ceadUFV
Coordenadoria de Educação Aberta e a Distância

UFV
Universidade Federal de Viçosa

CURSO TÉCNICO EM ADMINISTRAÇÃO

ESTATÍSTICA BÁSICA

Hernani Martins Júnior
Riciane Leão

Universidade Federal De Viçosa
Reitor: Demetrius David da Silva
Vice-Reitora: Rejane Nascentes

Coordenadoria de Educação Aberta e a Distância
Diretor: Francisco de Assis Carvalho Pinto

Organizadores:
Hernani Martins Júnior
Riciane Leão

Edição de Conteúdo e CopyDesk:
João Batista Mota

Layout:
Antônio dos Santos

Editoração Eletrônica:
Beatriz Fonseca

ceadUFV Coordenadoria de Educação Aberta e a Distância
UFV Universidade Federal de Viçosa

Apresentação

Há pouco mais de um século havia algumas profissões muito comuns. Duas que aqui trago para nossa reflexão foram: Acendedor de Lampião; e Guarda-chaves de Companhia Ferroviária. O acendedor de lampião era a pessoa que no entardecer corria pelas vias acendendo luminárias existentes nas ruas. Foi provavelmente uma profissão muito conhecida em Conceição do Mato Dentro nos idos de 1800. O Guarda-chaves era o cargo designado ao responsável pelas chaves de uma estação ferroviária, era responsável por guardar as chaves da estação.

Provavelmente nossos genitores mais remotos tiveram contato com estas profissões ou com estes profissionais. Hoje elas não mais existem, foram sufragadas pelas revoluções tecnológicas e não se fazem mais necessárias. A primeira deixou de existir pelo advento da energia elétrica e posteriormente pelos sensores de luminosidade que automatizam o processo da iluminação pública. A segunda também foi suplantada por sistemas eletrônicos, cartões magnetizados de acesso, cadastramento eletrônico, enfim, também não se faz mais necessária.

Tem sido assim desde sempre na história da humanidade e vai continuar sendo. Algumas coisas que agora são muito importantes e valiosas perderão o valor e depois a existência. Um breve olhar pela internet e é possível encontrar especialistas que conjecturam a respeito das profissões do futuro e as profissões que ficarão no passado, aquelas mais promissoras, que as vezes nem existem ainda e aquelas que deixarão de existir. Dentre as profissões apontadas como profissões do futuro está a do Estatístico que neste contexto recebe um novo nome: o Analista de *Big Data*. E é fácil saber o porquê desta profissão ser tão promissora. Se eventualmente você deseja comprar uma ferramenta pela internet, ou intenta comprar uma passagem aérea ou mesmo um produto de beleza, certamente você receberá reiteradas ofertas e mensagens de alerta nos acessos seguintes.

O que está havendo é uma descrição remota do seu comportamento como consumidor e isto realmente importa para o mundo empresarial. Se eu gosto de viajar e não gosto de comprar ferramentas, uma vez captado este meu comportamento, eu receberei alertas de viagens e certamente não receberei alertas de ofertas de ferramentas, pois meu perfil remoto já foi identificado por um programador analista de *big data*.

Esta sistemática possibilita maximizar as vendas e otimizar os esforços de marketing, melhorando o desempenho econômico das corporações, e é justamente por isto que a estatística se desponta como uma das profissões do futuro

Estatística é isto, é entender o estado das coisas, é entender como as coisas estão. Sua relação com a administração é íntima pois serve de ferramenta para o processo de tomada de decisões. Quem não mensura, quem não conhece uma situação não conseguirá administrar, pois faltará mecanismos que auxiliem o processo de tomada de decisão.

Neste curso nosso objetivo é desenvolver algumas ferramentas básicas nos permitam conhecer determinadas situações. Tal processo de conhecimento subsidia o processo de tomada de decisão. Começaremos com as noções de probabilidade, em seguidas com conceitos de População e Amostra, seguidos de técnicas de análise descritiva de dados. Terminamos nosso curso estudando algumas medidas de posição e de dispersão, que nos auxilia na descrição de populações, que melhoram o entendimento de determinadas situações, mais uma vez servindo de ferramenta para o processo de tomada de decisão.

As ferramentas aqui desenvolvidas possibilitam maior confiança nas atividades de gestão, permite um comportamento pautado pelo profissionalismo e pela ciência, ajuda a desconstruir mitos e intuições falsas. Que estas descobertas possam nos fazer crescer, no trabalho e na vida. Desejo a todos os alunos, bons estudos!

Professor Hernani Martins Júnior

Significado dos ícones da apostila

Para facilitar o seu estudo e a compreensão imediata do conteúdo apresentado, ao longo de todas as apostilas, você vai encontrar essas pequenas figuras ao lado do texto. Elas têm o objetivo de chamar a sua atenção para determinados trechos do conteúdo, com uma função específica, como apresentamos a seguir.

DESTAQUE: são definições, conceitos ou afirmações importantes às quais você deve estar atento.

GLOSSÁRIO: Informações pertinente ao texto, para situá-lo melhor sobre determinado termo, autor, entidade, fato ou época, que você pode desconhecer.

SAIBA MAIS: se você quiser complementar ou aprofundar o conteúdo apresentado na apostila, tem a opção de links na internet, onde pode obter vídeos, sites ou artigos relacionados ao tema.

PARA REFLETIR: vai fazer você relacionar um tópico a uma situação externa, em outro contexto

EXERCÍCIOS: são momentos para você colocar em prática o que foi aprendido.

Sumário

CAPÍTULO 01

Introdução à probabilidade

1. Histórico

A teoria da probabilidade estuda o comportamento probabilístico de determinados eventos, de um determinado acontecimento específico em situações em que diversos acontecimentos são possíveis. A teoria da probabilidade se desenvolveu como ferramenta de mensuração dos riscos para os jogos de azar, muito comuns na Europa durante a idade média. Neste período é que surge a necessidade de determinar as possibilidades de determinados eventos, quase sempre ligados aos jogos de azar. Os jogos envolviam apostas em dinheiro, desta forma aos apostadores muito importava a mensuração dos riscos ou a previsibilidade de seus lances.

A popularidade dos jogos de azar atraíram a curiosidade de muitos matemáticos, gerando grandes avanços no desenvolvimento da teoria da probabilidade. Os matemáticos italianos Pacioli, Cardano e Tartaglia no século XVI tiveram grande participação nas primeiras considerações referentes aos jogos e apostas. Em seguida muitos outros matemáticos se envolveram no estudo mais aprofundado da probabilidade, criando ferramentas muito utilizadas até os dias de hoje. Dentre os mais importante, podemos citar: Pascal; Bernoulli; Gauss; e Poisson.

Todos estes nomes são referências obrigatórias para o estudo da estatística e da probabilidade até os dias atuais. Seus teoremas e pressupostos continuam válidos e são de compreensão obrigatória para o estudioso desta área.

Pascal e Fermat foram responsáveis pela elaboração da base geral da teoria do cálculo das probabilidades e da análise combinatória, marcando o início da probabilidade como ciência. Para o estudo analisaram simulações envolvendo apostas nos jogos de dados, os estudos possibilitaram o levantamento de diversas hipóteses envolvendo os resultados constatados nos testes estatísticos.

Bernoulli deu continuidade nos estudos no fim do século XVIII, com trabalhos em sua maioria envolvendo os grandes números possibilitou apresentar as combinações, permutações e a classificação binomial.

Laplace em seus estudos formulou a regra de sucessão, dando a ele o nome do pai da probabilidade. Laplace utilizava como análise o sucesso ou fracasso de algo acontecer, um exemplo muito comum utilizado por ele foi a probabilidade de o sol nascerá amanhã. Para a análise Laplace usa como referência o dado de que o sol sempre nasce de manhã e nunca falhou, desta forma a chances do mesmo acontecer novamente é extremamente grande. O experimento parece um pouco absurdo, porém tem todo o suporte na teoria.

O matemático alemão Gauss estabelecia o método dos mínimos quadrados e a lei das distribuições das probabilidades. Seus estudos levaram à sistematização de uma das mais famosas distribuições de probabilidade, chamada distribuição normal ou também chamada distribuição gaussiana, carregando a insígnia de seu criador.

A probabilidade ganhou muito espaço no âmbito dos governos, empresas e organizações, ajudando a desenvolver estratégias de suporte no processo de tomada de decisão, analisando-se os riscos de cada escolha. O estudo da probabilidade tem seu uso em todos os ramos do saber: Administração, contabilidade, engenharia, medicina, política, biologia. Em todos os ramos a probabilidade é objeto de interesse, para definição de qual a melhor estratégia, de qual o melhor procedimento.

2. Espaço Amostral e Evento

Todo o indivíduo, em seu dia a dia, utiliza-se da probabilidade de forma intuitiva. Seja na escolha do barzinho de sexta à noite, seja pensando no clima, na temperatura ao longo do dia, enfim, seja nas coisas mais corriqueiras, naquelas que se pratica sem se dar conta, utiliza-se de um conjunto de experiências já vividas que constituem o conjunto das possibilidades possíveis, como meio pelo qual os indivíduos se nortearão no processo de tomada de decisões. Uma boa referência para este assunto pode ser encontrada em Meyer (1983).

Sempre que se trabalha com probabilidade também se trabalha com a incerteza. A escolha do barzinho do final de semana geralmente é feita com base em uma informação previamente acumulada, em alguma experiência vivida que subsidia o processo de tomada de decisão, gerando uma expectativa. Todavia o que se tem é uma expectativa, que pode se confirmar ou não segundo alguns fatores aleatórios. Por isto dizermos que a previsibilidade de determinado evento traz consigo algum nível de incerteza, geralmente ocasionados por fatores desconhecidos ou aleatórios.

Pode-se em um experimento aleatório em que há diversos possíveis eventos, é possível imaginar qual será o evento mais provável, e com qual probabilidade teórica este deve ocorrer. Isto não determina o resultado cabalmente, mas se comprova com a reiteração do experimento. Um bom exemplo disso é o lançamento de uma moeda honesta, onde a probabilidade de cada uma das faces é a mesma. Se realizarmos 1000 lançamentos sobre as mesmas condições espera-se que 500 deles seja cara e os outros 500 seja coroa, porém não podemos afirmar que esse seja o resultado do experimento, e sim apenas uma aproximação probabilística.

Nem todos os experimentos são puramente aleatórios como o lançamento de uma moeda honesta. Para exemplificar este fato, pense na seguinte pergunta:

Em um torneio de xadrez, quais as chances de que o jogador 1 vença o jogador 2?

Para que seja possível responder a essa pergunta é necessário conhecer as características de cada um deles, as quais condicionam o desempenho do jogador, ou seja, número de partidas jogadas, número de partidas ganhas e se são ou não do mesmo nível no torneio. Todos esses fatores influenciam diretamente o resultado final da partida. De qualquer forma, embora os eventos tenham probabilidades condicionadas a determinadas variáveis, o espaço amostral deste experimento é bem simples. Neste experimento apenas estes 3 eventos são possíveis, então S representa o espaço amostral do experimento: $\mathbf{S = \{Vitória\ de\ 1, Empate, Vitória\ de\ 2\}}$.

Via de regra para os cálculos de probabilidade é necessário o conhecimento de todos os possíveis desfechos, de todos os possíveis resultados, de todos os possíveis eventos. A conjunção de todos os possíveis eventos chamamos de espaço amostral, conjunto que por definição, reúne todos os possíveis eventos associados a um dado experimento aleatório. Para o caso dos jogadores de xadrez o espaço amostral é relativamente simples. Tem-se apenas 3 eventos possíveis: vitória do jogador 1; vitória do jogador 2; ou empate entre ambos.

Para o experimento aleatório lançamento de um dado o espaço amostral será $\mathbf{S = \{1,2,3,4,5,6\}}$, ou seja o conjunto S congrega todos os possíveis resultados, constitui portanto o espaço amostral para o experimento aleatório lançamento de um dado. Para o experimento aleatório lançamento de uma moeda o espaço amostral é ainda mais simples, será definido por $\mathbf{S = \{Cara, Coroa\}}$, neste caso $\mathbf{S}$ também é a congregação de todos os possíveis eventos do experimento aleatório lançamento de uma moeda.

Outros possíveis experimentos tendem a ter um espaço amostral mais complexo. Imaginem que o evento de interesse seja o número de clientes que chegam em uma empresa. Neste caso a variável de interesse é uma contagem do número de clientes, e os possíveis eventos serão congregados nos espaço amostral deste experimento aleatório dado por **S = {0, 1, 2, 3, 4, 5, 6, …, 10, 11, 12, … }**, neste exemplo o espaço amostral pode ser um conjunto de evento que tende ao infinito.

Congregação de todos os possíveis eventos constituem o espaço amostral, normalmente representado por **S** ou **Ω**. Embora neste caso possa haver uma relação entre eventos e probabilidades, há diversos fatores diferindo as probabilidades do fenômeno. Mas há, uma infinidade de experimentos que são fenômenos aleatórios, pura e simplesmente. O fenômeno aleatório é qualquer fenômeno cujo processo é governado por leis probabilísticas. Imagine o lançamento de dados, ou o lançamento de moedas, o sorteio de pedras de bingo, ou a roleta de um cassino. Todos estes experimentos têm seus resultados determinados pelo acaso, são puramente aleatórios, governados por leis probabilísticas. Para que se garanta a aleatoriedade do experimento, é necessário o controle de todas as outras variáveis que possam influir no resultado. A aleatoriedade constitui uma fonte de variação cuja causa é o acaso.

Para entender de forma clara os conceitos de Espaço Amostral e Evento comumente utilizamos o Diagrama de Venn. É uma sistematização gráfica que simplifica o entendimento sobre o tema ao mesmo tempo que mostra a relação desta matéria com a teoria dos conjuntos geralmente estuda nas escolas regulares.

Figura 1: Diagrama de Venn

O diagrama de Venn é um grande espaço contendo todos os indivíduos da população. Todos os possíveis eventos poderão ser demonstrados através de um diagrama de Venn.

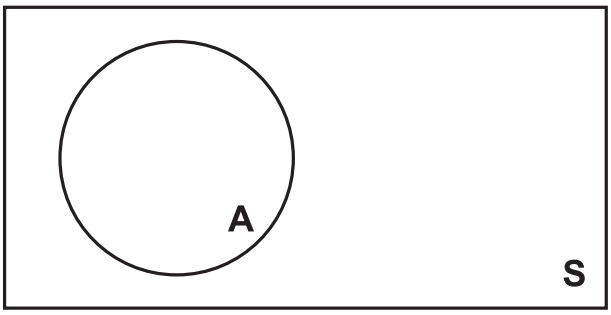


Figura 2: Diagrama de Venn demonstrando a ocorrência do evento A

Da figura 2 é possível observar que uma partição do espaço amostral **S** foi tomada. No círculo da figura temos indivíduos com características comuns e eventualmente podem compor um grupo de interesse ou simplesmente um evento **A**. A ideia aqui é trabalhar o conceito de forma livre, para que o mesmo possa ser aplicado em diversas áreas do conhecimento, em diversas áreas da gestão. Suponhamos que **S** seja o conjunto de trabalhadores de uma empresa, e **A** o conjunto de trabalhadores do sexo feminino. Assim **A** está contido em **S**, que na linguagem dos conjuntos equivale a dizer que **A** pertence a **S**.

Outra ideia que importa resgatar aqui é a ideia do conjunto complementar ou simplesmente evento complementar. Se **A** é um conjunto de indivíduos, todos os outros indivíduos pertencentes a **S** são chamados de A^c. Assim se juntarmos os dois conjuntos **A U** A^c temos o conjunto **S**, ou simplesmente o Espaço Amostral. Os elementos que participam de **A** não participam de A^c, assim como os elementos de A^c não participam de **A**, então concluímos que a intersecção entre os dois conjuntos é um conjunto vazio.

Em termos práticos se **A** representa os indivíduos mulheres, A^c representa todos os indivíduos que não são mulheres, todos os demais. É fácil ver que os indivíduos que participam do conjunto **A** não podem participar do evento A^c então dizemos que estes dois conjunto são mutuamente exclusivos, ou seja, os indivíduos de **A** são exclusivos de **A** assim como os indivíduos A^c de são exclusivos de A^c.

Desenvolvendo probabilidades

Considerando o Evento A demonstrado na Figura 2. Qual deve ser o valor máximo admitido e o valor mínimo admitido para a probabilidade dele ocorrer?

Agora vamos explorar mais um pouco o Diagrama de Venn. Considere que existam dois eventos de interesse **B** e **C** em **S**. Definiremos **B** como sendo o conjunto de funcionários que possuam apenas 2 filhos e **C** como o conjunto de funcionários que possuam apenas 3 filhos. Desta forma poderemos representar nossa nova situação em um diagrama de Venn, trazido na Figura 3.

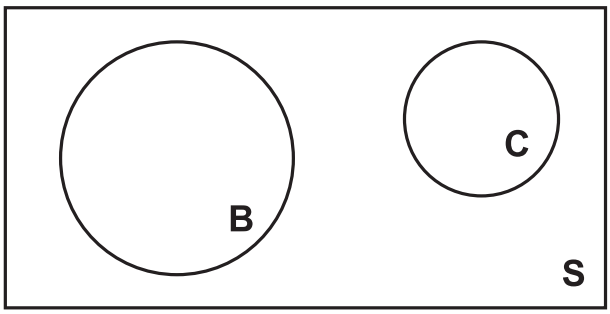


Figura 3: Eventos B e C expressos no Diagrama de Venn

Na figura 3 podemos que B e C também são conjuntos mutuamente exclusivos. Os funcionários que têm apenas dois filhos estão todos em B e os funcionários que possuem apenas 3 filhos estão todos em C, os dois conjuntos não compartilham nenhum dos indivíduos, portanto são conjuntos mutuamente exclusivos.

Desenvolvendo probabilidades

Considerando os Eventos expressos na Figura 3, qual é o evento mais provável dentre B e C?

Neste caso é fácil ver que o conjunto **B** tem maior probabilidade de ocorrer que o evento **C**. O evento **B** abarca uma proporção maior do espaço amostral, sendo assim retirando um indivíduo ao acaso de **S** é mais provável que ele pertença a **B** do que pertença a **C**.

Notem que neste caso é possível a ocorrência de outros eventos que não os eventos **B** e **C**, que neste caso não foram assinalados. Toda esta área de **S** que não pertence a B nem pertence a C representam outros funcionários da empresa que possuem número de filhos diferente de 2 e diferente de 3. Podemos imaginar que há funcionários que possuam 0 filhos, 1 filho, 4 filhos, 10 filhos, e assim por diante.

O diagrama de Venn será usado mais adiante para a elucidação do conceito de independência entre diferentes eventos. Por enquanto ele nos prestou para elucidar a noção de Espaço Amostral, Evento e Probabilidade.

Como vimos, uma forma bem didática de se enxergar a teoria da probabilidade é através da teoria dos conjuntos, disciplina vista e revista ao longo do ensino fundamental. Os eventos são acontecimentos de características peculiares dentre todos os possíveis acontecimentos. Se pensarmos, por exemplo, no clima de um determinado dia, os eventos poderão ser características peculiares deste dia: sol, geada, chuva, vento, etc. Assim numa multitude de possíveis eventos alguns acontecerão em detrimento de outros.

Se na estatística tem-se o espaço amostral como a congregação de todos os possíveis eventos, na teoria dos conjuntos o espaço amostral seria definido pelo conjunto, e os eventos seriam representados por subconjuntos. De forma análoga os subconjuntos são agregações de indivíduos com características peculiares, tomados dentre todos os possíveis indivíduos. Veja o esquema a seguir, seguindo o exemplo dado acima.

Resumindo!

Espaço amostral (S): conjunto de todos os indivíduos que constituirão todos os possíveis eventos.

Evento: Conjunto de indivíduos com características comuns entre si.

Probabilidade: Chance de ocorrência de determinado evento, dentre uma gama de eventos possíveis em S.

3. Definição de probabilidade

A frequência relativa está intimamente ligada com o conceito clássico de probabilidade, na verdade os dois conceitos se fundem. Pode ser exemplificada se pensarmos no seguinte evento, em uma empresa no setor de gestão de pessoas trabalham oito funcionários. Podemos efetuar um experimento aleatório escolhendo ao acaso um dos funcionários de uma empresa.

$$\mathbf{S = \{Camila, Pedro, Talita, Carlos, Paloma, Eduardo, Poliana, Juliana\};}$$

E podemos considerar como evento de interesse o sexo do funcionário escolhido. A pergunta natural que nos vem à mente é:

Qual a probabilidade de escolher um funcionário ao acaso e que ele seja do sexo feminino?

