Introduction to Statistical Modelling

STAT2507
Chapter 2-2
Describing Data with Numerical Measures

MEASURES OF RELATIVE STANDING

- How many standard deviation from the mean does the measurement lie?
- This is measured by z-score

$$Z_{score} = \frac{x - \bar{x}}{S}$$

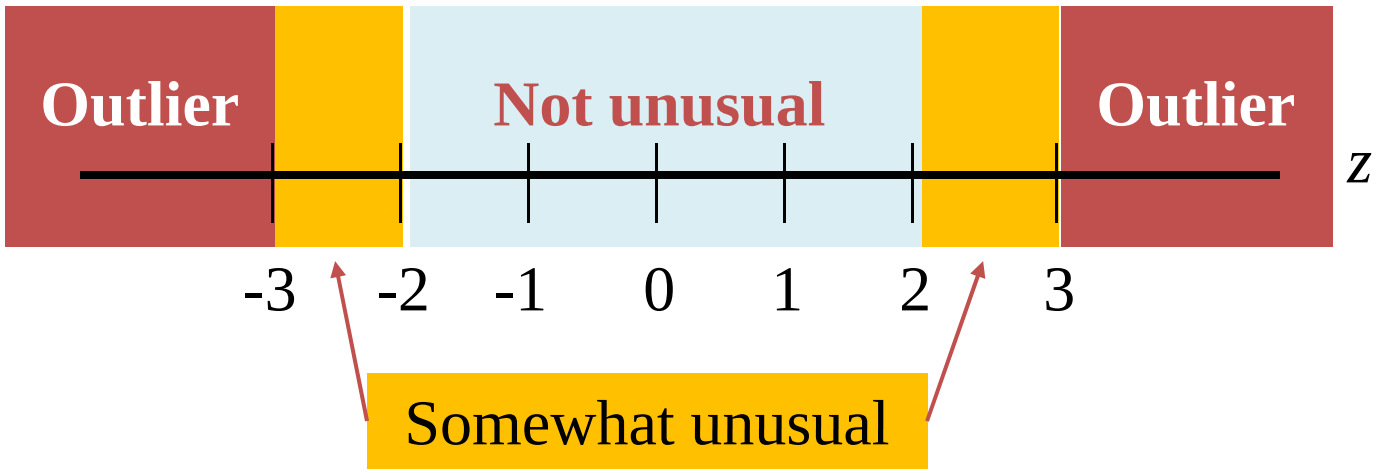
2.9) x = 9, $\bar{x}$ = 5, s=2. How many standard deviation does the x=9 from mean?

Measures of Relative Standing

- From Tchebysheff's theorem and Empirical Rule
  - At least 3/4 and more likely 95% of measurements lie within 2 standard deviations of the mean
  - At least 8/9 and more likely 99.7% of measurements lie within 3 standard deviations of the mean

