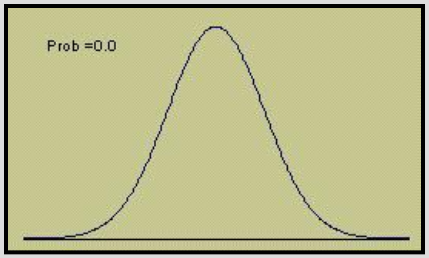
Using Measures of Center and Spread: Tchebysheff's Theorem

Given a number k greater than or equal to 1 and a set of n measurements, at least $1-(1/k^2)$ of the measurement will lie within k standard deviations of the mean.

- Can be used for either samples ($\bar{x}$ and s) or for a population (μ and σ).
- **Important results:**
  - If $k = 2$, at least $1 - 1/2^2 = 3/4$ of the measurements are within 2 standard deviations of the mean.
  - If $k = 3$, at least $1 - 1/3^2 = 8/9$ of the measurements are within 3 standard deviations of the mean.

Using Measures of Center and Spread: The 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.

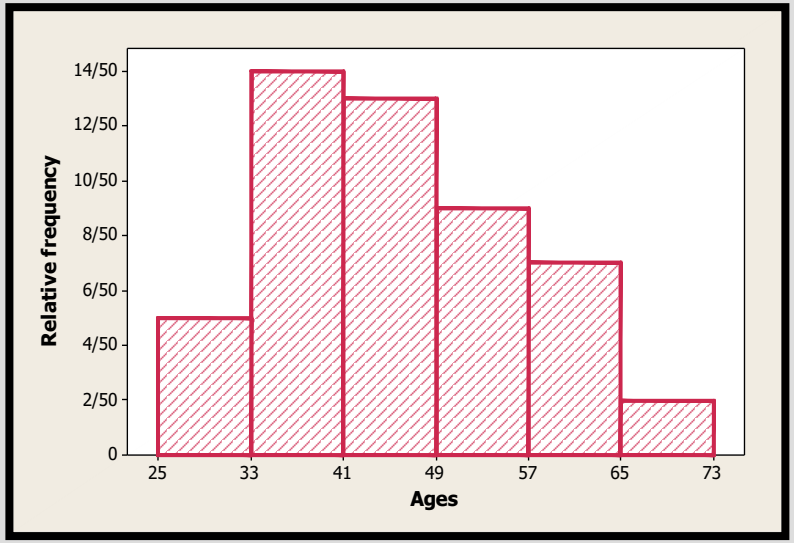
EXAMPLE

The ages of 50 tenured faculty at a state university.

34	48	**70**	63	52	52	35	50	37	43	53	43	52	44
42	31	36	48	43	**26**	58	62	49	34	48	53	39	45
34	59	34	66	40	59	36	41	35	36	62	34	38	28
43	50	30	43	32	44	58	53						

$$\bar{x} = 44.9$$

$$s = 10.73$$



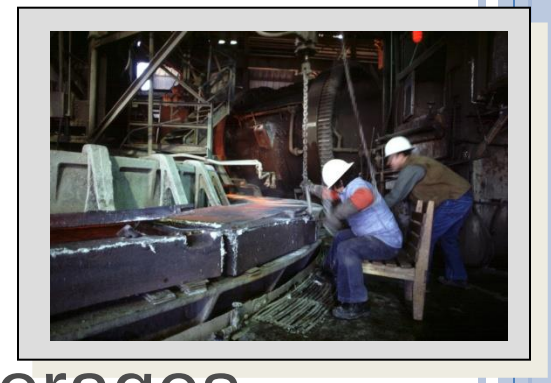
Shape? Skewed right

k	$\bar{x} \pm ks$	Interval	Proportion in Interval	Tchebysheff	Empirical Rule
1	44.9 ±10.73	34.17 to 55.63	31/50 (.62)	At least 0	≈ .68
2	44.9 ±21.46	23.44 to 66.36	49/50 (.98)	At least .75	≈ .95
3	44.9 ±32.19	12.71 to 77.09	50/50 (1.00)	At least .89	≈ .997

