



Newcastle
University

Sellafield Workshop

March 2022

Dr. Lee Fawcett and Dr. Joe Matthews,
Newcastle University

Lesson 2: Exploratory multivariate data analysis

Problem session 1

Question: A hidden outlier?

Suppose that you're the production supervisor at a company. Your process has been running smoothly until, suddenly, alarms are sounding, and you need to decide what to do!

Your production relies on observations on two main variables - temperature (°C) and acidity (pH) - both of which are recorded once per hour.

The data are stored on a webpage, and can be read into R using the following code:

```
1 #Reads the data into R from a webpage:
2 dd = read.table("https://www.mas.ncl.ac.uk/~nlf8/sellafield/experiment.txt")
```

The data on acidity are stored in the first column of `dd`; data on temperature are stored in the second column.

- (a) Produce time series plots for both acidity and temperature. For example:

```
1 plot(ts(dd[,1]), main="Plot of acidity values", xlab="Time", ylab="pH")
```

Can you detect anything unusual from your time series plots?

- (b) Now produce a scatterplot of acidity against temperature, using R's `plot` command, and comment.

Problem session 2

Question: Scatterplot matrices

Let's consider the construction of simple scatterplot matrices in R.

- (a) First, import the `heptathlon` dataset that we studied during the last session, and take a look around the dataset:

```
1 #Reads the heptathlon data into R from a webpage:
2 heptathlon=read.table("https://www.mas.ncl.ac.uk/~nlf8/sellafield/heptathlon.
3 txt", header=TRUE, row.names=1)
4 #Look at the head of the dataset:
5 head(heptathlon)
6
7 #Extract performances for the 100 metres hurdles:
8 heptathlon$H100M
```

- (b) Use R's `apply` command to produce histograms of performances on all 7 variables. Don't forget to make use of `par(mfrow=c(2,4))` to partition your plotting window!

- (c) Use the `pairs` command to re-produce the scatterplot matrix from the last session.
- (i) What is the nature of the association between performances in the 100 metres hurdles and the 800 metres? What about the relationship between performances in the 100 metres hurdles and the long jump?
 - (ii) How are performances in the shot put and the javelin throw associated?

Question: Correlation matrices

Let's consider the construction of correlation matrices in R. As demonstrated in the last session, execute the following code in R:

```
1 cor(heptathlon)
```

- (a) Can you explain how the correlation matrix supports your answers in question 2(c)?
- (b) Why do all the cells down the diagonal of the matrix contain 1.00000?
- (c) Which pair of variables seem to have the strongest association? What is the *direction* of this association?
- (d) For which pairs of variables might you say there is *no* association? Why should you be careful here?

Problem session 3

Question: Principal Components Analysis

The primary focus of the lecture was on the use of Principal Components Analysis (PCA) as a way of exploring multivariate datasets. Recall from the lecture that, for easier interpretation of the principal components, we should make sure the direction of each variable has the same interpretation. In the lecture, we re-scaled data on the running events to make high values correspond to better performances. The following code reads the heptathlon dataset into R (you would have already used this in an earlier question), and then re-scales the running variables so that – for all variables – larger values correspond to better performances:

```
1 > heptathlon=read.table("https://www.mas.ncl.ac.uk/~nlf8/sellafield/heptathlon.txt",
2   header=TRUE, row.names=1)
3 > heptathlon2 = heptathlon
4 > heptathlon2$H100M = max(heptathlon2$H100M) - heptathlon2$H100M
5 > heptathlon2$R200M = max(heptathlon2$R200M) - heptathlon2$R200M
6 > heptathlon2$R800M = max(heptathlon2$R800M) - heptathlon2$R800M
```

Now try the following code to re-produce the analysis from the lecture:

```
1 > (pca = prcomp(x=heptathlon2, scale=TRUE))
2 > plot(pca$x[,1], pca$x[,2], xlab="First PC", ylab="Second PC")
3 > text(pca$x[,1], pca$x[,2], labels=row.names(heptathlon2))
```

Question: Principal Components Analysis

Load the USArrests dataset into R by typing:

```
1 > USA=read.table("https://www.mas.ncl.ac.uk/~nlf8/sellafield/USArrests.txt", header=
2   TRUE)
3 #Look at the first 6 rows:
4 > head(USA)
5      State Murder Assault UrbanPop Rape Region
6 1  Alabama   13.2    236      58 21.2  South
7 2  Alaska    10.0    263      48 44.5  North
8 3  Arizona     8.1    294      80 31.0  South
9 4  Arkansas     8.8    190      50 19.5  South
10 5 California     9.0    276      91 40.6  West
11 6  Colorado     7.9    204      78 38.7  West
```

You should see arrests, per 100,000 residents, for assault, murder and rape in each of the 50 US states in 1983. Also given is the percent of the population living in urban areas, and a region identifier. Perform a PCA on this dataset, and use a score plot to detect any clustering.

