A IMPORTÂNCIA DA VARIABILIDADE

*Fernando Magalhães**

Quando fui convidado para apresentar uma comunicação, na cerimónia comemorativa do 129º aniversário do Instituto Superior de Contabilidade e Administração do Porto, de imediato pensei que a tarefa para a qual estava a ser convidado era um complicado desafio. A primeira etapa, talvez a mais delicada, consistia em decidir o tema que deveria abordar, uma vez que teria uma plateia cientificamente heterogénea e, verdade seja dita, a Estatística não é um tema que seduza facilmente os leigos.

A primeira tentação seria falar sobre sondagens (estávamos a acabar de sair de um processo eleitoral). Mas, ou escolheria falar de um conjunto de banalidades percetíveis por todos os que formavam a minha plateia, ou teria que optar por uma comunicação técnica e científica, o que, certamente, iria entediar aqueles a quem a Estatística e a Matemática pouco dizem.

Então, tomando em consideração de que a razão de ser da minha profissão (professor) é, acima de tudo, o aluno, lembrei-me que poderia exemplificar como é possível fazer uma aula de apresentação de uma unidade curricular de Estatística com o intuito de motivar os estudantes logo desde o início, mostrando que o medo e o receio com que eles, *a priori*, entram na primeira aula de Estatística, afinal, não tem um verdadeiro motivo para existir. Como lhes costumo dizer, tudo o que os seus colegas mais antigos provavelmente opinaram sobre a

---

* Fernando José Malheiro de Magalhães, Professor Coordenador do Instituto Superior de Contabilidade e Administração do Porto (ISCAP/IPP), fjmm@iscap.ipp.pt

dificuldade da Estatística, sobre a elevada taxa de reprovação na unidade curricular, e muito mais, não é verdade. Trata-se de uma piada que eles usam para praxar os caloiros... e estes acreditam!

A Estatística é, até um certo ponto, intuitiva, por ser uma ciência totalmente aplicada ao mundo real, o que facilita a sua compreensão. Contudo, muitos dos seus conceitos são semelhantes e, muitos deles, têm que ser adquiridos lentamente. Não é possível estudar Estatística numa semana, de manhã à noite, e obter um resultado satisfatório. Os conceitos estatísticos, por frequentemente conterem várias subtilezas na sua construção, têm que ser compreendidos devagar. Assim, na minha opinião, a chave para ter sucesso em Estatística assenta em dois fatores essenciais: ser assíduo às aulas, compreendendo os conceitos aí lecionados, e estudar continuamente, isto é, todas as semanas, sendo suficiente estudar apenas um pouco de cada vez.

Vejamos, então, como poderá ser uma aula de apresentação da unidade curricular de Estatística.

Como é óbvio, a primeira pergunta a fazer aos estudantes será: "O que é a Estatística?". A essa questão, usualmente, segue-se uma hesitação e um silêncio. Depois, lá começam a sair respostas tais como: "É aquilo da média", "É aquilo das probabilidades, dos arranjos e das combinações" ou "É para fazer previsões".

Nestas respostas, é interessante realçar alguns aspetos: em primeiro lugar, nota-se que os alunos, muitas vezes, não têm uma noção geral do que é a estatística, mas apenas uma noção dos conceitos definidos matematicamente sem que tenham completa noção do seu significado prático. Além disso, confundem, frequentemente, a Estatística com os diversos instrumentos matemáticos utilizados no seu desenvolvimento.

Aproveitando o facto de ter sido referido que a Estatística faz previsões, poder-se-á acrescentar que a ciência estatística está presente em tudo o que nos rodeia. As nossas decisões mais básicas do dia-a-dia são tomadas de acordo com as tendências e padrões que identificamos e que nos permitem prever o que vai acontecer e, assim, tentar tomar a melhor decisão.

Sendo a Estatística vista como uma ciência que permite "adivinhar" o desconhecido, terá que agradar a todos, visto que o ser humano é, por natureza, curioso. Assim, a possibilidade de prever o futuro ou adivinhar o desconhecido torna a Estatística, de facto, aliciante.

Analisando observações de um determinado fenómeno cujo comportamento é variável, a Estatística tem metodologias que permitem identificar tendências e padrões, possibilitando, assim, prever o que se irá observar no futuro.

Do mesmo modo, a Estatística faculta métodos que permitem concluir sobre o todo com base na análise de apenas uma parte. Ao todo, chama-se população e, à parte, amostra. Ora, como é óbvio, é frequente não ser possível ou viável analisar toda a população. Nesse caso, será estudada uma parte da população, isto é, a amostra e, a partir das observações nela contidas, inferir sobre o que se passa no todo, ou seja, na população.

Tal situação é ilustrada na figura seguinte.

