Análise Exploratória de Dados - Avaliação Presencial 02 (Gabarito) - 2024/01

Prof. Hugo Carvalho

02/07/2024

Questão 1:

a) O coeficiente de variação é uma quantidade que visa comparar o desvio padrão com a média, resultando em um número que nos permite interpretar "quão grande é um desvio padrão em relação à determinado valor de média". Dessa forma, ele pode ser interpretado também como uma medida de dispersão.

b) Tanto a média quanto o desvio padrão são medidos na mesma unidade, que será a unidade na qual o conjunto de dados x é apresentado. Portanto, ao fazer o quociente entre essas duas quantidades, elas se "cancelarão", resultando em uma quantidade adimensional.

c) Algumas vantagens do coeficiente de variação são: pelo fato de ser adimensional, sua comparação é simplificada, por não depender de unidades de medida; ademais, a comparação entre o coeficiente de variação de dois conjuntos de dados distintos (medidos potencialmende em unidades distintas) faz sentido; finalmente, pode-se apontar que o próprio fato de explicar uma variabilidade comparando-a naturalmente com uma medida de centralidade é também uma vantagem de tal métrica. Algumas desvantages, por outro lado, são: sensibilidade a valores discrepantes, já que a média e o desvio padrão o são; dificuldade de interpretação à medida que a média se aproxima de zero (dados centrados).

d) Nesse caso, o coeficiente de variação é dado por $500/100.000 = 0{,}005$, um valor bastante pequeno. Dessa forma, é possível concluir que a estimativa é relativamente acurada. Note que 500m pode parecer uma variabilidade muito grande, sem a informação da média. O coeficiente de variação faz adequadamente essa comparação.

e) No primeiro caso, o coeficiente de variação é $2/5 = 0{,}4$, indicando uma variabilidade moderada em torno da média; no segundo caso, o coeficiente de variação é igual a 2, indicando uma alta variabilidade em torno da média; no terceiro caso o coeficiente de variação não pode ser calculado, pois não pode-se dividir por zero; finalmente, no último caso, o coeficiente de variação será negativo, igual à -1, indicando que a média é negativa e que a variabilidade dos dados em torno da média é significativa.

Questão 2: A Figura 1 resulta da coleta de dados de um usuário da rede social Reddit, ao longo de um ano, quantificando seis sentimentos distintos: deprimido, ansioso, culpado, exausto, feliz e satisfeito. Nota-se que o seu diagnóstico de depressão crônica reflete na alta incidência dos sentimentos negativos (deprimido, ansioso, culptado e exausto), e na baixa incidência dos sentimentos positivos (feliz e satisfeito). De fato, percebe-se que em nenhum momento o usuário pontuou com nota máxima os sentimentos de feliz e satisfeito, mas eles foram pontuados diversas vezes com zero, tanto que tal valor sequer chega a ser um valor discrepante. Por outro lado, os sentimentos negativos nunca chegam a zero, mas atingem sempre pontuação máxima. Em particular o sentimento de exaustão atinge o valor máximo sem que ele seja um valor discrepante, indicando o alto desgaste mental que uma enfermidade desse porte traz para a pessoa. Pontos positivos do gráfico são que ele transmite de forma clara informações referentes a uma doença perigosa, e poderia servir como uma ferramenta educativa para explicar melhor o impacto da depressão crônica na vida de uma pessoa. Pontos negativos são: o uso das cores não é claro (seja na versão colorida ou na versão em tons de cinza), e poderia ser evitado; em alguns *box-plots* os valores dos quantis coincidem com a mediana, o que torna o gráfico de mais difícil leitura; finalmente, em alguns casos há também valores discrepantes coincidindo com o limite superior, algo que não deveria ocorrer, indicando que na confecção do gráfico houve algum tipo de arredondamento nos valores.