Data Analytics and Mining

Measures of variability, Z-score, Standardization, Outliers

Data Analytics and Mining, 2024

Majid Sohrabi
National Research University Higher School of Economics

October 18, 2024

Variability

Variability & Range

Variability provides a quantitative measure of the differences between scores in a distribution and describes the degree to which the scores are spread out or clustered together.

There are different measures of variability:

- Range
- Interquartile range
- Variance and standard deviation

Range, which is the distance covered by the scores in a distribution, from the smallest score to the largest score.

$$\text{range} = X_{max} - X_{min}$$

Calculating range:

$$\text{range} = X_{max} - X_{min}$$

$X = [3, 5, 6, 7, 4, 6, 5, 4, 6]$

$\text{range} = 7 - 3 = 4$

$X = [3, 5, 6, 7, 4, 6, 5, 4, 6, 100]$

$\text{range} = 100 - 3 = 97$

When to use the range?

The problem with using the range as a measure of variability is that it is completely **determined by the two extreme values and ignores the other scores in the distribution**.

Thus, a distribution with one unusually large (or small) score will have a large range even if the other scores are all clustered close together.

Because the range does not consider all the scores in the distribution, **it often does not give an accurate description of the variability for the entire distribution**. For this reason, the range is considered to be a crude and unreliable measure of variability.

The range can be calculated only for numerical and some ordinal variables.

Learning check

1. Which of the following sets of scores has the greatest variability?

a. 2, 3, 7, 12

b. 13, 15, 16, 17

c. 24, 25, 26, 27

d. 42, 44, 45, 46

3. How many scores in the distribution are used to compute the range?

a. only 1

b. 2

c. 50% of them

d. all of the scores

Interquartile range (IQR) & Percentiles

Interquartile range (IQR), also called the **midspread, middle 50%**, is a measure of statistical dispersion, being equal to the difference between **75th** and **25th percentiles**, or between upper and lower quartiles.

$$IQR = Q3 - Q1$$

The rank or **percentile rank** of a particular score is defined as the percentage of individuals in the distribution with scores at or below a particular value.

When a score is identified by its percentile rank, the score is called a **percentile**.

