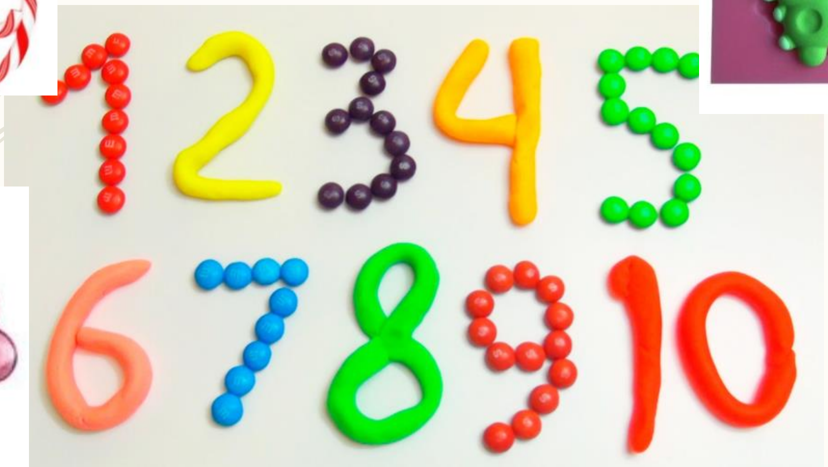
Goodies*



* Goodies related to animals, plants and numbers…

EMMA 2019 – 16 Novembro 2019, FCUL

"…Espera-se também, ao divulgar este acontecimento junto aos estudantes de Biologia do país, aproximá-los dos centros de investigação e dos investigadores que trabalham nesta área.

O objetivo é proporcionar uma oportunidade de networking ímpar e identificar estratégias que permitam criar uma verdadeira comunidade de investigadores com interesse na temática dos mamíferos marinhos, potenciando assim esforços conjuntos à escala nacional…."

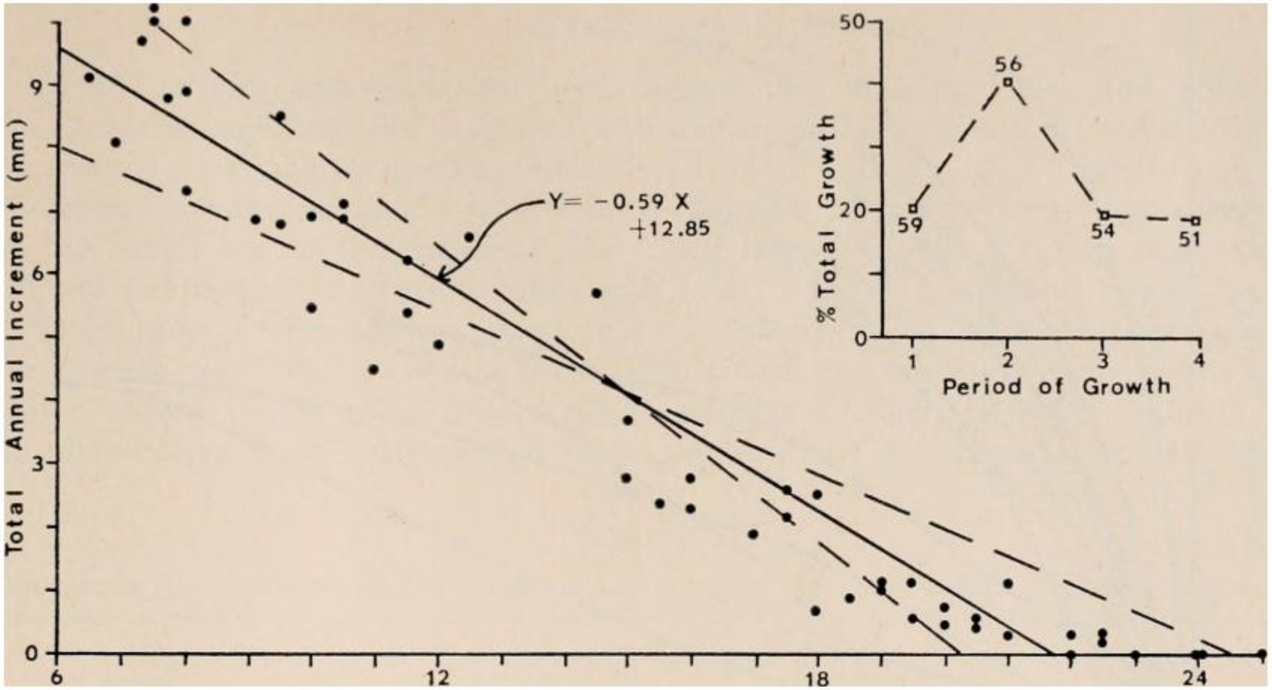
http://emma2019.campus.ciencias.ulisboa.pt/

Ecology for the Masses

Bob O'Hara: The Rise of the Ecological Modeller

Sam Perrin

3 days ago



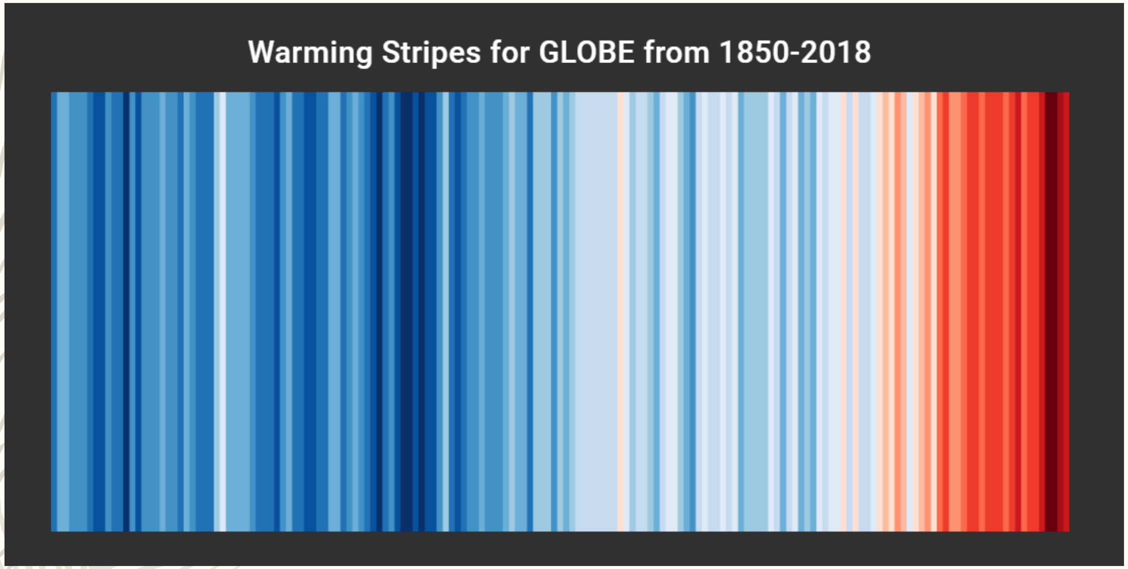
Non-scientists still often think of ecologists as field workers in cargo shorts, running around a grassland with a notebook and a tape measure. Whilst I'd be remiss to say this wasn't a percentage of us, the last two decades has seen the rise of ecological modelling, which has resulted in a new breed of ecologist. One who is capable of working almost exclusively with data, producing species distribution maps and population fluctuation graphs without leaving the office.

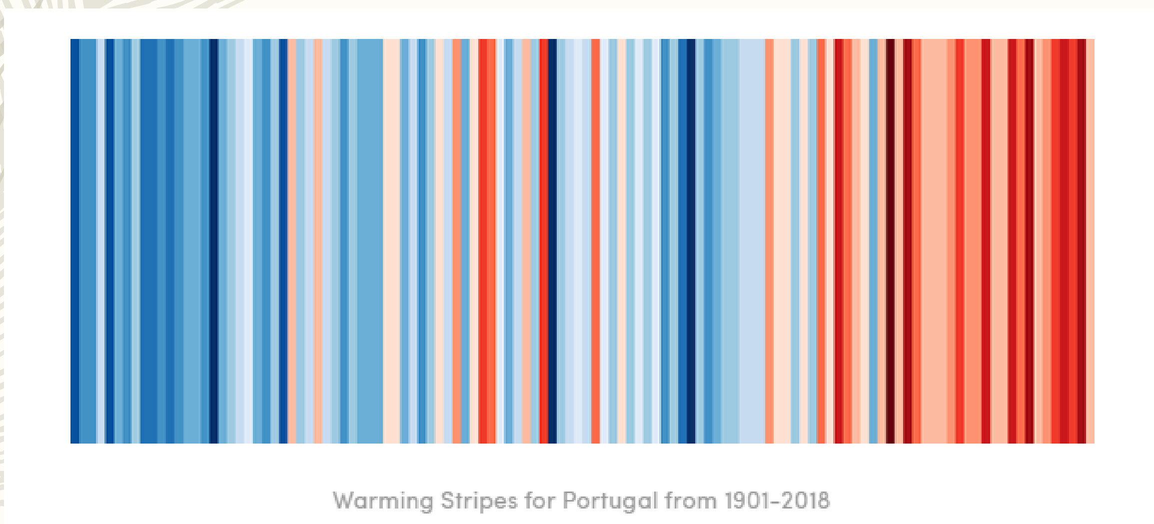
At the forefront of this group is Bob O'Hara, who has long claimed he plans to retire

https://ecologyforthemasses.com/2019/09/24/bob-ohara-the-rise-of-the-ecological-modeller/

https://showyourstripes.info/

#ShowYourStipes

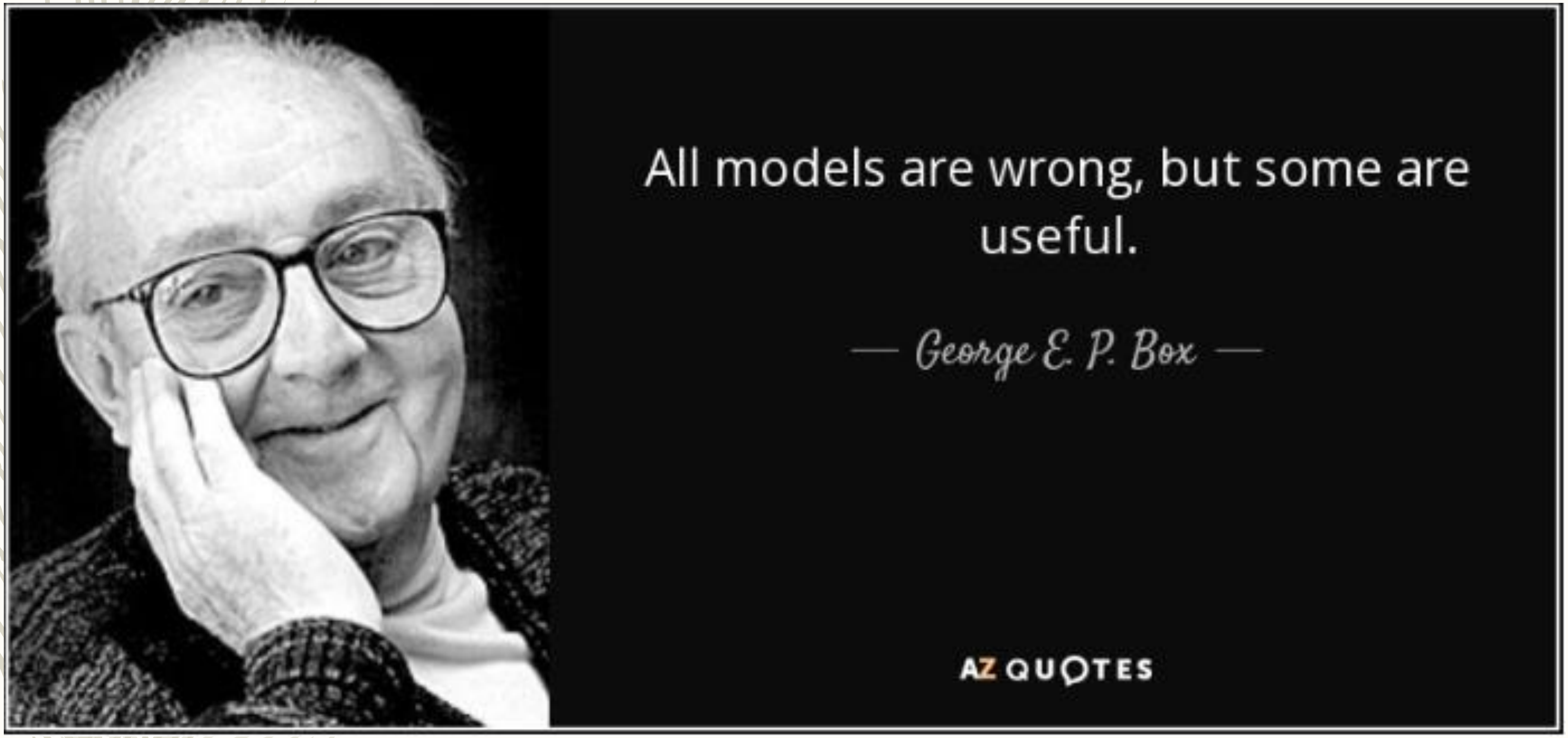




Annual average temperatures for GLOBE from 1850-2018 using data from UK Met Office.

