*Estatística descritiva: como conhecer os dados*

Estatística descritiva. Generalidades

A estatística descritiva é um ramo da estatística que nos ajuda a entender melhor os dados, com base em sua **descrição** e **análise**. Como parte do processo de Data Science, a estatística descritiva é um dos pilares da **Análise Exploratória de Dados** ou **EDA**, que veremos mais adiante neste curso.

Tipos de variáveis e tipos de dados

Nas aulas anteriores pudemos ver como os tipos de variáveis ​​funcionam no Python. Em estatística, **os tipos de variáveis ​​são análogos em conceito aos tipos de dados**. Isto ocorre na programação, entre outros motivos, porque **existem operações que podemos fazer com alguns deles, mas não com outros.** Podemos partir de uma noção básica em estatística sobre os tipos de variáveis ​​desta analogia. Em geral falaremos sobre uma grande divisão nos tipos de variáveis, que é a que será mais prática para Data Science: variáveis ​​quantitativas e qualitativas.

Variáveis quantitativas

Variáveis ​​quantitativas são variáveis ​​para as quais faz sentido somar, subtrair ou calcular a média, dependendo do caso. Em geral, são numéricas e pertencem ao conjunto dos números reais ou a um de seus subconjuntos. Por exemplo, as alturas das pessoas podem ser medidas em centímetros e as idades podem ser medidas em números inteiros, o que na realidade não acontece (nossa idade aumenta dia a dia, segundo a segundo), mas pode servir a propósitos práticos. Ter 20,4 anos pode ser muito importante para um estudo médico, mas muito impraticável para as idades dos consumidores de um produto.

De acordo com a noção que demos anteriormente, uma verificação rápida para averiguarmos se a variável é numérica é fazer esta pergunta**: "Se eu somar, subtrair ou calcular a média dos valores, o resultado faz sentido?"**. Se fizer sentido para qualquer um desses casos, estamos lidando com uma variável numérica. É importante notar que isso sempre **depende do problema** que estamos enfrentando. Por exemplo, é válido dizer que em média uma família tem 2,5 filhos. Para fins estatísticos é muito útil saber. Mas, por exemplo, dizer que em média os carros da marca Ford têm 4,5 portas não é nada útil. Para estes exemplos, "número de filhos por família" é uma variável numérica, enquanto "número de portas de carros", embora seja um número (3 portas, 5 portas, etc.) não é uma variável quantitativa por natureza. Variáveis ​​numéricas podem ser implementadas em Python com os tipos de dados int, float ou complex.

Séries temporais

As variáveis ​​do tipo data são um mundo à parte, tanto para a visão estatística quanto para a abordagem do tipo de dados de programação. Possuem algumas qualidades distintivas como a sequencialidade, a relação entre os dados com valores sucessivos e as características particulares da evolução temporal, como anos bissextos, estações do ano, periodicidade em escalas que não se encaixam perfeitamente (como as semanas no mês), etc. Embora essas questões possam ser pensadas com simples bom senso, na hora de realizar análises e tirar conclusões, principalmente com grandes volumes de dados, esse tipo de dado pode trazer problemas se o tratarmos manualmente. Felizmente, as linguagens de programação levam em conta essas situações e as resolvem com funções específicas, e de forma muito prática. Lembre-se de que vimos uma breve introdução à Série Temporal na Aula 04 com o Pandas.

Em termos estatísticos, uma série temporal pode ser considerada uma variável quantitativa, pois embora a média não seja uma operação que faça muito sentido, é possível estabelecer diferenças entre datas, que resultam em durações. Por exemplo, é prática comum no desenvolvimento de sistemas que a idade de uma pessoa seja calculada como a diferença entre a data de nascimento (carregada no banco de dados) e a data de hoje (calculada com uma função de linguagem de programação).