Percentiles

X
5
4
3
2
1

Percentiles

X	f
5	1
4	5
3	8
2	4
1	2

Percentiles

X	f	Cf
5	1	20
4	5	19
3	8	14
2	4	6
1	2	2

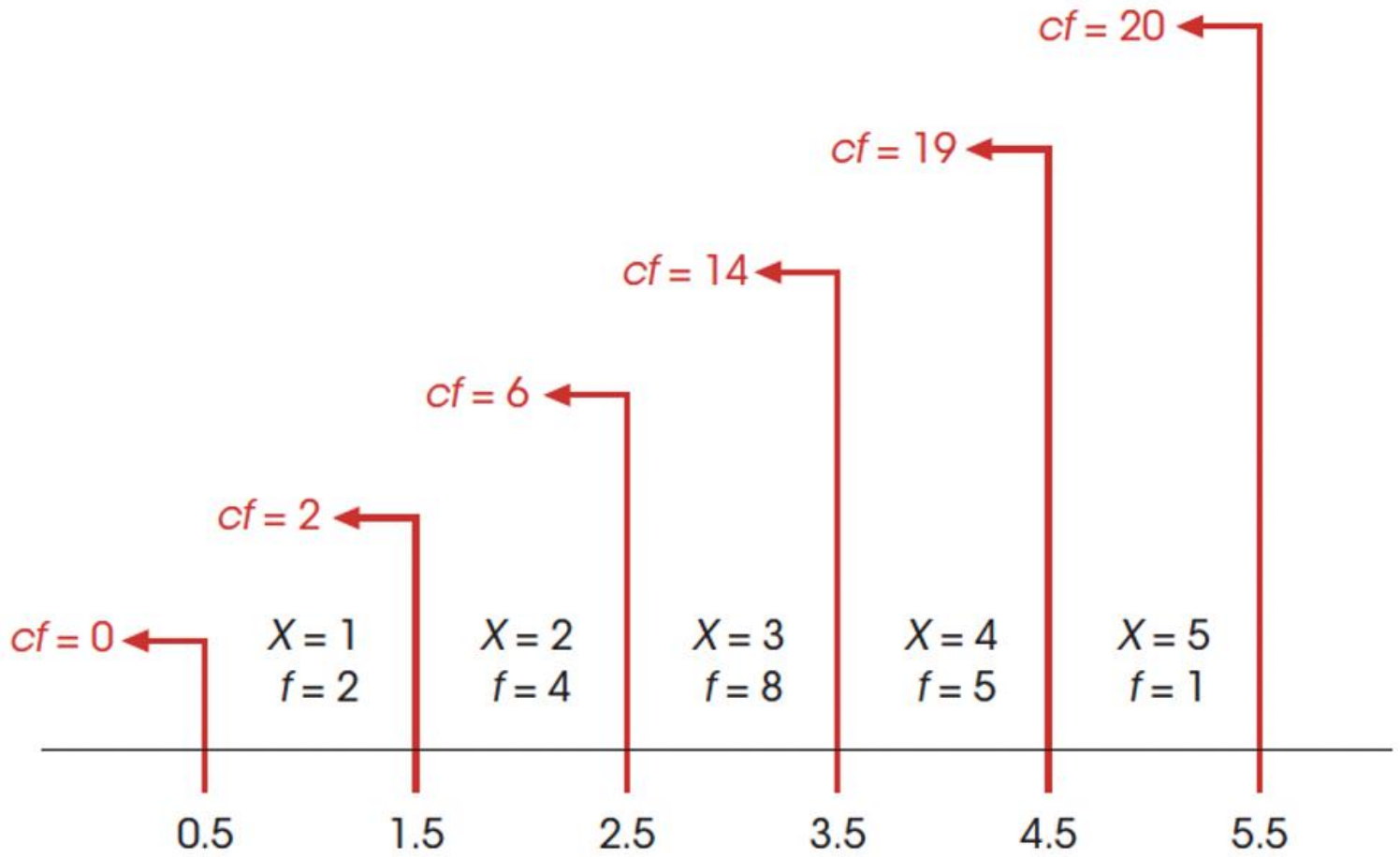
$$c\% = \frac{cf}{N}(100\%)$$

Cumulative percentage is cumulative frequency divided by the total number of observations and multiplied by 100%.

Percentiles

X	f	Cf	$c\%$
5	1	20	100%
4	5	19	95%
3	8	14	70%
2	4	6	30%
1	2	2	10%

Cumulative frequencies and upper limits



Learning check

X	f	Cf	c%
5	1	20	100%
4	5	19	95%
3	8	14	70%
2	4	6	30%
1	2	2	10%

1. What is the 95th percentile?

2. What is the percentile rank for X = 3?

Quartiles

In statistics, a **quartile** is a type of value that divides the number of data points into four parts, or quarters, of more or less equal size. The data must be ordered from smallest to largest to compute quartiles.

The three main quartiles are as follows:

- **The first quartile** (Q1) is defined as the middle number between the smallest number (minimum) and the median of the data set. It is also known as the lower or 25th empirical quartile, as 25% of the data is below this point.
- **The second quartile** (Q2) is the median of a data set; thus 50% of the data lies below this point.
- **The third quartile** (Q3) is the middle value between the median and the highest value (maximum) of the data set. It is known as the upper or 75th empirical quartile, as 75% of the data lies below this point.
- Along with the **minimum and maximum of the data (which are also quartiles)**, the three quartiles described above provide a five-number summary of the data

