An alternative to the variation coefficient

Carlo Bertoluzza[1], Rosa Casals[2], Gloria Naval[3], and Antonia Salas[4]

Abstract The aim of this paper is to introduce an invariant by translation coefficient different from the variation one (widely used in literature but not fulfilling that property) that allows us to study whether the mean is a good representation of the distribution or not. The value of this new coefficient for a normally distributed random variable is obtained in order to establish a criterion, similar to the one used in the symmetry or kurtosis coefficients, to decide the grade of representation of the mean.

1 Introduction

Why defining an alternative coefficient?

Variation coefficient is widely used in literature (see for example [1], [3]) in order to obtain a grade of representation of the mean for different distributions, since it is a relative dispersion measure providing the number of times that the standard deviation is contained in the mean of the corresponding distribution. However, this coefficient is not invariant by translations which, in our opinion, it is a quite significant issue. That is why we consider necessary to define a coefficient that evaluates the spread or distance of the values of the distribution with respect to a central one (allowing us to measure the grade of representation of this central value as a numerical summary of that distribution) that is invariant by changes in the origin of the distribution.

In order to show what we mean, let us introduce a numerical example.

Dipartimento di Informatica e Sistemistica, Universitá di Pavia (Italia) `retired Professor` · Dpto. Estadística e I.O. y D.M., Universidad de Oviedo `rmcasals@uniovi.es` · Dpto. Estadística e I.O. y D.M., Universidad de Oviedo `glorianaval@uniovi.es` · Dpto. Estadística e I.O. y D.M., Universidad de Oviedo `antonia@uniovi.es`

Example 1. Let us consider the height of a group of 10 individuals given in three different ways: X *(height in meters)*; $Y = 100X$ *(height in centimeters)*; $Z = Y - 100$ *(centimeters above one meter)*. The observed values, expressed in the three alternative ways, are shown in table 1.

Table 1 Data example 1

X	1.7	1.85	1.72	1.65	1.73	1.58	1.65	1.69	1.6	1.67
Y	170	185	172	165	173	158	165	169	160	167
Z	70	85	72	65	73	58	65	69	60	67

Being the same data should provide us with exactly the same goodness of representation of the mean as a numerical summary of this distribution, independently of the units (X *vs* Y) or the referential origin (X *vs* Z) considered to describe them. The variation coefficient for variable X is

$$\overline{x} = 1.684 \quad S_x^2 = 0.005164 \quad \Rightarrow \quad VC(X) = \frac{\sqrt{0.005164}}{1.684} = 0.042672789$$

and for variable Y is

$$\overline{y} = 168.4 \quad S_y^2 = 51.64 \quad \Rightarrow \quad VC(Y) = \frac{\sqrt{51.64}}{168.4} = 0.042672789$$

which means that a scale change does not affect it. However

$$\overline{z} = 68.4 \quad S_z^2 = 51.64 \quad \Rightarrow \quad VC(Z) = \frac{\sqrt{51.64}}{68.4} = 0.105059908$$

so translations do affect the value of this coefficient. This representation of the mean of Z can be understood as a worse one than that of X since the variation coefficient is greater for Z than for X. This coefficient can undergo dramatic changes for different expressions of the same set of values. Let us consider a new variable $T = Y - 170$ that provides the centimeters above/below 170. Despite the sign, the value of the variation coefficient is much greater for T than for Y

$$\bar{t} = -1.6 \quad S_t^2 = 51.64 \quad \Rightarrow \quad VC(T) = \frac{\sqrt{51.64}}{-1.6} = -4.49131106$$

This means, with the interpretation broadly given to this coefficient in literature, that the grade of representation of -1.6 as a numerical summary of the average difference between the height (in centimeters) of those individuals and 170 is significantly smaller than that of 168.4 centimeters as the average height of the same group of individuals. The same holds with 68.4 centimeters exceeding the meter as a representation of these data *(T vs Z)*. This makes no sense for us, since the information given by these distributions is intrinsically the same, although the values differ from one variable to another.

