UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
CADERNOS DE MATEMÁTICA E ESTATÍSTICA
SÉRIE B: TRABALHO DE APOIO DIDÁTICO

ESTATÍSTICA
UMA LINGUAGEM PARA DIALOGAR COM A INCERTEZA

JOÃO BEAL VARGAS

SÉRIE B, N° 54
PORTO ALEGRE, AGOSTO DE 2000

VARGAS, João Beal. Estatística, Uma Linguagem para Dialogar com a Incerteza. Departamento de Estatística, Instituto de Matemática (UFRGS). 2000.

Gráfica C.E.U.E.

AGRADECIMENTOS :

À LÚCIA GIRAFFA, do Centro de Informática da PUC / RS, pela capa.

À ELENICE, pela datilografia.

À MARILEI, pela digitação cirúrgica.

À MARY SHELLEY, pela idéia.

ÍNDICE

INTRODUÇÃO

Este trabalho é uma tentativa de elucidar um pouco algumas relações fundamentais entre uma metodologia para analisar dados denominada ESTATÍSTICA e a realidade de cada ser humano. É também uma tentativa de ensinar alguns conceitos estatísticos básicos. Não é um manual de receitas.

As expressões "distribuição verdadeira", "distribuição subjacente" e "distribuição populacional" são consideradas equivalentes aqui. Como pode haver distribuições estimadas, hipotizadas ou assumidas, além da já mencionada, apenas o termo "distribuição" pode ser ambíguo em muitas ocasiões. Como também pode haver parâmetros estimados, hipotizados ou assumidos, será usada a expressão "parâmetro verdadeiro", ao invés de simplesmente "parâmetro". Os termos e expressões que não estiverem definidos aqui devem ter seus significados procurados em outros textos, a não ser aqueles cuja clareza dispense tal tarefa.

O Apêndice 1 é um pequeno oásis para os que já se sentiram inconformados com alguma imbecilidade científica ... ou dos cientistas. O Apêndice 3 é uma parte da resposta à pergunta: "Qual é a maneira de pensar deste professor?", que meus alunos fazem para si mesmos no primeiro dia de aula.

LIÇÃO 1 - PROBABILIDADE

$44 = 12 + 32$ reforço

Um EXPERIMENTO é, a rigor, qualquer procedimento que pode ser repetido, ao menos na nossa imaginação, tantas vezes quantas desejarmos e que, em cada uma destas repetições, produz um resultado. Ficaria subentendido, mas vamos deixar explícito que todas as repetições, reais ou imaginárias, são realizadas sob as "mesmas" condições.

Um EXPERIMENTO DETERMINÍSTICO é qualquer experimento que, sempre que for realizado, produz o "mesmo" resultado.

Um EXPERIMENTO ALEATÓRIO é qualquer experimento que, sempre que for realizado, não produz necessariamente o "mesmo" resultado, apesar de ser sempre realizado sob as "mesmas" condições. Cada experimento aleatório tem um ESPAÇO AMOSTRAL, ou seja, um conjunto cujos elementos representam todos os possíveis resultados que podem ser produzidos pelo experimento. Assim, um espaço amostral é um conjunto com pelo menos 2 (dois) elementos.

Um EVENTO é qualquer subconjunto de um espaço amostral. Um EVENTO ELEMENTAR (ou SIMPLES) é qualquer evento com um único resultado. Um EVENTO COMPOSTO é qualquer evento com dois ou mais resultados.

Quando vamos realizar um experimento aleatório, não podemos predizer, com certeza, qual resultado ocorrerá. Por que? Porque há mais do que um resultado possível, isto é, há uma variabilidade nos resultados das realizações do experimento. À primeira vista pode parecer estranho, mas "variabilidade" e "incerteza" são, neste contexto, palavras sinônimas.

Qualquer ser humano dotado de um cérebro que esteja funcionando, após observar várias repetições de um experimento aleatório, começa mesmo "sem querer" a formular relações quantitativas entre as freqüências das ocorrências dos resultados que ele observou. O quociente entre o número de vezes que o resultado ocorreu e o número de vezes que o experimento foi realizado é denominado PROBABILIDADE (DO RESULTADO). Dizemos que um EVENTO OCORRE se, e somente se, o resultado (de uma realização do experimento aleatório) é um dos seus elementos. Assim, a PROBABILIDADE (DE UM EVENTO) é o quociente entre o número de vezes que o evento ocorreu e o número de vezes que o experimento foi realizado. Desta forma, representando o espaço amostral por Ω, os eventos por A, B, C,... e as probabilidades dos eventos por PA, PB, PC, ... temos as seguintes relações:

(1) $P(\varnothing) = 0$ e $P(\Omega) = 1$

(2) $0 \leq P(A) \leq 1$, para qualquer evento A

(3) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(4) $P(A) + P(A^c) = 1$

Verificamos, assim, que uma PROBABILIDADE é , matematicamente, uma função que, a cada evento de um espaço amostral associa um número. A rigor, há três maneiras de atribuirmos números (probabilidades) aos eventos: FREQÜENTISTA ou EMPÍRICA, AXIOMÁTICA ou DEDUTIVA e SUBJETIVA ou BAYESIANA.

REFORÇOS (Respostas na página 26)

1) Um provador de vinhos deve experimentar 3 tipos rotulados com A, B e C e ordená-los de acordo com sua qualidade. Se ele é um embusteiro, qual a probabilidade de que o vinho A seja classificado em 3º lugar e qual a probabilidade de que o vinho A seja classificado como melhor do que o vinho C?

2) Um dado é lançado duas vezes. Supondo que ele seja HONESTO (isto é, não viciado) qual a probabilidade de que a soma das faces voltadas para cima nos dois lançamentos seja maior ou igual a 9?

3) Qual a probabilidade de ocorrer no máximo uma cara no lançamento de 3 moedas EQUILIBRADAS?

4) Numa gaveta há 6 canetas, duas sem tinta e quatro com tinta. Três delas são retiradas no escuro, sucessivamente, AO ACASO e sem reposição. Qual a probabilidade de que sejam escolhidas 3 canetas com tinta? Qual a probabilidade de que as duas sem tinta estejam entre as 3 que foram retiradas?

5) A quantidade de calor desprendida numa reação química exotérmica específica pode ser considerada o resultado de um experimento? Em caso afirmativo, você considera tal experimento aleatório ou determinístico?

6) O que significa a frase "a probabilidade de um resultado ocorrer é 3/17"?

7) A probabilidade de um pedreiro ser alcóolatra é 0,1. Para construir um galpão em sua propriedade rural, um empresário necessita contratar 3 pedreiros. Qual é a probabilidade de que no máximo um deles seja alcóolatra?

8) Podemos ter $P(A \cup B) = P(A)$?

9) Uma urna tem 5 bolas numeradas de 1 a 5. Duas delas são extraídas ao acaso e sem reposição. Qual é a probabilidade de que elas apresentem números consecutivos (por exemplo, 34 e 43 são considerados consecutivos)? Se um prêmio no valor do maior número é oferecido, quais os valores possíveis para estes prêmios e qual a probabilidade de cada um?

10) No Reforço 7, quais as quantidades possíveis de pedreiros alcóolatras e suas probabilidades ?

11) Uma mulher se diz paranormal e um homem dela duvida. Para resolver a querela, 4 cartas são sorteadas ao acaso e com reposição de um BARALHO COMPLETO e a mulher deve adivinhar o naipe de cada carta. Se ela for uma embusteira, qual é a probabilidade de que ela acerte mais do que a metade? Quais os valores possíveis de acertos e suas probabilidades ?

12) Calcule $P(A^c)$, $P(B^c)$, $P(A \cap B^c)$ e $P(A \Delta B)$ supondo que $P(A) = 0{,}04$, $P(A \cup B) = 0{,}16$ e $P(A \cap B) = 0{,}01$. Os eventos A e B são (MUTUAMENTE) EXCLUSIVOS?

Uma VARIÁVEL ALEATÓRIA é uma função real definida num espaço amostral. Seus valores, associados com as probabilidades respectivas de suas ocorrências, são a DISTRIBUIÇÃO DA VARIÁVEL ALEATÓRIA. Vamos destacar, a seguir, alguns números obtidos da distribuição de uma v.a. que tem sido muito importantes nas diversas áreas do conhecimento humano. É preciso salientar que alguns deles podem não existir em várias situações.

Seja X uma variável aleatória com valores num conjunto enumerável C. Representamos por

(5) $p_X(x) = P\{X = x\} = P\{\omega \in \Omega : X(\omega) = x\}, \forall x \in C.$

A ESPERANÇA (ou MÉDIA) de X e a VARIÂNCIA de X são definidas por:

(6) $E(X) = \sum_{x \in C} x \cdot p_X(x).$ e $Var(X) = \sum_{x \in C} (x - E(X))^2 \cdot p_X(x).$

A fórmula a seguir simplifica mais os cálculos, mas não permite uma interpretação:

(7) $Var(X) = E(X^2) - (E(X))^2$, onde $E(X^2) = \sum_{x \in C} x^2 \cdot p_X(x).$

O DESVIO PADRÃO de X é, simplesmente, $Des(X) = \sqrt{Var(X)}$.

O DESVIO MÉDIO de X é definido por $Dm(X) = \sum_{x \in C} |x - E(X)| \cdot p_X(x).$

A MEDIANA de X, Med (X), é qualquer número real x tal que: $P\{X \le x\} \ge \frac{1}{2}$ e $P\{X \ge x\} \ge \frac{1}{2}$.

A MODA de X, Mod (X), é qualquer valor x do conjunto C tal que $p_X(x) \ge p_X(t)$, para todo $t \in C$, e $p_X(a) > p_X(a)$, para algum $a \in C$.0

Um VETOR ALEATÓRIO é qualquer função que associa, a cada elemento de um espaço amostral, uma lista ordenada de k números reais. Se k=2, temos um VETOR ALEATÓRIO BIDIMENSIONAL; se k=3, temos um VETOR ALATÓRIO TRIDIMENSIONAL; etc. A DISTRIBUIÇÃO DE UM VETOR ALEATÓRIO é formada por todas as listas que ele pode assumir, associadas com as probabilidades respectivas da ocorrência de cada uma destas listas.

Consideremos um experimento aleatório que foi realizado (uma única vez) e cujo resultado não é conhecido com exatidão: sabe-se apenas que ele é um dos resultados de um evento completamente especificado B. Esta informação modifica por inteiro a estrutura probabilística existente antes da realização do experimento: por exemplo, agora a probabilidade do evento B^c ter ocorrido é zero. Esta probabilidade modificada pela informação da ocorrência do evento B é denominada PROBABILIDADE CONDICIONAL DADO O EVENTO B e é expressa por

(8) $$P_B(A) = P(A / B) = P(A \cap B) / P(B).$$

A e B são EVENTOS INDEPENDENTES se, e somente se, $P(A \cap B) = P(A) \cdot P(B)$; esta condição implica $P(A / B) = P(A)$ e $P(B / A) = P(B)$. Duas variáveis aleatórias X e Y são denominadas VARIÁVEIS ALEATÓRIAS INDEPENDENTES se, e somente se, qualquer evento descrito apenas por uma delas for independente de qualquer outro evento descrito apenas pela outra, isto é, se $A \subset C_X$ e $B \subset C_Y$, então $\{X \in A\}$ e $\{Y \in B\}$ são eventos independentes, ou seja,

(9) $$P\{X \in A \text{ e } Y \in B\} = P\{X \in A\} \cdot P\{Y \in B\}.$$

Demonstra-se que uma condição necessária e suficiente para que isto ocorra é que a distribuição (conjunta) do vetor aleatório (X,Y) seja o produto da distribuição (marginal) de X pela distribuição (marginal) de Y, isto é,

(10) $$p_{X,Y}(x,y) = p_X(x) \cdot p_Y(y), \quad \forall\, x \in C_X \text{ e } \forall\, y \in C_Y.$$

O COEFICIENTE DE CORRELAÇÃO LINEAR das variáveis aleatorias X e Y é definido pela expressão:

(11) $$\text{Cor}(X,Y) = \frac{E(X \cdot Y) - E(X) \cdot E(Y)}{\text{Des}(X) \cdot \text{Des}(Y)}, \text{ onde } E(X \cdot Y) = \sum\sum x \cdot y \cdot p(x,y).$$

O numerador da expressão (14) é a COVARIÂNCIA DAS VARIÁVEIS ALEATÓRIAS X e Y, anotada Cov (X,Y). Demonstra-se que:

(12) Se X e Y são independentes, então $\text{Cor}(X,Y) = 0$.

(13) $-1 \leq \text{Cor}(X,Y) \leq 1$.

(14) $\text{Cor}(X,Y) = \text{Cor}(Y,X)$.

(15) $\text{Cov}(X,X) = \text{Var}(X)$.

(16) $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2 \cdot \text{Cov}(X,Y)$.

(17) $\text{Cov}(aX+b, cY+d) = a \cdot c \cdot \text{Cov}(X,Y)$.

(18) $\text{Cor}(aX+b, cY+d) = \text{sgn}(a \cdot c) \cdot \text{Cor}(X,Y)$.

(19) $\text{Cor}(X,Y) = +1$ se, e somente se, $P\{Y = aX+b\} = 1$, para algum $a > 0$.

(20) $\text{Cor}(X,Y) = -1$ se, e somente se, $P\{Y = aX+b\} = 1$, para algum $a < 0$.

REFORÇOS (Respostas na página 26)

13) Na tabela ao lado, temos observações referentes ao rendimento acadêmico global Y, expresso em 5 níveis, e ao número de disciplinas cursadas X. Calcula: $P\{X \leq 3\}$, $P\{X = 1$ ou $Y = D\}$, $P\{X \leq 3 | Y = A\}$, $P\{Y \neq A\}$, $P\{X = 3 | Y = D\}$ e $P\{X > 5, Y \neq D\}$. Estes dois critérios de classificação são independentes? Calcula EX, Des X, Mod X e Med X.

Y \ X	1	2	3	4	5	6	7
A	13	26	44	32	28	15	21
B	15	37	40	26	45	18	13
C	21	43	32	84	36	10	9
D	2	6	12	18	9	8	48
E	16	19	18	12	9	3	12

14) Sabemos que PA = PB = 0,004 e $P(A \cap B) = 0,000024$. Calcula $P(A \cup B)$, $A(A|B)$, $P(B|A)$, $P(A^c)$, $P(B^c)$. Os eventos A e B são independentes?

15) Num casarão há 20 cães e 15 gatos classificados de acordo com a tabela a seguir, que identifica a cor predominante. Um animal morreu. Qual é a probabilidade de que seja um gato? Qual é a probabilidade de que seja um gato, dado que sua cor predominante era branca? A cor predominante e o tipo de animal eram independentes? Antes do animal morrer, houve uma inspeção na qual 3 animais foram sorteados ao acaso. Qual é a probabilidade de que tenham sido inspecionados um gato e dois cães?

	branca	preta	outra
caẽs	10	7	3
gatos	8	5	2

16) Repita o Reforço 11, mas considerando que a mulher deve adivinhar apenas a cor da carta. Seja Y o número de acertos. Calcula EY, Des Y e Mod Y e Med Y.

17) Uma urna tem uma bola "1", três bolas "2", cinco bolas "3" e uma bola "4". Uma bola é sorteada ao acaso e seu número X é observado. Sem recolocá-la, outra bola é sorteada ao acaso e seu número Y é observado. Calcula o coeficiente de coordenação linear. X e Y são independentes?

18) No experimento do Reforço 17, seja Z a diferença em valor absoluto dos números das duas bolas sorteadas. Qual é a distribuição de Z?

19) No experimento do Reforço 1, qual é a probabilidade de que o vinho A seja classificado em 3º lugar, dado que o vinho B foi classificado em 1º lugar?

20) Qual a probabilidade de ocorrer no máximo uma cara em 3 lançamentos sucessivos de uma moeda cuja probabilidade de dar cara é 2/5 da de dar coroa?

21) Um dado honesto é lançado 2 vezes. Qual é a probabilidades de que a soma das faces seja pelo menos igual a 9 ? Qual a probabilidade de que a soma das faces seja pelo menos igual a 9, dado que sua diferença em valor absoluto é, no máximo, igual a 3 ?

22) Uma variável aleatória X tem a distribuição dada pela tabela a seguir. Calcula EX, Des X, Mod X e Med X.

x	-2	-1	0	1	2
$p_X(x)$	1/15	2/15	3/15	4/15	5/15

23) Sabe-se que 60% dos "gays" de Los Angeles são bigodudos, que 10% da população masculina da cidade é "gay" e que 20% dos heterossexuais do sexo masculino de Los Angeles são bigodudos. (Na prática, estas quantias são estimadas através de amostras e dificilmente seriam assim tão exatos). Você está em Los Angeles bebendo um "drink" no balcão de um bar e um bigodudo senta ao seu lado. Qual é a probabilidade de que ele seja "gay"?

24) Três cartas são retiradas sem reposição de um baralho completo. Qual é a probabilidade de que ocorra:

a) pelo menos um ás?
b) dois reis e um ás?
c) exatamente dois reis?
d) dois números e uma letra?

25) Uma moeda honesta e um dado equilibrado são lançados. Qual é a probabilidade de ocorrer cara e face par? Qual é a probabilidade de ter ocorrido cara, dado que ocorreu face par?

26) Quais são os objetos de estudo da Teoria das Probabilidades?

27) A observação de quanto vai chover na Esquina Democrática durante as 24 horas do dia 1º de maio do próximo ano é um experimento aleatório?

28) O que significa "a probabilidade de um evento ocorrer é 0,72"?

29) Uma variável aleatória Y tem a distribuição tabelada a seguir. Calcule EY, Des Y, Mod Y e Med Y.

y	-3	-1	0	1	3	5
$p_Y(y)$	0,15	0,25	0,05	0,10	0,20	0,25

30) Numa urna há 3 bolas brancas e 5 amarelas. Uma bola é retirada ao acaso e, sem que seja recolocada, outra é retirada também ao acaso. Qual é o seu espaço amostral para este experimento? Qual é a probabilidade do evento "a 1ª bola é branca"? Qual é a probabilidade do evento "a 2ª bola é branca"?Qual é a probabilidade condicional da 1ª bola ser branca dado que a 2ª é branca? Qual a probabilidade condicional da 2ª ser branca dado que a 1ª é branca?

31) Uma sacola tem 11 fichas com cor amarela de um lado e a cor branca do outro. Em cada lado há um número de acordo com a distribuição a seguir, onde "a" é o número do lado amarelo e "b" é o número do lado branco de cada ficha. Uma das fichas é sorteada ao acaso. As variáveis aleatórias A e B são independentes? Calcule $P\{A>3|B>2\}$, $P\{B=4 \mid A<3\}$ $P\{B<4 \mid A>2\}$ e Cor $[A, B]$.

(a, b)	(1,2)	(1,3)	(2,4)	(3,3)	(3,4)	(4,2)	(4,4)
nº de fichas	2	3	1	1	2	1	1

32) Um experimento aleatório gera vetores aleatórios (x, y) com as probabilidades expressas pela tabela a seguir. As variáveis aleatórias X e Y são independentes? Calcule o seu coeficiente de correlação. Calcule $P(X=2 \mid Y<0)$, $P(Y=0|X\neq 2)$, $P(X>0|Y\geq 0)$, $P(Y\neq 0|X\neq 0)$, $P(Y=0|X\geq 0)$, $P(Y=0|X<0)$ e $P(X>0|Y=3)$.

x \ y	-2	-1	0	1	2
0	0,08	0,04	0,01	0,04	0,08
1	0,10	-	0,20	-	0,10
2	0,07	0,07	0,07	0,07	0,07

33) A variável aleatória tem a distribuição a seguir. Calcule EW, Des W, Mod W e Med W.

w	-9	-6	-3	0	2	4	6
p_w (w)	0,25	0,20	0,10	0,05	0,10	0,20	0,10

34) A probabilidade de que um doente seja alérgico a determinada substância química é 0,1. Qual é a probabilidade de que dois doentes examinados não sejam alérgicos?

35) Cite um experimento aleatório de sua vida pessoal, seu espaço amostral e as probabilidades que você atribuiu aos resultados. Que probabilidade você usou ?

36) Uma entidade hospitalar trabalha com 40 médicos, 60 enfermeiros, 100 auxiliares e 50 encarregados das demais funções. Todos são obrigados a entrarem e saírem por uma única porta de serviço. Suponha que as pessoas saiam ao acaso e que você possa entrevistá-las ao saírem. Qual é a probabilidade de que sua primeira entrevista seja com um médico? Qual é a probabilidade de que suas três primeiras entrevistas sejam com enfermeiros?

37) Qual o tipo de probabilidade que as pessoas usam geralmente para tomar decisões em sua vida pessoal? Exemplifique.

38) Uma variável aleatória X tem a distribuição de probabilidade ao lado. Calcule EX, Res X, Med X e Mod X.

x	1	2	3	4	5
$p(x)$	0,2	0,25	0,3	0,1	0,15

39) Um dado honesto é lançado. Calcule a probabilidade de ocorrer face par dado que ocorreu uma face diferente de 6. Estes eventos são independentes? Calcule a probabilidade de ocorrer face par dado que ocorreu uma face superior a dois. Estes eventos são independentes?

40) Num hospital, os doentes são classificados através de duas variáveis: X (sexo: 1 - masculino, 2 - feminino) e Y (estado clínico: 0 - inspira cuidados, 1 - em observação, 2 - em recuperação), cuja probabilidade conjunta é:

x \ y	0	1	2
1	0,25	0,10	0,15
2	0,28	0,12	0,10

Calcule $P\{X = 1\}$, $P\{Y = 0\}$, $P\{X = 1, Y=0\}$, $P\{X = 1 | Y = 0\}$ e $P\{Y = 0 | X = 1\}$ X e Y são independentes? Embora não tenha sentido (por quê?), calcule Cor (X, Y).

41) Uma urna tem três bolas amarelas e quatro vermelhas. Três delas são retiradas sucessivamente ao acaso e sem reposição. Seja X o total de bolas amarelas retiradas. Escreva a distribuição de X. Calcule EX, Des X e Mod X. Calcule a possibilidade da última ser amarela, dado que ocorreram duas amarelas e vice-versa. Calcule a probabilidade da última ser amarela, dado que ocorreram duas vermelhas e vice-versa. Calcule a possibilidade da última ser amarela, dado que a primeira é vermelha e vice-versa.

42) Repita o Reforço 41 para três bolas amarelas e três vermelhas.

43) Repita o Reforço 41 para três bolas amarelas e duas vermelhas.

44) Considere o lançamento de um dado honesto.

a) Face par e face ímpar são eventos independentes?

b) Face par e face maior do que 4 são independentes?

c) Face par e face menor do que 3 são independentes?

Quais destes eventos são elementares?

LIÇÃO 2 - POPULAÇÃO E VARIÁVEIS

7 reforços

Uma POPULAÇÃO é qualquer conjunto bem definido com pelo menos dois elementos, ou seja, dado um elemento qualquer sabemos se ele pertence ou não à população. Podemos ter dificuldades com expressões tais como "famílias residentes em determinado município", "residências em determinada vila", "indústria nacional" etc., mas as ambigüidades de tais expressões devem ser eliminadas de início. Anotamos a população por U, já que também a denominamos UNIVERSO.

Uma VARIÁVEL é qualquer função (real) definida em um subconjunto com pelo menos dois elementos da população, isto é, em uma SUBPOPULAÇÃO. Qualquer aplicação com "valores" em algum conjunto de expressões do idioma português definida em uma subpopulação é um ATRIBUTO. Todo atributo pode ser transformado em variável através de uma codificação adequada e vários autores preferem usar a expressão "VARIÁVEL QUALITATIVA" ao invés do termo "atributo". Anotamos as variáveis e os atributos por X, Y etc. e os seus valores por x, y etc.

EXEMPLO 1: Seja U a população formada por todas as residências de Porto Alegre. Seja X o total de menores de 18 anos de cada residência em determinada data. Seja Y a renda residencial líquida de um determinado mês. Seja Z a profissão principal do chefe da família em determinada data. ▲

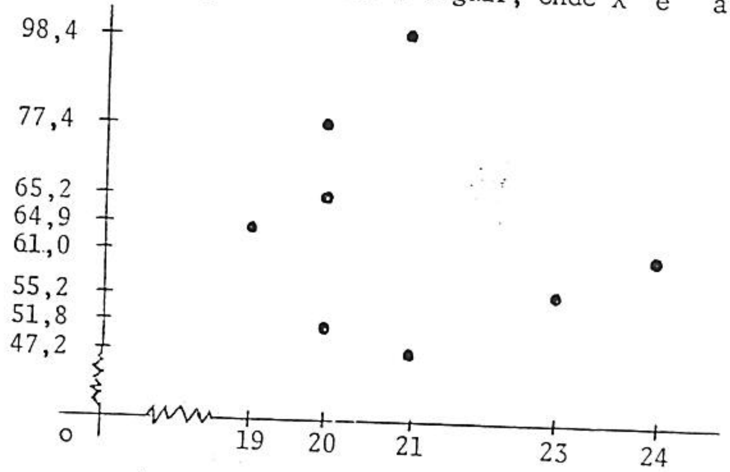
EXEMPLO 2: Seja U o universo formado por todos os alunos de determinado colégio. Sejam X_1 o sexo, X_2 o Q.I., X_3 a idade, X_4 a altura e X_5 a principal atividade profissional do pai. ▲

O TAMANHO DE UMA POPULAÇÃO é a sua quantidade de elementos, que anotamos por N. Uma POPULAÇÃO FINITA é qualquer população limitada em tamanho. Uma POPULAÇÃO INFINITA é qualquer população de tamanho ilimitado. Os exemplos típicos de populações infinitas são as populações resultantes de processos geradores de itens, tais como o arremesso de uma moeda, a retirada com reposição de bolas de uma urna, a fabricação de produtos em série por um equipamento, os nascimentos dos insetos de uma determinada espécie etc.

A DISTRIBUIÇÃO VERDADEIRA DE UMA VARIÁVEL é a função que, a cada $x \in \mathbb{R}^1$, associa sua freqüência relativa de ocorrência na subpopulação onde a variável está definida. Anotamos a freqüência relativa de ocorrência de um valor x na variável X por $f_X(x)$. A soma das freqüências relativas de todos os valores que não excedem o valor x, incluindo portanto a freqüência relativa do próprio, é a FREQUÊNCIA (RELATIVA) ACUMULADA do valor x na variável X, que anotamos por $F_X(x)$. Um PARÂMETRO VERDADEIRO DE UMA VARIÁVEL é qualquer característica numérica descritiva da distribuição verdadeira da variável tal como média, variância etc.

EXEMPLO 3: Considere a população U definida pela tabela a seguir, onde X é a idade e Y é o peso.

u	X(u)	Y(u)
Noé	20	65,2
Édson	20	77,4
Lúcia	23	55,2
Carlos	21	98,4
Alfonso	24	61,0
Júlia	19	64,9
Márcia	20	51,8
Luiz	21	47,2



$\sum X = 168$ $\mu_x = 21{,}0$

$\sum X^2 = 3548$ $\sigma_x = 1{,}6$

$\sum Y = 521{,}1$ $\mu_y = 65{,}14$

$\sum Y^2 = 35815{,}49$ $\sigma_y = 15{,}30$

$\sum XY = 10912{,}3$ $\rho_{x,y} = -0{,}1592$

x	19	20	21	23	24
$f_X(x)$	1/8	3/8	2/8	1/8	1/8
$F_X(x)$	1/8	4/8	6/8	7/8	1

▲

Para obtermos a distribuição verdadeira de uma variável é necessário medí-la em todos os elementos da subpopulação onde ela está definida. Assim, a obtenção da distribuição verdadeira e, conseqüentemente, dos parâmetros verdadeiros envolve fatores tais como custo, rapidez, precisão e outros que tornam esta tarefa impraticável em muitos casos e impossível em outros.

Vamos finalizar enfatizando algo um tanto óbvio, mas muito importante: a distribuição e os parâmetros de uma variável dependem do universo subjacente. Esta frase é, a rigor, sem sentido, uma vez que "número de filhos" não é uma variável, mas "número de filhos de uma mulher que atualmente fuma, cuja idade atual não supera 30 anos e é funcionária do Banrisul" é uma variável.

REFORÇOS (Respostas na página 15)

1) a) Todas as populações são finitas?
   b) Existe diferença entre população e universo?
   c) O que é uma subpopulação?
   d) Existe diferença entre variável e atributo?
   c) Na especificação de uma população pode haver elementos repetidos?

2) a) Uma variável existe "em si" ou somente "dentro de uma situação específica?
   b) Todas as variáveis do Exemplo 2 estão definidas com clareza ou alguma (s) pode(m) gerar controvérsias?

3) Considera as variáveis do Exemplo3. Calcula: $f_x(18)$, $f_x(25)$, $F_x(25)$, $F_x(4)$, $f_y(51,8)$, $f_y(51,9)$, $f_y(65,2)$, $f_y(72,0)$, $f_y(99,1)$, $F_y(45,0)$, $F_y(51,8)$, $F_y(51,9)$, $F_y(65,2)$, $F_y(72,0)$, $F_y(99,1)$.

4) Corrige cada frase a seguir, se necessário, ou escreva "está correta" :

a) Se 1/5 da população ainda não completou 15 anos, então F(15) = 4/5.

b) Se 1/3 da população já completou 18 anos, então F(18) < 1/3.

c) Se metade da população ainda não completou 21 anos, então F(21) >1/2.

d) Uma subpopulação é qualquer subconjunto da população.

e) A freqüência acumulada do valor x é a soma das freqüências dos valores maiores do que x.

f) A freqüência acumulada do valor x é a soma das freqüências dos valores menores do que x.

g) Se 1/3 da população já completou 18 anos, então F(19) ≥ 1/3.

h) Se 1/3 da população ainda não completou 18 anos, então F(20) > 1/3.

i)"Atributo" é uma expressão que designa qualquer variável.

j) A freqüência acumulada de uma variável no valor x é a soma dos valores da variável que são menores ou iguais a x.

5) Considera a variável Z: 4,5 3,2 1,5 8,6 2,3 5,1 6,2 7,3 4,5. Calcula as freqüências absolutas e absolutas acumuladas para cada um dos valores a seguir: 3,2 6,0 8,7 0,4 9,1 8,6 4,5.

6) Calcula as médias, os desvios padrões, os desvios médios, as medianas e as modas para as variáveis do Exemplo 3 e para a variável do Reforço 5.

7) Conceitue "população" e "variável".

RESPOSTAS

(1 a) Não. Por exemplo, todos os lançamentos possíveis de um dado. (1b) Neste texto, não. São sinônimos. (1c) É qualquer subconjunto da população com pelo menos dois elementos. (4d) Idem. (1d) Sim. Uma variável é uma função "quantitativa" definida numa subpopulação e um atributo é uma função "qualitativa" definida numa subpopulação. (1e) Não, pois uma população é um conjunto. (2 a) Somente "dentro de uma situação específica", i.e., após a definição da (sub)população onde ela está definida. (2b) A "principal atividade profissional do pai" pode não estar clara. As demais estão. (3) 0; 0; 1; 0; 1/8; 0; 1/8; 0; 0; 0; 2/8; 2/8; 6/8; 6/8; 1. (4 a) F(15) ≥ 1/5 (4b) F(18) ≥ 2/3 (4c) F(21) ≥1/2 (4e) A freqüência acumulada de uma variável no valor x é a soma das freqüências dos valores da variável menores ou iguais a x. (4f) Idem. (4j) Idem. (4g) F(19) ≥ 2/3 (4h) F(20) ≥ 1/3. (4i) Não. Somente variáveis "qualitativas". (5) f (0,4) = f (6,0) = f (8,7) = f (9,1) = 0; f (3,2) = f (8,6) = 1; f (4,5) = 2; F(0,4) = 0; F(3,2) = 3; F(4,5) = 5; F(6,0) = 6; F(8,6) = F(8,7) = F(9,1) = 9. (6) X (21,0; 1,6; 1,25; 20,5; 20) Y(65,14; 15,30; 11,40; 62,95; não existe) Z(4,8; 2,18; 1,78; 4,5;4,5)

LIÇÃO 3 - CENSO E AMOSTRAGENS $27 = 8 + 19$ reforços

Uma AMOSTRA é qualquer lista ordenada de elementos de uma população, com elementos repetidos ou não. O TAMANHO (ESPECÍFICO) DE UMA AMOSTRA é a quantidade de elementos (distintos) da lista ordenada. Anotamos uma amostra qualquer por "a", seu tamanho por "n (a)" e seu tamanho específico por " ν (a)".

Um ESPAÇO AMOSTRAL é qualquer conjunto de amostras que contenha pelo menos duas. Em cada situação consideramos um espaço amostral específico que anotamos por "A". Uma AMOSTRAGEM é qualquer processo bem definido de seleção de uma única amostra de A segundo algum critério.

EXEMPLO 1: Se a população é formada por todos os adultos de uma cidade e o espaço amostral por todas as listas sem repetições de tamanho 15, então o processo a seguir é uma amostragem: saímos para a rua às 6 horas da manhã e entrevistamos os primeiros 15 adultos residentes na cidade que encontramos. Este processo, técnica ou método de amostragem pode apresentar inconvenientes, tais como: a amostra assim selecionada é inadequada (para os objetivos da pesquisa) algumas pessoas talvez tenham que esperar para serem entrevistadas e não o façam, o que impossibilitaria a obtenção de uma amostra de acordo com o esquema planejado (isto é, inviabilizaria na prática esta amostragem) e a necessidade de acordar cedo. ▲

Uma AMOSTRAGEM PROBABILÍSTICA é qualquer processo de seleção por sorteio de uma das amostras de um espaço amostral de acordo com uma probabilidade P bem classificada nesse espaço amostral.

EXEMPLO 2: Seja U = $\{$Ana, Beatriz, Carla, Diana e Eva$\}$.

a	(A,B,C)	(B,A,C)	(E,A)	(A,E)	(A,A,E)	(A,A,B,B)
P(a)	0,20	0,10	0,25	0,25	0,15	0,05
n(a)	3	3	2	2	3	4
ν(a)	3	3	2	2	2	2

A probabilidade definida aqui é inteiramente arbitrária e não está baseada em qualquer critério. Ao invés dela, poderíamos ter definido qualquer outra. Pegamos um chapéu (ou uma boina) e colocamos dentro 100 pedacinhos de papel dos quais 20 têm neles escrito (A,B,C), 10 têm (B,A,C), 25 têm (E,A), 25 tem (A,E) 15 têm (A,A,E) e 5 têm (A,A,B,B). Selecionamos ao acaso um pedacinho de papel. ▲

A PROBABILIDADE DE INCLUSÃO de um elemento da população quando fazemos uma amostragem probabilística é simplesmente a probabilidade dele pertencer à amostra selecionada. Assim, IP(u) = $\sum_{u \in a}$ P(a). No Exemplo 2: IP (A) = 1, IP (B) = 0,35, IP (c) = 0,3, IP (D) = 0 e IP(E) = 0,65. Uma AMOSTRAGEM PROBABILÍSTICA

UNIFORME é qualquer amostragem probabilística que atribui a todas as amostras do espaço amostral a mesma probabilidade. Uma AMOSTRAGEM EQUIPROBABILÍSTICA é qualquer amostragem probabilística que atribui a todos os elementos da população a mesma probabilidade de inclusão.

Para uma amostragem probabilística necessitamos: (1) uma população, (2) um espaço amostral, (3) uma probabilidade neste espaço amostral e (4) um mecanismo que permita realizar um sorteio de acordo com essa lei de probabilidade. Para obtermos tal mecanismo, eliminando inconvenientes, tipo encontrarmos um chapéu (ou uma boina) e enrolarmos "ad nauseum" pedacinhos de papel, empregamos costumeiramente uma TABELA DE DÍGITOS ALEATÓRIOS.

Um CENSO pode ser visto como um processo "degenerado" de amostragem, onde o espaço amostral contém uma única amostra: uma listagem completa de todos os elementos da população, que fatalmente é selecionada.

Na vida científica e também na vida pessoal de cada um de nós, a maior parte do conhecimento empírico e das decisões que tomamos está fundamentada em alguns poucos exemplos, isto é, no resultado de alguma amostragem. Vamos supor dois búlgaros que passam temporadas no Brasil, um 10 dias e o outro 10 anos, e cada um deles escreve um livro sobre (falta de?) costumes do povo brasileiro; do ponto-de-vista técnico, a principal diferença entre ambos é que um deles trabalhou com uma amostra menor. Uma dona-de-casa ao ser mal atendida por um funcionário quando vai comprar, pela primeira vez, numa das lojas de determinada rede, pode concluir que o atendimento de todos os funcionários da rede é péssimo e decide nunca mais comprar nesta rede de lojas.

No outro extremo de um censo, também encontramos um processo "degenerado" de amostragem, onde o espaço amostral é o conjunto vazio, ou seja, o espaço amostral não contém amostras e, assim, nenhuma amostra é selecionada. Mesmo assim, o indivíduo se julga conhecedor do assunto e apto a tomar decisões. A respeito disso, há versos de Goethe:

"O que não percebeis, negais que exista;
o que não calculaste, é mentira;
o que vós não pesastes, não tem peso;
metal que não cunheis, dizeis que é falso."

Uma INFERÊNCIA ESTATÍSTICA é um julgamento, uma opinião, uma afirmação sobre o todo - a população - levando em conta os resultados obtidos na examinação de apenas uma parte deste todo: a amostra selecionada. Se ganhamos uma amostra grátis de um produto num supermercado, se provamos uma colher de sopa para sabermos se ela está boa de sal, se espiamos uma revista numa banca de jornais ou livraria com o intuito de decidir se a compramos ou não, se experimentamos roupas num provador de um magazine, se assistimos uns poucos minutos de um programa de TV antes de decidirmos se trocamos ou não de canal, se saímos com uma pessoa algumas vezes para decidirmos se iniciamos um namoro ou não, estamos fazendo uma amostragem. Assim, uma esposa pode esclarecer seu marido quando este exagera na bebida: "Meu amor, você não precisa esvaziar completamente a garrafa para conhecer o sabor do líquido que ela contém."

Qualquer esquema, método, técnica ou processo de amostragem que utilizemos pode nos assegurar que a amostra selecionada representa a população em todos os seus aspectos ou, em particular, representa a população nos aspectos específicos que nos interessam no momento? Infelizmente não. Denominamos ERRO AMOSTRAL a (eventual) diferença entre a amostra selecionada e a população em relação aos aspectos específicos que nos interessam no momento. Uma inferência estatística vem geralmente acompanhada de uma declaração probabilística sobre a quantidade de erro amostral existente no processo de amostragem desde que, evidentemente, a amostragem seja uma amostragem probabilística (vamos frisar que estas "declarações probabilísticas sobre a quantidade de erro amostral existente" se referem ao processo de amostragem, ou seja, a todas as amostras com suas respectivas probabilidades, e não à particular amostra selecionada). As amostragens não probabilísticas como as citadas no parágrafo anterior evidentemente não possibilitam qualquer tipo de declaração probabilística sobre a quantidade de erro amostral (porventura) existente. Assim, a Teoria das Amostragens e a Teoria das Probabilidades são fundamentais para a Teoria das Inferências Estatísticas.

Vamos agora enfatizar algo que é tão óbvio quanto importante: os termos "população" e "amostra" são relativos a uma situação específica. Portanto, os estudantes de uma turma podem ser pensados como uma população da qual selecionaremos uma amostra ou como uma amostra de todos os estudantes da Universidade.

REFORÇOS (Respostas na página 26)

1) Conceitue amostragem.

2) Numa amostragem probabilística, a amostra pode ser selecionada somente com dígitos aleatórios?

3) Todas as amostragens são probabilísticas?

4) Que probabilidade você usaria no espaço amostral do Exemplo 2? Quais seriam então, as probabilidades de inclusão de cada elemento?

5) Qual a probabilidade de inclusão de cada elemento da população do Exemplo 2 se for realizada uma amostragem probabilística uniforme em A?

6) As conclusões obtidas através de amostragem são recentes na história da humanidade?

7) Corrija as frases que estiverem incorretas e repita as que estiverem corretas:

a) A amostra deve ser selecionada da população;

b) Cada elemento da população tem a mesma probabilidade do que os demais de pertencer à amostra selecionada em qualquer amostragem.

c) Uma amostra fornece informações sobre a amostragem.

d) A amostragem do Exemplo 1 é eqüiprobabilística.

8) Que característica você exige para considerar uma amostragem "boa" ou "justa?"

As inferências estatísticas são baseadas em amostras. Estas são selecionadas por meio de amostragens. Os censos são alternativas para as amostragens. A primeira vista, uma coleta de dados realizada em toda a população é sempre preferível a uma realizada apenas numa parte da população. Na prática, entretanto, o oposto é freqüentemente verdadeiro porque:

1 - Um censo é impossível quando a população é infinita.

2 - Os ensaios podem ser destrutivos, isto é, a informação só é obtida com a destruição da unidade amostral como ocorre ao testarmos fósforos, munição, tempo de durabilidade de componentes etc. Nestas situações, um censo daria uma descrição precisa de uma população que não existe mais. Existe a anedota do português que veio para o Brasil e escreveu para seu amigo que o Brasil era muito moderno pois, entre outras coisas, tinha uns palitinhos que bastava riscar numa caixinha para fazer fogo, como ele podia comprovar com a caixinha que acompanhava a carta. A resposta foi que o invento não funcionava em Portugal ou que então estavam com defeito. Aí, em nova carta: "Impossível, testei todos."

3 - A rapidez. Se a informação que necessitamos tem urgência, uma coleta de dados realizada em toda a população pode terminar tarde demais, especialmente quando há muitas unidades ou elas estão muito dispersas. Por exemplo, inspecionarmos um por todos os morangos de um caminhão com 5 toneladas pode fazer com que as frutas fiquem deterioradas. Além disso, se uma população tende a mudar com o tempo, um censo seria uma combinação de várias populações. Por exemplo, um levantamento estadual para descobrir a percentagem dos habitantes que contraíram determinada doença contagiosa pode ser tão longo que, quando os resultados vierem e a ação médica apropriada é tomada, a doença já está tão espalhada que uma ação diferente se faz necessária, de fato; aqueles que fizeram o levantamento podem ter contribuído para a difusão da doença.

4 - O custo pode ser muito elevado quando fazemos um censo, especialmente quando há muitas unidades na população ou o exame de cada unidade é oneroso.

5 - A precisão pode não ocorrer quando um censo é realizado numa população muito grande, pois há muitos dados para serem coletados e, assim, uma equipe de trabalho muito grande é necessária, o que pode acarretar eventuais dificuldades de coordenação e de controle, uma baixa na qualidade do pessoal contratado e um aumento da ocorrência de ERROS NÃO-AMOSTRAIS, tais como erros de lançamento, de computação e de digitação.

6 - Uma amostragem permite uma abrangência de informação maior do que um censo. Um levantamento de dados está sempre sujeito a limites de orçamento e tempo. Assim, uma amostragem permite, com os mesmos recursos e no mesmo tempo, um levantamento mais amplo e mais apurado não só sobre as mesmas variáveis, mas também sobre outras variáveis adicionais que talvez interessem.

7 - A repetição. No caso extremo em que todos os elementos da população são idênticos (em relação às variáveis em estudo), então uma amostra formada por uma única unidade amostral fornecerá uma descrição completa de toda a população. Há muitas situações práticas nas quais os elementos da população são muito similares, o que faz com que a informação obtida através de um censo seja apenas um pouquinho maior do que a obtida através de uma amostragem de bem poucas unidades amostrais.

Apesar de todas as razões que acabamos de listar, há situações reais nas quais um censo é mais vantajoso. Entre elas, vamos destacar as seguintes:

a) A população é tão pequena que o custo dispendido, bem como o tempo para fazermos o levantamento dos dados na população são praticamente os mesmos dispendidos para o fazermos numa amostra. Por exemplo, quando a população é uma câmara com 36 vereadores ou a equipe de professores de um cursinho que tem ao todo 15 professores.

b) A amostra deve ser tão grande para que a precisão que desejamos seja obtida, que o custo e tempo adicionais dispendidos para fazermos o censo, isto é, o levantamento de dados em toda a população, são praticamente desprezíveis, além de serem compensados pela eliminação do erro amostral. Por exemplo, quando o tamanho da população é 415 e uma amostra de tamanho 380 é necessária para obtermos a precisão desejada.

c) Uma precisão total é necessária. Por causa da variabilidade amostral, nunca podemos ter certeza absoluta através de uma amostra. Um censo pode nos dar informações precisas e seguras desde que não ocorram erros na coleta dos dados, na compilação, nos cálculos etc. Por exemplo, uma instituição financeira não seleciona uma amostra de seus caixas para saber quanto dinheiro ela dispõe no final do dia: o fato do levantamento ser realizado em todos os caixas não elimina eventuais erros de contagem ou aritméticos, mas elimina o problema de decidir se algum particular grupo de caixas é representativo de todas elas e de como selecioná-lo. Aqueles que se preocupam com infidelidade conjugal ficariam satisfeitos em saber que ela não ocorre às quartas?

REFORÇOS (Respostas na página 26)

9) Que faz a palavra "eventual" entre parênteses na conceituação de "Erro Amostral?"

10) Que são amostragens não probabilísticas?

11) No espaço amostral do Exemplo 1, o tamanho de cada amostra é igual ao seu tamanho específico?

12) As noções de população e de amostra se referem a propriedades intrínsecas?

13) Um censo é, a rigor, uma amostragem? Quem formula uma opinião "na carona" é, de acordo com os versos de Goethe, um experimentador?

14) Cite quatro vantagens usuais das amostragens sobre censos.

15) Fale sobre uma situação prática na qual você prefere fazer uma amostragem a um censo e vice-versa.

16) Um censo deve ser realizado sempre que possível ou há situações práticas nas quais se pode obter melhores resultados com uma amostragem?

17) Corrija as frases que estiverem incorretas e repita as que estiverem corretas:

a) Numa amostra nunca há elementos repetidos;
b) Exames de sangüe são feitos por amostragens;
c) Uma amostra é qualquer processo que produz uma amostragem.

18) Cite um exemplo prático onde a rapidez impediria a realização de um censo devido à urgência.

19) Uma coleta de dados feita numa amostra bem selecionada, ou mesmo em toda a população, pode fornecer, na prática, conclusões bastante incorretas ?

20) Um candidato deseja avaliar, dois meses antes da eleição, sua campanha política através de um censo. Dependendo do resultado deste levantamento de dados, ele mudará ou intensificará sua estratégia. Comente.

21) Você deseja decidir se vai ou não assistir a um filme e pede detalhes a alguém que já o fez. A narrativa dele pode ser um censo ?

22) O uso de amostras nas pesquisas de sua área proffissional ocorre com que freqüência, na sua opinião? E os censos?

23) a) Num levantamento de dados por amostragem, o erro amostral é o único erro que pode ocorrer?

b) A realização de um censo é uma garantia de precisão absoluta?

c) A exigência de um treinamento altamente especializado para o pessoal de campo, isto é, aqueles que irão coletar os dados é um fator que recomenda um censo ou uma amostragem?

24) Preencha as lacunas com a (s) palavra (s) adequada (s):

Uma amostragem ________________ nos fornece condições de ______________ o erro amostral quando desejamos fazer ______________ que, portanto, se apoia na ______________ e na ______________. Para a seleção da ______________ é conveniente o uso de uma ________________ para assegurarmos que a seleção seja feita de acordo com a ________________.

25) Há um jeito prático de verificarmos se uma amostra representa a população ou esta palavra significa apenas uma esperança que temos?

26) É possível se conhecer a distribuição verdadeira de uma variável através de uma amostragem?