Lesson 4: Probability and probability models

Problem session 1

Question: Introductory ideas

Consider a spinner with the following quadrants:



- (a) Suppose X : Colour landed on after spin. Complete the probability mass function below.

x	Yellow	Blue	Green	Red
$P(X = x)$	1/4			

- (b) Explain why the table of results above represents a valid probability mass function.
- (c) Find:
- (i) $P(X = \text{Blue or } X = \text{Green})$
 - (ii) $P(X = \text{Yellow or } X = \text{Red})$
- (d) The spinner is spun twice. What is the probability that we see both a Blue and a Yellow?

Problem session 2

Question: Binomial and Poisson probabilities

Let's consider the construction of simple probability distributions in R. We'll start by re-visiting the two examples from the lecture:

1. The number of tropical storms observed per year
2. The number of rats developing tumours after exposure to *Ochratoxin A*

Recall that in example 1 the number of tropical storms observed per year was assumed Poisson with rate $\lambda = 1.92$, where λ was estimated from the mean of the sample. Try out the following code, making sure you follow the comments provided.

```

1 #Set up a vector for X, the number of tropical storms:
2 > (X = seq(0, 7, 1))
3 [1] 0 1 2 3 4 5 6 7
4
5 #Now apply the dpois command to X to find the PMF:
6 > dpois(X, 1.92)
7 [1] 0.146606962 0.281485367 0.270225953 0.172944610 0.083013413 0.031877150
8 [7] 0.010200688 0.002797903
9
10 #And the ppois command to find the CDF:
11 > ppois(X, 1.92)
12 [1] 0.1466070 0.4280923 0.6983183 0.8712629 0.9542763 0.9861535 0.9963541
13 [8] 0.9991520

```

(a) Using the R output, find:

- (i) $P(\text{Exactly 3 tropical storms})$
- (ii) $P(\text{Exactly 6 tropical storms})$
- (iii) $P(\text{At most 4 tropical storms})$
- (iv) $P(\text{Fewer than 4 tropical storms})$
- (v) $P(\text{At least 4 tropical storms})$

Recall that in example 2 the number of rats developing tumours was assumed Binomial with $n = 5$ and $p = 0.35$.

- (b) Use the code above for example 1 to obtain the PMF and CDF for the number of rats developing tumours, as shown during the lecture. *Hint: Try R's built-in commands `dbinom(X, n, p)` and `pbinom(X, n, p)`, to replace the commands used for the Poisson distribution above.*
- (c) Suppose we will roll a fair, six-sided die, 4 times, and we are interested in X : The number of sixes we will roll. Notice this could be modelled as a Binomial distribution:
 - We have $n = 4$ fixed trials
 - Each trial has two possible outcomes – “success” (get a 6) or “failure” (don't get a 6)
 - We should be able to assume independence between the trials – for example, getting a six on one trial has no bearing over whether we get a six in any other trial
 - The probability of success is $p = 1/6$
 - Therefore, we can say $X \sim \text{Bin}(4, 1/6)$

Use R in the sameway you did in the last example to find the full probability mass function for X . By hand, calculate the expected value and the variance of X .

Question: More Binomial and Poisson probabilities

- (a) A recent study published in the *Annals of Internal Medicine* suggests that far more people than previously thought are carrying variants of rare genetic diseases, which could force us to redefine what is considered a healthy genome. In fact, they estimate that 20% of people are likely to carry such variants.

For a sample of 10 randomly chosen subjects, use R to find the full probability distribution for X : Number of subjects carrying variants of rare genetic diseases.

- (b) Stored fruit is subject to insect damage. For a sample of 150 fruits, the average number of insects per fruit is 1.55. Use R to find the probability that a randomly selected piece of fruit had exactly 0, 1, 2 and 3 insects.

