05 | Exploratory Data Analysis

Ivan Corneillet

Data Scientist

GENERAL ASSEMBLY

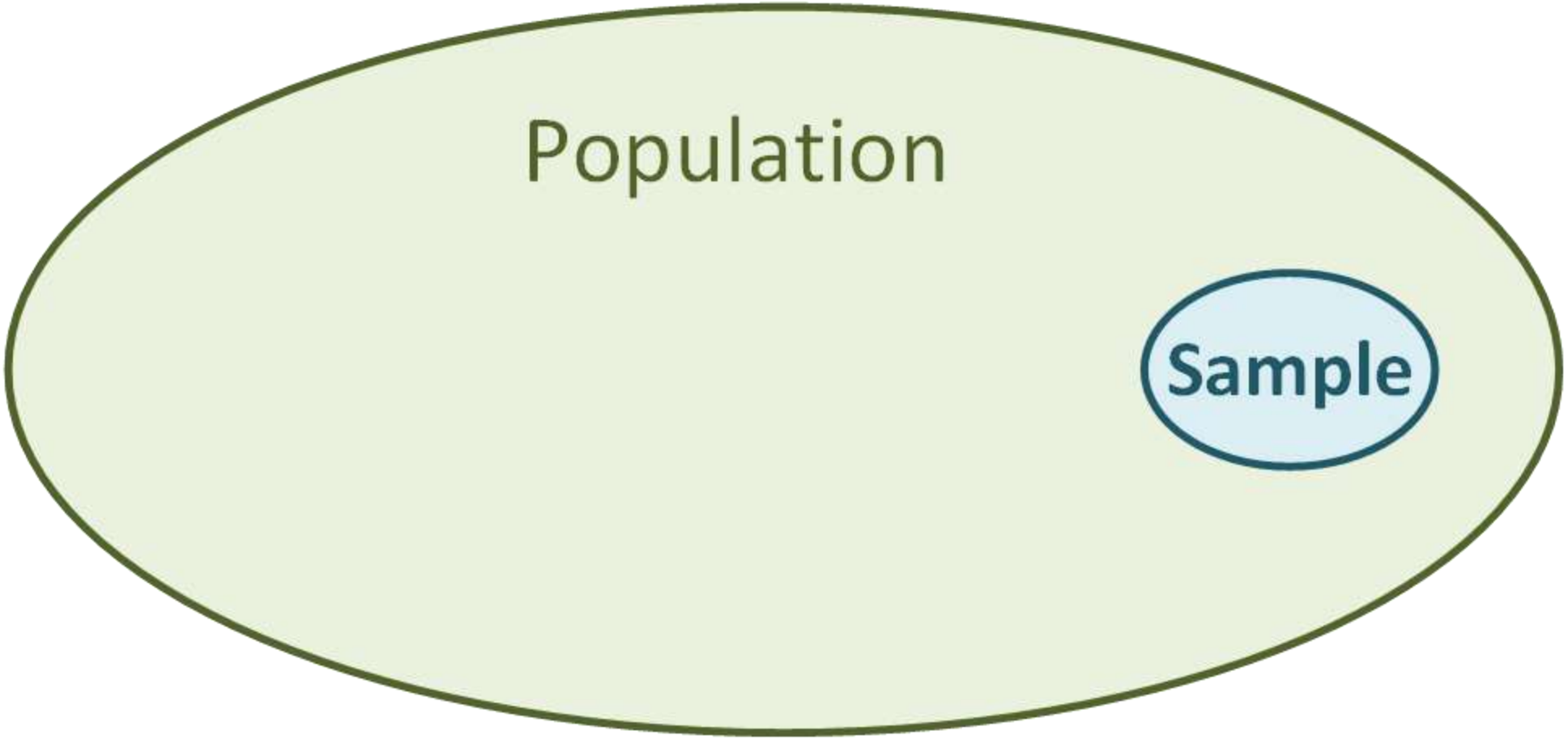
Learning Objectives

After this lesson, you should be able to:

- Identify variable types
- Use *pandas* to analyze datasets using basic summary statistics: mean, median, mode, max, min, quartile, inter-quartile range, variance, standard deviation, and correlation
- Create data visualizations – including boxplots, histograms, and scatter plots – to discern characteristics and trends in a dataset

Populations and Samples

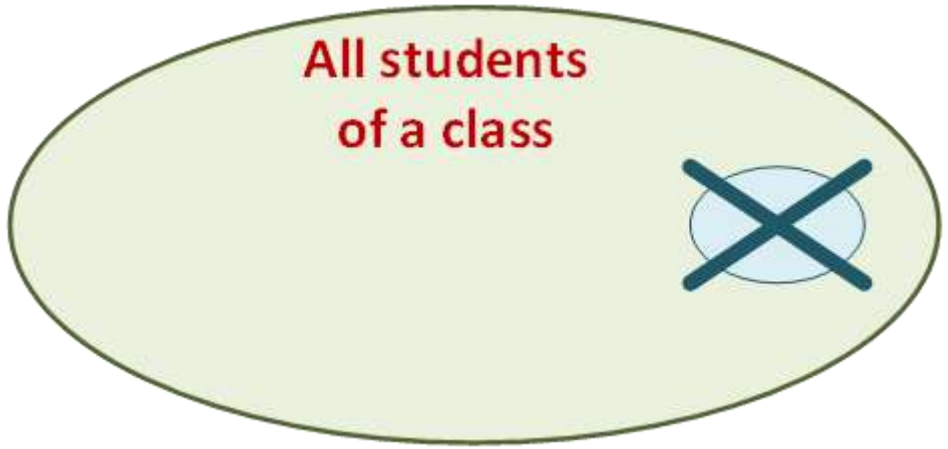
Populations and Samples



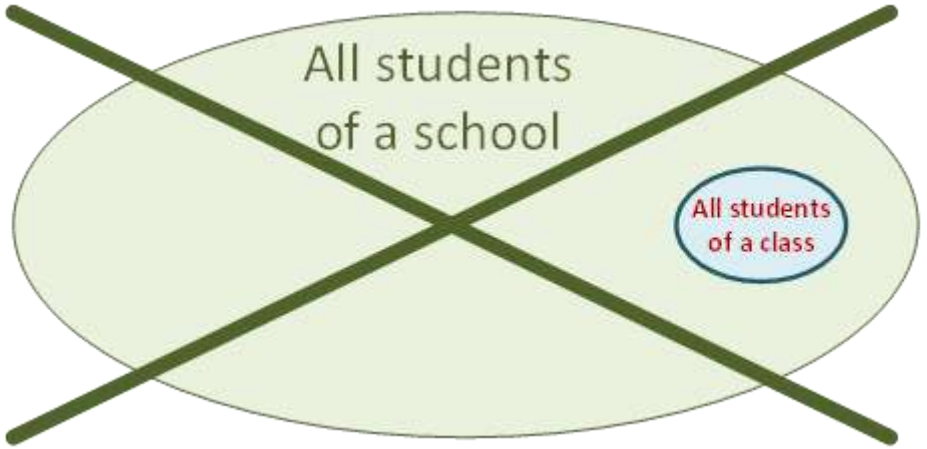
A dataset may be considered either as a population or a sample, depending on the reason for its collection and analysis

