

# MÓDULO 1

## Estatística Descritiva

### Comentários Aula 01: Variáveis

#### Tópico: quantitativa x qualitativa

**P1:** As variáveis **quantitativas** são sempre de intervalo e razão? E as **qualitativas** sempre nominais e ordinais?

**R:** A variável ordinal pode ser quantitativa ou qualitativa, dependendo de como foi medida. Exemplo: escala de nível de satisfação de 1 a 5 é quantitativa já que podemos tirar uma média ou mediana. Já as medalhas ouro, prata e bronze são qualitativas. As outras que você perguntou estão corretas!

**P3:** Professora, só tenho uma pequena dificuldade em diferenciar **quali de quantitativo** porque quando pegamos uma variável qualitativa, exemplo "grau de instrução", fazemos análises do tipo % de pessoas com ensino superior, etc. Diferenciar as variáveis em qualitativas e quantitativas seria só pra fins didáticos? Acabo não vendo muita função prática nisso

**R:** (Larissa) Olá Jadson, tudo bem? Ótimo questionamento. Então, diferenciar os tipos de amostras em qualitativa ou quantitativa nos ajuda a encontrar quais os possíveis testes estatísticos que podemos fazer com esse tipo de amostra. Como você mesmo falou, dados como 'grau de instrução'(qualitativo) utilizamos a frequência relativa (%) e dados como idade (quantitativos) podemos utilizar a média. Entende? :)

**P4:** Professora, a variável anos de estudo (base PNAD) é **quanti ou quali**?

**R:** (Letícia) Oi Juliana!! Tudo bem?  
A variável anos de estudo é quantitativa contínua =D

**P5:** Estou confusa em relação a data, seria **qual tipo de variável**? (data de entrada e data de saída, por ex)

**R:** (Aline) Olá, Yara! Tudo bem? =)  
De fato, a data é uma variável que confunde a gente porque depende de como ela é "medida" e qual o nosso objetivo. De modo geral pode ser classificada de diferentes formas:

- Caso a gente estiver lidando com dias da semana e meses do ano, então se enquadra melhor como uma variável categórica ordinal, porque há uma ordem/progressão dessas categorias.

- Mas se é data de entrada/saída nesse formato (dd/mm/aaaa), então poderíamos tratar como uma variável quantitativa. E quanto a ser quantitativa discreta ou contínua, acredito que depende dos dados. Por exemplo, imagine que eu tenha a data de entrada e saída diários durante um período contínuo de tempo (estou coletando sem interrupções por exemplo), então faria mais sentido ser tratada como variável contínua. Mas por outro lado, se eu tenho uma quantidade finita de datas de entrada/saída, seria mais adequado tratar como discreta.

É sempre bom analisar caso por caso e ver o que funciona melhor considerando os dados que você tem!

**P6:** Não entendi como é a **representação numérica** da "cor dos olhos". Seria verde - 2, castanho - 5...?

**R:** Alguns questionários definem valores para respostas, mas que são somente representações. Seria sim como seu exemplo (azul:1, verde:2, preto:3, castanho:4, etc). Outro exemplo é associar um valor numérico para cada cidade do Brasil. Mas na hora de analisar, a gente vê qual valor representa qual cidade, ao invés de ver somente os números e tirar a média.

**P7:** Boa Tarde. Estou começando a aula hoje.

Uma dúvida. Trabalho com gestão de pessoas e gostaria de saber a classificação para uma **variável salário**. A princípio, me pareceu contínua, mas levando em conta que todo salário vem de uma tabela salarial, organizada em níveis salariais e cada um com um valor numérico específico, poderia considerar como uma variável discreta? (Ex: o valor de R\$ 1000 corresponde ao nível 1, R\$ 1200, corresponde ao nível 2. ninguém vai ter salário de R\$ 1100 pois esse não existe na tabela). Em relação à escala de medição, seria escala razão? Já que podemos definir claramente o quanto um salário é melhor que outro? Mesmo levando em conta que jamais terei na minha base de dados um empregado com salário zero?

**R:** (Aline) Olá, Carlos! Te desejo muito boas-vindas ao curso! =D

No caso de trabalho formal com salários tabelados, poderíamos tratar a variável salário como discreta porque ela só poderia assumir valores específicos como você mencionou. Mas se estivermos fazendo uma pesquisa por exemplo, com profissionais informais ou autônomos que possuem renda variável, faria sentido classificá-la como contínua. A rigor, quando estamos falando de valor monetário temos uma variável contínua, mas o contexto deve ser levado em conta.

Sobre a escala de medida, seria de razão sim! Mesmo que na prática (legalmente) seria impossível existir um salário zerado, o valor zero é significativo porque não é possível ter

salários negativos e além disso, o zero indica "ausência de salário". Faz sentido? Espero ter ajudado! :)

**P8:** Olá, boa tarde!

Nível de satisfação mesmo sendo representado numericamente pode ser considerado uma **variável qualitativa**?

Por exemplo, caso 0 = nada satisfeito, 5 = pouco satisfeito e 10 = muito satisfeito.

Eu havia entendido que mesmo com a representação numérica o que vale é o que o número de fato representa, isto é, se é de fato um número resultado de uma mensuração ou uma representação de categoria.

**R:** Oi, Ana! Tudo bem?

Depende de como você vai utilizar essas informações na sua pesquisa/estudo. Se você trabalhar somente com as categorias, será uma variável qualitativa. Mas por outro lado, se você considerar apenas os números, seria considerado uma variável quantitativa ordinal porque há uma ordem intrínseca (10 é melhor do que o 5, que por sua vez é melhor do que 0). Faz sentido?

**P9: Nominal e ordinal** estão dentro da variável qualitativa? Ou não ou em ambas?

**R:** A variável nominal sempre será qualitativa. Já a ordinal pode ser quantitativa ou qualitativa, dependendo de como foi medida. Exemplo: escala de nível de satisfação de 1 a 5 é quantitativa já que podemos tirar uma média ou mediana\*. Já as medalhas ouro, prata e bronze são qualitativas.

\*Atenção: A literatura diverge muito quanto a isso, uns falam a média não pode ser usada por não ser uma escala contínua, devendo usar somente a mediana. Já eu sou do time que acredita que podemos fazer a média sim para facilitar uma comparação de escalas (é aqui que entra a teoria pura vs a aplicabilidade).

## **Tópico: Variável Quantitativa**

**P1:** Olá, professora.

Uma variável **quantitativa "contínua"** sempre pode ser identificada por um número decimal?

**R:** (Letícia) Oi Carlos! Tudo bem? =D

Nem sempre uma variável quantitativa contínua é acompanhada por um número decimal. Por exemplo, o peso de um saco de arroz, ele pode ter 1kg, 2 kg, ..., 5kg que são números inteiros, mas você também pode ter uma quantidade de número "quebrado", como 4,5kg ou 3,700kg, que fazem parte de uma escala contínua.

**Cont:** Oi Letícia.

Então, neste caso, o PESO seria categorizado como uma variável **quantitativa CONTÍNUA**, pois ele pode ser representado dentro de um intervalo de valores, embora também possa assumir valores inteiros?

**R:** (Letícia) Isso mesmo, a variável é quantitativa contínua, pois temos uma escala contínua para medi-la, entre 1 e 2, temos infinitos números nesse caso, não tendo como precisar um número exato, pode ser 1 kg, 1,5kg, 1,567kg, etc.

Já na variável quantitativa discreta existe uma quantidade finita e enumerável, exemplo na variável número de filhos, alguém pode ter 1,2, 3 filhos... mas não 1,5 ou 1,75 filhos.

**P2:** Professora, no caso de porcentagens, elas são tratadas como variáveis **discretas ou contínuas**? Imagine que tenhamos 5 defeitos em 100 unidades produzidas. Teremos 5% de defeitos. Isso é discreto ou contínuo?

**R:** (Larrie) Oi Claudio, tudo bem?

Depende. Geralmente porcentagem é um jeito de sumarizar resultados! Por exemplo, temos a frequência absoluta (que são os números em si) e as frequência relativas (que são as porcentagens).

Na maioria dos casos, se você for coletar variáveis que já sejam porcentagens, serão contínuas, por não serem números naturais.

Mas nesse exemplo que você deu, da quantidade de defeitos, perceba que a gente não pode ter "meio defeito" ou "1/4 de defeito", então seria uma variável discreta, ok?

**P3:** os valores dos salários é uma **variável quantitativa contínua**?

**R:** Isso mesmo! As variáveis numéricas que podem ter casas decimais (como o salário, que pode ser R\$1.234,56), são quantitativas contínuas :)

## **Tópico: Escalas de Medida**

**P1:** Professora, no exemplo final, sobre "a quantidade de chuva em SP em um ano", poderíamos entender de duas maneiras: 1) a chuva em milímetro, neste caso, seria uma **razão**. 2) podemos entender também como uma percepção, choveu muito ou choveu pouco. Neste segundo caso, seria um **nominal**. Faz sentido meu raciocínio aqui?

**R:** (Aline)

Olá, Claudio! Tudo bem?

Faz sim! Depende do que você quer mensurar. Se você fizesse uma pesquisa para perguntar a percepção das pessoas sobre a quantidade de chuvas em um determinado período, se foi "pouca" ou "muita", você estaria criando duas categorias, assim, a "percepção das pessoas

sobre o volume de chuvas" seria uma variável categórica nominal. Porém , se depois você quisesse saber com exatidão o volume médio de chuvas nesse mesmo período, você usaria o milímetro como unidade de medida, assim "média de chuvas" é uma variável quantitativa contínua de razão como você disse :)

**Cont:** Só uma dúvida. No caso da classificação choveu "muito", " pouco", "médio"...não seria uma variável categórica **ordinal**? Pergunto pois fiquei na duvida se essa classificação poderia ser colocada em ordem, embora não numérica. Obrigado

**R:** (Aline) Oi Eduardo! Boa noite!

Quando é categórica ordinal, há uma ordem intrínseca nas categorias, a gente não sabe quantificar, mas uma categoria seria melhor do que outra sabe? Por exemplo, se estivéssemos falando de medalhas (ouro, prata, bronze). No caso das chuvas, acredito que dependeria da pesquisa, de qual qual região do Brasil/mundo estou analisando, se estou avaliando produção agrícola ou a influência da chuva em acidentes de trânsito, etc. Então, o mais correto seria dizer que é uma variável ordinal mesmo. Faz sentido pra você? :)

### **Tópico: Escala de Medida Ordinal**

**P1:** Professora, no slide tem "Para **dados ordinais**, alguns cálculos numéricos são possíveis..." Poderia dar exemplo?

**R:** Claro, um exemplo seria a escala de pontuação Likert, que é ordinal, mas podemos tirar uma média para fins comparativos, já que a mediana poderia dar igual entre duas variáveis.

**P2:** Boa noite, professora,

Uma escala Likert que avalia desde o Discorda Totalmente até o Concorda Totalmente mas atribui números a essas respostas seria uma escala qualitativa?

**R:** (Letícia) Boa noite Hendi! Tudo bem? :D

Isso mesmo, a escala Likert é uma **escala qualitativa ordinal**, pois mesmo usando números, nesse caso são uma forma de representação da ideia de "Discorda Totalmente até Concorda Totalmente" e não uma contagem.

**P3:** Fiquei com uma dúvida em "**escalas ordinais**", onde numa frase diz: "... porque o nosso objetivo não é interpretá-las."

Porém no último parágrafo conclui assim: "... mas devemos ter cuidados com quais usar e como interpretá-los."

É interpretável ou não?

**R:** Sobre as interpretações, são duas coisas diferentes:

A primeira frase se refere às diferenças entre os ranques. A gente sabe que tem uma ordem, mas não sabe a distância de um ponto para o outro. Por exemplo, as medalhas de ouro, prata e bronze. Conhecemos essa ordem, mas não sabemos se o primeiro colocado ficou 1 milésimo de segundo na frente do segundo colocado ou se ficou horas a frente. Essa diferença/distância que não interpretamos. Outro exemplo, se eu falo que prefiro os chocolates: amargo, ao leite e branco nessa ordem, você não sabe se o amargo é disparado o melhor ou se ele tá ali quase empatado com o ao leite e vence por um pouquinho, na minha preferência. Você sabe da ordem, mas não consegue quantificá-la.

Sobre a segunda frase, me referi aos calculos numéricos, como por exemplo, uma média. Podemos sim tirar a média dessa variável, mas devemos ter cuidado ao interpretá-la.

**Cont:** MUITÍSSIMO obrigado pela explicação. Ficou bem claro.

Inclusive em relação à segunda frase, esse "cuidado ao interpretá-la" conseguimos entender melhor no vídeo seguinte, a aula 2, que eu já conclui e está bom demais! Muito obrigado por compartilhar todos os seus ensinamentos, amo o jeito que você explica!

**P4:** Uma pesquisa com **variáveis qualitativas ordinais**, como por exemplo, ótimo, bom, regular, ruim e péssimo pode ser considerada uma pesquisa quantitativa? Na faculdade que trabalho, uma pesquisa deste tipo tem como resultado a nota do professor, que é numérica: Professor Nota 10) e não conceitual (Professor Ótimo). Os conceitos são transformados em notas de 1 a 5 e tiradas as médias. Neste caso não seria intervalar?

**R:** Em teoria, essa é uma variável qualitativa ordinal e não deve-se tirar a média. Porém, para a escala likert, não há um consenso na literatura: alguns pesquisadores defendem a teoria e acham que a mediana deve ser usada, enquanto outros acreditam que a média pode ser utilizada para fins comparativos (eu vou com o 2o grupo, já que é seria forma central de melhor visualização e comparação do que seria a mediana).

**P5:** Olá professora, apenas para ver se eu entendi o conteúdo

No caso de notas de uma pesquisa de satisfação, onde 1 é muito insatisfeito e 5 é muito satisfeito, apesar de serem números ela é uma variável Qualitativa e a escala de medida seria a **ordinal**, correto?

É errado então eu fazer uma média das notas para concluir se, no geral, os clientes estão satisfeitos ou não com o serviço?

**R:** (Larrie) Olá!! Tudo bem?

Então Daniela, a escala de medida seria a ordinal mesmo. Essa sua pergunta é muito interessante porque é um ponto de discussão forte. O que temos é: se a variavel ordinal tiver muitos subníveis, pode trabalhar ela como se fosse numérica. Quando é uma pequena escala de 1 a 5, o que eu costumo fazer é trabalhar como variavel qualitativa mesmo (categórica). Mas isso pode ser muito subjetivo, depende muito do seu intuito. Ser errado não é, mas talvez dependendo do seu objetivo não seja o ideal. Mas errado, não é. Consegui esclarecer sua dúvida?

**Cont:** Acredito que ainda vamos ver isso mais pra frente pois no vidro a professora falou que ia mostrar as formas de analisar escalas ordinais, mas aonde eu trabalho a diretoria quer saber como está a pesquisa de satisfação dos clientes em nossos canais de atendimentos, fazemos uma média ponderada das notas para chegar numa nota média. Então, por exemplo, chegamos que no canal A teve média 4.6 e o canal B teve média 3.5, então o canal A estamos com um ótimo nível de satisfação pois os clientes estão mais próximos do muito satisfeito enquanto o canal B temos que melhorar pois a satisfação não está muito boa. Como você faria a análise para este caso, utilizaria ela como quantitativa ou como qualitativa?

**R:** Oi Dani, uma escala ordinal 1-5 seria considerada quantitativa. A literatura diverge muito quanto a isso, uns falam a média não pode ser usada por não ser uma escala contínua, devendo usar a mediana. Já eu sou do time que acredita que podemos fazer a média sim para facilitar a comparação (é aqui que entra a teoria pura vs a aplicabilidade).

Porém, em teoria, você não pode fazer um teste t de comparação, pois uma suposição desse teste é que os dados sejam normais e eles só podem ser normais se forem contínuos. Nesse caso (quando uma suposição é falha), usamos um teste não paramétrico (nesse caso seria o Mann-Whitney - em breve aqui na plataforma).

**P6:** Na **escala ordinal**, temos o conceito de magnitude, certo? Na escala de dor de 1 até 10, 10 representa mais dor do que 5. O que acontece é que não há como mensurar a distância entre uma e outra, é isso?

**R:** Oi Angellica, é isso mesmo. Além de ser subjetivo (o meu 5 pode ser diferente do seu 5), não temos como medir se a distância entre 5 e 6 é a mesma entre 6 e 7, por exemplo. Mas sabemos que 7 é maior que 6, que é maior que 5.

**P7:** Sobre **escalas ordinais**: dá para levar em conta o NPS (que seria o famoso: você indicaria um amigo para a empresa x?) ou pesquisa de satisfação, uma vez que satisfação é um conceito muito amplo, que não é o mesmo para mim e para você, por isso empresas acabam realizando pesquisas com mais perguntas para coletar os dados e poder chegar perto do que seriam os elementos da satisfação.

**R:** Exatamente, essas escalas são difíceis de padronizar, pois o que eu considero como "satisfeita" pode ser diferente para você. Mas olhando a variável isoladamente, ela é sim uma escala ordinal.

### **Tópico: Escala de Intervalo/ Razão**

**P1:** Professora, na aula 1 no minuto 18:46 aprox., vc diz que o tempo dos Ranking de corrida seria de **Escala de Razão**. Não entendi o pq dessa interpretação, uma vez que a variável é de tempo (**Esc. de Intervalo**) e a possibilidade do Zero existe, sendo assim ele é uma variável.

**R:** O tempo em relação ao horário é intervalo, já que 0 horas não significa "falta de" hora. No caso desse exemplo, o tempo é o tempo de corrida (duração). O Bolt levou 9,63 segundos para percorrer 100 metros. Essa é uma variável razão, já que 0 segundos seria sim "falta de", ou seja, um atleta que não correu.

**Cont:** Nesse caso, se a variável estiver em tempo total uso a escala de razão, mas se for horário relógio uso escala intervalar? Estou correta?

**R:** Sim, na verdade devemos sempre analisar o tipo da variável. Não é porque é "tempo" que vai ser sempre o mesmo tipo.

**P2:** Se a temperatura for medida em Kelvin, que tem um zero absoluto, ela pode ser considerada **intervalar**?

**R:** Ótima pergunta: a escala Kelvin é considerada **razão** por ter o zero absoluto. Já Celsius e Fahrenheit são intervalares.

Entender as variáveis pode não ser tão simples. Uma mesma variável, temperatura, tem medidas diferentes, dependendo de que escala estamos falando... Por isso devemos ter cautela ao analisar as variáveis.

**P3:** Ainda com **dúvidas de Intervalo e Razão**. Algum texto extra ?

**R:** Oi Ricardo, quer compartilhar aqui sua dúvida? Os livros didáticos falam sobre isso, mas eles dão exemplos mais simples, no geral. Algumas variáveis são difíceis de serem analisadas...

**Cont:**

Minha dúvida é se a **escala de Razão** é subconjunto da **escala de intervalo** uma vez que a diferença é o zero. Neste caso, porque duas escalas?

Outra pergunta, a escala Kelvin possui zero absoluto. Neste caso seria uma escala do tipo intervalo ?

**R:** Oi Ricardo! Não é um subconjunto, já que em uma inclui a ideia de "zero absoluto" e a outra não. Ou seja, uma variável pode ser razão ou ser intervalo (2 coisas separadas).

Ótima pergunta sobre a escala Kelvin: é considerada razão por ter o zero absoluto. Já Celsius e Fahrenheit são intervalares.

Viu como não é tão simples? Até temperatura pode ter medidas diferentes, dependendo de que escala estamos falando... Por isso devemos ter cautela ao analisar as variáveis.

**P4:** Por que o tempo em horas do jogo foi dado como **Razão** e o tempo em horas da chegada do cliente foi dado como **Intervalo**? [https://drive.google.com/file/d/102AU7oN7Reh-02eQGxtwcXgLdhWGA\\_92/view](https://drive.google.com/file/d/102AU7oN7Reh-02eQGxtwcXgLdhWGA_92/view)



**R:** (Letícia)

Olá Fernando,  
tudo bem?

O tempo em horas do jogo foi dado como razão pois o 0 significa "ausência de tempo", ou seja, o tempo passou a ser contado dali, a corrida ainda não havia iniciado.

Já o horário que o cliente entra na loja é medido em intervalo, pois 0:00 horas não significa que "não existia tempo", as 0:00 horas existe no relógio.

Espero ter te ajudado :D

**P5:** Olá, professora. Pode me explicar mais a **escala de razão**?

**R:** (Letícia) Olá Mineia, tudo bem? =D

Na escala de razão, o "0" possui um significado, como "falta de" algo, por exemplo, na variável número de filhos significa que não há filhos.

Diferente da escala de intervalo onde o zero absoluto existe, como no caso de medir uma temperatura por exemplo, quando dizemos que está fazendo  $0^{\circ}\text{C}$ , está sentindo frio, certo?! Então o  $0^{\circ}$  existe, você não deixa de sentir temperatura porque está em  $0^{\circ}$ .

**P6:** A **escala de razão** tanto pode ser englobada como uma variável quantitativa discreta (valores inteiros) quanto contínua (fracionária, ou seja, com os "meios") ?

**R:** (Letícia) Oi Carlos,

Sim, a escala de razão você pode usar nos dois tipos de variável, desde que faça sentido em um cálculo de razão, onde só não é possível a divisão por 0, os demais números podem ser aplicados nesse tipo de cálculo.

**P7:** Manipulações algébricas possíveis na **escala de intervalo** são válidas na **escala de razão**? Podemos então operar adição e subtração nas medidas de intervalo e razão mas multiplicação e divisão somente nas medidas de razão, é correto isso?

**R:** (Larissa) Oii Angellica. Tudo bem? É isso mesmo, os dados da escala de razão podem ser multiplicados, divididos, adicionados ou subtraídos, já os dados da escala de intervalo só podem ser adicionados e subtraídos.

**P8:** Professora, eu tive dificuldade de compreender no exemplo a diferença da escala de medida quantitativa **intervalar da razão**, quando você citou o exemplo da temperatura em que  $40^{\circ}$  não será o dobro de  $20^{\circ}$ , isso se refere a qualidade do calor e por isso não podemos inferir que o calor será o dobro apesar de ter dobrado o intervalo?

