Oct 19, 2023

Engineering Statistics

Week 2: Measures for the center and variation

Motivation Example

The strength of two materials are measured (in kg/cm^2) under five different conditions and given in the following table:

	I	II	III	IV	V
Material A	10	20	30	40	50
Material B	5	25	30	20	70

Which material is strengthen than the other? Why?

Example

The strength of two materials are measured under five different conditions and given in the following table:

	I	II	III	IV	V	Mean
Material A	10	20	30	40	50	30
Material B	5	25	30	20	70	30

Which material is strengthen than the other? Why?

Example

The strength of two materials are measured under five different conditions and given in the following table:

	I	II	III	IV	V	Mean	Deviation
Material A	10	20	30	40	50	30	15.8
Material B	5	25	30	20	70	30	24.2

Which material is strengthen than the other? Why?

Introduction

When we speak about a dataset, we refer to either a sample or a population. If statistical inference is our goal, we will ultimately wish to use **sample descriptive measures** to make inferences about the corresponding measures of population.

Descriptive measures

Measures for Central Tendency

Measures for Variability

Measures for Central Tendency

The central tendency of the set of measurements - that is, the tendency of the data to cluster, or center, about the certain numerical values.

Measures for Variability

The variability of the set of measurements - that is, the spread of the data.

Measures

Measures for Central Tendency

- mean
  - arithmetic mean
  - weighted mean
  - geometric mean
- median
- mod

Measures for Variability

- range
- mean absolute deviation
- variance
- interquartile range

Mean

The mean of a set of quantitative data is the sum of the measurements, divided by the number of observations contained in the dataset.

Computing the sample mean

Suppose that there is a following sample data with n observations:

$$x_1, x_2, \ldots, x_n$$

the sample mean is denoted by $\bar{x}$ and formulated as follows:

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

1.Mean

Notations

$\bar{x}$: sample mean

μ : population mean

Population mean μ is usually unknown in practice.

We will often use the sample mean to estimate (make an inference about) the population mean.

Properties of Arithmetic Mean

Property 1. Suppose that $\bar{X}$ represents the mean of a dataset, then, if constant value of a is added to each value of this dataset, the mean of the new dataset becomes $\bar{X} + a$, and similar; if constant value of a is subtracted from each value of this dataset, the arithmetic mean of the new dataset becomes $\bar{X} - a$.

Properties of Mean

Property 2. Suppose that $\bar{X}$ represents the mean of a dataset, then, if each value of the dataset is multiplied by a constant such as a ($a > 0$), the mean of the new dataset becomes $a \cdot \bar{X}$, and similar; if each value of the dataset is divided by a constant such as b ($b > 0$), the mean of the new dataset becomes $\bar{X}/b$.

Properties of Mean

Property 3. For a given dataset, the total of the deviations of the values from their arithmetic mean is zero.

$$\sum_{i=1}^{n} (X_i - \bar{X}) = 0$$

Properties of Mean

Property 4. For a given dataset, the sum of the squared deviations of the values from their arithmetic mean is minimum.

$$\sum_{i=1}^{n} (X_i - \bar{X})^2 \to min$$

2.Weighted Mean

$$\bar{X}_W = \frac{\sum_{i=1}^{n} w_i \cdot x_i}{\sum_{i=1}^{n} w_i}$$

w_i: weights for the each observation

$\sum_{i=1}^{k} w_i \cdot x_i$: weighted sum of all observations

3.Geometric Mean

- The geometric mean can be used in many fields, including business (interest rates, proportional growth), communication (aspect ratio of an image), computer science ,medicine, biology (growth rates), and social sciences (population growth).

- In summary, when the data is expressed in terms of percentages, ratios, and indexes, the geometric mean is used. The value of geometric mean is always less than the value of arithmetic mean in a data set.

$$\bar{X}_G = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$$

Example

The mean of strength of the following materials can be calculated as follows:

	I	II	III	IV	V
Material A	10	20	30	40	50
Material B	5	25	30	20	70

$$\bar{x}_A = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{5}(10 + 20 + 30 + 40 + 50) = \frac{1}{5} \times 150 = 30\ kg/cm^2$$

$$\bar{x}_B = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{5}(5 + 25 + 30 + 20 + 70) = \frac{1}{5} \times 150 = 30\ kg/cm^2$$

4.Median

The median of a quantitative data set is the middle observation when the observations are sorted in ascending (or descending) order.

Computing the sample mean

Sort the n observations from the smallest to the largest.

1. If n is odd, the sample median is the middle observation.
2. If n is even, the sample median is the mean of the two middle observations.