- Students of a class are a population if the analysis describes the distribution of scores in that class
- But they are a sample the analysis infers from their scores the scores of other students (e.g., all students from that school)

Descriptive Statistics

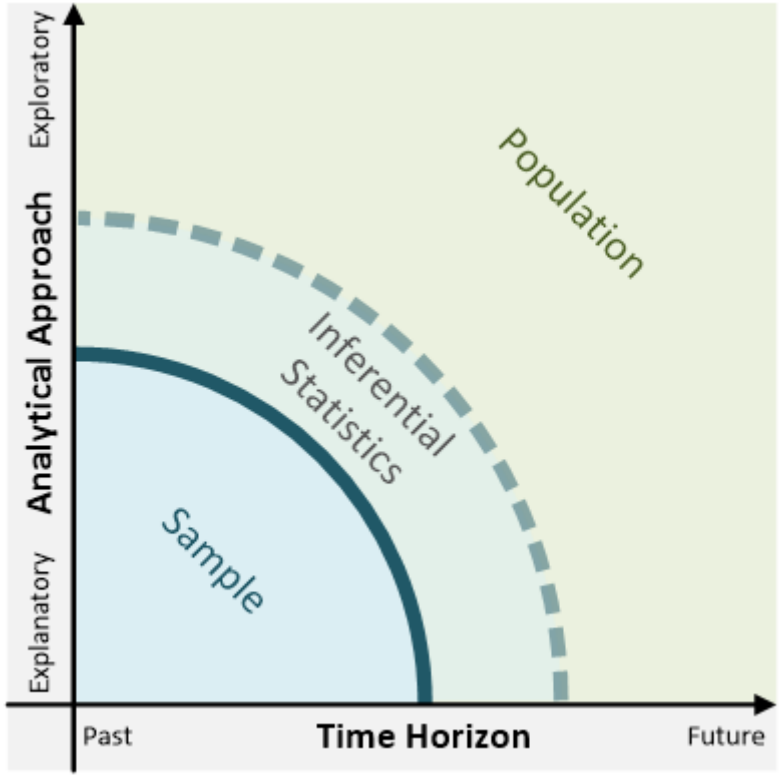


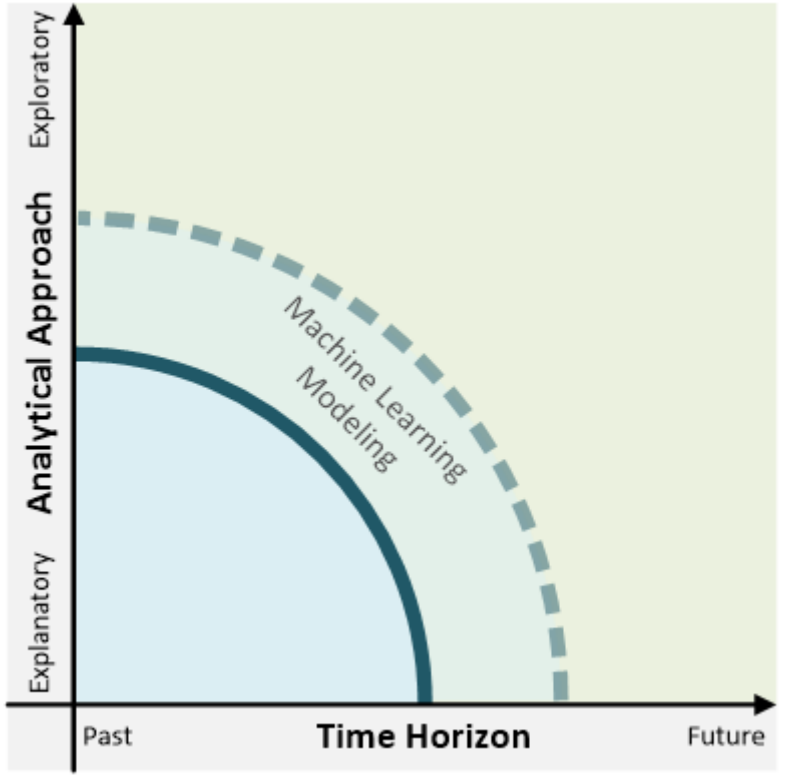
Inferential Statistics



Exploratory Data Analysis is concerned with Descriptive Statistics (e.g., “what happened last quarter?” and “how many units were sold?”)

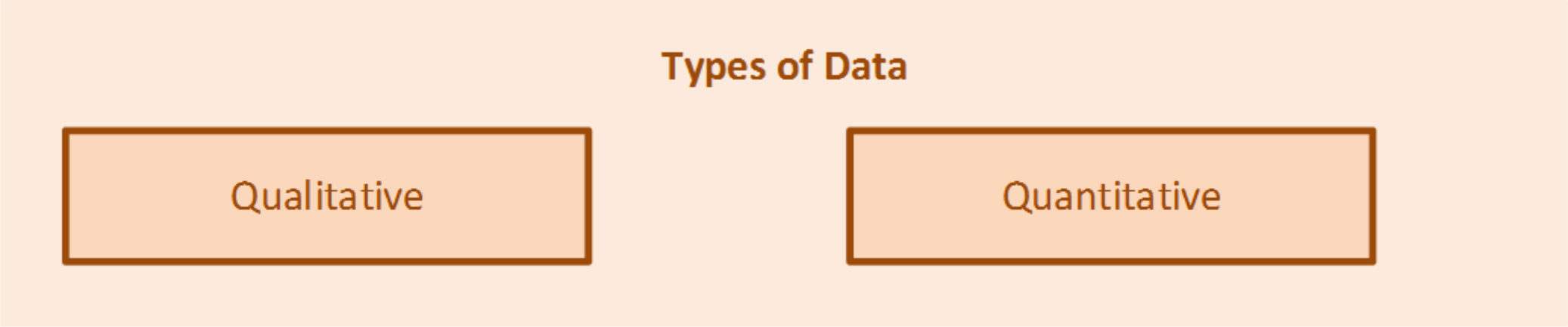
Machine Learning modeling concerns itself with Inferential Statistics (e.g., “what if …?”, “what will happen next?”, and “what if these trends continue?”)





