Fundamentos Estatísticos para Ciência dos Dados
Variáveis Aleatórias

Renato Martins Assunção

DCC, UFMG - 2015

Probabilidade e dados

- Probabilidade é um assunto de matemática.

Probabilidade e dados

- Probabilidade é um assunto de matemática.
- Estabelece um espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$ e faz cálculos (matemáticos) de probabilidade.

Probabilidade e dados

- Probabilidade é um assunto de matemática.
- Estabelece um espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$ e faz cálculos (matemáticos) de probabilidade.
- Estatística, data mining e machine learning são assuntos que lidam com dados.

Probabilidade e dados

- Probabilidade é um assunto de matemática.
- Estabelece um espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$ e faz cálculos (matemáticos) de probabilidade.
- Estatística, data mining e machine learning são assuntos que lidam com dados.
- Uma tabela cheia de números:

Probabilidade e dados

- Probabilidade é um assunto de matemática.
- Estabelece um espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$ e faz cálculos (matemáticos) de probabilidade.
- Estatística, data mining e machine learning são assuntos que lidam com dados.
- Uma tabela cheia de números: linhas são itens, colunas são atributos medidos nos itens.
- Como ligar estes dois assuntos?

Probabilidade e dados

- Probabilidade é um assunto de matemática.
- Estabelece um espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$ e faz cálculos (matemáticos) de probabilidade.
- Estatística, data mining e machine learning são assuntos que lidam com dados.
- Uma tabela cheia de números: linhas são itens, colunas são atributos medidos nos itens.
- Como ligar estes dois assuntos?
- O link é fornecido pelo conceito de *variável aleatória*.

Variáveis aletórias como redução de Ω

- O espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$ é a base matemática da probabilidade.

Variáveis aletórias como redução de Ω

- O espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$ é a base matemática da probabilidade.
- O espaço de probabilidade precisa atribuir probabilidade a todo evento $A \subset \Omega$.

Variáveis aletórias como redução de Ω

- O espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$ é a base matemática da probabilidade.
- O espaço de probabilidade precisa atribuir probabilidade a todo evento $A \subset \Omega$.
- Se Ω for muito complicado, podemos estar interessados apenas em *alguns* aspectos específicos do experimento aleatório.

Variáveis aletórias como redução de Ω

- O espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$ é a base matemática da probabilidade.
- O espaço de probabilidade precisa atribuir probabilidade a todo evento $A \subset \Omega$.
- Se Ω for muito complicado, podemos estar interessados apenas em *alguns* aspectos específicos do experimento aleatório.
- *Variáveis aleatórias* (v.a.) constituem a ferramenta para reduzir o espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$ ao mínimo necessário na prática.

Variáveis aletórias como redução de Ω

- O espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$ é a base matemática da probabilidade.
- O espaço de probabilidade precisa atribuir probabilidade a todo evento $A \subset \Omega$.
- Se Ω for muito complicado, podemos estar interessados apenas em *alguns* aspectos específicos do experimento aleatório.
- *Variáveis aleatórias* (v.a.) constituem a ferramenta para reduzir o espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$ ao mínimo necessário na prática.
- Variáves aleatórias são características numéricas do fenômenos aleatório.

Variáveis aletórias como redução de Ω

- O espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$ é a base matemática da probabilidade.
- O espaço de probabilidade precisa atribuir probabilidade a todo evento $A \subset \Omega$.
- Se Ω for muito complicado, podemos estar interessados apenas em *alguns* aspectos específicos do experimento aleatório.
- *Variáveis aleatórias* (v.a.) constituem a ferramenta para reduzir o espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$ ao mínimo necessário na prática.
- Variáves aleatórias são características numéricas do fenômenos aleatório.

Variáveis aletórias: formalismo

- Formalmente, variável aleatória é uma função matemática (mensurável) de Ω para $\mathbb{R}$.

$$\begin{aligned} X : \Omega &\rightarrow \mathbb{R} \\ \omega &\rightarrow X(\omega) \end{aligned}$$

Variáveis aletórias: formalismo

- Formalmente, variável aleatória é uma função matemática (mensurável) de Ω para $\mathbb{R}$.

$$\begin{aligned} X : \Omega &\rightarrow \mathbb{R} \\ \omega &\rightarrow X(\omega) \end{aligned}$$

- Isto é, uma variável aleatória é qualquer função matemática X que vai de Ω para $\mathbb{R}$.
- A condição de ser mensurável na definição acima é bastante técnica.
- Toda função "prática" é mensurável: toda função envolvendo logs, exponenciais, polinômios, funções trigonométricas, é mensurável.

Variáveis Aletórias e tabelas de dados

- Lembre-se da tabela de dados estatísticos:
  - nas linhas, itens ou indivíduos ou exemplos (tais como diferentes pacientes com câncer de um hospital, diferentes clientes de um banco). Cada linha representa uma diferente realização ou instanciação de elementos ω de Ω.
  - nas colunas, características ou atributos dos itens (sexo, idade e estágio do câncer; saldo médio na conta corrente, tempo como correntista)
- Informalmente, *variáveis aleatórias* (v.a.) são as representações *matemáticas ou probabilísticas* dessas colunas de atributos na tabela de dados estatísticos.

Variáveis Aletórias e tabelas de dados

- Lembre-se da tabela de dados estatísticos:
  - nas linhas, itens ou indivíduos ou exemplos (tais como diferentes pacientes com câncer de um hospital, diferentes clientes de um banco). Cada linha representa uma diferente realização ou instanciação de elementos ω de Ω.
  - nas colunas, características ou atributos dos itens (sexo, idade e estágio do câncer; saldo médio na conta corrente, tempo como correntista)
- Informalmente, *variáveis aleatórias* (v.a.) são as representações *matemáticas ou probabilísticas* dessas colunas de atributos na tabela de dados estatísticos.
- Como é a conexão entre a tabela de dados e o modelo probabilístico?

Espaço de probabilidade e tabela de dados

- Espaço de probabilidade: $(\Omega, \mathcal{A}, \mathbb{P})$
- Tabela de dados:

	spam	num_char	line_breaks	format	number
1	no	21,705	551	html	small
2	no	7,011	183	html	big
3	yes	631	28	text	none
⋮	⋮	⋮	⋮	⋮	⋮
50	no	15,829	242	html	small

Tabela: Quatro primeiras linhas da tabela `spam`. Fonte: OpenIntro Statistics Project, `https://www.openintro.org/stat/textbook.php`.

Espaço de probabilidade e tabela de dados

- Ω = conjunto de todos os e-mails já recebidos e a receber
- Matriz contém aenas uma AMOSTRA dos elementos de Ω
- Cada *LINHA* da tabela corresponde a um elemento distinto de Ω
- Cada coluna, representa diferente características ou medições sobre os e-mails
- EM GERAL, supomos que os diferentes e-mails (diferentes linhas) representam eventos independentes uns dos outros
- Em cada linha, os elementos medem características diferentes do MESMO ω
- Assim, os elementos *dentro de uma mesma linha* costumam estar associados/correlacionados (definição precisa mais tarde).

Variáveis Aletórias e tabelas de dados

- Os elementos numa coluna da tabela de dados é vista como instãncias de uma v.a.
- Toda v.a. pode ser pensada simplesmente como sendo a combinação de DOIS componentes:

Variáveis Aletórias e tabelas de dados

- Os elementos numa coluna da tabela de dados é vista como instãncias de uma v.a.
- Toda v.a. pode ser pensada simplesmente como sendo a combinação de DOIS componentes:
  - um conjunto de valores possíveis $\in \mathbb{R}$.

Variáveis Aletórias e tabelas de dados

- Os elementos numa coluna da tabela de dados é vista como instãncias de uma v.a.
- Toda v.a. pode ser pensada simplesmente como sendo a combinação de DOIS componentes:
  - um conjunto de valores possíveis $\in \mathbb{R}$.
  - probabilidades associadas a estes valores.

Variáveis Aletórias e tabelas de dados

- Os elementos numa coluna da tabela de dados é vista como instãncias de uma v.a.
- Toda v.a. pode ser pensada simplesmente como sendo a combinação de DOIS componentes:
  - um conjunto de valores possíveis $\in \mathbb{R}$.
  - probabilidades associadas a estes valores.
- As probabilidades vêm de um modelo $(\Omega, \mathcal{A}, \mathbb{P})$ que muitas vezes *não* precisa ser explicitamente apresentado. Isto facilita muito a vida.

Tipos de v.a.'s

- As v.a.'s são representadas por letras maiúsculas: $X, Y, W, U, Z, \ldots$.
- Temos três tipos básicos de *dados estatísticos* nas tabelas de dados estatísticos:
  - dados categóricos ou não-numéricos, que podem ser nominais (tais como sexo ou religião) ou ordinais (por exemplo, a resposta a uma pergunta como "Você confia muito, pouco ou nada nos membros do Congresso?")
  - dados numéricos discretos: número de filhos,número de requisições nas últimas duas horas.
  - dados numéricos contínuos: saldo na conta corrente, temperatura, índice de inflação.
- Estes dados são representados por dois tipos de variáveis aleatórias:
  - **V.A.s Discretas:** Para os dados categóricos ou numéricos discretos.
  - **V.A.s Contínuas:** Para os dados numéricos contínuos.

V.A. Discreta

- **V.A.s Discretas:** Para os dados categóricos ou numéricos discretos.
- Composta de duas LISTAS enumeráveis.
  - Uma lista de valores possíveis para a v.a.: $\{x_1, x_2, \ldots\}$
  - Uma lista com a probabilidade associada a cada um desses valores: $\{p(x_1), p(x_2), \ldots\}$
- Podemos representar as duas listas numa tabela:

Valores possíveis	x_1	x_2	x_3	$\ldots$
Probab assoc	$p(x_1)$	$p(x_2)$	$p(x_3)$	$\ldots$

- A lista de probabilidades deve ter valores ≥ 0 e eles devem somar 1.

Gráfico das probabilidades associadas

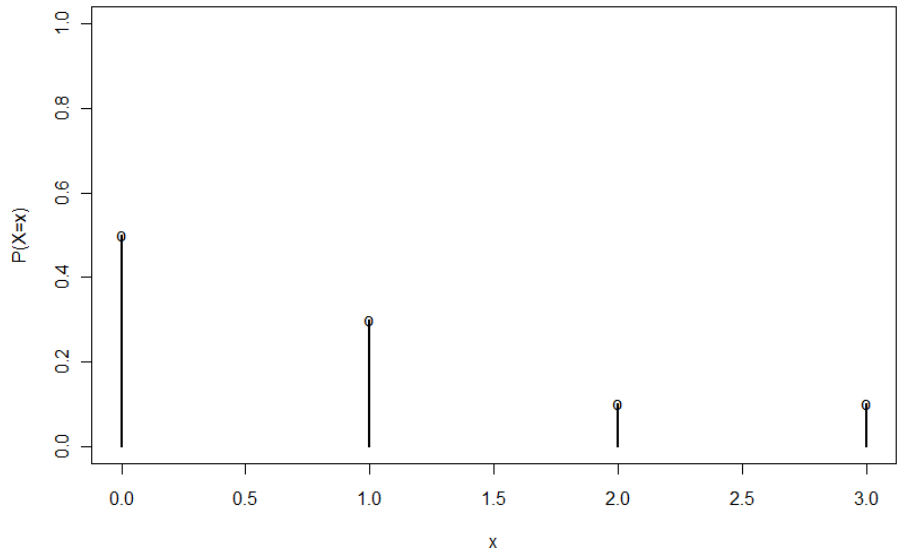


Figura: Função $\mathbb{P}(X = x_i)$ onde x_i é um dos valores possíveis de X. Também chamada de função massa de probabilidade. X tem valores possíveis $\{0, 1, 2, 3\}$ com probabilidades $\{0.5, 0.3, 0.1, 0.1\}$, respectivamente.

V.A.s discretas - exemplos

- Uma coluna da tabela de dados indica o sexo de um indivíduo ω escolhido de uma população.
- Arbitrariamente, associamos o valor 0 a MASC e 1 a FEM.
- Isto é, $X(\omega) = 0$ se ω for do sexo masculino e $X(\omega) = 1$ se ω for do sexo feminino.
- Para cada indivíduo ω olhamos apenas seu sexo, representado por $X(\omega) \in \{0, 1\}$.
- Para acabar a especificação dessa v.a. discreta, precisamos especificar a lista de probabilidades associada.
- Digamos, $\mathbb{P}(X = 0) = p(0) = 0.35$ e $\mathbb{P}(X = 1) = p(1) = 1 - 0.35 = 0.65$.

V.A.s discretas - exemplos

- Num posto de gasolina, monitora-se a cada 5 minutos durante as horas de pico o uso de suas 4 bombas de abastecimento de veículos.
- De 5 em 5 minutos, anota-se o número de bombas em uso.
- Os itens ou instâncias são os diferentes instantes de tempo.
- Os dados são numéricos discretos e,em cada instante, podem ser 0, 1, 2, 3 ou 4.
- Seja ω um dos instantes de tempo.
- $X(\omega)$ é o número de bombas em uso.
- É preciso também especificar as probabilidades de cada valor possível para X. Por exemplo:

Valores possíveis	0	1	2	3	4
Probab assoc	$p(0) = 0.32$	$p(1) = 0.42$	$p(2) = 0.21$	$p(3) = 0.04$	$p(4) = 0.01$

V.A.s discretas - exemplos

- Numa rede social, escolha n usuários-vértices ao acaso e conte o número de arestas incidentes de cada um deles. (seguidores).
- Os itens ou instâncias são os diferentes usuários
- Os dados são numéricos discretos e podem ser $0, 1, 2, 3, \ldots$ sem um limite máximo natural.
- Seja ω um dos usuários e $X(\omega)$ o seu número de seguidores.
- $X(\omega) \in \{0, 1, 2, 3, \ldots\} = \mathbb{N}$.
- Especificando as probabilidades (sem explicar de onde tiramos isto):

Val. pos. k	0	1	2	3	...	223	...
Probab $p(k)$	0.001	0.002	0.002	0.04	...	0.002	...

- A lista (infinita) de probabilidades deve ter valores ≥ 0 e eles devem somar 1. Isto é, $1 = \sum_{k=0}^{\infty} p(k)$.

V.A.s discretas - exemplos

- Pergunta-se a uma amostra de indivíduos (as instâncias) qual é a sua religião: católica, protestante, sem religião, outras religiões cristãs, espírita, outras.
- São seis categorias possíveis para cada resposta, claramente não numéricas e sem ordenação.
- Vamos representar esta coluna de dados com uma variável aleatória X.
- Como X é uma função de Ω para $\mathbb{R}$, arbitrariamente nós vamos associar um número a cada categoria da resposta.

V.A.s discretas - exemplos

- Seja $X(\omega)$ uma variável aleatória que, para cada indivíduo ω da população, associe um número da seguinte forma:

$$X(\omega) = \begin{cases} 1, & \text{se } \omega \text{ é católico} \\ 2, & \text{se } \omega \text{ é protestante} \\ 3, & \text{se } \omega \text{ não tem religião} \\ 4, & \text{se } \omega \text{ é de outras religiões cristãs} \\ 5, & \text{se } \omega \text{ é espírita} \\ 6, & \text{se } \omega \text{ é de alguma outra religião} \end{cases}$$

- A associação entre as categorias e os números correspondentes écompletamente arbitrária.
- Qualquer outra associação seria válida.

V.A.s discretas - exemplos

- Por exemplo, poderíamos ter definido:

$$X(\omega) = \begin{cases} -2, & \text{se } \omega \text{ é católico} \\ -1, & \text{se } \omega \text{ é protestante} \\ 0, & \text{se } \omega \text{ não tem religião} \\ 1, & \text{se } \omega \text{ é de outras religiões cristãs} \\ 5, & \text{se } \omega \text{ é espírita} \\ 999, & \text{se } \omega \text{ é de alguma outra religião} \end{cases}$$

- Na prática, com estes atributos não-numéricos, os valores da variável aleatória serão usados apenas como um rótulo (numérico) para a categoria.

