Unit Two
Descriptive Biostatistics

ASSOCIATE PROFESSOR DIANA ARABIAT

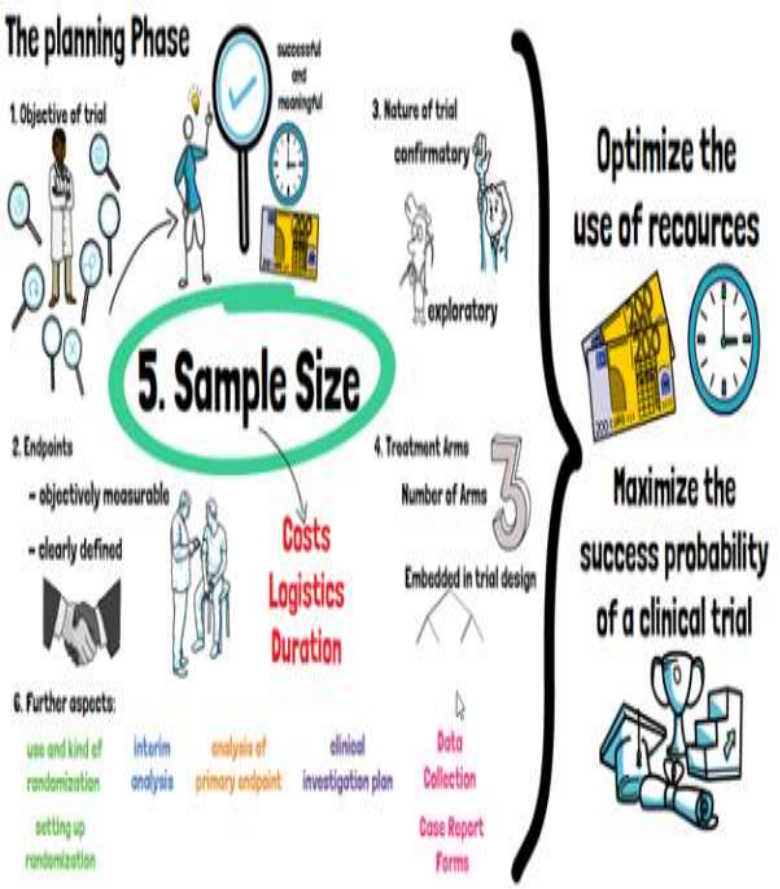
Recap

BIOSTATISTICS

What is the biostatistics?

A branch of applied math that deals with collecting, organizing and interpreting data using well-defined procedures.

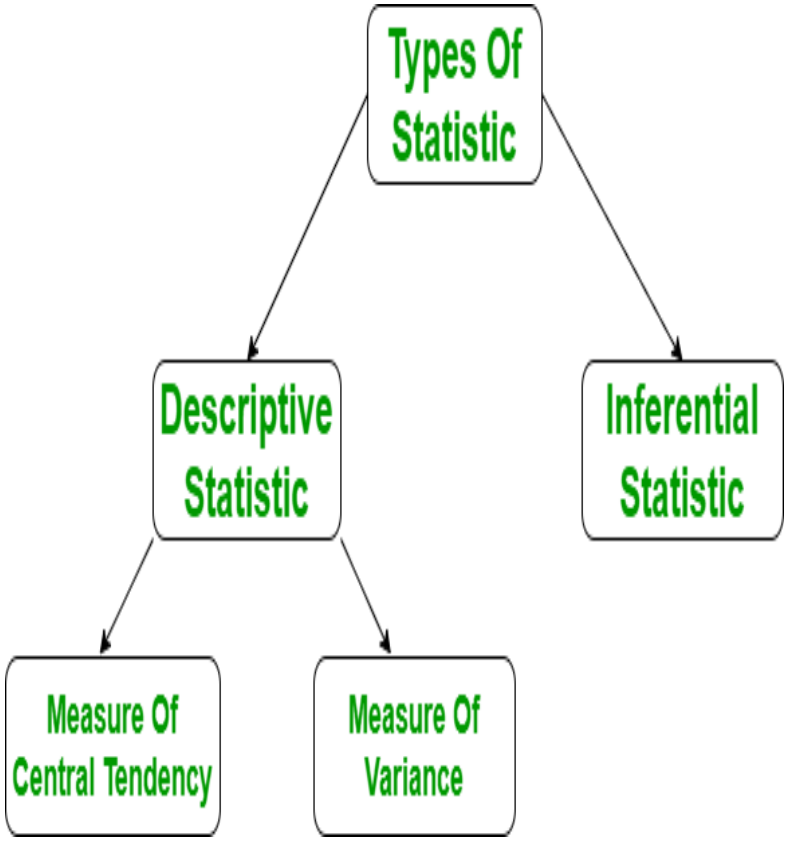
فرع من الرياضيات التطبيقية يتعامل مع جمع البيانات وتنظيمها وتفسيرها باستخدام إجراءات محددة جيداً.



TYPES OF BIOSTATISTICS:

- الإحصائيات الوصفية. يتضمن تنظيم البيانات وتلخيصها وعرضها لجعلها أكثر قابلية للفهم.

- الإحصائيات الاستدلالية. وهي توضح درجة ثقة إحصائية العينة التي تتنبأ بقيمة المَعْلَمة السكانية.



- Descriptive Statistics. It involves organizing, summarizing & displaying data to make them more understandable.
- Inferential Statistics. It reports the degree of confidence of the sample statistic that predicts the value of the population parameter.

What is Descriptive Biostatistics?

The best way to work with data is to summarize and organize them.

Numbers that have not been summarized and organized are called **raw data**.

أفضل طريقة للتعامل مع البيانات هي تلخيصها وتنظيمها.

تسمى الأرقام التي لم يتم تلخيصها وتنظيمها بالبيانات الأولية.

Raw data on number of smarphones owned per family

2	3	4	1	2	2	3	5	2	4
3	2	4	1	3	4	5	3	2	4
2	4	2	3	2	3	2	3	2	3
3	2	3	2	1	2	3	4	1	2
1	2	2	3	3	2	4	2	2	3

Quantitative raw data

Definition

Data is any type of information

Raw data is a data collected as they are received.

Organized data is data organized either in ascending, descending or in a grouped data.

البيانات هي أي نوع من المعلومات

البيانات الأولية هي البيانات التي يتم جمعها فور تلقيها.

البيانات المنظمة هي بيانات منظمة إما تصاعديًا أو تنازليًا أو في بيانات مجمعة.

Organizing Data

After collecting data, the first task for a researcher is to organize and simplify the data so that it is possible to get a general overview of the results.

Raw Data: Data which is not organized is called raw data.

Un-Grouped Data: Data in its original form is called Un-Grouped Data.

Note: Raw data is also called ungrouped data.

