
FIAP

FIAP

STATISTICAL COMPUTING WITH R

AULA 2

Prof. Reinaldo Borges Júnior

São Paulo - 2021

ESTATISTICA

Lembrando... EXERCÍCIO DA AULA PASSADA

Durante dois dias, em um cruzamento movimentado de ruas de São Paulo, foram registrados 50 valores de nível de ruído de tráfego, em decibéis (dB), os quais são apresentados na tabela a seguir:

Variável: ruído ou poluição sonora (dB).
Tipos de dados: dados brutos.

Obtenha, a partir da tabela, o **"rol"** e a **"amplitude total (AT)"**, dos dados colhidos.

Poluição sonora em decibéis – dados brutos

58,0	62,5	65,3	67,0	68,3	65,0	66,4	58,0	67,0	68,3
67,0	62,5	62,5	66,4	66,4	65,0	65,0	60,2	60,2	62,5
60,2	60,2	59,5	59,5	59,5	65,0	66,4	66,4	66,4	65,0
60,2	62,5	67,0	67,2	67,0	70,1	70,1	71,9	70,1	65,0
67,0	66,4	66,4	68,3	68,3	68,3	65,0	65,0	62,5	71,9

ESTATISTICA

Conceitos fundamentais

- Essa pesquisa pode ser qualitativa ou quantitativa, depende do plano e questão de pesquisa traçado pelo pesquisador.

- Enquanto no modelo qualitativo a questão pode incluir a opinião do pesquisador, na quantitativa o observador deve manter sua imparcialidade diante do fenômeno, registrando e verificando o que “os números estão dizendo”.

- Porém, com a crescente complexidade dos dados, podemos articular as duas modalidades – **qualitativa e quantitativa** – e é nesse ponto que o meio computacional se faz necessário porque são muitos dados para serem analisados.

- Assim podemos entender as razões das ações de cada pessoa em certo contexto. Nesse sentido, a pesquisa articulada, ou pesquisa híbrida (MCNEILL, 2018) ganha espaço, trazendo de forma consistente os questionamentos de ‘Por quê?’ e ‘Como?’.

- Os Parâmetros serão determinados pelo pesquisador e podem englobar **medidas de tendência centrais** como a média, a moda e a mediana dos valores. Um parâmetro estatístico auxilia na compreensão dos dados, pois dessa forma não preciso buscar os dados brutos a todo instante.

ESTATISTICA

Fases do trabalho estatístico

- Definir o Problema e o tipo de pesquisa;
- Definir o tipo de amostragem;
- Planejar a coleta de dados;
- Realizar Coleta de dados;
- Sistematizar ou organizar os dados (Seriação);
- Apresentar os dados;
- Analisar e interpretar os dados;
- Relacionar as inferências ao problema (em função do tipo de amostragem);
- Responder a questão (problema).

ESTATISTICA

Trabalho estatístico

- - Diferente de pesquisas científicas (ou acadêmicas) que procuram questões a serem dissertadas e amplamente descritas, as pesquisas estatísticas podem apresentar respostas dicotômicas (sim ou não, gosto ou não gosto, este ou aquele tipo de serviço etc.).

- Por exemplo, se desejo conhecer a preferência do público jovem (com idade entre 15 e 18 anos) a respeito de operadoras e planos de celulares, a resposta da pesquisa será objetiva e direta, não oferecendo uma demanda para interpretações. A pesquisa pode ser ampliada e com isso verificada algumas causas para essa preferência (preços, promoções, localidade, sistema operacional, recursos disponíveis etc.) cada novo aspecto trata de uma nova variável e por isso as pesquisas atuais, muitas vezes, se encontram no prisma do que conhecemos como Big Data.

ESTATISTICA

Trabalho estatístico

- Diferente de pesquisas científicas (ou acadêmicas) que procuram questões a serem dissertadas e amplamente descritas, as pesquisas estatísticas podem apresentar respostas dicotômicas (sim ou não, gosto ou não gosto, este ou aquele tipo de serviço etc.).

- Definir o problema ou a questão a ser analisada por uma pesquisa estatística é o primeiro passo, e a partir disso, posso traçar os caminhos que seguirei, como o tipo de pesquisa necessária e o tipo de amostragem que será representativa e útil para alcançar uma resposta.

- Nessa etapa é preciso delimitar o escopo da pesquisa, ou seja, qual o problema ou questão a ser respondida. Essa questão está ligada ao tipo de pesquisa – Quantitativa ou Qualitativa – e como consequência da questão eu também especifico qual a população que fará parte dela.

