WEBINAR ON STATTISTICS FOR MICROBIOLOGISTS

JAETECH GLOBAL

Statistical Methods for Antimicrobial Susceptibility Testing Data

Dr. Md. Kamruzzaman

Biostatistician, Seoul National University

August 05, 2023

By end of this presentation, you will be able to learn:

- About Statistics, and its type and application in Microbiological study
- Population and Sample
- Variable: Quantitative, Qualitative
- Graphical presentations: Bar, Histogram, Line graphs
- Descriptive statistics
- Correlation and regression
- Inferential statistics (Hypothesis testing)

What is Statistics?

- **Statistics:** Statistics is concerned with scientific methods for collecting, organizing, summarizing, presenting and analyzing sample data from a specified population of interest as well as drawing valid conclusion.
- Unfamiliar term: population and sample
- For example, *E. coli* infection is by eating contaminated food, such as: grounded beef.
- The WHO reports that a growing number of infections, including pneumonia, tuberculosis, and salmonellosis, are getting harder to treat as antibiotics become less effective.

Types of Statistics?

- **Types of Statistics:** There are two types of statistics:
    1. Descriptive statistics
    2. Inferential Statistics
- **Descriptive Statistics** Descriptive statistics consists of methods for organizing, displaying, and describing data by using tables, graphs, and summary measures.
- **Inferential Statistics:** Inferential statistics consists of methods that use sample results to help make decisions or predictions about a population

Population and Sample

- **Population**: A statistical population is the collection of all items of interest in a particular study.
- **Sample:** A sample is a representative part of population.
- **Example**: For example, *E. coli* infection is by eating contaminated food, such as: grounded beef.
  - ✓ Population: individual who eat contaminated food.
  - ✓ Sample: *E. coli* infected individual

Variable

Variable: A variable is a characteristics that can vary from one individual to another, time to time and place to place.

Example: Commonly used variables for AMR data are:

- ✓ Patient ID
- ✓ Age
- ✓ Sex
- ✓ Species
- ✓ Sample type
- ✓ Date of admission
- ✓ Organism
- ✓ Minimum inhibitory concentration (MIC).
- ✓ Sample Collection Date
- ✓ Department

Variable …

- Variable can be classified into ways
  - ✓ Qualitative variable (categorical Variable)
  - ✓ Quantitative variable (numerical variable)

Quantitative Variable

- **Quantitative variable**: A variable that can be measured numerically.
- For example,
  - ✓ number of patients in a hospital,
  - ✓ number of death in a hospital,
  - ✓ age of patients,
  - ✓ monthly income of patients etc.

- Quantitative variable can be further classified as: discrete and continuous variable.

Quantitative Variable

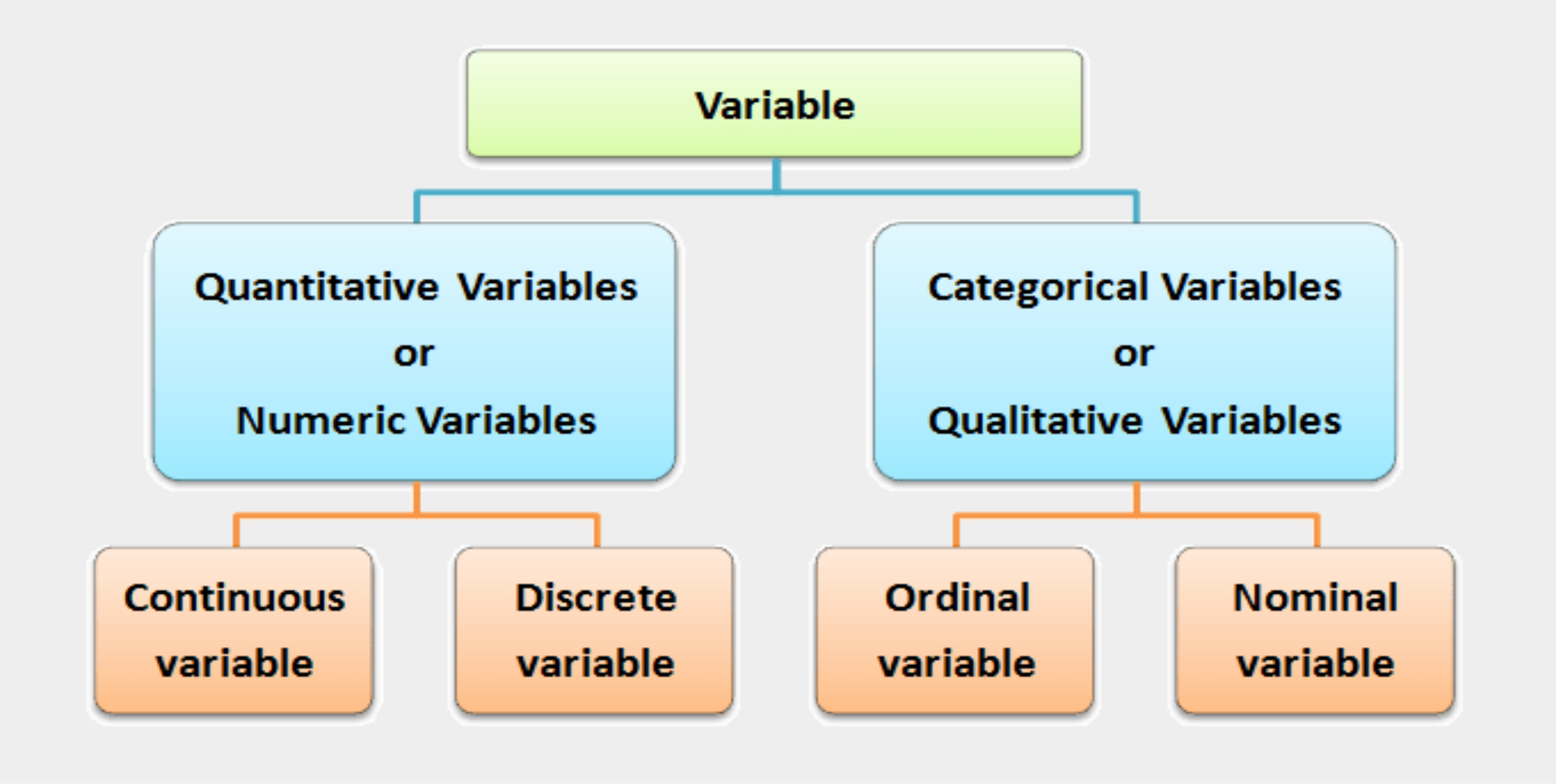
- **Discrete variable**: A variable can take only values at isolated points.
- For example,
  - ✓ the number of hospitalized patients,
  - ✓ the number of deaths attributable to resistant pathogens and
  - ✓ the number of different antimicrobials to which resistance is identified.

