Descriptive analysis of continuous variables

Tuan V. Nguyen
Professor and NHMRC Senior Research Fellow
Garvan Institute of Medical Research
University of New South Wales
Sydney, Australia

Some *old* words

"If it were not for the great variability among individuals, medicine might be a Science, not an Art"

William Osler, 1882

The Principles and Practice of Medicine

Normal (Gaussian) distribution

- Given a series of values *xi* (*i = 1, … , n*): *x1, x2, …, xn*, the mean is:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Study 1**: the color scores of 6 consumers are: 6, 7, 8, 4, 5, and 6. The mean is:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{6+7+8+4+5+6}{6} = \frac{36}{6} = 6$$

- **Study 2**: the color scores of 4 consumers are: 10, 2, 3, and 9. The mean is:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{10+2+3+9}{4} = \frac{24}{4} = 6$$

Variation

- The mean does not adequately describe the data. We need to know the ***variation*** in the data.
- An obvious measure is the sum of ***difference*** from the mean.

- For study 1, the scores 6, 7, 8, 4, 5, and 6, we have:

(6-**6**) + (7-**6**) + (8-**6**) + (4-**6**) + (5-**6**) + (6-**6**)

= 0 + 1 + 2 – 2 – 1 + 0

= 0

NOT SATISFACTORY!

Sum of squares

- We need to make the difference positive by squaring them. This is called "***Sum of squares***" (SS)
- For study 1: 6, 7, 8, 4, 5, 6, we have:

  $SS = (6-\mathbf{6})^2 + (7-\mathbf{6})^2 + (8-\mathbf{6})^2 + (4-\mathbf{6})^2 + (5-\mathbf{6})^2 + (6-\mathbf{6})^2$

  $= 10$

- For study 2: 10, 2, 3, 9, we have:

  $SS = (10-\mathbf{6})^2 + (2-\mathbf{6})^2 + (3-\mathbf{6})^2 + (9-\mathbf{6})^2 = 50$

- This is better!
- **But it does not take into account sample size *n*.**

Variance

- We have to divide the SS by sample size *n*. But in each square we use the mean to calculate the square, so we lose 1 degree of freedom.
- Therefore the correct denominator is *n*-1. This is called ***variance*** (denoted by s^2)

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}$$

- **Or, in the sum notation:**

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

Variance - example

- For study 1: 6, 7, 8, 4, 5, and 6, the variance is:

$$s^2 = \frac{(6-6)^2 + (7-6)^2 + (8-6)^2 + (5-6)^2 + (6-6)^2}{6-1} = \frac{10}{5} = 2$$

- For study 2: 10, 2, 3, 9, the variance is:

$$s^2 = \frac{(10-6)^2 + (2-6)^2 + (3-6)^2 + (9-6)^2}{4-1} = \frac{50}{3} = 16.7$$

- **The scores in study 2 were much more variable than those in study 1.**

Standard deviation

- The problem with variance is that it is expressed in **unit squared**, whereas the mean is in the actual unit. We need a way to convert variance back to the actual unit of measurement.

- We take the square root of variance – this is called "***standard deviation***" (denote by *s*)

- For study 1, s = sqrt(2) = 1.41

  For study 2, s = sqrt(16.7) = 4.1

Coefficient of variation

- In many studies, the standard deviation can vary with the mean (eg higher/lower mean values are associated with higher/lower SD)
- Another statistic commonly used to quantify this phenomenon is the ***coefficient of variation*** (CV).
- A CV expresses the *SD as percentage of the mean.* ***CV = SD/mean*100***
- For study 1, CV = 1.41 / 6 * 100 = 23.5%

  For study 2, CV = 4.1 / 6 * 100 = 68.3%

Summary statistics

- Summary statistics are usually shown in ***sample size***, ***mean*** and ***standard deviation***.
- In our examples

Study	N	Mean	SD
1	6	6.0	1.4
2	4	6.0	4.1

Implications of the mean and SD

- "*In the Vietnamese population aged 30+ years, the average of weight was 55.0 kg, with the SD being 8.2 kg*."
- What does this mean?

