Estimating Population from Samples

- **Sample mean is a random variable**
  $\Rightarrow$ Mean has some distribution

  $\therefore$ Multiple sample means have "mean of means"

- **Knowing distribution of means can estimate error**

Confidence Intervals

- Sample mean value is only an estimate of the true population mean
- Bounds c_1 and c_2 such that there is a high probability, $1-\alpha$, that the population mean is in the interval (c_1,c_2):

$$\text{Prob}\{ c_1 < \mu < c_2\} = 1-\alpha$$

where α is the significance level and
$100(1-\alpha)$ is the confidence level

Confidence Interval of Sample Mean

- **Knowing where 90% of sample means fall, we can state a 90% confidence interval**

- **Key is *Central Limit Theorem:***
  - **Sample means are normally distributed**
  - **Only if independent**
  - **Mean of sample means is population mean μ**
  - **Standard deviation (*standard error) is* $\sigma/\sqrt{n}$**

Confidence Interval of Sample Mean

Queremos c_1 e c_2 tal que: $P(c_1 \le \mu \le c_2) = 1 - \alpha$

Seja a media amostral $\bar{x}$. Pelo teorema central do limite temos $\bar{x} \sim N(\mu, \sigma/\sqrt{n})$

Entao: $P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \le z_{1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}$ $\qquad P\left(-z_{1-\frac{\alpha}{2}} \le \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \le z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$

$$P\left(-z_{1-\frac{\alpha}{2}}\sigma/\sqrt{n} \le \bar{x} - \mu \le z_{1-\frac{\alpha}{2}}\sigma/\sqrt{n}\right) = P\left(-z_{1-\frac{\alpha}{2}}\sigma/\sqrt{n} \le \mu - \bar{x} \le z_{1-\frac{\alpha}{2}}\sigma/\sqrt{n}\right) = 1 - \alpha$$

$$P\left(\bar{x} - z_{1-\frac{\alpha}{2}}\sigma/\sqrt{n} \le \mu \le \bar{x} + z_{1-\frac{\alpha}{2}}\sigma/\sqrt{n}\right) = 1 - \alpha$$

Se σ desconhecido, s - desvio padrao da amostra, serve como bom estimador

Estimating Confidence Intervals

- **Two formulas for confidence intervals**
  - **Large sample (over 30 observations from any distribution): *z*-distribution**
  - **Small sample from normally distributed population: *t*-distribution**

- **Common error: using *t*-distribution for non-normal population**
  - **Central Limit Theorem "often" saves us**

The z Distribution

- **Interval on either side of mean:**

$$\bar{x} \pm z_{1-\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

- **Significance level α is small for large confidence levels**
- **Tables of *z* are tricky: be careful!**

Example of z Distribution

- **35 samples: 10 16 47 48 74 30 81 42 57 67 7 13 56 44 54 17 60 32 45 28 33 60 36 59 73 46 10 40 35 65 34 25 18 48 63**
- **Sample mean $\bar{x}$ = 42.1. Standard deviation *s* = 20.1. n = 35**
- **90% confidence interval is**

$$42.1 \pm (1.645)\frac{20.1}{\sqrt{35}} = (36.5, 47.7)$$

The t Distribution

- **Formula is almost the same:**

$$\bar{x} \pm t_{\left[1-\alpha/2;n-1\right]}\left(\frac{s}{\sqrt{n}}\right)$$

- **Usable only for normally distributed populations!**
- **But works with small samples**

Example of *t* Distribution

- **10 height samples: 148 166 170 191 187 114 168 180 177 204**
- **Sample mean $\bar{x}$ = 170.5. Standard deviation *s* = 25.1, n = 10**
- **90% confidence interval is**

$$170.5 \pm (1.833)\frac{25.1}{\sqrt{10}} = (156.0, 185.0)$$

- **99% interval is (144.7, 196.3)**

Getting More Confidence

- **Asking for a higher confidence level widens the confidence interval**
  - **Counter-intuitive?**
- **How tall is Fred?**
  - **90% sure he's between 155 and 190 cm**
  - **We want to be 99% sure we're right**
  - **So we need more room: 99% sure he's between 145 and 200 cm**

Intervalos de Confiança e Testes de Hipótese

- **Teste de hipótese:**
  - **Hipótese nula H_0 versus hipótese alternativa H_a**
    - **H_0 = dois métodos A e B produzem resultados estatisticamente iguais**

      H_a = dois métodos produzem resultados estatisticamente diferentes

    H_o: $\mu_A = \mu_B$ H_A: $\mu_A \neq \mu_B$
  - **Computa alguma estatística dos dados que permita testar as hipóteses**

    Computa $\overline{x_A - x_B}$
  - **Faz referência a alguma distribuição que mostra como a estatística seria distribuída se a hipótese nula fosse verdadeira**
    - **Ex: já sabemos que a distribuição das médias segue uma Normal**

Intervalos de Confiança e Testes de Hipótese

- Com base na distribuição de referência, computa a probabilidade de se obter uma discrepância tão grande quanto a observada e H_o ainda ser verdadeira
  - p-value
- Quanto menor o p-value, menos provável é que a hipótese nula seja verdadeira e mais significativo (estatisticamente) o resultado é
  - Quanto menor o p-value, maior a chance de : $\mu_A \neq \mu_B$
- Rejeita hipótese nula se p-value < nível de significância α

- Intervalos de confiança e testes de hipótese: mesmo arcabouço

Distribuições Comuns de Variáveis Aleatórias Discretas

1. **Uniforme**
2. **Bernoulli**
3. **Binomial**
4. **Geometrica**
5. **Poisson**

Distribuição Discreta Uniforme

- **A v.a. discreta *X* que assume *n* valores discretos com probabilidade $p_X(i) = 1/n,\ 1 \le i \le n$**

- ***pmf*** $$p_X(x_i) = \begin{cases} 1/n, & se \quad x_i \in X \\ 0, & caso \quad contrário \end{cases}$$

- **CDF:** $$F(t) = \sum_{i=1}^{t} p_X(i) = \frac{t}{n}$$

Variável de Bernoulli

- **V.A gerada por um experimento único de Bernoulli tem um resultado binário {1, 0} ou {sucesso, falha}**
- **A v.a. binária X é chamada variável de Bernoulli tal que:**

Função de massa de probabilidade:

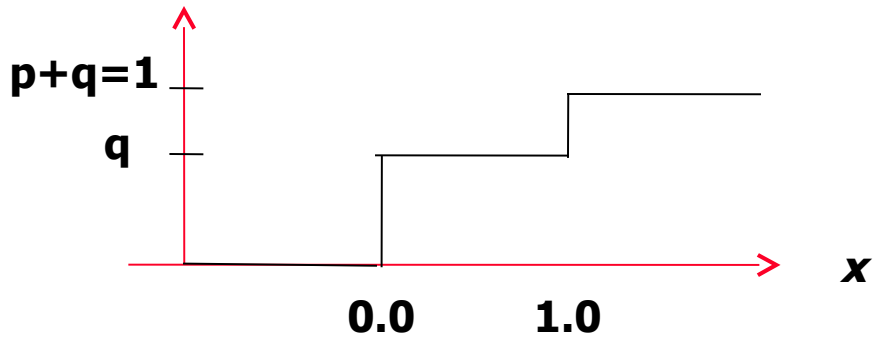
$$p = P(X = 1)$$

$$q = 1 - p = P(X = 0)$$

Distribuição de Bernoulli

- **CDF**

$$F(x) = \begin{cases} 0 & x < 0 \\ q & 0 \leq x < 1 \\ 1 & x \leq 1 \end{cases}$$



Binomial

- **A v.a. X representa o numero de sucessos em uma sequencia de experimentos de Bernoulli.**
- **Todos experimentos são independentes.**
- **Cada resultado é um "sucesso" ou "falha".**
- **A probabilidade de sucesso de um experimento é dado por *p*. A probabilidade de uma falha é *1- p*.**
- **Uso do modelo: número de processadores "down" num cluster; número de pacotes que chegam ao destino sem erro.**

