More on statistics

Assoc. Prof. Nguyen Manh Tuan

9/2022

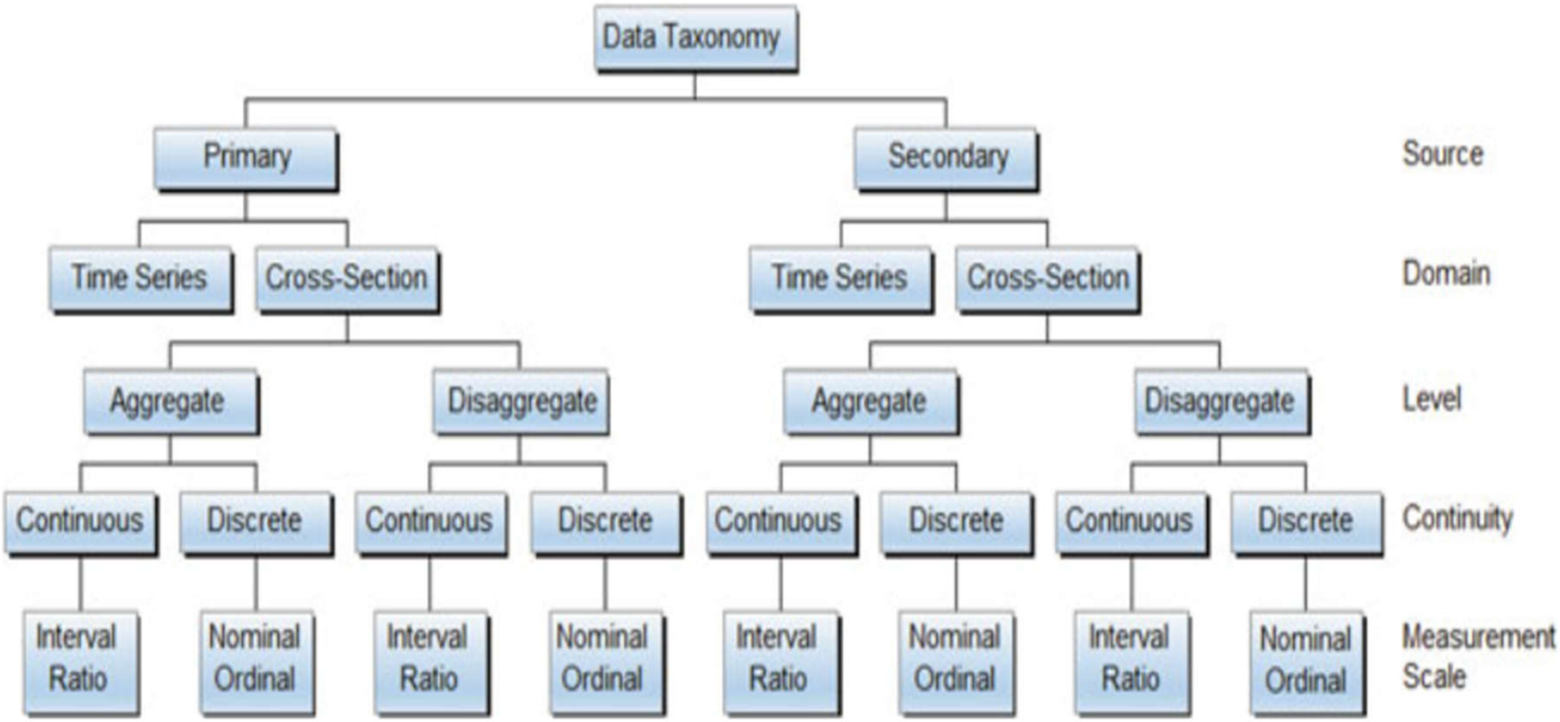


Fig. 2.1 A data taxonomy. Source: Paczkowski (2016). Permission to use granted by SAS Press

