Estatística Aplicada a Ciências Ambientais

Distribuições de Probabilidade

Daniel Detzel
detzel@ufpr.br

Agenda

Distribuições de probabilidade

- definições

Distribuições discretas

- processos de Bernoulli
- distribuição Geométrica
- distribuição Binomial

Hands-on work: espaço para trabalho

Distribuições de probabilidade

definições

Distribuições de probabilidade | definições

Considere o histograma obtido para a série de OD medida no posto IG3, rio Iguaçu, entre 17 jun 05 e 14 ago 2017

Qual é a probabilidade de a OD estar entre 2 e 3 mg/L?

Essa informação é válida para a população de OD no posto IG3?

Distribuições de probabilidade | definições

Para se chegar em estimativas representativas da população, é preciso assumir um modelo matemático teórico

O modelo é conhecido por distribuição de probabilidades e pode assumir diferentes formas

variáveis aleatórias discretas vs. contínuas

propriedades que comprovam que o modelo matemático é, de fato, uma distribuição de probabilidades

Para as próximas definições, considere a variável aleatória X

Distribuições de probabilidade | definições

Variáveis aleatórias discretas:

As probabilidades de ocorrência dos valores de X são representadas por:

$$p_X(x) \equiv \mathrm{prob}(X = x)$$

Lê-se: probabilidade de a variável aleatória X assumir o valor x

$p_X(x)$ representa a função massa de probabilidades (FMP)

analogia teórica ao histograma de frequências da amostra

Distribuições de probabilidade | definições

Variáveis aleatórias discretas: (cont.)

Propriedades de uma FMP:

$$\begin{cases} p_X(x) \geq 0, \text{ para qualquer valor de } x \\ \displaystyle\sum_{\text{todos } x} p_X(x) = 1 \end{cases}$$

Qualquer função que atenda a essas propriedades é uma FMP

Distribuições de probabilidade | definições

Variáveis aleatórias discretas: (cont.)

Da FMP é possível obter a função acumulada de probabilidades (FAP):

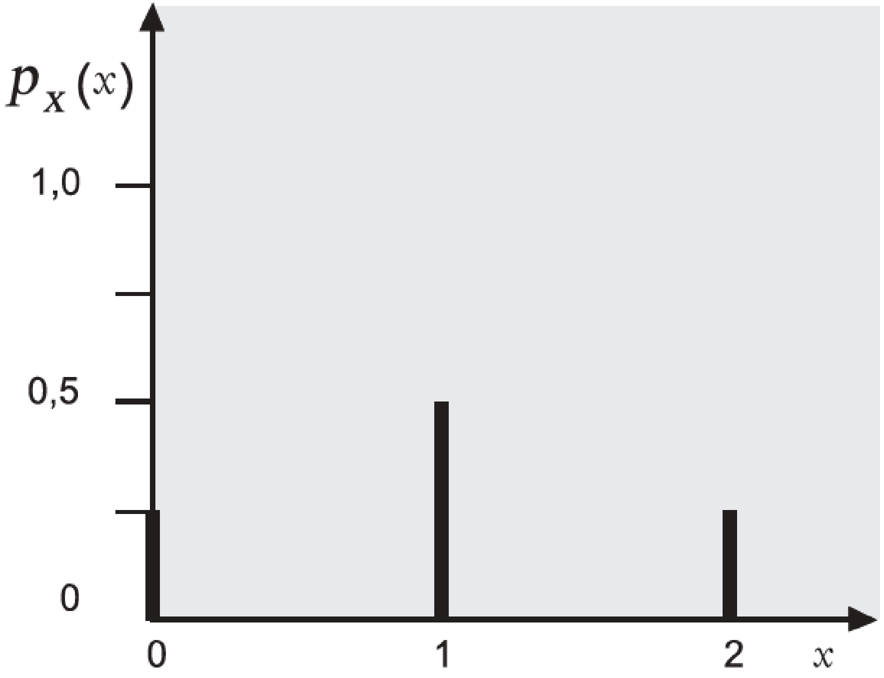
$$P_X(x) \equiv P(X \leq x) = \sum_{\text{todos } x_i \leq x} p_X(x_i) = 1$$

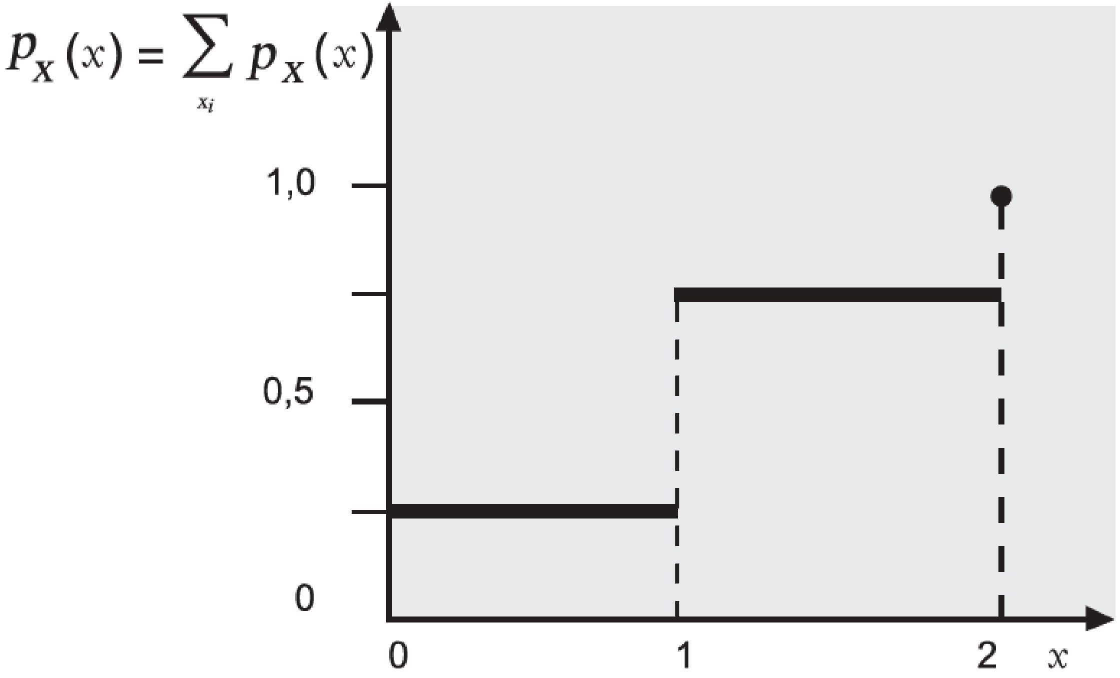
Denota a probabilidade de a variável aleatória X ser menor ou igual a x

