DISCOVERING STATISTICS USING R

Basic statistics with R – part II

For Bits - April 2019

veronique.storme@psb.ugent.be

https://github.com/vstorme/basicstat

GHENT UNIVERSITY

VIB-UGENT CENTER FOR PLANT SYSTEMS BIOLOGY

outline

- Originally based on the book “Discovering statistics using R” by Andy Field http://www.uk.sagepub.com/dsur/main.htm
- Topics
  - Correlation
  - Ordinary least-squares regression
  - Comparing 2 means
  - Comparing several means
  - Factorial anova
  - Randomised block design
  - Non-parametric tests
  - Categorical data

Correlation

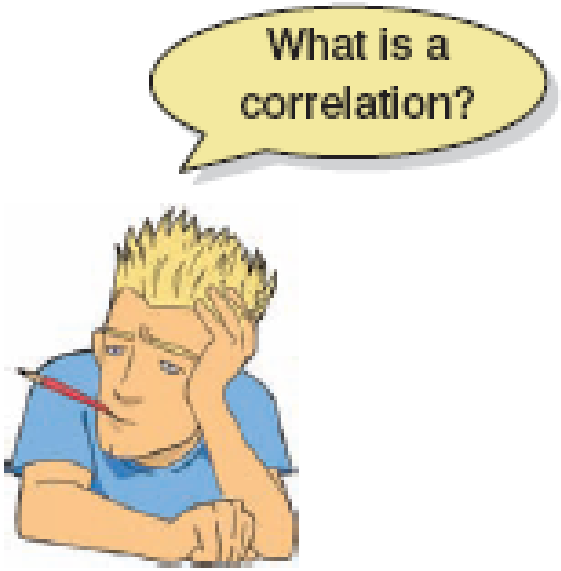
ANDY FIELD

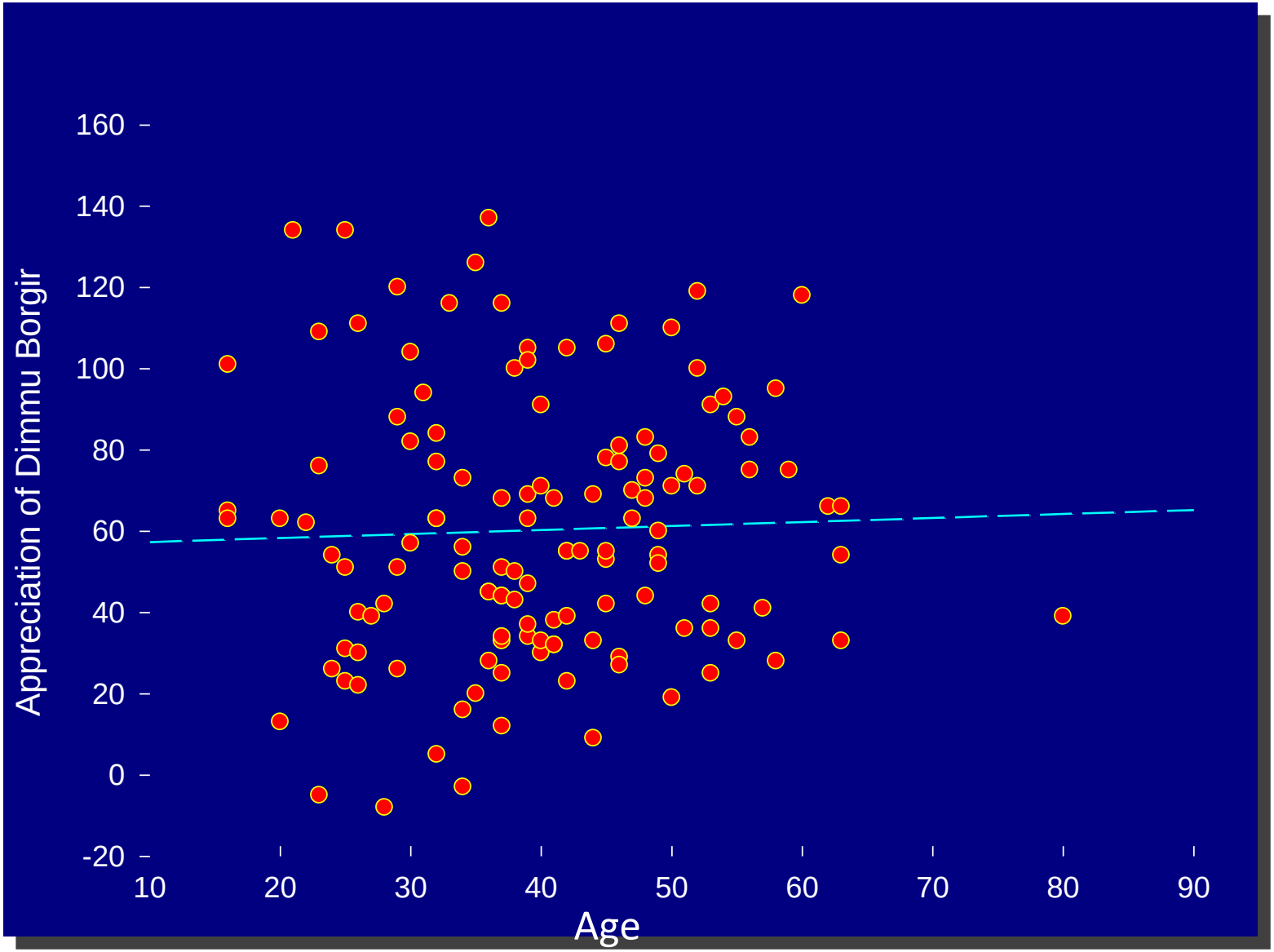
Aims

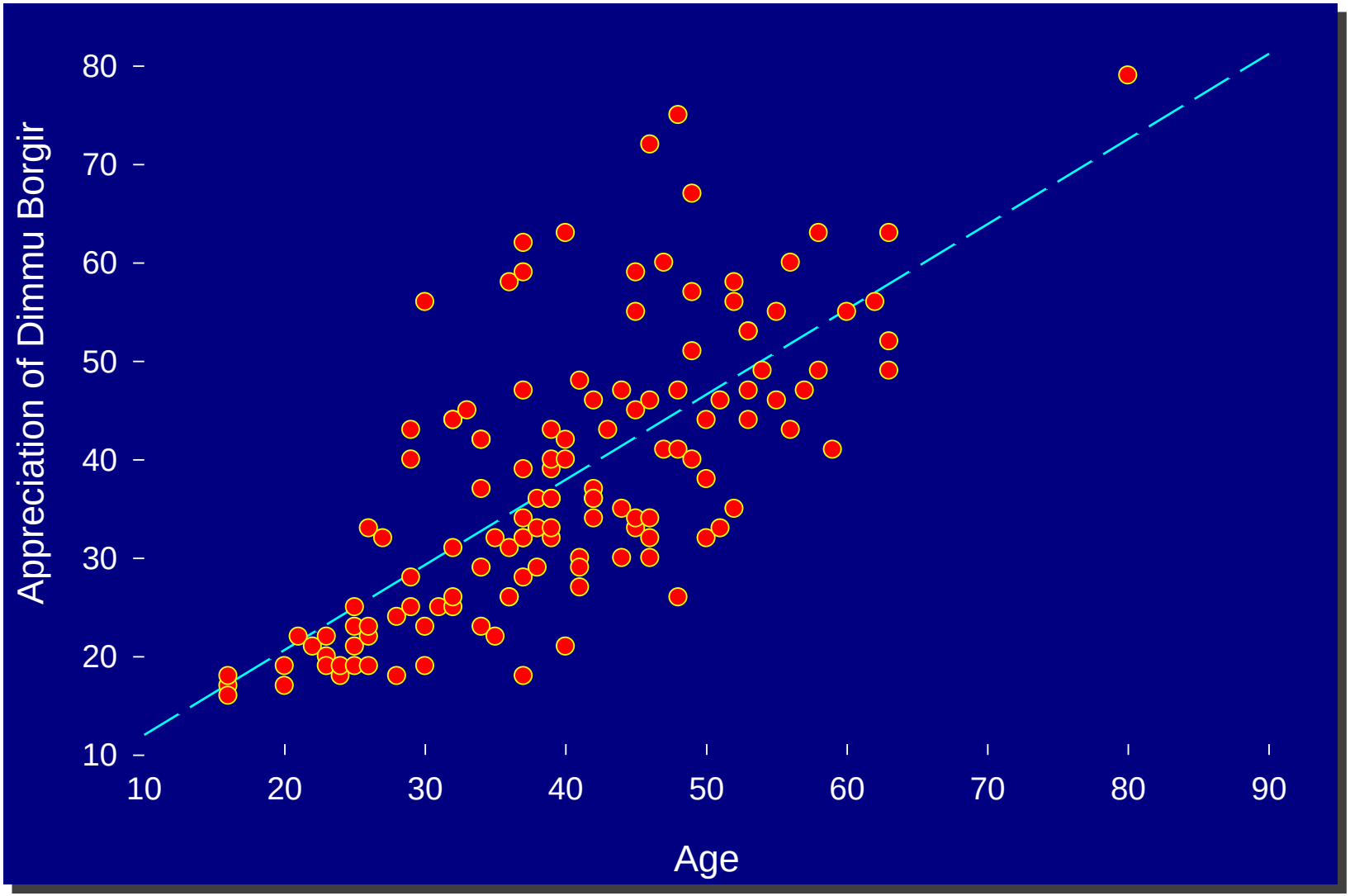
- Measuring relationships between paired data
  - Scatterplots
  - Covariance
  - Pearson's correlation coefficient
- Nonparametric measures
  - Spearman's rho
  - Kendall's tau
- Interpreting correlations
  - Causality

