Introduction to Generalized Linear Models

Filippo Gambarota

University of Padova

2022/2023

Outline

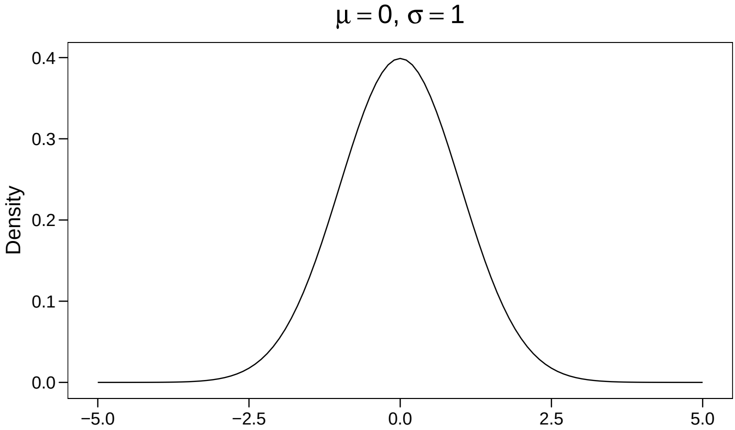
Beyond the Gaussian distribution

Quick recap about Gaussian distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Were μ is the **mean** and σ is the **standard deviation**

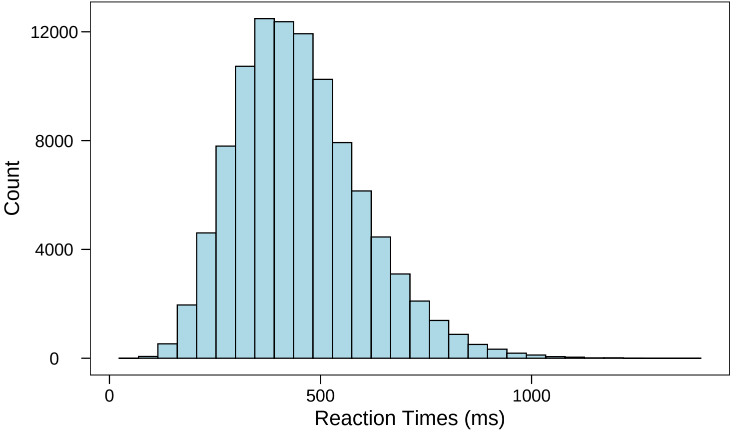
Quick recap about Gaussian distribution



But not always gaussian-like variables!

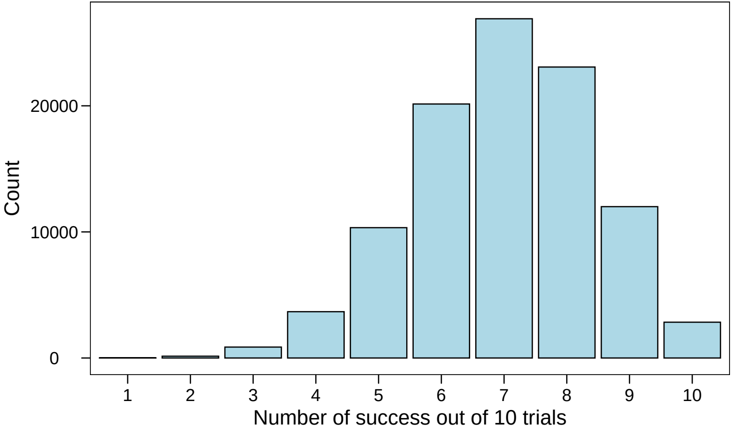
Reaction times

Measuring reaction times during a cognitive task. Non-negative and proably skewed data.

