INTRODUÇÃO

ESTATÍSTICA EXPERIMENTAL

O que é???

trata de metodologias para coleta, organização, análise e interpretação de dados obtidos em levantamentos amostrais ou em experimentos especialmente delineados para tal fim, com o objetivo de tomar melhores decisões.

ETAPAS ENVOLVIDAS EM UM EXPERIMENTO

- PLANEJAMENTO

- EXECUÇÃO E CONDUÇÃO DO EXPERIMENTO

- COLETA DOS DADOS

- ORGANIZAÇÃO DOS DADOS

- ANÁLISE DOS DADOS

- INTERPERATAÇÃO DOS RESULTADOS

- CONCLUSÕES OBTIDAS (DECISÕES Á TOMAR)

NA COLETA DOS DADOS...

AS VARIÁVEIS MEDIDAS OU OBTIDAS PODEM SER:

- VARIÁVEL
  - QUALITATIVA
    - ORDINAL
    - NOMINAL
  - QUANTITATIVA
    - DISCRETA
    - CONTÍNUA

EXPRESSA ORDEM Ex: produção de cana: alta, média ou baixa

ORDINAL

NOMINAL

EXPRESSA NOME Ex: nomes das variedades, cor de colmo

Resultante de contagem Ex: nº de perfilhos por touceira

DISCRETA

QUANTITATIVA

Números
ou
Valores

CONTÍNUA

Resultante de mensurações Ex: peso do colmo; TCH

Para cada tipo de variável há uma metodologia estatística mais adequada para a sua organização e análise.

DISTRIBUIÇÃO E MEDIDAS DE DISPERSÃO

Na prática utilizamos experimentos ou levantamentos de dados, cujos resultados não podemos prever com exatidão.

Mas, é sempre possível prever um conjunto de **possíveis resultados**, e existem **leis de probabilidade** que governam tais resultados.

O conjunto de todos os possíveis resultados forma a **população dos valores.**

APÓS A COLETA DOS DADOS DO EXPERIMENTO ...

... Os dados devem ser organizados e resumidos antes de proceder qualquer análise estatística de fato.

dependendo da sua natureza

↓

QUALITATIVO OU QUANTITATIVO

A representação dos dados podem ser feitas por tabelas, gráficos e medidas descritivas de dispersão e pela natureza da distribuição em que ocorrem!!

A representação de todos os possíveis valores e probabilidades associadas, é chamada de **distribuição dos valores respostas.**

idéia global da variável em estudo

Exemplos

✓Nesta sala .... Para qual time de futebol você torce???

Podemos resumir os resultados obtidos numa Tabela!!

✓Produtividade de diferentes variedades, nos 3 cortes

Podemos resumir os dados em um gráfico!!

As medidas de dispersão importantes para resumir uma distribuição de valores

Para compreensão dessas medidas, tomemos um exemplo de uma distribuição simples com 5 valores.

$Y=\{1, 2, 3, 5, 9\}$

MÉDIA ARITIMÉTICA

É simplesmente a soma de todos os valores da distribuição, dividido pelo número total deles.

No nosso exemplo: $Y=\{1, 2, 3, 5, 9\}$

MÉDIA (**μ**) = 1 + 2 + 3 + 5 + 9 = 20

MÉDIA (**μ)** = 20/ 5 = 4

PORTANTO... A MÉDIA DESSE CONJUNTO DE VALORES É IGUAL A 4!!!

A média é um exemplo de medida de posição!!

Pode representar todo um conjunto de valores (mensurações)

... Porém QUASE SEMPRE não é suficiente para caracterizar completamente a distribuição dos dados

NO CASO DO NOSSO EXEMPLO..... O **4** NÃO REPRESENTA BEM O CONJUNTO DE VALORES QUANDO COMPARADO COM OUTRO COJUNTO DE DADOS

$Y=\{1, 2, 3, 5, 9\}$ MÉDIA = 4

$Y=\{4, 3, 4, 5, 4\}$ MÉDIA = 4

Se olharmos apenas pela média poderíamos dizer que esses dois conjuntos de valores são iguais

... Mas olhando para as distribuições dos valores observamos que os valores **1** e **9** estão bem distantes dos demais.

DAÍ A IMPORTÂNCIA DAS MEDIDAS DE DISPERSÃO

- **variância;**
- **desvio padrão;**
- **coeficiente de variação**

MAIS UTILIZADAS

Com medidas de dispersão podemos observar as **diferenças individuais** dos dados em torno **do centro de distribuição (média).**

Variância (σ^2)

É a média da soma dos quadrados dos desvios em relação à própria média.