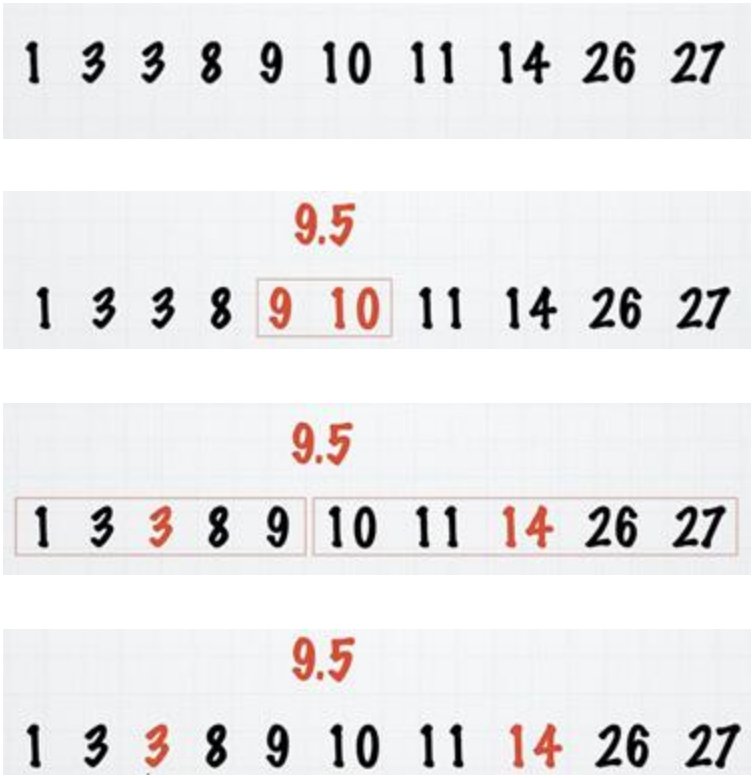
Computing interquartile range

1 3 3 8 9 10 11 14 26 27

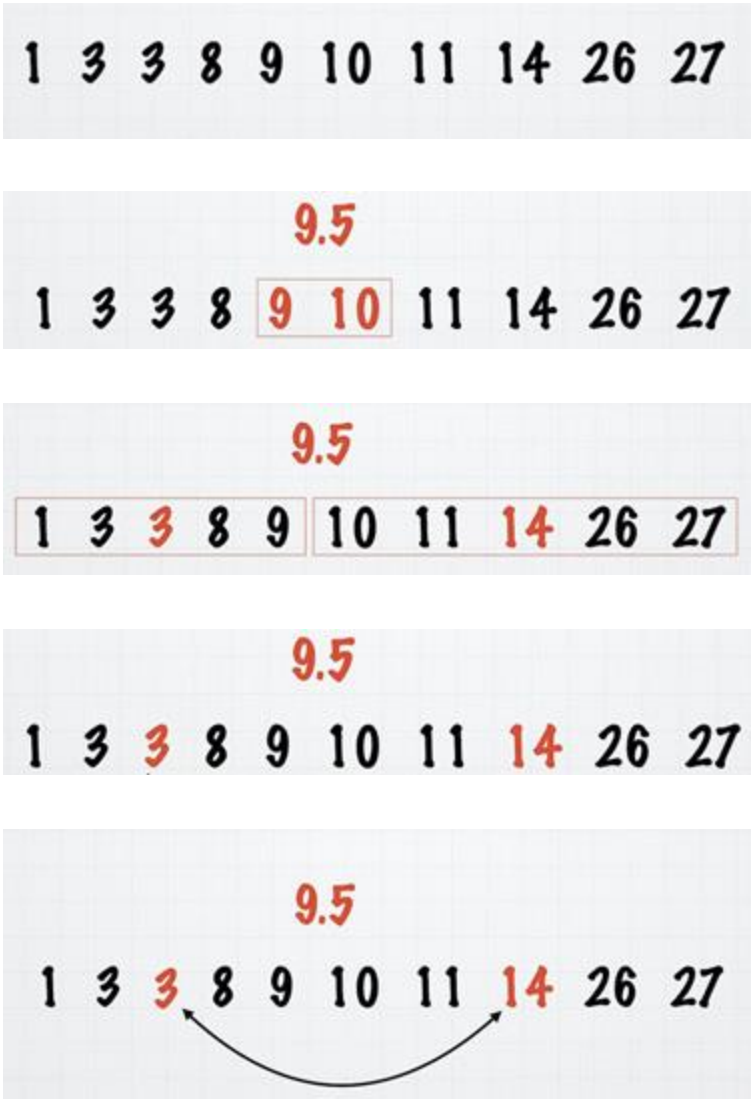
Computing interquartile range

Computing interquartile range

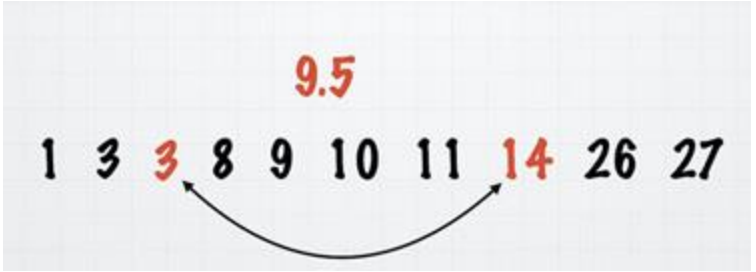
Computing interquartile range



Computing interquartile range

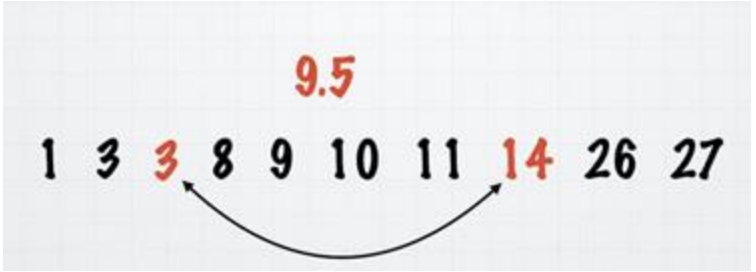


Computing interquartile range

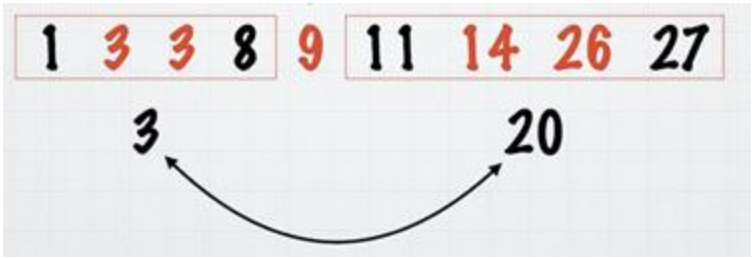


$$IQR = Q3 - Q1 = 14 - 3 = 11$$

Computing interquartile range



$$IQR = Q3 - Q1 = 14 - 3 = 11$$



$$IQR = Q3 - Q1 = 20 - 3 = 17$$

Computing interquartile range

$$\text{range} = X_{max} - X_{max}$$

$X = [3, 5, 6, 7, 4, 6, 5, 4, 6]$

$\text{range} = 7 - 3 = 4$

$X = [3, 5, 6, 7, 4, 6, 5, 4, 6, 100]$

$\text{range} = 100 - 3 = 97$

Computing interquartile range

$\text{range} = X_{max} - X_{max}$

$X = [3, 5, 6, 7, 4, 6, 5, 4, 6]$

$\text{range} = 7 - 3 = 4$

[3, 4, 4, 5, 5, 6, 6, 6, 7]

$X = [3, 5, 6, 7, 4, 6, 5, 4, 6, 100]$

$\text{range} = 100 - 3 = 97$

[3, 4, 4, 5, 5, 6, 6, 6, 7, 100]

Computing interquartile range

range = $X_{max} - X_{max}$

X = [3, 5, 6, 7, 4, 6, 5, 4, 6]

range = 7 – 3 = 4

IQR = Q3 – Q1 = 6 - 4 = 2

X = [3, 5, 6, 7, 4, 6, 5, 4, 6, 100]

range = 100 – 3 = 97

IQR = Q3 – Q1 = 6 - 4 = 2

When to use interquartile range?

