Estatística Básica

MEDIDAS RESUMO

Renato Dourado Maia
Instituto de Ciências Agrárias
Universidade Federal de Minas Gerais

Motivação Básica

- Se você estivesse num ponto de ônibus e alguém perguntasse sobre o **quanto** um **determinado ônibus demora para passar**, o que você responderia?
  - Com certeza você **não apresentaria uma tabela de frequências** e tampouco **um modelo teórico** para a variável aleatória de interesse!
- Quem perguntou deseja uma **resposta breve** que **sintetize a informação** e não uma completa descrição completa dos dados coletados ou de uma modelagem que eventualmente foi feita.

Motivação Básica

- Nesta unidade serão apresentadas algumas **medidas** que **sumarizam as informações disponíveis** sobre o **comportamento** de uma **variável**, seja para um **conjunto de dados** – toda a **população** ou uma **amostra**, seja para **variáveis aleatórias**.
- Essencialmente, estudaremos medidas que representam **tendência central** e **dispersão**.

Medidas de Posição (Tendência Central)

- Seja uma variável X com observações x_1, x_2, ...,x_n.
  - A **mediana** (md_{obs}) é o valor que ocupa a posição central dos dados ordenados.
  - A **moda** (mo_{obs}) é dada pelo valor mais frequente.
  - A **média** é dada pela soma das observações dividida pelo número de observações:

$$\bar{x}_{obs} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \underbrace{\frac{n_1 x_1 + n_2 x_2 + \cdots + n_k x_k}{n_1 + \cdots + n_k} = \frac{\sum_{i=1}^{k} n_i x_i}{n} = \sum_{i=1}^{k} \frac{n_i}{n} x_i}_{\textbf{Tabela de Frequências}} .$$

Exemplo 1

- Suponha que parafusos a serem utilizados em tomadas elétricas são embalados em caixas rotuladas como contento 100 unidades. Em uma construção, 10 caixas de um lote tiveram o número de parafusos contados, fornecendo os valores 98, 102, 100, 100, 99, 97, 96, 95, 99, 100. Determine, para o **número de parafusos por caixa**, a **média**, a **moda** e a **mediana**.

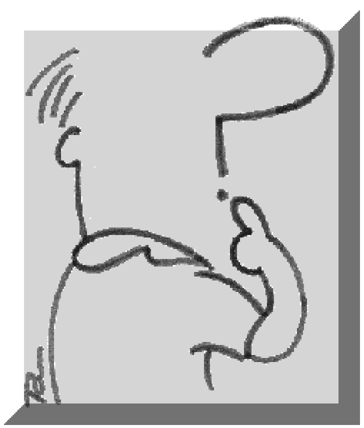
Exemplo 1

- **Dados ordenados**: 95, 96, 97, 98, 99, 99, 100, 100, 100, 102.

$$
\begin{cases}
\rightarrow md_{obs} = \dfrac{quinto\ elemento + sexto\ elemento}{2} = \dfrac{99+99}{2} = 99 \\
\rightarrow mo_{obs} = 100 \\
\rightarrow \bar{x}_{obs} = \dfrac{98+102+\cdots+100}{10} = \dfrac{986}{10} = 98{,}6
\end{cases}
$$

Pergunta

Quando utilizar cada uma das **medidas de posição**?

