Titolo

Filippo Gambarota

University of Padova

2022/2023

Outline

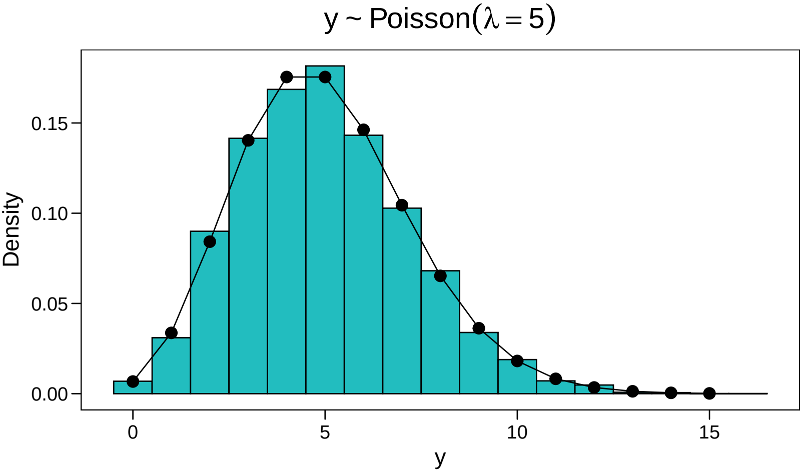
Poisson distribution

The Poisson distribution is defined as:

$$p(y) = \frac{e^{-\mu}\mu^y}{y!}$$

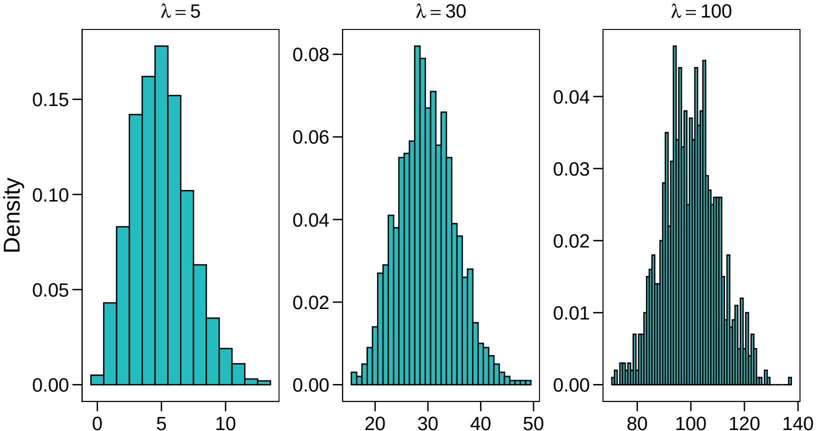
Where the mean is μ and the variance is μ

Poisson distribution

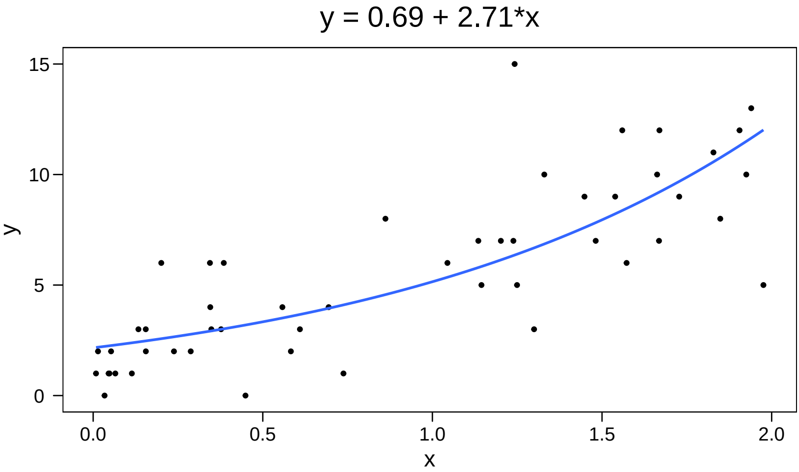


Poisson distribution

As the mean increases also the variance increase and the distributions is approximately normal:



Poisson distribution



Overdispersion

Overdispersion concerns observing a greater variance compared to what would have been expected by the model.

