      41.7
20     20 3 Units 31.0      40.3
> ancova<-aov(stopping$stopping~stopping$alcohol+stopping$age)
>
> summary(ancova)
                 Df Sum Sq Mean Sq F value   Pr(>F)
stopping$alcohol  3  456.1  152.04   25.61 3.76e-06 ***
stopping$age      1   21.9   21.91    3.69    0.074 .
Residuals        15   89.1    5.94
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Significant Effect For Amount of Alcohol (= <0.05)

109

TukeyHSD

> TukeyHSD(ancova,"stopping$alcohol")

```
R Console
> ancova<-aov(stopping$stopping ~ stopping$alcohol+stopping$age)
> TukeyHSD(ancova,"stopping$alcohol")
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = stopping$stopping ~ stopping$alcohol + stopping$age)

$`stopping$alcohol`
                 diff        lwr      upr     p adj
1 Unit-0 Units   3.70 -0.7417801  8.14178 0.1196368
2 Units-0 Units  9.42  4.9782199 13.86178 0.0001053
3 Units-0 Units 12.22  7.7782199 16.66178 0.0000052
2 Units-1 Unit   5.72  1.2782199 10.16178 0.0100402
3 Units-1 Unit   8.52  4.0782199 12.96178 0.0003028
3 Units-2 Units  2.80 -1.6417801  7.24178 0.3039307

Warning message:
In replications(paste("~", xx), data = mf) :
  non-factors ignored: stopping$age
>
```


110

Estimated Marginal Means

> library('emmeans’)
> emmeans(ancova, 'alcohol')

```
R Console
                 diff        lwr      upr     p adj
1 Unit-0 Units   3.70 -0.7417801  8.14178 0.1196368
2 Units-0 Units  9.42  4.9782199 13.86178 0.0001053
3 Units-0 Units 12.22  7.7782199 16.66178 0.0000052
2 Units-1 Unit   5.72  1.2782199 10.16178 0.0100402
3 Units-1 Unit   8.52  4.0782199 12.96178 0.0003028
3 Units-2 Units  2.80 -1.6417801  7.24178 0.3039307

Warning message:
In replications(paste("~", xx), data = mf) :
  non-factors ignored: stopping$age
> emmeans(ancova, 'alcohol')
Error in emmeans(ancova, "alcohol") : could not find function "emmeans"
> library('emmeans')
> emmeans(ancova, 'alcohol')
 alcohol emmean   SE df lower.CL upper.CL
 0 Units   30.6 1.13 15     28.2     33.0
 1 Unit    34.7 1.10 15     32.3     37.0
 2 Units   41.0 1.11 15     38.6     43.3
 3 Units   43.8 1.11 15     41.4     46.2

Confidence level used: 0.95
>
```


111

Presenting the Results from One-Way ANCOVA

A one-way between-groups analysis of covariance was conducted to compare the effect of amount of alcohol drunk on stopping distance on a group of adult drivers. The independent variable was the amount of alcohol drunk (none, 1 unit, 2 units, 3 units) and the dependent variable consisted of the stopping distance after an emergency stop in the car at 40 miles per hour. Participants’ age was used as the covariate in this analysis.
After adjusting for age, a significant difference (F(3, 15) = 25.61, p<.001) and was found between the amount of alcohol drunk and the stopping distance.

112

Comparison of Groups

A sample of male and female adults was selected and randomly allocated to four teaching groups using different teaching methods, after which they were given a test.

113

Two-Way ANCOVA

Method	Gender	Pre-Test Score	Final Test Score
Method 1	Male	25.00	31.20
Method 1	Male	24.00	27.40
Method 1	Male	23.50	29.80
Method 1	Male	24.70	33.50
Method 1	Male	24.20	34.00
Method 2	Male	24.50	31.20
Method 2	Male	26.00	33.70
Method 2	Male	27.00	34.30
Method 2	Male	22.00	39.20

