Statistics I

Railway Engineering Mathematics

Sheffield Hallam University

Lecture 23

Learning Outcomes

- Use EXCEL to obtain measures of central tendency for a data set: mean, median and mode.
- Use EXCEL to obtain measures of dispersion for a data set: range and standard deviation.

Introduction

Why study statistics?

- Whenever information is gathered about a process, the results of an experiment, financial patterns, the characteristics of standardised machine parts, etc, then it will be necessary to perform statistical calculations on those data in order to be able to interpret their meaning and infer conclusions.

Types of data

- **Qualitative** - nonnumeric data such as “favourite colour”, “hairstyle”, “blood type”.
- **Quantitative** - data that can be represented by a number. For example, “height”, “number of family members”.

Quantitative data are a collection of n measurements of a variable x, often written as $x_1, x_2, x_3, \ldots, x_n$. There are two types:

- **Discrete** – a variable that can be counted or that has a fixed set of values. For example, the number of visitors to a park (you can't have half or 0.2 of a person).
- **Continuous** – a variable that can be measured on a continuous scale. For example, “temperature” or “height”.

Measures of Central Tendency

There are three measures of central tendency:

- Mean: what we usually refer to as the average. EXCEL command: `=AVERAGE(...)`

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

- Median: this is the middle value in an ordered set of data: $\left(\frac{(n+1)}{2}\right)^{\text{th}}$ data point. EXCEL command: `=MEDIAN(...)`
- Mode: the most often occurring value. EXCEL: `=MODE(...)`

Example 1

The number of particles emitted by a radioactive source and detected by a Geiger counter in 40 consecutive period of 1 minute were measured/recorded as follows:

1	0	2	1	3	4	0	1	5	2
2	1	1	3	2	1	3	2	1	0
4	3	2	1	0	2	1	4	2	3
3	1	4	2	3	1	2	3	0	5

Summarise the data into a frequency table and find the mean, median and mode.

Example 1

x_i	Frequency f	fx_i	Cumulative frequency
0	5	0	5
1	11	11	16
2	10	20	26
3	8	24	34
4	4	16	38
5	2	10	40

$$\bar{x} = \frac{\sum fx_i}{\sum f} = \frac{81}{40} = 2.025$$

The median is the 20.5^{th} value, thus 2 from the CF.
The mode is 1.

Measures of Dispersion

There are several ways to measure how spread out the data is around the average:

- Range
- Interquartile range (IQR)
- Standard deviation

Range and Quartiles

The range of the data is simply the difference between the largest and smallest values.

Range

$$\text{Range} = \text{maximum value} - \text{minimum value}$$

EXCEL command: `=MAX(...)-MIN(...)`

Quartiles and IQR

Data can also be characterised by the upper and lower quartiles. Arrange the data values in increasing order, then . . .

Quartiles

The lower quartile (Q_1) is the median of the lower half of the data.

The upper quartile (Q_3) is the median of the upper half.

The difference between them is the **interquartile range** (IQR):

IQR

$$IQR = Q_3 - Q_1$$

EXCEL command:

```
=QUARTILE.EXC(...,3)-QUARTILE.EXC(...,1)
```

Example: Range and Quartiles

Example:

1 1 2 3 3 3 4 5 5 7 7 8 10 19 23

The range is the largest value minus the lowest value: $23 - 1 = 22$.

There are 15 data points, so Q_1 is the $\frac{1}{4}(n+1)^{\text{th}}$ value, i.e. the 4^{th} value. Thus: $Q_1 = 3$.

Q_3 is the $\frac{3}{4}(n+1)^{\text{th}}$ value, i.e. the 12^{th} value. Thus: $Q_3 = 8$.

And so we have: $IQR = Q_3 - Q_1 = 8 - 3 = 5$.

Measures of Dispersion

Population standard deviation: the data's average deviation from the mean (if one has access to **all** data).

$$\sigma = \sqrt{\frac{\sum(x-\overline{x})^2}{n}} \quad \text{ungrouped data}$$

EXCEL command: `=STDEV.P(...)`

OR Sample standard deviation: the data's average deviation from the mean (if one has access to a **sample** of all data).

$$\sigma = \sqrt{\frac{\sum(x-\overline{x})^2}{n-1}} \quad \text{ungrouped data}$$

EXCEL command: `=STDEV.S(...)`

Return to Example 1

Determine the population standard deviation for example 1 data using the step-by-step calculation and the in-built Excel function.

In this case,

$$n = 40$$

and we already determined that the mean of this sample is:

$$\bar{x} = 2.025$$

Return to Example 1

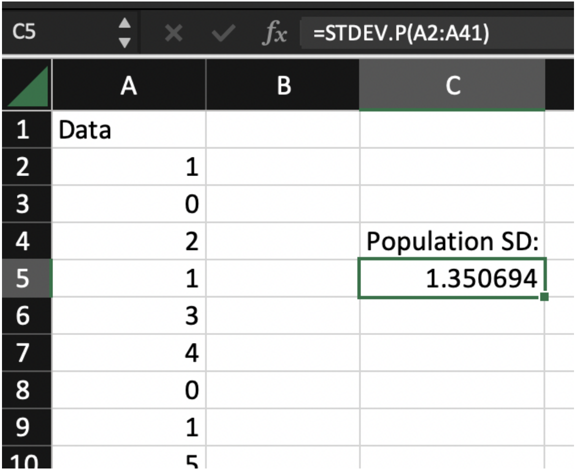
x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	f	$f(x_i - \bar{x})^2$
0	-2.025	4.1006	5	20.5030
1	-1.025	1.0506	11	11.5566
2	-0.025	0.0006	10	0.0060
3	0.975	0.9506	8	7.6048
4	1.975	3.9006	4	15.6024
5	2.975	8.8506	2	17.7013
				$\sum(x_i - \bar{x})^2 = 72.9741$

Then the sample standard deviation is:

$$\sigma = \sqrt{\frac{72.9741}{40}} = 1.351$$

Return to Example 1

In Excel:



C5 =STDEV.P(A2:A41)

	A	B	C
1	Data		
2	1		
3	0		
4	2		Population SD:
5	1		1.350694
6	3		
7	4		
8	0		
9	1		
10	5		