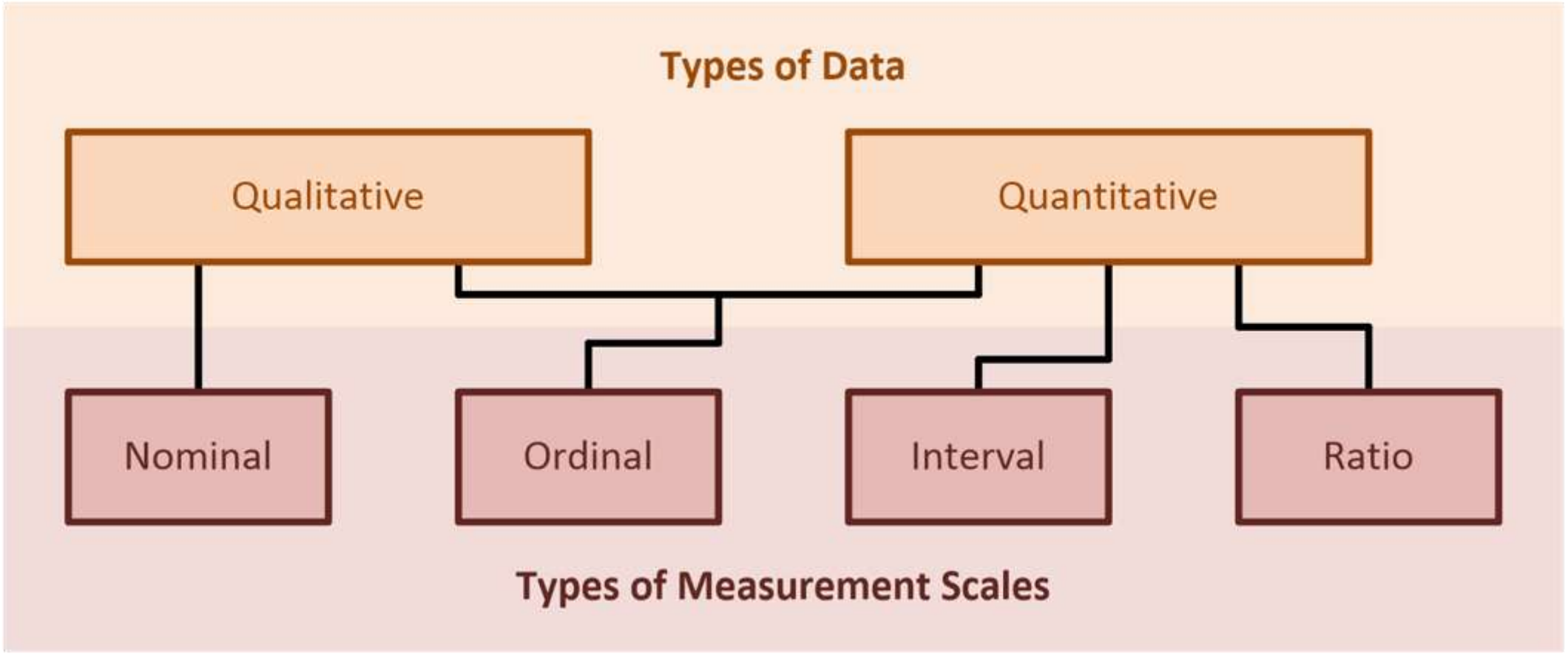
Types of Data
Types of Measurement Scales

Types of Data



- Qualitative Data
  - Uses descriptive terms to measure or classify something of interest, e.g., education level
- Quantitative Data
  - Uses numerical values to describe something of interest, e.g., age

