The following exercieses are from the following resource from page 109 to 112.

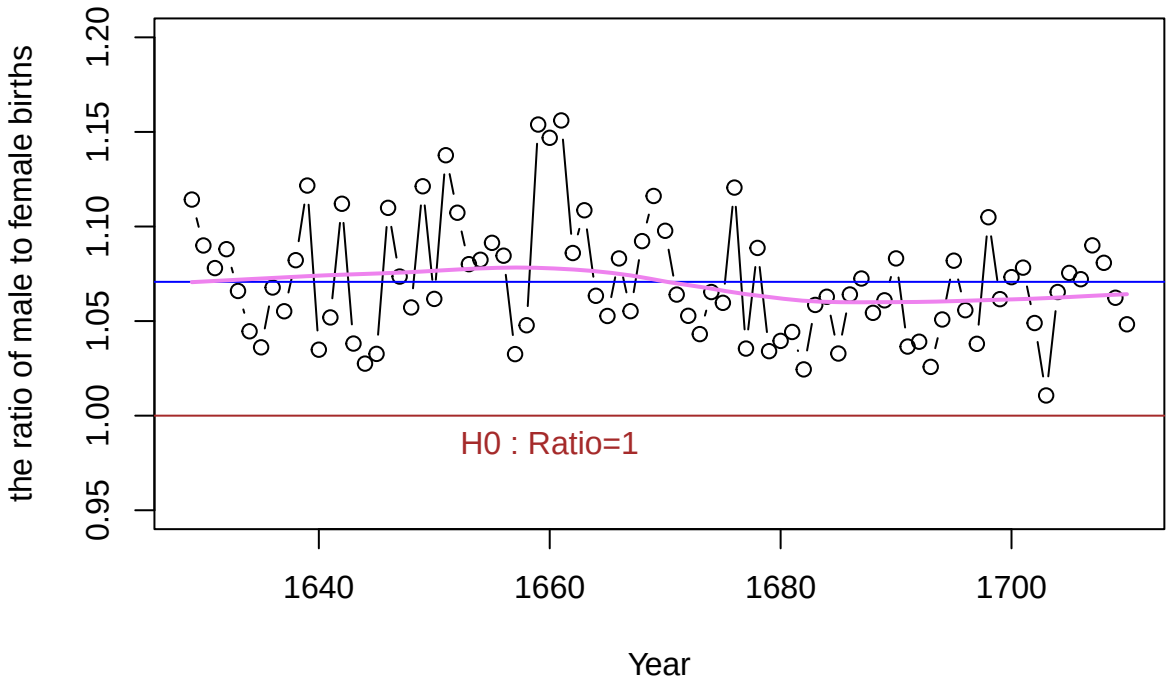
resource: [Friendly] Discrete Data Analysis with R: visualization and modeling techniques for categorical and count data, Michael Friendly, et al, ISBN: 978-1-4987-2583-5, CRC Press, 2015. [Available online at Seneca library]

Exercise 3.1 The Arbuthnot data in HistData (Example 3.1) also contains the variable Ratio, giving the ratio of male to female births.

(a) Make a plot of Ratio over Year, similar to Figure 3.1.

My Answer: What features stand out? The plot features horizontal lines and text annotation, effectively highlighting the null hypothesis and the average ratio. Simultaneously, the violet smooth line enhances the visualization by portraying the overall trend in the data, offering a concise yet comprehensive depiction of the male-to-female birth ratio dynamics over the years.Which plot do you prefer to display the tendency for more male births? The choice in presenting the trend towards more male births hinges on contextual considerations and the intended audience. The violet smooth line, by offering a clear visual representation of the overall trend, facilitates the identification of patterns in the data. However, if simplicity takes precedence, the scatterplot with horizontal lines could be deemed adequate. The decision ultimately relies on the specific analytical objectives and the level of detail suitable for effective communication with the targeted audience.

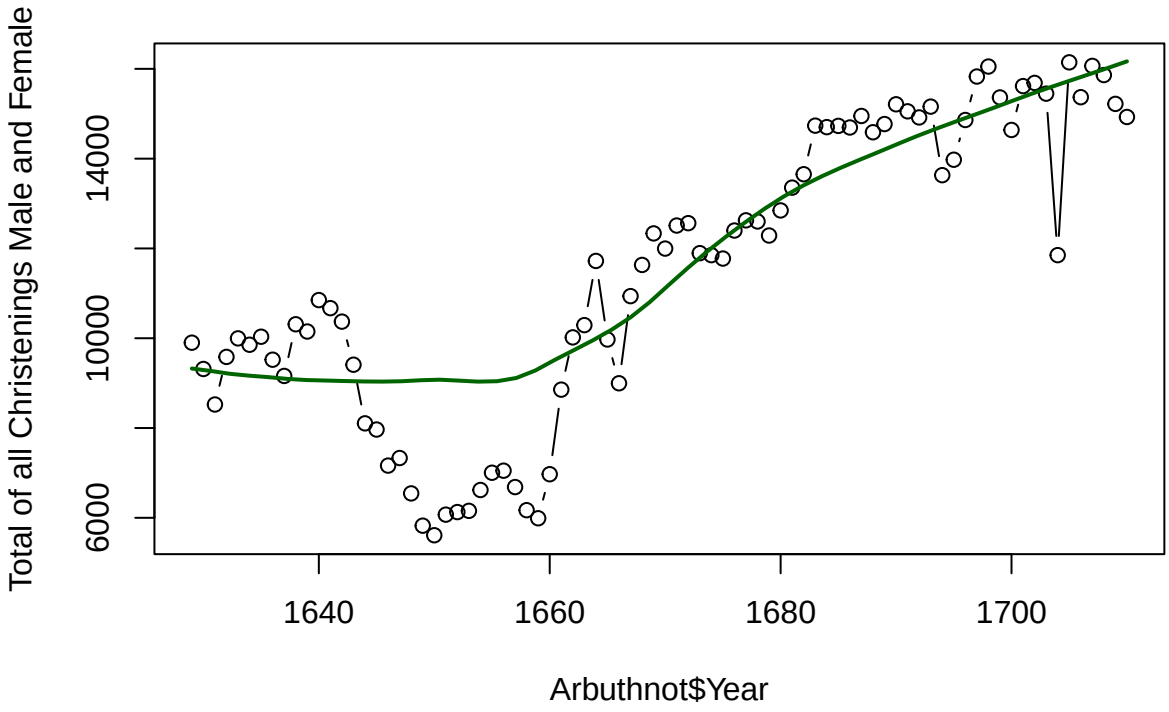
```r
# Access Arbuthnot from HistData package explicitly
data("Arbuthnot", package = "HistData")

# Extracting variables from Arbuthnot
Year <- Arbuthnot$Year
Ratio <- Arbuthnot$Ratio

# Create a plot for Ratio based on Year
plot(Year, Ratio, type='b', ylim=c(0.95, 1.2), ylab="the ratio of male to female births")

# Horizontal lines
abline(h=1, col="brown", lwd=1)
abline(h=mean(Ratio), col="blue")

# Text annotation
text(x=1660, y=1, "H0 : Ratio=1", pos=1, col="brown")

# Smooth line
Arb.smooth <- loess.smooth(Year, Ratio)
lines(Arb.smooth$x, Arb.smooth$y, col="violet", lwd=2)
```



(b) Plot the total number of christenings, Males + Females or Total (in 000s) over time.