Voltando ao nosso conjunto de valores

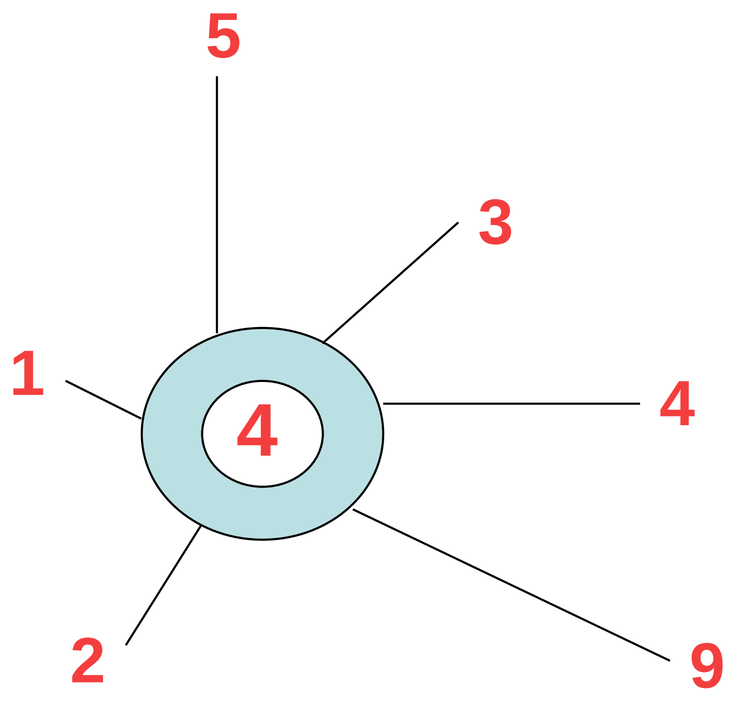
$$Y=\{1, 2, 3, 5, 9\}$$

Lembrando que a média = 4

$$\sigma^2 = ((1 - 4)^2 + (2 - 4)^2 + (3 - 4)^2 + (5 - 4)^2 + (9 - 4)^2)/5 = 8$$

A variância desse conjunto é igual a 8!!

A variância é a medida comumente usada para resumir a variabilidade de uma distribuição, pois **mede a concentração dos dados em torno de sua média.**



Comparando com o conjunto de dados 2...

$Y=\{4, 3, 4, 5, 4\}$

MÉDIA = 4

E A VARIÂNCIA NESSE CASO QUANTO SERÁ?????

$$\sigma 2 = ((4 - 4)^2 + (3 - 4)^2 + (4 - 4)^2 + (5 - 4)^2 + (4 - 4)^2)/5 = 0{,}4$$

AGORA É POSSIVÉL OBSERVAMOS A DIFERENÇA ENTRE OS DOIS CONJUNTOS DE DADOS

CONJUNTO 1

$Y=\{1, 2, 3, 5, 9\}$

MÉDIA = 4

VARIÂNCIA = 8

CONJUNTO 2

$Y=\{4, 3, 4, 5, 4\}$

MÉDIA = 4

VARIÂNCIA = 0,4

PORTANTO PODEMOS CONCLUIR QUE O CONJUNTO 1 POSSUI MAIOR VARIABILIDADE QUE O CONJUNTO 2

Desvio padrão (σ)

Corresponde à raiz quadrada da variância, portanto possui a mesma unidade da média.

No nosso exemplo $Y=\{1, 2, 3, 5, 9\}$

Lembrando que a variância = 8

$$\sigma = \sqrt{\sigma^2} = 2{,}83$$

O desvio padrão desse conjunto é igual a 2,83!!

É considerada uma medida básica de variabilidade, por ser expressa na mesma unidade de valores do conjunto de dados, facilitando a interpretação.

No caso do conjunto 2 $Y=\{4, 3, 4, 5, 4\}$

Lembrando que a variância = 0,4

$\sigma = \sqrt{\sigma^2} = 0{,}63$

O desvio padrão desse conjunto é igual a 0,63!!

Coeficiente de variação (CV)

- O coeficiente de variação é uma medida de variação relativa, a qual expressa o desvio padrão como uma porcentagem da média, ou seja, é o desvio padrão expresso na mesma unidade da μ (em %).

- É uma medida de variabilidade que deve ser usada quando se **compara variabilidades de diferentes conjuntos de dados.**

ENTÃO, NO CONJUNTO 1

DESVIO PADRÃO

$$CV = \frac{\sqrt{\sigma^2}}{\mu} . 100$$

MÉDIA

$$CV = \sqrt{8} / 4 \cdot (100) = 70{,}71\%$$

NO CONJUNTO 2

DESVIO PADRÃO

$$CV = \frac{\sqrt{\sigma^2}}{\mu} . 100$$

MÉDIA

$$CV = \sqrt{0{,}4} / 4 \cdot (100) = 15{,}81\%$$

ASSIM:

CONJUNTO 1 $\neq$ CONJUNTO 2

$\downarrow$ $\downarrow$

CV = 70,71% **CV =15,81**

PORTANTO O CONJUNTO **1** POSSUI MAIOR VARIABILIDADE DO QUE O CONJUNTO **2**

DISTRIBUIÇÕES DE PROBABILIDADE

São importantes para descreverem o comportamento de um conjunto de dados de variáveis aleatórias, sejam elas discretas ou contínuas.

FORMAM A BASE DA TEORIA ESTATÍSTICA

Assim temos

- **distribuições discretas de probabilidade**
- **distribuições contínuas de probabilidade**

Discutiremos aqui as principais distribuições discretas e contínuas de probabilidade!!!

Distribuições discretas de probabilidade (variáveis discretas)

- Distribuição Binomial
- Distribuição de Poisson

DISTRIBUIÇÃO BINOMIAL

EXEMPLOS DE SITUAÇÕES QUE ILUSTRAM **SUCESSO OU FRACASSO**; **SIM OU NÃO**, **MACHO OU FÊMEA**, **GERMINOU OU NÃO** ETC... REPETIDOS n VEZES

Para entendermos, vamos considerar um experimento, onde foram colocadas 4 gemas de cana-de-açúcar para germinar (brotar).

4 GEMAS

- mesma idade
- mesma variedade
- postas para brotar de forma isolada, de modo que uma não pudesse interagir na brotação da outra.
- substrato e condições de umidade iguais.

Variável medida Y= número de gemas brotadas

Como há 4 gemas, cada uma delas poderá germinar ou não.