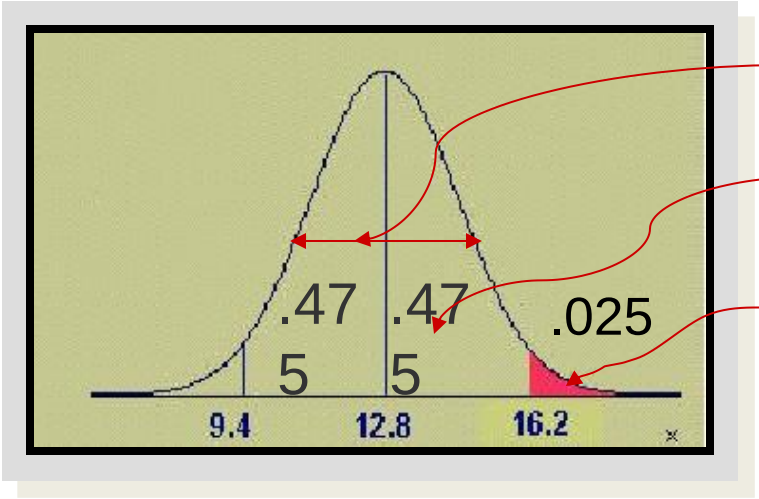
- Do the actual proportions in the three intervals agree with those given by Tchebysheff's Theorem?
- Do they agree with the Empirical Rule?
- Why or why not?

- **Yes. Tchebysheff's Theorem must be true for any data set.**
- **No. Not very well.**
- **The data distribution is not very mound-shaped, but skewed right.**

Example



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?



95% between 9.4 and 16.2

47.5% between 12.8 and 16.2

(50-47.5)% = 2.5% above 16.2

Approximating *s*

- From Tchebysheff's Theorem and the Empirical Rule, we know that

$$R \approx 4\text{-}6\ s$$

- To approximate the standard deviation of a set of measurements, we can use:

$$s \approx R/4$$

$$\text{or } s \approx R/6 \text{ for a large data set.}$$

Approximating *s*

The ages of 50 tenured faculty at state university.

•	34	48	**70**	63	52	52	35	50	37	43	53	43	52	44
•	42	31	36	48	43	**26**	58	62	49	34	48	53	39	45
•	34	59	34	66	40	59	36	41	35	36	62	34	38	28
•	43	50	30	43	32	44	58	53						

R = 70 – 26 = 44

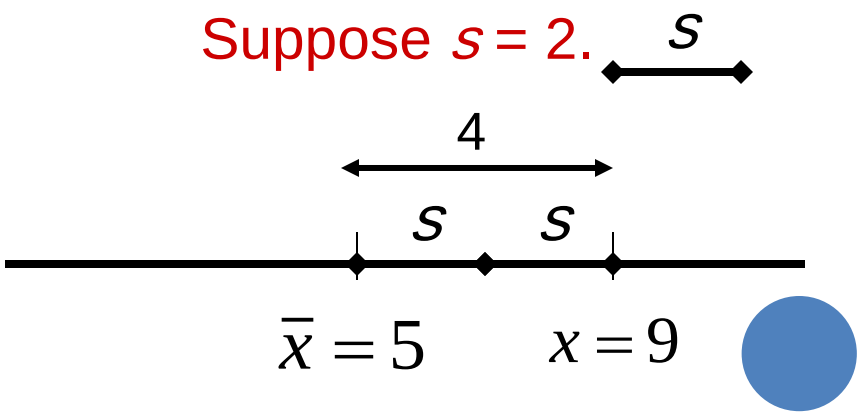
$$s \approx R/4 = 44/4 = 11$$

Actual *s* = 10.73

Measures of Relative Standing

- Where does one particular measurement stand in relation to the other measurements in the data set?
- How many standard deviations away from the mean does the measurement lie? This is measured by the ***z*-score.**