V.A.s discretas - exemplos

- Vamos voltar a especificação anterior,em que $X(\omega) \in \{1, \ldots, 6\}$.
- Para completar a especificação da v.a., precisamos também declarar as probabilidades associadas com cada categoria de religião (ou cada valor possível da v.a.).
- Por exemplo, usando os dados do IBGE, na década de 80, ao escolher um indivíduo ao acaso da população brasileira, temos as seguintes probabilidades:

Val. pos. k	1 (cat)	2 (pro)	3 (s.rel)	4 (out. cr.)	5 (esp)	6 (out)
Probab $p(k)$	0.75	0.15	0.07	0.01	0.01	0.01

V.A. - Ω e $\mathbb{R}$

- A atribuição de probabilidade a cada valor possível de uma v.a. X é consequência das probabilidades definidas no espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$.
- Por exemplo, lançe uma moeda 6 vezes, com C = cara e $\tilde{C}$ = coroa.

V.A. - Ω e $\mathbb{R}$

- A atribuição de probabilidade a cada valor possível de uma v.a. X é consequência das probabilidades definidas no espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$.
- Por exemplo, lançe uma moeda 6 vezes, com C = cara e $\tilde{C}$ = coroa.
- $\Omega = \{CCCCCC, \tilde{C}CCCCC, C\tilde{C}CCCC, \ldots, \tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}\}$

V.A. - Ω e $\mathbb{R}$

- A atribuição de probabilidade a cada valor possível de uma v.a. X é consequência das probabilidades definidas no espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$.
- Por exemplo, lançe uma moeda 6 vezes, com C = cara e $\tilde{C}$ = coroa.
- $\Omega = \{CCCCCC, \tilde{C}CCCCC, C\tilde{C}CCCC, \ldots, \tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}\}$
- Ω possui 36 elementos e $\mathbb{P}(\omega) = 1/36$.

V.A. - Ω e $\mathbb{R}$

- A atribuição de probabilidade a cada valor possível de uma v.a. X é consequência das probabilidades definidas no espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$.
- Por exemplo, lançe uma moeda 6 vezes, com C = cara e $\tilde{C}$ = coroa.
- $\Omega = \{CCCCCC, \tilde{C}CCCCC, C\tilde{C}CCCC, \ldots, \tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}\}$
- Ω possui 36 elementos e $\mathbb{P}(\omega) = 1/36$.
- Se não estivermos interessados na ordem em que os resultados aparecem, mas apenas no número total de caras, podemos focar apenas numa versão reduzida do espaço de probabilidade.

V.A. - Ω e $\mathbb{R}$

- A atribuição de probabilidade a cada valor possível de uma v.a. X é consequência das probabilidades definidas no espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$.
- Por exemplo, lançe uma moeda 6 vezes, com C = cara e $\tilde{C}$ = coroa.
- $\Omega = \{CCCCCC, \tilde{C}CCCCC, C\tilde{C}CCCC, \ldots, \tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}\}$
- Ω possui 36 elementos e $\mathbb{P}(\omega) = 1/36$.
- Se não estivermos interessados na ordem em que os resultados aparecem, mas apenas no número total de caras, podemos focar apenas numa versão reduzida do espaço de probabilidade.
- Definimos $X(\omega)$ como sendo o número de C's em ω

V.A. - Ω e $\mathbb{R}$

- Formalmente,

$$\begin{aligned} X : \Omega &\rightarrow \mathbb{R} \\ \omega &\rightarrow X(\omega) = \text{ número de } C\text{'s em } \omega \end{aligned}$$

V.A. - Ω e $\mathbb{R}$

- Formalmente,

$$\begin{aligned} X : \Omega &\rightarrow \mathbb{R} \\ \omega &\rightarrow X(\omega) = \text{ número de } C\text{'s em } \omega \end{aligned}$$

- Portanto, $X(\omega) \in \{0, 1, \ldots, 6\} \subset \mathbb{R}$.
- Estes são os valores possíveis da v.a. X.
- Cada um desses valores possíveis possui uma probabilidade que é induzida pelo espaço de probabilidade original $(\Omega, \mathcal{A}, \mathbb{P})$.

V.A. - Ω e $\mathbb{R}$

- Uma proposição acerca do valor em $\mathbb{R}$ de uma v.a. determina um evento A em Ω. NOTAÇÃO:

$$[X = 6] = \{\omega \in \Omega : X(\omega) = 6\} = \{\omega = (CCCCCC)\}$$

V.A. - Ω e $\mathbb{R}$

- Uma proposição acerca do valor em $\mathbb{R}$ de uma v.a. determina um evento A em Ω. NOTAÇÃO:

$$[X = 6] = \{\omega \in \Omega : X(\omega) = 6\} = \{\omega = (CCCCCC)\}$$

$$[X = 5] = \{\omega \in \Omega : X(\omega) = 5\} = \{(\tilde{C}CCCCC), (C\tilde{C}CCCC), \ldots, (CCCCC\tilde{C})\}$$

V.A. - Ω e $\mathbb{R}$

- Uma proposição acerca do valor em $\mathbb{R}$ de uma v.a. determina um evento A em Ω. NOTAÇÃO:

$$[X = 6] = \{\omega \in \Omega : X(\omega) = 6\} = \{\omega = (CCCCCC)\}$$

$$[X = 5] = \{\omega \in \Omega : X(\omega) = 5\} = \{(\tilde{C}CCCCC), (C\tilde{C}CCCC), \ldots, (CCCCC\tilde{C})\}$$

$$[X = 1] = \{\omega \in \Omega : X(\omega) = 1\} = \{(C\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}), (\tilde{C}C\tilde{C}\tilde{C}\tilde{C}\tilde{C}), \ldots, (\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}C)\}$$

V.A. - Ω e $\mathbb{R}$

- Uma proposição acerca do valor em $\mathbb{R}$ de uma v.a. determina um evento A em Ω. NOTAÇÃO:

$$[X = 6] = \{\omega \in \Omega : X(\omega) = 6\} = \{\omega = (CCCCCC)\}$$

$$[X = 5] = \{\omega \in \Omega : X(\omega) = 5\} = \{(\tilde{C}CCCCC), (C\tilde{C}CCCC), \ldots, (CCCCC\tilde{C})\}$$

$$[X = 1] = \{\omega \in \Omega : X(\omega) = 1\} = \{(C\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}), (\tilde{C}C\tilde{C}\tilde{C}\tilde{C}\tilde{C}), \ldots, (\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}C)\}$$

$$[X = 0] = \{\omega \in \Omega : X(\omega) = 0\} = \{\omega = (\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C})\}$$

V.A. - Ω e $\mathbb{R}$

- Uma proposição acerca do valor em $\mathbb{R}$ de uma v.a. determina um evento A em Ω. NOTAÇÃO:

$$[X = 6] = \{\omega \in \Omega : X(\omega) = 6\} = \{\omega = (CCCCCC)\}$$

$$[X = 5] = \{\omega \in \Omega : X(\omega) = 5\} = \{(\tilde{C}CCCCC), (C\tilde{C}CCCC), \ldots, (CCCCC\tilde{C})\}$$

$$[X = 1] = \{\omega \in \Omega : X(\omega) = 1\} = \{(C\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}), (\tilde{C}C\tilde{C}\tilde{C}\tilde{C}\tilde{C}), \ldots, (\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}C)\}$$

$$[X = 0] = \{\omega \in \Omega : X(\omega) = 0\} = \{\omega = (\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C})\}$$

- Mais notação:

$$[X \geq 5] = \{\omega \in \Omega : X(\omega) \geq 5\} = \{(CCCCCC), (\tilde{C}CCCCC), (C\tilde{C}CCCC), \ldots, (CCCCC\tilde{C})\}$$

V.A. - Ω e $\mathbb{R}$

- Uma proposição acerca do valor em $\mathbb{R}$ de uma v.a. determina um evento A em Ω. NOTAÇÃO:

$$[X = 6] = \{\omega \in \Omega : X(\omega) = 6\} = \{\omega = (CCCCCC)\}$$

$$[X = 5] = \{\omega \in \Omega : X(\omega) = 5\} = \{(\tilde{C}CCCCC), (C\tilde{C}CCCC), \ldots, (CCCCC\tilde{C})\}$$

$$[X = 1] = \{\omega \in \Omega : X(\omega) = 1\} = \{(C\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}), (\tilde{C}C\tilde{C}\tilde{C}\tilde{C}\tilde{C}), \ldots, (\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}C)\}$$

$$[X = 0] = \{\omega \in \Omega : X(\omega) = 0\} = \{\omega = (\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C})\}$$

- Mais notação:

$$[X \geq 5] = \{\omega \in \Omega : X(\omega) \geq 5\} = \{(CCCCCC), (\tilde{C}CCCCC), (C\tilde{C}CCCC), \ldots, (CCCCC\tilde{C})\}$$

ou então

$$[X \leq 1] = \{\omega \in \Omega : X(\omega) \leq 1\} = \{(\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}), (C\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}), (\tilde{C}C\tilde{C}\tilde{C}\tilde{C}\tilde{C}), \ldots, (\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}C)\}$$

V.A. - Ω e $\mathbb{R}$

- Uma proposição acerca do valor em $\mathbb{R}$ de uma v.a. determina um evento A em Ω. NOTAÇÃO:

$$[X = 6] = \{\omega \in \Omega : X(\omega) = 6\} = \{\omega = (CCCCCC)\}$$

$$[X = 5] = \{\omega \in \Omega : X(\omega) = 5\} = \{(\tilde{C}CCCCC), (C\tilde{C}CCCC), \ldots, (CCCCC\tilde{C})\}$$

$$[X = 1] = \{\omega \in \Omega : X(\omega) = 1\} = \{(C\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}), (\tilde{C}C\tilde{C}\tilde{C}\tilde{C}\tilde{C}), \ldots, (\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}C)\}$$

$$[X = 0] = \{\omega \in \Omega : X(\omega) = 0\} = \{\omega = (\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C})\}$$

- Mais notação:

$$[X \geq 5] = \{\omega \in \Omega : X(\omega) \geq 5\} = \{(CCCCCC), (\tilde{C}CCCCC), (C\tilde{C}CCCC), \ldots, (CCCCC\tilde{C})\}$$

ou então

$$[X \leq 1] = \{\omega \in \Omega : X(\omega) \leq 1\} = \{(\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}), (C\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}), (\tilde{C}C\tilde{C}\tilde{C}\tilde{C}\tilde{C}), \ldots, (\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}C)\}$$

- Eventos podem ser manipulados com união, interseção etc.

$$[X \leq 5 \text{ AND } X > 4] = \{\omega \in \Omega : X(\omega) \leq 5\} \cap \{\omega \in \Omega : X(\omega) > 4\} = \{(\tilde{C}CCCCC), (C\tilde{C}CCCC), \ldots, (CCCCC\tilde{C})\}$$

V.A. - Ω e $\mathbb{R}$

- Uma proposição acerca do valor em $\mathbb{R}$ de uma v.a. determina um evento A em Ω. NOTAÇÃO:

$$[X = 6] = \{\omega \in \Omega : X(\omega) = 6\} = \{\omega = (CCCCCC)\}$$

$$[X = 5] = \{\omega \in \Omega : X(\omega) = 5\} = \{(\tilde{C}CCCCC), (C\tilde{C}CCCC), \ldots, (CCCCC\tilde{C})\}$$

$$[X = 1] = \{\omega \in \Omega : X(\omega) = 1\} = \{(C\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}), (\tilde{C}C\tilde{C}\tilde{C}\tilde{C}\tilde{C}), \ldots, (\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}C)\}$$

$$[X = 0] = \{\omega \in \Omega : X(\omega) = 0\} = \{\omega = (\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C})\}$$

- Mais notação:

$$[X \geq 5] = \{\omega \in \Omega : X(\omega) \geq 5\} = \{(CCCCCC), (\tilde{C}CCCCC), (C\tilde{C}CCCC), \ldots, (CCCCC\tilde{C})\}$$

ou então

$$[X \leq 1] = \{\omega \in \Omega : X(\omega) \leq 1\} = \{(\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}), (C\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}), (\tilde{C}C\tilde{C}\tilde{C}\tilde{C}\tilde{C}), \ldots, (\tilde{C}\tilde{C}\tilde{C}\tilde{C}\tilde{C}C)\}$$

- Eventos podem ser manipulados com união, interseção etc.

$$[X \leq 5 \text{ AND } X > 4] = \{\omega \in \Omega : X(\omega) \leq 5\} \cap \{\omega \in \Omega : X(\omega) > 4\} = \{(\tilde{C}CCCCC), (C\tilde{C}CCCC), \ldots, (CCCCC\tilde{C})\}$$

Função distribuição acumulada de probabilidade

- A função distribuição acumulada de probabilidade da v.. X é a função matemática $\mathbb{F}(x)$ definida para todo $x \in \mathbb{R}$ e dada por

$$\begin{aligned} \mathbb{F}: \mathbb{R} &\rightarrow [0,1] \\ x &\rightarrow \mathbb{F}(x) = \mathbb{P}(X \leq x) \end{aligned}$$

- Esta função é simples e não possui NENHUMA informação adicional alé m daquela representada na lista de probabilidades.
- No entanto, ela é muito importante tanto na teoria quanto na prática de análise de dados.
- Por isto, vamos estudá-la com cuidado.
- Vamos começar calculando $\mathbb{F}(x)$ num caso particular.

Exemplo de $\mathbb{F}(x)$

- Suponha que temos uma v.a. aleatória discreta X com valores possíveis $\{0, 1, 2, 3\}$ e probabilidades associadas $p(k) = \mathbb{P}(X = k)$ dadas por

Valores possíveis k	1	2	3	4
Probab assoc $p(k)$	0.1	0.4	0.2	0.3

- Vamos calcular $\mathbb{F}(x) = \mathbb{P}(X \leq x)$ para alguns dos valores de x:
  - $\mathbb{F}(-1) = \mathbb{P}(X \leq -1) = 0$ pois não existe nenhuma chance de X ser menor ou igual a -1. O menor valor que X pode assumir é 1.
  - Pela razão acima, para qualquer $x < 1$, teremos $\mathbb{F}(x) = \mathbb{P}(X \leq x) = 0$.

Exemplo de $\mathbb{F}(x)$

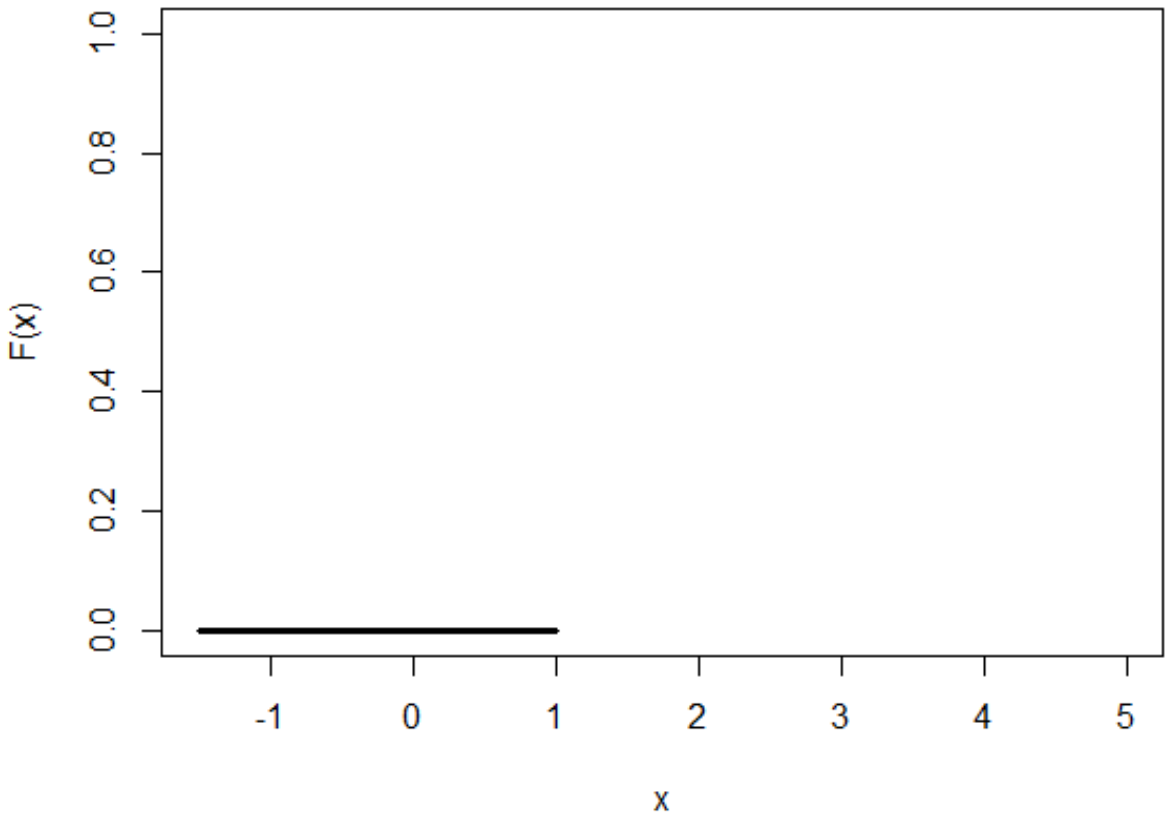


Figura: $\mathbb{F}(x) = \mathbb{P}(X \leq x)$ para $x < 1$.