Variáveis categóricas ou qualitativas

As variáveis ​​categóricas ou qualitativas, por sua vez, como o próprio nome indica, **servem efetivamente para categorizar os elementos**. Isso significa que poderíamos construir subconjuntos ou subgrupos de elementos de acordo com a referida variável. Aqui, por exemplo, encontramos alguns dados comuns que são usados ​​em data science para caracterizar pessoas, como sexo/gênero, ocupação/profissão e local de origem. Em geral, as variáveis ​​categóricas são dados de texto livre, portanto, em princípio, são fáceis de identificar. Como regra geral, **não é errado recorrer à pergunta anterior: "Se eu somar, subtrair ou calcular a média dos valores, o resultado faz sentido?"**. Quando a resposta for **"não" para todos os casos, a variável é categórica.**

Casos particulares

Em geral, as linguagens de programação oferecem ferramentas para detectar os tipos de **variáveis ​​de forma automática**, mas não custa sempre checar minuciosamente os tipos de variáveis, principalmente para detectar os casos mostrados abaixo.

Variáveis que são categóricas mas que são escritas como números

Variáveis ​​que são categóricas, mas escritas como um número

Como mencionamos anteriormente, é importante **levar em consideração a natureza dos dados** para verificar se faz sentido que a variável que eles representam seja numérica ou categórica. No exemplo das portas do carro, temos uma variável que pode ser escrita com números (para o exemplo dado: variável “Número de portas”, com valores possíveis 3, 4 e 5), mas cuja natureza não é numérica. Variáveis ​​categóricas, se forem números como neste caso, não fazem sentido quando somadas, subtraídas ou calculadas em média, mas **fazem sentido se forem contadas**. Por exemplo, pode ser muito interessante saber quantos carros têm 3 portas em comparação com quantos têm 4 ou 5.

Variáveis lógicas

As variáveis ​​lógicas são análogas aos tipos de **dados booleanos ou lógicos** (bool) do Python. Esse tipo de dado tem como valores possíveis true (True) e false (False). Essa variável tem seu tipo de dados específico em Python, porque pode ser usada para realizar operações específicas. No entanto, para efeitos de análise estatística, continua a ser uma **variável categórica.**

Variáveis de alta cardinalidade e metadados

Em Data Science, não é incomum que apareçam variáveis ​​que, em termos de seu tipo, podem ser numéricas ou textuais, mas que **não fazem sentido para serem somadas, subtraídas ou calculadas em média, ou contadas**. Exemplos típicos dessa situação são números de documentos ou qualquer tipo de identificação pessoal ou fiscal, números de série de produtos ou identificadores ou códigos de clientes. Em geral, esses valores, **mais do que categorizar ou quantificar** uma parte de todo o conhecimento que temos sobre um elemento, os **identificam** de alguma forma. Os identificadores **ajudam a localizar um item**, o que é essencial para seu armazenamento e manipulação em bancos de dados, mas não **tem utilidade prática para fins** **estatísticos**. Qualquer variável que cumpra essa função é um **metadados**, ou seja, um dado que descreve ou fornece informações sobre os dados em questão. Nesse sentido, os metadados **não são quantitativos nem categóricos** e é melhor **separar esse tipo de variável da análise estatística**.

Uma maneira rápida de identificar esses metadados é a partir de sua **cardinalidade**. No campo de Bancos de Dados, a cardinalidade é uma medida de **quantos dados** diferentes existem para uma **variável**. No caso apresentado das portas dos carros, a cardinalidade seria 3: carros 3 portas, 4 portas e 5 portas, ou seja, três valores diferentes. No caso de valores lógicos ou booleanos a cardinalidade é 2, e no caso de gêneros pode ser variável. Por exemplo, um usuário do Facebook pode escolher entre mais de 60 gêneros diferentes no aplicativo. As linguagens de programação calculam a cardinalidade da variável em relação ao número de elementos, e essa regra pode ser muito útil se não soubermos o significado específico da variável com a qual estamos trabalhando. Se a **cardinalidade de uma variável for próxima ao número de elementos**, então é possível que a variável seja um metadado, por exemplo, certamente haverá tantos números de documentos quanto pessoas. Por outro lado, 60 gêneros parecem um número grande para várias categorias quando vistos isoladamente, mas comparados aos 2,7 bilhões de usuários do Facebook, na verdade são muito poucos. O gênero em si é uma característica que pode ser inferida a partir do simples senso comum, mas nos casos em que o significado da variável não é tão claro, **a regra da cardinalidade serve como um guia muito bom**.

