MXB107 R Code Snippets

Dan Tran

2022-09-01

Contents

Assignment 1

This document is intended to help with R Programming only. For mathematical explanation, please refer to the weekly [Readings].

Set up the MXB107 package and load the `episodes` data set. The first line (commented) can be used to install dependencies (only when required by R).

```
# install.packages(c( "tidyverse", "knitr", "rmarkdown", "pander",
# "ggforce", "kableExtra", "gridExtra"))
install.packages("MXB107_1.0.0.2022.zip", repos = NULL, type="source")
```

```
## Installing package into 'C:/Users/autum/OneDrive - Queensland University of Technology (1)/Sessional
## (as 'lib' is unspecified)
```

```
library(MXB107)
```

```
## Loading required package: tidyverse

## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## Loading required package: knitr
##
## Loading required package: rmarkdown
##
## Loading required package: pander
##
## Loading required package: ggforce
##
## Loading required package: kableExtra

## Warning in !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output
## %in% : 'length(x) = 2 > 1' in coercion to 'logical(1)'

##
## Attaching package: 'kableExtra'
##
## The following object is masked from 'package:dplyr':
##
##     group_rows
##
## Loading required package: gridExtra
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

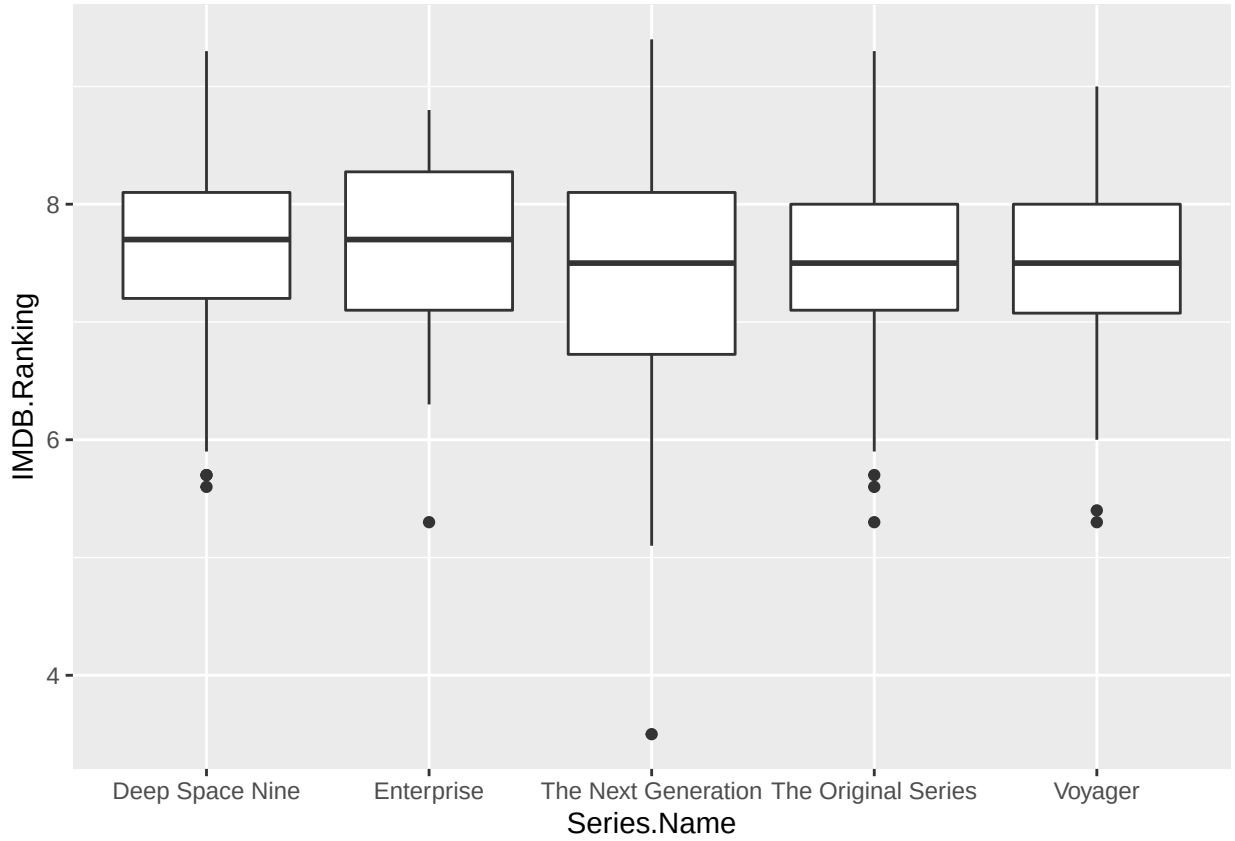
```
data(episodes)
```

Part 1: Summarising Data

Question 1

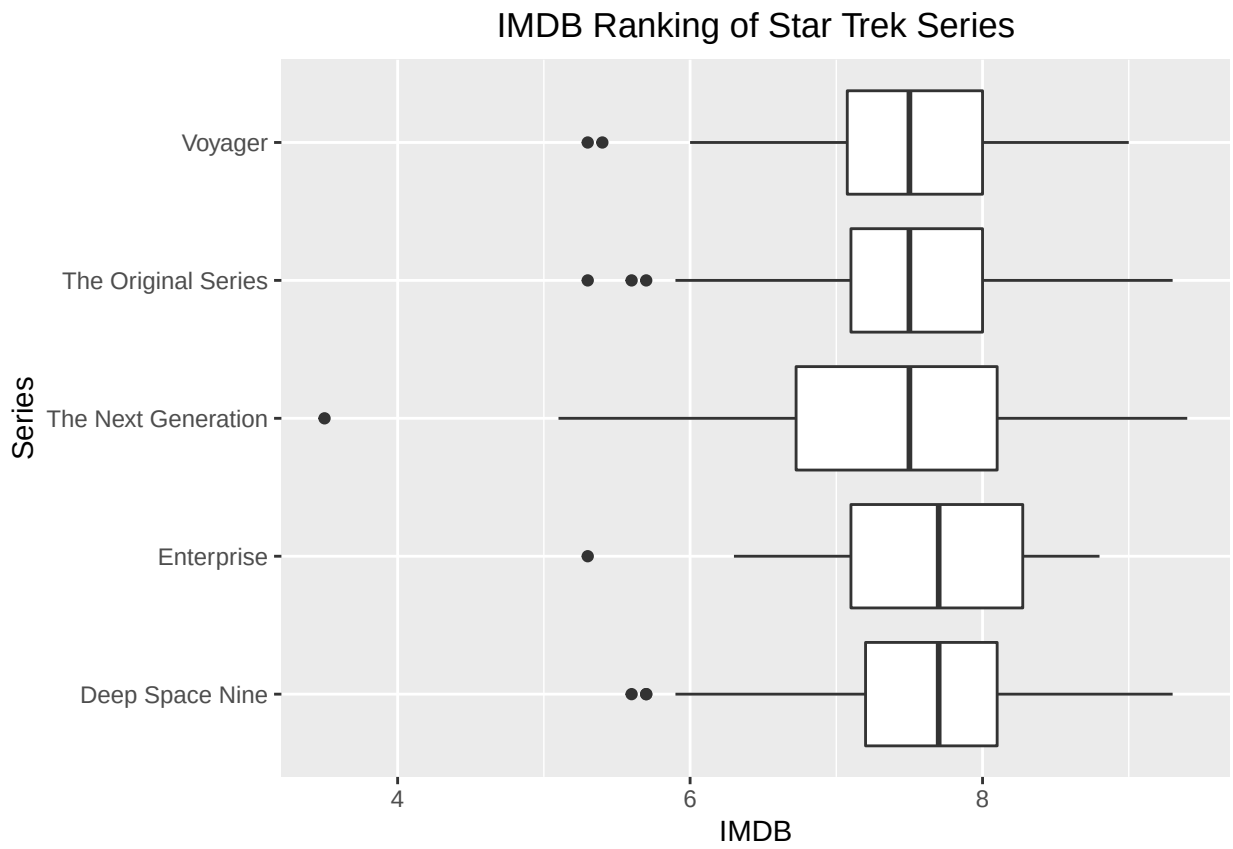
Shows the IMDB rankings for each series of *Star Trek*.

