

# Introduction to Biostatistics 1

Dr Kuhlula Maluleke  
School of Health Systems & Public Health  
University of Pretoria  
Email: [kuhlula.maluleke@up.ac.za](mailto:kuhlula.maluleke@up.ac.za)

Make today matter



UNIVERSITEIT VAN PRETORIA  
UNIVERSITY OF PRETORIA  
YUNIBESITHI YA PRETORIA

Faculty of  
Health Sciences

Fakulteit Gesondheidswetenskappe  
Lefapha la Disaense tša Maphelo



# Outline for this session

- What is statistics?
- Statistic versus parameter
- Descriptive versus Inferential statistics
- What is involved in Descriptive statistics?
- What is involved in Inferential Statistics?
- Descriptive statistics for categorical data
- Descriptive statistics for numerical data



# **Part I: Introduction to Statistics and Types of Data**

## Definition of statistics

***Statistics is the discipline that involves the collection, organization, analysis, interpretation and presentation of data***

# A statistic versus parameter

## – What is a statistic?

- *A statistic is a measure calculated from a sample (subset of population), e.g. sample mean, sample standard deviation, sample proportion*
- *We use Roman letters for statistics (e.g.  $s$  for sample standard deviation)*

## – What is a parameter?

- *A parameter is a measure calculated from a population, e.g. population mean ( $\mu$ ), population standard deviation ( $\sigma$ ), population proportion ( $\pi$ )*
- *We use Greek letters for parameters*

# Why sample or Representative sample?

- We always want to know about some measurements in the population
  - *E.g. proportion of people living with HIV in South Africa*
- However, because of limited resources we are not able to test everyone for HIV in South Africa
- Therefore, we select a subset of the population, called a sample, and calculate **statistics** from that sample
- How we select the sample is important for us to be able to generalize from **sample** to **population**

# Two branches of statistics

## **Descriptive statistics**

- *Descriptive statistics deals with describing, summarizing, and presenting data, to show patterns in the data*
- *Categorical data – frequency tables, pie/bar charts*
- *Numerical data – Mean (SD) or Median (IQR), Histograms, Box-and-whisker plots*

## **Inferential statistics**

- *Inferential statistics allows the use of sample data to generalize about the populations from which the samples were drawn*
- *95% Confidence Intervals (95%CI)*
- *Hypothesis Testing – P-values*

# What is involved in Descriptive statistics?

- Data analysis for categorical data
  - *tabulation of data*
  - *Frequency tables, Bar charts, Pie charts*
- Data analysis for numerical data
  - *measures of average or central tendency – mean, median, mode*
  - *measures of variation or spread - standard deviation, variance, percentiles, quartiles, ranges, interquartile ranges*

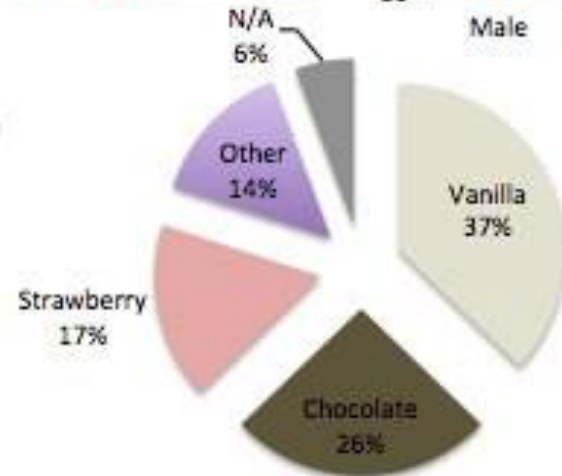


# Examples of descriptive statistics

	A	B	C	D
1	Respondent #	Age	Gender	Favorite Ice Cream Flavor
2	1	36	m	Vanilla
3	2	22	f	Chocolate
4	3	61	m	Strawberry
5	4	88	m	Other
6	5	31	m	N/A
7	6	53	m	N/A
8	7	30	f	Chocolate
9	8	64	f	Chocolate
10	9	18	m	Vanilla
11	10	16	f	Vanilla
12	11	83	m	Strawberry
13	12	16	f	Strawberry
14	13	94	m	Strawberry
15	14	55	m	Vanilla
16	15	42	f	Chocolate
17	16	18	f	Vanilla
18	17	21	f	Vanilla

Raw Data

Age	
Mean	42.6
Standard Dev.	21.9



Descriptive Statistics

# What is involved in Inferential statistics?

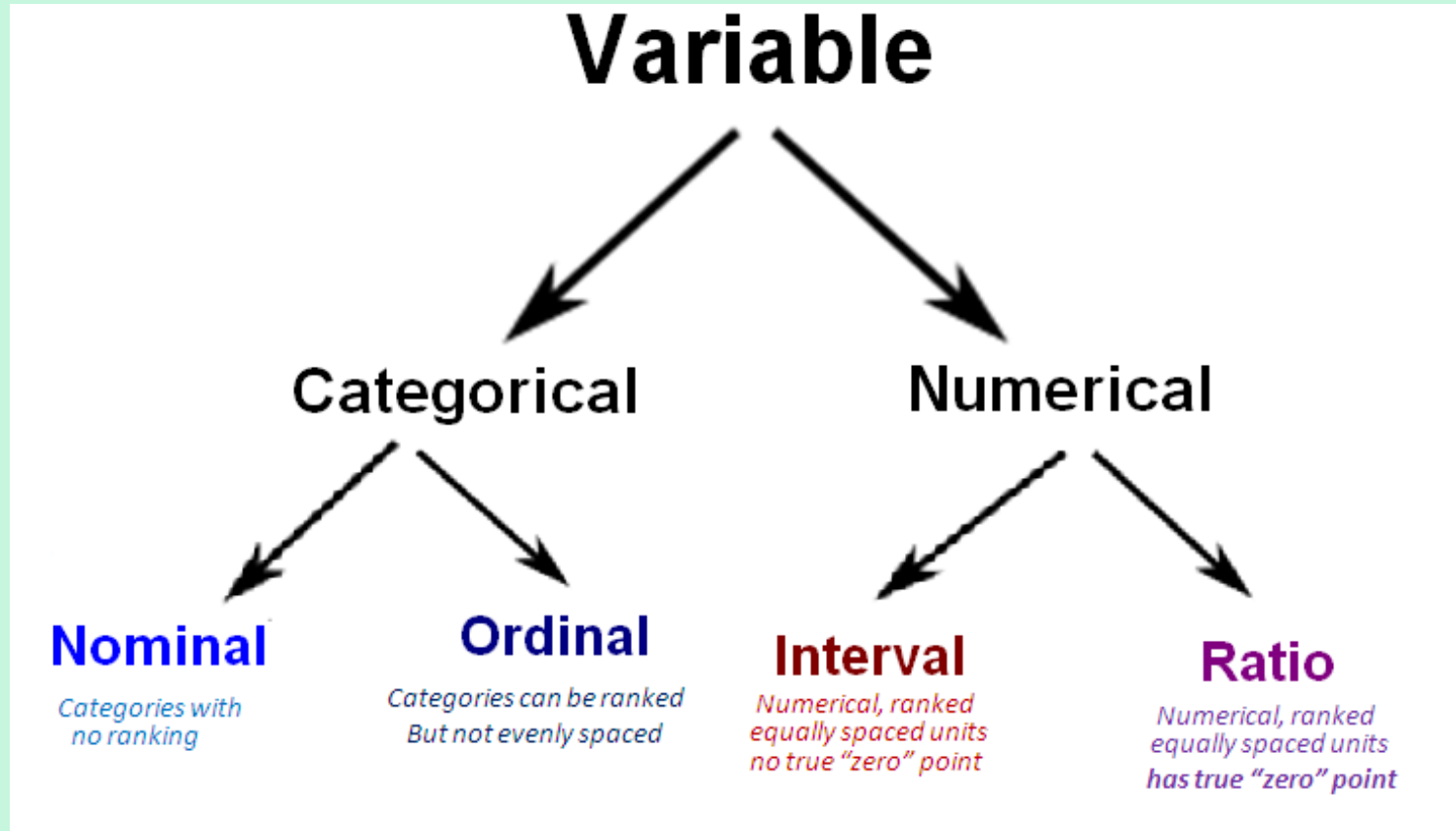
- Hypothesis testing
  - *State null and alternative hypothesis*
  - *Use sample data to calculate  $p$ -value, which is*
  - *Probability the observed is only due to chance*
  - *Make a generalization from sample to population (interpreting results)*
- 95% Confidence Intervals
  - *Use the sample to calculate 95%CI, which is*
  - *The interval in which we are 95% confident that the true population value lies within*

