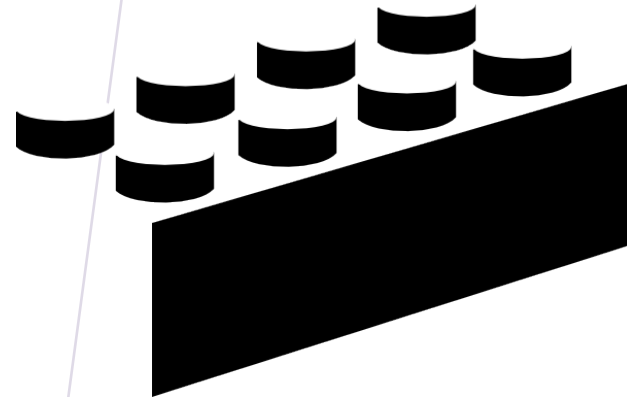


# ما هي البيانات

## What is Data

Based on [Udacity](#) Lesson

Mohammed Lotfy



## تُعَرَّف كلمة "بيانات" Data على أنها مجموعة

مميزة من المعلومات.

فمثلا في الجدول التالي بيانات  
مجموعة من الطلاب ونتائج  
اختبارهم في مادة الإحصاء:

الاسم الأول	الاسم الثاني	الدرجة
أحمد	رضا	٨٨
سمر	عادل	٩٣
سيف	كمال	٧٥

كل خلية في الجدول السابق تمثل  
بيان، ومن هذه البيانات نستطيع  
استنتاج معلومات، مثل: سمر  
حصلت على درجات أعلى من أحمد  
وسيف.



البيانات هي الأجزاء  
المكوّنة للمعلومات،  
وكل بيان يمثل جزء  
مستقل ومتميز عن  
غيره من البيانات، مثل  
بيت من المكعبات،  
فكل قطعة مكعبات  
هي بيان مستقل،  
ولكنها معا تكوّن شيئاً  
أكبر يسمى معلومات.

قد تفكرون في البيانات على أنها أرقام على جداول البيانات، ولكن البيانات قد تأتي بصيغ مختلفة، من نص إلى فيديو إلى جداول بيانات وقواعد بيانات إلى صور إلى صوت.



تتخذ البيانات أشكال متعددة،  
فقد تكون نصاً، أو فيديو أو  
صورة أو ملفاً صوتياً.



كما يمكن تخزين البيانات  
بطرق متعددة، فيمكن  
تخزينها في جداول  
البيانات على برامج مثل  
إكسيل

A black and white icon of a database cylinder, consisting of three horizontal bands and three small circles on the right side.

تعدّدت الاستفادة من البيانات  
إلى تقريبا **جميع المجالات**.  
وأصبح الاعتماد عليها وسيلة  
منتشرة عالميا في أوجه الحياة  
المختلفة.

يعتبر استخدام البيانات طريقة  
عالمية جديد.

تُستخدم البيانات لفهم كل واجهة في حياتنا وتطويرها، من  
الاكتشاف المبكر للأمراض، إلى الشبكات الاجتماعية التي  
تسمح لنا بالتواصل مع بعضنا البعض حول العالم.





وبغض النظر عن مجال دراستك،

سواء كان مجال التأمين والبنوك،

أو المجال الطبي، أو التعليم، أو

الزراعة أو الصناعة، يمكنك استخدام

البيانات لاتخاذ قرارات أفضل

وتحقيق أهدافك.

وسنساعدك على البدء بطريقة صحيحة للاستفادة من البيانات الخاصة

بمجالك في هذه الدورة التدريبية.

# أنواع البيانات

## Data Types

Based on [Udacity](#) Lesson

Mohammed Lotfy



أثناء جلوسي في المقهى، أشاهد الكلاب وهي تمر

من أمامي، وأتساءل: كم عددها؟



أتساءل إن كان هذا العدد يزداد

خلال أيام العمل، أم في عطلات

نهاية الأسبوع، ربما يختلف العدد

أيام الاثنين عن أيام الثلاثاء.



انتبهت أيضا إلى **سلالات** هذه الكلاب، وأتساءل إن كانت تمر سلالة كولي مثلا في يوم الاثنين أكثر من يوم الأربعاء.

وأتساءل أيضا أي من السلالات  
أشاهدها أكثر من غيرها؟ وهل  
هذه السلالة هي الأكثر مشاهدة  
من أمام جميع المقاهي؟ وإذا  
سرت في شارع آخر هل ستتغير  
السلالة الأكثر مرورا من أمامي؟



سنشرح في هذا الدرس نوعين رئيسيين من البيانات: بيانات كمية  
**Quantitative Data**، مثل عدد الكلاب، وبيانات مصنّفة [بيانات  
وصفية أو نوعية] **Categorical Data**، مثل السلالة.

1

تتخذ البيانات الكمية شكل أرقام تتيح لنا إجراء عمليات رياضية عليها.

+

0

4

9

÷



3

نستنتج من ذلك أنه ليس كل بيان على شكل أرقام هو بيان كمي، فرقم الهاتف يأخذ شكل أرقام، ولكن لا معنى لإجراء عمليات حسابية عليه، فلا معنى لجمع رقمي هاتفين مثلا، أو إيجاد متوسط هواتف مدينة ما! وكذلك مع بيانات مثل أرقام الحافلات وأرقام لاعبي كرة القدم وأرقام جلوس الطلاب وغيرها.



المحك الأساسي لاعتبار أن البيان كمي هو إمكانية إجراء العمليات الحسابية عليه.

لاحظنا ذلك في المثال السابق مع

عدد الكلاب، فإذا رأيت 5 كلاب

في يوم الاثنين و6 كلاب في يوم

الثلاثاء، فيمكنني القول أنني رأيت

11 كلبا  $\{5+6\}$  حتى الآن خلال

هذا الأسبوع.

في المقابل فإن البيانات

المصنفة (النوعية أو

الوصفية) تُستخدم في

تصنيف أو تمييز مجموعة

معينة من العناصر.



البيانات الوصفية هي بيانات تصف مجموعة محددة تشترك فيما بينها في صفة أو صفات مشتركة. مثل تصنيف أنثى وذكر، فبيان أنثى أو ذكر هو بيان وصفي، ومثل تقديرات الطلاب: امتياز، جيد جدا، جيد، وهكذا.

وقدر رأينا ذلك مع سلاطات الكلاب.

بشكل عام فإن أي بيان لا معنى لإجراء عمليات رياضية عليه هو بيان وصفي أو نوعي.

## Categorical Data



بيانات وصفية

السلالات

بودل  
وولف  
هاسكي

## Quantitative Data



بيانات كمية

عدد الكلاب

0  
1  
2

# البيانات الاسمية والرتبية

## Nominal and Ordinal

Based on [Udacity](#) Lesson

Mohammed Lotfy

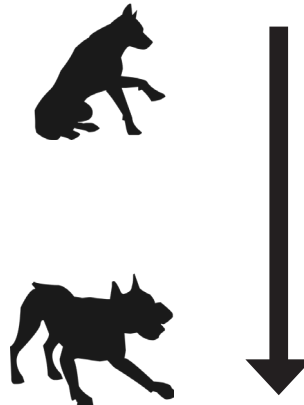
يمكننا تقسيم أنواع البيانات المصنّفة [البيانات الوصفية] أيضًا إلى ترتيب

مصنّف، [بيانات رتبية] Ordinal Data واسمية مطلقة [بيانات اسمية]

Nominal Data

أولاً، لنلقِ نظرة على بيانات الترتيب المصنّف [البيانات الرتبية]

هل تتذكر تلك الكلاب في المقهى؟ لنقل إنني أعطي تقييماً لمدى لطفها معي





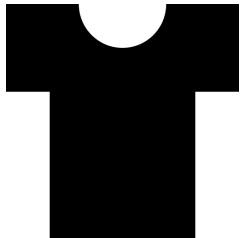
في بعض الأحيان أصافح  
الكلب ونصبح من أفضل  
الأصدقاء، وفي أحيان  
أخرى، يتبول على حذائي.  
وأنا أقيم هذه التفاعلات  
من الإيجابية جدًا إلى  
السلبية جدًا

وتُعرف هذه الفئات المصنفة كبيانات ترتيب مصنّف [بيانات رتبية]





## البيانات الرتبية يمكن ترتيبها من الأكبر فالأقل.



مقاسات الملابس

صغير  
متوسط  
كبير



تقييمك لمنتج أو خدمة



تقديرات الطلاب

ممتاز  
جيد جدا  
جيد  
مقبول

## ما يميز هذا النوع من البيانات شيئين:

الفرق بين كل رتبة والأخرى غير  
متساو، فالفرق بين الممتاز  
والجيد جدا لا يساوي بالضرورة  
الفرق بين الجيد جدا والجيد.



أنها بيانات تصنيفية، أي لا  
معنى لإجراء عمليات حسابية  
عليها كما ذكرنا في الدرس  
السابق.



ففي تقييمك لمنتج بعدد نجوم معين، فبالرغم من أن التقييم هنا يأخذ  
شكل رقم، إلا أننا لا يمكننا إجراء عمليات رياضية عليه، والفرق بين  
كل تقييم والآخر غير متساوي ولا يمكن قياسه حتى، فهو تقييم تقديري  
من جانبك.

النوع الآخر من البيانات  
التصنيفية هي التي **لا**  
**تخضع لترتيب محدد**،  
وتسمى **بيانات اسمية**،  
مثل بيان النوع سواء كان  
ذكرا أو أنثى، فلا يوجد هنا  
ترتيب محدد يجعلنا نضع  
نوعا ما قبل الآخر.



لاحظوا أن هذا يختلف  
عن السلالات، التي  
تُعرف على أنها بيانات  
اسمية مطلقة **[بيانات**  
**اسمية]**، ولا يوجد بها  
ترتيب وفقاً للمراتب.

# Categorical Data

## البيانات الوصفية

### Nominal Data



بيانات اسمية

السلالات

بودل  
وولف  
هاسكي

### Ordinal Data



بيانات رتبية

التقييم

إيجابي جدا  
إيجابي  
محايد  
سلبي  
سلبي جدا

# البيانات المتصلة والمنفصلة

## Continuous and Discrete Data

Based on [Udacity](#) Lesson

Mohammed Lotfy

يمكننا أيضًا تقسيم البيانات الكمية إلى أنواع

سأفترض أن معظم تفاعلاتي

الإيجابية تحدث مع الكلاب الأكبر

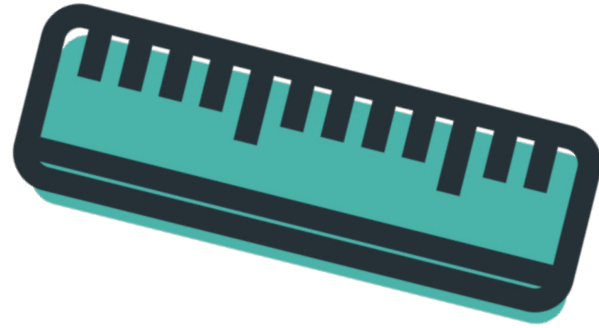
سنًا، لأنها تدربت وقتًا كثيرًا

يعتبر عمر الكلب من البيانات

الكمية المتصلة

**البيانات الكمية المتصلة Continuous Data** هي التي يمكن  
تجزئتها إلى أي قيم أصغر، مثل الطول بالكيلومتر يمكن تقسيمه  
إلى أمتار، والمتر إلى سنتيمترات والسنتيمتر إلى مليمتراً، وهكذا.

بسبب هذه التجزئة فإن كل قيمة والأخرى تكون متصلة بما قبلها أو بعدها.



**البيانات المنفصلة Discrete**  
**Data هي التي لا أستطيع تجزئتها**  
**إلى قيم أقل، وتُستخدم في العد.**

فمثلا عدد الأطفال في أسرة ما يجب أن يكون عدد صحيح، ٣ أو ٤ أطفال، ولا توجد أسرة لديها ٣ أطفال ونصف. فكل وحدة منفصلة عن الأخرى.

في حين أن عدد الكلاب  
التي أتفاعل معها هي من  
البيانات الكمية المنفصلة



في حياتنا اليومية، نستطيع جعل كل البيانات منفصلة.

فمن الأسهل علينا التعامل مع قيم صحيحة بدلا  
من التعامل مع القيم الدقيقة لكل بيان، فمثلا  
تقول أن عمر ك ٢٨ عاما فقط، ولا تذكر الشهور  
والأيام، فضلا عن الساعات والدقائق والثواني!

لذلك يكون من الصعب ملاحظة الفرق بين البيانات المنفصلة والمتصلة



## 7.3

يمكن أن تأخذ البيانات

المتصلة شكل أي قيمة

رقمية بما في ذلك

القيم العشرية، وحتى

الأعداد السالبة أحياناً.

يعتبر عمر الكلب في هذا الموقف مثالاً

على البيانات المتصلة، حيث يمكننا

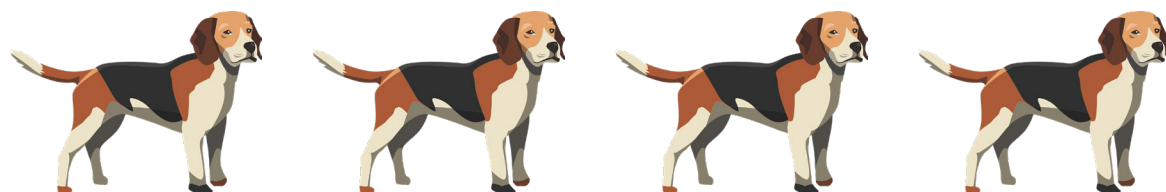
تقسيم هذا المتغير [عمر الكلب] إلى

أجزاء أصغر وأصغر، وسيبقى مع ذلك

أجزاء أقل لم نأخذها في الحسبان.



على سبيل المثال، يمكننا أن نتحدث عن العمر من حيث السنوات أو الشهور أو الأيام أو الساعات أو الدقائق أو الثواني، ولا تزال هناك وحدات أصغر [مثل الميلي ثانية وما هو أقل منها].



غير أن البيانات المنفصلة، مثل عدد الكلاب، تأخذ فقط شكل القيم القابلة للعدّ.

# Quantitative Data

## البيانات الكمية

### Discrete Data

 بيانات منفصلة

عدد الكلاب

### Continuous Data

بيانات متصلة



أعمار الكلاب

# ملخص أنواع البيانات

## Data Types Summary

Mohammed Lotfy

[facebook.com/mohammud.lotfy](https://facebook.com/mohammud.lotfy)

# أنواع البيانات

لا معنى للإجراء  
عمليات حسابية

Categorical

تصنيفية



يمكن إجراء عمليات  
حسابية عليها

Quantitative

كمية



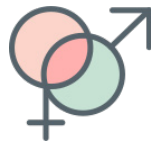
Ordinal

رتبية



Nominal

اسمية



Discrete

منفصلة



Continuous

متصلة



يمكن ترتيبها، من  
الأصغر للأكبر أو  
العكس

لا يمكن ترتيبها

لا يمكن تجزئتها  
إلى وحدات أقل،  
وتأخذ أعداد  
 صحيحة فقط

يمكن تجزئتها  
إلى وحدات أقل،  
وتقبل علامات  
عشرية



معرفة وتحديد أنواع البيانات هو أمر

هام، حيث يمكننا من فهم أنواع

التحليلات التي يمكن إجرائها على هذه

البيانات وكذلك أنواع الرسومات البيانية

الممثلة لها.

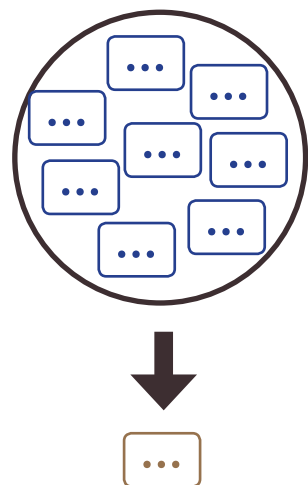


# إحصائيات التلخيص

## Summary Statistics

Mohammed Lotfy

[facebook.com/mohammud.lotfy](https://facebook.com/mohammud.lotfy)



تلخيص كم كبير من البيانات في عدد  
محدود من القيم يعبر عن مجموعة  
البيانات التي نقوم بتحليلها.

إحصائيات التلخيص  
في الدروس التالية

سندرس كيفية

استخدام الإحصاء لوصف

البيانات الكمية.

ستفهمون كيفية

جمع البيانات، والإجابة

عن الأسئلة باستخدام

هذه البيانات

الأسئلة الخاصة بالمشاكل والتحديات  
التي يعالجها تحليل البيانات.





كمثال للتحليل

الذي نجريه هنا في

"يوادسيتي"، فإننا نبحث

عن مقدار الوقت الذي

يستغرقه الطلاب لإكمال

برنامج ال **nanodegree**

شهادة تشمل مجموعة من الدورات المترابطة  
الخاصة بمجال محدد

في هذا الدرس، أتمنى أن

تصبحوا منتبهين لمهارات

تحليل البيانات الذي يتم

في الكواليس، وما تعنيه

الأرقام المستنتجة من عملية

التحليل.

نحاول توقع عدد الشهور أو الساعات التي سيستغرقها الطلاب لإكمال البرنامج.

باستخدام تحليل البيانات بطريقة صحيحة نستطيع  
توقع النتائج واتخاذ القرارات الصائبة على أساس هذا  
التوقع



متوسط الوقت هو وقت الطالب  
العادي، الذي لا يستغرق وقتاً  
طويلاً ولا قصيراً، بل في المنتصف  
بين هذا وذاك.

من الطرق التي يمكن أن نبدأ بها هي أن نسجل  
متوسط الوقت المستغرق لإكمال البرنامج.

إلا أن هذا لا يطلعنا

على كل شيء، أنا

على يقين من أن هناك

اختلافات في الوقت

المستغرق لإكمال

البرنامج حسب ما لدى

الطلاب من معرفة قبل

الالتحاق بالبرنامج.

1 3 4 6 9 11 12

حساب متوسط الوقت المستغرق لإتمام البرنامج لا يطلعنا  
على الصورة كاملة، فربما اختلافات وفروقه بين كل طالب  
والآخر قد تزيد أو تنقص، ولكنها موجودة.



قد تكون أقصر مدة

زمنية مطلوبة لإكمال

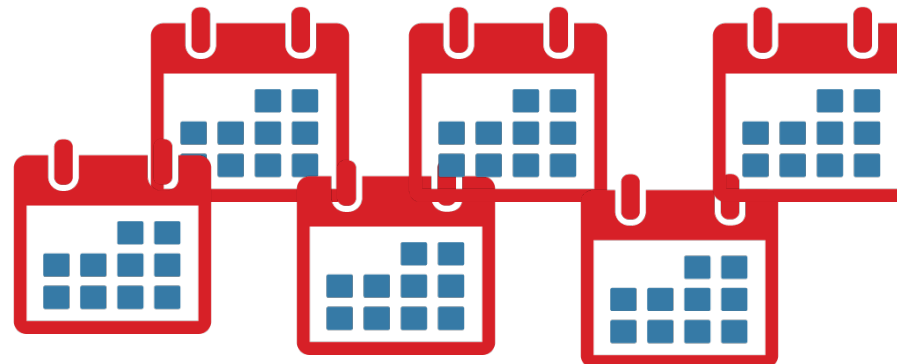
برنامج nanodegree

هي بضعة أسابيع.

كيف يكمل هؤلاء الطلاب الدورة التدريبية بهذه السرعة الفائقة؟

وقد تكون أطول

مدة هي عامين.



ما نسبة الطلاب

الذين يستغرقون

وقتًا أطول من

ثمانية أشهر؟



ما نسبة الطلاب

الذين يكملونه

بسرعة لا تتجاوز

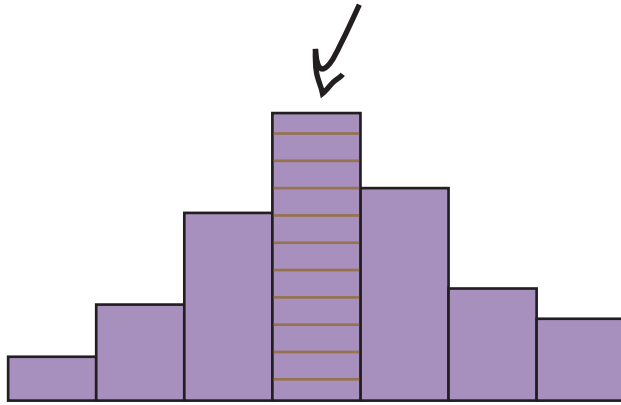
شهرين؟

هذه أسئلة تتعلق بمدى اختلاف وقت إنهاء  
البرنامج بين كل طالب والآخر. والإجابة عليها  
توضح لنا مدى تشتت أو تباعد البيانات عن  
بعضها.

## تمنحك مقاييس المركز Center

فكرة عن الطالب المتوسط

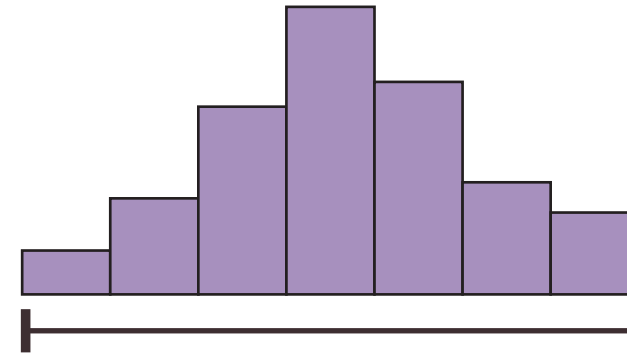
الطالب الذي يقع مستواه في المنتصف من  
الطلاب الذين ينتهوا من البرنامج بسرعة  
كبيرة والطلاب الذين ينتهوا من البرنامج في  
وقت كبير



وتمنحك مقاييس التشتت

**Spread** فكرة عن مدى اختلاف

الطلاب عن بعضهم البعض.

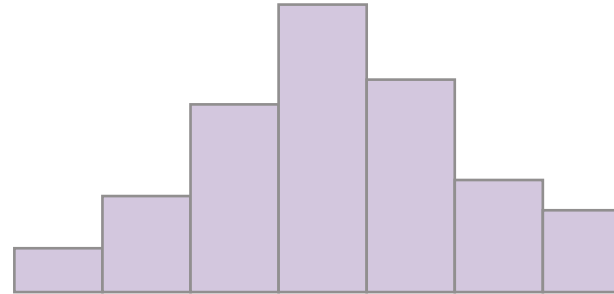


مدى تباعد أو تقارب وقت انتهاء البرنامج  
بين جميع الطلاب

قد تمنحنا الرسوم البيانية صورة كاملة عن مقدار الوقت الذي يستغرقه أي طالب

لإكمال البرنامج.

الصورة بألف كلمة، والرسوم  
البيانية تجعلك ترى مباشرة معنى  
البيانات التي لديك.



سنتعلم في الدروس

التالية كيفية استخدام هذه

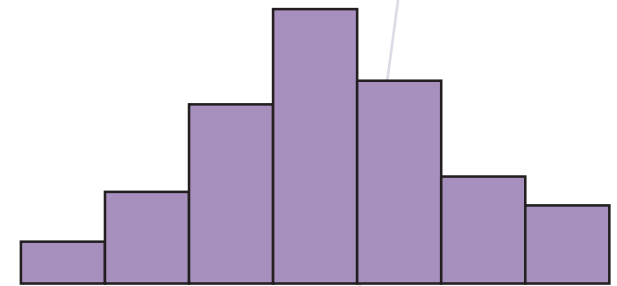
المقاييس لتكون مفيدة

ومفهومة للآخرين.

# حساب الوسط

## Calculating the Mean

Based on Udacity Lesson



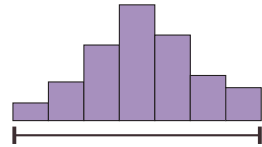
By: Mohammed Lotfy

عند تحليل كل من البيانات الكمية، سواء المنفصلة أو المتصلة، نناقش أربعة جوانب

رئيسية:

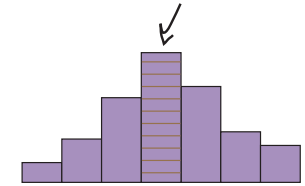
يساعدنا على تحديد مدى  
اختلاف القيم عن بعضها  
البعض، فلو كان التشتت  
كبيرا فهذا يعنى أن  
الفروق بين القيم كبير،  
والعكس إذا كان صغيرا.

الانتشار Spread



يبحث عن القيمة التي تتركز  
حولها جميع القيم، أى القيمة  
التي تقع في منتصف القيم،  
وتفيدنا في تلخيص القيم  
جميعها في قيمة واحدة، فتصبح  
هذه القيمة هي المتوسط  
Average.

المركز Center



هى قيم كبيرة جدا أو صغيرة  
جدا مقارنة ببقية القيم الأخرى،  
لذا يجب الانتباه لها لأنها تؤثر  
على عملية التحليل كما نرى فيما  
بعد.

القيم الشاذة

Outliers

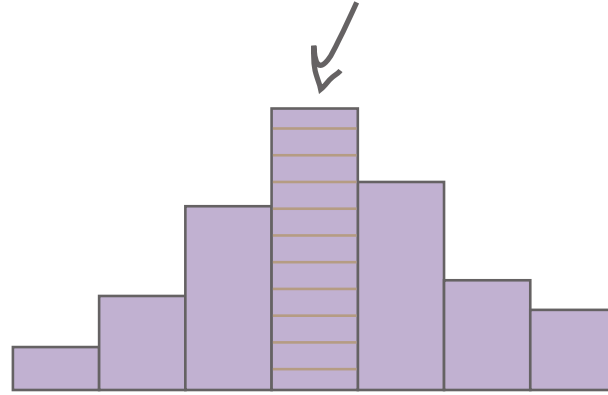


هو رسم يمثل مركز  
وانتشار مجموعة من  
البيانات بصورة مرئية  
وسهلة الفهم.

الشكل Shape







في هذا الدرس، سنركز

على مقاييس المركز.

سنشرح في هذا الدرس  
كيفية حساب المتوسط  
أو الوسيط Mean



المتوسط Mean

الوسيط Median

الوضع [المنوال] Mode

هناك ثلاثة مقاييس للمركز

مقبولة بشكل كبير:

لتوضيح كيفية حساب كل

من هذه القياسات فكروا

في هذا الجدول الخاص

بعدد الكلاب التي أراها أمام

المقهى خلال أسبوع ما:

الاثنين	الثلاثاء	الأربعاء	الخميس	الجمعة	السبت	الأحد
5	3	8	3	15	45	9

بالنظر إلى الجدول، نلاحظ أنني رأيت خمسة كلاب

يوم الاثنين، ثلاثة كلاب يوم الثلاثاء، ثمانية كلاب يوم

الأربعاء، وهكذا...



قد يسألك صديق: كم عدد الكلاب التي تتوقع رؤيتها في يوم معين؟

نلاحظ هنا أن السؤال عن التوقع، حيث يساعدنا التحليل على التنبؤ بما  
سيحدث والإجابة على أسئلة تتعلق بالمستقبل لاتخاذ قرارات صائبة.

قد تختار الإجابة عن هذا السؤال بطرق مختلفة، مثل: إن ذلك يعتمد على اليوم أو

الأسبوع.

يعتبر الوسيط طريقة سهلة وسريعة لحساب متوسط مجموعة من القيم، وتستطيع به التعبير عن كافة القيم، أو كافة مشاهدات الكلاب خلال الأسبوع، باستخدام قيمة واحدة فقط هي قيمة الوسيط

لكن عادة ما يكون هذا التوقع مقترن

بالوسيط **Mean** أو بمتوسط مجموعة

البيانات التي لدينا

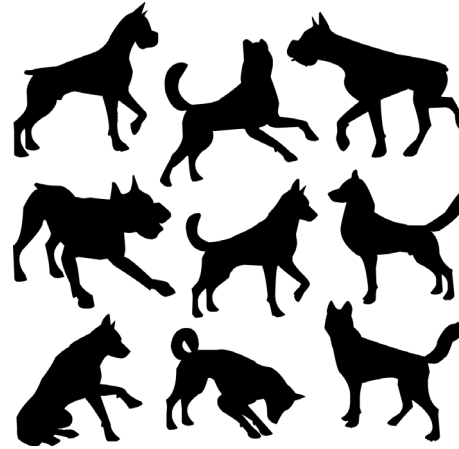
الأحد	السبت	الجمعة	الخميس	الأربعاء	الثلاثاء	الاثنين
9	45	15	3	8	3	5
↑	↑	↑	↑	↑	↑	↑
↓						
?						

فبدلاً من عرض كل تلك القيم، يمكننا تلخيصها في قيمة واحدة تعبر عن متوسط هذه القيم، وبالتالي تكون هذه القيمة هي متوسط عدد الكلاب المتوقع رؤيتها في أي يوم.

يتم حساب الوسط عن طريق قسمة مجموع كل القيم الموجودة في مجموعة البيانات

على عدد البيانات.

إجمالي عدد مشاهدات الكلاب  
خلال الأسبوع



الأحد	السبت	الجمعة	الخميس	الأربعاء	الثلاثاء	الاثنين
9	45	15	3	8	3	5

الوسط =

الأحد	السبت	الجمعة	الخميس	الأربعاء	الثلاثاء	الاثنين
9	45	15	3	8	3	5



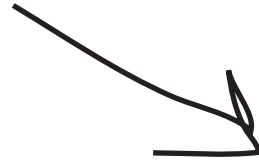
عدد أيام الأسبوع

$$\text{كلب } 12.57 = \frac{9 + 45 + 15 + 3 + 8 + 3 + 5}{7}$$

وهذا هو مجموع عدد الكلاب التي شوهدت في كل يوم،

مقسوم على عدد الأيام في الأسبوع.

12.57 كلب

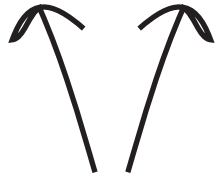


نجد أن الوسط هنا يعبر عن  
عدد الكلاب بقيم عشرية

لا يعتبر الوسط دائمًا هو أفضل قياس للمركز.

كما نرى فإن ناتج الوسط به رقم  
بعد العلامة العشرية، ولهذا غير  
منطقي، لأنه لا يوجد نصف  
كلب، فالقيم المستخدمة في العد  
يجب أن تكون قيم كمية منفصلة  
كما شرحنا سابقًا، وهذه إحدى  
مشكلات الوسط، بالإضافة إلى  
مشكلة أخرى في هذا المثال...

الأحد	السبت	الجمعة	الخميس	الأربعاء	الثلاثاء	الاثنين
9	45	15	3	8	3	5



ففي يومين فقط من

سبعة أيام، تم تسجيل

عدد كلاب أكبر من

الوسط المحدد لهم

في حين أن الوسط يجب أن يتوسط جميع البيانات، وهنا هو يقع بين قيمتين فقط، وبقية القيم أقل منه.

بمعنى آخر فإن الوسط يجب أن يكون تقريبا نصف عدد البيانات أكبر منه والنصف الآخر أقل منه، وهذا ليس الحال هنا.

لعلك لاحظت أن عدد الكلاب المشاهدة يوم السبت هو ٤٥ كلبا، وهي قيمة أكبر بكثير من بقية القيم. هل تتذكر القيم السابقة؟

لعل هذا هو سبب أن قيمة الوسط أكبر من معظم القيم. جرب إعادة حساب الوسط واستبدل عدد الكلاب في يوم السبت بقيمة قريبة من بقية القيم، ولتكن ٦، ولاحظ كيف تغيرت قيمة الوسط.



السبت

45



# حساب الوسيط

## Median Calculation

Based on Udacity Lesson

Mohammed Lotfy



المتوسط Mean



الوسيط Median

المنوال Mode

قد يكون الوسيط Median مقياسًا

أكثر ملاءمةً في هذه الحالة.