```
ggplot(episodes, aes(x = Series.Name, y = IMDB.Ranking))+
  geom_boxplot()
```



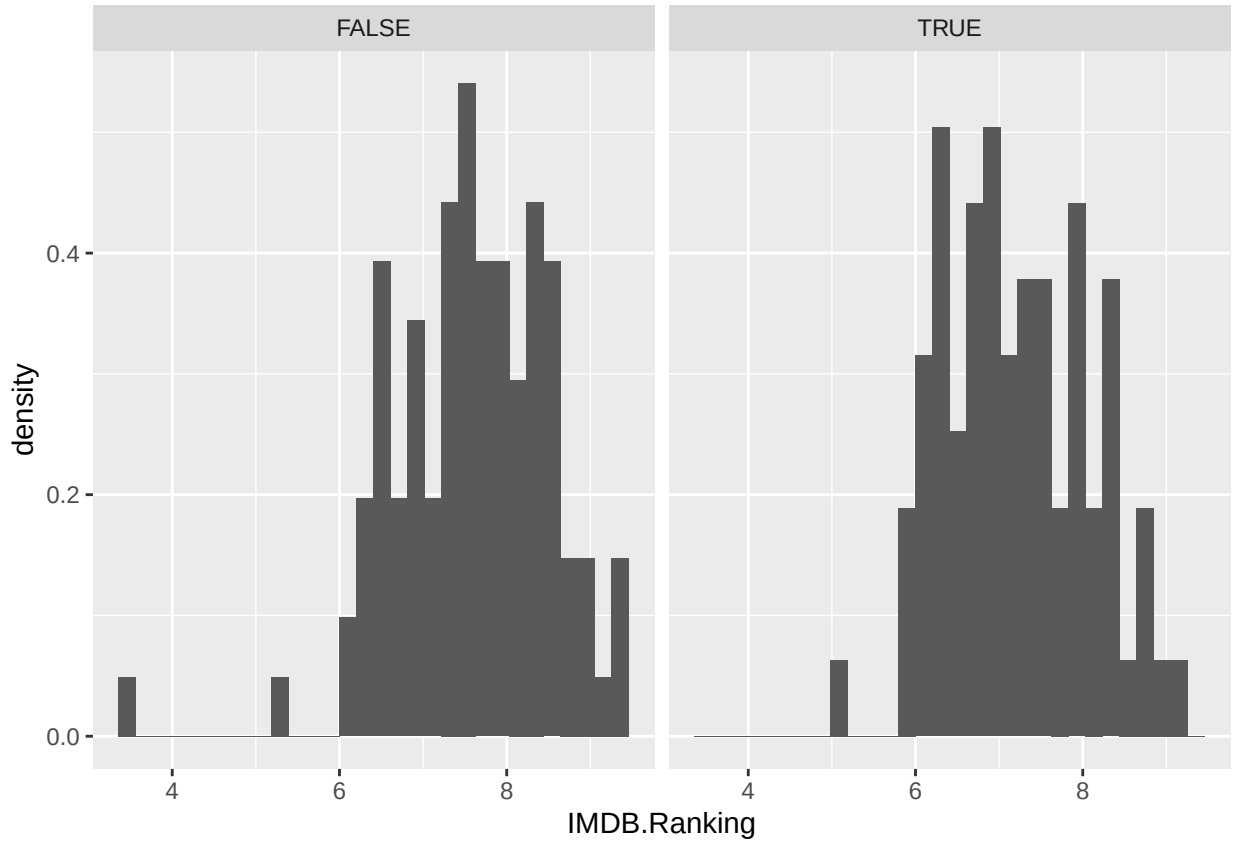
Add title, labels using `title()`, `xlab()` and `ylab()`. **Please change the labels and title in your assignment**

```
ggplot(episodes, aes(y = Series.Name, x = IMDB.Ranking))+ # Changed x and y order for horizontal
  geom_boxplot()+                                         # Add boxplot
  xlab("IMDB")+                                           # Add x-label
  ylab("Series")+                                         # Add y-label
  ggtitle("IMDB Ranking of Star Trek Series")+            # Add title
  theme(plot.title=element_text(hjust=0.5))               # Adjust title to middle
```



Create pair of histograms for IMDB rankings of **Star Trek: The Next Generation** based on Bechdel-Wallace Test status. The `aes(y=..density..)` is used to display the density rather than count.