4.Median

- Outlier is an observation that is **unusually large or small** relative to the other values.
- There are three possible reason:
    1. Recording or data entry error
    2. Observations from a different population
    3. Rare event

Mean vs. Median

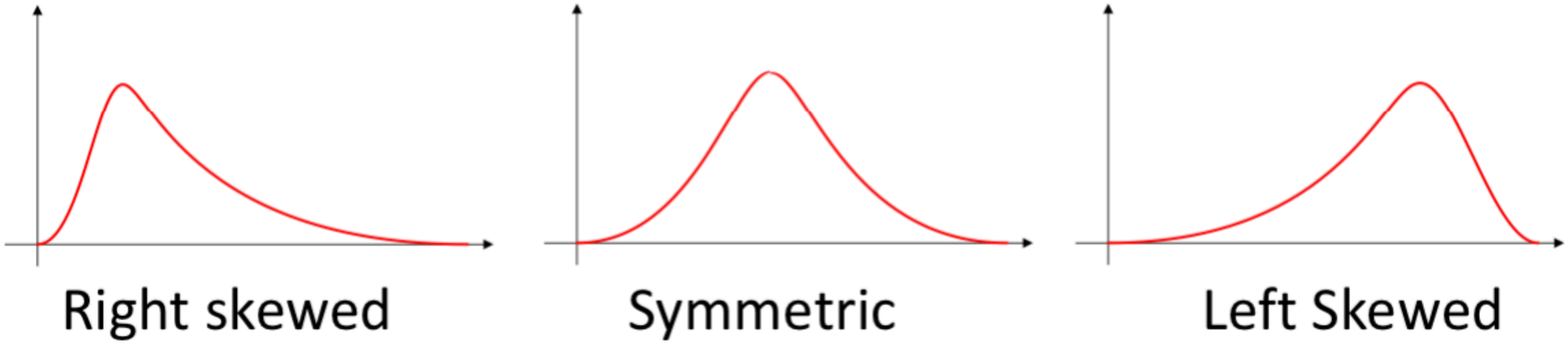
- In general, **outliers effect the mean more than the median**, since these values are used explicitly in the calculation of the mean.
- The median is not effected directly by outliers, since only the middle observation is explicitly used to calculate the median. In other words, **the median is less sensitive than the mean to outliers**.
- Therefore, median is more preferable to mean when there are outliers in the data.

5.Mode

The mode is the observation that occurs most frequently in the data.

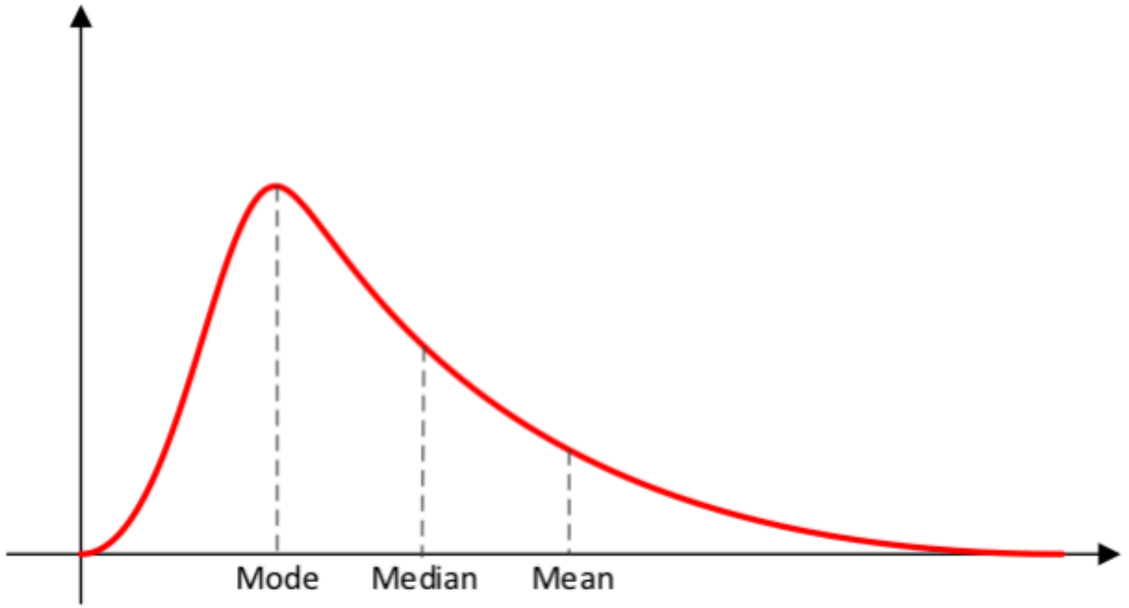
Skewness

A data is referred as skewed if one tail of the distribution has more extreme observations than the other tail.



