Tipos de dados

Há diferentes tipos de dados com os quais se pode trabalhar em Aprendizado de Máquina, e cada tipo vai implicar o uso de métodos e processos específicos por parte do programador. Há 3 grandes categorias: numéricos, categóricos e ordinais.

Numéricos

São coisas que podem ser medidas, representadas em números, comparadas por magnitude e manipuladas por métodos matemáticos. Há 2 subcategorias desse tipo de dado: discretos e contínuos.

Discretos

Dados numéricos discretos são aqueles representados por uma escala na qual sempre se tem finitos valores de um intervalo, geralmente inteiros. Não se pode ter quanta precisão se quer entre duas quantias, os valores saltam de um para o outro dado um valor mínimo pré-definido.

Exemplos: número de carros que uma família possui, número de hospitais numa cidade, idade em anos inteiros, tamanho da população de um país etc.

Contínuos

Dados numéricos contínuos são aqueles representados por uma escala na qual se pode ter infinitos valores dentro de um intervalo. Pode-se ter quanta precisão quiser na medida, pois os valores podem ser o quão pequenos se quer.

Exemplos: altura de uma pessoa, temperatura, velocidade, lucro, salário, quantidade de chuva etc.

Categóricos

Dados categóricos são aqueles separados por classes, grupos. Não se pode compará-los diretamente, pois não há uma relação pré-definida entre seus atributos como no caso de valores numa escala matemática. Podem até ser representados por números, como CPFs ou códigos de área telefônicos, mas a comparação matemática entre eles não necessariamente significa algo nesse contexto (dificilmente haveria utilidade em calcular o CPF médio de um grupo de pessoas, por exemplo).

Exemplos: classe social, nacionalidade, RG, time do coração, sexualidade, etnia, candidato de preferência, ocupação profissional etc.

Ordinais

São dados representados por uma escala de valores em que cada valor tem uma ordem específica dentro de um ranking, mas a diferença absoluta entre eles não significa nada dentro da análise. Ou seja, esses dados estabelecem uma relação de ordem entre certos valores, mas não indicam a magnitude das diferenças entre eles de forma precisa, algo que precisa ser feito através de outros tipos de dados.

Exemplos: classificação de um produto entre “ótimo”, “bom”, “médio”, “ruim”, “horrível” ou entre 1, 2, 3, 4 e 5 estrelas; classificação final dos candidatos em um vestibular; classificação de dor entre “muito forte”, “forte”, “moderada”, “fraca” e “muito fraca” ou numa escala inteira de 1 a 10 etc.

Média, mediana e moda

Média

A média de conjunto de dados é simplesmente a soma dos valores de todos os elementos no conjunto dividido pela quantidade de elementos. A média $\overline{X}$ de um conjunto de dados com n elementos se dá por:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Exemplo de cálculo:

C = {4, 1, 8, 6, 9}

$$\overline{X}_c = \frac{1+4+6+8+9}{5} = 5.6$$

Exemplos de uso da média: de filhos por pessoa num país, de renda mensal numa população, de gols por jogo de um jogador de futebol etc.

Mediana

A mediana de um conjunto de dados é o valor localizado no centro da sequência ordenada de forma crescente ou decrescente dos valores desse conjunto. Caso não haja um elemento exatamente centralizado, a mediana será a média do par de valores centralizado. A mediana tende a ser mais utilizada do que média por apresentar uma representação melhor do elemento típico do conjunto, já que é menos suscetível a outliers.

Exemplo de cálculo:

C = {4, 1, 8, 6, 9}

C_{cresc} = {1, 4, 6, 8, 9}

M = 6

Exemplo clássico de uso da mediana: determinar uma renda típica de um cidadão de um país sem a deformação da desigualdade.

Moda

É o valor mais comum, com maior número de aparições, dentro de um conjunto de dados. É uma medição geralmente feita com dados discretos. Exemplo de cálculo:

C = {3, 4, 6, 4, 8, 9, 8, 4, 10, 11, 4}

Mod = 4

Desvio Padrão e Variância

São as medidas mais usadas para quantificar a dispersão dos dados de um conjunto.

Variância (σ^2 ou s²)

É a média dos quadrados das somas das diferenças entre a média do conjunto e cada elemento.

A elevação ao quadrado faz com que resultados negativos do cálculo das diferenças não cancelem os positivos e, principalmente, dá mais peso a outliers, facilitando a identificação da existência deles através dessa medida e do desvio padrão. Por exemplo, a variância de uma população de dados com N elementos e média μ é:

$$\sigma^2 = \frac{\sum_{i=1}^{n} x_i - \mu}{N}$$

Caso se esteja lidando com uma amostra de tamanho n e média M da população, a variância s é dada por:

$$s^2 = \frac{\sum_{i=1}^{n} x_i - M}{n-1}$$

Exemplo de cálculo para uma população:

C = {3, 4, 5, 10, 9}

μ = (3 + 4+ 5 + 10 + 9)/5 = 6.2

$\sigma^2 = \frac{(3-6.2)^2+(4-6.2)^2+(5-6.2^2)+(10-6.2)^2+(9-6.2)^2}{5}$= 7.75

Desvio Padrão (σ ou s)

Medida mais usada para se falar da dispersão de dados. É a raiz quadrada da variância.

