Material de apoio às aulas de Estatística Aplicada I

Disciplina: INE 5111

Profa. Vera do Carmo Comparsi de Vargas

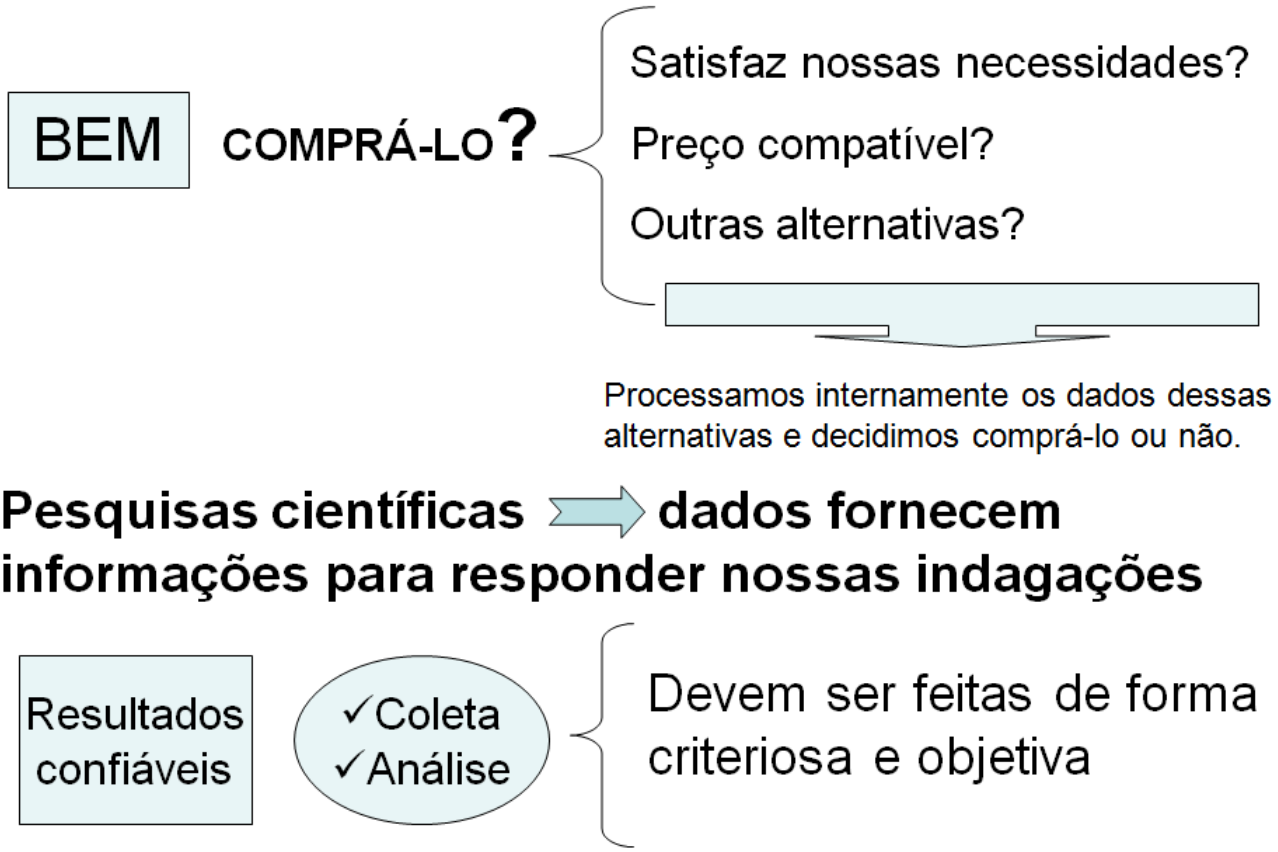
2020-1

Definição de Estatística

“Estatística é a Ciência que permite obter conclusões a partir de dados” (Paul Velleman**).**

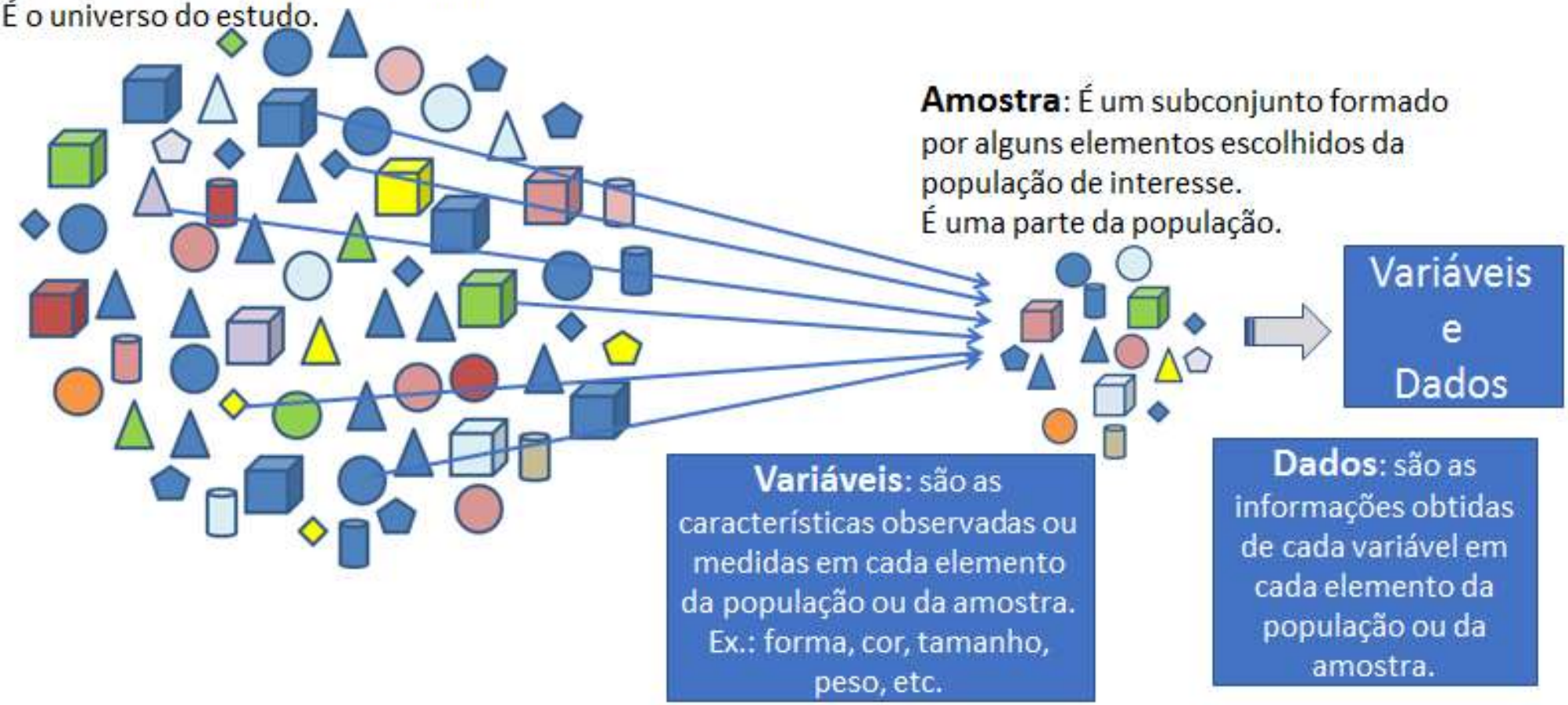
Ciência que parte de perguntas e desafios do mundo real:

Decisões do dia a dia baseadas em dados



População: É o conjunto de elementos que apresentam as características de interesse a serem investigadas para os quais desejamos que nossas conclusões sejam válidas.
É o universo do estudo.

Amostragem: Reúne os métodos necessários para coletar adequadamente amostras representativas e suficientes para que os resultados obtidos possam ser generalizados para a população de interesse.



Amostra: É um subconjunto formado por alguns elementos escolhidos da população de interesse.
É uma parte da população.