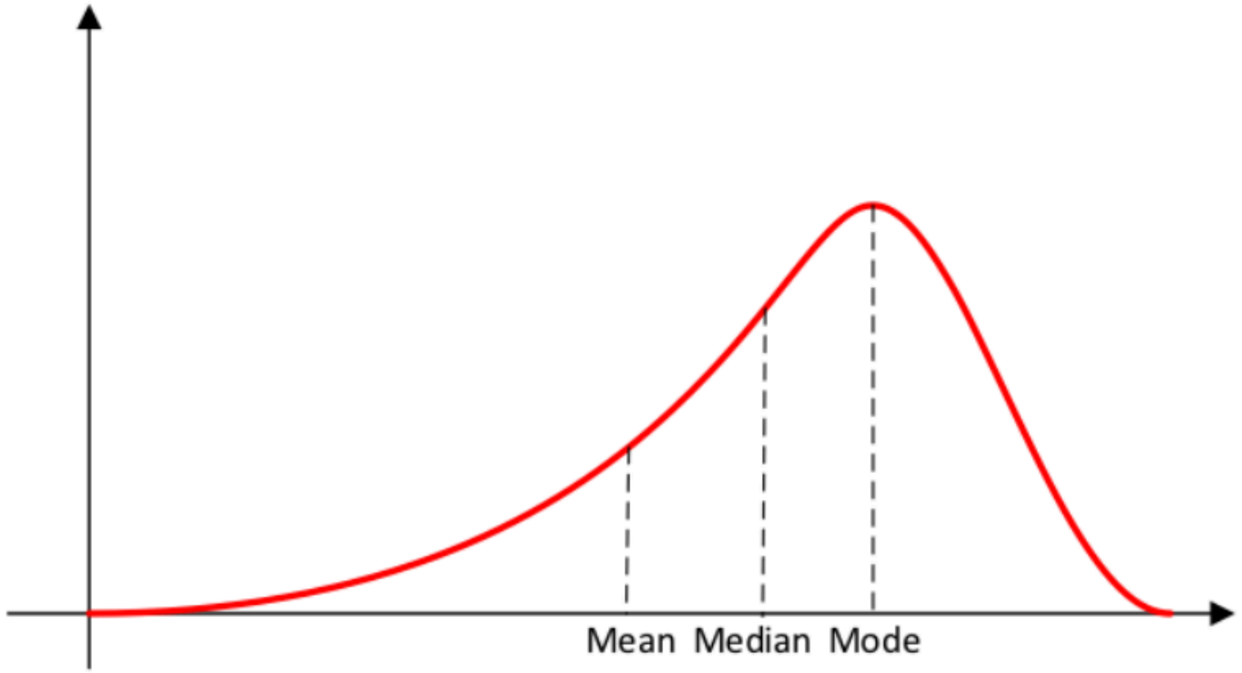
Right skewed data

$$\text{Mode} \leq \text{Median} \leq \text{Mean}$$



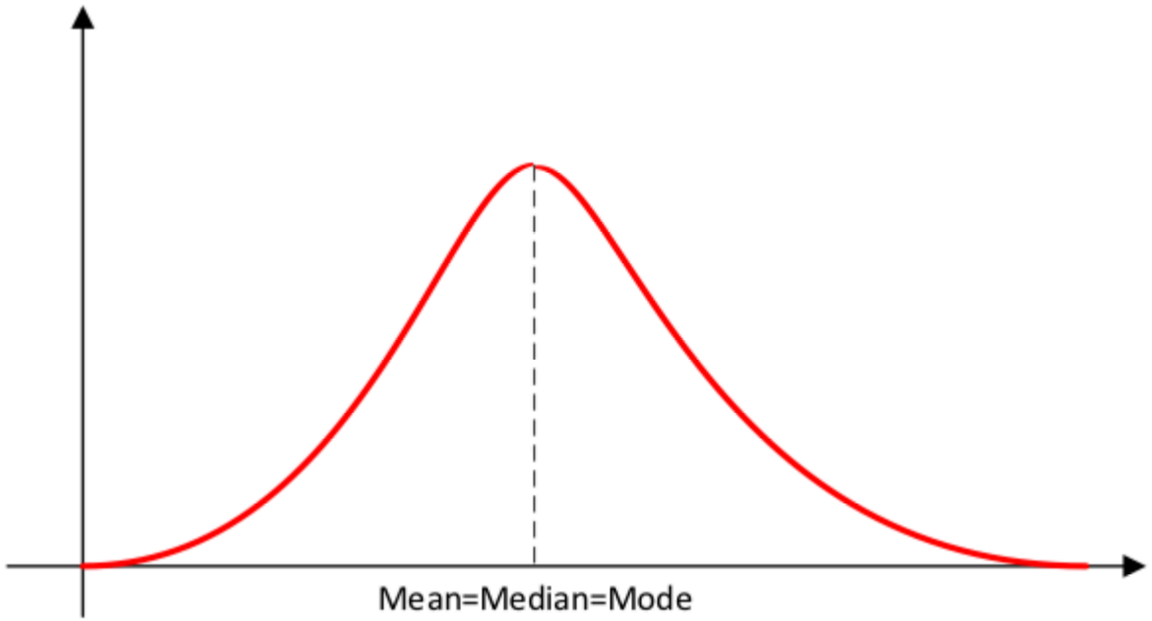
Left skewed data

$$\text{Mean} \leq \text{Median} \leq \text{Mode}$$



Symmetric data

Mean = Median = Mode



Measures of variability

- The mean alone does not provide a complete or sufficient description of data.
- While two data could have the same mean, the individual observations in one data could vary more from the mean than do the observations in the second data.

Range

The range of a quantitative data is equal to the difference between largest and smallest observations.

Computing range

Range = $x_{max} - x_{min}$

Example

The range of strength of the following materials can be calculated as follows:

	I	II	III	IV	V
Material A	10	20	30	40	50
Material B	5	25	30	20	70

$$r_A = max(x_i^A) - min(x_i^A) = 50 - 10 = 40\ kg/cm^2$$

$$r_B = max(x_i^B) - min(x_i^B) = 70 - 5 = 65\ kg/cm^2$$

If the last observation of Material B had taken the value 45, using range might not have been a good choice to reflect the change in the data.

Median Absolute Deviation

In order to avoid the disadvantages of the range, we need a measure of variability that is based on including all measurements in a data set.

Computing median absolute deviation

$$\text{MAD} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

Sample Variance

The sample variance for a sample of n observations is equal to the sum of the squared deviations from the mean, divided by $n - 1$.

Computing the sample variance

Suppose that there is a following sample data with n observations:

$$x_1, x_2, \ldots, x_n$$

the sample variance is denoted by s^2 and formulated as follows:

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

Example

The sample variance of strength of the following materials can be calculated as follows:

	I	II	III	IV	V
Material A	10	20	30	40	50
Material B	5	25	30	20	70

