

# Introduction to Generalized Linear Models

Filippo Gambarota

University of Padova

2022/2023

{Updated on 2023-04-24}

# Outline

1. Beyond the Gaussian distribution
2. Generalized Linear Models
3. Relevant distributions
4. Data simulation #extra

## Beyond the Gaussian distribution

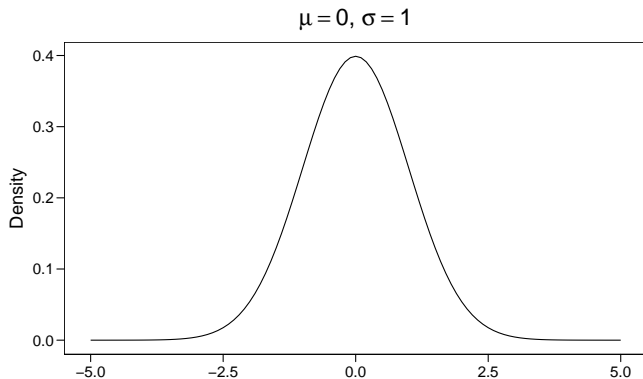
# Quick recap about Gaussian distribution

- The Gaussian distribution is part of the Exponential family
- It is defined with mean ( $\mu$ ) and the standard deviation ( $\sigma$ ) that are independent
- It is symmetric with the same value for mean, mode and median
- The support is  $[-\infty, +\infty]$

The Probability Density Function (PDF) is:

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

# Quick recap about Gaussian distribution



But not always gaussian-like variables!

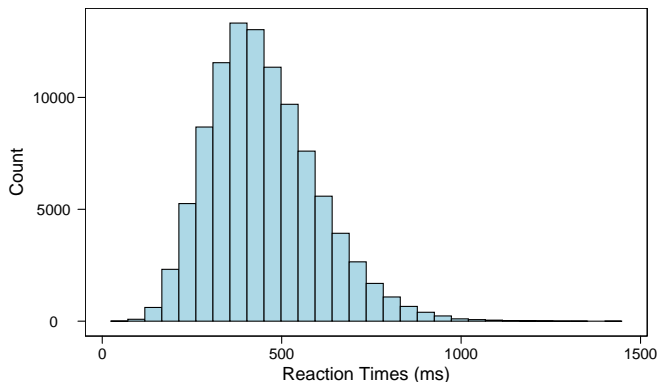
# Quick recap about Gaussian distribution

In fact, in Psychology, variables do not always satisfy the properties of the Gaussian distribution. For example:

- Reaction times
- Accuracy
- Percentages or proportions
- Discrete counts
- Likert scales

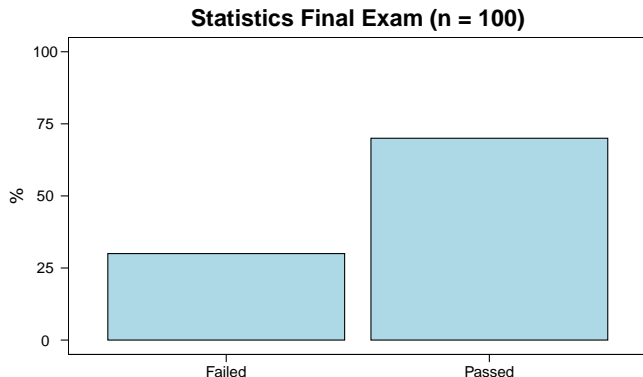
# Reaction times

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



# Binary outcomes

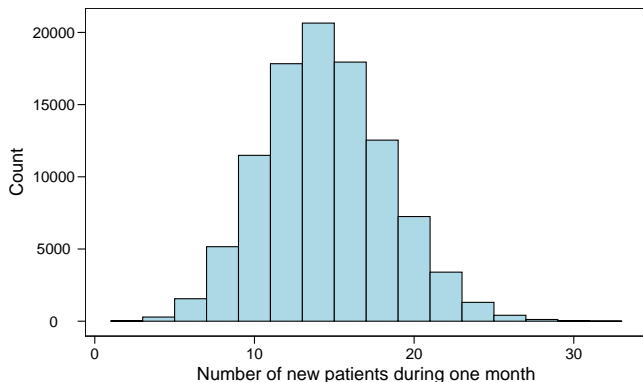
Counting the number of people passing the exam out of the total.  
Discrete and non-negative. A series of binary (i.e., *bernoulli*) experiments.





