

# Session 7.4: Trainee Notes

## Quantitative Data Distribution

### Overview

Summary statistics are used to convey most of the useful information about a large set of numbers (the data) in a simple but sensible fashion.

### Content

1. Measures of location
2. Measures of dispersion
3. Confidence intervals

### Learning Objective

After completing this session, you will be able to:

- Assess appropriate summary statistics for all types of data
- Present summary data in a useful and meaningful way

### 3. Measures of location

**Question:** What is the "average" reaction time?

There are three common measures of location:

#### (Arithmetic) mean

- Sum of variable divided by the sample size
- Denoted by  $\bar{x}$ , and called the sample mean

#### *Advantages*

- Uses all of the data
- Suitable for further analysis

#### *Disadvantages*

- Affected by a few extreme values

#### Median

- The middle of the ordered sample of data.
- Crudely speaking the median cuts the ordered data set in half: 50% below and 50% above.

#### *Advantages*

- Not affected by a few extreme values

#### *Disadvantages*

- Does not use all of the data
- Not particularly suitable for further analysis

#### Mode

- The single value that occurs most frequently
- The most common value

#### *Advantages*

- Simple

#### *Disadvantages*

- Bimodal and multimodal data is common
- Not particularly suitable for further analysis

### Exercise 7.4

Open the Excel file 'Copy of Datasets.xls'

Open 'Cohort data 2' worksheet

Calculate the mean, mode and median for 'age'

**data → data analysis→descriptive statistics**

## 4. Measures of dispersion

**Question:** how spread out are the reaction times?

There are three common measures of dispersion:

### Range

- Maximum - minimum

#### *Advantages*

- Easy to understand

#### *Disadvantages*

- Does not use all of the data
- Not particularly suitable for further analysis
- Affected by extreme values

### Interquartile range (IQR)

- Regard the median as the 50% point
- Describe spread using the 25% (lower quartile) and 75% (upper quartile) points,
- Interquartile range: third-first quartile
- It is the range over which the "middle 50%" of the sample is spread

#### *Advantages*

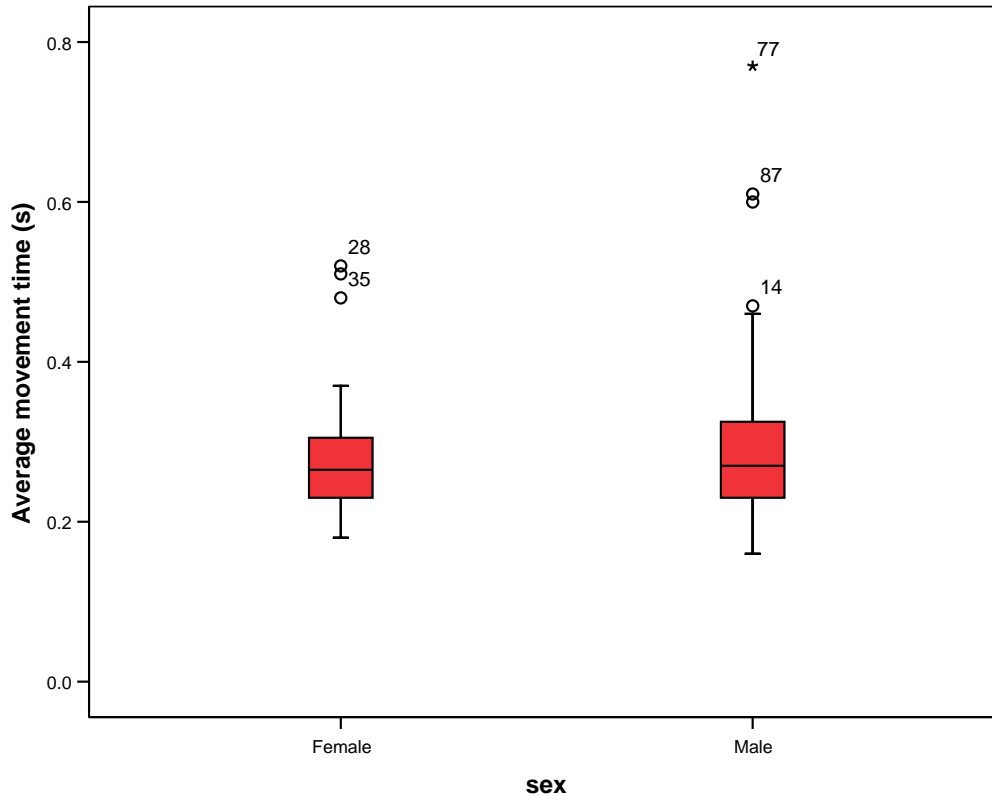
- Not affected by a few extreme values

#### *Disadvantages*

- It does not use all the data
- Not particularly suitable for further analysis

## Boxplots

Boxplots are a pictorial way of representing the quartile values. This is useful for comparing distributions.