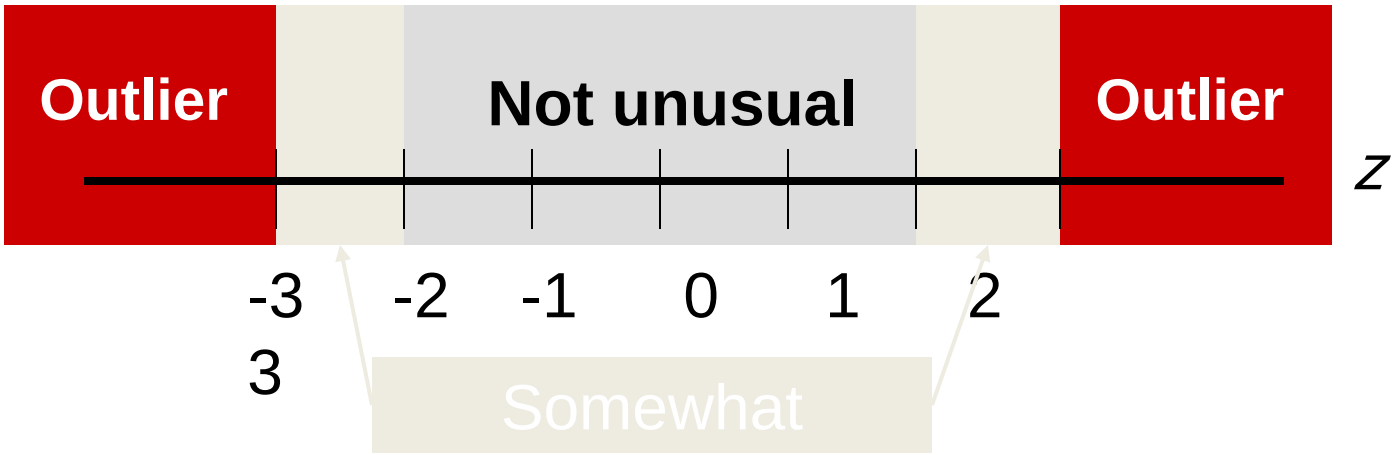
$$z\text{-score} = \frac{x - \bar{x}}{s}$$



$x = 9$ lies $z = 2$ std dev from the mean.

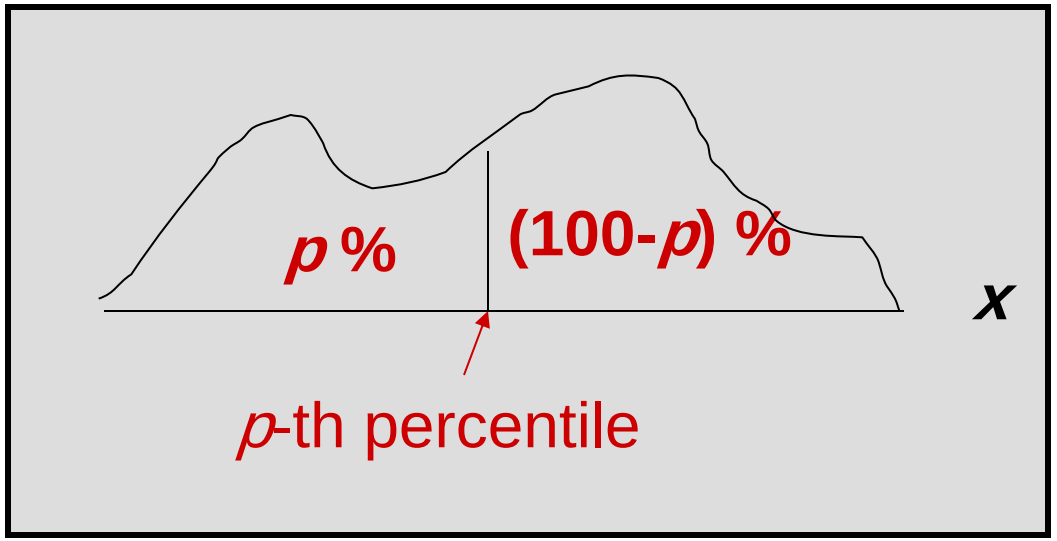
z-Scores

- From Tchebysheff's Theorem and the 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 between –2 and 2 are not unusual. *z*-scores should not be more than 3 in absolute value. *z*-scores larger than 3 in absolute value would indicate a possible **outlier.**



Measures of Relative Standing

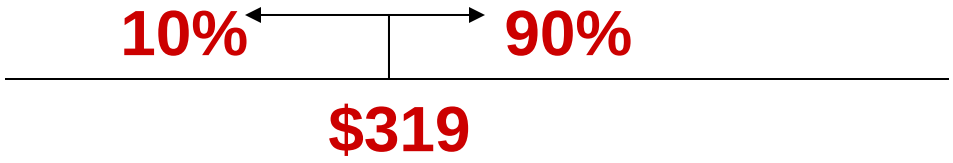
- How many measurements lie below the measurement of interest? This is measured by the ***p*th percentile.**



EXAMPLES

- 90% of all men (16 and older) earn more than $319 per week.