**R:** (Letícia) Oi Miykaella, tudo bem? =D

Na escala de razão, o "0" possui um significado, como "falta de" algo, por exemplo, na variável número de filhos significa que não há filhos.

Diferente da escala de intervalo onde o zero absoluto existe, como no caso de medir uma temperatura por exemplo, quando dizemos que está fazendo  $0^{\circ}\text{C}$ , está sentindo frio, certo?! Então o  $0^{\circ}$  existe, você não deixa de sentir temperatura porque está em  $0^{\circ}$ .

**P9:** Sinto que não compreendi totalmente a escala de medida de **intervalo e razão**.

A variável é de escala de medida intervalo quando os valores são numéricos e só faz sentido naquele contexto os valores serem somados ou subtraídos e o valor zero e/ou negativos podem ser quantificados /interpretados (têm sentido naquele contexto)?

A variável de razão - é possível realizar qualquer cálculo (soma, subtração, multiplicação, divisão). Ou seja, no contexto daquela variável, o zero significa ausência ou falta.

Parece que vai sempre depender do contexto da variável. Obrigado

**R:** Oi Pedro, é isso mesmo!

A escala intervalar teria como exemplo a temperatura ou o ano de nascimento. Alguém pode ter nascido no ano 0 e não significa ausência.

A escala razão seria número de filhos ou renda. Uma pessoa com 0 renda, não a tem, e uma pessoa que ganha 4 mil ganha o dobro do que uma que ganha 2 mil.

**Cont:** Fiquei a pensar no exemplo que deu do ano do nascimento e surgiu-me a questão da idade de uma pessoa. Também é **intervalar**? Um bebé pode ter zero anos (pode só ter meses), mas é um valor que pode ser quantificado (posso estar enganado, mas do meu ponto de vista, não há ausência de valor, porque zero é um número como outro qualquer neste contexto)? No entanto, também dá para fazer mais cálculos para além da soma e subtração. É possível verificar que uma pessoa de 20 anos tem o dobro da idade de uma de 10 anos e que uma de 20 anos tem três vezes menos idade do que uma de 60 anos. Não sei nestas situações o que é mais importante observar, se o valor do zero ou se é possível fazer todos os cálculos.

Desculpe ainda a confusão. Acho que quanto mais penso mais confuso fico :)

Obrigado

**R:** Não se preocupe, é normal ficar confuso! E isso é ótimo, pois está praticando e pensando sobre esses conceitos. Sobre a idade, é uma escala razão, já um bebé com 0 anos tem ausência de idade em anos. O que é diferente da temperatura  $0^{\circ}\text{C}$  que não significa ausência de temperatura em Celsius. Espero que tenha conseguido esclarecer, mas se tiver mais dúvidas, só perguntar :)

**P10:** Se a temperatura for medida em Kelvin, ela tem **escala de razão**? E se for, há alguma vantagem em converter temperaturas em outras escalas para Kelvin, e depois fazer análises ou modelos?

**R:** (Aline) Olá, João! Tudo bem? :) A escala de temperatura em Kelvin é de fato uma variável de escala de razão porque seu valor zero é absoluto, ou seja, nada pode ser medido a uma temperatura inferior a 0 K. Dentre as escalas de medidas, a de razão é a mais completa porque nos oferece todas as propriedades das outras e nos oferece a possibilidade de multiplicação, divisão e com isso podemos ter noções mais diretas da temperatura, como por exemplo, 200 K é duas vezes mais quente do que 100 K, o que não é possível de ser feito em Celsius.

A vantagem é que todos os tipos de análises estatísticas podem ser aplicados aos dados que estão na escala de razão. Mas eu diria que essa conversão depende mais do tipo de análise que está se querendo fazer. Em pesquisas científicas, a escala Kelvin é a mais preferida pelo seu aspecto mais direto/absoluto, mas para o dia a dia, a escala Celsius e Fahrenheit já é suficiente para nos dar a previsão e a média do dia por exemplo. E se for uma pesquisa de satisfação, por exemplo, da percepção dos clientes em relação à temperatura de um certo prato, talvez a escala ordinal já é o suficiente, porque estaríamos lidando com noções como 'adequado', 'muito quente', etc.

### **Tópico: Outros/ Exercícios**

**P1:** Gostaria de entender como realizar corretamente o exercício 2 do módulo 1

**R:** Oi Antonio! Claro, começando no ponto 0 horas com largura de 10, devemos selecionar 0-9, 10-19, e assim sucessivamente. Se dividir em 0-10, terá uma largura de 11.

**P2:** Boa tarde. Fiquei com duvida na resolução do exercício 5A. Como chegou no resultado de 87?

**R:** Oi Ana Paula, boa tarde! Como não temos os dados e somente a informação do boxplot, fazemos uma aproximação pelo gráfico. A mediana é o valor do tracinho central do boxplot. O boxplot que representa a área de negócios é o verde (Business). Vemos que o tracinho central está entre 80 e 90, mais próximo a 90, logo a mediana é aproximadamente \$87 (mil/ano).

**P3:** Boa tarde, Aline! Tudo bem?

Acabei de resolver os exercícios da aula 1, módulo 1, e fiquei com as seguintes dúvidas:

1. Por que "número de irmãos" tem escala de medida de razão?
2. De modo geral, o que podemos inferir quando a distribuição dos dados é assimétrica positiva e quando ela é assimétrica negativa? Em suma, por que é importante saber se ela é positiva ou negativa? O que isso me diz sobre os dados?
3. Qual a lógica de resolução do exercício 5b? Fique um pouco confusa.
4. Posso entender a amplitude interquartil como a distância de "pontos/ valores" entre Q3 e Q1?

5. Tive dificuldade em calcular o desvio padrão (exercício 7A). Não cheguei na resposta correta. edit: vi seu arquivo Excel com a lógica de resolução. Mas no cálculo da variância (célula G15), a divisão do somatório dos desvios foi por 4, não deveria ter sido por 5, que é o número de observações?

6. Como se calcula a correlação? Como o exercício 7b, módulo 1, chegou a um valor de porcentagem por exemplo?

7. como calcular a correlação? O gabarito do exercício 7D traz, inclusive, a resposta em porcentagem. Qual foi o raciocínio para se chegar à resposta?

No mais, as aulas são super didáticas e eu estou amando o curso! Um abraço.

**R:** (Aline) Boa tarde, Stefanie! Tudo jóia e você?

Que maravilha saber que você está amando o curso! Obrigada pela mensagem! :D

Vou te responder de acordo com o número das perguntas, tudo bem?

1) Normalmente são as variáveis contínuas que entram na categoria de escala de razão, mas a variável número de irmãos, mesmo sendo discreta pode ser considerada como de razão, pelas características a seguir:

\* Presença de zero absoluto ou verdadeiro: Ao dizer '0 irmãos', significa ausência de irmãos. Não é possível ter 0 irmãos ou valores negativos para essa variável;

\* É possível calcular soma, subtração, multiplicação e divisão e com isso fazer algumas avaliações, por exemplo: "João tem 3 irmãos, e eu tenho 1 irmão, logo, João tem o triplo de irmãos!".

\* É possível também calcular a média e desvio padrão, mas como estamos falando em quantidade de pessoas, faz mais sentido usar os valores inteiros nos resultados.

2)

Quando percebemos que há assimetria nos dados (positiva ou negativa) podemos já pensar em algumas coisas, por exemplo:

\* Pode ser um indício de que haja outliers;

\* Quando temos muita assimetria, significa que há muita variabilidade nos dados, então a média provavelmente não vai representar bem os meus dados porque ela é influenciada por valores extremos desse conjunto. Nesse caso, a mediana é uma medida resumo melhor para dados assimétricos;

\* Se há assimetria provavelmente os dados não seguem a distribuição normal, justamente porque uma das suas características é a simetria. E saber sobre a normalidade dos dados é requisito para fazer alguns testes estatísticos, por exemplo. Então, ao analisar a assimetria, você já tem uma intuição inicial sobre a normalidade dos dados;

- Se a assimetria for positiva, significa que a maior parte dos dados estão concentrados abaixo da média (os valores extremos 'puxam' a média para a cauda da direita). Aqui saberemos que a média é maior que a mediana;

- E ao contrário, se for assimetria negativa, os dados estão concentrados acima da média (os valores extremos 'puxam' a média para a cauda da esquerda). Aqui a média é menor que a mediana.

De forma geral, a assimetria, juntamente com a curtose, te ajuda a ter uma ideia geral do formato da distribuição dos dados e se você pode julgar se a média é uma boa medida para representá-los ou não. Se puder, aconselho ler esse artigo do blog que explica com mais detalhes sobre essas medidas: <https://blog.proffernandamacieli.com.br/assimetria-e-curtose-dos-dados/>

3)

Para recordar, os quartis dividem os nossos dados em 4 partes, sendo que o:

Q1 – valor que engloba até 25% dos dados;

Q2 (mediana) – valor que engloba até 50% dos dados

Q3 – valor que engloba até 75% dos dados

Nesse link, você pode ver uma imagem dessa ideia:

<https://drive.google.com/file/d/1bRjPvcL5nXclaeTX-tF0xqwj0VVAKDWY/view?usp=sharing>

No exercício, pelos boxplots apresentados, a gente pode ver que a mediana (Q2) dos salários de business está muito alinhada com o Q3 dos salários de direito, certo? Por ser o Q3, sabemos então que 75% é percentual de pessoas na área de direito que tem salário abaixo da mediana dos salários em negócios.

4) Exatamente, a amplitude ou intervalo interquartil é a diferença (distância) entre o valor do Quartil 3 e o valor do Quartil 1. No boxplot, representa o comprimento da caixa!

5) No exercício 7A, a divisão foi por 4 porque estamos lidando com amostras, então a fórmula da variância muda um pouquinho, ou seja, devemos dividir por  $n-1$  (total de observações menos 1), nesse caso:  $5-1=4$ . Vamos dividir por  $N$  (total de observações) só quando estivermos trabalhando com toda a população, o que é mais raro. No Excel, precisa ter atenção porque as fórmulas mudam, se for calcular desvio padrão para amostra é DESVPAD.A e se for para população é DESVPAD.P.

Nesse post do blog também é explicado sobre medidas de dispersão e são mostradas as fórmulas: [https://blog.proffernandamacieli.com.br/medidas\\_dispersao/](https://blog.proffernandamacieli.com.br/medidas_dispersao/)

Com isso, ficou mais claro resolver essa questão? Se tiver outra dúvida, é só dizer! :)

6 e 7)

A fórmula da Correlação compreende a fórmula da covariância e do desvio padrão juntas, dá uma olhadinha:

<https://drive.google.com/file/d/1EAMQjN4oi9v0V7dAFT04ERfYvcI9icTp/view?usp=sharing>

(Esse print é do livro Manual de Análise de Dados, Fávero).

Como a correlação positiva varia entre 0 e +1 e correlação negativa entre 0 e -1, você pode ver casos, como no resultado desse exercício, em que se multiplica o resultado da correlação por 100, apresentando em porcentagem. Isso seria só pra ajudar na interpretação, mas não é necessário.

Usando a função CORREL no Excel, você vai encontrar que a correlação é 0,57 (arredondado) e a interpretação seria a mesma, é uma correlação moderada :)

Uma observação: no exercício 7 , as variáveis de retorno anual para eletrônicos e energia já são dadas em porcentagem, então a média e o desvio padrão também estão nessa unidade de medida. Podemos dizer que o retorno médio é 9,92% e o desvio padrão é 47,07%.

Espero ter ajudado Stefanie,

**P4:** Boa tarde! Com relação ao exercício 7 eu acertei as médias, porém na hora de calcular o desvio-padrão eu errei. Gostaria que fossem apresentados os cálculos para saber onde eu errei

**R:** (Aline) Olá, Érika! Claro! :D

Pelo Excel podemos calcular o desvio padrão diretamente pela fórmula DESVPAD.A, considerando que estamos trabalhando com amostras, tudo bem?

No arquivo a seguir eu deixei também o passo a passo de como calcular o desvio padrão através da variância:

<https://docs.google.com/spreadsheets/d/1CWkju37OCih6HSQ4ozDjxWX6XIA4KUcL/edit?usp=sharing&ouid=102019378199879856305&rtpof=true&sd=true>

Espero ter ajudado! Se ficou alguma dúvida, fique à vontade para falar :)

## Comentários Aula 02: medidas de tendência central

### Tópico: média e mediana

**P1:** Boa tarde, qual seria o caso em que é melhor tirar a **média frente a mediana**?

**R:** (Larrie) Boa noite, rs (respondendo a noite). Tudo bem? Então Gabriel, em geral utiliza-se a média, geralmente a mediana é utilizada quando os dados possuem outliers que afetam bastante o resultado dos dados.

Tipo, imagina que você tem dados relacionados a idade de 10 indivíduos: 17,17,17,17,17,18,19,19,20,21.

Se fizermos a média dessa idade o resultado é 18,2 e a mediana 17,5. O que condiz com as idades que temos na amostra.

Mas e se adicionarmos um indivíduo com 80 anos. A média fica de 23,5 e a mediana de 18, nesse caso é melhor utilizar a mediana.

**P2:** Após concluir a aula fiz uma busca na internet sobre **média e mediana** e achei uma publicação no LinkedIn reforçando o que foi dito no vídeo e trazendo um outro exemplo sobre a diferença entre ambos os métodos.

<https://www.linkedin.com/pulse/m%C3%A9dia-ou-mediana-qual-dado-utilizar-na-an%C3%A1lise-descritiva-pablo-lopes/>

Com isso me surgiu uma dúvida professora. Para saber se eu posso usar a média, eu consigo no Excel, por exemplo, fazer:

Primeiro - uma função de mínimo e máximo, para achar o menor e o maior valor;

Segundo - depois fazer a média e verificar o quão distante essa média está dos valores menor e maior encontrado?

Existe uma convenção que diga que se o maior valor for X vezes maior que a média seja considerado um outlier? Ou que se o menor valor for X vezes menor que a média seja considerado um outlier?

**R:** Para saber se pode usar a média, tem que saber se tem outliers e se a distribuição da variável é simétrica. Se tiver outliers e/ou assimetria, a mediana é uma melhor medida.

Para verificar outlier pode ser pelo boxplot ou z-score (próximo módulo)

**P3:** Profª, supondo que eu esteja fazendo uma análise de alguma variável qualquer em alguma empresa, e a **média não seja significativa**, então escolho a mediana para ser mostrada em um dashboard, como eu comunico isso para os interessados? Explico que metade dos dados estão abaixo ou acima desse valor? É que dificilmente se escuta esse tipo de colocação, geralmente as pessoas falam "em média". Como você sugere que seja feita essa comunicação?

**R:** Oi Thais, boa pergunta. Realmente em muitos lugares a mediana não é usada e um dos motivos é o desconhecimento. As pessoas não entendem essa medida. Nesse caso, eu

apresentaria ambas informações. A média, velha e conhecida de casa, mas também a mediana. Sua interpretação está corretíssima, é uma medida que mostra que metade dos dados estão acima e metade abaixo, essa é a melhor forma de apresentar com fácil entendimento. Mas se você usar só a mediana, a audiência pode reclamar (sendo um gestor ou cliente), então mostre as duas, ok?

### Tópico: outliers

**P1:** Professora, sobre analisar os dados de salário observando que na amostra contém **Outliers**, talvez analisar com a Média Geométrica não seria até melhor que a mediana ou a moda ?

Fiz os cálculos aqui e estes foram os seguintes resultados:

Xg	R\$ 101.628,14 (Méd Geom.)
X	R\$ 154.285,71 (Méd Simp. )
Me	R\$ 90.000,00 (Mediana )
Mo	R\$ 40.000,00 (Moda )

**R:** A média geométrica também é uma boa medida para análise descritiva, mas tem 2 "probleminhas":

(1) é mais difícil de explicar/interpretar que a mediana. Se a audiência souber, ótimo, mas para o leigo pode dificultar e a ideia é facilitar;

(2) damos prioridade para usar a média se formos fazer inferência estatística (a partir do módulo 3) ou a mediana se formos fazer análise não paramétrica (que será um novo módulo em breve).

**P2:** Professora, no caso de termos o **outliers** como no exemplo de salário no qual o presidente ganha 500k. O melhor seria retirar esse outlier e fazer a média dos salários restantes ou fazer a mediana mesmo? Nesse caso ambos ficariam com valores muito parecidos, então fica fácil avaliar e escolher qual usar pq nosso N é pequeno. Mas no caso de termos uma amostra muito grande, como saberíamos qual das opções é a melhor?

**R:** (Larissa) Olá Americo, tudo bem? Boa pergunta. Então, se o seu conjunto de dados for grande suficientemente você pode remover o outlier que não terá tantos prejuízos para sua análise. Além da remoção ou utilização da mediana você também pode realizar a transformação dos seus dados, através da transformação logarítmica.

**P3:** Como classificar dados como **outlier**? Há algum critério numérico para tanto? Quão díspare e em relação a quê certos dados podem ser enquadrados como outliers?

**R:** (Aline) Olá, Hugo! Tudo bem?

Há sim cálculos para identificar quais dados seriam outliers ou não dentro de um conjunto de dados específico. O mais comum é utilizar o intervalo interquartil que pode ser visualizado



no gráfico de boxplot (ambos serão explicados com mais detalhes na próxima aula) . Mas já adiantando, o intervalo interquartil é a diferença entre o terceiro e primeiro quartil:  $IQR = Q3 - Q1$  (que é exatamente o tamanho da caixa no gráfico box plot).

Em seguida fazemos:

$Q1 - 1,5 * IQR$  -> todos os valores abaixo desse valor serão considerados outliers

$Q3 + 1,5 * IQR$  -> todos os valores acima desse valor serão considerados outliers

No livro 'Manual de Análise de Dados' , o professor Fávero traz uma nova versão desse cálculo que seria:

$Q1 - 3 * IQR$  e  $Q3 + 3 * IQR$ . Assim, todos os dados abaixo ou acima desses valores, respectivamente seriam outliers do tipo 'extremos'.

Há também outras formas de detectar outliers, usando Z-scores (que será mencionado mais a frente no curso) e também pelo teste de Dixon e teste de Grubbs se você quiser se aprofundar mais! :)

**P4:** Olá pessoal sou novo aluno da turma 9 , estou com dúvida de entender realmente sobre o **sentido estatístico do outlier**. Se for para explicar para uma criança como seria ...Eu sou leigo em estatística .

**R:** (Aline) Boa tarde, Antonio! Tudo bem? Seja muito bem-vindo!

Sua dúvida é muito importante, vamos lá!

O outlier é aquele valor que está muito afastado dos demais valores do conjunto de dados, certo? É o que conhecemos popularmente como o 'ponto fora da curva' (falamos isso geralmente para alguém que se destaca em alguma coisa :D)

Mas por que o tema de outlier é importante na estatística?

Bom, quando estamos fazendo análises, normalmente desejamos ter um número que represente o conjunto de dados que estamos trabalhando, o que será muito útil para:

- a) nos dar um resumo de todos aqueles dados;
- b) futuramente fazer testes estatísticos, como o de hipóteses, por exemplo.

Geralmente esse representante 'eleito' é a média aritmética simples. Porém, a média tem um ponto fraco que é ser muito influenciável por valores muito diferentes (os outliers). E isso pode nos levar a tirar conclusões errôneas da realidade, por exemplo, como foi mencionado no exemplo da aula, achar que os funcionários de uma empresa ganham um salário altíssimo, mas que infelizmente não é verdade.

Então, caso identifiquemos um ou mais outliers no conjunto de dados, temos que:

1º - Confirmar se esse outlier foi um erro de digitação, por exemplo;

2º - Caso seja um dado verdadeiro, se eu manter esse outlier no conjunto de dados, qual impacto isso terá na minha análise? Posso chegar a conclusões falhas sobre o assunto que estou estudando?

Fez mais sentido agora, Antonio? :) Fique à vontade para sempre fazer perguntas! Na próxima aula a prof<sup>a</sup> Fernanda vai explicar melhor como detectar e visualizar um outlier utilizando gráficos :)

## Comentários Aula 03: Histograma e Boxplot

### Tópico: boxplot

**P1:** Olá! Visualmente, o **box and whiskers plot** me lembra muito o intervalo de confiança que aparece em meta-análises. Eles são a mesma coisa?

**R:** Oi Jan! Eles parecem mas não são iguais. Você vai entender melhor sobre intervalo de confiança, como é calculado, e qual o propósito de ser feito no Módulo 4.

O boxplot não te dá uma confiança, mas uma representação gráfica dos seus dados, é como se você pegasse todos os seus dados, colocasse em ordem e dividisse em 4 fatias iguais - e é assim que vemos sua distribuição.

**P2:** Parabéns pela clareza e pela apresentação visual ! Por favor, eu tenho 2 dúvidas: i) no momento 24m, o histograma B apresenta **outliers** nos marcadores 9 e 10 do eixo X. No boxplot correspondente 1, os outliers surgem logo após a marcação 5 e se estende. Eu não entendi essa diferença nos gráficos; ii) O que define o **tamanho do box** ? Os 4 boxes apresentam tamanhos diferentes

**R:** (Larrie) Oi Fábio! Tudo beleza?

Antes de responder suas dúvidas, vou te indicar um texto super legal do nosso site que fala sobre como ler um boxplot. Todas as suas duvidas podem ser sanadas com esse texto (e indo além, até aprofundando mais no assunto para você entender de uma vez por todas como montar um bo Axplot)

O link é: <https://blog.proffernandamacieli.com.br/como-ler-um-boxplot/>

De qualquer forma, vou explicar aqui nos comentários sobre suas duvidas. Mas indico super você ler o texto também :)

Sobre a primeira dúvida: Os gráficos são equivalentes. O que define um outlier em um boxplot é: Limite Inferior =  $Q1 - 1,5 * IIQ$  e Limite Superior =  $Q3 + 1,5 * IIQ$ . Então tudo que vem além ou antes disso é outlier. Não tenho acesso aos dados dos gráficos da imagem, mas muito provavelmente se tivesse acesso, usando essa formula que te mostrei, os outliers surgiriam no 5 mesmo. No histogramas não temos como ver de maneira mais clara outliers, fica bem subjetivo. Só da pra ver isso mesmo no boxplot. Então os gráficos boxplot 1 e histograma b são equivalentes, só que no histograma você não consegue perceber de maneira tão clara que nem no boxplot que os outliers começam a partir do 5 mesmo.