Exemplo de $\mathbb{F}(x)$

- Exatamente no ponto $x = 1$, a função $\mathbb{F}(x)$ dá um salto.
- De fato, o evento $[X \leq 1]$ é idêntico a evento $[X = 1]$ já que não existe nenhum ω tal que $X(\omega) < 1$.
- Dessa forma, temos

$$\mathbb{F}(1) = \mathbb{P}(X \leq 1) = \mathbb{P}(X = 1) = 0.1$$

- Assim, a função $\mathbb{F}(x)$ salta de 0 para $x < 1$ para 0.1 no ponto $x = 1$.
- Para $x = 1.5$ temos

$$\mathbb{F}(1.5) = \mathbb{P}(X \leq 1.5) = \mathbb{P}(X = 1) = 0.1$$

- Na verdade, para qualquer x tal que $1 < x < 2$ temos $\mathbb{F}(x) = \mathbb{P}(X = 1) = 0.1$

Exemplo de $\mathbb{F}(x)$

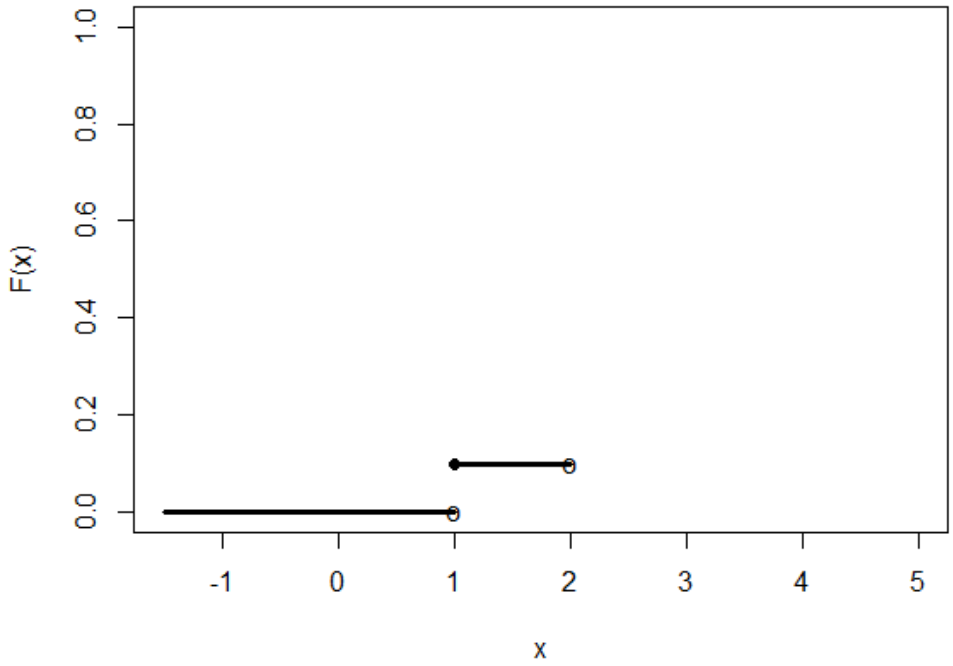


Figura: $\mathbb{F}(x) = \mathbb{P}(X \leq x)$ para $x < 2$.

Exemplo de $\mathbb{F}(x)$

- Exatamente no ponto $x = 2$, a função $\mathbb{F}(x)$ dá mais um salto.
- O evento $[X \leq 2]$ é idêntico a união de dois eventos disjuntos

$$[X = 1 \text{ ou } X = 2] = [X = 1] \cup [X = 2]$$

- Eles são disjuntos pois, pela definição de uma função matemática, não podemos ter um elemento $\omega \in \Omega$ tal que $X(\omega) = 1$ e, ao mesmo tempo, $X(\omega) = 2$.
- Assim, temos

$$\mathbb{F}(2) = \mathbb{P}(X \leq 2) = \mathbb{P}(X = 1) + \mathbb{P}(X = 2) = p(1) + p(2) = 0.1 + 0.4 = 0.5$$

- Veja que a altura do salto é igual a $p(2)$, a probabilidade $p(2) = \mathbb{P}(X = 2)$.
- Para qualquer x tal que $2 < x < 3$ temos

$$\mathbb{F}(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X = 1) + \mathbb{P}(X = 2) = p(1) + p(2) = 0.5$$

Exemplo de $\mathbb{F}(x)$

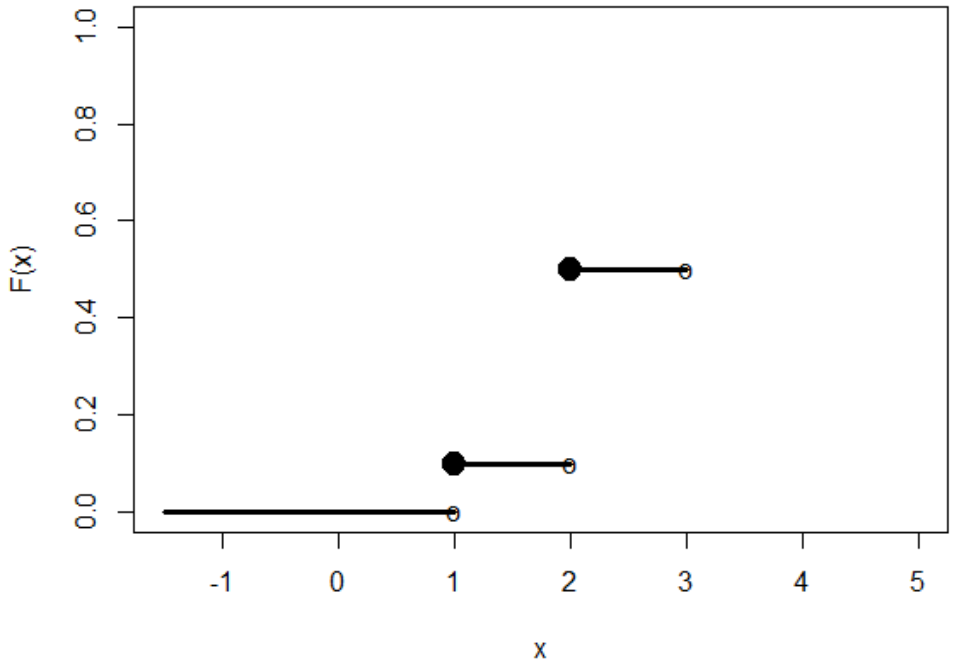


Figura: $\mathbb{F}(x) = \mathbb{P}(X \leq x)$ para $x < 3$.

Exemplo de $\mathbb{F}(x)$

- Continuando desta forma, vemos que $\mathbb{F}(x)$ vai dar saltos em $x = 3$ e e $x = 4$.
- A altura do salto em $x = k$ é igual à probabilidade $p(k) = \mathbb{P}(X = k)$.
- Quando escolhermos um valor x maior que todos os pontos possíveis de X teremos $\mathbb{F}(x) = 1$.
- Por exemplo, se $x = 4.5$, claramente teremos

$$\mathbb{F}(4.5) = \mathbb{P}(X \leq 4.5) = 1$$

  pois, com certeza, teremos $X \leq 4.5$ já que o maior valor possível de X é 4.
- O gráfico completo de $\mathbb{F}(x)$ é mostrado a seguir.

Exemplo de $\mathbb{F}(x)$

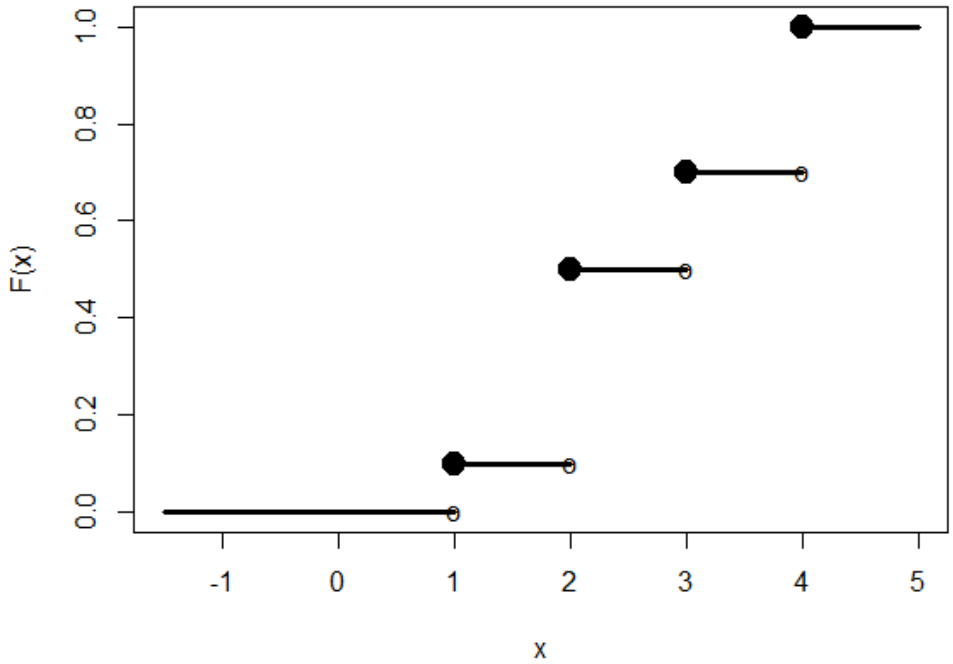


Figura: $\mathbb{F}(x) = \mathbb{P}(X \leq x)$.

Caso geral de $\mathbb{F}(x)$

- Suponha que temos uma v.a. aleatória discreta X com valores possíveis x_i e probabilidades associadas $p(x_i) = \mathbb{P}(X = x_i)$ dadas por

Valores possíveis	x_1	x_2	x_3	$\ldots$
Probab assoc	$p(x_1)$	$p(x_2)$	$p(x_3)$	$\ldots$

- A função distribuição acumulada de probabilidade é definida como:

$$\mathbb{F}(x) = \mathbb{P}(X \leq x) = \sum_{x_i \leq x} p(x_i)$$

- Isto é, $\mathbb{F}(x)$ é o valor acumulado (a soma) das probabilidades $p(x_i)$ dos pontos possíveis x_i que são menores ou iguais a x.

Esperança matemática $\mathbb{E}(X)$

- O valor esperado de uma v.a. discreta é uma soma dos seus valores possíveis ponderados pelas suas probabilidades respectivas.
- Suponha que temos uma v.a. aleatória discreta X com valores possíveis x_i e probabilidades associadas $p(x_i) = \mathbb{P}(X = x_i)$ dadas por

Valores possíveis	x_1	x_2	x_3	$\ldots$
Probab assoc	$p(x_1)$	$p(x_2)$	$p(x_3)$	$\ldots$

- Então, por definição, temos

$$\mathbb{E}(X) = \sum_i x_i p(x_i)$$

$\mathbb{E}(X)$

- $\mathbb{E}(X)$ é um valor teórico, matemático, associado com a distribuição de probabilidade da v.a X.
- Não é necessário nenhum dado estatístico para calcular $\mathbb{E}(X)$.
- Bastam as duas listas, a de valores possíveis e a de probabilidades associadas.
- $\mathbb{E}(X)$ não precisa ser igual a nenhum dos valores possíveis x_i da v.a. X.
- Qual o significado empírico deste número $\mathbb{E}(X)$? Como interpretá-lo na prática?

Interpretando $\mathbb{E}(X)$

- Suponha uma v.a. discreta X com valores possíveis x_i e probabilidades associadas $p(x_i) = \mathbb{P}(X = x_i)$.
- Temos uma enorme amostra de N instâncias independentes de X.
- Nesta amostra, x_i apareceu N_i vezes.
- Podemos estimar as probabilidades pela frequência relativa da ocorrência de x_i na amostra:

$$p(x_i) = \mathbb{P}(X = x_i) \approx \frac{N_i}{N}$$

- Assim,

$$\mathbb{E}(X) = \sum_i x_i p(x_i) \approx \sum_i x_i \frac{N_i}{N}$$

Interpretando $\mathbb{E}(X)$

- Encontramos
$$\mathbb{E}(X) = \sum_i x_i p(x_i) \approx \sum_i x_i \frac{N_i}{N}$$
- Como x_i apareceu N_i vezes na amostra, isto é o mesmo que somar todos os N valores da amostra e dividir por N.
- Isto é, se a amostra é grande, devemos ter o número teórico $\mathbb{E}(X)$ aproximadamante igual à média aritmética dos N elementos da amostra.

$\mathbb{E}(X)$

- Vamos reforçar: $\mathbb{E}(X)$ é um número real, uma constante, associada com a duas listas (valores possíveis e probabilidades associadas) que constituem uma v.a.
- $\mathbb{E}(X)$ não é, ela mesma, uma v.a.
- $\mathbb{E}(X)$ é apenas um resumo teórico da distribuição de X (ou um resumo das duas listas).
- É aproximadamente igual à média aritmética dos valores de uma grande amostra de instâncias de X.

Distribuições Especiais

- Existem infinitas distribuições de probabilidade.
- Dado um conjunto de valores possíveis, qualquer atribuição de números não-negativos que somem constiuem uma distribuição de probabilidade.
- Entretanto, algumas poucas atribuições recebem nomes especiais.
- Estas distribuições aparecem com frequência na análise de dados e são matematicamente tratáveis.
- Podemos pensar no analista de dados abordando um problema prático com um saco de distribuições de probabilidade bem conhecidas.
- Ele gostaria de não precisar inventar uma nova distribuição mas sim, de usar uma daquelas que já estão no seu embornal.
- Vamos ver algumas das mais populares agora.

Distribuição de Bernoulli

- É a distribuição mais simples: dois resultados possíveis apenas.
- $X(\omega)$ só assume dois valores possíveis: 0 ou 1.
- Isto é, $X(\omega) \in \{0, 1\}$ para todo $\omega \in \Omega$
- Definimos duas probabilidades
  - $p(0) = \mathbb{P}(X = 0) = \mathbb{P}(\omega \in \Omega : X(\omega) = 0)$
  - $p(1) = \mathbb{P}(X = 1) = \mathbb{P}(\omega \in \Omega : X(\omega) = 1)$
- Temos $p(0) + p(1) = 1$ o que implica que $p(1) = 1 - p(0)$.
- É comum escrever $p(1) = p$ e $p(0) = q$.

