MPEI 2023-2024

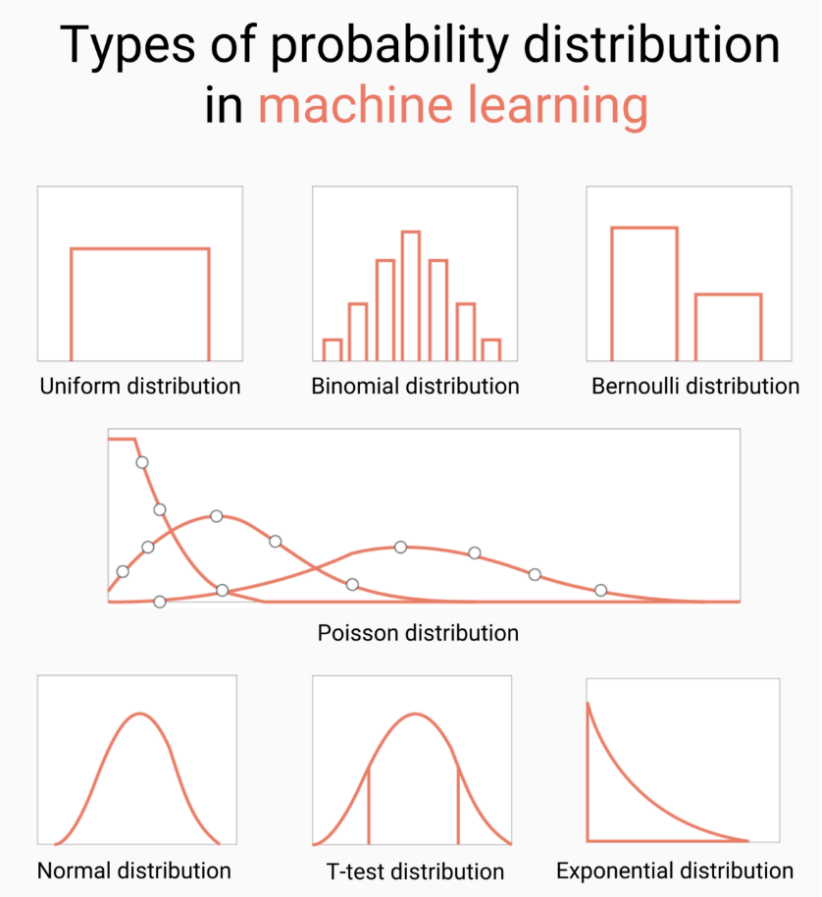
Variáveis aleatórias (continuação): Distribuições

Distribuições - Motivação

- As funções de massa de probabilidade e de densidade de probabilidade (para o caso contínuo) podem assumir as mais variadas formas
- Mas existe um conjunto de "formas" (distribuições) que aparecem repetidamente em muitos e variados problemas
  - Formam um conjunto de ferramentas base que é muito útil conhecer …

Existem muitas distribuições

- Discretas
  - **Bernoulli**
  - **Binomial**
  - **Poisson**
  - **Geométrica**
  - …
- Contínuas
  - **Uniforme**
  - **Normal**
  - **Exponencial**
  - Qui-quadrado
  - T de Student …



- Ver Wikipedia
  - https://en.wikipedia.org/wiki/List_of_probability_distributions
  - https://datasciencedojo.com/blog/types-of-statistical-distributions-in-ml/

Distribuições Discretas

Distribuição de Bernoulli

- Distribuição directamente relacionada com as experiências de Bernoulli
- Seja *A* um acontecimento relacionado com o resultado de uma experiência aleatória
- A variável de Bernoulli define-se como
- $I_A(\omega) = \begin{cases} 1 & se\ \omega \in A \\ 0 & caso\ contrário \end{cases}$

Distribuição de Bernoulli

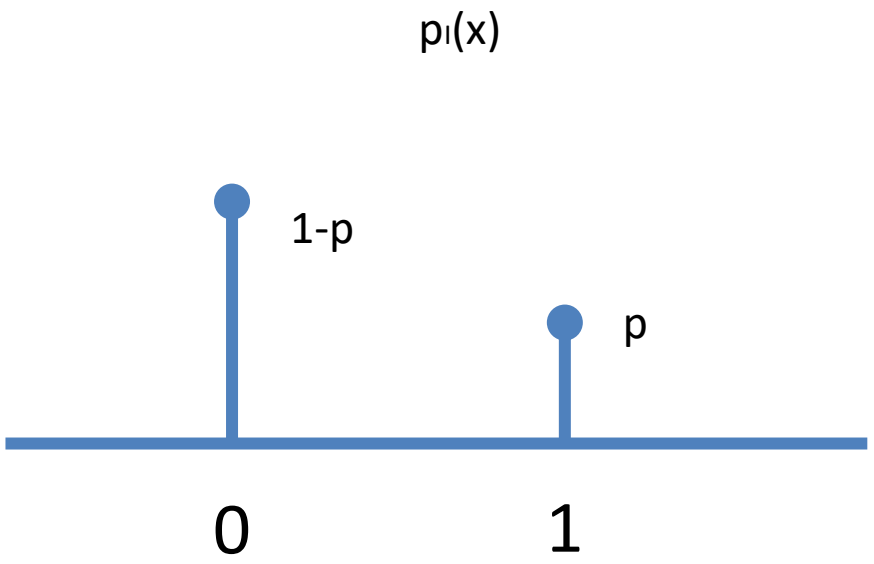
- O I que usamos para a designar resulta de ser usada muitas vezes como indicadora da ocorrência/não ocorrência de um evento

- Quando o evento ocorre a variável aleatória I assume o valor 1
  - caso contrário o valor 0

Distribuição de Bernoulli

- $S_I = \{0,1\}$
- p= Pr(A)

- $p_I(1) =$p
- $p_I(0) =$1-p



- Valor esperado E[I] ?
- Var(I) = ?

Distribuição de Bernoulli

- $\mathrm{E}[I] = \sum_i x_i p(x_i)$
$= 0 \times (1-p) + 1 \times p$
$= \boldsymbol{p}$
- $Var(I) = E[I^2] - (E[I])^2$
- $E[I^2] = 0^2 \times (1-p) + 1^2 \times p = \mathrm{p}$
- $Var(I) = p - p^2 = \boldsymbol{p\,(1-p)}$

Distribuição Binomial

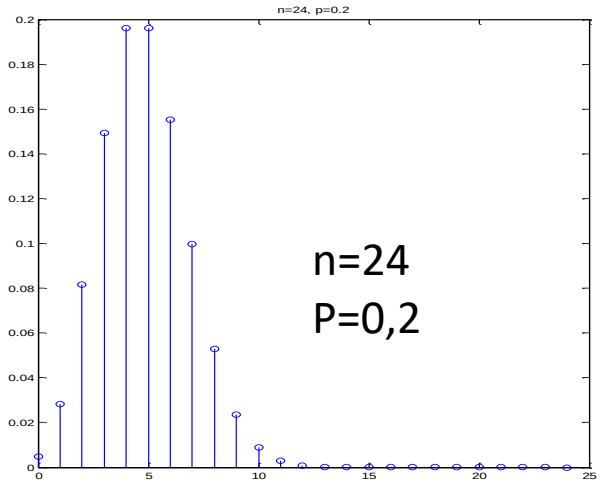
- Directamente relacionada com a Lei Binomial
- Seja X o número de vezes que um acontecimento A ocorre em n experiências de Bernoulli
  - isto é, X representa o número de sucessos em n experiências (observações)

