

# População e Amostra

- População (ou universo): todos os  **$N$**  membros de uma classe ou grupo.
  - Ex.: todos os processos executados numa máquina durante o período que esteve ativa.
- Amostra é uma parte da população, denotada por  **$n$** .
  - Ex.: todos processos executados pela máquina em 18/03/2005

## A interpretação da frequência e a distribuição da amostra

- Quando se faz inferências estatísticas do ponto de vista da frequência, nós assumimos que nossos dados são amostras de uma população inteira.
- A população é descrita pela média e variância da população que são desconhecidas.
- A amostra é descrita pela média e variância da amostra.
  - A média e variância da amostra proveêm estimativas sobre a média e variância da população inteira.
  - As estimativas são conhecidas com um grau de incerteza.

# Variável Aleatória

- Variável que assume valores de acordo com uma certa probabilidade.
- Variável usualmente denotada por letras maiúsculas, e valores particulares por letra minúscula.
- Exemplos:
  - Atraso numa rede
  - Tempo de execução de uma transação de consulta

# Função de Distribuição Cumulativa (CDF cumulative distribution function)

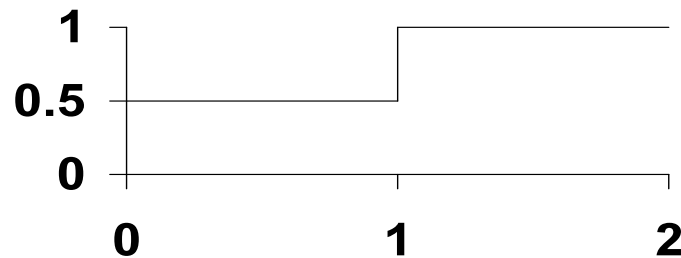
- Mapeia um valor para uma probabilidade cujo resultado é menor ou igual a ***a***:

$$F_x(a) = P(x \leq a)$$

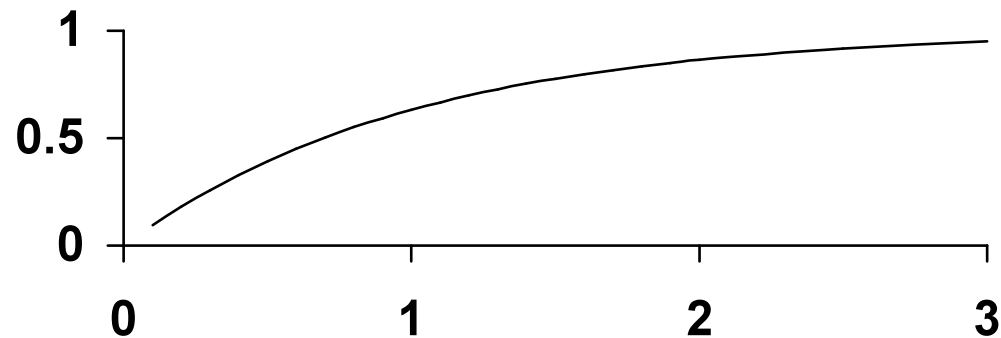
- Válida para variáveis contínuas e discretas
- Monotonicamente crescente
- Fácil de especificar, calcular, medir...

# Exemplos

- Jogada de uma moeda ( $T = 1$ ,  $H = 2$ ):



- Tempo entre chegadas de pacotes Exponencial:



# Função de Densidade de Probabilidade (pdf)

- Derivada da CDF (contínua):

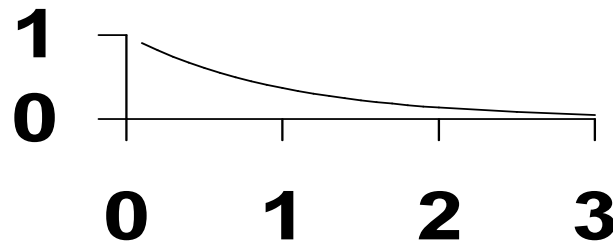
$$f(x) = \frac{dF(x)}{dx}$$

- Útil para determinar intervalos de probabilidades :

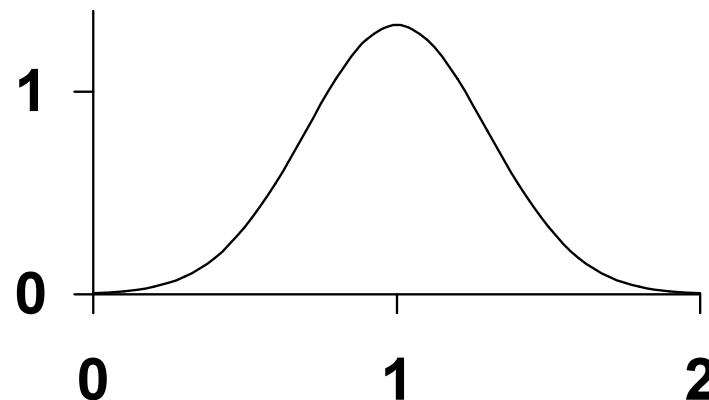
$$\begin{aligned} P(x_1 < x \leq x_2) &= F(x_2) - F(x_1) \\ &= \int_{x_1}^{x_2} f(x) dx \end{aligned}$$

