BDM600 - Lab 4 - Group 8

Ran Arino; Zubeka Dane Dang; Solmaz Heidar Nassab

2024-02-19

Credentials

All members participated in this lab assignment. The file is the merged version.

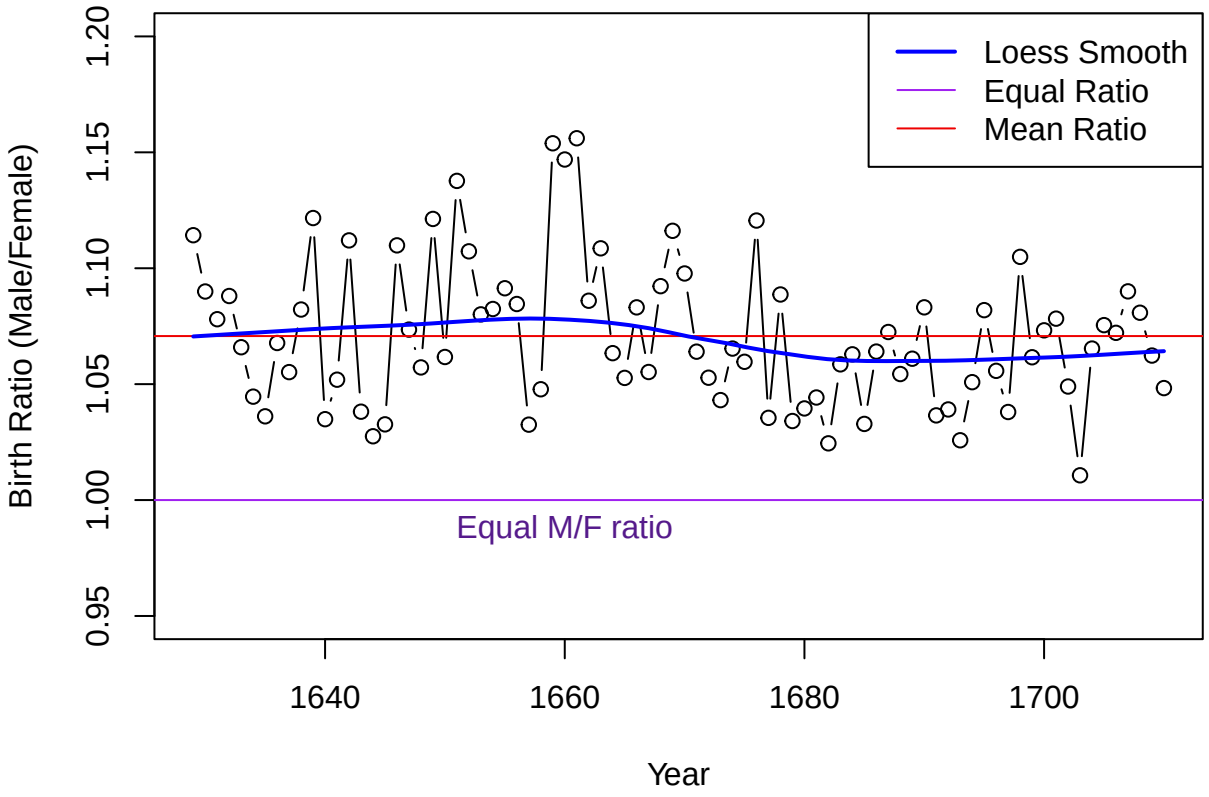
Start

```
# install libraries
library(ggplot2)
library(HistData)
library(grid)
library(gnm)
library(vcd)
library(vcdExtra)
library(MASS)
```

3.1

(a): Male ratio

```
data("Arbuthnot", package = "HistData")
# Setplot margins
par(mar = c(5, 4, 1, 1) + .1)
# Plot Ratio over Year
with(Arbuthnot, {
    plot(Year, Ratio, type='b', ylim = c(.95, 1.2), ylab = "Birth Ratio (Male/Female)")
    # Add horizontal lines
    abline(h = 1, col = "purple", lwd = 1)
    abline(h = mean(Ratio), col = "red2")
    # Add smoothed line
    Arb.smooth <- loess.smooth(Year, Ratio)
    lines(Arb.smooth$x, Arb.smooth$y, col = "blue", lwd = 2)
    # Add annotation
    text(x = 1660, y = 1, "Equal M/F ratio", pos = 1, col = "purple4")
    # Add legend
    legend("topright", legend = c("Loess Smooth", "Equal Ratio", "Mean Ratio"), col = c("blue", "purple
})
```



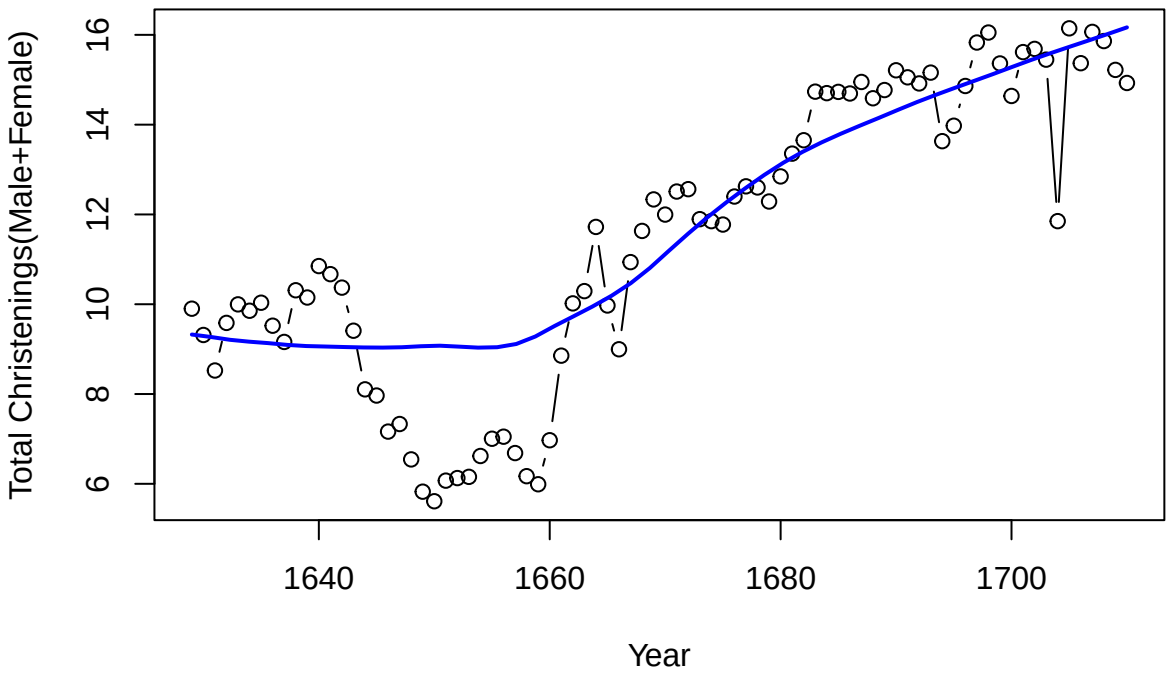
Observations: - Fluctuation shows the variablity in the male ratio. - Overall, male births are relatively larger than female. - Slight downtrend since the the data peaked in 1680.