- If the data are ***normally*** distributed, this means that the probability that *an individual randomly selected from the population with weight being w kg is*:

$$P(Weight = w) = \frac{1}{s\sqrt{2\pi}} \exp\left[\frac{-(w-\bar{x})^2}{2s^2}\right]$$

Implications of the mean and SD

- In our example, $x = 55$, $s = 8.2$
- The probability that *an individual randomly selected from the population with weight being 40 kg is*:

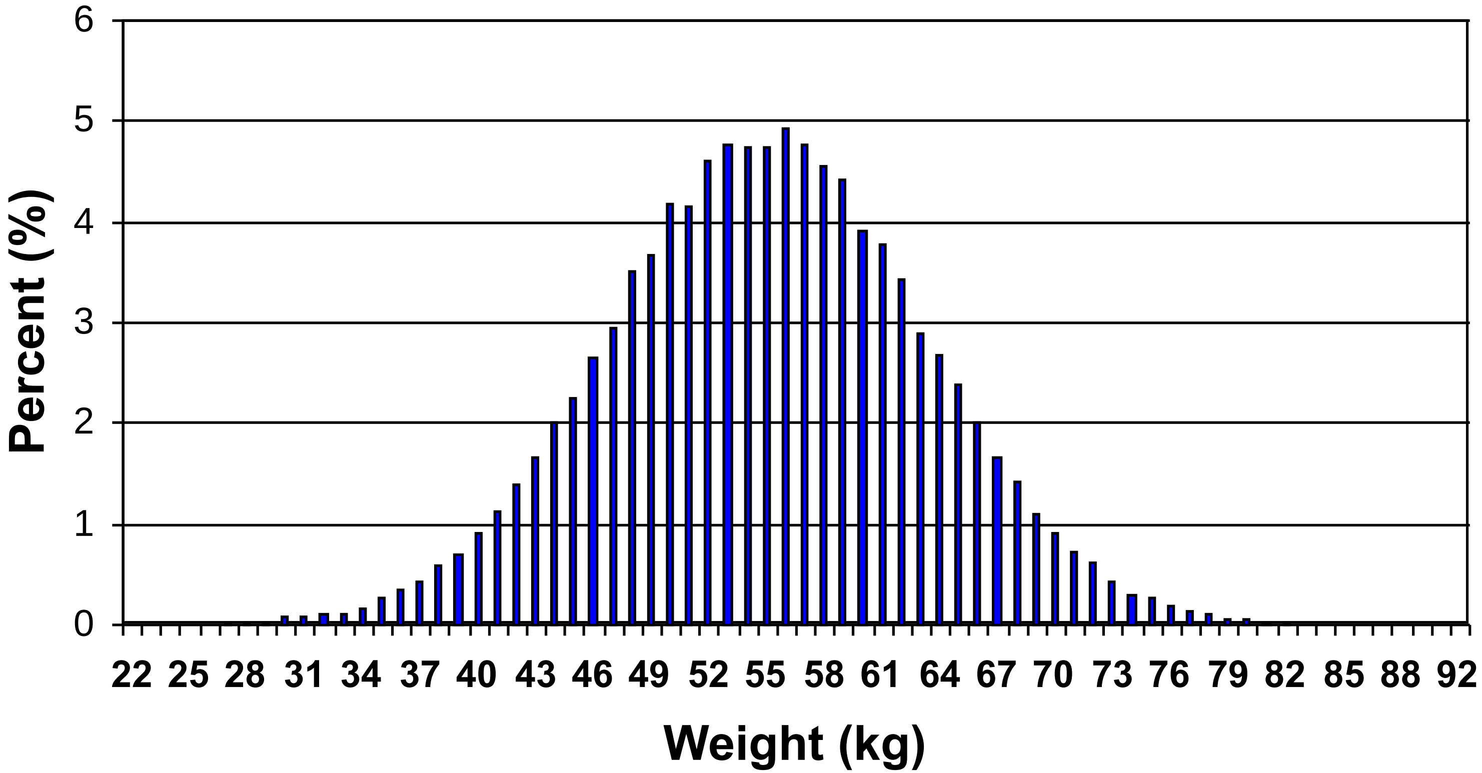
$$P(Weight = 40) = \frac{1}{8.2 \times \sqrt{2 \times 3.1416}} \exp\left[\frac{-(40-55)^2}{2 \times 8.2 \times 8.2}\right] = 0.009$$

$$P(Weight = 50) = \frac{1}{8.2 \times \sqrt{2 \times 3.1416}} \exp\left[\frac{-(50-55)^2}{2 \times 8.2 \times 8.2}\right] = 0.040$$

$$P(Weight = 80) = \frac{1}{8.2 \times \sqrt{2 \times 3.1416}} \exp\left[\frac{-(80-55)^2}{2 \times 8.2 \times 8.2}\right] = 0.0004$$