27) Corrija cada frase a seguir, se necessário:

a) A amostragem substitui a população.

b) Os censos são realizados mais rapidamente do que as amostragens.

c) Os censos possibilitam menor custo do que as amostragens.

d) Uma conclusão totalmente correta pode ser obtida regularmente com amostras.

e) A população é representada por uma amostragem.

f) Uma inferência estatística é uma opinião a respeito da amostra.

g) Um censo é possível quando a população é infinita.

h) Devemos usar amostragens em ensaios destrutivos.

i) As amostras produzem conclusões consideradas totalmente corretas.

j) Na prática são usadas somente amostragens probabilísticas.

k) Os atributos estão definidos na amostra selecionada.

l) Todos os elementos da população tem a mesma probabilidade de inclusão nas amostragens probabilísticas.

m) Na nossa vida pessoal raramente tomamos decisões com base em resultados de amostragens.

n) Os censos nos possibilitam inferências estatísticas totalmente corretas.

LIÇÃO 4 - AMOSTRAGENS ALEATÓRIAS SIMPLES 8 reforços

Há várias formas de selecionarmos elementos de uma população. Uma AMOSTRAGEM ALEATÓRIA SIMPLES COM REPOSIÇÃO é uma técnica que executamos assim: selecionamos completamente ao acaso um elemento da população. Este é o primeiro elemento da amostra. A seguir, selecionamos novamente ao acaso um elemento da população, que pode eventualmente ser o mesmo selecionado na primeira vez. Este é o segundo elemento da amostra. E assim sucessivamente. Uma AMOSTRAGEM ALEATÓRIA SIMPLES SEM REPOSIÇÃO é uma técnica que executamos com apenas uma diferença da amostragem aleatória simples com reposição: nenhum elemento da população pode ser selecionado mais do que uma vez, ou seja, em cada etapa do processo de seleção é selecionado ao acaso um elemento da população entre aqueles que ainda não foram selecionados. Uma AMOSTRA ALEATÓRIA SIMPLES (COM REPOSIÇÃO ou SEM REPOSIÇÃO) é qualquer amostra selecionada através de uma amostragem aleatória simples (com reposição ou sem reposição). Vamos usar "a.a.s." como abreviatura quer de amostragem aleatória simples quer de amostra aleatória simples.

Na população infinita dos itens gerados num processo de produção industrial, podemos selecionar os elementos para a amostra simplesmente na ordem em que vão saindo da linha de produção. Se o processo de produção é estável, cada elemento saído da linha de produção tem a mesma probabilidade do que os demais de apresentar qualquer um dos resultados possíveis e esta amostragem é equivalente a uma a.a.s. com reposição, já que do ponto de vista probabilístico é como se estivéssemos observando lançamentos sucessivos de uma mesma moeda de um mesmo dado. Podemos fazer o mesmo raciocínio quando registramos a quantidade de carros que abastecem num posto de gasolina em períodos de 24 horas, a quantidade diária de atendimentos num hospital, os tempos de atendimento de fregueses nas caixas de um supermercado etc.

Na população finita dos empregados de uma firma, podemos elaborar uma listagem de todos os elementos da população. Atribuímos números consecutivos de 1 até N aos elementos de acordo com a listagem e sorteamos ao acaso números de 1 até N, um por um, até obtermos a quantidade de elementos desejada para a amostra, omitindo os números que já foram previamente sorteados caso a a.a.s. seja sem reposição. Os números sorteados identificam os elementos da população selecionados para a amostra. Podemos proceder da mesma maneira quando selecionamos estudantes das escolas de engenharia do Paraná, livros de uma biblioteca, anunciantes de um jornal, lojas de calçados de Novo Hamburgo, pacientes internados em um hospital, motoristas de uma empresa de ônibus etc.

A reposição ou não de um elemento da amostra na população antes da seleção do próximo elemento da amostra faz realmente diferença somente quando a população é finita. Se o tamanho da amostra é pequeno em relação ao tamanho da população, a não-reposição dos elementos já selecionados terá um efeito insignificante nas probabilidades de seleção dos elementos remanescentes e as duas amostragens aleatórias simples, a com reposição e a sem reposição, podem ser consideradas praticamente equivalentes. Por outro lado, se o tamanho da amostra é relativamente grande em relação ao tamanho da população, as probabilidades de seleção diferem substancialmente nos dois casos. Uma regra que tem sido empregada é repor os elementos quando a FRAÇÃO AMOSTRAL $f = n/N$, onde "n" é o tamanho da amostra, exceder a 1/20.

Uma a.a.s., seja com reposição ou sem reposição, atribui a cada elemento da população a mesma probabilidade de inclusão. Assim, a probabilidade de seleção de um dos elementos de uma subpopulação qualquer é proporcional ao seu tamanho. Portanto, subpopulações maiores terão uma probabilidade maior de ter elementos na amostra do que subpopulações menores e as subpopulações do mesmo tamanho terão probabilidades iguais. Alguns autores expressam esta idéia afirmando que uma a.a.s. tem a tendência a selecionar uma AMOSTRA REPRESENTATIVA, ao passo que outros utilizam a expressão AMOSTRA PROPORCIONAL. Para nós estas duas expressões não terão sentido algum, pois não as conceituaremos.

Uma a.a.s. sem reposição geralmente fornece amostras mais informativas do que uma a.a.s. com reposição, mas a utilização desta facilita os cálculos teóricos por causa da independência entre as seleções sucessivas. Assim, a a.a.s. com reposição é a técnica de amostragem mais fácil de ser tratada do ponto-de-vista matemático e a maioria dos métodos de inferência estatística existentes foram elaborados para serem aplicados a dados coletados numa amostra selecionada através desta técnica. Caso isto não tenha ocorrido, podemos verificar se a amostra selecionada pode ser considerada equivalente a uma a.a.s. com reposição através de um TESTE DE ALEATORIEDADE ou procurar métodos de inferência estatística especiais para a amostragem específica que foi utilizada:

OUTRA AMOSTRAGEM ⟶ MÉTODOS ESPECIAIS

⇕

A.A.S. COM REPOSIÇÃO ⟶ MÉTODOS COMUNS

No parágrafo anterior utilizamos a expressão "amostra aleatória simples com reposição" e podemos usá-la, é claro, mas sabendo que quem é aleatório é o processo de seleção, isto é, a amostragem e não a amostra. A palavra "aleatória" é, neste texto, sinônimo de "probabilística": poderíamos usar a expressão "amostragem probabilística simples". A palavra "aleatória" não significa que selecionamos os elementos através de um processo caótico tipo "este-sim-este-não". Tais amostragens obviamente não são amostragens probabilísticas e vamos denominá-las AMOSTRAGENS ACIDENTAIS. Há autores que usam a palavra "acidental" como sinônimo da palavra "aleatória". Nós não.

REFORÇOS (Respostas na página 26)

1) Cite uma vantagem das a.a.s. com reposição sobre as a.a.s. sem reposição e vice-versa.

2) a) Qual é a menor fração amostral possível ?

b) As amostragens acidentais são probabilísticas ?

c) Para que servem os Testes de Aleatoriedade ?

3) a) Selecione ao acaso e sem reposição 5 empregados de uma empresa a partir de um fichário onde eles estão numerados consecutivamente de 1 até 185.

b) Faça o mesmo para um fichário de 1 até 3282.

4) a) Usa uma amostragem proporcional para selecionar uma amostra de tamanho 10 de uma população de tamanho 120. Idem para uma amostragem representativa.

b) A fração amostral pode ser maior do que 1, ou equivalentemente, o tamanho da amostra pode ser maior do que o tamanho da população?

5) a) Diga uma diferença entre uma "amostra aleatória" e uma "amostra acidental".

b) Com que freqüência são realizadas a.a.s. nas pesquisas de sua área profissional?

6) Um lote de peças deve ser aceito ou rejeitado após a inspeção de algumas peças. Os inspetores não são treinados em Estatística e se recusam a recolocar as peças já inspecionadas no lote. Que tipo de a.a.s. eles preferem realizar?

7) Cite um exemplo prático onde é bastante difícil selecionarmos uma a.a.s. e explique porque.

8) Você considera os procedimentos a seguir como (probabilisticamente) equivalentes a a.a.s.?

a) Para estudar a imagem de uma escola na comunidade, cada criança da escola recebe um questionário que deve ser preenchido pelo pai, mãe ou responsável;

b) Para medir a receptividade pública do último pronunciamento presidencial, um repórter de um jornal entrevista 25 pessoas numa esquina do centro de São Paulo;

c) Para predizer a opinião pública a respeito do último empréstimo feito pelo Governo do Estado, 10 deputados são selecionados ao acaso e sem reposição.

LIÇÃO 1

(1) 1/3 e 0,5 (2) 5/18 (3) 0,5 (4) 1/5 e 1/5
(5) Sim, pq. pode ser repetido. Determinístico, pq. Pode ser calculada sem a realização do experimento. (6) Se o experimento for repetido um **grande** número de vezes, o resultado ocorrerá **aproximadamente** 3/17 das vezes.
(7) 0,972 (8) Sim. Por exemplo, sempre que $B \subset A$.
(9) 2/5 e $P\{R = r\} = (r-1)/10, \forall r \in \{2, 3, 4, 5\}$.
(10) $P\{Q = 0\} = 0{,}729$ $P\{Q = 1\} = 0{,}243$ $P\{Q = 2\} = 0{,}027$ $P\{Q = 3\} = 0{,}001$
(11) $P\{A = 0\} = 81/256$ $P\{A = 1\} = 108/256$ $P\{A = 2\} = 54/256$
$P\{A = 3\} = 12/256$ $P\{A = 4\} = 1/256$ $P\{A > 2\} = 13/256 \cong 0{,}0508$
(12) 0,96 0,87 0,03 0,15 Não, pois $P\{A \cap B\} \neq 0$ implica $A \cap B \neq \varnothing$.
(13) 0,43 0,21 83/179 621/800 12/103 101/800 Não, pq. ... 3,92 1,78 4 4.
(14) 0,007976 0,006 0,006 0,996 0,996 Não pq. ...
(15) 3/7 4/9 Não, pq. ... 570/1309
(16) 5/16 $p(0) = p(4) = 1/16$ $p(1) = p(3) = 4/16$ $p(2) = 6/16$ 2 1 2 2
(17) $-1/9(\neq 0)$, o que implica que X e Y não são independentes.
(18) $p(0) = 13/45$ $p(1) = 23/45$ $p(2) = 8/45$ $p(3) = 1/45$ (19) 0,5 (20) 275/343
(21) 5/18 1/3 (22) 2/3 1,25 2 1 (23) 0,25
(24) 4804/22100 24/22100 264/22100 10080/22100 (25) 0,25 0,5
(26) Os experimentos considerados aleatórios
(28) Se o experimento for repetido um **grande** número de vezes, o evento ocorrerá **aproximadamente** 72% das vezes. (29) 1,25 ; 2,86 ; - 1 e 5 (bimodal) ; 1
(30) 21/56 21/56 6/21 6/21 (31) 1/8 1/6 2/5 $0{,}3673(\neq 0)$, logo não são indep.
(32) Não, pq. ... 0 ; 7/18 ; 21/65 ; 51/64 ; 48/75 ; 0,28 ; ? ; ?
(33) –2,15 5,56 -9 -3 (34) 0,81 (35) ... (36) 4/25 1711/128650 0,013824
(37) Subjetiva. ... (38) 2,75 1,30 3 3 (39) 0,4 Não, pq. ... 0,5 Sim, pq. ...
(40) 0,5 0,53 0,25 25/53 25/50 Não, pq. ... Não tem sentido, pq. ...(-0,0955)
(41) $p(0) = 4/35$ $p(1) = 18/35$ $p(2) = 12/35$ $p(3) = 1/35$; 1,29 ; 0,70 ; 1 ; 2/3 ; 8/5 ; 1/3 ; 2/5 ; 0,5 ; 2/3.
(42) $p(0) = p(3)$ 1/20 $p(1) = 9/20$; 1,5 ; 0,67; 1 e 2 ; 2/3 3/5 1/3 0,3 3/5 3/5
(43) $p(1) = 0{,}3$ $p(2) = 0{,}6$ $p(3) = 0{,}1$; 1,8 0,6 2 ; 2/3 2/3 1/3 1/6 0,75 0,5
(44) Não, pq. ... Sim, pq. ... Sim, pq. ... Nenhum. Todos são compostos.

LIÇÃO 3

(5) $\mathbb{P}(A) = 1$ $\mathbb{P}(B) = \mathbb{P}(E) = 1/2$ $\mathbb{P}(C) = 1/3$ $\mathbb{P}(D) = 0$ (6) Não. Por exemplo, a conclusão de que o fogo queima. (7 a) do espaço amostral (7b) em qualquer amostragem equiprobabilística (7c) a população (7d) não é probabilística (12) Não. Se referem a uma propriedade extrínseca: o nosso modo de enfocar a situação. (17 a) pode haver (17b) CORRETA (24) probabilística – inferências estatísticas – Teoria das Probabilidades – Teoria das Amostragens – amostragem probabilística – tabela de dígitos aleatórios – probabilidade especificada (26) Não. Somente através de censo. (27 a) o censo (27b) lentamente (27c) exigem maior (27d) ocasionalmente (27e) uma amostra (27f) da população (27h) CORRETA (27i) Os censos (27k) em subpopulações (27l) nas amostragens equiprobabilísticas (27m) freqüentemente (27n) Os censos nem sequer possibilitam inferências estatísticas.

LIÇÃO 4

(2 a) 1/N (4 a) As duas expressões não têm significado técnico para nós. (4b) Sim, embora seja uma estupidez. Por exemplo nas a.a.s. com reposição. (5 a) A aleatória é probabilística e a acidental não. (6) Sem reposição. (8) Nenhum deles. Aqui ocorre os chamados PROBLEMAS DE COBERTURA: (a) os pais ou responsáveis são apenas parte da comunidade, (b) os passantes naquela esquina são apenas parte da população brasileira, (c) os deputados são apenas parte da população do Estado (embora há quem a considere plenamente representada por eles...).

LIÇÃO 5 - OUTRAS TÉCNICAS DE AMOSTRAGEM $50 = 13 + 37$ reforço

Há outras formas de selecionarmos elementos de uma população além das a.a.s.. Vamos tentar descrever as que são consideradas mais importantes. Se vamos coletar dados nos elementos de uma amostra selecionada através de uma destas técnicas "alternativas", ela precisa ser planejada e executada com muito cuidado. Devemos conhecê-la bem teoricamente, especialmente nos aspectos relativos à análise e à interpretação dos resultados através de seus métodos especiais de inferência estatística.

As amostragens probabilísticas são realizadas de tal forma que conhecemos a probabilidade de cada amostra a ser selecionada, Por causa disso, podemos determinar "a quantidade de variabilidade amostral". Neste sentido a amostragem é "objetiva", ou "mensurável" e uma estimativa do erro amostral pode ser obtida. As a.a.s. são probabilísticas. As amostragens por julgamento e as de voluntários, descritas a seguir, são não-probabilísticas. Nestas, a variabilidade amostral não pode ser calculada e, conseqüentemente, não pode ser estimada. Portanto, devemos usar amostragens probabilísticas sempre que for possível, apesar de existirem situações nas quais as amostragens não-probabilísticas são alternativas toscas, porém úteis, para as amostragens probabilísticas e, em algumas situações especiais, são as únicas amostragens que podem ser realizadas.

AMOSTRAGENS POR JULGAMENTO

Quando a fração amostral é pequena, uma a.a.s. pode produzir uma amostra que não represente bem a população, ao passo que alguém bastante familiarizado com a população pode especificar habitualmente quais são suas unidades mais representativas. Por exemplo, uma cadeia de restaurantes deseja experimentar uma nova técnica de atendimento, digamos o fornecimento de bandejas aquecidas a domicílios e escritórios. O custo de tal experimento recomenda que somente dois dos restaurantes sejam utilizados para o levantamento dos dados e análise. Os restaurantes diferem consideravelmente em termos de tamanho, localização,clientela e lucro. Ao invés de escolhermos ao acaso os dois restaurantes que experimentalmente atenderão também com a nova técnica, pode ser mais sábio confiarmos no conhecimento da administração para a seleção dos dois restaurantes mais típicos da cadeia. Estas amostragens também têm sido denominadas AMOSTRAGENS SUBJETIVAS, INTENCIONAIS ou PROPOSITAIS.

AMOSTRAGENS DE VOLUNTÁRIOS

Há situações em que é muito difícil obtermos unidades de população que estejam dispostas a participarem do experimento. Na área médica, via-de-regra os experimentos devem ser realizados com os pacientes disponíveis ou com os que foram persuadidos a serem voluntários. Nenhum grupo pode ser considerado realmente uma amostra representativa da espécie humana. Portanto, é sempre arriscado tentarmos obter conclusões sobre a humanidade em geral com base em tais estudos, mas alguns resultados podem ser suficientemente reveladores a ponto de exigirem amostragens probabilísticas para validá-los experimentalmente.

AMOSTRAGENS SISTEMÁTICAS

São amostragens executadas assim: dividimos uma listagem da população em grupos consecutivos de mesmo tamanho e, após selecionarmos ao acaso alguns elementos do primeiro grupo, selecionamos também os elementos que ocupam as mesmas posições em todos os demais grupos. Por exemplo, se $N = 180$ e $n = 15$, então $F = 1/f = N/n = 180/15 = 12$, o que significa que um elemento de cada grupo consecutivo de 12 elementos deve ser selecionado. Uma tabela de dígitos aleatórios deve ser usada apenas para a seleção do (s) elemento (s) do primeiro grupo, já que a seleção dos demais é automática. Por exemplo, se selecionamos o número 11, os demais serão $11 + 12 = 23$, $23 + 12 = 35$, ..., $167 + 12 = 179$. Para utilizarmos esta técnica de amostragem devemos estar seguros de que a listagem que utilizamos não apresenta os elementos com alguma periodicidade(em relação às características que pretendemos estudar). Por exemplo, para estudarmos a natalidade em determinada cidade, o mês de novembro ocorre geralmente cerca de nove meses após o período de Carnaval. Uma listagem de nomes em ordem alfabética pode ter grupos, porque nomes de diferentes etnias começam com certas letras ou com certas combinações de letras. Ao selecionarmos sistematicamente casas quando a listagem está baseada na ordem das casas nas ruas, podemos correr o risco de que o número de casas em cada quadra seja aproximadamente igual: as casas de esquina devem ter mais valor, maiores impostos, mais ruído etc., assim como seus moradores devem ter mais renda, mais preocupações com a política municipal, maiores dificuldades de audição etc.

AMOSTRAGENS ESTRATIFICADAS

São amostragens nas quais dividimos a população em subpopulações, portanto subconjuntos bem definidos com pelo menos dois elementos, disjuntas denominadas ESTRATOS. Estas populações são geralmente formadas por elementos similares em algum aspecto, não necessariamente algum que desejamos estudar. Entretanto, quanto mais homogêneos forem os elementos de cada estrato em relação aos aspectos que desejamos estudar, melhor será a qualidade das inferências estatísticas realizadas com base na amostra selecionada. Uma amostra, ou mais precisamente uma SUBAMOSTRA, é selecionada em cada estrato através de amostragens não necessariamente iguais. Por exemplo, podemos selecionar uma a.a.s. sem reposição de tamanho 82 no primeiro estrato, uma amostragem sistemática de tamanho 46 no segundo estrato, um censo no terceiro estrato etc. Quando conseguimos agrupar a população em estratos homogêneos em relação a algum aspecto que desejamos estudar, a variabilidade dentro de cada estrato é bem menor do que a existente em toda a população e para detectá-la com a mesma precisão é suficiente uma amostra de tamanho bem menor. Este raciocínio pode ser aplicado a um caso extremo para que possamos entendê-lo melhor: se todos os elementos de um estrato forem idênticos em relação a algum aspecto que desejamos estudar, então um único elemento do estrato informa com precisão como é todo o estrato. Portanto, quanto mais similares forem os elementos de um estrato, menor será o tamanho da amostra necessária para obtermos uma inferência estatística com uma qualidade determinada.

AMOSTRAGENS POR CONGLOMERADOS

São amostragens nas quais inicialmente toda a população é "coberta" por alguns grupos de elementos não necessariamente disjuntos nem necessariamente homogêneos denominados CONGLOMERADOS. Alguns destes conglomerados são selecionados e todos os seus elementos são incluídos na amostra. Idealmente, cada conglomerado pode ser visto como uma minipopulação. De fato, se cada conglomerado é perfeito, no sentido de representar com precisão toda a população, necessitamos de somente um deles. Infelizmente isto ocorre muito raramente na prática porque os conglomerados são geralmente formados por elementos que possuem uma certa proximidade física e, assim, são freqüentemente homogêneos. Estas amostragens oferecem vantagens de planejamento e de economia. Por exemplo, para estudarmos as preferências políticas de marinheiros seria muito custoso caso eles fossem selecionados através de uma a.a.s. e é bem mais fácil selecionarmos alguns navios e, neles entrevistarmos todos os marinheiros. Geralmente é impossível ou impraticável agruparmos os elementos da população em conglomerados bem heterogêneos e, assim, geralmente necessitamos selecionar uma grande quantidade de conglomerados.

AMOSTRAGENS EM ESTÁGIOS MÚLTIPLOS

São amostragens que executamos em várias etapas como, por exemplo, quando desejamos entrevistar todas as donas-de-casa do Estado a respeito de um produto de limpeza ou toda a população a respeito do hábito de tomar chimarão: no primeiro estágio selecionamos alguns municípios, no segundo estágio alguns distritos destes municípios, no terceiro estágio algumas vias destes distritos, no quarto estágio alguns trechos destas vias e, finalmente, no quinto estágio algumas residências destes trechos.

REFORÇOS (Respostas na página 55)

1) a) Qual é o significado da palavra "alternativas" no primeiro parágrafo desta lição?

b) Uma amostragem por julgamento permite uma atribuição objetiva ao valor do erro amostral?

c) A quantidade de variabilidade amostral é essencial para que possamos fazer inferências estatísticas? Quais são os tipos de amostragens em que podemos avaliá-la?

2) Diga uma diferença entre as amostragens estratificadas e as amostragens por conglomerados.

3) O que são amostragens probabilísticas? Em que situações devemos usá-las?

4) a) Selecione uma amostra sistemática de tamanho 10 de uma população de tamanho 85.

b) Numa amostragem sistemática de tamanho 135 de uma listagem com 7830 unidades amostrais, a 6402ª unidade foi selecionada. Quais são a primeira e a última?

c) Qual é a quarta amostra sistemática de tamanho 10 da população $U= (u_1, \ldots u_{50})$? Qual é a probabilidade dela ser sorteada?

d) Quantas amostras sistemáticas de tamanho 12 existem numa população de tamanho 384? Quais são os elementos da sexta amostra? Qual é a probabilidade de inclusão de cada elemento da população?

5) a) Diga uma diferença entre estratos e conglomerados.

b) Existe uma técnica de amostragem que tem ... vantagem sobre todas as outras?

6) Estabeleça a correspondência entre as colunas:

a - amostragem aleatória simples	( ) poucos elementos selecionados
b - amostragem sistemática	( ) listagem dos elementos
c - amostragem estratificada	( ) listagem aleatória dos elementos
d - amostragem por conglomerados	( ) subpopulações homogêneas
e - amostragem proposital	( ) elementos fisicamente próximos

7) Considera $U = \{ u_1, \ldots, u_{60} \}$. Escreva uma possível amostra por conglomerados de tamanho 6. A amostra $(u_{11}, u_{23}, u_{35}, u_{47}, u_{59})$ parece ser selecionada através de qual técnica de amostragem?

8) Um estabelecimento comercial deseja fazer um levantamento de dados sobre sua imagem num bairro onde há 42 quarteirões, cada um com um número de residências que varia de 8 até 93. Como você selecionaria 10 quarteirões e, a seguir, 3 residências de cada quarteirão? Você preferiria selecionar as 30 residências em uma única etapa?

9) À primeira vista, qual parece ser uma técnica de amostragem apropriada para cada situação a seguir?

a) Os elementos da população estão muito espalhados e deseja-se minimizar despesas com deslocamentos.

b) Deseja-se fazer um levantamento de dados relativos a lazer, tais como, a percentagem da renda empregada, tipos de atividades, quantidade de tempo dispendida semanalmente etc.

c) Deseja-se estudar o grau de escolaridade de marinheiros.

d) Deseja-se estudar a prática de ginástica ou atividades desportivas dos operários de uma grande indústria.

e) Idem para uma pequena indústria.

f) Deseja-se verificar a aceitação de um programa de TV nas diversas classes sociais de uma grande cidade.

g) O experimento será realizado com cadáveres.

h) Deseja-se estudar a variável salário entre os empregados de todas as indústrias de alimentos do Rio Grande do Sul.

i) O experimento será realizado com animais.

j) Deseja-se estudar (um) a (possível) relação entre o aproveitamento escolar de crianças até 12 anos e a carga semanal de TV a que são submetidas em um município com menos de 30 mil habitantes.

k) O experimento será realizado com pacientes terminais.

l) Deseja-se determinar as probabilidades das faces de um dado.

10) Considera $U = \{u_1, \ldots, u_{60}\}$. Escreva uma possível amostra sistemática de tamanho 6. A amostra $(u_{11}, u_{12}, u_{13}, u_{14}, u_{15})$ parece ter sido selecionada através de qual técnica de amostragem?

11) Descreva uma amostragem em 4 estágios de tamanho 2000 dos estudantes de nível superior do Estado.

12) Uma AMOSTRAGEM ESTRATIFICADA PROPORCIONAL é uma amostragem estratificada na qual todos os estratos têm a mesma fração amostral. Numa população com 4 estratos de tamanhos $N_1 = 80$, $N_2 = 175$, $N_3 = 2000$ e $N_4 = 15$, foi realizada uma amostragem estratificada proporcional com $n_2 = 35$. Qual foi o tamanho da amostra?

13) Que tipo de amostragem é o sorteio de algumas turmas quando a população é formada por todos os alunos de um colégio?

As seis técnicas de amostragem descritas nesta lição junto com as três vistas na Lição 4 não são exclusivas. Muito pelo contrário, elas formam aquilo que os matemáticos usualmente denominam "uma base", isto é, toda amostragem pode ser pensada como uma certa combinação ou mistura de algumas delas.

Uma amostragem por julgamento apresenta a vantagem de não necessitar da elaboração de uma listagem na maioria das vezes, o que a torna mais rápida e menos custosa do que as outras técnicas.

Uma amostragem sistemática de uma população finita exige o uso de uma listagem da população e, portanto, tem o mesmo tipo de problemas envolvendo listagens para as demais amostragens. Se os elementos da população estão na listagem sem nenhum critério, as amostragens sistemáticas são equivalentes à a.a.s..

Uma amostragem estratificada tem a grande vantagem de que a amostragem realizada dentro de cada estrato é uma amostragem desta subpopulação. Ora, uma seleção de elementos de toda a população não nos garante uma boa seleção em alguma subpopulação específica, muito menos para todos os estratos. Assim, usando amostragens estratificadas, podemos obter conclusões estatísticas em separado para cada estrato, o que dá a estas amostragens um atrativo especial. Tal interesse pode não existir quando o levantamento de dados é planejado, mas se ele surgir posteriormente, não será necessário um novo planejamento amostral nem a conseqüente coleta de dados. Dentro de cada estrato usualmente realizamos a.a.s. mas isto não é obrigatório: em alguns estratos podemos fazer até um censo. Por exemplo, nos balanços de uma empresa sabemos que 15% dos itens representam cerca de 75% do faturamento; assim, é aconselhável fazermos um censo nos estoques destes poucos itens e selecionarmos amostras aleatórias simples nos estratos formados pelos demais.

Uma amostragem por conglomerados é usualmente utilizada quando não há possibilidade de obtermos uma listagem dos elementos da população (ou algo que seja equivalente a uma listagem), o que acontece geralmente quando estudamos alcóolatras, homossexuais, favelados, prostitutas, toxicômanos etc. As amostragens por conglomerados são especialmente adequadas para levantamento de dados em amostras de populações de acesso difícil como estas já mencionadas, mas podemos utilizá-las também em outras situações. Por exemplo, uma listagem de todas as residências de uma cidade pode ser muito extensa e requerer uma ordem judiciária para ser obtida, ao passo que qualquer mapa atualizado da cidade permite uma listagem pequena e rápida de seus quarteirões.

Uma amostragem em estágios múltiplos pode usar em alguns estágios uma seleção de elementos (unidades) e em outros uma seleção de conglomerados. Por exemplo, quando a população de interesse é formada por todos os proprietários de automóveis de um estado, uma a.a.s. deverá incluir proprietários espalhados amplamente através do estado e, além das despesas, a coordenação da coleta de dados poderá ser um tanto difícil. Conglomerados de distritos ou cidades, por outro lado, contém proprietários de automóveis em áreas concentradas e, assim, a coleta de dados se torna bem menos custosa e a coordenação desta coleta bem mais fácil. Uma seleção aleatória de alguns destes conglomerados é o primeiro estágio. As amostragens dentro de cada conglomerado, isto é, o segundo estágio, poderão ser aleatórias simples, estratificadas ou novamente por conglomerados, desde que o número de proprietários de automóveis dentro de um particular distrito ou cidade seja grande demais para um censo.

Geralmente estudamos diversas variáveis ao mesmo tempo. Pode acontecer que cada uma delas, devido às suas características específicas, possua uma técnica de amostragem mais adequada. Nestas situações, é geralmente melhor realizarmos logo uma a.a.s. ou alguma técnica que possa ser considerada praticamente equivalente a ela.

Nos levantamentos de dados feitos com o grande público, onde muitos entrevistadores precisam ser contratados, há a possibilidade de alguns falsificarem várias, senão todas, as entrevistas. Não é surpresa alguma isto acontecer em países onde os estudantes têm o hábito de "colar" nas provas. Por esta razão, é sempre aconselhável a verificação de uma certa quantidade de entrevistas antes de pagá-los.

REFORÇOS (Respostas na página 55)

14) Cita, sem explicar, uma característica exclusiva de cada uma das seguintes técnicas de amostragem : amostragem de voluntários, a.a.s., amostragem em estágios múltiplos, amostragem por julgamento.

15) Reescreva corretamente as frases incorretas:

   a) Todos os elementos dos conglomerados selecionados fazem parte da amostra.

   b) Estratos são subconjuntos disjuntos da amostra estratificada.

   c) Todos os estratos têm a mesma probabilidade de serem sorteados numa amostragem estratificada.

16) Descreva uma amostragem em três estágios de tamanho 240 dos universitários que se matricularam na UFRGS em cursos de graduação neste semestre letivo.

17) Na população $U = \{u_1, \ldots, u_{112}\}$, qual é a sétima amostra sistemática de tamanho 8? Qual é a probabilidade de que ela seja sorteada? Qual é a probabilidade de inclusão de cada elemento da população nesta amostragem?

18) a) A população é formada por todos os alunos de graduação da UFRGS deste semestre letivo. Os alunos que, num dado momento, estão tendo aula num determinado prédio, formam um estrato ou um conglomerado?

b) Na mesma população de (a), decidiu-se usar uma amostragem estratificada onde os estratos serão as 4 áreas: Tecnológica, Biológica, Humanas e Artes. Qual é a principal vantagem desta técnica? A quantidade de elementos selecionados em cada estrato deve ser a mesma? Qual estrato deve fornecer mais elementos para a amostra, caso você tenha respondido negativamente?

19) A amostra $(u_1, u_2, u_3, u_{35}, u_{36}, u_{37}, u_{38}, u_{39})$ pretende representar a população $U = \{u_1, \ldots, u_{68}\}$. Que tipo de amostragem parece ter sido utilizada?

20) Na população $U = \{u_1, \ldots, u_{120}\}$, qual é a 4ª amostra sistemática de tamanho 5 ? A amostra $(u_{12}, u_{36}, u_{23}, u_{14}, u_6, u_{29}, u_{54}, u_{42}, u_{76}, u_{61}, u_{78}, u_{59}, u_{118}, u_{101}, u_{86}, u_{93}, u_{81}, u_{110})$ parece ter sido obtida através de qual técnica de amostragem ? Idem para $(u_9, u_{10}, u_{11}, u_{12}, u_{89}, u_{90}, u_{91}, u_{92})$. Qual é a 6ª amostra sistemática de tamanho 6 ?

21) Uma população foi dividida em 5 estratos de tamanhos $N_1 = 2000$, $N_2 = 3000$, $N_3 = 7000$, $N_4 = 2800$ e $N_5 = 2200$. Desejamos fazer uma amostragem estratificada proporcional de tamanho 400. Quantos elementos devemos selecionar de cada estatuto? Qual a fração amostral de cada estrato? Se fizéssemos outra amostragem estratificada, selecionando a mesma quantidade de elementos de cada estrato, quais seriam as frações amostrais de cada um deles?

22) Uma população foi dividida em 3 estratos de tamanhos $N_1 = 240$, $N_2 = 360$ e $N_3 = 100$. Através de uma amostragem estratificada você pretende selecionar 35 elementos da população através de a.a.s. com reposição. Se as variâncias dos estratos forem $\sigma_1^2 = 1$, $\sigma_2^2 = 2$ e $\sigma_3^2 = 4$, qual deve ser a sua REPARTIÇÃO AMOSTRAL, isto é, quantos elementos você deve selecionar em cada estrato?

23) Em linhas gerais, ordene estas amostragens da pior para a melhor: estratificada, por julgamento e aleatória simples (não é necessário justificar).

24) Em linhas gerais, ordene estas amostragens da melhor para a pior: estratificada, de voluntários e aleatória simples (não é necessário justificar).

25) Qual destas amostragens em linhas gerais é considerada pior: de voluntários ou aleatória simples? Justifique.

26) Qual destas amostragens em linhas gerais é considerada melhor: sistemática ou aleatória simples? Justifique.

27) a) Uma amostragem por conglomerados é considerada um processo sofisticado de amostragem?

b) O que é uma amostragem de voluntários? Quando deve ser usada?

c) O que é uma amostragem estratificada? Quando deve ser usada?

28) a) Quais as amostragens geralmente não nos possibilitam fixar de antemão o tamanho exato da amostra?

b) As amostragens aleatórias simples são sempre o melhor esquema de seleção de amostras?

c) Qual a técnica de seleção de uma amostra que mais vantagens oferece na seleção?

29) As amostragens probabilísticas são as únicas que podem ser usadas?

30) "Algumas gotas de sangue são recolhidas na ponta de um dos dedos de uma pessoa quando se deseja determinar sua quantidade de glóbulos brancos." Isto é uma amostragem? De que tipo? É probabilística? É representativa?

31) "Para pesquisarmos a opinião dos moradores de uma cidade a respeito da administração de um prefeito, escolhemos 200 moradores de uma região beneficiada por um projeto deste prefeito." Isto é uma amostragem? De que tipo? É probabilística? É representativa?

32) Numa população de tamanho 1242 foi feita uma seleção amostral por conglomerados que resultou em 138 unidades. Qual é a fração amostral? Quantas amostras sistemáticas existem com a mesma fração?

33) Quais são as duas principais vantagens das amostragens estratificadas?

34) Existe uma técnica de amostragem universalmente melhor?

35) Você deseja entrevistar funcionários de um hospital ou operários de uma fábrica. Esperá-los na porta de serviço é uma amostragem? De que tipo? É probabilística? É representativa? Em termos práticos, você considera bom este método para fazer o levantamento de dados?

36) Considere a população $U = \{u_1, \ldots, u_{48}\}$. Qual é o segundo conglomerado de tamanho 4? Qual é a segunda amostra sistemática de tamanho 6? Em cada caso, qual é a fração amostral e qual a probabilidade de inclusão de cada elemento da população?

37) a) O que é uma amostragem equiprobabilística?

b) O que é uma amostragem estratificada proporcional?

c) O que é uma amostragem uniforme?

38) À primeira vista, qual parece ser a técnica de amostragem mais indicada para a população formada por todos os camionistas do Brasil?

39) Na população $U = \{u_1, ..., u_{70}\}$, qual é uma possível amostra sistemática de tamanho 7 ? E uma amostra estratificada de tamanho 10, onde $U_1 = \{u_1, ..., u_{35}\}$ e $U_2 = \{u_{36}, ..., u_{70}\}$?

40) Na população $U_1 = \{u_1,... , u_{80}\}$, qual é uma possível amostra de tamanho 10 formada por 2 conglomerados ? $(u_3, u_{23}, u_{43}, u_{63})$ parece ter sido obtida usando que tipo de amostragem ?

41) Seleciona uma amostra sistemática de tamanho 10 de uma população de tamanho 125.

42) Uma amostra sistemática de tamanho 15 de uma população de tamanho 275 contém as unidades u_{100}, u_{141}, u_{183}. Qual é esta amostra ?

43) $(u_3, u_{13}, u_{23}, u_{33})$ parece ser uma amostragem sistemática de uma população com $N = 40$?

44) $(u_3, u_{13}, u_{23}, u_{33})$ parece ser uma amostra sistemática de uma população com $N = 50$?

45) Uma população é separada em 3 estratos tais que $N_1 = 40$, $N_2 = 200$ e $N_3 = 360$. Supõe-se que há praticamente a mesma variabilidade nos 3 estratos. Deseja-se uma amostra estratificada de tamanho 60. Quantos elementos devem ser selecionados de cada estrato ?

46) Em que situações você prefere usar uma amostragem de voluntários ?

47) Amostragens intencionais são usadas geralmente em populações grandes ?

48) Seleciona uma amostra sistemática de tamanho 12 de uma população com $N = 300$.

49) Há alguma característica comum a todas as amostragens ?

50) Decidiu-se realizar uma amostragem estratificada. A utilização da mesma fração amostral em todos os estratos proporcionará as melhores inferências estatísticas ?

LIÇÃO 6 - LISTAGENS OU EQUIVALENTES

20 reforços

Numa população finita, um processo de seleção verdadeiramente aleatório de uma amostra depende de uma listagem exata das unidades amostrais, sejam elas elementos ou conglomerados da população.

Qualquer processo de seleção através de uma listagem envolve geralmente a atribuição de números consecutivos às unidades amostrais listadas e sorteios verdadeiramente ao acaso de alguns destes números, que identificarão as unidades que irão compor a amostra selecionada. Podemos usar moedas, dados, cartas, bolas ou fichas para realizarmos tais sorteios. Por exemplo, se a população tem tamanho 781, podemos colocar fichas numeradas de 1 até 781 numa caixa e, após misturarmos bem, sorteamos fichas desta caixa de acordo com o esquema probabilístico planejado. Na prática, entretanto, tais estratégias não são absolutamente seguras de reproduzir tal esquema com exatidão: como ter certeza de que as fichas foram mesmo bem misturadas, que as cartas foram mesmo bem embaralhadas e que nenhuma ficou grudada, que uma das bordas do dado não está mais gasta, desfavorecendo duas faces? Além disso, se uma grande quantidade de sorteios é necessária, eles são enfadonhos e consomem muito tempo. É preferível destarte que a seleção das unidades amostrais seja realizada através de uma tabela de dígitos aleatórios.

Uma listagem não é, em geral, um censo, mas apenas um meio de identificarmos as unidades amostrais. A menos que as variáveis que desejamos estudar já estejam incluídas nas listagens, elas deverão ser medidas nos elementos das amostra, se fizermos uma amostragem, ou da população, de fizermos um censo. Por exemplo, uma listagem dos cinemas dos municípios da Grande São Paulo não informa o total de ingressos vendidos em cada cinema no mês passado ou num feriado específico. Uma listagem dos automóveis registrados numa central de trânsito não informa quais os que têm alarma contra roubo, de que tipo e marca. Uma listagem dos estoques de ações numa bolsa de valores não informa o balanço das empresas. Portanto, a função de uma listagem é, em princípio, apenas nos possibilitar a seleção das unidades amostrais nas quais iremos coletar (isto é, medir) os dados.

Se conseguimos uma listagem exata das unidades amostrais, então a seleção da amostra através de uma tabela de dígitos aleatórios é bastante simples. No lugar de uma listagem, podemos usar qualquer outro tipo de MOLDE (FRAME) que sirva para identificar as unidades amostrais. Por exemplo, o fichário dos livros de uma biblioteca. Uma listagem é destarte apenas um tipo de molde. Em várias situações práticas temos ao nosso dispor um molde incompleto ou desatualizado e sua utilização nos faria atribuir probabilidade de inclusão zero aos elementos da população que não estão no molde.

A obtenção de uma listagem ou de qualquer outro tipo de molde pode ser impossível, dispender muito dinheiro ou consumir muito tempo. Por exemplo, se a população é formada pelos moradores de Porto Alegre que possuem Carteira de Habilitação para dirigir algum veículo motorizado, sabem dirigir bicicleta e já visitaram um zoológico, então é mais aconselhável utilizarmos alguma técnica especial de amostragem do que tentarmos elaborar uma listagem.

Nos casos em que as unidades amostrais não podem ser claramente identificadas é inútil a elaboração de qualquer molde. Por exemplo, se desejamos fazer um levantamento de dados a respeito das possibilidades de aproveitamento do lixo ou da poluição de um rio. A alternativa é selecionarmos locais, ao invés de elementos, e pensarmos a população como formada por quadrados de área, cubos de água etc. Tais amostragens são as AMOSTRAGENS POR ÁREA.

REFORÇOS (Respostas na página 55)

1) O que é necessário, na maioria das amostragens, para que possamos selecionar uma amostra rigorosamente de acordo com a técnica escolhida ?
2) Quais são as duas razões principais para selecionarmos unidades amostrais através de mecanismos aleatórios ?
3) A identificação dos alunos nos locais onde sentam em um sala de aula é um molde ? Aleatório?
4) O fichário de uma biblioteca, na prática, nos permite realizar a amostragem planejada ?
5) Quais são os melhores moldes ?
6) "Unidades amostrais" e "elementos da população" significam a mesma coisa ?
7) Cita uma população onde é fácil a elaboração de um molde e diga o porquê.
8) Sempre devemos elaborar uma listagem ?
9) O que é uma Tabela de Digitos Aleatórios ? Para que são usadas na Amostragem ? Elas são indispensáveis ?
10) O que poderia substituir uma listagem dos pacientes de um hospital ?
11) Uma amostragem em estágios múltiplos exige um molde de todas as unidades populacionais ?
12) As amostragens sistemáticas dispensam moldes ?
13) As amostragens por área exigem um molde das unidades populacionais ?
14) Numa amostragem por conglomerados é necessário um molde de toda a população ?
15) Exemplifique dois moldes que não são listagens.
16) Uma listagem completa e atualizada da população equivale a um censo ?
17) Cita uma situação prática na qual a dificuldade de obtenção de um molde prejudica a seleção de uma amostra.
18) As listagens são indispensáveis nas amostragens probabilísticas ?
19) Qual é o papel das listagens na Amostragem ?
20) Cita os dois problemas mais freqüentes dos moldes usados na prática.

Lição 7 – PROBABILIDADE (2ª parte)

23 reforços

7.1 – Alguns modelos discretos unidimensionais

Modelo Binomial $n \in \{1, 2, \ldots\}$ e $0 < p < 1$

É usado para uma seqüência de n experimentos dicotômicos (1-SUCESSO; 0-FRACASSO), independentes e com a mesma probabilidade de sucesso p, ou seja, para n ENSAIOS DE BERNOULLI. A v.a. X é o total de sucessos nos n ensaios.

$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \forall x \in \{0,1,\ldots,n\} \qquad E(X) = np \qquad Var(X) = np(1-p)$$

Modelo Hipergeométrico $N \in \{1, 2, \ldots\}$, $K \in \{1, \ldots, N\text{-}1\}$ e $n \in \{1, \ldots, N\text{-}1\}$

É usado quando são selecionados ao acaso e sem reposição n elementos de um conjunto finito com N elementos, dos quais exatamente K apresentam certa característica, denominada SUCESSO. A v.a. X é o total de sucessos entre os n selecionados.

$$p_X(x) = \binom{K}{x}\binom{N-K}{n-x}\binom{N}{n}^{-1}, \forall x \in \{\text{máx}(0, n-N+K), \ldots, \text{mín}(n,K)\}$$

$$E(X) = n\frac{K}{N} \qquad Var(X) = n\frac{K}{N}\left(1-\frac{K}{N}\right)\frac{N-n}{N-}$$

Modelo Poisson $\lambda > 0$

É utilizado para contagem de acontecimentos ocasionais em porções de mesmo tamanho de um universo. A v.a. X é o total de acontecimentos em cada uma destas porções.

$$p_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}, x \in \{0,1,\ldots\} \qquad E(X) = \lambda \qquad Var(X) = \lambda$$

7.2 – Variáveis Aleatórias Contínuas

A FUNÇÃO DE DISTRIBUIÇÃO (ACUMULADA) de uma v.a. X é definida por

(1) $$F_X(x) = P\{X \le x\} = P\{\omega \in \Omega : X(\omega) \le x\}, \forall x \in \Re^1.$$

Uma v.a. X é uma VARIÁVEL ALEATÓRIA CONTÍNUA se, e somente se, existe uma função real não negativa f_X definida em $\Re^1$ tal que

(2) $$F_X(x) = \int_{-\infty}^{x} f_X(t) \cdot dt, \forall x \in \Re^1.$$

Esta função é uma (FUNÇÃO) DENSIDADE DE PROBABILIDADE da v.a. X . Suas propriedades principais são:

(3) $\int_{-\infty}^{+\infty} f_X(x) \cdot dx = 1$ (4) $P\{a < X < b\} = \int_a^b f_X(x) \cdot dx, \forall a, b \in \Re^1$, tal que $a < b$.

(5) $P\{X = x\} = \int_x^x f_X(t) \cdot dt = 0, \forall x \in \Re^1$ (6) $F'_X(x) = f_X(x)$, se F_X é diferenciável em x.

Se $g : \Re^1 \to \Re^1$ uma função e X uma v.a. contínua, então g(X) é uma v.a. contínua e

(7) $$E(g(x)) = \int_{-\infty}^{+\infty} g(x) \cdot f_X(x) \cdot dx.$$

Os parâmetros das v.a. contínuas são definidos de modo análogo aos das v.a. discretas, com a substituição de p_X por f_X e de $\sum$ por $\int$.