Vamos aos dados:

- Conjunto de possibilidades: temos primeiramente o espaço amostral determinado por oito funcionários $\mathbf{S = \{Camila, Pedro, Talita, Carlos, Paloma, Eduardo, Poliana, Juliana\}}$;
- Conjunto de possibilidades favoráveis: funcionários do sexo feminino que representa um subconjunto do espaço amostral, ou seja, um evento $\mathbf{E = \{Camila, Talita, Paloma, Poliana, Juliana\}}$;

$$\text{Sendo assim temos:} \quad fr = \frac{5\ casos\ favoráveis}{8\ casos\ possíveis} = \frac{5}{8}$$

No exemplo acima há uma probabilidade esperada para determinado evento e o desenrolar do experimento aleatório será controlado por estas probabilidades. Esta operação de comparação

entre o número de casos favoráveis dentre o número de casos possíveis chamamos **Definição Clássica de Probabilidade**.

Veremos também a **Definição Axiomática de Probabilidade** que embora usada para cursos mais formais, nos ajuda a entender conceitos importantes envolvendo a teoria das probabilidades. De acordo com a teoria axiomática de probabilidade, podemos dizer que existe probabilidade se alguns axiomas são satisfeitos.

Já que estamos nos familiarizando com o temos utilizaremos Ω (Símbolo Grego para S) para designar o Espaço Amostral, assim avançamos com maior formalismo.

A partir dessas observações podemos definir que as probabilidades como uma função que atribui possibilidade de ocorrência aos eventos de um espaço amostral **Ω**, portanto, se **A** é um evento de **Ω**, **P(A)** é a probabilidade de **A**, desde que as seguintes condições sejam satisfeitas conforme demonstrado em Magalhães (2006):

$$\text{Axioma I} \qquad P(\Phi) = 0$$

O símbolo **ϕ** corresponde ao conjunto vazio que no mundo dos eventos representa a ocorrência de um evento impossível, ou seja, que não pode ocorrer no espaço amostral considerado.

$$\text{Axioma II} \qquad P(\Omega) = 1°$$

Como **Ω** representa o espaço amostral, sua probabilidade deverá ser igual a 1, é a maior probabilidade possível, diz respeito ao evento que abrange todo o espaço amostral considerado.

$$\text{Axioma III} \qquad 0 \leq P(A) \leq 1°$$

A probabilidade de um determinado evento, sempre estará entre zero e um (0 e 1).

$$\text{Axioma IV} \qquad \sum_{i=1}^{i=n} P\,(Ai) = 1\,, \text{Ai como todos os subconjuntos exclusivos de } \Omega$$

Desta definição é possível ver que não há probabilidade maior que 1, não há probabilidade menor que 0, e que a soma das probabilidades de todos os possíveis subconjuntos mutuamente exclusivos do Espaço Amostral somam no máximo 1.

Considerando-se três eventos:

a. Funcionário ser do sexo feminino - neste evento leva-se em consideração apenas um grupo de indivíduos com características peculiares, qual seja, o sexo feminino.
b. O nome começar com a letra **P** - aqui o evento constitui uma outra forma de partição do conjunto. Todo nome começa com uma letra, e podem ser diversas, mas consideraremos pertencente ao evento B somente aqueles nomes iniciados com a letra **P**.
c. O nome começar com a letra **C** - um outro evento representado por outra partição do conjunto, utilizando da mesma variável utilizado no evento **B**, porém selecionando indivíduos diferentes.

Da associação entre os diferentes eventos, pode-se chegar a outros eventos:

- **A** ∩ **B** (**A** ***interseção com*** **B**) é os eventos em que **A** (***feminino***) **e** **B** (***iniciado com a letra*** **P**) ocorrem simultaneamente. Farão parte deste evento apenas dois indivíduos: Paloma e Poliana. Então equivale a dizer, na notação dos conjuntos que: **A** ∩ **B** = {***Paloma, Poliana***}.
- **A** ∩ **C** (**A** ***união com*** **C**) é o evento em que **A** (***feminino***) **ou** **C** (***iniciado com a letra C***) ocorrem. Neste caso não há exclusão, e sim inclusão, qualquer dos eventos satisfarão a condição.

Se um dos dois eventos ocorrerem (ou ambos): {*Camila, Talita, Carlos, Paloma, Poliana, Juliana*}.

- **A e A^c** são dois subconjuntos muito utilizados.

Para a melhor compreensão dos eventos podemos utilizar o **Diagrama de Venn**, na Figura 4, para mostrar a relação entre eventos.

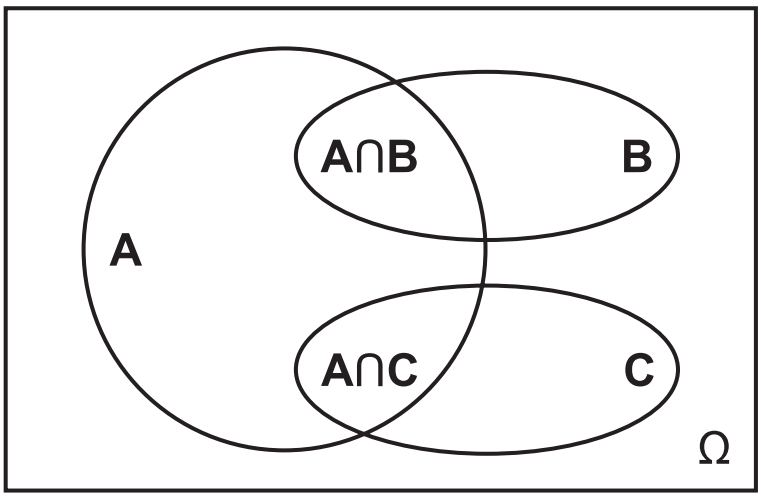


Figura 4: Evento A (mulheres) em associação com Evento B (Nome começado com P) e Evento C (nome começado com C).

Neste caso da Figura 4 os conjuntos possuem sobreposições, representadas pela interseção dos conjuntos. Isto indica que o evento **A** não é independente do evento **B** assim como o evento **A** não é independente do evento **C** pois não se constituem de conjuntos mutuamente exclusivos. Há elementos de **A** que pertencem a **B** assim como há elementos de **A** que pertence a **C**, assim, tanto **A** e **B** quanto **A** e **C** não são exclusivos entre si. Todavia, como podemos observar os eventos **B** e **C** não possuem intersecção, portanto são mutuamente exclusivos e como tal são chamados de eventos independentes entre si.

A partir deste exemplo é possível desenvolver a noção de independência entre eventos que veremos na próxima secção.

Resumindo!

Definição Clássica de probabilidade: Uma probabilidade é uma Razão do tipo:

$$Pr(X) = \frac{Num.\ Elementos\ X}{Num.\ Elementosde\ \Omega}$$

Definição Axiomática de probabilidade: Uma probabilidade existe se atende aos seguintes axiomas:

Axioma I $P(\Phi) = 0$, Axioma II $P(\Omega) = 1$ Axioma III $0 \leq P(A) \leq 1°$

Axioma IV $\sum_{i=1}^{i=n} P(Ai) = 1$, ***Ai*** como todos os subconjuntos exclusivos de **Ω**

4. Independência

Considerando-se dois eventos qualquer: a probabilidade de um dos eventos ocorrerem, ou seja, a probabilidade da União, é dada pela probabilidade do primeiro evento ocorrer somada à probabilidade do segundo evento ocorrer subtraída a probabilidade de ocorrência da intersecção dos dois eventos. Tal teorema é representado abaixo:

$$P(A \cup C) = P(A) + P(C) - P(A \cap C)$$

Todavia, se os eventos são mutuamente exclusivos a intersecção entre eles é representada por um conjunto vazio, e pelo **Axioma I** temos que esta probabilidade é sempre zero, $P(\Phi) = 0$. Desta forma caso os conjuntos sejam mutuamente exclusivos, estamos falando de eventos **independentes**, então a probabilidade de qualquer um dos eventos ocorrem, (Probabilidade da União ocorrer) fica simplificadamente dada por:

$$P(A \cup C) = P(A) + P(C)$$

Já que $P(A \cap C)$ é igual à $P(\Phi) = 0$.

Para ilustrar esta questão considere dois eventos mutuamente exclusivos $\mathbf{A}$ e $\mathbf{C}$ pertencentes ao espaço amostral. Conforme demonstrado na Figura 5.

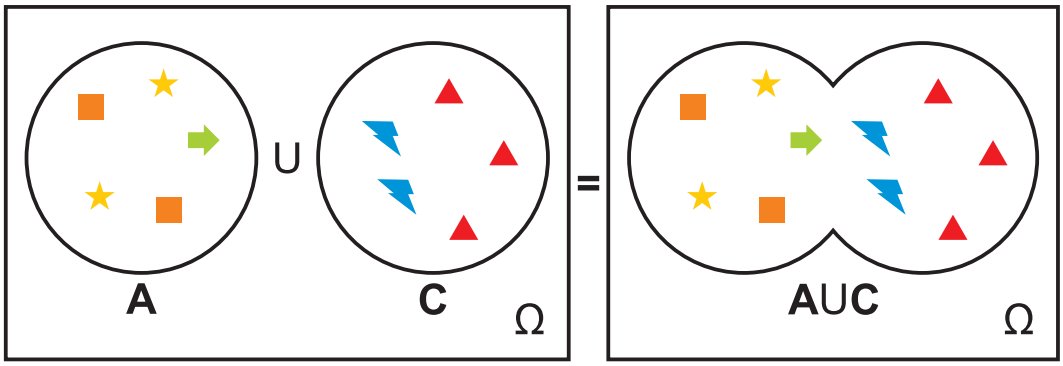


Figura 5: Evento A, Evento C e Evento AUC

$$P(A \cup C) = P(A) + P(C)$$

Podemos notar que A e C são mutuamente exclusivos, pois não apresentam elementos em comum, ou seja, não compartilham elementos em seus conjuntos, sendo assim a probabilidade da união dos conjuntos é igual à soma das probabilidades individuais. Esta é a famosa **Regra do OU.** Quando um evento **OU** outro evento satisfazem a condição e estes eventos são independentes, basta somar as probabilidades dos eventos individuais.

Resumindo!

Veja a Regra do OU: Se A e B são eventos independentes a probabilidade de A ou de B é igual à soma das probabilidades P(A∪C) = P(A) + P(C).

Para casos onde os eventos não são mutuamente exclusivos, na regra da soma devemos nos atentar para que a intersecção esteja sendo contadas duas vezes, caso visto no segundo tópico onde a funcionária Camila se enquadra nos dois eventos (ser do sexo feminino e seu nome começar com a letra C). Sendo assim deve se retirar a probabilidade da intersecção.

$$P(AC) = P(A) + P(C) - P(A \cap C)$$

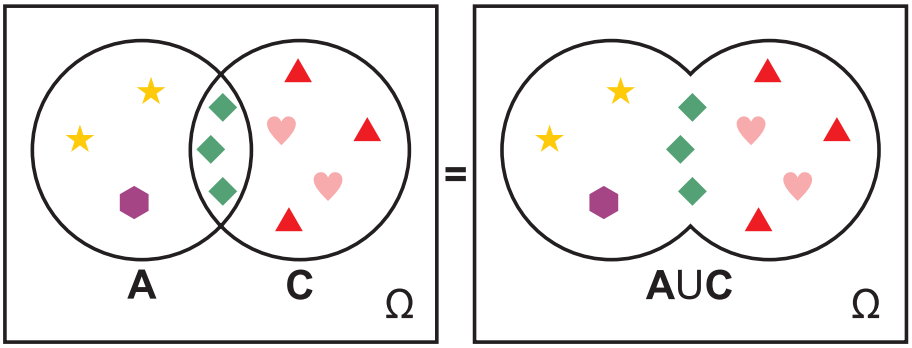


Figura 6: Evento A, Evento C e Evento AUC

Nesse caso os conjuntos A e C apresentam termos em comum, desta forma quando se deseja aplicar a regra da soma, é importante subtrair a interseção (A∩C), caso contrário os elementos verdes serão considerados duas vezes, a primeira quando considerado na probabilidade do evento A e a segunda quando considerado na probabilidade do evento B. Observem que no conjunto A∪C existem apenas 3 elementos verdes.

Vamos para mais um exemplo para uma compreensão mais completa dos conceitos descritos acima. Imaginemos um dado honesto, onde seu espaço amostral é dado por S = {1, 2, 3, 4, 5, 6}.

Sabendo que o dado apresenta seis faces igualmente prováveis, podendo ser definida utilizando a relação de número de casos favoráveis / número de casos possíveis. 998692151… Ronaldo Rufino...

$$P(1) = \frac{num.\ de\ faces\ com\ 1}{num.\ total\ de\ faces} = \frac{1}{6}$$

Portanto, podemos aplicar o mesmo conceito para todas as demais faces, logo:

$$P(2) = \frac{1}{6};\ P(3) = \frac{1}{6};\ P(4) = \frac{1}{6};\ P(5) = \frac{1}{6};\ P(6) = \frac{1}{6};$$

Agora vamos analisar um pouco mais esses dados.

Podemos definir qual a probabilidade de que ao lançar o dado a face que sair seja de um número par?

Para definir o valor da probabilidade desse evento devemos levar em conta todos os eventos que atendem as necessidades do exercício, ou seja, que a face seja de número par. Sendo assim temos que a face pode ser de número igual a 2 **ou** 4 **ou** 6. Como todos os eventos deste experimento são mutuamente exclusivos a probabilidade individual de cada termo de interesse deve ser somada, não havendo nenhuma intercessão a ser subtraída.

$$P(2U4U6) = P(2) + P(4) + P(6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} \qquad P(PARES) = \frac{3}{6} = \frac{1}{2}$$

Portanto temos que a probabilidade de que a face que aparece após o lançamento seja par é de 0,5.

Podemos definir qual a probabilidade de que ao lançar o dado a face que sair seja de um número maior ou igual a 5?

Usando os mesmos conceitos utilizados para exercício anterior temos que:

$$P(face \geq 5) = P(5) + P(6) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

Ou seja, a cada 3 lançamentos, 1 das faces deve ser maior ou igual a 5.

A partir dessa análise podemos definir facilmente qual é a probabilidade de que ao lançar um dado de que o número que sair na face seja menor ou igual a 4.

Utilizando conceitos citados acima podemos definir que a probabilidade total de um espaço amostral é igual a 1, sendo assim podemos excluir da totalidade os termos que não atendem às nossas necessidades, ou seja, a face ser igual a 5 ou 6. Temos então que:

$$P(1U2U3U4) = P(face \leq 4) = 1 - P(face \geq 5) = 1 - \frac{1}{3} = \frac{2}{3}$$

A compreensão sobre o conteúdo descrito acima é extremamente útil para o estudo da probabilidade, auxiliando na compreensão e dedução de eventos posteriores em estudo.

5. Principais regras e modelos

Nesta secção já poderemos avançar um pouco mais. Já é possível entender a regra do **E** e a regra do **OU**. Mas para isto é necessário entender o conceito de **Probabilidade condicional.**

Assim como é na vida é na Estatística. As condições nunca são perfeitas, as coisas se misturam os eventos dependem uns dos outros e desta forma um evento interfere na ocorrência do outro.

Imaginemos que estejamos estudando a variável acidentes de trabalho. Com base em estudos desta natureza foi possível perceber que há um risco bem maior de ocorrência de acidentes de trabalho nos momentos finais da jornada do trabalho. Nestes momentos o funcionário fica mais desatento quanto às normas de segurança diante da ansiedade pelo fim da jornada. Casos como este mostram que exitem variáveis que em conjunto condicionam determinado evento. Neste problema podemos ter diversas variáveis condicionando o evento acidade de trabalho. Só para citar algumas: Horas de descanso; horas trabalhadas no dia; nível de estresse, condição familiar, etc.

A probabilidade condicional se trata do interesse em saber as chances de algo acontecer, sabendo que um evento anterior já tenha acontecido. Quer outro exemplo? Qual a probabilidade de chover em Conceição do Mato Dentro, sabendo estamos em Dezembro. Ou seja, qual a probabilidade da ocorrência do evento A condicionado à ocorrência prévia de B. A representação utilizada para o cálculo é dado por P(A|B) – lê-se probabilidade de A dado B (lembre-se que o símbolo | não corresponde a uma divisão e sim a uma condição de que outro evento já aconteceu) sendo expresso da seguinte forma:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Significa dizer que a probabilidade de ocorrência de A condicionada à ocorrência de B será igual à probabilidade da intersecção entre A e B, dividida pela probabilidade de ocorrência de B (o evento que já ocorreu).

Se for de interesse do estudo o contrário também pode ser obtido, onde estuda-se a probabilidade do evento B ocorrer, visto que A já ocorreu.

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

Vale ressaltar que a operação de interseção é comutativa, ou seja, a ordem dos fatores não alterará o resultado. Portanto:

$$P(A \cap B) = P(B \cap A$$

Para entender melhor vamos a um exemplo: Imagine que deseja-se estudar como a renda da população interfere no habito de posse de carro próprio, sendo assim, foram levantados em um grupo amostral de interesse a renda e se o individuo tem ou não carro próprio. Os resultados obtidos estão descritos na tabela abaixo:

	Méd – baixa (MB)	Alta (A)	Total
Tem carro (C)	3.582	6.246	9.830
Não tem carro (NC)	8.496	1.058	9.554
Total	12.078	7.304	19.384

Podemos considerar, portanto que o espaço amostral (S) corresponde a conjunto de 19.384 pessoas, isso significa que este representa a totalidade da probabilidade, ou seja, P(S) = 1.

Agora vamos entender melhor os dados:

- C = Pessoa que tem carro próprio;
- NC = Pessoa que não tem carro próprio;
- MB = Pessoa de classe média-baixa;
- A = Pessoa de classe alta;
- C∩MB = Pessoa de classe média-baixa que tem carro próprio.
- NC ∪ A = Pessoas que não tem carro próprio ou são de classe alta.

Com essas considerações pode-se determinar alguns eventos, vamos aos cálculos:

$$P(MB) = \frac{nºdepessoasdeclassemédia - baixa}{nºtotaldepessoasentrevistadas} = \frac{12.078}{19.384} = 0.6231$$

Ou seja, o numero de pessoas de classe médio-baixa representa 62,31% do grupo amostral.

$$P(C) = \frac{nºdepessoasquetemcarropróprio}{nºtotaldepessoasentrevistadas} = \frac{9.830}{19.384} = 0.5071$$

Portanto, 50,71% dos entrevistados possuem carro próprio, independente da classe que o individuo faça parte.

Podemos utilizar o conceito de eventos complementares para determinar o valor de outras probabilidades, como por exemplo, o número de entrevistados de não tem carro próprio, e isso pode ser representado a seguir:

$$P(NC) = 1 - 0.5071 = 0,4929$$

Portanto, ter carro (C) é complementar de não ter carro próprio (NC). É importante lembrar que neste caso os eventos são mutuamente exclusivos, pois os indivíduos só podem estar em um dos grupos, não havendo evento compartilhamento entre os eventos.

Já para determinar a probabilidade de um individuo possuir carro e ser de classe média-baixa, ou seja $P(CUMB)$. Note que neste caso os eventos não são mutuamente exclusivos, existem contribuintes que são comuns nas duas situações ao mesmo tempo, desta forma deve-se subtrair a intercessão entre os eventos para que seus dados não sejam contados duas vezes.

$$P(CUMB) = P(C) + P(MB) - P(C \cap MB)$$

Onde:

$$P(C \cap MB) = \frac{nºdepessoascarrocpróprio de classe.med - bai}{nºtotaldepessoasentrevistadas} = \frac{3.582}{19.384} = 0.1848$$

Portanto, temos:

$$P(CUMB) = P(C) + P(MB) - P(C \cap MB) = 0{,}5071 + 0{,}6231 - 0{,}1848 = 0{,}9454$$

Qual a probabilidade de ao sortear um indivíduo do exemplo anterior, sabe-se que esse indivíduo é de classe alta, qual a probabilidade de ele não possuir carro?

