Titolo

Filippo Gambarota

University of Padova

2022/2023

Outline

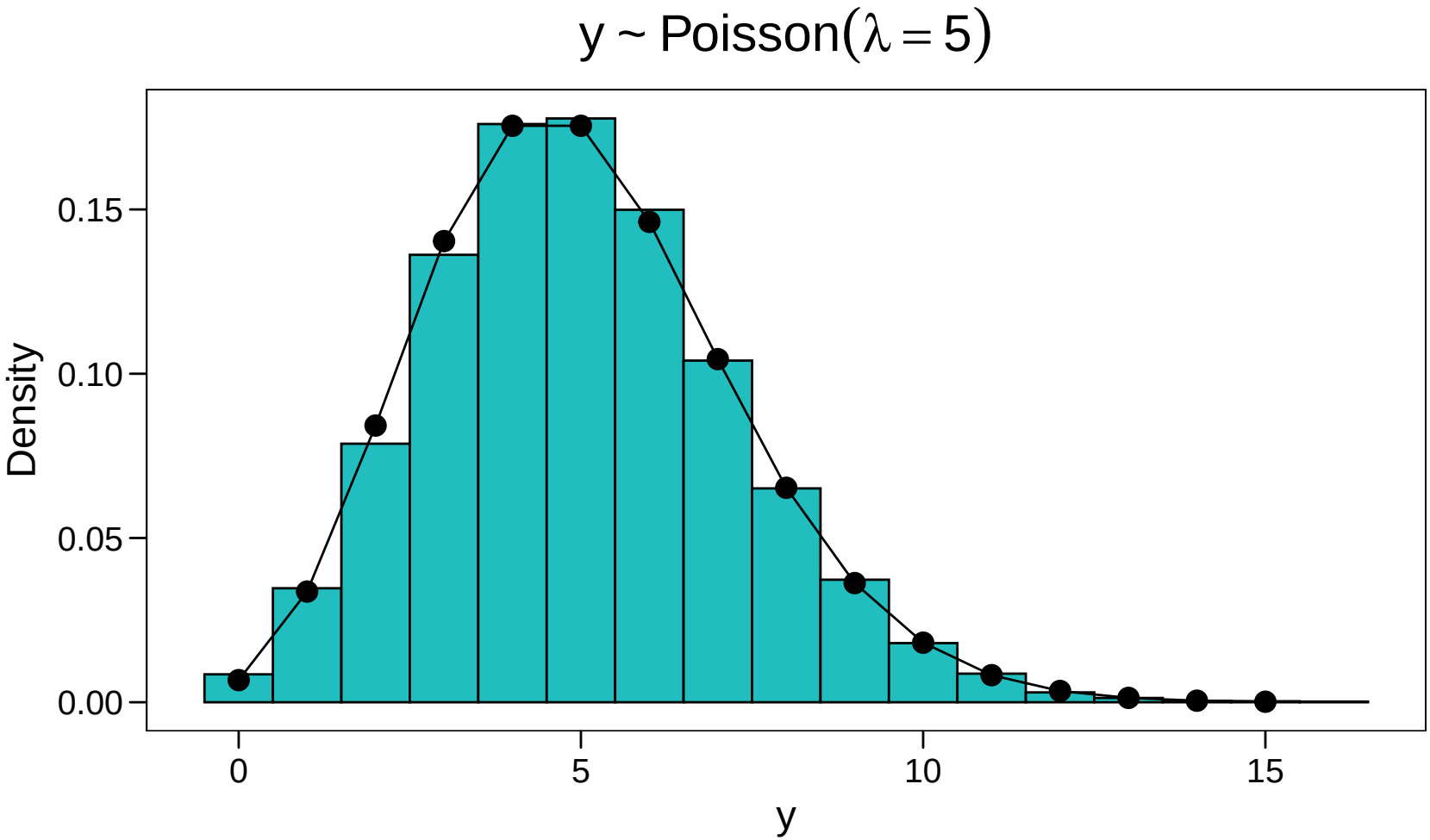
Poisson distribution

The Poisson distribution is defined as:

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

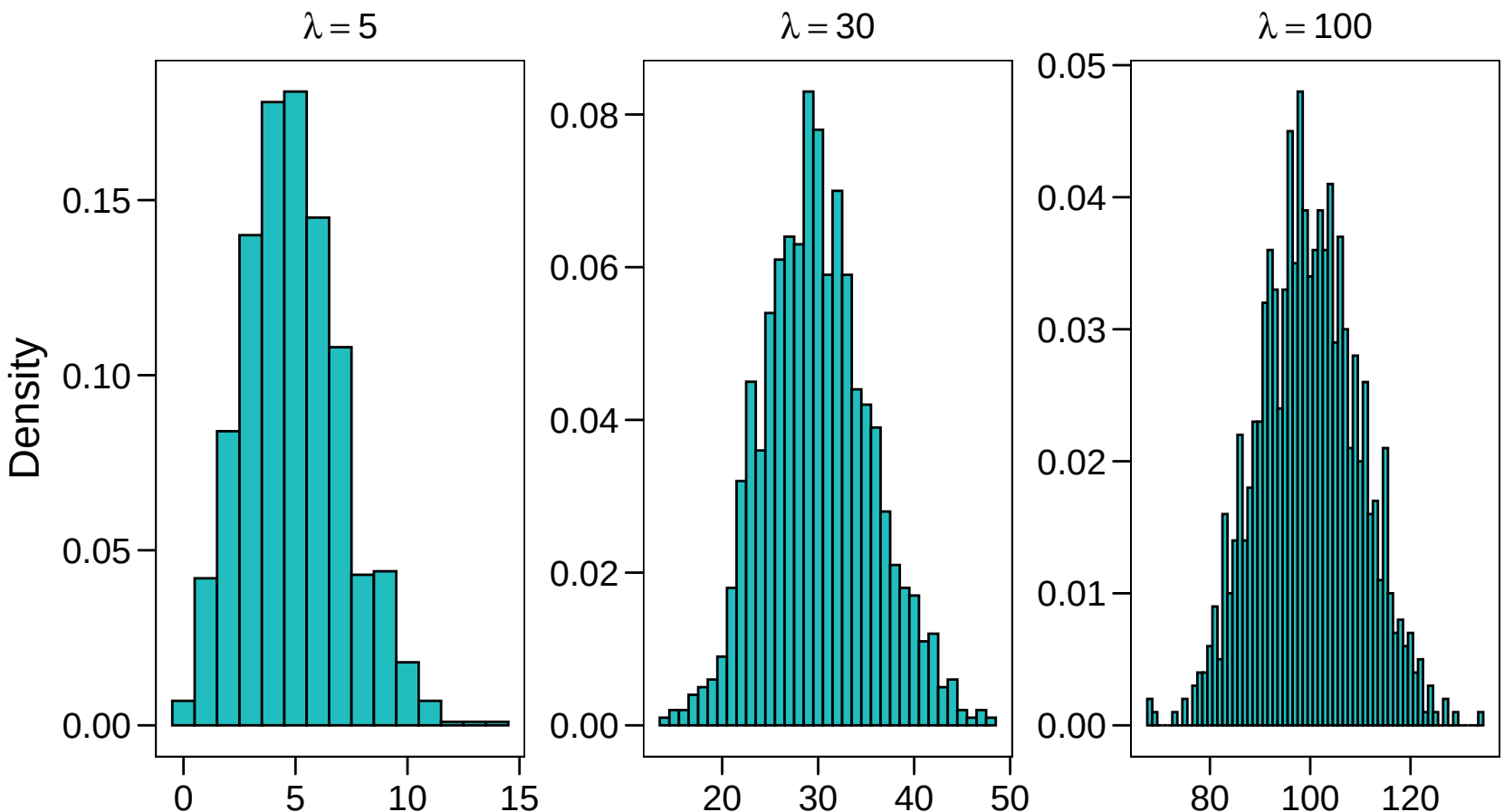
Where the mean is μ and the variance is μ

