Estimating from sums of samples of unequal size

MIGUEL A. S. CASQUILHO and JOÃO A. BRANCO
IST, Universidade de Lisboa, 1049-001 Lisboa, Portugal

Sums are very frequent in the industrial activities, e.g., in packages or bags or packets of items. These are routinely measured for confirmation due to commercial reasons, thus giving rise to numerous data. So, it would be advantageous to be able to estimate the distribution parameters: in the typical case of Gaussian behaviour, the mean and standard deviation of the individual items from their sums. The main difficulty stems from the fact that the sums may come from samples of different or even unknown sizes. This study identifies the various possibilities, from known and equal sample sizes to unknown and different sample sizes, and solves them or suggests solutions.

Keywords: *sample sums, estimation, Monte Carlo.*

1. Fundamentals and scope

Numerosas actividades industriais produzem artigos que são vendidos em conjuntos por motivos operacionais ou comerciais, sendo importante o peso de cada conjunto. No entanto, o peso de cada artigo individual, w, é tipicamente uma característica decisiva, que deve respeitar os limites de especificação inferior e superior que lhe são impostos, limites estes de natureza quer legal quer operacional. Assim, para que os artigos respeitem as especificações, convirá aproveitar toda a informação disponível para estimar os parâmetros da distribuição de w, isto é, a média e o desvio-padrão no caso Gaussiano, que se considera neste estudo.

O peso de cada artigo é monitorizado (*on line*) no processo de fabrico e também habitualmente (*off line*) por amostragem, para cumprir objectivos legais, comerciais e de satisfação do cliente, mas a pesagem de conjuntos é muito mais fácil. Para além do peso individual, w, imposto na produção e em princípio estatisticamente controlado durante o processo, constata-se que, em muitos casos, uma informação suplementar é obtida, o peso, T, de cada conjunto de n artigos, cada um com peso w_i, não medido,

$$T = \sum_{i=1}^{n} w_i \qquad \{1\}$$

Assume-se neste estudo que T e n de cada conjunto são conhecidos, mas n é constante ou variável. A pesagem do conjunto destina-se a verificar a quantidade vendida, mas faculta acessoriamente um substancial acervo de dados de somas dos pesos. Embora estes dados sejam habitualmente apenas usados para a referida verificação, a sua exploração poderá contribuir para a estimação dos parâmetros da distribuição. Quando os conjuntos têm a mesma dimensão (n constante), a estimação dos parâmetros é trivial, mas já assim não é quando os conjuntos têm diferentes dimensões (n variável). Dos dois casos, adiante clarificados, a análise do segundo é o objectivo deste estudo.

O primeiro caso, de conjuntos com n constante, pode-se encontrar em muitas indústrias. No ramo alimentar, são exemplos: embalagem com uma dúzia de ovos duma certa classe de tamanho; caixas com iogurtes, com garrafas de bebida, de óleo alimentar, conservas. Nas indústrias de tintas, colas, detergentes, são exemplos: caixas ou paletes com latas de certa capacidade nominal. Noutras indústrias, como as da

Dr. Casquilho is an Assistant Professor (retired) in the Department of Chemical Engineering at Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal. His email address is mcasquilho@tecnico.ulisboa.pt.

Dr. Branco is an Associate Professor (retired) in the Department of Mathematics at Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal. His email address is jbranco@math.tecnico.ulisboa.pt.

construção civil, os pesos individuais não são cruciais, mas, por motivos operacionais, podem sê-lo os dos conjuntos (palete com tijolos).

O segundo caso, de conjuntos com *n* variável, aparece, por exemplo, na indústria química. Um caso típico (observado por um dos autores na sua passagem por esse sector), que motiva o presente estudo, é a expedição de camiões carregados com sacos de adubo de peso 50 kg [ADP, 2013] em número variável, dependendo das encomendas dos clientes. Neste tipo de indústria, um camião é sempre pesado (à entrada e à saída) para verificar se a diferença de peso da carga corresponde à encomenda, tentando-se também indirectamente confirmar o número de sacos, por vezes centenas, susceptível de erro de contagem.

Além do exemplo dos sacos de adubo, encontrou-se, na indústria alimentar, um exemplo homólogo, que se adopta neste estudo por simplicidade, duma embalagem de bolachas com as seguintes características nominais: peso de 800 g; subdivisão em 4 pacotes (de cerca de 200 g); e cerca de 36 bolachas por pacote. Observou-se que o número de artigos (bolachas) em cada um dos 4 pacotes apresentava diferenças, o que torna este problema similar ao das cargas de sacos de adubo. Tanto nos conjuntos de sacos de adubo como nos de bolachas, assume-se que (Eq. {1}): os pesos individuais, w_i, não são conhecidos (a não ser por prévia amostragem geral); e para cada conjunto, são conhecidos T e o respectivo valor de n.