Primeiramente, atente-se a informação dada, sabe-se que o indivíduo sorteado faz parte do grupo de pessoas de classe alta. A estrutura da probabilidade pode ser entendida como P(O que quero saber | O que sei), sendo assim tem-se que se deseja calcular $\mathrm{P(NC|A)}$, vamos então aos cálculos:

$$P(NC|A) = \frac{nºdepessoasquenãotemcarropropriodeclassealta}{nºtotaldepessoasentrevistadasdeclassealta} = \frac{1.058}{7.304} = 0.1448$$

Note que o grupo amostral foi reduzido, utilizando somente dados dos entrevistados pertencentes a classe alta, ou seja, a segunda coluna. Como foi informado que o individuo sorteado é de classe alta o espaço amostral que era de 19.384 entrevistados passou a ser mais restrito com apenas 7.304 entrevistados, todos de classe alta com ou sem carro próprio.

Vemos, portanto que $\mathrm{P(NC|A)}$ é dada por:

$$P(NC|A) = \frac{P(NC \cap A)}{P(A)}$$

Esse novo conceito aprendido nos possibilita avançar um pouco mais no conteúdo, definindo a seguir a regra do produto ou **Regra do produto e Eventos independentes**

A partir da probabilidade condicionada, conteúdo definido do tópico anterior, pode-se obter a chamada regra do produto, regra a qual possibilita definir a probabilidade da intercessão entre dois eventos A e B de um espaço amostral definido por: $P(A \cap B)$

Temos pelo conceito de probabilidade condicionada que:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Portanto podemos dizer quer, a probabilidade da intercessão de A e B é dada por:

$$P(A \cap B) = P(A|B).P(B)$$

Quando fala-se que dois eventos A e B são independentes, significa dizer então que $\mathrm{P(A|B) = P(A)}$ ou $\mathrm{P(B|A) = P(B)}$, visto que um evento não interfere no outro.

Sendo assim, se for sabido que A e B são independentes e fazendo a substituição $\mathrm{P(A|B) = P(A)}$ é verdade dizer que:

$$P(A \cap B) = P(A).P(B)$$

Em alguns conteúdos essa regra é conhecida também como a redra do "e", onde quando enunciamos uma probabilidade condicionada utilizamos o termo e como ligante entre os eventos, veja:

Notem que a intersecção entre A e B implica a simultaneidade dos eventos, então os dois eventos têm que ser observados ao mesmo tempo, implica o perfazimento das duas condições.

Resumindo

Veja a Regra do E: Se A e B são eventos independentes a probabilidade de A E B é igual ao PRODUTO das probabilidades
P(AUC) = P(A) ▪ P(C)

Para compreender melhor os conceitos expressos acima, vamos analisar uma situação onde podemos aplicá-los. Considere o seguinte exemplo:

Dado uma urna com três bolas brancas e quatro bolas pretas. Retira-se duas bolas ao acaso, uma após a outra, sem reposição. Deseja-se saber qual a probabilidade de que as duas bolas retiradas sejam da mesma cor?

Para a melhor compreensão do problema vamos dispor as possíveis combinações utilizando o diagrama de árvore, como é visto na figura a seguir:

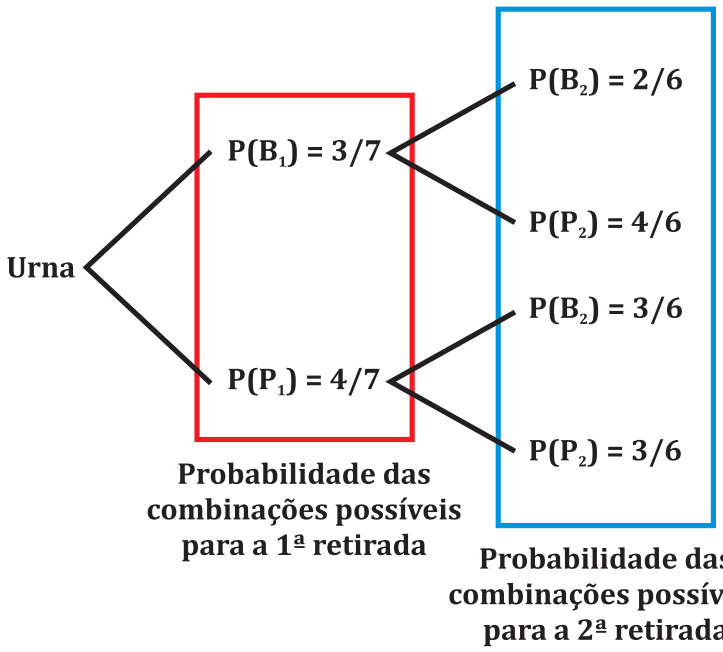


Figura 7: Representação da distribuição de probabilidade para uma urna Fonte: elaborado pelo autor deste livro

Pode-se analisar pelo diagrama todas as possíveis combinações de retirada, com suas respectivas probabilidades. Para determinar o valor da probabilidade de cada dos eventos podemos seguir o seguinte raciocínio:

$$P(B_1) = \frac{nºdebolasBrancanaprimeirarodada}{nºdebolasnaurnanaprimeirarodada} = \frac{3}{7}$$

$$P(V_1) = \frac{4}{7}$$

Para a segunda retirada temos que nos atentar que o nímero de bolas dentro da urna terá diminuído em 1 unidade (seja branca ou preta), pois não há reposição. E que o número de bolas pretas e brancas varia de acordo com o resultado da rodada anterior (condicionante). Sendo assim tem-se:

$$P(B_2) = \frac{nºdebolasBrancanasegundarodada}{nºdebolasnaurnanasegundarodada}$$

Para a primeira situação onde a primeira retirada é uma bolinha branca temos as seguintes combinações:

$$P(B_2) = \frac{2}{6}$$

$$P(V_2) = \frac{4}{6}$$

Já para a segunda combinação onde a primeira retirada foi uma bolinha da cor preta tem-se:

$$P(B_2) = \frac{3}{6}$$

$$P(V_2) = \frac{3}{6}$$

Todos os eventos descritos acima são independentes ou exclusivos, pois cada uma das bolas só pode ser retirada uma vez e apenas uma das combinações pode ocorrer.

Sendo assim é possível determinar a probabilidade de cada uma das quatro combinações possíveis. Retornando ao exemplo, deseja-se saber qual a probabilidade de que ao retirar aleatoriamente as bolinhas tenham a mesma cor. É importante destacar que não é evidenciado a cor da bolinha, apenas que tenham a mesma cor, sendo assim temos que considerar as duas combinações possíveis:

Define-se que $P(A)$ = Probabilidade das bolas serem da mesma cor, logo tem-se que:

$$P(A) = P(B_1 \cap B_2) + P(P_1 \cap P_2)$$

$$P(B_1 \cap B_2) = P(B_1).P(B_2) = \frac{3}{7}.\frac{2}{6} = \frac{1}{7}$$

$$P(P_1 \cap P_2) = P(P_1).P(P_2) = \frac{4}{7}.\frac{3}{6} = \frac{2}{7}$$

$$P(A) = \frac{1}{7} + \frac{2}{7} = \frac{3}{7}$$

Note que utilizamos duas regras apendidas no conteúdo, onde temos probabilidade condicionada e regra da soma, também conhecidas como regra do e (*) e ou (+). Podemos notar observando a descrição a seguir:

$$P(A) = P(B_1 \cap B_2) + P(P_1 \cap P_2)$$

Assim podemos calcular a probabilidade de como sendo a soma das probabilidades dos eventos 'duas brancas' e 'duas pretas'.

Exercícios Teoria da Probabilidade

1) Considerando o conhecimento adquirido nesta secção analise o seguinte conjunto de dados expresso em S e respondas as questões. S={2/3; 5/6; 7/2; 1/10, 2/245; 3/1, 2; 1}

a) Neste conjunto há valores que podem designar probabilidades e outros que não o podem. Quais destes valores não podem ser considerados probabilidades?

b) Justifique sua resposta.

2) Considerando que a probabilidade de um casal possuir um filho homem é ½ e que a probabilidade do filho ser mulher é ½, e que os nascimentos sucessivos dos filhos são eventos independentes. Responda as questões seguintes:

a) Qual a probabilidade de o primeiro filho ser Homem ou Mulher?

b) Qual a probabilidade do havendo dois filhos do casal o primeiro ser homem e o segundo ser mulher?

c) Qual a probabilidade de que tendo o casal 3 filhos, os três sejam mulheres?

d) Qual a probabilidade de que tendo o casal 3 filhos, pelo menos 2 sejam mulheres.

e) Qual a probabilidade de que tendo o casal 5 filhos, os cinco sejam homens?

3) Defina Espaço Amostral.

4) Defina Evento.

5) Construa os espaços amostrais para os seguintes eventos:

a) Lançamento de uma moeda honesta.

b) Lançamento simultâneo de duas moedas honestas.

c) Lançamento de simultâneo de três moedas honestas.

d) Considerando a variável Número de Caras, descreva o suporte nos números Naturais para os experimentos pedidos em 5.1 em 5.2 e em 5.3.

e) Para cada valor possível para a variável em 5.4, construa uma tabela em que haja todos os possíveis eventos da variável Número de Caras e suas respectivas probabilidades.

6) Considere o Lançamento de um dado honesto.

a) Quais os possíveis eventos deste experimento?

b) Quais são as probabilidades destes possíveis eventos?

c) Qual a probabilidade de sair a face 5?

d) Qual a probabilidade de sair a face 5 ou a face 6?

e) Considerando dois lançamentos consecutivos do dado como sendo lançamentos independentes, qual a probabilidade sair duas vezes a face 5.

f) Considerando dois lançamentos consecutivos do dado como sendo lançamentos independentes, qual a probabilidade de sair pelo menos uma vez a face 5.

g) Considerando dois lançamentos consecutivos do dado como sendo lançamentos independentes, qual a probabilidade de sair no máximo uma vez a face 5.

7) Suponha um evento A com probabilidade p=0,6; um evento B com probabilidade p=0,3; e que a probabilidade de A e B ocorrer simultaneamente é p=0,2. Calcule a probabilidade do Evento A U B.

8) Considerando o experimento Lançamento de uma Moeda, calcule a probabilidade do Evento Cara U Coroa.

9) Considerando o experimento Lançamento de um Dado calcule a probabilidade do evento 1 U 2 U 3.

10) Considerando o evento lançamento de uma moeda, qual a probabilidade do evento Cara intersecção Coroa.

11) Suponha um evento A com probabilidade p=0,4; um evento B com probabilidade p=0,3; e que a probabilidade de A e B ocorrer simultaneamente é p=0,2. Calcule:

a) A probabilidade do Evento A U B.
b) Sabendo que o Evento A já ocorreu, qual a probabilidade de ocorrência do Evento B?
c) Sabendo que o Evento B já ocorreu, qual a probabilidade do Evento A ocorrer?

CAPÍTULO 02

População e amostra

1. Conceitos

Para dar início ao conteúdo precisamos definir bem o que é **população** e **amostra**, vamos ao trabalho. População pode ser definida com o conjunto de pessoas, eventos ou itens ao qual você deseja conhecer. As populações podem ser compostas por indivíduos ou por objetos. Podemos citar como exemplo e populações compostas por indivíduos as pessoas pertencentes a um plano de saúde, habitantes de uma cidade ou bairro, proprietários de imóveis, membros de uma equipe de futebol, eleitores de um município ou estado, estudantes, trabalhadores de determinado ramo da indústria, rebanho de determinado estado, componentes da fauna de determinado bioma, etc. Mas como dito a população pode ser composta de elementos inanimados, como por exemplo a produção de aço de uma indústria, os computadores de determinada empresa, peças de um automóvel, queijos de determinada região, árvores de determinada floresta e assim por diante.

Uma das maneiras primárias de se conhecer uma população é através de um **Censo** que consiste em levantar e conhecer **todos** os indivíduos componentes de determinada população. Acontece que muitas das vezes as populações são sobremodo grandes e há grande dificuldade de se acessar todos os indivíduos da população. Imagine por exemplo a dificuldade que é realizar um censo da população do Brasil, um país com dimensões continentais, com diferentes configurações geográficas. São mais de 220 milhões de habitantes espalhados nos mais diferentes rincões das variadas regiões do país. Temos montanhas, pampas, pantanal, florestas, caatinga, grandes metrópoles, e em todos estes ambientes é possível encontrar pessoas que necessariamente deveriam ser levantas na hipótese de um censo. Este procedimento envolve um enorme contingente de pessoas, demanda treinamento e é demorado, além do mais todo este contingente de pessoas e a complexidade de gerenciamento do procedimento torna o censo um procedimento caro, e muitas das vezes impossível.

Desta forma, em muito dos casos a coleta de dados de uma população inteira se torna inviável, e para sanar este problema utilizamos a **amostragem**, uma técnica que permite conhecer uma população através de estimadores.

A **amostra** é a parte ou fração que é realmente estudada e utilizada como representante da população, sendo assim é muito importante a determinação do grupo de amostra, o mesmo deve atender a critérios estatísticos e metodológicos. Estes parâmetros variam de acordo com o método de pesquisa adotado e ao tipo de população de estamos trabalhando.

Na imagem abaixo (Figura 8) temos um exemplo visual do que seria a população e a amostra:

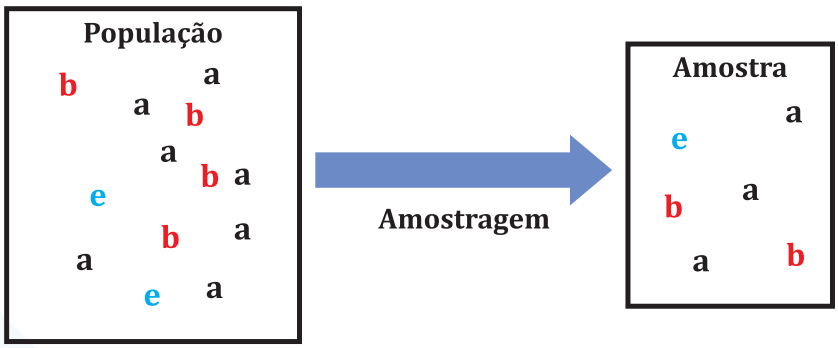


Figura 8: Esquema de uma população e uma amostra

Conforme podemos ver a população é o conjunto do TODO, enquanto que a amostra é parte da população. Neste exemplo temos uma População com N=13 e uma Amostra com n=6. Para fins didáticos os tamanhos da população e da amostra foram diminuídos, mas na prática estes conjuntos costumam ser maiores.

Acontece que quando procedemos uma Amostragem e tomamos esta Amostra como base para se conhecer a população, temos apenas uma ideia de como seria a População. É como se víssemos uma imagem por um espelho e não diretamente. Daí a razão de que não podemos considerar as grandezas obtidas na amostra como se fossem da mesma natureza das grandezas populacionais.

População e Amostra

Amostra é o meio pelo qual se intenta conhecer uma população.
Problema!!! uma população pode originar uma enorme combinação de Amostras…..

2. Parâmetro, estimador e estimativa

A população é fixa e determinada e como tal seus valores de referência não variam, daí por serem valores fixos, confiáveis, convencionamos chamá-los de **Parâmetros.**

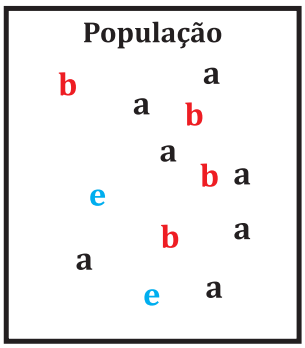
No caso da amostra, cada amostragem que fizermos de forma aleatória gera uma amostra diferente, por que os indivíduos sorteados para compor a amostra nem sempre serão os mesmos, desta forma temos uma amostra que varia em função de aspectos aleatórios. A função utilizada na amostra para estimar o valor do parâmetro populacional, convencionamos chamar de **Estimador.** Um nome bem apropriado para uma função que estima o valor do parâmetro.

A leitura de uma amostra nos dá apenas uma **Estimativa** do parâmetro, e se amostramos novamente a população a amostra obtida será diferente e como tal a estimativa obtida a partir dela também será diferente.

Voltemos ao Exemplo da Figura 8. Vamos tentar descrever como seria nossa população se fôssemos fazer um **senso**.

Descrição da população

- população composta de letras;
- tamanho da população N=13;
- Três cores presentes: Preto; Vermelho e Azul;
- Proporção de letras pretas prop=7/13;
- Proporção de letras vermelhas prop=4/13;
- Proporção de letras azuis prop=2/13.



Neste esquema podemos ver uma descrição completa da população, uma descrição que a retrata fidedignamente. Todavia na prática o que temos é uma amostra e pela amostra a descrição seria um pouco diferente, conforme mostrado na Tabela 1.

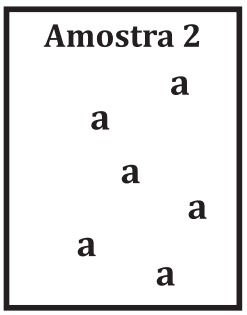
Se o processo de amostragem é bem feita obtemos uma **amostra representativa** que de fato descreve bem a população. Um exemplo de amostra representativa é esta apresentada na figura 8. Cujas **Estimativas** são bem próximas dos **Parâmetros**.

Descrição da população	Descrição obtida pela Amostra
• População composta de letras; • Tamanho da população N=13; • Três cores presentes: Preto; Vermelho e Azul; • Proporção de letras pretas prop=7/13; • Proporção de letras vermelhas prop=4/13; • Proporção de letras azuis prop=2/13;	• População composta de letras; • Tamanho da Amostran=6; • Três cores presentes: Preto; Vermelho e Azul; • Proporção de letras pretas prop=3/6; • Proporção de letras vermelhas prop=2/6; • Proporção de letras azuis prop=1/6.

Mas devido a fatores aleatórios cada amostra tende a se aleatória e pode fornecer diferentes estimativas. Nada impede que a amostra obtida seja diferente. Como por exemplo a obtida nesta segunda amostra.

Descrição obtida pela Amostra 2

- População composta de letras;
- Tamanho da Amostra=6;
- Todos da cor Preta;
- Proporção de letras pretas prop=1;
- Proporção de letras vermelhas prop=0;
- Proporção de letras azuis prop=0.



Não é difícil ver que neste caso as **Estimativas** obtidas são bem diferentes dos valores **Paramétricos**. Esta discrepância pode de dar por acaso ou por um processo de amostragem defeituoso. Se amostrarmos mal pode ser que tomemos apenas alguns elementos específico para a composição de nossa amostra. Isto seria catastrófico, pois distorce a realidade populacional. Todavia, nada impede, que mesmo que haja um processo de amostragem criterioso, tenhamos uma amostra que não representa bem a realidade da população.

Para bem descrever uma população utilizamos **Variáveis** diversas, geralmente tomada por conveniência do pesquisador. Cada área da ciência possui suas variáveis de interesse. Uma criança quando nasce é mensurada segundo algumas variáveis, geralmente mede-se o peso a altura, a pressão sanguínea, o tempo gestacional e a frequência cardíaca. Se estamos numa indústria de metalurgia nossas variáveis de interesse são diferentes, poderiam ser: quantidade produzida; custo da matéria-prima; resistência de uma barra de aço à tração, percentual de pureza, número de funcionários, dentre outras. Se estamos trabalhando na recuperação de uma área degradada as variáveis serão próprias deste tipo de atividade: tipos de contaminantes presentes na área; número de espécies da fauna; números de espécies da flora; pluviosidade; declividade; temperatura, dentre outras variáveis ambientais.

3. Tipos de Variáveis descritivas

Para o estudo da população inteira (censo) ou em uma amostragem precisamos definir muito bem as variáveis que serão importantes para o trabalho de levantamento. São estas variáveis que descreverão a população segundo os critérios de interesse do pesquisador. Em relação à sua natureza classifica-se as variáveis como: qualitativas (ordinais ou nominais) e quantitativas (discretas ou contínuas). A classificação é extremamente importante, pois a mesma permite que você defina, posteriormente, o tipo de teste estatístico e as distribuições de probabilidade a serem utilizadas.

Variável qualitativa: As variáveis qualitativas são aquelas que não podem ser representadas numericamente, a informação de interesse é uma característica do individuo analisado, bom

exemplos de variável qualitativa são: Cor dos olhos, classe social, estado civil, nome da empresa, sexo, cor da pele, sexualidade e etc. Essa variável pode ser dividida em duas variáveis distintas, nominal ou ordinárias, que serão explicadas a seguir.

Nominal: São relacionados a características que não apresenta, nenhuma hierarquia ou ordem, como por exemplo, sexo dos funcionários, estado civil, naturalidade e etc.