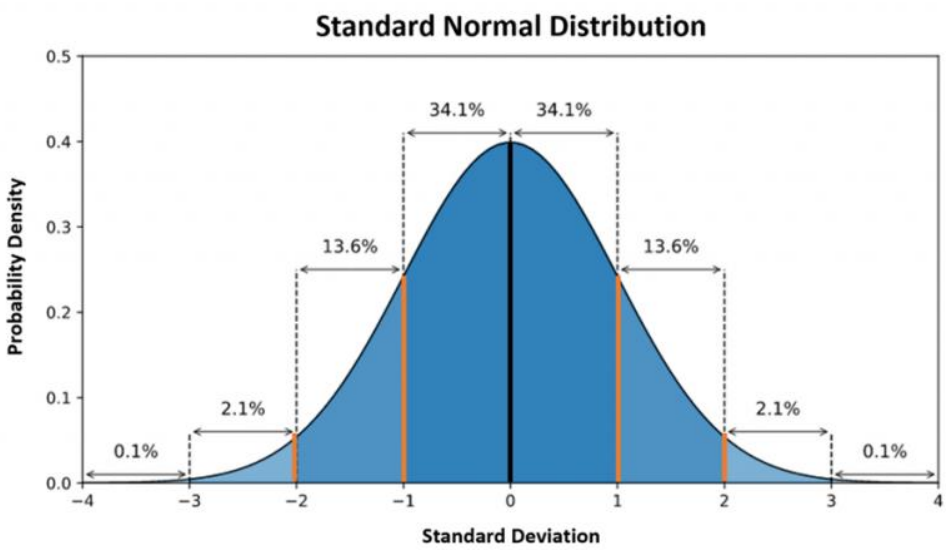
Um ponto é considerado ou não um outlier dependendo de quantos desvios padrões ele está da média do conjunto de dados.

A notação com sigma é para o populacional e com o s é para o amostral, seguindo a mesma lógica da variância.

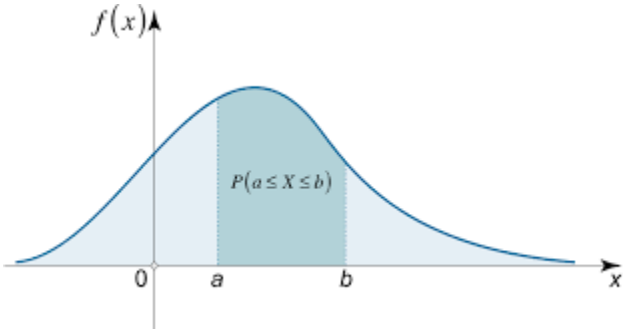
Função de Densidade de Probabilidade

Função que retorna a probabilidade de o valor de input ocorrer. Para se ter a probabilidade de um intervalo de valores ocorrerem, o que é algo mais usual de se calcular, deve-se calcular a área abaixo desse intervalo no gráfico da Função de Densidade.

Exemplo de Função de Densidade:



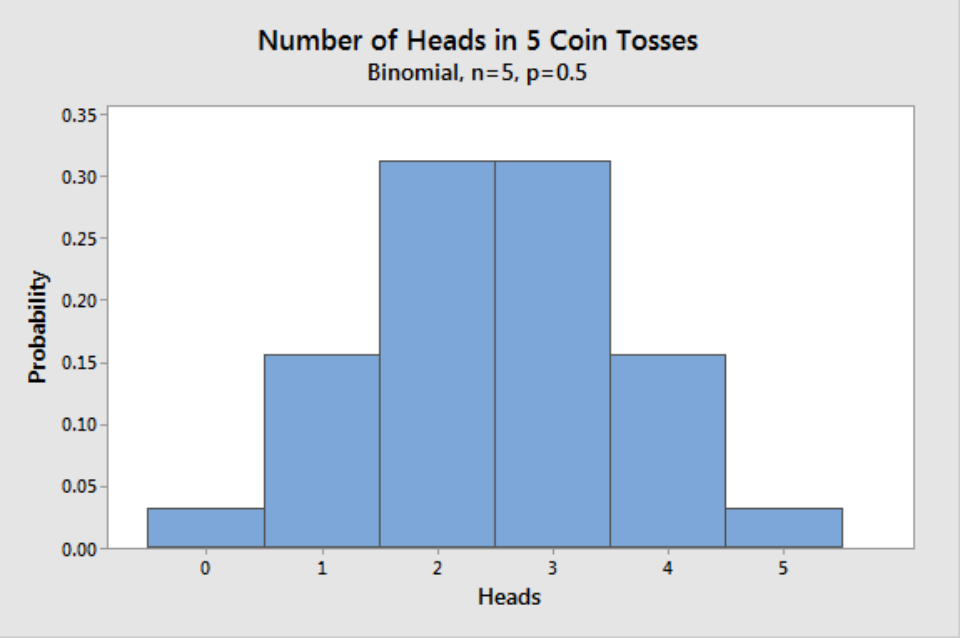
Probabilidade de um valor dentro do intervalo [a, b] ocorrer:



Função de Massa de Probabilidade

É análoga à Função de Densidade de Probabilidade, mas para dados discretos. Muito parecida com um histograma, já que o valor de probabilidade de um valor para o outro tem um salto, não é contínuo.