Implications of the mean and SD

- The distribution of weight of the entire population can be shown to be:



Z-scores

- Actual measurements can be converted to z-scores
- A z-score is the ***number of SDs from the mean***

$$Z = \frac{x - \bar{x}}{s}$$

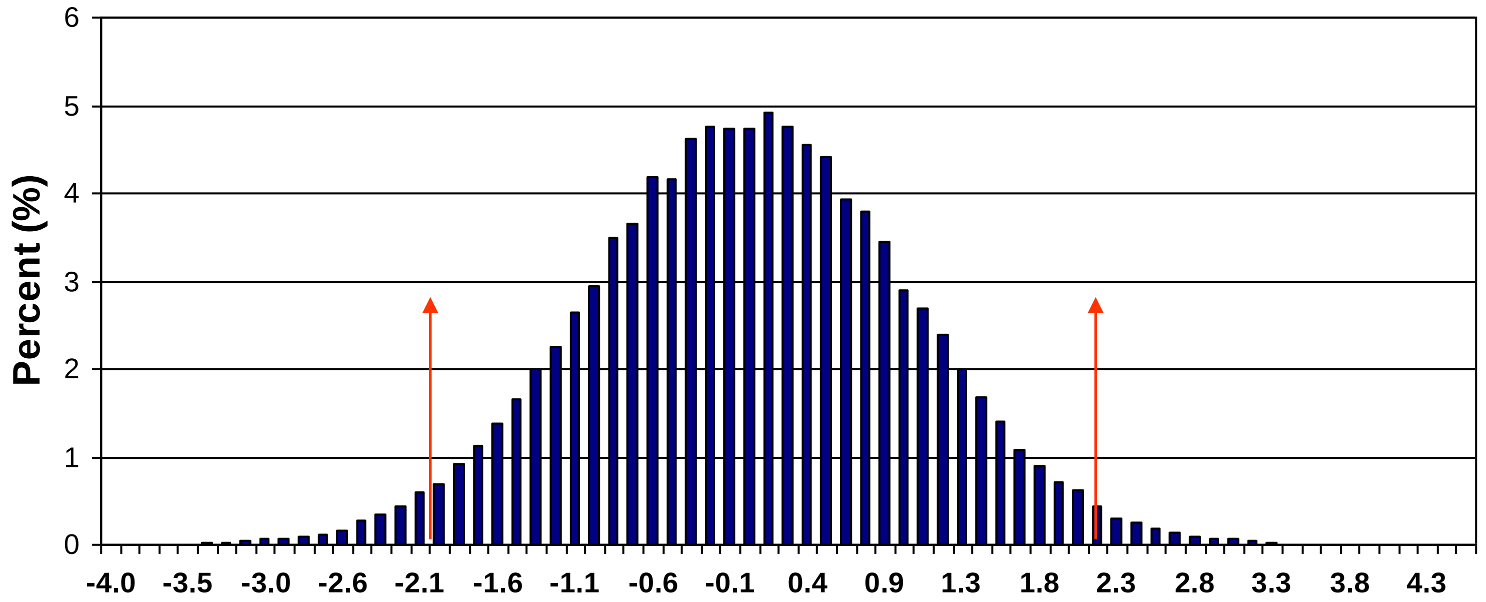
- A weight = 55 kg → z = (55 - 55)/8.2 = 0 SDs
- A weight = 40 kg → z = (40 - 55)/8.2 = -1.8 SDs
- A weight = 80 kg → z = (80-55)/8.2 = 3.0 SDs

Z-scores = Standard Normal Distribution

- A z-score is unitless, allowing comparison between variables with different measurements
- Z-scores have mean 0 and variance of 1.
- Z-scores → Standard Normal Distribution

Z-scores and area under the curve

- Z-scores and weight – another look:



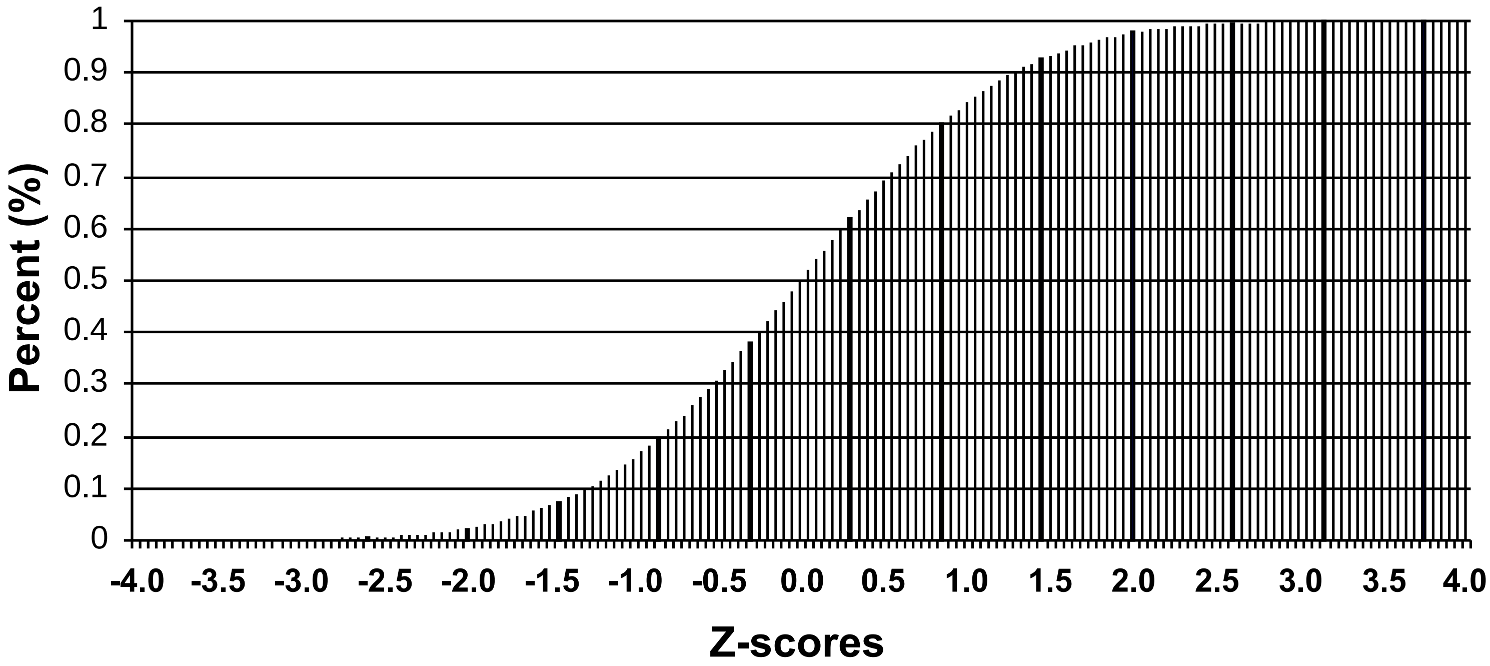
- Area under the curve for z $\leq$ -1.96 = 0.025
- Area under the curve for -1.0 $\leq$ z $\leq$ 1.0 = 0.6828
- Area under the curve for -2.0 $\leq$ z $\leq$ 2.0 = 0.9544
- Area under the curve for -3.0 $\leq$ z $\leq$ 3.0 = 0.9972

95% confidence interval

- A sample of n measurements ($x_1, x_2, \ldots, x_n$), with mean x and standard deviation s.
- 95% of the individual values of x_i lies between $x\text{-}1.96s$ and $x+1.96s$

- Mean weight = 55 kg, SD = 8.2 kg
- 95% of individuals' weight lies between 39 kg and 71 kg.

Cumulative probability (area under the curve) for Z-scores



Z ≤	-3	-2.5	-2.0	-1.5	-1.0	-0.5	0	0.5	1.0	1.5	2.0	2.5	3.0
Prob	.0013	.006	.0227	.0668	.1587	.3085	.5000	.6915	.8413	.9332	.9772	.9938	.9987

Standard error (SE)

$$SE = \frac{s}{\sqrt{n}}$$

- SE = standard error
- *s* : standard deviation
- *n*: sample size

What does it mean ?

