Medidas de Tendência Central

Média

Quando queremos um único número que resuma uma variável quantitativa de um banco de dados pensamos logo na média.

- Qual é a nota média dos alunos de uma turma?
- Por que homens têm uma renda meedia superior a das mulheres?
- Qual é a média de gols que um atacante faz por jogo?

A média é a mais simples medida estatística. Basta somar os valores de uma distribuição numérica e dividir pelo número de casos (observações).

média = soma de valores/ número de casos

Na distribuição composta por nove números:

- 78, 91, 94, 98, 99, 101, 103, 105, 114

A média é (883/ 9) = 98

Mediana

A mediana é menos conhecida, mas é uma medida de tendência central usada, sobretudo, quando existem valores extremos na distribuição.

Quando ordenamos uma distribuição de casos do menor para o maior número (ou vice-versa), a mediana é o número que divide a distribuição ao meio; ou seja, metade dos números estão acima, metade abaixo.

Na distribuição abaixo a mediana é 99

- 78, 91, 94, 98, **99**,101, 103, 105, 114

No caso de a distribuição ter um número par de dados, as duas observações centrais são somadas e dividas por dois

Na distribuição abaixo a mediana é 100 = 99 + 101/2

- 78, 91, 94, 98, **99**, **101**, 103, 105, 114, 121.

Exercício

Use uma calculadora e calcule: a média e a mediana se acrescentarmos uma observação de valor 1200 na distribuição abaixo:

- 78, 91, 94, 98, 99, 101, 103, 105, 114, 121.

Qual dos dois valores parece mais represntataivo dos valores da distribuição?

Resposta

Moda

A moda é valor que aparece com mais frequência na distribuição

Na distribuição abaixo a moda é 101, já que aparece duas vezes

- 78, 91, 94, 98, 99, **101**, **101**, 103, 105, 114

⚠️Cuidado com a média

A média é sensível a observações com valores extremos:

- Uma pessoa com renda extremamente alta pode oferecer um número enviezado da renda

de um grupo
- A presença da cidade São Paulo envieza a média de morados nos municípios da reginao Metropolitana do estado de São Paulo

Nesses casos, melhor utilizar a mediana

A mediana não é a mensagem

texto do biólogo americano, Stephen Jay Gould (1941-2002)

Minha vida recentemente se deparou, de uma maneira muito pessoal, com duas das célebres frases de Mark Twain. Uma eu deixarei para o fim deste ensaio. A outra (por vezes atribuída a Disraeli), identifica três tipos de inverdades, cada uma pior que a outra: mentiras, mentiras descaradas, e estatísticas.

Considere o exemplo clássico de estender a verdade com números — um caso bastante relevante à minha estória. A estatística reconhece diferentes medidas para "na média", ou para a tendência central. Média aritmética é nosso conceito usual de "na média" — adicione todos os itens e divida pelo número que os compartilha (100 doces coletados por cinco crianças no próximo Halloween irão dar 20 para cada uma num mundo justo). A mediana, uma medida diferente da tendência central, é o ponto no meio do caminho. Se eu colocar em fila indiana cinco crianças em ordem de tamanho, a criança mediana é mais baixa que duas, e mais alta que as outras duas (que podem ter dificuldade em obter a fração justa dos doces). Um político no poder pode dizer com orgulho, "A renda média de nossos cidadãos é de 15.0 por ano." O líder da oposição pode responder "Mas metade de nossos cidadãos ganha menos que 10.0 por ano." Ambos estão corretos, mas nenhum cita a estatística com objetividade fria. O primeiro invoca a média aritmética, o segundo a mediana. (Médias são mais altas que medianas nesses casos porque um milionário vale mais que centenas de pobres ao determinar a média, mas ele pode contrabalancear apenas um mendigo no cálculo da mediana).