Conceitos básicos

Toda vez que pegamos um conjunto de dados, assumimos que eles representam um viés da realidade. Por exemplo, ao fazer uma pesquisa por telefone, temos apenas os dados dos entrevistados, não de todas as pessoas. Ao fazer um levantamento de indicadores demográficos (altura, peso, idade), consideramos apenas os dados das pessoas pesquisadas. Mesmo que queiramos ver os dados de nossa empresa e tenhamos acesso ao banco de dados completo, se fizermos data science queremos saber o que pode acontecer com os dados que ainda não temos: por exemplo, quais clientes podem comprar mais, ou quais são mais propensos a cometer fraude.

Nesse sentido, sempre há dados aos quais não temos acesso, porque não podemos levantá-los ou porque ainda não aconteceram. No entanto, são precisamente esses dados que serão objeto de nossas previsões e estimativas da mão da estatística e do data science. É por isso que é necessário distinguir entre esses conjuntos de dados.

Em termos estatísticos, chamaremos cada elemento **individual** que pode ser descrito com um conjunto de variáveis ​​representadas com dados, ou simplesmente observação. O universo de observações que consiste em todas as unidades de observação possíveis, incluindo aquelas às quais não temos acesso, ou que ainda não ocorreram, é chamado de **população**. Por fim, o conjunto de observações que contém as unidades de observação às quais temos acesso é chamado de **amostra**. A partir da amostra aplicaremos ferramentas estatísticas para obter conclusões sobre a população.

Medidas resumo

Muitas vezes surgem situações em que obtemos amostras na forma de conjuntos de dados, e depois de acomodá-los para poder introduzi-los adequadamente em nosso código (um tópico que veremos em profundidade na classe correspondente a Data Wrangling), temos que dar uma olhada nas variáveis, a fim de interpretar as características do conjunto de dados e, assim, obter conclusões preliminares sobre a população da amostra. No processo de data science, essas conclusões podem ser usadas para tomar decisões sobre quais algoritmos aplicar ou quais considerações fazer em relação aos dados.

Dois exemplos comuns dessa análise prévia podem ser vistos nos dados econômicos. No marketing é conhecida a "regra de Pareto", que diz que 80% do lucro é dado por 20% dos clientes. Embora nem sempre seja exatamente assim, é muito comum que isso aconteça, e nesse sentido a análise de clientes muitas vezes exige um corte (deixar os clientes de fora da análise) para considerar apenas os mais importantes, e esse corte é feito à base de métodos de estatística descritiva. Por outro lado, sempre que falamos de salários ou distribuições de riqueza, sabe-se que em geral grandes somas de dinheiro estão concentradas em poucas pessoas. Dessa forma, calcular uma média simples para entender qual é o salário mais representativo para a população é uma má escolha, pois seria o equivalente a medir a média das alturas de uma sala de jardim de infância, ... junto com alguns jogadores de basquete! As diferenças de dinheiro são geralmente muito grandes (como as diferenças de altura no caso anterior), e é necessário analisá-las com ferramentas mais apropriadas que a média. Veremos essas questões a seguir.

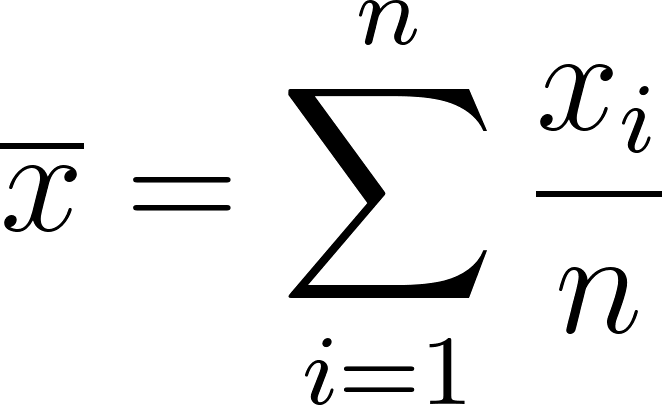


Para ordenar as medidas a seguir, faremos uma distinção entre variáveis ​​qualitativas e quantitativas. Tenhamos sempre presente neste sentido que a correta caracterização da variável é essencial para saber qual a medida a aplicar, e assim compreender melhor os dados.