ESTATISTICA

Trabalho estatístico

- A definição de um problema estatístico se traduz em uma pergunta muito simples: **“O que quero saber?”**

- Para alcançar a resposta, necessito estruturar e planejar o caminho da pesquisa (metodologia) restringindo quais informações eu preciso e como deverei consegui-las para resolver meu problema (coleta de dados).

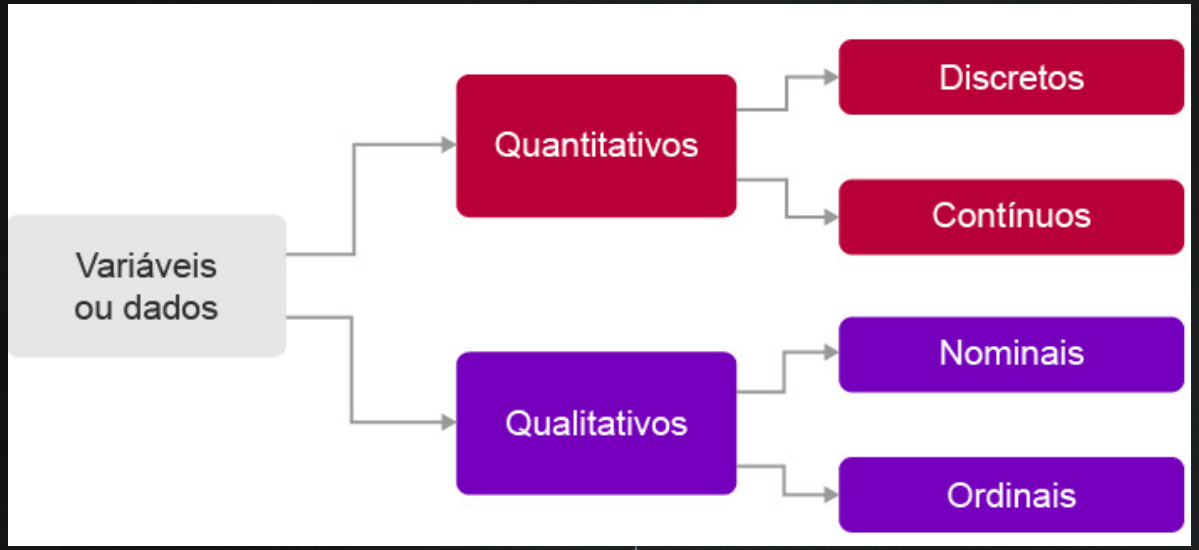
- Assim, uma pesquisa poderá ser quantitativa ou qualitativa. Esses métodos de pesquisa representam um ponto de vista do investigador sobre a organização do estudo, ou o: **“Como irei proceder para responder o que quero saber?”**.

- Na pesquisa quantitativa, a opinião do investigador deve ser excluída, enquanto na qualitativa pode estar integrada ao estudo.

- A escolha do tipo de pesquisa auxilia na definição do tipo de dados (ou variáveis) que serão utilizados, como observado na Figura a seguir, do tipo de variável estatística.

ESTATISTICA

Diagrama do tipo de variável estatística



ESTATISTICA

Trabalho estatístico

- Enquanto a **escala quantitativa** está relacionada com características que **podem ser medidas ou contadas** com características numéricas (por exemplo: massa, preço de venda, pressão arterial, número de filhos), a **escala qualitativa** será relacionada com a descrição de uma característica. Nesse caso, a variável não pode ser medida mas pode ser observada.

