A NATUREZA DA ESTATÍSTICA

Landim, Flávia – flavia@im.ufrj.br[1] (IM/UFRJ)
Soto, Ezequiel – cheque@impa.br[2] (IMPA – doutorando em Computação Gráfica)
Rocha, Nei – rocha@im.ufrj.br[3] (IM/UFRJ)
Rangel, Letícia – leticiagranjel@gmail.com[4] (CAP/UFRJ)
Leal, Vanessa – vanesamatosleall@gmail.com[5] (SME Angra dos reis e Mesquita – RJ)
Silva, Alexandre – alexandre.silva@uniriotec.br[6] (UNIRIO)

1. INTRODUÇÃO[1]

A Estatística está presente no mundo contemporâneo, chega aos cidadãos em todos os meios de comunicação e é a ferramenta por excelência no tratamento de modelagem de fenômenos aleatórios (não determinísticos). Diariamente somos confrontados com informações estatísticas sobre temas que variam de Economia à Educação, de filmes a esportes, de comida à medicina, e de pesquisas de opinião a comportamento social. Tais informações orientam decisões em nossas vidas pessoais e permitem-nos exercer nossas responsabilidades como cidadãos. (Franklin, C. A., 2007, GAISE).

A relevância do raciocínio estatístico e do conhecimento para o efetivo funcionamento na sociedade da informação levou à introdução do termo **Letramento Estatístico**: "A capacidade de compreender e avaliar criticamente resultados estatísticos que permeiam a vida diária, acompanhada da capacidade de apreciar como o pensamento estatístico pode contribuir em decisões públicas e privadas, profissionais e pessoais." (Batanero, Borovcnik, 2016)

Vivemos cercados de incertezas. A todo momento somos bombardeados por informações sobre pesquisas científicas comprovando (estatisticamente) que tal substância causa uma patologia, ou sobre pesquisas de opinião, índices de pobreza, características sobre o envelhecimento da população, e outros temas de natureza incerta. Num mundo assim, é importante ter espírito crítico para informações sujeitas à incerteza a fim de poder interpretá-las e, quando necessário, poder escolher, entre diferentes opções, aquela que parece melhor diante da incerteza. Nesse sentido, a Estatística é uma disciplina fundamental para todos os estudantes e, certamente, com grande responsabilidade para a formação crítica do cidadão, pois ela é usada nas mais variadas áreas do conhecimento tais como: Medicina, Economia, Política, Direito, Psicologia, Engenharia, Educação, entre outras.

Mas afinal o que é Estatística? Uma possível resposta é: “a arte e ciência de coletar, analisar, apresentar e interpretar dados, para que se tomem decisões sob incerteza”.

Existem formas diferentes de abordar a Estatística como uma disciplina. A abordagem voltada para “modelagem” será abordada nesse material. O propósito fundamental é, a partir de um conjunto de dados coletados, visualizar estruturas. Os primeiros passos nesse sentido envolvem a organização dos dados em tabelas de frequências e a construção de gráficos.

No que segue serão apresentadas atividades introdutórias para definir os diferentes tipos de variáveis e apresentar os tipos de gráficos adequados em cada caso.

[1] O material apresentado nessas notas faz parte do capítulo “A Natureza da Estatística” do Projeto “Livro Aberto de Matemática” (www.umlivroaberto.com) para o Ensino Médio.

2. Pesquisa sobre a prática de esportes e atividade física (IBGE/PNAD-2015)

A Pesquisa Nacional por Amostra de Domicílios (PNAD), realizada pelo IBGE (https://www.ibge.gov.br/estatisticas-novoportal/sociais/populacao/9127-pesquisa-nacional-por-amostra-de-domicilios.html), obtém informações anuais sobre características demográficas e socioeconômicas da população, como sexo, idade, educação, trabalho e rendimento, e características dos domicílios. Com periodicidade variável, a PNAD obtém informações sobre migração, fecundidade, entre outras, tendo os domicílios como unidade de coleta da informação. Temas específicos abrangendo aspectos demográficos, sociais e econômicos também são investigados.

Um aspecto fundamental da Estatística praticado nessa pesquisa é a forma na qual a amostra, subconjunto da população, é selecionada. Essa seleção é cuidadosamente planejada de modo que seja adequado estender os resultados obtidos na amostra para a população.

Para que os resultados de uma amostra possam ser estendidos para a população, é necessário planejar com cuidado como a amostra será selecionada, pois o critério de seleção da amostra depende da estrutura da população. Por exemplo, para saber se o feijão cozinhando na panela está bem temperado, basta provar uma pequena colherada. Por quê? Partimos do pressuposto de que todos os ingredientes foram bem misturados e, assim, a mistura é homogênea.

Quando dispomos de dados provenientes de um subconjunto da população sempre podemos descrever os dados nos restringindo apenas ao subconjunto. Se quisermos estender nossas conclusões para a população, será necessário o uso de outras tecnologias que permitam calcular as incertezas associadas a essas extensões.