Preferred plots: the line plots is preferable if the focus is on the relative comparison of make to female as a ratio.

(b): Total

```
Arbuthnot$TotalNum <- Arbuthnot$Males + Arbuthnot$Females

with(Arbuthnot,{
  plot(Year, Total, type = "b", ylab = "Total Christenings(Male+Female)")
  # Add smoothed line
  Arb.smooth <- loess.smooth(Year, Total)
  lines(Arb.smooth$x, Arb.smooth$y, col = "blue", lwd = 2) })
```



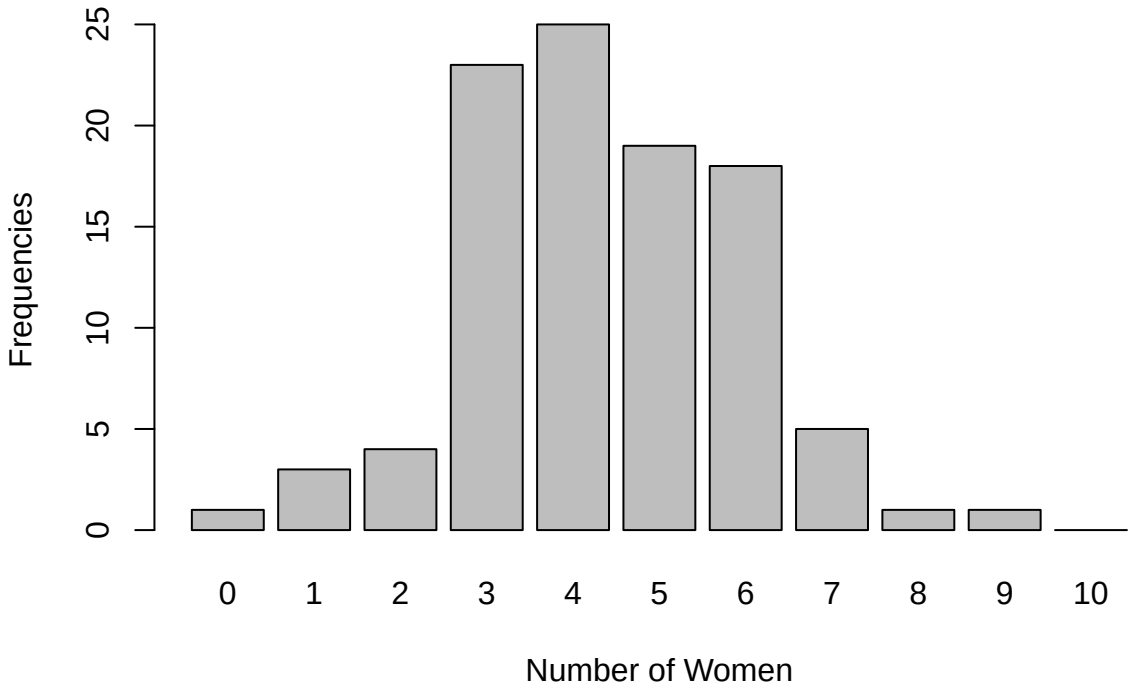
Observations: - Clear uptrend since 1660 after the bottom around 1650-1660. - Huge rise occurred from 1661 to 1664. - The unsual behavior could be the sudden decline in 1704.

3.3

```
# install data
data("WomenQueue")
```

(a): show frequency

```
barplot(WomenQueue, xlab = "Number of Women", ylab = "Frequencies")
```



(b): check GOF for binomial

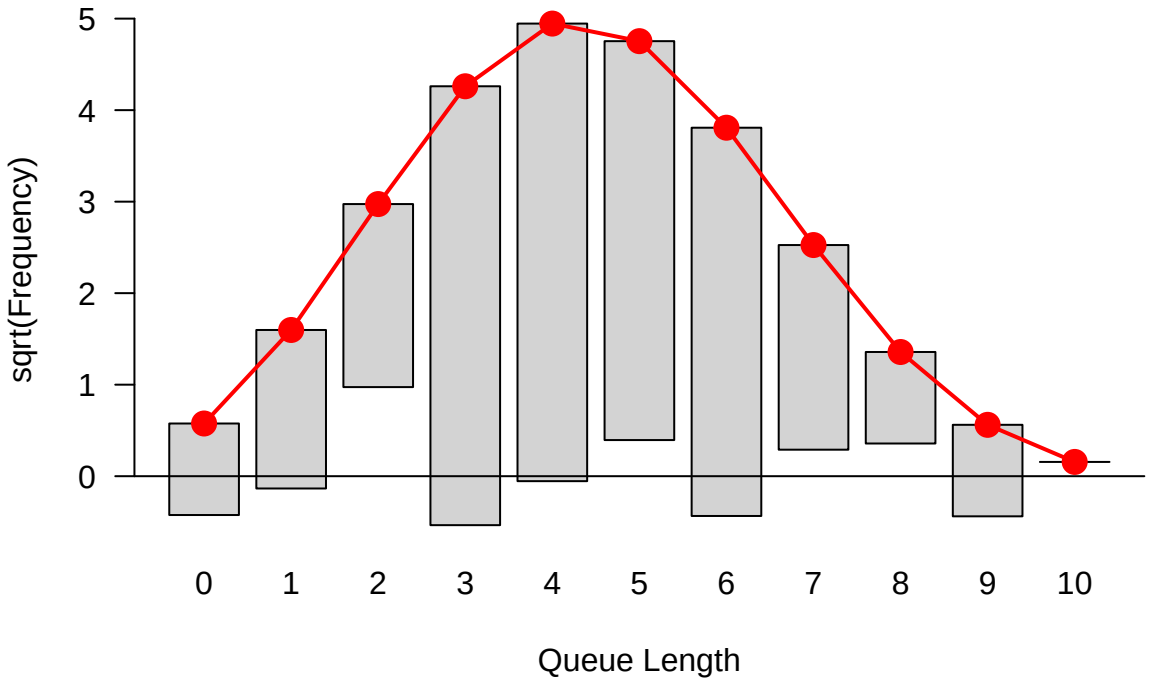
```
gof_fit01 <- goodfit(WomenQueue, type = "binomial", par = list(size = 10))
summary(gof_fit01)
```

```
##
##   Goodness-of-fit test for binomial distribution
##
##                       X^2 df  P(> X^2)
## Likelihood Ratio 8.650999  8 0.3725869
```

