

SUMMARIZING DATA

Session Objectives

- **Measures of central tendency**
 - **Mean, Median, mode**
- **Measures of dispersion – Range, standard deviation, Standard error**

Descriptive Measures for continuous data

- **Central tendency measures** – They are computed to give a “center” around which the measurements in the data are distributed.
- **Variation or variability measures** – They describe data spread or how far away the measurements are from the center.

Statistics related to continuous variables

- **Mean**
- **Median**
- **Mode**
- **Range**
- **Standard Deviation**
- **Standard Error**

Measures of Central Tendency

Central tendency measures

- **Mean – The average value
Affected by extreme values**
- **Median – The middle value
Not affected by extremes**
- **Mode – Most frequently occurring
observation, there may be
more than one mode.**

Mean

- **Average**

- **Arithmetic Mean** $= (\bar{x})$
 $= \frac{\text{sum of individual values}}{\text{number of observations}}$
 $= \frac{\sum x}{n}$

Exercise

- The diastolic blood pressure of 10 individuals was 83, 75, 81, 79, 71, 95, 75, 77, 84, 90.

-

- **Arithmetic**

$$\text{Mean} = \frac{83+75+81+79+71+95+75+77+84+90}{10}$$

$$= \frac{810}{10}$$

$$= 81$$

Median

- The data are first arranged in an ascending or descending order of magnitude
- Middle observation is located, which is called median.
- **If the number of values is odd,**
Median = middle value
- **If the number of values is even,**
Median = average of the two middle values

Median divides the data into two equal parts with 50% of the observations above the median and 50% below it.

Diastolic Blood Pressure
83
75
81
79
71
95
75
77
84

Unsorted

Diastolic Blood Pressure
71
75
75
77
79 Median
81
83
84
95

Sorted in ascending order

Exercise

- **Exercise: 1** odd no (11) of observations

- 11, 13, 15, 12, 10, 9, 2, 8, 12, 11, 10

- **Median**

- 8, 9, 10, 10, 11, 11, 12, 12, 12, 13, 15

- **Exercise: 2** even no (12) of observations

- 11, 13, 15, 12, 10, 9, 12, 8, 12, 11, 10, 12

- Arranged in ascending order

- 8, 9, 10, 10, 11, 11, 12, 12, 12, 12, 13, 15

$$\text{median} = \frac{11+12}{2}$$

Mode

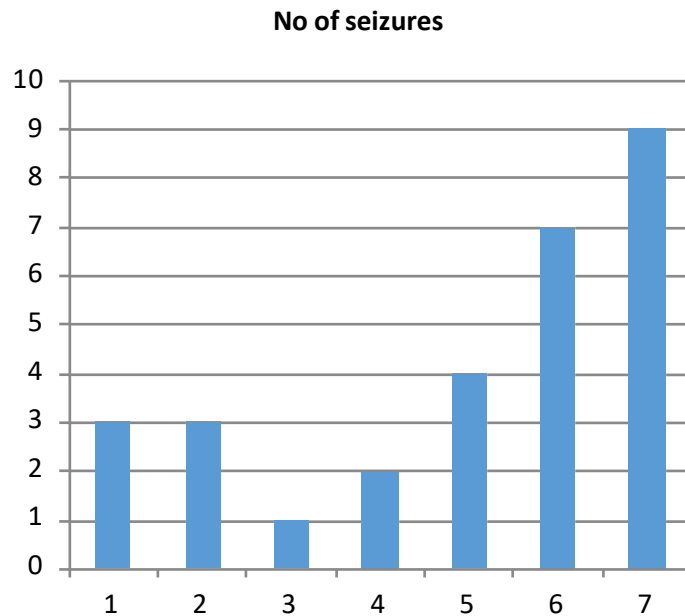
- Most frequent observation.
- The value that appears most frequently in the data set.