راجع مثال الدرس  
السابق

فالوسيط هو قيمة

تقسّم مجموعة البيانات

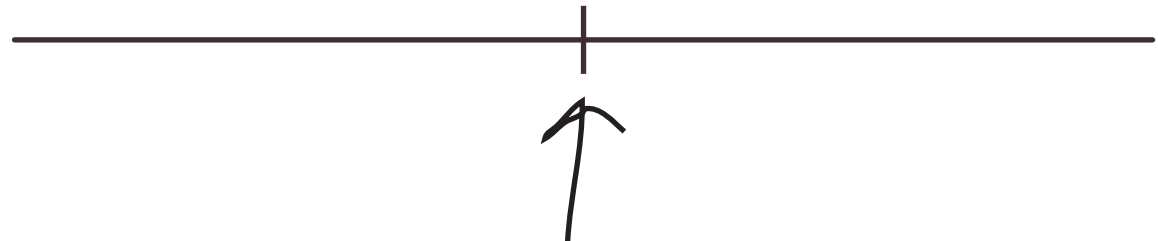
على أن يكون ٥٠٪ من

القيم أكبر منه بينما الـ

٥٠٪ المتبقية أصغر منه.

50%

50%



الوسيط

الأحد	السبت	الجمعة	الخميس	الأربعاء	الثلاثاء	الاثنين
9	45	15	3	8	3	5

3   3   5   **8**   9   15   45

بالنسبة إلى هذه المجموعة من البيانات فإن الوسيط يساوي ٨.

سنتعرف بعد قليل على كيفية حساب قيمة الوسيط

وهذه إجابة أفضل كثيرًا من ١٢

كلبًا ونصف التي حصلنا عليها

من الوسط Mean.

ولا يقتصر الأمر على وجود القيمة ٨ في منتصف البيانات

مما يجعله يمثل مركز البيانات  
بطريقة أفضل من الوسيط في  
هذا المثال



ولكنها أيضا لا تقسم أيًا من كلابنا نصفين.



وكملاحظة بخصوص حساب الوسيط، فإنه يعتمد على عدد القيم أو البيانات،

هل عددها زوجي أم فردي.

لننظر في هذين المثالين لتوضيح هذه النقطة:

1 2 3 3 5 8 10

5 8 3 2 1 3 10 105

أول ما علينا فعله هو ترتيب هذه القيم من الأصغر إلى الأكبر

1 2 3 3 5 8 10

في المثال الأول القيم مرتبة تصاعديا لذا فلن نحتاج إلى ترتيبها بأنفسنا

في المثال الأول نتعامل مع 7 قيم

وبالترميز يمكننا

أن نشير للرقم 7

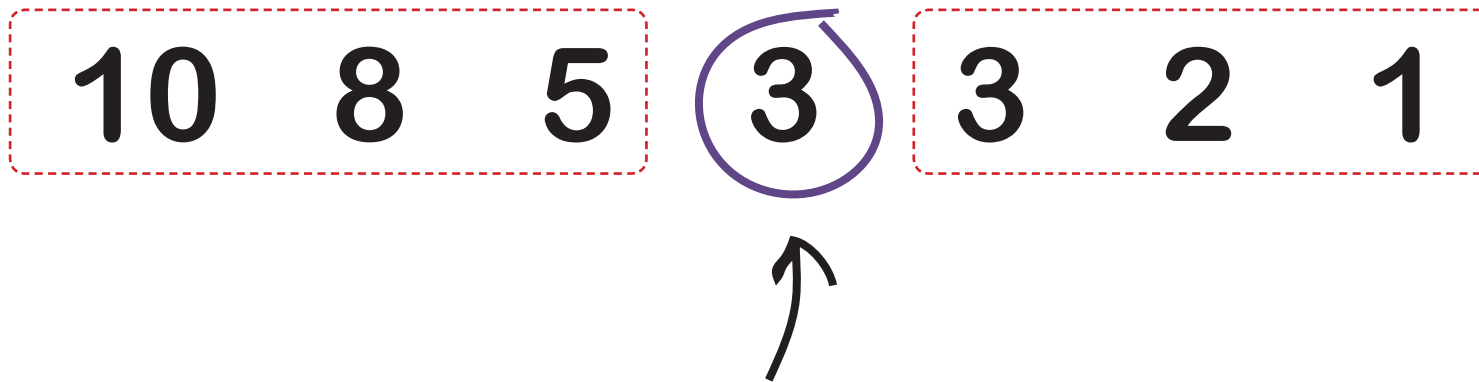
بالرمز  $n$ .

$$n = 7$$

تصور أن الرمز  $n$  هو وعاء نضع به قيمة ما، والقيمة هي 7 في هذا المثال.



ولأن هذا العدد **فردى** تكون قيمة الوسيط  
تمامًا هي القيمة التي في **المنتصف**.



وفي هذا المثال، الوسيط **Median** هو 3.

في المثال الثاني  $n$  تساوي عددًا زوجيًا

$$n = 8$$

105 10 3 1 2 3 8 5

مجددًا، يجب علينا أولاً ترتيب هذه القيم

105 10 8 5 3 3 2 1

نظرًا لعدم وجود قيمة واحدة في منتصف البيانات، فسوف

نحسب متوسط القيمتين **Mean** اللتان في منتصف القيم،

ليكون متوسطهما هو قيمة الوسيط.



$$4 = \frac{5 + 3}{2}$$

← مجموع القيمتين

← عدد هـم



105 10 8 5 4 3 3 2 1

الوسيط



ولاحظ، أن في هذه الحالة فإن قيمة الوسيط

**Median** ليست قيمة من قيم مجموعة البيانات

التي لدينا.



# المنوال

## Mode

Based on Udacity Lesson

Mohammed Lotfy

المتوسط Mean

الوسيط Median



المنوال Mode

يعطينا المقياس الثالث

للمركز القيمة الأكثر

شيوعاً في مجموعة

البيانات التي لدينا.

المنوال هو القيمة التي لها أكبر تكرار في مجموعة البيانات.

في مجموعة البيانات

هذه، المنوال هو

القيمة 3

الاثنين	الثلاثاء	الأربعاء	الخميس	الجمعة	السبت	الأحد
5	3	8	3	15	45	9

في حال تكرار قيمتين بنفس القيمة فإن مجموعة البيانات يكون لها منوالان، وتسمى في هذه الحالة **bimodal**، أي مزدوجة النوال  
في هذا المثال النوال هو: 3 و 5



في حال عدم تكرار أية قيمة، أو تكرار جميع القيم بنفس القيمة، فإن مجموعة البيانات لا يكون لها أي نوال



# الترميز

# Notation

Mohammed Lotfy  
@mohammud.lotfy

## ذكرنا سابقا الجوانب الأربعة الأساسية لتحليل البيانات الكمية:

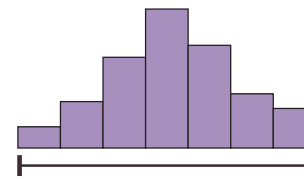
القيم الشاذة Outliers



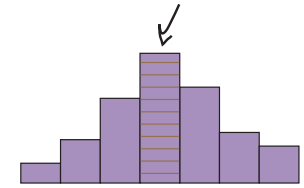
الشكل Shape



الانتشار Spread



المركز Center



المتوسط Mean

الوسيط Median

الوضع [المنوال] Mode

واطلعنا على مقاييس

المركز عن طريق شرح:

وقبل أن نتطرق إلى مقاييس

الانتشار من المهم فهم الترميز.



قد تعتقد أنك لست على دراية به، ولكنك تستخدم الترميز طوال الوقت!

الترميز هو استخدام الرموز والعلامات، مثل +، للدلالة على معنى ما، فالرمز + يعني إضافة رقمين إلى بعضهما البعض. وهذه الرموز مفهومة على مستوى العالم ولا تحتاج إلى ترجمة، فناطقى أية لغة يفهمون معنى +، حتى وإن اختلف نطقها من لغة إلى أخرى.



فكر في هذا المثال، 3 + 5،

زائد هي كلمة عربية

وهذا الرمز [+ ] من

الترميز، وهو لغة عالمية.

يعد الترميز لغة

رياضيات مشتركة

تستخدم للتواصل.

يستخدم الترميز في كتابة المعادلات وحل المسائل الرياضية  
(والإحصاء هي جزء تطبيقي للرياضيات)، ولأنه مفهوم  
على مستوى اللغات المختلفة، فهو يستخدم للتواصل بين  
الباحثين والعلماء على اختلاف لغاتهم.

سواء أكنتم تتحدثون العربية أو الأسبانية أو اليونانية أو غيرها من اللغات، يمكنكم  
العمل معًا باستخدام الترميز كلغة مشتركة لحل المشكلات.

سنعمل معًا على بعض

الأمثلة لضمان أنك أتقنت

تمامًا هذا المفهوم.

وكتعلم أي لغة جديدة قد يكون الترميز مخيفًا

في أول الأمر، ولكنه أداة ضرورية لتناقل

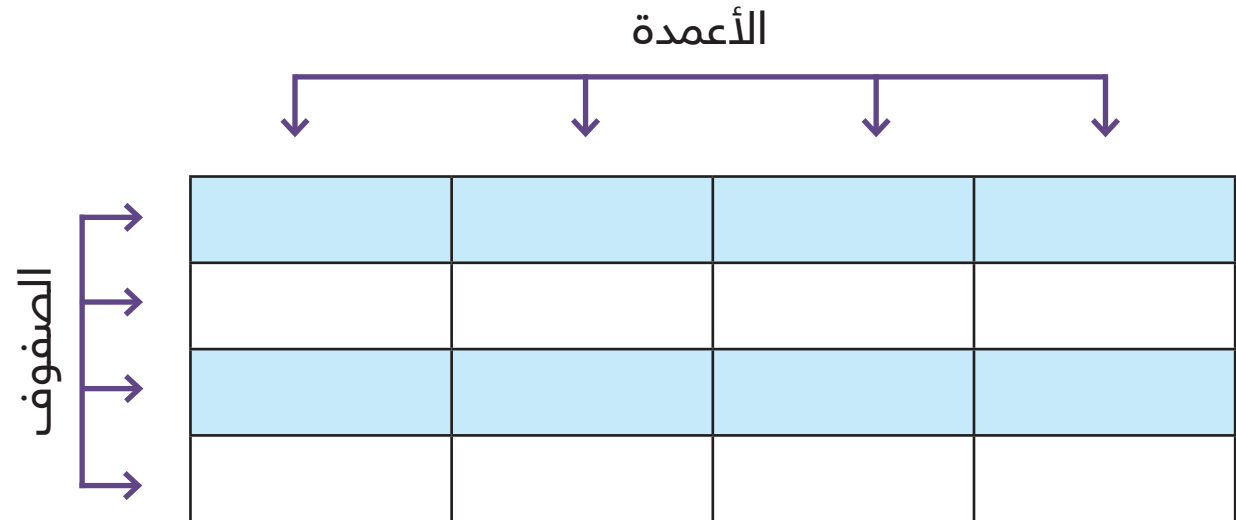
الأفكار المتعلقة بالبيانات.



# المتغيرات العشوائية

## Random Variables

لأن هذا هو أول مثال [لفهم الترميز **Notation**]، لنطبق فكرة التدوين [الترميز] الجديدة هذه على شيء استخدمناه من قبل، وهو جداول البيانات.



جداول البيانات هي طريقة  
شائعة للاحتفاظ بالبيانات  
في حياتنا العملية.

في جدول البيانات، يوجد لدينا صفوف وأعمدة.

ولفهم أفضل لكيفية استخدام جداول البيانات للاحتفاظ بالبيانات لنأخذ مثالاً:



افترضوا أنني أدير مدونة

صغيرة عن أفضل

مغامراتي مع الكلاب

وأسوأها في المقهى

يبيع المقهى أيضاً الحلي المتعلقة

بتلك المغامرات للبيع، بداية من

ألعاب الرمي والإحضار والمقاود

وصولاً إلى حقائب الكلاب.



قبل جمع البيانات عادة ما نبدأ بسؤال أو بعدة أسئلة:

كم عدد **الأشخاص** الذين يزورون موقعي؟

كم من **الوقت** يقضيه الزوار في موقعي؟

هل هناك **اختلافات** في عدد الزوار حسب يوم الأسبوع؟

كم عدد الزوار الذين **يشتررون** شيئاً ما من خلال المدونة؟

للإجابة عن هذه الأسئلة، افترضوا أننا نتعقب تاريخ الزيارة، ويوم الأسبوع الخاص بالزيارة، ومقدار الوقت المستغرق على الموقع، وما إذا كان الزائر سيشترى شيئاً أم لا.

الوقت المستغرق

على الموقع

هل تمت عملية شراء؟

يمكننا أن نمثل

كل معلومة على

شكل عمود

التاريخ	اليوم	الوقت	شراء

يقترن العمود في

مجموعة البيانات

لدينا بما يسمى

"متغير عشوائي".

التاريخ	اليوم	الوقت	شراء
---------	-------	-------	------

وشرح مفهوم المتغير العشوائي باللغة العادية أمر معقد:

"رمز نائب عن القيم

الممكنة لعملية ما".

ولكن بالتدوين [الترميز]، يكون أمرًا بسيطًا:

في التدوين [الترميز]، هو  $X$ .



يمكنك تصور المتغير العشوائي على أنه كوب يمكنك ملئه بما تشاء. الوعاء نفسه هو المتغير العشوائي، وما تضعه بداخله هو القيمة.

$X$



يمكن التعبير عن المتغير العشوائي

(الوعاء) برمز ما، وليكن  $X$ .

بالنسبة إلى موقعنا الإلكتروني، يعتبر كل من تاريخ  
الزيارة، ويوم الأسبوع الخاص بالزيارة، ومقدار  
الوقت المستغرق على الموقع، وما إذا كان الزائر  
سيشتري شيئاً أم لا، كلها متغيرات.

			
شراء	الوقت	اليوم	التاريخ



لنفترض أن لدينا زائرًا يوم الخميس الموافق ١٥ يونيو، يتصفح الزائر موقعنا لمدة خمس دقائق ولا يشتري شيئًا، ثم يقوم زائر آخر بزيارة الموقع في اليوم نفسه لمدة ١٠ دقائق ويشتري شيئًا.

لاحظ كيف تمت  
إضافة كل من  
هذين الشخصين  
إلى جدول البيانات.

التاريخ	اليوم	الوقت	شراء
١٥ يونيو	الثلاثاء	٥	لا
١٥ يونيو	الثلاثاء	١٠	نعم

تمت إضافة كل زيارة لكل شخص في  
صف مستقل

قد يكون لدينا عدد زوار أكبر بكثير، ويمكننا تحديث جدول البيانات لدينا وفقًا لذلك:

التاريخ	اليوم	الوقت	شراء
١٥ يونيو	الثلاثاء	٥	لا
١٥ يونيو	الثلاثاء	١٠	نعم
١٦ يونيو	الجمعة	٧	نعم
١٦ يونيو	الجمعة	٩	نعم
١٦ يونيو	الجمعة	١٢	لا



على سبيل المثال، للإجابة

عن سؤال الوقت الذي

يستغرقه الزوار على

موقعنا، نحتاج إلى أن نلقي

نظرة على هذا العمود.



التاريخ	اليوم	الوقت	شراء
١٥ يونيو	الثلاثاء	٥	لا
١٥ يونيو	الثلاثاء	١٠	نعم
١٦ يونيو	الجمعة	٧	نعم
١٦ يونيو	الجمعة	٩	نعم
١٦ يونيو	الجمعة	١٢	لا

عند استخدام جداول البيانات، نحلل

أحيانًا عمود كامل للإجابة عن الأسئلة

المثيرة للاهتمام.

عن طريق قيم عمود الوقت  
نستطيع مثال الإجابة عن  
سؤال: ما هي أطول مدة قضاها  
زائر على الموقع؟

وللإجابة عن السؤال الخاص

بكم عدد عمليات الشراء

التي تتم من خلال مدونتنا،

نحتاج إلى أن نلقي نظرة

على هذا العمود.

للإجابة عن السؤال الخاص بهل

هناك اختلافات في عدد الزيارات

حسب يوم الأسبوع، نحتاج إلى أن

نلقي نظرة على هذا العمود.



التاريخ	اليوم	الوقت	شراء
١٥ يونيو	الثلاثاء	٥	لا
١٥ يونيو	الثلاثاء	١٠	نعم
١٦ يونيو	الجمعة	٧	نعم
١٦ يونيو	الجمعة	٩	نعم
١٦ يونيو	الجمعة	١٢	لا

رياضيًا، نعبر عادة عن عمود ما أو متغير عشوائي ما باستخدام

حرف كبير capital letter.

عادة ما نستخدم الحرف الكبير X

X

Y Z

لكن يمكننا بكل

سهولة استخدام

Y أو Z أو أي

حرف آخر كبير.

يمكننا أن نفترض أن المتغير

العشوائي  $X$  يشير إلى مقدار

الوقت الذي يستغرقه الزائر

على موقعنا.



$X$  يشير إلى الكوب الذي من الممكن أن يحتوى على أية قيمة للوقت، ولا يشير إلى القيمة ذاتها.

$X$

التاريخ	اليوم	الوقت	شراء
١٥ يونيو	الثلاثاء	٥	لا
١٥ يونيو	الثلاثاء	١٠	نعم
١٦ يونيو	الجمعة	٧	نعم
١٦ يونيو	الجمعة	٩	نعم
١٦ يونيو	الجمعة	١٢	لا

بعبارة أخرى فإن  $X$  يعبر عن جميع القيم المحتملة للوقت.

الوقت
0
10
7
9
12

ومن ثمَّ، يرتبط **X** بهذا العمود بأكمله.

**Y**

ويمكننا أيضا أن نفترض أن لدينا متغيرًا عشوائيًا هو **Y**، يشير إلى ما إذا

كان الزائر سيشترى شيئًا من الموقع أم لا.

التاريخ	اليوم	الوقت	شراء
١٥ يونيو	الثلاثاء	0	لا
١٥ يونيو	الثلاثاء	10	نعم
١٦ يونيو	الجمعة	7	نعم
١٦ يونيو	الجمعة	9	نعم
١٦ يونيو	الجمعة	12	لا

ومن ثمَّ، يرتبط

**Y** بهذا العمود

بأكمله.

# المتغيرات العشوائية والمشاهدات

Random and Observed Variables

Mohammed Lotfy  
@mohammud.lotfy



فكرة مجردة بمعنى أن  $X$  أو  $Y$   
 يمثلان أية قيمة محتملة، ولا  
 يمثلان قيمة محددة، فهما يمثلان  
 احتمال قائم وليس تحدد معين  
 لقيمة ما.

الوقت
0
10
7
9
12

لاستكمال ما سبق نقول بأن الحرفان الكبيران  $X$  و  $Y$   
 يمثلان متغيران عشوائيان. وهذه فكرة مجردة.

يمكن أن يأخذ مقدار الوقت الذي يقضيه  
 الفرد على موقع الويب الخاص بنا العديد من  
 القيم المختلفة.

إذًا، لا يعتبر الحرف الكبير  $X$  عددًا، بل هو مجموعة كاملة من القيم الممكنة. ويمكننا التفكير في الحرف  $X$  كنائب عن كل تلك القيم الممكنة.



كما قلنا قبل قليل، فإن  $X$  هو كوب يمكنك وضع أية قيمة به، والقيم التي يمكنك وضع إحداها به هي القيم الممكنة له، فالوعاء  $X$  الذي يشير إلى الوقت الذي يقضيه الزائر على الموقع له قيم محتملة أو ممكنة هي أي رقم يمثل عدد الدقائق.

عندما ننظر إلى **ناتج محددة** للمتغير العشوائي

نشير إليه **بحرف صغير small letter**.

الوقت		
0	○.....	X
10	○.....	X
7	○.....	X
9	○.....	X
12	○.....	X

نعبر عن كل قيمة من القيم الممكنة

للمتغير العشوائي برمز على شكل حرف

صغير small letter.

غالبًا ما يكون الحرف الصغير

متبوعًا **بحرف منخفض**،

يساعدنا على إلحاق تدوين

[ترميز] بكل قيمة محددة في

مجموعة البيانات.

X<sub>1</sub>

بالنسبة إلى مجموعة البيانات هذه، نقول إن الوقت الذي يقضيه الفرد على موقعنا يعبر عن مقدار ما، يمكننا تدوينه بحرف  $X$  كبير.

يقضي الزائر الأول  
خمس دقائق على  
الموقع والذي يمكننا  
تدوينه بالحرف  $x_1$ .

 $x_1$  $X$ 

التاريخ	اليوم	الوقت	شراء
١٥ يونيو	الثلاثاء	٥	لا
١٥ يونيو	الثلاثاء	١٠	نعم
١٦ يونيو	الجمعة	٧	نعم
١٦ يونيو	الجمعة	٩	نعم
١٦ يونيو	الجمعة	١٢	لا

لاحظ أن هذا حرف  $x$  صغير.

X

ويقضي الزائر الثاني

١٠ دقائق على

الموقع والذي يمكننا

تدوينه بالحرف  $X_2$ . $X_2$ 

التاريخ	اليوم	الوقت	شراء
١٥ يونيو	الثلاثاء	٥	لا
١٥ يونيو	الثلاثاء	١٠	نعم
١٦ يونيو	الجمعة	٧	نعم
١٦ يونيو	الجمعة	٩	نعم
١٦ يونيو	الجمعة	١٢	لا

مرة أخرى، لأنها قيمة ملاحظة فإنها تُكتب بحرف  $X$  صغير.

قيمة تم رصدها وتسجيلها  
لحدث محدد، مثل مدة زيارة  
معينة لزائر معين في وقت معين

قد تستمر التسمية حتى نصل إلى

الزائر الأخير في مجموعة البيانات،

الذي نسميه الزائر رقم  $n$ .

نرمز إلى عدد القيم في أي عمود بالحرف  $n$

التاريخ	اليوم	الوقت	شراء
١٥ يونيو	الثلاثاء	٥	لا
١٥ يونيو	الثلاثاء	١٠	نعم
١٦ يونيو	الجمعة	٧	نعم
١٦ يونيو	الجمعة	٩	نعم
١٦ يونيو	الجمعة	١٢	لا

$x_1$

$x_2$

...

$x_n$

ويمكننا تسمية

هذا الزائر  $x_n$ .

وهذه أيضا قيمة بحرف صغير.

يشير المتغير  $x_n$  إلى آخر قيمة في عمود ما، بنقض النظر عن عدد القيم، فهو ليس له قيمة ثابتة، قد يكون القيمة الخامسة في عمود مكون من خمسة قيم، أو القيمة الألف في عمود مكون من ألف قيمة.

التدوين [الترميز] هو أداة ضرورية لنا لتناقل الأفكار الرياضية.

عندما نلاحظ قيمة معينة للمتغير

العشوائي نستخدم حرفاً صغيراً

مع حرف منخفض يشير إلى

[ترتيب] القيمة المعينة للمتغير

العشوائي الذي نقصده.

$x_2$

لقد ناقشنا الآن فكرة أن الحروف

الكبيرة تستخدم كتدوين [كترميز]

لمتغيرات عشوائية.

$X$

قد تكون التدوينات [الترميز] مربكة، فقبل أن نتعمق أكثر، تحقق من فهمك لها.

وتوجد اختبارات في الوحدة التالية لضمان أنك أتقنت المفاهيم التي شرحناها.

لا بد من وجود طريقة أفضل

There Must Be a Better Way

Mohammed Lotfy  
@mohammud.lotfy



ليس الغرض من هذا الفيديو

إعادة تعلم كيفية حساب

الوسط، بل دراسة التدوين

[الترميز] باستخدام مقياس

تعرفه بالفعل [الوسط].

فيما سيأتي سندمج ما نعرفه عن كيفية

حساب **الوسط** مع التدوين [الترميز].

الوقت
١٥
١٠
٥
١٠
٣

لننظر إلى مقدار **الفترة الزمنية** التي يستغرقها شخص ما

على موقعنا الإلكتروني **بالدقائق**.

تخيلوا أننا نجمع مقدار  
الفترة الزمنية التي  
يستغرقها خمسة  
أشخاص على موقعنا  
الإلكتروني.

الوقت
١٠
١٠
٥
١٠
٣

الوقت	
١٠	○ ..... $x_1$
١٠	○ ..... $x_2$
٥	○ ..... $x_3$
١٠	○ ..... $x_4$
٣	○ ..... $x_5$

من الفيديو الأخير  
تعلمنا أنه يمكننا  
تسمية القيم  
بهذه الطريقة:

$x_1$  للقيمة الأولى،  $x_2$  للقيمة الثانية،  $x_3$  للقيمة الثالثة، وهكذا.

الوقت		
١٥	○	$x_1$
١٠	○	$x_2$
٥	○	$x_3$
١٠	○	$x_4$
٣	○	$x_5$

تخيل أننا نرغب في جمع أول عددين معًا.

يمكننا كتابة هذا بالتدوين [الترميز]  $x_2 + x_1$

$$\begin{array}{ccc} 15 & + & 10 \\ \downarrow & & \downarrow \\ x_1 & + & x_2 \end{array}$$

حيث تساوي في هذا المثال  $10 + 15$

إذا أردنا جمع أكثر من

قيمتين، فستكون متابعة

العملية نفسها أمرًا مضجرًا.

$$x_1 + x_2 + x_3 + \dots + x_{100}$$

تخيل إذا كان لدينا ١٠٠ قيمة، سيكون لدينا  $x_1$  زائد  $x_2$  زائد  $x_3$  وهكذا، وسيكون علينا

متابعة العملية هذه حتى  $x_{100}$ .

لا بد أن يكون قد توصل شخص ما إلى طريقة

أفضل لتدوين هذا. وبالفعل هناك طريقة أفضل.

# التجميع

# Summation

Mohammed Lotfy  
@mohammud.lotfy

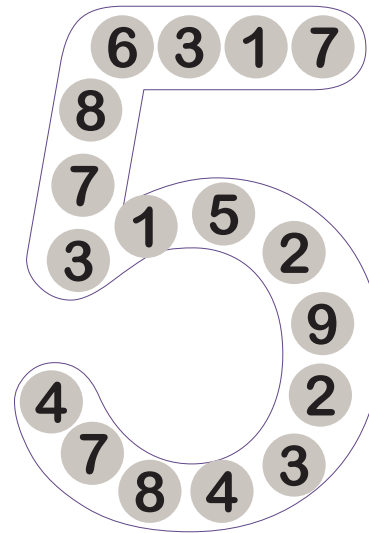
هناك طرق شائعة

لتدوين [لترميز]

معظم التجميعات

.Aggregations

التجميع هو طريقة لاختصار أعداد كثيرة إلى أعداد أقل،  
وعادة ما يكون عددًا واحدًا فقط.



المتوسط Mean

الوسيط Median

الوضع [المنوال] Mode

تتضمن التجميعات المشاعة مقاييس المركز، التي شرحناها

سابقًا.

تأخذ كل من هذه التجميعات أعدادا كثيرة لتكون النتيجة هي قيمة واحدة تقدم معلومات عن مجموعة البيانات.

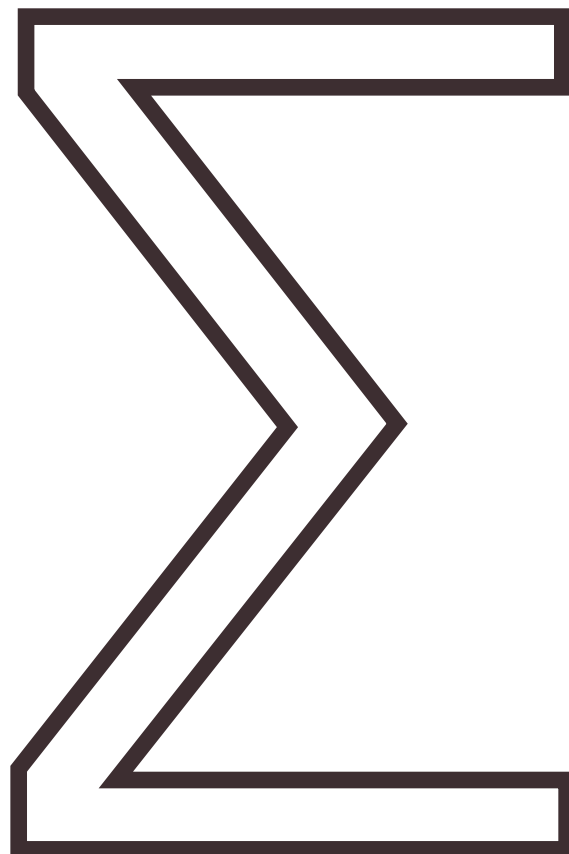
فمثلا لو كان لدينا ١٠ قيم خاصة بدرجات ١٠ طلاب في امتحان ما، نستطيع التعبير عن هذه القيم العشر برقم واحد، ولكن الوسيط Mean، وهذا الوسيط هو تجميع Aggregation لهذه القيم.

Α α Β β Γ γ Δ δ  
 Ε ε Ζ ζ Η η Θ θ  
 Ι ι Κ κ Λ λ Μ μ  
 Ν ν Ξ ξ Ο ο Π π  
 Ρ ρ Σ σ ς Τ τ Υ υ  
 Φ φ Χ χ Ψ ψ Ω ω

تعتبر الأبجدية اليونانية مصدرا

شائعا لكتابة التدوين [الترميز].

على غرار الإنجليزية، توجد أحرف صغيرة وكبيرة في الأبجدية اليونانية.



SUM

لجمع قيم كثيرة معًا نستخدم رمزًا

يونانيًا يسمى **Sigma**

A	α	B	β	Γ	γ	Δ	δ
E	ε	Z	ζ	H	η	Θ	θ
I	ι	K	κ	Λ	λ	M	μ
N	ν	Ξ	ξ	O	ο	Π	π
P	ρ	Σ	σς	T	τ	Υ	υ
Φ	φ	X	χ	Ψ	ψ	Ω	ω

على وجه التحديد، نستخدم **Sigma** المكتوب بحرف كبير.



نستخدم الرمز بشكل عام بالطريقة التالية:



ستلاحظون أنه بدلاً من كتابة قيم  $x$  متعددة، وكل منها  
برمز سفلي مختلف، نكتب  $x$  واحدة مع الرمز السفلي  $i$ .

وهنا، يعتبر  $i$  عنصرًا نائبًا يخبرنا عن [ترتيب] قيم  $x$  التي سيتم تجميعها.

لذا عند جمع أول قيمتين

فقط، نريد أن تكون  $i$  قيمة

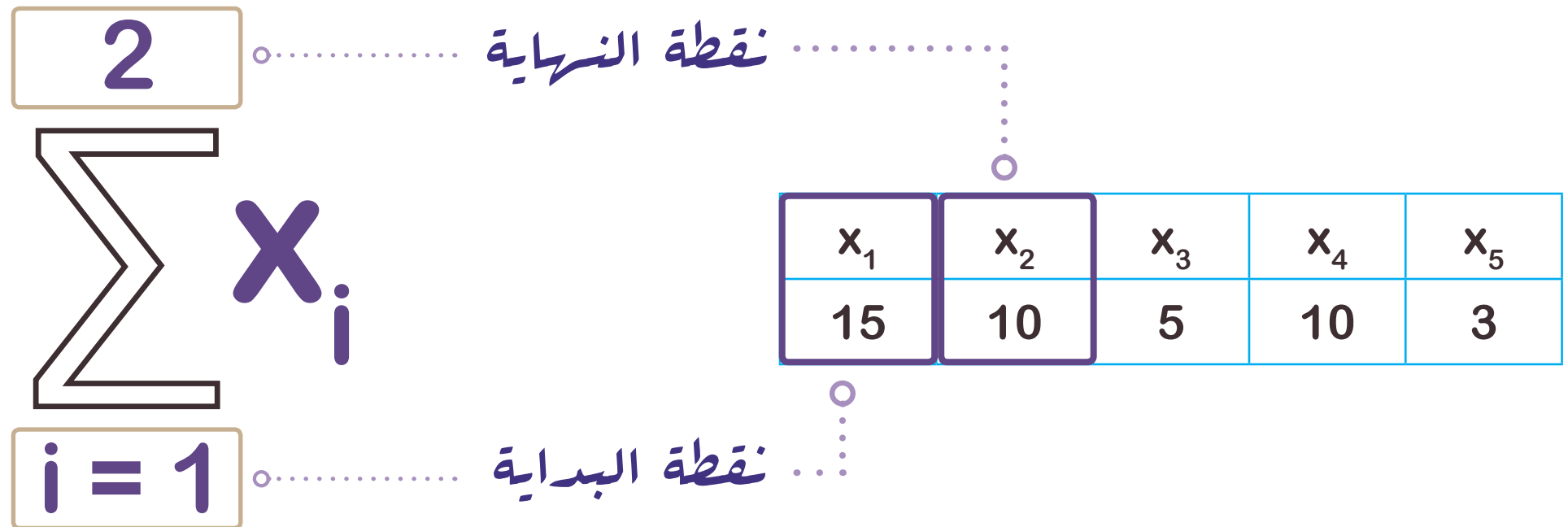
قدرها واحد أو اثنان.