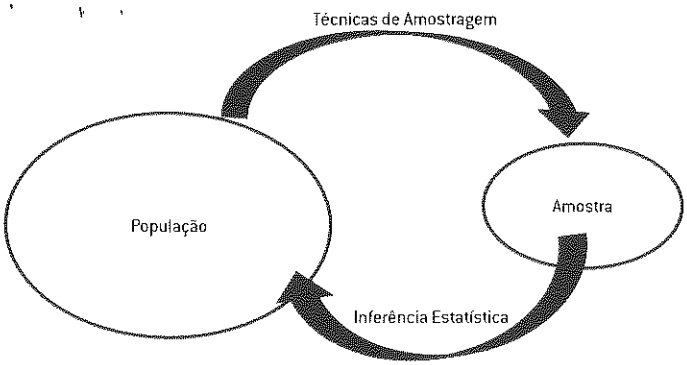


Figura 1

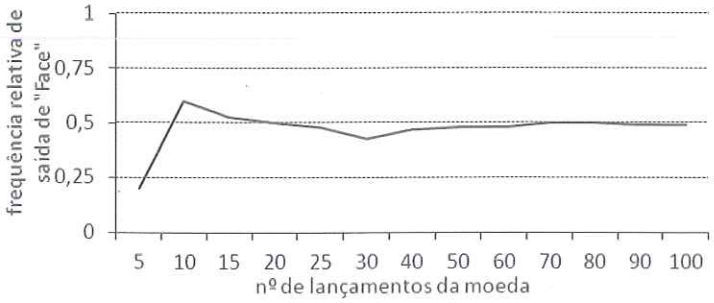
A possibilidade de prever o futuro ou "adivinhar" aquilo que não é observado pode fazer pensar que o estatístico é uma espécie de bruxo ou adivinho. Ora, o que distingue o estatístico do bruxo? O bruxo prevê ou adivinha não dizendo qual é a confiança que tem nas suas adivinhas. Por outro lado, o estatístico consegue quantificar a confiança que tem nas suas previsões ou adivinhas. Tal é conseguido através

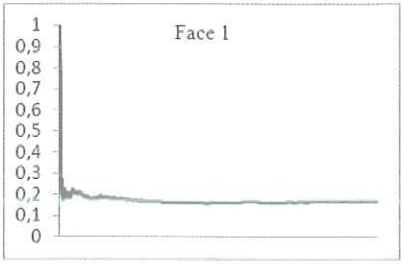
do recurso ao conceito de medida de probabilidade. Fazendo uso da Teoria da Probabilidade, o estatístico estabelece o grau de confiança (ou a probabilidade de erro) das suas previsões.

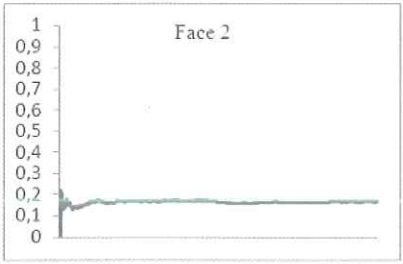
Assim sendo, o conceito de probabilidade é fundamental para a construção de qualquer método estatístico. Ora, o que é a probabilidade?

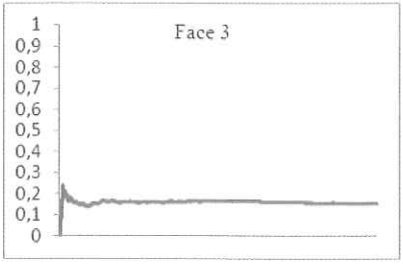
No início do século XIX, Laplace define o conceito de probabilidade para o caso em que, ao realizar uma experiência, os resultados possíveis de obter são em número finito e igualmente possíveis (equiprováveis). Em tal situação, a probabilidade de ocorrência de um acontecimento é obtida pelo quociente entre o número de casos favoráveis a que o acontecimento se realize e o número de resultados possíveis da experiência. Por exemplo, ao lançar uma moeda regular, a probabilidade de ocorrer "Face" é ½=0.5; ao lançar um dado vulgar e não viciado, a probabilidade de que cada uma das faces do dado fique voltada para cima é 1/6. Note-se que estas probabilidades foram obtidas no pressuposto de que os resultados elementares (as faces 1, 2, 3, 4, 5, 6) são equiprováveis. E se tal não acontecer? Nesse caso, a definição clássica de Laplace não pode ser utilizada para calcular as probabilidades.

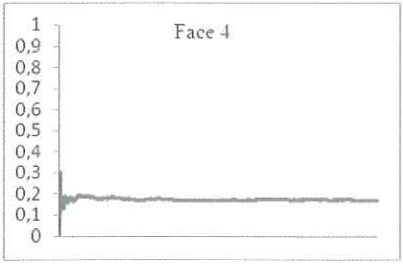
No início do século XX, surge a definição frequencista de probabilidade que contorna o problema de os resultados elementares da experiência não serem equiprováveis. Segundo esta definição, a probabilidade de um acontecimento é o valor para o qual tende a frequência relativa do acontecimento à medida que o número de repetições da experiência aumenta. Nas figuras seguintes, ilustra-se a obtenção do valor da probabilidade de saída de "Face" quando se lança uma moeda e a obtenção dos valores das probabilidades de saída de cada um dos lados de um dado, segundo a definição frequencista.

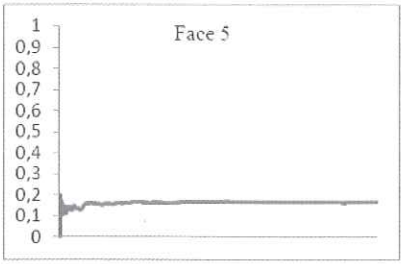

frequência relativa de saída de "Face"
1
0,75
0,5
0,25
0
5 10 15 20 25 30 40 50 60 70 80 90 100
nº de lançamentos da moeda

