Adrião Simões Ferreira da Cunha
Estaticista Oficial Aposentado, Antigo Vice-Presidente do Instituto Nacional de Estatística de Portugal
Lisboa, 8 de Fevereiro de 2023

ALGUMAS NOTAS SOBRE A ESTATÍSTICA

As estatísticas são um instrumento poderoso de conhecimento da Sociedade, essenciais à tomada de decisão, à definição e avaliação de estratégias e até ao próprio debate político. Por isso considero que é grande a responsabilidade associada aos desenvolvimentos metodológicos e à produção e utilização de dados, quer pelos organismos oficiais, quer por outras entidades de análise e investigação.

Esta responsabilidade é cada vez maior dado que vivemos num mundo dependente de informação. A natureza global da Sociedade, nomeadamente na sua dimensão económica, assenta em larga medida na rapidez de acesso ao conhecimento e na credibilidade e relevância da informação que o suporta.

Um dos principais objetivos da comunidade estatística deve ser permitir informação de qualidade, cobrindo as mais diversas áreas, quer temáticas quer geográficas. Neste domínio os progressos têm sido notáveis. A qualidade do trabalho efetuado é hoje amplamente reconhecida e o grau de disseminação, até em consequência do progresso tecnológico, tem aumentado de forma significativa.

Há contudo ainda muito por fazer. Recordo por exemplo no âmbito dos Objetivos de Desenvolvimento do Milénio as fragilidades existentes a nível da produção estatística sobre a qualidade de vida nos países menos desenvolvidos e a dificuldade que isso representa para o desenho e para a avaliação de políticas de desenvolvimento e de combate à pobreza.

Outro grande objetivo deve ser o de privilegiar o rigor. É importante que a produção de informação não ceda à conveniência ou às tendências do momento. É fundamental que o trabalho estatístico seja produzido e transmitido com clareza e independência. Só assim podemos falar verdadeiramente de contributos para o conhecimento.

Creio também que a comunidade académica desta área do saber tem o dever de colaborar no processo de educação sobre a importância e o papel da análise estatística como forma de melhorar a nossa compreensão do mundo e de questionar pontos de vista pré-concebidos.

Independente da sua qualidade as estatísticas são por vezes usadas com fins demagógicos, o que pode ser conseguido através da manipulação dos resultados e da deturpação da informação recolhida. A comunidade científica deve contribuir para contrariar o mau uso, intencional ou não, das estatísticas.

A Estatística é um instrumento fundamental do avanço científico, quer nas ciências naturais quer nas ciências humanas. O teste de teorias e o avanço da ciência acompanham em muito o grau de sofisticação e de conhecimento que a Teoria das Probabilidades, e a Estatística em geral, vão proporcionando.

Muitos laureados com o prémio Nobel da Economia receberam este prémio pelas suas contribuições para a econometria, e pelos avanços que, por essa via, foram gerados em outras áreas científicas.

O facto de muitos destes laureados relativamente recentes demonstra por seu turno a crescente importância de metodologias estatísticas rigorosas para o avanço da ciência económica. Acredito que o mesmo se aplica de resto a praticamente todas as áreas do saber.

O desenvolvimento de bases teóricas e metodológicas robustas é fundamental para progredir na qualidade e no rigor da informação produzida. Creio que o debate científico não deve esquecer o papel cada vez mais relevante que a Estatística desempenha no contexto social, político e económico.

Da formulação e da correta utilização da Estatística dependem decisões que afetam a vida – presente e futura - das pessoas, a credibilidade das instituições e a capacidade de se exercer uma vigilância adequada sobre o desempenho daqueles que decidem e governam. Estou certo de que a comunidade estatística está ciente do impacto do seu trabalho e do muito que pode contribuir para estes objetivos.

Portugal progrediu muito nos últimos anos no que respeita à produção e à qualidade das estatísticas. O trabalho que os investigadores têm realizado e o esforço técnico que tem sido feito, em instituições como o Instituto Nacional de Estatística, no sentido de assegurar uma informação mais clara, fiável e independente tem sido notável. Este é um trabalho que deve prosseguir e ser aprofundado pelos países.

CONCEITOS ESTATÍSTICOS FUNDAMENTAIS

A Estatística é um ramo da matemática que lida com a recolha, análise, interpretação e apresentação de massas de dados numéricos.