POSSIVÉIS RESULTADOS

zero (nenhuma brotada)

uma

duas

três gemas brotadas

quatro (todas brotadas).

y_i

Dessa forma podemos escrever: Y={0, 1, 2, 3, 4}

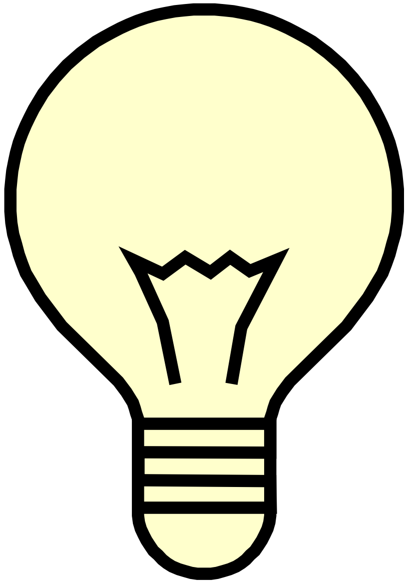
DISTRIBUIÇÃO DOS VALORES

Qual a chance, em um experimento isolado, de sair qualquer valor particular y dos possíveis descritos em Y?

seja, p= ½ (50%), a probabilidade de cada gema brotar (gb)

Então

a probabilidade de cada gema não brotar (nb) é (1 – p) = ½ (50%).



Para Y=0 temos, (nb), (nb), (nb), (nb) e a probabilidade desse resultado é P(Y=0) = P(0) = ½ x ½ x ½ x ½ = **1/16**

Para Y=1, qualquer uma das 4 gemas deverá brotar, ou seja, ser (gb) e as outras 3 deverão ser (nb). Há, portanto 4 possibilidades P(Y=1)= P(1) = **4 x 1/16**, onde 4 é o número de seqüências com 3 gb e uma nb.

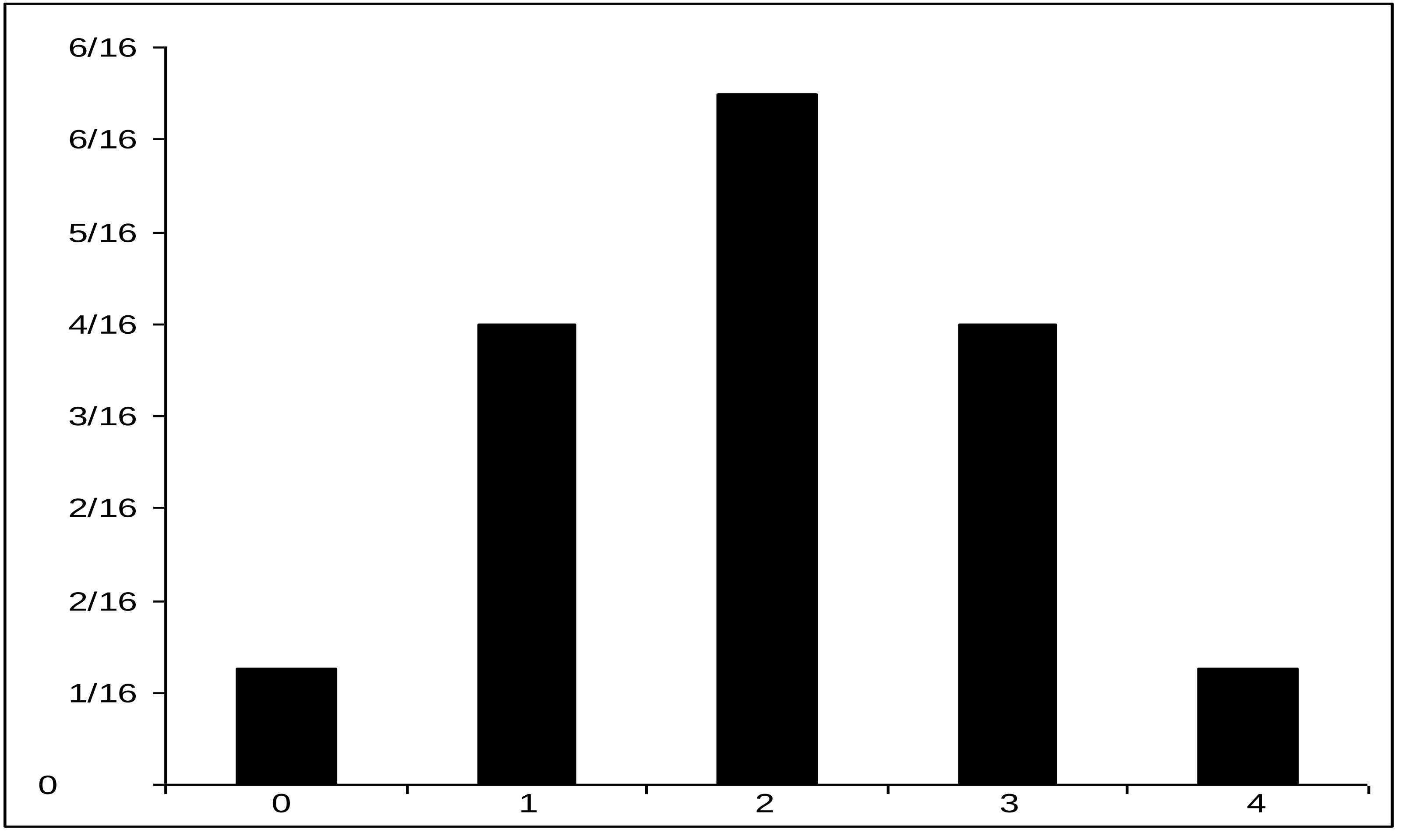
Para Y=2, temos quaisquer 2 gemas que brotarão (gb) e 2 gemas que não brotarão (nb), assim P (Y=2) = P(2) = **6 x 1/16**