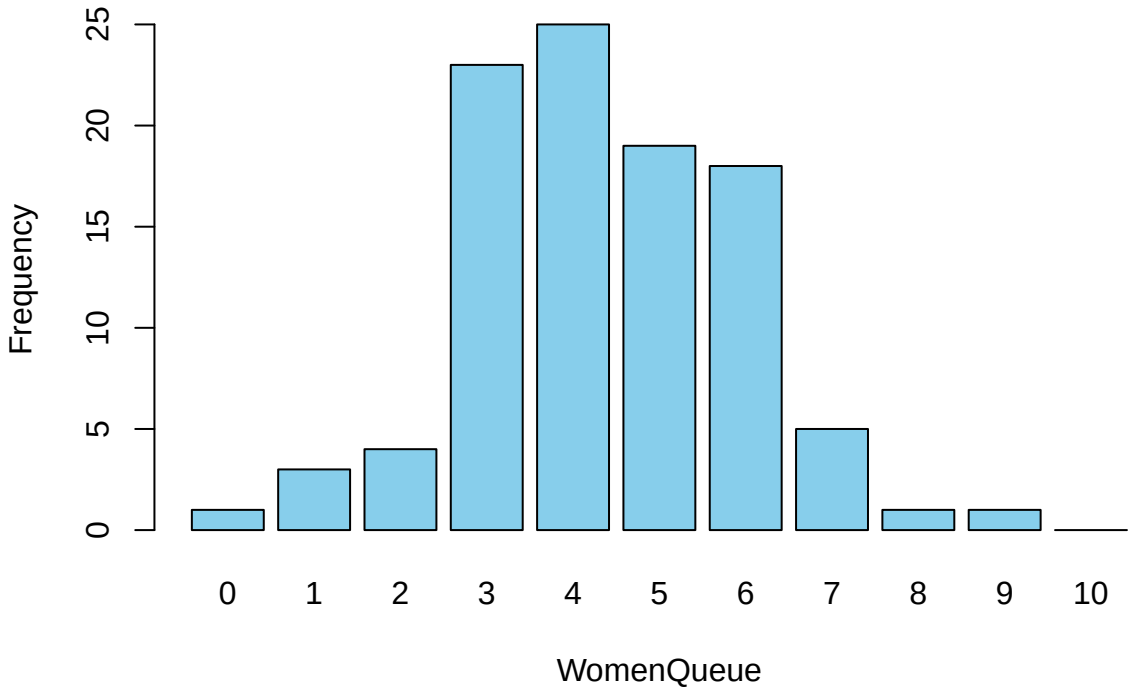
My Answer: What unusual features do you see? An intriguing observation is the sudden decline in both male and female christenings between 1640 and 1660. Additionally, a noteworthy trend is the continuous increase in the total number of christenings for both genders as the years progress.

```
# total number of Christenings
Arbuthnot$Total <- Arbuthnot$Males + Arbuthnot$Females
plot(Arbuthnot$Year, Arbuthnot$Total, type='b', ylab="Total of all Christenings Male and Female")
Arb.smooth <- loess.smooth(Arbuthnot$Year, Arbuthnot$Total)
lines(Arb.smooth$x, Arb.smooth$y, col="darkgreen", lwd=2)
```



Exercise 3.3 Use the data set WomenQueue to: (a) Produce plots analogous to those shown in Section 3.1 (some sort of bar graph of frequencies).

```
data("WomenQueue", package = "vcd")
barplot(WomenQueue, xlab = "WomenQueue", ylab = "Frequency", col = "skyblue")
```



(b) Check for goodness-of-fit to the binomial distribution using the goodfit() methods described in Section 3.3.2.

My Answer: With a likelihood ratio chi-squared statistic of 8.651 and 8 degrees of freedom, the resulting p-value is 0.3726. Given that this p-value exceeds the common significance level of 0.05, we lack sufficient evidence to reject the null hypothesis. In this scenario, the null hypothesis likely posits that the observed data adheres to a binomial distribution with the specified parameters. The test indicates that the observed data aligns well with the anticipated distribution, and any discrepancies might be attributed to random variability.

```
library(vcd)
```

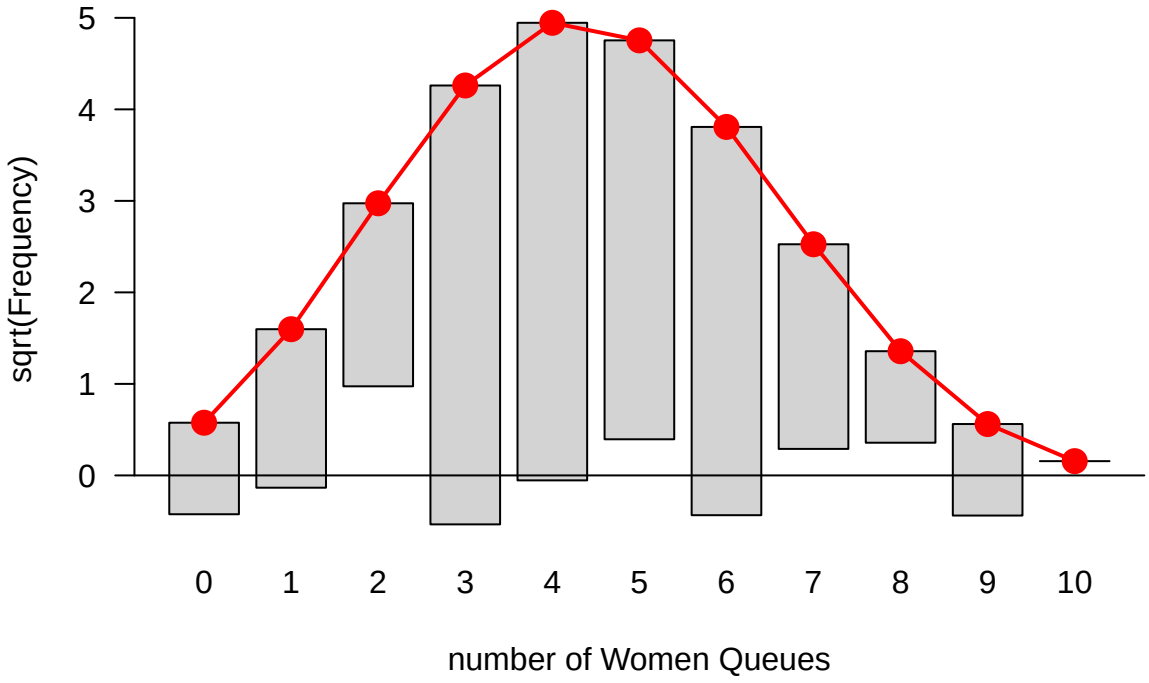
```
## Loading required package: grid
```

