2017 HAWAII UNIVERSITY INTERNATIONAL CONFERENCES
SCIENCE, TECHNOLOGY & ENGINEERING, ARTS, MATHEMATICS & EDUCATION JUNE 8 - 10, 2017
HAWAII PRINCE HOTEL WAIKIKI, HONOLULU, HAWAII

ON CHEBYSHEV'S INEQUALITY IN ELEMENTARY STATISTICS - AN ORIGINAL PROOF

GARRISON, JOSEPH M.
DEPARTMENT OF MATHEMATICS
MIDDLE GEORGIA UNIVERSITY
COCHRAN, GEORGIA.

Prof. Joseph M Garrison
Department of Mathematics
Middle Georgia State University
Cochran, Georgia.

On Chebyshev's Inequality in Elementary Statistics - An Original Proof

Synopsis:

This paper will present an original proof of Chebyshev's Inequality and attempt to show that the inequality is extremely valuable in statistics, can be understood with minimal effort and can be proved in an understandable way in an elementary statistics course.

On Chebyshev's Inequality in Elementary Statistics-An Original Proof

When students in my elementary statistics classes encounter Chebyshev's Inequality, it is routinely perceived as being incomprehensible and furthermore of little value. And when the course ends the illusion generally remains. This paper presents a comprehensible proof appropriate for an elementary statistics course.

Common questions from students are:

- What does Chebyshev's Inequality really do?
- When using Chebyshev's rule, should I use the population standard deviation, sample standard deviation, or can I use either?
- Of what value is Chebyshev's Inequality?
- Why must k be larger than one?

Purpose:

This paper will present an original proof of Chebyshev's Inequality and attempt to show that the inequality is extremely valuable in statistics, can be understood with minimal effort and can be proved in an understandable way in an elementary statistics course.

History:

Known as the founding father of Russian mathematics, Pafnuty Chebyshev first mentioned and proved the inequality in a paper published in 1867. However, it first appeared fourteen years earlier and stated without proof in 1853 in a paper published by his friend and colleague, Irénée-Jules Bienaymé http://en.wikipedia.org/wiki/Chebyshev%27s_inequality#History. In 1884 one of Chebyshev's two doctoral students, Andrey Andreyevich Markov, presented an elementary proof of the theorem in his Doctoral thesis "About Some Applications of Algebraic ContinuousFractions"(Wikipedia).[6]http://en.wikipedia.org/wiki/Chebyshev%27s_inequalityhttp://en.wikipedia.org/wiki/Chebyshev%27s_inequality#History A crater on the moon and an asteroid, Asteroid 2010 Chebyshev, are named after him.
http://www.statisticshowto.com/pafnuty-lvovich-chebyshev/

Chebyshev's Inequality:

Chebyshev's Inequality states that not more than $\frac{1}{k^2}$ of any distribution can be more than k standard deviations from the mean or at least $1-\frac{1}{k^2}$ must be within k standard deviations of the mean.

Example:
The average height of an adult male is 69.7 inches with a standard deviation of 2.76 inches.
The average height of an adult female is 63.8 inches with a standard deviation of 2.36 inches.

A two standard deviation range: $1-\frac{1}{k^2} = 1-\frac{1}{2^2} = 75\%$

Adult male:

$[69.7 - 2(2.76), 69.7 + 2(2.76)] =$

$[64.18 \text{ inches}, 75.22 \text{ inches}]$

Adult female:

$[63.8 - 2(2.36), 63.8 + 2(2.36)] =$

$[59.08 \text{ inches}, 68.52 \text{ inches}]$

$\left\{\begin{matrix}95.44\% \text{ guaranteed by empirical rule} \\ 75\% \text{ guaranteed by Chebyshev's rule}\end{matrix}\right\}$

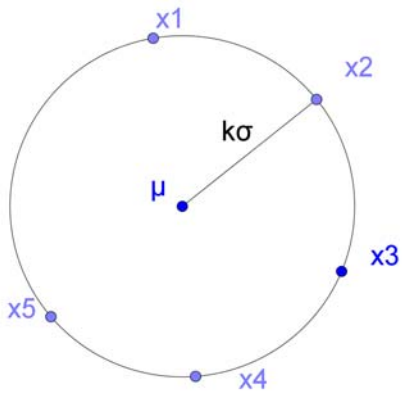
Consider a non-empty distribution containing N data values.

Let N_1 be the number of data values whose distance from the mean is $\geq k$ standard deviations.

Then $\frac{N_1}{N} \leq \frac{1}{k^2}$.

In order to maximize the number of data values at a distance $\geq k\sigma$ from the mean, we divide the data set into two parts , N_1 of the data values being precisely $k\sigma$ from the mean and the remaining $N - N_1$ data values being at the mean.

The total variance $\left(\text{the total sum of the squares}\right)$ is then assigned to the outliers.



