

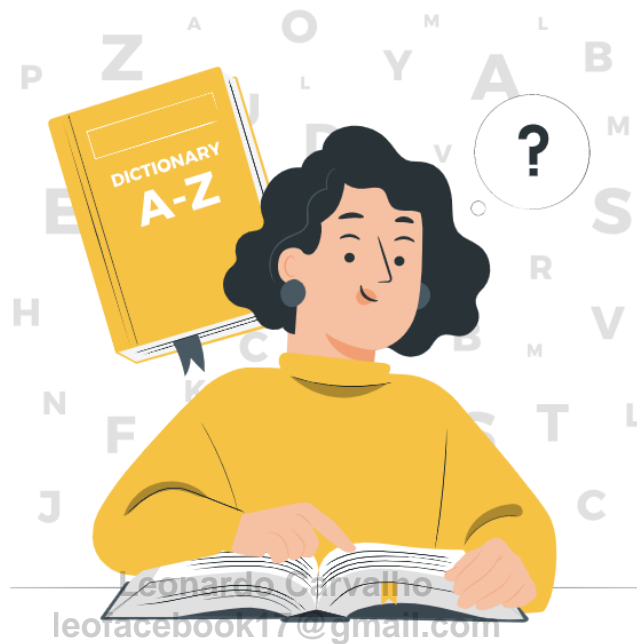
E.B.A - ESTATÍSTICA DO BÁSICO AO AVANÇADO

COM RENATA BIAGGI

ESTATÍSTICA DESCRITIVA



GLOSSÁRIO DE SÍMBOLOS



ITEM DE UMA SEQUÊNCIA

Se temos um conjunto de dados, cada item pode ser representado como sendo x_1, x_2, x_3 , etc. Por exemplo, temos o conjunto 10, 21, 32, 43, 58, então:

$x_1 = 10, x_2 = 21, x_3 = 32, x_4 = 43, x_5 = 58$

SOMATÓRIO - Σ

Em matemática, somatório ou somatória é a adição de uma sequência de quaisquer tipos de números. O resultado é sua soma ou total.



$$\sum_{i=1}^n x_i = x_1 + x_2 + \cdots + x_n$$

O índice i significa a partir de qual elemento devemos começar a soma. O símbolo n indica que, nesse caso, vamos somar a sequência toda - que tem um total de n elementos.

Vamos a alguns exemplos. Vamos supor que queremos somar os números 1, 2 e 3, ou seja, temos um total de 3 elementos para somar:

$$\sum_{i=1}^3 x_i = x_1 + x_2 + x_3 = 1 + 2 + 3 = 6$$

Muitas vezes vemos o somatório em expressões como:

$$\sum_{n=1}^3 2n - 1$$

Leonardo Carvalho
leofacebook17@gmail.com

Esse somatório indica que devemos somar a expressão $2 \cdot n - 1$, substituindo o n por cada um dos valores da série. Logo, para a série 1, 2 e 3 temos

$$\begin{aligned} & \sum_{n=1}^3 2n - 1 \\ &= \underbrace{[2(1) - 1]}_{n=1} + \underbrace{[2(2) - 1]}_{n=2} + \underbrace{[2(3) - 1]}_{n=3} \\ &= 1 + 3 + 5 \\ &= 9 \end{aligned}$$

PRODUTÓRIO - Π

O produtório é a multiplicação de uma sequência de objetos matemáticos (números, funções, vetores, matrizes, etc.), chamados fatores, que tem como resultado o produto. É uma operação análoga ao somatório, embora seja menos utilizada quanto esse último. É representado pela letra grega pi maiúscula (Π).

$$\prod_{i=1}^n x_i = x_1 \times x_2 \times \dots \times x_n$$

OUTROS SÍMBOLOS IMPORTANTES

| Símbolo | Parâmetro |
|------------|----------------------------|
| μ | Média populacional |
| σ | Desvio-padrão populacional |
| σ^2 | Variância populacional |
| p | Proporção populacional |
| \bar{x} | Média amostral |
| s | Desvio-padrão amostral |
| s^2 | Variância amostral |
| \hat{p} | Proporção amostral |
| α | Nível de significância |

COMO USAR ESSA APOSTILA



Bem-vindos(as) ao nosso curso de E.B.A - Estatística do básico ao avançado! Sou a professora Renata Biaggi e estarei com vocês de perto nessa jornada em busca de mais conhecimento técnico e de tomadas de decisões bem fundamentadas.

Como vocês bem sabem, durante nosso curso teremos aulas online com bastante teoria e bastante prática. Teremos também exercícios extra-aulas e essa apostila, que escrevi pensando em guiar e complementar os estudos de vocês.

A apostila é composta por todos os tópicos que veremos em sala de aula de uma forma muito detalhada, teórica e exemplificada. Minha sugestão é que vocês leiam o tópico referente a aula **antes** de assistirem as aulas referentes. Isso vai possibilitar que vocês tenham um desempenho muito maior.

Ao longo da apostila vocês vão encontrar várias citações de autores, bem como indicações de livros e blogs para que vocês complementem os estudos

- especialmente para os tópicos nos quais não vamos nos aprofundar durante o curso. Novamente, recomendo fortemente que todo esse material seja lido antes de cada aula, beleza?

Antes de partirmos para o conteúdo de fato, vamos alinhar algumas expectativas. O intuito do nosso curso não é aprender como usar linguagens de programação ou ferramentas de uma forma geral. Queremos aqui focar em **estatística e teste de hipótese**, entretanto para isso precisamos do auxílio de alguma interface para fazermos cálculos de uma forma mais rápida e para lidar com todos os nossos dados. Por isso, os dois próximos tópicos vão se dedicar a abordar de forma bastante rápida as principais ferramentas que guiarão esse curso: Python e Excel. Depois desses dois tópicos, entraremos no conteúdo que aborda a nossa tão amada ~~matemática~~ matemática.

Outro ponto importantíssimo de ressaltarmos é que nesta apostila vamos detalhar os cálculos de todos os exemplos para demonstrarmos como usar cada fórmula. Na prática, dificilmente calcularemos qualquer coisa "na mão". Softwares como Excel ou linguagens como Python já tem esses cálculos intrínsecos, salvando bastante nosso tempo. Da a importância dos capítulos introdutórios, uma vez que precisaremos manipular tais ferramentas para que elas façam todos os cálculos por nós.

Então Renata, pra quê eu preciso ver todas as fórmulas e entender os exemplos de cálculo? Porque aqui nós **não seremos ferramenteiros!** Nosso intuito é sair do curso entendendo de fato cada conceito que dá suporte às suas análises e modelos preditivos para que, quando vocês se depararem com situações complexas reais, vocês saibam agir sozinhos e possam pensar criticamente no que fazer.

Preparados? Bora lá!

1. INTRODUÇÃO À ESTATÍSTICA



Em alguma fase de seu trabalho, você já deve ter se deparado com um problema que só poderia ser resolvido analisando e entendendo um conjunto de dados relevantes aos seus estudos. De forma geral, esses conjuntos de dados coletados precisam ser transformados em informações, para compará-los com outros resultados, ou ainda para julgar sua adequação a alguma teoria. Podemos dizer que a essência da estatística.

A estatística, de forma formal, é descrita como a ciência que se preocupa em desenvolver e estudar métodos para coletar, analisar, interpretar e apresentar dados empíricos. É um campo altamente interdisciplinar; a pesquisa em estatística encontra aplicabilidade em praticamente todos os campos de conhecimento humano - business, área da saúde, sociologia, entre outros.

Temos basicamente duas divisões quanto ao assunto é estatística: a descrição e a inferência. A **descrição** é a forma que temos de resumir os nossos dados - e dedicamos alguns ao longo da apostila para realizar essa função.

Já a **inferência** está em tudo que tange à incerteza. Existem muitas situações que encontramos na ciência (ou mais geralmente na vida) em que o resultado é incerto. Em alguns casos, a incerteza é porque o resultado em questão ainda não foi determinado (por exemplo, podemos não saber se

choverá amanhã), enquanto em outros casos a incerteza é porque, embora o resultado já tenha sido determinado, não estamos cientes disso (por exemplo, podemos não saber se passamos em um exame específico).

Nesse contexto de incertezas, a probabilidade desempenha um papel fundamental e se dedica justamente a tentar medir de forma matemática quão improvável é um evento acontecer. Qualquer esforço de medição ou coleta de dados está sujeito a várias fontes de variação. Com isso quero dizer que, se a mesma medida/experimento fosse repetida com dados diferentes, a resposta provavelmente mudaria. Os estatísticos tentam entender e controlar (quando possível) as fontes de variação em qualquer situação - e a probabilidade entra com força nisso.

Dentro do cenário de incerteza, como fazemos para tomar uma decisão correta? Como manipular os dados de forma correta e tirar os insights adequados? A estatística e todos os seus campos vão nos ajudar nessa missão.

Leonardo Carvalho
leofacebook17@gmail.com

2. TIPOS DE DADOS E ALGUMAS REPRESENTAÇÕES



TIPOS DE DADOS

Leonardo Carvalho
leofacebook17@gmail.com

Dados são valores atribuídos a algo. Estes valores não precisam ser necessariamente números. Eles também podem ser, por exemplo, conceitos ou posições em um mapa. Dados podem ser medidos ou mensurados por meio de instrumentos, mas também podem ser atribuídos de forma arbitrária (ou seja, por opinião - por exemplo, pesquisas de satisfação).

Dividimos os dados em 2 categorias: numéricos (também chamados de quantitativos) e categóricos (também chamados de qualitativos). Os dados numéricos ainda podem ser divididos em discretos e contínuos, enquanto os categóricos podem ser divididos em nominal e ordinal.

Tipos de dados



Variável qualitativa nominal são valores que expressam atributos, sem nenhum tipo de ordem. Ex: cor dos olhos, sexo, estado civil, presença ou ausência.

Variável qualitativa ordinal são valores que expressam atributos, porém com algum tipo de ordem, ou grau. Ex: grau de escolaridade (1º grau, 2º grau, 3º grau, pós-graduação...); resposta de um paciente (nenhuma melhora, alguma melhora, muita melhora); classe social (alta, média, baixa).

Variável quantitativa discreta são valores observados somente em pontos isolados ao longo de uma escala de valores - ou seja, temos uma quantidade finita de dados. São valores positivos inteiros (incluindo o zero). Ex: Número de filhos(0, 1, 2, ...); Número de faltas; alunos com notas abaixo de 5,0.

Variável quantitativa contínua são valores em qualquer ponto fracionário ao longo de um intervalo especificado de valores - ou seja, temos uma quantidade quase infinita de dados. De forma geral, são números com casas depois da vírgula. Ex: temperatura do corpo; altura (em metros).