A questão maior que comunmente cria desconfiança ou desdenho pela ciência estatística é mais perturbadora. Muitas pessoas fazem uma separação infeliz e inválida entre coração e mente, sentimento e intelecto. Em algumas tradições contemporâneas, incentivadas por atitudes estereotipicamente centradas no sul da Califórnia, o sentimento é exaltado como sendo mais "real" e a única base apropriada para agir — se a sensação é boa, faça — enquanto o intelecto é minimizado como sendo apêndice de um elitismo fora de moda. A Estatística, nessa dicotomia absurda, acaba muitas vezes se tornando o símbolo do inimigo. Como Hilaire Belloc escreveu, "A Estatística é o triunfo do método quantitativo, e o método quantitativo é a vitória da esterilidade e da morte."

Esta é uma estória pessoal da estatística, propriamente interpretada, como profundamente doce e

portadora de vida. Ela declara uma guerra santa à desclassificação do intelecto contando uma pequena estória sobre a utilidade de conhecimento acadêmico, árido, científico. O coração e a cabeça são pontos focais de um só corpo, de uma só personalidade.

Em julho de 1982 eu descobri que sofria de mesotelioma abdominal, um câncer raro e grave normalmente associado com a exposição a abestos. Quando acordei depos da cirurgia , fiz minha primeira pergunta à minha médica e quimioterapeuta: “Qual a melhor literatura técnica sobre mesotelioma ?” Ela respondeu, com um toque de diplomacia (a única vez que fugiu da franqueza direta), que a literatura médica não continha nada que realmente valesse a pena ler.

Claro que manter um intelectual longe da literatura funciona tão bem quanto recomendar castidade ao Homo Sapiens, o mais sexuado de todos os primatas. Assim que pude andar, fui direto para a biblioteca médica Countway, em Harvard, e digitei mesotelioma no programa de computador de pesquisas bibliográficas. Uma hora mais tarde, cercado da mais recente literatura sobre mesotelioma abdominal, eu entendi com um nó na garganta por quê minha médica havia me dado aquele conselho humano. A literatura não poderia ser mais brutalmente clara: mesotelioma é incurável, com uma mortalidade mediana de apenas oito meses após o diagnóstico. Eu fiquei sentado atordoado por cerca de quinze minutos, então sorri e disse para mim mesmo: ah, é por isso que não me deram nada para ler. Então minha cabeça voltou a funcionar, ainda bem.

Se um pouco de aprendizado puder ser, em algum momento, algo perigoso, eu tinha encontrado o exemplo clássico. Atitude claramente importa ao lutar com o câncer. Não sabemos por quê (de minha visão materialista clássica, eu suspeito de estados mentais reforçando o sistema imunológico). Mas equalize pessoas com câncer por idade, sexo, saúde, status socio-econômico, e, em geral, aqueles com atitudes positivas, com forte propósito e vontade de viver, com compromisso para lutar, com uma resposta ativa de auxiliar seu próprio tratamento, e não apenas uma aceitação passiva do que o médico diz, tenderão a viver por mais tempo. Alguns meses mais tarde eu perguntei a Sir Peter Medawar, meu guru científico pessoal e prêmio Nobel em imunologia, qual a melhor receita para ter sucesso contra o câncer. “Uma personalidade sanguínea”, ele respondeu. Por sorte (pois uma pessoa não consegue se reconstruir no curto prazo e para um propósito definido), eu sou, se alguma coisa, de temperamento constante e confiante exatamente como receitado.

Vê-se o dilema para médicos humanos: se a atitude importa tão criticamente, deveria uma conclusão tão sombria ser propagandeada, especialmente sabendo que tão poucas pessoas possuem entendimento de estatística para avaliar o quê essas afirmações realmente signficam ? De meus anos de experiência com a evolução em pequena escala dos caramujos terrestres das Bahamas, tratada quantitativamente, eu desenvolvi esse entendimento técnico — e estou convencido que ele teve um papel central em salvar a minha vida. Conhecimento é poder, como diz o provérbio.

O problema pode ser posto de forma breve: O que significa “mortalidade mediana de 8 meses” em nosso vernáculo ? Eu suspeito que a maior parte das pessoas, sem treinamento em estatística, leria

nosso vernáculo? Eu suspeito que a maior parte das pessoas, sem treinamento em estatística, leria essa afirmação como "Eu provavelmente vou estar morto em oito meses" — exatamente a conclusão que deveria ser evitada, porque não é verdadeira, e porque a atitude importa tanto.