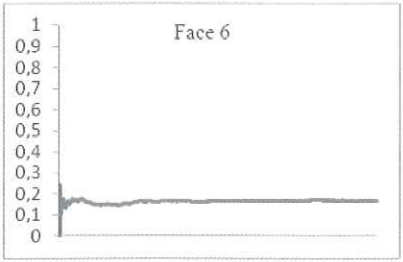

Face 1
1 0,9 0,8 0,7 0,6 0,5 0,4 0,3 0,2 0,1 0


Face 2
1 0,9 0,8 0,7 0,6 0,5 0,4 0,3 0,2 0,1 0


Face 3
1 0,9 0,8 0,7 0,6 0,5 0,4 0,3 0,2 0,1 0


Face 4
1 0,9 0,8 0,7 0,6 0,5 0,4 0,3 0,2 0,1 0


Face 5
1 0,9 0,8 0,7 0,6 0,5 0,4 0,3 0,2 0,1 0


Face 6
1 0,9 0,8 0,7 0,6 0,5 0,4 0,3 0,2 0,1 0

De notar que estas duas definições de probabilidade, a de Laplace e a frequencista, dão a mesma probabilidade de ocorrência de um acontecimento no caso de os resultados da experiência serem equiprováveis.

Assim, a definição frequencista de probabilidade permite calcular a probabilidade de um qualquer acontecimento, mesmo que os resultados elementares não sejam equiprováveis. Por exemplo, se a moeda fosse viciada, para calcular a probabilidade de o lado "Face" ficar voltado para cima, ter-se-ia que repetir o lançamento de uma moeda um grande número de vezes e ir calculando, lançamento a lançamento, a frequência relativa de saída de "Face", até à sua estabilização num determinado valor. Esse valor, para o qual tende a frequência relativa é considerado, para esta moeda, a probabilidade de ocorrência do lado "Face" quando é lançada ao ar.

Ora, na vida prática, não faz sentido repetir a experiência um número infinito de vezes. Assim sendo, quantas vezes se repete? Esta pergunta remete para uma questão muito frequente. Qual deve ser a dimensão da amostra? Com efeito, os resultados obtidos nos sucessivos lançamentos da moeda (ou do dado) podem ser vistos como as observações que constituem uma amostra para tentar entender a probabilidade de ocorrência de saída de "Face". O tamanho de uma amostra deve ter em consideração se o que estamos interessados em estudar apresenta grande ou pequena variabilidade entre os elementos da população de onde a amostra vai ser recolhida. De qualquer modo, a definição do tamanho da amostra e como se procede à sua recolha fazem parte do capítulo da Amostragem, um campo imenso dentro da Estatística, que não cabe no âmbito desta aula.

Assim, repetindo a experiência de lançar a moeda um número préfixado de vezes, a frequência relativa de saída de "Face" não é mais do que uma estimativa da probabilidade de ocorrer "Face" e, sendo uma estimativa, está sempre acompanhada por um erro. Analogamente, quando se analisa uma amostra para tentar concluir sobre a população, esta inferência está, também, sujeita a um erro ou um risco. Ora, a Teoria da Probabilidade é usada em Estatística para quantificar esse risco.

Voltemos à situação ilustrada na figura 1. No processo de Inferên-

cia Estatística, ao tirar conclusões sobre a população com base, apenas, no que se observou na amostra, estamos, necessariamente, a cometer erros e a correr riscos de más conclusões que, através da Teoria da Probabilidade, podem ser quantificado e controlados.

Ora, nesta aula, vamos focar a nossa atenção em alguns aspetos relativos à análise da amostra. Com efeito, a análise da amostra é, talvez, o ponto mais importante da Estatística e é algo que está presente em tudo o que fazemos diariamente. Por exemplo, mesmo na situação mais usual do nosso dia-a-dia, quando aquecemos um prato de sopa no micro-ondas e queremos ver se já está à temperatura desejada, não comemos a sopa toda para, no fim, concluir que a deveríamos ter aquecido mais. O que fazemos é provar uma colher de sopa e decidir, com base na temperatura da sopa contida nessa colher, se devemos aquecer mais a sopa ou não. Ora, o que acabamos de fazer, o que é? Não é mais do que analisar uma amostra e concluir sobre a população com base no que se observa na amostra.

Desde o início do estudo de métodos estatísticos (na verdade, desde o ensino secundário), aprende-se, obtida uma amostra de dados, a organizá-los em tabelas, construir gráficos e calcular uma série de medidas. Infelizmente, muitas vezes, os alunos aprendem a calcular um vasto conjunto de medidas mas não adquirem o hábito de interpretar e criticar os valores que obtêm. Calculam a média, a moda, a mediana, variância, desvio-padrão, etc., mas muitas vezes não têm a noção da informação que estas medidas comportam.

As noções fundamentais da Estatística podem, na minha opinião, resumir-se a três conceitos: distribuição, localização e variabilidade. Assim, vejamos.

Considere-se um estudo em que queremos analisar o peso dos alunos do ISCAP. Para isso, recolhemos uma amostra de estudantes. Seja essa amostra constituída pelos alunos que estão dentro da sala de aula. Assim, começa-se por pesar os alunos em questão.

Os pesos encontrados estão representados na seguinte figura.

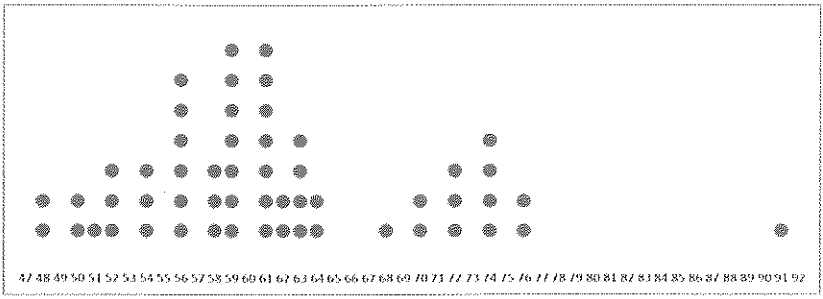


Figura 2

Estes valores podem ser apresentados numa tabela como a que se segue.