114

Method	Gender	Pre-Test Score	Final Test Score
Method 1	Male	25.00	31.20
Method 1	Male	24.00	27.40
Method 1	Male	23.50	29.80
Method 1	Male	24.70	33.50
Method 1	Male	24.20	34.00
Method 2	Male	24.50	31.20
Method 2	Male	26.00	33.70
Method 2	Male	27.00	34.30
Method 2	Male	22.00	39.20
Method 2	Male	25.00	36.00
Method 3	Male	24.30	42.10
Method 3	Male	24.60	40.60
Method 3	Male	28.00	43.40
Method 3	Male	27.00	37.90
Method 3	Male	26.00	39.00
Method 4	Male	25.00	46.70
Method 4	Male	23.00	45.10
Method 4	Male	28.00	43.20
Method 4	Male	28.50	41.70
Method 4	Male	23.60	40.30
Method 1	Female	24.60	35.10
Method 1	Female	29.00	32.30
Method 1	Female	25.90	35.40
Method 1	Female	23.70	38.10
Method 1	Female	30.00	35.90
Method 2	Female	25.00	32.80
Method 2	Female	24.00	35.90
Method 2	Female	23.50	36.10
Method 2	Female	24.70	40.90
Method 2	Female	24.20	38.20
Method 3	Female	24.50	38.10
Method 3	Female	26.00	35.90
Method 3	Female	27.00	40.20
Method 3	Female	22.00	39.80
Method 3	Female	25.00	37.60
Method 4	Female	24.30	39.80
Method 4	Female	24.60	31.70
Method 4	Female	28.00	41.20
Method 4	Female	27.00	39.10
Method 4	Female	26.00	37.90

115

115

Comparison of Groups (t-test)

A sample of adults was selected and randomly allocated to four teaching groups using different teaching methods, after which they were given a test.

Males/Females

One continuous variable – Final Test Score

103

116

Comparison of Groups (t-test)

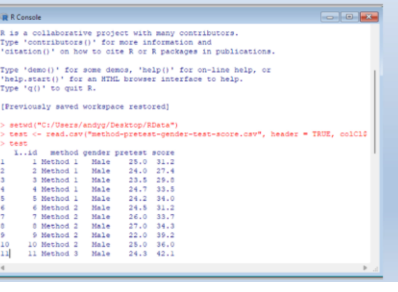
> test <- read.csv("method-pretest-gender-test-score.csv", header = TRUE, colClasses = c("numeric","factor","factor","numeric", "numeric"))

104

117

Comparison of Groups (t-test)

> test <- read.csv("method-pretest-gender-test-score.csv", header = TRUE, colClasses = c("numeric","factor","factor","numeric", "numeric"))



104

118

Comparison of Groups (t-test)

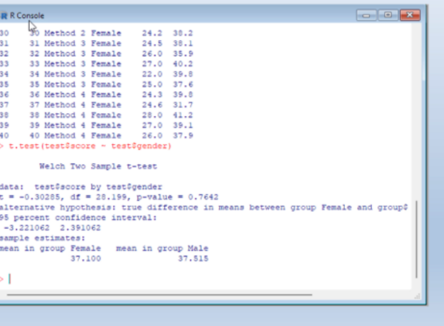
> t.test(test$score ~ test$gender)

105

119

Comparison of Groups (t-test)

> t.test(test$score ~ test$gender)



105

120

Data for Analysis of Variance (One-Way ANOVA)

A sample of adults was selected and randomly allocated to four teaching groups using different teaching methods, after which they were given a test.

Group 1: Method 1

Group 2: Method 2

Group 3: Method 3

Group 4: Method 4

One categorical variable – Method

One continuous variable – Final Test Score

121

121

Analysis of Variance (One-Way ANOVA)

```
> anova1<-aov(test$score ~ test$method)
> summary(anova1)
```

```
R Console
37    37 Method 4 Female    24.6  31.7
38    38 Method 4 Female    28.0  41.2
39    39 Method 4 Female    27.0  39.1
40    40 Method 4 Female    26.0  37.9
> t.test(test$score ~ test$gender)

        Welch Two Sample t-test

data:  test$score by test$gender
t = -0.30285, df = 28.199, p-value = 0.7642
alternative hypothesis: true difference in means between group Female and group$
95 percent confidence interval:
 -3.221062  2.391062
sample estimates:
mean in group Female   mean in group Male
              37.100               37.515

> anova1<-aov(test$score ~ test$method)
> summary(anova1)
            Df Sum Sq Mean Sq F value   Pr(>F)
test$method  3  344.2  114.75   11.13 2.58e-05 ***
Residuals   36  371.0   10.31
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> |
```