Variáveis e Dados

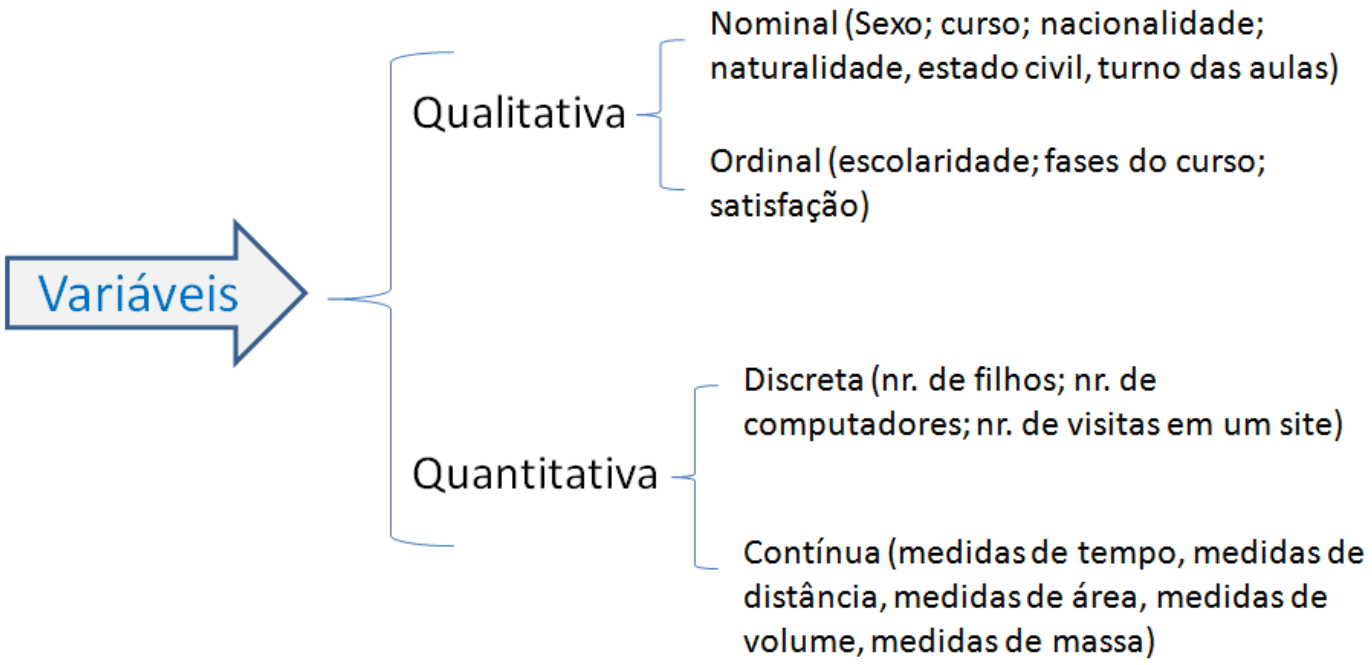
Variáveis: são as características observadas ou medidas em cada elemento da população ou da amostra. Ex.: forma, cor, tamanho, peso, etc.

Dados: são as informações obtidas de cada variável em cada elemento da população ou da amostra.

Conjunto de dados:

Identificação	Forma	Cor	Tamanho (área cm^2)	Peso (g)
Item 1	Círculo	Azul	0,05	2,8
Item 2	Cubo	Amarelo	0,08	4
Item 3	Triângulo	Verde	0,03	2,5
Item 4	Pentágono	Lilás	0,025	1,8
Item 5	Cilindro	Marrom	0,07	1,5

Classificação das variáveis:



POPULAÇÃO:

-**População finita: quando se conhece o seu tamanho (N)**

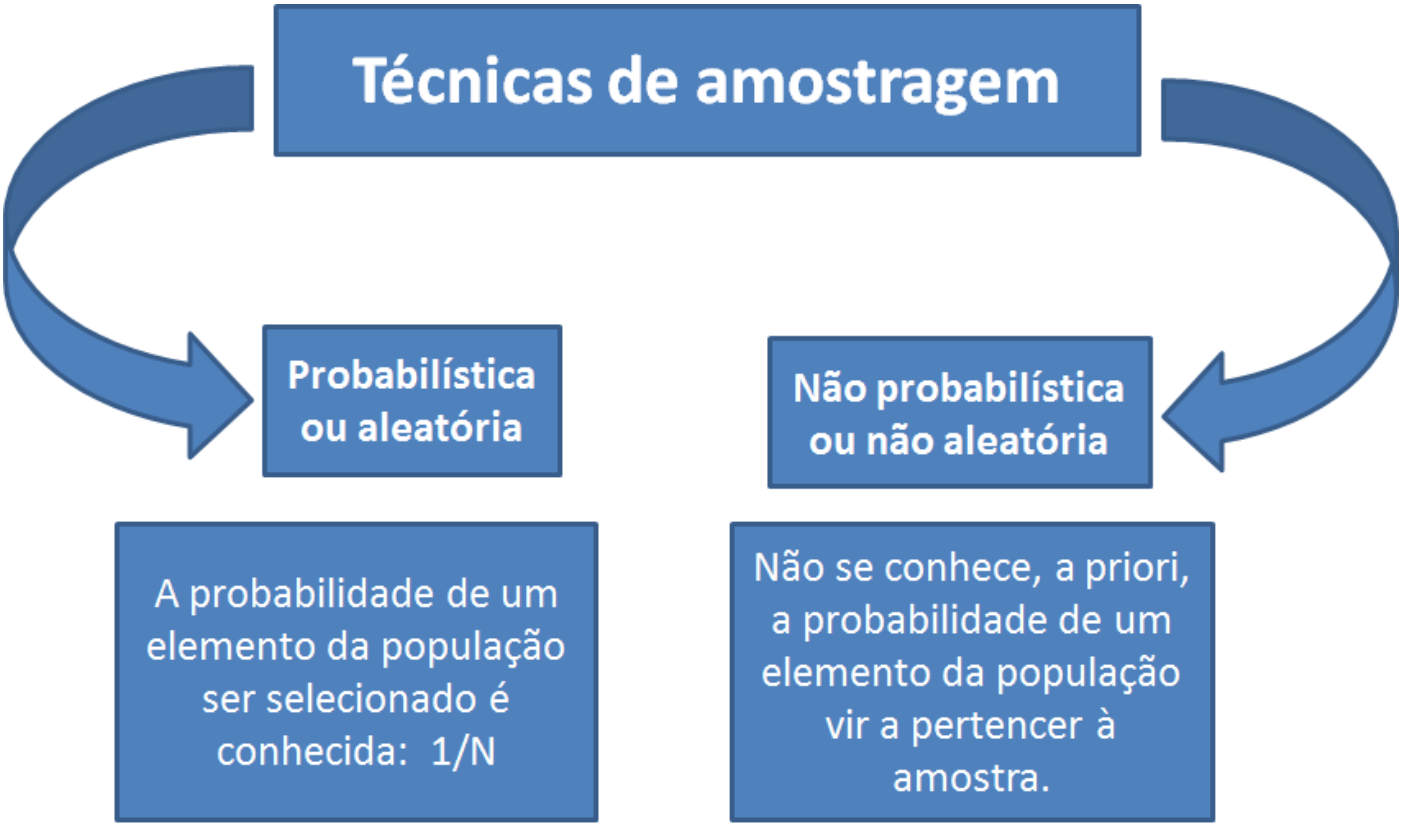
- **População infinita: não há como determinar o seu tamanho (N)**

Amostra (n) é um subconjunto finito e representativo da população.

Censo é o exame de **todos** os elementos que compõem uma população.

Amostragem X Censo

Amostragem é mais vantajosa:	**Censo é mais vantajoso:**
- População infinita; - Atualização; - Testes destrutivos; - Confiabilidade dos dados; - Tipo de informação; - Operacionalidade.	- População pequena; - Precisão; - Dispõe de informação completa.


Técnicas de amostragem
Probabilística ou aleatória
Não probabilística ou não aleatória
A probabilidade de um elemento da população ser selecionado é conhecida: 1/N
Não se conhece, a priori, a probabilidade de um elemento da população vir a pertencer à amostra.

Características das amostragens

Aleatórias ou probabilísticas

É o processo de retirada das unidades de observação de uma população na qual cada unidade tem a mesma oportunidade de integrar a amostra;

Não há influência do pesquisador;

Equivale a um sorteio;

Permite a avaliação objetiva da variabilidade amostral ou erro amostral.

Não aleatórias ou não probabilísticas

Há influência do pesquisador, julgamento ou intencionalidade;

Não permite a avaliação objetiva da variabilidade amostral ou erro amostral.

Os resultados podem conter vieses e incertezas.

Tipos de amostragens aleatórias

- Aleatória simples
- Aleatória sistemática
- Aleatória estratificada uniforme
- Aleatória estratificada proporcional
- Aleatória por conglomerados

Amostragem aleatória simples

Sorteia-se, com igual probabilidade, um elemento qualquer da população. Repete-se o processo até obter o tamanho da amostra (n).