7.3 – Alguns modelos contínuos unidimensionais

Modelo Uniforme ou Retangular $-\infty < \alpha < \beta < +\infty$

$f(x) = (\beta - \alpha)^{-1} \cdot I_{[\alpha\,;\,\beta]}(x)$ $\quad E(X) = (\alpha + \beta)/2 \quad Var(X) = (\beta - \alpha)^2/12$

Modelo Exponencial $\alpha > 0$

$f(x) = \alpha e^{-\alpha x} \cdot I_{[0;+\infty]}(x) ; F(x) = (1 - e^{-\alpha x}) \cdot I_{[0;+\infty]}(x) ; E(X) = 1/\alpha; Var(X) = 1/\alpha^2$

Modelo Gaussiano ou Normal $-\infty < \mu < +\infty$ e $\sigma > 0$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \qquad E(X) = \mu \qquad Var(X) = \sigma^2$$

A NORMAL PADRÃO (ou REDUZIDA), que tem $\mu = 0$ e $\sigma = 1$, é a única tabelada e:

(6) $X \sim N(\mu\,;\,\sigma) \Rightarrow Z = (X - \mu)/\sigma \sim N(0\,;\,1)$

(7) $Z \sim N(0\,;\,1) \Rightarrow X = \mu + \sigma Z \sim N(\mu\,;\,\sigma)$

REFORÇOS (Respostas na página 72)

1) Sabe-se que $X \sim B(5\,;\,0{,}4)$. Calcule $P\{X = 3\}$, $P\{X \geq 4\}$ e $P\{X \leq 3\}$.
2) Num teste tipo certo / errado com 10 questões, qual é a probabilidade de um aluno que esteja respondendo ao acaso acerte 8 ou mais ? E num teste com 5 alternativas ?
3) Qual a probabilidade de ocorrer exatamente 2 caras em 6 lançamentos de uma moeda com probabilidade 0,35 de ocorrer coroa num único lançamento ?
4) Numa moeda cuja probabilidade de ocorrer cara num único lançamento é 2/5 da probabilidade de ocorrer coroa, qual a probabilidade de ocorrer 2 caras ou 3 coroas em 10 lançamentos ?
5) Uma gaveta tem 10 lápis, dos quais 3 não tem ponta. Se 5 lápis são escolhidos ao acaso de uma só vez, qual a probabilidade de que exatamente 2 estejam sem ponta ? Escreve a média, a moda, a mediana, o desvio médio e o desvio padrão do total de lápis sem ponta entre os selecionados.
6) Uma urna contém 4 bolas amarelas e 6 brancas. São extraídas ao acaso e sem reposição 3 bolas, das quais X são amarelas. Calcule $P(X \geq 2)$, E(X), Des (X), Dm (X), Med(X) e Mod (X).
7) Um professor distribui uma lista com 12 problemas. A prova será formada por 4 deles sorteados ao acaso. Um aluno não consegue resolver 3. Se cada problema da prova vale 2,5 pontos, qual é a nota esperada, a nota modal, o desvio padrão e qual a probabilidade dela exceder a 6,0 ?
8) Uma urna tem 3 bolas vermelhas e 5 brancas. São retiradas simultaneamente e ao acaso 3 bolas, das quais X são brancas. Sejam $Y = 3X$ e $Z = X^2$. Escreve, para cada uma destas variáveis, a função de probabilidade, a esperança, a variância, a moda, o desvio médio e a mediana.

9) Em certa empresa, o número médio mensal de acidentes é 6. Qual a probabilidade de ocorrer num mês qualquer somente 1 acidente ? E 5 ou mais ? E menos de 3 ?

10) O número de defeitos dos trechos de um cabo tem distribuição de Poisson. Se há, em média, 10 defeitos a cada 100 metros, qual a probabilidade de que 1 metro tenha pelo menos um defeito ?

11) A média do fluxo de veículos, em determinado local de certa rodovia, é igual a 30 veículos por minuto. Calcule as probabilidades de que: (a) passem 2 ou mais veículos num intervalo de 2 segundos, (b) nenhum veículo passe num intervalo de 30 segundos.

12) Na revisão de um livro de Estatística foi observada uma média de 2 erros por página. Escolhida uma página ao acaso, qual a probabilidade de que ela: (a) tenha somente um erro, (b) não tenha erros, (c) tenha 5 erros ou mais.

13) Em cada caso a seguir, calcule a constante α para que f seja uma densidade de probabilidade, a função de distribuição, a média, a moda, a mediana, o desvio padrão e o desvio médio:

(a) $f(x) = \alpha x^2 I_{[-1;2]}(x)$

(b) $f(x) = \alpha x^{-1/2} I_{(0;1)}(x) + \alpha x I_{[1;4]}(x)$

(c) $f(x) = \alpha x^{-2} I_{[\alpha;+\infty)}(x)$

(d) $f(x) = (\alpha/3) I_{[0;1]}(x) + (2\alpha/3) I_{[5;6]}(x)$

(e) $f(x) = \alpha x^2 (1 - x^3) I_{[-1;1]}(x)$

(f) $f(x) = \alpha (1 - |x|) I_{[-2;2]}(x)$

(g) $f(x) = \alpha x^{-5} I_{[1;+\infty)}(x)$

(h) $f(x) = \alpha (1 - x^2) I_{[-1;1]}(x)$

(i) $f(x) = \alpha e^{-|x|}$

(j) $f(x) = \alpha x^3 I_{[-1;2]}(x)$

14) Seja $X \sim N(-2{,}1 ; 3{,}9)$. Calcule $P\{X > 0{,}7\}$, o 1° e o 3° quintis e $P\{-3{,}4 < X < -0{,}7\}$.

15) Sabe-se que X é uma v.a. com esperança 3 e variância 2,4. Calcule $P\{X \le 2\}$, supondo que X tem distribuição : (a) Binomial, (b) Normal.

16) Calcule Dm (X), Med (X) e Mod (X) para : $U(\alpha ; \beta)$, $E(\alpha)$ e $N(\mu ; \sigma)$.

17) Seja $X \sim B(5 ; 0{,}2)$ e $Y \sim N(500 ; 10)$. Calcule y tal que $P\{Y > y\} = P\{X = 3\}$.

18) O tempo T de duração de um componente tem distribuição Exponencial com média igual a 100 horas. Calcule $P\{T > 120\}$, $P\{T > 120 \mid T > 100\}$, Med (T), Des (T), o 1° e o 4° quintis.

19) O tempo de vida T, medido em horas, de certas bactérias tem distribuição Exponencial com variância igual a 10.000 horas². Calcule $P\{50 < T < 150\}$, $P\{T < 150 \mid T > 100\}$ e Mod (T).

20) Seja $X \sim U(5 ; 15)$. Calcule $P\{X < 8 \mid X < 9\}$, $P\{X > 9 \mid X > 8\}$, Med (X), Mod (X) e Var (X).

21) Seja X uma v.a. com distribuição de Poisson com média igual à variância de Y, que tem distribuição Uniforme sobre [12 ; 18]. Calcule $P\{X > 2\}$, $P\{13 < Y < 17 \mid Y > 14\}$ e Med(Y).

22) O tempo de vida T de certos micróbios tem distribuição Uniforme com mediana igual a 20 horas e amplitude igual a 12 horas. Calcule $P\{T > 25\}$, $P\{T > 20 \mid T > 18\}$ e Des (T).

23) A flutuação das cotações das ações de certa empresa é Normal com média de 1,16 e variância de 0,0081. Você deseja comprá-las na faixa inferior de 20% de seu preço. Qual a importância máxima que você paga por cada ação? Você exige para vendê-las o valor mínimo de 1,30. Qual a probabilidade de que você possa vende-las num instante qualquer ?

Inicialmente devemos definir a população, depois devemos selecionar a amostra e depois devemos medir as variáveis nos elementos selecionados para a amostra. Assim obtemos os dados em seu estado bruto. Para assimilarmos mais claramente as informações que eles contêm, podemos tratá-los, isto é, dar a eles algum "polimento", usando alguma técnica de ESTATÍSTICA DESCRITIVA. Tais técnicas são essencialmente de três tipos: gráficas, tabelares ou numéricas. Um número extraído de uma amostra é denominado (UM VALOR DE) UMA ESTATÍSTICA.

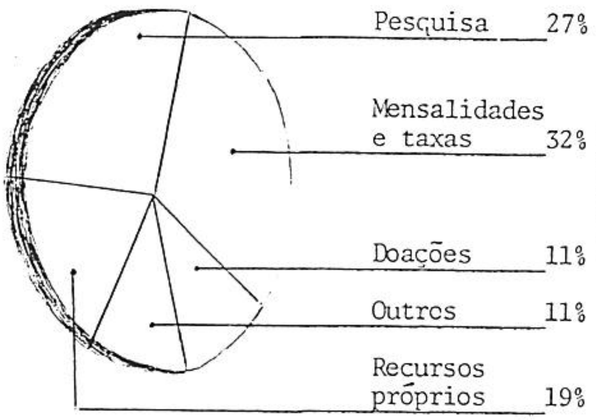
Acreditamos que as representações e leituras gráficas sejam bastante claras para quem estiver lendo este trabalho. Para quem precisar usar gráficos com freqüência e tiver dificuldades para entendê-los, recomendamos ler sobre o assunto na bibliografia existente, na qual destacamos como bastante acessível o livro ESTATÍSTICA FUNDAMENTAL, de Aljocyr Pesca, Ed. Sulina.

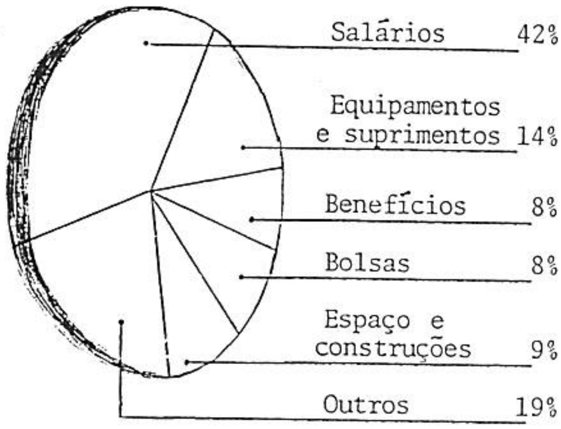
Nos exemplos e reforços a seguir serão abordadas tabelas e métodos numéricos usados para descrever os dados. Antes, entretanto, é necessário que façamos um "lembrete" filosófico: nenhuma descrição é completa, reproduz totalmente o ente descrito, mas apenas fornece uma idéia de alguns dos seus aspectos. Assim, em toda a descrição há algo que se perde: informação, ou seja, conhecimento.

DE ONDE VEM E PARA ONDE VAI O DINHEIRO

(Dados relativos ao ano letivo de setembro de 88 a setembro de 89)

RECEITA US$ 952.4 MILHÕES





Fonte: Departamento de Notícias e Assuntos Públicos da Universidade Harvard.

ESTATÍSTICA DESCRITIVA é a parte da Estatística cujo objetivo é descrever, reunir, sumarizar, apresentar, tabular, agrupar e representar graficamente dados coletados, quer através de uma amostragem, quer através de um censo.

EXEMPLO 1 - Os dados a seguir são os escores de 43 alunos em duas avaliações consecutivas.

1)	5,3	6,7
2)	dez	-
3)	-	-
4)	zero	6,4
5)	-	-
6)	9,5	-
7)	6,5	6,0
8)	-	-
9)	1,5	4,4
10)	5,0	8,0
11)	1,5	6,0
12)	8,8	5,7
13)	5,0	6,2
14)	zero	2,2
15)	zero	4,2
16)	-	-
17)	3,5	4,4
18)	3,5	6,5
19)	3,5	5,0
20)	4,0	7,5
21)	2,0	6,7
22)	7,5	7,5
23)	2,5	7,5
24)	3,5	5,2
25)	dez	8,0
26)	3,5	5,8
27)	0,5	7,0
28)	-	3,5
29)	3,5	4,5
30)	2,0	4,8
31)	6,0	5,7
32)	6,0	6,7
33)	2,5	5,8
34)	9,8	5,0
35)	4,5	6,3
36)	-	-
37)	zero	5,7
38)	5,0	4,5
39)	9,3	9,0
40)	8,0	8,4
41)	zero	7,4
42)	zero	2,0
43)	zero	2,4

x	$f(x)$	$F(x)$
zero	7	7
0,5	1	8
1,5	2	10
2,0	2	12
2,5	2	14
3,5	6	20
4,0	1	21
4,5	1	22
5,0	3	25
5,3	1	26
6,0	2	28
6,5	1	29
7,5	1	30
8,0	1	31
8,8	1	32
9,3	1	33
9,5	1	34
9,8	1	35
dez	2	37

$n_x = 37$

$\sum x = 153,7$

$\sum x^2 = 1022,81$

y	$f(y)$	$F(y)$
2,0	1	1
2,2	1	2
2,4	1	3
3,5	1	4
4,2	1	5
4,4	2	7
4,5	2	9
4,8	1	10
5,0	2	12
5,2	1	13
5,7	3	16
5,8	2	18
6,0	2	20
6,2	1	21
6,3	1	22
6,4	1	23
6,5	1	24
6,7	3	27
7,0	1	28
7,4	1	29
7,5	3	32
8,0	2	34
8,4	1	35
9,0	1	36

$n_y = 36$

$\sum y = 208,6$

$\sum y^2 = 1308,62$

média: $\bar{x} = (\sum_i x) / n_x = 153{,}7 : 37 = 4{,}15$

$\bar{y} = (\sum_i y) / n_y = 208{,}6 : 36 = 5{,}79$

variância: $\text{var } x = \sum_i (x - \bar{x})^2 / n_x = (\sum_i x^2) / n_x - \bar{x}^2$

$= 1022{,}81 : 37 - 4{,}15^2 = 10{,}3873$

$\text{var } y = \sum_i (y - \bar{y})^2 / n_y = (\sum_i y^2)/n_y - \bar{y}^2$

$= 1308{,}62 : 36 - 5{,}79^2 = 2{,}7750$

desvio padrão: $\text{des } x = (\text{var } x)^{1/2} = 10{,}3873^{1/2} = 3{,}22$

$\text{des } y = (\text{var } y)^{1/2} = 2{,}7750^{1/2} = 1{,}67$

desvio médio: $\text{dm } x = \sum_i |x - \bar{x}| / n_x = 99{,}47 : 37 = 2{,}69$

$\text{dm } y = \sum_i |y - \bar{y}| / n_y = 47{,}02 : 36 = 1{,}31$

amplitude: $a(x) = x^*_n - x^*_1 + 1 \text{ u.m.} = 10{,}0 - 0{,}0 + 0{,}1 = 10{,}1$

$a(y) = y^*_n - y^*_1 + 1 \text{ u.m.} = 9{,}0 - 2{,}0 + 0{,}1 = 7{,}1$

moda: $\text{mod } x$ = zero (UNIMODAL)

$\text{mod } y \in \{5{,}7;\ 6{,}7;\ 7{,}5\}$ (TRIMODAL)

freqüência acumulada: $F_x(4{,}9) = 22/37 = 0{,}595$

$F_y(4{,}9) = 10/36 = 0{,}278$

$P_x[3{,}0;\ 6{,}0] = F_x(6{,}0) - F_x(2{,}9) = (28 - 14)/37 = 14/37 = 0{,}378$

$P_y[3{,}0;\ 6{,}0] = F_y(6{,}0) - F_y(2{,}9) = (20 - 3)/36 = 17/36 = 0{,}472$

$P_x(7{,}0;\ \text{dez}] = F_x(\text{dez}) - F_x(7{,}0) = (37 - 29)/37 = 8/37 = 0{,}216$

$P_y(7{,}0;\ \text{dez}] = F_y(\text{dez}) - F_y(7{,}0) = (36 - 28)/36 = 8/36 = 0{,}222$

coeficiente de correlação linear de Pearson entre x e y :

$n_{x,y} = 35$

$\sum_i xy = 872{,}32$

$\sum_i x = 153{,}7 - \text{dez} - 9{,}5 = 134{,}2$

$\sum_i x^2 = 1022{,}81 - \text{dez}^2 - 9{,}5^2 = 832{,}56$

$\sum_i y = 208{,}6 - 3{,}5 = 205{,}1$

$\sum_i y^2 = 1308{,}62 - 3{,}5^2 = 1296{,}37$

$$\text{cor}(x,y) = \frac{n \sum_i xy - (\sum_i x)(\sum_i y)}{[n(\sum_i x^2) - (\sum_i x)^2]^{1/2}\ [n(\sum_i y^2) - (\sum_i y)^2]^{1/2}}$$

$$= \frac{35 \times 872{,}32 - 134{,}2 \times 205{,}1}{[35 \times 832{,}56 - 134{,}2^2]^{1/2}\ [35 \times 1296{,}37 - 205{,}1^2]^{1/2}} = 0{,}4956$$

▲

EXEMPLO 2 - "Num estudo dos sedimentos praiais do trecho Laguna - Imbituba foram colhidas amostras de areias praiais e de areias eólicas (areias de dunas). Constatou-se que ambas eram finas (0,125 - 0,250mm) e que, nesta área, as medidas de dispersão não podem ser usadas como características de distinção entre os dois tipos, ao contrário do que ocorre na costa riograndense. Constatou-se também que as praiais tem assimetria negativa e as eólicas assimetria positiva no Rio Grande do Sul e em Santa Catarina, o que faz com que esta medida seja empregada na classificação destes dois ambientes. A curtose não serve por causa da sua grande instabilidade." Vamos tentar explicar alguns destes termos com números totalmente artificiais.

ROL BRUTO DE DADOS:

5,6	4,2	4,6	5,5	5,7	2,3	6,2	5,1	2,9
6,1	8,7	7,6	2,0	4,4	5,3	6,5	4,1	4,5
3,4	1,0	6,5	3,0	5,4	6,4	7,1	4,7	4,6
3,2	5,4	3,7	3,7	4,4	5,2	6,0	5,9	5,1

$n = 36 \qquad \sum x = 176{,}0 \qquad \sum x^2 = 950{,}06$

$\sum x^3 = 5499{,}302 \qquad \sum x^4 = 33699{,}0578$

DADOS AGRUPADOS EM CLASSES:

i	Ci	fi	Fi	$\bar{x}i$
1	1,0 ⊢ 2,0	1	1	1,45
2	2,0 ⊢ 3,0	3	4	2,45
3	3,0 ⊢ 4,0	5	9	3,45
4	4,0 ⊢ 5,0	8	17	4,45
5	5,0 ⊢ 6,0	10	27	5,45
6	6,0 ⊢ 7,0	6	33	6,45
7	7,0 ⊢ 8,0	2	35	7,45
8	8,0 ⊢ 9,0	1	36	8,45

$n = 36 \qquad \sum \bar{x}_i f_i = 178{,}2 \qquad \sum \bar{x}_i^2 f_i = 967{,}09$

$\sum \bar{x}_i^3 f_i = 5616{,}5955 \qquad \sum \bar{x}_i^4 f_i = 34424{,}33523$

ROL ORDENADO DE DADOS:

1,0	2,0	2,3	2,9	3,0	3,2	3,4	3,7	3,7
4,1	4,2	4,4	4,4	4,5	4,6	4,6	4,7	5,1
5,1	5,2	5,3	5,4	5,4	5,5	5,6	5,7	5,9
6,0	6,1	6,2	6,4	6,5	6,5	7,1	7,6	8,7

média: $\mu = \dfrac{\sum x}{n}$

desvio padrão: $\sigma = \left[\frac{\sum (x-\mu)^2}{n}\right]^{1/2} = \left[\frac{\sum x^2}{n} - \mu^2\right]^{1/2}$

coef. de assimetria: $\alpha = \frac{\sum [(x-\mu)/\sigma]^3}{n} = \left[\frac{\sum x^3}{n} - 3\mu \frac{\sum x^2}{n} + 2\mu^3\right] \times \sigma^{-3}$

coef. de curtose: $K = \frac{\sum [(x-\mu)/\sigma]^4}{n} = \left[\frac{\sum x^4}{n} - 4\mu \frac{\sum x^3}{n} + 6\mu^2 \frac{\sum x^2}{n} - 3\mu^4\right] \times \sigma^{-4}$

ROL BRUTO DE DADOS:

$\mu = 176{,}0 : 36 = 4{,}89$

$\sigma = \left[950{,}06 : 36 - 4{,}89^2\right]^{1/2} = 1{,}58$

$\alpha = \left[\frac{5499{,}302}{36} - 3 \times 4{,}89 \times \frac{950{,}06}{36} + 2 \times 4{,}89^3\right] \times 1{,}58^{-3} = -0{,}15$

$K = \left[\frac{33699{,}0578}{36} - 4 \times 4{,}89 \times \frac{5499{,}302}{36} + 6 \times 4{,}89^2 \times \frac{950{,}06}{36} - 3 \times 4{,}89^4\right] \times 1{,}58^{-4} = 3{,}16$

DADOS AGRUPADOS EM CLASSES:

$u = 178{,}2 : 36 = 4{,}95$

$\sigma = \left[967{,}09 : 36 - 4{,}95^2\right]^{1/2} = 1{,}54$

$\alpha = \left[\frac{5616{,}5955}{36} - 3 \times 4{,}95 \times \frac{967{,}09}{36} + 2 \times 4{,}95^3\right] \times 1{,}54^{-3} = -0{,}09$

$K = \left[\frac{34424{,}33523}{36} - 4 \times 4{,}95 \times \frac{5616{,}5955}{36} + 6 \times 4{,}95^2 \times \frac{967{,}09}{36} - 3 \times 4{,}95^4\right] \times 1{,}54^{-4} = 2{,}75$

Estas quatro medidas assim definidas o foram através de MOMENTOS, isto é, de médias aritméticas de certas potências de dados. Além de exigirem um certo trabalho para serem calculadas, os (eventuais) erros de mensuração ficam elevados às mesmas potências e aumentam as oportunidades para a ocorrência de erros de digitação. Por estas razões, vários pesquisadores preferem usar medidas definidas através das SEPARATRIZES (ou QUANTIS), que por sua vez são definidos através das ESTATÍSTICAS DE ORDEM. Por exemplo, na Geologia os pesquisadores têm preferido usar as definições de Folk & Ward:

média: $\mu = \frac{x_{0,16} + x_{0,50} + x_{0,84}}{3}$

desvio padrão: $\sigma = \frac{x_{0,84} - x_{0,16}}{4} + \frac{x_{0,95} - x_{0,05}}{6{,}6}$

coef. de assimetria: $\alpha = \frac{x_{0,84} + x_{0,16} - 2x_{0,50}}{2(x_{0,84} - x_{0,16})} + \frac{x_{0,95} + x_{0,05} - 2x_{0,50}}{2(x_{0,95} - x_{0,05})}$

coef. de curtose: $K = \frac{x_{0,95} - x_{0,05}}{2{,}44(x_{0,75} - x_{0,25})}$

ROL BRUTO DE DADOS:

$$x_p = x^*_{int(k)} + (k - int(k))\left(x^*_{1+int(k)} - x^*_{int(k)}\right), \text{ onde } k = (n+1)p.$$

$x_{0,05}$: k = 37x0,05 = 1,85 ∴ $x_{0,05}$ = 1,0 + 0,85 (2,0 - 1,0) = 1,85

$x_{0,16}$: k = 37x0,16 = 5,92 ∴ $x_{0,16}$ = 3,0 + 0,92 (3,2 - 3,0) = 3,18

$x_{0,25}$: k = 37x0,25 = 9,25 ∴ $x_{0,25}$ = 3,7 + 0,25 (4,1 - 3,7) = 3,80

$x_{0,50}$: k = 37x0,50 = 18,5 ∴ $x_{0,50}$ = 5,1 + 0,5 (5,1 - 5,1) = 5,10

$x_{0,75}$: k = 37x0,75 = 27,75 ∴ $x_{0,75}$ = 5,9 + 0,75 (6,0 - 5,9) = 5,98

$x_{0,84}$: k = 37x0,84 = 31,08 ∴ $x_{0,84}$ = 6,4 + 0,08 (6,5 - 6,4) = 6,41

$x_{0,95}$: k = 37x0,95 = 35,15 ∴ $x_{0,95}$ = 7,6 + 0,15 (8,7 - 7,6) = 7,77

$$\mu = (3,18 + 5,10 + 6,41)/\ 3 = 4,90$$

$$\sigma = \frac{6,41 - 3,18}{4} + \frac{7,77 - 1,85}{6,6} = 1,70$$

$$\alpha = \frac{6,41 + 3,18 - 2 \times 5,10}{2 \times (6,41 - 3,18)} + \frac{7,77 + 1,85 - 2 \times 5,10}{2 \times (7,77 - 1,85)} = -0,14$$

$$K = \frac{7,77 - 1,85}{2,44\ (5,98 - 3,80)} = 1,11$$

DADOS AGRUPADOS EM CLASSES:

$$x_p = l_i + \frac{k - F_{i-1}}{f_i}\left(L_i - l_i\right), \text{ onde } k = np,$$ l_i e L_i são os limites inferior e superior.

$x_{0,05}$: k = 36 x 0,05 = 1,80 ∴ $x_{0,05}$ = 1,95 + 0,80 (2,95 - 1,95)/ 3 = 2,22

$x_{0,16}$: k = 36 x 0,16 = 5,76 ∴ $x_{0,16}$ = 2,95 + 1,76 (3,95 - 2,95)/ 5 = 3,30

$x_{0,25}$: k = 36 x 0,25 = 9,00 ∴ $x_{0,25}$ = 3,95

$x_{0,50}$: k = 36 x 0,50 = 18,00 ∴ $x_{0,50}$ = 4,95 + 1,00 (5,95 - 4,95)/10 = 5,05

$x_{0,75}$: k = 36 x 0,75 = 27,00 ∴ $x_{0,75}$ = 5,95

$x_{0,84}$: k = 36 x 0,84 = 30,24 ∴ $x_{0,84}$ = 5,95 + 3,24 (6,95 - 5,95)/ 6 = 6,49

$x_{0,95}$: k = 36 x 0,95 = 34,20 ∴ $x_{0,95}$ = 6,95 + 1,20 (7,95 - 6,95)/ 2 = 7,55

$$\mu = (3,30 + 5,05 + 6,49)\ /\ 3 = 4,95$$

$$\sigma = \frac{6,49 - 3,30}{4} + \frac{7,55 - 2,22}{6,6} = 1,61$$

$$\alpha = \frac{6,49 + 3,30 - 2 \times 5,05}{2 \times (6,49 - 3,30)} + \frac{7,55 + 2,22 - 2 \times 5,05}{2 \times (7,55 - 2,22)} = -0,08$$

$$K = \frac{7,55 - 2,22}{2,44 \times (5,95 - 3,95)} = 1,09$$

▲

Cada Xp separa os dados em dois grupos: o primeiro contendo a fração p dos valores mais inferiores e o segundo contendo a fração 1 - p dos mais superiores. Assim, por exemplo, os QUARTIS são as 3 separatrizes que separam os dados em quatro grupos iguais ($X_{0,25}$, $X_{0,5}$ e $X_{0,75}$) e os PERCENTIS são as 99 separatrizes que os separam em 100 grupos iguais ($X_{0,01}$, $X_{0,02}$,... e $X_{0,99}$). Há situações nas quais é preferível desconsiderarmos as porções mais extremas dos dados e assim surgem, por exemplo, a AMPLITUDE INTERQUALÍTICA $a_{0,5}$ (X) = $X_{0,75} - X_{0,25}$ e a AMPLITUDE INTERPERCENTÍLICA $a_{0,98}$ (X) = $X_{0,99} - X_{0,01}$.

REFORÇOS (Respostas na página 72)

1) Na tabela ao lado temos a distribuição de freqüências do número de empregados de 61 escritórios imobiliários. Calcule a média, o desvio padrão, a MEDIANA (isto é, a separatriz que divide os dados em dois grupos iguais, ou melhor, contendo frações iguais), a freqüência relativa acumulada no ponto 10, o 8º decil e a amplitude.

i	Ci	fi
1	0 ⊢ 3	6
2	3 ⊢ 6	13
3	6 ⊢ 9	21
4	9 ⊢ 12	18
5	12 ⊢ 15	3

2) A rapidez perceptiva de 11 pacientes sob o efeito de determinado sedativo foi 2,6 3,1 1,8 3,0 1,9 2,7 2,3 2,5 2,4 2,2 2,6. Calcule a média, a moda, o desvio médio, o desvio padrão, a mediana, a amplitude interquartílica, a 3ª estatística de ordem e o posto da 3ª observação. Calcule a mediana e a amplitude interquartílica dos escores padronizados.

3) Em 12 protótipos de um minicarro em fase de lançamento, o consumo médio em Km/l de cada um está listado a seguir: 21,2 24,4 23,7 25,1 24,8 22,3 24,3 25,2 22,7 23,4 21,8 24,9. Calcule a média, o desvio médio, a mediana, a amplitude interquintílica, a menor e a maior estatística de ordem.

4) Você herda um apartamento em Punta e pergunta a cinco corretores seu preço. Qual destes 5 preços deve ser considerado o preço justo ? Como você mediria a estabilidade do mercado ? Que tipos de medidas você usou ? Calcule-as para : 23 41 27 19 e 24 (em milhares de dólares). Estes valores acima estão ordenados de acordo com as idades dos corretores, da seguinte forma : o mais idoso avaliou em 23 mil dólares. Calcule um coeficiente de correlação e interprete-o. Que tipo de medida você usou agora ?

5) O que é correlação? O que é um coeficiente de correlação?

6) Um hospital realizou um levantamento dos honorários cobrados por médicos para certa cirurgia: 202 234 226 276 161 210 263 225 180 191 232 237 274 324 301 322 306 289 293 249 278. Calcule a mediana e a amplitude interquartílica. Calcule a 5ª estatística de ordem e o posto de x_{16}.

7) Considere os dados do Reforço 6 como sendo ordenados segundo os números de registro dos médicos no CRM. Calcule um coeficiente de correlação e interprete-o.

8) Os dados listados a seguir são as quantidades médias aproximadas do número de cigarros consumidos por médicos e enfermeiros fumantes durante seus períodos de plantão em determinado hospital. A duração de cada plantão é de 8 horas e os médicos e enfermeiros que participaram do experimento foram sorteados ao acaso. (a) Tem sentido calcularmos um coeficiente de correlação? b) Compare os dois grupos através de suas médias. (c) Compare-os através da soma de seus postos quando os grupos são misturados.

x (enfermeiros):	12	6	14	20	12	5	14	7	9	11
y (médicos):	15	14	18	21	4	32	6	8	12	15

9) Uma disciplina do Departamento de Matemática é destinada a estudantes de Psicologia e de Ciências Sociais. Entre os matriculados, seis deles apresentaram os resultados:

x (nota da 1ª avaliação)	6,5	3,5	-	dez	zero	3,0
y (nota da 2ª avaliação)	2,5	6,0	5,0	9,0	4,0	-

Calcule a correlação entre a notas da primeira e da segunda avaliação. Você calcularia a correlação dos resultados finais entre os dois cursos?

10) Os dados a seguir foram obtidos através de uma seleção ao acaso de todos os casos de AIDS registrados no Rio Grande do Sul. Trata-se da idade dos pacientes no momento em que a doença foi diagnosticada: 17 45 38 27 6 48 11 57 34 22. Determine a mediana, o 3º quintil, o desvio padrão, o desvio médio e o menor escore padronizado.

11) Durante 50 dias escolhidos ao acaso do ano passado, determinado hospital teve os óbitos a seguir:

14	10	12	9	9	12	12	7	6	4
11	12	10	10	9	10	8	10	8	14
6	9	5	13	6	9	8	10	15	15
11	7	10	11	8	7	7	10	13	17
10	7	12	11	5	12	10	7	7	9

Calcule a média, o desvio padrão, a moda, a mediana, a amplitude, o 3º quartil, o 4º quintil, a variância, o total e a 44ª estatística de ordem.

12) Na tabela ao lado temos a distribuição de frequências dos tempos de hospitalização (em horas) de 400 pacientes de determinada clínica de repouso. Determine:

a) a freqüência relativa da 5ª classe;
b) o limite superior real da 6ª classe;
c) a freqüência absoluta acumulada da 7ª classe;
d) o ponto médio da 8ª classe. Calcule também a mediana e faça o histograma.

i	Ci	fi
1	300 - 399	14
2	400 - 499	46
3	500 - 599	58
4	600 - 699	76
5	700 - 799	68
6	800 - 899	62
7	900 - 999	48
8	1000 - 1099	22
9	1100 - 1199	6

13) Cite duas medidas de posição, duas de dispersão e duas de associação.

14) Os valores a seguir são as alturas de 6 crianças de 11 anos: 1,28 1,65 1,35 1,13 1,68 1,42.

a) obtenha as estatísticas de ordem e os postos;
b) obtenha a média e o desvio padrão;
c) obtenha os valores padronizados;
d) obtenha a mediana e o DESVIO MEDIANO, isto é, a mediana das diferenças em valor absoluto entre cada valor e a mediana.

15) Foi feito um levantamento do QI de 7 homens e de 7 mulheres que trabalham numa certa firma:

x (homens):	93	48	124	101	68	93	69
y (mulheres):	101	48	101	102	41	67	85

Com base nas médias qual o sexo deve ser considerado com maior QI? E com base nas medianas? E com base na soma dos postos?

16) Os dados coletados a seguir são as idades das 9 primeiras pessoas com idade inferior a 21 anos que entraram para assistir uma sessão vespertina do circo: 3 7 7 11 15 15 19 11 5. Calcule a média e o desvio padrão. Calcule também a mediana e o desvio mediano.

17) Oito jovens fumantes e oito jovens não fumantes foram selecionados ao acaso numa escola e realizaram testes físicos que expressaram os resultados:

x (fumantes):	93	29	65	71	47	71	47	69
y (não fumantes):	91	34	60	79	47	71	71	84

Com base nestes dados, você afirmaria que o fumo já precocemente atua no estado atlético das pessoas? Responda comparando médias, medianas e somas dos postos.

18) Os dados a seguir são índices anuais de poluição sucessivos nos últimos 12 anos: 10 12 17 15 13 20 17 19 21 23 23 8. Calcule o Coeficiente de Correlação de Spearman. Que conclusão você pode obter deste número?

19) a) A moda é uma medida de dispersão ?
b) As médias são medidas de posição ?
c) As amplitudes medem associação ?
d) O que é um parâmetro ?
e) Define coeficiente de variação.
f) Os coeficientes de correlação de Pearson e de Spearman são equivalentes ?
g) Escreve duas medidas de centralização.
h) Para que usamos tabelas estatísticas ?

20) Um curso especializado em preparar profissionais de alto nível para ingresso em um concurso público preparou uma turma de 20, que obtiveram os resultados expressos pela distribuição de frequências ao lado. Calcule o desvio padrão ; o desvio médio e o rol ordenado.

i	Ci	fi
1	35 ⊢ 45	1
2	45 ⊢ 55	3
3	55 ⊢ 65	3
4	65 ⊢ 75	8
5	75 ⊢ 85	4
6	85 ⊢ 95	1

21) A média dos escores de uma prova foi 6,72 e o desvio padrão foi 0,91. O escore padronizado de um aluno foi 1,35. Qual foi o seu escore bruto?

22) a) As modas são momentos ou separatrizes ?
b) O 65º percentil de 6,0 7,3 2,1 5,3 9,3 7,5 é 7,0 ?
c) As medianas são medidas de posição ?
d) Para que usamos gráficos estatísticos ?
e) Escreve uma medida de associação.
f) A moda é uma medida de assimetria ?
g) As freqüências acumuladas são momentos ?
h) Define variância relativa.
i) O que é um coeficiente de assimetria ?
j) O que é uma estatística ?
l) Escreve uma relação entre os quartis de um rol de dados simétricos.
m) Todo quantil é também uma separatriz ?
n) Existem outras propriedades descritas pelas medidas estatísticas além de posição, dispersão e associação ?

23) Em um momento inesquecível, as idades de um grupo de pessoas estão expressas na distribuição de frequências ao lado. Calcule a média, a mediana, o desvio padrão e a amplitude interquintílica da variável idade naquele momento e 5 anos após.

i	Ci	fi
1	10 ⊢— 14	32
2	14 ⊢— 18	65
3	18 ⊢— 22	78
4	22 ⊢— 26	25

24) A variabilidade relativa de uma variável é expressa geralmente pelo seu COEFICIENTE DE VARIAÇÃO, isto é, pelo quociente entre o desvio padrão e a média. Numa fuga do convencional, podemos medí-la pelo quociente entre o desvio mediano e a mediana que denominaremos, digamos, COEFICIENTE DE INCONSTÂNCIA. Calcule-os para X: 62 62 87 65 58 92 88 79 e para y: 14,3 20,8 22,7 30,5 41,2 28,2 24,2 30,0.

25) Considere a distribuição de frequências a seguir:

y \ X	10 ⊢— 15	15 ⊢— 20	20 ⊢— 30	30 ⊢— 45	45 ⊢— 60
68 ⊢— 72	2	6	8	7	4
72 ⊢— 76	3	8	5	6	1
76 ⊢— 80	1	12	5	2	2
80 ⊢— 84	/	4	3	4	3

Calcule o Coeficiente de Correlação de Pearson e o desvio médio de X.

26) Na tabela ao lado estão os escores de 33 funcionários de uma organização que expressam a disposição para o trabalho segundo um escore de 0 a 100 com a precisão de décimos. Calcule a média, o desvio padrão, o valor da frequência relativa acumulada do escore 60 e os quartis.

i	Ci	fi
1	25,65 - 38,05	6
2	38,05 - 50,45	6
3	50,45 - 62,85	2
4	62,85 - 75,25	6
5	75,25 - 87,65	4
6	87,65 - 100,05	9

27) Foi realizado um levantamento do número aproximado de horas semanais de TV assistidas por crianças de 10 anos de idade em determinada cidade. Calcule a média, o desvio padrão, a mediana e a amplitude interdecílica.

i	Ci	fi
1	12 - 14	15
2	14 - 16	21
3	16 - 18	56
4	18 - 20	78
5	20 - 22	30

28) Um vestibulando que ainda não decidiu qual curso irá fazer, entrevista 10 geólogos empregados, a respeito de salários: 22,0 235,2 45,1 37,7 156,1 25,7 132,9 11,1 87,6 52,5. Calcule a média, a mediana, o desvio padrão e a amplitude interquartílica. Calcule o posto de x_8.

29) a) Sabe-se que $\sigma_X^2 = 36$, $\mu_X = 2$, $\sigma_Y^2 = 64$, $\mu_Y = 3$, $\sigma_Z^2 = 81$ e $\mu_Z = 9$. Qual é a variável com maior variância relativa e a com menor coeficiente de variação ? Quais são estes valores ?

b) Os pesos e as alturas de 4 irmãos são (75;1,65), (80;1,70), (85;1,65) e (80;1,60). Qual das duas variáveis é mais dispersa ?

c) Que são medidas de dispersão ?

d) Escreve uma relação entre os quintis de um rol de dados simétrico.

30) O que é um histograma? Para que serve?

31) Para $k \in \{1,\ldots, 99\}$, $x_{k/100}$ é o k - ésimo percentil da variável X. Sabendo que $X_{0,75} - X_{0,5} = 16,31$ e que $X_{0,5} - X_{0,25} = 15,27$, quanto vale a amplitude semi-interquartílica? Como você define a amplitude semi-interdecílica? E a amplitude semitotal? Como estes números se relacionam?

32) Calcule as estatísticas do Reforço 31 a partir dos dados do Reforço 28.

33) Considere o coeficiente de assimetria definido por $\alpha = \left[(x_n^* - x_{0,5})^3 - (x_{0,5} - x_1^*)^3\right]/2$.

Calcule-o para os dados: 2,1 3,2 9,1 9,3 4,9 9,5 7,0.

34) Para os dados agrupados em classe ao lado, calcule o desvio padrão e o desvio médio. Faça o histograma e o polígono de freqüências.

i	Ci	Fi
1	68 ⊢ 72	8
2	72 ⊢ 76	20
3	76 ⊢ 80	35
4	80 ⊢ 84	40

35) Calcule a variância relativa e o coeficiente de variação para a distribuição de freqüências:

x_i	0	1	2	3	4	5
F_i	10	28	58	68	76	80

36) Um grupo de amigos tem, hoje, as idades a seguir: 20 16 22 18 22 24 28 30 32 34 26 28. Calcule a moda, a média, a mediana, o desvio padrão, o desvio médio, a amplitude e a amplitude interdecílica destas idades hoje e daqui a dois anos.

37) Considere os seguintes dados emparelhados:

x	9,5	0,5	3,5	7,5	dez	/	2,5	dez	/	3,5	3,5	/
y	/	7,0	6,5	7,5	8,0	3,5	5,8	/	/	5,0	6,5	/

Calcule as médias, as medianas, os primeiros e os terceiros quartis, as freqüências acumuladas relativas até o valor 4,9, os desvios padrões, os desvios médios, as amplitudes, as amplitudes interquartílicas, as modas e as freqüências relativas entre 3 e 7 para cada uma das variáveis. Calcule também os coeficientes de correlação de Pearson e de Spearman.

38) Se M é o maior valor numérico de um rol de dados e m é o menor, então M-m é uma medida estatística? Em caso afirmativo, qual o seu nome?

39) Considere os três grupos de dados a seguir. Qual deles é o mais homogêneo (compare as variâncias) e qual deles é relativamente o mais homogêneo (compare as variâncias relativas)?

$m = 200$	$n = 50$	$\bar{z} = 8$
$\sum x_i f_i = 4000$	$\sum y_i g_i = 500$	$\sum z_i h_i = 3200$
var $X = 25$	$\sum y_i^2 g_i = 5450$	$\sum z_i^2 h_i = 32000$

40) Um rol de 10 pares de dados apresentou $\sum x_i = 18,3$ $\sum y_i = 31,5$ $\sum x_i^2 = 97,59$ $\sum y_i^2 = 205,21$ $\sum x_i y_i = 121,11$. Calcule o coeficiente de correlação de Pearson e, a partir dele, o de Spearman.

41) Considere os dados emparelhados a seguir:

x	7,0	2,0	4,0	4,5	5,5	6,0	/	3,5
y	6,6	/	5,5	5,0	9,0	6,5	8,0	2,3

Calcule a média e o desvio padrão de X. Calcule a mediana e a amplitude interquartílica de y. Calcule dois coeficientes de correlação.

42) Calcule a média e o desvio padrão pelas fórmulas de Folk & Ward para os dados do Reforço 20. Calcule os coeficientes de assimetria e de curtose.

43) Calcule a média, o desvio padrão, o 6º decil e a amplitude das idades dos campeões mundiais de vôlei de 1990 (Itália): Gardini (25), Martinelli (25), Di Giorgi (29), Tofoli (24), Masciarelli (27), Anastasi (30), Bracci (24), Bernardi (22), Cantagalli (24), Zorzi (25), Zucchetta (27) e Giani (20). Calcule a moda, a mediana, o desvio medio, a 3ª estatística de ordem, o maior escore padronizado e o posto de Zorzi.

LIÇÃO 5

(1 a) Que são consideradas as mais simples, as mais usadas e as mais "naturais". (1b) Não, pois não é probabilística. (1c) Não, mas é essencial para que possamos avaliar a qualidade das inferências estatísticas. Ela pode ser avaliada nas amostragens probabilísticas. (2) Na estratificada selecionamos algumas unidades de todos os estratos. Na por conglomerados selecionamos todas as unidades de alguns conglomerados. (3) São amostragens onde há uma probabilidade de seleção especificada para todas as amostras do espaço amostral. Devemos usá-las sempre que desejarmos fazer uma inferência estatística. (4b) u_{22} e u_{7794}. (4c) (u_4, u_9, u_{14}, u_{19}, u_{24}, u_{29}, u_{34}, u_{39}, u_{44}, u_{49}) e 1/5. (4d) 32, (u_6, u_{38}, u_{70}, u_{102}, ..., u_{358}) e $IP(u) = 1/32$, $\forall\, u \in U$.
(5 a) Estratos são necessariamente disjuntos e conglomerados não. (5b) Sim: a a.a.s. com reposição, pois a maioria das técnicas elementares de inferência estatística foi elaborada para esta amostragem. (6) e-a-b-c-d (7b) sistemática. (8 a h) em estágios múltiplos. (9 b d f j) estratificada. (9c) por conglomerados. (9e) sistemática. (9 g i k) de voluntários. (9l) aleatória simples. (10b) por conglomerados (12) 454.
(13) Depende. Pode ser por conglomerados se todos os alunos da turma sorteada fizerem parte da amostra. Se alguns alunos de cada turma forem selecionados, trata-se do primeiro estágio de uma amostragem em dois estágios. (15 a) CORRETA. (15b) da população. (15c) ...têm elementos selecionados numa ... (17) (u_7, u_{21}, u_{35}, u_{49}, u_{63}, u_{77}, u_{91}, u_{105}), 1/14 e 1/14. (18 a) Depende da amostragem, pois não são noções intrínsecas. (18b) Pode fornecer inferências estatística também para cada estrato. Não, mas pode ser. Aquele que tiver maior variabilidade das variáveis em estudo. (19) por conglomerados.
(20 a) (u_4, u_{28}, u_{52}, u_{76}, u_{100}) (20b) estratificada com 3 estratos de tamanho 60 e $n_1 = n_2 = n_3 = 6$. (20c) por conglomerados. (20d) (u_6, u_{26}, u_{46}, u_{66}, u_{86}, u_{106}) (21 a) 47, 70, 165, 66 e 52 com $f_i \cong 2/85$ (21b) 1/25, 2/75, 2/175, 1/35 e 2/55 com $n_i = 80$ (22) 5, 10 e 20.
(27 a) Não. Ao contrário, é bastante tosco. (28 a) As amostragens por conglomerados. (28b) Não. Se assim fosse, não se estudaria as demais. (28c) A amostragem sistemática. (29) Não. Mas deveriam ser usadas sempre. (30) Sim, não é um censo. Intencional. Não é probabilística. "Representativa" não tem significado técnico para nós, i.e., a pergunta não tem sentido; porém, ela tem sido considerada "representativa" no significado comum da palavra. (31) Idem. (32) 1/9 e 9. (34) Evidentemente não. Se existisse uma não dispenderíamos tempo estudando as demais. (35 a) Se não entrevistar todos é uma amostragem. (35b) É uma amostragem acidental de conglomerados. (35c) Não é probabilística. (35d) "Representativa" não tem significado técnico para nós, i.e., a pergunta não tem sentido; porém, no significado comum da palavra, a representação depende dos fluxos de entrada e saída e dos intervalos de tempo escolhidos pelos entrevistadores. (35e) Não. Considero péssimo, pois ao entrar ou sair do trabalho os funcionários e operários dificilmente estão dispostos a serem entrevistados.
(36) {u_5, u_6, u_7, u_8}, (u_2, u_{10}, u_{18}, u_{26}, u_{34}, u_{42}), 1/12 e 1/8, 1/12 e 1/8. (38) Em estágios múltiplos. (40b) Sistemática. (42) (u_{18}, u_{31}, u_{45}, u_{73}, u_{86}, u_{100}, u_{128}, u_{141}, u_{155}, u_{183}, u_{196}, u_{210}, u_{238}, u_{251}, u_{265}) (43) Não. Parece ser uma **amostra** sistemática de uma população com **N = 40**. (44) Idem. (45) $n_1 = n_2 = n_3 = 20$ (46) Em nenhuma, já que é considerada a pior amostragem. (47) Não. Como é usado o conhecimento profundo de uma pessoa sobre todos os elementos da população, ela não pode ser grande. (49) Sim. Todas são utilizadas com o objetivo de selecionar uma amostra que represente bem a população. (50) Ocasionalmente sim. A regra é: selecionar mais elementos dos estratos com mais variabilidade ~~Assim, as melhores frações amostrais dos estratos dependem diretamente de suas variabilidades e inversamente de seus tamanhos.~~

LIÇÃO 6

(1)Um molde exato e um mecanismo aleatório. (2) Rapidez e segurança de estar sendo reproduzido o esquema probabilístico definido. (3) Sim, é um molde. Dependendo da situação é aleatório. (4) Não. Geralmente não está atualizado. (5) Depende da situação. Em geral, considera-se as listagens. (6) Não. Por exemplo: a população é formada por todos os eleitores do Brasil e as unidades amostrais do primeiro estágio são os estados. (8) Não. Ela pode ser substituída por outro molde, já existente ou elaborado por nós. (10) A planilha dos leitos, desde que cada paciente tenha um leito... (11) Não. Somente das unidades amostrais necessárias em cada estágio. (12) Não. Necessitam muito de moldes. (14) Não, mas um molde de todos os conglomerados. (18) Não. Podem ser substituídas por moldes. (19) Permitir a realização da amostragem. (20) Não ter todos os elementos da população e ter elementos que não (mais) pertencem à população.

Consideremos uma população U, um espaço amostral A e uma probabilidade P neste espaço amostral. Denominamos a terna (U, A, P) um PLANEJAMENTO AMOSTRAL. Denominamos ESTATÍSTICA a qualquer função (real) definida em todo o espaço amostral. Assim, uma estatística não é um processo de obtenção de uma nova mensuração em cada amostra, mas sim uma transformação (função) numérica dos resultados das mensurações já realizadas.