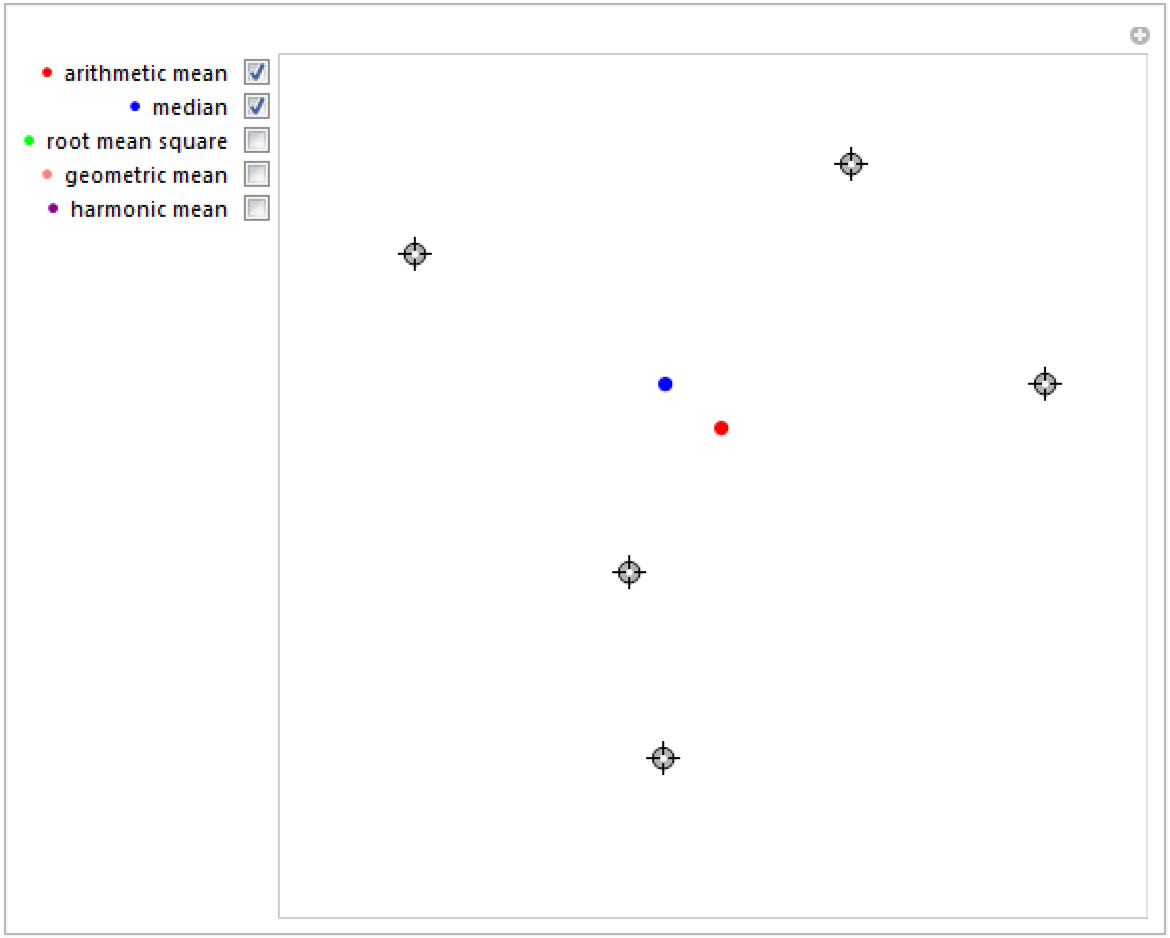


Medidas de Posição (Tendência Central)

- As **medidas de posição** podem ser utilizadas **em conjunto** para **auxiliar** na **análise** dos dados. Em determinadas situações, entretanto, **uma** pode ser **mais conveniente** do que as **outras**.
  - Se há valores **discrepantes**, a **mediana** é **menos afetada** do que a **média**.
  - Se há **muitas observações**, a mediana é mais **difícil** de ser calculada, mesmo considerando a utilização de **computadores**, pois o **processo de ordenação é custoso**.
  - Há distribuições **multimodais** – conjuntos de dados com **mais do que uma moda**.

