PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE ENGENHARIA
CURSO DE PÓS-GRADUAÇÃO LATO SENSU – ESPECIALIZAÇÃO
Engenharia de Processos e de Sistemas de Produção

# Estatística Aplicada à Engenharia

Prof. Hélio Radke Bittencourt

Curso de Especialização em
Engenharia de Processos e de Sistemas de Produção

Abril de 2014

# Sumário



# Calendário de aulas

Aula 1 25/04
Conceitos Básicos, Estatística Descritiva

Aula 2 26/04
Probabilidade, Amostragem e distribuições amostrais

Aula 3 10/05
Intervalos de Confiança, Cálculo do *n*

As aulas com o Professor Lori Viali ocorrerão em 09/05, 16/05 e 17/05.

# Cap. 1 CONCEITOS BÁSICOS

A **Estatística** pode ser definida como o conjunto de ferramentas para coleta, organização, análise e interpretação de dados experimentais. O objeto de estudo em Estatística é um conjunto de dados que pode constituir uma **população** ou uma **amostra**.

**População** é um conjunto finito ou infinito de elementos.

**Amostra** é um subconjunto da população. Geralmente buscamos amostras representativas. Uma **amostra representativa** é aquela que mantém as características da população.

Questões:

1) Qual o número aproximado de eleitores no Rio Grande do Sul? Qual o tamanho amostral das pesquisas eleitorais realizadas pelos institutos?

2) Como se chama a investigação de toda a população? Por que, por exemplo, o IBGE não realiza _______________ anualmente?

## 1.1 Tipos de Variáveis

Já vimos que a Estatística trabalha com conjunto de dados formado por elementos. Nós não trabalharemos diretamente com os elementos que formam o conjunto de dados, mas sim com suas características. **Variáveis** são características dos elementos que formam o conjunto de dados.

Vamos imaginar um carro e listar variáveis associadas a ele:

As variáveis podem ser classificadas em **qualitativas** ou **quantitativas**: as variáveis qualitativas expressam uma classificação em categorias e, por isso, também são chamadas de categóricas. As variáveis quantitativas expressam quantidades numéricas e se dividem em **discretas** e **contínuas**. As variáveis discretas assumem apenas determinados valores num dado conjunto enumerável, enquanto as variáveis contínuas podem assumir, ao menos teoricamente, qualquer valor num dado intervalo numérico.

Na prática todas as variáveis são discretas, devido à limitação dos instrumentos de mensuração.

As variáveis qualitativas ainda podem ser classificadas em **nominais** ou **ordinais**.

O **nível nominal de mensuração** é caracterizado por números que apenas diferenciam ou rotulam as categorias.

O **nível ordinal de mensuração** envolve números que, além de diferenciar, hierarquizam as categorias.

Figura – Resumo dos tipos de variáveis e escalas de mensuração

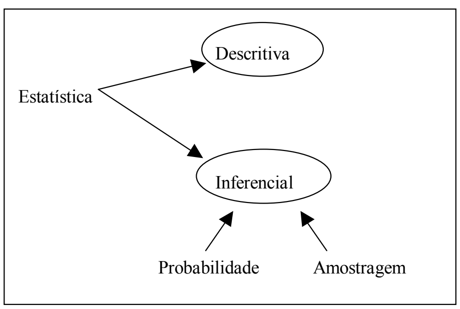
## 1.2 Estatística Descritiva e Inferencial

A estatística é um conjunto de ferramentas utilizadas para a coleta, tabulação, análise e interpretação de um conjunto de dados experimentais. A Estatística pode ser dividida em duas grandes áreas: **Descritiva** e **Inferencial**.

A **estatística descritiva** é aquela que costumamos encontrar com maior freqüência em jornais, revistas, relatórios, etc. Essa parte da estatística utiliza números para descrever fatos. Seu foco é a representação gráfica e o resumo e organização de um conjunto de dados, com a finalidade de simplificar informações. Nessa categoria se enquadram as médias salariais, taxas de inflação, índice de desemprego, etc.

A **estatística inferencial** consiste na obtenção de resultados que possam ser projetados para toda população a partir de uma amostra da mesma. Ela fundamenta-se na teoria da amostragem e no cálculo de Probabilidades. Essa é a área mais importante da Estatística.

Figura - Esquema geral de um curso de Estatística




**Probabilidade** é um ramo da Matemática que estuda fenômenos aleatórios.

**Amostragem** é o nome dado ao conjunto de procedimentos e técnicas para extração de elementos da população para compor a amostra.

# Cap. 2 ESTATÍSTICA DESCRITIVA

Apenas para lembrar: na **estatística descritiva** nos preocupamos em apresentar a informação de conjuntos de dados de forma resumida, mas, ao mesmo tempo, útil para o pesquisador.

## 2.1 Tabelas de freqüência

Tabelas de freqüência são encontradas em jornais informativos (Zero Hora, Correio do Povo, etc.), relatórios técnicos, monografias, dissertações, teses e revistas científicas. As tabelas de freqüência **simples** apresentam de forma concisa o número de ocorrências (absoluta e relativa) dos valores de uma variável

Uma tabela de freqüência genérica tem a seguinte configuração:

Tabela 1 – Tabela de freqüências genérica

| $i$ | $x_i$ | $f_i$ | $fr_i$ | $F_i$ | $Fr_i$ |
|---|---|---|---|---|---|
| 1 | $x_1$ | $f_1$ | $fr_1$ | $F_1$ | $Fr_1$ |
| 2 | $x_2$ | $f_2$ | $fr_2$ | $F_2$ | $Fr_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| k | $x_k$ | $f_k$ | $fr_k$ | $F_k$ | $Fr_k$ |
| | $\Sigma$ | n | 100,0% | - | - |

A notação utilizada é a seguinte:

**X** é uma variável qualquer

***x*** é um particular valor da variável X

**i** é um índice útil para enunciar as expressões matemáticas

**k** é o número de linhas da tabela

Os componentes da tabela de freqüências são:

**Freqüência absoluta ($f_i$):** número de ocorrências do valor $x_i$.

**Freqüência relativa ($fr_i$):** percentual de ocorrências do valor $x_i$

**Freqüência absoluta acumulada ($F_i$):** número de ocorrências até o valor $x_i$.

**Freqüência relativa acumulada ($Fr_i$):** percentual de ocorrências até o valor $x_i$

Como calcular as freqüências:

**Tabela de freqüência cruzada s**ão representações tabulares da freqüência de ocorrência de duas variáveis de maneira simultânea. São também chamadas de tabelas de contingência. Uma tabela cruzada tem *l* linhas e *c* colunas.

Exemplo – Tabela de contingência genérica

| X | Y | | | | |
|---|---|---|---|---|---|
| | $y_1$ | $y_2$ | | $y_c$ | |
| $x_1$ | $f_{11}$ | $f_{21}$ | $\cdots$ | $f_{1c}$ | $f_{1\bullet}$ |
| $x_2$ | $f_{12}$ | $f_{22}$ | $\cdots$ | $f_{2c}$ | $f_{2\bullet}$ |
| | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $x_l$ | $f_{l1}$ | $f_{l2}$ | $\cdots$ | $f_{lc}$ | $f_{l\bullet}$ |
| | $f_{\bullet 1}$ | $f_{\bullet 2}$ | $\cdots$ | $f_{\bullet c}$ | $f_{\bullet\bullet}$ |

## 2.2 Medidas de Tendência Central

São valores que trazem informação sobre a região em torno da qual os dados estão posicionados. As medidas de tendência central mais utilizadas são: Média, Mediana e Moda.

Para apresentar as medidas de tendência central vamos utilizar o seguinte exemplo.

**Banco de dados do ENADE 2011 (Cursos de Engenharia de Produção)**

### 2.2.1 – Média Aritmética ($\mu$ , $\overline{X}$ )

A **média aritmética** é definida como a soma de todas observações da variável X, dividida pelo número de elementos do conjunto de dados. Freqüentemente a média aritmética é o valor que melhor representa um conjunto de dados.

Quando os dados **não** estão organizados na forma de uma tabela de freqüências e, portanto, estão na forma **isolada**, as expressões genéricas para encontrar a média são:

População

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Amostra

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Quando os dados estão organizados na forma de uma tabela de freqüências deve-se **ponderar** os diferentes valores $x_i$ pelas respectivas freqüências $f_i$. Procedendo desta forma o cálculo da média aritmética torna-se mais simples e rápido.

População

$$\mu = \frac{\sum_{i=1}^{k} x_i \times f_i}{N}$$

Amostra

$$\overline{X} = \frac{\sum_{i=1}^{k} x_i \times f_i}{n}$$

### 2.2.2 – Mediana (Md)

A **mediana** é o valor que divide o conjunto de dados ordenado em duas partes com igual número de observações. Para calcular a mediana iremos utilizar uma nova notação. Seja $x_{[1]}, x_{[2]}, \ldots, x_{[n]}$ um conjunto de dados **ordenado** (ordem não-decrescente), onde o valor entre colchetes representa a posição no conjunto ordenado.