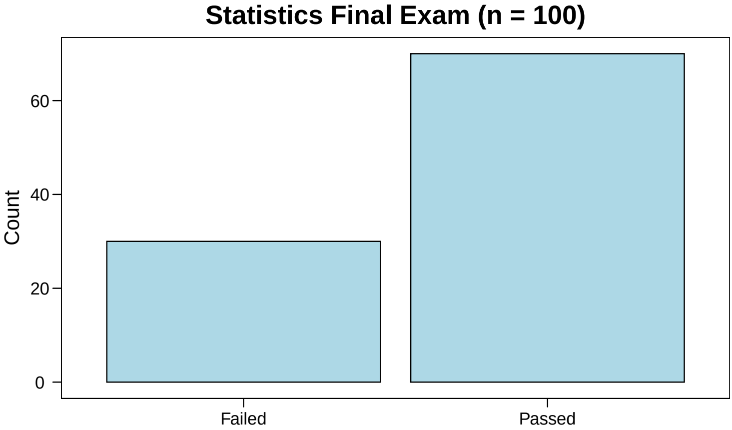


Binary outcomes

Counting the number of people passing the exam out of the total.
Discrete and non-negative. A series of binary (i.e., *bernoulli*) experiments.

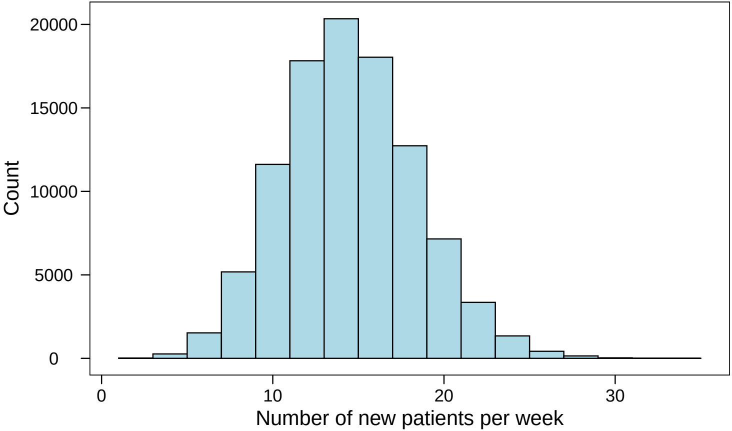


Binary outcomes

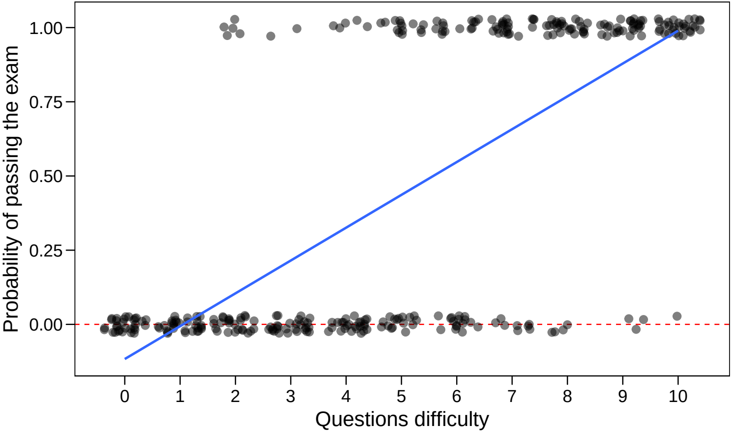


Counts

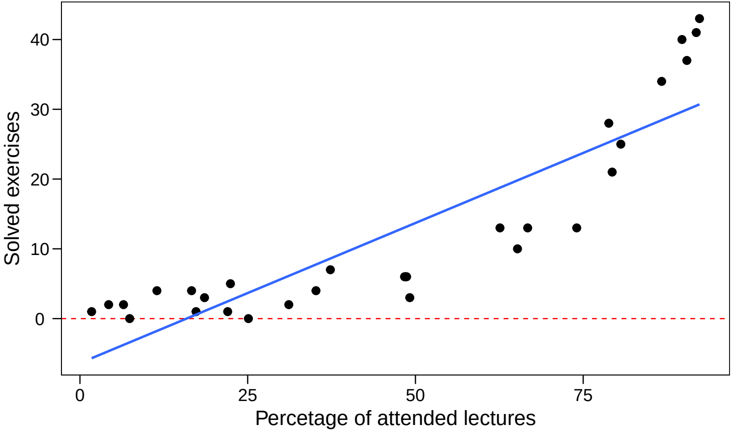
Counting the number of new patients per week. Discrete and non-negative values.