Distribuição Binomial

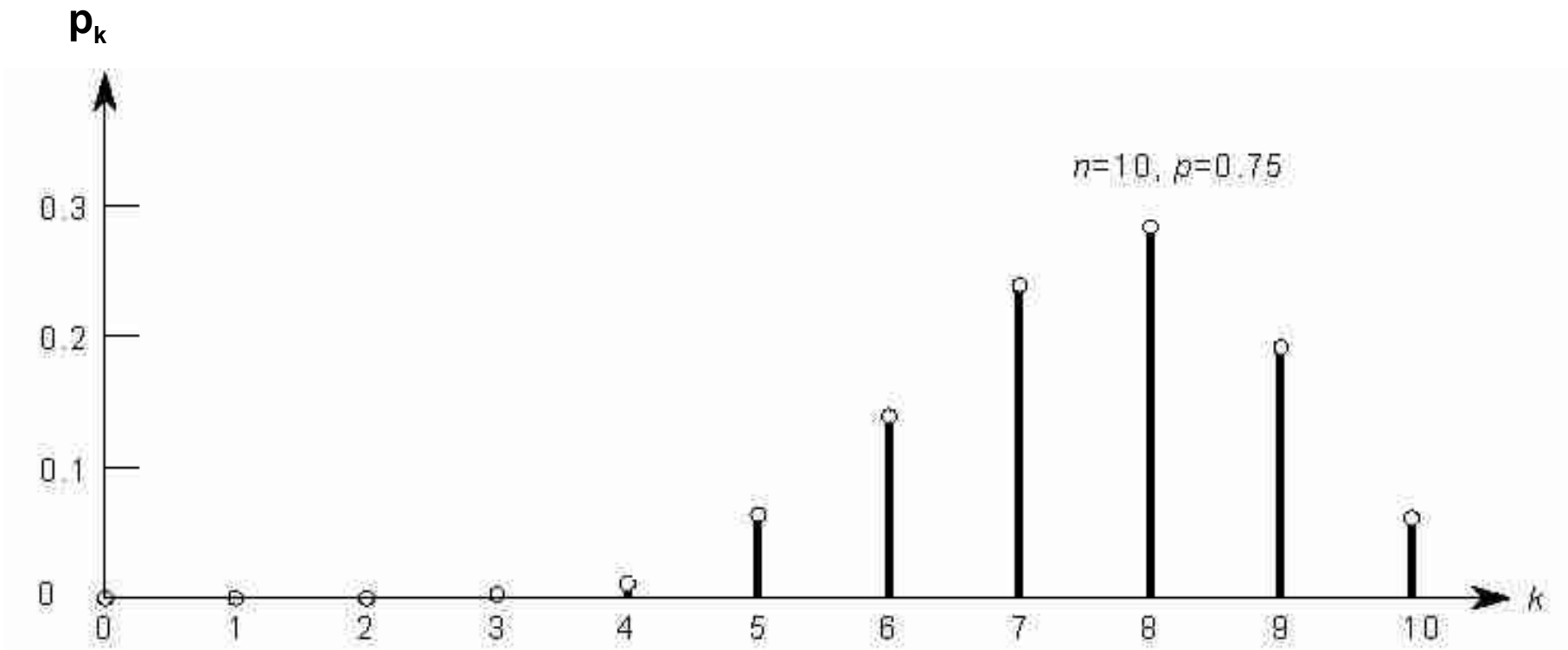
A distribuição binomial com parâmetros $n \geq 0$ and $0 < p < 1$, is

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

A média e variância da binomial são:

$$\mu = np \qquad \sigma^2 = np(1-p)$$

V.A. Binomial: pmf



Distribuição Geométrica

- **Número de experimentos até incluir o 1º sucesso.**

- **Em geral , *S* pode ter um tamanho infinitamente contável**

$$S = \{\underbrace{0, 0 \ldots, 0}_{i\text{-}1 times}, 1\} = \{0^{i-1}1 | i = 1, 2, 3, \ldots\}$$

- **Definir a v.a *Z* ($\in$ S):**
- **Por causa da independência:**

$$p_Z(i) = q^{i-1}p = p(1-p)^{i-1}, \ i = 1, 2, 3, \ldots$$

$$F_Z(t) = \sum_{i=1}^{\lfloor t \rfloor} p(1-p)^{i-1} = 1 - (1-p)^{\lfloor t \rfloor}, \ t \geq 0$$

$$E(x) = \frac{1}{p}$$

Geométrica

- **A distribuição geometrica é a única distribuição discreta que exibe a propriedade *MEMORYLESS*.**
- **Resultados futuros são independentes de eventos passados.**
- **Exemplo:**
  - **Z: numero de experimentos ate sucesso. Ja observamos n experimentos: todos com falhas.**
  - ***Y:* numero de experimentos adicionais necessarios ate que um sucesso ocorra, i.e. $Z = n+Y$ ou $Y=Z-n$**

Geométrica: ausência de memória

- ***Y=Z-n***

$$= P(Y = i \mid Z > n)$$

$$= P(Z - n = i \mid Z > n)$$

$$= P(Z = n + i \mid Z > n)$$

$$= \frac{P(Z = n + i \; and \; Z > n)}{P(Z > n)}$$

$$= \frac{P(Z = n + i)}{P(Z > n)} = \frac{P(Z = n + i)}{1 - F_Z(n)} = \frac{p_Z(n + i)}{1 - F_Z(n)}$$

$$= \frac{pq^{n+i-1}}{1 - (1 - q^n)} = pq^{i-1} = p_Z(i)$$

VA Poisson

- Número de eventos *independentes* que ocorrem em um intervalo de tempo
  (veja discussão em Ross, 4.8)
- Número de chegadas em um servidor em 1 hora
- Número de erros de impressão em uma página de um livro

  λ = # médio de eventos que ocorrem no período

- Aproximação para VA Binomial com n grande e p pequeno (Ross)

  Se X = Binomial(n,p), $X \approx \text{Poisson}(\lambda = np)$

VA Poisson: Aplicacao

- A V.A. de Poisson é boa para modelar vários fenômenos, como o número de transações que chegam a um servidor em uma hora, ou o número de queries que chegam a uma máquina de busca em 1 minuto ou número de pacotes que chegam num roteador em 1 segundo.
- Muito comumente usado para modelar chegada de sessões de usuários
  - servidores Web, multimídia, banco de dados, ftp, e-mail
- Sessões são iniciadas por usuários
  - Chegada de duas sessões tendem a ser independentes: Poisson é uma boa aproximação
- Contra-exemplo:
  - Chegada de *requisições* em um servidor Web
  - Premissa de independência não é válida: existe dependência entre requisições para o arquivo HTML e as imagens embutidas nele

Poisson

- **Uma v.a. de Poisson X tem sua pmf::**

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda} \qquad x = 0, 1, 2, \ldots$$

Onde $\lambda > 0$ é uma constante

$$E(X) = Var(X) = \lambda$$

Search Algorithms: Is the Web-Graph a Random graph?

- **Random graph $G_{n,p}$:**
  - **n nodes**
  - **every directed edge occurs with probability p**

- **Is the Web-graph a random graph $G_{n,p}$?**

- **In a random graph. degrees are distributed according to a Poisson distribution**

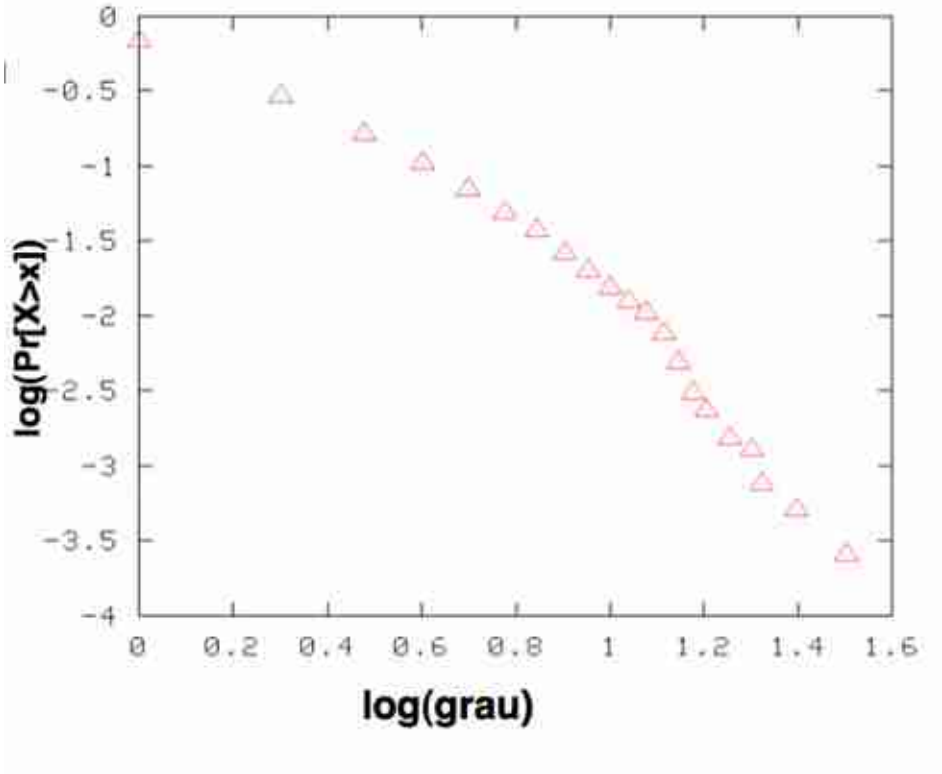
$$\mathbf{Prob}[X = k] = \frac{e^{-\lambda}\lambda^k}{k!}$$

Is the Internet (router-level) a random graph?