A média amostral $\bar{X}$, a mediana amostral $\tilde{X}$, a variância amostral var X, a amplitude amostral a (X), o desvio médio dm X, a função distribuição acumulada amostral de um valor $C \in \mathbb{R}^1$, que representamos por $\hat{F}_X$ (c) são geralmente estatísticas. Matematicamente, uma estatística é uma variável aleatória e pode, portanto, ter esperança, variância, mediana e todas as demais características numéricas das variáveis aleatórias. A distribuição (probabilística) de uma estatística depende, é evidente, do planejamento amostral definido.

EXEMPLO 1: Seja U = (Ana, Carmen, Teresa, Jorge, Bruno, Mauro, Davi), X a a idade dos homens e Y a das mulheres. Temos N = 7, $N_x = 4$ e $N_y = 3$.

U	Ana	Carmen	Teresa	Jorge	Bruno	Mauro	Davi
X (u)	/	/	/	23	18	22	22
Y (u)	23	23	22	/	/	/	/

Vamos considerar o espaço amostral A e a probabilidade P definidos pela tabela seguinte:

a	ACT	JMBA	AATC	JAJCC	BMTC	DACT	DMAC	ABDJ
P (a)	0,05	0,10	0,05	0,10	0,15	0,15	0,20	0,20

Aqui X_1* e y_2* não são estatísticas. Tampouco são estatísticas $\bar{X}$, $\tilde{X}$ e $\hat{F}_X$ (22,5), mas $\bar{y}$, $\tilde{y}$ e $\hat{F}_y$ (22,5) são estatísticas (neste planejamento amostral, é claro) cujos valores amostrais estão na tabela seguinte:

a	ACT	JMBA	AATC	JAJCC	BMTC	DACT	DMAC	ABDJ
$\bar{y}$	22 2/3	23	22,75	23	22,5	22 2/3	23	23
$\tilde{y}$	23	23	23	23	22,5	23	23	23
$\hat{F}y$(22,5)	1/3	0	1/4	0	1/2	1/3	0	0

As distribuições (probabilísticas) destas estatísticas são:

$\bar{y}$	22,5	22 2/3	22,75	23
P ($\bar{y}$)	0,15	0,20	0,05	0,60

E $\bar{y}$ = 22,85
Var $\bar{y}$ = 0,0391

$\tilde{y}$	22,5	23
P ($\tilde{y}$)	0,15	0,85

E $\tilde{y}$ = 22,93
Var $\tilde{y}$ = 0,0319

$\hat{F}y$ (22,5)	0	1/4	1/3	1/2
P ($\hat{F}y$ (22,5)	0,60	0,05	0,20	0,15

E $\hat{F}_y$ (22,5) = 0,15
Var $\hat{F}_y$ (22,5) = 0,0391 ▲

EXEMPLO 2: Na tabela a seguir está a relação dos jogadores de futebol convocados para a seleção da Alemanha Ocidental, suas alturas X, seus pesos Y, suas idades Z e o número de partidas já disputadas T. Copa 78 - Argentina).

a) Considerando os 21 jogadores uma população, temos:

$N = N_x = N_y = N_{x,y} = 21 \qquad \sum x_i = 3754 \qquad \sum x_i^2 = 67{,}1540$

$\sum X_i Y_i = 2763{,}31 \qquad \sum Y_i = 1544 \qquad \sum Y^2_i = 114002$

Portanto, podemos calcular os parâmetros seguintes:

$\mu_x = \sum x_i/N = 37{,}54 : 21 = 1{,}79$ e $\mu_y = 73{,}52$

$\sigma_x = (\sum x^2_i/N - \mu_x^2)^{1/2} = (67{,}1540 : 21 - 1{,}79^2)^{1/2} = 0{,}05$ e $\sigma_y = 4{,}79$

$$\rho_{x,y} = \frac{\sum x_i Y_i/N - \mu_x \cdot \mu_y}{\sigma_x \cdot \sigma_y} = \frac{2763{,}31 : 21 - 1{,}79 \times 73{,}52}{0{,}05 \times 4{,}79} = -\,0{,}06$$

i	Ui	Xi	Yi	Zi	Ti
1	Maier	1,83	77	33	74
2	Franke	1,84	76	29	3
3	Burdenski	1,80	73	26	0
4	Vogts	1,68	67	30	78
5	Kaltz	1,84	77	24	5
6	Russman	1,85	76	26	1
7	Dietz	1,78	76	29	12
8	Nogly	1,84	85	30	2
9	Magath	1,72	73	23	1
10	Bonhof	1,80	72	25	22
11	Hoeness	1,81	76	25	35
12	Stielike	1,74	68	22	6
13	Tenhagen	1,82	83	24	1
14	Seel	1,73	72	28	4
15	Fischer	1,78	73	27	2
16	Abramczik	1,80	72	21	2
17	Beer	1,73	67	30	14
18	Müller	1,83	70	23	6
19	Rummenigge	1,82	74	21	3
20	Flöhe	1,75	72	29	26
21	Holzenbein	1,75	65	31	27

b) Considerando estes 21 jogadores como uma amostra (intencional, é claro), podemos calcular as estatísticas seguintes:

$n = n_z = n_t = n_{z,t} = 21$ $\qquad \sum z_i = 556$ $\qquad \sum z_i^2 = 14964$

$\sum z_i t_i = 9447$ $\qquad \sum t_i = 324$ $\qquad \sum t_i^2 = 15160$

$\bar{z} = 26,48$ $\qquad$ des $z = 3,40$ $\qquad$ cor $(z,t) = 0,55$

$\bar{t} = 15,43$ $\qquad$ des $t = 22,00$

$\tilde{z} = 26$

$a\ (z) = 33 - 21 + 1 = 13$

c	18	21	25	30	35	40
$\hat{F}_z\ (c)$	0	2/21	9/21	19/21	1	1

$\tilde{t} = 5$

$a\ (t) = 78 - 0 + 1 = 79$

c	0	5	10	20	30	50
$\hat{F}_t\ (c)$	1/21	11/21	13/21	25/21	18/21	19/21

c) Considerando novamente estes 21 jogadores como uma população, uma possível amostra sistemática de tamanho 7 está na tabela seguinte:

U	Franke	Kaltz	Nogly	Hoeness	Seel	Beer	Flöhe
X (u)	1,84	1,84	1,84	1,81	1,73	1,73	1,75
Y (u)	76	77	85	76	72	67	72
Z (u)	29	24	30	25	28	30	29
T (u)	3	5	2	35	4	14	26

Para esta amostra, temos as seguintes estatísticas:

$\bar{x} = 1,79$ $\qquad$ des $x = 0,05$ $\qquad \tilde{x} = 1,81$ $\qquad a\ (x) = 0,12$

$\bar{y} = 75$ $\qquad$ des $y = 5,18$ $\qquad \tilde{y} = 76$ $\qquad a\ (y) = 19$

$\bar{z} = 27,86$ $\qquad$ des $z = 2,23$ $\qquad \tilde{z} = 29$ $\qquad a\ (z) = 7$

$\bar{t} = 12,71$ $\qquad$ des $t = 12,07$ $\qquad \tilde{t} = 5$ $\qquad a\ (t) = 34$

Aqui não tem sentido calcularmos os coeficientes de correlação entre X e Z, X e T, Y e Z, Y e T. Tais correlações são denominadas CORRELAÇÕES ESPÚRIAS.

$\sum xy = 941,95$ $\qquad$ cor $(x,y) = 0,82$

$\sum zt = 2428$ $\qquad$ cor $(z,t) = -\ 0,27$

d) Cons: lerando o planejamento amostral denominado amostragem sistemática de tamanho 7 sobre U.

a	$(u_1; u_4, \ldots 4_{19})$	$(u_2, u_5, \ldots u_{20})$	$(u_3, u_6, \ldots, u_{21})$
P (a)	1/3	1/3	1/3
$\bar{t}$	27,43	12,71	6,14
des t	31,48	12,07	8,81
$\tilde{t}$	12	5	2
a (t)	78	34	28

$E(\bar{t}) = 27{,}43 \times 1/3 + 12{,}71 \times 1/3 + 6{,}14 \times 1/3 = 15{,}43$

$E(\bar{t}^2) = 27{,}43^2 \times 1/3 + 12{,}71^2 \times 1/3 + 6{,}14^2 \times 1/3 = 317{,}2162$

$Var(\bar{t}) = 317{,}2162 - 15{,}43^2 = 79{,}1313$

$E(\text{des } t) = 17{,}45$ $\quad$ $Var(\text{des } t) = 100{,}1482$

$E(\tilde{t}) = 6{,}33$ $\quad$ $Var(\tilde{t}) = 17{,}5556$

$E(a\ (t)) = 46{,}67$ $\quad$ $Var(a\ (t)) = 496{,}8889$ ▲

REFORÇOS (Respostas na página 72)

1) Diga uma diferença entre parâmetro e estatística.

2) Como calculamos a média da variância amostral?

3) A distribuição de uma estatística depende apenas da população? Verifique sua resposta comparando a distribuição da estatística $\bar{y}$ do Exemplo 1 com as distribuições da mesma estatística para outros dois planejamentos amostrais com a mesma população, o mesmo espaço amostral, mas com um que poderia ser denominado "amostragem ao acaso" e com outro que poderia ser denominado "amostragem proporcional ao tamanho".

4) Na população U =($u_1, \ldots, u_5$) a variável Y mede 5 6 8 7 6 . Uma a.a.s sem reposição de tamanho 3 é selecionada, Qual é a variância relativa da função distribuição acumulada amostral do valor 6,5?

5) Preencha as lacunas com a (s) palavra(s) apropriada (s): A mediana amostral é uma ________________. Por esta razão ela não é uma ___________, ao contrário dos ________.

6) Na população U = ($u_1, \ldots, u_6$) a variável Y mede 5 4 3 3 6 6. Uma a.a.s. sem reposição de tamanho 2 deve ser selecionada. Qual é o coeficiente de variação da mediana amostral?

7) Repita o Reforço 6 para uma amostragem sistemática de tamanho 2.

8) Considere a população do Exemplo 2. Qual é a probabilidade de que uma a.a.s. com reposição de tamanho 3 seja formada por Vogts, Hoeness e Beer? E se for a a.a.s. sem reposição? E se for sistemática?

9) Uma moeda honesta é lançada 15 vezes e apresenta a face "cara" exatamente 6 vezes.

   a) Quem é a população?
   b) Qual é o seu tamanho?
   c) Qual é o tamanho da amostra?
   d) Qual é a proporção populacional?
   e) Qual é a proporção amostral?
   f) Qual era o número esperado de caras antes da realização do experimento?

10) Uma moeda é lançada 50 vezes e apresenta exatamente 25 caras. Responda às perguntas do Reforço 9.

11) Como calculamos a moda da mediana amostral?

12) Considere a 6ª amostra sistemática de tamanho 3 da população do Exemplo 2. Calcule $\bar{x}$, $\bar{y}$, $\bar{z}$ e $\bar{t}$. Estes valores são, a rigor, parâmetros ou estatísticas?

13) Quais das amostragens dos Exemplos 1 e 2 são equiprobabilísticas? Quais delas são uniformes?

14) As quantidades mencionadas nos Reforços 2 e 11 são parâmetros ou estatísticas?

15) Estatísticas são números obtidos da população?

16) As estatísticas são conjuntos ?

17) Qual é o erro amostral no Reforço 9 ? E no Reforço 10 ?

18) Muitas pessoas dizem que uma estatística é um número obtido da amostra selecionada. A rigor, isto é correto ? Se não, qual é uma denominação adequada para tal número ?

 42 reforços

Sempre desejamos conhecer os parâmetros verdadeiros das variáveis com exatidão e certeza. Isto só pode ser conseguido através de um censo, isto é, através da medição das variáveis em todos os elementos da população onde elas estão definidas, embora um censo não seja, na prática, uma garantia para tal grau de conhecimento, em virtude dos ERROS NÃO AMOSTRAIS: erros de mensuração, erros de de registro e erros de cálculo.

Quando um censo é impossível, impraticável ou demasiado inconveniente, tudo o que podemos fazer é "estimar" os parâmetros verdadeiros. Em Estatística a palavra "estimar" é usada sempre se referindo a algum parâmetro verdadeiro de alguma variável e significa "obter um único número supostamente próximo" deste parâmetro verdadeiro. (As exceções ao "sempre" ocorrem fora do nível destas notas). Infelizmente, não é extremamente difícil escutar alguém dizendo "estimar X" ou "estimar $\bar{y}$" ou outro absurdo da mesma natureza: X é uma variável e $\bar{y}$ é uma estatística.

Na prática podemos estimar um parâmetro verdadeiro de uma variável usando uma única amostra, isto é, através de uma ESTIMAÇÃO ESTATÍSTICA, ou não usando amostra alguma, isto é, através de uma ESTIMAÇÃO NÃO ESTATÍSTICA. Uma estimação não estatística é basicamente uma adivinhação, mas aqui esta palavra não deve ser interpretada pejorativamente, pois todo estatístico admite a existência de situações nas quais uma estimação não estatística pode ser muito melhor do que uma estimação estatística. Obviamente nos limitaremos às estimações estatísticas.

Um ESTIMADOR é qualquer função (real) definida em todo o espaço amostral. Do ponto-de-vista matemático, esta definição é idêntica à definição do termo "estatística". Estes entes matemáticos, as funções (reais) definidas em todo o espaço amostral, são"estatísticas" quando os usamos para resumir, condensar, simplificar ou reunir informações da amostra selecionada em único número e são "estimadores" quando os usamos para obter números supostamente próximos dos parâmetros verdadeiros das variáveis. Uma ESTIMATIVA é qualquer número obtido pela aplicação de um estimador a uma amostra.

AMOSTRA $\Rightarrow$ $\boxed{\text{ESTIMADOR}}$ $\Rightarrow$ ESTIMATIVA

Na literatura estatística é bastante comum o uso de letras gregas (μ, σ, ρ, τ etc.) para representar parâmetros verdadeiros e o uso de letras gregas com acento circunflexo ($\hat{\mu}$, $\hat{\sigma}$, $\hat{\rho}$, $\hat{\tau}$ etc.) para representar tanto seus estimadores como também suas estimativas, o que pode causar alguma confusão aos menos atentos. Aliás, nos livros escritos em inglês, há apenas uma palavra para designar estas duas idéias: a palavra "estimate".

EXEMPLO 1: Desejamos estimar μ, a média verdadeira das idades dos estudantes de graduação, matriculados numa universidade em determinado semestre letivo usando uma amostra aleatória simples sem reposição de tamanho 100. Podemos usar um destes estimadores: $\hat{\mu}_1$ (a média amostral), $\hat{\mu}_2$ (a média amostral dos 50 valores centrais), $\hat{\mu}_3$ (a mediana amostral), $\hat{\mu}_4$ (o maior valor amostral) ou $\hat{\mu}_5$ (a metade da média das idades dos pais dos 100 estudantes e, caso algum pai já tenha falecido, considera-se sua idade igual a zero). Qual destes 5 estimadores é o "melhor"? Que significa "melhor"? Há certamente mais uma boa quantidade de estimadores. O "melhor" entre os 5 mencionados é também o "melhor" entre todos os demais?

O problema da escolha do "melhor" estimador está bastante desenvolvido na literatura estatística, evidentemente através de conceitos técnicos e resultados teóricos que os usuários dos métodos estatísticos geralmente não apreciam muito. Precisamos ser sinceros e esclarecer que é muito difícil existir um estimador "melhor" do que todos os outros, a não ser em certas situações bem específicas (de interesse matemático, na maioria das vezes). Assim, uma análise estatística teórica visa quase sempre sugerir o uso de um estimador "bom" ou "ótimo", a palavra pode ser qualquer uma, já que é necessário explicitar, matematizar, definir, conceituar o que ela significa. A forma tradicional de medirmos a precisão de um estimador T para um parâmetro verdadeiro θ é através do ERRO QUADRÁTICO MÉDIO: $Eqm(T, \theta) = E\left[(T-\theta)^2\right] = \text{Var}\,T + (ET-\theta)^2$. Um estimador T é um ESTIMADOR NÃO VICIADO (ou NÃO TENDENCIOSO) PARA O PARÂMETRO θ se, e somente se, $E\left[T-\theta\right] = 0$, ou equivalentemente, $ET = \theta$. O número $E\left[T-\theta\right] = ET - \theta$ é o VÍCIO DO ESTIMADOR T PARA O PARÂMETRO θ. Como o objetivo de usarmos um estimador T é obter uma estimativa o mais próxima possível de θ, é bastante agradável sabermos que o estimador T que estamos usando seja não viciado ou que tenha um vício muito pequeno e, além disso, tenha uma variância pequena: assim seu erro quadrático médio será pequeno. Observemos que se T é um estimador não viciado para o parâmetro θ, então seu erro quadrático médio fica reduzido à sua variância.

Neste enfoque elementar não nos preocuparemos com o problema que acabamos de mencionar nem com os poucos conceitos técnicos mencionados para ilustrá-lo. Nos satisfaremos em usar uma idéia bem simples e bem intuitiva: a média verdadeira de uma variável é estimada pela média amostral desta variável, o desvio padrão verdadeiro pelo desvio padrão amostral, a correlação verdadeira de duas variáveis pela correlação amostral destas duas variáveis etc. Estes estimadores obviamente não são os únicos, nem sempre são os "melhores" e, em certos casos, são até "ruins" ou "péssimos". (O que quer dizer "ruim"? O que significa "péssimo"?) Eles são, entretanto, os mais comumente utilizados e é por esta razão que a amostra selecionada deve ser "representativa" da população.

EXEMPLO 2 - A listagem a seguir fornece o número de faltas X e a soma dos pontos Y obtidos pelos 50 alunos de uma turma de MAT 219 alguns anos atrás.

i	Xi	Yi	i	Xi	Yi	i	Xi	Yi	i	Xi	Yi	i	Xi	Yi
1	9	34	11	6	65	21	12	38	31	8	33	41	3	74
2	13	15	12	3	52	22	7	19	32	4	21	42	4	86
3	11	21	13	9	29	23	0	61	33	3	47	43	5	46
4	11	17	14	5	40	24	0	63	34	0	49	44	2	70
5	8	51	15	1	79	25	11	24	35	4	38	45	5	71
6	1	50	16	6	64	26	6	44	36	10	30	46	8	33
7	15	32	17	3	39	27	8	34	37	5	38	47	2	56
8	4	34	18	0	42	28	4	72	38	8	61	48	1	60
9	10	38	19	4	49	29	4	38	39	12	51	49	5	49
10	7	55	20	3	45	30	6	74	40	2	38	50	4	47

Sorteamos ao acaso e sem reposição 5 elementso e obtivemos a amostra $(u_{24}$ u_{45}, u_{30}, u_3, $u_{15})$.

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
24	0	63	0	3969	0
45	5	71	25	5041	355
30	6	74	36	5476	44
3	11	21	121	441	231
15	1	79	1	6241	79
$\sum$	23	308	183	21168	1109

Com esta amostra, obtivemos as estimativas seguintes:

$$\hat{\mu}_x = \bar{x} = 23 : 5 = 4,6$$

$$\hat{\mu}_y = \bar{y} = 308 : 5 = 61,6$$

$$\hat{\sigma}_x = \text{des } x = \left[183/5 - 4,6^2\right]^{1/2} = 15,44^{\ 1/2} = 3,9$$

$$\hat{\sigma}_y = \text{des } y = \left[21168/5 - 61,6^2\right]^{1/2} = 439,04^{\ 1/2} = 21,0$$

$$\hat{\rho}_{x,y} = \text{cor } (x, y) = \frac{1109 : 5 - 4,6 \times 61,6}{3,9 \times 21,0} = -0,7477$$

▲

REFORÇOS (Respostas na página 73)

1) Calcule os parâmetros estimados no Exemplo 2.

2) O que é tradicionalmente considerado um bom estimador?

3) Escolhe a 4^{a} amostra sistemática de tamanho 5 da população do Exemplo 2. Estime os 5 parâmetros agora com base nesta amostra. Por que, por exemplo, $cor(x,y) \neq \rho_{x,y}$? Como é denominada a diferença $cor(x,y) - \rho_{x,y}$?

4) Considere a população do Exemplo 2. Calcule as medianas e amplitudes verdadeiras de X e de Y. Considere a a.a.s. do Exemplo 2 e a amostra sistemática do Reforço 3. Com base em cada uma, estime estes parâmetros.

5) As amostragens aleatórias simples produzem sempre estimadores mais eficientes do que as outras amostragens?

6) Qual é a diferença essencial entre estimadores e estatísticas?

7) Considere a variável Y do Exemplo 2 e o planejamento amostral descrito a seguir. Calcule a média, o desvio padrão e o Eqm da mediana amostral.

$a_1 = (u_1, u_{11}, u_{21}, u_{31}, u_{41})$ $\qquad P(a_1) = 0,25$

$a_2 = (u_{10}, u_{20}, u_{30}, u_{40})$ $\qquad P(a_2) = 0,2$

$a_3 = (u_{11}, u_{22}, u_{33}, u_{44})$ $\qquad P(a_3) = 0,2$

$a_4 = (u_1, u_2, u_{24}, u_{25}, u_{26}, u_{49}, u_{50})$ $\qquad P(a_4) = 0,35$

8) Numa análise estatística dos 5 estimadores do Exemplo 1 aplicados a uma micropopulação de 1000 estudantes, cuja média verdadeira das idades é 21,62 obtivemos os resultados da tabela a seguir:

i	1	2	3	4	5
$E\,\hat{\mu}_i$	21,62	21,62	20,34	28,75	21,76
$Des\,\hat{\mu}_i$	0,18	0,13	0,10	0,02	0,08

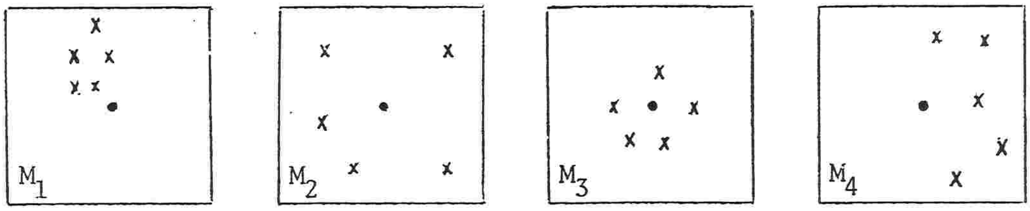
Quais deles são não viciados? Qual deles é o melhor e qual é o pior (nesta população)?

9) No Exemplo 2, aplique a MÉDIA AMOSTRAL 0,4 - PODADA, isto é, a média dos 3 valores centrais, na amostra sorteada para estimar μ_x e μ_y. Estas estimativas são melhores?

10) Se um experimento aleatório é repetido até que um evento ocorra 21 vezes, como você estima a probabilidade de ocorrência do evento?

11) Na população $U = (u_1, \ldots, u_5)$, os valores da variável Y são 6, 3, 8, 2 e 9. Calcule a média, o desvio padrão e o E q m da média amostral para uma a.a.s. com reposição de tamanho 2 e para uma a.a.s. sem reposição de tamanho 2. Qual amostragem é melhor?

12) Um atirador em 4 momentos psicológicos distintos deu 5 tiros em um alvo, conforme as figuras a seguir. Quais as conclusões que você obtém?



13) Uma moeda é lançada 50 vezes e resulta 36 caras. Estime a probabilidade de ocorrer cara em um único lançamento da moeda.

14) Existe algum tipo de relação entre uma média amostral e a média populacional? Em caso afirmativo, qual é esta relação e qual o seu nome?

15) Conceitue "distribuição amostral de um estimador".

16) As 6 primeiras amostras sistemáticas de tamanho 5 da população do Exemplo 2 formam um espaço amostral. A cada uma delas atribuímos probabilidade 1/6. Calcule o coeficiente de variação do coeficiente de correlação amostral entre X e Y.

17) Considere o Exemplo 2 da Lição 8. Você decide usar uma amostragem sistemática de tamanho 3 para estimar o total da variável T. Qual estimador você usará? Calcule sua média, seu desvio padrão e seu E q m.

18) "Se realizamos uma amostragem aleatória simples sem reposição de tamanho n de uma população finita de tamanho N na qual a variável X está definida em todos os elementos da população, então $E(\bar{X}) = \mu_X$ e $Var(\bar{x}) = [(N-n)/(N-1)]\,\sigma_X^2/n$". Use este Teorema para repetir o Reforço 17, substituindo amostragem sistemática por a.a.s. sem reposição. Qual amostragem é melhor ?

19) Conceitue "parâmetro". Conceitue "estimativa".

20) Como se calcula a moda de um estimador?

21) "Um estimador não viciado é um dos melhores estimadores que podemos usar". Comente brevemente.

22) Quando fazemos uma a.a.s. com reposição e desejamos estimar a variância de uma variável X definida em toda a população, são não tendenciosos:

$$\hat{\sigma}_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{e} \quad \hat{\sigma}_2^2 = \frac{1}{2(n-1)}\sum_{i=1}^{n-1}(x_{i+1} - x_i)^2 .$$

Com base na amostra 17,2 10,7 16,6 13,4 6,0 9,6 obtida de uma população com média 12 e variância 15, qual deles nos dá uma estimativa melhor? É este estimador o melhor dos dois? É o melhor de todos?

23) "A média de um estimador é sempre igual ao parâmetro que desejamos estimar." Comente esta frase sob a ótica da amostragem.

24) A tabela a seguir é formada pelo número de pessoas, x_1, a renda semanal, x_2, e as despesas semanais com alimentos, y, de 8 famílias de operários:

i	1	2	3	4	5	6	7	8
x_{1i}	2	3	3	5	4	7	2	4
x_{2i}	62	62	87	65	58	92	88	79
y_i	14,3	20,8	22,7	30,5	41,2	28,2	24,2	30,0

Estime:

a) a despesa semanal média em alimentos por família;

b) a despesa semanal média em alimentos por pessoa;

c) a percentagem da renda que é gasta em alimentos.

25) Para que serve o desvio padrão de um estimador?

26) Como se calcula a média de uma variável? Como se calcula a média de um estimador?

27) A média amostral de uma variável é um estimador não viciado da média populacional desta variável?

28) No Reforço 3 da Lição 9, qual dos 3 esquemas amostrais torna $\bar{y}$ melhor para estimar μ_y?

29) No Exemplo 2, a variável X é o número de faltas e poderia ter assumido valores entre 0 e 25, pois a chamada foi realizada 25 vezes. Assim, se Z é o número de presenças, então Z = 25 - X. É fácil demonstrar que

$$\text{Cor}(aX + b, cY + d) = \text{sgn}(ac) \cdot \text{Cor}(X,Y). \text{ Calcule } \hat{\rho}_{y,z} \text{ e } \rho_{z,y}.$$

30) Numa pesquisa, a técnica de amostragem que será utilizada deve ser escolhida apenas em função das variáveis em estudo ou também em função dos objetivos do pesquisador? Responda com base no Reforço 28.

31) Uma a.a.s. de tamanho 3, sem reposição, é selecionada da população U= (u_1, ..., u_9), onde a variável Y assume os valores 3, 11, 19, 15, 7, 9, 13, 13 e 9. Qual é o coeficiente de variação de $\bar{y}_{\gamma}$? (Dica: use o Teorema do Reforço 18). Idem para a amostragem sistemática de tamanho 3. Idem para a amostragem por conglomerados de tamanho 3, com um só conglomerado sorteado. O coeficiente de variação serve, neste caso, para hierarquizar os planejamentos amostrais? Há alguma reordenação da listagem que melhoraria algum destes esquemas amostrais?

32) As estimativas sempre resultam próximas dos parâmetros?

33) Uma população é formada por 5 elementos: a, b, c, d, e. Os elementos representados por vogais são casados. Construa a distribuição amostral da proporção de casados em um a.a.s. sem reposição de tamanho 3.

34) O que você entende por erro amostral de uma estimativa? Você pode medi-lo?

35) Uma estimação estatística difere de uma estimação não estatística? O que você entende por uma "estimação subjetiva"? E por "estimação dogmática?" Quais delas, a rigor, podem ser consideradas científicas?

36) Uma fábrica de pneus realiza estudos de durabilidade. Uma amostra apresentou os resultados a seguir (em x 1000 Km rodados): 20,2 23,4 22,6 27,6 16,1 21,0 26,3 22,5 18,0 19,1 23,2 23,7 27,4 32,4 30,1 32,2 30,6 28,9 29,3 24,9 27,8. Calcule a média e a média 0,5 - podada (isto é, a média após a eliminação de 25% da cauda de cada amostra). Qual destes estimadores produziu uma melhor estimativa para a média verdadeira?

37) Existe algum tipo de relação entre uma proporção amostral e a proporção populacional? Em caso afirmativo, qual é esta relação e qual é o seu nome?

38) Uma a.a.s. de tamanho 6 apresentou os valores a seguir para uma variável X: 2,9 6,4 3,6 8,1 5,5 e 4,8. Estime a amplitude e o 6º decil.

39) Considere U = {u_1, ..., u_8} com os valores x respectivamente iguais a 1,7 5,6 6,3 4,4 8,1 3,8 5,8 e 7,5. Considere A = {a_1, a_2, a_3}, P(a_1) = P{(u_1, u_2)} =1/4, P(a_2) = {(u_3, u_4)} =1/4 e P(a_3) = P{(u_5, u_6, u_7, u_8)} =1/2. Calcule Eqm$(\bar{x}, \mu)$.

40) A variabilidade de um estimador depende do planejamento amostral ?

41) Para que serve o Eqm de um estimador ?

42) Uma estimação estatística difere de uma estimação não estatística ?

21 reforços

"Não há questões estatísticas de rotina; há rotinas estatísticas questionáveis." (D. Cox)

Numa COLETA DE DADOS, também denominada LEVANTAMENTO DE DADOS, o pesquisador deve: (1) definir precisamente a população, (2) escolher os meios (mais) adequados para medir as variáveis que irá estudar e (3) definir explicitamente os objetivos de seu trabalho.

Estas etapas não são propriamente estatísticas, mas o pesquisador pode; e até deve, contar com o estatístico para elaborá-las. Ao estatístico compete: (4) a escolha das técnicas estatísticas mais adequadas para que os objetivos sejam alcançados e (5) o planejamento de uma amostragem apropriada para as técnicas escolhidas, para a precisão desejada e para as tabelas estatísticas existentes.

O "e até deve", no parágrafo anterior, é porque um pesquisador pouco experiente com coleta e análise de dados pode fazer com que estas atividades despendam tempo e recursos financeiros excessivos. Infelizmente, estes não são os únicos perigos, pois pode acontecer que um pesquisador elabore as etapas (1), (2) e (3) incompatíveis entre si ou impossíveis de serem realizadas na prática. Há também situações nas quais os dados coletados excedem a quantidade necessária para uma boa análise.

Quando uma variável é mencionada, deve estar claramente especificada a população onde ela varia. As variáveis não estão "soltas no ar", como as cabeças daqueles que usam a Estatística com um mínimo, se tanto, de bom senso e de atenção às exigências teóricas.

A função da Estatística evidentemente não é impedir que as variáveis variem na população, mas: (1) tentar revelar exatamente ou aproximadamente como elas variam na população, (2) tentar revelar exatamente ou aproximadamente alguns de seus parâmetros verdadeiros ou (3) medir a verossimilhança de algumas hipóteses feitas a respeito das distribuições verdadeiras de algumas variáveis. A palavra "hipótese" tem, aqui, um significado diferente da palavra "suposição". Esta é usada para afirmações que aceitamos e pronto. Aquela é usada para afirmações que aceitaremos ou rejeitaremos após uma coleta e análise de dados.

11.1 – Raciocínios

Há dois processos lógicos de raciocínios: a dedução (geral $\Rightarrow$ particular) e a indução (particular $\Rightarrow$ geral), também denominada inferência. Um processo dedutivo coreto conduz a uma conclusão certa, isenta de erros e deturpações, denominada DEDUÇÃO. Num processo indutivo correto, a conclusão sobre um todo, denominada INDUÇÃO ou GENERALIZAÇÃO ou INFERÊNCIA, é incerta e pode estar afetada por erro, pois existe possibilidade de atribuirmos ao todo propriedades exclusivas de alguma de suas partes. Uma INFERÊNCIA (ou GENERALIZAÇÃO, ou INDUÇÃO) ESTATÍSTICA é, na maioria das vezes, realizada simultaneamente com uma avaliação probabilística do seu erro. Esta avaliação probabilística informa (mede) a qualidade da conclusão inferida.

Uma indução estatística se diferencia de uma indução não estatística pela incerteza, ou seja, pela variabilidade. "Todas as chamas queimam" e "todos os pingos d'água não queimam" são induções não estatísticas porque presumimos que todas as chamas são idênticas em relação ao efeito que nos interessa, queimar ou não, e o mesmo vale para todos os pingos d'água.

11.2 – Mensurações

As ciências tentam explicar e predizer fenômenos observaveis com base em leis bastante gerais. Nas ciências mais avançadas, tais leis expressam relações quantitativas de certas propriedades do que está sendo estudado. Para que tais leis sejam formuladas é necessário que estas propriedades possam ser expressas por números. O processo no qual um cientista representa propriedades por números é denominado MENSURAÇÃO.

Algumas propriedades, como comprimento e peso, são medidas com métodos tão seculares que até parecem naturais hoje em dia. Outras, como utilidade e inteligência, são medidas com métodos tão recentes que parecem pouco naturais e muito arbitrários.

Algumas críticas às mensurações nas Ciências Humanas são pouco claras em relação aos seus significados, tais como: a mensuração de atributos humanos é mais complicada do que a de atributos físicos, não se pode obter uma precisão tão grande na mensuração de atributos humanos com na de atributos físicos, alguns atributos humanas não podem ser quantificados adequadamente, é muito difícil encontrar um padrão universal para ser a unidade etc.

Cada mensuração fornece certa quantidade de informação, revela certo tipo de estrutura e possui determinado grau de precisão. Os vários tipos de mensuração e seus significados constituem a TEORIA DA MENSURAÇÃO, cuja fundamentação lógica é devida à Hölder, num trabalho publicado em 1901, embora a mensuração seja tão velha como a Ciência.

Ao medimos uma propriedade, construímos um modelo numérico para um pedacinho do mundo. Um sistema formal, numérico ou não, só pode ser considerado um modelo para um pedacinho do mundo se ele representa sua estrutura em seus aspectos essenciais. Uma MENSURAÇÃO é um modelo (numérico) e, portanto, deve expressar o que ocorre naquele pedacinho do mundo. Uma mensuração não é apenas um modo de atribuirmos números a alguma propriedade dos elementos de uma população de acordo com regras bem especificadas.

Há quatro problemas fundamentais em mensurações: REPRESENTAÇÃO (Todos os atributos podem ser medidos ? Se não, quais condições são necessárias para fazermos uma mensuração e quais são suficientes ?), UNICIDADE (Num processo específico de mensuração, os números são determinados univocamente? Se não, quanta liberdade há na sua escolha ?), SIGNIFICADO (Num processo específico de mensuração, que conclusões podem ser obtidas das medidas e que afirmações feitas com base nestas medidas têm significado ?) e ESCALA (Como construir uma escala numérica ? Como converter informações em números ? Como manipular os erros de mensuração causados, por exemplo, pela imprecisão das balanças ou pelas inconsistências dos julgamentos de preferência ?).

Há quatro níveis de escala de mensuração: (1) O mais pobre é formado pelas mensurações da ESCALA NOMINAL, que têm somente a função de classificar os elementos em categorias exaustivas e exclusivas. (2) Quando os elementos não somente diferem de categoria para categoria, mas estas estão totalmente ordenadas segundo algum critério, temos uma mensuração na ESCALA ORDINAL. Os elementos dentro de cada categoria são considerados equivalentes. (3) Quando as diferentes categorias, além de estarem ordenadas, possuem suas distâncias claramente especificadas, a mensuração está na ESCALA INTERVALAR, desde que o ponto zero seja relativo, i.e., ele não indica a ausência total do que está sendo medido. (4) A escala de mensuração mais completa é a ESCALA DE RAZÃO, que é caracterizada por mensurações com um zero verdadeiro, além das demais características da escala intervalar. Portanto, estas mensurações permitem igualar, além das distâncias, também as razões.

11.3 – Dicas

Se todas as etapas são bem planejadas, geralmente obtemos boas conclusões, mesmo com técnicas de inferência bastante simples, mas levantamentos de dados mal planejados geralmente nos levam a grandes enganos, mesmo com técnicas de inferência sofisticadíssimas.

Ao planejarmos um questionário, quer para ser aplicado numa amostra quer na população, devemos verificar se todos os quesitos relevantes para os objetivos da pesquisa fazem parte dele e vice-versa. Devemos evitar a tendência de incluir perguntas desnecessárias, pois acarretam desperdício de tempo na aplicação do questionário e na compilação das respostas. Além disso, podem diminuir a qualidade de algumas respostas aos quesitos relevantes.

Uma boa amostragem não é suficiente para que nosso trabalho produza resultados satisfatórios. É imprescindível que as mensurações utilizadas nas variáveis em estudo sejam adequadas: perguntas corretas em questionários, instrumentos de laboratório aferidos etc.

Toda amostra dá informações (incompletas) sobre a população. Os métodos de inferência estatística procuram utilizar ao máximo tais informações. Se, após planejarmos uma coleta de dados e sua posterior análise, concluímos que a amostra deve ter um certo tamanho, é prudente selecionarmos uma amostra de um tamanho um pouco maior. Assim, além de obtermos mais informações (!?), nos resguardamos contra o extravio de algumas unidades amostrais e contra as NÃO-RESPOSTAS.

Uma AMOSTRAGEM PILOTO é uma amostragem realizada com a finalidade de, a partir da amostra selecionada, planejarmos o levantamento de dados em alguns ou em todos os seus aspectos. Alguns exemplos são: escolha das variáveis (o que medir), escolha da mensuração (como medir), tipo de amostragem, tamanho da amostra e qualidade da técnica de inferência que empregaremos. Em certos casos, a amostra piloto é útil até para (re)definirmos a população. Em geral, quanto maior for nossa ignorância a respeito da população, mais útil, mais importante e mais necessária será a amostra piloto. Eventualmente, ela pode ser a amostra definitiva.

REFORÇOS (Respostas na página 73)

1) Quais são as 3 tarefas primordiais para fazermos uma coleta de dados ?
2) Um pesquisador não precisa planejar seus experimentos com bastante cuidados porque há técnicas estatísticas para quaisquer situações ?
3) Variáveis são funções definidas em espaços amostrais ?
4) Existe diferença entre hipótese e suposição ?
5) A amostra pode ser selecionada antes do planejamento da pesquisa ?
6) Qual é a etapa inicial do planejamento de uma pesquisa ?
7) A Estatística é formada exclusivamente por técnicas indutivas ?
8) Qual é uma diferença entre as induções estatísticas e as induções não estatísticas ?
9) A Estatística é formada exclusivamente por técnicas dedutivas ?
10) Para que servem as amostras ?
11) Existem deduções estatísticas ?
12) Conceitue mensuração. Cite e explique um problema que pode ocorrer numa mensuração.
13) Conceitue escala ordinal.
14) Por que o nível de mensuração intervalar não é o mais completo ?
15) Numa coleta de dados onde você irá trabalhar com 3 variáveis, uma amostra de tamanho 42 fornecerá quantas observações ?
16) O que é uma amostra piloto ?
17) Ao planejarmos uma coleta de dados, devemos aproveitar a ocasião para mensurarmos todas as variáveis que imaginarmos ?
18) Uma boa amostragem assegura uma boa inferência estatística ?
19) Uma coleta de dados pode ser feita de qualquer maneira, pois existem técnicas estatísticas apropriadas para concertar todas as falhas que nela ocorrerem ?
20) O que são amostragens piloto ?
21) Uma amostra realmente representativa da população assegura uma boa indução estatística ?

LIÇÃO 7

(1) 0,23040; 0,08704; 0,91296 (2) 0,054688; 0,000078 (3) 0,0951 (4) 0,2557 (5) 1,5; 1 e 2; $\forall$ $x \in [1 ; 2]$; 2/3; $\sqrt{7/12}$ (6) 1/3; 1,2; $\sqrt{0,56}$; 0,6; 1; 1 (7) 7,5; 7,5; 1,85; 42/55 (9) 0,0149; 0,7149; 0,0620 (10) 0,0952 (11) 0,2642; e^{-15} (12) 0,2707; 0,1353; 0,0527 (14) 0,2358; -5,3838; -1,1133; 0,2699 (15) 0,3980; 0,2578 (17) 516,34 (18) 0,3012; 0,8187; 69,315; 100; 22,314; 160,944 (19) 0,3834; 0,3935; 0 (20) 1/3; 7/9; 10; não existe; 25/12 (21) 0,5768; ¾; 15 (22) 1/12; ¾; $2\sqrt{3}$ (23) 1,08; 0,0668

(13)	a	b	c	d	e	f	g	h	i	j
α	1/3	2/19	$\alpha > 0$	1	3/2	$\nexists$	4	3/4	1/2	$\nexists$
E (X)	5/4	130/57	$\nexists$	23/6	-3/7	$\nexists$	4/3	0	0	$\nexists$
Mod(X)	2	$\nexists$	α	[5;6]	- 1	$\nexists$	1	0	0	$\nexists$
Med(X)	$\sqrt[3]{7/2}$	$\sqrt{13/2}$	2α	21/4	-0,745	$\nexists$	$\sqrt[4]{2}$	0	0	$\nexists$
Var(X)	51/80	1,551	$\nexists$	203/36	102/245	$\nexists$	2/9	1/5	2	$\nexists$
Dm(X)	297/512	1,066	$\nexists$	20/9	0,5443	$\nexists$	9/32	3/8	1	$\nexists$

LIÇÃO 8

(1) 6,95; 3,14; 7; 52/61; 10; 15 (2) 2,464; 2,6; 0,312; 0,387; 2,5; 0,5; 2,2; 1; 0,094; 1,293 (3) 23,65; 1,142; 24; 2,88; 21,2; 25,2 (4) A mediana; ...; centralização e dispersão; 2,4; ...; -0,2; ...; associação (6) 249; 73,5; 210; 20 (7) 503/770 $\cong$ 0,6532 (8 a) Não, pois ... (8b) $\bar{x} = 11$ e $\bar{y} = 14,5$ (8c) $\tilde{x} = 11,5$ e $\tilde{y} = 14,5$ (8d) $\sum r(x) = 89,5$ e $\sum r(y) = 120,5$ (9 a) 0,5497 (9b) Não, pois ... (10) 30,5; 36,4; 15,92; 13,9; -1,539; (11) 9,68; 2,80; 10; 10; 14; 12; 12; 7,8176; 484; 13 (12) 0,17; 899,5; 372; 1049,5; 708,32; ... (14 a) 1,13, 1,28; 1,35; 1,42; 1,65; 1,68 e 2; 5; 3; 1; 6; 4 (14b) 1,4183 e 0,1954 (14c) –0,71; 1,19; -0,35; -1,48; 1,34; 0,01 (14d) 1,385 e 0,180 (15 a) $\bar{x} = 85,1$ e $\bar{y} = 77,9$ (15b) $\tilde{x} = 93$ e $\tilde{y} = 85$ (15c) $\sum r(x) = 55,5$ e $\sum r(y) = 49,5$ (16) 10,33; 4,99; 11; 4 (17 a) $\bar{x} = 61,5$ e $\bar{y} = 67,1$ (17b) $\tilde{x} = 67$ e $\tilde{y} = 71$ (17c) $\sum r(x) = 61$ e $\sum r(y) = 75$ (18) 68/143 $\cong$ 0,4745 (20) 12,29; 9,9; impossível (21) 7,9 (23) 17,92; 18,15; 3,62; 6,74 e 22,92; 23,15; 3,62; 6,74 (24) 0,1746 e 0,2827, 0,1667 e 0,1546 (25) -0,001639; 10,25 (26) 65,292; 23,345; 0,4106; 42,70; 68,02; 88,68 (27) 17,87, 2,20; 18,21; 6,19 (28) 80,59; 48,8; 68,96; 113,925; 1 (29 a) vr (x) = 9 e cv (z) = 1 (29b) cv (x) $\cong$ 0,044 e cv (y) $\cong$ 0,021 (29d) $x_{4/5} - x_{3/5} = x_{2/5} - x_{1/5}$ (32) 56,9625; 107,55; 112,1 (33) –51,012 (34) 3,78; 3,3; ...; ... (35) 0,425 e 0,652 (36) 22 e 28; 25; 25; 5,39, 4,67; 19; 16,8 e 24 e 30; 27; 27; 5,39; 4,67; 19; 16,8 (39) 25; 9; 16 e 0,0625; 0,09; 0,25 (40) 0,7700; impossível (41) 4,643; 1,552; 6,5; 3; 0,6759; 27/35 $\cong$ 0,7714 (42) 66,36; 12,85; -0,1981; 0,9834 (43) 25,17; 2,67; 25; 11; 24 e 25; 25; 37/18 $\cong$ 2,06; 24; 7.

LIÇÃO 9

(1) Um parâmetro é um número e uma estatística é uma função. (4) 1/9 (5) estatística – constante – parâmetros (6) aprox. 0,177 (7) aprox. 0,091 (8) 2/3087, 1/1330 e 0. (9b) ∞ (9c) 15 (9d) 0,5 (9e) 0,4 (9f) 7,5 (10b) ∞ (10c) 50 (10d f) não sei (10e) 0,5 (12 a) 1,806666... 77 26,3333... 9,3333... (12b) Não são parâmetros, pois foram obtidos de uma amostra. Não são estatísticas, pois não são funções. São os valores das estatísticas para a amostra considerada. (13) A do Exemplo 1 e a do Exemplo 2b não são equiprobabilísticas nem uniformes. A do Exemplo 2c e a do Exemplo 2d são sistemáticas, portanto são equiprobabilísticas e uniformes. (14) Nenhum. (15) Não. São funções definidas no espaço amostral. (16) Idem. (17 a) –0,1 (17b) Não sei. (18) Não. Valor da estatística na amostra selecionada.