Eu não estava, claro, muito feliz, mas não li a afirmação de acordo com esse vernáculo fatalista. Meu treinamento técnico formou uma perspectiva diferente para "8 meses de mortalidade mediana". O ponto é sutil, mas profundo — pois ele incorpora o modo de pensar caracterísitico do meu campo de estudo, biologia evolucionária e história natural.

Nós ainda carregamos a bagagem histórica da herança Platônica que busca essências puras e fronteiras definidas. (Por isso nós esperamos encontrar um "começo da vida" ou uma "definição de morte" sem ambiguidade, apesar da natureza frequentemente se apresentar como um contínuo irredutível). Essa herança Platônica, com sua ênfase em distinções claras e entidades imutáveis separadas, nos leva a ver medidas estatísticas de tendência central de forma errada, na verdade de forma oposta à interpretação apropriada ao nosso mundo real de variações, tonalidades, e continuidade. Ou seja, nós olhamos médias e medianas como "dura realidade," e a variação que permite o seu cálculo como uma massa de medidas transientes e imperfeitas dessa verdadeira essência. Se a mediana é a realidade e a variação em torno da mediana um artifício para seu cálculo, então "provavelmente vou estar morto em oito meses" pode passar como uma interpretação razoável.

Mas todos biólogos evolucionários sabem que a própria variação é a única essência irredutível da natureza. Variação é a dura realidade, não um conjunto de medidas imperfeitas da tendência central. Médias e medianas são a abstração. Portanto, eu olhei as estatísticas do mesothelioma bem diferente — e não só porque eu sou um otimista que tende a ver a rosca ao invés do buraco, mas primariamente porque sei que a própria variação é a realidade. Eu precisava me colocar dentro da variação.

Quando eu soube da mediana de oito meses, minha primeira reação intelectual foi: ótimo, metade das pessoas vive mais que isso; agora, quais são as minhas chances de estar naquela metade? Eu li por uma hora nervosa e furiosa, e conclui, com alívio: danadas de boas. Eu possuia cada uma das características que conferiam a probabilidade de uma vida mais longa: Eu era jovem; minha doença havia sido identificada em um estágio relativamente inicial; eu iria receber o melhor tratamento médico disponível; eu tinha um mundo pelo qual viver; eu sabia como ler os dados corretamente e não me desesperar.

Outro ponto técnico adicionou ainda mais consolo. Eu imediatamente me dei conta de que a distribuição da variação em torno da mediana de oito meses quase que certamente seria aquilo que os estatísticos chamam de "skewed para a direita". (Em uma distribuição simétrica, o perfil da variação à esquerda da tendência central é uma imagem espelhada da variação para a direita. Em distribuições enviesadas, a variação para um lado da tendência central é mais esticada — enviesada

para esquerda quando se extende para a esquerda, enviesada para direita quando se estica para a direita.) A distribuição da variação tinha que ser enviesada para direita, eu raciocinei. Afinal, a esquerda da distribuição continha uma fronteira irrevocável no zero (pois o mesothelioma só pode ser identificado no momento da morte ou antes). Portanto, não há muito espaço para a metade baixa (ou esquerda) da variação — ela precisa estar comprimida entre zero e oito meses. Mas a metade alta (ou direita) pode se estender por anos e anos, mesmo que ninguém sobreviva no fim. A distribuição tinha que ser enviesada para a direita, e eu precisava saber o quão longe a rabeira da direita chegava — pois eu já havia concluído que meu perfil favorável me fazia um bom candidato para aquela parte da curva.

A distribuição era, de fato, fortemente enviesada para direita, com uma rabeira longa (embora pequena) que se extendia por muitos anos além da mediana de oito meses. Eu não vi nenhuma razão para eu não estar naquela pequena rabeira, e pude respirar um longo suspiro de alívio. Meu conhecimento técnico havia ajudado. Eu havia lido o gráfico corretamente. Eu fiz a pergunta certa e encontrei as respostas. Eu tinha conseguido, bem provavelmente, o mais precioso presente — tempo substancial. Eu não teria que parar e seguir imediatamente a injunção de Isaías para Ezequias — coloque sua casa em ordem pois tu irás morrer, e não viver. Eu teria tempo para pensar, para planejar, e para lutar.