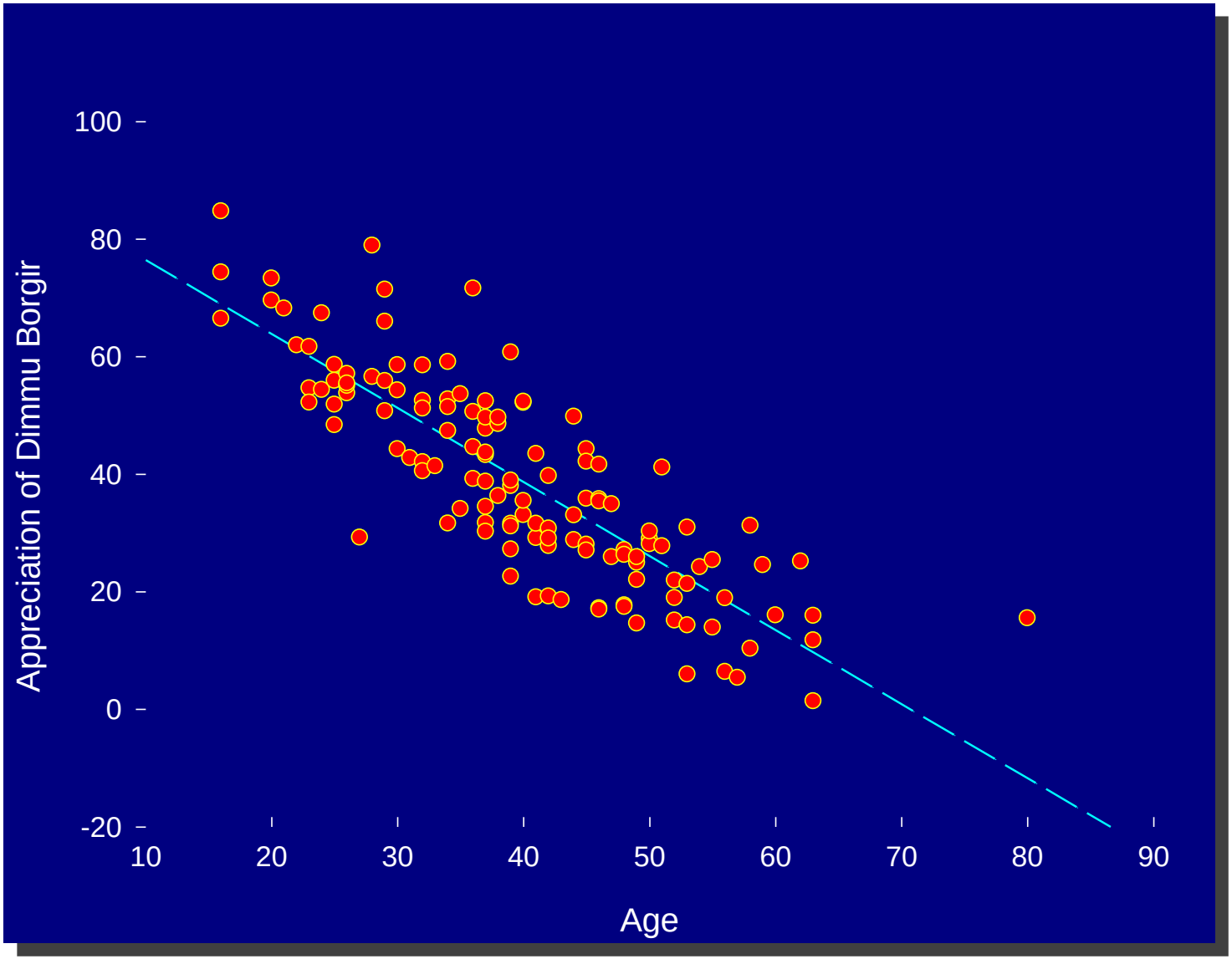
What is a Correlation?

- It is a way of measuring the extent to which two variables are related.
- It measures the pattern of responses across variables.




160
140
120
100
80
60
40
20
0
-20
Appreciation of Dimmu Borgir
10
20
30
40
50
60
70
80
90
Age


Appreciation of Dimmu Borgir
80
70
60
50
40
30
20
10
10
20
30
40
50
60
70
80
90
Age


Appreciation of Dimmu Borgir
100
80
60
40
20
0
-20
10
20
30
40
50
60
70
80
90
Age

Good statistical practice

- Descriptive (summary) statistics
- Graphical display
- Statistical model
- Testing the assumptions

Summary statistics

- {base}
  - > summary(data)
- {mosaic}
  - > favstats(~ var, data=df)
- use the apply-type function for dataframes to get summary of multiple variables
  - > dfapply(df, favstats, select = is.numeric)

