Probabilidade e Estatística

Probabilidade:

- Estudo das chances de ocorrência de um resultado (porcentagem)

Estatística:

- Ciência que utiliza as teorias probabilísticas para explicar a frequência da ocorrência de eventos

Regras da Probabilidade:

- Probabilidades são sempre positivas

- A soma das probabilidades de um evento é 1

- A probabilidade condicionada de dois eventos é o produto da probabilidade do primeiro e a do outro, condicionada ao primeiro

Probabilidade Condicional:

P(A/B): A probabilidade de A dado a condição B

P(A/B) = P(A e B)/P(B)

Fórmula de Bayes

P (A/B) = (P(B/A) * P(A)) / P(B)

- A ideia central da estatística Bayesiana é atualizar a certeza de uma pessoa após ela ser exposta a novas evidências

Probabilidade (A, B):

(A, B) = (B / A) * p(A)

Abordagem Clássica

P (A) = Ocorrências do evento / Total de eventos

Mais matemática

Abordagem Frequentísta

P (A) = lim Ocorrência do evento / Total

n → infinito (n tendendo ao infinito)

Variáveis:

*Tipos de variáveis:

Dados:

- Categorizados

Escala Nominal / Escala Ordinal

Ex: Lista de nomes, lista de menções

- Quantitativos

Escala Intervalar

Ex: Lista de salários, preços diários, número de usuários

- Seção transversal

Todas as observações são realizadas ao mesmo tempo (em uma época definida)

Ex: Censo (10 anos)

- Série temporal

Observações realizadas ao longo do tempo

* Variáveis aleatórias

Definição: Variável com valor incerto

- Discretas: Passos discretos

Ex: Velocímetro (não existe 59, 9 km/h)

- Contínuas: O valor não é definido

- Como calcular o valor esperado por uma variável aleatória:

$$E(X) = x1p1 + x2p2 + ...$$

*Controle do fluxo de execução

- Em python o controle de fluxo é feito por identação

Variância:

- Dado um conjunto de dados, a variância é uma medida de dispersão que mostra o quão distante cada valor desse conjunto está do valor central (médio)

- Quanto menor é a variância, mais próximos os valores estão da média; mas quanto maior ela é, mais os valores estão distantes da média

- Considere que X1, X2, ..., Xn são os n elementos de uma amostra e que X é a média aritmética desses elementos. O cálculo da variância amostral é dado por:

$$\text{Var. amostral} = \frac{(x_1 - x)^2 + (x_2 - x)^2 + (x_3 - x)^2 + \ldots + (x_n - x)^2}{n - 1}$$

Se, em contrapartida, quisermos calcular a variância populacional, consideraremos todos os elementos da população, e não apenas de uma amostra. Nesse caso, o cálculo possui uma pequena diferença.

Observe:

$$\text{Var. populacional} = \frac{(x_1 - x)^2 + (x_2 - x)^2 + (x_3 - x)^2 + \ldots + (x_n - x)^2}{n}$$

Desvio Padrão

- O desvio padrão é capaz de identificar o “erro” em um conjunto de dados, caso quiséssemos substituir um dos valores coletados pela média aritmética

- O desvio padrão aparece junto à média aritmética, informando o quão “confiável” é esse valor. Ele é apresentado da seguinte forma:

$$\text{média aritmética } (x) \pm \text{desvio padrão } (dp)$$

- O cálculo do desvio padrão é feito a partir da raiz quadrada positiva da variância. Portanto:

$$dp = \sqrt{var}$$

Teorema Central do Limite:

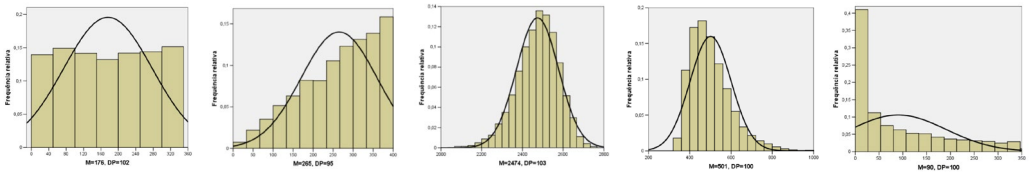
- Afirma que a soma (S) de N variáveis aleatórias independentes (X), com qualquer distribuição e variâncias semelhantes, é uma variável com distribuição que se aproxima da distribuição de Gauss (distribuição normal) quando N aumenta