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

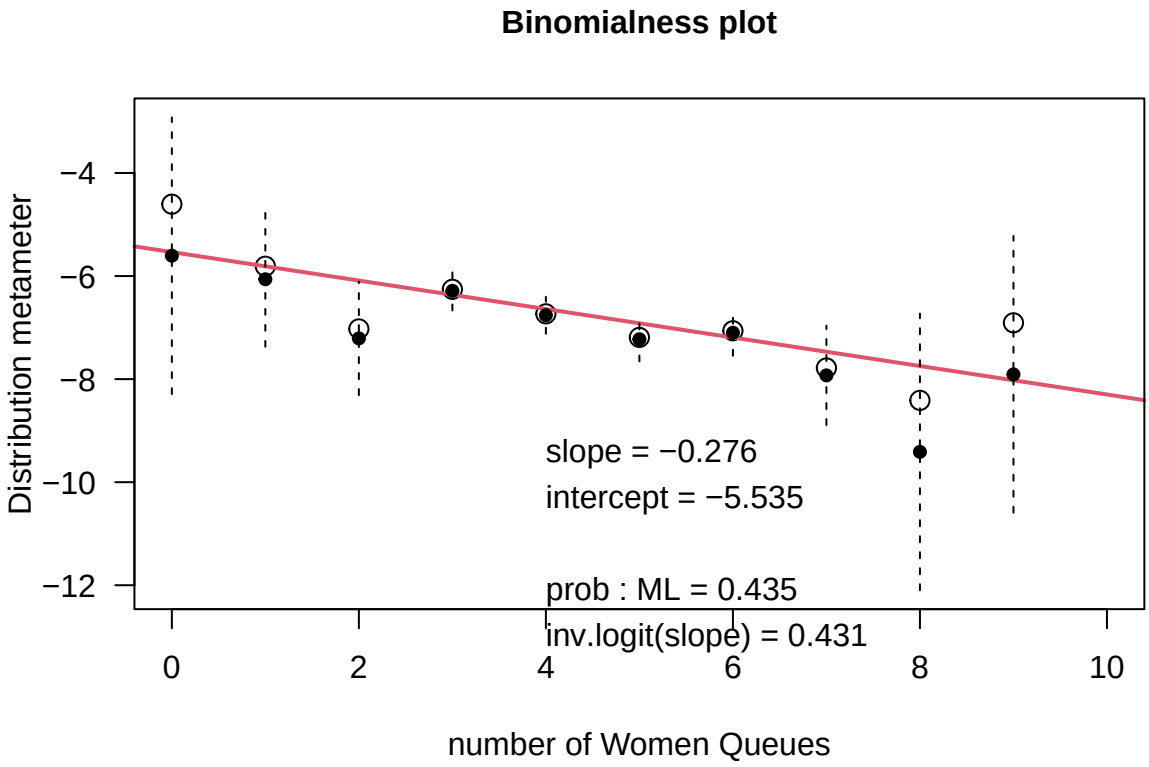
```
##
##   Goodness-of-fit test for binomial distribution
##
##                       X^2 df  P(> X^2)
## Likelihood Ratio 8.650999  8 0.3725869
```

(c) Make a reasonable plot showing departure from the binomial distribution.

```
plot(goodness_of_fit, xlab = "number of Women Queues")
```




```
distplot(WomenQueue, type = "binomial", size=10, xlab = "number of Women Queues")
```



(d) Suggest some reasons why the number of women in queues of length 10 might depart from a binomial distribution, Bin(n = 10; p = 1/2).

My Answer: 1. Unequal Gender Distribution (p Not Equal 1/2): If women (or men) are more prevalent in these queues, it would result in a departure from the expected binomial distribution with an equal probability (p = 1/2). An imbalance in gender composition within the queues could influence the observed distribution. 2. Non-Independence of Observations: The assumption of independence in a binomial distribution implies that each observation (individual joining the queue) is not influenced by or dependent on others. If people tend to join lines in groups or if there is some form of correlation between individuals in the queues, the independence assumption is violated, and this could lead to deviations from the expected binomial distribution. In real world the length of line and who are already in the line will influence people, so the observatons are not totally independent.

Exercise 3.4 Continue Example 3.13 on the distribution of male children in families in Saxony by fitting a binomial distribution, Bin(n = 12; p = 1/2), specifying equal probability for boys and girls. [Hint: you need to specify both size and prob values for goodfit().]

(a) Carry out the GOF test for this fixed binomial distribution. What is the ratio of xˆ2/df? What do you conclude?

My Answer: The ratio of xˆ2 to degrees of freedom (df) is calculated as part of the goodness-of-fit test for a binomial distribution using the Saxony data. In this specific analysis, the Pearson chi-squared statistic is 249.1954 with 12 degrees of freedom, resulting in a xˆ2/df ratio of approximately 20.77. Similarly, the likelihood ratio chi-squared statistic is 205.4060 with 12 degrees of freedom, yielding a xˆ2/df ratio of around 17.12. The extremely low p-values associated with both statistics (2.013281e-46 for Pearson and 2.493625e-37 for Likelihood Ratio) suggest that the observed data significantly departs from the expected binomial

distribution. The high x^2/df ratios further indicate a substantial discrepancy between the observed and expected frequencies, leading to the rejection of the null hypothesis and implying that the binomial model fits poorly to the Saxony data.

```
# Conduct goodness-of-fit test for Saxony data using a binomial distribution
Saxony_goodness_of_fit <- goodfit(Saxony, type = "binomial", par = list(size = 12, prob = 0.5))

# Obtain the summary of the goodness-of-fit test
summary_stats <- summary(Saxony_goodness_of_fit)
```

```
## Warning in summary.goodfit(Saxony_goodness_of_fit): 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
```

```
# Calculate the chi-squared per degree of freedom
chi_squared_per_df <- summary_stats[,"X^2"] / summary_stats[,"df"]

# Display the result
chi_squared_per_df
```

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

(b) Test the additional lack of fit for the model Bin(n = 12; p = 1/2) compared to the model Bin(n = 12; p = ^p) where ^p is estimated from the data.