As expressões genéricas para encontrar a mediana são:

**n ímpar** **n par**

### 2.2.3 – Moda (Mo)

A **moda** é definida como o valor mais freqüente de um conjunto de dados. É possível que o conjunto seja **bimodal** (duas modas) ou até mesmo **multimodal** (três os mais modas).

$$Mo = \{x_i\} \text{com maior } f_i$$

**Exemplo – Encontrar as MTCs para o CPC contínuo e seus 8 componentes no EXCEL**

| |
|---|
| **= MÉDIA(??:??)** |
| **=MED(??:??)** |
| **=MODO(??:??)** |

**Considerações sobre as MTC**

1. A média é a MTC mais influenciada por valores extremos, entretanto é a medida mais "rica", porque considera todos valores do conjunto de dados.

2. A mediana não é afetada por valores extremos.

3. A moda é a MTC mais "pobre", porque considera apenas os valores mais freqüentes.

4. Existem outros tipos de média que não são tão afetadas por valores extremos, entretanto não toleram, por exemplo, o valor **zero**.

Média harmônica

$$\overline{X}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Média geométrica

$$\overline{X}_G = \sqrt[n]{x_1 \times x_2 \times \ldots \times x_n}$$

Pode-se estabelecer a seguinte relação entre as médias:

$$\overline{X}_h \leq \overline{X}_G \leq \overline{X}$$

**Exemplo – Aplicando a Média Harmônica aos dados da planilha Vestibular**

| | Aluno A | Aluno B | Aluno C |
|---|---|---|---|
| Matemática | 500 | 840 | 160 |
| Física | 501 | 420 | 540 |
| Química | 497 | 430 | 530 |
| Geografia | 498 | 470 | 550 |
| História | 502 | 450 | 555 |
| Biologia | 503 | 450 | 565 |
| Literatura | 499 | 490 | 550 |
| Língua Portuguesa | 500 | 450 | 550 |
| | | | |
| Média | | | |
| Mediana | | | |
| Média sem Matemática | | | |
| | | | |
| Média Harmônica | | | |

## 2.3 Separatrizes

São valores que separam o conjunto de dados ordenado em partes com igual número de observações.

A **Mediana** é, portanto, uma separatriz porque divide o conjunto de dados em duas partes iguais.

Min |-----------------------|-----------------------| Máx
**Md**

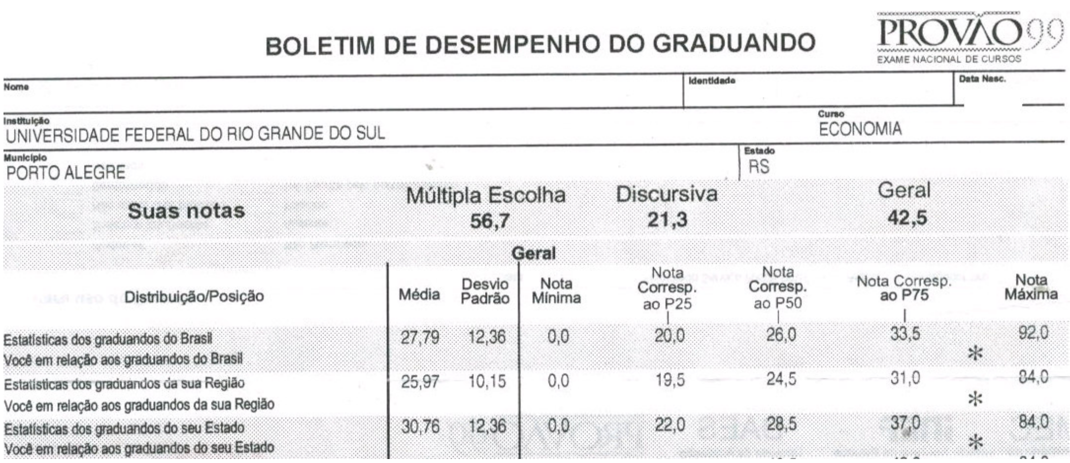
Os **Quartis** ($Q_i$) dividem o conjunto de dados em 4 partes iguais.

Min |-----------------------|-----------------------| Máx

Os **Percentis** ($P_i$) dividem o conjunto de dados em 100 partes iguais.

Min |-----------------------|-----------------------| Máx

**Exemplo – Boletim de Desempenho do Provão do MEC**



BOLETIM DE DESEMPENHO DO GRADUANDO

PROVÃO 99
EXAME NACIONAL DE CURSOS

Nome | Identidade | Data Nasc.

Instituição: UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Curso: ECONOMIA
Município: PORTO ALEGRE
Estado: RS

| Suas notas | Múltipla Escolha | Discursiva | Geral |
|---|---|---|---|
| | 56,7 | 21,3 | 42,5 |

Geral

| Distribuição/Posição | Média | Desvio Padrão | Nota Mínima | Nota Corresp. ao P25 | Nota Corresp. ao P50 | Nota Corresp. ao P75 | | Nota Máxima |
|---|---|---|---|---|---|---|---|---|
| Estatísticas dos graduandos do Brasil | 27,79 | 12,36 | 0,0 | 20,0 | 26,0 | 33,5 | | 92,0 |
| Você em relação aos graduandos do Brasil | | | | | | | * | |
| Estatísticas dos graduandos da sua Região | 25,97 | 10,15 | 0,0 | 19,5 | 24,5 | 31,0 | | 84,0 |
| Você em relação aos graduandos da sua Região | | | | | | | * | |
| Estatísticas dos graduandos do seu Estado | 30,76 | 12,36 | 0,0 | 22,0 | 28,5 | 37,0 | | 84,0 |
| Você em relação aos graduandos do seu Estado | | | | | | | * | |

**Exemplo – Estabelecer uma régua de percentis para o CPC contínuo**

| *Mín* | *P10* | *P20* | *P30* | *P40* | *P50* | *P60* | *P70* | *P80* | *P90* | *Máxi* |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | |

**=PERCENTIL(??:??)**

## 2.4 Medidas de Variabilidade

São medidas que complementam as MTC trazendo informação sobre a dispersão existente no conjunto de dados. Para introduzi-las vamos recorrer a um exemplo onde temos três diferentes empresas e a variável X investigada é o salário. Todas empresas têm seis funcionários.

**Tabela – Distribuição salarial em três empresas**

| | Empresa A | Empresa B | Empresa C |
|---|---|---|---|
| | **1000**<br>**1000**<br>**1000**<br>**1000**<br>**1000**<br>**1000** | **1000**<br>**800**<br>**1000**<br>**1200**<br>**1300**<br>**700** | **1000**<br>**200**<br>**300**<br>**1000**<br>**2100**<br>**1400** |
| Média ( $\overline{X}$ ) | | | |
| Moda (Mo) | | | |
| Mediana (Md) | | | |

| | | | |
|---|---|---|---|
| | | | |
| | | | |
| | | | |
| | | | |

**Questões**

1 – O que aconteceu com as MTC na tabela acima?

2 – As três empresas são iguais em relação a distribuição salarial?

3 – O que diferencia uma empresa da outra?

**A partir de agora aprenderemos a calcular medidas capazes de quantificar a variabilidade existente num conjunto de dados**

### 2.4.1 – Amplitude (R, do termo *Range*)

É a diferença entre o maior e o menor valor de um conjunto de dados.

$$R = máx\{x_i\} - mín\{x_i\}$$

### 2.4.2 – Variância ($\sigma^2$ , $s^2$)

**A variância é uma medida da variação em torno da média.** Por definição, variância é a média dos quadrados dos desvios em torno da média.

População

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Amostra

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

A variância, ao contrário da Amplitude, considera todos elementos do conjunto de dados no seu cálculo. Quanto maior for a variação dos valores do conjunto de dados, maior será a variância.

Quando os dados estão organizados na forma de uma **tabela de freqüências**, deve-se ponderar os quadrados dos desvios pela freqüência. Esse procedimento facilita o cálculo.

População

$$\sigma^2 = \frac{\sum_{i=1}^{k} (x_i - \mu)^2 \times f_i}{N}$$

Amostra

$$s^2 = \frac{\sum_{i=1}^{k} (x_i - \overline{X})^2 \times f_i}{n-1}$$

### 2.4.3 – Desvio-padrão ($\sigma$, $s$)

O desvio-padrão é a raiz quadrada positiva da variância. Essa medida corrige o problema de unidade que surge na variância. **O desvio-padrão também é uma medida da variação em torno da média.**

População

$$\sigma = \sqrt{\sigma^2}$$

Amostra

$$s = \sqrt{s^2}$$

O desvio-padrão expressa a variação média do conjunto de dados em torno da média, para mais ou para menos.

### 2.4.4 – Coeficiente de Variação (CV)

O CV é a razão entre o desvio-padrão e a média de um conjunto de dados. Ele expressa a variação relativa (%) presente no conjunto de dados em relação à média.

População

$$CV = \frac{\sigma}{\mu} \times 100\%$$

Amostra

$$CV = \frac{s}{\overline{X}} \times 100\%$$

Quanto maior o CV, mais heterogêneos serão os dados.