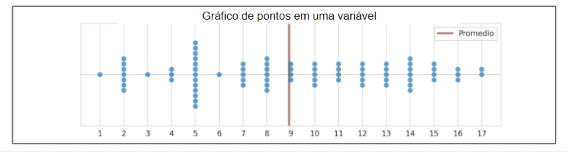
Variáveis quantitativas

Média

A média ou média aritmética é definida como a soma de todos os valores divididos pela quantidade de dados. Na fórmula:

[](https://www.codecogs.com/eqnedit.php?latex=%20%5Coverline%7Bx%7D%20%3D%20%5Csum_%7Bi%3D1%7D%5En%20%5Cfrac%7Bx_i%7D%7Bn%7D%20#0)

O parâmetro n representa o número de valores. O parâmetro i representa um índice que varia de 1 a n. A média é representada por. A média é a medida mais conhecida, e visa mostrar a posição mais central de todos os dados, descrevendo sua localização. Por exemplo, é útil saber que a altura média das crianças na sala do jardim é de 100 centímetros e que a altura média dos jogadores de basquete é de 195 centímetros. Os dois grupos estão bem definidos, pois suas médias são muito diferentes. Com isso, podemos dizer que em geral as posições dos dados para os dois grupos são diferentes entre si.



Mediana e quartis

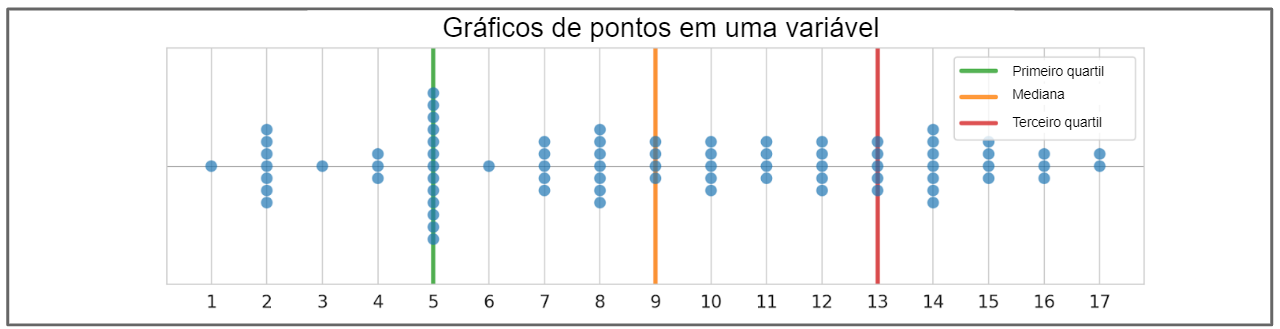
Muitas vezes a média não é uma medida suficiente para descrever os dados. Pode haver valores muito distantes do "centro" dos dados, ou todos os dados podem estar muito dispersos, ou o "centro" pode ser "executado" em algum lugar, como é muito comum com os salários: muitos trabalhadores ganham salários-mínimos ou próximo do mínimo, e pouquíssimas pessoas cobram valores exorbitantes. Nesses casos, a média não é um número claro para descrever o conjunto de dados.

Uma medida para resolver esta situação é a mediana. A mediana é calculada da seguinte forma:

1. Classificar os dados do menor para o maior. Se houver valores repetidos, basta colocar a quantidade de vezes que aparecem. A quantidade de dados originais permanecerá, mas desta vez em ordem.
2. Com os dados ordenados, vamos agora contar a quantidade de dados.
   1. Se a quantidade de dados for ímpar, vamos encontrar o valor que está exatamente no meio dos dados. Essa é a mediana.
   2. Se a quantidade de dados for par, haverá dois valores no centro dos dados. Vamos calcular a média desses dois valores. Esta será a mediana.

