Ch2. Describing Data with Numerical Measures

Graphs are extremely useful for the visual description of a data set. However, they are not always the best tool when you want to make inferences about a population from the information contained in a sample. For this purpose, it is better to use numerical measures to construct a mental picture of the data.

윤 연 옥

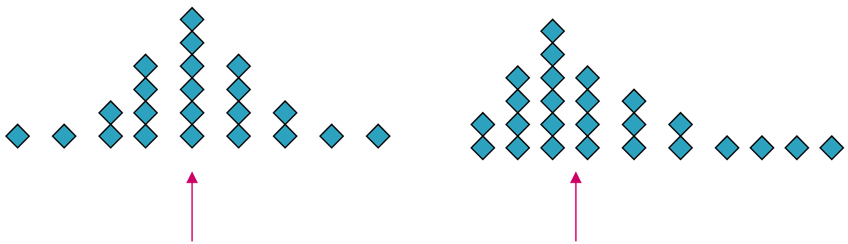
Describing Data with Numerical Measures

- Graphical methods may not always be sufficient for describing data.

- **Numerical measures** can be created for both **populations** and **samples.**
  - A **parameter** is a numerical descriptive measure calculated for a population.
  - A **statistic** is a numerical descriptive measure calculated for a sample.

2.1 Measures of Center

A measure along the horizontal axis of the data distribution that locates the center of the distribution.



Mean

- The **mean or average** of a set of measurements is the sum of the measurements divided by the total number of measurements

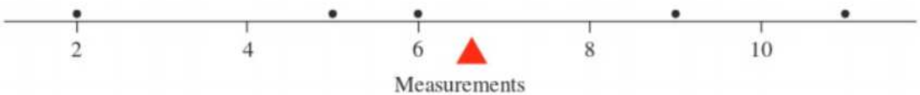
$$\bar{x} = \frac{\sum x_i}{n} \qquad \text{where } \sum x_i = x_1 + x_2 + \cdots + x_n$$

n = number of measurements

❖ If we were able to enumerate the whole population, the **population mean** would be called μ (the Greek letter "mu").

Example: set of data 2, 9, 11, 5, 6

$$\bar{x} = \frac{\sum x_i}{n} = \frac{2+9+11+5+6}{5} = \frac{33}{5} = 6.6$$



Median

- The **median** of a set of measurements is the middle measurement when the measurements are ranked from smallest to largest.
- The **position of the median** is **0.5(n+1)** once the measurements have been ordered.

Example1

The set: 2, 9, 11, 5, 6 $n=5$

- sort : 2, 5, 6, 9, 11
- Position : .5(n+1) = .5(5+1)=3rd

Median= 3rd largest measurement = 6

Example 2

Data set: 2, 9, 11, 5, 6, 27 $n=6$

- sort : 2, 5, 6, 9, 11, 27
- Position : $.5(n+1) = .5(6+1) = 3.5^{th}$

Median= (6+9)/2 = 7.5 : the average of the 3^{rd} and 4^{th} measurements

Tips

Example1 : mean=6.6 median =6

Example2 : mean=10 median=7.5 ← *contains large value(outliers)*

✓ The mean is more easily affected by extremely large or small values than the median

✓ The median is often used as a measure of center when the distribution is skewed

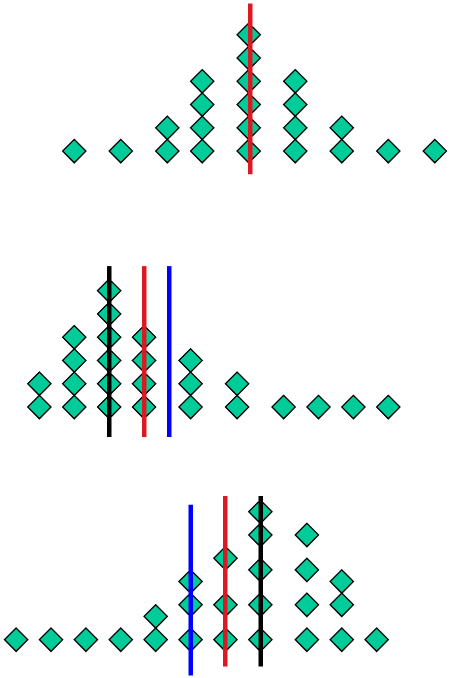
Mode

- The **mode** is the measurement which occurs most frequently.

