Análise de Dados com Linguagem Python

Distribuições de Probabilidade

Em teoria da probabilidade e em estatística, uma distribuição de probabilidade descreve o comportamento aleatório de um fenômeno dependente do acaso.

O estudo dos fenômenos aleatórios começou com o estudo dos jogos de azar – jogos de dados, sorteios de bolas de urna e cara ou coroa eram motivações para compreender e prever os experimentos aleatórios. Essas abordagens iniciais são fenômenos discretos, o que significa que o número de resultados possíveis é finito ou contável. Entretanto, certas questões revelam distribuições de probabilidade com suporte infinito não contável. Por exemplo, quando o lançamento de uma moeda tende ao infinito, o número de coroas aproxima-se de uma distribuição normal.

Flutuações e variabilidade estão presentes em quase todo valor que pode ser medido durante a observação de um fenômeno, independente de sua natureza, além disso quase todas as medidas possuem uma parte de erro intrínseco. A distribuição de probabilidade pode modelar incertezas e descrever fenômenos físicos, biológicos, econômicos, entre outros. O domínio da estatística permite o encontro das distribuições de probabilidade adaptadas aos fenômenos aleatórios.

Teoricamente uma descrição de probabilidade descreve a característica aleatória de uma experiência aleatória. O conceito de experiência aleatória surgiu para descrever um processo real de natureza experimental, em que o acaso intervém com resultados possíveis bem identificados. Por exemplo, em um lançamento de um dado não viciado (um evento aleatório) os resultados podem ser um número entre 1 e 6 com igual probabilidade (de acordo com a distribuição de probabilidade, há a mesma chance de saírem os seis resultados com probabilidade igual a um sexto).

Se os possíveis resultados dos fenômenos forem números contáveis, a distribuição de probabilidade é chamada discreta. Dar a distribuição de probabilidade significa dar a lista de valores possíveis com suas probabilidades associadas. Ela é dada por meio de uma fórmula, uma tabela de valores, uma árvore de probabilidade ou funções.

Em um contexto mais amplo, se os números dos resultados possíveis de um fenômeno aleatório forem finitos (contáveis ou incontáveis) em vez de infinitos, a distribuição de probabilidade descreve a distribuição de probabilidade dos resultados possíveis, mas caracterizados como funções (funções densidade, funções distribuição, entre outros) ou como medidas.

Confira este material complementar:

https://files.cercomp.ufg.br/weby/up/335/o/Um_passeio_hist%C3%B3rico_pelo_in%C3%ADcio_da_teoria_das_probabilidades-Mariana_Feiteiro_Cavalari_e_Ang%C3%A9lica_R._Cal%C3%A1bria.pdf

Análise de Dados com Linguagem Python

Distribuições Discretas

Distribuições de probabilidade em aplicações mais comuns são distribuições discretas e distribuições contínuas. Vejamos as distribuições discretas.

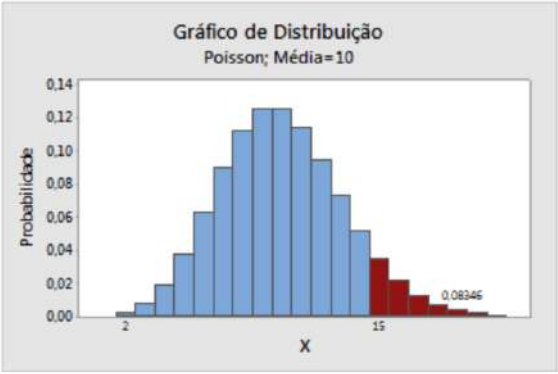
Uma distribuição discreta descreve a probabilidade de ocorrência de cada valor de uma variável aleatória discreta. Uma variável aleatória discreta é uma variável aleatória que tem valores contáveis, como uma lista de inteiros não negativos.

Com uma distribuição de probabilidade discreta, cada valor possível da variável aleatória discreta pode ser associado a uma probabilidade diferente de zero. Deste modo, uma distribuição de probabilidade discreta é, por vezes, apresentado em forma de tabela.