# Exemplos de pdf

- Tempo entre chegadas exponential:

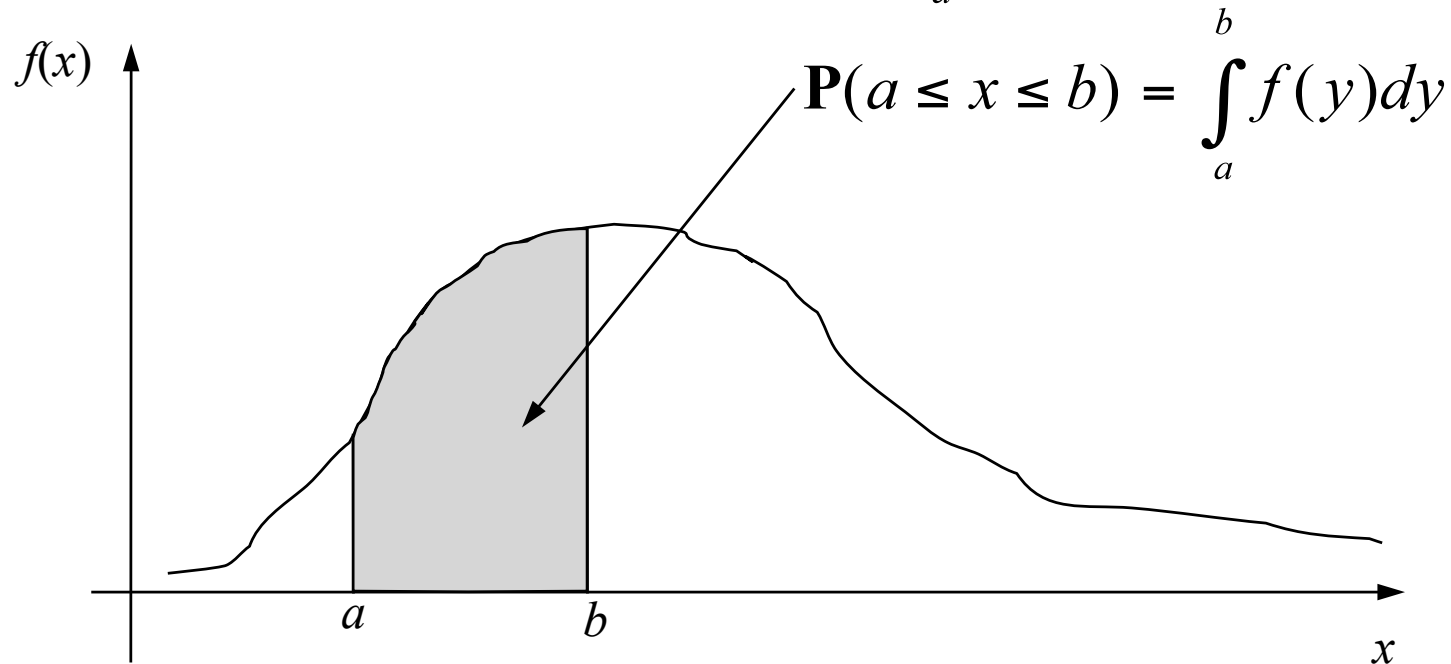


- Distribuição Normal (Gauss):



# CDF's and PDF's

For an interval  $I = [a, b]$ :  $\mathbf{P}(X \in I) = \int_a^b f(y)dy = F(b) - F(a)$





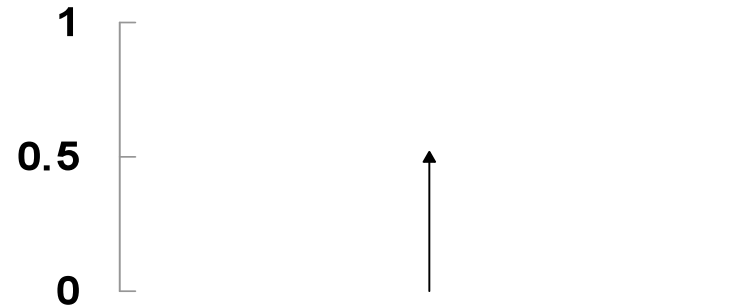
# Função de Massa de Probabilidade (pmf)

- CDF não são diferenciáveis para variáveis aleatórias discretas
- pmf serve com substituto:  $f(x_i) = p_i$  onde  $p_i$  é a probabilidade que  $x$  irá assumir o valor  $x_i$