BUREAU OF LABOR STATISTICS



$319 is the 10th percentile.

50th Percentile ≡ Median

25th Percentile ≡ Lower Quartile (Q_1)

75th Percentile ≡ Upper Quartile (Q_3)

Quartiles and the IQR

- The **lower quartile** **(Q_1)** is the value of x which is larger than 25% and less than 75% of the ordered measurements.
- The **upper quartile** **(Q_3)** is the value of x which is larger than 75% and less than 25% of the ordered measurements.
- The range of the "middle 50%" of the measurements is the **interquartile range,**

$$IQR = Q_3 - Q_1$$

Calculating Sample Quartiles

- The **lower and upper quartiles (Q_1 and Q_3),** can be calculated as follows:
- The **position of Q_1** is $.25(n + 1)$
- The **position of Q_3** is $.75(n + 1)$

once the measurements have been ordered. If the positions are not integers, find the quartiles by interpolation.

EXAMPLE

The weights of 18 patients in kgs:

40 60 65 65 65 68 68 70 70

70 70 70 70 74 75 75 90 95

Position of Q_1 = .25(18 + 1) = 4.75

Position of Q_3 = .75(18 + 1) = 14.25

✓Q_1 is 3/4 of the way between the 4th and 5th ordered measurements, or

$Q_1 = 65 + .75(65 - 65) = 65.$

EXAMPLE

The weights of 18 patients in kgs:

40 60 65 65 65 68 68 70 70

70 70 70 70 74 75 75 90 95

Position of Q_1 = .25(18 + 1) = 4.75

Position of Q_3 = .75(18 + 1) = 14.25

✓Q_3 is 1/4 of the way between the 14th and 15th ordered measurements, or

$Q_3 = 74 + .25(75 - 74) = 74.25$

✓and

$IQR = Q_3 – Q_1 = 74.25 - 65 = 9.25$

Using Measures of Center and Spread: The Box Plot

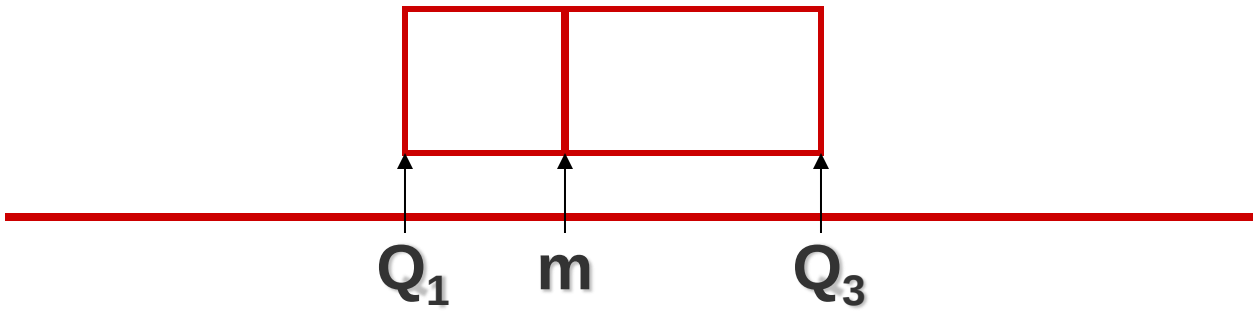
The Five-Number Summary:

Min Q_1 Median Q_3 Max

- Divides the data into 4 sets containing an equal number of measurements.
- A quick summary of the data distribution.
- Use to form a **box plot** to describe the **shape** of the distribution and to detect **outliers**.