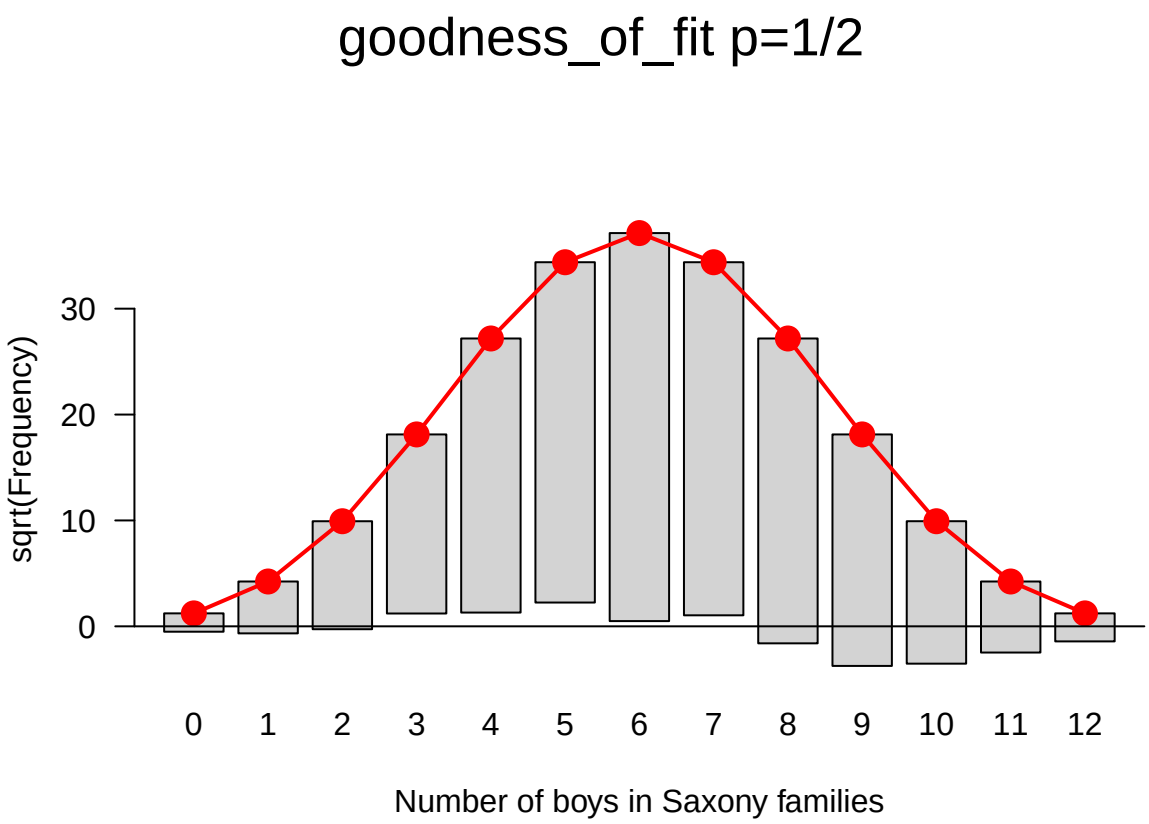
My Answer: The outcome reveals a substantial deviation from the expected distribution, albeit an improvement over the previous model with a fixed probability of 0.5. Rejecting the null hypothesis suggests that the binomial model incorporating the estimated probability (^p) fails to sufficiently capture the observed data. The low p-value shows the statistical evidence against this model, emphasizing the presence of unaccounted factors or patterns not accommodated by the binomial distribution. In conclusion it this model is better fit than previous one, but it is not still a good fit.

```
Saxony_goodness_of_fit2 <- goodfit(Saxony, type = "binomial", par = list(size = 12))
summary(Saxony_goodness_of_fit2)
```

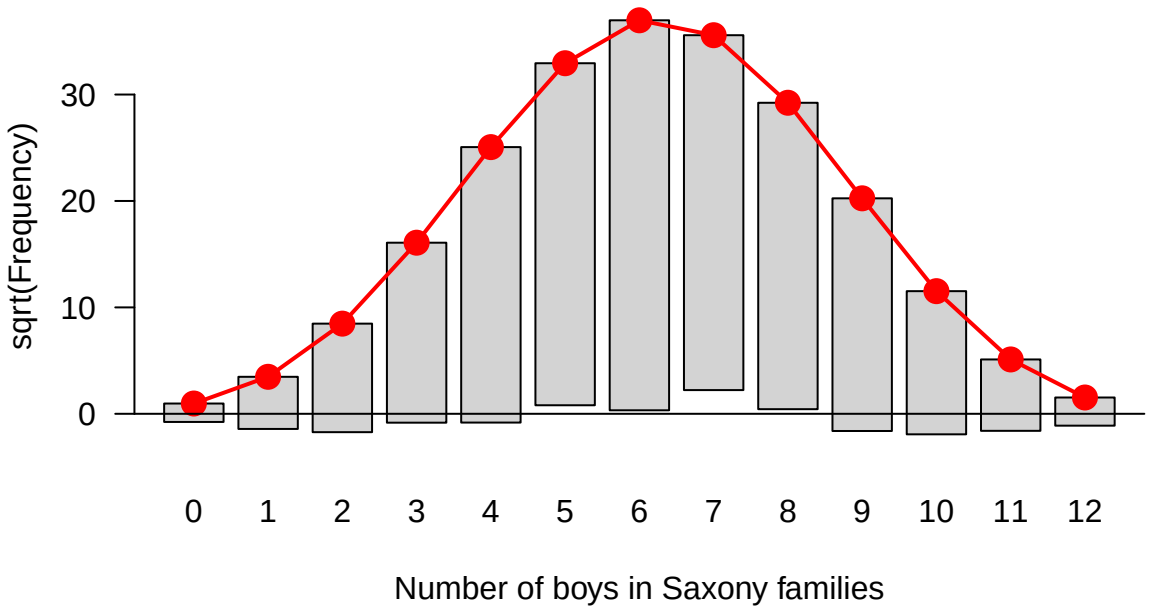
```
##
##   Goodness-of-fit test for binomial distribution
##
##                     X^2 df     P(> X^2)
## Likelihood Ratio 97.0065 11 6.978187e-16
```

(c) Use the plot.gootfit() method to visualize these two models.

```r
plot(Saxony_goodness_of_fit, main = "goodness_of_fit p=1/2", xlab = "Number of boys in Saxony families")
```




```r
plot(Saxony_goodness_of_fit2, main = "goodness_of_fit p=^p", xlab = "Number of boys in Saxony families")
```



Exercise 3.6 Mosteller and Wallace (1963, Table 2.4) give the frequencies, nk, of counts k = 0,1,... of other selected marker words in 247 blocks of text known to have been written by Alexander Hamilton. The data below show the occurrences of the word upon, that Hamilton used much more than did James Madison.

```
count <- 0 : 5
Freq <- c(129, 83, 20, 9, 5, 1)
```

(a) Read these data into R and construct a one-way table of frequencies of counts or a matrix or data frame with frequencies in the first column and the corresponding counts in the second column, suitable for use with goodfit().

```
# Create a data frame with frequencies and corresponding counts
my_df <- data.frame(Freq = Freq, count = count)

# Create a one-way table using xtabs
my_df_oneWay <- xtabs(Freq ~ count, data = my_df)
my_df_oneWay
```

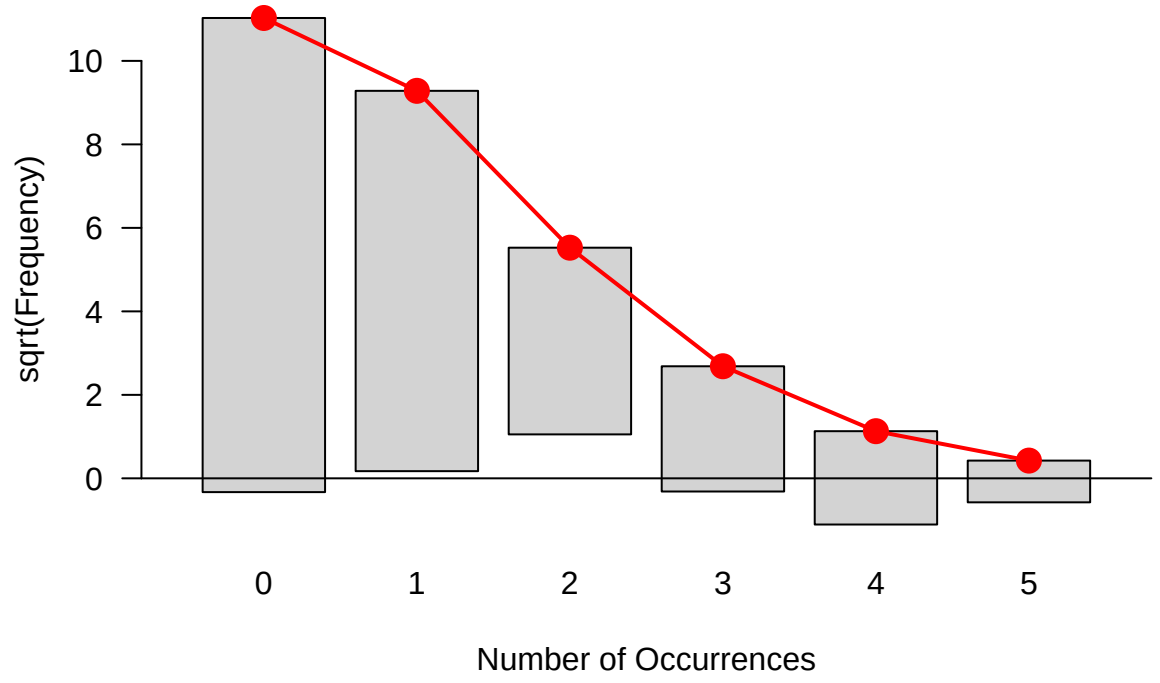
```
## count
##   0   1   2   3   4   5
## 129  83  20   9   5   1
```