$x_1$   $x_2$

والآن قد تسأل: كيف يُظهر هذا التدوين [الترميز] القيم التي نجمعها معًا؟

إذا كنا نريد جمع أول قيمتين فقط، فسنكتب

شيئًا مثل هذا:



لاحظ أن القيمة الموجودة في الجزء السفلي تعطينا نقطة البداية.

وهذا هو الشكل الذي نود أن تكون قيمة  $x$  الأولى عليه حيث  $i$  يساوي واحدًا

وتعطينا القيمة الموجودة في الجزء العلوي نقطة نهاية للقيمة التي نتوقف عندها،

النهاية لدينا هي اثنان [القيمة الثانية].

تخبرنا **Sigma** بأننا نريد أن نجمع بدءًا من هذه القيمة

بالأسفل، مرورًا بجميع القيم حتى نصل إلى القيمة

التي بالأعلى.

$$\sum_{i=1}^2 x_i = x_1 + x_2$$

$x_1$	$x_2$	$x_3$	$x_4$	$x_5$
15	10	5	10	3

$$= 15 + 10 = 25$$

التدوين [الترميز] بهذه الطريقة الجديدة، باستخدام علامة التجميع [Sigma]، هو نفسه

كالتدوين [الترميز] بهذه الطريقة:  $x_1 + x_2$ .

يؤدي إلى نفس النتيجة ولكن  
بطريقة مختصرة في الكتابة.



# تدوين [ترميز] الوسط

Notation for the Mean

Mohammed Lotfy  
@mohammud.lotfy

الآن إذا أردنا جمع كل القيم في المثال

الأصلي لدينا، فلن نعد بحاجة إلى

كتابة جميع قيم  $x$ .

$x_1$	$x_2$	$x_3$	$x_4$	$x_5$
15	10	5	10	3

$$\sum_{i=1}^5 x_i = x_1 + x_2 + x_3 + x_4 + x_5$$

بدلاً من ذلك، يمكننا كتابة عملية الجمع بدءاً من  $i = 1$  وانتهاءً بالقيمة الخامسة.

هذه هي أفضل طريقة عندما تريد الجمع على نطاق أوسع، 10 أو 20 أو 100 قيمة.

$$\sum_{i=1}^{10} x_i$$

$$\sum_{i=1}^{20} x_i$$

$$\sum_{i=1}^{100} x_i$$

لم نعد بحاجة إلى كتابة جميع قيم  $x$ .

ومع ذلك، يمكننا أن نكون أكثر كفاءة.



في كل مرة تتغير مجموعة البيانات

لدينا، يتعين عليّ أن أغير هذا العدد

في الأعلى للإشارة إلى عدد القيم التي

أجمعها.

$$\sum_{i=1}^2 x_i$$

إذا كنا نريد بالفعل أن نجمع

كل القيم لدينا، يمكننا

استبدال هذه القيمة العليا

بـ  $n$ . والآن، سيناسب الرمز

لدينا أي مجموعة بيانات.

$$\sum_{i=1}^n x_i$$

جمع كل القيم

جميع القيم في عمود ما في مجموعة البيانات  $n =$

حتى الآن، لدينا طريقة لجمع كل القيم التي لدينا، بغض النظر عن عددها في مجموعة البيانات.

لإنهاء حساب الوسط  $mean$ ، نحتاج إلى قسمة هذا

المجموع على عدد القيم الموجودة في مجموعة البيانات،

وما هذا العدد إلا الرمز  $n$ .

ستلاحظون عادة أن وسط mean

مجموعة البيانات يحمل مثل هذا

التدوين [الترميز] الذي نطقه

.x-bar

$\bar{x}$

ويتم حسابه باستخدام التدوين الذي ينص على جمع كل القيم الموجودة في مجموعة البيانات، ثم قسمتها على عدد القيم الموجودة فيها.

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

## ملحوظة

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

جميع كل القيم

ثم قسمتها على عدد القيم

ولهذه هي طريقة حساب الوسيط **mean** كما نأخذنا من قبل.

# ما هي مقاييس الانتشار؟

What Are Measures of Spread?

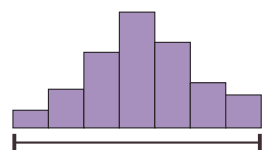
Mohammed Lotfy  
@mohammud.lotfy

## كيف تبعد القيم عن بعضها البعض

في الدروس القادمة  
سنشرح مقاييس الانتشار،  
وعلاقتها بالشكل والقيم  
الشاذة

يساعدنا على تحديد مدى  
اختلاف القيم عن بعضها  
البعض، فلو كان التشتت  
كبيرا فهذا يعني أن  
الفروق بين القيم كبير،  
والعكس إذا كان صغيرا.

## الانتشار Spread



هي قيم كبيرة جدا أو صغيرة  
جدا مقارنة ببقية القيم الأخرى،  
لذا يجب الانتباه لها لأنها تؤثر  
على عملية التحليل كما نرى فيما  
بعد.

## القيم الشاذة

## Outliers



## المتوسط Mean الوسيط Median الوضع [المنوال] Mode

في الدروس  
السابقة دررنا  
مقاييس المركز

## المركز Center

يبحث عن القيمة التي تتمركز  
حولها جميع القيم، أي القيمة  
التي تقع في منتصف القيم،  
وتفيدنا في تلخيص القيم  
جميعها في قيمة واحدة، فتصبح  
هي المتوسط Average.



## الشكل Shape

هو رسم يمثل مركز  
وانتشار مجموعة من  
البيانات بصورة مرئية  
وسهلة الفهم.



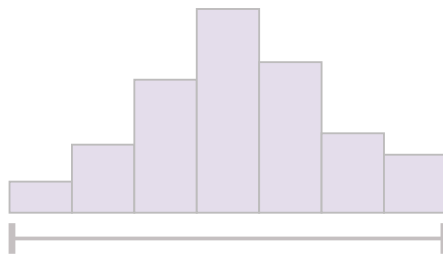
تشتمل قيم قياس الانتشار على:

① النطاق [المدى] Range

② النطاق الإرباعي [المدى الربيعي] Interquartile Range

③ التباين Variance

④ الانحراف المعياري Standard Deviation



# المدرجات التكرارية

## Histograms

Mohammed Lotfy  
[@mohammud.lotfy](https://www.instagram.com/mohammud.lotfy)



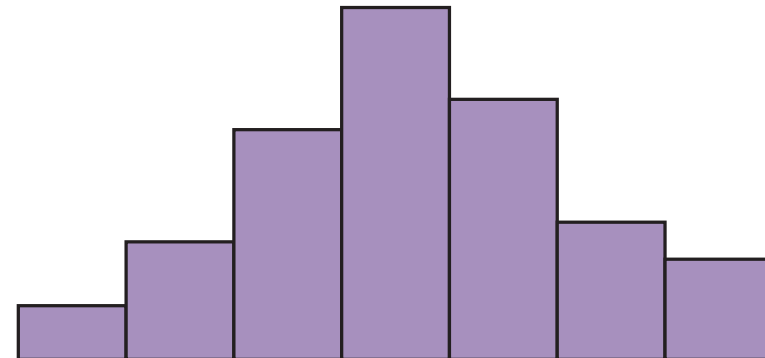
## من الأسهل فهم انتشار [تشتت] البيانات بطريقة مرئية

أكثر الوسائل شيوعاً لتمثيل

البيانات الكمية Quantitative

Data هي المدرجات التكرارية

Histograms



لفهم كيفية إنشاء المدرج التكراري Histogram

لنفترض وجود البيانات التالية:

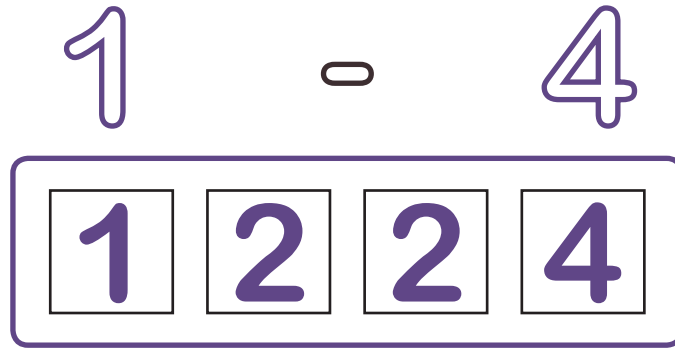
1 2 2 4 5 7 8 9 12 15

أولاً: نحتاج إلى تقسيم ما لدينا من بيانات إلى فئات [مجموعات] Bins

وتستطيع تحديد هذه الفئات بنفسك

في هذا المثال سنقوم بتقسيم البيانات إلى 4 فئات bins. كل فئة تتسع لأربع قيم بتكراراتها

الفئة الأولى، تبدأ من ١ وتنتهى بـ ٤

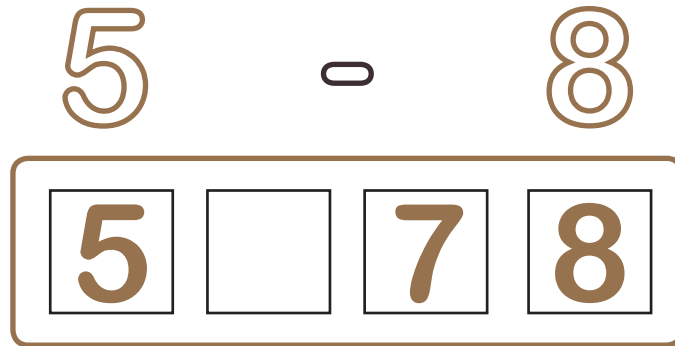


جميع القيم التي تقع في هذا النطاق ستكون في هذه الفئة



لاحظ تكرار العدد ٢ مرتين فتم وضعه مرتين

الفئة الثانية، تبدأ من ٥ وتنتهى بـ ٨

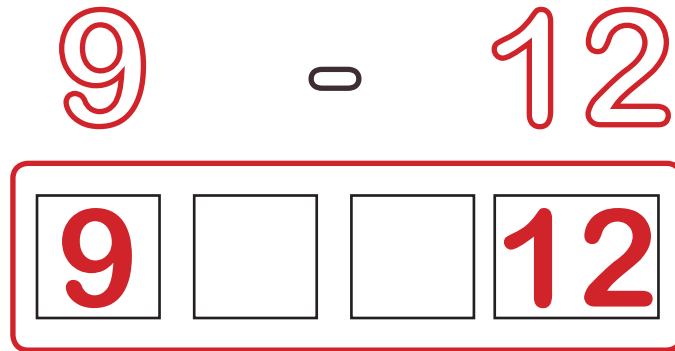


جميع القيم التي تقع في هذا النطاق ستكون في هذه الفئة



لا توجد سوى ٣ قيم فقط تقع في هذا النطاق

الفئة الثالثة، تبدأ من ٩ وتنتهى بـ ١٢



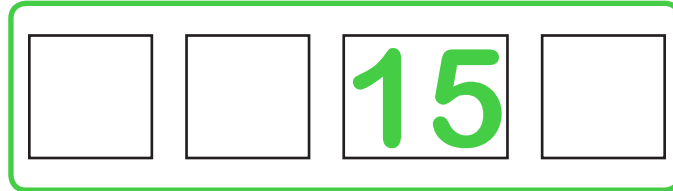
جميع القيم التي تقع في هذا النطاق ستكون في هذه الفئة



لا توجد سوى قيمتين فقط تقع في هذا النطاق

الفئة الثالثة، تبدأ من ١٣ وتنتهى بـ ١٦

$$13 - 16$$



جميع القيم التي تقع في هذا النطاق ستكون في هذه الفئة

1 2 2 4 5 7 8 9 12 15

قيمة واحدة فقط تقع في هذا النطاق

ملاحظة هامة: يتم وضع كل قيمة بتكراراتها داخل النطاق الخاص بها

فليس معنى أن النطاق من ١ - ٤ أنه لن يتسع إلا إلى ٤ قيم فقط، بل يمكن أن يكون به ١٠ قيم، إذا كانت القيم كالتالي:

1 - 4

1 1 1 2 2 2 2 4 4 4

الآن تم توزيع كل القيم على الفئات أو النطاقات bins الخاصة بها

1 2 2 4

5 7 8

9 12

15

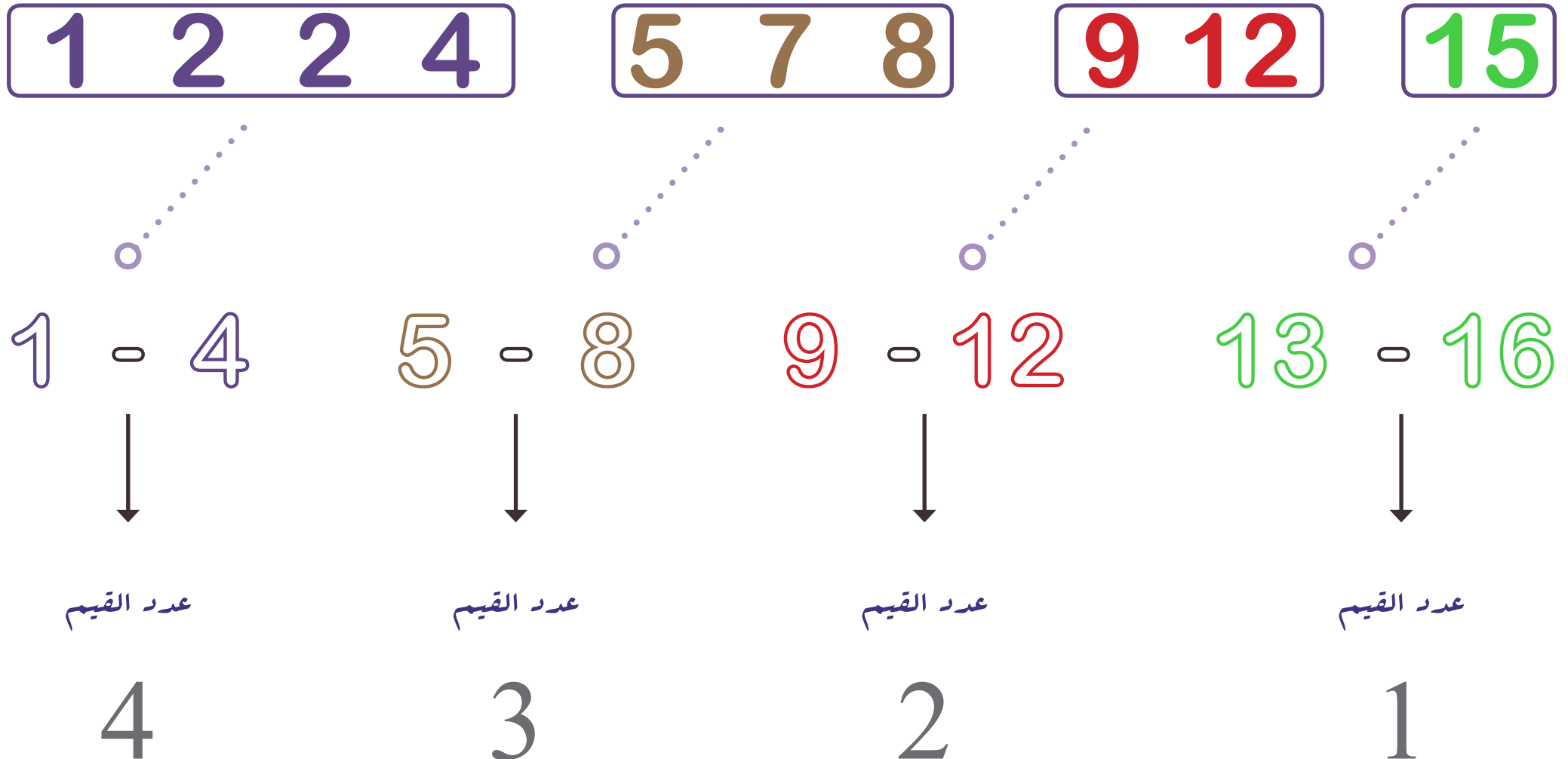
1 - 4

5 - 8

9 - 12

13 - 16

بعد ذلك نستطيع تمثيل كل فئة في المدرج التكراري histogram، بحيث يتم التعبير عن كل فئة على شكل عمود، وطول العمود يتحدد بعدد القيم الموجودة في كل فئة bin

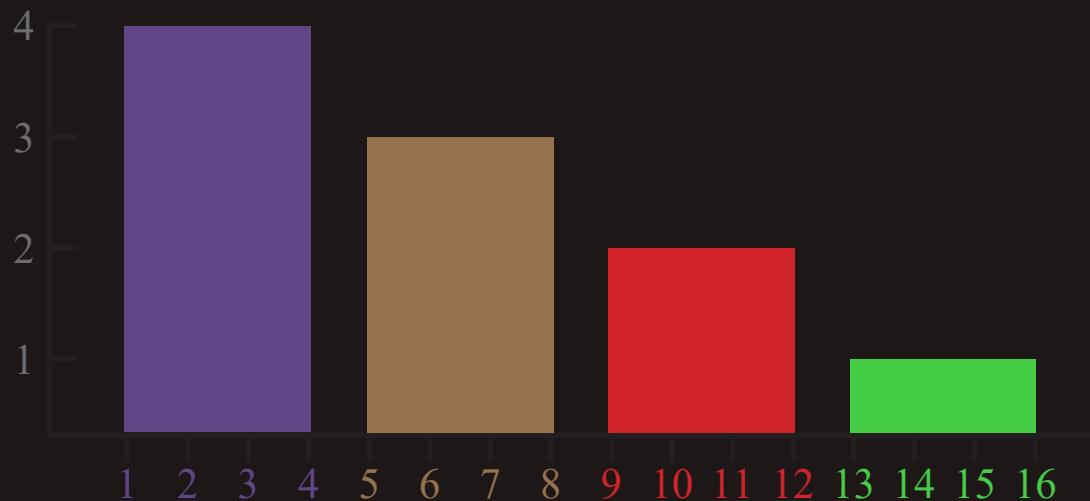




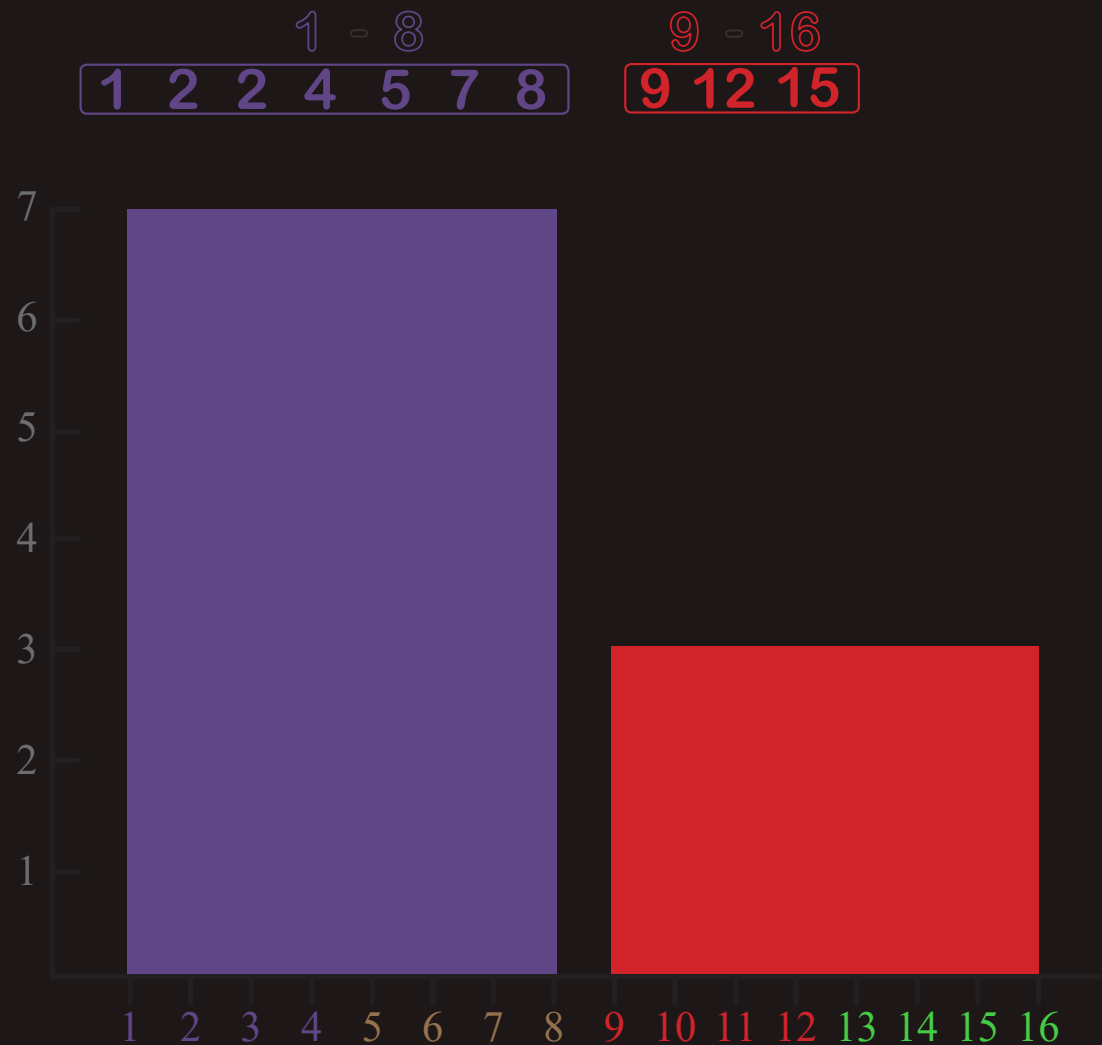
كل نطاق bin يمثل عمود

1 - 4	5 - 8	9 - 12	13 - 16
1 2 2 4	5 7 8	9 12	15

عدد القيم في كل نطاق bin



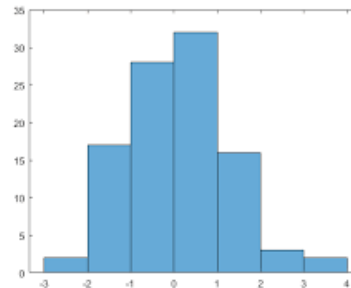
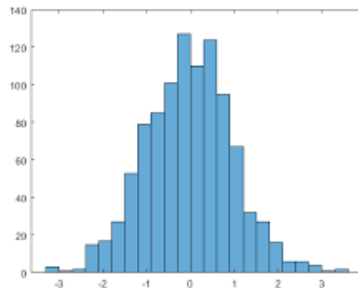
## تغيير النطاقات bins سيؤدي إلى تغيير الرسم



فكما نرى عندما قمنا  
بزيادة مدى النطاق  
فإن عدد الأعمدة أصبح  
أقل، وفقدنا شكل توزيع  
البيانات قليلاً

## لا يوجد اختيار واحد صحيح لتحديد النطاق bin

وفي معظم الحالات سيقوم البرنامج الذي تعمل عليه  
بتحديد النطاقات bins لك

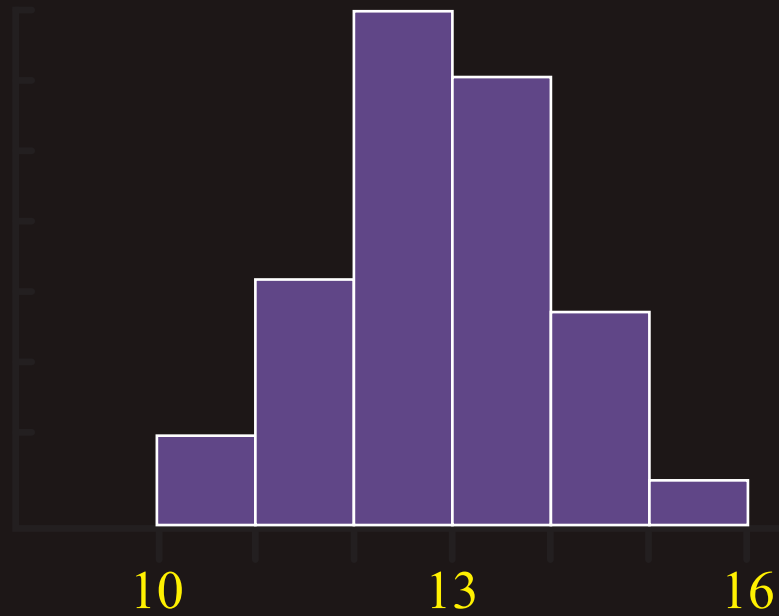


تعدد برامج التحليل الإحصائي،  
فمنها الإكسيل Excel و SPSS  
و Matlab، بالإضافة إلى لغات  
البرمجة الخاصة بالتحليل الإحصائي  
مثل R أو Python، وكل منها  
يسهل عليك كثيرا الحسابات  
والرسومات الخاصة بتحليل البيانات

# ما الفرق؟

## What is the Difference?

Mohammed Lotfy  
[@mohammud.lotfy](https://twitter.com/mohammud.lotfy)