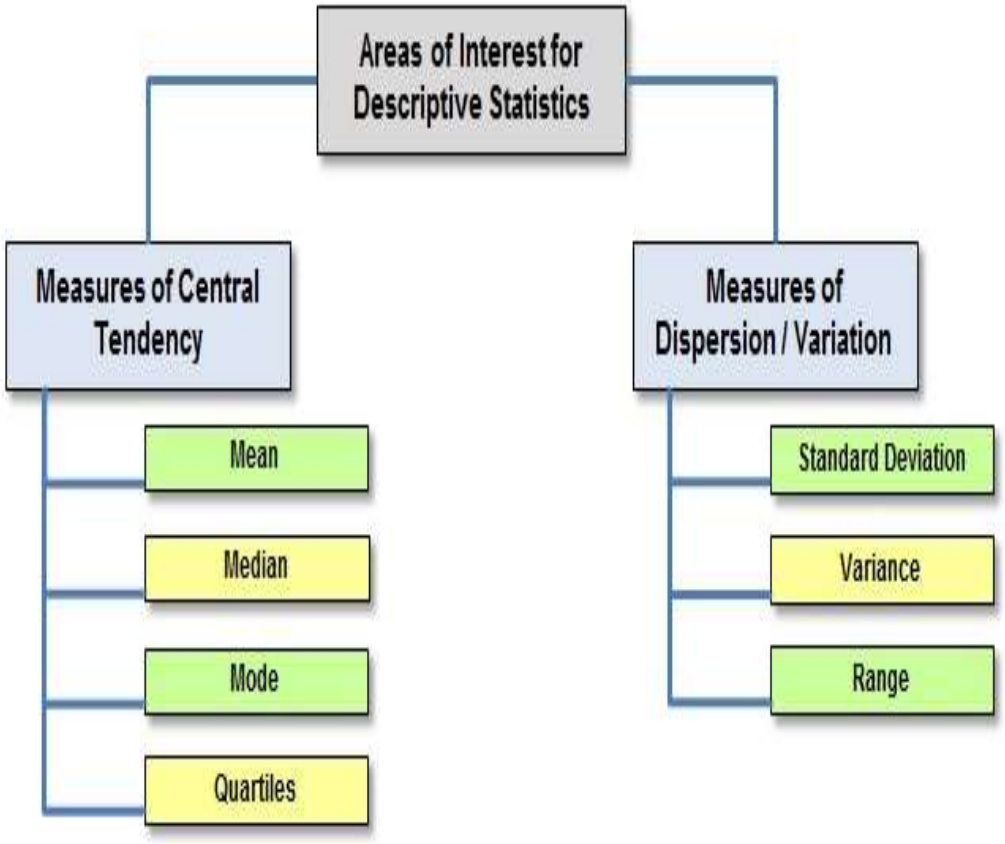
Descriptive Measures

A *descriptive measure* is a single number that is used to describe a set of data.

Descriptive measures include *measures of central tendency* and *measures of dispersion*.

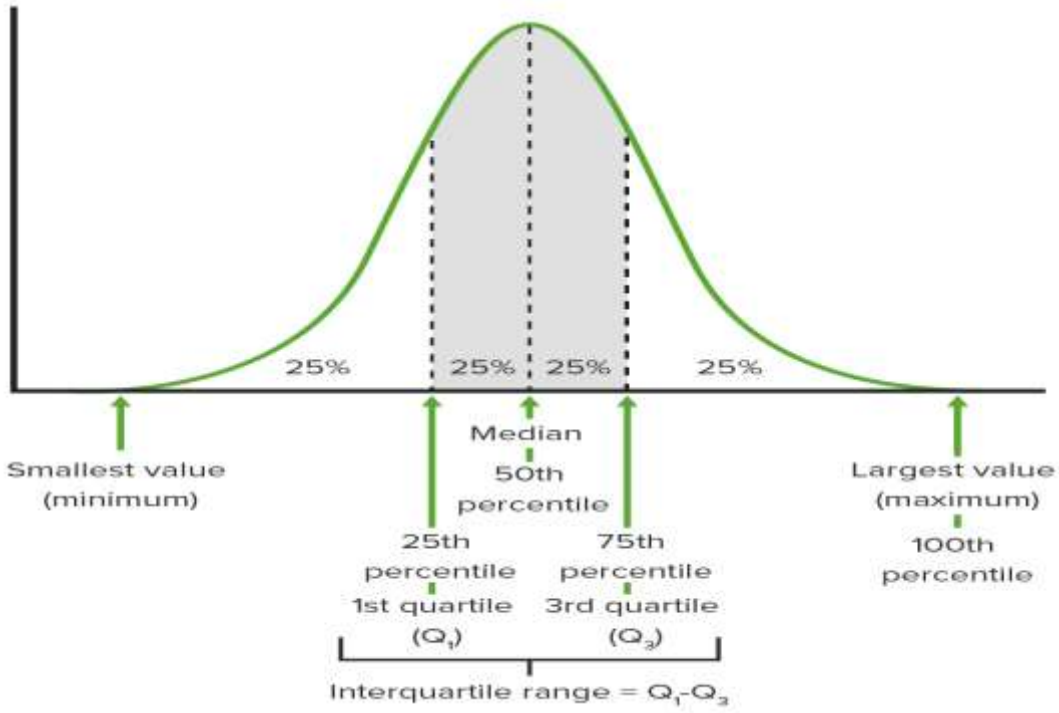
المقياس الوصفي هو رقم واحد يستخدم لوصف مجموعة من البيانات.

وتشمل التدابير الوصفية مقاييس الاتجاه المركزي ومقاييس التشتت.



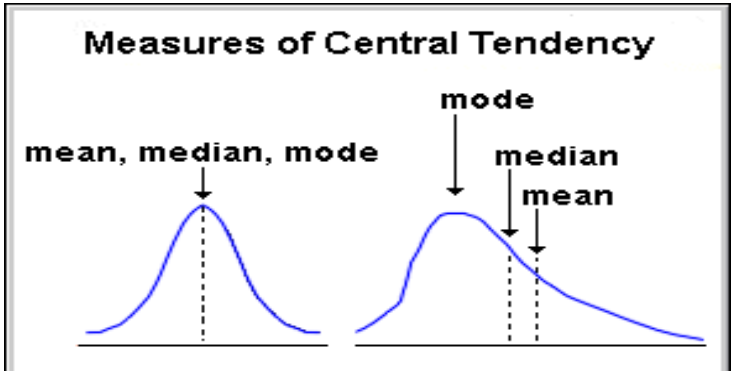
Measures of Dispersion

- Range
- Interquartile range
- Variance
- Standard Deviation
- Coefficient of Variation



Measures of Location

- Measures of central tendency: Mean; Median; Mode



- Measures of noncentral tendency - Quantiles Quartiles; Quintiles; Percentiles

Measures of Location or Central Tendency

This idea of Central tendency refers to the extent to which all the data values group around a typical or central value.

تشير فكرة الاتجاه المركزي هذه إلى المدى الذي تتجمع فيه جميع قيم البيانات حول قيمة نموذجية أو مركزية

Measures of Location

It is a property of the data that they tend to be clustered about a center point.

Measures of *central tendency* (i.e., central location) help find the approximate center of the dataset.

ومن خصائص البيانات أنها تميل إلى التجمع حول نقطة مركزية.

تساعد مقاييس الاتجاه المركزي (أي الموقع المركزي) في العثور على المركز التقريبي لمجموعة البيانات.

Researchers usually do not use the term average, because there are three alternative types of average.

عا دة لا يستخدم الباحثون مصطلح المتوسط، لأن هناك ثلاثة أنوا ع بديلة للمتوسط.

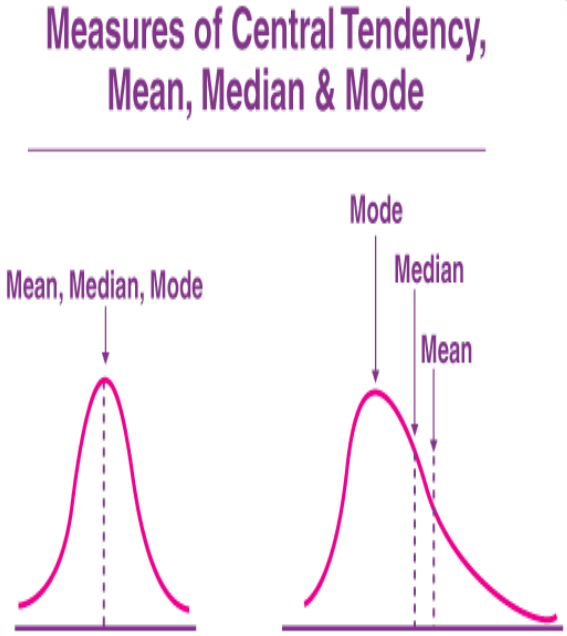
These include the *mean*, the *median*, and the *mode*.