https://dominicroye.github.io/en/2018/how-to-create-warming-stripes-in-r/

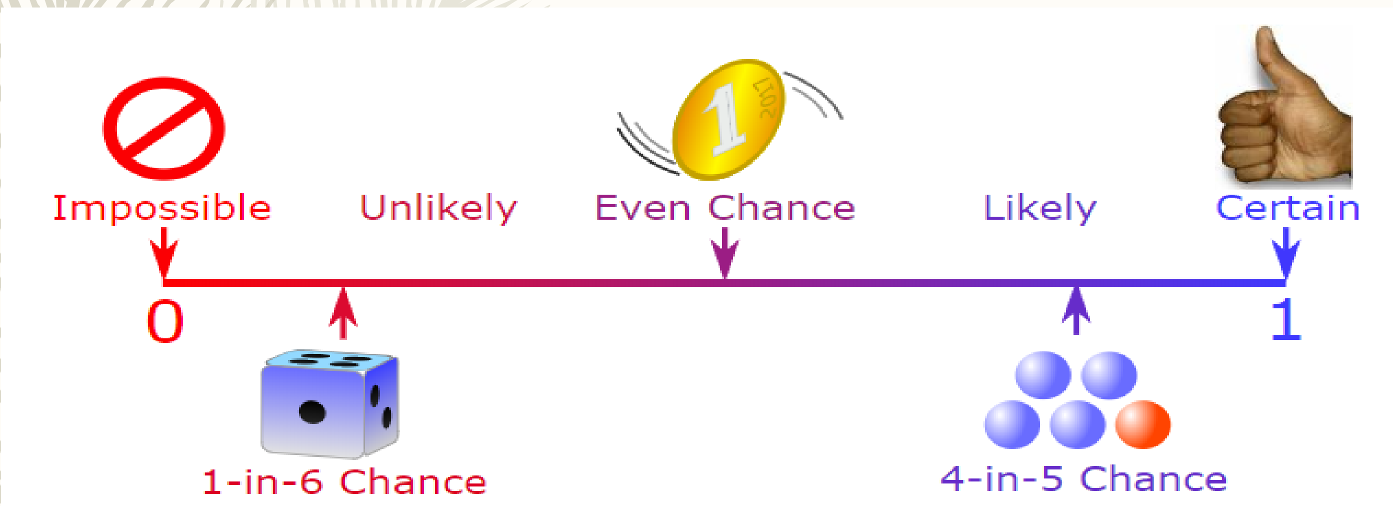
Ecologia Numérica - Aula Teórica 4 – 29-09-2018



https://www.azquotes.com/quote/534227

Probabilidades

- A probabilidade pode tomar valores entre 0 e 1



- O que significa uma probabilidade intermédia?

*A probabilidade de chover amanhã é 0.25?!**

** Mas, de notar, se eu avaliar depois de amanhã, ou choveu ou não!*

Notação e terminologia

- Designemos o evento por A. A probabilidade de um evento é geralmente escrita da seguinte forma

$$P(A) \text{ or } Pr(A)$$

- O complementar de determinado evento é $\bar{A}$ (tudo menos aquele evento).

$$P(\bar{A}) = 1 - P(A)$$

Probabilidades

- Uma probabilidade de 0.25 significa que será 3 vezes mais provável que não chova amanhã do que chova.

$$P(\text{não chover}) = 1 - P(\text{chover}) = 0.75$$

$$0.75/0.25 = 3$$

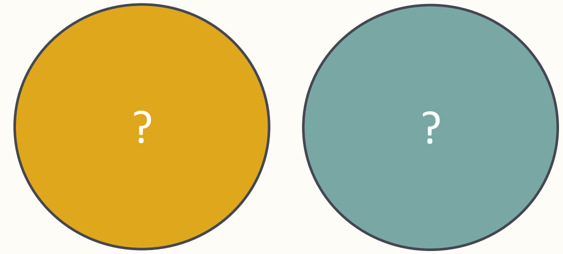
- Uma determinada probabilidade pode ser interpretada como uma proporção da concretização desse evento numa base temporal alargada.

A união de dois eventos consiste em tudo aquilo que estiver incluído em A ou B ou ambos.

Se

- *A = {chover amanhã}*
- *B = {chover amanhã e depois de amanhã}*
- *C = {3 peixes por arrasto}*
- *D = {4 ou 5 peixes por arrasto}*

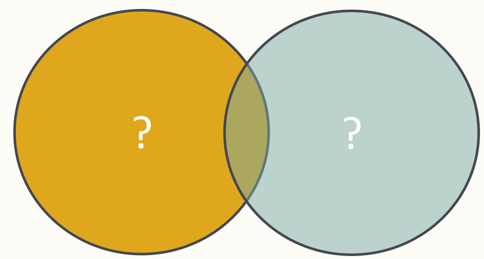
Então



- $A \cup B$ = *{chover nos próximos dois dias}*
- $C \cup D$ = *{3 a 5 peixes por arrasto}*

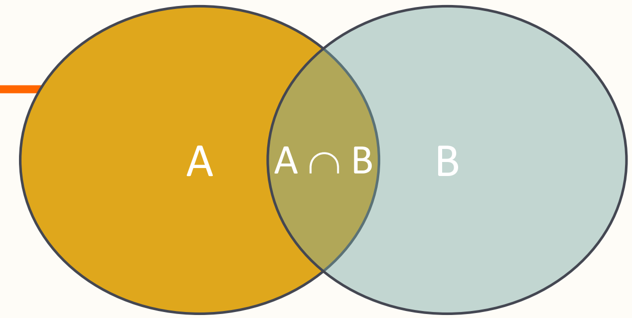
$P\{A \cup B\} \neq P\{A\} + P\{B\}$,

$P\{C \cup D\} = P\{C\} + P\{D\}$,



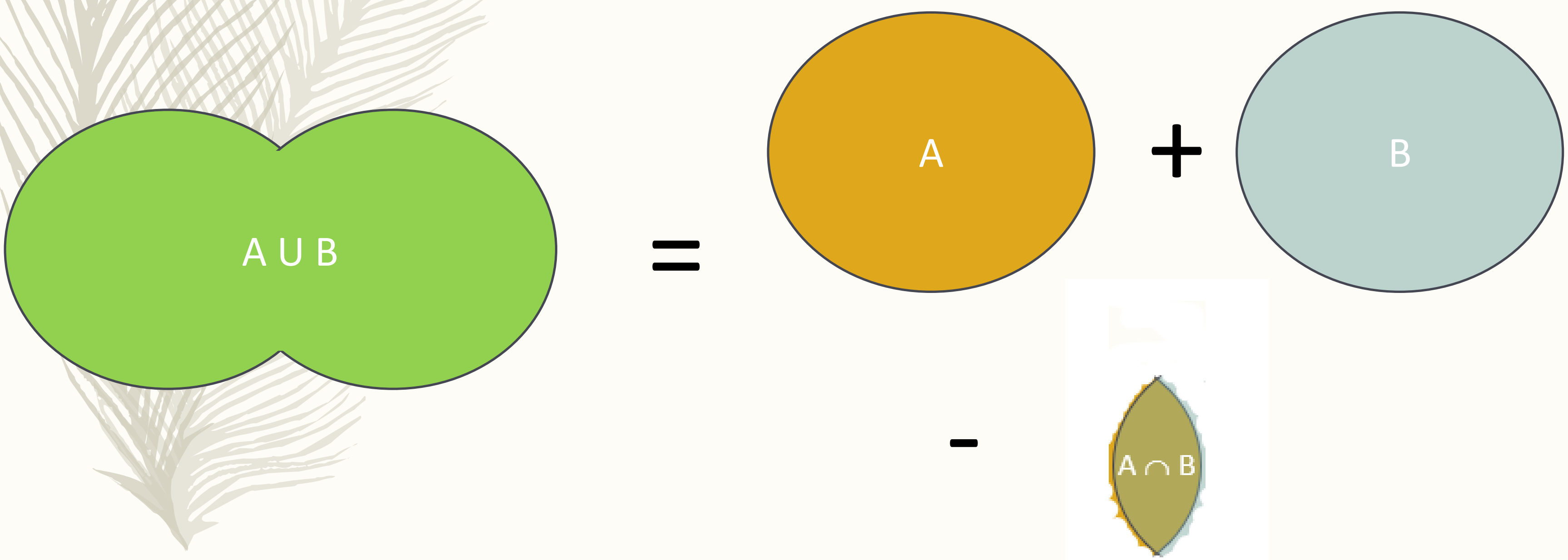
porque apenas C e D são mutuamente exclusivos, enquanto que A e B se intersectam!

tipos de variáveis
revisões sobre probabilidades



$$P\{A \cup B\} \neq P\{A\} + P\{B\}$$

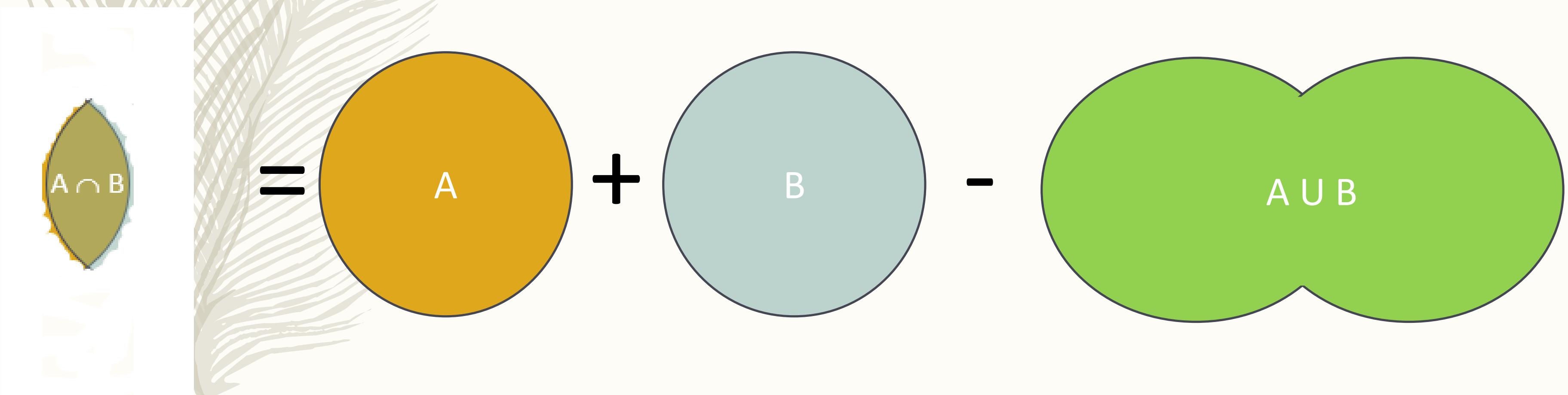
$$P\{A \cup B\} = P\{A\} + P\{B\} - P\{A \cap B\}$$



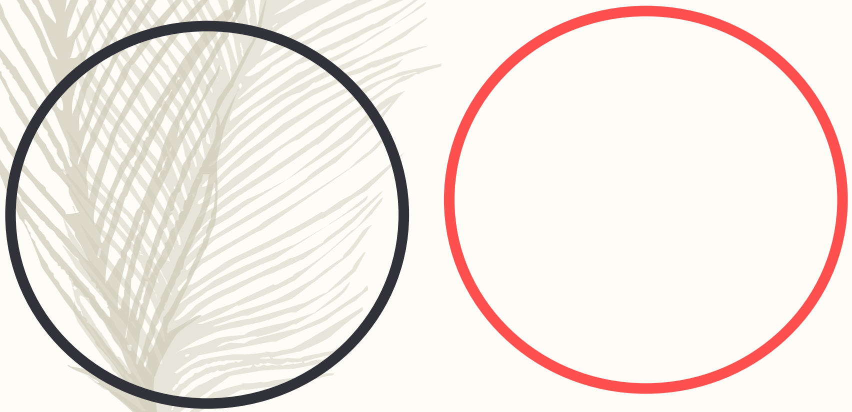
tipos de variáveis
revisões sobre probabilidades

$$P\{A \cup B\} = P\{A\} + P\{B\} - P\{A \cap B\}$$

$$P\{A \cap B\} = P\{A\} + P\{B\} - P\{A \cup B\}$$



$$P\{C \cup D\} = P\{C\} + P\{D\}$$



$$P\{C \cap D\} = \{\emptyset\}$$

Probabilidade condicional e independência

- Se a probabilidade de um evento for alterada pela ocorrência de outro evento, então os eventos não são independents.