Exercise

11, 13, 15, **12**, 10, 9, **12**, 8, **12**, 11, 10

Mode = 12

Number of seizures/month: 3, 3, 1, 2, 4, 7, 9



•Mean? 4.1

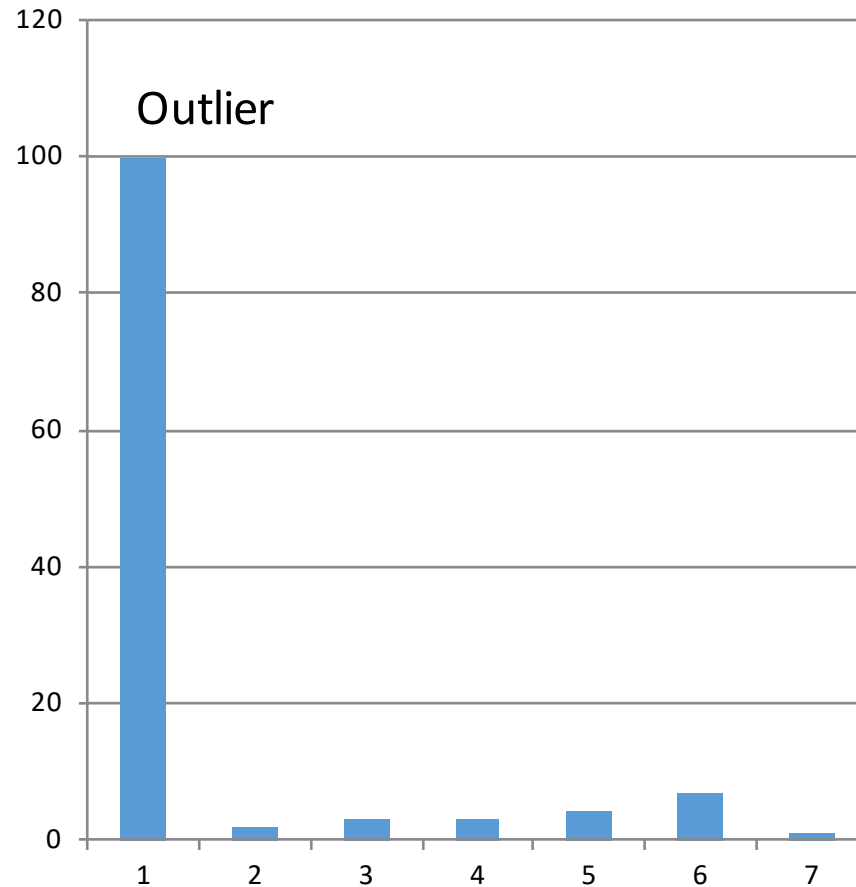
•Median? 3

•Mode? 3

What's wrong with a mean?

- Mean is sensitive to outliers (values far from the middle of the distribution)
 - Provides a falsely high or low measure of central tendency when outliers exist.
 - In such cases (look at your data), use the median as the preferred measure of central tendency.

Number of seizures/month: 100,2,3,3,4,7,1



•Mean? 17.14

•Median? 3

•Mode? 3

Measures of Dispersion

Measures of dispersion

- “**Dispersion**” also called variability, scatter, spread)
- Measure how spread out a set of data is.
- Dispersion is the scatteredness of the data series around its average.

Measures of dispersion

- **Range**
- **Standard deviation**
- **Variance**
- **Interquartile range**

Range

- **The difference between the values of the two extreme items of a series.**
- **i.e Difference between the maximum & minimum value in a set of observations.**

Exercise

- For example, from the following record of diastolic blood pressure of 10 individuals -

93, 75, 81, 79, 71, 90, 75, 95, 77, 94.

- Highest value = 95
- Lowest value = 71.
- **The Range is expressed as = $95-71=24$
& 71 to 95 .**

Characteristics of Range

- Simplest and most crude measure of dispersion.
- Affected by the extreme values.
- Gives an idea of the variability very quickly.

Standard deviation

- Tells us how individual values are deviated from and around the mean in the sample.
- Provides an index of variability.


Characteristics of Standard Deviation

- Very satisfactory and most widely used measure of dispersion.
- If SD is small, there is a high probability for getting a value close to the mean and
- If it is large, the value is farther away from the mean.
- It is less affected by fluctuations of sampling.

How to determine a SD

1. Calculate the mean
2. Calculate the difference between each value and the mean
3. Square each of the differences and sum them
4. Divide the sum by one less than the number of observations (if $n \leq 30$) and no. of observations (if $n > 30$).