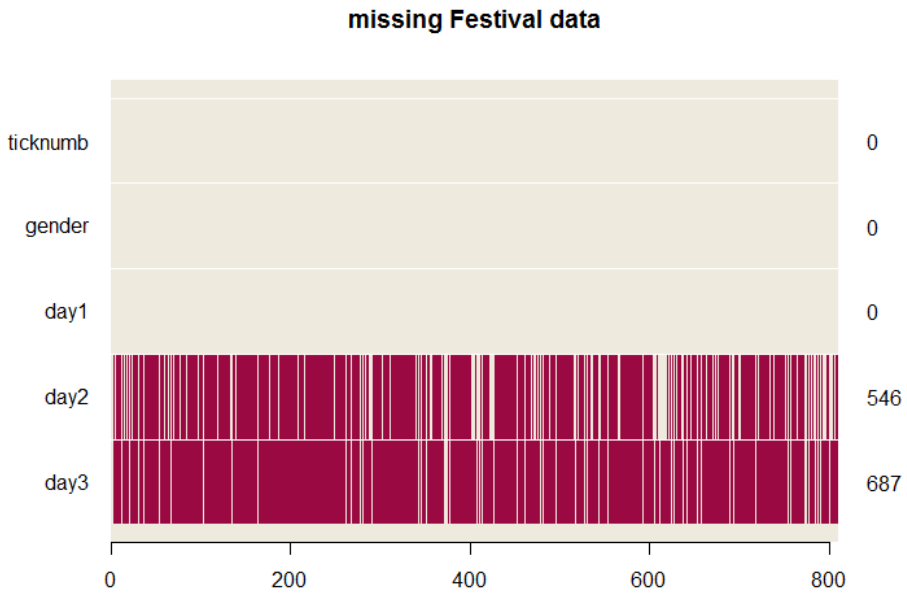
Summary statistics

- {plyr}
  - > summarise(df,

```
Nobs = sum(!is.na(var)),
Nmiss = sum(is.na(var)),
mean = mean(var, na.rm=TRUE),
sd = sd(var, na.rm=TRUE),
se   = sd/sqrt(Nobs),
t = qt(0.975, Nobs-1),
lower = mean - t*se,
upper = mean + t*se)
```

Missing data

- DescTools package
- > PlotMiss(festivalData, main="missing Festival data", clust=FALSE)



Scatterplots using {ggplot2}

- 2 variables

```
> scatter <- ggplot(df, aes(xvar, yvar))
> scatter + geom_point() +
    geom_smooth(method = "lm", colour =  "Blue", se = T, level=0.95) +
  labs(x = "xlabel", y = "ylabel")
```

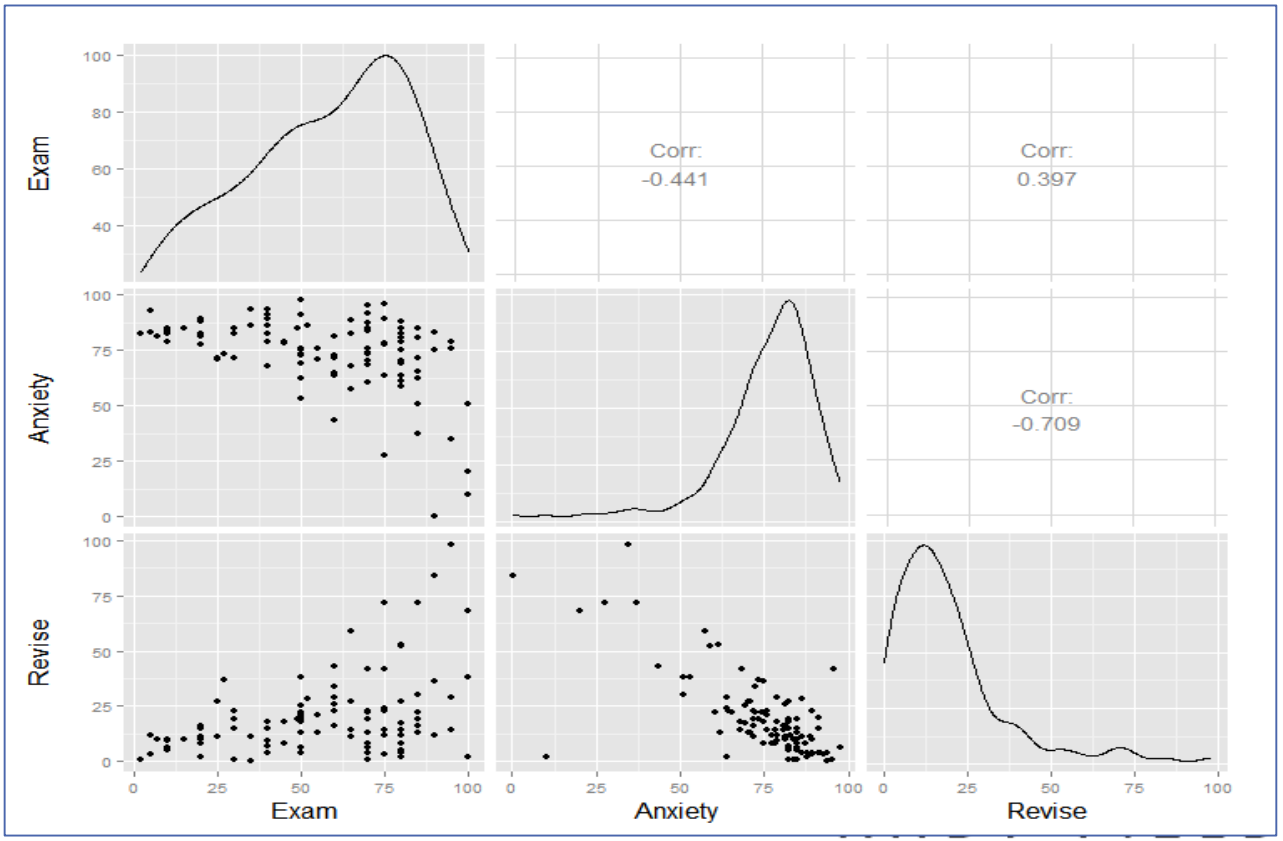
- More than 2 variables {ggplot2 and Ggally}

```
> ggpairs(df[, c("var1", "var2", "var3")],
      diag=list(continuous="densityDiag"), axisLabels='show')
```

