STATISTICS FOR DATA SCIENCE

Chebyshev's Inequality

Prof. Uma D

Department of Computer Science and Engineering

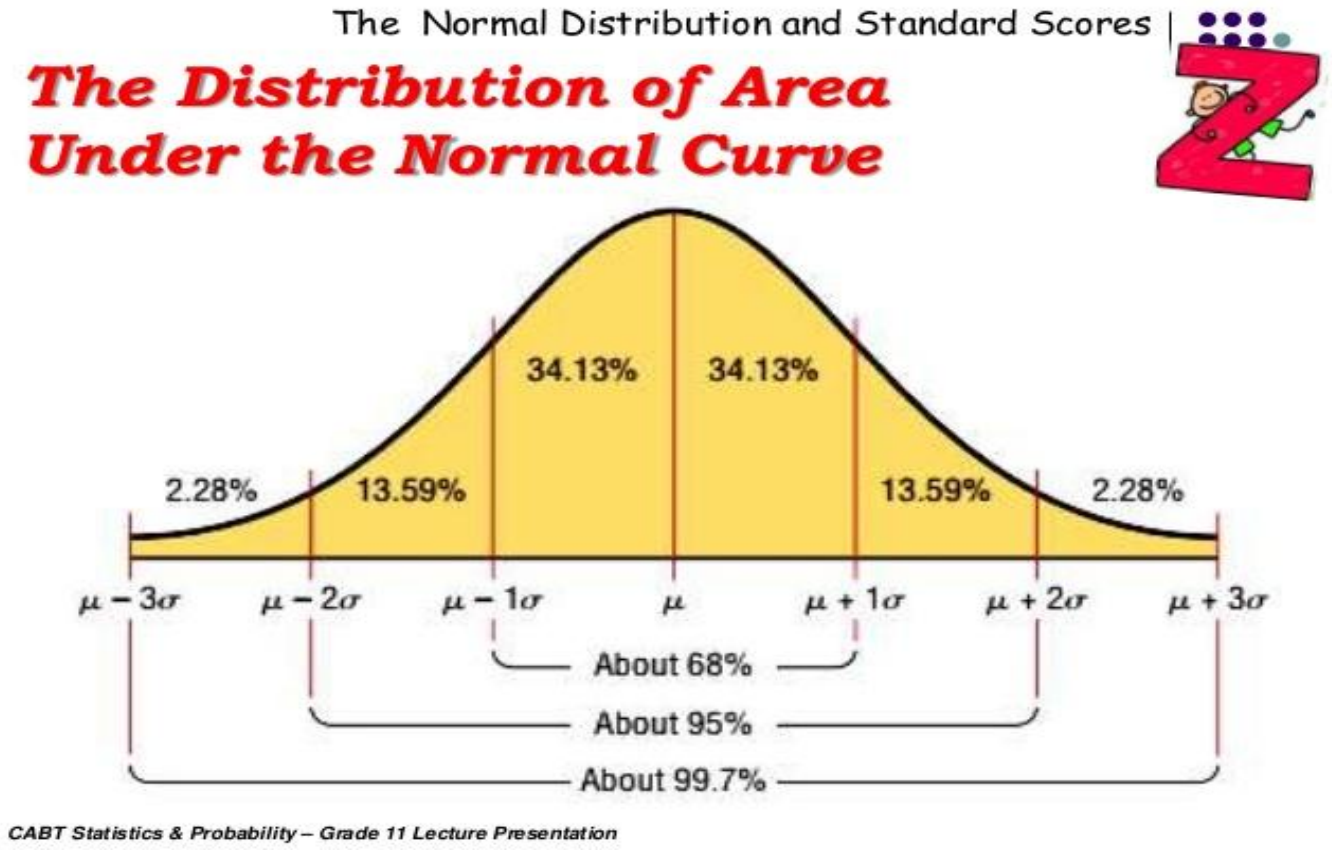
STATISTICS FOR DATA SCIENCE

Chebyshev's Inequality

Prof. Uma D

STATISTICS FOR DATA SCIENCE

Normal Distribution



Source Image: www.slideshare.net

Normal Distribution

68 – 95 -99.7 rule is when data is normally distributed.

$$Pr(\mu - \sigma < x < \mu + \sigma) \simeq 0.6827$$

$$Pr(\mu - 2\sigma < x < \mu + 2\sigma) \simeq 0.9545$$

$$Pr(\mu - 3\sigma < x < \mu + 3\sigma) \simeq 0.9973$$

When data is not normally distributed??

Chebyshev's inequality provides a way to know what fraction of data falls within **K standard deviations** from the mean for any data set.

Statement of Chebyshev's Inequality

Chebyshev's inequality states that at least $1 - 1/K^2$ of data from a sample must fall within K standard deviations from the mean, where K is any positive real number greater than one.

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Chebyshev's Inequality is used to describe the percentage of values in a distribution within an interval centered at the mean.

Source Image:ThoughtCo.com

Illustration of the Inequality

For K = 2 we have 1-1/K2 = 1 - 1/4 = 3/4 = 75%.

Chebyshev's inequality says that **at least 75%** of the data values of any distribution must be within **two standard deviations** of the mean.

For K = 3 we have 1 – 1/K2 = 1 - 1/9 = 8/9 = 89%.

Chebyshev's inequality says that **at least 89%** of the data values of any distribution must be within **three standard deviations** of the mean.

Chebyshev's Inequality

Note:

In practical usage, in contrast to the 68–95–99.7 rule, which applies to normal distributions,

Chebyshev's inequality is weaker, stating that a minimum of just 75% of values must lie within two standard deviations of the mean and 89% within three standard deviations.

Only the case $k > 1$ is useful.

When $k \leq 1$ the right hand $1/ k\ 2 \geq 1$ and the inequality is trivial as all probabilities are ≤ 1.

Statement of Chebyshev's Inequality

Chevyshev's Inequality can also be stated as follows:

Chebyshev's inequality relates mean and standard deviation by providing a bound on the probability that a Random Variable takes on a value that differs from its mean by K standard deviation or more is never greater than 1/ k 2

Note:

Chebyshev's bound is generally much larger than the actual probability.

Hence should only be used when the distribution of the random variable is unknown.

Problem 1

Computers from a particular company are found to last on average for three years without any hardware malfunction, with standard deviation of two months. At least what percent of the computers last between 31 months and 41 months?

$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$ $\mu = 3$ years $= 36$ months

$|31-36| = 5$ months $|41-36| = 5$ months $\sigma = 2$ months

$5 \geq k\sigma \Rightarrow k = \frac{5}{\sigma} = \frac{5}{2} = 2.5$ $k = 2.5$

$1 - \frac{1}{k^2} = 1 - \frac{1}{(2.5)^2} = 84\%$

Problem 2

What is the smallest number of standard deviations from the mean that we must go if we want to ensure that we have at least 50% of the data of a distribution?

$1 - \frac{1}{k^2} = 0.50 \Rightarrow \frac{1}{k^2} = 0.50$ $\quad k^2 = \frac{1}{0.5} \Rightarrow k^2 = 2$

$k = \sqrt{2} = 1.4$

Conclusion: At least 50% of the data is within approximately 1.4 standard deviations from the mean.

Do It Yourself !!!

The length of a metal pin manufactured by a certain process has mean 50 mm and standard deviation 0.45mm.

What is the largest possible value for the probability that the length of the metal pin is outside the interval [49.1 , 50.9] mm?

THANK YOU

Prof. Uma D

Department of Computer Science and Engineering