Probability and Statstics

Dr. Faisal Bukhari

Associate Professor

Department of Data Science

Faculty of Computing and Information Technology

University of the Punjab

Textbooks

- **Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer

- **Elementary Statistics: Picturing the World,** 6th Edition, Ron Larson and Betsy Farber

- **Elementary Statistics**, 13th Edition, Mario F. Triola

Reference books

- ❑ **Probability and Statistical Inference, Ninth Edition,** Robert V. Hogg, Elliot A. Tanis, Dale L. Zimmerman

- ❑ **Probability Demystified**, Allan G. Bluman

- ❑ **Practical Statistics for Data Scientists: 50 Essential Concepts,** Peter Bruce and Andrew Bruce

- ❑ **Schaum's Outline of Probability,** Second Edition, Seymour Lipschutz, Marc Lipson

- ❑ **Python for Probability, Statistics, and Machine Learning,** José Unpingco

References

Readings for these lecture notes:

- **Schaum's Outline of Probability, Second Edition (Schaum's Outlines)** by by Seymour Lipschutz, Marc Lipson

- **Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer

- Reference: http://www.mathsisfun.com/data/standard-deviation.html

These notes contain material from the above resources.

Distribution-free Result [1]

Chebyshev's Theorem: The probability that any random variable X will assume a value within **k standard deviations** of the mean is **at least $1 – 1/k^2$**.

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

Distribution-free Result [2]

Example 1: A random variable X has a mean μ = 8, a variance σ^2= 9, and an unknown probability distribution. Find

(a) $P(-4 < X < 20)$,

(b) $P(|X - 8| \geq 6)$.

Solution:

μ $= 8$ and **σ** $= 3$

$\mu - k\sigma$ $= 8 - k(3) =$ **$8 - 3k$**

$\mu + k\sigma$ $= 8 + k(3) =$ **$8 + 3k$**

$8 - 3k =- 4 \Rightarrow - 3k = -12 \Rightarrow k = 4$

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

(a) $P(-4 < X < 20) = P[8 - (4)(3) < X < 8 + (4)(3)]$

$$\geq 1 - \frac{1}{4^2}$$

$$\geq \frac{15}{16}$$

$$\geq 0.9375$$

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

$\mu = 8$ and $\sigma = 3$

(b) $P(|X - 8| \geq 6) = 1 - P(|X - 8| < 6) = 1 - P(-6 < X - 8 < 6)$

$= 1 - P(-6 + 8 < X < 6 + 8)$

$\mu - k\sigma = 8 - k(3) = 2$

or

$\mu + k\sigma = 8 + k(3) = 14$

$8 + 3k = 14 \Rightarrow 3k = 6 \Rightarrow k = 2$

$\Rightarrow P(|X - 8| \geq 6) \quad = 1 - P(2 < X < 14)$

$\Rightarrow P(|X - 8| \geq 6) \quad = 1 - P[8 - (2)(3) < X < 8 + (2)(3)]$

$\because P(\mu - k\sigma < X < \mu + k\sigma) \leq 1 - \frac{1}{k^2}$

Multiply both sides by -1

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

Add 1 on both sides

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

$\Rightarrow \mathbf{1 - P(\mu - k\sigma < X < \mu + k\sigma) \leq \frac{1}{k^2}}$

$$P(|X - 8| \geq 6) = \mathbf{1 - P[8 - (2)(3) < X < 8 + (2)(3)]}$$

$$\because \mathbf{1 - P(\mu - k\sigma < X < \mu + k\sigma) \leq \frac{1}{k^2}}$$

$$\Rightarrow P(|X - 8| \geq 6) \leq \frac{1}{2^2}$$

$$\Rightarrow \mathbf{P(|X - 8| \geq 6)} \leq \frac{1}{4}$$

Distribution-free Result [3]

According to Chebyshev's, the probability that a random variable assumes a value within 2 standard deviations of the mean is **at least 3/4**.