Seja

A = {rain today}, B = {rain tomorrow}, C = {rain in 90 days time}

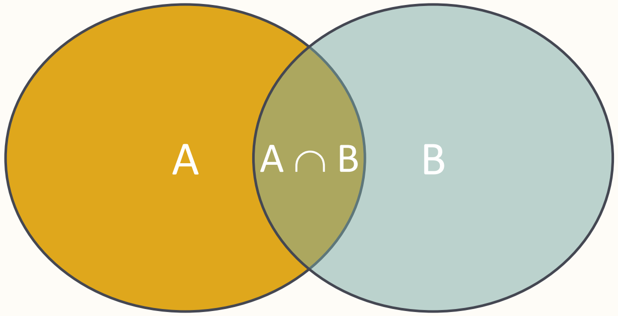
- É provável que o conhecimento de que A ocorreu possa alterar o valor de P para o evento B, mas não para o caso C.

Probabilidade condicional e independência

- Probabilidade de A, dado B, a.k.a. condicional a B
  - P(A|B):

$$P(A|B) \neq P(B) \text{ – excepto se A e B forem independentes!}$$

$$P(A|B) = P(A \cap B)/P(B)$$



Probabilidade condicional e independência

- Diz-se que dois acontecimentos são acontecimentos independentes se a probabilidade da sua intersecção for igual ao produto das suas probabilidades

$$P(A \cap C) = P(A) \times P(C)$$

Notícia (que ouvi na radio há alguns anos): 90% dos acidentes ocorrem a menos de 50 km de casa (i.e. perto, vs. longe)
Interpretação (que o lucutor fêz): numa viagem, os primeiros 50 km e os ultimos 50 km são os mais perigosos!!
Explicação: ainda estamos distraidos ou já estamos cansados, faz sentido!

P(perto) = 0.99
P(longe) = 0.01

P(acidente|perto de casa)=0.0005
P(acidente|longe de casa)=0.0005

P(acidente perto) = 0.99

P(perto|acidente) = 0.99
P(longe|acidente) = 0.01

P(acidente) = 0.0005

P(acidente e perto) ≈ P(perto)*P(acidente)

Ou seja… os acontecimentos são independentes… (…mas serão mesmo???)

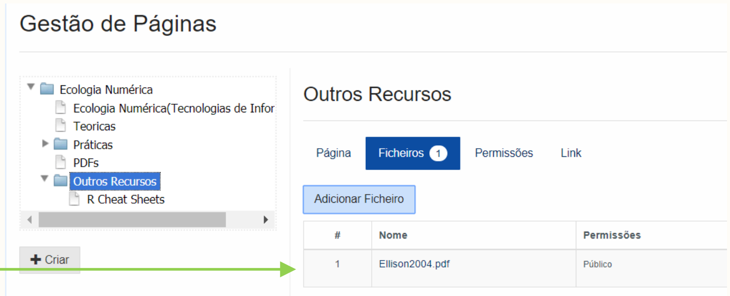
Probabilidade

– *Axiomática - frequencista*

baseada em Axiomas e teoremas de probalidade, frequência relativa em infinitas repetições de um acontecimento

– *Subjectiva - Bayesiana*

avaliação pessoal da possibilidade de ocorrência de um evento (teoria da decisão/ teoria da decisão Bayesiana*).



*Leitura opcional: Ellison, A. M. 2004 Bayesian inference in ecology *Ecology Letters* **7**: 509-520

Variáveis aleatórias

Uma **variável aleatória** é uma quantidade para a qual sabemos quais os valores possíveis, mas só sabemos o valor realizado depois de o observar, e esse valor é o resultado de um processo aleatório.

Os valores de determinadas variáveis estão sujeitos a processos aleatórios ou estocásticos que afectam essas variáveis – estas não são completamente previsíveis e, portanto, não são determinísticas. São as denominadas variáveis aleatórias.

São exemplo a temperatura da água, salinidade, turbidez, caudal de um rio, o número de animais em quadriculas de $1m^2$, (na realidade toda e qualquer medição que possam fazer!) etc.

Variáveis aleatórias

Qualquer medição com erro é ela propria uma variável aleatória, porque o erro é também ele aleatório.

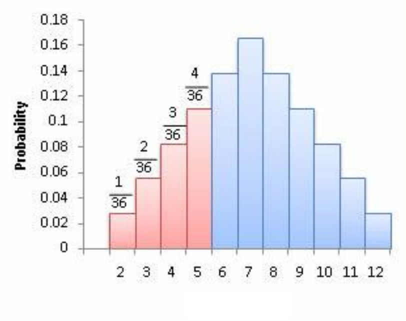
Frequentemente são medidos valores de variáveis que variam ao longo do tempo e/ou espaço. A esse conjunto de variáveis aleatórias chamamos um **processo estocático**.

Distribuições de probabilidade

Se medirmos uma variável aleatória muitas vezes, podemos construir uma distribuição de valores possíveis para essa variável.

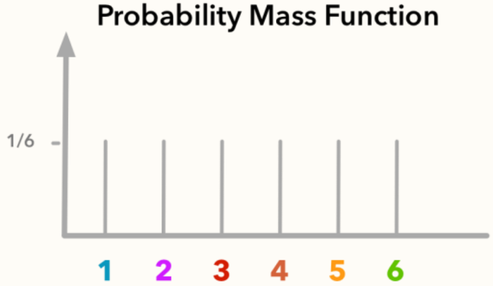
Se fosse possível registar múltiplos valores de uma variável nas mesmas condições obteríamos a distribuição de probabilidade dessa variável.

Uma variável aleatória é inequivocamente caracterizada pela sua distribuição:



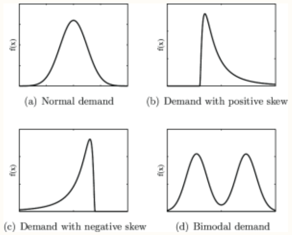
v.a. discreta:

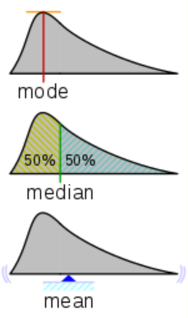
função **massa** de probabilidade



v.a. continua:

função distribuição de **densidade** de probabilidade (em inglês, *probability density function*, ou pdf)





Distribuições de probabilidade discretas

Uma função massa de probabilidade associa um valor de probabilidade a cada valor (discreto) que a variável pode tomar.

Exemplo 1. Variável aleatória que pode apenas tomar um de dois valore possiveis (Bernoulli, um caso específico da Binomial). Exemplo: Chover / Não chover.

P(Chover) = 0.2,

P(Não chover) = 0.8

Estas probabilidades constituem a função massa de probabilidade desta variável

Distribuições de probabilidade contínuas

As variáveis aleatórias contínuas podem tomar todos os valores possíveis (infinitos!) dentro de determinada amplitude de valores. Assim, não é possível associar valores de probabilidade a valores pontuais da variável.

Se X = temperatura da água de um rio

$P(X=x) = 0$ ou $P(X=2) = 0$ ou $P(X=3.193454) = 0$!

Distribuições de probabilidade contínuas

- Para as variáveis aleatórias contínuas temos uma função densidade de probabilidade, a qual permite calcular a probabilidade de observar a variável num determinado intervalo de valores.

- Esta probabilidade é calculada como sendo a área abaixo da função (a "curva" ou "linha") que define a função densidade, entre os valores de interesse.

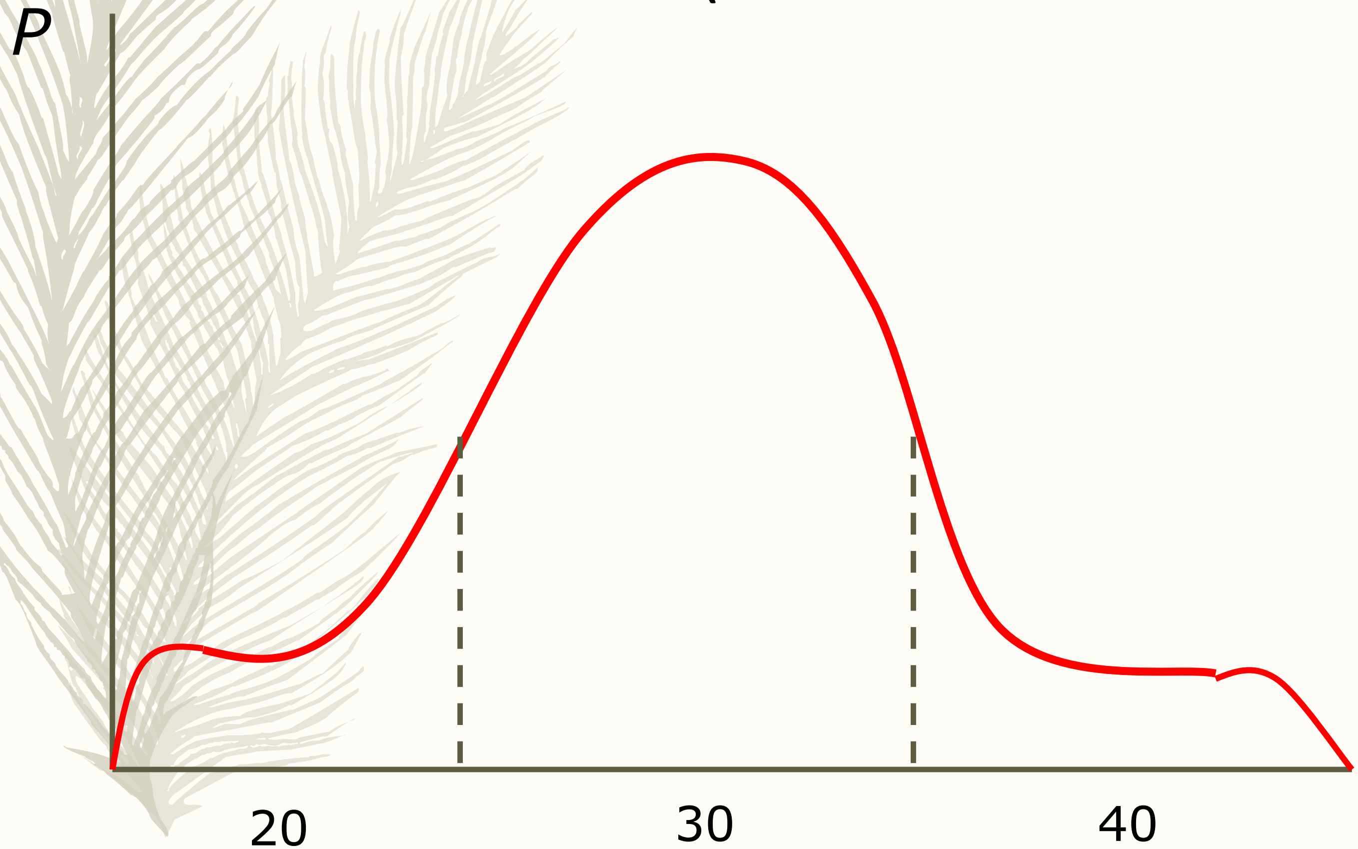
- A área total definida pela função é igual a 1.