Exemplo:



Só tem como cara ou coroa cair um número inteiro de vezes, portanto a escala dos dados é discreta.

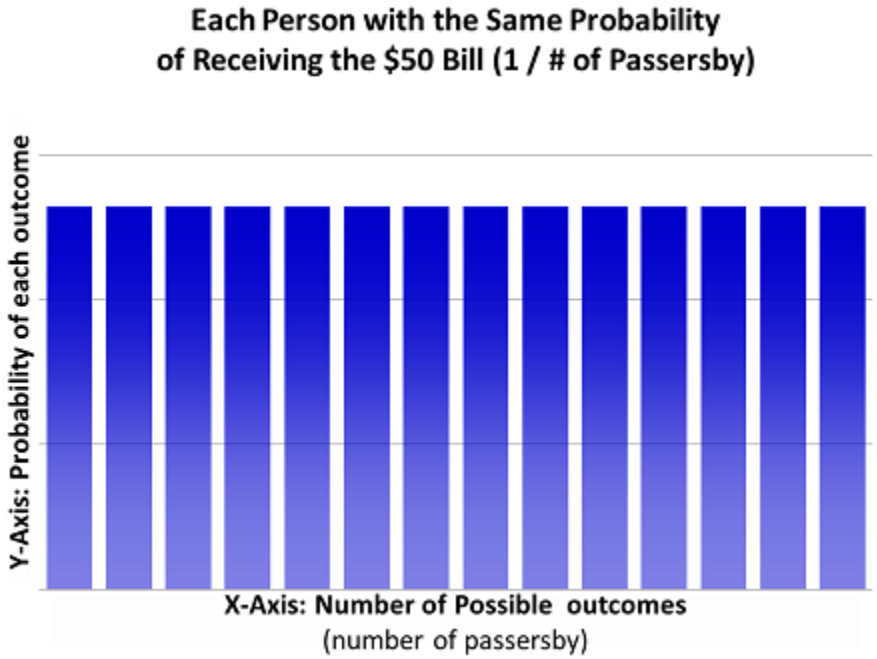
Tipos de distribuições

Há certos padrões de distribuição de probabilidade que se repetem muito na realidade podem ser separados em grupos, são as principais: uniforme, gaussiana, Poisson, binomial.

Uniforme

Quando todos os eventos do universo estudado têm a mesma (ou aproximadamente a mesma) probabilidade de ocorrerem. O gráfico da função de distribuição fica nivelado, formando uma área retangular.

Exemplo:

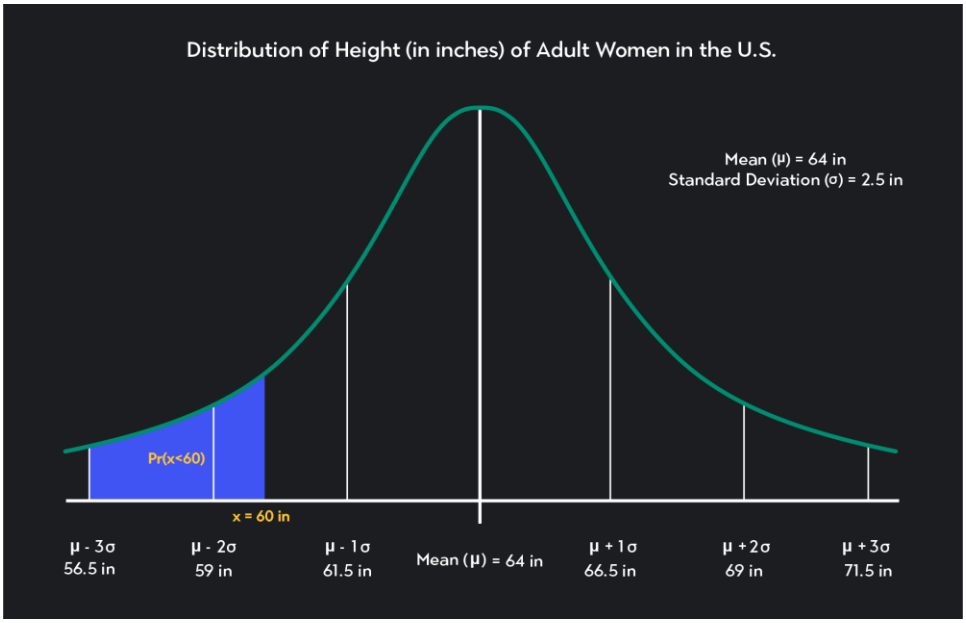


Probabilidade de cada participante de um sorteio de $50 ganhar o prêmio.