Types of Measurement Scales

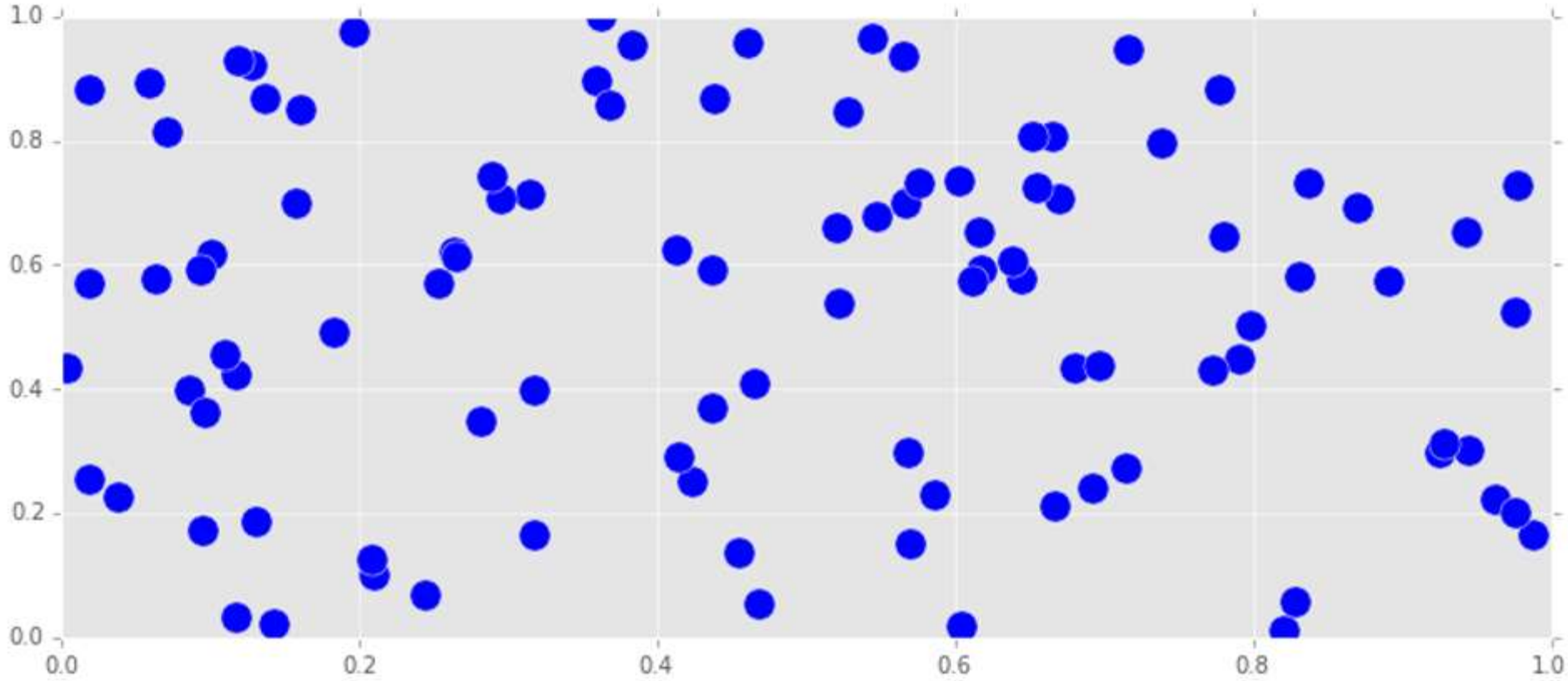


Types of Measurement Scales (cont.)

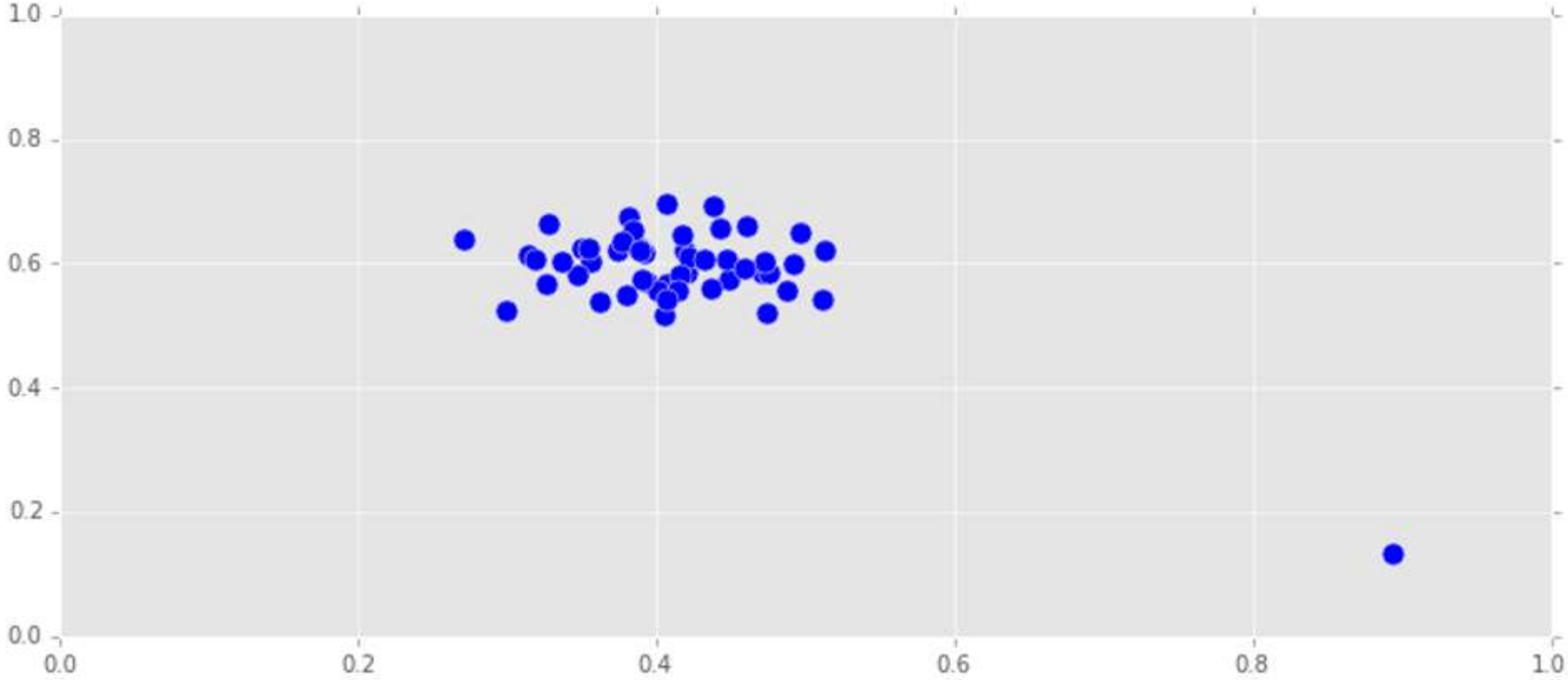
	Nominal	Ordinal	Interval	Ratio
e.g.	*Gender*	*Movie ratings*	*Temperature*	*Salary*
Categorize?	✓ (male, female)	✓	✓	✓
Rank-order?	✘	✓ (★<2★<3★<4★)	✓	✓
Add and subtract?	✘	✘ (4★−3★≠★)	✓ (75°C is 50°C warmer than 25°C)	✓
Multiply and divide?	✘	✘ (4★ not 4× better than 1★)	✘ (75°C not 3× as warm as 25°C) (0°C doesn't mean no temperature!)	✓ (Salary of $200K is 2× that of $100K) ($0 means no salary ☹)

Measures of Central Tendency
Measures of Dispersion

How would you summarize this data?