Para cada tipo de variável existem técnicas apropriadas para resumir as informações. Entretanto, verificaremos que técnicas usadas num caso podem ser adaptadas para outros.

Para finalizar, cabe uma observação sobre variáveis **qualitativas**. Em algumas situações podem-se atribuir valores numéricos às várias qualidades ou atributos (ou, ainda, classes) de uma variável qualitativa e depois proceder-se à análise como se esta fosse quantitativa, desde que o procedimento seja passível de interpretação.

Existe um tipo de variável qualitativa para a qual essa quantificação é muito útil: a chamada **variável booleana**. Uma variável booleana é aquela que pode assumir apenas dois valores. Esses valores geralmente são 0, como ausência, ou 1, como presença. Vamos supor que trabalhamos em um banco e precisamos coletar os dados de transações feitas no cartão de crédito de uma pessoa para analisarmos se houve compras fraudulentas - ou seja, compras feitas por um fraudador que estava tentando roubar o dinheiro do dono do cartão de crédito. Podemos classificar essas transações como 0 quando a transação não é uma fraude e 1 quando aquela transação é uma fraude. Essa é uma variável booleana.

REPRESENTAÇÕES TABULAR DE FREQUÊNCIA DE CADA TIPO DE DADO

Quando se estuda uma variável, um dos grandes interesses é conhecer o comportamento dessa variável, analisando a ocorrência de seus possíveis valores. Vamos dar uma olhadinha em alguns tipos de **representação de frequência**, começando pelas variáveis categóricas.

Para essa seção, vamos usar alguns exemplos da referência (Bussab, W, Morettin, P.). Vamos começar com os **dados categóricos**. Suponhamos que fizemos uma pesquisa com 36 funcionários da seção de "orçamentos" de uma empresa para identificar o nível de escolaridade deles (ou seja, nosso dado de interesse é o nível de escolaridade - uma variável categórica). Nosso resultado foi:

| Grau de instrução | Frequência n_i | Proporção f_i | Porcentagem $100 f_i$ |
|-------------------|---------------------|--------------------|--------------------------|
| Fundamental | 12 | 0,3333 | 33,33 |
| Médio | 18 | 0,5000 | 50,00 |
| Superior | 6 | 0,1667 | 16,67 |
| Total | 36 | 1,0000 | 100,00 |

A coluna "frequência" indica quantos funcionários tem cada um dos níveis. Por exemplo, 12 funcionários da nossa pesquisa estão no nível fundamental. A coluna "proporção" divide a frequência encontrada pelo total que coletamos. Voltando ao exemplo do "fundamental", dos 36 funcionários que coletamos, 12 tinham ensino fundamental - ou seja, a frequência de ocorrência do nível fundamental na nossa pesquisa é $12/36 = 0,333$. A coluna "porcentagem" multiplica a coluna "proporção" por 100.

As **proporções**, muitas vezes chamadas de **densidade**, são muito úteis quando se quer comparar resultados de duas pesquisas distintas. Por exemplo, suponhamos que se queira comparar a variável grau de instrução para empregados da seção de orçamento com todos os funcionários da empresa. Digamos que a empresa tenha 2.000 empregados.

| Grau de instrução | Frequência n_i | Porcentagem $100 f_i$ |
|-------------------|---------------------|--------------------------|
| Fundamental | 650 | 32,50 |
| Médio | 1.020 | 51,00 |
| Superior | 330 | 16,50 |
| Total | 2.000 | 100,00 |

Não podemos comparar diretamente as colunas das frequências das duas tabelas acima, pois os totais de empregados são diferentes nos dois casos. Mas as colunas das porcentagens são comparáveis, pois reduzimos as frequências a um mesmo total (no caso 100).

Quaisquer que sejam nossas **categorias (nominal ou ordinal)**, podemos construir a tabela de frequências da mesma forma. Essa forma de construção também se aplica a **variáveis numéricas discretas**. Como essas variáveis são números inteiros com quantidade finita, podemos apenas contar a ocorrência de cada um e, para a frequência, dividir pelo total (em quantidade). Por exemplo, vamos supor que coletamos também a quantidade de filhos que cada um dos 36 funcionários tem. Nessa coleta, vimos que 3 desses funcionários tem 4 filhos - a frequência para a quantidade "4 filhos" seria então $3/36 = 0,0833 = 8,33\%$. Faríamos isso para todas as quantidades de filho coletadas.

Por outro lado, a construção de tabelas de frequências para **variáveis numéricas contínuas** precisa de certo cuidado. Por exemplo, a construção da tabela de frequências para a variável salário, usando o mesmo procedimento acima, não resumirá as 36 observações num grupo menor, pois os salários são muitos distintos um do outro (são números com casas depois da vírgula e com alta variação - há quase que infinitas possibilidades). A solução empregada é agrupar os dados por faixas de salário - isso é chamado de **discretização** ou **binarização** - ou seja, pegamos uma variável contínua e a transformamos em variável discreta. Na tabela abaixo transformamos cada salário em uma faixa de valor. A primeira faixa indica salários de 4 mil até 8 mil, a segunda faixa salarial de 8 mil até 12 mil, e assim por diante.

| Classe de salários | Frequência n_i | Porcentagem $100f_i$ |
|--------------------|---------------------|-------------------------|
| 4,00 ┤ 8,00 | 10 | 27,78 |
| 8,00 ┤ 12,00 | 12 | 33,33 |
| 12,00 ┤ 16,00 | 8 | 22,22 |
| 16,00 ┤ 20,00 | 5 | 13,89 |
| 20,00 ┤ 24,00 | 1 | 2,78 |
| Total | 36 | 100,00 |

Da mesma forma que antes, agora sim podemos contar quantos funcionários se enquadram em cada uma das faixas. Com isso, somos capazes de calcular a porcentagem final por faixa de valor.

Se fizermos desse modo, ao resumir os dados referentes a uma variável contínua, perde-se algumas informações. Por exemplo, não sabemos quais são os oito salários da classe de 12 a 16, a não ser que investiguemos a tabela original. Sem perda de muita precisão, poderíamos supor que todos os oito salários daquela classe fossem iguais ao ponto médio da referida classe, isto é, 14. Voltaremos a este assunto mais para frente, quando falarmos de medidas de tendência central.

A escolha dos intervalos é arbitrária e a familiaridade com seu tipo de business lhe indicará quantas e quais classes (intervalos) devem ser usadas. Ferramentas como Excel e Python já fazem isso instantaneamente quando decidimos optar por ver os dados na forma de gráfica (histograma).

Agora vamos nos dedicar um pouco a entender essas formas de representar graficamente a frequência e esse tal de histograma para as variáveis numéricas.

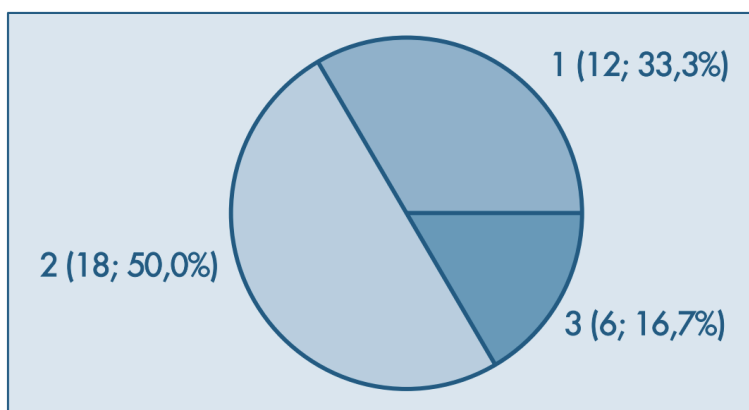
REPRESENTAÇÃO GRÁFICA DE FREQUÊNCIA PARA VARIÁVEIS QUALITATIVAS

Existem vários tipos de gráficos para representar variáveis qualitativas. Vários são versões diferentes do mesmo princípio, logo nos limitaremos a apresentar dois deles: gráficos em barras e de “pizza”.

GRÁFICO DE PIZZA

Vou apresentá-lo aqui por uma questão didática, mas nossa maior referência em visualização de dados condena veementemente o uso dos gráficos de pizza em uma apresentação pois *“é difícil ler os gráficos de pizza. Quando os segmentos têm tamanhos parecidos, é difícil (se não impossível) dizer qual é o maior.”* (Knafllic, C.N.). Quando a autora diz isso, não significa que os gráficos são complexos. Ela apenas quer dizer que esses gráficos podem acabar não passando a mensagem que você deseja. Recomendo fortemente a leitura da bibliografia Knafllic, C.N. caso você tenha interesse em melhorar suas apresentações.

Vamos ao gráfico de pizza. Esse tipo de gráfico destina-se a representar a porcentagem de cada categoria na base de dados. Consiste num círculo de raio arbitrário, representando o todo, dividido em setores, que correspondem às categorias. Para ilustrar, vamos usar como exemplo o grau de instrução dos empregados de “orçamento”, exemplificada nas tabelas acima

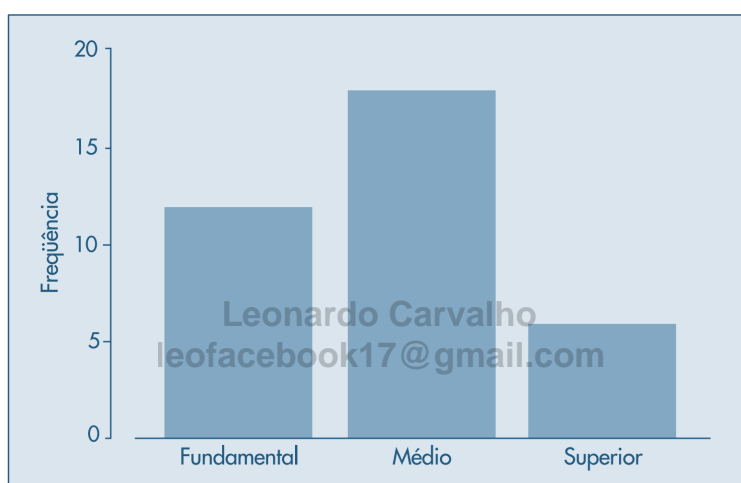


1 = Fundamental, 2 = Médio e 3 = Superior

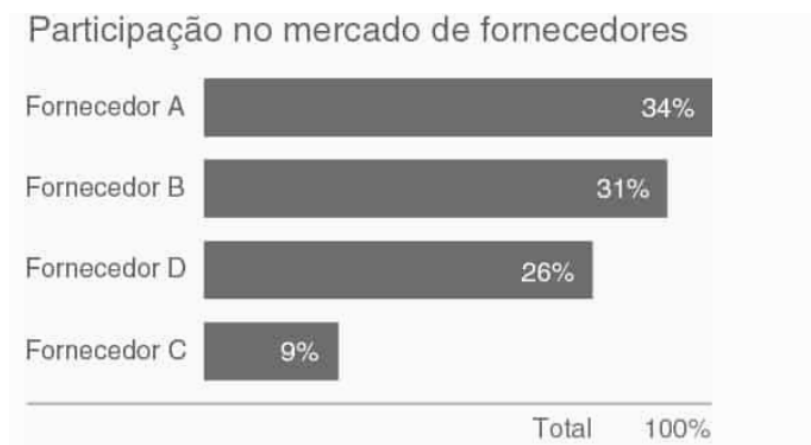


GRÁFICO DE BARRAS

Uma ótima alternativa ao gráfico de pizza é o gráfico de barras, que pode ser horizontal ou vertical. Vamos tomar a variável Y: grau de instrução dos empregados de "orçamento", exemplificada nas tabelas acima. O gráfico em barras consiste em construir retângulos ou barras, em que uma das dimensões é proporcional à magnitude a ser representada (frequência ou proporção), sendo a outra arbitrária, porém igual para todas as barras (mesma largura para todas as barras). Essas barras são dispostas paralelamente umas às outras, horizontal ou verticalmente.