Example

- set: 2, 4, 9, 8, 8, 5, 3
  - The mode is 8, which occurs twice
- The set: 2, 2, 9, 8, 8, 5, 3
  - There are two modes—8 and 2 (bimodal)
- The set: 2, 4, 9, 8, 5, 3
  - There is no mode (each value is unique).

Mean, Median & Mode



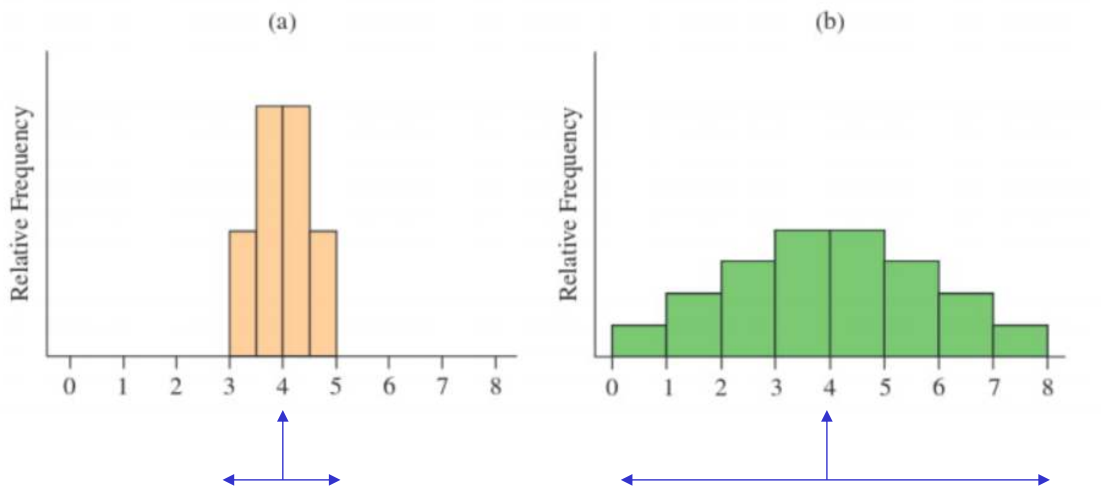
Symmetric: Mean = Median = Mode

Skewed right: Mode< Median< Mean

Skewed left: Mean < Median < Mode

2.2 Measures of Variability

- A measure along the horizontal axis of the data distribution that describes the **spread** of the distribution from the center.



The Range

- The **range, R**, of a set of n measurements is the difference between the largest and smallest measurements.

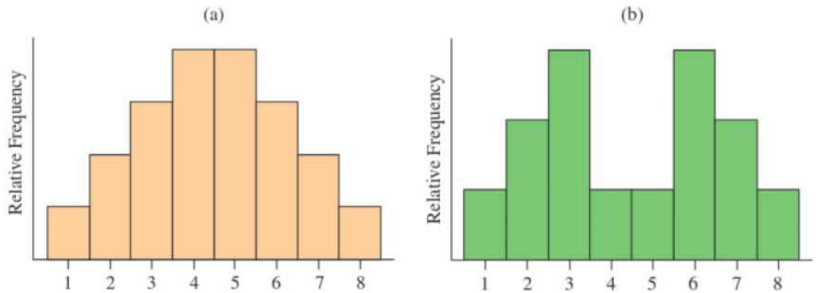
Example: A botanist records the number of petals on 5 flowers:

5, 12, 6, 8, 14

The range is **R = 14 – 5 = 9.**

✓ **Quick and easy, but only uses 2 of the 5 measurements**

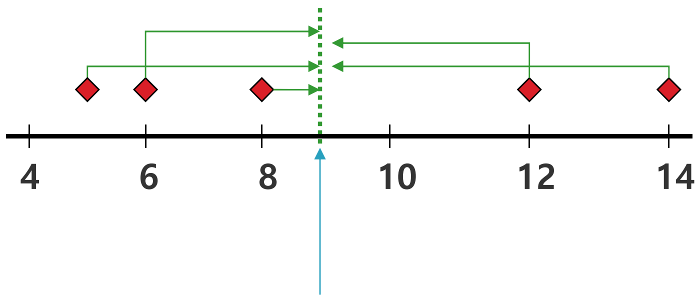
Distribution with equal range with unequal variability



The Variance

- The **variance** is measure of variability that uses all the measurements. It measures the average deviation of the measurements about their mean.