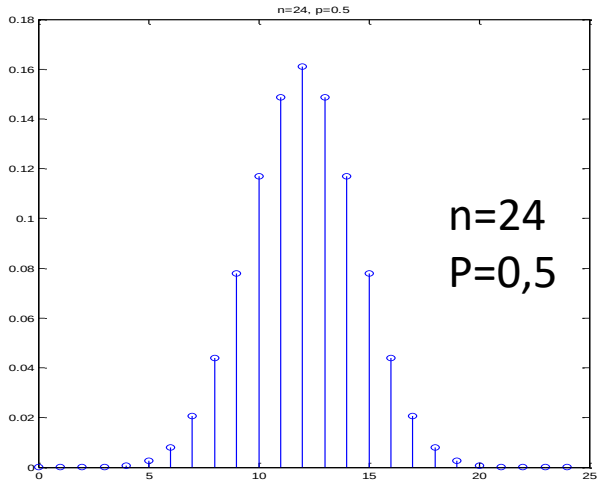
- $X = \sum_{j=1}^{n} I_j \quad \rightarrow S_X = \{0,1,2,..,n\}$

Pode ser vista como a soma de n v.a. De Bernoulli

Distribuição Binomial

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





- $F_X(x) = \sum_{k=0}^{\lfloor x \rfloor} \binom{n}{k} p^k (1-p)^{n-k}$

Distribuição Binomial – Média e Variância

- Fácil derivar usando o facto de termos **n variáveis de Bernoulli** independentes, que designamos por I_i

- $E[X] = E[\sum I_i] = \sum E[I_i] \quad pq\,?$
$= p + p + \cdots + p = \boldsymbol{n}\, p$
- De forma similar

$\text{Var}(X) =$
$Var(\sum I_i) = \sum Var(I_i) = \cdots = \boldsymbol{n}\, p\,(1-p)$

(as variáveis aleatórias I_i são independentes)

Distribuição Binomial - Exemplos

- Têm distribuição Binomial, por exemplo:
  - Número de peças defeituosas num lote de um determinado tamanho (ex: 50 peças)
  - Número de respostas certas num exame de verdadeiro falso
  - Número de clientes que efectuaram compras em 100 que entraram numa loja

Distrib. Binomial - Áreas de aplicação

- A distribuição surge em muitas áreas cientifico-tecnológicas:
  - **Informática**:
    - "acerto" e "falha" é uma interpretação possível para detetores de SPAM, testes a métodos/funções de um programa, procura de informação na web …
  - Medicina:
    - Por exemplo, os resultados "cura" ou "não cura" são importantes na indústria farmacêutica
  - Engenharia de produção:
    - Muitas vezes as medidas de controlo de qualidade são baseadas na distribuição binomial
      - aplica-se a qualquer situação industrial em que o resultado é binário e os resultados de ensaio são independentes e com probabilidades constantes
  - Indústria Militar:
    - "acerta" "falha" é muitas vezes a interpretação do lançamento de um míssil ou de uma missão

Exemplo 1: Transmissão digital

- Um sistema de transmissão digital envia um pacote de 1 kByte através de canal com ruído sendo a probabilidade de erro de cada bit 10^{-3} (ou seja 1 bit em cada mil).
- Considerando que os erros são independentes, determine:
  - Probabilidade de haver 1 erro ?
  - Probabilidade de haver erro ?

Exemplo 2 – segurança de aviões

- Considere que um motor de avião pode falhar com probabilidade p e que as falhas em motores distintos são independentes.
- Se um avião se despenha quando mais do que 50% dos motores falham, é mais seguro voar num avião de 4 motores ou de 2 motores ?
- Faz parte de um dos guiões Práticos
- Como resolver ?

- Sugestão: calcular a probabilidade de cair um avião com 2 motores, repetir para o de 4 motores e comparar os resultados (será função da probabilidade de falha de um motor)

Possível resolução

- O de 2 motores despenha-se se os 2 motores falharem. Qual a probabilidade de 2 falhas em 2 motores ?
- $p2 = p_X(2, n = 2) = \binom{2}{2} p^2 (1-p)^{2-2} = p^2$

- O de 4 despenha-se se 3 ou 4 falharem. Qual a probabilidade ?
- $p4 = p_X(3, n = 4) + p_X(4, n = 4)$
- $= \binom{4}{3} p^3 (1-p)^{4-3} + \binom{4}{4} p^4 (1-p)^{4-4}$
- $= 4\,p^3(1-p) + p^4 = 4\,p^3 - 3\,p^4$

- Relação entre $p2$ e $p4$
- $\frac{p4}{p2} = 4p - 3p^2 = p\,(4-3p)$
  - NOTE que depende de p

...

p	p2	p4	p4/p2
0,01	0,0001	0,00000397	0,0397
0,02	0,0004	0,00003152	0,0788
0,03	0,0009	0,00010557	0,1173
0,04	0,0016	0,00024832	0,1552
0,05	0,0025	0,00048125	0,1925
0,06	0,0036	0,00082512	0,2292
0,07	0,0049	0,00129997	0,2653
0,08	0,0064	0,00192512	0,3008
0,09	0,0081	0,00271917	0,3357
0,1	0,01	0,0037	0,37
0,2	0,04	0,0272	0,68
0,3	0,09	0,0837	0,93
0,4	**0,16**	**0,1792**	**1,12**
0,5	0,25	0,3125	1,25
0,6	0,36	0,4752	1,32
0,7	0,49	0,6517	1,33
0,8	0,64	0,8192	1,28
0,9	0,81	0,9477	1,17

p	p2	p4	p4/p2
0,3	0,09	0,0837	0,93
0,31	0,0961	0,091458	0,9517
0,32	0,1024	0,099615	0,9728
0,33	0,1089	0,10817	0,9933
0,34	**0,1156**	**0,117126**	**1,0132**
0,35	0,1225	0,126481	1,0325
0,36	0,1296	0,136236	1,0512
0,37	0,1369	0,146387	1,0693
0,38	0,1444	0,156934	1,0868
0,39	0,1521	0,167873	1,1037

O que significam p4/p2 < 1 ?

é mais seguro voar num avião de 4 motores ou de 2 motores ?

Exemplo de aplicação III

- According to the U.S. Census Bureau, approximately 6% of all workers in Jackson, Mississippi, are unemployed.
- In conducting a random telephone survey in Jackson, what is the probability of getting two or fewer unemployed workers in a sample of 20?

- De: Business Statistics, Ken Black, 6th ed, John Willey & Sons (cap 5)

Resolução

- 6% desempregado => $p = 0{,}06$
- Tamanho da amostra é 20 => $n = 20$
- 94% têm emprego => $1 - p = 0{,}94$
- x é o número de sucessos que se pretende

- Qual é a probabilidade de termos 2 ou menos desempregados na amostra de 20 ?