LIÇÃO 10

(1) $\mu_x = 5{,}64$ $\mu_y = 46{,}32$ $\sigma_x = 3{,}72$ $\sigma_y = 16{,}93$ $\rho = -0{,}5230$ (2) Um que tenha Eqm pequeno. (3 a) (u_4, u_{14}, u_{24}, u_{34}, u_{44}) (3b) $\hat{\mu}_X = 3{,}6$ $\hat{\mu}_Y = 47{,}8$ $\hat{\sigma}_X = 4{,}13$ $\hat{\sigma}_Y = 18{,}63$ $\hat{\rho} = -0{,}8672$ (3c) Por causa das "**flutuações amostrais**", i.e., as amostras diferem entre si e da população. (3d) Erro amostral. (4 a) Med (X) = 5 A(X) = 16 Med (Y) = 45,5 A(Y) = 72 (4b) 5 12 71 59 (4c) 2 12 49 54 (5) Não. Depende da população, da variável, do parâmetro, da outra amostragem e do conceito de "estimador mais eficiente". (6) Nenhuma. Ambos são funções definidas no espaço amostral. (7) 4,61 6,37 40,95 (8 a) $\hat{\mu}_1$ e $\hat{\mu}_2$ (8b) $\hat{\mu}_2$ e $\hat{\mu}_4$ (9 a) $\overline{x}_{0,6} = 4{,}0$ e $\overline{y}_{0,6} = 69{,}3$ (9b) Piores, pois seus erros amostrais são maiores. (10) 21/t, onde t é o total de repetições do experimento. (11 a) 5,6 1,93 3,72 (11b) 5,6 1,67 2,79 (11c) A a.a.s. sem reposição, pois seu Eqm é menor. (13) 0,72 (freqüentista) ou 0,5 (subjetiva) (14) Sim. A média amostral é um número que supomos próximo da média populacional. (Assim, esta relação não é "real" nem matemática, mas mental, ou seja, é uma crença que temos.) A média amostral é uma **estimativa** da média populacional. (15) É a relação de todas as estimativas deste estimador (para todas as amostras do espaço amostral) com as respectivas probabilidades delas ocorrerem. (16) –0,1935 (17) $\hat{\tau} = 7\sum t_i$ 324 256,64 65866 (18) $E(\hat{\tau}) = 324$ $Des(\hat{\tau}) = 253{,}1$ $Eqm(\hat{\tau}, \tau) = 64{,}0152$. Esta é melhor, pois seu $Eqm(\hat{\tau})$ é menor. (22 a) $\hat{\sigma}_1^2 = 18{,}647$ e $\hat{\sigma}_2^2 = 15{,}502 \Rightarrow \hat{\sigma}_2^2$ nos dá a melhor estimativa. (22b) Não sei. Para isto é necessário compará-los em todas as amostras. (22c) Não sei. Para isto é necessário compará-lo com todos os outros em todas as amostras do espaço amostral. (24) 26,49 7,06 35,73% (27) Pode ser ou não. Depende do esquema amostral. (28) A "amostragem ao acaso", pois seu Eqm é menor. (29) 0,7477 e 0,5230. (30) Depende também dos objetivos do pesquisador. Por exemplo: No Reforço 28, o pesquisador deseja estimar μ_y com o estimador $\overline{y}$. Cada esquema amostral produz um Eqm, dos quais o menor é o da "amostragem ao acaso". Para outro objetivo, isto poderia não acontecer. (33) p(0) = 0,1 p(1/3) = 0,6 p(2/3) = 0,3 (36) 25,11 25,3 Não sei, não a conheço. Se a conhecesse, porque iria estimá-la? (37) Veja a resposta do Reforço 14. (38) 5,3 e 5,68 (39) 1,17125 (40) Sim, vide o Reforço 28. (41) Para expressar uma medida de sua qualidade. (42) Sim, a estimação estatística tem base numa amostra e a estimação não estatística tem base no "conhecimento" de quem a faz.

LIÇÃO 11

(2) Ao contrário. Um pesquisador precisa sempre planejar seus experimentos com o maior cuidado possível, porque, entre outras razões, qualquer técnica estatística pode ser utilizada somente em certas situações específicas. (3) Não. Em subpopulações. (5) Pode, mas nunca deve, embora muitos ignorantes freqüentemente o façam. (7) Não. Por técnicas dedutivas, indutivas, descritivas e seletivas (amostragens). (9) Idem. (8) As induções estatísticas são as realizadas em situações consideradas aleatórias e as não estatísticas em situações consideradas determinísticas. (10) Para permitir que façamos induções estatísticas. (11) Sim. Por exemplo, quando calculamos parâmetros. (13) É uma escala de mensuração na qual os números não apenas diferenciam as categorias, mas também as ordenam. (14) Porque lhe falta o zero absoluto. (15) No máximo 126. (18) Não, são também importantes a mensuração e a técnica de inferência utilizadas. Além disso, uma "boa" amostragem não garante uma "boa" amostra. (21) Idem. (20) São amostragens que produzem amostras piloto.

 $40 = 32 + 8$ reforços

12.1 – Introdução

Consideremos uma população finita de tamanho N. Consideremos uma variável X definida em toda a população que assume apenas os valores 0 e 1. Se X = 1, então dizemos que este elemento da população é um SUCESSO; caso contrário, dizemos que ele é um FRACASSO. Na verdade, a variável X simplesmente particiona a população em dois grupos mutuamente exclusivos: os "sucessos" e os "fracassos."

Vamos representar por τ o total de sucessos na população, por per a percentagem de sucessos na população e por p a proporção de tais sucessos. Assim:

$$p = \tau / N, \qquad \text{per} = 100 \times p \qquad \text{e} \qquad \tau = N \times p.$$

Consideremos o espaço amostral formado por todas as amostras aleatórias simples sem reposição de tamanho n da população. Assim, existem $N \times (N-1) \times \ldots \times (N-n+1)$ amostras equiprováveis . Para cada uma destas amostras, vamos representar por T a sua quantidade de sucessos. Assim, a estatística T tem distribuição Hipergeométrica:

$$E(T) = np, \qquad \text{Var}(T) = \frac{N-n}{N-1} np(1-p) \qquad \text{e} \qquad P\{T = t\} = \frac{\binom{\tau}{t}\binom{N-\tau}{n-t}}{\binom{N}{n}}.$$

12.2 – Estimadores por Ponto

É natural estimarmos a proporção de sucessos na população pela proporção de sucessos na amostra. Assim: $\hat{p} = T/n$, $\hat{p}\text{er} = 100 \times \hat{p}$ e $\hat{\tau} = N \times \hat{p}$.

TEOREMA 1. $E(\hat{p}) = p$, $E(\hat{p}\text{er}) = \text{per}$, $E(\hat{\tau}) = \tau$, $\text{Var}(\hat{p}) = \dfrac{N-n}{N-1} \times \dfrac{p(1-p)}{n}$,

$$\text{Var}(\hat{p}\text{er}) = 100^2 \, \text{Var}(\hat{p}) \quad \text{e} \quad \text{Var}(\hat{\tau}) = N^2 \text{Var}(\hat{p}).$$

Uma medida da qualidade de um estimador não viciado é sua variância (ver pág. 62). Pelo Teorema 1, as variâncias de $\hat{p}$, $\hat{p}\text{er}$ e $\hat{\tau}$ podem ser calculadas somente se o parâmetro p é conhecido. Ora, se p é conhecido, não há problema de inferência estatística; portanto, não há necessidade de selecionarmos uma amostra.

Numa situação prática, tudo que podemos fazer é estimar as variâncias destes estimadores. Os estimadores naturais para estas variâncias são:

$$\hat{\text{V}}\text{ar}(\hat{p}) = \frac{N-n}{N-1} \times \frac{\hat{p}(1-\hat{p})}{n}, \quad \hat{\text{V}}\text{ar}(\hat{p}\text{er}) = 100^2 \, \hat{\text{V}}\text{ar}(\hat{p}) \quad \text{e} \quad \hat{\text{V}}\text{ar}(\hat{\tau}) = N^2 \hat{\text{V}}\text{ar}(\hat{p}).$$

É claro que podemos utilizá-los, mas os seguintes são melhores:

$$v(\hat{p}) = \left(1 - \frac{n}{N}\right) \times \frac{\hat{p}(1-\hat{p})}{n-1}, \qquad v(\hat{p}er) = 100^2 v(\hat{p}) \qquad e \qquad v(\hat{\tau}) = N^2 v(\hat{p}).$$

TEOREMA 2. Os estimadores $\hat{p}$, $\hat{p}er$ e $\hat{\tau}$ têm DISTRIBUIÇÕES ASSINTÓTICAS normais, isto é, quando o tamanho da amostra é grande, suas distribuições são aproximadamente normais.

EXEMPLO 1 - Consideremos uma listagem de 3042 nomes com endereços. Uma a.a.s. sem reposição de tamanho 200 apresentou 38 endereços desatualizados. Vamos estimar o total de endereços desatualizados.

$$\hat{p} = T/m = 38/200 = 0{,}19 \therefore \hat{\tau} = N\hat{p} = 3042 \times 0{,}19 = 577{,}98 \therefore \hat{\tau} = 578$$

$$\rho(\hat{\tau}) = N\left[\left(1-\frac{m}{N}\right)\frac{\hat{p}(1-\hat{p})}{m-1}\right]^{1/2} = 3042 \times \left[\left(1-\frac{200}{3042}\right) \times \frac{0{,}19 \times 0{,}81}{199}\right]^{1/2} = 81{,}77$$

Como $P\{\lvert\hat{\tau} - E(\hat{\tau})\rvert \leq 3 \times Des(\hat{\tau})\} \doteq 0{,}9974$, segue que $\lvert\hat{\tau} - \tau\rvert$ é "certamente" menor ou igual a $3 \times \rho(\hat{\tau})$, isto é, τ está "certamente" entre os valores $\hat{\tau} - 3 \times \rho(\hat{\tau})$ e $\hat{\tau} + 3 \times \rho(\hat{\tau})$.

Aqui neste exemplo, τ está "certamente" compreendido entre $577{,}98 \pm 3 \times 81{,}77 = [333;\ 823]$, isto é, o total de endereços desatualizados na listagem é "certamente" maior ou igual a 333 e menor ou igual a 823. ▲

EXEMPLO 2 - Consideremos uma populaçao com 800 elementos. Uma a.a.s. sem reposição de tamanho 100 apresentou 29 elementos com determinada característica. Estamos interessados em estimar a percentagem de elementos da população que apresentam a característica mencionada.

$$\hat{p} = T/m = 29/100 = 0{,}29 \therefore \hat{p}er = 100 \times \hat{p} = 100 \times 0{,}29 = 29\%$$

$$\rho(\hat{p}er) = 100 \times \rho(\hat{p}) = 100 \times \left[\left(1 - \frac{100}{800}\right) \times \frac{0{,}29 \times 0{,}71}{99}\right]^{1/2} = 4{,}27\%$$

$\hat{p}er \pm 3 \times \rho(\hat{p}er) = 29\% \pm 3 \times 4{,}27\% = [16{,}20\%;\ 41{,}80\%]$, isto é, a percentagem populacional dos que apresentam a característica em questão está "certamente" entre as percentagens 16,20% e 41,80%. ▲

12.3 – Estimadores por Conjunto

Quando usamos um estimador por ponto, temos uma grande precisão e também uma grande incerteza, pois a estimativa é um número (nada é mais preciso do que um número) supostamente (nada é mais incerto do que uma suposição) próximo do parâmetro. Vamos trocar um tanto de precisão por uma boa quantidade de certeza, que denominamos CONFIANÇA. Como? Substituindo a estimativa pontual por uma estimativa por conjunto.

EXEMPLO 3 - Em uma a.a.s. sem reposição de 200 professores de uma universidade com 2000 professores, 120 são favoráveis a uma greve e 57 são contrários. Vamos usar o COEFICIENTE DE CONFIANÇA $\gamma = 0,95$ para obtermos um (a ESTIMATIVA POR) INTERVALO DE CONFIANÇA 0,95 para o total de professores favoráveis à greve.

$$\boxed{\begin{array}{l} P\left\{-1,960 \leq \dfrac{\hat{\tau} - E(\hat{\tau})}{Des(\hat{\tau})} \leq +1,960\right\} \doteq 0,95 \\ P\left\{\hat{\tau} - 1,960 \times Des(\hat{\tau}) \leq E(\hat{\tau}) \leq \hat{\tau} + 1,960 \times Des(\hat{\tau})\right\} \doteq 0,95 \\ P\left\{\hat{\tau} - 1,960 \times \rho(\hat{\tau}) \leq \tau \leq \hat{\tau} + 1,960 \times \rho(\hat{\tau})\right\} \doteq 0,95 \end{array}}$$

$$\hat{p} = T/n = 120/200 = 0,6 \;\therefore\; \hat{\tau} = N\hat{p} = 2000 \times 0,6 = 1200$$

$$\rho(\hat{\tau}) = N \times \left[\left(1 - \frac{n}{N}\right) \times \frac{\hat{p}(1-\hat{p})}{n-1}\right]^{1/2} = 2000 \times \left[\left(1 - \frac{200}{2000}\right) \frac{0,6 \times 0,4}{199}\right]^{1/2} = 65,89$$

$$\hat{\tau} \pm 1,960 \times \rho(\hat{\tau}) = 1200 \pm 1,960 \times 65,89 \;\therefore\; \boxed{\underset{0,95}{1071 \overset{\tau}{\longmapsto} 1329}}$$ ▲

EXEMPLO 4 - Se, no Exemplo 3, a amostra tivesse apresentado 57 professores favoráveis à greve, então: $\hat{p} = 0,285$ $\hat{\tau} = 570$ $\rho(\hat{\tau}) = 60,72$ $\therefore \boxed{\underset{0,95}{451 \overset{\tau}{\longmapsto} 689}}$ ▲

Nos Exemplos 3 e 4 aplicamos o mesmo método, um (ESTIMADOR POR) INTERVALO DE CONFIANÇA 0,95 PARA τ, em duas amostras e obtivemos duas estimativas por intervalo de confiança 0,95 para τ. Se aplicarmos este estimador numa 3ª amostra, obteremos uma 3ª estimativa por intervalo. Podemos aplicar este estimador em qualquer a.a.s. sem reposição de tamanho 200. Se nós o aplicarmos a todas 2000 x 1999 x ... x 1801 amostras, obteremos uma grande quantidade de intervalos, dos quais aproximadamente 95% deles "capturaram" o parâmetro, isto é, τ pertencerá a aproximadamente 95% deles. Assim, o coeficiente de confiança 0,95 significa que "com probabilidade 0,95 o método é eficiente", isto é, "com probabilidade 0,95 a amostra sorteada produz um intervalo (aleatório) ao qual o parâmetro (fixo) pertença", isto é, "se usarmos este método uma grande quantidade de vezes, então em aproximadamente 95% das vezes o parâmetro pertencerá ao intervalo obtido."

No Exemplo 3 substituímos "uma estimativa (pontual) para τ é 1200" por "um (a estimativa por) intervalo de confiança 0,95 para τ é $[1071; 1329]$". No Exemplo 4 substituimos "uma estimativa (pontual) para τ é 570" por "um (a estimativa por) intervalo de confiança 0,95 para τ é $[451; 689]$". Assim, como dizíamos antes destes dois exemplos, sacrificamos a precisão para ganharmos confiança. Como tudo na vida, com exceção da infância e das mordomias escolares e acadêmicas, para se lucrar aqui há que se perder ali ou vice-versa.

Há situações práticas nas quais estamos interessados em estimativas (por ponto) e há situações práticas nas quais estamos interessados em (estimativas por) intervalos de confiança. Há também situações práticas nas quais estamos interessados em limites de confiança.

EXEMPLO 5 - Para os dados do Exemplo 3, podemos obter um (a ESTIMATIVA POR) LIMITE INFERIOR DE CONFIANÇA 0,95 PARA τ da seguinte maneira:

$$\boxed{\begin{array}{l} P\left\{\dfrac{\hat{\tau} - E(\hat{\tau})}{Des(\hat{\tau})} \leqslant 1{,}645\right\} \doteq 0{,}95 \\ P\left\{E(\hat{\tau}) \geqslant \hat{\tau} - 1{,}645 \times Des(\hat{\tau})\right\} \doteq 0{,}95 \\ P\left\{\tau \geqslant \hat{\tau} - 1{,}645 \times \rho(\hat{\tau})\right\} \doteq 0{,}95 \end{array}}$$

$$\hat{\tau} - 1{,}645 \times \rho(\hat{\tau}) = 1200 - 1{,}645 \times 65{,}89 = 1091{,}61 \;\therefore\; \boxed{1092 \underset{0,95}{\overset{\tau}{\longmapsto}} 2000}$$

EXEMPLO 6 - Para os dados do Exemplo 4, podemos obter um (a ESTIMATIVA POR) LIMITE SUPERIOR DE CONFIANÇA 0,95 para τ da seguinte maneira:

$$\boxed{\begin{array}{l} P\left\{\dfrac{\hat{\tau} - E(\hat{\tau})}{Des(\hat{\tau})} \geqslant -1{,}645\right\} \doteq 0{,}95 \\ P\left\{E(\hat{\tau}) \leqslant \hat{\tau} + 1{,}645 \times Des(\hat{\tau})\right\} \doteq 0{,}95 \\ P\left\{\tau \leqslant \hat{\tau} + 1{,}645 \times \rho(\hat{\tau})\right\} \doteq 0{,}95 \end{array}}$$

$$\hat{\tau} + 1{,}645 \times \rho(\hat{\tau}) = 570 + 1{,}645 \times 60{,}72 = 660{,}88 \;\therefore\; \boxed{0 \underset{0,95}{\overset{\tau}{\longmapsto}} 669}$$

Compete ao pesquisador decidir qual tipo de (estimativa por) conjunto de confiança lhe é mais adequado (a). Nisso, o estatístico pode ajudá-lo.

O Teorema 2 afirma que os estimadores $\hat{p}$, $\hat{p}er$ e $\hat{\tau}$ têm distribuição aproximadamente normal quando n é "suficientemente grande". A validade desta aproximação pode ser encontrada na literatura específica. Vamos proceder assim: se $n < 47$, então usamos a DISTRIBUIÇÃO t de STUDENT com $n - 1$ GRAUS DE LIBERDADE. Há uma bela teoria envolvendo a criação desta família de distribuições, que pode ser encontrada em livros de Estatística Matemática, mas vamos entendê-la apenas como um tributo que pagamos por usar amostras pequenas.

EXEMPLO 7 - Uma a.a.s. sem reposição de 28 crianças apresentou 11 filhos de pais separados. Vamos obter um intervalo de confiança 0,9 para a percentagem de crianças da população que são filhos de pais separados.

$\hat{p} = T/\ n = 11/28 = 0,39\ \therefore\ \hat{p}er = 100 \times \hat{p} = 100 \times 0,39 = 39,29\%$

$$\rho\ (\hat{p}er) = 100 \times \rho\ (\hat{p}) = 100\text{x} \left[(1-\frac{n}{N})\frac{\hat{p}\,(1-\hat{p})}{n-1}\right]^{1/2} = 100\text{x}\left[1\text{x}\frac{0,39\text{x}0,61}{27}\right]^{1/2} = 9,40\%$$

$$\hat{p}er \pm 1,7033\text{x}\rho\,(\hat{p}er) = 39,29 \pm 1,7033 \times 9,40\ \therefore\ \boxed{23,28\% \underset{0,9}{\overset{per}{\longmapsto}} 55,30\%}$$

▲

EXEMPLO 8 - Com os dados do Exemplo 7, vamos obter um intervalo de confiança 0,99 para o parâmetro per.

$$\hat{p}er \pm 2,7707\ \text{x}\rho(\hat{p}er) = 39,29 \pm 2,7707 \times 9,40\ \therefore \boxed{13,24\% \underset{0,99}{\overset{per}{\longmapsto}} 65,33\%}$$

▲

No Exemplo 7, a amplitude do intervalo de confiança 0,9 é 2 x 1,7033x9,40% = 32,02%. No Exemplo 8, onde o coeficiente de confiança é maior, a precisão é menor, pois a amplitude do intervalo é 2 x 2,7707 x 9,40% = 52,08%. Isto é indesejável, mas real. A finitude do ser humano recomenda que sejamos indulgentes ao julgarmos suas criações. Entre elas se encontra, sem dúvida, a metodologia estatística.

12.4 – Testes de Hipóteses

Suponhamos que uma moeda é honesta. Com o objetivo de verificarmos se esta hipótese é verdadeira, lançamos a moeda, vamos dizer, 100 vezes e obtemos 36 caras. Que fazemos? ACEITAMOS a hipótese ou a REJEITAMOS? Se tivéssemos obtido 48 caras nos 100 lançamentos, aceitaríamos a hipótese; se, ao invés disso, tivéssemos obtido 5 caras nos 100 lançamentos, rejeitaríamos a hipótese.

Vamos usar p para representar a probabilidade de ocorrer cara em um lançamento da moeda. A "hipótese" é também denominada HIPÓTESE NULA e é representada por H_o. Assim, $H_o: p = 0,5$. A ALTERNATIVA para esta hipótese é que a moeda não é honesta. A "alternativa" é também denominada HIPÓTESE ALTERNATIVA e é representada por H_1. Assim, $H_1:\ p \neq 0,5$.

Vamos usar T para representar o número de sucessos na amostra, ou seja, o número de caras nos 100 lançamentos da moeda. O VALOR ESPERADO DA ESTATÍSTICA T SOB H_o, isto é, o valor esperado de T quando H_o é realmente verdadeira, é $n\ p_o$= 100 x 0,5 = 50.

Intuitivamente, se o VALOR OBTIDO para T e o valor esperado para T sob H_o estiverem próximos, a amostra está "coerente" com a hipótese e, portanto, ela PODE ser aceita. Caso contrário, ela DEVE ser rejeitada. Outro modo de proceder é calcularmos uma "distância" entre $\hat{p}$ e p_o, o VALOR HIPOTÉTICO de p. Se esta distância é pequena, podemos aceitar H_0. Caso contrário, devemos rejeitá-la. Afinal, a amostra representa a realidade e, se nela não confiássemos, por que teríamos tido o trabalho de selecioná-la e de fazermos mensurações? De certa forma, a amostra "é" a realidade e a hipótese é apenas uma idéia nossa a respeito da realidade considerada. Elas se ajustam? Esta pergunta é respondida através de um TESTE DA HIPÓTESE.

A DISTÂNCIA CARTESIANA entre $\hat{p}$ e p_0 é $|\hat{p} - p_0| = |0{,}36 - 0{,}50| = 0{,}14$. Esta distância é "grande"? Ela é "significativa"? Em Inferência Estatística não trabalhamos com distâncias cartesianas, mas com outra, que denominamos DISTÂNCIA PROBABILÍSTICA. Esta é sempre expressa por um número entre 0 e 1, o que ajuda bastante a classificá-la em pequena ou grande.

TEOREMA 3. $\hat{p}$ tem DISTRIBUIÇÃO NULA ASSINTÓTICA Normal com $E_0(\hat{p}) = p_0$ e

$$Des_0(\hat{p}) = \left[\frac{N-n}{N-1} \times \frac{p_0(1-p_0)}{n}\right]^{1/2}, \text{ isto é, } \quad Z = \frac{\hat{p} - p_0}{Des_0(\hat{p})} \to N(0;1).$$

Como $\hat{p} = 0{,}36$ e $p_0 = 0{,}50$ implicam $Z = -2{,}80$ e como $d_0\,(-2{,}80;\,0) \doteq 0{,}4974$ é, sem dúvida, grande, rejeitamos H_0. Em outras palavras, a diferença entre $\hat{p}$ e p_0 é significativa. Acontece que tornou-se uma convenção medir as distâncias probabilísticas do valor obtido para (uma d) as estremidades e não do valor obtido para o valor esperado como fizemos. Quando uma distância probabilística é medida do valor obtido para (uma d) as extremidades "de forma correta", ela é denominada NÍVEL CRÍTICO AMOSTRAL e é representada por α. Para o nosso resultado: $\alpha(T) = \alpha(36) \doteq 2 \times 0{,}0026 = 0{,}0052$ ou $\alpha(\hat{p}) = \alpha(0{,}36) \doteq 0{,}0052$ ou $\alpha(Z) = \alpha(-2{,}80) \doteq 0{,}0052$. Esta distância probabilística, ou melhor, este nível crítico amostral, é pequeno. Portanto, $Z = -2{,}80$ está próximo de $-\infty$, isto é, $\hat{p}$ é significativamente menor do que p_0 e, assim, rejeitamos Ho.

EXEMPLO 9 - Uma a.a.s. sem reposição de 186 eleitores de certa zona eleitoral apresentou cerca de 54,84% que votarão em determinado candidato. Vamos testar a hipótese de que ele não obterá a maioria dos votos nesta zona eleitoral. Seja p a proporção de eleitores favoráveis ao candidato. Assim, H_0: $p \leq 0{,}5$ e H_1: $p > 0{,}5$. A distância entre a amostra e a hipótese é a distância entre $\hat{p} = 0{,}5484$ e $p_0 = 0{,}5$: $\alpha(Z) = \alpha(1{,}32) \doteq 0{,}0934$. Se consideramos este nível crítico amostral pequeno, então 1,32 está próximo de $+\infty$, isto é, $\hat{p}$ é significativamente maior que p_0 e, portanto, a alternativa deve ser verdadeira. Logo, rejeitamos H_0. Se consideramos este nível crítico amostral grande, então 1,32 está próximo de 0, isto é, $\hat{p}$ não é significativamente maior que p_0 e, portanto, a hipótese pode ser verdadeira. Logo, aceitamos H_0. ▲

EXEMPLO 10 - Com os dados do Exemplo 9, vamos testar a hipótese de que o candidato venceria as eleições na zona eleitoral mencionada. Assim, H_o: $p > 0,5$ e H_1: $p \leqslant 0,5$. A distância entre a amostra e a hipótese é 0, pois $\hat{p} = 0,5484$ está "mergulhada" na hipótese. Portanto, ela pode ser verdadeira. Logo, aceitamos H_o. ▲

O que é afinal, um nível crítico amostral "grande" e o que é um nível crítico amostral "pequeno"? Por que dizemos "a hipótese pode ser verdadeira" e "a hipótese deve ser falsa"? Por que sempre nos referimos à hipótese quer aceitando-a quer rejeitando-a? Esta pergunta tem resposta imediata: é uma convenção.

O que é, afinal, um nível crítico amostral "grande" e o que é um nível crítico amostral "pequeno"? Esta decisão é subjetiva e pessoal, mas isto não é um empecilho para as tomadas de decisões estatísticas. Há uma forma alternativa ao n.c.a. para tomarmos decisões estatísticas: antes de coletarmos a amostra, ao planejarmos o experimento, fixamos arbitrariamente uma probabilidade pequena α, o NÍVEL DE SIGNIFICÂNCIA DO TESTE. Se o valor obtido para uma estatística ou para um estimador estiver numa região "adequada" cuja probabilidade sob H_o for menor ou igual ao nível de significância do teste, então rejeitamos H_o; em caso contrário, aceitamos H_o. Há ainda outra expressão com significado parecido ao de n.s. e ao de n.c.a.: TAMANHO DO TESTE, que não iremos estudar, mas que também é representado por α .

EXEMPLO 11 - Numa pesquisa de mercado há 2000 compradores em potencial para um novo produto e 57 dos 150 entrevistados garantiram que o comprariam. Vamos usar o n.s. $\alpha = 0,05$ para testar a hipótese de que pelo menos a metade da população compraria o tal do produto. Seja τ o total dos que comprariam. Assim, H_o: $\tau \geqslant 1000$ e H_1 : $\tau < 1000$.

A hipótese nula deve ser rejeitada para valores estimados de τ extremamente (isto é, significativamente) distante dela, ou seja, quando $\hat{\tau} << 1000$; em outras palavras, quando $Z = (\hat{\tau} - \tau_o) / Des_o(\hat{\tau})$ for significativamente pequeno, isto é, quando $Z \leqslant -1,645$.

$$\hat{p} = T/n = 57/150 = 0,38 \quad \therefore \quad \hat{\tau} = N\,\hat{p} = 2000 \times 0,38 = 760$$

$$Des_o(\hat{\tau}) = N \times Des_o(\hat{p}) = 2000 \times \left[\frac{2000 - 150}{2000 - 1} \times \frac{0,5 \times 0,5}{150}\right]^{1/2} = 78,55$$

$$Z = (\hat{\tau} - \tau_o)/ Des_o(\hat{\tau}) = (760 - 1000)/ 78,55 = -3,06 \quad \therefore \text{ rejeitamos } H_o.$$

▲

EXEMPLO 12 - Na situação do Exemplo 11, vamos testar a hipótese de que menos do que 580 comprariam o tal produto. Assim, H_o : $\tau < 580$ e H_1: $\tau \geqslant 580$. A REGIÃO DE REJEIÇÃO agora é $\{Z \geqslant +1,645\}$.

$$Z = (\hat{\tau} - \tau_o)/ Des_o(\hat{\tau}) = (760 - 579)/ 71,25 = 2,54 \quad \therefore \text{ rejeitamos } H_o.$$ ▲

EXEMPLO 13 - Na situação do Exemplo 11, vamos testar a hipótese de que exatamente 700 comprariam o tal produto. Assim, H_0: $\tau = 700$ e H_1: $\tau \neq 700$. A REGIÃO CRÍTICA agora é $\{Z \leqslant -1{,}960\} \cup \{Z \geqslant 1{,}960\}$.
$Z = (\hat{\tau} - \tau_0)/ Des_0(\hat{\tau}) = (760 - 700)/ 74{,}93 = 0{,}80 \therefore$ aceitamos H_0.

▲

EXEMPLO 14 - Na situação do Exemplo 11, vamos testar a hipótese de que exatamente 800 comprariam o tal produto. Assim, H_0: $\tau = 800$ e H_1: $\tau \neq 800$. A REGIÃO DE REJEIÇÃO (ou REGIÃO CRÍTICA), é a mesma do Exemplo 13 e $Z = (760 - 800)/ 76{,}96 = -0{,}52 \therefore$ aceitamos H_0. ▲

A conclusão experimental do Exemplo 11 é "$\tau < 1000$"? A do Exemplo 12 é "$\tau \geq 580$"? A do Exemplo 13 é "$\tau = 700$"? A do Exemplo 14 é "$\tau = 800$?" Não, não, não e não. Quais são elas?

REFORÇOS (Respostas na página 89)

1) Responda à última pergunta da Seção 4.

2) Escreva as regiões críticas dos Exemplos 11 - 14 em função da Estatística $\hat{\tau}$.

3) Quais seriam as regiões críticas dos exemplos 9 - 14 em função da estatística T para o n.s. 0,01?

4) Qual é uma diferença entre região crítica e região de rejeição?

5) O uso do nível de significância pressupõe que, sendo H_0 verdadeira, não deve ocorrer um valor para a estatística do teste que seja uma "aberração" na DISTRIBUIÇÃO NULA desta estatística. Qual é o nome da região formada pelos valores considerados "aberrantes" sob H_0? O(s) valor(es) que separa(m) tal região da sua região complementar é (são) denominado(s) VALOR(ES) CRÍTICO(S). Quais são os valores críticos dos Exemplos 9 - 14? Quais são os valores críticos do Reforço 3?

6) O resultado do Exemplo 3 fornece uma "evidência conclusiva" de que a maioria dos professores é favorável à greve? Obtenha um intervalo de confiança 0,99 para τ. Com este resultado, responda à pergunta.

7) Uma moeda honesta, lançada 1000 vezes, apresentou 529 caras. Obtenha um intervalo de confiança 0,9 para a probabilidade de ocorrer cara em um único lançamento.

8) Os fatores $1 - n/N$ e $(N - n)/(N - 1)$ são denominados FATORES DE CORREÇÃO DAS POPULAÇÕES FINITAS (C.P.F.) e devemos omiti-los quando a população é infinita ou quando fazemos a.a.s. com reposição. Eles são geralmente omitidos quando planejamos uma pesquisa porque assim as fórmulas ficam mais simples e os tamanhos de amostra ficam um pouco maior. Obtenha um intervalo de confiança 0,9 para a probabilidade de ocorrer um sucesso num experimento no qual uma a.a.s. com reposição de tamanho 80 apresentou 2 sucessos.

9) Com os dados do Exemplo 11, teste Ho: $\tau \neq 600$.

10) Uma moeda é lançada e obtemos 226 caras. Você aceita ou rejeita a hipótese de que ela é honesta?

11) O que é distribuição de uma estatística? O que é distribuição nula de uma estatística? O que é distribuição assintótica de uma estatística?

12) O total de alunos de um curso é 1048. Numa turma há 45 alunos e 32 são favoráveis a certa proposta. Considere este conglomerado uma a.a.s. sem reposição e obtenha uma estimativa para o total de alunos favorável à proposta. Obtenha um limite inferior de confiança 0,99 para este total. Teste a hipótese de que este total é precisamente 800 ao n.s. 0,05.

13) Uma amostra (?) de 30 enfermeiros selecionada dos 300 que trabalham em um hospital apresentou 11 fumantes. Obtenha um intervalo de confiança 0,9 para o total de fumantes.

14) Uma amostra (?) de tamanho 60 é obtida de um condomínio com 280 unidades. Na amostra, 41 já são possuidores de algum convênio que lhes possibilita assistência médica. Obtenha um limite superior de confiança 0,99 para o total de unidades que já possui algum convênio...

15) De uma a.a.s. sem reposição de 290 domicílios de uma área urbana com 14.828 domicílios, 147 domicílios são habitados por seus proprietários e, destes, 141 têm filhos; 34 dentre os inquilinos da amostra não têm filhos. Considere a taxa de ocupação igual a 100%. Para as famílias que moram em casas alugadas, estime a percentagem das que têm filhos e calcule o erro padrão desta estimativa. Estime o total de imóveis alugados para pessoas sem filhos e calcule o erro padrão desta estimativa.

16) Repita o Reforço 15 com a informação adicional de que o total de imóveis alugados é 7526.

17) Um fabricante garante que pelo menos 95% dos equipamentos que produz atende às especificações. Um exame de 200 unidades selecionadas (?) de um lote de 1500 unidades apresentou 18 que não atendiam às especificações.

   a) Dê uma estimativa não pontual adequada com coeficiente de confiança 0,95.
   b) Teste a hipótese adequada com nível de significância 0,1.

18) Uma moeda foi lançada 100 vezes e obteve-se 75 caras. Você aceita a hipótese de que ela é honesta? Qual é um intervalo de confiança 0,9 para a probabilidade de ocorrer cara?

19) Uma a.a.s. sem reposição de tamanho 60 é obtida de um condomínio com 180 unidades. Na amostra, 41 unidades são favoráveis à construção de uma piscina. Estime o total de unidades favoráveis. Calcule o erro padrão de sua estimativa. Teste a hipótese de que pelo menos a metade do condomínio é favorável. Obtenha uma estimativa por conjunto adequada com $\gamma = 0,95$.

20) Uma a.a.s. de 1000 residências apresentou 108 com aquecimento solar. Obtenha um intervalo de confiança 0,9 para a proporção de residências com aquecimento solar na população. Teste a hipótese Ho: $p = 0,1$.

21) Uma amostra (?) de 25 gestantes apresentou 18 fumantes. Obtenha um IC-0,99 para a percentagem de gestantes fumantes e teste a hipótese deste parâmetro ser no máximo igual a 60% ao n.s. 0,01.

22) Um experimento realizado 60 vezes apresentou 18 sucessos. Estime por ponto a probabilidade de ocorrer sucesso. Obtenha um limite superior de confiança 0,95 para este parâmetro. Teste a hipótese dele ser 0,4 com $\alpha = 0,05$.

23) A osteomielite é uma infecção que freqüentemente pode ser tratada. Num experimento em que foi testado o efeito do oxigênio pressurizado em pacientes portadores desta doença, 44 dos 70 pacientes que receberam o tratamento ficaram livres da doença. Estime a percentagem de eficiência do tratamento com coeficiente de confiança 0,9. Obtenha também um limite inferior de confiança.

24) Uma amostra (?) de 27 pacientes de AIDS do sexo masculino apresentou 22 que tiveram relações homossexuais nos últimos dois anos. Estime a proporção de aidéticos homossexuais com coeficiente de confiança 0,95. Teste a hipótese de que esta proporção é no mínimo igual a 0,75 ao nível de significância 0,05.

25) Estudantes ocidentais que visitaram recentemente a China ouviram relatos de médicos chineses que afirmam que a acupuntura é eficiente em 90% dos casos em que é utilizada como anestesia em cirurgias. Numa amostra(?) de tamanho 48, somente 73% demonstraram efeitos anestésicos satisfatórios. Estas observações são inconsistentes com as afirmações dos médicos chineses?

a) Escreva a hipótese nula e a hipótese alternativa.

b) Qual é a região crítica para o nível de significância 0,05?

c) Qual é o valor crítico?

d) Qual é a decisão estatística?

e) Qual é a conclusão experimental?

f) Qual é o nível crítico amostral?

26) Para controlar a disseminação da gripe suína, foi ordenada nos E.E.U.U. em 1976 uma vacinação em massa. Logo depois surgiram numerosos casos de paralisia oriundos da Síndrome de Guillain-Barre. Difundiu-se então a notícia de que a vacina estava causando esta paralisia. O programa de vacinação foi suspenso para que fosse feita uma análise dos dados. De 220 milhões de pessoas, aproximadamente 50 milhões tinham sido vacinadas e, das 383 que tinham contraído esta paralisia, 202 haviam sido vacinadas. Você acha que existia dependência entre a vacinação e este tipo de paralisia? Responda as questões do Reforço 25.

27) Verifique que $p(1-p) \leqslant 1/4$. Em decorrência disso, Var $\hat{p} \leqslant [(N-n)/4n](N-1)$. Refaça os Exemplos 1 - 8 usando este limitante superior ao invés da estimativa $\rho(\hat{p})$. Esta técnica oferece uma vantagem e uma desvantagem. Quais são?

28) Uma urna tem 5 bolas brancas e 9 bolas amarelas. Quatro bolas são retiradas ao acaso e sem reposição. Seja T o número de bolas brancas retiradas. Escreva a distribuição de T. Calcule ET e Var T.

29) Existem três razões para a palavra "certamente" estar entre aspas nos Exemplos 1 e 2. Quais são elas?

30) Duas estatísticas são denominadas ESTATÍSTICAS EQUIVALENTES quando o valor de uma delas pode ser obtido diretamente do valor da outra sem necessidade das mensurações realizadas na amostra e vice-versa, ou seja, quando uma delas é uma função da outra e vice-versa. T e $\hat{p}$ são equivalentes? Z e $\hat{p}$ são equivalentes? $\hat{p}$er e $\hat{\tau}$ são equivalentes?

31) Tecnicamente, o que acontece nas duas situações em que as c.p.f. são omitidas? (Veja o Reforço 8).

32) Um dado é lançado 240 vezes e a face 6 ocorre 34 vezes.

a) Estime a probabilidade de ocorrência do resultado 6 num único lançamento.

b) Estime este parâmetro com coeficiente de confiança 0,9 .

c) Teste a hipótese de que o parâmetro é menor ou igual a 0,1 ao nível de significância 0,05.

d) Qual é o nível crítico amostral?

12.5 – Inferências para Duas Proporções

Consideremos dois estratos finitos de tamanhos N_1 e N_2 de uma determinada população. Consideremos uma variável X_1 definida em todo o primeiro estrato que assume apenas os valores 0 e 1 e uma variável X_2 definida em todo o segundo estrato que assume, também, apenas os valores 0 e 1. Consideremos as situações nas quais o parâmetro de interesse é $p_1 - p_2$, a diferença entre as proporções de sucessos nos estratos.

Consideremos o espaço amostral formado por todas as amostras compostas de uma a.a.s sem reposição de tamanho n_1 do primeiro estrato e de uma a.a.s sem reposição de tamanho n_2 do segundo estrato. O estimador "natural" de $p_1 - p_2$ é, sem dúvida, $\hat{p}_1 - \hat{p}_2$.

TEOREMA 4. $E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$ e $Var(\hat{p}_1 - \hat{p}_2) = Var(\hat{p}_1) + Var(\hat{p}_2)$. (Nas aplicações, esta variância é estimada por $v(\hat{p}_1 - \hat{p}_2) = v(\hat{p}_1) + v(\hat{p}_2)$.)

TEOREMA 5. A distribuição assintótica de $\hat{p}_1 - \hat{p}_2$ é Normal, i.e., quando os tamanhos das amostras são grandes, $\hat{p}_1 - \hat{p}_2$ tem distribuição aproximadamente Normal. (Se usamos a distribuição t de Student, os graus de liberdade são $n_1 + n_2 - 2$.)

TEOREMA 6. Sob a hipótese $H_0 : p_1 - p_2 = p_0$, a estatística $Z = \dfrac{\hat{p}_1 - \hat{p}_2 - p_0}{s_0(\hat{p}_{01} - \hat{p}_{02})}$ tem distribuição assintótica $N(0;1)$. O algorítmo para se obter Z é

$$\hat{p}_{01} = \frac{T_1 + T_2 + n_2 p_0}{n_1 + n_2} \qquad \hat{p}_{02} = \frac{T_1 + T_2 - n_1 p_0}{n_1 + n_2}$$

$$v_0(\hat{p}_{0i}) = \left(1 - \frac{n_i}{N_i}\right)\frac{\hat{p}_{0i}(1 - \hat{p}_{0i})}{n_i - 1} \qquad s_0(\hat{p}_{01} - \hat{p}_{02}) = \left[v_0(\hat{p}_{01}) + v_0(\hat{p}_{02})\right]^{1/2}$$

(Se usamos a t de Student, os graus de liberdade são $n_1 + n_2 - 2$.)

EXEMPLO 15 – Numa terapia que visa diminuir a ansiedade em certo tipo de pacientes, usou-se músicas de Alberto Nepomuceno e de J.S.Bach. No grupo A, 68 de 120 pacientes diminuiram significativamente sua ansiedade. No grupo B, 89 de 175. Em nenhum paciente houve aumento.

(a) Uma estimativa por ponto para $p_A - p_B$ é $\hat{p}_A - \hat{p}_B = 68/120 - 89/175 = 0{,}0581$.

(b) Uma estimativa por intervalo de confiança 0,99 é $\hat{p}_A - \hat{p}_B \;\pm\; 2{,}576 \times s(\hat{p}_A - \hat{p}_B)$, onde

$$\boxed{s(\hat{p}_A - \hat{p}_B) = [v(\hat{p}_A - \hat{p}_B)]^{1/2} = [v(\hat{p}_A) + v(\hat{p}_B)]^{1/2}}$$

$s(\hat{p}_A - \hat{p}_B) = (1 \times 68/120 \times 52/120 \times 1/119 + 1 \times 89/175 \times 86/175 \times 1/174)^{1/2} = 0{,}0592$

$0{,}0581 \pm 2{,}576 \times 0{,}0592 \;\therefore\; \boxed{-0{,}0943 \;\overset{p_A - p_B}{\underset{0{,}99}{\vdash\!\!-\!\!-\!\!-\!\!\dashv}}\; 0{,}2105}$

(c) Vamos testar a hipótese de que as músicas dos dois compositores são igualmente eficientes versus a alternativa de que não são, i.e., H_0: $p_A - p_B = 0$ vs. H_1: $p_A - p_B \neq 0$. Aqui, $p_0 = 0$.

$\hat{p}_{0A} = \hat{p}_{0B} = (T_A + T_B)/(n_A + n_B) = (68+89)/(120+175) = 157/295 = 0{,}5322$

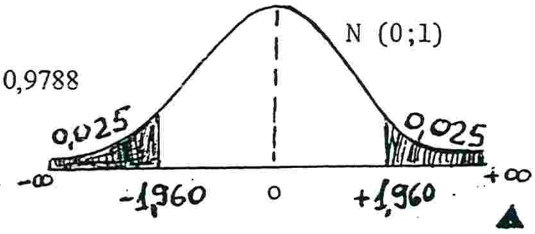
$v_0(\hat{p}_A) = 1 \times 157/295 \times 138/295 \times 1/119 = 0{,}0021$ $\quad v_0(\hat{p}_B) = 1 \times 157/295 \times 138/295 \times 1/174 = 0{,}0014$

$s_0(\hat{p}_A - \hat{p}_B) = (0{,}0021 + 0{,}0014)^{1/2} = 0{,}0594$

$Z = (\hat{p}_A - \hat{p}_B - p_0)/s_0(\hat{p}_A - \hat{p}_B) = 0{,}0581 \div 0{,}0594 = 0{,}9788$

Aceitamos H_0 ao n.s. 0,05 .

O n.c.a. é $\alpha\,(0{,}9788) \doteq 2 \times (1 - 0{,}8365) = 0{,}3270$.



▲

EXEMPLO 16 – Duas vacinas, produzidas com a finalidade de evitar certa doença em jacarés de estimação, foram eficientes, respectivamente, em 15 de 21 e em 12 de 23. Uma estimativa por intervalo para a diferença entre as eficiências das vacinas com $\gamma = 0{,}95$ é obtida assim:

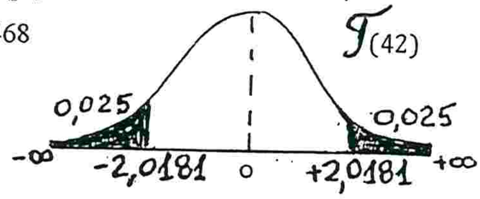
$\hat{p}_1 - \hat{p}_2 = 15/21 - 12/23 = 0{,}7143 - 0{,}5217 = 0{,}1925$

$v(\hat{p}_1) = 15/21 \times 6/21 \times 1/20 = 0{,}0102$ $\quad v(\hat{p}_2) = 12/23 \times 11/23 \times 1/22 = 0{,}0113$

$s(\hat{p}_1 - \hat{p}_2) = (0{,}0102 + 0{,}0113)^{1/2} = 0{,}1468$

$0{,}1925 \pm 2{,}0181 \times 0{,}1468$

$\boxed{-0{,}1037 \;\overset{p_1 - p_2}{\underset{0{,}95}{\vdash\!\!-\!\!-\!\!\dashv}}\; 0{,}4888}$



▲

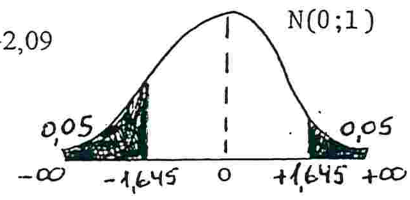
EXEMPLO 17 – Um produto para higiene bucal diminuiu efetivamente cáries em 48 de 80 crianças com dentes-de-leite e em 69 de 92 com dentição definitiva. Vamos testar a hipótese de mesma eficiência nos dois tipos de dentição com $\alpha = 0{,}1$. Aqui, H_0: $p_1 = p_2$ vs. H_1: $p_1 \neq p_2$ e $p_0 = 0$.

$\hat{p}_1 = 48/80 = 0{,}6$ $\qquad \hat{p}_2 = 69/92 = 0{,}75$ $\qquad \hat{p}_{0i} = (48+69)/(80+92) = 117/172 = 0{,}6802$

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - p_0}{\left[\dfrac{117}{172} \times \dfrac{55}{172} \times \left(\dfrac{1}{79} + \dfrac{1}{91}\right)\right]^{1/2}} = \frac{-0{,}15}{0{,}0717} = -2{,}09$$

Rejeitamos H_0 ao n.s. 0,1 .

O n.c.a. é $\alpha\,(-2{,}09) \doteq 2 \times (1 - 09821) = 0{,}0358$.



▲

EXEMPLO 18 - Uma a.a.s. sem reposição de 48 rapazes de um colégio com 1068 rapazes apresentou 44 que nunca experimentaram cigarros. Uma a.a.s. com reposição de 32 moças das 624 do colégio apresentou 25 que nunca experimentaram cigarros. Vamos inicialmente obter um intervalo de confiança 0,99 para $per_1 - per_2$:

$$\hat{per}_1 - \hat{per}_2 = 100 \times (44/48 - 25/32)\% = 91{,}67\% - 78{,}13\% = 13{,}54\%$$

$$v(\hat{per}_1) = \left(1 - \frac{48}{1068}\right) \times \frac{44}{48} \times \frac{4}{48} \times \frac{1}{47} \times (100\%)^2 = 15{,}5225\ (\%^2)$$

$$v(\hat{per}_2) = \frac{25}{32} \times \frac{7}{32} \times \frac{1}{31} \times (100\%)^2 = 55{,}1285\ (\%^2)$$

$$\rho(\hat{per}_1 - \hat{per}_2) = \left[15{,}5225 + 55{,}1285\right]^{1/2} = 8{,}4054\%$$

$$13{,}54\% \pm 2{,}576 \times 8{,}4054\% \quad \therefore \quad \boxed{-8{,}11\% \underset{0{,}99}{\overset{per_1 - per_2}{\longmapsto}} 35{,}19\%}$$

Vamos agora usar o n.s. 0,05 para testarmos $H_0: per_1 - per_2 = 0$ vs. $H_1: per_1 - per_2 > 0$.