Standard deviation

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$


x = Observed Value

\bar{x} = Mean of all Observed Values

$$\left[\frac{(x_1 + x_2 + \dots + x_n)}{\text{no. of Observed Values}} \right]$$

n = number of Observation

\sum = Sum

Standard deviation

- The diastolic blood pressure was as follows : 83, 75, 81, 79, 71, 95, 75, 77, 84, 90 of 10 individuals.

x	$(x - \bar{x})$	$(x - \bar{x})^2$
83	2	4
75	-6	36
81	0	0
79	-2	4
71	-10	100
95	-14	196
75	6	36
77	4	16
84	3	9
90	9	81
$\Sigma x = 810$		$\Sigma (x - \bar{x})^2 = 482$
n = 10		

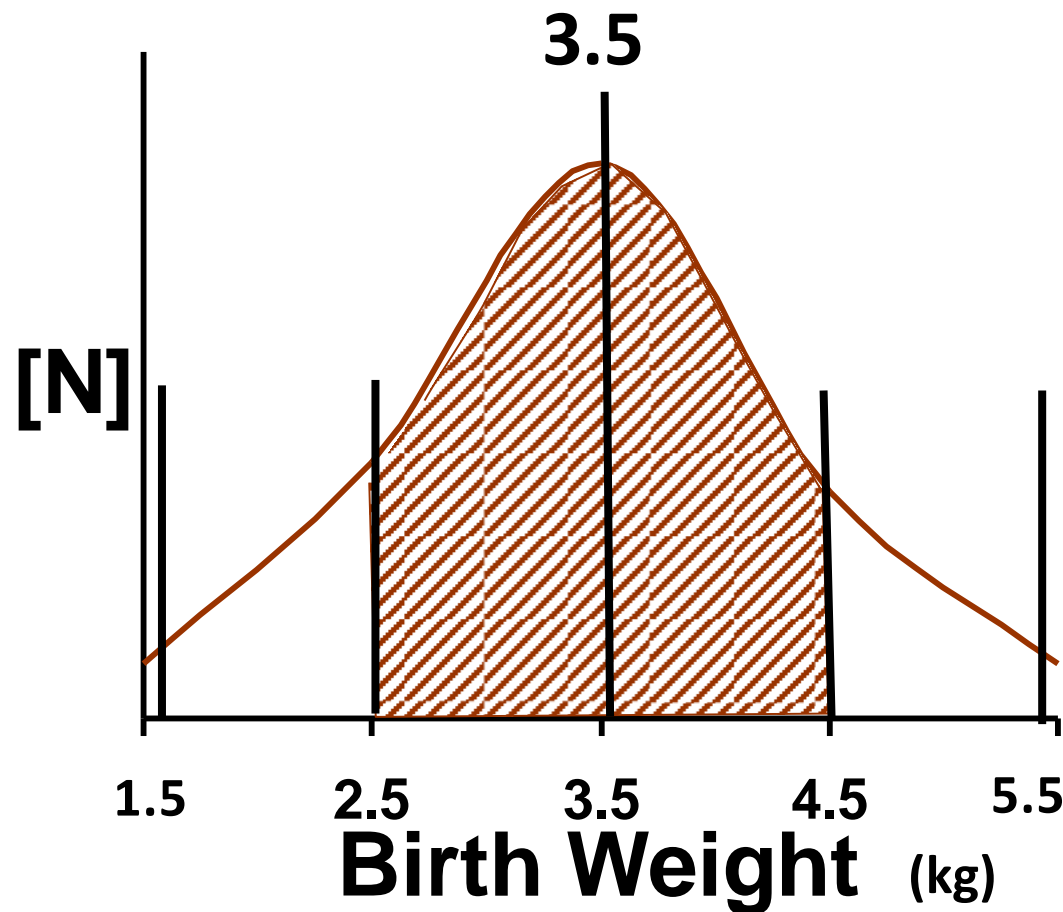
$$\text{Mean} = \frac{810}{10} = 81$$

$$\begin{aligned} \text{S.D.} &= \sqrt{\frac{\Sigma (x - \bar{x})^2}{n-1}} = \sqrt{\frac{482}{10-1}} = \sqrt{\frac{482}{9}} \\ &= \sqrt{53.55} = 7.31 \end{aligned}$$

Uses of the standard deviation

- The standard deviation enables us to determine, with a great deal of accuracy, where the values of a frequency distribution are located in relation to the mean.

Standard Deviation (SD) – for ‘Normal distribution’