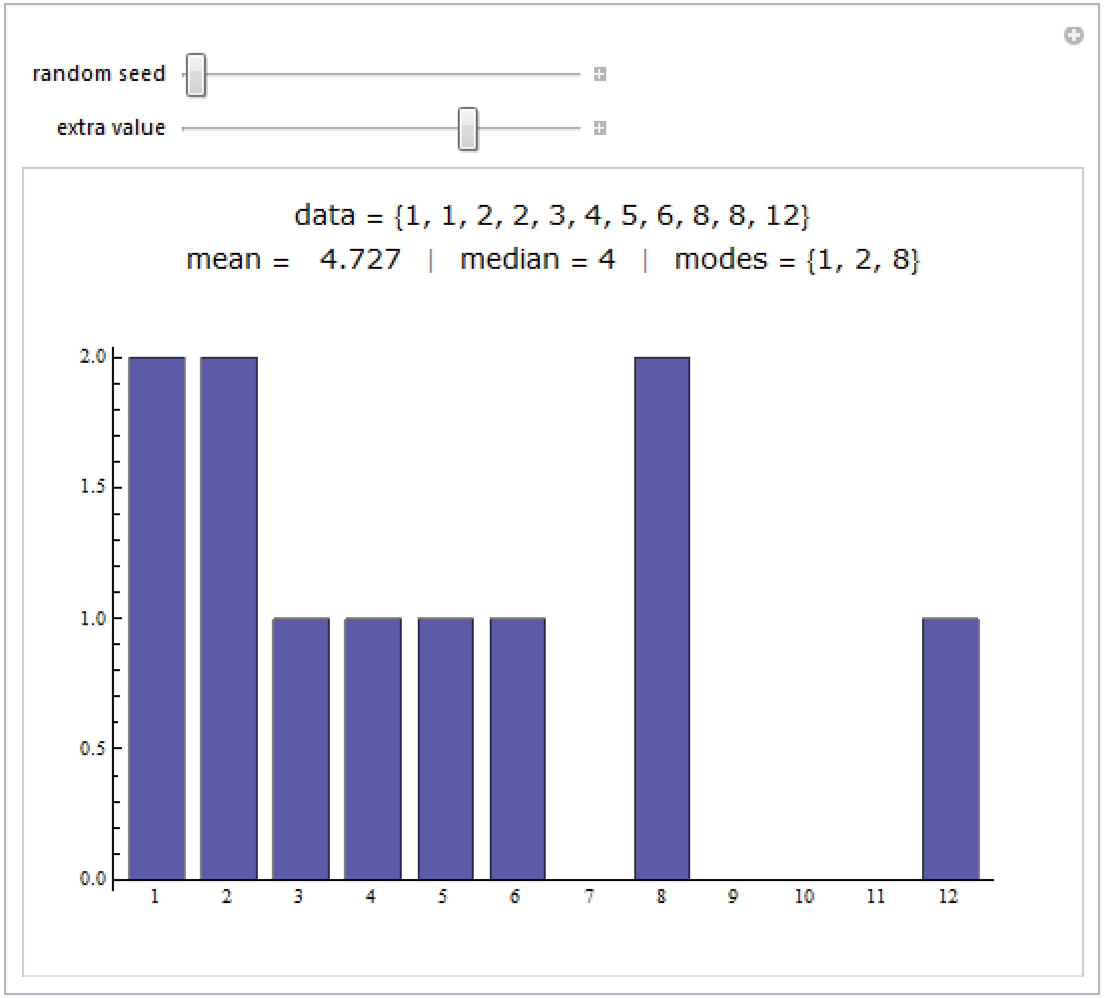
"Basic Statistics of Movable Points" from the Wolfram Demonstrations Project

http://demonstrations.wolfram.com/BasicStatisticsOfMovablePoints



"Mean, Median, Mode" from the Wolfram Demonstrations Project

http://demonstrations.wolfram.com/MeanMedianMode/



Medidas de Posição (Tendência Central)

- Seja uma VA X com possíveis valores x_1, x_2, ...,x_k e correspondentes probabilidades p_1, p_2, ..., p_k.
  - A **média**, **valor esperado** ou **esperança matemática** de X, denotada por $E(X)$, μ_X, ou simplesmente μ, é dada pela expressão:

$$E(X) = \sum_{i=1}^{k} x_i p_i .$$

  - A **moda** (Mo) de X é(são) o(s) valor(es) da variável com **maior probabilidade**, ou seja:

$$P(X = Mo) = max(p_{1,} p_{2,} \cdots, p_k) .$$

Medidas de Posição (Tendência Central)

- Seja uma VA X com possíveis valores $x_1, x_2, \ldots, x_k$ e correspondentes probabilidades $p_1, p_2, \ldots, p_k$.
  - A **mediana** de X (Md) é o valor que satisfaz às seguintes condições:

$$P\left(X \geq Md\right) \geq 0{,}5 \quad e \quad P\left(X \leq Md\right) \geq 0{,}5.$$

  - Em algumas situações, as desigualdades são satisfeitas por **qualquer valor num certo intervalo** e, nesse caso, **toma-se a mediana como o ponto médio do intervalo**.

Exemplo 2

- Determine a **média**, e **moda** e a **mediana** para a VA X com a **função discreta de probabilidade** apresentada na tabela a seguir.