$$s_A^2 = \frac{1}{n_A - 1} \sum_{i=1}^{n_A} (\bar{x}_A - x_i)^2 = \frac{1}{5-1}[(30-10)^2 + (30-20)^2 + (30-30)^2 + (30-40)^2 + (30-50)^2 = \frac{1}{4}(400 + 100 + 0 + 100 + 400) = \frac{1000}{4} = 250$$

$$s_B^2 = \frac{1}{n_B - 1} \sum_{i=1}^{n_B} (\bar{x}_B - x_i)^2 = \frac{1}{5-1}[(30-5)^2 + (30-25)^2 + (30-30)^2 + (30-20)^2 + (30-70)^2 = \frac{1}{4}(625 + 25 + 0 + 100 + 1600) = \frac{2350}{4} = 587.5$$

Sample Standard Deviation

The sample standard deviation s, is defined as the positive square root of the sample variance s^2.

Computing the sample standard deviation

Suppose that there is a following sample data with n observations:

$$x_1, x_2, \ldots, x_n$$

the sample variance is denoted by s and formulated as follows:

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

Sample Standard Deviation

- To compute the variance requires squaring the distances, which then changes the unit of measurement to square units.
- The standard deviation, which is the square root of variance, restores the data to their original measurement unit.
- If the original measurements were in feet, the variance would be in feet squared, but the standard deviation would be in feet.
- The standard deviation measures the average spread around the mean.

Example

The sample standard deviation of strength of the following materials can be calculated as follows:

	I	II	III	IV	V
Material A	10	20	30	40	50
Material B	5	25	30	20	70

$$s_A = \sqrt{\frac{1}{n_A - 1}\sum_{i=1}^{n_A}(\bar{x}_A - x_i)^2} = \frac{1}{5-1}\sqrt{[(30-10)^2 + (30-20)^2 + (30-30)^2 + (30-40)^2 + (30-50)^2} = \sqrt{\frac{1}{4}(400 + 100 + 0 + 100 + 400)} = \sqrt{\frac{1000}{4}} = 15.8 \quad kg/cm^2$$

$$s_B = \sqrt{\frac{1}{n_B - 1}\sum_{i=1}^{n_B}(\bar{x}_B - x_i)^2} = \sqrt{\frac{1}{5-1}[(30-5)^2 + (30-25)^2 + (30-30)^2 + (30-20)^2 + (30-70)^2} = \sqrt{\frac{1}{4}(625 + 25 + 0 + 100 + 1600)} = \sqrt{\frac{2350}{4}} = 24.2 \quad kg/cm^2$$

Coefficient of Variation

The coefficient of variation (CV) is a measure of relative dispersion that express the standard deviation as a percentage of the mean (only provided the mean is positive).

