1. **Qual a diferença entre estatística descritiva e estatística inferencial? Forneça exemplos de aplicação para cada uma delas.**

2. **Explique o conceito de média, mediana e moda. Quando você usaria cada uma dessas medidas de tendência central?**

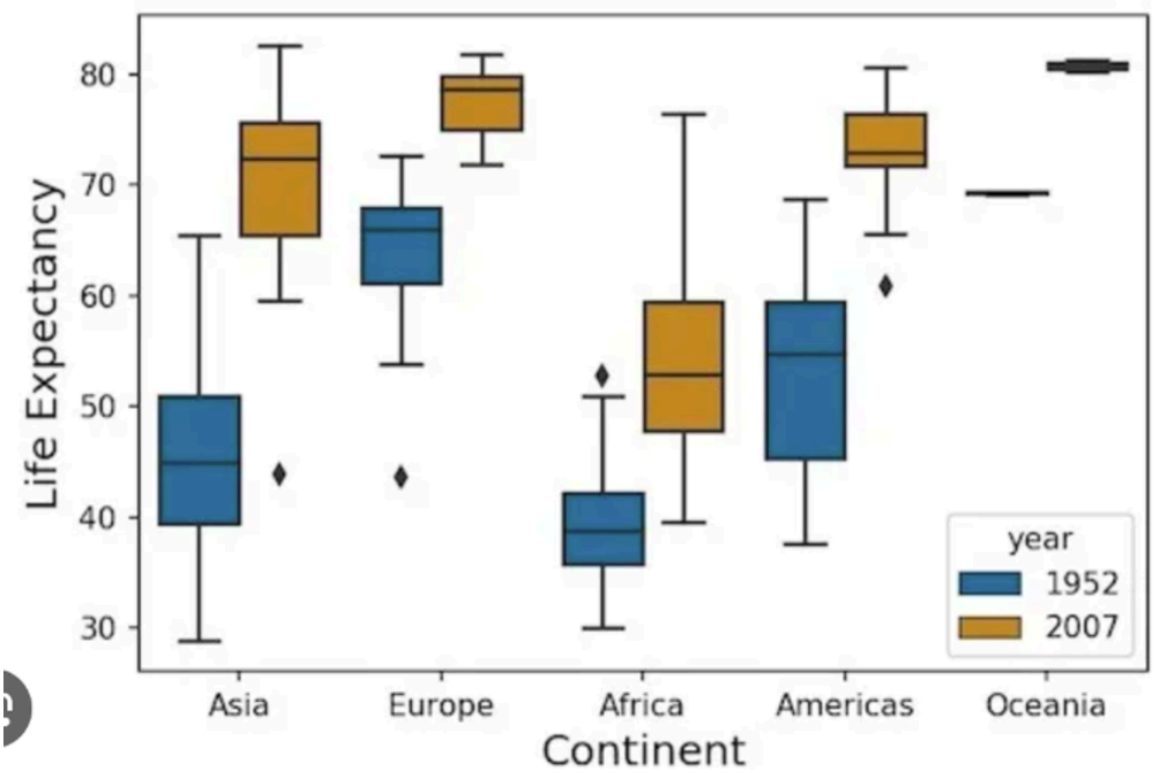
3. **Quais são as principais medidas usadas para descrever a variabilidade dos dados?**

4. **Suponha que você tenha um conjunto de dados com uma distribuição assimétrica à direita. Como você interpretaria essa assimetria e que medida de tendência central seria mais apropriada para descrever os dados?**

5. **Como a variância e o desvio padrão são utilizados para medir a dispersão dos dados? Qual é a relação entre essas duas medidas?**

6. **O que são outliers? O que fazer com eles? Quais tipos de medidas eles podem afetar?**

7. **Quais conclusões você tira com o gráfico abaixo?**



8. **Um plus: Na análise de dados, é comum encontrar dados faltantes. O que é isso? Por que eles existem? Como eles afetam nossa análise? O que fazemos com eles?**

Respostas

Respostas

1. **Qual a diferença entre estatística descritiva e estatística inferencial? Forneça exemplos de aplicação para cada uma delas.**

A estatística descritiva é uma área da estatística que se concentra na descrição e análise de dados, com o objetivo de resumir e apresentar informações de maneira facilmente compreensível. Ela é amplamente utilizada em áreas como negócios, finanças, ciências sociais e pesquisa de mercado. Alguns exemplos de aplicação incluem a determinação da média de idade dos clientes de um banco, a identificação do produto mais vendido em uma loja ou a análise do desempenho de uma equipe em uma temporada. Já a estatística inferencial é uma área da estatística que envolve a análise de dados para fazer inferências sobre uma população com base em uma amostra. Ela é usada para tirar conclusões e fazer previsões sobre um grande grupo de pessoas ou coisas com base em um pequeno grupo. Um exemplo de aplicação seria a determinação da margem de erro em uma pesquisa de opinião pública ou a previsão do número de vendas de um produto com base em dados históricos.