- **Continuous variable**: It can take any value on some interval.
- For example,
  - ✓ MIC value,
  - ✓ zone diameter

Qualitative Variable …

- **Nominal:** Categorical variables with no inherent order or ranking sequence.
- For example,
  - ✓ patients name,
  - ✓ patients ID,
  - ✓ patient gender.
- **Ordinal**: Variables with an inherent rank or order.
- For example,
  - ✓ disease severity: mild, moderate, severe;
  - ✓ AMR : resistance, susceptible and intermediate.

Variable summary



Data

- **Source of Data**
  - Primary Data: Firsthand data collected by the researcher himself
  - Secondary Data: Collected from any source.
- **Types of Data**
  - Cross-sectional data
  - Time series data (Longitudinal Data)

Study Design

- ❑ **Cross-sectional data:**
  - ✓ For example: In 2020, *Shigella* species isolates from urban Dhaka and rural Matlab were tested for resistance to all clinically relevant antibiotics in Bangladesh.
- ❑ **Time-series data:**
  - ✓ From 2000 to 2012, Shigella species isolates from urban Dhaka and rural Matlab were tested for resistance to all clinically relevant antibiotics in Bangladesh.

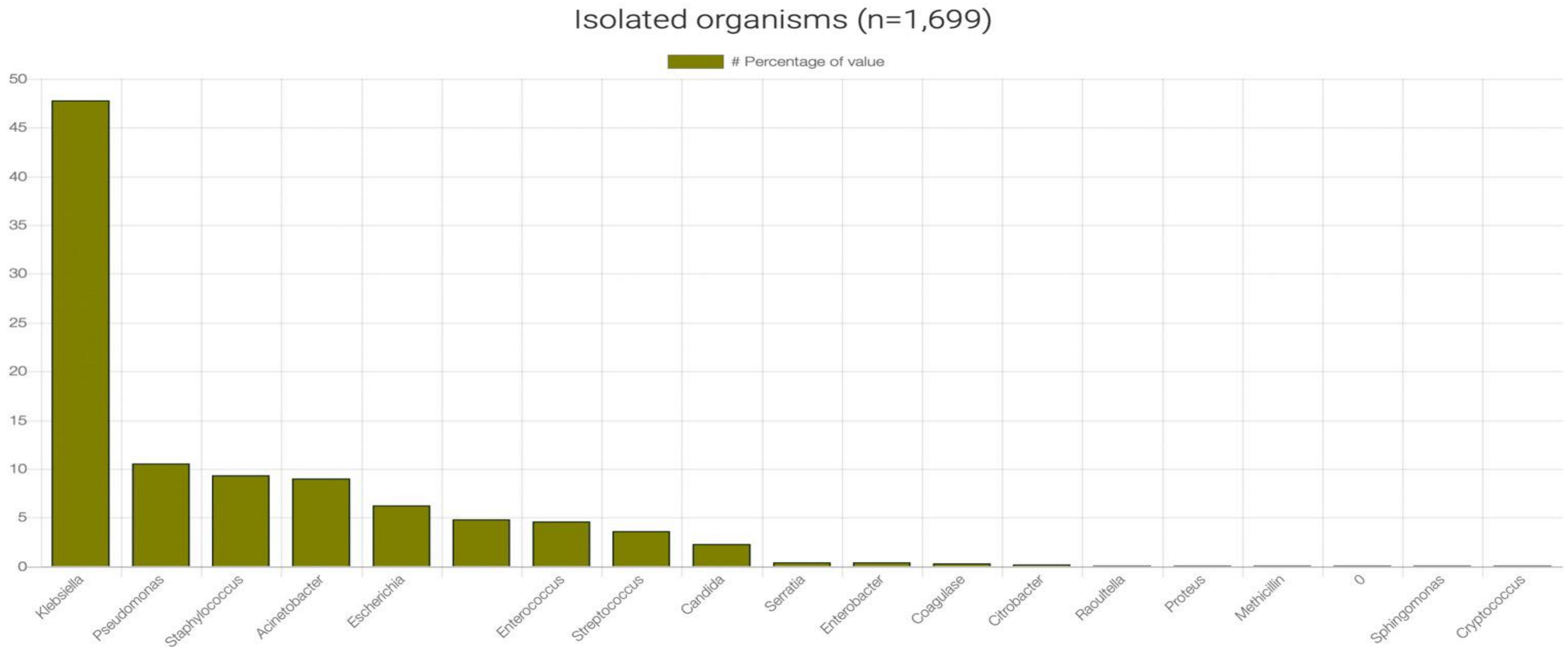
Data Presentation

- Data Presentation: Tables and Graphs
- Tables of data does not become easy and attractive to general people.
- Creating graphs of tables provide the simplest and most efficient displays.