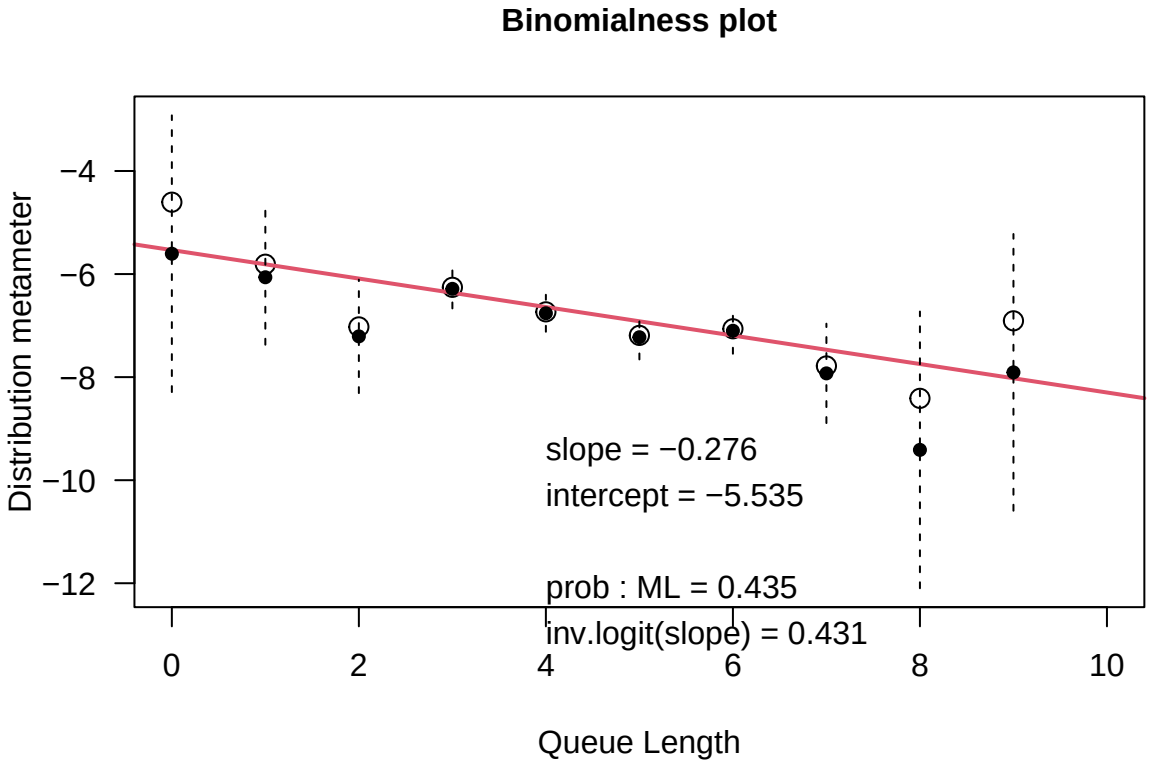
The goodness-of-fit test shows that the WomenQueue data will fit the binomial distribuiton by 37% of chance.

(c): Resonable plots

```
plot(gof_fit01, xlab = "Queue Length")
```




```
distplot(WomenQueue, type = "binomial", size = 10, xlab =  "Queue Length")
```



(d): Why does frequency distribution depart from binomial?

- It may be affected by various factors, including society, economy, and condition.
- The distribution may be affected by sampling biases or measurement errors.
- Thus, the actual results departs from the ideally expected chance of 50%.

3.4

(a): GOF test

```
data("Saxony")
gof_fit02 <- goodfit(Saxony, type = "binomial", par = list(size = 12, prob = .5))
ss1 <- summary(gof_fit02)
```

```
## Warning in summary.goodfit(gof_fit02): Chi-squared approximation may be
## incorrect
```

```
##
##   Goodness-of-fit test for binomial distribution
##
##                       X^2 df     P(> X^2)
## Pearson          249.1954 12 2.013281e-46
## Likelihood Ratio 205.4060 12 2.493625e-37
```

```r
# ratio of chi-square / df
ss1[,"X^2"] / ss1[,"df"]
```

```
##          Pearson Likelihood Ratio
##         20.76629         17.11717
```

Observations: - The extremely small p-values showed that the data differs from the binomial distribution under the specified parameters. - In terms of the ratios of chi-square, both are significantly greater than 1, which indicates a large deviation of the observed data from what we are expecting under the binomial distribution.

(b): Test additional lack of fit

```r
gof_fit03 <- goodfit(Saxony, type = "binomial", par = list(size = 12))
ss2 <- summary(gof_fit03)
```

```
##
##   Goodness-of-fit test for binomial distribution
##
##                       X^2 df     P(> X^2)
## Likelihood Ratio 97.0065 11 6.978187e-16
```

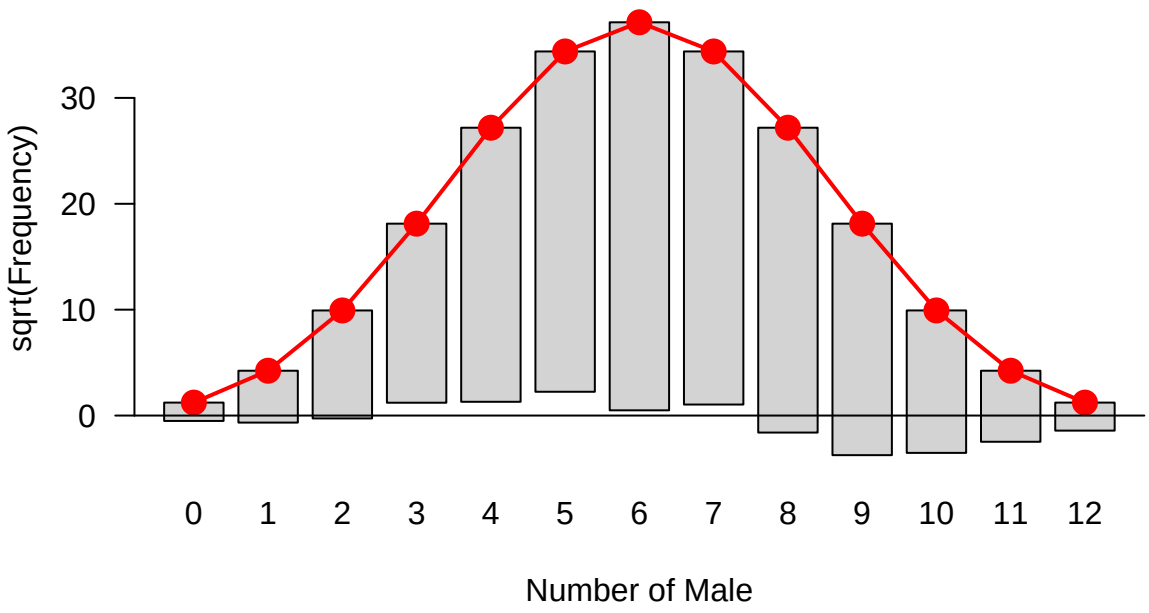
```r
# ratio of chi-square / df
ss2[,"X^2"] / ss2[,"df"]
```

```
## [1] 8.818773
```