- Neste tipo de problemas o importante e muitas vezes o mais difícil é identificar o p, n e x

Resolução

$$n = 20$$
$$p = 0{,}06$$
$$q = 1 - p = 0{,}94$$
$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$$
$$= 0{,}2901 + 0{,}3703 + 0.2246 = 0{,}8850$$

$$P(X = 0) = \frac{20!}{0!(20-0)!}(0{,}06)^0(0{,}94)^{20-0} = (1)(1)(0{,}2901) = 0{,}2901$$

$$P(X = 1) = \ldots$$

$$P(X = 2) = \ldots.$$

Distribuição Geométrica

- Seja X o número de vezes que é necessário repetir uma experiência de Bernoulli até obter um sucesso
  - Prob. Sucesso: p prob. Falha = 1-p

- $p_X(k) = \boldsymbol{p}(1-p)^{k-1}\ , k = 1,2,3, \ldots$

  Porque teremos k-1 insucessos e depois sucesso

- $F_X(x) = \sum_{k=0}^{\lfloor x \rfloor} p(1-p)^{k-1}$

Exemplo de aplicação – Helpdesk UA

- Problema:
- Considere o serviço de atendimento via telefone do Helpdesk da UA.
- Supondo que a probabilidade de se conseguir contactar o suporte é p=0,1 (só ao fim de 10 tentativas ☹).
- Determine a probabilidade de necessitar de menos de 3 chamadas até conseguir expor o seu problema ?

- Solução:
- $\Pr(n^{\underline{o}}\ chamadas < 3) =$
  $\Pr(1\ chamada\ OU\ 2\ chamadas)$
- $= p(1-p)^{1-1} + p(1-p)^{2-1} = p(2-p) = 0{,}19$

Distribuição Geométrica – Média e Variância

- Demontra-se que:

- $E[X] = \frac{1}{p}$
  - Resultado de $\sum_{i=1}^{\infty} i \;\; p(1-p)^{i-1}$
  - Intuitivo: no exemplo do Helpdesk, por exemplo, quanto mais provável atenderem menos chamadas teremos que fazer (em média)

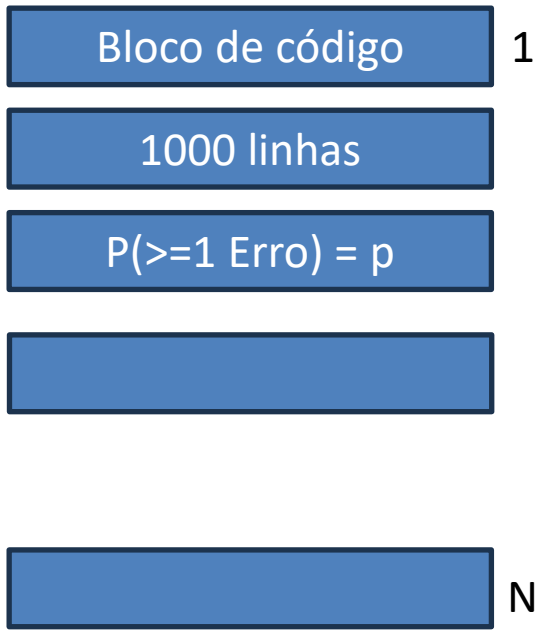
- $Var(X) = (1-p)/p^2$

Dist. Binomial para valores de n elevados

- Consideremos o seguinte cenário:

- Probabilidade de haver pelo menos um erro num conjunto de **1000 linhas** de código é **p** (p<1)
  - Não nos interessa o número de erros, apenas se existe algum ou não

- Se tivermos **N** blocos de 1000 linhas a probabilidade de **k** blocos terem erros segue a distribuição **Binomial com parâmetros N e p**



Prob(k blocos com erro) = Bin (N,p)

(continuação)

- Se analisarmos 100 linhas, e considerarmos a distribuição dos erros uniforme, a probabilidade desce para p/10
  - Teríamos então uma Binomial com parâmetros 10 N e p/10
- Teoricamente temos a forma de cálculo mas basta N ser um número moderado e 10N começa a ser elevado e os cálculos complicados [mesmo em computador]

- Exemplo: Blocos de 100 linhas; 1000 blocos ; p=0,98/10

- Qual a probabilidade do número de blocos com erro ser inferior ou igual a 100 ?

$$P = F_x(100) = \sum_{k=0}^{100} \binom{1000}{k} 0{,}098^k 0{,}902^{1000-k}$$

(continuação)

- Se reduzirmos mais o tamanho dos blocos
  - N aumenta
  - p diminui
  - As coisas ainda se complicam mais em termos de cálculo

- No limite teremos apenas um bloco de uma linha que vai ter, ou não, um erro

Distribuição de Poisson

- Em casos em que temos condições similares ao exemplo:
  - uma variável Binomial,
  - n cresce e p decresce
  - $np \to \lambda > 0$

- Para n grande pode fazer-se as seguintes aproximações:

$$\mathrm{p} \cong \frac{\lambda}{n} \qquad\qquad 1 - p \cong 1 - \frac{\lambda}{n}$$

- E calculando o limte ($n \to \infty$):

$$\lim_{n\to\infty} P(X = k) = \lim_{l\to\infty} \binom{n}{k} p^k (1-p)^{n-k} =$$

- = ... = $\frac{\lambda^k e^{-\lambda}}{k!}$

- $p_X(k) = \frac{\lambda^k e^{-\lambda}}{k!}$

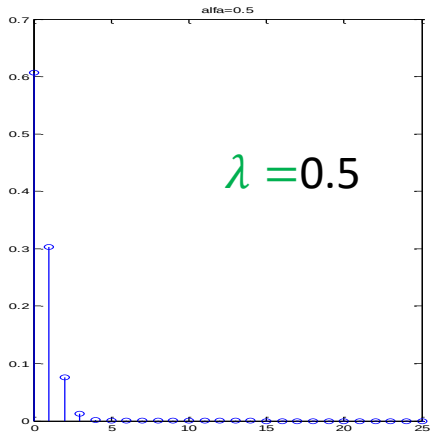
é a função de massa de probabilidade da distribuição de Poisson

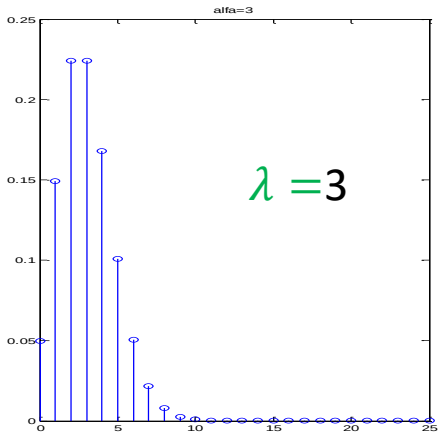
Distribuição para vários valores do parâmetro λ

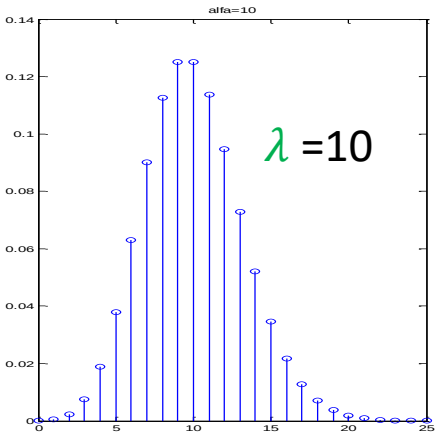
- Função de probabilidade:

$$p_X(k) = \Pr(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, k = 0,1,2,3 \ \ldots$$

- Tem apenas um parâmetro, o lambda







Distribuição de Poisson: Média e Variância

- $E[X] = \lambda$
  - Relembre que λ é aproximado por np e o valor esperado da Binomial é np

- $Var(X) = \lambda$

Distribuição de Poisson

- A distribuição de Poisson foca-se apenas no número de ocorrências (discreto) num intervalo de tempo contínuo (ou região do espaço).