- Em 1999, foi publicado um resultado supreendente sobre a topologia de roteadores da Internet [FFF99]

- Distribuição segue lei de potência

- Medidas de traceroute posteriores confirmaram resultados

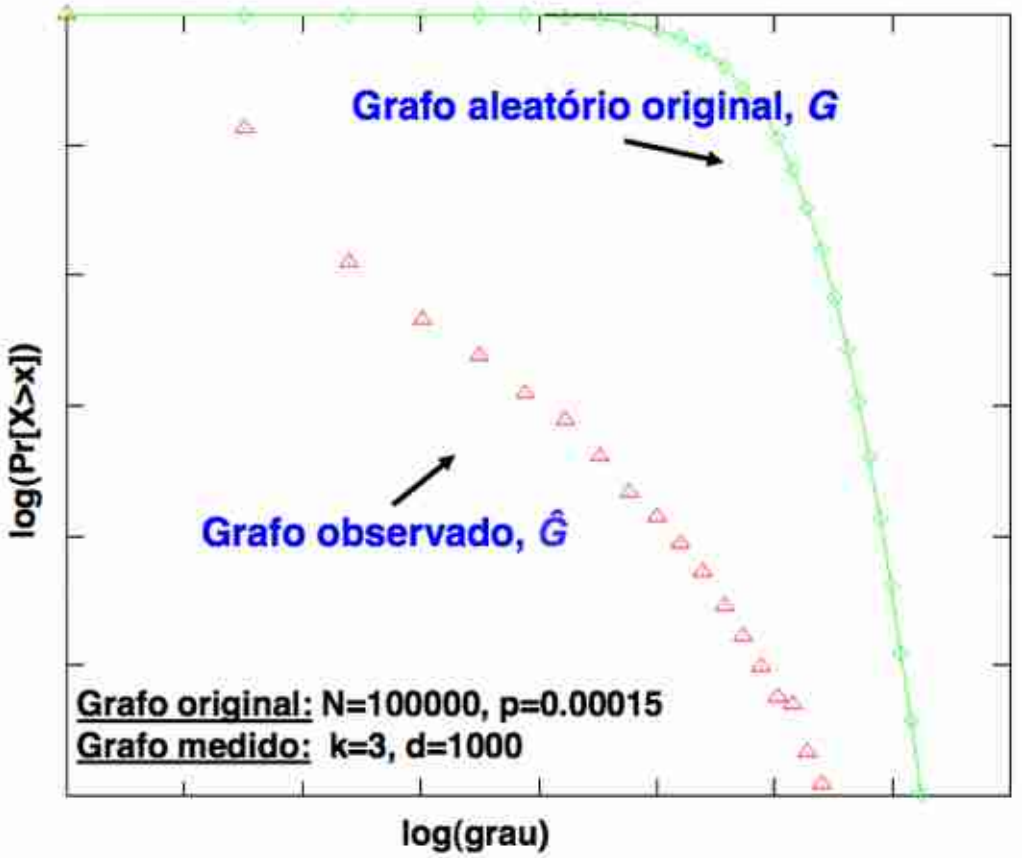


Slide inspirado na apresentação de Bruno Ribeiro: "Entendendo Redes Complexas através de Informação Experimental Incompleta: Mitos, Erros e Acertos", WPerformance 2010

Rede de roteadores da Internet

- **Porém, em 2003, mostrou-se que método traceroute produz amostras tendenciosas [Byers 2003]:**

- **Verificado teoricamente**



Slide inspirado na apresentação de Bruno Ribeiro: "Entendendo Redes Complexas através de Informação Experimental Incompleta: Mitos, Erros e Acertos", WPerformance 2010

Exercícios

1. **Considere que o número de mails que chegam a um servidor de mails no intervalo *t* segundos é distribuído como Poisson com parâmetro *0.3t.* Calcule a seguintes probabilidades:**
   - **Exatamente tres mensagens chegarão num intervalo de 10 seg.**
   - **No máximo 20 msgs chegarão num período de 20seg.**
   - **O número de msgs num intervalo de 5 seg está entre 3 e 7 mails.**

2. **A probabilidade de um *query* falhar (não ser bem sucedido) é $10^{(-4)}$. Qual a probabilidade de falharem mais de 3 queries numa sequência de 1000 *queries*?**

Solução do Exercício 1

$$P(X_t = k) = \frac{(0.3t)^k}{k!} e^{-0.3t}$$

1) $P(X_{10} = 3) = 0.224$

2) $P(X_{20} \leq 20) = 0.973$

2) $$P(3 \leq X_5 \leq 7) = \sum_{k=3}^{7} \frac{(1.5)^k}{k!} e^{(-1.5)} = 0.1909$$

Solução do Exercício 2

Evento falha de uma query: distribuição Bernoulli com p = 10^{-4}

Evento # de falhas: distribuição Binomial com n = 1000 e p = 10^{-4}

$$P(\#\ falhas > 3) = \sum_{4}^{1000} \binom{1000}{i} (10^{-4})^{i} (1-10^{-4})^{1000-i}$$

$$P(\#\ falhas > 3) = 1 - \sum_{0}^{3} \binom{1000}{i} (10^{-4})^{i} (1-10^{-4})^{1000-i} = 3.825 * 10^{-6}$$

Distribuições Discretas

- Zipf(α)
  - Comumente usada quando a distribuição é altamente concentrada em poucos valores
    - Popularidade de arquivos em servidores Web/multimídia
      - 90% dos acessos são para 10% dos arquivos
    - Popularidade de palavras na língua inglesa
  - Seja i, o elemento que ocupa a i-esima posição no ranking de concentração

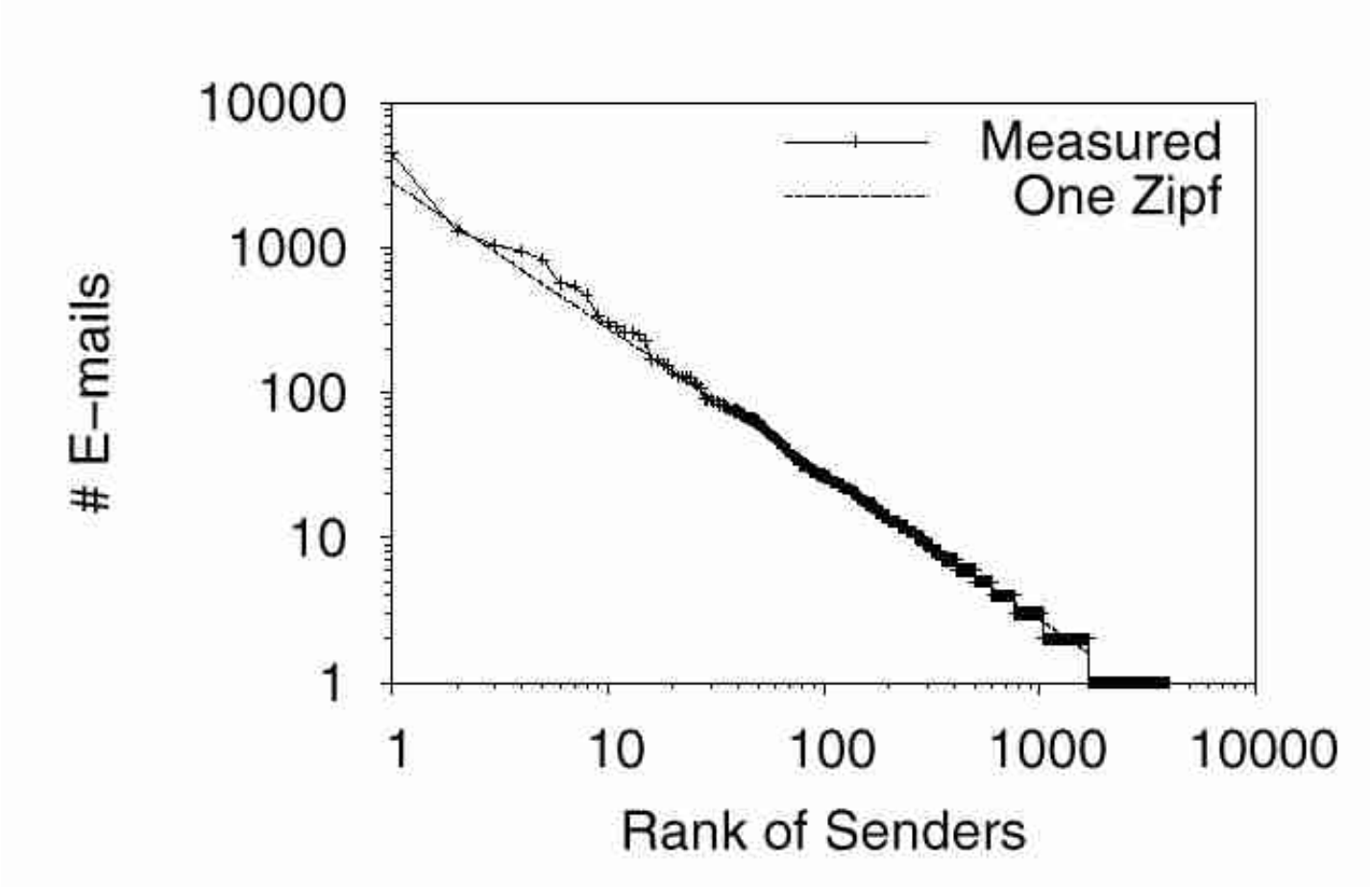
$$P(X = i) = \frac{C}{i^{\alpha}} \qquad i = 1,2,...$$

