Descriptive Statistics

Dr. Moh Moh Hlaing
Deputy Director/Head
Nutrition Research Division
Department of Medical Research

- **STATISTICS The** statistics is a field of study concerned with (1) the collection, organization, summarization, and analysis of data; and (2) the drawing of inferences about a body of data when only a part of the data is observed.

- **BIOSTATISTICS** The tools of statistics are employed in many fields—business, education, psychology, agriculture, and economics, to mention only a few. When the data analyzed are derived from the biological sciences and medicine, we use the term biostatistics to distinguish this particular application of statistical tools and concepts.

Using Statistics (Two Categories)

- **Descriptive Statistics**
  - ✓ Collect
  - ✓ Organize
  - ✓ Summarize
  - ✓ Display
  - ✓ Analyze
- **Inferential Statistics**
  - ✓ Predict and forecast value of population parameters
  - ✓ Test hypothesis about value of population parameter based on sample statistic
  - ✓ Make decisions

Types of statistics:

1. Descriptive (which *summarize some characteristic* of a sample)
   - Measures of central tendency
   - Measures of dispersion
   - Measures of skewness

2. Inferential (which test for significant *differences* between groups and/or significant *relationships* among variables within the sample

- There are several techniques for organizing and summarizing data so that we may more easily determine what information they contain.
- The ultimate in summarization of data is the calculation of a single number that in some way conveys important information about the data from which it was calculated.
- Such single numbers that are used to describe data are called descriptive measures.

Variables and Data

✏ Variables are the characteristics you're studying. Data are the values of those characteristics that you record.

✏Some of the variables may have produced numerical data, while other variables produced categorical data.

Variables	Data
Hb level	9.9,10.5,10.9,11.5,12.0 etc..
Household member	3,4,5,7,8 etc……
Blood Group	A, B, O , AB
Nutritional status	Low, Medium, High

Descriptive statistics: Measure of central Tendency

- A single value that is considered to be typical of the set of data as a whole.
- Measures of central tendency convey information regarding the average value of a set of values.
- A measure of central tendency is a ***single number*** that can be used to represent a set of data.

There are three different methods for measuring central tendency;