107

122

Analysis of Variance (One-Way ANOVA)

```
TukeyHSD(anova1)
```

```
R Console
              37.100               37.515

> anova1<-aov(test$score ~ test$method)
> summary(anova1)
            Df Sum Sq Mean Sq F value   Pr(>F)
test$method  3  344.2  114.75   11.13 2.58e-05 ***
Residuals   36  371.0   10.31
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> TukeyHSD(anova1)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = test$score ~ test$method)

$`test$method`
                  diff        lwr       upr     p adj
Method 2-Method 1 2.56 -1.3066912  6.426691 0.2978684
Method 3-Method 1 6.19  2.3233088 10.056691 0.0006686
Method 4-Method 1 7.40  3.5333088 11.266691 0.0000537
Method 3-Method 2 3.63 -0.2366912  7.496691 0.0723277
Method 4-Method 2 4.84  0.9733088  8.706691 0.0093145
Method 4-Method 3 1.21 -2.6566912  5.076691 0.8336921

> |
```

123

Data for Analysis of Variance (Two-Way ANOVA)

A sample of adults was selected and randomly allocated to four teaching groups using different teaching methods, after which they were given a test.

Group 1: Method 1

Group 2: Method 2 Males/Females

Group 3: Method 3

Group 4: Method 4

Two categorical variables – Method & Gender

One continuous variable – Final Test Score

124

124

Analysis of Variance (Two-Way ANOVA)

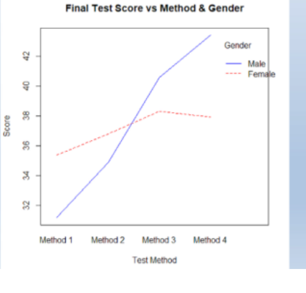
```
> anova2<-aov(test$score ~ test$method +
test$gender + test$method:test$gender)
> summary(anova2)
```

```
R Console
> TukeyHSD(anova1)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = test$score ~ test$method)

$`test$method`
                  diff        lwr       upr     p adj
Method 2-Method 1 2.56 -1.3066912  6.426691 0.2978684
Method 3-Method 1 6.19  2.3233088 10.056691 0.0006686
Method 4-Method 1 7.40  3.5333088 11.266691 0.0000537
Method 3-Method 2 3.63 -0.2366912  7.496691 0.0723277
Method 4-Method 2 4.84  0.9733088  8.706691 0.0093145
Method 4-Method 3 1.21 -2.6566912  5.076691 0.8336921

> anova2<-aov(test$score ~ test$method + test$gender + test$method:test$gender)
> summary(anova2)
                        Df Sum Sq Mean Sq F value   Pr(>F)
test$method              3  344.2  114.75  15.910 1.65e-06 ***
test$gender              1    1.7    1.72   0.239   0.6284
test$method:test$gender  3  138.5   46.17   6.401   0.0016 **
Residuals               32  230.8    7.21
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> |
```

125

Analysis of Variance (Two-Way ANOVA)

```
> interaction.plot(test$method, test$gender, test$score,
col=c("red", "blue"), trace.label="Gender", xlab="Test
Method", ylab="Score", main= "Final Test Score vs Method &
Gender")
```



111

126

Analysis of Variance (Two-Way ANOVA)

> TukeyHSD(anova2)

```
R Console
$`test$gender`
                diff      lwr     upr    p adj
Male-Female 0.415 -1.31488 2.14488 0.628413

$`test$method:test$gender`
                                 diff          lwr         upr     p adj
Method 2:Female-Method 1:Female  1.42  -4.08199387  6.92199387 0.9895070
Method 3:Female-Method 1:Female  2.96  -[illegible].54199387  8.46199387 0.6606975
Method 4:Female-Method 1:Female  2.58  -2.92199387  8.08199387 0.7917676
Method 1:Male-Method 1:Female   -4.18  -9.68199387  1.32199387 0.2481314
Method 2:Male-Method 1:Female   -0.48  -5.98199387  5.02199387 0.9999911
Method 3:Male-Method 1:Female    5.24  -0.26199387 10.74199387 0.0710472
Method 4:Male-Method 1:Female    8.04   2.53800613 13.54199387 0.0010074
Method 3:Female-Method 2:Female  1.54  -3.96199387  7.04199387 0.9832375
Method 4:Female-Method 2:Female  1.16  -4.34199387  6.66199387 0.9969137
Method 1:Male-Method 2:Female   -5.60 -11.10199387 -0.09800613 0.0436885
Method 2:Male-Method 2:Female   -1.90  -7.40199387  3.60199387 0.9478426
Method 3:Male-Method 2:Female    3.82  -1.68199387  9.32199387 0.3513162
Method 4:Male-Method 2:Female    6.62   1.11800613 12.12199387 0.0097756
Method 4:Female-Method 3:Female -0.38  -5.88199387  5.12199387 0.9999982
Method 1:Male-Method 3:Female   -7.14 -12.64199387 -1.63800613 0.0043336
Method 2:Male-Method 3:Female   -3.44  -8.94199387  2.06199387 0.4817462
Method 3:Male-Method 3:Female    2.28  -3.22199387  7.78199387 0.8754904
Method 4:Male-Method 3:Female    5.08  -0.42199387 10.58199387 0.0874134
```