Since we discard the extreme values in a distribution when computing the interquartile range, we can say that this measure is more relevant when we want to describe the entire distribution.

Since the IQR as well as a median is computed based on the ranks of the values, it makes sense to choose this measure of variability when you are reporting the median.

As well as the median, the IQR is less affected by extreme values than other measures of variability.

The IQR can be calculated only for numerical and some ordinal variables (e.g. level of satisfaction on the scale from -5 to 5).

Deviation, Variance, Standard deviation

Deviation is the distance from the mean:

$$\text{deviation score} = \mathbf{X - \mu}$$

Variance equals the mean of the squared deviations. Variance is the average squared distance from the mean.

$$\text{variance} = \mathbf{(X - \mu)^2 / N}$$

Standard deviation is the square root of the variance and provides a measure of the standard or average distance from the mean.

Computing the standard deviation

Score X
1
9
5
8
7

Computing the standard deviation

Score X	Deviation $X - \mu$
1	−5
9	3
5	−1
8	2
7	1

Computing the standard deviation

Score X	Deviation $X - \mu$	Squared Deviation $(X - \mu)^2$
1	−5	25
9	3	9
5	−1	1
8	2	4
7	1	1
		40 = the sum of the squared deviations

Computing the standard deviation

Score X	Deviation $X - \mu$	Squared Deviation $(X - \mu)^2$
1	−5	25
9	3	9
5	−1	1
8	2	4
7	1	1
		40 = the sum of the squared deviations

variance = 40 / 5 = 8

standard deviation = $8^{0.5}$ = 2.83

Variance and std. formulas for population and sample

$$\text{population standard deviation} = \sigma = \sqrt{\sigma^2} = \sqrt{\frac{SS}{N}}$$

$$\text{population variance} = \sigma^2 = \frac{SS}{N}$$

Population variance is represented by the symbol σ^2 and equals the mean squared distance from the mean. Population variance is obtained by dividing the sum of squares (SS) by N.

Population standard deviation is represented by the symbol σ and equals the square root of the population variance.

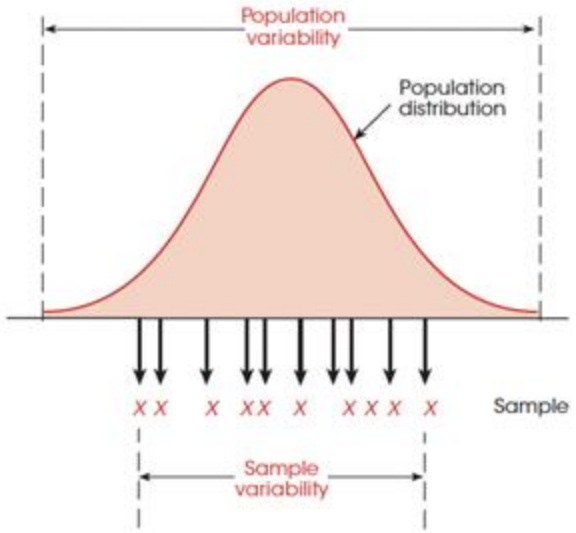
$$\text{sample variance} = s^2 = \frac{SS}{n-1}$$

$$\text{sample standard deviation} = s = \sqrt{s^2} = \sqrt{\frac{SS}{n-1}}$$

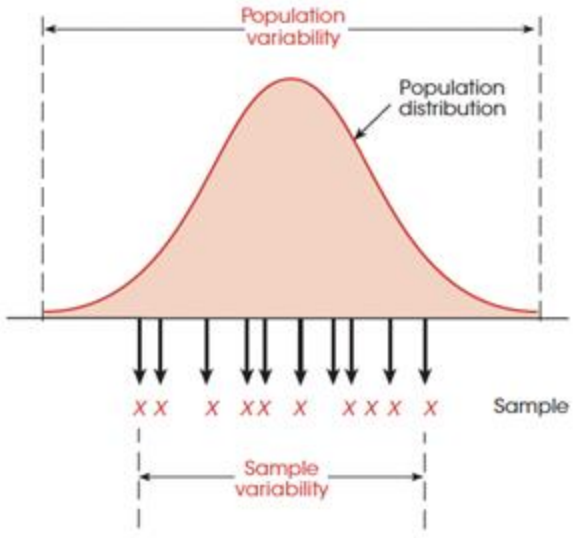
Sample variance is represented by the symbol s^2 and equals the mean squared distance from the mean. Population variance is obtained by dividing the sum of squares (SS) by n – 1.

Sample standard deviation is represented by the symbol s and equals the square root of the population variance.

Why is n-1?


Population variability
Population distribution
Sample
Sample variability

Why is n-1?



