

# Dados longitudinais: introdução a *generalized estimating equations* (GEE)

Pedro S.R. Martins

Grupo de estudos em estatística 2023



# Primeiras coisas primeiro



**Produção e Divulgação Científica**  
www.cientifica.com  
http://naruhodo.b9.com.br

**Científica & Podcast Naruhodo**


HOME VIDEOS PLAYLISTS COMMUNITY CHANNELS ABOUT

**BEM VINDO(A) AO CANAL CIENTÍFICA**

**Apresentação Canal Científica e Podcast Naruhodo no YT**  
Científica & Podcast Naruhodo • 3.7K views • 1 year ago  
Canal Científica Produção Científica com base em vídeos de Estatística Aplicada, Metodologia e Epistemologia www.cientifica.com Podcast Naruhodo! Naruhodo! é o podcast pra quem tem...

**Lives Estatística Nível I Psicobio UNIFESP 2021** ▶ PLAY ALL

Thumbnail	Video Title	Views	Time
	Estatística Psicobio I 2021 #01 - Teoria da Medida...	3.1K views	1:53:20
	Estatística Psicobio I 2021 #02 - Tipos de Variável e...	1.4K views	2:35:21
	Estatística Psicobio I 2021 #03 - Medidas Descritivas...	1.1K views	2:29:37
	Estatística Psicobio I 2021 #04 - TCL e Intervalos de...	988 views	2:31:33
	Estatística Psicobio I 2021 #05 - Intervalos de...	790 views	2:29:33



**Produção e Divulgação Científica**  
www.cientifica.com  
http://naruhodo.b9.com.br

**Científica & Podcast Naruhodo**

PLAY ALL

**Lives - Disciplina Estatística aplicada a Psicobiologia Nível II - UNIFESP**  
28 videos • 5,713 views • Last updated on Dec 7, 2020

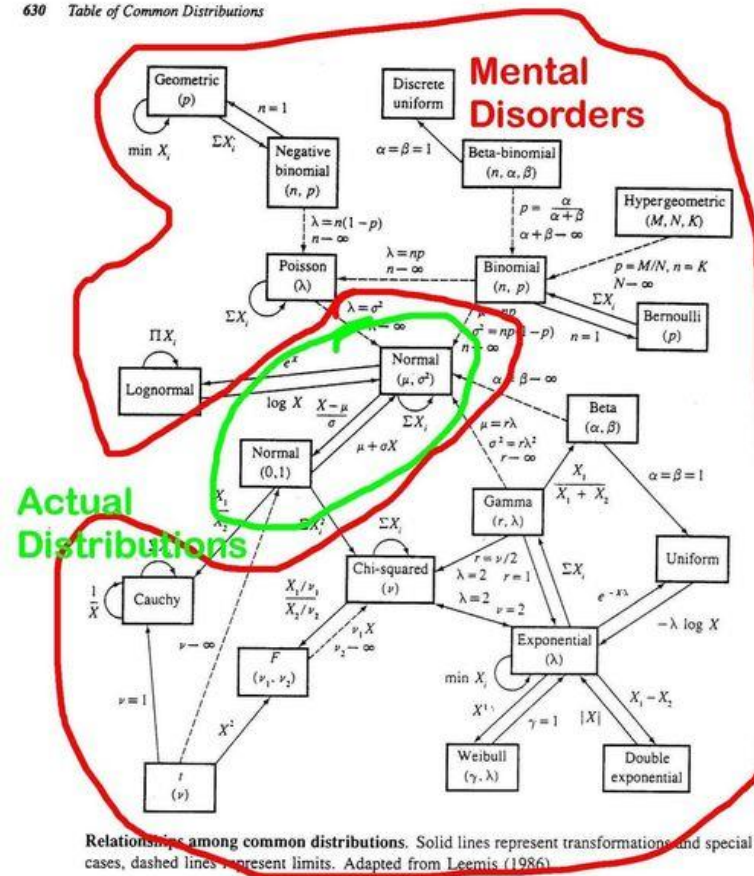
Thumbnail	Video Title	Views	Time
	Estatística Psicobio II - Revisão Regressão e o início do pensamento científico		2:19:30
	Estatística Psicobio II - Escrevendo modelos - Distribuições para variáveis discretas Parte I		2:40:56
	Estatística Psicobio II - Distribuições para variáveis discretas Parte II - Bin, BinNeg, Poisson		2:29:25
	Estatística Psicobio II - Distribuições para variáveis contínuas		2:32:12
	Estatística Psicobio II - GEE - Generalized Estimated Equations - Parte I		2:35:51
	Estatística Psicobio II - GEE - Generalized Estimated Equations - Parte II		2:24:38