```
filter(episodes,Series.Name=="The Next Generation")%>%
  ggplot(aes(x=IMDB.Ranking))+
  geom_histogram(aes(y=..density..), bins = 30)+
  facet_wrap(~Bechdel.Wallace.Test)
```



Question 2

Find the total number of rows in the `episodes` table. The `<-` operator assigns the value of the right hand side to the variable on the left-hand side (can be named anything, not just `n`).

```
n <- nrow(episodes)
n # To display the value of a variable, call it like this, or print(n)
```

```
## [1] 704
```

Find the mean IMDB ranking. **`episodes$IMDB.Ranking` selects the `IMDB.Ranking` column from the `episodes` dataframe.**

```
mean <- mean(episodes$IMDB.Ranking)
mean
```

```
## [1] 7.55071
```

Find the sum of squared distances from the mean IMDB Ranking, with `mean` already found above. Equivalent to

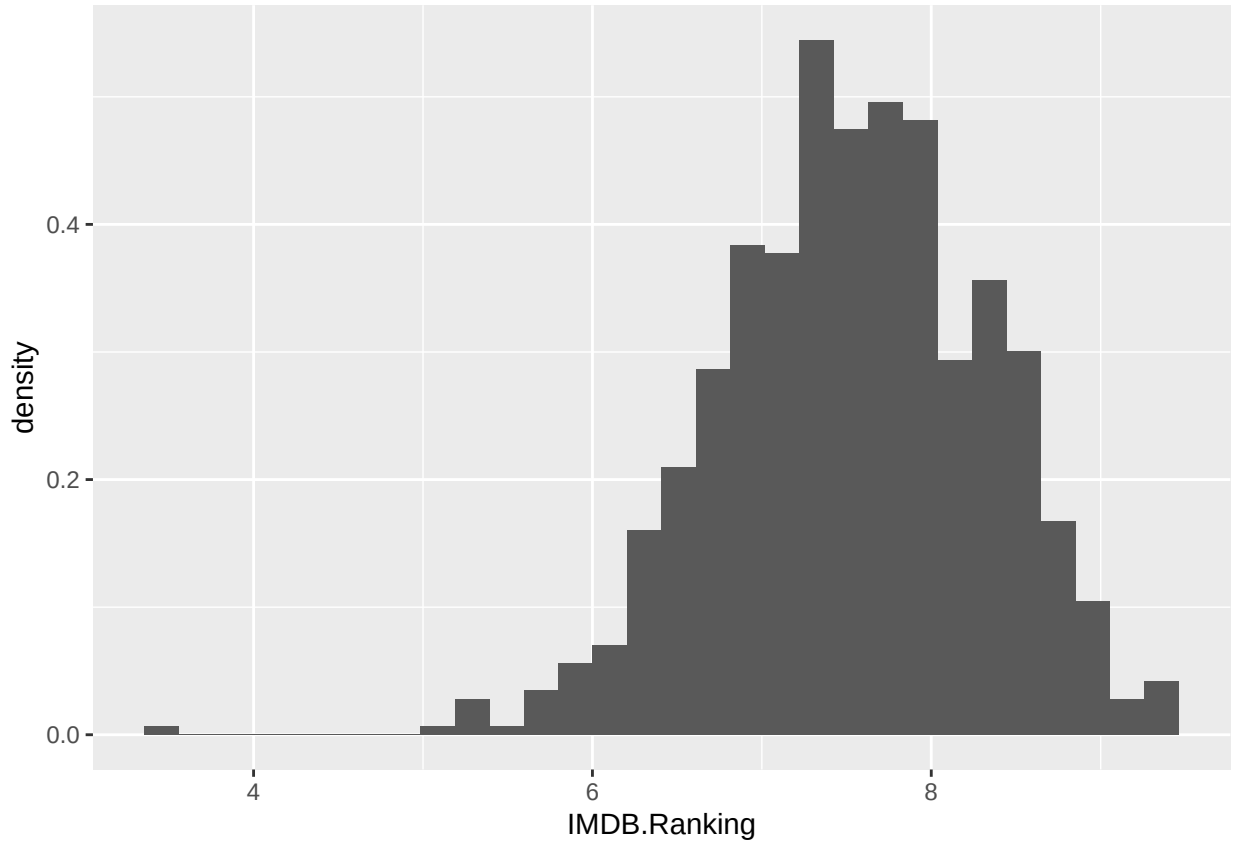
$$\sum_{i=1}^{n} (x_i - \bar{x})^2$$