Também podemos representá-lo de forma horizontal, e com percentuais. O exemplo abaixo foi retirado da referência Knafllic, C.N.



REPRESENTAÇÃO GRÁFICA DE FREQUÊNCIA PARA VARIÁVEIS QUANTITATIVAS

A representação gráfica mais comum para uma variável quantitativa também é o gráfico de barras. Primeiro, vamos dar uma olhada nas variáveis **quantitativas discretas**.

Vamos supor que perguntamos a algumas pessoas na faixa de 20 a 35 anos quantos filhos elas têm. Esse dado é um valor discreto, pois a quantidade de filhos é finita e não tem-se casas decimais. As respostas dessas pesquisas foram 0 (nenhum filho), 1 (um filho), 2 (dois filhos) e assim por diante.

Da mesma forma que para as variáveis quantitativas, uma ótima forma de representação nesse caso são gráficos de barras. No eixo X temos a quantidade de filhos e no Y quantos funcionários tem cada uma dessas quantidades.

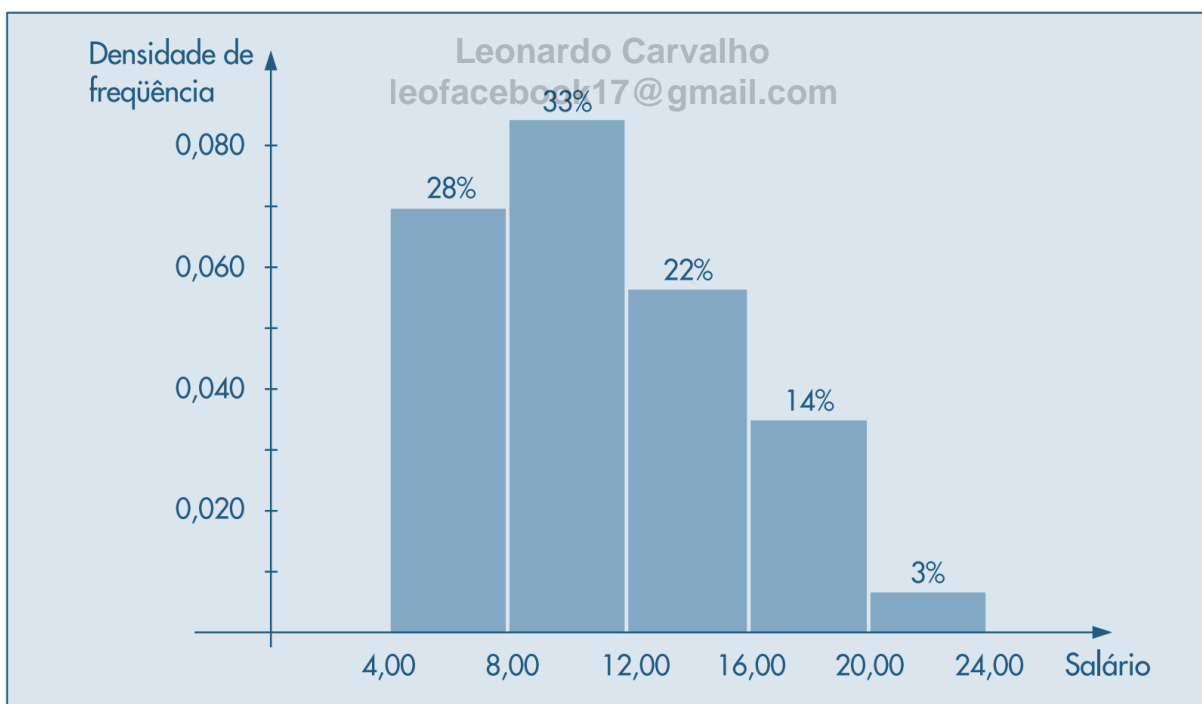


O artifício usado acima para representar uma **variável contínua** faz com que se perca muito das informações nela contidas. Uma alternativa a ser usada nestes casos é o gráfico conhecido como **histograma**.

O histograma é um gráfico de barras continua, com as bases proporcionais aos intervalos das classes e a área de cada retângulo proporcional à respectiva frequência. Pode-se usar tanto a frequência absoluta, n_i (contador), como a relativa, f_i (densidade de frequência).

Vamos ver abaixo um exemplo de distribuição de frequências da variável S, salário dos empregados da seção de orçamentos da Companhia MB.

| Classes de salários | Ponto médio s_i | Frequência n_i | Porcentagem $100 f_i$ |
|---------------------|----------------------|---------------------|--------------------------|
| 4,00 ┤ 8,00 | 6,00 | 10 | 27,78 |
| 8,00 ┤ 12,00 | 10,00 | 12 | 33,33 |
| 12,00 ┤ 16,00 | 14,00 | 8 | 22,22 |
| 16,00 ┤ 20,00 | 18,00 | 5 | 13,89 |
| 20,00 ┤ 24,00 | 22,00 | 1 | 2,78 |
| Total | — | 36 | 100,00 |



Para facilitar o entendimento, foi colocada acima de cada retângulo a respectiva percentagem das observações (arredondada). Assim, por meio da figura, podemos dizer que 28% dos funcionários têm um salário entre 4 e 8,

33% dos funcionários entre 8-12 e assim por diante. Esse gráfico também é muito útil para entender os percentuais acumulados. Podemos somar os 2 primeiros retângulos e dizer que 61% ($28\% + 33\%$) dos empregados têm salário inferior a 12. Olhando os 2 últimos retângulos, podemos dizer que 17% ($14\% + 3\%$) possuem salário superior a 16.

Sabendo que a soma dos percentuais (f_i) deve ser 100% (conforme mostra a tabela), dizemos que a soma de cada um desses retângulos deve ser 100%. Esse conhecimento será especialmente importante quando abordarmos **probabilidades**.

Leonardo Carvalho
leofacebook17@gmail.com

3. MEDIDAS DA ESTATÍSTICA DESCRITIVA



Leonardo Carvalho
leofacebook17@gmail.com

As estatísticas descritivas resumem os dados de um grupo que você escolher. Ou seja, ela basicamente descreve como está sua distribuição.

Estatísticas descritivas descrevem uma amostra. Você simplesmente pega um grupo no qual está interessado, registra dados sobre cada dado do conjunto e, em seguida, usa estatísticas e gráficos resumidos para apresentar as propriedades do grupo. Com estatísticas descritivas, **não há incerteza** porque você está **descrevendo** apenas os dados que você realmente mede.

Por exemplo, se você medir a altura de dois grupos de pessoas, você conhece as médias precisas para ambos os grupos e pode afirmar sem incerteza qual deles tem a média mais alta. Você **não** está tentando inferir propriedades sobre uma população maior.

Além de frequência numérica como mostrado anteriormente, podemos representar e entender os números a partir de medidas únicas - medidas centrais e medidas de dispersão. Vamos agora entender as principais medidas da estatística descritiva.

OUTLIERS

Antes de prosseguirmos com as medidas da estatística descritiva, é importante que façamos a introdução de um conceito que será muito falado no curso: outliers.

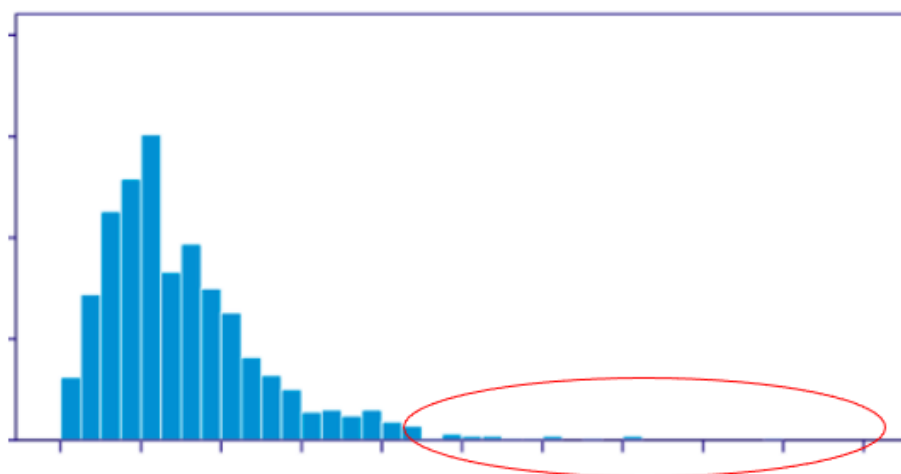
Chamamos de outliers aqueles valores que são muito discrepantes de um conjunto de dados. Por exemplo, se coletarmos a altura de 1000 mulheres brasileiras aleatórias, teríamos algo como:

1.51 m, 1.52 m, 1.52 m, 1.55 m, ..., 1.79 m, 1.80 m, 1.90 m

Em que a média é 1.60 m. Observando os dados acima, vemos que a mulher mais alta tem 1.90 m e destoa bastante da média. Dizemos então que esse dado é um outlier.

Muitas vezes usamos a palavra outliers para valores destoantes que são possíveis de acontecer ou valores destoantes impossíveis de acontecer. No nosso caso, é perfeitamente possível encontrarmos uma mulher com 1.90 m de altura. Contudo, se por acaso nessa série encontrássemos uma mulher com 4.0 m, saberíamos que é um outlier impossível de acontecer - ou seja, provavelmente houve um erro quando computaram esse dado.