Mean Birth-wt = 3.5 kg
Std Dev. = 1.0 kg

Mean ± 1 SD

$3.5 \pm 1\text{kg}$

2.5 – 4.5 kg = 68%

Mean ± 2 SD

$3.5 \pm 2\text{ kg}$

1.5 – 5.5 kg = 95%

Variance

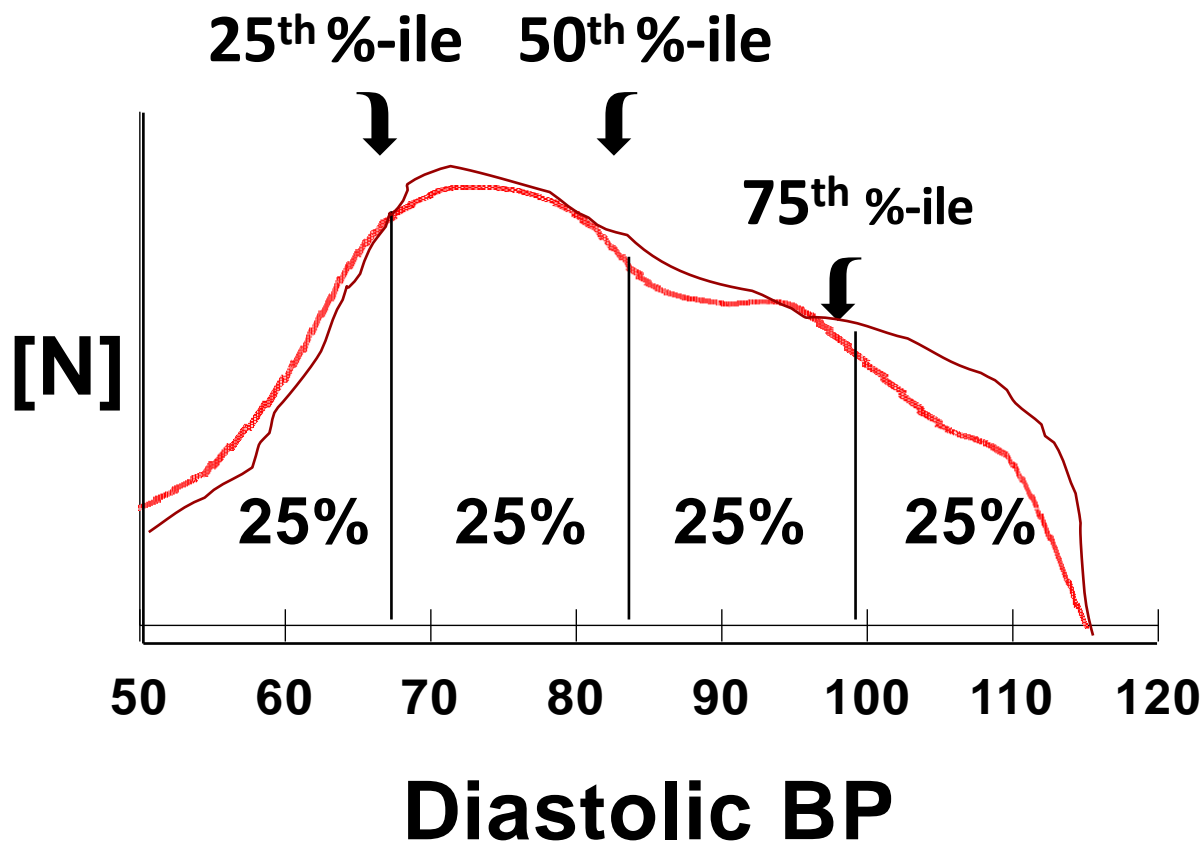
- **Variance = (SD)²**
- **Indicates the degree of variability among the observations for a given variable.**

Percentiles

- **The percentile is a number such that most $p\%$ of the measurements are below it and at most $100 - p$ percent of data are above it.**
- **Ex – if in a certain data the 85th percentile is 520 means that 15% of the measurements in the data are above 520 and 85% of the measurements are below 520.**

