University of Jordan
Chemical Engineering Department
905509 Statistical Quality Control

Review of Statistical Concepts

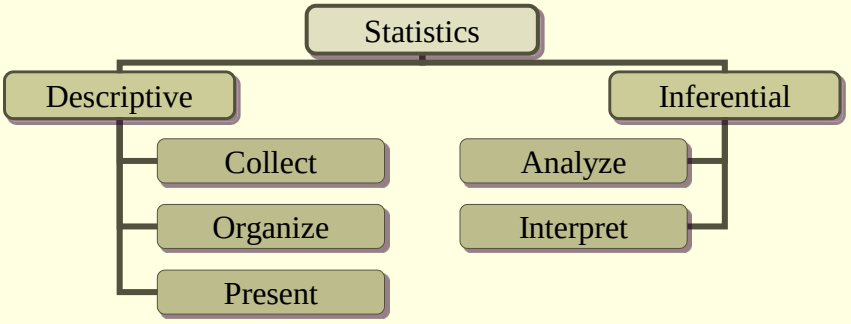
Dr. Ali Kh. Al-Matar
aalmatar@ju.edu.jo

Outline

- Review of statistics
- Measures of central tendency
- Measures of dispersion
- Measures of Symmetry

What is Statistics?

- Statistics is the science dealing with data to assist in making more effective decisions in the face of uncertainty

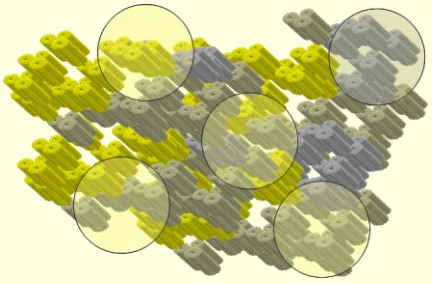


Why Study Statistics?

- Numerical information is everywhere!
- Statistical methods are used to make decisions that affect our lives.
- To understand why decisions are made and how such decisions affect us.

Population & Sample

- **Population** is the collection consisting of all possible outcomes of an experiment, measurement, or observation.
- **Sample** is a subset or a part of the population.



Measures of Central Tendency: Mean

- Mean

Sample *Population*

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

- Weighted Mean

Sample *Population*

$$\overline{X}_w = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i} \qquad \mu_w = \frac{\sum_{i=1}^{N} w_i X_i}{\sum_{i=1}^{N} w_i}$$

Central Tendency: Median

- The mean is very sensitive to outliers.
- Median is defined as a rigorous estimator of central tendency.
  - Sort the data in ascending order
  - The median is defined such that 50% of the values are above, and 50% are below it.

$$X = \begin{cases} Value((n+1)/2), & n \text{ is odd} \\ Average(Value(n/2), Value(n+1)/2))), & n \text{ is even} \end{cases}$$

Central Tendency: Mode

- Mode: the value of the most frequently encountered observation.
- Mathematically, it is the most frequent value i.e., the maximum of the distribution.

Measures of Dispersion

- Averages provide information about *central tendency* but it does not provide any info about the *spread* of the data
- Range is the difference between the maximum and minimum values of the random variable we are interested in

$$Range = Max - Min$$

Dispersion: Variance and Standard Deviation

- Variance

$$Sample: s^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}$$

$$Population: \sigma^2 = \frac{\sum_{i=1}^{n}\left(X_i - \mu\right)^2}{N}$$

- Standard Deviation

$$Sample: s = +\sqrt{s^2}$$

$$Population: \sigma = +\sqrt{\sigma^2}$$

Measures of Dispersion II

- Coefficient of Variation (CV) is the ratio of standard deviation to the arithmetic mean. Usually, it is expressed in percent

$$CV = 100\frac{s}{\bar{X}}\%$$

- It is a good measure for comparing different values of means and standard deviations

Measures of Symmetry: Skewness

- Measures of dispersion give some values about the spread of a distribution.
- They don't provide any info about the shape of the distribution around the mean or median.
- Coefficient of skewness (Sk) is defined to alleviate such lack of info