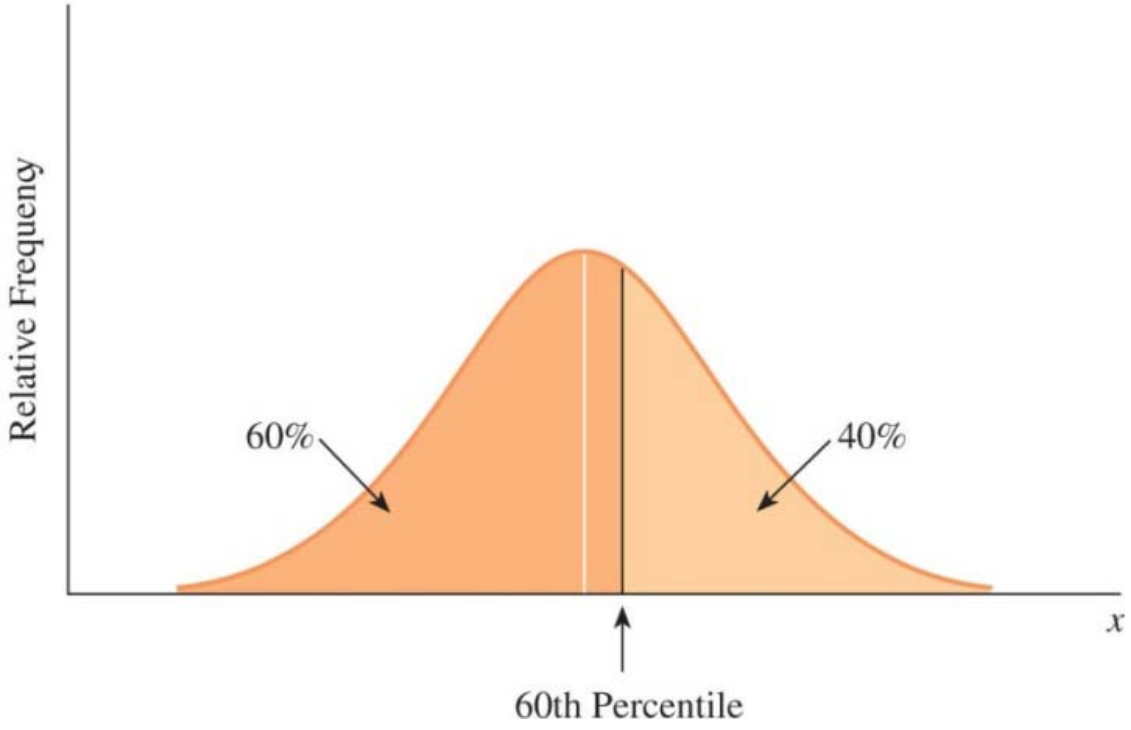
Z-SCORES

- |*z*-score|≤ 2 (-2≤ Z-score≤ 2) are not unusual
- 2 < z-score ≤ 3 and -3≤ z-score < -2 are somewhat unusual
- |z-score| > 3 **outlier**



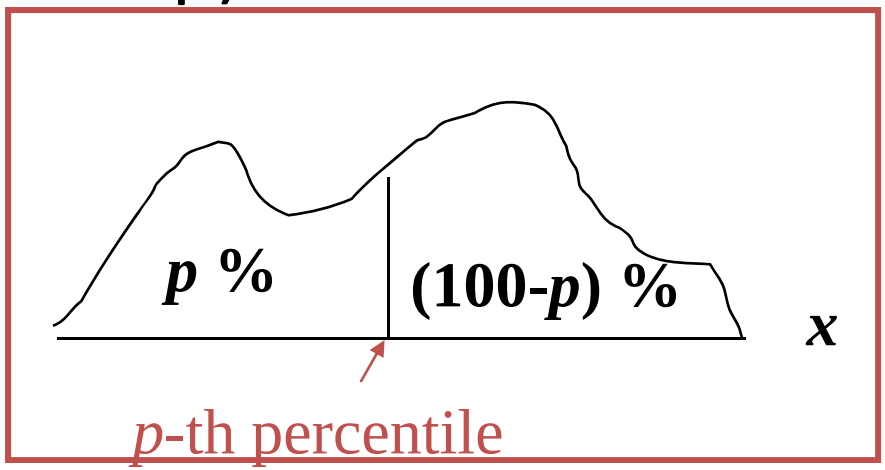
MEASURES OF RELATIVE STANDING

- A **percentile** is another measure of relative standing and most often used for large datasets



PERCENTILE

- A set of n measurements on the variable x has been arranged in order of magnitude. The p^{th} percentile is the value of x that is greater than p% of the measurement and is less than the remaining (100 – p)%

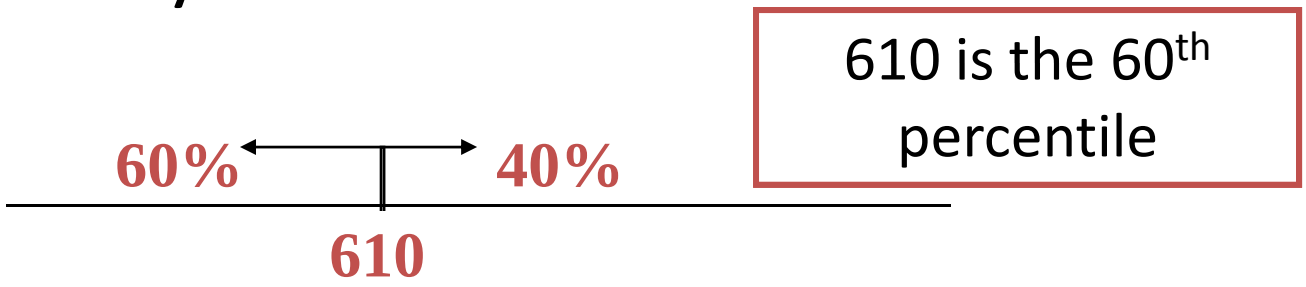


EXAMPLE

- Suppose you have been notified that your score of 610 on the Verbal Graduate Record Examination placed you at the 60^{th} percentile in the distribution of scores. 4000 students took this examination. Where does you score of 610 stand in relation to the scores of others who took the examination?

EXAMPLE CONT'D

- Scoring at the 60th percentile mean that 60% of all examination scores were lower than yours and 40% were higher.
- Total number of students = 4000. So, 2400 of them scored lower than you and 1600 scored higher than you.