Com uma distribuição discreta, ao contrário de uma distribuição contínua, é possível calcular a probabilidade de que X é exatamente igual a algum valor. Por exemplo, você pode usar a distribuição discreta de Poisson para descrever o número de visitas de clientes em um dia. Suponha que o número médio de visiatas por dia seja de 10 e você queira saber a probabilidade de receber 5, 10 e 15 visitas de clientes em um dia:

x	P (X = x)
5	0,037833
10	0,12511
15	0,034718

Também é possível visualizar uma distribuição discreta em um gráfico de distribuição para ver as probabilidades entre intervalos.



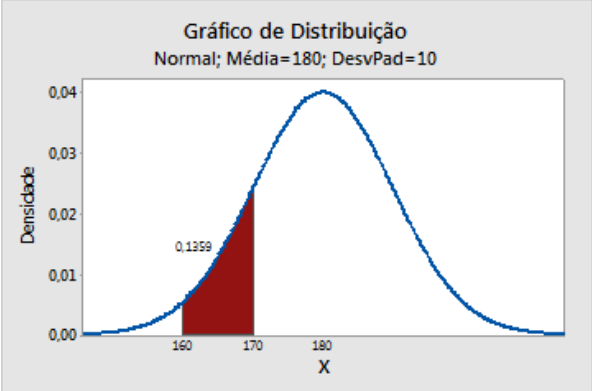
As barras sombreadas no gráfico acima representam o número de ocorrências quando as visiatas de clientes diárias forem 15 ou mais. A altura das barras somam 0,08346; por conseguinte, a probabilidade de que o número de visiatas por dia seja de 15 ou mais é 8,35%.

Análise de Dados com Linguagem Python

Distribuições Contínuas

A distribuição contínua descreve as probabilidades dos possíveis valores de uma variável aleatória contínua. Uma variável aleatória contínua é uma variável aleatória com um conjunto de valores possíveis (conhecidos como a intervalos) que é infinito e incontável.

As probabilidades de variáveis aleatórias contínuas (X) são definidas como a área sob a curva da sua distribuição. Assim, apenas as faixas de valores podem ter uma probabilidade diferente de zero. A probabilidade de que uma variável aleatória contínua seja igual a algum valor é sempre zero. Vejamos um exemplo com a variável aleatória peso (neste exemplo peso de um homem).



A região sombreada sob a curva, neste exemplo, representa o intervalo entre 160 e 170 libras (75 a 80 Kg). A área deste intervalo é 0,136; por conseguinte, a probabilidade de um homem selecionado aleatoriamente pesar entre 160 e 170 libras é de 13,6%. Toda a área sob a curva equivale a 1,0.

No entanto, a probabilidade de que X seja exatamente igual a algum valor é sempre zero porque a área sob a curva em um único ponto, que não tem nenhuma largura, é zero. Por exemplo, a probabilidade de um homem pesar exatamente 190 libras para a precisão infinita é zero. É possível calcular uma probabilidade não nula de que um homem pese mais do que 190 libras, ou menos do que 190 libras, ou entre 189,9 e 190,1 libras, mas a probabilidade de que ele pese exatamente 190 libras é zero.

Análise de Dados com Linguagem Python

Como Avaliar a Distribuição de Dados?

Você pode avaliar uma distribuição de dados através de gráficos, estatísticas descritivas ou comparação com uma distribuição teórica.

Gráficos

Os gráficos, como histogramas podem dar uma visão instantânea para a distribuição de um conjunto de dados. Os histogramas podem ajudá-lo a observar:

- Se os dados se agrupam em torno de um único valor ou se os dados têm vários picos ou modas.
- Se os dados estão espalhados finamente sobre uma grande variedade ou se os dados estão dentro de um pequeno intervalo.
- Se os dados são assimétricos ou simétricos.

Estatísticas Descritivas

As estatísticas descritivas que descrevem a tendência central (média, mediana) e dispersão (variância, desvio padrão) de dados com valores numéricos adicionam uma camada de detalhes e podem ser utilizadas para fazer comparações com outros conjuntos de dados.

Distribuições Teóricas

Finalmente, algumas distribuições comuns podem ser identificadas e são referidas pelo nome, como as distribuições normal, Weibull e exponencial. A distribuição normal, por exemplo, é sempre em forma de sino e simétrica em torno de um valor médio.