Distribuições de probabilidade | definições

Variáveis aleatórias discretas: (cont.)

Exemplo gráfico:





Imagens: Naghettini e Pinto (2007, p. 67)

Distribuições de probabilidade | definições

Variáveis aleatórias contínuas:

Para o caso contínuo, o equivalente à FMP é a função densidade de probabilidades (FDP)

equivalente a um gráfico de densidades para uma amostra de tamanho infinito

$$f_X(x)$$

A densidade não denota probabilidade

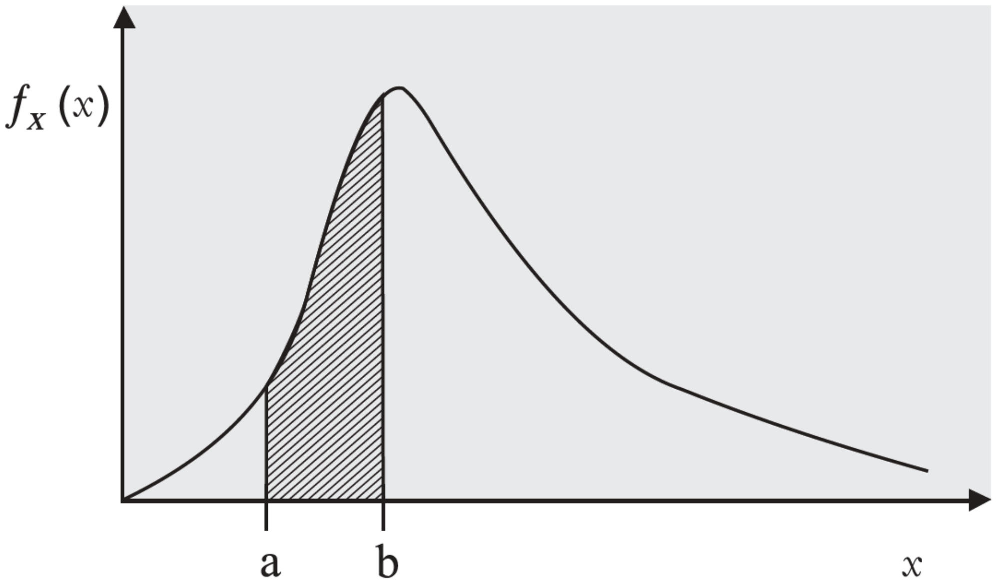
a probabilidade é calculada com base na área sob a curva da função

Variáveis aleatórias contínuas: (cont.)

Assim, para a probabilidade de a variável aleatória estar entre dois limites a e b, calcula-se:

$$\text{prob}(a < X < b) = \int_a^b f_X(x)\ dx$$

Imagem: Naghettini e Pinto (2007, p. 68)



Importante: a probabilidade de a variável aleatória assumir exatamente um valor qualquer é zero. Ou seja, $\text{prob}(X = x) = 0$

Distribuições de probabilidade | definições

Variáveis aleatórias contínuas: (cont.)

Propriedades de uma FDP:

$$\begin{cases} f_X(x) \geq 0, \text{ para qualquer valor de } x \\ \displaystyle\int_{-\infty}^{+\infty} p_X(x)\,dx = 1 \end{cases}$$