Graphs of Qualitative data	Bar diagram
	Multiple bar diagram
	Pie Chart
Graphs of Quantitative data	Histogram
	Line graph (Time Series data)

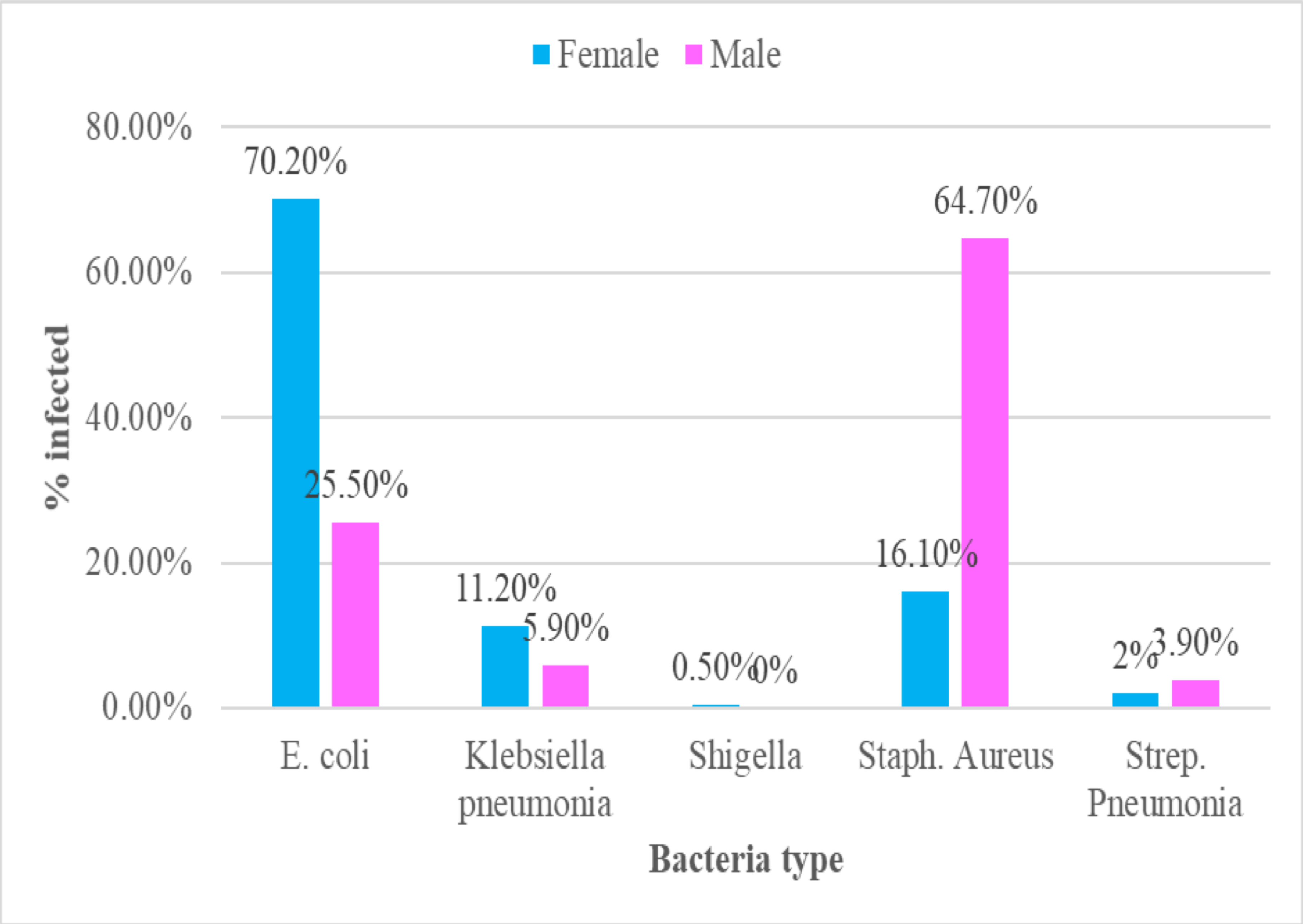
Bar Diagram

- Bar diagram is also known as Bar chart.
- Bar chart is use for categorical variable. Each bar for each category.



Multiple Bar Diagram

- Used for comparing two or more groups corresponding to a common variate value.