# Summary so far ....

- Statistics has two branches
  - Descriptive statistics
  - Inferential statistics
- Descriptive statistics deals with describing, summarizing, and presenting data
- Inferential statistics deals with generalizing from sample to population
- In most studies, we will do both descriptive and inferential statistics

# Types of Data

# Different types of data



## A “variable”

- Any characteristic that can take on *different values* for different individuals or items.
  - Age varies from person to person
  - Height, gender, wellness/illness etc
- A characteristic that we measure and describe in epidemiology

## Scales of measurement of variables



# Variables – scales of measurement

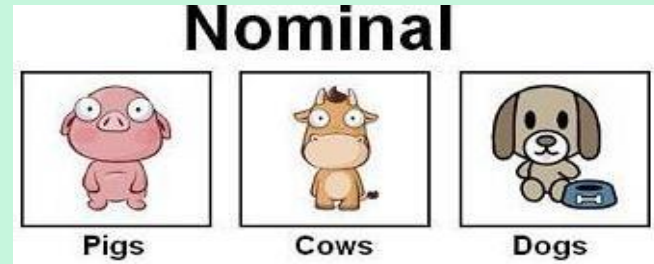
- Variables are classified into *four (4) types* depending on type of scale used to characterize them
- Variable can be measured using the
  1. **‘nominal’** or
  2. **‘ordinal’** or
  3. **‘interval’** or
  4. **‘ratio’** scale



# Nominal scale variable

- Has two or more categories
  - But there is *no basis for ordering or ranking* the categories
  - One category is *not any better or worse* than the other

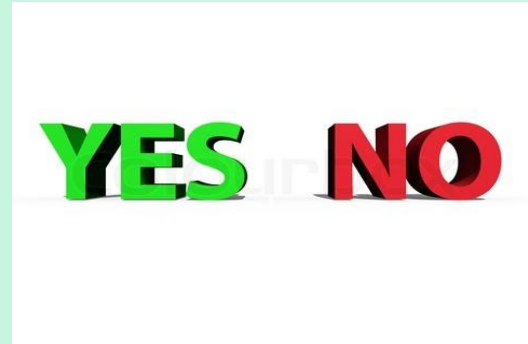
- Examples
  - Province of residence
  - Sex (male, female)
  - Marital status



- Each observation falls into one *mutually exclusive* and *exhaustive category*

# Binary variable

- A nominal variable with *two mutually exclusive* categories is also called
  - Binary variable
- Examples:
  - Alive or dead,
  - Response to question (Yes/No)
  - Sex (male/female),
  - Vaccination status (vaccinated/unvaccinated)

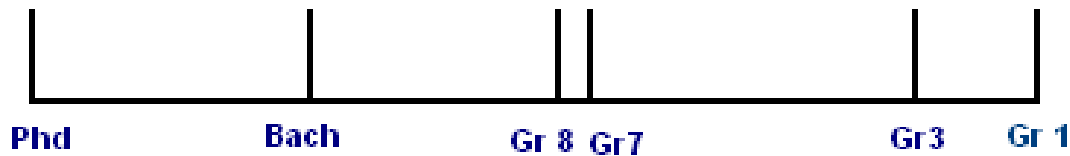


# Ordinal scale variable

- Variable has two or more categories
  - Categories can be ordered along a pre-established dimension
  - But no way of knowing how different the categories are from one another
  - We do not have equal intervals between the items
- Categories *can be ranked* but are *not evenly spaced*

## Ordinal scale variable

- Example:
  - Level of education can be categorized into – Tertiary, High School, Primary school, None, etc
  - But can we say that those who attended “high school” are *twice* as educated as those who only went to “primary”?

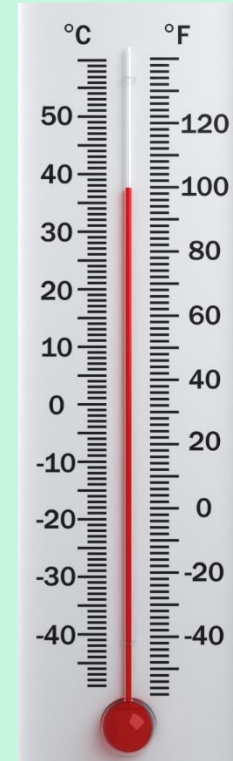


## Ordinal scale variable

- Considered stronger scale compared to nominal
  - Provides information about *relative ranking* of observations
  - Gives some idea on which observation is *better* than the other (s)
  - But does not tell us about the absolute magnitude of the difference between categories

# Interval scale variable

- Variable has a *numerical value*,
- Distance between units on the scale is equal over all levels (equally spaced units)
- Has no true (absolute) zero point
- Example: Temperature
  - Difference between 20 and 40 °C, is the same as the distance between 55 and 75 °C
  - But 0 °C does not mean absence of temperature



## Ratio scale variable

- Considered the strongest scale
- Has a numerical value
- Distance between units on the scale is equal over all levels (equally spaced)
- Has a meaningful true zero point –
  - Zero (0) of measurement indicates absence of variable
- Allows for the interpretation of ratio comparisons

# Ratio scale variable - weight

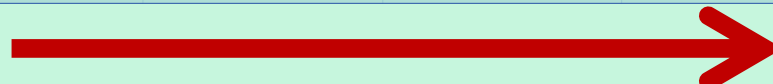
- *Equal intervals*
  - Difference between 3kgs & 5kgs is the same as between 8kgs & 10kgs
- *Ratio comparison*
  - We can say that 10kgs is twice as heavy as 5kgs
- *True zero point*
  - Zero (0kgs) indicates absence of variable





# Scales of measurement - summary

Provides:	Nominal	Ordinal	Interval	Ratio
"Counts," aka "Frequency of Distribution"	✓	✓	✓	✓
Mode, Median		✓	✓	✓
The "order" of values is known		✓	✓	✓
Can quantify the difference between each value			✓	✓
Can add or subtract values			✓	✓
Can multiple and divide values				✓
Has "true zero"				✓



Increasing level of information

# Summary for types of data & scale of measurement

- Type of data determines the statistical methods to be used
- We have numerical and categorical data
- Numerical data are either measured or counted
- Categorical data are observed
- Scales of measurement start from nominal, ordinal, interval, to ratio scale

## Exercise

What types of variables are the following?