X	-5	10	15	20
p_i	0,3	0,2	0,4	0,1

$$\begin{cases} \rightarrow \mu = \sum_{i=1}^{4} x_i p_i = (-5)\times(0{,}3) + 10\times 0{,}2 + 15\times 0{,}4 + 20\times 0{,}1 = 8{,}5 \\ \rightarrow Mo = 15 \\ \rightarrow Md \in [10{,}15] \Rightarrow Md = 12{,}5 \end{cases}$$

Medidas de Posição (Tendência Central)

	Conjunto de Dados	Variável Aleatória
Valores	$\begin{array}{ccccc} \boldsymbol{X} & x_1 & x_2 & \ldots & x_k \\ \boldsymbol{n_i} & n_1 & n_2 & \ldots & n_k \end{array}$	$\begin{array}{ccccc} \boldsymbol{X} & x_1 & x_2 & \ldots & x_k \\ \boldsymbol{p_i} & p_1 & p_2 & \ldots & p_k \end{array}$
Média	$\bar{x}_{obs} = \frac{1}{n} \sum_{i=1}^{k} n_i x_i$	$\mu = \sum_{i=1}^{k} x_i p_i$
Mediana	$md_{obs} = valor\ central$	$P(X \geq Md) \geq 0{,}5 \quad e$ $P(X \leq Md) \geq 0{,}5$
Moda	$mo_{obs} = valor\ com\ maior\ frequência$	$Mo = valor\ com\ maior\ probabilidade$

Pergunta

A **renda *per capita*** é **a-**
propriada para **compa-**
rar a **distribuição de**
renda de dois países?

Medidas de Dispersão

- As **medidas de posição** podem **esconder** valiosas **informações**, pois elas **não capturam a variabilidade** dos valores da variável em estudo.
  - **Diferentes** conjuntos de dados, por exemplo, podem ter **medidas de posição idênticas**.
  - Nesses casos, é provável que existam **diferenças** em relação à **dispersão dos dados**, isto é, quanto à **maneira como os valores de cada conjunto se espalham**.
- Para quantificar a **dispersão dos dados**, são utilizadas **medidas de dispersão**.

Medidas de Dispersão

Medidas de Dispersão – Conjunto de Dados

- A **amplitude** (Δ) é definida como a **diferença** entre o **maior** e o **menor** valor do conjunto de dados.
- **Desvio mediano**:

$$desvio\ \ mediano = \frac{1}{n} \sum_{i=1}^{n} \left| x_i - md_{obs} \right|.$$

- **Desvio médio**:

$$desvio\ \ médio = \frac{1}{n} \sum_{i=1}^{n} \left| x_i - \bar{x}_{obs} \right|.$$

Medidas de Dispersão – Conjunto de Dados

- Os **desvios mediano** e **médio** utilizam a **função módulo**, cuja derivada é **descontínua**, o que **dificulta** o tratamento matemático:
  - O que se pode dizer sobre a **derivada** da **função módulo**?
- Em função disso, é comum a utilização de uma **medida de dispersão** que tenha **propriedades matemáticas mais interessantes**.

Medidas de Dispersão – Conjunto de Dados

- A **variância** de uma variável X referente a um **conjunto de dados**, denotada por var_{obs}, pode ser calculada por meio da seguinte expressão:

$$var_{obs} = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \bar{x}_{obs} \right)^2 .$$

- O **desvio-padrão** (dp_{obs}), que tem a **mesma unidade dos dados originais**, é definido como a **raiz quadrada positiva da variância**.

Medidas de Dispersão – Conjunto de Dados

- A expressão a seguir **facilita** o cálculo da **variância**:

$$var_{obs} = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 \right) - \bar{x}_{obs}^2 .$$

- Para o caso de os dados estarem disponíveis numa **tabela de frequências**, são utilizadas as seguintes expressões para o cálculo da **variância**:

$$var_{obs} = \frac{1}{n} \sum_{i=1}^{k} n_i \left(x_i - \bar{x}_{obs} \right)^2 = \left(\frac{1}{n} \sum_{i=1}^{k} n_i x_i^2 \right) - \bar{x}_{obs}^2 .$$

Medidas de Dispersão – Variável Aleatória

- Seja uma VA X com $P(X_i = x_i) = p_i$, $i = 1, 2, \ldots, k$, e média μ. A **variância** de X é a **ponderação**, pelas respectivas **probabilidades**, dos **desvios relativos à média**, elevados ao quadrado, isto é,

$$Var(X) = \sum_{i=1}^{k} (x_i - \mu)^2 p_i.$$

- Normalmente, a **variância** por σ^2 e o **desvio-padrão** por σ.

