ASPECTS ON STATISTICAL APPROACH OF POPULATION HOMOGENEITY

Alexandru ISAIC-MANIU

PhD, University Professor, Department of Statistics and Economic Prognosis
University of Economic Studies, Bucharest, Romania
(**Co)Author of the books:** Proiectarea statistica a experimentelor (2006), Enciclopedia calitatii (2005), Dictionar de statistica generala (2003), Statistica pentru managementul afacerilor (2000), Metoda Weibull (1983)

E-mail: al.isaic-maniu@csie.ase.ro, **Web page:** http://www.amaniu.ase.ro

Viorel Gh. VODĂ

PhD, Senior Scientific Researcher, Institute of Mathematical Statistics and Applied Mathematics of the Romanian Academy, Bucharest, Romania
Co-author of the books: Proiectarea statistica a experimentelor (2006), Dictionar de statistica generala (2003), Manualul calitatii (1997)

E-mail: von_voda@yahoo.com

Abstract: *In this article we emphasize the manner in which this statistical indicator - the variation coefficient (v) - could help the inference on measurable characteristics generated by technological processes. Our interest lies upon the so-called SPC- Statistical Process Control; the main result obtained is the following: if the coefficient of variation is known, then the statistical distribution of capability index is of ALPHA-type distribution (Družinin). We also put into light some links between (v) and Taguchi's quality loss function.*

Key words: *variation coefficient; signal-to-noise ratio; Alpha distribution; capability index; quality loss function*

1. Introduction

It is well-known that when we desire to compare the spread (dispersion) in two sets of data if we choose to do this straightforwardly by comparing the two standard deviations - this may lead to fallacious conclusions. This may be due to the fact that the variables/characteristics involved are measured in different units. Furthermore, if the same unit of measurement is employed, one may still see a large difference between the two means.

Such situations may occur with data obtained from various areas of interest - from technology to biostatistics.

To deal with cases like those, we need a measure (an indicator) of **relative variation** rather than **absolute** variation. The coefficient of variation, which expresses the standard deviation as a percentage of the mean, is just such an indicator: since the mean and standard deviation have the same measurement unit (as the original data, in fact) this coefficient of variation is independent of any unit of measurement.

Theoretically, if X is a measurable characteristic - that is a continuous random variable (c.r.v.) with finite mean-value and variance $(E(x) < +\infty, \quad Var(x) < +\infty)$, then the ratio

$$V = \frac{\sqrt{Var(x)}}{E(x)}, \qquad \text{where } E(x) \neq 0 \tag{1}$$

is the coefficient of variation (c.v.) associated to X. Let us notice that the request for finitude is mandatory since there are distributions for which this condition is not fullfiled. For instance, the Cauchy distribution:

$$x : f(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}, \quad x \in R \tag{2}$$

"has no average value" - since

$$E(x) = \int_R x \cdot f(x) dx = \frac{1}{2\pi} \int_R \frac{2x}{1+x^2} dx = \frac{1}{2\pi} \ln(1+x^2) \Big|_{-\infty}^{+\infty} = \infty - \infty \quad \text{(a "senseless" form)} \tag{3}$$

or in the case of inverse Rayleigh variable:

$$X : f(x;a) = 2ax^{-3} \exp(-a/x^2), \quad x > 0, \quad a > 0 \tag{4}$$

where the variance is $Var(x) = +\infty$ (see Treyer, 1976 [17] or Bârsan-Pipu *et al.*, 1999 [2]).

2. Some properties related to sample coefficient of variation

As it is well-known, in practice we work with the sample coefficient of variation $\hat{V}$, that is:

$$\hat{V} = S/\overline{x}, \text{ where } \overline{x} = \frac{1}{n}\sum x_i \quad \text{and} \quad S^2 = \sum (x_i - \overline{x})^2/(n-1) \tag{5}$$

On the other hand, some authors (see Mc. Kay, 1932 [13] consider also the form:

$$\hat{V}_0 = S_0/\overline{x}, \text{ where and } \quad S_0{}^2 = \sum (x_i - \overline{x})^2/n \tag{6}$$

which provides that the statistics $B \cdot \hat{V}_0^2/(1+\hat{V}_0^2)$ where $B = n(1+1/\hat{V}^2)$ is chi-square distributed with (n-1) degrees of freedom (χ^2_{n-1}).

Johnson and Welch (1940, [12]) proved that $\sqrt{n}/\hat{V}$ has a non-central t-distribution with $(n-1)$degrees of freedom and $\sqrt{n}/V$ as noncentrality parameter and the underlying variable x is normally distributed $N(\mu, \sigma^2)$.