C é a constante de normalização

Zipf: lei das Potências

Distribuição Zipf

- **Modela popularidade dos remetentes de e-mails para a UFMG**



Distribuições de Variáveis Aleatórias Contínuas

- **Normal**
- **Exponencial**
- **Weibull**
- **Lognormal**
- **Pareto**
- **....**

Distribuições de Variáveis Aleatórias Contínuas

- **Variáveis aleatórias contínuas**
  - **Assumem um intervalo infinito de diferentes valores**
  - **W= percentual de crescimento do PIB em 2005**
  - **V=tempo para retornar a resposta de um "query"**
  - **Valores específicos-particulares de uma v.a. contínua tem probabilidade 0**
  - ***Intervalos de valores tem probabilidade ≠ 0***

Distribuição Normal (Gaussiana)

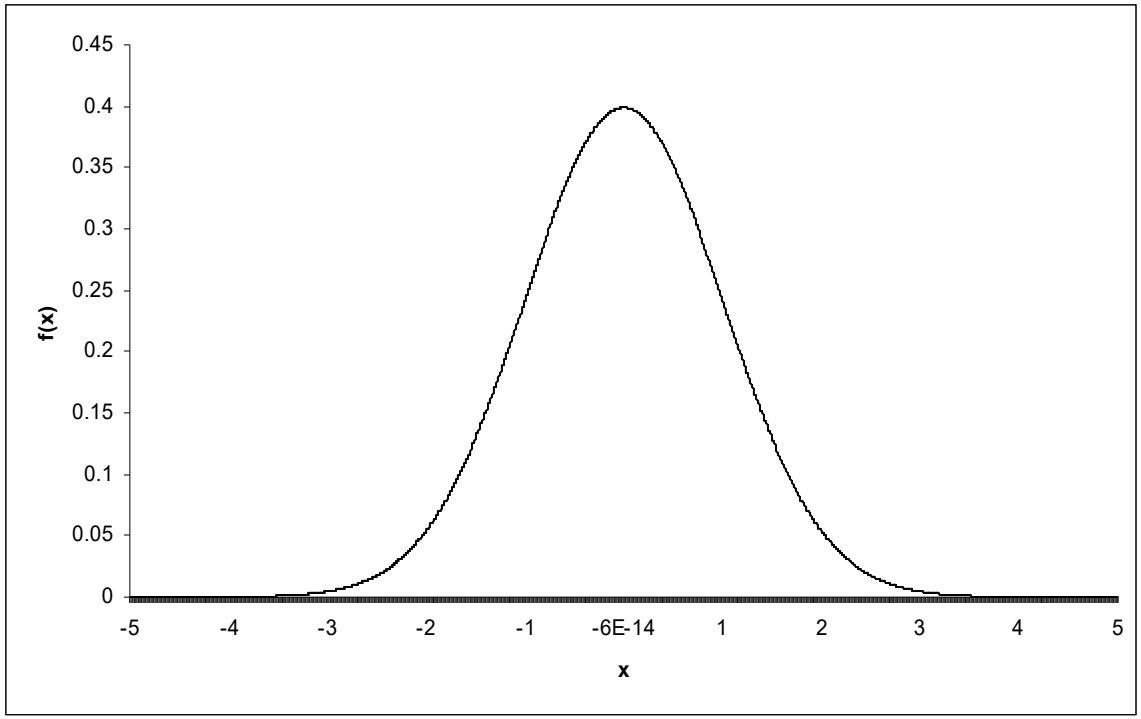
- **Distribuição mais comum na análise de dados**
- **pdf is:**

