2.3.2. Uma forma sistemática de fazer AED

Para organizar a análise exploratória de um conjunto de dados, uma boa idéia é fazer uma lista de perguntas sobre as características mais importantes de uma distribuição e depois procurar respondê-las, estudando os gráficos e medidas. Estas características podem ser divididas em três grupos:

1) a forma da distribuição
2) sua posição, dispersão, e valores discrepantes
3) sua granularidade

Discutiremos a seguir estes três grupos.

2.3.2.1. A forma da distribuição

O objetivo principal da maioria dos gráficos que vimos (como o de *pontos*, o de *ramo-e-folhas* e o *histograma*) é mostrar a *distribuição* dos dados, isto é, como os valores da variável estão distribuídos ao longo das escalas. A forma da distribuição na amostra nos sugere como deve ser a forma da distribuição na população (o que será depois útil na *Inferência*); além disso, pode fornecer indicíos que permitam avaliar a qualidade dos dados, e identificar os erros os que existirem neles.

Feito um gráfico, é preciso examinar primeiro a forma geral, e depois os detalhes. Para organizar a análise pode ser útil fazer uma série de perguntas sobre a forma da distribuição, como por exemplo:

- *Quantos aglomerados há? Apenas um, ou vários aglomerados separados?*
- *Se há apenas um aglomerado, quantas modas ele tem?*
- *A distribuição neste aglomerado é simétrica ou assimétrica?*

A grande maioria das variáveis têm distribuições com apenas um aglomerado, e apenas uma moda; se encontramos na amostra distribuições com mais de um aglomerado, ou mais de uma moda, isto pode ser um sinal de que algo saiu errado (um exemplo está na Fig. **3** da seção **2.1.1**, onde dados de homens e de mulheres foram misturados num mesmo gráfico, o que resultou em duas modas). A maioria das ferramentas estatísticas mais usadas foram desenvolvidas para lidar com esta forma de distribuição; quando uma distribuição é bimodal, por exemplo, mesmo uma medida tão simples quanto a *média aritmética* pode não ser muito útil. Na Inferência, várias técnicas (por exemplo os testes t, ANOVA, modelos de regressão) exigem que a distribuição da variável seja unimodal e simétrica, ou pelo menos aproximadamente simétrica (na verdade, estes métodos existem que a distribuição seja *gaussiana*, o que é um tipo de distribuição simétrica que veremos na seção **3.4.4)**.

2.3.2.2. Posição, dispersão, valores discrepantes

Procuramos aqui responder a perguntas como:

- *Em que faixa de valores se encontra a maioria dos dados ? Isto é plausível?*
- *Quais são os pontos máximos e mínimos?*
- *Há pontos discrepantes ? Estes valores são plausíveis?*
- *O que pode ter causado estas discrepâncias ?*

A *localização* e a *dispersão* de uma distribuição podem ser estimadas visualmente a partir dos gráficos, e depois confirmadas por medidas numéricas, escolhidas de acordo com a forma da distribuição e o objetivo do estudo. Para distribuições muito assimétricas, por exemplo, pode ser melhor usar a mediana e o intervalo quartílico, ao invés da média e da variância.

Os pontos discrepantes devem ser localizados *antes* de fazermos qualquer análise estatística, por duas razões. Primeiro, porque um valor discrepante pode indicar que houve um erro de medição ou de registro (isto é, alguém mediu errado, ou anotou errado o valor). Segundo, porque a maioria das ferramentas estatísticas clássicas (por exemplo, a variância e o coeficiente de correlação) são muito afetadas por valores discrepantes, e podem levar a resultados inteiramente distorcidos.

Quando o valor discrepante está obviamente errado e a amostra é grande, podemos simplesmente descartá-lo. Em alguns casos, o valor é claramente um erro de medição; em outros, pode indicar que houve um erro de registro. A decisão entre aceitar ou rejeitar o ponto discrepante, contudo, é sempre difícil, pois pode afetar o resultado do experimento. Às vezes, o ponto discrepante é justamente o que tem a informação mais importante da base de dados. Um exemplo disto ocorreu durante a análise da camada de ozônio sobre a Terra, acima do Pólo Sul: o primeiro artigo alertando sobre o buraco na camada foi publicado em 1985, mas dados que indicavam sua existência já tinham obtidos por satélites em 1979. Estes dados, no entanto, tinham sido excluídos automaticamente pelos programas de computador que faziam as análises, já que eram muito menores que os valores mínimos que se imaginavam possíveis, e tinham sido rejeitados como erros de medição [1].