- https://www.r-graph-gallery.com/portfolio/ggplot2-package/

Scatterplots

– ggpairs(examData[, c("Exam", "Anxiety", "Revise")], diag=list(continuous="density"), axisLabels='show') {Ggally}



Scatterplots using {mosaic}

- > xyplot(**yvar** ~ **xvar**, data=**df**)

Assumptions

- Variables are continuous
- Linear relationship
- Variables should be approximately normally distributed (bivariate normal)

Assessing Normality

- We don't have access to the sampling distribution so we usually test the observed data
- Central limit theorem
  - If $N > 30$, the sampling distribution **of the mean** is normal anyway
- Graphical displays
  - Q-Q plot
  - Histogram
- Values of skew/kurtosis
  - 0 in a normal distribution
- Formal normality tests
  - H_0: data follows a normal distribution
  - H_A: data does not follow a normal distribution

Histograms

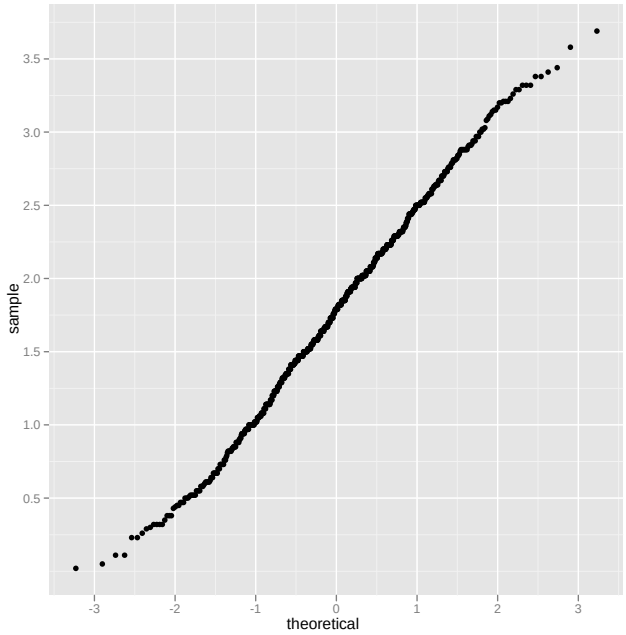
- Histograms plot:
  - Used for plotting continuous data (*x*-axis)
  - Shows the frequency (*y*-axis) distribution of the data
- Histograms help us to identify:
  - The shape of the distribution
    - Skewness (lack of symmetry)
    - Kurtosis
    - Spread or variation in scores
  - Unusual scores

Normal quantile plot (QQ-plot)

- The sample quantiles are plotted against the theoretical quantiles of a normal distribution
- A straight line indicates a normal distribution

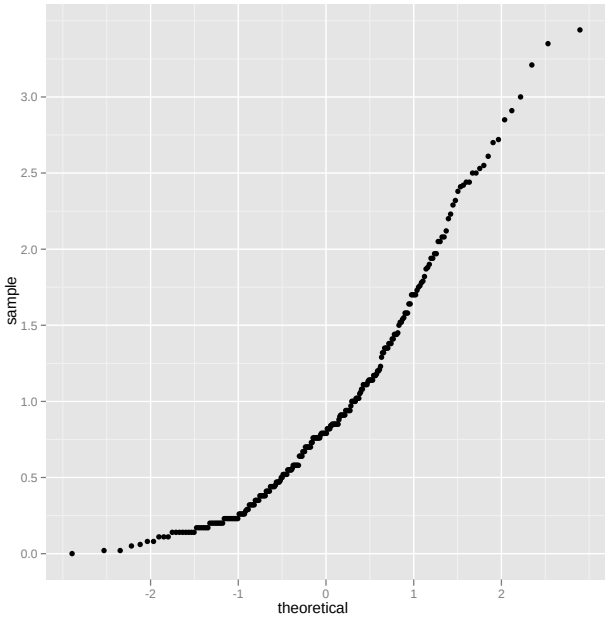
The Q-Q Plot

Hygiene Scores: Day 1



Normal

Hygiene Scores: Day 2



Not Normal

Histogram {ggplot2}

```
> plot2 <- ggplot(df, aes(var))
  plot2 + geom_histogram(binwidth = 200) +
geom_rug(sides="t") +
labs(x = "varlabel", y = "Frequency")
```

