Journal of Mathematics and Statistics 6 (1): 47-51, 2010
ISSN 1549-3644
© 2010 Science Publications

# Boundary Distributions with Respect to Chebyshev's Inequality

[1]Peter Bias, [2]Shawn Hedman and [2]David Rose
[1]Department of Economics,
[2]Department of Mathematics,
Florida Southern College, USA

**Abstract:** Variables whose distributions achieve the boundary value of Chebyshev's inequality are characterized and it is found that non-constant variables with this property are symmetric discrete with at most three values. Nevertheless, the bound of Chebyshev's inequality remains optimal for the class of continuous variables.

**Key words:** Chebyshev's inequality, k-boundary variable, k-condensed variable, nearly k-boundary variable

## INTRODUCTION

The familiar inequality of Chebyshev encountered in most elementary statistics courses affirms that for every random variable x with mean $\mu_x$ and standard deviation $\sigma_x$ and for every $k \geq 1$, the probability that x lies strictly within k standard deviations of the mean is at least $1-\frac{1}{k^2}$. Equivalently:

$$P\left(\left|x-\mu_x\right| \geq k\sigma_x\right) \leq \frac{1}{k^2}$$

The inequality is trivial if $k \geq 1$ so a proof for k>1 suffices. Elegant short proofs of this inequality can be found in standard texts such as Hogg and Craig (1995) and Larson and Farber (2009). The proof given below and found in (Rui, 1973) does not require separate cases for discrete and continuous variables.

**Proof:** Define a new variable by:

$$y = \begin{Bmatrix} k\sigma_{x,}^2 & \text{if } \left|x-\mu_x\right| \geq k\sigma_x \\ 0, & \text{if } \left|x-\mu_x\right| < k\sigma_x \end{Bmatrix}$$

Then:

$$\begin{aligned} & y \leq (x-\mu_x)^2 \Rightarrow E(y) = \sum yP(y) = \\ & k^2\sigma_x^2(\left|x-\mu_x\right| \geq k\sigma_x) \leq E((x-\mu_x)^2) = \sigma_x^2 \end{aligned}$$

Dividing by $k^2\sigma_x^2$ yields the result.

To show that the bound in Chebyshev's inequality cannot be improved, the following example is provided in (Hogg and Craig, 1995).

**Example 1:** Let x have the following distribution:

| x | P(x) |
|---|---|
| $-1$ | $\frac{1}{8}$ |
| 0 | $\frac{3}{4}$ |
| 1 | $\frac{1}{8}$ |

It is evident (by symmetry) that $\mu_x = 0$ and that $\sigma_x^2 = \sum x^2P(x) = \frac{1}{4}$ so that $\sigma_x = \frac{1}{2}$. Thus, for k = 2, $k\sigma_x = 1$ and $P(\left|x-\mu_x\right| < k\sigma_x) = P(\left|x\right| < 1) = \frac{3}{4} = 1-\frac{1}{2^2}$.

The following questions arise:

- What other variables might satisfy Chebyshev's boundary condition?
- Must all such boundary variables have unimodal distributions as in the example above?
- Can infinitely many distributions satisfy the boundary condition for the same k?

These questions are answered in the next section where k-boundary variables are introduced.

**Corresponding Author:** Peter Bias, Department of Economics, Florida Southern College, USA

## k--BOUNDARY VARIABLES

**Definition 2:** For a fixed $k \geq 1,$, a variable x is k-boundary if $P(|x-\mu_x| \geq k\sigma_x) = \frac{1}{k^2}$.

The 2-boundary variable given above has a unimodal distribution. We now give an example of a uniform k-boundary variable suggesting that the shape of a k-boundary variable may depend upon the value of k.