```
x <- episodes$IMDB.Ranking
sum_of_squared_distances <- sum((x - mean)^2)
sum_of_squared_distances
```

```
## [1] 423.3796
```

Display the histogram of the episode's IMDB rankings.

```
episodes %>%
  ggplot(aes(x = IMDB.Ranking))+
  geom_histogram(aes(y = ..density..), bins = 30)
```



Part 2: Computing Basic Probabilities for Events

Question 1,2 & 3

Basically counting things

Find the number of episodes passing the Bechdel Wallace Test.

```
pass_count <- sum(episodes$Bechdel.Wallace.Test == TRUE)
pass_count
```

```
## [1] 366
```

The condition `episodes$Bechdel.Wallace.Test == TRUE` can be changed to any other condition, such as: Number of episodes from *Star Trek: The Original Series*.

```
original_count <- sum(episodes$Series.Name == "The Original Series")
original_count
```

```
## [1] 80
```

Or the number of episodes from *Star Trek: The Original Series* and pass the Bechdel-Wallace Test. You can apply multiple conditions using the `&` operator between the two comparisons.

```
original_pass_count <- sum(episodes$Series.Name == "The Original Series" &
                           episodes$Bechdel.Wallace.Test == TRUE)
original_pass_count
```

```
## [1] 5
```

Question 4

More advanced counting by accounting for categorical data

The number of episodes of each season for each series, that already passed the Bechdel Wallace Test. *Don't worry if the last two lines don't make sense.*

```
episodes%>%
  filter(Bechdel.Wallace.Test == TRUE)%>% # Remove this filter for all episodes
  group_by(Series.Name, Season)%>%
  tally()%>%
  pivot_wider(names_from = Series.Name, values_from = n)%>%
  bind_rows(summarise_all(., ~sum(., na.rm=TRUE)))%>% # Total column
  mutate(Total = rowSums(.[setdiff(names(.),"Season")], na.rm = TRUE)) # Total row
```

Season	Deep Space Nine	Enterprise	The Next Generation	The Original Series	Voyager	Total
1	13	11	10	NA	14	48
2	17	9	9	2	24	61
3	19	9	9	3	17	57
4	12	9	12	NA	24	57
5	13	NA	16	NA	23	52
6	12	NA	10	NA	23	45
7	14	NA	12	NA	20	46
28	100	38	78	5	145	366

Part 3: Modelling with Probability Distributions

Question 1 & 2

In R, given any distribution (Binomial, Geometric, Poisson, Normal, etc.), the Probability Density Function $Pr(X = x)$ *can be found using* `d<distribution name>`*, and the cumulative probability can be found using* `p<distribution name>`*. I show the Geometric distribution as an example.*

Given $X \sim Geom(0.5)$, find $Pr(X = 2)$.

```r
dgeom(2, prob = 0.5)
```

```
## [1] 0.125
```

Given $X \sim Geom(0.5)$, find $Pr(X \leq 2)$.