Assessing Skew and Kurtosis

- {DescTools}

  > Desc(df$var, main="", plotit=TRUE)

Formal normality test

- Shapiro-Wilk test

```
> shapiro.test(df$var)
```

Testing bivariate normality

- {MVN}

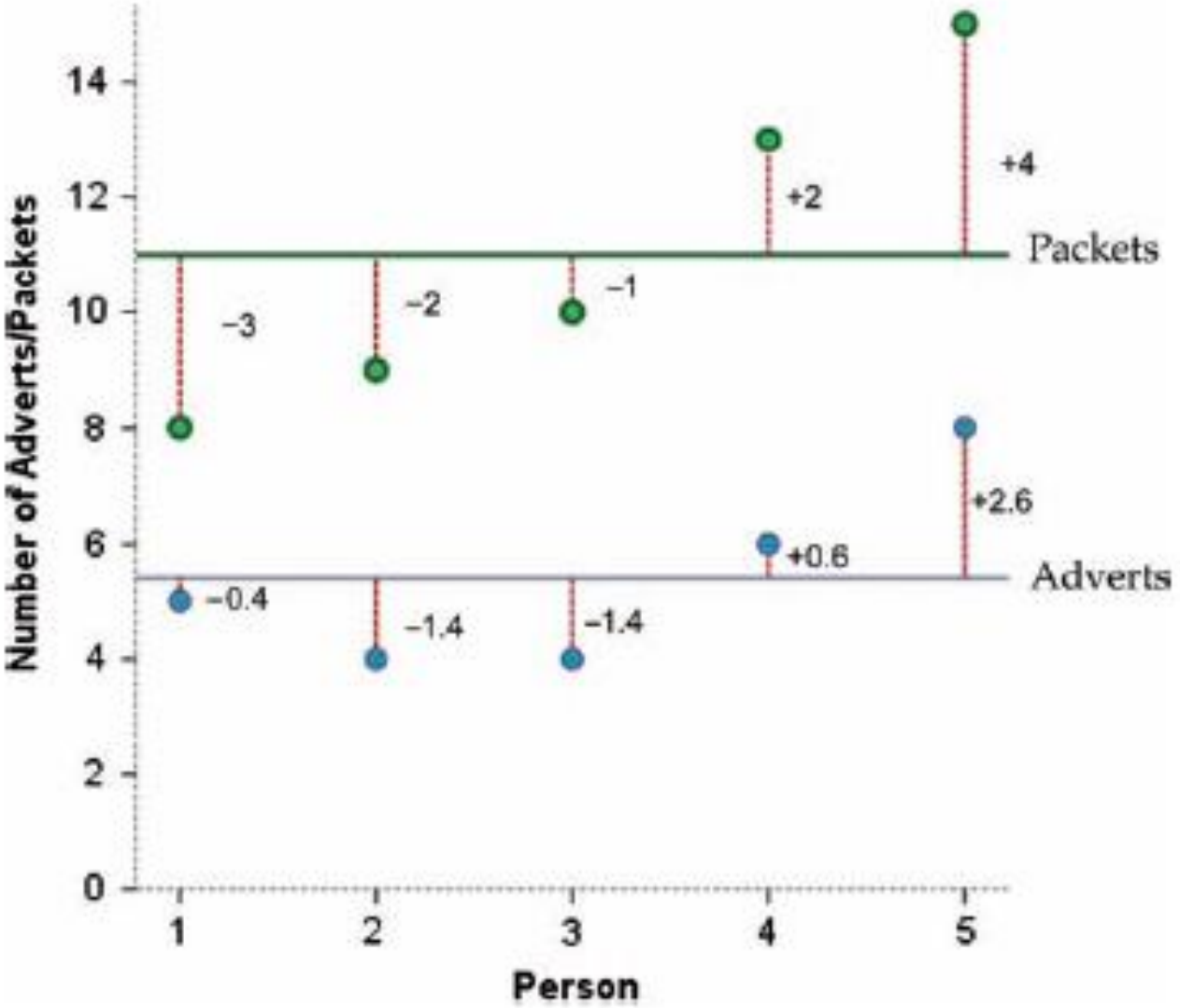
```
> result = mvn(data = df[,2:4],
             mvnTest = "hz",
             univariateTest = "AD",
             univariatePlot = "histogram",
             multivariatePlot = "qq",
multivariateOutlierMethod = "adj",
        showOutliers = TRUE, showNewData = FALSE)

> result$multivariateNormality

> result$univariateNormality
```

Measuring Relationships

- We need to see whether as one variable increases, the other increases, decreases or stays the same.
- This can be done by calculating the covariance.
  - We look at how much each observation deviates from the mean.
  - If both variables deviate from the mean by the same amount, they are likely to be related.