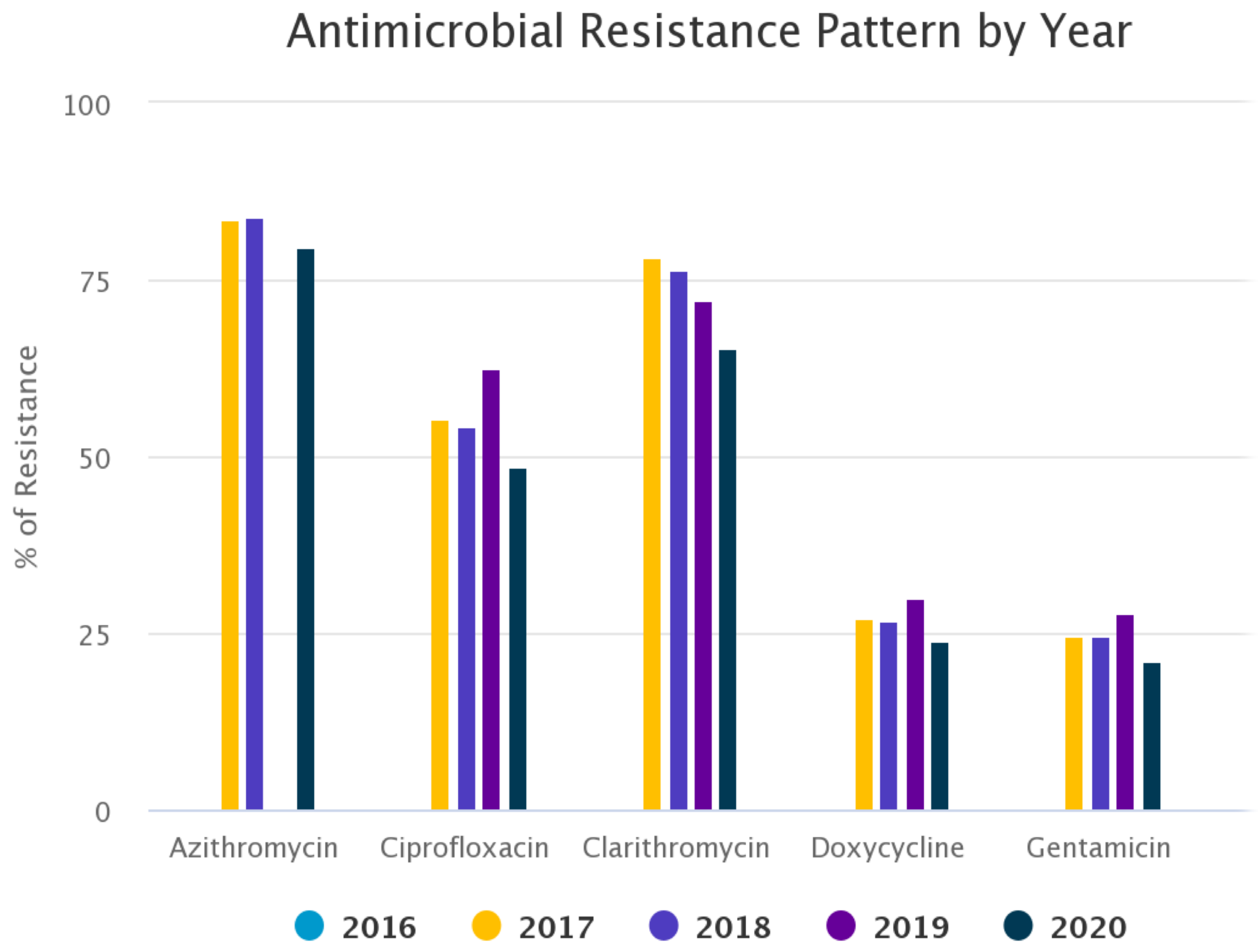


Figure: Percentage of female and male patients infected by type of bacteria

Pie Chart

- Pie chart is for the percentage distribution

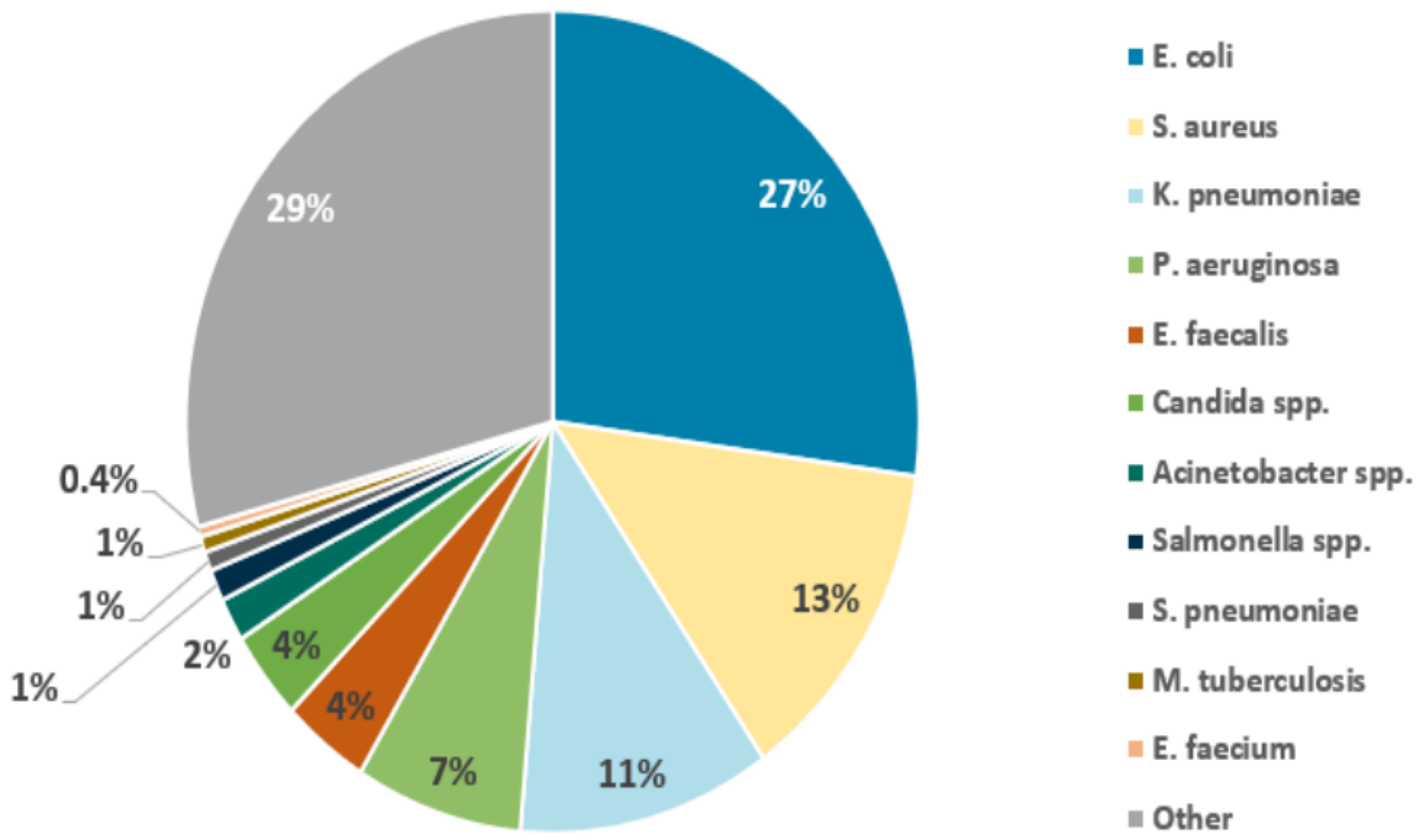


Figure: Distribution of reported AMR priority pathogens, UAE, 2020, by pathogen (n=128,128)
Source: National AMR surveillance report 2020, UAE

Line Graph

- Line graph particularly used for numerical data if we wish to show time series data