Na PNAD 2015 foi realizada a investigação de um tema específico chamado "Suplemento de Práticas de Esporte e Atividade Física" no qual foram investigadas as pessoas moradoras de 15 anos ou mais de idade, em seu tempo livre, no período de referência de 365 dias, com o objetivo de quantificar aquelas que praticaram algum esporte ou atividade física no período considerado bem como a sua percepção quanto a isso. As informações levantadas nessa pesquisa foram obtidas por meio de um questionário no qual se perguntou:

i) Se a pessoa moradora havia praticado esporte, e em caso afirmativo, a respectiva modalidade.

ii) Independente da resposta anterior, também se perguntou se a pessoa praticava alguma atividade física que não considerava como esporte, informando, em caso positivo, também a modalidade.

iii) Outras informações levantadas nessa pesquisa foram: motivação para a prática da atividade física, local onde é praticada a atividade, frequência na qual a atividade é praticada, duração da atividade; e a participação em competições.

iv) Também foram levantadas informações sobre as pessoas que responderam que não praticavam atividade física. Perguntou-se o motivo de não o fazerem e se haviam praticado anteriormente, caso em que se perguntou a modalidade praticada, a idade em que parou de praticar e a causa da interrupção.

v) Além dessas informações, a pesquisa investigou também a avaliação da população sobre a opção de o poder público investir no desenvolvimento de atividades físicas e esportivas ou em outra área (saúde, educação, etc.) na vizinhança de seu domicílio.

Responda os itens a seguir.

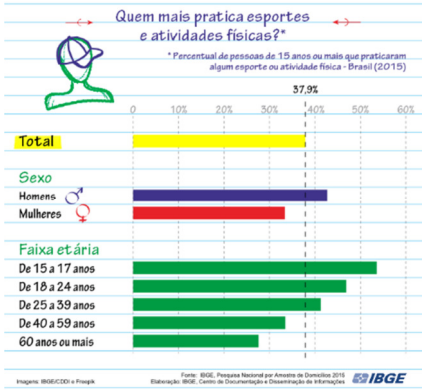
a) Liste pelo menos oito variáveis investigadas na PNAD e no "Suplemento de Práticas de Esporte e Atividade Física" da PNAD 2015, baseando-se no texto apresentado.

b) Das variáveis citadas no item anterior, quais delas apresentam respostas não numéricas?

c) Das variáveis citadas no item a), quais delas apresentam respostas numéricas?

2.1 Análise de infográfico

A seguir, um infográfico produzido pelo IBGE, usando os dados do Suplemento Prática de Esporte e Atividade Física da PNAD 2015 é apresentado.



PNAD - Infográfico 1

a) Segundo a pesquisa, qual a porcentagem de pessoas de 15 anos ou mais que praticaram algum esporte ou atividade física no período de um ano?

b) O título genérico deste infográfico, a saber, "Quem mais pratica esportes e atividades físicas? - Percentual de pessoas de 15 anos ou mais que praticaram algum esporte ou atividade física-Brasil (2015)", diz respeito à população brasileira de 15 anos ou mais ou à amostra coletada?

c) Com base nas recomendações médicas sobre a prática de atividades físicas para se ter boa saúde, como você avalia o resultado obtido na pesquisa para a população brasileira de 15 anos ou mais?

d) Considerando homens e mulheres separadamente, percebe-se alguma diferença com relação à prática de atividades físicas? Em caso afirmativo, descreva a(s) diferença(s) observada(s).

e) Considerando as faixas etárias discriminadas no infográfico, percebe-se alguma diferença com relação à prática de atividades físicas? Em caso afirmativo, descreva a(s) diferença(s) observada(s).

3. ORGANIZANDO AS IDEIAS

Alguns conceitos importantes da Estatística serão apresentados a seguir.

3.1 Conceitos Básicos

Em geral, a palavra **população** representa um conjunto de habitantes de um determinado lugar. No entanto, em Estatística, população tem um sentido mais amplo e pode ser definida como o conjunto de todos os elementos com pelo menos uma característica em comum. Observe que é exatamente essa característica em comum que vai definir o universo (a população) de uma pesquisa.

Assim, em Estatística, a população não precisa ser um conjunto de pessoas, pode ser o conjunto de parafusos fabricados por uma indústria, o conjunto de animais de certa espécie que vivem em uma região, todos os estudantes universitários de um país, etc.

Amostra: é um subconjunto não vazio da população.

Cada uma das unidades investigadas em um estudo estatístico é denominada elemento. Assim, cada domicílio e seus residentes são elementos na atividade da PNAD.

Cada característica observada de um elemento é uma variável estatística. Assim, na atividade da PNAD, estão presentes várias variáveis estatísticas de interesse do domicílio e de seus residentes tais como local, número de cômodos, número de residentes; sexo, idade e rendimento dos residentes, etc.

Suponha que se deseja investigar a opinião dos estudantes de um colégio quanto à modificação da lista de produtos vendidos na cantina para outros mais saudáveis, trocando refrigerantes por sucos naturais entre outros. Para isso, a direção da escola irá entrevistar cinco alunos sorteados de cada uma de suas 40 turmas. Nesse exemplo, a população corresponde a todos os estudantes deste colégio e, a amostra, aos 200 estudantes que foram entrevistados. Cada estudante entrevistado é um elemento e, a variável de interesse é a opinião do estudante: "a favor" ou "contra" à mudança. Num estudo desse tipo, costuma-se registrar também outras variáveis como sexo, idade, ano de ensino, turno, etc.