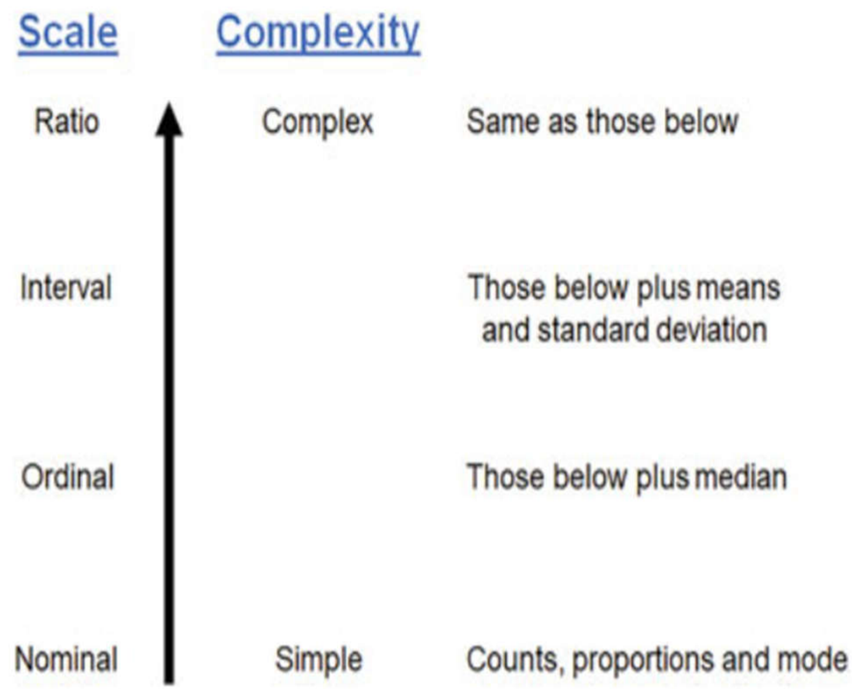


Fig. 2.2 Measurement scales attributed to Stevens (1946). Source for this chart: Paczkowski (2016). Permission to use granted by SAS Press

- Nominal scale(counts, proportions, mode)
  - “Buy/Don’t buy”
  - Black, brown, blue, red
- Ordinal scale (median, percentiles)
  - Entry-level, middle, executive-level
- Interval scale (mean, sd) (distance between values is meaningful; but the origin is meaningless because it can be changed)
  - 80F/40F = 2 but 80F is twice as hot as 40F?
  - C = (F-32)*5/9; 40F = 4C; 80F = 27C; 27C/4C # 2
- Ratio scale (fixed zero as an origin)
  - Sales

Percentiles, Quartiles, and Box-Plots

Percentiles

- *Percentiles* are data that have been divided into 100 groups.
- For example, you score in the 83rd percentile on a standardized test. That means that 83% of the test-takers scored below you.
- *Deciles* are data that have been divided into 10 groups.
- *Quintiles* are data that have been divided into 5 groups.
- *Quartiles* are data that have been divided into 4 groups.

Percentiles, Quartiles, and Box Plots

Percentiles

- Percentiles may be used to establish ***benchmarks*** for comparison purposes (e.g. health care, manufacturing, and banking industries use 5th, 25th, 50th, 75th and 90th percentiles).
- Quartiles (25, 50, and 75 percent) are commonly used to assess financial performance and stock portfolios.
- Percentiles can be used in employee merit evaluation and salary benchmarking.

Percentiles, Quartiles, and Box Plots

Quartiles

- ***Quartiles*** are scale points that divide the sorted data into four groups of approximately equal size.

<table>
<tr><th></th><th>Q_1</th><th></th><th>Q_2</th><th></th><th>Q_3</th><th></th></tr>
<tr><td>⇦Lower 25%⇨</td><td>|</td><td>⇦Second 25%⇨</td><td>|</td><td>⇦Third 25%⇨</td><td>|</td><td>⇦Upper 25%⇨</td></tr>
</table>

- The three values that separate the four groups are called Q_1, Q_2, and Q_3, respectively.

Percentiles, Quartiles, and Box Plots

Quartiles

- The second quartile Q_2 is the <u>*median*</u>, a measure of *central tendency*.

			Q_2			
⇦ Lower 50% ⇨			|	⇦ Upper 50% ⇨		

- Q_1 and Q_3 measure *dispersion* since the <u>*interquartile range*</u> $Q_3 - Q_1$ measures the degree of spread in the middle 50 percent of data values.

	Q_1		Q_3	
⇦Lower 25%⇨	|	⇦ Middle 50% ⇨	|	⇦Upper 25%⇨

Percentiles, Quartiles, and Box Plots

Quartiles – The method of medians

- The first quartile Q_1 is the median of the data values below Q_2, and the third quartile Q_3 is the median of the data values above Q_2.

