MSCI152: Introduction to Business Intelligence and Analytics

Lecture 7: Measures of Spread

Lancaster University Management School

Overview

- Descriptive Measures: **Spread**

Summary Statistics

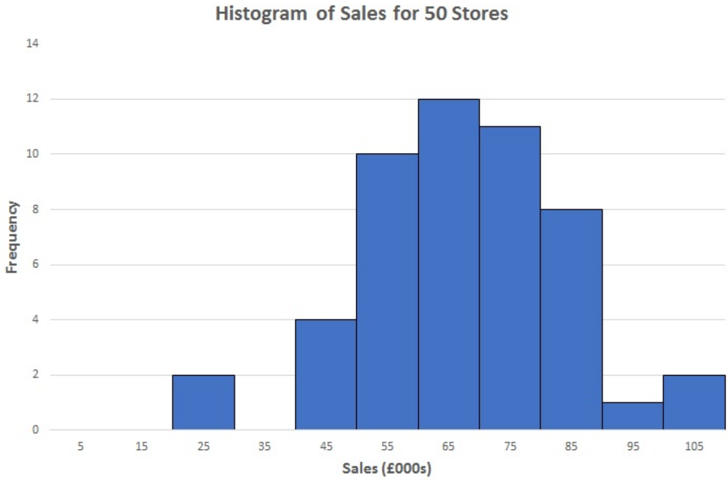
Location:

- Mean
- Median
- Mode

Spread:

- **Standard deviation**
- **Range**
- **Percentiles and Quartiles**

How spread out is this data?



Variability

Variability: How much difference is there between the values in the data

The more variable the data, the less relevant the average **may be**

E.g.: What would happen in a hospital which was equipped to treat the average number of emergency admissions per day?

E.g.: Average monthly demand for a product is 10,000 units. Should I plan to produce 10,000 units each month?

Sample Variance

Variance is important in some statistical methods:

- sum of the squared difference of each value from the mean
- variance is in **squared** units of what you are measuring

Sample Variance, s^2:

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

Excel: VAR.S()

First let us unpick the numerator on the R.H.S. of this equation

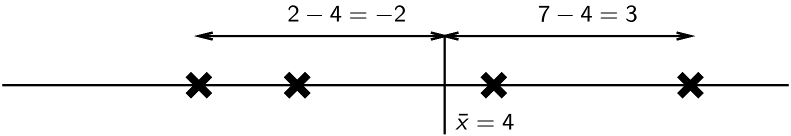
$\sum(x_i - \bar{x})^2$

What does $\sum(x_i - \bar{x})^2$ mean? Given data

$$x_1 = 2, \quad x_2 = 3, \quad x_3 = 7, \quad x_4 = 3, \quad x_5 = 5.$$

The mean, $\bar{x} = \frac{\sum x_i}{n} = \frac{20}{5} = 4$. Think about the difference

$$(x_i - \bar{x})$$



$\sum(x_i - \bar{x})^2$

Given data

$$x_1 = 2, \quad x_2 = 3, \quad x_3 = 7, \quad x_4 = 3, \quad x_5 = 5.$$

We have

$$\begin{aligned}
\sum_{i=1}^{5}(x_i - \bar{x})^2 &= (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + (x_4 - \bar{x})^2 + (x_5 - \bar{x})^2 \\
&= (2-4)^2 + (3-4)^2 + (7-4)^2 + (3-4)^2 + (5-4)^2 \\
&= (-2)^2 + (-1)^2 + 3^2 + (-1)^2 + 1^2 \\
&= 4 + 1 + 9 + 1 + 1 \\
&= 16
\end{aligned}$$

So the variance of these data is

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{16}{4} = 4$$

Sample Standard Deviation

Standard deviation is used more often in practice

- standard deviation is the square-root of the variance
- so is in the same units as what you are measuring

Sample Standard Deviation, s:

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

Excel: STDEV.S()

Alternative formula easier to calculate by hand:

$$s = \sqrt{\frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)}}$$

Sample Standard Deviation

Standard deviation is a measure of "*typically*" how far away the values are from the mean.

The higher the standard deviation, the more spread out the values are.

