Por que a fórmula para a variância amostral tem $n-1$ no seu denominador

Suponhamos que alguém colocou bolas idênticas numa urna, sendo que cada bola carrega um dos números $k_1, k_2, \ldots, k_M$. Denotaremos por p_i a proporção das bolas da urna que carregam o número k_i. Com isso, a seguinte tabela

k_1	k_2	$\cdots\cdots$	k_{M-1}	k_M
p_1	p_2	$\cdots\cdots$	p_{M-1}	p_M

é exatamente aquilo ao que chamamos por "distribuição de frequencia relativa por atributo "numero carregado" da população de bolas na urna". Esse distribuição será chamada no que se segue por "distribuição populacional".

Recordo, para as necessidades dos argumentos a vir, que a quantia

$$k_1p_1 + k_2p_2 + \cdots + k_{M-1}p_{M-1} + k_Mp_M \tag{1}$$

calculada com valores da distribuição populacional, chama-se *média populacional*; denotaremos essa por μ.

Outrossim recordo, também para as necessidades futuras, que a quantia

$$(k_1-\mu)^2p_1 + (k_2-\mu)^2p_2 + \cdots + (k_{M-1}-\mu)^2p_{M-1} + (k_M-\mu)^2p_M \tag{2}$$

chama-se *variância populacional*; denotaremos essa por σ^2.

Imagine agora que não sabemos nada sobre a distribuição populacional: não sabemos nem M, nenhum dos k_i e nenhum dos p_i.

Imagine que desejamos estimar o valor de σ^2, quer dizer, desejamos estimar a variânica populacional, e para o fim dessa estimação, somos autorizados fazer amostra simples com reposição. Isso quer dizer, que podemos retirar ao acaso uma bola da urna e observar o número carregada pela bola retirada, e podemos repetir esse procedimento tantas vezes quantas desejamos, desde que após cada ratirada, a bola seja devolvida à urna, e todas as bolas sejam bem misturadas.

Então, ao fixar o número de retiradas a serem feitas e ao dentor esse por n (esse n é o que chama-se *o tamanho da amostra* a ser feita), sugerimos o seguinte procedimento: após fazer n retiradas e ver os números

$$x_1, x_2, \ldots, x_n \tag{3}$$

fazer a média delas, isso é,

$$\frac{1}{n}\left\{x_1 + x_2 + \cdots + x_n\right\} \tag{4}$$

e usar esse para calcular o valor da expressão

$$\frac{1}{n-1}\left\{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_n-\bar{x})^2\right\} \tag{5}$$

onde $\bar{x}$ é a notação para a média. O valor da expressão (5) é a nossa estimativa para a variância populacional.

O presente texto é sobre o porque o denominador da expressão (5) é $n-1$ e não n, como seria de se esperar com base em considerações intuitivas e genéricas.

Para que possamos dar a a explicação ao $n-1$, precisamos olhar na amostra antes dela acontecer. Desse ponto de vista,

o número a ser visto na 1-a retirada é uma variável aleatória, a qual denotamos por X_1,
o número a ser visto na 2-a retirada é uma variável aleatória, a qual denotamos por X_2,
$\vdots$ $\vdots$ $\vdots$ $\vdots$ $\vdots$ $\vdots$
o número a ser visto na n-ésima retirada é uma variável aleatória, a qual denotamos por X_n

E então, a média (4) adquira a expressão

$$\frac{1}{n}\left\{X_1 + X_2 + \cdots + X_n\right\} \tag{6}$$

Fica claro que isso é uma variávela aleatória; denotamos ela por $\bar{X}$. Ainda mais, do nosso ponto de vista de antes, a expressão (5) adquira a expressão

$$\frac{1}{n-1}\left\{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2\right\} \tag{7}$$

Para justificar que a fórmula sugerida para estimar σ^2 é boa, vou mostrar que

$$I\!E\left[\frac{1}{n-1}\left\{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2\right\}\right] = \sigma^2 \tag{8}$$

o que interpreta-se dizendo que

$$\text{o estimador acerta em média no valor estimado} \tag{9}$$

Não pretendo detalhar a relação dessa interpretação com a fórmula à qual ela se relaciona[1].

Para completarmos o objetivo de mostrar a igualdade (8), precisamos dos seguintes fatos:

Fato 1: A distribuição de cada variável aleatória X_i é exatamente a distribuição populacional, quer dizer,

x	k_1	k_2	$\cdots\cdots$	k_{M-1}	k_M
$I\!P[X_i = x]$	p_1	p_2	$\cdots\cdots$	p_{M-1}	p_M

e, como consequencia disso,

[1]Na verdade, tal detalhamento foi feito na sala de aula.

Fato 2:

$$\mathbb{E}[X_i] = \mu, \qquad \text{Var}[X_i] = \sigma^2, \text{ para cada } X_i \tag{10}$$

Fato 3: As variávei aleatórias $X_1, X_2, \ldots, X_n$ são independentes em conjunto.

Confere que você entende bem os Fatos 1-3: O primeiro deles segue-se da definição de experimento aleatório, da construção de modelo probabilístico e da definição de variável aleatória como a expressão do resultado a vir num experimento aleatório. É importante entender que a suposição, segundo a qual não conhecemos nada sobre a distribuição populacional, não impede da mesma ser a distribuição de cada variável aleatória que definimos; é só não conhecemos a distribuição. O comentário do Fato 2 vem no mesmo sentido: apesar de não conhecer os valores da média populacional e da variância populacional, sabemos que esses valores coincidem com, respectivamente, a esperança e a variância de cada uma das variáveis aleatórias que definimos. O Fato 3 segue-se de um comentário que fiz quando expliquei o conceito de independencia entre variáveis aleatórias. É suficiente que você acredite na independência. Creio que isso não é difícil a ser concebido pois as bolas retiradas são devolvidas na urna, e, portanto, o resultado de i-ésima retirada não pode dependenr dos resultados das retiradas anteriores.