The meaning of SE

- Consider a population of 10 people:

  130, 189, 200, 156, 154, 160, 162, 170, 145, 140

- Mean μ = 160.6 cm

- We *repeated* take random samples, each sample has 5 people:

The meaning of SE

- We *repeated* take random samples, each sample has 5 people:

1st sample: 140, 160, 200, 140, 145 mean $x1 = 157.0$

2nd sample: 154, 170, 162, 160, 162 mean $x2 = 161.6$

3rd sample: 145, 140, 156, 140, 156 mean $x3 = 147.4$

4th sample: 140, 170, 162, 170, 145 mean $x4 = 157.4$

5th sample: 156, 156, 170, 189, 170 mean $x5 = 168.2$

6th sample: 130, 170, 170, 170, 170 mean $x6 = 162.0$

7th sample: 156, 154, 145, 154, 189 mean $x7 = 159.6$

8th sample: 200, 154, 140, 170, 170 mean $x8 = 166.8$

9th sample: 140, 170, 145, 162, 160 mean $x9 = 155.4$

10th sample: 200, 200, 162, 170, 162 mean $x10 = 178.8$

….

SD of x1, x2, x3, …, x10 is the SE

Use of SD and SE

Let the *population mean* be μ (we do not know μ). Let the *sample mean* be *x* and SD be *s*.

- 68% *individuals* in the population will have values from x–*s* to x+*s*
- 95% *individuals* in the population will have values from x–2*s* to x+2*s*
- 99% *individuals* in the population will have values from x–3*s* to x+3*s*

Let the *population mean* be μ (we do not know μ). Let the *sample mean* be *x* and SE be *se*.

- 68% *averages* from repeated samples will have values from x–*se* to x+*se*
- 95% *averages* from repeated samples will have values from x–2*se* to x+2*se*
- 99% *averages* from repeated samples will have values from x–3*se* to x+3*se*

Central location: Median

- The median is the value with a *depth* of (n+1)/2
- When *n* is even, average the two values that straddle a depth of (n+1)/2
- For the 10 values listed below, the median has depth (10+1) / 2 = 5.5, placing it between 27 and 28. Average these two values to get median = 27.5

05 11 21 24 27 28 30 42 50 52

↑

median

Average the adjacent values: *M* = 27.5

More examples of medians

- Example A: 2 4 6

  Median = 4

- Example B: 2 4 6 8

  Median = 5 (average of 4 and 6)

- Example C: 6 2 4

  Median $\neq$ 2

(Values must be ***ordered*** first)

The median is *robust*

The median is more resistant to skews and outliers than the mean; it is more *robust*.

This data set has a mean of 1636:

1362 1439 1460 1614 1666 1792 1867

Here's the same data set with a data entry error "outlier" **(*highlighted*).** This data set has a mean of 2743:

1362 1439 1460 1614 1666 1792 *9867*

The median is 1614 in both instances, demonstrating its robustness in the face of outliers.

Mode

- The mode is the most commonly encountered value in the dataset
- This data set has a mode of 7
  {4, 7, 7, 7, 8, 8, 9}
- This data set has no mode
  {4, 6, 7, 8}
  (each point appears only once)
- The mode is useful only in large data sets with repeating values

