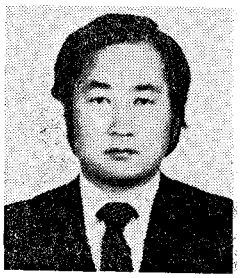
데이타의 整理方法



金 源 炅
〈韓國工業標準協會 專門委員〉

分布의 數量的인 表示法

1. 平均値

우리 직장에서는 물론 일반가정에서도 흔히 平均値를 使用한다. 예를 들면, 우리나라 남자 어린이의 평균 체중, 어느 부서의 평균 생산량이 1천 5백상자 혹은 건전지 수명이 8시간 등의 식으로 평균해서 이야기하고 있는 경우가 많다.

데이터의 분포 즉 집단으로서의 데이터의 성질을 數量的으로 나타내려면, 분포의 中心과 분포의 散布 두가지를 생각할 수 있다. 분포의 中心과 散布의 두가지만 결정되면 거의 모습이 정해진다. 이 분포의 中心位置를 나타내는 尺度로서는 평균치, 메디안 밋드렌지, 모오드 등이 있지만 평균치를 제일 많이 쓰고 있다.

평균치에는 算術平均, 幾何平均, 調和平均 등이 있지만 일반적으로 평균치라고 하면 算術平均을 말한다. "平均値"란 「데이터를 전부 합해 그것을 데이터의 개수로 나눈 값을 말하며 $\bar{x}$(엑스바아)로 나타내며 試料에서 구한 평균치이기 때문에 "試料平均"이라고도 부른다. n개의 데이터를 $x_1, x_2, x_3, \cdots x_n$으로 나타내면, 이 평균치 $\bar{x}$는 $\bar{x}=\frac{x_1+x_2+\cdots x_n}{n}$이 된다. $x_1+x_2+\cdots+x_n$라 쓰기 번거롭기 때문에 이것을 $\sum_{i=1}^{n} x_i$로 나타내기도 하며 i가 첫번째 데이터로부터 시작, n번째까지의 모든 데이터를 合算한다는 뜻을 가지고 있다.

Σ는 그리이스 문자로 「시그마」라고 읽으며 로오마자의 S에 해당한다. 영어에서는 덧셈을 나타내는 기호로써 쓰이고 있다. 이 Σ라는 기호를 사용하게 되면 $\bar{x}=\frac{\sum_{i=1}^{n} x_i}{n}=\frac{\text{데이터의 합계}}{\text{데이터수}}$로 표현할 수가 있다.

즉 Σ는 취할 수 있는 모든 값인 $i=1$에서부터 $i=n$까지에 이르는 모든 것을 합치는 것을 생략하여 간단히 하면 $\bar{x}=\frac{\sum x}{n}$로 나타낸다.

2. 平方合, 分散, 標準偏差 및 範圍

평방합, 분산, 표준편차 및 범위는 산포의 크기를 數量的으로 나타낸 척도라고 한다. 그 의미와 계산을 예를 들어 보면 어떤 회사의

1, 2, 3개 라인으로 각 5명씩 랜덤으로 선택하여 어떤 순간의 포장능률의 숫자를 조사해 본 결과 다음과 같은 데이터를 얻었다.

〈表 1〉 능률의 데이터

(단위 : 상자)

라인	포 장 상 자	평균치 $\bar{x}$	메디안 $\tilde{x}$	미드렌지 M_1
1	10, 10, 10, 10, 10	$\frac{50}{5}=10$	10	$\frac{10+10}{2}=10$
2	6, 7, 10, 13, 14	$\frac{50}{5}=10$	10	$\frac{6+14}{2}=10$
3	3, 6, 10, 14, 17	$\frac{50}{5}=10$	10	$\frac{3+17}{2}=10$

이 3개 라인의 능률은 平均値 $\bar{x}$도, 메디안 $\tilde{x}$도 그리고 밋드렌지 M_1도 똑같이 다 10상자이었다. 그러나 이 3개 라인은 한눈으로 봐도 분명히 差가 있다. 그 어디가 어떻게 다른가?

그것은 분포의 중심이 똑 같으면서도 산포의 크기가 다른 것을 알 수가 있다. 즉 1라인은 모든 데이터가 10상자로 전연 산포가 없는 데이터가 되어 있다. 그런데 2라인은 10상자를 중심으로 6상자에서 14상자까지의 사이이고 또한 3라인은 3상자에서 17상자까지의 범위에 걸쳐서 5개의 데이터가 산포되고 있다.

이것을 수량적으로 잘 평가할 수 있는 방법이 없겠는가, 그럴려면 산포를 표현하든지, 다시 말해서 尺度를 定義하면 된다. 이들에 대해서 하나하나 생각해 보기로 한다.

(1) 平方合 S

산포의 크기를 숫자로 표시하려면 개개의 값 x가 평균치 $\bar{x}$에서 어느 정도 떨어져 있는가를 나타내는 尺度를 생각하면 된다. 개개의 데이터 x_i와 평균치 $\bar{x}$와의 차, 즉 $x_i-\bar{x}$를 偏差라고 말한다.

앞에서 언급한 1, 2, 3라인의 데이터에 대해서 偏差의 合을 구하면 다 같이 제로(0)가 된다. 3라인의 경우를 예로 들면, $\sum(x_i-\bar{x})=(3-10)+(6-10)+(10-10)+(14-10)+(17-10)+(17-10)=(-7)+(-4)+0+4+7=0$ 이 된다.