**Considerações sobre as Medidas de Variabilidade (MV)**

1. A **Amplitude** é a MV mais “pobre”, porque considera apenas os dois valores extremos do conjunto de dados.

2. A **Variância** não é interpretada na prática devido ao problema da unidade, que está ao quadrado.

3. O **Desvio-padrão** é a MV mais conhecida, sendo amplamente utilizada.

4. Dentre as MV estudadas, sugere-se que o **CV** seja utilizado para comparação da variabilidade entre diferentes conjuntos de dados. Por não ter unidade, o CV pode ser utilizado até mesmo para comparar a variabilidade entre variáveis expressas em diferentes unidades.

**Exemplo – Diferenciando as empresas**

Encontrar as medidas de variabilidade para diferenciar as três empresas.

| |
|---|
| **=MAXIMO(??:??) – MINIMO(??:??)** |
| **=VAR(??:??)** |
| **=DESVPAD(??:??)** |
| **=DESVPAD(??:??)/MEDIA(??:??)** |

**Exemplo – Encontrar a média e o desvio-padrão do CPC contínuo por Dep. Administrativa (Tipo)**

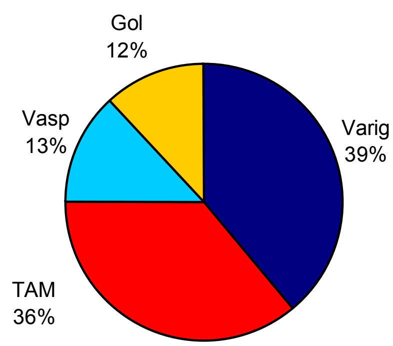
## 2.5 Análise Gráfica

O tipo de gráfico adequado para cada variável depende do tipo de variável. Segue uma relação de exemplos de variáveis e tipos de gráficos adequados.

**Variável Qualitativa Nominal (com poucas categorias)**

**GRÁFICO DE SETORES**

Figura – Marca mais lembrada no quesito “Empresa Aérea”.

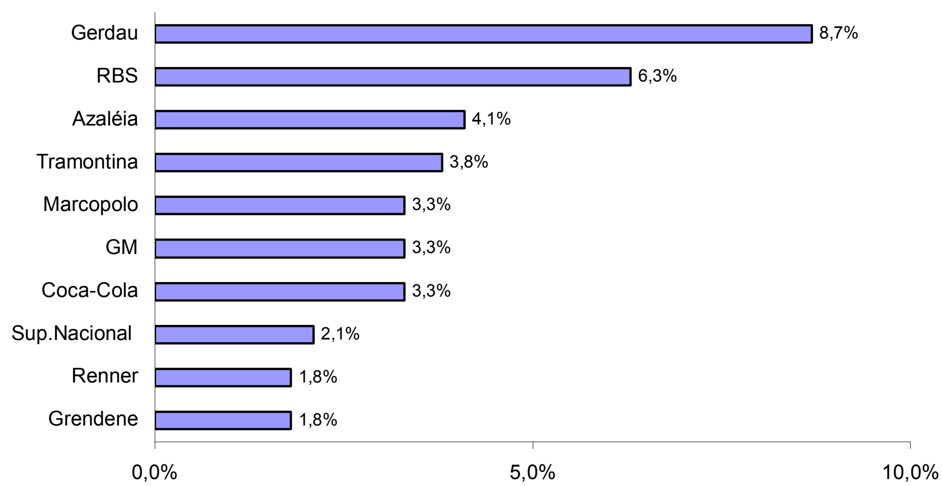



Fonte: Dados de lembrança fictícios. O gráfico foi construído de acordo com a participação no mercado em 2002.

Variável Qualitativa Nominal (com muitas categorias):

**GRÁFICO DE BARRAS**

Figura – As 10 empresas gaúchas mais lembradas

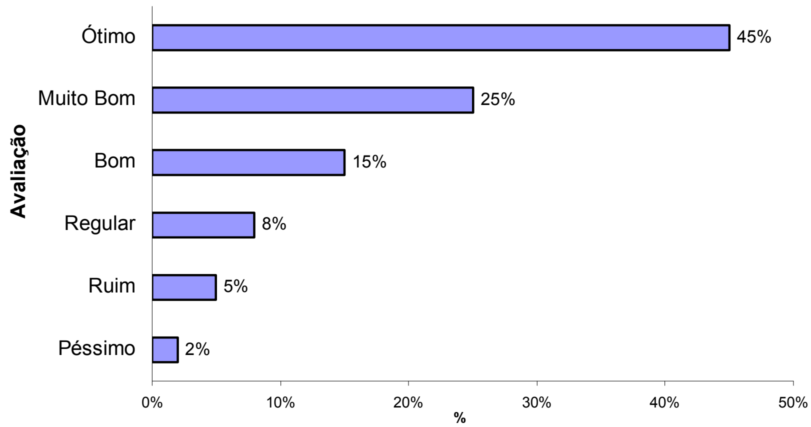



Fonte: Top of Mind – Revista Amanhã – 2002
Obs.: As outras empresas citadas somam 63,3%, entretanto individualmente não ultrapassam 1,7%.

Variável Qualitativa Ordinal:

**GRÁFICO DE BARRAS**

Figura – Avaliação do atendimento em um restaurante

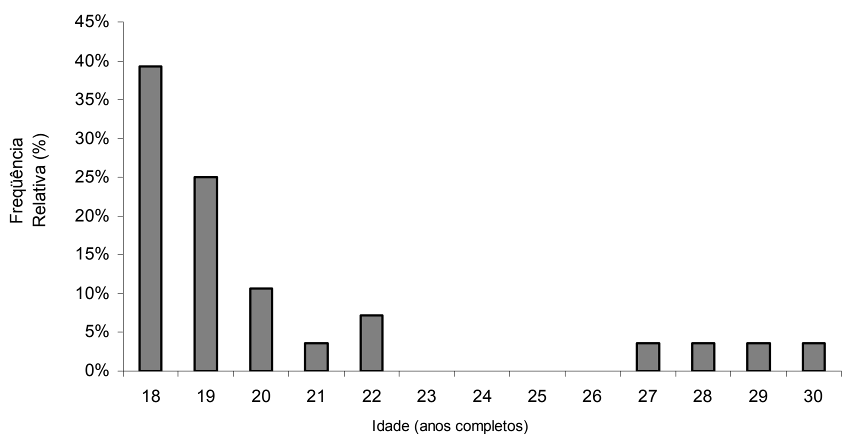



Fonte: Dados fictícios.
Base: 100 observações.

Variável Quantitativa Discreta

**GRÁFICO DE COLUNAS**

Figura – Distribuição da turma por idade

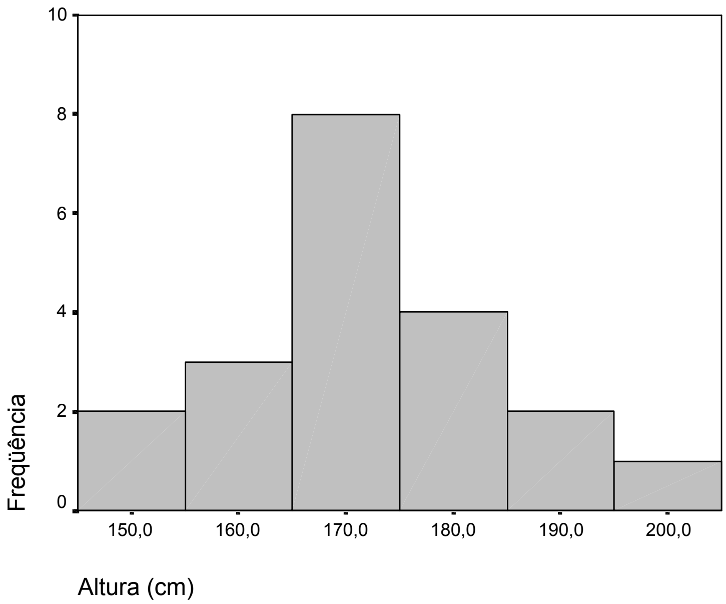



Fonte: Dados coletados na turma de Estatística para ADM – Empreendorismo e Sucessão.
Base: 28 alunos

Variável Quantitativa Contínua

**HISTOGRAMA**

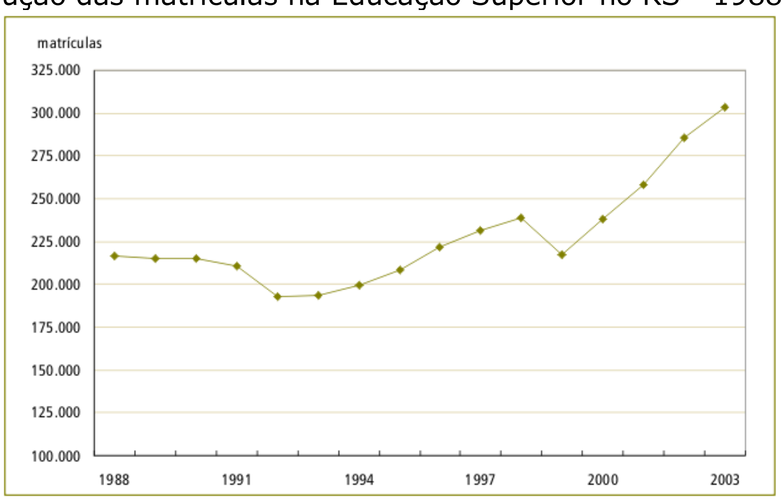
Figura – Distribuição de uma turma por altura