# Counts

Counting the number of new hospitalized patients during one month in different cities. Discrete and non-negative values.



**Should we use a linear model for these variables?**

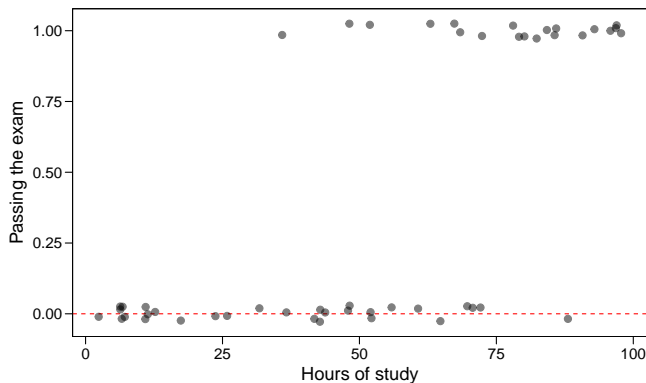
# Should we use a linear model for these variables?

Let's try to fit a linear model on the probability of passing the exam as a function of the hours of study:

student	study.hours	passing
1	65	0
2	72	0
3	97	1
4	7	0
...	...	...
47	48	1
48	36	1
49	11	0
50	70	0

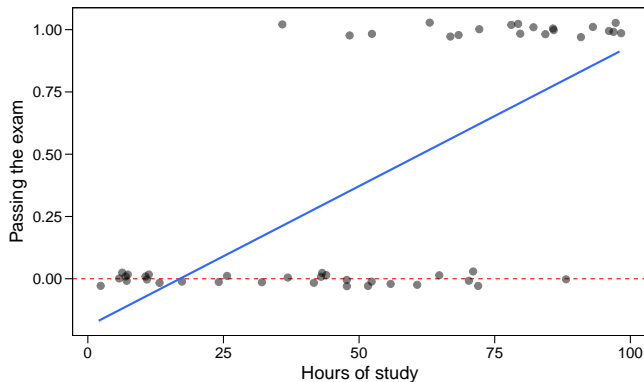
# Should we use a linear model for these variables?

Let's plot the data:



# Should we use a linear model for these variables?

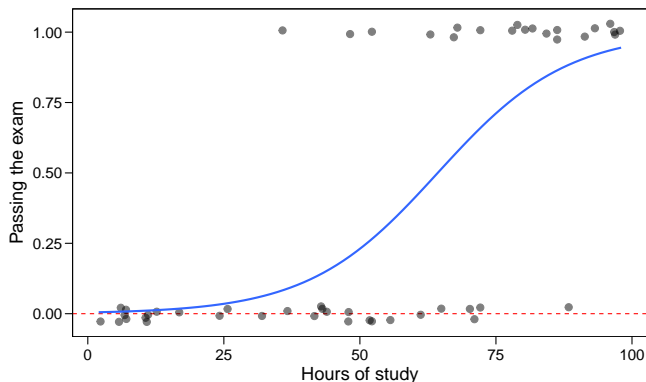
Let's fit a linear model passing `~ study_hours` using `lm`:



Do you see something strange?

# Should we use a linear model for these variables?

A little **spoiler**, the relationship should be probably like this:

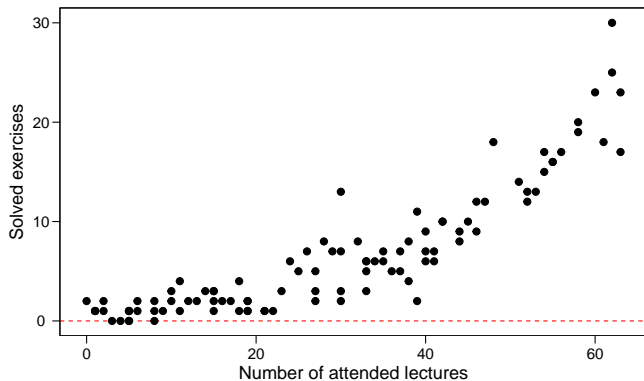


# Should we use a linear model for these variables?

Another example, the number of solved exercises in a semester as a function of the number of attended lectures:

student	attended.lectures	nsolved
1	38	4
2	16	2
3	42	10
4	33	6
...	...	...
97	11	4
98	41	6
99	33	5
100	45	10