Medidas de Dispersão – Variável Aleatória

- A **variância** de definida no *slide* anterior pode ainda ser considerada como o **valor esperado** de uma **nova variável aleatória**, o **desvio ao quadrado**, isto é,

$$Var(X) = E\left[(X-\mu)^2\right].$$

- Essa expressão pode ser **reescrita** da seguinte forma:

$$Var(X) = E\left(X^2\right) - \mu^2 = \sum_{i=i}^{k} p_i x_i^2 - \mu^2 .$$

Medidas de Dispersão

	Conjunto de Dados	Variável Aleatória
Valores	$\begin{array}{ccccc} \boldsymbol{X} & x_1 & x_2 & \dots & x_k \\ \boldsymbol{n_i} & n_1 & n_2 & \dots & n_k \end{array}$	$\begin{array}{ccccc} \boldsymbol{X} & x_1 & x_2 & \dots & x_k \\ \boldsymbol{p_i} & p_1 & p_2 & \dots & p_k \end{array}$
Variância	$\begin{cases} var_{obs} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{k} n_i \left(x_i - \bar{x}_{obs} \right)^2 \\ var_{obs} = \left(\dfrac{1}{n} \displaystyle\sum_{i=1}^{k} n_i x_i^2 \right) - \bar{x}_{obs}^2 \end{cases}$	$\begin{cases} \sigma^2 = E\left[\left(X - \mu \right)^2 \right] = \displaystyle\sum_{i=1}^{k} p_i \left(x_i - \mu \right) \\ \sigma^2 = E\left(X^2 \right) - \mu^2 = \displaystyle\sum_{i=i}^{k} p_i x_i^2 - \mu^2 \end{cases}$

Propriedades da Média e da Variância

Conjunto de Dados	Variável Aleatória
$Y = aX + b$	$Y = aX + b$
$\bar{y}_{obs} = a\,\bar{x}_{obs} + b$	$E(Y) = aE(X) + b$
$var_{obs}(Y) = a^2\, var_{obs}(X)$	$Var(Y) = a^2\, Var(X)$

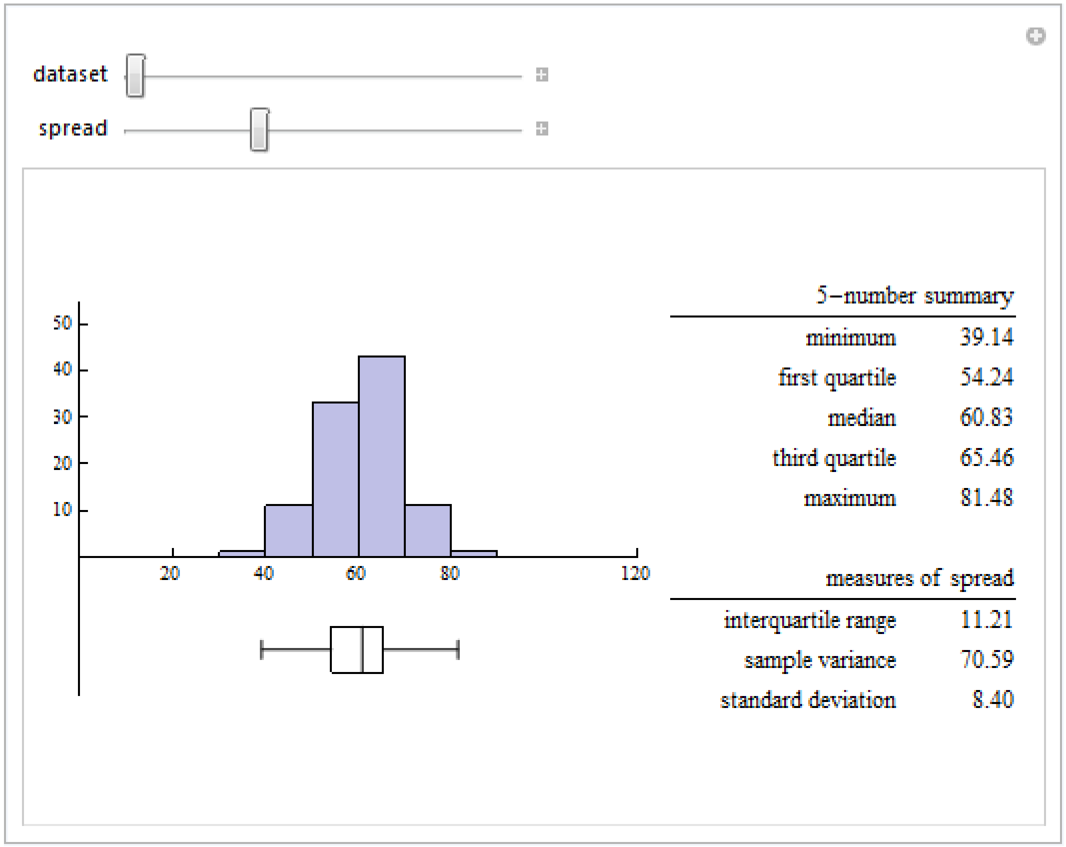
Exercício interessante: demonstrar essas propriedades...

