MSMS - 106

Ananda Biswas

Practical 04

✎ Fit a binomial distribution to the given dataset.

x	0	1	2	3	4	5	6	7	8
f	5	9	22	29	36	25	10	3	1

Also perform a χ^2 goodness of fit test.

➜ *Fitting a Binomial Distribution*

Here $n = 8$. $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}$; $\hat{p} = \dfrac{\bar{x}}{n}$.

```
x <- 0:8; n <- length(x)-1
freq <- c(5, 9, 22, 29, 36, 25, 10, 3, 1)
```

```
weighted_mean <- function(x, weight){
  xw <- 0
  w <- 0
  for (i in 1:length(x)) {
    xw <- xw + x[i] * weight[i]
    w <- w + weight[i]
  }
  return(xw / w)
}
```

```
x_bar <- weighted_mean(x, freq)
x_bar

## [1] 3.557143
```

$\bar{x} = 3.5571429$. So $\hat{p} = 0.4446429$. Now we fit $Bin(8, 0.4446429)$ distribution to the given data.

Now $P(X = 0) = (1 - \hat{p})^8 = 0.0090485$ and

$$P(X = i + 1) = \frac{n - i}{i + 1} \cdot \frac{p}{1 - p} \cdot P(X = i) \;\; \forall i = 0(1)n - 1.$$

Also, expected frequency of $i = k \cdot P(X = i)\ \forall i = 0(1)n$, where $k = \sum_{i=0}^{n} f_i$ is the total frequency.

```
p <- x_bar / n
```

```
probabilities <- c((1 - p)^n)

i <- 1
while (i <= 8) {
  probabilities[i+1] <- ((n-(i-1)) / (i)) * (p / (1-p)) * probabilities[i]

  i <- i + 1
}
```

```
total_frequency <- 0

for (i in 1:length(freq)) {
  total_frequency <- total_frequency + freq[i]
}
```

```
expected_frequencies <- c()

for (i in 1:9) {
  expected_frequencies[i] <- probabilities[i] * total_frequency
}
```

Here is our fit.

```
df <- data.frame(x = x,
                 observed = freq,
                 expected = expected_frequencies)
df

##   x observed   expected
## 1 0        5  1.2667967
## 2 1        9  8.1140163
## 3 2       22 22.7375088
## 4 3       29 36.4092584
## 5 4       36 36.4385263
## 6 5       25 23.3394033
## 7 6       10  9.3432660
## 8 7        3  2.1373204
## 9 8        1  0.2139038
```

```
sum(df$observed); sum(df$expected)

## [1] 140
## [1] 140
```