<table>
<tr><th></th><th>Q_1</th><th></th><th>Q_2</th><th></th><th>Q_3</th><th></th></tr>
<tr><td>⇦Lower 25%⇨</td><td>|</td><td>⇦Second 25%⇨</td><td>|</td><td>⇦Third 25%⇨</td><td>|</td><td>⇦Upper 25%⇨</td></tr>
</table>

For first half of data, 50% above, 50% below Q_1.

For second half of data, 50% above, 50% below Q_3.

Percentiles, Quartiles, and Box Plots

Method of Medians

- For small data sets, find quartiles using ***method of medians***:

Step 1: Sort the observations.

Step 2: Find the median Q_2.

Step 3: Find the median of the data values that lie below Q_2.

Step 4: Find the median of the data values that lie above Q_2.

Percentiles, Quartiles, and Box Plots

Method of Medians

Example:

A financial analyst has a portfolio of 12 energy equipment stocks. She has data on their recent price/earnings (P/E) ratios. To find the quartiles, she sorts the data, finds Q_2 (the median) halfway between the middle two data values, and then finds Q_1 and Q_3 (medians of the lower and upper halves, respectively) as illustrated in Figure 4.25.

FIGURE 4.25 **Method of Medians**

Company	*Sorted P/E*
Maverick Tube	7
BJ Services	22
FMC Technologies	25
Nabors Industries	29
Baker Hughes	31
Varco International	35
National-Oilwell	36
Smith International	36
Cooper Cameron	39
Schlumberger	42
Halliburton	46
Transocean	49

Q_1 is between x_3 and x_4 so
$Q_1 = (x_3 + x_4)/2 = (25 + 29)/2 = 27.0$

Q_2 is between x_6 and x_7 so
$Q_2 = (x_6 + x_7)/2 = (35 + 36)/2 = 35.5$

Q_3 is between x_9 and x_{10} so
$Q_3 = (x_9 + x_{10})/2 = (39 + 42)/2 = 40.5$

Source: Data are from *BusinessWeek*, November 22, 2004, pp. 95–98.

Percentiles, Quartiles, and Box Plots

Example: P/E Ratios and Quartiles

- So, to summarize:

	Q_1		Q_2		Q_3	
⇦Lower 25%⇨ of *P/E* Ratios	**27**	⇦Second 25%⇨ of *P/E* Ratios	**35.5**	⇦Third 25%⇨ of *P/E* Ratios	**40.5**	⇦Upper 25%⇨ of *P/E* Ratios

- These quartiles express central tendency and dispersion. What is the interquartile range?

Percentiles, Quartiles, and Box Plots

Quartiles – Excel

Quartile	Percent Below	Excel Quartile Function	Excel Percentile Function	Interpolated Position in Data Array
Q_1	25%	=QUARTILE.EXC(Data,1)	=PERCENTILE.EXC(Data,.25)	$.25n + .25$
Q_2	50%	=QUARTILE.EXC(Data,2)	=PERCENTILE.EXC(Data,.50)	$.50n + .50$
Q_3	75%	=QUARTILE.EXC(Data,3)	=PERCENTILE.EXC(Data,.75)	$.75n + .75$

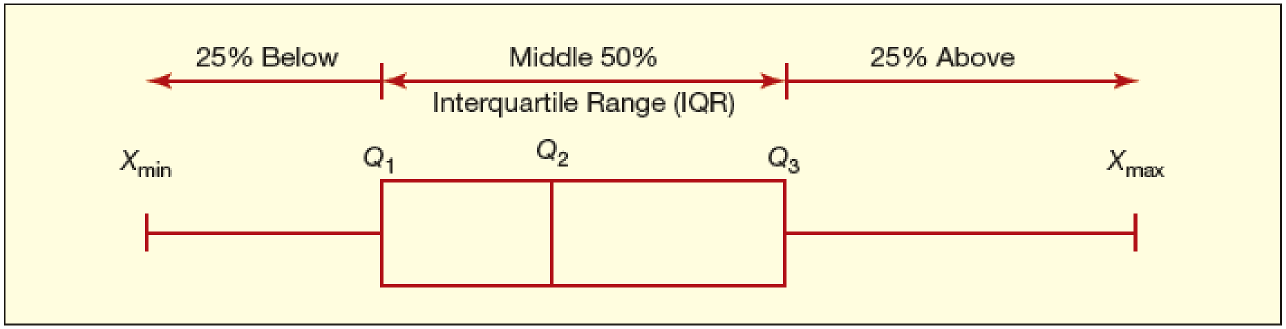
Percentiles, Quartiles, and Box Plots

Box Plots

A useful tool of *exploratory data analysis* (EDA) is the **box plot** (also called a *box-and-whisker plot*) based on the **five-number summary**:

$x_{min}, Q_1, Q_2, Q_3, x_{max}$

The box plot is displayed visually, like this.