Qualquer função que atenda a essas propriedades é uma FDP

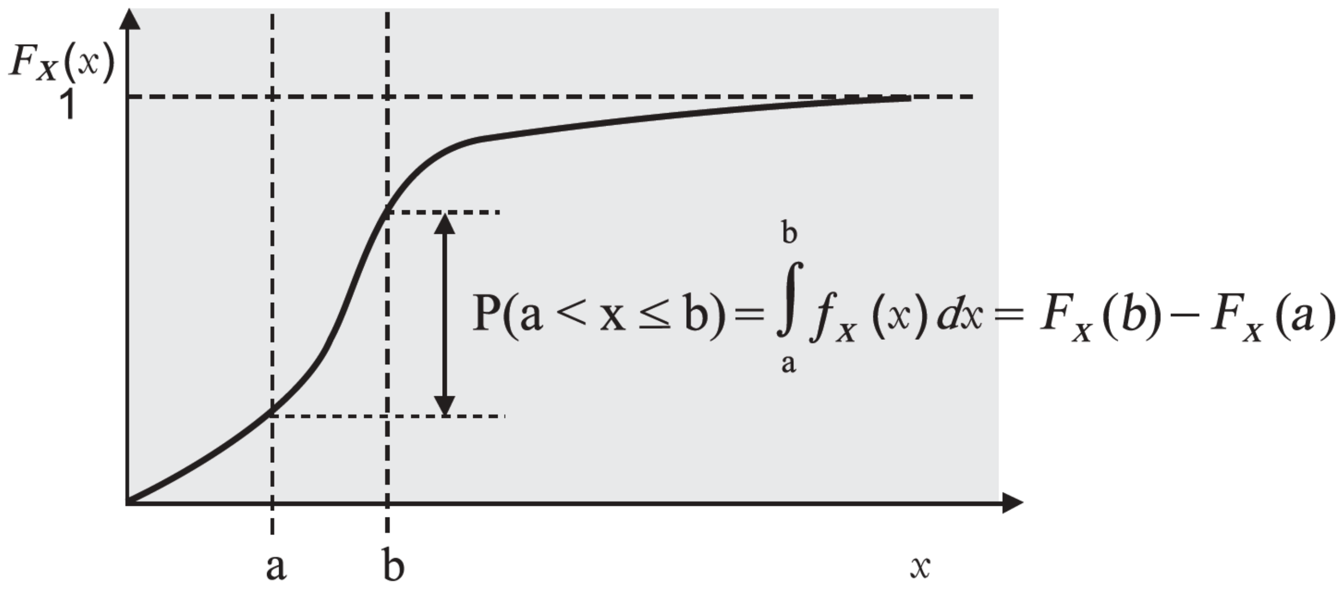
Distribuições de probabilidade | definições

Variáveis aleatórias contínuas: (cont.)

Para o caso contínuo, o equivalente à FAP é a função densidade de probabilidades acumulada (FDA)

$$F_X(x) = \int_{-\infty}^{x} f_X(x)\ dx$$

Imagem: Naghettini e Pinto (2007, p. 68)



Denota a probabilidade de não excedência de x, ou seja, $\mathrm{prob}(X \le x)$

Distribuições de probabilidade | definições

Em resumo:

Caso discreto

FMP: $p_X(x) = \text{prob}(X = x)$
Denota a probabilidade de a v.a. assumir x

FAP: $P_X(x) = \text{prob}(X \leq x)$
Denota a probabilidade de não excedência da v.a.

Caso contínuo

FDP: $f_X(x)$
Denota densidade da probabilidade associada a x

FDA: $F_X(x) = \text{prob}(X \leq x)$
Denota a probabilidade de não excedência da v.a.

Distribuições de probabilidade | definições

Formas de FDPs:

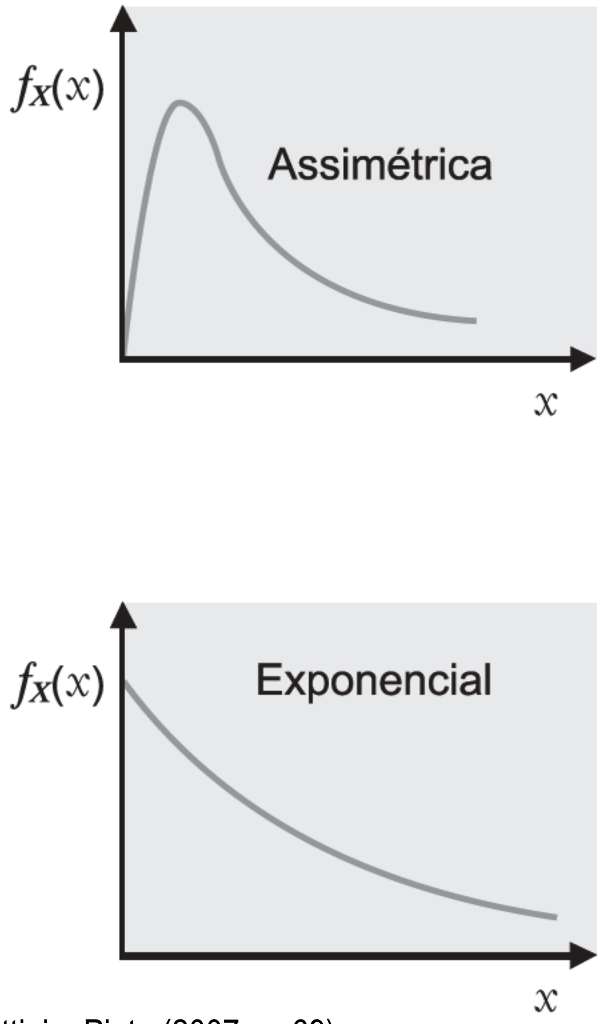
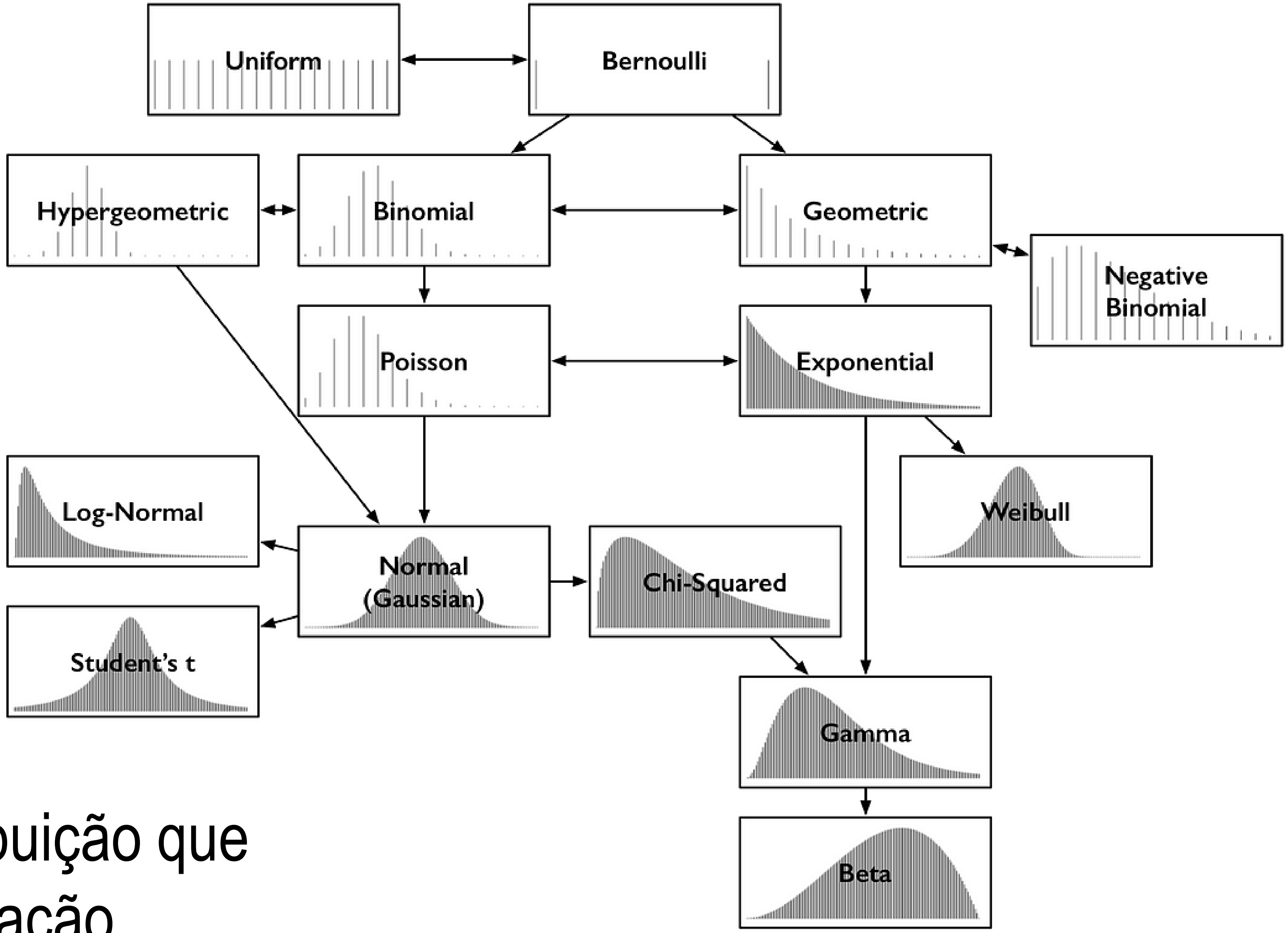


Imagem: Naghettini e Pinto (2007, p. 69)

Distribuições de probabilidade | definições



Desafio: encontrar a distribuição que melhor representa a população

Imagem https://medium.com/@srowen/common-probability-distributions-347e6b945ce4

DISTRIBUIÇÕES DE PROBABILIDADE

distribuições discretas

Distribuições de probabilidade | distribuições discretas

Distribuições discretas mais utilizadas:

- Binomial
- Geométrica
- Binomial negativa
- Poisson
- Hipergeométrica
- Multinomial
- etc.

Foco será dado nas distribuições Binomial e Geométrica

Distribuições de probabilidade | processos de Bernoulli

Processos de Bernoulli fundamentam as distribuições Binomial e Geométrica (além da binomial negativa)

Seja um experimento com dois resultados possíveis:

S: sucesso, com probabilidade de ocorrência p
F: falha, com probabilidade de ocorrência $1-p$

O espaço amostral é formado pelo conjunto {S,F}

Se a esse experimento for associada uma variável aleatória X com valores 1 (sucesso) e 0 (falha), ele segue um processo de Bernoulli

Distribuições de probabilidade | processos de Bernoulli

Dois tipos de variáveis aleatórias discretas Y podem ser associadas a ele:

Geométrica: quando Y se refere ao nº de repetições independentes que precisam acontecer para que um único sucesso ocorra

ex.: quantas vezes um dado precisa ser lançado para que apareça um número 6?

Binomial: quando Y se refere ao nº de sucessos em N repetições independentes

ex.: quantos números 6 aparecem em 100 lançamentos de um dado?