F.N. David (1949 [3]) gave some approximations to the first four moments of $\hat{V}$, assuming that x has a normal variance and the mean value of $V = \sigma/\mu$ is not (very) large.

Iglewicz, Myers and Howe (1968 [9]) provided some approximations for the percentiles $\hat{V}_p$ of $\hat{V}$, assuming also normality of X and imposing to the value of V, the restriction $V \le 0.5$. The percentiles are obtained from the equation $\Pr ob\{\hat{V} \le \hat{V}_p\} = 1 - p$ via $\chi^2_{n-1;p}$ - the quantiles of chi-squared distribution.

Two years later the same Iglewicz and Myers (1970 [10]) gave a simpler version of $\hat{V}_p$ as:

$$\hat{V}_p \approx \sqrt{n/(n-1)} \cdot \sqrt{\chi^2_{n-1;p} / \left(B - \chi^2_{n-1;p}\right)} \tag{7}$$

where B has been defined above.

Ivan and Văduva in a paper in Romanian (see [11, 1969]) proposed a series expansion of the density of $\hat{V}$ as follows:

$$f(x;\gamma,\delta) = \frac{2\exp(-\delta^2/2)}{\sqrt{\pi}\cdot\Gamma(\gamma/2)} \cdot x^{\gamma-1} \cdot \sum_{K=0}^{\infty} \frac{(2\delta^2)^K}{(2K!)} \cdot \frac{\Gamma\left(\frac{\gamma+1}{2}+K\right)}{(1+x^2)^{\frac{\gamma+1}{2}+K}}$$

where $\gamma = n-1$, $\delta = |\mu|\sqrt{n}/\sigma$ and $\Gamma(\bullet)$ is the well-known Gamma function. Their formula - in spite of the fact that is an "exact" one, is very cumbersome to use.

Warren (1982 [19] proposes the use of the exact relationship of $\hat{V}$ to noncentral t, or better - he says - the normal approximation to noncentral t.

Anders Hald (1952, [8]) considered a normal variable $N(\mu,\sigma^2)$ with known $V = \sigma/\mu$ and proved that one may deduce the following approximations:

$$E(\hat{V}) \approx V \quad \text{and} \quad Var(\hat{V}) \approx \frac{V^2}{2(n-1)}\left(1+2V^2\right) \tag{8}$$

where n is the sample size used to estimate $\mu : \bar{x} = n^{-1}\sum x_i$, x_i being the sample measurements on $X \in N(\mu,\sigma^2)$.

For large n and small values of V, the sample coefficient of variation may be considered as approximately normally distributed with mean V and variance $V^2/2(n-1)$. We discarded the term $V^4/(n-1)$ which is negligible in the above assumptions.

3. Some new inferences

Let X be a measurable quality characteristic of class $N(\mu,\sigma^2)$, where $V = \sigma/\mu$ is assumed to be known ($V = V_0$ - a positive value). Therefore, $\sigma = V_0\mu$ and the law becomes:

$$X : f(x; V_0, \mu) = \frac{1}{\mu\left(V_0\sqrt{2\pi}\right)} \cdot \exp\left\{-\frac{(x-\mu)^2}{2V_0^2\mu^2}\right\} \tag{9}$$

$x \in R, \quad \mu > 0, \quad V_0 > 0$

The parameter μ is easily estimated by the sample mean $\bar{x}$ and therefore $\hat{\sigma} = V_0 \cdot \bar{x}$, where $\bar{x} = n^{n-1}\sum x_i$.

It is important to notice that MLE (Maximum Likelihood Estimator) for μ in this case has a "ugly" form and it is obtained as the positive root of a second degree equation (a similar case when we have the law $N(\lambda,\lambda)$ - that is mean value is equal to the variance, has been investigated in Stoichițoiu-Vodă, 2002, [16, page 255]).

Tziafetas (1989, [18, page 965]) states that if $x \in N(\mu,\sigma^2)$ then its associated Gini's coefficient G has the value $(\sigma/\mu)\cdot\sqrt{\pi}$ - in our case $V_0\sqrt{\pi}$. Let us remember that Corrado Gini (1884 - 1965) has proposed in 1912 (see [7]) what is called now "Gini coefficient" which is in fact half of the **relative mean difference**:

$$G = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right| \tag{10}$$

which is an **even location free** statistic (see Patel and Read 1996 [15] page 280).

H.A. David (1968 [43]) discovered that (10) does appear in an old paper of Friedrich Robert Helmert (1843 - 1917) published 1876 in a German astronomical journal (see also David and Edwards, 2001, [5] for an English version of Helmert's article).