Em um histograma, os outliers podem ser facilmente vistos através da cauda longa do gráfico:



Falaremos mais sobre outliers quando abordarmos sobre Boxplots.

NOTA IMPORTANTE: OUTLIERS PODEM SER VALORES MUITO PEQUENOS OU MUITO GRANDES. POR EXEMPLO, SE EXISTISSE NA SÉRIE UMA MULHER DE 1.30 M DE ALTURA, ELA TAMBÉM SERIA UM OUTLIER.

Leonardo Carvalho
leofacebook17@gmail.com

TENDÊNCIA CENTRAL

Uma medida de tendência central é um valor único que tenta descrever um conjunto de dados identificando sua posição central. São também chamadas de medidas de localização central.

Pense em como você descreve um número. Geralmente, a descrição é feita de acordo com seu valor. Por exemplo, para descrever o número 2, você pode mostrar dois dedos ou dizer $1 + 1 = 2$. Mas como você descreveria um grupo de dados? Neste caso, não adiantaria muito usar os dedos, e somar os números seria impossível. Usando medidas de tendência central, você pode descrever um grupo de dados em um único valor.

Existem três medidas de tendência central muito utilizadas no dia-a-dia: a média, a mediana e a moda.

Média — é o valor médio dos dados, calculado através da divisão da soma dos valores com o número total de valores.

Mediana — é o valor do meio na série de dados, quando os valores estão dispostos em ordem crescente ou decrescente.

Moda — é o valor mais comum (o que mais se repete) em uma série de dados.

Média, mediana e moda são todas medidas válidas de tendência central, mas sob diferentes condições. A escolha da medida de tendência central varia de acordo com o uso e necessidade. Nas próximas seções, veremos cada uma delas, aprenderemos como calculá-las e em quais condições utilizá-las.

MÉDIA ARITMÉTICA

A média aritmética é a mais popular, e muitas vezes é simplesmente chamada de “média”.

Para calculá-la, some os valores de todos os termos e depois divida pelo número de termos. Exemplo:

Qual é a média de 2, 4, 6, 8 e 10? Leonardo Carvalho
leofacebook17@gmail.com

Solução:

Primeiro, some todos os números.

$$2 + 4 + 6 + 8 + 10 = 30$$

Agora, divida por 5 (número total de observações).

$$\text{Média} = 30/5 = 6$$

MÉDIA PONDERADA

A média ponderada é calculada quando determinados valores fornecidos em um conjunto de dados são mais “importantes” que os outros (possuem maior peso ou maior influência no resultado).

A fórmula é escrita como sendo:

$$W = \frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i}$$

Em que W é a média ponderada e w_i é o peso dado àquele conjunto de dados.

Exemplo 1. Na escola, a média anual de cada matéria é calculada de acordo com os princípios da média ponderada. Considerando que o peso das notas esteja relacionado com o bimestre em questão, determine a média anual de sabendo que as notas em Matemática foram iguais a:

1º Bimestre: Nota 6- peso 1

2º Bimestre: Nota 7 - peso 2

3º Bimestre: Nota 8 - peso 3

4º Bimestre: Nota 9 - peso 4

$$W = \frac{6*1 + 7*2 + 8*3 + 9*4}{1+2+3+4} = \frac{80}{10} = 8$$

leofacebook17@gmail.com

MÉDIA GEOMÉTRICA/HARMÔNICA

A média geométrica é definida como a raiz n -ésima (n-ésima; de grau n) do produto de cada valor, ou seja, de n números no conjunto de dados fornecido.

A média geométrica pode ser aplicada em qualquer conjunto de dados estatístico, mas normalmente ela é empregada na geometria. Há também aplicação em problemas da matemática financeira que envolvam taxa percentual acumulada, ou seja, porcentagem sob porcentagem. Além de ser a média mais conveniente para dados que se comportam como uma progressão geométrica. Sua fórmula é expressa como:

$$\left(\prod_{i=1}^n x_i \right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \cdots x_n}$$

Essa fórmula pode ser bastante assustadora, mas vamos dar um exemplo para ficar mais claro.

Exemplo 1. Qual o valor da média geométrica entre os números 3, 8 e 9?

Como temos 3 valores, iremos calcular a raiz cúbica do produto.

$$M_G = \sqrt[3]{3 \cdot 8 \cdot 9} = \sqrt[3]{216} = 6$$

Simple, né? Agora vamos dar um exemplo do mundo financeiro

Exemplo 2: Um investimento rende no primeiro ano 5%, no segundo ano 7% e no terceiro ano 6%. Qual o rendimento médio desse investimento?

Para resolver esse problema devemos encontrar os fatores de crescimento.

- 1.º ano: rendimento de 5% → fator de crescimento de 1,05 (100% + 5% = 105%)
- 2.º ano: rendimento de 7% → fator de crescimento de 1,07 (100% + 7% = 107%)
- 3.º ano: rendimento de 6% → fator de crescimento de 1,06 (100% + 6% = 106%)

$$M_G = \sqrt[3]{1,05 \cdot 1,06 \cdot 1,07} = \sqrt[3]{1,19091} = 1,05996$$

Para encontrar o rendimento médio devemos fazer:

$$1,05996 - 1 = 0,05996$$

Assim, o rendimento médio dessa aplicação, no período considerado, foi de aproximadamente 6%.

Notem que a média geométrica penaliza valores muito baixos. É o caso de quando temos um zero em uma das medidas. Multiplicando qualquer coisa por zero, a média geométrica será zero também (você já devem ter tido aquele querido professor que faz a média do semestre se baseando em médias geométricas, certo?).

Na prática, tirando o mundo financeiro, a média geométrica é pouco usada.

MEDIANA

A mediana, em estatística, é o valor médio da lista de dados fornecida, quando organizados em uma ordem. A disposição dos dados ou observações pode ser feita em ordem crescente ou decrescente. Exemplo: A mediana de {2, 3, 4} é 3. Em matemática, a mediana também é um tipo de média, que é usada para encontrar o valor do centro (é encontrado ordenando-se todos os dados e escolhendo o que está no centro). Para um conjunto de dados, pode ser considerado como o valor "intermediário".

A mediana é menos afetada por discrepâncias e dados distorcidos. A característica básica da mediana na descrição de dados — em comparação com a média — é que ela não é distorcida por uma pequena proporção de valores extremamente grandes ou pequenos e, portanto, fornece uma melhor representação de um valor "típico".

A renda mediana, por exemplo, pode ser uma maneira melhor de sugerir o que é uma renda "típica", pois a distribuição de renda geralmente é muito distorcida, especialmente no Brasil (poucos ganham MUITO).

A mediana é de importância central em estatísticas robustas, pois é a estatística mais resistente, tendo um ponto de ruptura de 50%; a mediana não é um resultado arbitrariamente grande ou pequeno, desde que não mais da metade dos dados estejam contaminados.

Para determinar o valor mediano em uma sequência de números, esses números devem primeiro organizados, em ordem de valor do menor para o maior (mais comum) ou do maior para o menor.

Se houver uma quantidade ímpar de números, o valor mediano é o número que está no meio, com a mesma quantidade de números abaixo e acima. Se houver uma quantidade par de números na sequência, o par do meio deve ser determinado, somado e dividido por dois para que se encontre o valor mediano.

Sua fórmula ~~assustadora~~ é dada por:

$$\text{Med}(X) = \begin{cases} X[\frac{n}{2}] & \text{if } n \text{ is even} \\ \frac{(X[\frac{n-1}{2}] + X[\frac{n+1}{2}])}{2} & \text{if } n \text{ is odd} \end{cases}$$

Vamos a alguns exemplos.

Exemplo 1. Qual a mediana das alturas dos jogadores de um time de vôlei onde as alturas são: 1,97m; 1,87m; 1,99m; 2,01m; 1,83m?

Organizando os valores em ordem crescente:

1,83m; 1,87m; **1,97m**; 1,99m; 2,01m;

Verificamos que a quantidade de dados é ímpar. A mediana é, portanto, o valor do meio depois da ordenação. Logo, a mediana é **1,97m**

Exemplo 2. Calcule o valor da mediana da seguinte amostra de dados: (32, 27, 15, 44, 15, 32).

Primeiro precisamos colocar os dados em ordem, assim temos:

15, 15, 27, 32, 32, 44

Como essa amostra é formada por 6 elementos, que é um número par, a mediana será igual a média dos elementos centrais, ou seja:

$$M_d = \frac{27 + 32}{2} = \frac{59}{2} = 29,5$$

A mediana às vezes é usada em oposição à média, quando há valores discrepantes na sequência que podem distorcer o resultado. A mediana de uma sequência é menos afetada por **outliers** do que a média.

MODA

A moda é o valor que aparece com mais frequência em um conjunto de valores de dados.

Assim como a média e a mediana, a moda é uma forma de expressar, em um número (geralmente) único, informações importantes sobre uma variável aleatória ou uma população.

Por exemplo, na sequência abaixo, 16 é a moda, pois aparece mais vezes no conjunto do que qualquer outro número:

3, 3, 6, 9, 16, 16, 16, 27, 27, 37, 48

Um conjunto de números pode ter mais de uma moda (isso é conhecido como bimodal — quando houver duas modas) se houver vários números que ocorram com igual frequência e mais vezes do que os outros no conjunto.

3, 3, 3, 9, 16, 16, 16, 27, 37, 48

No exemplo acima, tanto o número 3 quanto o número 16 são modas, pois cada um ocorre três vezes e nenhum outro número ocorre tão frequentemente.

Se nenhum número em um conjunto ocorrer mais de uma vez, esse conjunto não terá moda, como o exemplo abaixo::

3, 6, 9, 16, 27, 37, 48

Um conjunto de números com duas modas é chamado de bimodal; um conjunto de números com três modas é trimodal; e qualquer conjunto de números com mais de uma moda é multimodal.

A moda é a medida de tendência mais aplicada para séries de dados categóricos (“qualitativos”, que não possuem um ordenamento natural ou relações de grandeza entre si).

MEDIDAS DE DISPERSÃO

O resumo de um conjunto de dados por uma única medida representativa de posição central esconde toda a informação sobre a dispersão do conjunto de observações. Por exemplo, suponhamos que cinco grupos de alunos submeteram-se a um teste, obtendo-se as seguintes notas:

grupo A (variável X): 3, 4, 5, 6, 7

grupo B (variável Y): 1, 3, 5, 7, 9

grupo C (variável Z): 5, 5, 5, 5, 5

grupo D (variável W): 3, 5, 5, 7

grupo E (variável V): 3, 5, 5, 6, 6

Vemos que a média em cada um dos grupos é 5,0. A identificação de cada uma destas séries por sua média (5, em todos os casos) nada informa sobre suas diferentes **variabilidades**. Notamos, então, a conveniência de serem criadas medidas que resumam quão dispersos são os dados de um conjunto de observações e que nos permita, por exemplo, comparar conjuntos diferentes de valores, como os dados acima, segundo algum critério estabelecido.