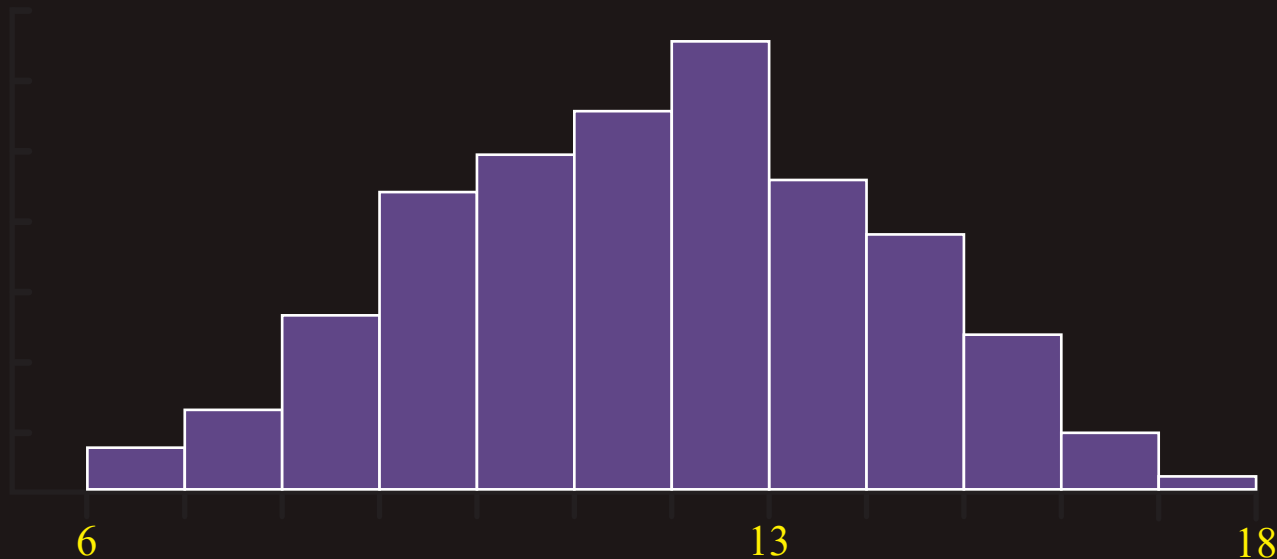


هذان مدرجان تكراريان two

histograms للمقارنة بين

عدد الكلاب التي رأيتها خلال

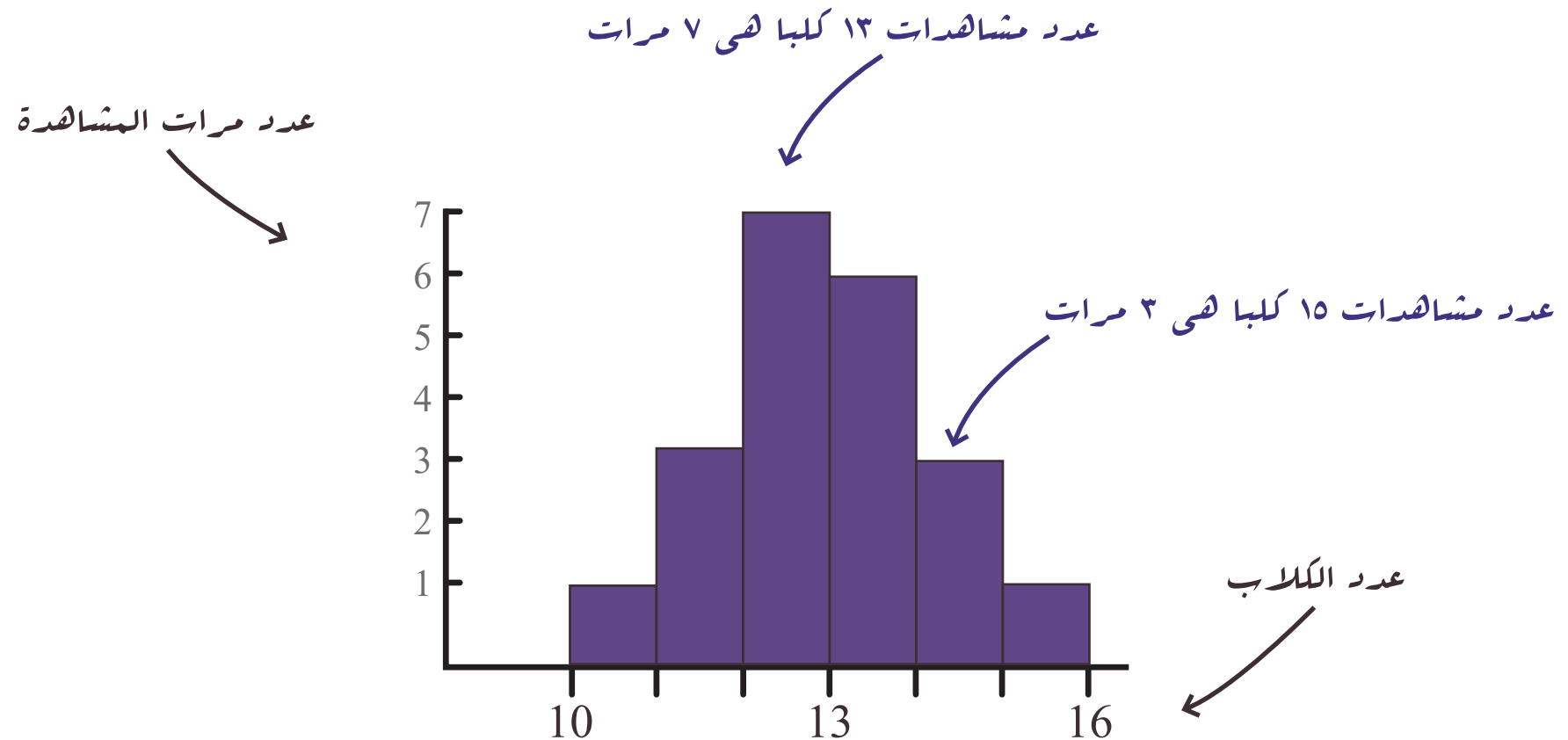
أيام الأسبوع،

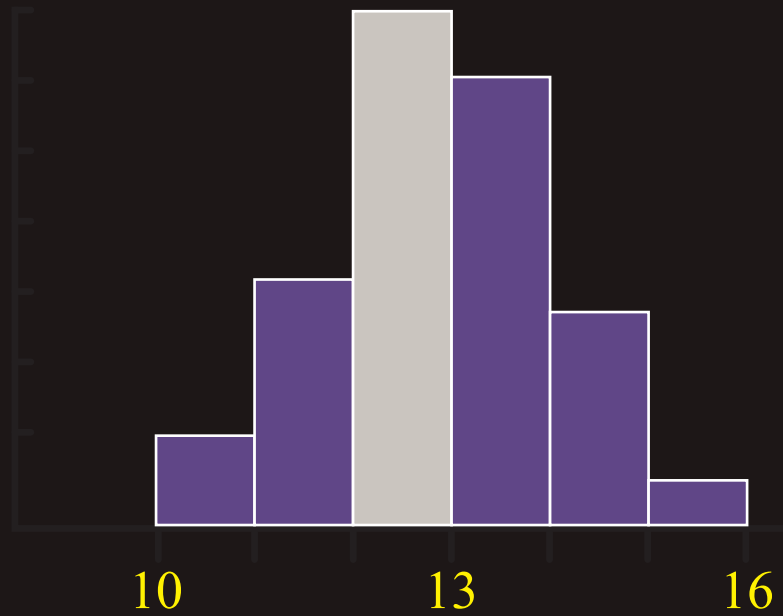


والتي رأيتها في

عطلات نهاية الأسبوع

كيف نقرأ المدرج التكراري؟



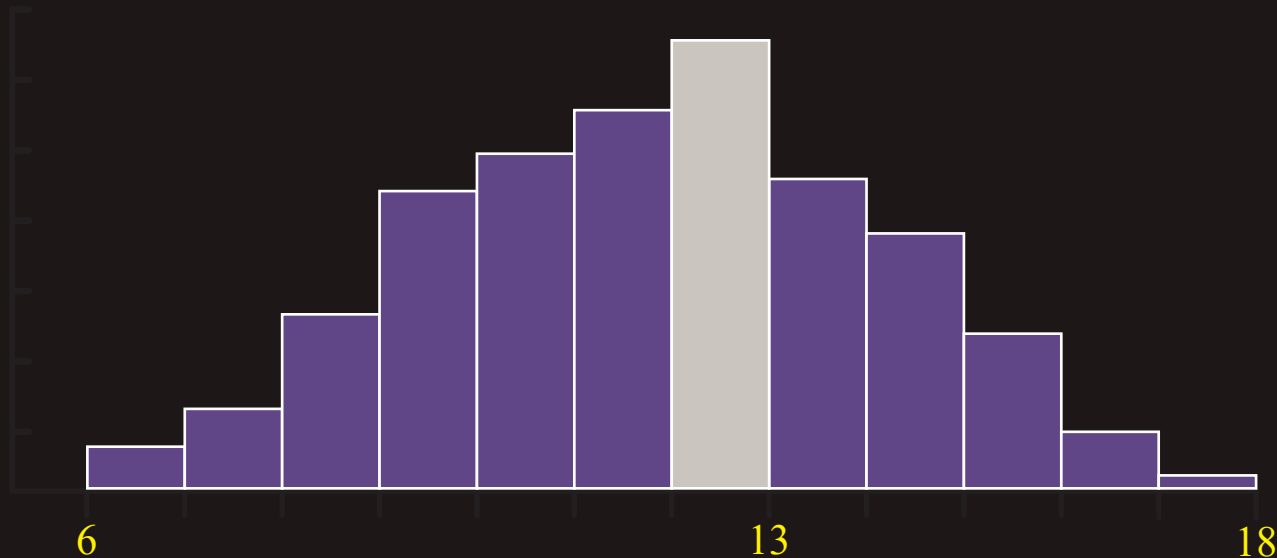


ستلاحظون أن أطول فئة

bin لكل من أيام الأسبوع

وعطلات نهاية الأسبوع هي

عند قيمة 13 كلها



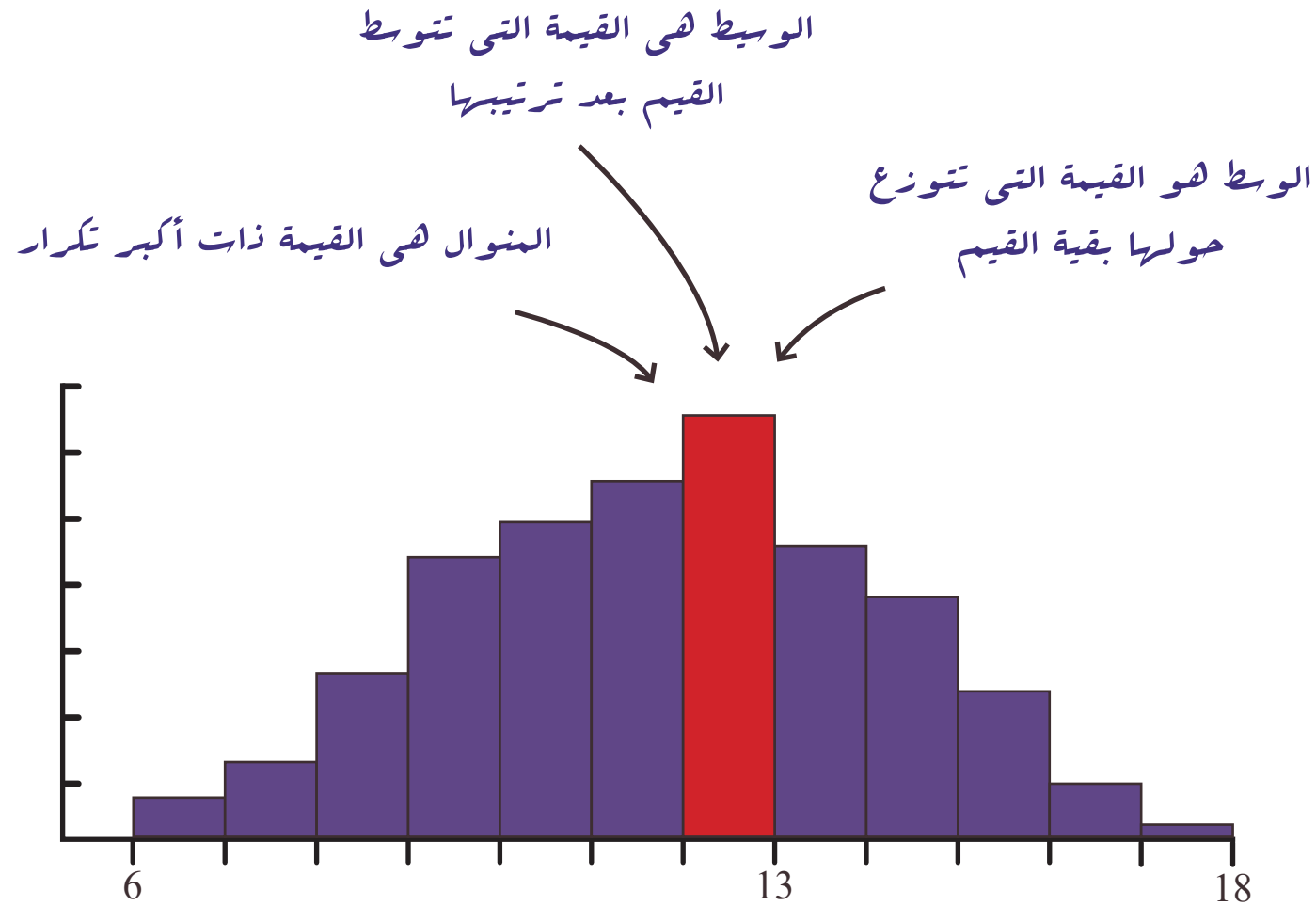
لذا فإن عدد الكلاب التي أتوقع رؤيتها في أيام الأسبوع هو نفسه التي أتوقع رؤيتها في عطلات نهاية الأسبوع

بمعنى آخر

مقاييس المركز measures of center هي نفسها في كلتا الحالتين

الوسط mean = الوسيط median = المنوال mode = 13 كلبا تقريبا





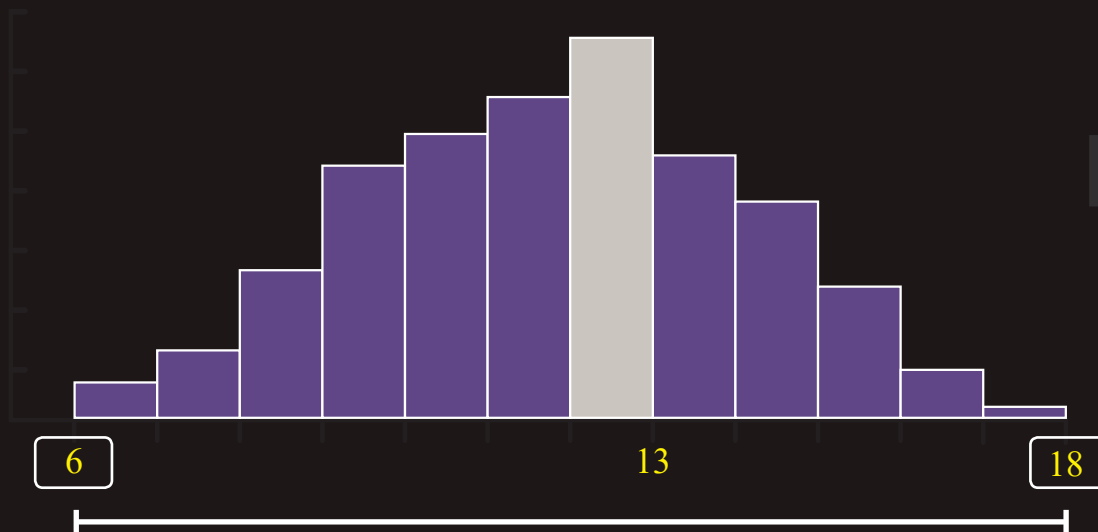
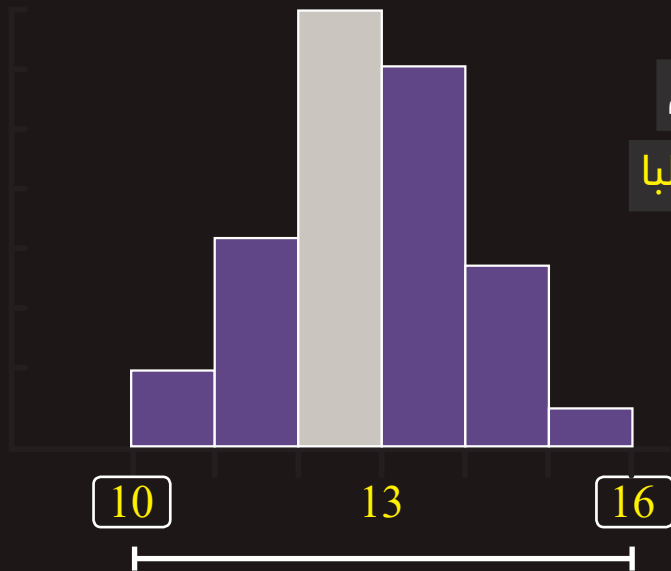
ولكن هناك شيئاً مختلفاً بشأن هذين التوزيعين

الاختلاف هو في

كيفية تشتت أو انتشار

Spread البيانات في

كل مجموعة



# التلخيص ذو الخمسة أرقام

## The 5 Number Summary

Mohammed Lotfy  
[@mohammud.lotfy](https://www.instagram.com/mohammud.lotfy)

إحدى الطرق الأكثر شيوعًا لقياس انتشار البيانات spread of data هي النظر إلى الملخص ذي الخمسة أرقام The 5 Number Summary ، الذي يوفر لنا القيم لحساب المدى ومدى بين ربعين [المدى الإرباعي].

يتكون الملخص ذو الخمسة أرقام من خمس قيم:

- ① الحد الأقصى [أكبر قيمة] ..... Maximum
- ② الربع الثالث [الإرباعي الثالث] ..... Third Quartile
- ③ الربعي الثاني [الإرباعي الثاني] (بالوسيط) ..... Second Quartile (Median)
- ④ الربعي الأول [الإرباعي الأول] ..... First Quartile
- ⑤ الحد الأدنى [أقل قيمة] ..... Minimum

لدينا مجموعة البيانات التالية:

5, 8, 3, 2, 1, 3, 10

أولاً: نرتب القيم:

**1**, 2, 3, 3, 5, 8, **10**

بمجرد ترتيب القيم، يسهل تحديد القيمة الكبرى **maximum value** والصغرى

**minimum value**.

## حساب الوسيط:

الوسيط هو القيمة التي تتوسط القيم بعد ترتيبها

1, 2, 3, 3, 5, 8, 10



القيمة الصغرى  
Minimum



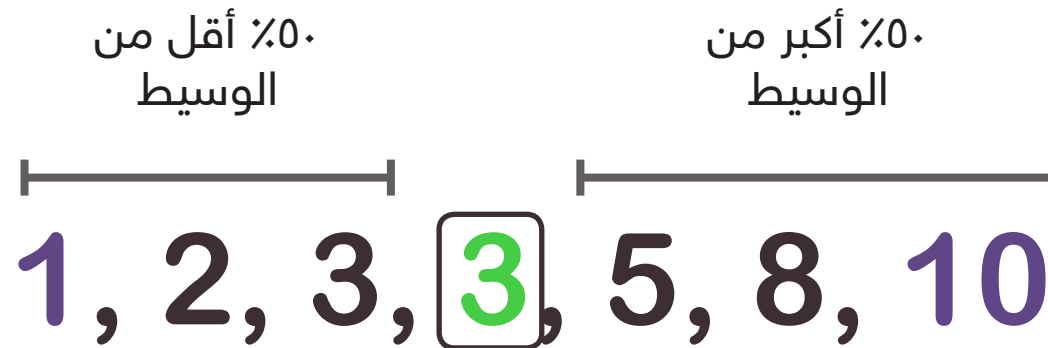
الوسيط  
Median



القيمة الكبرى  
Maximum

نُطلق على الوسيط أيضًا  $Q_2$  أو الربيعي الثاني [الإرباعي الثاني] لأن 50% من بياناتنا أو ربعي [نصف] البيانات أقل من هذه قيمة الوسيط.

والنصف الآخر أكبر من الوسيط



Q<sub>2</sub>  
الوسيط  
Median

Second Quartile  
=  
Median

القيمتان المتبقيتان لإكمال الملخص ذو الخمسة أرقام هما  $Q_1$  و  $Q_3$ .

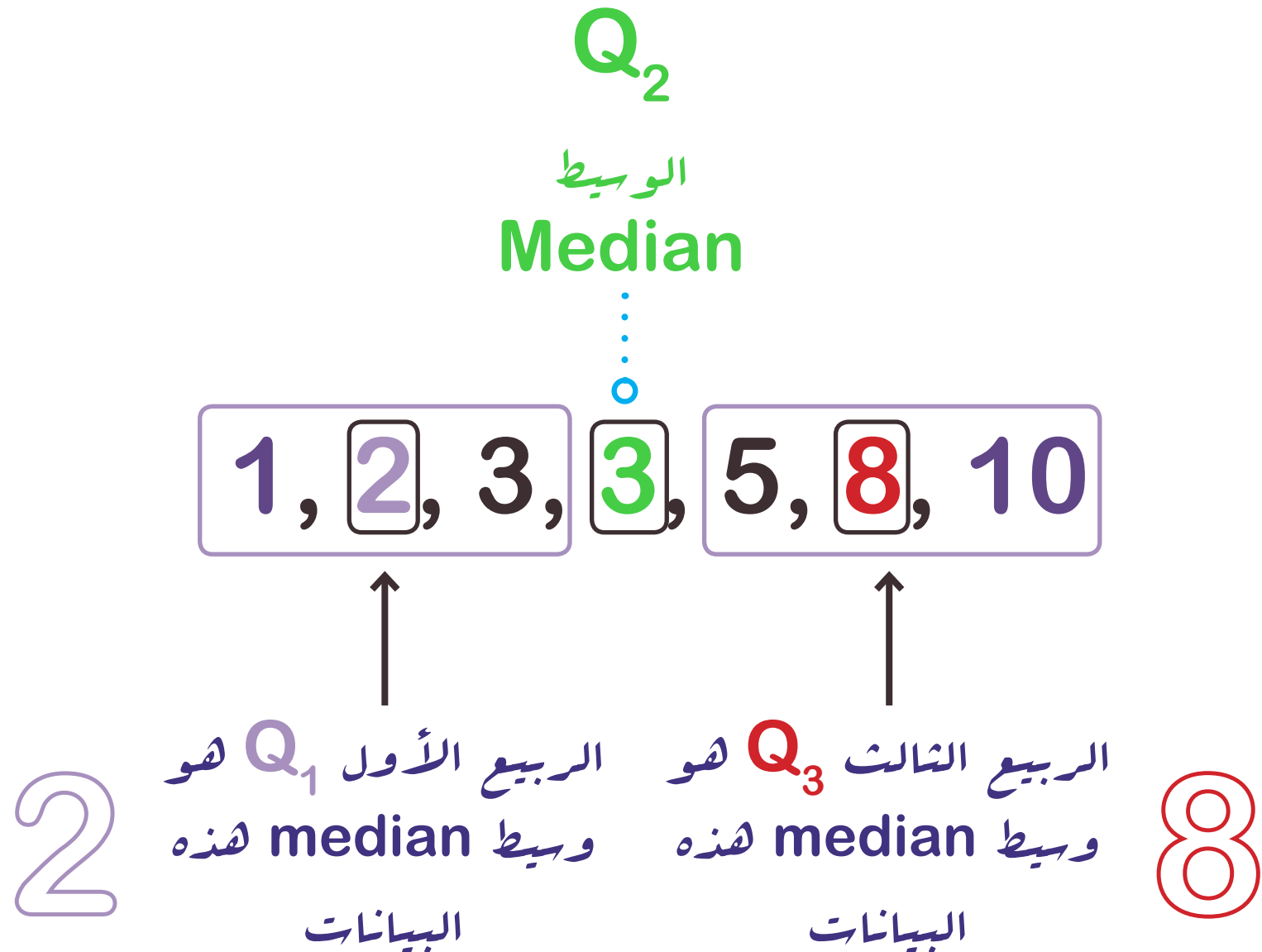
يمكن اعتبار هاتين

القيمتين بأنهما

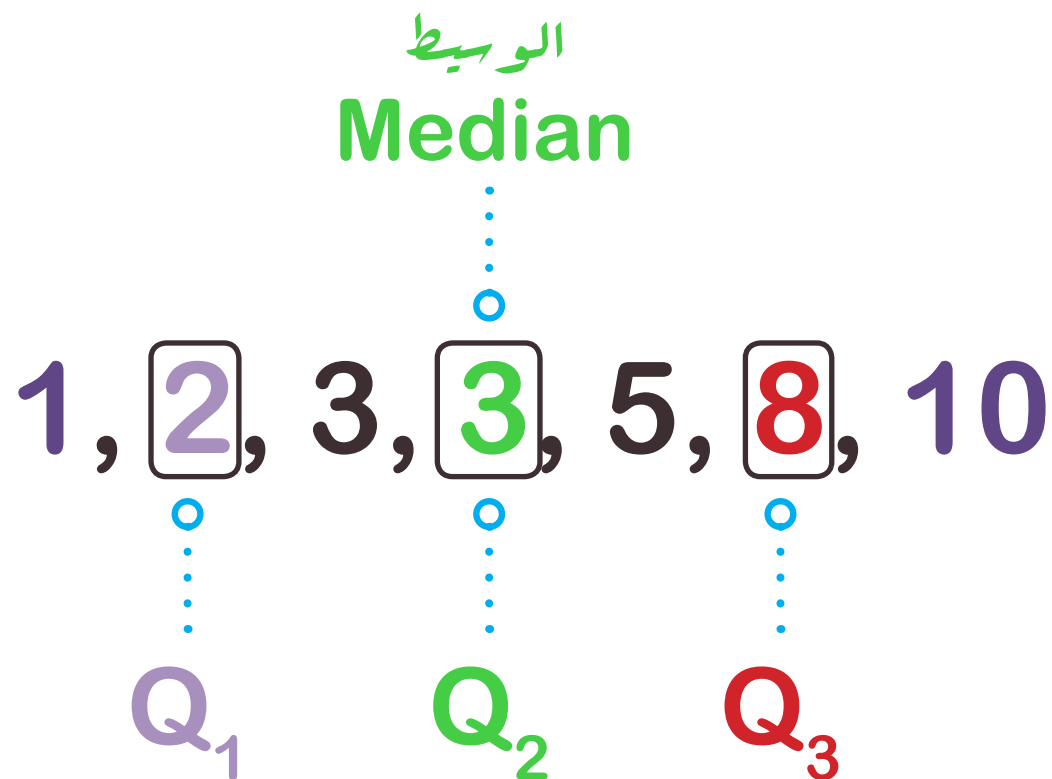
وسيطان للبيانات

الواقعة على كل

طرف من  $Q_2$ .





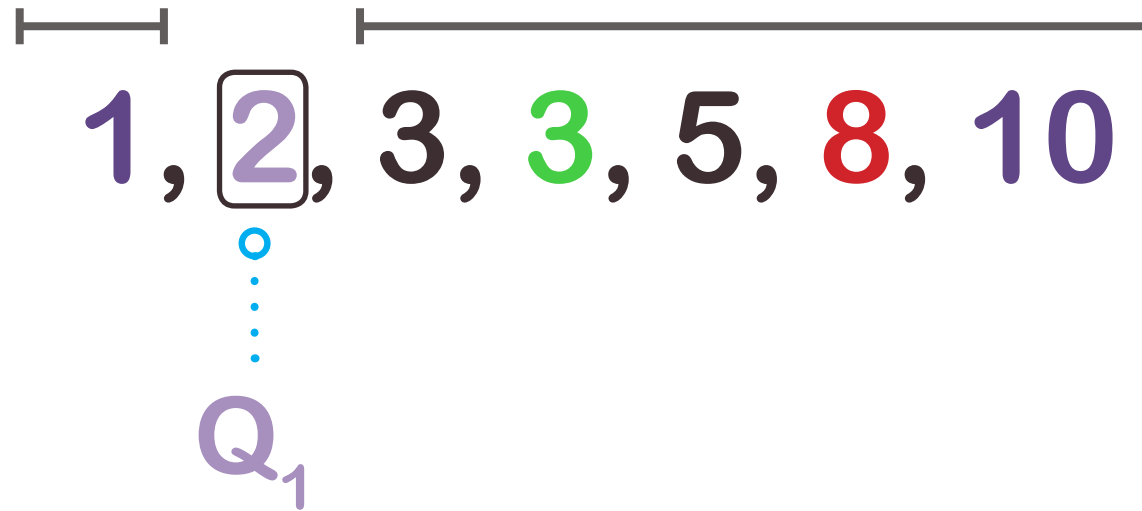


لاحظ أن  $Q_2$  لم تكن ضمن أي من هاتين المجموعتين المستخدمة لحساب

$Q_1$  أو  $Q_3$ .

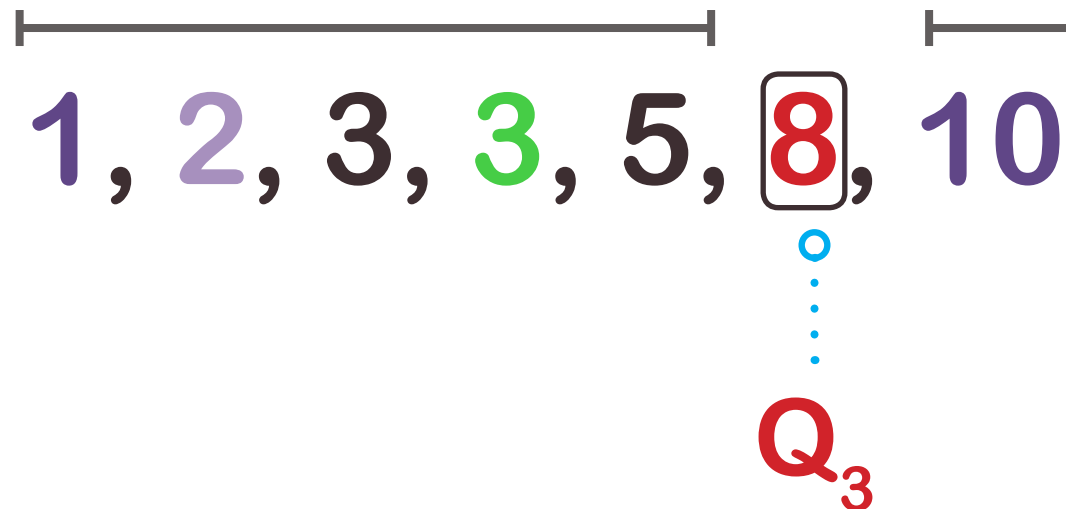
25% أقل من  $Q_1$

75% أكبر من  $Q_1$



75% أقل من  $Q_3$

25% أكبر من  $Q_3$



وهذه هي القيم التي تعبر عن الملخص ذي الخمسة

أرقام *The 5 Number Summary*

الوسيط  
Median

1 , 2 , 3 , 3 , 5 , 8 , 10

$Q_2$

$Q_3$

$Q_1$

Minimum

Maximum

لنفكر في مثال آخر في مجموعة البيانات لها عدد زوجي من القيم:

5, 8, 3, 2, 1, 3, 10, 105

مرة أخرى نحتاج إلى ترتيب القيم أولاً:

1, 2, 3, 3, 5, 8, 10, 105

Minimum

Maximum

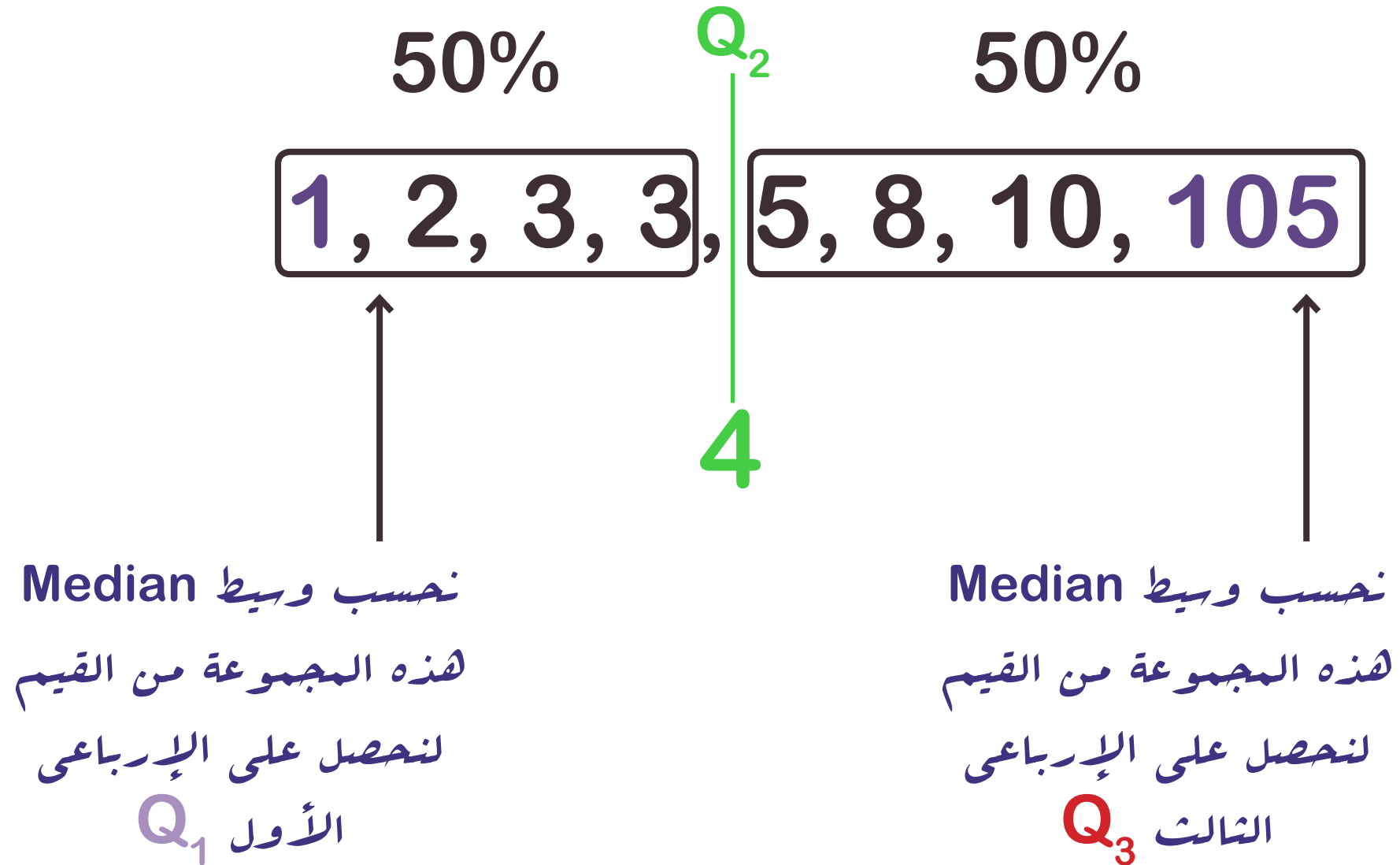
يمكنك تحديد القيمة الكبرى والصغرى بسهولة.

تذكر أنه عندما يكون عدد القيم زوجي فإن الوسيط **Median** يكون متوسط القيمتين اللتان في منتصف البيانات.