$\mathbb{E}(X)$ no caso Bernoulli

- Se $p(1) = p$ e $p(0) = q$, temos

$$\mathbb{E}(X) = 1 \times p + 0 \times (1 - p) = p$$

- Observe que $\mathbb{E}(X) = p$ não é igual a nenhum valor possível de X, que são apenas 0 ou 1 e tipicamente $0 < p < 1$.
- Se tivermos uma grande amostra de instâncias de X, cada uma delas igual a 0 ou 1, devemos ter $\mathbb{E}(X) = p$ aproximadamente igual a média aritmética dos valores 0 ou 1 observados.
- Mas uma média aritmética deste tipo é apenas a proporção de 1's na amostra.
- Isto é, como obviamente esperado, devemos ter

$$\mathbb{E}(X) \approx \hat{p} = \frac{1}{N} \sum_i x_i$$

Distribuição Binomial

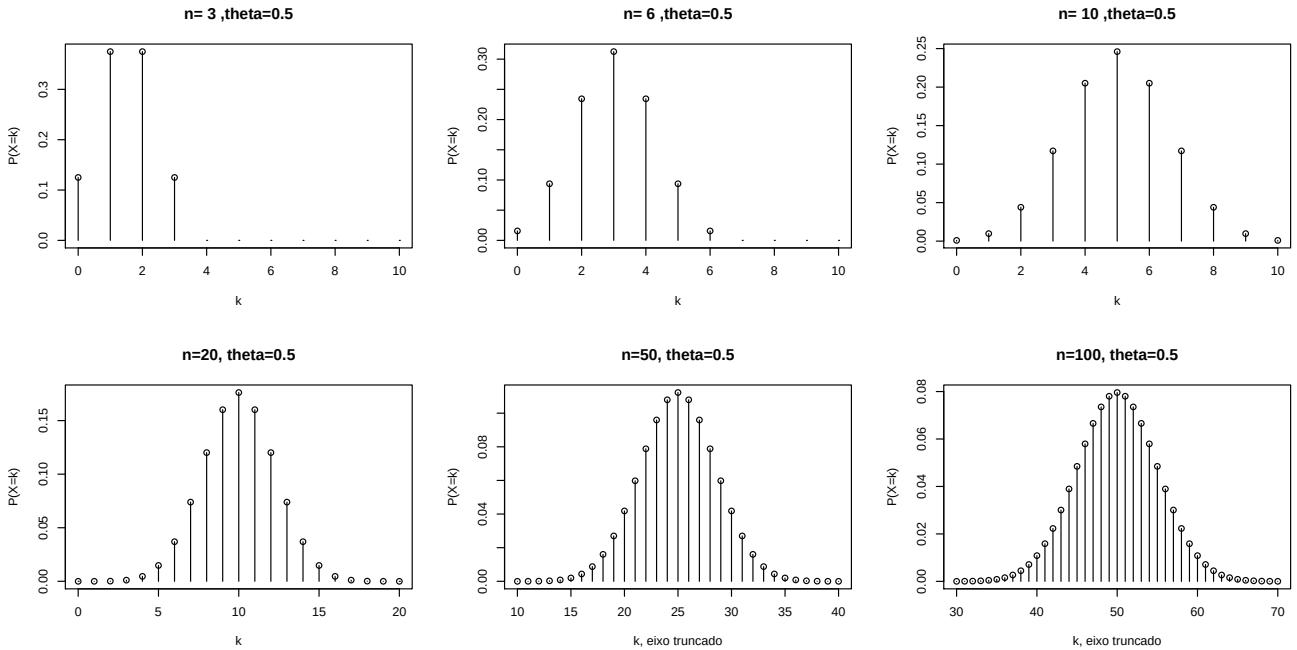
- n repetições *independentes* de um experimento binário (de Bernoulli): sucesso ou fracasso.
- Probabilidade de sucesso é constante e igual a $\theta \in [0, 1]$.
- X conta o número total de sucessos: $X \sim \text{Bin}(n, \theta)$
- Lista de valores possíveis: $0, 1, 2, \ldots, n$
- Lista de probabilidades associadas: $(1-\theta)^n, n\theta(1-\theta), \ldots, \theta^n$
- Fórmula geral:

$$\mathbb{P}(Y = k) = \frac{n!}{k!(n-k)!}\theta^k(1-\theta)^{n-k}$$

- Temos $\mathbb{E}(Y) = n\theta$ e $DP = \sqrt{\mathbb{V}(Y)} = \sqrt{n\theta(1-\theta)}$
- A forma da distribuição binomial depende de θ e de n

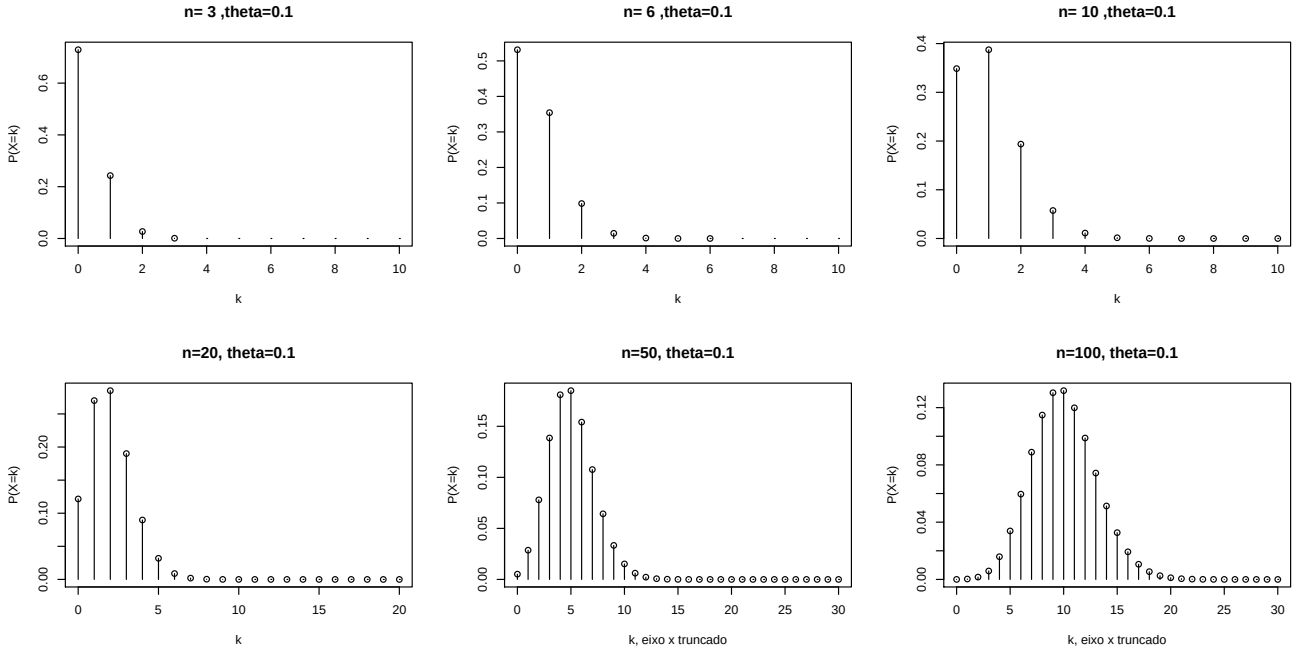
Bin(n, θ) depende de n e θ

Distribuição Binomial, $\theta = 1/2$, variando n



Bin(n, θ) depende de n e θ

Distribuição Binomial, $\theta = 0.1$, variando n



Testes de soros ou vacinas

- Doença em gado com incidência de 25%.
- Testando uma vacina, recentemente descoberta: injetamos em n animais sadios.
- Como avaliar o resultado? Para cada animal, sucesso = sadio com probab 0.75
- Para uma vacina absolutamente inócua, a probab de que k dos n animais não sejam contaminados é
  $\mathbb{P}(Y = k) = n!/(k!(n-k)!)0.75^k 0.25^{n-k}$
- Para $k = n = 10$, essa probab vale $\mathbb{P}(Y = 10) = 0.75^{10} = 0.056$.
- Para $k = n = 12$, ela vale somente $\mathbb{P}(Y = 12) = 0.75^{12} = 0.032$.
- Assim,
  - se num total de 10 ou 12 animais, nenhum é contaminado,
  - teremos uma forte indicação de que o soro teve algum efeito embora esse resultado não se constitua em prova conclusiva.

Testes de vacinas

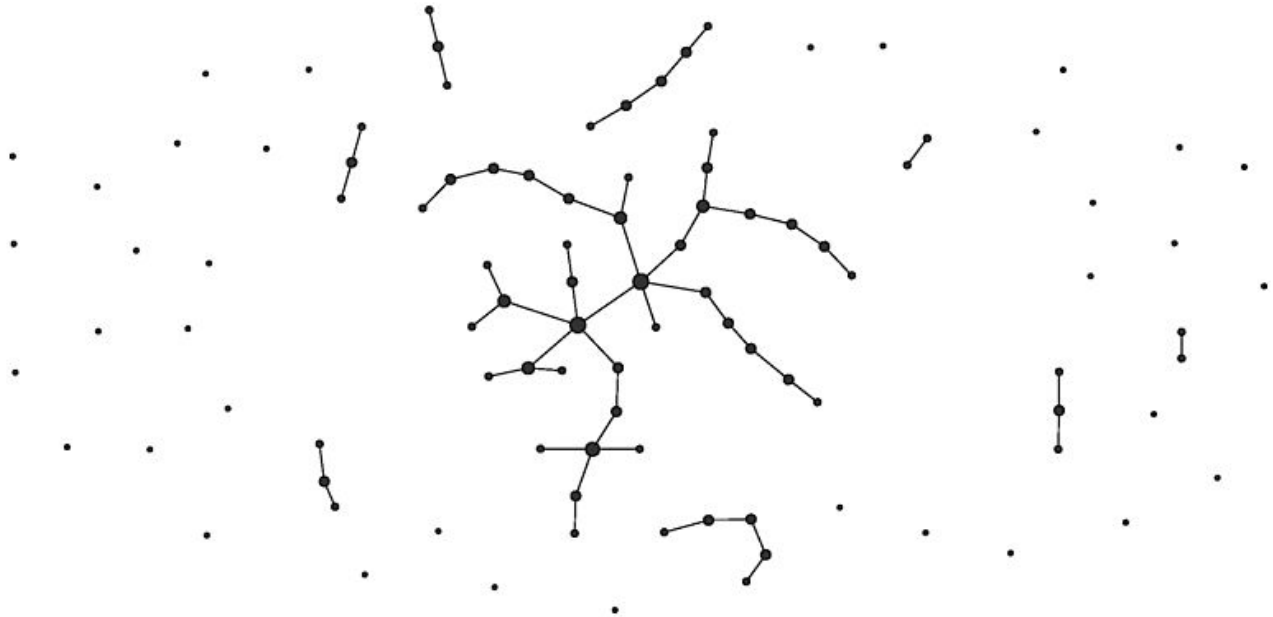
- Vimos que, com $n = 10$, tivemos $\mathbb{P}(Y = 10) = 0.056$.
- Sem a vacina, a probabilidade de que dentre 17 animais, *no máximo um* deles fique infectado é igual a $\mathbb{P}(Y \leq 1) = \mathbb{P}(Y = 0) + \mathbb{P}(Y = 1) = 0.75^{17} + 17 \times 0.75^{16} \times 0.25 = 0.0501$.
- Portanto, a evidência a favor da vacina é *mais forte* quando há 1 contaminado em 17 do que quando há 0 em 10!
- Para $n = 23$, temos $\mathbb{P}(X \leq 2) = 0.0492$.
- Assim, 2 ou menos infectados em 23 é, outra vez, uma evidência mais forte em favor da vacina, do que 1 em 17 ou 0 em 10.

Binomial em redes sociais

- Modelo de Erdös-Rényi para grafos sociais.
- n vértices formando $n(n-1)/2$ pares de possíveis arestas não-direcionadas.
- Para cada PAR de vértices, jogue uma "moeda": se cara, conecte por aresta.
- $\mathbb{P}(\text{cara}) = \theta$
- Moeda é lançada independentemente
- Fixe um vértice qualquer e seja Y o número de conexões.
- Então $Y \sim \text{Bin}(n-1, \theta)$.
- Veja que $\mathbb{E}(Y) = (n-1)\theta \approx n\theta$.

Um exemplo de Erdös-Rényi

- Um grafo gerado pelo modelo binomial de Erdós and Rényi com $\theta = 0.01$ e $n = 100$.



Alguns resultados

- Apenas como curiosidade (nosso curso não é sobre redes sociais), Erdös e Rényi provaram vários resultados matemáticos sobre os grafos aleatórios supondo que $n \to \infty$
- Considere $n\theta \approx \mathbb{E}(Y)$, o número esperado de vizinhos de um vértice qualquer.
- Se $n\theta > 1$: o grafo terá um componente gigante da ordem de n e o segundo maior componente será $\leq O(\log(n))$
- Se $n\theta > 1$: o grafo gerado quase certamente não terá um componente conectado maior que $O(\log(n))$
- Se $n\theta > (1+\epsilon)\log(n)$: o grafo quase certamente será completamente conectado
- Se $n\theta < (1-\epsilon)\log(n)$: o grafo quase certamente terá vértices isolados
- Etc, etc, etc...

Grafo segue o modelo de Erdös-Rényi?

- Como saber isto?
- O que podemos fazer?
- Uma maneira óbvia é comparar a distribuição do número de vizinhos realmente observada no grafo real com a distribuição ESPERADA sob o modelo de Erdös-Rényi.
- Como medir a distância entre o que observamos e o que esperamos?
- Temos uma resposta genérica: o teste qui-quadrado (daqui a pouco).

Distribuição Multinomial

- A multinomial é uma generalização da binomial.
- A binomial conta o número de sucessos em n repetições de um experimento binário.
- Em cada repetição temos duas categorias para classificar o resultado: sucesso ou fracasso.
- Quando tivermos mais de duas categorias em cada repetição, teremos a multinomial.
- Na multinomial, também repetimos um experimento n independentemente.
- Entretanto, em cada experimento existem k possibilidades e não apenas duas, como na binomial.
- O resultado do experimento é contar quantas vezes cada uma das k possibilidades apareceu nas n repetições.

Multinomial: o exemplo canônico

- Imagine que um dado desbalanceado é lançado n vezes.
- Em cada repetição ocorre uma "categoria": $1, 2, 3, 4, 5$ ou 6.
- As probabilidades de cada categoria são: $\theta_1, \theta_2, \ldots, \theta_6$.
- Ao fim dos n lançamentos teremos:

$$\begin{aligned} N_1 &= \text{no. lanç. na cat. } 1 \\ N_2 &= \text{no. lanç. na cat. } 2 \\ \vdots &= \vdots \\ N_6 &= \text{no. lanç. na cat. } 6 \end{aligned}$$

- Resultado é um VETOR aleatório multinomial com 6 posições contando o número de ocorrência de cada categoria. NOTAÇÃO:

$$(N_1, N_2, \ldots, N_6) \sim \mathcal{M}(n; \theta_1, \ldots, \theta_6)$$

Binomial como Multinomial

- A binomial pode ser vista como um caso simples da multinomial.
- Seja $X \sim \text{Bin}(n, \theta)$, onde X é o número de sucessos em n repetições de um experimento binário.
- De forma bastante redundante, poderíamos registrar o fenômeno aleatório na forma do número de sucesso e do número de fracassos: $(X, n - X)$.
- Este vetor é uma multinomial com duas categorias.
- Na nossa notação:

$$(X, n - X) \sim \mathcal{M}(n; \theta, 1 - \theta)$$

Multinomial: suporte

- Voltando ao caso do dado desbalanceado lançado n vezes:

$$\mathbf{N} = (N_1, N_2, \ldots, N_6) \sim \mathcal{M}(n; \theta_1, \ldots, \theta_6)$$

- Qual o suporte deste vetor aleatório $\mathbf{N}$?
- Para qualquer sequência de lançamentos, o resultado $\mathbf{N}$ será um vetor $(n_1, \ldots, n_6)$ de inteiros ≥ 0 com $n_1 + \ldots n_6 = n$.
- Assim, existe um número finito (mas bem grande) de valores possíveis para $\mathbf{N}$.

Multinomial: probabilidades associadas

- Quais as probabilidades associadas com os elementos do suporte?
- Vamos calcular um caso particular antes de dar a fórmula geral.
- Usando $n = 8$ lançmentos do dado, vamos calcular a probabilidade

$$\mathbb{P}(\mathbf{N} = (2, 0, 2, 1, 0, 3))$$

- Isto é, a chance de rolar o dado 8 vezes e terminar tendo a face 1 duas vezes, a face 2 nenhuma vez, a face 3 duas vezes, a face 4 uma vez, a face 5 zero vezes e a face 6 três vezes.
- Existem várias sequências ω de 8 lançamentos que levam ao resultado acima.