في عالم مثالي، سيكون المتوسط والوسيط والمنوال هو نفسه.

- المتوسط (ليس جزءًا من مجموعة البيانات بشكل عام)
- الوسيط (قد يكون جزءًا من مجموعة البيانات)
- الوضع (دائمًا جزء من مجموعة البيانات)

In a perfect world, the mean, median & mode would be the same.

- Mean (generally not part of the data set)
- Median (may be part of the data set)
- Mode (always part of the data set)



Commonly Used Symbols

For a Sample

$\bar{x}$ sample mean

s^2 sample variance

s sample standard deviation

For a Population

μ population mean

σ^2 population variance

σ population standard deviation

Central Tendency Measures		
Measure	**Formula**	**Description**
Mean	$\sum x/n$	Balance Point
Median	n+1/2 Position	Middle Value when ordered
Mode	None	Most frequent

Notes on Sample Mean $\bar{X}$

Formula

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Summation Sign

- In the formula to find the mean, we use the "summation sign" — $\sum$
- This is just mathematical shorthand for "add up all of the observations"

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + X_3 + \ldots\ldots + X_n$$

Notes on Sample Mean

Also called *sample average* or *arithmetic mean*

Mean for the sample = $\bar{X}$ *or* M, *Mean for population = mew* (μ)

Uniqueness: For a given set of data there is one and only one mean.

Simplicity: The mean is easy to calculate.

Sensitive to extreme values

التفرد: لمجموعة معينة من البيانات يوجد وسط واحد فقط.

البساطة: المتوسط سهل الحساب.

The Median

الوسيط هو القيمة الوسطى للبيانات المطلوبة

للحصول على الوسيط، يجب علينا أولاً إعادة ترتيب البيانات في مصفوفة مرتبة (بترتيب تصاعدي أو تنازلي).

The median is the middle value of the *ordered* data

To get the median, we must first rearrange the data into an ***ordered array*** (in ascending or descending order).

بشكل عام، نقوم بترتيب البيانات من القيمة الأقل إلى القيمة الأعلى.

ولذلك، فإن الوسيط هو قيمة البيانات بحيث يكون نصف الملاحظات أكبر والنصف الآخر أصغر. وهي أيضًا النسبة المئوية الخمسين.

إذا كان n فرديًا، فإن الوسيط هو الملاحظة الوسطى للمصفوفة المرتبة. إذا كانت n زوجية، فهي في منتصف المسافة بين الملاحظتين المركزيتين.

Generally, we order the data from the lowest value to the highest value.

Therefore, the median is the data value such that half of the observations are larger and half are smaller. It is also the 50th percentile.

If n is odd, the median is the middle observation of the ordered array. If n is even, it is midway between the *two* central observations.

The Median

- The median has 3 interesting characteristics:
  - 1. The median is not affected by extreme values, only by the number of observations.
  - 2. Any observation selected at random is just as likely to be greater than the median as less than the median.
  - 3. Summation of the absolute value of the differences about the median is a minimum:

$$\sum_{i=0}^{n} |X_i - Median| = minimum$$

Mean vs. Median

Advantage: The median is less affected by extreme values.

Disadvantage:

- The median takes no account of the precise magnitude of most of the observations and is therefore less efficient than the mean
- If two groups of data are pooled the median of the combined group can not be expressed in terms of the medians of the two original groups but the sample mean can.

$$\bar{x}_{pooled} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

الميزة: الوسيط أقل تأثراً بالقيم المتطرفة.

العيب:
- لا يأخذ الوسيط في الاعتبار الحجم الدقيق لمعظم الملاحظات وبالتالي فهو أقل كفاءة من المتوسط

- إذا تم تجميع مجموعتين من البيانات، فلا يمكن التعبير عن متوسط المجموعة المدمجة من حيث متوسطات المجموعتين الأصليتين ولكن يمكن التعبير عن متوسط العينة.

Comparison of the Mode, the Median, and the Mean

In a normal distribution, the mode , the median, and the mean have the same value.

The mean is the widely reported index of central tendency for variables measured on an interval and ratio scale.

The mean takes each and every score into account.

It also the most stable index of central tendency and thus yields the most reliable estimate of the central tendency of the population.

في التوزيع الطبيعي، يكون للمنوال والوسيط والمتوسط نفس القيمة.

المتوسط هو مؤشر الاتجاه المركزي المُعلن عنه على نطاق واسع للمتغيرات المقاسة على مقياس الفاصل والنسبة.

المتوسط يأخذ كل درجة في الاعتبار.

وهو أيضًا المؤشر الأكثر استقرارًا للنزعة المركزية، وبالتالي يعطي التقدير الأكثر موثوقية للنزعة المركزية للسكان.

Comparison of the Mode, the Median, and the Mean

يتم سحب المتوسط دائمًا في اتجاه الذيل الطويل، أي في اتجاه الدرجات القصوى.

بالنسبة للمتغيرات التي انحرفت بشكل إيجابي (مثل الدخل)، يكون المتوسط أعلى من المنوال أو الوسيط. بالنسبة للمتغيرات المنحرفة سلبا (مثل العمر عند الوفاة) يكون المتوسط أقل.

The mean is always pulled in the direction of the long tail, that is, in the direction of the extreme scores.

For the variables that positively skewed (like income), the mean is higher than the mode or the median. For negatively skewed variables (like age at death) the mean is lower.

When there are extreme values in the distribution (even if it is approximately normal), researchers sometimes report means that have been adjusted for outliers.

To adjust means one must discard a fixed percentage (5%) of the extreme values from either end of the distribution.

عندما تكون هناك قيم متطرفة في التوزيع (حتى لو كانت طبيعية تقريبًا)، يقوم الباحثون أحيانًا بالإبلاغ عن المتوسطات التي تم تعديلها للقيم المتطرفة.

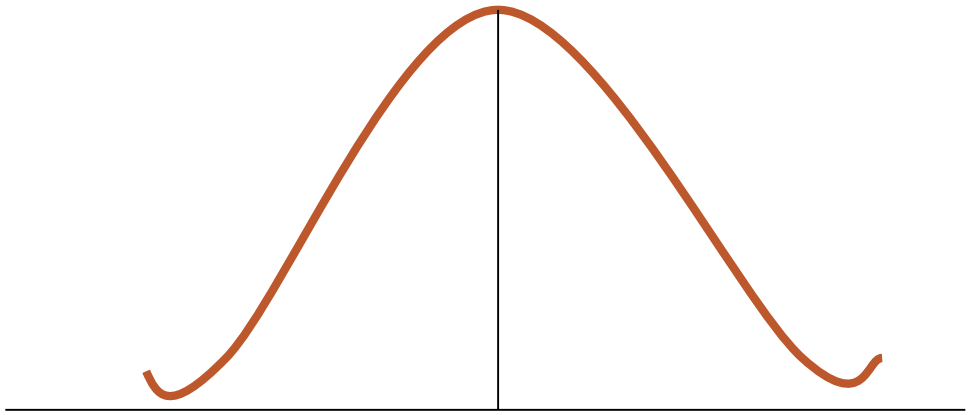
للضبط يعني أنه يجب على المرء تجاهل نسبة ثابتة (5٪) من القيم المتطرفة من أي من طرفي التوزيع.