The Mean

The Median and

The Mode

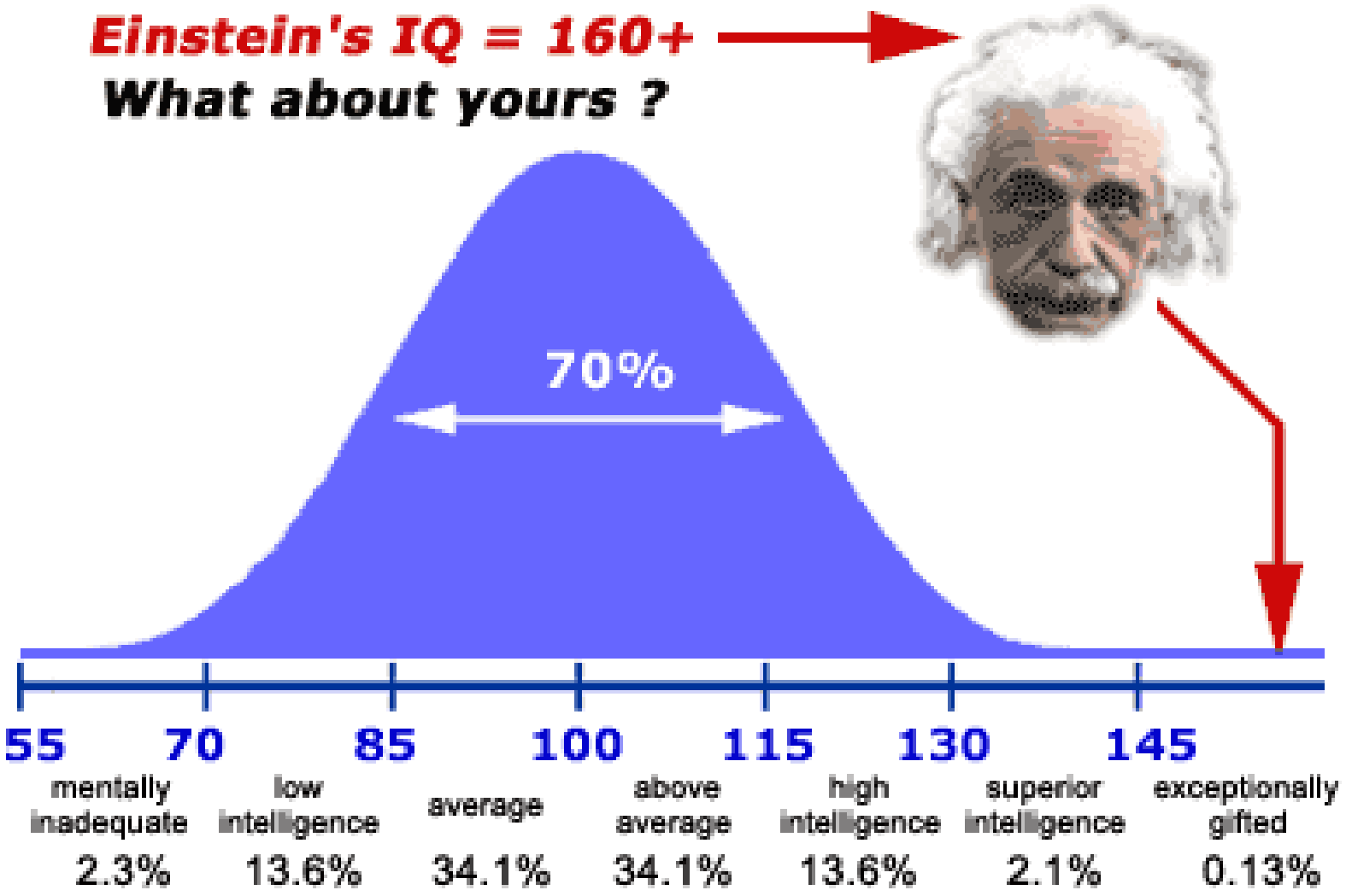
- The **mean** is the sum of all the data values divided by the number of values.
- The **median** is the middle number when the data are arranged in order.
- The **mode** is the value that occurs most frequently in the data.

Mean

- The 'average' score—sum of all individual scores divided by the number of scores
- has a number of useful statistical properties
  - however, can be sensitive to extreme scores ("outliers")
- many statistics are based on the mean

- a "trimmed mean" may be better for descriptive purposes

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



Source: www.wilderdom.com/.../L2-1UnderstandingIQ.html

The Median

- The **second** measure of central tendencies
- The goal of the median is to locate the **midpoint** of the distribution.
- 50^{th} percentile…
- There are no specific **symbols** or **notions** to identify the median
- more resistant to effects of outliers…
- First reorder the data set from the smallest to the largest
- Mark off high and low values until you reach the **middle**.
- If there 2 middles, add them and **divide** by 2

Mode

- Mode is the most frequent value or score in the distribution.
- It is defined as that value of the item in a series.
- It is denoted by the capital letter Z.
- Highest point of the frequencies distribution curve.

Examples of Measures of Central Tendency:

- For the data 1,2,3,4,5,5,6,7,8

The measures of central tendency are;

- **Mean =5**
- **Median =** 5
- **Mode =** 5

- The ***mean*** is a good summary for values that represent magnitudes, like test marks and the cost of something.

The ***median*** is best used when ranking people or things, like heights or when extreme values might affect the mean.

