Tutorial 4 Worksheet AY 22/23 Sem 1

DSA2101

Practice with Web Scraping

Read in the English Premier League (football) scores from http://www.worldfootball.net for the seasons 2010/2011 and 2021/2022 (just the two seasons). You will probably have to navigate to England > Schedule or Results and Tables to see the table. Your data frame should have the following columns at this point:

X1	X2	X3	X4	X5	X6	X7	season
14/08/2010	12:45	Tottenham Hotspur	-	Manchester City	0:0 (0:0)	NA	2010-2011
	15:00	Wigan Athletic	-	Blackpool FC	0:4 (0:3)	NA	2010-2011
	15:00	Bolton Wanderers	-	Fulham FC	0:0 (0:0)	NA	2010-2011
	15:00	Wolverhampton Wanderers	-	Stoke City	2:1 (2:0)	NA	2010-2011
	15:00	Aston Villa	-	West Ham United	3:0 (2:0)	NA	2010-2011
	15:00	Blackburn Rovers	-	Everton FC	1:0 (1:0)	NA	2010-2011

Cleaning up the dataset

If you are having problems scraping the tables, work with the prepared dataset on Canvas: `epl_raw_2010_2021.rds`. As you can see, there are issues with the dataset. Address the following issues

1. Fill up the empty date fields with the most recent date above. For instance, Rows 2 - 6 should all contain `14/08/2010` in column `X1`.
2. Split up the column with scores - the scores in parentheses are half-time scores. The scores outside parentheses are the full-time scores of each match.

Your final table should resemble this, with 760 rows in it.

date	time	home	away	scores	season	final_home	final_away	half_home	half_away
14/08/2010	12:45	Tottenham Hotspur	Manchester City	0:0 (0:0)	2010-2011	0	0	0	0
14/08/2010	15:00	Wigan Athletic	Blackpool FC	0:4 (0:3)	2010-2011	0	4	0	3
14/08/2010	15:00	Bolton Wanderers	Fulham FC	0:0 (0:0)	2010-2011	0	0	0	0
14/08/2010	15:00	Wolverhampton Wanderers	Stoke City	2:1 (2:0)	2010-2011	2	1	2	0

It is common, at least to begin with, to model the number of goals scored in a football match using a Poisson distribution. With continuous distributions, we can use qq-plots to inspect how well the distribution fits to the data. But these are not appropriate for discrete distributions. In this tutorial, we shall see how we can use something known as a *rootogram* to assess the fit visually.

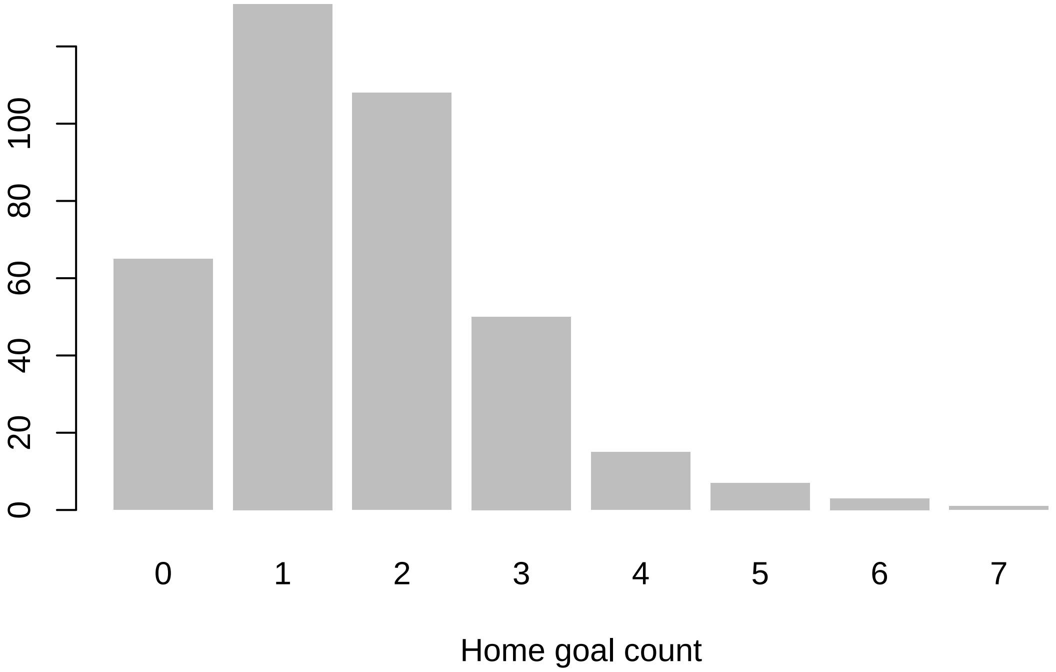
A rootogram displays the observed counts against the expected counts from a fitted distribution.