A Estatística é usada em quase todos os aspetos da Ciência de Dados. É usada para analisar, transformar e limpar dados, avaliar e otimizar algoritmos de aprendizagem de máquina e também é usada na apresentação de perceções e descobertas.

O campo da Estatística é extremamente amplo e pode ser difícil determinar exatamente o que precisa aprender e em que ordem. Mas o facto é que nem tudo é necessário em Ciência de Dados e não é necessário graduação em Estatística para trabalhar como Cientista de Dados.

Neste artigo apresento **8 Conceitos Estatísticos Fundamentais Para *Data Science*** que precisa entender ao estudar ou trabalhar com Ciência de Dados. Estas não são técnicas particularmente avançadas, mas são uma seleção dos requisitos básicos que precisa saber antes de passar para a aprendizagem de métodos mais complexos.

1- Amostragem

Em Estatística todos os dados brutos que pode ter disponíveis para um teste é conhecido como população. Por uma série de razões não é viável medir os padrões e tendências em toda a população. As estatísticas permitem-nos tomar uma amostra, realizar alguns cálculos sobre o conjunto de dados e usando a probabilidade e algumas suposições podemos com um certo grau de certeza compreender as tendências para toda a população ou prever eventos futuros.

Digamos por exemplo que queremos entender a prevalência de uma doença como o cancro de mama em toda a população de um país. Por razões práticas não é possível rastrear toda a população. Em vez disso podemos pegar numa amostra aleatória e medir a prevalência entre esses dados. Supondo que a nossa amostra seja suficientemente aleatória e representativa de toda a população, podemos obter uma medida de prevalência e fazer inferências sobre toda a população.

2- Estatística Descritiva

A Estatística Descritiva, como o nome sugere, ajuda-nos a descrever os dados. Em outras palavras, permite-nos compreender as características. Aqui o objetivo não é prever algo, fazer suposições ou inferência, mas simplesmente fornecer uma descrição da aparência da amostra de dados que temos.

As estatísticas descritivas são normalmente calculadas a partir dos dados. Isso inclui as medidas de tendência central, como:

Média – o valor médio dos dados.
Mediana – o valor central se ordenarmos os dados em ordem crescente e dividirmos exatamente pela metade.
Moda – o valor que ocorre com mais frequência.

3- Distribuições

As estatísticas descritivas são úteis, mas muitas vezes podem ocultar informações importantes sobre o conjunto de dados. Por exemplo se um conjunto de dados tem números que são muito maiores que os outros a média pode ser distorcida e não nos dará uma representação verdadeira dos dados.

Uma distribuição pode ser representada por um gráfico, geralmente um histograma, que exibe a frequência com que cada valor aparece num conjunto de dados. Este tipo de gráfico fornece-nos informações sobre a dispersão e a assimetria dos dados.

Uma distribuição geralmente formará um gráfico semelhante a uma curva, que pode ser inclinada mais para a esquerda ou direita.

Uma das distribuições mais importantes é a *distribuição normal*, comumente chamada de *curva em sino* devido ao seu formato. É de forma simétrica com a maioria dos valores agrupados em torno do pico central e os valores mais distantes distribuídos igualmente em cada lado da curva. Muitas variáveis na natureza formarão uma distribuição normal, como a altura das pessoas e as pontuações de QI (Coeficiente de Inteligência). A distribuição normal de uma variável é a suposição de vários algoritmos de *Machine Learning*.

4- Probabilidade

Probabilidade em termos simples é a probabilidade de um evento ocorrer. Em Estatística um evento é o resultado de uma experiência que pode ser algo como o lançamento de um dado ou os resultados de um teste AB (consiste em dividir o tráfego de uma determinada página em 2 versões: a atual e uma desafiante, e depois mede-se qual das versões apresenta maior taxa de conversão).

A probabilidade de um único evento é calculada dividindo o número de eventos pelo número total de resultados possíveis. Considere, por exemplo, conseguir um 6 ao lançar um dado. Como existem 6 resultados possíveis, a possibilidade de rolar um 6 é 1/6 = 0,167, e às vezes isso também é expresso como uma porcentagem, então 16,7%.

Os eventos podem ser *independentes* ou *dependentes*.