# Modelos generalizados

- Retomando a ideia

630 Table of Common Distributions



# Leitura recomendada

## CAPÍTULO 5

## CHAPTER 5

### Variáveis Aleatórias e Distribuições de Probabilidade

*Aquilo a que chamamos acaso não é, e não pode deixar de ser, senão a causa ignorada de um efeito conhecido.*  
Voltaire

Ao final deste capítulo, você será capaz de:

- Compreender os conceitos relativos as variáveis aleatórias discretas e contínuas.
- Calcular a esperança, a variância e a função de distribuição acumulada de variáveis aleatórias discretas e contínuas.
- Descrever os principais tipos de distribuição de probabilidades para variáveis aleatórias discretas: uniforme discreta, Bernoulli, binomial, geométrica, binomial negativa, hipergeométrica e Poisson.
- Descrever os principais tipos de distribuição de probabilidades para variáveis aleatórias contínuas: uniforme, normal, exponencial, Gama, qui-quadrado ( $\chi^2$ ), *t* de Student e *F* de Snedecor.
- Determinar a distribuição mais adequada para determinado conjunto de dados.

5.1. **Fávero, L. P., & Belfiore, P. (2017). *Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®*. Elsevier Brasil.**

mente, a distribuição de probabilidade da mesma variável para a população (Martins e Domingues, 2011).

### Special random variables

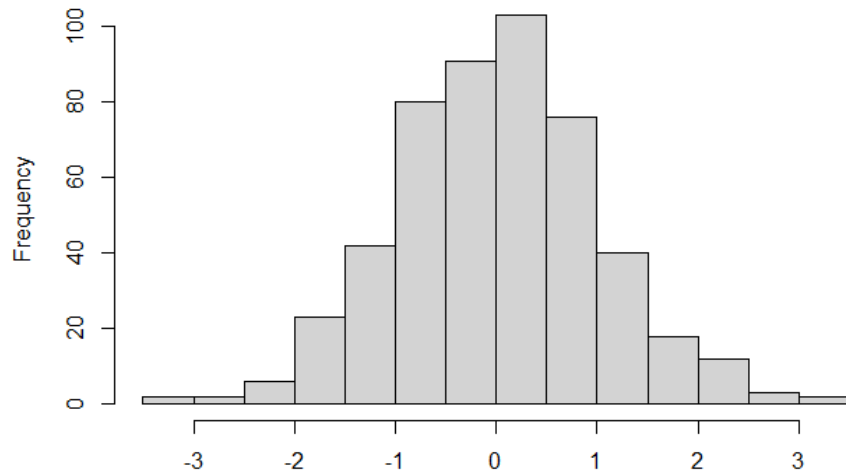
Certain types of random variables occur over and over again in applications. In this chapter, we will study a variety of them.

#### 5.1 The Bernoulli and binomial random variables

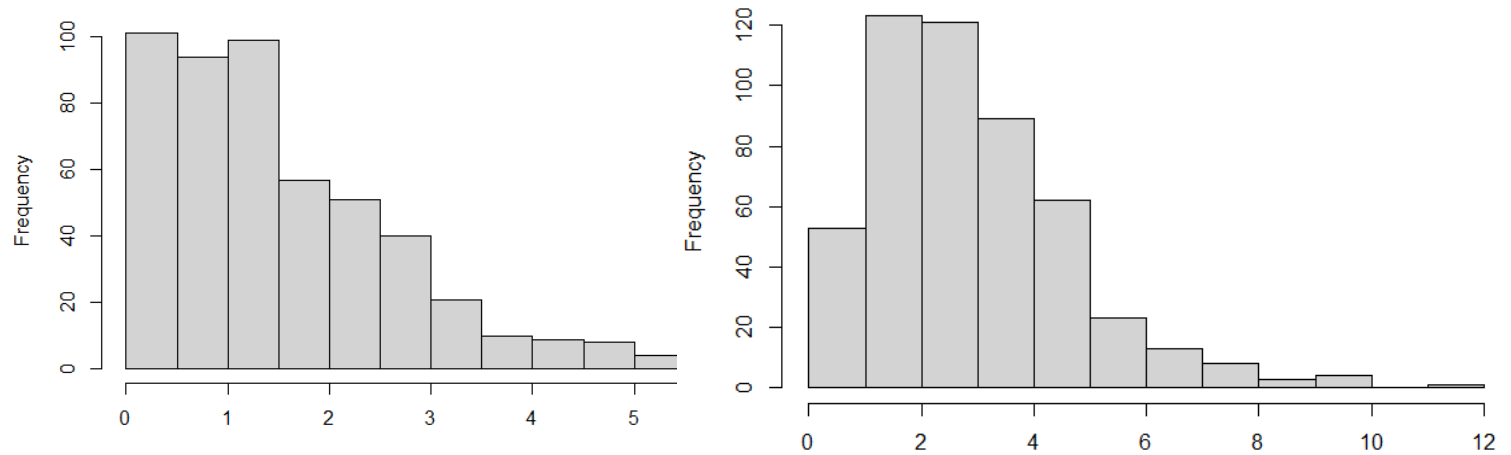
Ross, S. M. (2020). *Introduction to probability and statistics for engineers and scientists*. Academic press.