1, 2, 3, 3, 5, 8, 10, 105

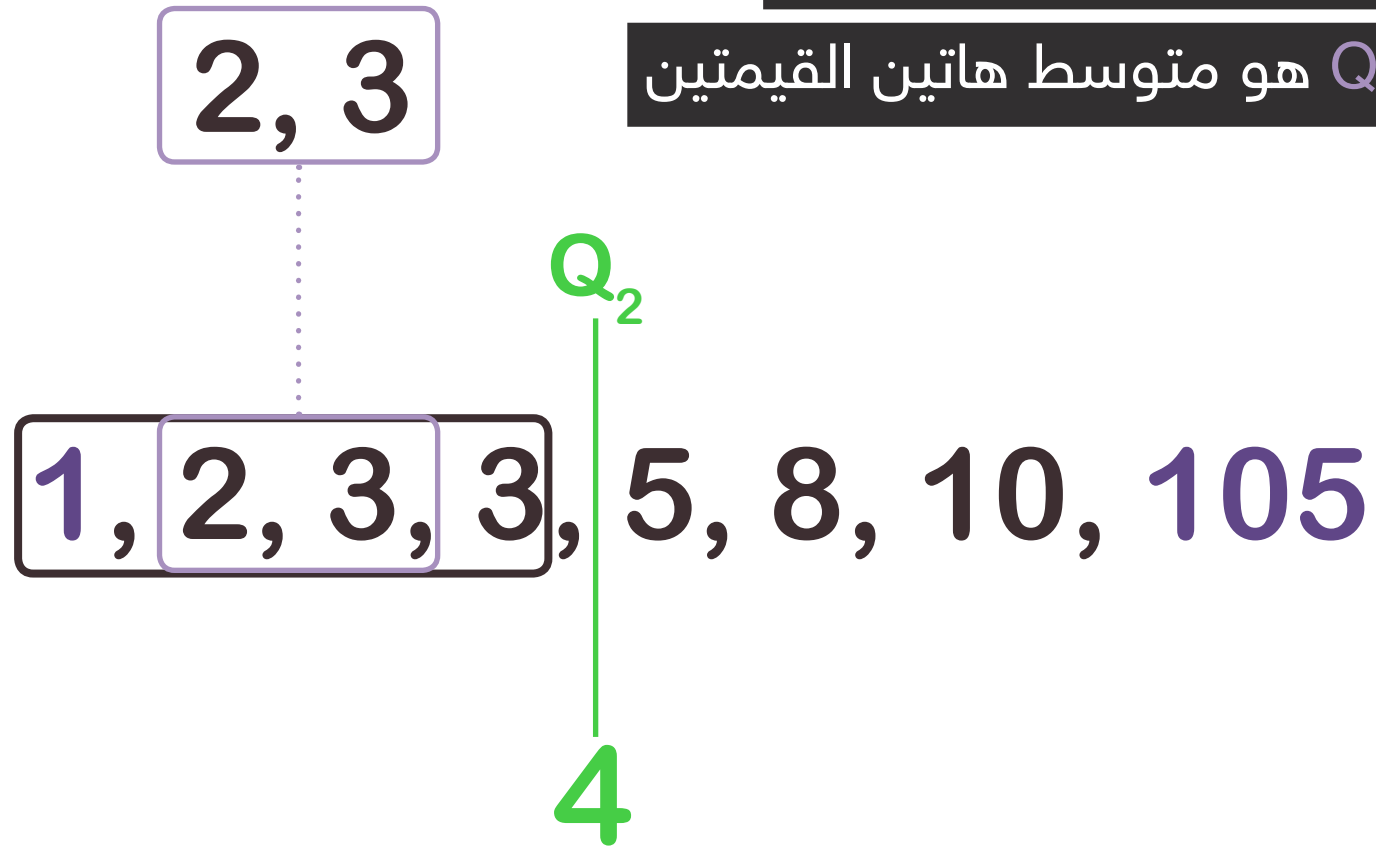
$$Q_2 = (3 + 5)/2 = 4$$

لإيجاد  $Q_1$  و  $Q_3$  نقسم مجموعة البيانات عند الوسيط Median إلى نصفين:



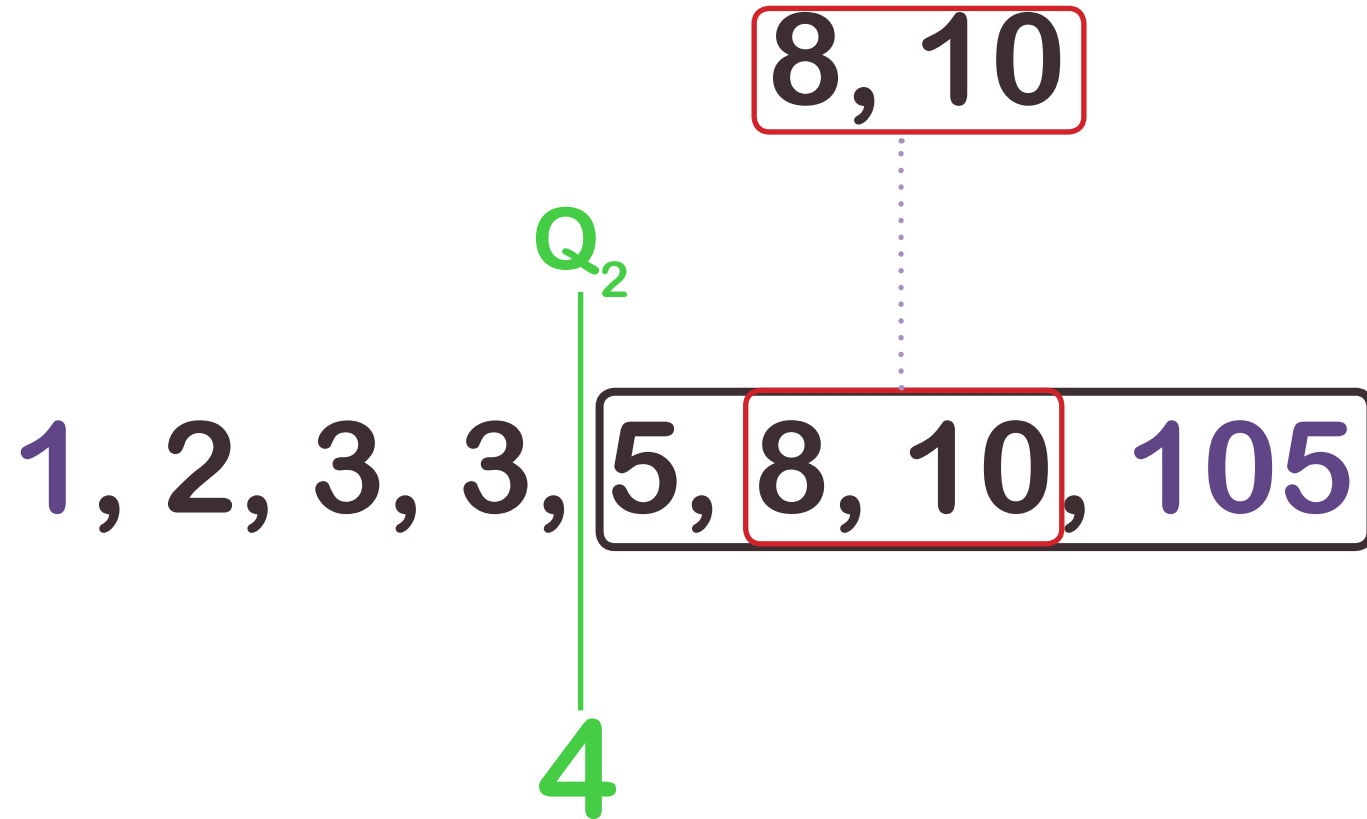
بالنسبة إلى مجموعة البيانات هذه،

سيكون  $Q_1$  هو متوسط هاتين القيمتين



$$Q_1 = (2 + 3)/2 = 2.5$$

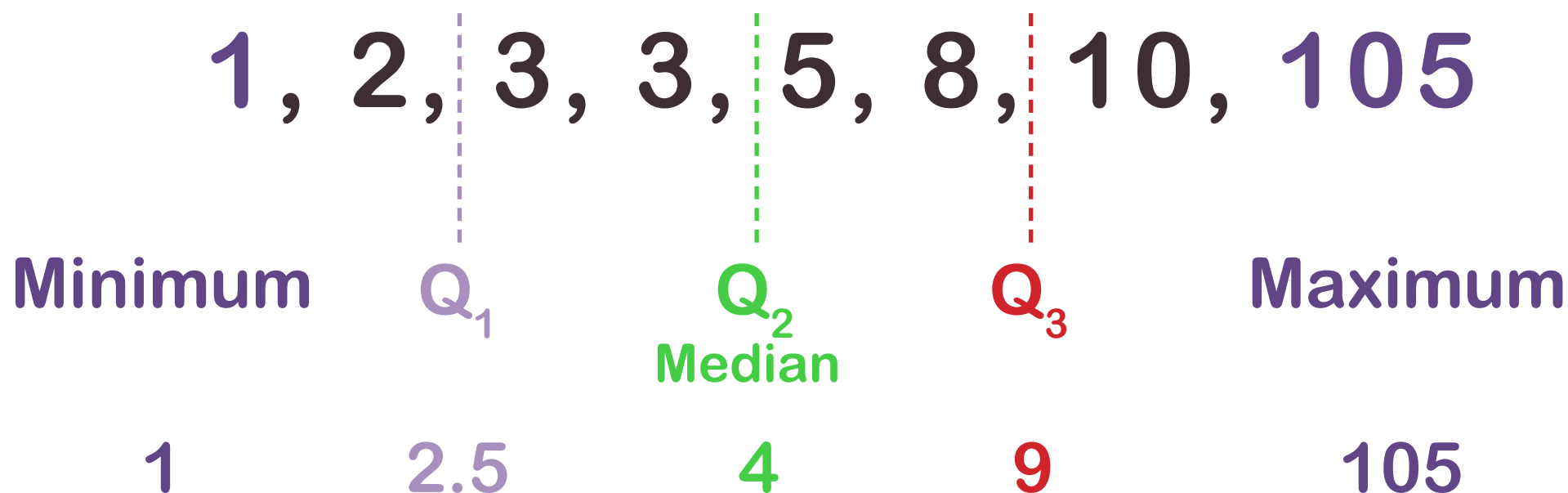
وسيكون  $Q_3$  هو متوسط هاتين القيمتين:



$$Q_3 = (8 + 10)/2 = 9$$



يقدم إلينا هذا الملخص ذو الخمسة أرقام **The 5 Number Summary** كما يلي:



بمجرد أن نحسب كل القيم الخاصة بالملخص ذي الخمسة أرقام لا يمثل إيجاد المدى

Rabge ومدى بين ربعين [المدى الإرباعي] Interquartile Range مشكلة

المدى = أكبر قيمة - أقل قيمة

$$\text{Range} = \text{Maximum} - \text{Minimum}$$

المدى الإرباعي = الإرباعي الثالث - الإرباعي الأول

$$\text{Interquartile Range (IQR)} = Q_3 - Q_1$$

1, 2, 3, 3, 5, 8, 10

لمجموعة البيانات الأولى:

المدى =  $10 - 1 = 9$

$$\text{Range} = 10 - 1 = 9$$

1, 2, 3, 3, 5, 8, 10, 105

لمجموعة البيانات الثانية:

المدى =  $105 - 1 = 104$

$$\text{Range} = 105 - 1 = 104$$

1, 2, 3, 3, 5, 8, 10

لمجموعة البيانات الأولى:

المدى الإرباعي = 8 - 2 = 6

Interquartile Range (IQR) = 8 - 2 = 6

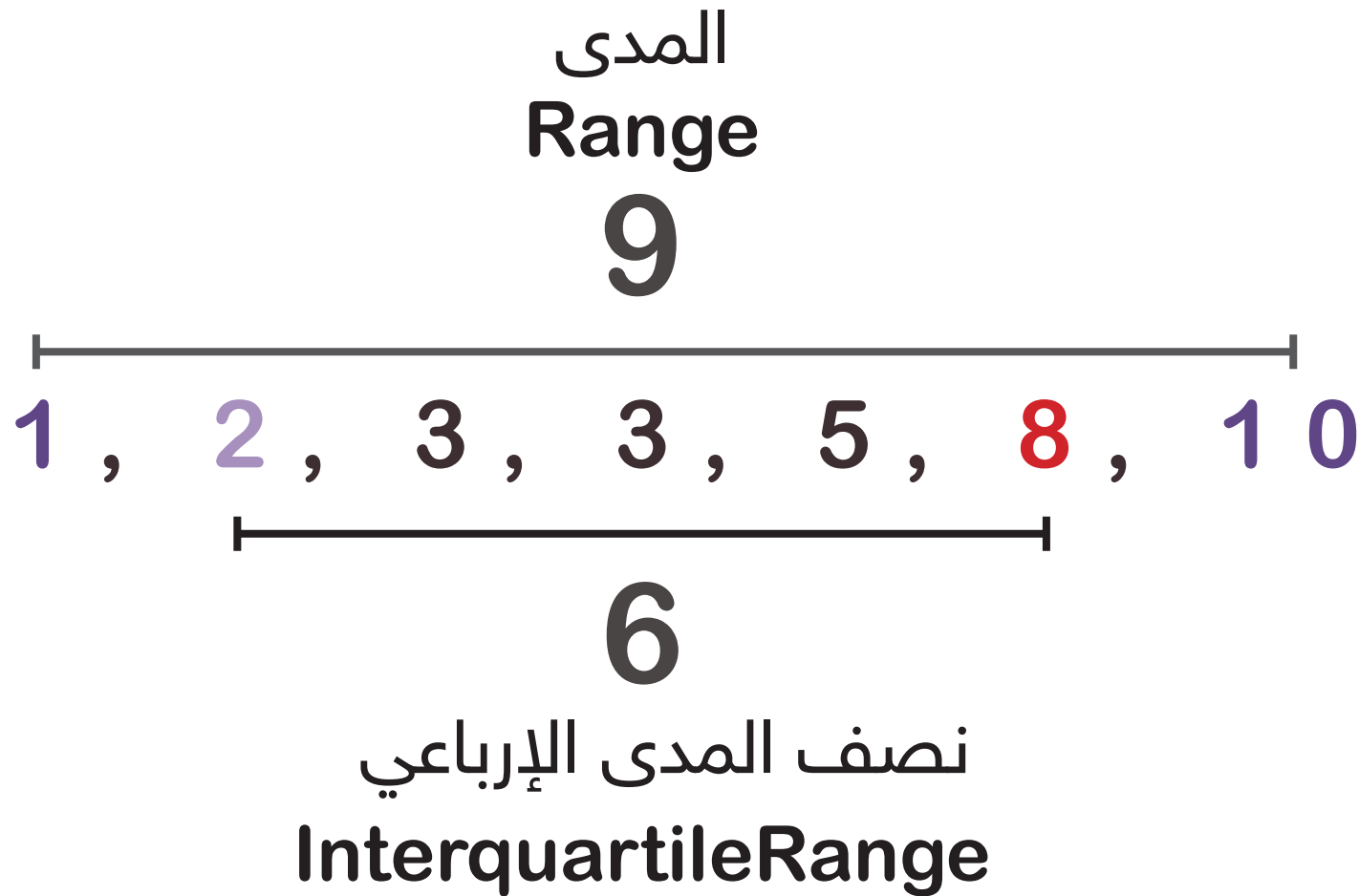
1, 2, 2.5, 3, 3, 5, 8, 9, 10, 10.5

لمجموعة البيانات الثانية:

المدى الإرباعي = 9 - 2.5 = 6.5

Interquartile Range (IQR) = 9 - 2.5 = 6.5

مجموعة البيانات الأولى:



## مجموعة البيانات الثانية:

ملاحظة أن المدى  
range يتأثر  
بالقيم الشاذة، في  
حين أن المدى  
الإرباعي لا يتأثر  
بها.

للمزيد راجع  
الشرح المفصل

ب عنوان: مقاييس

الانتشار

Measures of  
Spread

المدى  
Range  
104

1, 2, 3, 3, 5, 8, 10, 105  
2.5 9

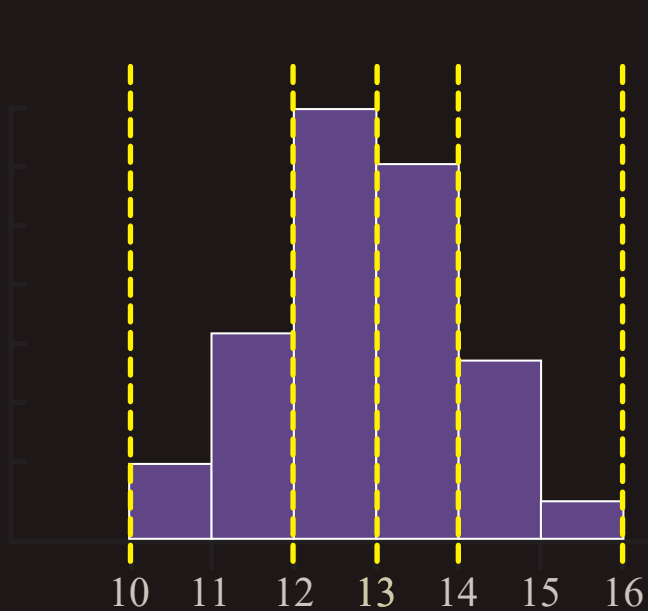
7.5

نصف المدى الإرباعي  
InterquartileRange

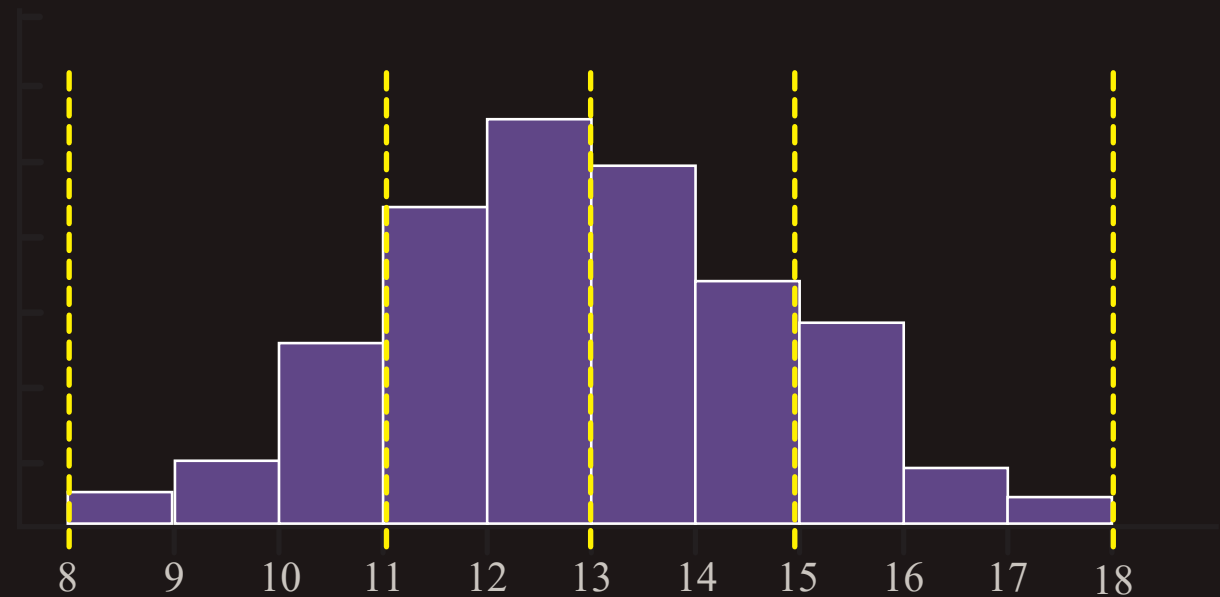
# Box Plots

Mohammed Lotfy  
**@mohammud.lotfy**

بالرجوع إلى التوزيعات التي وجدناها لعدد الكلاب التي أراها، يمكننا تمييز القيم  
للملخص ذي الخمسة أرقام كالتالي:



بقية أيام الأسبوع

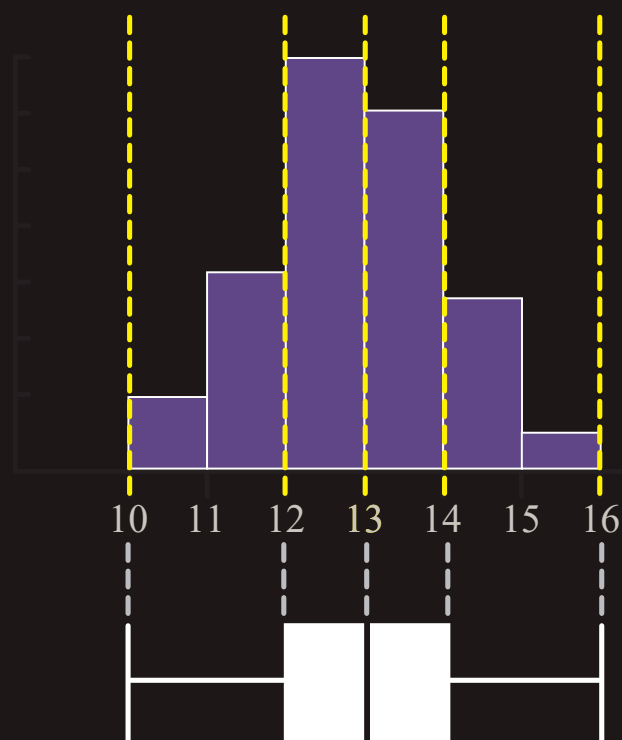


نهاية الأسبوع

إذا أخذنا في الاعتبار فقط هذه العلامات، سينتج مخططًا معروفًا للبيانات يعرف

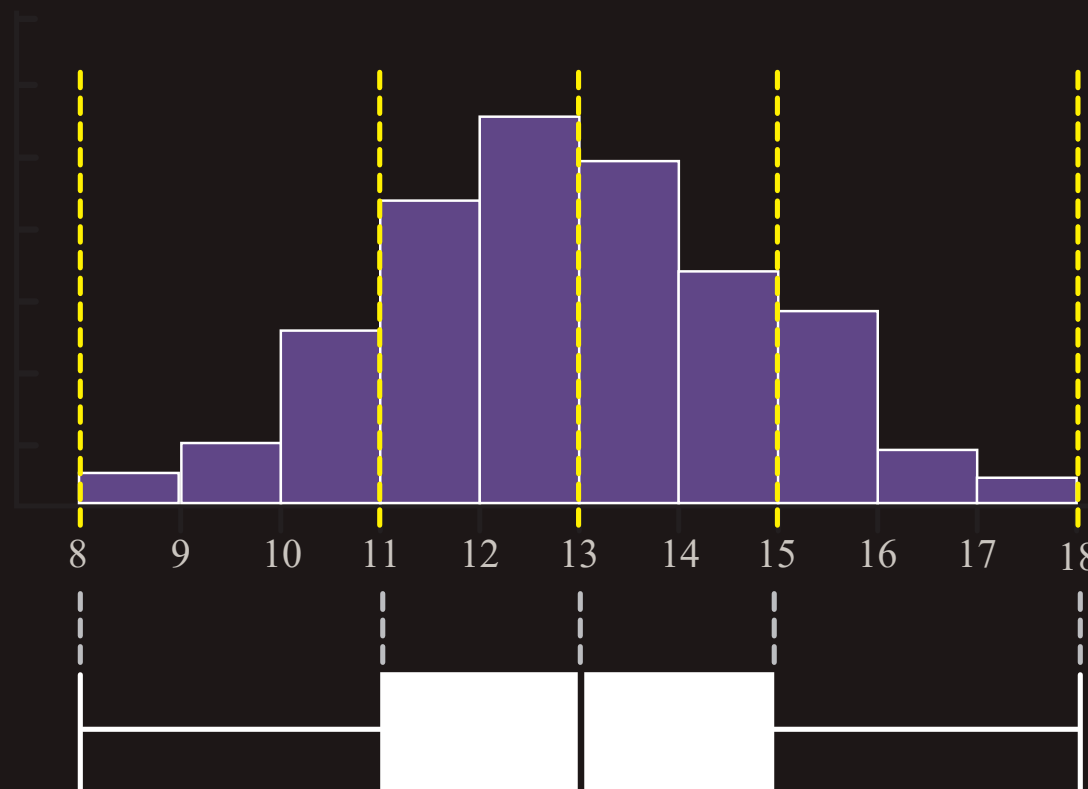
باسم `box plot`.





Minimum Q<sub>1</sub> Q<sub>2</sub> Q<sub>3</sub> Maximum

10 12 13 14 16

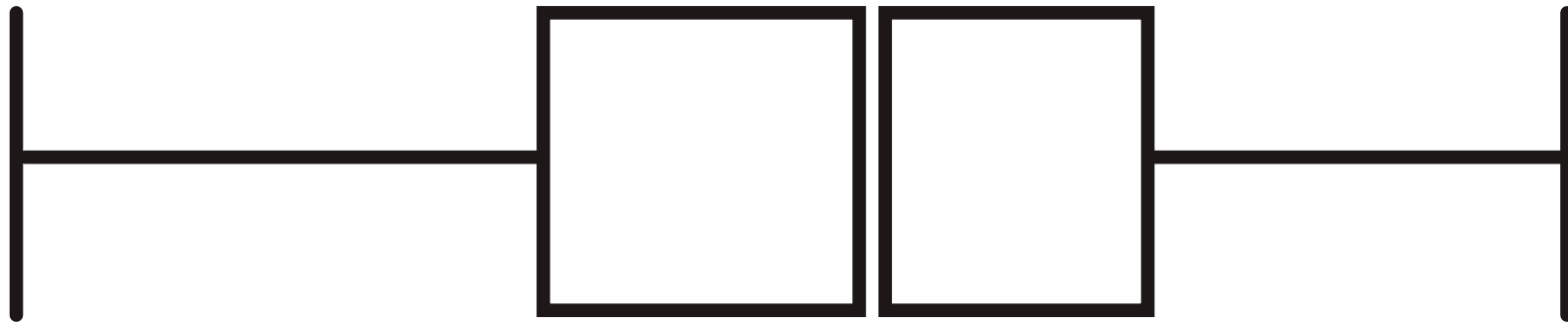


Minimum Q<sub>1</sub> Q<sub>2</sub> Q<sub>3</sub> Maximum

8 11 13 15 18

من كلٍ من المدرج التكراري histogram والملخص ذي الخمسة أعداد، يمكننا أن نرى سريعاً أن عدد الكلاب التي أراها في العطلات الفروق بينها أكبر من الفروق بين عدد الكلاب التي أراها في أيام الأسبوع.

من الممكن أن يكون box plot مفيدًا لسرعة المقارنة بين انتشار مجموعتي بيانات من خلال بعض القياسات الرئيسية، مثل الإرباعيات  $Q_1$ ،  $Q_2$  and  $Q_3$ ، وأكبر وأصغر قيمة  $maximum$  and  $minimum$  value.

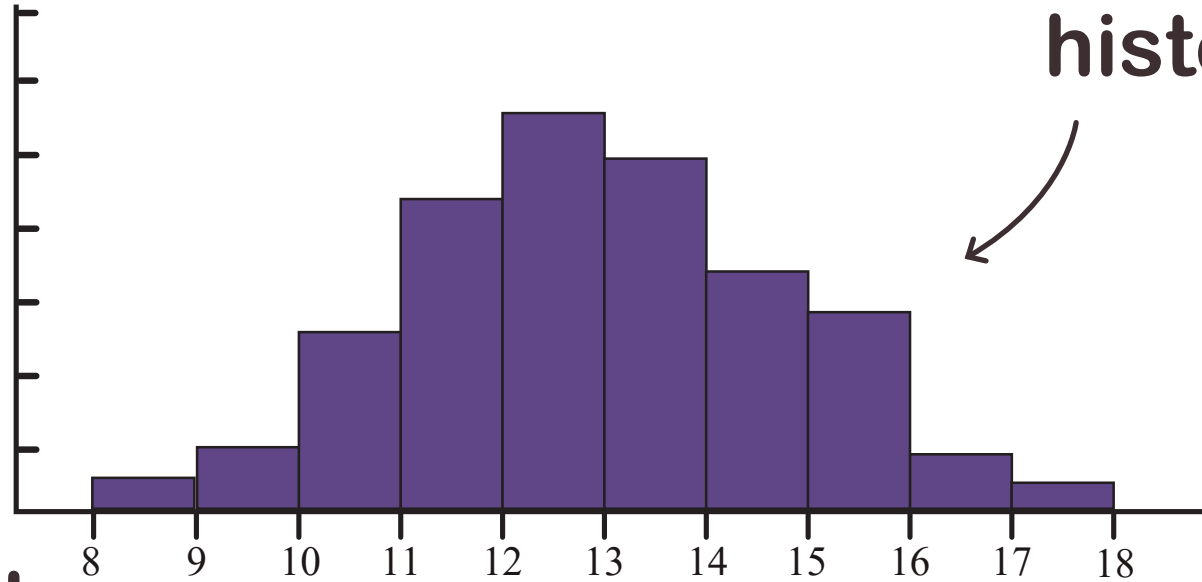


Minimum

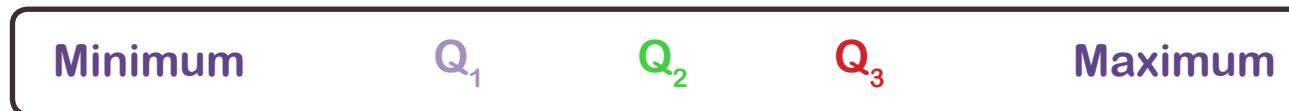
 $Q_1$  $Q_2$  $Q_3$ 

Maximum

المدرج التكراري  
histogram



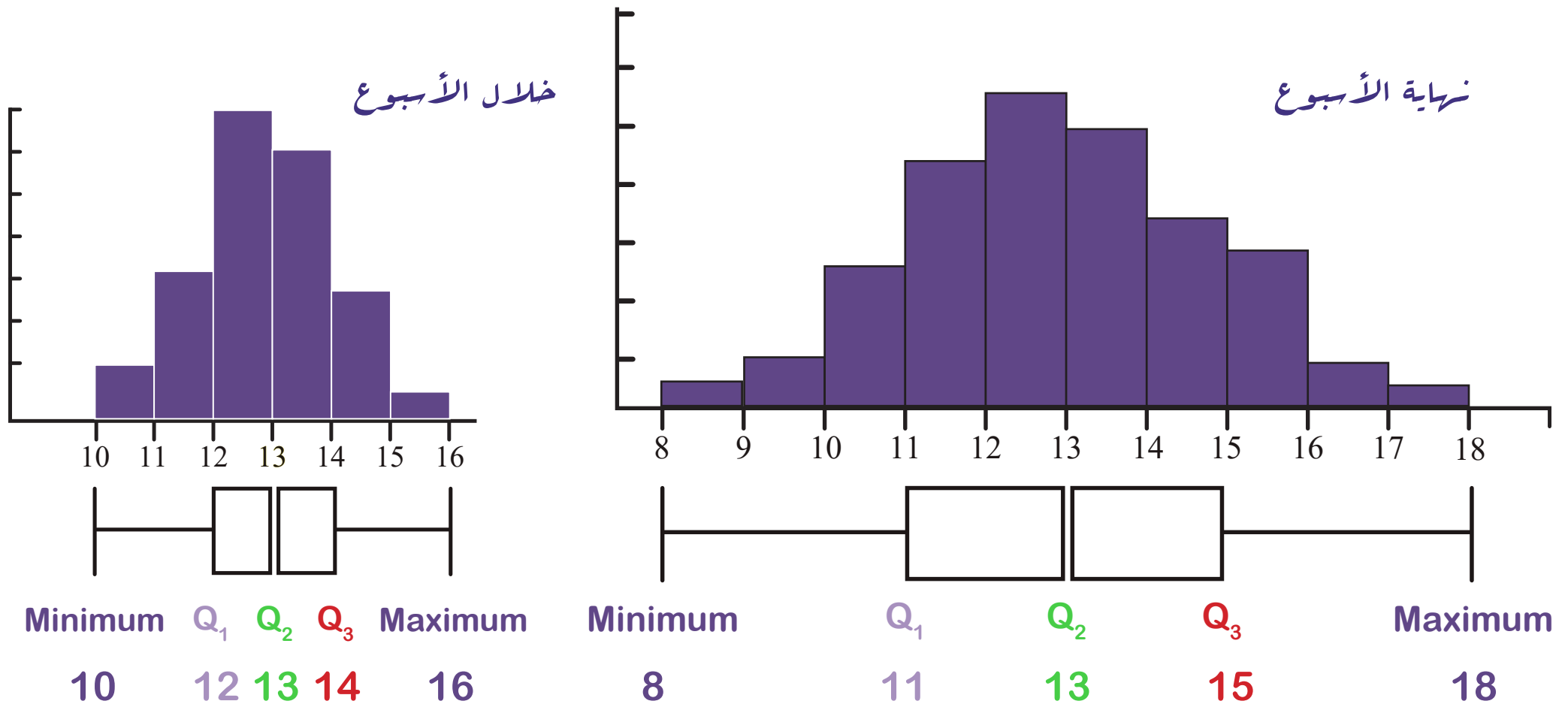
Box plot



الملخص ذو الخمسة أرقام

The 5 Number Summary

يساعدنا كل من المدرج التكرار Histogram والملخص ذو الخمسة أرقام the 5 Number Summary وال Box plot في معرفة شكل البيانات ومدى انتشار وتشتت القيم عن بعضها البعض

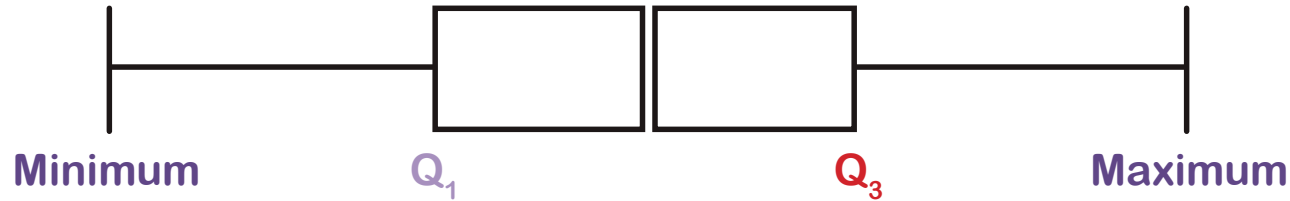


مقدار التشتت أو الفروق بين القيم لعدد الكلاب المشاهدة في عطلات نهاية الأسبوع أكبر منه بين عدد الكلاب المشاهدة خلال أيام الأسبوع الأخرى.

ن نهاية الأسبوع

المدى الإربعاعى

كل من المدى  
والمدى



الإربعاعى

المدى

لمشاهدات نهاية

الأسبوع أكبر

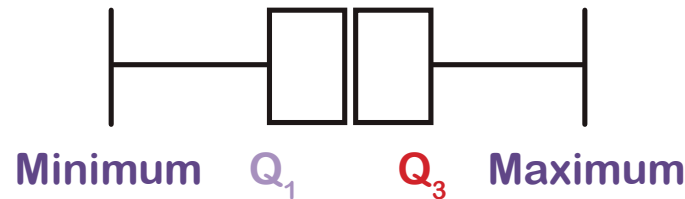
من مشاهدات

بقية أيام

الأسبوع

خلال الأسبوع

المدى الإربعاعى



المدى

في ا لدرس القادم سنرى كيف يمكننا  
التعبير عن التشتت Spread من خلال  
قيمة واحدة فقط.