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

- **$-\infty \leq x \leq +\infty$**
- **Média é μ , desvio padrão σ**

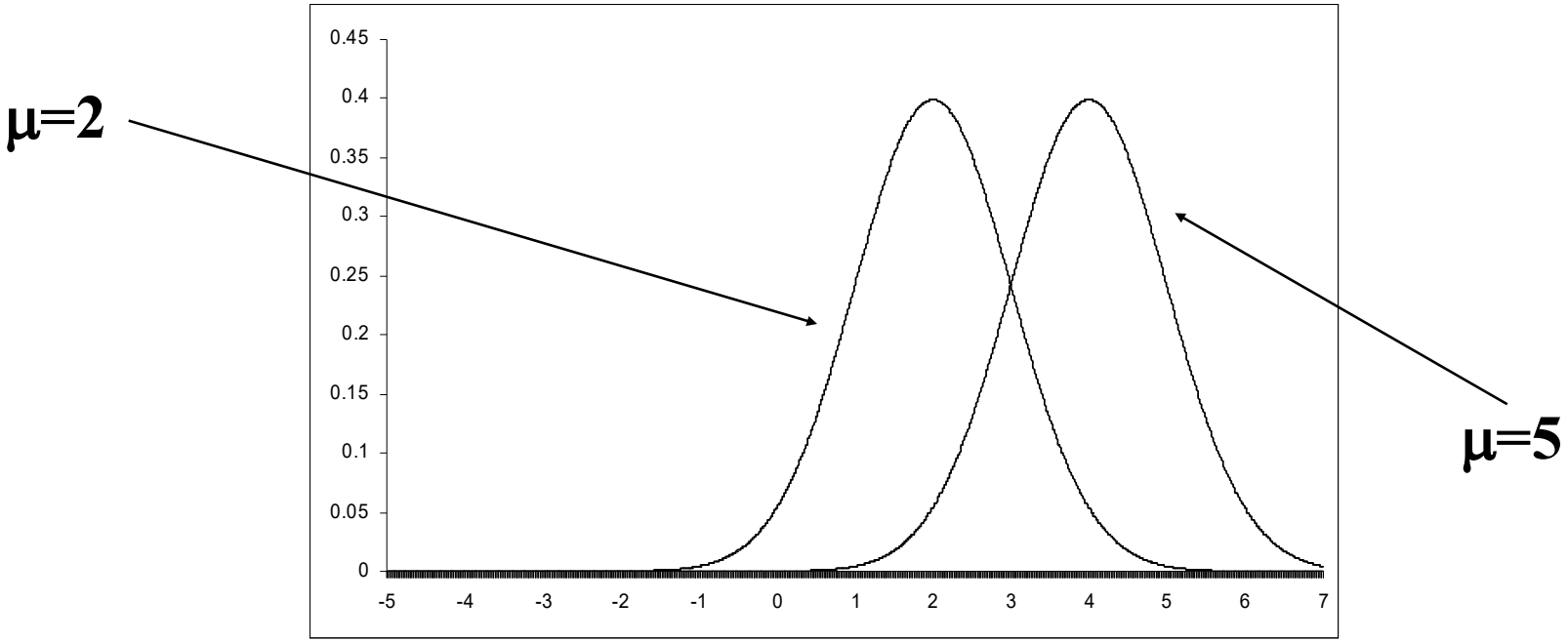
Normal

- **Função de densidade para $\mu=0$, $\sigma=1$**



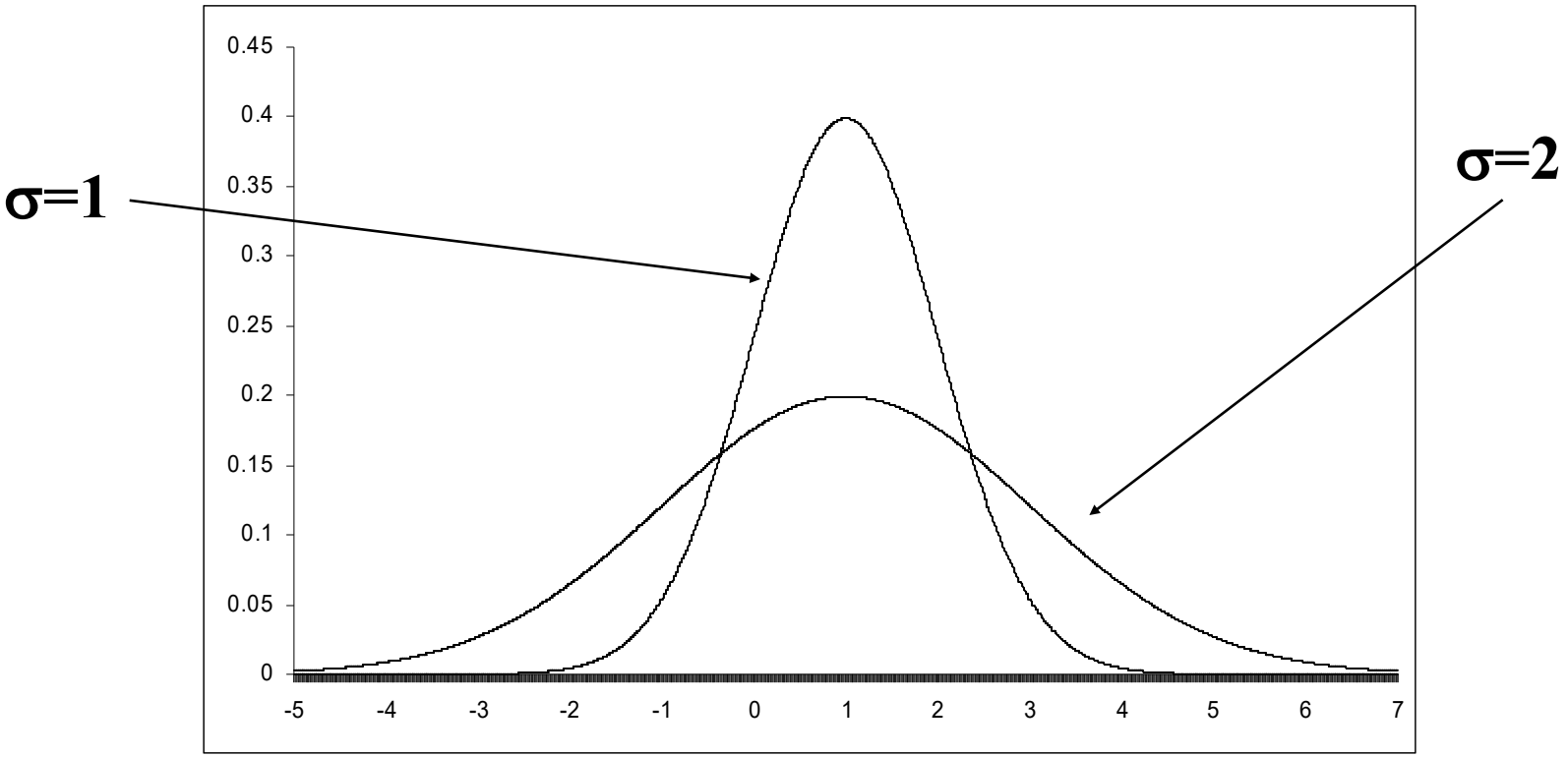
Normal

- **Função de densidade para σ=1**



Normal

- **Funções de densidade para μ=1**



Notação para Distribuições Gaussianas

- **Geralmente denotada $N(\mu,\sigma)$**
- **Normal unitária é $N(0,1)$**
- **Se x tem $N(\mu,\sigma)$, $\frac{x-\mu}{\sigma}$ tem $N(0,1)$**

- **O α-quantil de uma normal unitária $z \sim N(0,1)$ é denotado por z_α tal que**

$$\left\{P\left(\frac{x-\mu}{\sigma}\right) \le z_\alpha\right\} = \left\{P(x) \le \mu + z_\alpha \sigma\right\} = \alpha$$

Distribuicao Exponencial

- **Quantidade de tempo até que determinado evento ocorra**

$$f_X(x) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0$$

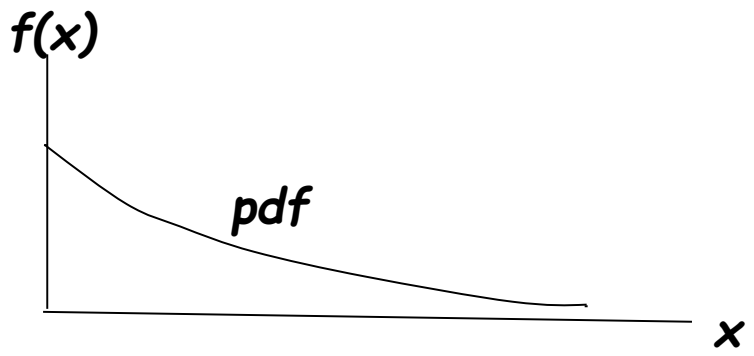
$$F_X(x) = 1 - e^{-\lambda x} \quad \text{for } x \geq 0$$

λ = taxa de chegadas

1/ λ = tempo médio entre chegadas

Exemplo: v.a. exponencial

- **pdf:** $f(x) = \lambda e^{-\lambda x}, x \geq 0$
- **CDF:** $F(x) = 1 - e^{-\lambda x}$



- V.A. muito frequentemente usada em computacao
- Modelos:
  - Tempo entre duas submissões de queries
  - Tempo de execução de processos
  - Tempo entre chegadas de pacotes em um roteador
  - Tempo entre chegadas de sessões em um servidor

Distribuicao Exponencial

$$P(X \leq 1/\lambda) = 1 - e^{-\lambda \times 1/\lambda} = 1 - 1/e$$

$$E(X) = 1/\lambda$$

$$Var(X) = 1/\lambda^2 \Rightarrow SD(X) = 1/\lambda \Rightarrow$$

$$CV(X) = 1$$

$$CV = 1 \not\Rightarrow \text{exponencial}$$

Distribuições Exponencial e Poisson

- Seja uma distribuição Poisson que denote o número de eventos N em um intervalo de tempo t
- Seja T_1 o momento do 1o evento
- Seja T_n o tempo entre o (n-1)-esimo e o n-esimo eventos
- Sequência $\{T_n, n=1, 2, ...\}$: tempos entre chegadas

$$P(T_1 > t) = P(N(t) = 0) = e^{-\lambda t} \Rightarrow T_1 \sim \text{exponencial}(\lambda)$$

$$P(T_2 > t \mid T_1 = s) = \text{Prob (0 eventos em } (s, s+t) \mid T1 = s)$$

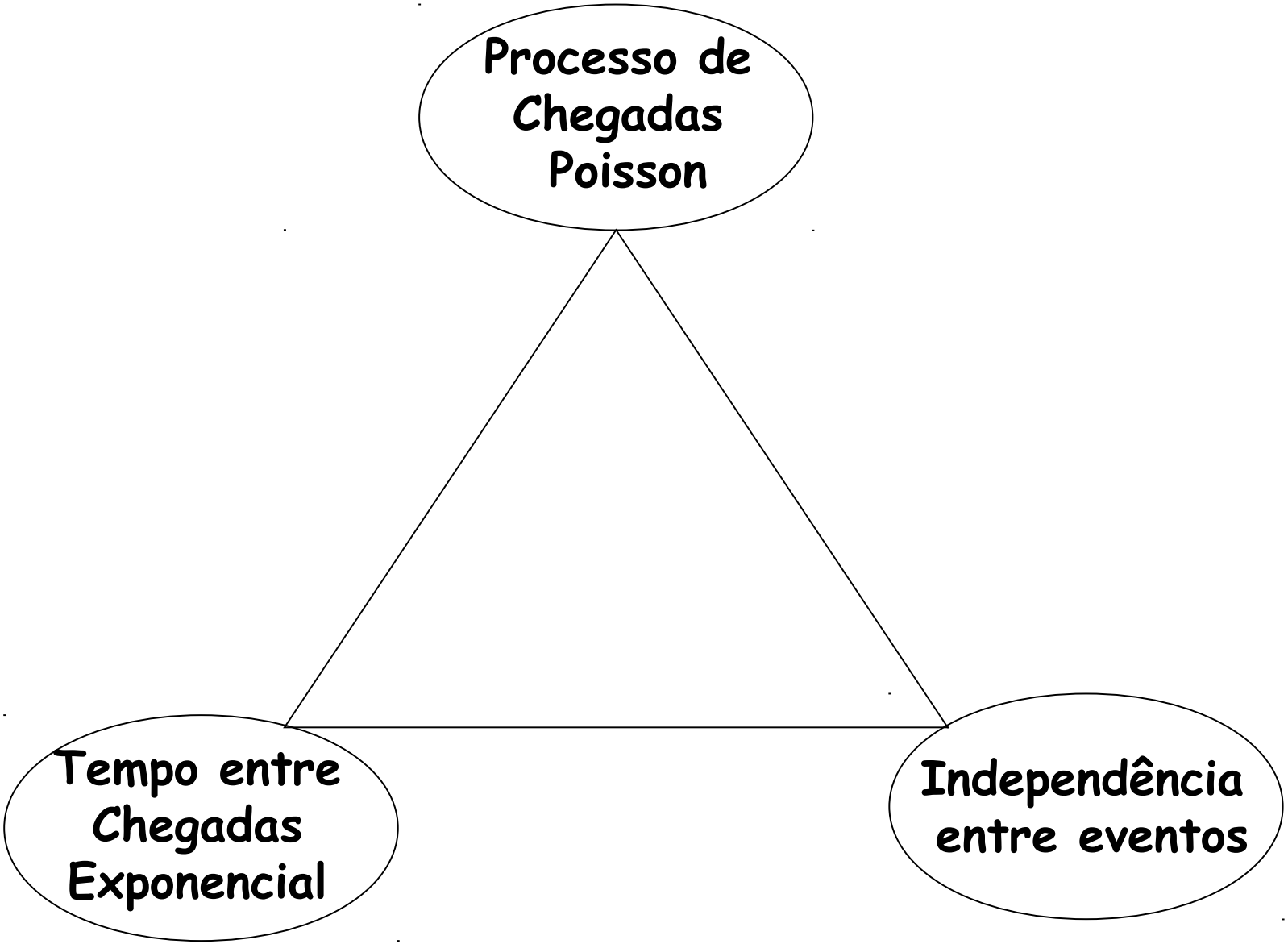
$$= \text{Prob (0 eventos em } (s, s+t))$$