2. **Explique o conceito de média, mediana e moda. Quando você usaria cada uma dessas medidas de tendência central?**

A média é a medida de tendência central mais comum, obtida pela soma de todos os valores em um conjunto de dados e dividindo pelo número total de valores. Ela é usada para descrever um conjunto de dados quando se deseja saber o valor central típico. A média é sensível a valores extremos e pode ser distorcida por eles. A mediana é a medida de tendência central que divide um conjunto de dados em duas partes iguais, sendo que metade dos valores são maiores e metade são menores que ela. Ela é usada quando há valores extremos no conjunto de dados ou quando a distribuição dos dados é assimétrica. A mediana não é afetada por valores extremos. A moda é a medida de tendência central que representa o valor mais frequente em um conjunto de dados. Ela é usada para identificar o valor mais comum ou representativo em um conjunto de dados. É usada especialmente para dados categóricos

3. **Quais são as principais medidas usadas para descrever a variabilidade dos dados?**

Amplitude (Range): É a diferença entre o maior e o menor valor no conjunto de dados. Embora seja simples de calcular, a amplitude pode ser sensível a valores extremos e não fornece informações detalhadas sobre a distribuição dos dados.

Variância: A variância mede o quão distantes os valores estão da média. Valores mais dispersos têm uma variância maior. No entanto, a unidade da variância é o quadrado da unidade dos dados originais, o que pode dificultar a interpretação direta.

Desvio Padrão: O desvio padrão é a raiz quadrada da variância e é uma medida de dispersão mais comumente usada. Ele tem a mesma unidade dos dados originais, o que facilita a interpretação. Um desvio padrão maior indica maior dispersão dos valores em relação à média.

Intervalo Interquartil (IQR): O IQR é a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1) de um conjunto de dados. Ele fornece uma medida da dispersão dos valores que não é afetada por valores extremos, sendo útil quando há presença de outliers.

Coeficiente de Variação: O coeficiente de variação é calculado como o desvio padrão dividido pela média, multiplicado por 100. Ele fornece uma medida relativa de variabilidade, permitindo comparar a dispersão entre diferentes conjuntos de dados, mesmo que eles tenham unidades diferentes.

4. **Suponha que você tenha um conjunto de dados com uma distribuição assimétrica à direita. Como você interpretaria essa assimetria e que medida de tendência central seria mais apropriada para descrever os dados?**

Uma distribuição assimétrica à direita é caracterizada pelo fato de que a **maior parte dos dados está concentrada à esquerda**, enquanto a cauda longa se estende para a direita. Isso significa que há valores maiores do que a média (ou a mediana) que puxam a distribuição para essa direção.

Nesse cenário, a média seria influenciada por esses valores extremamente altos, resultando em um valor maior do que a mediana. A mediana, por outro lado, seria uma medida mais apropriada de tendência central para descrever os dados, uma vez que ela não é tão afetada por valores extremos quanto a média. A mediana representa o valor no centro da distribuição, de forma que metade dos valores esteja abaixo dela e a outra metade acima.

A média pode ser distorcida para cima pela presença de valores extremos, o que não refletiria adequadamente a concentração da maioria dos dados à esquerda. Portanto, em distribuições assimétricas à direita, a mediana é geralmente preferida como medida de tendência central, pois ela oferece uma visão mais robusta da posição central dos dados, independentemente da presença de valores extremos.

5. **Como a variância e o desvio padrão são utilizados para medir a dispersão dos dados? Qual é a relação entre essas duas medidas?**

A variância e o desvio padrão são medidas que quantificam a dispersão ou variabilidade dos dados em relação à média. Eles indicam o quão distantes os valores individuais estão da média do conjunto de dados. Vamos explorar como essas medidas são calculadas e a relação entre elas:

Variância: A variância é uma medida da média dos quadrados dos desvios de cada valor em relação à média do conjunto de dados. Ela é calculada usando a seguinte fórmula:

$$\text{Variância}(\sigma^2) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

onde:

- n é o número de observações no conjunto de dados.
- xi é cada valor individual no conjunto de dados.
- xbarra é a média dos valores do conjunto de dados.