The fact that a sample tends to be less variable than its population means that sample variability gives a biased estimate of population variability. This bias is in the direction of underestimating the population value rather than being right on the mark.

Fortunately, the bias in sample variability is consistent and predictable, which means it can be corrected. This is n – 1 in a denominator.

When to use Std.?

Since the standard deviation is computed with the distribution mean in mind, we use this as a pair for mean when describing our distribution.

As well as the mean, the standard deviation might be affected by extreme values.

You can compute the standard deviation only for numerical variables.

Learning check

1. Standard deviation is probably the most commonly used value to describe and measure variability. Which of the following accurately describes the concept of standard deviation?

a. the average distance between one score and another

b. the average distance between a score and the mean

c. the total distance from the smallest score to the largest score

d. one half of the total distance from the smallest score to the largest score

2. What is the variance for the following set of scores? 2, 2, 2, 2, 2

a. 0

b. 2

c. 4

d. 5

3. What is the standard deviation for the following population of scores?
Scores: 1, 3, 7, 4, 5

a. 20

b. 5

c. 4

d. 2

1. If sample variance is computed by dividing by n, instead of $n - 1$, how will the obtained values be related to the corresponding population variance.

a. They will consistently underestimate the population variance.

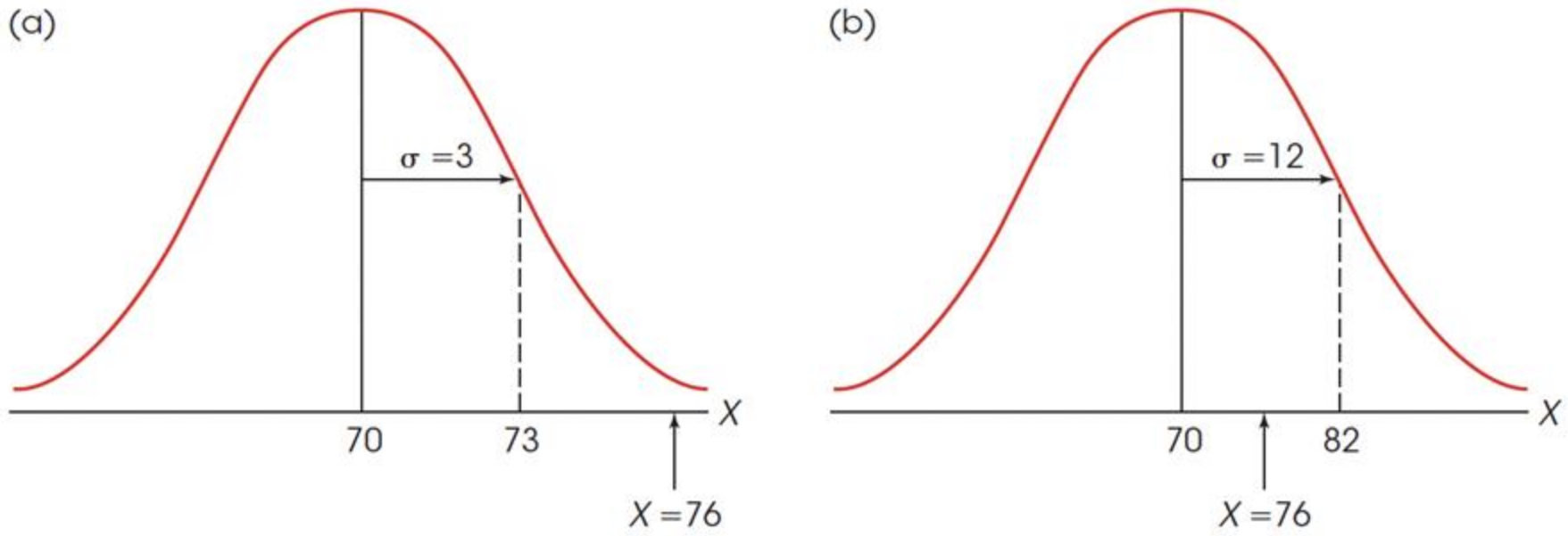
b. They will consistently overestimate the population variance.

c. The average value will be exactly equal to the population variance.

d. The average value will be close to, but not exactly equal to, the population variance.

Z-score

How to compare values from different distributions?



What is the z-score?

The purpose of **z-scores**, or standard scores, is to identify and describe the exact location of each score in a distribution.

A score by itself does not necessarily provide much information about its position within a distribution. These original, unchanged scores that are the direct result of measurement are called raw scores.

To make raw scores more meaningful, they are often transformed into new values that contain more information. This transformation is one purpose for z-scores.

Why do we need z-score?

In summary, the process of transforming X values into z-scores serves two useful purposes:

1. Each z-score tells the exact location of the original X value within the distribution.
2. 2. The z-scores form a standardized distribution that can be directly compared to other distributions that also have been transformed into z-scores.

Learning check

1. If your exam score is $X = 60$, which set of parameters would give you the best grade?

a. $\mu = 65$ and $\sigma = 5$

b. $\mu = 65$ and $\sigma = 2$

c. $\mu = 70$ and $\sigma = 5$

d. $\mu = 70$ and $\sigma = 2$

3. Last week Sarah had a score of $X = 43$ on a Spanish exam and a score of $X = 75$ on an English exam. For which exam should Sarah expect the better grade?

a. Spanish

b. English

c. The two grades should be identical.

d. Impossible to determine without more information

Z-score?