Sobre a segunda dúvida: o que define o tamanho do box é o intervalo interquartil! A dispersão dos dados pode ser representada pelo intervalo interquartil (IIQ) que é a diferença entre o terceiro quartil e o primeiro quartil (tamanho da caixa). Cada box tem um tamanho diferente porque cada gráfico apresenta um intervalo interquartil diferente.

Ficou mais claro? Depois me conta se entendeu e se gostou do texto que te indiquei :)

**P3:** Professora, é possível fazer **Boxplot** no Excel ?

**R:** Nas versões mais recentes, sim! Não tem essa opção dentro do nosso plugin "análise de dados", mas você pode ir em Inserir > Gráficos, e no botão do histograma tem a opção boxplot.

### Tópico: quartis

**P1:** No exercício 6, apesar de já citado, **encontrei valores diferentes para o primeiro e o terceiro quartis** (67,5 e 85,5). Existe alguma forma de cálculo diferente da média aritmética simples quando o número de elementos é par?

**R:** Boa pergunta! Na verdade tem uma continha, pois já que o número de elementos é par, a mediana é aquele valor "invisível" no meio. Nesse exemplo os dois valores no meio são iguais, mas e se não fossem?

Existe uma continha chata que dá para fazer na mão, mas minha dica é usar no excel a função de quartil: "**=QUARTIL.EXC(matrix;quarto)**", onde 'matriz' é a coluna dos dados e 'quarto' é o quartil, ou seja, 1 para o 1o, 2 para o 2o, etc.

**P2:** Não ficou claro para mim como se calculam os quartis, especialmente **Q1 e Q3**. Poderiam me explicar? Obrigada!

**R:** (Aline) Bom dia Letícia! Tudo bem? Claro, vamos lá! :)

Começando pelas definições: os quartis Q1, Q2 e Q3 dividem o conjunto de dados em 4 partes iguais:

Q1 = 1/4 ou 25% dos seus dados ficam sobre ou abaixo desse primeiro quartil

Q2 (Mediana) = 1/2 ou 50% dos seus dados ficam sobre ou abaixo do segundo quartil

Q3 = 3/4 ou 75% dos seus dados ficam sobre ou abaixo do terceiro quartil

Para encontrar os quartis a gente sempre usa o cálculo da mediana. Relembrando, a mediana é o valor que divide o conjunto exatamente no meio. Se for um conjunto ímpar você já identifica esse número. Se o conjunto for par, a gente calcula a média dos dois valores que dividem igualmente o conjunto.

Vamos ver um exemplo para encontrar os quartis!

Supondo que a gente tem o seguinte conjunto de dados {7 20 16 6 58 9 20 50 23 33 8 10 15 16 104}

\* A primeira coisa a se fazer é ordenar os dados em ordem crescente = {6 7 8 9 10 15 16 16 20 20 23 33 50 58 104}

\* Em seguida, a gente calcula Q2, que é a mediana desse conjunto. Como temos um grupo ímpar com 15 valores, basta a gente conferir o número que irá dividir o conjunto pela metade.

Então o Q2 será o número 16 que divide o grupo em dois, cada uma com 7 valores. Os subgrupos são: { 6, 7, 8, 9, 10, 15, 16} e { 20, 20, 23, 33, 50, 58, 104}.

\* O Q1 será a mediana dos valores que estão à esquerda de Q2 (isto é, o primeiro subgrupo) que é = { 6, 7, 8, 9, 10, 15, 16}. Novamente, como temos um conjunto ímpar, só precisamos achar o valor que irá dividir esse subgrupo pela metade. Então, o Q3 será o número 9 porque ele divide o subgrupo igualmente nas metades { 6, 7, 8} e {10, 15, 16}.

\* E o Q3 será a mediana dos valores que estão à direita de Q2 (o segundo subgrupo) que é = {20 20 23 33 50 58 104}. Como sabemos que é um conjunto ímpar, basta encontrar o número que divide esse subgrupo pela metade também. Assim, Q3 é o número 33 porque ele divide o subgrupo igualmente nas metades { 20, 20, 23} e {50, 58, 104}.

Então os quartis são: Q1 = 9, Q2 = 16, Q3 = 33. Segue um esquema pra você visualizar: <https://drive.google.com/file/d/1bEwK1STZMyUiEuCAZy7oncLGXggzhtF-/view?usp=sharing>

Esse post do blog da profª Fernanda também explica esse assunto, Leticia! <https://blog.proffernandamaciel.com.br/como-ler-um-boxplot/>

Espero que tenha te ajudado =)

**P3:** Professora, fiquei com dúvida em relação ao Q4 que não aparece. **O Q4 seria o máximo?** Fiquei confuso nessa parte do quartil.

**R:** É isso mesmo! Não medimos porque é o valor máximo.

**Cont:** Professora, nesse caso seria o valor máximo desconsiderando possíveis outliers de maior valor (caso houvesse algum)? Ou os outliers, caso existam, entram nessa consideração do valor máximo?

**R:** Sim, o valor máximo é o maior valor que não é um outlier. Se houver outliers nos dados, eles entram como asteriscos, como mostrado no exemplo dos ganhadores do Oscar.

**P4:** Gostaria de saber sobre a **interpretação do boxplot** (último exercício).

Eu poderia interpretar pela proporção?

Pensei nas porcentagem pelos **quartis**: mínimo ao Q1 = 25%, até Q2 = 50% e assim por diante)

Pensei em falar que cerca de 75% das atrizes ganham óscar com idade inferior aos atores. Está certo?

**R:** Você pode interpretar pela proporção sim, porém nesse exemplo seria: cerca de 75% das atrizes ganharam o oscar com idade inferior a mediana dos atores. Você pode observar que o Q3 das atrizes está praticamente na mesma altura do Q2 dos atores.

Ou você pode falar que 25% das atrizes que ganharam o oscar têm idade inferior aos atores, já que o Q1 das atrizes está aproximadamente na altura do valor mínimo dos atores.

**P5:** Professora, sobre a interpretação do boxplot, focando na caixa, também seria correto afirmar que entre o **Quartil 2 e o Quartil 3** do boxplot de atores (Oscar), há uma maior variabilidade de dados entre a mediana e o terceiro quartil, que seria uma maior variabilidade da idade de atores que ganharam o Oscar entre 42 e 50 anos. Certo?

**R:** O boxplot divide os dados em 4 partes iguais, então quando a área da caixa é maior, há sim mais dispersão (variabilidade). Outra interpretação seria: um quarto dos atores que ganham o Oscar está entre 39 e 42 anos e um quarto está entre 42 e 50 anos.

**Cont:** Tudo bem: "a área da caixa é maior..., há mais dispersão"! Mas me refiro apenas entre 42 e 50 anos. Posso afirmar que há maior variabilidade entre as idades de atores entre 42 e 50 anos que ganham o oscar, certo?

**R:** Focando na caixa, sim. Porque pensando no boxplot como um todo, eu diria que há maior variabilidade no último quarto: entre 50 e 61 (aprox).

**P6:** Olá Fernanda.

Estava a ver no seu blog o exemplo que utilizou para qualquer os **quartis** dos dados e não percebi como chegou àqueles resultados (Q1, Q2, Q3). Há algum cálculo que seja necessário fazer? Pensei que era só colocar os números por ordem (do menor para o maior) e dividi-los em 4 partes (min;25-50%,50-75%-max) e o nº que correspondesse à sua respetiva percentagem era o selecionado ao quartil. No entanto, não estava a perceber como era suposto fazer quando não desse para dividir exatamente o mesmo número de números em 4 partes.

Exemplo 1: para os seguintes dados: 7, 9, 16, 36, 39, 45, 45, 46, 48, 51

Mínimo= 7

Q1 = 14,25

Q2 (mediana) = 42

Q3 = 46,50

Máximo= 51

Não sei se está confusa a minha dúvida.

**R:** Oi Pedro,

O seu raciocínio está totalmente correto. Só que as vezes não temos exatamente os valores de corte nos dados. Nesse exemplo no blog, temos 10 dados, e não temos um valor exatamente no meio para ser a mediana. Então ela será a média dos dois valores centrais (39 e 45).

Na verdade existe uma conta para calcular cada quartil, mas minha dica é usar no excel a função de quartil: "**=QUARTIL.EXC(matrix;quarto)**", onde 'matriz' é a coluna dos dados e 'quarto' é o quartil, ou seja, 1 para o 1o, 2 para o 2o, etc. Teste aí no seu Excel e me avise se não tiver conseguido que eu te ajudo!

**P7:** outra dúvida que tenho é que parece que no box plot, pelo menos nas figuras os quartis não são iguais.

Poderia me confirmar se o quartil 1 é de 0% a 25% o 2 de 26% a 50% o 3 de 51% a 75% e o quarto de 76% a 100%?

**R:** (Aline) Olá novamente Kelvin!

Temos na verdade, 3 quartis (Q1, Q2 e Q3) e com eles conseguimos dividir os nossos dados em 4 partes (até 25%, até 50%, até 75% e acima de 75%). Essa imagem ilustra a ideia: [https://sweet.ua.pt/pedrocruz/bioestatistica/\\_images/aed-pdf-111.png](https://sweet.ua.pt/pedrocruz/bioestatistica/_images/aed-pdf-111.png)

Normalmente quando a gente constrói um boxplot usando algum software/linguagem de programação, automaticamente os quartis são calculados, o mais importante acredito é realmente entender o que eles representam.

O boxplot reflete a distribuição dos dados, então nem sempre você verá um gráfico totalmente 'certinho', ou seja, com todas as suas partes simétricas. Como foi abordado na aula, se os seus dados são assimétricos, isso será refletido na forma do seu boxplot, mas a questão é que, os quartis sempre devem ser respeitados. Então, o Q1 é o valor que delimita 25% dos dados, o Q2 (a mediana), delimita 50% dos dados, e o Q3, delimita 75 % dos seus dados. Outra coisa que pode acontecer é quando algum dos quartis são iguais, isso gera um boxplot com aparência estranha (achatado por exemplo). Nesse post aqui há uma ótima explicação sobre esse assunto: <https://fernandafperes.com.br/blog/interpretacao-boxplot/>

Recomento também o artigo de blog da professora Fernanda, nele tem uma introdução do assunto e ensina como comparar um boxplot com o outro (ou quando você precisa analisar um grupo deles) :<https://blog.proffernandamaciel.com.br/como-ler-um-boxplot/>

## Tópico: outliers

**P1:** Professora, fiquei com uma dúvida na parte de **detecção de outliers**. Gostaria de saber porque utilizamos o fato " $1,5 \times \text{IQR}$ " para encontrar os limites superior e inferior. Por que não usamos " $1 \times \text{IQR}$ " ou " $2 \times \text{IQR}$ "? Seria uma convenção ou uma demonstração matemática nos faz chegar a esse valor?

**R:** Sim, tem uma explicação matemática em torno da fórmula. Ao usar o 1,5, ficamos bem próximos a 3 desvios-padrão da média. Na aula de regra empírica (próximo módulo), vamos cobrir essa parte de 3 desvios-padrão e por que depois disso podemos considerar possíveis outliers.

**P2:** Professora, parabéns pela clareza no curso.

Poderíamos dizer que a Média é uma medida de tendência central adequada para os casos analisados quando não detectamos a presença de **outliers** - através da fórmula " $\text{IQR}$ "? Ou uma coisa não implica na outra?

**R:** (Larrie) Olá, boa noite! Tudo bem?

Paulo, a média é sempre nossa primeira opção de medida de tendência central quando temos distribuições simétricas (que se aproximem de uma curva normal). Então, quando não temos outliers, talvez nossa distribuição será simétrica/normal ou algo próximo a isso, mas não é certeza. A gente não define simetria/normalidade com a presença ou não de outliers pela fórmula do IQR, importante lembrar isso!

>>>>>> Mas como saber se nossa distribuição é simétrica? <<<<<<<<<

Para saber se os dados se comportam de maneira normal, existem alguns métodos. Os testes de normalidade é um deles. Eles são utilizados para verificar se a distribuição de probabilidade associada a um conjunto de dados pode ser aproximada pela distribuição normal.

Resumidamente, o teste de Shapiro-Wilk é um teste específico para normalidade, usados mais em pequenas amostras, enquanto o método utilizado pelo Kolmogorov-Smirnov é mais geral e menos poderoso (o que significa que rejeita corretamente a hipótese nula de normalidade com menos frequência). Ambas as estatísticas tomam a normalidade como hipótese nula e estabelecem uma estatística de teste com base na amostra.

Apesar de muito utilizados, os testes estatísticos de normalidade vem sendo muito criticados, por perderem força com amostras muito grandes. Outro meio de avaliar a normalidade é por meio de um histograma, olhando se a distribuição dos dados no gráfico segue um formato de sino (a famosa distribuição gaussiana). Desde que o histograma não apresente inconsistências com a distribuição normal no histograma, é recomendável a avaliação dos estimadores de simetria e curtose, que representam aspectos ligados à forma do histograma: desviado para a esquerda/direita (simetria) ou apiculado/ achatado (curtose); ambas as medidas se aproximam do zero quando os dados são normais.

Outro gráfico que pode ser utilizado para avaliar a distribuição, é o gráfico Q-Q Plot. Diagramas quantil-quantil (diagramas Q-Q) são representações gráficas das proporções dos dados da amostra original em comparação com os quantis esperados para uma distribuição normal. Nesses casos, o diagrama Q-Q deve, idealmente, se apresentar como uma linha diagonal caso os dados sejam próximos à distribuição normal.

Consegui tirar sua dúvida? Estou a disposição :)

**P3:** Olá! No boxplot a mediana, Q1 e Q3 consideram ou desconsideram os **outliers**? Pelo que pareceu o Max e Min desconsideram os outliers correto?

**R:** (Larissa) Oi Caroline, boa pergunta. O dado é considerado outlier quando ele está acima ou abaixo do Q1 ou Q3;

**Cont:** Sim, mas ao calcular a mediana o excel, por exemplo, considera ou desconsidera os outliers na construção de um boxplot? Exemplo, em um boxplot a mediana é 10, esse número desconsiderou os outliers?

**R:** (Larissa) Oi Caroline, na minha primeira resposta informei que o que era considerado outlier estava abaixo ou acima do Q1 e do Q3, nesse caso, é abaixo do limite inferior ou acima do limite superior. Para encontrar um outlier você utiliza a fórmula abaixo:



$Q1 - 1,5 * IQR$  (para o limite inferior)  
 $Q3 + 1,5 * IQR$  (para o limite superior)

Fica mais claro de entender assim, né?

Outra coisa, em relação a sua segunda pergunta (sobre o excel e a mediana). Na verdade, a mediana não sofre tanto com a influência de valores outliers, então, talvez não faça sentido você considerar ou não os outliers quando for realizar essa métrica.

Veja  esse  exemplo:  
[https://www.canva.com/design/DAFJ\\_bGhESM/cwzw4r9YXC6M66m3jl5PkA/view?utm\\_content=DAFJ\\_bGhESM&utm\\_campaign=designshare&utm\\_medium=link2&utm\\_source=sharebutton](https://www.canva.com/design/DAFJ_bGhESM/cwzw4r9YXC6M66m3jl5PkA/view?utm_content=DAFJ_bGhESM&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton)

Temos dois conjunto de dados. O primeiro com valores próximos e o segundo adicionado um valor bem discrepante. Veja que mesmo com esse valor alto (que é 80), a média mudou bastante deixando de ser representativa pois não há nenhum indivíduo próximo dos 23 anos, porém a mediana ainda ficou representativa daqueles dados.

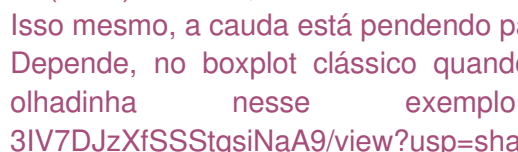
**P4:** Os **outliers** devem ser sempre excluídos das análises? Qual a melhor estratégia para identificá-los? Histogramas? testes de normalidade?

**R:** (Larissa) Ótima pergunta Kelli.

Uma das formas de identificar outliers é através da visualização dos dados através de um boxplot, histograma ou gráfico de dispersão. Sobre a sua primeira pergunta "Os outliers devem ser sempre excluídos das análises?" a resposta é depende. Se o seu conjunto de dados for grande o suficiente você pode eliminar esse valor, mas caso contrário você pode realizar uma transformação logarítmica ou utilizar a mediana em seus testes.

**P5:** Esse boxplot das atrizes é assimétrico negativo? A gente não considera os **outliers** como "cauda", certo?

**R:** (Aline) Bom dia, Amandha! Tudo bem? :)

Isso mesmo, a cauda está pendendo para a esquerda, portanto tem assimetria negativa. Depende, no boxplot clássico quando há outliers eles fazem parte da cauda, dá uma olhadinha  nesse exemplo: [https://drive.google.com/file/d/19fv\\_d1osmSn-3lV7DJzXfSSStqsiNaA9/view?usp=sharing](https://drive.google.com/file/d/19fv_d1osmSn-3lV7DJzXfSSStqsiNaA9/view?usp=sharing)

Aqui, como o comprimento da cauda direita é bem maior que a da esquerda nos indica que provavelmente há outliers, mas não temos como saber só pela visualização.

No vídeo a professora usou o boxplot modificado, que usa asteriscos para identificar os outliers, então fica bem separadinho a cauda (sem outliers) e os outliers representados separadamente.

Tem um resumo sobre esse assunto no blog da prof. Fernanda se você quiser conferir <https://blog.proffernandamaciel.com.br/como-ler-um-boxplot/> :)

**P6:** Na linguagem popular, ao falar sobre alguém extremamente bom em algo, fala-se "ele(a) é fora da curva"! Certamente isso veio da estatística. Tecnicamente, isso significaria um **outlier** de um gráfico fosse um bloxpot?

**R:** É isso mesmo, Jadson! Uma pessoa "fora da curva" vem do outlier ;)

**P7:** A lógica e equação para identificar os **outliers** são os mesmos para qualquer tipo de distribuição (normal, lognormal,...)?

**R:** (Letícia) Olá Jadson! Tudo bem? =D  
Sim, você pode usar a mesma lógica e equação.

**P8:** Para a construção do boxplot, consideramos todo o conjunto de dados (inclusive os outliers) para encontrar Q1, Q2, Q3 e IQR, e assim que obtermos esses respectivos valores podemos verificar quem é de fato **enquadrado como outlier**, certo?

**R:** (Aline) Oi, Ana! Bom dia, tudo bem? :)  
Exatamente, a sua descrição está correta! No início não tem como ter certeza absoluta se um valor é outlier ou não, então utilizamos todo o conjunto de dados para calcular os quartis e o intervalo interquartil (IQR). A detecção dos outliers no boxplot é possível justamente através das fórmulas:  
 $Q1 - 1.5 * IQR$ . (qualquer valor abaixo desse resultado é um outlier)  
 $Q3 + 1.5 * IQR$  (qualquer valor acima desse resultado é um outlier)

A professora Fernanda tem um vídeo bem informativo sobre outliers, depois dá uma olhadinha: <https://www.youtube.com/watch?v=o3uTAZyROI8>

Espero ter te ajudado :D

**P9:** A mediana nesse caso leva em consideração os outliers, certo? - Na análise dos ganhadores do Oscar

**R:** O Gabriela! A gente considera os outliers na contagem, mas o seu valor não entra no cálculo. Por exemplo, se tivermos 15 observações, a mediana vai ser o valor que está na posição 8, independente se tivermos outliers ou não. Por isso falamos que a mediana é uma medida que não considera os outliers (em relação ao seu valor).

**Tópico: histograma/ boxplot**

**P1:** Olá, professora! No vídeo, sua fala sempre relaciona **boxplot a histograma** (entendo que pelo contexto da aula), considerando o uso de classes. Mas no caso de variáveis discretas, com distribuição de frequência sem histograma (usando um simples gráfico de barras), também é comum o uso do boxplot para entendimento da variabilidade dos dados, certo??

**R:** Oi Pedro! Eu relaciono o histograma com o boxplot por serem usados quando temos uma variável quantitativa, portanto podemos usar com uma variável discreta sim.

**P2:** No ponto 12:35 da aula, a **distribuição é assimétrica positiva**, porém dá para confundir por que há a frequência relativa 0.22 antes de 0.44, e aí então que passa a diminuir. Num primeiro momento achei que fosse simétrica. Esta distribuição já não teria que começar a diminuir desde a primeira frequência, isto é, a primeira frequência relativa não teria que ter um valor maior do que 0.44 ?

**R:** Oi Fabricio, boa pergunta! Não precisa diminuir desde o primeiro intervalo. O que vai determinar a assimetria é a "cauda": se existe uma cauda (a frequência reduzindo) à direita, é assimétrica positiva. Esse gráfico foi um exemplo com poucos dados que realmente está quase simétrico, mas se você tem uma grande quantidade de dados, essa diferença fica mais evidente.

**P3:** Professora, na explicação comparando atores e atrizes do Oscar, você falou que os atores têm uma **distribuição "simétrica à direita"**. Pensei que os termos direita e esquerda seriam somente para a distribuição assimétrica.

**R:** Sim, é assimétrica!! :)

## **Tópico: histograma**

**P1:** Como eu defino um critério de agrupamento quando vou fazer um **histograma**?

**R:** Não existe um jeito certo ou uma fórmula. Na verdade, o critério de escolha é bem arbitrário, e por isto o histograma não é a melhor forma de visualizar a distribuição dos dados. Por exemplo, a forma que você escolhe para agrupar pode dar uma distribuição e a forma que eu escolho dar outra. Por isso, a melhor forma para visualização é o boxplot.

**P2:** Eu posso ter um **histograma para variáveis qualitativas categóricas**, caso eu defina a frequência de acontecimento delas dentro das observações e fazer a razão com o total de observações?