Com eventos *dependentes*, um evento anterior influencia o evento subsequente. Digamos que temos um pacote de leite e queremos determinar a probabilidade de escolher aleatoriamente um pacote vermelho. Todas as vezes que removêssemos um leite do pacote, a probabilidade de escolher o vermelho mudaria devido ao efeito de eventos anteriores.

Os eventos *independentes* não são afetados por eventos anteriores. No caso do pacote de leite se cada vez que selecionamos um o colocamos de volta no pacote, a probabilidade de selecionar vermelho permaneceria a mesma todas as vezes.

Se um evento é independente ou não é importante, pois a maneira como calculamos a probabilidade de vários eventos muda dependendo do tipo.

A probabilidade de vários eventos independentes é calculada simplesmente multiplicando a probabilidade de cada evento. No exemplo do lançamento de dados, digamos que quiséssemos calcular a probabilidade de lançar um 6 três vezes. Isso seria parecido com o seguinte:

1/6 = 0,167
1/6 = 0,167
1/6 = 0,167
0,167 * 0,167 * 0,167 = 0,005

O cálculo é diferente para eventos dependentes, também conhecido como *probabilidade condicional*. Se tomarmos o exemplo do M&M, imagine que temos um pacote com apenas duas cores vermelho e amarelo, e sabemos que o pacote contém 3 vermelhos e 2 amarelos e queremos calcular a probabilidade de escolher dois vermelhos em uma fileira.

Na primeira escolha, a probabilidade de escolher um vermelho é 3/5 = 0,6. Na segunda escolha, removemos um M&M, que por acaso era vermelho, então nosso segundo cálculo de probabilidade é 2/4 = 0,5. A probabilidade de escolher dois vermelhos em uma fileira é, portanto, 0,6 * 0,5 = 0,3.

5- Viés

Como discutimos anteriormente, usamos amostras de dados para fazer estimativas sobre todo o conjunto de dados. Da mesma forma para modelagem preditiva usaremos alguns dados de formação e tentaremos construir um modelo que possa fazer previsões sobre novos dados.

Viés é a tendência de um modelo estatístico ou preditivo de super ou subestimar um parâmetro. Isso geralmente se deve ao método usado para obter uma amostra ou à forma como os erros são medidos.

Há vários tipos de vieses comumente encontrados nas estatísticas. Eis uma breve descrição de dois.

Viés de seleção – ocorre quando a amostra é selecionada de forma não aleatória. Em Data Science, um exemplo pode ser interromper um teste AB mais cedo quando o teste está em execução ou selecionar dados para treinar um modelo de aprendizagem de máquina de um período de tempo que pode mascarar os efeitos sazonais.

Viés de confirmação – ocorre quando a pessoa que realiza alguma análise tem uma suposição predeterminada sobre os dados. Nessa situação, pode haver uma tendência de gastar mais tempo examinando variáveis que provavelmente apoiarão essa suposição.

6- Variância

Como discutimos anteriormente neste artigo a média é uma medida de tendência central. A variância mede a distância de cada valor no conjunto de dados da média. Essencialmente é uma medida da dispersão dos números em um conjunto de dados.

O desvio padrão é uma medida comum de variação para dados que têm uma distribuição normal. É um cálculo que fornece um valor para representar a extensão da distribuição dos valores. Um desvio padrão baixo indica que os valores tendem a ficar muito próximos da média, enquanto um desvio padrão alto indica que os valores estão mais dispersos.

Se os dados não seguem uma distribuição normal, outras medidas de variância são usadas. Normalmente, o *intervalo interquartil* é usado. Essa medida é derivada primeiro ordenando os valores por classificação e, em seguida, dividindo os pontos de dados em quatro partes iguais, chamadas *quartis*. Cada quartil descreve onde 25% dos pontos de dados se encontram de acordo com a mediana. O intervalo interquartil é calculado subtraindo a mediana dos dois quartos centrais, também conhecidos como Q1 e Q3.

7- Conflito de Escolha Viés/Variância

Os conceitos de viés e variância são muito importantes em *Machine Learning*. Quando construímos um modelo de aprendizagem de máquina, usamos uma amostra de dados conhecida como conjunto de dados de formação. O modelo aprende padrões nesses dados e gera uma função matemática que é capaz de mapear o rótulo de destino correto ou valor (y) para um conjunto de entradas (X).