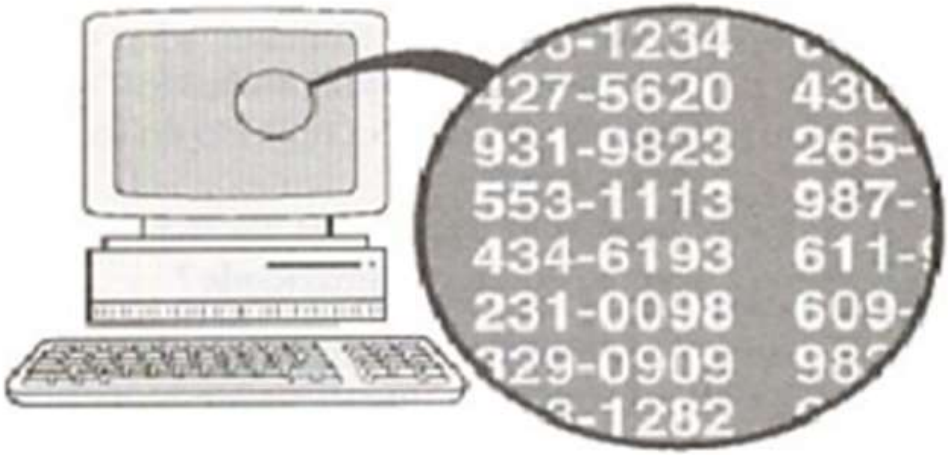


Tabela 1 Números aleatórios

59 58 48 36 47	92 85 05 08 65	47 49 10 41 05	10 75 59 75 99	17 28 97 99 75
53 26 21 50 21	37 93 85 52 86	86 22 75 34 37	69 85 25 03 78	50 26 18 25 10
07 02 16 58 67	05 32 93 87 84	31 30 62 78 60	59 90 24 22 07	74 43 43 56 91
92 87 67 56 36	58 58 16 88 16	17 83 52 09 99	86 17 20 95 93	01 46 77 18 11
90 57 05 58 96	84 33 68 15 87	28 18 08 76 89	94 60 94 48 76	92 93 49 13 91
24 26 56 02 33	33 21 75 54 04	96 28 85 78 11	54 01 92 86 36	65 19 45 97 79
20 09 49 50 27	33 86 85 59 39	02 25 60 56 26	01 11 24 44 15	58 00 54 54 09
22 74 50 39 12	83 91 03 38 78	85 56 78 41 44	26 04 12 13 50	38 15 61 02 51
10 45 36 09 86	07 68 31 98 41	98 17 56 93 84	16 01 48 99 36	44 61 71 69 67
09 82 11 18 29	96 19 12 47 26	26 01 14 78 55	33 11 13 56 95	68 66 57 90 33
04 63 02 45 50	61 91 02 14 07	57 36 29 12 74	89 47 84 89 69	13 85 22 66 83
55 93 05 63 30	40 05 51 03 31	68 15 33 85 87	94 80 24 96 62	31 38 95 35 38
66 15 07 64 38	16 44 52 26 42	34 65 99 71 63	87 22 04 62 15	76 94 00 00 77
96 31 72 41 94	47 03 44 73 77	96 17 02 97 50	26 67 60 63 57	66 81 92 03 20
07 10 58 83 63	35 47 34 05 38	92 26 05 33 40	91 23 43 68 72	29 74 60 67 01
04 47 64 02 49	10 52 21 00 80	40 56 68 97 32	43 46 70 65 08	96 52 25 29 44
56 24 53 31 96	65 42 53 27 78	23 30 61 34 18	56 59 23 69 27	83 66 60 03 12
98 15 27 91 71	24 15 28 61 91	83 49 05 82 54	53 59 30 25 19	36 31 31 56 58
36 96 23 77 26	79 74 28 12 16	08 88 07 28 71	45 43 40 07 66	11 26 38 51 87
66 01 53 03 67	92 27 27 17 54	31 23 30 42 83	85 78 21 68 34	86 33 77 84 40
48 07 09 48 65	92 33 41 97 63	48 97 19 86 81	10 85 42 84 49	03 82 01 82 88
95 44 86 84 32	09 03 56 46 96	64 51 33 75 10	29 00 99 23 82	92 31 77 08 17
91 73 15 42 46	72 21 07 34 11	92 70 89 58 54	11 30 93 38 29	00 53 93 14 09
08 35 79 86 83	06 89 37 82 12	81 14 08 82 04	91 88 04 86 36	18 10 09 78 99
37 20 97 09 96	86 34 77 09 31	04 38 18 79 61	68 66 47 40 35	40 16 50 22 54
79 14 72 97 40	90 98 64 42 25	72 95 89 98 59	03 73 02 95 47	34 85 74 60 90
58 55 07 49 26	08 02 70 20 14	57 17 20 89 16	07 86 05 38 61	69 48 78 18 62
77 93 74 07 34	23 49 25 23 87	43 93 35 93 02	80 94 57 16 22	73 67 28 75 37
91 82 56 78 91	47 22 60 09 32	67 02 21 71 61	12 83 08 40 00	52 23 47 46 58
53 66 43 91 44	19 05 53 26 31	89 52 31 98 20	03 70 03 61 07	52 79 97 75 92
91 03 23 35 58	48 22 68 98 07	12 20 88 41 89	19 00 56 88 74	96 71 20 52 46
70 35 43 62 20	81 20 95 72 99	80 91 40 17 51	26 71 79 23 17	01 25 48 07 82
93 85 01 86 56	78 48 74 55 63	62 09 64 35 47	08 70 04 66 86	08 91 83 42 94
75 40 86 33 31	96 06 26 53 07	41 58 96 29 23	17 71 66 60 72	07 18 47 73 75
37 15 68 73 37	31 76 55 39 13	49 61 13 83 90	53 47 54 53 52	80 30 40 35 21
35 88 34 83 04	71 67 75 40 83	99 97 96 83 32	16 04 27 99 31	49 80 34 34 95
73 06 78 79 97	28 86 29 45 91	76 44 64 99 81	33 95 06 94 26	85 78 57 43 12
94 70 05 36 32	38 44 59 60 01	13 74 03 30 33	24 79 77 71 87	41 57 07 96 68
09 65 41 62 93	63 28 60 59 28	29 08 69 81 67	60 57 53 64 28	12 24 35 23 49
12 39 50 50 09	22 70 54 75 38	78 56 79 26 62	79 37 83 33 92	33 30 61 41 90

Nota: Os espaços entre os números são apenas para facilitar a leitura, mas os números podem ser lidos com a quantidade de algarismos que se queira.

Exemplo:

1-Gustavo	2-Pedro	3-Giovanni	4-Caroline	5-Luan	6-Vitória	7-Olivia	8-Enzo
9-Antonella	10-Matheus	11-Paulo	12-Clarice	13-Marcelo	14-Breno	15-Camila	16-Beatriz
17-Melissa	18-Bento	19-Evellyn	20-Juliana	21-Otávio	22-Levi	23-Luana	24-João
25-Eduardo	26-Esther	27-Renan	28-Raul	29-Paulo	30-Vicente	31-Ana	32-Carlos

Quadro 1 – Estudantes do Curso A, da Universidade X, matriculados no semestre i do ano W.

Números aleatórios

5960	6399	6363	6773	6823	5192	1886	1364	9956	4243	7071	2125	5118	6461	2176	1800	8438	5675	7955	4428
0971	0750	8091	8095	6815	4081	3247	6153	4743	3677	1483	2910	4194	3889	7575	9145	5603	7091	8129	3059
5676	8029	1302	7434	0025	3581	3653	2382	7284	0597	6008	9438	0210	7016	3488	8372	6579	6397	9760	7807

Quadro 2 – Números aleatórios

Selecionar 3 alunos para serem os representantes da turma

No software R para gerar número aleatório digitar a função no RGui:

sample(1:32,3)

entre 1 e 32 gerar 3 números

Amostragem aleatória sistemática (periódica)

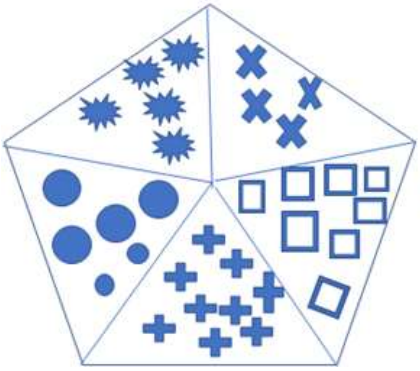
É o processo de seleção das unidade de observação que é realizado periodicamente utilizando um intervalo de seleção (i.s.).
O i.s. é determinado, para população finita, através da divisão do tamanho da população (N) pelo tamanho da amostra (n) a selecionar.

$$i.s. = N/n$$

Escolhe-se um ponto de partida (aleatório) e, sistematicamente, selecionam-se os demais.

Amostragem aleatória estratificada

A população é caracterizada por subgrupos , denominados estratos. Dentro dos estratos há homogeneidade e entre os estratos há heterogeneidade.