- Esta distribuição não tem um número de experiências (n) como na Binomial
  - As ocorrências são independentes das outras ocorrências

Aproximação de Poisson à distribuição Binomial

- Problemas envolvendo a distribuição Binomial em que ***n* é grande e o valor de p é pequeno**, gerando desta forma **eventos raros**, são os candidatos à utilização da distribuição de Poisson

- Regra prática ("rule of thumb") :
  - Se $n>20$ e $np <=7$ a aproximação de Poisson é suficientemente próxima para ser usada em vez da Binomial

Aproximação de Poisson à distribuição Binomial

- Procedimento para aproximar a Binomial por Poisson:
  1. Calcular a média da Binomial ***μ = np***
  2. Como μ é o valor esperado da Binomial, passa a ser o **λ** (=E[X]) de Poisson
  3. Usar a fórmula de Poisson (ou uma tabela)

$$p_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Aplicações da Distribuição de Poisson

- As distribuições de Poisson surgem em experiências onde se verificam as seguintes propriedades:
  - O número de resultados que ocorrem num determinado intervalo de tempo ou região é independente do número que ocorre em qualquer outro intervalo temporal ou região espacial disjunta
  - A probabilidade que um resultado ocorra durante um intervalo ou região infinitesimal é proporcional ao comprimento do intervalo ou dimensão da região e não depende das ocorrências fora desse intervalo ou região
  - A probabilidade de haver mais que um resultado numa região infinitesimal é desprezável

Exemplo de aplicação

- Bank customers arrive randomly on weekday afternoons at an average of 3.2 customers every 4 minutes.
- What is the probability of having more than 7 customers in a 4-minute interval on a weekday afternoon?

- De: Business Statistics, Ken Black, 6th ed, John Willey & Sons (cap 5)

Resolução

- Consideremos que o número de clientes (em intervalos de 4 minutos) é representado pela variável aleatória X

- Pretendemos

  P("X > 7 clientes /4 minutos")

- λ= ?

- λ = 3,2 [nº médio de clientes em 4 minutos]

Resolução (continuação)

- A solução requer que calculemos para k = 8, 9, 10, 11, 12, 13, 14, ... até o valor ser aproximadamente zero
  - Ou usemos o complemento e calculemos k =0,1,2,3,4,5,6,7

- Depois é só somar as probabilidades

- O resultado (0,0168) mostra que é pouco provável que um banco que tem em média 3,2 clientes a cada 4 minutos receba mais de 7 clientes num período de 4 minutos
  - TPC: confirmar este valor

Aplicações

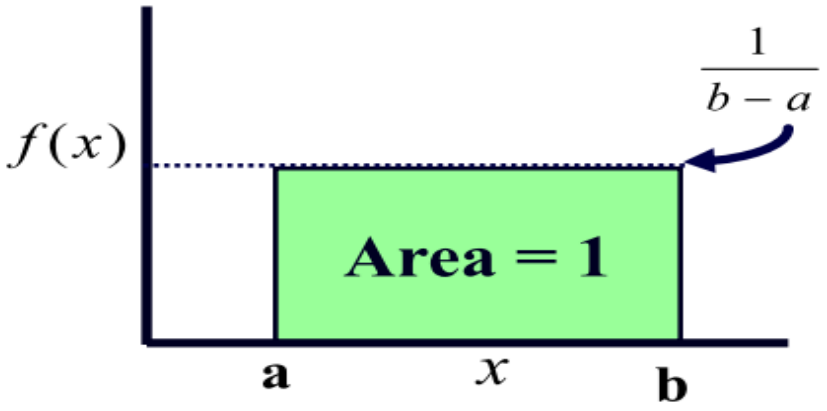
- Este tipo de probabilidades são muito úteis para os gestores de Bancos (e outras instituições com atendimento ao público) dimensionarem postos de atendimento

- A distribuição de Poisson é também muito útil na modelação da chegada de mensagens (ou outros tipos de eventos) em redes de computadores

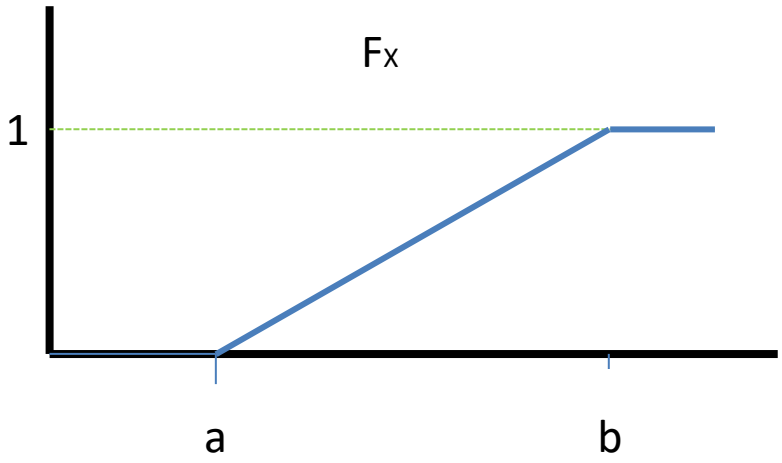
Distribuições contínuas

Distribuição uniforme

- U(a,b) é definida por:
- $f_X(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & caso\ contrário \end{cases}$



- $F_X(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \le x \le b \\ 1 & x > b \end{cases}$



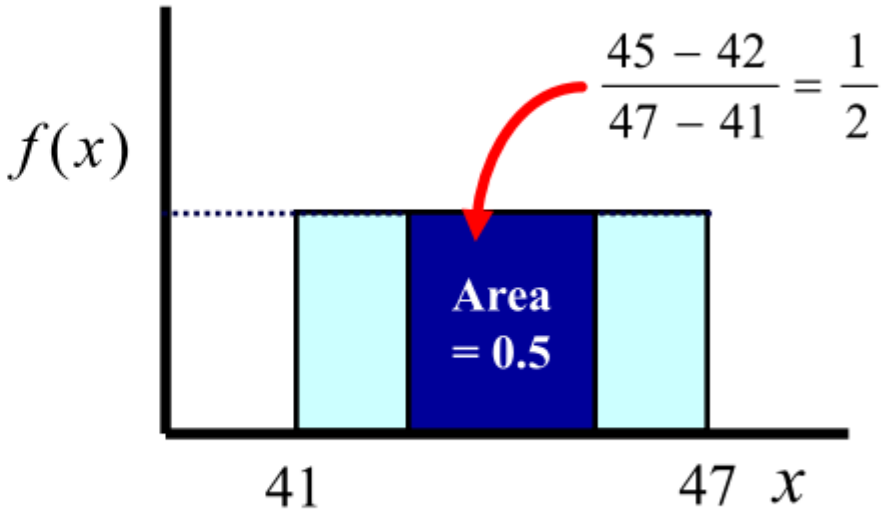
- $E[X] = \frac{a+b}{2}$
- $Var(X) = \frac{(b-a)^2}{12}$