Ao gerar esta função de mapeamento o modelo usará um conjunto de suposições para melhor aproximar o alvo. Por exemplo, o algoritmo de regressão linear assume uma relação linear (linha reta) entre a entrada e o destino. Essas suposições geram viés no modelo.

Como cálculo, o viés é a diferença entre a previsão média gerada pelo modelo e o valor verdadeiro.

Se tivéssemos de treinar um modelo usando diferentes amostras de dados de formação obteríamos uma variância nas previsões que são retornadas. A variância na aprendizagem de máquina é uma medida de quão grande é essa diferença.

Em aprendizagem de máquina o viés e a variância são o erro geral esperado para as previsões. Num mundo ideal teríamos baixo viés e baixa variância. No entanto na prática minimizar o viés geralmente resultará num aumento na variância e vice-versa. A compensação de viés/variância descreve o processo de equilibrar esses dois erros para minimizar o erro geral de um modelo.

8- Correlação

Correlação é uma técnica estatística que mede as relações entre duas variáveis. A correlação é considerada linear (formando uma linha quando exibida em um gráfico) e é expressa como um número entre +1 e -1, conhecido como *coeficiente de correlação*.

Um coeficiente de correlação de +1 indica uma correlação perfeitamente positiva (quando o valor de uma variável aumenta o valor da segunda variável também aumenta), um coeficiente de 0 indica nenhuma correlação e um coeficiente de -1 indica uma correlação negativa perfeita.

Importa ainda ressaltar que correlação não implica causalidade. O facto de haver correlação entre duas variáveis não significa que uma causa a ocorrência da outra. Para afirmar isso teríamos que realizar estudos adicionais e uma análise de causalidade.

A Estatística é um campo amplo e complexo e este artigo pretende ser uma breve introdução a algumas das técnicas estatísticas mais comumente usadas em *Data Science*.

A ESTATÍSTICA É A "MÃO" QUE GOVERNA O MUNDO

Assistimos provavelmente ao período da História em que há mais dados estatísticos disponíveis e maior desinformação à volta deles.

Opiniões e ciência confundem-se. Infelizmente uma das formas mais frequentemente usadas para tentar impor uma ideia sem fundamento é pelo meio de argumentos quantitativos porque a maior parte das pessoas tem alguma dificuldade em desconstruir números.

Estou certo de que passa despercebido à maior parte das pessoas como de facto a Estatística está na base de muitas das decisões que nos afetam a vida no dia-a-dia. Desde o GPS que prevê a hora a que vamos chegar a uma reunião até aos anúncios que nos são apresentados na Internet, tudo é controlado de uma forma ou de outra por análises estatísticas mais ou menos sofisticadas.

Quem nunca ouviu falar em termos como *big data*, *artificial intelligence*, *machine learning*, *business intelligence*? Estes termos tão pomposos e potencialmente intimidantes que quase tornam mais inteligentes aqueles que os usa m ou neles trabalham, não passam de diferentes roupagens para modelos estatísticos que utilizam dados para fazer previsões que são depois usadas para tomar decisões em face de incerteza. E é assim que por exemplo a nossa operadora de telecomunicações nos sugere outra(s) série(s) com base naquela que acabámos de ver e até parece interessante!

Com base nas avaliações que vamos fazendo eles descobrem o que gostamos. Aquelas 5 estrelinhas que vamos colocando nas avaliações dos programas a que assistimos vão permitir à operadora, com base em vários milhares de estrelinhas, dos muitos milhares de clientes, perceber os gostos de grupos de pessoas com perfis de visualização semelhantes que no limite poderão estar disponíveis para pagar um canal *premium* que terá mais séries direcionadas para as preferências. Aí sim temos operadoras contentes, que usaram *big data*, que ajustaram um algoritmo de *machine learning*, e estão a fazer *business intelelligence*, ou, diria eu, Estatística.

A origem da palavra Estatística vem da ciência que estuda o Estado, mais concretamente os números que o Estado precisaria de compreender para ser mais eficiente. É essa vertente que dá origem à área das Estatísticas Oficiais e que muitas vezes nos examinam a vida frequentemente sem sequer sabermos bem como.