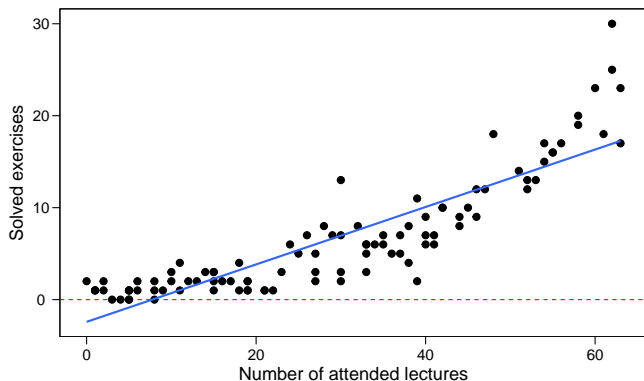
# Should we use a linear model for these variables?





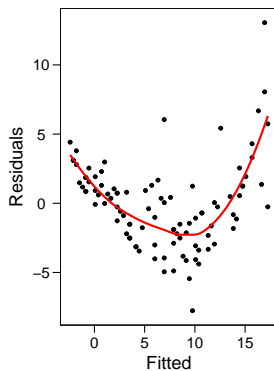
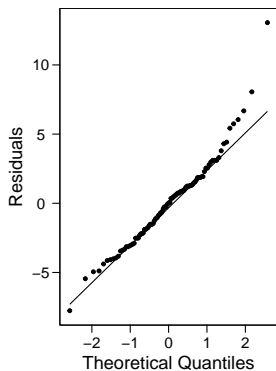
# Should we use a linear model for these variables?

Again, fitting the linear model seems partially appropriate but there are some problems:



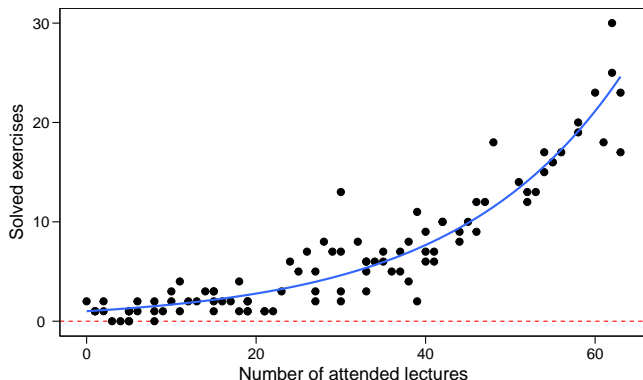
# Should we use a linear model for these variables?

Also the residuals are quite problematic:



# Should we use a linear model for these variables?

Another little spoiler, the model should consider both the support of the y variable and the non linear pattern. Probably something like this:



# So what?

Both linear models somehow capture the expected relationship but there are serious fitting problems:

- impossible predictions
- poor fitting for non-linear patterns

As a general rule in statistics:

*"All models are wrong, some are useful."*

— George Box

# We need 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**

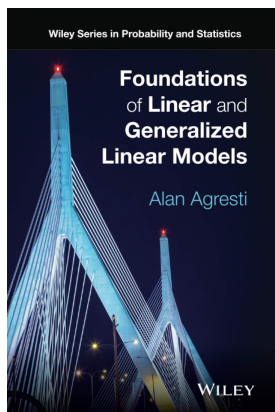
Let's switch to Generalized Linear Models!

## Generalized Linear Models

# Main references

For a detailed introduction about GLMs

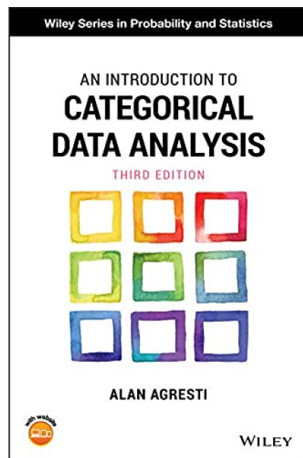
- Chapters: 1 (intro), 4 (GLM fitting), 5 (GLM for binary data)