- **Estratificada uniforme**
- **Estratificada proporcional**

Análises

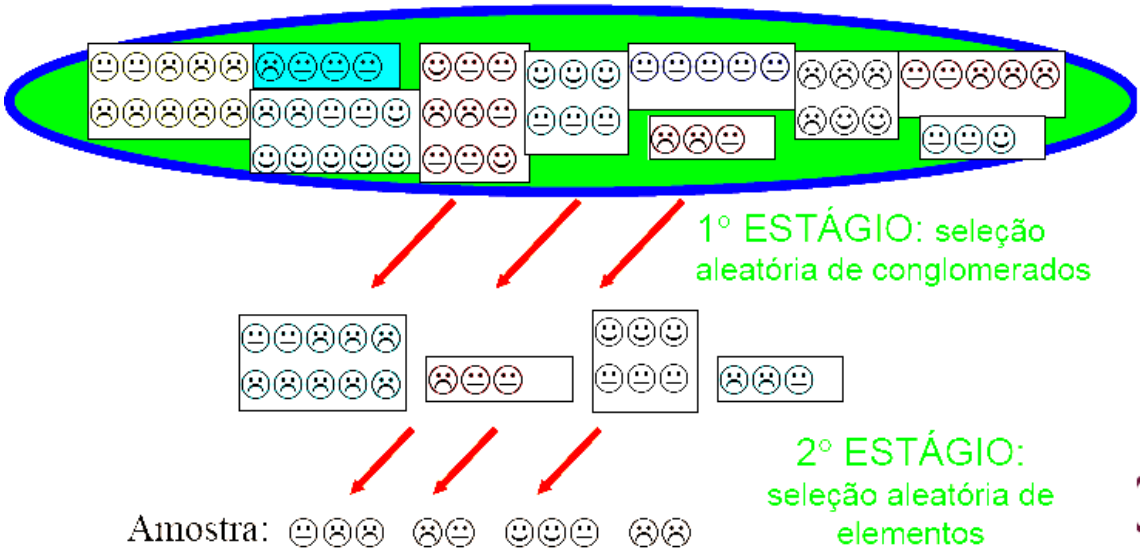
Dentro dos grupos
Homogeneidade das unidades

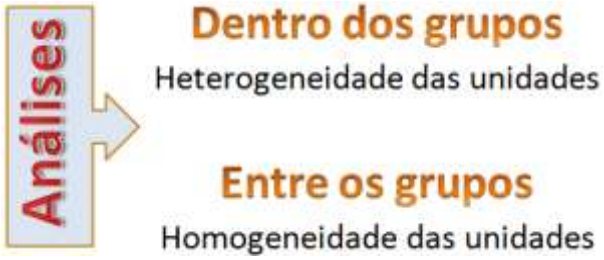
Entre os grupos
Heterogeneidade das unidades

Amostragem aleatória por conglomerado

Conglomerado ou agrupamentos (*clusters*).
Os elementos da população são selecionados aleatoriamente de forma natural por grupos (*clusters*).
Dentro dos grupos há heterogeneidade e entre os grupos há homogeneidade .

No primeiro estágio selecionam-se os conglomerados. Num segundo estágio observam-se todos os elementos dos conglomerados escolhidos, ou se faz nova seleção entre os extraídos no primeiro estágio e assim sucessivamente. Todas as seleções devem ser aleatórias.





Exercícios: Lista 1

Tamanho da amostra:

Depende:

1. **Do objetivo da pesquisa (uma média, uma proporção ou várias proporções)**
    - **Uma média: Ex.: O objetivo da pesquisa é saber o IA médio da turma**
    - **Uma proporção: Ex.: O objetivo da pesquisa é saber a satisfação dos estudantes com o seu curso**
    - **Várias proporções: Ex.: O objetivo da pesquisa é conhecer o perfil dos estudantes do curso**

2. **Da população a ser amostrada (o tamanho e as características)**
3. **Do nível de confiança**
    - **Definido pela comunidade científica da área para a qual a pesquisa se enquadra. Pesquisas de opinião usam o nível de confiança de 95% (de modo geral)**
4. **Da margem de erro**
    - **Definido pelo pesquisador. Pesquisas eleitorais têm usado 2%, 3%...**
    - **Implicação: quanto menor a margem de erro, mais precisas são as estimativas dos resultados, porém necessita-se de uma amostra grande**

Fórmula de cálculo para determinar o tamanho da amostra para o caso de amostragem aleatória simples

- Tamanho inicial da amostra:

$$n_0 = \frac{z^2}{4\,E_0^2}$$

- Tamanho final da amostra:

a) população grande (N > 20 n_0): nesse caso o tamanho final é igual ao tamanho inicial, isto é, $n = n_0$

b) população não muito grande (N ≤ 20 n_0): nesse caso o tamanho final é determinado usando a seguinte fórmula:

$$n = \frac{N.n_0}{N + n_0}$$

Exercícios:

1) Determine o tamanho necessário de amostras aleatórias simples para estimar várias proporções, nos casos a seguir:

a) nível de confiança de 95% e margem de erro de 2%

b) nível de confiança de 90% e margem de erro de 2%

c) nível de confiança de 95% e margem de erro de 3%

d) nível de confiança de 90% e margem de erro de 3%

Para cada caso a) b) c) e d) considere os tamanhos das populações: N=100; N=1.000; N=10.000; N=100.000

Como classificar uma pesquisa?

De acordo com Gil (2002)[1] as pesquisas podem ser classificadas quanto aos objetivos e quanto aos procedimentos técnicos.

Quanto aos objetivos: a) exploratórias; b) descritivas e c) explicativas

Quanto aos procedimentos técnicos utilizados: a) bibliográfica; b) documental; c) experimental; d) ex-post fact; e) estudo de coorte, f) levantamento; g) estudo de campo; h) estudo de caso; i) pesquisa-ação e j) pesquisa participante.

EXERCÍCIO

Pesquisar na Internet o que é cada um dos tipos de pesquisas acima citados.

[1] GIL, Antonio Carlos. **Como elaborar projetos de pesquisa.** 4. ed. São Paulo: Atlas, 2002. Disponível em < https://professores.faccat.br/moodle/pluginfile.php/13410/mod_resource/content/1/como_elaborar_projeto_de_pesquisa_-_antonio_carlos_gil.pdf>

EXEMPLO DE PROJETO DE PESQUISA:

Tema: Estudantes de Estatística

Delimitação do tema

Quem? Onde? Quando? Quanto?

Restringir dentro de um contexto

O que é possível estudar sobre o tema escolhido?

Conhecimentos prévios em Estatística; Interesses sobre Estatística; Perfil (características gerais) e tantos outros

Escolher UMA ideia. Meu interesse é o perfil dos estudantes

Fazer novamente a pergunta: O que é possível estudar sobre o perfil dos estudantes de estatística?

Perfil econômico; social; acadêmico; técnico; empreendedor...

Escolher UMA ideia. Meu interesse é o perfil acadêmico dos estudantes de estatística

O que é possível estudar sobre o perfil acadêmico?

Foco nos estudos; Materiais que utiliza; Métodos de estudo; Satisfação com o curso...

Escolher UMA ideia. Meu interesse é a satisfação dos estudantes de estatística com o seu curso

Tem que se ter no mínimo DUAS ideias para se trabalhar

Assim, o tema de pesquisa é

O perfil acadêmico dos estudantes de estatística, disciplina INE5111, e a satisfação com o curso que está fazendo na UFSC em 2019-1

Problema de pesquisa:

Existe uma relação entre o perfil acadêmico e a satisfação dos estudantes de estatística com o seu curso?

Definição de termos:

Perfil acadêmico: o curso que faz; o turno; a fase que está; estágio/trabalho na área do curso; atividades extras (participação em eventos, centro acadêmico, cursos complementares, língua estrangeira....); tempo que dedica para estudos...

Satisfação com o curso: com as salas de aula; com os laboratórios; com os professores; com os conteúdos programáticos das disciplinas, com os métodos de ensino...

Hipóteses de pesquisa:

1) A satisfação dos estudantes difere com relação:
    a. às salas de aula;
    b. aos laboratórios;
    c. aos professores;

d. aos conteúdos programáticos das disciplinas e a
e. os métodos de ensino.

2) Os estudantes com experiências de estágios ou de trabalho na área do curso são mais satisfeitos.

3) Os estudantes menos satisfeitos têm IA médio menor.

4) As notas médias das avaliações são equivalentes a média do IA

5) Não há diferença no grau de satisfação entre os estudantes do diurno e do noturno

Objetivo geral:

Verificar se há uma relação entre o perfil acadêmico e a satisfação dos estudantes de estatística com o seu curso

Objetivos específicos:

1) Verificar se existem diferenças entre as satisfações dos estudantes com relação:
   a. às salas de aula;
   b. aos laboratórios;
   c. aos professores;
   d. aos conteúdos programáticos das disciplinas e a
   e. os métodos de ensino.

2) Investigar se há associação entre o grau de satisfação com estágio/trabalho na área do curso.

3) Analisar se há relação entre a satisfação e o IA.

4) Examinar as relações existentes entre as notas médias das avaliações e o IA médio.

5) Averiguar se o grau de satisfação é o mesmo para os estudantes do diurno e do noturno.