Normal (Gaussiana)

Distribuição em forma de sino: quanto mais afastado da média/mediana, menos provável é de o evento ocorrer. Ou seja, no universo estudado há certos intervalos e valores têm consideravelmente mais probabilidade de ocorrer do que outros.

Exemplo:

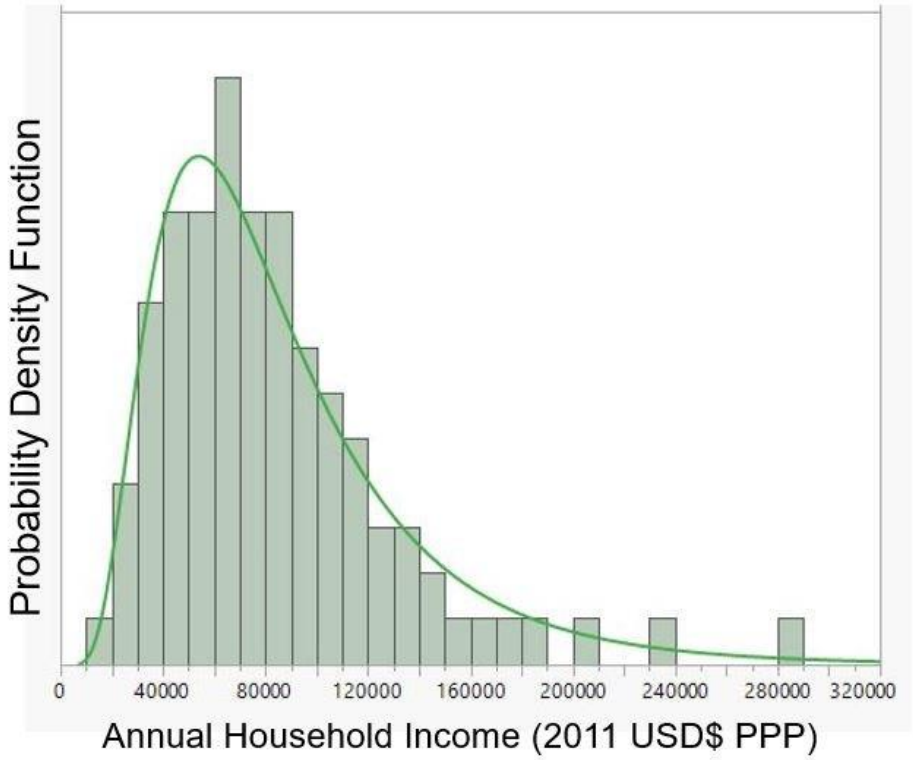


Distribuição da altura de mulheres nos EUA.

Exponencial (Power Law)

Distribuição muito enviesada para algum lado do universo de dados, com uma queda ou aumento acintoso de probabilidade de ocorrência de um intervalo para outro.

Exemplo:

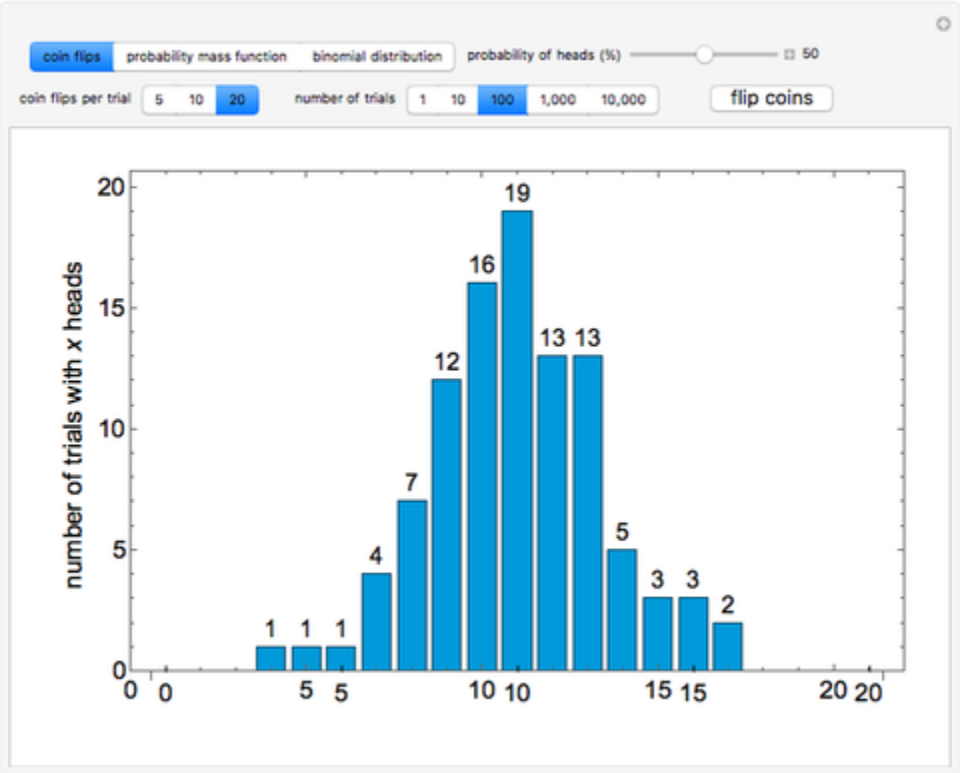


Distribuição de renda anual nos EUA.