Constructing a Box Plot

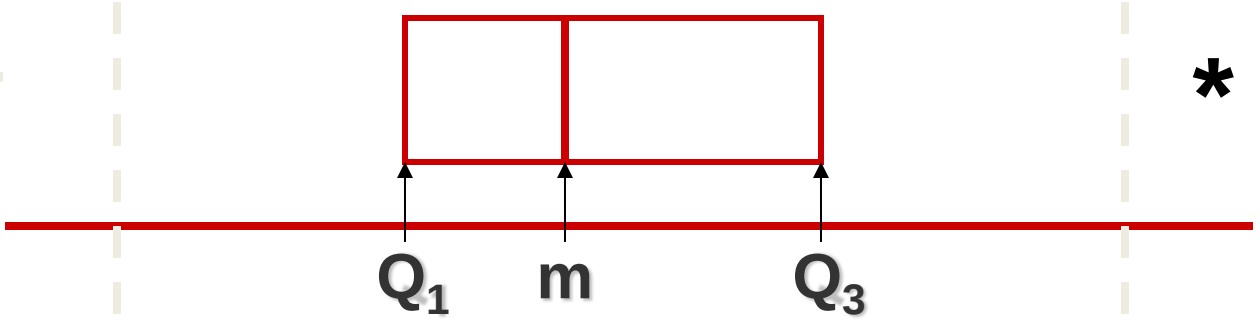
- ✓Calculate Q_1, the median, Q_3 and IQR.
- ✓Draw a horizontal line to represent the scale of measurement.
- ✓Draw a box using Q_1, the median, Q_3.



Constructing a Box Plot

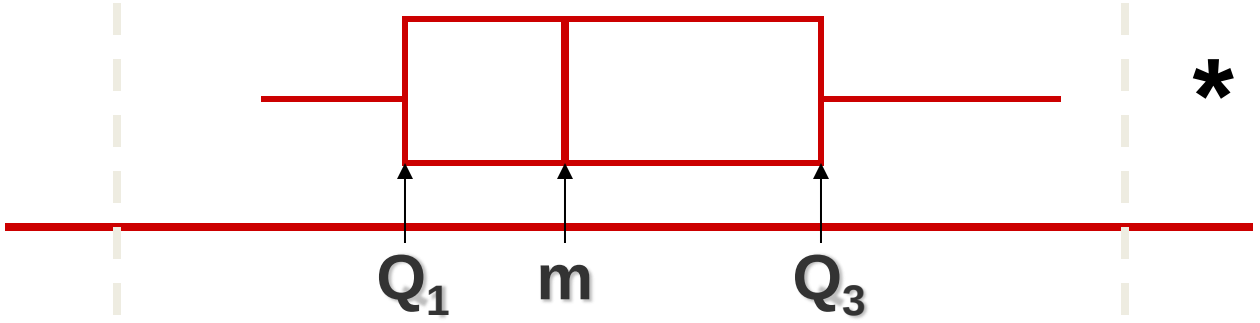
- ✓Isolate outliers by calculating
  - ✓Lower fence: Q_1-1.5 IQR
  - ✓Upper fence: Q_3+1.5 IQR

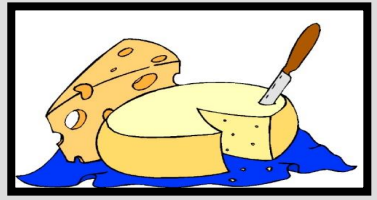
- ✓Measurements beyond the upper or lower fence is are outliers and are marked (*)