Distribuição Geométrica:

A variável geométrica Y está associada ao nº de experimentos requeridos para que um único sucesso ocorra

quando ele acontece, $Y = y$

isso significa que ocorreram $y - 1$ falhas antes

$$p_Y(y) = p(1-p)^{y-1}, y = 1,2, \dots$$

$$P_Y(y) = \sum_{i=0}^{y} p(1-p)^{i-1}, y = 1,2, \dots$$

onde

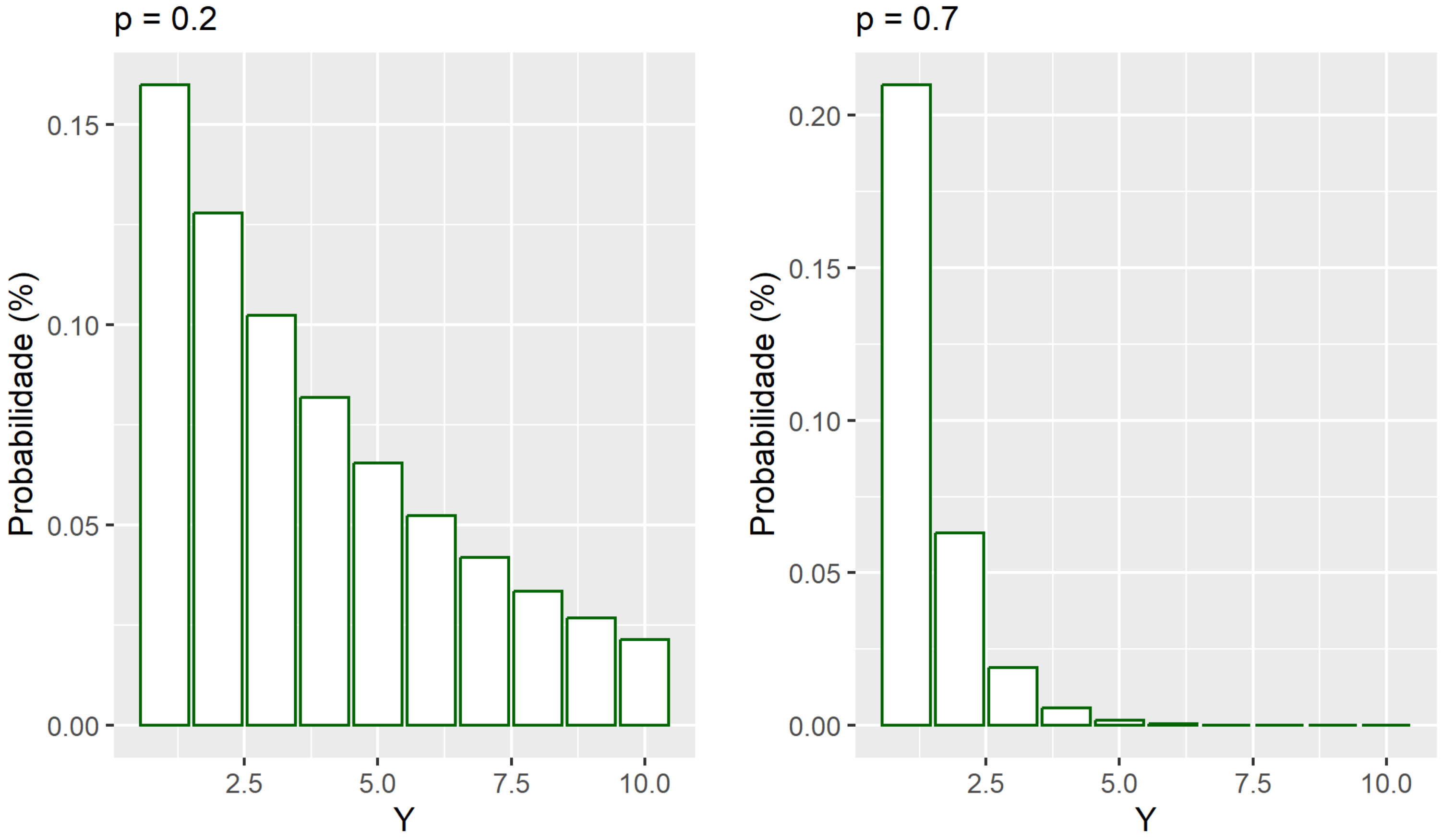
p probabilidade de sucesso

Distribuições de probabilidade | distribuição Geométrica

O valor esperado e a variância da variável geométrica são:

$$E[Y] = \frac{1}{p}$$

$$VAR[Y] = \frac{1-p}{p^2}$$



Exemplo: Se num rio ocorre uma cheia por ano e a probabilidade de que a cheia seja catastrófica é 5%, qual o número médio de anos que se deve esperar para observar uma nova cheia catastrófica?

Solução:
O enunciado pede o número médio de anos.

Das definições anteriores, sabe-se que a média de uma variável aleatória é o seu valor esperado.

$$\mu_Y = E[Y]$$

A distribuição geométrica tem como parâmetro:

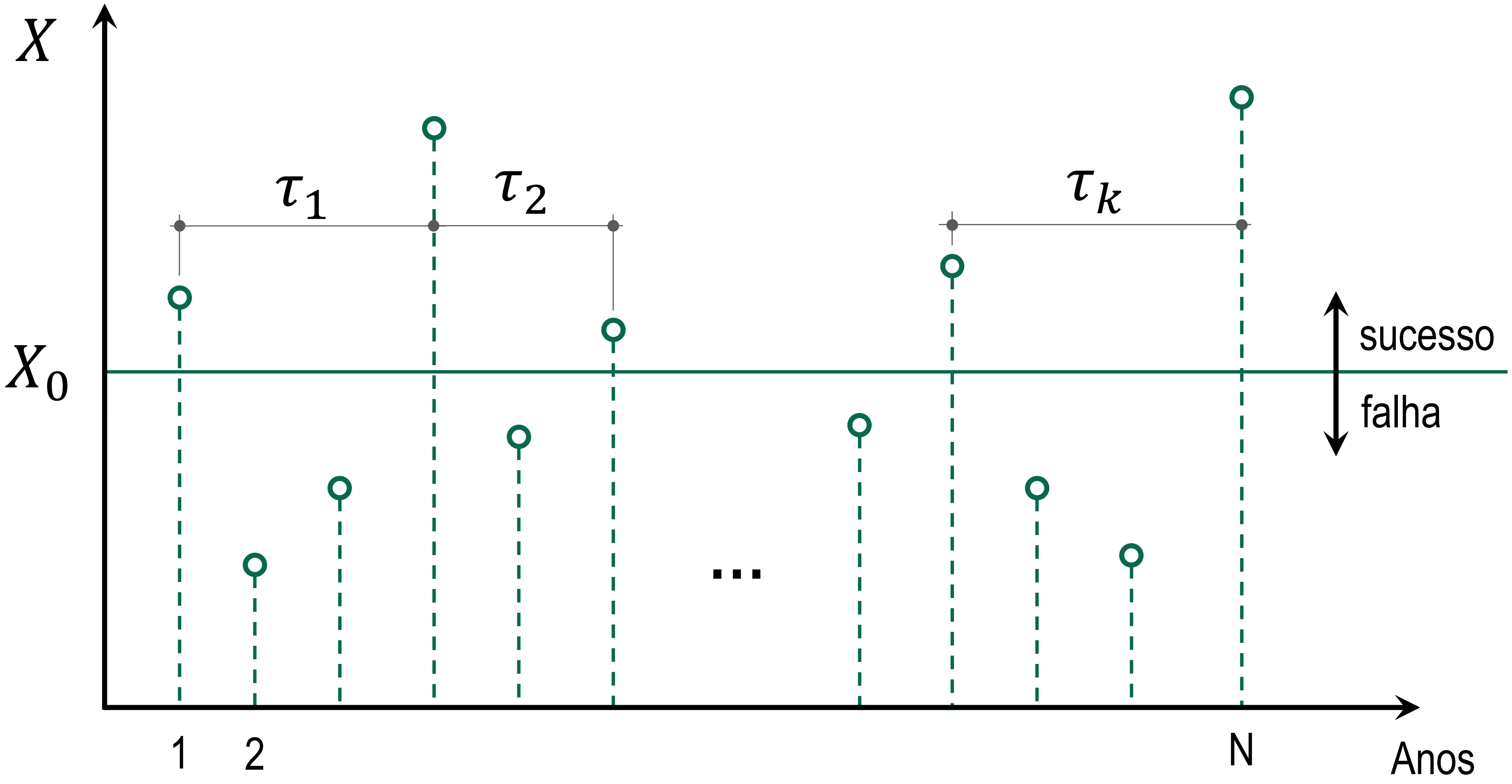
$$E[Y] = \frac{1}{p} \rightarrow E[Y] = \frac{1}{0,05} = 20$$

Assim, o número médio de anos para que a cheia catastrófica ocorra é de 20 anos.

Este é um resultado extremamente importante para a engenharia!

Distribuições de probabilidade | processos de Bernoulli

Considere uma série qualquer, com um limiar acima do qual os valores são considerados sucessos e abaixo falhas (análise *peak over threshold*)



τ_1: 3 anos
τ_2: 2 anos
(...)
τ_k: 3 anos

Distribuições de probabilidade | processos de Bernoulli

A variável τ é chamada de tempo de recorrência

Supondo que $N = 50$ anos e que foram observados 5 sucessos nesse período, tem-se que $\bar{\tau} = 10$ anos

lê-se: a variável X é superada, em média, uma vez a cada 10 anos

Nessas condições, $\tau \sim G(p)$, onde "~" denota "tem distribuição"

lê-se: a variável τ tem distribuição geométrica com parâmetro p

Portanto, é possível trabalhar com o parâmetro da distribuição para definir o tempo de retorno T

Distribuições de probabilidade | processos de Bernoulli

Assim, o tempo de retorno T é o valor esperado do tempo de recorrência τ:

$$T = E[\tau] = \frac{1}{p}$$

T expressa o número médio de anos para que um evento catastrófico ocorra

Importante:

Nessa definição, sucesso é o evento catastrófico

T não é um tempo cronológico

T não é uma previsão!!

Distribuições de probabilidade | distribuição Geométrica

Exemplo: Qual é a probabilidade de que uma cheia de 10 anos de retorno ocorra pela primeira vez no quinto ano depois do início da operação de uma obra hidráulica?

Solução:

$T = 10$ anos

$y = 5$

$$T = \frac{1}{p} \rightarrow p = \frac{1}{T} = \frac{1}{10} = 0{,}10$$

Distribuições de probabilidade | distribuição Geométrica

FMP distribuição geométrica

$$p_Y(y) = p(1-p)^{y-1} = 0{,}10(1-0{,}10)^{5-1} = 0{,}06561$$

Portanto, a probabilidade de que uma cheia com 10 anos de retorno ocorra no quinto ano de operação da obra é de 6,56%.

No R

```
dgeom(y-1,prob = p)
dgeom(4,prob = 0.1)
[1] 0.06561
```

Distribuição Binomial:

A variável binomial Y está associada ao nº de sucessos entre as N possibilidades de um experimento

agora, $Y = y = 0,1,2, \dots, N$

cada ponto do espaço amostral terá y sucessos e $N - y$ falhas:

Possibilidade 1:

F	S	S	F	F	...	F	S
	1	2					y

Possibilidade 2:

S	S	F	F	S	...	S	F
1	1			1		y	

(...)