Um último ponto sobre distribuições estatísticas. Elas se aplicam somente para um conjunto determinado de circunstâncias — neste caso a sobrevivência com mesothelioma sob métodos convencionais de tratamento. Se as circunstâncias mudam, a distribuição pode se alterar. Eu fui colocado em um protocolo experimental de tratamento e, com sorte, estarei no primeiro grupo amostral de uma nova distribuição com mediana alta e uma rabeira direita esticando até uma morte natural em idade avançada.

Parece-me que se tornou excessivamente chique olhar a aceitação da morte como algo equivalente à posse de dignidade intrínseca. Claro que eu concordo com o pastor do Eclesiastes, que há um tempo para amar e um tempo para morrer — e quando meus dias terminarem espero encarar o fim calmamente, de minha própria maneira. Na maior parte das situações, no entanto, prefiro o ponto de vista mais marcial de que a morte é o terrível inimigo — e não acho nada criticável naqueles que se enfurecem com o apagar das luzes.

As espadas da batalha são muitas, e nenhuma mais efetiva do que o humor. Minha morte foi anunciada em um encontro de meus colegas na Escócia, e eu quase experimentei o delicioso prazer de ler o meu obituário escrito por um de meus melhores amigos (que suspeitou e checou; ele também é um estatístico, e não esperava me encontrar tão longe na rabeira direita.) Ainda asssim, o incidente me deu as primeiras gargalhadas depois do diagnóstico. Pense só, eu quase pude repetir uma das frases mais famosas de Mark Twain: as notícias de minha morte foram fortemente exageradas.

Epílogo por Steve Dunn

Muitos me escreveram para perguntar o que aconteceu com Stephen Jay Gould. Infelizmente, o Dr. Gould faleceu em Maio de 2002, com 60 anos de idade. O Dr. Gould viveu 20 anos muito produtivos após seu diagnóstico, portanto excedendo sua mediana de oito meses por um fator de trinta! Apesar dele ter morrido de câncer, aparentemente não foi mesothelioma, mas sim um segundo câncer não relacionado.

Em Março de 2002, o Dr. Gould publicou sua Obra Prima de 1342 páginas, "A Estrutura da Teoria Evolucionária". É justo que Gould, um dos cientistas e escritores mais prolíficos do mundo, tenha conseguido completar a expressão definitiva de seu trabalho científico e sua filosofia — e bem a tempo. O texto é longo demais e denso demais para quase qualquer leigo — mas os trabalhos de Stephen Jay Gould vão continuar vivos. Especialmente, espero, "A Mediana Não É a Mensagem".

Texto original de Stephen Jay Gould e Steven Dunn traduzido para o português em Maio de 2007, por Ricardo Castro. Original em inglês: The Median Isn't the Message.

Quantis

O quantil é definido como segmento de tamanho igual de uma determinada população. Uma das métricas mais comuns em análise estatística, a mediana, é na verdade apenas o resultado da divisão de uma população em dois quantis. Um quintil é um dos 4 valores que dividem os dados em 5 partes iguais, cada uma sendo 1/5 (20%). Uma população dividida em tercis tem 3 partes iguais, enquanto uma dividida em quartis tem 4 partes.

Os quantis são mensurados de formas de diferentes, mas sempre parte de uma distribuição de dados ordenado do menor para o maior valor, ou vice-versa.

Quartil

Um quartil divide os dados em três pontos – um quartil inferior, a mediana e o quartil superior – para formar quatro grupos do conjunto de dados.

Assim como a mediana divide os dados na metade, de modo que 50% da medida fique abaixo da mediana e 50% acima dela, o quartil divide os dados em quartos para que 25% das medidas sejam menores que o quartil inferior, 50 % sejam menores que a mediana e 75% sejam menores do que o quartil superior.

