Practice 1: Number of Breaks

Ignacio Almodóvar Cárdenas & Luis Ángel Rodríguez García

22/02/2022

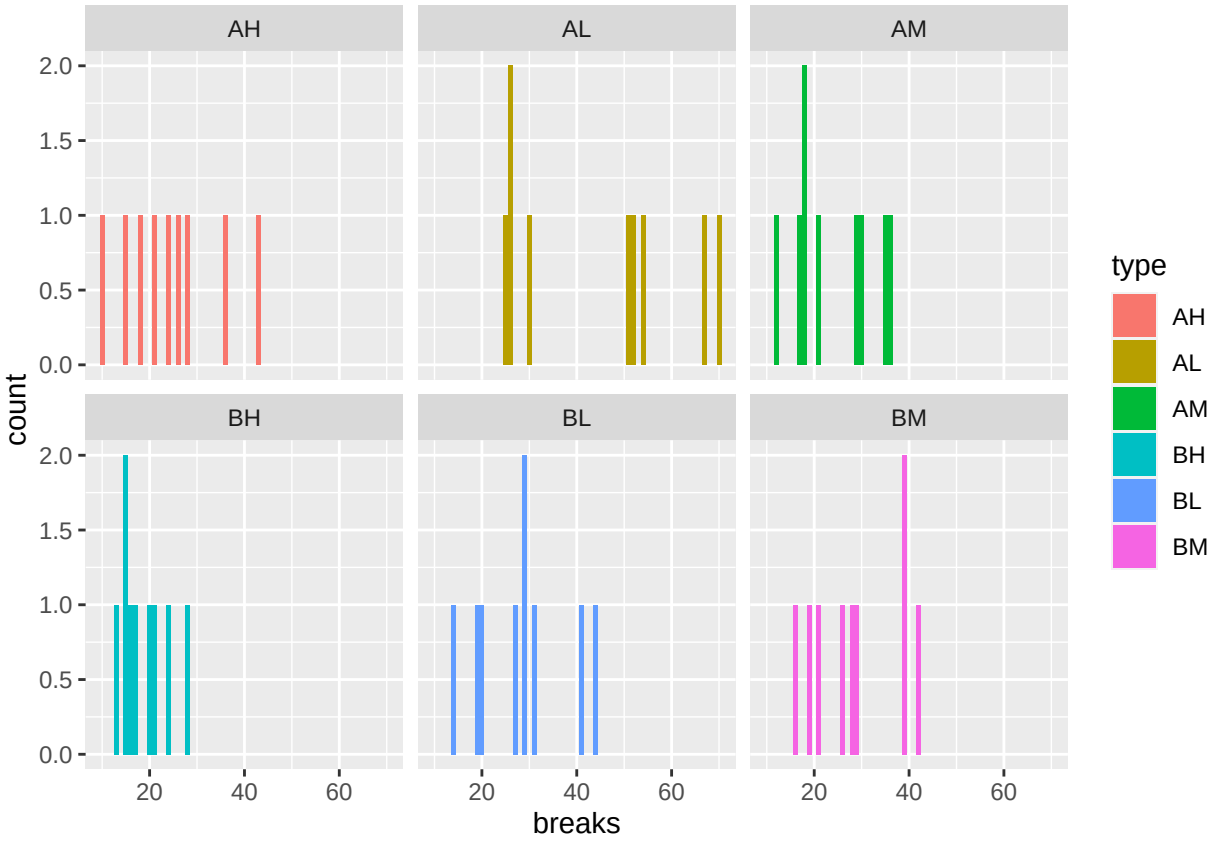
Introduction

We want to provide a conditional estimation on the number of breaks for different types of threads. Therefore, as the data provided contains observations from different groups of threads, we are going to split them into six different groups. This way we can apply Bayesian Inference to obtain insights for each type of group.

The dataset contains 3 variables:

- Number of breaks (continuous variable)
- Wool (categorical variable that defined two types of the wool: A and B)
- Tension (categorical variable that indicates the tension of the thread: L -low-, M -medium- and H -high-)

As we just mentioned, we are going to split the data into 6 different groups. Therefore we will have one group for wool A with low tension, another for wool A with medium tension, wool A with high tension and so on.



In the plots above, we can see the probability mass function of each wool type and tension. in the case of the

wool A and tension *high*, there is not a unique mode. In the reaming types associated with their tensions there are just a few modes (or just one).