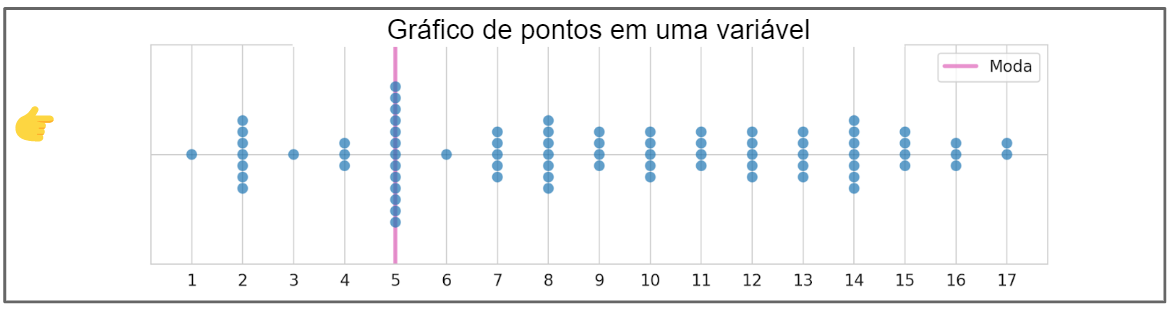
Vamos pensar por um segundo o que significa ter a mediana de um conjunto de dados. Se tivéssemos que estar no meio dos dados ordenados para chegar à mediana, podemos dizer com segurança que metade dos dados é menor ou igual à mediana e que a outra metade dos dados é maior ou igual à mediana. Desta forma, dizemos que 50% dos dados possuem valores menores que a mediana, e da mesma forma os 50% restantes dos dados possuem valores maiores que a **mediana**.

Com essa ideia em mente, podemos estender o conceito de mediana para os valores que atingem 25% e 75% dos dados. Chamaremos esses valores de primeiro e terceiro quartis, pois representam um quarto e três quartos dos dados, respectivamente. Nesse sentido, a mediana equivale a dois quartos dos dados, com os quais também a chamaremos de segundo quartil. A mediana é mais compreensível quando falamos de salários, pois podemos dizer a partir de seu uso que uma certa porcentagem da população ganha mais (ou menos) que um determinado valor. Se esse valor for igual, por exemplo, à cesta básica de consumo, então o percentual da população que ganha menos que esse valor será considerado “abaixo da linha da pobreza”, como exemplo de como esse indicador econômico é medido. Os valores percentuais derivados da mediana que se situam entre 0 e 100 são genericamente chamados de percentis. Nesse sentido, podemos dizer, por exemplo, que o primeiro quartil corresponde ao percentil 25 e a mediana corresponde à porcentagem de 50.