Variável Discreta	Valor Esperado	Variância
Uniforme (1, k)	$\frac{1+k}{2}$	$\frac{k^2-1}{12}$
Bernoulli (p)	p	$p(1-p)$
Binomial (n, p)	np	$np(1-p)$
Geométrica (p)	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$
Poisson (λ)	λ	λ
Hipergeométrica (n, m, r)	$\frac{rm}{n}$	$\frac{rm(n-m)(n-r)}{n^2(n-1)}$

Exercício interessante: calcular...

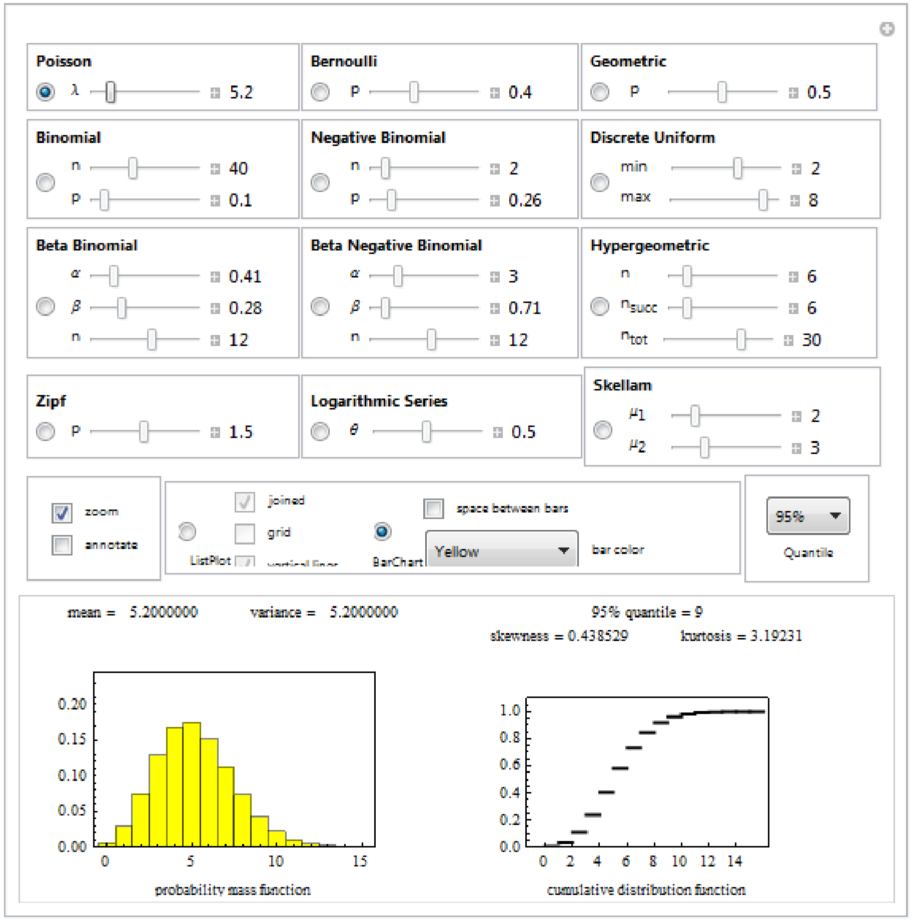
"Descriptions of Univariate Data" from the Wolfram Demonstrations Project