According to Zítek (1954 [20]), Helmert's statistic has been used by the astronomer Halger von Andrae in 1872 as an estimator of the so-called "probable error" $(0{,}6745\cdot\sigma)$ as:

$$\hat{\sigma} = \frac{\sqrt{\pi}}{n(n-1)}\sum_{i=1}^{[n/2]}(n-2i+1)\cdot W_{(i)}$$

where [n/2] is the largest integer in n/2 and $W_{(i)}$ is the quasi-range of order i, namely $W_{(i)} = x_{(n-i+1)} - x_{(i)}$, where $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ are ordered sample values.

It is important to notice that *from an economical point of view*, a known coefficient of variation in the normal case is not of much use since this means that the measure of income inequality (or wealthy inequality) is always constant - which is not the case in real life (there is a paper in this respect belonging to Gini himself: "Measurement of inequality and incomes" published in 1921 - see Morgan, 1962 [14]).

The assumption of a known V is useful in process capability theory. Let [LSL, USL] be the interval of specifications imposed to the characteristic X (LSL = Lower Specifications Limit; USL = Upper Specifications Limit) and therefore, the potential index of the process is:

$$Cp = \frac{USL - LSL}{6\sigma} = \left(\frac{USL - LSL}{6V_0}\right)\cdot\frac{1}{\mu} \tag{11}$$

since we did assume that $V = \sigma/\mu = V_0(>0)$. The estimator of C_p is hence:

$$\hat{C}p = \left(\frac{USL - LSL}{6V_0}\right)\cdot\frac{1}{\overline{x}} = K\cdot\frac{1}{\overline{x}} \tag{12}$$

where k is the constant $(USL - LSL)/6V_0$

It follows that the inference on $\hat{C}_p$ is transferred to the inference on $1/\overline{x}$ - that is on the reciprocical of $\overline{x}$ - the sample average (where $\overline{x} > 0$).

Since $\overline{x}$ is normally distributed $N(\mu,\sigma^2/n)$, we may write:

$$f\left(x,\mu,\sigma^2/n\right)=\frac{\sqrt{n}}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2/n}\right\}=\frac{\sqrt{n}}{V_0\mu\sqrt{2\pi}}\exp\left\{-\frac{n(x-\mu)^2}{2V_0\mu^2}\right\}, \quad x>0,\quad \mu>0 \tag{13}$$

The distribution of $1/\overline{x}$ has to be evaluated now:

$$\Pr ob\left\{\frac{1}{\overline{x}}<u\right\}=\Pr ob\left\{\frac{1}{u}<\overline{x}\right\}=1-\Pr ob\left\{\overline{x}\le\frac{1}{u}\right\} \tag{14}$$

where we did assume that $\overline{x}>0$. Therefore, we have:

$$F(u)=1-\Pr ob\left\{\overline{x}\le\frac{1}{u}\right\}=1-\int_0^{1/u} f\left(x;\mu,\sigma^2/n\right)dx \tag{15}$$

where if we take the derivative, we obtain the density function of the variable $1/\overline{x}$:

$$f(u)=F'(u)=\frac{\sqrt{n}}{\left(v_0\mu\sqrt{2\pi}\right)\cdot u^2}\cdot\exp\left\{-\frac{n\left(\frac{1}{u}-\mu\right)^2}{2v_0^2\mu^2}\right\},\quad u>0,\quad v_0,\mu>0 \tag{16}$$

Hence, we did reach the following (interesting-we dare to say) result, namely:

> The distribution of the estimated potential index $\hat{C}_p$ in the case when the variation coefficient is known, is of Družinin Alpha type distribution (see Dorin *et al*, 1994 [6], p. 110 - 117) - that is the distribution of a left truncated normal variable, the truncation point being $x_\tau=0,\quad x>x_\tau=0$

Another intervention of V in capability evaluation is the following: let x be a measurable characteristic, normally distributed with unknown V and consider that we have only one specification, namely LSL = 0. In this case, the capability index $\hat{C}_{pk}$ is:

$$\hat{C}_{pk}=\min\left\{\frac{|\overline{x}\text{ - LSL}|}{3s},\frac{|\overline{x}\text{ - USL}|}{3s}\right\}=\frac{|\overline{x}\text{ - LSL}|}{3s}=\frac{1}{3}\cdot\left(\frac{\overline{x}}{s}\right) \tag{17}$$

that is $\hat{C}_{pk}=(1/3)\cdot(1/\hat{v})$, where $\hat{v}=s/\overline{x}$. This "inverse" $(1/\hat{v})$ is known as signal-to-noise ratio (the empirical one), used by Genichi Taguchi (see [1]) in his theory of experimentation.