# Main references

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

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

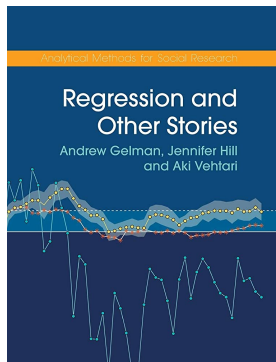




# Main references

Great resource for interpreting Binomial GLM parameters:

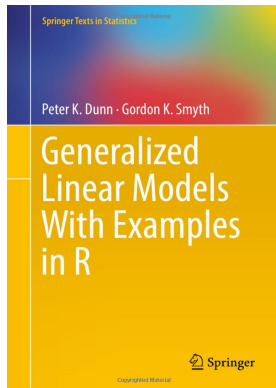
- Chapters: 13-14 (Binomial Logistic GLM), 15 (Poisson and others GLMs)



# Main references

Detailed GLMs book. Very useful especially for the diagnostic part:

- Chapters: 8 (intro), 9 (Binomial GLM), 10 (Poisson GLM and overdispersion)



# 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 continuous 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, etc.

# Systematic Component

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

$$\eta = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

# Link Function

- The **link function**  $g(\mu)$  is the function that connects the expected value (i.e., the mean  $\mu$ ) of the probability distribution (i.e., the random component) with the *linear combination* of predictors  $g(\mu) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$ . The inverse of the link function  $g^{-1}$  map the linear predictor ( $\eta$ ) into the original scale.

$$g(\mu) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$
$$\mu = g^{-1}(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)$$

# Link function

- The simplest **link function** is the **identity link** where  $g(\mu) = \mu$  and correspond to the standard 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.



# Common link functions

Family	Link	Range
gaussian	identity	$(-\infty, +\infty)$
binomial	logit	$\frac{0, 1, \dots, n_i}{n_i}$
binomial	probit	$\frac{0, 1, \dots, n_i}{n_i}$
poisson	log	$0, 1, 2, \dots$

## 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:

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

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

# Bernoulli and Binomial

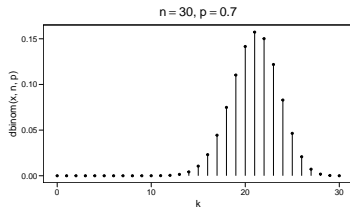
The simplest situation for a Bernoulli trial is a coin flip. In R:

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

```
## [1] 0
```

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

```
## [1] 6
```

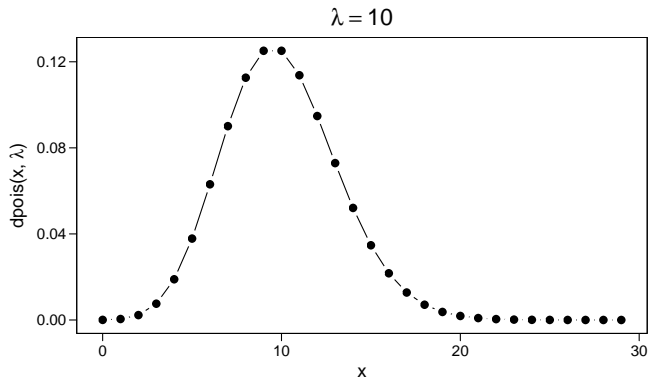


# Poisson distribution

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  $\lambda$ .

Data simulation #extra

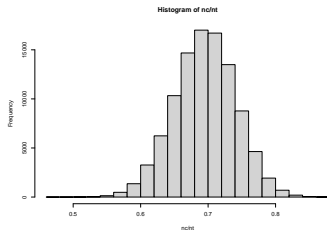


# Data simulation #extra

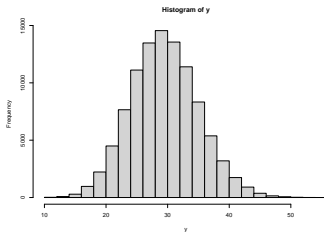
- During the course we will try to simulate some data. Simulating data is an amazing education tool to understand a statistical model.
- By simulating from a **generative model** we are doing **Monte Carlo Simulations** [1]

# Data simulation #extra

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




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



# References

- [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](mailto:filippo.gambarota@unipd.it)

 [github.com/filippogambarota](https://github.com/filippogambarota)