Number of Adverts/Packets
14
12
10
8
6
4
2
0
1
2
3
4
5
Person
Packets
Adverts
+4
+2
−1
−2
−3
+2.6
+0.6
−0.4
−1.4
−1.4

Revision of Variance

- The variance tells us by how much scores deviate from the mean for a single variable.
- It is closely linked to the sum of squares.
- Covariance is similar – it tells by how much scores on two variables differ from their respective means.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}{n-1}$$

Covariance

- Calculates the error between the mean and each subject's score for the first variable (*x*).
- Calculate the error between the mean and their score for the second variable (*y*).
- Multiply these error values.
- Add these values and you get the cross product deviations.
- The covariance is the average cross-product deviations:

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Problems with Covariance

- It depends upon the units of measurement.
  - E.g. the covariance of two variables measured in miles might be 4.25, but if the same scores are converted to kilometres, the covariance is 11.
- One solution: standardize it!
  - Divide by the standard deviations of both variables.
- The standardized version of covariance is known as the correlation coefficient.
  - It is relatively unaffected by units of measurement.

The Correlation Coefficient

$$r = \frac{cov(x,y)}{s_x s_y}$$

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Correlation: Example

- Anxiety and exam performance
- Participants:
  - 103 students
- Measures
  - Time spent revising (hours)
  - Exam performance (%)
  - Exam Anxiety (the EAQ, score out of 100)
  - Gender

General Procedure for Correlations Using **R**

- To compute basic correlation coefficients there are three main functions that can be used:
  - cor{stats}, default NA in presence of missing values
  - cor.test{stats}, missing pairs are deleted by default
  - rcorr{Hmisc} , missing values are deleted in pairs

Function	Pearson	Spearman	Kendall	*p*-values	CI	Multiple Correlations?	Comments
cor()	✓	✓	✓			✓	
cor.test()	✓	✓	✓	✓	✓		
rcorr()	✓	✓		✓		✓	2 d.p. only

Correlations using **R**

- Pearson correlations:

  > cor(**examData**, use = "complete.obs", method = "pearson")

  > cor.test(**examData$Exam**, **examData$Anxiety**, method = "pearson")

  > rcorr(**examData**, type = "pearson")

Pearson Correlation Output

	Exam	Anxiety	Revise
Exam	1.0000000	-0.4409934	0.3967207
Anxiety	-0.4409934	1.0000000	-0.7092493
Revise	0.3967207	-0.7092493	1.0000000

Reporting the Results

- The time spent revising was negatively correlated with exam anxiety, $r = -.71$ ($p < .001$).

- What about?
  - Exam performance was significantly correlated with exam anxiety, $r = -.44$, and time spent revising, $r = .40$ ($p<0.001$);

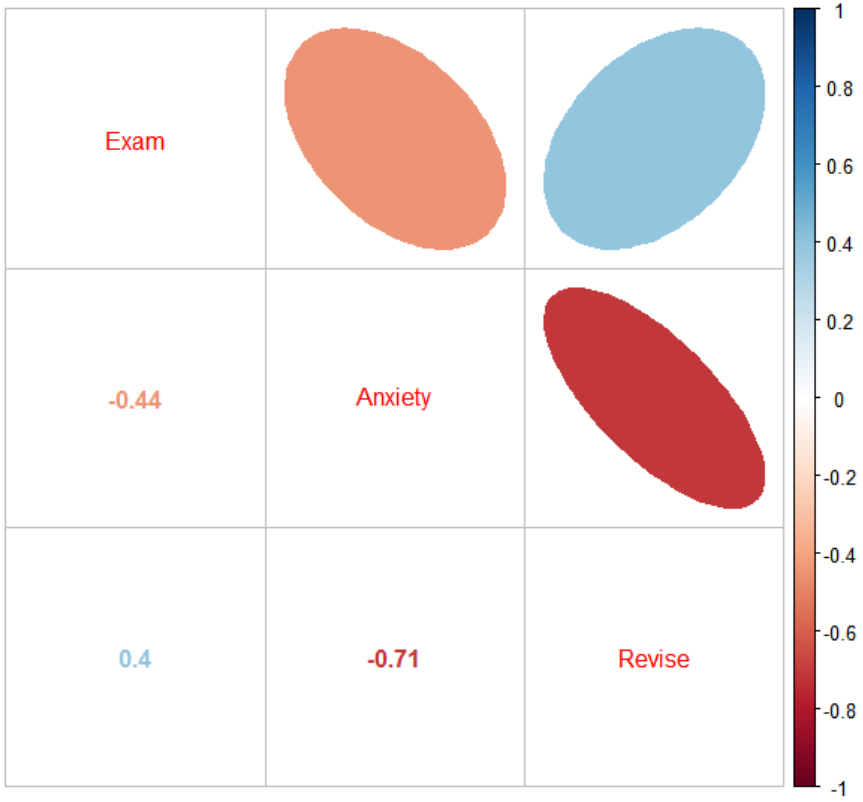
Things to Know about the Correlation

- It varies between -1 and +1
  - 0 = no relationship
- It measures the strength of a linear relation
- Rejecting H_0 indicates only great confidence that ρ is not exactly zero.
- A *p*-value does not measure the magnitude of the association.
- Sample size affects the *p*-value.
- Not robust against outliers
- Coefficient of determination, r^2
  - By squaring the value of *r* you get the proportion of variance in one variable shared by the other.