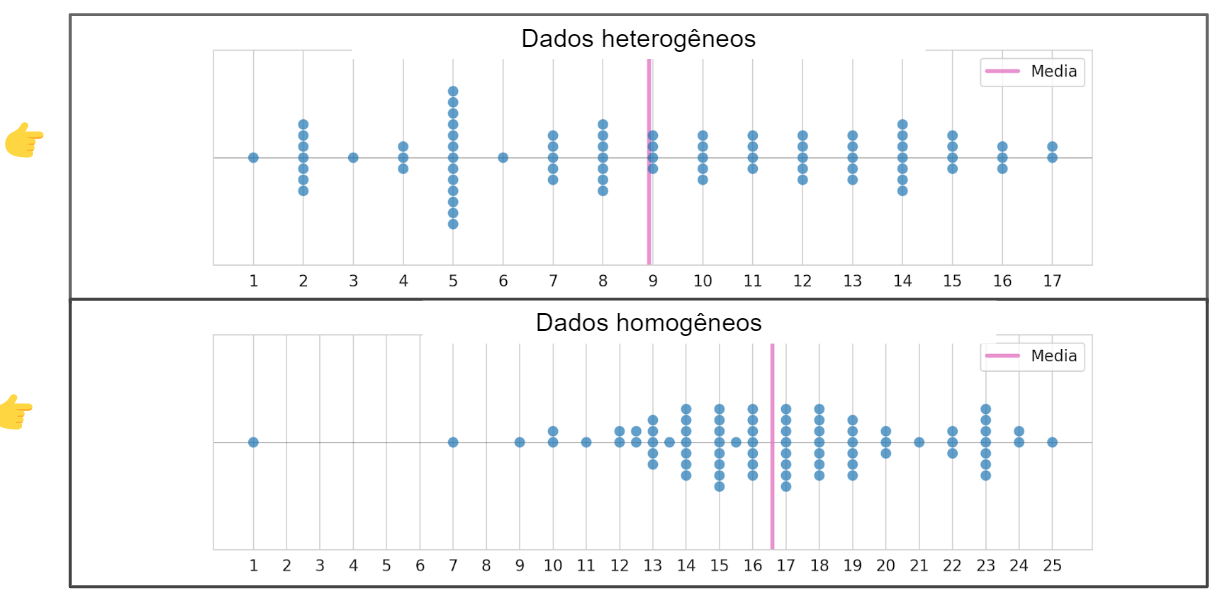
Moda

A moda é o valor que ocorre com mais frequência em um conjunto de dados. É obtida a partir de uma simples contagem dos dados, calculando qual valor aparece mais vezes. É importante notar que esta medida pode ser usada tanto para variáveis ​​quantitativas quanto quantitativas. No caso em que a moda é calculada para variáveis ​​quantitativas, é mais prático trabalhar com valores inteiros, pois nesse caso haverá mais valores que possivelmente se repetem, e nesse caso faz mais sentido saber qual é o mais frequente.

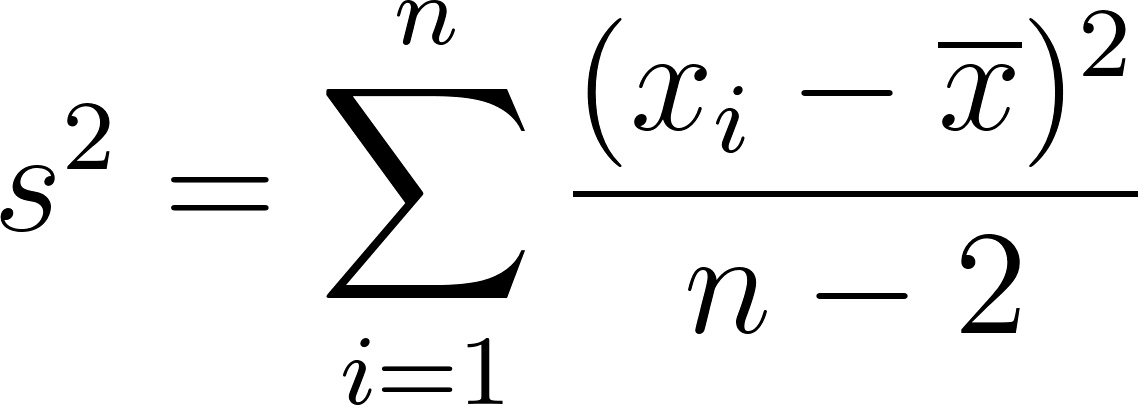


Variância

As medidas acima servem para "localizar" os dados. Conhecendo os valores da média, mediana e moda, podemos obter um bom resumo sobre a posição dos dados. Por outro lado, uma vez conhecida a posição de uma variável, também é conveniente saber como é sua forma. Consideraremos neste sentido dois tipos de formas: dados homogêneos e heterogêneos. Se os dados são mais homogêneos, significa que geralmente estão agrupados mais próximos da média. Por outro lado, se os dados são heterogêneos, significa que, em geral, estão longe da média. Mostramos alguns exemplos nas próximas duas figuras.