# Diferentes distribuições (arranhando a superfície)

Distribuição normal



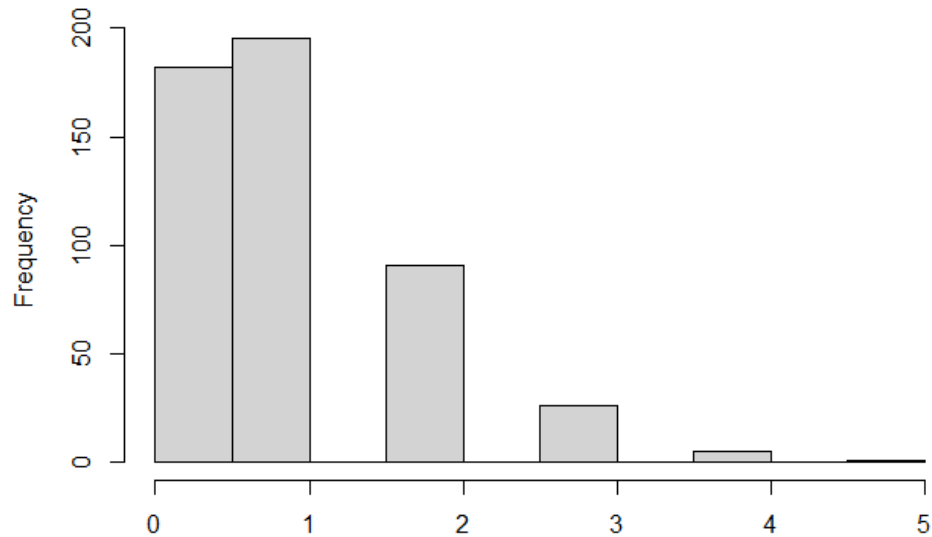
Distribuição Gamma



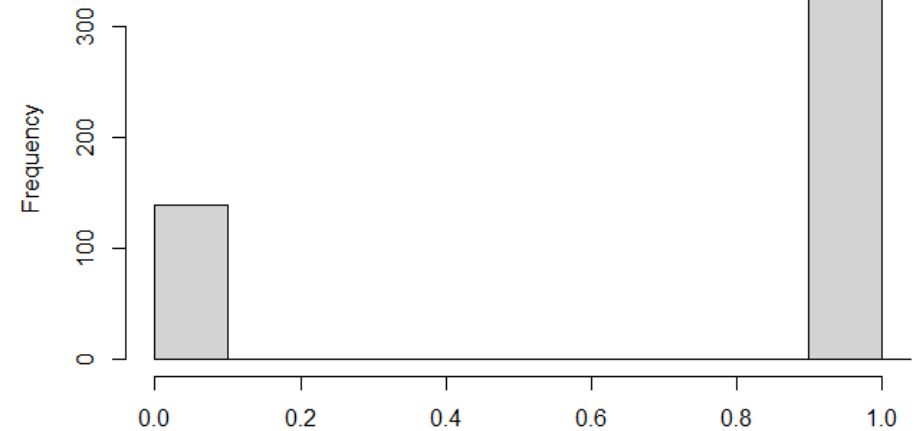
```
rnorm(500,0,1) %>% hist(., main = NULL) ;  
rgamma(500,shape = 2) %>% hist(., main = NULL);  
rgamma(500,shape = 1.5) %>% hist(., main = NULL);  
rgamma(500,shape = 3) %>% hist(., main = NULL)
```

# Diferentes distribuições

Distribuição poisson



Distribuição binomial



```
rpois(500,1) %>% hist(., main = NULL);  
rbinom(500,prob = .7, size = 1) %>% hist(., main = NULL)
```