2.3.2.3. Granularidade (ou resolução)

Perguntas a responder:

- *Qual é a granularidade (resolução) dos dados?*
- *Esta granularidade era esperada, ou usual para este tipo de dados?*

A *granularidade* (ou *resolução*) indica o nível de detalhe com que os dados foram medidos ou registrados. Quanto maior o nível de detalhe, menor a granularidade e maior a resolução da variável. Quando uma variável resulta de *medições*, ela é em geral uma variável contínua, ao contrário das variáveis resultantes de *contagens*, que são discretas. Variáveis contínuas, teoricamente, poderiam ser medidas com precisão infinita (um número ilimitado de dígitos depois da vírgula); na prática porém, todo valor é arredondado. Isto pode ser exigido pelo instrumento de medida (nenhum deles tem precisão infinita), ou pela utilização que se pretende fazer dos dados; por exemplo, o peso de pessoas são geralmente da-

dos em quilogramas. Seria tecnicamente possível pesar alguém com a precisão de gramas, ou de frações de gramas, mas esta precisão seria inútil nas aplicações práticas, já que o peso de uma pessoa varia continuamente durante o dia. A granularidade pode também ser determinada pela unidade de medida. As alturas de pessoas são medidas na maioria dos países em centímetros; nos EUA, porém, são medidas em polegadas (uma polegada equivale a 2,54 cm), e portanto o resultado é menos preciso, e a granularidade dos dados é maior. Numa variável que resulta de *contagens*, a granularidade pode ser definida pela maneira como a contagem é feita. Por exemplo, se analisamos um conjunto de dados sobre o pulso em repouso dos pacientes de uma amostra e notamos que todos os valores registrados são múltiplos de quatro, podemos inferir que quem coletou os dados contou as pulsações de cada paciente por 15 segundos e depois multiplicou os resultados por 4. Esta técnica torna a coleta de dados mais rápida, mas evidentemente diminui a resolução dos valores; a observação da granularidade às vezes nos permite, portanto, inferir ou confirmar a forma de coleta dos dados.

Moore [2] cita o exemplo de um artigo de revisão publicado que supostamente analisava "um conjunto de 20 estudos, com 57 por cento reportando resultados significativos, dos quais 42 % concordam em uma conclusão, enquanto os 15 % remanescentes favorecem outra conclusão, freqüentemente a conclusão oposta." O autor observa que esta frase não faz sentido, pois se há 20 artigos, os valores possíveis das porcentagens estarão necessariamente espaçados em intervalos de 5 por cento; é possível falarmos em "55 % dos 20 estudos" (11 estudos) ou "60 % dos 20 estudos" (12 estudos); mas não de "57 % dos estudos" (11,4 estudos?!). A frase portanto não faz sentido, pois o resultado mencionado não é coerente com a granularidade esperada.

Nota: Os termos "granularidade" e "resolução" vêm da fotografia analógica, onde os *grãos* eram os pontos de pigmento na superfície da foto; os termos passaram a ser usados também para designar a quantidade de *pixels* usados nas imagens digitais. A Fig. 1 mostra uma mesma foto, com níveis crescentes de granularidade (e níveis decrescentes de resolução).

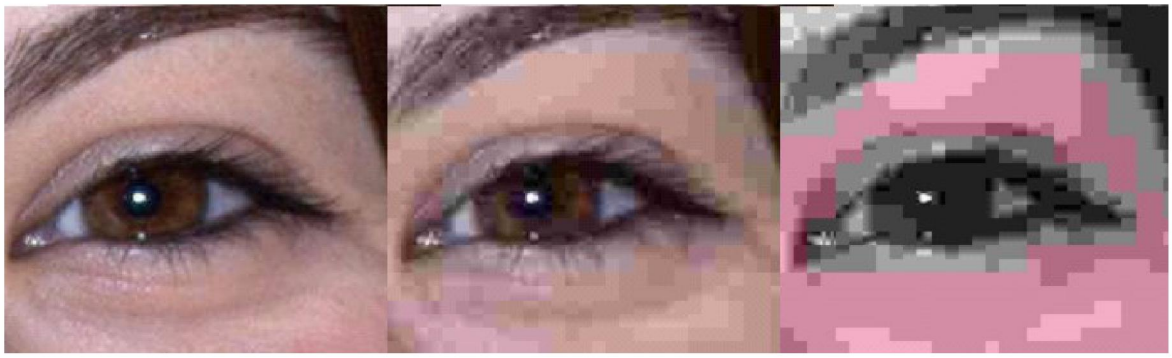

Figura 1. Três versões de uma fotografia, com diferentes granularidades

---

Referências

[1] W. Burroughs (ed.). *Climate into the 21st century*. Cambridge University Press, 2003
[2] Moore, D.S. *Statistics – Concepts and Controversies.* 3rd. ed. New York: W. H. Freeman & Co., 1991.