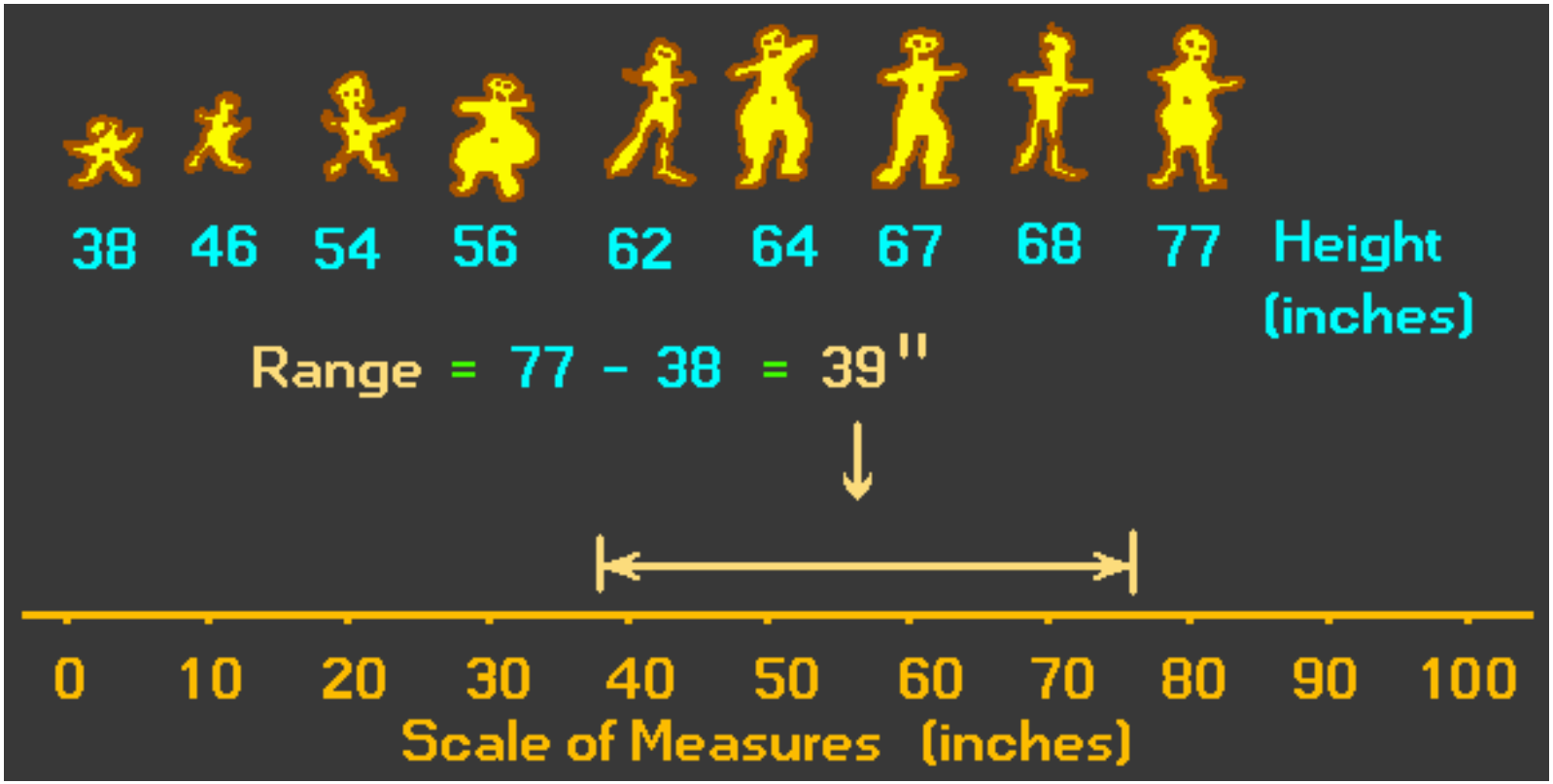
The ***mode*** is best used when finding out the most popular dress size or the most popular brand of chocolate.