- Data : 5, 12, 6, 8,14



Mean $\bar{x} = \frac{\sum x_i}{n} = \frac{45}{5} = 9$

The Variance

- The **variance of a population** of *N* measurements is the average of the squared deviations of the measurements about their mean μ.

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

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

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

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: $s = \sqrt{s^2}$

Two ways to Calculate the Sample Variance

(1) Use Definition formula

$$s^2 = \frac{\sum(x_i - \overline{x})^2}{n-1} = \frac{60}{4} = 15$$

$$s = \sqrt{s^2} = \sqrt{15} = 3.87$$

(2) Use the Calculational Formula

$$s^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1} = \frac{465 - \frac{45^2}{5}}{4} = 15$$

$$s = \sqrt{s^2} = \sqrt{15} = 3.87$$

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

Some Notes

- The value of s is **ALWAYS positive**.
- The larger the value of s^2 or s, the larger the variability of the data set.
- **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 σ.

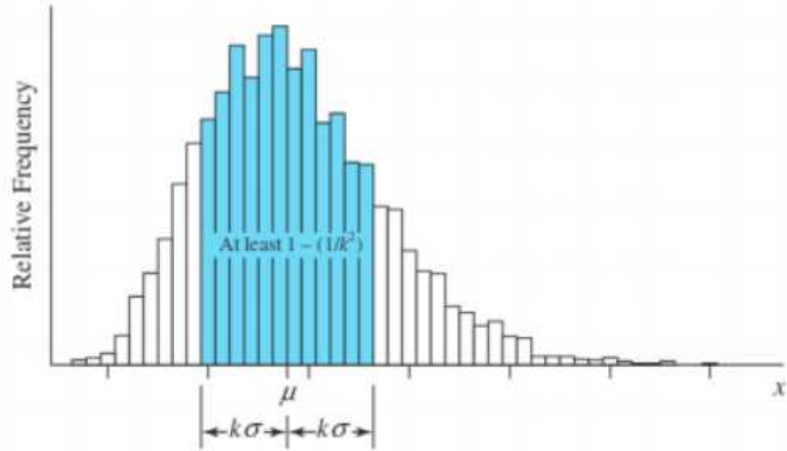
2.3 Understanding and Interpreting the Standard Deviation

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.

Ex) Tchebysheff's Theorem

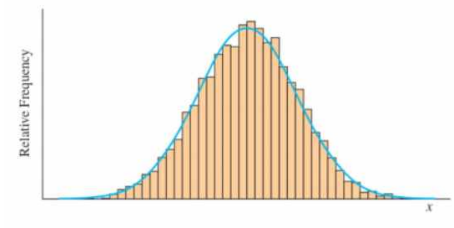


k	$1-(1/k^2)$
1	$1-1=0$
2	$1-1/4=3/4$
3	$1-1/9=8/9$

Not helpful

Empirical Rule

- Given a distribution of measurements that is approximately mound-shaped:

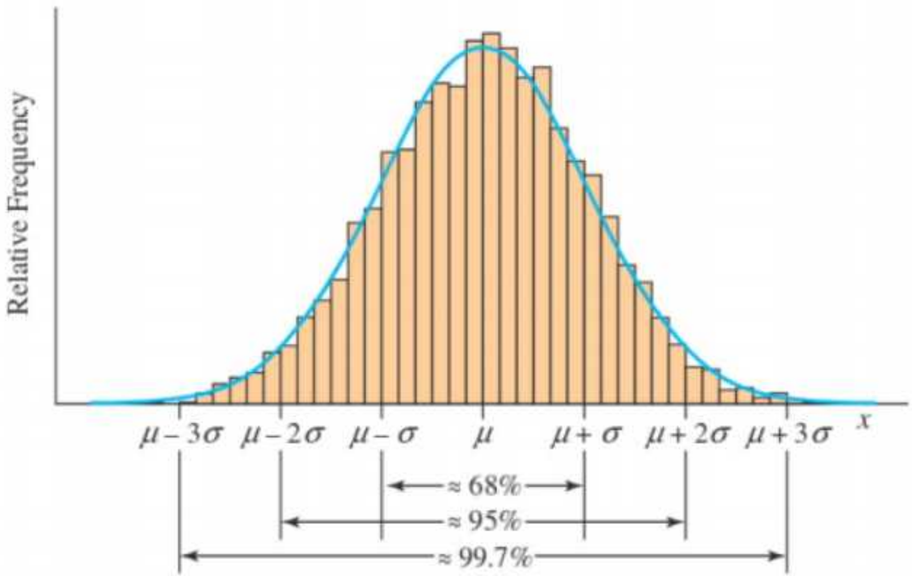


✓ The interval $\mu \pm \sigma$ contains approximately 68% of the measurements.

✓ 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.

Ex) Empirical Rule



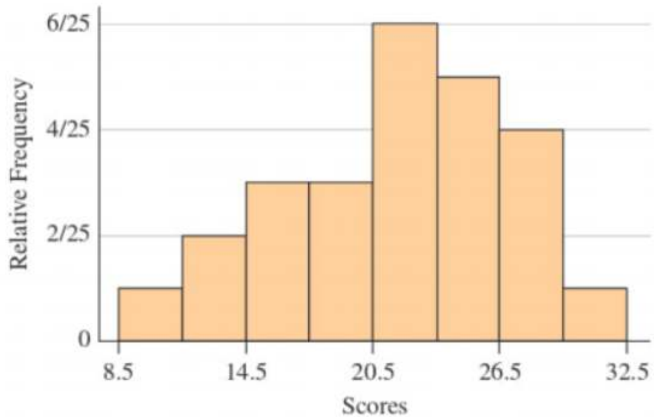
Example

- Lesson Plan Assessment Scores(0~34)

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

$\bar{x} = 21.6\,, s = 5.5$

Relative frequency histogram



Shape?

Nearly mound-shaped

k	$\bar{x} \pm ks$	Interval	Proportion in interval	Tchebysheff	Empirical Rule
1	21.6 ± 5.5	16.1~27.1	16/25(.64)	At least 0	≈ .68
2	21.6 ± 11.0	10.6~32.6	24/25(.96)	At least .75	≈ .95
3	21.6 ± 16.5	5.1~38.1	25/25(1.0)	At least .89	≈ .997

- Tchebysheff's Theorem applies to any set of measurements—sample or population, large or small, mound-shaped or skewed.

  - It gives a **lower bound** to the fraction of measurements to be found in an interval constructed as $\bar{x} \pm ks$

  - it always be satisfied, but it is a **very conservative**

- The Empirical Rule is a "rule of thumb" that can be used as a descriptive tool **only when the data tend to be roughly mound-shaped**

  - this rule will give you a **more accurate estimate**

Approximating s using Range

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:

$$s \approx R/4$$

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

Ex1) Data : 5, 12, 6, 8, 14

$$R = 14 - 5 = 9 \ \Rightarrow\ s \approx \frac{R}{4} = \frac{9}{4} = 2.25$$

Actual $s = 3.87$ is a little larger than out estimate.

Ex 2) Lesson Plan Assessment Scores

$$R = 31.9 - 10.2 = 21.7 \Rightarrow\ s \approx \frac{R}{4} = \frac{21.7}{4} = 5.4$$

Actual $s = 5.5$ is very close approximation.