**Example 3:** Let $k = \frac{\sqrt{6}}{2}$ and let x have the given distribution:

| x | P(x) |
|---|---|
| $-1$ | $\frac{1}{3}$ |
| $0$ | $\frac{1}{3}$ |
| $1$ | $\frac{1}{3}$ |

By symmetry, $\mu_x = 0 \Rightarrow \sigma_x^2 = \sum x^2 P(x) = \frac{2}{3} \Rightarrow \sigma_x = \frac{\sqrt{6}}{3}$. Thus, $k\sigma_x = 1$ and $P(|x-\mu_x| \geq 1) = \frac{2}{3} = \frac{1}{\frac{3}{2}} = \frac{1}{k^2}$.

The following example shows that Chebyshev's bound cannot be improved for any $k \geq 1$.

**Example 4:** Let $k \geq 1$. and let x $x$ have the following distribution:

| x | P(x) |
|---|---|
| $-1$ | $\frac{1}{2k^2}$ |
| $0$ | $1-\frac{1}{k^2}$ |
| $1$ | $\frac{1}{2k^2}$ |

Then, $\mu_x = 0, \sigma_x = \frac{1}{k}$ and $P(|x-\mu_x| < k\sigma_x) = P(|x| < 1) = 1 - \frac{1}{k^2}$. If k = 1, we may assume that the variable is two-valued since then P(0) = 0.

In this general example the distributions are symmetric and U-shaped and hence bimodal for $0 < k < \frac{\sqrt{6}}{2}$ and are symmetric and cap-shaped and thus unimodal for $k < \frac{\sqrt{6}}{2}$. These shapes remain unchanged under linear transformation of the variable. It is a consequence of the Lemma below that the k-boundary condition is invariant under linear transformation, confirming that indeed, infinitely many examples of k-boundary variables exist for each $k \geq 1$. In particular, z-transforms of k-boundary variables yield standard k-boundary variables. Our characterization of k-boundary variables will follow from the fact that there is a unique non-constant standard k-boundary variable for each $k \geq 1$. Constant variables and those with infinite standard deviation must be separately considered because they do not have a z-transform. We note that constant variables are 1-boundary since they have standard deviation 0 and variables with infinite standard deviation cannot be k-boundary for any finite k. It follows that non-constant k-boundary variables for $k \geq 1$ have z-transforms. We continue with the simple but important probability invariance for linear transforms.

**Lemma 5:** If $y = mx + b$ with $m \neq 0$, then for each $k > 0, P(|x-\mu_x| < k\sigma_x) = P(|y-\mu_y| < k\sigma_y)$.

**Proof:** For the linear transform y of x, y = mx+b, $\mu_y = m\mu_x + b$ and $\sigma_y^2 = \sigma_{mx+b}^2 = \sigma_{mx}^2 = m^2\sigma_x^2 \Rightarrow \sigma_y = |m|\sigma_x$. Now, $P(|y-\mu_y| < k\sigma_y) = P(|mx - m\mu_x| < k|m|\sigma_x) = P(|m||x-\mu_x| < |m|k\sigma_x) = P(|x-\mu_x| < k\sigma_x)$.

It follows immediately that any linear transform of a k-boundary variable is also a k-boundary variable and in particular, linear transforms of the two and three-valued discrete k-boundary variables exhibited above are k-boundary variables. In fact, the converse is true. Every non-constant k-boundary variable is a linear transform of one of the variables from the example above. For each k>1, every variable with a k-boundary distribution is a linear transform of the variable whose k-boundary distribution is given in Example 4 above. Also, it may be shown that the only 1-boundary variables are the constant variables and the two-valued uniform discrete variables. Evidently, all variables with asymmetric distributions or having at least four values including continuous variables, do not satisfy the boundary condition for Chebyshev's inequality. To establish this claim we first show that for each k≥1, the standard non-constant k-boundary variable is unique.