Percentiles - for non-normally distributed data

...% of data that fall below a specific value

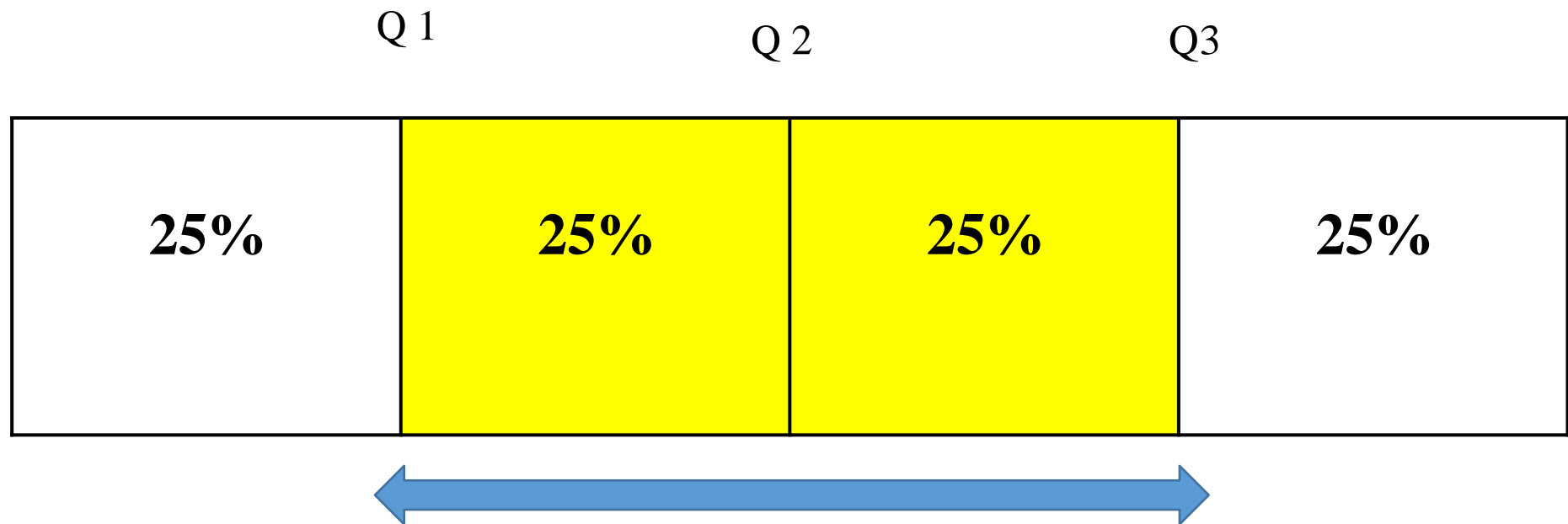


50th percentile is the
MEDIAN.

The **25th** to the **75th**
percentile is the
**INTERQUARTILE
RANGE (IQR)**.

INTERQUARTILE RANGE

“Interquartile range” is from Q1 to Q3.

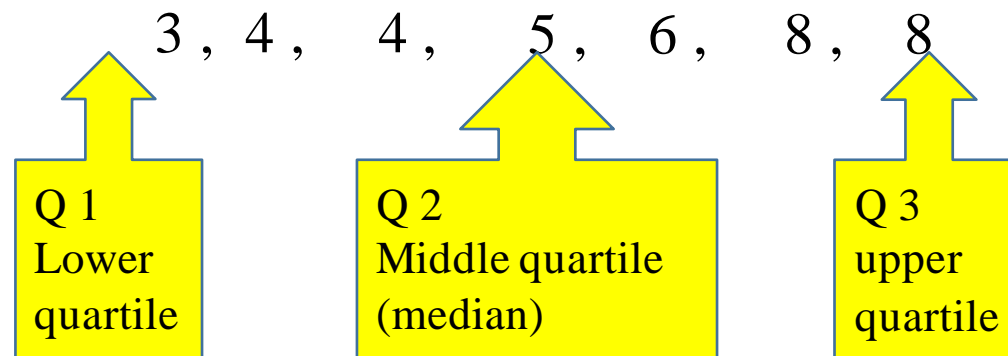


$$\text{interquartile range} = Q\ 3 - Q\ 1$$

To calculate it just subtract quartile 1 from quartile 3

Example: 5, 8, 4, 4, 6, 3, 8.

- First put the list of numbers in order.
- Then cut the list into 4 equal parts.
- The quartiles are the cuts.



Quartile (Q1) = 4

Quartile (Q2) = median = 5

Quartile (Q3) = 8

Interquartile range is $Q3 - Q1 = 8 - 4 = 4$

Standard Error

- If we take a random sample (n) from the population, and similar samples over and over again we will find that every sample will have a different mean (\bar{x}).
- If we make a frequency distribution of all the sample means drawn from the same population, we will find that the distribution of the mean is nearly a normal distribution and the mean of the sample means practically the same as the population mean (μ).

- This is a very important observation that the sample means are distributed normally about the population mean (μ).
- The standard deviation of the means is a measure of the sample error and is given by the formula σ/\sqrt{n} which is called the standard error or the standard error of the mean.