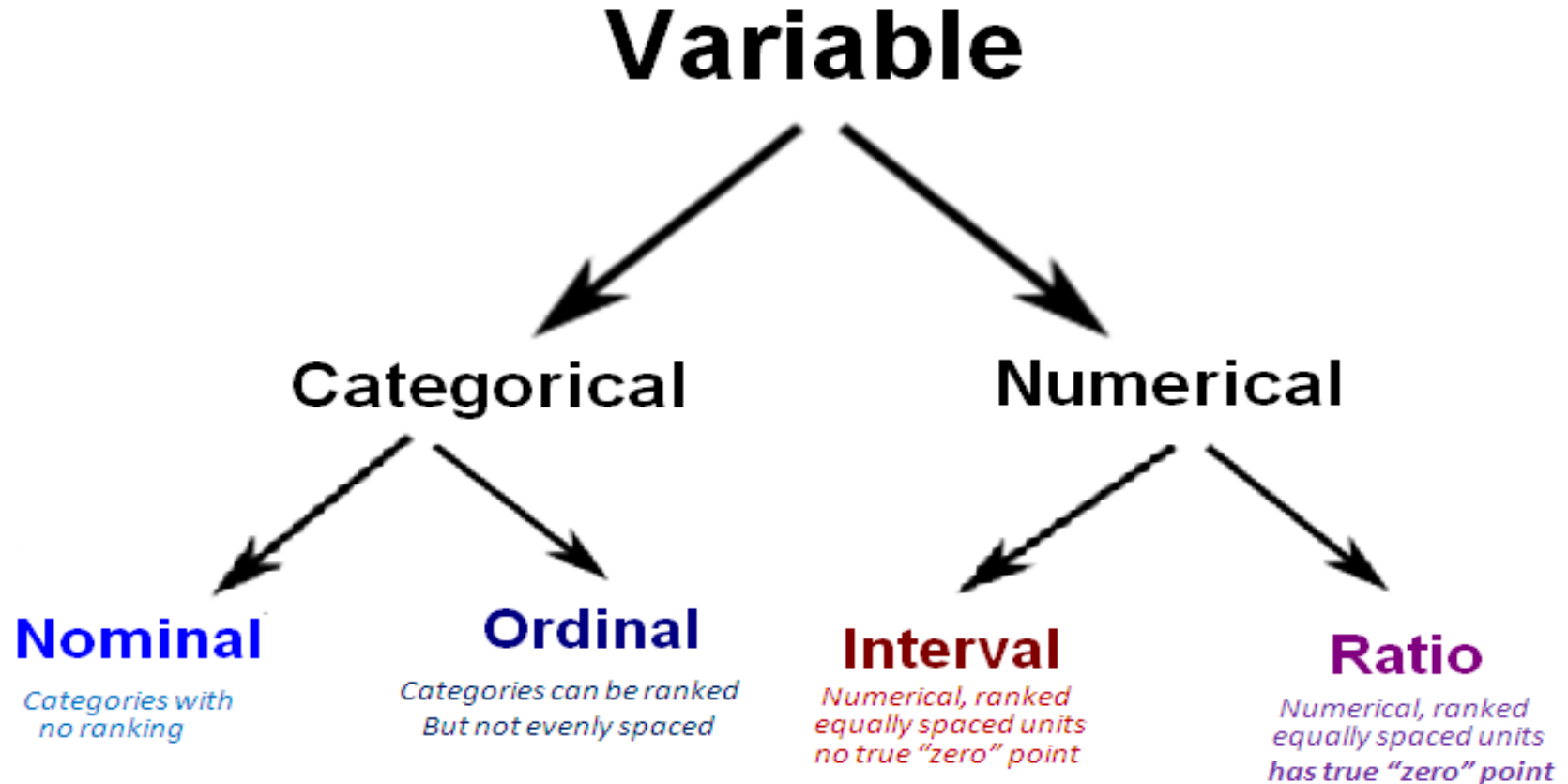
1. Number of episodes of disease in a patient per year?
2. Viral load level in a patient on antiretroviral therapy?
3. Patient's marital status?
4. Severity haemophilia (mild/moderate/severe)?
5. Reduction in blood pressure after antihypertensive treatment?
6. Sex (Male/Female)

# **Part II: Descriptive statistics for categorical and numerical data**

# **Descriptive statistics for categorical data**

***The type of statistical methods to be used depend on the type of data***

## Recall different types of data



# Categorical variables

Several categories with no order

- blood group: A, B, AB, O

Several categories with order

- health status: poor/ moderate/ good/ excellent

Two categories only (also called a binary variable)

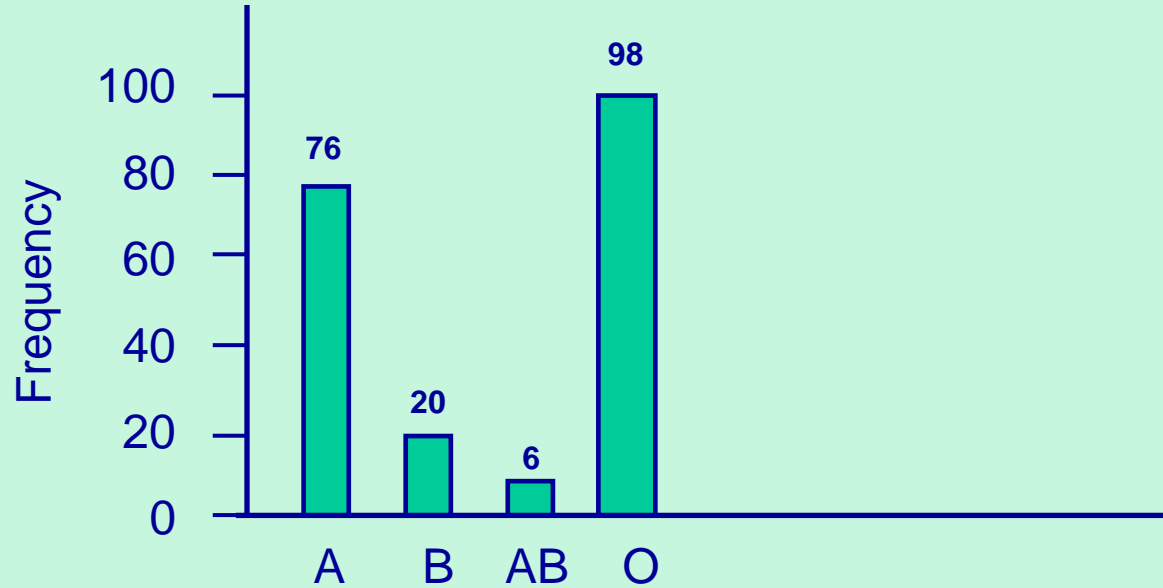
- dead/ alive

## Frequency distribution of blood group

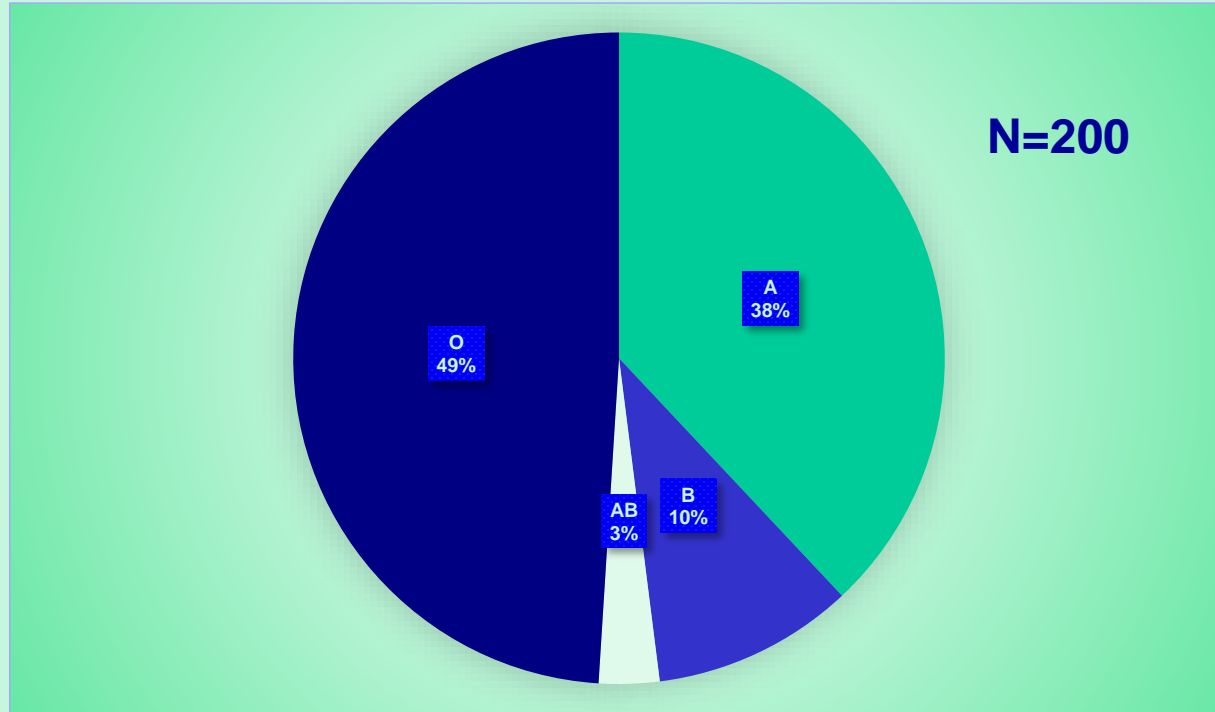
Blood group	Frequency	%
A	76	38
B	20	10
AB	6	3
O	98	49
Total	200	100