Na distribuição: 78, 91, 94, 98, 99, 101, 103, 105, 114, 117, 121

- O 94 é o quartil inferior, também conhecido como Q1
- O 101 é a mediana: metade dos números estão abaixo dele, metade acima
- O 114 é o quartil superior, também conhecido como o Q3

78, 91, **94**, 98, 99, **101**, 103, 105, **114**, 117, 121

Observe que 25% dos casos estão abaixo de 94; 50% dos casos estão abaixo de 101; 75% dos casos estnao abaixo de 114

Quintil

Um quintil é um valor estatístico de um conjunto de dados que representa 20% de uma determinada população, portanto, o primeiro quintil representa o quinto mais baixo dos dados (1% a 20%); o segundo quintil representa o segundo quinto (21% a 40%) e assim por diante.

Qual é o percentual de renda que cada quintil recebe no Brasil?

1. Ordene as famílias (ou indivíduos) do menor para o maior.
2. Calcule os quatro valores que dividirão a distribuição em 5 faixas
3. Some a renda de todos os indivíduos de um determinado quintil e calcule quanto esse resultado representa do precentual total
4. Em 2015, por exemplo, os 20% mais pobres ficavam com 3,6% da renda; enquanto os 20% mais ricos ficavam com 56% da renda

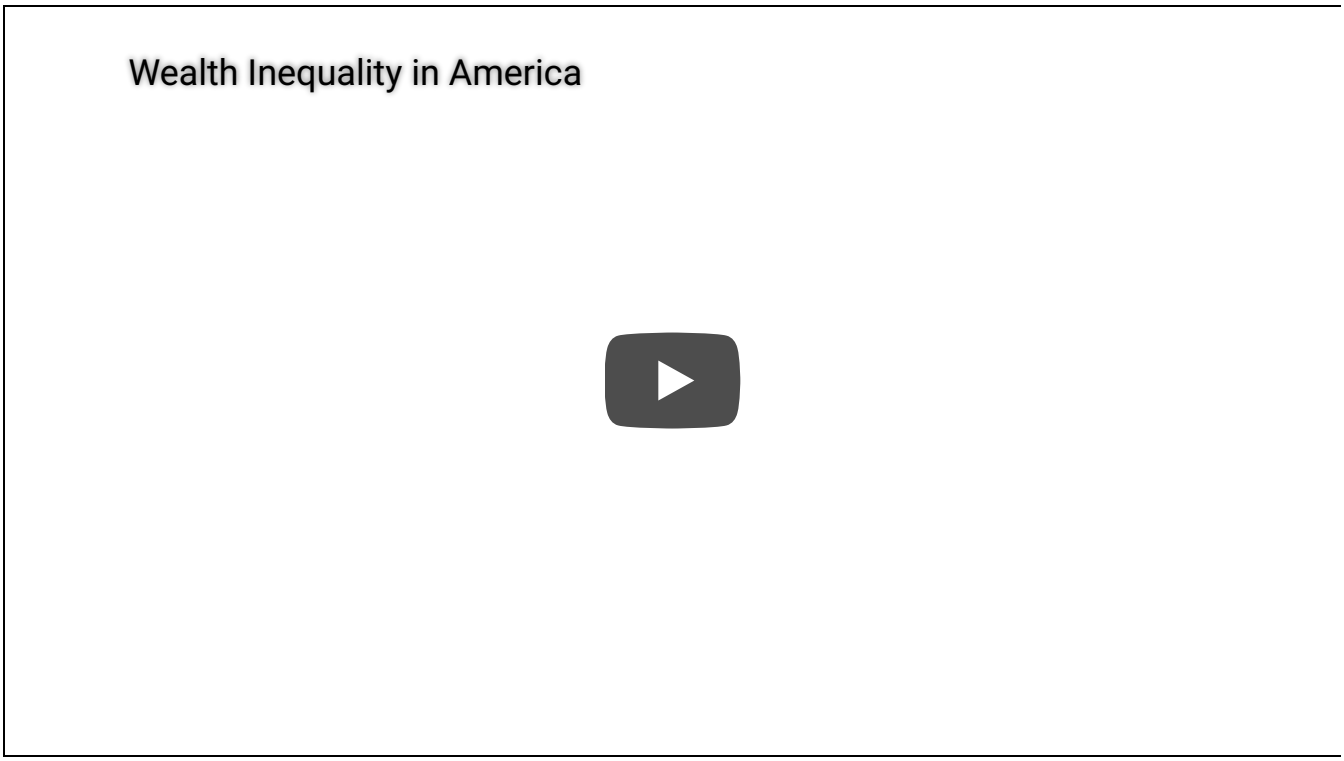
Veja o gráfico com os valores do Brasil e de outros países: Our world in data