População: Alunos de graduação do Curso X da UFSC (Por exemplo: O número de alunos do curso X é 850 alunos)

Tamanho da população: N= 850

Amostra:

Para definição da amostra devemos responder duas perguntas: quantos e quem vamos necessitar para compor nossa amostra

- *Quantos*

Tamanho da amostra (calcular com margem de erro - por exemplo: 2,5% - e nível de confiança de 95% (z=1,965).

Usando esses dados na fórmula para o **tamanho inicial da amostra** obtém 1545 alunos.

$$n_0 = \frac{z^2}{4.E^2} = \frac{1{,}965^2}{4.(0{,}025^2)} = 1545$$

Ajustando para o tamanho da população: N=850, o **tamanho final da amostra** é 549 alunos.

$$n = \frac{N.n_0}{N + n_0} = \frac{850\,.\,1545}{850 + 1545} = 549$$

Como essa amostra pode ser inviável pelo tempo disponível e a falta de outros recursos necessários, o pesquisador aumenta a margem de erro para 3%, por exemplo, e recalcula o tamanho da amostra.

- *Quem*

Técnica de amostragem (descrever as técnicas que serão utilizadas para selecionar a amostra).

A população (o curso X da UFSC, por exemplo, tem alunos que estudam no turno diurno e no turno noturno e nas fases iniciais há número maior de alunos. Então se devem levantar informações sobre quantos alunos tem em cada turno e em cada fase para calcular os respectivos percentuais).

Em virtude dessas características da população necessita-se de uma amostra que apresente, nas mesmas proporções, cada característica da população e por isso usa-se a amostragem estratificada proporcional. O quadro 6 apresenta um exemplo de cálculo para determinar os percentuais na população de cada estrato e esses mesmos percentuais são aplicados para determinar o número de estudantes necessários em cada um dos estratos para compor a amostra.

Quadro 6 – Exemplo de cálculo para determinar o tamanho da amostra para a técnica de amostragem estratificada proporcional para a população de 850 alunos do Curso X da UFSC – 2018/1 por estrato: diurno, noturno, fases iniciais e fases finais.

		Quantidade de alunos do Curso X da UFSC – 2018/1			
		Diurno		Noturno	
		Fases inicias	Fases finais	Fases inicias	Fases finais
População	N=850	270	180	240	160
% da população		0,3175	0,2118	0,2835	0,1882
Amostra	**n=549**	**175**	**117**	**155**	**104**

Desse modo na amostra deverá ter 175 alunos do diurno nas fases iniciais; 117 alunos do diurno nas fases finais; 155 alunos do noturno nas fases iniciais e 104 alunos do noturno nas fases finais.

O que vamos perguntar aos estudantes?

Variáveis

As variáveis serão posteriormente transformadas nas perguntas para o nosso questionário. Essas devem ser estabelecidas em função dos objetivos e das hipóteses de pesquisa. O quadro 7 apresenta na primeira coluna os objetivos e as hipóteses e na segunda coluna as respectivas variáveis propostas para cada objetivo/hipótese.

Quadro 7 – Exemplo de variáveis propostas para alcançar os objetivos e verificar as hipóteses de pesquisa

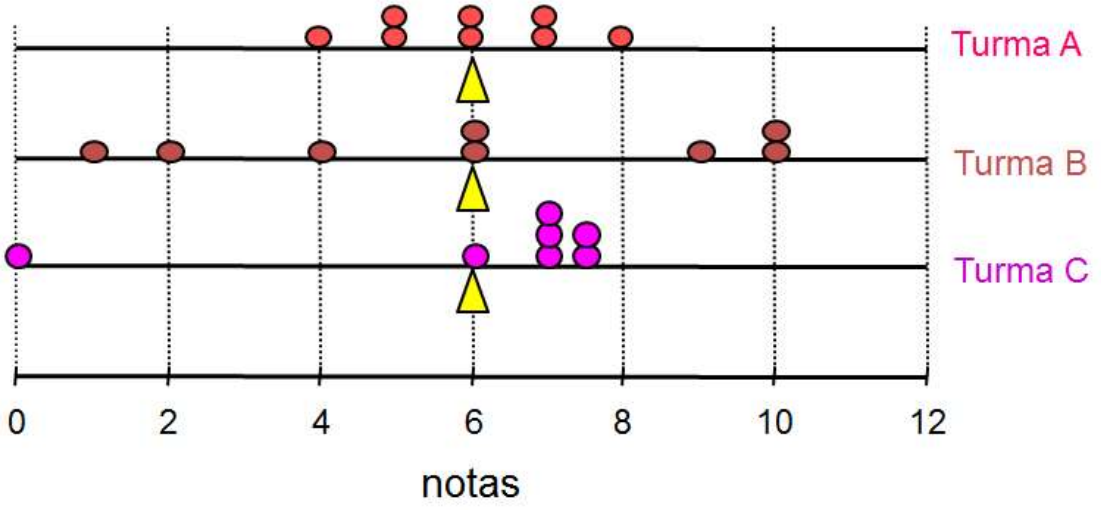
Variáveis para atender ao objetivo geral	Satisfação com o programa – grade curricular
	Satisfação com os métodos de ensino
	Satisfação com os professores
	Satisfação com as salas de aula
	Satisfação com os laboratórios
	Curso
	Turno
	Fase
	Estágio
	Trabalho
	Participação em eventos, centro acadêmico
	Sexo
Variáveis para atender ao objetivo específico 1	Satisfação com o programa – grade curricular
	Satisfação com os métodos de ensino
	Satisfação com os professores
	Satisfação com as salas de aula
	Satisfação com os laboratórios
Variáveis para atender ao objetivo específico 2	Satisfação
	Estágio
	Trabalho
Variáveis para atender ao objetivo específico 3	Satisfação
	IA
Variáveis para atender ao objetivo específico 4	Notas das avaliações (provas, trabalhos)
	IA
Variáveis para atender ao objetivo específico 5	Satisfação
	Turno

EXERCÍCIO

Elaborar um projeto de pesquisa seguindo o exemplo dado. (Tema livre)

Estatísticas descritivas:

Turmas	Notas de alguns alunos								média	dp	p	n
A	4	5	5	6	6	7	7	8	6	1,31	0,625	8
B	1	2	4	6	6	9	10	10	6	3,51	0,7	8
C	0	6	7	7	7	7,5	7,5		6	2,69	0,857	7



Medidas de tendência central

$$\begin{cases} a)\ Média \\ b)\ Mediana \\ c)\ Moda \end{cases}$$

Média aritmética simples:

$$\bar{x} = \frac{\sum x_i}{n}$$

Medidas de localização (ou separatrizes)

$$\begin{cases} Quartis \\ Decis \\ Percentis \end{cases}$$

Valores discrepantes (outliers)

Desvio entre quartis (dq): $dq = Q_3 - Q_1$

Esperado para mínimo: $= Q_1 - 1{,}5(dq)$

Esperado para máximo: $= Q_3 + 1{,}5(dq)$

Medidas de dispersão

a) Amplitude
b) Desvio entre quartis
c) Variância
d) Desvio padrão
e) Coeficiente de variação

Variância: $S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$ ou $S^2 = \frac{\sum x_i^2 - n.\bar{x}^2}{n-1}$