Ordinal: Ao contrário da variável qualitativa nominal, a ordinal apresenta hierarquia e um ordenamento bem definido, bons exemplos disso são a hierarquia do funcionário em uma empresa, ranking das empresas em relação ai nível de faturamento (primeira, segunda, terceira e etc), período no qual o aluno se encontra matriculado, ordem de nascimento do indivíduo, etc.

Variável quantitativa: Está ligada a observações que podem ser contadas e representadas numericamente, exemplos disso são: quantidade de água consumida em um mês, número de proprietários que não pagaram IPTU, número de pessoas atendidas em um estabelecimento, dentre outras. Essa variável também se divide em duas bem distintas, a discreta e continua, sejamos a seguir:

Discreta: estão relacionados a observações contáveis, que estão associadas a valores representado pelo universo dos números Naturais, geralmente são números inteiros e positivos. Geralmente descrevem variáveis de contagem, que expressam contagens em números inteiros. Podemos exemplificar a variável discreta com o número de pessoas que pagaram seus empréstimos em dia, pessoas residentes em uma cidade, número de funcionários em uma empresa, número de ligações em uma central telefônica, número de pessoas atendidas numa central de saúde, número de panes em sistemas eletrônicos, etc.

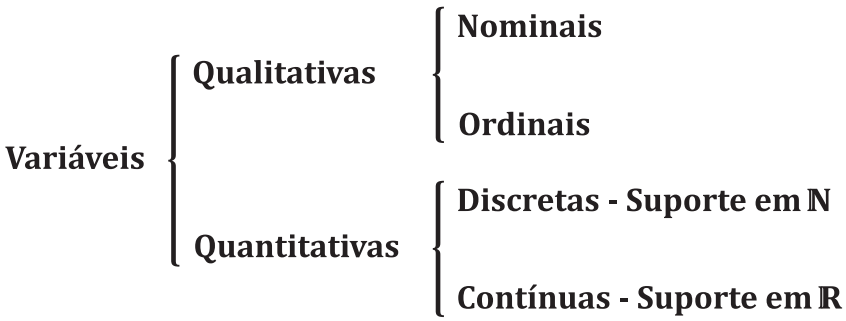


Figura 9: Tipos de variáveis

Continua: As variáveis contínuas são variáveis quantitativas, expressam quantidades, porém com suporte nos números Reais. Pode ser representada por qualquer número dentro de um intervalo numérico. Bons exemplos da variável contínua são tempo de espera na fila de um hospital, peso das crianças de uma escola infantil, altura de um grupo de pessoas, peso de determinada embalagem, resistência a tração de determinada barra metálica, percentual de pureza de determinada liga metálica, dimensões de uma barra de aço produzida por uma máquina, etc. Todas estas variáveis serão descritas por quantidades cujo suporte será o suporte dos Reais.

Agora com esses novos conceitos aprendidos referente a classificação das variáveis, vamos retomar o estudo de população e amostragem. Como já foi falado, amostragem é a seleção de elementos de uma população, com a condição de que essa amostra represente de forma real a população em estudo.

A seguir vamos discorrer sobre as vantagens da utilização do estudo da população por amostras representativas (Aquela que mantêm as características da população de onde a amostra foi retirada) em relação ao censo (avaliação de toda a população).

Exercícios População e Amostra

1) **Defina População.**
2) **Defina Amostra.**
3) **Defina amostragem.**
4) **Qual a diferença entre Parâmetro, Estimador e Estimativa.**
5) **Qual a diferença entre variáveis qualitativas e quantitativas?**
6) **Qual a diferença entre variáveis quantitativas discretas e variáveis quantitativas contínuas.**
7) **Cite 6 exemplos para cada tipo de variável apresentada nesta secção.**
8) **Utilizando o conjunto de dados Milsa apresentado no Anexo 1, responda:**
   a) Quantas variáveis foram avaliadas?
   b) Qual o tamanho da Amostra?
   c) Classifique as variáveis conforme o tipo qualitativa nominal, qualitativa ordinal, quantitativa discreta, quantitativa contínua.
9) **Quantos valores é possível encontrar para um parâmetro?**
10) **Quantas estimativas são possíveis de se encontrar para população?**

CAPÍTULO 03

Amostragem

Segundo Bolfarine e Bussab (2005) discorrem profundamente a respeito das técnicas de amostragem e dos princípios norteadores desta ciência. Dizem que tanto melhor será a amostragem quanto maior for o nosso conhecimento a respeito da população. Tal afirmação de que é importante conhecer a população para se fazer uma boa amostragem parece um paradoxo já que a amostragem só importa quando não conhecemos uma determinada população e a queremos conhecer de forma a otimizar tempo e dinheiro.

Ao utilizar a amostragem o número de indivíduos analisados é **reduzido**, desta forma há uma **redução** significativa nos **custos**. Ao se tratar de grandes populações o estudo de apenas uma fração representativa da população é suficiente para obter os resultados necessários de forma eficaz.

Além do custo podemos evidenciar também a vantagem quanto ao tempo de aplicação da pesquisa, com um número mais reduzido de indivíduos o tempo de aplicação será mais curto e o gerenciamento da pesquisa também é reduzido. Visto que normalmente a necessidade desses resultados seja sempre o mais rápido possível.

A qualidade nos resultados tende a ser melhor, visto que com um número reduzido de entrevistados, o número de entrevistadores também passa a ser mais reduzido, desta forma é possível um melhor preparo desses entrevistadores, possibilitando a aplicação de treinamento e alinhamento entre os entrevistadores.

Por fim é importante evidenciar que o número mais reduzido também possibilita a supervisão mais eficiente e efetiva dos dados, proporcionando, portanto, em alguns casos resultados mais exatos a partir de uma amostra do que por um censo.

Após a essa análise podemos dizer que as amostras devem apresentar uma característica muito importante, a representatividade. Portanto, ao trabalhar com amostragem, a escolha da amostra deve conter as características do meio que a mesma foi retirada.

Bom, mas como vamos saber se amostra é realmente representativa?

Após definir a realização de uma pesquisa deve-se selecionar uma amostra advinda da população, para que essa amostragem seja efetiva podemos utilizar o plano de amostragem que nada mais é do que o plano que determina o número de amostras a serem retiradas de uma população, também é muito importante definir as unidades de amostragem. Além de definir também o tipo de amostragem a ser trabalhada.

As unidades amostrais podem ser os próprios elementos da população, quando é possível o acesso aos mesmos ou outra forma que possibilite a chegada até eles. Para entender melhor o que é unidade de amostragem imaginem que deseja-se obter o perfil socioeconômico utilizando como população os domicílios de uma cidade. A unidade amostral será cada um dos domicílios, que corresponderá aos elementos da população.

Para a amostragem, temos dois tipos principais: as amostragens **probabilísticas** e as **não probabilísticas**. Vejamos:

- **Amostragem probabilística**: São aplicadas em casos onde os elementos da população podem ser mensurável e representado por uma probabilidade ou chance diferente de zero. Podemos exemplificar utilizando uma creche, imagine que temos 50 crianças de 0 a 3 anos,

e desejamos realizar uma amostragem quanto ao tempo de permanência dos mesmos no espaço, sendo assim, são sorteados 10 dentre as 50 crianças. Isso só é possível, pois a população é finita e totalmente acessível.

- **Amostragem não probabilística**: São denominadas amostragem não probabilística quando não é possível determinar a probabilidade ou a chance de um elemento/individuo da população faça parte da amostra. Um bom exemplo é a amostragem realizada em munícios ou estados, onde temos indivíduos a qual temos acesso e outros que não temos, o que não temos acesso não entram para participar da amostra. Ou seja, os participantes são selecionados e não sorteados.

Podemos concluir, portanto que a utilização da amostragem probabilística é mais eficaz para garantir a representatividade da amostra, pois fatores aleatórios (o acaso) seriam os únicos responsáveis por eventuais discrepâncias entre a população e a amostra. Os dados discrepantes também são levados em consideração para inferências estatísticas e cálculos de possíveis margens de erro de previsão.

Para entendermos ainda mais sobre amostragem vamos detalhar os tipos de amostragem probabilísticos a seguir.

1. Amostragens probabilísticas

Como já vimos, a amostragem probabilística é caracterizada por ser possível definir o erro do processo de estimação assim como o poder dos testes. Conforme já dissemos o processo de amostragem traz consigo aleatoriedade, e as estimativas dos parâmetros vêm com incerteza. Por isto é sempre preferível a escolha de um método probabilístico de amostragem São 4 tipos básicos associados à amostragem probabilística. O esquema abaixo ilustra os principais métodos:

Amostragem Probabilística
- **Aleatória Simples**
- **Sistemática**
- **Aleatória Estratificada**
- **Aleatória Por Conglomerado**

1.1. Amostragem Aleatória (casual) Simples

A amostragem aleatória simples (AAS) é aplicada para quando a população apresenta homogeneidade em relação à variável em estudo e apresenta uma população finita. Para a seleção dos indivíduos da amostragem geralmente, atribuímos um número a cada indivíduo da população, e em seguida realizamos um sorteio aleatório com reposição. Para o sorteio aleatório todos os indivíduos têm a mesma probabilidade de serem sorteados e pertencer à amostra, equivale a dizer que uma condição para a amostragem aleatória simples é que o **espaço amostral** seja **equiprovável**. Vale ressaltar que se a população for muito pequena e em casos que não há reposição, a condição de independência não é satisfeita.

Para exemplificar imagine que você queira amostrar um número de pessoas que estão processo seletivo de uma empresa com N = 1000 candidatos. Como a população é finita, enumeramos cada um dos N candidatos e sorteia-se n = 100 deles. É o procedimento mais simples dentre todos os procedimentos de amostragem probabilística.

1.2. Amostragem Sistemática

Assim como na AAS a amostragem sistemática a variável de interesse é homogênea, porém os elementos da população devem ser ordenados de acordo com a variável de interesse e divididos em períodos (ciclos) de igual dimensão. Esta atitude é importante para garantir que os resultados não estejam subestimados em relação à população em estudo.

É importante evidenciar ainda que a variável não deve apresentar ciclos de variação coincidentes com os ciclos de retirada, pois isso faz com que a amostragem não seja representativa.

Veja a seguir o procedimento para a amostragem sistemática:

1. Obtém-se o tamanho da população (**N**);
2. Calcula-se o tamanho da amostra (**n**) – veremos isso mais adiante; e
3. Encontra-se o intervalo de retirada **k = N/n**:
   - Se **k** for fracionário, deve-se aumentar **n** até tornar o resultado inteiro;
   - Se **N** for um número primo, excluem-se por sorteio alguns elementos da população para tornar **k** inteiro;
4. Sorteia-se o ponto de partida (um dos **k** números do primeiro intervalo), usando uma tabela de números aleatórios, ou qualquer outro dispositivo (isso precisa ser feito para garantir que todos os elementos da população tenham chance de pertencer à amostra); e
5. A cada **k** elementos da população, retira-se um para fazer parte da amostra, até completar o valor de **n**.

Para a melhor compreensão do conceito acima vamos desenvolver um exemplo, veja:

Imagine que uma empresa de produtos profissionais para cabelo deseja saber a opinião dos seus clientes comerciais sobre o seu produto na cidade de Ouro Preto. Suponhamos que são ao todo **6.719** clientes comerciais, e para a amostragem são necessários pelo menos **350** elementos. Primeiro organiza-se em uma fila em ordem alfabética todos os clientes e em seguida definimos os intervalos.

$$K = \frac{N}{n} = \frac{6.719}{350} = 19{,}1971$$

Como o valor de K obtido é um valor fracionado atitudes devem ser feito. Aumentar o tamanho da amostra não sana nosso problema, visto que 6.719 é um número primo. O valor da amostra também não pode ser reduzido, visto que este é o valor mínimo de elementos, portanto deve permanecer igual a 350. O que nos resta fazer é reduzir o tamanho da população e refazermos a operação, portanto teremos:

$$K = \frac{N}{n} = \frac{6.700}{350} = 19{,}14$$

Sendo assim temos que a cada 19 clientes um é retirado para participar da amostra. Para definir os números sorteados devemos definir o ponto de partida: um número de 1 a 100 (do 1º ao 100º cliente). Imagine que o resultado do sorteio foi o número 7, logo a amostra será dada pelos seguintes números: {7, 26, 45, 64, 83, 102, 121, 140, ...}.

1.3. Amostragem Estratificada

Diferentemente da amostragem aleatória simples e da amostragem Sistemática, na amostragem estratificada não há uma homogeneidade em relação às variáveis de interesse. Como por exemplo, em uma pesquisa eleitoral para prefeito de uma cidade, é esperado que as respostas dos

eleitores sejam diferentes de acordo com local que o eleitor mora, classe social, escolaridade, sua profissão e conhecimento quanto aos candidatos. Porém, é comum encontrar homogeneidade dentro de cada um dos grupos (estratos) citados.

Ou seja, a população apresenta heterogeneidade entre os elementos, porém em cada estrato existe uma homogeneidade nas respostas. É importante evidenciar que os estratos devem ser mutuamente exclusivos entre si, portanto, cada elemento deve pertencer apenas a um dos grupos. Para garantir que a amostragem seja representativa devemos garantir que cada um dos grupos sejam neles representativos. Logo, se utilizássemos uma amostragem aleatória simples é possível que um número maior de elementos pertencentes a um determinado estrato seja entrevistado, desta forma a amostragem não será representativa.

Veja a seguir ima imagem que exemplifica o que foi expresso a cima:

A amostragem estratificada pode ser dividida em dois tipos principais quanto a proporção do estrato, veja a seguir:

- **Proporcional:** é quando o número de elementos selecionados em um estrato é proporcional ao tamanho do estrato em relação a população, ou seja, se o estrato corresponde a 20% da população, o número de elementos selecionados devem ser referentes a 20% dos elementos do estrato.
- **Uniforme:** é quando o número de elementos selecionados é igual em cada estrato, ou seja, são selecionados 25% de elementos no estrato A, B, C e D, sendo os conjuntos não proporcionais entre si.

Como visto, para a utilização da amostragem estratificada é necessário um conhecimento muito elevado da população, de forma a possibilitar a divisão clara dos estratos.

1.4. Amostragem por conglomerados

Apesar de a amostragem estratificada apresentar resultados com ótima qualidade, a sua implementação é dificultada devido a falta de informações sobre a população em alguns casos, impossibilitando a estratificação. Quando nos deparamos com esse problema, podemos contorná-lo utilizando o esquema de amostragem por conglomerados.

Os conglomerados são definidos em função da experiência do gestor ou pesquisador. É comum definir os conglomerados de forma arbitraria, a partir de fatores geográficos, como por exemplo, bairros, quarteirões ou até mesmo pequenos municípios. Conglomerados são subgrupos mutuamente exclusivos da população em estudo, que individualmente reproduz a população, ou seja, cada elemento do subgrupo é muito homogêneo entre os demais. Este tipo de amostragem é comumente utilizado pra populações muito grandes, como em pesquisas de nível estadual ou nacional.

Porém, como é feita a divisão dos conglomerados e a seleção da amostra? Vejamos a seguir os procedimentos para a obtenção desses subgrupos.

1. Divide-se a população em conglomerados;
2. Sorteiam-se os conglomerados (usando tabela de números aleatórios ou qualquer outro método não viciado); e
3. Pesquisam-se todos os elementos dos conglomerados sorteados, ou sorteiam-se elementos deles.

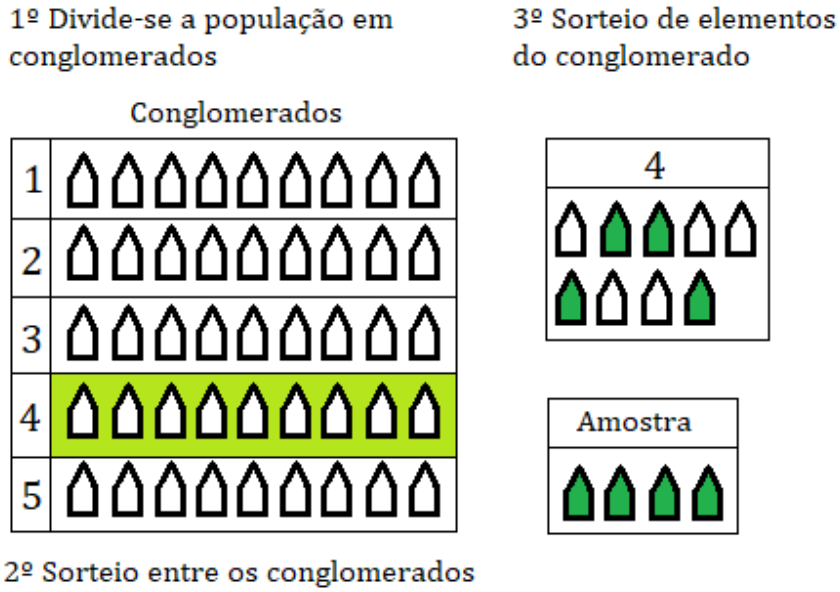


Figura 10: Esquema amostral por conglomerados

Um exemplo da utilização da amostragem por conglomerados é a Pesquisa Nacional por Amostra de Domicílios (PNAD) do Instituto Brasileiro de Geografia e Estatística (IBGE), o qual coleta informações demográficas e socioeconômicas sobre a população brasileira. A estrutura utilizada pela PNAD é composta por três estágios:

1. Primeiro estagio: amostras de municípios (conglomerados) para cada uma das regiões geográficas do Brasil;
2. Segundo estágio: setores censitários sorteados em cada município (conglomerado sorteado); e
3. Terceiro estágio: domicílios sorteados em cada setor censitário.

Bom, você deve estar se perguntando, e quando não for possível garantir a probabilidade de todo elemento da população pertencer à amostra? Então esse é o momento de partirmos para a amostragem não probabilística.

2. Amostragens NÃO probabilísticas

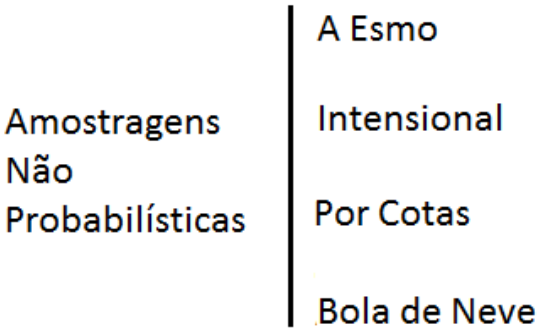
Quando trabalhamos com uma amostra probabilística é necessário ter de posse a listagem de elementos da população, isso exige acesso a todos os elementos da população, e em alguns casos essas informações são muito difíceis de obter ou até mesmo impossível. Desta forma se torna inviável a utilização da amostragem probabilística, sendo assim, podemos recorrer a amostragem não probabilística.

Ao utilizar a amostragem não probabilística não se sabe a priori, a probabilidade que um elemento da população tem em pertencer à amostra. Neste caso não é possível calcular o erro decorrente da generalização dos resultados da amostra (Erro amostral), desta forma não é possível conhecer os parâmetros da população com o rigor metodológico da estatística.

Esse tipo de amostragem tem algumas utilizações, sobretudo quando não é possível lançar mão de uma metodologia amostral probabilística. É utilizado em:

1. Estudos preliminares e em projetos de pesquisa;
2. Quando a metodologia de pesquisa é qualitativa; ou
3. Quando a população não está bem definida ou não pode ser enumerada.

Assim como na amostragem probabilística, a amostragem não probabilística também apresenta alguns tipos principais de amostragem, vejamos o esquema a seguir.



2.1. Amostragem a esmo

Para a amostragem a esmo os interessados nas informações procura ser o mais aleatório possível, porém sem realizar um sorteio formal, porque em muitos casos esta atividade é muito trabalhosa. Um critério muito importante para a amostragem a esmo é que a população dever ser homogenia, quanto mais homogenia for a população, mais podemos a equivalência com a amostragem aleatória simples.

Para exemplificar, imagine que temos uma caixa com 5.000 balas homogêneas entre si, no qual queremos retirar 100 balas para compor a amostra, se fossemos utilizar a AAS, o processo seria muito trabalhoso. Então retiramos os elementos a esmo, sem nenhuma norma ou critério prévio.

Esse tipo de amostragem também pode ser utilizado quando a população for formada por material contínuo (gases, líquidos, minérios), bastando homogeneizar o material e retirar a amostra.