The population variance of a distribution is: $\sigma^2 = \dfrac{\sum_{i=1}^{N}\left(x_i - \mu\right)^2}{N}$

The total sum of the squares is the numerator: $SS = \sum_{i=1}^{N}\left(x_i - \mu\right)^2$

$$N\sigma^2 = \sum_{i=1}^{N}\left(x_i - \mu\right)^2$$

For each of the data values $\left| x_i - \mu \right| = k\sigma$

For each of the remaining $N - N_1$

data values $\left| x_i - \mu \right| = 0.$ Therefore $\left(x_i - \mu\right)^2 = k^2\sigma^2$

$N_1\left(x_i - \mu\right)^2 = N_1 k^2\sigma^2$

Let the total sum of the squares = SS=$N_1 k^2\sigma^2$

$$\frac{\sum\left(x_i - \mu\right)^2}{N} = \sigma^2$$

$$\sum\left(x_i - \mu\right)^2 = N\sigma^2 = SS$$

Therefore $N_1 k^2\sigma^2 = N\sigma^2; \left(N \neq 0, \text{ and } \sigma \neq 0\right)$

$N_1 k^2 = N$

$$\frac{N_1}{N} = \frac{1}{k^2}$$

Therefore the fraction of outliers is less than. $\dfrac{1}{k^2}$.

This argument holds when this unique distribution
from Candide's world is possible.

Example:

Let $x_i = \left\{ \begin{matrix} 1 \text{ for i = 1 to 50} \\ -1 \text{ for i =1 to 50} \end{matrix} \right\}$

$\mu = 0, \sigma^2 = 1, \sigma = 1$

If $k = 2$ then $1 - \frac{1}{k^2} = \frac{3}{4}$

$\mu \pm k\sigma = [2, -2]$

and Chebyshev obviously is valid.

More Generally:

Given a population of N data values with variance σ^2.
Let N_1 be the number of values $\geq k\sigma$
from the mean.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \qquad N\sigma^2 = \sum_{i=1}^{N}(x_i - \mu)^2 = \text{sum of the squares}$$

$$\sum_{i=1}^{N_1}(k\sigma + c_i)^2 \leq N\sigma^2$$

$$\sum_{i=1}^{N_1}k^2\sigma^2 + \sum_{i=1}^{N_1}2k\sigma c_i + \sum_{i=1}^{N_1}(c_i)^2 \leq N\sigma^2$$

$$N_1k^2\sigma^2 \leq N\sigma^2 - \sum_{i=1}^{N_1}2k\sigma c_i - \sum_{i=1}^{N_1}(c_i)^2$$

$$\sum_{i=1}^{N_1} k^2\sigma^2 + \sum_{i=1}^{N_1} 2k\sigma c_i + \sum_{i=1}^{N_1} (c_i)^2 \le N\sigma^2$$

$$N_1 k^2\sigma^2 \le N\sigma^2 - \sum_{i=1}^{N_1} 2k\sigma c_i - \sum_{i=1}^{N_1} (c_i)^2$$

$$\frac{N_1}{N} \le \frac{1}{k^2} - \frac{\sum_{i=1}^{N_1} 2k\sigma c_i}{Nk^2\sigma^2} - \frac{\sum_{i=1}^{N_1} (c_i)^2}{Nk^2\sigma^2} \le \frac{1}{k^2}$$

or the proportion of the population
that can can be $\ge k\sigma$ from the mean $\le \frac{1}{k^2}$.

Note that given σ^2 does not effect $\frac{1}{k^2}$.

$$\frac{N_1}{N} \leq \frac{1}{k^2}$$
$$-\frac{N_1}{N} \geq -\frac{1}{k^2}$$
$$1-\frac{N_1}{N} \geq 1-\frac{1}{k^2}$$

or the proportion of the population that must be $\leq k\sigma$ from the mean is $\geq 1-\frac{1}{k^2}$.

Argument that k must be greater than 1:

$$\frac{N_1}{N} \leq \frac{1}{k^2}$$

If k ≤ 1 then $\frac{1}{k^2} \geq 1$

This is implying that the fraction of outliers is some number ≤ 1, yielding no new information!

Conclusion:

- Chebyshev's inequality can be used on any distribution (there were no assumptions about a normal population in the proof).
- As the proof demonstrates, the population standard deviation is used.
- Chebyshev's inequality sets exact boundaries on the dispersion of a distribution. The inequality does not approximate, but guarantees that at least $1-\frac{1}{k^2}$ of the data lies within k standards deviations of the population mean.
- The theorem has been used in proving other theorems such as the Law of Large Numbers.
- The proof presented uses no mathematics higher than that required for an elementary statistics course.

REFERENCES:

Derbyshire, J. *Prime Obsession: Bernhard Riemann and the Greatest Unsolved Problem in Mathematics.* New York: Penguin, 2004.

http://www.statisticshowto.com/what-is-chebyschevs-inequality/

http://www.statisticshowto.com/pafnuty-lvovich-chebyshev/

http://en.wikipedia.org/wiki/Chebyshev%27s_inequality#History