Distribution Characteristics

Mode: Peak(s)

Median: Equal areas point

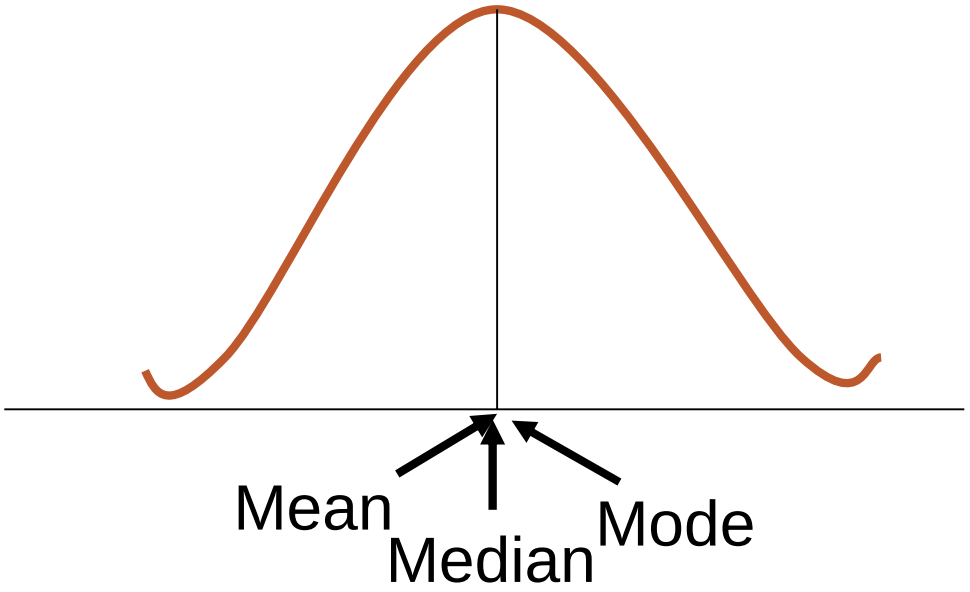
Mean: Balancing point



Shapes of Distributions

Symmetric (Right and left sides are mirror images)

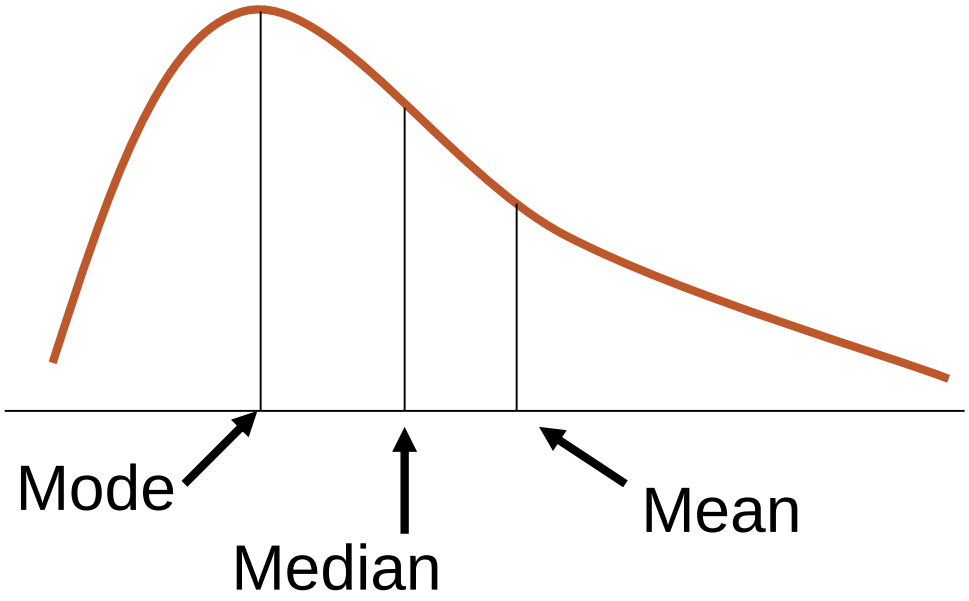
- Left tail looks like right tail
- Mean = Median = Mode



Shapes of Distributions

Right skewed (positively skewed)

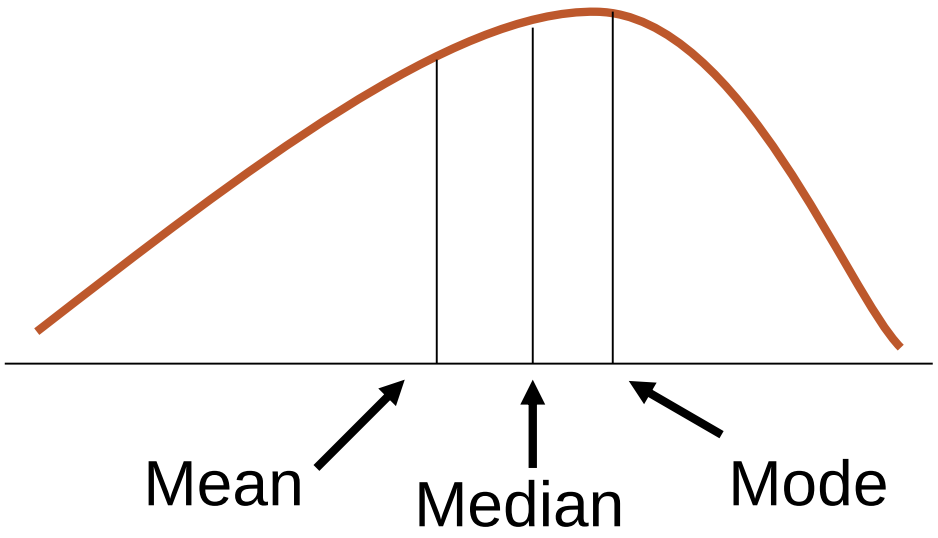
- Long right tail
- Mean > Median



Shapes of Distributions

Left skewed (negatively skewed)

- Long left tail
- Mean < Median



Quantiles

Measures of non-central location used to summarize a set of data

Examples of commonly used quantiles:

- Quartiles
- Quintiles
- Deciles
- Percentiles

Quartiles

Quartiles split a set of ordered data into four parts.

- Imagine cutting a chocolate bar into four equal pieces… How many cuts would you make? (yes, 3!)

Q_1 is the First Quartile

- 25% of the observations are smaller than Q_1 and 75% of the observations are larger

Q_2 is the Second Quartile

- 50% of the observations are smaller than Q_2 and 50% of the observations are larger. Same as the Median. It is also the 50th percentile.

Q_3 is the Third Quartile

- 75% of the observations are smaller than $Q_3$and 25% of the observations are larger

Quartiles

A quartile, like the median, either takes the value of one of the observations, or the value halfway between two observations.

- If n/4 is an integer, the first quartile (Q1) has the value halfway between the (n/4)th observation and the next observation.
- If n/4 is not an integer, the first quartile has the value of the observation whose position corresponds to the next highest integer.

210
220
225
225
225
235
240
250
270
280

$Q_1 = 225$, the 3rd observation

Q_2 = Median - 230

Q_3=250, 3rd observation from the bottom

- The method we are using is an approximation. If you solve this in MS Excel, which relies on a formula, you may get an answer that is slightly different.