Note: the notations ρ(rho) or r are used to denote the Pearson correlation

Visualise correlations

- corrplot.mixed(C$r, lower = "number", upper = "ellipse") {corrplot}

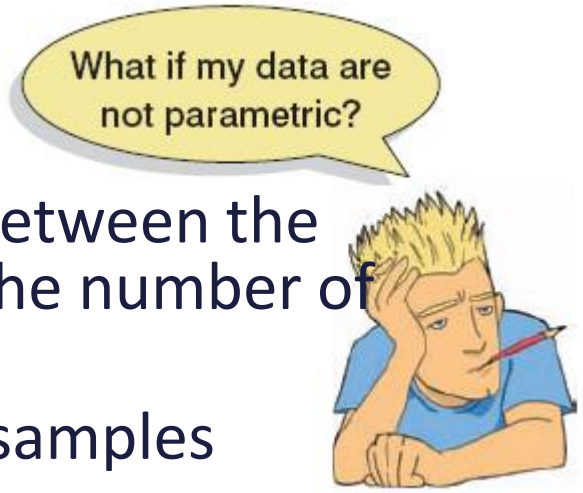


Correlation and Causality

- The third-variable problem:
  - In any correlation, causality between two variables cannot be assumed because there may be other measured or unmeasured variables affecting the results.
- Direction of causality:
  - Correlation coefficients say nothing about which variable causes the other to change.

Non-parametric Correlation

- Spearman's rho rank-order correlation
  - Pearson's correlation on the ranked data
  - is a statistical measure of the strength of a monotonic relationship between paired data
- Kendall's tau
  - Based on ranked data
  - is proportional to the difference between the number of concordant pairs and the number of discordant pairs
  - Better than Spearman's for small samples
  - When there are lots of ties



Spearman's rho

- Takes also values from -1 to +1
  - 1 indicates a perfect association of ranks
  - cor(grade.english, grade.math , method = "spearman")
  - cor(Rx, Ry , method = "pearson")

English (mark)	Maths (mark)	Rank (English)	Rank (maths)
56	66	9	4
75	70	3	2
45	40	10	10
71	60	4	7
62	65	6	5
64	56	5	9
58	59	8	8
80	77	1	1
76	67	2	3
61	63	7	6

Spearman's Rho

- World's Biggest Liar competition
  - 68 contestants from past competitions
  - Measures
    - Where they were placed in the competition (first, second, third, etc.)
    - Creativity questionnaire (maximum score 60)
  - cor(liarData$Position, liarData$Creativity, method = "spearman")
  - cor.test(liarData$Position, liarData$Creativity, alternative = "less", method = "spearman")
  - rcorr(as.matrix(liarData[, 1:2]),type="spearman")
- Application
  - To assess performance of F1 progeny in different environments

Spearman's Rho
Output

Spearman's rank correlation rho

data: liarData$Position and liarData$Creativity

S = 71948.4, p-value = 0.0008602

alternative hypothesis: true rho is less than 0

sample estimates:

rho

-0.3732184

Kendall's Tau

- To carry out Kendall's correlation on the World's Biggest Liar data simply follow the same steps as for Pearson and Spearman correlations but use *method = "kendall"*:

```
cor(liarData$Position, liarData$Creativity,
method = "kendall")
```

Kendall's Tau

- The output is much the same as for Spearman's correlation.

```
        Kendall's rank correlation tau

data:  liarData$Position and liarData$Creativity
z = -3.2252, p-value = 0.0006294
alternative hypothesis: true tau is less than 0
sample estimates:
       tau
-0.3002413
```

DEMO CORRELATIONS

Open the program Ch6_correlations.R

Exercises

- Task 1
  - A student was interested in whether there was a positive relationship between the time spent doing an essay and the mark received. He got 45 of his friends and timed how long they spent writing an essay (hours) and the percentage they got in the essay (essay). He also translated these grades into their degree classifications (grade): in the UK, a student can get a first-class mark (the best), an upper second, a lower second, a third, a pass or a fail (the worst). Using the data in the file **EssayMarks.dat** find out what the relationship was between the time spent doing an essay and the eventual mark in terms of percentage and degree class. Draw a scatterplot too.

Exercises

- Tip Task 1
  - The grade needs to be numerically coded
    - > essayData$grade2[essayData$grade=="First Class"]=1
    - > essayData$grade2[essayData$grade=="Upper Second Class"]=2
    - > essayData$grade2[essayData$grade=="Lower Second Class"]=3
    - > essayData$grade2[essayData$grade=="Third Class"]=4

Exercises

- Task 2
  - Use the dataset **grades.csv** to determine whether previous expertise with mathematics determines whether a student will do well on a statistics course. The degree classifications are as described above. Math grades are coded 1 to 6, with 1 being the best grade.