Fonte: Alunos de uma turma de Estatística I. Gráfico construído no software SPSS.
Base: 20 observações

Séries Temporais - GRÁFICO DE LINHAS

Figura - Evolução das matrículas na Educação Superior no RS - 1988 a 2003

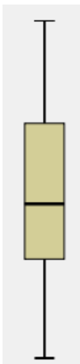



Fonte: MEC/INEP (extraído do Atlas Sócio Econômico do RS

## Exemplo – Construir no Excel

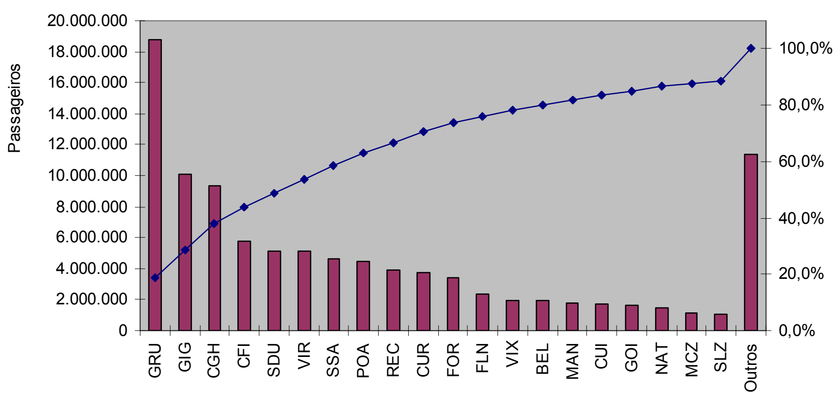
Um gráfico de setores (var. qualitativa poucas categorias)
Um gráfico de barras (var. qualitativa com muitas categorias)
Um gráfico de barras (var. qualitativa ordinal)
**Histograma (melhor fazer no Minitab ou SPSS)**

## OUTRO TIPO DE GRÁFICO 1 – Box-plot



## Exemplo - Fazer um Box-plot do CPC Contínuo por Den. Acadêmica

## OUTRO TIPO DE GRÁFICO 2 – Gráfico do Pareto




Fonte: Infraero, Jul2013 (exceto GRU e VIR, cujos dados são de 2012). Total de 66 aeroportos.

## Exemplo - Fazer um Pareto para as matrículas nas Engenharias

# Cap. 3 – Um pouco de Probabilidade

## 3.1 Principais conceitos

Probabilidade é o ramo da matemática que trata de **fenômenos aleatórios**. A observação de um fenômeno aleatório por parte do homem é chamada de **experimento aleatório**.

Características de um experimento aleatório:

1ª) Não se conhece um particular valor do experimento antes dele ser executado, porém podemos descrever todos os possíveis resultados - as possibilidades;

2ª) Quando o experimento é repetido algumas vezes, os resultados ocorrem de uma forma aparentemente acidental. Mas quando o número de repetições aumenta, uma regularidade aparecerá. E esta regularidade que torna possível construir um modelo matemático útil para análise do experimento.

**Exemplos de fenômenos aleatórios:**

1) Condições meteorológicas
2) Produção de arroz anual numa cidade
3) Resultado de uma cirurgia
4) Lançamento de uma moeda
5) Resultados de loterias

**Exemplos de experimentos aleatórios:**

E1: Jogue um dado e observe o n.º na face de cima.
E2: Jogue uma moeda 3 vezes e observe o número de caras obtido.
E3: Jogue uma moeda 3 vezes e observe a seqüência de caras e coroas obtida.
E4: Uma mulher está grávida de gêmeos. O sexo dos bebês será verificado.
E5: O número de alunos matriculados na turma de "Est Apl à Psicologia" é verificado
E6: A temperatura de um paciente é verificada pela enfermeira.

Nos seis exemplos anteriores não somos capazes de precisar o resultado, entretanto conseguimos listar os possíveis resultados.

**Espaço amostral** de um experimento aleatório é o conjunto de todos os resultados possíveis do experimento. É denotado por $S$ ou $\Omega$.

**Exemplos de espaços amostrais relacionados aos experimentos anteriores.**

$S_1$ =

$S_2$ =

$S_3$ =

$S_4$ =

$S_5$ =

$S_6$ =

Um **evento** é um subconjunto de *S*. Em particular, *S* e ∅ (conjunto vazio) são eventos; *S* é dito o evento certo e ∅ o evento impossível.

**Exemplo de eventos no lançamento de um dado**

*S* = {1,2,3,4,5,6}

| | |
|---|---|
| A: ocorre um n.º par | A = {2,4,6} |
| B: ocorre a face 6 | B = {6} |
| C: ocorre um n.º maior que 6 | C = ∅ |
| D: ocorre nº 6 ou nº par | D = {2,4,6} |
| E: ocorre nº par ou nº ímpar | E = {1,2,3,4,5,6} = *S* |

É possível realizar **operações com eventos** que nada são do que operações com conjuntos já estudadas no Ensino Fundamental.

**Operações com eventos**

Sejam A e B dois eventos associados a um espaço amostral *S*.

1) **União**: $A \cup B \rightarrow$ A ocorre <u>ou</u> B ocorre ou ambos ocorrem

2) **Interseção**: $A \cap B \rightarrow$ A ocorre <u>e</u> B ocorre

3) **Complementar**: $A^c$ ou A $\rightarrow$ não ocorre A

Duas definições importantes:

1) Dois eventos A e B são **excludentes** ou **mutuamente exclusivos** se a ocorrência de um impedir a ocorrência de outro. Em outras palavras, não podem ocorrer simultaneamente.

2) Eventos ou resultados **equiprováveis** têm a mesma probabilidade de ocorrência.

**Exemplo – Lançamento de um dado e uma moeda, ambos honestos**

Escreva o espaço amostral. Os resultados são todos equiprováveis? Qual a probabilidade de um particular par (x,y) ser selecionado. Assinale os seguintes eventos:

### 3.1.1 Conceitos de probabilidade

**⇒ Conceito Axiomático**

Seja A um evento de S. A probabilidade de ocorrência de A, denotada por P(A), deverá satisfazer os seguintes axiomas (propriedades fundamentais):

Axioma 1: $0 \le P(A) \le 1$

Axioma 2: $P(S) = 1$

**⇒ Conceito clássico**

Esse conceito só é válido se todos resultados de S forem **equiprováveis.** Para casos assim a probabilidade de ocorrência do evento A é obtida por:

$$P(A) = \frac{n(A)}{n(S)}$$

n(A) é o número de resultados favoráveis ao evento A
n(S) é o número total de resultados em S

**Exemplos – Conceito clássico**

1) Mega-sena, Lançamento de moedas e dados honestos.

### ⇒ Conceito freqüentista

Neste conceito a probabilidade é vista como um limite**.** Para casos assim a probabilidade de ocorrência do evento A é obtida por:

1º) O experimento é repetido *n* vezes.

2º) Observa-se a freqüência relativa de ocorrência de um certo resultado A:

$f_r$(A) = $\frac{n(A)}{n}$, onde n(A) é o nº de vezes em que ocorre o resultado A em *n* realizações do experimento.

3º) Probabilidade como limite. A medida que *n* aumenta, a fr(A) converge para a real probabilidade P(A).

$$P(A) = \lim_{n\to\infty} \frac{n(A)}{n}$$

*Exemplos – Conceito freqüentista*

1) Verificando se um dado é honesto.

2) Encontrando a probabilidade de ocorrência de um acidente aéreo.

3) Qual a probabilidade de uma criança nascer com *Síndrome de Down*?

### 3.1.2 Probabilidade Condicional

A probabilidade de ocorrência de um evento pode ser influenciada pela ocorrência de um evento paralelo. Considere que A e B são eventos de um mesmo espaço amostral S. Chamaremos de **P(A|B)** a probabilidade de ocorrência do evento A dado que o evento B já ocorreu.

Graficamente:

Olhando para o desenho podemos estabelecer as seguintes relações:

P(A|B) = P(B|A) =

**Exemplo – Escolhendo alguém na sala de aula**

Suponha que um aluno da turma será sorteado. Após saber o resultado o professor faz algumas perguntas utilizando probabilidade condicional.

**Exemplo – Máquina X Conformidade**

| Máquina | Resultado | | Total |
|---|---|---|---|
| | Conforme | Não conforme | |
| A | 60 | 25 | 85 |
| B | 40 | 65 | 105 |
| Total | 100 | 90 | 190 |

Resolver as seguintes probabilidades:

### 3.1.3 Independência

Dois eventos A e B são considerados independentes se a ocorrência de um não interfere na probabilidade de ocorrência do outro:

$$P(A|B) = P(A) \qquad \text{e} \qquad P(B|A) = P(B)$$

Isolando a intersecção na expressão de probabilidade condicional obtemos:

$$P(A \cap B) = P(A) \times P(B)$$

Esse conceito é fundamental para aplicações em Estatística.