## Bar chart of blood group



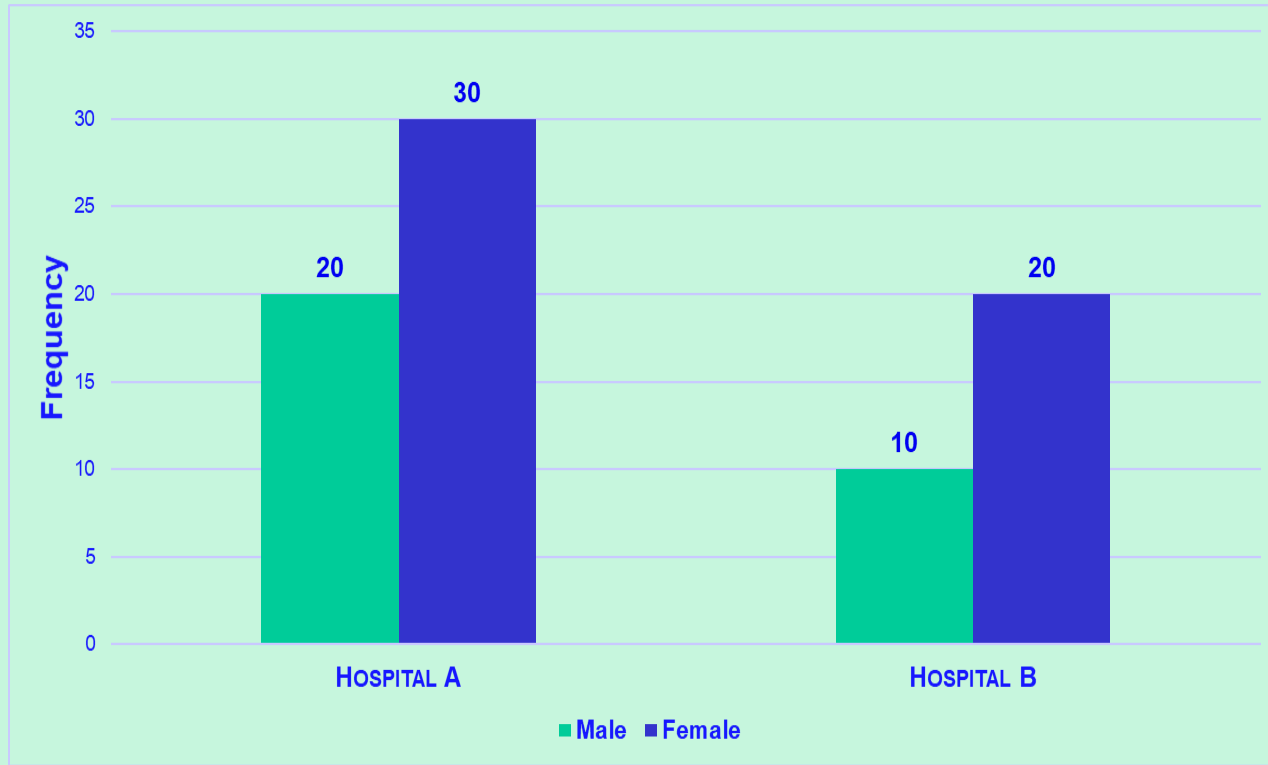
## Pie chart for blood group



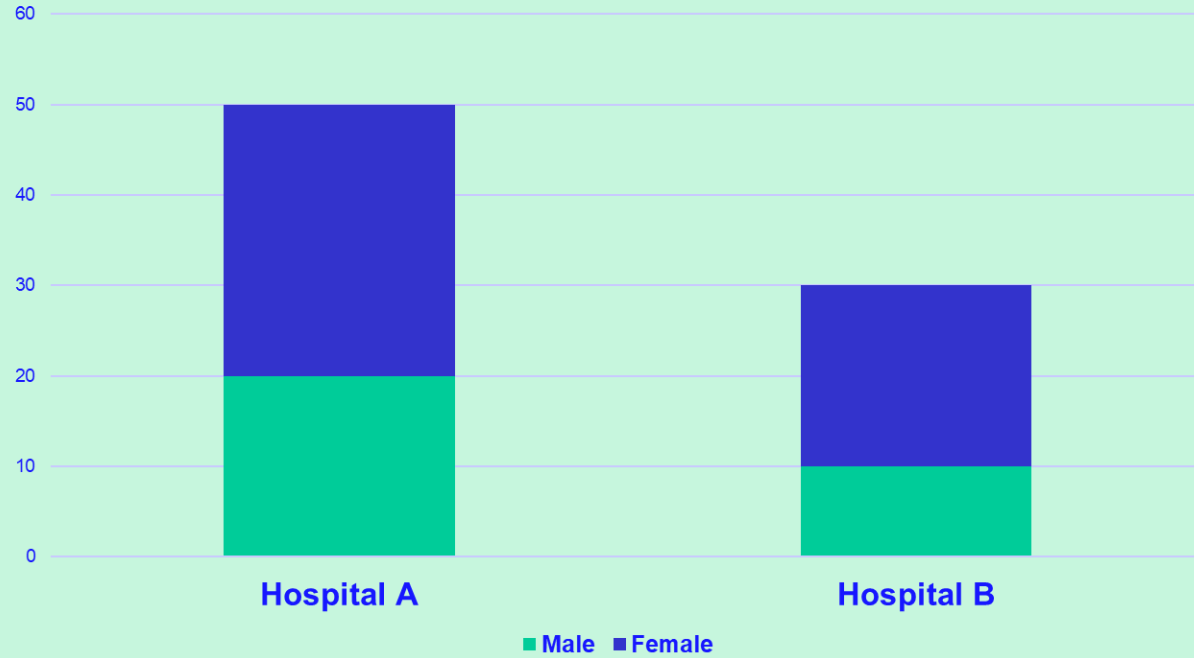
## Two categorical variables

	Hospital A	Hospital B
Male	20	10
Female	30	20

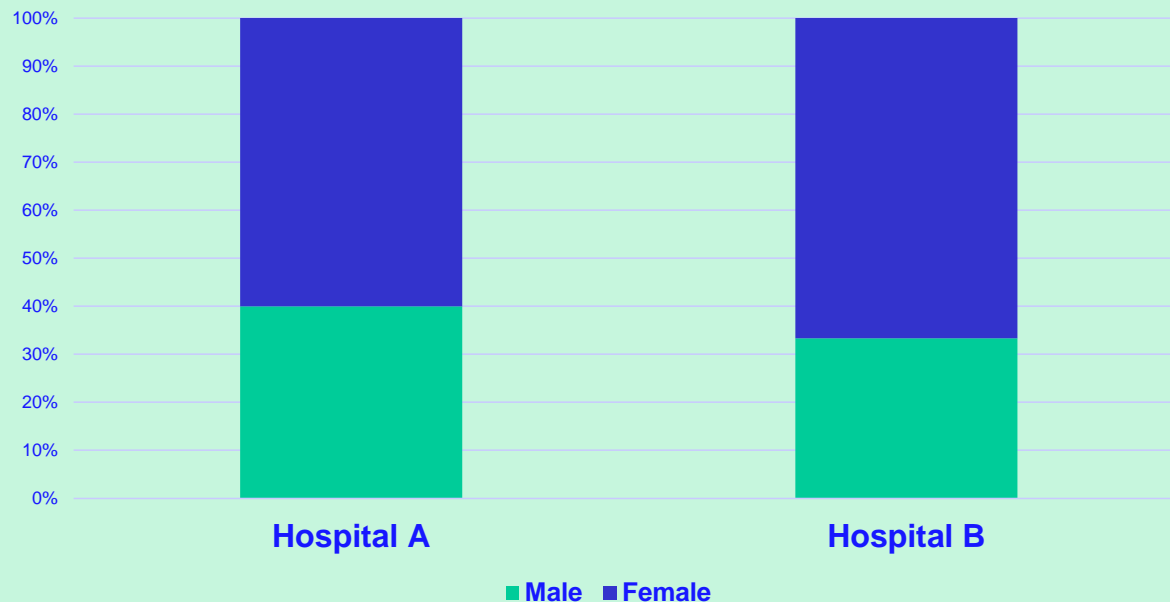
# Clustered bar chart



# Stacked bar chart



# 100% Stacked bar chart



# **Descriptive statistics for numerical data**

***The type of statistical methods to be used depend on the type of data***

**Now let's consider numerical  
variables...**



# Numerical variables

## Measured (continuous variable)

Where a measuring instrument is used e.g.

- birth weight in kg

Variable can take any values within a range

## Count (discrete variable)

Variable can only take certain discrete values, e.g.

- Number of hospital admissions last year

$0, 1, 2, 3, \dots$

# Continuous measured variable

An example is head circumference in new-born babies, measured in cm.

