The Mean Value of a Sample

Klaus Pommerening

September 2020 – last change September 17, 2020

Suppose we take a sample A of $n = \#A$ elements from the "population" $\{1, \ldots, N\}$. We ask for the distribution of the mean value of this sample. In other words, we consider all n-element subsets $A \subseteq \{1, \ldots, N\}$ the sums of whose elements we denote by

$$\Sigma(A) = \sum_{a \in A} a.$$

The mean value of A is $\Sigma(A)/n$, and we ask for the distribution of all these "sample" mean values, in particular the variance—intuitively obvious is the expectation of this distribution, namely identical with the mean value $(N+1)/2$ of $\{1, \ldots, N\}$.

The frequency function

$$\begin{aligned} F_{N,n} \colon \mathbb{Z} &\longrightarrow \mathbb{Z}, \\ F_{N,n}(x) &= \#\{A \subseteq \{1, \ldots, N\} \mid \#A = n,\ \Sigma(A) = x\}, \end{aligned}$$

indirectly describes the distribution of the mean values by the formula

$$p_{N,n}(t) = \frac{1}{\binom{N}{n}} F_{N,n}(nt) \quad \text{for } t \in \mathbb{R}. \tag{1}$$

This is the fraction of all $\binom{N}{n}$ subsets $A \subseteq \{1, \ldots, N\}$ with n elements that have sum nt, or have mean value $t = \Sigma(A)/n$.

Example For $N = 4$ and $n = 2$ the subsets A are

$$\begin{aligned} A &= \{1,2\} \quad \text{with} \quad \Sigma(A) = 3, \\ A &= \{1,3\} \quad \text{with} \quad \Sigma(A) = 4, \\ A &= \{1,4\} \quad \text{with} \quad \Sigma(A) = 5, \\ A &= \{2,3\} \quad \text{with} \quad \Sigma(A) = 5, \\ A &= \{2,4\} \quad \text{with} \quad \Sigma(A) = 6, \\ A &= \{3,4\} \quad \text{with} \quad \Sigma(A) = 7. \end{aligned}$$

Hence the frequency function is

$$F_{4,2}(x) = \begin{cases} 2 & \text{for } x = 5, \\ 1 & \text{for } x = 3, 4, 6, 7, \\ 0 & \text{otherwise.} \end{cases}$$

There are some obvious formulas:

Lemma 1 *Let $N \in \mathbb{N}$. Then:*

(i) $F_{N,n}(x) = 0$ *constant if* $n > N$.

(ii) $F_{N,0}(x) = 1$ *for* $x = 0$, *and* $F_{N,0}(x) = 0$ *for* $x \neq 0$.

(iii) $F_{N,1}(x) = 1$ *for* $1 \leq x \leq N$, *and* $F_{N,1}(x) = 0$ *for* $x \leq 0$ *or* $x \geq N + 1$.

(iv) (*Symmetry*) $F(x) = F_{N,n}(n\,(N + 1) - x)$ *for all* $x \in \mathbb{Z}$.

(v) $F_{N,2}(x) = \lfloor \frac{x-1}{2} \rfloor$ *for* $3 \leq x \leq N + 1$,
$F_{N,2}(x) = \lfloor \frac{2N+1-x}{2} \rfloor$ *for* $N + 1 \leq x \leq 2N - 1$,
$F_{N,2}(x) = 0$ *otherwise.*

(vi) (*Recursion 1*) $F_{N,n}(x) = F_{N-1,n}(x) + F_{N-1,n-1}(x - N)$ *for* $1 \leq n < N$.

(vii) (*Recursion 2*) $F_{N,n}(x) = F_{N-1,n}(x - n) + F_{N-1,n-1}(x - n)$ *for* $1 \leq n < N$.

Proof. (i) There are no subsets with n elements.
(ii) The empty set has sum 0.
(iii) For a one-element subset A we have $\Sigma(A) = x$ if and only if $A = \{x\}$.
(iv) Consider the bijective map

$$\varphi \colon \{1, \ldots, N\} \longrightarrow \{1, \ldots, N\}, \quad a \mapsto N + 1 - a,$$

that reverses the order of $\{1, \ldots, N\}$. It induces a bijection

$$\Phi \colon \mathcal{P}(\{1, \ldots, N\}) \longrightarrow \mathcal{P}(\{1, \ldots, N\})$$

on the power set via the assignment $\Phi(\{a_1, \ldots, a_n\}) = \{\varphi a_1, \ldots, \varphi a_n\}$. Then

$$\Sigma(\Phi(A)) = \sum_{a \in A} \varphi a = \sum_{a \in A} N + 1 - a = n \cdot (N + 1) - \Sigma(A).$$

Hence Σ takes the value $n\,(N + 1) - x$ exactly as many times as the value x.