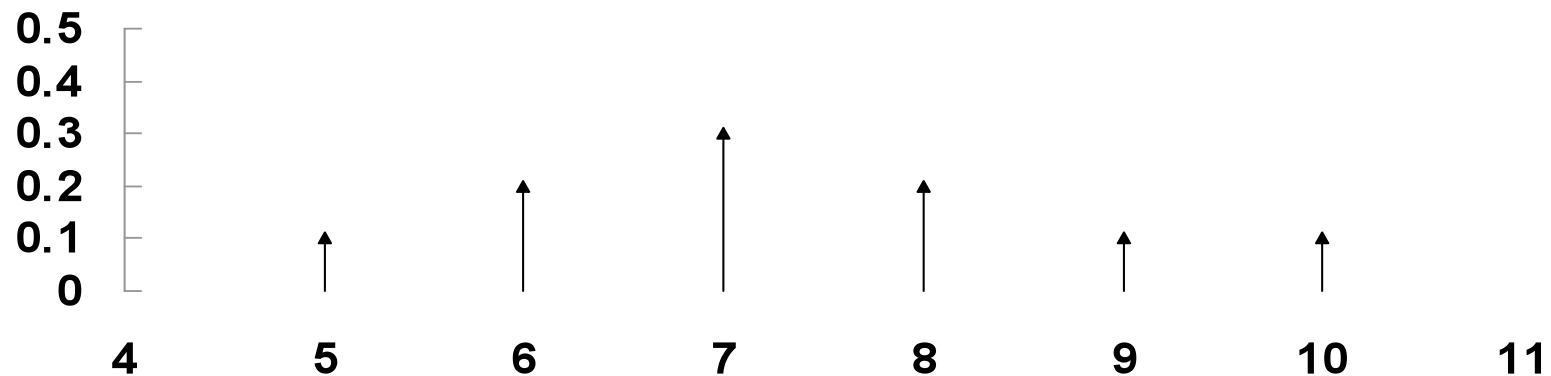
$$\begin{aligned} P(x_1 \leq x \leq x_2) &= F(x_2) - F(x_1) \\ &= \sum_{x_1 < x_i \leq x_2} p_i \end{aligned}$$

# Exemplos de pmf

- Jogada de moeda:



- Tamanho típico de uma turma de Pós:



# Expectância, Média ou Esperança Matemática

- Média

$$\mu = E(x) = \sum_{i=1}^n p_i x_i = \int_{-\infty}^{\infty} x f(x) dx$$

- Somatório se discreto
- Integral se contínuo

## Variância & Desvio Padrão

- $\text{Var}(x) = E[(x - \mu)^2] = \sum_{i=1}^n p_i (x_i - \mu)^2$   
 $= \int_{-\infty}^{+\infty} (x_i - \mu)^2 f(x) dx$
- Usualmente denotada por  $\sigma^2$
- Raíz quadrada  $\sigma$  é chamado de ***desvio padrão***

# Coeficiente de Variação (C.V.)

- Quociente do desvio padrão pela média:

$$C.V. = \frac{\sigma}{\mu}$$

- Indica quão bem a média representa a variável

# Covariância

- Dados  $x$ ,  $y$  com médias  $\mu_x$  e  $\mu_y$ , sua covariância é:

$$\begin{aligned}\text{Cov}(x, y) &= \sigma_{xy}^2 = E[(x - \mu_x)(y - \mu_y)] \\ &= E(xy) - E(x)E(y)\end{aligned}$$

- Alta covariância implica que  $y$  afasta da média sempre que  $x$  também o faz.

# Covariância

- Para variáveis independentes,

$$E(xy) = E(x)E(y)$$

$$\text{então } \text{Cov}(x,y) = 0$$

- Reverso não é verdade:  $\text{Cov}(x,y) = 0$  não implica em independência.
- Se  $y = x$ , covariância reduz-se à variância

# Coeficiente de Correlação

- Covariância normalizada:

$$\text{Correlação}(x, y) = \rho_{xy} = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y}$$

- Sempre varia entre -1 e 1
- Correlação de 1  $\Rightarrow x \sim y$ ,  $-1 \Rightarrow x \sim \frac{1}{y}$



# Média e Variância de Somas

- Para qq variável aleatória,

$$E(a_1x_1 + a_2x_2 + \cdots + a_kx_k)$$

$$= a_1E(x_1) + a_2E(x_2) + \cdots + a_kE(x_k)$$

- Para variáveis independentes,

$$\text{Var}(a_1x_1 + a_2x_2 + \cdots + a_kx_k)$$

$$= a_1^2 \text{Var}(x_1) + a_2^2 \text{Var}(x_2) + \cdots + a_k^2 \text{Var}(x_k)$$

# Quantil

- Valor de  $x$  no qual a CDF assume um valor  $\alpha$  é chamado *a-quantil* or *100 $\alpha$ -percentil*, denoted by  $x_\alpha$ .

$$P(x \leq x_\alpha) = F(x_\alpha) = \alpha$$

- Se 90-esimo percentil score no GRE foi 1500, então 90% da população obteve 1500 ou menos.

# Mediana

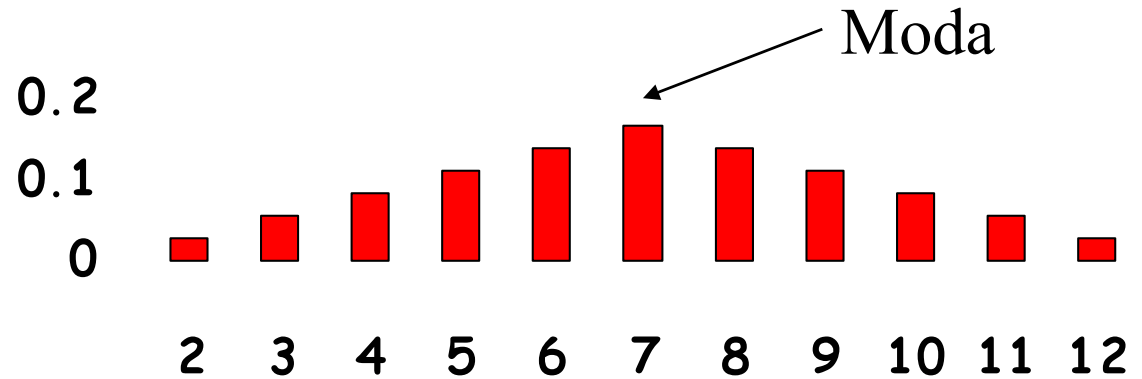
- 50-percentil (0.5-quantil) de uma variável aleatória
- Alternativa a média
- Por definição, 50% da população é sub-mediana, 50% super-mediana
  - Muitos queries rápidos (lentos)
  - Muitas pessoas ricas (pobres)?