Decil

Se divirmos uma distribuição em dez segmentos temos os decis. O procedimento é o mesmo: o primeiro decil representa o décimo mais baixo dos dados (1% a 10%); o decil o seguinte representa o segundo décimo (11 a 20% dos dados).

Imagine uma turma em que os alunos tenham tirado diferentes notas entre 2 e 10. Digamos que a nota do primeiro decil seja 3,5. Isso significa que os alunos que tiraram até 3,5 estão no primeiro decil. Se a nota que divide o decil superior é 9, isso significa que os alunos que estão nessa faixa tirararm pelo menos 9.

Esse vídeo usa os quintis e decis para mostrar a desigualdade de renda nos Estados Unidos.



Percentil

Uma distribuição numérica também pode ser segmentada em 99 percentis, representados por p1, p2, p3... p99, que dividem os dados em cem partes com cerca de 1% dos casos em cada um.

Um jovem com 1,80 metros está no 90 percentil da altura de uma dada população, significa que 90% jovens estão abaixo desse patamar e 10% dos jovens têm pelo menos 1,80 de altura.

⚠️ **Programas estatísticos e percentil**

Todos os programas estatísticos (R, Python, Julia, Stata, SPPS) calculam o os quintis de uma distribuição de maneira relativamente simples.

O usuário pode definir um percentil que ele gostaria de conhecer, ou trabalhar com os quintis mais comuns (quartil, quintil e decil).

O uso mais frequente dos quintis está associado a uma outra medida para compará-los, particularmente nos estudos sobre a desigualdade; por exemplo:

- Qual é a renda média de cada quintil?
- Qual é a mediana de votos gastos de campanha dos deputados em cada decil?

- Qual é o percentual de renda somada das pessoas de um quintil sobre a renda agregada de toda a população?

Um gráfico para observar os quintis

Um gráfico pouco usado, mas útil para observar a distribuição de observações é o ECDF (*Empirical Cumulative Distribution Function*).

No gráfico abaixo, observamos uma distribuição de cinco valores (-2, -1, 0, 1, 2). Fixe o olhar na linha em um ponto do eixo horizontal (x) e trace uma linha até o eixo vertical (y). O ponto em que a a linha imaginaria corta y observamos o percentil.

Para o valor 0, por exemplo, o percential é de cerca de 0,48; ou seja, cerca de 48% de casos encontram-se abaixo de zero

Medida de Dispersão

A média, a mediana e a moda mostra uma medida com um sumário geral dos dados. Um outro conjunto de medidas revela o padrão de dispersão de dados.

Podemos imaginar que existe uma variação maior da idade dos torcedores que comparecem a um jogo no Maracanã, do que dos frequentadores de um show de hip-hop, ou de um baile da terceira idade.

Que medida usar para mensurar a variabilidade de uma população?

Desvio Padrão

O desvio padrão é a medida de dispersão mais comum em estatística. Se tivermos que apresentar uma estatística que resuma a dispersão dos dados, geralmente a escolhida será o desvio padrão. Como o próprio nome sugere, o desvio padrão informa qual é o desvio “normal” dos dados. Na verdade, ele calcula o desvio médio dos valores em relação à média . Quanto maior o desvio padrão, mais dispersos são os dados; quanto menor o desvio padrão, mais os dados estão centrados em torno da média.

A fórmula do desvio padrão é denotado pela letra :

Como você pode ver na fórmula, o desvio padrão é o desvio médio dos dados em relação à média μ. Observe que o uso do quadrado da diferença entre as observações e a média é para evitar que diferenças negativas sejam anuladas pelas diferenças positivas.