محمد لطفي

# مقدمة إلى الانحراف المعياري

Introduction to Standard Deviation

Mohammed Lotfy  
@mohammud.lotfy

الطريقة الأكثر شيوعًا التي يستخدمها المحترفون لقياس انتشار **spread** مجموعة

من البيانات باستخدام قيمة واحدة هي استخدام الانحراف المعياري **Standard**

**Deviation** أو التباين **Variance**.

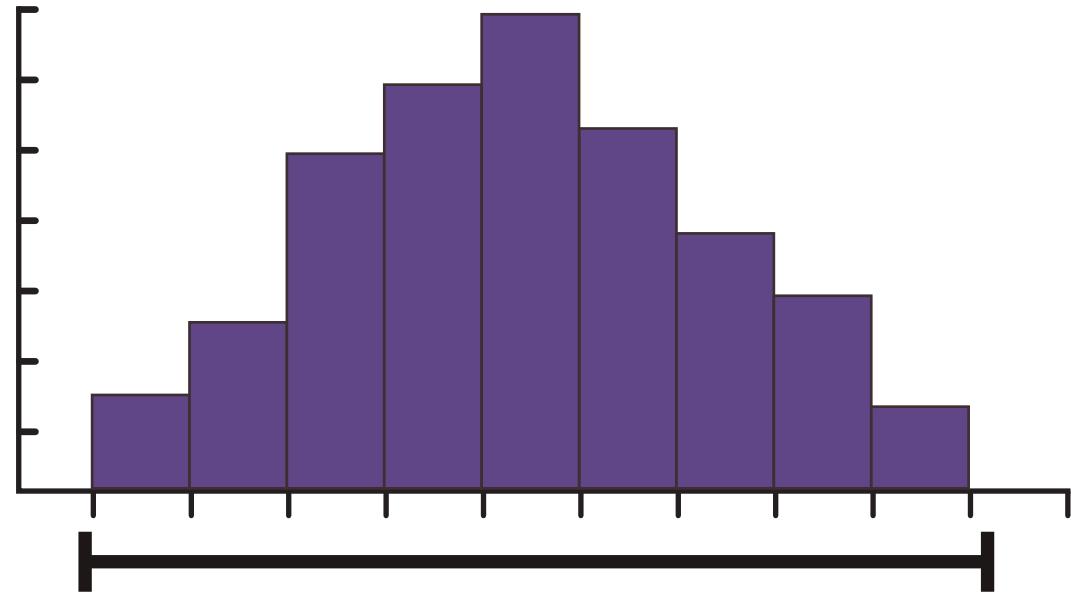
هنا، سنركز على الانحراف

المعياري **Standard**

**Deviation**، لكننا أثناء

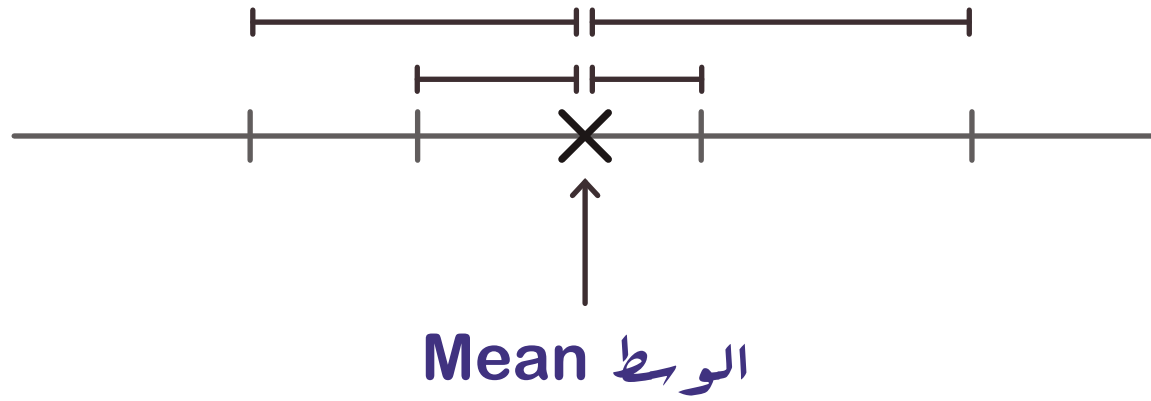
حسابه سنتعلم كيفية

حساب التباين **Variance**.



يمكننا باستخدام الانحراف المعياري  
بقيمة واحدة معرفة مدى تشتت مجموعة  
من البيانات





يُخبرنا الانحراف المعياري عن

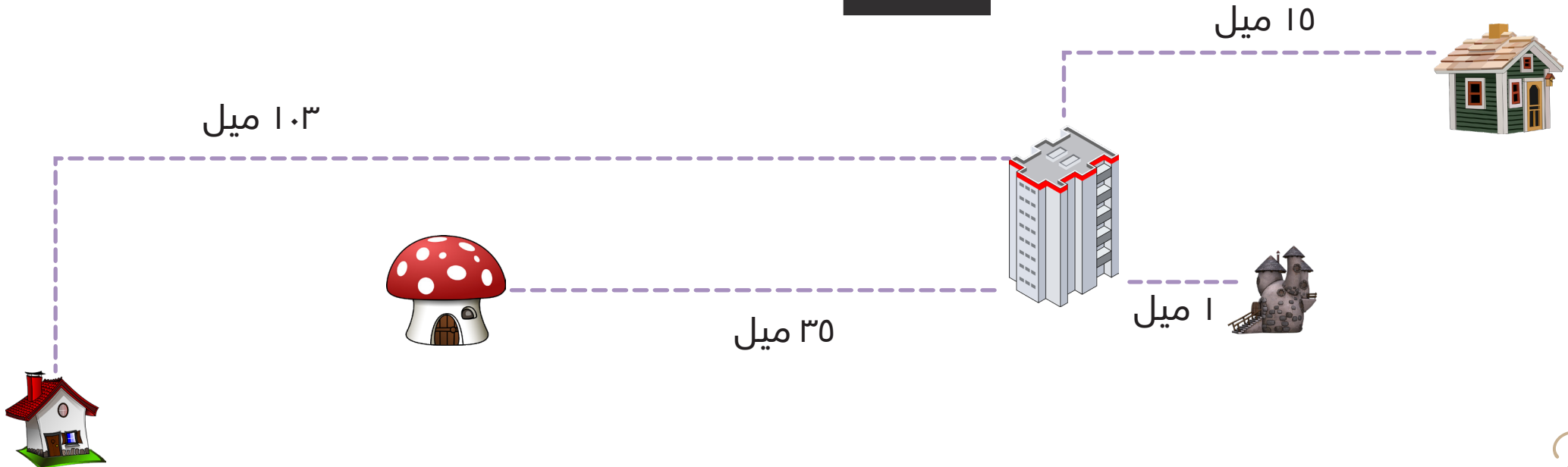
مدى ابتعاد كل قيمة في

مجموعة البيانات عن متوسط

Mean هذه البيانات.

خيلوا أننا نريد أن نعرف إلى أي مدى يُبعد بيت مجموعة من الموظفين عن مكان

عملهم:



يمكننا تجميع Aggregate كل هذه المسافات معًا لإيضاح أن متوسط Mean

المسافة التي يبعدها موقع الموظفين عن عملهم يساوي 38.5 ميلًا.

الوسط  
Mean

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

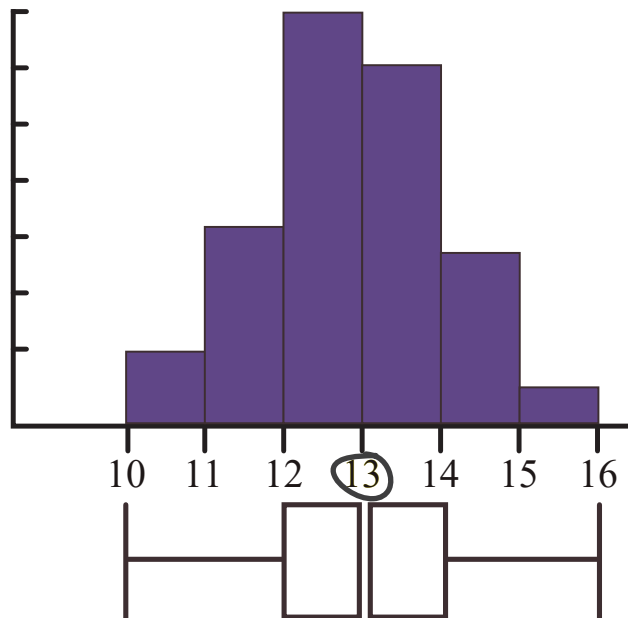
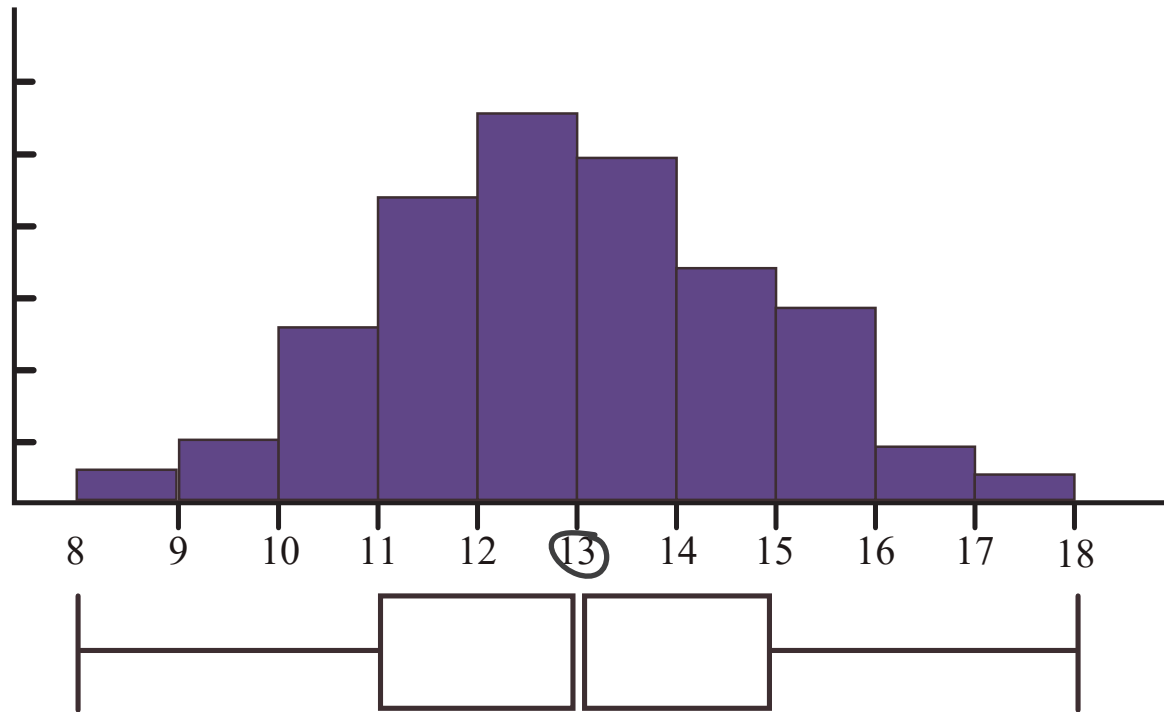
مجموع القيم

عدد القيم

أى أنه يمكننا تلخيص  
جميع المسافات من  
المنزل إلى العمل  
في قيمة واحدة تمثل  
وسط mean كل هذه  
المسافات

$$\bar{x} = \frac{15 + 35 + 1 + 103}{4} = 38.5$$

من الممكن أن تشتركا مجموعتي  
بيانات في قيمة المتوسط (راجع  
درس Box Plot)، ولكنهما  
تختلفان في الانتشار spread،  
سواء تم قياسه عن طريق المدى  
range أو المدى الإرباعي  
interquartile range أو  
الانحراف المعياري



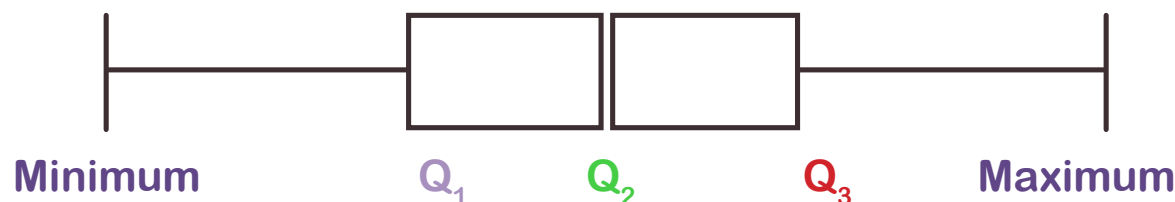
لكل من مجموعتي  
البيانات نفس  
المتوسط:

الوسط = الوسيط =  
المنوال = 13

ولكنها يختلفان في  
تشتت وانتشار البيانات  
عن هذا المتوسط كما  
يوضحه الـ Box Plot

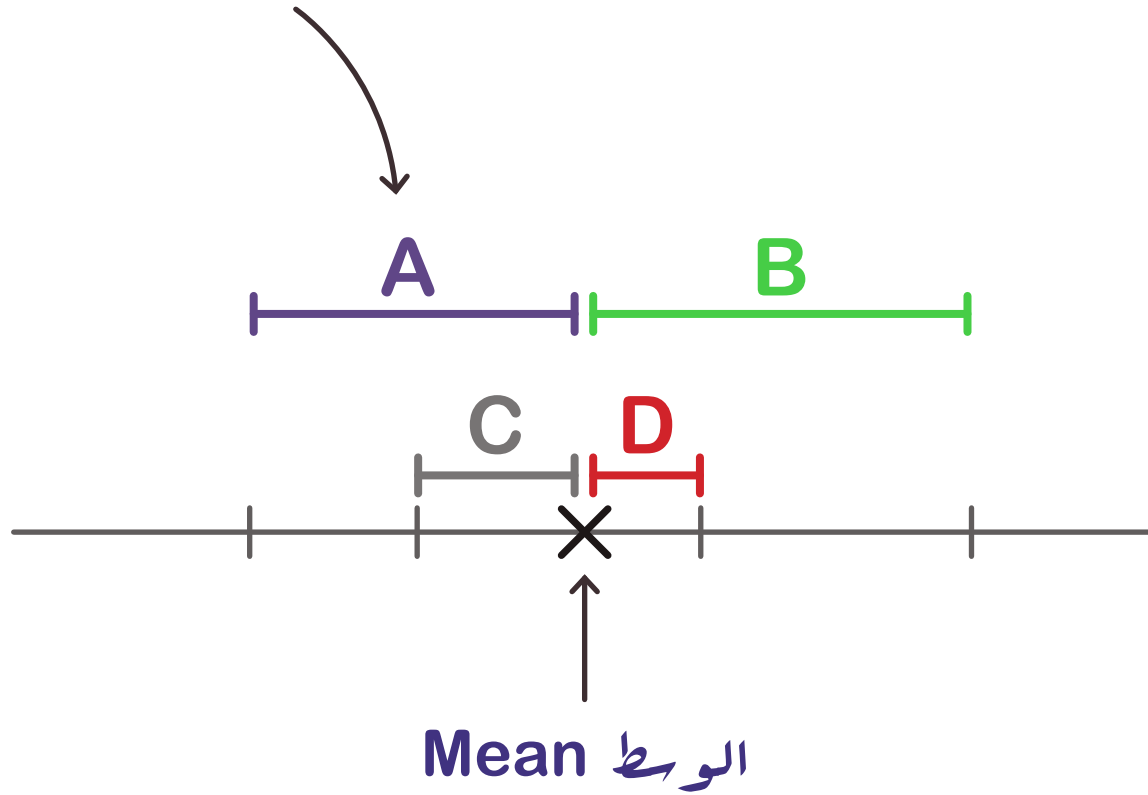
الآن نريد أن نعرف كيف تختلف vary المسافة بين المنزل والعمل من موظف إلى آخر.

يمكننا استخدام الملخص ذو الخمسة  
أرقام كوصف لهذه البيانات



ولكن إذا أردنا قيمة واحدة فقط للتعبير عن الانتشار Spread فربما نختار الانحراف المعياري **Standard Deviation**.

تعبّر الحروف عن المسافة بين كل قيمة والوسط  
mean



الانحراف المعياري standard

deviation هو عبارة عن

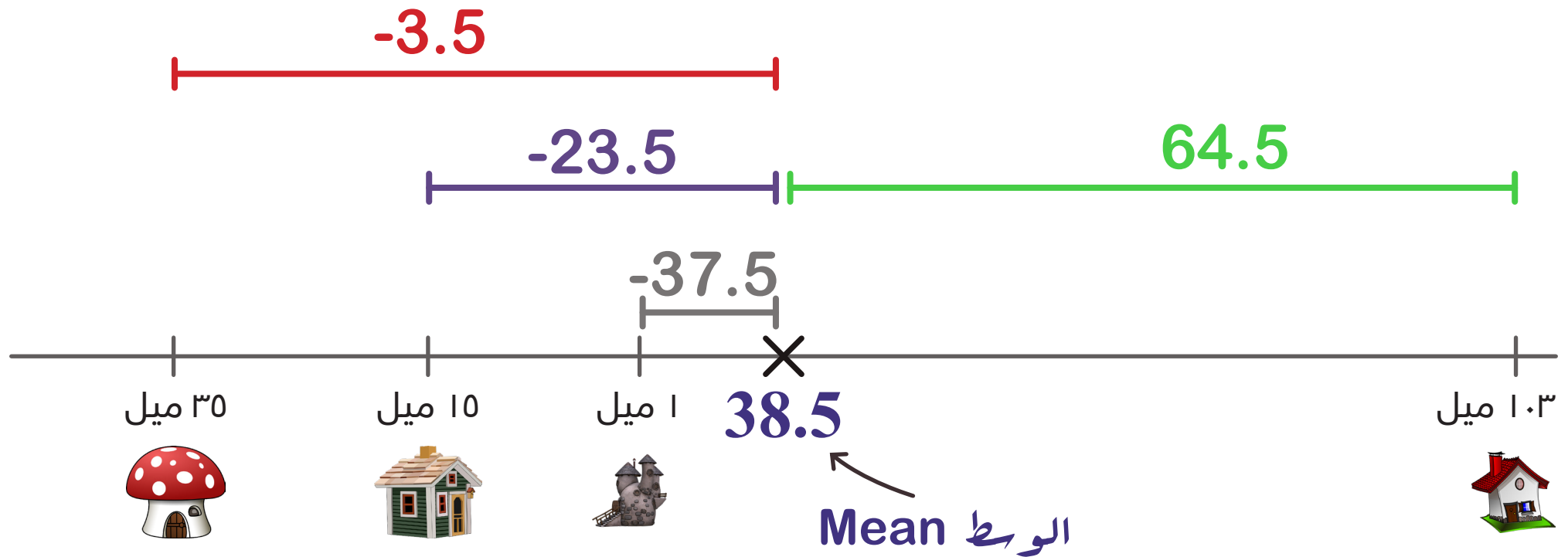
متوسط اختلاف كل قيمة عن

الوسط Mean الخاص بهذه

القيم.

الانحراف المعياري يعبر عن متوسط القيم A ، B ، C ، و D

بالنسبة إلى هذا المثال يشير ذلك إلى [متوسط] اختلاف المسافة التي يبعدها كل شخص عن العمل عن متوسط Mean جميع المسافات.



لذا، يصبح مثل متوسط average كل تلك المسافات بين منازل الموظفين وعملهم.

قم بحساب متوسط الفروق  
بين كل منزل وموقع العمل،  
ماذا تلاحظ؟



# حساب الانحراف المعياري

Standard Deviation Calculation

Mohammed Lotfy  
@mohammud.lotfy



حساب الانحراف المعياري

إن الحساب اليدوي

للالانحراف المعياري (أو

لغيره من المقاييس)

يمنحك **حدسًا** (أو إحساسًا)

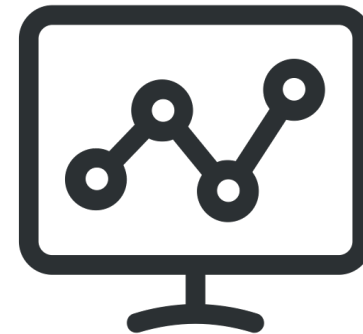
لما تقوم به.

هذا الحدس ضروري

لفهم البيانات بشكل جيد،

ولاختيار التحليل المناسب

لكل حالة أو موقف.



تخيلوا أننا لدينا مجموعة بيانات لها أربع قيم هي:

10 14 10 6

نحسب الوسط mean

الوسط  
Mean

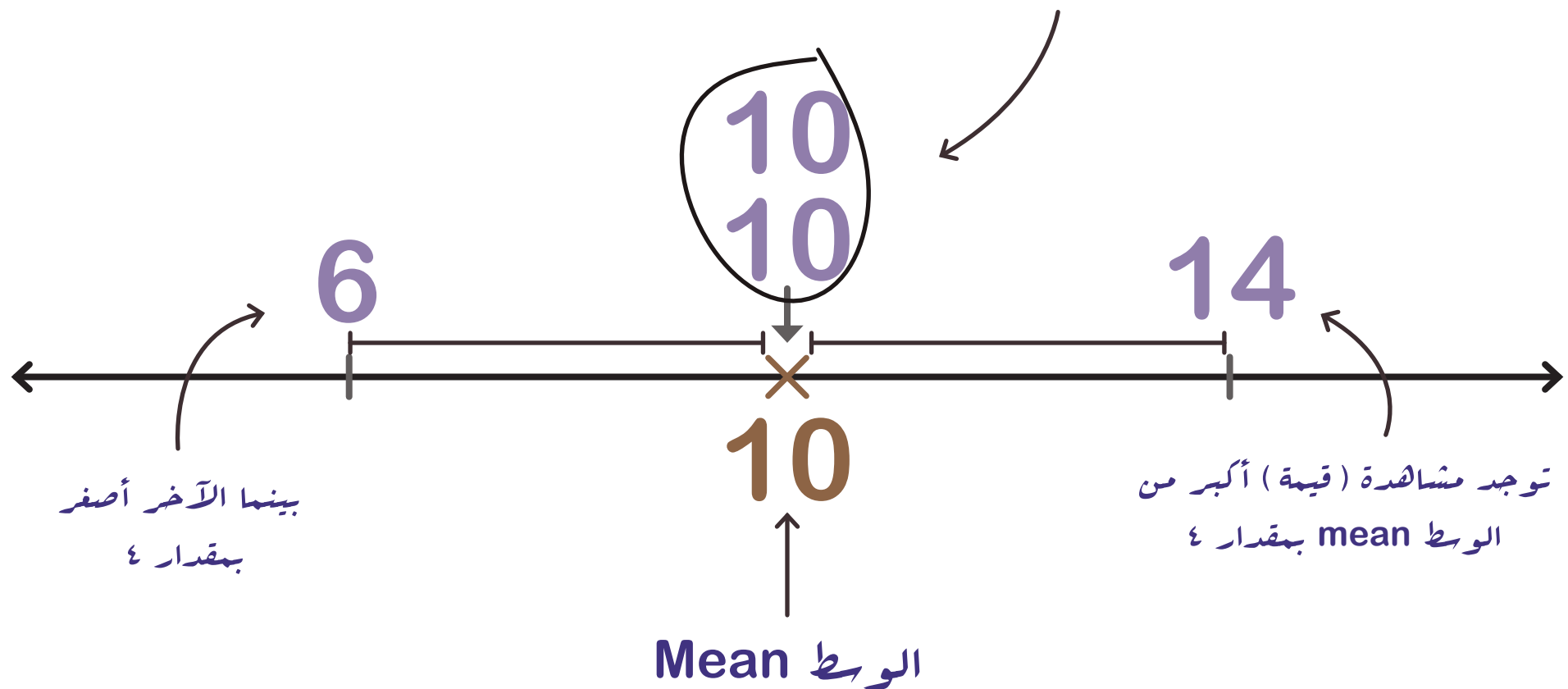
$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{40}{4} = 10$$

مجموع القيم

عدد القيم

نريد بعد ذلك النظر إلى مسافة كل مشاهدة [أو قيمة] من هذا الوسط.

توجد مشاهدتان two observations (أو قيمتان)  
يساويان الوسط mean بالضبط، لذا المسافة بينهما وبين  
الوسط mean تساوي صفرًا



## نحسب الفرق بين كل قيمة والوسط mean

2

يتم التعبير عن الفروق بين كل مشاهدة  
observation والوسط mean بالترميز:

$$x_i - \bar{x}$$

$$10 - 10 = 0$$

$$14 - 10 = 4$$

$$10 - 10 = 0$$

$$6 - 10 = -4$$

عند حساب الوسط mean للفروق بين كل قيمة والوسط الخاص بهذه القيم، فإن الفروق الموجبة ستكون دائما، ولأي مجموعة بيانات، مساوية للفروق السالبة

$$4 = -4$$

مما سيجعل دائما مجموع هذه الفروق يساوي صفرا

$$0 + 4 + 0 + -4 = 0$$

وبالتالي فإن قيمة وسط mean هذه الفروق يساوي صفرا هو الآخر

$$\bar{x} = \frac{0}{4} = 0$$

وبالطبع فإن قيمة صفر لا تعد  
قياسا مفيدا للتشتت أو الانتشار  
.spread

حيث سيشير الصفر إلى أن كل القيم متساوية أو إلى أنه لا يوجد انتشار spread.

بدلاً من ذلك، نحتاج إلى جعل كل هذه القيم موجبة.

### 3 نربّع الفروق بين القيم والوسط mean

الطريقة التي

تمكننا من القيام

بذلك أثناء حساب

الانحراف المعياري

هي تربيع كل القيم.

$$\begin{array}{rclclcl}
 x_i & - & \bar{x} & & & \\
 10 & - & 10 & = & 0^2 & = & 0 \\
 14 & - & 10 & = & 4^2 & = & 16 \\
 10 & - & 10 & = & 0^2 & = & 0 \\
 6 & - & 10 & = & -4^2 & = & 16
 \end{array}$$



إذا قمنا بذلك هنا فستصبح القيم السالبة لدينا تساوي 16.



الآن يمكننا أن نحسب متوسط **mean** هذه القيم.

4 نحسب متوسط الفروق بين القيم والوسط **mean**

$$\frac{0 + 16 + 0 + 16}{4} = 8$$

ما قمنا بحسابه الآن هو متوسط مربع  
المسافة بين كل قيمة وبين الوسط **mean**

ويسمى بالتباين **variance**

يمكن التعبير عن التباين variance بالمعادلة التالية:

التباين

مجموع مربع الفرق بين القيم والوسط mean

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

عدد القيم

هذه الصيغة تعبر عن  
القسمة على عدد القيم n

## نحسب الجذر التربيعي للتباين Variance

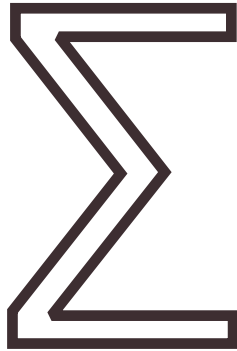
5

$$\sigma = \sqrt{8} = 2.83$$

ويرمز له بالحرف الصغير  
sigma

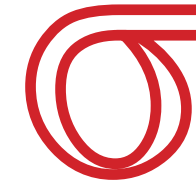
وهذا هو الانحراف المعياري  
**Standard Deviation**

الانحراف المعياري standard deviation هو متوسط مقدار المسافة التي تبعتها كل نقطة في مجموعة البيانات لدينا عن الوسط mean.



الحرف الكبير upper case سيجمما

يعبر عن الجمع Sum



الحرف الصغير lower case

سيجمما يعبر عن الانحراف المعياري

Standard Deviation

$$\sigma = \sqrt{\text{variaance}}$$

ما قمنا به لحساب الانحراف المعياري هو التالي:

## Standard Deviation (SD) Calculation:

حساب متوسط القيم

1

حساب الفرق بين كل قيمة ومتوسط القيم

2

تربيع هذا الفرق

3

حساب متوسط هذه الفروق

4

حساب الجذر التربيعي لهذا المتوسط

5

# لماذا الانحراف المعياري؟

Why Standard Deviation?

Mohammed Lotfy

@mohammud.lotfy

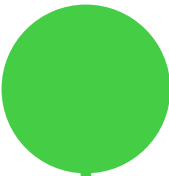
الانحراف المعياري يُستخدم للوصول إلى عدد واحد لمقارنة انتشار مجموعتي بيانات.

من الجيد أن تكون قادرًا على الحكم على مدى انتشار البيانات من مجموعة إلى أخرى دون الحاجة إلى استخدام جدول بكامل القيم التي لدينا.

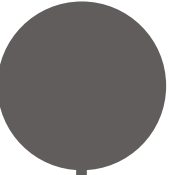
وهكذا فإن لدينا طريقة لتحديد أي مجموعة بيانات هي الأكثر انتشارًا.

يمكننا فقط مقارنة الانحراف المعياري standard deviation لمجموعة ما بالانحراف المعياري لمجموعة أخرى.





إنَّ وجود عدد واحد يُبسِّط كمية المعلومات التي يحتاج الشخص الذي تقدم إليه التقرير إلى فهمها.



إنَّ وجود قيمة واحدة أيضًا له ميزة أخرى من حيث ما يُعرف باسم الإحصائيات الاستنتاجية. لكن هذا خارج موضوعنا الحالي.

# نقاط أخيرة هامة

Important Final Points

Mohammed Lotfy

@mohammud.lotfy

## أولاً

يستخدم الانحراف المعياري للمقارنة بين انتشار spread المجموعات المختلفة لتحديد أيها أكثر انتشاراً.

## ثانياً

فعند المقارنة بين أسعار الأسهم، يعتبر سعر السهم الذي يتغير بانحراف معياري كبير بمرور الوقت أكثر خطورة من سعر السهم الذي يتغير بانحراف معياري أقل.

عندما تتعلق البيانات بالمال والاقتصاد فإن ارتفاع الانحراف المعياري مرتبط بارتفاع المخاطر.

## ثالثا

لتكون المقارنة عادلة، يجب أن تكون البيانات جميعها بوحدات القياس نفسها.

إذا كنت تقيس بالدولار في مجموعة بيانات، وباليورو في غيرها، فمن غير العادل مقارنة مجموعتي البيانات لتحديد ما له انتشار أكبر.

## أخيرا

يوجد للتباين وحدات قياس مربعة للقياسات الأصلية.

إذا كنت تقيس العائد بالدولار، سيتضمن التباين وحدات الدولار مربع (دولار<sup>2</sup>)، وهي غير مفيدة بشكل عملي.

لهذا السبب، غالبًا ما يعتبر الانحراف المعياري، وهو الجذر التربيعي للتباين، قياسًا أكثر فائدة لقياس الانتشار؛ حيث إنه يتعامل بنفس وحدات مجموعة البيانات الأصلية.

إذا كنت تقيس العائد بالدولار فإن الانحراف المعياري أيضًا يعبر عنه بوحدات من الدولار.