Se X for uma variável aleatória continua, então existe uma função densidade probabilidade f(x) que a caracteriza tal que

- $f(x) \geq 0$ para todo o valor de x
- $\int_{-\infty}^{+\infty} f(x)dx = 1$
- $P(X=k) = 0$ para todo o X

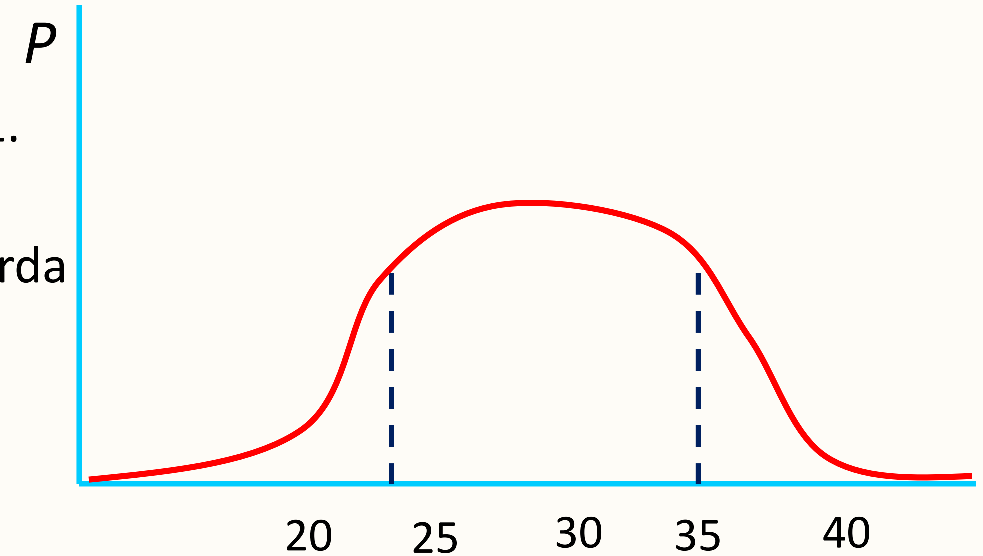
Exemplo:

Função densidade de probabilidade para a temperatura máxima do ar (… em Julho … em Lisboa)



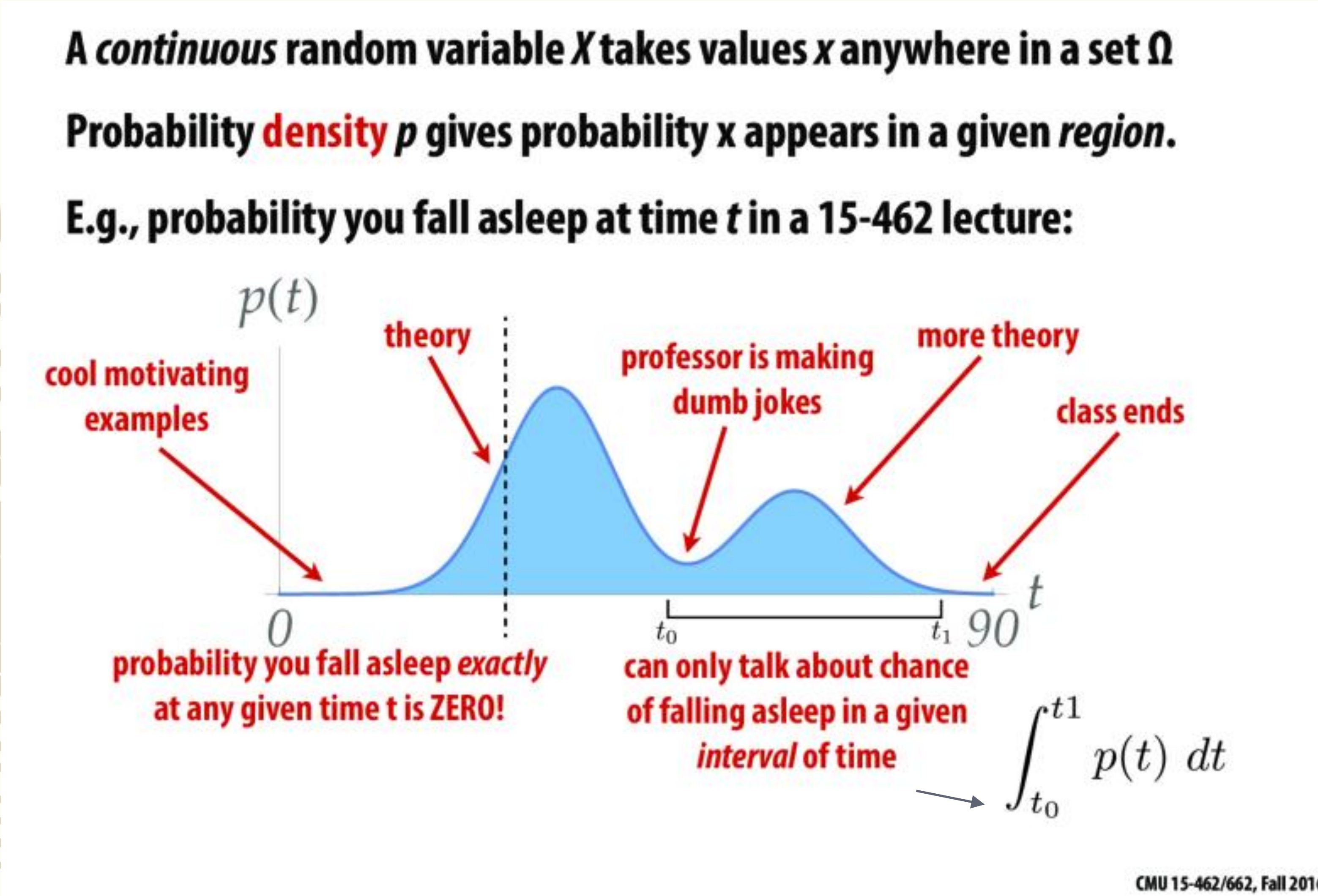
tipos de variáveis
revisões sobre probabilidades

Exemplo: Função densidade de probabilidade para a temperatura máxima (Tm) do ar



- A área total definida pela curva é 1.
- A área definida pela curva à esquerda de 20ºC é P(Tm) < 20ºC.
- A área entre 24ºC e 35ºC é P(24ºC < Tm <35ºC).
- A área à direita de 35ºC é a probabilidade da Tm exceder 35ºC.

Blatantly stolen from http://15462.courses.cs.cmu.edu/fall2016/lecture/variancereduction/slide_005



Note-se que, embora útil, este exemplo é estranho. Quase certamente, que fez isto não se apercebeu que o que definiu foi a variável aleatória T que representa o momento em que um aluno adormece numa aula, mas que para isto ser uma variável aleatória bem definida, ou seja o integral de p(t) entre 0 e 90 min = 1, o aluno tem de adormecer sempre durante a aula.

Podíamos usar a probabilidade condicional para garantir que a variável aleatória fica bem definida: representa a probabilidade, dado que um aluno adormece na sala de aula, adormecer no minuto *t*.

Famílias de distribuições de probabilidade

- O número de diferentes distribuições de probabilidade é infinito.

- As variáveis aleatórias com que trabalhamos são, quase invariavelmente, únicas!

- No entanto, é frequente agrupar conjuntos de distribuições semelhantes em famílias de distribuições.

Famílias de distribuições de probabilidade

Exemplos de famílias de distribuições discretas:

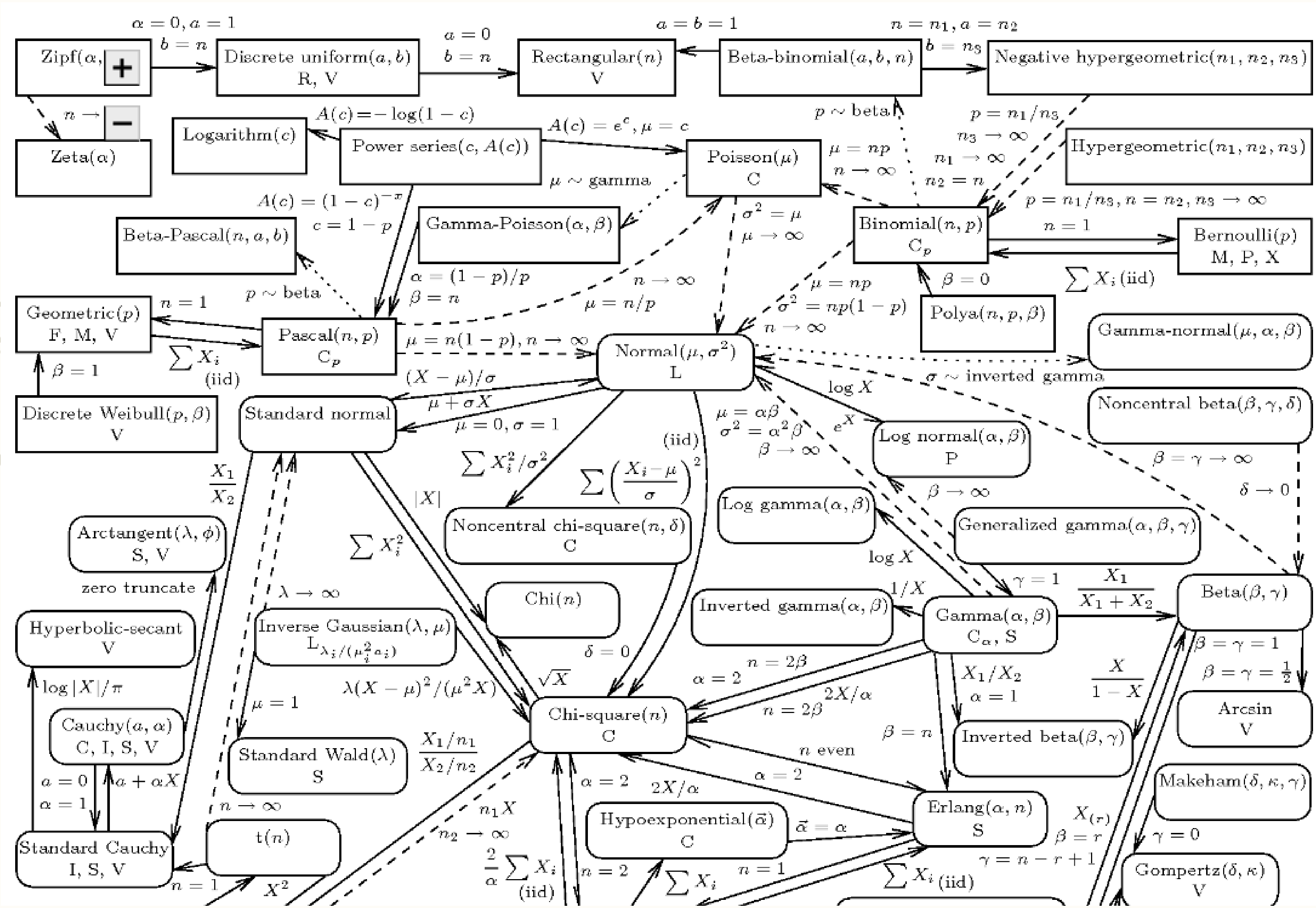
- Binomial
- Bernoulli
- Multinomial
- Binomial negativa
- Poisson
- Hipegeométrica

Famílias de distribuições de probabilidade

Exemplos de famílias de distribuições contínuas:

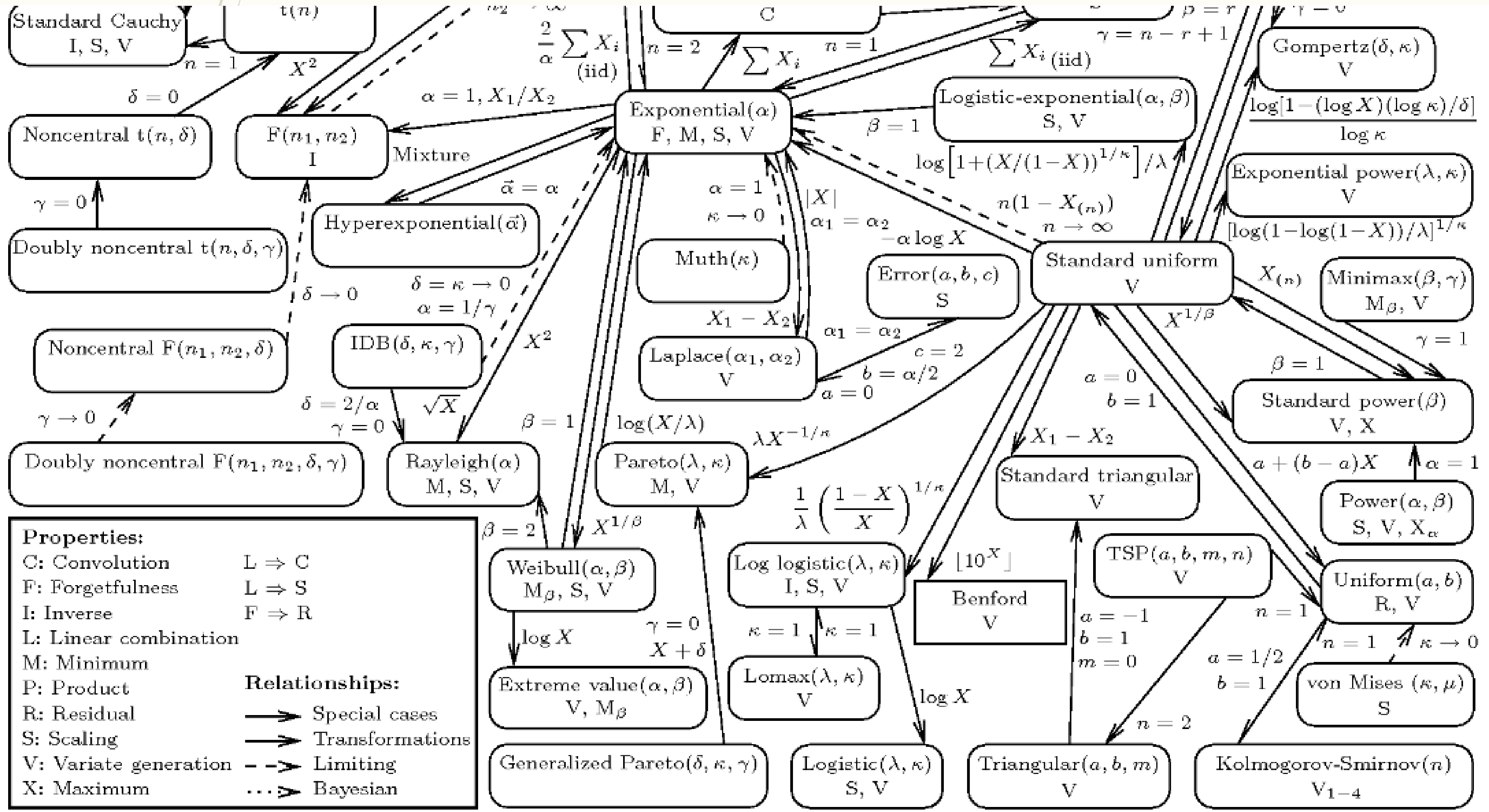
Normal (Gaussiana)
Exponencial
Gama
Beta
Log-normal
Weibull

Podem parar de stressar, nunca vão conhecer todas as distribuições



Continua…

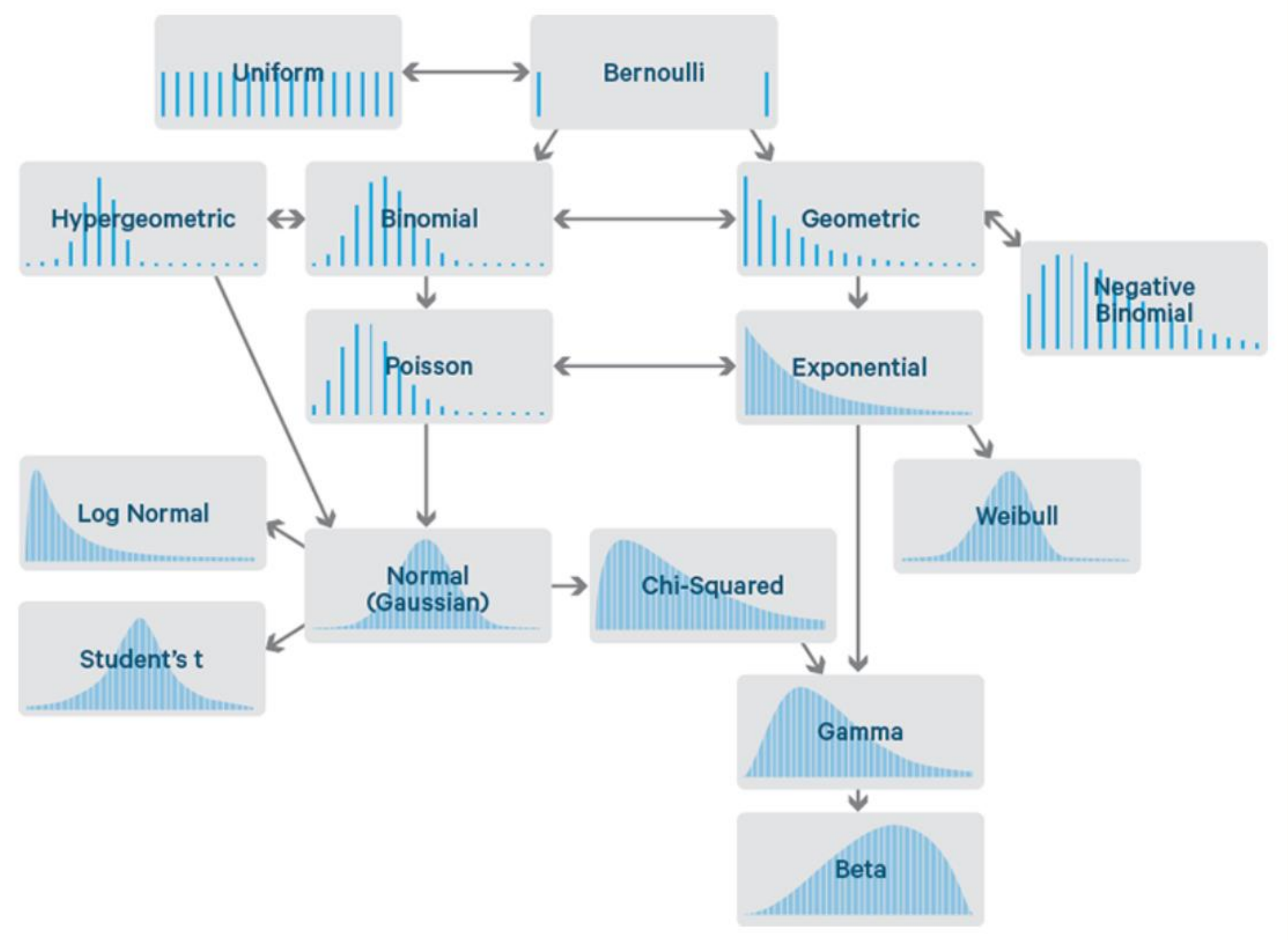
Continuação



http://www.math.wm.edu/~leemis/chart/UDR/UDR.html

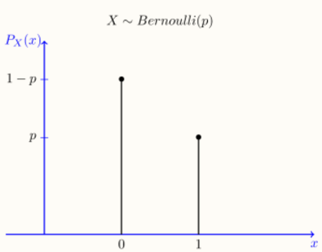
Mas aqui fica uma entrada num blog que ajuda a perceber as relações entre as mais comuns:

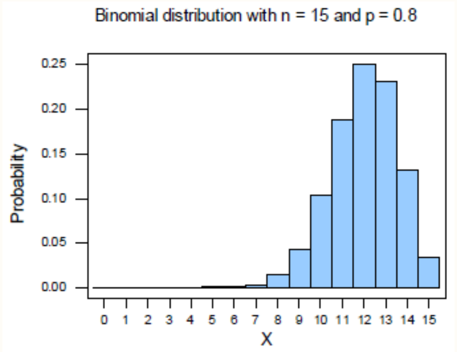
[Common Probability Distributions: The Data Scientist’s Crib Sheet]



Distribuição Binomial

- Associada a fenómenos em que a resposta é binária:
  - sucesso / insucesso
  - presente / ausente
  - sim / não
  - macho / fêmea
  - verdadeiro / falso
  - castanho / amarelo





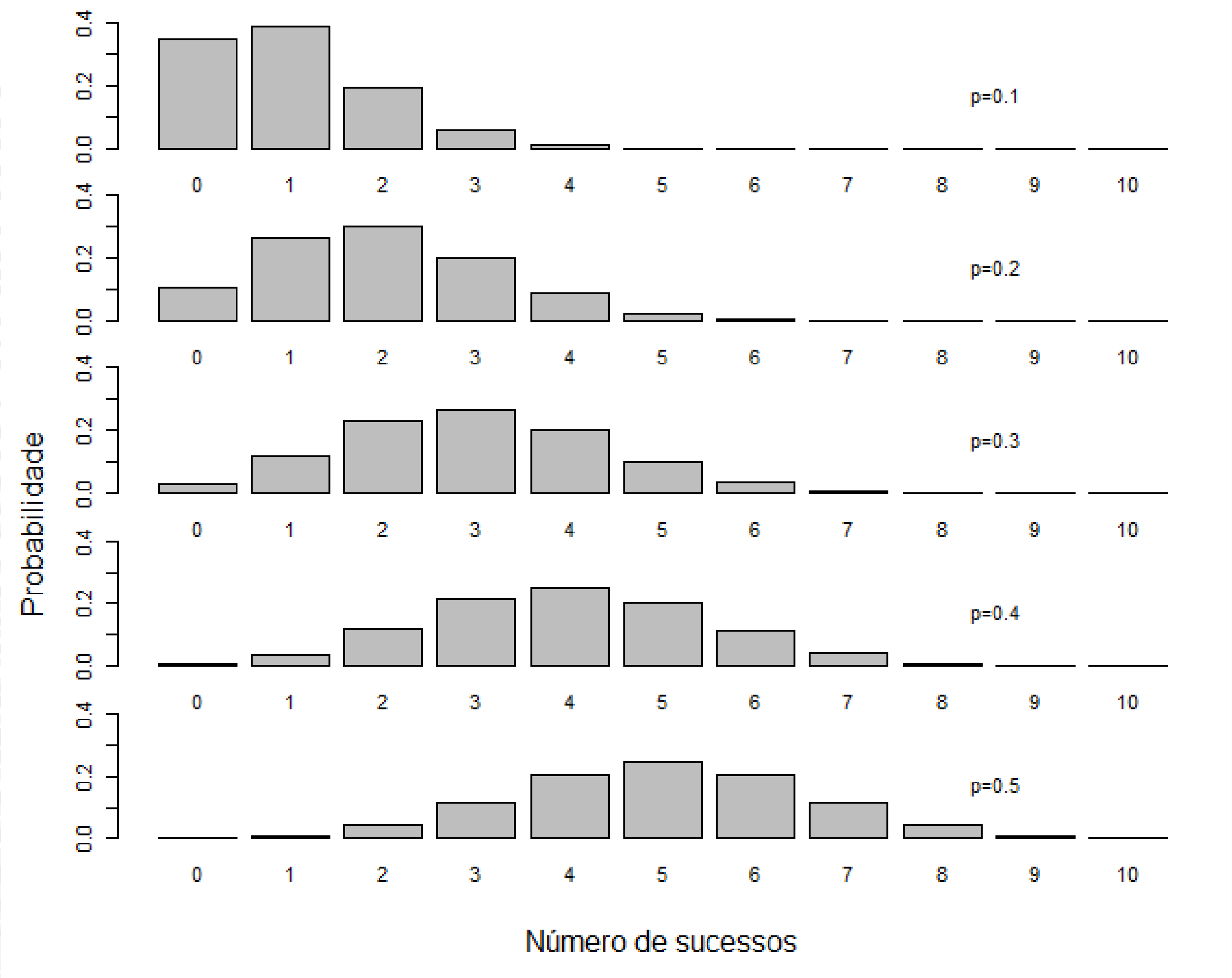
- A probabilidade de sucesso (p) é a mesma em cada realização.
- Uma variável aleatória binomial traduz o nº de sucessos (X) em *n* testes ou provas.
- Os pressupostos são a independência das provas e p constante.
- Se só tivermos uma prova dizemos que é uma distribuição de Bernoulli (2 valores possíveis, sucesso ou insucesso).
- Qualquer binomial pode ser decomposta numa soma de Bernoulli’s!