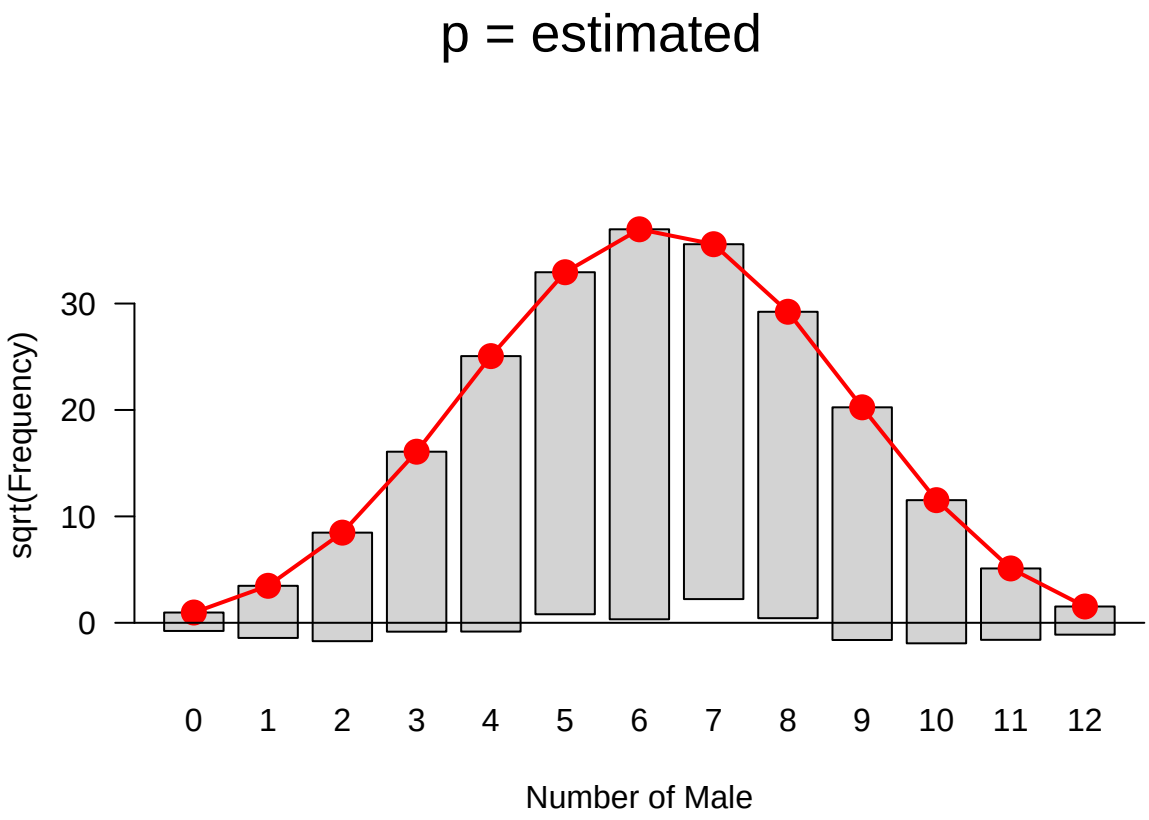
Observations: - The p-value was still small, so the result remained to suggest the observed data is different from the expected binomial distribution. - However, compared to the assumed conditions ($p = 0.5$) previously, both the p-value and statistics value slightly increased.

(c): Visualization

```r
plot(gof_fit02, main = "p = 0.5", xlab = "Number of Male")
```




```
plot(gof_fit03, main = "p = estimated", xlab = "Number of Male")
```



3.6

(a): Construct One-way table

```
counts <- 0:5
frequencies <- c(129, 83, 20, 9, 5, 1)

# Combine into a data frame
data <- data.frame(counts, frequencies)
# Construct one-way table
table <- xtabs(frequencies ~ counts, data=data)
table
```

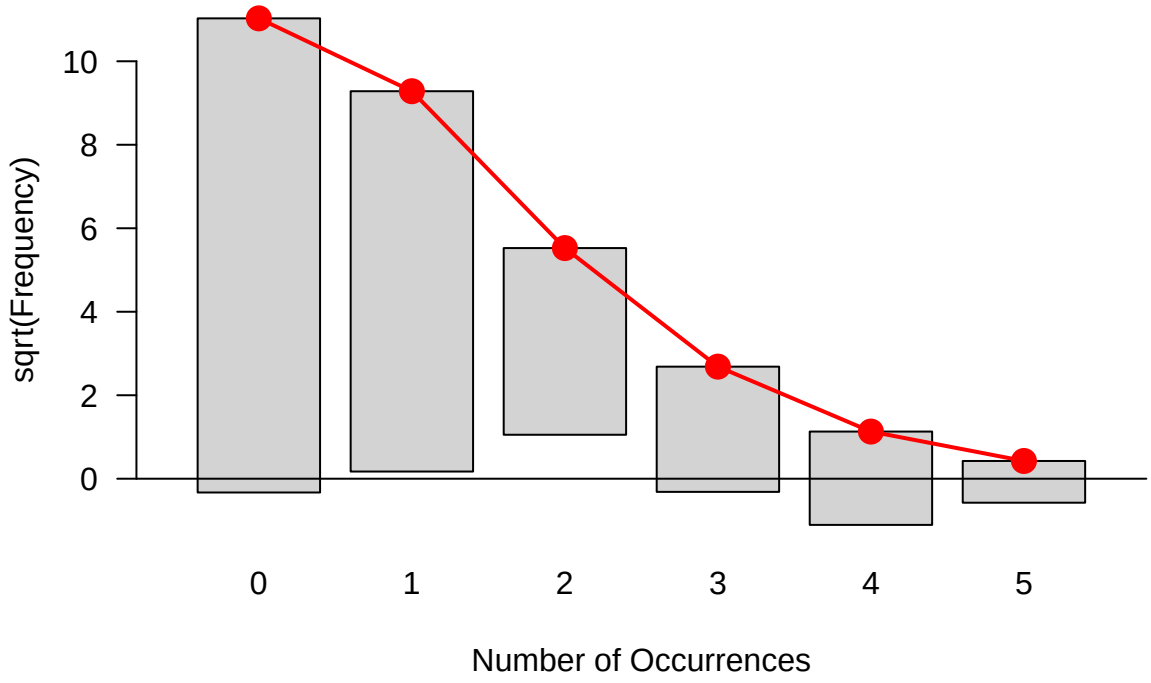
```
## counts
##   0   1   2   3   4   5
## 129  83  20   9   5   1
```