Exemplo

- P(42<=X<=45) com U(41,47)

$$P(x_1 \le X \le x_2) = \frac{x_2 - x_1}{b - a}$$

$$P(42 \le X \le 45) = \frac{45-42}{47-41} = \frac{1}{2}$$



Função rand() do Matlab

- A função rand() do Matlab gera números obedecendo a uma distribuição uniforme

  – Com $a = 0$ e $b = 1$

- Para ter U(a,b) basta usar:

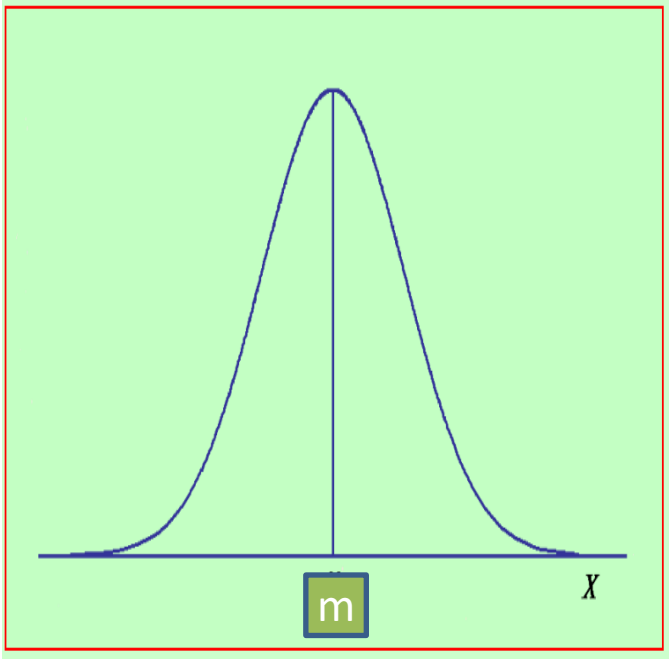
  a+rand()*(b-a)

Distribuição Normal (ou Gaussiana)

- Uma V.A. diz-se normal ou Gaussiana se

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

- Frequentemente usa-se a notação $N(m, \sigma^2)$
- Curva em forma de sino
- Simétrica em torno da média (m) e com alargamento σ



Distribuição Normal

- Função de distribuição acumulada

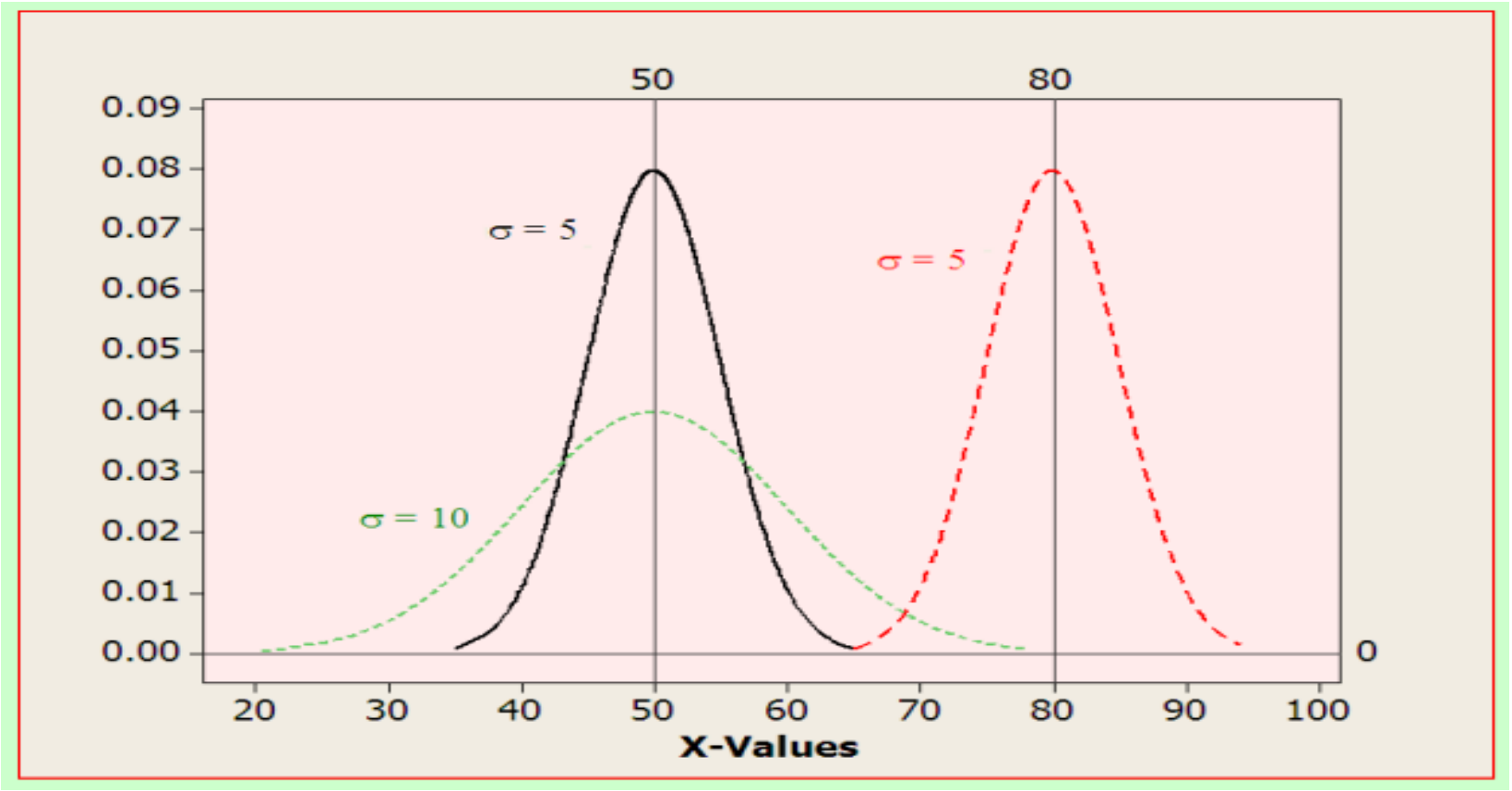
$$F_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(t-m)^2}{2\sigma^2}} dt$$

- $E[X] = m$
- $Var(X) = \sigma^2$

- Nota: é muito comum utilizar-se μ em vez de m para representar a média

Família de curvas

- Variando os 2 parâmetros …



Gaussiana normalizada

- Como existe um número infinito de combinações para *m* e *σ* pode gerar-se uma família infinita de curvas
  - Sendo pouco prático lidar com esta situação, em especial antes da existência de computadores

- Foi desenvolvido um mecanismo pelo qual qualquer distribuição normal pode ser convertida numa distribuição única, a Gaussiana normalizada $N(0,1)$

- A fórmula de conversão é:

$$z = \frac{x-m}{\sigma}$$

Ou seja, subtrair a média e dividir pelo desvio padrão

Gaussiana normalizada

- Função densidade de probabilidade:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x)^2}{2}}$$