```r
pgeom(2, prob = 0.5)
```

```
## [1] 0.875
```

For other distributions, please refer to the documentation [here], or a slightly more visually engaging example [here].

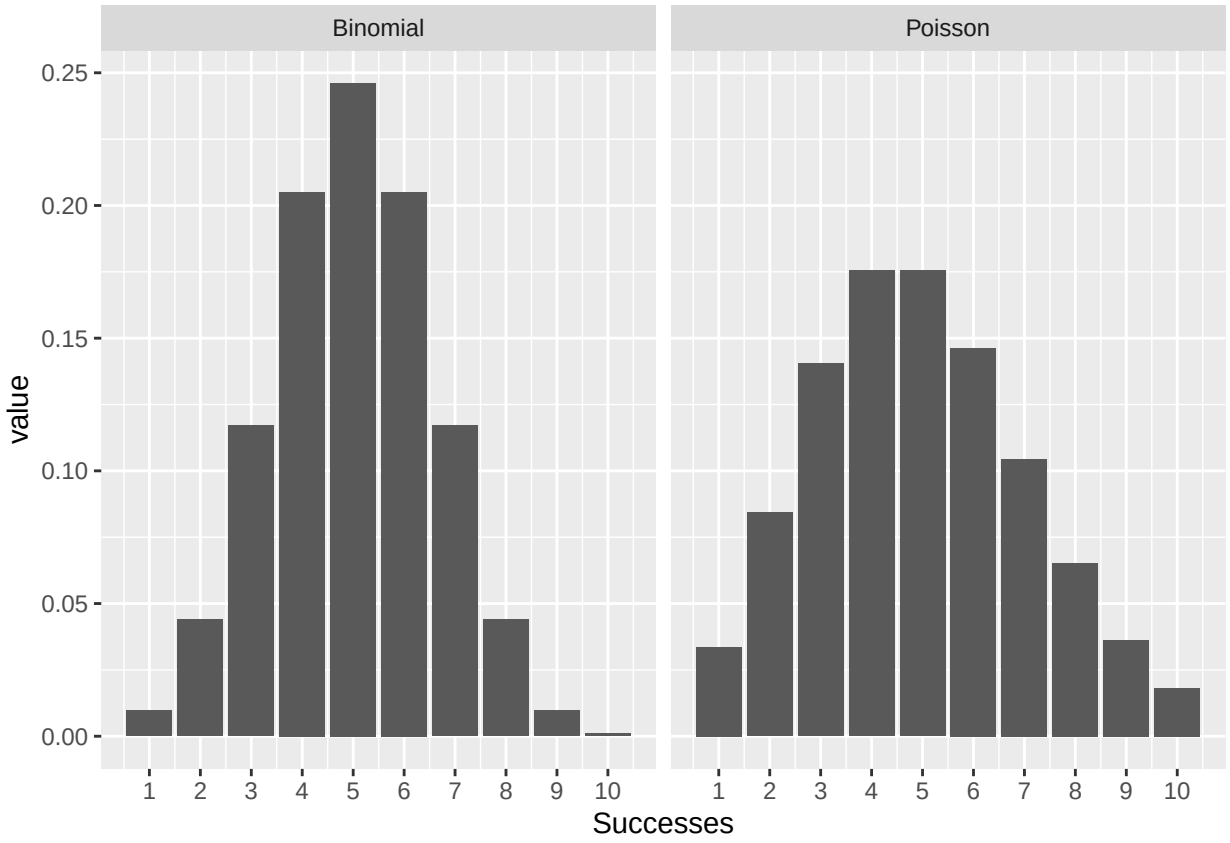
Question 3

Plot the probability distributions for the number of successes out of 10 trials, with a success probability of 0.5; using the Binomial $X_B \sim Binom(10, 0.5)$ and Poisson $X_P \sim Pois(10 \times 0.5)$.

```r
n <- 10
p <- 0.5

number_of_successes <- 1:n

# Generate the dataframe with 3 columns: Successes, Binomial, Poisson
data <- data.frame(Successes = number_of_successes,
                   Binomial = dbinom(number_of_successes,n,p),
                   Poisson = dpois(number_of_successes,n*p))
# Plot side-by-side plots
data %>%
  pivot_longer(cols = -c(Successes), names_to = "Distribution") %>%
  ggplot(aes(x = Successes, y = value))+
  scale_x_continuous(breaks=data$Successes)+
  geom_bar(stat="identity")+
  facet_wrap(~Distribution)
```



Extra

Embed an image

Use `![](path\to\image)`. Replace the `path\to\image` with the actual path to your image. For example. I have a `QUT_Logo.jpg` in the same folder as my `R Markdown` file, which can be added using `![](QUT_Logo.jpg)`.

R Markdown Cheatsheet

Access RStudio's R Markdown Cheatsheet [Here].