An estimate of the overdispersion can be done calculating the ratio of squared pearson residuals and the degrees of freedom of the model .

$$P = \frac{\sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i}}\right)^2}{df}$$

Without overdispersion the ratio is aproximately 1, with overdispersion the ratio is greater than 1.

Testing overdispersion

There are multiple ways of testing the overdispersion. The first is using the P statistics computed in the slide before and calculate a p value based on the χ^2 distribution with $df = n - p$ degrees of freedom with n is the number of observations and p the number of model coefficients. A p value lower than the α level suggest evidence for overdispersion.

```
fit <- glm(y ~ x, data = dat, family = poisson())
(overdisp <- sum(residuals(fit, type = "pearson")^2)/fit$df.re
```

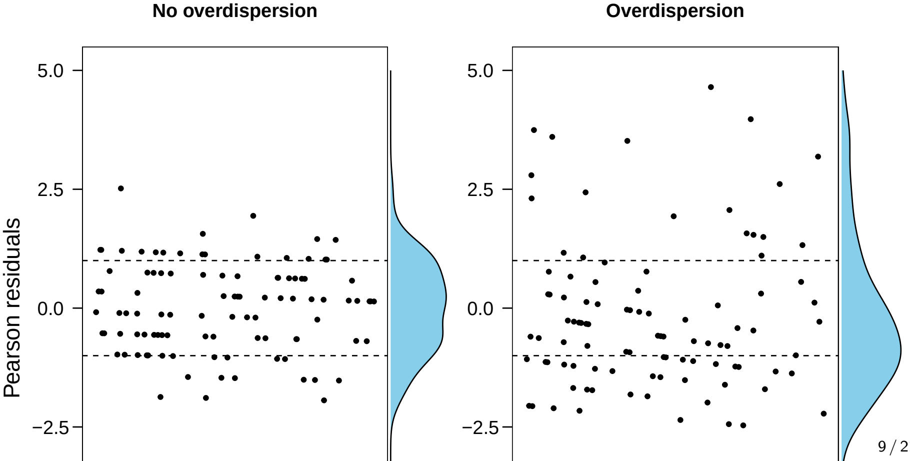
```
## [1] 1.090281
```

```
performance::check_overdispersion(fit)
```

```
## # Overdispersion test
##
##        dispersion ratio =  1.090
##   Pearson's Chi-Squared = 52.333
##                 p-value =   0.31
```

Overdispersion intuition

If data are generated from a Poisson model, the mean $\mu = \lambda$ and the variance $\sigma^2 = \lambda$. For this reason the standardized (pearson) residuals should be normally distributed with mean 0 and standard deviation 1. In the presence of overdispersion, the residuals will be larger:



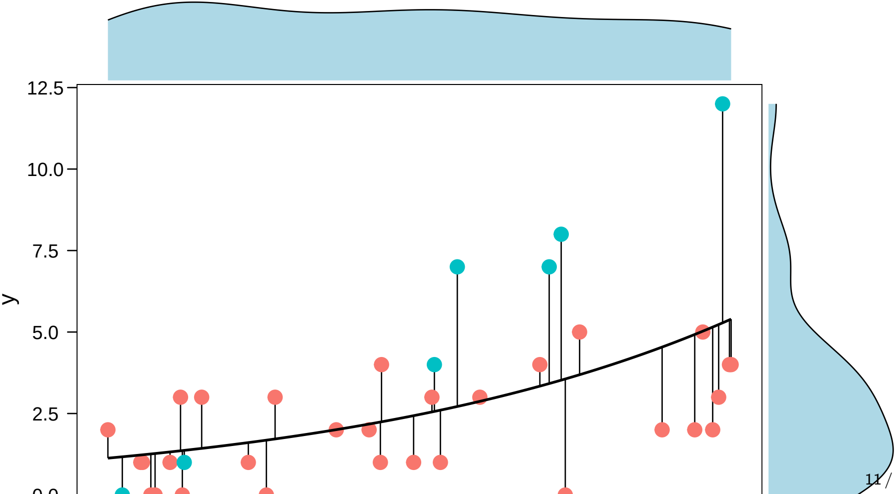
Causes of overdispersion

There could be multiple causes for overdispersion:

- outliers or anomalous obervations that increases the observed variance
- missing important variables in the model

Outliers or anomalous data

This (simulated) dataset contains $n = 30$ observations coming from a poisson model in the form $y = 1 + 2x$ and $n = 7$ observations coming from a model $y = 1 + 10x$.