Para facilitar, imagine uma população de apenas 3 adultos, com as seguintes alturas (em cm):

160,4 , 175,8 e 181,5

A média é 172,6. O desvio padrão é calculado da seguinte maneira:

O desvio padrão das alturas desses três adultos é de 8,91 cm. Isso significa que, em média, a altura dos adultos se desvia da média em 8,91 cm.

💡**Propriedades do desvio padrão**

- O valor do desvio padrão é sempre positivo (nunca negativo).
- O valor do desvio padrão aumenta dramaticamente com a inclusão de um ou mais valores extremos (outliers).
- A unidade do desvio padrão é a mesma unidade dos dados originais.

Z-escore e a padronização

O Z-escore é uma forma de padronizar os dados, de modo que cada observação "perca" a sua unidade original e seja transformada em desvios-padrão. Para isso, basta subtrair cada valor da média e dividir pelo desvio padrão da distribuição.

onde, é a observação, a média e é o desvio padrão.

Observe que um z-score negativo indica que o valor está abaixo da média, enquanto um z-score

positivo mostra que o valor está acima da média.

Uso do z-score apara padronizar resultados do pentatlo

Dois atletas (A e B) competem em uma corrida de 800 metros, cuja média de todos os cometidores foi de 137 segundos, com desvio padrão de 5 segundos. O corredor A completou a corrida em 129 segundos, enquanto o corredor B gastou 140 segundos. Qula é o z-escore de cada um?

- O corredor A:

- O corredor B:

Os desempenho dos atletas A e B no salto a distância. A média é de 6 metros e desvio-padrão de 30 cm. Qual é o z-score do atlea A que saltou 6.60cm e do atleta B que saltou 5.84 cm.

- O saltador A:

- O saltador B:

Desse modo, é possivel padronizar (transformar em desvios-padrão em relação à média) valores de diferentes distribuições.

Exercício

- Um aluno de economia tirou nota 7 em cálculo (média 5 e desvio padrão de 3).
- Um aluno de ciências sociais tirou nota 9 em teoria antropológica (média 8 e desvio padrão de 1).

Quem obteve um melhor desempenho relativo?

```
• Enter cell code...
```

💡Benefícios da padronização

- Os valores padronizados são convertidos das unidades originais para a unidade estatística de desvio padrão da média.
- assim, podemos comparar valores que são medidos em diferentes escalas, com diferentes unidades e extraídos de diferentes populações.

Z-escore e a curva normal

Imagine que nós empilhassemos as observações de uma determinada população (por exemplo, as notas de uma turma, ou a altura das pessoas que compareceram ao último Fla-Flu) já padronizadas. Esse empilhamento das observação produziria um gráfico semelhante a uma curva normal,que tem o formato de um sino.

A curva normal é a distribuição mais conhecida da estatística e tem uma propriedade: sabemos quantos casos estão abaixo de cada segmento da curva, quando observamos os desvios-padrão.

- 68,2 % dos casos estão a 1 desvio-padrão em relação à média
- 95,4% dos casos estão a 2 desvios-padrão em relação à média
- 99.7% dos casos estão a 3 desvios-padrão em relação à média

Curva normal e a regra dos desvios

Se a população de casos, se distribui em um formato de uma curva normal, e sabemos o z-score de uma observação, é possivel conhecer em que percentil da distribuição ele está.

O saltador A do exemplo do exemplo anterior está a 2 desvios-padrão em relação à média. No gráfico acima podemos observar que a faixa vermelha contempla 2,2 % (2,1% + 0,1 %) dos casos; ou seja, o saltador A está entre os top 97,8 % da distibuição.

Os livros antigos de estatística traziam uma tabela em anexo, onde era possivel fazer a conversão entre o z-score eo percential de um caso.

⚠️ **A Regra dos dois desvio de Leo Monastério**

Nunca brigue se o adversário estiver a mais de dois desvios padrão de você em qualquer dimensão: conhecimento, ideologia, inteligência ou porte físico