Análise de Dados com Linguagem Python

Coeficiente de Correlação

O coeficiente de correlação é uma medida estatística da força da relação entre os movimentos relativos de duas variáveis.

Os valores variam entre -1 e 1. Um coeficiente calculado maior que 1 ou menor que -1 significa que houve um erro na medição de correlação. Uma correlação de -1 mostra uma correlação negativa perfeita, enquanto uma correlação de 1 mostra uma correlação positiva perfeita. Uma correlação de 0 indica que não há relação linear entre o movimento de duas variáveis.

Por exemplo, um coeficiente de correlação poderia ser calculado para determinar o nível de correlação entre o preço do petróleo bruto e o preço das ações de uma empresa produtora de petróleo, como a Exxon Mobil Corporation. Uma vez que as empresas petrolíferas obtêm lucros maiores à medida que os preços do petróleo sobem, a correlação entre as duas variáveis é altamente positiva.

Existem vários tipos de coeficientes de correlação, mas o mais comum é a correlação de Pearson, que mede a força e a direção da relação linear entre duas variáveis. Ele não pode capturar relacionamentos não lineares entre duas variáveis e não pode diferenciar entre variáveis dependentes e independentes.

A força da relação varia em grau com base no valor do coeficiente de correlação. Por exemplo, um valor de 0,2 mostra que há uma correlação positiva entre duas variáveis, mas é fraca e provavelmente sem importância. Os analistas em alguns campos de estudo não consideram as correlações importantes até que o valor ultrapasse pelo menos 0,8. No entanto, um coeficiente de correlação com um valor absoluto de 0,9 ou mais representaria uma relação muito forte.

Análise de Dados com Linguagem Python

Calculando e Interpretando a Correlação

O coeficiente de correlação é calculado determinando primeiro a covariância das variáveis e, em seguida, dividindo essa quantidade pelo produto dos desvios padrão dessas variáveis. Ou, de forma mais simples em Python, usamos a função corr() em um dataframe do Pandas para calcular a correlação entre todas as variáveis, sempre em pares (a correlação mede a relação entre duas variáveis).

Interpretamos a correlação para compreender o comportamento de uma variável em relação a outra. Se a correlação é positiva, isso significa que se o valor de uma variável aumenta, o valor da outra variável também aumenta. Na correlação negativa, aumenta o valor de uma variável, diminui o valor da outra.

Mas atenção: correlação não implica causalidade. Não podemos estabelecer uma relação de causa com base na correlação (para estabelecer relação de causa usamos análise causal). Por exemplo:

Considere a venda de sorvetes e a temperatura. Se aumenta a temperatura, aumenta a venda de sorvetes, certo? Isso indica que as variáveis têm uma correlação positiva, mas não podemos afirmar que uma variável causa a outra. Para relação de causa devemos fazer outras análises, incluindo a análise causal.

Análise de Dados com Linguagem Python

Projeto 6

Market Basket Analysis

Este é um projeto bem especial. Usaremos uma das técnicas de análise mais úteis para a área de varejo, o Market Basket Analysis (MBA) ou Análise de Cesta de Compras.

O MBA é uma técnica intensiva em termos matemáticos e estatísticos e aplicaremos a técnica sem construir uma única fórmula matemática. Usaremos pacotes e funções da Linguagem Python. Também faremos um extenso trabalho de análise de dados.

Todos os scripts estão ao final do capítulo

Análise de Dados com Linguagem Python

Projeto 6

Definição do Problema

Nosso objetivo deste projeto é identificar associação entre produtos de uma cesta de compras. Quem compra o Produto A, também compra o Produto B? Qual a força dessa relação, se ela existir?

Vamos analisar dados de milhões de transações aplicando Market Basket Analysis.

Análise de Dados com Linguagem Python

Projeto 6

Fonte de Dados

Para este projeto usaremos dados disponíveis no link abaixo:

https://www.kaggle.com/c/instacart-market-basket-analysis/data

Você deve criar sua conta no Kaggle e fazer o download de todos os arquivos. Descompacte os arquivos e coloque na pasta dados, dentro da mesma pasta onde estiver o Jupyter Notebook do Projeto 6. O artigo abaixo explica em detalhes como os dados foram gerados:

https://tech.instacart.com/3-million-instacart-orders-open-sourced-d40d29ead6f2