Multinomial: probabilidades associadas

- Por exemplo, se os 8 lançamentos forem

$$\omega = (3, 1, 6, 6, 1, 4, 6, 3)$$

teremos

$$\mathbf{N}(\omega) = (N_1(\omega), \ldots, N_6(\omega)) = (2, 0, 2, 1, 0, 3)$$

- Esta não a única sequência com estas contagens mas vamos nos concentrar nela(por hora).
- Qual é a probabilidade $\mathbb{P}(\omega)$ desta sequência de 8 lançamentos?
- Como os lançamentos são independentes teremos:

$$\begin{aligned}
\mathbb{P}(\omega = (3, 1, 6, 6, 1, 4, 6, 3)) &= \mathbb{P}(\text{sair 3 no 1o. E sair 1 no 2o. E ... sair 3 no 8o.}) \\
&= \mathbb{P}(\text{sair 3 no 1o.})\,\mathbb{P}(\text{sair 1 no 2o.}) \ldots \mathbb{P}(\text{ sair 3 no 8o. }) \\
&= \theta_3\ \theta_1\ \theta_6\ \theta_6\ \theta_1\ \theta_4\ \theta_6\ \theta_3 \\
&= \theta_1^2\ \theta_2^0\ \theta_3^2\ \theta_4^1\ \theta_5^0\ \theta_6^3
\end{aligned}$$

Multinomial: probabilidade mais geral

- Generalizando, se a sequência ω de n lançamentos tiver
  - n_1 aparições da face 1
  - n_2 aparições da face 2
  - $\vdots$
  - n_6 aparições da face 6

  teremos

$$\mathbb{P}(\omega) = \theta_1^{n_1} \quad \theta_2^{n_2} \quad \theta_3^{n_3} \quad \theta_4^{n_4} \quad \theta_5^{n_5} \quad \theta_6^{n_6}$$

Multinomial: probabilidade mais geral

- Voltando aos $n = 8$ lançamentos do dado, vamos calcular a probabilidade
$$\mathbb{P}(\mathbf{N} = (2, 0, 2, 1, 0, 3))$$
- Seja A o evento formado portodos os ω (seq. de $n = 8$ lanc.) tais que existem 2 1's, 0 2's, 2 3's, 0 4's, e 3 6's.
- Como calculamos antes, todo $\omega \in A$ terá a mesma probabilidade
$$\mathbb{P}(\omega) = \theta_1^2 \ \ \theta_2^0 \ \ \theta_3^2 \ \ \theta_4^1 \ \ \theta_5^0 \ \ \theta_6^3$$
- Assim,
$$\mathbb{P}(\mathbf{N} = (2, 0, 2, 1, 0, 3)) = \mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\omega) = C \times \theta_1^2 \ \ \theta_2^0 \ \ \theta_3^2 \ \ \theta_4^1 \ \ \theta_5^0 \ \ \theta_6^3$$
onde C é o número de sequências de tamanho 8 onde colocamos um elemento de $\{1, 2, \ldots, 6\}$ em cada posição e em que temos exatamente 2 1's, 0 2's, ..., 3 6's.

Multinomial: probabilidade mais geral

- Este número de possibilidades é igual a

$$\binom{8}{2,0,2,1,0,3} = \frac{8!}{2!0!2!1!0!3!} = 1680$$

- É o número de permutações distintas do vetor

$$\omega = (3,1,6,6,1,4,6,3)$$

- Generalizando para qualquer n e k catgorias, se

$$\mathbf{N} = (N_1, N_2, \ldots, N_k) \sim \mathcal{M}(n; \theta_1, \ldots, \theta_k)$$

então

$$\mathbb{P}(\mathbf{N} = (n_1, n_2, \ldots, n_k)) = \frac{n!}{n_1! n_2! \ldots n_k!} \theta_1^{n_1} \quad \theta_2^{n_2} \ldots \theta_k^{n_k}$$

Exemplos de Multinomial

- Suponha que uma amostra de $n = 22343$ indivíduos escolhidos independentemente da população brasileira e classificados em $k = 6$ categorias de religião.

Categorias i	Católica	Protestante	Sem Relig	Espírita	Outras Crist.	Outras
θ_i	0.75	0.15	0.07	0.01	0.01	0.01
N_i	16692	3398	1568	241	221	223

- As contagens aleatórias do número de pessoas em cada categoria seguem uma distribuição multinomial

$$\mathbf{N} = (N_1, N_2, \ldots, N_6) \sim \mathcal{M}(22343; (0.75, 0.15, 0.07, 0.01, 0.01, 0.01))$$

Exemplos de Multinomial

- Suponha que uma amostra de $n = 538$ indivíduos escolhidos independentemente dentre pacientes com linfoma de Hodgkins (um tipo de câncer do sistema linfático) são classificados em 12 categorias de acordo com sua resposta a um certo tratamento e seu tipo histológico:

	Resposta			
Tipo Histológico	Positiva	Parcial	Sem Resposta	Total
LP	74	18	12	104
NS	68	16	12	96
MC	154	54	58	266
LD	18	10	44	72
Total	314	98	126	538

- As contagens aleatórias do número de indivíduos em cada categoria seguem uma distribuição multinomial

$$\mathbf{N} = (N_1, N_2, \ldots, N_{12}) \sim \mathcal{M}(538; (\theta_1, \ldots, \theta_{12})$$

Classificação de textos

- Imagine uma grande coleção de textos (chamada de *corpus*), tais como artigos de jornal, em que cada documento é classificado em um de 3 tópicos: *esporte*, *política* ou *outros*.
- Esta coleção é classificada manualmente exigindo uma grande quantidade de homem-hora de trabalho.
- O objetivo agora é criar uma regra de classificação automática em uma dessas 3 categorias de outros textos não considerados, tais como textos a serem escritos no futuro.
- Uma maneira de fazer isto usa a distibuição multinomial para modelar os textos.

Textos como sequência de palavras

- Considere um dos tópicos. Por exemplo, *esporte*.
- Vamos pensar num *modelo generativo* de um texto de *esporte*.
- O texto será gerado como uma sequência de palavras escolhidas ao acaso de uma lista de palavras.
- A escolha das palavras é feita independentemente uma das outras.
- Ignorando alguns detalhes práticos, vamos chamar uma lista ordenada de D palavras distintas da língua portuguesa de *vocabulário*.
- Associamos uma probabilidade θ_i à palavra i do vocabulário.

Um dado com N faces

- Podemos imaginar um "dado" com D faces, cada face representando uma das palavras distintas do vocabulário.
- A chance de escolher a palavra i é a probabilidade θ_i do "dado" produzir a face i.

palavra-face	1	2	...	D	Soma
probabilidades	θ_1	θ_2	...	θ_D	1

- O texto é gerado rolando o "dado" sucessivamente e independentemente.
- Assim, um texto de esportes poderia ser gerado a partir desse modelo produzindo, por exemplo, *gol, Neymar, jogo, gol, cheio, gol, etc.*

Modelo não é realista

- Obviamente, é muito pequena a chance deste modelo generativo gerar um texto minimamente similar a um texto real de jornal.
- A sequência de palavras a ser gerada dificilmente terá a estrutura sintática do português ou um sentido semântico.
- É surpreendente que seja útil um modelo tão simples e tão flagrantemente falso como modelo para geração de textos da realidade.
- A lista de probabilidades vai variar de tópico para tópico.
- A lista de D palavras distintas (o vocabulário) é a mesma para todos os tópicos.
- Entretanto, cada tópico vai atribuir probabilidades diferentemente.
- No tópico *esporte*, as palavras *gol*, *jogador*, *rede* terão probabilidades θ_i maiores que sob os tópicos *política* e *outros*.

Obtendo as probabilidades de cada tópico

- As probabilidades das palavras de cada tópico são obtidas a partir das frequências simples calculadas na coleção manualmente rotulada.
- Tome todos os documentos da coleção que foram classificados como *esporte*.
- Coloque todas as palavras usadas no texto num saco de palavras (modelo *bag of words*).
- Se a palavra *jogo* aparecer 523 vezes ao longo dos textos, 523 palavras *gol* serão colocadas no saco de palavras.
- Conte quantas vezes cada uma das D palavras do vocabulário aparece no saco de palavras e divida pelo númerototal de palavras dentro do saco de maneira a obter proporções que somam 1.
- Por exemplo, se a palavra *gol* aparece 1.5% das vezes dentro do saco *esporte* então $\theta_{gol} = 0.015$.
- Isto é repetido para cada tópico criando um saco de palavras diferente e portanto diferentes probabilidades θ_i.

Evitando $\theta_i = 0$

- Geralmente, ao fim desse pocesso de estimação de probabilidades, muitas palavras do vocabulário terão $\theta_i = 0$.
- A razão é que, por exemplo, a palavra *inefável* pode não ter aparecido nem uma única vez na coleção *esporte*.
- Isto indicaria que esta palavra nunca poderia aparecer no futuro num texto de *esporte*.
- Queremos evitar esta impossibilidade futura da palavra aparecer num texto de *esporte*.
- A impossibilidade é devido a $\theta_i = 0$ nestas palavras e é criada pela flutuação estatística em palavras com probabilidades pequenas.
- Uma coleção de textos de *esporte*, mesmo que bem grande, terá zero ocorrências de muitas palavras que, embora improváveis num texto de *esporte*, não são impossíveis.

Uma solução

- Uma solução simples é colocar uma cópia de cada uma das D palavras distintas do dicionário no saco de palavras, além das próprias palavras vinda da coleção de textos de *esporte*.
- Suponha que existam D palavras distintas e a coleção de textos tem um total de M palavras.
- A palavra i aparece m_i vezes na coleção, onde m_i pode zer igual a zero.
- Ao invés de estimar θ_i pela fração m_i/M, use o estimador $\hat{\theta}_i = (m_i + 1)/(M + D)$.
- Este estimador é chamado de estimador de Laplace.
- Veja que, se $m_i = 0$, teremos $\hat{\theta}_i = 1/(M + D)$, um valor bem pequeno mas estritamente maior que zero.

As probabilidades de cada tópico

- Suponha que as probabilidades das D palavras distintas do vocabulário foram estimadas usando o estimador de Laplace em cada coleção de textos, de *esporte*, de *política*, e *outros*.
- O resultado está numa tabela:

palavra	1	2	...	D	Soma
θ_{1i}, esporte	θ_{11}	θ_{12}	...	θ_{1D}	1
θ_{2i}, política	θ_{21}	θ_{22}	...	θ_{2D}	1
θ_{3i}, outros	θ_{31}	θ_{32}	...	θ_{3D}	1

- Um novo texto aparece e desejamos classificá-lo automaticamente numa das três categorias.
- Usamos a distribuição multinomial para isto.

Classificando um novo texto

- O novo texto tem M palavras ao todo, algumas repetidas várias vezes ao longo do texto:

$$\text{Novo texto } = (x_1, x_2, x_3, \ldots, x_M)$$

  one x_j é a palavra j do novo texto.
- Usando o modelo *bag of words*, qual a probabilidade deste novo texto ter sido escrito usando as probabilidades θ_{1i} do tópico *esporte*?
- Seja N_i o número aleatório de vezes que a palavra i do dicionário vaiaparecer no novo texto.
- Se o texto é de esportes e o modelo *bag of words* for válido, temos uma multinomial para estas contagens:

$$\mathbf{N} = (N_1, N_2, \ldots, N_D) \sim \mathcal{M}\left(M; (\theta_1, \ldots, \theta_D)\right)$$

A probabilidade do texto

- Suponha que as contagens efetivamente observadas no novo texto foram os inteiros n_1, n_2, etc.
- Calculamos agora a probabilidade de observarmos este novo texto DADO QUE O TÓPICO É ESPORTE:

$$\mathbb{P}\left(\mathbf{N} = (n_1, n_2, \ldots, n_D) | \text{ esporte } \right) = \frac{n!}{n_1! n_2! \ldots n_D!} \theta_{11}^{n_1} \quad \theta_{12}^{n_2} \ldots \theta_{1D}^{n_D}$$

- Fazemos o mesmo cálculo para os outros dois tópicos:

$$\mathbb{P}\left(\mathbf{N} = (n_1, n_2, \ldots, n_D) | \text{ política } \right) = \frac{n!}{n_1! n_2! \ldots n_D!} \theta_{21}^{n_1} \quad \theta_{22}^{n_2} \ldots \theta_{2D}^{n_D}$$

e

$$\mathbb{P}\left(\mathbf{N} = (n_1, n_2, \ldots, n_D) | \text{ outros } \right) = \frac{n!}{n_1! n_2! \ldots n_D!} \theta_{31}^{n_1} \quad \theta_{32}^{n_2} \ldots \theta_{3D}^{n_D}$$

- Observe que a constante multinomial envolvendo os fatoriais é a mesma nos três casos.

Evitando calcular a constante

- Podemos fixar uma das categorias-tópico como referência e comparar as probabilidades relativamente a esta categoria-base.
- Por exemplo, suponha que fixemos a categoria *esporte* e calculamos duas razões.
- A primeira delas:

$$
\begin{aligned}
r_{p.e} &= \frac{\mathbb{P}\left(\mathbf{N} = (n_1, n_2, \ldots, n_D) \middle|\ \text{política}\ \right)}{\mathbb{P}\left(\mathbf{N} = (n_1, n_2, \ldots, n_D) \middle|\ \text{esporte}\ \right)} \\
&= \left(\frac{\theta_{21}}{\theta_{11}}\right)^{n_1} \cdots \left(\frac{\theta_{2D}}{\theta_{1D}}\right)^{n_D} = \prod_{i=1}^{D} \left(\frac{\theta_{2i}}{\theta_{1i}}\right)^{n_i}
\end{aligned}
$$

- Note que a constante desapareceu.
- Note também que, se a palavra i não aparecer no novo texto (e portanto $n_i = 0$), então o fator $(\theta_{2i}/\theta_{1i})^{n_i}$ é igual a 1 e não precisa ser calculado.

Usando as razões

- Calculamos também a segunda razão:

$$r_{o.e} = \frac{\mathbb{P}\left(\mathbf{N} = (n_1, n_2, \ldots, n_D) | \text{ outro } \right)}{\mathbb{P}\left(\mathbf{N} = (n_1, n_2, \ldots, n_D) | \text{ esporte } \right)}$$

- Suponha que $r_{p.e}$ seja maior que 1. Por exemplo $r_{p.e} = 4.3$.
- Isto significa que a chance de ter estas contagens n_1, n_2, etc. (isto é, ter este novo texto) quando o tópico é *política* é 4.3 vezes maior que a mesma chance quando o tópico é *esporte*:

$$\mathbb{P}\left(\mathbf{N} = (n_1, n_2, \ldots, n_D) | \text{ política } \right) = 4.3 \;\; \mathbb{P}\left(\mathbf{N} = (n_1, n_2, \ldots, n_D) | \text{ esporte } \right)$$

- Se $r_{p.e}$ for menor que 1, o raciocínio é o oposto. Por exemplo, $r_{p.e} = 0.1$, a probabilidade do texto sob o tópico *política* é 10 vezes menor que a mesma probabilidade sob o tópico *esporte*.

Tomando decisões

- Uma regra de decisão então pode ser a seguinte:
  - Caso $\max\{r_{p.e}, r_{o.e}\} < 1$, atribua o novo texto à categoria de referência *esporte*.
  - Caso $\max\{r_{p.e}, r_{o.e}\} > 1$, atribua o novo texto à categoria-numerador que leva ao máximo das razões.
  - Caso $\max\{r_{p.e}, r_{o.e}\} = 1$, não existe evidência suficiente no novo texto para alocar a uma das categorias. Pode-se escolher ao acaso uma das categorias que compõem uma razões que é igual a 1.
- Esta regra de decisão funciona bem em vários casos mas em algumas situações ela pode (e deve) ser melhorada.
- A questão está relacionada à diferença entre $\mathbb{P}(A|B)$ e $\mathbb{P}(B|A)$.

Qual probabilidade queremos?

- Nossa regra de decisão é baseada na comparação de probabilidades como

$$\mathbb{P}\left(\mathbf{N} = (n_1, n_2, \ldots, n_D) | \text{ esporte }\right)$$

- Mas o que realmente gostaríamos de saber é o valor da probabilidade inversa:

$$\mathbb{P}\left(\text{ esporte } | \mathbf{N} = (n_1, n_2, \ldots, n_D)\right)$$

- A primeira probabilidade calcula a chance de ter o texto novo DADO que ele foi escrito na categoria *esporte*.
- A segunda probabilidade calcula a chance de que o texto seja da categoria *esporte* DADO que ele possui a configuração de palavras observada.
- Em geral, estas probabilidades não são iguais e, na verdade, podem ser bem diferentes.