It is used to compare the variabilities of two different data having different means or units.

Computing the coefficient of variation

$$CV = \frac{s}{\bar{x}} \times 100 \text{ if } \bar{x} > 0$$

Example

The coefficient of variation of strength of the following materials can be calculated as follows:

	I	II	III	IV	V
Material A	10	20	30	40	50
Material B	5	25	30	20	70

$$CV_A = \frac{s_A}{\bar{x}_A} = \frac{\sqrt{\frac{1}{n_A - 1}\sum_{i=1}^{n_A}(\bar{x}_A - x_i)^2}}{\frac{1}{n}\sum_{i=1}^{n} x_i} = \frac{15.8}{30} = 0.527$$

$$CV_B = \frac{s_B}{\bar{x}_B} = \frac{\sqrt{\frac{1}{n_B - 1}\sum_{i=1}^{n_B}(\bar{x}_B - x_i)^2}}{\frac{1}{n}\sum_{i=1}^{n} x_i} = \frac{24.2}{30} = 0.808$$

Chebyshev's Theorem and The Empirical Rule

If we are comparing the variability of two samples selected from a population, the sample with the larger standard deviation is the more variable of the two.

Thus, we know how to interpret the standard deviation on a relative or comparative basis, but we haven't explained how it provides a measure of variability for a single sample.

Chebyshev's Theorem and The Empirical Rule

Chebyshev's theorem enables us to make statements about the proportion of data values that must be within a specified number of standard deviation of the mean.

Chebyshev's theorem

At least $(1 - 1/k^2)$ of the data values must be within k standard deviations of the mean, where k is any values greater than 1. In other words, the percent of observations that lie within the interval $(\bar{x} - ks, \bar{x} + ks)$ is at least $100(1 - 1/k^2)\,\%$ where $\bar{x}$ and s are the mean and the standard deviation of the data.

Chebyshev's Theorem and The Empirical Rule

Empirical rule

If the distribution of the data is bell-shaped, then the following empirical rule can be used.

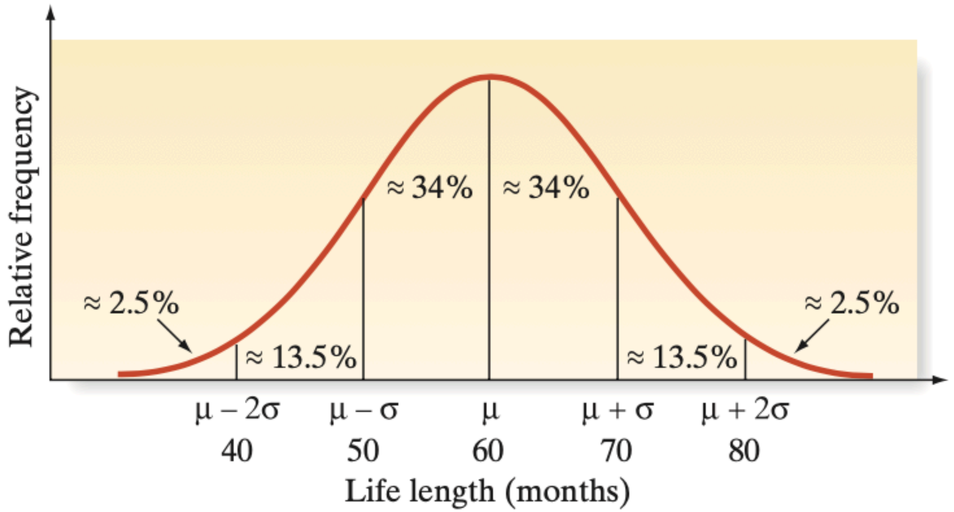
- 68% will be within the range $(\bar{x} - s, \bar{x} + s)$
- 95% will be within the range $(\bar{x} - 2s, \bar{x} + 2s)$
- 99.7% will be within the range $(\bar{x} - 3s, \bar{x} + 3s)$

Example

A manufacturer of automobile batteries claims that the average length of life for its grade A battery is 60 months. However, the guarantee on this brand is for just 36 months. Suppose the standard deviation of the life length is known to be 10 months and the frequency distribution of the life-length data is known to be mound shaped.

