Problem Set 0: Intro to R Markdown

Statistics 139 Teaching Staff Linh Vu Collaborator: Brice Laurent

Due: September 15, 2023

This is an R Markdown (Rmd) document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. It was designed to simplify creating HTML documents, but we will be using just PDF documents.

The content of an R Markdown document is created in the R Studio script editor. Formatting commands in the text are converted to a PDF output when you click the *Knit* button, located on the toolbar at the top of the script editor.

In R Studio, the menu item File/New File/R Markdown choice produces a dialog box for output type. To create a new document, select document from the left side of the dialog box, enter a name and title, and choose PDF. A template will open in the script editor, with file extension .Rmd. We will provide templates for problem set solutions, and you will start by simply opening the file.

Let's get started!

The first six lines of this file are referred to as the **header**. R Markdown is very particular about the form of the header - the three dashed lines above and below lines 2-5 must appear exactly as in this document, and there must be a blank space between the colon and the descriptive text. Also, the title, author, and date fields must be kept within double quotes. The output line specifies the output format as PDF; you will never need to change that line for the purposes of this class.

Using R Markdown to knit a PDF

1. First, rename this file to include your first initial and last name – e.g., stat139_pset0_jxenakis.rmd.
2. In this document, edit the header to include your name. Click *Knit PDF*. This should produce a PDF file located in the same folder as the Rmd file, with a name like stat139_pset0_jxenakis.pdf. Note that the file name for a PDF created from an Rmd document will be the same, except with a different file extension. The file name and title of the document seen in the header can be different.

If you did not get a PDF file, stop and try to diagnose the problem by looking at the error messages in the R Markdown section of the Console. The error messages are not always helpful, so if you cannot solve the issue, ask the teaching staff for help or talk with another student.

Plain text is prepared in paragraphs, as in the first part of this document. *Text* enclosed in asterisks is *italicized* in the PDF output. **Text** enclosed in double asterisks appears in **bold font**. There must be no space between the asterisks and the enclosed text.

Using text and Markdown language in R Markdown

3. Write a brief paragraph describing previous coursework in statistics (at Harvard or elsewhere) and share your motivation for taking Statistics 139 this semester along with any reservations. *Knit* the document. Note that each time you *knit* the document, the output overwrites the previous version.

I've taken Stat 110 and Stat 111, two theoretical Statistics courses. I've also taken Stat 100 (Intro to Data Science) and Stat 108 (Intro to Statistical Computing), so I'm quite familiar with R and enjoy coding in R. I'm taking Stat 139 mainly because it is required for Stat concentrators, but I'm also excited to learn more about modeling.

Bulleted lists are produced using the formatting commands:

- Item 1
- Item 2
  - Item 2a
  - Item 2b
- Item 3

The list must be preceded by a blank line, and 4 spaces should be used before sub-items.

4. Write a bulleted list giving your class year at Harvard, your concentration or field of study if you have one, and the name of your dorm/house. If you are a grad student, you are Dudley House. Under the entry for your dorm/house, prepare sub-items with the name of your neighborhood (e.g. Yard, River, Quad) and the name(s) of your Resident Dean (you can find resident deans here. *Knit* the document and inspect it to make sure the PDF is correctly produced.

- Junior/class of 2025
- Statistics
- Lowell House
  - River Central
  - David Laibson and Nina Zipser

Using Math mode (aka, LaTeX) in R Markdown

Math mode is similar to LaTeX. In order to call Math mode, you can use the $ symbol just like LaTeX, so you an create in line math symbols like α and $\hat{\beta}$, or you can create a separate formula like:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

5. Write a separate formula to define the variance of a continuous random variable X as seen as the second formula under 'Variance of a continuous random variable' on this webpage.

$$var(X) = \sigma_X^2 = E[(X-\mu)^2]$$

, where

$$\mu = E(X)$$

6. Matrix algebra by hand and in R. Let:

$$\vec{c} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}, \quad \vec{x} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \quad \vec{\mu} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \mathbf{\Sigma} = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$$

Calculate the following by hand using math mode (aka, LaTeX), and confirm the calculations in an R chunk:

(a) $\vec{c}^T(\vec{x}-\vec{\mu})$

$$\begin{bmatrix}0.5 & 0.5\end{bmatrix} \cdot \begin{bmatrix}1\\1\end{bmatrix} = 1$$

(b) $\mathbf{\Sigma}^{-1}$

$$det[\Sigma] = 2 \cdot 4 - 1 \cdot 1 = 7$$

$$\frac{1}{7} \cdot \begin{bmatrix}4 & -1\\-1 & 2\end{bmatrix} = \begin{bmatrix}\frac{4}{7} & \frac{-1}{7}\\\frac{-1}{7} & \frac{2}{7}\end{bmatrix}$$

(c) $\vec{c}^T\mathbf{\Sigma}\vec{c}$

$$\begin{bmatrix}0.5 & 0.5\end{bmatrix} \cdot \begin{bmatrix}2 & 1\\1 & 4\end{bmatrix} \cdot \begin{bmatrix}0.5\\0.5\end{bmatrix} = \begin{bmatrix}0.5 & 0.5\end{bmatrix} \cdot \begin{bmatrix}1.5\\2.5\end{bmatrix} = 2$$

(d) $\vec{c}^T(\vec{x}-\vec{\mu})(\vec{c}^{\,T}\mathbf{\Sigma}\vec{c})^{-1/2}$

Using results from (a) and (c), we get:

$$1 \cdot 2^{-\frac{1}{2}} = \frac{1}{\sqrt{2}}$$

(e) In 30 words or less, briefly interpret the calculation in part (d) (assuming $\vec{x}$ is a data vector, $\vec{\mu}$ is a mean vector, and $\mathbf{\Sigma}$ is a covariance matrix).

- For (e), the formula involves subtracting the mean from the data and dividing that by the variance, so this looks like a form of standardizing the data.

```
# initiate matrices
c <- matrix(c(0.5, 0.5), ncol = 1)
x <- matrix(c(2, 3), ncol = 1)
mu <- matrix(c(1, 2), ncol = 1)
sigma <- matrix(c(2, 1, 1, 4), ncol = 2)

# part a
t(c) %*% (x-mu)
```

```
##      [,1]
## [1,]    1
```

```
# part b
solve(sigma)
```

```
##            [,1]       [,2]
## [1,]  0.5714286 -0.1428571
## [2,] -0.1428571  0.2857143
```

```r
# part c
t(c) %*% sigma %*% c
```

```
##      [,1]
## [1,]    2
```

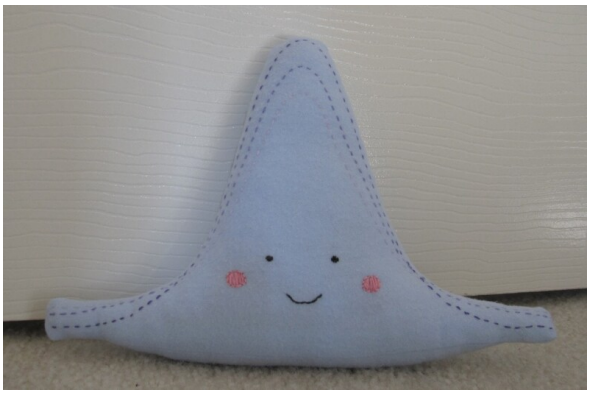
```r
part_c <- t(c) %*% sigma %*% c

# part d
t(c) %*% (x-mu) %*%  part_c^(-0.5)
```

```
##           [,1]
## [1,] 0.7071068
```

To start a new page in the PDF document, enter the text 'newpage' preceded by a backslash (just like in latex), as in. . . (new page coming in the PDF!)

Including hand-written work.

You can also easily include **scanned**, hand-written answers (or any images on your computer) in your knitted pdf. To do this, you need to use the lines `![caption here]('directory/image.png')` (do not forget the exclamation point!). Here are two examples, side-by-side:




7. Draw a picture, a diagram, or your signature by hand and scan it (easiest to use a scanning ap on your phone like 'CamScanner', Adobe's 'Scan', etc.) and include it here. Edit the size so that it looks reasonable in the knitted pdf.