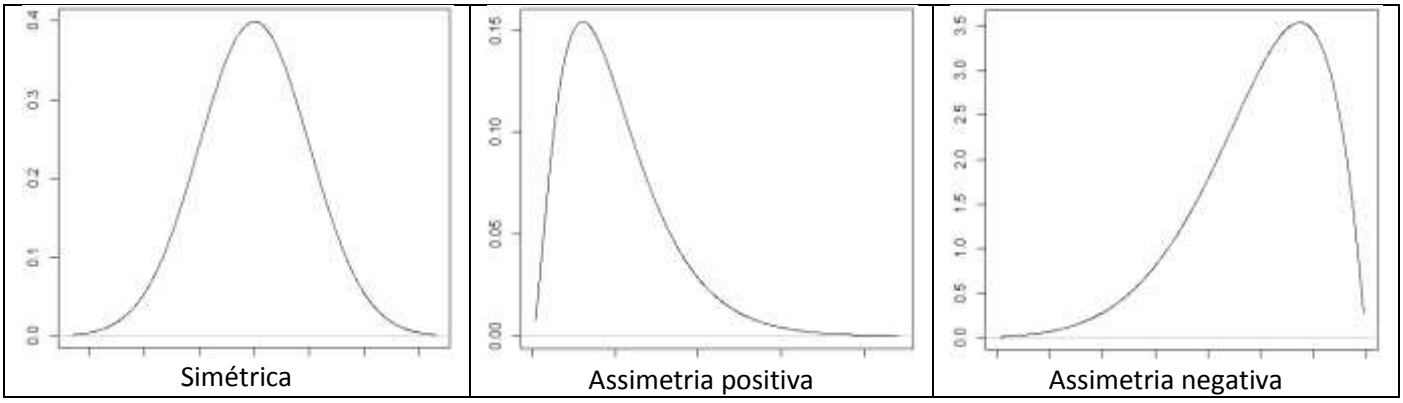
Desvio padrão: $s = \sqrt{s^2}$

Coeficiente da variação: $CV = \frac{s}{\bar{x}}.100$

Medidas de forma da distribuição

{ *Assimetria*
{ *Curtose*

Simetria

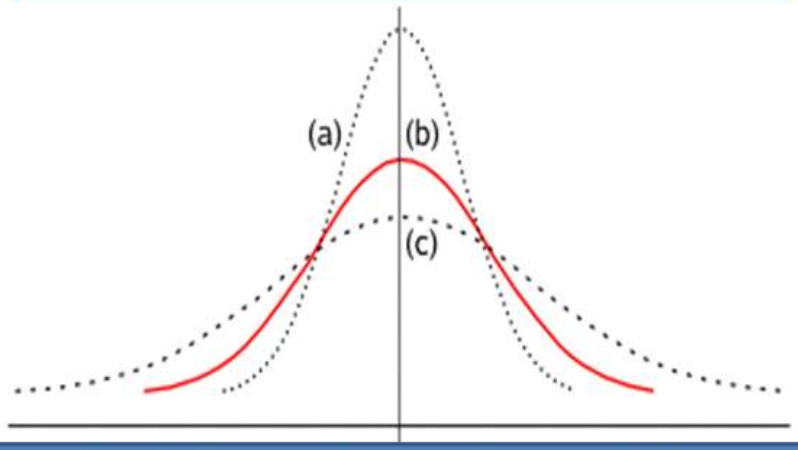


Coeficiente de assimetria = 0 : a distribuição dos dados é simétrica

Coeficiente de assimetria < 0 : a distribuição dos dados não é simétrica, apresenta assimetria negativa

Coeficiente de assimetria > 0 : a distribuição dos dados não é simétrica, apresenta assimetria positiva

Curtose



Coeficiente > 3 curva Leptocúrtica:
- Curva a (elevada);

Coeficiente = 3 curva Mesocúrtica:
- Curva b (normal);

Coeficiente < 3 curva Platicúrtica:
- Curva c (achatada).

Exercícios: Lista 2

Exercícios: Lista 3

Exercícios: Lista 4

Exercícios: Lista 5

Probabilidade

A probabilidade pode ser definida como a chance de algo acontecer ou não.

A probabilidade pode ser quantificada por números de zero a um (0 a 1). Uma probabilidade igual a zero, significa que não há chances de algo acontecer e igual a um, significa que é de certeza que algo acontece.

Qual é chance de alguém acertar na Mega Sena?

A probabilidade de alguém acertar na Mega Sena é

$$\frac{1}{50.063.860} = 0{,}00000002$$

Entre uma das formas de se calcular probabilidades os modelos teóricos de probabilidades são bastante empregados na prática. Existem modelos teóricos apropriados para as variáveis aleatórias discretas e para as variáveis aleatórias contínuas. Assim, quando se deseja determinar a probabilidade de ocorrer algum evento, verificam-se as características da variável que está relacionada a esse evento e comparam-se tais características com as propriedades dos modelos teóricos de probabilidades. Com a escolha do modelo adequado, determinam-se as probabilidades desejadas.

O modelo teórico de probabilidades mais empregado para análises de dados é a Distribuição de Probabilidade Normal

A Distribuição de Probabilidade Normal tem uma função muito importante para a inferência estatística.

O que é inferência estatística?

É comum se ter o interesse em pesquisar as características de uma população e essa ser muito grande para observar todos os elementos, ou itens, ou objetos, ou pessoas que a compõem. Para viabilizar essa pesquisa retira-se uma amostra e, por meio da amostra, é possível conhecer as características da referida população.

Para garantir que a amostra seja representativa, empregam-se as técnicas de amostragem e para garantir que a amostra seja significativa, calcula-se o tamanho da amostra. Entre os tipos de técnicas de amostragem são as aleatórias que permitem fazer a inferência.

As características de interesses na população são chamadas de parâmetros e são desconhecidas. Na amostra essas características são chamadas estimativas. Por exemplo:

CANDIDATOS	AMOSTRA
A	20%
B	15%
C	10%
...	55%

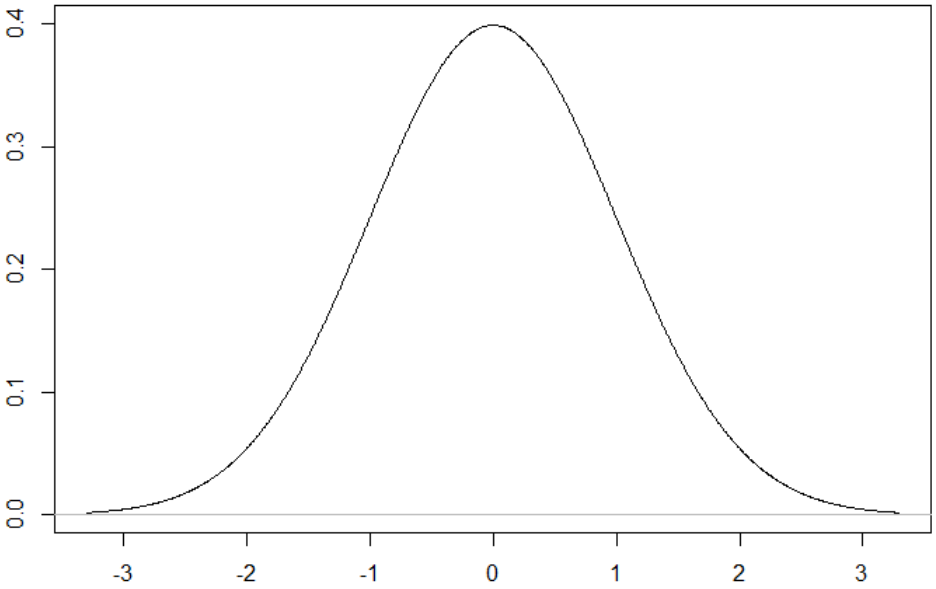
Margem de erro: 2%

INFERÊNCIA

CANDIDATOS	AMOSTRA	POPULAÇÃO
A	20%	De 18% a 22%
B	15%	De 13% a 17%
C	10%	De 8% a 12%
...	55%	

Voltando a conhecer mais sobre a Distribuição de Probabilidade Normal

A Distribuição de Probabilidade Normal divide a inferência estatística em duas: estatística paramétrica e não paramétrica. Para se saber que técnicas empregar na inferência, antes se necessita saber o comportamento da variável que se está estudando.



Curva da distribuição normal padrão:

- é simétrica (simetria)

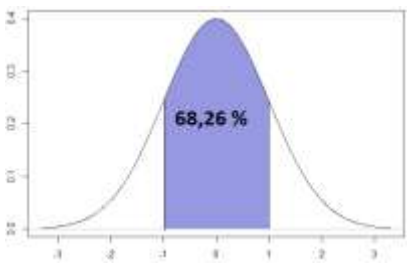
- centrada na média

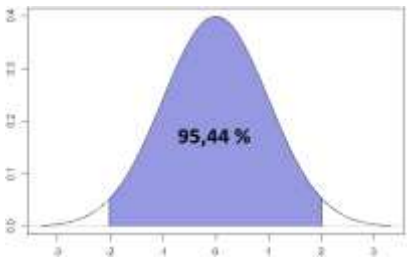
- nas proximidades da média se concentra as maiores probabilidades

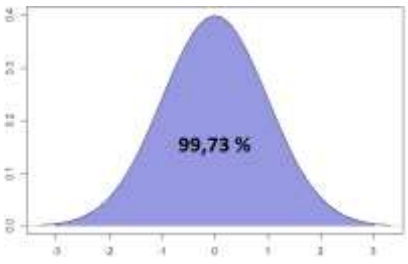
- os afastamentos da média são em unidades de desvio padrão

- é o desvio padrão que serve de referência para que a curva seja mais achatada ou mais alongada (Curtose)

- se uma variável X se aproxima de uma distribuição de probabilidade normal, então se diz que:







A média mais ou menos um desvio padrão contém 68,25% dos dados

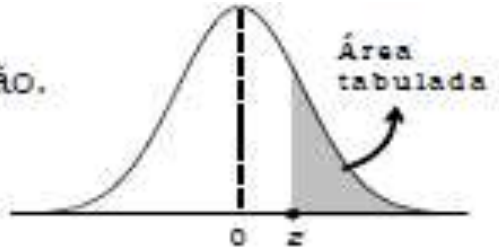
A média mais ou menos dois desvios padrões contém 95,44% dos dados

A média mais ou menos três desvios padrões contém 99,73% dos dados

Porque se chama Normal Padrão?

