A measure of deviation of frequency of a particular item in a sample from the expected frequency

Subhrajit Bhattacharya [†]
Dept. of Mechanical engineering,
IIT Kharagpur,
INDIA.
email: subhrajit83@yahoo.co.in

The deviation of obtained frequency from the expected frequency of a particular item in a sample (of size n) collected from a population (of size N) is measured, and its expected value is found out.

[†] The author is a second year student in the department of Mechanical engineering in IIT Kharagpur.
Postal address for correspondence: 53/1, Canal Street, Kolkata - 700048, INDIA.

Let the population be of size N of which x individuals are of a particular type (say 'A').
Hence, proportion of 'A' in the population = (x / N)

Let a sample of size n be taken.
Hence, expected number of 'A' in the sample = (x / N).n

Let, f = frequency of 'A' obtained in a particular sample.
Hence, $f \in \{ 0, 1, \ldots, \min(n, x) \}$

Hence, $\Delta_f = [f - (x / N).n]^2$ is a measure of how much deviated is the actual frequency f from the expected (x / N).n of individual 'A'.
We take the squared value so that $\Delta_f \geq 0$

The mean value of $\Delta_f = E(\Delta_f)$ gives a measure of mean deviation.
Hence,

$$E(\Delta_f) = \sum_{f=0}^{\min(n,\, x)} \left[\Delta_f \cdot P(\text{ In the sample exactly } f \text{ items are of type 'A' }) \right]$$

Now,
P(In the sample exactly f items are of type 'A')
= (no. of ways in which exactly f of the n chosen items can be of type 'A') / ${}^{N}P_{n}$
$= ({}^{x}C_{f} \cdot {}^{N-x}C_{n-f} \cdot n!) / {}^{N}P_{n}$
$= ({}^{x}C_{f} \cdot {}^{N-x}C_{n-f}) / {}^{N}C_{n}$

$E(\Delta_f) =$

$$\delta(n, x) = \sum_{f=0}^{\min(n,\, x)} \left[\left(f - \frac{x}{N} n \right)^2 \cdot \frac{{}^{x}C_{f} \cdot {}^{N-x}C_{n-f}}{{}^{N}C_{n}} \right]$$

$$= \frac{1}{{}^{N}C_{n}} \sum_{f=0}^{\min(n,\, x)} \left[\left(f - \frac{x}{N} n \right)^2 \cdot {}^{x}C_{f} \cdot {}^{N-x}C_{n-f} \right]$$

If $n \leq x$, then, $\min(n, x) = n$
If $n > x$, ${}^{x}C_{f} = 0$ for all $f > x$, i.e. we can add terms after $f = x$ to the summation series whose net contribution will turn out to be 0.
Hence, we may write,

$$\delta(n, x) = \frac{1}{{}^{N}C_{n}} \sum_{f=0}^{n} \left[\left(f - \frac{x}{N} n \right)^2 \cdot {}^{x}C_{f} \cdot {}^{N-x}C_{n-f} \right]$$

<u>Notes</u>:

$\delta(n, x)$ is a measure of deviation of obtained frequency from the expected frequency of a particular item 'A' in a sample (of size n) collected from a population (of size N in which exactly x items are 'A')

1. $\delta(n, 0) = 0$
2. $\delta(0, x) = 0$
3. When $n = N$ (i.e. sample is the population), some simple mathematical works will give $\delta(N, x) = 0$, which is an expected result, for when the sample is the population, no deviation of sample proportion can occur from population proportion.
4. When $N \to \infty$ and $(n/N) = p$ doing some mathematical works we get $\delta = 3x.2^{x-2} + 2^{x}.(1 - p.x^{2})$
5. $[\delta(n, x)]^{1/3}/n$ gives a measure of proportion deviation. It can be used to compare deviation measures among samples of different sizes.