**Exemplo –** A probabilidade de ter um carro roubado no período de 1 ano é de 0,6%. Escreva o espaço amostral e calcule as probabilidades para todos possíveis resultados de uma família que tem dois carros.

## 3.2 – Variáveis aleatórias discretas – Distribuição Binomial

O exercício acima pode ser resolvido pela *Distribuição Binomial*. Sempre que um experimento que assume apenas dois possíveis resultados em cada repetição for repetido *n* vezes e que a probabilidade de sucesso é constante em cada repetição podemos modelar o número de sucessos pela distribuição Binomial.

X = número de sucessos, variando de 1 até n
p = probabilidade de sucesso em cada repetição
1-p = probabilidade de fracasso em cada repetição
n = número de repetições

*Expressão genérica da Binomial*

$$\boxed{P(X = x) = \frac{n!}{x!(n-x)!} \times p^x \times (1-p)^{n-x}}$$

O número esperado ou **esperança** de sucessos na distribuição Binomial é facilmente encontrado. Intuitivamente, responda as perguntas a seguir:

1) Se lançarmos uma moeda honesta 100 vezes, qual o número esperado de caras?

2) Se lançarmos um dado 600 vezes, qual o número esperado de faces "5".

3) Em uma frota que 1000 carros, qual o número esperado de carros roubados no período de 1 ano?

$$\boxed{E(X) = n \times p}$$

### Exercício – Inspeção por amostragem

Suponha que lotes com uma quantidade muito grande de peças são produzidos e que a proporção de peças defeituosas nesses lotes é estimada em 2%. Numa amostra aleatória de 30 itens para inspeção:

a) Qual modelo descreve X=número de peças defeituosas na amostra?
b) Esboce o modelo teórico (tabela e gráfico).
c) Faça uma simulação com um lote de N=5000 no Excel e construa a distribuição empírica de X.

### Exercício – Aceitação por amostragem

Construa uma Curva Característica de Operação (CCO) para um plano de amostragem para aceitação / rejeição de um lote. Siga as instruções do professor.

## 3.3 Variáveis aleatórias contínuas

### 3.3.1 Conceitos

As variáveis contínuas podem, ao menos teoricamente, assumir qualquer valor num intervalo numérico. Sendo assim fica impossível representarmos variáveis contínuas da mesma forma que as variáveis discretas.

**Importante**

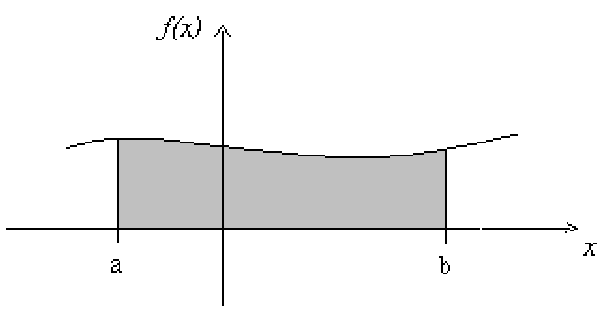
As variáveis contínuas são representadas por **curvas**, chamadas de **função densidade de probabilidade**, e a área sob essa função representa a probabilidade de ocorrência. Nas variáveis contínuas **não** existe a probabilidade de ocorrência de um valor exato, mas sim de intervalos.

A **função densidade de probabilidade**, denotada por $f_X(x)$, é a função que indica o comportamento probabilístico da variável aleatória contínua X. A função densidade de probabilidade deverá satisfazer as seguintes condições:

a) $f(x) \geq 0$, para todo $x \in R$.

b) Área total sob a curva deve ser igual a 1.

A área sob a curva $f_X(x)$ nos informa a probabilidade de ocorrência de valores da variável X.




Supondo que o gráfico acima represente a função de probabilidade de uma variável aleatória X. Como sabermos a probabilidade de ocorrência de valores entre *a* e *b* ?

**Exemplo – Gerador de Números pseudo-aleatórios (Distribuição Uniforme)**

Suponha que o gerador de números pseudo-aleatórios do Excel siga uma distribuição Uniforme no intervalo de [0 ; 1].
a) Esboce graficamente a função densidade de probabilidade para esse caso.
b) Encontre a probabilidade de ocorrer um número maior que 0,8.

### 3.3.2 A Distribuição Normal ou Curva de Gauss

A **distribuição Normal** ou **Gaussiana** é, sem dúvida, o modelo probabilístico mais conhecido. Várias técnicas estatísticas necessitam da suposição de que os dados se distribuam normalmente para serem utilizadas. Na natureza uma grande quantidade de variáveis apresentam tal distribuição.

Uma v.a.c. X tem distribuição normal com parâmetros $\mu$ e $\sigma$ se sua função densidade de probabilidade é dada por:

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

onde $\mu$ e $\sigma$ são parâmetros,

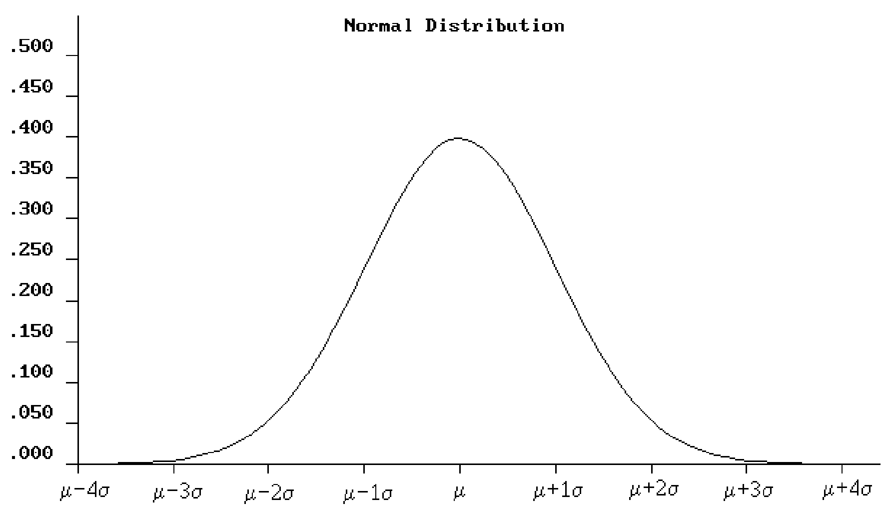
$-\infty < \mu < +\infty \ ;\ \ \sigma > 0$

**Notação**

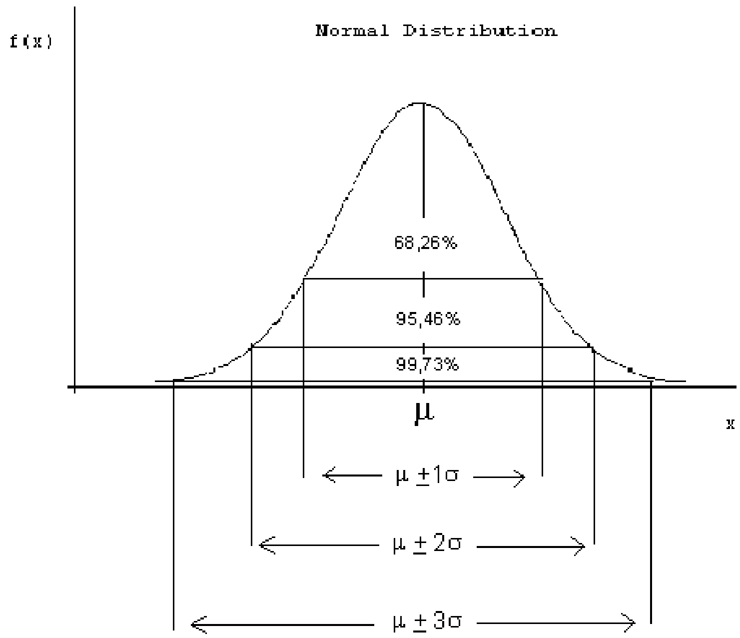
$$X \sim N(\mu,\sigma)$$

X tem distribuição Normal com média $\mu$ e desvio-padrão $\sigma$.

Os parâmetros da Normal são a média e o desvio-padrão, que permitem infinitas curvas normais com diferentes formatos (mas sempre simétricas). O gráfico da $f_X$ é apresentado a seguir:



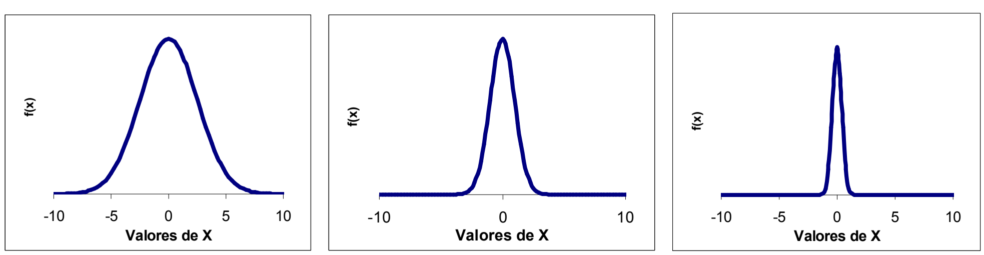
A distribuição Normal, independentemente dos valores dos parâmetros, apresenta **sempre** a seguinte relação:




## Entendendo os parâmetros da Normal:

A média $\mu$ informa o centro da distribuição. É um parâmetro de locação.