Antes das facilidades proporcionadas pela tecnologia de computação, não era fácil calcular as probabilidades de uma variável, usando o modelo matemático da distribuição normal. Para resolver essa dificuldade foi criada uma tabela onde constam os valores de probabilidades, considerando que o valor da média é zero e o valor do desvio padrão é um. Com isso, qualquer pessoa, com facilidade, encontra os valores que necessita, usando a tabela da distribuição normal padronizada: N (0; 1).

Tabela 4 Distribuição normal padrão.



	segunda decimal de z									
z	0	1	2	3	4	5	6	7	8	9
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2842	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0722	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0352	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
2,9	0,0019	0,0018	0,0017	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
3,0	0,00135									
3,5	0,000 233									
4,0	0,000 031 7									
4,5	0,000 003 40									
5,0	0,000 000 287									

Na prática, como vamos usar a Distribuição de Probabilidade Normal?

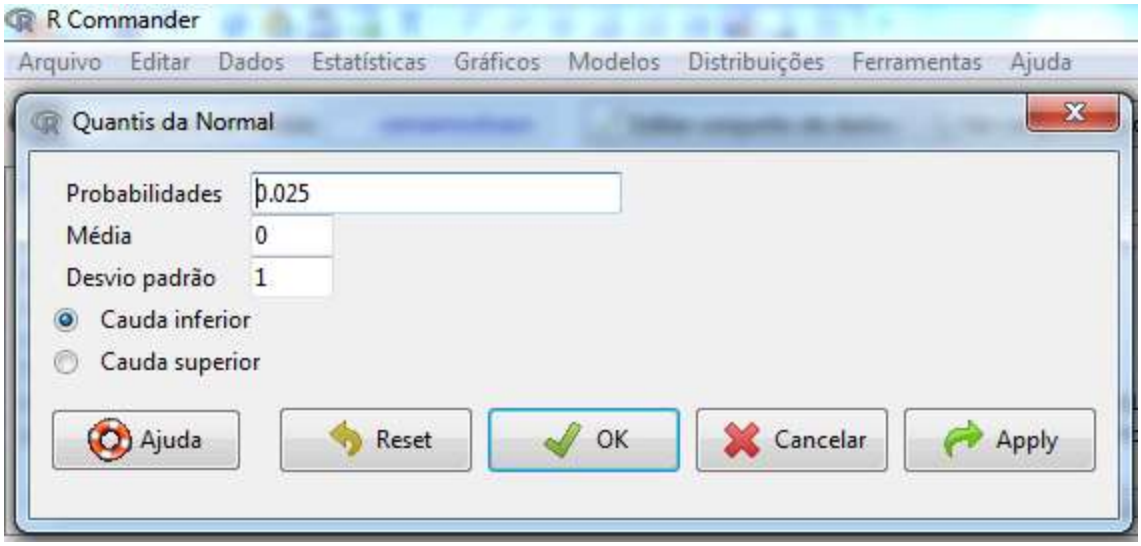
No cálculo do tamanho da amostra

$$n_0 = \frac{z^2}{4.E^2}$$

Z é um valor obtido da distribuição de probabilidade normal para o nível de significância desejado.

Por exemplo, para o nível de significância de 95%

No software R essa distribuição de probabilidades encontra-se no menu: Distribuições > Distribuições Contínuas > Distribuição Normal > Quantis da Normal

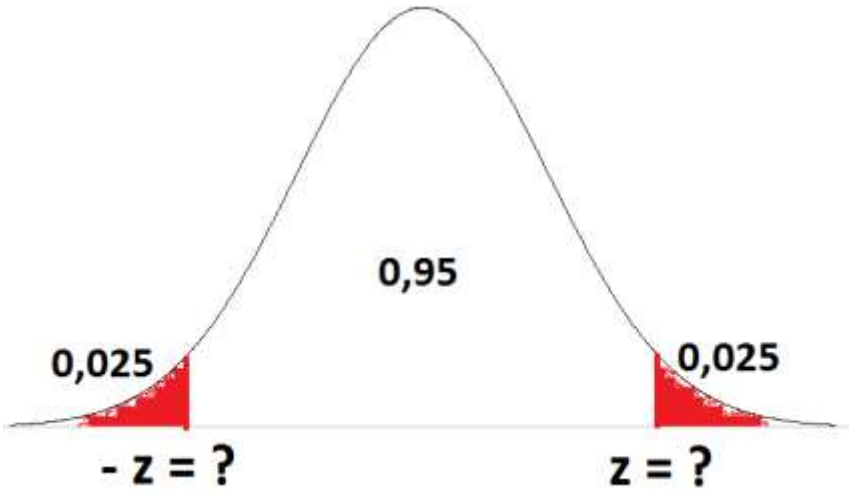


Informar a probabilidade de acordo com o nível de confiança para a qual se deseja.

A média 0 (zero) e o desvio padrão 1 (um) já estão preenchidos e com isso se tem a distribuição normal padrão, que calcula o valor de z para a fórmula do tamanho da amostra.

É necessário entender a lógica do programa para informar corretamente os valores de probabilidades.

No caso de 95% para o nível confiança

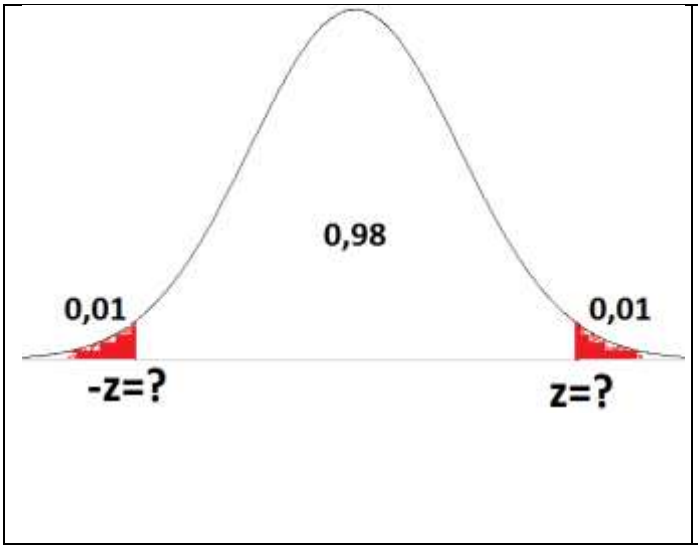


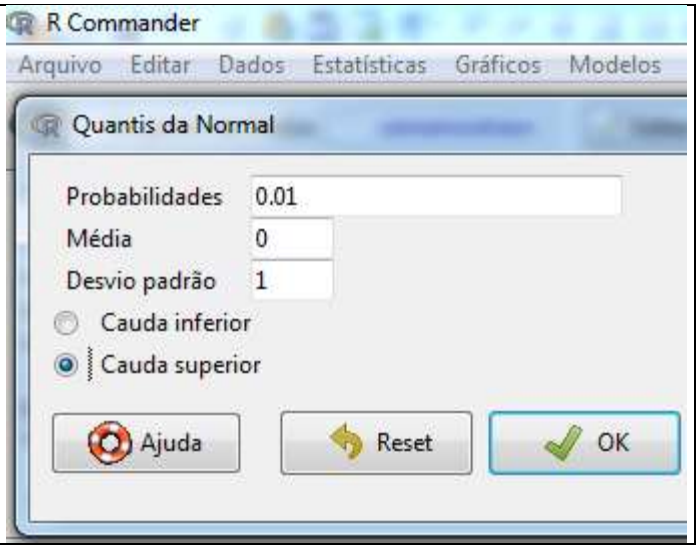
O valor 0,95 de probabilidade (divide-se 95 por 100 para tirar o símbolo %. E porque probabilidade é um número entre 0 e 1) é colocado no centro da distribuição normal. A área total sob a curva é igual a 1 o que corresponde a todos os valores de probabilidade. No ponto mais alto da curva tem-se a média da distribuição normal, acima e abaixo da média as probabilidades diminuem à medida que os valores no eixo horizontal se distanciam desse ponto central. Isso significa que em torno da média concentra a maior probabilidade. Como a distribuição é simétrica, as áreas correspondentes às duas caudas são iguais. Na figura a área pintada de vermelho é igual a 0,025 para cada uma das caudas. Para determinar esse valor calcula-se: 1 – 0,95 = 0,05 dividido por 2. É o que falta dos 0,95 para completar 1, distribuídos para os dois lados. O valor que deve ser informado no software é o que corresponde a essa probabilidade, no exemplo, 0.025. O resultado será o valor de z. Para a cauda direita (superior) o seu valor é positivo e para a esquerda (inferior) é negativo.