**R:** Boa pergunta! Na verdade o histograma só pode ser usado quando a variável é contínua. Ao tirar a frequência de variáveis categóricas, você estará na verdade fazendo um gráfico de barras.

É normal o gráfico de barras e o histograma serem confundidos por serem bem parecidos.

**P3:** Fiquei com dúvida no exercício 2 "Uma professora pergunta aos seus alunos quantas horas, em média, eles trabalham por semana. As respostas são: 21, 30, 0, 0, 15, 0, 15, 12, 10, 32.5, 0, 15, 10, 15, 5, 25. Um **histograma** começando no ponto 0 horas e usando largura da barra de 10 seria:"

Fiz o histograma de 0-10, 11-20, 21-30, 31 +. Há 7, 5, 3, 1 ocorrências em cada barra, respectivamente. Essa maneira de fazer não está correta? O histograma da resposta correta não representa essa distribuição. ?

**R:** (Larrie) @Edna B.Oii, tudo bem?

Os intervalos do histograma são: 0-9, 10-19, 20-29, 30-39.

Porque a 0-10 temos um intervalo com 11 números, o que não foi pedido na questão.

Quando fazemos dessa forma, a resposta fica correta!

**P4:** Os intervalos de cada barra são abertos? Por que na teoria nunca se terá um valor exato? Tipo o 400 no primeiro **histograma** da aula?

**R:** (Aline) Bom dia, Murilo! Como criador do histograma você decide qual será o intervalo de cada classe, não há teoricamente um intervalo certo ou errado porque depende dos dados que você está trabalhando. No histograma sobre os valores das casas, observamos que os intervalos começam entre 300 - 400 (mil dólares), porque casas mais baratas não eram de interesse e não foram consideradas no banco de dados.. E esse intervalo poderia ser diferente, de 50 em 50 (mil dólares) ou de 10 em 10 (mil dólares), por exemplo, depende de quão detalhado você quer que o histograma seja. De forma geral é bom se perguntar, a) qual a magnitude dos meus dados? b) quero visualizar uma distribuição geral ou mais detalhada dos meus dados?

Sugiro a leitura desse post do nosso blog: <https://blog.proffernandamaciel.com.br/como-ler-um-histograma/> em que é demonstrado como o histograma muda visualmente dependendo do intervalo de classe escolhida! :)

**P5:** Professora, neste tema de **histogramas**, há o conceito de curtose, que não domino bem. Será abordado adiante? O quanto este parâmetro é relevante nas análises de uma distribuição?

**R:** (Aline) Boa noite, Claudio! No momento não temos uma aula específica sobre curtose, mas consideramos as sugestões de todos vocês para aulas futuras! Além disso, os temas de assimetria e curtose serão abordados em breve no blog da prof. Fernanda :)

A curtose é apresentada em toda tabela com as estatísticas descritivas que você pede sobre um conjunto de dados e em poucos minutos você já consegue tirar algumas informações a partir dela. A curtose indica se os dados estão mais concentrados ou espalhados. Visualmente percebemos a curtose a partir do "achatamento" da curva de distribuição e numericamente o valor de referência vem da curva normal, cuja curtose  $K = 3$ . Caso você tenha um tempinho, recomendo muito essa vídeo aula: [https://www.youtube.com/watch?v=gNDNB\\_XVXiM&t=295s](https://www.youtube.com/watch?v=gNDNB_XVXiM&t=295s), em que é abordado em detalhes sobre o que é assimetria e curtose. Espero que seja útil para você!

**P6:** Professora, no segundo exercício da lista, o primeiro sobre **histogramas**, eu não entendi o porquê daquela resposta ser a certa. Pode me ajudar?

**R:** (Leticia) Boa noite Suelen!

No exercício as classes possuem 10 elementos, então como começamos do 0 fica da seguinte forma:

0-9 (5 elementos)

10-19 (7 elementos)

20-29 (2 elementos)

30-39 (2 elementos)

Sendo assim, a resposta correta é a letra B.

Espero que tenha te ajudado! :D

**P7:** Professora, e **histogramas declusterizados**? Só utilizo apenas se estiver trabalhando com dados espaciais?

**R:** Oi Marcelo, nunca trabalhei com histogramas declusterizados, mas já ouvi sobre sobre a relação com dados espaciais sim. Esse site pode te ajudar: <https://towardsdatascience.com/why-data-scientists-should-decluster-their-geospatial-datasets-8425b0b2453f>

**Cont:** Professora, sobre histogramas declusterizados, têm também essa explicação do Prof. Clayton Deutsch, da Universidade de Alberta, à respeito da utilização de histogramas declusterizados.

[http://claytonvdeutsch.com/wp-content/uploads/2021/02/CAG\\_Notes.pdf](http://claytonvdeutsch.com/wp-content/uploads/2021/02/CAG_Notes.pdf).

**P8:** Posso fazer um **histograma** para dados sem intervalos de classes?

**R:** (Larissa) Oi Mineia, ótima pergunta! Nem sempre nossos dados possuem intervalos de classe bem já definidos, mas você pode defini-los. Para isso você precisa:

1. Calcular a amplitude dos seus dados ( R: maior valor - menor valor)
2. Escolher o número de classes que vão ser utilizadas no histograma
3. Calcular o intervalo das classes (Se você escolheu 10 classes e a amplitude foi 50 cm, o intervalo seria 5 cm)

No geral, você tem que observar quais são seus dados para poder definir o melhor intervalo que vai ser representado no histograma.

**Cont:** Ok, então para o histograma preciso definir intervalos. Isso quer dizer que não são todos os dados que podem ser expressos através de um histograma? Por exemplo dados discretos não podem?

**R:** Podem sim, o histograma é utilizado para representar tantos dados discretos quanto dados contínuos.

**Cont:** Larissa, acabei de assistir a aula inaugural e a Fernanda falou que histograma é feito somente para dados contínuos.

**R:** Oii Mineia, tudo bem? Esse é mais um caso em que a Professora Fernanda fala que na prática é bem diferente. Em teoria você utiliza o histograma para dados contínuos, mas na prática se você conseguir analisar corretamente seus dados você consegue fazer um histograma de dados discretos.

**P9:** Bom dia, como eu faço esse **gráfico com a média, moda e mediana** ?

**R:** (Aline) Bom dia, Taiana! Tudo bem? :)

Uma boa maneira de comparar visualmente ao mesmo tempo a média, moda e mediana seria utilizando o histograma! Esse gráfico nos dá a distribuição dos dados e geralmente vamos ver que:

a) a média poderá estar localizada mais à esquerda ou à direita quando os dados são assimétricos. E quando os dados são simétricos (a curva normal em forma de sino), a média estará exatamente na barra central;

b) a moda é bem fácil de identificar porque ela é sempre a barra mais alta do histograma;

c) e a mediana como não é afetada pelos outliers geralmente aparece mais ao centro do histograma (entre a média e a moda) em distribuições assimétricas. E quando for a distribuição normal padrão, a mediana estará localizada exatamente ao meio porque ela será igual a média.

Segue uma imagem pra visualizar melhor:  
<https://drive.google.com/file/d/1QPriNty8yUp80QvIZkul8CoU7W7nEPKs/view?usp=sharing>

Há também o boxplot e o gráfico de violino que utilizam apenas a mediana. Pode parecer um pouco confuso agora, mas na próxima aula a professora Fernanda vai falar bastante sobre o histograma e o boxplot! Você consegue fazer esses gráficos no Excel, Power BI, Tableau e também em outros sites de visualização de dados, como o Flourish (<https://flourish.studio/>).

Consegui te ajudar? :D

### **Tópico: outros/exercícios**

**P1:** Boa noite. Estou fazendo o exercício de histograma numero 2 e não aparece nenhum desses que está na letra A, B e C. Está correto ou falta alguma resposta?

**R:** (Aline) Olá, Patricia! Está correto, o gabarito para essa questão é a letra B :)

Você pode seguir o seguinte passo a passo para resolver esse exercício:

Nesse exercício foi te dado a média de horas de trabalho/semana de um grupo de alunos. Em seguida foi pedido para você encontrar o histograma que melhor representa esse grupo de valores, correto? Então há basicamente 3 etapas:

1) O primeiro passo é definir o intervalo numérico de cada categoria que formarão as caixas (bins) do histograma. No exercício foi pedido para você começar com 0 e considerando uma largura de 10, ou seja, você só pode ter 10 números representados em cada categoria. Então seria:

\* 0 - 9

\* 10 - 19

\* 20 - 29

\* 30 - 39 (Podemos parar por aqui porque o maior valor nesse conjunto é 32,5)

2) Em seguida, precisamos verificar a frequência de valores em cada uma das categorias anteriores. É a frequência que indicará a altura de cada coluna no histograma. Nesse caso, basta organizar os dados e contar mesmo. Se você tivesse uma base de dados muito grande aí teria que usar alguma programa como o Excel/Power BI para criar alguma formatação/coluna condicional usando as categorias que a gente definiu no passo anterior. Então aqui, a gente tem:

\* 0 - 9 = frequência 5 (porque aqui se enquadram os valores 0,0,0,0,5)

\* 10 - 19 = frequência 7 (porque aqui se enquadram os valores 15,15,12,10,15,10,15)

\* 20 - 29 = frequência 2 (porque aqui se enquadram os valores 21 e 25)

\* 30 - 39 = frequência 2 (porque aqui se enquadram os valores 30 e 32,5)

3) Por fim, agora vamos ver qual histograma é o mais correto nesse caso. A escala usada nos histogramas é de 1 em 1, então a gente precisa encontrar um histograma com 4 colunas, em que a coluna mais alta é a de frequência 7, a segunda coluna mais alta é a de frequência 5, e as duas últimas colunas com frequência 2. Por isso, escolhemos a letra B). Uma outra forma mais rápida de encontrar é ver a quantidade de colunas. Nas letra A) e C) os histogramas tem 6 e 5 colunas respectivamente, então "só de olhar" a gente poderia eliminar essas opções.

Espero ter ajudado Patricia! =D

**P2:** Olá Fernanda e colegas! Eu não compreendi como se chegou na resposta do exercício 2 da lista (histograma).

Há uma concentração maior de alunos que trabalham 0 e 15 horas, com uma quantidade menor trabalhando no intervalos entre as duas e acima de 15 horas.

Nesse sentido, eu faria um histograma assimétrico à direita, onde a primeira barra seria a mais alta, junto com outra mais ao centro do histograma. Eu escolhi a resposta A, embora não fosse exatamente o que eu imaginei antes de analisar as opções.

**R:** Oi Daniel, você pode seguir o seguinte passo a passo para resolver esse exercício:

1) O primeiro passo é definir o intervalo numérico de cada categoria que formarão as caixas (bins) do histograma. No exercício foi pedido para você começar com 0 e considerando uma largura de 10, ou seja, você só pode ter 10 números representados em cada categoria. Então seria:

\* 0 - 9

- \* 10 - 19
- \* 20 - 29
- \* 30 - 39 (Podemos parar por aqui porque o maior valor nesse conjunto é 32,5)

2) Em seguida, precisamos verificar a frequência de valores em cada uma das categorias anteriores. É a frequência que indicará a altura de cada coluna no histograma. Nesse caso, basta organizar os dados e contar mesmo. Se você tivesse uma base de dados muito grande aí teria que usar alguma programa como o Excel/Power BI para criar alguma formatação/coluna condicional usando as categorias que a gente definiu no passo anterior. Então aqui, a gente tem:

- \* 0 - 9 = frequência 5 (porque aqui se enquadram os valores 0,0,0,0,5)
- \* 10 - 19 = frequência 7 (porque aqui se enquadram os valores 15,15,12,10,15,10,15)
- \* 20 - 29 = frequência 2 (porque aqui se enquadram os valores 21 e 25)
- \* 30 - 39 = frequência 2 (porque aqui se enquadram os valores 30 e 32,5)

3) Por fim, agora vamos ver qual histograma é o mais correto nesse caso. A escala usada nos histogramas é de 1 em 1, então a gente precisa encontrar um histograma com 4 colunas, em que a coluna mais alta é a de frequência 7, a segunda coluna mais alta é a de frequência 5, e as duas últimas colunas com frequência 2. Por isso, escolhemos a letra B). Uma outra forma mais rápida de encontrar é ver a quantidade de colunas. Nas letra A) e C) os histogramas tem 6 e 5 colunas respectivamente, então "só de olhar" a gente poderia eliminar essas opções.

**P3:** Bom dia Fernanda,

Nos gráficos bimodais ou multimodais, para cálculo da moda, será sempre o valor com maior frequência na mesma, certo? O interesse de saber se é bimodal ou multimodal será mais como informação para saber quais são as classes/grupos que mais se repetem no nosso conjunto de dados?

**R:** Oi Pedro,

Sim, nos gráficos bi e multimodais, verificamos as classes de maior frequência. Mas não precisa ser literalmente o número de modas. Por exemplo, um gráfico que tenha valores com frequências ao redor de 5, mas tenha um valor repetido 20 vezes e outro 25 vezes, seria um gráfico bimodal - mesmo que a moda aqui seja somente o valor que se repetiu 25 vezes. Avaliamos o formato, que tem dois picos.

**P4:** Boa noite, muito boa a aula. Senti falta de mostrar alguns cálculos e regras.

1 - Como determinar as classes, intervalos abertos e fechados.

2 - Como calcular o Q1 e Q3.

3 - Nos exemplos de histograma o final da classe e o mesmo valor do início da classe seguinte, por ex: 10-20, 20-30, 30-40. Porém nas explicações dos comentários sobre o exercício 2 está sendo orientado dividir as classes da seguinte forma 0-9, 10-19, 20-29. Quando usar cada formato? A segunda forma se dar por serem exclusivamente valores inteiros?



**R:** Oi Lucas, obrigada! Fico feliz que tenha gostado :)

Respondendo as suas dúvidas:

1 - Depende da sua área e objetivo. No dia a dia dos negócios, usamos um valor que "faz sentido", como no exemplo da aula que eu separei as notas em blocos de 10. Mas na teoria tem a regra de Sturges, que diz que o número ideal de bins (classes) é:  $\log_2(n) + 1$  (arredondando para cima).

2 - Existe uma continha chata que dá para fazer na mão, mas minha dica é usar no excel a função de quartil: "`=QUARTIL.EXC(matriz;quarto)`", onde 'matriz' é a coluna dos dados e 'quarto' é o quartil, ou seja, 1 para o 1o, 2 para o 2o, etc.

3 - Quando temos valores descritos como 10-20, 20-30, assumimos que o primeiro 20 é excluído e o segundo é incluído, então na verdade estamos descrevendo 10-19, 20-29 (se for discreto) ou 10-19.999999..., 20-29.999999... (se for contínuo). Se a variável for contínua, arredondamos para uma melhor visualização. Porém, depende da sua audiência. Para evitar ambiguidades, pode ser escrito como  $[10-20[$ ,  $[20-30[$ , que são os símbolos matemáticos para incluído e excluído, mas sua audiência pode não saber.

**P5:** olá! estou fazendo o exercício n.06 Modulo 1

Quando eu calculo o limite superior o valor apresentado é 115 mas as notas vão até 99, no cálculo isso pode acontecer? ou o resultado precisa ser até o limite dos dados? entendo que não devo considerá-lo como outlier, quanto ao outlier considere apenas o menor valor do conjunto de dados, confirmei no boxplot e foi o que apareceu também. Fiquei um pouco confusa nesta aula vou tentar absorver outros materiais colocados aqui nos comentários, se tiver mais algum que se encaixe na minha dúvida, pode me passar também. Obrigada!

**R:** Oi, Cristina! Tudo bem? :D

O seu cálculo está correto! O limite superior e inferior não precisam ser necessariamente valores que estão no seu conjunto de dados. Eles apenas representam valores de referência para você encontrar possíveis outliers.

O método de encontrar outliers através do boxplot usa como base o Intervalo Interquartil – IQR (também chamado de Amplitude Interquartil). Esse intervalo é justamente o terceiro quartil menos o primeiro quartil ( $Q3 - Q1$ ) e é exatamente o comprimento da caixa do boxplot.

O limite superior é o maior valor que não é outlier e para achá-lo você precisa fazer:

Limite superior =  $Q3 + (1,5 * IQR)$ . Qualquer valor acima do limite superior pode ser considerado um outlier.

Já o limite inferior é o menor valor que não é um outlier e você encontra ao fazer:

Limite inferior =  $Q1 - (1,5 * IQR)$ . Aqui, qualquer valor abaixo do limite inferior será considerado um outlier.

Caso os dados fiquem dentro dos limites significa que não há outliers. No exercício só temos um que é o valor 25, justamente por estar abaixo do limite inferior = 37.

Se puder conferir, temos dois posts no blog sobre o tema dessa aula, acredito que pode ajudar:

- Como ler um boxplot: <https://blog.proffernandamacieli.com.br/como-ler-um-boxplot/>
- Como ler um histograma: <https://blog.proffernandamacieli.com.br/como-ler-um-histograma/>

Fique à vontade para enviar uma nova mensagem caso ainda tenha dúvidas :)

**P6:** Tive bastante dificuldade para entender a questão número 2 da folha de exercícios, me explica, por favor?

**R:** (Aline) Oi, Kelvin! Tudo bem com você? Com certeza! :)

Nesse exercício foi te dado a média de horas de trabalho/semana de um grupo de alunos. Em seguida foi pedido para você encontrar o histograma que melhor representa esse grupo de valores, correto? Então há basicamente 3 etapas:

1) O primeiro passo é definir o intervalo numérico de cada categoria que formarão as caixas (bins) do histograma. No exercício foi pedido para você começar com 0 e considerando uma largura de 10, ou seja, você só pode ter 10 números representados em cada categoria. Então seria:

\* 0 - 9

\* 10 - 19

\* 20 - 29

\* 30 - 39 (Podemos parar por aqui porque o maior valor nesse conjunto é 32,5)

2) Em seguida, precisamos verificar a frequência de valores em cada uma das categorias anteriores. É a frequência que indicará a altura de cada coluna no histograma. Nesse caso, basta organizar os dados e contar mesmo. Se você tivesse uma base de dados muito grande aí teria que usar alguma programa como o Excel/Power BI para criar alguma formatação/coluna condicional usando as categorias que a gente definiu no passo anterior. Então aqui, a gente tem:

\* 0 - 9 = frequência 5 (porque aqui se enquadram os valores 0,0,0,0,5)

\* 10 - 19 = frequência 7 (porque aqui se enquadram os valores 15,15,12,10,15,10,15)

\* 20 - 29 = frequência 2 (porque aqui se enquadram os valores 21 e 25)

\* 30 - 39 = frequência 2 (porque aqui se enquadram os valores 30 e 32,5)

3) Por fim, agora vamos ver qual histograma é o mais correto nesse caso. A escala usada nos histogramas é de 1 em 1, então a gente precisa encontrar um histograma com 4 colunas, em que a coluna mais alta é a de frequência 7, a segunda coluna mais alta é a de frequência 5, e as duas últimas colunas com frequência 2. Por isso, escolhemos a letra B). Uma outra forma mais rápida de encontrar é ver a quantidade de colunas. Nas letra A) e C) os histogramas tem 6 e 5 colunas respectivamente, então "só de olhar" a gente poderia eliminar essas opções.

**P7:** Bastante didática e esclarecedora a parte do Histograma. Gostaria de deixar como sugestões 2 pontos que valeria a pena acrescentar: 1) Regra de Sturges e 2) A ordem é sempre média - mediana - moda ou para um lado ou para outro.

**R:** (Aline) Olá, Andréa! Ficamos muito felizes em saber disso! =D

1) Muito bem lembrado! A regra de Sturges normalmente é citada como uma forma de calcular a quantidade de classes no histograma. Para quem deseja saber mais, segue aqui um artigo simples sobre o tema: <https://www.statology.org/sturges-rule/> e também uma calculadora que calcula a quantidade de classes usando essa regra : <https://www.statology.org/sturges-rule-calculator/>

2) Isso mesmo!

\* Se for assimétrico para a esquerda, encontraremos a sequência: média (que é 'puxada' para a esquerda), a mediana, e a moda (que representa o ponto mais alto do histograma);

\* Se for assimétrico para a direita, será: a moda (ponto mais alto), mediana e a média (nesse caso, 'deslocada' para a direita);

\* E caso seja uma distribuição perfeitamente simétrica, exatamente ao meio estará a média, mediana e moda porque são iguais.

## Comentários Aula 04: Medidas de Dispersão

### Tópico: desvio padrão

**P1:** Professora, boa noite. Ao namorar os dados acabo observando que faz mais sentido usar a Méd. Geom. ao invés da Simples, a minha dúvida é se poderia fazer o **cálculo do DP** com a Xg ou teria que ser com a X (simples) msm?

**R:** Oi Álisson, existe uma medida chamada desvio padrão geométrico, que é o desvio padrão considerando o uso da média geométrica. Porém, a partir do módulo 3 quando entrarmos em inferência, será importante o uso da média e desvio padrão "normais", mas para uma análise exploratória, pode usar a geométrica sim.

Para mais informações: [http://en.wikipedia.org/wiki/Geometric\\_standard\\_deviation](http://en.wikipedia.org/wiki/Geometric_standard_deviation)

Aqui mostra como calcular no excel:  
[https://excelribbon.tips.net/T011208\\_Calculating\\_a\\_Geometric\\_Standard\\_Deviation.html](https://excelribbon.tips.net/T011208_Calculating_a_Geometric_Standard_Deviation.html)

**P2:** Oi Professora, boa noite! Tudo bem?

Poderia explicar por gentileza se existe **diferença entre desvio absoluto, desvio médio e desvio padrão**? Se sim, qual é? E quando usar cada um deles? Obrigada!

**R:** Oi Leticia! Até onde sei, existem o desvio absoluto médio e o desvio padrão. O desvio absoluto médio mede a distância de cada observação e a média, soma tudo e divide por n. O problema dessa medida (que não é muito utilizada) é que se você tiver uma média 5, e os valores 2 e 8, o cálculo seria  $(2-5) + (8-5)$  dividido por 2, mas o resultado seria  $-3+3 = 0$ . Ou seja, não tem desvio (o que está errado). Então como o positivo e o negativo se cancelam, usamos o desvio padrão, que é medido com a elevação ao quadrado, como foi ensinado nessa aula :)

**Cont:** Professora, fiquei curioso e resolvi procurar a demonstração da fórmula (DMA), A diferença da Var. em relação a Média é em módulo ( $|X_i - \bar{X}|$ ), que por definição é quase que improvável obter ZERO como resultado da operação DMA.