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

$$\Longrightarrow P(\mu - 2\sigma < X < \mu + 2\sigma) \geq 1 - \frac{1}{2^2}$$

$$\Longrightarrow P(\mu - 2\sigma < X < \mu + 2\sigma) \geq \frac{3}{4} \text{ ans}$$

Distribution-free Result [4]

Here $x_1 = \mu - 2\sigma$ and $x_2 = \mu + 2\sigma$

$$Z = \frac{x - \mu}{\sigma}$$

At $x_1 = \mu - 2\sigma$

$$\Longrightarrow \; Z_1 = \frac{(\mu - 2\sigma) - \mu}{\sigma}$$

$$= -2$$

At $x_2 = \mu + 2\sigma$

$$\Longrightarrow Z_2 = \frac{(\mu + 2\sigma) - \mu}{\sigma}$$

$$= 2$$

$$\begin{aligned} P(\mu - 2\sigma < X < \mu + 2\sigma) &= P(-2 < Z < 2) \\ &= P(Z < 2) - P(Z < -2) \\ &= 0.9772 - 0.0228 \\ &= \mathbf{0.9544} \end{aligned}$$

$\Longrightarrow$ Which is a much stronger statement than that given by **Chebyshev's theorem**.

Measures of Location: The Sample Mean and Median [1]

- ❑ **Measures of location** are designed to provide the analyst with some **quantitative values** of where the center, or some other location, of data is located.

- ❑ **Sample mean**: Suppose that the observations in a sample are $x_1, x_2, \ldots, x_n$. The **sample mean**, denoted by $\bar{x}$, is

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$= \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Measures of Location: The Sample Mean and Median [2]

- ❑ **Sample median:** Given that the observations in a sample are $x_1, x_2, \ldots, x_n$, arranged in **increasing order** of magnitude, the sample median is

$$x = \left\{ \begin{array}{ll} x_{(n+1)/2,} & if \;\; n \; is \; odd \\ \frac{1}{2}(x_{n/2} \; + \; x_{n/2 \, + \, 1}), & if \;\; n \; is \; even \end{array} \right\}$$

Measures of Location: The Sample Mean and Median [2]

Example: Find mean and median of the following data: 1.7, 2.2, 3.9, 3.11, and 14.7.

Solution:

$$\overline{x} = \frac{\sum_{i=1}^{n} \mathrm{x_i}}{n}$$

$$= \frac{x_1 + \ x_2 + . \ . \ . + x_n}{n}$$

$\overline{x}$ = 5.12 ans

Measures of Location: The Sample Mean and Median [2]

$$x = \left\{ \begin{array}{ll} x_{(n+1)/2}, & if \;\; n \; is \; odd \\ \frac{1}{2}(x_{n/2} \; + \; x_{n/2 \; + \; 1}), & if \;\; n \; is \; even \end{array} \right\}$$

$$x = 3.9$$

Arithmetic Mean (Average)

The arithmetic mean is calculated by summing all values in a dataset and dividing by the count of values.

Example:

For values 5, 10, 15, the mean is (5 + 10 + 15) / 3 = 10.

Applications:

- Feature scaling and normalization in machine learning
- Assessing model performance consistency in evaluations
- Time complexity analysis of algorithms

Median

The median is the middle value in an ordered dataset, which is useful for datasets with outliers or skewed distributions.

Example:

For the dataset 2, 4, 6, 8, the median is $\frac{4+6}{2}$ = 5

Applications:

○ Robust statistic for non-normal data distributions

○ Used in model performance metrics where outliers are minimized

○ Common in image processing for noise reduction

Mode

The mode is the most frequently occurring value in a dataset. It is used often with categorical data.

Example:

For values 1, 2, 4, 4, 5, the mode is 4.

Applications:

- Data imputation for missing categorical values
- Identifying common terms in text analysis (NLP)
- Useful in anomaly detection for repetitive patterns

Geometric Mean

The geometric mean is calculated by multiplying all values and taking the nth root, where n is the count of values.

Example:

For values 2, 8, the geometric mean is $\sqrt{2 \times 8}$ = 4

Applications:

- ○ Used in compounded growth rate analysis
- ○ Applied in NLP for average accuracy/error rate across tasks
- ○ Log-transformations in multiplicative relationships

Harmonic Mean

The harmonic mean is the reciprocal of the average of the reciprocals of values.

Example:

For values 1, 4, 4, harmonic mean = $\frac{3}{\frac{1}{1}+\frac{1}{4}+\frac{1}{4}}$ = 2

Applications:

- F1 score calculation in classification (ML)
- Used in network latency analysis for average response time
- Emphasizes lower values in data, making it useful for certain rate-based data

Weighted Mean

Weighted mean gives different values different levels of importance based on assigned weights.