Taguchi also considered the so-called average-loss function associated to quality, that is $L(y)=k\cdot E\left[(y-T)^2\right]$ where y is a measured value of the characteristic and T is its target value (or "optimal" value). In practice, we deal with

$$L(y)=k\cdot E\left[s^2+(\overline{y}-T)^2\right] \tag{18}$$

where $\overline{y}=n^{-1}\sum y_i$, $s^2=n^{-1}\sum(y_i-\overline{y})^2$, y_i, $i=1,n$ are measured value and k - a constant depending on the actual problem investigated.

If T = 0, then $L(y)=k\left[s^2+\overline{y}^2\right]=ks^2\left[1+\left(\frac{\overline{y}}{s}\right)^2\right]$ and we see that $1/\overline{v}^2$ appears:

$$L(y)=ks^2\left[1+(1/\overline{v})^2\right] \tag{19}$$

If V is known (V = V_0), the loss-function depends only on variability (s). On the other hand, if we have $\overline{y}=T$, then:

$$L(y)=k\cdot s^2=k\overline{y}^2\cdot\left(\frac{s^2}{\overline{y}^2}\right)=k\cdot y^2\cdot\left(\overline{v}^2\right) \tag{20}$$

which means that the loss-function depends only on location $(\overline{y})$ if V is known.

Taguchi states that if $\overline{y}$ is very close to the target value (T), then, the standard deviation could be expressed as:

$$s_0=s\cdot\frac{T}{\overline{y}} \text{ (we have } s_0\approx s \text{ if } T\approx\overline{y}\text{)} \tag{21}$$

and hence, the loss-function becomes $L(y)=K\cdot T\cdot\hat{V}^2$.

Since the statistics $\overline{y}$ and s^2 are independent, we may write:

$$E[L(y)]=K\cdot T^2\cdot E\left(\frac{s^2}{\overline{y}^2}\right)=KT^2\cdot E(s^2)\cdot E\left(\frac{1}{\overline{y}^2}\right) \tag{22}$$

Since in general, we have $E\left(\frac{1}{x}\right)>\frac{1}{E(x)}$, we obtain finally the inequality:

$$E[L(y)]\geq KT^2\cdot\frac{E(s^2)}{E(\overline{y}^2)} \tag{23}$$

If the characteristic is normal $N(\mu,\sigma^2)$, the mean-values of sample statistics $\overline{y}^2$ and s^2 are already known and are found in Patel - Read (1996, [15]).

Practical conclusions are the following:

(i) if the process is perfectly centered on the mean-value (which is just the target value), the loss is null (the ideal case);

(ii) the loss cannot be infinitely large: it is more or less significant - depending on the distance of the characteristic's values from its target (this distance may appear also due to the uncertainty in measurements);

(iii) if the variation coefficient is known, the loss-function depends on variability or location - and this fact is a consequence of the particular choice of that loss-function.

4. The use of V for parameter estimation

If we have an arbitrary continuous random variable x with $f(x;\theta),\quad x \in R,\quad \theta = (\theta_1, \theta_2)$ as its density function which involves usually two unknown parameters θ_1 and θ_2, in most of the cases, V - the coefficient variation depends only on one of these parameters (this is not the case of the normal law!).

For instance, if we consider the log-normal law:

$$x : f(x;\mu,\sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-(\ln x - \mu)^2 / 2\sigma^2\right],\quad x > 0,\quad \mu,\sigma > 0 \tag{24}$$

since $E(x) = \exp(\mu + \sigma^2 / 2)$ and $Var(x) = \left(e^{\sigma^2} - 1\right) \cdot \exp(\mu + \sigma^2 / 2)$

we have

$$V = \frac{\sqrt{Var(x)}}{E(x)} = \sqrt{e^{\sigma^2} - 1} \tag{25}$$

which depends only on (σ).

If we take X as a modified Gamma variable (see Dorin et al., 1994, page 110):

$$X : f(x;\theta,k) = \left(\frac{k}{\theta}\right) \frac{x^{k-1}}{\Gamma(k)} \exp - (kx/\theta),\quad x \ge 0,\quad \theta,k > 0 \tag{26}$$

where

$\Gamma(k) = \int_0^\infty u^{k-1} e^{-u} du$ (Gamma function)

we have $E(x) = \theta,\quad Var(x) = \theta^2 / k$ and hence $V = 1/\sqrt{k}$.