2.4 Measures of Relative Standing

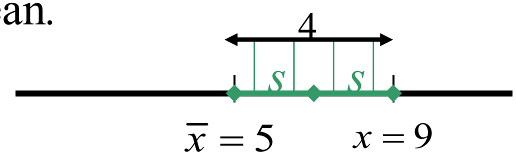
z-scores

- 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.**

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

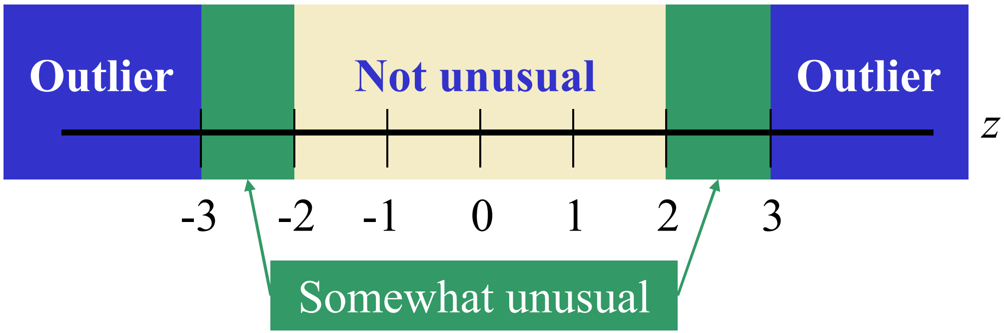
Ex) Suppose $\bar{x}=5$, $s = 2$.

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



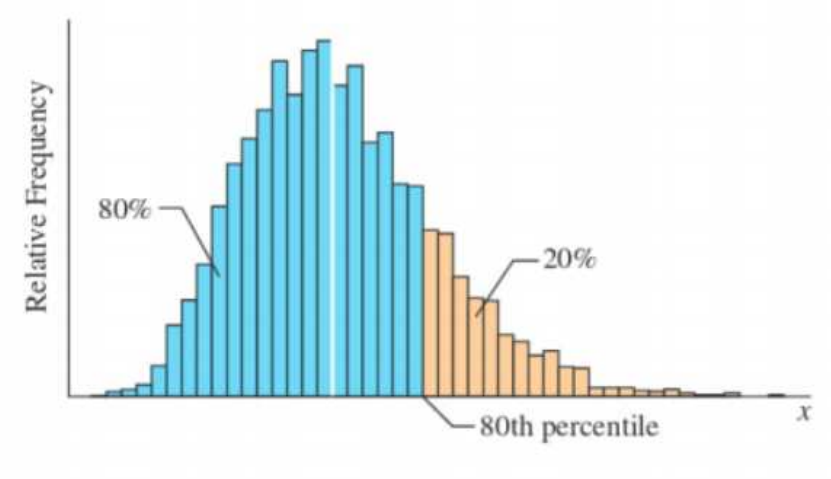
Outliers and z-scores

- 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.**



P^{th} percentile

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

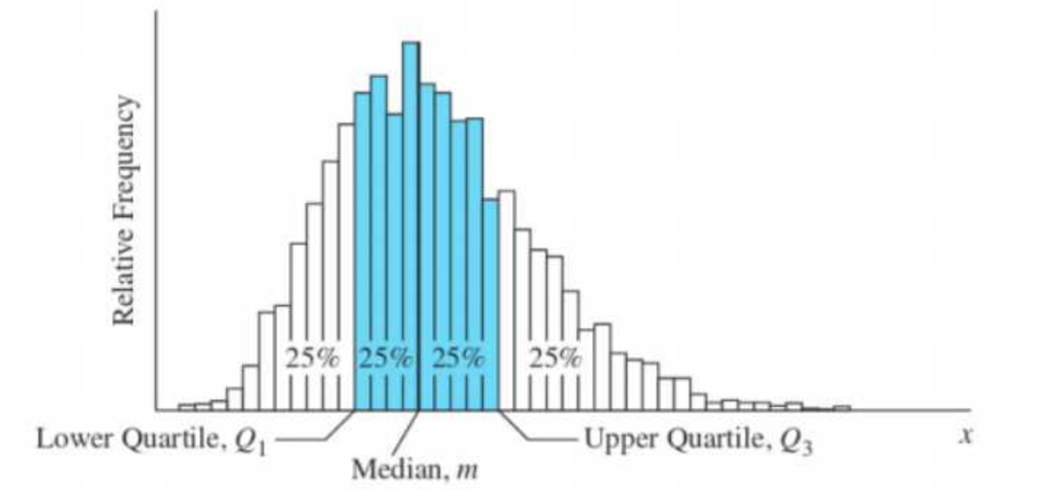


Ex) 90% of all men earn more than \$400 per week

$\Rightarrow$ \$400 is the 10^{th} percentile

Quartiles and IQR

- 50^{th} Percentile = Median
- 25^{th} Percentile = Lower Quartile(Q_1)
- 75^{th} Percentile = Upper Quartile(Q_3)
- The range of the "middle 50%" of the measurements is the **interquartile range**, IQR = $\mathbf{Q_3 - Q_1}$



