Introduction to Statistical Modelling

STAT2507
Chapter 2-1
Describing Data with Numerical Measures

IMPORTANCE OF NUMERICAL MEASURES

- Graphs methods may not always be sufficient for describing data
- Hard to measure the similarities and differences between graphs
- Difficult to describe the degrees of differences
- One way to over come these problems is to use numerical measures which can be calculated for either a sample or population.

NUMERICAL MEASURES

- A **parameter** is a numerical descriptive measure calculated for a **population**
- A **statistics** is a numerical descriptive measure calculated for a **sample**

NUMERICAL MEASURES

- Measures of Centre : locates the centre of distribution along horizontal axis
  - Mean
  - Median
  - Mode
- Measures of Variability: provides information about the spread of the data
  - Range
  - Variance
  - Standard Deviation

Numerical Measures

- Measures of Relative Standing: provides position of one observation relative to others
  - Z-score
  - Percentile (large data sets)
  - Inter-quartile Range
- The Five-number summary and the Box plot

MEASURES OF CENTRE: MEAN

- The arithmetic mean or average of a set of n measurements is equal to the sum of the measurements divided by n.
- Population Mean:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

where x_i- measurements on all units in the population

- Sample Mean: $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Measures of Centre: Mean

- Example: Find the mean of n = 5 measurements 2, 9, 11, 5, 6

$$\text{mean} = \frac{2+9+11+5+6}{5}$$

$$= 6.6$$

- An important use of $\bar{x}$ is as an estimator of unknown population mean μ. But, sample mean changes from sample to sample.
- When there are extremely small or extremely large observations in the sample, the sample mean is drawn toward the direction of the extreme measurements

CENTRE OF MEASURE: MEDIAN

- Median m of set of n measurements is the value of x that falls in the middle position when the measurements are ordered from smallest to largest
- Position of median in ordered data is 0.5(n+1). If this value ends in 0.5, average of two adjacent values is the median
- Median is less sensitive to extreme values or outliers

Centre of Measure: Median

- If the distribution is strongly skewed by one or more extreme values, use median rather than mean as centre of measure.
- If the distribution is symmetric, mean and median are equal

EXERCISE

2.1) Find the median of following data.

a) The set : $2, 4, 9, 8, 6, 5, 3$

b) The set: 2, 4, 9, 8, 6, 5

Measure of centre: Mode

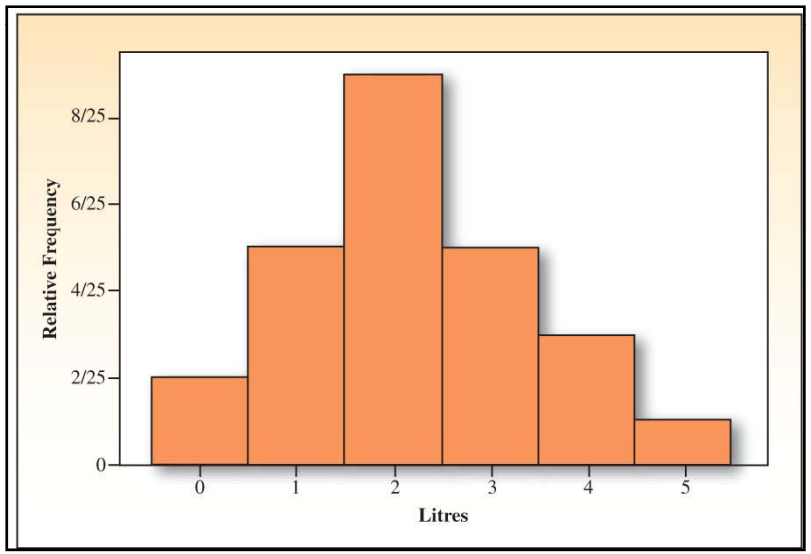
- **Mode**: the measurement which occurs most frequently
  - In the set: 2, 4, 9, 8, 8, 5, 3
    - The mode is **8**, which occurs twice
  - In the set: 2, 2, 9, 8, 8, 5, 3
    - There are two modes—**8** and **2** (bimodal)
  - In the set: 2, 4, 9, 8, 5, 3
    - There is no mode; each value is unique

EXERCISE

2.2) The number of litres of milk purchased by 25 households: 0 0 1 1 1 1 1 2 2 2 2 2 2 2 2 2 3 3 3 3 3 4 4 4 5. Find the mean, median, mean. Total milk purchased = 55.