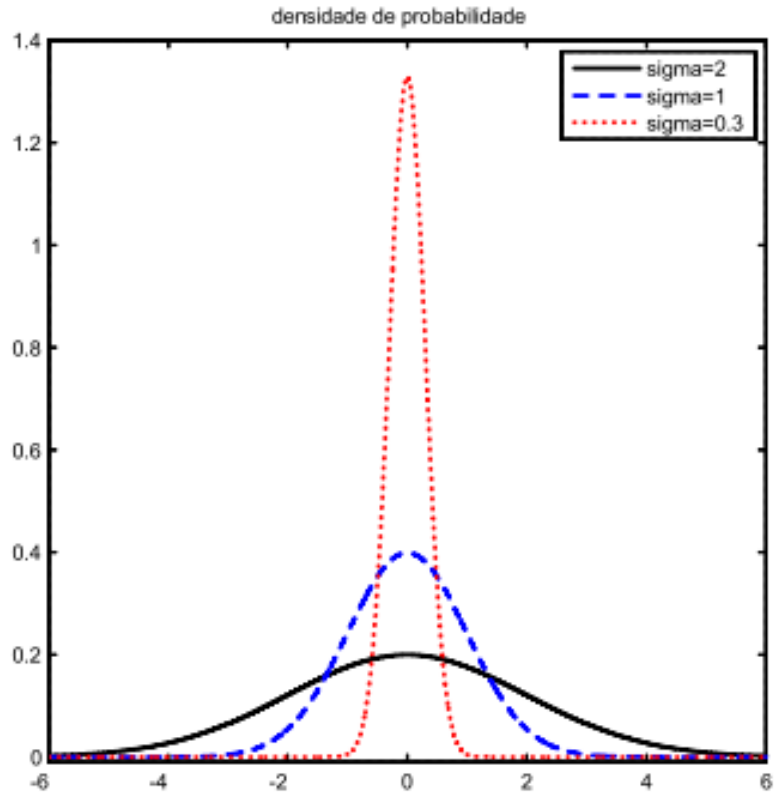
- A função de distribuição acumulada $\Phi(x)$:

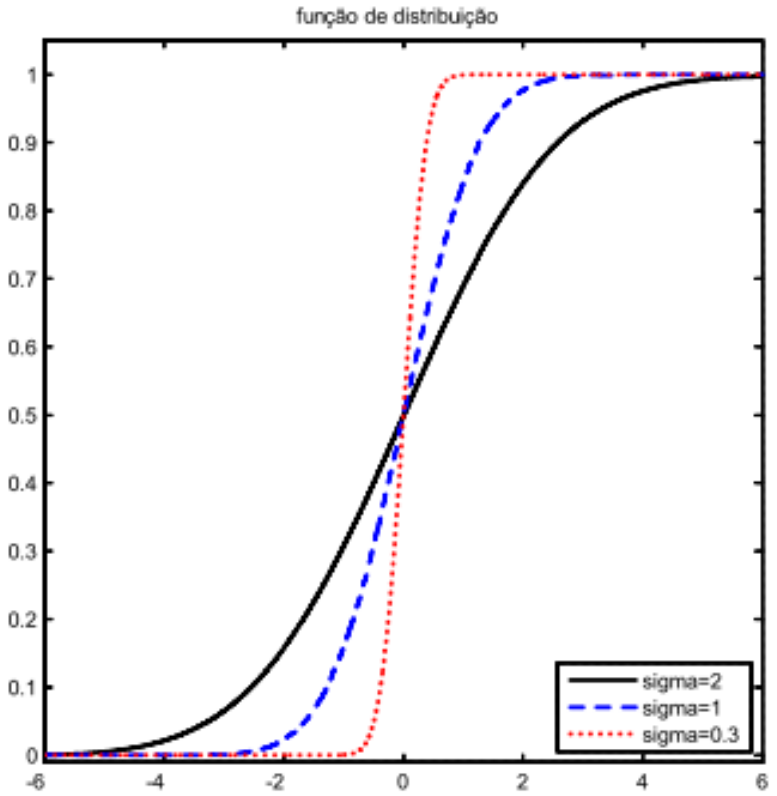
$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t)^2}{2}} dt$$

- $\Phi(x)$ encontra-se frequentemente tabelada

- Outras tabelas comuns são as de $\boldsymbol{Q(x) = 1 - \Phi(x)}$

Gaussiana normalizada





Função distribuição acumulada

- A função de distribuição (acumulada) de N(m,σ^2) pode ser expressa em termos de $\Phi(x)$

$$F_X(x) = \Phi\left(\frac{x-m}{\sigma}\right)$$

Exemplo de uso de $Q(x)$

- Uma empresa, monopolista do mercado de um determinado produto, tem uma procura mensal *X* que segue uma distribuição normal N(75,100). Determine P[78<*X*<80]

$$P\left[78 \leq X \leq 80\right] = P\left[\frac{78-75}{10} \leq U_X \leq \frac{80-75}{10}\right] =$$
$$= P\left[0.3 \leq U_X \leq 0.5\right] = Q(0.3) - Q(0.5) = 0.074$$

Distribuição Normal

- É muito provavelmente a mais conhecida e utilizada de todas as distribuições (contínuas)

- Adequa-se/ajusta-se a muitas características humanas
  - Altura, peso, velocidade, resultados de testes de inteligência, esperança de vida…
- Também se adequa a muitas outras coisas da natureza
  - Árvores, animais etc têm muitas características que seguem a distribuição normal

- Surge quando vários efeitos acumulados e independentes se sobrepõem

Distribuição Normal e a Binomial

- Demonstra-se que a função massa de probabilidade da Binomial de média $m = np$ e $\sigma^2 = np(1-p)$ com m não muito pequeno e n elevado pode ser aproximada por:

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(k-m)^2}{2\sigma^2}}$$

- Ou seja a distribuição normal
  - Desde que $m = np$ e variância igual $np(1-p)$

Distribuição exponencial

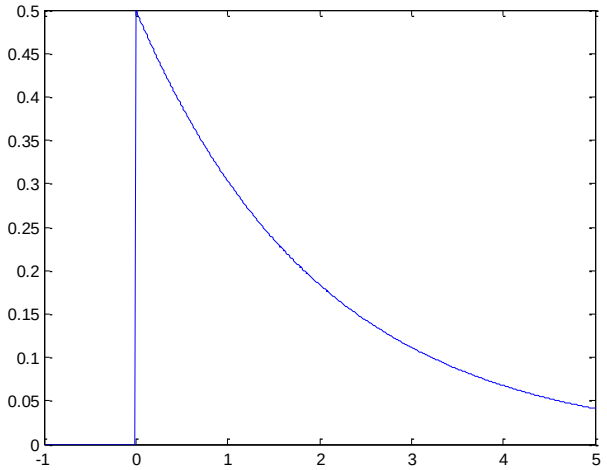
- Surge frequentemente em problemas envolvendo filas de espera e fiabilidade
  - Exemplos:
    - Tempo até um computador avariar
    - Tempo entre chegada de utentes à urgência de um Hospital
- É não negativa (prob. 0 para $x<0$)
- Está relacionada com a distribuição (discreta) de Poisson
  - Se o número de acontecimentos que ocorrem num intervalo seguem distribuição de Poisson, o tempo entre eles segue distribuição exponencial