Binomial

Distribuição muito parecida com a gaussiana, com a diferença de que aparece quando o universo de dados é discreto.

Exemplo:



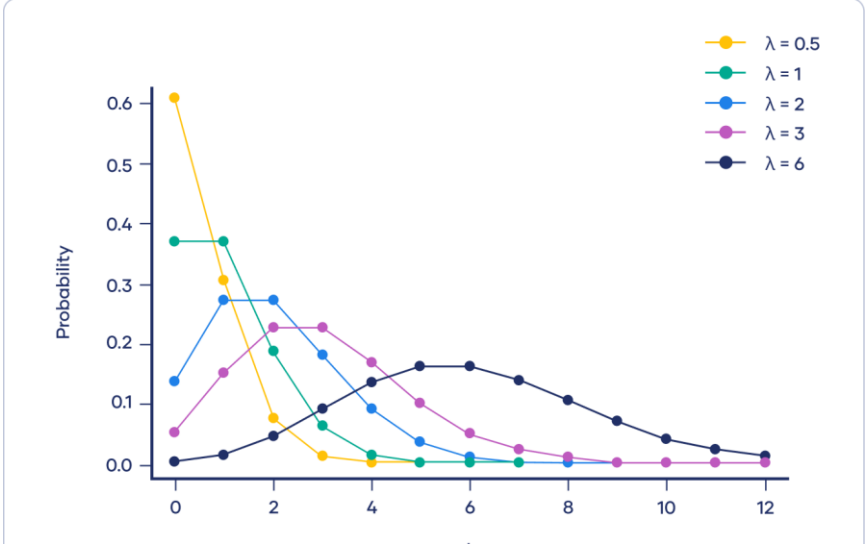
Quantas vezes (ordenada) saiu x caras (abscissa) num experimento de 20 jogadas de moedas.

Poisson

Distribuição que mede a probabilidade de um evento ocorrer k vezes (ou seja, aparece com dados discretos) em algum ponto ou intervalo de tempo baseado nas seguintes informações: média λ de vezes que o evento ocorre por unidade de tempo; conhecimento de que os eventos são independentes; conhecimento de que a probabilidade do evento ocorrer em cada instante é sempre a mesma. Considerando esses dados, tem-se que a probabilidade de um evento ocorrer k vezes é:

$$P(k;\lambda) = \frac{\lambda^k \times e^{-\lambda}}{k!}$$

Exemplos:

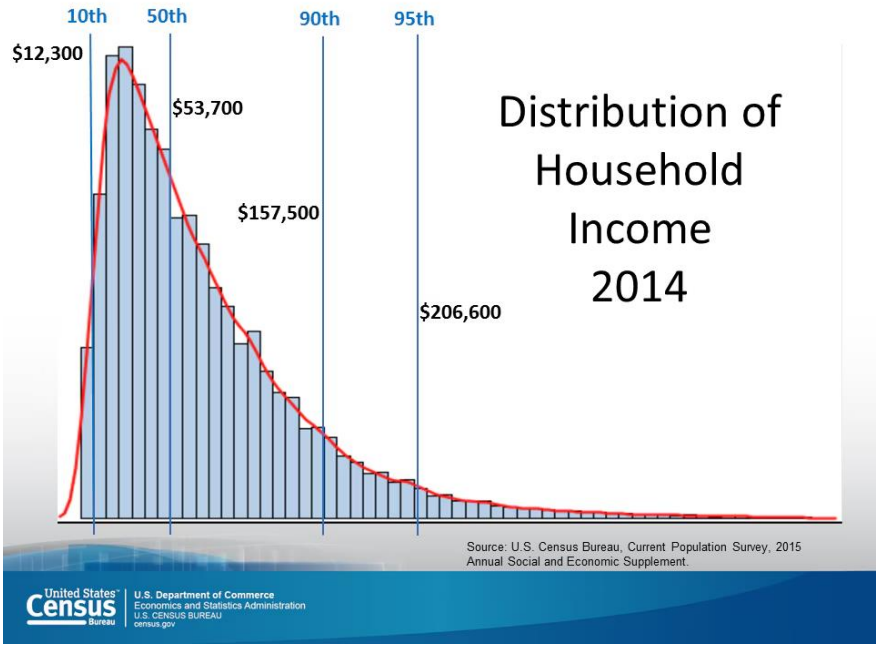


Percentil e momento

Percentil

Dado um conjunto ordenado de valores, o X° percentil é o valor tal qual X% da amostra ou população é menor ou igual.