Para Y=3, quaisquer 3 gemas brotarão (gb) e uma não brotará (nb), P(Y=3) = P(3) = **4 x 1/16**

Para Y=4, todas as gemas deverão brotar, (gb), ou seja, P(Y=4) = P(4) = **1/16.**

TEMOS UM TOTAL DE 16 RESULTADOS POSSIVÉIS

Podemos mostrar a distribuição de probabilidades através do gráfico:



A FUNÇÃO ABAIXO REPRESENTA A DISTRIBUIÇÃO BINOMIAL:

probabilidade $\longleftarrow$ $$\mathbf{P\ (y) = C_{n,y}\ p^{y}\ (1\text{-}p)^{n\text{-}y}}$$

y= 0, 1, 2, 3, 4

A distribuição binomial é dependente de **"n" e de "p",** pois é necessário ter **n repetições independentes** de um experimento em que **p é constante** e o resultado de cada experimento isoladamente não interfere no resultado de outro experimento, ou seja, os experimentos são independentes.

POR EXEMPLO

SE FOREM REALIZADOS 1000 EXPERIMENTOS IGUAIS A ESSE (com brotação de 4 gemas de cana-de-açúcar)

A freqüência esperada de nenhuma gema brotar será:

f (0)= p(0) x 1000 = 1/16 x 1000 = 62,5

E assim para as outras probabilidades:

f (1)= p(1) x 1000 = 4/16 x 1000= 250
f (2)= p(2) x 1000 = 6/16 x 1000= 375
f (3)= p(3) x 1000 = 4/16 x 1000= 250
f (4)= p(4) x 1000 = 1/16 x 1000= 62,5

+ **1000 experimentos**

PROPRIEDADES BINOMIAL

- Há n repetições independentes de um experimento simples (sucesso/fracasso) com probabilidade p constante.
- X= nº de sucessos, $P(x) = C_{n,x}\, p^x\, (1-p)^{(n-x)}$, $x=1,2,3,...,n$;
- Média de X = np;
- Variância de X = np(1 – p)

Sobredispersão em relação à binomial

- 0corre quando há agregação, reboleiras, contágio etc
- Seja um experimento repetido 10 vezes em que se colocou para germinar 6 toletes com 3 gemas. Resultados: 12,8,11,14,11,9,9,14,12,7; média =10,7; variância=5,79.
- Esperado binomial: Média de X = np; ou seja, p=10,7/18=0,59.
- Variância de X = np(1 – p); ou seja 18. 0,59.(1-0,59)=4,34.
- Sobredispersão de 30%, 5,79/4,34.
- **Nesse caso, é melhor avaliar cada tolete como a unidade sim/não.**

Subdispersão em relação à binomial

- Ocorre quando há repulsão, canibalismo etc
- Seja um experimento repetido 10 vezes em que se colocou para desenvolver 6 lagartinhas em cada tolete. Resultados: 1,2,1,1,1,0,1,1,0,0; média =0,8; variância=0,40.
- Esperado binomial: Média de X = np; ou seja, p=0,8/6=0,13.
- Variância de X = np(1 – p); ou seja 6. 0,13.(1-0,13)=0,67.
- Subdispersão de 40% : (0,67- 0,40)/0,67=0,40.
- **Nesse caso, é melhor avaliar cada tolete como a unidade sim/não.**

DISTRIBUIÇÃO DE POISSON

- Esta distribuição ocorre naturalmente quando se deseja **contar o número de eventos de certo tipo, que ocorrem em um intervalo de tempo, ou superfície ou volume.**

- Uma outra ocorrência de variáveis Poisson surge da aproximação da binomial quando **''n'' é grande e ''p'' é pequeno**.

n (repetições) **será grande** quando for maior ou igual a **50**

p (probabilidade) **será pequeno** quando for menor ou igual a **10**

EXEMPLO DE APLICAÇÃO

- NÚMERO DE COLMOS / touceira DE UMA VARIEDADE DE CANA

- NÚMERO DE INSETOS (PRAGAS) POR UNIDADE DE ÁREA EM UM CANAVIAL

- Quando se semeiam volumes de sementes pequenas de difícil individualização,como é o caso da cana-de-açúcar ou de algumas hortaliças.

Se distribuirmos as sementes com uma medida que em média, solta 5 sementes por linha ou por área (cm^2, por exemplo). **Qual a probabilidade de ter 4 sementes em uma linha , após a semeadura?**

PARA ESSE CASO AS PROBABILIDADES PODEM SER ESTUDADAS POR POISSON

$P(y) = (e^{-\mu}\ \mu^{y}) / y!$, onde y= 0, 1, 2, 3...

e = aproximadamente 2,718...

Portanto para o nosso exemplo, a probabilidade de se ter 4 sementes em uma linha é:

$P(4) = (e^{-5}\ 5^{4}) / 4! = 0{,}175$ ou 17,5%

Da mesma forma podemos calcular a probabilidade de ter 3 ou menos sementes ou ainda 6 ou mais sementes.

Exemplos:

Para 3 sementes = **$P(3) = (e^{-5}\ 5^{3}) / 3! = 0{,}140$ ou 14%**

Para 6 sementes = **$P(6) = (e^{-5}\ 5^{6}) / 6! = 0{,}146$ ou 14,6%**

E assim por diante...