(eventos Poisson são independentes)

$$= e^{-\lambda t} \Rightarrow T_2 \sim \text{exponencial}(\lambda)$$

$\Rightarrow T_1, T_2, \ldots, T_n$ são independentes e têm mesma distribuição exponencial(λ)

Distribuições Exponencial e Poisson



Distribuição Exponencial

- **Exponencial (λ) :**

$$P(X \le t+x \mid X > t) = \frac{P([X \le t+x] \cap [X > t])}{P(X > t)}$$

$$= \frac{P(t < X \le t+x)}{1 - P(X \le t)}$$

$$= \frac{P(X \le t+x) - P(X < t)}{1 - P(X \le t)}$$

$$= \frac{1 - e^{-\lambda(t+x)} - (1 - e^{-\lambda t})}{1 - (1 - e^{-\lambda t})}$$

$$= \frac{1 - e^{-\lambda t} e^{-\lambda x} - 1 + e^{-\lambda t}}{e^{-\lambda t}}$$

$$= \frac{e^{-\lambda t}(1 - e^{-\lambda x})}{e^{-\lambda t}} = 1 - e^{-\lambda x} = P(X \le x)$$

Propriedade sem memória (memoryless)

Propriedade Memoryless

- Distribuição exponencial é a única distribuição contínua que tem a propriedade memoryless

- Por sua vez, distribuição geométrica é a única discreta que tem a propriedade memoryless

Outras Distribuições Contínuas

- **Weibull**
- **Lognormal**
- **Pareto**

Distribuição de Weibull

A VA contínua *T* tem uma distribuição de *Weibull* se:

$$f(t) = \lambda \alpha t^{\alpha - 1} e^{-\lambda t^{\alpha}}$$

$$F(t) = 1 - e^{-\lambda t^{\alpha}}$$

Onde os parâmetros satisfazem

$$t \geq 0 \quad \lambda > 0 \; \alpha > 0$$

Distribuição Lognormal

Uma VA X tem uma *distribuição lognormal* se a VA $Y = \ln(X)$ (ou $X = e^{Y}$) tem uma distribuição normal com parâmetros μ e σ

$$f(x;\mu,\sigma) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma x} e^{-[\ln(x)-\mu]^2 / (2\sigma^2)} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Muito utilizada para modelar duracao de sessao de usuarios em servicos web

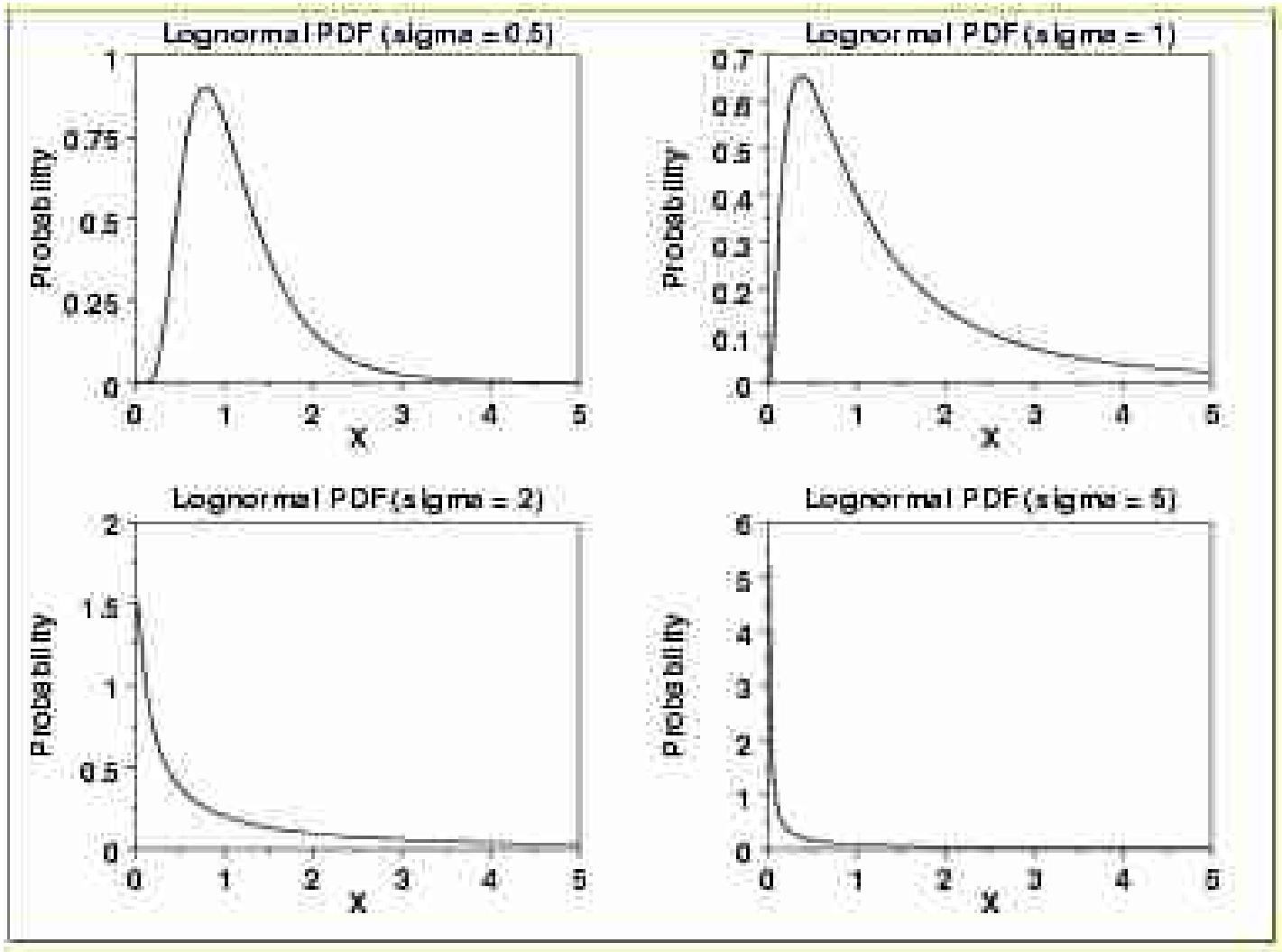
Média e Variância

A média e variância de uma VA X que tem uma distribuição lognormal são:

$$E(X) = e^{\mu+\sigma^2/2} \qquad V(X) = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right)$$