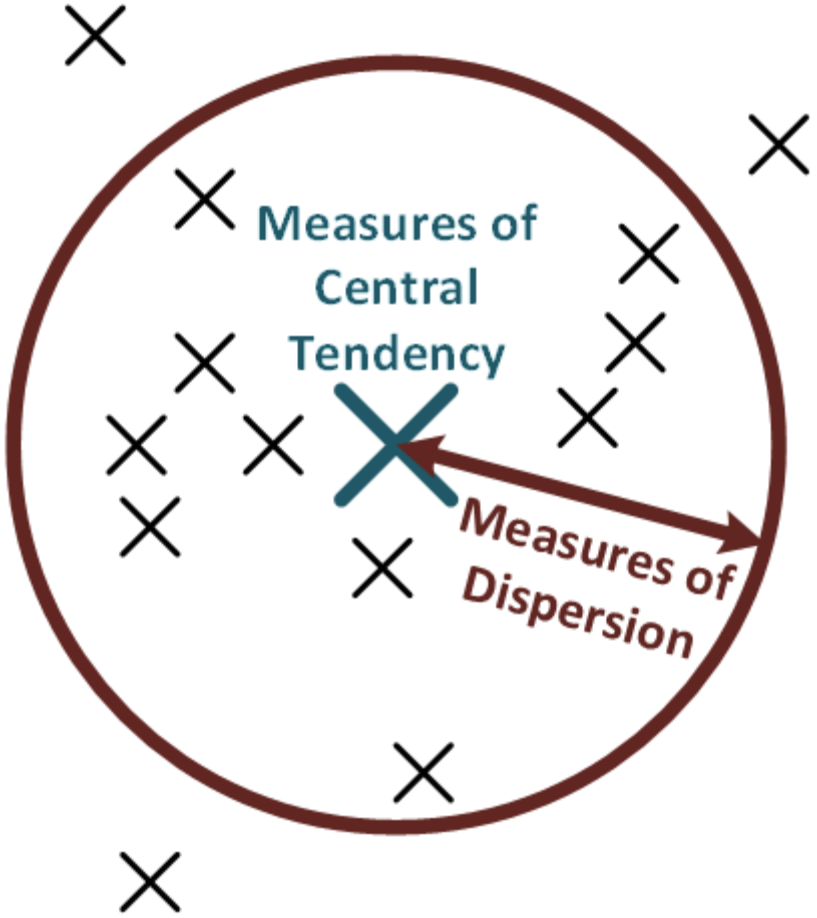
How would you summarize this data? (cont.)



Mean, Median, and Mode | Trade-offs

	Value is in the dataset	Value is easy to compute	Value is resistant to outliers	Corresponding measure of Dispersion	Used extensively by mathematical models
Mean	☹ (Unlikely)	☺	☹	☺ (Variance, standard deviation)	☺
Median	😐 (50% chance)	😐 (need to rank the values)	☺	☺ (Interquartile Range)	☹
Mode	☺ (Always)	☹ (Need to count and rank the count)	☺	☹ (Not really)	☹ (Mode might not be defined or you might have multiple values)

Measures of Central Tendency and Measures of Dispersion



- Measures of Central Tendency
  - (Or measures of location)
  - Answer the question: “What’s the typical or common value for a variable?”
  - Mean, Median, Mode
- Measures of Dispersion
  - (Or measures of variability/spread)
  - Answer the question: “How far do values stray from the typical value?”
  - Variance, Standard Deviation, Range, Interquartile Range (IQR)

Mean, Variance, and Standard Deviation

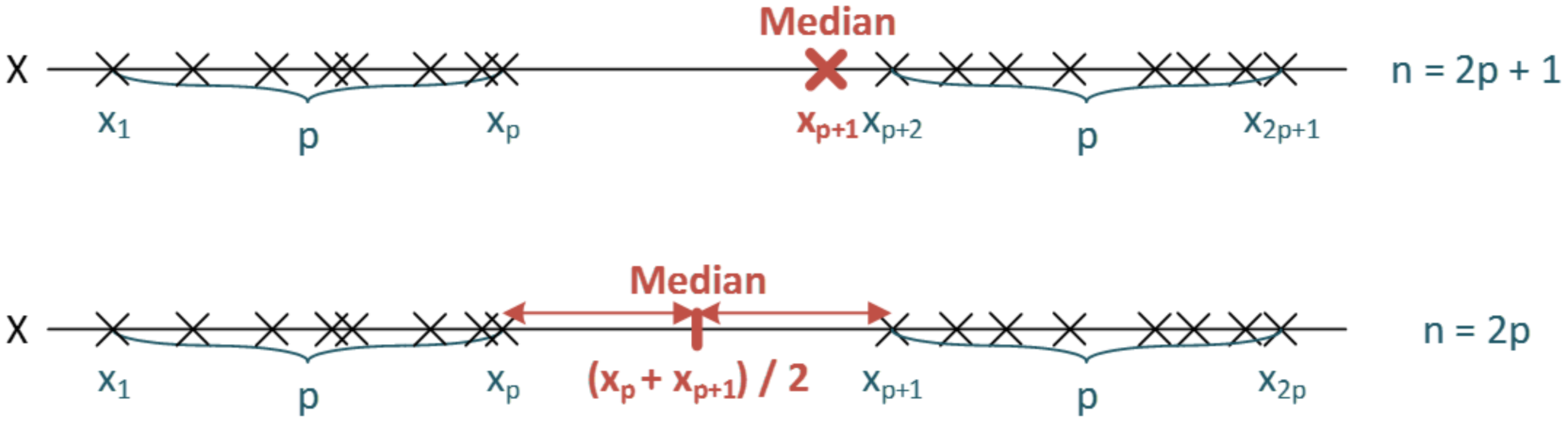
Ordinal ✘	Nominal ✘	Interval ✓	Ratio ✓

	Population	Sample
(Arithmetic) Mean *(a.k.a., the first moment)* (Mean has unit of X:$[X]$)	$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = E[X^1]$ (mu)	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (x-bar)
Variance *(a.k.a., the second moment)* $[X^2]$	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$ $= E[(X-\mu)^2]$ (sigma-squared)	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Standard Deviation $[X]$	$\sigma = \sqrt{\sigma^2}$ (sigma)	$s = \sqrt{s^2}$