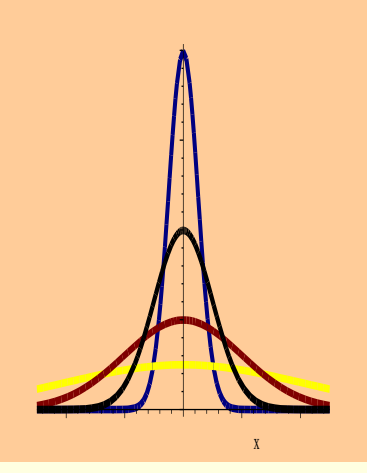
$$Sk = \frac{3(mean - median)}{s}$$

Useful Excel Functions

- Average
- Stdev
- Mode
- Median
- Skew
- MAX and MIN

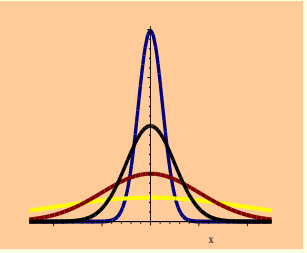
Normal (Gaussian) Distribution

- Most important and widely used distribution
- Occurs naturally due to "**central limit theorem**"
- Describes
  - Error distribution
  - Height, and weight of individuals
  - Life expectancy of certain products, people etc.
  - Good approximation for many distributions



Characteristics of the Normal Distribution

- Bell shaped and has a single peak (unimodal) at the center of the distribution.
- Mean = Median = Mode.
- Symmetrical around the mean
- Falls asymptotically.



1σ	68.26%
2σ	95.44%
3σ	99.74%

Mathematics of the Normal Distribution

- The normal distribution is characterized by two parameters:
  - Mean (μ).
  - Standard deviation (σ).

$$f(x,\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

$$F(x,\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]dx$$

Standard Normal Distribution

- The standard normal distribution has a mean of 0 and standard deviation equal to unity.
- The standard normal distribution is useful since there is no closed form integral for the cumulative pdf. It reduces all families of normal distributions to one using the z transform.

The z Transform

- Used to convert between the actual normal distribution and standard normal distributions

$$z = \frac{X - \mu}{\sigma}$$

Useful Excel Function

- NORMSDIST(z) : standard normal distribution
- NORMSINV(p) : inverse standard normal distribution
- NORMINV(p,m,s): inverse general
- NORMDIST(X,m,s,Cumulative) : general

Example 1

A life test on a large number of type D alkaline batteries revealed that the mean life for a particular use before failure is 19.0 hours. The distribution of lives approximated a normal distribution. The standard deviation was 1.2 hours.

1. About 68% of the batteries failed between what two values?
2. About 95% of the batteries failed between what two values?
3. Virtually all of the batteries failed between what two values?

Solution

μ=19
σ= 1.2

1. 68% will fail between ±1σ
19 ± 1(1.2) = [17.8,20.2]

2. 95% will fail between ±2σ
19 ± 2(1.2) = [16.6,21.4]

3. Virtually all will fail at about ±3σ
19 ± 3(1.2) = [15.4,22.6]

Example 2

A tire manufacturer wants to set a minimum mileage guarantee on its new MX100 tire. Tests revealed that the mean mileage is 47900 with a standard deviation of 2050 miles and a normal distribution. The manufacturer wants to set the minimum guaranteed mileage so that no more than 4% of the tires will have to be replaced. What minimum guaranteed mileage should the manufacturer announce?

Solution

Use the z transform

$$z = \frac{X - 47900}{2050}$$

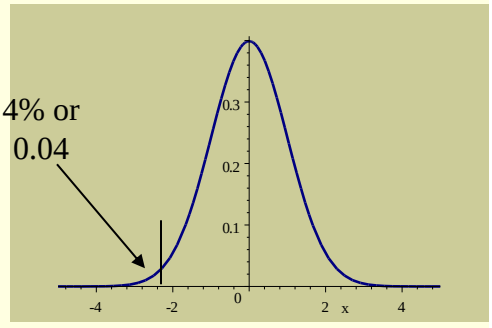
Two unknowns z and X

Use NORMSINV

z=-1.75

Now z is known , find X as

X= 44312



z_α Notation

- In statistical inference need arises for certain small tail areas under the standard normal curve
- z_α will refer to the value on the measurement axis for which α of the area under the z curve lies to the right of z_α .
- z_α is usually referred to as the “critical values”