PROPRIEDADES POISSON

- Há repetições independentes de um experimento simples (sucesso/fracasso) com probabilidade p constante, EM UM LIMITADO DE MÉDIA CONSTANTE.
- Y= nº de sucessos, $\mathbf{P(y) = (e^{-\mu}\,\mu^{y}) / y!}$, onde y= 0, 1, 2, 3... ,e = aprox. 2,718...
- Média de Y = $\boldsymbol{\mu} =$ np = Variância de Y

Distribuições contínua de probabilidade (variáveis contínuas)

DISRTRIBUIÇÃO NORMAL

- A distribuição normal é a mais comum para as variáveis contínuas.

- Trata-se de uma das mais usadas na prática para análise de dados e para fazer inferências, uma vez que muitas variáveis que se encontram na natureza se distribuem de acordo com a distribuição normal.

- Grande parte das distribuições se aproximam da normal conforme a amostra (n) cresce. Ex: binomial

A distribuição normal representa bem as respostas, quando estas são influenciadas por muitos fatores

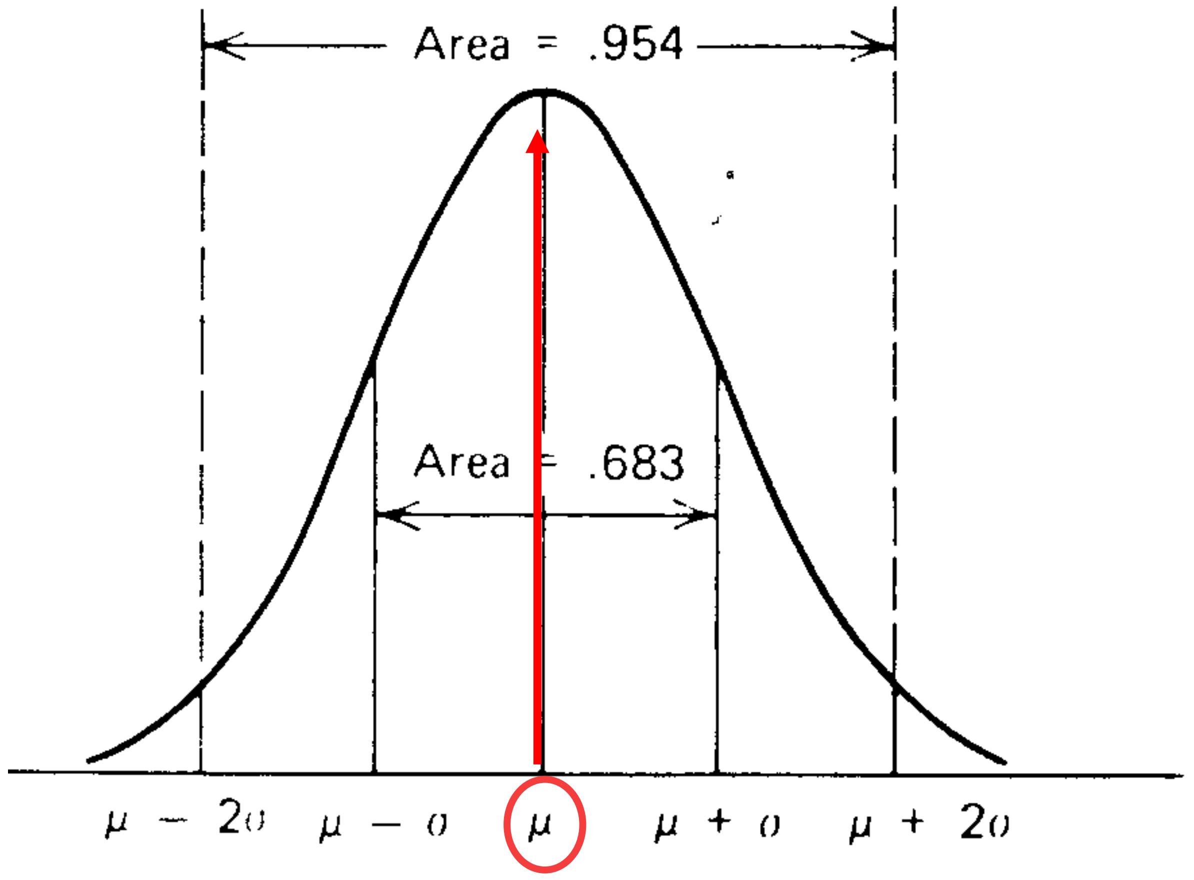
exemplo

são variáveis influenciadas por diversos fatores: genéticos, climáticos, manejo, etc.

altura ou peso dos colmos de uma cultura de cana-de-açúcar

Assim, os diferentes colmos de uma cultura (supondo mesmo local e mesma variedade) tipicamente se distribuem segundo uma distribuição normal.

A **maior parte dos valores estarão próximos da média** geral, mas eventualmente há colmos bem mais pesados ou bem mais altos que a média, assim como bem mais leves ou bem mais baixos.