Comparison of Mean, Median, Mode

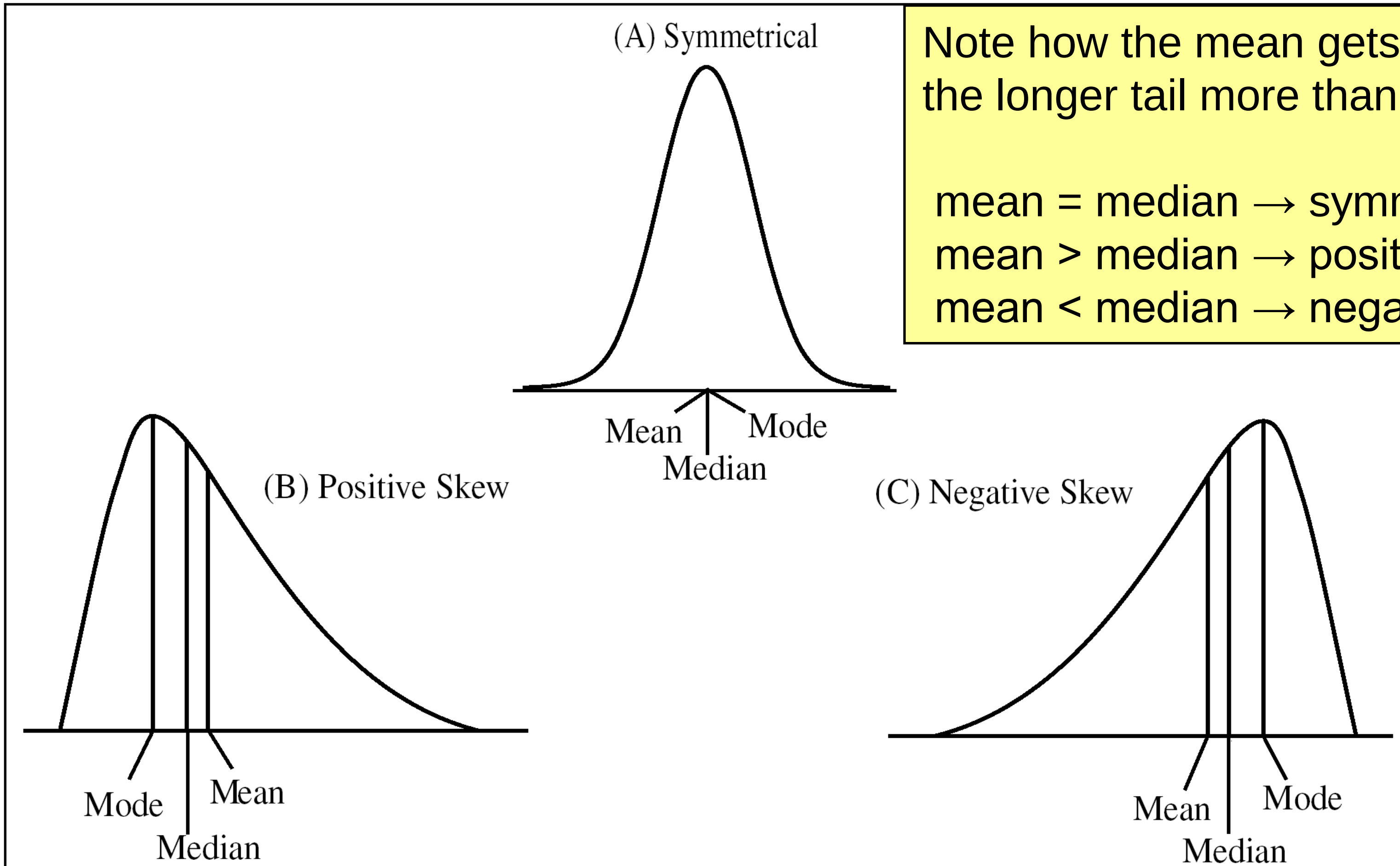


FIGURE 4.4 Effect of a skew on the mean, median, and mode.

Note how the mean gets pulled toward the longer tail more than the median

mean = median → symmetrical distrib
mean > median → positive skew
mean < median → negative skew

Spread: Quartiles

- **Quartile 1 (Q1):** cuts off bottom 25% of data
- **Quartile 3 (Q3):** cuts off top 25% of data
  = median of the to half of the data set
- **Interquartile Range (IQR)** = Q3 – Q1

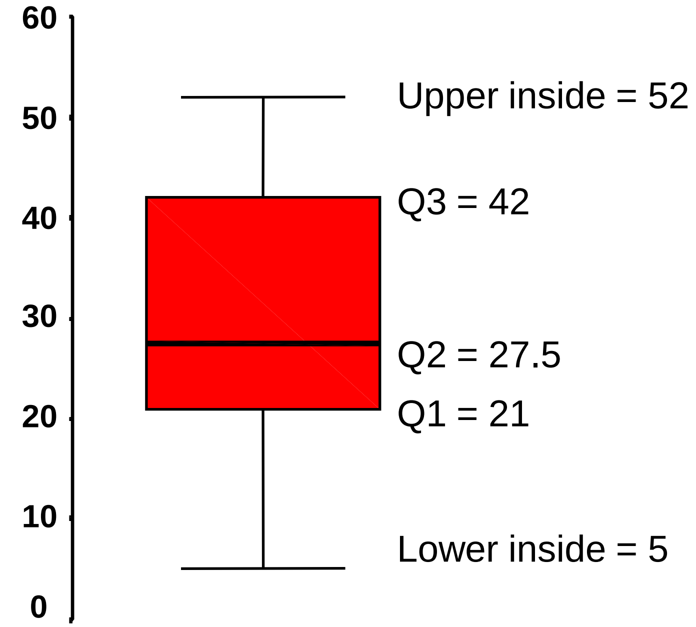
```
05   11   21   24   27     28   30   42   50   52
          ↑               ↑              ↑
          Q1            median           Q3
```