Descriptive statistics: Measure of dispersion (Measures of Variability)

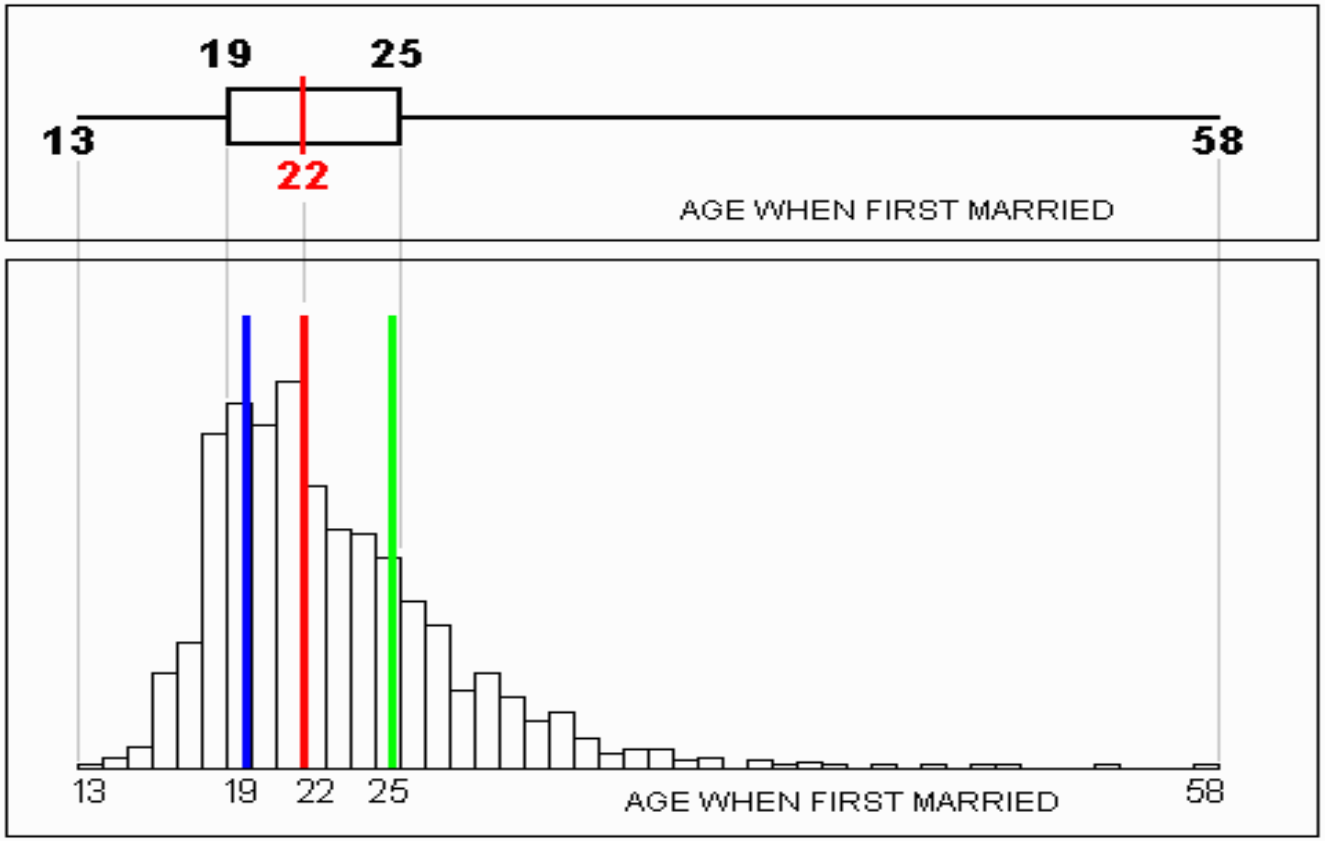
- **Range**
  - ✓ Difference between maximum and minimum values
- **Interquartile Range**
  - ✓ Difference between third and first quartile $(Q_3 - Q_1)$
- **Variance**
  - ✓ Average* of the squared deviations from the mean
- **Standard Deviation**
  - ✓ Square root of the variance

* Definitions of population variance and sample variance differ slightly.