- Cannot be negative, larger value means more deviation

Sensitive to extreme values

Units are the same as the original data

- e.g., if the data is in cm then the standard deviation is in cm

Standard Deviation: Example

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
10	-10	100
12	-8	64
12	-8	64
14	-6	36
21	1	1
29	9	81
42	22	484
140	**SUM**	830

$$\bar{x} = \frac{\sum x_i}{n} = \frac{140}{7} = 20$$

e.g. for x_2,

$$(x_2 - \bar{x}) = (12 - 20) = -\mathbf{8}$$

and

$$(x_2 - \bar{x})^2 = (-8)^2 = \mathbf{64}$$

The standard deviation is

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{830}{6}} = \mathbf{11.76}$$

Standard Deviation: Alternative Formula

	x_i	$\bar{x}^2$
	10	100
	12	144
	12	144
	14	196
	21	441
	29	841
	42	1764
SUM	140	3630

We have,

$$n = 7, \ \sum x_i = 140, \ \sum x_i^2 = 3630$$

The standard deviation is then

$$\begin{aligned} s &= \sqrt{\frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)}} \\ &= \sqrt{\frac{7 \times 3630 - (140)^2}{7(7-1)}} \\ &= \sqrt{\frac{25,410 - 19,600}{42}} \\ &= \sqrt{138.333} \\ &= \mathbf{11.76} \end{aligned}$$

Calculating Spread for Populations

To calculate the standard deviation or variance for a population (i.e., we have a complete set of the data), divide by N rather than by $n-1$. In Excel: VAR.P() and STDEV.P().

Usually, Greek letter σ denotes the standard deviation for a population.

In our data set the standard deviation is

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{830}{6}} = 11.76$$

If this data was the population

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}} = \sqrt{\frac{830}{7}} = 10.89$$

If the sample n is large, this makes very little difference

Range

Range $=$ Maximum value $-$ Minimum value

Example:

$$10,\ \ 12,\ \ 12,\ \ 14,\ \ 21,\ \ 29,\ \ 42$$

Range $= 42 - 10 = \mathbf{32}$

Excel: MAX() - MIN()

Comparing Distributions

Coefficient of Variation, CV, is a ratio given by

$$CV = \frac{\text{Standard Deviation}}{\text{Mean}} = \frac{s}{\bar{x}}$$

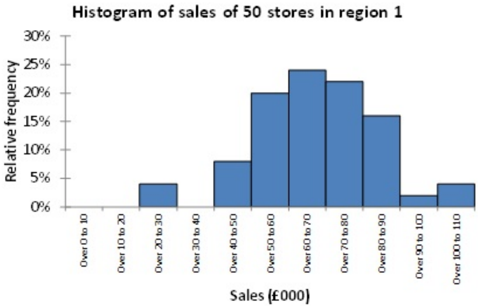
This can be useful in comparing populations with values of different magnitudes or different units

- For our data,

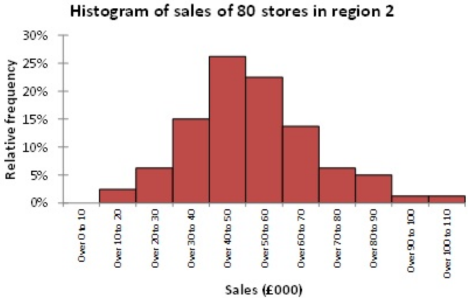
$$CV = \frac{11.76}{20} = 0.59 \text{ (or 59\%)}$$

EXCEL: STDEV.S() / AVERAGE()

Comparing region 1 and region 2



Region 1:
Mean = 67.7
Median = 68.7
Mode = Over 60 to 70
St. dev. = 16.7
CV = 25%



Region 2:
Mean = 51.3
Median = 49.9
Mode = Over 40 to 50
St. dev. = 17.6
CV = 34%

Percentiles

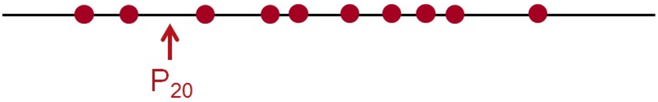
A **percentile value** P_k is the value where:

- $k\%$ of ordered observations are less than this value; and
- $(100 - k)\%$ ordered observations are more than this value
- Note: The **median** is P_{50}

e.g. For P_{20}

- 20% of ordered observations are less than the percentile value
- 80% of ordered observations are more than the percentile value

So, P_{20} for a sample of 10 ordered observations:



Calculating percentiles

There are various ways of calculating percentiles (and quartiles)

We will use the method from the textbook on the next slides

- Note that this method is consistent with the median calculation earlier

Excel: PERCENTILE.EXC(.,k), where k is the kth percentile.