O desvio-padrão $\sigma$ informa o formato da curva.




Os cálculos integrais envolvendo a distribuição Normal são bastante complicados. Felizmente, veremos a seguir uma relação que facilita muito nossa vida.

**Exemplo – Aplicação prática**

A altura de mulheres adultas no RS segue uma distribuição Normal com média de 165cm e desvio-padrão de 6cm.

a) Qual a probabilidade de uma mulher ter entre 159 e 171cm?

b) Qual a probabilidade de uma mulher ter entre 153 e 177cm?

c) Qual a probabilidade de uma mulher ter mais de 177cm?

d) Qual a probabilidade de uma mulher ter menos de 180cm?

**Distribuição Normal-padrão ou Normal reduzida**

Seja X uma variável aleatória normalmente distribuída com quaisquer parâmetros média $\mu$ e desvio-padrão $\sigma$. Se realizarmos a seguinte transformação obteremos uma nova variável Z com média 0 e desvio-padrão 1:

$$X \sim N(\mu,\sigma) \quad \rightarrow \quad Z = \frac{X-\mu}{\sigma} \quad \rightarrow \quad Z\ (0,1)$$

**Qualquer variável com distribuição Normal pode ser padronizada para a Normal.** A distribuição Normal padronizada (Z) é tabelada.

| |
|---|
| O valor de **Z** indica quantos desvios acima ou abaixo nós estamos em relação à média. |

**Exemplo – Aprendendo a usar a tabela**

1) Calcule:

a) $P(Z < 1,24) =$

b) $P(Z < 1,67) =$

c) $P(Z > 2,12) =$

# Cap. 4 Amostragem & Distribuições Amostrais

## 4.1 Conceitos Básicos

**Amostragem** é o nome dado ao conjunto de procedimentos e técnicas para extração de elementos da população para compor a amostra. O objetivo da amostragem é obter uma amostra representativa da população em estudo. Um **Censo** é a investigação da população completa.

Por que trabalhar por amostragem?

__________________________________________

__________________________________________

__________________________________________

__________________________________________

A **fração de amostragem *f*** é a razão entre o tamanho amostral (n) e o tamanho populacional (N). Não existem regras fixas para tamanho de amostra, ou seja, cada caso merece um cuidado especial. Frases como "10% da população é ideal", quase sempre não são verdadeiras.

$$f = \frac{n}{N}$$

As técnicas de amostragem se dividem em: **probabilísticas** e **não-probabilísticas**. As técnicas probabilísticas são aquelas onde todos os elementos da população têm uma probabilidade não nula de seleção; é possível associar probabilidade de seleção a todos elementos que compõem a população. Nas técnicas não-probabilísticas não podemos garantir que todos elementos têm probabilidade de serem selecionados para a amostra.

## 4.2 Principais técnicas de amostragem probabilística

Geralmente as técnicas probabilísticas produzem melhores resultados do que as não probabilísticas, tendo uma melhor receptividade pela comunidade científica. Neste tipo de amostragem, a seleção dos elementos envolve a utilização de algum **dispositivo aleatório** para seleção das unidades amostrais, pois, desta forma, estamos garantindo que todos os elementos da população tiveram sua "chance" de serem selecionados.

| |
|---|
| **Exemplo de dispositivos aleatórios:**<br>Função Aleatório do Excel, Random, Urna, Dado, Moeda. |

### 4.2.1 Amostragem Aleatória Simples (AAS)

Apesar de ser uma forma extremamente simples de seleção de elementos da população, a AAS é considerada uma das melhores técnicas de amostragem. Na AAS cada elemento da população tem igual probabilidade de seleção.

Etapas:

1) Enumerar a população de 1 até $N$.
2) Sortear $n$ números no intervalo de 1 até $N$. (pode ser COM ou SEM reposição)
3) Compor a amostra com os elementos selecionados.

Probabilidade de seleção de um elemento na AAS:

Número de amostras possíveis SEM reposição:

Número de amostras possíveis COM reposição:

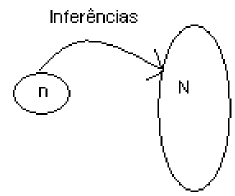
**Exemplo – Amostra n=2 da população N=5**

Verificar quantas amostras são possíveis SEM reposição da população de tamanho 5 verificando também as probabilidades de seleção de cada unidade.

**A B C D E**

## 4.3 Distribuições Amostrais & Estimação por ponto e por intervalo

Agora que já conhecemos seus dois pilares – Probabilidade e Amostragem – iniciaremos a parte de **Inferência Estatística**. Inferir significa transcender resultados amostrais para toda população.




No item a seguir faremos a distinção entre características populacionais e amostrais.

### 4.3.1 Parâmetros e Estimadores

Um **parâmetro** é um valor que descreve uma característica da população. Um **estimador** é uma função de dados amostrais que gera **estimativas** para um parâmetro. A palavra **estatística** pode ser utilizada como um sinônimo de estimador.

Considere $x_1$ , $x_2$ , ... , $x_n$ uma amostra aleatória de $n$ elementos de uma população. Denotaremos por $\theta$ um parâmetro qualquer e por $\hat{\theta}(x)$ ou simplesmente $\hat{\theta}$ um estimador de $\theta$. Os parâmetros geralmente são denotados por letras gregas e é bastante comum utilizar o "chapéu" para diferenciar os estimadores dos parâmetros.

### 4.3.2 Propriedades dos Estimadores

Considere $x_1$ , $x_2$ , ... , $x_n$ uma amostra de $n$ elementos de uma população com média $\mu$ e desvio-padrão $\sigma$ , sendo os $x_i$ independentes e identicamente distribuídos. Um estimador $\hat{\theta}(x)$ será considerado uma variável aleatória pois, cada uma das possíveis amostras de tamanho $n$ da população provavelmente gerará uma diferente estimativa.

| |
|---|
| Um estimador $\hat{\theta}(x)$ é uma variável aleatória porque depende da amostra (aleatória) selecionada. |

Um bom estimador do parâmetro $\theta$ deverá apresentar algumas propriedades desejáveis. Iremos discutir quatro propriedades dos estimadores:

1ª) **Não tendenciosidade**: um estimador é não-tendencioso se a sua esperança coincide com o valor do parâmetro. Em outras palavras, uma estimador é não-tendencioso se as suas estimativas, em média, são corretas, orbitando em torno de $\theta$.

Exemplo:
E( $\overline{X}$ ) =

2ª) **Consistência**: um estimador é consistente se a sua variabilidade diminui a medida que o tamanho amostral aumenta. Esta é uma propriedade altamente desejável quando aliada a não-tendenciosidade.

Se $\hat{\theta}(x)$ é consistente, então $\lim_{n\to\infty} Var(\hat{\theta}) = 0$.

Exemplo:

Var( $\overline{X}$ )=

| |
|---|
| ***Importante:***<br>O desvio-padrão de um estimador também é chamado de Erro-padrão.<br>$DP(\hat{\theta}) = \sqrt{Var(\hat{\theta})} = Erro - padrão$ |

3ª) **Eficiência**: um estimador $\hat{\theta}_1$ é mais eficiente que $\hat{\theta}_2$ se Var( $\hat{\theta}_1$ ) < Var( $\hat{\theta}_2$ ) para um mesmo tamanho amostral *n*.

4ª) **Suficiência**: um estimador $\hat{\theta}$ é suficiente se utilizar o máximo de informação disponível na amostra para estimar o parâmetro $\theta$.

### 4.3.3 Métodos de Estimação

Neste texto discutiremos dois diferentes métodos de estimação de parâmetros. Um terceiro método, chamado de mínimos quadrados, será trabalhado no item 5.4.

1º) **Método da Máxima Verossimilhança (MáxVer) ou *Maximum Likelihood* (*ML*):** é o método computacionalmente mais empregado para estimação de parâmetros, consistindo na maximização da função de verossimilhança $\ell(\theta)$ ou de seu logaritmo ln $\ell(\theta)$.

A função de verossimilhança retrata a probabilidade de uma particular amostra $x_1, x_2, \ldots, x_n$ ter sido selecionada em função do(s) parâmetro(s) que regem a função de probabilidade do fenômeno. Assim sendo, a estimativa de máxima verossimilhança para $\theta$ será aquela que tornar máxima a função. De acordo com Meyer (1969), as estimativas de MáxVer podem ser tendenciosas, mas são consistentes, assintoticamente eficientes e, a medida que *n* aumenta, os estimadores de MáxVer seguem distribuição aproximadamente normal.

$$\ell(\theta) = f(x_1,\theta) f(x_2,\theta) \ldots f(x_n,\theta) = \prod_{i=1}^{n} f(x_i,\theta)$$

Para entender melhor a função de verossimilhança, tente lembrar da regra da independência:

$$P(A \cap B) = P(A) \times P(B)$$
$$P(A \cap B \cap C) = P(A) \times P(B) \times P(C)$$
$$P(\bigcap_{i=1}^{n} A_i) = \prod_{i=1}^{n} P(A_i)$$

Exemplo – URNA (Costa Neto, 1977):

Considere uma urna com S bolas Pretas e 10-S bolas Brancas e quatro bolas **COM** reposição serão extraídas da caixa. Estimar por MáxVer o valor de S sabendo que, numa extração aleatória ocorreu uma bola Preta e três Brancas.