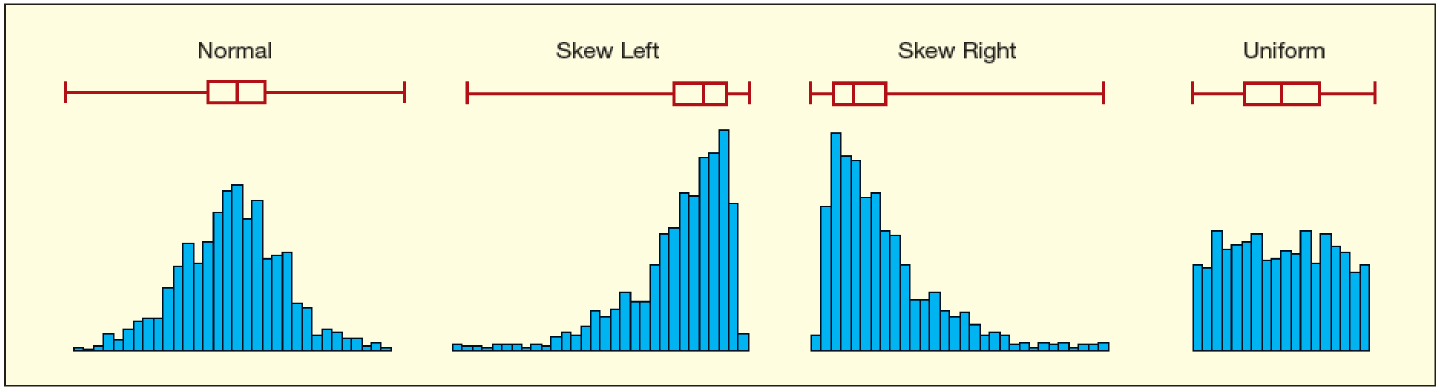
Percentiles, Quartiles, and Box Plots

Box Plots

- A box plot shows *variability* and *shape.*

FIGURE 4.27

Sample Boxplots from Four Populations (n = 1000)



Percentiles, Quartiles, and Box Plots

Box Plots: *Fences and Unusual Data Values*

- Use quartiles to detect unusual data points by defining ***fences*** using the following formulas:

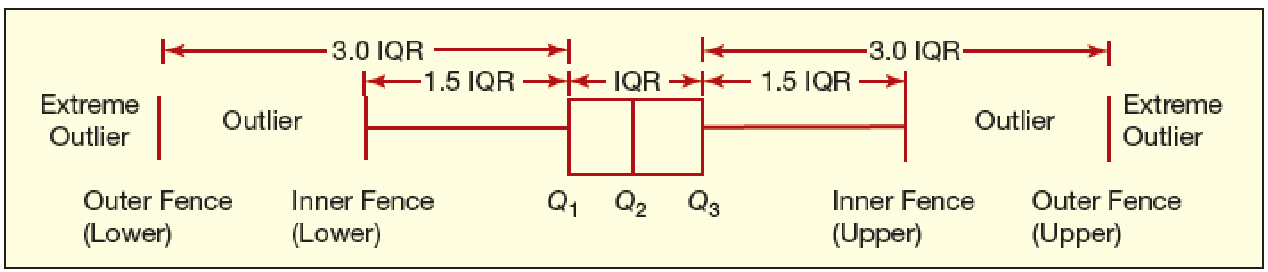
	Inner fences	*Outer fences:*
Lower fence	$Q_1 - 1.5\ (Q_3 - Q_1)$	$Q_1 - 3.0\ (Q_3 - Q_1)$
Upper fence	$Q_3 + 1.5\ (Q_3 - Q_1)$	$Q_3 + 3.0\ (Q_3 - Q_1)$

Percentiles, Quartiles, and Box Plots

Box Plots: *Fences and Unusual Data Values*

- Values outside the inner fences are ***unusual*** while those outside the outer fences are ***extreme outliers***. Here is a visual illustrating the fences:

A diagram helps to visualize the fence calculations. To get the fences, we merely add or subtract a multiple of the IQR from Q_1 and Q_3.