Parâmetro: é uma característica numérica da população.

Estimador: é uma função que produz estimativas de parâmetros usando as variáveis observadas na amostra.

Voltando ao exemplo anterior, sobre a modificação da lista de produtos da cantina, temos que o parâmetro corresponde à proporção dos estudantes desse colégio que são favoráveis à mudança (na maioria das vezes não acessível, a menos que se realize um censo). O estimador desse parâmetro corresponderá à proporção de estudantes favoráveis à mudança na amostra, que resultará numa estimativa do parâmetro.

As etapas da análise estatística podem ser divididas em duas estruturas básicas: **Estatística Descritiva** e **Estatística Inferencial**. A primeira corresponde a uma exploração das informações que podem ser retiradas dos dados amostrais de modo a reconhecer estruturas que possibilitem futuramente inferir sobre parâmetros de interesse. A segunda consiste em estabelecer modelos probabilísticos para que se possa fazer afirmações sobre a população com algum nível de confiança. Em resumo, a Estatística Descritiva é uma espécie de arqueologia dos dados observados e, a Estatística Inferencial, a indução das informações obtidas da amostra para características da população não observada em sua totalidade.

A PNAD faz uso de inferência estatística, pois ela investiga uma amostra de domicílios em algumas cidades brasileiras, mas propõe estimativas para as características da população brasileira.

Quando se realiza um censo, levantamento de dados de toda a população, não existe a necessidade de fazer uma inferência estatística. No entanto, muitas vezes a realização de um censo é inviável, por várias razões como custo muito alto, tempo muito longo, entre outras.

Proposições são elementos importantes na construção de toda a ciência. No que se refere à natureza da Estatística, em contraponto à natureza da Matemática, podemos destacar dois tipos de proposições.

Uma proposição é dita matemática se é possível classificá-la em verdadeira ou falsa, ainda que essa afirmação seja uma conjectura não provada. Assim, a proposição

"O quadrado de um número par é par." é uma proposição matemática, pois sabemos que ela é verdadeira. Da mesma forma, a proposição

"O triângulo de lados 6, 4 e 3 é um triângulo retângulo." é uma proposição matemática, pois sabemos que é falsa.

Por outro lado, uma proposição estatística é uma afirmação sobre a qual nunca teremos condição de afirmar se é verdadeira ou falsa mas apenas aferir um nível de confiança para ela.

A proposição "Uma moeda, que ao ser lançada 10 vezes, resulta em 10 coroas, não é uma moeda equilibrada." é uma proposição estatística, pois existe a possibilidade de em 10 lançamentos de uma moeda equilibrada obtermos 10 coroas, embora isso seja pouco provável de ocorrer.

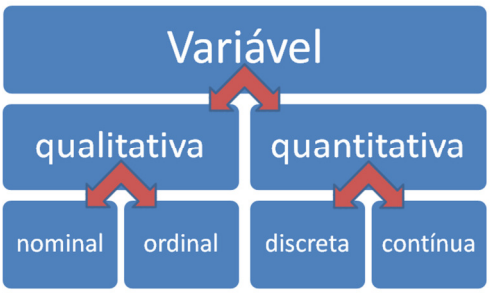
Observação: Uma moeda é dita ser equilibrada se as probabilidades de se obter cara e coroa são iguais. Caso contrário, a moeda é dita ser não-equilibrada.

Se lançarmos 100 vezes essa mesma moeda e obtivermos 8 caras, teremos mais evidências para aceitar a proposição de que não seja equilibrada, mas ainda assim não poderemos afirmar que a proposição seja verdadeira. Proposições desse tipo que envolvem um nível de confiança sobre sua veracidade são proposições de natureza estatística.

3.2 Classificação de variáveis

A classificação das variáveis estudadas é importante, pois as técnicas e procedimentos estatísticos de análise de dados dependem do tipo de variável investigado. Nesse sentido é importante reconhecer a natureza de cada variável investigada para posterior tratamento da informação obtida. Por exemplo, se estamos estudando a modalidade de atividades físicas praticadas pelos brasileiros de 15 anos ou mais, não faz sentido calcular média, pois ela não assume valores numéricos.

Existem dois tipos principais de variáveis (qualitativas e quantitativas), que se subdividem, por sua vez, em duas categorias, conforme a figura a seguir.



Variável qualitativa: uma variável estatística é qualitativa se as possíveis respostas para ela são atributos não numéricos. A maior parte das variáveis identificadas no "Suplemento de Práticas de Esporte e Atividade Física" da PNAD/2015, representa variáveis qualitativas.

Variável qualitativa nominal: uma variável qualitativa é nominal quando não existe nenhuma ordenação natural das respostas associadas à variável. Exemplos de variáveis nominais: bairro de residência, tipo sanguíneo, modalidade de atividade física que pratica, etc.