http://demonstrations.wolfram.com/DescriptionsOfUnivariateData



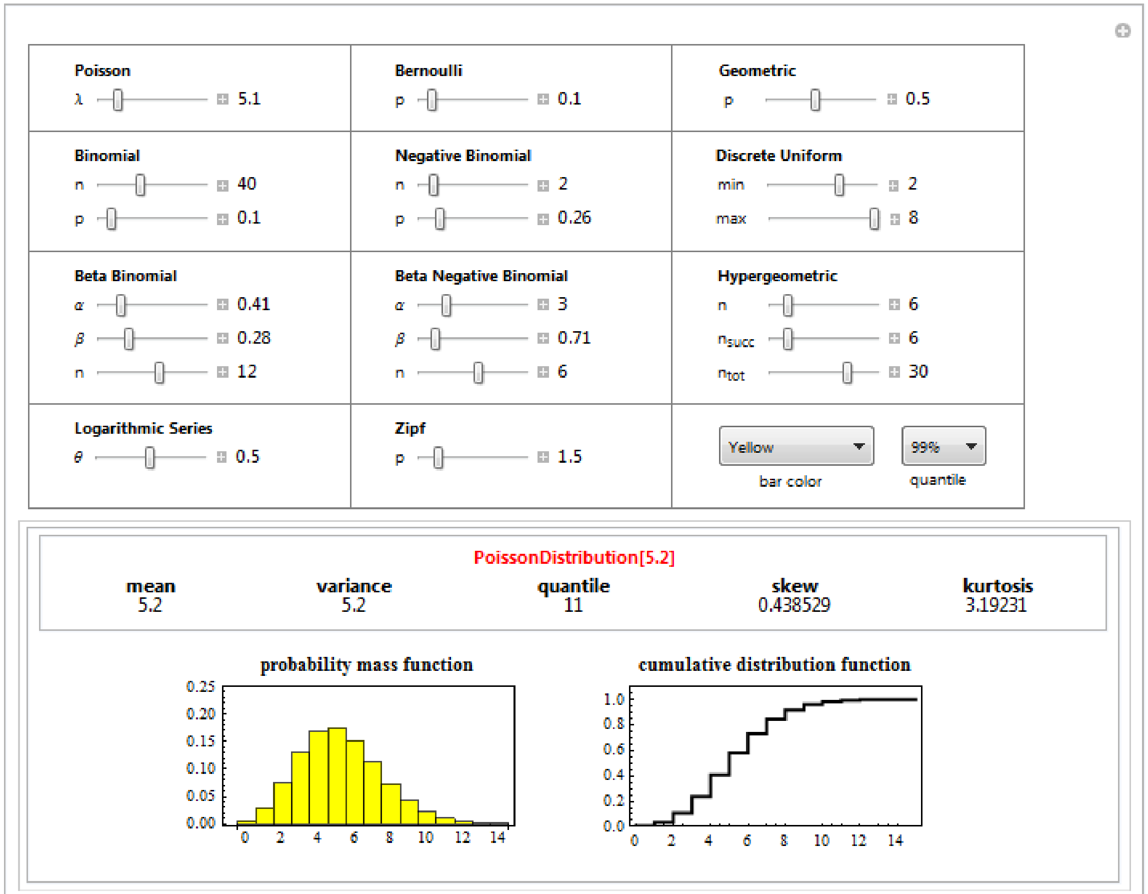
"Illustrating the Use of Discrete Distributions" from the Wolfram Demonstrations Project

http://demonstrations.wolfram.com/IllustratingTheUseOfDiscreteDistributions



"Mathematica 7's Discrete Distributions" from the Wolfram Demonstrations Project

http://demonstrations.wolfram.com/Mathematica7sDiscreteDistributions/



That's All Folks!

Próxima aula: exercícios do capítulo 4...