**Theorem 6:** For each k≥1, there is a unique non-constant k-boundary variable z having $\mu_z = 0$ and $\sigma_z = 1$.

**Proof:** For each k≥1, the existence of such a standard k-boundary variable has already been established since z-transforms are linear transforms and hence preserve the k-boundary condition. For the uniqueness, we consider the discrete and continuous variable cases separately. Firstly, assume that z is a non-constant discrete k-boundary variable for some k≥1 with $\mu_z = 0$ and $\sigma_z = 1$. Then:

$$1 = \sigma_z^2 = \sum z^2 P(z) = \sum_{|z|<k} z^2 P(z) + \sum_{|z|<k} z^2 P(z) \geq \sum_{|z|<k} z^2 P(z) +$$

$$k^2 \sum_{|z|\geq k} P(z) = \sum_{|z|<k} z^2 P(z) + k^2\left(\frac{1}{k^2}\right) = \sum_{|z|<k} z^2 P(z) + 1 \geq 1$$

It follows that $\sum_{|z|<k} z^2 P(z) = 0$ so that in particular, $0 < |z| < k \Rightarrow P(z) = 0.$ Also then, $1 = \sum_{|z|\geq k} z^2 P(z) = \sum_{|z|\geq k} k^2 P(z)$ implies that $P(z = 0)$ if $|z| > k$. So, the only possible values for z are –k, 0 and k. But, 0 cannot be the only value since $\sigma_z = 1 > 0$. And, since $\mu_z = 0$, both values-k and k must be assumed. Further, the distribution must be symmetric since:

$$0 = \mu_z = -kP(-k) + 0p(0) + kP(k) = k(P(k) - P(-k))$$

with $k > 0 \Rightarrow P(-k) = P(k)$.

Then, $1 = \sigma_z^2 = (-k)^2 P(-k) + 0^2 P(0) + k^2 P(k) = 2k^2 P(k)$ implies that $P(k) = P(-k) = \frac{1}{2k^2}$. It also follows that $P(0) = 1 - \frac{1}{k^2}$. Further, if k = 1, -k and k are the only values of z each with probability $\frac{1}{2}$.

We now show that no continuous variable has a k-boundary distribution. To evoke a contradiction suppose that k≥1 and z is a continuous variable with k-boundary distribution and that $\mu_z = 0$ and $\sigma_z = 1$. Let f(z) be the probability density function for z and assume that f has at most finitely many points of essential discontinuity. (Less stringent assumptions on f could be made.) Firstly:

$$1 = \sigma_z^2 = \int_{+\infty}^{+\infty} z^2 f(z)dz = \int_{|z|\leq k} z^2 f(z)dz + \int_{|z|\geq k} z^2 f(z)dz \geq$$

$$\int_{|z|\leq k} z^2 f(z)dz + k^2 \int_{|z|\geq k} f(z)dz = \int_{|z|\leq k} z^2 f(z)dz + k^2\left(\frac{1}{k^2}\right) =$$

$$\int_{|z|\leq k} z^2 f(z)dz + 1 \geq 1 \Rightarrow \int_{|z|\leq k} z^2 f(z)dz = 0 \Rightarrow f(z) = 0$$

for $|z| \leq k$ except on a finite set (or a set of measure 0) $\Rightarrow \int_{|z|\leq k} f(z)dz = 0$. Then, $\int_{|z|\geq k} f(z)dz \Rightarrow f(z) - 0$ $1 = \int_{|z|\geq k} z^2 f(z)dz = \int_z f(z)dz =$ for $|z| \geq k$ except at finitely many points (or a set of measure $0) \Rightarrow 1 = \int_{|z|\geq k} f(z)dz = 0$. This contradiction shows that continuous k-boundary variables do not exist.

**Corollary 7:** If k>1 and x and y are each k-boundary variables, then y = mx + b for some m>0 and for some b.

**Proof:** By the uniqueness of standard k-boundary variables, x and y have the same z-transform. In particular, $\frac{x - \mu_x}{\sigma_x} = \frac{y - \mu_y}{\sigma_y} \Rightarrow y = mx + b$ with $m = \frac{\sigma_y}{\sigma_x} > 0$ and $b = \frac{1}{\sigma_x}\begin{vmatrix} \sigma_x & \sigma_y \\ \mu_x & \mu_y \end{vmatrix}$.