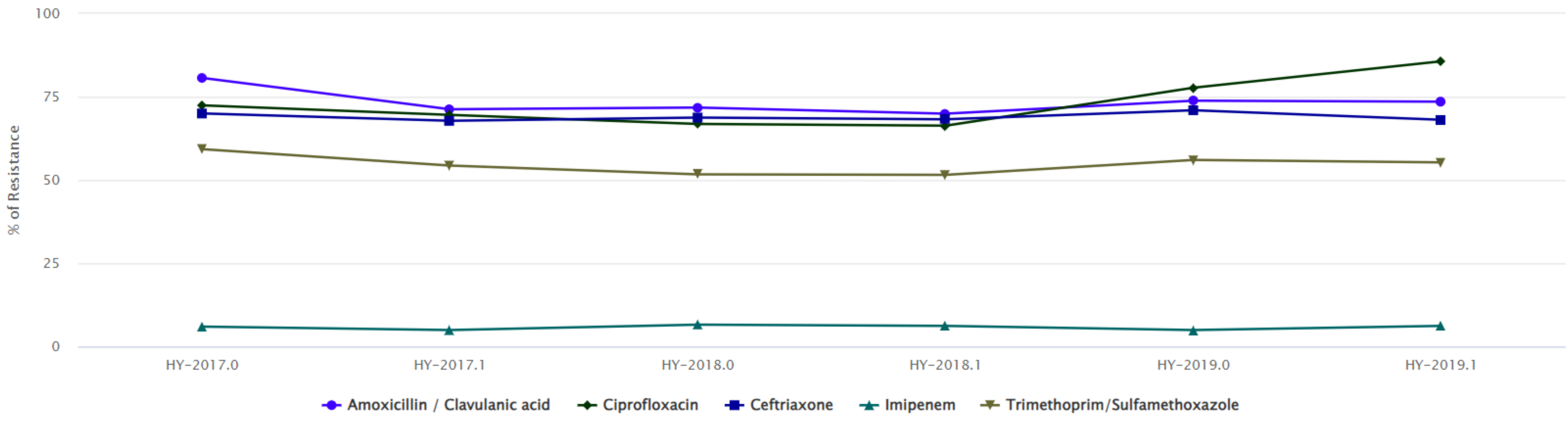


Figure: Half yearly trends of *E. coli* over the period 2017-2019

Histogram

- Histogram is the most common graphical presentation of the frequency distribution.
- A histogram is constructed by placing
  - ✓The class boundaries on the horizontal axis of a graph and
  - ✓The frequencies on a vertical axis
- Draw a histogram using the following dataset. Consider, this data is represent the age of the Dengue patients is Dhaka city