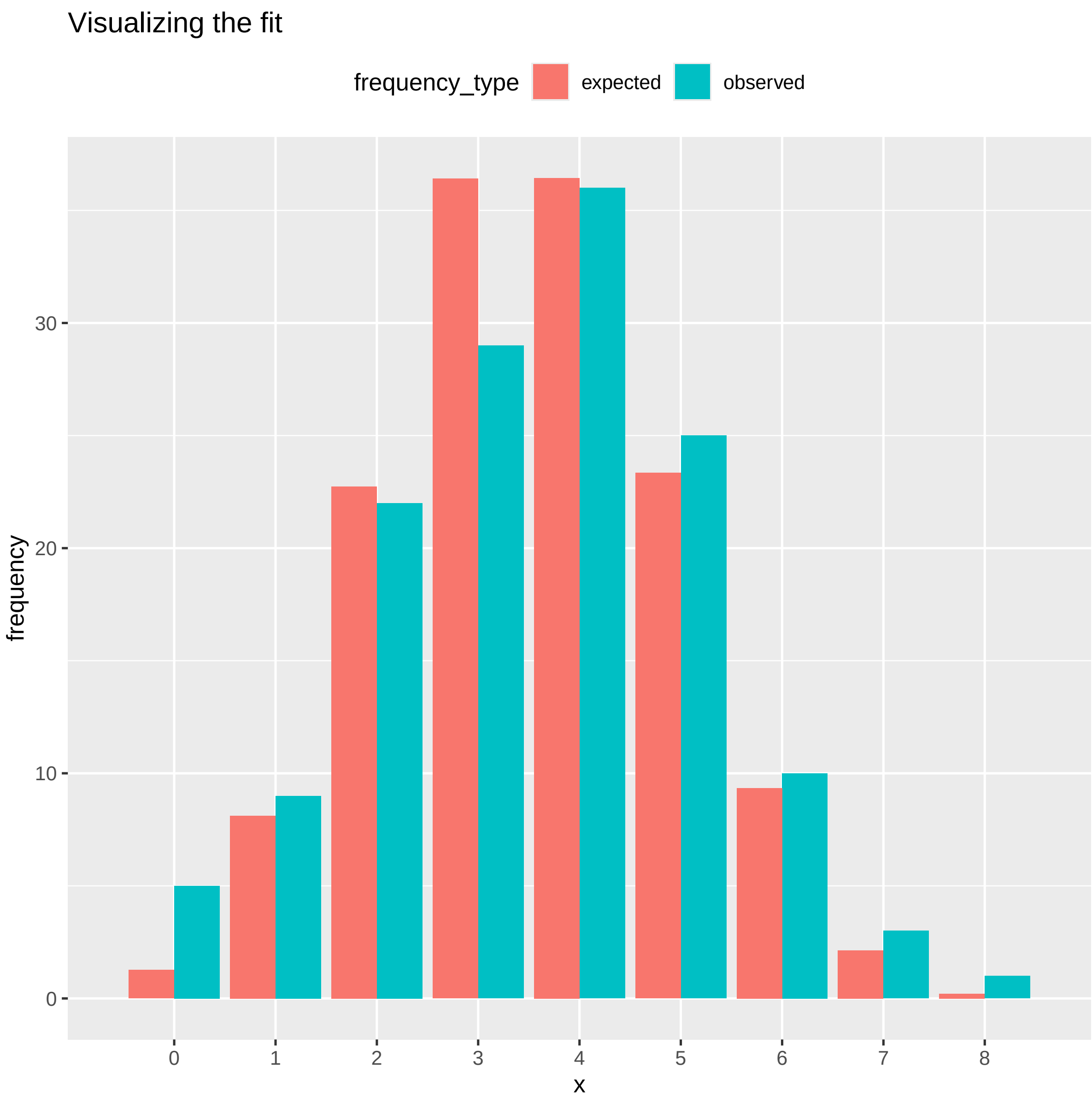
Total expected frequency and total observed frequency are also equal.

A visualization of the fit will be great.

```
library(tidyverse)
```

```
df_melted <- df %>%
                pivot_longer(cols = c("observed", "expected"),
                             names_to = "frequency_type",
                             values_to = "frequency")
```

```
df_melted %>%
  ggplot(aes(x = x, y = frequency, fill = frequency_type)) +
  geom_col(position = "dodge") +
  scale_x_discrete(limits = x) +
  labs(title = "Visualizing the fit") +
  theme(legend.position = "top")
```



⊕ χ^2 ***Goodness of fit test***

$\chi^2 = \sum_{i=1}^{m} \frac{(f_i - kp_i)^2}{kp_i}$ where m is the number of classes, f_i is the observed frequency of i-th class,

p_i is the theoretical probability of belonging to i-th class, k is total frequency.

In large sample, $\chi^2 \sim \chi^2_{m-1}$.

We also must have expected frequency greater than or equal to 5 for each class.

Here, in order to achieve so, we shall combine similar categories $x = 0$ & 1; $x = 6$ & 7 & 8.

```
new_df <- data.frame(x = c("0, 1", "2", "3", "4", "5", "6, 7, 8"),
                     observed = c(df[1, 2] + df[2, 2],
                                  df[3:6, 2],
                                  df[7, 2] + df[8, 2] + df[9, 2]),
                     expected = c(df[1, 3] + df[2, 3],
                                  df[3:6, 3],
                                  df[7, 3] + df[8, 3] + df[9, 3]))
```

Now we have

```
new_df

##         x observed  expected
## 1    0, 1       14  9.380813
## 2       2       22 22.737509
## 3       3       29 36.409258
## 4       4       36 36.438526
## 5       5       25 23.339403
## 6 6, 7, 8       14 11.694490
```

See that each of the expected frequencies is greater than or equal to 5. Number of classes m is 6. Now we perform χ^2 goodness of fit test.

```
observed_chi_sq <- 0

for (i in 1:dim(new_df)[1]) {
  d <- new_df$observed[i] - new_df$expected[i]
  e <- new_df$expected[i]
  observed_chi_sq <- observed_chi_sq + (d^2) / e
}
```

```
observed_chi_sq; qchisq(0.05, 5, lower.tail = FALSE)

## [1] 4.384173
## [1] 11.0705
```

Observed $\chi^2 = 4.3841734 < \chi^2_{0.05,5} = 11.0704977$. So we fail to reject the null hypothesis of goodness of fit test and conclude that there is not enough evidence to claim that the given data is not from a Binomial population.