Note that the precision of the measurement depends on the instrument used

Creating a frequency distribution needs some preliminary work

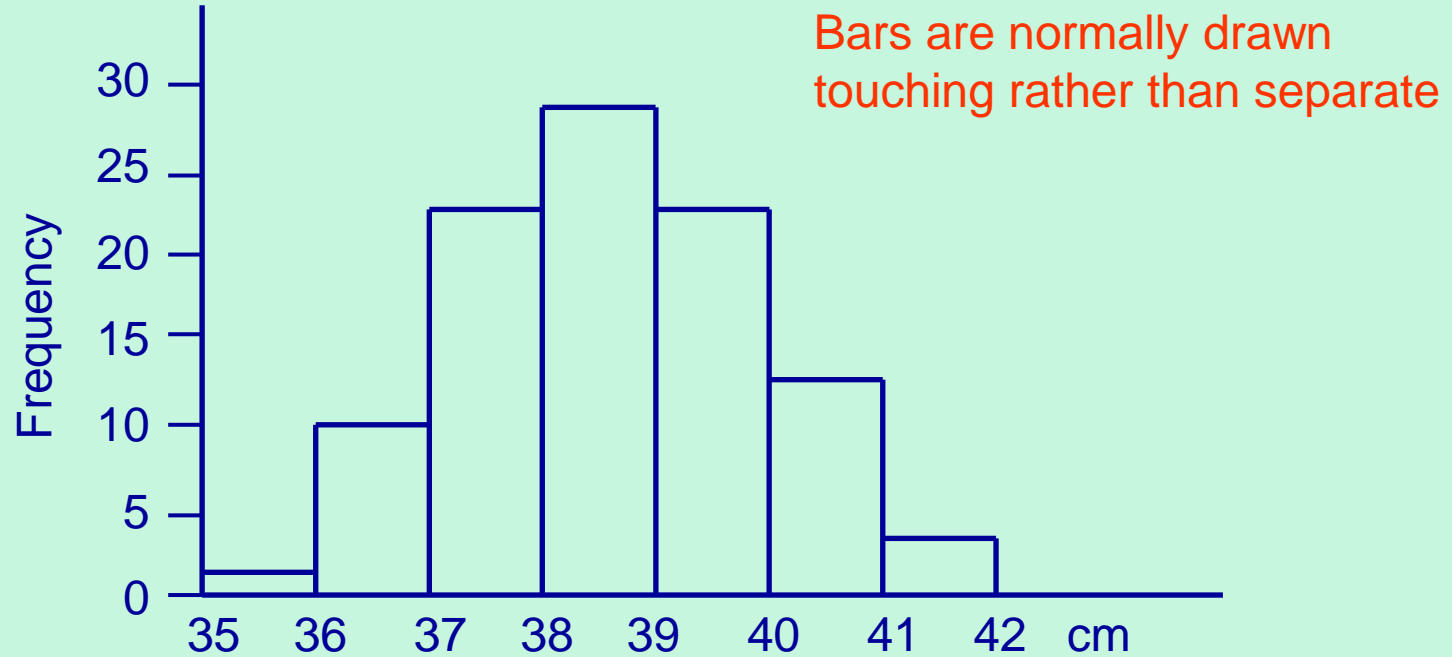
## Head circumference of 100 babies (cm)

38.6, 39.1, 38.0, 36.2, 37.7, 39.0, 38.4, 40.3, 39.8, 38.9,  
37.5, 39.5, 37.2, 40.4, 37.5, 38.8, 37.5, 36.8, 37.3, 39.2,  
38.1, 37.9, 38.8, 39.0, 39.3, 40.7, 38.8, 38.1, 37.5, 38.1,  
38.3, 38.7, 38.5, 39.0, 38.1, 38.5, 39.2, 37.9, 40.6, 36.2,  
37.2, 36.2, 37.4, 35.8, 41.0, 39.4, 38.5, 38.6, 40.9, 39.5,  
39.9, 39.6, 39.4, 38.9, 38.6, 36.1, 39.0, 38.5, 40.1, 36.4,  
40.4, 38.6, 39.7, 41.0, 37.5, 39.5, 38.1, 37.7, 40.7, 40.6,  
37.6, 40.4, 37.5, 38.5, 38.5, 36.5, 37.8, 40.0, 37.5, 36.8,  
38.0, 36.8, 39.0, 36.2, 37.7, 37.8, 37.5, 37.1, 38.6, 38.7,  
41.0, 39.5, 38.3, 39.1, 40.9, 37.3, 38.7, 38.8, 39.3, 39.0

## Frequency distribution

Head circumf (cm)	Frequency
35.0 - 35.9	1
36.0 - 36.9	10
37.0 - 37.9	23
38.0 - 38.9	28
39.0 - 39.9	23
40.0 - 40.9	12
41.0 - 41.9	3
Total	100