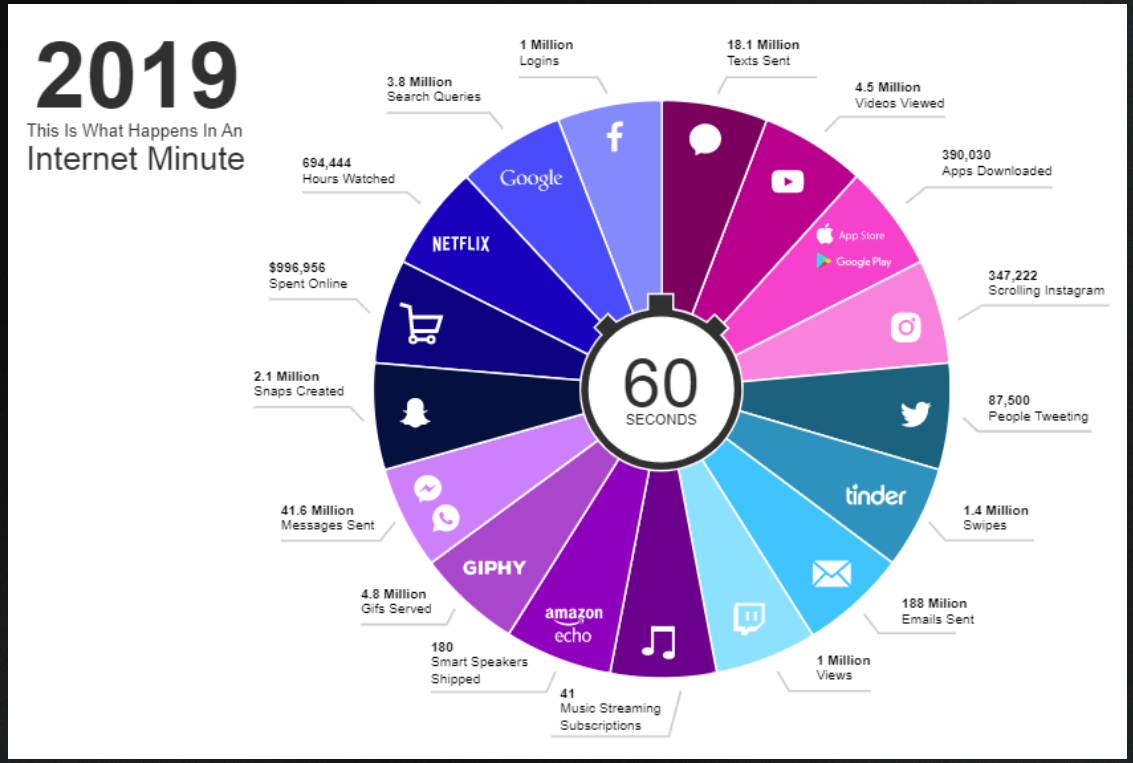
- Por exemplo (escala qualitativa): os tipos de defeitos em um produto (sensor) no final da linha de montagem, o nível educacional (ensino fundamental, ensino médio, graduação e pós-graduação), sexo (masculino ou feminino), cor da pele etc.

- A quantitativa ainda pode ser classificada em **discreta para valores inteiros**, como número de usuários “logados” em um game ou **contínuas**, como salário médio de profissionais em Tecnologia da Informação no Brasil.

- E a qualitativa em **ordinal**, quando apresenta algum tipo de ordenação como grau de escolaridade ou **nominais**, quando não existe ordem pré-determinada como sexo ou etnia.