Peso	Nº de Alunos
45-55	11
55-65	31
65-75	10
75-85	2
85-95	1
Total	55

A escolha dos intervalos dos pesos é arbitrária, havendo, contudo, algumas normas e sugestões bem acolhidas pela maioria dos estatísticos. Esta distribuição de frequências pode ser apresentada num gráfico da seguinte forma.

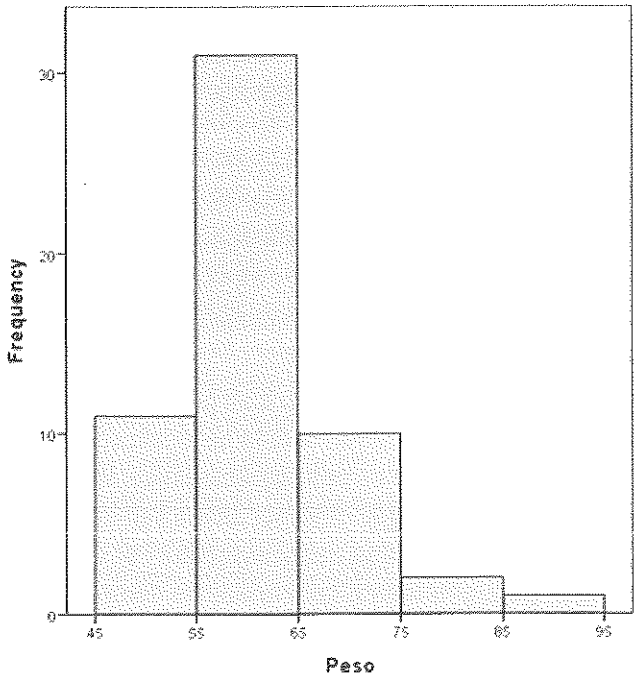


Figura 3

Qualquer uma destas apresentações dos dados começa a contar algo sobre os pesos dos alunos. Mostra como estes pesos se distribuem pelos possíveis valores. Rapidamente se observa que há uma grande concentração de valores entre os 55 e os 65 Kg, há poucos estudantes com peso superior a 75 Kg, etc.

Chegou agora a altura de escolher um valor que se considere característico ou representativo do peso de um aluno do ISCAP. Talvez o valor que surge logo na nossa mente seja a média aritmética, que neste caso é

$$\bar{x} = \frac{1}{55}\sum_{i=1}^{55} x_i = 61.455$$

Mas, afinal, o que representa o peso médio? Este valor é aquele que seria o peso de um aluno caso todos os estudantes pesassem o mesmo. Neste caso, cada um dos 55 estudantes pesaria 61.455Kg.

O valor médio pode, do ponto de vista físico, ser entendido como o centro de gravidade do sistema composto pelo conjunto das observações. Imagine-se uma reta sobre a qual vão sendo colocados pesos unitários simbolizando cada uma das observações. A Figura 2 ilustra esta situação. Então, o valor médio (61.455) representa o ponto onde deve ser colocado um suporte para que a reta, com os pesos em cima, se mantenha em equilíbrio.

Sendo a média o valor que caracteriza o peso dos alunos incluídos nesta amostra, esta é, então, uma medida de localização. Indica onde, numa reta, se situa ou localiza o peso característico. Mas será a média uma medida de localização que deva ser usada sem reservas? Será a média sempre uma boa medida de localização? Por vezes, não é, tal como se ilustra frequentemente com anedotas como aquela que diz que se uma pessoa come, num dia, dois frangos e outra pessoa não come nenhum, está tudo bem porque, em média, cada um comeu um frango; ou aquela que diz que o estatístico que tem a cabeça dentro de um forno e os pés num balde de gelo está, em média, com uma temperatura confortável. O que aqui falta avaliar é se existem observações extremas e analisar a variabilidade nelas presente.

A média é uma boa medida de localização desde que o conjunto de dados não apresente valores muito extremos (baixos ou altos). No conjunto acima, há um estudante que se destaca por pesar 91Kg, muito mais do que a generalidade dos estudantes pesados. Ora, este valor extremo pode influenciar muito a média, arrastando-a para um valor que não caracteriza o fenómeno "peso" em estudo. Para compreender bem o efeito de um valor extremo sobre a média de um conjunto de dados, considere-se a seguinte amostra (muito pequena, mas fácil e suficiente para apresentar a situação pretendida):

12 14 11 11 13

cujo valor médio é 12.2. Suponhamos, agora, que um dos valores é

extremamente diferente dos restantes, o que até pode ser devido a um erro de digitação. Seja, por exemplo,

12 74 11 11 13

cuja média é 24.2. Ora, o valor médio, que praticamente duplicou, foi fortemente influenciado pelo facto de uma observação com o valor 14 ter sido alterada para uma observação com o valor 74. Isto mostra que a média, embora muito usada para caracterizar estatisticamente um fenómeno, é muito influenciável por valores extremos. Diz-se, então, que a média é uma medida pouco robusta.

Mas, então, como proceder para contornar situações em que possam existir observações extremas?