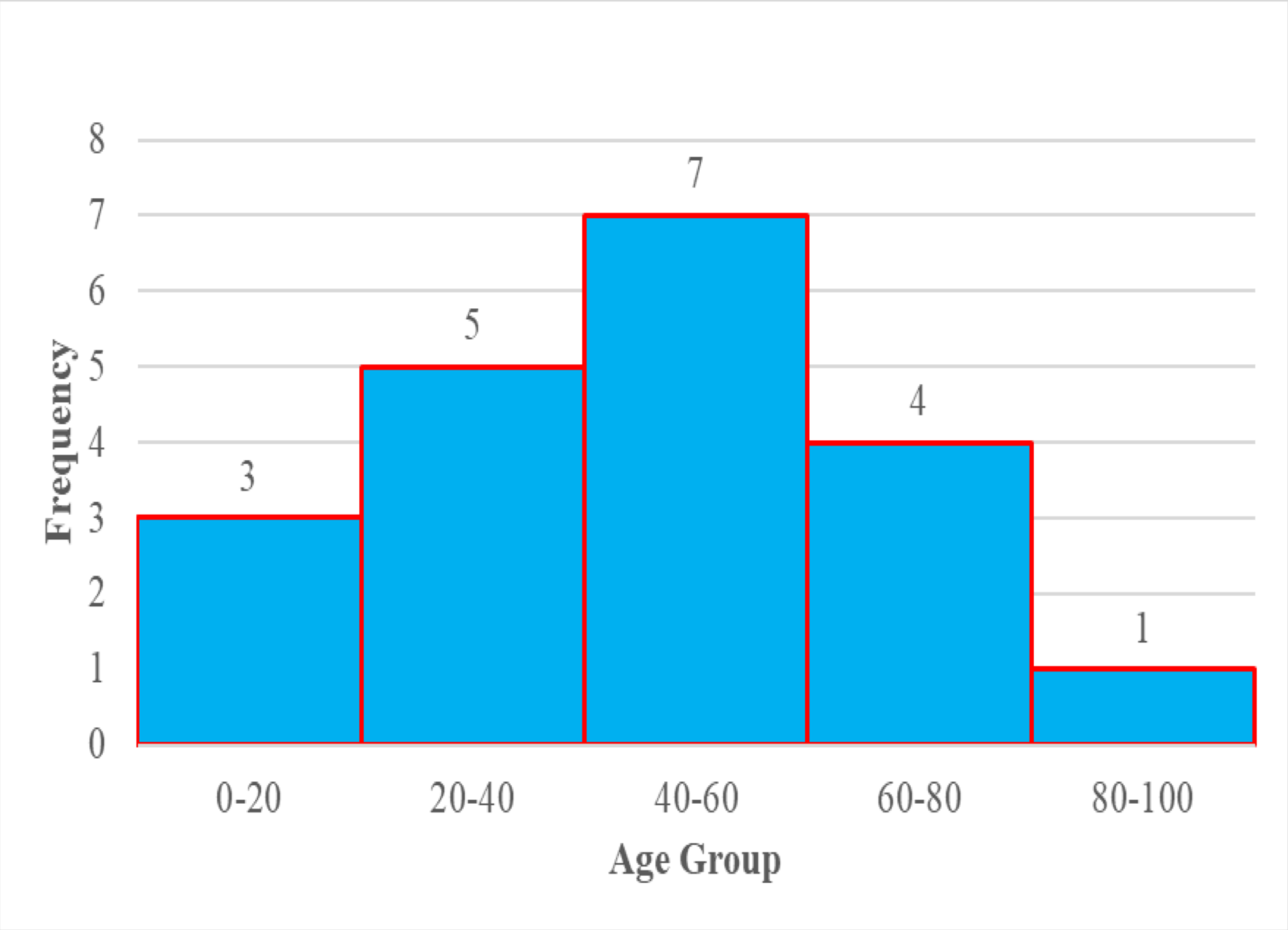
39	41	22	38	46	55	65	78	83	18
28	54	53	61	10	16	29	58	55	66

Histogram …

- Frequency Distribution

Age Group	Frequency	Observations
0-20	3	10, 16, 18
20-40	5	22, 28, 29, 38, 39
40-60	7	41, 46, 53, 54, 55, 55, 58
60-80	4	61, 65, 66, 78
80-100	1	83

- Histogram



Descriptive Statistics

- From the decision making point of view, numerical values or indices are useful for summarizing and describing data.
- There are two types of indices that are specially useful.
  - ✓ **Measures of central tendency:** Mean, Median and Mode
  - ✓ **Measures of dispersion:** Range, standard deviation.
- ▪ For scientific paper writing we usually use: Mean and standard deviation
- Summary statistics include the mean, standard deviation, median, maximum and minimum. Summary statistics can be calculated to summarize quantitative variables in a dataset.
- Graphical Representation of summary statistics: Box-and-Whisker plot.

Measures of Central Tendency: Mean

- The mean of the following data set:

16 10 12 10 9 14 13

is

(16 + 10+ 12+ 10+ 9+ 14+ 13)/7 =12.

- The general formula of the mean of a set of numbers $x_i, \mathrm{i} = 1,2, \dots, \mathrm{n}$ is

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

- Mean is affected by outlier.

Measures of Central Tendency: Median

- Middle most observation.
- Data set: 16 10 12 10 9 14 13
- Data arrange in ascending order: 9 10 10 12 13 14 16 => Median = 12
- If the number of sample is even, then take the mean of the two middle value.
- Consider the following ordered set of values: 5 6 **8 9** 12 15
- Its median is (8+9)/2 = 8.5
- Median is insensitive to extreme value (outlier)

Measures of Central Tendency: Mode

- Middle most observation.
- Data set: 16 10 12 10 9 14 13
- Data arrange in ascending order: 9 10 10 12 13 14 16 => Median = 12
- If the number of sample is even, then take the mean of the two middle value.
- Consider the following ordered set of values: 5 6 **8 9** 12 15
- Its median is (8+9)/2 = 8.5
- Median is insensitive to extreme value (outlier)

Measures of Dispersion: Range

- It is the difference between the *maximum* and the *minimum* values of the sample
- Dataset 1 (D1): 12 13 13 14 15 14 (Min = 12, Max = 15)
- Dataset 1 (D2): 5 9 12 15 20 20 (Min = 5, Max = 20)
- The range of D1 is: 15 - 12 = 3
- The range of D2 is: 20 - 5 = 15
- The measure is very vulnerable to extreme values

Measures of Dispersion: Standard Deviation

- Standard deviation (SD) is the most common measure of dispersion.
- $\text{SD} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$Standard deviation (SD) is the most common measure of dispersion.