ESTATISTICA

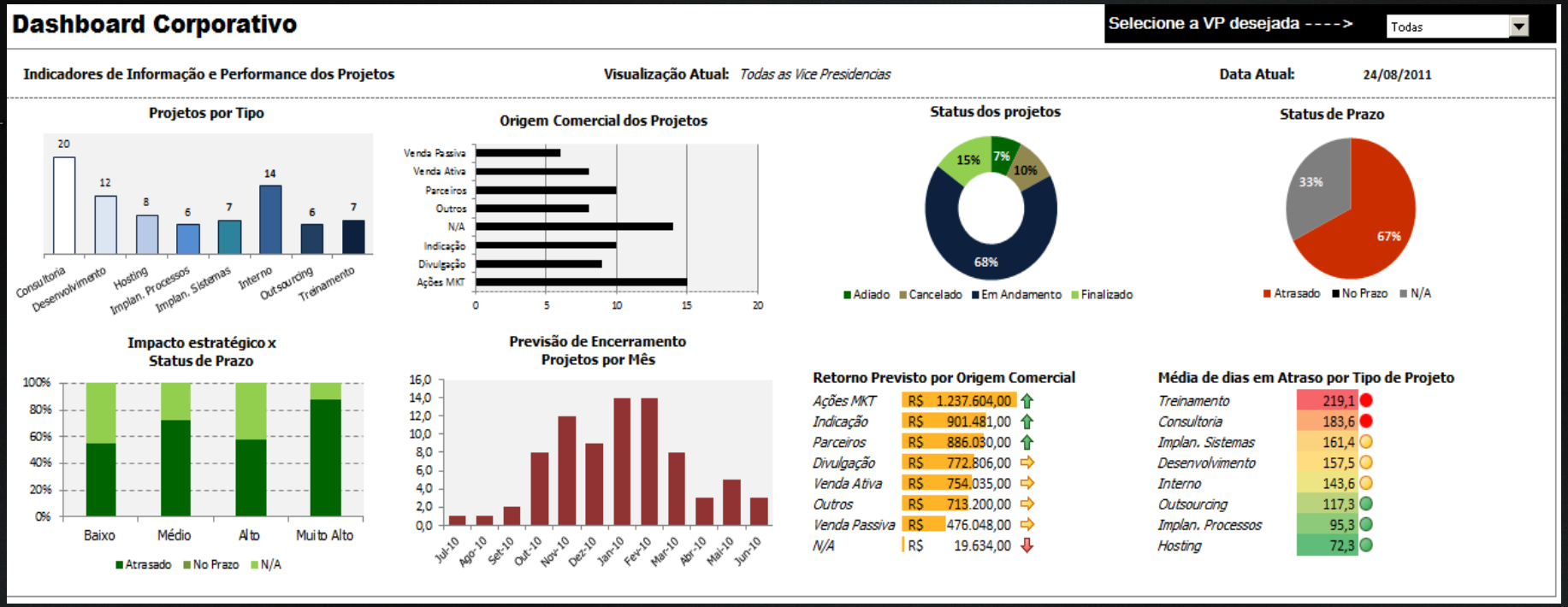
Infográfico



- Virou uma febre! O **infográfico** agrada a todos e é uma excelente forma de apresentar dados coletados em uma amostragem, bem como trazer informações adicionais que sejam relevantes para a compreensão de um fenômeno pelo leitor. Se você não sabe bem o que é um infográfico, pergunto: **Você sabe o que acontecia em 60 segundos na Internet em 2019?**

ESTATISTICA

Dashboards Corporativos



ESTATISTICA

Distribuição de frequências quantitativas discretas

- A **distribuição de frequências** consiste em uma forma de organização dos dados, agrupando-os em classes ou categorias e verificando a frequência (ocorrência, número de observações) em cada uma delas.

- Para isso, existe a necessidade de saber classificar uma variável. Afinal, como vimos na aula anterior, diferentes variáveis oferecem diferentes formas de tratamento estatístico.

- Em uma distribuição de frequência, estou interessado em resumir os dados para poder visualizar e analisar a amostra e seus dados gerados.

- Assim, vou utilizar alguns exemplos e apresentar outros conceitos importantes da Estatística Descritiva.

ESTATISTICA

Exemplo 1

Uma pesquisa foi realizada com a finalidade de saber o número de pessoas por família em uma região de São Paulo, capital. Uma amostra de 25 famílias foi entrevistada. E as respostas foram anotadas:

4 5 4 6 5 3 4 4 6 7 3 4 5 4 6 5 5 5 4 4 3 4 5 4 4

- **Variável e Tipo:** número de pessoas por família é Quantitativa Discreta.
- **Rol (para a lista de dados brutos):** 3 3 3 4 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 6 6 6 7
- **Classes ou Grupos:** identificamos todos os possíveis valores de respostas, ou seja: {3, 4, 5, 6, 7}.

ESTATISTICA

Exemplo 1

- **Frequência (ou frequência absoluta fi):** é o número de ocorrência (repetições) da resposta, desse modo, teremos a seguinte organização (Tabela “Número de pessoas por família").

Número de pessoas	Frequência
3	3
4	11
5	7
6	3
7	1
Total	25

ESTATISTICA

Exemplo 2

Foi perguntado para vinte (20) alunos de uma classe, qual país gostariam de conhecer e foram anotadas as respostas:

França; Portugal; França; Espanha; Estados Unidos;
Canadá; Espanha; Portugal; Estados Unidos; Estados Unidos;
Alemanha; Holanda; França; Estados Unidos; Canadá;
Canadá; Estados Unidos; França; Alemanha; África do Sul.

- **Variável:** País
- **Pesquisa:** Qualitativa nominal

ESTATISTICA

Exemplo 2

Vamos organizar a tabela de frequência (Tabela “País que se deseja conhecer”):

País	Frequência
África do Sul	1
Alemanha	2
Canadá	3
Espanha	2
EUA	5
França	4
Holanda	1
Portugal	2
Total	20

ESTATISTICA

Distribuição de frequências para variáveis quantitativas contínuas

- Quando preciso trabalhar com variáveis do tipo contínuas, percebemos que a possibilidade de obtermos valores repetidos é menor.

- São alguns exemplos de variáveis contínuas: Altura, Peso, Diâmetro, Número de Salários Mínimos, renda per capita; tempo de permanência em um website etc.