(b) Fit and plot the Poisson model for these frequencies.

```
my_df_poisson <- goodfit(my_df, type="poisson")
summary(my_df_poisson)
```

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

```
plot(my_df_poisson)
```

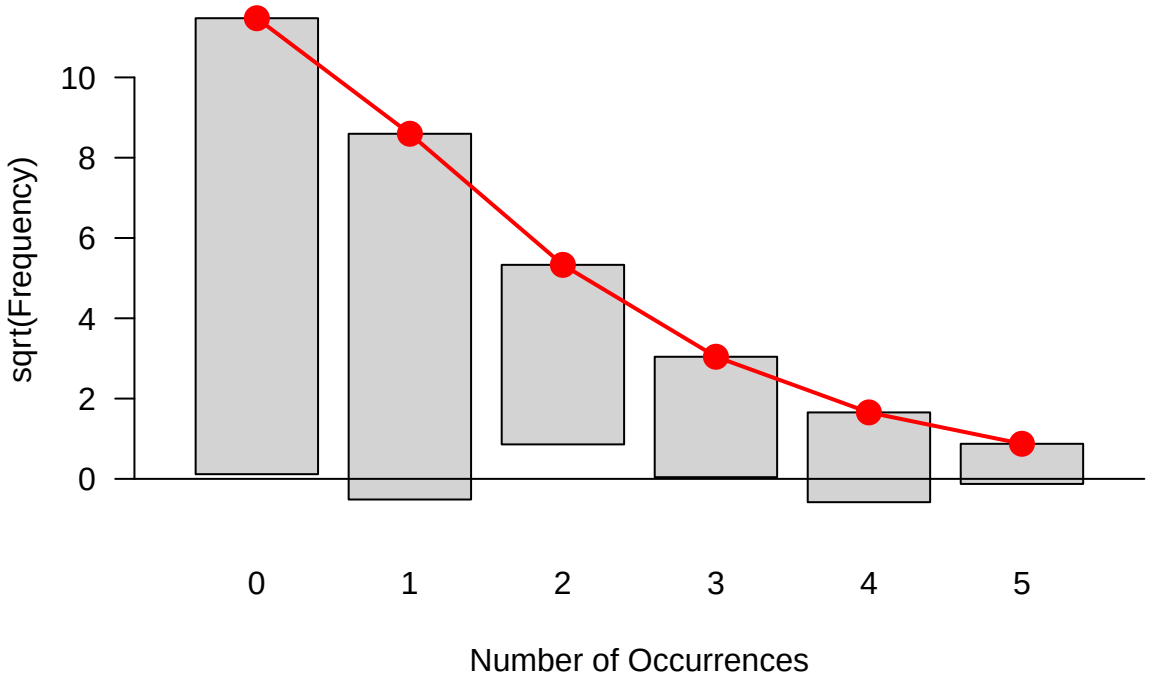


(c) Fit and plot the negative binomial model for these frequencies.

```
my_df_nbinomial <- goodfit(my_df, type="nbinomial")
summary(my_df_nbinomial)
```

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

```
plot(my_df_nbinomial)
```



(d) What do you conclude?

My Answer:

- Poisson Distribution:
- Likelihood Ratio chi-squared statistic: 13.13892
- Degrees of freedom: 4
- P-value: 0.0106
- Negative Binomial Distribution:
- Likelihood Ratio chi-squared statistic: 6.030625
- Degrees of freedom: 3
- P-value: 0.1101

-Result: A smaller p-value (closer to 0) suggests stronger evidence against the null hypothesis (that the data fits the distribution). In this case, the p-value for the Poisson distribution is 0.0106, indicating a statistically significant departure from the expected distribution. On the other hand, the p-value for the negative binomial distribution is 0.1101, which is larger, suggesting less evidence against the null hypothesis.

So, the comparison of p-values suggests that the negative binomial model provides a better fit to the data than the Poisson model.

Exercise 3.7 The data frame Geissler in the vcdExtra package contains the complete data from Geissler's (1889) tabulation of family sex composition in Saxony. The table below gives the number of boys in families of size 11. boys 0 1 2 3 4 5 6 7 8 9 10 11 Freq 8 72 275 837 1,540 2,161 2,310 1,801 1,077 492 93 24

(a) Read these data into R.

```
data("Geissler", package = "vcdExtra")
boys11 <- subset(Geissler, size == 11)
xtabs(Freq ~ boys, data = boys11)
```

```
## 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) Following Example 3.13, use goodfit() to fit the binomial model and plot the results. Is there an indication that the binomial does not fit these data?

My Answer: Negative and large positive residuals on the plot reveal disparities between the observed and expected counts, indicating potential shortcomings in the fit of the binomial model with parameters estimated by maximum likelihood to the underlying distribution of the data. Examining the graphical representation, especially at counts 5, 6, and 7, where observed values markedly differ from expected values, suggests potential issues in the model's ability to accurately capture these specific data points. The negative residuals in this region hint at potential overfitting. Additionally, in other count categories such as 2, 3, 9, and 10, the presence of large positive residuals suggests potential underfitting, indicating that the model may not adequately account for these observations.

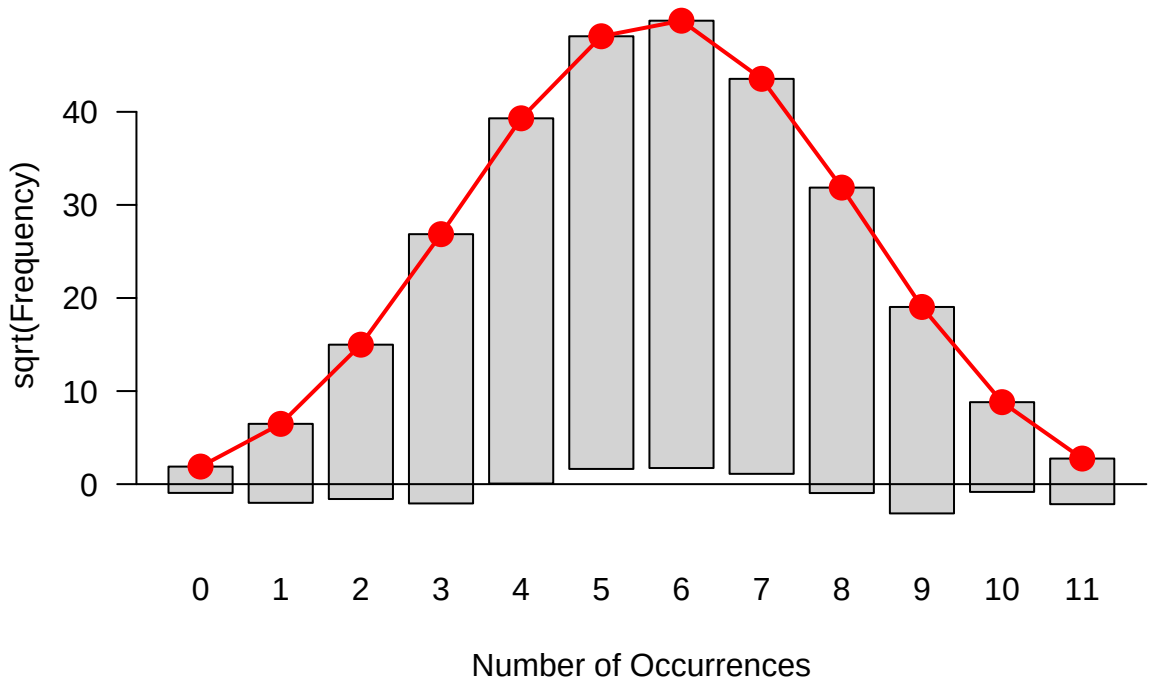
```
boys11.tab <- xtabs(Freq ~ boys, data=boys11)
binomial_boys11 <- goodfit(boys11.tab, type="binomial", par=list(size=11))
binomial_boys11
```