PERCENTILE AND QUARTILES

- 50^{th} Percentile — Median/Second Quartile (Q_2)
- 25^{th} Percentile — Lower Quartile/First Quartile (Q_1)
- 75^{th} Percentile — Upper Quartile/Third Quartile (Q_3)

MEASURES OF RELATIVE STANDING

- A set of n measurements on the variable x has been arranged in order of magnitude
  - **Lower Quartile (first quartile, Q_1):** is the value of x that is greater than ¼ of the measurements and is less than the remaining ¾ ; 25^{th} percentile - Q_1
  - Lower Quartile, Q_1, is the value of x in position, $P_1 = 0.25(n+1)$

MEASURES OF RELATIVE STANDING

- A set of n measurements on the variable x has been arranged in order of magnitude
  - **Median (second quartile, Q_2):** is the value of x that is greater than ½ of the measurements and is less than the remaining ½; 50^{th} percentile - Q_2
  - Median, Q_2, is the value of x in position, $P_2 = 0.5(n+1)$

MEASURES OF RELATIVE STANDING

- A set of n measurements on the variable x has been arranged in order of magnitude
  - **Upper Quartile (third quartile, Q_3):** is the value of x that is greater than ¾ of the measurements and is less than then remaining ¼; 75^{th} percentile - Q_3
  - Upper Quartile, Q_3, is the value of x in position, $P_3=0.75(n+1)$

MEASURES OF RELATIVE STANDING

- When 0.25(n+1) and 0.75(n+1) are not integers, quartiles are found by interpolation, using the values in the two adjacent positions.
- The range of the "middle 50%" of the measurements is the **interquartile range**, **IQR** $= Q_3 - Q_1$
- Useful for large data sets.

EXAMPLE

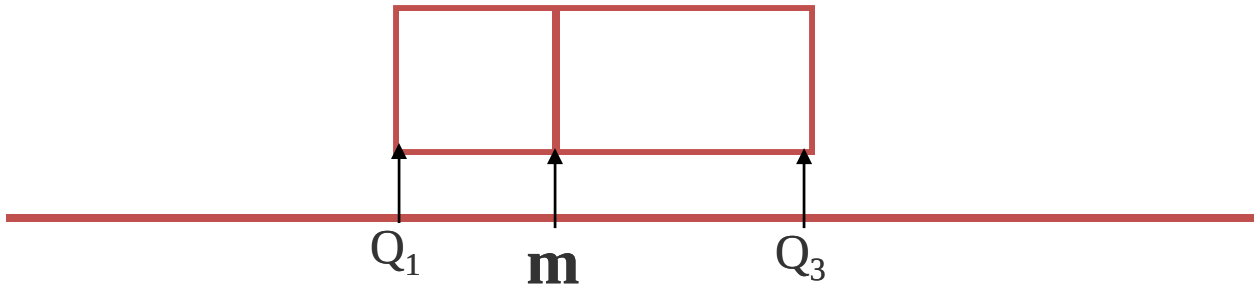
2.10) The prices ($) of 18 brands of walking shoes: 40, 60, 65, 65, 65, 68, 68, 70, 70, 70, 70, 70, 70, 74, 75, 75, 90, 95. Find the lower and upper quartiles of these measurements

THE BOX PLOT

- Divide the data into 4 sets containing an equal number of measurements
- A quick summary of the data distribution used to form a **box plot** to describe the **shape** of the distribution and to detect **outliers**
- The Five-number Summary: Minimum, Q_1, Median, Q_3, Maximum

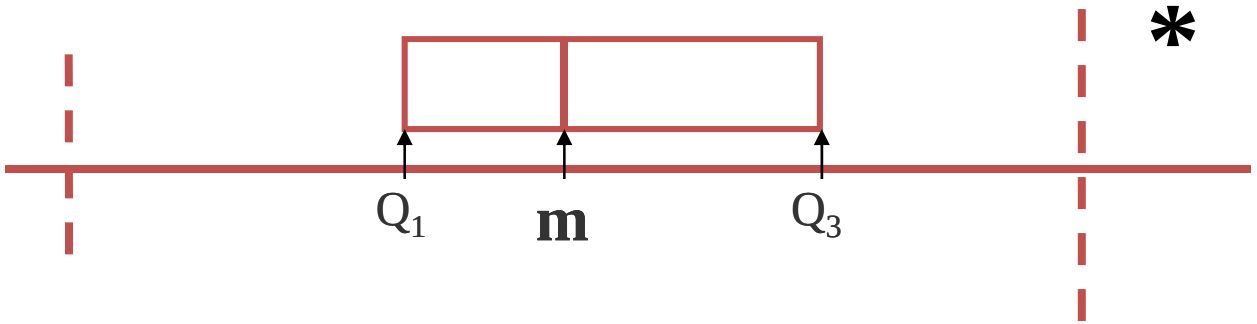
THE BOX PLOT

- Inter Quartile Range (IQR) = $Q_3 - Q_1$
- Construct the box plot
  - Draw a horizontal line to represent the scale of the measurement
  - Draw a box using Q_1, Median, and Q_3