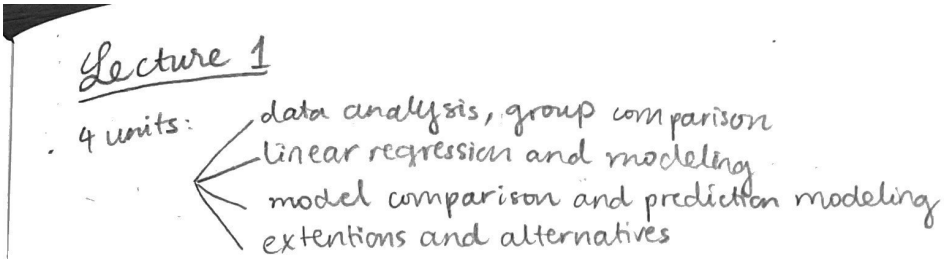


Figure 1: handwrite

Using R with R Markdown

The real power of R Markdown is that it allows for short R programs to be included in the Rmd file, with both the program and its output automatically being produced in the PDF document.

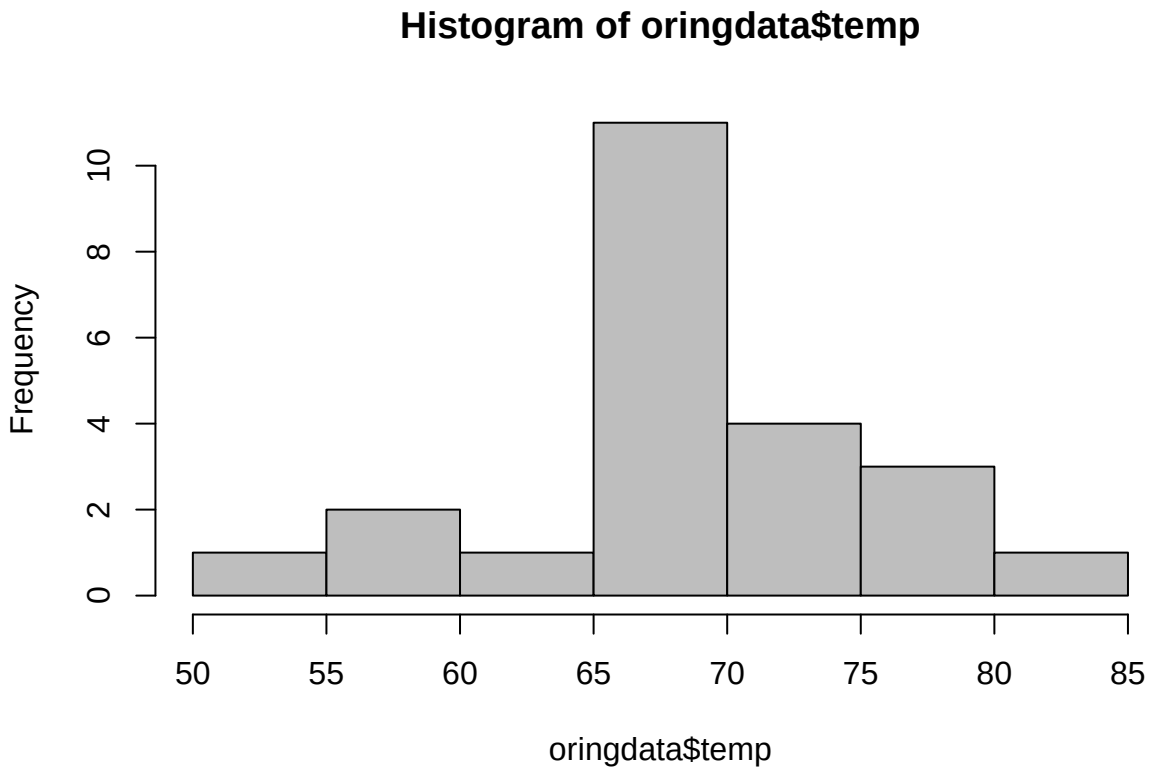
R programs in an Rmd file are located in **code chunks**. You can embed an R code chunk by either typing the three apostrophes followed by an "r" enclosed in braces, then the additional three apostrophes to close the chunk, or simply press the *R* button from the *Insert* menu on the right of the toolbar on top of the script editor.