Variável qualitativa ordinal: uma variável qualitativa é ordinal quando é possível estabelecer uma relação de ordem entre as respostas associadas a ela. Por exemplo,

nível de instrução da mãe com as respostas possíveis: Ensino Fundamental completo, Ensino Médio completo, Ensino Superior incompleto e Ensino Superior completo. Podemos perceber que quem tem Ensino Médio completo tem maior nível de instrução de quem tem Ensino Fundamental completo.

Variável quantitativa: uma variável é quantitativa se as respostas para ela são numéricas. Exemplos de variáveis quantitativas são idade, peso, altura, temperatura, número de irmãos, número de horas semanais dedicadas à prática de atividade física.

Variável quantitativa discreta: as variáveis discretas resultam de uma contagem ou são variáveis cuja quantidade de valores possíveis é finita. Por exemplo, o número de atendimentos em um Pronto-Socorro nos finais de semana, o número de erros de impressão na página de um livro, número de irmãos, etc.

Variável quantitativa contínua: as variáveis quantitativas contínuas em geral resultam de uma medição. Por exemplo, altura, peso, temperatura, etc.

Observação: Uma variável quantitativa pode ser tratada como qualitativa, por exemplo, a idade trabalhada em faixas etárias torna-se uma variável qualitativa ordinal. No entanto, se consideramos a idade em anos completos temos uma variável quantitativa. Por outro lado, também podemos transformar uma variável qualitativa em quantitativa. Considere a variável "prática de atividades físicas" que tem como respostas "Sim" ou "Não". Esse tipo de variável com apenas duas respostas é chamado variável binária e tem uma representação numérica natural. Podemos atribuir o número 1 para a resposta "Sim" e o número 0 para a resposta "Não". Essa estratégia permite somar todas as respostas. Observe que a soma representará o número de pessoas na amostra que praticam atividade física e a "média" representará a proporção de pessoas na amostra que praticam atividade física.

3.3 Gráficos para Variáveis Qualitativas

A primeira etapa na descrição de um conjunto de informações de uma variável qualitativa é organizar as respostas obtidas em função das frequências nas quais elas ocorreram. Essas informações de frequências podem ser representadas pela frequência absoluta (contagem de casos), frequência relativa ou porcentagem (contagem de casos sobre o número total de observações). Para esses tipos de variáveis os gráficos mais comuns são o gráfico de barras em que as barras com larguras iguais são igualmente espaçadas e, seus comprimentos, são dados pelas respectivas frequências, cada barra representando uma resposta da variável qualitativa.

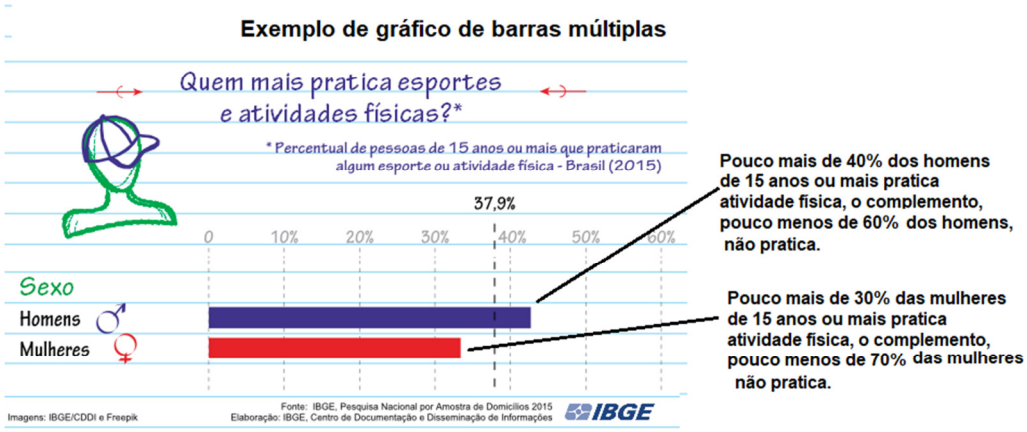
Em geral, se a variável for ordinal dispomos as respostas em ordem crescente. Se a variável é nominal, podemos dispor as respostas em ordem decrescente de frequência.

Gráficos de área também são usados para variáveis qualitativas e, o mais comum, é o gráfico de setores. Cada setor representa uma resposta da variável qualitativa e a medida da área do setor em relação à medida da área do círculo é igual à frequência relativa na qual a resposta ocorreu.

Outra possibilidade de gráfico de área é o gráfico de retângulos no qual o retângulo maior é subdividido em retângulos cujas áreas relativas correspondem às porcentagens das respostas que eles representam.

Observação: Quando estamos trabalhando com variáveis qualitativas usamos a escala da frequência (absoluta, relativa, porcentagem) na construção de gráficos para representar a distribuição de frequências das respostas dadas à variável sob

investigação. As representações gráficas mais comuns são os gráficos de barras e gráficos de setores. Para comparações da mesma variável em grupos diferentes é comum usar o gráfico de barras múltiplas com frequências relativas ou porcentagens.