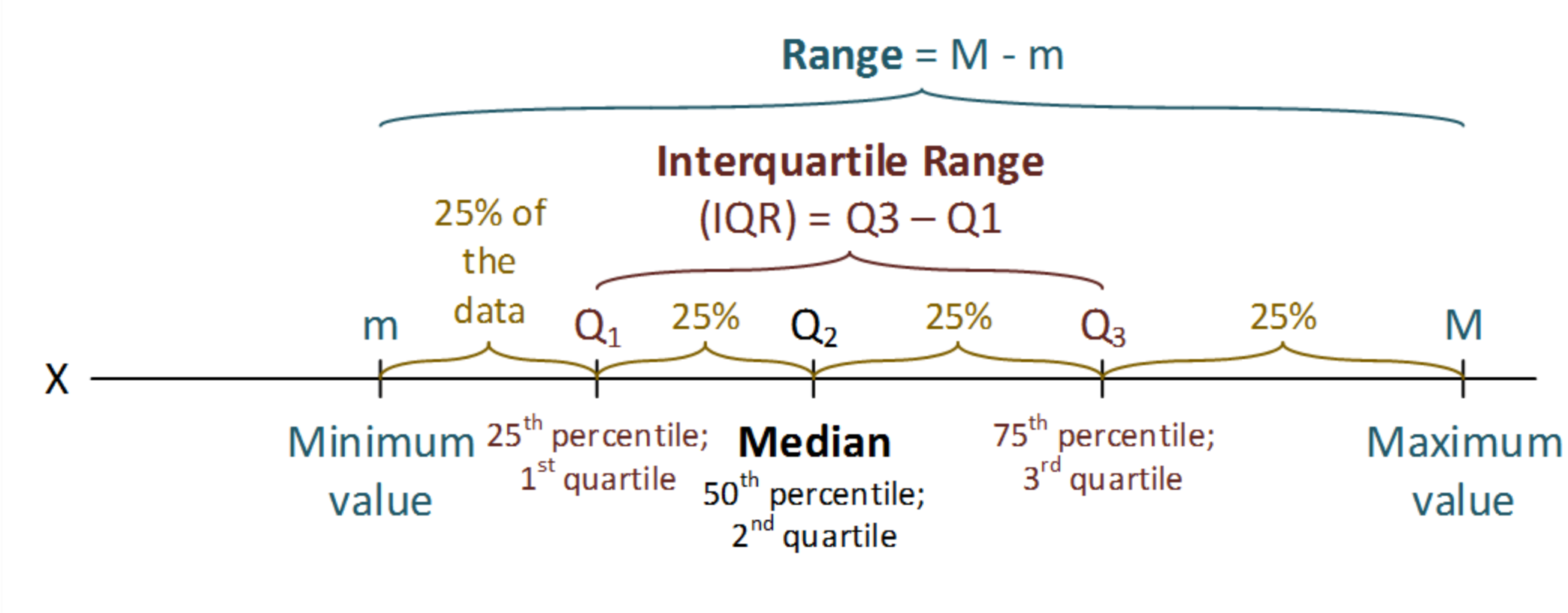
(mean, variance, and standard deviations are based on the values of x_i)

Median, Range, and Interquartile Range

Median


Median
X
x_1
p
x_p
x_{p+1} x_{p+2}
p
x_{2p+1}
n = 2p + 1
Median
X
x_1
p
x_p
$(x_p + x_{p+1}) / 2$
x_{p+1}
p
x_{2p}
n = 2p

Median, Range, and Interquartile Range



Median, Range, and Interquartile Range (cont.)

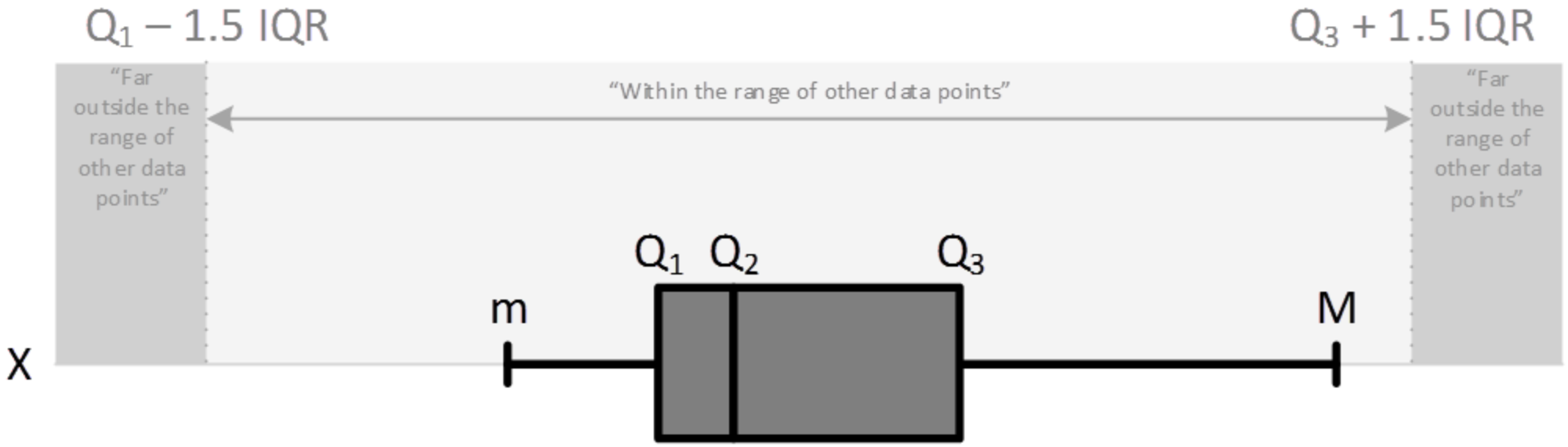
Nominal ✘	Ordinal ✘	Interval ✓	Ratio ✓

Median	$median = \begin{cases} x_{p+1} \; if \; n = 2p+1 \\ \dfrac{x_p + x_{p+1}}{2} \; if \; n = 2p \end{cases}$
Range	$range = x_n - x_1$
Percentile	$q_k = \begin{cases} x_{\lceil p \rceil} \; if \; p = \dfrac{nk}{100} \; not \; integer \\ \dfrac{x_p + x_{p+1}}{2} \; otherwise \end{cases}$
Quartile	$Q_1 = q_{25}; Q_3 = q_{75}$
Interquartile Range	$IQR = Q_3 - Q_1$