Formally speaking:

Let X_i be the number of **home goals** scored in a game in the 2010-2011 season, i ranging from 1 to 380. We assume that $X_i \sim Pois(\lambda)$. We are going to find the Maximum Likelihood Estimate (MLE) for λ and then assess how well that distribution fits to the data.

To prepare the data for the rootogram, perform the following steps:

1. Extract the data for the 2010-2011 season and store it in a data frame `s1011`.
2. Tabulate the number of matches where 0 home goals, 1 home goal, 2 home goals, etc. were scored. In other words, let $N_j = \sum_{i=1}^{380} I(X_i = j)$, for $j = 0, 1, 2, \ldots, 7$. These are the observed counts. At this point, you should be able to make this bar chart:



3. Compute the mean number of **home goals** scored in this 2010-2011 season. Store it in a (scalar) variable `lam_est`. This is the MLE for λ.

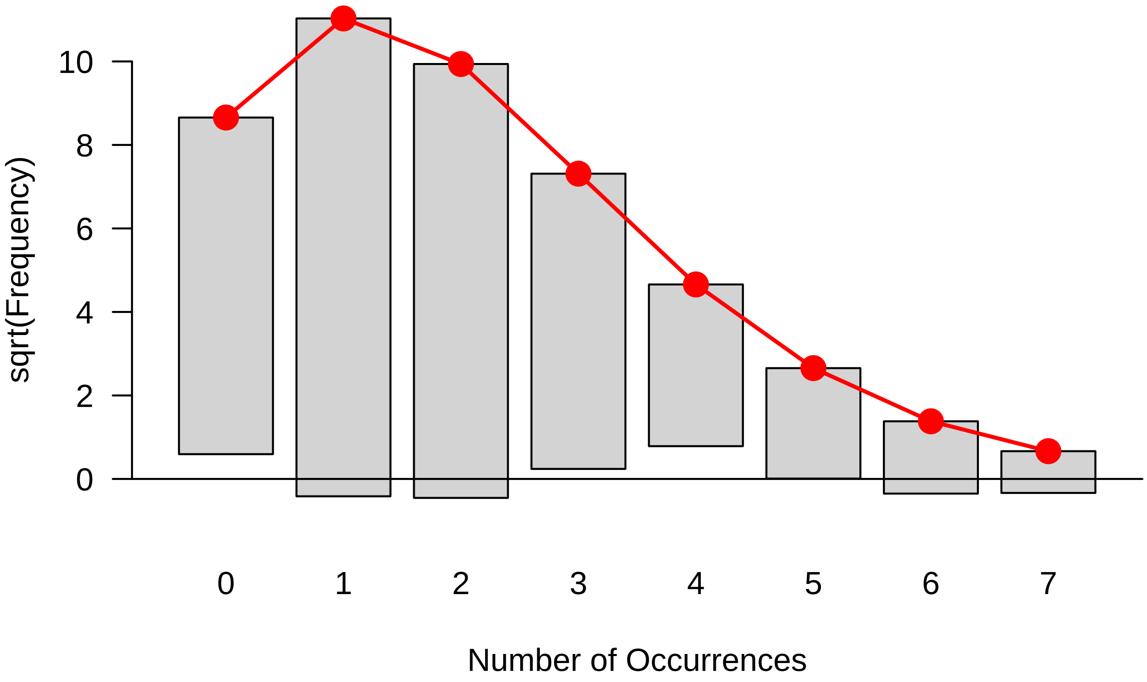
4. Using the estimated lambda, compute the expected number of games with 0, 1, 2, etc. home goals. In other words, compute

$$\hat{N}_j = 380 \times P(X_i = j \mid \lambda = \hat{\lambda}), \quad j = 0, 1, \ldots, 7$$

   Store your vector as `exp_goal_count`.

5. Now, we can construct the rootogram with:

```
library(vcd)
rootogram(table(s1011$final_home), exp_goal_count)
```



The gray bars are the observed counts from earlier. The red lines represented the counts using the fitted distribution. The mis-alignments with the horizontal axis denote residuals. With the rootogram, we can tell where our fitted model is not appropriate, and in which direction.

Construct rootograms for the home and away goals in both seasons, and think about the following:

- Is the Poisson appropriate for all of them?
- We can visually assess things now, but how can we make things more objective, i.e. to decide if the fit is acceptable or not?
- Why is it a square-root scale on the vertical axis?