Distribuição Binomial

- É improvável que uma binomial seja uma boa descritora do nº de dias chuvosos, num período de 10 dias consecutivos, por causa da não-independênia destes eventos. (ou seja, teremos algo provavelmente mais próximo do 0 ou 10 do que valores intermédios)

- Provavelmente seria apropriada para descrever o nº de "Janeiros" sem neve num período de 20 anos, se, e só se, pudessemos aceitar a independência interanual destes eventos.

tipos de variáveis
revisões sobre probabilidades



Função massa de probabilidade de uma Binomial com 10 provas e probabilidade p=0.1,0.2,0.3,0.4,0.5

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

`dbinom(k,n,p)`

Desafio: tentem reproduzir a figura da página anterior usando a função acima!

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

`choose(n,k)`

No tutoriual de introdução ao R e Rstudio vimos a função `choose`. É isto que ela faz, combinações de `n`, `k` a `k`, ou por outras palavras, o coeficiente binomial!

Um exemplo com fauna Portuguesa do uso da distibuição binomial para modelar presença-ausência de lontras

Generalized linear mixed model (GLMM) com resposta binomial

OTTER RESPONSE TO LARGE DAMS CONSTRUCTION 69

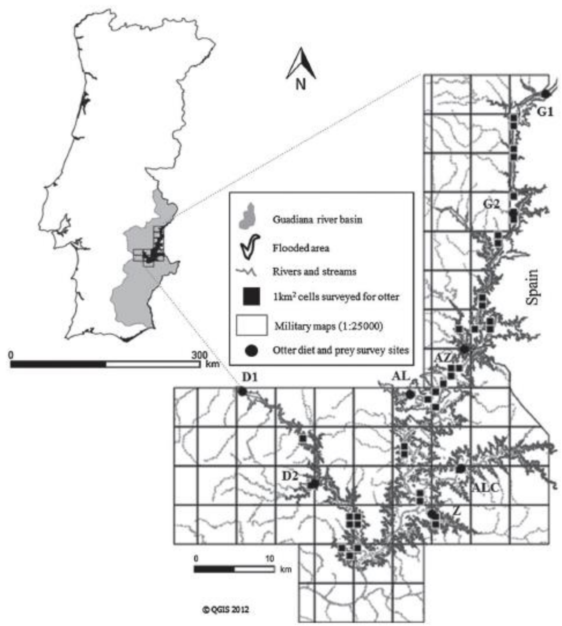


Figure 1. Location of the Alqueva Reservoir in southern Portugal, showing the 25 km^2 and the 1 km^2 (black squares) otter survey grid cells and the sites (bl circles) where otter diet and prey were assessed. Two sites (G1 and G2) were located in the main Guadiana River. Other sites were located in tributaries of Guadiana River: the Azevel stream (AZ), Álamo stream (AL), Degebe stream (D1 and D2), Alcarrache stream (ALC) and Zebro stream (Z).

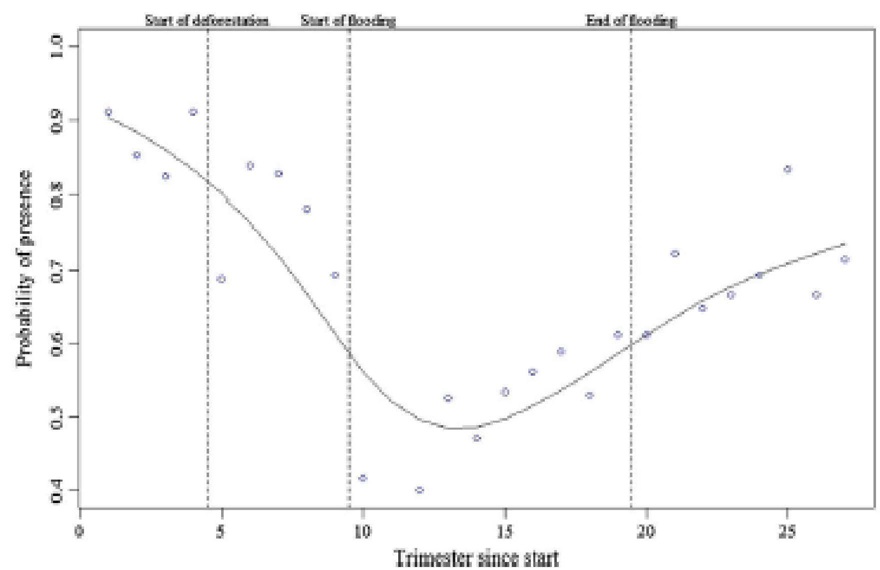


Figure 3. Probability of otter presence as a function of trimester in the flooded area of the Alqueva Reservoir. Data are represented by points and fitted model by a black line.

Pedroso, N. M.; Marques, T. A. & Santos-Reis, M. 2014 The response of otters to environmental changes imposed by the construction of large dams *Aquatic Conservation: Marine And Freshwater Ecosystems* **24**: 66–80

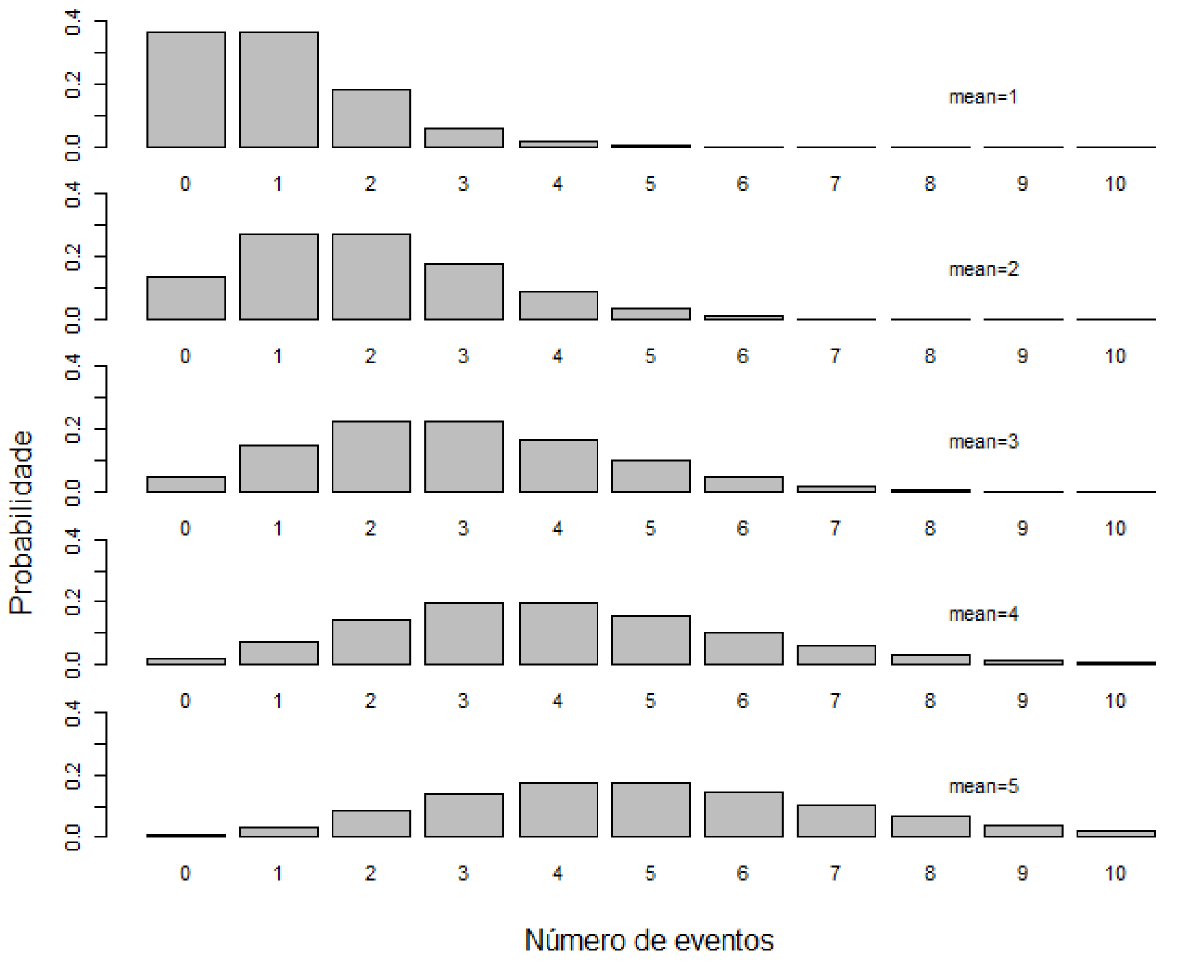
Distribuição Poisson

- A distribuição Poisson é frequentemente usada para descrever contagens
- Por exemplo:
  - Nº de tempestades num ano (contagem por unidade de tempo)
  - Nº de animais num km^2 (contagem por unidade de espaço)
- Os pressupostos são que os eventos ocorrem aleatoriamente num ritmo/taxa relativamente constante
- A Poisson é uma boa aproximação à Binomial quando n é grande e p pequeno.
- Tem apenas um parâmetro, a média (geralmente representada por λ)
- A media é igual à variância !

Distribuição Poisson

$$P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \mathrm{x}\geq 0, \lambda > 0$$

Distribuição Poisson



Função massa de probabilidade de uma Poisson com media λ

Distribuição Normal (Gaussiana)

A distribuição Normal (também conhecida como Gaussiana) é uma das distribuições contínuas mais utilizadas no âmbito da estatística inferencial.

Tem uma forma em sino, que é definida por 2 parâmetros:

- A média (μ na população, $\bar{x}$ na amostra) – a distribuição é simétrica em torno da média
- O desvio-padrão (σ na população, s_x ou s na amostra) – determina a dispersão da distribuição.

Distribuição Normal (Gaussiana)

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

-Inf < x < +Inf, -Inf < μ < +Inf, $\sigma > 0$

Distribuição Normal (Gaussiana)

Cerca de 2/3 da ditribuição está compreendida entre 1 σ em torno da média, e ~95% entre 2 σ para cada lado em relação à média.