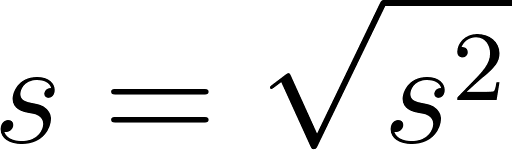
Uma medida estatística que funciona muito bem para isso é a variância. Consiste em medir as distâncias de todos os dados até a média, elevando cada distância ao quadrado, fazendo a soma e dividindo pela quantidade de dados menos duas unidades. Na fórmula:

[](https://www.codecogs.com/eqnedit.php?latex=%20s%5E2%20%3D%20%5Csum_%7Bi%3D1%7D%5En%20%5Cfrac%7B(x_i%20-%20%5Coverline%7Bx%7D)%5E2%7D%7Bn-2%7D%20#0)

O número n representa o número de valores. O número i representa um índice que varia de 1 a n. A variância é representada por s2. Não teremos que usar essa fórmula "manualmente" para entender como funciona a variância, mas é bom pensar nesse número da seguinte forma: é uma espécie de média das distâncias dos dados em relação à média dos espetáculos. Ou seja, quanto mais longe os dados estiverem da média, maiores serão as distâncias e, portanto, a variância será maior. Por outro lado, se os dados estiverem muito próximos da média, suas distâncias serão menores e o valor da variância será menor.

Desvio padrão

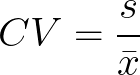
A variância tem um problema: como sua fórmula envolve uma operação de elevar ao quadrado, seu resultado será expresso em unidades ao quadrado. Por exemplo, se estamos medindo alturas em centímetros, a variância será dada em centímetros quadrados, o que não faz muito sentido se quisermos interpretar o valor em relação aos dados. Para isso, basta aplicar a raiz quadrada na variância e obter um valor nas unidades da variável. Esse novo valor é chamado de desvio padrão e é simbolizado pela letra s. Na fórmula:

[](https://www.codecogs.com/eqnedit.php?latex=%20s%20%3D%20%5Csqrt%7Bs%5E2%7D%20#0)

Para o caso anterior de medição de alturas em centímetros, o desvio padrão também será expresso em centímetros, e dará uma ideia de quão longe os dados estão, em média, da média amostral.

Coeficiente de variação

Representa a variabilidade *dos dados em relação à média*. Quanto menor o CV mais homogêneo é o conjunto de dados.Um CV é baixo (< 0.25) indica que um conjunto de dados é razoavelmente homogêneo.



Variáveis qualitativas

No caso das variáveis ​​qualitativas, como mencionamos anteriormente, os cálculos que fazem sentido são aqueles relacionados à contagem das observações e sua categorização. De acordo com o que foi visto, então, nos interessam as medidas listadas abaixo, entre outras.

* A contagem dos dados totais (valor de n).
* A contagem dos dados por categoria, ou seja, para cada valor possível da variável, o número de observações.
* O valor mais frequente, ou seja, aquele com mais observações. Este é o cálculo da moda, como vimos anteriormente.

Distribuições de variáveis

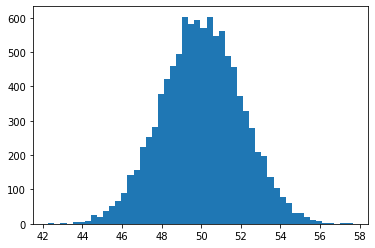
Introdução

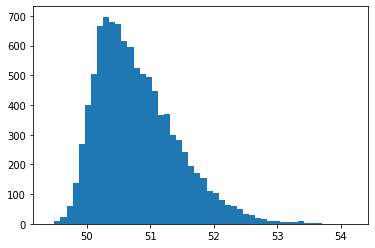
Indo um pouco mais fundo na linguagem estatística, veremos que se as variáveis ​​quantitativas possuem uma posição e uma forma, então cada uma delas pode ser comparada e/ou diferenciada de outras variáveis ​​com posições e/ou formas diferentes. Dizemos então que cada variável tem uma distribuição diferente.

Uma maneira de pensar sobre distribuição como um conceito é a seguinte. Vamos colocar todos os valores possíveis da variável em uma linha, que chamaremos com a letra x. Agora vamos representar graficamente um ponto para cada valor que ocorre em nossos dados, empilhando-os quando houver mais de um valor repetido. Desta forma, a altura de cada pilha de pontos representa o número de observações para cada valor.