Poisson distribution

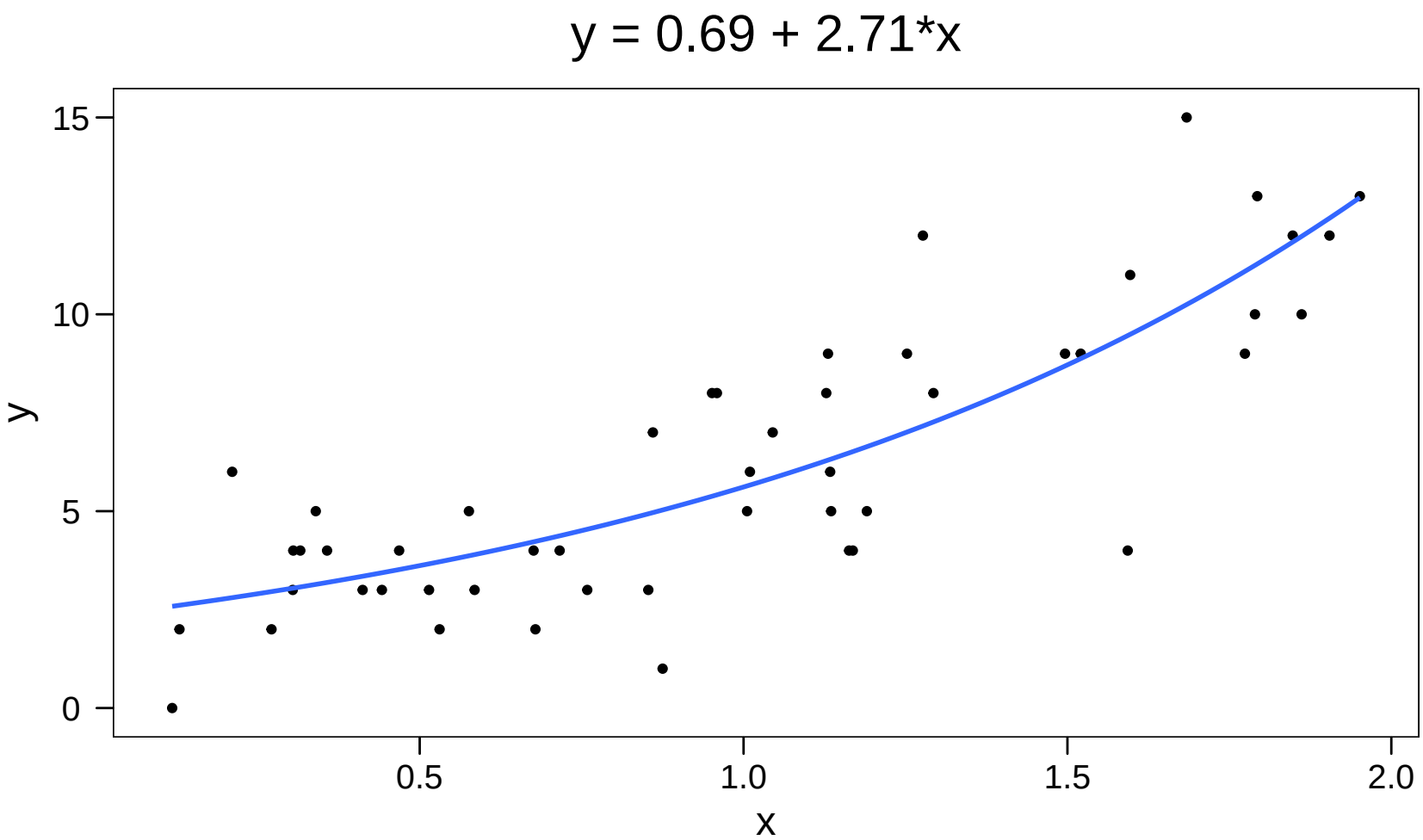


Poisson distribution

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] 0.7181773
```

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