# Outras distribuições

- Gaussiana invertida
- Binomial negativa
- Tweedie
- Multinomial



# Generalized estimating equations

*Biometrika* (1986), 73, 1, pp. 13–22  
Printed in Great Britain

13

## Longitudinal data analysis using generalized linear models

BY KUNG-YEE LIANG AND SCOTT L. ZEGER

*Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205, U.S.A.*

### SUMMARY

This paper proposes an extension of generalized linear models to the analysis of longitudinal data. We introduce a class of estimating equations that give consistent estimates of the regression parameters and of their variance under mild assumptions about the time dependence. The estimating equations are derived without specifying the joint distribution of a subject's observations yet they reduce to the score equations for multivariate Gaussian outcomes. Asymptotic theory is presented for the general class of estimators. Specific cases in which we assume independence,  $m$ -dependence and exchangeable correlation structures from each subject are discussed. Efficiency of the proposed estimators in two simple situations is considered. The approach is closely related to quasi-likelihood.

*Some key words:* Estimating equation; Generalized linear model; Longitudinal data; Quasi-likelihood; Repeated measures.

### 1. INTRODUCTION

Longitudinal data sets, comprised of an outcome variable,  $y_{it}$ , and a  $p \times 1$  vector of covariates,  $x_{it}$ , observed at times  $t = 1, \dots, n_i$  for subjects  $i = 1, \dots, K$  arise often in applied sciences. Typically, the scientific interest is either in the pattern of change over time, e.g. growth, of the outcome measures or more simply in the dependence of the outcome on the covariates. In the latter case, the time dependence among repeated measurements for a subject is a nuisance. For example, the severity of respiratory disease along with the nutritional status, age, sex and family income of children might be observed once every three months for an 18 month period. The dependence of the outcome variable, severity of disease, on the covariates is of interest.

With a single observation for each subject ( $n_i = 1$ ), a generalized linear model (McCullagh & Nelder, 1983) can be applied to obtain such a description for a variety of continuous or discrete outcome variables. With repeated observations, however, the correlation among values for a given subject must be taken into account. This paper presents an extension of generalized linear models to the analysis of longitudinal data when regression is the primary focus.



Kung-Yee Liang



Scott L. Zeger

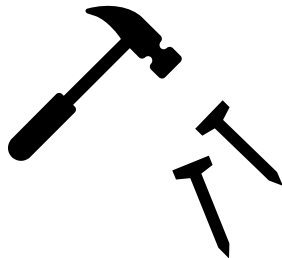
KUNG-YEE LIANG, SCOTT L. ZEGER, Longitudinal data analysis using generalized linear models, *Biometrika*, Volume 73, Issue 1, April 1986, Pages 13–22, <https://doi.org/10.1093/biomet/73.1.13>



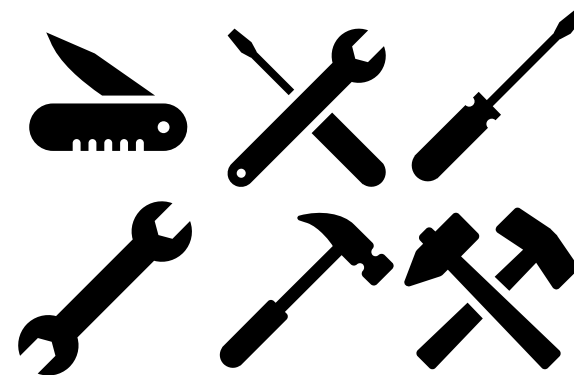
# Generalized estimating equations

- **Definição:** GEE é uma extensão de modelos lineares generalizados para estimar estimativas médias para a população controlando para a dependência entre medidas repetidas
- GEE assume um formato de equação em que a variável dependente pode ser de qualquer distribuição, e a interpretação varia de acordo com a função de ligação que escolhemos

ANOVA



GEE



# Usos GEE

- Podemos pensar na GEE como uma regressão longitudinal
  - Ou seja, erros correlacionados
- Mas, para isso, é necessário **conhecer a variável dependente**

- VD** {
- Variável dicotômica: regressão logística
  - Variável contínua: (1) distribuição normal (2) distribuição gamma
  - Variável de contagem: regressão Poisson
  - ...