محمد لطفي

# أشكال التوزيعات

## Shapes of Distributions

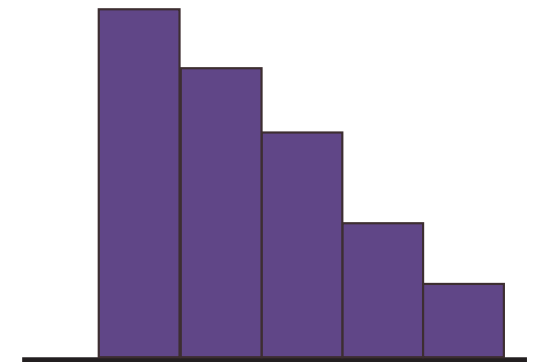
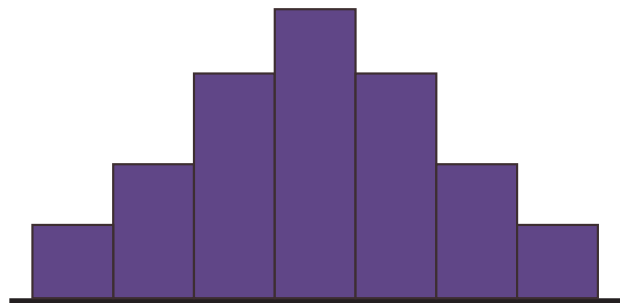
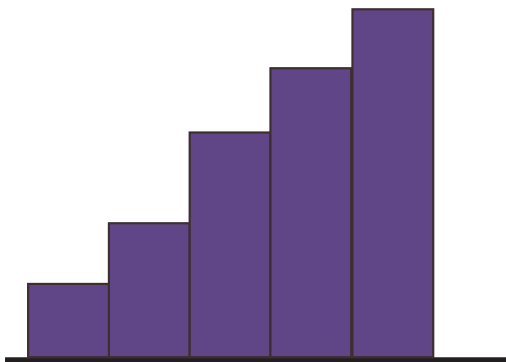
Mohammed Lotfy  
[@mohammud.lotfy](https://twitter.com/mohammud.lotfy)

بعد أن ناقشنا كيفية إنشاء مدرج تكراري

histogram، يمكننا استخدام هذا لتحديد

الشكل **shape** المعبر عن البيانات

لدينا هنا ثلاثة مدرجات تكرارية تظهر شكل ثلاث مجموعات من البيانات

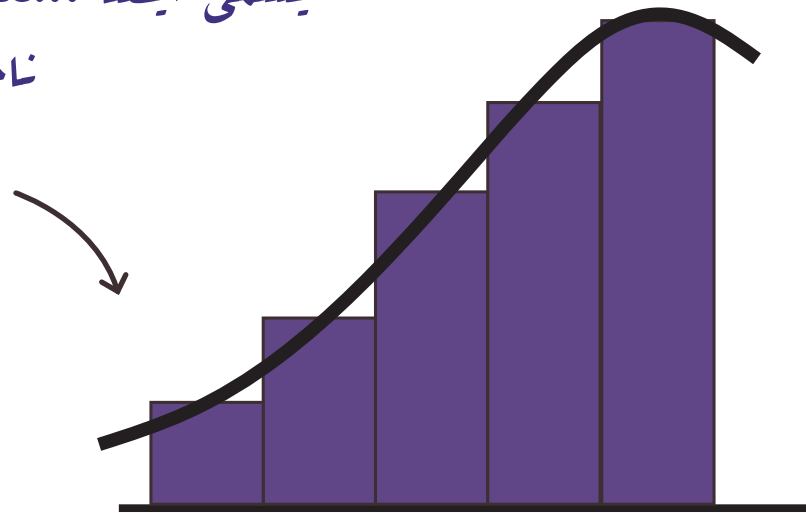


يعتبر المدرج التكراري الذي به فئات bins أقصر على اليسار وفئات أطول على

اليمين شكلاً منحرفاً إلى اليسار [توزيع ملتوي التواء سالبا أو الملتوي إلى اليسار] left

skewed shape

يسمى أيضا left tail لوجود ذيل tail  
ناحية اليسار



وقلة لها قيم صغيرة

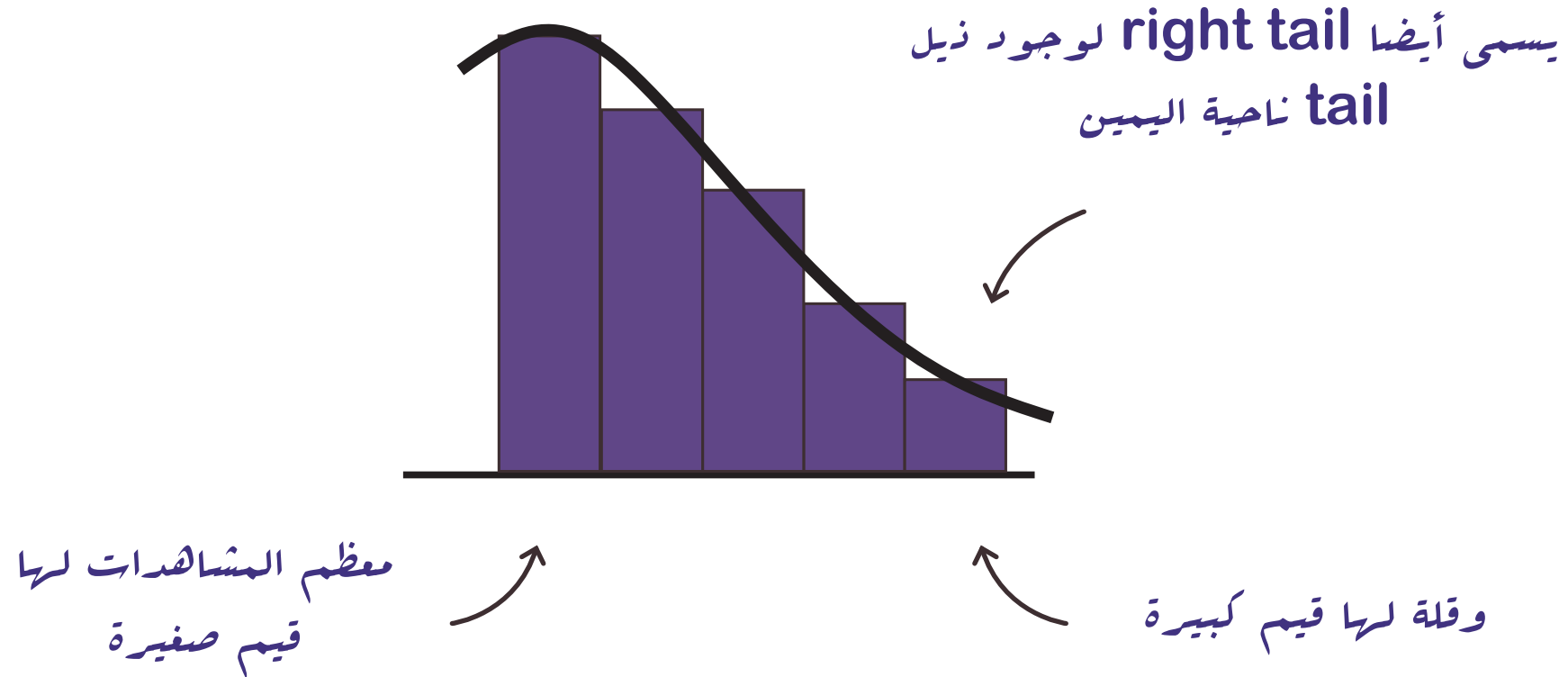
معظم المشاهدات لها  
قيم كبيرة



المدرج التكراري الذي به فئات bins أقصر على اليمين وفئات أطول على اليسار يعتبر

شكلًا منحرفًا إلى اليمين [توزيع ملتوي التواء موجباً أو الملتوي إلى اليمين] right

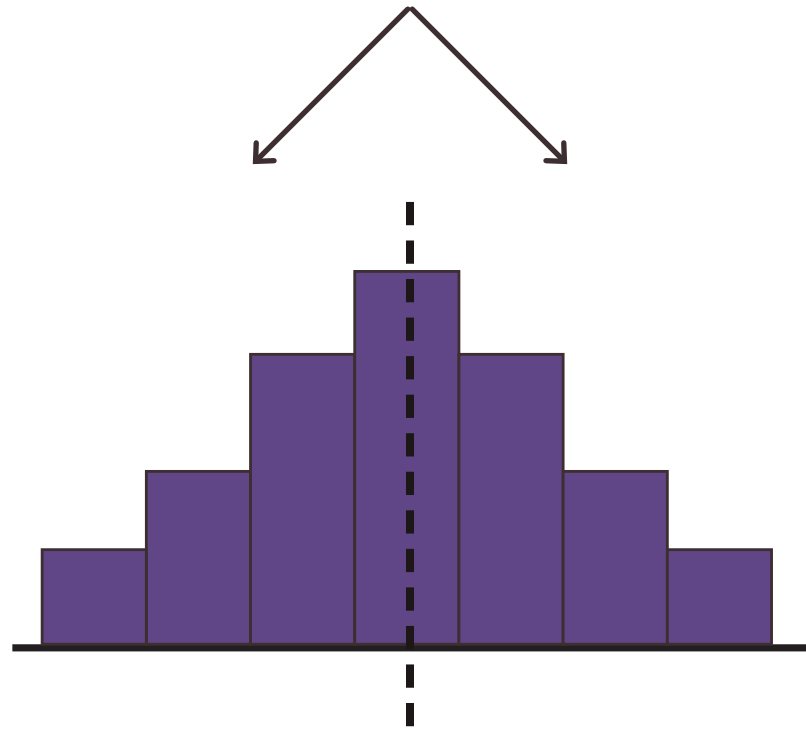
skewed shape



أي توزيع distribution يمكنك فيه رسم خط في المنتصف، ويكون فيه الجانب

الأيمن مطابقًا للجانب الأيسر يعتبر متناظرًا [متماثلًا] symmetric

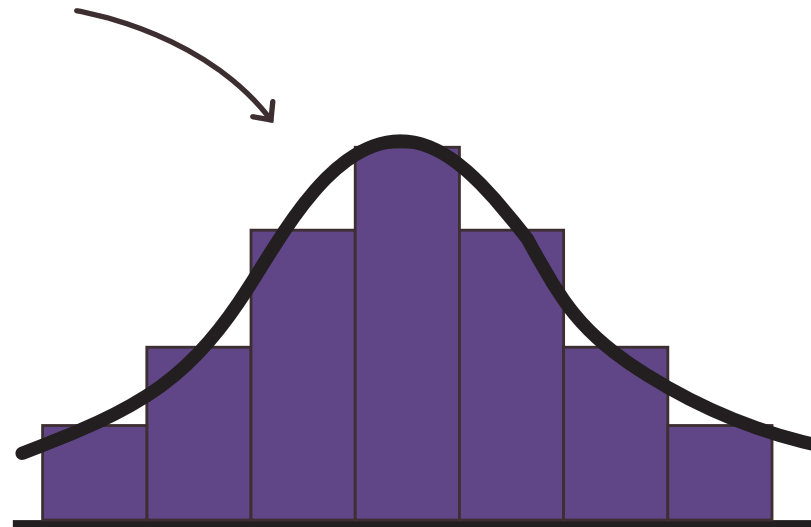
كل نصف صورة متماثلة مع النصف الآخر



يُعرف أحد أشهر التوزيعات المتناظرة بالتوزيع الطبيعي normal distribution

ويُسمى أيضًا منحنى ناقوسي bell curve [منحنى على شكل جرس].

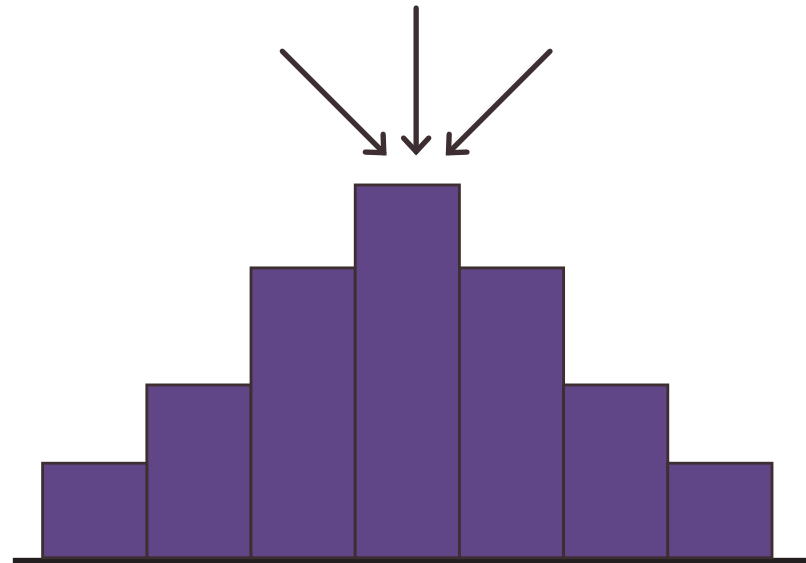
يعبر عن التوزيع الطبيعي normal  
distribution منحنى يأخذ شكل جرس bell



يمكن أن يوضح لنا شكل التوزيع shape of distribution الكثير  
عن مقاييس المركز measures of center والانتشار spread

في التوزيعات المتناظرة المتوسط mean = الوسيط median = الوضع mode [المنوال]

كلٌّ من هذه المقاييس توجد في المركز center



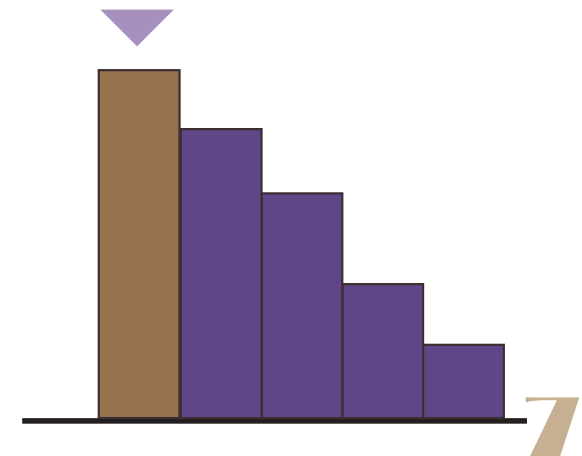
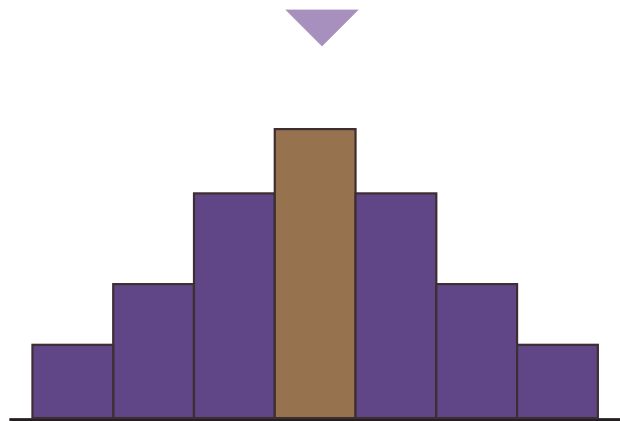
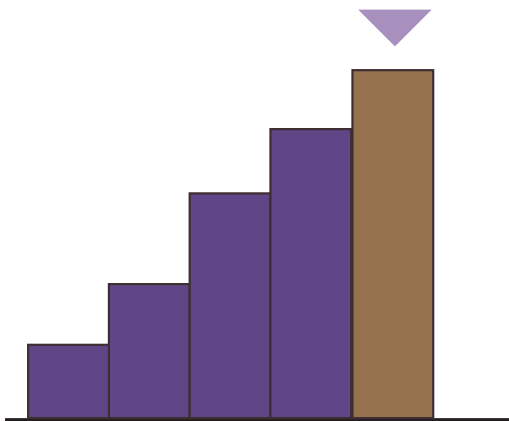
الوضع mode [المنوال] دائما

هو أطول شريط bar في

المدرج التكراري histogram.

# المنوال

# The Mode

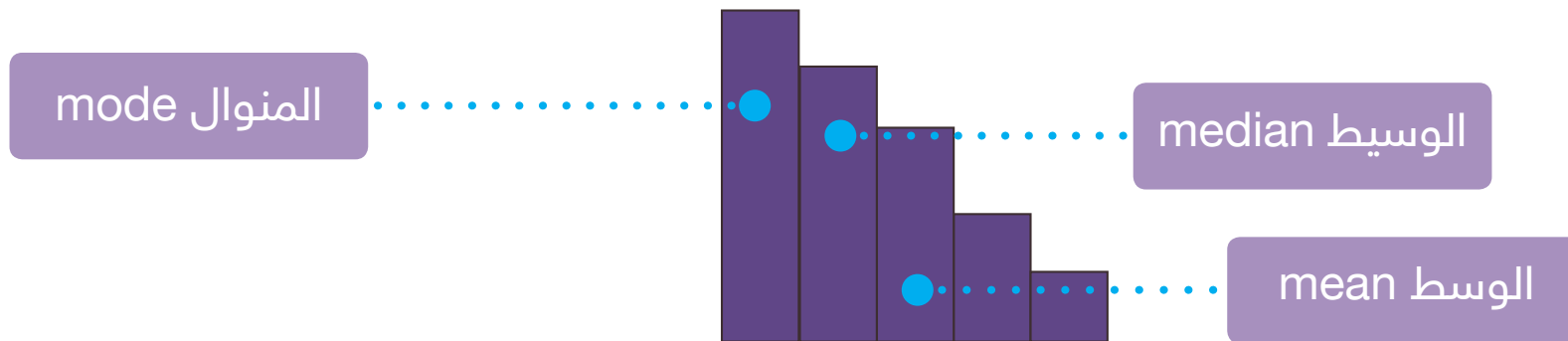


عندما يكون لدينا توزيعات منحرفة skewed distributions [توزيعات

ملتوية] تتم إزاحة الوسط mean ناحية ذيل tail التوزيع distribution

بينما يظل الوسيط median قريبًا من الوضع mode [المنوال]

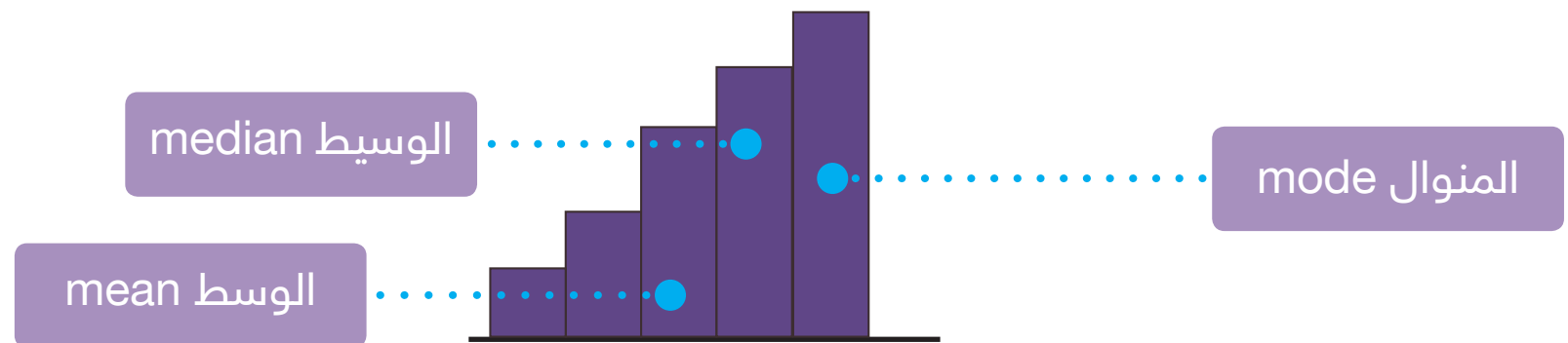
على سبيل المثال، في هذا التوزيع المنحرف إلى اليمين right skewed distribution [ملتوي التواء موجب] ستتم إزاحة المتوسط الحسابي mean إلى أعلى [نحو القيمة الأكبر]



وينتج عن ذلك  
متوسط mean أكبر  
من الوسيط median

في التوزيع المنحرف إلى اليسار right skewed distribution  
[ملتوي التواء سالبا] يتم سحب المتوسط mean إلى أسفل [نحو  
القيمة الأقل]

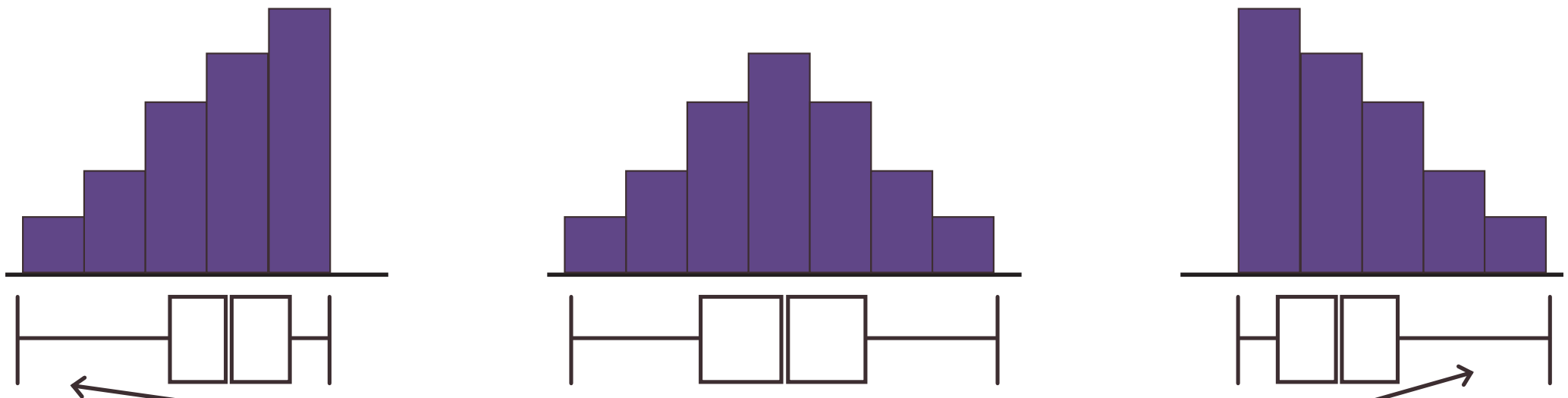
وينتج عن ذلك  
متوسط mean أقل  
من الوسيط median



لربط هذا بالتمثيل المرئي للمدرج التكراري histogram فلنعد إلى

الملخص ذي الخمسة أعداد 5 number summary

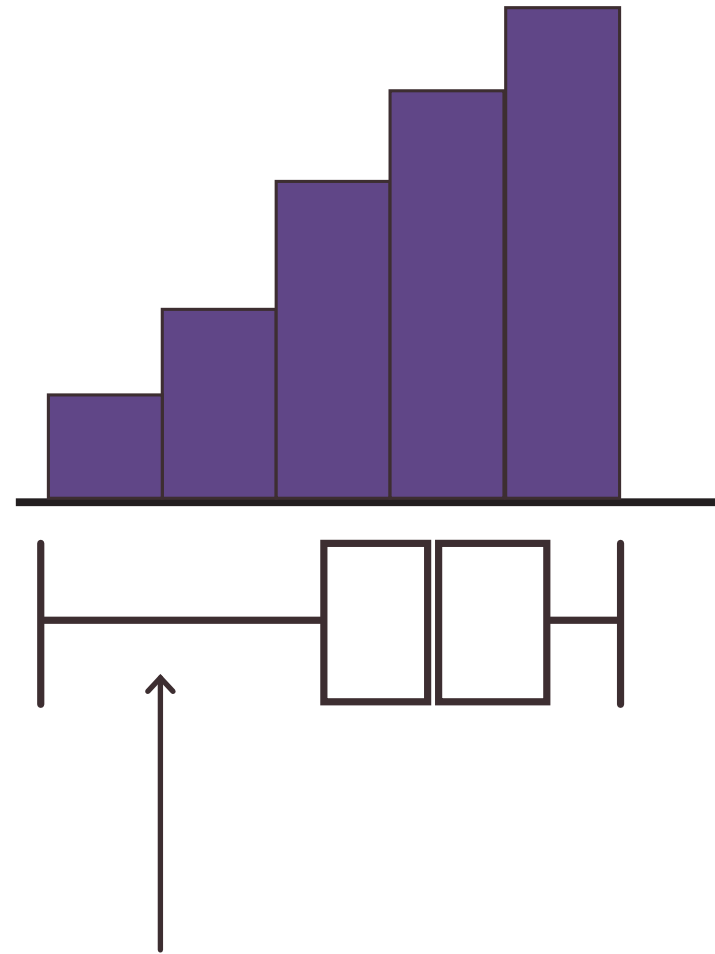
يوجد هنا الـ box plots المعبرة عن كل مدرج تكراري



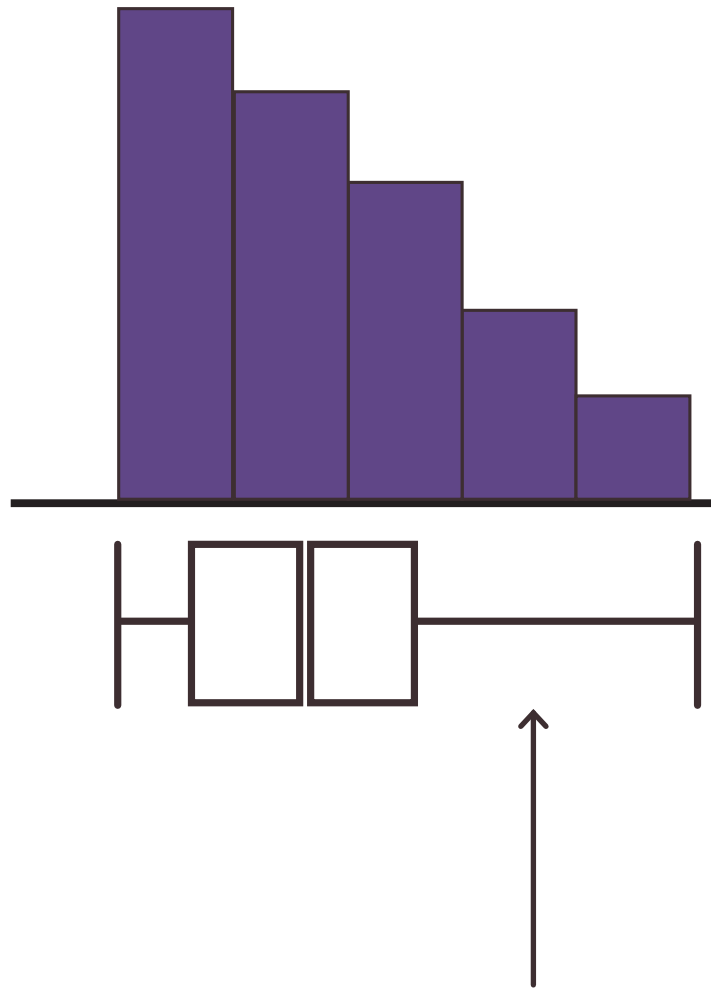
لاحظ كيف تتمدد الزوائد whiskers [الأطراف] في اتجاه الانحراف skew لكل توزيع

من التوزيعات المنحرفة [الملتوية] skewed distributions





أي أن الزوائد [الأطراف] الأطول تكون ناحية اليسار للتوزيع المنحرف  
إلى اليسار [الملتوي التواء سالباً] left skewed distribution

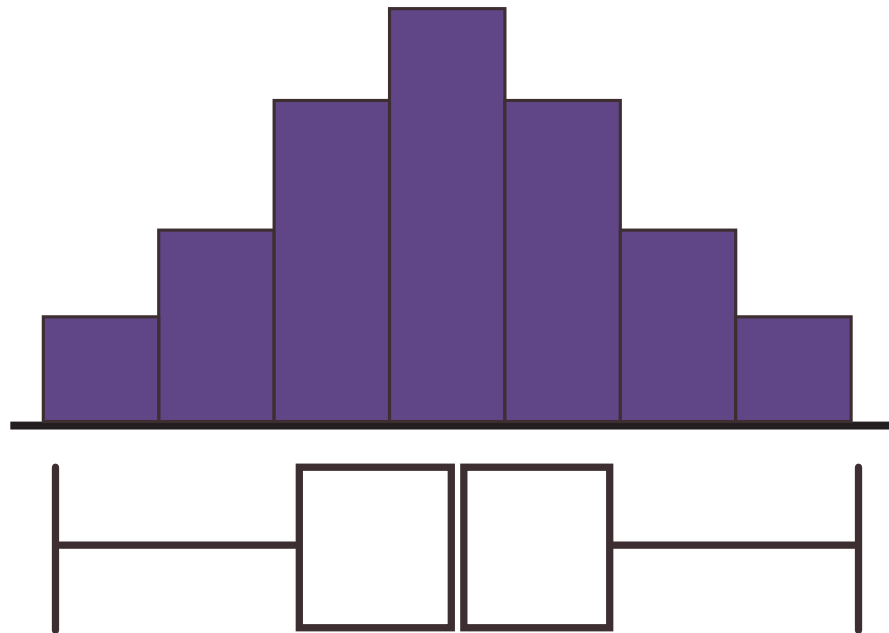


وتكون ناحية اليمين للتوزيع المنحرف إلى اليمين [الملتوي التواء موجبا]

right skewed distribution

ولكن المدرج التكراري المتناظر symmetric histogram يعبر عنه أيضا box plot

متماثل أو متناظر.



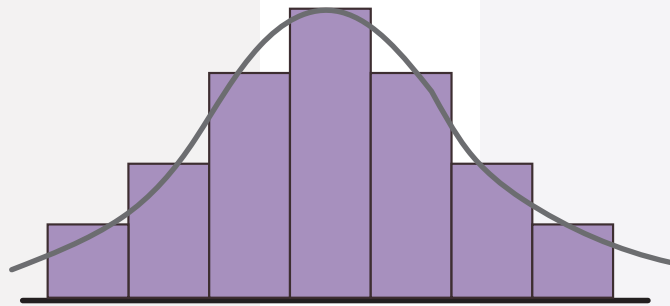
# أشكال لبيانات واقعية

The Shape for Data in the World

Mohammed Lotfy  
[@mohammud.lotfy](https://twitter.com/mohammud.lotfy)

إذا كنت تتعامل مع البيانات، فيمكنك دائمًا إنشاء مخطط رسومي plot سريع لرؤية الشكل shape وفهم البيانات بشكل أفضل.

تشتمل بعض أمثلة البيانات التي تأخذ تقريبًا شكل جرس bell curve [التوزيع الاعتيادي normal distribution] على:



- ◀ الارتفاعات والأوزان
- ◀ درجات الاختبارات القياسية
- ◀ كميات هطول الأمطار
- ◀ وسط توزيع ما mean of a distribution
- ◀ أخطاء في عمليات التصنيع

التوزيعات الاعترالية normal distributions تتميز بوجود معظم القيم حول المتوسط، وقلة من القيم تكون كبيرة جدا أو صغيرة جدا

درجات الاختبارات القياسية



معظم درجات الأفراد في اختبارات، مثل اختبار الذكاء، تترواح حول المتوسط، والقلة من يتميزون بذكاء عالٍ جدا أو منخفض جدا

الارتفاعات والأوزان



معظم أطوال وأوزان الأشخاص مثلا تترواح حول المتوسط، والقلة من يتميزون بطول أو وزن عالٍ جدا أو منخفض جدا

وسط توزيع ما mean of a distribution  $\bar{x}$

كميات هطول الأمطار



معظم كميات الأمطار تترواح حول المتوسط، والقلة تتسم بالشدة حد السيول الجارفة، أو بالندرة كما في الصحراء

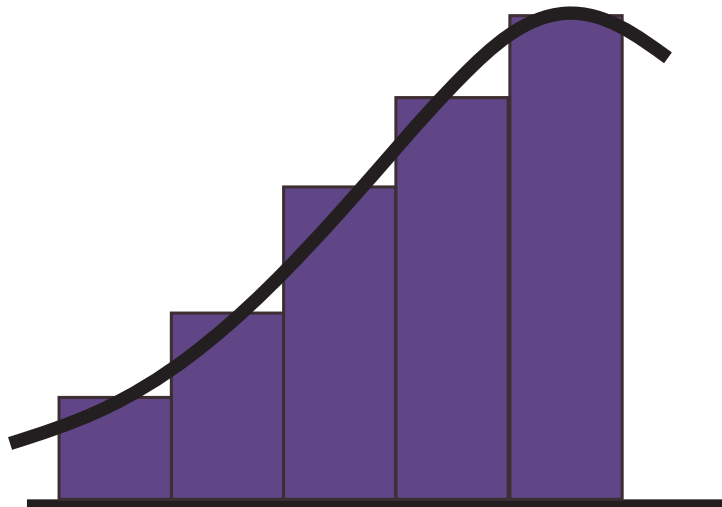
أخطاء في عمليات التصنيع



معظم أخطاء التصنيع تترواح حول المتوسط، والقلة تكون بلا أخطاء أو بها أخطاء كثيرة

تشتمل بعض البيانات التي تتبع توزيعات منحرفة إلى اليسار **left skewed data**

[الملتوية التواء سالبا] على:



◀ تقديرات الاختبارات GPA

◀ عمر الوفاة

◀ تغيرات أسعار الأصول

التوزيعات الملتوية التواء سالبا left skewed distributions تتميز بأن معظم القيم أكبر من المتوسط، وقلة من القيم تكون صغيرة جدا

عمر الوفاة



معظم الأشخاص يموتون في سن كبيرة، وقلة  
لهم من يموتون في عمر مبكر

تقديرات الاختبارات GPA



معظم تقديرات الطلاب بشكل عام في  
أى اختبار تكون أكبر من درجة النجاح  
(المتوسط)، وقلة لهم من يرسبون

تغيرات أسعار الأصول

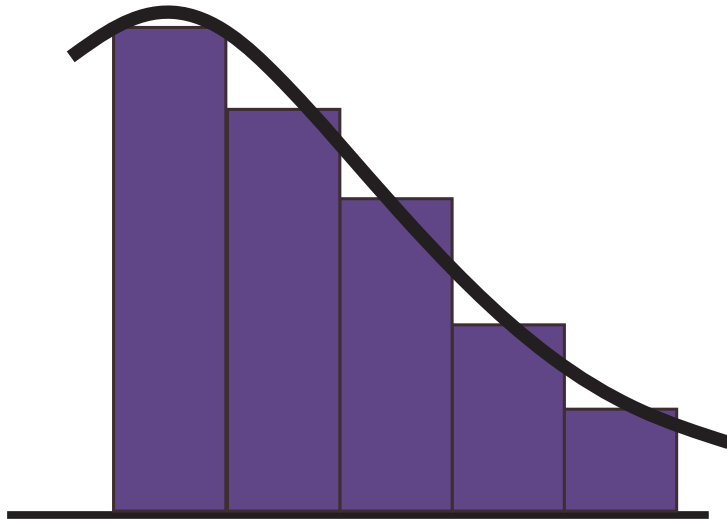


غالباً ما يكون تغير أسعار أصول الشركة، مثل  
السندات والأسهم والمباني والأراضي، في  
ارتفاع كبير



تشتمل بعض البيانات التي تتبع توزيعات منحرفة إلى اليمين [ملتوية التواء موجبا]

right skewed distribution على:



◀ كمية الدواء المتبقي في مجرى الدم  
مع مرور الوقت

◀ توزيع الثروة

◀ القدرات الرياضية لدى الأشخاص

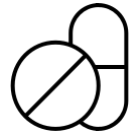
التوزيعات الملتوية التواء موجبا right skewed distributions تتميز بأن معظم القيم أقل من المتوسط، وقلة من القيم تكون كبيرة جدا

توزيع الثروة



معظم الأشخاص لا يملكون ثروات طائلة،  
والقليلون لهم من يتمتعون بالثراء الهائل

كمية الدواء المتبقي في  
مجرى الدم مع مرور الوقت



معظم الدواء المتبقى في الدم يكون بكميات  
قليلة بمرور الوقت

القدرات الرياضية لدى الأشخاص



معظم الأشخاص قدراتهم الرياضية متواضعة،  
و القلة لهم من يتمتعون بقدرات رياضية  
عالية

على الرغم من أن التوزيعات

المنحرفة إلى اليمين right

skewed distribution والمنحرفة

إلى اليسار left skewed

distribution والمتماثلة

symmetric distribution هي أكثر

التوزيعات شيوعا...