Distribuições de probabilidade | distribuição Binomial

Ou seja, os sucessos e as falhas podem ser combinados de C_y^N maneiras

$$p_Y(y) = \binom{N}{y} p^y (1-p)^{N-y}, y = 1,2,\ldots$$

$$P_Y(y) = \sum_{i=0}^{y} \binom{N}{i} p^i (1-p)^{N-i}, y = 1,2,\ldots$$

onde

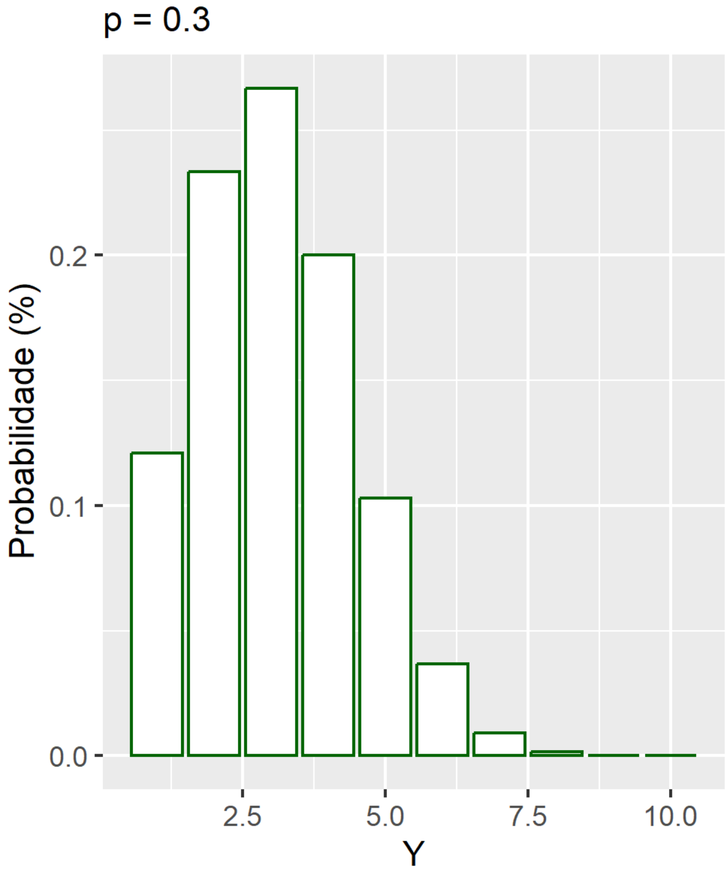
p probabilidade de sucesso

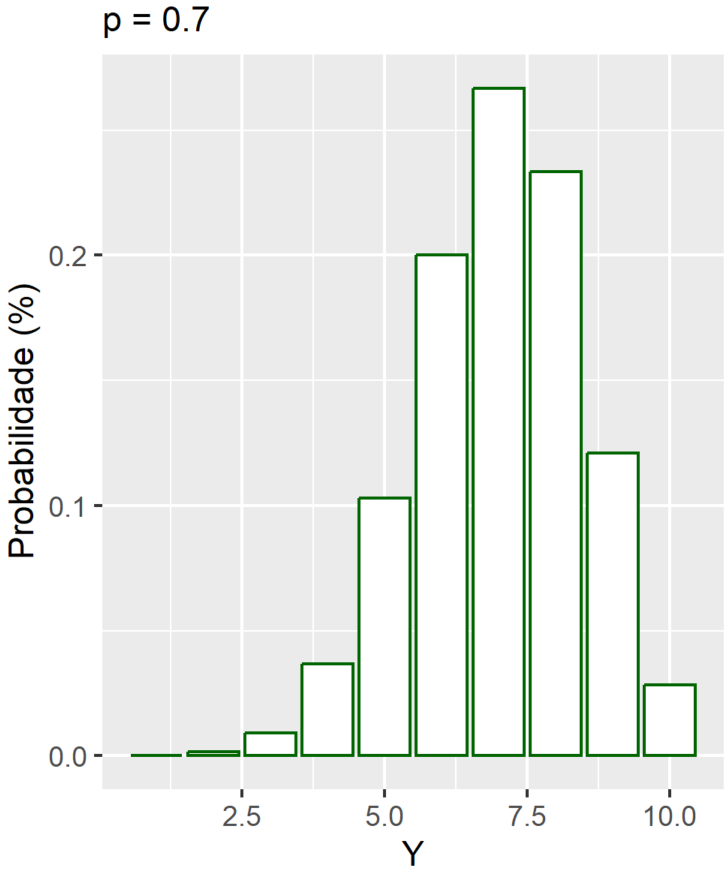
Distribuições de probabilidade | distribuição Binomial

O valor esperado e a variância da variável binomial são:

$$E[Y] = N \cdot p$$

$$VAR[Y] = N \cdot p \cdot (1 - p)$$





Distribuições de probabilidade | distribuição Binomial

Exemplo: Em média, quantas vezes uma cheia de 10 anos de retorno irá ocorrer em 40 anos de operação de uma obra hidráulica? Qual é a probabilidade de que exatamente esse número ocorra em 40 anos?

Solução:
Como visto, uma cheia de 10 anos de retorno tem $p = 0{,}10$.