Technically, the corollary fails for k = 1 since then, x could be a constant variable while y is a two-valued variable. It also follows from the theorem above that for k>1, the distribution for each k-boundary variable is uniquely determined by its shape.

Now that k-boundary distributions have been completely determined, in the next section we briefly consider the dual problem. To what extent can the bound in Chebyshev's inequality be wrong for a given distribution? In other words, for which distributions will 100% of a population lie within k standard deviations of the mean?

## k--CONDENSED VARIABLES

**Definition 8:** For k>0, a variable x is k-condensed if $P(|x - \mu_x| < k\sigma_x) = 1$, or equivalently, if $P(|x - \mu_x| \geq k\sigma_x) = 0$.

Firstly, there are variables having 100% of all values within k standard deviations for each k>1. In particular, if x is the symmetric variable with two values -1 and 1, then $\mu_x = 0$ and $\sigma_x = 1$, so that $P(|x - \mu_x| < k\sigma_x) = P(|x| \leq 1) = 1$. The case for k = 1 is different. We leave it to the reader to show that it is impossible for any variable with finite variance to have all values strictly within one standard deviation from the mean. However, in this case, the probability that a variable is strictly within even a fraction of a standard deviation of the mean can be arbitrarily close to 1.

Consider the sequence of variables $x_n$ with distributions:

| $x_n$ | $P(x_n)$ |
|---|---|
| $-1$ | $\frac{1}{2n}$ |
| $0$ | $1-\frac{1}{n}$ |
| $1$ | $\frac{1}{2n}$ |

For each $n = 1, 2..., \mu_x = 0$ and $\sigma_{x_n} = \frac{1}{\sqrt{n}}$ so that for all $0 < k \leq 1, P_n = P(|x_n - \mu_{x_n}| < k\sigma_{x_n}) = P(|x_n| < \frac{k}{\sqrt{n}}) = 1 - \frac{1}{n}$. Clearly, for all $0 < k \leq 1, \lim_{n\to\infty} P_n = 1$. Remarkably, also $\lim_{n\to\infty} \sigma_{x_n} = 0$. In some sense, each variable $x_n$ is nearly 1-condensed.

On the other hand, allowing the pathology of infinite standard deviation, yields the existence of k-condensed variables for all $0 < k \leq 1$.

**Example 9:** Let x be the continuous variable whose probability density function is given by:

$$f(x) = \begin{cases} -\frac{1}{x^3} & \text{if } x \leq -1 \\ 0 & \text{if } -1 < x < 1 \\ \frac{1}{x^3} & \text{if } 1 \leq x \end{cases}$$

Then:

$$\int_{+\infty}^{+\infty} f(x)dx = 1, \mu_x = 0$$

and

$$\sigma_x^2 = \int_{+\infty}^{+\infty} x^2 f(x)dx = 2\lim_{t\to\infty}\int_1^{+\infty} \frac{1}{x}dx = \lim_{t\to\infty} \ln(t^2) = +\infty$$

Thus, $\sigma_x = +\infty$ and $P(|x - \mu_x| < k\sigma_x) = P(|x| < +\infty) = 1$ for all $0 < k \leq 1$.

We return now to the question of whether Chebyshev's inequality can be improved if restricted to important subclasses of random variables. Of course it can! A Chebyshev type inequality for the subclass of all continuous uniform variables, or the subclass of all normal variables will have a much higher bound. But, such restriction thwarts the intended universality of Chebyshev's inequality. As a compromise, can the inequality be improved if applied only to the broad class of all continuous variables? The next section is devoted to the demonstration that the bound of Chebyshev's inequality, though never achieved by any member of the class, is still optimal for the class of continuous variables.