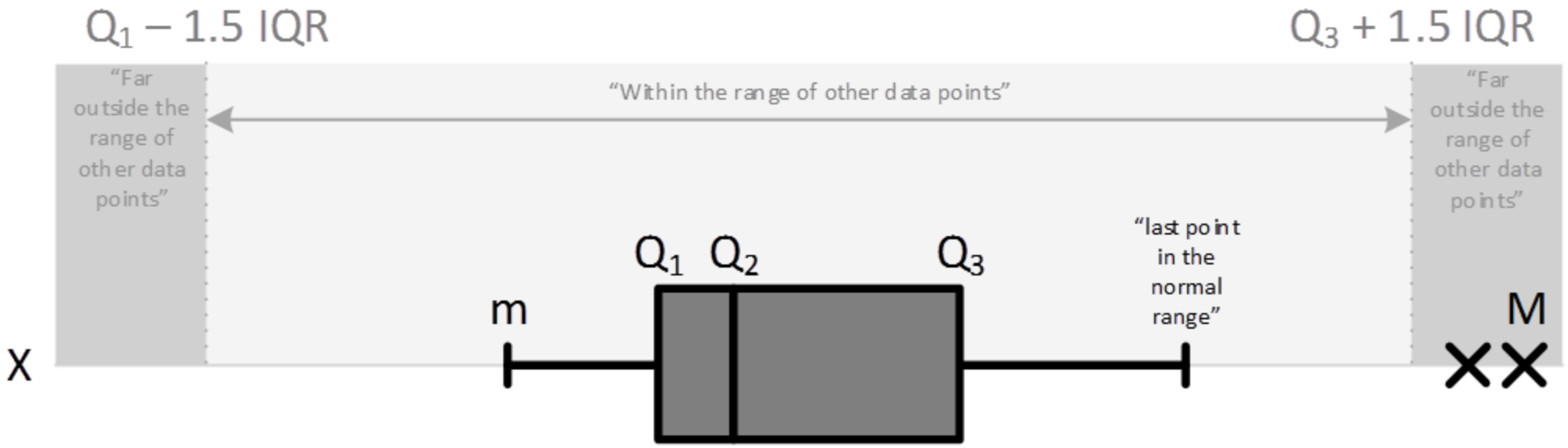
(median, range, and interquartile range are based on the ranks of x_i; x_i ranked from smallest to largest)

Boxplots

Boxplot #1 | Median, Range, Interquartile Range; no Outliers



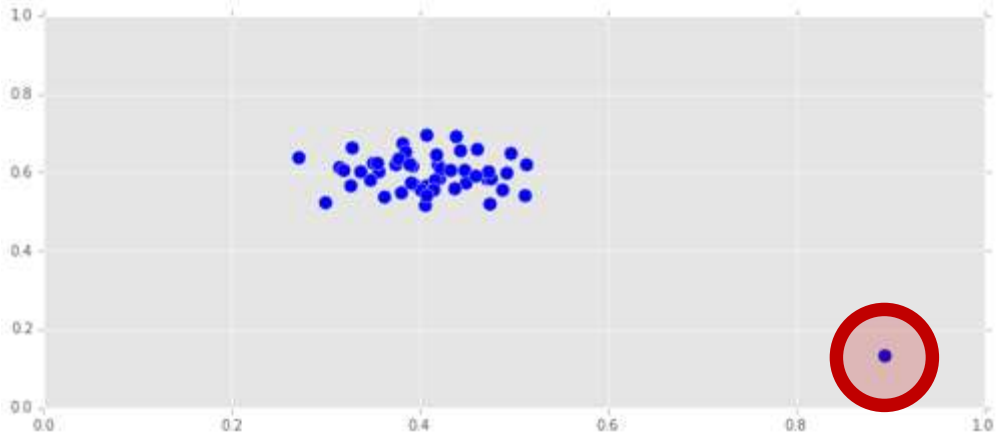
Boxplot #2 | Median, Range, Interquartile Range; with Outliers

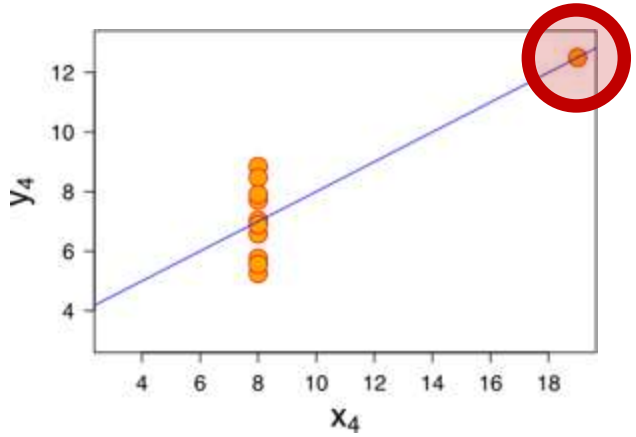


Outliers

Think twice before discarding outliers; they might be the most important points of your dataset

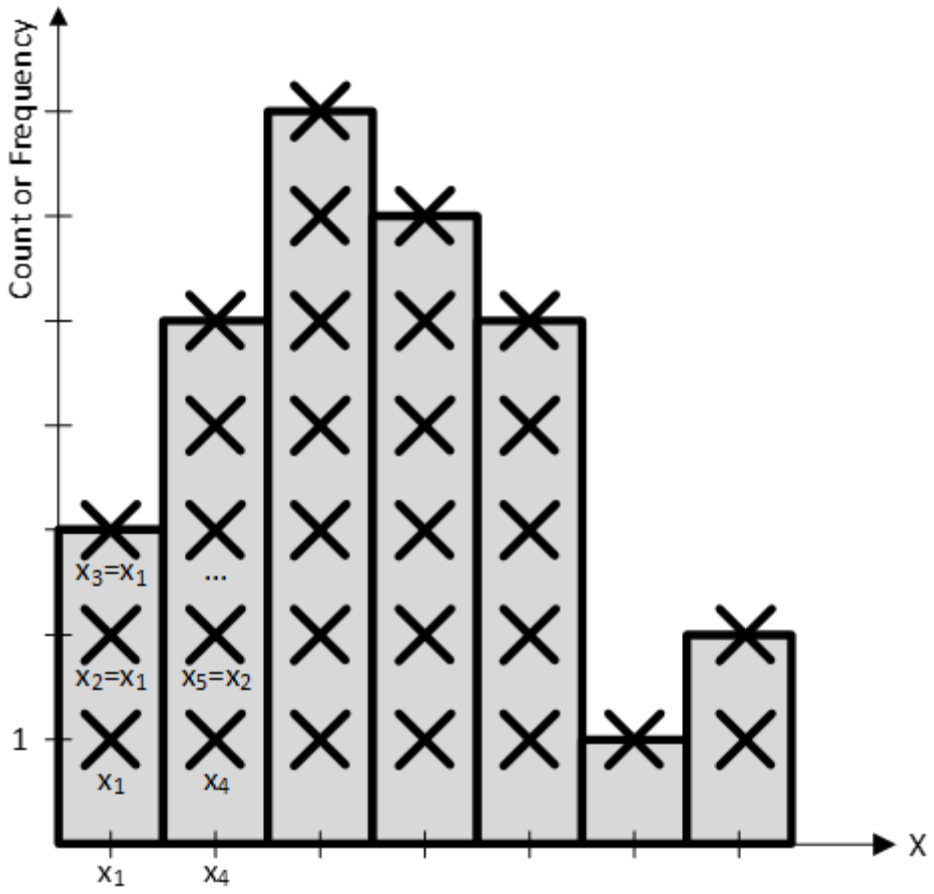
- Outliers are values that are “far” from the central tendency
- No formal definition among statisticians on how to define outliers (how do you define “far”?)
- However, general agreement that they be identified and dealt with appropriately (e.g., keep or discard)
    - They might be the most important points of your dataset