Example 2. Let us consider a normally distributed random variable $X \sim N(\mu, \sigma)$. Assume that we are working with a given value of the variance (say 1 for sake of simplicity). The bell-shape density function that characterizes the distribution is exactly the same, independently of the location of the mean. Hence, a constant value for a coefficient giving the goodness of representation of the mean was to be expected. However, as $VC(X) = \frac{SD(X)}{E(X)}$, that is, $VC(X) = \frac{\sigma}{\mu}$, this coefficient increases as the mean decreases. That is the same as saying about the mean that the closer to zero, the less representative of the distribution.

2 Definition of a new coefficient

Is it possible to express how spread out the values of a distribution are with respect to a central value, and, consequently, how good this central value is as a representation of the distribution, in such a way that translation changes do not affect that goodness of representation (as it happens with the variation coefficient)?

We may agree that moving the values to another location keeping static the "distances" between them and their relative positions between each other should not affect the representation of the mean as a numerical summary of the data, since the dispersion of the data is exactly the same in both locations, being the value of the referential point (mean) the only noticeable modification. "Moving the values to another location" has modified the mean, while "keeping static their relative positions between each other" does not change the variance of the distribution. So, as the variance averages the squared distances of the data with respect to their mean, why not introducing some invariant by translations measure in the denominator of the "variation coefficient" so that the value it takes is exactly the same wherever the data are located? Doing so, the corresponding value of the ratio will show how good the mean is to summarize the values of a distribution just paying attention to the relative positions within the data but not to the actual location with respect to a given origin.

Definition 1. For any variable X, the ***representation coefficient of X*** is the ratio between its standard deviation and its range, that is

$$RC(X) = \frac{D(X)}{R(X)} = \frac{D(X)}{max(X) \text{ - } min(X)}$$

This coefficient is well defined for any nondegenerate variable X and takes nonnegative values smaller than 1. It is quite easy to see that the closer to one, the less representative the mean.

A more general representation coefficient

Nevertheless, outliers can strongly affect the value of this coefficient, so it may be better to exclude that part of the distribution on calculating a grade of the representation of the mean. We intend to present a coefficient which is not so sensitive to the presence of extreme values of the distribution. There are different ways of detecting outliers but, regardless of the method, any value that is really far from the rest of the observations is said to be an outlier. For instance, $\mu \pm 2.7\sigma$ are the outliers cutoffs given in [4] when working on a normal distribution.

For any $r \in (0,1)$ let $P_{100r}(X)$ denote the $100r-th$ percentile of X, that is, $r = P(X \leq P_{100r}(X))$

Definition 2. The **$100r$% trimmed representation coefficient of X** is defined as

$$RC_{100r}(X) = \frac{D(X)}{P_{100(1-\frac{r}{2})}(X) - P_{\frac{100r}{2}}(X)} \ \forall r \in [0, 0.5]$$

where $P_0(X) = min(X)$ and $P_{100}(X) = max(X)$ (so this coefficient is a generalization of the latter one).

Remark 1. This coefficient is generally well defined. Once again there are some exceptions, since the denominator is zero for any value of r when X is degenerate. It is also zero for not so large values of r when X doesn't take too many different values (then its "extreme" percentiles coincide, that is, it is "almost sure" degenerate; median and mode will coincide in that case and the mean is rarely going to be chosen as a representation of such a variable).

Remark 2. In general, any of these trimmed representation coefficients compares the value of a dispersion measure with respect to the mean with a dispersion measure which does not refer to any particular central value and that does not take into account outliers (to the extent the experimenter wants to). Then, it can be used to measure how good the mean is as a numerical summary of the distribution.

Remark 3. Although it can be defined for any value of $r \in [0,1]$, it is not sensitive to eliminate too many "outliers"; so, it makes not much sense to calculate it for $r > 0.5$. For instance, the $100r$% trimmed representation coefficient equals the 100(1-r)% one in absolute value.

Property 1. **(Invariance by translations).** The trimmed representation coefficient is invariant by translations, that is, $RC_{100r}(X+k) = RC_{100r}(X)\ \forall r \in [0,1]$ and $\forall k \in \mathbb{R}$.