(v) By (iv) we need to consider only the case $3 \leq x \leq N+1$. Let $1 \leq a \leq N-1$. The two-element sets A with $\min A = a$ contribute 1 to the frequencies of x for $x = 2a + 1, \ldots, a + N$, and 0 otherwise. That is, the value $x = \Sigma(A)$ occurs once for a with $2a + 1 \leq x$, or equivalently, $a \leq \lfloor \frac{x-1}{2} \rfloor$.

(vi) Let $A \subseteq \{1, \ldots, N\}$ be an n-element set. We distinguish two cases:

Case 1. The subsets $A \subseteq \{1, \ldots, N-1\}$ contribute $F_{N-1,n}(x)$ times the sum x.

Case 2. The subsets with $N \in A$ have $\Sigma(A) = \Sigma(A - \{N\}) + N$, hence contribute $F_{N-1,n-1}(x-N)$ times the sum x.

(vii) Again we distinguish two cases according to whether $1 \in A$.

Case 1. The subsets $A \subseteq \{2, \ldots, N\}$ contribute $F_{N-1,n}(x-n)$ times the sum x: consider the sets $A' = \{a-1 \mid a \in A\}$ with sum $\Sigma(A') = \Sigma(A) - n$.

Case 2. If $1 \in A$ we consider the $(n-1)$-element set $\bar{A} = \{a-1 \mid a \in A, a \neq 1\}$ with $\Sigma(A) = \Sigma(\bar{A}) + (n-1) + 1$. These sets A contribute $F_{N-1,n-1}(x-n)$ times the sum x.

$\diamond$

As applications of the first recursion formula we prove some sum formulas by induction.

Corollary 1 *For* $1 \leq n \leq N$ *we have*

$$\sum_{x \in \mathbb{Z}} F_{N,n}(x) = \binom{N}{n}.$$

Proof. Start with $N = 1$, hence also $n = 1$. Then the lefthand side has one non-zero summand: $F_{1,1}(1) = 1$. The righthand side is 1.

Now assume that $N \geq 2$. Then by (vi) and induction

$$\begin{aligned}\sum_{x \in \mathbb{Z}} F_{N,n}(x) &= \sum_{x \in \mathbb{Z}} F_{N-1,n}(x) + \sum_{x \in \mathbb{Z}} F_{N-1,n-1}(x-N) \\ &= \binom{N-1}{n} + \binom{N-1}{n-1} = \binom{N}{n}.\end{aligned}$$

$\diamond$

This proof serves as exercise for the evaluation of more complex sums below. From a practical point of view it is dispensable since the corollary simply states that the set $\{1, \ldots, N\}$ has exactly $\binom{N}{n}$ subsets of size n.

Corollary 2 *For* $1 \leq n \leq N$ *we have*

$$\sum_{x \in \mathbb{Z}} x \cdot F_{N,n}(x) = \frac{n\,(N+1)}{2} \cdot \binom{N}{n}.$$

Proof. Again we start with $N = 1$, hence also $n = 1$. Then the lefthand side has one non-zero summand: $1 \cdot F_{1,1}(1) = 1$. The righthand side is $\frac{1 \cdot 2}{2} \cdot \binom{1}{1} = 1$.

Now assume that $N \geq 2$. Then by (vi) and induction

$$\begin{aligned}
\sum_{x \in \mathbb{Z}} x \cdot F_{N,n}(x) &= \sum_{x \in \mathbb{Z}} x \cdot F_{N-1,n}(x) + \sum_{x \in \mathbb{Z}} (x - N) \cdot F_{N-1,n-1}(x - N) \\
&\quad + \sum_{x \in \mathbb{Z}} N \cdot F_{N-1,n-1}(x - N) \\
&= \frac{n\,N}{2} \cdot \binom{N-1}{n} + \frac{(n-1)\,N}{2} \cdot \binom{N-1}{n-1} + N \cdot \binom{N-1}{n-1} \\
&= \frac{n\,N}{2} \cdot \binom{N-1}{n} + \frac{n\,N}{2} \cdot \binom{N-1}{n-1} + \frac{N}{2} \cdot \binom{N-1}{n-1} \\
&= \frac{n\,N}{2} \cdot \binom{N}{n} + \frac{N}{2}\frac{n}{N} \cdot \binom{N}{n} \\
&= (N+1) \cdot \frac{n}{2} \cdot \binom{N}{n}.
\end{aligned}$$

$\diamond$

Again this was a redundant proof: The corollary says that the "mean of the means $\Sigma(A)/n$" of all $\binom{N}{n}$ subsets A of size n is $(N+1)/2$. This however results directly from the symmetry of the distribution.

Corollary 3 *For $1 \leq n \leq N$ we have*

$$\sum_{x \in \mathbb{Z}} x^2 \cdot F_{N,n}(x) = \frac{n\,(N+1)}{12} \cdot (3nN + N + 2n) \cdot \binom{N}{n}.$$

Proof. Here too we start with $N = 1$, hence also $n = 1$. Then the lefthand side has one non-zero summand: $1^2 \cdot F_{1,1}(1) = 1$. The righthand side is $\frac{1 \cdot 2}{12} \cdot (3+1+2) \cdot \binom{1}{1} = 1$.