Um possível método consiste em excluir as observações mais altas e as observações mais baixas e calcular a média dos restantes valores. Chama-se a este processo o cálculo de uma média aparada. Como exemplo de uma média aparada, temos a classificação atribuída pelo Júri numa competição de patinagem artística ou de ginástica rítmica. Usualmente, a nota mais alta e a nota mais baixa são ignoradas e a nota final resulta da média das classificações não excluídas.

Um outro procedimento consiste em calcular uma medida de localização que seja pouco influenciada por observações extremas. Trata-se da mediana. Ora, o que é a mediana? A mediana de um conjunto de dados representa o valor que está no meio desse conjunto. Assim, chamem-se cinco pessoas para tentar encontrar a mediana das alturas. A pessoa que está no meio do conjunto é a terceira pessoa. Trata-se de uma pessoa com a altura mediana? Usualmente, não. Para encontrar a mediana, será necessário começar por colocar as pessoas em fila, da mais baixa à mais alta (ordenar as observações de modo crescente) e selecionar a pessoa que está no meio. A mediana será a terceira altura do conjunto ordenado das cinco alturas. Assim, a mediana é definida como o valor que divide o conjunto de observações em duas partes de igual tamanho. E se o conjunto de observações tiver um número par de elementos? Nesta situação, e supondo que temos quatro pessoas, teríamos que começar por ordenar as observações e o valor que divide

o conjunto de dados em duas partes de igual tamanho seria a altura de um individuo imaginário que estaria entre o segundo e o terceiro valor da amostra ordenada. Geralmente, convenciona-se que esse indivíduo imaginário terá uma altura correspondente à média aritmética das alturas do segundo e do terceiro elementos. A esse valor chama-se, então, a mediana do conjunto de observações.

Note-se que a mediana é uma medida de localização mais robusta do que a média, uma vez que, mesmo que substituíssemos o indivíduo mais alto por um gigante, o valor mediano mantinha-se.

Assim, na análise de um conjunto de dados, o cálculo da média deve ser acompanhado pelo cálculo da mediana e estas duas medidas devem ser interpretadas em conjunto. Ambas fornecem detalhes importantes sobre a população em estudo.

Depois de vermos a importância de conhecer a distribuição dos dados e de ter uma ideia sobre a localização destes, interessa analisar a sua variabilidade. Com efeito, a variabilidade talvez seja a informação estatística mais importante que um conjunto de dados contém. Se não existisse variabilidade nas observações, não se usariam metodologias estatísticas, mas sim matemáticas. A estatística só se utiliza quando as observações, por serem influenciadas pelo acaso, apresentam uma variabilidade que não podemos controlar.

Como é óbvio, conclusões obtidas a partir de uma amostra de observações de um fenómeno que apresenta pequena variabilidade são mais fiáveis do que aquelas que se obtêm a partir de uma amostra de observações relativas a um fenómeno com grande variabilidade. Assim, conhecer e poder quantificar a variabilidade de um conjunto de dados é importante, ou mesmo fundamental, para poder realizar um bom estudo estatístico.

A distribuição dos pesos dos alunos da turma (Figura 1) ilustra como estes variam. Mas como pode ser quantificada esta variabilidade através de uma medida interpretável?

Para estudar a variabilidade de um conjunto de dados e por facilidade de apresentação, considere-se a situação seguinte.

São distribuídos, a 16 pessoas, bilhetes que têm escrito o nome

do jogo que as pessoas jogaram (jogos A, B, C e D) e o prémio que cada pessoa recebeu. Os bilhetes têm os seguintes valores:

Jogo A:	20€	20€	20€	140€
Jogo B:	50€	50€	50€	50€
Jogo C:	30€	30€	70€	70€
Jogo D:	10€	30€	40€	120€

Com base nos prémios recebidos por estas pessoas, queremos decidir qual dos jogos será preferível jogar. Ora, quando jogamos, o objetivo é ganhar o máximo possível. Assim, poderíamos começar por calcular qual é o prémio que, em média, se ganhou em cada um dos jogos, para optar por aquele que tem um prémio médio maior. Ora, o prémio médio em cada um dos jogos é

Jogo 1: $\frac{20+20+20+140}{4} = 50€$

Jogo 2: $\frac{50+50+50+50}{4} = 50€$

Jogo 3: $\frac{30+30+70+70}{4} = 50€$

Jogo 4: $\frac{10+30+40+120}{4} = 50€$

verificando-se que, todos os jogos, nas amostras recolhidas, apresentaram igual ganho médio. Contudo, pelos valores observados, constata-se que os ganhos, nos quatro jogos, não são idênticos. Por exemplo, enquanto no Jogo 2, o ganho foi sempre de 50€, no Jogo 1, este variou entre 20€ e 140€. A figura 4 mostra os ganhos observados em cada um dos quatro jogos.

Ora, tendo estes quatro conjuntos de prémios igual valor médio, em que diferem eles? Como se pode ver na Figura 4, os quatro conjuntos de valores diferem na variabilidade que eles apresentam. Enquanto no Jogo 2 os valores não apresentam qualquer variabilidade, nos outros jogos os prémios variam, nuns mais e noutros menos. Este fator, variabilidade, pode ser decisivo para um jogador escolher o jogo que vai jogar. Por exemplo, um jogador muito cauteloso optaria pelo Jogo 2, uma vez que parece que um prémio de 50€ está garantido. Por outro lado, um jogador que goste de arriscar talvez optasse pelo Jogo

1 pois, se tudo corresse bem, ele poderia ganhar 140€. É claro que, tendo azar, ganharia apenas 20€.