A **z-score** specifies the precise location of each X value within a distribution.

The sign of the z-score (+ or −) signifies whether the score is above the mean (positive) or below the mean (negative). The numerical value of the z-score specifies the distance from the mean by counting the number of standard deviations between X and μ.

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

$$X = \mu + z\sigma$$

Learning check

1. Of the following z-score values, which one represents the most extreme location on the left-hand side of the distribution?

a. $z = +1.00$

b. $z = +2.00$

c. $z = -1.00$

d. $z = -2.00$

2. Of the following z-score values, which one represents the location closest to the mean?

a. $z = +0.50$

b. $z = +1.00$

c. $z = -1.00$

d. $z = -2.00$

3. For a population with $\mu = 100$ and $\sigma = 20$, what is the z-score corresponding to $X = 105$?

a. +0.25

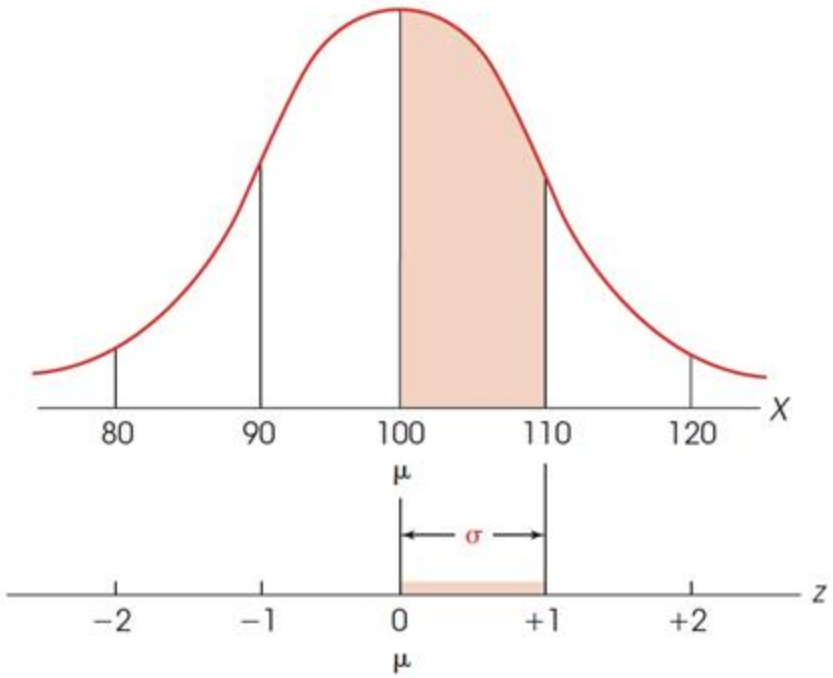
b. +0.50

c. +4.00

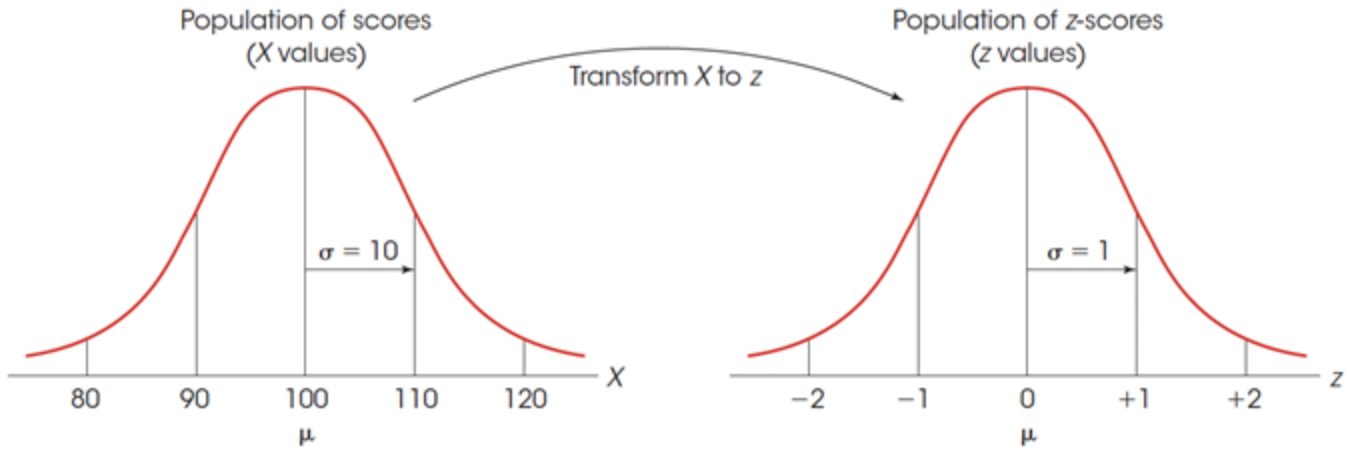
d. +5.00

Standardized distribution

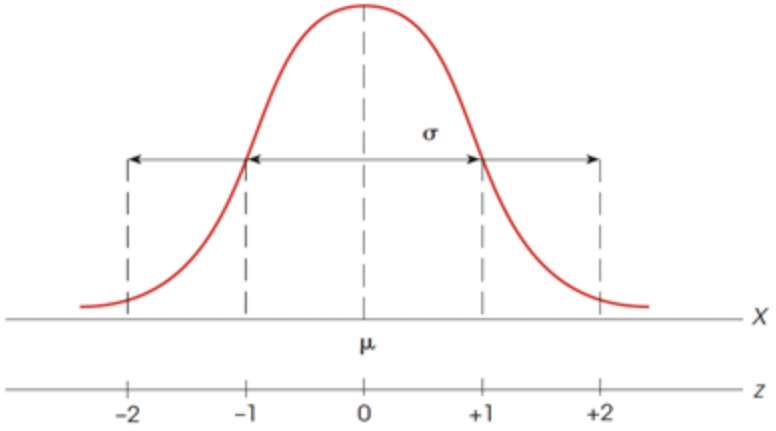
A **standardized distribution** is composed of scores that have been transformed to create predetermined values for μ and σ. Standardized distributions are used to make dissimilar distributions comparable.



Using z-score to standardize distribution



Standardized z-distribution:



A **standardized z-distribution** always has a mean of 0 and a standard deviation of 1.

Learning check

1. A population with $\mu = 85$ and $\sigma = 12$ is transformed into z-scores. After the transformation, the population of z-scores will have a standard deviation of _____

a. $\sigma = 12$

b. $\sigma = 1.00$

c. $\sigma = 0$

d. cannot be determined from the information given

2. A population has $\mu = 50$ and $\sigma = 10$. If these scores are transformed into z-scores, the population of z-scores will have a mean of _____ and a standard deviation of _____.