Distribuição Lognormal

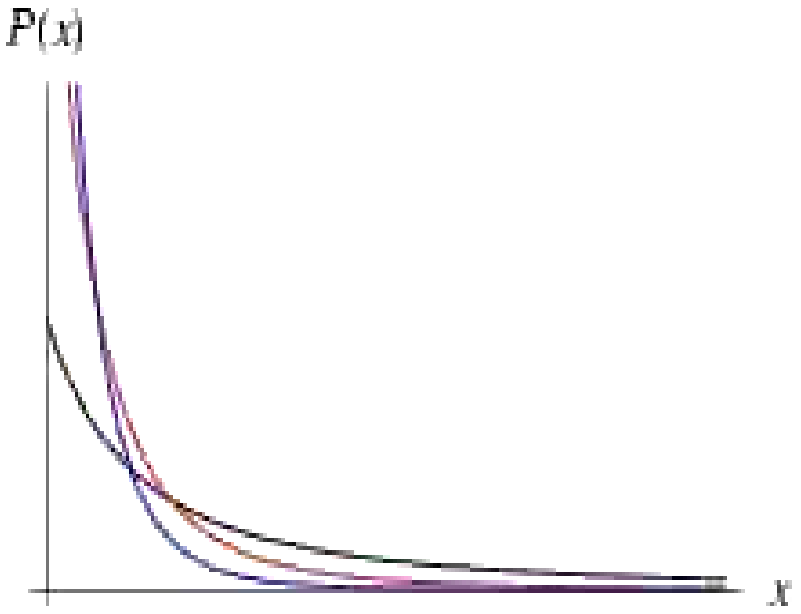
$\mu=1$

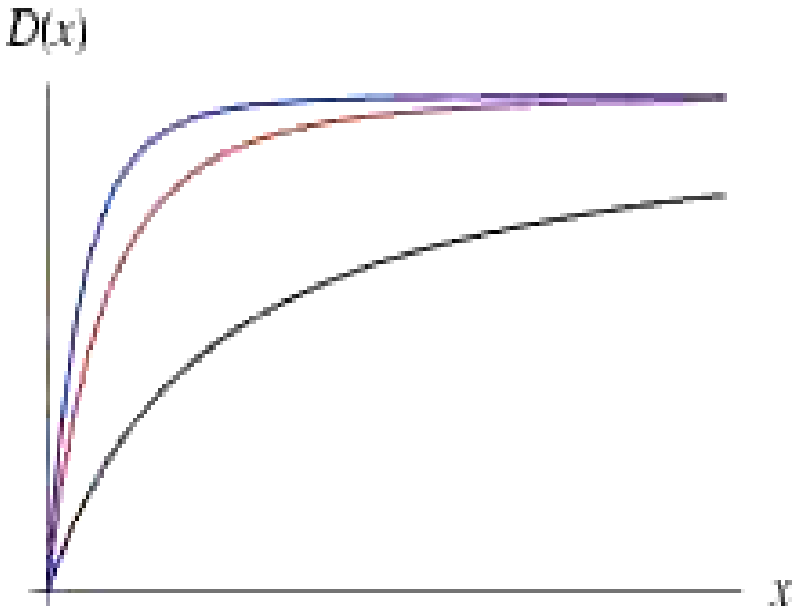


Distribuição de Pareto

Uma das distribuições *heavy tailed.*

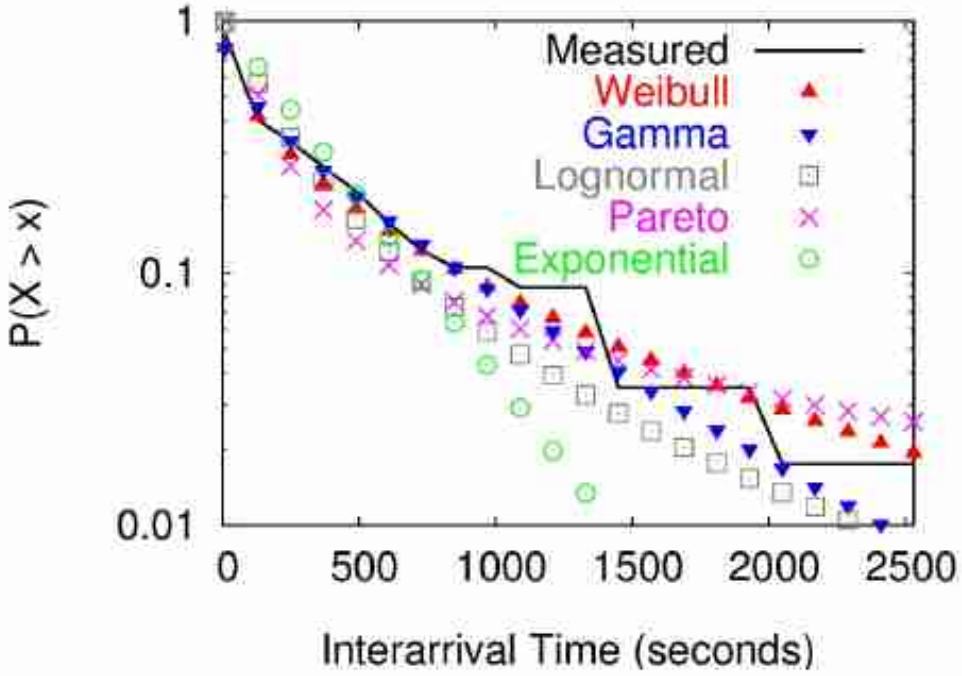
$$f(x) = \frac{ab^a}{x^{(a+1)}} = ab^a x^{-(a+1)} \qquad x \geq 1$$





Session Arrival Process

- Depends on the workload and on file size range



eTeach

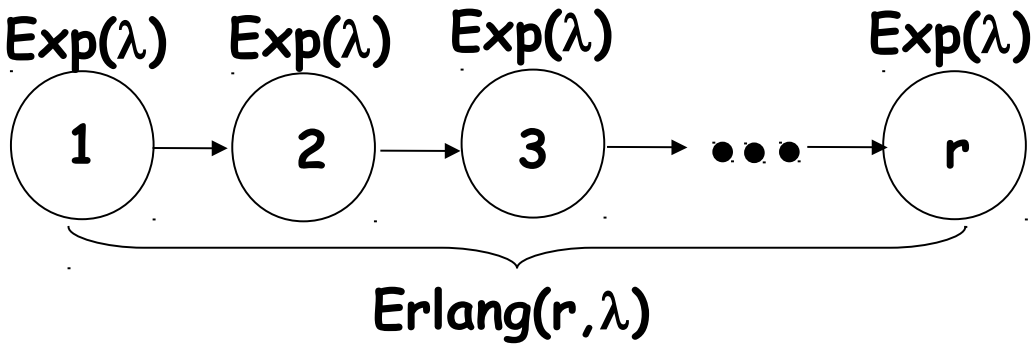
- Best Fitted Distribution of Session Inter-Arrival Times:
  - eTeach: Weibull and Lognormal
  - TV/UOL and Radio/UOL : Exponential
  - ISP/Audio: Pareto (body and tail)

Erlang: Soma de Exponenciais

- Genericamente: X_1, X_2, ... X_r, todas independentes e seguindo exponencial(λ):

$$Z = X_1 + X_2 + \ldots X_r \sim \text{Erlang de n estágios}$$

- Ex: tempo de processamento dividido em várias (r) etapas. A duração de cada etapa é exponencialmente distribuída com mesmo λ



- Se $X_i \sim$ exponencial (λ_i), onde λ_i são diferentes

$Z = X_1 + X_2 + \ldots X_r \sim$ Hipoexponencial

Distribuição de Erlang

- **As pdf e CDF de uma variavel X que tem distribuicao Erlang com parametros λ e r são:**

$$f(x) = \frac{\lambda^r x^{r-1} e^{-\lambda x}}{(r-1)!} \qquad F(x) = 1 - \sum_{k=0}^{r-1} \frac{(\lambda x)^k}{k!} e^{-\lambda x}$$

for x > 0 and r = 1, 2 , …

Exercícios

- **O tempo de CPU de um *query* típico medida em ms segue uma distribuição de Erlang de três estágios com $\lambda = 0.5$. Determine qual a probabilidade que a demanda de CPU da *query* excederá 1 milisegundo.**

- **O tempo de vida em dias de um componente de software é modelado por uma distribuição de Weibull com $\alpha = 2$. A partir de um grande número de componentes, foi observado que 15% dos componentes que duraram mais de 90 dias falharam antes de 100 dias. Determine o parâmetro λ**

Solução #1

O tempo de CPU de um *query* típico medida em ms segue uma distribuição de Erlang de três estágios com λ = ½. Determine qual a probabilidade que a demanda de CPU da *query* excederá 1 milisegundo.

$$F(x) = 1 - \sum_{k=0}^{r-1} \frac{(\lambda x)^k}{k!} e^{-\lambda x}$$

r = 3 e λ = ½

$$F_X(x) = 1 - \left(\sum_{k=0}^{2} \frac{(\frac{1}{2}x)^k}{k!}\right) e^{-\frac{1}{2}x}$$

$$P(X > 1) = 1 - P(X \le 1) = 1 - F_X(1)$$

$$= (1 + \frac{1}{2} + \frac{1}{8}) e^{(-\frac{1}{2})} = 0.9856$$

Solução #2

- **O tempo de vida em dias de um componente de software é modelado por uma distribuição de Weibull com $\alpha = 2$. A partir de um grande número de componentes, foi observado que 15% dos componentes que duraram mais de 90 dias falharam antes de 100 dias. Determine o parâmetro λ.**

$$F(x) = 1 - e^{-\lambda x^{\alpha}}$$

$$F(x) = 1 - e^{-\lambda x^{2}}$$

$$P(X < 100 \mid X > 90) = 0.15$$

$$= \frac{P(90 < X < 100)}{P(X > 90)}$$

$$= \frac{F(100) - F(90)}{1 - F(90)}$$

$$= \frac{e^{-\lambda (90)^2} - e^{-\lambda (100)^2}}{e^{-\lambda (90)^2}} = 0.15$$

$$\lambda = 0.00008554$$

Distribuição dos Mínimos

- Sistema composto de n componentes. Sistema funciona se todos componentes estão operando corretamente
- Tempo de falha : $X_1, X_2, \ldots, X_n$ exponencial (λ)
- Tempo de de vida do sistema $Z = \min (X_1, X_2, \ldots, X_n)$

  $P(Z \leq z) = P$ (pelo menos um $X_i \leq z$) = ?

  P (exatamente um $X_i \leq z$) = ?

Distribuição dos Mínimos

- Sistema composto de n componentes. Sistema funciona se todos componentes estão operando corretamente
- Tempo de falha : X_1, X_2, ..., X_n exponencial (λ)
- Tempo de de vida do sistema Z = min (X_1, X_2, ..., X_n)

  P(Z ≤ z) = P (pelo menos um $X_i \leq z$) = ?