Histograms

Histograms. $x_1 = x_2 = x_3 < x_4 = x_5 \ldots$

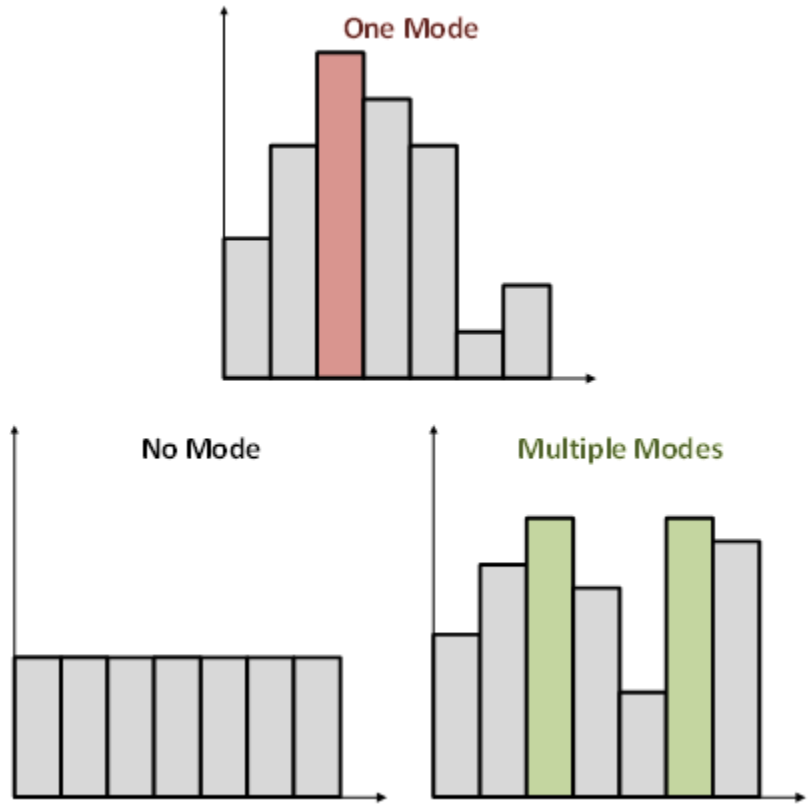


Mode

Modes and Histograms

Nominal ✓	Ordinal ✓	Interval ✓	Ratio ✓

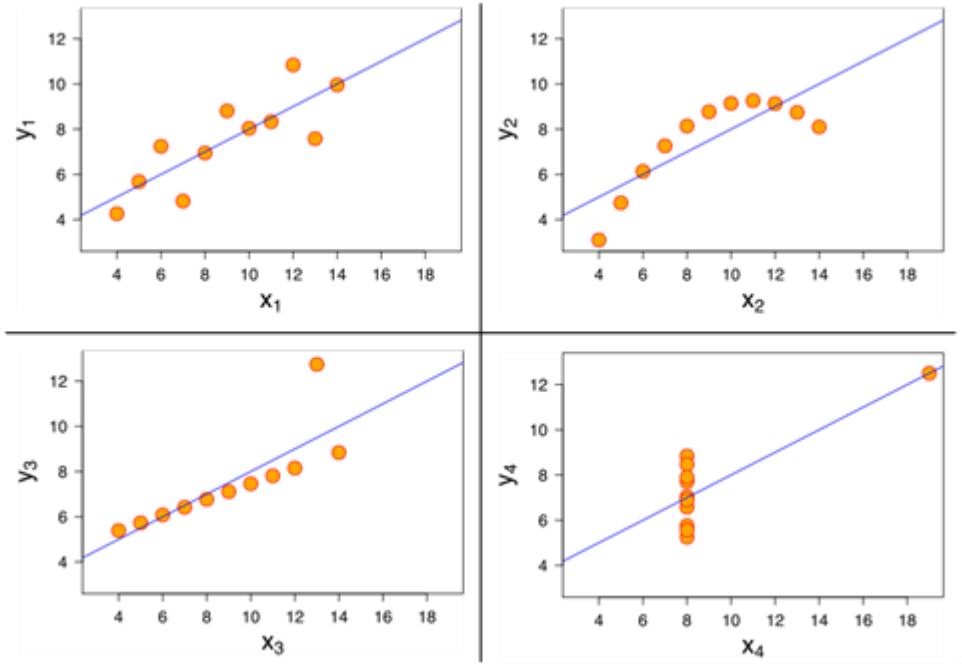
- The Mode is the value(s) that occur(s) most often



Plot the Data!

Don’t rely on basic statistic properties and **plot the data!** 4 datasets (Anscombe’s Quartet) that have nearly identical simple statistical properties, yet are very different

Scatter plot appears to be a simple linear relationship, corresponding to two variables correlated and following the assumption of normality.



Not distributed normally; while an obvious relationship between the two variables can be observed, it is not linear, and the linear correlation is not relevant.

Distribution is linear, but with a different regression line, which is offset by the one outlier which exerts enough influence to alter the regression line.

Example when one outlier is enough to produce a high correlation coefficient, even though the relationship between the two variables is not linear.

Property	Value
Mean of x_i	9
Sample variance of x_i	11
Mean of y_i	7.50
Sample variance of y_i	4.122 or 4.127
Correlation between x_i and y_i	0.816
Linear regression line in each case	$y_i = 3.00 + 0.500\ x_i$

(Linear) Correlation

Correlation

- A measure of strength and direction for a **linear association** between two random variables

$$\rho_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

- $\rho = 0$ means that the two variables don't have a linear association
  - It doesn't imply that they are independent!

Correlation (cont.)