Distribuição Normal

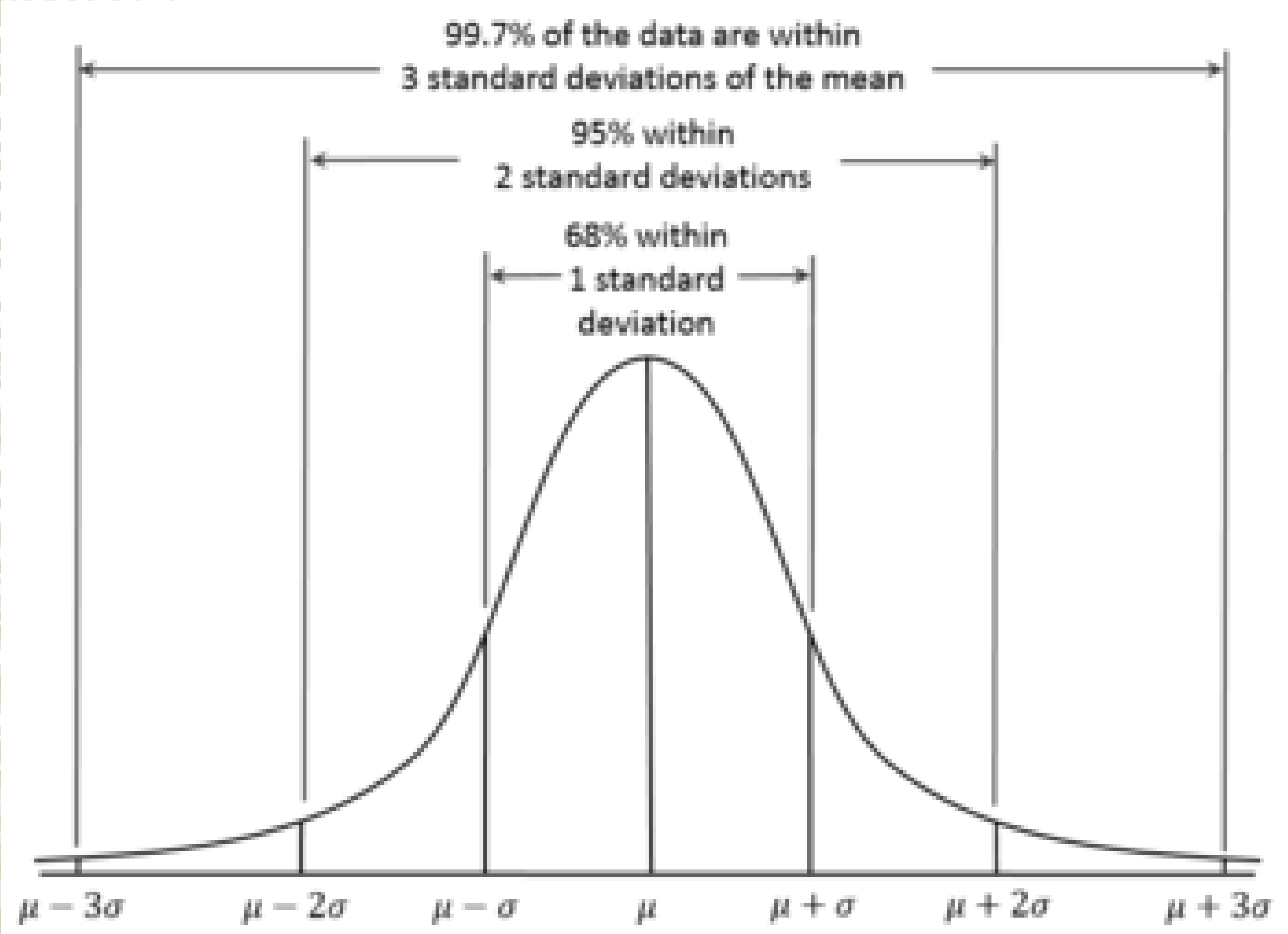
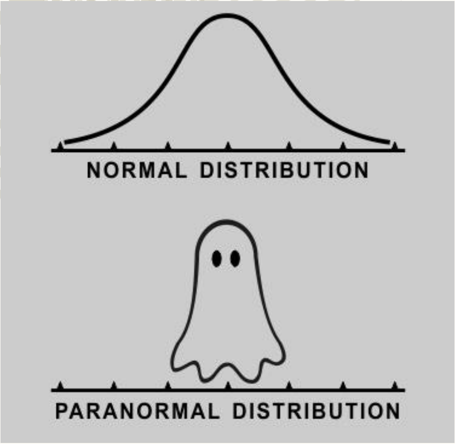
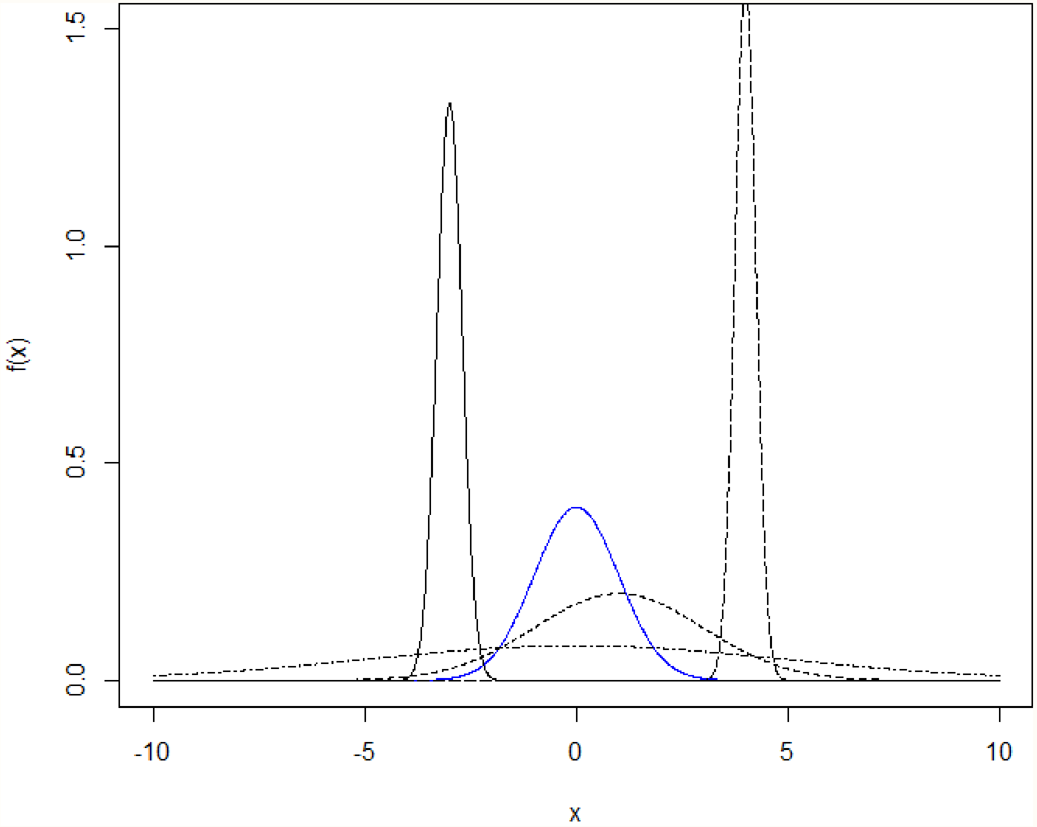


Image stolen from: [Wikipedia](). Desafio: fazer uma versão melhorada e não desfocada desta imagem em R. Eu colocarei a imagem nos slides e assim quem o fizer pode adicionar uma linha ao CV e dizer que ajudou na preparação do material de Ecologia Numérica.

Distribuição Normal (Gaussiana)

Porque é que a distribuição Normal é tão importante na estatística?

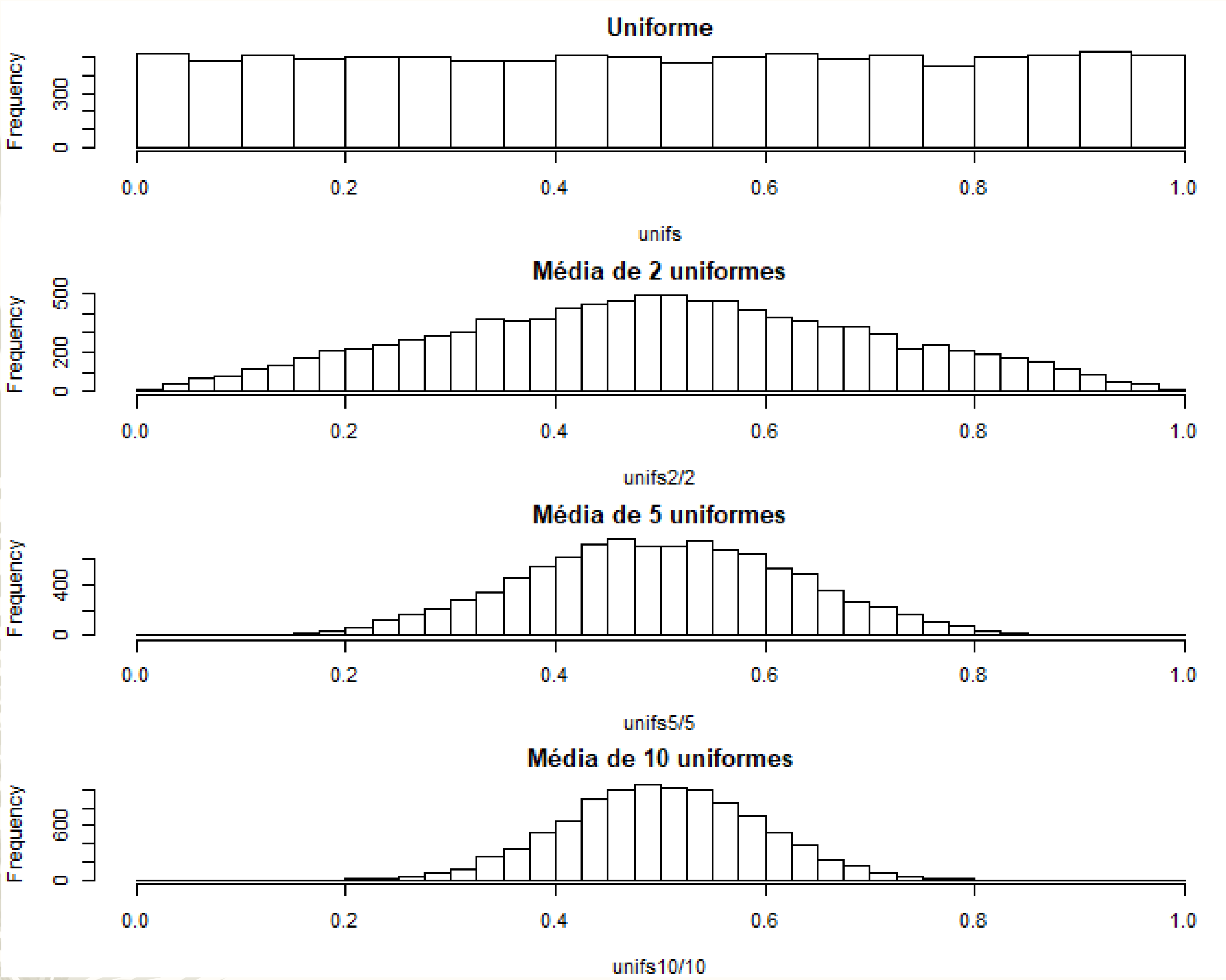




Teorema do Limite Central (TLC)

Se forem recolhidas amostras aleatórias de dimensão *n* de uma população normal, a distribuição das médias destas amostras será uma distribuição Normal.

As distribuições das médias de populações não-normais tenderão para a normalidade à medida que *n* aumenta.



Exemplo de uma média de uniformes a convergir para Gaussiana

Teorema do limite central (TLC)

A variância da população das médias decrescerá à medida que *n* aumenta:

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$$

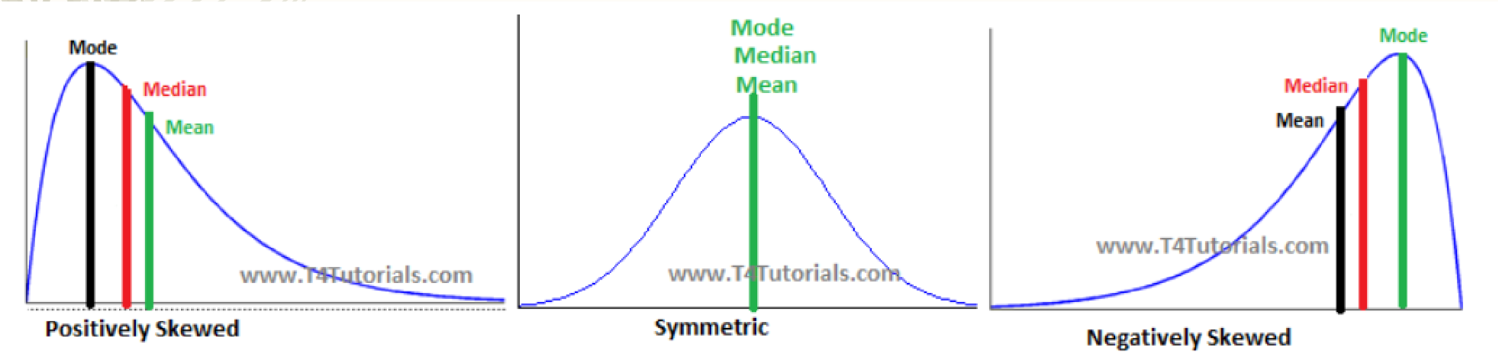
Variância da média

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Desvio padrão da média
(aka erro padrão ou standard error)

Outras distribuições contínuas

- Muitas variáveis apresentam desvios à normalidade pela falta de simetria.
- Um tipo comum de desvio é o chamado *skewness* (assimetria), que é verificado quando uma das caudas da distribuição é muito mais deprimida e alongada que a outra.
- A assimetria (skewness) positiva é a mais comum (cauda alongada na parte direita da distribuição).