Other Quantiles

على غرار ما تعلمناه للتو عن الربعيات، حيث تقوم 3 أرباع بتقسيم البيانات إلى 4 أجزاء متساوية،

- هناك 9 أعشار تقسم التوزيع إلى 10 أجزاء متساوية (أعشار).
- هناك أربعة شرائح تقسم السكان إلى 5 أجزاء متساوية.

Similar to what we just learned about quartiles, where 3 quartiles split the data into 4 equal parts,

- There are 9 *deciles* dividing the distribution into 10 equal portions (tenths).
- There are four *quintiles* dividing the population into 5 equal portions.
- … and 99 percentiles (next slide)

وفي كل هذه الحالات، فإن الاتفاقية هي نفسها. النقطة، سواء كانت ربعية أو عشرية أو مئوية، تأخذ قيمة إحدى الملاحظات أو لها قيمة في منتصف المسافة بين ملاحظتين متجاورتين. ليس من الضروري أبدًا تقسيم الفرق بين ملاحظتين بشكل أكثر دقة.

In all these cases, the convention is the same. The point, be it a quartile, decile, or *percentile*, takes the value of one of the observations or it has a value halfway between two adjacent observations. It is never necessary to split the difference between two observations more finely.

Percentiles

نستخدم 99 نسبة مئوية لتقسيم مجموعة البيانات إلى 100 جزء متساوٍ.

تستخدم النسب المئوية في تحليل نتائج الاختبارات الموحدة. على سبيل المثال، قد تبدو درجة 40 في اختبار موحد بمثابة درجة سيئة، ولكن إذا كانت النسبة المئوية 99، فلا تقلق بشأن إخبار والديك.

ما هي النسبة المئوية Q1؟ Q2 (الوسط)؟ Q3؟

سنستخدم دائمًا برامج الكمبيوتر للحصول على النسب المئوية.

We use 99 *percentiles* to divide a data set into 100 equal portions.

Percentiles are used in analyzing the results of standardized exams. For instance, a score of 40 on a standardized test might seem like a terrible grade, but if it is the 99^{th} percentile, don't worry about telling your parents. ☺

Which percentile is Q1? Q2 (the median)? Q3?

We will always use computer software to obtain the percentiles.

Measures of Dispersion

إنه يشير إلى مدى انتشار الدرجات.

بمعنى آخر، مدى تشابه أو اختلاف المشاركين عن بعضهم البعض في المتغير. وهي إما عينة متجانسة أو غير متجانسة.

لماذا نحتاج إلى النظر في مقاييس التشتت؟ خذ بعين الاعتبار هذا المثال:

شركة على وشك شراء رقائق كمبيوتر يجب أن يبلغ متوسط عمرها 10 سنوات. لدى الشركة خيار بين اثنين من الموردين. من يجب أن يشتروا الرقائق؟ يأخذون عينة من 10 شرائح من كل مورد ويختبرونها. انظر البيانات في الشريحة التالية.

It refers to how spread out the scores are.

In other words, how similar or different participants are from one another on the variable. It is either homogeneous or heterogeneous sample.

Why do we need to look at measures of dispersion?

Consider this example:

A company is about to buy computer chips that must have an average life of 10 years. The company has a choice of two suppliers. Whose chips should they buy? They take a sample of 10 chips from each of the suppliers and test them. See the data on the next slide.

Measures of Dispersion

We see that supplier B's chips have a longer average life.

However, what if the company offers a 3-year warranty?

Then, computers manufactured using the chips from supplier A will have no returns while using supplier B will result in 4/10 or 40% returns.

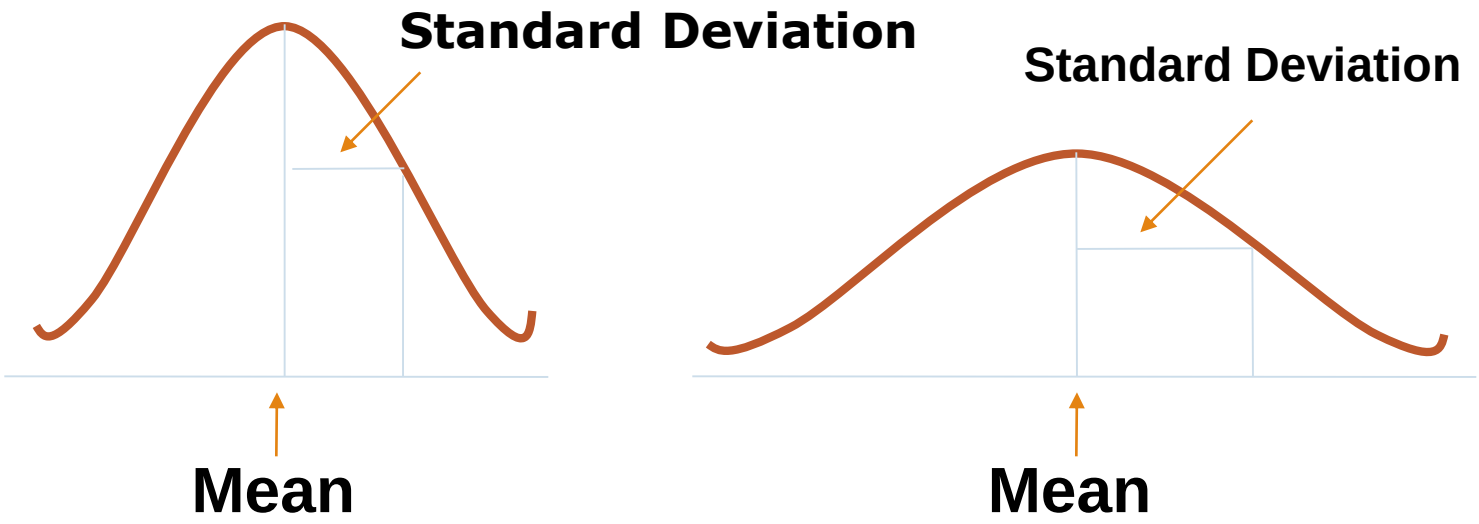
Supplier A chips (life in years)	Supplier B chips (life in years)
11	170
11	1
10	1
10	160
11	2
11	150
11	150
11	170
10	2
12	140
$\overline{X}_A = 10.8$ years	$\overline{X}_B = 94.6$ years
$\text{Median}_A = 11$ years	$\text{Median}_B = 145$ years
$s_A = 0.63$ years	$s_B = 80.6$ years
$\text{Range}_A = 2$ years	$\text{Range}_B = 169$ years

Measures of Dispersion

We will study these five measures of dispersion

- Range
- Interquartile Range
- Standard Deviation
- Variance
- Coefficient of Variation
- Relative Standing.

Normal Distribution



The Range

هو أبسط قياس للتباين، وهو الفرق بين أعلى درجة وأدنى درجة في التوزيع.

في البحث، غالبًا ما يظهر النطاق على أنه الحد الأدنى والحد الأقصى للقيمة، دون درجة الفرق الملخصة. ويقدم ملخصًا سريعًا لتقلب التوزيع. كما يوفر معلومات مفيدة حول التوزيع عندما تكون هناك قيم متطرفة.

النطاق له قيمتان، وهو غير مستقر إلى حد كبير.

Is the simplest measure of variability, is the difference between the highest score and the lowest score in the distribution.

In research, the range is often shown as the minimum and maximum value, without the abstracted difference score.

It provides a quick summary of a distribution's variability.

It also provides useful information about a distribution when there are extreme values.

The range has two values, it is highly unstable.

The Range

Range = Largest Value – Smallest Value

Example: 1, 2, 3, 4, 5, 8, 9, 21, 25, 30

Answer: Range = 30 – 1 = 29.