EXERCISE CONT'D

- The data is approximately mound-shaped

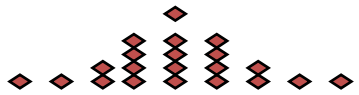


EXTREME VALUES

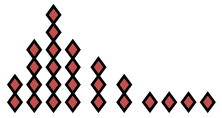
- The mean is more easily affected by extremely large or small values than the median
- When a distribution is symmetric, the mean and the median are equal
- The median is often used as a measure of centre when the distribution is skewed

EXTREME VALUES

- If a distribution is skewed to the right, the mean shifts to the right; if a distribution is skewed to the left, the mean shifts to the left
  - Symmetric: Mean ≈ Median



  - Skewed right: Mean > Median



  - Skewed left: Mean < Median

EXERCISE

2.3) The number of Tim Hortons within 5 kilometres (km) of city centre (downtown) in cities of South-western Ontario is given below:

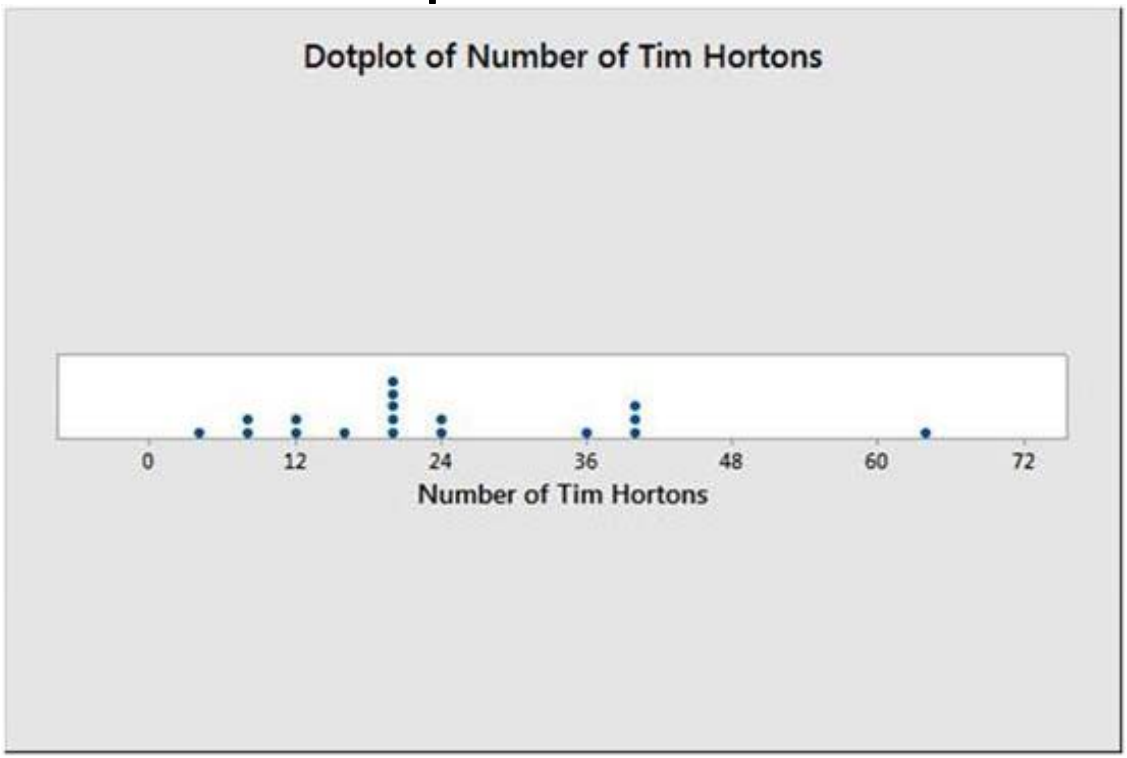
22	16	12	19	40	40
21	10	20	19	4	65
39	20	6	34	23	7

a) Find the mean, the median and the mode

b) Compare the median and the mean. What can you say about the shape of the distribution?