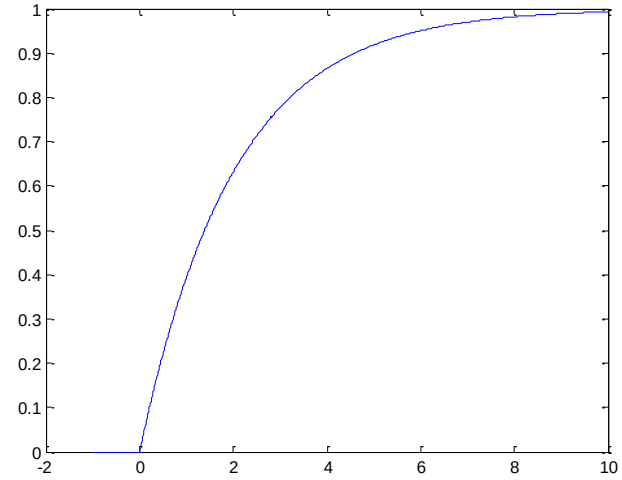
Distribuição exponencial

- $f_X(x) = \begin{cases} 0, & x < 0 \\ \lambda e^{-\lambda x}, & x \geq 0 \end{cases}$

- $F_X(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\lambda x}, & x \geq 0 \end{cases}$

- $E[x] = \frac{1}{\lambda}$
- $\mathrm{Var}(x) = \frac{1}{\lambda^2}$





Exemplo de aplicação

- A vida útil, em milhares de horas, de um componente de um robô é uma variável aleatória com distribuição exponencial de valor médio 10 (milhares de horas)
- Qual a probabilidade de um desses componentes selecionado ao acaso durar menos de 4000 horas?
- $\lambda = \frac{1}{10} = 0{,}1$

$$P[X < 4] = \int_0^4 0.1e^{-0.1x}dx = F_X(4) = 0.33$$

Outras distribuições

Distribuição dos primeiros dígitos

- Em 1881, um matemático e astrónomo americano, Simon Newcomb, percebeu que as primeiras páginas dos livros de logaritmos das bibliotecas estavam mais gastas que o resto, intrigado, investigou o assunto e…
- percebeu que em amostras aleatórias de dados reais o dígito 1 aparece quase 1/3 das vezes
  - Em lugar dos 1/9 se seguissem uma distribuição uniforme (discreta)

- Mais tarde, em 1938, o físico **Frank Benford** após uma investigação mais profunda chegou à mesma conclusão

Lei/Distribuição de Benford

- Função probabilidade →

$$P(d) = \log_{10}(d+1) - \log_{10}(d) = \log_{10}\left(\frac{d+1}{d}\right) = \log_{10}\left(1+\frac{1}{d}\right).$$

d	*P*(*d*)	Valor Relativo de P(d)
1	30.1%	
2	17.6%	
3	12.5%	
4	9.7%	
5	7.9%	
6	6.7%	
7	5.8%	
8	5.1%	
9	4.6%	

- A Lei/Distribuição de Benford, também conhecida como a "Lei dos Primeiros Dígitos", é uma ferramenta muito poderosa e muito simples que aponta suspeitas de fraudes, erros de digitação etc

- Mais info:
  - https://pt.wikipedia.org/wiki/Lei_de_Benford
  - http://gigamatematica.blogspot.pt/2011/07/lei-de-benford.html

Lei/Distribuição de Zipf

- George Kingsley **Zipf**, linguista da Universidade de Harvard, analisou a obra monumental de James Joyce, *Ulisses*, e contou as palavras distintas, ordenando-as por frequência

- Verificou que:
  - a palavra mais comum surgia 8000 vezes;
  - a décima, 800 vezes;
  - a centésima, 80 vezes;
  - a milésima, 8 vezes.

Lei de Zipf

- A **Lei de Zipf** é uma lei empírica que rege a dimensão, importância ou frequência dos elementos de uma lista ordenada
  - formulada na década de 1940 por Zipf

- Numa lista ordenada, o membro n teria uma relação de valor com o 1º da lista segundo 1/n

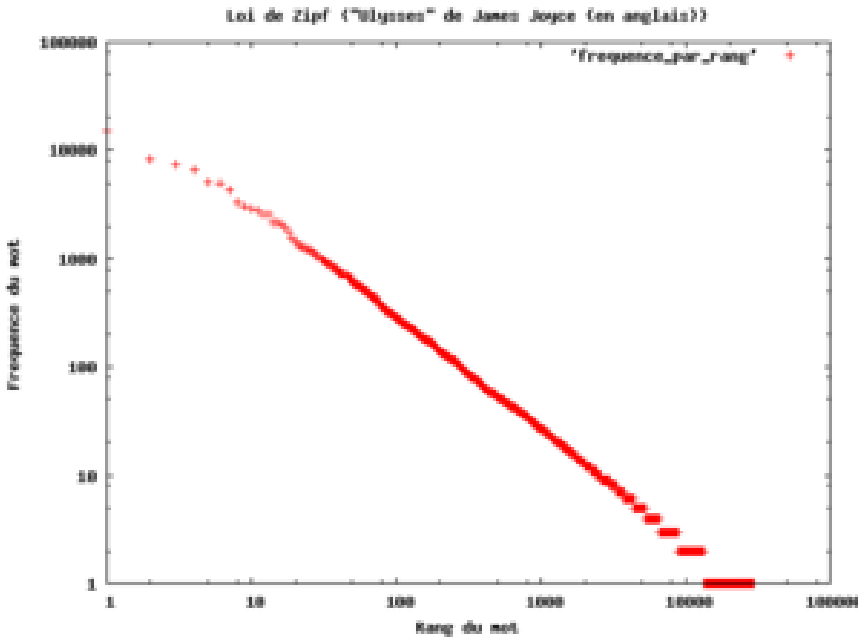
- Mais info: https://pt.wikipedia.org/wiki/Lei_de_Zipf

Lei de Zipf

- A probabilidade de ocorrência $p(n)$ de uma palavra esteja ligada à sua ordem n na ordem das frequências por:

$$p(n) = \frac{K}{n}$$

- Sendo K uma constante dependente da língua
- $p(n)$ é estimada com base na contagem de ocorrências de palavras num texto ou conjunto de textos



Frequência das palavras em função da ordem na versão original de Ulisses *de James Joyce*.
De: Wikipedia