- A média de S é o somatório das médias de X

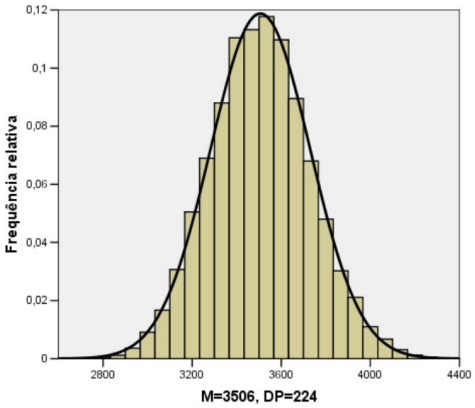
- A variância de S é o somatório das médias de X

Exemplo:

Para a soma de cinco variáveis aleatórias distribuídas arbitrariamente através do Método de Monte Carlo



A variável SOMA das 5 variáveis, em acordo com o TCL, apresenta distribuição aproximadamente gaussiana ou normal:



M = 176 + 265 + 2474 + 501 + 90 = 3506

- O teorema Central do Limite (TCL), aplicado às médias amostrais de uma variável aleatória (X) com qualquer distribuição e variância finita, implica que as médias amostrais apresentam distribuições tendendo à distribuição normal conforme o número de observações nas amostras (n) cresce

- A média das médias amostrais é igual à média de X

- O desvio padrão das médias amostrais é igual ao desvio padrão de X dividido pela raiz quadrada de n

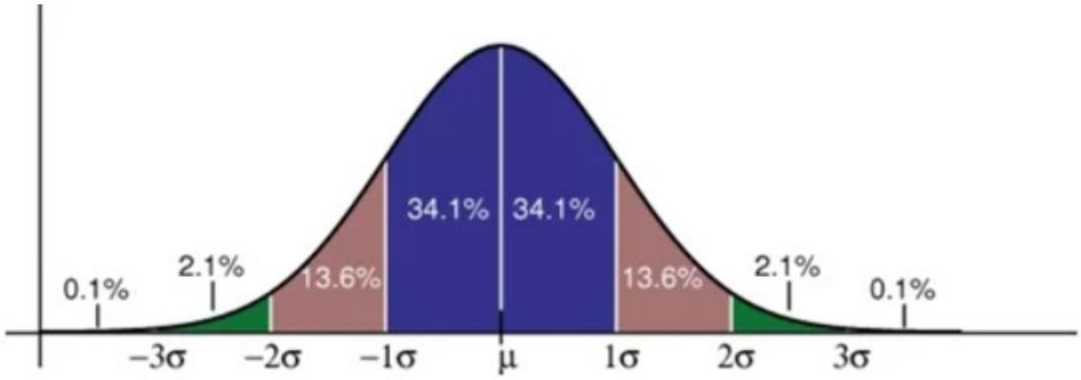
Teoria dos grandes números

- A teoria afirma que quanto maior o número de elementos no estudo, mais a probabilidade calculada se aproxima da probabilidade real

Exemplo:

No lançamento de 100 vezes de 1 moeda pode dar 54 vezes cara e 46 coroa. Dessa forma a probabilidade calculada é 54% para dar cara. Mas se for jogada infinitas vezes, é esperado, teoricamente, que a probabilidade calculada seja 50% (igual a probabilidade real).

Curva normal



- Na curva normal a média está sempre no meio da curva

Moda

- A moda é o termo que mais se repete em uma distribuição

Mediana

- É uma ideia geométrica da distribuição

Z score

(X – média) / desvio padrão = z score

Correlação e regressão

- Quando duas (ou mais) variáveis possuem relação entre si, dizemos que há um grau de correlação entre elas, por exemplo, o tamanho de uma circunferência e o tamanho de seu raio

- No caso da circunferência e seu raio, a correlação é perfeita, uma vez que o tamanho da circunferência é dado por uma fórmula claramente dependente do tamanho do raio ($C = 2\pi R$)

- O lançamento de dois dados honestos é um evento completamente independente um do outro. Nesse caso, não se espera qualquer relação entre os resultados. Logo não há correlação entre eles.