**R:** Sim, é isso mesmo, desculpe pela outra mensagem. Usamos o desvio padrão/ variância ao invés dessa medida, pois faz mais sentido matemático (para derivação da fórmula é melhor usar quadrado que valores absolutos).

**P3:** Oi Fernanda, bom dia, tudo bem?

Deixa eu aproveitar que o tema dessa aula para tirar uma dúvida contigo que eu sempre tive. É com relação ao desvio-padrão.

Eu entendi sua explicação, só não entendo **com INTERPRETAR um desvio-padrão**.

Já li e ouvi que desvio-padrão serve para verificar a homogeneidade dos dados: quanto mais longe de zero, menos homogêneo; quanto mais próximo, mais homogêneo.

No entanto, eu nunca consegui entender a partir de quanto é longe e partir de quanto não é.

Deixa eu dar um caso bem concreto: estou trabalhando na análise do curso do qual eu faço parte. Há um indicador que diz assim: "número de orientações CONCLUÍDAS por docente e desvio-padrão, a fim de verificar a homogeneidade da distribuição das orientações"

Pois bem, eu fiz o levantamento dos números de orientação por docente, e foram os seguintes:

Docente A: 5 orientações  
Docente B: 6 orientações  
Docente C: 4 orientações  
Docente D: 4 orientações  
Docente E: 3 orientações  
Docente F: 5 orientações  
Docente G: 2 orientações  
Docente H: 3 orientações  
Docente I: 4 orientações  
Docente J: 1 orientação  
Docente K: 2 orientações  
Docente L: 2 orientações  
Docente M: 7 orientações  
Docente N: 4 orientações

Pronto, diante disso encontrei um desvio-padrão de: 1,68.

Aí me ocorrem 3 perguntas:

- 1) Como interpreto isso? Posso dizer que a distribuição foi homogênea?
- 2) A partir de quanto não seria homogênea? A partir de "3" de "3,5" de "5,2"?
- 3) É preciso colocar algum tipo de unidade de medida? Ou seja, quando eu for escrever no relatório eu digo simplesmente que o desvio-padrão foi de 1,68 ou digo que foi de 1,68 "orientações"?

Tem outro indicador, ligeiramente diferente do que mencionei acima, que é o número de orientações EM ANDAMENTO.

Fiz as contas e, nesse caso, deu o desvio-padrão de 2,24.

Aí, a mesma dúvida, isso é uma distribuição homogênea ou não? Qual a referência para eu dizer se um desvio-padrão é homogêneo?

**R:** Oi Itamar, tudo bem?

Na verdade, o desvio padrão é uma medida do quanto as observações estão distantes da média. Apesar do senso de homogeneidade fazer sentido, não é isso que o desvio padrão mede. Portanto, não existe um "corte" onde eu possa ver até que ponto os dados são homogêneos ou heterogêneos.

Tendo dito isso, o desvio padrão não é bem uma medida que seja interpretada, já que tudo que ela mede é a dispersão em volta da média. Mas você poderia utilizar para comparar a dispersidade entre dois conjuntos de dados, por exemplo: o conjunto de dados 1 tem o desvio

padrão de 4,5, com as observações mais dispersas do que o conjunto de dados 2, que tem o desvio padrão de 2,3.

Quanto à unidade de medidas, o desvio padrão tem a mesma do que você está medindo, ou seja, no seu exemplo seriam 1,68 orientações.

**Cont:** Obrigado pelo retorno Fernanda. É impressão minha ou a utilidade do desvio-padrão está em comparar dois conjuntos?

**R:** Não somente! Veremos em outras aulas como a combinação da média com o desvio padrão é importante para várias análises estatísticas (além da aplicação em fórmulas como intervalo de confiança e teste de hipóteses).

**Cont:** Ajudando um pouco, o desvio padrão é usado muito (ou pelo menos foi usado muito) no que se chamava CEP - Controle Estatístico de Processo. É a tal regra 68-95-99,7. Quando começaram a fazer controle de qualidade do processo e não mais dos produtos (só separar o que está fora da especificação), adotaram a regra dos 3 sigma - os 99,7. Por essa regra, seria estatisticamente esperado de um processo que se encontrasse ok que produzisse uma não conformidade a cada 370 itens (0,27%). Isso foi lá nos anos 60 no Japão e foi responsável pela recuperação da indústria no pós guerra. Quando parei de mexer com isso, lá no fim do século passado (olha quem é o senhor idoso aqui), já se falava em 4 sigma (1 falha a cada 1000 itens). Em resumo, vc tinha uma linha do tempo com retas representando a média e a média  $\pm 1s$ ,  $2s$  e  $3s$ . E aí alimentava com as medidas periódicas. A depender das alterações (havia um monte de regrinhas), afirmava-se que o processo estava sob controle ou que havia uma alteração nele. A professora pode confirmar, mas as regrinhas nada mais eram do que formas mais simples do que fazer teste de hipótese para saber se a média das novas medidas tinha se desviado da média histórica.

**R:** Obrigada pelo comentário, William! Vamos aprender o que é a regra empírica (68-95-99,7) no módulo 3.

**P4:** Olá professora, ainda estou na dúvida sobre como eu interpreto o **desvio padrão**, o que ele significa no momento da minha análise descritiva.

No seu exemplo, tenho uma média de 500 mil reais e um desvio padrão de 50 mil reais, mas o que isso significa? Que a distância dos meus dados é por volta de 50 mil, seria isso? Se eu tivesse um desvio de 100 mil reais, os meus dados estão mais dispersos que no primeiro caso? E qual a importância de eu entender a dispersão dos dados nesse momento da análise descritiva?

Até hoje a única coisa que eu uso o desvio padrão é para equalizar dados com medidas diferentes para poder compará-los

**R:** Olá Daniela! Tudo bem?

Então, sobre o que é desvio padrão:

O desvio padrão é uma medida que expressa o grau de dispersão de um conjunto de dados. Ou seja, o desvio padrão indica o quanto um conjunto de dados é uniforme. Quanto mais próximo de 0 for o desvio padrão, mais homogêneo são os dados.

Então se você tiver um desvio padrão de 50 mil, quer dizer que os seus dados se distanciam em média da sua tendência central de 50 mil unidades. Se você tem um desvio padrão de 100 mil e outro de 50 mil, o de 100 mil possui dados mais dispersos sim.

A importância de ver sua dispersão é realmente saber se seus dados são mais homogêneos ou heterogêneos. Por exemplo, você trabalha em uma fábrica de produzir sucos. Se a gente pegar 10 garrafas de suco e fizer uma média, digamos que tenha 500 ml ( o que é previsto na embalagem ). Mas seus clientes estão reclamando que algumas garrafas estão vindo vazias demais ou cheias demais. Nesse momento você pode se perguntar: mas a média não tá certa? 500ml? Sim, está. Porém possui uma variação muito grande na quantidade de suco que vem, o que significa uma perda do processo de controle de qualidade! Esse é um exemplo dentre muitos onde o desvio padrão pode ser aplicado. Em aulas futuras você pode ver também que se eu tenho uma distribuição normal, dá para conhecer/prever toda (isso mesmo, toda) sua amostra a partir do desvio padrão e da média!

Ficou mais claro para você?

**P5:** Olá. Eu fiquei em dúvida, de como o **desvio padrão** soluciona o problema de outliers, sendo que eles estão sendo considerados no momento em que calculamos a média.

**R:** Oi Ana, o desvio padrão não soluciona o problema de outliers, mas podemos identificá-los quando temos a média e o desvio padrão, pela regra empírica ou z-score (que vemos no módulo 2).

Eu falei um pouco sobre outliers no encontro inaugural, como um complemento dessas aulas. A aula gravada está dentro do módulo de encontros :)

**P6:** Boa noite! Gostaria de uma ajuda com relação a outliers. Tenho um dataset com dados de mais de 15 anos, sendo que o **desvio padrão**, assimetria positiva, curtose Leptocúrtica, sendo todas essas medidas bastante grande. Mesmo aplicando a regra de  $3,0 \times$  o desvio padrão X IQR, acabo perdendo muitos dados. Qual seria a saída?

**R:** Oi Marcos, por que você quer retirar os outliers? Na maioria dos casos, a gente não os elimina, a não ser que sejam erros

**Cont:** Na verdade não quero retirar. Fiz essa pergunta, pois gostaria de saber qual é o procedimento caso haja a necessidade de serem erros mesmo. Já que para calcular , através dos testes, seja Shapiro, seja Kolmogorov-Smirnov é aplicado nos resíduos, correto?

**R:** Se forem erros, podem sim ser retirados. Os testes mencionados são testes de normalidade de resíduos que fazemos na análise de regressão linear (mas não necessariamente tem a ver com outliers, então não entendi muito bem).

**Cont:** Sim Prof.Fernanda, foi falha minha de não mencionar que todo o conteúdo desta discussão é para se fazer uma regressão linear

**R:** Para a regressão linear: quando temos outliers, fazemos uma análise com eles e outra retirando eles. Então, comparamos os dois modelos (usando o R2 e o se). Nem sempre temos um modelo melhor quando retiramos os outliers, então essa é uma análise que pode ser feita.

**P7:** algumas dúvidas:

1) como interpretar um desvio padrão maior que a média?

2) quando se fala que uma coisa dista X desvio pad4oes de outra, a que isso se refere do ponto de vista estatístico? aos dados já normalizados?

**R:** Boa tarde, Hugo! Peço desculpas pela demora em te responder, você trouxe um questionamento interessante na primeira pergunta que eu ainda não havia lido a respeito :)

1)Essa comparação relativa entre a média e o desvio pode ser vista por exemplo no coeficiente de variação ( $C.V = \text{desvio}/\text{média} * 100$ ). Nesse caso, se o nosso numerador é maior, o coeficiente de variação sempre será menor, indicando mais homogeneidade dos dados. Mas falando de uma interpretação mais direta, acredito ser importante não perder de vista que o desvio padrão, independente se é maior ou menor que a média, sempre trará a magnitude da dispersão dos dados, se estão concentrados ou não. E isso pode ter diferentes significados a depender de quais dados estamos trabalhando. Essa tema estava sendo debatido aqui: <https://www.quora.com/What-does-it-mean-when-the-standard-deviation-is-higher-than-the-mean-What-does-that-tell-you-about-the-data> e deixo aqui um trecho que pode ser útil:

" Digamos que seus dados representem distâncias acima e abaixo do nível do mar. Sua média neste caso pode ser zero - nível do mar - e seu desvio padrão pode ser 20 pés. Isso indicaria que a maioria de suas medições está entre 20 pés acima e 20 pés abaixo do nível do mar. Por outro lado, e se seus dados representassem as idades dos moradores de um condomínio em Palm Beach? Nesse caso, sua média pode ser 85 e seu desvio padrão pode ser 10, indicando que a maioria dos residentes tem entre 75 e 95 anos.

No primeiro caso, o desvio padrão é maior que a média. No segundo caso, é menor. Mas, em última análise, seu tamanho relativo pouco importa - é o que eles dizem sobre a estrutura dos dados, a maneira como são distribuídos, que é importante. Usando essas informações, você pode começar a fazer inferências sobre os dados. Por exemplo, no primeiro conjunto de dados, você poderia determinar se acima do nível do mar, um certo ponto estava significativamente mais elevado do que todos os outros, ou seja, se representava uma anomalia estatística que valia a pena investigar - com base em quantos desvios padrão da média ele estava localizado"

2) É importante lembrar que quando você ler que um dado está a 'x desvios padrão' de distância, será sempre em relação à média daquela amostra ou população, porque é uma medida que sempre trará a dispersão dos dados em relação ao seu centro, representado pela média. Além disso, o desvio padrão é uma medida de dispersão que se aplica para qualquer distribuição, não apenas a dados normalmente distribuídos. No caso da distribuição normal padrão, se destaca a Regra Empírica que nos traz com exatidão a distância em desvios padrão que os dados estarão da média. No caso: 68% dos dados estão a 1 desvio padrão de distância, 95% estão a 2 desvios padrão de distância e 99% estará até 3 desvios padrão de distância. Minha resposta faz sentido pra sua pergunta, Hugo? :)



## Tópico: variância e desvio padrão

**P1:** Uma dúvida: pelo valor da amplitude não conseguimos saber bem sobre a dispersão dos dados. Isso ficou claro para mim o motivo. Mas, apenas olhando o valor da **variância** ou, principalmente, do **desvio**, de que forma eles nos dizem algo sobre a dispersão? Obrigada desde já!

**R:** Oi Silvia! O desvio padrão por si só realmente não trás muita informação. Mas é uma medida de dispersão muito usada em outras fórmulas e aplicações. Ficará mais claro nos próximos módulos, como por exemplo quando falamos da distribuição normal e a regra empírica ou o z-score.

**P2:** Olá. É correto interpretar que a **variância** é apenas um passo para o cálculo do **desvio** padrão, mas ela em si não tem aplicação direta, por se tratar de uma unidade de medida diferente (unidade ao quadrado) da unidade de medida dos dados (por exemplo, como citado, idade<sup>2</sup>)?

**R:** (Larissa) Oi Márcio, tudo bem? O desvio padrão e a variância são medidas de dispersão, ou seja, são parâmetros utilizados na estatística para calcular o quanto os dados de um conjunto de valores podem variar. Sendo que a variância representa o quão distante os valores estão da média, só que as vezes a variância por si só não mostra realmente como está a distribuição daqueles dados quando existem outliers influenciando a média. Nesse caso, entra o desvio padrão para solucionar esse problema. Pense no desvio padrão como é o “erro” se quiséssemos substituir um dos valores coletados pelo valor da média. Entendeu a diferença?

**Cont:** Oi, Larissa, obrigado por responder. Ainda assim não consigo imaginar uma aplicação prática / interpretação da variância. Seria possível me dar um exemplo? Obrigado.

**R:** (Fernanda) Oi Márcio, tanto a variância quanto o desvio padrão não têm uma aplicação prática ou interpretação direta. Mas são medidas muito úteis para outras análises, principalmente quando usamos a inferência estatística. Eu falei sobre isso na seção FAQ do nosso encontro inaugural, dá uma olhadinha lá, está dentro do módulo de encontros ;)

**P3:** Excelente aula, professora!

Na resposta do exercício 7 - c, a justificativa para indicar qual o fundo foi mais arriscado é a **variância**, porém, gostaria de saber se posso justificar em cima do **desvio padrão** visto que são diretamente proporcionais ou é mais adequado falar da variância mesmo ?

**R:** Pode analisar pelo desvio padrão também! :)

**P4:** Olá!

O que significa o n na fórmula da **variância**.  
Não consegui pegar.

**R:** Olá, Cláudia! Tudo bem? :)

O 'n' minúsculo se refere ao tamanho da amostra. Se for 'N' maiúsculo é o tamanho da população. Temos um post de blog sobre o tema, recomendo a leitura [https://blog.proffernandamaciel.com.br/medidas\\_dispersao/](https://blog.proffernandamaciel.com.br/medidas_dispersao/)

**P5:** Olá, professora. Tudo bem? Três dúvidas aqui.

1) O **desvio padrão** serve somente para "normalizar" a **variância**? Dado que, como mencionou, se aplica o desvio padrão à variância para tirar o "elevado ao quadrado" e permitir uma melhor leitura do estudo... é isso?

2) Para que eu entenda melhor gostaria de fornecer um exemplo. Tenho dados da variação da bolsa de valores dos últimos 10 anos. Faço, portanto, os devidos cálculos para ter a média, variância e o desvio padrão. Existe alguma interpretação de probabilidade dentro do desvio padrão? Usando da distribuição normal como exemplo, teria então 68% de chance de ocorrer as variações "inputadas" no estudo até 1 desvio padrão, 95% de chance de acontecer algo que englobe até 2 desvios padrões e o que exceder demais seria chance de calda, outliers?

3) Faz sentido/posso usar a mediana ao invés da média para extrair variância e em seguida o desvio padrão para dados muito dispersos? (Desculpe se for uma pergunta muito sem sentido essa kkkkk)

**R: (Aline)** Boa tarde, Jefferson! Vamos para suas dúvidas ;D

1) É importante ter em mente que tanto a variância quanto o desvio padrão tem exatamente a mesma função, que é indicar a dispersão dos dados, e no caso do preço de ações, seria a volatilidade desses ativos. A vantagem do desvio padrão é justamente que ele nos traz essa variação na mesma unidade de medida da média e não elevado ao quadrado (que é matematicamente necessário para não anular o cálculo da variância), mas que não nos ajuda a interpretar a realidade. Imagine um relatório dizendo que a variação no preço de uma ação na última semana foi de "R\$ 72,5 reais elevado ao quadrado", é um pouco difícil de trabalhar com dados nesse magnitude.

2) Isso mesmo! Se você estiver trabalhando com dados que seguem uma distribuição normal, então você pode aplicar a chamada Regra Empírica (que será tratada nas próximas aulas) que nos dá todas as possibilidades de variação naquela distribuição, em que até 99,7% dos dados estarão dispersos entre 1 e 3 desvios-padrão. Segue um vídeo bem curtinho da prof<sup>a</sup> Fernanda sobre o tema: <https://www.youtube.com/watch?v=JXs5VBePCwE> . Para outras distribuições, pesquise sobre o Teorema de Chebyshev , ele não é tema do curso mas pode ser útil/interessante de aprender também :)

3) De forma alguma Jefferson! Essa pergunta é muito válida sim! Mas eu receio que a explicação é mais matemática do que 'estatística' propriamente dita e necessita de entender alguns outros conceitos, como o do método dos mínimos quadrados. Nesse link a seguir, há uma resposta com indicação de outros links para aprofundar no assunto:

<https://stats.stackexchange.com/questions/177052/using-the-median-for-calculating-variance?rq=1> De forma bem simplificada, a média seria uma medida que minimiza os erros no cálculo da variância. Já a mediana é mais utilizada em uma medida de dispersão chamada 'Median Absolute Deviation' (MAD), para casos em que se há muitos outliers no conjunto de dados. Em português eu não encontrei muita informação a respeito, mas esse vídeo em inglês traz um exemplo <https://www.youtube.com/watch?v=V4Vq1krcUmE> e uma interpretação. Pode ser interessante para o seu caso!

**Cont: (Aline)** Oi Jefferson! A resposta da Aline ficou bem completa, mas eu queria complementar algo sobre o ponto 3. Não podemos usar mediana com variância, como ela disse, por uma questão matemática mesmo.

A mediana geralmente vem acompanhada do IQR, que é aquele calculo para o boxplot - tanto que no boxplot utilizamos a mediana como medida central, já que é um gráfico de posição.

Ou seja, temos aqui 2 casais: média + variância (ou o desvio padrão) e mediana + IQR (ou o MAD).

### Tópico: amplitude

**P1:** Professora, para o **cálculo da amplitude** os valores máximos e mínimos são os mesmos que o do boxplot? Pois nele os outliers estão fora desse limite

**R:** Oi Matheus, ótima pergunta. Para a amplitude, usamos o máximo - mínimo. No boxplot, o primeiro valor é o mínimo e o último é o máximo, caso não haja outliers. Se houver, o "max" passa a ser o limite superior e/ou o "min" passa a ser o limite inferior (você encontra esses valores usando a fórmula para detecção de outliers). Ou seja, para calcular a amplitude de um boxplot, você vai olhar o valor máximo mesmo, nem que seja um outlier.

**P2:** Oi, Larissa. Obrigado pelo retorno.

Eu quero dizer assim, digamos que seja exigido o informe da **amplitude**. Eu faço o processo que você falou primeiro e depois informo a amplitude, com outliers excluídos ou informo a amplitude considerando os outliers. Não sei se existe uma regra, um padrão.

**R:** (Letícia) Oi Jadson! tudo bem?

Você pode calcular a amplitude, tratar ou excluir os outliers conforme a Larissa orientou, e após isso você apresenta o informe com os dados já tratados.

**P3:** É correto eliminar os outliers para observar a **amplitude** depois, ou ela deve contemplar sempre os outliers?

**R:** (Larissa) Oi Jadson, tudo bem? O ideal é utilizar o intervalo interquartil para observar a presença de outliers. Caso você tenha outliers em sua amostra, você deve realizar o tratamento desses outliers ou a remoção deles. E posteriormente você pode verificar o intervalo interquartil novamente para verificar se ainda existem esses outliers.

**Cont:** Mas se precisarmos informar a amplitude devemos considerar o valor mínimo e máximo, mesmo que sejam outliers, certo?

**R:** (Letícia) Isso mesmo Marília, para o cálculo da amplitude é necessário calcular a diferença entre o maior valor e o menor valor do seu conjunto de dados. =D

### Tópico: outros

**P1:** Fernanda, algumas dúvidas práticas que sempre me "atortentam", pois nas análises ocorrem frequentemente. Comecei a fazer meu projeto e a variável idade (de gestantes) pelo teste de Shapiro-Wilk mostra que não existe normalidade ( $p < 0,05$ ). A análise da assimetria (0,83) e curtose (3,11) mostram valores aceitáveis de normalidade. O Box Plot mostra a presença de outlier. O histograma mostra assimetria à direita. Considerando isso:

- 1) Considero a variável normal ou não? Em minha prática eu iria optar por trabalhar com dados de mediana e intervalo interquartil. Até porque eu não sei como excluir os outliers de forma segura. kkkk
- 2) Caso não seja normal, como sei quais dados devo excluir para eliminar os outliers?
- 3) Já recebi a orientação que a variável idade pode ser sempre usada como tendo distribuição normal. Existe algum tipo de critério ou isso não faz sentido?