95% confidence interval

- **Approximately 2 standard errors above and below the estimate**
- **The range within which 95% of estimates from multiple samples would be expected to lie**
- **Regarded as the range within which the “true population” value probably lies (with 95% certainty)**

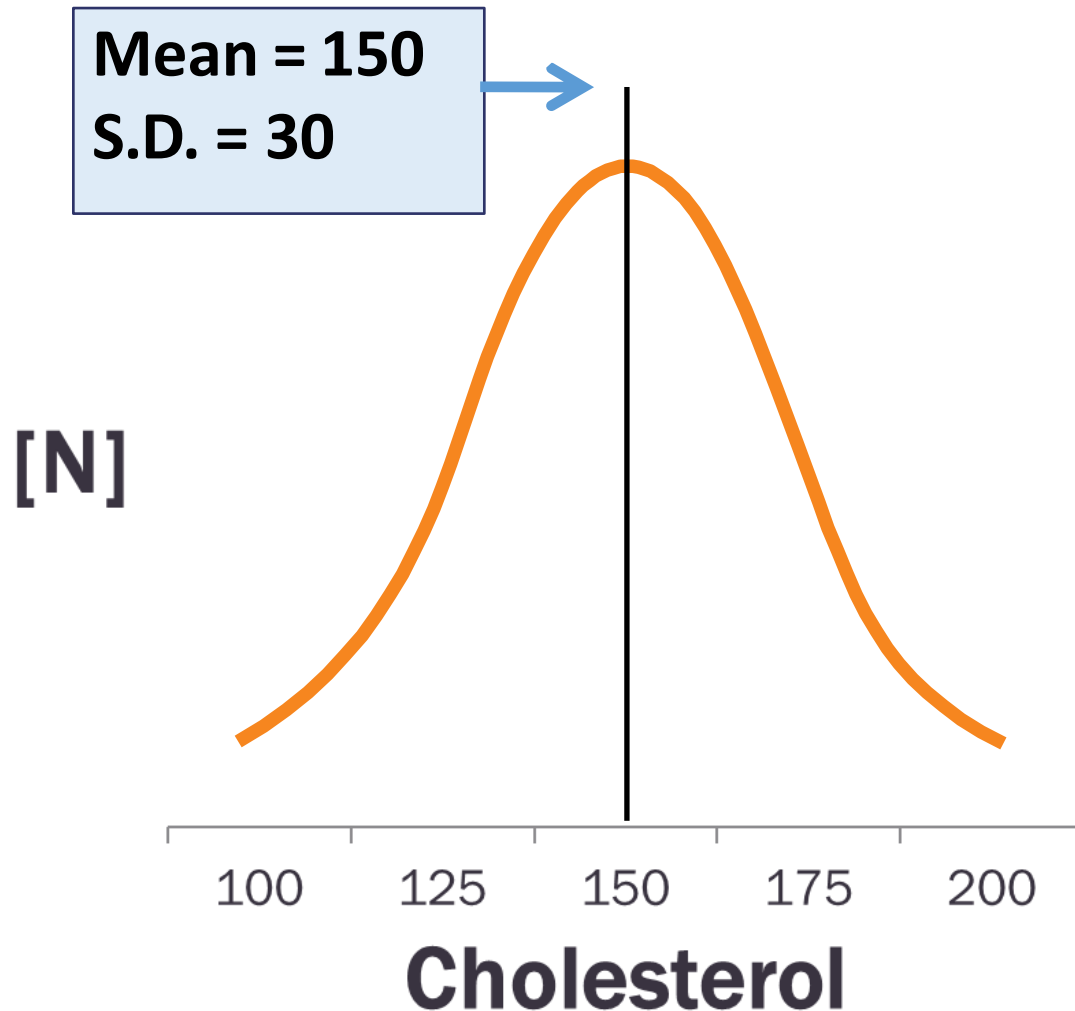
95% confidence interval of the mean

The SEM is used to describe a 95% confidence interval for an observed mean. (95% CI = Mean \pm 2 SEM)

This confidence interval narrows with larger sample size.

Since $SE = SD * \frac{1}{\sqrt{n}}$

95% CI of the mean



If based on 4 values,
95% CI is mean \pm 2 SE

$$150 \pm 2 \times 30 / \sqrt{4}$$

$$150 \pm 2 \times 15$$

$$120 - 180$$

If based on 100 values,
95% CI is mean \pm 2 SE

$$150 \pm 2 \times 30 / \sqrt{100}$$

$$150 \pm 2 \times 3$$

$$144 - 156$$

Interpreting Estimates with Confidence Intervals

- **Confident that 95% of all sample means based on the given sample size will fall within the range of the CI.**

-
- **Look at your data**
 - **For continuous data, summarize with mean (for central tendency) and SD (for dispersion) only for normal bell – shaped distributions (otherwise, use median and percentiles)**
 - **Interpret mean with confidence interval while inferring to population**
 - **For categorical data, use proportions**