Proof. Trivial since $D(X+k) = D(X)\ \forall k$ and $P(X+k \leq y) = P(X \leq y-k)\ \forall y \ \Rightarrow \ P_{100r}(X+k) = k + P_{100r}(X)$. Hence $P_{100(1-\frac{r}{2})}(X+k) - P_{\frac{100r}{2}}(X+k) = P_{100(1-\frac{r}{2})}(X) - P_{\frac{100r}{2}}(X)$ □

Property 2. **(Absolute invariance by scale).** The trimmed representation coefficient is invariant by change of scale in absolute value, that is, $|RC_{100r}(kX)| = |RC_{100r}(X)|\ \forall r \in [0,1]$ and $\forall k \in \mathbb{R}$.

Proof. Trivial since *D(kX)= kD(X)* $\forall k$. On the other hand, $P(kX \leq y) = P(X \leq \frac{y}{k})\ \forall y\ \forall k > 0 \ \Rightarrow\ P_{100r}(kX) = kP_{100r}(X)$ and $P(kX \leq y) = P(X \geq \frac{y}{k})\,\forall y\ \forall k < 0 \ \Rightarrow\ P_{100r}(kX) = kP_{100(1-r)}(X)$. Hence $|P_{100(1-\frac{r}{2})}(kX) - P_{\frac{100r}{2}}(kX)| = |k|\ |P_{100(1-\frac{r}{2})}(X) - P_{\frac{100r}{2}}(X)|$ □

Example 1 (continued). For the data of the height used above, one can obtain

$$RC_{10}(X) = \frac{\sqrt{0.005164}}{1.85 - 1.58} = 0.266151766 \quad RC_{10}(Y) = \frac{\sqrt{51.64}}{185 - 158} = 0.266151766$$

$$RC_{10}(Z) = \frac{\sqrt{51.64}}{85 - 58} = 0.266151766 \quad RC_{10}(T) = \frac{\sqrt{51.64}}{15 - (-12)} = 0.266151766$$

Analogously, $RC_{50}(X) = \frac{\sqrt{0.005164}}{1.72-1.65} = RC_{50}(Y) = RC_{50}(Z) = RC_{50}(T)$. Invariance holds for any r.

3 The trimmed representation coefficient for a normally distributed random variable

Invariance of the trimmed coefficient with respect to the parameters

Let $X \sim N(0,1)$ and let $P_{100r}(X)$ be the $100r$-*th* percentile of X. Then for $Y \sim N(\mu, \sigma)$

$$\begin{aligned} r = P(Y \leq P_{100r}(Y)) &= P(\frac{Y-\mu}{\sigma} \leq \frac{P_{100r}(Y)-\mu}{\sigma}) \\ \Rightarrow\ P_{100r}(Y) &= \mu + \sigma P_{100r}(X) \end{aligned}$$

which means that the $100r\%$ trimmed representation coefficient for a normal distribution does not depend on its parameters.

$$\begin{aligned} RC_{100r}(N(\mu,\sigma)) &= \frac{\sigma}{(\mu + \sigma P_{100(1-\frac{r}{2})}(X)) - (\mu + \sigma P_{\frac{100r}{2}}(X))} \\ &= \frac{1}{P_{100(1-\frac{r}{2})}(X) - P_{\frac{100r}{2}}(X)} = RC_{100r}(N(0,1)) \end{aligned}$$

Value of the trimmed representation coefficient for some usual 100r%

Let us use R for obtaining the corresponding percentiles of the normal distribution. As we have just proved, we can reduce our calculations to the standard normal.

The $5th$ and $95th$ percentiles of $N(0,1)$ are

```
> qnorm(c(.95,.05), mean=0, sd=1, lower.tail=TRUE)
[1] 1.644854 -1.644854
```

so the 10% trimmed representation coefficient of $N(0,1)$ takes value

```
> 1.644854+1.644854
[1] 3.289708
> 1/3.289708
[1] 0.3039783
```

Proceeding in an analogous way we can obtain other percentages the values shown in table 2.

Table 2 Values of the trimmed representation coefficient for the standard normal distribution

$100r\%$	1%	2%	2.5%	5%	10%	20%	50%
RC_{100r}	0.142857	0.214929	0.2230746	0.2551067	0.3039783	0.3901519	0.7413008

As it can be seen, this coefficient takes quite similar values for $r \in [0.02, 0.05]$ which are the most usual percentages for outliers cutoffs in literature.