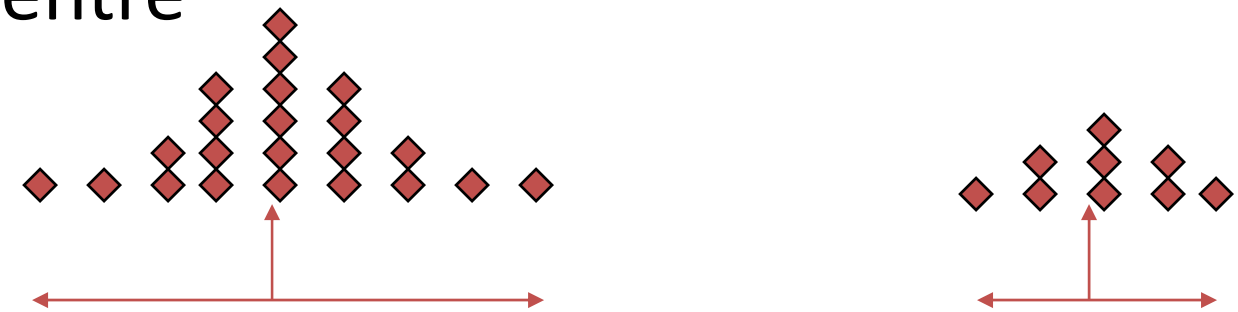
EXERCISE

c) Draw a dotplot for the data. Does this confirm your conclusion about the shape of the distribution from part b?



MEASURES OF VARIABILITY

- Datasets with the same centre look different because of the way number spread out from the centre.
- Measures of Variability: a measure along the horizontal axis of the data distribution that describes the **spread** of the distribution from the centre



THE RANGE

- **Range (R)**: the difference between the largest and smallest measurements in a set
  - **Example**: a botanist records the number of petals on five flowers: **5, 12, 6, 8, 14**
- The range is **R = 14 – 5 = 9**
- Range uses only 2 of n (=5) measurements
- Range is easy to calculate and interpret, but is an adequate measure of variation for small sets of data.

THE VARIANCE

- **Variance**: a measure of variability that uses all the measurements; it measures the average deviation of the measurements about their mean
  - Variance of a population

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

THE VARIANCE

- The variance of a sample of n measurement is the sum of squared deviations of the measurement about their mean, divided by (n-1)

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

THE VARIANCE

- Computing formula for calculating s^2

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

THE STANDARD DEVIATION

- In calculating the variance, we squared all of the deviations, and in doing so changed the scale of the measurements
- To return this measure of variability to the original units of measure, we calculate the **standard deviation**, the positive square root of the variance
  - Population Standard Deviation

$$\sigma = \sqrt{\sigma^2}$$

  - Sample Standard Deviation

$$S = \sqrt{S^2}$$

EXERCISE

2.4) Calculate the standard deviation of these 5 measurements: 5, 12, 6, 8, 14

	x_i	x_i^2	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
	5	25	-4	16
	12	144	3	9
	6	36	-3	9
	8	64	-1	1
	14	196	5	25
Sum	45	465	0	60

EXERCISE

2.5) An article in Archaeometry involved an analysis of 26 samples of Romano-British pottery found at four different kiln sites in the United Kingdom. The samples were analyzed to determine their chemical composition. The percentage of iron oxide in each of five samples collected at the Island Thorns site was: 1.28, 2.39, 1.50, 1.88, 1.51

EXERCISE

a) Calculate the range
b) Calculate the sample variance and the standard deviation using the computing formula
c) Compare the range and the standard deviation. The range is approximately how many standard deviations?