즉 偏差의 합에서는 偏差의 플러스 값과 마이너스 값이 서로 상쇄되어 그 總和가 항상 0이 되는 것이다. 따라서 偏差의 합 또는 평균치를 가지고 산포를 평가할 수는 없다. 그래서 偏差의 부호를 없애는 방법을 생각해야 한다. 이 방법으로는 여러 가지를 생각할 수 있는데, "偏差의 絕對値"를 취하는 것도 그러한 방법 중의 한 가지이다. 그 밖에 偏差를 제곱하는 것도 하나의 방법이다. 여기서는 이 방법을 택하여 곱하기로 하면 母集團의 상태도 쉽게 알 수 있기 때문이다. 3라인의 데이터에 대한 偏差의 제곱을 계산하면,

$$\begin{aligned}\sum(x_i-\bar{x})^2&=(3-10)^2+(6-10)^2+(10-10)^2\\&\quad+(14-10)^2+(17-10)^2\\&=(-7)^2+(-4)^2+0^2+(4)^2+(7)^2\\&=130\text{상자}\end{aligned}$$

가 된다. 이것을 「平方合」이라고 하며 S로 표시한다. 平方合이란 개개 측정치의 평균치로부터의 偏差의 제곱합이다. 1라인은 $S=0$(상자)^{2}이며 2라인의 데이터는 $S=50$(상자)^{2}이 된다. 이들 平方合의 값은 3개라인의 데이터의 산포의 크기에 대응하는 것을 알수가 있다 즉 $S=\sum_{i=1}^{n}(x_i-\bar{x})^2$로 구해진다.

(2) 分散 V

平方合은 분명히 分布의 散布를 나타내지만 데이터의 수 n에 의해서 영향을 받는다는 결점을 가지고 있다. 데이터의 수가 많아지면

당연히 平方合 S도 커지게 마련이다. 이래서는 불편하기 때문에 n의 영향을 없애기 위하여 $(n-1)$로 나눈다. 데이터 $x_1, x_2, x_3 \cdots x_n$이 있을 때 平方合 S를$(n-1)$로 나눈 값을 分散, 또는 不偏分散이라고 하며 이때 $(n-1)$을 自由度(degress of freedom)라고 부르며, 보통 ϕ(타이)라는 기호로 나타낸다.

$$V=\frac{S}{\phi}=\frac{S}{n-1}$$

(3) **標準偏差 s**

分散 V는 偏差의 제곱의 代表値와 같으며 偏差의 代表를 구하려면 分散 V를 平方根으로 구한 「標準偏差 s」를 사용하면,

$$s=\sqrt{V}=\sqrt{\frac{S}{n-1}}$$

표준편차란 표준적인 偏差라는 의미이며 모집단의 표준편차—모표준편차—에 대응하는 값(統計量)이다. 앞에 나온 1,2,3개라인의 표준 편차를 구하면,

1라인 : $s_1=0$(상자)

2라인 : $s_2=\sqrt{50/4}=\sqrt{12.5}=3.5$ (상자)

3라인 : $s_3=\sqrt{130/4}=\sqrt{32.5}=5.7$(상자)

이렇게 平方合 S는 分散 V의 단위가 (상자)^{2}식으로 잘 이해할 수 없는 것에 비해 표준편차 s에서는 평균치와 같은 단위인 (상자)로 나타나 편리하다.

표준편차의 계산식을 봐서 알 수 있듯이 「표준편차가 작다」고 하는 것은 즉, 전체의 산포가 작다는 것은 분포가 평균치의 주위에 모여 있다는 것이다. 반대로 「표준편차가 크다」는 것은 분포가 평균치로부터 멀리 떨어져 있는 것을 의미한다.

(4) **範圍 R**

산포의 크기로 1조의 데이터 $x_1x_2 \cdots x_n$중의 최대치에서 최소치를 뺀 것을 말한다. 최대치를 x_{max}. 최소치를 x_{min}이라하면 $R=x_{max}-x_{min}$이 된다. 2라인은 $R=14-6=8$(상자) 3라인은 $R=17-3=14$(상자)의 계산이 된다.

3. 랜덤샘플링(random sampling)

母集團이 分布를 지니고 있다는 사실은 母集團에서 샘플을 取할때 랜덤샘플링이 되도록 주의해야 한다는 것을 의미한다.

가령 製品을 보았을 때, 經驗者가 보면 좋은 것과 나쁜 것이 區別이 되고, 作業者는 샘플로서 좋은 것을 選擇하고 싶어 할 것이다. 또 檢査員은 나쁜 것을 샘플링할려고 할 것이다. 이렇게 되어서는 工程을 代表하는 샘플이라고 할 수 없다. 우리는 工程의 상태를 알기 위하여 샘플을 取하는 것이지만, 이것은 일반적으로 選擇하여 取하는 것보다 랜덤하게, 즉 손에 잡히는 대로 샘플을 取하는 것이 더욱 좋다.

또 랜덤샘플링을 했을 때 샘플의 平均値나 範圍, 標準偏差 등이 어떤 값이며 어떤 分布를 가질 것인가 하는 것이 統計學에서는 잘 알려져 있다. 따라서 랜덤샘플링을 함으로서 統計量의 分布法則에 의거하며 그 데이터를 判斷해 가자는 것이다.

앞에서 말한 바와 같이 좋은 것과 나쁜 것을 選擇해서 행하는 샘플링은 랜덤화될 수 없다.

실제로 一定한 間隔을 두고 샘플을 取하면 대개 랜덤샘플링이 될 수 있다. 이 랜덤샘플링이란 統計的 手法을 活用해서 品質管理나 그 밖의 다른 管理를 實施해 갈 때 아주 重要한 것으로서, 確實히 이루어져야 할 것이다.

〈다음호에 계속〉