(b): Fit data to Poisson

```
gof_fit04 <- goodfit(table, type = "poisson")
summary(gof_fit04)
```

```
##
##   Goodness-of-fit test for poisson distribution
##
##                       X^2 df   P(> X^2)
## Likelihood Ratio 13.13892  4 0.01061657
```

```
plot(gof_fit04)
```

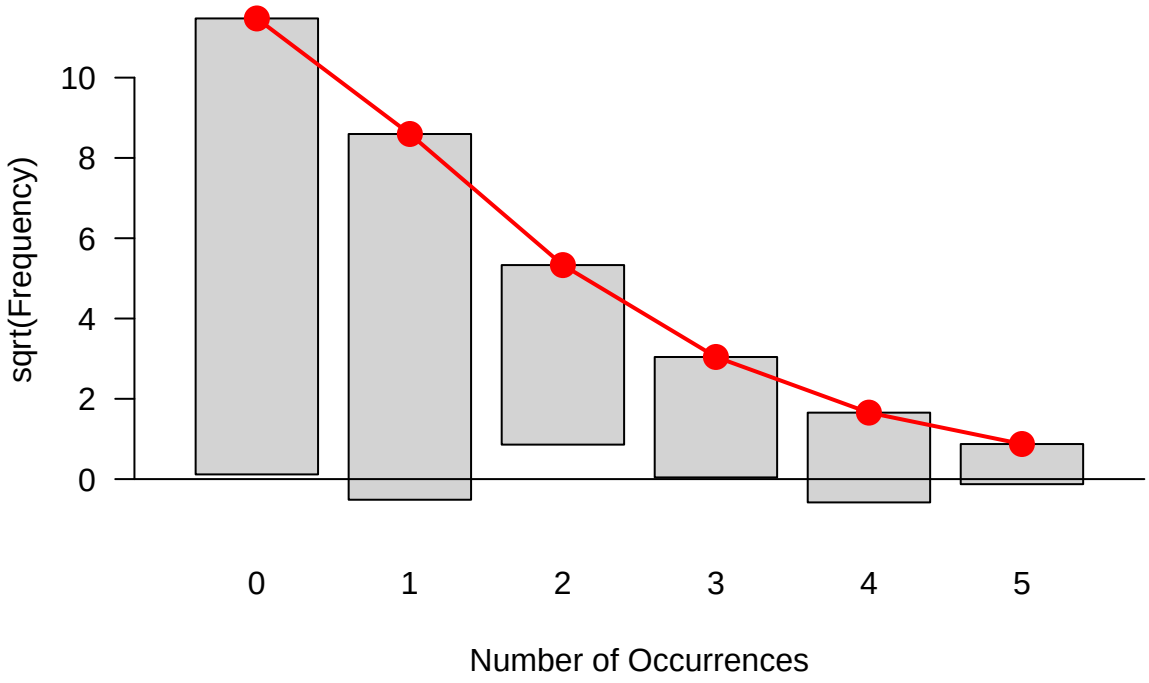


(c): Fit data to Negative Binomial

```
gof_fit05 <- goodfit(table, type = "nbinomial")
summary(gof_fit05)
```

```
##
##   Goodness-of-fit test for nbinomial distribution
##
##                       X^2 df  P(> X^2)
## Likelihood Ratio 6.030625  3 0.1101297
```

```
plot(gof_fit05)
```



(d): Conclusion

- Our defined data did not perfectly fit neither Poisson and Negative Binomial Distribution.
- Both p-values are relatively small, which means that the probability of fitting the data into each distribution is significantly low.
- However, the observed data might fit to the negative binomial distribution with 10% chance.

3.7

(a): Load data

```
data("Geissler", package = "vcdExtra")
size11 <- subset(Geissler, Geissler$size == 11)
size11_boys <- xtabs(Freq ~ boys, data=size11)
size11_boys
```

```
## boys
##    0    1    2    3    4    5    6    7    8    9   10   11
##    8   72  275  837 1540 2161 2310 1801 1077  492   93   24
```

(b): Distribution fit

```
gof_fit06 <- goodfit(size11_boys, type = "binomial")
```

```
## Warning in goodfit(size11_boys, type = "binomial"): size was not given, taken
## as maximum count
```