This property - namely the dependence on only one of the distribution's parameters - allows the use of V to estimate both these parameters, as follows:

(i) tabulate the quantity $V = g(k)$ - as in the previous case $V = 1/\sqrt{k}$ (Gamma) or $V = g(\sigma) = \sqrt{e^{\sigma^2} - 1}$ (log - normal) a.s.o. for a suitable range of values for k, σ, etc.;

(ii) draw a random sample $x_1, x_2, \ldots, x_n$ from X population and compute the sample coefficient of variation $\hat{V} = s / \overline{x}$;

(iii) search in the table, the obtained value of $\hat{V}$ and read the corresponding value of k or σ or whatever parameter may be there: this value is the estimation of σ for instance in log - normal law. - say $\hat{\sigma}$;

(iv) equating E(x) with the sample mean $\overline{x}$, one obtains the relationship:

$$\overline{x} = \exp\{\mu + \hat{\sigma}^2 / 2\} \tag{27}$$

where from an estimation of μ is extracted:

$$\hat{\mu} = \ln \overline{x} - \sigma^2 / 2 \tag{28}$$

The procedure is valid for any distribution for which V depends only on one parameter: Weibull, Gama, Alpha one some examples. Normal law as we know does not fulfill this request since $V = \sigma / \mu$ and Beta variable is another case.

References

1. Alexis, J. **Metoda Taguchi în practica industriala. Planuri de experiente (translation from French),** Editura TEHNICA, Bucuresti, Colectia MQM, 1999
2. Barsan-Pipu, N., Isaic-Maniu, Al. and Voda, V. Gh. **Defectarea. Modele statistice cu aplicatii**, Editura ECONOMICA, Bucuresti, 1999
3. David, F. N. **Note on the application of Fisher's K-statistics,** Biometrika, vol. 36, pag. 389 – 393, 1949
4. David, M., A. **Gini's mean difference rediscovered,** Biometrika, vol. 36, pag. 232 – 240, 1968
5. David, H. A. and Edwards, A.W.F. **Annotated Readings in the History of Statistics,** Springer Verlag, Berlin-New York, 2001
6. Dorin, Al. C., Isaic-Maniu, Al. and Voda, V. Gh. **Probleme statistice ale fiabilitatii,** Editura ECONOMICA, Bucuresti, 1994
7. Gini, C. **Variabilità e Mutabilitá,** Tipografia di Paolo Cuppini, Bologna, 1912
8. Hald, A. **Statistical Theory with Engineering Applications,** John Wiley and Sons, Inc., New York, 1952
9. Iglewicz, B., Myers, R.H. and Howe, R. B. **On the percentage points of the sample coefficient of variation,** Biometrika, vol. 55, pag. 580 – 581, 1968
10. Iglewicz, B. and Myers, R.H. **Comparisons of approximations to the percentage points of the sample coeficient of variation,** Technometrics, Vol. 12, pag. 166-169, 1970
11. Ivan, C. and Văduva, I. **Asupra repartiției coeficientului de variabilitate,** Stud. Cerc. Mat. tom 21, nr. 7, pag. 1047 – 1062, 1969
12. Johnson, N. L. and Welch, B. L. **Applications of the non-central t distribution,** Biometrika, vol. 31, pag. 362 – 389, 1940
13. McKay, A. **Distribution of the coefficient of variation and the extended „t" distribution,** J. Roy. Statist. So. vol. 95, pag. 695 – 698, 1932
14. Morgan, J. **The anatomy of income distribution,** The Review of Economics and Statistics, vol. 44, pag 270 – 283, 1962
15. Patel, J. K. and Read, C. B. **Handbook of the Normal Distribution (second edition, revised and expanded),** Marcel Dekker Inc., New York, 1996
16. Stoichițoiu, D. G. and Vodă, V. Gh. **The same mean-value and variance - some theoretical and practical consequences in product quality and realiability analysis,** Proceedngs of the 8th Int. Conf. „Quality, Reliability Mainteinablity", edited by Eurocer Building, SRAC, 2002
17. Treyer, V. N. **Vseobschye termofluctuantzyonnye osnovy kineticheskoy teorii dolgovechnosty i nadežnosty mashin i priborov,** STAQUAREL ,76, Prague, vol. I, pag. 261 – 268, 1976
18. Tziafetas, G. N. **A formula for the Gini coefficient and its decomposition,** Biometrical Journal (Akad. Verlag – Berlin), vol. 31, nr. 9, pag. 961 – 967, 1989
19. Warren, W. G. **On the adequacy of the chi-squared approximations for the coefficient of variation,** Communications in Statistics, Ser. B., vol. 11, pag. 659 – 666, 1982
20. Zitek, Fr. **O pewnych estymatorach odchylenia standardowego,** Zastosowania Matematyki (Warszawa), vol. I, nr. 4, pag. 342 – 353, 1954