```
##
## Observed and fitted values for binomial distribution
## with parameters estimated by 'ML'
##
##  count observed      fitted pearson residual
##      0        8    3.561571        2.3518439
##      1       72   41.947913        4.6400157
##      2      275  224.572436        3.3650364
##      3      837  721.362873        4.3054683
##      4     1540 1544.755904       -0.1210049
##      5     2161 2315.602347       -3.2128030
##      6     2310 2479.362698       -3.4013219
##      7     1801 1896.217320       -2.1866129
##      8     1077 1015.159294        1.9409187
##      9      492  362.317259        6.8129887
##     10       93   77.588097        1.7496804
##     11       24    7.552287        5.9850291
```

```r
summary(binomial_boys11)
```

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

```r
plot(binomial_boys11)
```

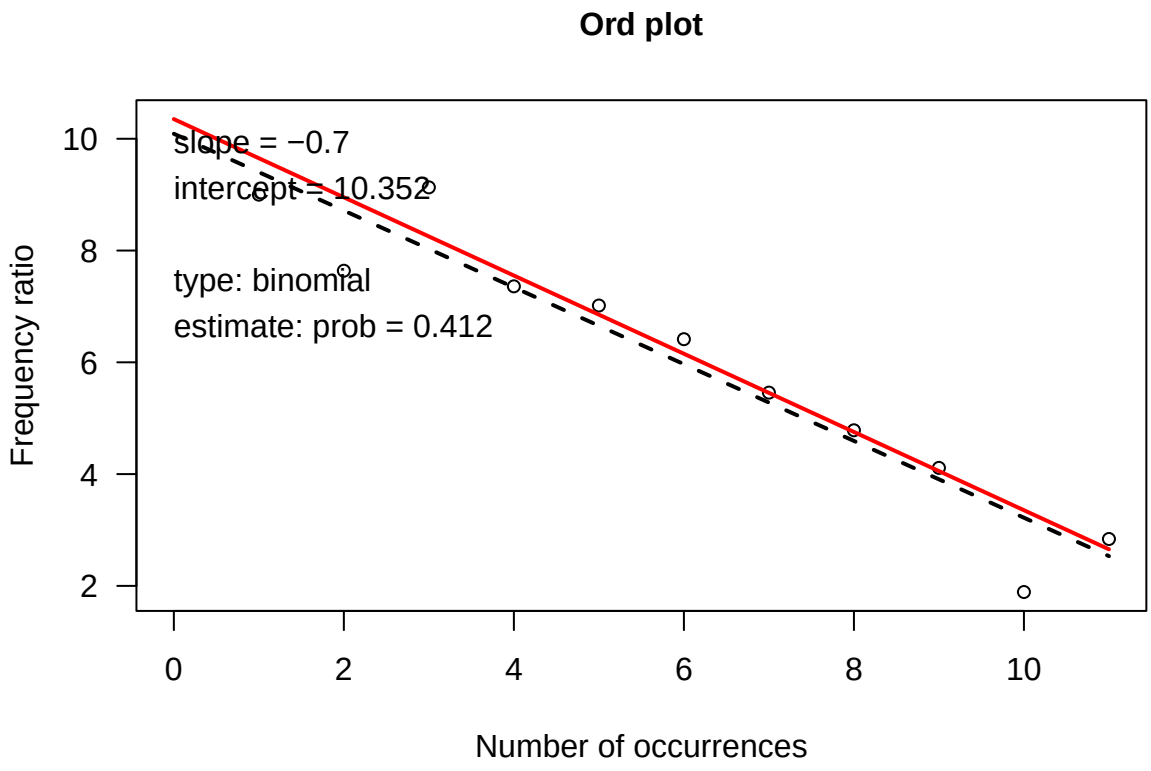


(c) Diagnose the form of the distribution using the methods described in Section 3.4.

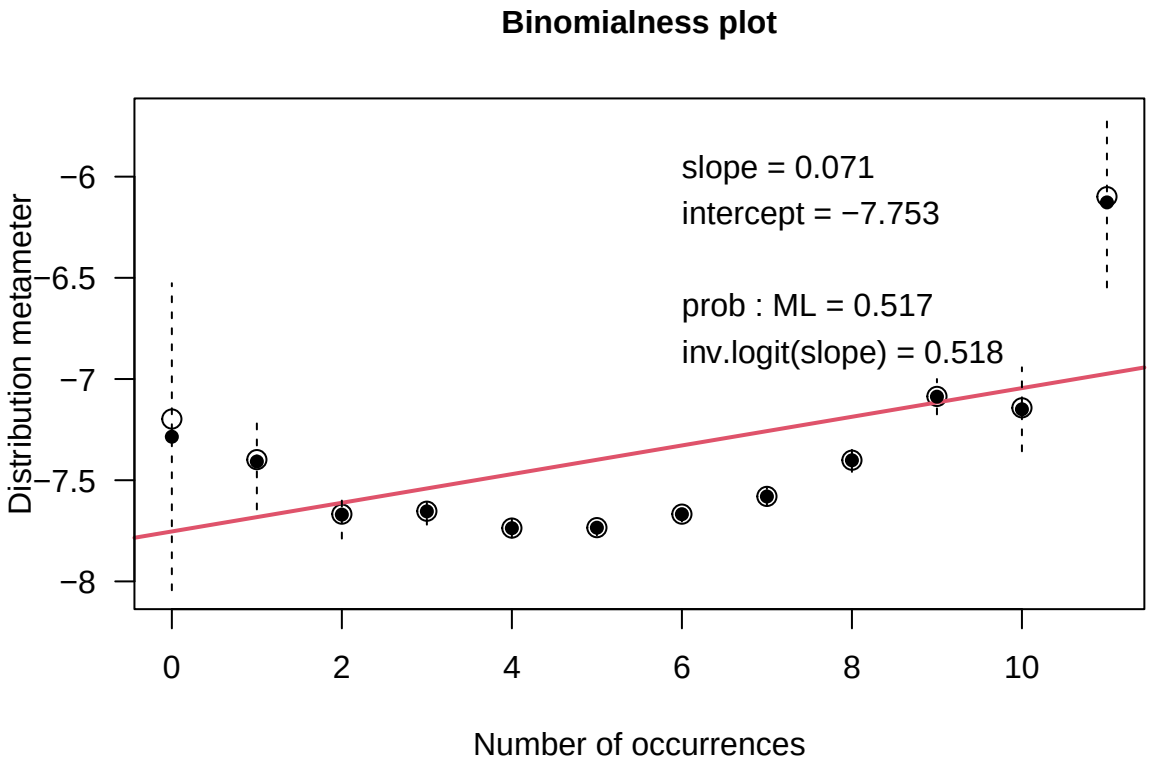
Summary of Section 3.4. The Ord plot is a diagnostic tool used to assess the form of a discrete distribution when substantive knowledge about a plausible mechanism for generating the data is lacking. It relies on a linear relationship that holds for Poisson, binomial, negative binomial, and logarithmic series distributions. The slope and intercept of the plot distinguish these distributions. A positive slope indicates a negative binomial or logarithmic series, while a zero or negative slope suggests a Poisson distribution. A crucial point is the application of weighted least squares fit, considering smaller counts with less weight, for a more accurate diagnosis.

My Answer: The Ord plot suggests that, the distribution closest in form is the binomial. However, the binomialness distribution plot contradicts this by revealing that the binomial model is not a suitable representation for the data.