127

Data for Analysis of Variance (One-Way ANCOVA)

A sample of adults was selected and randomly allocated to four teaching groups using different teaching methods, after which they were given a test.

Group 1: Method 1

Group 2: Method 2 — Pre-Test/Final Test

Group 3: Method 3

Group 4: Method 4

One categorical variable – Method

Two continuous variables – Pre-Test Score & Final Test Score

128

128

Analysis of Variance (One-Way ANCOVA)

> ancova1<-aov(test$score ~ test$method + test$pretest)

```
R Console
>
> ancova1<-aov(test$score ~ test$method + test$pretest)
> summary(ancova1)
             Df Sum Sq Mean Sq F value   Pr(>F)
test$method   3  344.2  114.75  10.828 3.52e-05 ***
test$pretest  1    0.1    0.13   0.012    0.914
Residuals    35  370.9   10.60
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

129

Analysis of Variance (One-Way ANCOVA)

> TukeyHSD(anova1)

```
R Console
>
> ancova1<-aov(test$score ~ test$method + test$pretest)
> summary(ancova1)
             Df Sum Sq Mean Sq F value   Pr(>F)
test$method   3  344.2  114.75  10.828 3.52e-05 ***
test$pretest  1    0.1    0.13   0.012    0.914
Residuals    35  370.9   10.60
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> TukeyHSD(anova1)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = test$score ~ test$method)

$`test$method`
                  diff        lwr       upr     p adj
Method 2-Method 1 2.56 -1.3066912  6.426691 0.2978684
Method 3-Method 1 6.19  2.3233088 10.056691 0.0006686
Method 4-Method 1 7.40  3.5333088 11.266691 0.0000537
Method 3-Method 2 3.63 -0.2366912  7.496691 0.0723277
Method 4-Method 2 4.84  0.9733088  8.706691 0.0093145
Method 4-Method 3 1.21 -2.6566912  5.076691 0.8336921

>
```

130

Data for Analysis of Variance (Two-Way ANCOVA)

A sample of adults was selected and randomly allocated to four teaching groups using different teaching methods, after which they were given a test.

Group 1: Method 1

Group 2: Method 2 — Pre-Test/Final Test

Group 3: Method 3 — Male/Female

Group 4: Method 4

Two categorical variables – Method & Gender

Two continuous variables – Pre-Test Score & Final Test Score

131

131

Analysis of Variance (Two-Way ANOVA)

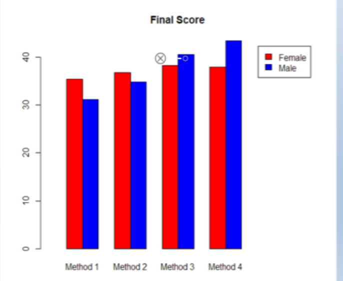
> tapply(test$score, list(test$gender, test$method), mean)