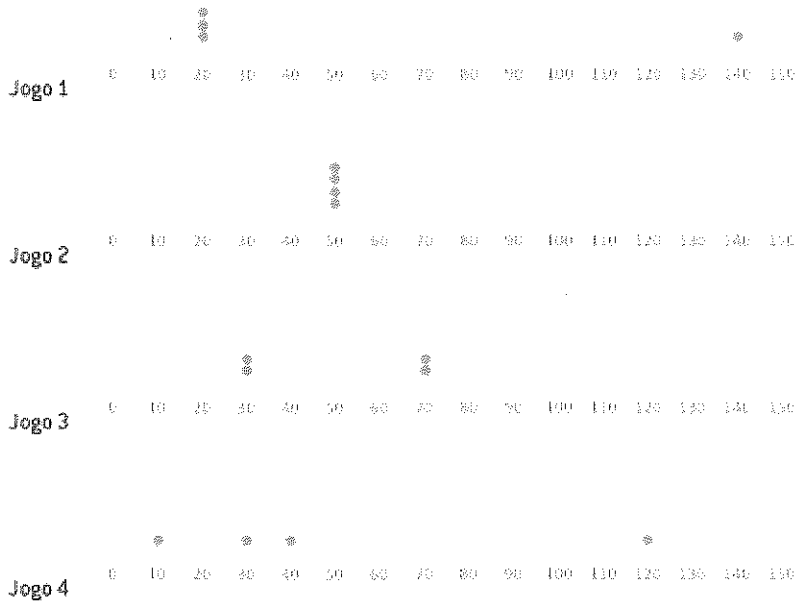


Figura 4

Este exemplo mostra que a variabilidade de um fenómeno está diretamente relacionada com a noção de risco. Assim, é importante tentar definir uma medida que avalie a variabilidade existente num conjunto de valores. Tal medida, idealmente, tomaria o valor zero quando não existe variabilidade e, quanto mais variabilidade houver, maior deveria ser o valor assumido pela medida.

Olhando para a Figura 4, nota-se que a variabilidade está relacionada com o espaçamento entre os diversos números registados. Variando os valores observados entre um valor mínimo e um valor máximo, surge como medida natural da variabilidade, o conceito de amplitude:

$$A = \max(x_i) - \min(x_i)$$

Com efeito, esta medida, no caso do Jogo 2, em que não há variabilidade, toma o valor zero e, quanto maior for a variabilidade, parece haver tendência para que a amplitude seja maior:

Jogo 1: $A = 140 - 20 = 120$
Jogo 2: $A = 50 - 50 = 0$
Jogo 3: $A = 70 - 30 = 40$
Jogo 4: $A = 120 - 10 = 110$

Como já era de esperar, a variabilidade dos prémios relativos ao Jogo 2 é nula; além disso, observando os valores obtidos para as amplitudes, seremos levados a concluir que no Jogo 1 é onde há maior variabilidade, logo seguido do Jogo 4.

Porém, a amplitude, medida de variabilidade extremamente fácil de calcular, é muito pouco robusta uma vez que a amplitude considera, para o seu cálculo, apenas dois valores (o mínimo e o máximo) ignorando todos os outros. Ora, sendo os dados a maior fonte de riqueza da Estatística, não é admissível que estejamos a ignorar a informação contida em cada uma das observações.

Assim, é importante continuar a procurar uma medida de variabilidade que considere todas as observações do conjunto.

Quando se está a avaliar a variabilidade de um conjunto de dados, é preciso definir variabilidade estatisticamente. Este conceito pode ser visto como sendo o modo como as observações se desviam em relação a uma medida de localização, por exemplo, a média. As oscilações das observações relativamente ao prémio médio estão ilustradas na Figura 5.

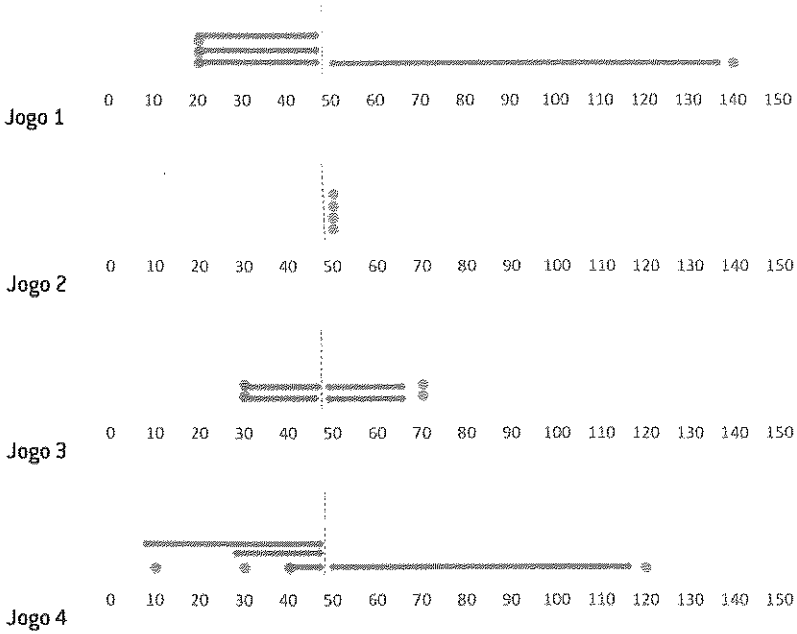


Figura 5

Podemos, então, pensar em definir uma medida de variabilidade que considere todos os desvios dos prémios observados em relação ao prémio médio, calculando as correspondentes diferenças. Assim

Jogo 1: 20-50=-30
20-50=-30
20-50=-30
140-50=90

Jogo 2: 50-50=0
50-50=0
50-50=0
50-50=0

Jogo 3: 30-50=-20
30-50=-20
70-50=20
70-50=20

Jogo 4: 10-50=-40
30-50=-20
40-50=-10
120-50=70

Como queremos que a medida de variabilidade tome em consideração as 4 diferenças para cada um dos jogos, podemos pensar em somar essas diferenças. Ora, somando, para cada um dos jogos, as 4 diferenças, concluímos que essa soma dá, sempre, o valor zero. Por que razão? Isso acontece porque quando consideramos as observações à esquerda da média a diferença dá negativa e quando consideramos observações à direita da média a diferença dá positiva, donde resulta, sempre, que a soma das diferenças seja nula.