Como escolher entre o gráfico de setores ou o gráfico de barras para representar a distribuição de frequências de uma variável qualitativa? Se o número de respostas diferentes é grande, maior que 4, ou se as diferenças nas frequências das respostas são pequenas, por exemplo uma tem porcentagem 22% e a outra tem porcentagem 25%, o gráfico de setores não será adequado, pois pequenas diferenças de ângulos não são perceptíveis, enquanto que no gráfico de barras é fácil perceber pequenas diferenças de comprimento das barras. Para fazer comparações múltiplas o gráfico de setores não é adequado. Observe que no infográfico I, os gráficos separados por sexo e por faixa etária, são gráficos de barras múltiplas.

4. GRÁFICOS PARA VARIÁVEIS QUANTITATIVAS CONTÍNUAS

4.1 ATIVIDADE: CONSTRUÇÃO DO HISTOGRAMA

Um arranjo de oito radiotelescópios (A, B, C, D, E, F, G e H) como ilustrado na figura detectou sinais cujos oito registros de tempo para cada radiotelescópio se encontram no quadro a seguir.

A	B	C	D	E	F	G	H
3,03	4,37	5,04	5,73	4,03	5,37	6,04	6,74
3,38	4,46	5,11	5,84	4,38	5,46	6,11	6,84
3,69	4,55	5,19	5,95	4,60	5,55	6,19	6,96
3,78	4,63	5,29	6,08	4,78	5,64	6,29	7,08
3,92	4,71	5,36	6,23	4,92	5,72	6,36	7,23
4,04	4,79	5,45	6,41	5,04	5,79	6,45	7,40
4,16	4,87	5,54	6,62	5,16	5,87	6,54	7,63
4,27	4,95	5,64	6,97	5,26	5,95	6,64	7,97

Como construir uma tabela de frequências desses dados uma vez que os registros de tempo são todos distintos? Como você faria para visualizar o comportamento de uma variável com estas características?

A natureza quantitativa de uma variável contínua pode muitas vezes levar a resultados que praticamente não se repetem. Eles podem ser todos diferentes, como é observado no exemplo. Com o objetivo de identificar alguma estrutura no comportamento deste tipo de variável é necessário agrupar os valores em intervalos de classe, o que permite analisar a sua distribuição de frequências.

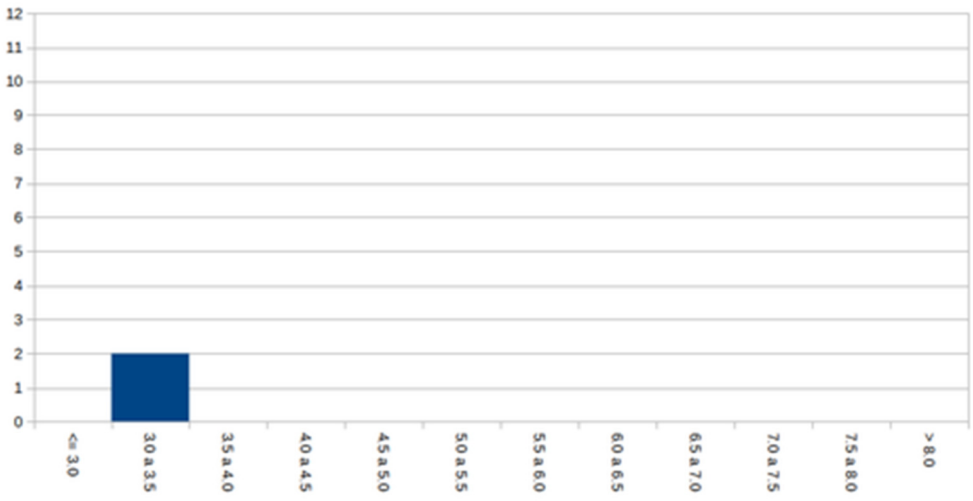
Complete o quadro a seguir, que utiliza intervalos de amplitude 0,5 começando em 3,0. Observe que cada intervalo no quadro é fechado à esquerda e aberto à direita.

Intervalo de classe	Número de observações
[3,0 ; 3,5[	
[3,5 ; 4,0[	
[4,0 ; 4,5[	
[4,5 ; 5,0[	
[5,0 ; 5,5[	
[5,5 ; 6,0[	
[6,0 ; 6,5[	
[6,5 ; 7,0[	
[7,0 ; 7,5[	
[7,5 ; 8,0[	

Para visualizar o comportamento desses dados, iremos construir um gráfico, chamado histograma, que é composto por retângulos adjacentes cujas alturas representam a frequência de observações no intervalo correspondente. A base de cada retângulo corresponde aos limites do intervalo definido no agrupamento dos dados.

Observação: O uso da escala de frequência como altura dos retângulos no histograma é possível apenas quando os comprimentos dos intervalos de classe são iguais. Quando esses comprimentos são desiguais, usa-se a escala da densidade de frequência dada pela razão entre frequência e comprimento do intervalo.

b) Complete a figura a seguir com os demais retângulos do histograma.