- Dependendo da técnica ou do aparelho de medição, a precisão pode aumentar, fazendo com que os valores se tornem quase únicos (ou pouco se repitam). Exemplo: Quantos valores existem entre as alturas (cm) de 1,50 e 1,60? Infinitos valores, pois se trata de números reais (conjunto dos reais) e quanto melhor a precisão do sistema de medição (coleta de dados), mais valores encontrados entre esses dois.

- Então, ao organizar a nossa tabela de frequência, usaremos um recurso da **Estatística Descritiva** denominado **intervalo de classe**. Desse modo, as classes ou categorias não serão valores absolutos (inteiros) como nos exemplos anteriores, mas intervalos (de classe) com mesmas amplitudes (tamanho do intervalo) de classe (h).

ESTATISTICA

Exemplo 3

Veja como ficou o exemplo dos dados de poluição sonora da aula anterior, vou organizar os dados em um rol:

58	60,2	62,5	62,5	65	66,4	66,4	67	68,3	70,1
58	60,2	62,5	65	65	66,4	66,4	67	68,3	70,1
59,5	60,2	62,5	65	65	66,4	66,4	67	68,3	70,1
59,5	60,2	62,5	65	65	66,4	67	67	68,3	71,9
59,5	60,2	62,5	65	65,3	66,4	67	67,2	68,3	71,9

ESTATISTICA

Exemplo 3

OBSERVAÇÃO TEÓRICA:

- Acredito que você possa me perguntar: “Professor, eu trabalho com banco de dados, e essa história de rol é apenas uma indexação que posso utilizar para classificar meus dados facilitando sua organização.”

- **Eu respondo: Muito bem! Você está correto! Sabe porque?**

- Muitos processos da Estatística datam de um período no qual o computador não estava na vida das pessoas ou dos pesquisadores (ou era para poucos).

- Então, muitas técnicas ainda **estão se adaptando aos novos contextos**, otimizando processos e facilitando a utilização de um número cada vez maior de informações, e aí temos a Cloud Computing; Big Data; Machine Learning, Inteligência Artificial, etc.

ESTATISTICA

Exemplo 3

OBS: Indicamos a fórmula de Sturges (i) se o número de elementos
da amostra for elevado (n > 50). E para quantidades ou espaços amostrais pequenos, a segunda equação (ii).

- Imagine agora que seja possível coletar uma quantidade muito grande de informações sobre a poluição sonora em determinado cruzamento de São Paulo. Então, além do rol (indexação dos dados), eu preciso de uma maneira para sintetizar e analisar os dados. Preciso de uma tabela de frequência. Vamos fazer isso:

- **Primeiro Passo:** o Rol.

- **Segundo Passo:** determinar o **número de classes (k)** e o **intervalo de classe (h)** para a tabela de frequência.

$$(i) k = 1 + 3{,}322 . \log(n)$$

$$(ii)\ k = \sqrt{n}$$

OBS: Não existem regras rígidas para determinar o número de classes, apenas seguir alguns parâmetros:
- Não escolher um número muito pequeno de classes, senão a amplitude de classe será muito grande.
- Não escolher um número muito grande de classes, senão a amplitude de classe será muito pequena.
- Existem algumas sugestões para o cálculo do número de classes e, consequentemente, da amplitude de classe, uma delas é a **fórmula de Sturges (i)** e a outra é o **critério da raiz quadrada (ii),** sendo **n o número de elementos** da amostra coletada.

ESTATISTICA

Exemplo 3

- **Terceiro Passo:** O **intervalo de classe (h)** é calculado pela **razão** entre **Amplitude Total (AT)** ou **Amplitude Amostral (AA)** e **Número de Classes (k)**. Então:

Valor mínimo: menor valor do rol (X_{Min})
Valor máximo: maior valor do rol (X_{Max})
Amplitude total ou Amplitude Amostral: **$AT = X_{Max} - X_{Min}$**

$$h = \frac{AT}{k} = \frac{X_{Max} - X_{Min}}{k} = \frac{71{,}9 - 58}{k}$$

- Calculando o número de classes (k), com a amplitude de classe (AT) para os dados do rol:

$$k = \sqrt{50} \cong 7{,}071 \cong 7$$

$$h = \frac{71{,}9 - 58}{7} = \frac{13{,}9}{7} \cong 2$$

ESTATISTICA

Exemplo 3

OBS: Tiboni (2010) aconselha que os valores devem sofrer arredondamento de acordo com a coerência dos dados. Se, por exemplo, os dados são todos múltiplos de 5, então, é natural que o número de classes e a amplitude de classe também sejam múltiplos de 5, é uma questão de avaliação e bom senso. Por isso, não vou adotar uma regra específica de arredondamento, apenas a "conveniência e a coerência".