REMARKS

- s and σ are always greater than or equal to zero
- The larger the value of S^2 or s, the greater the variability of the data set
- If s is equal to zero, all the measurements must have the same value

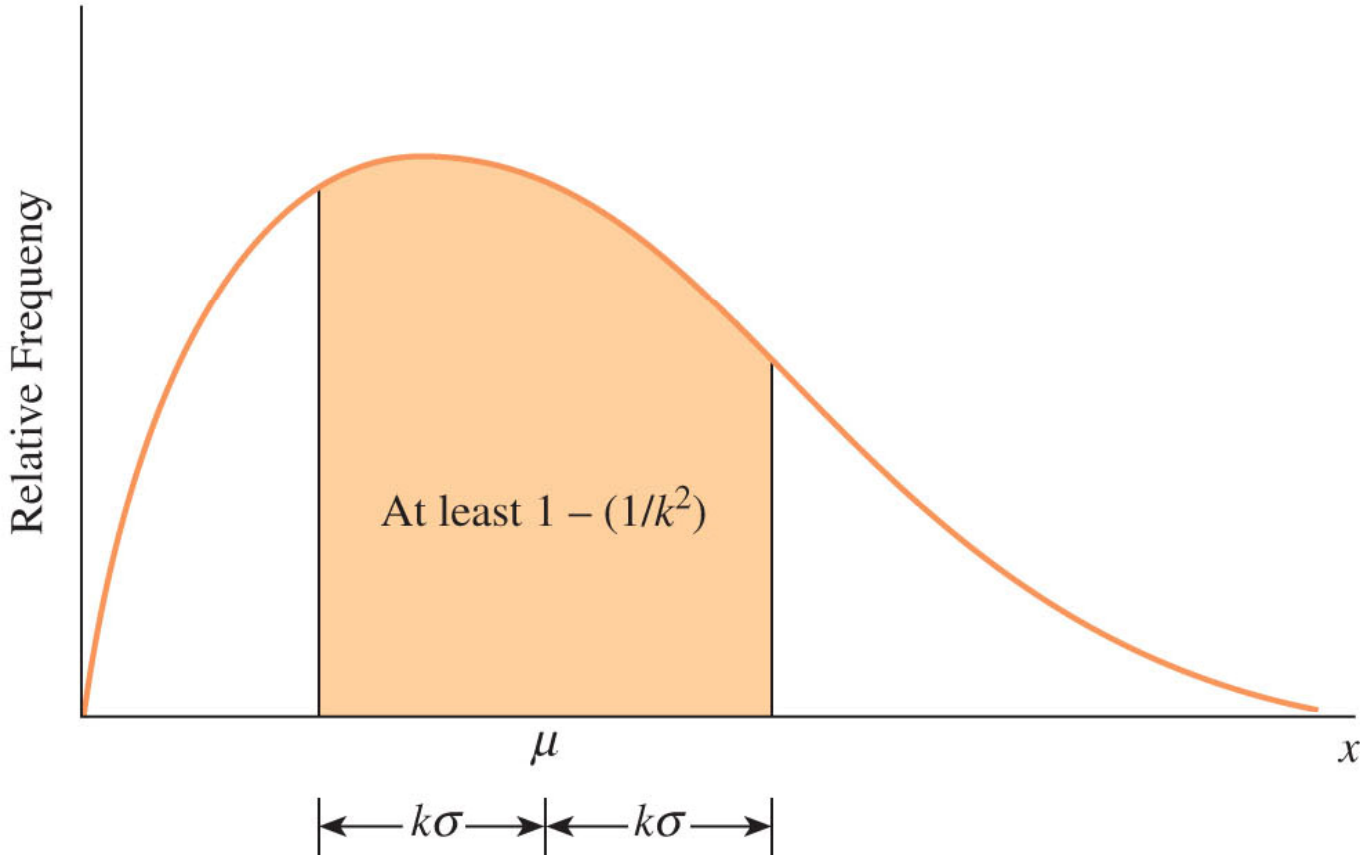
REMARKS

- In order to measure the variability in the same units as the original observations, we compute the standard deviation.
- Why divide by n-1?
  - The sample standard deviation s is often used to estimate the population standard deviation σ
  - Dividing by n – 1 gives us a better estimate of s