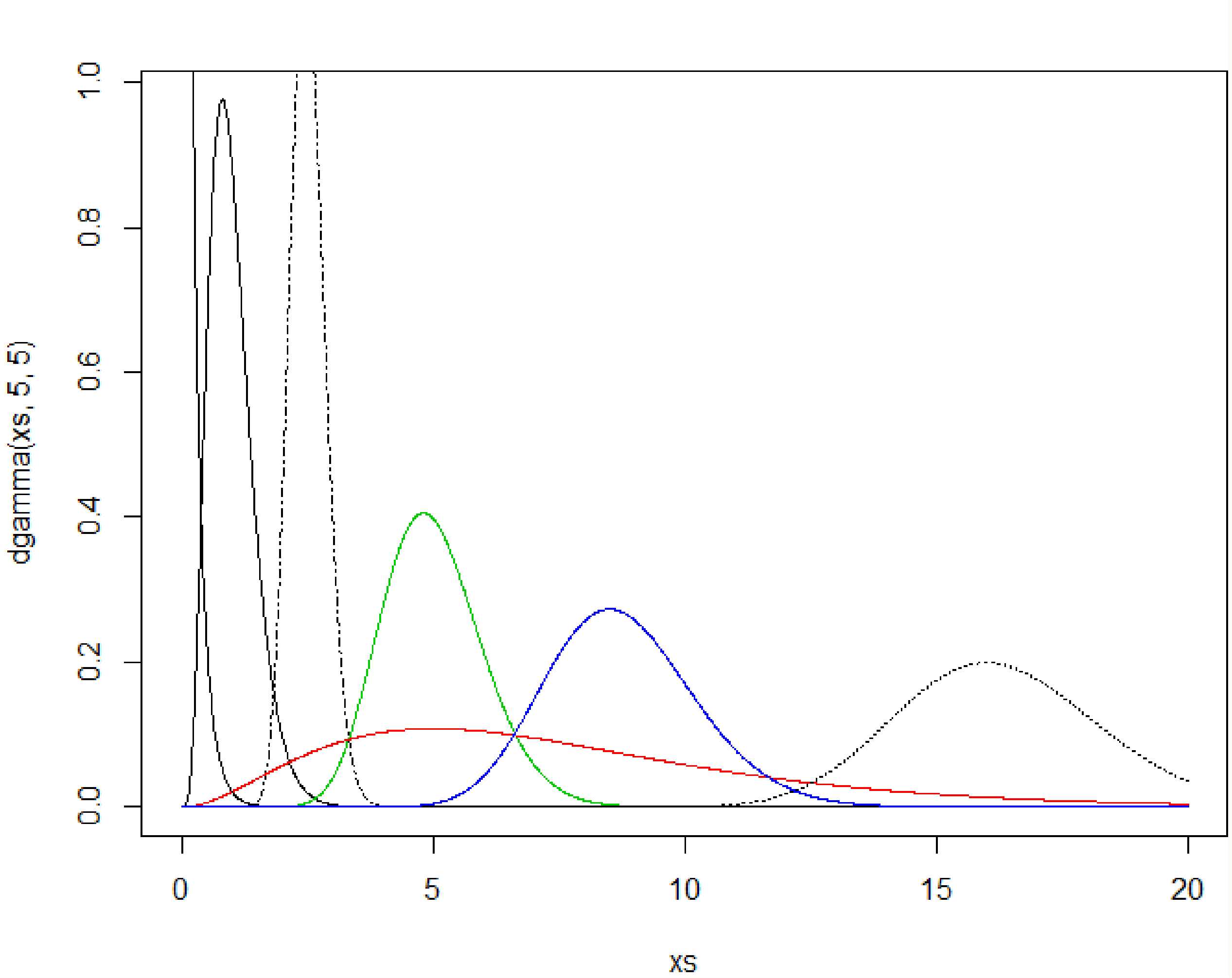
Famílias de distribuições assimétricas

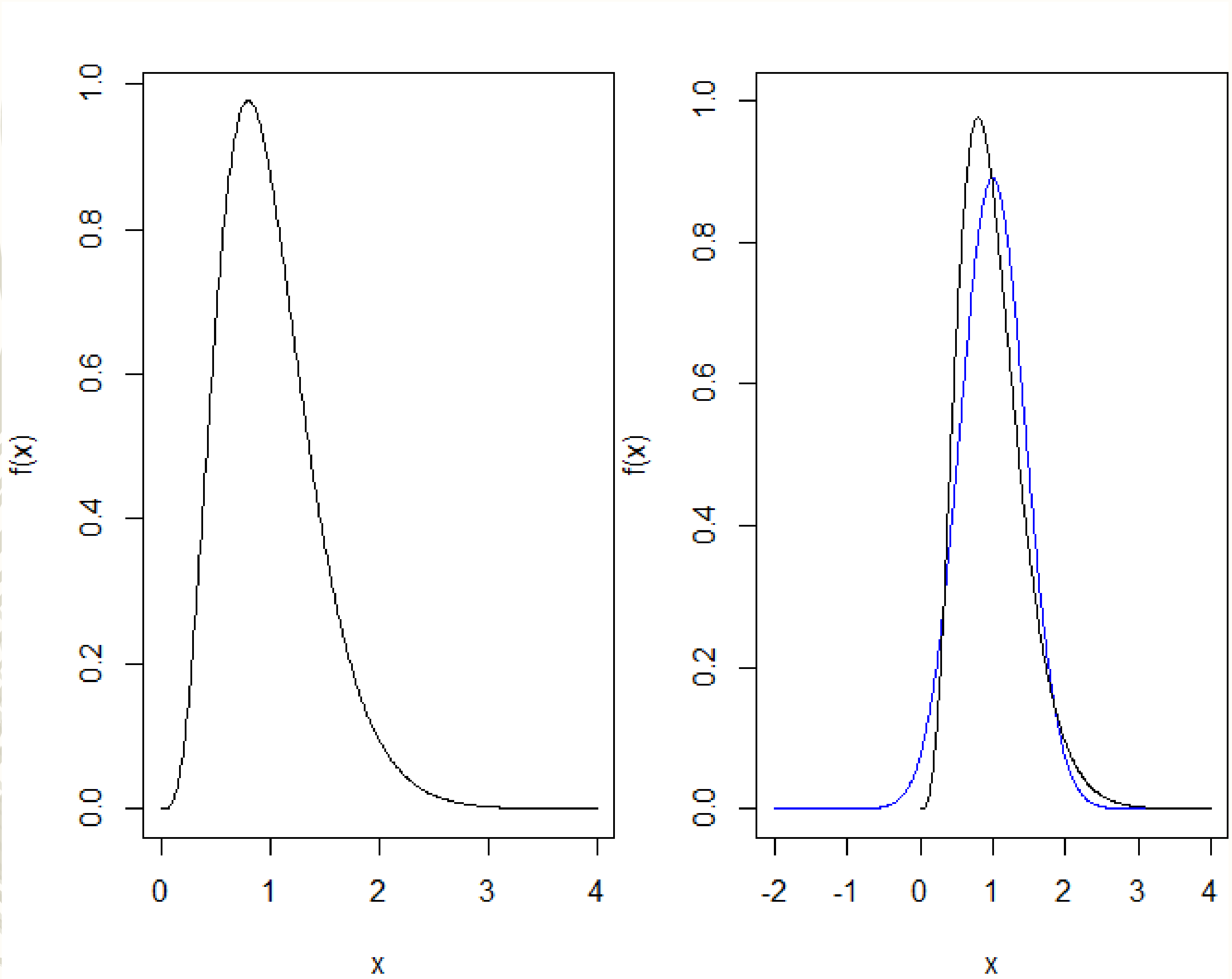
- Existe um grande número de distribuições assimétricas, muitas das quais se enquadram na família exponencial (não confundir com a variável exponencial!)

  Weibull
  Gamma
  Log-normal

- Estas distribuições são definidas por dois ou mais parâmetros que lhes podem conferir formas muito diferenciadas.
- São todas com suporte estritamente positivo (i.e. $x>0$)

Gamma - Uma distribuição com assimetria positiva





Uma Gamma(5,5) e a comparação com uma Gaussiana com a mesma média e desvio padrão

A distribuição beta

- Uma distribuição com suporte (i.e. os valores que a variável aleatória pode tomar) em (0,1)

- Uma distribuição extremamente flexível
  (e para que pode servir?)

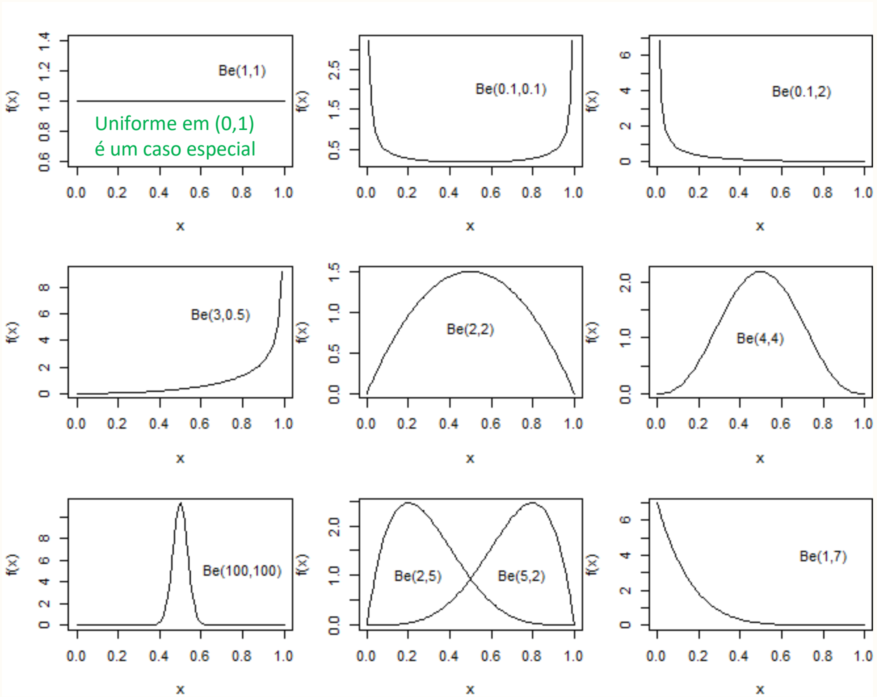
Received: 12 October 2018 | Accepted: 24 May 2019
DOI: 10.1111/2041-210X.13234

REVIEW

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression

Jacob C. Douma[1,2] | James T. Weedon[3,4]



Se X for Beta então 0 < x < 1!

Probabilidades e estatísticas

A determinação de probabilidades está associada a uma população, caracterizada através duma distribuição de probabiblidades, e consiste na previsão do que poderá acontecer quando retirada uma certa amostra.

As estatísticas são determinadas a partir de amostras da população e servem para descrever os dados ou para inferir e tecer considerações sobre a população donde eram provenientes as amostras.

Distribuições no R

Distributions {stats} R Documentation

Distributions in the stats package

Description

Density, cumulative distribution function, quantile function and random variate generation for many standard probability distributions are available in the **stats** package.

Details

The functions for the density/mass function, cumulative distribution function, quantile function and random variate generation are named in the form `dxxx`, `pxxx`, `qxxx` and `rxxx` respectively.

For the beta distribution see `dbeta`.

For the binomial (including Bernoulli) distribution see `dbinom`.

For the Cauchy distribution see `dcauchy`.

For the chi-squared distribution see `dchisq`.

For the exponential distribution see `dexp`.

For the F distribution see `df`.

For the gamma distribution see `dgamma`.

For the geometric distribution see `dgeom`. (This is also a special case of the negative binomial.)

For the hypergeometric distribution see `dhyper`.

For the log-normal distribution see `dlnorm`.

For the multinomial distribution see `dmultinom`.

For the negative binomial distribution see `dnbinom`.

For the normal distribution see `dnorm`.

No R: `?Distributions`

… a lista continua, e depois há muitas outras distribuiçãos menos usuais em packages adicionais

Distribuições no R

No R: `?Distributions`

As funções:

Cada distribuição é indexada por diferentes parâmetros

Densidade:
d*****: dbinom, dpois, dnorm, dgamma, dbeta, etc…
Argumentos: quantil + parâmetros que definem a distribuição
Resultado: Qual o valor da função densidade de probabilidade no quantil q

```
dnorm(-2,mean=0,sd=1)
[1] 0.05399097
```

Função de distribuição

p*****: pbinom, ppois, pnorm, pgamma, pbeta, etc…
Argumentos: quantil + parâmetros que definem a distribuição
Resultado: Qual o valor da função de distribuição cumulativa no quantil q

```
pnorm(0,mean=0,sd=1)
[1] 0.5
```

Distribuições no R

As funções:

Função quantil

q*****: qbinom, qpois, qnorm, qgamma, qbeta, etc…
Argumentos: probabilidade + parâmetros que definem a distribuição
Resultado: Qual o quantil associado a uma determinada probabilidade

```
qnorm(0.975,mean=0,sd=1)

[1] 1.959964

qnorm(0.5,mean=0,sd=1)
[1] 0
```

Função geradora de números pseudo-aleatórios

r*****: rbinom, rpois, rnorm, rgamma, rbeta, etc…
Argumentos: número de observações + parâmetros que definem a distribuição