Desvio Padrão: O desvio padrão é a raiz quadrada da variância e fornece uma medida da dispersão dos valores em relação à média. Ele é calculado usando a fórmula:

$$\text{Desvio Padrão}(\sigma) = \sqrt{\text{Variância}(\sigma^2)}$$

A relação entre a variância e o desvio padrão é simplesmente a raiz quadrada. O desvio padrão é uma medida mais intuitiva e interpretável, pois tem a mesma unidade dos dados originais. Isso facilita a compreensão da dispersão em termos das unidades em que os dados estão expressos.

Em termos de interpretação, um valor maior de variância ou desvio padrão indica maior dispersão dos valores em relação à média. Valores menores indicam que os dados estão mais agrupados em torno da média.

Em resumo, a variância e o desvio padrão são medidas essenciais para quantificar a dispersão dos dados e entender como os valores se espalham em relação à média. O desvio padrão é frequentemente preferido porque tem uma interpretação mais intuitiva, mas ambos são importantes para a análise estatística e a compreensão das características dos dados.

6. **O que são outliers? O que fazer com eles? Quais tipos de medidas eles podem afetar?**

Outliers, também conhecidos como valores atípicos, são observações que se desviam significativamente do restante dos dados em um conjunto. Essas observações são muito diferentes dos outros valores e podem ser tanto extremamente altas quanto extremamente baixas em relação à média do conjunto de dados. Outliers podem ser resultados de erros de medição, processos incomuns ou eventos raros, entre outras razões.

A presença de outliers pode afetar a análise estatística e as conclusões que podem ser tiradas a partir dos dados. Eles podem distorcer as medidas de tendência central e a estimativa da variabilidade dos dados. Alguns dos principais efeitos dos outliers incluem:

1. **Média e Desvio Padrão:** Os outliers podem aumentar artificialmente a média, puxando-a na direção do valor atípico. Além disso, a presença de outliers pode aumentar significativamente a variância e, consequentemente, o desvio padrão.
2. **Regressão Linear:** Outliers podem influenciar negativamente a qualidade de um modelo de regressão linear, levando a ajustes inadequados e previsões imprecisas.
3. **Testes Estatísticos:** Alguns testes estatísticos assumem que os dados estão distribuídos normalmente e não têm valores extremos. A presença de outliers pode violar essas suposições, resultando em resultados incorretos ou tendenciosos.
4. **Análise Descritiva:** Outliers podem distorcer a interpretação dos resultados em análises descritivas e gráficos, dificultando a identificação de padrões reais nos dados.

O tratamento de outliers depende do contexto e dos objetivos da análise. Aqui estão algumas abordagens comuns:

1. **Identificação e Remoção:** Identificar outliers através de métodos estatísticos, como o uso de gráficos de caixa (box plots) ou cálculos de distância entre os valores e, se justificado, removê-los do conjunto de dados. No entanto, a remoção de outliers deve ser realizada com cautela e justificação sólida. Somente devemos remover aquilo que de fato não faz sentido ou é um erro.

2. **Transformações:** Em alguns casos, é possível aplicar transformações aos dados para reduzir o impacto dos outliers. Por exemplo, a transformação logarítmica pode suavizar a influência de valores extremos.

3. **Tratar Separadamente:** Dependendo do contexto, os outliers podem ser tratados separadamente em análises específicas, ou até mesmo podem ser o foco da análise se forem de interesse para o estudo.

4. **Utilizar Métodos Robustos:** Em análises onde outliers são uma preocupação, utilizar métodos estatísticos robustos que são menos sensíveis a valores extremos pode ser uma abordagem válida.

7. Quais conclusões você tira com o gráfico abaixo?

Esse gráfico demonstra a expectativa de vida em diversos continentes em 1952 e 2007. Numa breve observação é possível criar a hipótese que a expectativa de vida aumentou em todos os continentes em 2007, mas como há um overlap entre os boxplots, seria necessário um teste de hipóteses.

Ásia: Em 1952, o limite inferior de expectativa de vida foi 30 anos, e o limite superior em torno de 68 anos. Já a mediana é em torno de 45 anos. O Q3 é levemente maior, mas não consigo afirmar se há uma distribuição à direita ou se é próximo de uma distribuição normal, nesse caso, precisaria observar os valores numéricos de skew e kurtosis. Em 2007, o limite inferior de expectativa de vida foi 60 anos e o limite superior mais de 80 anos. A mediana está em torno de 72. Neste caso, há uma distribuição à esquerda, pois o Q1 é maior. Também é possível notar que houve um outlier inferior aos 45 anos, idade que foge um pouco das observações.