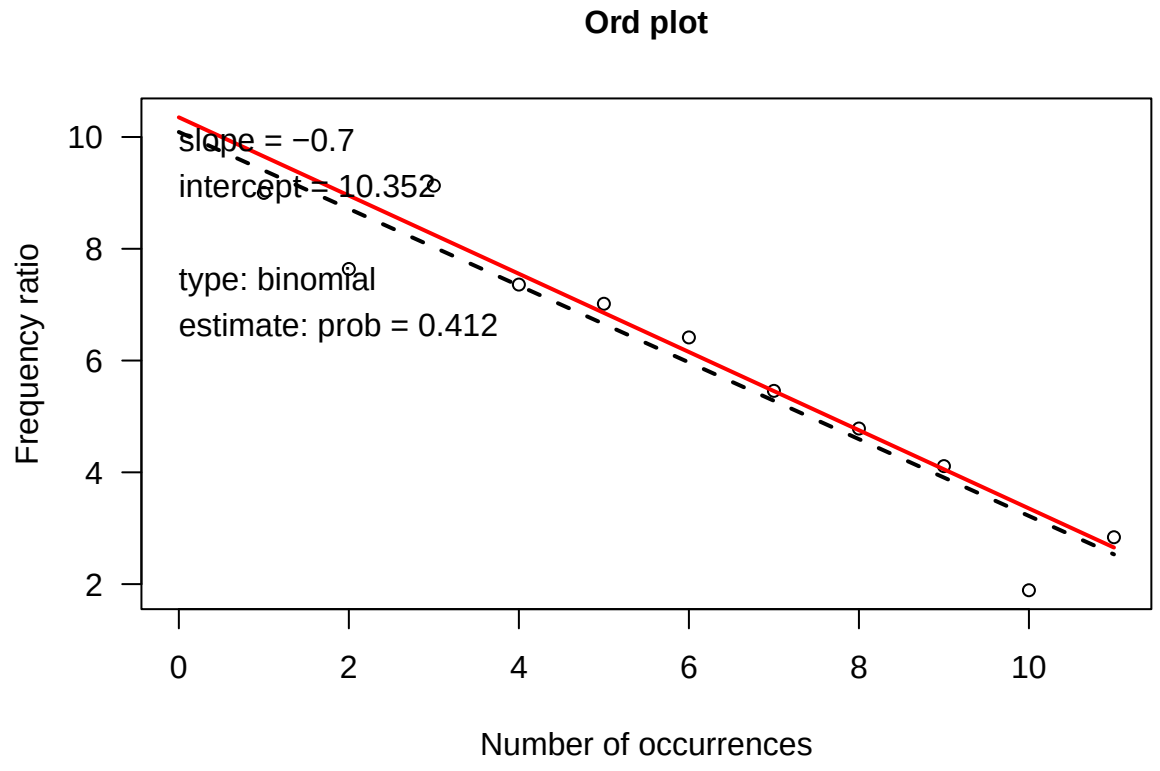
```
summary(gof_fit06)
```

```
##
##   Goodness-of-fit test for binomial distribution
##
##                       X^2 df     P(> X^2)
## Likelihood Ratio 148.0892 10 9.212554e-27
```

The p-value is significantly small, which means that the probability of fitting this data into binomial distribuiton is close to zero.

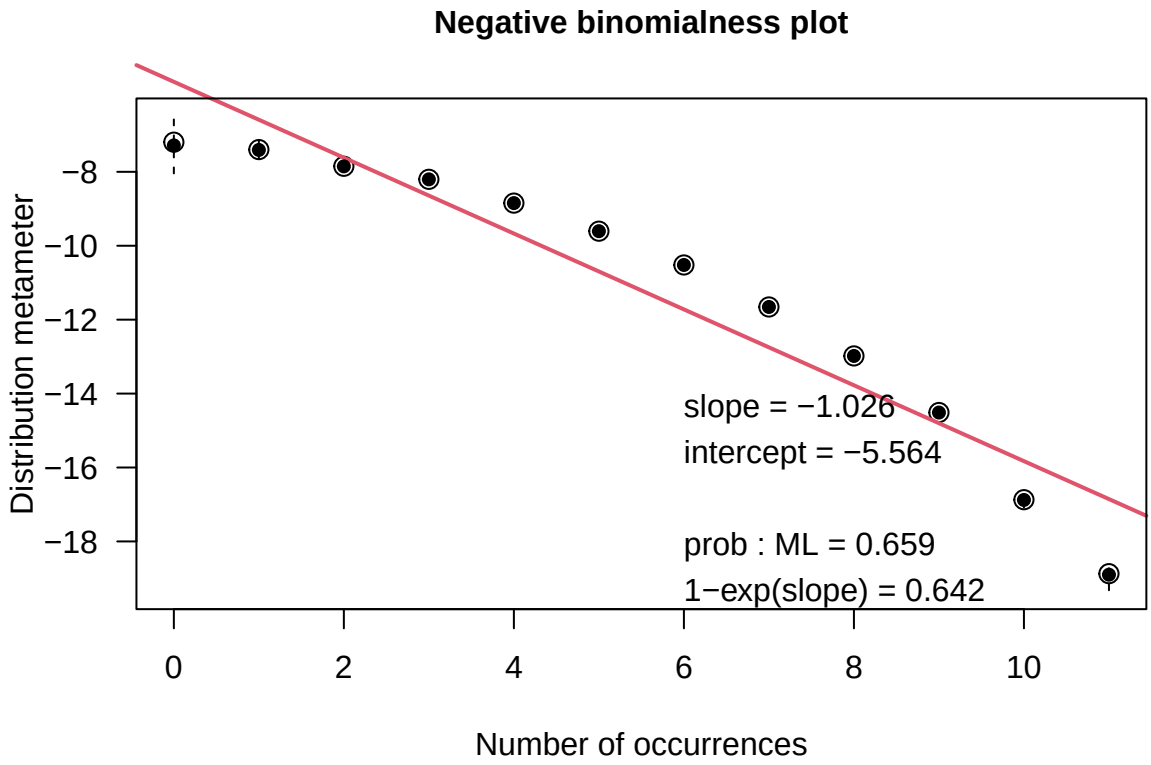
(c): Diagnose the distribution form

```
Ord_plot(size11_boys)
```



(d): Fit negative binomial

```
distplot(size11_boys, type = "nbinomial", size = 11)
```



The negative binomial distribution does not perfectly fit the data due to the divergence between an ideal line and actual points.

3.8

```
data("Bundesliga", package = "vcd")
BL1995 <- xtabs(~ HomeGoals + AwayGoals,data= Bundesliga, subset=(Year == 1995))

BL.df <- as.data.frame(BL1995, stringsASFactors = FALSE)
BL.df <- within(BL.df, {
  HomeGoals <- as.numeric(HomeGoals)
  AwayGoals <- as.numeric(AwayGoals)
  TotalGoals <- HomeGoals + AwayGoals
})
BL.df
```

```
##    HomeGoals AwayGoals Freq TotalGoals
## 1          1         1   26          2
## 2          2         1   19          3
```

```
## 3          3          1   27          4
## 4          4          1   14          5
## 5          5          1    3          6
## 6          6          1    4          7
## 7          7          1    1          8
## 8          1          2   16          3
## 9          2          2   58          4
## 10         3          2   23          5
## 11         4          2   11          6
## 12         5          2    5          7
## 13         6          2    1          8
## 14         7          2    0          9
## 15         1          3   13          4
## 16         2          3   20          5
## 17         3          3   20          6
## 18         4          3   10          7
## 19         5          3    3          8
## 20         6          3    0          9
## 21         7          3    0         10
## 22         1          4    5          5
## 23         2          4    5          6
## 24         3          4    5          7
## 25         4          4    4          8
## 26         5          4    0          9
## 27         6          4    1         10
## 28         7          4    1         11
## 29         1          5    0          6
## 30         2          5    4          7
## 31         3          5    1          8
## 32         4          5    2          9
## 33         5          5    0         10
## 34         6          5    0         11
## 35         7          5    0         12
## 36         1          6    1          7
## 37         2          6    0          8
## 38         3          6    1          9
## 39         4          6    0         10
## 40         5          6    0         11
## 41         6          6    0         12
## 42         7          6    0         13
## 43         1          7    0          8
## 44         2          7    1          9
## 45         3          7    1         10
## 46         4          7    0         11
## 47         5          7    0         12
## 48         6          7    0         13
## 49         7          7    0         14
```