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 $.75(n+1)$

✓ *If the positions are not integers, find the quartiles by interpolation.*

Example) Data: 16, 25, 4, 18, 11, 13, 20, 8, 11, 9

(1) Rank the n=10 measurements from smallest to largest:

4, 8, 9, 11, 11, 13, 16, 18, 20, 25

(2) Calculate

- Position of Q_1 = .25(n+1)= .25(10+1)= 2.75
- Position of Q_3 = .75(n+1)= .75(10+1)= 8.25

(3) Interpolation

- Q_1 = 8 + .75(9- 8) = 8 + .75 = 8.75
- Q_3 = 18 +.25(20-18) = 18+ .5 = 18.5

(4) IQR = Q_3 - Q_1 =18.5 -8.75 = 9.75

The Five-Number Summary and 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

(1) Calculate Q_1, the median, Q_3 and IQR.

(2) Draw a horizontal line to represent the scale of measurement.

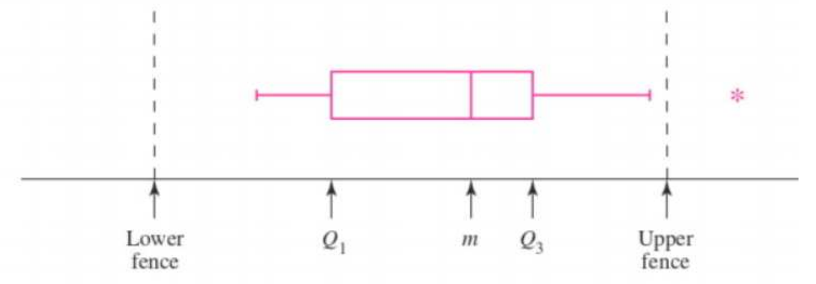
(3) Draw a box using Q_1, the median, Q_3.

(4) Isolate outliers by calculating

- Lower fence: Q_1-1.5 IQR, Upper fence: Q_3+1.5 IQR

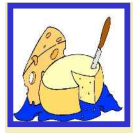
(5) Measurements beyond the upper or lower fence is are outliers and are marked (*).

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



Example

Data: the amounts of sodium per slice(in milligrams) for each of eight brands of regular American cheese.(n=8)



340, 300, 520, 340, 320, 290, 260, 330

(1) ranked from smallest to largest:

260, 290, 300, 320, 330, 340, 340, 520

(2) Calculate median, Q_1, and Q_3

$.5(n+1)= .5(9)= 4.5, \quad .25(n+1)=.25(9)=2.25, \quad .75(n+1)=.75(9)=6.75$

so that $m=(320+330)/2=325$, $Q_1 =290+.25(10)=292.5$, and $Q_3=340$.

(3) IQR and upper and lower fences

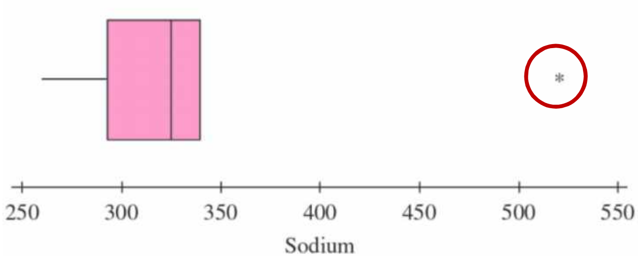
$IQR= Q_3 - Q_1 = 340 - 292.5 = 47.5$

Lower fence: $292.5-1.5(47.5) = 221.25$

Upper fence: $340 +1.5(47.5) = 411.25$

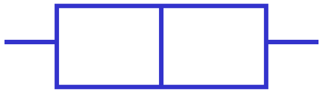
✓ *The value 520 is the only outlier*

Box plot



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: $\bar{x} = \dfrac{\sum x_i}{n}$

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.

Key Concepts

II. Measures of Variability

1. Range: R = largest – smallest
2. Variance
   a. Population of N measurements: $\sigma^2 = \dfrac{\sum (x_i - \mu)^2}{N}$
   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

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

   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.

IV. Measures of Relative Standing

1. Sample *z*-score:
2. *p*th 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: $\text{IQR} = Q_3 - Q_1$

Key Concepts

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.
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.