Europa: Em 1952, o limite inferior de expectativa de vida foi em torno de 55 anos, sendo a mediana em 68 e o limite superior 72. Também se trata de uma distribuição à esquerda, mais pessoas viveram menos. Também há um outlier em torno dos 42. Em 2007, há o mesmo padrão de distribuição, mas o limite superior foi 71, o superior 82 e a mediana está em torno de 78. Observar-se um aumento considerável na expectativa de vida.

África: Em 1952, no continente africano o limite inferior de expectativa de vida era em torno de 30 anos, o superior 50, e acima disso era um outlier. A mediana estava em torno de 40. Já em 2007, o limite inferior foi 40, o superior 78 e a mediana 53. A distribuição de ambos aparenta ser à direita, pois o Q3 é maior, principalmente em 2007, indicando que neste continente, mais pessoas viveram com mais idade.

Américas: Em 1952, o limite inferior de expectativa foi 38 e o superior 68, sendo a mediana 58. É uma distribuição à esquerda, ou seja, mais pessoas vivem menos. Em 2007, o inferior foi 68, com um outlier inferior em volta dos 60 e o limite superior foi 80. Neste caso, não consigo afirmar a distribuição, precisaria ver o valor numérico.

Na Océania, os boxplots parecem uma linha, isso indica que as observações estão concentradas numa única idade/idade muito próxima ou não há observações o suficiente para o boxplot representar. Sendo assim, não é possível muita análise, mas como os boxplots não se sobrepõem, é possível afirmar que a expectativa de vida aumentou de 1952 a 2007.

8. Um plus: Na análise de dados, é comum encontrar dados faltantes. O que é isso? Por que eles existem? Como eles afetam nossa análise? O que fazemos com eles?

Dados faltantes, também conhecidos como valores ausentes ou dados ausentes, referem-se a situações em que um ou mais valores em um conjunto de dados estão ausentes ou não foram coletados. Isso pode ocorrer por diversas razões, como erros de medição, falhas de coleta, recusas dos participantes em responder ou até mesmo dados que simplesmente não foram registrados.

A presença de dados faltantes pode ter implicações significativas na análise de dados e na interpretação dos resultados. Alguns dos efeitos dos dados faltantes incluem:

1. **Viés e Representatividade:** Dados faltantes podem introduzir viés nos resultados, pois os valores ausentes podem não ser representativos da população original. Isso pode levar a conclusões errôneas ou enganosas.
2. **Redução da Precisão:** A presença de dados faltantes pode reduzir a precisão das estimativas e das análises, afetando a confiabilidade dos resultados.
3. **Limitações na Análise:** Alguns métodos estatísticos e algoritmos de análise podem não funcionar adequadamente quando há valores ausentes, o que pode limitar as opções de análise disponíveis.
4. **Amostra Reduzida:** A presença de muitos dados faltantes pode reduzir o tamanho efetivo da amostra, o que pode impactar a validade estatística dos resultados.

Para lidar com dados faltantes, existem várias abordagens que podem ser consideradas:

1. Entendimento do nulo: É importante entender por que os dados estão faltando. Se houver um padrão ou um mecanismo subjacente para os dados ausentes, isso pode influenciar a escolha da abordagem de tratamento.
2. **Exclusão de Dados:** Esta é uma abordagem simples, porém arriscada. Envolve a exclusão de observações com dados ausentes do conjunto de dados. Isso pode ser apropriado se a quantidade de dados ausentes for pequena e não sistemática, mas pode levar à perda de informações valiosas.
3. **Preenchimento de Dados:** Nessa abordagem, os valores faltantes são estimados ou preenchidos com valores substitutos. Isso pode ser feito usando técnicas como a média, mediana, regressão, imputação múltipla ou algoritmos mais avançados, dependendo do contexto e da natureza dos dados.
4. **Análise por Subconjunto:** Em alguns casos, é possível realizar análises apenas com o subconjunto de dados completo, evitando a necessidade de lidar diretamente com os dados faltantes. Isso pode ser viável se os dados ausentes forem limitados a algumas variáveis.
5. **Criação de coluna booleana**: criar uma nova coluna ao lado indicando se aquele valor está ou não nulo (0 para não nulo e 1 para nulo). Assim, levamos a informação para o modelo de machine learning