```
R Console
>
>
> tapply(test$score, list(test$gender, test$method), mean)
       Method 1 Method 2 Method 3 Method 4
Female    35.36    36.78    38.32    37.94
Male      31.18    34.88    40.60    43.40
>
```

132

Analysis of Variance (Two-Way ANOVA)

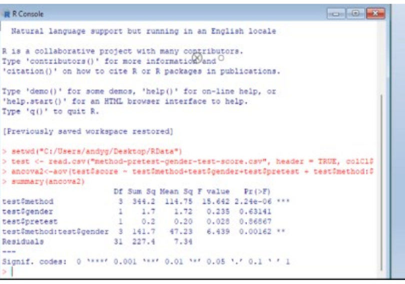
> plot <- tapply(test$score, list(test$gender, test$method), mean)
> barplot(plot, main="Final Score", beside=TRUE, xlim=c(0,16), legend=TRUE, col=c("red","blue"))



133

Analysis of Variance (Two-Way ANCOVA)

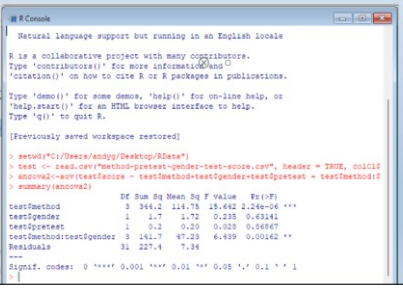
> test <- read.csv("method-pretest-gender-test-score.csv", header = TRUE, colClasses = c("numeric","factor","factor","numeric", "numeric"))



134

Analysis of Variance (Two-Way ANCOVA)

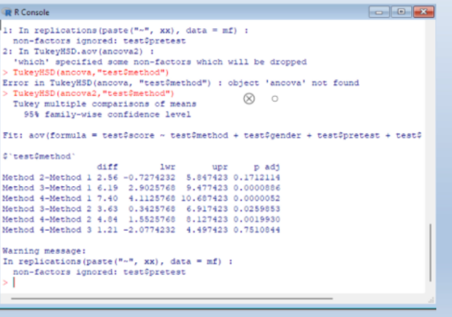
> ancova2<-aov(test$score ~ test$method+test$gender+test$pretest + test$method:test$gender)
> summary(ancova2)



135

Analysis of Variance (Two-Way ANCOVA)

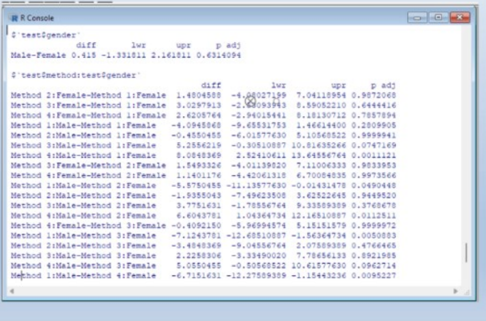
> TukeyHSD(ancova2,"test$method")



136

Analysis of Variance (Two-Way ANCOVA)

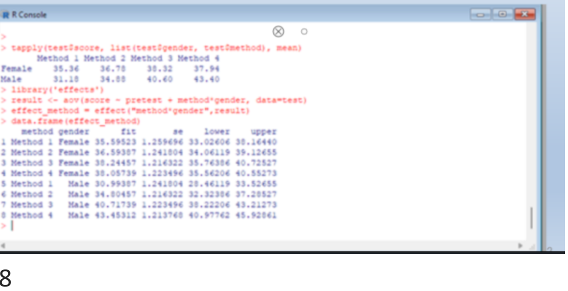
> TukeyHSD(ancova2)



137

Analysis of Variance (Two-Way ANCOVA)

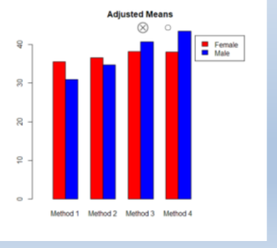
library("effects")
result <- aov(score ~ pretest + method*gender, data=test
effect_method = effect("method*gender",result)
ata.frame(effect_method)



138

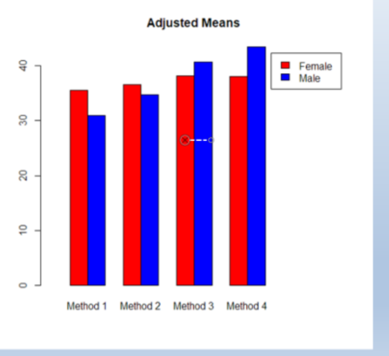
Analysis of Variance (Two-Way ANCOVA)