# Moda

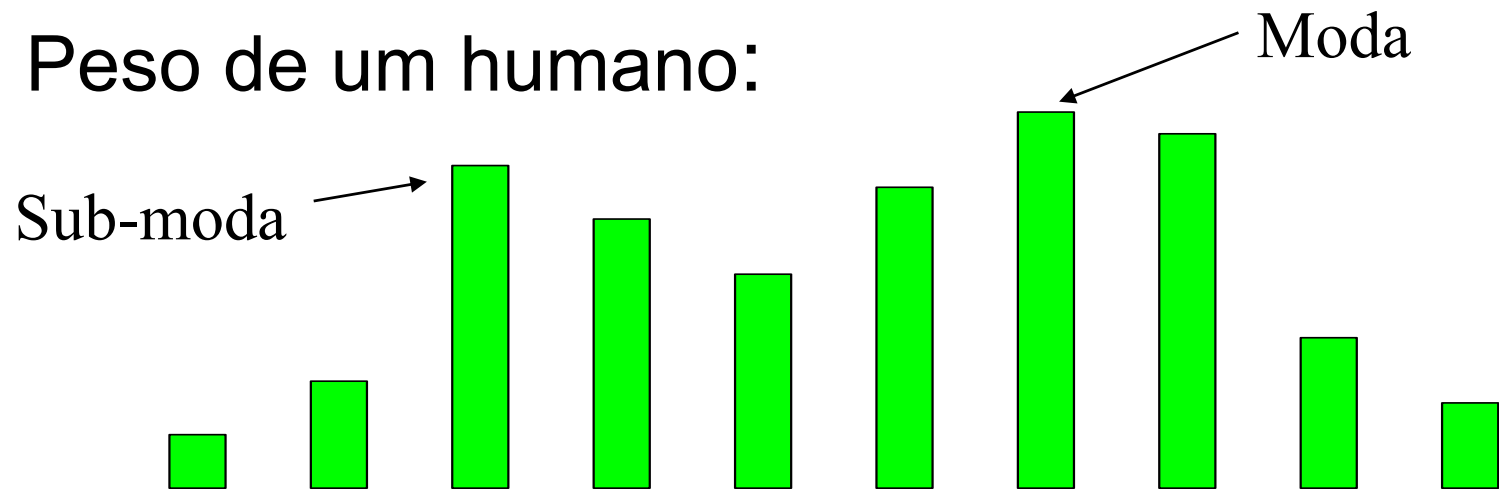
- Valor mais provável, i.e.,  $x_j$  com a maior probabilidade  $p_j$ , or  $x$  no qual pdf/pmf é máximo
- Não necessariamente definido (empate)
- Algumas distribuições são bi-modais (ex: altura dos humanos tem uma moda para homens e uma moda para mulheres.)

# Exemplos de Moda

- Dois dados:



- Peso de um humano:



# 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  $\mu$ , desvio padrão  $\sigma$

# 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  $\alpha$ -quantil de uma normal unitária  $z \sim N(0, 1)$  é denotado por  $z_\alpha$  tal que

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

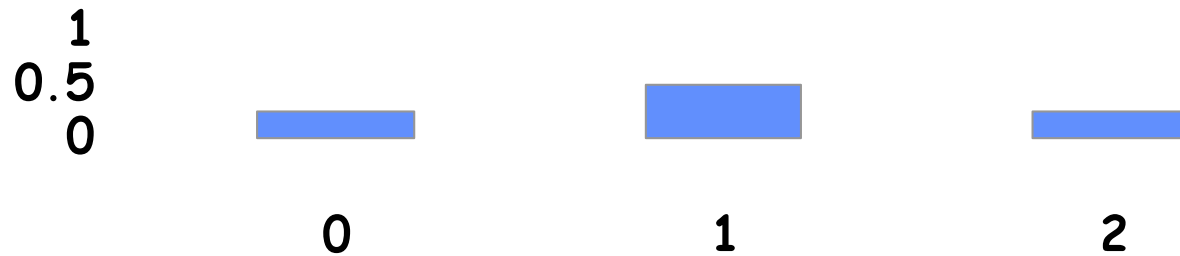
# Por que uma Distribuição de Gauss é tão Popular?

- Se  $x_i \sim N(\mu_i, \sigma_i)$  e todos  $x_i$  independentes, então  $\sum \alpha_i x_i$  é normal com média  $\sum \alpha_i \mu_i$  e variância  $\sigma^2 = \sum \alpha_i^2 \sigma_i^2$
- A soma/media de um número grande de observações independentes de qualquer distribuição é uma distribuição normal (Teorema Central do Limite)
  - ⇒ Erros experimentais, tipicamente a soma de varios componentes, podem ser modelados como uma distribuição normal.