Problem session 3

Question: The Normal distribution

Load the female haematocrit values into R, and store them in the vector `pcv`, by typing:

```
1 pcv = scan("https://www.mas.ncl.ac.uk/~nlf8/sellafeld/haem.txt")
```

- (a) Use appropriate graphical and numerical techniques to summarise the dataset, and to help to confirm an assumption of Normality – with $\mu = 42.15$ and $\sigma = 3.44$.
- (b) Use the following R code to find the probability that a randomly chosen female student will have a haematocrit value of less than 45%:

```
1 #Recall, the mean of our sample is 42.15 and the standard deviation is 3.44:
2 > pnorm(45, 42.15, 3.44)
3 [1] 0.796303
```

- (c) Now, using similar code, find the probability that a randomly chosen female student will have a haematocrit value (i) less than 48%; (ii) less than 53%.
- (d) What is the probability of a randomly chosen female student having a haematocrit measurement of *more than* 45%?
- (e) Suppose we are told that male haematocrit values are Normally distributed, with $\mu = 45$ and $\sigma = 4$. To produce synthetic data for 126 male students, type the following code (note the switch from the `p` prefix to the `r` prefix – this switches from probability generation to random generation!):

```
1 pcv_male = rnorm(126, 45, 4)
```

Explore the simulated dataset by producing a histogram. Also find the mean and standard deviation of your simulated sample. What is the probability that a male student will have a haematocrit value less than 48%? How does this compare to that for a female student?

Question: Normal Q-Q Plots

- (a) Sticking with the female Haematocrit values, let's produce a Q-Q plot to formally assess Normality in the data:

```
1 qqnorm(pcv)
2 qqline(pcv)
```

How would you interpret this plot?

- (b) The *Exponential distribution* is a probability distribution for continuous data that are very left-skewed (rather than symmetric, as with Normally-distributed data). This distribution has a single rate parameter λ controlling the extent of asymmetry in the distribution. Let's look at what a normal Q-Q plot looks like for exponentially-distributed data!

```
1 #Simulate a synthetic dataset of 500 values from an exponential distribution
   with rate 2:
2 X = rexp(1000, 2)
3 #Now produce a Normal Q-Q plot:
4 qqnorm(X)
5 qqline(X)
```

Lesson 6: Analysis of Variance

Problem session 1

Question: One-Way ANOVA

The following data are from an experiment to compare the effects of four catalysts on the yield of a chemical process:

Catalyst	Observations				
A	4.85	4.89	5.60	5.08	3.42
B	9.43	6.77	5.39	7.39	
C	4.67	5.49	5.96	3.71	5.51
D	5.50	4.90	6.00	5.81	4.68

Type the data into a stacked dataframe in R in the following way:

```
1 yield = c(4.85, 4.89, 5.60, 5.08, ..., 4.68)
2 catalyst = c(rep("A", 5), rep("B", 4), rep("C", 5), rep("D", 5))
```

- Using any techniques you have learnt in this course so far, check the assumption that the four samples are Normally distributed.
- Also, informally check the underpinning assumption of equal variances in the four populations from which these samples have been drawn.
- Write down your null hypothesis.
- Now perform a one-way Analysis of Variance using the following code, and interpret the output.

```
1 > fit = aov(yield ~ catalyst)
2 > summary(fit)
3 > (tukeytest = TukeyHSD(fit))
4 > plot(tukeytest)
```

Question: Randomised Blocks ANOVA

Recall the suncream example from the lecture:

In an experiment to determine the relative effectiveness of four brands of sun cream (Miksa *et al.*, 2016), six volunteers each tried four creams (A, B, C and D) on various parts of their legs. They were then subjected to ultra-violet radiation and an estimate of the degree of burning was made (higher figures indicate greater burning).

The results are given below, along with some summaries:

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6	Total	mean	s.d.
A	8.6	8.8	7.9	6.4	6.8	8.1	46.6	7.77	0.969
B	5.9	6.2	6.6	6.6	7.4	6.3	39.0	6.50	0.514
C	6.7	9.0	8.4	8.6	6.9	7.2	46.8	7.80	0.982
D	8.2	7.4	8.4	6.9	7.5	7.5	45.9	7.65	0.554

Read the suncream data used in the lecture into R using the following code:

```
1 > suncream = read.table("https://www.mas.ncl.ac.uk/~nlf8/sellafield/suncream.txt",  
  header = TRUE)  
2 > suncream2 = stack(suncream)
```

- (a) Repeat the analysis from the lecture, checking for differences in the performances of the four suncreams.
- (b) Add an extra term into the model to check the significance of a blocking factor (the 'subject' could be a blocking factor here), and perform the randomised blocks ANOVA in R. What do you conclude?