```
Ord_plot(boys11.tab)
```



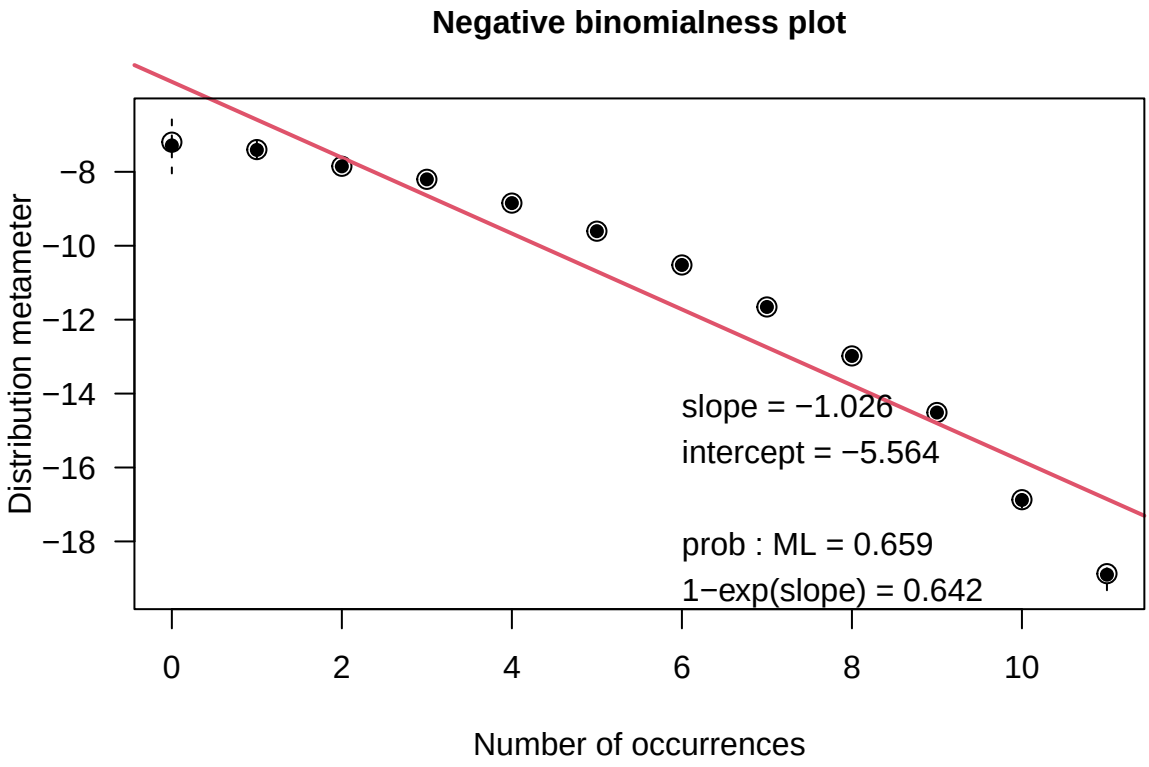

```
distplot(boys11.tab, type="binomial", size=11)
```



(d) Try fitting the negative binomial distribution, and use distplot() to diagnose whether the negative binomial is a reasonable fit.

My Answer: The negative binomial is also not a suitable option too, because this is evident from the size of the residuals.

```
distplot(boys11.tab, type = "nbinomial", size = 11)
```




```
goodfit(boys11.tab, type = "nbinomial", par = list(size = 11))
```

```
##
## Observed and fitted values for nbinomial distribution
## with parameters estimated by 'ML with size fixed'
##
##  count observed    fitted pearson residual
##      0        8  109.1179        -9.680106
##      1       72  409.1061       -16.666663
##      2      275  836.6319       -19.417109
##      3      837 1235.6735       -11.341369
##      4     1540 1474.0729         1.717134
##      5     2161 1507.2574        16.838875
##      6     2310 1369.9457        25.398102
##      7     1801 1133.9697        19.808193
##      8     1077  869.6233         7.032252
##      9      492  625.7336        -5.346206
##     10       93  426.5468       -16.150032
##     11       24  277.5495       -25.399866
```

Exercise 3.8 The data frame Bundesliga gives a similar data set to that for UK soccer scores (UKSoccer) examined in Example 3.9, but over a wide range of years. The following lines calculate a two-way table, BL1995, of home-team and away-team goals for the 306 games in the year 1995.

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

```
##           AwayGoals
## HomeGoals  0  1  2  3  4  5  6
##         0 26 16 13  5  0  1  0
##         1 19 58 20  5  4  0  1
##         2 27 23 20  5  1  1  1
##         3 14 11 10  4  2  0  0
##         4  3  5  3  0  0  0  0
##         5  4  1  0  1  0  0  0
##         6  1  0  0  1  0  0  0
```

(a) As in Example 3.9, find the one-way distributions of HomeGoals, AwayGoals, and TotalGoals = HomeGoals + AwayGoals.

```
# Convert BL1995 to a data frame
df_BL1995 <- as.data.frame(BL1995, stringsAsFactors = FALSE)

# Ensure numeric type for HomeGoals, AwayGoals, and calculate TotalGoals
df_BL1995$home_goals <- as.numeric(df_BL1995$HomeGoals)
df_BL1995$away_goals <- as.numeric(df_BL1995$AwayGoals)
df_BL1995$total_goals <- df_BL1995$home_goals + df_BL1995$away_goals

# Create frequency tables for Home, Away, and Total goals
home <- xtabs(Freq ~ home_goals, data = df_BL1995)
away <- xtabs(Freq ~ away_goals, data = df_BL1995)
total <- xtabs(Freq ~ total_goals, data = df_BL1995)

home
```

```
## home_goals
##   0   1   2   3   4   5   6
##  61 107  78  41  11   6   2
```

```
away
```

```
## away_goals
##   0   1   2   3   4   5   6
##  94 114  66  21   7   2   2
```

```
total
```

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

(b) Use goodfit() to fit and plot the Poisson distribution to each of these. Does the Poisson seem to provide a reasonable fit?