-Abaixo as saídas do Stata.

```
. swilk idade
```

#### Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
idade	360	0.93347	16.669	6.662	0.00000

```
. sum idade, d
```

idade					
Percentiles		Smallest			
1%	19.49076	18.8501			
5%	20.1848	19.28542			
10%	20.73785	19.32375	Obs	360	
25%	22.33402	19.49076	Sum of Wgt.	360	
-----					
50%	25.35524		Mean	26.40251	
		Largest	Std. Dev.	5.113515	

75%	29.73854	40.78302		
90%	33.48118	41.99042	Variance	26.14803
95%	36.34908	42.45037	Skewness	.8330088
99%	40.78302	42.77892	Kurtosis	3.114008

**R:** Oi novamente, Lalucha! Mais uma vez peço desculpas pela demora, mas suas perguntas não são triviais (o que é muito legal)! Espero que meus comentários te ajudem :D

1) O teste de Shapiro-Wilk é bem robusto e pelos resultados dele, você pode considerar que os dados não são normais. Na minha pesquisa, eu encontrei um pergunta muito similar a sua aqui (<https://stats.stackexchange.com/questions/129417/if-my-histogram-shows-a-bell-shaped-curve-can-i-say-my-data-is-normally-distrib#:~:text=Age%20can%20not%20be%20from,bell%2Dshaped%20distributions%20out%20there.>) Nesse caso, a pessoa descreve que também está lidando com a variável idade e pelo histograma parecia que era normal, mas quando foi feito o teste (Kolmogorov-Smirnov), foi detectado que não há normalidade. Na explicação oferecida, é lembrado que há outras distribuições que podem ter um formato parecido com um 'sino' e também serem simétricas como a normal. Avaliar o histograma pode nos dar uma pista, mas não é suficiente para a gente 'bater o martelo' e dizer se há normalidade ou não. O mesmo eu diria para a assimetria e a curtose (elas nos dizem que os dados são aproximadamente simétricos / possuem leve assimetria e com um 'pico' mais elevado do que a normal).

Outro ponto é que se a idade fosse normal padrão, significaria que teríamos idades negativas, o que não faz sentido. Acredito que independente da variável, o ideal é ter o respaldo de algum teste que é reconhecido, como o do Shapiro Wilk =)

Outro ponto é que se a idade fosse normal padrão, significaria que teríamos idades negativas, o que não faz sentido. Acredito que independente da variável, o ideal é ter o respaldo de algum teste que é reconhecido, como o do Shapiro Wilk =)

2) Acredito que a primeira coisa a se pensar é: você precisa que os dados sejam normais? Por exemplo, você precisará realizar algum tipo de teste e qual seria o tipo de teste que irá fazer depois? Aqui no curso, no módulo 10, "Projetos e Apresentações", tem uma aula, na verdade é um PDF, chamado 'Roadmap – qual teste usar'? E nele a profª Fernanda indica alguns testes possíveis dependendo do seu objetivo, dá uma olhadinha se puder :). Se for necessário fazer uma regressão, o pressuposto seria a normalidade dos resíduos, então a princípio, os dados não seguirem a distribuição normal não é um problema.

Uma vez que você já saiba que tipo de análise/teste gostaria de fazer e sendo os dados não normalmente distribuídos, duas sugestões seriam:

1) Verificar o teste não-paramétrico equivalente ao teste que você queria fazer. Testes não-paramétricos não pedem o pressuposto de normalidade e não são influenciados pelos outliers).

2) Outra opção seria calcular o logaritmo da variável. Ao fazer isso, esse nova variável 'logaritmizada', apresenta uma distribuição normal e você pode continuar a análise do jeito que tinha pensado (seguindo o pressuposto da normalidade). Nesse vídeo é mostrado como fazer esse cálculo usando o SPSS, mas imagino que exista essa opção no Stata também : ) <https://www.youtube.com/watch?v=RWP1xmPenAI>

3) Sobre os outliers, eles podem ser a causa dos dados não serem normais, mas não é a única. O único caso que posso dizer que se pode realmente deletar os outliers sem receio seriam se eles fossem resultado de algum erro de digitação/erro de medição, aí não teria dúvidas.

Se puder Lalucha, assista ao vídeo que a profª Fernanda fez sobre como lidar com outliers: [https://www.youtube.com/watch?v=\\_YWAb-RHL8g](https://www.youtube.com/watch?v=_YWAb-RHL8g). De forma geral, é necessário:

- Primeiro verificar se eles não foram resultado de algum erro de digitação ou de medição;
- Verificar a magnitude de outliers em relação à quantidade de dados no total (em porcentagem). Se forem muitos outliers e eles forem retirados, possivelmente vai afetar os resultados da análise, então seria bom ter cautela nesse caso (e possivelmente não deletá-los);
- Se a quantidade de outliers não é tão significativa assim, em seguida, você poderia retirar todos os outliers e comparar os resultados (olhar as medidas descritivas por exemplo, média, mediana, desvio padrão). Se não houver alteração significativa, então poderia considerar deletá-los, mas é em último caso mesmo.

De forma geral, veja qual seria o próximo passo da sua análise e se o pressuposto da normalidade é realmente importante.

3)

Lalucha, nos meus estudos eu já vi citarem o peso e a altura como aproximadamente normalmente distribuídos, mas como eu mencionei, se a idade fosse normalmente distribuída poderia ter o caso de 'idades negativas', então não acho que seja algo a se esperar que aconteça. Acredito que o mais certo é a gente recorrer aos testes pra nos ajudar a chegar numa conclusão :)

## Comentários Aula 05: Covariância e Correlação

### Tópico: causalidade

**P1:** Professora, fiquei com dúvida sobre Correlação x Causalidade, existe algum método que comprove a **causalidade** entre variáveis? Ou isso é algo plenamente "respondido" somente pela análise dos dados gerais?

**R:** Existem métodos para provar causalidade, mas eles são bem complexos e não vemos neste curso. É importante entender que a correlação não nos mostra causalidade. Alguns métodos para verificar causalidade são Variáveis Instrumentais (modelo econométrico) e estudos de experimentos.

**P2:** Muito bacana esse vídeo. A curiosidade sobre as relações de causalidade das variáveis da natureza/vida talvez tenha sido a grande motivação dos seres humanos no desenvolvimento da vida.

Tem um livro muito interessante sobre a **causalidade**. "O livro do porquê: a nova ciência de causa e efeito" de Judea Pearl e Dana Mackenzie

**R:** E até hoje a causalidade é muito estudada, pois não é fácil de ser medida. Ainda temos curiosidades sobre a relação entre muitos fatos...

Obrigada pela dica! Já estudei muito pelos artigos do Judea Pearl, mas não conheço esse livro, vou olhar!

**P3:** Sempre escutei a expressão "relação não é **causalidade**!" e quando vi a explicação do tubarão e o site do Tyler Vigen ficou muito mais claro de compreender.

**R:** Isso aí, Deyvid! O pessoal adora esses exemplos, fica bem mais claro, né? :)

### Tópico: correlação

**P1:** Ótima aula, prof. Fernanda. Quanto a **correlação**, tenho uma dúvida constante: quando encontra-se uma correlação muito próxima a 1 ou -1, eu devo rever novamente os meus dados e o número de observações para que não haja qualquer resultado enviesado? (levando em consideração dados e pesquisas na área de business). Muito obrigado.

**R:** É muito difícil na "vida real" termos correlações próximos a -1 ou 1, então é bom ver se os dados não são um subconjunto um do outro, ou calculados de fórmulas parecidas. Se forem dados independentes, não há viés, mas é bom saber dessa correlação caso vá fazer uma modelagem (por exemplo, na regressão não podemos usar dados altamente correlacionados como variáveis independentes, pois aí sim haverá um viés na sua análise).

**P2:** Boa tarde Professora Fernanda. Um verdadeiro prazer em lhe conhecer. Estou gostando demais das suas aulas. MUITÍSSIMO obrigado mesmo pelo excelente conteúdo e a sua forma tão didática de explicar as coisas.

Eu gostaria, se me permite, lhe fazer uma pergunta. Como eu posso determinar a **correlação** quando se trata de várias variáveis que afetam a uma variável final, devo determinar a correlação de uma a uma com a variável principal, ou posso determinar uma correlação geral? Por exemplo, vamos supor que eu vendo Suco de Limão (eis aqui o meu "Y") e que o preço final do meu suco depende de o preço da água mineral (X1), do limão (X2) e do açúcar (X3). Minha correlação a ser calculada a partir do Excel, por exemplo, eu teria que fazer de Y com X1, de Y com X2 e assim por diante, certo?? Ou seria melhor considerar uma média ponderada a partir da composição da minha receita??

Desde já, muito obrigado

**R:** Oi Roberto, obrigada pelo carinho! Fico feliz que esteja gostando das aulas!

Quando a sua pergunta, a correlação só se faz entre 2 variáveis, portanto sim, você deve fazer o Y com X1, Y com X2, etc. Porém a melhor análise para o seu caso seria a regressão múltipla, assim você pode ver exatamente a influência de cada variável no seu Y, além de poder prever o preço do Y dado os preços de X1, X2 e X3.

**P3:** Olá. Existe algum parâmetro/escala para interpretação da correlação? Podemos adotar algo do tipo:

+1 = positivo perfeito

+0,7 a 0,99 = positivo muito forte

0,5 a 0,69 = positivo forte

0,3 a 0,49 = positivo moderado

0,1 a 0,29 = positivo fraco

0,01 a 0,09 = positivo muito fraco

0 nenhum

-0,01 a 0,09 = negativa muito fraco ... etc

**R:** (Letícia) Olá Márcio! Tudo bem? Você pode adotar conforme abaixo:

-1 = negativa perfeita

-0,9 = negativa alta

-0,5 = negativa moderada

0 = sem correlação

0,5 = positiva moderada

0,9 = positiva alta

1 = positiva perfeita

Se o resultado que você obter estiver em algum dos intervalos, você pode usar da aproximação.



**P4:** Boa tarde, gostaria de ter uma ideia de criar uma **correlação** entre a quantidade de vendas ou preço dos produtos e quantidade de vendas. Há essa possibilidade de construir isso?

**R:** Com certeza, Fyllype! Na aula de Excel eu mostro como fazer isso. Use esses seus dados e replique o que ensino na aula, depois me conte o que encontrou :)

**P5:** Boa noite!

Eu poderia fazer a correlação apenas entre variáveis ordinais com ordinais? Se tenho variáveis nominais (mesmo que estejam representadas por números) não tenho como relaciona-las ou tem alguma forma? Fiquei bem confusa.

Exemplo 1, onde acho que se aplica a correlação :

Em escala de 1 a 10 o quanto você se acha informado sobre AI?

Em escala de 1 a 10 quanto você acha que a AI poderia ser usada no processo de educação?

Exemplo 2, onde não consigo enxergar uma forma de correlacionar, já que os números são apenas representação numérica de uma escala nominal:

Em escala de 1 a 10 o quanto você se acha informado sobre AI?

Quando você pensa em AI o que você sente? ( 1 opção escolhida por entrevistado (ID)

1 - curiosidade

2-Medo

3-Indiferença

4-Confiança

segue uma pequena amostra da minha base

"ID"	"Q1 - information about AI "	"Q5.Feelings"	"Q7.Utility_grade"
1	8	1	9
2	7	1	6
3	5	1	6
4	5	1	9
5	4	1	8
6	5	2	6
7	7	1	10
8	6	1	8
9	6	1	8
10	4	1	7
11	4	1	10
12	6	1	4
13	9	4	10
14	9	1	9
15	7	3	10
16	6	3	4
17	1	3	3
18	8	2	8

**R:** Olá, Cristina! Como vai? :)

Para fazer a correlação de Pearson (tema da aula), as variáveis precisam ser quantitativas contínuas e também há alguns requisitos, por exemplo, que haja linearidade entre elas. Então se temos alguma variável nominal (mesmo que representada por números) não é indicado utilizar a Correlação de Pearson.

Nos exemplos que você trouxe, eu percebi que todas as variáveis são ordinais, correto? Nesse caso é possível usar a Correlação de Spearman (há uma aula no módulo 9) ou Correlação de Kendall, ambas são não paramétricas e também pedem alguns pressupostos, por exemplo, que a relação entre as variáveis seja monotônica (isto é, que 1) à medida que o valor de uma variável aumenta, o mesmo acontece com o valor da outra variável; ou 2) à medida que o valor de uma variável aumenta, o valor da outra variável diminui.).

\* Nesse link (em inglês) há um resumo muito bom sobre os três tipos de correlação: <https://datascience.stackexchange.com/questions/64260/pearson-vs-spearman-vs-kendall>

\* Nesse vídeo (em português) há uma demonstração de como fazer no R todas as três correlações [https://www.youtube.com/watch?v=7UGWOHF8k0Q&ab\\_channel=FernandaPeres](https://www.youtube.com/watch?v=7UGWOHF8k0Q&ab_channel=FernandaPeres) que mencionei!

Espero que te ajude, Cristina!

**P6:** Boa noite! Estou amando suas aulas profa. Fernanda! Só queria ter te conhecido quando eu estava quebrando a cabeça no meu doutorado.

Tentei fazer uma **correlação** com a tabela abaixo, mas o Excel fica dizendo que tem dados não numéricos. Então como fazer para correlacionar o IQA (índice de qualidade de água) de uma bacia hidrográfica no semiárido com os períodos seco e chuvoso abaixo?

IQA	Chuvoso	Seco
Ótima	0	1
Boa	7	7
Aceitável	4	6
Ruim	4	5
Péssima	0	0

**R:** Oi, Érika! Tudo bem? Em nome da profª Fernanda e equipe te agradeço muito por esse retorno tão positivo! É uma satisfação saber que o curso tem sido tão proveitoso para você! :D

Sobre sua pergunta, o IQA está como uma variável qualitativa ordinal e as outras duas "Chuvoso" e "Seco" são quantitativas, mas não consigo dizer se são ordinais ou contínuas apenas pelos dados que você trouxe.

No caso da Correlação Linear de Pearson que é o tema da aula, só é possível calculá-la se todas as variáveis são quantitativas contínuas, então nesse caso:

1) Colocando a variável IQA como quantitativa ordinal e se a variável "Chuvoso" e "Seco" também forem ordinais, poderia ser utilizado a Correlação de Spearman ou de Kendall (no módulo 9, há uma aula sobre a Correlação de Spearman);

2) Se considerar a variável IQA na sua forma padrão, isto é qualitativa, então nesse caso, quando temos a junção de uma variável qualitativa nominal com várias categorias e uma variável quantitativa (variáveis "Chuvoso" e "Seco", não calculamos a correlação, mas fazemos uma comparação entre médias dos grupos/categorias através da ANOVA. Esse assunto é o tópico do módulo 8, tudo bem? =)

**P7:** Olá! No exemplo do sorvete e do tubarão, onde há uma **correlação**, mas não uma relação de causa, necessariamente há uma terceira variável que conecta as duas? Nesse caso seria o aumento da temperatura, mas é possível que não exista uma outra variável e seja apenas coincidência? Como descobrir se há uma terceira variável que é responsável pela correlação entre outras duas? Obrigado!

**R:** Olá, Alexandre! Tudo bem?

A princípio, apenas pela correlação não podemos afirmar se há uma terceira variável ou se seria apenas coincidência (a temperatura pode ser essa terceira variável, mas seria necessário fazer um estudo sobre isso).

Mas como pode haver uma terceira ou mais variáveis que estão influenciando as variáveis correlacionadas é por isso que não podemos falar em causalidade. Através da correlação também não seria possível definir o sentido da causalidade (se o consumo de sorvete causa o ataque de tubarões ou vice-versa).

O estudo de causalidade é algo bem mais complexo, vou deixar algumas referências para você dar uma olhadinha:

Na área da Economia/Econometria temos:

- Variáveis Instrumentais, nesse vídeo há uma explicação geral sobre o tema (em português): [https://www.youtube.com/watch?v=I5ZNIMHoxzM&ab\\_channel=CanalDoPorQu%C3%A](https://www.youtube.com/watch?v=I5ZNIMHoxzM&ab_channel=CanalDoPorQu%C3%A)  
A%3F

- Temos também o trabalho do Clive Granger que ganhou um Nobel de Economia, ele trabalhou com causalidade em séries temporais: [https://www.youtube.com/watch?v=kJcvjPMHk58&ab\\_channel=CanalDoPorQu%C3%A](https://www.youtube.com/watch?v=kJcvjPMHk58&ab_channel=CanalDoPorQu%C3%A)  
A%3F

- Temos também o estudo de causalidade no contexto de experimentos [https://www.youtube.com/watch?v=w7du7FicPQo&ab\\_channel=Cient%C3%ADstica%26PodcastNaruhodo](https://www.youtube.com/watch?v=w7du7FicPQo&ab_channel=Cient%C3%ADstica%26PodcastNaruhodo) esse vídeo é uma aula do curso II de Estatística Aplicada à Psicobiologia da Unifesp (recomendo muito o curso, inclusive!)

E alguns materiais de leitura:

- 'The Book of Why' - <https://www.amazon.com.br/Book-Why-Science-Cause-Effect/dp/1541698967> - infelizmente não há uma edição em português brasileiro (só Portuguesa) talvez seria mais interessante achar a versão em inglês mesmo;

- Há o livro 'Causal Inference - What If' - É gratuito e disponível aqui: <https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/>;

- E também um artigo falando sobre o tema de forma geral: Methods and tools for causal discovery and causal inference:  
<https://wires.onlinelibrary.wiley.com/doi/epdf/10.1002/widm.1449>

**P8:** Olá !! Dúvida sobre **Correlação**: Considerando que a Correlação está relacionada a duas variáveis caso eu queira representar essas duas variáveis em um gráfico para eu demonstrar uma suposta correlação (também por imagem), onde as linhas subiriam e desceriam em movimentos parecidos, como devo proceder com os dados caso as duas variáveis tenham grandezas muito distantes, para que ambas apareçam no gráfico ?

Ex: Var1=12, Var2=14.200...

Sabendo que no gráfico, as Var1, estarão quase sobre a linha de Zero pois serem muito inferiores.

**R:** Bom dia, Henry! Tudo bem? :)

Para mostrar as linhas das duas variáveis ao mesmo tempo da forma que você mencionou, poderia ser feito um gráfico de linha com eixo secundário. E na construção desse gráfico é preciso ajustar as escalas porque provavelmente uma das variáveis não precisaria começar em 0 exatamente. Vou deixar um tutorial para Excel que mostra exatamente como fazer isso: <https://www.youtube.com/watch?v=Pf2owArn0TE>

Uma sugestão que te dou é também fazer o gráfico de dispersão que plota as duas variáveis juntas na forma de pontos e é bem fácil de ver visualmente se há algum tipo de correlação entre elas. Vou deixar um tutorial também para Excel: <https://www.youtube.com/watch?v=L9rxESk8ApA> . Esse vídeo faz parte do post de blog sobre esse tipo de gráfico, recomendo a leitura!

Espero ter te ajudado :)

### **Tópico: correlação espúria**

**P1:** Professora, a **correlação espúria** seria o mesmo que causalidade então?

R: Não, pelo contrário! A correlação espúria ocorre quando duas coisas parecem ter uma relação mas não têm. Pode acontecer por coincidência, como no site que eu mostro na aula, ou por um terceiro fator que impacta os dois analisados, como o exemplo do sorvete e tubarão.

Somente observando a correlação, não temos como encontrar causalidade. Pode até haver, mas a correlação não nos diz isso.

**P2:** Sempre que vejo esses exemplos de **correlação espúria**, eu lembrei do caso do jogador de futebol Ramsey kkkkk. Diversas vezes que ele marcou um gol, um famoso morreu! Pura coincidência kkk tem até meme com isso.

Tem até artigo explicando a correlação usando ele como exemplo kkk

esse é o meme: <https://knowyourmeme.com/memes/the-ramsey-effect>

Artigos: <https://saneeya.wordpress.com/tag/aaron-ramsey/> <https://competence.rosengroup.com/mod/page/view.php?id=4549>

**P3:** ótimas explicações. Há algum tipo de testes que possa ser aplicados para identificar **correlações espúrias**?

**R:** Oi Robson,

Infelizmente não tem um teste que prove se a correlação é espúria ou não. É algo percebido pelo senso comum ou conhecimento do tópico/teoria.

Eu gosto de um trecho de um post que fala sobre isso: "Geralmente é difícil diagnosticar uma correlação espúria, uma vez que a teoria de uma pessoa é a teoria da conspiração ou coincidência de outra pessoa. O exemplo recente mais famoso disso foi o debate sobre se o aquecimento global é uma consequência das ações humanas ou não. No século 20, um debate semelhante ocorreu sobre se o uso do tabaco causava câncer de pulmão.

A principal ferramenta para diagnosticar se uma correlação é espúria ou não é examinar a qualidade da teoria por trás dela. No caso do tabaco e do câncer de pulmão, apenas uma explicação clara para o mecanismo biológico que levou o fumo ao câncer de pulmão resolveu o debate.

Uma abordagem mais orientada por dados para diagnosticar a correlação espúria é usar técnicas estatísticas para examinar os resíduos. Se os resíduos apresentarem autocorrelação, isso sugere que alguma variável-chave pode estar faltando na análise."

Lembrando que a análise sugerida não detecta correlação espúria, mas sinaliza que uma variável pode estar faltando (como pode não estar). Só nos dá uma pista.

O texto completo está aqui: <https://www.displayr.com/learn-what-is-spurious-correlation/#:~:text=A%20more%20data%2Ddriven%20approach,be%20missing%20from%20the%20analysis.>

### **Tópico: covariância**

**P1:** Professora a correlação é dada pelo  $r$  que varia de -1 a +1. E a **covariância** é dada pelo que?

**R:** O símbolo da covariância é um sigma e varia de - infinito a + infinito. Mas a covariância não é muito usada, já que só nos diz a direção da associação e não quão forte ou fraca ela é. A correlação é recomendada por dar as duas informações.

**Cont:** Grato! Pelo q entendi a **correlação** é usada para variáveis quantitativas, existem análogos para variáveis qualitativas e para qualitativas com quantitativas?

**R:** Não, mas existe correlação para dados ordinais (Spearman), que é uma correlação não paramétrica. A análise de quanti com quali é a ANOVA que será visto no módulo 8.

**P2:** Outra dúvida, é sobre as equações para calcular a covariância e a correlação. Fui pesquisar, mas achei mais de uma e não sei dizer qual é a certa. Você poderia me indicar um vídeo ou link para saber mais sobre? Obrigado!