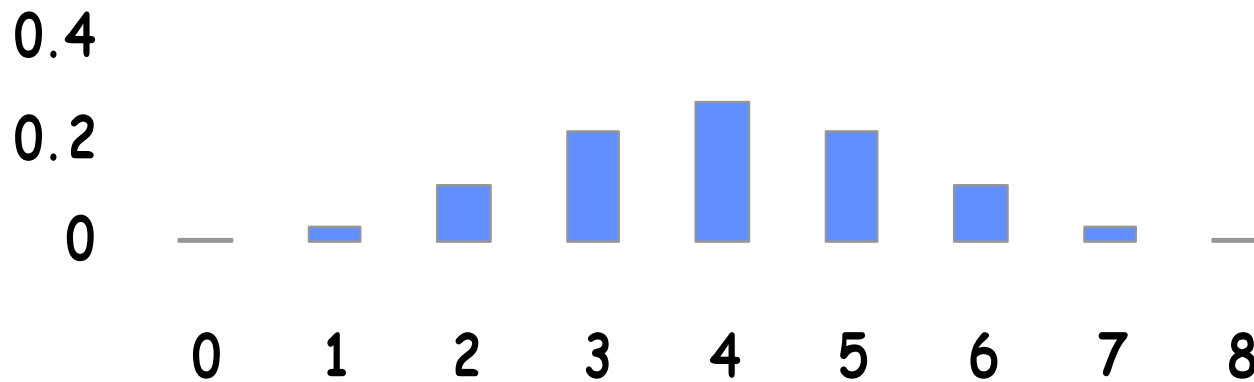


# Teorema Central do Limite

- Sum of 2 coin flips (H=1, T=0):



- Sum of 8 coin flips:



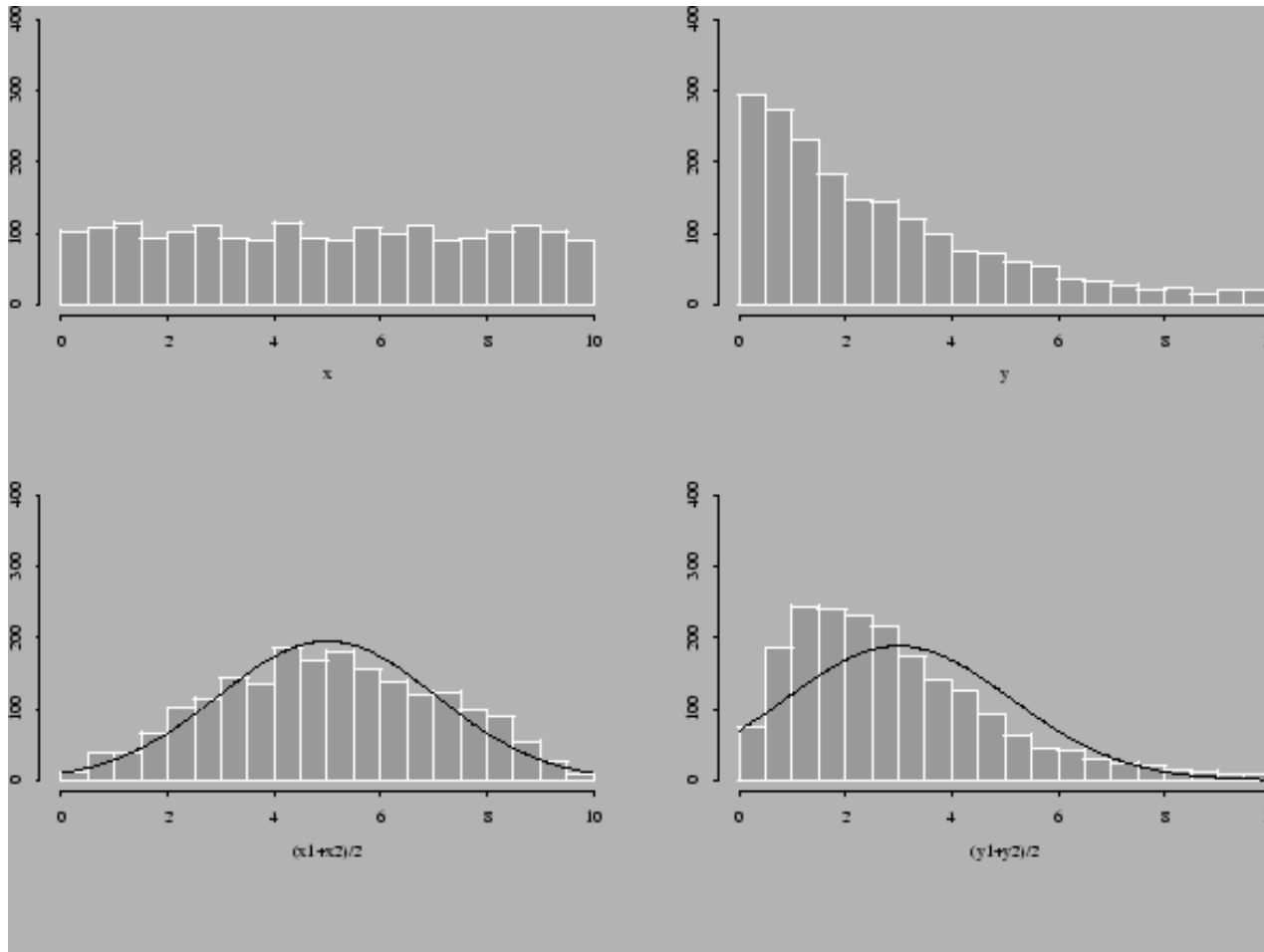
# Teorema Central do Limite

- Se  $A$  é uma amostra de tamanho  $n$  de uma população, com distribuição arbitrária, mas média  $\mu$  e desvio padrão  $\sigma$ :

A distribuição da média amostral é aproximadamente Normal com média  $\mu$  e desvio padrão

$$\frac{\sigma}{\sqrt{n}}$$

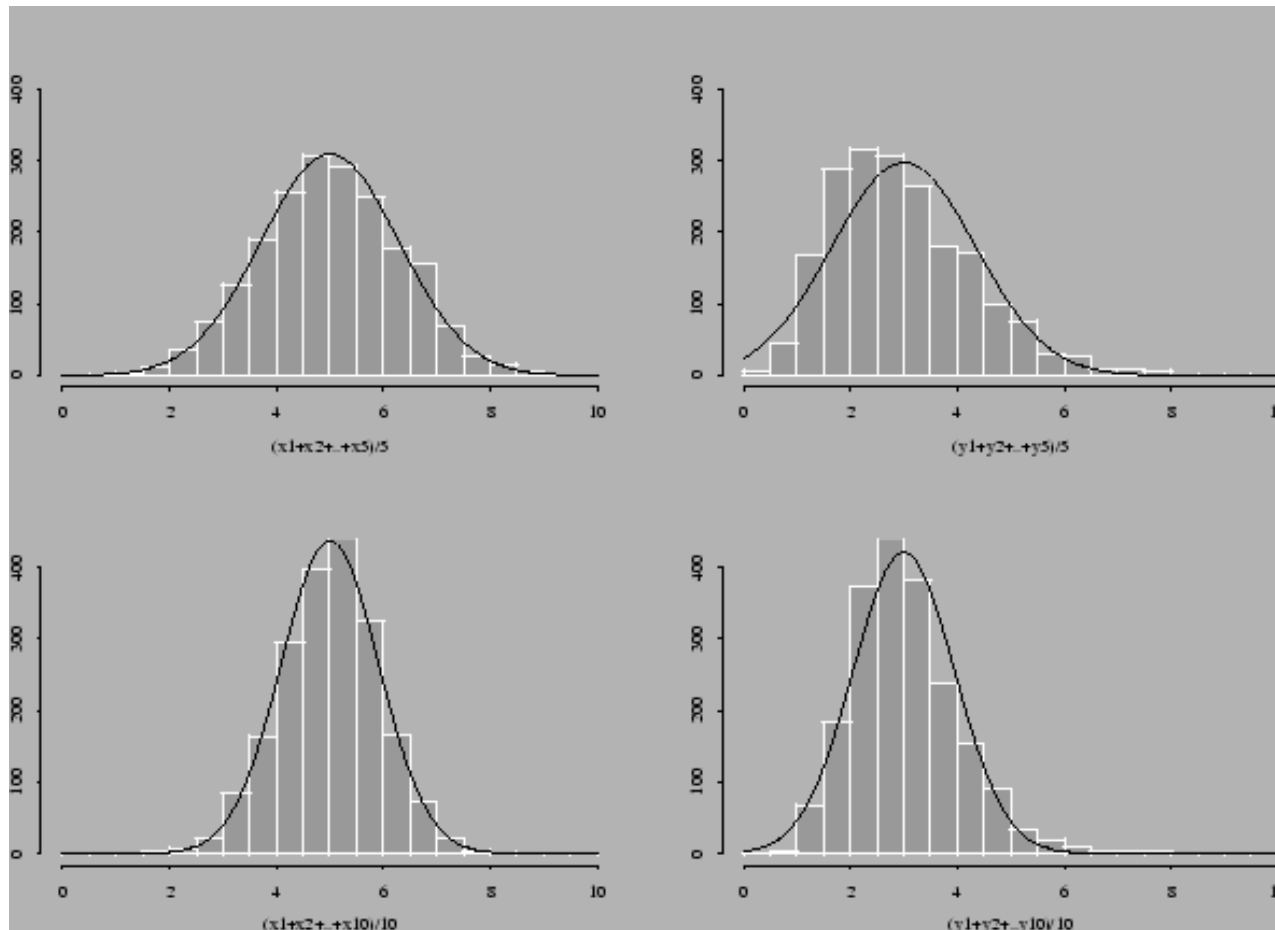
# Teorema Central do Limite



Duas  
distribuicoes  
arbitrarias

Distribuicao  
da media de  
uma amostra de  
2 elementos

# Teorema Central do Limite



Distribuição da  
média de uma  
amostra de 5  
elementos

Distribuição  
da média de  
uma amostra de  
10 elementos

Médias populacionais são 5 e 3

# Measured Data

But, we don't know  $F(x)$  – all we have is a bunch of observed values – *a sample*.