Calculating Percentiles

Finding a certain percentile P_k

- e.g. P_{20} is the 20$^{\text{th}}$ percentile value

Position of percentile is: $P_k\% \times (n+1)$

E.g., If $n = 10$, the position is: $20\% \times 11 = 0.2 \times 11 = 2.2$

- So we want the 2.2$^{\text{th}}$ observation
- But we only have the 2$^{\text{nd}}$ and 3$^{\text{rd}}$ observations

Take value that is 0.2 (i.e., 20%) between 2$^{\text{nd}}$ and 3$^{\text{rd}}$ value using **linear interpolation**.

Example: Calculating Percentiles

Calculating P_{20} for ordered data:

$$8, \; 9, \; 12, \; 15, \; 16, \; 19, \; 21, \; 29, \; 34, \; 42$$

$n = 10$, so the position of P_{20} is

$$20\% \times 11 = 0.2 \times 11 = 2.2^{\text{th}} \text{ observation}$$

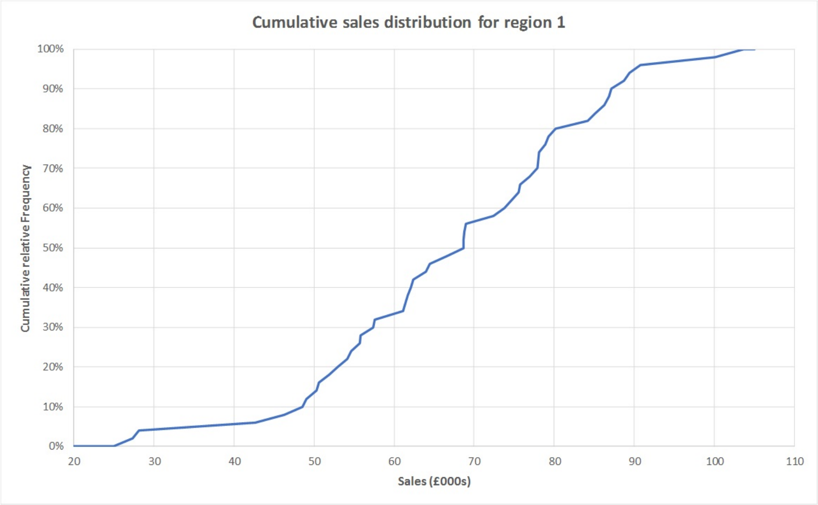
The value of P_{20} is

$$9 + 0.2 \times (12 - 9) = \mathbf{9.6}$$

Note: P_{20} lies 20% between "9" and "12".

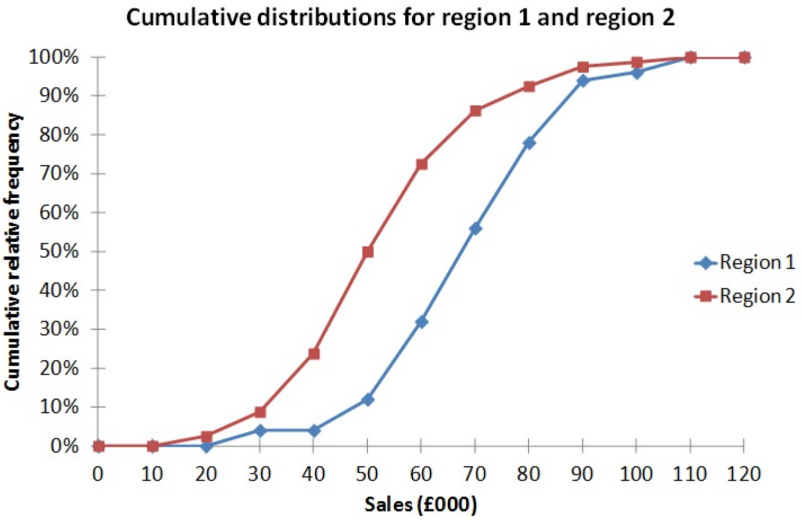
- The difference of the two is $(12 - 9) = 3$.
- 20% of 3 is $\mathbf{0.6}$.
- Adding this amount to "9" gives us P_{20}

Cumulative Frequency Chart (Ogive)



Excel XY (scatter) chart plotting cumulative frequency against the each data point

Cumulative Frequency Chart (Ogive)

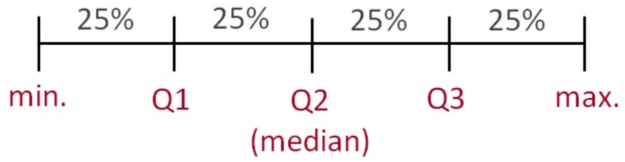


Excel XY (scatter) chart plotting cumulative frequency against the end point of each interval

Quartiles

Quartiles are just the 25, 50 and 75 percentiles, P_{25}, P_{50}, P_{75}, called **Q1**, **Q2**, **Q3**.