Histograma dos dados coletados pela grade de radiotelescópios

c) Calcule a média dos registros e localize-a no histograma construído, sabendo que a soma dos 64 registros de tempo é 351,95. Comente sobre a localização da média no histograma construído?

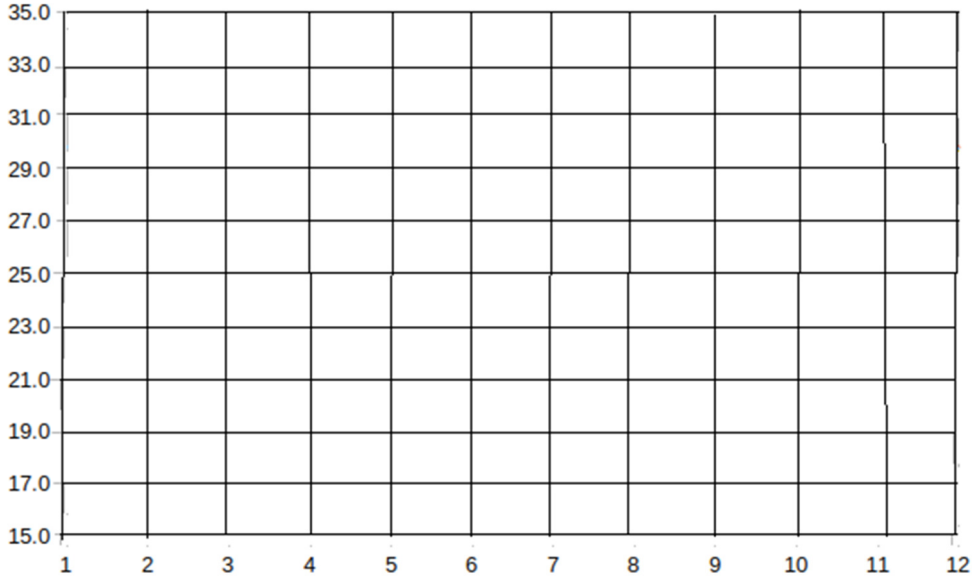
4.2 ATIVIDADE: GRÁFICOS DE LINHA

Dados coletados ao longo do tempo (como a informação meteorológica) são conhecidos como séries de dados temporais ou, apenas, séries temporais, já que correspondem a variáveis que mudam continuamente ao longo do tempo e a informação só é útil se sabemos o momento em que foram realizadas as medições.

O quadro a seguir fornece a média das temperaturas máximas para cada mês nos anos de 1991 a 2000 da cidade de Porto Alegre em graus centígrados.
(Fonte: http://www.inmet.gov.br/portal/index.php?r=bdmep/bdmep)

Mês	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
1	30,23	30,43	31,34	30,33	30,74	29,89	32,09	29,13	30,65	30,63
2	31,03	31,48	29,28	28,85	29,46	29,78	29,62	28,26	29,56	29,93
3	30,55	30,05	28,22	28,05	29,12	28,67	28,63	27,20	31,64	27,85
4	26,15	25,52	27,66	25,51	26,22	27,03	26,56	24,03	24,00	26,32
5	25,31	21,44	23,29	24,33	21,95	22,94	22,95	22,00	21,51	21,78
6	20,32	22,68	19,13	20,09	20,45	17,76	19,42	19,60	18,87	21,50
7	19,75	16,91	17,97	20,41	21,60	16,99	20,67	20,47	18,78	17,59
8	21,81	20,50	21,90	21,28	21,55	22,59	23,06	19,77	21,94	20,85
9	23,99	22,14	20,83	25,21	22,62	21,40	22,32	21,22	22,65	22,25
10	26,17	26,16	26,40	24,60	24,17	25,34	23,27	25,19	23,07	24,02
11	26,93	27,16	28,07	26,53	28,93	28,40	26,51	28,24	26,36	26,87
12	30,60	29,95	29,73	32,05	30,44	29,87	30,28	28,91	29,08	29,51

a) Escolha dois anos diferentes e marque com um ponto as temperaturas do quadro na grade quadriculada usando o mês como abscissa (x) e a temperatura como ordenada (y). Utilize cores diferentes para a série de cada ano.



b) Una os pontos correspondentes ao mesmo ano (mesma série) de meses consecutivos com um segmento e observe o resultado. Você percebe algum comportamento similar para a temperatura em anos diferentes?
c) Compare seu gráfico com o de colegas que escolheram outros anos (ou acrescente séries de outros anos ao seu gráfico). O que você percebe com relação à temperatura nos meses iniciais, intermediários e finais do ano? A que se deve esse comportamento da temperatura?

5. ORGANIZANDO AS IDEIAS

Dois tipos de gráficos para representar variáveis quantitativas contínuas foram apresentados: o histograma e o gráfico de linha.

O **histograma** é uma representação gráfica da distribuição de frequências de uma variável quantitativa contínua agrupada em intervalos usando retângulos adjacentes. Cada retângulo no histograma corresponde a um intervalo considerado e a razão da área desse retângulo em relação à área total do histograma deve ser igual à frequência relativa de casos desse intervalo.