Example

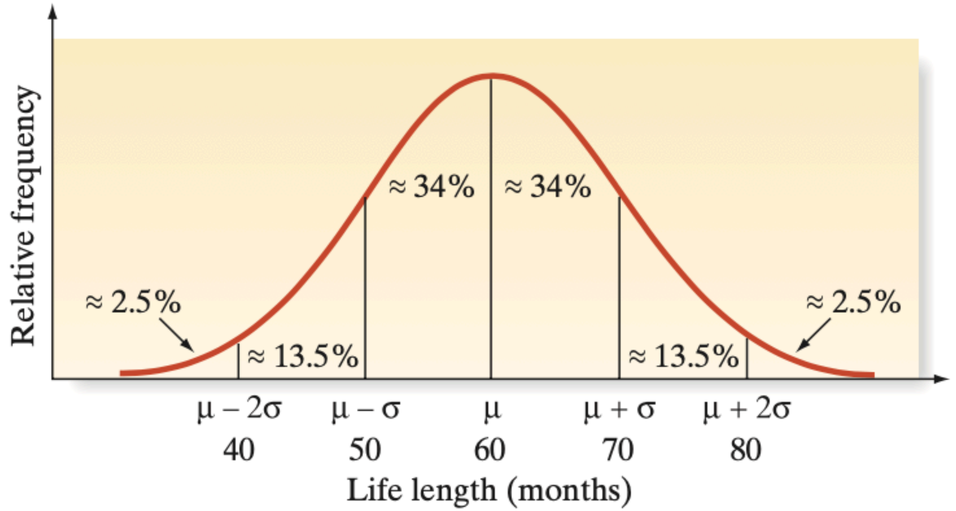
A. Approximately what percentage of the manufacturer's grade A batteries will last more than 50 months, assuming that the manufacturer's claim is true?



It is easy to see in the figure that the percentage of batteries lasting more than 50 months is approximately 34% (between 50 and 60 months) plus 50% (greater than 60 months). Thus, approximately 84% of the batteries should have a life exceeding 50 months.

Example

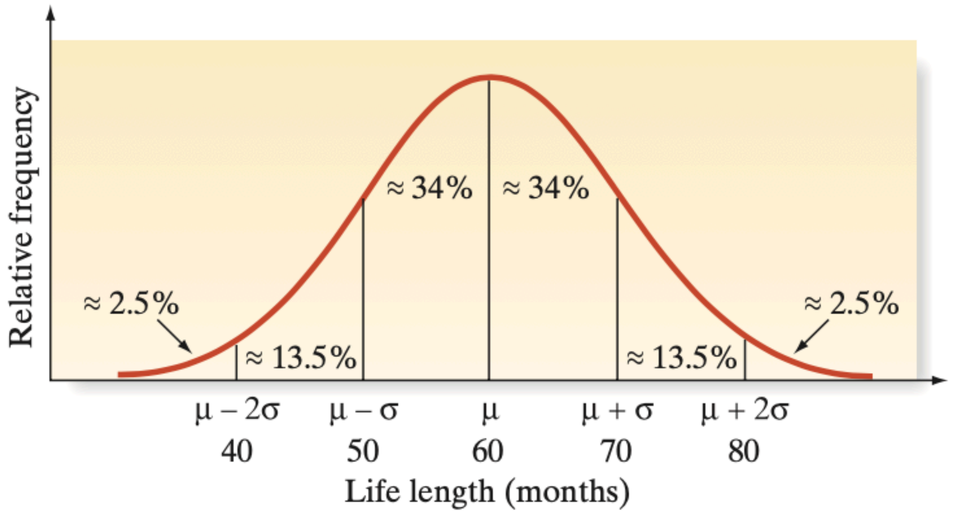
B. Approximately what percentage of the manufacturer's batteries will last less than 40 months, assuming that the manufacturer's claim is true?



The percentage of batteries that last less than 40 months can also be easily determined from the figure: Approximately 2.5% of the batteries should fail prior to 40 months, assuming that the manufacturer's claim is true.

Example

C. Suppose your battery lasts 37 months. What could you infer about the manufacturer's claim?



If you are so unfortunate that your grade A battery fails at 37 months, you can make one of two inferences: Either your battery was one of the approximately 2.5% that fail prior to 40 months, or something about the manufacturer's claim is not true. Because the chances are so small that a battery fails before 40 months, you would have good reason to have serious doubts about the manufacturer's claim. A mean smaller than 60 months or a standard deviation longer than 10 months would each increase the likelihood of failure prior to 40 months.*

Z-scores

- z– scores can be used to identify outliers.
- Recall that the empirical rule allows us to conclude that for data with a bell-shaped distribution, almost all the data values will be within 3 standard deviations of the mean.
- Hence, any measurement with $|z| > 3$ is considered as a potential outlier.

Z-scores

The sample z-score of an observation x_i is:

Computing the z-score

$$z = \frac{x_i - \bar{x}}{s}$$

where $\bar{x}$ and s are the sample mean and standard deviation, respectively.