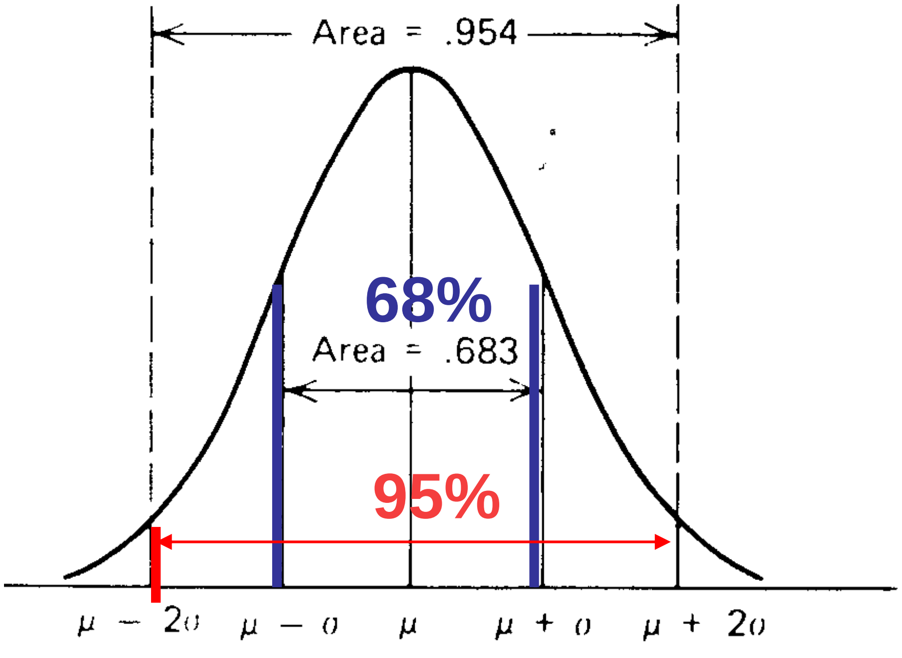


A distribuição normal pode ser descrita totalmente pela média e pelo desvio padrão.

Demonstra-se que

68% dos valores são esperados no intervalo média ± 1 desvio padrão.

95% dos valores no intervalo, são necessários a média ± 2 desvios padrões.



Para calcular a probabilidade entre dois valores dentro da curva utilizamos o auxilio da tabela normal padrão que disponibiliza essas probabilidades.

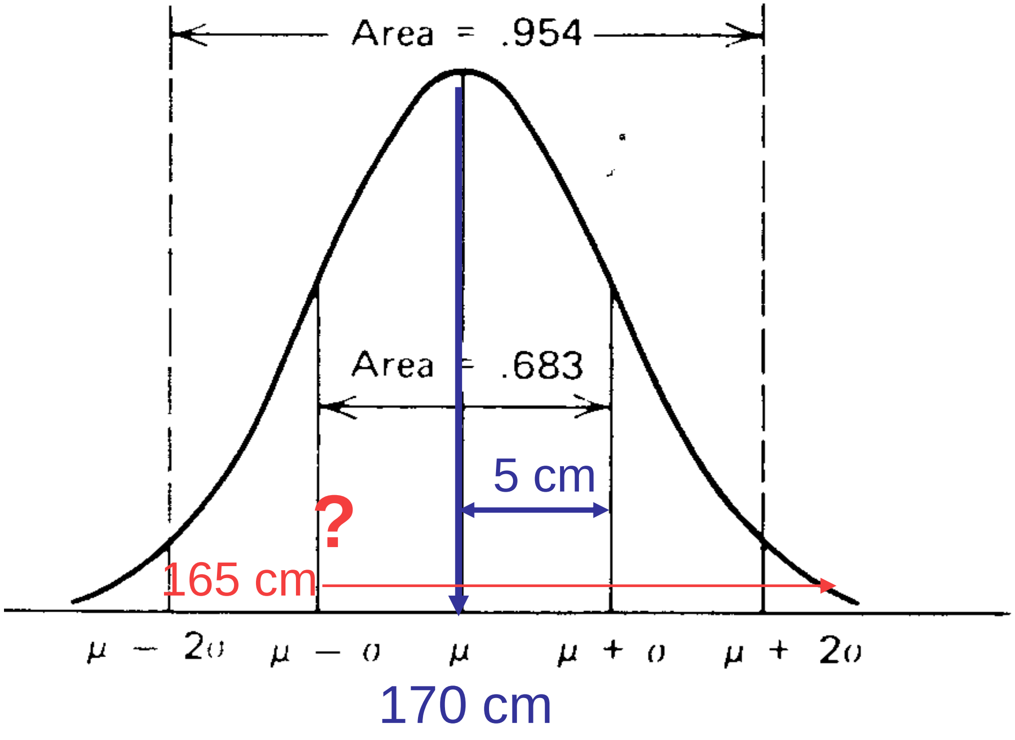
Usa-se a padronização da variável Y em uma variável Z (variável normal padronizada), assim:

$$Z = \frac{Y - \mu}{\sigma}$$

Z tem distribuição normal com média 0 e variância 1.

Há tabelas os cálculos das probabilidades usando Z, vejamos o exemplo a seguir.

Sabendo que a altura média de 10.000 colmos de variedade de cana-de-açúcar são distribuídos normalmente, com média (μ) 170 cm e desvio padrão (σ) 5 cm.



(a) qual é o número esperado de colmos com altura superior a 165 cm?

Solução

P(Y > 165) = $$P\left(\frac{Y-\mu}{\sigma} > \frac{165-170}{5}\right) = P(Z > -1)$$

P(Z > -1) = **0,8413**

Portanto, o número esperado de colmos com altura superior a 165 cm é (10.000 x 0,8413) ≅ **8.413 colmos.**

AMOSTRAGEM

O que vem a ser??

Porque é necessária??

Na realização de qualquer estudo **quase nunca** é possível examinar **todos os elementos da população** de interesse, seja por questões de tempo, economia ou da forma de análise.

EXEMPLO

Se o interesse é examinar a qualidade dos colmos de cana-de-açúcar antes da colheita, não podemos analisar todos os colmos, por questão de viabilidade e de preservação do canavial.

A solução é **selecionar parte dos elementos** (amostra de colmos), analisá-la e inferir propriedades para o todo (população de colmos).

POPULAÇÃO

É o conjunto de indivíduos (objetos), tendo pelo menos uma variável comum observável, ou seja, é constituída por todos os valores possíveis com a distribuição conhecida ou não.