# Histogram of head circumference



# Measure of average

We want this to represent all the observations in a single value

**Mean and median** are common measures

**Mode** is a measure that is rarely useful

# Measures of average

**Mean:**  $\frac{\text{sum of values}}{\text{Number of obs}} = \frac{\sum x_i}{N}$

**Median:** value of middle observation when all observations are arranged in size order

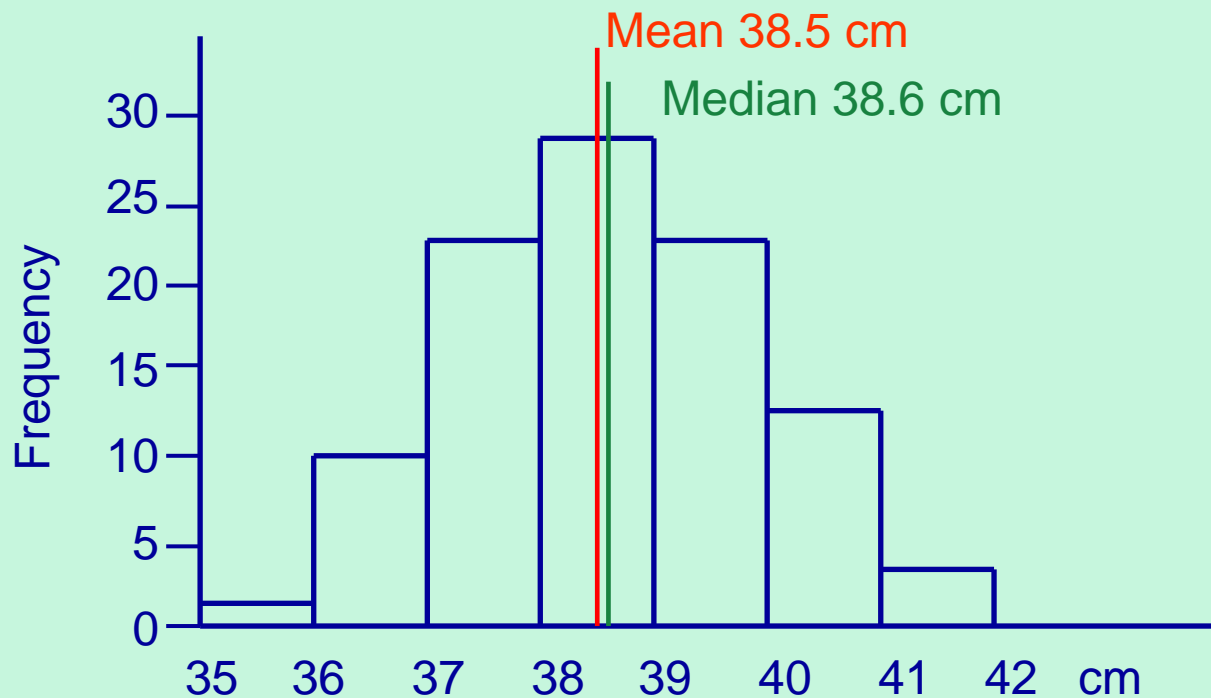
3,4,5,6,8,8,9,10,11,11,12,14,15,16,16,18,19,21



median of these 18 observations is midway between the 9th and 10th observation

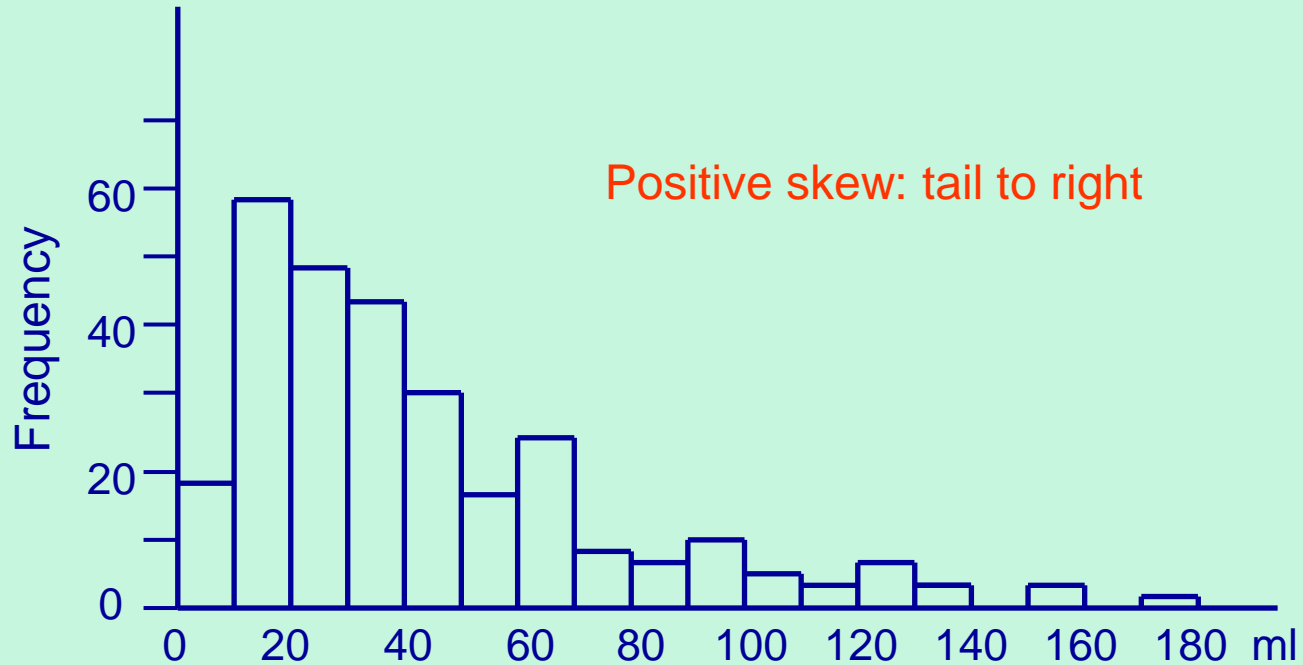
**Mode:** most common value

# Histogram of head circumference

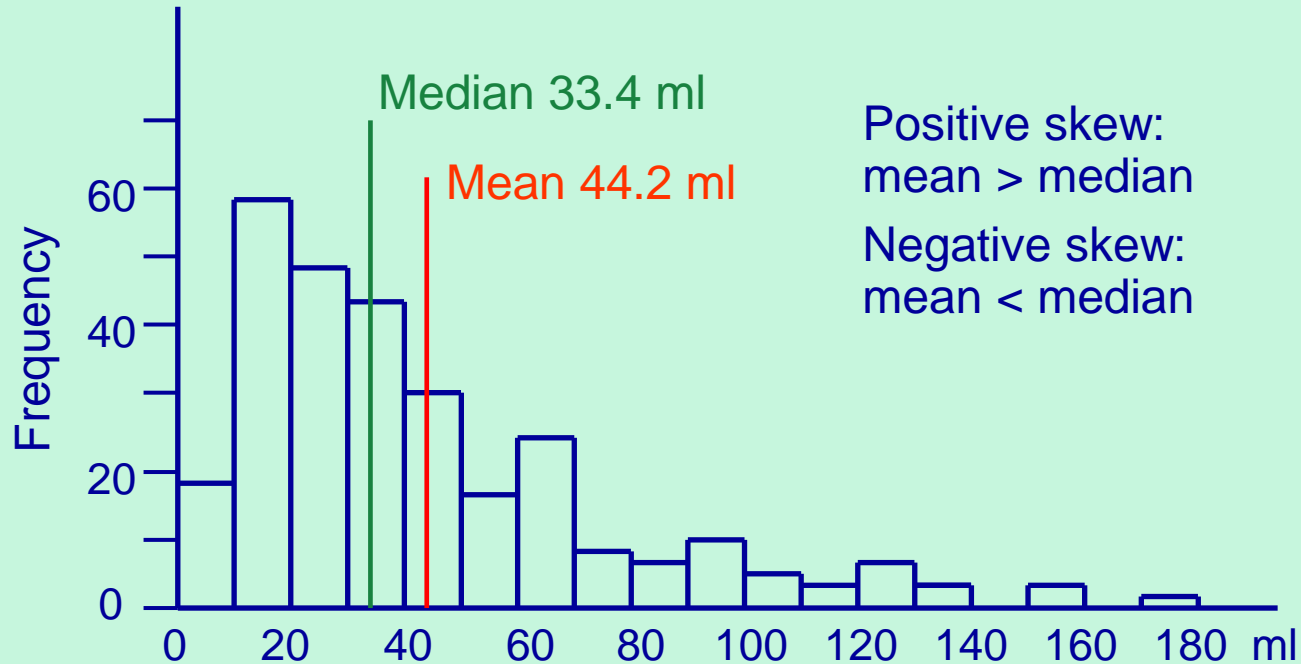




# Compare histogram of menstrual blood loss (ml)



# Histogram of menstrual blood loss (ml)



# Mean vs. median (1)

**Mean** is calculated using arithmetic, so variable should be in an interval scale (but we sometimes calculate means for ordinal scales)

If used with discrete variables the mean can produce odd results. e.g. mean number of children in family = 3.2

## Mean vs. median (2)

**Mean** is much affected by outlying values - median is not. For very skewed data, the median is the more stable measure and gives a better picture of the 'average' than the mean.