A **variabilidade/dispersão** é considerada significativa quando a variação ou falta de uniformidade no tamanho dos itens de uma série é grande. Se a

variabilidade for menor, a dispersão é menos significativa. Se todos os dados forem idênticos, a dispersão é nula.

ALCANCE

A medida mais simples de **dispersão absoluta** é o **alcance**. Ele é apenas o maior ponto de dados menos o menor. Podemos escrever isso como $R = H - L$, sendo H o valor máximo encontrado na amostra e L o valor mínimo. Curiosidade: a letra R foi usada para referenciar a palavra inglesa "range", que pode ser usada para definir a amplitude dos dados.

Por exemplo, se as notas dos estudantes de matemática em uma prova é dada por:

Aluno A: 2 pontos

Aluno B: 2 pontos

Aluno C: 4 pontos

Aluno D: 5 pontos

Aluno E: 8 pontos

Aluno F: 8 pontos

Aluno G: 9 pontos

Leonardo Carvalho
leofacebook17@gmail.com

O alcance nesse caso seria $9 - 2 = 7$ pontos.

A medida de alcance pode ser interessante para entendermos, nesse caso, se o aluno com pior nota está em um "patamar" muito diferente do aluno com melhor nota. Nesse caso temos uma diferença de 7 pontos entre eles, o que é grande considerando que as notas variam de 0 a 10.

DESVIOS E VARIÂNCIA

Um critério frequentemente usado para entender a dispersão de um conjunto é aquele que mede a dispersão dos dados **em torno de sua média** - ou seja, quanto cada um dos dados difere da média do seu grupo. Vamos voltar ao seguinte exemplo de notas um teste de grupos de pessoas:

grupo A (variável X): 3, 4, 5, 6, 7
grupo B (variável Y): 1, 3, 5, 7, 9
grupo C (variável Z): 5, 5, 5, 5, 5
grupo D (variável W): 3, 5, 5, 7
grupo E (variável V): 3, 5, 5, 6, 6

A média de cada grupo, como já calculamos antes, é 5. Olhando para o grupo C, vemos que os integrantes são muito homogêneos - ou seja, todos ali tiraram 5. Se observarmos por outro lado o grupo B, vemos que temos pessoas que tiraram uma nota super alta (nota 9) e alguém que não estudou e tirou 1 - ou seja, números que estão bem distantes da média do grupo como um todo.

Para analisarmos quão distantes cada um dos pontos está da média e compilamos isso em um único valor, duas medidas são as mais usadas:

desvios e variância.

Leonardo Carvalho
leofacebook17@gmail.com

Para o grupo A acima os desvios de cada ponto em relação a média (x_i - média) são: -2, -1, 0, 1, 2. Se somarmos cada um desses elementos, teríamos que a soma dos desvios é igual a zero, que teoricamente indicaria que não há nenhum desvio na série. Mas isso não é verdade, certo? Vemos que os elementos não têm o mesmo valor! Isso acontece pois valores desviando para menos (valores menores que a média) se tornam negativos, enquanto valores desviando para mais (valores maiores que a média) tem sinal positivo. Somando-os, anulamos os efeitos e ficamos com uma ideia errada do desvio da série.

Para resolver esse problema temos duas opções: (a) considerar o total dos desvios em valor absoluto (módulo, ignorando o sinal); (b) considerar o total dos quadrados dos desvios, o que faria com que o sinal negativo se tornasse positivo. Para o grupo A teríamos, respectivamente,

$$\sum_{i=1}^5 |x_i - \bar{x}| = 2 + 1 + 0 + 1 + 2 = 6,$$

$$\sum_{i=1}^5 (x_i - \bar{x})^2 = 4 + 1 + 0 + 1 + 4 = 10.$$

O uso desses totais pode causar dificuldades quando comparamos conjuntos de dados com números diferentes de observações, como os conjuntos A e D acima. Desse modo, é mais conveniente exprimir as medidas como médias (dividindo pela quantidade de dados), isto é, o **desvio médio** e a **variância** são definidos por

Desvio médio

$$dm(X) = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

Leonardo Carvalho
leofacebook17@gmail.com

Variância

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

Em que x_i é a representação de cada dado, \bar{x} é a média de todos os dados juntos e n é o tamanho da amostra.

Para o grupo A temos

$$dm(X) = 6/5 = 1,2,$$

$$var(X) = 10/5 = 2,0$$

Para o grupo D temos

$$\begin{aligned} \text{dm}(W) &= 4/4 = 1,0, \\ \text{var}(W) &= 8/4 = 2,0. \end{aligned}$$

Podemos dizer, então, que, de acordo com o desvio médio, o grupo D é mais homogêneo que A, enquanto ambos são igualmente homogêneos, segundo a variância.

A variância é uma medida de dimensão igual ao quadrado da dimensão dos dados (por exemplo, se os dados são expressos em cm , a variância será expressa em cm^2), pode causar problemas de interpretação. Costuma-se usar, então, o **desvio padrão**, que é definido como a raiz quadrada positiva da variância.

$$\text{dp}(X) = \sqrt{\text{var}(X)}$$

Leonardo Carvalho
leofacebook17@gmail.com

Para o grupo A o desvio padrão é

$$\text{dp}(X) = \sqrt{\text{var}(X)} = \sqrt{2} = 1,41$$

Uma interpretação mais intuitiva para o desvio padrão é entendê-lo com um uma medida de erro caso aproximarmos (ou estimarmos) todos os valores da série dados pela sua média. Quanto erraríamos se representássemos toda nossa série de dados pela média.

Por ser calculado com os quadrados dos desvios da média, tanto erros positivos quanto erros negativos contribuem para o valor final da medida.

Algumas observações sobre os desvios.

- O desvio padrão mede a variação dos dados com relação à média e tem a mesma unidade de medida que o conjunto de dados.
- Os desvios são sempre maior ou igual a 0. Quando $s = 0$, o conjunto de dados não apresenta variação (todos os elementos têm o mesmo valor).

- À medida que os valores se afastam da média (isto é, estão mais dispersos), o valor do desvio aumenta.

Veremos mais tarde que a variância de uma amostra será calculada usando-se o denominador $n - 1$, em vez de n . Para grandes amostras isso fará pouquíssima diferença.

NOTA IMPORTANTE: TANTO A VARIÂNCIA COMO O DESVIO MÉDIO SÃO MEDIDAS DE DISPERSÃO CALCULADAS EM RELAÇÃO À MÉDIA DAS OBSERVAÇÕES. ASSIM COMO A MÉDIA, A VARIÂNCIA (OU O DESVIO PADRÃO) SÃO AFETADOS EXAGERADAMENTE SE EXISTEM OUTLIERS EM NOSSA SÉRIE. OU SEJA, NOSSO DESVIO AUMENTA MUITO SE TEMOS APENAS 1 VALOR DISCREPANTE.

COEFICIENTE DE VARIAÇÃO

O **coeficiente de variação (CV)**, a medida análoga de dispersão relativa, é apenas o desvio padrão dividido pela média aritmética. Para dar como uma porcentagem em vez de uma proporção, multiplique por 100%. A fórmula do CV pode ser vista abaixo:

$$CV = \frac{\sigma}{\mu}$$

Em que μ é a média de todos os dados e σ é o desvio-padrão.

O coeficiente de variação é muito interessante pois ele não está atrelado a unidade da medida e não precisamos ter ideia da grandeza da média para entender se o valor é grande ou pequeno.

Por exemplo, vamos supor que coletamos dados de salário mensal de um grupo de pessoas. Se eu te disser que o desvio-padrão é R\$ 500 você não tem ideia se esse desvio é grande ou pequeno, certo? Se a média for R\$ 1.000 reais, esse desvio parece enorme. Agora se a média for R\$ 100.000, esse desvio é bem menor. Logo, para entender o desvio-padrão, precisamos sempre olhar a média e fazer algumas continhas na nossa cabeça para entender quão grande é o desvio.

Com o coeficiente de variação, não precisamos olhar dois números separadamente e fazer continhas na nossa cabeça.

- Coeficiente de variação para média = R\$ 100.000 e desvio-padrão = R\$ 500

$$CV = R\$ 500 / R\$ 100.000 = 0,005 = 0,5\%$$

- Coeficiente de variação para média = R\$ 1.000 e desvio-padrão = R\$ 1.000

$$CV = R\$ 500 / R\$ 1.000 = 0,5 = 50\%$$

Percebem que só olhando para o CV já conseguimos entender que os dados são muito mais discrepantes no segundo caso?

O CV também é muito útil quando queremos entender qual distribuição tem maior variabilidade mas não estamos comparando coisas comparáveis. Por exemplo, se quisermos comparar uma distribuição de salário e uma distribuição de altura, como são medidas diferentes (uma em reais e outra em metros) não podemos comparar nem a média nem o desvio. Por outro lado, como o CV não possui unidade, podemos comparar os CVs dessas duas distribuições e entender qual delas é mais dispersa.

COMBINAÇÃO DE VARIÁVEIS

Podemos formar novas distribuições combinando variáveis aleatórias. Se soubermos a média e o desvio-padrão das distribuições originais, podemos usar essas informações para calcular a média e o desvio-padrão da distribuição resultante.

Podemos combinar médias diretamente, mas não podemos fazer isso com desvios-padrão. Podemos combinar variâncias, desde que seja razoável assumir que as variáveis sejam independentes.



| | Média | Variância |
|-------------------------|-------------------------|--|
| Somando: $T = X + Y$ | $\mu_T = \mu_X + \mu_Y$ | $\sigma_T^2 = \sigma_X^2 + \sigma_Y^2$ |
| Subtraindo: $D = X - Y$ | $\mu_D = \mu_X - \mu_Y$ | $\sigma_D^2 = \sigma_X^2 + \sigma_Y^2$ |

Aqui estão alguns fatos importantes sobre a combinação de variâncias:

- Certifique-se de que as variáveis sejam independentes ou que seja razoável assumir independência antes de combinar variâncias.
- Mesmo quando estamos subtraindo duas variáveis aleatórias, ainda somamos suas variâncias; subtrair duas variáveis aumenta a variabilidade total dos resultados.
- Podemos calcular o desvio-padrão de distribuições combinadas tirando a raiz quadrada das variâncias combinadas.