Now assume that $N \geq 2$. Note that

$$x^2 = [(x - N) + N]^2 = (x - N)^2 + 2N\,(x - N) + N^2.$$

Thus by (vi) and induction

$$
\begin{aligned}
\sum_{x\in\mathbb{Z}} x^2\cdot F_{N,n}(x) &= \sum_{x\in\mathbb{Z}} x^2\cdot F_{N-1,n}(x) + \sum_{x\in\mathbb{Z}} (x-N)^2\cdot F_{N-1,n-1}(x-N)\\
&\quad + 2N\cdot\sum_{x\in\mathbb{Z}}(x-N)\cdot F_{N-1,n-1}(x-N) + N^2\cdot\sum_{x\in\mathbb{Z}} F_{N-1,n-1}(x-N)\\
&= \frac{n\,N}{12}\cdot\binom{N-1}{n}\cdot[3n\,(N-1)+(N-1)+2n]\\
&\quad + \frac{(n-1)\,N}{12}\cdot\binom{N-1}{n-1}\cdot[3\,(n-1)(N-1)+(N-1)+2\,(n-1)]\\
&\quad + 2N\cdot\binom{N-1}{n-1}\cdot\frac{(n-1)\,N}{2} + N^2\cdot\binom{N-1}{n-1}\\
&= \frac{n\,N}{12}\cdot\frac{N-n}{N}\cdot\binom{N}{n}\cdot[\underbrace{3nN-3n+N-1+2n}_{3nN+N-n-1}]\\
&\quad + \frac{(n-1)\,N}{12}\cdot\frac{n}{N}\cdot\binom{N}{n}\cdot[\underbrace{3nN-3n-3N+3+N-1+2n-2}_{3nN-2N-n}]\\
&\quad + N\cdot\frac{n}{N}\cdot\binom{N}{n}\cdot[nN-N] + N^2\cdot\frac{n}{N}\cdot\binom{N}{n}\\
&= \frac{n}{12}\cdot\binom{N}{n}\cdot\left[3nN^2+N^2-nN-N-3n^2N-nN+n^2+n\right]\\
&\quad + \frac{n}{12}\cdot\binom{N}{n}\cdot\left[3n^2N-2nN-n^2-3nN+2N+n\right]\\
&\quad + \frac{n}{12}\cdot\binom{N}{n}\cdot[12nN-12N+12N]\\
&= \frac{n}{12}\cdot\binom{N}{n}\cdot\left[3nN^2+N^2+5nN+N+2n\right]\\
&= \frac{n}{12}\cdot\binom{N}{n}\cdot[3nN\,(N+1)+N\,(N+1)+2n\,(N+1)]\,.
\end{aligned}
$$

◇

As a first test, using Formula 1 and Corollary 1, we evaluate the (finite) sum

$$
\sum_{t\in\mathbb{R}} p_{N,n}(t) = \frac{1}{\binom{N}{n}}\sum_{t\in\mathbb{R}} F_{N,n}(nt) = \frac{1}{\binom{N}{n}}\sum_{x\in\mathbb{R}} F_{N,n}(x) = 1,
$$

the expected result.

Secondly in the same way, using Corollary 2, we evaluate the formula for the mean value

$$\sum_{t\in\mathbb{R}} t\, p_{N,n}(t) = \frac{1}{\binom{N}{n}} \sum_{t\in\mathbb{R}} t\, F_{N,n}(nt) = \frac{1}{\binom{N}{n}} \frac{1}{n} \sum_{x\in\mathbb{R}} x\, F_{N,n}(x) = \frac{N+1}{2}\,,$$

as we knew already.

Finally as a new result we derive a closed expression for the variance of the sample mean:

$$\begin{aligned}
\sum_{t^2\in\mathbb{R}} t\, p_{N,n}(t) - \left(\frac{N+1}{2}\right)^2 &= \frac{1}{\binom{N}{n}} \sum_{t\in\mathbb{R}} t^2\, F_{N,n}(nt) - \left(\frac{N+1}{2}\right)^2 \\
&= \frac{1}{\binom{N}{n}} \frac{1}{n^2} \underbrace{\sum_{x\in\mathbb{R}} x^2\, F_{N,n}(x)}_{\frac{n\,(N+1)}{12}\binom{N}{n}(3nN+N+2n)} - \left(\frac{N+1}{2}\right)^2 \\
&= \frac{N+1}{12\,n}\cdot(3nN+N+2n) - \frac{N+1}{12\,n}\cdot(3nN+3n) \\
&= \frac{(N+1)(N-n)}{12\,n}.
\end{aligned}$$

By this we have proved the main result of this article:

Proposition 1 *Let $1 \le n \le N$. Consider the distribution $p_{N,n}$ of the mean values of samples of size n from the set $\{1,\ldots,N\}$. This distribution has the variance*

$$V(p_{N,n}) = \frac{(N+1)(N-n)}{12\,n}.$$