Range



Source: www.animatedsoftware.com/ statglos/sgrange.htm



Source: http://pse.cs.vt.edu/SoSci/converted/Dispersion_I/box_n_hist.gif

Variance (S^2)

- Average of squared distances of individual points from the mean
  - sample variance
- High variance means that most scores are far away from the mean.
- Low variance indicates that most scores cluster tightly about the mean.
- The amount that one score differs from the mean is called its deviation score (deviate)
- The sum of all deviation scores in a sample is called the *sum of squares*

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

Standard Deviation (SD)

- A summary statistic of how much scores vary from the mean
- Square root of the Variance

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

Coefficient of Variation

- In some situations we may be interested in a descriptive statistic that indicates how large the standard deviation is relative to the mean.
- This measure is called the **coefficient of variation** and is usually expressed as a percentage.

COEFFICIENT OF VARIATION

$$\left(\frac{\text{Standard deviation}}{\text{Mean}} \times 100\right)\%$$

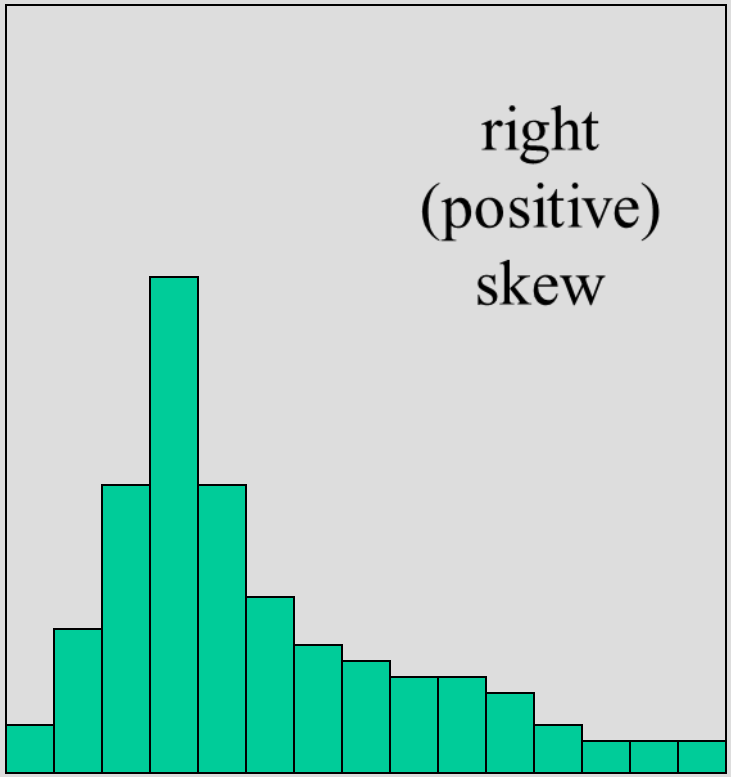
Skewness of distributions

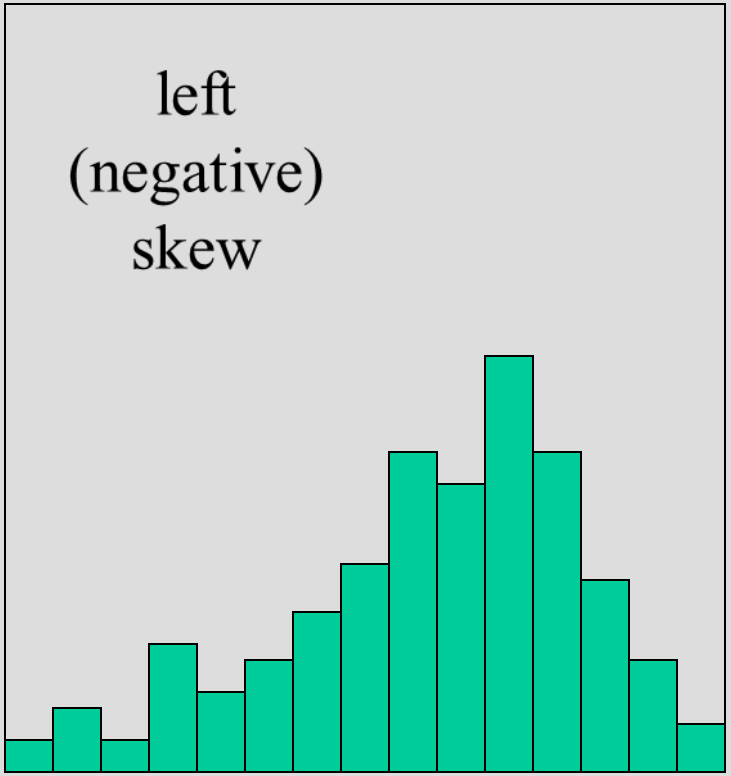
- Measures look at how lopsided distributions are—how far from the ideal of the normal curve they are
- When the median and the mean are different, the distribution is skewed. The greater the difference, the greater the skew.
- Distributions that trail away to the left are negatively skewed and those that trail away to the right are positively skewed
- If the skewness is extreme, the researcher should either transform the data to make them better resemble a normal curve or else use a different set of statistics—nonparametric statistics—to carry out the analysis

- **Interquartile Range:-**A measure of variability that overcomes the dependency on extreme values is the **interquartile range (IQR)**.

INTERQUARTILE RANGE

$$\text{IQR} = Q_3 - Q_1$$


right
(positive)
skew


left
(negative)
skew

Measures of Position