Como mencionamos quando falávamos sobre a moda, não é prático medir a frequência de ocorrência dos dados quando os valores não são inteiros. Apesar disso, é possível construir pequenos intervalos que separam os números em intervalos consecutivos. Desta forma podemos considerar cada intervalo como um "valor" com sua própria entidade, e desta forma podemos ter uma ideia mais fiel sobre a forma dos dados. Dessa forma, representamos graficamente a distribuição dos dados em um gráfico chamado histograma. Os histogramas são muito convenientes para visualizar rapidamente a posição e a dispersão dos dados. Veremos essa ferramenta gráfica em profundidade na aula 7.

De acordo com as formas das distribuições, podemos fazer suposições que nos ajudam a entender melhor os dados. Uma característica a ter em mente ao analisar distribuições é se elas são simétricas ou assimétricas. O histograma na primeira figura abaixo mostra uma distribuição simétrica enquanto a distribuição na segunda figura mostra uma distribuição assimétrica à esquerda, como poderia acontecer com os salários de uma população: muitas pessoas com baixos salários e poucas pessoas com altos salários.

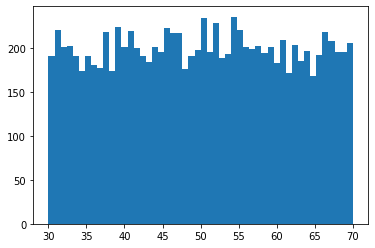




Existem muitas distribuições com nome próprio, que têm aplicações particulares nas diversas áreas de estatística aplicada e do Data Science. Duas distribuições muito importantes são a distribuição uniforme e a distribuição normal. Veremos essas distribuições a seguir.

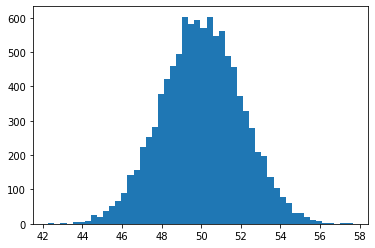
A distribuição uniforme

A proposição das distribuições parte de uma ideia de um caminho teórico “perfeito”, ao qual os dados se ajustam em maior ou menor grau. Se todos os valores possíveis aparecerem aproximadamente o mesmo número de vezes, estamos falando de uma distribuição uniforme. É útil ter em mente que, se gerarmos números aleatoriamente, seria conveniente que eles seguissem essa distribuição, pois isso garantiria que cada número tenha a mesma probabilidade de aparecimento. As funções de geração de números aleatórios que as calculadoras e planilhas possuem seguem essa distribuição. Na figura podemos ver um histograma de dados na forma de uma distribuição uniforme.

**

A distribuição normal

Muitos processos e variáveis ​​no mundo real seguem uma distribuição com uma forma particular chamada distribuição normal. Esta distribuição é composta por pontos que se agrupam simetricamente em torno de um valor médio e cuja variância não ultrapassa uma certa distância da média. Abaixo está um histograma com dados agrupados de acordo com uma distribuição normal.



Esses valores e características são profundamente estudados e desenvolvidos no campo da estatística. A importância da distribuição normal está em sua aparição em múltiplos campos do mundo real. Se pudermos verificar que a distribuição dos dados é aproximadamente normal, podemos usar suas propriedades e salvar muitas suposições sobre o comportamento dos dados.

Como uma ideia geral sobre a distribuição normal, consideremos as seguintes propriedades muito úteis que são cumpridas quando os dados apresentam essa distribuição.

* Os dados normais são simétricos em relação à média
* A média, mediana e moda têm aproximadamente o mesmo valor.
* A chamada “regra empírica” é geralmente cumprida a partir da qual:
  + 68% dos dados estão a cerca de 1 desvio padrão da média
  + 95% dos dados estão a cerca de 2 desvios padrão da média
  + 99,7% dos dados estão a cerca de 3 desvios padrão da média

Com isso, qualquer dado que esteja além de 3 vezes o valor do desvio padrão longe da média pode ser considerado um valor extremo ou discrepante. Veremos os valores extremos mais tarde.