(a): Find one-way distribution

```
# for HomeGoals
HomeGoals_dist <- xtabs(Freq ~ HomeGoals, data = BL.df)
```

```
HomeGoals_dist
```

```
## HomeGoals
##   1   2   3   4   5   6   7
##  61 107  78  41  11   6   2
```

```
# for AwayGoals
AwayGoals_dist <- xtabs(Freq ~ AwayGoals, data = BL.df)
AwayGoals_dist
```

```
## AwayGoals
##   1   2   3   4   5   6   7
##  94 114  66  21   7   2   2
```

```
# for TotalGoals
TotalGoals_dist <- xtabs(Freq ~ TotalGoals, data = BL.df)
TotalGoals_dist
```

```
## TotalGoals
##  2  3  4  5  6  7  8  9 10 11 12 13 14
## 26 35 98 62 39 29 10  4  2  1  0  0  0
```

(b): Fit Poisson

```
# Home
gof_fit07 <- goodfit(HomeGoals_dist, type = "poisson")
summary(gof_fit07)
```

```
##
##   Goodness-of-fit test for poisson distribution
##
##                       X^2 df     P(> X^2)
## Likelihood Ratio 70.7216  5 7.251639e-14
```

```
# Away
gof_fit08 <- goodfit(AwayGoals_dist, type = "poisson")
summary(gof_fit08)
```

```
##
##   Goodness-of-fit test for poisson distribution
##
##                        X^2 df     P(> X^2)
## Likelihood Ratio 97.97289  5 1.413068e-19
```

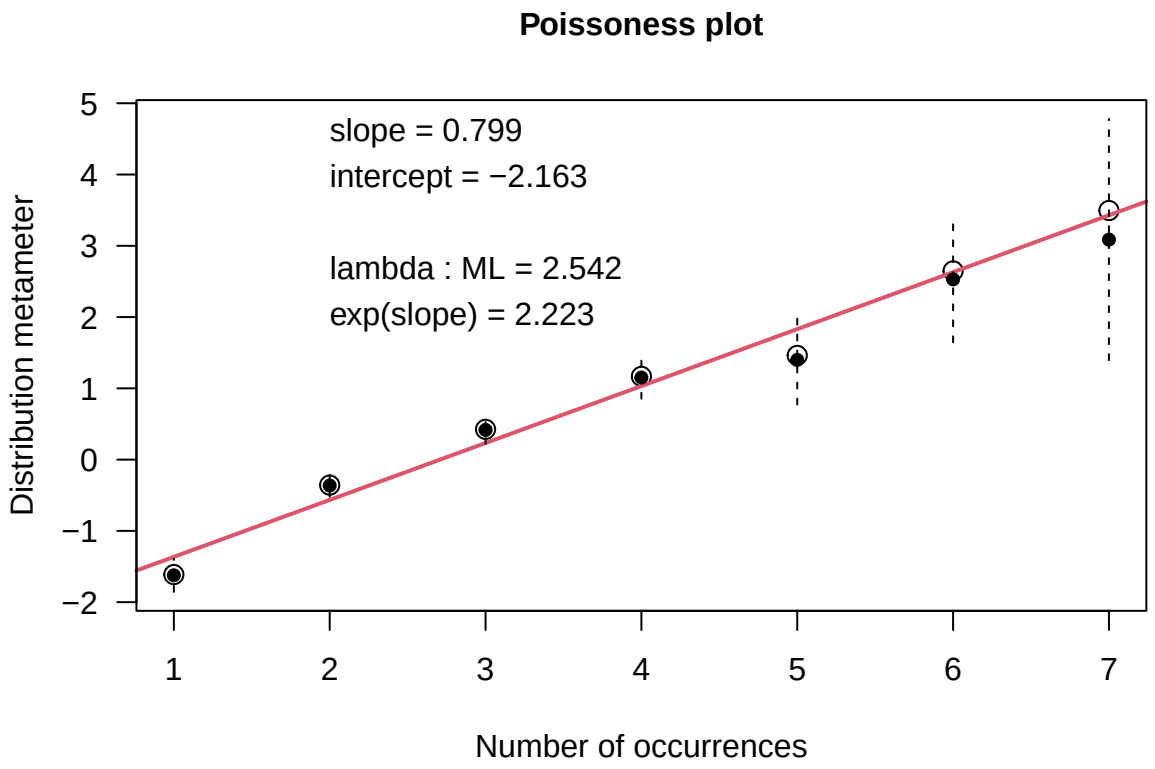
```
# Total
gof_fit09 <- goodfit(TotalGoals_dist, type = "poisson")
summary(gof_fit09)
```

```
##
##   Goodness-of-fit test for poisson distribution
##
##                       X^2 df     P(> X^2)
## Likelihood Ratio 72.55804  8 1.518508e-12
```