Bayesian inference

Once we have our groups well defined, we can start to apply Bayesian Inference. As we want to have more knowledge about the mean number of breaks during weaving for different configurations, we can represent our problem with a ***Poisson* model**, being the count the number of breaks. Therefore,

$$Y_{1_{AL}}, \ldots, Y_{n_{AL}} | \theta_{AL} \sim i.i.d. Poisson(\theta_{AL})$$

$$Y_{1_{AM}}, \ldots, Y_{n_{AM}} | \theta_{AM} \sim i.i.d. Poisson(\theta_{AM})$$

$$\vdots$$

$$Y_{1_{BH}}, \ldots, Y_{n_{BH}} | \theta_{BH} \sim i.i.d. Poisson(\theta_{BH})$$

We also know that for a poisson distribution, both mean and variances are:

$$\mathbb{E}[Y|\theta] = \theta \quad \mathbb{V}ar[Y|\theta] = \theta$$

Prior distribution

However, we do not have any information about θ, it is some unknown number between 0 and ∞. Then, we have to defined this parameter using other methods but not insights that come from the sample. The only knowledge we have about it is that from the prior we can assume that:

$$\theta \sim gamma(a, b)$$

In order to implement it, we have to assign values for both parameters a and b. For a Gamma distribution we know that:

$$\mathbb{E}[\theta] = \frac{a}{b} \quad \mathbb{V}ar[\theta] = \frac{a}{b^2}$$

From the prior we have to guess the mean number of breaks. As we do not have explicit information from θ, we have to set it in a rough way. Let's suppose that an expert says that in average, the number of breaks of the wool is around 23 with a standard deviation of 1. Consequently, we are ready to calculate the parameters based on the previous formula of the expected value (we plug in the value 1 to b):

$$\begin{cases} \frac{a}{b} = 23 \\ \frac{a}{b^2} = 1 \end{cases} \tag{1}$$

$$a = 529 \quad b = 23$$

Now, setting b as 23 we get that a is 529, then we have our prior set. This means that in 23 wool there are 529 breaks.

Posterior distribution

Once we have defined our prior, the next step is to obtain the posterior knowing that the posterior distribution of θ os the gamma distribution as it follows:

$$\{\theta|Y_1,\ldots,Y_n\} \sim gamma(a+\sum_{i=1}^{n} Y_i, b+n)$$

In order to calculate the conditional estimation of the mean and the variance we are going to use the following formulas:

$$\mathbb{E}[\theta|Y_1,\ldots,Y_n] = \frac{a+\sum y_i}{b+n} \quad \mathbb{V}ar[\theta|Y_1,\ldots,Y_n] = \frac{a+\sum y_i}{b+n}\frac{b+n+1}{b+n}\frac{1}{b+n+1}$$

Let's compute all the things above using R and summarize the results in Table 1.

Table 1: Posterior results

	n	**sy**	**mu**	**pos.mean**	**pos.ci.lw**	**pos.ci.up**	**pos.variance**
AL	9	401	44.56	29.06	27.22	30.96	0.91
AM	9	216	24.00	23.28	21.64	24.98	0.73
AH	9	221	24.56	23.44	21.79	25.14	0.73
BL	9	254	28.22	24.47	22.78	26.21	0.76
BM	9	259	28.78	24.62	22.94	26.37	0.77
BH	9	169	18.78	21.81	20.22	23.46	0.68

In the table above, we can see the first column related to the sample size of each group, the secon associated to the sum of the breaks from the sample of this group, the third is the mean of the breaks from the sample, the forth (low) and the fifth (up) are the confidence intervals associated and the third is the variance associated to the parameter θ for each group.

We can see that the AL breaks are those ones with the larger mean and the higher variance, in contrast we have the BH breaks that are associated with the smaller mean and the shorter variance. Comparing these results with the insights shown in the bar plots of the introduction section, we conclude that the *a posteriori* information are well informative by the sample data (something that we want as we have to be as impartial as possible).