Calculando a probabilidade inversa

- Podemos usar a regra de Bayes para inverter as probabilidades de interesse.
- Por exemplo,

$$\mathbb{P}\Big(\text{ esporte } \Big|\mathbf{N} = (n_1, n_2, \ldots, n_D)\Big) = \frac{\mathbb{P}\Big(\mathbf{N} = (n_1, n_2, \ldots, n_D)\Big|\text{ esporte }\Big)\ \mathbb{P}(\text{ esporte })}{\mathbb{P}(\mathbf{N} = (n_1, n_2, \ldots, n_D))}$$

e

$$\mathbb{P}\Big(\text{ política } \Big|\mathbf{N} = (n_1, n_2, \ldots, n_D)\Big) = \frac{\mathbb{P}(\mathbf{N} = (n_1, n_2, \ldots, n_D)|\text{ política })\ \mathbb{P}(\text{ política })}{\mathbb{P}(\mathbf{N} = (n_1, n_2, \ldots, n_D))}$$

- Observe que os denominadores do lado direito das duas expressões são idênticos e vão desaparecer se tomarmos as razões das probabilidades:

$$\begin{aligned}\frac{\mathbb{P}(\text{ política }|\mathbf{N} = (n_1, n_2, \ldots, n_D))}{\mathbb{P}(\text{ esporte }|\mathbf{N} = (n_1, n_2, \ldots, n_D))} &= \frac{\mathbb{P}(\mathbf{N} = (n_1, n_2, \ldots, n_D)|\text{ política })\ \mathbb{P}(\text{ política })}{\mathbb{P}(\mathbf{N} = (n_1, n_2, \ldots, n_D)|\text{ esporte })\ \mathbb{P}(\text{ esporte })} \\ &= r_{p.e}\ \frac{\mathbb{P}(\text{ política })}{\mathbb{P}(\text{ esporte })}\end{aligned}$$

Decisão a priori

- Repetindo:

$$\frac{\mathbb{P}(\text{ política } | \mathbf{N} = (n_1, n_2, \ldots, n_D))}{\mathbb{P}(\text{ esporte } | \mathbf{N} = (n_1, n_2, \ldots, n_D))} = r_{p.e} \, \frac{\mathbb{P}(\text{ política })}{\mathbb{P}(\text{ esporte })}$$

- Concluímos que a decisão deve ser baseada na razão $r_{p.e}$ calculada anteriormente MAS corrigida pelo produto da razão $\mathbb{P}(\text{ política }) / \mathbb{P}(\text{ esporte })$
- Esta razão calcula quão frequente é a apariação de um texto de *política* em relação à frequência de um texto de *esporte*.
- Por exemplo, suponha que textos de *esporte* sejam 100 vezes mais frequentes que textos de *política* de forma que $\mathbb{P}(\text{ política }) / \mathbb{P}(\text{ esporte }) = \frac{1}{100}$.

Atualizando a priori

- A priori, sem olhar o novo texto, sabemos que existem muito mais artigos de *esporte* que de *política*.
- Uma regra de decisão razoável, que decida *a priori*, sem nem olhar o texto novo, é alocar qualquer texto que surja à categoria *esporte*.
- De posse do novo texto, olhando a configuração das palavras, podemos mudar a nossa regra de decisão *a priori* alocando o texto novo a *política*.
- Mas *esporte* é tão frequente que vamos fazer isto apenas se a evidência a favor de *política* no novo texto for bem forte.
- Por exemplo, se $r_{p.e} = 1.1$, há alguma mas não muita evidência favorável a *política*.
- Afinal, isto significa que a chance de ter a configuração de palavras do texto novo quando o tópico é *política* é apenas 10% maior que a mesma chance quando o tópico é *esporte*.

Atualizando a priori

- Com *esporte* sendo 100 vezes mais frequente que *política* em geral e com o novo texto tendo $r_{p.e} = 1.1$ obtemos

$$\frac{\mathbb{P}(\text{ política } | \mathbf{N} = (n_1, n_2, \ldots, n_D))}{\mathbb{P}(\text{ esporte } | \mathbf{N} = (n_1, n_2, \ldots, n_D))} = r_{p.e} \frac{\mathbb{P}(\text{ política })}{\mathbb{P}(\text{ esporte })} = \frac{1.1}{100} = 0.011$$

- Isto é, a chance de ser de *política* continua sendo aprox. 100 vezes menor que a chance de ser de *política* mesmo sendo $r_{p.e} > 1$.
- Continuamos a atribuir o texto a *esporte*.

Um problema numérico

- Seja $(\theta_1, \ldots, \theta_D)$ um vetor onde θ_i é a probabilidade da palavra i do dicionário ser usada num texto de certo tópico (esportes, digamos).
- As probabilidades somam 1.
- Um texto específico é analisado e você obtem as contagens $n_1, \ldots, n_D$ de modo que n_i é o número de vezes que a palavra i do dicionário apareceu neste texto.
- Dado que o texto é realmente de esportes, a probabilidade de que ele tenha estas contagens é dada pelo modelo multinomial:

$$\mathbb{P}\left(\mathbf{N} = (n_1, n_2, \ldots, n_D) | \text{ esporte } \right) = \frac{n!}{n_1! n_2! \ldots n_D!} \theta_1^{n_1} \quad \theta_2^{n_2} \ldots \theta_D^{n_D}$$

- Já vimos que a constante não precisa ser calculada e portanto seu problema é obter o valor numérico da expressão

$$\theta_1^{n_1} \quad \theta_2^{n_2} \ldots \theta_D^{n_D}$$

Um problema numérico

- Queremos

$$\theta_1^{n_1} \ \ \theta_2^{n_2} \ldots \theta_D^{n_D} = \prod_{i=1}^{D} \theta_i^{n_i}$$

- O número D de palavras do vocabulário é muito grande.
- A maioria das probabilidades θ_i são números próximos de zero.
- O produto de muitas delas elevadas à potência n_i rapidamente leva ao limite de precisão numérica da máquina.
- E o produto é transformado em 0. Um exemplo ilustrativo:

```
p1 = runif(1000) # mil numeros aleatorios entre 0 e 1
# padronizando para que p1 tenha probabilidades somando 1
p1 = p1/sum(p1)
# gerando 1000 contagens entre 1 e 100 ao acaso
contagens = sample(1:100, 1000, replace=T)
# calculando theta^n
aux = (p1)^contagens
# obtendo o seu produto: retorna 0
prod(aux)
[1] 0
```

Tomando logs

- O truque para fazer este cálculo é usar logaritmos.
- Na escala log, produtos são transformados em somas e, por isto, costumam ficar muito mais estáveis numericamente.

$$\begin{aligned}
\log \mathbb{P}(\text{texto}) &= \log\left(\mathbb{P}\left(\mathbf{N} = (n_1, n_2, \ldots, n_D) | \text{ esporte }\right)\right) \\
&= \text{cte.} + \log\left(\theta_1^{n_1}\right) + \log\left(\theta_2^{n_2}\right) + \ldots + \left(\theta_D^{n_D}\right) \\
&= \text{cte.} + n_1 \log\left(\theta_1\right) + n_2 \log\left(\theta_2\right) + \ldots + n_D\left(\theta_D\right) \\
&= \text{cte.} + \sum_{i=1}^{D} n_i \log\left(\theta_i\right)
\end{aligned}$$

- Na prática, a constante pode ser ignorada (e não precisa ser calculada) pois ela será a mesma em todos os tópicos (esportes, política, etc). Em R:

```
lp1 = log(p1)
aux = contagens*lp1
sum(aux)
# sendo mais sintetico em R, em uma unica linha de comando
sum(contagens * log(p1))
```

Resolvendo o problema

- Em geral, queremos calcular a probabilidade de observar um certo texto dado que ele é de política *dividida* pela probabilidade de observar este mesmo texto dado que ele é de esportes.
- Esta razão é igual a $r_{p.e}$:

$$\begin{aligned} r_{p.e} &= \frac{\mathbb{P}\left(\mathbf{N} = (n_1, n_2, \ldots, n_D) \middle|\ \text{política}\ \right)}{\mathbb{P}\left(\mathbf{N} = (n_1, n_2, \ldots, n_D) \middle|\ \text{esporte}\ \right)} \\ &= \left(\frac{\theta_{21}}{\theta_{11}}\right)^{n_1} \cdots \left(\frac{\theta_{2D}}{\theta_{1D}}\right)^{n_D} \\ &= \prod_{i=1}^{D} \left(\frac{\theta_{2i}}{\theta_{1i}}\right)^{n_i} \end{aligned}$$

- O mesmo truque de tomar logaritmos se aplica aqui.

Resolvendo o problema

- A constante já desapareceu e usando logaritmos, temos

$$\log(r_{p.e}) = \sum_{i=1}^{D} n_i \log\left(\frac{\theta_{2i}}{\theta_{1i}}\right)^{n_i} = \sum_{i=1}^{D} n_i \left(\log(\theta_{2i}) - \log(\theta_{1i})\right)$$

- Se `p1` e `p2` sao os vetores de probabilidades dos dois tópicos em R, basta escrever
- `sum(contagens * (log(p1)-log(p2)))`
- Este valor está na escala log. Assim, $r_{p.e} > 1$ implica em $\log(r_{p.e}) > 0$ enquanto que $r_{p.e} < 1$ implica em $\log(r_{p.e}) < 0$.

Decaimento atômico

- Uma massa atômica emite partículas radioativas.
- Um contador registra o número de partículas atingindo uma placa num intervalod de 7.5 segundos
- Valores possíveis para a contagem: $0, 1, 2, \ldots, \infty$?
- Probabilidades associadas: empírico ou teórico.
- Um modelo teórico para a emissão de partículas

Um modelo teórico

- Hipótese 1: A probabilidade da chegada de k partículas num intervalo de tempo depende apenas do comprimento do intervalo
- Hipótese 2: Os néméros de partículas em intervalos de tempo disjuntos são v.a.'s independentes
- Hipótese 3: As partículas chegam sozinhas e não simultaneamente.
- Pode-se provar que um sistema estocástico com estas propriedades necessariamente terá $\mathbb{P}(Y = k) = \frac{\lambda^k}{k!} e^{-\lambda}$
- para $k = 0, 1, 2, \ldots$
- e onde λ é uma constante positiva associada com o massa radioativa.

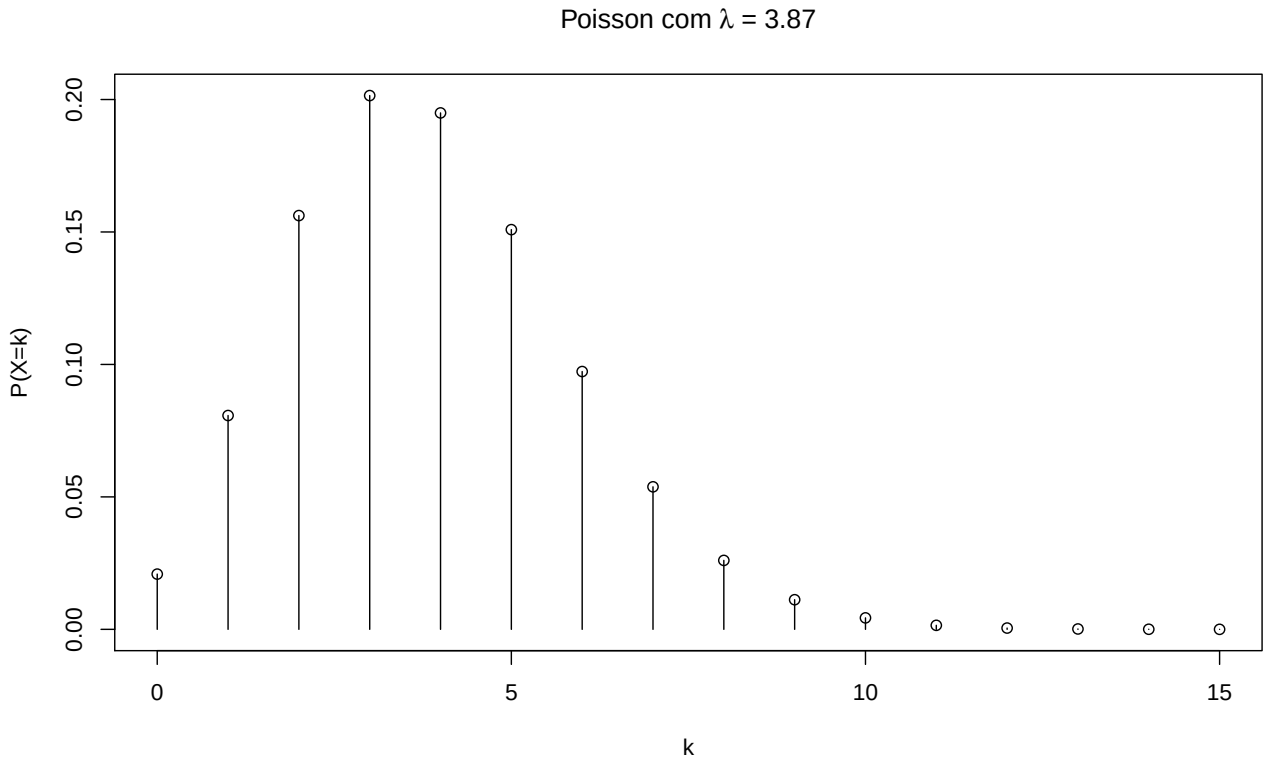
Distribuição de Poisson

- Temos uma constante $\lambda > 0$.
- Os valores possíveis são $0, 1, 2, \ldots$ e as probabilidades associadas são
- $\mathbb{P}(Y = 0) = \frac{\lambda^0}{0!} e^{-\lambda} = e^{-\lambda}$
- $\mathbb{P}(Y = 1) = \frac{\lambda^1}{1!} e^{-\lambda} = \lambda\ e^{-\lambda}$
- $\mathbb{P}(Y = 2) = \frac{\lambda^2}{2!} e^{-\lambda}$
- $\mathbb{P}(Y = 3) = \frac{\lambda^3}{3!} e^{-\lambda}$
- $\mathbb{P}(Y = 4) = \frac{\lambda^4}{4!} e^{-\lambda}$
- Etc.
- De maneira geral: $\mathbb{P}(Y = k) = \frac{\lambda^k}{k!} e^{-\lambda}$
- Pode-se provar que $\mathbb{E}(Y) = \lambda$.

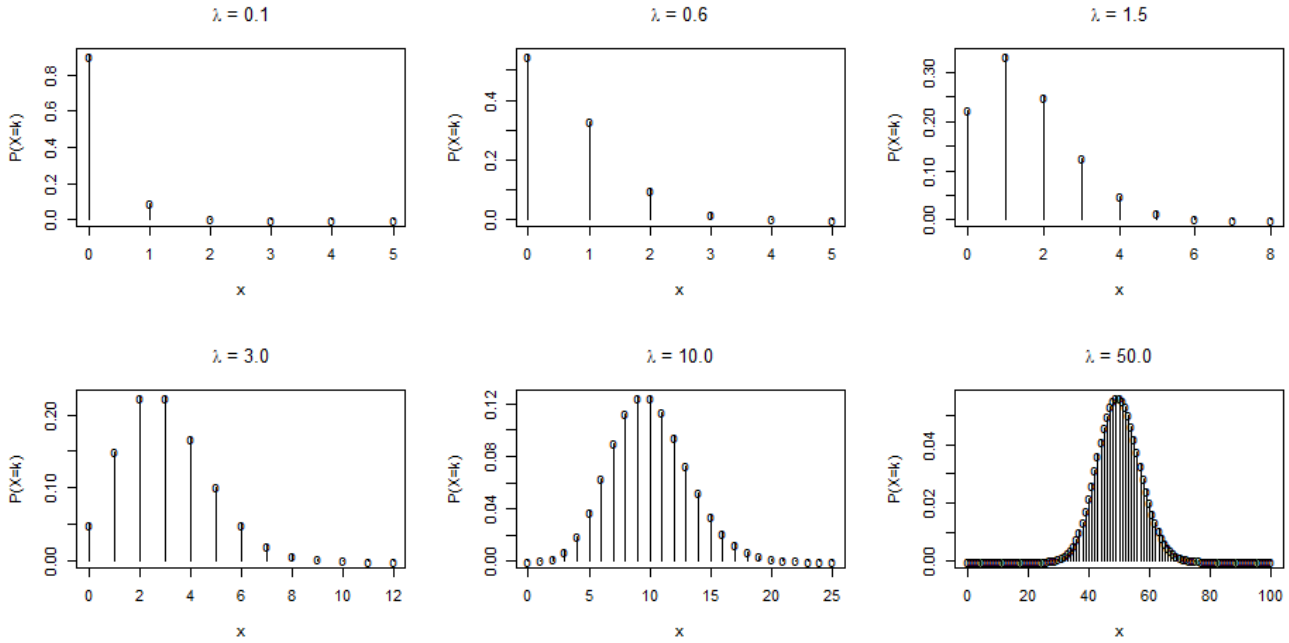
Um caso particular

- Por exemplo, se $\lambda = 3.87$, então
- $\mathbb{P}(0 \text{ partículas em 7.5 segundos}) = \mathbb{P}(Y = 0) = e^{-3.87} = 0.021$
- $\mathbb{P}(Y = 1) = 3.87 \times \ e^{-3.87} = 0.081$
- $\mathbb{P}(Y = 2) = 3.87^2/2! \times \ e^{-3.87} = 0.156$
- $\mathbb{P}(Y = 3) = 3.87^3/3! \ e^{-3.87} = 0.201$
- Etc.
- Temos $\mathbb{E}(Y) = 3.87$.