Z-scores

The sample z-score of an observation x_i is:

Computing the z-score

$$z = \frac{x_i - \bar{x}}{s}$$

where $\bar{x}$ and s are the sample mean and standard deviation, respectively.

Variability

Numerical Measures for Variability

- range
- mean absolute deviation
- variance
- interquartile range

Numerical Measures for Relative Standing

- percentiles
- quartiles
- interquartile range
- boxplot

Variability

Numerical Measures for Variability

- range
- mean absolute deviation
- variance
- interquartile range

Numerical Measures for Relative Standing

- percentiles
- quartiles
- interquartile range
- boxplot

Numerical Measures of Relative Standing

- Descriptive measures of the relationship of a measurement to the rest of the data are called measures of relative standing.
- One measure of the relative standing of a measurement is its percentile ranking.

1.Percentiles

The p-th percentile is a value such that at least p percent of the observations are less than or equal to this value and at least $(100 - p)$ percent of the observations are greater than or equal to this value.

1.Percentiles

Computing the p-th percentile

Step 1. Arrange the data in ascending order (smallest value to largest value)

Step 2. Compute an index i

$$i = \frac{p}{100} \cdot n$$

where p is the percentile of interest and n is the number of observations.

Step 3.

(a) If i is not an integer, round up. The next integer greater than i denotes the position of the p-th percentile.

(b) f i is an integer, the p-th percentile is the average of the values in positions i and $i + 1$.

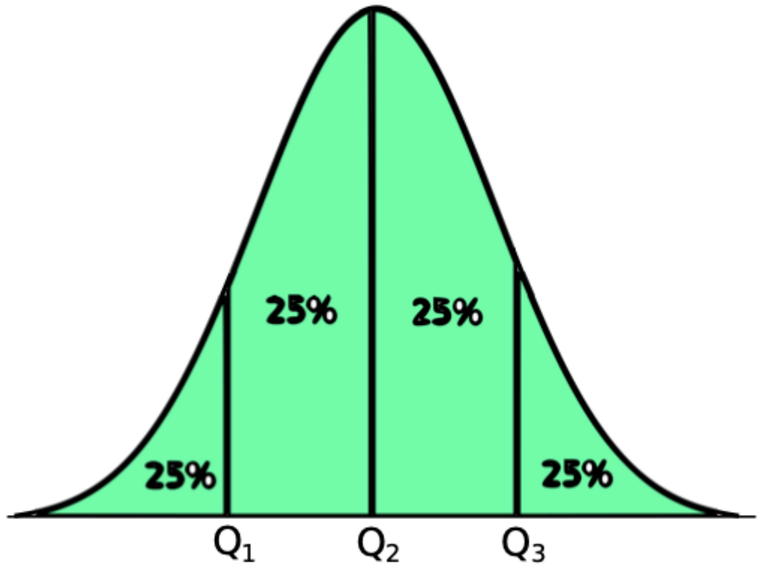
1.Percentiles

Example. Find $85\,\%$ percentile of the following data.

3310, 3355, 3450, 3480, 3480, 3490, 3520, 3540, 3550, 3650, 3730, 3925

2.Quartiles

Percentiles that partition a data set into four categories, each category containing exactly $25\,\%$ of the observations, are called quartiles.



Q_1: first (lower) quartile, or 25th percentile

Q_2: second quartile (median), or 50th percentile

Q_3: third (upper) quartile, or 75th percentile

2.Quartiles

Example. Find the quartiles of the following data.

3310, 3355, 3450, 3480, 3480, 3490, 3520, 3540, 3550, 3650, 3730, 3925

2.Quartiles

Five-Number Summary

The five-number summary refers to the five descriptive measure: minimum, first quartile, median, third quartile, maximum

$$min < Q_1 < median < Q_3 < max$$

3.Interquartile range (IQR)

The interquartile range (IQR) is a measure of variability that is not as sensitive to the presence of outliers as the standard deviation (variance).

Computing the interquartile range

$$IQR = Q_3 - Q_1$$

3.Interquartile range (IQR)

- The IQR is also a measure of the sample variability. The smaller IQR refers the less variability in the data.
- The advantage of IQR over variance (or standard deviation) is that IQR is not affected by the extreme values (or outliers).

3.Interquartile range (IQR)

Example. Find the interquartile range of the following data.

3310, 3355, 3450, 3480, 3480, 3490, 3520, 3540, 3550, 3650, 3730, 3925

4.Boxplot

- A boxplot is a graphical summary of data that is based on the five-number summary.
- A key development of a box plot is the computation of the median, the quartiles Q_1 and Q_3, and the interquartile range $IQR = Q_3 - Q_1$.