Vamos a um exemplo. Aproximadamente 1,7 milhões de estudantes fizeram a prova SAT em 2015. Cada estudante recebeu uma nota em análise crítica e em matemática. Aqui está o resumo das estatísticas para cada parte do teste em 2015:

| Matéria | Média | Desvio-padrão |
|-----------------|------------------|---------------------|
| Análise crítica | $\mu_{CR} = 495$ | $\sigma_{CR} = 116$ |
| Matemática | $\mu_M = 511$ | $\sigma_M = 120$ |
| Total | $\mu_T = ?$ | $\sigma_T = ?$ |

Considere que a nota final deve ser composta pela média de ambas as matérias - matemática e análise crítica. Considerando essas duas matérias,

qual é a média e o desvio-padrão que representaria a **nota geral (soma das duas notas)** dos alunos?

A média da soma é simplesmente a soma das médias. Portanto, a média da nota geral é $511 + 495 = 1006$

Já o desvio-padrão não pode ser somado diretamente. Para fazermos isso, precisamos somar as variâncias e, então, tirar a raiz quadrada.

$$s_p^2 = s_1^2 + s_2^2$$

Em que s_p^2 é a variância das notas somadas, s_1^2 é a variância do conjunto 1, s_2^2 é a variância do conjunto 2, n_1 é o tamanho do conjunto 1 e n_2 é o tamanho do conjunto 2.

Aplicando a raiz quadrada nos dados, temos o que chamamos de **desvio padrão agrupado** (ou seja, o desvio padrão de 2 amostras).

$$s_p = \sqrt{116^2 + 120^2} = 166.9$$

MEDIDAS SEPARATRIZES: QUARTIS E PERCENTIS

Tanto a média como o desvio padrão podem não ser medidas completas para representar um conjunto de dados, pois, além de serem afetados exageradamente por valores extremos, não conseguimos ter uma ideia da simetria da distribuição. Para contornar isso, usamos as medidas separatrizes.

As medidas separatrizes são usadas em estatística para dividir o número total de observações de uma distribuição em certo número de partes iguais. As mais comumente usadas são: Quartis e Percentis. **É importante observar**

que os dados devem ser classificados em ordem crescente (uso mais comum) ou decrescente antes de calcular os valores da partição.

Os Quartis dividem os dados em quatro partes iguais; os Decis os dividem em dez partes iguais; e os Percentis os dividem em cem partes iguais. Esses valores de separatrizes são usados para fragmentar uma distribuição em partes menores, tornando-as mais fáceis de medir, analisar e entender.

Os **quartis** dividem um conjunto de dados em **quatro** partes iguais. São três quartis que dividem todos os dados, com um quarto dos valores de dados em cada parte: Primeiro Quartil (Q1), Segundo Quartil (Q2) e Terceiro Quartil (Q3). O percentil refere-se ao percentual de dados totais acumulados em uma determinada porção. Por exemplo, o percentil 25 nos diz que 25% dos dados estão concentrados ali.

Os Q1, Q2 e Q3 também são chamados de quartil inferior, quartil médio (ou mediana) e quartil superior, respectivamente.

Leonardo Carvalho

leofcarvalho17@gmail.com

Vamos entender o conceito de cada quartil e percentil:

1. O **primeiro quartil (Q1)** separa a primeira parte de um quarto ($1/4$) dos dados da parte superior de três quartos ($3/4$), ou seja, 25% dos dados ficarão abaixo de Q1 e 75% ficarão acima dele. Esse conjunto também é chamado de **percentil 25** (25% dos dados concentrados nesse conjunto)
2. O segundo quartil (Q2) divide os dados em duas partes iguais. Ele separa a primeira metade dos dados da segunda metade, ou seja, 50% dos dados estão abaixo dele e os 50% restantes estão acima dele. O segundo quartil também é chamado de **mediana** dos dados ou de **percentil 50**.
3. O **terceiro quartil (Q3)** separa as três primeiras partes dos dados da última, ou seja, 75% dos dados ficarão abaixo dele e 25% acima dele. O Q3 também é chamado de **percentil 75**.

Suponha que tenhamos os seguintes valores de uma variável X: 15, 5, 3, 8, 10, 2, 7, 11, 12. Aqui temos uma sequência de 9 valores



Ordenando os valores, obtemos as estatísticas de ordem $x_1 = 2, x_2 = 3, \dots, x_9 = 15$, ou seja, teremos $2 < 3 < 5 < 7 < 8 < 10 < 11 < 12 < 15$.

Usando a definição de mediana dada, teremos que mediana = $Q_2 = 8$. Ou seja, 50% dos dados assumem valor até 8.

Suponha que queiramos calcular os dois outros quartis, Q_1 e Q_3 . A ideia é dividir os dados em quatro partes: 2 3 5 7 8 10 11 12 15, em que o número 8 destacado é nossa mediana (número que divide dos dados no meio).

Uma possibilidade razoável para obter o Q_1 é considerar a mediana dos primeiros quatro valores para , ou seja

$$q_1 = \frac{3 + 5}{2} = 4,$$

Leonardo Carvalho
leofacebook17@gmail.com

Portanto, dizemos que 25% dos dados estão concentrados até o valor 4.

E a mediana dos últimos quatro valores para obter Q_3 , ou seja,

$$q_3 = \frac{11 + 12}{2} = 11,5.$$

Portanto, dizemos que 75% dos dados assumem o valor de até 11.5.

Vamos observar agora uma outra sequência X: 15, 5, 3, 8, 10, 2, 7, 11, 12, 67

Ordenando os valores, obtemos as estatísticas de ordem $x_1 = 2, x_2 = 3, \dots, x_{10} = 67$, ou seja, teremos $2 < 3 < 5 < 7 < 8 < 10 < 11 < 12 < 15 < 67$.

Usando a definição de mediana dada, teremos que mediana = $Q_2 = (8 + 10)/2 = 9$

Um raciocínio similar ao anterior será aplicado. Porém, agora como temos 2 dados centrais (tivemos que fazer a média para encontrar a mediana), podemos abstrair os valores 8 e 10 para formar um único valor chamado de mediana. Nesse caso, teríamos:

$$2 < 3 < 5 < 7 < 9 < 11 < 12 < 15 < 67$$

Da mesma forma que o anterior Q1 será a mediana a partir do valor 2 até 9, ou seja, o valor central nesses dados.

$$2 < 3 < 5 < 7 < 9$$

Nesse conjunto, o número 5 é o valor central. Portanto, Q1 = 5. Ou seja, 25% dos dados assumem até o valor 5.

Analogamente, Q3 será a mediana a partir de 9 até o valor 67.

$$9 < 11 < 12 < 15 < 67$$

Logo, no conjunto 15, 5, 3, 8, 10, 2, 7, 11, 12, 67, o Q3 é 12.

RESISTÊNCIA DOS DADOS

Leonardo Carvalho
leofacebook17@gmail.com

Dizemos que uma medida de localização ou dispersão é **resistente** quando for pouco afetada por mudanças de uma pequena porção dos dados. A mediana é uma medida resistente, ao passo que a média não o é.

Para ilustrar este fato, considere as populações dos 30 municípios do Brasil. Se descartarmos Rio de Janeiro e São Paulo, a média das populações dos 28 municípios restantes é 100,6 e a mediana é 82,1. Para todos os dados, a média passa a ser 145,4, ao passo que a mediana será 84,3.

Note que a média aumentou bastante, influenciada que foi pelos dois valores maiores, que são muito discrepantes da maioria dos dados. Mas a mediana variou pouco. O desvio padrão também não é uma medida resistente.

Os valores de quartis/percentis também são bastante resistentes e com eles pode-se ter uma boa ideia da simetria da distribuição dos dados.

SIMETRIA



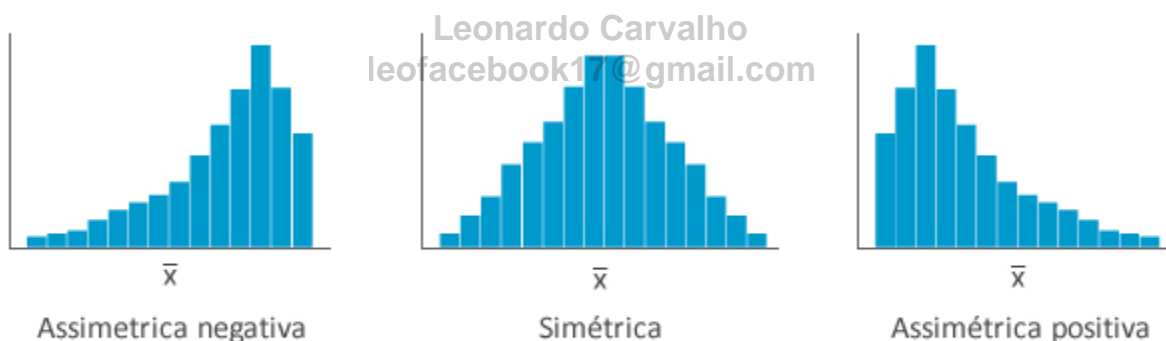
Devido a resistência dos dados, em vários casos as medidas descritas acima podem não ser suficientes e, então, precisamos de medidas de simetria.

Quando uma distribuição é **simétrica** significa que as observações estão igualmente distribuídas em torno da média (metade acima e metade abaixo).

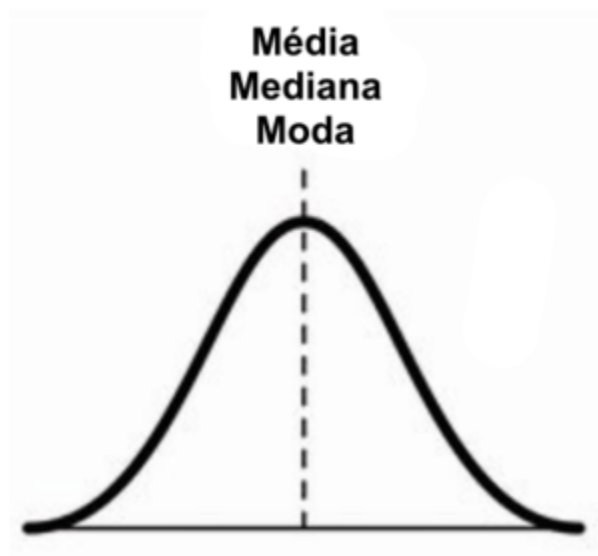
A assimetria de uma distribuição pode ocorrer de duas formas:

- quando os valores concentram-se à esquerda (assimetria com concentração à esquerda ou assimetria com cauda à direita);
- quando os valores concentram-se à direita (assimetria com concentração à direita ou com assimetria cauda à esquerda);

Ao definir a assimetria de uma distribuição, algumas pessoas preferem se referir ao lado onde está a concentração dos dados. Porém, outras pessoas preferem se referir ao lado onde está faltando dados (cauda). As duas denominações são alternativas.