إلا أنه يمكن أن تكون البيانات في العالم الحقيقي غير منتظمة،

ولا تتبع أي من هذه التوزيعات

# الشكل والقيم المتطرفة

## The Shape and Outliers

Mohammed Lotfy

@mohammud.lotfy

في هذا الدرس، نريد إلقاء نظرة على الجانب الأخير المستخدم لوصف المتغيرات

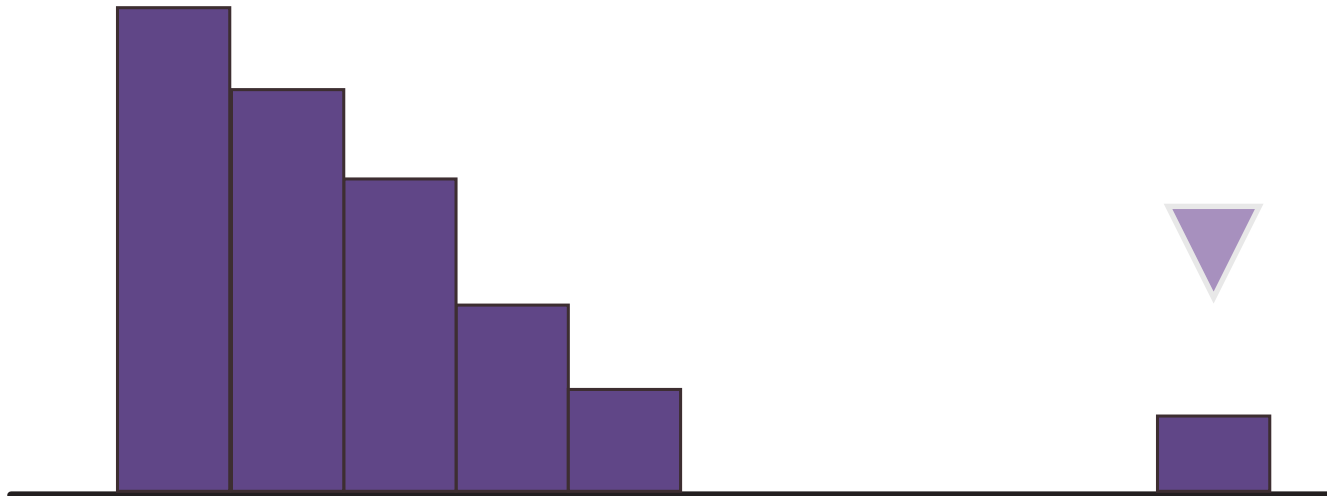
الكمية

المركز Center

الانتشار Spread

الشكل Shape

القيم المتطرفة Outliers



القيم الخارجية [القيم الشاذة  
أو المتطرفة] **outliers** هي  
نقاط البيانات [القيم] التي  
تقع **بعيدًا جدًا** عن بقية القيم  
في مجموعة البيانات

لإيضاح تأثير القيم الخارجية [القيم الشاذة أو المتطرفة] outliers في كيفية عرض تقارير عن الإحصاءات statistics التي خلصنا إليها، فلنفترض هذا المثال:

تخيل أنني اخترت عشرة رواتب سنوية لرجال أعمال، تسع منها كالتالي:

155 15 24 56 105 53 92 68 45

القيم بالآلاف دولار

والقيمة العاشرة هي راتب الرئيس التنفيذي لـ Facebook

1.6 مليار دولار

160 مليون دولار

يمكننا حساب وسط mean رواتب

رجال الأعمال وفقا لهذه البيانات ليكون

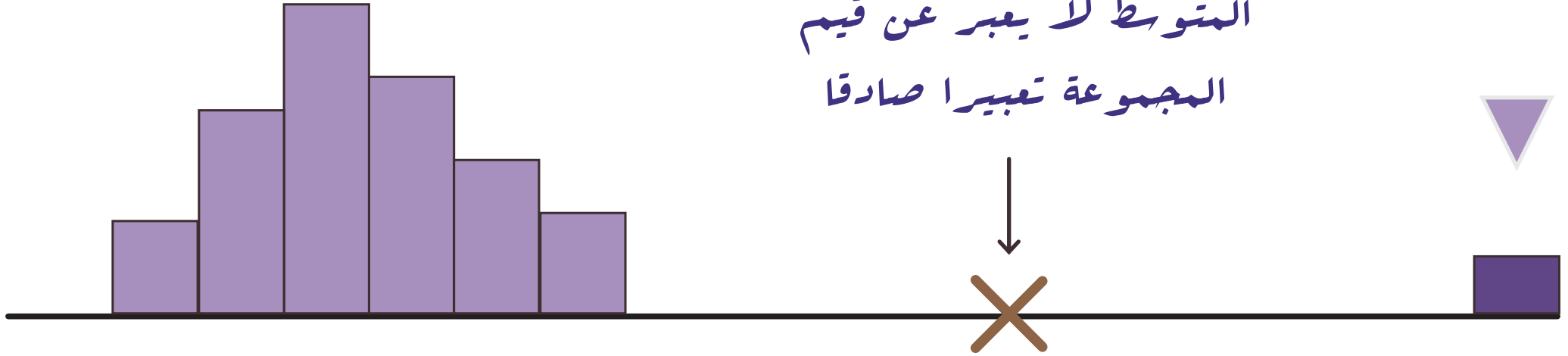
تقريبًا 160 مليون دولار في السنة

هذه قيمة مضللة للغاية

حرفيًا، لم يحصل أي من رجال الأعمال على هذا الراتب. لم تقترب

أي من الرواتب العشرة من هذا المبلغ

بسبب القيمة الشاذة فإن  
المتوسط لا يعبر عن قيم  
المجموعة تعبيرا صادقا



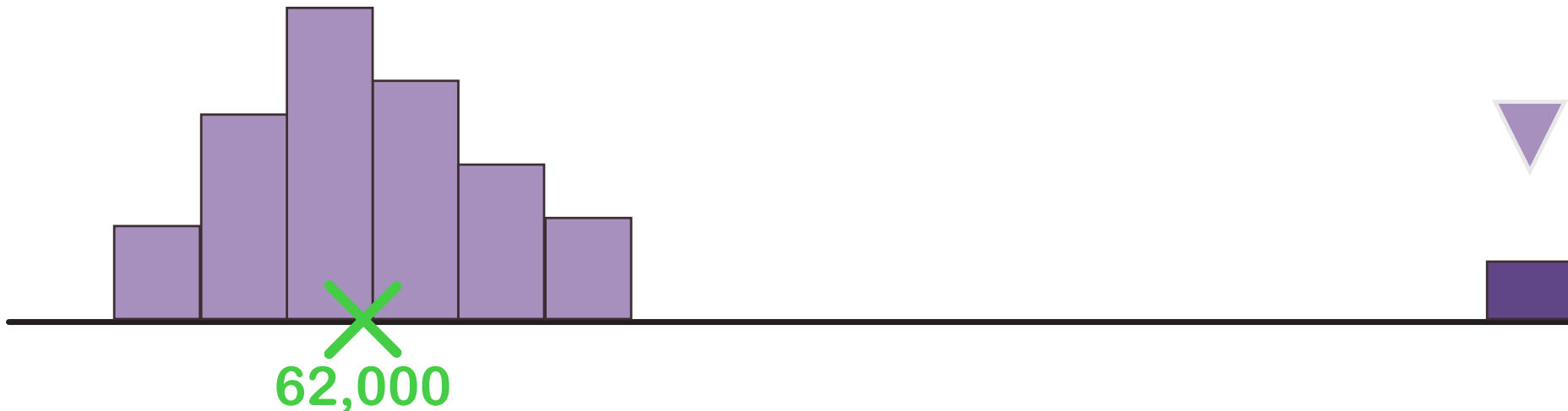
لأن الوسط mean يتأثر بالقيم الشاذة  
outliers، فإن قيمة ١,٦ مليار دولار قد  
عملت على رفع قيمة الوسط mean  
بشكل كبير، لتصبح غير معبرة عن  
معظم قيم الرواتب.



في هذه الحالة فإن القياس الأفضل للمركز سيكون الوسيط median

لأن الوسيط median لا يتأثر بالقيم  
المتطرفة outliers فإنه يعتبر طريقة أفضل  
لحساب المركز center في هذه الحالة.

والوسيط median هنا هو 62,000 دولار في العام



ويعتبر مؤشر أفضل لما يحتمل أن يربحه أي من رجال الأعمال وفقًا للبيانات التي لدينا.

لا يعتبر الانحراف المعياري standard deviation قياسًا جيدًا أيضًا عن الانتشار

spread في هذه الحالة

من الطبيعي أن يزيد الانحراف  
المعياري standard deviation هو  
الآخر لأنه يعتمد على الوسط mean  
في حسابه

فقيمة الانحراف المعياري هنا تقريبا 480 مليون دولار

وهي قيمة كبيرة تشير إلى أن أرباح رجال الأعمال مشتتة بدرجة هائلة،  
لكن ذلك ليس صحيحًا أيضًا

فقيم رواتب معظم رجال الأعمال متقاربة جدًا من بعضها البعض، في حين يوجد  
راتب واحد فقط هو المشتت عنها بدرجة هائلة.

# التعامل مع القيم المتطرفة

## Working with Outliers

Mohammed Lotfy

@mohammud.lotfy

كيف ينبغي لنا أن نتعامل مع هذه القيم الشاذة outliers في الواقع العملي؟

◀ على الأقل ينبغي أن نلاحظ وجودها.

◀ نحتاج إلى أن ندرك تأثيرها في الإحصائيات الملخصة للبيانات، مثل الوسط mean أو المدى range.

◀ ففي حال وجود قيم شاذة outliers، يزيد الوسط mean والانحراف المعياري standard deviation بدرجة كبيرة.



إذا كانت القيم الشاذة عبارة عن أخطاء مطبعية أو أخطاء في إدخال البيانات فيجب تصحيح هذا الخطأ وإزالة القيم الشاذة.

وإذا كنا نعلم قيمتها الصحيحة فيمكننا تعديلها

هناك حقل كامل يتناول هذا الموضوع يسمى  
اكتشاف القيم الشاذة anomaly detection



في الحالات المشابهة لمثال رواتب رجال الأعمال [الدرس السابق] ربما نحاول أن نفهم ما الذي كان مختلفًا تمامًا بشأن القيمة الشاذة مقارنة بالأفراد الآخرين؟

كيف أصبح رجل الأعمال هذا شخصًا في قمة النجاح؟

ولماذا تكون أرباحه كبيرة جدًا عند مقارنته بغيره؟

3

القياسات المرتبطة بالملخص ذي خمسة أعداد، مثل:  
الحد الأدنى minimum والحد الأقصى maximum  
والإرباعيات الأول  $Q_1$  والثاني  $Q_2$  والثالث  $Q_3$   
تكون أكثر فائدة عندما تكون القيم الشاذة موجودة

4

يوضح مثال رواتب رجال الأعمال أنك بحاجة إلى توخي  
الحذر بشأن كيفية مشاركة نتائجك وتقديم استنتاجاتك  
باستخدام إحصائيات تلخيص البيانات summary  
statistics عندما تكون لدينا قيم شاذة outliers

بعض الإحصائيات  
statistics مضللة أكثر  
من غيرها

التعبير عن البيانات برقم  
واحد فقط قد يكون  
مضللاً للغاية فيما  
يتعلق بالبيانات

إذا كنت مستخدماً للمعلومات information المستنتجة من البيانات data ، فمن المهم معرفة كيفية طرح الأسئلة الصحيحة المتعلقة بالإحصائيات statistics من

حولك





# نصائح للتعامل مع القيم المتطرفة

Advices for Working with Outliers

Mohammed Lotfy  
[@mohammud.lotfy](https://twitter.com/mohammud.lotfy)

إذا كنت الشخص الذي يعد التقارير فأليك بعض المبادئ التوجيهية عند تحليل البيانات:

أولا

مثل البيانات بشكل مرئي **plot** your data

## ثانيا

إذا كانت لديك قيم خارجية **outliers** فحدد **كيف ستتعامل معها**

قد يتطلب ذلك خبيرا في المجال الذي تنتمي إليه البيانات [ المجال الطبي أو المالي مثلا]

هل ينبغي إزالتها؟  
هل ينبغي إصلاحها؟  
هل ينبغي الاحتفاظ بها؟

## ثالثا

إذا كنت تعمل مع البيانات التي يتم توزيعها بشكل **اعتيادي** normally distributed ، أي شكل الجرس bell-shaped data الذي رأيناه من قبل

فسيتمكنك معرفة كل **التفاصيل الصغيرة** عن البيانات باستخدام الوسط **mean** والانحراف المعياري **standard deviation** فقط

قد يبدو هذا أمرا غريبا ، ولكنه صحيح

## رابعاً

ومع ذلك، إذا كانت البيانات الخاصة بنا منحرفة [ملتوية] **skewed** فإن **الملخص ذا الخمسة أعداد** يقدم معلومات عن مجموعات البيانات هذه أكثر مما يقدم المتوسط الحسابي mean والانحراف المعياري standard deviation

مرة أخرى، الملخص الأكثر فائدة والذي يمكنك الحصول عليه يكون مرئياً **visual** في

كثير من الأحيان

# الإحصاء الوصفي مقابل الإحصاء الاستدلالي

Descriptive vs. Inferential Statistics

Mohammed Lotfy  
[@mohammud.lotfy](https://twitter.com/mohammud.lotfy)

جميع الموضوعات التي تم تناولها حتى الآن تغطي ما يسمى

بالإحصاء الوصفي Descriptive Statistics

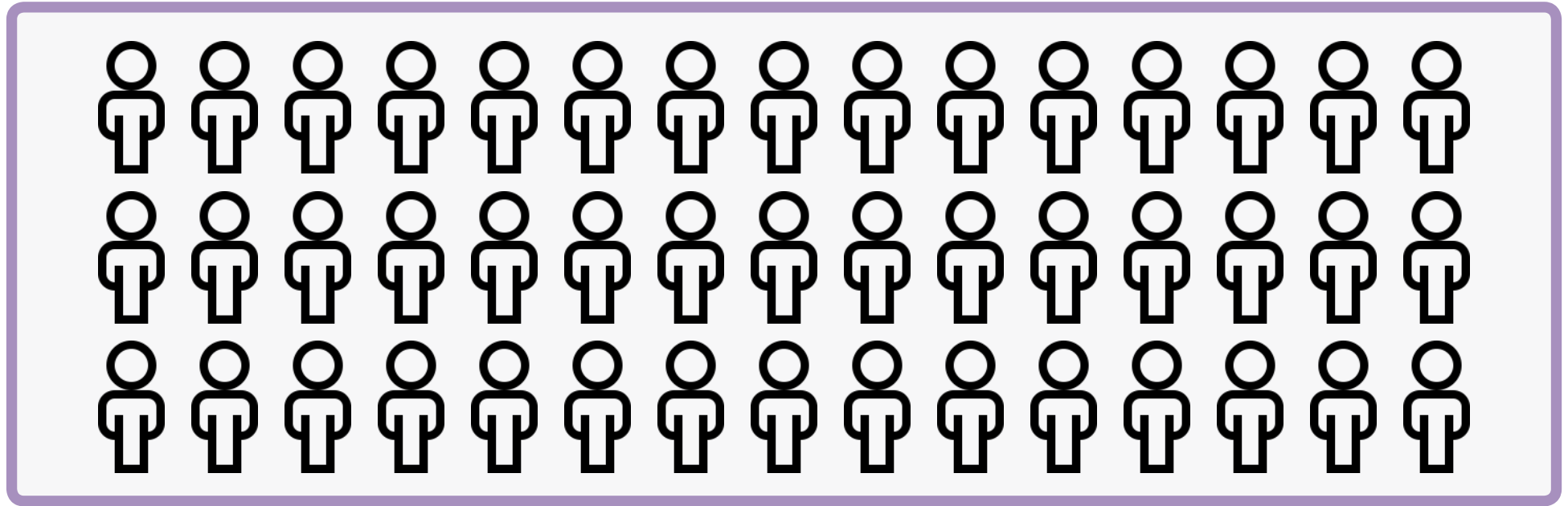
والتي تهدف إلى **وصف** البيانات التي تم جمعها

يوجد فرع كامل آخر من الإحصاء يُعرف باسم الإحصاء الاستنتاجية

[الإحصاء الاستدلالية] Inferential Statistics

يهدف إلى **استخلاص الاستنتاجات** حول مجتمع **population** الأفراد استنادًا فقط إلى عينة sample من من ذلك المجتمع

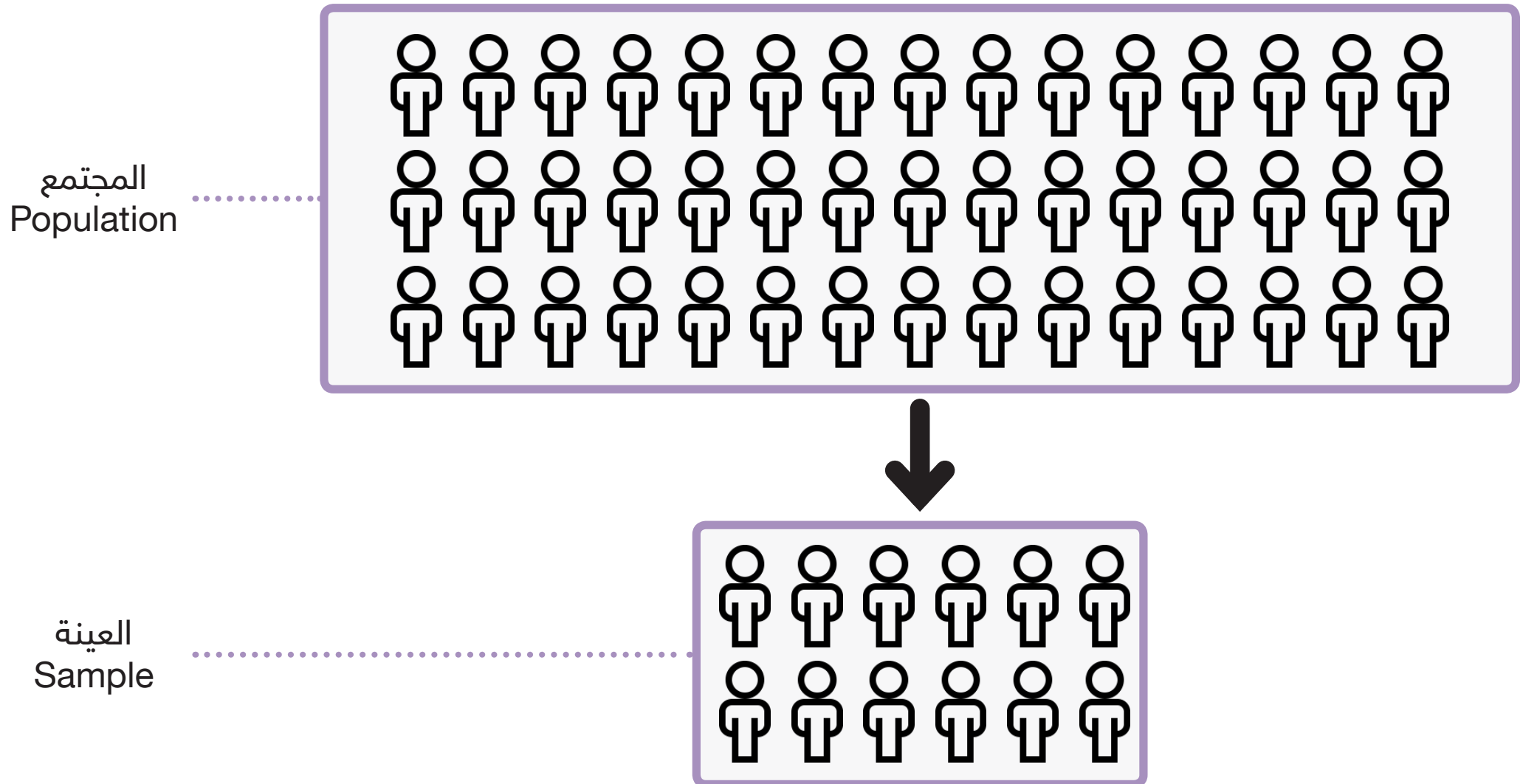
يشير مصطلح المجتمع population إلى جميع الأفراد أو الوحدات التي ندرسها  
بدراستها



فلو كنت تهدف إلى دراسة سلوك معين للأطفال من عمر ٣ إلى ٥ سنوات في جمهورية مصر العربية، فإن جميع أطفال مصر يمثلون المجتمع population الخاص بدراستك



ولأنه يصعب دراسة جميع أفراد المجتمع population فإنه يتم اختيار مجموعة جزئية من هذا المجتمع، تسمى العينة sample، لدراستها واستنتاج النتائج الخاصة بالمجتمع من خلالها.



لنفترض أنني أريد استنتاج نسبة من يشربون

القهوة بين جميع طلاب Udacity



فقمت بإرسال بريد إلكتروني إلى جميع خريجي  
Udacity والطلاب الحاليين، طارًا فيه السؤال  
التالي: هل تشربون القهوة؟



لنفترض أن القائمة كانت تحتوي على 100,000 بريد إلكتروني

المجتمع  
Population

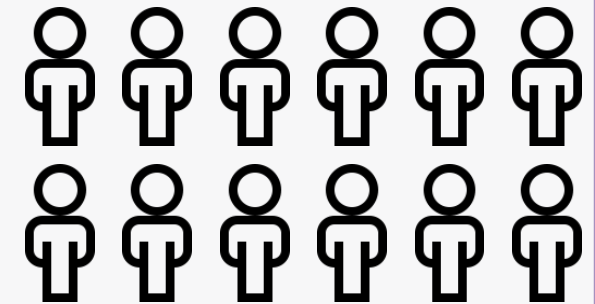


لسوء الحظ، ليس كل من أرسلتُ إليهم

رسائل البريد الإلكتروني قاموا بالرد،

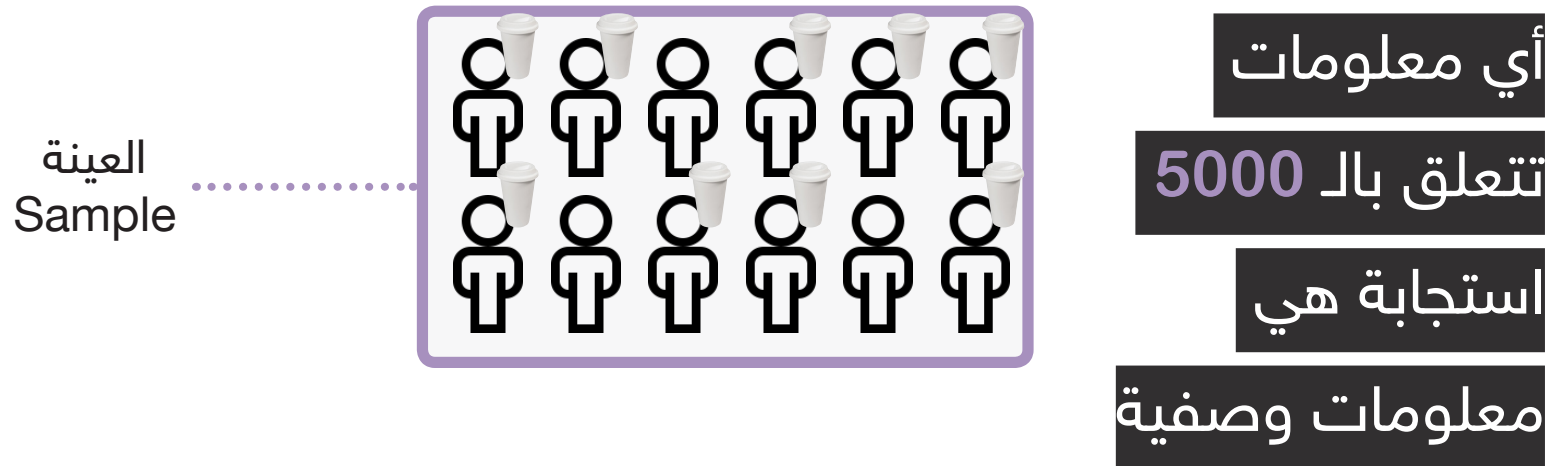
ومن ثمّ، تلقيت 5000 رد فقط

العينة  
Sample



توصلت إلى أن نسبة 73% من الأفراد الذين قاموا بالرد على رسالة البريد الإلكتروني

قالوا إنهم يشربون القهوة



أي أنسها تصف الـ 5000 طالب الذين  
قاموا بالرد على السؤال

تتعلق الإحصاء الوصفية Descriptive Statistics بوصف بيانات العينة التي لدينا



إن الإحصاء الاستنتاجية [الإحصاء الاستدلالية] Inferential Statistics تتعلق

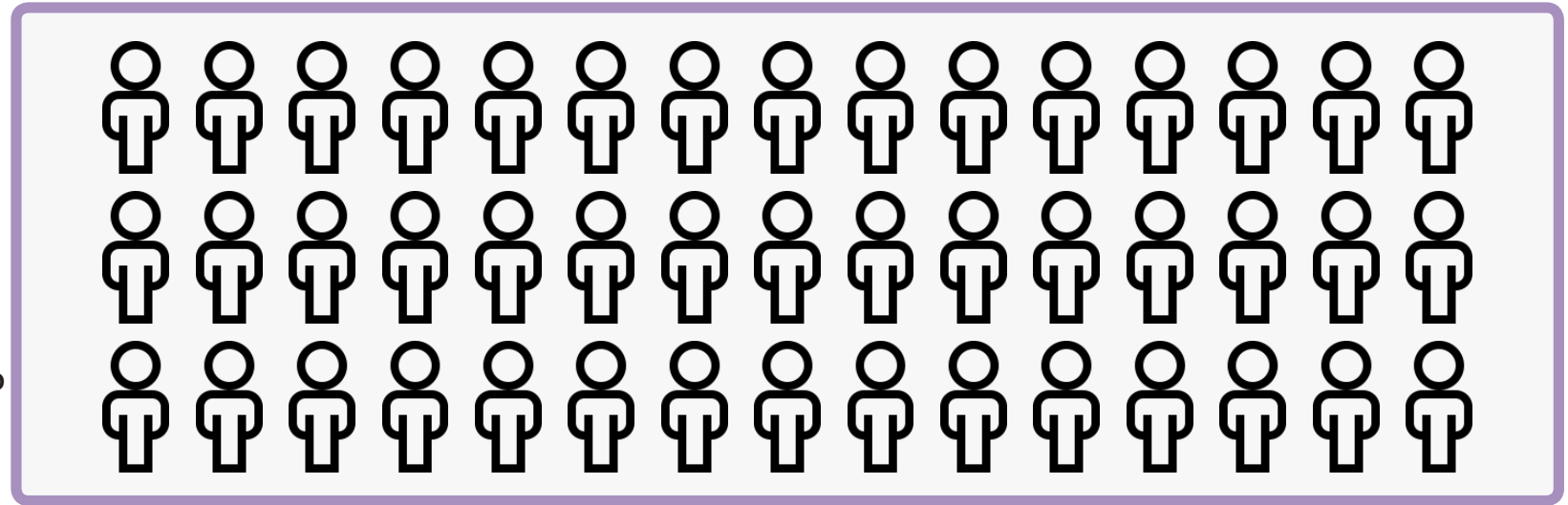
باستخلاص الاستنتاجات فيما يتعلق بعاتات تناول جميع الـ 100,000 طالب، وذلك

باستخدام البيانات المستخلصة من الـ 5000 طالب فقط.

أى أن الإحصاء الاستدلالي تستخدم نتائج الإحصاء الوصفي المتعلقة بالعينة  
Sample للوصول إلى استنتاجات خاصة بالمجتمع Population



المجتمع  
Population



3

نستطيع استنتاج  
معلومة تتعلق بالمجتمع  
population parameter، وتسمى  
مَعْلَمَة parameter

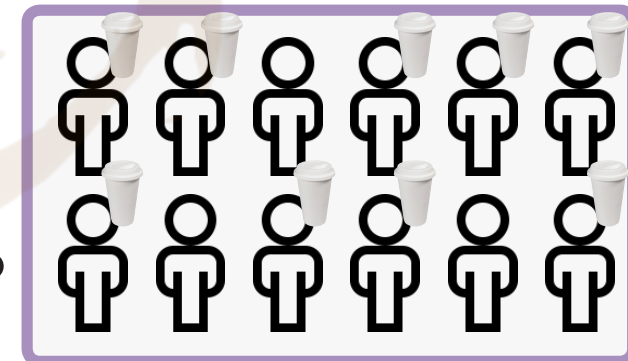
باستخدام الإحصاء الاستدلالي  
Inferential Statistics



العينة  
Sample

1

2



باستخدام الإحصاء الوصفي  
Descriptive Statistics

المعلومات التي نحصل عليها من  
العينة تسمى إحصاءة statistic

73%



المجتمع

**Population** 100,000

العينة

**Sample** 5,000

الإحصاءة

**Statistic** 73%

المعلمة

**Parameter**

نسب من يشربون القهوة  
في المجتمع المكون من 100,000 طالب



نقوم باختيار مجموعة جزئية من المجتمع population التي نطلق عليها اسم **العينة**

**sample**

في الحالة التي لدينا، نقصد ٥٠٠٠ طالب



يُطلق على أي ملخص رقمي محسوب من العينة اسم **الإحصاء statistic**

وهي تمثل نسبة ٧٣٪ من الـ ٥٠٠٠ طالب الذين يشربون القهوة



يُعرف الملخص الرقمي للمجتمع باسم **المَعْلَمَة parameter**

في هذا المثال **لا نعرف هذه القيمة** حتى الآن



يُعرف **استخلاص الاستنتاجات** المتعلقة بالمعلمة parameter، استنادًا إلى

الإحصاءات statistics، باسم الاستنتاج **[الاستدلال inference]**