Most datasets we will use in this course will be placed on the Canvas site where you can download them, or will be placed directly on Posit Cloud. The following steps will allow you to open a data set in R:

A dataset called 'ORingData.csv' has been placed on Posit Cloud (and on Canvas). This dataset contains information on O-ring damage (a measure of the total number of incidents of O-ring erosion, heating, and blow-by) and temperature at lift-off, in degrees Fahrenheit (F), for all 23 previous flights of the space shuttle *Challenger* prior to its explosion on January 28, 1986. Use this dataset to answer the questions below.

The following code will load the data set and produce a histogram. In order for the output to be presented, you need change the **eval=FALSE** option to **eval=TRUE** or **eval=T**. *Knit* the document after this change, and the histogram should now be visible in the PDF output.

```
oringdata=read.csv("./data/ORingData.csv")
hist(oringdata$temp,col="gray")
```



8. Answer parts (a) and (b) with R chunks (include justification/explanation in plain text when prompted) and part (c) with a paragraph in plain text.

(a) Calculate the following summary statistics for both the `temp` and `oring_damage` variables: sample mean, sample standard deviation, min, median, max, and the 1st and 3rd quartiles. Also calculate the proportion of launches that had any O-ring damage at all.

- 30.4% of launches had O-ring damage

```
# sum stats of temp
round(c(summary(oringdata$temp), "SD" = sd(oringdata$temp)), 2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      SD
##   53.00   67.00   70.00   69.17   74.00   81.00    6.93
```

```r
# sum stats of oring_damage
round(c(summary(oringdata$oring_damage), "SD" = sd(oringdata$oring_damage)), 2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      SD
##    0.00    0.00    0.00    1.43    3.00   11.00    2.68
```

```r
# prop of launches that had O-ring damage
mean(oringdata$oring_damage > 0)
```

```
## [1] 0.3043478
```

(b) Split the observations into two groups: the launches with no O-ring damage vs. launches with some O-ring damage. Use summary statistics and graphics to explore whether there is evidence of a difference in temperature on flights when there was no damage to the O-rings vs. flights for when there was any damage at all. Comment on the results without performing a formal hypothesis test.

- Both the numerical and graphic summaries show that the temperature on flights tended to be lower when there was some O-ring damage (vs when there was not). It seems plausible that the lower the temperature at lift-off, the more likely it is that O-ring damage occurs.
- One caveat is that there are only 23 data points and there are still overlapping temperatures recorded for both flights with and without O-ring damage (as seen in scatterplot and boxplot), so the evidence for the trend is not particularly strong.

```r
# subset data
no_damage <- oringdata[oringdata$oring_damage == 0,]
damage <- oringdata[oringdata$oring_damage > 0,]

# numerical summary
summary(no_damage$temp)
```