De maneira geral, encontramos uma perfeita simetria na distribuição normal (que vamos falar mais a frente), em que os dados estão mais concentrados em um ponto central e quanto mais distante da média, menor é a frequência dos dados. Ao traçarmos uma linha no meio da curva teremos dois lados espelhados. Se você tiver acesso às demais medidas descritivas, irá verificar que a média, a mediana e a moda também são iguais.



Leonardo Carvalho
leofacebook17@gmail.com

E por que é importante sabermos se os dados são simétricos ou não?

Suponha que você faça parte do time de treinamento de uma empresa de telemarketing que está tentando entender a performance geral dos atendentes para dar treinamentos direcionados. Se você se basear apenas em média ou desvio-padrão, pode perder informações valiosas.

Por exemplo, se você plotar o gráfico de performance para os atendentes na forma de histograma e obtiver uma curva assimétrica negativa, poderá constatar que há poucos atendentes que possuem uma baixa performance quando comparado com a maioria e que provavelmente estão causando grandes desvios e baixas médias de performance geral. Uma possível solução de business para isso seria dar treinamentos para essas pessoas específicas ao invés de fazê-lo para o time inteiro.

Por outro lado, se você plotar o gráfico performance e constatar uma assimetria positiva, verá que a maioria dos seus atendentes tem uma performance baixa. Com isso, poderá pensar em medidas de melhoria de processo ou até entender se aqueles com performance muito elevada estão

trapaceando o sistema de alguma forma (isso infelizmente acontece e como conhecedores de dados precisamos analisá-los de forma completa).

Mas será que precisamos sempre olhar os gráficos para entender isso? Nem sempre! Existem métricas específicas que traduzem assimetrias de uma forma numérica. Isso é muito útil especialmente se você tiver tantas métricas de performance e tantos times que olhar gráficos um a um seria quase inviável.

SKEW (ASSIMETRIA)

Uma das formas adotadas para o cálculo da assimetria é através do **coeficiente de assimetria de Fisher**. O coeficiente vem a partir do terceiro momento de ordem superior em torno da média através de uma função geradora de momentos. Momentos são medidas resumo de uma distribuição, sendo 1º momento = média (valor esperado), 2º momento = variância, 3º momento = assimetria e 4º momento = curtose.

Leonardo Carvalho
leofacebook17@gmail.com

O Skew (assimetria) é medido por:

$$\text{Skew} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

A interpretação, será:

- Skew = 0, a distribuição é simétrica;
- Skew > 0, a distribuição é assimétrica positiva (à direita);
- Skew < 0, a distribuição é assimétrica negativa (à esquerda).

Dizemos que os dados são aproximadamente normais (e, portanto, simétricos) se $-1 < \text{Skew} < 1$.

Vamos supor que para o caso dos atendentes de telemarketing o skew = 2,78. Podemos fazer a seguinte interpretação: o sinal positivo significa que a distribuição é assimétrica à direita e como 2,78 é maior que o intervalo de referência ($-1 < \text{Skew} < 1$), os dados apresentam alto grau de assimetria.

NOTA IMPORTANTE: *PODEMOS USAR A FUNÇÃO SKEW PARA CALCULAR NO GOOGLE SHEETS*

CURTOSE

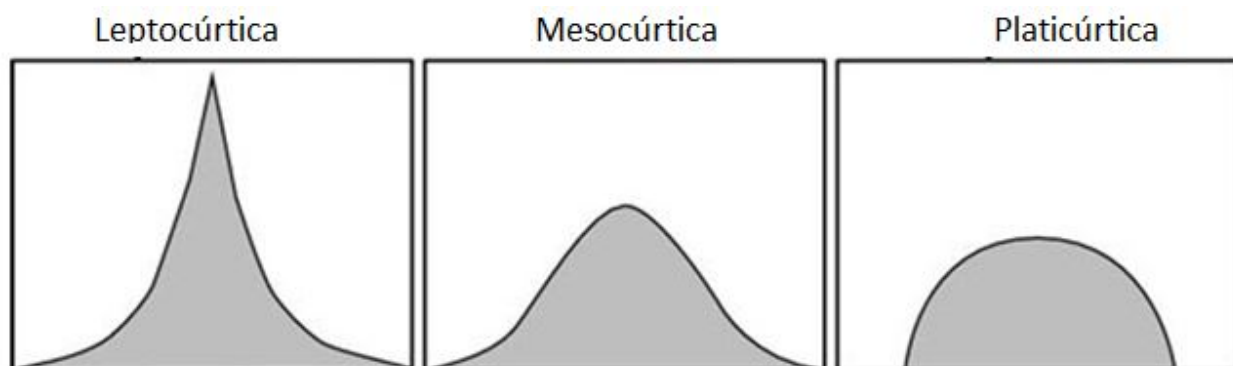
A curtose (kurtosis em inglês) representa o grau de achatamento da distribuição, isto é, quão espalhados os dados estão em torno da média. Novamente, usamos a curva normal padrão como referência e podemos interpretar a curtose por meio de gráficos ou numericamente. Pode ser classificada em três tipos:

a) Mesocúrtica: que é própria curva normal padrão

Leonardo Carvalho
leofacebook17@gmail.com

b) Platicúrtica: possui grau de achatamento maior que da curva normal padrão, o que nos indica que os dados estão mais espalhados (logo, o desvio padrão também é maior).

c) Leptocúrtica: seu grau de achatamento é menor que o da curva normal padrão (curva mais pontiaguda), indica que os dados estão mais concentrados (desvio padrão menor)



A curtose pode ser calculada pelo coeficiente de curtose de Fisher, que neste caso utiliza o quarto momento de ordem superior ao redor da média:

$$\text{Kurtosis} = \left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_j - \bar{x}}{s} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Assim, se:

- Curtose = 0, a curva é normal padrão, isto é, mesocúrtica
- Curtose > 0, grau de achatamento baixo, a curva é leptocúrtica
- Curtose < 0, grau de achatamento alto, a curva é platicúrtica

Em alguns programas estatísticos, como o STATA, é comum encontrar a curtose da distribuição normal como $K = 3$. Neste caso, a interpretação é a mesma. Isto é:

Leonardo Carvalho
leofacebook17@gmail.com

- $K = 3$, curva normal padrão
- $K > 3$, curva leptocúrtica
- $K < 3$, curva platicúrtica

Como interpretar na prática? Vamos supor que para o caso dos atendentes de telemarketing encontramos curtose = 10,82, o que nos indica que a curva é leptocúrtica, isto é, é menos achatada que a curva normal – o pico da distribuição é mais acentuado – então sabemos que os dados estão mais concentrados em um determinado ponto.

NOTA IMPORTANTE: PODEMOS USAR A FUNÇÃO KURT PARA CALCULAR NO GOOGLE SHEETS

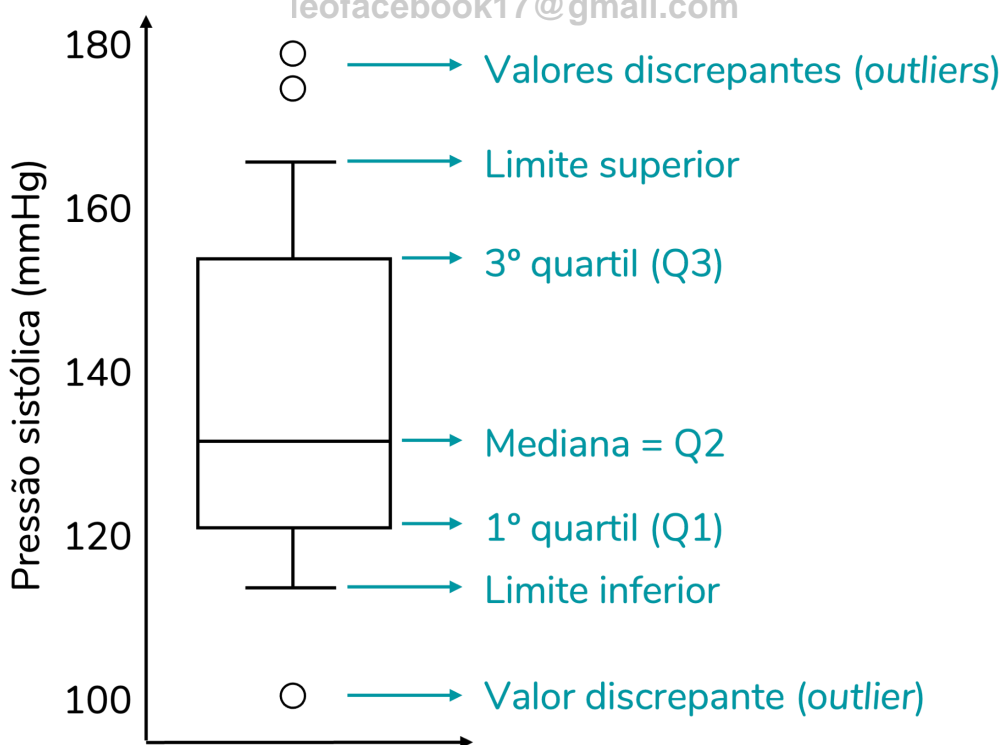
BOXPLOTS

Outra forma muito interessante para entendermos a distribuição e simetria dos dados é através dos quartis e boxplots/. Vamos supor uma distribuição com vários números. Suponhamos que ordenamos de forma crescente esses números, da mesma forma que fizemos no exemplo dos quartis. O número de menor valor será chamado de x_1 , o número de maior valor será chamado de x_n e os quartis serão chamados de q_1 , q_2 , q_3 . Consideramos que os **dados são simétricos** quando:

- a) $q_2 - x_1 = x_n - q_2$;
- b) $q_2 - q_1 = q_3 - q_2$;
- c) $q_1 - x_1 = x_n - q_3$;

A diferença $q_2 - x_1$ é chamada **dispersão inferior** e $x_n - q_2$ é a **dispersão superior**. A condição (a) nos diz que estas duas dispersões devem ser aproximadamente iguais, para uma distribuição aproximadamente simétrica.

Podemos representar graficamente essas informações de quartis em um boxplot. No exemplo abaixo coletamos a pressão sistólica de 100 pessoas e plotamos os dados em um boxplot.