Constructing a Box Plot

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



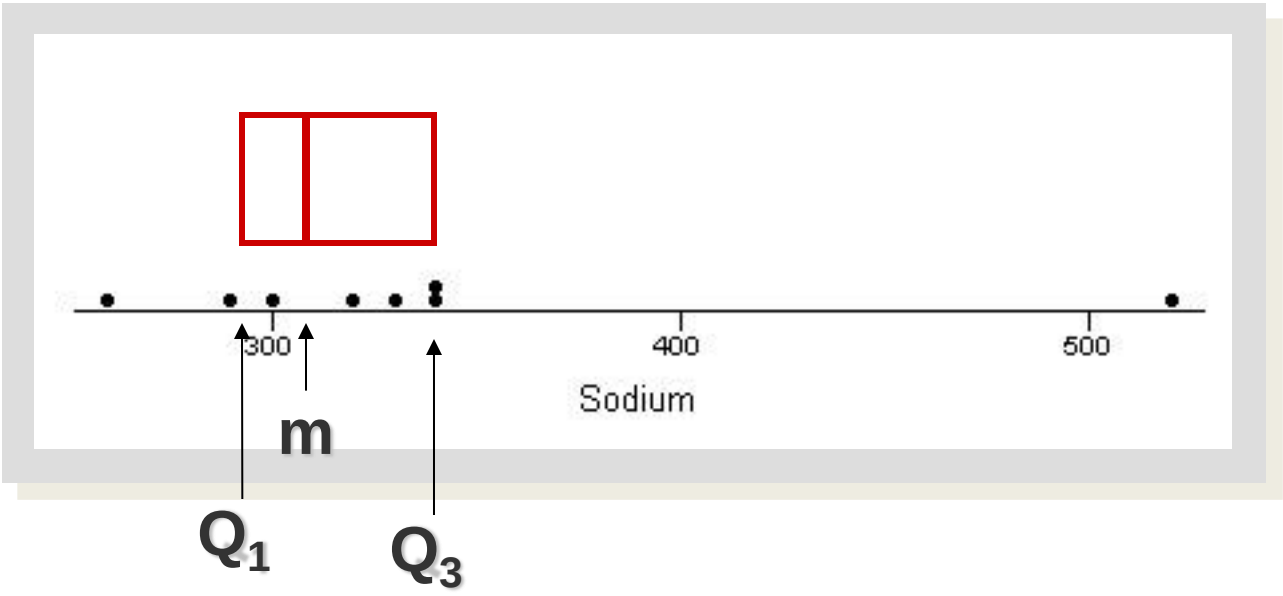


EXAMPLE

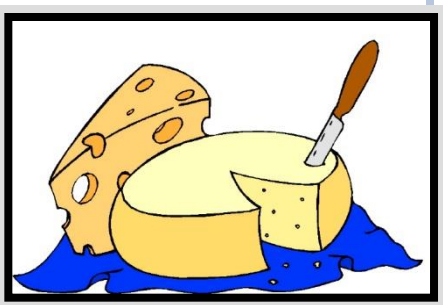
Amount of sodium in 8 brands of cheese:

260 290 300 320 330 340 340 520

$Q_1 = 292.5$ $m = 325$ $Q_3 = 340$



Example

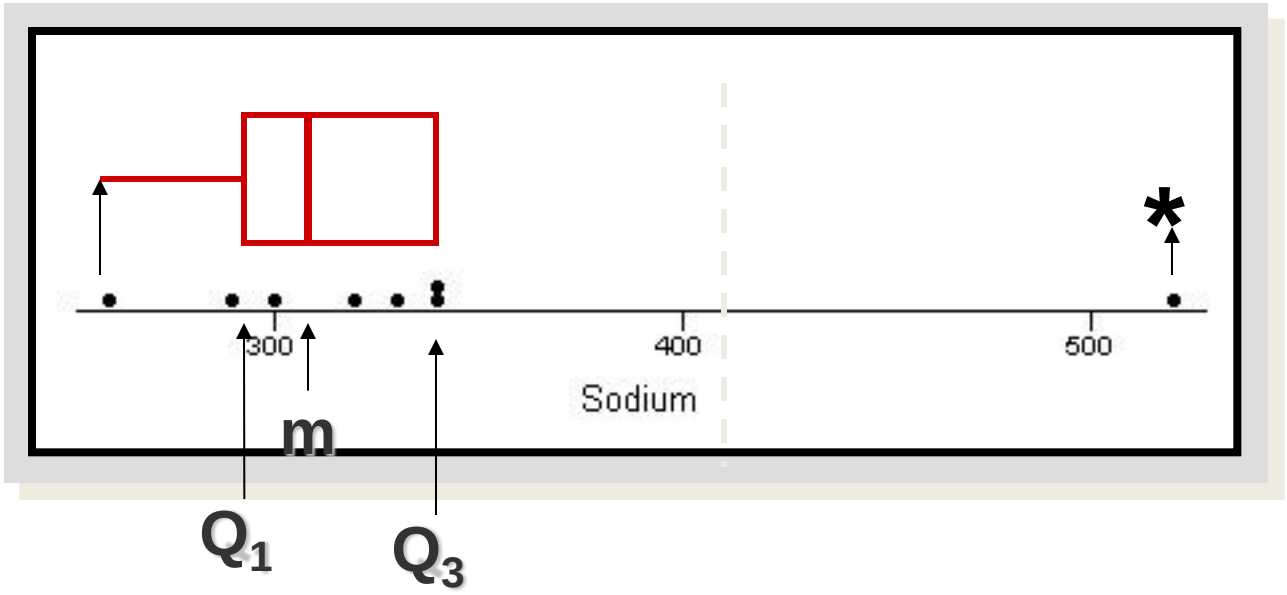


IQR = 340-292.5 = 47.5

Lower fence = 292.5-1.5(47.5) = 221.25

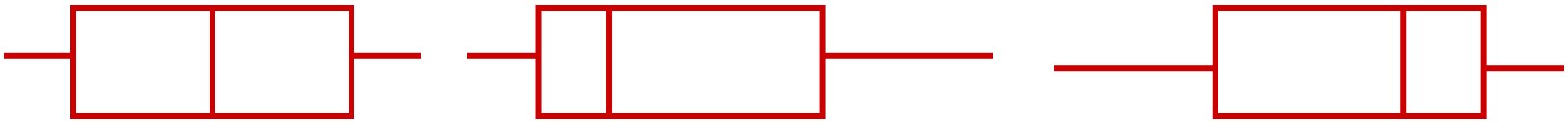
Upper fence = 340 + 1.5(47.5) = 411.25

Outlier: $x = 520$



Interpreting Box Plots

- ✓ Median line in center of box and whiskers of equal length—symmetric distribution
- ✓ Median line left of center and long right whisker—skewed right
- ✓ Median line right of center and long left whisker—skewed left



Key Concepts

I. Measures of Center

1. Arithmetic mean (mean) or average

   a. Population: μ

   b. Sample of size n: $\bar{x} = \frac{\sum x_i}{n}$

2. Median: **position** of the median = $.5(n+1)$

3. Mode

4. The median may preferred to the mean if the data are highly skewed.

II. Measures of Variability

1. Range: R = largest – smallest

KEY CONCEPTS

2. Variance

   a. Population of N measurements:

   $$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

   b. Sample of n measurements:

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

3. Standard deviation

   $$\text{Population standard deviation} : \sigma = \sqrt{\sigma^2}$$

   $$\text{Sample standard deviation} : s = \sqrt{s^2}$$

4. A rough approximation for s can be calculated as $s \approx R/4$.

   The divisor can be adjusted depending on the sample size.

Key Concepts

III. Tchebysheff's Theorem and the Empirical Rule

1. Use Tchebysheff's Theorem for any data set, regardless of its shape or size.

   a. At least $1-(1/k^2)$ of the measurements lie within k standard deviation of the mean.

   b. This is only a lower bound; there may be more measurements in the interval.

2. The Empirical Rule can be used only for relatively mound-shaped data sets.

- Approximately 68%, 95%, and 99.7% of the measurements are within one, two, and three standard deviations of the mean, respectively.

Key Concepts

IV. Measures of Relative Standing

1. Sample z-score:
2. pth percentile; $p\%$ of the measurements are smaller, and $(100 - p)\%$ are larger.
3. Lower quartile, Q_1; **position** of $Q_1 = .25(n + 1)$
4. Upper quartile, Q_3; **position** of $Q_3 = .75(n + 1)$
5. Interquartile range: $\text{IQR} = Q_3 - Q_1$

V. Box Plots

1. Box plots are used for detecting outliers and shapes of distributions.
2. Q_1 and Q_3 form the ends of the box. The median line is in the interior of the box.

Key Concepts

3. Upper and lower fences are used to find outliers.

    a. **Lower fence:** $Q_1 - 1.5(\text{IQR})$

    b. **Upper fence:** $Q_3 + 1.5(\text{IQR})$

4. **Whiskers** are connected to the smallest and largest measurements that are not outliers.

5. Skewed distributions usually have a long whisker in the direction of the skewness, and the median line is drawn away from the direction of the skewness.