Problem session 2

Question: Factorial experiments

Run the following code in R to re-produce the analysis on the rats data we have just examined in the lecture. The R code is available separately, so you can just copy-and-paste this into your R terminal. Make sure you understand what's going on each step of the way!

```
1 #read in the data from a webpage:  
2 > rats = read.table("https://www.mas.ncl.ac.uk/~nlf8/rats.txt", header = TRUE)  
3  
4 #Make sure R knows that the source, and level, of protein are both factors:  
5 > rats$Source = as.factor(rats$Source)  
6 > rats$Level = as.factor(rats$Level)  
7  
8 #Produce some summary plots – checking for Normality along the way!  
9 > par(mfrow=c(1,2))  
10 #Boxplots for the data by source of protein:  
11 > boxplot(rats$Gains~rats$Source)  
12 #And now by level of protein:  
13 > boxplot(rats$Gains~rats$Level)  
14  
15 #How about Normal Q-Q plots from last week? Easy!  
16 > par(mfrow=c(2,3))  
17 > qqnorm(rats$Gains[rats$Source==1], main="Source = Beef")  
18 > qqline(rats$Gains[rats$Source==1])  
19 > qqnorm(rats$Gains[rats$Source==2], main="Source = Cereal")  
20 > qqline(rats$Gains[rats$Source==2])  
21 > qqnorm(rats$Gains[rats$Source==3], main="Source = Pork")  
22 > qqline(rats$Gains[rats$Source==3])  
23  
24 > qqnorm(rats$Gains[rats$Level==1], main="Level = High")  
25 > qqline(rats$Gains[rats$Level==1])  
26 > qqnorm(rats$Gains[rats$Level==2], main="Level = Low")  
27 > qqline(rats$Gains[rats$Level==2])  
28  
29 #Let's now produce some numerical summaries – we can use any function we like, but  
  let's check the assumption of equal variances:  
30 > aggregate(rats$Gains, list(rats$Source), FUN = sd)  
31 > aggregate(rats$Gains, list(rats$Level), FUN = sd)  
32  
33 #Now let's do the ANOVA!  
34 > fit = aov(rats$Gains~rats$Source+rats$Level+rats$Source*rats$Level)  
35 > summary(fit)
```

```

1 #And finally , an interaction plot:
2 > par(mfrow=c(1,1))
3 > interaction.plot(x.factor = rats$Source, #x-axis variable
4                   trace.factor = rats$Level, #variable for lines
5                   response = rats$Gains, #y-axis variable
6                   fun = mean, #metric to plot
7                   ylab = "Weight gain (g)",
8                   xlab = "Source of protein",
9                   col = c("pink", "blue"),
10                  lty = 1, #line type
11                  lwd = 2, #line width
12                  trace.label = "Level of protein")

```

Lesson 8: Further Topics in Regression

Problem session 1

Question: Logistic regression

Import the dataset `potter.txt` into R by typing the following code:

```
1 hp = read.table("https://www.mas.ncl.ac.uk/~nlf8/sellafield/potter.txt", header=TRUE)
```

Take a look around the data. You should see that there are 30 rows, each corresponding to a participant in a recent survey investigating whether or not self-declared Harry Potter fans 'identify' with the main protagonist in the franchise.

The column labelled 'Potter' is the main response variable, taking 1 if the survey participant said they did identify with Harry Potter, and 0 otherwise. Column 2 shows the Age of the participant, and column 3 takes the value 0 if the participant was male, and 1 if they were female.

Notice that the main response variable is binary, so we will perform a binary logistic regression in R.

- (a) Use the `glm` function in R, with the argument `family = binomial`, to perform a multiple binary logistic regression.
- (b) Are both Age and sex important predictors of the probability that a fan identifies with Harry Potter?
- (c) Use a process of backwards elimination to find the best binary logistic regression model. How might you interpret the regression coefficients in your resulting model?
- (d) Suppose a new 19-year-old Harry Potter fan is recruited to the study. How likely is it that he will identify with Harry Potter?