Example:

If x_1=10, x_2=20, weights w_1=1, w_2=3, then weighted mean = (10 × 1 + 20 × 3) / (1+3) = **17.5**

Applications:

- Ensemble learning where model weights vary based on accuracy
- Weighted loss functions in imbalanced class datasets
- Important in time series where recent data is more relevant

Combined Mean

The **combined mean** is **the mean of two or more groups** with **different sizes**, useful in cross-validation or distributed data.

Example:

If Group A (mean=10, size=5) and Group B (mean=20, size=5), combined mean = 15.

Applications:

- Calculating overall performance in cross-validation
- Useful in distributed data analysis across multiple machines
- Helpful in analyzing customer segment data in business intelligence

Trimmed Mean [1]

- ❑ A trimmed mean is computed by **“trimming away”** a certain percent of both the largest and the smallest set of values. For example, the 10% trimmed mean is found by **eliminating the largest 10% and smallest 10%** and computing the average of the remaining values.

Trimmed Mean [2]

Example: Find the 10% trimmed mean for no nitrogen and nitrogen for the given data.

No Nitrogen	Nitrogen
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

Solution:

Without-nitrogen group the 10% trimmed mean is given by

$\bar{x}_{tA\ (10)}$ = (0.32 + 0.37 + 0.47 + 0.43 + 0.36 + 0.42 + 0.38 + 0.43)/8

= .39750

10% trimmed mean for the with-nitrogen group we have

$\bar{x}_{tA\ (10)}$= (0.43 + 0.47 + 0.49 + 0.52 + 0.75 + 0.79 + 0.62 + 0.46)/8

= .56625.

Trimmed Mean [3]

- ❑ The trimmed mean is, of course, more **insensitive** to **outliers** than the sample mean but not as insensitive as the median.

- ❑ On the other hand, the trimmed mean approach makes use of more information than the sample median. Note that the **sample median** is, indeed, a special case of the **trimmed mean** in which all of the sample data are eliminated apart from the middle one or two observations.

Measures of Variability

❑ **Sample range** = x_{max} - x_{min}

OR

It is the difference between the maximum and the minimum value in the data.

❑ The **sample variance**, denoted by s^2, is given by

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

OR

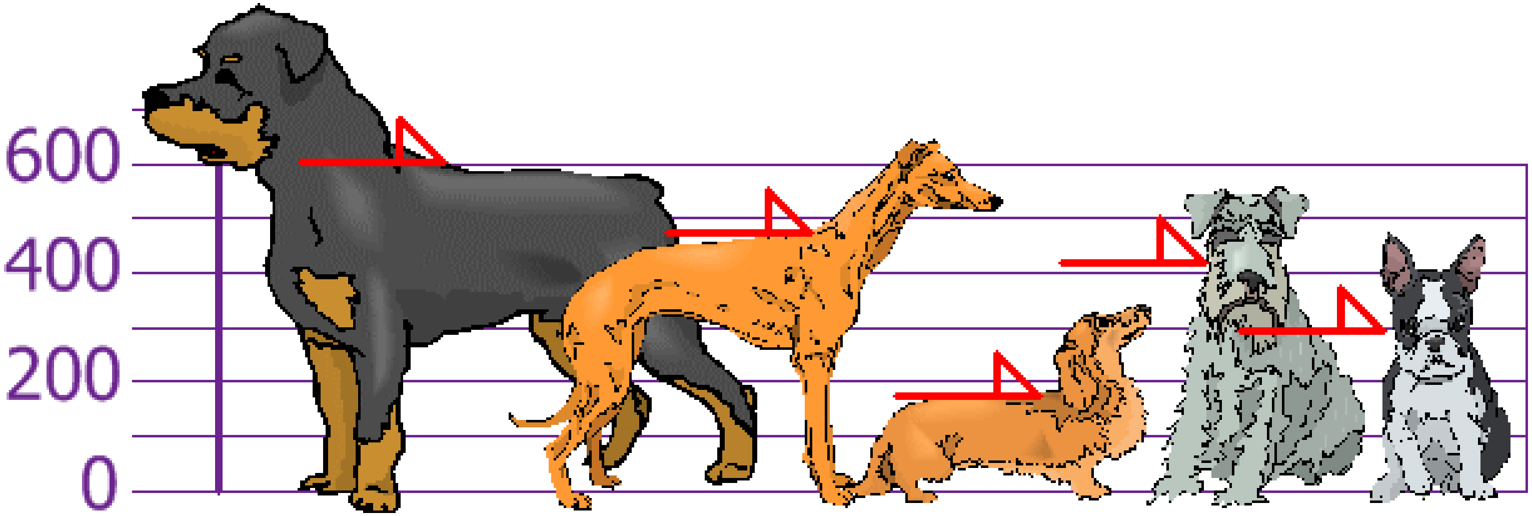
It is the average of the **squared** differences from the Mean.