Contudo, note-se que, para avaliar o grau de variabilidade das observações, não nos interessa se a observação está à direita ou à esquerda da média. O que na verdade interessa é o tamanho do desvio da observação em relação à média, não sendo, por isso, relevante o sinal da diferença. Assim, poder-se-ia resolver o problema considerando os valores absolutos das diferenças. Ter-se-ia, então

Jogo 1: |20-50|=30
|20-50|=30
|20-50|=30
|140-50|=90
Soma= 30+30+30+90=180

Jogo 2: |50-50|=0
|50-50|=0
|50-50|=0
|50-50|=0
Soma= 0+0+0+0=0

Jogo 3: |30-50|=20
|30-50|=20
|70-50|=20
|70-50|=20
Soma= 20+20+20+20=80

Jogo 4: |10-50|=40
|30-50|=20
|40-50|=10
|120-50|=70
Soma= 40+20+10+70=140

Neste exemplo, a soma permite concluir sobre a variabilidade de cada um dos jogos quando comparado com os outros, uma vez que, para todos os jogos, foram recolhidas 4 observações. Se tal não sucedesse, a soma não permitiria concluir nada. Considere-se um jogo como o Jogo 3, que apresenta pequena variabilidade quando comparado com o Jogo 4. Se tivéssemos recolhido uma amostra com 50 prémios do Jogo 3 e, relativamente ao Jogo 4, tivéssemos poucas observações, seria natural que a soma dos tamanhos dos desvios observados fosse maior para o Jogo 3 do que para o Jogo 4. Com efeito, não é de estranhar que a soma de muitas parcelas pequenas possa, facilmente, ser maior do que a soma de poucas parcelas grandes. Assim, é importante relativizar as somas que fizemos em função do número de observações de cada conjunto. Tal é feito calculando o tamanho médio do desvio, isto é, dividindo a soma obtida pelo número de observações de cada conjunto.

Surge, assim, a definição de desvio médio:

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

onde n é o número de observações do conjunto e $\bar{x}$ é a média do conjunto.

No nosso exemplo teríamos:

Jogo 1: DM=180/4=45
Jogo 2: DM=0/4=0
Jogo 3: DM=80/4=20
Jogo 4: DM=140/4=35

Recordando que a introdução do valor absoluto ou módulo no cálculo do desvio médio teve como objetivo tirar o sinal da diferença entre o valor observado e o valor médio, pode pensar-se que, alternativamente, poderíamos ter calculado o quadrado das diferenças. Com efeito, ao calcular o quadrado das diferenças, o sinal desaparece. No exemplo que estamos a considerar, teríamos:

Jogo 1: $(20-50)^2=30^2=900$
$(20-50)^2=30^2=900$
$(20-50)^2=30^2=900$
$(140-50)^2=90^2=8100$
Soma= 900+900+900+8100=10800

Jogo 2: $(50-50)^2=0^2=0$
$(50-50)^2=0^2=0$
$(50-50)^2=0^2=0$
$(50-50)^2=0^2=0$
Soma= 0+0+0+0=0

Jogo 3: $(30-50)^2=20^2=400$
$(30-50)^2=20^2=400$
$(70-50)^2=20^2=400$
$(70-50)^2=20^2=400$
Soma= 400+400+400+400=1600

Jogo 4: $(10-50)^2=40^2=1600$
$(30-50)^2=20^2=400$
$(40-50)^2=10^2=100$
$(120-50)^2=70^2=4900$
Soma= 1600+400+100+4900=7000

As comparações entre estas somas permitem fazer, também, comparações entre as variabilidades existentes nos respetivos conjuntos de observações. Contudo, tal como anteriormente, a comparação das somas só é possível porque todos os conjuntos têm o mesmo número de observações. Para prevenir situações em que as amostras têm tamanhos diferentes, importa relativizar estas somas com base na dimensão das amostras. Para isso, deveríamos dividir estas somas pelo número de observações que formam cada conjunto.

Surge assim a definição de uma nova medida de variabilidade: a variância. Esta medida é definida por

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

A variância, ao longo do tempo, foi ganhando preferência em relação ao desvio-médio, anteriormente apresentado, por ser, frequentemente, preferível a transformação quadrática relativamente à transformação em módulo. Tal prende-se, essencialmente, com o facto de a transformação quadrática ser muito "bem comportada" do ponto de vista matemático, ao contrário da função módulo. Esta função, por exemplo, não é derivável na origem.