Ex: Todos os colmos de uma determinada variedade de cana ("X"), que são plantados no estado de São Paulo.

Nesse caso, como na maioria das situações na agricultura, é impossível trabalhar com todos os valores!!

AMOSTRAS

AMOSTRA

É qualquer subconjunto da população.

Ex: 100 colmos da variedade “X” de cada região do estado de São Paulo, onde ela é plantada

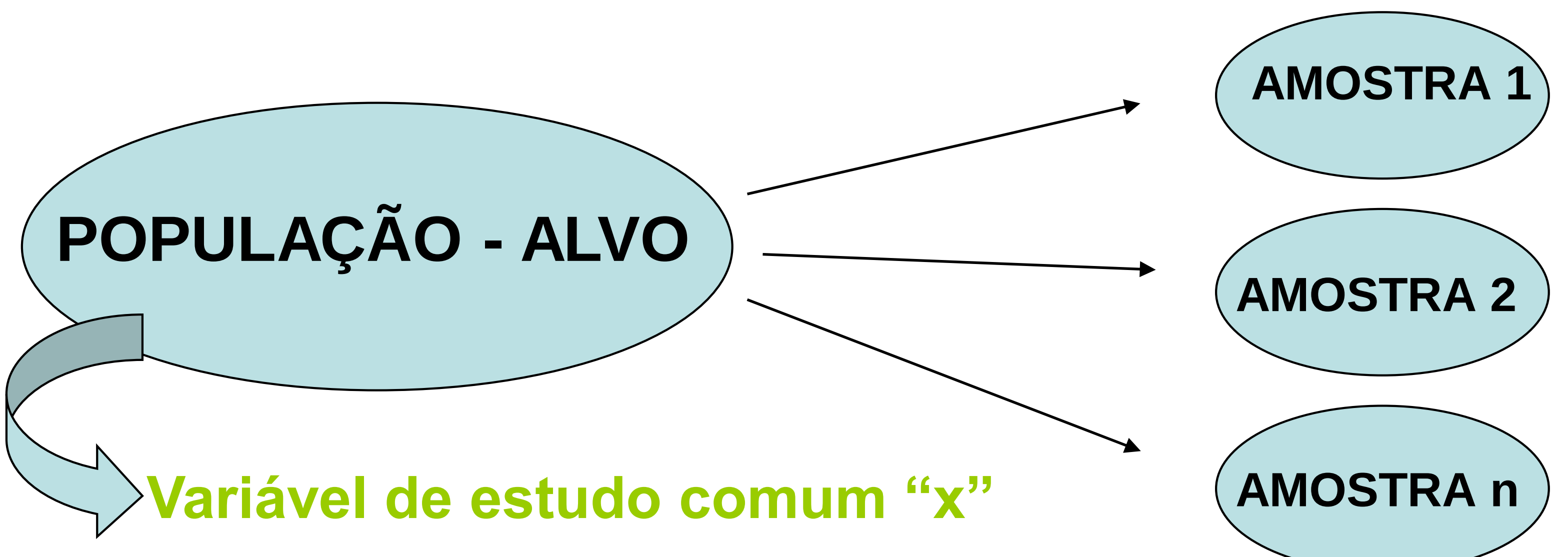
Felizmente, podemos analisar uma amostra, pois há leis que governam as relações entre os valores amostrais e os valores da população da qual a amostra foi extraída, desde que a **amostragem seja bem feita.**

AMOSTRAGEM

É A MANEIRA COMO OBTEMOS A AMOSTRA.

IMPORTANTE

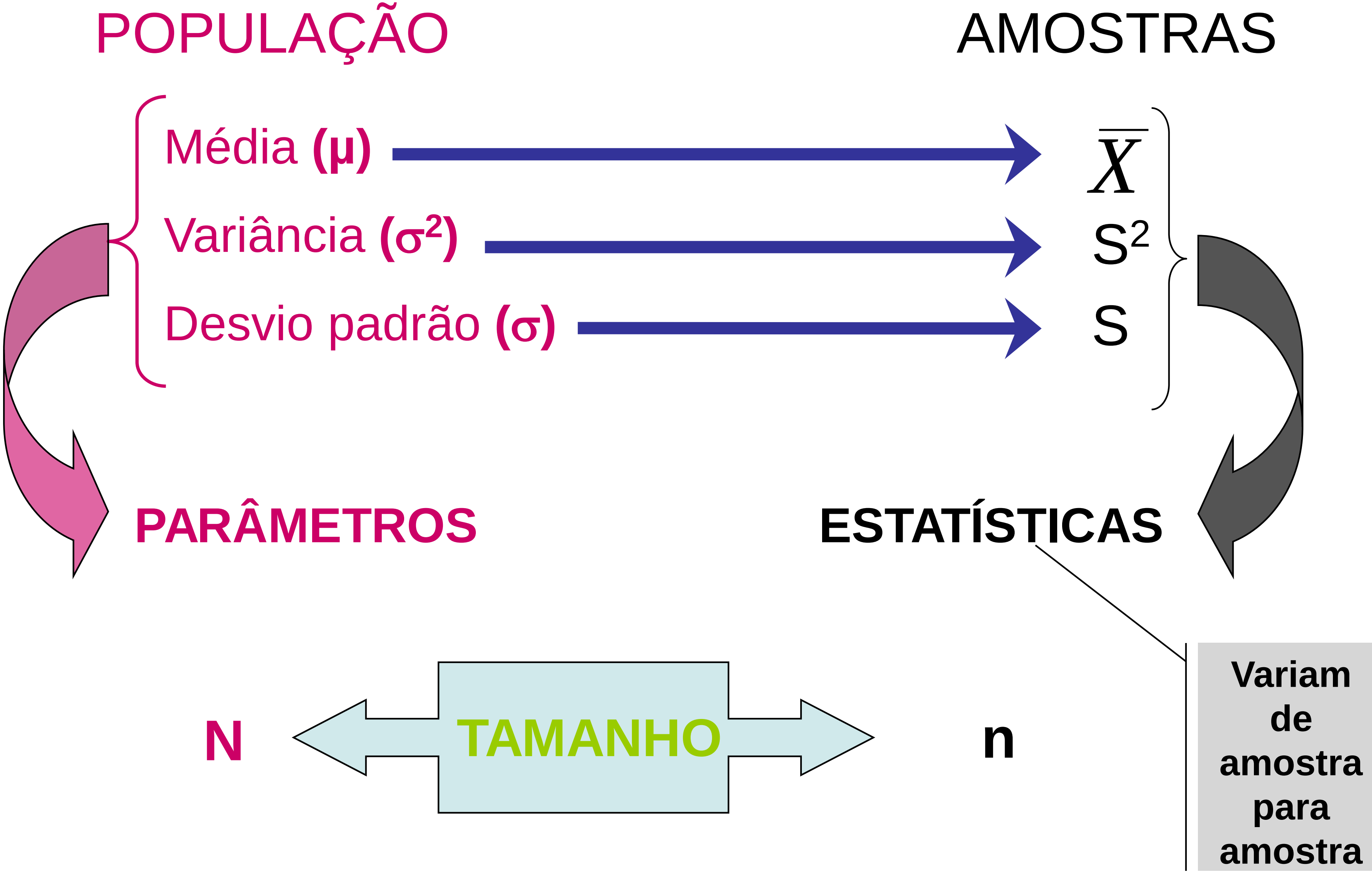
DEVE SER BEM REALIZADA!!