TCHEBYSHEFF'S THEOREM

- Given a number k >= 1, and a set of n measurements, at least [1-(1/k^2)] of the measurements will lie within k standard deviations of their mean
- Tchebysheff's theorem applies to any set of measurements and can be used to describe either a sample or population

TCHEBYSHEFF'S THEOREM



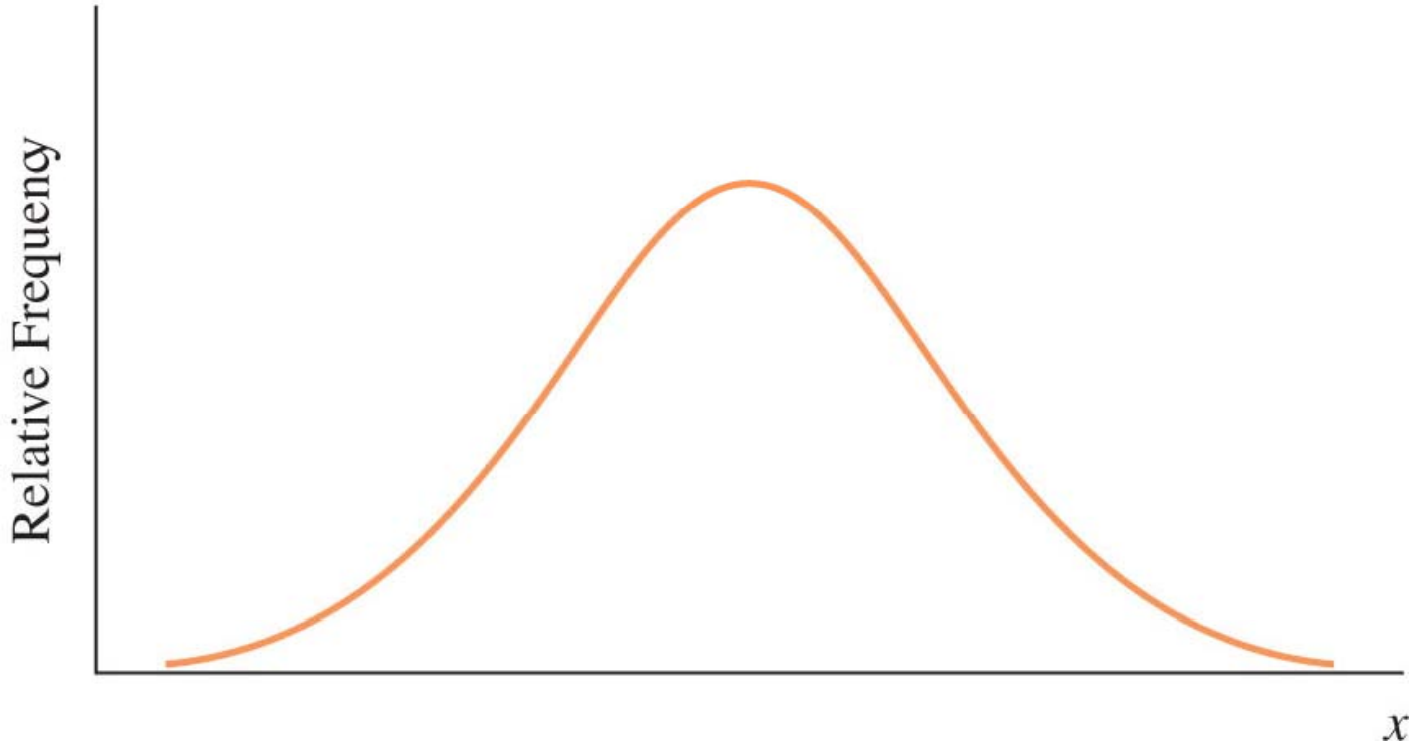
TCHEBYSHEFF'S THEOREM

- It can be used for sample or population
  - If k=1, at least none of measurements lie with 1 standard deviation of the mean
  - If k = 2, at least $1 - (½)^2 = 3/4$ of the measurements are within 2 standard deviations of the mean – in the interval $\mu-2\sigma$ to $\mu+2\sigma$
  - If k = 3, at least $1 - (⅓)^2 = 8/9$ of the measurements are within 3 standard deviations of the mean – in the interval $\mu-3\sigma$ to $\mu+3\sigma$