Vamos agora introduzir notações que ajudarão a simplificar a escrita da demostração que pretendemos fazer. Em primeiro lugar, observe que

$$X_i - \bar{X} = (X_i - \mu) - (\bar{X} - \mu) = (X_i - \mu) \; - \; \frac{(X_1 - \mu) + (X_2 - \mu) + \ldots + (X_n - \mu)}{n}$$

Por isso, ao introduzir

$$Y_i = X_i - \mu, \text{ para cada } i$$

podemos reescrever nosso objetivo da seguinte maneira:

$$\text{provar que } \mathbb{E}\left[\frac{1}{n-1}\left\{(Y_1 - \bar{Y})^2 + \cdots + (Y_n - \bar{Y})^2\right\}\right] = \sigma^2 \tag{11}$$

onde

$$Y_1, Y_2, \ldots, Y_n \text{ são variáveis aleatórias independentes em conjunto,} \tag{12}$$

$$\mathbb{E}[Y_i] = 0 \text{ e } \text{Var}[Y_i] = \sigma^2 \text{ para cada } i, \tag{13}$$

$$\bar{Y} \text{ é a notação para } \frac{1}{n}\left\{Y_1 + Y_2 + \cdots + Y_n\right\} \tag{14}$$

sendo que (12) é a consequencia do Fato 3, e (13) é a consequencia do Fato 2; óbvio, que poderiamos desenhar a distribuição de cada Y_i a partir da distribuição das X's, mas não fizeram isso pois estas não serão importantes para as contas a seguir.

O segundo passo na nossa presente derivação de fatos e notações auxiliares é o fato de que

$$\mathbb{E}[Y_i^2] = \sigma^2 \tag{15}$$

Esse segue-se da expanção $\text{Var}[Y_i] = I\!E[Y_i^2] - (I\!E[Y_i])^2$ junto com as duas relações da Eq. (13). Já no terceiro passo deduzimos que

$$I\!E[Y_i Y_j] = I\!E[Y_i]\, I\!E[Y_j] = 0 \times 0 = 0, \text{ para todos } i \text{ e } j \text{ diferentes entre si} \tag{16}$$

Observe que a primeira igualdade da sequencia (16) sustenta-se pela independência entre Y_i e Y_j, a qual, por sua vez, segue-se da independência das Y's em conjunto. Essa conta é o único lugar onde usamos a independência, mas o resultado da conta é a chave principal para tudo, de modo que a ausência da independência quebraria a conta e, em sequencia, toda a demonstração.

Agora, juntamos (15) e (16) para derivar que

$$\begin{aligned}
&I\!E\left[(Y_1 + Y_2 + \ldots + Y_n)^2\right] = I\!E\left[Y_1^2 + Y_2^2 + \ldots + Y_n^2 + 2Y_1Y_2 + \ldots + 2Y_{n-1}Y_n\right] = \\
&= \left(I\!E[Y_1^2] + \ldots + I\!E[Y_n^2]\right) \; + \; (2I\!E[Y_1Y_2] + \ldots + I\!E[Y_{n-1}Y_n]) = \\
&= (\sigma^2 + \ldots + \sigma^2) \; + \; (0 + \ldots + 0) = n\sigma^2
\end{aligned}$$

Esse resultado será usado para fechar a demonstração em conjunto com o resultado da seguinte conta, puramente algébrica:

$$\begin{aligned}
&\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = \\
&= \left(Y_1^2 - 2Y_1 \times \bar{Y} + \left(\bar{Y}\right)^2\right) + \ldots + \left(Y_n^2 - 2Y_n \times \bar{Y} + \left(\bar{Y}\right)^2\right) = \\
&= \left(Y_1^2 + \ldots + Y_n^2\right) \; - 2n\bar{Y} \times \bar{Y} \; + \; n\left(\bar{Y}\right)^2 = \\
&= \left(Y_1^2 + \ldots + Y_n^2\right) \; - \; n\left(\bar{Y}\right)^2 \; = \; \left(Y_1^2 + \ldots + Y_n^2\right) \; - \; \frac{1}{n}\left(Y_1 + \ldots + Y_n\right)^2
\end{aligned}$$

Portanto,

$$\begin{aligned}
&I\!E\left[\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2\right] = I\!E\left[Y_1^2 + \ldots + Y_n^2\right] \; - \; \frac{1}{n} I\!E\left[(Y_1 + \ldots + Y_n)^2\right] = \\
&= I\!E\left[Y_1^2\right] + \ldots + I\!E\left[Y_n^2\right] \; - \; \frac{1}{n}\left(n\sigma^2\right) = \\
&= n\sigma^2 - \sigma^2 = (n-1)\sigma^2
\end{aligned}$$

Agora ficou claro que (já passando de volta para as variáveis aleatórias X's)

$$I\!E\left[\frac{1}{n-1}\left\{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2\right\}\right] = \sigma^2 \tag{17}$$

enquanto que, se tomassemos n e vez de $n-1$ no denominador, teriamos

$$I\!E\left[\frac{1}{n}\left\{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2\right\}\right] = \frac{n-1}{n}\sigma^2 \tag{18}$$