Percentiles, Quartiles, and Box Plots

Box Plots: *Fences and Unusual Data Values*

- For example, consider the P/E ratio data:

	Inner fences	*Outer fences:*
Lower fence:	107 – 1.5 (126 –107) = 78.5	107 – 3.0 (126 –107) = 50
Upper fence:	126 + 1.5 (126 –107) = 154.5	126 + 3.0 (126 –107) = 183

There is one outlier (170) that lies above the *inner fence. There are no* extreme outliers that exceed the *outer fence.*

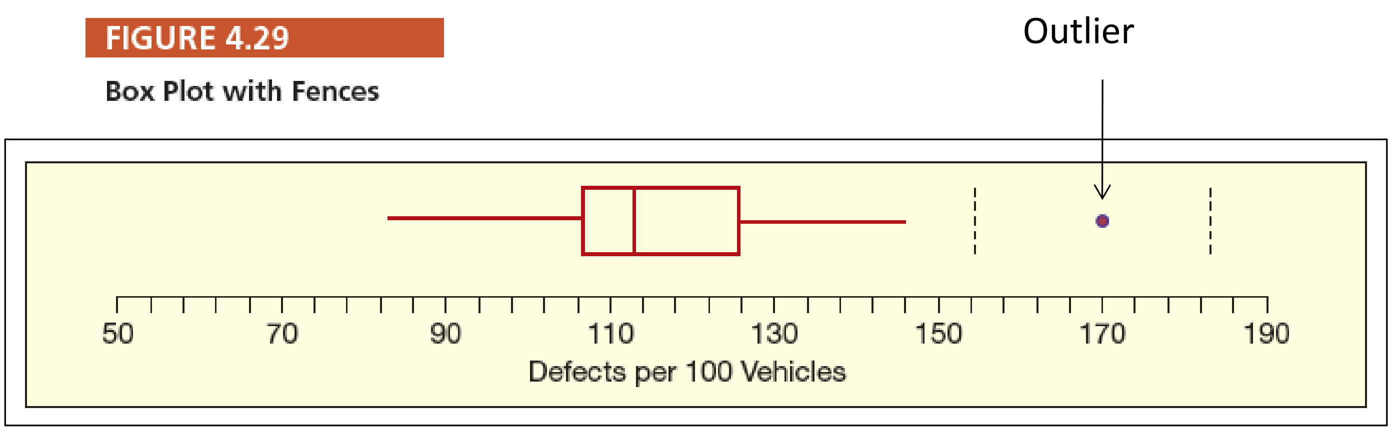
Percentiles, Quartiles, and Box Plots

Box Plots: *Fences and Unusual Data Values*

- Truncate the whisker at the fences and display unusual values and outliers as dots.

FIGURE 4.29

Box Plot with Fences



- Based on these fences, there is only one outlier.

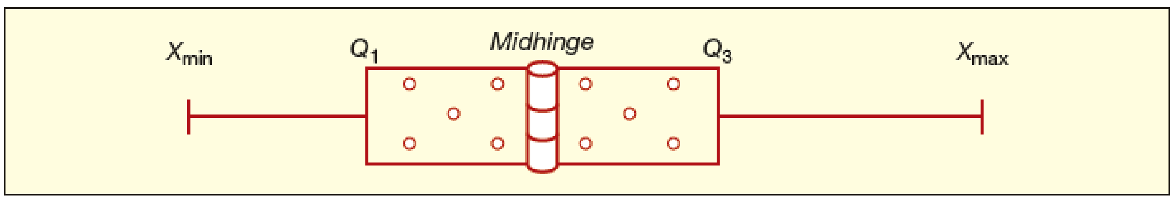
Percentiles, Quartiles, and Box Plots

Box Plots: *Midhinge*

Quartiles can be used to define an additional measure of center that has the advantage of not being influenced by outliers. The **midhinge** is the average of the first and third quartiles:

$$\text{Midhinge} = \frac{Q_1 + Q_3}{2}$$

The name "midhinge" derives from the idea that, if the "box" were folded at its halfway point, it would resemble a hinge:



Since the midhinge is always exactly *halfway* between Q_1 and Q_3 while the median Q_2 can be *anywhere* within the "box," we have a new way to describe skewness:

$$\begin{aligned}
&\text{Median} < \text{Midhinge} && \Rightarrow \text{Skewed right (longer right tail)}\\
&\text{Median} \cong \text{Midhinge} && \Rightarrow \text{Symmetric (tails roughly equal)}\\
&\text{Median} > \text{Midhinge} && \Rightarrow \text{Skewed left (longer left tail)}
\end{aligned}$$