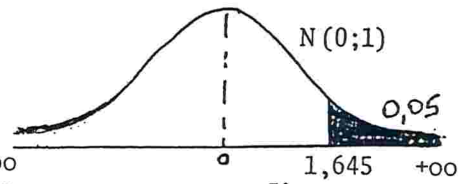
$$\hat{per}_{01} = \hat{per}_{02} = 100\% \times \left[(44 + 25)/(48 + 32)\right] = 100\% \times (69/80)$$

$$V_0(\hat{per}_1) = (100\%)^2 \times \left(1 - \frac{48}{1068}\right) \times \frac{69}{80} \times \frac{11}{80} \times \frac{1}{47} = 24{,}0987\ (\%)^2$$

$$V_0(\hat{per}_2) = (100\%)^2 \times 1 \times \frac{69}{80} \times \frac{11}{80} \times \frac{1}{31} = 38{,}2560\ (\%)^2$$

$$\rho_0(\hat{per}_1 - \hat{per}_2) = \left[24{,}0987\ (\%)^2 + 38{,}2560\ (\%)^2\right]^{1/2} = 7{,}8965\%$$

$$Z = \frac{\hat{per}_1 - \hat{per}_2 - per_0}{\rho_0\ (\hat{per}_1 - \hat{per}_2)} = \frac{13{,}54\%}{7{,}90\%} = 1{,}71$$



Rejeitamos H_0 ao n.s. 0,05.

O n.c.a. é $\alpha(1{,}71) \doteq 0{,}0436$.

Vamos agora testar $H_0: per_1 - per_2 \leq 5\%$ vs. $H_1: per_1 - per_2 > 5\%$.

$$\hat{per}_{01} = 100\% \times \left[(44 + 25 + 32 \times 0{,}05) / (48 + 32)\right] = 88{,}25\%$$

$$\hat{per}_{02} = 100\% \times \left[(44 + 25 - 48 \times 0{,}05) / (48 + 32)\right] = 83{,}25\%$$

$$V_0(\hat{per}_1) = \left(1 - \frac{48}{1068}\right) \times \frac{88{,}25 \times 11{,}75}{47} = 21{,}0709\%$$

$$V_0(\hat{per}_2) = 83{,}25 \times 16{,}75 / 31 = 44{,}9819\%$$

$$\rho_0(\hat{per}_1 - \hat{per}_2) = \left[21{,}0709 + 44{,}9819\right]^{1/2} = 8{,}1273\%$$

$$Z = \frac{\hat{per}_1 - \hat{per}_2 - per_0}{\rho_0\ (\hat{per}_1 - \hat{per}_2)} = \frac{13{,}54\% - 5\%}{8{,}13\%} = 1{,}05 \quad \therefore$$ Aceitamos H_0 ao n.s. 0,05.

O n.c.a. é $\alpha(1{,}05) \doteq 0{,}1469$.

▲

REFORÇOS (Respostas na página 89)

33) Numa pesquisa realizada com mulheres fumantes, na faixa etária dos 25 aos 33 anos, foram selecionadas (?) 50 casadas, das quais 22 eram fumantes, e 50 solteiras, das quais 35 eram fumantes. Teste H_0: $p_c - p_s \leq -0{,}05$ vs. H_1: $p_c - p_s > -0{,}05$ ao n.s. 0,1. Qual o n.c.a. ? Obtém um limite superior de confiança 0,9 para o parâmetro $p_c - p_s$.

34) Em cada um dos exemplos 15 à 18, obtém um LIC-0,9 e um LSC-0,99 para o parâmetro.

35) Com os dados do Exemplo 16, teste as hipóteses a seguir e explicite o n.c.a. .

a) H_0: $p_1 \leq p_2$ vs. H_1: $p_1 > p_2$ ao n.s. 0,05 .

b) H_0: $p_1 = p_2$ vs. H_1: $p_1 \neq p_2$ ao n.s. 0,05 .

c) H_0: $p_1 \geq p_2$ vs. H_1: $p_1 < p_2$ ao n.s. 0,05 .

36) Com os dados do Exemplo 17, teste as hipóteses a seguir e explicite o n.c.a. .

a) H_0: $p_1 - p_2 \leq -0{,}07$ vs. H_1: $p_1 - p_2 > -0{,}07$ ao n.s. 0,01.

b) H_0: $p_1 - p_2 = -0{,}07$ vs. H_1: $p_1 - p_2 \neq -0{,}07$ ao n.s. 0,01.

c) H_0: $p_1 - p_2 \geq -0{,}07$ vs. H_1: $p_1 - p_2 < -0{,}07$ ao n.s. 0,01.

d) Obtém um IC-0,9 para $p_1 - p_2$.

37) Com os dados do Exemplo 18, teste as hipóteses a seguir e explicite o n.c.a. .

a) H_0: $per_1 - per_2 = 5\%$ vs. H_1: $per_1 - per_2 \neq 5\%$ ao n.s. 0,05.

b) H_0: $per_1 - per_2 \geq 5\%$ vs. H_1: $per_1 - per_2 < 5\%$ ao n.s. 0,05.

38) Numa amostra de 21 arquitetos, 15 lêem habitualmente. Numa amostra de 23 biólogos, 12 lêem habitualmente. Obtém um LIC-0,9 para a diferença das proporções e teste a hipótese dela ser no máximo 0,1 ao n.s. 0,05. Qual o NÍVEL DESCRITIVO AMOSTRAL ?

39) Entre as últimas 28 casas de luxo construídas em Caxias do Sul, 5 tinham poços artesianos. Entre as últimas 25 casas de luxo construídas em Pelotas, 10 tinham poços artesianos.

a) Obtém um LIC-0,9 para cada uma das percentagens de casas de luxo que estão sendo construídas com poços artesianos em cada cidade.

b) Teste a hipótese de que a percentagem pelotense supera a caxiense em precisamente 10% versus a alternativa bilateral ao n.s. 0,1. Qual o n.d.a. ?

40) Num colégio há 312 rapagões e 240 moçoilas. A fração amostral 1/12 foi aplicada em cada estrato para produzir a.a.s. sem reposição. Tanto entre os rapagões selecionados como entre as moçoilas selecionadas, precisamente 6 apresentaram determinado atributo.

a) Obtém um LIC-0,95 e um IC-0,95 para per_m - per_r .

b) Teste H_0: $per_m = per_r$ vs. H_1: $per_m \neq per_r$ ao n.s. 0,05. Qual o n.d.a. ?

LIÇÃO 12

(1 a) O total de compradores em potencial deve ser menor do que 1000. (1b) ... deve ser maior do que 580. (1c) ... pode ser igual a 700. (1d) ... pode ser igual a 800. (2) $\{\hat{\tau} \leq 870\}$, $\{\hat{\tau} \geq 697\}$, $\{\hat{\tau} \leq 553\} \cup \{\hat{\tau} \geq 847\}$, $\{\hat{\tau} \leq 649\} \cup \{\hat{\tau} \geq 951\}$. (3) $\{T \geq 109\}$, $\{T \leq 78\}$, $\{T \leq 61\}$, $\{T \geq 56\}$, $\{T \leq 38\} \cup \{T \geq 67\}$, $\{T \leq 45\} \cup \{T \geq 75\}$. (4) Nenhuma . São expressões com o mesmo significado. (5 a) Região Crítica ou Região de Rejeição. (5b) Nos Exemplos 9 e 10 não há valores críticos, pois não há n.s. . (5c) 109; 78; 61; 56; 38 e 67; 45 e 75. (6 a) Não, pois usamos uma amostra e uma técnica que acerta com prob. aprox. 0,95. (6b) [1030 ; 1070] (6c) Continua não sendo, mas a confiança é maior. (7) Não é necessário, pois p = 0,5 é conhecido. (8) [0 ; 0,054] (9) Esta H_0 não pode ser testada, pois não há valores estimados de τ significativamente distantes dos seus valores da hipótese nula. (10) Impossível, pois não temos n. (12 a) 745 (12b) 577 (12c) $t_o = -0{,}8425$; $t_c = \pm 2{,}0154$; aceito. (13) [67 ; 153] (14) 226 (15 a) 76,22% e 3,54% (15b) 1738 e 277,85 (16 a) 76,22% e 3,54% (16b) 1738 e 277,85 (17 a) LSC=94,11% (17b) $t_o = -2{,}7871$; $t_c = -1{,}282$; rejeito. (18 a) $t_o = 5$; rejeito. (18b) [0,6784 ; 0,8216] (19) 123; 8,90; aceito porque $\hat{\tau} = 123$ está "mergulhado" em H_0; LIC=109 (20 a) [0,092 ; 0,124] (20b) $t_0 = 0{,}843 \leq 1 \Rightarrow$ aceito (21 a) [46,37% ; 97,63%] (21b) $t_o = 1{,}225$; $t_c = 2{,}4922$; aceito (22 a) 0,3 (22b) 0,398 (22c) $z_o = -1{,}581$; $z_c = \pm 1{,}960$; aceito (23) [53,29% ; 72,43%]; 55,40% (24) [0,658 ; 0,971]; $\hat{p}$ está "mergulhado em $H_0 \Rightarrow$ aceito (25) per = 90% vs per < 90%; $(-\infty\,;\,-1{,}645]$; -1,645; z_o 3,926 $\Rightarrow$ rejeito; A acumpultura deve ser eficiente em menos do que 90% dos casos em que é utilizada como anestesia em cirurgias; 0,0000 (26) p = 50/220 vs $p \neq 50/220$; $(-\infty\,;\,-1{,}960] \cup [+1{,}960\,;+\infty)$; - 1,960 e + 1,960; $z_o = 14{,}017 \Rightarrow$ rejeito ; A proporção de vacinados entre os doentes deve ser maior do que a proporção de vacinados na população; 0,0000 (27) [267 ; 889]; [14,96% ; 43,04%]; [1069 ; 1331]; [439 ; 701]; 1090; 680; [23,19% ; 55,38%]; [13,11% ; 65,47%]; mais segurança; menor precisão (28) p (0) = 126/1001; p (1) = 420/1001; p (2) = 360/1001; p (3) = 90/1001; p (4) = 5/1001; 10/7; 450/637 (29) Existe 0,0026 de prob. de erro. Não estão sendo usados os desvios padrões verdadeiros, mas estimativas dos mesmos. A distribuição dos estimadores não são exatamente normais, mas assintoticamente normais. (30) Sim, pois $\hat{p} = T/n$. Sim, pois $Z = (\hat{p} - p_0) / Des_0(\hat{p})$. Sim, pois $N \times \hat{p}er = 100 \times \hat{\tau}$. (31) Temos $N = +\infty$ e, assim, elas se tornam exatamente iguais a 1. (32 a) 0,142 (32b) [0,105 ; 0,179] (32c) $z_o = 2{,}1517$; $z_c = +1{,}645$; rejeito (32d) 0,0158 (33 a) $z_o = -2{,}102$; $z_c = +1{,}282$; aceito (33b) 0,9821 (33c) –0,136 (34) –0,0177 e 0,1957; 0,0014 e 0,5475; -0,2415 e 0,0161; 2,77% e 33,09% (35 a) $t_o = 1{,}2799$; $t_c = +1{,}6820$; aceito; aprox. 0,1 (35b) $t_c = \pm 2{,}0181$; aceito; aprox. 0,2 (35c) $t_c = -1{,}6820$; aceito; aprox. 0,9 (36 a) $z_o = -1{,}114$; $z_c = +2{,}326$; aceito; 0,8665 (36b) $z_c = \pm 2{,}576$; aceito; 0,2670 (36c) $z_c = -2{,}326$; aceito; 0,1335 (36d) [-0,2675 ; -0,0325] (37 a) $z_o = 1{,}051$; $z_c = \pm 1{,}960$; aceito; 0,2938 (37b) $z_c = -1{,}645$; aceito; 0,8531 (38 a) 0,0014
0,25 (38b) $t_o = 0{,}6199$; $t_c = +1{,}6820$; aceito; maior do que 0,5 (39 a) 8,17% e 26,82% (39b) $z_o = 0{,}96$ $z_c = \pm 1{,}645$; aceito; 0,3370 (40 a) –14,75%; [-19,08% ; +32,92%] (40b) $t_o = 0{,}541$; $t_c = \pm 2{,}0154$; aceito; maior do que 0,5.

LIÇÃO 13 – INFERÊNCIAS PARA MÉDIAS E VARIÂNCIAS

41 reforços

13.1 – Inferências para Uma Média

Consideremos uma população finita de tamanho N e uma variável X definida em toda a população. Vamos representar por τ o total populacional, por μ a média populacional e por σ^2 a variância populacional. Logo:

$$\tau = \sum X_i \quad , \quad \mu = \frac{\sum X_i}{N} = \frac{\tau}{N} \quad e \quad \sigma^2 = \frac{\sum (X_i - \mu)^2}{N} = \frac{\sum X_i^2}{N} - \mu^2 .$$

Consideremos o espaço amostral formado por todas as amostras aleatórias simples sem reposição de tamanho n da população. Assim, existem $N \times (N - 1) \times ... \times (N - n + 1)$ amostras equiprováveis. Para cada uma delas, sejam:

$$\bar{x} = \frac{\sum x_i}{n} \quad e \quad s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{\sum x_i^2 - (\sum x_i)^2 / n}{n-1} = \frac{\sum x_i^2 - n \cdot \bar{x}^2}{n-1} .$$

TEOREMA 1. $E(\bar{x}) = \mu$ e $Var(\bar{x}) = \frac{N-n}{N-1} \times \frac{\sigma^2}{n}$.

A média amostral $\bar{x}$ é um estimador natural para a média populacional μ. Já sabemos da importância das variâncias e dos desvios padrões dos estimadores. Var $(\bar{x})$ depende de σ^2, que depende de μ, que é desconhecido; aliás, μ é precisamente o parâmetro que desejamos estimar com o estimador $\bar{x}$. Portanto, a variância e o desvio padrão de $\bar{x}$ não podem se calculados e precisam ser estimados. Vamos fazer isso com os estimadores:

$$v(\bar{x}) = \left(1 - \frac{n}{N}\right) \times \frac{s^2}{n} \qquad e \qquad s(\bar{x}) = \sqrt{v(\bar{x})} \ .$$

TEOREMA 2. O estimador $\bar{x}$ tem distribuição assintótica Normal, isto é, quando o tamanho da amostra é grande, sua distribuição é aproximadamente Normal.

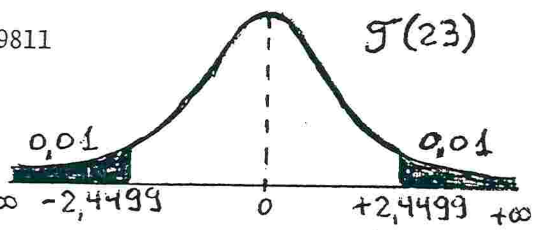
EXEMPLO 1 - Vamos obter um intervalo de confiança 0,98 para a média de uma variável que apresentou os escores a seguir em uma a.a.s. sem reposição: 58 69 60 52 65 71 59 57 61 62 63 58 70 64 63 59 67 62 44 93 25 47 31 73.

$n = 24$, $\sum x_i = 1433$ e $\sum x_i^2 = 89811$

$\bar{x} = (\sum x_i)/n = 1433/24 = 59,7$

$s^2 = \left[\sum x_i^2 - (\sum x_i)^2/n\right] / (n - 1)$

$= (89811 - 1433^2/24)/23 = 184,74$



$s(\bar{x}) = \left[(1 - n/N) s^2/n\right]^{1/2} = (184,74/24)^{1/2} = 2,77$

$\bar{x} \pm 2,4999 \times s(\bar{x}) = 59,7 \pm 2,4999 \times 2,77$ ∴

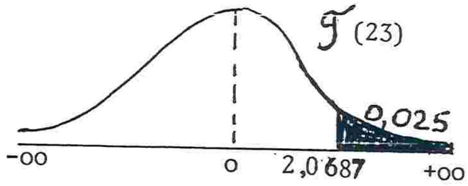
$$\boxed{52,8 \underset{0,98}{\overset{\mu}{\longmapsto}} 66,6}$$

▲

EXEMPLO 2 - Com os dados do Exemplo 1, vamos testar $H_0: \mu \leq 50$ vs $H_1: \mu > 50$ ao n.s. $\alpha = 0,025$.

$$t = \frac{\bar{x} - 50}{\rho(\bar{x})} \sim \mathcal{T}(23)$$

$= (59,7 - 50)/2,77 = 3,4992$



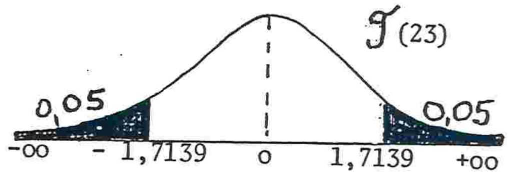
Rejeitamos H_0 ao n.s. 0,025.

$\alpha(3,4992) < \alpha(2,8073) = 0,005$.

▲

EXEMPLO 3 - Com os dados do Exemplo 1, vamos testar $H_0: \mu = 65$ vs $H_1: \mu \neq 65$ ao n.s. $\alpha = 0,1$.

$t = (59,7 - 65)/2,77 = -1,9073$



Rejeitamos H_0 ao n.s 0,1.

$0,05 < \alpha(-1,9073) < 0,1$.

▲

13.2 - Inferências para Duas Médias

Quando trabalhamos com duas variáveis tais como, por exemplo, o resultado de um teste aplicado antes de algum tratamento e o resultado do mesmo teste aplicado após o tratamento, ou tais como, por exemplo, o resultado de um teste aplicado após dois tratamentos distintos, geralmente desejamos compará-las.

Quando desejamos comparar duas variáveis, podemos fazer isso através da diferença de suas médias. Assim, se X e Y são as variáveis, o parâmetro de interesse é $\mu_x - \mu_y$. Se trabalhamos com DADOS EMPARELHADOS, tais como o primeiro exemplo do parágrafo anterior, definimos uma nova variável $D = X - Y$ e, assim, uma inferência a respeito de $\mu_x - \mu_y$ é uma inferência a respeito de uma média, $\mu_d = \mu_x - \mu_y$, que acabamos de estudar na seção anterior. Vamos considerar, a seguir, as situações que trabalhamos com DADOS INDEPENDENTES, tais como o segundo exemplo do parágrafo anterior.

Consideremos dois estratos finitos de tamanhos M e N de uma determinada população. Consideremos uma variável X definida em todo o primeiro estrato e uma variável Y definida em todo o segundo estrato. Consideremos o espaço amostral formado por todas as amostras compostas de uma a.a.s. sem reposição de tamanho m do primeiro estrato com uma a.a.s. sem reposição de tamanho n do segundo estrato. O estimador natural para o parâmetro de interesse é $\bar{x} - \bar{y}$. Aqui, não tem sentido calcularmos diferenças entre as observações de X e as observações de Y.

TEOREMA 3. $E(\bar{x} - \bar{y}) = \mu_x - \mu_y$ e $Var(\bar{x} - \bar{y}) = Var(\bar{x}) + Var(\bar{y})$.

TEOREMA 4. A distribuição assintótica de $\overline{x}-\overline{y}$ é Normal. (Se usamos a distribuição t de Student, os graus de liberdade são $m+n-2$.)

EXEMPLO 4 - Sejam $m = 25 \quad \sum x_i = 117,9 \quad \sum x_i^2 = 815,47$

$n = 12 \quad \sum y_i = 35,8 \quad \sum y_i^2 = 207,34$

$\bar{x} = 117,9/25 = 4,72 \qquad V(\bar{x}) = 1 \times s_x^2/m = 10,81/25 = 0,4324$

$\bar{y} = 35,8/12 = 2,98 \qquad V(\bar{y}) = 1 \times s_y^2/n = 9,14/12 = 0,7616$

$\bar{x} - \bar{y} = 4,72 - 2,98 = 1,73$

$s(\bar{x}-\bar{y}) = [V(\bar{x}) + V(\bar{y})]^{1/2} = [0,4324 + 0,7616]^{1/2} = 1,09$

Assim, um intervalo de confiança 0,99 para $\mu_x - \mu_y$ é $\overline{x}-\overline{y} \pm 2,7328 \times s(\overline{x}-\overline{y}) =$

$= 1,73 \pm 2,7238 \times 1,09$

$$\boxed{-1,24 \;\underset{0,99}{\overset{\mu_x - \mu_y}{\longleftrightarrow}}\; +4,71}$$

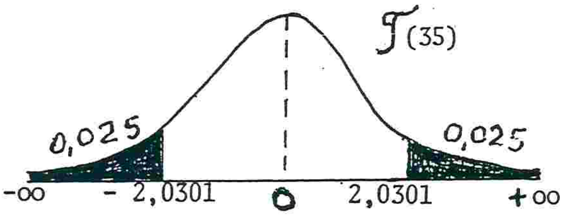
e um limite superior de confiança 0,99 para $\mu_x - \mu_y$ é $\overline{x}-\overline{y} + 2,4377 \times s(\overline{x}-\overline{y}) =$

$= 1,73 + 2,4377 \times 1,09 = 4,40.$

Para testarmos $H_0: \mu_x - \mu_y = \mu_o$ vs $H_1: \mu_x - \mu_y \neq \mu_o$, onde μ_o é um número especificado, ao n.s. 0,05, devemos usar a estatística

$t = [(\bar{x} - \bar{y}) - \mu_o] / s(\bar{x} - \bar{y})$

e a região de rejeição (ou região crítica) assinalada no desenho.



μ_o	- 2	- 1	0	1
t	3,4159	2,5008	1,5856	0,6705
decisão	rejeitamos Ho	rejeitamos Ho	aceitamos Ho	aceitamos Ho

Como é que é? Como é que é? Rejeitamos que $\mu_x - \mu_y = -2$! Rejeitamos que $\mu_x - \mu_y = -1$! Tudo bem ! Entretanto, aceitamos que $\mu_x - \mu_y = 0$ e aceitamos que $\mu_x - \mu_y = 1$??!! Sim, é exatamente isso. Veremos a explicação deste "intrigante enigma" na Lição 16. ▲

13.3 - Inferências para Uma Variância

Consideremos uma população não necessariamente finita. Consideremos uma variável X definida em toda a população, mas cuja distribuição verdadeira é Normal. Consideremos o espaço amostral formado por todas as a.a.s. com reposição de tamanho n.

TEOREMA 5. $E(s^2) = \sigma^2$ e $Var(s^2) = 2 \times \sigma^4/(n-1)$.

TEOREMA 6. A distribuição assintótica de s^2 é Normal. A distribuição exata de $Q=(n-1)\times s^2/\sigma^2$ é Qui-quadrado com $n-1$ graus de liberdade.

EXEMPLO 5 - O salário médio semanal de uma amostra (?) de 30 empregados de uma empresa de construção civil é Cr\$ 1.800,00 e o desvio padrão desta amostra é Cr\$ 120,00. Vamos testar H_0: $\sigma \leq$ Cr\$ 100,00 vs H_1: $\sigma >$ Cr\$ 100,00.

$$q = (n-1)s^2/\sigma_0^2 = 29 \times 120^2/100^2 = 41{,}76$$

α	0,1	0,05	0,025	0,01
q_c	39,087	42,557	45,722	49,588
decisão	rejeitamos Ho	aceitamos Ho	aceitamos Ho	aceitamos Ho

Aqui nossa decisão depende do nível de significância (arbitrariamente) escolhido. Isto certamente não nos faz ficar muito contentes. O que nos deixaria contentes? Um valor obtido para Q "bem próximo" de 29, e assim aceitaríamos "tranqüilamente" H_0, ou um valor obtido para Q "bem próximo" de + oo (maior do que 52,336), e assim rejeitaríamos "tranqüilamente" H_0.

Um intervalo de confiança 0,95 para σ é

$$16{,}047 \leq \frac{29 \times 120^2}{\sigma^2} \leq 45{,}722 \quad \therefore \quad \boxed{\text{Cr\$ } 95{,}57 \;\underset{0,95}{\overset{\sigma}{\longmapsto}}\; \text{Cr\$ } 161{,}32.}$$

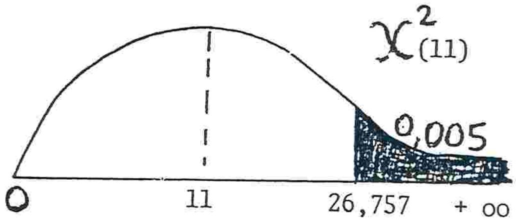
▲

EXEMPLO 6 - Desejamos verificar se uma máquina de envasar, que supostamente está regulada para envasar 150 gramas de ervilha por lata, está efetivamente regulada, o que ocorre quando o COEFICIENTE DE VARIAÇÃO é no máximo 0,02. Uma amostra (?) de 12 latas apresentou pesos x_i tais que $\sum x_i = 1916{,}4$ gr e $\sum x_i^2 = 307072{,}73$ gr^2.

$$\text{Coef.Var.}(X) = \sigma/\mu = \sigma/150 \leq 0{,}02 \iff \sigma \leq 3 \text{ gr}$$

$$\begin{cases} H_0: & \sigma \leq 3 \\ H_1: & \sigma > 3 \end{cases} \qquad q = \frac{(n-1)s^2}{\sigma_0^2} = \frac{\sum x_i^2 - (\sum x_i)^2/n}{\sigma_0^2} = 113{,}739$$

Aqui rejeitamos "$\sigma \leq 3$" tranqüilamente mesmo. Devemos parar a produção e regular a máquina de envasar, sob o risco de que ela esteja envasando muito pouco em bastante latas e purê em outras tantas. Se, entretanto, esta decisão for incorreta, a interrupção do envasamento e a regulagem serão realizadas desnecessariamente.



Vamos obter agora uma estimativa pontual para σ :

$$\hat{\sigma} = s = \left[\frac{\sum x_i^2 - (\sum x_i)^2/n}{n-1}\right]^{1/2} = (93{,}0591)^{1/2} = 9{,}65 \text{ gr.}$$

Como já "constatamos" acima, 9,65 >> 3.

Vamos agora obter um limite inferior de confiança 0,95 para σ e um de confiança 0,99.

$$P\left\{\frac{(n-1)s^2}{\sigma^2} \leqslant 19{,}675\right\} = 0{,}95 \therefore \frac{1023{,}65}{\sigma^2} \leqslant 19{,}675 \therefore \sigma \geqslant 7{,}21 \text{ gr.}$$

$$P\left\{\frac{(n-1)s^2}{\sigma^2} \leqslant 24{,}725\right\} = 0{,}99 \therefore \frac{1023{,}65}{\sigma^2} \leqslant 24{,}725 \therefore \sigma \geqslant 6{,}43 \text{ gr.}$$

Vamos novamente observar que maior confiança implicou maior imprecisão. Podemos colocar nossas inferências como "$\sigma \geqslant 7{,}21$ gr" $\wedge$ "$\sigma \geqslant 6{,}43$ gr."? ▲

REFORÇOS (Respostas na página 108)

1) Responda à pergunta no final do Exemplo 6.

2) Conceitue erro padrão de uma estimativa.

3) Var$(\bar{x})$ é um parâmetro?

4) Uma a.a.s. de tamanho 37 das lâmpadas produzidas por determinada indústria foi testada quanto à sua duração. Os resultados obtidos foram uma média de 501,8 horas e um desvio padrão de 3,0 horas. É viável crer, a 1% de significância, que a duração média de todas as lâmpadas produzidas pela referida indústria supera 500 horas?

5) No Reforço 4 estime a probabilidade de que uma lâmpada dure mais do que 500 horas.

6) Uma população de 8 pessoas tem duas pessoas canhotas. Seja Y a quantidade de pessoas canhotas em uma a.a.s. sem reposição de tamanho 4 desta população. Uma a.a.s. de uma v.a. X com distribuição normal apresentou valores (aproximados para) 20 21 22 23 24 25 26 22 24 23. Encontre uma estimativa para o valor x tal que $P\{X > x\} = P\{Y = 2\}$.

7) Comparando-se as médias dos pesos perdidos por pessoas submetidas a duas dietas, deseja-se testar a hipótese de igualdade ao n.s. 0,01. Qual é a região crítica para uma alternativa bilateral? Qual é o n.c.a. para m = n = 40, $\bar{x} = 10$, $\bar{y} = 8$, $s_x^2 = 4{,}3$ e $s_y^2 = 5{,}7$?

8) Por que usamos um limite inferior de confiança no Exemplo 6?

9) Durante 309 dias úteis, sorteados ao acaso, verificou-se que houve 1328 acidentes de trabalho no setor A de uma indústria. Durante 150 dias, sorteados ao acaso, verificou-se que houve 684 acidentes de trabalho no setor B. Suponhamos que X, o número de acidentes de trabalho por dia no setor A, e que Y, o número de acidentes de trabalho por dia no setor B, tenham distribuição de Poisson. Desta forma, $EX = Var = \lambda_A$ e $EY = Var\ Y = \lambda_B$. Portanto, $\bar{x}$ é um estimador "natural" tanto da média de X como de sua variância e o mesmo ocorre com $\bar{y}$ em relação a Y. Estime $\lambda_A - \lambda_B$ pontualmente e por intervalo com $\gamma = 0,99$. Teste a hipótese de que não há diferença entre a incidência de acidentes nos dois setores.

10) Uma a.a.s. sem reposição de 20 domicílios, selecionados em uma zona rural de 10.000 domicílios, apresentou os números a seguir de pessoas por domicílio: 6 4 8 10 5 3 2 5 7 6 9 5 4 6 7 8 6 5 3 5. Estime o total de habitantes por ponto e por intervalo com $\gamma = 0,95$. Teste a hipótese do número total de habitantes ser maior ou igual a 50.000.

11) Uma a.a.s. sem reposição de um grupo de 102 alunos do Curso A apresentou as idades a seguir. Obtenha um intervalo de confiança 0,95 para a idade média e um para a variância de todos os alunos do grupo.

x_i	17	18	19	20	21
m_i	2	8	6	5	3

12) Para testar se a vitória de sua equipe no futebol aos domingos influencia a eficiência profissional às segundas, 10 operários foram selecionados para serem observados nas duas situações. Os escores obtidos (em uma escala que mede eficiência profissional) estão a seguir. Que conclusão podemos obter ao nível de significância 0,025?

operário	1	2	3	4	5	6	7	8	9	10
x_i (vitória)	12	16	15	13	16	10	15	17	14	12
y_i (derrota)	10	17	18	16	19	12	17	15	17	14

13) Uma a.a.s. sem reposição de um grupo de 368 alunos do Curso B apresentou as idades a seguir.

a) Teste a hipótese de que a média das idades deste grupo é igual a média das idades do grupo do Reforço 11.

b) Teste a hipótese de que as proporções de alunos com menos de 18 anos são iguais.

y_j	17	18	19	20	21
m_j	2	12	20	12	2

14) Alguns pesquisadores pensam que a Vitamina C pode ser útil para reduzir os depósitos formadores de colesterol situados na parte inferior das paredes arteriais e, portanto, pensam que a Vitamina C contribui também para a redução da possibilidade de ataques cardíacos. Os níveis de colesterol de cada uma das 49 pessoas que os possuíam acima do normal foram anotados, antes e depois de uma dieta na qual cada pessoa ingeriu diariamente 500 mg de Vitamina C. Os dados obtidos para a queda do nível de colesterol apresentaram média 643 mg/l e desvio padrão 189 mg/l. Com $\gamma = 0,95$ estime a média por intervalo, limite inferior e posterior.

15) Realizou-se uma experiência a fim de comparar dois tipos de dieta, A e B, destinadas a redução de peso. Dois grupos, cada um com 30 pessoas gordas, fizeram cada um,uma das dietas. Os dados observados para a queda de peso foram $\bar{y}_A = 21,3$ $\rho_A = 2,6$ $\bar{y}_B = 13,4$ $\rho_B = 1,9$. A designação dos sujeitos para cada uma das dietas foi aleatória. Estime $\mu_A - \mu_B$ por intervalo, por limite inferior e por limite superior com o coeficiente de confiança 0,99.

16) Realizou-se um experimento para determinar o efeito do fumo sobre a pressão sangüínea em uma amostra de 25 universitários. Mediu-se a pressão de cada um por ocasião do término do curso e 5 anos depois. Registrando-se a diferença em mm de mercúrio obteve-se a média 9,7 e o desvio padrão 5,8. Descreva a população. Estime o aumento médio de pressão para o período de 5 anos com $\gamma = 0,9$.

17) O número de períodos de chuva por semana em determinada região foi observado durante 80 semanas consecutivas. (Isto pode ser considerada uma boa amostragem?) A soma das observações foi 336 e a soma de seus quadrados foi 1764. Usando $\gamma = 0,95$:

a) Obtenha um limite inferior de confiança para μ.
b) Estime o total de períodos de chuva em 3 anos.

18) Para testar H_0: $\sigma^2 = 16$ vs H_1: $\sigma^2 \neq 16$, comente sobre a adequabilidade ou não de cada um destes testes: $\{\rho \neq 4\}$, $\{\rho < 3\}$ e $\{\rho < 2,5\} \cup \{\rho > 6,5\}$.

19) Os valores a seguir são a queda da capacidade de raciocínio de indivíduos sob o efeito de maconha: 2,6 3,1 1,8 3,0 1,9 2,7 2,3 2,5 2,4 2,2 2,6. Obtenha um intervalo de confiança 0,95 para a média. Teste H_0: $\mu \leq 2,2$ vs H_1: $\mu > 2,2$ ao n.s. 0,05. Qual é o nível crítico amostral?

20) No Exemplo 5, o que significa "bem próximo de 29" e o que significa "bem próximo de + ∞"?

21) Uma a.a.s. de tamanho 46 apresentou $\sum x_i = 562{,}4$ e $\sum x_i^2 = 11364{,}28$. O total de elementos da população é 418. Obtenha um limite inferior de confiança para o total da variável na população com $\gamma = 0{,}95$ e teste a hipótese deste parâmetro ser no mínimo igual a 4.000 unidades ao n.s. 0,05.

22) A decisão do Exemplo 6 pode estar errada?

23) O que aconteceria se tivéssemos aceito Ho erradamente no Exemplo 6?

24) A tabela ao lado contém a distribuição amostral do número de empregados de 61 escritórios de um edifício com 384 escritórios.

i	Ci	f_i
1	0 ⊢ 3	6
2	3 ⊢ 6	13
3	6 ⊢ 9	21
4	9 ⊢ 12	18
5	12 ⊢ 15	3

a) Estime o total de empregados com $\gamma = 0{,}95$.

b) Teste a hipótese de que os escritórios pequenos ($X \leq 5$) não superam 25% da população com $\alpha = 0{,}05$.

c) Teste a hipótese de que os escritórios grandes ($X \geq 9$) superam 25% da população, também com o nível de significância 0,05.

25) Repita os Exemplos 5 e 6 usando a distribuição assintótica de ρ^2 ao invés da exata.

26) No Reforço 21, obtenha um intervalo de confiança 0,9 para σ^2. Teste $H_0: \sigma^2 \leq 81$ vs $H_1: \sigma^2 > 81$ ao n.s. 0,1. Qual é o nível crítico amostral?

27) No Reforço 17, teste a hipótese de que a variância é no máximo 9 com $\alpha = 0{,}05$.

28) No Reforço 24, obtenha um limite inferior, um limite superior e um intervalo, todos de confiança 0,9 para σ^2 e para σ.

29) No Reforço 19, teste $H_0: \sigma = 1$ vs $H_1: \sigma \neq 1$ e $K_0: \sigma \geq 1$ vs $H_1: \sigma < 1$ com $\alpha = 0{,}05$. Escreva as regiões críticas.

30) Uma amostra de "pacientes" da gripe X e outra de "pacientes" da gripe Y apresentaram os tempos de incubação do vírus listados a seguir. Estime $\mu_X - \mu_Y$ com $\gamma = 0{,}95$; teste a hipótese "$\sigma_X \leq 2$" e obtenha um limite inferior de confiança 0,9 para σ_X.

x (dias) :	7	2	15	18	19	6	13	10	8	22	16	18
y (dias) :	9	6	9	20	8	3	8	23	5	20	7	4

31) Os escores de eficiência em um hospital para enfermeiros e enfermeiras sorteados ao acaso e avaliados, estão listados a seguir. Teste a hipótese de que os dois grupos "têm a mesma eficiência" com $\alpha = 0{,}1$.

enfermeiros: 74,4 53,9 32,5 52,8 19,2 59,6 39,0 44,9 74,2

enfermeiras: 69,6 74,0 68,4 65,0 38,4 65,8 68,4 52,8 72,8

Qual é o n.c.a.? Estime os desvios padrões de cada grupo. A partir destas estimativas, e sem considerar a significância de sua diferença, ou melhor, de sua razão (porque médias são comparadas através de diferenças e variâncias o são através de razões), uma vez que não estudamos nenhuma técnica de inferência para isso, que conclusão pode ser obtida?

32) Teste $H_0: \mu \leq 50$ e $K_0: \sigma = 50$ ao n.s. 0,1 com os dados a seguir, explicitando as regiões de rejeição.

22,0 235,2 45,1 37,7 156,1 25,7 132,9 11,1 87,6 52,5

33) Uma a.a.s. de tamanho 25 obtida de uma variável distribuída normalmente em toda a população apresentou $\sum x_i = 125{,}36$ e $\sum x_i^2 = 716{,}2474$. Teste a hipótese $\sigma^2 \geq 7$ ao n.s. 0,025. Explicite a região crítica. Nesta mesma pesquisa, outra variável era do tipo sucesso-fracasso e apresentou 18 sucessos. Teste a hipótese $p \geq 0{,}6$ com o mesmo n.s. e com correção de continuidade. Explicite a região crítica.

não

34) Oito veículos sorteados ao acaso percorreram um trajeto padrão equipados com pneus radiais (X) e com pneus comuns (Y). O rendimento do combustível (em km/ℓ) está tabulado a seguir. Você diria que o tipo dos pneus influencia o rendimento? Teste $H_0: \mu_x \geq 18$ vs $H_1: \mu_x < 18$ e $K_0: \sigma_y \leq 4$ vs $K_1: \sigma_y > 4$.

veículos	1	2	3	4	5	6	7	8
x	26,5	14,3	12,7	20,2	15,1	16,9	23,4	16,4
y	25,8	14,5	12,1	19,9	15,1	15,8	23,0	16,0

Para responder a pergunta acima, use as diferenças "cruas" e as diferenças percentuais.

35) Uma a.a.s. de 25 domicílios teve sua área medida, o que resultou $\sum x_i = 1253{,}6\ m^2$ e $\sum x_i^2 = 71624{,}74\ m^4$.

a) Estime a média e o desvio padrão.

b) Ache um limite superior de confiança 0,99 para cada um.

c) Ao n.s. 0,01 teste $H_0: \mu = 40 m^2$ e $K_0: \sigma = 10\ m^2$ contra as respectivas alternativas bilaterais.

36) Uma a.a.s. sem reposição de 27 pessoas teve sua agressividade medida antes e depois de serem submetidas a uma situação de ansiedade tal como,por exemplo, uma prova de Estatística. Os escores resultantes foram tais que $\sum x_i = 1215$ e $\sum x_i^2 = 58674$ no pré-teste e $\sum y_i = 972$ e $\sum y_i^2 = 37999$ no pós-teste.

a) Estime por intervalos de confiança 0,99 os escores médios de agressividade da população antes e depois.

b) Teste a hipótese de que esta situação de ansiedade não aumenta nem diminui os escores médios de agressividade da população ao n.s. 0,01.

c) Obtenha limites superiores de confiança 0,99 para os desvios padrões dos escores de agressividade na população antes e depois.

37) Uma a.a.s. de tamanho 28 apresentou $\sum x_i = 151,76$ e $\sum x_i^2 = 872,4784$.

a) Obtenha um limite inferior de confiança 0,99 e um intervalo de confiança 0,95 para o desvio padrão.

b) Ao n.s. 0,01 teste a hipótese do desvio padrão ser maior ou igual a 1 e ao n.s. 0,05 teste a hipótese dele ser igual a 1.

38) Com os dados do Reforço 32, obtenha intervalos de confiança 0,9 para os parâmetros μ e σ, assim como limites inferiores e superiores de confiança 0,9.

39) O desvio padrão de uma a.a.s. com reposição de uma variável com distribuição supostamente normal pode ser calculado com os dados: $n = 15$, $\sum x_i = 18.600$ horas, $\sum x_i^2 = 23148375$ horas^2.

a) Obtenha um IC - 0,95 para σ.

b) Teste a hipótese $\sigma = 70$ horas ao n.s. 0,05.

40) Uma a.a.s. de tamanho 46 de uma variável apresentou $\sum x_i = 562,4$ e $\sum x_i^2 = 11.364,28$. Teste H_0: $\mu = 10$ vs H_1: $\mu \neq 10$ e K_0: $\sigma \leq 6$ vs K_1 : $\sigma > 6$ com $\alpha = 0,05$. Em ambos os casos, escreva ou represente graficamente as regiões críticas.

41) Um grupo com 50 pessoas apresentou uma média de 7,82 horas com desvio padrão 0,24 horas. Outro grupo, com 100 pessoas apresentou uma média de 6,75 horas com desvio padrão 0,30 horas. Considere cada um destes grupos a.a.s. das respectivas populações.

a) Dá uma estimativa por intervalo para a diferença de médias com $\gamma = 0,99$.

b) Teste a hipótese de que não há diferenças entre as médias com $\alpha = 0,05$.

14.1 – Testes de Ajustamento

Consideremos um espaço de probabilidade (Ω, P) e uma PARTIÇÃO PROBABILÍSTICA $\{A_1,...,A_k\}$, i.e.,

(1) $A_1,...,A_k$ são eventos exclusivos

(2) $A_1 \cup ... \cup A_k = \Omega$

(3) $P(A_i) > 0$, para todo i $\in$ {1, ..., k}.

Consideremos (não na prática, mas na nossa imaginação) uma seqüência infinita de realizações independentes de um experimento aleatório com espaço de probabilidade (Ω, P). Para cada $n \geq 1$ e cada i $\in$ {1, ..., k}, seja $O_n(i)$ o número de vezes que ocorreu o evento A_i nas n primeiras realizações. Para cada $n \geq 1$, a distribuição do vetor aleatório $(O_n(1),...,O_n(k))$ é denominada DISTRIBUIÇÃO MULTINOMIAL $(n;p_1;...;p_k)$, onde $p_i = P(A_i)$, para todo i $\in$ {1, ..., k}. É obvio que

(4) $$\sum_{i=1}^{k} O_n(i) = n, \text{ para todo } n \geq 1.$$

EXEMPLO 1 – Um dado honesto é lançado infinitas vezes. Temos $\Omega = \{1,2,3,4,5,6\}$ e $P\{\omega\} = 1/6$, para todo $\omega \in \Omega$. Sejam $A_1 = \{1\}$, $A_2 = \{2,3\}$ e $A_3 = \{4,5,6\}$. Temos k = 3, $p_1 = 1/6$, $p_2 = 2/6$ e $p_3 = 3/6$. ▲

Para cada $n \geq 1$ e cada i $\in$ {1, ..., k}, a estatística $O_n(i)$ tem DISTRIBUIÇÃO BINOMIAL $(n;p_i)$. Portanto, $E[O_n(i)] = n \times p_i$.

Se temos uma distribuição Multinomial com $p_1,...,p_k$ desconhecidos e nosso interesse é testar a hipótese simples $H_0 : p_1 = p_{01},...,p_k = p_{0k}$, onde $p_{01},...,p_{0k}$ são números especificados, então fazemos $E_n(i) = E_0[O_n(i)] = n \times p_{0i}$, para todo i $\in$ {1, ..., k} . Assim, H_0 é rejeitada se, e somente se, for significativamente grande o valor da estatística

(5) $$T_n = \sum_{i=1}^{k} \frac{[O_n(i) - E_n(i)]^2}{E_n(i)} = \left[\sum_{i=1}^{n} \frac{O_n^2(i)}{E_n(i)}\right] - n,$$

cuja distribuição nula assintótica é Qui-quadrado com k-1 graus de liberdade. Consideraremos esta aproximação satisfatória quando estiver satisfeita a condição

(6) $E_n(i) \geq 5$, para todo i $\in$ {1, ..., k}.

EXEMPLO 2 – Se, na prática o dado do Exemplo 1 é lançado "apenas" 240 vezes, então $H_0 : p_1 = 1/6$, $p_2 = 2/6$ e $p_3 = 3/6$ apresenta E (1) = 40, E (2) = 80 e E (3) = 120 , devendo ser rejeitada ao n.s.a. 0,01 se, e somente se, $T \geq 9{,}210$. ▲

EXEMPLO 3 - No estudo do temperamento das pessoas que escolhem determinada profissão, observou-se que há predominantemente 4 tipos. Estudos realizados há 10 anos atrás mostram que os temperamentos codificados como A, B, C, D estavam nas proporções 4 : 12 : 5 : 4. Uma amostra atual de 77 destes profissionais apresentou, respectivamente, 18, 36, 13, 10 em cada grupo.

i	p_{oi}	E (i)	O (i)	O^2/E
1	0,16	12,32	18	26,30
2	0,48	36,96	36	35,06
3	0,20	15,40	13	10,97
4	0,16	12,32	10	8,12
	1	77	77	80,455

Como T_{77} = 80,455 - 77 = = 3,455 é próximo de 3 que é o valor esperado sob Ho, aceitamos a hipótese. Para rejeitá-la precisaríamos pelo menos $T_{77} \geqslant 6,251$. ▲

14.2 – Tabelas de Contingência

Consideremos uma população cujos elementos estão classificados segundo duas partições probabilísticas, uma com "l" categorias e outra com "c" categorias. A população fica particionada em $l \cdot c$ categorias que satisfazem a (1) e a (2), mas podem não satisfazer a (3).

Desejamos testar a hipótese composta de que as duas partições probabilísticas são independentes, i.e., de que a informação de que um elemento da população pertence a uma determinada categoria de uma delas não modifica a probabilidade dele pertencer a determinada categoria da outra, quaisquer que sejam estas categorias. Vamos ilustrar isto.

	B_1	B_2	B_3
A_1	60	42	18
A_2	30	21	9

Independência

	B_1	B_2	B_3
A_1	75	27	18
A_2	15	36	9

Dependência

Se conhecemos toda a população, não é necessário testarmos a hipótese. Apenas verificamos se ela é verdadeira ou falsa. Portanto, estamos interessados somente em tabelas de contingência amostrais. A hipótese de independência é rejeitada se, e somente se, a dependência encontrada na amostra for significativamente grande. Para verificarmos isso, fazemos

$$\hat{E}_n(i,j) = n \times \hat{P}_0(i,j) = n \times \hat{P}(A_i) \times \hat{P}(B_j) = n \times \frac{O_n(A_i)}{n} \times \frac{O_n(B_j)}{n} = \frac{O_n(A_i) \times O_n(B_j)}{n}$$

e usamos a estatística

(7) $$T_n = \sum_{i=1}^{l} \sum_{i=1}^{c} \frac{\left[O_n(i,j) - \hat{E}_n(i,j)\right]^2}{\hat{E}_n(i,j)} = \left[\sum_{i=1}^{l} \sum_{i=1}^{c} \frac{O_n^2(i,j)}{\hat{E}_n(i,j)}\right]^2 - n,$$

cuja distribuição nula assintótica é Qui-quadrado com $(l-1) \times (c-1)$ graus de liberdade. Consideraremos esta aproximação satisfatória quando estiver satisfeita a condição

(8) $$\hat{E}_n(i,j) \geq 5, \quad \text{para todo } i \in \{1,\ldots,l\} \text{ e para todo } j \in \{1,\ldots,c\}.$$

EXEMPLO 4 - Desejamos verificar se existe alguma relação entre aptidão mental e aptidão manual em certos indivíduos. Uma amostra (?) destes indivíduos apresentou a tabela de contingência a seguir, onde os valores esperados estão colocados entre parênteses ao lado dos valores obtidos (o que é bastante prático e comum).