2.2. Amostragem intencional (por julgamento)

Para a amostragem intencional o pesquisador seleciona os elementos que farão parte da amostra. A seleção é baseada em fatores importantes para a sua análise, o pesquisador escolhe quais indivíduos ele julga ser representativo da população. Esta amostragem é muito comum em estudos qualitativos, em estudos de marketing, em estudos de mercado, casos em que o pesquisador está focado em um grupo específico e bem definido.

Podemos exemplificar a utilização da amostragem intencional em uma situação que se deseja obter a aceitação de uma nova marca de vinho a ser inserido no mercado na cidade de Belo hori-

zonte. Apenas pessoas que façam uso desta bebida e tenham condições financeiras (classe social de maior poder aquisitivo) de adquirir o produto poderão fazer parte da amostra em estudo.

É importante ressaltar que a amostragem se não for feito de forma muito criteriosa pode apresentar uma amostra viciada, da qual não representa a população em estudo, pois baseia-se apenas nas preferências do pesquisador, que pode se enganar voluntária ou involuntariamente.

2.3. Amostragem por cotas

A Amostragem por cotas se assemelha a amostra por estratificação proporcional, onde a população é dividido em uma grande quantidade de grupos, porém nessa amostragem não há sorteio para a seleção dos elementos da amostra e sim, uma seleção de forma proporcional (cota) à população do grupo.

Um exemplo bem conhecido desse tipo de amostra são as pesquisas eleitorais, onde a população de eleitores é dividida de acordo com alguns critérios como sexo, nível de instrução, faixas de renda entre outros aspectos, e obtém-se cotas proporcionais ao tamanho dos grupos, dados que podem ser obtidos através do IBGE. Portanto, se em uma pesquisa o grupo em estudo é composto por 42% de homens e 58% de mulheres, quando for realizada a pesquisa com uma amostra de 100 entrevistados, 42 deles devem ser homens enquanto 58 deles devem ser representados por mulheres, esse critério deve ser seguido pelo entrevistador de forma a garantir a proporcionalidade das cotas.

É importante evidenciar que os resultados da amostragem por cotas não têm validade estatística, visto que não contemplam o princípio de aleatoriedade na seleção da amostra.

Lembrando que as metodologias de amostragem mostradas acima, são as mais utilizadas no estudo e aplicação da estatística, porém existem muitas outras utilizadas em casos ainda mais específicos.

Até o momento muito foi falado sobre a amostra de uma população, portanto no próximo tópico descreveremos como deve ser feito o calculo para definir o tamanho de uma amostra probabilística.

2.4. Amostragem Bola de Neve

Há casos em que a amostragem serve para monitoramento de determinada situação. Imaginem que você seja responsável por monitorar uma área sob risco potencial de enchentes. Seu comportamento amostral será diferente a depender do período climático. No período das secas suas amostragens dos níveis dos rios pode ser menos intensas, mas começadas as chuvas, você precisa de um estado de atenção maior, estará em estado de alerta, e sendo assim, ao começarem as chuvas você aumentará sua intensidade amostral. Se durante o período das secas você amostrava uma vez por mês, iniciado o período chuvoso, suas amostras poderão ser retiradas semanalmente ou mesmo diariamente. Pois é muito importante a informação mais recente para o processo de tomada de decisão.

Suponha que você monitore uma lavoura de Soja verificando a infestação de determinado inseto praga da lavoura. Ao andar pelos talhões você raramente encontra um inseto, então decide por uma amostragem esparsa, que lhe dá uma ideia geral da infestação do inseto. Todavia se a presença do inseto se confirmar reiteradas vezes, você estará em estado de alerta, e então precisa de uma informação mais acurada e menos vaga. Então você aumenta sua intensidade de amostragem de acordo com o cenário observado. Se inicialmente você fazia uma amostra a cada 10 mil plantas, agora pode ser levado a tomar uma amostra a cada 2 mil plantas.

3. Cálculo do tamanho de uma amostra probabilística

Um dos aspectos mais controversos da técnica de amostragem é a determinação do tamanho da amostra, pois envolve conceitos como a inferência estatística, probabilidade e a própria teoria da amostragem. Sendo assim, apresentaremos uma visão simplificada para obter-se o**tamanho mínimo** de uma **amostra aleatória simples** considerando os 4 aspectos listados abaixo:

1. O interesse na proporção de ocorrência de um dos valores de uma variável qualitativa na população – a variável diz respeito a uma **proporção**;
2. A **confiabilidade** dos resultados da amostra deve ser aproximadamente igual a 95% (ou seja, há 95% de probabilidade de que a proporção populacional do valor da variável qualitativa esteja no intervalo definido pelos resultados da amostra);
3. Estamos fazendo uma **estimativa exagerada** do tamanho de amostra; e
4. Não vamos nos preocupar com aspectos financeiros relacionados ao tamanho da amostra (embora, obviamente, seja uma consideração importante).

O **erro amostra** tolerável (E_0), já citado anteriormente é o primeiro passo para o calculo do tamanho da amostra. O erro amostral é definido como sendo o valor máximo admitido de erro da estimativa da característica estudada da população.

Este conceito é muito visto por nós em pesquisas eleitorais, é normal ouvir o seguinte enunciado "O candidato X está com 29% das intenções de voto, com marguem de erro de 2% para mais ou para menos". Isso quer dizer que o erro amostral tolerado pela pesquisa é de 2%, então a porcentagem de eleitores que declaram o voto ao candidato X é de 29% ± 2%.

É intuito pensar que quando menos for o valor do erro amostral tolerado, maior deverá ser o tamanho da amostra. É possível notar isso a partir da equação para obter a primeira estimativa do tamanho da amostra:

$$n_0 = \frac{1}{E_0^2}$$

Sendo E0 o erro amostral tolerado escolhido pelo pesquisador, e n0 é a primeira estimativa de tamanho de amostra. Sendo conhecido o tamanho da população N, é possível efetuar a primeira correção da estimativa, dado pela seguinte equação:

$$n = \frac{Nxn_0}{N + n_0}$$

Utilizemos como exemplo as eleições a prefeito em uma cidade de 300.000 habitantes, considerando um alto grau de confiabilidade e um erro amostral tolerável de 2%. Vamos aos cálculos:

1. Começamos determinando a primeira estimativa n_0:

$$n_0 = \frac{1}{E_0^2} = \frac{1}{0{,}02^2} = 2.500$$

2. Em seguida aplica-se a correção para relacionar a população em estudo:

$$n = \frac{Nxn_0}{N + n_0} = \frac{300.000x2.500}{300.000 + 2.500} = 2.479{,}338$$

Portanto define-se que o tamanho mínimo da amostra é dado por 2480 eleitores entrevistados para que se garanta um erro amostral máximo de 2%. Observe que a amostra detém 0,826% da população em estudo. Podemos observar também que poderíamos ter utilizado a primeira esti-

mativa de primeira, pois a correção não efetuou uma mudança muito significativa no tamanho da amostra e garantiria um pequeno ganho na qualidade da amostra.

Para que você tenha uma visão macro da relação da amostra em relação ao tamanho da população para um erro amostral admitido de 2%, temos o gráfico a seguir:

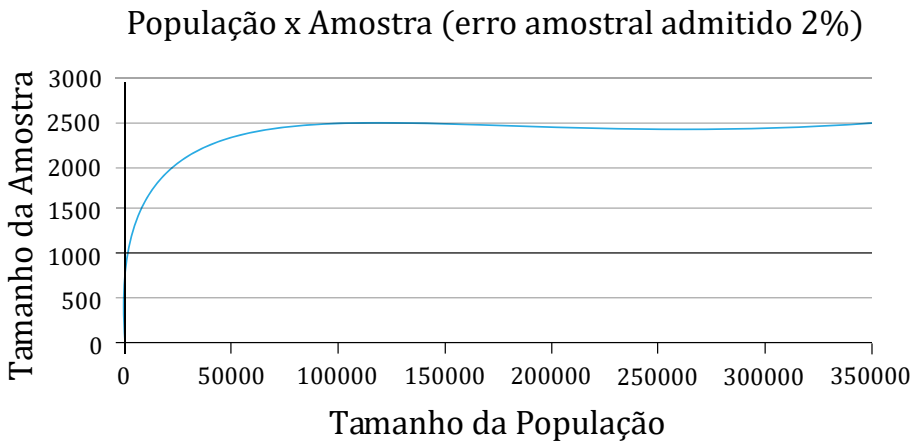


Figura 11: Simulação do Erro Amostral em Função dos Tamanhos das: Amostra e População

Ao analisar o gráfico é perceptível que a quanto maior a população em estudo (principalmente a partir de 300.000 elementos) o tamanho da amostra tende a se estabilizar se aproximando cada vez mais do valor de 2500 elementos, coincidentemente o valor da primeira estimativa n_0 para erro amostral de 2%. Isso evidência que não é necessário utilizar 50% da população para obter uma amostra representativa.

É importante alerta que para a amostragem estratificada o cálculo do tamanho da amostra deve ser feito para cada um dos estratos de forma independente. Se isso não for feito o erro final da amostra será superior ao previsto pelo pesquisador.

Exercícios Amostragem

1) **Quais são os objetivos de se estudar amostragem?**
2) **O que é amostra piloto?**
3) **Em que situações são utilizadas a amostra piloto?**
4) **Quais as vantagens de se fazer uma amostragem?**
5) **Por que existe o aspecto da Aleatoriedade na amostra e não o há na população?**
6) **Quais são as dificuldades inerentes ao Senso?**
7) **Porque a teoria diz que uma amostragem apresenta é mais confiável que um senso?**
8) **Qual a diferença entre amostragem probabilística e amostragem não probabilística?**
9) **Quais são as condições que possibilitam o uso da Amostragem Aleatória Simples.**
10) **Cite uma vantagem de se usar uma Amostragem Aleatória por Conglomerado. Justifique sua resposta.**

CAPÍTULO 04

Tabelas e gráficos

1. Tabelas de Frequências

Quando realizamos uma pesquisa, coletamos dados que costumamos chamar de dados brutos, são geralmente grandes planilhas em que os dados se encontram como foram coletados, sem nenhuma sistematização, sem nenhum tratamento prévio. Um bom exemplo para a compreensão desse conceito é o número de empresas cadastradas em 20 cidades da região da mata no ano de 2019 (os dados são fictícios e elaborados pelo autor do material). Os dados serão chamados ao longo do texto para se demonstrar conceitos estatísticos e serão indicados por **Exemplo 1**, estão apresentados na tabela a seguir, na forma em que foram coletados sendo assim denominados como dados brutos.

Esse tipo de dado normalmente não traz ao leitor muitas informações, sendo necessário organizá-los de forma a melhorar a obtenção de informações de interesse. Então teremos:

Tabela 1: Número de Empresas por cidades – Exemplo 1.

Cidade	Número de empresas	Cidade	Número de empresas
1	84	11	106
2	92	12	102
3	105	13	100
4	98	14	102
5	87	15	81
6	97	16	81
7	88	17	98
8	100	18	100
9	105	19	90
10	89	20	99

Uma primeira análise que fazemos diante de um problema desta natureza é saber a classificação das cidades quanto ao número de empresas, isto implica em criar um ordenamento, colocando as cidades numa ordem quanto ao número de empresas. Pode-se ver da tabela acima que há cidades com mais empresas e cidades com menos empresa, mas abstrair esta informação diretamente da tabela não é uma tarefa fácil.

Todavia se ordenamos as cidades quanto ao número de empresas esta informação é facilmente obtida.

Tabela 2: Número de Empresas por cidade – ordenados em ordem crescente.

Cidade	Número de empresas	Cidade	Número de empresas
15	81	17	98
16	81	20	99
1	84	13	100
4	87	18	100
5	88	8	100
7	88	14	102
10	89	12	102
19	90	3	105
2	92	9	105
6	97	11	106

Ao observar a Tabela 2 vemos que a simples organização dos dados em ordem aumenta muito o nível de informação dos dados, dando mais visibilidade. É possível observar que o menor valor do conjunto de dados é o 81 e o maior o 106, sendo assim, um dado que já pode ser observado de primeira é a amplitude total, a diferença entre o maior e o menor, entre o máximo e mínimo deste conjunto de dados, que neste caso é de 25 unidades.

Podemo observar que as cidades 15 e 16 são as que possuem menor número de empresas, assim como as cidades 3, 9 e 11 são as que possuem maior número de empresas. Também é possível ver as cidades que possuem um número equivalente de empresas com é o caso das cidades, 15 e 16, ou com as cidades 5 e 7, ou com as cidades 13, 18 e 8, e assim por diante. A sistematização dos dados melhora a percepção da realidade captada.

Podemos ainda utilizar a tabela de simples frequência, onde os dados de mesmo valor são agrupados, e uma segunda coluna com a frequência absoluta dos dados é criada. A frequência absoluta, nada mais é do que a quantidade de vezes que aquele número aparece no conjunto, para exemplificar esse conceito vamos passar os dados anteriores para uma tabela de frequências absoluta que nada mais é que junção dos dados repetidos, não deixando de indicar a frequência com que ocorrem.

Tabela 3: Exemplo 1 Sistematizado em classes

Frequência Observada (Frequencia Absoluta)	Número de Empresas
2	81
1	84
1	87
2	88
1	90
1	92
1	97
2	98
1	99
3	100
2	102
2	105
1	106

Nesta nova sistematização é possível ter uma ideia da distribuição de frequências, ou seja, como que o número cidades (a frequência) varia em relação ao número de empresas. É importante ressaltar, que nesta nova conformação, podemos ver sob novo olhar a mesma informação, todavia aqui perdemos a indexação quanto à cidade especificamente. Vemos por exemplo que existem 3 cidades com 100 empresas, mas não conseguimos mais identificar, neste quadro, quais são estas três empresas. Isto mostra um aspecto da sistematização dos dados: resumir, simplificar, significa perder algumas informações, mas o que conta é que se ganhe objetividade e clareza naquilo que se quer demonstrar.

Vejamos mais um exemplo o qual chamaremos **Exemplo 2**: Suponha que você trabalhe na secretaria de assistência social do município e precisa planejar as ações para o próximo exercício financeiro. Para subsidiar suas estimativas você utilizou um banco de dados que se encontrava disponível na própria secretaria. Os dados diziam respeito ao número de encaminhamentos feitos pela secretaria em períodos anteriores. Os dados diziam respeito a atendimentos mensais mas o estagiário esqueceu de colocar a referência dos meses nas fichas e com o manuseio delas acabou-

-se perdendo a ordem original delas. O levantamento dizia respeito a um período de 60 meses, mas um dia o mesmo estagiário deixou a janela aberta, foi o suficiente para as crianças da vizinhança usarem parte das fichas para fazerem aviõezinhos.

Você chegou do treinamento que fazia, foi logo remanejado para a Secretaria de Assistência Social pois o caos havia se instalado por lá. Estava uma verdadeira bagunça e demandava de gente competente e qualificada para organizar a repartição. De toda a desorganização você conseguiu sintetizar numa tabela alguns dados. O período levantado contava com 22 meses, e você convencionou chamar a variável "Número de encaminhamentos" de X. estes dados convencionamos chamar

Exemplo 2.

$$X=\{4, 4, 2, 5, 1, 4, 1, 4, 2, 2, 0, 6, 5, 4, 2, 2, 3, 3, 4, 9, 4, 2\}$$

Buscando visualizar melhor as informações construiu uma tabela de frequências cujas classes você mesmo as definiu:

Tabela 4: Exemplo 2 – Encaminhamentos na Secretaria de Assistência social – sistematização por classes e frequência.

Classes – Números de Encaminhamentos	[0,2]	]2,4]	]4,6]	]6,8]	]8,10]	]10,12]
Frequência Absoluta	9	9	3	0	1	0
Frequência Relativa	0,41	0,41	0,13	0	0,05	0

Com a Tabela 4 já é possível ter uma ideia mais clara do comportamento desta variável – Número de Encaminhamentos. Observe que a conclusão a que chegamos emerge de um período amostral, e retrata a realidade daquele período, suas conclusões podem dar uma ideia da sistemática dos encaminhamentos, mas como já sabemos, contará com o ingrediente incerteza.

Da Tabela 4 podemos ver que é muito comum meses com até 4 encaminhamentos. Também não é raro acontecer de ter meses com o número de encaminhamentos entre 5 e 6. Apenas excepcionalmente tiveram mais que 7 encaminhamentos.

Se fosse eu o gestor da secretaria eu estaria muito preocupado com uma estrutura de atendimento para até 6 pessoas. Como podemos observar das Frequências Relativas da Tabela, uma estrutura como esta atenderia a demanda de aproximadamente 95% das vezes, precisaria também de alguma medida emergencial para situações específicas, mas que como não são muito comuns não necessariamente necessitam uma estrutura disponível.

Este exemplo nos dá uma ideia da praticidade da informação disponível em uma distribuição de frequências em como se dá sua relação com a teoria das probabilidades.

2. Distribuições de Frequências

Quando trabalhamos com um número grande de dados de variáveis quantitativas, a utilização de uma tabela de simples frequência pode não ser muito útil, já que é muito improvável que haja valores coincidentes, já que o suporte desta variável são os números Reais, contando com infinitas possibilidades. As classes são intervalos numéricos nos quais os dados da variável analisada são agrupados.

Ao distribuir os dados observados em classes podemos contar quantos elementos pertencem a cada uma das classes. Obtendo portanto a frequência de classe. A tabulação dos dados de forma

agrupada em classes em conjunto com suas respectivas frequências, é chamado de distribuição de frequência. Cada uma das classes da tabela devem ter um limite inferior e superior, que delimita o intervalo da classe. Vale ressaltar que a amplitude de cada um dos intervalos devem ser iguais entre si.

Porém, como são determinados os intervalos das classes?

Primeiro precisamos conhecer os tipos de intervalo existente. A baixo são listados e explicados alguns deles.

- **Intervalos abertos:** os limites da classe (inferior e superior) não pertencem a mesma.
- **Intervalos fechados:** os limites da classe (superior e inferior) pertencem à classe em questão.
- **Intervalos mistos:** um dos limites pertence à classe e o outro não.

Podemos utilizar qualquer um dos tipos de intervalo, porém o que é mais utilizado para esse tipo tabela e que vamos adotar para a resolução dos problemas é o intervalo misto, que é apresentado abaixo:

$$[23{,}5\ ;\ 25{,}5[$$

(o 23,5 pertence ao intervalo e o 25,5 não pertence ao intervalo)

Os valores de 23,5 e 25,5 foram escolhidos arbitrariamente, somente para demonstrar o formato do intervalo. Para entender melhor, desenvolveremos uma tabela de distribuição de frequência por classes com os dados do número de empresas registradas na região da mata no ano de 2019. Visto no tópico anterior.

Para dar início a construção da tabela é necessário determinar o número de classes (k) em que os dados serão agrupados. Ele é determinado a partir do número de observações (n) do conjunto em estudo, conforme podemos ver a seguir:

$$k = \sqrt{n}, para\, n \leq 100$$

$$k = 5logn, para\, n > 100$$

Tendo que o número de cidades observadas pela pesquisa é $n=20$, o número de classes será definido por $k = \sqrt{n}$.

$$k = \sqrt{n} = \sqrt{20} = 4{,}47$$

Como o valor encontrado para k não é um número inteiro, devemos arredondá-lo para o valor inteiro imediatamente acima, ou seja, $k = 5$. Vale lembrar que as classes também podem ser determinadas de forma arbitraria, sem o uso dessa regra.

Após determinar o número de classes (k) partimos para a determinação da amplitude do intervalo de classe (c). Para determinar a amplitude do intervalo de classes precisamos da amplitude total dos dados (AT) que é a diferença entre o maior valor e o menor valor observado.

$$AT = 106 - 81 = 25$$

Vamos então determinar o valor da amplitude da classe: $c = \frac{A}{k}$

Substituindo os valores referentes ao problema em estudo temos: $c = \frac{25}{5} = 5$

Sendo assim a amplitude das classes é de 5 unidades, ou melhor 5 empresas cadastradas. Neste caso o valor de c foi um valor inteiro, mas em muitos outros esse valor pode ser fracionado e em

muitos casos com muitas casas decimais, isso dificulta a montagem da tabela, sendo aconselha-se que seja feito um arredondamento nas casas decimais de forma a ser coerem com os dados originais (mesmo número de casas decimais), e o arredondamento deve ser feito sempre para um valor maior que o obtido, de forma a garantir que nenhum elemento do conjunto seja deixado de fora.