- To identify the position of a data value in a data set, using various measures of position, such as percentiles, deciles, and quartiles.

Percentiles: If data is ordered and divided into 100 parts, then cut points are called Percentiles. 25th percentile is the Q1, 50^{th} percentile is the Median (Q2) and the 75th percentile of the data is Q3.

Deciles: If data is ordered and divided into 10 parts, then cut points are called Deciles

Percentiles

- Percentiles divide the data set in 100 equal groups.

Eg. If a data value is located at the 80th percentile, it means that 80% of the values fall below it in the distribution and 20% fall above it.

Quartiles

- Quartiles divide the distribution into four
- groups, separated by Q1, Q2, and Q3.
- Q1 is the same as the 25th percentile.
- Q2 is the same as the 50th percentile or the
- median.
- Q3 is the same as the 75th percentile.

Eg.

Find Q1, Q2, and Q3 for the data set: 15, 13, 6, 5, 12, 50, 22, 18.

- Arrange in order: 5, 6, 12, 13, 15, 18, 22, 50
- Find the median or Q2. This is an even set of data, so find the two in the middle and find their midpoint. (13+15)/2 = 28 /2 = 14.
- Find Q1. This is the median of the numbers less than 14, or 5, 6, 12, and 13. (6+12)/2 = 18/2 = 9.
- Find Q3. Median of 15, 18, 22, and 50. (18+22)/2 = 40/2 = 20.

Interquartile Range (IQR)

- The difference between Q1 and Q3 (Q3 - Q1).
- Used to identify outliers.
- Used as a measure of variability in exploratory analysis.

Deciles

- Divide a distribution into 10 groups. Denoted D1, D2, D3, etc.
- Correspond to P10, P20, P30, etc.

Phases in data analysis

Descriptive phase

- Frequency tables
- Measures of central tendency for numerical variables: mean (SD), median (IQR), mode
- Measures of dispersion for numerical variables: range, percentile, variance, standard deviation
- Cross tabulation of two categorical variables
- Scatter plot for two continuous variables

Analytic phase

- Bivariate analysis
- Multivariate analysis

References

- Wayne W. Daniel. BIOSTATISTICS: A Foundation for Analysis in the Health Sciences,10th Edition.