4 Interpretation of the trimmed representation coefficient

The value of this coefficient for the normal distribution can be used as a reference to determine whether the mean as a representation of the distribution is good enough or not, analogously as, for example, the kurtosis coefficient $\gamma_2(Y) = \frac{\mu_4(Y)}{[D(Y)]^4} - 3$ greater than 0 means that the distribution is sharper than the bell-shaped one, with $\mu_4(Y)$ being the $4th$ central moment of Y.

Property 3. The closer to zero the (trimmed) representation coefficient of X, the better its mean as a representation of the distribution of X.

Definition 3. The mean of a variable X is said to be more representative than the mean of the normal distribution for a $100r\%$ if $RC_{100r}(X) \leq RC_{100r}(N(0,1))$.

As we have just obtained, the trimmed representation coefficient of a normal distribution increases as the order (r) increases. So, the fewer values included in the interval of reference of the denominator, the larger the coefficient.

Whenever $RC_{100r}(X) > 0.7413008$ it can be said that the mean of X is not representative for the distribution since the central 50% of the observations of a distribution is the minimal set of values to be considered on studying their representation by means of a central value.

Example 1 (continued). As we have just obtained, $RC_{10}(X) = 0.266151766 < 0.3039783 = RC_{10}(N(0,1))$, so, the mean of these data is a better representation of the distribution than that of the Gauss one. On the other side, if one pay attention just to the data included in the box of a box-whiskers diagram, one can conclude that those data are more disperse than those from a normal distribution since $RC_{50}(X) = 1.026585384 > 0.7413008 = RC_{50}(N(0,1))$.

5 Trimmed representation coefficient for other theoretical distributions

Exponential distribution

Let us consider $f(x) = \lambda e^{-\lambda x}$ $x > 0$ the density function of an exponential random variable X.

Property 4. The trimmed representation coefficient of an exponentially distributed random variable does not depend on the value of its parameter.

Proof. On calculating percentiles for an exponentially distributed random variable X of parameter λ, one obtain

$$r = P(X < P_{100r}) = \int_0^{P_{100r}} \lambda e^{-\lambda x} dx = 1 - e^{-\lambda P_{100r}} \Rightarrow P_{100r} = -\frac{ln(1-r)}{\lambda}$$

Hence

$$RC_{100r}(Exp(\lambda)) = \frac{\frac{1}{\lambda}}{\frac{1}{\lambda}\ln(1-\frac{r}{2}) - \frac{1}{\lambda}\ln(\frac{r}{2})} = \frac{1}{\ln(1-\frac{r}{2}) - \ln(\frac{r}{2})}$$

□

As we have proved that the value of the parameter is not essential for calculating the trimmed representation coefficient, for the sake of simplicity we will use R and calculate the coefficient for an exponentially distributed random variable of rate equal to one.

```
> qexp(c(.95,.05,.75,.25), rate=1, lower.tail=TRUE)
[1] 2.99573227 0.05129329 1.38629436 0.28768207
> 1/(2.99573227-0.05129329)
[1] 0.3396233
> 1/(1.38629436-0.28768207)
[1] 0.9102392
> qexp(c(0.9,0.1,0.975,0.025), rate=1, lower.tail=TRUE)
[1] 2.30258509 0.10536052 3.68887945 0.02531781
> 2.30258509-0.10536052
[1] 2.197225
> 1/2.197225
[1] 0.4551195
> 3.68887945-0.02531781
[1] 3.663562
> 1/3.663562
[1] 0.2729584
> qexp(c(0.99,0.01,0.995,0.005), rate=1, lower.tail=TRUE)
[1] 4.605170186 0.010050336 5.298317367 0.005012542
> 4.605170186-0.010050336
[1] 4.59512
> 1/4.59512
[1] 0.2176222
> 5.298317367-0.005012542
[1] 5.293305
> 1/5.293305
[1] 0.1889179
```

The values of the trimmed representation coefficient for an exponential distribution are shown in table 3. So, we can conclude that the mean of an exponential distribution is slightly less representative of its distribution than that of the normal one.