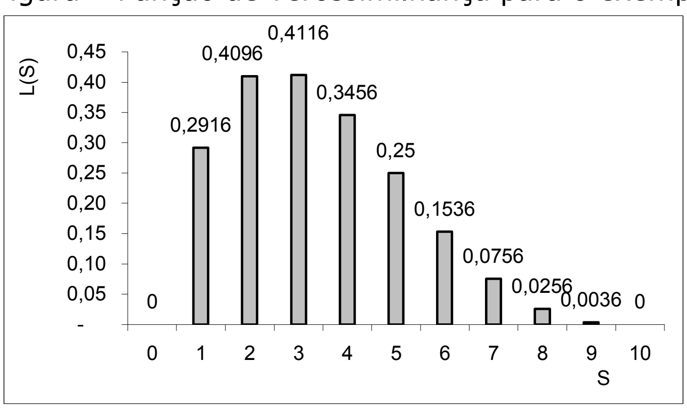
Como n=10 bolas, logo sabemos que o valor de S irá variar entre 0 e 10 e, por ser extração COM reposição, podemos modelar este fenômeno pela distribuição Binomial com parâmetros n=4 e p=S/10.

Suponha que X=número de bolas pretas na amostra, assim a função de verossimilhança $\ell(S)$ será baseada na binomial e ficará única e exclusivamente como função de S, visto que X=1 a partir dos dados da amostra n=4.

$$\ell(S) = P(X=1) = C_4^1 \left(\frac{S}{10}\right)^1 \left(\frac{10-S}{10}\right)^{4-1} = 4 \times \frac{S}{10} \times \left(\frac{10-S}{10}\right)^3 = 4 \times \frac{S}{10} \times \frac{(10-S)^3}{1000} = \frac{S(10-S)^3}{2500}$$

Mostraremos graficamente o processo de maximização de $\ell(S)$ na forma de um gráfico próprio para esta função discreta.

Figura – Função de verossimilhança para o exemplo da urna

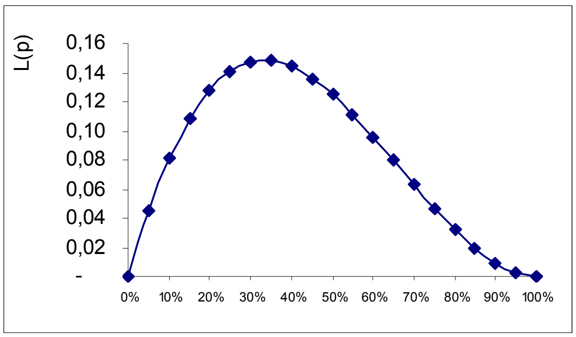


Exemplo – PRÊMIOS OCULTOS

Uma promoção distribui picolés gratuitos dentre aqueles que encontrarem o vale-brinde no palito. Suponha que você deseje estimar a taxa de premiação, ou seja, qual a proporção *p* de picolés vendidos que têm prêmio. Suponha que, na terceira tentativa você encontre um prêmio.

Perceba que, embora tenhamos outro problema que envolva variáveis discretas, neste caso não podemos enumerar todas possibilidades para *p* visto que não sabemos o total de picolés produzidos. De qualquer forma, obviamente, $0<p<1$.

Figura – Função de verossimilhança para o exemplo do Prêmio Oculto




Exemplo – Estimador de MáxVer para $\lambda$ na Poisson

Encontar o estimador MáxVer para $\lambda$ sabendo que $X \sim$ Poisson ($\lambda$). Neste caso é muito mais simples maximizar o logaritmo natural da função de verossimilhança.

2º) **Método dos Momentos**: Este método foi proposto pelo inglês Pearson em 1894 e consiste em igualar os momentos amostrais aos momentos da distribuição populacional.

Exemplo – Distribuição Normal

Média amostral (1º momento amostral) = $\overline{X}$

Variância amostral (2º momento amostral) = $\hat{\sigma}^2$

Na distribuição Normal:

$E(X) = \mu$ e $Var(X) = \sigma^2$

Portanto $\overline{X}$ pode ser usado para estimar $\mu$ e $\hat{\sigma}^2$ para estimar $\sigma^2$. Sabemos, entretanto, que $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}{n}$ é um estimador viciado para $\sigma^2$.

Exemplo – Poisson
Sabendo que $X \sim$ Poisson ($\lambda$), encontrar pelo método dos momentos um estimador para $\lambda$.

## 4.4 Distribuição amostral da média ( $\overline{X}$ )

Já sabemos que o estimador $\overline{X}$ é uma variável aleatória, mas ainda não estudamos o seu comportamento probabilístico. Vejamos um exemplo:

Exemplo – População N=6 e Amostra n=3

| A | B | C | D | E | F |
|---|---|---|---|---|---|
| 10 | 20 | 30 | 40 | 50 | 60 |

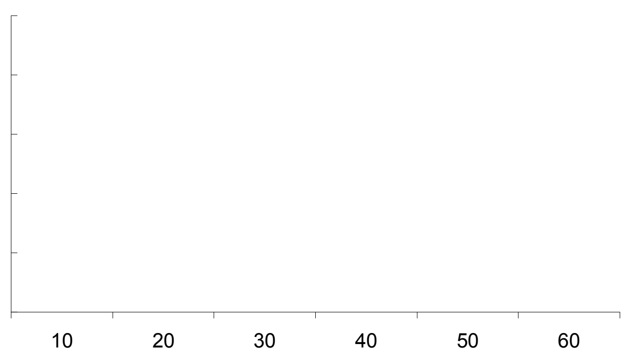
Parâmetros; $\mu$ = 35 $\sigma$ =17,078

**Quadro – Possíveis estimativas para $\mu$**

| *i* | *Amostra* | $\overline{X}_i$ | *i* | *Amostra* | $\overline{X}_i$ |
|---|---|---|---|---|---|
| **1** | **ABC** | | **11** | **BCD** | |
| **2** | **ABD** | | **12** | **BCE** | |
| **3** | **ABE** | | **13** | **BCF** | |
| **4** | **ABF** | | **14** | **BDE** | |
| **5** | **ACD** | | **15** | **BDF** | |
| **6** | **ACE** | | **16** | **BEF** | |
| **7** | **ACF** | | **17** | **CDE** | |
| **8** | **ADE** | | **18** | **CDF** | |
| **9** | **ADF** | | **19** | **CEF** | |
| **10** | **AEF** | | **20** | **DEF** | |

$$\overline{\overline{X}} = \frac{\sum_{i=1}^{20} \overline{X}_i}{20} \qquad \sigma^2_{\overline{X}} = \frac{\sum_{i=1}^{20} \left(\overline{X}_i - \overline{\overline{X}}\right)}{20}$$

Graficamente:




O objetivo deste exemplo foi demonstrar o funcionamento do teorema do limite central, formalizado a seguir.

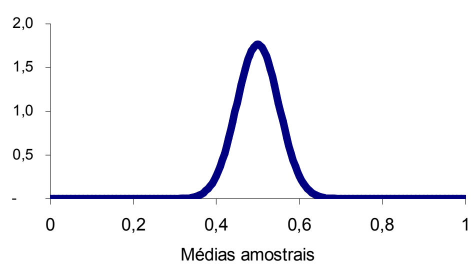
**Teorema do Limite Central**

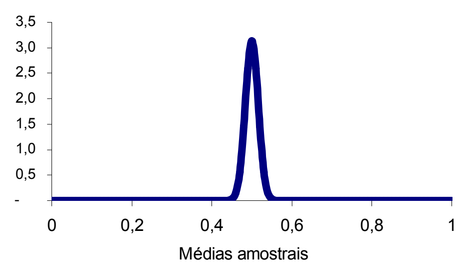
Para populações infinitas:
$\overline{X}$, a medida que n$\to$infinito, tem distribuição Normal com Média=$\mu$ e Desvio-padrão $= \dfrac{\sigma}{\sqrt{n}}$

Quando N é finito:
$\overline{X}$, a medida que n$\to$N, tem distribuição Normal com Média=$\mu$ e Desvio-padrão $= \dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}}$

Exemplo – População com média 0,5

Considere uma população infinitamente grande com média $\mu = 0{,}5$. Vamos avaliar as distribuições amostrais da média amostral $\overline{X}$ com n = 30 e 300.




n = 30 n = 300

Percebemos claramente que com o aumento do tamanho amostral a distribuição de $\overline{X}$ fica cada vez mais concentrada em torno do parâmetro $\mu$. Isso quer dizer que, quanto maior amostra, maior a possibilidade de acertarmos o valor do parâmetro.