❑ The **sample standard deviation**, denoted by s, is the positive square root of s^2, that is, $s = \sqrt{s^2}$

Note: It should be clear to the reader that the sample standard deviation is, in fact, a measure of variability. Large variability in a data set produces relatively large values of $\sum(\mathrm{x_i} - \overline{\mathrm{x}})^2$and thus a large sample variance. The quantity n − 1 is often called the **degrees of freedom associated with the variance** estimate. In this simple example, the degrees of freedom depict the number of **independent pieces of information** available for computing variability

Mean, Variance, and S.D [1]

- **Example:** You and your friends have just measured the heights of your dogs (in millimeters):

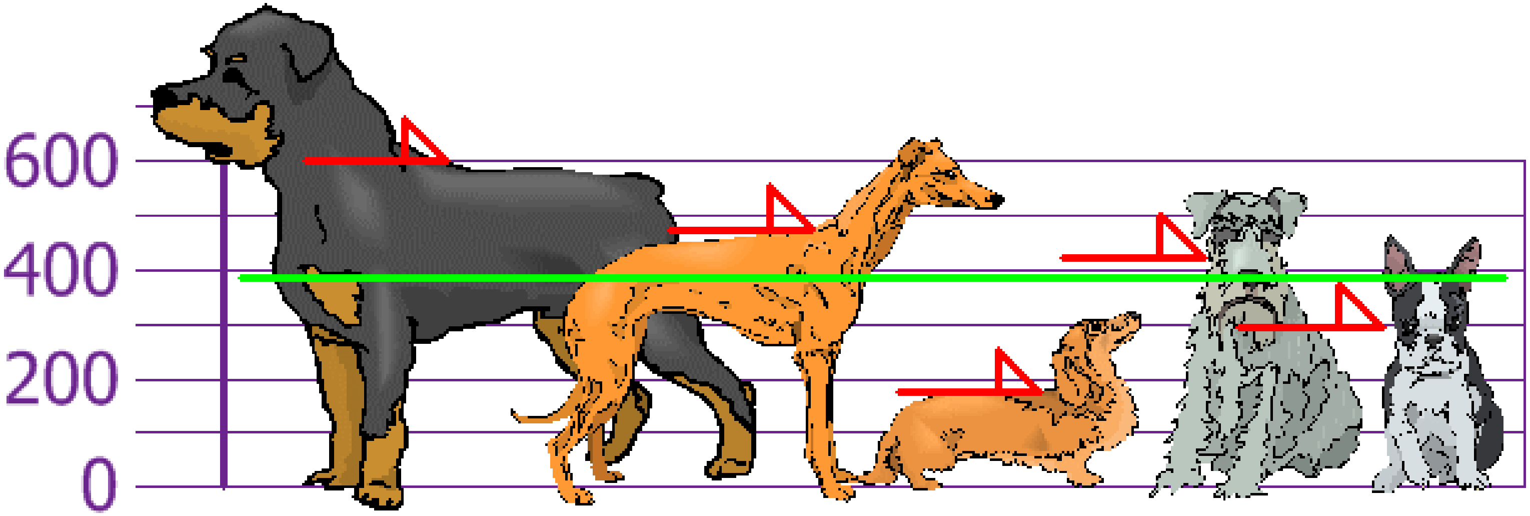


- The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm. Find out the Mean, the Variance, and the Standard Deviation.