Table 3 Values of the trimmed representation coefficient for the exponential distribution with mean one

$100r\%$	1%	2%	5%	10%	20%	50%
RC_{100r}	0.1889179	0.2176222	0.2729584	0.3396233	0.4551195	0.9102392

Uniform distribution

Let us consider $X \sim \mathcal{U}(a, b)$, that is $f(x) = \frac{1}{b-a}$ if $a < x < b$

Property 5. The trimmed representation coefficient of a uniform distributed random variable does not depend on the value of its parameters.

Proof. On calculating percentiles for a uniformly distributed random variable X on the interval *(a,b)*, one obtain

$$r = P(X < P_{100r}) = \int_a^{P_{100r}} \frac{1}{b-a} dx = \frac{P_{100r} - a}{b-a} \;\Rightarrow\; P_{100r} = a + r(b-a)$$

Hence,

$$RC_{100r}(U(a,b)) = \frac{\sqrt{\frac{(b-a)^2}{12}}}{a + (1-\frac{r}{2})(b-a) - a - \frac{r}{2}(b-a)} = \frac{1}{(1-r)\sqrt{12}}$$

□

As we have just obtained that the trimmed representation coefficient of a uniformly distributed random variable depends on the order, we can obtain for any interval *(a,b)*, the values given in table 4.

Table 4 Values of the trimmed representation coefficient for the (continuous) uniform distribution

$100r\%$	1%	2%	5%	10%	20%	50%
RC_{100r}	0.291591045	0.294566463	0.303868562	0.320750149	0.360843918	0.577350269

So, whenever at most 10% of the extreme observations of a distribution are excluded, the mean is said to be a better representation of the data of a normal distribution than of a uniform one.

Some well-known discrete distributions

Finally, let us use R in order to obtain some percentiles from a Poisson of mean 1 and also from a Binomial with 10 trials and p = *0.5*. The values obtained are shown below, and summarized in table 5.

```
> qpois(c(0.25,.75), lambda=1, lower.tail=TRUE)
[1] 0 2
> qpois(c(0.1,.9), lambda=1, lower.tail=TRUE)
[1] 0 2
> qpois(c(0.05,.95), lambda=1, lower.tail=TRUE)
[1] 0 3
> qpois(c(0.025,.975), lambda=1, lower.tail=TRUE)
[1] 0 3
```

```
> qpois(c(0.01,.99), lambda=1, lower.tail=TRUE)
[1] 0 4
> qpois(c(0.005,.995), lambda=1, lower.tail=TRUE)
[1] 0 4
> qbinom(c(.005,.995), size=10, prob=0.5, lower.tail=TRUE)
[1] 1 9
> qbinom(c(.01,.99), size=10, prob=0.5, lower.tail=TRUE)
[1] 1 9
> qbinom(c(.025,.975), size=10, prob=0.5, lower.tail=TRUE)
[1] 2 8
> qbinom(c(.05,.95), size=10, prob=0.5, lower.tail=TRUE)
[1] 2 8
> qbinom(c(.1,.9), size=10, prob=0.5, lower.tail=TRUE)
[1] 3 7
> qbinom(c(.25,.75), size=10, prob=0.5, lower.tail=TRUE)
[1] 4 6
```

Table 5 Values of the trimmed representation coefficient for some discrete distributions

$100r\%$	1%	2%	5%	10%	20%
$RC_{100r}(\mathcal{P}(1))$	0.25	0.25	$\frac{1}{3}$	$\frac{1}{3}$	0.5
$RC_{100r}(\mathcal{B}(10,0.5))$	0.197642353	0.197642353	0.263523138	0.3952844707	0.790569415

All these values are larger than the corresponding ones for the normal distribution, so these variables are worse represented by their means than the normal one.

References

1. Rodríguez Muñiz, LJ *et al* (2011) Métodos Estadísticos para Ingeniería. Garceta Grupo Editorial, Madrid
2. Peña, D (2002) Análisis de datos multivariantes. McGrawHill/Interamericana de España, Madrid
3. McPherson, G (1990) Statistics in Scientific Investigation. Its Basis, Application and Interpretation. Springer, New York
4. Jobson, JD (1991) Applied Multivariate Data Analysis, Vol. I: Regression and Experimental Design. Springer, New York