Exemplo:



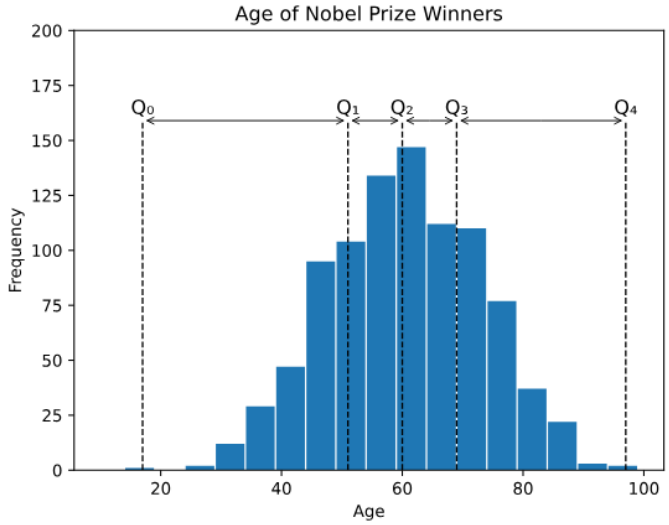
Nesse conjunto de dados, o 50° percentil indica que 50% da população americana em 2014 ganhava $53.700 ou menos por ano. Isso também poderia ser informado pela mediana, ou seja, o 50° percentil de um conjunto de dados é equivalente à sua mediana.

Quartil

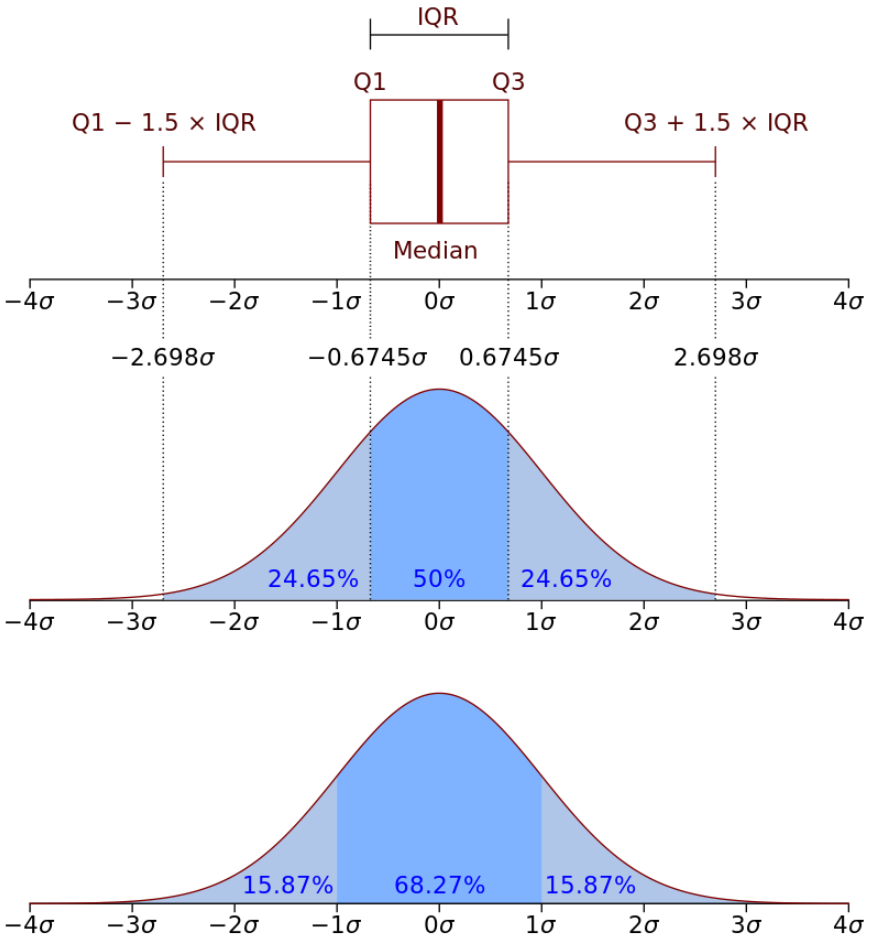
Percentis também são usados para se falar de quartis, que são os percentis que dividem os dados em 4 intervalos, ou seja, há 3 quartis.

O 1° quartil (Q1) é o 25° percentil, o 2° quartil (Q2) é o 50° percentil (equivalente à mediana) e o 3° quartil (Q3) é o 75° percentil.

A diferença entre Q3 e Q1 é a Amplitude Interquartil (Interquartile Range), que é um indicador de dispersão dos dados nos 50% centrais da distribuição.



Divisão dos quartis no conjunto das idades em anos inteiros de vencedores do Prêmio Nobel.



Exemplo de medida da Amplitude Interquartil.