- Q1 is also called the **Lower Quartile**, [QUARTILE.EXC(.,1)]
- Q2 is the **Median** [MEDIAN() or QUARTILE.EXC(.,2)]
- Q3 is also called the **Upper Quartile** [QUARTILE.EXC(.,3)]



Divides the sorted data into 4 equal parts

The 5 values: min, Q1, Q2, Q3, max are called the **5 number summary**

Quartiles: Example 1

$$10, \quad 12, \quad 12, \quad 14, \quad 21, \quad 29, \quad 42$$

$n = 7$. Hence $n + 1 = 8$

Q1: Obs. $= 25\% \times 8 = 0.25 \times 8 = 2^{\text{nd}}$. Therefore

$$Q1 = 12$$

Q2: Obs. $= 50\% \times 8 = 0.5 \times 8 = 4^{\text{th}}$. Therefore

$$Q2 = 14$$

Q3: Obs. $= 75\% \times 8 = 0.75 \times 8 = 6^{\text{th}}$. Therefore

$$Q3 = 29.$$

Quartiles: Example 2

$$10, \quad 12, \quad 12, \quad 14, \quad 21, \quad 29, \quad 42, \quad 67$$

$n = 8$. Hence $n + 1 = 9$

Q1: Obs. $= 25\% \times 9 = 0.25 \times 9 = 2.25^{\text{th}}$. Therefore

$$Q1 = 12 + 0.25 \times (12 - 12) = 12$$

Q2: Obs. $= 50\% \times 9 = 0.5 \times 9 = 4.5^{\text{th}}$. Therefore

$$Q2 = 14 + 0.5 \times (21 - 14) = 17.5$$

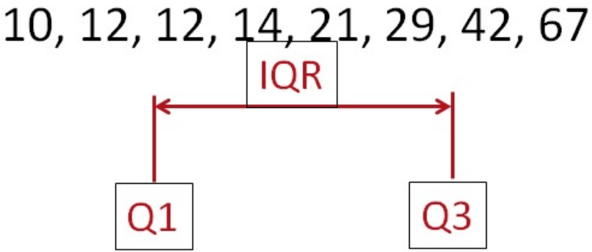
Q3: Obs. $= 75\% \times 9 = 0.75 \times 9 = 6.75^{\text{th}}$. Therefore

$$Q3 = 29 + 0.75 \times (42 - 29) = 38.75.$$

Inter-quartile range (IQR)

- $IQR = Q3 - Q1$
- This is the width of the middle 50% of the data
- In example on previous slide:

$$IQR = 38.75 - 12 = 26.75$$



Excel: QUARTILE.EXC(.,3) - QUARTILE.EXC(.,1)

Boxplot

A **boxplot** (or box-and-whisker-diagram) is a graph of a data set that consists of:

- a box from Q1 to Q3
- a vertical line showing the median
- lines from the sides of the box going to the last observation apart from outliers
- a special symbol (such as an asterisk) is used to identify outliers

Boxplot: 5 Number Summary and Outliers

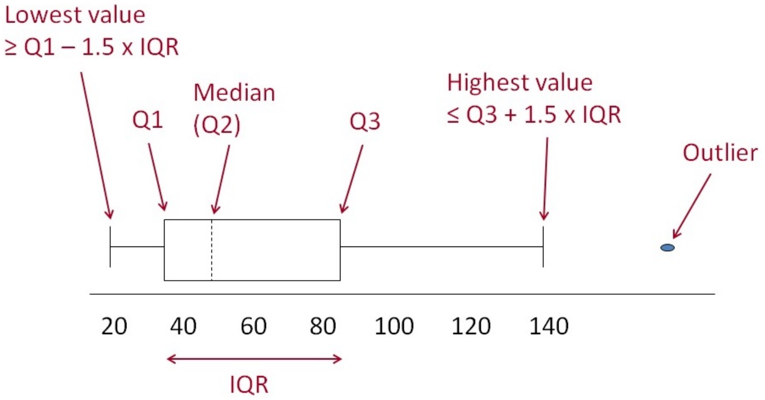
For a set of data, the **5-number summary** is:

- minimum value
- lower quartile, Q1
- median, Q2
- upper quartile, Q3
- maximum value

For a boxplot, a data point is an **outlier** if it is:

- above Q3 by an amount greater than $1.5 \times IQR$ or
- below Q1 by an amount greater than $1.5 \times IQR$

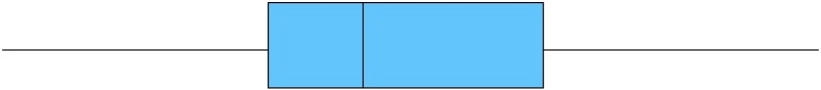
Boxplot



Lowest value
≥ Q1 – 1.5 x IQR
Median
(Q2)
Q1
Q3
Highest value
≤ Q3 + 1.5 x IQR
Outlier
20
40
60
80
100
120
140
IQR

Creating a Boxplot

- Create a scale covering the smallest to largest values
- Mark the location of the five numbers
- Draw a rectangle beginning at Q1 and ending at Q3
- **Check** if there are outliers; if yes, then mark all outliers and mark the smallest and largest values that are not outliers
- Draw a line in the box representing the median
- Draw lines from the ends of the box to the smallest and largest values that are not outliers

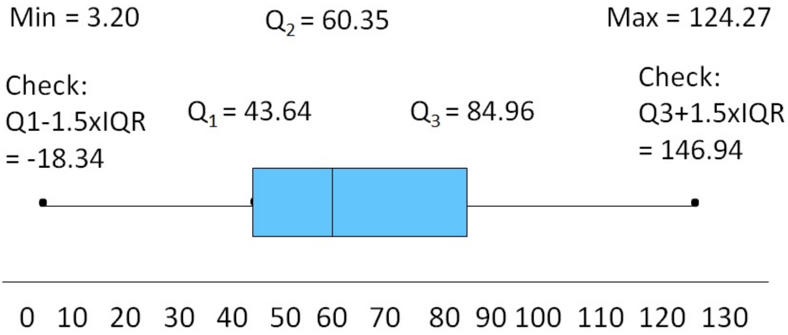


Boxplot Example 1

Data has observations $\{3.20, 5.15, \ldots, 124.27\}$ with:

- smallest observation $= 3.20$
- Q1 $= 43.64$
- Q2 (median) $= 60.35$
- Q3 $= 84.96$
- largest observation $= 124.27$

Boxplot Example 1

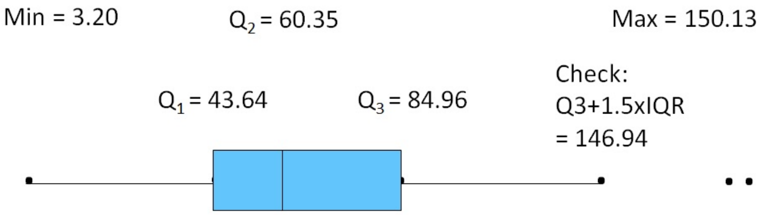


Boxplot Example 2

Data has observations $\{3.20, 5.15, \ldots, 124.27, 148.33, 150.13\}$ with:

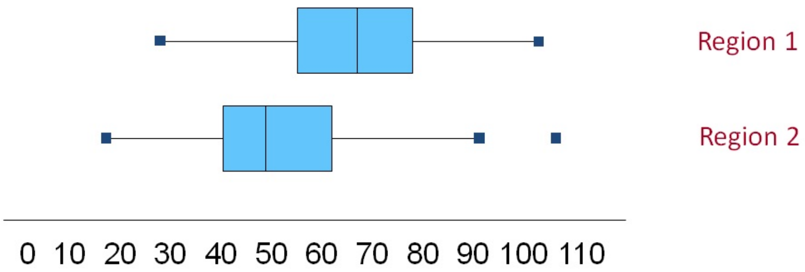
- smallest observation $= 3.20$
- Q1 $= 43.64$
- Q2 (median) $= 60.35$
- Q3 $= 84.96$
- largest observation $= 150.13$

Boxplot Example 2



So, 148.33 and 150.13 are outliers
highest that isn't an outlier is 124.27

Boxplots for Sales Data



Wrap up

Here we:

- Discussed summary statistics on **spread**

Next time:

- Anna will introduce relationships between variables: correlation and regression.