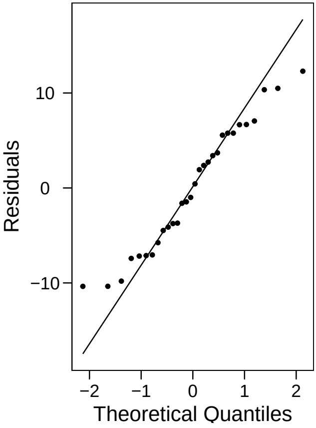
Should we use a linear model for these variables?

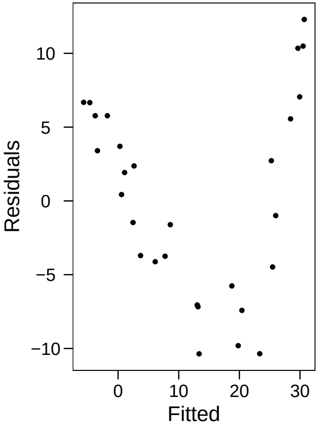


Should we use a linear model for these variables?



Should we use a linear model for these variables?





A new class of models

- We need that our model take into account the **features of our response variable**
- We need a model that, **with appropriate transformation**, keep **properties of standard linear models**
- We need a model that is **closer to the true data generation process**

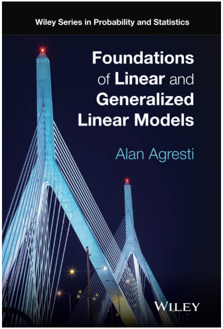
Let's switch to Generalized Linear Models!

Generalized Linear Models

Main references

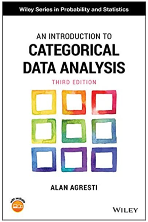
For a detailed introduction about GLMs

- Chapters: 1 (intro), 4 (GLM fitting), 5 (GLM for binary data)



For a basic and well written introduction about GLM, especially the Binomial GLM

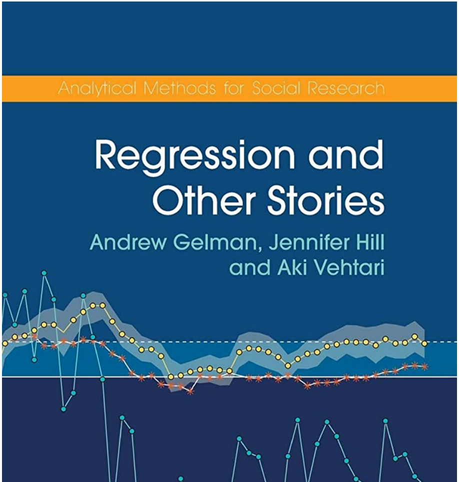
- Chapters: 3 (intro GLMs), 4-5 (Binomial Logistic Regression)



Main references

Great resource for interpreting Binomial GLM parameters:

- Chapters: 13-14 (Binomial Logistic GLM), 15 (Poisson and others GLMs)



General idea

- models that assume distributions other than the normal distributions
- models that considers non-linear relationships

Recipe for a GLM

- **Random Component**
- **Systematic Component**
- **Link Function**

Random Component

The **random component** of a GLM identify the response variable Y and the appropriate probability distribution. For example for a numerical and continous variable we could use a Normal distribution (i.e., a standard linear model). For a discrete variable representing counts of events we could use a Poisson distribution.

Systematic Component

The **systematic component** or *linear predictor* of a GLM is the combination of explanatory variables i.e. $\beta_0 + \beta_1 x_1 + ... + \beta_p x_p$.

Link Function

The **link function** $g(\mu)$ is the function that connects the expected value (i.e., the mean μ) of the probability distribution (i.e., the random component) with the *linear combination* of predictors $g(\mu) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$

The simplest **link function** is the **identity link** where $g(\mu) = \mu$ and correspond to the classic linear model. In fact, the linear regression is just a GLM with a **Gaussian random component** and the **identity** link function.