Outliers or anomalous data

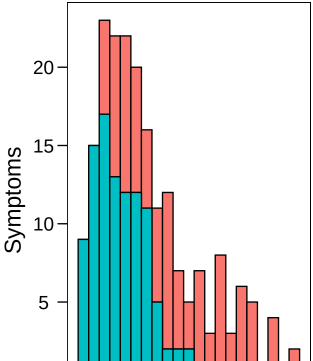
Clearly the sum of squared pearson residuals is inflated by these values producing more variance compared to what should be expected.

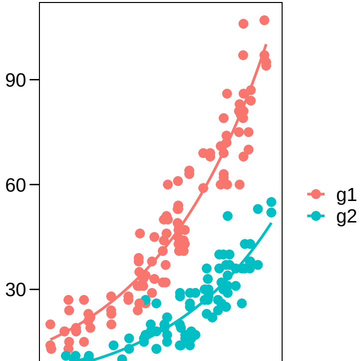
```
##     mean      var
## 2.756757 6.689189
```

```
## # Overdispersion test
##
##        dispersion ratio =  1.515
##   Pearson's Chi-Squared = 53.019
##                 p-value =  0.026
```

Missing important variables in the model

When important predictors in the model are missing, there could be more variance than expected from the Poisson model. For example, we have the relationship between number of symptoms during a month (y) and the self-report distress measure (x).





Missing important variables in the model

Despite unlikely, let's imagine to model this data using a `y ~ distress` model ignoring that there are two groups. The group is missing thus the model is estimating the average relatioship between `y ~ distress` across groups.

```
scat <- dat |>
    ggplot() +
    geom_point(aes(x = distress, y = y, color = group),
               size = 3,
               show.legend = FALSE) +
    stat_smooth(aes(x = distress, y = y),
                method = "glm",
                color = "black",
                se = FALSE,
                method.args = list(family = poisson())) +
    mytheme() +
    theme(axis.title.y = element_blank(),
          legend.title = element_blank())
```

Missing important variables in the model

Fitting the model without group will increase the observed variance creating a source of overdispersion:

```
##
## Call:
## glm(formula = y ~ distress, family = poisson(link = "log"),
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.8397  -2.2311  -0.0039   1.9859   5.3730
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.4138     0.0355   68.00   <2e-16 ***
## distress      1.9460     0.0508   38.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
##
```

Missing important variables in the model

Fitting the correct model would solve the issue taking into account that the overdispersion is no longer present when the group is considered:

```
##
## Call:
## glm(formula = y ~ distress + group, family = poisson(link =
##     data = dat)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.93239  -0.71182  -0.08319   0.60298   2.94727
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.32780    0.03573   65.14   <2e-16 ***
## distress     1.96465    0.05060   38.83   <2e-16 ***
## group1       0.79057    0.02542   31.10   <2e-16 ***
## ---
```

Why worring about overdispersion?

Overdispersion is mainly problematic because of the influence on **parameters standard error**

Dealing with overdispersion

Dealing with overdispersion

If all the variables are included and no outliers are present, the phenomenon itself contains more variability respect to the Poisson model predictions. There are two main approaches to deal with the situation:

- quasi-poisson model
- poisson-gamma model AKA negative-binomial model

Quasi-poisson model

The quasi-poisson model allow estimating an extra parameter that is the overdispersion

```
## lambda = 20.000, var = 20.000, theta = 10.000

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   14.00   19.00   20.01   25.00   72.00

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00   17.00   20.00   19.99   23.00   43.00

## [1] 20.00989

## [1] 60.04267

## [1] 19.99419

## [1] 20.00976
```

References