```
summary(goodfit(home))
```

```
##
##   Goodness-of-fit test for poisson distribution
##
##                       X^2 df  P(> X^2)
## Likelihood Ratio 3.480285  5 0.6263728
```

```
summary(goodfit(away))
```

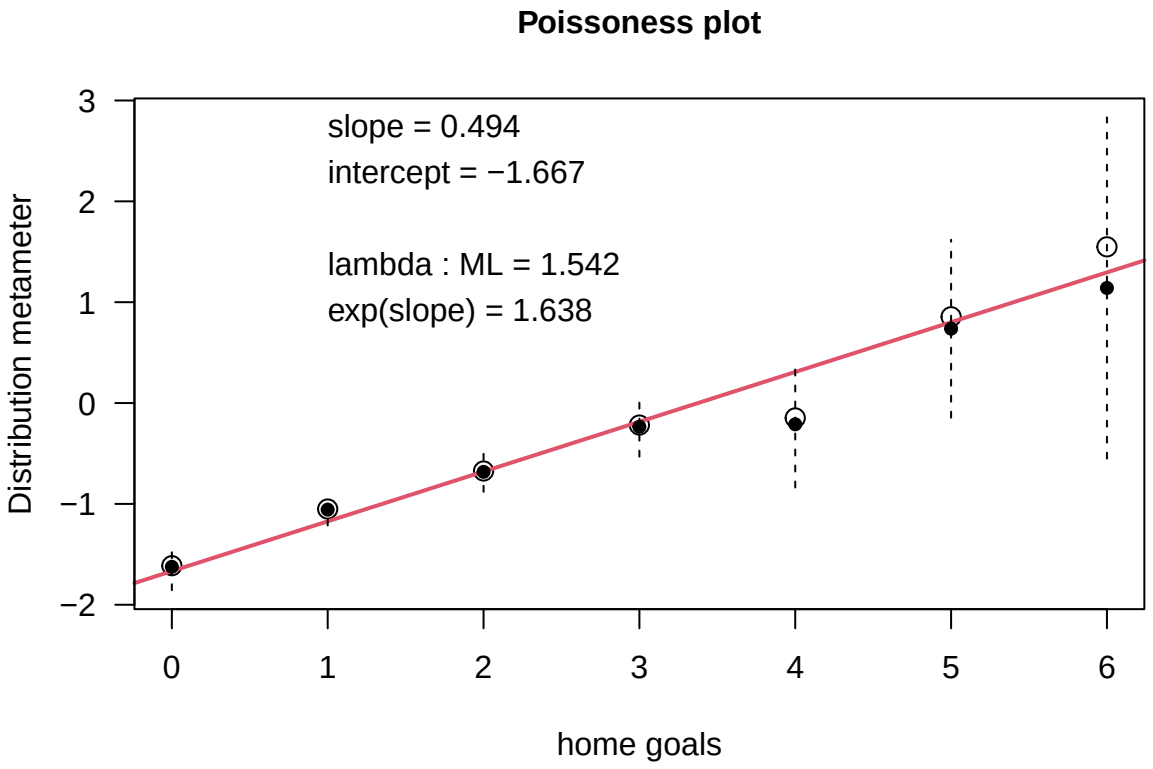
```
##
##   Goodness-of-fit test for poisson distribution
##
##                      X^2 df  P(> X^2)
## Likelihood Ratio 4.86855  5 0.4321324
```

```
summary(goodfit(total))
```

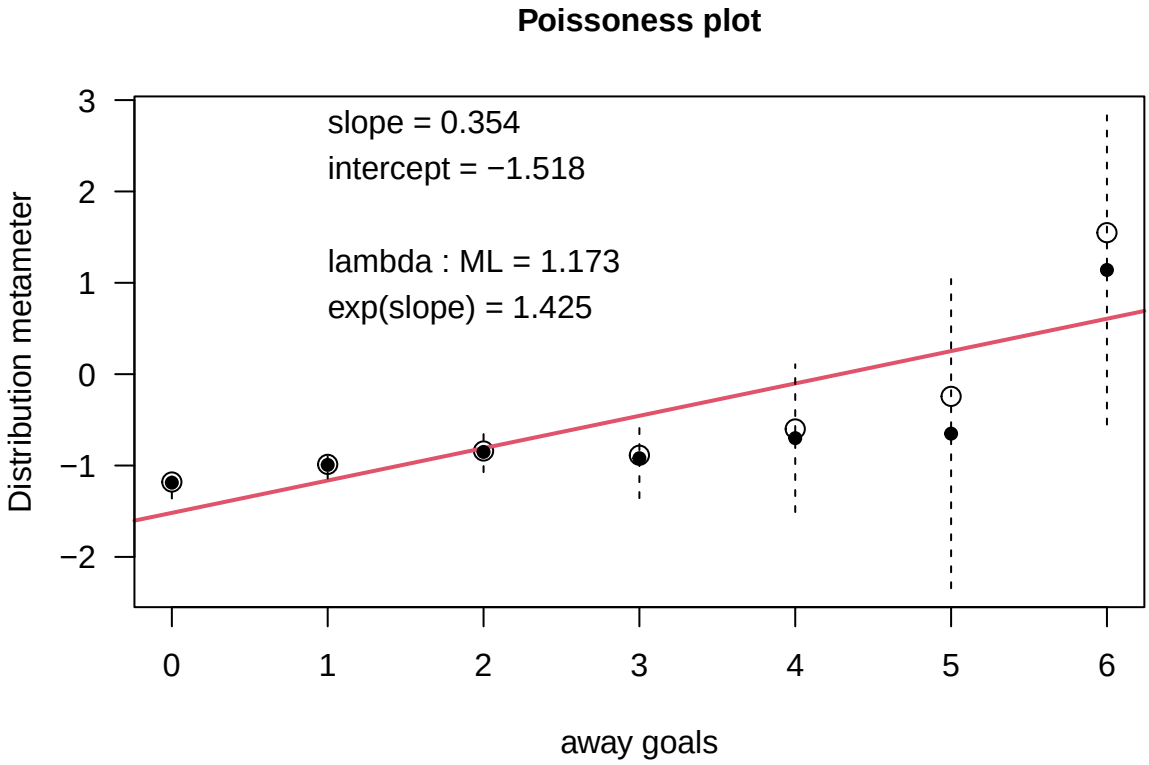
```
##
##   Goodness-of-fit test for poisson distribution
##
##                       X^2 df   P(> X^2)
## Likelihood Ratio 19.84905  8 0.01092249
```

(c) Use distplot() to assess fit of the Poisson distribution.

```
distplot(home, xlab="home goals")
```




```
distplot(away, xlab="away goals")
```



(d) What circumstances of scoring goals in soccer might cause these distributions to deviate from Poisson distributions? **My Answer:** In soccer, deviations from Poisson distributions in goal-scoring patterns can arise due to factors that challenge the assumptions of the Poisson model. The Poisson distribution assumes (a) independent events and (b) constant probabilities. However, in soccer, the probability of scoring a goal is likely not constant for all pairs of teams, as various factors such as team strength, strategies, player skills, and match circumstances can significantly influence goal-scoring dynamics, leading to deviations from the Poisson distribution.