Momento

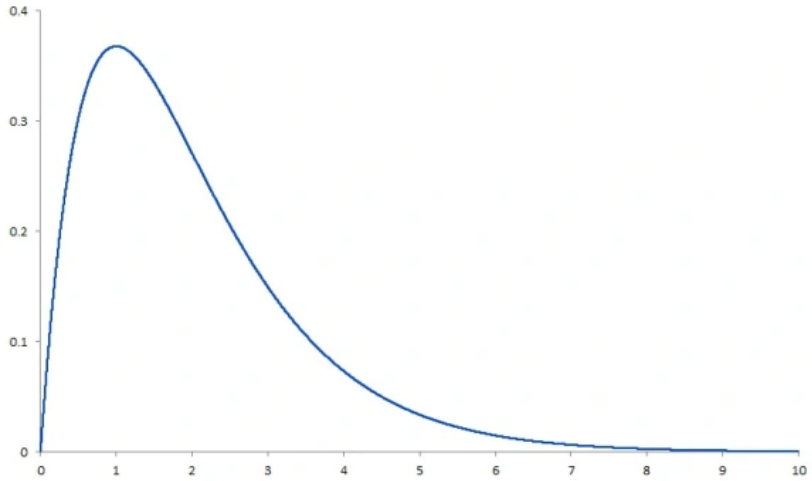
São medidas quantitativas que descrevem características de uma distribuição de probabilidade. O primeiro, segundo, terceiro e quarto são, respectivamente: média, variância, assimetria e curtose.

Média e variância já foram descritas anteriormente neste documento.

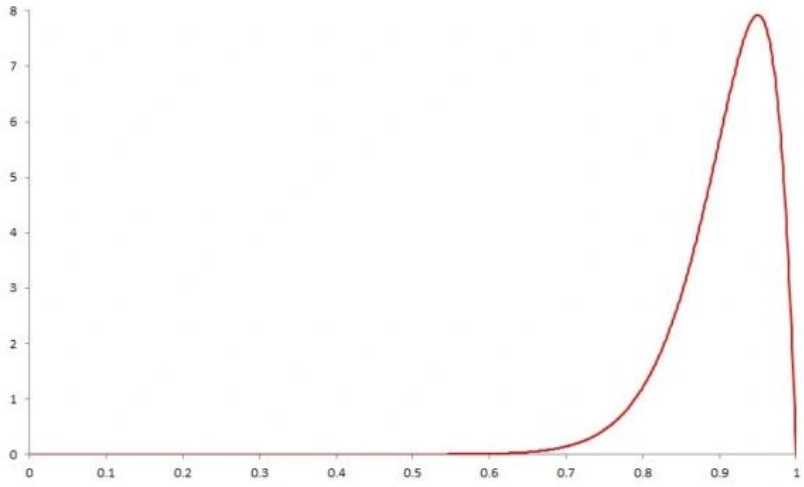
Assimetria

Uma distribuição pode ser concentrada para um dos lados, deixando o outro mais “magro” em comparação, formando uma cauda. Se essa cauda for no lado direito do gráfico, chamamos de assimetria positiva, se for para a esquerda, chamamos de assimetria negativa.

Exemplos:



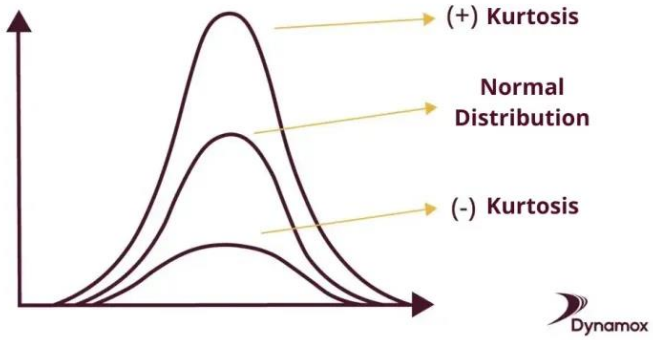
Distribuição com assimetria positiva.



Distribuição com assimetria negativa.

Curtose

É uma medida do quão concentrada no centro é uma distribuição em comparação com uma normal. Se for menos, a curtose é negativa, se for mais, é positiva:



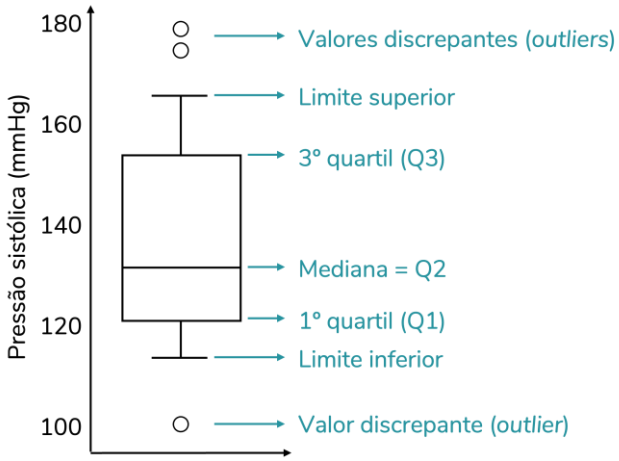
Matplotlib

Biblioteca com funções para a plotagem de dados. Suporta os mais diferentes tipos de gráficos, como de barras, histograma, de disperção, boxplot etc.