Mental \ Manual	Boa	Regular	Fraca	
Boa	56 (54)	80 (63)	44 (63)	180
Regular	15 (27)	45 (31,5)	30 (31,5)	90
Fraca	49 (39)	15 (45,5)	66 (45,5)	130
	120	140	140	400

$$T = \sum_{i=1}^{3} \sum_{j=1}^{3} \frac{O_m^2(i,j)}{\hat{E}_m(i,j)} - m = \left[\frac{56^2}{54} + \frac{80^2}{63} + \ldots + \frac{66^2}{45,5}\right] - 400 = 53,827$$

Aqui $(\ell - 1)\times(c - 1) = (3 - 1)\times(3 - 1) = 4$ e, como $P\{T \geq 14,860\} = 0,005$, a hipótese nula de independência deve ser rejeitada, ou seja, deve haver uma dependência (de que tipo?) entre as aptidões manual e mental dos indivíduos (da população em questão, é óbvio). Esta amostra sugere uma dependência do seguinte tipo (sugere, note bem !):

Mental \ Manual	Boa	Regular	Fraca
Boa	=	+	−
Regular	−	+	=
Fraca	+	−	+

▲

A hipótese nula de independência parece ser, à primeira vista, "simples", isto é, conter uma única idéia, mas a hipótese alternativa de dependência parece "composta", mesmo à primeira vista. Ambas são compostas.

14.3 – Testes de Homogeneidade

Consideremos uma população cujos elementos estejam classificados segundo um único critério que a particiona em c categorias e que também esteja particionada em k estratos.

Desejamos testar a hipótese composta de que os estratos são homogêneos, isto é, de que a partição probabilística seja a mesma em todos os estratos. Para isso realizamos uma a.a.s. com reposição de m_1 elementos no primeiro estrato, ..., uma a.a.s. com reposição de m_k elementos no k estrato. Os resultados desta amostragem podem ser dispostos numa tabela semelhante a uma tabela de contingência (o que é bastante prático e comum), mas que não é uma tabela de contingência. A partir daí, tudo o que se segue é semelhante, embora nossos propósitos e o planejamento do experimento sejam diferentes. Portanto, a distribuição nula assintótica da estatística T_n é qui-quadrado com $(k - 1)(c - 1)$ graus de liberdade.

EXEMPLO 5 - Numa disciplina de determinada universidade, um grupo de 100 alunos faz um curso assistindo aulas regulares e outro grupo de 100 alunos faz o mesmo curso em casa através de fichas de estudo. Os resultados são:

	A	B	C	D	E	
Estrato 1 (Regulares)	15 (12)	25 (21,5)	32 (30,5)	17 (22,5)	11(13,5)	100
Estrato 2 (Fichas)	9 (12)	18 (21,5)	29 (30,5)	28 (22,5)	16(13,5)	100
	24	43	61	45	27	200

$$T = \sum_{i=1}^{k} \sum_{j=1}^{c} \frac{O_n^2(i,j)}{\hat{E}_n(i,j)} - (n_1 + \dots + n_k) = \left[\frac{15^2}{12} + \frac{25^2}{21,5} + \dots + \frac{16^2}{13,5}\right] - 200 = 6,402$$

Portanto, aceitamos Ho. (Para rejeitarmos, só se $T \geqslant 7,779$). ▲

REFORÇOS (Respostas na página 136)

1) Arbitre valores para $O_{240}(1)$, $O_{240}(2)$ e $O_{240}(3)$ no Exemplo 2 e teste a hipótese. Qual é o n.c.a.a.?

2) Prove que $\displaystyle\sum_{i=1}^{k} \frac{[O_n(i) - E_n(i)]^2}{E_n(i)} = \left[\sum_{i=1}^{k} \frac{O_n^2(i)}{E_n(i)}\right] - n.$

3) Verifique, em cada um dos exemplos desta lição, se $E_n(i) \geqslant 5$ para todo $i \in \{1, \dots, k\}$.

4) Estime as proporções de cada temperamento no Exemplo 3. No seu entender, suas estimativas são muito diferentes dos valores atribuídos pela hipótese a tais proporções? Por que calcular o valor da estatística?

5) Para aplicarmos os 3 testes qui-quadrado desta lição devemos usar qual tipo de amostragem?

6) Uma amostra de 500 pacientes que sofrem determinada enfermidade apresentou 290 do grupo 0, 83 do grupo A, 102 do grupo B e 25 do grupo AB. Você aceita a hipótese "$p_O = 1/2$, $p_A = p_B = 1/5$, $p_{AB} = 1/10$" ao nível de significância assintótico 0,01? Qual é o nível crítico amostral assintótico?

7) Teste a hipótese de independência ao n.s.a. 0,025. Qual é o n.c.a.a.? Em caso de rejeição da hipótese nula, qual é o tipo de dependência sugerida pela amostra?

	Caprichosos	Regulares	Desleixados
Aprovados	56	72	16
Reprovados	14	28	14

8) Os dados a seguir classificam casos de tumores no fígado x uso de anticoncepcionais por via oral. Que conclusão você pode obter com $\alpha = 0,05$? Qual é a decisão estatística? O que a amostra sugere?

	Usuárias	Já usaram	Nunca usaram
Benigno	128	72	46
Maligno	69	88	97

9) Nos Testes de Homogeneidade, os tamanhos dos estratos são importantes?

10) Alguma das técnicas desta lição serve para testar diferença de proporções?

11) Uma a.a.s. com reposição de 200 universitários apresentou 51 casados. Teste a hipótese de que a percentagem de universitários casados é 20% com o n.s.a. 0,05. Qual é o n.c.a.a.?

12) É razoável supormos que os acidentes de trânsito em determinada cidade não apresentam "preferência" por certos dias da semana ao n.s.a. 0,1? Dados: 48 (DOM), 66 (SEG), 54 (TER), 55, 52, 71, 39 (SÁB).

13) Suponhamos que, na situação do Reforço 12, os dados tenham sido misturados, isto é, temos os números 39, 48, 52, 54, 55, 66 e 71, mas não sabemos a quais dias da semana eles correspondem. A hipótese poderia, ainda assim, ser testada? E se este "rodízio" tivesse sido feito no Reforço 6?

14) Qual é a principal desvantagem dos Testes Qui-Quadrado?

15) Considerando os dados do Reforço 12, teste a hipótese de que nos fins-de-semana ocorrem mais acidentes (naquela determinada cidade) do que nos outros dias.

16) Construída a distribuição de freqüências de uma a.a.s. com reposição de uma variável, verificamos graficamente que há uma semelhança com a distribuição Poisson de média 3. Utilizamos 7 classes e encontramos o valor 15,314 para a estatística adequada para testar esta impressão visual. Qual é a decisão ao n.s.a. 0,05? E ao n.s.a. 0,01?

17) Um grupo de 600 pessoas que possuem determinada enfermidade é separado ao acaso em dois grupos com uma tabela de dígitos aleatórios. Os dois grupos são tratados de modo idêntico quanto à alimentação, recreação, horas de repouso etc. Entretanto, o grupo A recebe determinado tratamento terapêutico e o grupo B não. Deseja-se testar a hipótese de que a terapia auxilia na cura desta enfermidade. Podemos usar um teste χ^2? Em caso afirmativo, diga qual e use-o para os dados: 205 dos 317 pacientes do grupo B curaram-se, o que foi conseguido por 218 do grupo A.

18) Uma a.a.s. com reposição de 200 universitários apresentou os dados a seguir (rendimento acadêmico x interesse por futebol). Que conclusão podemos obter?

	Grêmio	Inter	Nenhum	Sem interesse
Bom	38	20	1	1
Regular	60	21	4	5
Fraco	12	29	3	6

19) Teste a independência entre nível de maconha x comportamento anti-social predominante ao n.s.a. 0,05. Há um n.s.a. que faria você mudar sua conclusão? Qual é o nível crítico amostral assintótico?

Consumo	Agressivo	Desligado	Nada
Exagerado	10	11	10
Médio	5	7	6
Eventual	7	15	8

20) A distribuição das rendas (expressas em alguma unidade monetária) familiares de certo bairro está tabelada ao lado. Você admite uma distribuição uniforme para estes dados?

A	O
450 ⊢ 600	57
600 ⊢ 800	65
800 ⊢ 1000	76
1000 ⊢ 1200	54
1200 ⊢ 1350	73

21) Você admitiria uma distribuição normal para os dados do Reforço 20?

22) Um professor obteve dos seus alunos no semestre passado a distribuição de conceitos a seguir. Através de um teste χ^2, qual é a conclusão experimental?

	A	B	C	D	E
Psicologia	2	9	11	13	11
Engenharia	0	7	25	10	14
Arquitetura	0	4	25	10	5

23) Um cientista afirma que, em certa região, o número de dias secos é o triplo do número de dias úmidos, que por sua vez é o dobro do número de dias chuvosos. Suas observações foram 93 dias secos, 35 chuvosos e 58 úmidos. Qual é sua conclusão experimental?

24) Um pesquisador afirma que o tempo de vida, medido em horas, de determinadas bactérias obedece a uma das distribuições da Família Exponencial. Mais precisamente, se X representa o tempo de vida de uma destas bactérias, ele afirma que $p_1 = P\{X < 100\} = 0,551$, $p_2 = P\{100 < X < 200\} = 0,248$, $p_3 = P\{200 < X < 300\} = 0,111$ e $p_4 = P\{X > 300\} = 0,090$. A observação de 254 bactérias resultou $O(1) = 128$, $O(2) = 73$, $O(3) = 28$ e $O(4) = 25$. Qual é sua conclusão experimental?

25) Após anotar todos os resultados do lançamento de um dado durante alguns anos, um jogador conclui que: $p_1 = 0,21$ $p_2 = 0,17$ $p_3 = 0,15$ $p_4 = 0,14$ $p_5 = 0,18$ $p_6 = 0,15$. Outro dado foi arremessado 240 vezes e apresentou $O_1 = 50$ $O_2 = 49$ $O_3 = 35$ $O_4 = 38$ $O_5 = 34$ $O_6 = 34$. Você acha que o jogador pode/deve usar este dado com as mesmas estratégias do outro? Responda usando o n.s.a. 0,05 e o n.c.a.a.

26) Numa pesquisa de mercado, os clientes foram classificados em classe econômica x opinião sobre os preços da empresa que contratou a pesquisa. Qual a sua conclusão experimental ao n.s.a. 0,05? Qual o n.c.a.a.?

	Muito caro	Caro	Igual	Regular	Barato	Muito barato	Não sabe
A	3	7	12	21	16	1	/
B	3	12	15	12	14	/	4
C	4	2	20	25	7	/	2

27) Sobre a qualidade do atendimento, os clientes do Reforço 26 expressaram as opiniões ao lado. Idem. Idem.

	Ótimo	Bom	Regular	Ruim
A	43	6	9	2
B	26	8	22	4
C	22	15	16	7

28) Os casos registrados de acidentes de trabalho de uma a.a.s. com reposição são: 98 (SEG), 146 (TER), 114 (QUA), 87 (QUI), 155 (SEX). Estes dados referem-se somente aos dias de ocorrência dos acidentes em empresas da construção civil, que usam a semana inglesa em virtude de acertos com o sindicato. Usando o n.c.a.a., teste a hipótese de que os acidentes ocorrem ao acaso.

29) Foram selecionados ao acaso e com reposição 400 crianças de escolas de um determinado padrão. A tabela ao lado expressa o desempenho escolar atribuído pelos pais e pelos professores destas crianças. Você aceita a hipótese de independência? Qual é o n.c.a.a.?

Pais \ Professores	Bom	Regular	Fraco
Bom	56	80	44
Regular	15	45	30
Fraco	49	15	66

30) Foram selecionados (?) 200 enfermos portadores de certa enfermidade para serem submetidos a 3 tipos de tratamento, cujos resultados estão tabulados a seguir. Você aceita a hipótese de que os tratamentos produzem resultados homogêneos? O que a amostra sugere?

Resultado	Bom	Regular	Insatisfatório
Tratamento A	38	60	12
Tratamento B	20	21	29
Tratamento C	2	9	9

31) Em determinado estado , as doenças A, B, C e D são responsáveis por 15%, 21%, 18%, e 14% dos óbitos, respectivamente. Um estudo das causas de 308 óbitos ocorridos em certa cidade revelou 43, 76, 85 e 21 óbitos, respectivamente. Qual a sua conclusão experimental?

32) Numa amostra (?) composta de 336 portadores de determinada moléstia, seus portadores estão classificados quanto ao grupo sangüíneo e fator RH.

a) Teste a hipótese de independência ao n.s.a. 0,05. Qual é o n.c.a.a.?

	0	A	B	AB
+	104	68	62	45
-	22	19	11	5

b) A hipótese de que a doença se encontra distribuída nos grupos sangüíneos na proporção 4 : 3: 3: 2 pode ser aceita ao n.s.a. 0,05? Qual é o n.c.a.a.?

33) Um estudo visando determinar a eficiência de um soro contra a artrite resultou na comparação de dois grupos de pacientes de 200 doentes cada, cujos resultados estão tabelados ao lado. Qual é a conclusão experimental? Qual é o n.c.a.a.?

	Soro	Placebo
Melhoraram	117	74
Não melhoraram	83	126

34) Um departamento de saúde pública acredita que a incidência de 3 doenças raras ocorram com a mesma freqüência anual em determinada região. Esta hipótese pode/deve ser aceita/rejeitada com base nos registros do ano passado: 16 casos da doença A, 23 da B e 9 da C?

LIÇÃO 13

(1) Não : "σ deve ser não inferior à 7,21 gr." e "σ deve ser não inferior à 6,43 gr.". (2) É uma estimativa do Eqm do estimador utilizado (para produzir a estimariva). (3) Não, porque $\overline{X}$ é uma estatística e não uma variável. (4) t_o = 3,650; t_c = 2,4345; rejeito . (5) 0,7257 (6) 21,56 (7) $(-\infty ; -2{,}576] \cup [+2{,}576 ; +\infty)$; 0,0000 (8) Só assim podemos ter "evidências" estatísticas de que $\sigma \geq 3$. (9 a) –0,2623 (9b) [-0,8045 ; 0,2800] (9c) t_o = -1,259; α = 0,2076; aceito. (10) 57000; [47395 ; 66605] ; A estimativa está "dentro" da hipótese nula. Portanto, aceito. (11) [18,52 ; 19,40]; [0,8656 ; 2,8196] (12) t_o = -2,423; t_c = -2,2622; rejeito. (13 a) t_o = -0,169; $\alpha \cong$ 0,8650; aceito. (13b) t_o = 0,800 ; $\alpha \cong$ 0,4238; aceito. (14) [590,0 ; 696,0]; 598,6; 687,4 (15) [6,39 ; 9,41]; 6,53; 9,27 (16 a) Todos os universitários que fumavam por ocasião do término do curso e continuaram até 5 anos após o término. (16b) [7,72 ; 11,68] (17) depende; 3,81; [583 ; 727] (18) Não; não; sim (19 a) [2,19 ; 2,74] (19b) t_o = 2,156; t_c = 1,8125; rejeito; 0,025 < α < 0,05. (20 a) Digamos que entre 23,567 e 33,711. (20b) Digamos que maior do que 52,336. (21 a) 4135,36 (21b) A estimativa para o total está "dentro" da hipótese. Portanto, aceito. (22) Sim, mas a probabilidade disto ocorrer é bastante inferior à 0,005. (23) A máquina continuaria desajustada. (24 a) [2389 ; 2949] (24b) z_o = 1,207; z_c = 1,645; aceito. (24c) A estimativa 0,3443 está "dentro" da hipótese. Portanto aceito. (25 a) z_o = 1,675 (1,282 $\Rightarrow$ rejeito, 1,645 $\Rightarrow$ rejeito, 1,960 $\Rightarrow$ aceito, 2,326 $\Rightarrow$ aceito); [97,50 ; 172,26] (25b) z_o = 21,9040 > > 3 $\Rightarrow$ rejeito; 9,65; 7,40; 6,84 (26 a) [72,7963 ; 146,6199] (26b) q_o = 55,411 ; q_c = 57,505; aceito; 0,1 < α < 0,25. (27) Esta variável é discreta, logo esta técnica não pode ser usada. (28) 8,1130 e 2,85; 13,0102 e 3,61; [7,6506 ; 14,0498] e [2,77 ; 3,75] (29 a) q_o = 1,645; $[0 ; 3{,}247] \cup [20{,}483 ; +\infty)$; rejeito. (29b) q_o = 1,645; [0 ; 3,940]; rejeito. (30 a) [-2,85 ; +8,18] (30b) q_o = 104,917; q_c = 26,757; rejeito (30c) 4,93 (31) t_o = -1,9249; t_c = $\pm$ 1,7459 ; rejeito; 0,05 < α < 0,1 ; s_x = 18,36 e s_y = 11,36 $\Rightarrow$ os enfermeiros parecem ser um grupo mais heterogêneo do que as enfermeiras. (32 a) t_o = 1,3308 ; $[1{,}3830 ; +\infty)$; aceito (32b) q_o= 19,022 ; $[0 ; 3{,}325] \cup [16{,}919 ; +\infty)$; rejeito; $\alpha \cong$ 0,05 (33 a) q_o = 12,520 ; [0 ; 12,401] ; aceito (33b) t_o = 1,2247 ; $[2{,}0639 ; +\infty)$; aceito. (34 a) t_o = 2,8201 ; 0,01 < α < 0,025; rejeito. (34b) q_o = 9,585 ; 0,1 < α < 0,25 ; aceito. (35 a) 50,14 e 19,11 (35b) 59,67 e 28,41 (35c) t_o = 2,6542; t_c = $\pm$ 2,7969 ; aceito ; q_o = 87,642 ; q_{c1} = 9,886 e q_{c2} = 45,559 ; rejeito.(36 a) [38,37 ; 51,63] e [30,25 ; 41,75]. (36b) Como estes dados são emparelhados não podemos usar o teste de Student sem conhecermos os valores d_i. Entretanto, como os IC-0,99 se sobrepõem, a igualdade entre as médias pode ser aceita. (36c) 18,11 e 15,70 (37 a) 1,03 e [1,08 ; 1,85] (37b) s = 1,36 $\geq$ 1 $\Rightarrow$ aceito ; q_o = 49,9392 ; q_{c1}= 14,573 e q_{c2} = 43,194 ; rejeito. (38) [38,45; 122,73] ; 48,80; 112,38; [53,02 ; 119,59] ; 56,91; 106,81 (39 a) [56,84 ; 122,43] (39b) q_o = 17,219; q_{c1} = 5,629 e q_{c2} = 26,119; aceito. (40 a) t_o = 1,5118 ; $(-\infty ; -2{,}0141] \cup [+2{,}0141 ; +\infty)$; aceito. (40b) q_o = 124,676 $[61{,}656 ; +\infty)$; rejeito (41 a) [0,95 ; 1,19] (41b) z_o = 23,62 ; z_c = $\pm$ 1,960 ; rejeito.

LIÇÃO 15 – REGRESSÃO E CORRELAÇÃO 13 reforços

Há muitas situações nas quais estamos interessados em estudar simultaneamente duas variáveis definidas numa mesma população.

PROBLEMA 1. Desejamos funções f e g definidas em $\Re^1$ que expressam o valor médio de uma variável para cada valor fixado da outra: $f(x)=E[Y \mid X=x]$ e $g(y)=E[X \mid Y=y]$.

A função f é denominada FUNÇÃO DE REGRESSÃO DE Y EM RELAÇÃO À X e a função g tem denominação análoga. Se conhecemos uma destas funções, sabemos apenas a média de uma das variáveis para cada valor fixado da outra.

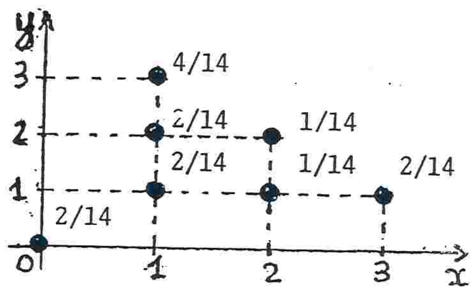
PROBLEMA 2. Desejamos funções lineares f_0 e g_0 definidas em $\Re^1$ que estejam mais próximas de f e de g, respectivamente, do que as demais funções lineares.

A "proximidade" do Problema 2 deve ser definida clara e convenientemente. Em geral, a distância usada é $d(f,f_0)=\{E([f(X)-f_0(X)]^2)\}^{1/2}$, definida analogamente para g e g_0. Como os resultados também são análogos, de agora em diante vamos trabalhar explicitamente com apenas uma das funções. Para a outra, é só fazermos as substituições correspondentes.

O Problema 2 é denominado PROBLEMA DE REGRESSÃO LINEAR A DUAS VARIÁVEIS. A função linear $f_0(X)=a_0 \cdot X+b_0$ é denominada RETA DE REGRESSÃO DE Y EM RELAÇÃO À X. Seus coeficientes são expressos pelas fórmulas:

$$a_0=\frac{E(XY)-E(X)\cdot E(Y)}{Var(X)} \quad \text{e} \quad b_0=\frac{E(Y)\cdot E(X^2)-E(X)\cdot E(XY)}{Var(X)}.$$

EXEMPLO 1 – Considere o vetor aleatório bidimensional (X,Y) com a distribuição de probabilidade da figura. Temos:



$f(0)=0\times 1=0$

$f(1)=1\times 2/8+2\times 2/8+3\times 4/8=9/4$

$f(2)=1\times 1/2+2\times 1/2=3/2$

$f(3)=1\times 1=1$

$g(0)=0\times 1=0$

$g(1)=1\times 2/5+2\times 1/5+3\times 2/5=2$

$g(2)=1\times 2/3+2\times 1/3=4/3$

$g(3)=1\times 1=1$

$E(XY)=30/14$

$E(X)=18/14$ $\quad E(X^2)=34/14$ $\quad Var(X)=152/196$

$E(Y)=23/14$ $\quad E(Y^2)=53/14$ $\quad Var(Y)=213/196$

$$a_0 = \frac{30/14 - 18/14 \times 23/14}{152/196} = \frac{6}{152} \qquad b_0 = \frac{23/14 \times 34/14 - 18/14 \times 30/14}{152/196} = \frac{242}{152}$$

$$c_0 = \frac{30/14 - 23/14 \times 18/14}{213/196} = \frac{6}{213} \qquad d_0 = \frac{18/14 \times 53/14 - 23/14 \times 30/14}{213/196} = \frac{264}{213}$$

$$f_0(x) = \frac{3x + 121}{76} \qquad g_0(y) = \frac{2y + 88}{71}$$

Aqui X e Y não são variáveis independentes, pois nem suas funções de regressão nem suas retas de regressão são funções constantes. ▲

O conhecimento dos valores verdadeiros de a_0 e de b_0 pode ser obtido somente com a distribuição verdadeira (i.e., populacional) conjunta de X e Y. Ora, se conhecemos esta distribuição, não temos problema de inferência estatística. Se, entretanto, todo o nosso conhecimento é uma distribuição amostral conjunta de X e Y, os valores verdadeiros de a_0 e de b_0 não podem ser calculados. Tudo o que podemos fazer são inferências a respeito deles.

PROBLEMA 3. Desejamos estimadores $\hat{a}, \hat{b}, \hat{c}, \hat{d}$ para os parâmetros a_0, b_0, c_0, d_0 de f_0 e g_0.

EXEMPLO 2 (ESPIONAGEM ADMINISTRATIVA) – Uma amostra de 135 operários de certa fábrica tem seu tempo de trabalho X, em anos completados, e seu salário Y, em certa unidade financeira, tabulados a seguir. Vamos estimar as retas de regressão f_0 e g_0, e os números $f_0(8)$, $f_0(9)$, $f_0(10)$, $g_0(12)$, $g_0(13)$, $g_0(14)$.

$\sum x_i = 543 \quad \sum x_i^2 = 2603$

$\hat{V}ar(X) = 3{,}1032$

$\sum y_i = 1035 \quad \sum y_i^2 = 8257$

$\hat{V}ar(Y) = 2{,}3852$

y \ x	0	1	2	3	4	5	6	
5	1	5	4	3				
6	2	2	8	7	3			
7	1	1	1	6	9	2	1	
8		1	1	5	12	13	7	2
9				2	3	7	6	4
10					1	3	5	2
11						1	3	1

$\sum x_i y_i = 4436 \quad \hat{f}_0(x) = 0{,}65 \cdot x + 5{,}05 \quad \hat{f}_0(8) = 10{,}25 \quad \hat{f}_0(9) = 10{,}90 \quad \hat{f}_0(10) = 11{,}55$

$\hat{g}_0(y) = 0{,}85 \cdot y - 2{,}48 \quad \hat{g}_0(12) = 7{,}72 \quad \hat{g}_0(13) = 8{,}57 \quad \hat{g}_0(14) = 9{,}42$ ▲

A reta de regressão de Y em relação à X nada informa sobre a dispersão conjunta de X e Y em torno dela. Esta reta é, apenas, a reta (verdadeira ou estimada) mais próxima da função de regressão de Y em relação à X. Na Figura 2 a seguir, f_0 é muito mais explicativa da variabilidade de Y dado X do que na Figura 1 a seguir.

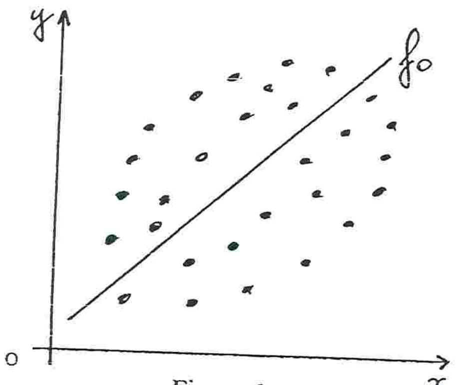


Figura 1

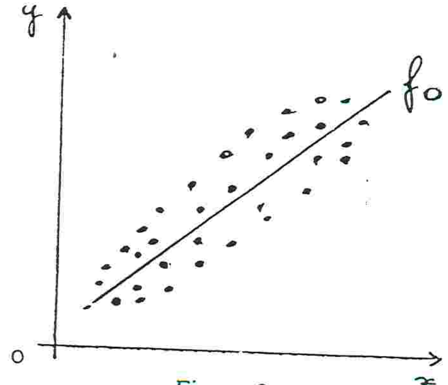


Figura 2

PROBLEMA 4. Desejamos uma medida da dispersão conjunta de X e Y em torno de f_0 e de g_0.

TEOREMA 1. $$\mathrm{Var}(Y) = E\left[(Y - f_0(X))^2\right] + E\left[(f_0(X) - E(Y))^2\right].$$

A parcela $E\left[(Y - f_0(X))^2\right]$ é o RESÍDUO, ou a VARIAÇÃO DE Y NÃO EXPLICADA POR f_0, e a outra parcela é a VARIAÇÃO DE Y EXPLICADA POR f_0. A dispersão conjunta de X e Y em torno de f_0 é a CORRELAÇÃO LINEAR ENTRE (?) X e Y, usualmente medida por

$$\rho(X, Y) = \mathrm{sgn}(a_0) \times \left\{E\left[(f_0(X) - E(Y))^2\right] / \mathrm{Var}(Y)\right\}^{1/2}.$$

Este número é o COEFICIENTE DE CORRELAÇÃO LINEAR ENTRE (?) X e Y. Como $\rho^2(X,Y)$ = (VARIAÇÃO DE Y EXPLICADA POR f_0) / (VARIAÇÃO TOTAL DE Y), deduzimos que $-1 \leq \rho(X,Y) \leq +1$.

TEOREMA 2. $$\rho(X, Y) = a_0 \times \frac{\mathrm{Des}(X)}{\mathrm{Des}(Y)} = \frac{E(XY) - E(X) \cdot E(Y)}{\mathrm{Des}(X) \cdot \mathrm{Des}(Y)}.$$

A última expressão do Teorema 2 é usada para o cálculo de $\rho(X,Y)$ e sua simetria justifica o nome "coeficiente de correlação linear entre X e Y" ao invés de "coeficiente de correlação linear de Y em relação à X".

É bastante comum, na prática desconhecermos a distribuição verdadeira (i.e., populacional) conjunta de X e Y. Usualmente, todo o nosso conhecimento é uma distribuição amostral conjunta de X e Y e, portanto, não podemos calcular o coeficiente de correlação linear entre X e Y. Tudo o que podemos fazer é inferência a seu respeito. Por exemplo, ele pode ser estimado (por ponto) através de

$$\hat{\rho}(X,Y)=\frac{n\sum x_i y_i-(\sum x_i)\cdot(\sum y_i)}{\left[n\sum x_i^2-(\sum x_i)^2\right]^{1/2}\cdot\left[n\sum y_i^2-(\sum y_i)^2\right]^{1/2}}.$$

TEOREMA 3. Se $(X_1, Y_1),\ldots,(X_n, Y_n)$ for uma a.a.s. com reposição de uma DISTRIBUIÇÃO NORMAL BIVARIADA com $\rho = 0$, então a estatística $T=\frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}}$ tem distribuição de Student com $n-2$ graus de liberdade.

EXEMPLO 3 – No Exemplo 1, vamos calcular $\rho(X, Y)$.

$$\rho(X,Y)=\frac{E(XY)-E(X)\cdot E(Y)}{Des(X)\cdot Des(Y)}=\frac{30/14-18/14\times 23/14}{(152/196\times 213/196)^{1/2}}=0{,}0333\,.\ \blacktriangle$$

EXEMPLO 4 – No Exemplo 2, vamos estimar $\rho(X, Y)$ e testar H_0: $\rho = 0$ vs. H_1: $\rho > 0$.

$$\hat{\rho}(X,Y)=\frac{135\times 4446-543\times 1035}{(135\times 2603-543^2)^{1/2}\times(135\times 8257-1035^2)^{1/2}}=0{,}7433\,.$$

$$t=\frac{0{,}7433\times\sqrt{135-2}}{\sqrt{1-0{,}7433^2}}=12{,}81\ \therefore\ \alpha(12{,}81)=0{,}0000\ \therefore\ \text{rejeito}.\ \blacktriangle$$

EXEMPLO 5 – Vamos testar H_0: $\rho = 0$ vs. H_1: $\rho \neq 0$ com $n = 8$ e $\hat{\rho} = 0{,}7666$.

$$t=\frac{0{,}7666\times\sqrt{8-2}}{\sqrt{1-0{,}7666^2}}=2{,}92\ \therefore\ 0{,}02<\alpha(2{,}92)<0{,}05\ \therefore\ \text{rejeito}.\ \blacktriangle$$

REFORÇOS (Respostas na página 114)

1) Existem métodos de Inferência Estatística para obtermos conjuntos de confiança para os parâmetros da reta de regressão e para o coeficiente de correlação linear? Existem métodos de Inferência Estatística para testarmos hipóteses a respeito destes parâmetros?

2) Existem problemas de regressão a mais de duas variáveis? Existem problemas de regressão não lineares a duas variáveis?

3) Uma a.a.s. com reposição de tamanho 25 apresentou $\sum x_i = 125{,}36$ $\sum x_i^2 = 716{,}2474$ $\sum y_i = 168{,}45$ $\sum y_i^2 = 1623{,}4556$ $\sum x_i y_i = 874{,}1373$. Estime $E[X|Y = 8{,}76]$.

4) Na tabela abaixo estão as freqüências relativas da ocorrência dos valores de duas variáveis numa população. Calcule a função de regressão e a reta de regressão de U em relação à V. As variáveis U e V são independentes ? Justifique.

u \ v	-2	-1	0	1	2
0	0,16	0,07	0,02	0,01	/
1	0,04	0,15	0,09	0,02	0,02
2	0,08	0,02	0,10	0,10	0,12

x \ y	0	1	2
-1	1/25	4/25	3/25
0	2/25	5/25	2/25
1	3/25	4/25	1/25

5) As variáveis X e Y têm a distribuição de probabilidade conjunta da tabela acima. Elas são independentes ? Obtenha as funções de regressão e as retas de regressão. Calcule ρ (X,Y). Calcule os Eqm_s de f e f_0 como previsores de Y. Idem para g e g_0 como previsores de X.

6) Obtém fórmulas para $Eqm(Y, f(X))$ e $Eqm(Y, f_0(X))$.

7) Considere os 50 estudantes do Exemplo 2 na página 63 uma população. Calcule $f, f_0, \rho, Eqm(Y, f(X)), Eqm(Y, f_0(X)), d(Y, f(X))$ e $d(Y, f_0(X))$.

8) Considere a amostra de jogadores de futebol tabulada a seguir, onde Z é a idade e T é o total de gols feitos em seus clubes atuais. Estime o coeficiente de correlação linear, as duas retas de regressão e a média do total de gols feitos em seus clubes atuais por jogadores com a idade de 27 anos. Teste $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$.

z	25	22	24	28	27	21	30	23	21
t	22	6	1	4	2	2	14	6	3

9) Considere a amostra do Reforço 34 da página 98. Qual a sua estimativa para o rendimento médio de combustível com pneus radiais dos veículos que rendem 18,7 km/l com pneus comuns ? Teste $H_0 : \rho = 0$ versus $H_1 : \rho > 0$.

10) Uma amostra de 6 empresas de engenharia apresentou, no mês passado, a receita X e a despesa com pessoal Y, expressas em certa unidade monetária, tabuladas a seguir. Estime a média com despesa com pessoal das empresas com receita igual a 60. $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$.

x	62	62	87	65	58	92
y	14,3	20,8	22,7	30,5	41,2	28,2

11) Em quais situações as retas de regressão de Y em relação à X são bons previsores de Y ?

12) Se Y é o preço unitário de certo equipamento ao consumidor, X é a oferta (em número de fábricas que o produzem) e $\hat{f}(x) = 1000\ (10 + 2/x)$, o que significa $\hat{f}(5)$?

13) X é um indicador quantitativo contínuo geral da saúde de pessoas com 70 anos ou mais. Y é um indicador do número de horas semanais em frente à TV. Uma a.a.s. apresentou $\hat{\rho} = -0,425$. Teste $H_0 : \rho = 0$ ao n.s. 0,1. Qual o n.d.a. ? Qual o significado prático do erro eventual ?

(de tamanho 18)

LIÇÃO 14

(5) A rigor, não. Somente a.a.s. com reposição. Entretanto, ... (6) $t = 21{,}830 > 11{,}345 \Rightarrow$ rejeito ; $\alpha << 0{,}005$ (7) $t = 7{,}407 > 7{,}378 \Rightarrow$ rejeito ; $\alpha \cong 0{,}025$ (8) $t = 37{,}340 > 5{,}991 \Rightarrow$ rejeito; $\alpha << 0{,}005$ (9) Não, porque devemos fazer a.a.s. com reposição. (10) Sim, o Teste de Homogeneidade com $k = c = 2$. (11) $t = 3{,}781 < 3{,}841 \Rightarrow$ aceito ; $\alpha \cong 0{,}0526$ (12) $t = 12{,}582 > 10{,}645 \Rightarrow$ rejeito ; $\alpha \cong 0{,}05$ (14) Exigem amostras de tamanho grandes, pois usam distribuições nulas assintóticas. (15 a) $\hat{p} = 87/385 = 0{,}23$; $H_0 : p \leq 2/7 = 0{,}29$; aceito (15b) $H_0 : p > 2/7$; $t = -2{,}5947$; $\alpha \cong 0{,}0048 \Rightarrow$ rejeito (16) $\alpha = 0{,}05$: $t_c = 12{,}592 \Rightarrow$ rejeito ; $\alpha = 0{,}01$: $t_c = 16{,}812 \Rightarrow$ aceito (17 a) Sim, o Teste Qui-quadrado de Homogeneidade. (17b) $t = 10{,}988 > 7{,}879 \Rightarrow$ rejeito ; $\alpha < 0{,}005$ (18) $t = 28{,}693 > 14{,}860 \Rightarrow$ rejeito ; $\alpha << 0{,}005$ (19) $t = 1{,}475 < 9{,}488 \Rightarrow$ aceito ; Nenhum, pois $t < E_0(T) = 4$. Também porque $\alpha \cong 0{,}8282$. (20) $t = 12{,}214$; $0{,}01 < \alpha < 0{,}025$; Não admito. (21) $t = 34{,}538 > 10{,}597 \Rightarrow$ rejeito ; $\alpha << 0{,}005$ (22) $t = 13{,}492 > 12{,}592 \Rightarrow$ rejeito ; $0{,}025 < \alpha < 0{,}05$ (23) $t = 24{,}411 > 10{,}597 \Rightarrow$ rejeito ; $\alpha << 0{,}005$ (24) $t = 2{,}813 < E_0(T) = 3 \Rightarrow$ aceito ; $\alpha < 0{,}5$ (25) $t = 4{,}326 < E_0(T) = 5 \Rightarrow$ aceito ; $\alpha \cong 0{,}5$; Sim, pode. (26) $t = 14{,}724 > 12{,}592 \Rightarrow$ rejeito ; $0{,}01 < \alpha < 0{,}025$ (27) $t = 21{,}146 > 12{,}592 \Rightarrow$ rejeito ; $\alpha << 0{,}005$ (28) $t = 29{,}250 > 14{,}860 \Rightarrow$ rejeito ; $\alpha << 0{,}005$ (29) $t = 53{,}827 > 14{,}860 \Rightarrow$ rejeito ; $\alpha << 0{,}005$ (30) $t = 28{,}693 > 14{,}860 \Rightarrow$ rejeito ; $\alpha << 0{,}005$ (31) $t = 31{,}768 > 14{,}860 \Rightarrow$ rejeito ; $\alpha << 0{,}005$ (32 a) $t = 3{,}398 < 7{,}815 \Rightarrow$ aceito ; $\alpha > 0{,}25$ (32b) $t = 3{,}940 < 7{,}815 \Rightarrow$ aceito ; $\alpha > 0{,}25$ (33 a) $t = 0{,}658 < 9{,}348 \Rightarrow$ aceito ; $\alpha \cong 0{,}7676$ (33b) Sim: amostras pequenas, amostras mal selecionadas, intervenção "malévola" do acaso. (34) $t = 6{,}125$; $0{,}025 < \alpha < 0{,}05$; rejeito ; $\alpha \cong 0{,}0476$.

LIÇÃO 15

(1) Sim. É só procurá-los na literatura específica. (2 a) Sim. Por exemplo, $z = ax + by + c$. (2b) Sim. Por exemplo, $y = ax^b + c$. (3) 5,13 (4 a) $g(-2) = 0{,}71$; $g(-1) = 0{,}79$; $g(0) = 1{,}38$; $g(1) = 1{,}69$; $g(2) = 1{,}86$ (4b) $0{,}32\,v + 1{,}28$ (4c) Não, pois nem g nem g_0 são constantes. (5 a) Não, pois $P\{X = 1 / Y = 0\} = 3/6 \neq P\{X = 1 / Y = 2\} = 1/6$ (5b) $f(-1) = 5/4$; $f(0) = 1$; $f(1) = 3/4$; $-x/4 + 1$ (5c) $g(0) = 1/3$; $g(1) = 0$; $g(2) = -1/3$; $-(y+1)/3$ (5d) $-0{,}2887$ (5e) 11/25; 11/25 ; 44/75 ; 44/75 (6) $\mathrm{Eqm}(Y, f(X)) = E(Y^2) + E(f^2(X)) - 2 \times E(Y \cdot f(X))$; $\mathrm{Eqm}(Y, f_0(X)) = (1 - \rho^2) \cdot \mathrm{Var}(Y)$ (7) ...; $-2{,}38\,x + 59{,}75$; $-0{,}52$; 155,28 ; 208,16 ; 12,5 ; 14,4 (8 a) 0,3431 ; $0{,}16\,t + 23{,}49$; $0{,}75\,z - 11{,}75$; 8,5 (8b) $t_0 = 0{,}9664 < 1 \Rightarrow$ aceito (9 a) 19,1 (9b) $t_0 = 29{,}42 >> 3{,}7074 \Rightarrow$ rejeito (10 a) 27,10 (10b) $t_0 = 0{,}2353 < 1 \Rightarrow$ aceito (11) Quando ρ estiver próximo de -1 ou de $+1$. (12) É uma estimativa do preço unitário médio ao consumidor do equipamento quando houver 5 fábricas produzindo-o. (13) $t_0 = -1{,}8781$; $t_c = \pm 1{,}7459$; rejeito ; entre 0,05 e 0,1 ; ...

Um conhecimento de Matemática ajuda bastante um investigador ou um profissional a analisar as informações que possui e também a tomar decisões com uma lógica mais aprimorada. Para que isto aconteça não é necessário tornar-se um virtuose, mas é imprescindível entender certas noções e resultados básicos.

Qualquer ciência que trabalhe com mensurações não pode prescindir do "manual que ensina a operar estas ferramentas": a Matemática. Além disto, os processos mentais empregados pelos matemáticos em seus problemas são formalmente semelhantes aos utilizados pelos cientistas em seus trabalhos e aos utiiizados pelos seres humanos em geral em seus raciocínios a respeito das coisas mundanas.

A terminologia matemática e sua simbologia proporcionam uma economia substancial de espaço e uma pureza absoluta de significado. Ao usarmos matemática em um problema específico, não somos obrigados a usar quer a terminologia quer a simbologia já existentes; podemos criar termos e símbolos quando isto for conveniente, mas sempre especificando com precisão seus significados. A Matemática não é sua terminologia nem sua simbologia, mas as idéias que elas transmitem.

Até pouco atrás, a abordagem de todo problema científico era determinística. Em muitos problemas esta abordagem conduz a bons resultados, mas em alguns a INCERTEZA, isto é, a VARIABILIDADE DAS MENSURAÇÕES, isto é, a IMPREVISIBILIDADE DOS RESULTADOS, deve ser considerada. Nestas situações, ou seja, quando é usada uma abordagem aleatória, qualquer metodologia utilizada pertence a uma destas duas áreas da Matemática: ou à PROBABILIDADE ou à ESTATÍSTICA. (Vamos considerar a Probabilidade como parte da Estatística daqui para a frente, apenas para simplificar as coisas).

Vista de outro ângulo, a Estatística é o corpo de conhecimentos que trata dos métodos para manejar dados numéricos. Tais métodos são empregados para reunir, classificar, analisar e interpretar dados numéricos, além de obter conclusões a partir deles. A Estatística se ocupa dos dados em si, sem relevar o processo que os origina; assim, por exemplo, uma pessoa que possui uma listagem ou uma tabulação de vários resultados de arremessos de um dado honesto pode concluir, mesmo sem conhecer um dado, que se trata de um processo que gera ao acaso números de 1 a 6.

A ESTATÍSTICA DESCRITIVA trata dos métodos usados para descrever os dados obtidos, isto é, trata da compreensão de tais dados. A INFERÊNCIA ESTATÍSTICA, ou ESTATÍSTICA INDUTIVA, trata da obtenção de conclusões que transcendem os dados obtidos e que se aplicam a um conjunto maior de dados denominado POPULAÇÃO ou UNIVERSO.

O que é precisamente certo para a amostra não pode ser considerado mais do que aproximadamente certo para a população, e esta aproximação pode ser excelente ou boa ou não ter valor algum. Entretanto, se não entendemos bem o que os dados significam, a utilização de qualquer metodologia estatística, por mais popular, bonita ou sofisticada que seja, é completamente inútil. Esta ênfase é importante porque os números geralmente não são sagrados (veja a Lição 11) e estão sujeitos a ambigüidades que podem nos privar de uma parte substancial de toda a informação que julgamos contenham.

Em qualquer situação que usamos Inferência Estatística estão presentes uma população, uma ou mais variáveis e uma amostragem. Esta, a rigor, deve ser probabilística.

Quando estimamos um parâmetro θ, a qualidade de um estimador **T** é medida geralmente pelo ERRO QUADRÁTICO MÉDIO $\mathbf{Eqm\ (T,\theta) = E\,[\,(T-\theta)^2\,]}$. É facilmente demonstrado que $\mathbf{Eqm\ (T,\theta) = Var\,(T) + [\,E\,(T) - \theta\,]^2}$. O número $\mathbf{E\,(T) - \theta}$ é o VÍCIO do estimador T. Portanto, para ESTIMADORES NÃO VICIADOS, i.e., para estimadores tais que $\mathbf{E(T) = \theta}$, o erro quadrático médio é igual a variância. Os estimadores $\hat{p}, \hat{p}\mathrm{er}, \hat{\tau}$ na Lição 12 e $\hat{\mu}, \hat{\tau}$ na Lição 13 são não viciados e seus desvios padrão são estimados pelos erros padrão. Para estimadores viciados, os erros padrão devem considerar, evidentemente, as magnitudes dos seus vícios.

EXEMPLO 1 - O parâmetro p da Lição 12 pode ser estimado por $\hat{p} = (T+1)/(n+2)$, que nunca produz as estimativa 0 e 1.

$$\text{vício}(\hat{p}, p) = E(\hat{p}) - p = E\left(\frac{T+1}{n+2}\right) - p = \frac{E(T+1)}{n+2} - p = \frac{np+1}{n+2} - p = \frac{1-2p}{n+2}$$

$$\mathrm{Var}(\hat{p}) = \mathrm{Var}\left(\frac{T+1}{n+2}\right) = \frac{\mathrm{Var}(T)}{(n+2)^2} = \frac{N-n}{N-1} \times \frac{np(1-p)}{(n+2)^2}$$

$$\mathrm{Eqm}(\hat{p}, p) = \frac{N-n}{N-1} \times \frac{np(1-p)}{(n+2)^2} + \left(\frac{1-2p}{n+2}\right)^2$$

▲

EXEMPLO 2 - O parâmetro σ^2 da Lição 13 pode ser estimado por $\hat{\sigma}^2 = \sum (x_i - \bar{x})^2 / n$. Como $n \cdot \hat{\sigma}^2 = (n-1) \cdot s^2$, temos: $\text{vício}(\hat{\sigma}^2, \sigma^2) = -\sigma^2 / n$, $\text{Var}(\hat{\sigma}^2) = 2(n-1)\sigma^4 / n^2$ e $\text{Eqm}(\hat{\sigma}^2, \sigma^2) = 2(n-1)\sigma^4 / n^2 + \sigma^4 / n^2 = (2n-1)\sigma^4 / n^2$. ▲

Quando um estimador se torna cada vez mais preciso com o aumento do tamanho da amostra, ele é denominado um ESTIMADOR CONSISTENTE. Em outras palavras, um estimador consistente é um estimador que não erra com amostras de tamanho infinito e erra por quantidades desprezíveis com amostras de tamanho grande. Evidentemente, isto não é uma conceituação precisa da noção de "consistência de um estimador". Uma conceituação precisa desta noção envolve um instrumental matemático um tanto acima daquele que estamos aqui usando, mas podemos esclarecer que, quando o **Eqm** do estimador se anula para amostras de tamanho infinito (como ocorre nos dois exemplos acima), o estimador é consistente.

Quando estimamos um parâmetro θ através de um estimador por intervalo $[T_1 ; T_2]$, a qualidade do estimador (por intervalo) é medida geralmente por alguma combinação do coeficiente de confiança γ e da amplitude média das estimativas por intervalo $E[T_2 - T_1]$. O instrumental matemático a que nos limitamos nos fez mencionar apenas o coeficiente de confiança. É claro que, entre vários estimadores por intervalo de mesma confiança, escolhemos aquele que tiver amplitude média menor.

Testamos uma hipótese quando desejamos verificar se os dados obtidos em um experimento, isto é, da amostra, estão em harmonia com alguma afirmação específica sobre a distribuição populacional das variáveis em estudo, tal como: a média de X é 5, X e Y são variáveis independentes, X e Y têm a mesma variância, a proporção de parafusos defeituosos é menor do que 3/144 etc. Esta afirmação é denominada HIPÓTESE (NULA). Sua negação (no universo em questão) é denominada (HIPÓTESE) ALTERNATIVA.