What is a sample?

- Example: How tall is a human?
  - Could measure every person in the world (actually even that's a sample)
  - Or could measure every person in this room
- Population has *parameters*: fixed, typically unknown
- Sample has *statistics*
  - Drawn from population
  - Inherently erroneous

# Central Tendency

- Sample mean –  $\bar{x}$  (arithmetic mean)
  - Take sum of all observations and divide by the number of observations
- Sample median
  - Sort the observations in increasing order and take the observation in the middle of the series
- Sample mode
  - Plot a histogram of the observations and choose the midpoint of the bucket where the histogram peaks

# Indices of Dispersion

- Measures of how much a data set varies
  - Range
  - Sample variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- And derived from sample variance:
  - Square root -- standard deviation,  $s$
  - Ratio of sample mean and standard deviation – CV  
 $s / \bar{x}$
- Percentiles
  - Specification of how observations fall into buckets

# Interquartile Range

- Yet another measure of dispersion
- The difference between Q3 and Q1
- Semi-interquartile range -

$$SIQR = \frac{Q_3 - Q_1}{2}$$

- Often interesting measure of what's going on in the middle of the range



# Determining a Distribution for a Data Set

- If a data set has a common distribution, that's the best way to summarize it
  - Saying a data set is uniformly distributed is more informative than just giving its sample mean and standard deviation
- So how do you determine if your data set fits a distribution?
  - Plot a histogram
  - Quantile-quantile plot
  - Statistical methods

# Quantile-Quantile Plots

- Most suitable for small data sets
- Basically -- guess a distribution
- Plot where quantiles of data should fall in that distribution
  - Against where they actually fall in the sample
- If plot is close to linear, data closely matches that distribution

# Obtaining Theoretical Quantiles

- We need to determine where the quantiles should fall for a particular distribution
- Requires inverting the CDF for that distribution

$$q_i = F(x_i) \qquad x_i = F^{-1}(q_i)$$

- Then determining quantiles for observed points
- Then plugging in quantiles to inverted CDF

# Inverting a Distribution

- Many common distributions have already been inverted (how convenient...)
- For others that are hard to invert, tables and approximations are often available (nearly as convenient)


# Is Our Example Data Set Normally Distributed?

- Our example data set was  
-17, -10, -4.8, 2, 5.4, 27, 84.3, 92, 445, 2056
- Does this match the normal distribution?
- The normal distribution doesn't invert nicely
  - But there is an approximation for  $N(0,1)$ :  
$$x_i = 4.91 \left[ q_i^{0.14} - (1 - q_i)^{0.14} \right]$$
  - Or invert numerically

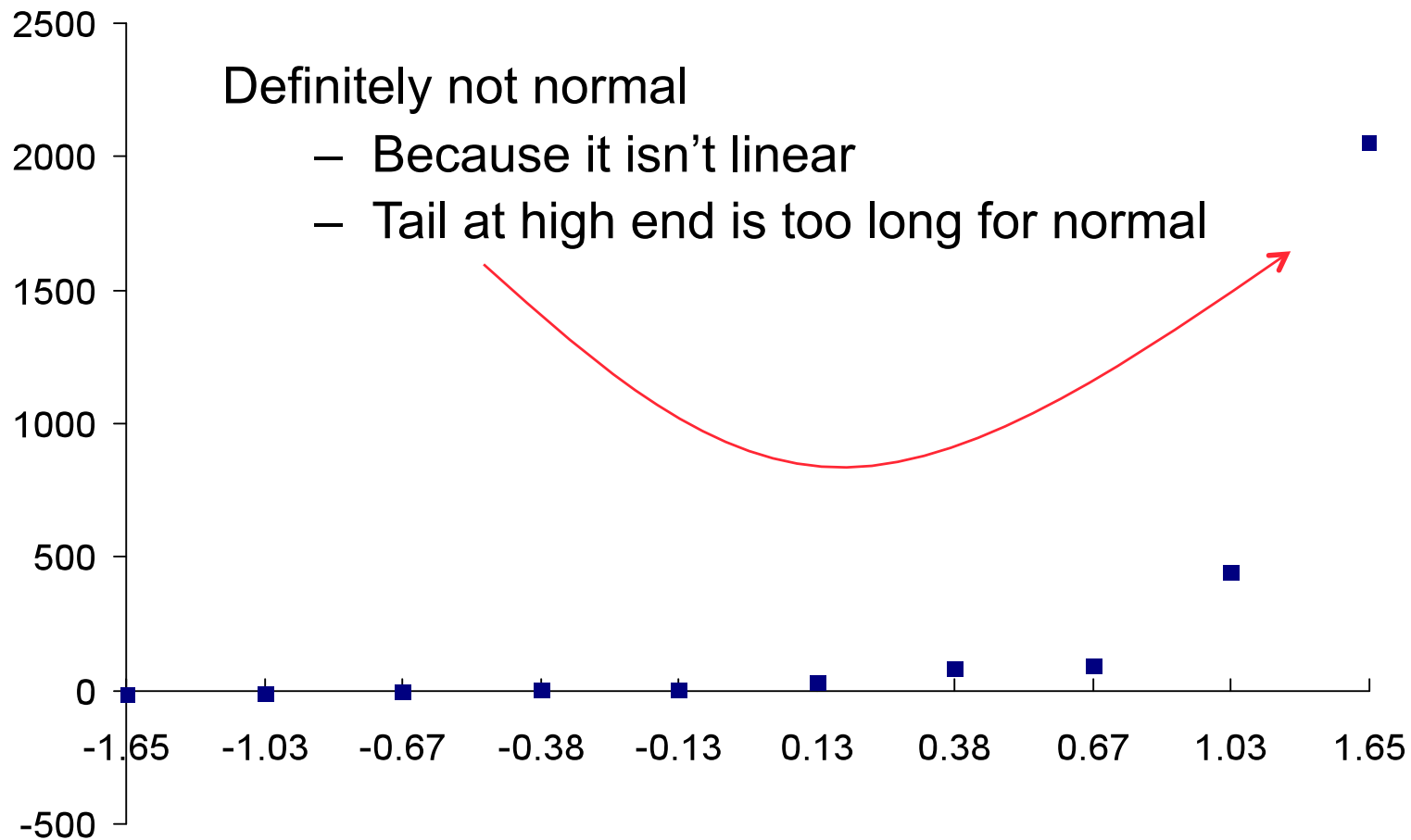
# Data For Example Normal Quantile-Quantile Plot