- Quando se analisa somente a relação entre duas variáveis, diz-se que a correlação (e regressão) é simples

- Quando se analisa a relação entre mais de duas variáveis, diz-se que a correlação (e regressão) é múltipla

- Assim, Correlação e Regressão são técnicas que visam estimar a relação que possa existir entre variáveis

- Correlação resume o grau de relacionamento entre as variáveis em um escala de (-1 a 1)

- Regressão tem como resultado uma fórmula matemática que descreve o relacionamento entre variáveis

Coeficientes de correlação

1) Coeficiente de correlação de Matthews: Fornece uma medida de precisão entre as variáveis. Muito utilizada em Data Mining para construção de matriz de confusão

2) Coeficiente de correlação de Kendall: Verifica a semelhança de dados ordinais. Muito utilizada em psicologia e Machine Learning.

3) Coeficiente de correlação de Wilcoxon: Utilizado quando não há certeza sobre a normalidade das variáveis (estatística não-paramétrica). Ou seja, quando não há uma distribuição normal.

4) Coeficiente de correlação de Spearman (MAIS UTILIZADO): Utiliza-se da ordem das observações e não seus valores. Pode ser utilizado para qualquer tipo de relação entre variáveis, não somente a linear

Interpretação dos dados

Pela função describe():

- describe() é uma função que oferece estatísticas descritivas básicas sobre um DataFrame ou uma Series, incluindo a média, desvio padrão , valores mínimo e máximo, quartis, entre outros

1) Media e Mediana:

Se a média e a mediana são aproximadamente iguais, isso pode sugerir uma distribuição simétrica

2) Desvio Padrão:

Um desvio padrão alto indica que os dados têm uma dispersão maior em relação à média, o que pode sugerir uma distribuição mais ampla ou mais variável

3) Valores Mínimo e Máximo:

Podem fornecer uma ideia da amplitude dos dados. Se a diferença entre o valor máximo e mínimo for grande, pode indicar uma distribuição ampla

4) Quartis

Os quartis podem ajudar a entender como os dados estão distribuídos em diferentes partes do conjunto. Por exemplo, se há uma grande diferença entre o terceiro quartil (75%) e o primeiro quartil (25%), isso pode indicar uma distribuição não uniforme

Pelo boxplot:

- É possível usar o boxplot para ter uma ideia da distribuição dos dados de maneira visual. O boxplot é uma representação gráfica que fornece informações sobre a mediana, os quartis, os possíveis outliers e a amplitude dos dados. Ele pode ser útil para identificar a simetria, dispersão e possíveis valores atípicos na distribuição dos dados

- Sua interpretação inclui:

1) A linha no meio do retângulo representa a mediana

2) A caixa (ou retângulo) inferior representa o primeiro quartil (25%) e o superior o terceiro quartil (75%)

3) As linhas verticais (“whiskers”) estendem-se do quartil inferior ao mínimo e do quartil superior ao máximo (a menos que existam outliers)

4) Os pontos fora das whiskers podem ser considerados outliers, dependendo do critério de detecção de outliers

Pelo histograma dos Z-scores:

- O histograma dos Z-scores pode ajudar a identificar a distribuição dos dados padronizados em relação à média e ao desvio padrão. Normalmente, espera-se que esse histograma se assemelhe a uma distribuição normal padrão, com a maioria dos valores centrados em torno de 0 e com um desvio padrão de 1, se os dados originais seguirem uma distribuição normal ou se forem aproximadamente normalizados

Z-score (continuação):

- O Z-score é uma medida estatística que descreve a relação de um valor com a média de um grupo de valores. Ele é medido em termos de desvios padrão da média. Se a pontuação Z for 0, indica que a pontuação do ponto de dados é idêntica à pontuação média.

- Basicamente, o Z-score é o número de desvios padrão em relação à média de um ponto de informação.

- Ele também pode ser interpretado como uma proporção do número de desvios padrão abaixo ou acima da população, o que significa uma pontuação bruta.

1) Padronização:

→ O Z-score padroniza os dados, transformando-os em uma escala com média zero e desvio padrão igual a um.

2) Interpretação:

→ O resultado do Z-score mostra a posição relativa do dado em relação aos outros. Um Z-score positivo indica que o valor está acima da média, enquanto um Z-score negativo indica que está abaixo da média.

3) Aplicações:

→ O Z-score é usado em várias áreas, como finanças, estatística e psicologia, para comparar resultados de testes ou estudos com uma população “comum” e entender onde um valor se encontra em relação à média.

Resumo:

→ Em resumo, o Z-score nos ajuda a contextualizar e comparar dados, considerando variabilidade da população.