# Covariáveis/preditores

- Basicamente, dois tipos de covariáveis

## Invariante no tempo (time-invariant)

- Variáveis estáveis (sexo, idade em um do pontos, outros)
- Variáveis medidas apenas uma vez devido ao design do experimento

## Variável com o tempo (time-varying)

- Variáveis que mudam ao longo do tempo
- Variáveis medidas várias vezes ao longo do experimento

# Pressupostos GEE

- As respostas na variável dependente são correlacionadas (ou seja, um indivíduo responde mais de uma vez)
- Covariáveis existem como um função linear ou não linear, podendo interagir entre si
- A homogeneidade da variância não é satisfeita
- O usuário especifica a matriz de correlação correta dos erros (working correlation matrix)
  - (a) especificar a função de ligação
  - (b) conhecer a distribuição da variável dependente
  - (c) conhecer a distribuição das correlações

# Matriz de correlação dos erros

1. Independente (sem correlação entre tempos)
2. Intercambiável (exchangeable, a.k.a simétrica): as correlações são as mesmas entre os tempos
3. Autorregressiva de ordem 1 (AR1): assume que as correlações das observações são dependentes dos valores anteriores apenas, por meio de um processo autorregressivo
4. Matriz não estruturada: não assume *constraints* para as correlações par a par entre os tempos

# Independente

$$\begin{array}{c} \text{T1} \\ \text{T2} \\ \text{T3} \end{array} \begin{array}{ccc} \text{T1} & \text{T2} & \text{T3} \\ \left[ \begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array} \right] \end{array}$$

Dados transversais normalmente

# Intercambiável (simétrica)

$$\begin{array}{c} \text{T1} \\ \text{T2} \\ \text{T3} \end{array} \begin{array}{ccc} \text{T1} & \text{T2} & \text{T3} \\ \left[ \begin{array}{ccc} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{array} \right] \end{array}$$

Efeitos lineares

Correlação é a mesma entre todos os tempos

Efeito similar à ANOVA



# Autorregressiva de ordem 1 (AR1)

$$\begin{array}{c} \text{T1} \\ \text{T2} \\ \text{T3} \end{array} \begin{array}{ccc} \text{T1} & \text{T2} & \text{T3} \\ \left[ \begin{array}{ccc} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{array} \right] \end{array}$$

Assume que as correlações das observações são dependentes dos valores anteriores apenas, por meio de um processo autorregressivo

A correlação entre T1 e T3 é menor que entre T1 e T2

Efeitos logarítmicos, exponenciais e não monotônicos

# Não estruturada

$$\begin{array}{c} \text{T1} \\ \text{T2} \\ \text{T3} \end{array} \begin{array}{ccc} \text{T1} & \text{T2} & \text{T3} \\ \left[ \begin{array}{ccc} 1 & \rho_{1-2} & \rho_{1-3} \\ \rho_{1-2} & 1 & \rho_{2-3} \\ \rho_{1-3} & \rho_{2-3} & 1 \end{array} \right] \end{array}$$

Não sei qual o padrão de mudança e não quero chutar

# Ajuste do modelo

- não tem um índice de ajuste aproximado ou absoluto, porque os modelos são feitos em procedimentos de estimação
- Entretanto, é possível ver testes de qui quadrado para as VI, análise do resíduo (**o resíduo deve aproximar uma normal!**)
- Comparação de modelos usando **QIC** (mesma lógica do AIC, quanto menor melhor)

# Link de ligação

- Como interpretar os resultados?
- Basicamente, como regressão logística, linear ou poisson?

# Referências

- Wilson, J. R., Vazquez-Arreola, E., & Chen, D. G. (2020). *Marginal models in analysis of correlated binary data with time dependent covariates*. Springer International Publishing.
- Guimarães, L. S. P., & Hirakata, V. N. Uso do Modelo de Equações de Estimações Generalizadas na análise de dados longitudinais. Rev HCPA [Internet]. 2012 [cited 2016 May 15]; 32 (4): 503-11.
- Hardin, J. W., & Hilbe, J. M. (2012). *Generalized Estimating Equations*. CRC Press.
- Ma, Y., Mazumdar, M., & Memtsoudis, S. G. (2012). Beyond repeated-measures analysis of variance: Advanced statistical methods for the analysis of longitudinal data in anesthesia research. *Regional Anesthesia and Pain Medicine*, 37(1), 99–105. <https://doi.org/10.1097/AAP.0b013e31823ebc74>
- de Melo, M. B., Daldegan-Bueno, D., Menezes Oliveira, M. G., & de Souza, A. L. (2022). Beyond ANOVA and MANOVA for repeated measures: Advantages of generalized estimated equations and generalized linear mixed models and its use in neuroscience research. *European Journal of Neuroscience*, 56(12), 6089-6098.