Desemprego, inflação ou défice, que podem afundar ou salvar uma economia não são mais do que previsões de modelos estatísticos. Quando se diz que o desemprego num país está em 12% não se perguntou a cada um dos cidadãos se estavam ou não desempregados. Foi recolhida uma amostra e com base nessa amostra foram feitas inferências sobre a população. O processo de amostragem é por isso fundamental, e se facilitarmos este processo as consequências são desconhecidas.

Se quero saber quantas pessoas já foram expostas ao COVID-19 não posso usar uma amostra de conveniência. Usando como amostra um grupo de pessoas que foi fazer análises de livre e espontânea vontade, como no estudo recente do Instituto Nacional de Saúde Dr. Ricardo Jorge, de Portugal, e voluntários, como num estudo recente do Instituto de Medicina Molecular também de Portugal, não se consegue à partida interpretar com robustez os resultados obtidos porque alguém que foi fazer análises o fez certamente por uma razão, e quem se dá ao trabalho de se oferecer como voluntário para um estudo terá também a sua justificação. Em nenhum dos casos os grupos serão uma amostra representativa da população nacional: é a Estatística que permite compreender isso.

A pergunta pode parecer pouco urgente. Afinal todos conseguimos imaginar as consequências terríveis de voltar a viver sem democracia, sem sistemas de proteção social ou sem eletricidade. Mas estatísticas? A sua importância pode não ser assimilada de forma tão imediata. Mas na realidade perdê-las seria como voltar a uma espécie de analfabetismo massificado. Viveríamos completamente perdidos. Em que mundo viveríamos sem produção de estatísticas? Sem esta informação sobre as nossas vidas estaríamos "à deriva".

Viveríamos completamente perdidos. Precisamos cada vez mais de estatísticas porque nos dão a informação que precisamos sobre o que somos, como somos, o que fazemos, os reflexos dos nossos comportamentos. Sem estatísticas viveríamos à deriva com base nas nossas opiniões. Não seria bom para uma Sociedade que se quer cada vez mais desenvolvida e que queremos que avance. E para saber se estamos a avançar as estatísticas também são essenciais.

É uma visão dramática e para a entendermos melhor é importante perceber como as estatísticas afetam a nossa vida. Pode começar no topo da pirâmide, nas decisões dos Governos, mas desce todos os degraus até ao nosso dia-a-dia. Ou seja, ajudam-nos desde saber quantos desempregados existem ou se o investimento público caiu até conhecer as cidades com as casas mais baratas ou quanto é que estamos a ajudar o ambiente ao substituir banhos de imersão por duches.

Por definição as estatísticas são instrumentos de conhecimento e sem essa informação não se conhece a realidade nem se sabe o que se passa ao nosso lado. Quem toma decisões tem de saber sobre que realidade as vai tomar. Se não tiver dados fiáveis vão sair más decisões. E sobre quem vão recair essas más decisões? Sobre os cidadãos. Pensa-se sempre nos Governos, mas também vai ao nível individual. Posso tomar uma decisão e refletir-se no bem-estar do meu vizinho.

Sem estatísticas estaríamos dependentes daquilo que uma pessoa poderosa diz ser a verdade o que retiraria poder à Sociedade. No passado havia muita gente totalmente analfabeta. Essas pessoas estavam à mercê dos outros, daquilo que diziam que estava escrito nas paredes dos autocarros. Num mundo sem estatísticas teríamos regimes não democráticos em que certas pessoas se sentem portadoras da verdade. A democracia precisa muito de informação credível.

No entanto mesmo com estatísticas, e há cada vez mais a serem produzidas, os últimos anos fizeram muitos questionar se não entrámos no que muitas vezes se designa como Sociedade "pós-facto", em que os dados são distorcidos e apresentados de forma a confirmar a nossa narrativa. Uma situação agravada pela nova paisagem mediática que permite o consumo de informação que não contradiz a nossa visão do mundo. O que podem os estaticistas fazer para contrariar esta trajetória?

A palavra é **educação**. Reforço da literacia estatística. É importante que as escolas se compenetrem que as estatísticas devem entrar no seu mundo e que os Ministérios da Educação cumpram o seu papel. Despertar [nas crianças] o interesse por conhecer os números, saberem "desmontá-los" e depois formular opiniões. Precisamos de comunicar. Em vez de falar em formação bruta de capital fixo, falar em investimento. Vamos estar a comunicar com mais pessoas.