**Median** can be calculated even if you don't know precisely what the extreme values are

**Mean** is much more useful than median in further statistical calculations

# Measures of variation

There are several measures to encapsulate the variation in a set of observations

**Range, percentiles, and interquartile range** are all derived from ranking the observations

**Standard deviation** is calculated from the values of the observations

# Range

The difference between the maximum and minimum values in a dataset.

E.g. for head circumference

Maximum value = 41.0 cm

Minimum value = 35.8 cm

The range = max value – min value  
 $= 41.0 - 35.8 = 5.2$

# Percentiles

Arrange the N observations in order of size.

Common percentiles are 5th, 25th, 50th, 75th, 95th

The 25th percentile is the value with rank 25% of  $N+1$

e.g. Suppose you have 23 observations:

1,3,3,3,5,5,6,7,8,8,8,9,10,11,12,13,14,14,15,17,20,22,25

25th  
percentile:  
value with  
rank 6

50th  
percentile:  
value with  
rank 12

75th  
percentile:  
value with  
rank 18

## Interquartile Range (IQR)

The 25th percentile is often referred to as  $Q_1$  and the 75th percentile as  $Q_3$

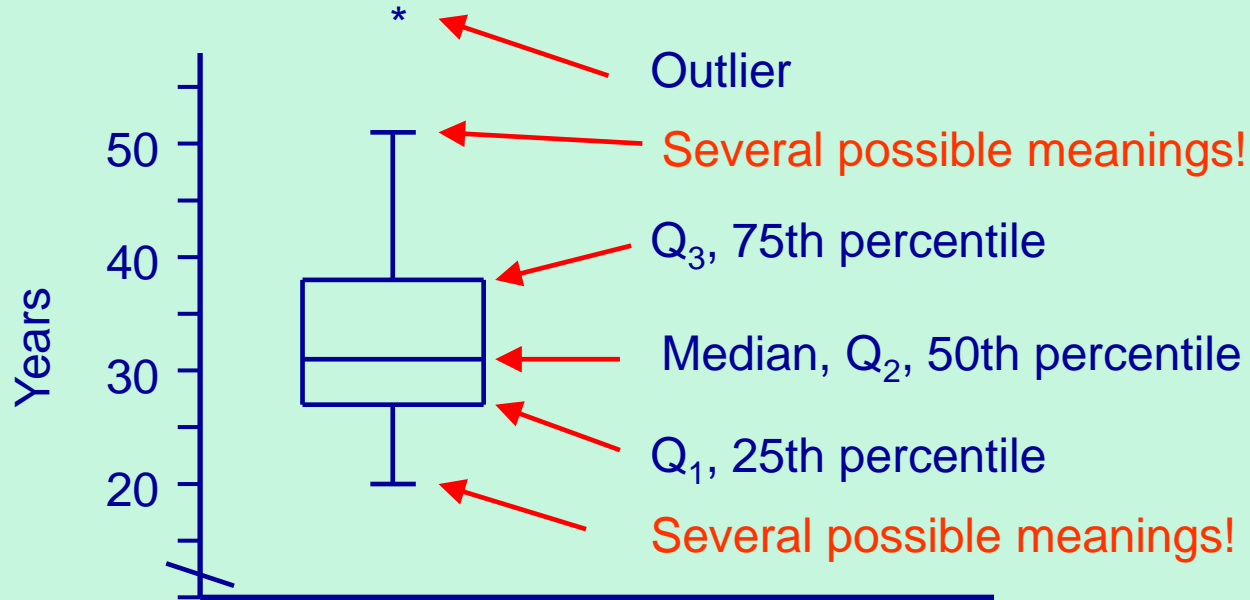
The IQR is the range from  $Q_1$  to  $Q_3$ : in our example 5 to 14

The IQR includes the middle 50% of observations, and excludes extreme (perhaps unrepresentative) values



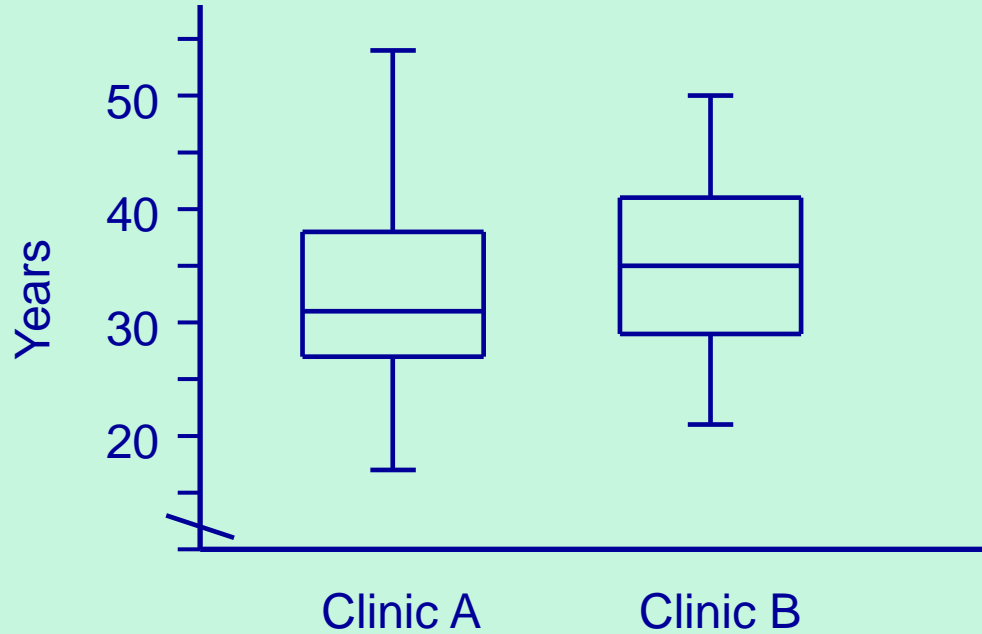
# Box-and-whisker plot

Age distribution of patients in HIV clinic



# Box-and-whisker plot

Age distribution of patients in two HIV clinics



## Standard deviation (SD)

1. Find the mean  $\bar{x}$
2. Find the deviation of each observation from the mean  $(x_i - \bar{x})$
3. Square each deviation  $(x_i - \bar{x})^2$
4. Add up the squared deviations  $\Sigma(x_i - \bar{x})^2$
5. Divide by  $n-1$ . This is the **variance**

6. Take the square root. **SD** = 
$$\sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}}$$

## Calculating SD

x	$(x-\bar{x})$	$(x-\bar{x})^2$
10	$(10-14)=-4$	$(-4)^2=16$
12	$(12-14)=-2$	$(-2)^2=4$
20	$(20-14)=6$	$(6)^2=36$
Total		$\Sigma(x-\bar{x})^2= 16+4+36=56$

$$\text{Mean} = \bar{x} = (10+12+20)/3 = 42/3 = 14$$

$$\text{Variance} = [ \Sigma(x-\bar{x})^2 ] / (n-1) = 56 / (3-1) = 56/2 = 28$$

$$\text{Standard Deviation} = \text{SD} = \sqrt{28} = 5.29$$

## How do you interpret the SD? (1)

The SD has no immediately obvious interpretation, but understanding it comes with experience