## 4.5 Distribuição amostral da proporção ( $\hat{p}$ )

O estimador $\hat{p}$ teoricamente pode ser descrito pela distribuição Binomial. Vamos definir uma variável binária X assumindo os valores 0 (característica ausente) e 1 (característica presente). Considere uma população infinita com parâmetro *p.* O estimador imediato de *p* é a proporção amostral $\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$. Se assumirmos população infinita (equivalente a amostragem com reposição) a distribuição amostral de *n* $\hat{p}$ será descrita pela Binomial com parâmetros (*n ; p*). No caso de populações finitas a distribuição de $\hat{p}$ será Hipergeométrica.

Exemplo – População N=6 e Amostra n=3

| | | | | | |
|---|---|---|---|---|---|
| A | B | C | D | E | F |
| 0 | 1 | 0 | 1 | 1 | 0 |

**Quadro – Possíveis estimativas para p**

| *i* | *Amostra* | $\hat{p}_i$ | *i* | *Amostra* | $\hat{p}_i$ |
|---|---|---|---|---|---|
| **1** | **ABC** | | **11** | **BCD** | |
| **2** | **ABD** | | **12** | **BCE** | |
| **3** | **ABE** | | **13** | **BCF** | |
| **4** | **ABF** | | **14** | **BDE** | |
| **5** | **ACD** | | **15** | **BDF** | |
| **6** | **ACE** | | **16** | **BEF** | |
| **7** | **ACF** | | **17** | **CDE** | |
| **8** | **ADE** | | **18** | **CDF** | |
| **9** | **ADF** | | **19** | **CEF** | |
| **10** | **AEF** | | **20** | **DEF** | |

$$\bar{\hat{p}} = \frac{\sum_{i=1}^{20} \overline{X}_i}{20}$$

Para n > 30 a distribuição de $\hat{p}$ pode ser aproximada pela Normal, com média igual ao parâmetro *p* e desvio-padrão dado por $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

# Cap. 5 Inferência Estatística: Intervalos de Confiança

A estimação por **ponto** e por **intervalo** constitui um tipo de inferência estatística, pois buscamos encontrar o verdadeiro valor do parâmetro populacional a partir de uma amostra.

## 5.1 – Estimação por ponto

Visa estimar o valor do parâmetro através de estimativas pontuais (únicas). A vantagem é ser de fácil e rápida interpretação, mas, em contrapartida, a probabilidade de acerto "na mosca" é praticamente nula.

Mas por que a probabilidade de acertar "na mosca" é praticamente zero? Você se lembra que os estimadores podem ser encarados como variáveis aleatórias contínuas e que, nesse tipo de variável, não é possível calcular a probabilidade de ocorrência para valores exatos, mas sim para intervalos.

## 5.2 – Estimação por intervalo de confiança

Consiste em cercar o valor da estimativa pontual por uma região cuja probabilidade de conter o verdadeiro parâmetro seja conhecida.

**NOTAÇÕES que serão utilizadas a partir de agora**

$\alpha$ (alfa) = nível de significância 1 - $\alpha$ = nível de confiança

$t_{n-1;\frac{\alpha}{2}}$ = valor da distribuição t de Student com *n-1* graus de liberdade e área $\frac{\alpha}{2}$ à direita.

$z_{\frac{\alpha}{2}}$ = valor da distribuição normal padrão com área $\frac{\alpha}{2}$ à direita

### 1°) Intervalo de Confiança para $\mu$ (teórico)

Conhecendo o teorema do limite central podemos construir intervalos de confiança para a média populacional. Para isso basta cercarmos a estimativa pontual $\overline{X}$ por um intervalo cuja probabilidade de conter o parâmetro seja conhecida.

$P(a < \mu < b) = 1 - \alpha$ O intervalo [a;b] tem 1-$\alpha$ de probabilidade de conter $\mu$.

Mas quais os valores de *a* e *b*? O conhecimento do comportamento probabilístico de $\overline{X}$ possibilita encontramos um intervalo de confiança para $\mu$.

$$P\left( a < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < b \right) = 1-\alpha$$

**I.C. para μ com 1-α de confiança =** $\left[ \overline{X} \pm z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \right]$

Na fórmula de IC acima percebemos a presença de um parâmetro (σ). Se estamos procurando um intervalo de confiança para μ é porque NÃO conhecemos μ. É praticamente impossível conhecermos σ e não conhecermos μ. Por isso esse resultado acaba sendo INÚTIL na prática.

## 2°) Intervalo de Confiança para μ (prático)

Ao substituirmos o parâmetro σ por seu estimador *s* , a distribuição amostral de $\overline{X}$ deixa de ter uma distribuição Normal e passa a ter uma distribuição *t de Student*. Desta forma os Intervalos de confiança podem ser utilizados em situações práticas.

**I.C. para μ com 1-α de confiança =** $\left[ \overline{X} \pm t_{n-1,\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}} \right]$

Quando a população é finita, faz-se necessário o fator de correção para populações finitas:

**I.C. para μ com 1-α de confiança =** $\left[ \overline{X} \pm t_{n-1,\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} \right]$

**Exemplo – Diâmetro de Aros**

Aros são produzidos numa indústria metalúrgica. O diâmetro nominal é de 15cm. Numa amostra aleatória de 31 aros, o diâmetro médio encontrado foi de 15,3cm com um desvio-padrão de 0,8cm.

Construa um IC 95% para o verdadeiro diâmetro médio dos aros nesta linha de produção.

## 3º) Intervalo de Confiança para uma proporção populacional p

A estimativa pontual para uma proporção é dada diretamente pela proporção amostral.
Já vimos que para n>30, a distribuição de $\hat{p}$ pode ser aproximada pela distribuição Normal (p; $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$). Vamos deduzir a expressão do intervalo:

*I.C. para p com 1-$\alpha$ de confiança* = $\left[\hat{p} \pm z_{\frac{\alpha}{2}} \times \sqrt{\frac{\hat{p}\times(1-\hat{p})}{n}}\right]$

O fator de correção $\sqrt{\frac{N-n}{N-1}}$ é acrescentado em caso de populações infinitas.

*I.C. para p com 1-$\alpha$ de confiança* = $\left[\hat{p} \pm z_{\frac{\alpha}{2}} \times \sqrt{\frac{\hat{p}\times(1-\hat{p})}{n}} \times \sqrt{\frac{N-n}{N-1}}\right]$

onde $z_{0,05}$ =1,645 (90%)

$z_{0,025} = 1{,}96$ (95%)

$z_{0,005} = 2{,}576$ (99%)

**Exemplo –** Proporção de reclamações no período de garantia

Uma empresa vendeu 1500 unidades de uma máquina com garantia de 2 anos. Uma amostra de 150 clientes (leia-se 150 máquinas) revelou que 15 tinham utilizado os serviços de garantia. Construa um Intervalo de confiança 95% para a verdadeira proporção de clientes que utilizou serviços de garantia.

## 5.3 Dimensionamento de Amostras

Dimensionar uma amostra é verificar qual o *n* necessário para atingir uma precisão pré-definida, fixando-se um nível de confiança e fazendo algumas suposições. Na realidade, dimensionar a amostra é controlar a *margem de erro* do estudo.

**Margem de erro = semi-amplitude do Intervalo de Confiança**

As expressões a seguir são de fácil aplicação, mas é necessário ter algum conhecimento prévio sobre a variável de interesse para ser possível calcular *n*. Os valores de *t* , $\overline{X}$ , s e $\hat{p}$ podem ser obtidos em amostras piloto ou em estudos anteriores, ou ainda, em último caso, por intuição.

Deduções das expressões de cálculo do tamanho amostral:

**Para médias:**

$$n = \left( \frac{t_{n-1;\frac{\alpha}{2}} \times s}{erro} \right)^2$$

**Para proporções:**

$$n = \frac{z_{\frac{\alpha}{2}}^2 \times \hat{p} \times (1 - \hat{p})}{erro^2}$$

Esses tamanhos amostrais precisam ser ajustados em caso em populações finitas:

$$n_{finito} = \frac{N \times n}{N + n}$$

Exercícios:

1) No exercício do diâmetro dos aros, qual deveria ser o *n* para reduzir a margem de erro para 0,15cm, mantendo a confiança em 95%.

2) No exercício da garantia, qual deveria ser o *n* para reduzir a margem de erro para 2% mantendo a confiança em 95%.

3) Diferencie nível de confiança de margem de erro.

4) A relação entre a margem de erro e o *n* é linear?

## Exercícios

**1) Segundo a FEE, o rendimento médio real dos ocupados na RMPA foi estimado em R$ 1395,00 em Abril/2011. Considerando um desvio-padrão correspondente a 50% da média, qual deverá ser o tamanho amostral *n* para estimar esse rendimento médio com margem de erro de R$ 100,00.**

**2) Construir uma curva que mostre o tamanho amostral *n* como função do *erro* usando os dados do exercício anterior.**

**3) Numa amostra de n=120 porto-alegrenses obteve-se um total de 30 pessoas que nasceram em outras cidades.**
**Construa um intervalo de confiança BINOMIAL e outro usando a fórmula baseada na aproximação pela NORMAL para a proporção de poto-alegrenses que não nasceram em POA.**

**4) Extrair 500 amostras de tamanho n=30 da população N=10.000 do arquivo Excel. Construir intervalos para cada estimativa de x-barra e contar quantos contém o parâmetro.**