O passo seguinte é saber se os INE responsáveis por produzir e divulgar estatísticas oficiais têm os meios adequados para responder a estes novos desafios. Há cada vez mais informação a circular muita vinda de fontes não-oficiais.

Os INE estão divididos entre aproveitar melhor esse novo potencial e lidar com as suas limitações. Ao mesmo tempo exige-se que produzam mais dados e mais indicadores e a uma velocidade maior. É uma guerra cada vez mais dura por vezes sem que sejam dadas mais armas para a combater.

Será que os INE têm os recursos suficientes para responder a estes desafios. O problema algumas vezes não é financeiro mas de "insuficiência de recursos humanos", pelo que os INE devem fazer um esforço de auto-formação.

MINI DICIONÁRIO ESTATÍSTICO

A

Amostra: é um subconjunto finito da população que se supõe representativo desta.

Amostra Amodal: amostra que não tem moda.

Amostra Bimodal: amostra que tem duas modas.

Amostra Imparcial: amostra em que todos os elementos tiveram igual oportunidade de fazer parte da mesma.

Amostra Multimodal: amostra que tem mais do que duas modas.

Amostra Representativa: é a que deve conter em proporção todas as características qualitativas e quantitativas da população.

Amostragem Aleatória Simples: é a em que qualquer elemento da população tem a mesma probabilidade de ser escolhido.

Amostragem Estratificada: é aquela em que a população está dividida em estratos ou grupos diferenciados.

Amostragem Sistemática: é aquela em que os elementos são escolhidos a partir de uma regra previamente estabelecida.

Amplitude de um Conjunto de Dados: é a diferença entre o maior valor e o menor valor desse conjunto. Se os dados estiverem agrupados em classes a amplitude é a diferença entre o limite superior da última classe e o limite inferior da primeira.

[Fórmula]

Para um intervalo do conjunto de dados de [a,b], onde x_1= a e x_n= b

$$x_n - x_1 = b - a$$

Atributos Qualitativos: são atributos que estão relacionados com uma qualidade e apresentam-se com várias modalidades.

Atributos Quantitativos: são atributos aos quais é possível atribuir uma medida e apresentam-se com diferentes intensidades ou valores.

C

Censo: é uma operação estatística que resulta da observação de todos os indivíduos da população relativamente a diferentes atributos pré-definidos.

Classe Mediana ($\widetilde{x}$): é a classe para dados classificados que contem a Mediana (neste caso considera-se como Mediana o valor da variável estatística que corresponde a $n/2$, quer n seja par, quer n seja ímpar).

Classe Modal: é a classe para dados classificados que aparece com maior frequência.

Coeficiente de Correlação Linear (*r*): é a medida estatística que permite calcular o valor numérico correspondente ao grau de dependência entre duas variáveis, o qual varia entre -1 e 1.

[Fórmula]

$$r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left[\sum_{i=1}^{n} x_i\right]\left[\sum_{i=1}^{n} y_i\right]}{n}}{\sqrt{\left[\sum_{i=1}^{n} x_i^2 - \frac{\left[\sum_{i=1}^{n} x_i\right]^2}{n}\right] \times \left[\sum_{i=1}^{n} y_i^2 - \frac{\left[\sum_{i=1}^{n} y_i\right]^2}{n}\right]}}$$

Correlação: é a relação ou dependência entre as duas variáveis de uma distribuição bidimensional.

Correlação Fraca ou Nula: quando o Diagrama de Dispersão não permite o ajustamento de nenhuma reta, o que significa que $r=0$. Diz-se então que não existe nenhuma relação entre as variáveis da Distribuição Bidimensional.

Correlação Negativa Forte: quando a reta de regressão obtida a partir do Diagrama de Dispersão tem declive negativo. A correlação é negativa quando r varia entre -1 e 0 e será tanto mais forte quanto r se aproxima de -1.

Correlação Negativa Perfeita ou Linear: quando a reta de regressão obtida a partir do Diagrama de Dispersão tem declive negativo com $r=-1$.

Correlação Positiva Forte: quando a reta de regressão obtida a partir do Diagrama de Dispersão tem declive positivo. A correlação é positiva quando r varia entre 0 e 1 e será tanto mais forte quanto r se aproxima de 1.