$i$	$q_i = 100(i-0.5)/n$	$y_i$	$x_i$
1	0.05	-17	-1.64684
2	0.15	-10	-1.03481
3	0.25	-4.8	-0.67234
4	0.35	2	-0.38375
5	0.45	5.4	-0.1251
6	0.55	27	0.1251
7	0.65	84.3	0.383753
8	0.75	92	0.672345
9	0.85	445	1.034812
10	0.95	2056	1.646839

**Sample  
quantile**



# Example Normal Quantile-Quantile Plot



# Is Our Example Data Set Normally Distributed?

- As diferenças entre os valores medidos de um sistema e os preditos por um modelo (erros) foram quantificadas para 8 predições.

Os valores encontrados foram:

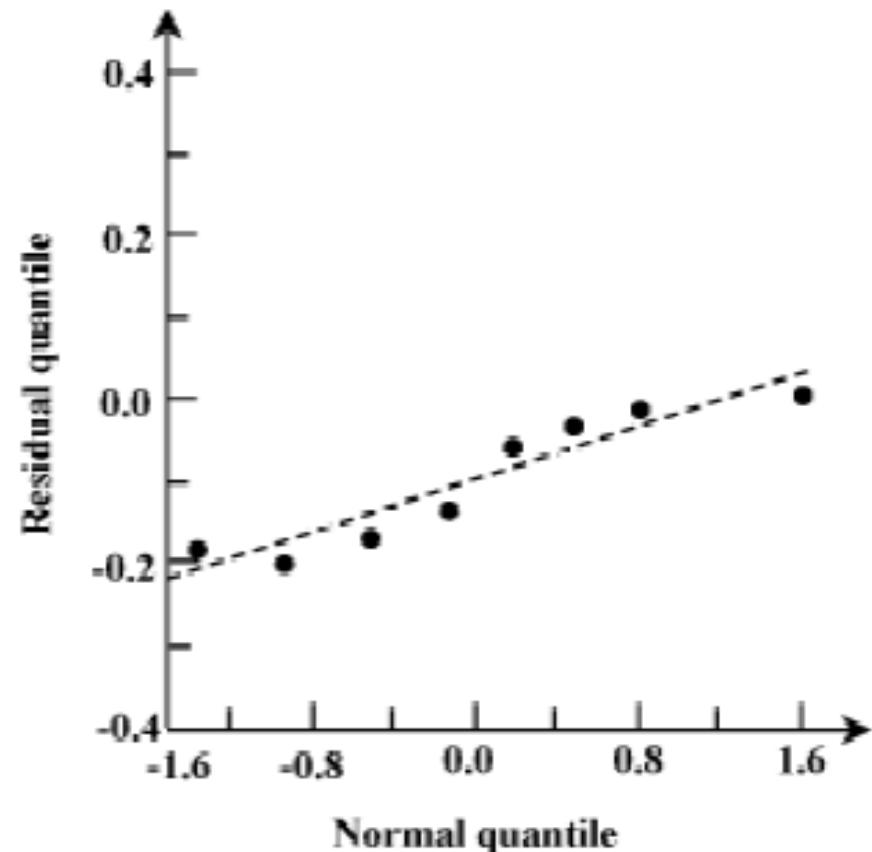
-0.04, -0.19, 0.14, -0.09, -0.14, 0.19, 0.04, 0.09

Estes erros são normalmente distribuídos?



# Is Our Example Data Set Normally Distributed?

$i$	$q_i = \frac{i-0.5}{n}$	$y_i$	$x_i$
1	0.0625	-0.19	-1.535
2	0.1875	-0.14	-0.885
3	0.3125	-0.09	-0.487
4	0.4375	-0.04	-0.157
5	0.5625	0.04	0.157
6	0.6875	0.09	0.487
7	0.8125	0.14	0.885
8	0.9375	0.19	1.535



# Desvios da Normal