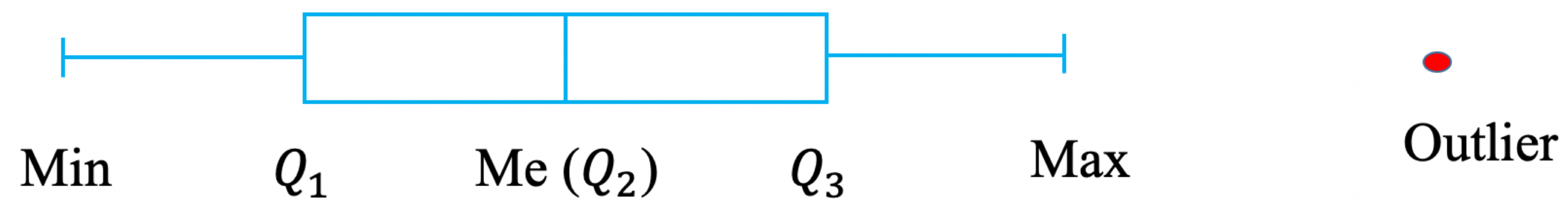
$$\text{SD} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Measures of Dispersion: Quartile

- Data can divide into four parts that cover the total range.
- The first quartile (Q_1) is the first 25% of the data.
- The second quartile (Q_2) or median is between the 25% and 50% points in the data.
- The third quartile (Q_3) is the 25% of the data lying between the median and the 75% cut point in the data.
- Interquartile Rang, IQR = $Q_3 - Q_1$

Five Number Summary and Box-and-Whisker plot

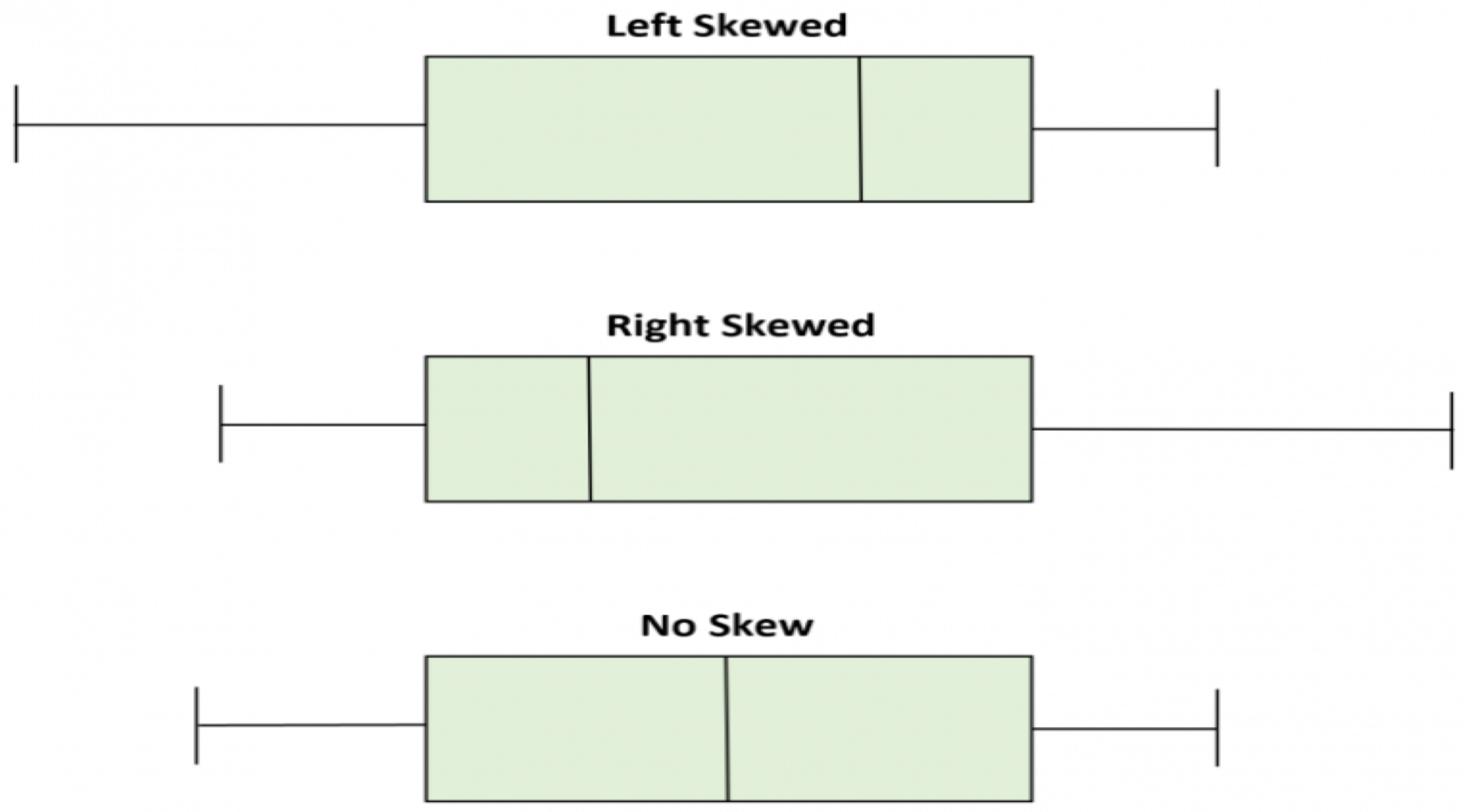
- A Box-and-Whisker plot (**box plot)** is a plot of the five number summary of a dataset, which includes:
  - ✓ The minimum value
  - ✓ The first quartile, Q_1
  - ✓ The median value, Q_2
  - ✓ The third quartile Q_3
  - ✓ The maximum value