Note-se que, regra geral, o valor da variância de um conjunto de dados não é interpretável. Com efeito, ao definirmos a variância, já "estragámos" o valor que da sua definição resulte: começamos por considerar o tamanho dos desvios absolutos das observações em relação à média e, depois, por conveniência matemática, elevamos os tamanhos dos desvios ao quadrado. Logo, a variância pode ser vista como a média dos quadrados dos desvios das observações em relação à média do conjunto. Ora, isto não é uma grandeza que possa ser interpretável. Então, se assim é, para que serve calcular a variância? Dados dois desvios absolutos, se o primeiro é menor que o segundo, em tamanho, então o quadrado do primeiro é menor que o quadrado do segundo. Ou seja, a relação mantém-se após a transformação qua-

drática. Assim, a variância deve ser usada para comparar as variabilidades de dois ou mais conjuntos de dados e não para dizer se um só conjunto apresenta grande ou pequena variabilidade.

Relativamente ao exemplo considerado, temos as variâncias (V) seguintes:

Jogo 1: V=10800/4=2700
Jogo 2: V=0/4=0
Jogo 3: V=1600/4=400
Jogo 4: V=7000/4=1750

Mas a variância, como medida de variabilidade, tem um problema: as suas unidades. No exemplo do jogo, os prémios estão expressos em Euros e, por isso, o prémio médio vem em Euros; assim, a diferença entre cada valor observado e o prémio médio está em euros; elevando esta diferença ao quadrado, o resultado vem em Euros^2; somadas as parcelas, esta soma vem em Euros^2 e dividindo pelo número de observações, conclui-se, por fim, que a variância vem expressa em Euros^2. Mas que unidade é esta? O que significa 1 Euro^2? Nada! Assim, surgiu a vontade de definir uma medida que avalie a variabilidade e que não tenha este problema de unidades. Como resolver? Calcule-se a raiz quadrada e o resultado virá em Euros novamente. Surge, assim, a medida talvez mais usada em estatística para avaliar a variabilidade: o desvio-padrão que não é mais do que

$$DP = \sqrt{V}$$

Mas faz sentido calcular a raiz quadrada só para resolver o problema das unidades? E o valor propriamente dito o que vai significar? Mais uma vez, nota-se que o desvio-padrão, em regra geral, só deve ser calculado para comparar a variabilidade de dois ou mais conjuntos de dados e, se $V_X < V_Y$, então, tem-se $DP_X < DP_Y$ e, por isso, as conclusões obtidas à custa do desvio-padrão são as mesmas que se obtiveram com a variância, agora com a vantagem de que as unidade já não são estranhas.

Será que a procura de uma medida eficaz para avaliar a variabilidade de um conjunto de dados já está completa?

O desvio-padrão é, sem dúvida, a medida de variabilidade mais usada. Contudo, também tem as suas fragilidades. Considere-se a seguinte situação. Dois médicos fizeram um estudo sobre o peso dos seus doentes. Para isso, entre outras medidas, calcularam o desvio-padrão dos pesos, com vista a avaliar a variabilidade destes. Suponhamos que o médico A encontrou um desvio-padrão de 2Kg no conjunto dos pesos dos seus doentes e que o médico B encontrou um desvio padrão de 0.8Kg. Olhando para este valores, conclui-se que os pesos dos doentes do médico A apresentam maior variabilidade do que os pesos dos doentes do médico B, uma vez que 2 > 0.8. Suponhamos, agora, que recebemos a informação de que o médico A é um médico de medicina interna que trabalha com adultos e que o médico B é um neonatalogista. Continuamos com a mesma opinião sobre as variabilidades dos pesos nos dois grupos? Certamente que não: uma variação na ordem dos 800 gramas num recém-nascido é muito mais significativa do que uma variação na ordem dos 2 quilos num adulto. Ora, isto acontece porque, nos dois grupos, as médias dos pesos são completamente diferentes. Assim, importa ver qual a variabilidade existente por unidade média do peso. Surge assim uma outra medida de variabilidade: o coeficiente de variação, definido por

$$CV = \frac{DP}{\bar{x}}$$

O coeficiente de variação deve ser usado quando se pretende comparar as variabilidades de dois ou mais conjuntos de dados que têm valores médios significativamente diferentes.

Mas, além disso, note-se que o coeficiente de variação é uma medida sem unidades. Por isso, também é útil quando se pretende comparar as variabilidades de dois ou mais conjuntos de dados que estão expressos em unidades diferentes e não convertíveis.

Atrás foi dito que a variância, o desvio padrão e o coeficiente de

variação devem ser usados apenas para comparar duas ou mais populações relativamente à sua variabilidade. Esta afirmação é, até certo ponto, forçada. Ela foi feita com base no que se acabou de apresentar. Contudo, o valor assumido por estas medidas e, em especial, pelo desvio-padrão, pode ser, até certo ponto, interpretado.

No caso particular em que se sabe que a população é normalmente distribuída, pode afirmar-se que o intervalo que é obtido pela média mais ou menos duas vezes o desvio padrão contém cerca de 95% da população. Note-se que esta interpretação só é válida se a suposição de normalidade da população se verificar.

De um modo mais geral, existe uma importante desigualdade, cuja explicação sai fora do contexto desta apresentação, chamada desigualdade de Chebyshev, cuja expressão é

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

onde μ e σ representam, respetivamente, a média e o desvio-padrão da população, que permite concluir, por exemplo, que, qualquer que seja a população em estudo, o intervalo definido pela média mais ou menos dois desvios-padrão contém, pelo menos, 75% dos elementos da população.

O objetivo desta aula foi mostrar como os conceitos mais básicos da Estatística podem ser definidos de um modo intuitivo e como esta ciência é apelativa, não sendo assim tão complicada a sua aprendizagem.