In this diagram, outliers are marked as circles and extreme values are marked as stars.

## Variance and standard deviation

- These quantify how spread out the values are about the mean.
- The (sample) **variance,  $s^2$** , is defined to be the "average" of the squared differences from the mean:

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

- Note the use of the divisor  $n-1$ .

An undesirable feature of the variance is that its units of measurement are squared.

- (Sample) **standard deviation,  $s = \sqrt{s^2}$**
- Same measurement units as the original data.
- Think of the standard deviation as describing how much (on "average") the individual times differ from the mean.

### *Advantages*

- Uses every value
- It is suitable for further analysis

### *Disadvantages*

- Affected by a few extreme observations
- Tends to be useful only when the distribution of data is approximately normal

### *Standard Deviation and Empirical Rules*

For almost any set of quantitative data:

- **60%-75%** of the data will be located within a distance of **one** standard deviation of the mean  $\rightarrow x \pm s$
- **90%-98%** of the data will be located within a distance of **two** standard deviations of the mean  $\rightarrow x \pm 2s$
- **99%-100%** of the data will be located within a distance of **three** standard deviations of the mean  $\rightarrow x \pm 3s$

#### **Example**

For the move\_ave variable  $\bar{x} = 0.290$  and  $s = 0.0893$

$$x \pm 2s = 0.290 \pm 0.179 = 0.111 \text{ to } 0.469$$

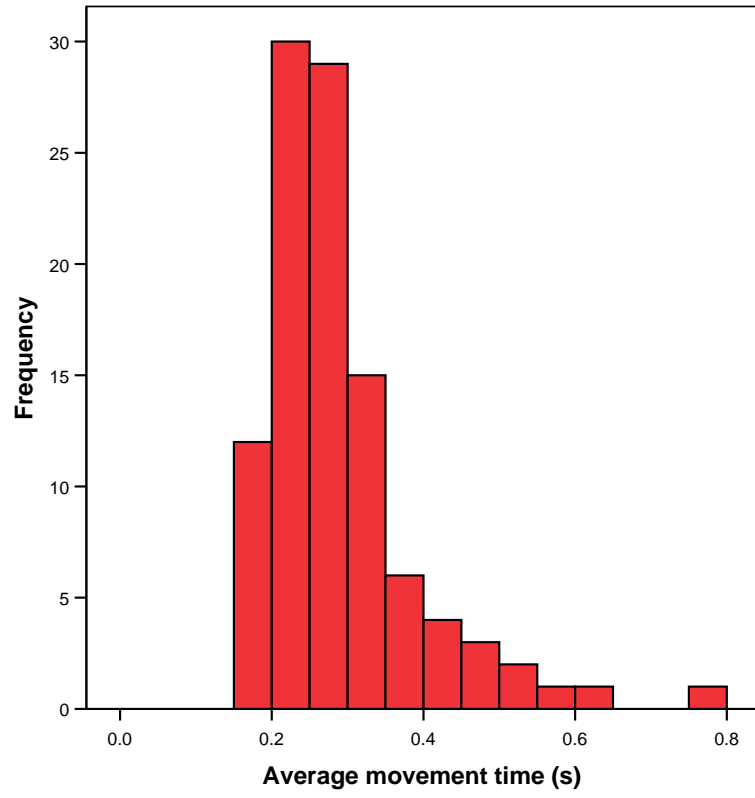
This interval contains 95.2% of the sample of 147 reaction times.

### **Recommendations on the use of Summary Statistics**

Common types of distributions:

- Roughly symmetrical and unimodal (single "peak")
  - Use mean and standard deviation.
- Markedly skew (and unimodal)
  - Use median and interquartile range.