Abaixo se mostra a saída do R para o que foi solicito no menu anterior. Veja que o valor de z é igual a -1,959964. Quando se apreendeu a calcular o tamanho da amostra o número usado foi 1,96 (com arredondamento para duas casas decimais), porque foi obtido por meio da tabela da distribuição normal. Observe que no menu acima aparece acionado o botão correspondente à cauda inferior, por isso o resultado do z é negativo.

```
> qnorm(c(0.025), mean=0, sd=1, lower.tail=TRUE)
[1] -1.959964
```

Veja outro exemplo: Para o nível de confiança de 98%.





```
> qnorm(c(0.01), mean=0, sd=1, lower.tail=FALSE)
[1] 2.326348
```

Na tabela da distribuição normal, para o nível de confiança de 0,98 o valor de z é igual a 2,33.

Como usar a tabela da distribuição normal?

Os números que estão no corpo da tabela correspondem às probabilidades e a primeira coluna corresponde ao valor de z, sendo esse um número inteiro com uma casa decimal.

A segunda casa decimal para os valores de z correspondem às colunas que se apresentam no corpo da tabela. Por exemplo, se quiser ver a probabilidade correspondente a z=1,25. Os números 1,2 aparecem na primeira coluna e o valor 5 corresponde a sexta coluna da tabela. Assim, a probabilidade é:

O outro modo de uso da tabela é quando se quer saber o valor de z tendo o valor da probabilidade, como exemplificado para os níveis de confiança de 95% e 98%.

O valor de z para 0,95, do mesmo modo como apresentado anteriormente, a probabilidade é 0,025. Para saber o valor de z, deve-se encontrar no corpo da tabela o número mais próximo desse 0,025:

Encontrado o número 0,025 no corpo da tabela, verifica-se na horizontal que está na linha 1,9 e na vertical na coluna 6. Assim, o valor de z é 1,96.

Continuando com a INFERÊNCIA

Outro uso dessa distribuição de probabilidades é na inferência. A inferência estatística pode ser feita por meio de estimação de parâmetros ou testes de hipóteses. Porém, antes de qualquer aplicação de técnicas de inferência é necessário verificar se a variável X se aproxima de uma distribuição normal. Para fazer essa verificação há o modo gráfico e os testes.

Há dois gráficos que são usados: o histograma e o gráfico de comparação de quantis.

Para confirmar o que é sugerido nos gráficos, existem os testes de normalidade

Exercícios: Lista 6

TESTES DE HIPÓTESES

No projeto foram estabelecidas as hipóteses de pesquisa que devem ser verificadas com os dados. Isto é, se aquela suposição inicial é confirmada ou não, com base nos dados levantados.

Inicialmente devem ser estabelecidas duas hipóteses para cada hipótese de pesquisa elaborada no projeto. Essas duas hipóteses são chamadas de hipótese nula e hipótese alternativa.

H_0 – é a hipótese nula

H_1 – é a hipótese alternativa

A hipótese nula é elaborada com o sentido de que os valores encontrados nos dados (as estimativas) são iguais aos valores da população (os parâmetros).

A hipótese alternativa é elaborada com o sentido de contradizer a hipótese nula, ou seja, os valores obtidos com a pesquisa não são iguais aos da população.

Em resumo a formulação das hipóteses pode ser descrita como:

H_0 : sempre usa a igualdade

H_1 : usa a diferença, ou maior, ou menor

Testes unilaterais e bilaterais

Em função da hipótese alternativa, H_1, os testes são classificados em unilaterais e bilaterais.

Quando a hipótese alternativa, H_1, expressa a diferença, o teste é bilateral

Quando a hipótese alternativa, H_1, se traduz por menor ou maior, o teste é unilateral à esquerda ou à direita, respectivamente.

Nível de significância do teste

O nível de significância do teste corresponde à probabilidade de rejeitar H_0, sendo H_0 verdadeira. Em outras palavras, é o risco de tomar uma decisão errada na conclusão de um teste de hipóteses.

O nível de significância é simbolizado pelo α (alpha). O valor de α, em geral, é 0,05. É obtido da diferença: 1 menos o valor para o nível de confiança estabelecido para a pesquisa.

O risco de tomar uma decisão errada:

Decisão		Realidade	
		H_0 verdadeira	H_0 falsa
	Aceitar H_0	**OK**	Erro Tipo II (β)
	Rejeitar H_0	Erro Tipo I (α)	**OK**

A realidade representa o parâmetro da população que é desconhecido. Esse parâmetro pode ser conhecido se forem pesquisados todos os objetos de investigação da população. Como só é possível pesquisar uma amostra, será por meio dessa amostra que se fará a inferência, isto é: concluir, com base na amostra, o que é na população, decidindo por aceitar ou rejeitar a hipótese nula. Assim, se estabelece H_0, que tanto pode ser verdadeira como falsa. Os dados levantados pela pesquisa levam o pesquisador a tomar uma decisão: aceitar H_0 ou rejeitar H_0. Essa decisão é tomada com base no nível de significância do teste.

Regra para a decisão de um teste de hipóteses

Com base no p valor informado pelo software R e comparado esse ao nível de significância, o α, decide-se por aceitar ou rejeitar H_0, de acordo com as seguintes regras:

$p \leq \alpha \rightarrow$ rejeita-se H_0

$p > \alpha \rightarrow$ aceita-se H_0

Ao rejeitar H_0, a conclusão é que os dados mostram evidência de que a hipótese alternativa é comprovada. Ao aceitar H_0, a conclusão é que os dados não comprovam o que a hipótese alternativa pressupõe.

Tipos de testes:

I) Variáveis qualitativas

- Teste do Qui-quadrado.

Quando o teste envolver uma ou mais variáveis qualitativas será usado o teste do Qui-quadrado.

O teste do Qui-quadrado é empregado para verificar se existe alguma relação entre variáveis qualitativas. Por exemplo, será que há diferença no grau de satisfação entre homens e mulheres quanto ao laboratório de informática?

Se os dois grupos (homens e mulheres) estão igualmente satisfeitos, então não se comprova uma relação entre essas duas variáveis, porque a satisfação com o laboratório de informática é independente do sexo. Nesse caso não há relação entre essas duas variáveis.

Outro exemplo: será que há uma relação entre as variáveis: turno de estudos e situação de trabalho? É possível que os estudantes do turno noturno, em maior quantidade, são trabalhadores; enquanto que no diurno os estudantes não trabalham.

Se os dois grupos (trabalha e não trabalha) forem dependentes dos turnos, então haverá mais pessoas que trabalham estudando a noite e menos pessoas que não trabalham estudando de dia. Desse modo verifica-se que há uma relação entre as variáveis, pois estudar a noite depende de trabalhar ou não.

A confiabilidade do resultado do teste Qui-quadrado requer o atendimento de alguns pressupostos:

1) Quando $n \leq 20$ (n = tamanho da amostra) é necessário que as frequências esperadas em cada célula sejam ≥ 5.

2) Quando $n > 20$ (n = tamanho da amostra) é aceitável que as frequências esperadas, no máximo em 20% das células, sejam < 5 e não poderá haver frequências esperadas < 1.

Se não são forem atendidos os pressupostos em tabelas 2X2 usa-se o teste exato de Fisher.

II) Variável qualitativa e quantitativa

É necessário testar se a variável quantitativa se aproxima da distribuição Normal ou não, pelo teste de Shapiro-Wilk. Se a variável se aproxima da distribuição Normal usa-se teste paramétrico, caso contrário usa-se teste não paramétrico.

As categorias de respostas da variável qualitativa determinam quantos grupos/amostras independentes para o teste a ser usado, conforme o quadro abaixo:

Aplicação	Teste paramétrico	Teste não paramétrico
Duas amostras independentes	Teste *t* ou teste *z*	Teste de Wilcoxon
Várias amostras independentes	Análise de variância (ANOVA – teste F)	Teste de Kruskal-Wallis

III) Variáveis quantitativas

a) Se for o caso de comparação das médias resultantes entre as duas variáveis é necessário testar se as variáveis quantitativas se aproximam da distribuição Normal ou não, pelo teste de Shapiro-Wilk. Basta que uma delas não se aproxime da distribuição Normal para usar teste não paramétrico. Para aplicar teste paramétrico exige-se que TODAS as variáveis - que estejam envolvidas no teste - se aproximem da distribuição Normal. O teste será conforme o quadro abaixo para amostras/grupos dependentes (pareados ou dados emparelhados).

Aplicação	Teste paramétrico	Teste não paramétrico
Duas amostras dependentes (pareados ou dados emparelhados)	Teste *t* ou teste *z*	Teste de Wilcoxon
Várias amostras dependentes (pareados ou dados emparelhados)	Análise de variância (ANOVA – teste F – para vários fatores)	Teste de Friedman

b) Se for o caso de verificar se há relação entre as duas variáveis quantitativas usa-se a correlação linear de Pearson e o teste de correlação para avaliar se a correlação é significativa ou não.

Exercícios: Lista 7

Exercícios: Lista 8