Devemos escolher uma entre estas duas decisões: aceitar a hipótese ou rejeitar a hipótese. Esta escolha é sempre baseada em mensurações realizadas na amostra selecionada. Uma técnica que transforma qualquer amostra numa destas duas decisões é denominada um TESTE. O conjunto das amostras, ou equivalentemente, o conjunto de valores de uma estatística, que um teste associa à decisão de rejeitar a hipótese é denominada REGIÃO CRÍTICA DO TESTE. É óbvio que a região crítica determina o teste.

Quando a hipótese nula, representada por H_0, é verdadeira e nós a rejeitamos, ocorre o que denominamos ERRO DO TIPO I.

		NATUREZA	
		H_0 é verdadeira	H_0 é falsa
DECISÃO	Aceitamos H_0	OK	ERRO DO TIPO II
	Rejeitamos H_0	ERRO DO TIPO I	OK

Quando H_0 é falsa e nós a aceitamos ocorre o que é denominado ERRO DO TIPO II. A hipótese alternativa é representada por H_1.

EXEMPLO 3 - Especificou-se para um concreto a resistência de 150 kgm / m^2. Vamos retirar 16 corpos de prova. Se definimos o problema como H_0: $\mu \geqslant 150$ vs H_1: $\mu < 150$, a conseqüência de cometermos Erro do Tipo I será deixarmos de usar um concreto que tem a resistência especificada. Por outro lado, se definimos o problema como H_0: $\mu < 150$ vs H_1: $\mu \geq 150$, as conseqüências dos erros se invertem. ▲

EXEMPLO 4 - Um fabricante de conservas deseja que sua máquina de envasar coloque 100 gramas de ervilha em cada lata. O desvio padrão tolerado é de 2 gramas. Se definimos o problema como H_0: $\mu = 100$ vs H_1: $\mu \neq 100$, a conseqüência de cometermos Erro do Tipo I será interromper a produção e ajustar, ou contratar alguém para ajustar, a máquina de envasar sem necessidade. Esta é também a conseqüência de cometermos Erro do Tipo I ao testarmos K_0: $\sigma \leq 2$ vs K_1: $\sigma > 2$. Nestes dois problemas, a conseqüência de cometermos Erro do Tipo II é, evidentemente, deixar ou que a máquina envase pouca ervilha ou envase purê de ervilhas ou ambos. ▲

Evidentemente desejamos usar um teste cujas probabilidades de erros de ambos os tipos sejam próximas de zero. Desafortunadamente, isto só existe quando o tamanho da amostra é muito grande, o que é um inconveniente prático. A teoria estatística dos testes de hipótese resolve este impasse da seguinte forma: arbitramos um limitante pequeno para a probabilidade de cometermos Erro do Tipo I e não nos preocupamos com a probabilidade de cometermos Erro do Tipo II, que pode eventualmente ser pequena também. Este limitante pequeno para a probabilidade de rejeitarmos H_0 quando ela é verdadeira é denominado NÍVEL DE SIGNIFICÂNCIA.

A escolha do nível de significância é totalmente arbitrária, apesar de que o uso parece ter consagrado o valor 0,05. Não há um limitante preciso válido em todas as situações para a probabilidade tolerada de cometermos Erro do Tipo I. Os valores 0,1 0,025 0,01 0,005 têm sido também bastante utilizados. A seleção de um nível de significância entre os valores aqui citados permite que as tabelas necessárias para realizarmos os testes sejam reduzidas. Um nível de significância quase tão "popular" como 0,05 é o nível de significância 0,01.

Uma HIPÓTESE SIMPLES é uma hipótese que especifica completamente a distribuição das variáveis a que se refere. Caso contrário, ela é uma HIPÓTESE COMPOSTA. (Aqui a palavra "hipótese" se refere tanto à hipótese nula quanto à hipótese alternativa.)

A qualidade de um teste $\{a \in A : T(a) \in C\}$, ou equivalentemente, da região crítica C, é geralmente medida pela FUNÇÃO PODER $Fpo(C) = P\{a \in A : T(a) \in C\} = P(C)$. Os testes ideais, que existem muito raramente, são aqueles que satisfazem:

$Fpo(C) = 0$, se H_0 é verdadeira e

$Fpo(C) = 1$, se H_1 é verdadeira.

Os testes bons são aqueles que estão próximos dos testes ideais.

Vamos observar a importante relação entre a função poder de um teste e os dois tipos de erro existentes:

$Fpo(C) = P$ (Erro do Tipo I), se H_0 é verdadeira e

$Fpo(C) = 1 - P$ (Erro do Tipo II), se H_1 é verdadeira.

Através desta relação constatamos mais uma vez que os testes bons são aqueles com função poder pequena se H_0 é verdadeira e com função poder grande se H_1 é verdadeira.

Como o Erro do Tipo I é controlado pelo nível de significância, resta a preocupação com o Erro do Tipo II. Por esta razão, vários livros e artigos consideram a função poder apenas se H_1 é verdadeira. De qualquer forma, um TESTE PODEROSO é um teste com função poder grande se H_1 é verdadeira.

O cálculo da função poder raramente é uma tarefa fácil e geralmente é trabalho para os estatísticos matemáticos. Vários livros e artigos práticos, entretanto, mencionam a função poder com o objetivo de esclarecer a qualidade de um teste estatístico ou com o objetivo de comparar dois ou mais testes estatísticos.

REFORÇOS (Respostas na página 136)

1) Se um pesquisador utiliza a metodologia (s.f., ato de dirigir o espírito na direção da verdade) estatística denominada Teste de Hipótese, qual das duas decisões é merecedora de sua confiança?

2) As inferências estatísticas diferem das inferências subjetivas e das inferências dogmáticas?

3) Uma comunicação matemática com a linguagem comum é impossível?

4) As hipóteses H_o e H_1 podem ser verdadeiras simultaneamente?

5) Ao testar uma hipótese com o nível de significância 0,05 um pesquisador aceitou-a. Você sabe algo a respeito da probabilidade dele ter tomado a decisão correta?

6) Qual é a relação entre Erro do Tipo I e Erro do Tipo II?

7) Qual enfoque é mais limitado, o determinístico ou o aleatório?

8) Uma amostra de 40 mães de pacientes aidéticos submetidas ao TAT possibilitou a obtenção de [64,3%; 70,1%] como um intervalo de confiança 0,99 para a percentagem de mães que apresentam intensa negação maníaca. Qual é o significado deste resultado?

9) O que, tecnicamente, são bons estimadores e bons testes?

10) Presume-se que as crianças de um certa faixa etária em Porto Alegre sejam submetidas a uma carga semanal de televisão maior do que as de Florianópolis. Explique sucintamente como se poderia "comprovar" isso através de uma inferência estatística.

11) Nos livros de Estatística Não Paramétrica geralmente se encontra uma frase do tipo: "Quando ambos podem ser aplicados, os testes paramétricos usualmente são mais poderosos do que os testes não paramétricos." Qual é a utilidade desta informação?

12) Qual é a relação entre erro quadrático médio e erro padrão?

13) Que tipos de estimadores estudamos nestas notas?

14) O parâmetro está dentro do intervalo de confiança?

15) Existe uma decisão estatística confiável?

16) Como se pode "comprovar" uma hipótese de pesquisa através de uma inferência estatística?

17) O que significa a frase: " [2; 8] é um intervalo de confiança 0,94 para o parâmetro θ"?

18) Na prática, como avaliamos a precisão de um estimador por ponto?

19) Porque não atribuímos valores lógicos às inferências estatísticas ?

20) Diga duas utilidades para as estatísticas.

21) O que medem os coeficientes de confiança?

22) Se 553 caras forem obtidas em 1000 lançamentos de uma moeda honesta, qual é um intervalo de confiança assintótica 0,914 para a probabilidade de ocorrer cara em um único lançamento?

23) Estatísticas equivalentes (veja o Reforço 30 da Lição 12) são iguais na forma e diferentes no conteúdo?

24) Há nesta lição alguma frase equivalente à "as palavras são apenas dedos apontando a lua; deve-se olhar a lua e não os dedos"?

25) Na assembléia de uma empresa votam 604 acionistas que costumam não faltar. Para que determinadas mudanças sejam realizadas é necessário que, no mínimo, 2/3 da assembléia as aprovem. Um dos participantes decide propor uma destas mudanças. Se ele entrevista 58 participantes antes da sessão, que quantidade amostral lhe "assegura" a aprovação?

26) O PROBLEMA DUAL de um problema estatístico de testar uma hipótese é o problema que tem as mesmas hipóteses, porém invertidas. Quais são os problemas duais H_0^* vs. H_1^* e K_0^* vs. K_1^* no Exemplo 4?

27) A rejeição da hipótese nula implica na veracidade da hipótese alternativa?

28) O que é necessário para um pesquisador obter um bom resultado com uma técnica estatística?

29) Como deve ser a função poder de um teste bom para cada um dos problemas a seguir?

a) $\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$ b) $\begin{cases} K_0 : \mu \leq \mu_0 \\ K_1 : \mu > \mu_0 \end{cases}$ c) $\begin{cases} L_0 : \mu = \mu_0 \\ L_1 : \mu < \mu_0 \end{cases}$

30) Qualquer conclusão estatística pode ser enunciada com alguma medida do erro que ela possa ter? De que tipo são estas medidas?

31) Nos Exemplos 3 e 4, caso as variáveis subjacentes tenham distribuição normal, há alguma hipótese simples?

32) Uma a.a.s. com reposição de tamanho 150 de uma variável apresentou $\sum x_i$ = 482,6. Obtenha, se possível, uma estimativa para a média e uma para o desvio padrão. Com a suposição de que a distribuição verdadeira é normal, de que modo você obteria uma estimativa para a probabilidade da variável exceder sua média?

33) O modelo normal explica com uma aproximação bastante satisfatória muitas variáveis que existem na prática. Nestes casos, quando se deseja estimar o maior valor da variável, qual deve ser o parâmetro e como se poderia estimá-lo?

34) Conceitua erro padrão de uma estimativa.

35) Um estimador com variância pequena é um bom estimador?

36) Estimador pontual ou por conjunto?

37) Um comprador de tijolos acredita que a qualidade dos tijolos entregues por determinado fabricante esteja diminuindo. A pressão média de esmagamento sempre foi de 400 lb/pol^2 com desvio padrão de 20 lb/pol^2. Monte dois problemas de teste de hipóteses, um referente à média e outro referente ao desvio padrão. Quais são as conseqüências práticas de cada tipo de erro em cada um dos problemas? Qual é o problema dual de cada um dos problemas?

38) Vários japoneses jovens (com 30 anos no máximo) que emigram para o Brasil se nacionalizam. Em uma consulta ao conculado japonês, você é informado de que a percentagem dos que se nacionalizam é cerca de 80%. Você acha que esta informação é um tanto exagerada. Monte um problema de teste de hipótese. Qual é a conseqüência prática de cada tipo de erro? Qual é o problema dual?

39) Qual é a utilidade da função poder? Qual é a sua relação com o nível de significância?

40) O que são métodos de inferência estatística? Exemplifique.

41) Na Zero Hora de 4/3/91, de um artigo de Moacyr Scliar: "A doença representa, sobretudo, o imprevisível. Um técnico pode garantir o conserto de um aparelho, mas o médico não pode fazer o mesmo em relação ao organismo enfermo. No confronto com nossa finitude e fragilidade ... quando o médico cai na defensiva, o resultado é pior: uma profusão de exames inúteis, ou, ao contrário, procedimentos que deveriam ser realizados e não o são". Discuta o texto e explicite suas impressões sobre ele.

LIÇÃO 17 – MODELOS

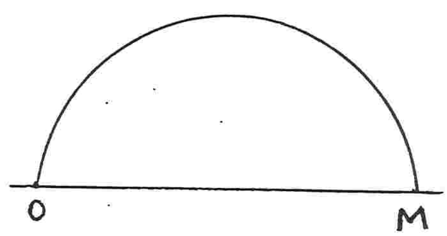
17.1 – Teoria Matemática do Aprendizado

A partir de 1950, a Teoria Matemática do Aprendizado desenvolveu-se de acordo com um ponto-de-vista mais abrangente que utiliza o enfoque probabilístico, hoje bastante comum nas Ciências Humanas.

O modelo matemático mais antigo de aprendizado apareceu em 1907. Ele foi o precursor de vários outros que também foram criados baseados na suposição de que a aquisição de conhecimentos acontece com taxa constante, portanto não considerando que ela seja proporcional ao que já foi aprendido. Também foram criados modelos que consideravam a taxa de aquisição de conhecimentos proporcional ao que já foi aprendido e modelos que consideravam a taxa de aquisição de conhecimentos proporcional ao produto do que já foi aprendido pelo que ainda falta aprender.

Um MODELO é uma estrutura puramente formal e uma TEORIA é uma interpretação dada a um fragmento da realidade, a um pedacinho do mundo real. Teorias diferentes podem conduzir a um mesmo modelo e, portanto, os sistemas matemáticos podem ser encarados como teorias portáteis, que podem ser levadas para qualquer área do conhecimento humano.

Em geral, uma CURVA DE APRENDIZADO representa o erro cumulativo em função do número de tentativas. A aceleração destas curvas é quase sempre negativa, ou seja, elas são do tipo da figura ao lado. Tais modelos podem também ser usados para descrever outros sistemas como, digamos, uma epidemia que se espalha rapidamente quando há muitas pessoas saudáveis, mas que vai diminuindo rapidamente na medida em que a quantidade de pessoas saudáveis vai se esgotando. Tais sistemas são geralmente autocatalíticos, pois estimulam-se e consomem seus recursos.



Os modelos iniciais criados na Teoria Matemática do Aprendizado não proviam equações empíricas que se ajustavam à grande quantidade de dados já disponíveis. O primeiro trabalho que forneceu um embasamento para a curva de aprendizado foi de Thurstone (1928). As variáveis principais que ele considerou foram a prática, representada por tentativas, e a aquisição, representada pela probabilidade de êxito. Seu trabalho foi generalizado por Gulliksen (1943), que utilizou um parâmetro para a punição e outro para a recompensa, numa curva cuja equação é $U = \alpha\left(1 - (\beta/\omega + \beta)^{\alpha/\beta}\right)$, onde U é o erro cumulativo, ω é o número de tentativas, α é o parâmetro devido à punição e β é o devido à recompensa. Gulliksen realizou seu experimento com 7 cobaias e esta equação deu um ajuste quase perfeito. Esta equação abrange os casos especiais nos quais o aprendizado ocorre somente através de punições.

17.2 – Psicologia Matemática

A Teoria da Informação, a Teoria dos Jogos e a Teoria das Decisões Estatísticas são áreas relativamente novas da Matemática que contribuíram para esta abordagem da Psicologia, que tem se caracterizado principalmente pelo uso de modelos matemáticos para representar fenômenos psicológicos.

Entre as opiniões sobre o que é a Psicologia Matemática estão: é a última esperança da psicologia científica, é um saco de fraudes com alguma utilidade, é uma formulação sofisticada do óbvio. Vamos encará-la como uma tentativa de usar métodos matemáticos para investigar problemas de psicologia. A Psicologia Matemática não é, portanto, definida em termos de conteúdos como a Psicologia do Aprendizado, a Psicologia Social ou a Psicologia da Percepção, mas sim como um estilo de investigação empregado em algumas destas áreas ou em várias delas. Mais ainda, não é um estilo uniforme, pois usa uma grande variedade de métodos.

Vamos relacionar a seguir alguns experimentos clássicos onde variáveis e relações psicológicas foram medidas e expressas de forma matemática, seus autores e o ano de publicação.

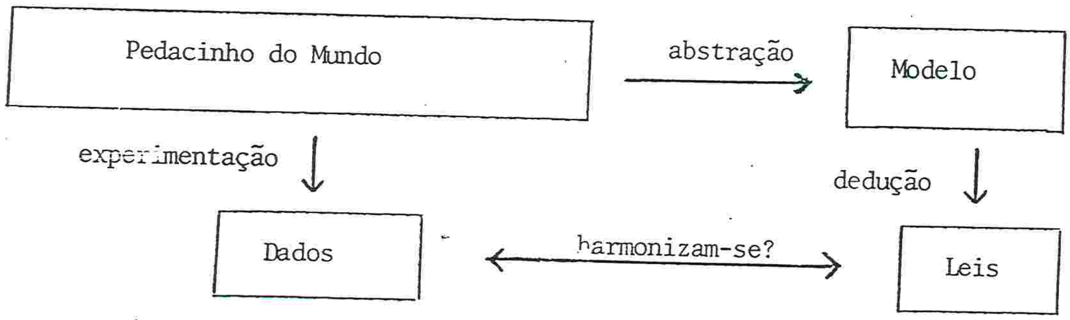
1. Sensações x Impulsos Físicos, Fechner (1860);
2. Inteligência, Spearman (1904);
3. Atitude, Thurstone (1928);
4. Psicologia Social, Lewin (1936);
5. Princípios Básicos do Aprendizado, Hull (1943).

17.3 – Modelos Matemáticos

Da mesma forma que diversas linguagens podem ser usadas para expressar a mesma idéia, diversos modelos formais podem ser usados para representar o mesmo fenômeno empírico. Há modelos geométricos, algébricos, probabilísticos etc.

Pedacinho do Mundo	⟷	Modelo
Real	⟷	Abstrato
Seres	⟷	Entes
Relações entre os Seres	⟷	Relações entre os Entes

Após concluído um modelo, um investigador obtém com o auxílio da lógica conclusões formais que são válidas no modelo. Se os dados experimentais se harmonizam com as leis do modelo, ele é aceito.



Do contrário, ele deve ser reformulado. A essência desse diagrama é caracterizar a investigação como uma tentativa de construção de um modelo de um pedacinho do mundo, o qual é testado através da comparação de suas leis (conseqüências) com os dados observados. Tais testes têm aspectos lógicos e aspectos estatísticos. Por exemplo, o modelo de Poisson pode ser aceito para um fenômeno aleatório discreto com resultados possíveis 0, 1, 2, ... somente se a média e a variância dos dados observados foram aproximadamente iguais, uma vez que no modelo elas são iguais.

Nas ciências experimentais um modelo pode apenas ser rejeitado com base nos dados e nunca provado. A adoção de um modelo acarreta certas conseqüências: suas leis lógico-formais. Se os dados experimentais não se harmonizam com tais conseqüências, o modelo é inadequado e não errado. Se elas se harmonizam, o modelo é aceitável e não certo.

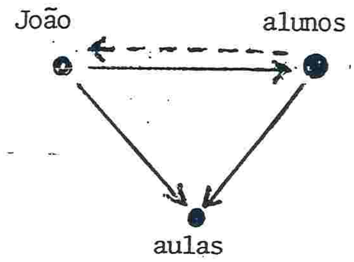
Uma relação muito importante entre modelos, $M_2 \subseteq M_1$, pode ser expressa por qualquer uma das seguintes sentenças equivalentes: M_1 é mais fraco do que M_2, M_1 é mais geral do que M_2, M_2 é mais forte do que M_1, M_2 é um caso especial de M_1. Por exemplo, se M_1 trata a variável X como contínua, M_2 a trata como tendo distribuição normal e M_3 a trata como tendo distribuição normal com média 65, então $M_3 \subseteq M_2 \subseteq M_1$. É óbvio que: se um modelo mais geral é rejeitado, então todos os seus casos especiais são também rejeitados.

A Matemática tem sido muito importante para o desenvolvimento científico, mas nem todas as suas aplicações foram frutíferas. As vantagens principais do uso da linguagem matemática são: sua generalidade (uma vez que é abstrata), sua precisão (possui uma semântica inequívoca) e sua possibilidade de permitir derivações lógicas (revelando desta forma algumas propriedades de seus elementos básicos e/ou algumas relações entre estes elementos básicos que podem estar bem ocultas).

Uma mensuração é uma correspondência entre um sistema empírico e um sistema numérico. O sistema numérico representa o empírico quando as relações que consideramos relevantes entre os elementos deste sistema são refletidas por relações entre os números. Por exemplo, se x pesa 8 kg e y pesa 6 kg, o fato de " x pesar mais do que y " é refletido por " $8 > 6$".

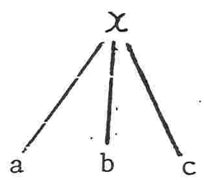
Sistemas formais não numéricos também podem ser modelos de sistemas empíricos. Há 3 tipos de tais modelos que se destacam atualmente nas Ciências Humanas: os provenientes da Teoria dos Gráficos, os da Teoria dos Conjuntos e os da Simulação. No que se segue nos limitaremos aos sistemas formais não numéricos provenientes da Teoria dos Gráficos.

Um GRÁFICO é uma coleção de pontos e linhas. Quando um gráfico é usado como um modelo (não numérico), os entes são representados por pontos e as relações entre os entes por linhas entre os pontos. As linhas contínuas geralmente representam relações positivas e as linhas descontínuas geralmente representam relações negativas. Por exemplo, nesta figura temos representado o sistema formado pelo professor João, que gosta das aulas e gosta de seus alunos, que gostam das aulas mas não gostam dele.



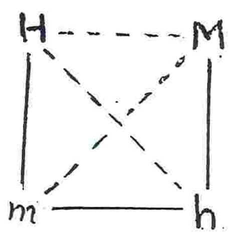
Um GRÁFICO NÃO DIRIGIDO é um cujas relações são todas simétricas, isto é, $x R y \Longleftrightarrow y R x$ para toda relação R. Caso contrário, temos um GRÁFICO DIRIGIDO. Nos gráficos não dirigidos, não é necessário uma seta de x para y e outra de y para x, basta uma linha entre x e y. Quando há apenas uma relação, temos um GRÁFICO NÃO SINALIZADO. Quando há duas ou mais, temos um GRÁFICO SINALIZADO. Nos casos mais comuns há exatamente duas, das quais uma é convencionada como sendo positiva e a outra como negativa.

EXEMPLO 1 - Um chefe x pode comunicar-se pelo interfone com três subordinados, que não podem comunicar-se entre si. Temos um gráfico não dirigido e não sinalizado.



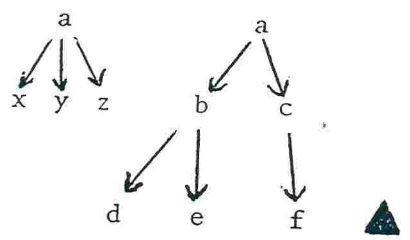
▲

EXEMPLO 2 - Um casal (H, M) convida um amigo h e uma amiga m para jantar. A anfitriã deseja sentar os 4 de tal modo que ela não sente ao lado de seu marido e nem dois homens nem duas mulheres sentem um ao lado do outro. Temos um gráfico não dirigido e sinalizado.

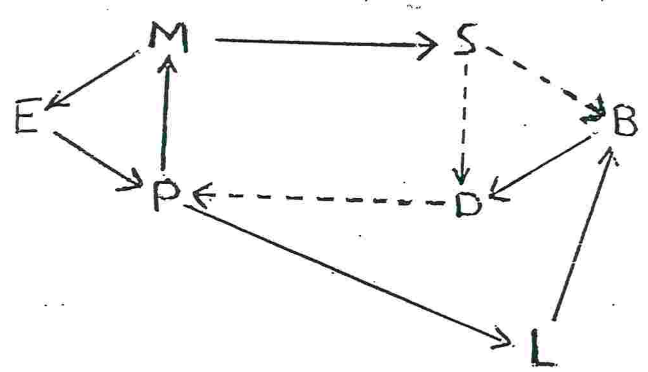


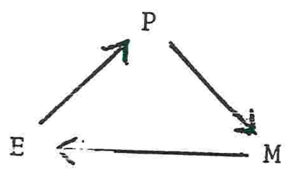
▲

EXEMPLO 3 - As hierarquias em instituições são exemplos de gráficos dirigidos e não sinalizados.

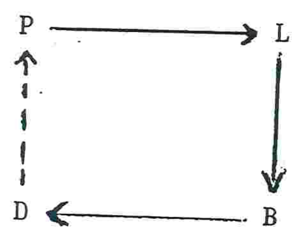


EXEMPLO 4 - Vamos representar por P o número de pessoas em uma cidade, por L a quantidade de lixo por unidade de área, por B a quantidade de bactérias por unidade de área, por D o número de doenças, por M a modernização, por E a emigração para a cidade e por S os recursos sanitários. Vamos considerar positiva a relação " x incrementa y " e negativa a relação " x inibe y". Temos aqui um gráfico dirigido e sinalizado.





ciclo com auto-alimentação positiva



ciclo com auto-alimentação negativa

Um ciclo tem AUTO-ALIMENTAÇÃO POSITIVA se, e somente se, ele tem um número par de relações negativas. Um ciclo tem AUTO-ALIMENTAÇÃO NEGATIVA se, e somente se, ele tem um número ímpar de relações negativas.

17.4 – Modelos Estatísticos

Sempre que usamos alguma técnica de Inferência Estatística estamos pressupondo a existência de algum MODELO ESTATÍSTICO, isto é, de algum conjunto de distribuições (multivariadas, se estamos estudando mais de uma variável, que é o que quase sempre ocorre) das quais uma é, exatamente ou aproximadamente, a distribuição verdadeira (do vetor formado por todas as variáveis em estudo numa ordem determinada). A escolha do modelo pode estar fundamentada em estudos teóricos, em conveniências matemáticas, em levantamentos já realizados das mesmas características em populações similares, em levantamentos já realizados de características similares na mesma população, em realizações passadas do mesmo experimento etc.

EXEMPLO 5 - Vamos considerar a população formada por todos os alunos matriculados em um determinado colégio, a variável X como o QI dos alunos da segunda série do segundo grau e a variável Y como o QI dos alunos da terceira série do segundo grau. Pode ser razoável supormos que, se a distribuição verdadeira de Y é aproximdamente Normal, uma conclusão que pode ter sido obtida por levantamentos já realizados em função dos exames vestibulares em vários anos anteriores, então a distribuição verdadeira de X também é aproximadamente Normal, possivelmente com outra média, mas com o mesmo desvio-padrão. ▲

Há autores que omitem ou não dão relevância aos modelos estatísticos, mas o fato é que qualquer técnica de Inferência Estatística é construída teoricamente pressupondo algum modelo e tem validade assegurada, em princípio, apenas neste modelo. Em termos práticos, qualquer técnica de Inferência Estatística pode ser empregada se, e somente se, alguma das distribuições (multivariadas) do modelo ao qual a técnica se aplica na teoria é, exatamente ou aproximadamente, a distribuição verdadeira (do vetor formado por todas as variáveis em estudo numa ordem determinada).

EXEMPLO 6 - Nem a altura nem o peso de qualquer população finita de seres humanos tem distribuição exatamente Normal, pois não podem assumir quer infinitos valores quer valores negativos. ▲

As técnicas estatísticas denominadas TESTES DE AJUSTAMENTO têm como objetivo verificar se determinado modelo estatístico e dados coletados podem ser considerados ajustados, isto é, se o modelo expressa satisfatoriamente o modo como o vetor ... varia na população, isto é, se uma das distribuições do modelo expressa satisfatoriamente tal variabilidade. No caso de vários modelos se ajustarem aos dados ou, conforme preferirmos, dos dados se ajustarem a vários modelos, surge o problema da escolha do melhor modelo, mas é necessário tornar preciso tecnicamente o significado da palavra "melhor".

Um modelo estatístico é uma estrutura matemática e, portanto, pode ser criado sem que uma de suas distribuições (multivariadas) seja, exatamente ou aproximadamente, a distribuição verdadeira (do vetor formado por todas as variáveis em estudo numa ordem determinada) em alguma população. Além disso, o fato de um modelo estatístico ser útil, isto é, satisfatoriamente adequado, para representar a variabilidade de determinadas características em determinadas populações não implica que ele seja útil apenas para tais características nem que ele seja útil para estas características em todas as populações.

EXEMPLO 7 - O Modelo Geométrico é exato para representar o número de realizações independentes de um experimento de Bernoulli até a ocorrência do primeiro sucesso, mas pode também ser adequado para representar o número de acidentes diários em determinadas indústrias nos dias em que há acidentes, o número de filhos dos casais (de determinada população) que têm filhos etc. ▲

EXEMPLO 8 - Consideremos uma população formada por 5 indivíduos de 1,64m de altura, 5 indivíduos de 1,65m de altura, ..., 5 indivíduos de 1,89m de altura. A distribuição verdadeira da característica altura nesta população é aproximadamente Uniforme com $\alpha = 1,635$m e $\beta = 1,895$m. Portanto, o Modelo Normal é inadequado, apesar de ser o mais geralmente utilizado quando estudamos a característica altura em populações humanas. ▲

As preocupações teóricas e práticas com o uso afoito, indiscriminado e dogmático do Modelo Normal para representar a variabilidade de variáveis têm motivado boa parte dos estudos teóricos da Estatística nas últimas décadas. Uma parte considerável dos resultados obtidos em tais estudos, junto com resultados de estudos teóricos motivados por outras razões, têm sido divulgados sob o polêmico nome de ESTATÍSTICA NÃO PARAMÉTRICA. Outra parte não tão considerável, sob o nome de ROBUSTEZ.

APÊNDICE 1 - UM ESBOÇO DE UMA CRÍTICA À CIÊNCIA

Assim como falamos em biosfera, em estratosfera e em ionosfera, também podemos falar em mitosfera, com a intenção de nos referir à atmosfera intelectual e afetiva própria de cada época. Com objetivos que parecem estar bem definidos, o homem comum é levado a aceitar como evidentes, ou pelo menos como definitivamente adquiridas, teorias e idéias de outros homens, alguns já há muito sepultados.

Idéias não são fatos e as teorias não são a realidade. As teorias são apenas modelos considerados satisfatórios para representar determinadas situações, mas nenhuma delas pode ser considerada uma verdade suprema, uma verdade evidente por si mesma ou uma verdade demonstrada. Acontece que os leigos, após ouvirem inúmeras vezes que as verdades científicas são a realidade, acabam acreditando.

A ciência chega somente até os fenômenos de experiência e não a um em si qualquer. Como,então, poderia captar a natureza íntima das coisas? Um exemplo bastante elucidativo do que é a ciência está na seguinte anedota: "Um cientista tirou uma perna de uma aranha. Mandou que a aranha andasse e ela andou. Retirou outra perna da aranha. Mandou novamente que ela andasse e ela andou.Continuou retirando as pernas, uma por uma, e mandando que a aranha andasse, o que ela fazia cada vez mais penosamente. Ao ser retirada sua última perna, a aranha não obedeceu à ordem de andar. O cientista então divulgou ao mundo sua descoberta experimental: as aranhas ficam surdas sem as pernas."

Quem não aceita parte da mitosfera de sua época não deve se sentir anormal e tampouco ser rejeitado pelos demais, pois o direito de pensar não deve ser concedido somente àqueles que pensam o que nós pensamos. Além disso,muitos dos que hoje são considerados grandes cientistas foram dissidentes em suas épocas. O oposto também é deplorável: a supervalorização do novo e o desprezo do tradicional. No aspecto social, podemos citar um recorte do caderno "Letras & Livros" do Correio do Povo de 3/4/82: "... neste fim-de-século, cuja principal característica política é a insubmissão a todos os valores que anteriormente ordenaram e hierarquizaram a sociedade humana".

Cada ciência é uma linguagem, isto é, um sistema de sinais que tenta descrever a realidade oculta sob a aparência das coisas. Esta realidade é lentamente desvendada pela investigação e jamais atingida de maneira absoluta. Cada ciência possui não apenas esta função externa ou descritiva, mas também uma função interna ou reflexiva, e será tão mais perfeita quanto mais inequívoca a sua semântica e quanto mais apurada a sua sintaxe. Em conseqüência disso, cada ciência deve ser entendida primeiro em seu aspecto sintático para depois ter seu aspecto semântico utilizado. Outrossim, cabe aqui salientar enfaticamente a forte relação existente entre pensamento e linguagem, a ponto de não se saber se é o pensamento que precede as palavras ou se são as palavras que estimulam e fecundam o pensamento. Esta idéia foi aproveitada por G. Orwell no livro "1984".

Existem livros que endeusam a Ciência (Arquitetos de Idéias, E.R.Trattner Ed. Globo, 1967) e há os que criticam a Ciência à maneira católica (Psicoanalisis y Marxismo, L. Jugnet, Ed. Cruz y Fierro, 1977) ou à maneira filosofia oriental (É a Ciência uma Nova Religião? , E. Bono, Ed. Civilização Brasileira, 1971). Vamos exercitar um pouco nosso senso crítico através de um exemplo: "Psicologistas experimentais estão estudando (incomodando, perturbando, matando) animais com o objetivo de saber mais sobre a sede. Já há alguns fatos (por honestidade, deve-se dizer idéias ou teorias) aceitos e numerosas hipóteses (várias delas logo serão 'comprovadas' e passarão a ser fatos). O processo supostamente simples de responder à sede bebendo algum líquido não é tão elementar como se tem pensado (??????); ele é tão complicado que mesmo décadas de investigação científica (a palavra 'científica' aqui imprime respeitabilidade a tais investigações) não tem conseguido obter todos os seus fatos (que durarão precisamente até serem substituídos por novos fatos recém descobertos)".

O que transcrevemos a seguir é a introdução de um livro de Isaac Asimov: "Este livro foi publicado pela primeira vez em 1956 e a descrição da superfície de Mercúrio corresponde às convicções astronômicas daquela época. Desde 1956, porém, os conhecimentos do Sistema Solar progrediram muito por causa do uso de radar e mísseis. Em 1956, acreditava-se que Mercúrio sempre só apresentasse o mesmo lado ao Sol e que, por conseqüência, havia um lado continuamente iluminado pelo Sol e um outro lado lado sempre escuro, com regiões fronteiriças onde a iluminação e a escuridão se alternavam. Em 1965, os astrônomos estudaram reflexos de raios radar da superfície de Mercúrio e, com grande surpresa, constataram que isso não correspondia à realidade. Enquanto Mercúrio girava em torno do Sol em 88 dias, também revolvia em volta de seu próprio eixo em 59 dias. Isso significava que todos os lados de Mercúrio ficavam expostos ao Sol numa certa época e que, afinal, não existia nenhum 'lado escuro'. Espero que os leitores gostem deste livro apesar disso, mas faço voto que eles não tomem ao pé-da-letra e aceitem como fatos uma parte do material que era 'certo' em 1956, mas que agora já está superado".

Agora vamos citar Fernando Pessoa: "A única coisa certa em Ciência Social é que não há ciência social (...). Não há uma lei social até hoje descoberta; há só teorias e especulações que, por definição, não são ciência. E onde não há ciência, não há universalidade".

O trecho a seguir é de Millor Fernandes (Veja, 1/12/1982): "Toda hora eu vejo nos jornais, revistas, televisão e na rua, pessoas muito 'livres' de preconceitos e ... E, no entanto, todas acham que a Terra gira em torno do Sol. Por que? Porque Galileu 'provou' isso um dia. Mas provou para quem? Pode ser que tenha provado para os cientistas. O homem comum e mesmo nós, intelectuais em geral, aceitamos apenas. Sem pensar. Preconceituosamente. Como antes de Galileu aceitávamos que o Sol girava em torno da terra. Mas entre Galileu - de cujas 'provas' nunca tomamos conhecimento - e a realidade, que literalmente salta (gira) aos nossos olhos, temos que acreditar é em nossos olhos. Nossos olhos vêem, com absoluta certeza, que o Sol nasce ali e morre ali, girando em torno de uma Terra parada. O resto, só com provas em contrário".

"Não há verdades comprovadas, mas há mentiras evidentes". Esta frase pode ser uma síntese da Ciência, pois a experimentação é capaz de derrubar uma verdade, mas não pode colocar outra em seu lugar. Quem o faz é o cientista, que nem sempre é um idealista completamente desinteressado em dinheiro e sem um ego para satisfazer. O modo como os homens da ciência buscam o conhecimento talvez seja bem ilustrado através do seguinte conto: "Um homem viu Nasrudin procurando algo no jardim e lhe perguntou o que havia perdido. Ao obter a resposta, abaixou-se também para ajudar Nasrudin a procurar chaves. Após um certo tempo, perguntou:

- Onde foi exatamente que o senhor as perdeu?
- Dentro de casa.
- Então por que está procurando-as aqui?
- Há mais luz aqui no jardim do que dentro de casa."

Novamente Millor Fernandes (Correio do Povo, 10/7/83): 'Meu amigo, um tanto esquerdista, tendo ouvido falar nas admiráveis experiências do sábio Pavlov, resolveu praticá-las também. Colocou seu cão preso numa gaiola, até o animal ficar esfomeado. Quando o animal já estava uivando de fome, meu amigo chegou perto dele, deu três latidos breves e colocou o prato de comida na frente do cão. O cão, esfomeado que estava, devorou a comida. A intenção científica do meu amigo era, claro, ensinar o cão a pedir comida com três latidos seguidos. Treinou isso durante um mês, dois meses, três meses. Diariamente chegava perto do cão, dava três latidos no mesmo tom, e deixava o prato, que o cão devorava. Mas, afinal, desistiu de criar o reflexo condicionado no animal; parece que este nunca tinha realmente ouvido falar de Pavlov: não aprendia mesmo. Um dia meu amigo chegou junto dele, irritado com a sua incapacidade de aprendizagem e deixou o prato, exclamando: 'Toma, imbecil. Come o que quiser.' Mas quando voltou, no dia seguinte, o cachorro não tinha comido coisa alguma. E, com o passar dos dias, meu amigo descobriu uma coisa terrível: o cão, agora, se recusa a comer sem que o meu amigo lata três vezes seguida." E Millor Fernandes finalizou com a moral da história: "A Ciência é para os cientistas."

APÊNDICE 2 - REFLEXÕES

1. O objetivo da Teologia é averigüar a índole de Deus. É com o apoio da Estatística que as leis da esfera social podem ser estudadas e codificadas, e certos aspectos da índole de Deus ali revelados. O estudo da Estatística é, portanto, um serviço religioso.

2. Mensuração não necessariamente significa progresso. Falhando a possibilidade de medir o que se deseja, a lascívia por medição pode, por exemplo, simplesmente causar a mensuração de algo parecido - talvez esquecendo a diferença - ou causar a desconsideração de alguns fatores porque eles não podem ser medidos ...

3. Por uma pequena amostra, nós podemos julgar o pedaço inteiro.

4. Vamos pensar ... Uma pessoa entra neste mundo justo agora ... O Sol seria provavelmente, o primeiro objetivo que chamaria sua atenção; mas após tê-lo perdido na primeira noite, ela ignoraria completamente se o veria novamente alguma vez ... Mas ela veria um segundo aparecimento do Sol, ou seja, um primeiro retorno, e, quando o Sol sumisse de novo, deveria surgir nela uma esperança de um segundo retorno. Esta esperança aumentaria com o número de retornos ... Mas nenhum número finito de retornos poderia ser suficiente para produzir uma certeza absoluta.

5. A curva descrita por uma simples molécula de ar ou vapor é regulada de uma maneira tão exata como uma órbita planetária; a única diferença entre elas é a que vem da nossa ignorância.

6. A grande tragédia da ciência: o assassinato de uma hipótese belíssima por um fato feioso.

7. Nasrudin estava caminhando na rua principal de uma cidade, jogando a esmo migalhas de pão. Seus conhecidos perguntaram:

   - Que fazes Nasrudin?
   - Mantenho os tigres afastados.
   - Não têm havido tigres por aqui faz centenas de anos.
   - Claro. Funciona, não?

8. As palavras são apenas dedos apontando a Lua. Deve-se olhar a Lua e não os dedos.

9. A utilidade de uma xícara está em sua capacidade de esvaziar-se.

10. Um pai diz para seu filho que vê tudo em dobro:

- Filho, tu vês dois ao invés de um.
- Como pode ser? Se assim fosse, pareceria ter quatro luas lá em cima ao invés de duas.

11. Os cientistas persuadiram os possessores de que possuiriam ainda mais e os governantes de que governariam ainda mais se recorressem a eles, conquistando rapidamente uma posição acima da riqueza e do poder. De que forma? Em primeiro lugar, introduzindo em tudo uma infinita complexidade.

12. Examine as condutas dos governos. Seus métodos fariam vergonha a qualquer assembléia de acionistas e arrastariam para a falência qualquer empresa particular. Qual é o atrativo que há para um gênio em vender seu cérebro aos seus governantes medíocres?

13. As monarquias pretendiam ter um poder de origem sobrenatural. O rei, os figalgos, os ministros e os responsáveis tudo faziam para serem notados através de seus vestuários, suas residências etc. Após a Revolução Francesa, o poder impregna-se de teorias abstratas e o governo oculta-se. Nas democracias modernas é difícil definir com exatidão quem governa, quem é quem e quem decide o quê. O mundo moderno disfarça os seus chefes como os insetos que parecem ramos ou folhas de árvores.

14. É preciso prestigiar o investigador isolado. De fato, acabamos por acreditar que o progresso dos conhecimentos já não é possível sem equipes numerosas, sem uma enorme aparelhagem, sem um financiamento considerável.

15. A última finalidade da ciência é a transmutação do próprio cientista, um lento caminhar em direção à libertação do espírito.

16. Um dos grandes mitos deste século é a certeza que os homens civilizados têm de saber tudo a respeito do Universo em que vivem.

17. O conhecimento científico não é objetivo. É como a civilização, uma conspiração. Repudiam-se uma quantidade de fatos porque eles alteram os raciocínios correntes. Vivemos em um regime de inquisição em que a arma usada é o desdém acompanhado de risos. Não temos liberdade para duvidar da ciência.

18. As universidades não foram feitas para ensinar o que está nos livros, por mais avançados que estes sejam. Elas foram feitas para preparar gente capaz de escrever os livros de amanhã, gente capaz de responder às perguntas que os livros de hoje não respondem.

APÊNDICE 3 - CONSIDERAÇÕES PESSOAIS

O objetivo de um curso universitário não é informar, pois qualquer quantidade de anos é insuficiente para alguém adquirir toda a informação existente na sua profissão e, ademais, ela está arquivada em algum lugar, podendo ser localizada quando se fizer necessário. O objetivo é formar, ou seja, entre outras coisas, dar a capacidade de analisar e de usar adequadamente tal informação.

A Estatística como receituário está nos livros, mas deve ser ensinada em aula como um modo de pensar, isto é, como uma linguagem. Aprender novas técnicas é fácil, mas aprender novas idéias é sempre trabalhoso.

Cada disciplina deve ter sua própria razão de ser e não deve ser neutra na formação do aluno. Nunca deve ser vista apenas como um instrumento, ou apenas como um pré-requisito para outra disciplina.

Todo ser humano deve selecionar, entre os estímulos que recebe do mundo exterior, aqueles aos quais vai responder. Um universitário deve ser bem mais criterioso do que os demais, pois o estudo deve ser freqüente e natural. Uma média semanal de duas horas é o mínimo admissível para cada disciplina.

A seriedade profissional de um professor nas avaliações nada tem a ver ver com a camaradagem pessoal existente com seus alunos. Além disso, o processo de avaliação deve ser o mais contínuo possível e não a etapa final, da mesma forma que os instrumentos de avaliação devem ser apenas meios e não os fins.

As creches são destinadas a entreter crianças, A Universidade, ao menos explicitamente, não se destina a entreter pós-adolescentes ou pessoas de faixas etárias posteriores, apesar de estar infestada de telas-e-teclas, similares a brinquedos eletrônicos. Agora, porém, as brincadeiras são sérias ... e "de verdade". Afinal, acreditamos que somos adultos.

Os estudantes que entendem as avaliações meramente como "chances" de fazer pontos, ao invés de estudarem permanentemente com afinco, somente "dão uma olhada", nas poucas horas que antecedem as provas, na (parte da) matéria que (supõem) irá "cair".

O interesse demonstrado após as avaliações deve ser focado na dicotomia cognitiva "o que já sei – o que não sei e preciso aprender" e não na dicotomia lúdica "o que acertei – o que não acertei". A nota obtida, nestes casos, mede a eficácia da estratégia utilizada para substituir o verdadeiro estudo. Se ela não funcionou, será, mesmo, necessário (começar a) estudar ?

Um professor deve tentar provocar mudanças de atitudes em seus alunos, mesmo que (quase todos) eles (acreditem que) estejam vivendo num mundo que condiciona as pessoas ao recebimento de informações de descarte imediato.

LIÇÃO 16

(1) A rejeição de H_0, porque... (2) Sim, porque são baseadas em observações da realidade. (3) Não, mas é menos precisa e mais dispendiosa. (4) Não, porque uma é a negação da outra. (5) Não, porque o n.s. se refere somente à probabilidade de erro quando H_0 é rejeitada. (6) Nenhuma, porque um ocorre quando H_0 é verdadeira e o outro ocorre quando H_1 é verdadeira. (7) O determinístico, que pode ser considerado um caso especial do aleatório. (8) A percentagem de mães deve pertencer ao intervalo, pois foi empregada uma técnica que acerta com probabilidade 0,99. (9 a) Os que têm Eqm pequeno. (9b) Os que têm Fpo grande quando H_1 é verdadeira. (10) Testando $H_0 : \mu_P = \mu_F$ vs. $H_1 : \mu_P > \mu_F$ e rejeitando H_0. (11) Quando ambos podem ser aplicados, devemos usar os não-paramétricos, porque... (12) O erro padrão ao quadrado é uma estimativa do Eqm. (13) Por ponto, por intervalo, por limite inferior e por limite superior. (14) (a) Só Deus sabe. (b) Acreditamos que sim. (c) Deve estar, afinal confiamos no coeficiente de confiança. (15) Não. O que acontece é que a rejeição de H_0 é quase totalmente confiável. (16) Fazendo-a coincidir com a hipótese alternativa de algum teste e rejeitando a hipótese nula. (17) Ver (8) acima. (18) Através do erro padrão. (19) Porque não sabemos se estão certas ou erradas. (20) Estimar parâmetros e testar hipótese. (21) A qualidade dos estimadores por conjunto. (22) Não tem sentido obtê-lo, pois o verdadeiro valor do parâmetro p é conhecido (= ½). (23) Não, é o contrário: são iguais no conteúdo e diferentes na forma. (24) Sim: "A matemática não é sua terminologia nem sua simbologia, mas as idéias que elas transmitem". (25) 49 ou mais. (27) Não. Implica que **acreditamos** na sua veracidade, mas admitindo a possibilidade de estarmos enganados. (28) Escolher uma técnica de amostragem adequada, escolher uma técnica de inferência adequada, não ser infeliz na seleção da amostra, usar um processo de mensuração adequado, não cometer erros de mensuração, de registro, de cálculo e de interpretação. (30) Nem todas. São de natureza probabilística. (31) Não. Todas são compostas. (32 a) 3,22 (32b) Impossível, pois falta Σx_i^2. (32c) Ver (22) acima. (33 a) $\theta = \mu + 3\sigma$ (33b) $\hat{\theta} = \hat{\mu} + 3\hat{\sigma} = \bar{x} + 3s$ (34) É uma estimativa da raiz quadrada do Eqm do estimador que a produziu. (35) Depende. Se for não viciado ou se tiver vício pequeno, a resposta é afirmativa. Se tiver vício grande, a resposta é negativa. (36) Depende da situação. (39 a) Medir a qualidade de um teste de hipótese, ou de sua região crítica, quando H_1 é verdadeira. (39b) O n.s. supera ou maximiza a Fpo quando H_0 é verdadeira. (40 a) São métodos que produzem, a partir de uma amostra, uma afirmação a respeito de toda a população. (40b) Os estimadores por ponto. (41) Nota-se uma analogia entre as duas atitudes extremas do médico e o tipo de erro que ele quer evitar com as decisões estatísticas e seus possíveis erros...