Pros:

- Easy to calculate

Cons:

- Value of range is only determined by two values
- The interpretation of the range is difficult.
- One problem with the range is that it is influenced by extreme values at either end.

الايجابيات:

- سهولة الحساب

سلبيات:

- يتم تحديد قيمة النطاق بقيمتين فقط
- تفسير النطاق صعب.
- إحدى مشكلات النطاق هي أنه يتأثر بالقيم المتطرفة عند كلا الطرفين.

Standard Deviation

- The standard deviation, *s*, measures a kind of "average" deviation about the mean. It is not really the "average" deviation, even though we may think of it that way.
- Why can't we simply compute the average deviation about the mean, if that's what we want?

$$\frac{\sum_{i=1}^{n}(X_i - \overline{X})}{n}$$

- If you take a simple mean, and then add up the deviations about the mean, as above, this sum will be equal to 0. Therefore, a measure of "average deviation" will not work.

Standard Deviation

- Instead, we use:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

- This is the "definitional formula" for standard deviation.
- The standard deviation has lots of nice properties, including:
  - By squaring the deviation, we eliminate the problem of the deviations summing to zero.
  - In addition, this sum is a minimum. No other value subtracted from X and squared will result in a smaller sum of the deviation squared. This is called the "least squares property."
- Note we divide by (n−1), not n. This will be referred to as a loss of one degree of freedom.

Standard Deviation

The smaller the standard deviation, the better is the mean as the summary of a typical score. E.g. 10 people weighted 150 pounds, the SD would be zero, and the mean of 150 would communicate perfectly accurate information about all the participants wt. Another example would be a heterogeneous sample 5 people 100 pounds and another five people 200 pounds. The mean still 150, but the SD would be 52.7.

In normal distribution there are 3 SDs above the mean and 3 SDs below the mean.

كلما كان الانحراف المعياري أصغر، كان المتوسط أفضل كملخص للنتيجة النموذجية. على سبيل المثال 10 أشخاص يبلغ وزنهم 150 رطلاً، سيكون SD صفراً، وسيقوم متوسط 150 بتوصيل معلومات دقيقة تماماً عن جميع المشاركين بالوزن. مثال آخر سيكون عينة غير متجانسة مكونة من 5 أشخاص 100 جنيه وخمسة أشخاص آخرين 200 جنيه. لا يزال المتوسط 150، لكن SD سيكون 52.7.

في التوزيع الطبيعي هناك 3 SDs فوق المتوسط و 3 SDs تحت المتوسط.

Standard Deviation: N vs. (n-1)

Note that $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}}$ and $s = \sqrt{\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}}$

- You divide by N only when you have taken a census and therefore know the population mean. This is rarely the case.
- Normally, we work with a sample and calculate sample measures, like the sample mean and the sample standard deviation:
- The reason we divide by n-1 instead of n is to assure that *s* is an *unbiased* estimator of σ.
  - We have taken a shortcut: in the second formula we are using the sample mean, $\overline{X}$, a statistic, in lieu of μ, a population parameter. Without a correction, this formula would have a tendency to understate the true standard deviation. We divide by n-1, which increases *s*. This makes it an *unbiased estimator* of σ.
  - We will refer to this as "losing one degree of freedom" (to be explained more fully later on in the course).

Variance

- The variance, s^2, is the standard deviation (s) squared. Conversely, $s = \sqrt{variance}$.

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

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

This is what computer software (e.g., MS Excel or your calculator key) uses.

Coefficient of Variation (CV)

المشكلة في s 2 و s هي أنهما، مثل المتوسط، في الوحدات "الأصلية".

وهذا يجعل من الصعب مقارنة التباين بين مجموعتين من البيانات الموجودة في وحدات مختلفة أو حيث يكون حجم الأرقام مختلفًا تمامًا في المجموعتين. على سبيل المثال،

- لنفترض أنك ترغب في مقارنة سهمين أحدهما بالدولار والآخر بالين؛ إذا كنت تريد معرفة أيهما أكثر تقلبًا، فيجب عليك استخدام معامل الاختلاف.
- ليس من المناسب أيضًا مقارنة مخزونين بأسعار مختلفة إلى حد كبير حتى لو كان كلاهما في نفس الوحدات.
- سيكون الانحراف المعياري للسهم الذي يتم بيعه بحوالي 300 دولار أمريكي مختلفًا تمامًا عن الآخر الذي يبلغ سعره حوالي 0.25 دولار أمريكي.

سيكون معامل الاختلاف مقياسًا أفضل للتشتت في هذه الحالات من الانحراف المعياري (انظر المثال في الشريحة التالية).

The problem with s^2 and s is that they are both, like the mean, in the "original" units.

This makes it difficult to compare the variability of two data sets that are in different units or where the magnitude of the numbers is very different in the two sets. For example,

- Suppose you wish to compare two stocks and one is in dollars and the other is in yen; if you want to know which one is more volatile, you should use the coefficient of variation.
- It is also not appropriate to compare two stocks of vastly different prices even if both are in the same units.
- The standard deviation for a stock that sells for around $300 is going to be very different from one with a price of around $0.25.

The *coefficient of variation* will be a better measure of dispersion in these cases than the standard deviation (see example on the next slide).

$$CV = \frac{s}{\bar{x}}(100\%)$$

Coefficient of Variation (CV)

السيرة الذاتية من حيث النسبة المئوية. ما نقوم بحسابه فعليًا هو النسبة المئوية لمتوسط العينة التي تمثل الانحراف المعياري. إذا كانت قيمة CV 100%، فهذا يشير إلى أن متوسط العينة يساوي الانحراف المعياري للعينة. وهذا من شأنه أن يوضح أن هناك قدرًا كبيرًا من التباين في مجموعة البيانات. ومن الواضح أن 200% سيكون أسوأ.

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

CV is in terms of a percent. What we are in effect calculating is what percent of the sample mean is the standard deviation. If CV is 100%, this indicates that the sample mean is equal to the sample standard deviation. This would demonstrate that there is a great deal of variability in the data set. 200% would obviously be even worse.

The Interquartile range (IQR)

The Interquartile range (IQR) is the score at the 75th percentile or 3rd quartile (Q3) minus the score at the 25th percentile or first quartile (Q1). Are the most used to define outliers.

It is not sensitive to extreme values.

النطاق الرباعي (IQR) هو النتيجة عند المئين الخامس والسبعين أو الربع الثالث (Q3) مطروحًا منها النتيجة عند المئين الخامس والعشرين أو الربع الأول (Q1). هي الأكثر استخدامًا لتحديد القيم المتطرفة.

أنها ليست حساسة للقيم المتطرفة.

Standard Scores

There are scores that are expressed in terms of their relative distance from the mean. It provides information not only about rank but also distance between scores.

It often called Z-score.

هناك درجات يتم التعبير عنها من حيث بعدها النسبي عن المتوسط. فهو يوفر معلومات ليس فقط عن الرتبة ولكن أيضًا عن المسافة بين الدرجات.

غالبًا ما يطلق عليه Z-score.

Z Score

Is a standard score that indicates how many SDs from the mean a particular values lies.

Z = Score of value – mean of scores divided by standard deviation.

هي درجة قياسية تشير إلى عدد الانحراف المعياري (SD) من المتوسط الذي تكمن فيه قيم معينة.

Z = درجة القيمة – متوسط الدرجات مقسومًا على الانحراف المعياري.