```
## # Overdispersion test
##
##        dispersion ratio =  0.718
##   Pearson's Chi-Squared = 34.473
##                 p-value =  0.929
```

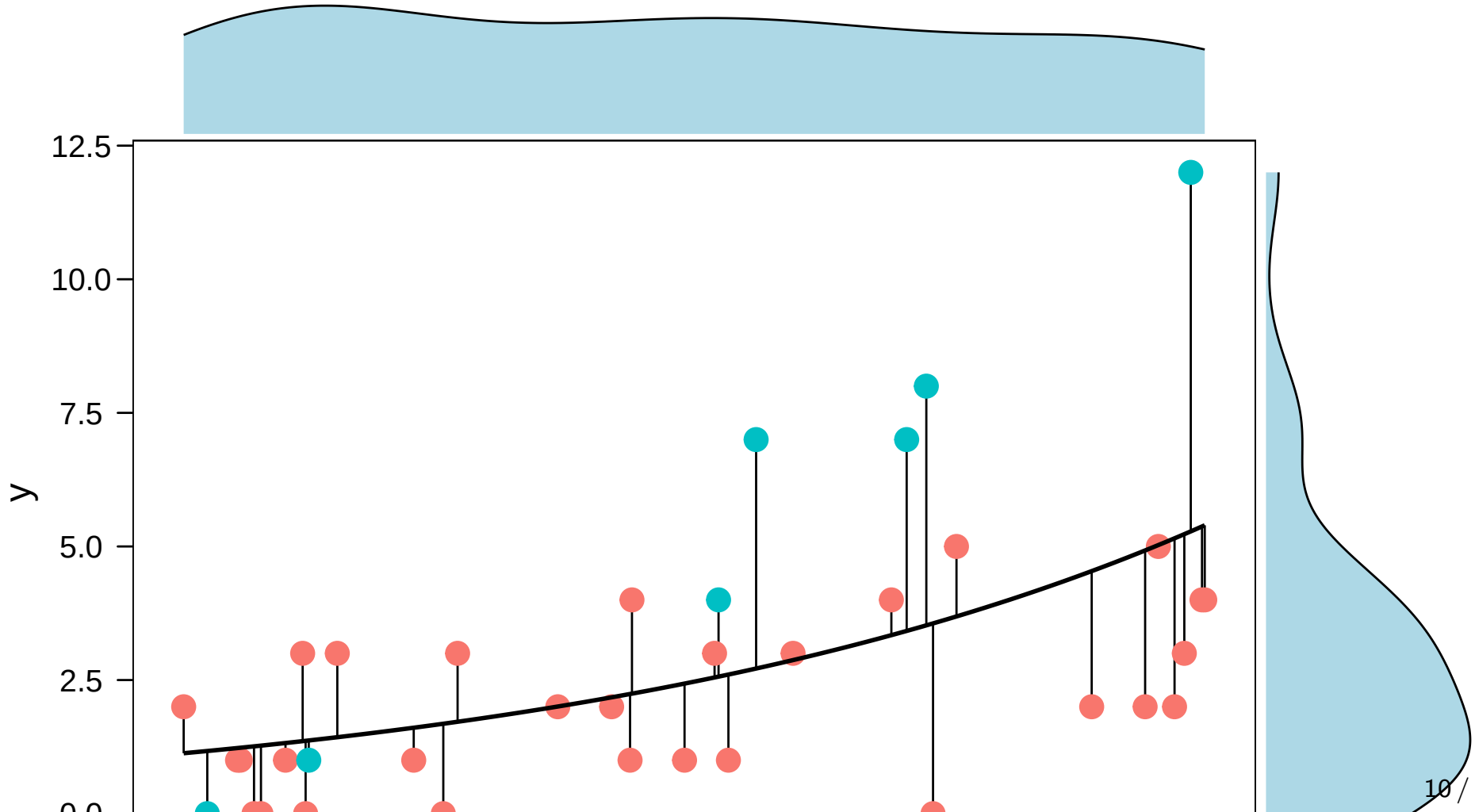
Causes of overdispersion

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
```

References