What summary statistics might we use with the reaction time example?



## 5. Confidence intervals for estimates

An estimated statistic has little meaning without knowing how certain we can be about that value.

You should estimate any statistic as a **confidence interval** which is *highly likely* to contain the statistic and not just a point.

This interval takes into account the size of the sample and the underlying variability.

The most common statistic is of course the mean, and the easiest confidence interval to calculate is the confidence interval around the mean of some normally distributed data. The **standard error** is used as the measure of underlying variability.  $s.e. = \sigma/\sqrt{n}$

The 95% confidence interval of mean  $\mu$  is derived by:

$$\bar{x} \pm 1.96 * \sigma/\sqrt{n}$$

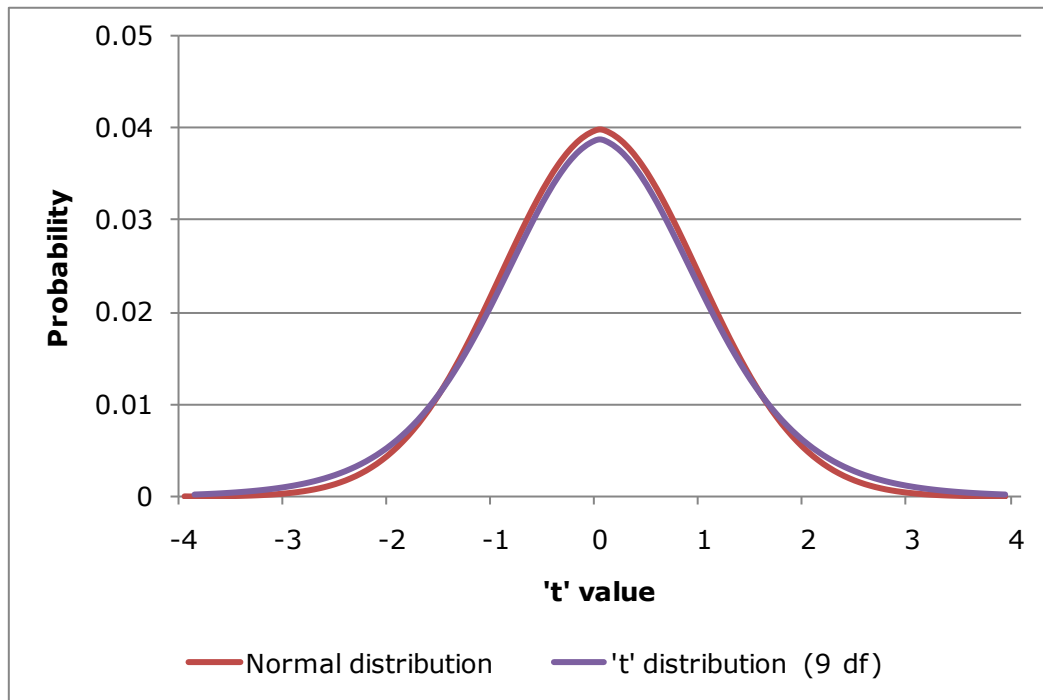
However,  $\sigma^2$  is not usually known and so is estimated by  $s^2$ . The 95% confidence interval for  $\mu$  is then:

$$\bar{x} \pm t_{n-1} * s/\sqrt{n}$$

where  $t_{n-1}$  is the 5% level for the t-distribution with (n-1) degrees of freedom. We use the t-distribution as  $\sigma^2$  is estimated. As n gets large, the t distribution gets more and more similar to the normal distribution

**'t' values for alpha (2-sided)**

Degrees of freedom	0.05	0.01
5	2.57	4.03
10	2.23	3.17
15	2.13	2.95
20	2.09	2.85
25	2.06	2.79
30	2.04	2.75
35	2.03	2.72
40	2.02	2.70
45	2.01	2.69
50	2.01	2.68
infinity	1.96	2.58



## Interpretation

If one obtained many different samples, all from the same population, then 95% of the samples would contain  $\mu$ , i.e. there is a 95% chance that a single interval contains  $\mu$ .

The use of a 95% confidence interval is most common, but you could have a 90% or 99% interval, a 90% interval is narrower than a 95% interval – you are less sure that the interval contains the population mean ( $\mu$ ).

The calculation is the same as described above, but the multiplier is different.