STATISTICS FOR DATA SCIENCE

Chebyshev's Inequality

Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

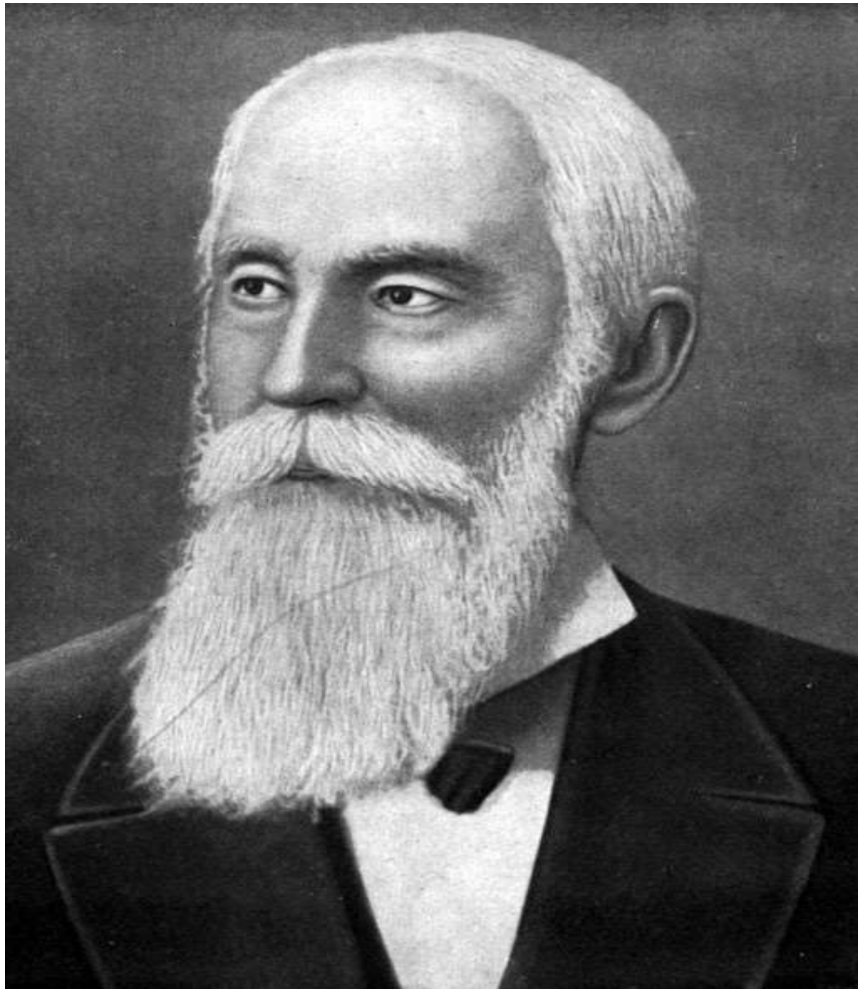
Chebyshev’s Inequality

Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

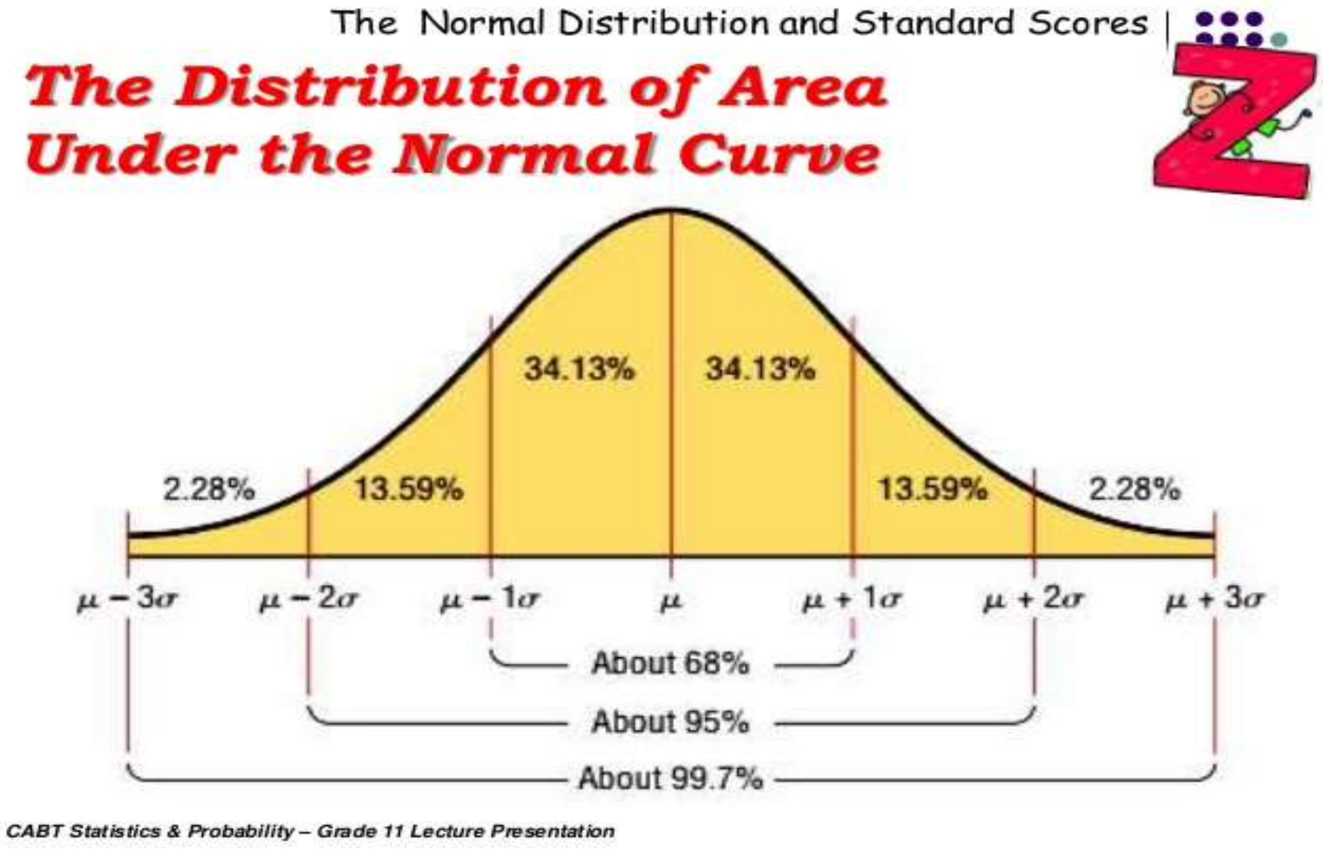
Topics to be covered...

- Chebyshev’s Inequality
- 68–95–99.7 rule : When Data is distributed Normally?
- When Data is not distributed Normally?
- Statement of Chebyshev’s Inequality
- Examples

CHEBYSHEV'S INEQUALITY



Normal Distribution



Source Image: www.slideshare.net

Normal Distribution

Shorthand used to remember the percentage of values that lie within a band around the mean in a **normal distribution** with a width of one, two and three standard deviations, respectively

Normal Distribution

$$\Pr(\mu - \sigma \leq x \leq \mu + \sigma) \approx 0.6827$$
$$\Pr(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 0.9545$$
$$\Pr(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \approx 0.9973$$

When Data is not distributed Normally?

- But if the data set is not distributed normally, then a different amount could be within one standard deviation.

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

- The inequality has great utility because it can be applied to any probability distribution in which the mean and variance are defined.