Com os dados obtidos já é possível a construção dos das classes e seus respectivos intervalos. A primeira classe sempre se inicia com o menor valor do conjunto (limite inferior) e o limite superior é definido a partir da soma do menor valor do conjunto mais a amplitude da classe. Vamos construir os intervalos a seguir:

1ª classe → [81 ; 86[

Para as classes seguintes o valor do limite inferior é igual ao limite superior da classe anterior. Já o limite superior é igual ao limite inferior da classe mais a amplitude da classe (c). Vejamos:

2ª classe → [86 ; 91[

3ª classe → [91 ; 96[

4ª classe → [96 ;101[

5ª classe → [101 ; 106[

Com base nessas determinações já podemos estruturar a tabela de distribuição de frequência por classes, considerando dados do **Exemplo 1**.

Tabela 5: Sistematização das classes – Exemplo 1.

Intervalos	Frequência absoluta (F_a)
[81 ; 86[	?
[86 ; 91[	?
[91 ; 96[	?
[96 ; 101[	?
[101 ; 106[	?

Podemos observar que em cada uma das classes (linhas) a uma interrogação quanto ao valor da frequência absoluta, nosso próximo passo é determinar o valor da frequência absoluta (F_a), e de mais duas frequências que são de nosso interesse, são elas a frequência relativa (F_r) e a frequência acumulada (F_{ac}).

Mas como determinamos essas frequências e o que elas significam?

A frequência absoluta (F_a) corresponde ao número de observações encontradas dentro do intervalo em uma determinada classe. A frequência relativa (F_r) é a proporção do número de observações em uma determinada classe em relação ao total de observações. Essa frequência é normalmente representada em termos de porcentagem, sendo assim o valor resultante da divisão entre a frequência absoluta pelo número de observações deve ser multiplicado por 100.

$$F_{ri} = \frac{F_{ai}}{\sum_{i=1}^{n} F_{ai}}$$

A frequência acumulada (F_{ac}) nada mais é que a soma do valor das frequências até a sua respectiva classe. A frequência acumulada pode ser aplicada tanto para a frequência absoluta onde seu valor máximo ou final deve ser igual ao número de observações do experimento, quando para a frequência relativa onde seu valor máximo é igual a 100%.

Para determinar a frequência absoluta da primeira classe, por exemplo, contaremos os temos que então dentro do intervalo de 81,00 a 85,99, visto que o termo de valor 86 pertence apenas a

segunda classe e não deve ser considerado na primeira. Nesse caso, por exemplo, a primeira classe é composta pelos termos de valor 81, 81 e 84, totalizando uma frequência de valor igual a 3. A partir desse exemplo, vamos então à determinação do valor de cada uma dessas frequências e os dispor na tabela 6:

Tabela 6: Sistematização das classes – Exemplo 1 – com frequências Absoluta e Relativa

Intervalos	Frequência absoluta (F_a)	Frequência relativa (F_r)
[81 ; 86[	3	15
[86 ; 91[	4	20
[91 ; 96[	1	5
[96 ; 101[	7	35
[101 ; 106[	5	25
Total	20	100

A disposição dos dados em forma de distribuição de frequência facilita o calculo manual de várias medidas estatísticas que são importantes para a compreensão dos dados, além de sua apresentação em formato gráfico.

Para determinar o valor da frequência acumulada faremos a soma acumulativa das classes, ou seja, a terceira classe, por exemplo, será a soma da primeira classe + segunda classe + terceira classe. Ela pode ser entendida como se na terceira classe tivéssemos todos os elementos no intervalo de [81 ; 96[. Vejamos a seguir na Tabela 6:

Tabela 7: Sistematização das classes – Exemplo 1 – com frequências Absoluta, Absoluta Acumulada, Relativa e Relativa Acumulada

Intervalos	Frequência absoluta (F_a)	Frequência absoluta acumulada ($F_{a\,ac}$)	Frequência relativa (F_r)	Frequência relativa acumulada ($F_{ra\,c}$)
[81 ; 86[	3	3	15	15
[86 ; 91[	4	7	20	35
[91 ; 96[	1	8	5	40
[96 ; 101[	7	15	35	75
[101 ; 106[	5	20	25	100
Total	20	-	100	-

A partir dessas relações a compreensão e análise dos dados é facilitada, agilizando a obtenção de resultados para a compreensão dos dados.

É comum na estatística o estudo de metodologias para a construção de tabelas de frequência. Acima utilizamos uma metodologia para definir o número de classes e a amplitude das classes. Mas ao longo da minha vivência como professor e pesquisar na área, é cada vez mais incomum estes procedimentos de cálculo. A classificação das classes, assim como a amplitude de classe acaba sendo arbitrada pelo pesquisador, que via de regra, tem interesse em determinadas disposições.

Estas metodologias eram muito comuns no início do desenvolvimento da ciência estatística mas foram perdendo importância com o advento de novas tecnologias para a análise de dados, sobretudo com o advento dos computadores que facilitam sobremaneira a construção de gráficos e tabelas.

Após esse aprendizado é importante se questionar se o uso de tabelas é mesmo a melhor forma de representar dados de uma pesquisa. No próximo tópico veremos que podemos utilizar gráficos para representar esses resultados, vamos continuar?

3. Gráficos

Os gráficos são representações visuais de dados e informações numéricas que são utilizados para facilitar a interpretação dos mesmos. Assim como nas tabelas os dados contidos nos gráficos são referentes a tudo aquilo que pode ser medido ou quantificado. Devido a representação gráfica a compreensão do conteúdo é mais logico, por atribuir um significado mais concreto aos dados.

Para que o gráfico seja realmente representativo ele deve respeitar alguns requisitos básicos, vejamos a seguir:

- **Simplicidade**: deve apresentar o essencial, sem excesso de cores ou desenhos;
- **Clareza**: o gráfico deve apresentar referências para todos os dados expressos nele, para garantir uma leitura correta dos valores; títulos, eixos, nomes de variáveis são alguns dos elementos importantes para a clareza de um gráfico;
- **Veracidade:** apresentar de forma verdadeira os dados, sem apresentar distorções na escala por exemplo.

Existem vários tipos de gráficos, a escolha de qual gráfico utilizar está ligada ao tipo de dado que você tem pretende transmitir. Esses gráficos podem ser obtidos utilizando-se planilhas eletrônicas, como o Excel ou a planilha Calc do OpenOffice. Cada um deles apresenta um conjunto de vantagens e desvantagens, a seguir veremos todos os detalhes referentes a cada tipo gráfico.

Para saber como utilizar a planilha Calc do pacote OpenOffice nas distribuições de frequências e de gráficos, acesse o site: <http://www.ufpa.br/dicas/open/calc-ind.htm>. Acesso em: 07 jan. 2019.

3.1. Gráfico em linhas

Os gráficos em linhas ou também chamados como gráficos de segmento são utilizados para apresentar valores numéricos em determinado espaço de tempo. Este tipo de gráfico é utilizado para dados do tipo de distribuição simples. É possível visualizar a evolução ou diminuição do fenômeno em estudo no decorrer do tempo.

Uma de suas vantagens é encontrada na possibilidade de comparar mais de um fenômeno ao decorrer de um intervalo de tempo, expressos em um único gráfico. A desvantagem está na dificuldade em identificar a continuidade da variação.

Vamos utilizar os dados da tabela a seguir para gerar um gráfico de linhas, a Tabela 8 refere-se ao número de acidentes no trânsito nas 3 primeiras semanas do ano de 2019.

Tabela 8: Dados Acidentes de Trânsito – Exemplo 3

Dias da semana	1ª Semana	2ª Semana	3ª Semana	Total
Domingo	1865	2391	1736	5992
Segunda	1729	1964	1601	5294
Terça	1320	1537	1422	4279
Quarta	1398	1620	1315	4333
Quinta	1454	1532	1376	4362
Sexta	1673	2058	1528	5259
Sábado	1577	1841	1587	5005
Total	11016	12943	10565	34524

Ao observar os dados é possível notar que a segunda semana do ano de 2019 apresentou maior índice de acidentes no trânsito, se comparado às outras duas semanas. É possível evidenciar também que os dias de domingo e sexta feira apresentam maior número de acidentes se comparados aos outros dias da semana.

Graficamente temos o problema 3 conforme a Figura 12:

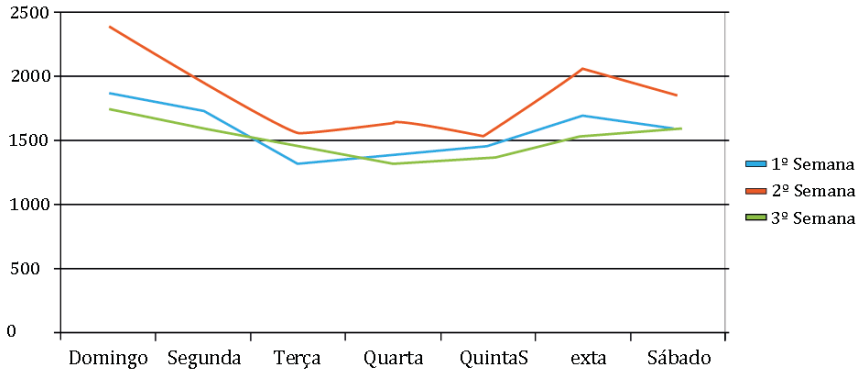


Figura 12: Gráfico de Linhas para o Exemplo 3 - Acidentes de trânsito

Vamos utilizar outro exemplo que chamaremos **Exemplo 4**. Considerando que tenhamos uma situação em que avaliamos indivíduos quanto à renda e à escolaridade. A renda é dada em unidades de Mil Reais e a escolaridade é medida em anos equivalentes à série em que o indivíduo deixou de estudar. Os dados do Exemplo 4 são apresentados na Tabela 9.

Tabela 9: Renda e Escolaridade dos Indivíduos.

Indiv	1	2	3	4	5	6	7	8	9	10	11
Rda	3,7	2,8	6,3	7,2	12,1	1,2	2,1	1,5	3,2	1.1	0.9
Esc	12	11	16	18	20	4	6	7	11	12	8

Esta mesma situação pode ser melhor visualizada em dois gráficos agora apresentados com outros atributos gráficos:

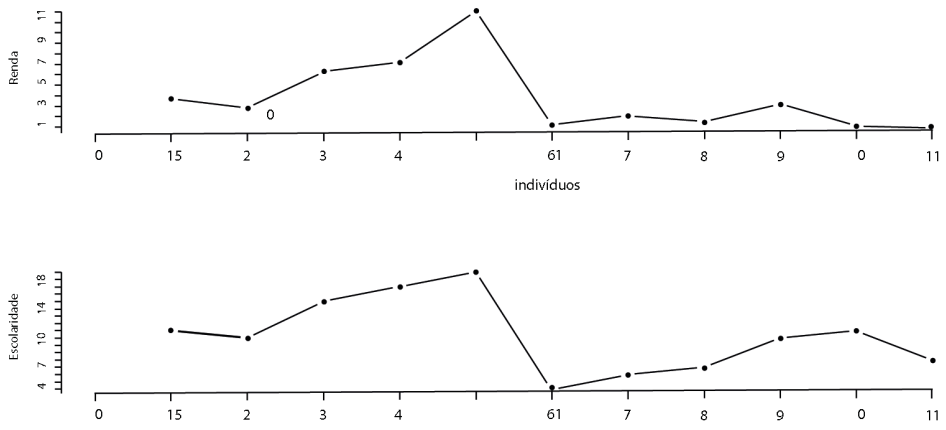


Figura 13: Grafico de Linhas mostrando a renda e a escolaridade de 11 indivíduos

Muito mais que numa tabela de frequência a informação é mais facilmente visualizada com o auxílio de um gráfico. Podemos ver que além da variação das duas variáveis ao longo dos indivíduos podemos inclusive observar a associação entre as duas variáveis. Quando observados conjuntamente, conforme expresso na Figura 13, ver-se-á que a renda recebida está associada com o nível de escolaridade.

3.2. Gráfico de Colunas

Os gráficos de coluna são aplicados em casos que se deseja representar dados variáveis qualitativas. Sua estrutura é muito simples, são retângulos verticais em que no eixo das abscissas ou eixo horizontal (x) coloca-se os valores das variáveis e no eixo das ordenadas ou eixo vertical (y) são dispostas as frequências observadas para cada uma das colunas. Todas as colunas devem ter base de mesma largura, separadas por um espaçamento simples iguais de coluna para coluna.

Este tipo de gráfico é geralmente utilizado para descrever variáveis qualitativas. Utilizando os dados do Exemplo 3 podemos construir o seguinte gráfico com o total de acidentes em cada uma das semanas.

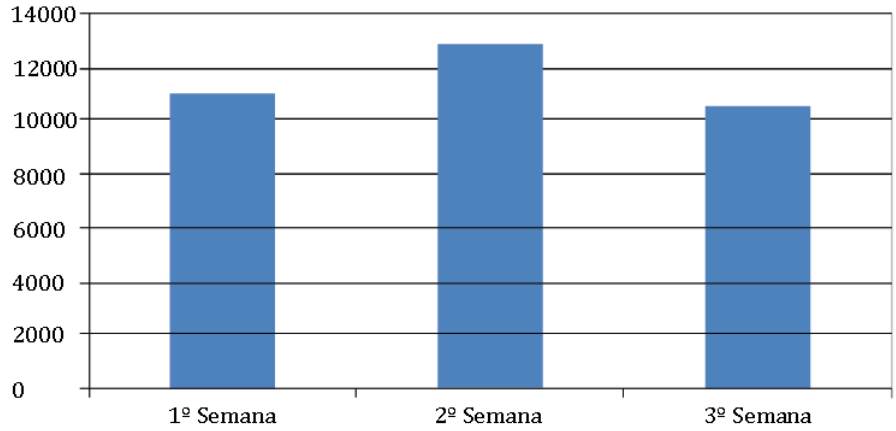
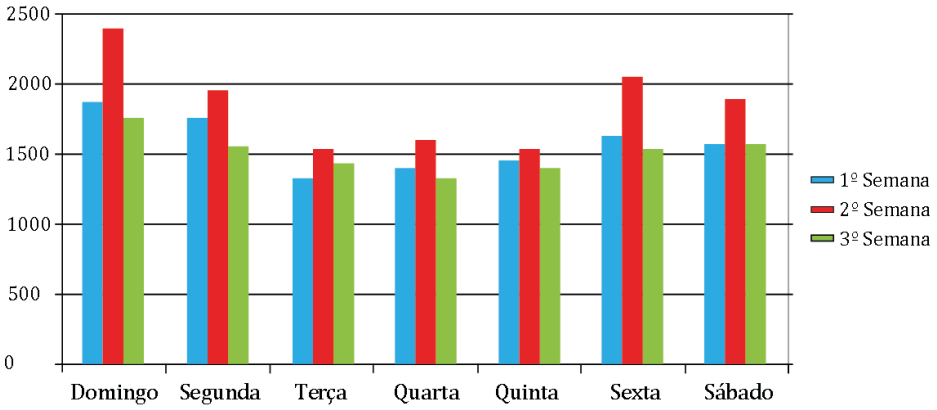


Figura 14: Gráfico de colunas referente ao Número de acidentes no transito nas 3 primeiras semanas do ano de 2019

Neste gráfico (Gráfico 14) fica evidente que a semana que apresenta maior número de acidentes é a segunda semana. Porém não é possível saber quais dias da semana mais colaboram para esse número mais elevado de acidentes.

Existe ainda o gráfico de barras agrupadas, que segue os mesmos princípios do de barras simples, porém ele compara os dados de mesma variável. Sua desvantagem está no fato em que não podemos realizar essa comparação para muitas variáveis, já que a representação deixaria de ver pratica. Veremos isso na Figura 15, gráfico a seguir relacionando os acidentes por dia da semana.



Fgura 15: Gráfico de colunas agrupadas referente ao Número de acidentes no transito nas 3 primeiras semanas do ano de 2019

A partir desse gráfico é mais fácil a comparação entre os dias da semana quanto às três semanas de interesse.

É importante destacar que apesar do nome do gráfico ser 'gráfico de colunas' o gráfico pode utilizar também representações como cilindros fechados, pirâmides ou até mesmo cones em representações em 3 dimensões. Veja os exemplos a seguir.

3.3. Gráfico de Barras

Os gráficos de barras seguem as mesmas definições do gráfico de colunas, seu diferencial é encontrado na representação gráfica, onde as barras são horizontais fixadas no eixo das ordenadas (y) apresentando as suas variáveis, e no eixo horizontal ou eixo das abscissas são representados os valores das frequências observadas. Utilizaremos o mesmo **Exemplo 3** para criar o gráfico disposto na Figura 16, veja:

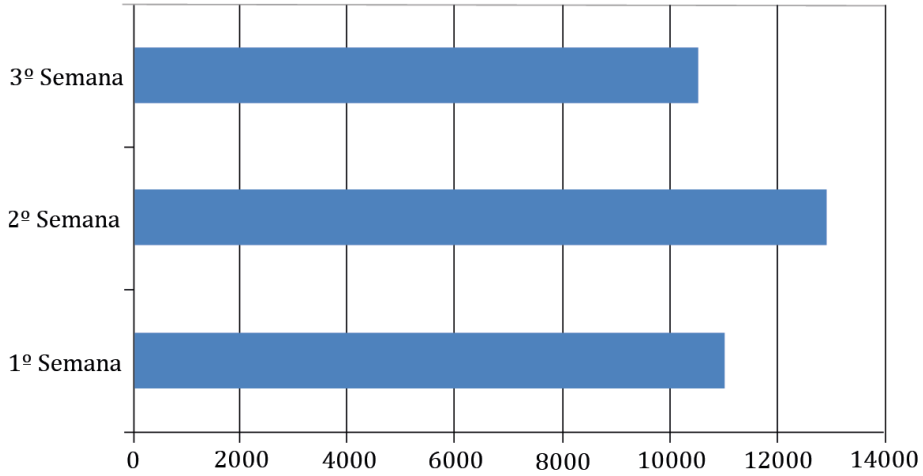


Figura 16: Gráfico de barras referente ao Número de acidentes no transito nas 3 primeiras semana do ano de 2019

As mesmas conclusões encontradas no gráfico de colunas podem ser identificadas no gráfico de barras, pois a única mudança entre os dois tipos de gráfico é o sistema de eixos de referência.

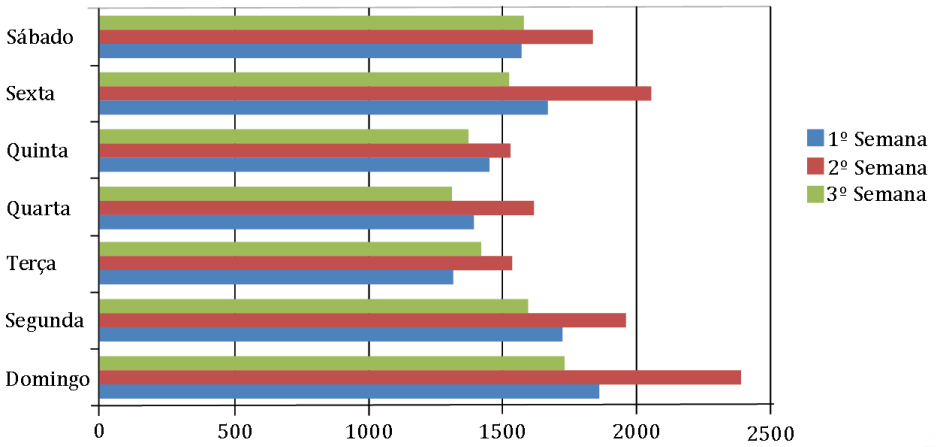


Figura 17: Gráfico de barras agrupadas referente ao Número de acidentes no transito nas 3 primeiras semana do ano de 2019 dividido por dias da semana

3.4. Gráfico de Pizza

O gráfico de Pizza também conhecido como gráfico de setores representa as variáveis em fatias ou setores, acompanhado do valor real dos dados ou em porcentagem de partes de um total, sendo o uso da porcentagem mais comum.

O círculo completo de raio qualquer representa o todo (360° = 100%), ou seja, engloba todo o conjunto de dados. O círculo é dividido em partes proporcionais à porcentagem representante da variável (X° = frequência relativa percentual).

A vantagem encontrada nesse tipo de gráfico é a fácil visualização das proporções entre as variáveis, é comumente utilizado quando as proporções são mais importantes do que o valor real. A desvantagem está em não permitir a comparação entre dois grupos de dados.