There are multiple **random components** and **link functions** for example with a $0/1$ binary variable the usual choice is using a **Binomial** random component and the **logit** link function.

Relevant distributions

Binomial distribution

The probability of having k success (e.g., 0, 1, 2, etc.) out of n trials with a probability of success p is:

$$f(n,k,p) = Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

The np is the mean of the binomial distribution and $np(1-p)$ is the variance.

Bernoulli distribution

The **binomial** distribution is just a repetition of k **Bernoulli** trials. A single Bernoulli trial is:

$$
\begin{aligned}
f(x,p) = p^x(1-p)^{1-x} \\
x \in \{0,1\}
\end{aligned}
$$

The mean is p and the variance is $p(1-p)$

Bernoulli and Binomial

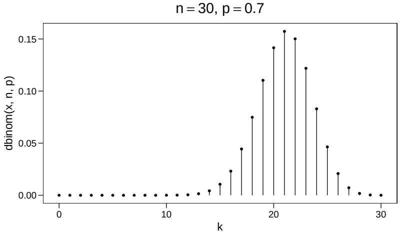
The simplest situation for a Bernoulli trial is a coin flip. In R:

```
n <- 1
p <- 0.7
rbinom(1, n, p) # a single bernoulli trial
```

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

```
n <- 10
rbinom(1, 10, p) # n bernoulli trials
```

```
## [1] 9
```

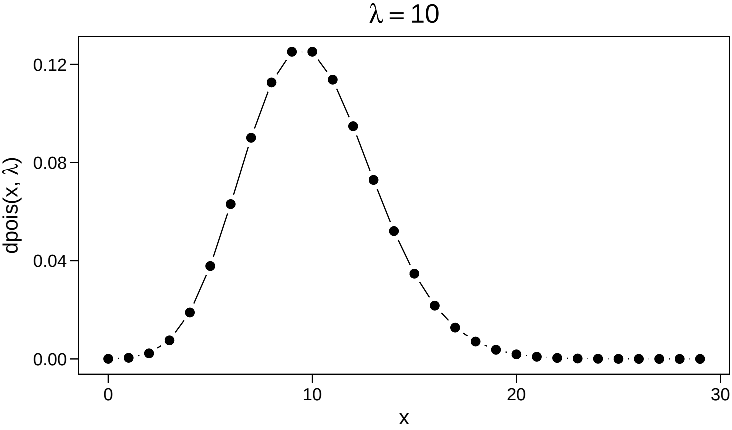


Poisson distribution

The number of events k during a fixed time interval (e.g., number of new user on a website in 1 week) is:

$$f(j, \lambda) = Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Poisson distribution



The mean and also the variance is λ.

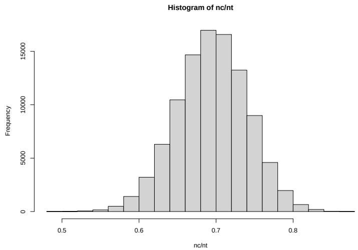
Data simulation `#extra`

Data simulation #extra

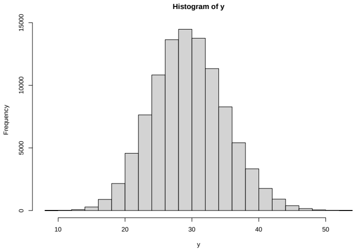
- During the course we will try to simulate some data. Simulating data is an amazing education tool to understand a statistical model.
- By simulating from a **generative model** we are doing **Monte Carlo Simulations** [1]

Data simulation #extra

```r
n <- 1e5 # number of experiments
nt <- 100 # number of subjects
p <- 0.7 # probability of success
nc <- rbinom(n, nt, p)
```




```r
n <- 1e5 # number of subjects
lambda <- 30 # mean/variance
y <- rpois(n, lambda)
```



[1] J. E. Gentle, "Monte carlo methods for statistical inference," in *Computational statistics*, J. E. Gentle, Ed., New York, NY: Springer New York, 2009, pp. 417–433. doi: 10.1007/978-0-387-98144-4_11.

filippo.gambarota@unipd.it

github.com/filippogambarota