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.

$$P(\mu_X - k\sigma_X < X < \mu_X + k\sigma_X) \geq 1 - \frac{1}{k^2}$$

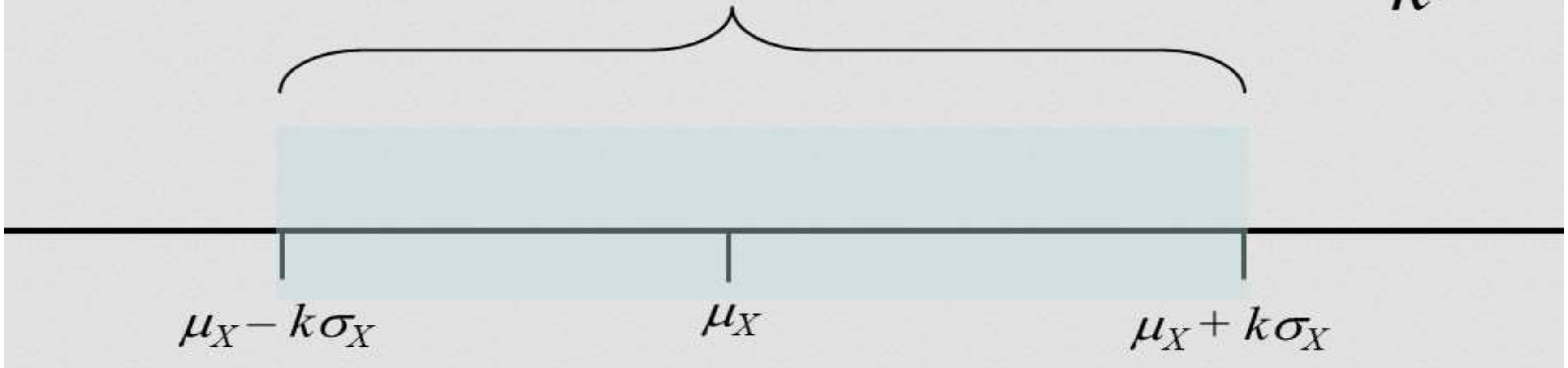


Illustration of the Inequality

To illustrate the inequality, we will look at it for a few values of K:

For K = 2 we have $1-1/K^2$ = 1 - 1/4 = 3/4 = 75%. So 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 $1 – 1/K^2$ = 1 - 1/9 = 8/9 = 89%. So Chebyshev's inequality says that at least 89% of the data values of any distribution must be within three standard deviations of the mean.

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.

Statement of Chebyshev's Inequality

- 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 $1/k^2$

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

- 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.

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.

Example

- 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?

Solution

Mean lifetime = 3 years = 36 months.

Standard Deviation = 2 months

To find % of the computers last from 31 months to 41 months.

| 31 – mean | = | 31 – 36 | = 5 months

| 41 – mean | = | 41 – 36 | = 5 months

K = 5/standard deviation = 5/2

=> K = 2.5

By Chebyshev's inequality,

at least $1 - 1/(2.5)^2 =$ 84% of the computers last from 31 months to 41 months.

Example

- 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?

Solution

Mean = 50 mm

Standard deviation = 0.45 mm

To find $P(X \leq 49.1 \text{ or } X \geq 50.9) \leq 1/K^2$

Find K:

$|49.1 - \text{mean}| = |49.1 - 50| = 0.9$

$|50.9 - \text{mean}| = |50.9 - 50| = 0.9$

$K = 0.9 / \text{Standard deviation} = 0.9 / 0.45 \quad K = 2$

By Chebyshev's inequality,

$P(X \leq 49.1 \text{ or } X \geq 50.9) \leq 1/K^2 \leq 1/4 \leq 0.25$

Example

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?

Solution

Here we use Chebyshev's inequality and work backward.

$1-1/K^2 = 0.50$

$1/K^2 = 0.50$

$K^2 = 1/0.5$

$K^2 = 2$

$K = \text{sqrt}(2)$

$K = 1.4$

By Chebyshev's inequality,

So at least 50% of the data is within approximately 1.4

THANK YOU

Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

Department of Computer Science and Engineering