Apendix

Code

```
knitr::opts_chunk$set(echo = TRUE)

library(ggplot2)
library(dplyr)
library(magrittr)
library(datasets)
library(scales)
library(kableExtra)

data <- warpbreaks

data <- data %>%
  mutate(type = paste(wool, tension, sep="")) %>%
  select(c('breaks', 'type'))

ggplot(data, aes(x=breaks, fill= type)) +
  geom_bar() +
  facet_wrap(~type)
a <- 529
b <- 23

breaks.al <- data %>% dplyr::filter(type=="AL")
breaks.am <- data %>% dplyr::filter(type=="AM")
breaks.ah <- data %>% dplyr::filter(type=="AH")
breaks.bl <- data %>% dplyr::filter(type=="BL")
breaks.bm <- data %>% dplyr::filter(type=="BM")
breaks.bh <- data %>% dplyr::filter(type=="BH")

n.al <- nrow(breaks.al)
sy.al <- sum(breaks.al$breaks)
mu.al <- mean(breaks.al$breaks)

n.am <- nrow(breaks.am)
sy.am <- sum(breaks.am$breaks)
mu.am <- mean(breaks.am$breaks)

n.ah <- nrow(breaks.ah)
sy.ah <- sum(breaks.ah$breaks)
mu.ah <- mean(breaks.ah$breaks)

n.bl <- nrow(breaks.bl)
sy.bl <- sum(breaks.bl$breaks)
mu.bl <- mean(breaks.bl$breaks)

n.bm <- nrow(breaks.bm)
sy.bm <- sum(breaks.bm$breaks)
mu.bm <- mean(breaks.bm$breaks)

n.bh <- nrow(breaks.bh)
sy.bh <- sum(breaks.bh$breaks)
```

```r
mu.bh <- mean(breaks.bh$breaks)

n <- c(n.al, n.am, n.ah,
       n.bl, n.bm, n.bh)
sy <- c(sy.al, sy.am, sy.ah,
        sy.bl, sy.bm, sy.bh)
mu <- c(mu.al, mu.am, mu.ah,
        mu.bl, mu.bm, mu.bh)

posterior.mean <- function(sy, n){
  return( (a+sy) / (b+n) )
}

posterior.variance <- function(sy, n) {
  return( ((a+sy) / (b+n)) * ((b+n+1) / (b+n)) * (1 / (b+n+1)) )
}

posterior.ci <- function(sy, n) {
  qgamma(c(0.025,0.975), a+sy, b+n)
}

# Posterior mean
pos.mean.al <- posterior.mean(sy.al, n.al)
pos.mean.am <- posterior.mean(sy.am, n.am)
pos.mean.ah <- posterior.mean(sy.ah, n.ah)
pos.mean.bl <- posterior.mean(sy.bl, n.bl)
pos.mean.bm <- posterior.mean(sy.bm, n.bm)
pos.mean.bh <- posterior.mean(sy.bh, n.bh)
pos.mean <- c(pos.mean.al, pos.mean.am, pos.mean.ah,
              pos.mean.bl, pos.mean.bm, pos.mean.bh)

# CI 95% for each type
pos.ci.al <- posterior.ci(sy.al, n.al)
pos.ci.am <- posterior.ci(sy.am, n.am)
pos.ci.ah <- posterior.ci(sy.ah, n.ah)
pos.ci.bl <- posterior.ci(sy.bl, n.bl)
pos.ci.bm <- posterior.ci(sy.bm, n.bm)
pos.ci.bh <- posterior.ci(sy.bh, n.bh)
pos.ci.lw <- c(pos.ci.al[1], pos.ci.am[1], pos.ci.ah[1],
           pos.ci.bl[1], pos.ci.bm[1], pos.ci.bh[1])
pos.ci.up <- c(pos.ci.al[2], pos.ci.am[2], pos.ci.ah[2],
           pos.ci.bl[2], pos.ci.bm[2], pos.ci.bh[2])

# Posterior variance
pos.variance.al <- posterior.variance(sy.al, n.al)
pos.variance.am <- posterior.variance(sy.am, n.am)
pos.variance.ah <- posterior.variance(sy.ah, n.ah)
pos.variance.bl <- posterior.variance(sy.bl, n.bl)
pos.variance.bm <- posterior.variance(sy.bm, n.bm)
pos.variance.bh <- posterior.variance(sy.bh, n.bh)
pos.variance <- c(pos.variance.al, pos.variance.am, pos.variance.ah,
                  pos.variance.bl, pos.variance.bm, pos.variance.bh)
```

```r
pos.res.matrix <- cbind(n, sy, mu, pos.mean,
                        pos.ci.lw, pos.ci.up, pos.variance)
rownames(pos.res.matrix) <- c("AL", "AM", "AH", "BL", "BM", "BH")

knitr::kable(pos.res.matrix, caption = "Posterior results", digits = 2) %>%
  column_spec(1, bold=TRUE) %>% row_spec(0, bold=TRUE) %>%
  kable_styling(latex_options = "HOLD_position")
```