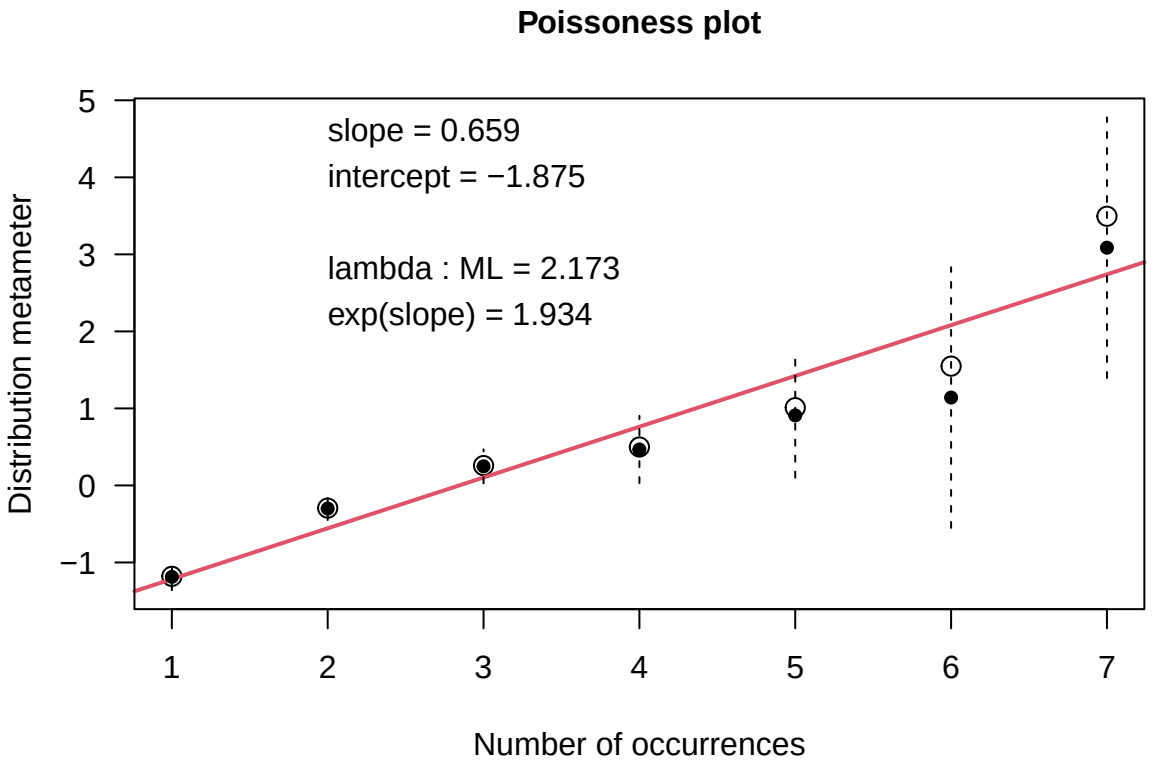
- Home: The data may reasonably fit the Poisson distribution compared to the other two distributions.
- Away: The data might fit the Poisson distribution, but its chance is relatively low.
- Total: The data does not fit the Poisson distribution.

(c): Use distplot()

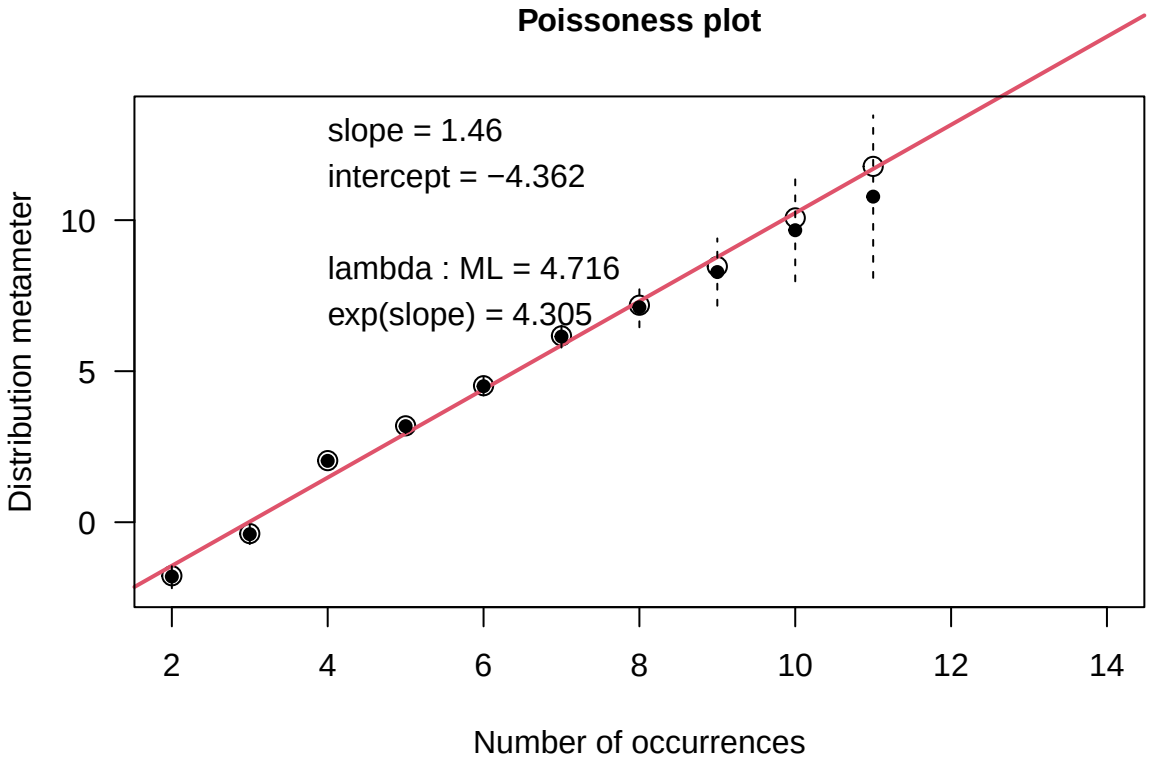
```
distplot(HomeGoals_dist, "poisson")
```




```
distplot(AwayGoals_dist, "poisson")
```




```
distplot(TotalGoals_dist, "poisson")
```



- Home: There are smaller differences between points and a line, the data is likely to fit the Poisson distribution. - Away: The difference between points and a line is bigger than Home, so it might conclude that the data fails to fit the Poisson distribution. - The points are close to the ideal line, so it could indicate the data in total goals might follow the Poisson distribution. However, considering that the significantly lower p-value shows huge discrepancies between observed data nad the expected values based on Poisson, so the data does not reasonably fits to Poisson.

(d): What circumstances of scoring goals in soccer might cause these distributions to deviate from Poisson distributions?

- Dependence of Goals: Goals are not always independent; the occurrence of one can affect the likelihood of another.
- Team's strategy: The team sometimes employs a significant defensive or offensive strategy, which could cause the huge number of goals.
- External Factors: Weather, pitch conditions, and crowd support can influence game dynamics in ways not accounted for by a constant rate of occurrence.
- Red cards: If the number of red cards in each game sinificantly affect the number of goals, causing variance that is not followed by a simple Poisson model.
- SKills: If there is no difference between two teams in terms of their skills, most matches would have low scores, which could cause a large deviation from the Poisson distribution. On the other hand, if there are huge difference between two teams (where half of teams are significantly strong team, but the rest of teams are significantly weak), it could cause the skewed games with larger scores in one team when strong vs weak team but with lower scores when strong vs strong team.