**R:** Olá, Alexandre! Claro! :)

Segue as fórmulas apresentadas no livro "Manual de Análise de Dados" (Fávero). Você pode ver os prints que tirei do livro pelos links:

- **Fórmula da Covariância:**  
<https://drive.google.com/file/d/1PweR3lpy5ipMPoKdyaRIIb30cPWKDWj3/view?usp=sharing>

- **Fórmula da Correlação de Pearson:**  
<https://drive.google.com/file/d/1EAMQjN4oi9v0V7dAFTto4ERfYvcI9icTp/view?usp=sharing>

## **Tópico: outros**

**P1:** a diferença entre os pontos de dispersão do gráfico e a reta representam o **erro padrão**? se sim, os valores individuais de cada erro padrão pode indicar um possível outlier? (considerando um p-value: 0,05)

**R:** Essa diferença é um resíduo e não erro padrão (veremos no módulo de Regressão) ;) E não identificamos outliers dos resíduos.

**P2:** Um caso muito citado fala sobre um supermercado que posicionou pacotes de fraldas perto de cervejas pois concluiu que homens que iam comprar fraldas para seus bebês também compravam cerveja. Isto seria um exemplo de correlação somente? Ou daria pra dizer que se trata de causalidade(acredito que não) ?

**R:** Esse caso é um exemplo de market basket analysis (análise de associação) que é visto em data mining. Se trata de medir a correlação para entender quais itens são vendidos juntos. A história é que fizeram essa análise no Walmart e encontraram que cerveja e fraldas eram compradas juntas, como um "combo", e interpretaram que já que os pais não podem mais sair para um bar por causa do bebê, eles estavam comprando mais cerveja ao comprar fraldas. As duas vendas estão associadas, mas não temos como determinar uma causalidade (eles foram comprar fralda e lembraram da cerveja ou foram comprar cerveja e já que tavam lá resolveram levar mais um pacote de fralda para casa?).

Porém, apesar de muito citada, um colega que já trabalhou no Walmart disse que logisticamente é impossível um mercado posicionar fralda na área que vende álcool, e também não teria sentido colocar cerveja na área de bebês. Eu também já ouvi falar que essa história é uma lenda urbana (foi inventada para fins ilustrativos).

**P3:** Boa tarde ... Por favor, a regressão múltipla está contemplada nesse curso ? Se não, teria algum material para me indicar ? Desde já, agradeço.

**R:** (Aline) Boa tarde, Fábio! Obrigada você por perguntar! :D

Nos módulos 6 e 7 do curso você encontrará todo o conteúdo relacionado à regressão linear e logística. Regressão linear múltipla é o tópico da aula 4 do módulo 6!

**P4:** Sabe o que gostaria de entender ... Sobre os principais tipos de distribuição e suas aplicações. Estou com dificuldades de encontrar algo confiável na internet. Outro dia, o professor mostrou um gráfico sobre a renda em alguns países, e ele disse "esse gráfico segue a distribuição Chi Quadrado". O que ele queria dizer ?

**R:** (Larrie) Oi, tudo bem Fábio?

Esse é um assunto bem extenso, sabia? Acho que não consigo te responder de maneira objetiva e efetiva em apenas um comentário :/

Mas, te indico um vídeo muito bom da UNIVESP que fala das principais distribuições de probabilidade:

<https://www.youtube.com/watch?v=j3Zbup0KMxY>

Também fica ligado que no blog da Prof. Fernanda, vamos abordar individualmente cada tipo de distribuição, ok?

Qualquer dúvida, estou a disposição.

**P5:** Olá! Ainda falando sobre variabilidade (não sei se esse é o termo correto para o que estou procurando). Existe alguma relação entre variabilidade e tamanho de amostra? Se sim, como isso afeta a análise e os resultados? Obrigada.

**R:** (Letícia) Oi Angellica, tudo bem?

Chamamos de variabilidade a quantidade de conjuntos e dados diferentes dentro do seu banco de dados, caso seja esse mesmo o conceito que esteja pensando, existe sim relação com o tamanho da amostra, pois quanto maior o tamanho da amostra maior a variabilidade.

E quanto maior o banco de dados melhor fica a sua análise, pois a análise da amostra vai se tornando mais precisa.

Espero que tenha te ajudado :)

## Comentários: Aplicando com o Excel

### Tópico: excel

**P1:** Olá Fernanda, um bom aproveitamento que eu obtive nesta aula, foi fazendo os cálculos estatísticos através de fórmula, comparando com o resultado obtido do plug-in Análise de Dados. Um fundamento importante nela foi a respeito da **moda** nas notas das meninas. Se observamos, veremos que se trata de multimodal (76,89,84,88,74), porém a Análise de Dados traz 76 no campo Modo. Bom, já vemos que tem algo errado aí, pois se a assimetria é negativa, a teoria diz que a moda deveria ser maior do que a média (81,68), mas na verdade, isso acontece porque este campo foi programado para fazer a fórmula Modo, esta fórmula traz a primeira moda encontrada dentro do seu range, no caso 76. O correto aqui seria utilizar a fórmula modo.multi de forma matricial, pois assim ele traria todas as modas das notas das meninas. Abs!!

**R:** @Gustavo S.Excelente, Gustavo! Primeiro, muito bom ir testando os cálculos para comparar. Também uma ótima dica sobre a função "modo.mult". Quando temos dados quantitativos, a moda não é uma medida tão interessante quando a média e mediana, mas sempre bom notar esses detalhes. Obrigada pela colaboração!

**P2:** Professora, boa noite. Gostei da aula, foi bem clara. Eu fiquei com uma dúvida. A **correlação** (pelo plug-in para as análises de dados) é a correlação paramétrica de Pearson (olhei em ajuda). No entanto, os dados precisam ser normais, certo? Tem como fazer a correlação de Spearman (não paramétrico) no Excel? Não sei se terá isso mais a frente na disciplina. Obrigado!

**R:** @Carlos C.Sim, essa é a correlação de Pearson que é a "clássica". O Excel ainda não faz análises não paramétricas (só se fizer "na mão", criando tabelinhas e fórmulas) e no curso não veremos esse tipo de análise. Eu tenho um vídeo ensinando a ideia do Spearman, como fazer "na mão" e usando o R Studio, dá uma olhada: <https://youtu.be/vWCOZ1P8uiQ>

**P3:** Outra dúvida: o excel tem alguma função que permite juntar as colunas do gráfico fazendo com que este fique com mais "cara" de **histograma**? Muito obrigado.

**R:** @Gilmar J.Você pode também fazer o gráfico do histograma diretamente em inserir > gráficos. Desta forma as colunas ficam juntinhas.

**P4:** Adorei saber desse resumo na ferramenta análise de dados, sempre usei a **função** na célula "=média(B1:B12)" para calcular essas medidas descritivas. Vem até mesmo o erro padrão! Aprendendo sempre, esse é o lema!!!

**R:** @ElisiaEsse plugin é maravilhoso, facilita a vida sem ter que fazer um monte de fórmula! Só não entendi ainda por que já não vem incluído automaticamente no Excel... É tão escondido que ninguém sabe, mas estou aqui para isto ;)



**P5:** Olá, Fernanda! Na construção do **histograma** vi que vc usou uma qtde de classes e uma amplitude aleatórias. O mais indicado seria calcular a quantidade de classes e a amplitude? Ou isso depende do interesse do pesquisador?

**R:** @Camila P.Oi Camila,

Em teoria, o número de classes seria a raiz quadrada no número de observações. E para saber a amplitude, basta calcular o valor máximo - o valor mínimo e dividir pelo número de classes. Isso faria com que, por exemplo, as notas das meninas fosse dividida a cada 8: 62-70, 70-78, 78-86... Você pode usar desta forma, mas a ideia é que sejam exemplos menos "teóricos" e mais aplicados a vida real. Não faz mais sentido visualizar de 10 em 10 neste caso? Então depende do seu problema ou interesse. E para visualizar a distribuição, o boxplot é mais recomendado, mesmo que se faça o histograma com as classes separadas da forma "correta" pela teoria ;)

**P6:** Uma aplicação importantíssima da **COVARIÂNCIA** pode ser vista na teoria das finanças, Mathematical Financial Optimization, publicado no Portifólio Selection Efficient Diversification of Investments, de Harry Markowitz, prêmio Nobel de economia em 1990.

**R:** Oi, Cláudio, obrigada pela contribuição. Nesse caso, tanto a covariância quanto a correlação podem ser usadas, já que a correlação é uma função monotônica da covariância

**P7:** Professora, sobre a primeira planilha que tem uma aba de Salários de professores, seria possível calcular a **correlação** entre o salário e o gênero e/ou faixa etária? Faria sentido essa análise para identificar possíveis discrepâncias de salário entre gêneros ou faixa etária?

**R:** @Mariana B.Pode calcular a correlação entre duas variáveis contínuas, lembrando de verificar se existe uma relação linear entre elas. Ou seja, não tem como fazer com gênero, mas salário e idade sim (se tiverem uma relação linear).

Mas essa análise não mostra discrepância, mostraria se o salário aumenta conforme a idade, e quão forte é essa relação. Acho que a análise que você quer verificar é o que iremos aprender no módulo de regressão ;)

**P8:** Estou adorando as aulas. Está sendo uma boa oportunidade para aprender e revisar algumas coisas. Fiquei com duas dúvidas a partir dos exemplos de aplicação no Excel da parte 1:

1) Sobre o **número de classes e tamanhos dos intervalos**:

Você escolheu intervalos a partir de 50, com amplitude 9 para cada intervalo, para mostrar como construir o histograma na parte 1 da aplicação no Excel. Sei que foi apenas um exemplo prático, mas gostaria de aproveitar para tirar uma dúvida que, acho, é até mais profissional do que técnica: Aprendi que, para definir o número de classes (intervalos) e sua amplitude, deveríamos (1) aplicar a regra de Sturges ( $k = 1 + 3,3\log(n)$ , sendo  $k$  o número de classes e  $n$  o número total de dados) e (2) definir a amplitude de cada classe com a divisão da amplitude total das frequências pelo número de classes encontrado. Na prática, no dia-a-dia, isso é muito cobrado? Ou prevalece mesmo o bom-senso? Os "seniors" cobram esse tipo de coisa?

Ou se recorre a tal procedimento apenas quando o volume dos dados nos deixa um pouco mais "perdidos"?

2) No exemplo do **histograma** das notas e meninos, posso dizer que ele é bimodal, porque há uma variação positiva na cauda, no intervalo "61-70"?

**R:** (Larrie) @Bruno G.Oiii, tudo bem Bruno?

Então, a regra de Sturges é conhecida porém não muito aplicada na prática. No dia a dia, raramente eu aplico, mas nas minhas provas de estatística, eu tenho que aplicar. Eu recorro mais a isso quando tenho muitos dados. Outra regra que pode ser aplicada é tirar a raiz quadrada do tamanho da amostra. Eu particularmente não consideraria bimodal pq os picos não são da mesma altura. Bimodal tem que ter dois picos distintos de alturas parecidas.

**Cont:** @Larrie M.Olá, Larrie, muitíssimo obrigado pelas respostas! Quando você fala de aplicar em provas, é contexto de faculdade, né?

**R:** @Bruno G.Oii! Isso mesmo :)

**P9:** Bom dia. Na **correlação**, no exemplo apresentado, devemos observar a quantidade da amostra para saber se o valor é estatisticamente aceito, correto?

**R:** (Larrie) Boa tarde! Tudo bem?

O que quer dizer com estatisticamente aceito?  $p < 0,05$ ?

Bem, não é apenas olhando o coeficiente de correlação, nem olhando ele + amostra. Em alguns outros programas como SPSS, STATA e Excel, o coeficiente de correlação já vem junto com o seu respectivo valor de p.

Já da para ter uma noção básica pelo próprio valor do coeficiente, quanto mais próximo de 1, mais provável que o valor de p seja  $< 0,05$

Para calcular o valor de p do coeficiente de correlação vc vai usar a seguinte formula:  
$$= \text{TDIST} ( ( \text{pearson\_cell} * \text{sqrt} ( N-2 ) ) / \text{sqrt} ( 1 - ( * \text{pearson\_cell} \text{pearson\_cell} ) ) ) , N, 2)$$

Esse site explica melhor como fazer isso:

<http://ptcomputador.com/Software/microsoft-access/135546.html>

Ele explica direitinho como fazer, parece um pouco difícil mas não é. Eu particularmente prefiro usar o SPSS para achar o valor de p do coeficiente de correlação.

**P10:** Olá, sou novo no curso, por isso não sei se mais na frente será mostrado, mas acredito que seria interessante também mostrar um pouco de estatística descritiva bivariada e multivariada. Seria interessante. Fica a sugestão caso, não exista nenhuma aula falando sobre isso. Também gostaria de saber, entender melhor como interpretar aquele valor -0.7 da **assimetria**, o valor em si. Varia de quanto a quanto? Essas coisas. E gostaria, também, de entender a **curtose**. Obrigado.

**R:** (Aline) Bom dia, Leandro!

Muito obrigada pela dica! A sua sugestão sobre estatística descritiva bivariada/multivariada abrange conceitos de vetor de médias e matrix de covariância, por exemplo?

Nos próximos módulos do curso a professora Fernanda apresenta análises bivariadas e multivariadas nas aulas de correlação e regressão linear simples e múltipla :)

Tanto a assimetria e a curtose nos ajudam a perceber a forma da distribuição dos dados. Então, em uma distribuição perfeitamente simétrica, como a normal, os dados se concentram na média e a frequência deles diminui a medida que distanciamos dela, seria portanto assimetria igual a zero.

Mas na vida real a gente provavelmente vai encontrar distribuições assimétricas, tem dois casos principais:

a) assimetria positiva ou à direita, os dados se concentram abaixo da média e a gente percebe uma cauda à direita,

b) assimetria negativa ou à esquerda, os dados se concentram acima da média e a cauda à esquerda.

Então para interpretar o número que indica assimetria, verifique duas coisas:

1) o sinal do número: se positivo, assimetria à direita, se negativo, assimetria à esquerda;

2) se o número está perto de zero ou não, o ideal é que a assimetria fique no intervalo entre -1 e +1, quanto mais perto de zero, mais simétrico e portanto, a média é mais representativa.

No exemplo, foi -0,7, então é assimetria à esquerda e os dados apresentam uma assimetria aceitável.

A curtose também nos ajuda a perceber o quanto os dados estão concentrados ou espalhados, visualmente a gente percebe se a curva da distribuição fica mais "achatada" ou não. A curtose pode ser grande: dados concentrados em torno da média e grau de "achatamento" baixo ou pequena: dados espalhados e grau de "achatamento" alto. É desejável ter curtoses elevadas, porque isso nos indica que os dados estão próximos da média e portanto ela é uma boa medida descritiva para os dados. A curva normal tem curtose  $K = 3$  e esse número pode ser uma referência. Eu recomendo muito que você assista essa aula: [https://youtu.be/gNDNB\\_XVXiM](https://youtu.be/gNDNB_XVXiM), é bem grandinha mas vale a pena!

**P11:** Olá. Na edição do intervalos (min 14:58 video parte 1), a primeira classe vai até 50 (inclusive)? Seria mais correto anotar  $\leq 50$ ?

**R:** (Letícia) Olá Márcio, tudo bem?

Sim, está correto!

Você pode anotar dessa forma.

**Cont:** também achei oportuno o  $\leq 50$ .

**R:** (Letícia) Olá Carlos, boa noite! Tudo bem? Sim, você pode escrever dessa forma sem problemas, a notação é correta.

**P12:** Boa noite. Sou leigo e estou tendo o primeiro contato com a análise de dados. Me ocorreu uma dúvida sobre qual função do Excel seja mais adequada, é uma dúvida bem

simples. Agradeço se alguém puder explicar a diferença e qual aplicabilidade das funções disponíveis (uso Excel - Microsoft 365):

I) Desvpad.A

II) Desvpad.P

III) Desvpada

IV) Desvpadpa

V) DESVPAD (Aqui neste tem até um sinal de alerta sobre compatibilidade entre versões do Excel)

VI) DESVPADP (com o mesmo sinal de alerta)

VII) Bddesvpa

Desde já deixo meu agradecimento e estou gostando do conteúdo. Parabéns Prof<sup>a</sup>.

**R:** Oi Plínio, obrigada e fico feliz que esteja gostando!

Quanto a sua dúvida, o Excel tem 2 fórmulas base para desvio padrão: uma considerando o cálculo da amostra (I, III, V - fazem a mesma coisa) e outra considerando a população (II, IV, VI). A função bddesvpa eu não conhecia, mas pelo que vi aqui também considera o desvio padrão da amostra.

Recomendo assistir a aula de análise de dados no Excel, onde eu mostro como fazer a estatística descritiva de uma forma bem simples ao invés de usar as fórmulas :)

**P13:** Olá prof @Fernanda M. tive dificuldade no exercício 7 para achar os valores de desvio padrão e correlação, não sei se são os números negativos. Não consegui chegar no resultado.

**R:** Oi Danielle, sem problemas, eu te ajudo! Você está usando o plugin "análise de dados"? Nele você clica em "estatística descritiva" e ele te dá todas as informações descritivas como média e desvio padrão. Depois você clica de novo no plugin para a correlação (alternativamente, pode usar a função =correl e selecionar cada coluna). Tenta dessa forma e me avise se deu certo! :)

**P14:** Boa noite, Prof e tutores.

No exercício minha moda das meninas é 89. Olhando os dados vi que é Multimodal. Estou certo?

**R:** Oi Vinicius, boa tarde! Os dados são multimodais sim. Acontece que o Excel informa somente o primeiro valor encontrado, logo, olhar a moda usando essa tabela não é o ideal.

De qualquer forma, quando trabalhamos com dados numéricos, é mais comum usarmos média ou mediana, enquanto usamos a moda para variáveis categóricas.

### Tópico: dúvidas fim de bloco (exercícios e dúvidas gerais)

**P1:** Professora e o que significa a **curtose** na análise descritiva? Para que ela serve?

**R:** @AUGUSTO G. Muita gente fala que mede quão "curta" a curva da distribuição é, mas essa interpretação está errada. A curtose é uma medida de quão grossa é a cauda da distribuição (que veremos no próximo módulo), e é calculada para ver a possibilidade de outliers, mas não é uma medida muito usada por ser simplista (por exemplo, o teste de normalidade Jarque-Bera).

**P2:** Acho que estou em apuros para fazer meu **projeto**. Aparentemente entendi todas as aulas do módulo um, mas não sei como aplicá-las com meus dados.

Meus dados não são simples como os apresentados nos exemplos das aulas. Cito um exemplo (acho que o mais simples que tenho): Desenvolvi 6 pré-misturas para pães (enriquecidas com 6 farinhas diferentes) e mais o controle (totalizando 7) e tive que fazer análise de vida de prateleira ao longo de 6 meses. Desses sete produtos, verifiquei como o índice de oxidação lipídica se comportou, como os ácidos graxos se comportaram (vários ácidos graxos, como a umidade se comportou, a atividade de água (dentro outros itens). Não faço ideia de como fazer. Não é algo simples como "notas de meninos e meninas". Queria muito uma luz.

**R:** Olá Midori,

Na primeira parte do projeto, você só vai apresentar seus dados e fazer uma análise descritiva. Uma ideia seria mostrar a análise de vida média das farinhas, assim como as médias das outras medidas (índice de oxidação, ácidos graxos, etc) e interpretar os valores encontrados. Tiveram a umidade adequada, dentro do esperado? Teve algum outlier? Qual o desvio padrão (qual tipo de farinha teve mais variabilidade)? Os valores fazem sentido, dado os ingredientes da sua pré-mistura (era esperado)? E comparando com o controle, ele ficou acima, abaixo, no meio, quando comparado às pré-misturas?

Nesse momento você está explorando e entendendo os seus dados, namorando, vendo o que faz sentido e contando uma história. Espero ter ajudado :)

**P3:** Olá, tudo bem? Estou realizando os **exercícios** do módulo 1 e fiquei com uma dúvida. Na primeira questão, pede-se que se determine os tipos de variáveis e na letra c a resposta para ano na universidade aparece como qualitativa ordinal. No entanto, foi esclarecido que os dados ordinais têm categorias que podem ser colocadas em ordem/ranking, como se uma medida fosse melhor do que outra de alguma forma. Nesse caso, como poderíamos afirmar

que o ano na universidade se encaixaria aí? Não seria uma variável quantitativa, contínua e de intervalo? Obrigada!

**R:** Oi Liza, o ano da universidade seria o 1o (calouro), 2o, 3o, e assim por diante. Existe uma ordem já que você vai para o 2o depois que completa o 1o (na teoria - sabemos que na vida real isso pode ser mais complexo, rs). Não é contínua pois você não declara que está no 2o ano e meio, por exemplo.

**Cont:** Obrigada, professora! Eu havia interpretado ano como o ano de ingresso na universidade, ex: 2000, 2011 etc. Mas agora entendi o contexto.

**P4:** Olá! Uma última dúvida sobre os **exercícios**: não identifiquei na aula a explicação para o cálculo do que seria uma amplitude interquartil (IQR), limite inferior e limite superior. Poderiam esclarecer melhor este ponto? Obrigada!

**R:** (Larissa) Oii Liza, tudo bem?

A amplitude interquartil é calculada com base no cálculo de quartis, sendo o primeiro quartil (inferior), o quartil intermediário (mediana), o terceiro quartil (superior). E a diferença entre o quartil superior e o quartil inferior determina o intervalo interquartil.

Os quartis inferior e superior, Q1 e Q3, são definidos como os valores abaixo dos quais estão um quarto e três quartos, respectivamente, dos dados. Para saber quais são você precisa ordenar os dados do menor para o maior e contar o número de observações para fazer os seguintes cálculos:

$$n+1/4 = Q1$$

$$n+1/2 = \text{mediana}$$

$$3(n+1)/4 = Q3$$

A amplitude interquartil vai ser =  $Q3 - Q1$

Vou dar um exemplo:

0, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 6, 6, 7, 8, 10 (n = 19)

A mediana é o  $(19+1) / 2 = 10$  valor, i.e. 3 crianças.