- **Quarto Passo:** criar as classes com as respectivas frequências (absolutas ou fi):

i	Classes	Frequência (fi)
1	58 |— 60	5
2	60 |— 62	5
3	62 |— 64	6
4	64 |— 66	9
5	66 |— 68	15
6	68 |— 70	5
7	70 |— 72	5
	$\sum_{i=1}^{7} f_i$	50

ESTATISTICA

Exemplo 3

- **Quinto Passo:** Construir uma ampliação da tabela de frequência (Tabela de distribuição de frequência completa) e vou explicar o significado de cada coluna e como foi produzida.

i	Classes	Frequência (f_i)	$\bar{x}_i$	fr_i	fr_i(%)	Fac_i	$Frac_i$	$Frac_i$ (%)
1	58 ⊢ 60	5	59	0,10	10	5	0,10	10
2	60 ⊢ 62	5	61	0,10	10	10	0,20	20
3	62 ⊢ 64	6	63	0,12	12	16	0,32	32
4	64 ⊢ 66	9	65	0,18	18	25	0,50	50
5	66 ⊢ 68	15	67	0,30	30	40	0,80	80
6	68 ⊢ 70	5	69	0,10	10	45	0,90	90
7	70 ⊢ 72	5	71	0,10	10	50	1,00	100
	Total	50		1,00	100			

ESTATISTICA

Exemplo 3

- Ao escrever as classes de minha tabela, utilizei uma notação específica, (l_i|— L_i) que representa os limites inferiores e superiores de cada classe. Nesse caso, excluí o limite inferior e incluí o limite superior (Quadro “Tipos de representação de intervalos”).

- O ponto médio da classe é calculado da seguinte maneira: $\bar{x} = \frac{l_i + L_i}{2}$

- A frequência relativa é o quociente entre cada frequência de classe pelo total de elementos da amostra: $fr_i = \frac{f_i}{\sum_{i=1}^{n} f_i}$.

- Para frequência relativa percentual fr(%), basta fazer a frequência relativa multiplicada por 100.

ESTATISTICA

Exemplo 3

- Para ilustrar como calcular as **frequências acumuladas (Faci)**, apresento um fragmento de minha tabela de distribuição ao lado:

- A **frequência relativa acumulada** é o quociente entre cada frequência acumulada de classe pelo total de elementos da amostra: $fr_i = \frac{fac_i}{\sum_{i=1}^{n} f_i}$ e, para frequência acumulada relativa percentual (%), basta fazer a frequência acumulada relativa multiplicada por 100.

Frequência (f_i)	Fac_i	Operação realizada
5	5	Fac1 = f1
5	10	Fac2 = f1 + f2 = 5 + 5 = 10
6	16	Fac3 = f2 + f3 = 10 + 6 = 16
9	25	Fac4 = 16 + 9 = 25
15	40	Fac5 = 25 + 15 = 40
5	45	Fac6 = 40 + 5 = 45
5	50	Fac7 = 45 + 5 = 50
50		

ESTATISTICA

Exemplo 3

Mas, por que fazer uma tabela tão completa e cheia de colunas específicas?

- É importante elencar algumas considerações a respeito da tabela de frequência:
- É uma tabela de dados estatísticos e, portanto, não é uma tabela para apresentação de dados ao público.
- Por ser extremamente completa, oferece, de modo direto e objetivo, uma série de respostas a perguntas sobre a amostra analisada.
- Se a amostra utilizada for probabilística, então, a tabela aponta inferências sobre a população correspondente.
- Informação é tudo! Como afirmei, para qualquer tomada de decisões, é preciso possuir informações corretas e consistentes.

ESTATISTICA

Exercício

Vamos treinar um pouco?

Foram coletadas as alturas de 26 pessoas de uma quadra de um bairro da Zona Norte. Construa uma tabela de distribuição de frequência (completa, como no EXEMPLO 3) dessas alturas. A seguir os dados:

1.59; 1.60; 1.57; 1.55; 1.60;
1.63; 1.57; 1.61; 1.58; 1.59;
1.62; 1.56; 1.58; 1.61; 1.61;
1.61; 1.63; 1.62; 1.65; 1.58;
1.61; 1.61; 1.60; 1.60; 1.60; 1.57.

OBRIGADO

/reinaldoborgesjunior

FIAP MBA+


FIAP