Q1 = 21, Q3 = 42, and IQR = 42 – 21 = 21

Box plot

Data: 05 11 21 24 27 28 30 42 50 52

- 5 pt summary: {5, 21, 27.5, 42, 52};
  box from 21 to 42 with line @ 27.5
- IQR = 42 – 21 = 21.
  FU = Q3 + 1.5(IQR) = 42 + (1.5)(21) = 73.5
  FL = Q1 – 1.5(IQR) = 21 – (1.5)(21) = –10.5
- None values above upper fence
  None values below lower fence
- Upper inside value = 52
  Lower inside value = 5
  Draws whiskers



Box plot

Seven metabolic rates:

1362 1439 1460 1614 1666 1792 1867

1. 5-point summary: 1362, **1449.5, 1614, 1729**, 1867

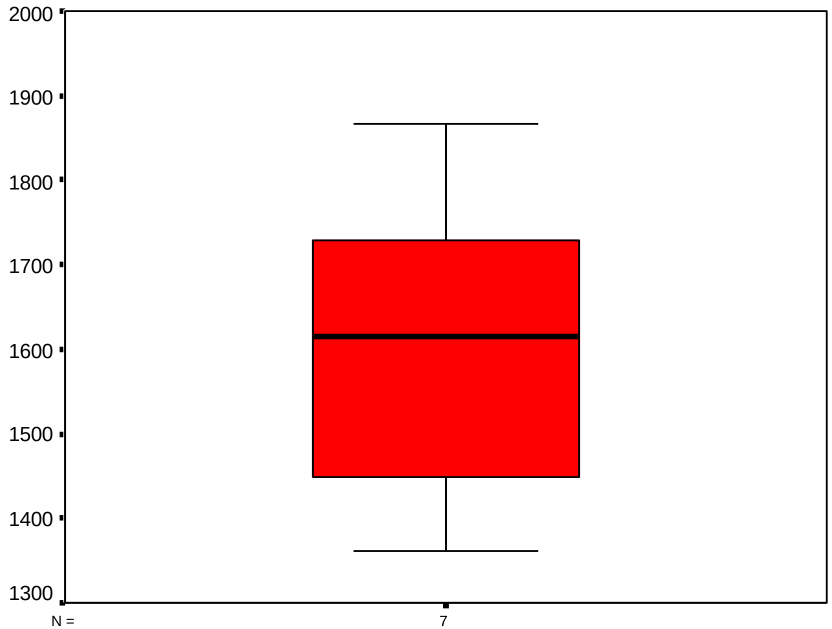
2. IQR = 1729 – 1449.5 = 279.5

F_U = Q3 + 1.5(IQR) = 1729 + (1.5)(279.5) = 2148.25

F_L = Q1 – 1.5(IQR) = 1449.5 – (1.5)(279.5) = 1030.25

3. None outside

4. Whiskers end @ **1867** and **1362**



Data source: Moore,

Report of statistical summary

- Always report a measure of central location, a measure of spread, and the sample size
- Symmetrical mound-shaped distributions $\Rightarrow$ report mean and standard deviation
- Non-normally distributed data: median, interquartile ranges

Summary

- *Mean* indicates the typical value of sample values.
- *Standard deviation* indicates the between-subjects variability of sample values;
- *Standard deviation* indicates the variability among sample means = *standard deviation of the means*.
- (There is no such thing called "standard error of the means" (SEM))
- *Coefficient of variation* indicates the relative variability (about the mean) among subjects within a sample.
- *95% confidence interval* <u>loosely</u> means the probable values of a sample with 95% probability.