EMPIRICAL RULE

- Given a distribution of measurements that is approximately **mound-shaped**:
  - The interval $\mu \pm \sigma$ contains approximately 68% of the measurements
  - The interval $\mu \pm 2\sigma$ contains approximately 95% of the measurements
  - The interval $\mu \pm 3\sigma$ contains approximately 99.7% of the measurements

Mound-Shaped Distribution



EXERCISE

2.6) The length of time for a worker to complete a specified operation averages 12.8 minutes with a standard deviation of 1.7 minutes. If the distribution of times is approximately mound-shaped, what proportion of workers will take longer than 16.2 minutes to complete the task?

APPROXIMATING S

- Tchebysheff's theorem and Empirical rule can be used to detect **gross** errors in the calculation of s. These tools tells that most of measurements lie within the two std of the mean. Therefore, R ≈ 4-6 S
- To approximate the standard deviation of a set of measurements, we can use

S ≈ R/4 or S ≈ R/6 for a large data set

EXERCISE

2.7) 25 lesson plan were scored on a scale of 0 to 34:

26.1	26.0	14.5	29.3	19.7
22.1	21.2	26.6	**31.9**	25.0
15.9	20.8	20.2	17.8	13.3
25.6	26.5	15.7	22.1	13.8
29.0	21.3	23.5	22.1	**10.2**

$\overline{x}$ = 21.6, S = 5.5

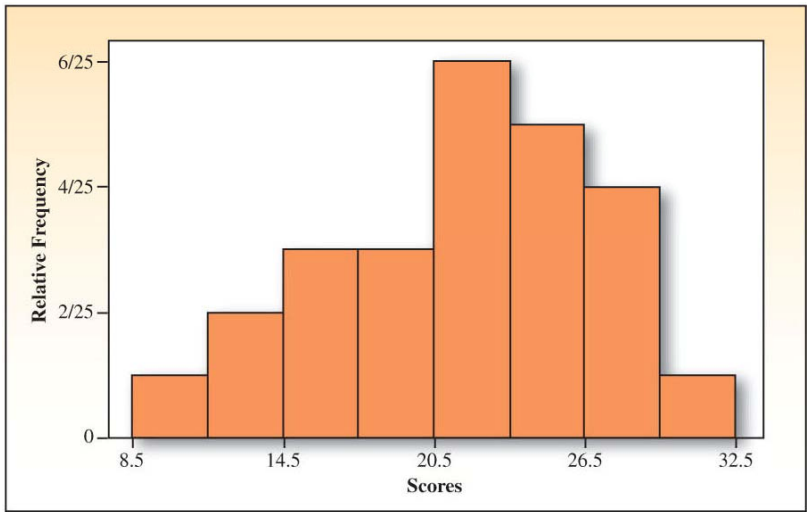
EXERCISE

k	$\bar{x} \pm ks$	Interval	Proportion in Interval	Tchebysheff	Empirical
1	21.6±5.5	16.1,17.1	16/25=0.64	At least 0	≈0.68
2	21.6±11	10.6,32.6	24/25=0.96	At least 0.75	≈0.95
3	21.6±16.5	5.1,38.1	25/25=1	At least 0.89	≈0.997

EXERCISE

Calculate the approximate standard deviation.

Histogram for the data



EXERCISE

2.8) Suppose that the resting breathing rates for university –age students have a relative frequency distribution that is mound-shaped with mean equal to 12 and a standard deviation of 2.3 breaths per minute. What fraction of all students would have breathing rates in the following intervals?

EXERCISE

a) 9.7 to 14.3 breaths per minute
b) 7.4 to 16.6 breaths per minute
c) More than 18.9 or less than 5.1 breaths per minute.