## NEARLY k--BOUNDARY VARIABLES

For any fixe d>0 and $k \geq 1$, we want to find a non-constant continuous variable x for which:

$$1 - \frac{1}{k^2} \leq P(|x - \mu_x| < k\sigma_x) \leq \left(1 - \frac{1}{k^2}\right) + d$$

After some experimentation and with the help of the Maple symbolic software program, we considered a family of continuous variables parametrized by a with 0<a<1 whose symmetric continuous probability density functions are piecewise linear with graphs consisting of the x-axis except where three isosceles triangles rise from their bases on the x-axis. The base width of each triangle is 2a and 0 is the center of the base of the central triangle. The area of the central triangle for each variable of the family is $1 - \frac{1}{k^2} + d$ where it is assumed that $d < \frac{1}{k^2}$. The two outer triangles are congruent sharing an area complementary in 1 to that of the central triangle. The outer triangles have bases 1 unit removed from 0. The family of variables $x_a$ with probability density functions $f_a$ as described here are defined as follows. For a fixed $k \geq 1$ and for each 0<a<1, $f_a(x_a) = 0$ except that:

$$f_a(x_a) = \begin{cases} \frac{(1-dk^2)(x_a+1+2a)}{2a^2k^2} & -1-2a \leq x_a \leq -1-a \\ \frac{(dk^2-1)(1+x_a)}{2a^2k^2} & -1-a \leq x_a \leq -1 \\ \frac{[(d+1)k^2-1](a+x_a)}{a^2k^2} & -a \leq x_a \leq 0 \\ \frac{[1-(d+1)k^2](x_a-a)}{a^2k^2} & 0 \leq x_a \leq a \\ \frac{(1-dk^2)(x_a-1)}{2a^2k^2} & 1 \leq x_a \leq 1+a \\ \frac{(dk^2-1)(x_a-1-2a)}{2a^2k^2} & 1+a \leq x_a \leq 1+2a \end{cases}$$

By Maple, $\int_{+\infty}^{+\infty} f_a(x_a)dx_a = 1$ and $\mu_{x_a} = 0$ Also, by Maple:

$$\begin{aligned}\sigma_{x_a}^2 = {} & \frac{(1-dk^2)[(1+a)^4-(1+2a)^4]}{8a^2k^2} + \\ & \frac{(1-dk^2)(1+2a)[(1+2a)^3-(1+a)^3]}{6a^2k^2} + \\ & \frac{(dk^2-1)[1-(1+a)^4]}{8a^2k^2} + \frac{(dk^2-1)[(1+a)^3-1]}{6a^2k^2} + \\ & \frac{[(d+1)k^2-1]a^2}{12k^2} - \frac{[1-(d+1)k^2]a^2}{12k^2} + \\ & \frac{(1-dk^2)[(1+a)^4]}{8a^2k^2} + \frac{(dk^2-1)[(1+a)^3-1]}{6a^2k^2} + \\ & \frac{(dk^2-1)[(1+2a)^4-(1+a)^4]}{8a^2k^2} + \\ & \frac{(dk^2-1)(1+2a)[(1+a)^3-(1+2a)^3]}{6a^2k^2}\end{aligned}$$

Further, with Maple we have that $\lim_{a\to 0}(k\sigma_{x_a}) = \sqrt{1-dk^2}$ which is real and less than 1 for $0 < d \le \frac{1}{k^2}$. Thus, we may choose a>0 so that $k\sigma_{x_a} \le 1$. It follows that for such an $a, P(\left|x_a - \mu_{x_a}\right| < k\sigma_{x_a}) = \left(1-\frac{1}{k^2}\right) + d$, the area under the central triangle. Since d can be chosen arbitrarily near 0, there is no bound larger than $1-\frac{1}{k^2}$ for which Chebyshev's inequality will hold for all continuous variables.

## REFERENCES

Hogg, R.V. and A.T. Craig, 1995. Introduction to Mathematical Statistics. 5th Edn., Prentice Hall, ISBN: 0-02-355722-2, pp: 70.

Larson, R. and B. Farber, 2009. Elementary Statistics Picturing the World. 4th Edn., Prentice Hall, ISBN: 10: 0-13-206290-9, pp: 89.

Rui, Z.Y., 1973. Modern Probability Theory. Harper and Row, ISBN: 06-047327-4, pp: 217.