Lei de Zipf – Inglês escrito

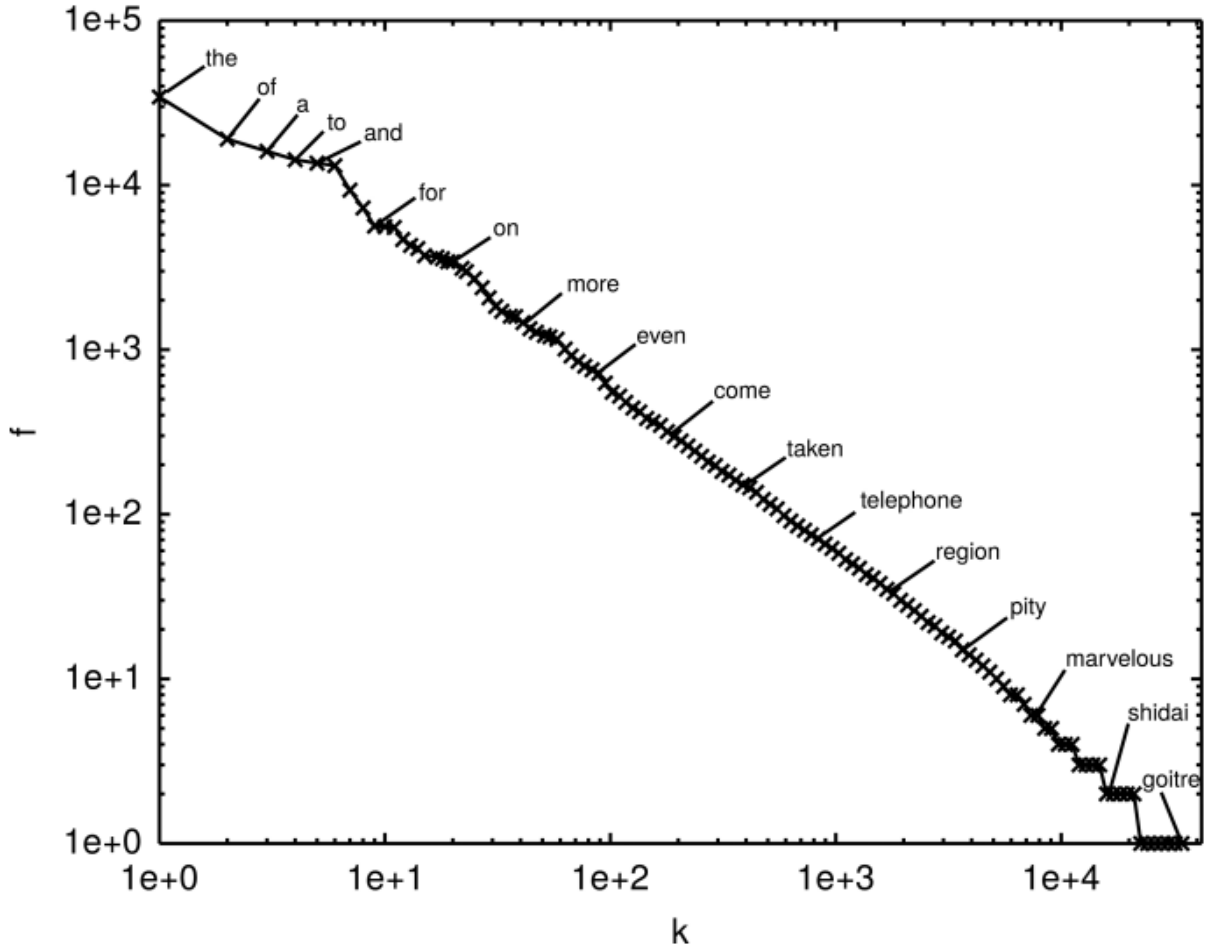


Figura 1: Lei de Zipf no Inglês escrito (dados do OANC). *Rank* (k) versus frequência de ocorrência (f).

Exemplo de aplicação da Lei de Zipf

- Aplicação na área da segurança:

Revista Brasileira de Ensino de Física, vol. 38, nº 1, 1313 (2016)
www.scielo.br/rbef
DOI: http://dx.doi.org/10.1590/S1806-11173812125

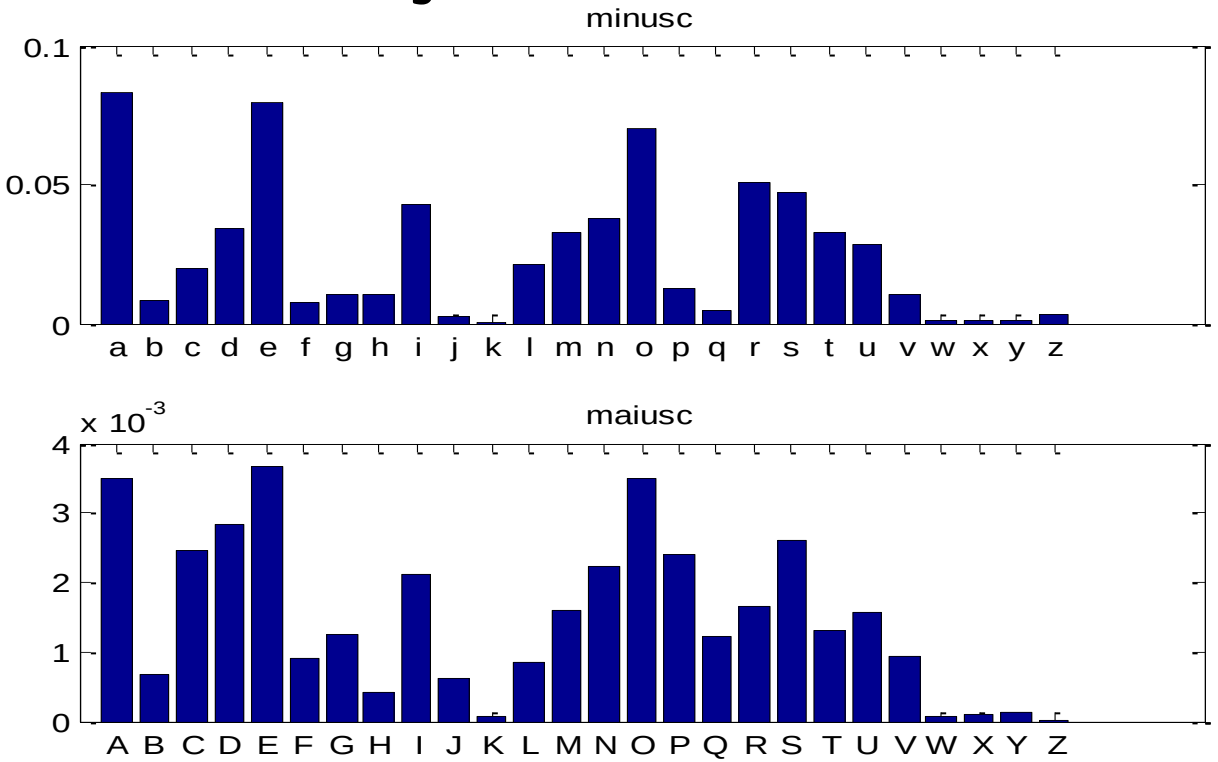
Influência da lei de Zipf na escolha de senhas

Influence of Zipf's law on the password choices

Leonardo Carneiro de Araújo*[1], João Pedro Hallack Sansão[1], Hani Camille Yehia[2]

- Artigo em PDF disponível em http://www.scielo.br/pdf/rbef/v38n1/1806-9126-rbef-38-01-S1806-11173812125.pdf

Outra distribuição:
Distribuição das letras em Português …



char	prob.
a	0.083
b	0.008
c	0.020
d	0.034
e	0.079
f	0.007
g	0.010
h	0.011
i	0.043

…..

- Probabilidades estimadas usando o texto pg21209.txt do projecto Gutemberg

Mais informação

- Capítulo 5 ("Estudo de Algumas Distribuições") do livro "Métodos Probabilísticos para Engenharia Informática"
  - Francisco Vaz e António Teixeira, Sílabo

- Material online
  - Lectures:
    - http://www.stat.berkeley.edu/~stark/SticiGui/Text/randomVariables.htm