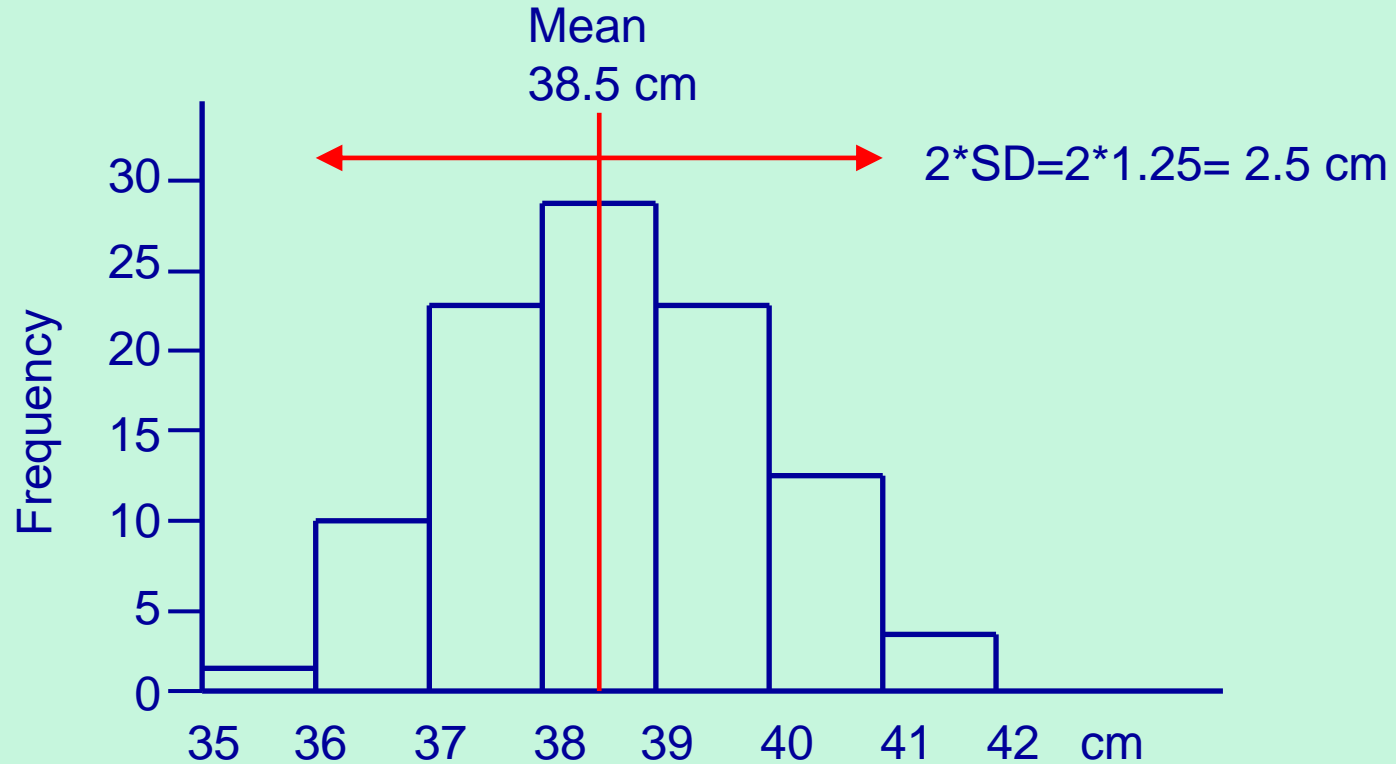
The bigger the SD the more variability in the data

## How do you interpret the SD? (2)

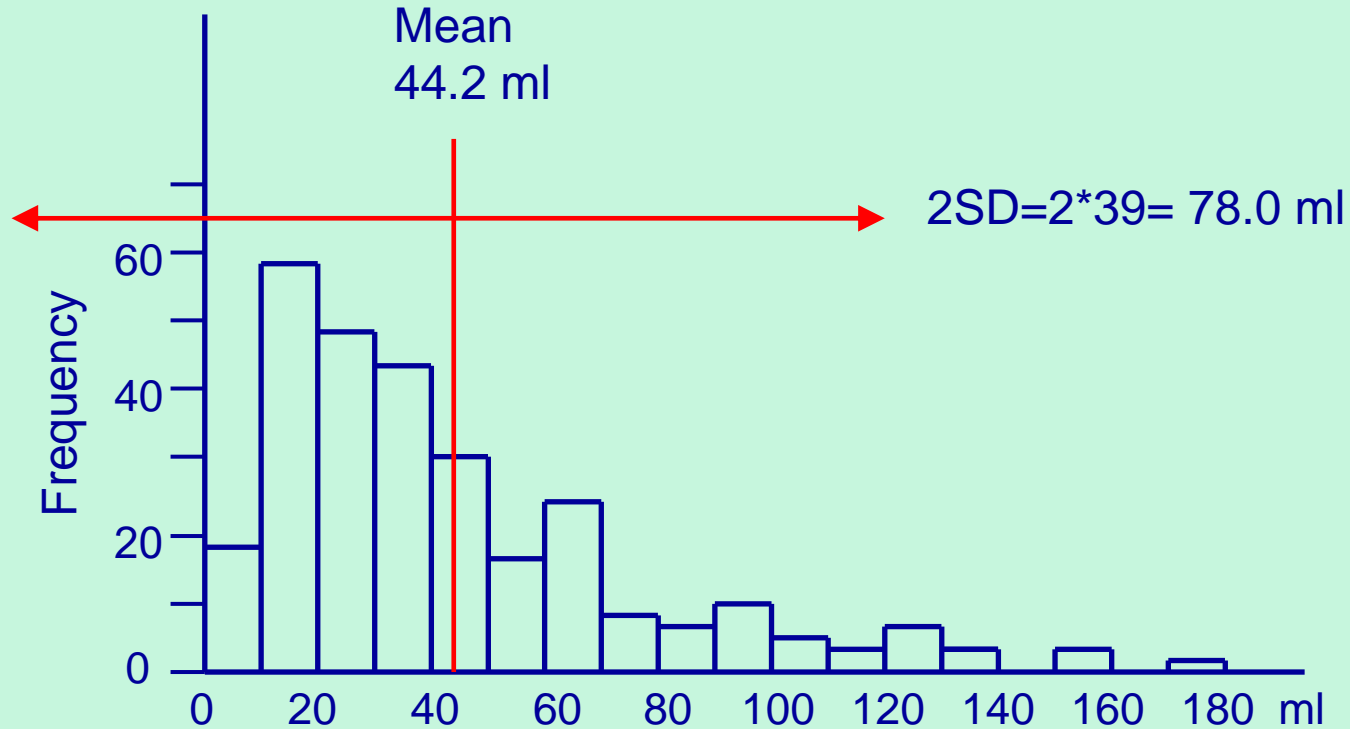
If the distribution of the data is symmetrical, the large majority of a set of observations will usually be within 2 SD of the mean

If the data come from a Normal distribution, 95% of observations will be within 2 SD of the mean (more on this next session)

# Head circumference (cm), SD 1.25



## Menstrual blood loss (ml), SD 39





## SD in a skew distribution

For the blood loss data,  
mean + 2SD = 122.2 ml, but  
mean - 2SD = -33.8 ml, an impossible value!

Simply from knowing the mean and SD you  
can tell that this distribution is skewed

## Comparison of IQR and SD

For describing the variability of the data, the SD is useful for symmetrical distributions, but the IQR is more informative for very skewed distributions.

For further statistical calculations the SD is essential.

# Summary

- There are two branches of statistics: Descriptive and Inferential
- Descriptive statistics is what we have covered – deals with data summary and presentation
- The type of analysis depends on whether variables are categorical or numerical
- Categorical data are summarized using frequency tables, pie charts and bar charts
- For normally distributed numerical data, use Mean and SD
- For skewed data, use Median and IQR

# Thank You



UNIVERSITEIT VAN PRETORIA  
UNIVERSITY OF PRETORIA  
YUNIBESITHI YA PRETORIA