a. 50 and 10

b. 50 and 1

c. 0 and 10

d. 0 and 1

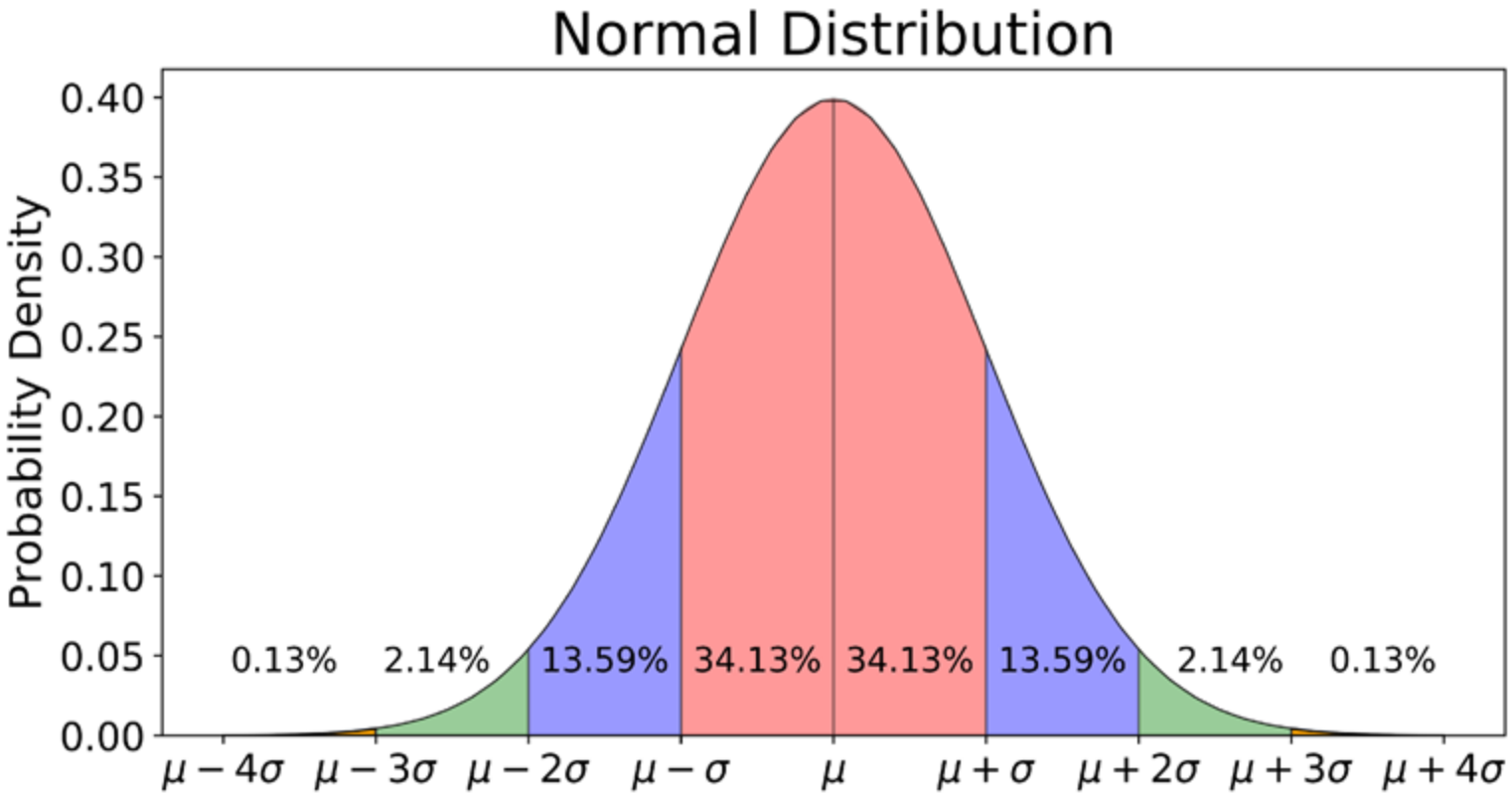
Outliers and standard deviation

In statistics, an **outlier** is a data point that differs significantly from other observations. An outlier can cause serious problems in statistical analyses.

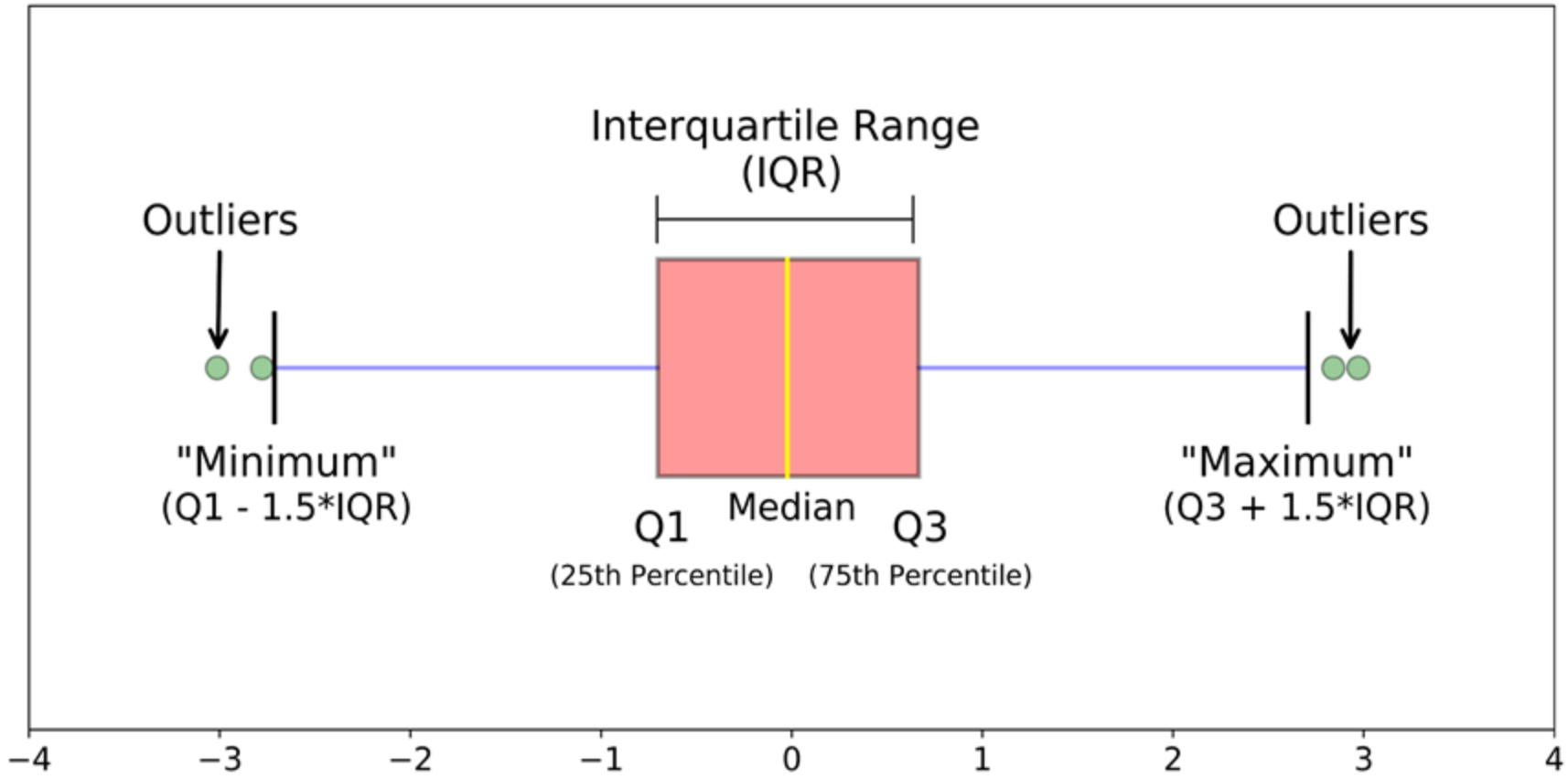
There are different formal definitions for outliers, but the most common ones are those:

- any value that is further than 2-3 standard deviations from the mean is an outlier.
- any value that is further than 3 standard deviations from the mean is a strong outlier.

Proportions and the standardized distribution



Outliers and IQR



Thank you!

Majid Sohrabi

msohrabi@hse.ru

@MSOHRABI_CS