O quartil inferior é 5 e o quartil superior é 15, i.e. 2 e 6, portanto a amplitude inter-quartil é 4.

**P5:** O que seria o "erro padrão"?

**R:** (Larissa) Oi Jadson, o erro padrão é uma medida da precisão da média amostral calculada, ou seja, possui a função de quantificar a certeza com a qual a média calculada a partir de uma amostra aleatória estima a média verdadeira da população. Deu pra entender?

**P6:** Não sei se entendi. O erro padrão não deveria ser medida de dispersão? Poderia informar a equação?

Essa seria o que as pesquisas eleitorais chamam de "margem de erro", onde o candidato pode ter + ou - X% do valor apresentado?

**R:** (Letícia) Oi Jadson,

Uma medida de dispersão calcula a distância entre um dado e a média, já o erro padrão mostra a distância entre a média amostral e a média populacional, e pode ser calculada pela divisão do desvio padrão pela raiz quadrada do tamanho da amostra.

Quanto ao estudo das margens de erro do exemplo que mencionou, ele exige uma maior profundidade na análise, que ficará mais claro ao estudar intervalo de confiança..

**P7:** Aline, muito obrigada! me ajudou sim, outra dúvida: como esta base faz parte do meu primeiro trabalho de estatística descritiva, eu poderia utilizar as correlações de Kendal ou Spearman neste trabalho ?

**R:** Imagina, Cristina! Que bom que te ajudou! =D

Minha sugestão seria para você incluir na parte 2 do trabalho. Como essas duas correlações são do tipo não-paramétricas, seria interessante você ter tempo de estudar sobre esse conceito primeiro e entender as vantagens e desvantagens. Por exemplo, quando estamos falando de Estatística não-paramétrica, deixamos de trabalhar com as variáveis originais para utilizar os chamados 'postos' e isso interfere na robustez dos testes e correlações, embora sejam mais flexíveis porque não pedem os requisitos dos testes e correlações paramétricas. No módulo 9, o foco é exatamente neste tema!

Mas claro, se você sentir à vontade de já trazer essas correlações na parte 1, fique à vontade também! :)

### **Tópico: boxplot excel (saco)**

**P1:** O Excel permite fazer boxplot

[https://drive.google.com/file/d/1\\_0KT0UmXUasu-cOlJrNjo1VTMdqA467A/view](https://drive.google.com/file/d/1_0KT0UmXUasu-cOlJrNjo1VTMdqA467A/view)

**R:** (Letícia) Boa noite Fernando, tudo bem?

Permite sim, obrigada por compartilhar conosco!

**P2:** Olá, prof Fernanda e equipe.

Quando você diz que o BoxPlot não é possível fazer você se refere apenas a esse plug in de 'Análise de Dados'? Quando vou em 'Inserir' > 'Inserir gráficos estatísticos' há a opção de inserir um BoxPlot lá. Seria possível comentar um pouco sobre as limitações dessa opção e como ficaria a comparação usando os dados do exemplo de notas de meninos e meninas?

Obrigada,

Anazélia

**R:** (Larrie) Olá, tudo bem, Anazélia?

Da para fazer boxplot no Excel mas ele é muito muito incompleto. As principais medidas para análise de dados (mediana, quartis, outliers, limite máximo e mínimo). E se a gente fizer isso

no Excel em meninos e meninas, não dá para tirar muitas conclusões pois não aparece as principais medidas. Mas ficariam um pouco parecido pois a mediana e quartis não se distanciam muito.

Em breve no curso vamos falar como faz no R e no SPSS, que são plataformas que fazem boxplots bons e bonitos.

**P3:** O Excel faz bloxplot, mas ele chama de gráfico "Caixa e Caixa Estreita". Basta selecionar os dados e clicar em inserir gráficos.

**R:** Sim, eu acabo mostrando em outra aula :)

**P4:** A versão mais atual do excel já faz boxplot (caixa estreita).

**R:** @Murillo F. Verdade! Eu acabo mostrando isso em outra aula mais para frente... Obrigada por pontuar nessa aula também.

### **Tópico: outros**

**P1:** Muito bom! Este módulo de aulas complementares será liberado posteriormente?

**R:** Sim, muito em breve! :)

**P2:** Ótima aula, professora. Dúvida: em uma **série temporal**, como eu posso realizar a correlação no excel? devo tirar a média de cada variável ou há algum outro método? Obrigado.

**R:** Oi Gilmar, eu não trabalho com séries temporais, então realmente não sei. Alguém por aí sabe dar a dica?

**P3:** Oi Fernanda...vc comentou na PARTE 2 sobre aulas complementares....seria mais adiante esta aula? Obrigado.

**R:** @Maximilian F. Oi Max, é que esse vídeo foi inserido depois que os alunos de outra turma tiveram acesso a essas aulas. Vocês terão acesso em breve, dentro do módulo extra

**P4:** Esse achado maravilhoso da Análise descritiva vai me poupar muito tempo daqui em diante! Fazia tudo fórmula por fórmula.

**R:** @Ayla B. A melhor coisa é descobrir o que nos poupa tempo! :)



**P5:** Onde fica a aula extra que mostra a análise em Python?

**R:** @Claudicelia O.Vão entrar no módulo extra nos próximos dias ;)

**P6:** E o boxplot? Vamos aprender a fazer?

**R:** @Maurício P.Vou mostrar no próximo encontro :)

**P7:** Olá. Uma dúvida sobre o Excel. No casos exercícios e do andar do curso, eu consigo seguir utilizando o Google Sheets ou fica muito difícil? Tem alguma opção ao Excel que seja satisfatória? Obrigado.

**R:** (Aline) Boa tarde, Fabiano!

Embora a professora Fernanda tenha feito os exercícios práticos no Excel, eu creio que é possível você acompanhar no Google Sheets. Algumas fórmulas serão diferentes, confira essa lista: [https://support.google.com/docs/topic/3105600?hl=pt-BR&ref\\_topic=3046366](https://support.google.com/docs/topic/3105600?hl=pt-BR&ref_topic=3046366), mas quando você tiver alguma dúvida, pode enviar por aqui ou pelo discord, que podemos te ajudar!

Uma sugestão seria você usar alguma extensão para te ajudar com os cálculos, eu usei a Logic Sheet para fazer o exercício que está na aula "Aplicando com Excel" [https://workspace.google.com/u/0/marketplace/app/logic\\_sheet\\_ferramentas\\_de\\_an%C3%A1lise\\_de\\_da/796322869198](https://workspace.google.com/u/0/marketplace/app/logic_sheet_ferramentas_de_an%C3%A1lise_de_da/796322869198). Com ela eu consigo gerar uma tabela com as estatísticas descritivas, dentre outras opções! Espero que tenha ajudado :)

## Comentários: Primeiros passos em R

**P1:** Larissa, instalei o R e RStudio com sucesso. Contudo **não consigo abrir o RStudio**. Vc poderia me ajudar?

**R:** (Larissa) Oii Marcus, qual o erro que está dando? Ou você não está conseguindo achar o programa? Você poderia enviar um print do erro? (Você pode colocar o print no drive e compartilhar o link aqui, ou falar pelo grupo do discord que a professora disponibiliza no inicio do curso)

**Cont:** Olá Larissa ! Conseguir abrir e reproduzir alguns comandos da sua aula. Mas ao tentar criar dados retornou a msg "Error in c(1,2,3,4): objeto interpretável como fator". Também observei que os resultados não foram apresentados no campo Enviroment

**R:** Oii Marcus, no caso você dados <- c(1,2,3,4) , colocou desse jeito?

**P2:** Tenho algumas **dúvidas**:

- Na moda, os NA já são desconsiderados automaticamente? Não preciso pedir pra excluir?
- Não entendi o que ela quis dizer com quantil, é quartil? Se for, então eu posso considerar que os 4 quartis são: 18-20 , 20-28, 28-33 e 33-45?
- Antes de instalar o pacote, se eu não tiver, o R me diz qual pacote eu preciso?

**R:** (Larissa) Oii, tudo bem?

Sobre a primeira pergunta: Não precisa pedir pra excluir não, nesse caso vai retornar só os resultados que possuem algum valor.

Segunda pergunta: é quartil sim, para calcular o quartil você utiliza a função quantile(Meninas) ou quantile(Meninos) isso se você tiver utilizado a função attach(nomedabasededados) antes, se não você utiliza quantile(nomedabasededados\$Meninas) , sendo que o nome da base de dados é o nome que você colocou.

Terceira pergunta: então, já aconteceu comigo algumas vezes do R sugerir qual o pacote eu poderia instalar que tinha a função que eu queria, mas nem sempre isso acontece. Mas se você souber a função que você quer usar e não souber o pacote, pode colocar no google: 'nomedafunção' pacote em R, que você consegue achar.

**P3:** Há 2 alternativas para não precisar instalar localmente o R e o RStudio:

1. Usar a versão na cloud (<https://posit.cloud/> --& Posit Cloud é o novo nome do RStudio Cloud) - tem versão gratuita e outras pagas.

2. Usar o Google Colab:

- pode acessar diretamente por este link: <https://colab.research.google.com/#create=true&language=r>

- ou acessar pelo link "normal": <https://colab.research.google.com/>, e dentro do notebook ir em Runtime &gt; Change Runtime Type &gt; Selecionar R em Runtime Type.

A primeira opção, do RStudio na cloud, tem a mesma visualização da versão no desktop com as áreas: Editor, Environment, Console e Output. Já a versão do Google Colab é um notebook mesmo, menos visual - mas bastante útil...

**R:** Ótima dica, Adriana! Obrigada :)

**P4:** Olá, eu preciso manter o R instalado mesmo depois de instalar o RStudio?

**R:** Oi, Daiane! Tudo bem? =D

É necessário manter o R instalado (e atualizar se necessário com o tempo), porque o RStudio é "apenas" um ambiente de trabalho com um visual mais amigável para o usuário, ele é totalmente dependente do R. Tecnicamente o RStudio é uma IDE (ambiente de desenvolvimento integrado), isso é bem comum, por exemplo, o Python geralmente é utilizado com uma IDE chamada VS Code (e há outras também). :)

**P4:** Nossa, o meu não aparece essa parte do script no canto superior esquerdo... :O

**R:** Oi, Camila! Tudo bem? :)

Você se refere a parte que escrevemos os códigos, certo? Em cada área na parte superior normalmente encontramos um botão de minimizar/maximizar. No seu está aparecendo também?

**Cont:** Ah sim! deu certo hehe muito obrigada! Aline, outra dúvida. O RStudio contempla todas as "funções" do R? Eu consigo desinstalar o R e ficar somente com o RStudio? Obrigada!

**Cont:** Maravilha, Camila! =D

O RStudio funciona como uma forma visual mais amigável para utilizar as ferramentas que tem no R, então sem o R instalado ele (RStudio) não funcionaria. Tecnicamente chamam o RStudio de IDE (ambiente de desenvolvimento integrado), que é justamente esses softwares que são interfaces gráficas mais fáceis de utilizar. Então é só deixar o R sempre instalado (e ir atualizando conforme necessário), mas seu ambiente de trabalho vai ser no RStudio tá bom? :)

**P5:** Boa noite! Eu desistalei a versão anterior que eu tinha do R e do R Studio para instalar a versão mais atual no meu computador que é no ambiente Windows. O R abriu corretamente, mas o R Studio, toda vez que eu clico no ícone aparece o instalador perguntando se quer instalar e não abre o programa. O que eu faço? A versão que apareceu para download foi 4.3.0-win.exe

**R:** Bom dia, Érika!

Entendo. Mesmo que você aceite instalar novamente, essa mensagem continua aparecendo depois, certo?

Aparentemente o RStudio não está conseguindo localizar o R que já foi instalado. Érika, você poderia desinstalar tanto o R como o R Studio e reiniciar o processo de instalação?

1) Nesse link: <https://cran.r-project.org/bin/windows/base/> clique em "Download R-4.3.0 for Windows" para instalar a versão mais atualizado do R;

2) Nesse link: <https://posit.co/download/rstudio-desktop/>, na opção 2, clique para instalar o RStudio.

Por favor, me diga se isso te ajudou! Se quiser compartilhar algum print, você pode colocá-lo no Google Drive ou One Drive e enviar o link da imagem aqui, tudo bem? :)

**P6:** É possível importar dados \*.xlsx?

**R:** Sim, Ronaldo, é possível! Segue um tutorial em como fazer: <https://www.youtube.com/watch?v=eRPKYrfce9w>

Se preferir, segue uma descrição do passo a passo:

Importar arquivos Excel (extensão .xlsx ou .xls)

"Para isto usamos um pacote que já está instalado no R. Basta ir no Files/Import Dataset/From Excel ou ir na aba Environment e depois na opção Import Dataset escolhendo a opção From Excel. Além de um preview dos dados, aparecerá a opção de manter ou renomear o arquivo de dados, escolher uma worksheet da planilha Excel, informar o range da planilha onde está o banco de dados, limitar o número de linhas a serem importadas, definir o número de linhas iniciais que devem ser desconsideradas, informar se a planilha tem ou não cabeçalho, redefinir o tipo de cada variável, informar códigos usados para missing values, escolher a opção de pular determinadas colunas da base de dados. A medida que opções são feitas, é possível visualizar ou mesmo copiar a linha de comando relativa aquele conjunto de opções. Pode ser útil copiar e colar o código para que não seja necessário usar a caixa de diálogo em uma próxima vez que aquela base de dados for analisada" Fonte: [https://www.ime.usp.br/~fmachado/MAE399/ManualR.nb.html#:~:text=Importar%20arquivos%20Excel%20\(extens%C3%A3o%20.&text=Para%20isto%20usamos%20um%20pacote,escolhendo%20a%20op%C3%A7%C3%A3o%20From%20Excel](https://www.ime.usp.br/~fmachado/MAE399/ManualR.nb.html#:~:text=Importar%20arquivos%20Excel%20(extens%C3%A3o%20.&text=Para%20isto%20usamos%20um%20pacote,escolhendo%20a%20op%C3%A7%C3%A3o%20From%20Excel).

**P7:** Ola, por favor peço que disponibilizem os arquivos de dados para repetir o que foi feito na aula. Obrigado.

**R:** Olá, Ronaldo! Como vai? :)

Obrigada por trazer essa sugestão! Os scripts dos próximos módulos já estão disponíveis, vou verificar se há possibilidade de disponibilizar para essa aula em específico também. Agradeço muito a sugestão! Envio uma resposta assim que tiver uma atualização =D

**Cont:** Olá novamente Ronaldo, tudo bem? A Larissa (autora dos vídeos) nos informou que não possui um script para essa aula em específico porque a ideia era apenas demonstrar algumas funções iniciais do R para entrada de dados. Ela sugeriu utilizar algum arquivo em Excel que você tivesse ou consultar o seguinte link: [https://fernandomayer.github.io/ce083-2016-2/04\\_Entrada\\_de\\_dados.html](https://fernandomayer.github.io/ce083-2016-2/04_Entrada_de_dados.html). Através dele é possível ver as diferentes formas de entrada de dados no R e há arquivos para baixar e praticar também! Qualquer dúvida, pode nos falar tá bem? =)

**P8:** Oi Larissa, tudo bem?

Quando você usa o + para quebra de linha, você inclui ele no final da linha e também no início da linha seguinte para indicar que é continuação, ou se apenas incluir no final da primeira linha o R já entende que a linha posterior é continuação?

Obrigada!

**R:** Bom dia Leticia! :D Isso mesmo, o sinal de '+' você acrescenta somente no final da linha e automaticamente o R cria a quebra de linha com uma indentação no código (que é esse espaço recuado para a direita), indicando que faz parte do mesmo 'bloco'. Uma outra forma de melhorar a visualização é habilitando a opção chamada 'Soft-wrap', facilita bastante! No menu principal você clica em 'Tools', em seguida: 'Global Options', depois: 'Code' e por último marque a caixa de seleção 'Soft-wrap R source files' e finalize clicando em 'Apply'. Espero ter te ajudado!

## Comentários: Aplicação em R

**P1:** Pessoal, como interpreto o **valor da curtose**? Entendi que curtose seria a distribuição dos valores na curva ("grau de achatamento"). Mas, como interpretar os valores na saída do R?

Existe algum coeficiente da curtose (ex: k) padronizado para comparação? Eu achei algumas coisas discrepantes na internet, podendo ser  $k=0$  ou  $k=0.263$ .

Exemplo: Se for  $k=0.263$ .

Curva simétrica: coeficiente de curtose = 0.263

Curva pontiaguda: coeficiente de curtose < 0.263

Curva achatada: coeficiente de curtose > 0.263

Obrigado!

**R:** (Larissa) Oii carlos, tudo bem? Nesse texto, na parte em que fala sobre curtose, tem explicando sobre sua dúvida: <https://medium.com/psicodata/usando-r-para-avaliar-a-normalidade-14482a27cf60>

**P2:** 1 - Eu acho que eu perdi onde a prof explica sobre curtose, erro padrão...

2 - Assimetria é a que foi explicada no histograma e no boxplot (esquerda... direita...)?

**R:** (Larissa) Boa tarde Suelen, tudo bem?

Referente a explicação sobre assimetria e curtose, vou deixar abaixo o link desse tema no blog da professora Fernanda para te ajudar na compreensão:

<https://blog.proffernandamacieli.com.br/assimetria-e-curtose-dos-dados/>

Sobre a explicação de assimetria, é essa mesmo que você mencionou, está no slide 4 da aula de Histograma e Boxplot.

Já o erro padrão, ele é a variação entre a média amostral e a média populacional, você pode calcular ele dividindo o desvio padrão pela raiz quadrada do tamanho da amostra

Espero ter te ajudado :)

**P3:** Ao **importar a base de dados** dos preços das casas tive um problema com o nome da pasta. Apareceu a seguinte msg no Console:

'Error: mixing Unicode and octal/hex escapes in a string is not allowed'

Verifiquei que o problema estava no nome da pasta que a base de dados estava inserida, pois estava com assento.

Há uma forma de resolver isso, sem que eu precise modificar o nome da pasta ou trocar o arquivo de lugar?

**R:** (Larissa) Oii Marília, tudo bem? Não, você terá que ajustar o nome do arquivo ou importar utilizando outra forma:

```
read.delim("clipboard",dec="," ,header=T)
```

Só que nesse formato você tem que copiar no ctrl c todos os dados e depois dar enter no código. Entende?

**P4:** Olá!! No PDF do exercício não tem o **site para coletar os dados**. Fico no aguardo.

**R:** (Larissa) Oi Mineia tudo bem? Assim que você abrir o R, você vai executar o seguinte comando:

```
data(iris)
```

Esse comando vai abrir o conjunto de dados para você conseguir realizar as análises.

**P5:** Olá boa noite! Não consegui baixar o "readxl" pareceu a mensagem a versão requerida não pode ser encontrada.

**R:** Oi Danielle, tudo bem?

Nesse caso, sugiro instalar o pacote para ler o Excel separadamente, que pode ser que funcione. Antes de importar os dados, use esses comandos:

```
install.packages("readxl")  
library(readxl)
```

Aí você tenta importar a base novamente. Espero que funcione dessa forma :)

**P6:** Olá!

Quando tento importar o arquivo em Excel, aparece uma mensagem: "Required package versions could not be found:

```
readxl 0.1 is not available  
Rcpp 0.11.5 is not available
```

Check that `getOption("repos")` refers to a CRAN repository that contains the needed package versions.

Poderia me ajudar?

**R:** Oi Ana Luíza, tudo bem?

Seguindo os passos do vídeo era para abrir automaticamente... Nesse caso, sugiro instalar o pacote para ler o Excel separadamente, que pode ser que funcione. Antes de importar os dados, use esses comandos:

```
install.packages("readxl")  
library(readxl)
```

Aí você tenta importar a base novamente. Espero que funcione dessa forma :)

**P7:** Olá, tudo bem?

Pra mim não aparece nenhum arquivo com o nome "Análise de dados Excel", no ZIP tem apenas um arquivo chamado "Módulo 1 Script" e se eu tento importar, também não me mostra essa opção chamada "Import Dataset"

**R:** Olá, Camila! Tudo jóia e você? :D

Entendi, para resolver essa questão siga esse passo a passo, por favor:

1) Escolha ou crie uma pasta no seu computador que será o diretório de trabalho para as aulas do R. Nessa pasta, já deixe salvo o arquivo 'AnaliseDadosExcel.xlsx', é o mesmo que a professora utilizou na aula 'Aplicando com Excel' (o arquivo está disponível na aba 'Materiais' dessa aula);

2) No RStudio selecione essa pasta como diretório de trabalho assim como a Larissa fez no vídeo. No menu principal clique em "Session", depois "Set Working Directory" e "Choose Directory".

3) Agora você verá o arquivo do Excel na janela do lado direito onde está escrito "Files". É só importar o arquivo: clique no nome do arquivo e em seguida "Import Dataset". A Larissa faz isso no minuto 1:33.

4) Agora pode rodar o scrip normalmente!

Funcionou? :)

**P8:** As funções std.error, skewness e kurtosis não funcionaram, mesmo instalando o pacote.  
Mensagem: Error in std.error(AnaliseDadosExcel\$Meninas) :  
could not find function "std.error"

Error in skewness(AnaliseDadosExcel\$Meninas) :  
could not find function "skewness"

Error in kurtosis(AnaliseDadosExcel\$Meninas) :



could not find function "kurtosis"

Será que estou fazendo algo errado?  
Como resolver?

**R:** Olá, Cláudia! Tudo bem?!

Sei que a sua dúvida já foi resolvida no Discord, mas vou deixar uma sugestão que pode ser útil para você ou outros estudantes futuramente :)

Depois que é feita a instalação do pacote que se quer utilizar é necessário carregá-lo. Se você quiser conferir se os pacotes estão carregados, digite esse comando `(.packages())` que no console irá aparecer a lista dos pacotes já carregados.

Se não estiver carregado, é necessário executar o comando para o carregamento usando a função `library( )`. Entre parênteses colocamos o nome do pacote que queremos, então nesse caso seria: `library(plotrix)`. Para executar basta ir na linha que está o comando anterior e apertar control + enter.

=D