4.Boxplot

Drawing a boxplot

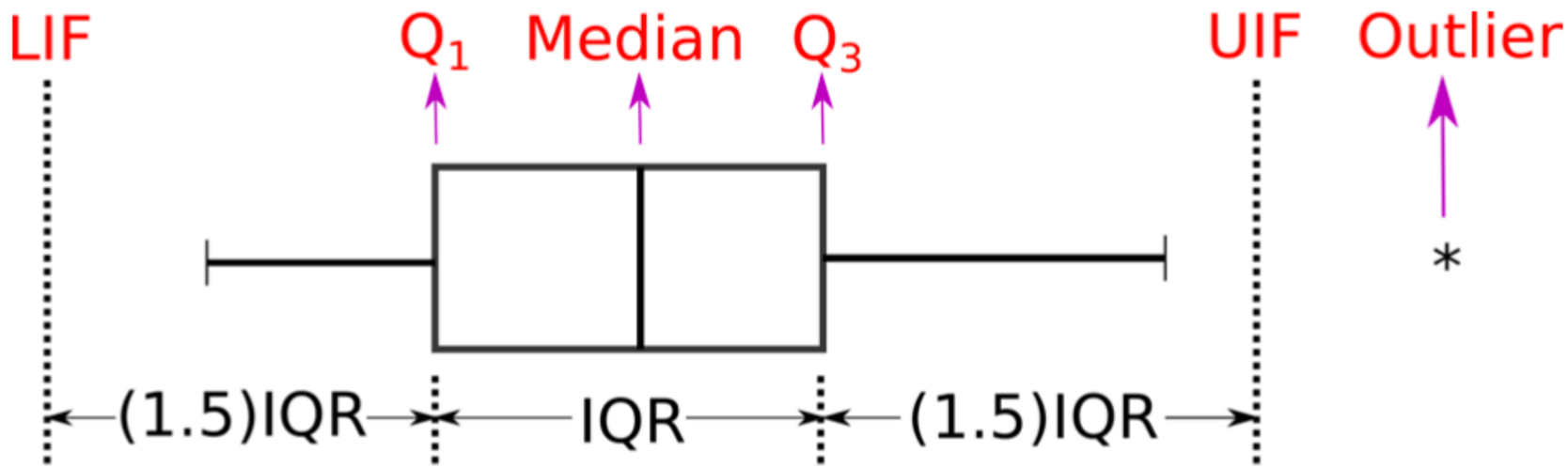
Step 1. A rectangular (the box) is drawn with the ends (the hinges) drawn at the lower and upper quartiles Q_1 and Q_3. The median of the data is shown in the box, usually by a line.

Step 2. The points at distances $1.5 \cdot IQR$ from each hinge mark inner fences of the data set. Lines (the whiskers) are drawn from each hinge to the most extreme observation inside the inner fence. Thus,

$$\text{Lower Inner Fence (LIF)} = Q_1 - 1.5 \cdot IQR$$
$$\text{Upper Inner Fence (UIF)} = Q_3 + 1.5 \cdot IQR$$

Step 3. One symbol (e.g., *) is used to represent observations beyond the inner fences.

4.Boxplot



4.Boxplot

Example. Draw the boxplot of the following data.

3310, 3355, 3450, 3480, 3480, 3490, 3520, 3540, 3550, 3650, 3730, 3925

4.Boxplot

Outliers

- An observation is an outlier if it is more than $1.5 \cdot IQR$ away from the nearest quartile (the nearest end of the box)
- An observation x_i is an outliers if $x_i < Q_1 - 1.5 \cdot IQR$ or $x_i > Q_3 + 1.5 \cdot IQR$.

4.Boxplot

Example. Draw the boxplot of the following data and determine the outlier(s) if any:

2, 3, 3, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7, 7,15

4.Boxplot

- The line inside the box represents the center of the distribution of the data.
- The length of the box is the IQR. Since the IQR is a measure of the sample variability, the length of the box can be useful for the comparison of variabilities of two samples.

4.Boxplot

Example. Two cities provided the following information on public school teachers' salaries.

City	min	first quartile	median	third quartile	max
A	38.400	44.000	48.300	50.400	56.300
B	39.600	46.500	51.200	55.700	61.800

a) Draw a boxplot for the salaries in City A.

b) Draw a boxplot for the salaries in City B.

c) Are there larger differences at the lower or the higher salary levels? Explain.

Course materials

You can download the notes and codes from:

https://github.com/mcavs/ESTUMatse_2022Fall_EngineeringStatistics

Contact

Do not hesitate to contact me on:

https://twitter.com/mustafa__cavus

https://www.linkedin.com/in/mustafacavusphd/

mustafacavus@eskisehir.edu.tr