Notem que valores discrepantes (outliers) são representados por bolinhas e eles não entram na conta dos quartis. **O "limite inferior" para o boxplot não é o valor mínimo que temos nesse conjunto para os casos em que o valor mínimo do nosso conjunto seja um outlier.** Isso **não** significa que, caso queiramos escrever as estatísticas como mínimo, máximo, média, mediana e quartis, nós não consideramos esse outlier. A descrição das estatísticas do conjunto sempre vai considerar todos os valores, porém a representação no boxplot tem essa particularidade.

E como esse outlier é calculado? A partir dos cálculos de limite inferior e superior.

- Limite Inferior = Primeiro Quartil - $1,5 * (\text{Terceiro Quartil} - \text{Primeiro Quartil})$
- Limite Superior = Terceiro Quartil + $1,5 * (\text{Terceiro Quartil} - \text{Primeiro Quartil})$

Leonardo Carvalho

Caso haja um dado menor do que o limite inferior ou maior que o limite superior, o boxplot considera isso um outlier.

Para exemplificar, vamos considerar que coletamos as notas de 12 alunos de um curso. Ordenamos essas idades em ordem crescente para facilitar a

explicação.

| Posição | Idade |
|-----------------|-------|
| 1 ^a | 18 |
| 2 ^a | 19 |
| 3 ^a | 21 |
| 4 ^a | 21 |
| 5 ^a | 21 |
| 6 ^a | 22 |
| 7 ^a | 22 |
| 8 ^a | 22 |
| 9 ^a | 23 |
| 10 ^a | 23 |
| 11 ^a | 24 |
| 12 ^a | 27 |

Leonardo Carvalho
leofacebook17@gmail.com

O primeiro passo é calcular os quartis/percentis. Para encontrar o percentil 25, primeiramente precisamos encontrar em qual posição devemos buscar o valor. Podemos chegar a essa posição, multiplicando o percentil que queremos pelo tamanho da amostra e dividindo por 100.

Posição do Percentil 25 = Percentil * Tamanho da Amostra / 100 = $25 * 12 / 100 = 300 / 100 = 3$

Na posição 3, temos a idade de 21 anos. Sendo assim, o percentil 25 dessa amostra é 21 anos. Isso significa que pelo menos 25% dos indivíduos dessa amostra têm no máximo 21 anos.

E se o cálculo da posição de determinado percentil não resultar em um número inteiro? Podemos tirar a média dos dois valores intermediários, como mostramos no exemplo dos quartis. Lembre-se, dificilmente faremos esses

cálculos "na mão" pois usaremos ferramentas como Excel e Python. Porém, para que haja um entendimento completo dos conceitos, é importante que vocês vejam como o cálculo é realizado.

Fazendo os cálculos dos outros quartis temos:

| Variável | Mínimo | 1º Quartil | 2º Quartil | 3º Quartil | Máximo |
|----------|--------|------------|------------|------------|--------|
| Idade | 18 | 21 | 22 | 23 | 27 |

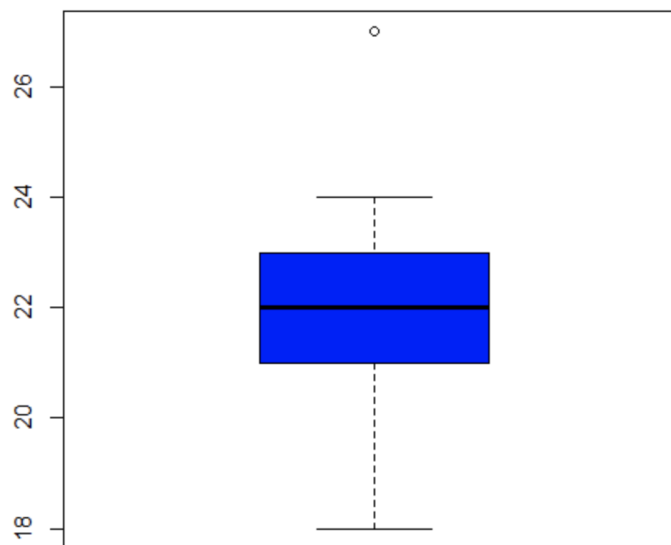
Agora calculando os limites para o boxplot:

Limite inferior = Primeiro Quartil - $1,5 * (\text{Terceiro Quartil} - \text{Primeiro Quartil}) = 21 - 1,5*(23-21) = 18$

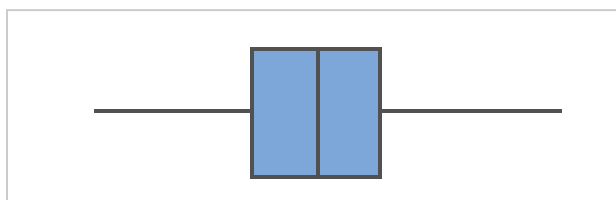
Limite superior = Terceiro Quartil + $1,5 * (\text{Terceiro Quartil} - \text{Primeiro Quartil}) = 23 + 1,5*(23-21) = 26$

Vemos aqui que não temos outliers inferiores (nossa idade mínima é 18, que coincide com o limite inferior. Porém, por outro lado, o limite superior é 26, e a idade máxima é 27. Ou seja, teremos outliers superiores.

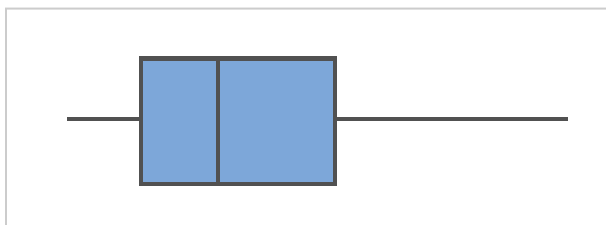
No gráfico Boxplot temos:



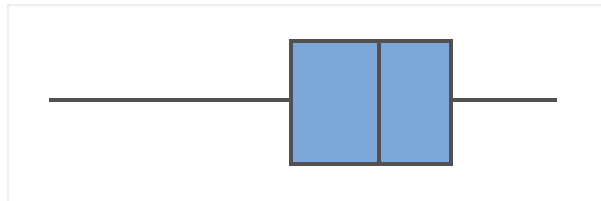
Também usamos o boxplot para entender a simetria dos nossos dados. Um conjunto de dados que tem uma distribuição **simétrica**, terá a linha da mediana no centro do retângulo, como o boxplot abaixo:



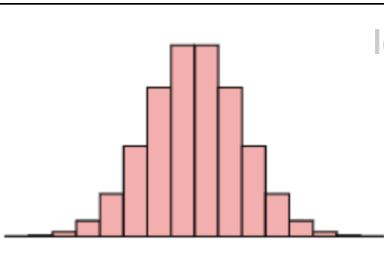
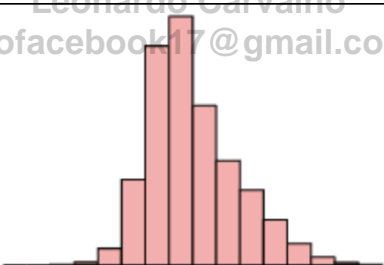
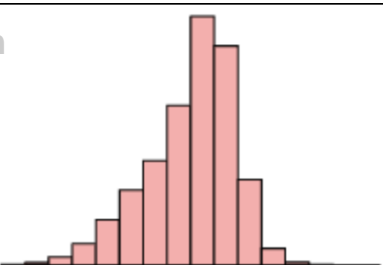
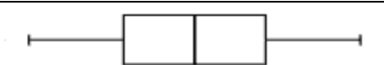
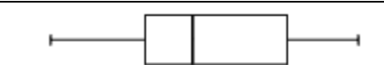
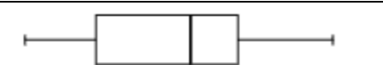
Quando a linha da mediana está próxima ao primeiro quartil, ou seja, quando existe uma cauda mais longa em números maiores (Q3 ou limite superior alongados), os dados são **assimétricos positivos**



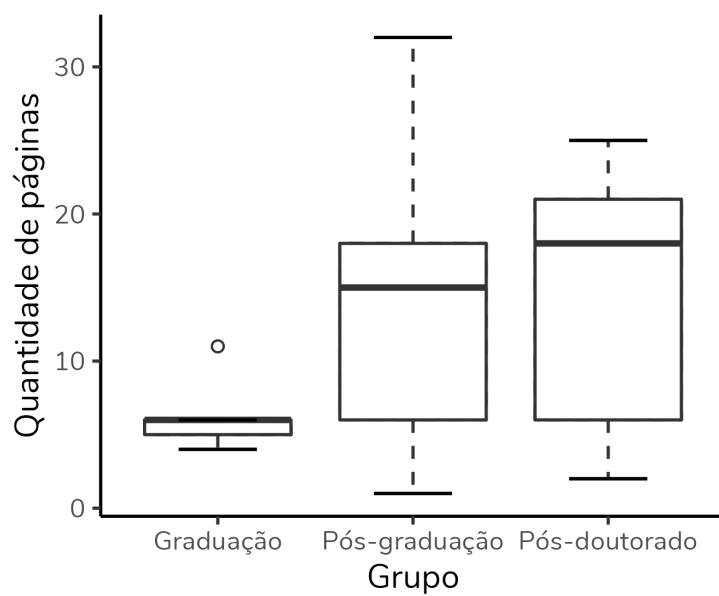
Quando a linha da mediana está próxima ao terceiro quartil, ou seja, quando existe uma cauda mais longa em números menores (Q1 ou limite inferior alongados), os dados são **assimétricos negativos**.



Também conseguimos fazer esse mesmo paralelo com histogramas

| Symmetric | Skewed right (positive) | Skewed left (negative) |
|---|--|---|
|  |  |  |
|  |  |  |

Os Boxplots são extremamente úteis para analisarmos nossos dados. Abaixo temos um 3 boxplots comparativos sobre a quantidade de páginas que alunos costumam ler quando estão na graduação, pós-graduação e pós-doutorado.



Baseado em todas as informações anteriores, qual informação você tira sobre esses dados?

Vamos ver todos os detalhes em aula!

Leonardo Carvalho
leofacebook17@gmail.com



REFERÊNCIAS BIBLIOGRÁFICAS



Leonardo Carvalho
leofacebook17@gmail.com

Bussan, W., Morettin, P. - Estatística Básica - Editora Saraiva - 2010 - 6th ed

Frost, J. - Hypothesis testing - An intuitive guide for making data driven decisions - Jim Frost - 2020 - 1st ed

Huyen, C. - Designing machine learning systems - Editora O'Reilly - 2022 - 1st ed

Knaflic, C.N - Storytelling com dados: Um guia sobre visualização - Editora Alta Books - 2019

Larson, R., Farber B. - Estatística Aplicada - Editora Pearson - 2016 - 6th ed