Do valor esperado da distribuição binomial, tem-se:

$$E[Y] = N \cdot p = 40 \cdot 0{,}10 = 4$$

Distribuições de probabilidade | distribuição Binomial

Portanto, uma cheia com tempo de retorno de 10 anos irá ocorrer 4 vezes em 40 anos (o que é esperado...)

Contudo, a probabilidade de exatamente 4 cheias com $T = 10$ anos acontecerem em 40 anos é:

$$p_Y(y) = \binom{N}{y} p^y (1-p)^{N-y} = \binom{40}{4} 0{,}10^4 (1-0{,}10)^{40-4} = 0{,}2059$$

Dessa forma, a probabilidade é apenas 20,6%.

Distribuições de probabilidade | distribuição Binomial

O resultado permite concluir que em aproximadamente 80%* de toda as possíveis combinações de sucessos e falhas em 40 anos, as cheias de 10 anos de retorno não irão ocorrer exatamente 4 vezes.

$$^{*}\; 100 \times (1 - 0{,}2059) \cong 80\%$$

<u>No R</u>

```
dbinom(x = valor, size = anos, prob = p)
dbinom(x = 4, size = 40, prob = 0.1)
[1] 0.20588
```

Distribuições de probabilidade | distribuição Binomial

Trabalhando com o conceito de tempo de retorno no âmbito da distribuição Binomial, é possível estimar a probabilidade de que um evento de sucesso ocorra pelo menos uma vez em N anos

lembrando que sucesso é o evento catastrófico

nesse caso, N denota a vida útil do projeto

$$p_Y(y \geq 1) = 1 - p_Y(y = 0) = 1 - \binom{N}{0} p^0 (1-p)^{N-0}$$

Distribuições de probabilidade | distribuição Binomial

Como se sabe, $p = 1/T$. Substituindo na equação, chega-se a:

$$p_Y(y \geq 1) = 1 - \left(1 - \frac{1}{T}\right)^N$$

Esta equação expressa o risco hidrológico R:

$$R = 1 - \left(1 - \frac{1}{T}\right)^N$$

Exemplo: Em uma usina hidrelétrica, a vida útil da obra é de 50 anos e o vertedouro foi projetado para uma cheia com tempo de recorrência igual a 100 anos. Avaliar o risco do projeto.

Solução:

$T = 100$ anos

$N = 50$ anos

$$R = 1 - \left(1 - \frac{1}{100}\right)^{50} = 0{,}40$$

Em conclusão, o risco do projeto é de 40%.

Esse é considerado um valor alto. Alternativamente, pode-se recalcular T para um dado risco a ser assumido. Para $R = 5\%$:

$$0{,}05 = 1 - \left(1 - \frac{1}{T}\right)^{50} \rightarrow T \cong 975$$

Assim, o tempo de retorno para risco de 5% e 50 anos de vida útil do projeto é de 975 anos.

Distribuições de probabilidade | distribuição Binomial

<u>Exercício (proposto)</u>: Após a análise de 10 amostras de água, um laboratório fez uma contagem da presença da bactéria *E. Coli*. Os resultados, expressos em centenas de organismos por 100 ml de água foram:

$$\{17, 21, 25, 23, 17, 26, 24, 19, 21, 17\}$$

Sendo N o número total de organismos nas amostras e Y o número apenas de *E. Coli* (10^2/100ml), estimar a probabilidade de encontrar $Y = y = 20$.

Solução:

A probabilidade é calculada por meio da FMP da distribuição Binomial:

$$p_Y(y) = \binom{N}{y} p^y (1-p)^{N-y}$$

Entretanto, a única informação fornecida foi $y = 20$. As demais são estimadas com base nos parâmetros da distribuição, substituindo os valores populacionais ($E[Y]$, $VAR[Y]$) por estimativas amostrais ($\bar{y}$, s_y^2):

$$E[Y] = \bar{y} = N \cdot p$$
$$VAR[Y] = s_y^2 = N \cdot p \cdot (1-p)$$

Da amostra fornecida, tem-se:

$\bar{y} = 21$ (10^2/100 ml)
$s_y^2 = 11{,}78$ (10^2/100 ml)^2

Substituindo $N \cdot p$ por $\bar{y}$ na equação da variância, tem-se:

$$s_y^2 = \bar{y} \cdot (1 - p) \rightarrow (1 - p) = \frac{s_y^2}{\bar{y}} = \frac{11{,}78}{21}$$

$$\therefore p = 0{,}439$$

Distribuições de probabilidade | distribuição Binomial

Retornando à equação do valor esperado, chega-se a:

$$N = \frac{\bar{y}}{p} = \frac{21}{0{,}439} \cong 48$$

Por fim:

$$p_Y(20) = \binom{48}{20} 0{,}439^{20} (1 - 0{,}439)^{48-20} = 0{,}11$$

Assim, a probabilidade de encontrar 20 *E. Coli* (10^2/100ml) é de 11%.

Revisão

Distribuições de probabilidades são funções matemáticas com propriedades definidas e que expressam:

- probabilidades (ou densidades): funções massa e densidade de probabilidades
- probabilidades de excedência: funções de distribuição e densidade acumuladas

Distribuições discretas

- aplicadas a variáveis aleatórias discretas
- destacam-se as que são derivadas de processos de Bernoulli

Distribuição Geométrica

- nº de experimentos até ocorrer o sucesso

Distribuição Binomial

- nº de sucessos entre as diferentes possibilidades de um experimento

Estatística Aplicada a Ciências Ambientais

Daniel Detzel
detzel@ufpr.br