A amostra deve representar bem a população de interesse para que as inferências sejam corretas

O PROCESSO DE ESCOLHA DEVE SER ALEATÓRIO (CASUAL)

Exemplo clássico: SORTEIO PELA URNA OU GLOBO


POPULAÇÃO
AMOSTRAS
Média **(μ)**
Variância **(σ²)**
Desvio padrão **(σ)**
$\bar{X}$
S^2
S
PARÂMETROS
ESTATÍSTICAS
N
TAMANHO
n
Variam de amostra para amostra

2 TIPOS DE AMOSTRAGEM ALEATÓRIA SIMPLES

- **COM REPOSIÇÃO**
- **SEM REPOSIÇÃO**

Amostragem aleatória simples **com reposição**

Os elementos da amostra (n) são selecionados um de cada vez, a partir dos elementos da população (N), **repondo o elemento sorteado** na população antes do próximo sorteio.

Qualquer elemento pode ser sorteado mais do que uma vez.

As n seleções são independentes e **cada elemento na população tem a mesma probabilidade de inclusão na amostra.**

PROPRIEDADES

Denominador da média da população = N

Denominador da média amostral = **n** - - - - - - - → **NÃO CORREÇÃO**

Denominador da variância da população = N

Denominador da variância amostral = **n-1** - - - - - - - → **CORREÇÃO**

Uma vantagem prática deste tipo de amostragem é que, em algumas situações, é uma conveniência importante não ser necessário averiguar se qualquer elemento nos dados está incluído na amostra mais de uma vez.

Amostragem aleatória simples **sem reposição**

A amostra pode ser obtida por n seleções em que, em cada passo (seleção), **todos os elementos não selecionados da população, têm igual chance de seleção.**

Descarta-se seleções repetidas e continua até que n elementos distintos sejam obtidos.

Propriedades COM (A) OU SEM (B) REPOSIÇÃO

(AouB) O valor esperado para média amostral é a Média **(μ)**

(A) O valor esperado para a variância da média amostral é $\boldsymbol{\sigma^2 = (\sigma^2/n)}$.

(B) VARIÂNCIA PRECISA SER CORRIGIDA $\boldsymbol{\sigma^2 = (\sigma^2/n)}$. $[(N-n)/(N-1)]$.

Para população GRANDE (N grande)

NÃO CORREÇÃO: SEM =COM REPOSIÇÃO

ESSAS PROPRIEDADES SÃO FUNDAMENTAIS PARA DEFINIR O TAMANHO DA AMOSTRA

ASSIM: M$\boldsymbol{\sigma}^2$**= (I$\boldsymbol{\sigma}^2$/n)**=VALOR FIXADO.

M$\boldsymbol{\sigma}^2$**= VARIÂNCIA DA MÉDIA**

I$\boldsymbol{\sigma}^2$ =VARIÂNCIA INDIVIDUAL

EXEMPLO : SEJA **I$\boldsymbol{\sigma}^2$ = 64.** Qual o n AMOSTRA ALEATÓRIA para que M$\boldsymbol{\sigma}^2$=4?

FÁCIL VER QUE n=16

PROPRIEDADE DO CV

- NO CASO, COM INDEPENDÊNCIA
- CV (MEDIA)= CV (INDIVIDUAL)/ $\mathbf{n^{1/2}}$
- EXEMPLO:Se amostras compostas, com n1=4 tradagens para um atributo de solos, mostra CV=30%, qual o n2 para CV= 10%
- 30 = CV (4)= CV (INDIVIDUAL)/ $4^{1/2}$
- CV (INDIVIDUAL) = 60
- 10 = CV (INDIVIDUAL)/ $(n2)^{1/2}$; n2=36