```
> means <- data.frame(effect_method)
> plot <- tapply(means$fit, list(means$gender,
means$method), mean)
> barplot(plot, main="Final Score, controlling for Pre-Test",
beside=TRUE, xlim=c(0,20), legend=TRUE)
```



139

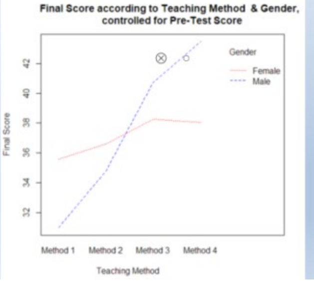
Analysis of Variance (Two-Way ANCOVA)



140

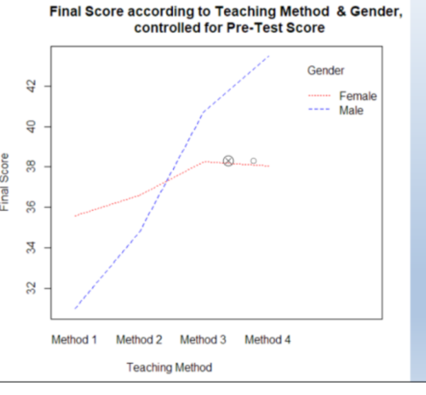
Analysis of Variance (Two-Way ANCOVA)

```
> interaction.plot(means$method, means$gender, means$fit,
col=c("red", "blue"), trace.label="Gender", xlab="Teaching
Method", ylab="Final Score", main= "Final Score according to
Teaching Method & Gender, \n controlled for Pre-Test Score")
```



141

Analysis of Variance (Two-Way ANCOVA)



142

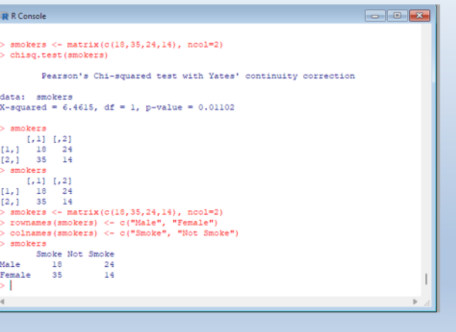
Presenting the Results from Two-Way ANCOVA

A 2 by 2 between-groups analysis of covariance was conducted to assess the effectiveness of four methods of teaching English for male and female participants. The independent variables were the teaching method (method 1, 2, 3 or 4) and gender. The dependent variable was scores on the final test, administered following completion of the intervention programs. Scores on the pre-test administered prior to the commencement of the programs were used as a covariate to control for individual differences.
After adjusting for pre-test scores, there was a significant interaction effect: F (1, 31) = 6.439, p < .001, with a large effect size (partial eta squared = .384). These results suggest that males and females respond differently to the four methods. Males showed a more substantial increase on the final test score after participation the class using methods 1 and 2 than methods 3 & 4, whereas females did not differ so much.

143

143

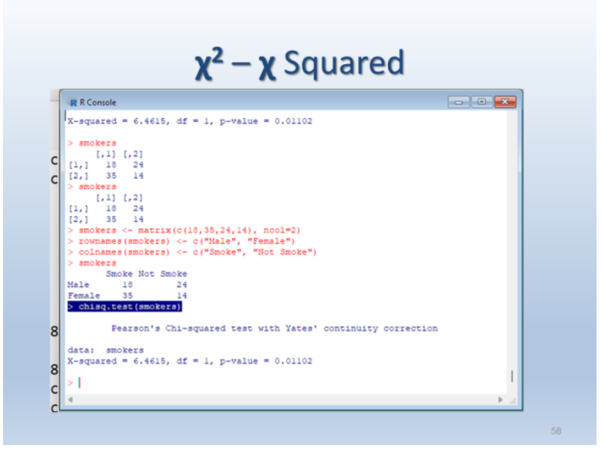
χ^2 – χ Squared


R Console

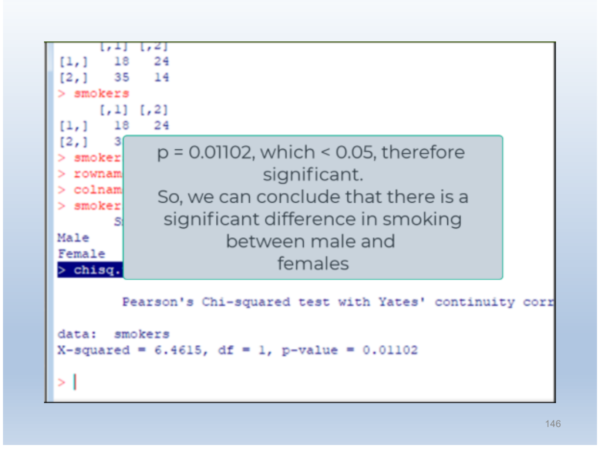
```
> smokers <- matrix(c(18,35,24,14), ncol=2)
> chisq.test(smokers)

