Sequencing: Working with count data

Koen Van den Berge

7/5/2021

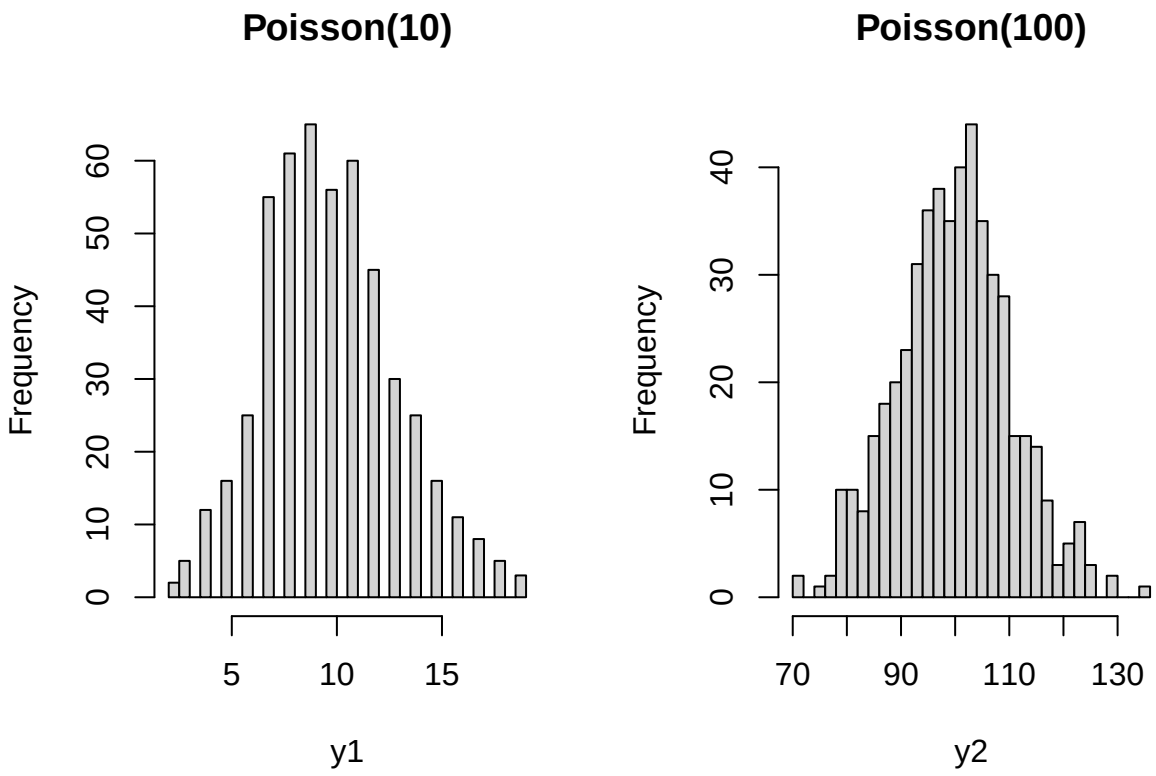
Contents

1 The Poisson distribution

- The Poisson distribution is a typical count distribution that is generally popular and fairly easy to work with. It is defined by a single parameter: its mean μ. For a Poisson distributed random variable Y_i with observations $i \in \{1, \dots, n\}$, its variance is equal to its mean. That is, if $Y_i \sim Poi(\mu)$, then $E(Y_i) = Var(Y_i) = \mu$.
- This immediately shows an important feature of count data: the **mean-variance relationship**. Indeed, in count data, the variance will always be a function of the mean.
- This is quite intuitive. Consider the following example. You have two bird cages, where in one bird cage there are 10 birds, while in the other there are 100 birds. You let a sample of people count the number of birds in either one of the cages. It seems unlikely that a person in front of the 10-bird cage would come up with an estimate of 5, while it seems quite likely that someone in front of the 100-bird cage would come up with an estimate of 95. Even though the difference from the true value is the same, the exact value has an impact on the plausible deviation around it.

```
set.seed(11)
y1 <- rpois(n=500, lambda=10)
y2 <- rpois(n=500, lambda=100)

par(mfrow = c(1,2))
hist(y1, main="Poisson(10)", breaks=40)
hist(y2, main="Poisson(100)", breaks=40)
```



- While being fairly simple, the Poisson distribution is also quite restrictive. Indeed, the restriction that the variance equals the mean rarely holds up in practice.
- As we will see, several extensions of the Poisson distribution have been developed, in order to circumvent its restrictive assumptions.

2 Properties of (RNA-seq) count data

In this lecture, we will introduce working with count data, using a real bulk RNA-seq dataset from Haglund *et al.* (2012). We will be importing this dataset using the data package parathyroidSE from Bioconductor.

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
if(!"SummarizedExperiment" %in% installed.packages()) BiocManager::install("SummarizedExperiment")
# install package if not installed.
if(!"parathyroidSE" %in% installed.packages()) BiocManager::install("parathyroidSE")

library(parathyroidSE)
library(SummarizedExperiment)

# import data
data("parathyroidGenesSE", package="parathyroidSE")
# rename for convenience
se <- parathyroidGenesSE
rm(parathyroidGenesSE)
```

- Count data are inherently

```
y <- assays(se)$counts[1,]
hist(y, breaks = ncol(se),
     xlab = "Gene expression")
```

2.1 Mean-variance relationship

2.2 Relative uncertainty, offsets and count scaling

3 Variance-stabilizing transformations

4 Generalized linear models