O **gráfico de linha** é uma representação útil quando os dados são uma série temporal, ou seja, os dados são coletados ao longo do tempo. Esse gráfico é construído marcando-se no plano Cartesiano os pontos (x,y) em que abscissa x representa o tempo e, a ordenada y, a variável quantitativa. Os pontos consecutivos são unidos por segmentos.

Quantos intervalos de classe considerar no agrupamento dos dados?

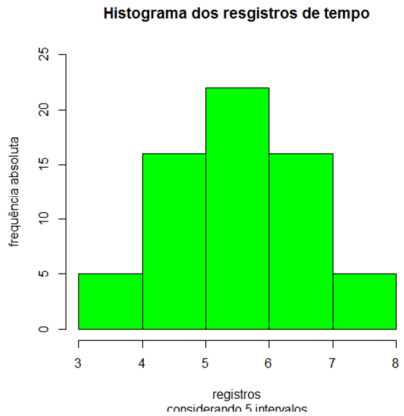
Quando existe a necessidade de agrupar os dados em intervalos, uma questão que se coloca é: quantos intervalos usar para que se possa reconhecer estruturas de frequências nesse conjunto? Não existe uma única resposta para essa questão. No entanto, devemos evitar tanto usar um número reduzido de intervalos, quanto usar um número grande de intervalos. Por exemplo, se usarmos um único intervalo, o histograma seria representado por um único retângulo que nada informaria sobre o comportamento dos dados. Por outro lado, se o número de intervalos for igual ou superior ao número de observações, o histograma potencialmente teria apenas classes com uma única observação e o objetivo de visualizar estruturas dos dados em análise se perderia.

Embora não exista uma resposta única sobre quantos intervalos considerar, alguns autores sugerem usar o número inteiro mais próximo da raiz quadrada do número de observações, outros sugerem usar de 5 a 15 intervalos de amplitudes iguais. No GeoGebra, por exemplo, a função que constrói histogramas permite trabalhar com 3 a 20 intervalos.

Em que situações há a necessidade de considerarmos intervalos de amplitudes desiguais?

Normalmente, na primeira construção dos intervalos consideramos sempre intervalos de amplitudes iguais. Mas pode acontecer, nesse agrupamento, intervalos vazios ou intervalos com um número muito grande de observações. Quando essas situações ocorrem recomenda-se juntar dois intervalos consecutivos no primeiro caso ou subdividir o intervalo no segundo caso.

Observação: O gráfico de barras não é um histograma, apesar de suas representações serem parecidas. Os gráficos de barras são úteis para descrever a distribuição de frequências de uma variável qualitativa. Nesse gráfico só há um eixo com escala que corresponde aos valores das frequências das categorias (respostas) da variável. As barras podem ser tanto verticais como horizontais e são apresentadas de forma igualmente espaçada. Cada barra representa uma resposta da variável qualitativa e a altura da barra corresponde à frequência daquela resposta. O mais comum é dispor as respostas em ordem decrescente de frequência. Esse tipo de gráfico também pode ser usado para representar uma variável quantitativa discreta, sendo que nesse caso, as posições das barras correspondem aos valores assumidos pela variável. Pela natureza discreta da variável, as barras não são adjacentes e, pela natureza quantitativa da variável, o posicionamento das barras não é livre. Os histogramas são úteis para representar a distribuição de frequências de uma variável quantitativa contínua cujos valores foram agrupados em intervalos. No histograma, o eixo das abscissas (horizontal) representa a escala da variável contínua e, o eixo das ordenadas (vertical) representa a escala da frequência ou densidade de frequência que é definida como a razão entre a frequência e a amplitude do intervalo.



Não podemos variar livremente a posição dos intervalos nesse gráfico. Ele revela uma estrutura importante desses dados, a saber, os registros de tempo ocorrem com maior frequência nos intervalos intermediários (de 4 a 6) e com frequência bem menor nos intervalos extremos (de 3 a 4 e de 7 a 8).

6. Noções básicas sobre seleção de amostras

Quando queremos estender nossas observações provenientes de uma amostra para a população é importante ter cuidado na sua seleção, pois ela deve ser representativa da população. Embora não seja nosso objetivo aqui descrever métodos variados de seleção de amostras, cabe destacar que existem dois tipos principais de seleção: os probabilísticos e os não probabilísticos.

O primeiro tipo é fundamental para que seja possível avaliar a incerteza das conclusões devido à amostragem tais como margem erro e nível de confiança. Nesse tipo de seleção de amostra, conhecemos a probabilidade de seleção dos elementos da população na amostra. Entre os métodos probabilísticos mais comuns destacam-se

a) amostragem aleatória simples: todas as amostras de igual tamanho têm probabilidades iguais de serem selecionadas.
b) amostragem estratificada: a população é dividida em grupos de elementos homogêneos (similares nas características a serem investigadas) e os grupos são heterogêneos entre si. A amostra é composta por amostras aleatórias simples de cada grupo, em geral, proporcionalmente aos tamanhos dos grupos.

c) amostragem por conglomerados: a população é subdividida em conglomerados (subpopulações). Uma amostra aleatória simples de conglomerados é obtida e, em seguida, todos os elementos dos conglomerados escolhidos são observados.
d) amostragem sistemática: toda a população deve estar catalogada numa lista, por exemplo, lista dos alunos matriculados numa escola em ordem alfabética. Suponha que a lista contenha 1000 alunos e que se deseja obter uma amostra de tamanho 50. Para isso divide-se 1000 por 50 obtendo-se 20 blocos de 50 alunos. Sorteia-se ao acaso um número de 1 a 20, por exemplo, o número 9. Seleciona-se o aluno de número 9 e, depois, os próximos elementos são selecionados de 20 em 20 como uma Progressão Aritmética de razão 20 e primeiro termo 9.