        Pearson's Chi-squared test with Yates' continuity correction

data:  smokers
X-squared = 6.4615, df = 1, p-value = 0.01102

> smokers
     [,1] [,2]
[1,]   18   24
[2,]   35   14
> smokers
     [,1] [,2]
[1,]   18   24
[2,]   35   14
> smokers <- matrix(c(18,35,24,14), ncol=2)
> rownames(smokers) <- c("Male", "Female")
> colnames(smokers) <- c("Smoke", "Not Smoke")
> smokers
       Smoke Not Smoke
Male      18        24
Female    35        14
> 
```

57

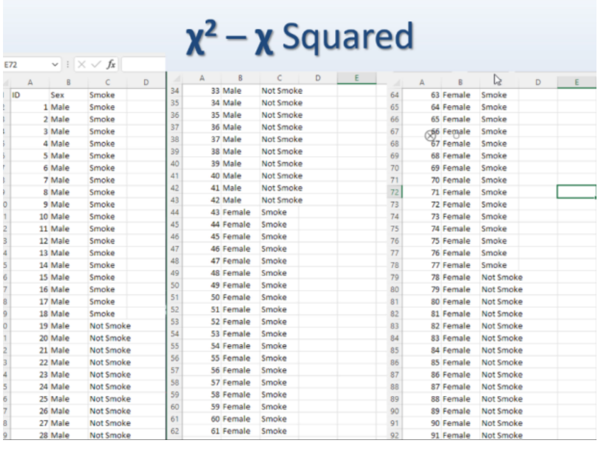
144



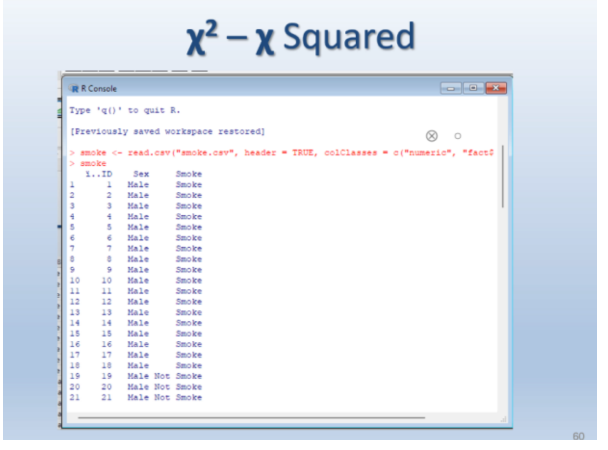
145



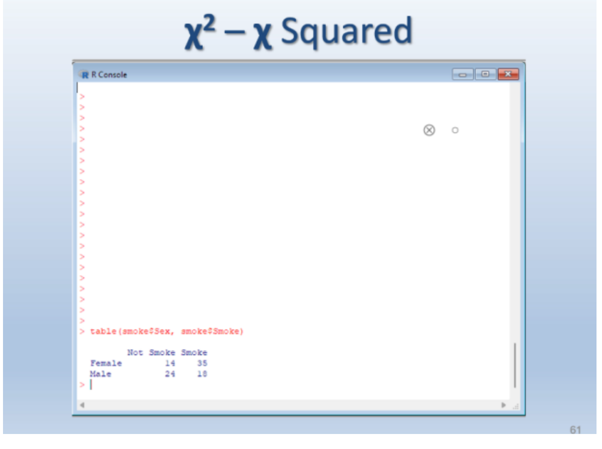
146



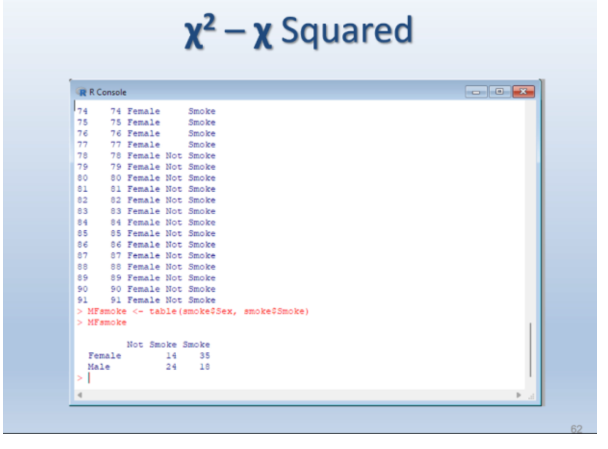
147



148



149



150

χ^2 – χ Squared