Correlação Positiva Perfeita ou Linear: quando a reta de regressão obtida a partir do Diagrama de Dispersão tem declive positivo com $r=1$.

D

Dados Classificados: são valores que uma dada variável pode tomar dentro de certo intervalo. Estes dados são classificados ou agrupados em classes.

Dado Estatístico: é o resultado da observação de um atributo/variável qualitativa ou quantitativa.

Dados Simples: são valores associados a uma variável e cuja representação é feita através duma tabela.

Definição do Problema: é a primeira fase do estudo estatístico e consiste na definição e formulação correta do problema a ser estudado.

Desvio Médio (*d*): é a média aritmética do valor absoluto da diferença entre cada valor e a média no caso dos dados não classificados. No caso dos dados classificados tem que se entrar em conta com a frequência absoluta de cada observação.

[Fórmula]

Desvio Médio (com *n*= n.º de **Dados não classificados**

$$d = \frac{\sum_{i=1}^{n} \left| x_i - \bar{x} \right|}{n}$$

Desvio Médio (com *n*= n.º de **Dados classificados**

$$d = \frac{\sum_{i=1}^{n} \left(f_i \times \left| x_i - \bar{x} \right| \right)}{n}$$

Desvio Padrão (σ): é a raiz quadrada positiva da variância.

[Fórmula]

$$\sigma = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

Diagrama de Dispersão: é a representação num referencial ortonormado de um conjunto de pares ordenados de valores (x, y), onde cada par ordenado corresponde a uma observação.

Distribuição Bidimensional: é a representação de uma variável bidimensional (x_i, x_j), com $1 \leq i \leq n$ e x_i e x_j duas variáveis unidimensionais.

Diagrama de Caule-e-Folhas: o mesmo que Separador de Frequências.

Diagrama de Extremos e Quartis: é um diagrama que representa os valores extremos e os quartis de uma variável estatística.

Distribuição de Frequências: o mesmo que Tabela de Frequências.

E

Estatística: é o método que ensina a recolher, classificar, apresentar e interpretar um conjunto de dados numéricos.

Estatística Descritiva: é o ramo da Estatística que tem por finalidade descrever certas propriedades relativas a um conjunto de dados.

Estatística Indutiva: é o ramo da Estatística que procura inferir propriedades da população a partir de propriedades verificadas numa amostra da mesma.

F

Fenómenos Independentes: são fenómenos respeitantes à mesma variável que não têm qualquer ligação um com o outro.

Frequência Absoluta (f_i): é o número de vezes que o valor de determinada variável é observado.

Frequência Absoluta Acumulada (F_i): é a soma das frequências absolutas anteriores com a frequência absoluta deste valor.

[Fórmula]

$$F_i = \sum_{i=1}^{n} f_i$$

Frequência Relativa (f_{ri}): é o quociente entre a frequência absoluta do valor da variável e o número total de observações.

[Fórmula]

$$f_{r_i} = \frac{f_i}{n}$$

Frequência Relativa Acumulada (F_{ri}): é a soma das frequências relativas anteriores com a frequência relativa desse valor.

[Fórmula]

$$F_{r_i} = \sum_{i=1}^{n} f_{r_i}$$

Função Cumulativa: é a função que indica para cada valor real x a frequência absoluta (ou relativa) de observações com intensidade menor ou igual a x. A representação gráfica desta função é em forma de escada.

G

Gráfico Circular: é representado por um círculo que está dividido em setores cujas amplitudes são proporcionais à frequência que lhe corresponde.

Gráfico de Barras: é constituído por barras horizontais ou verticais de comprimento proporcional à frequência.

H

Histograma: é um gráfico de barras em que a área destas é proporcional à frequência, não havendo espaço entre as mesmas. Só se utiliza em variáveis quantitativas contínuas.

M

Média Aritmética Simples ($\bar{x}$): é o quociente da soma de todos os dados não classificados pelo número desses dados.

[Fórmula]

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Média Aritmética Ponderada ($\bar{x}$): é o quociente entre o somatório do produto de cada dado classificado pela sua frequência absoluta e o número desses dados.

Fórmula]

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{n} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_n x_n}{n}$$

Mediana ($\tilde{x}$): é o valor da variável, para dados não classificados, que ocupa a posição central da distribuição.