  P (exatamente um $X_i \leq z$) = ?

$$P(\text{exatamente um } X_i \leq z) = \binom{n}{1} F_X(z)\left(1 - F_X(z)\right)^{n-1}$$

$$= \binom{n}{1} \left(1 - e^{-\lambda z}\right)\left(1 - (1 - e^{-\lambda z})\right)^{n-1}$$

Distribuição dos Mínimos

- $P(Z \leq z) = P$ (pelo menos um $X_i \leq z$)

Distribuição dos Mínimos

- **P(Z ≤ z) = P (pelo menos um $X_i \leq z$)**

$$P(\text{pelo menos um } X_i \leq z) = \sum_{j=1}^{n} \binom{n}{j} \left(F_X(z)\right)^j \left(1 - F_X(z)\right)^{n-j}$$

$$= \sum_{j=1}^{n} \binom{n}{j} \left(1 - e^{-\lambda z}\right)^j \left(e^{-\lambda z}\right)^{n-j}$$

$p = (1 - e^{-\lambda z})$

Z tem distribuição exponencial com parâmetro λn

$$= \sum_{j=1}^{n} \binom{n}{j} p^j (1-p)^{n-j}$$

$$= \sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} - \binom{n}{0} p^0 (1-p)^n$$

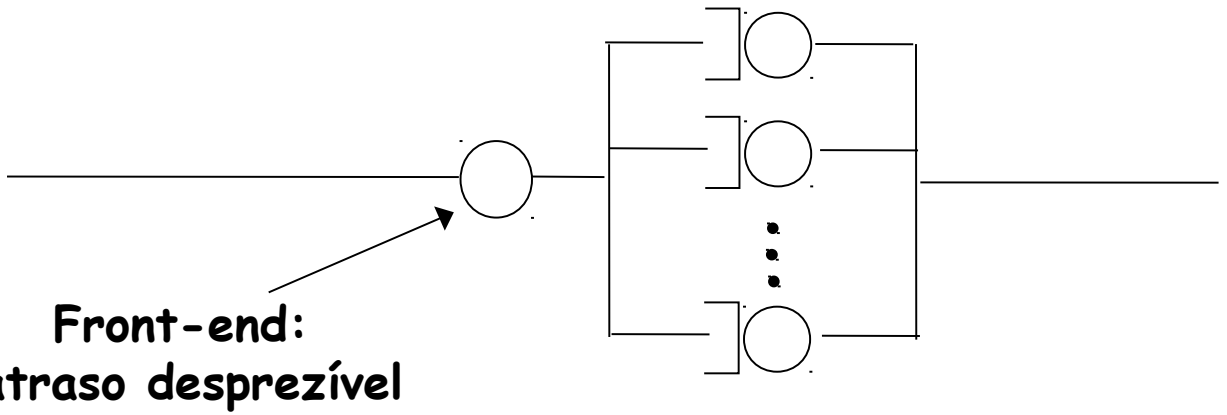
$$= 1 - (1-p)^n = 1 - \left(1 - \left(1 - e^{-\lambda z}\right)\right)^n = 1 - e^{-\lambda n z}$$

Distribuição dos Máximos

- n tarefas independentes : X_1, X_2,....., X_n: exponencial (λ)
- Tempo de resposta = tempo de execução da tarefa mais longa

$$Z = \max (X_1, X_2, \ldots, X_n)$$

  - Ex: tempo de resposta de máquina de busca composta de n processadores executando em paralelo. Cada máquina processa consulta em uma partição do dicionário



Distribuição dos Máximos

- **n tarefas independentes : X_1, X_2,....., X_n: exponencial (λ)**
- **Tempo de resposta = tempo de execução da tarefa mais longa**

$$Z = \max (X_1, X_2, \ldots, X_n)$$

$$\begin{aligned} P(Z \le z) &= P(\max(X_i) \le z) \\ &= P(X_1 \le z \cap X_2 \le z \cap \ldots \cap X_n \le z) \\ &= P(X_n \le z)P(X_2 \le z)\ldots P(X_n \le z) \\ &= (1-e^{-\lambda z})(1-e^{-\lambda z})\ldots(1-e^{-\lambda z}) = (1-e^{-\lambda z})^n \end{aligned}$$

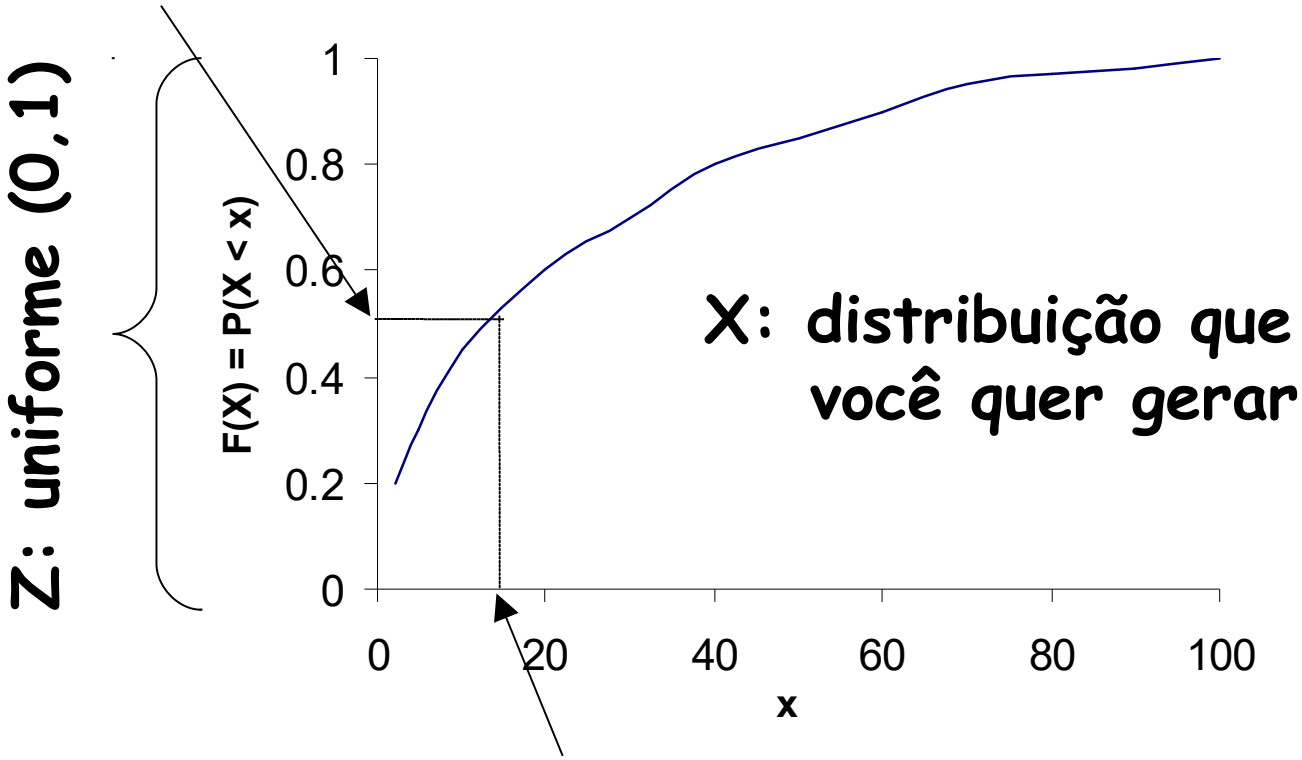
Gerando Distribuições

- Como gerar amostras de uma distribuição a partir de um gerador de números aleatórios uniformemente distribuídos

  (Unix: random(), drand48())?

Gerando Distribuições

Gerador de números aleatórios
retorna valor entre 0 e 1. Ex: 0.52



Aplicando o número aleatório na função inversa de F(X), consegue-se gerar um ponto amostral

Gerando Distribuições

- **Teorema da Transformada Inversa: Seja X uma variavel contínua com CDF F(x). Então:**

  Z = F(X) ~ uniforme (0,1)

Prova:

$$G(Z) = P(Z \le z) = P(F(X) \le z)$$

$$= P(Z \le F^{-1}(z)) = F(F^{-1}(z)) = z$$

Gerando Distribuições

Ex: geração de amostras de uma distribuição exponencial

$Z = F(X) = 1 - e^{-\lambda x}$ (CDF da exponencial)

$X' = F^{-1}(X) =$

$X' = - 1/\lambda \ln(1 - Z)$, onde $Z \sim$ uniforme(0,1)

$F(Z \leq z) = z$

Qual a distribuição de X' ?

Gerando Distribuições

Ex: geração de amostras de uma distribuição exponencial

$F(X) = 1 - e^{-\lambda x}$

$X' = F^{-1}(X) = - 1/\lambda \ln(1 - Z)$, onde $Z \sim$ uniforme(0,1)

$$F(Z \leq z) = z$$

$$F(X') = P(X' \leq y) = P(- 1/\lambda \ln(1 - Z) \leq y)$$

$$= P (\ln(1 - Z) \geq -\lambda y)$$

$$= P(1 - Z \geq e^{-\lambda y})$$

$$= P(Z \leq 1 - e^{-\lambda y}) = 1 - e^{-\lambda y}$$

$\Rightarrow X'$ é exponencial

O mesmo procedimento pode ser utilizado para gerar amostras de diferentes distribuições, partindo da inversa da CDF da distribuição desejada