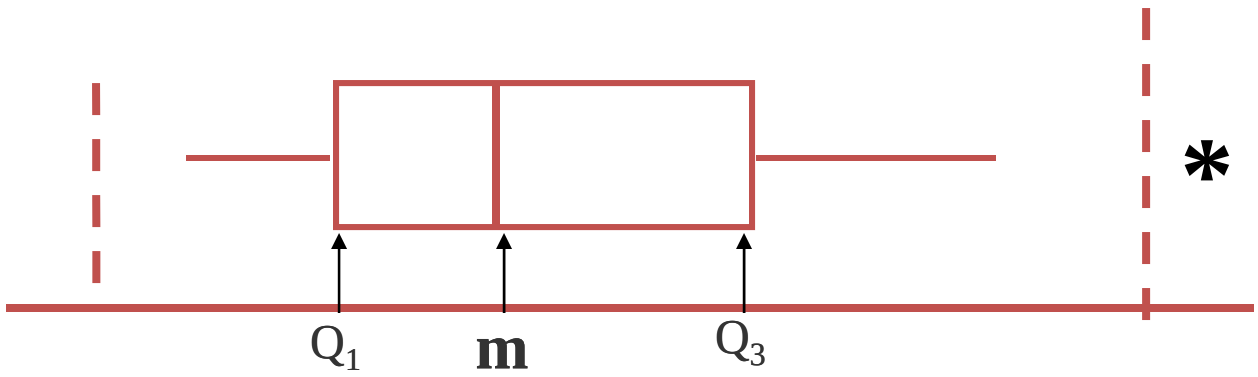
BOX PLOT

- Isolate outliers by calculating (Inner fence):
  - Lower fence: Q_1-1.5 IQR
  - Upper fence: Q_3+1.5 IQR
- Measurements beyond the upper or lower fence are outliers and are marked (*)



BOX PLOT

- Draw "whiskers" connecting the largest and smallest measurements that are NOT outliers to the box

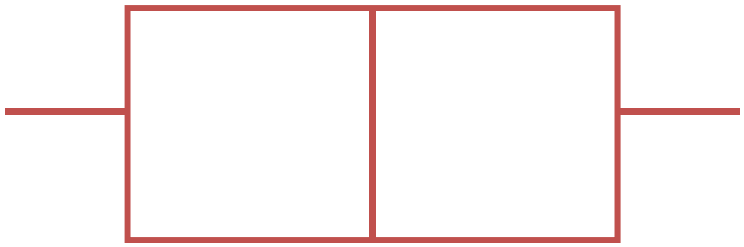


Box Plot

- Outer fence:
  - Lower outer fence: Q_1 - 3 IQR
  - Upper outer fence: Q_3 + 3 IQR

Interpreting Box Plot

- Symmetric distribution: Median line in the centre of box and whiskers of equal length, mean ≈ median



Interpreting Box Plot

- Right skewed distribution : most values are small with few exceptionally large ones that pull mean to the right. Mean > Median. Longer tail on the right.

INTERPRETING BOX PLOT

- Left skewed distribution: Most values are large with few exceptionally small ones that pull the mean to the left. Mean< Median

EXERCISE

2.11)Data (n = 50):

0.2	0.2	0.3	0.4	1.0	1.2	1.3	1.4	1.6	1.6	2.0	2.1	-
2.4	2.4	2.7	3.3	3.5	3.7	3.9	4.1	4.3	4.4	5.6	5.8	6.1
6.6	6.9	7.4	7.4	8.2	8.2	8.3	8.7	9.0	9.6	9.9	11.4	12.6
13.5	14.1	14.7	16.7	18.0	18.0	18.4	19.2	23.1	24.0	26.7	32.3	

Find the median, lower quartile, upper quartile for the data. Use these descriptive measures to construct a box plot for the data. Use box plot to describe the data distribution.

SUMMARY

- Measures of the centre
  - Mean
    - Population mean
    - Sample mean
  - Median
  - Mode

SUMMARY

- Measures of Variability
  - Range: R
  - Variance
    - Population Variance
    - Sample Variance
  - Standard Deviation
    - Population Standard deviation
    - Sample Standard deviation

SUMMARY CONT'D

- Tchebysheff's Theorem and Empirical Rule
- Measures of Relative Standing
  - Sample z-score
  - P^{th} percentile
  - Lower Quartile, Q_1
  - Upper Quartile, Q_3
  - Inter Quartile Range, IQR = $Q_3 - Q_1$

SUMMARY CONT'D

- The Five-Number Summary and Box Plots
  - Five-number summary
  - Box plots
  - Upper and lower fences
  - Outliers
  - Whiskers