[Fórmula]

n par

$$\tilde{x} = \frac{x_{\frac{n}{2}} + x_{\left(\frac{n}{2}+1\right)}}{2}$$

n impar

$$\tilde{x} = x_{\left(\frac{n+1}{2}\right)}$$

Medidas de Dispersão: é um conjunto de medidas (Amplitude, Variância e Desvio Padrão) utilizadas no estudo da variabilidade duma determinada distribuição, permitindo obter uma informação mais completa acerca da "forma" da mesma.

Medidas de Localização: é um conjunto de medidas (Média, Mediana, Moda e Quartis) que representam de uma forma global um conjunto de dados.

Medidas de Tendência Central: o mesmo que Medidas de Localização.

Moda (m): observação que ocorre com maior frequência numa amostra.

N

Núvem de Pontos: o mesmo que Diagrama de Dispersão.

O

Organização dos Dados: consiste em "resumir" os dados através da sua contagem e agrupamento.

P

Pictogramas: são gráficos onde se utilizam figuras ou símbolos alusivos ao problema em estudo.

Planificação do Problema: consiste na determinação de um processo para resolver o problema e em especial como obter informações sobra a variável em estudo.

Polígono de Frequências: são gráficos com aspeto de linhas quebradas. Constroem-se unindo por segmentos de reta os pontos médios das bases superiores dos retângulos de um histograma.

População: é um conjunto de seres com uma dada característica em comum e com interesse para o estudo.

Q

Quartis (Q_1 e Q_3): são os valores que dividem a distribuição em 4 partes iguais.

[Fórmula]

Quando o índice *i* dos x_i **é um número inteiro**:

n par:

$$Q_1 = x_{\frac{n+2}{4}} \quad Q_3 = x_{\frac{3n+2}{4}}$$

n ímpar:

$$Q_1 = x_{\frac{n+1}{4}} \quad Q_3 = x_{3\times\left(\frac{n+1}{4}\right)}$$

Quando o índice *i* dos x_i **não é um número inteiro**, calcula-se como nos exemplos seguintes:

$$\frac{n+2}{4} = 3{,}5 \Rightarrow Q_1 = \frac{x_3 + x_4}{2}$$

$$3\times\left(\frac{n+1}{4}\right) = 32{,}5 \Rightarrow Q_3 = \frac{x_{32} + x_{33}}{2}$$

R

Recenseamento: o mesmo que Censo.

Recolha de Dados: é a primeira etapa depois de definido o problema em estudo.

Reta de Regressão: é a reta traçada sobre uma dada Nuvem de Pontos, sendo um modelo matemático que pretende descrever a relação existente entre duas variáveis unidimensionais de uma distribuição bidimensional.

[Fórmula]

$$y = ax + b$$

onde

$$b = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

$$a = y - bx$$

Relações Estatísticas: são relações que se podem estabelecer entre determinadas variáveis de um problema em estudo.

S

Separador de Frequências: é um tipo de tabela que permite ter uma perceção imediata do aspeto global dos dados sem perda da informação contida na coleção dos dados inicial.

Somatório (∑): representa de forma abreviada uma soma.

[Fórmula]

$$\sum_{i=1}^{n} g_i(x) = g_1(x) + g_2(x) + \ldots + g_{n-1}(x) + g_n(x)$$

Onde (x) representa uma expressão, cuja variável é *x,* que varia consoante o índice *i* varia de 1 até *n*.

Sondagem: é o estudo estatístico que se baseia numa parte da população, isto é, numa amostra que deve ser representativa dessa população.

T

Tabela de Frequências: são tabelas onde se apresentam os dados por classes e as frequências respetivas.

Tamanho da Amostra: é o número de elementos que constituem uma dada amostra.

U

Unidade Estatística ou Indivíduo: é cada um dos elementos da população.

V

Variância (σ^2): é a medida que permite avaliar o grau de dispersão dos valores da variável em relação à média.

[Fórmula]

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

Variáveis Contínuas: são as variáveis que podem tomar qualquer valor de um determinado intervalo.

Variáveis Discretas: são as variáveis que podem tomar um número finito ou uma infinidade numerável de valores.

Variáveis Qualitativas: o mesmo que Atributos Qualitativos.

Variáveis Quantitativas: o mesmo que Atributos Quantitativos.