Pretende-se, no que se segue, estimar os parâmetros da distribuição de Gauss dos artigos individuais. Após uma breve digressão pelo caso de *n* constante, observa-se o caso de *n* variável, finalizando com algumas conclusões e conjecturas baseadas em simulação via Monte Carlo.

2. Conjuntos com *n* constante

Em conjuntos com *n* constante (comum aos vários conjuntos) ou em conjuntos com *n* variável, quando os pesos individuais são conhecidos, a estimação dos parâmetros, μ e σ, torna-se um problema trivial. O conjunto está representado na Figure 1, onde simbolicamente os artigos soltos e acessíveis pretendem indicar que cada peso é conhecido.

FIGURE 1. Pesos conhecidos.

De facto, será simplesmente

$$\hat{\mu} = \overline{X} \qquad \hat{\sigma} = S \tag{2}$$

em que a média $\overline{X}$ e o desvio-padrão são calculados sobre a totalidade dos valores, todos conhecidos.

Quando, nos referidos conjuntos (*n* constante), apenas as somas são conhecidas, mas não os valores individuais, o problema pode ainda ser reduzido ao caso anterior (Eq. {2}). Os conjuntos de artigos estão representados na Figure 2, onde os pacotes (fechados) indicam que cada peso não está acessível e não é conhecido; contudo, os valores de n são os mesmos.

FIGURE 2. Somas conhecidas (pesos individuais não conhecidos).

As somas, T, são os dados experimentais obtidos. É, evidentemente,

$$\mu(T) = n\mu \qquad \sigma(T) = n\sigma \tag{3}$$

No entanto, as correspondentes médias, $\overline{X} = T/n$, são de tratamento mais intuitivo, pelo que será

$$\mu(\overline{X}) = \mu \qquad \sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}} \tag{4}$$

Haverá, simplesmente, neste caso de n constante, uma amostra duma certa dimensão, seja m, de valores de $\overline{X}$, à qual se aplicará a estimação habitual, conforme Eq. {2}, dando, atendendo a Eq. {4},

$$\hat{\mu} = \mu(\overline{X}) \qquad \hat{\sigma} = \sigma(\overline{X})\sqrt{n} \tag{5}$$

Em resumo, com n constante e conhecendo apenas somas, tudo o que há a fazer é considerar as m médias como se fossem valores individuais e proceder à estimação básica.

Na Table 1, para ilustração numérica, inclui-se um exemplo com 4 amostras de dimensão 10, com dados simulados duma distribuição Gaussiana reduzida.

Table 1 - Pesos individuais simulados

i	w_{i1}	w_{i2}	w_{i3}	w_{i4}
1	-0,0615	-1,0816	0,5299	-0,3250
2	0,2342	-0,2763	1,1235	-1,0183
3	0,2200	-1,3308	-1,0680	1,0717
4	-0,4650	-0,9999	0,8644	0,0076
5	0,1962	-0,6997	0,2796	1,1208
6	1,2753	0,7537	-0,5438	0,7411
7	-0,0949	0,2901	-0,4854	1,0709
8	-1,6252	0,2138	1,8794	0,8228
9	1,7439	1,0900	-1,1481	1,9923
10	-0,7012	0,5114	-0,3144	0,0621

Considerando que os dados (Table 1) são conhecidos, é simplesmente

$$\hat{\mu} = \overline{X} = 0{,}1464 \qquad \hat{\sigma} = S = 0{,}9178 \qquad \{6\}$$

(notando que estes resultados provêm de números com mais algarismos significativos do que os da Table 1).

Considerando agora que os dados individuais (Table 1) não são conhecidos, mas apenas as 4 somas, será, na Table 2, sucessivamente,

Table 2 – Somas e médias das várias amostras

	Amostra 1	Amostra 2	Amostra 3	Amostra 4
Soma	0,7217	-1,5293	1,1171	5,5460
Média	0,0722	-0,1529	0,1117	0,5546

dando $\sigma\left(\overline{X}\right) =$ 0,2961 e, com n = 10,

$$\hat{\mu} = \overline{X} = 0{,}1464 \qquad \hat{\sigma} = \hat{\sigma}\sqrt{n} = 0{,}9362 \qquad \{7\}$$

em que $\hat{\mu} = \overline{X} = 0{,}1464$ é naturalmente idêntico ao valor da Eq. {6}. Recorde-se que os dados provêm duma distribuição Gaussiana reduzida, pelo que as estimativas devem ser comparadas, respectivamente, com 0 e 1.

3. Conjuntos com *n* variável

###Reescrito até aqui. Falta:
Traduzir, etc., etc.