```
R Console
81     81 Female Not Smoke
82     82 Female Not Smoke
83     83 Female Not Smoke
84     84 Female Not Smoke
85     85 Female Not Smoke
86     86 Female Not Smoke
87     87 Female Not Smoke
88     88 Female Not Smoke
89     89 Female Not Smoke
90     90 Female Not Smoke
91     91 Female Not Smoke
> MFsmoke <- table(smoke$Sex, smoke$Smoke)
> MFsmoke

         Not Smoke Smoke
  Female        14    35
  Male          24    18
> chisq.test(MFsmoke)

        Pearson's Chi-squared test with Yates' continuity correction

data:  MFsmoke
X-squared = 6.4615, df = 1, p-value = 0.01102

> |
```

63

151

χ^2 – χ Squared

```
R Console
81     81 Female Not Smoke
82     82 Female Not Smoke
83     83 Female Not Smoke
84     84 Female Not Smoke
85     85 Female Not Smoke
86     86 Female Not Smoke
87     87 Female Not Smoke
88     88 Female Not Smoke
89     89 Female Not Smoke
90     90 Female Not Smoke
91     91 Female Not Smoke
> MFsmoke <- table(smoke$Sex, smoke$Smoke)
> MFsmoke

         Not Smoke Smoke
  Female        14    35
  Male          24    18
> chisq.test(MFsmoke)

        Pearson's Chi-squared test with Yates' continuity correction

data:  MFsmoke
X-squared = 6.4615, df = 1, p-value = 0.01102

> |
```

63

152

Describing Data using Statistics

Description

- Central Tendency
- Dispersion

Inference

– Exploring Relationships
  - Correlation
  - Regression

– Comparing Groups
  - t-Test
  - ANOVA
    - One-Way
    - Two-Way
    - ANCOVA

153

153

References & Further Reading

Brown, J. D. & Rogers, T. S. (2002). *Doing second language research*. Oxford: Oxford University Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* New York: Routledge

Dörnyei, Z. (2007). *Research methods in applied linguistics.* Oxford: Oxford University Press.

De Vries, A. & Meys, J. (2015). *R for dummies* (2nd ed.). London: Wiley.

De Winter, J. C. F. & Dodou, D. (2012). Five point Likert items: t-test versus Mann-Whitney-Wilcoxon. *Practical Assessment, Research and Evaluation, 15*(11).

Field, A., Miles, J. & Field, Z. (2013). *Discovering statistics using R.* London: Sage.

Hand, D. J. (2008) *Statistics: A very short introduction*. Oxford: Oxford University Press.

Hatch, E. & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. Boston, MA: Heinle & Heinle.

Kabacoff, R. I. (2015). *R in action* (2nd ed.). Shelter Island, NY: Manning.

Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R* (2nd ed.). London: Routledge.

Lowie, W. & Seton, B. (2013). *Essential statistics for applied linguistics*. London: Palgrave Macmillan.

Popham, W. J. & Sirotnik, K. A. (1973). *Educational statistics: Use and interpretation* (2nd ed.). New York: Harper & Row

Salkind, N. J. (2011). *Statistics for people who (think they) hate statistics* (4th ed.). London: Sage.

Saunders, M. & Lewis, P. (2012). *Doing research in business and management*. London: FT/Prentice Hall.

Saunders, M., Lewis, P. & Thornhill, A. (2016). *Research methods for business students* (7th ed.). London: FT/Prentice Hall.

Shavelson, R. (1988). *Statistical reasoning for the behavioral sciences*. Boston: Allyn and Bacon.

Speigelhalter, D. (2020). *The art of statistics: Learning from data.* London: Pelican.

154

154