Existem vários tipos de representação do gráfico de setores, um deles pode ser visto a baixo, relacionando a quantidade de acidentes quanto aos dias da semana nas 3 primeiras semanas do ano de 2019.

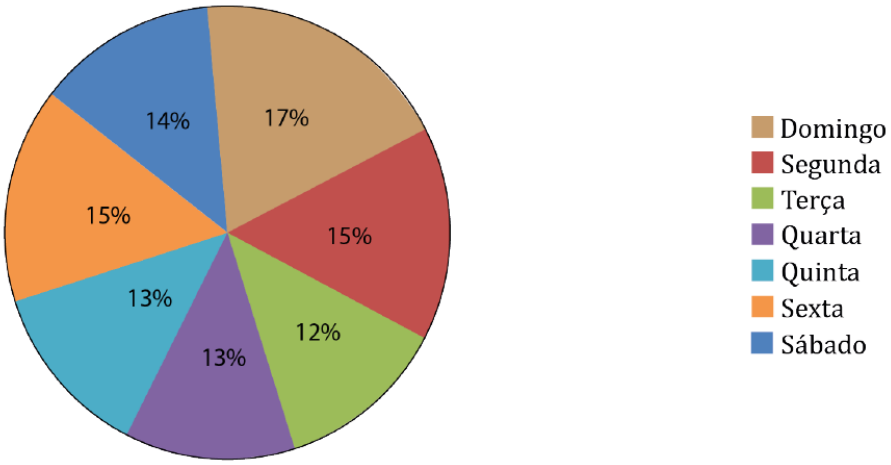


Figura 18: Gráfico de Pizza referente ao Número de acidentes no transito nas 3 primeiras semana do ano de 2019 dividido por dias da semana

Ao observar o Gráfico 18 fica nítido que os dias que apresentam maior índice de acidentes são os dias de domingo com 17%, e segunda e sexta com 15% cada. É fácil identificar também o dia com menor índice, ao observar o gráfico vemos que o dia de terça-feira apresenta apenas 12% do total dos acidentes.

3.5. Histograma

O Histograma é utilizado em dados apresentados em agrupamento por classes, onde cada coluna representa uma das classes. Sua estrutura se assemelha muito a um gráfico de colunas, porém as colunas não apresentam espaçamento entre ela, pois o fim de uma coluna e o início da coluna subsequente, ou seja, coincidem.

Em alguns casos o histograma é o único gráfico capaz de representar algumas situações, devido a sua forma única de apresentar os dados. A partir dele é possível ter uma visão ampla da distribuição dos dados quanto a suas além de ser possível observar os pontos de crescente e decrescente dos dados. Vejamos alguns exemplos: Figura 19 e Figura 20.

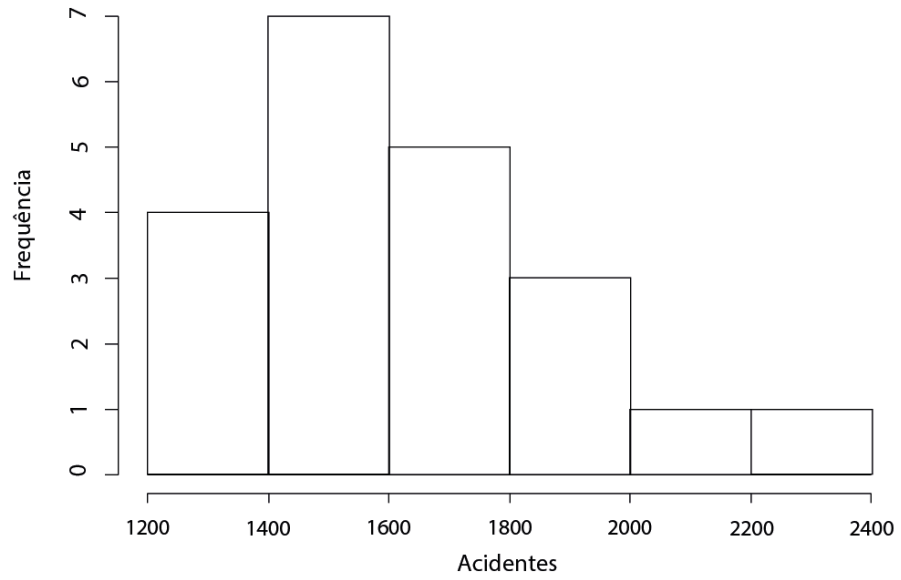


Figura 19: Histograma para o problema Acidentes de Trânsito – Dados do Exemplo 3

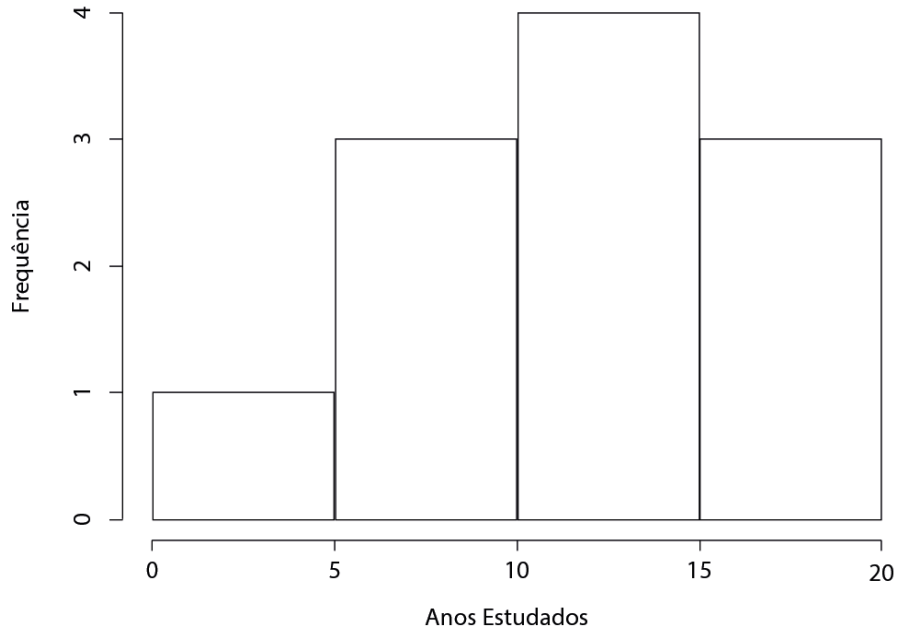


Figura 20: Histograma para a Variável Escolaridade – Dados do Exemplo 2

Embora esta função não esteja implementada na maioria dos programas computacionais que comumente encontramos, do ponto de vista da Estatística e da Teoria da Probabilidade é o tipo de gráfico mais utilizado. Representa como uma aproximação das às variáveis contínuas e possibilita a construção de **polígonos de frequência**, imprescindível para o entendimento de f**unções de distribuições de probabilidade** contínuas, que nada mais é que a base da teoria da probabilidade moderna.

É comumente utilizado em conjunto com o Histograma o Polígono de frequência, ele utiliza o ponto médio das classes e o valor da frequência, criando um gráfico de linhas sobrepondo o histograma, conforme demonstrado na Figura 21.

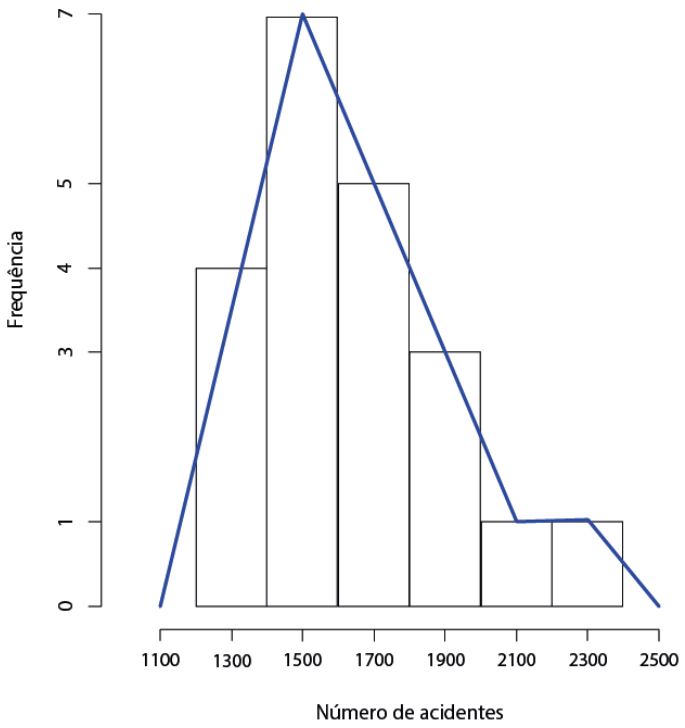


Figura 21: Dados Exemplo 3 - dispostos em um histograma associado com polígono de frequências

Exercícios T

1) Considerando a Variável Estado Civil do conjunto de dados Milsa apresentado no Anexo1, faça as questões a seguir:

a) Construa um gráfico de pizza para a variável Estado Civil.
b) Construa um gráfico de pizza para a variável Região.
c) Construa um gráficos de colunas para a variável Instrução.
d) Faça um histograma para a variável Salário.
e) Faça uma tabela de frequência para a variável Anos.
f) Faça um histograma para a tabela de frequência construída em 1.5.

2) Considerando o Banco de dados do Anexo 2, construa um gráfico de Linhas, considerando as 4 variáveis em questão, referente ao fechamento das 4 maiores bolsas europeias.

3) Com base no Exercício 2, diga, na sua opinião, qual das bolsas teve o melhor desempenho no período avaliado.

4) Com base no Exercício 2, diga, qual ação teve maior valorização no período.

5) Considerando o Banco de Dados do Anexo 3, constua um gráfico de linhas, considerando as 4 variáveis em questão: Valor no fechamento das 4 maiores bolsas europeias.

6) Com base no Exercício 5, diga, qual ação teve maior valorização no período.

7) Com base no Exercício 5, diga, qual ação teve maior valorização no período.

8) Com base no Banco de Dados do Anexo 3, construa 4 histogramas de frequência: um para cada bolsa.

CAPÍTULO 05

Medidas de posição e de dispersão

Como já dissemos, a estatística busca verificar o estado das coisas, verificar como as coisas se encontram. Para descrever as "coisas" que significam situações diversas que o profissional pesquisador encontrará, primeiramente lançamos mão das **Variáveis** conforme já foi visto ao longo do texto. Tivemos o contrato com diversas variáveis, só para relembrar algumas: Dias da Semana, Número de Acidentes, Escolaridade em anos estudados, Renda e Milhares de Reais, Número de encaminhamentos, dentre outras. As variáveis qualitativas podem subsidiar a discussão numa abordagem descritiva qualitativa, no entanto as variáveis quantitativas expressam quantidades que quase sempre são objeto de mensuração para a comparação entre diferentes situações.

Na tentativa de representar um conjunto de dados de interesse através de grandeza numérica, a primeira coisa que nos vem à cabeça é um número que especifique a posição do conjunto em uma escala de valores possíveis da variável. Estas grandezas são chamadas de medida de posição, que tem como objetivo principal definir o centro de uma distribuição de frequência, ou seja, um valor numérico aos quais os demais dados do conjunto rodeiam.

As medidas de posição mais conhecidas e mais utilizadas são a média, a mediana e a moda. Mas também as outras medidas de posição menos utilizadas, alguns exemplos deles são os quartis, os decis e os percentis, porém sua definição é um pouco diferente, não se encontra um número central, mas constituem estatísticas de ordem associadas à posição dos indivíduos numa sequência monótona, representam os valores ocorrentes em determinados trechos do conjunto.

Para nós ajudar nos estudos desse conteúdo utilizaremos dois conjuntos de dados descritos a seguir a partir de tabelas. Vejam a seguir os dados do **Exemplo 5** e do **Exemplo 6**:

Para o estudo da produção de leite na fazenda A levantou-se a quantidade de leite produzido em cada mês do ano de 2019. Esses resultados são dados a partir da soma da produção diária durante cada mês, sendo assim, os dados do Exemplo 5 estão dispostos na Tabela 10 na unidade mil litros por mês (Mil L/mês).

Tabela 10: de distribuição da produção de leite na fazenda A (em Mil L/mês)

Mês	Jan.	Fev.	Mar.	Abr.	Mai.	Jun.	Jul.	Ago.	Set.	Out.	Nov.	Dez.
Produção	24,6	22,8	21,0	20,2	19,7	18,1	16,0	15,5	16,8	19,3	20,9	23,4

A situação do **Exemplo 6** é a seguinte: A escola X de ensino fundamental e médio a partir do 6º ano desejou levantar a quantidade de alunos de acordo com a idade de cada um na data de matrícula do início do ano de 2019. Para a disposição dos dados foi utilizados a Tabela 11 de distribuição de frequência por agrupamento simples com intervalos de 2 anos. Vejamos a seguir:

Tabela 11: Distribuição por agrupamento por classes de idades dos alunos da escola X a partir do 6º ano.

Intervalo de idades dos alunos	Frequência absoluta (nº de alunos)
[10 ; 12[	114
[12 ; 14[	157
[14 ; 16[	202
[16 ; 18[	187
[18 ; 20[	106
[20 ; 22[	41

Já dispostos os dados podemos iniciar os estudos das medidas de tendência central, cada um deles serão descritos separadamente e exemplificados utilizando os dados das tabelas acima. Vamos começar?

1. Média Aritmética

Dentre as medidas de posição citadas anteriormente, a média aritmética é a mais comum e mais compreensível delas. Sua vasta utilização está ligada a facilidade dos seus cálculos, além de dar ao interessado um dado muito importante para analise.

A média pode ser do tipo simples ou ponderado. Iremos começar pelo estudo de **média aritmética simples**, que é dado pela relação da soma dos valores observados dividido pelo número de valores do conjunto. Seja um conjunto de *n* valores de uma variável quantitativa X, a média é dada pela seguinte equação:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Onde temos que:

1. x_i é um dos valores do conjunto;
2. $\sum_{i=1}^{n}$ é a soma de todos os valores do conjunto; e
3. $\boldsymbol{n}$ é o tamanho do conjunto $\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \cdots + x_n$
4. n é o tamanho do conjunto.

Para compreendermos melhor a utilização dessa equação vamos realizar os cálculos referentes aos dados da produção de leite na fazenda A. Temos que o valor de *n* é igual a 12. Vamos lá:

$$\sum_{i=1}^{12} x_i = 24{,}6 + 22{,}8 + 21{,}0 + 20{,}2 + 19{,}7 + 18{,}1 + 16{,}0 + 15{,}5 + 16{,}8 + 19{,}3 + 20{,}9 + 23{,}4$$

$$\sum_{i=1}^{12} x_i = 238{,}3.\ 10^3 L$$

Temos, portanto que foram produzidos durante todo o ano de 2019, $238{,}3.\ 10^3$ L/mês de leite na fazenda A em estudo, para definirmos a média de produção por anos, devemos dividir o somatório pela quantidade de termos n, igual a 12. Logo temos que a média é:

$$\bar{x} \sum_{i=1}^{n} x_i = \frac{238{,}3}{12} = 19{,}9.\ 10^3 \frac{L}{mês}$$

Como podemos interpretar esse resultado?

Podemos interpretar o resultado da média como sendo a quantidade de leite produzida em cada mês, caso a produção fosse igual em todos os 12 meses do ano. A média elimina as oscilações mês a mês e nos dá um valor central de referência em torno do qual os demais valores oscilam.

Obtemos então que são produzidos em média $19{,}9.\ 10^3$ Litros de leite por mês na fazenda de estudo, onde há meses em que esse valor será superior e em outro inferior ou até iguais a média, porém oscilam em torno da média.

2. Média Ponderada

Todavia, há casos em que os dados são distribuídos quanto à sua frequência relativa ou frequência em classes, para esses casos utiliza-se um novo conceito de média, o da**média aritmética ponderada**, na qual a frequência dos dados é diferente de 1 (um), logo a equação é dada a seguir:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_1 . fa_i}{n}$$

Onde:

k é o número de classes de variáveis agrupadas;

x_i é o ponto médio da classe i;

$\sum_{i=1}^{k} x_1 . fa_i$ é a soma da multiplicação entre o ponto médio i e a frequência absoluta i de cada uma das classes;

$$\sum_{i=1}^{k} x_i . fa_i = x_1 . fa_1 + x_2 . fa_2 + x_3 . fa_3 + \cdots + x_k . fa_k$$

fa_i é referente a frequência absoluta da classe i; e

$\sum_{i=1}^{k} fa_i$ é a soma da frequência de cada uma das classes i.

$$\sum_{i=1}^{k} fa_i = fa_1 + fa_2 + fa_3 + \cdots + fa_k$$

n é o número de classes.

Utilizaremos os dados do **Exemplo 6** referentes à escola X de ensino fundamental e médio para a aplicação e compreensão da média aritmética ponderada.

Primeiro precisamos definir o valor do ponto médio x_i de cada uma das classes. O ponto médio é dado pela média simples dos dois extremos do intervalo da classe desejada, ou seja, para a primeira classe temos que:

$$\bar{x}_1 = \frac{10 + 12}{2} = 11$$

Vamos adicionar à Tabela 11 uma coluna com os valores referentes ao ponto médio de cada uma das classes, conforme expresso na Tabela 12.

Tabela 12: Distribuição por agrupamento de classes de idades dos alunos da escola X a partir do 6º ano

Intervalo de idades dos alunos (anos)	Ponto Médio (x_i)	Frequência absoluta (nº de alunos)
[10 ; 12[	11	114
[12 ; 14[	13	157
[14 ; 16[	15	202
[16 ; 18[	17	187
[18 ; 20[	19	106
[20 ; 22[	21	41

Tendo o valor do ponto médio podemos determinar a média aritmética ponderada. Vamos, portanto aos cálculos?

Primeiro determinamos o valor do numerador da equação:

$$\sum_{i=1}^{6} x_i . fa_i = (11 \times 114) + (13 \times 157) + (15 \times 202) + (17 \times 187) + (19 \times 106) + (21 \times 41)$$

$$\sum_{i=1}^{6} x_i . fa_i = 12379$$

Em seguida, determinamos o denominador, que é o somatório das frequências, vejamos:

$$\sum_{i=1}^{6} fa_i = 114 + 157 + 202 + 187 + 106 + 41$$

$$\sum_{i=1}^{6} fa_i = 807$$

Agora podemos determinar o valor final da média aritmética ponderada efetuando a divisão entre os resultados. Sendo assim temos:

$$\bar{x} = \frac{\sum_{i=1}^{6} x_i . fa_i = \frac{12379}{807} \cong 15{,}34 anos}{n}$$

A unidade referente à média será sempre a mesma da variável (x_i), neste caso a unidade será anos.

É importante destacar que a utilização do ponto médio para determinar o valor da média em um conjunto agrupado por classes nem sempre representa bem a amostra, pois ao agrupar os dados perdemos algumas informações quanto à distribuição dos dados, generalizando os resultados em classes. Na verdade o que estamos fazendo é uma aproximação pelo ponto médio da classe ponderando-se pela frequência da respectiva classe.

Um problema encontrado na utilização da média aritmética (simples ou ponderada) é devido ao fato de utilizar todos os valores do conjunto, e em alguns casos pode haver valores discrepantes, esses valores podem destorcer o resultado final da média, pois não representa a amostra.

Outro problema comum encontrado na utilização da média aritmética é que o valor obtido pelo calculo não pode ser assumido pela variável. Podemos observar isso no exercício resolvido a cima, onde o valor obtido apresenta casas decimais, porém sabemos que a medida em anos é sempre inteira. Esse tipo de resultado é comum em variáveis discretas

Vale lembrar que utilizamos $\bar{x}$ quando trabalhamos com o calculo da média de uma amostra, já quando trabalhamos com uma população completa devemos utilizar o símbolo μ para nos referir à média. A forma de calcular é a mesma para as duas situações, porém as notações são diferentes.

3. Mediana

A mediana é o ponto que divide o os elementos ordenados (crescente ou decrescente) do conjunto em duas partes iguais. Ou seja, metade dos termos se encontra antes da mediana, e a outra metade depois.

Diferentemente da média aritmética os valores discrepantes do conjunto pouco afetam no resultado da mediana.