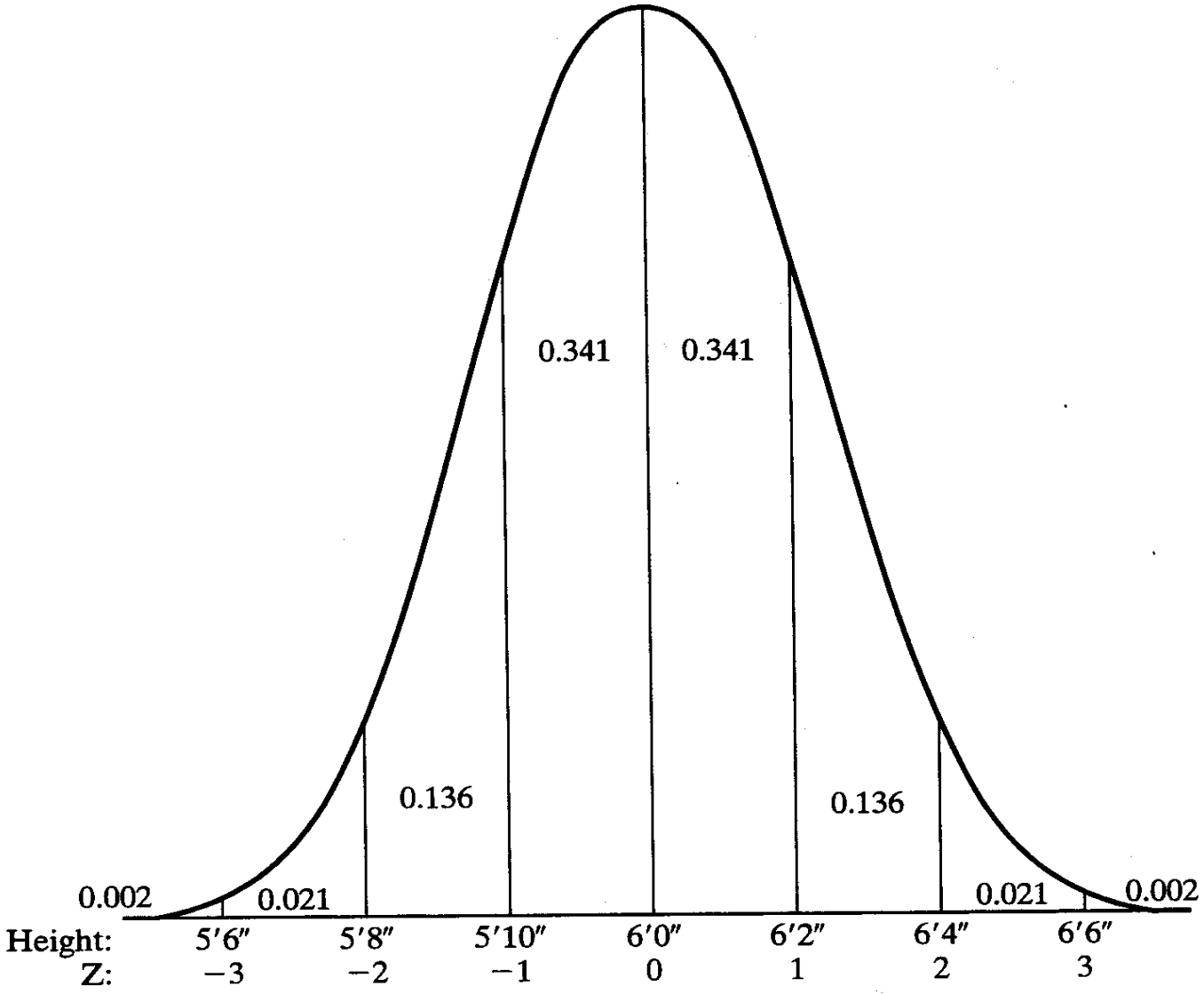
Standard Normal Scores

How many standard deviations away from the mean are you?

Standard Score (Z) =

$$\frac{\textit{Observation} - \textit{mean}}{\textit{Standard deviation}}$$

"Z" is normal with mean 0 and standard deviation of 1.


0.341
0.341
0.136
0.136
0.002
0.021
0.021
0.002
Height:
5′6″
5′8″
5′10″
6′0″
6′2″
6′4″
6′6″
Z:
−3
−2
−1
0
1
2
3

Standard Normal Scores

A standard score of:

Z = 1: The observation lies one SD above the mean

Z = 2: The observation is two SD above the mean

Z = -1: The observation lies 1 SD below the mean

Z = -2: The observation lies 2 SD below the mean

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

Standard Normal Scores

Example: Male Blood Pressure, mean = 125, s = 14 mmHg

- BP = 167 mmHg

$$Z = \frac{167 - 125}{14} = 3.0$$

- BP = 97 mmHg

$$Z = \frac{97 - 125}{14} = -2.0$$

What is the Usefulness of a Standard Normal Score?

It tells you how many SDs (s) an observation is from the mean

Thus, it is a way of quickly assessing how "unusual" an observation is

Example: Suppose the mean *BP* is 125 mmHg, and standard deviation = 14 mmHg

- Is 167 mmHg an unusually high measure?
- If we know Z = 3.0, does that help us?

ويخبرك بعدد SD (S) الملاحظة من المتوسط

وبالتالي، فهي طريقة للتقييم السريع لمدى "غرابة" الملاحظة

مثال: لنفترض أن متوسط ضغط الدم هو 125 ملم زئبقي، والانحراف المعياري = 14 ملم زئبق

- هل يعتبر 167 ملم زئبق مقياسًا مرتفعًا بشكل غير عادي؟
- إذا علمنا أن Z = 3.0 فهل يساعدنا ذلك؟

Standardizing Data: Z-Scores

يمكننا تحويل الدرجات الأصلية إلى درجات جديدة مع x
0 = و 1 = s.
سيعطينا هذا رقمًا نقيًا بدون وحدات قياس. ▸
أي درجة أقل من المتوسط ستكون الآن سلبية.
أي درجة في المتوسط ستكون 0.
أي درجة أعلى من المتوسط ستكون إيجابية.

- We can convert the original scores to new scores with $\overline{X} = 0$ and $s = 1$.
- This will give us a pure number with no units of measurement.
- Any score below the mean will now be negative.
- Any score at the mean will be 0.
- Any score above the mean will be positive.

Standardizing Data: Z-Scores

بغض النظر عن ما تقوم بقياسه، فإن درجة Z التي تزيد عن +5 أو أقل من -5 ستشير إلى درجة غير عادية للغاية.

بالنسبة للبيانات الموحدة، إذا تم توزيعها بشكل طبيعي، فإن 95٪ من البيانات ستكون بين ±2 انحراف معياري عن المتوسط.

إذا كانت البيانات تتبع التوزيع الطبيعي

- 95% من البيانات ستكون بين -1.96 و+1.96.
- 99.7% من البيانات تقع بين -3 و+3.
- 99.99% من البيانات تقع بين -4 و+4.

No matter what you are measuring, a Z-score of more than +5 or less than – 5 would indicate a very, very unusual score.

For standardized data, if it is normally distributed, 95% of the data will be between ±2 standard deviations about the mean.

If the data follows a normal distribution,

- 95% of the data will be between -1.96 and +1.96.
- 99.7% of the data will fall between -3 and +3.
- 99.99% of the data will fall between -4 and +4.