A função de probabilidade



Poisson: variando λ



Um pouco de realidade

- Esta fórmula matemática bate com a realidade?
- Rutherford, Chadwick e Ellis (1920)
- "Repetiram" o experimento um grande número de vezes: 2608 intervalos de tempo consecutivos de 7.5 segundos cada um
- Sejam $y_1 = 4, y_2 = 3, y_3 = 0, \ldots, y_{2608} = 4$ a contagem de partículas emitidas em cada intervalo.
- Vamos assumir que eles são os valores instanciados das v.a.'s i.i.d $Y_1, Y_2, \ldots, Y_{2608}$, todas com distribuição Poisson(λ).
- Se este modelo Poisson para a emissão de partículas estiver correto o que podemos esperar nas contagens observadas?
- Vamos comparar os valores teóricos $\mathbb{P}(Y = k) = \frac{\lambda^k}{k!} e^{-\lambda}$ com a frequência observada de intervalos com contagens iguais a k

Comparando os dados com o modelo Poisson

- Vamos calcular $\mathbb{P}$(emitir k partículas em 7.5 segundos) usando o modelo de Poisson.
- Isto é, vamos calcular $\mathbb{P}(Y = k) = \frac{\lambda^k}{k!} e^{-\lambda}$.
- Não usaremos os dados aqui.
- A seguir, vamos calcular a proporção dos 2608 intervalos em que obtivemos k partículas.
- Veja que não usamos nenhum modelo aqui, apenas os dados observados.
- Se o modelo estiver correto, estes dois valores devem ser parecidos para todo k.

Um último passo

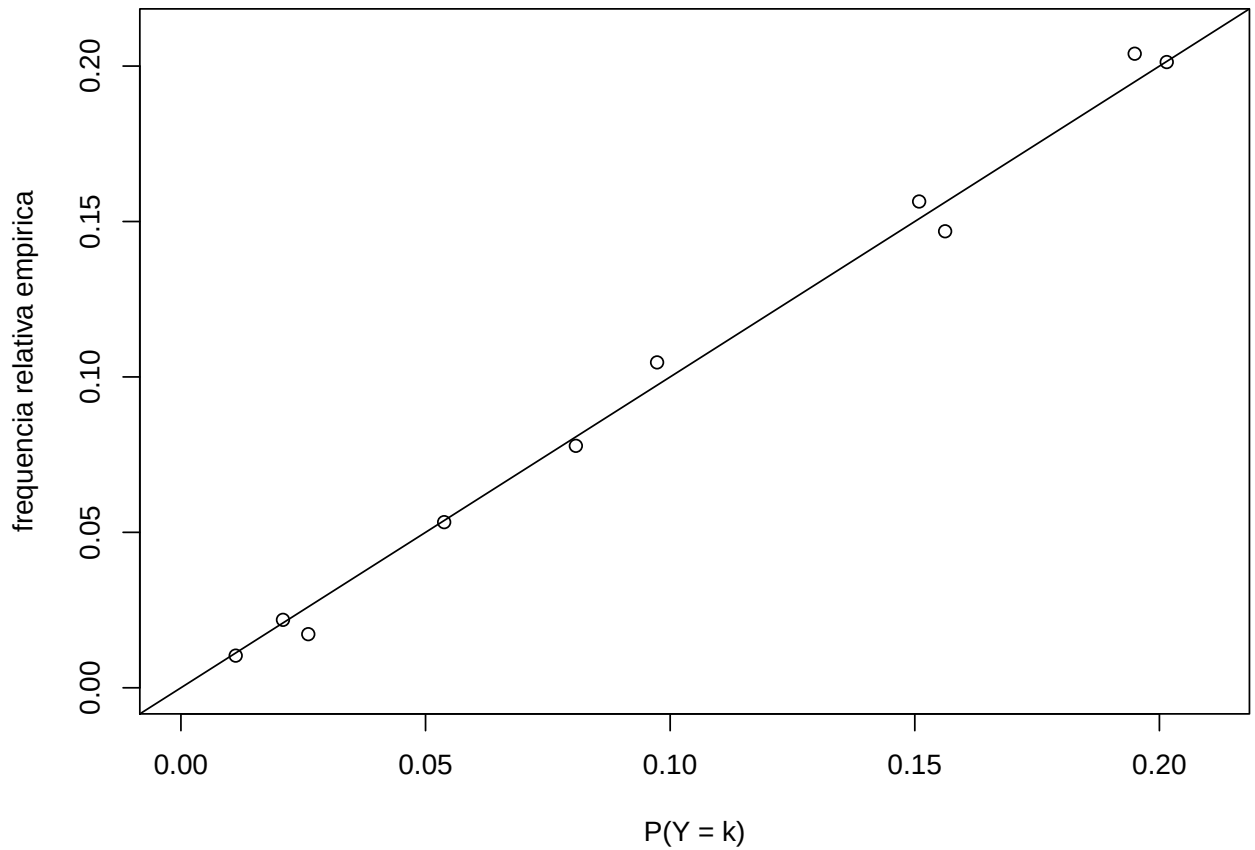
- Na verdade, para calcular $\mathbb{P}(Y = k) = \frac{\lambda^k}{k!} e^{-\lambda}$ precisamos de algum valor para λ.
- No caso Poisson, $\mathbb{E}(Y_i) = \lambda$.
- Assim, λ é o valor esperado de Y ou típico.
- Calcule a média aritmética das observções como uma aproximação para λ:
- Use $\hat{\lambda} = (y_1 + y_2 + \ldots y_{2608})/2608 = 3.87$
- Assim, vamos calcular $\mathbb{P}(Y = k) = \frac{3.87^k}{k!} e^{-3.87}$

A comparação

k	$\mathbb{P}(Y = k) = \frac{3.87^k}{k!} e^{-3.8}$	Frequência empírica no experimento
0	0.02086	$57/2608 = 0.02186$
1	0.08072	$203/2608 = 0.07784$
2	0.15619	0.14686
3	0.20149	0.20130
4	0.19495	0.20399
5	0.15089	0.15644
6	0.09732	0.10968
7	0.05381	0.05329
8	0.02603	0.01725
9	0.01119	0.01035

Tabela: Probabilidades teóricas obtidas através do modelo de Poisson e frequências empíricas obtidas através dos dados.

Comparação visual



Distribuição de Poisson

- Duas situações em que a distribuição de Poisson aparece:
- Quando contamos número de ocorrências sem um limite claro para o número máximo
  - número de colisões no tráfego de BH durante o ano.
  - número de automóveis entrando na UFMG entre 7 e 8 da manhã
  - número de consultas médicas que um cliente de um plano de saúde faz durante o ano
- Aproximação para uma binomial $X \sim \text{Bin}(n, \theta)$ com um núemro máximo possível n muito grande e θ muito pequeno:
  - número de mortos por câncer de esôfago durante o ano em BH.
  - número de apólices de uma carteira com 2 ou mais sinistros durante o ano.
  - número de sinistros de uma apólice específica durante um ano.

Bombas em Londres

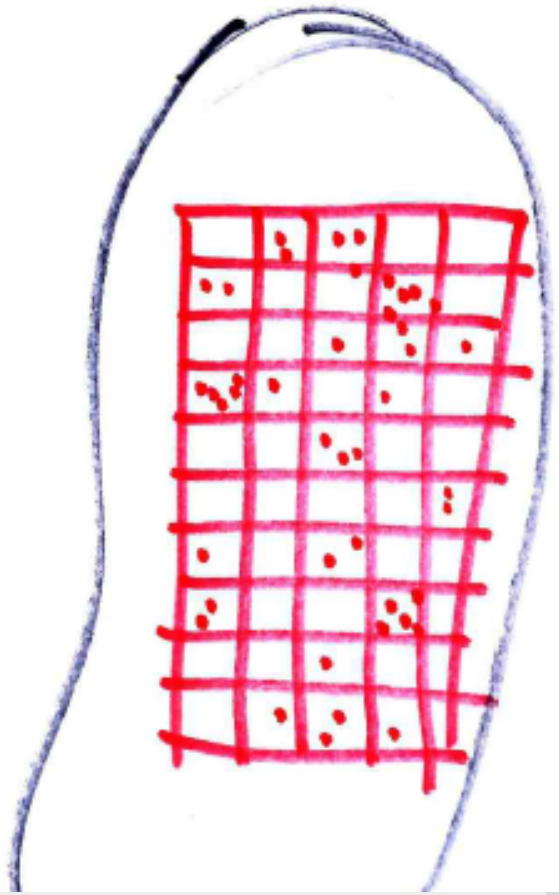
- WW2: Bombas voadoras V2 atiradas do continente europeu através do Canal da Mancha em Londres
- Segredos militares: os alemães tinham mira?
- SE OS ALEMÃES NÃO possuiam mira, o número de bombas num pequeno quadrado seria uma v.a. com distribuição Poisson(λ), mesmo λ onde quer que estivesse o quadrado
- Por quê?

Bombas em Londres

- Fixe um pequeno quadrado no mapa.
- Seja X o úmero de bombas no quadrado.
- Um grande número de bombas,
- Pequena probabilidade θ de atingir um quadrado específico.
- Bombas lançadas independentemente.
- Aproximação da binomial pela Poisson com $\lambda = n\theta$.
- SE NÃO TEM MIRA, a probabilidade θ é a MESMA para todo pequeno quadrado.
- n quadradinhos com contagens $Y_1, \ldots, Y_n$ iid Poisson(λ).
- Este modelo teórico ajusta-se aos dados? Se sim, evidência a favor da hipótese de não mira.

Bombas em Londres

- Foram mapeados os locais atingidos por bombas.
- Área foi dividida em $N = 576$ quadradinhos, $0.25 km^2$.



Bombas em Londres

- Foram mapeados os locais atingidos por bombas.
- Área foi dividida em $N = 576$ quadradinhos, $0.25 km^2$.
- Seja Y_i o número de bombas no quadradinho i.
- $Y_1, \ldots, Y_n$ são iid Poisson(λ)???
- Vamos calcular as probabilidades $\mathbb{P}(Y = 0)$, $\mathbb{P}(Y = 1)$, $\mathbb{P}(Y = 2)$, etc usando os modelo de Poisson (não vamos usar os dados aqui).
- A seguir, vamos calcular a proporção dos quadradinhos em que $Y = 0$, $Y = 1$, $Y = 2$, etc usando apenas os dados empíricos.
- Vamos então comparar as probabilidades teóricas de Poisson com as frequência baseadas apenas nos dados.
- Se forem similares, os dados são compatíveis com o modelo.

Bombas em Londres

- O número total de bombas em Londres é 537 e existem $N = 576$ quadradinhos.
- O número médio de bombas por quadradinho é $\hat{\lambda} = 537/576 = 0.9323$.
- Seja Y_i o número de bombas no quadradinho i.
- $Y_1, \ldots, Y_n$ são iid Poisson(λ) com $\lambda = 0.9323$???

Bombas em Londres

k	N_k	$N_k/576$	$\mathbb{P}(Y=k)$
0	229	0.398	0.394
1	211	0.366	0.367
2	93	0.161	0.171
3	35	0.061	0.053
4	7	0.012	0.012
≥ 5	1	0.002	0.003
Total	576	1	1

Tabela: k é o número de bombas num quadradinho, N_k é o número de quadradinhos que foram atingidos por k bombas, $N_k/576$ é a proporção de quadradinhos atingidos por k bombas e $\mathbb{P}(Y=k) = 0.9323^k/k!e^{-0.9323}$ é a probabilidade de uma v.a. Poisson($\lambda = 0.9323$) ser igual a k.

Distribuição geométrica

- Y é o número de fracassos em uma sequência de ensaios independentes de Bernoulli até que o primeiro sucesso seja observado.
- Em cada ensaio a probabilidade de sucesso é θ
- $Y = 0$ significa que o primeiro ensaio foi um sucesso
- Temos $\mathbb{P}(Y = 0) = \mathbb{P}(S) = \theta$
- $Y = 1$ significa que o primeiro ensaio foi um fracasso e osegundo foi um sucesso.
- Assim, $\mathbb{P}(Y = 2) = \mathbb{P}(FS) = (1 - \theta)\theta$

Distribuição geométrica

- $\mathbb{P}(Y = k) =$??
- $\mathbb{P}(Y = 0) = \mathbb{P}(S) = \theta$
- $\mathbb{P}(Y = 1) = \mathbb{P}(FS) = (1 - \theta)\theta$
- $\mathbb{P}(Y = 2) = \mathbb{P}(FFS) = (1 - \theta)^2\theta$
- $\mathbb{P}(Y = 3) = \mathbb{P}(FFFS) = (1 - \theta)^3\theta$
- De forma geral: $\mathbb{P}(Y = k) = (1 - \theta)^k\theta$, para $k = 0, 1, 2, \ldots$
- Pode-se mostrar que $\mathbb{E}(Y) = 1/\theta$ quando Y é geométrica com parâmetro de sucesso igual a θ.

Distribuição geométrica

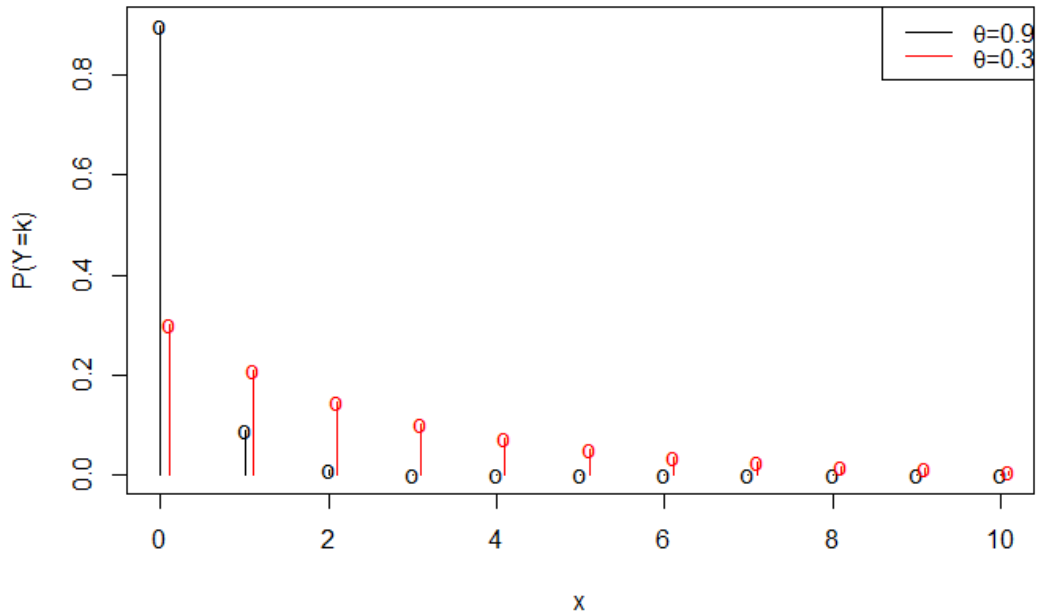


Figura:

Distribuição de Zipf ou de Pareto

- $X \in \{1, 2, 3, \ldots, N\}$
- N pode ser finito ou infinito.
- $\mathbb{P}(X = k) = \frac{C}{k^{1+\alpha}}$ com $\alpha > 0$.
- C é uma constante garantindo que as probabilidades somem 1:

$$\begin{aligned} 1 &= \mathbb{P}(X = 1) + \mathbb{P}(X = 2) + \mathbb{P}(X = 3) + \ldots \\ &= C\left(\frac{1}{1^{1+\alpha}} + \frac{1}{2^{1+\alpha}} + \frac{1}{3^{1+\alpha}} + \ldots\right. \\ &= C\sum_{k=1}^{N} \frac{1}{k^{1+\alpha}} \end{aligned}$$

o que implica que

$$C = \frac{1}{\sum_{k=1}^{N} \frac{1}{k^{1+\alpha}}}$$

- O que *realmente* importa: $\mathbb{P}(Y = k) \propto 1/k^{1+\alpha}$, inversamente proporcional a uma potência de k.