A implementação de suas principais funcionalidades está no arquivo código_aula.ipynb.

Box Plot

Tipo de gráfico muito útil para a visualização da dispersão de um conjunto de dados.



A caixa no centro é a Amplitude Interquartil (contém metade dos dados do conjunto), dentro dela está a mediana. Acima e abaixo da caixa há linhas verticais que se estendem até pequenas linhas horizontais, estes são os intervalos dos dados não discrepantes. Os valores discrepantes (outliers) geralmente são definidos como sendo aqueles com valor 1.5*IIQ acima de Q3 ou 1.5*IIQ abaixo de Q1.

Seaborn

Biblioteca para visualização de dados que, além de ter funções de plotagem próprias, muda as configurações das da matplotlib (assumindo que esteja importada) e deixa a visualização mais agradável e moderna.

Seaborn consegue plotar os mais diferentes tipos de gráfico, como boxplot, de dispersão, histograma, swarmplot etc.

Para mais detalhes e implementação, olhar o arquivo código_aula.ipynb.

Covariância e Correlação

Covariância

Suponha que há 2 conjuntos de dados de mesmo tamanho N organizados em 2 vetores, A e B. Calcula-se a diferença de cada valor em A para a sua média, formando um novo vetor com esses valores. Faz-se o mesmo com B. O produto escalar entre esses 2 vetores dividido por N (ou N-1, se forem amostras) é a covariância entre os dados dos vetores A e B. Matematicamente é equivalente ao cosseno entre os vetores de diferenças da média, ou seja, determina o quão juntos os dados de A e B variam da média.

$$cov_{x,y} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$$

Uma covariância próxima a zero indica que não há muita relação entre os dados dos 2 conjuntos, mas uma grande indica que sim. Mas o quão grande? É isso que a correlação determina.

Correlação

É a covariância dividida pelo desvio padrão de ambos os conjuntos de dados. Isso normaliza o valor da covariância para que se possa ter uma referência mais global do que é relação grande, moderada ou pequena entre 2 variáveis.

$$r = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2 (y_i - \overline{y})^2}}$$

Valores próximos a 1 indicam forte correlação positiva (o aumento do valor de uma variável tende a ser acompanhado pelo aumento no valor da outra).

Valores próximos a –1 indicam forte correlação negativa (o aumento do valor de uma variável tende a ser acompanhado pela diminuição do valor da outra).

Valores próximos a 0 indicam correlação fraca ou inexistente entre as variáveis.

OBS: correlação por si só não implica causalidade, só experimentos controlados e randomizados podem determinar causalidade. A correlação pode ser usada como motivadora e/ou ferramenta para esses experimentos.

Probabilidade Condicional

Diz respeito à probabilidade de um certo evento ocorrer dado que um certo outro ocorreu.

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

A fórmula acima diz que a probabilidade de A ocorrer dado que B ocorreu (P(A|B)) é igual à probabilidade de A e B ocorrerem simultaneamente (P(A, B)) dividida pela probabilidade de A ocorrer (P(A)).

Se A e B forem eventos independentes, P(A|B) deve ser aproximadamente igual à P(A), já que a ocorrência de B não influencia nem diz nada a respeito da ocorrência de A. Se forem diferentes, sabemos que há alguma relação de dependência entre os dois eventos.

Caso não sejam independentes, P(A, B) não é igual a P(A)*P(B), pois P(A) afeta P(B) e vice-versa.

Essas são formas também de detectar dependência entre duas variáveis.

Teorema de Bayes

Teorema que estabelece uma relação entre a probabilidade de A ocorrer dado que B ocorreu, com a probabilidade de B ocorrer dado que A ocorreu e com as probabilidades de cada evento ocorrer incondicionalmente:

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

Um insight importante a ser derivado desse conceito é que a probabilidade de algo que depende de B depende muito também da probabilidade incondicional desse mesmo evento e de B.

Exemplo:

Considere um teste para detecção de uma certa droga na corrente sanguínea de uma pessoa que tem 99% de precisão geral, isto é, identifica corretamente um usuário da droga 99% das vezes e corretamente retorna negativo para 99% dos não usuários. Aparentemente parece um teste muito bom, mas se considerarmos que apenas 0.3% da população usa a droga em questão, a história muda:

Considere evento A = ser um usuário da droga, evento B = testar positivo para a droga.

P(B) = probabilidade de testar positivo e ser um usuário + probabilidade de testar positivo e não ser um usuário

P(B) = 0.99*0.003 + 0.01 * 0.997 = 0.013

$P(A|B) = \frac{P(A)P(B|A)}{P(B)} = \frac{0.003 \times 0.99}{0.013} = 0.228 = 22.8\%$

Ou seja, há apenas 22.8% de chance de, caso alguém teste positivo para a droga, essa pessoa ser de fato usuária. É uma precisão muito baixa de algo que inicialmente parecia muito bom.

Só porque P(A|B) é alta (99%), não significa que P(B|A) também seja. Há de se considerar as probabilidades de base de A e B.