Major reasons for using Index (descriptive statistics)

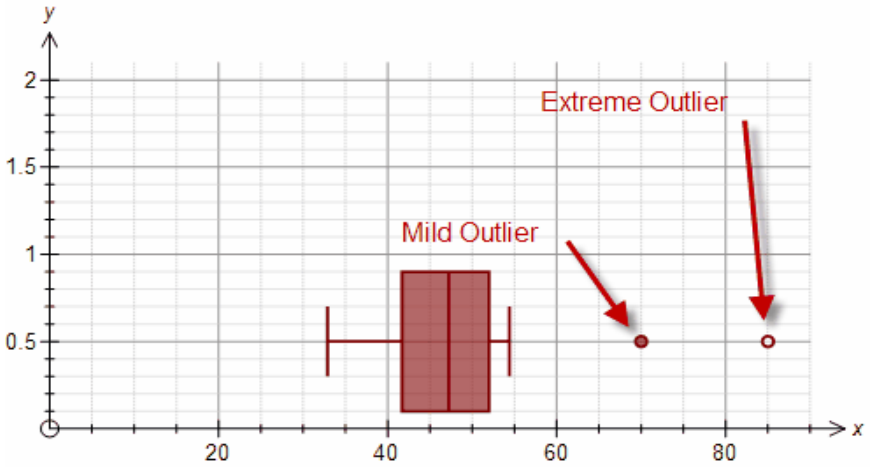
فهم البيانات.

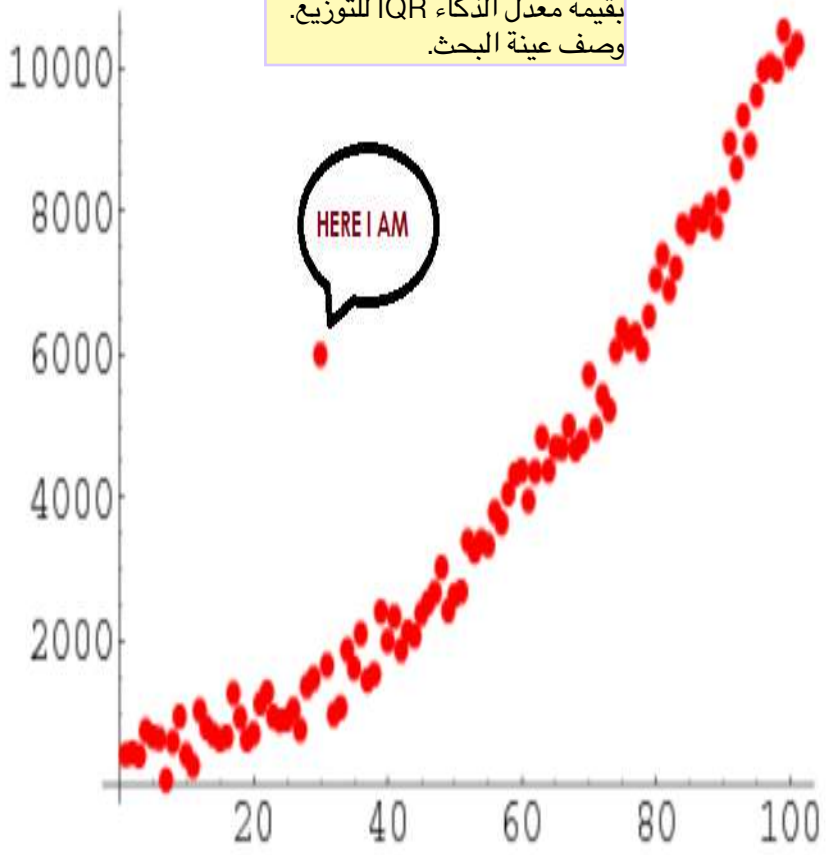
تقييم القيم المتطرفة. غالبًا ما يتم تحديد القيم المتطرفة فيما يتعلق بقيمة معدل الذكاء IQR للتوزيع.
وصف عينة البحث.

Understanding the data.

Evaluating outliers. Outliers are often identified in relation to the value of a distribution's IQR.

Describe the research sample.



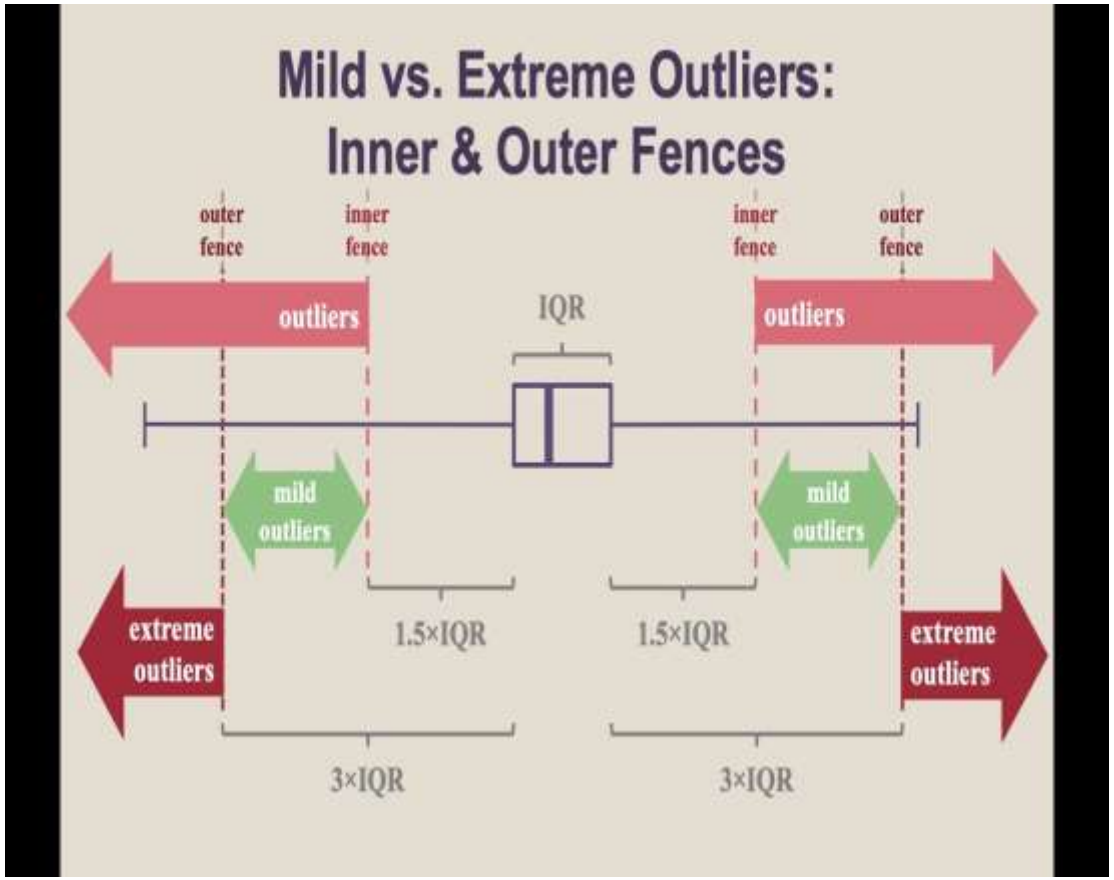


القيمة المتطرفة المعتدلة هي قيمة بيانات تقع بين 1.5 و3 أضعاف معدل الذكاء IQR أقل من Q1 أو أعلى من Q3.

القيمة المتطرفة هي قيمة بيانات تزيد عن ثلاثة أضعاف IQR أقل من Q1 أو أعلى من Q3.

A mild outlier is a data value that lies between 1.5 and 3 times the IQR below Q1 or above Q3.

Extreme outlier is a data value that is more that three times the IQR below Q1 or above Q3.




THE END OF UNIT 2