Gráfico de duas Zipf ou Pareto

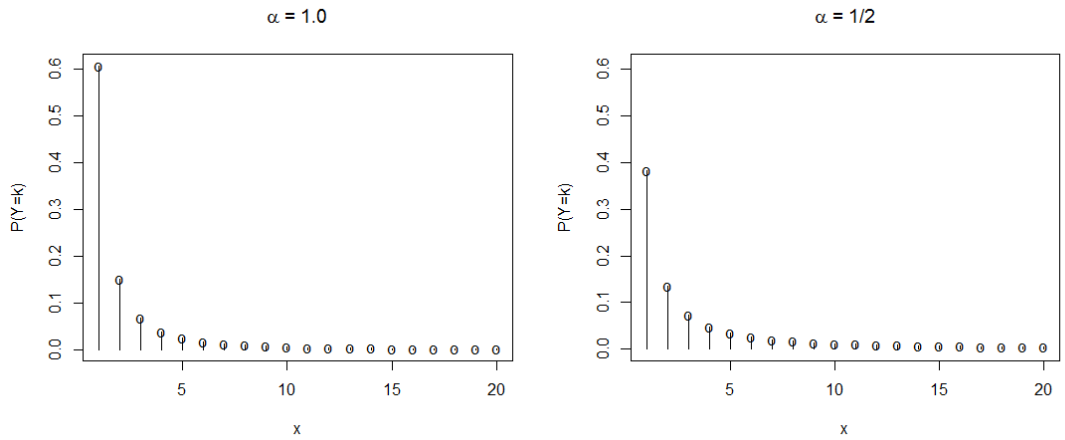


Figura: Probabilidades $\mathbb{P}(Y = k) = c/k^{1+alpha}$ de uma distribuição Pareto com $\alpha = 1$ (esquerda) e de uma Pareto com $\alpha = 1/2$ (direita). A escala do eixo vertical é a mesma nos dois casos.

Zipf clássica

- $\mathbb{P}(Y = k) \propto 1/k$ com $k = 1, 2, \ldots, N$ e N finito.
- Assim, temos
  - $\mathbb{P}(Y = 1) \propto 1$
  - $\mathbb{P}(Y = 2) \propto 1/2$
  - $\mathbb{P}(Y = 3) \propto 1/3$, etc.
- Exemplo: frequência de palavras.
- Considere um longo texto (ou vários textos num único documento).
- Algumas palavras aparecem pouco, são raramente usadas.
- Outras aparecem com grande frequência

Frequência de palavras em portugues

- Por exemplo, em portugues brasileiro temos:

palavra	posto (rank)	frequência
de	1	79607
a	2	48238
ser	27	4033
amor	802	174
chuva	2087	70
probabilidade	8901	12
iterativo	14343	6
algoritmo	21531	3

Tabela: Posto (ou rank) de algumas palavras e frequência de sua aparição por milhão de palavras em textos de português brasileiro

- Retirado de `www.linguatec.pt`.

Experimento

- Imagine o seguinte experimento: escolha uma palavra completamente ao acaso do texto.
- Seja Y o posto (ou rank) da palavra escolhida ao acaso.
- Por exemplo, se $Y = 1$, a palavra escolhida é a mais frequente.
- Se $Y = 2$, a palvra escolhida é a 2a. mais frequente.
- Se a distribuição de Zipf for um bom modelo devemos ter

$$\mathbb{P}(Y = k) \approx \frac{C}{k^{1+\alpha}}$$

com $\alpha \approx 0$.

Na escala log

- Se tivermos

$$\mathbb{P}(Y = k) \approx \frac{C}{k^{1+\alpha}},$$

ao tomarmos log dos dois lados teremos:

$$\log\left(\mathbb{P}(Y = k)\right) \approx \log(C) - \theta \log(k)$$

- Assim, um plot de $\log\left(\mathbb{P}(Y = k)\right)$ versus $\log(k)$ deveria ser aproximadamente uma linha reta com intercepto $\log(C)$ e inclinação $\theta \approx 1$.

Verificando...

- Pela visão frequentista de probabilidade, sabemos que $n_k = 10^6 \mathbb{P}(Y = k)$ é aproximadamente igual á frequência (por milhão) da palavra de posto k
- Então, se o modelo Zipf for adequado,

$$\log(n_k) \approx \log(10^6) + \log(\mathbb{P}(Y = k)) \approx \big(\log(10^6) + \log(C)\big) - \theta \log(k)$$

Português brasileiro

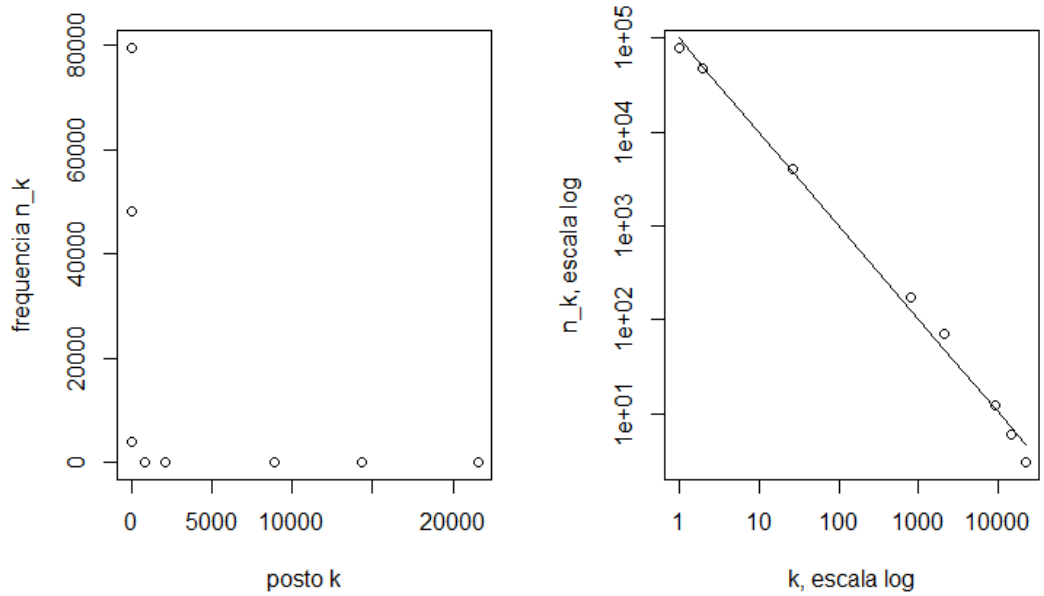


Figura: Gráfico do posto k versus a freqência n_k para algumas palavras do português (esquerda). O gráfico da direita mostra os mesmos dados num gráfico log-log (isto é, os pontos são $(\log(k), \log(n_k))$. A reta $\log(n_k) = 11.51 - 0.999 \log(k)$).

Pareto

- Zipf é um caso particular da distribuição de Pareto.
- Pareto é muito comum em estudos da web.
- Uns poucos sites possuem milhões de páginas mas milhões de sites possuem apenas umas poucas páginas.
- Poucos sites contêm milhões de links, enquanto a maioria não possui mais que uma dezena de links.
- Milhões de usuários visitam uns poucos sites dando pouca atenção a milhões de outros.

Pareto ou power-law

- A distribuição de Pareto também é chamada de lei de potência (power-law) por quê a probabilidade de k é o inverso de uma potência de k.
- Pareto dá uma probabilidade finita e apreciável a números k muito grandes, muito maiores que a imensa maioria dos valores muito frequentes.
- Nas distribuições binomial, Poisson ou geométrica, némeros muio maiores que a maioria são muito improvéveis, quase impossíveis.

Como identificar uma Pareto?

- A maneira mais efetiva de checar informalmente se a distribuição de Paretoé um bom modelo para os dados é olhar para o gráfico de $\log(k)$ versus o log da frequência $\log(n_k)$.
- Se o modelo Pareto for adequado,devemos observar aproximadamente uma linha reta neste plot.
- Entretanto, os valores muito altos e pouco frequentes de k vão gerar muito ruído na extremidade esquerda do gráfico.

Gerei 1000 Pareto($\alpha = 1$)

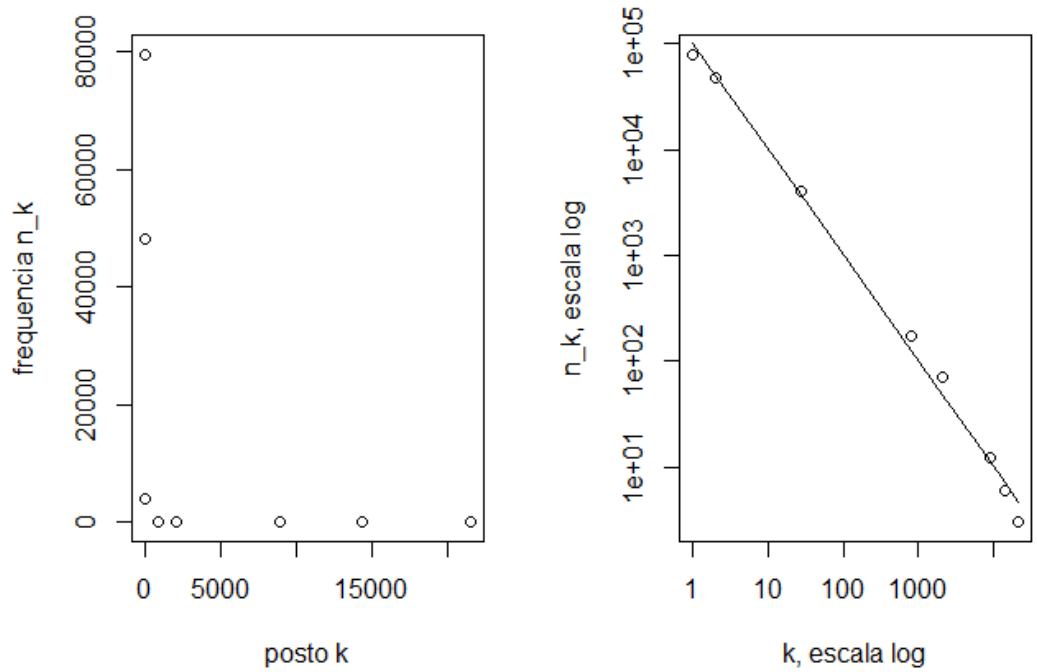


Figura: Esquerda: gráfico dos valores gerados sequencialmente. Direita: Gráfico de k versus sua frequência, escala log nos dois eixos.

Bins?

- Agregando e contando a frequência em pequenos intervalos (bins) não resolve o problema.
- Uma solução parcial 'e calcular bins logaritmicos: $(1, 2)$, $(3, 4)$, $(5, 8)$, $(9, 16)$, $(17, 32)$, etc.
- Falta figura ...

Função $\mathbb{P}(Y \geq k)$

- A melhor maneira de visualizar se a distribuição Pareto é um bom modelo é plotar a função $\mathbb{P}(Y \geq k)$ versus k, ambas na escala log.
- A razão é que, no modelo Pareto,

$$\begin{aligned} \mathbb{P}(Y \geq k) &= \mathbb{P}(Y = k) + \mathbb{P}(Y = k+1) + \ldots \\ &= \sum_{l=k}^{\infty} \frac{C}{k^{1+\alpha}} \\ &\approx \frac{C^*}{k^{\alpha}} \end{aligned}$$

- Assim, tomando log dos dois lados, temos

$$\log(\mathbb{P}(Y \geq k)) \approx \log(C^*) - \alpha \log(k)$$

- O plot de $\log(\mathbb{P}(Y \geq k))$ versus $\log(k)$ dará aproximadamente uma linha reta se o modelo for adequado.

Gráfico

- Para cada k estime $\mathbb{P}(Y \geq k)$ pela proproção de elementos da amostra que são maiores ou iguais a k.
- Podemos também simplesmente contar o número absoluto n_k de elementos da amostra que são maiores ou iguais a k.
- O plot de $\log(n_k)$ versus $\log(k)$ deveria ser aproximadamente uma linha reta.

Gerei 1000 Pareto($\alpha = 1$)

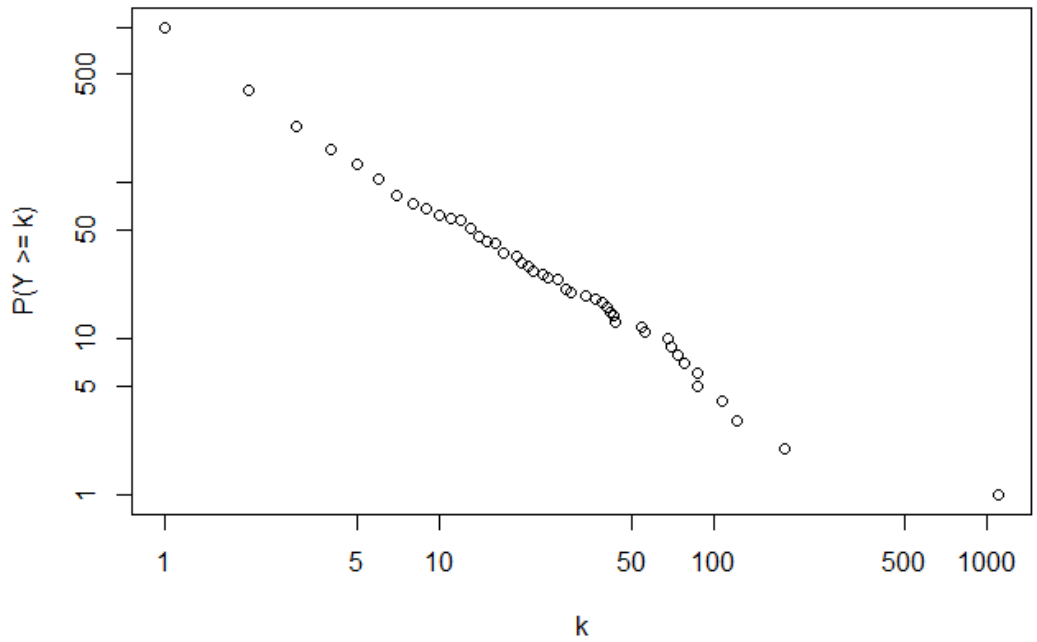


Figura: eixos.

Poisson $\times$ geométrica $\times$ Pareto

- Qual a diferença mais relevante?
- Todas são distribuições sobre os inteiros positivos
- Diferença está no comportamento *na cauda*:
  - Poisson tem cauda curta, valores com probabilidades significativas estão concentrados em uma faixa estreita torno de $\mathbb{E}(Y) = \lambda$.
  - Pareto gera facilmente valores muito grandes, ordens de grandeza maiores que $\mathbb{E}(Y)$.
  - Geométrica é um caso intermediário.

Comparando as três

- Poisson:
$$\frac{\mathbb{P}(Y=k+1)}{\mathbb{P}(Y=k)} = \frac{e^{-\lambda}\lambda^{k+1}/(k+1)!}{e^{-\lambda}\lambda^{k}/k!} = \frac{\lambda}{k+1} \to 0$$
se $k \to \infty$. Isto é $\mathbb{P}(Y=k+1) <<< \mathbb{P}(Y=k)$ se k é grande.
- Geométrica:
$$\frac{\mathbb{P}(Y=k+1)}{\mathbb{P}(Y=k)} = \frac{(1-\theta)^{k+1}\theta}{(1-\theta)^{k}\theta} = 1-\theta < 1\,,$$
constante em k. Isto é, $\mathbb{P}(Y=k+1) = (1-\theta)\mathbb{P}(Y=k)$, uma queda geométrica ou exponencial.
- Pareto:
$$\frac{\mathbb{P}(Y=k+1)}{\mathbb{P}(Y=k)} = \left(\frac{k}{k+1}\right)^{\theta} \to 1$$
Isto é $\mathbb{P}(Y=k+1) \approx \mathbb{P}(Y=k)$ se k é grande, uma queda muito lenta.