Mean, Variance, and S.D [2]

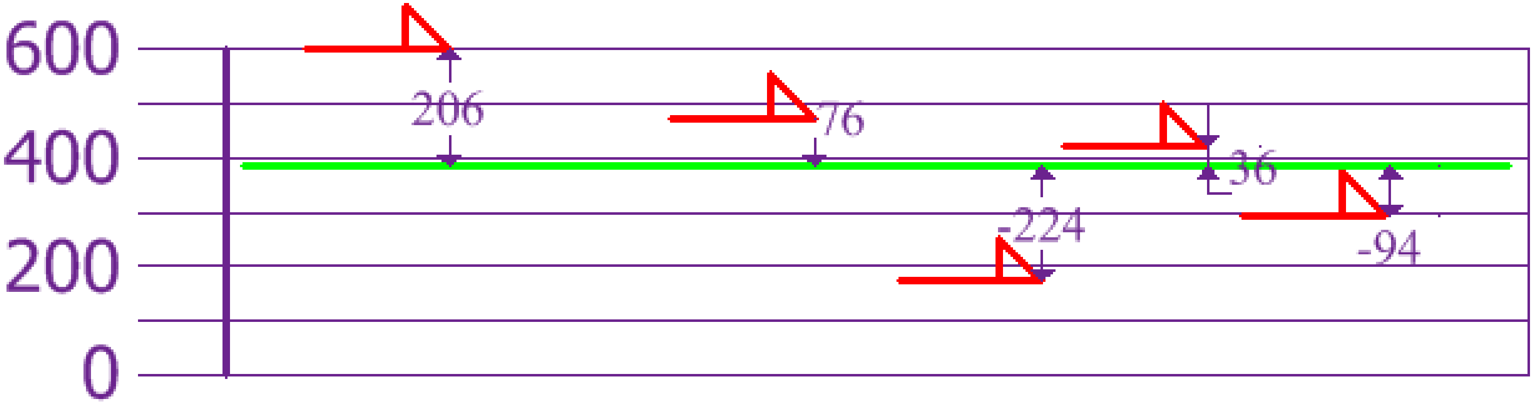
Solution:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{600 + 470 + 170 + 430 + 300}{5} = \frac{1970}{5} = 394$$



Mean, Variance, and S.D [3]

- ❑ Now, we calculate each dogs difference from the Mean:



Mean, Variance, and S.D [4]

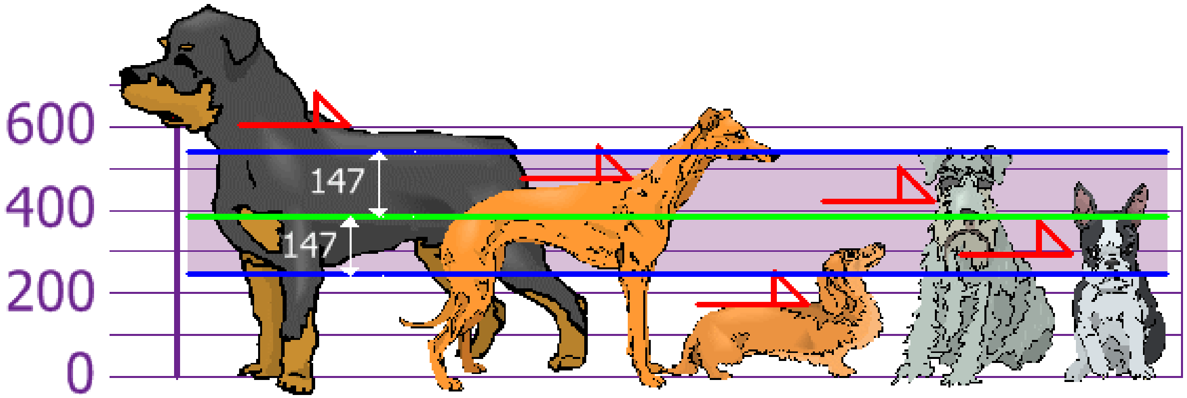
To calculate the Variance, take each difference, square it, and then average the result:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$$

$$s^2 = \frac{206^2 + 76^2 + (-244)^2 + 36^2 + (-94)^2}{5}$$

$$s^2 = \frac{108{,}520}{5} = 21,\ 704$$

$$s = \sqrt{s^2} = 147$$



Mean, Variance, and S.D [5]

- So, using the **Standard Deviation** we have a "standard" way of **knowing what is normal**, and what is **extra large or extra small**.

- Sample Variance = 108,520 / **4** = **27,130**

- Sample Standard Deviation = $\sqrt{27{,}130}$ = **164** (to the nearest mm)

Units for Standard Deviation and Variance:

We use the term average squared deviation even though the definition makes use of a division by degrees of freedom **n – 1** rather than n.

- ❑ Of course, if n is large, the difference in the denominator is inconsequential.
- ❑ As a result, the **sample variance** possesses units that are **the square of the units** in the observed data whereas the sample **standard deviation** is found in **linear units**.