Um exemplo simples da determinação da mediana de um conjunto pode ser visto na imagem a seguir.

$$\{1, 3, 4, 5, 6, 6, 8, 9, 11\}$$

Md

Para determinar a posição da mediana, primeiramente devemos ordenar os elementos do conjunto de forma crescente ou decrescente, e em seguida, determinar se o conjunto apresenta um número par ou impar de elementos, isso afetará em qual equação deveremos utilizar.

Quando o conjunto apresenta um número impar de elementos a determinação da mediana é mais simples e direto, sendo dada por um dos elementos do conjunto.

$$M_d = x_i = x_{\left(\frac{n+1}{2}\right)}$$

Onde:

n é o número de elementos do conjunto;

x_i é o termo central do conjunto;

i é a posição do termo central, onde $i = \left(\frac{n+1}{2}\right)$;

Quando o conjunto que apresenta um número par de elementos é necessário realizar uma média dos dois valores centrais, suas posições podem ser determinada pela seguinte equação:

$$M_d = \frac{x_i + x_{i+1}}{2} = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$$

Onde:

n é o número de elementos do conjunto;

x_i é o termo central do conjunto;

i é a posição do termo central, onde $i = \left(\frac{n}{2}\right)$;

Agora vamos aplicar esse aprendizado nas tabelas fornecidas no inicio da unidade. Primeiro iremos determinar a mediana da tabela de distribuição da produção de leite na fazenda A. Primeiro precisamos ordenar os termos em ordem crescente, matematicamente chamamos os números ordenados por sequência monótona:

$$X = \{15{,}5 \quad 16{,}0 \quad 16{,}8 \quad 18{,}1 \quad 19{,}3 \quad 19{,}7 \quad 20{,}2 \quad 20{,}9 \quad 21{,}0 \quad 22{,}8 \quad 23{,}4 \quad 24{,}6\}$$

Ao contar os termos temos que $n = 12$, sendo ***n*** um valor par devemos aplicar a seguinte equação:

$$M_d = \frac{x_i + x_{i+1}}{2} = \frac{x_{\left(\frac{12}{2}\right)} + x_{\left(\frac{12}{2}+1\right)}}{2} = \frac{x_6 + x_7}{2}$$

Já sabemos qual a posição dos termos que devemos efetuar a média, portanto vamos enumerar os termos para identificarmos o 6º e 7º termo do conjunto.

{15,5	16,0	16,8	18,1	19,3	19,7	20,2	20,9	21,0	22,8	23,4	24,6}
1º	2º	3º	4º	5º	6º	7º	8º	9º	10º	11º	12º

Identificados os termos de interesse, já podemos substitui-los na equação.

$$M_d = \frac{x_6 + x_7}{2} = \frac{19{,}7 + 20{,}2}{2} = 19{,}95.10^3 \frac{L}{mês}$$

Ao compararmos o valor da média (19,9.) e da mediana (19,95) é possível observar que os valores estão muito próximos entre si. Isso demostra que o conjunto apresentar termos que representam bem a amostra, e não há termos discrepantes.

Para dados distribuídos em uma tabela de agrupamento por classes devemos utilizar os dados da frequência para determinar a posição da mediana. Vamos utilizar a Tabela 13 com os dados do **Exemplo 6**(escola X), para nos ajudar a identificar a classe em que a mediana se encontra iremos adicionar a tabela uma coluna com a frequência relativa acumulada (visto na unidade 3).

Tabela 13: Tabela de distribuição por agrupamento de classes de idades dos alunos da escola X a partir do 6º ano com frequência absoluta

Intervalo de idades dos alunos (anos)	Ponto Médio (x_i)	Frequência absoluta (nº de alunos)	Frequência absoluta acumulada
[10 ; 12 [	11	114	114
[12 ; 14 [	13	157	271
[14 ; 16[	15	202	473
[16 ; 18[	17	187	660
[18 ; 20[	19	106	766
[20 ; 22[	21	41	807

Temos, portanto que a primeira classe detém os 114 primeiros termos, a segunda classe detém do termo 115 à 271, e assim por diante. Vamos então determinar a posição da mediana. Sabemos que a soma dos termos (soma da frequência) é igual a 807, um número impar, portanto devemos utilizar a seguinte equação:

$$M_d = x_i = x_{\left(\frac{807+1}{2}\right)=x_{404}}$$

Determinamos então que o valor da mediana é dado pelo 404º termo do conjunto. Sendo assim ele se encontra na terceira classe (linha) de intervalo [14 ; 16[. Porém, não é possível determinar com exatidão qual é o valor da idade desse termo, então utilizaremos outra equação para determinar o valor da mediana. Mas já sabemos que é algo entre 14 e 16 anos.

4. Medidas de dispersão

A medida de dispersão ou também chamadas de medidas de variabilidade tem o como objetivo medir o quão próximos uns dos outros estão os valores de uma amostra ou grupo. A partir de cálculos é possível obter valores numéricos que sintetizam a variabilidade.

Utilizaremos os dados do **Exemplo 7** a seguir para o estudo das medidas de variação. Os dados são das turmas A e B do 8º ano, referente a uma prova de matemática aplicada com o valor máximo de 10 pontos, estão expressos na Tabela 14.

Tabela 14: Tabela de distribuição de notas das turmas A e B

Turma	A	B
Notas	9,3	3,0
	8,0	8,6
	4,2	9,1
	5,0	5,3
	7,0	4,8
	8,5	4,4
	6,4	7,9

Nesta unidade iremos tratar de medidas de dispersão como a amplitude total ou intervalo, variância, desvio padrão e o coeficiente de variação. Vamos começar?

5. Variância (S^2)

A variância é vista como sendo a medida de dispersão mais importante para o estudo de dados, pois engloba em seus cálculos todos os dados da amostra ou população, seus cálculos são fáceis e representa uma estatística de fácil compreensão e aplicação.

A variância utiliza a diferença entre o dado analisado e a média aritmética do conjunto para encontrar os chamados erros ou desvios encontrados na distribuição dos dados. O erro de cada um dos elementos é encontrado a partir da seguinte relação:

$$e_i = x_i - \bar{x}$$

Sendo x_i um elemento qualquer da amostra ou população e $\bar{x}$ a média aritmética de todos os elementos. É importante ressaltar que o somatório do erro de todos os elementos de uma amostra ou população deve ser sempre igual a zero, ou seja:

$$\sum_{i=1}^{n} e_i = 0 \blacksquare$$

Embora a base do cálculo da variância seja os desvios em relação a média, eles não podem ser utilizados como medida descritiva pois sempre somam 0, independentemente do número de elementos do conjunto e idependentemente do padrão de dispersão dos dados. Porém, se elevamos a diferença ao quadrado, assim eliminamos os sinais negativos dos desvios em relação à media, fazendo com que a soma destas diferenças seja sempre positiva e reflitam o padrão de dispersão dos dados. O próximo passo é fazer uma ponderação para o tamanho da amostra. A equação que representa a variância é a seguinte:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Onde:

x_i é um elemento qualquer da amostra ou população;

$\bar{x}$ é a média aritmética de todos os elementos;

n é o número de elementos da amostra e

$(n - 1)$ é chamado de grau de liberdade.

Variância com valores baixos, próximos de zero, significa que os dados observados estão muito próximos a média, ou seja, a pouca variação entre os elementos. Já para valores altos de variância os dados da amostra ou população estão muito distantes da média, portanto a uma variação muito grande entre os elementos.

Algumas propriedades da variância são importantes para a melhor compreensão, então vamos conhecê-las:

Variância de Constante - A variância de uma constante k é nula: $S^2(k) = 0, k = constante$;

Soma de Constante aos Dados - Somar ou ao subtrair uma constante k a todos os dados, não altera a variância: $Z = X + K$, $S^2(Z) = S^2(X)$;

Multiplicação dos dados por Constante - Multiplicando todos os dados por uma constante k, a variância é multiplicada por: $Z = X.K$, $S^2(Z) = K^2.S^2.(X)$;

Vamos então aplicar o que aprendemos sobre variância até aqui para resolvermos o problema referente as notas, apresentado ao inicio deste tópico.

Primeiro precisamos determinar a média aritmética das notas para a turma A e B.

$$\bar{x}_A = \frac{48{,}4}{7} \cong 6{,}91 \quad \text{e} \quad \bar{x}_B = \frac{43{,}1}{7} \cong 6{,}16$$

Vamos agora determinar a valor da variância para cada uma das turmas:

$$S_A^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_A)^2}{n} \qquad S_A^2 = \frac{(9{,}3 - 6{,}91)^2 + (8{,}0 - 6{,}911)^2 + \cdots + (6{,}4 - 6{,}9)^2}{7-1} = 3{,}45$$

$$S_B^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_B)^2}{n} \qquad S_B^2 = \frac{(3{,}0 - 6{,}16)^2 + (8{,}6 - 6{,}16)^2 + \cdots + (7{,}9 - 6{,}16)^2}{7-1} = 5{,}55$$

Se descuidadamente você ao levantar as notas dos alunos use os limites 0 e 100 como escala de referência implicaria dizer que o invés de você considerar a nota 9,3 ; 8,0; 4,2; etc estaria considerando as notas multiplicadas pela constante 10, de tal forma que teríamos 93, 80, 42, e assim por diante. Notem que estamos falando da multiplicação por uma constante da variável observada. Assim seguindo a **Propriedade 3** teríamos que a variância seria $K^2.S^2(X) = 10^2.3{,}45 = 345$.

Se considerarmos a mesma situação para a turma B, a variância seria $K^2.S^2(X) = 10^2.5{,}55 = 555$

Observem que quando multiplicamos um conjunto de dados por uma constate alteramos o padrão de dispersão dele isto reflete na magnitude da variância. Ao contrário do que ocorre quando somamos uma constate a um conjunto de dados, caso em que altera-se apenas a locação da média, mas não se altera o padrão de dispersão dos dados, por isto, conforme se depreende da **Propriedade 2** a soma de uma constante não altera a variância dos dados.

Utilizando os Conceitos

Considerando as turmas A e B do Exemplo 7 qual delas apresenta maior uniformidade entre as notas? **Resposta:** Turma A – possui menor variância que Turma B.

Vale lembrar que a unidade utilizada para os resultados da variância sempre estarão ao quadrado (KM^2, mm^2, Kg^2, e etc), isso causa confusão na compreensão do resultado. Para resolver esse problema, podemos utilizar o desvio padrão, que veremos a seguir.

6. Desvio Padrão

Assim como a variância, o desvio padrão é uma medida de distribuição dos dados em torno da média amostral. Tanto em probabilidade quanto em estatística, o desvio padrão é usado para expressar outros conceitos matemáticos importantes como o coeficiente de correlação, o coeficiente de variação ou a alocação ótima de Neyman, dentre outros.

O desvio é representado pela raiz quadrada positiva da variância, veja a seguir:

$$S = \sqrt[2]{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}} = \sqrt[2]{S^2}$$

Assim como para a variância, o desvio padrão também apresenta algumas propriedades importantes, vamos conhecê-las agora:

Soma de Constante - Ao somar ou ao subtrair uma constante k a todos os dados, o desvio padrão não se altera: $Z = X + K$, $S(Z) = S(X)$;

Multiplicação por Constante - Multiplicando todos os dados por uma constante k, o desvio padrão fica multiplicado por k: $Z = X.K$, $S(Z) = S(X)$.

Agora vamos aplicar esses conhecimentos para determinar o valor do desvio padrão para as notas da turma A e B.

$$S_A = \sqrt[2]{S_A^2} = \sqrt[2]{3{,}45} \cong 1{,}86 \qquad S_B = \sqrt[2]{S_B^2} = \sqrt[2]{5{,}55} \cong 2{,}36$$

Agora partimos para o última medida de dispersão a ser estudado neste material, o coeficiente de variação.

Exercícios Medidas de Posição e Dispersão

1) Calcule a Média para os dados apresentados no Exemplo1, Exemplo 2, Exemplo 3, Exemplo 4, e Exemplo 5.

2) Calcule a Mediana para os dados apresentados no Exemplo1, Exemplo 2, Exemplo 3, Exemplo 4, e Exemplo 5.

3) Calcule Variância para os dados apresentados no Exemplo1, Exemplo 2, Exemplo 3, Exemplo 4, e Exemplo 5.

4) Calcule o Desvio Padrão para os dados apresentados no Exemplo1, Exemplo 2, Exemplo 3, Exemplo 4, e Exemplo 5.

5) Calcule a média de todas as variáveis quantitativas presente no conjunto de dados do Anexo 1.

6) Calcule a Variância de todas as variáveis quantitativas presente no conjunto de dados do Anexo 1.

7) Calcule a Média das séries de fechamento das bolsas europeias dispostas no banco de dados do Anexo 2.

8) Calcule a Média das séries de fechamento das bolsas europeias dispostas no banco de dados do Anexo 3.

9) Comparados os resultados do Anexos 2 e do Anexo 3: alguma das bolsas europeias reduziu o seu valor?

10) Calcule a variância e do desvio padrão de todas as 4 variáveis do Anexo 2.

11) Calcule a variância e do desvio padrão de todas as 4 variáveis do Anexo 3.

12) Sabendo que o Anexo 2 e o Anexo 3 lidam com as mesmas bolsas de valores, compare os resultados do exercício 10 com os resultados do exercício 11.

13) As variâncias e os desvio padrões obtido no exercício 10 e no exercício 11, podem ser utilizadas para medir a instabilidade dos índices nos períodos respectivos. Pergunta-se:

a) Em 10 qual das bolsas apresentava maior instabilidade?
b) Em 11 qual das bolsas apresentava maior instabilidade?

14) Considerando a produção de leite apresenta no Exemplo 5. Suponha que ao você tomar nota da produção, você errou a escala, e colocou um zero a mais em todas as observações o que fez com que todas as observações ficassem multiplicadas por 10. Crie este conjunto de dados (multiplicado por 10) e calcule a média, a variância e o desvio padrão

15) Comente os resultados obtidos no exercício 14 à luz das propriedades apresentadas nesta secção.

16) Considerando que ao baixar o arquivo com os dados do Anexo 3, as ações da DAX (Bolsa de valores alemã) foram coladas em uma célula que já estava formatada. Assim todos os valores ficaram somados com 300, pois esta era a operação da formatação. Sendo assim crie o novo conjunto de dados (DAX+300) e calcule a média, a variância e o desvio padrão.

17) Comente os resultados obtidos no Exercício 16 à luz das propriedades apresentadas nesta secção.

Referências

ACTION, Portal. http://www.portalaction.com.br. Acesso em 07 de Janeiro de 2020.

BOLFARINE, Heleno; BUSSAB Wilton. Elementos de. Amostragem. São Paulo: Ed. Blucher, 2005.

BUSSAB Wilton de O. e MORETTIN Pedro A., *Estatística Básica*, Saraiva, Sao Paulo, 9ed, 2017.

MAGALHÃES, Marcos Nascimento. Probabilidade e Variáveis Aleatórias. São Paulo: EdUSP. 2006.

MEYER, Paul L. Probabilidade: Aplicações a Estatística. LTC, 1983.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2018. URL https://www.R-project.org/. Acesso em 12 dezembro 2019.

Conjunto de dados Milsa[1]

Funcionário	Est.civil	Inst	Filhos	Salário	Anos	Meses	Região
1	solteiro	1o Grau	NA	4.00	26	3	interior
2	casado	1o Grau	1	4.56	32	10	capital
3	casado	1o Grau	2	5.25	36	5	capital
4	solteiro	2o Grau	NA	5.73	20	10	outro
5	solteiro	1o Grau	NA	6.26	40	7	outro
6	casado	1o Grau	0	6.66	28	0	interior
7	solteiro	1o Grau	NA	6.86	41	0	interior
8	solteiro	1o Grau	NA	7.39	43	4	capital
9	casado	2o Grau	1	7.59	34	10	capital
10	solteiro	2o Grau	NA	7.44	23	6	outro
11	casado	2o Grau	2	8.12	33	6	interior
12	solteiro	1o Grau	NA	8.46	27	11	capital
13	solteiro	2o Grau	NA	8.74	37	5	outro
14	casado	1o Grau	3	8.95	44	2	outro
15	casado	2o Grau	0	9.13	30	5	interior
16	solteiro	2o Grau	NA	9.35	38	8	outro
17	casado	2o Grau	1	9.77	31	7	capital
18	casado	1o Grau	2	9.80	39	7	outro
19	solteiro	Superior	NA	10.53	25	8	interior
20	solteiro	2o Grau	NA	10.76	37	4	interior
21	casado	2o Grau	1	11.06	30	9	outro
22	solteiro	2o Grau	NA	11.59	34	2	capital
23	solteiro	1o Grau	NA	12.00	41	0	outro
24	casado	Superior	0	12.79	26	1	outro
25	casado	2o Grau	2	13.23	32	5	interior
26	casado	2o Grau	2	13.60	35	0	outro
27	solteiro	1o Grau	NA	13.85	46	7	outro
28	casado	2o Grau	0	14.69	29	8	interior
29	casado	2o Grau	5	14.71	40	6	interior
30	casado	2o Grau	2	15.99	35	10	capital
31	solteiro	Superior	NA	16.22	31	5	outro
32	casado	2o Grau	1	16.61	36	4	interior
33	casado	Superior	3	17.26	43	7	capital
34	solteiro	Superior	NA	18.75	33	7	capital
35	casado	2o Grau	2	19.40	48	11	capital
36	casado	Superior	3	23.30	42	2	interior

1 Bussab e. Morettin (2017)

Dados – Mercado de Ações Europeu

Conjunto de dados Mercado de Ações Europeu por R Core Team (2018)[2]. Adaptação 20 primeiras observações.

```
> EuStockMarkets[1:20,]
          DAX    SMI    CAC   FTSE
 [1,] 1628.75 1678.1 1772.8 2443.6
 [2,] 1613.63 1688.5 1750.5 2460.2
 [3,] 1606.51 1678.6 1718.0 2448.2
 [4,] 1621.04 1684.1 1708.1 2470.4
 [5,] 1618.16 1686.6 1723.1 2484.7
 [6,] 1610.61 1671.6 1714.3 2466.8
 [7,] 1630.75 1682.9 1734.5 2487.9
 [8,] 1640.17 1703.6 1757.4 2508.4
 [9,] 1635.47 1697.5 1754.0 2510.5
[10,] 1645.89 1716.3 1754.3 2497.4
[11,] 1647.84 1723.8 1759.8 2532.5
[12,] 1638.35 1730.5 1755.5 2556.8
[13,] 1629.93 1727.4 1758.1 2561.0
[14,] 1621.49 1733.3 1757.5 2547.3
[15,] 1624.74 1734.0 1763.5 2541.5
[16,] 1627.63 1728.3 1762.8 2558.5
[17,] 1631.99 1737.1 1768.9 2587.9
[18,] 1621.18 1723.1 1778.1 2580.5
[19,] 1613.42 1723.6 1780.1 2579.6
[20,] 1604.95 1719.0 1767.7 2589.3
```

```
> EuStockMarkets[1841:1860,]
          DAX    SMI    CAC   FTSE
 [1,] 6186.09 8400.8 4368.9 6179.0
 [2,] 6184.10 8412.0 4322.1 6132.7
 [3,] 6081.11 8340.7 4220.1 5989.6
 [4,] 6043.82 8229.2 4235.9 5976.2
 [5,] 6040.58 8205.7 4205.4 5892.3
 [6,] 5854.35 7998.7 4139.5 5836.1
 [7,] 5867.52 8093.0 4122.4 5835.8
 [8,] 5828.74 8102.7 4139.2 5844.1
 [9,] 5906.33 8205.5 4197.6 5910.7
[10,] 5861.19 8239.5 4177.3 5837.0
[11,] 5774.38 8139.2 4095.0 5809.7
[12,] 5718.70 8170.2 4047.9 5736.1
[13,] 5614.77 7943.2 3976.4 5632.5
[14,] 5528.12 7846.2 3968.6 5594.1
[15,] 5598.32 7952.9 4041.9 5680.4
[16,] 5460.43 7721.3 3939.5 5587.6
[17,] 5285.78 7447.9 3846.0 5432.8
[18,] 5386.94 7607.5 3945.7 5462.2
[19,] 5355.03 7552.6 3951.7 5399.5
[20,] 5473.72 7676.3 3995.0 5455.0
```

2 Dados dizem respeito ao índice de fechamento das 4 maiores Bolsas europeias. Alemanha, Suíça, França e Reino Unido. Fonte: https://www.rdocumentation.org/packages/datasets/versions/3.6.2/topics/EuStockMarkets

Coordenadoria de
Educação Aberta e a Distância