MSMS - 106

Ananda Biswas

Practical 05

✎ Fit a poisson distribution to the given data-set.

x	0	1	2	3	4	5	6	7	8	9 or more
f	162	193	115	83	44	24	19	8	2	0

Also perform a χ^2 goodness of fit test.

➜ *Fitting a Poisson Distribution*

The P.M.F. of a Poisson distribution is $P(X=x) = e^{-\lambda} \cdot \dfrac{\lambda^x}{x!}\ ; x = 0, 1, 2, 3, \dots\ , \lambda > 0.$

We estimate parameter λ as $\hat{\lambda} = \bar{x} = \dfrac{\sum\limits_{i=0}^{\infty} x_i f_i}{\sum\limits_{i=0}^{\infty} f_i}$.

```r
x <- 0:9
freq <- c(162, 193, 115, 83, 44, 24, 19, 8, 2, 0)
```

```r
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)
}
```

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

## [1] 1.775385
```

$\bar{x} = 1.7753846$. So $\hat{\lambda} = 1.7753846$. Now we fit $Poisson(1.7753846)$ distribution to the given data.

Now $P(X=0) = e^{-1.7753846} = 0.1694183$ and

$$P(X=i+1) = \frac{\lambda}{i+1} \cdot P(X=i);\ i = 0, 1, 2, \ldots.$$

Also, expected frequency of $i = k \cdot P(X=i);\ i = 0, 1, 2, \ldots$, where $k = \sum_{i=0}^{\infty} f_i$ is the total frequency.

```
lambda <- x_bar
```

```
probabilities <- c(exp(-lambda))

i <- 1
while (i <= 8) {
  probabilities[i+1] <- (lambda / x[i+1]) * 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
}
```

And $P(X \geq 9) = 1 - P(X \leq 8) = 9.9084863 \times 10^{-5}$.

Here is our fit.

```
df <- data.frame(x = c(0:8, "9 or more"),
                 observed = freq,
                 expected = c(expected_frequencies,
                              total_frequency * (1 - sum(probabilities))))
df

##            x observed     expected
## 1          0      162 110.12187946
## 2          1      193 195.50869061
## 3          2      115 173.55156074
## 4          3       83 102.70692364
## 5          4       44  45.58607303
## 6          5       24  16.18656255
## 7          6       19   4.78956235
## 8          7        8   1.21475933
## 9          8        2   0.26958313
## 10 9 or more        0   0.06440516
```

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

## [1] 650
## [1] 650
```

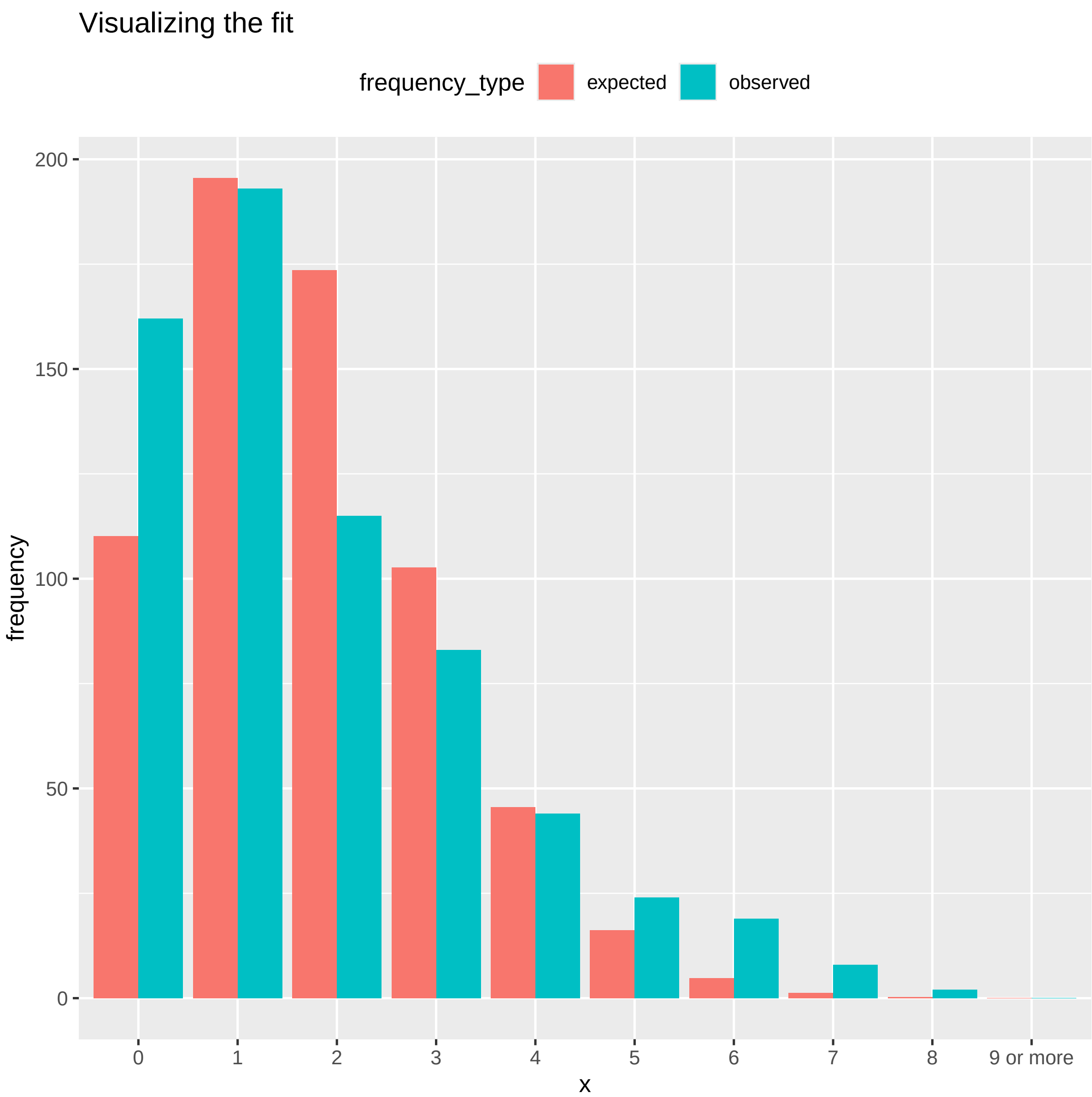
Total expected frequency and total observed frequency are also equal.

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") +
  labs(x = "x", 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, $\chi^2 \sim \chi^2_{m-1-u}$, where u is the number of parameters estimated from the data.

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

Here, in order to achieve so, we shall combine last 4 classes.

```
new_df <- data.frame(x = c(0:5, "6 or more"),
                     observed = c(df[1:6, 2], sum(df[7:10, 2])),
                     expected = c(df[1:6, 3], sum(df[7:10, 3])))
```

Now we have

```
new_df

##           x observed  expected
## 1         0      162 110.12188
## 2         1      193 195.50869
## 3         2      115 173.55156
## 4         3       83 102.70692
## 5         4       44  45.58607
## 6         5       24  16.18656
## 7 6 or more       29   6.33831
```

See that each of the expected frequencies is greater than or equal to 5. Number of classes m is 7. 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] 132.8571
## [1] 11.0705
```

Observed $\chi^2 = 132.857145 > \chi^2_{0.05,5} = 11.0704977$. So we reject the null hypothesis of goodness of fit test and conclude that the given data is not from a Poisson population.