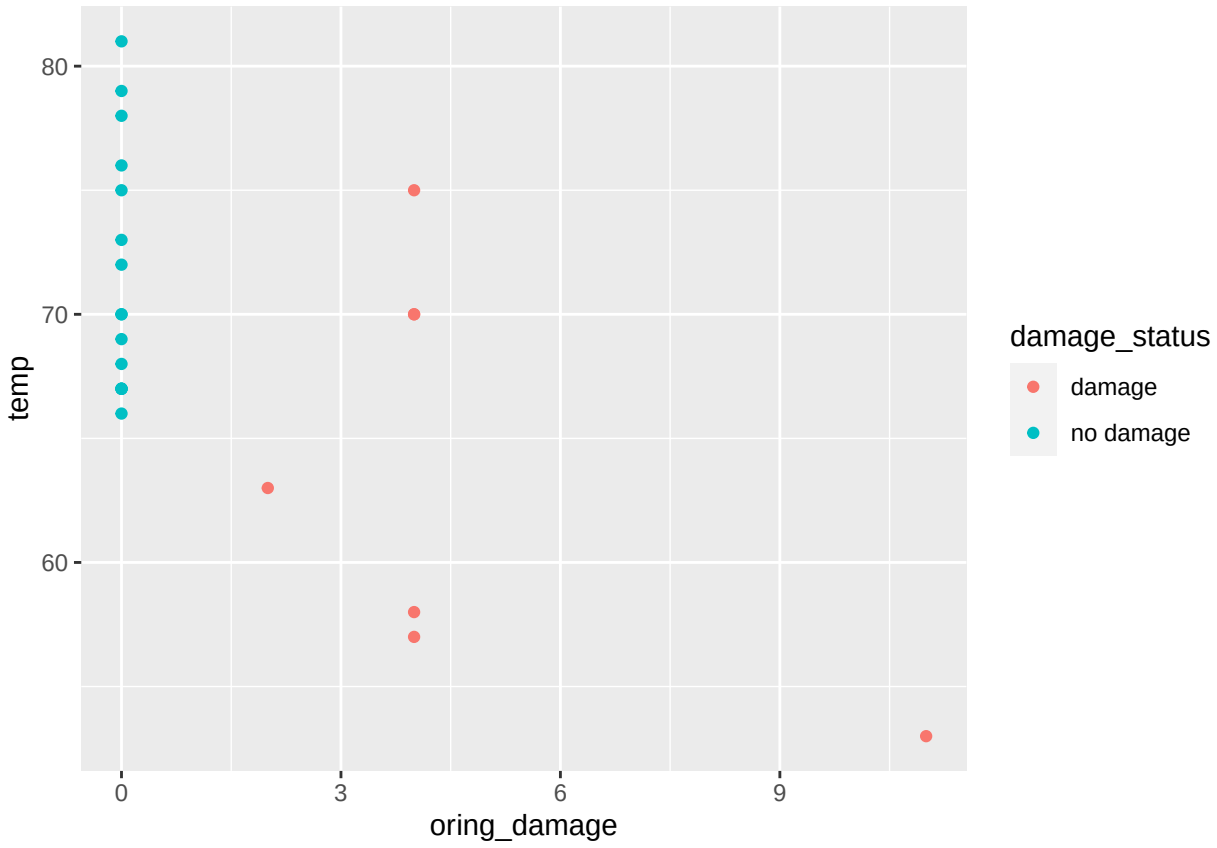
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   66.00   67.00   70.00   71.56   75.25   81.00
```

```r
summary(damage$temp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   53.00   57.50   63.00   63.71   70.00   75.00
```

```r
# graphical summary
boxplot(no_damage$temp, damage$temp,
        names = c("no O-ring damage", "some O-ring damage"),
        ylab = "temp at lift-off (F)")

library(ggplot2)
```


```
oringdata$damage_status <- ifelse(oringdata$oring_damage > 0, "damage", "no damage")
ggplot(oringdata, aes(oring_damage, temp, col = damage_status)) +
  geom_point()
```



(c) It is the morning of January 28, 1986, and the temperature at lift-off is predicted to be around 29 degrees F. NASA would like your professional opinion on whether they should launch the rocket later in the morning. Provide a one paragraph (3-6 sentences) summary of whether they should go ahead with the launch.

NASA should not launch the rocket because 29 degrees F is lower than temperature at lift-off of any historical flight with no O-ring damage. Among previous flights without any O-ring damage, the lowest temperature at lift-off was 66 degrees F, and 29 is way below that, making O-ring damage more likely. And 29 degrees F is even lower than the temperature at lift-off of any previous flight, with or without O-ring damage, making it hard to correctly gauge the safety of the flight.

9. Please share you thoughts on the "Quotes for Discussion" slide from the Unit 0 slides. We will revisit these quotes at the end of the semester and see how our thinking on them may have evolved.

I think the quotes are about how Statistics can be used in life. Even though statistical analysis is imperfect and is subjected to misuse and models cannot accurately represent real life, Statistics as a tool can still shed light on many issues. The quotes all try to get at the fact that we need to strike a balance between enjoying the various benefits of statistical tools and being aware of their shortcomings.