Os casos mais comuns de métodos não probabilísticos são amostragem por conveniência e amostragem por julgamento. A amostragem por conveniência caracteriza-se por não ter um plano particular de amostragem. O objetivo nesse caso não seria generalizar conclusões e sim descrever as características principais do grupo de estudo. Nas amostras por julgamento, os elementos da amostra são escolhidos por um especialista no assunto sob investigação. A grande desvantagem dos métodos não probabilísticos é a impossibilidade de avaliar incertezas devido à amostragem.

Exemplo: A direção de uma escola de Ensino Médio deseja realizar uma pesquisa para conhecer a opinião de seus 520 alunos sobre a antecipação em 30 minutos dos horários de seus turnos. Para tanto, devem decidir entre as seguintes estratégias de seleção de amostra.

a) 40 alunos considerando os primeiros a chegar à escola na segunda-feira. Temos, nesse caso, uma amostra de conveniência, pois a probabilidade de seleção dos alunos não é determinada no plano de amostragem: selecionar os 40 primeiros. Observe também que, esse esquema de seleção não parece razoável para essa pesquisa, pois é possível resultar numa resposta viesada, isto é, tendendo a favorecer a mudança de horário por considerar apenas os primeiros a chegar, não representando necessariamente a opinião da maioria dos 520 alunos da escola.

b) 40 alunos escolhidos a partir do cadastro de 520 alunos matriculados da seguinte forma: como 520/40=13, sorteia-se ao acaso um número entre 1 e 13, por exemplo 8; e, depois, seleciona-se do cadastro, os alunos nas posições 8, 21, 34, 47, e, assim sucessivamente de 13 em 13, até o aluno de posição 515 no cadastro de alunos, totalizando 40 observações. Temos, nesse caso, uma amostra sistemática cuja probabilidade de seleção é conhecida, a saber, 1/13.

c) 40 alunos sendo 16 do primeiro ano, 14 do segundo ano e 10 do terceiro ano, escolhidos ao acaso, tendo em mente que na escola 40% dos alunos são de primeiro ano, 35% dos alunos são de segundo ano e 25% dos alunos são de terceiro ano. Trata-se de uma amostra estratificada proporcionalmente ao tamanho dos estratos que correspondem aos anos do Ensino Médio. As probabilidades de seleção de amostra são conhecidas.

d) 40 alunos de uma turma do segundo ano na qual está um filho do diretor. Trata-se de uma amostra de conveniência e que pode resultar num resultado duplamente viesado, tanto pelo fato de que a turma escolhida não representa necessariamente a maioria dos estudantes da escola, quanto pela presença do filho do diretor nessa turma que pode influenciar o resultado.

7. Medidas Resumo: continuação

Após organizar os dados em tabelas e gráficos, os próximos passos envolvem resumir a informação obtida por meio de algumas medidas. Por exemplo, a partir de um conjunto de dados quantitativos pretende-se responder as seguintes questões:

a) É possível encontrar valor(es) para resumir as observações? Qual(is) seria(m) este(s) valor(es)? Como encontrá-lo(s)?
b) Como medir se os dados estão "próximos" ou "afastados" uns dos outros?
c) Como você classifica a forma do gráfico construído para representar os dados?
d) Existe algum valor muito diferente dos demais? Como identificá-lo?

As respostas a essas questões são tratadas no segundo capítulo de Estatística do Projeto Livro Aberto de Matemática, “Medidas de Posição e Dispersão”.

Referências Bibliográficas:

Batanero, C., & Borovcnik, M. (2016). Statistics and probability in high school. Springer.

Bussab, W. O. & Morettin, P. A. (2017). Estatística Básica. Saraiva. Nona edição.

IBGE (2017) <https://vamoscontar.ibge.gov.br/> Acesso em: 29 ago. 2017.

Franklin, C. A. (2007). Guidelines for assessment and instruction in statistics education (GAISE) report: A pre-K--12 curriculum framework. American Statistical Association.

Rossman, Allan J., and Beth L. Chance. (1998). Workshop Statistics:: Discovery With Data and Minitab. Springer Science & Business Media.

Sugestão de vídeos:

O Prazer da Estatística - https://www.youtube.com/watch?v=nB5l9OW2eyo
O que é Estatística? - https://www.youtube.com/watch?v=-Wm9cxiXUe0
Ação, Reação, Correlação - http://m3.ime.unicamp.br/recursos/1043