- To see the data pattern of particular quantitative variable, such as age of patients.

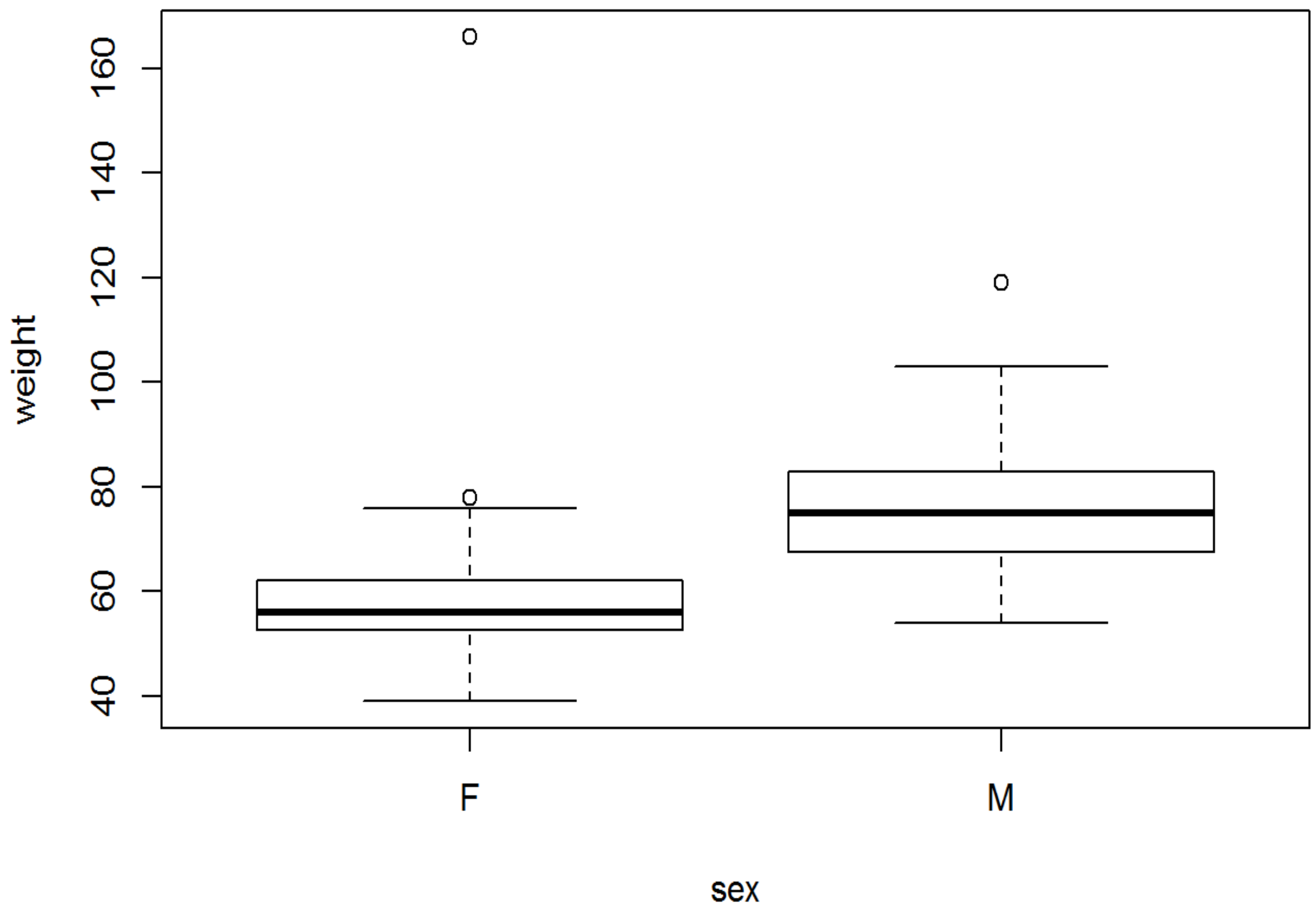
Box-and-Whisker plot

- We can determine whether or not a distribution is skewed based on the location of the median value in the box plot.



Box-and-Whisker plot …

- Boxplot is use for compare the distribution of other data sets.



▪ Boxplot for comparing the male and female weight

Correlation and Regression

- Effect of antimicrobial consumption on *E. coli* resistance.
  - ✓ Correlation between antimicrobial consumption (AMC) and antimicrobial resistance (AMR) in *E. coli* at a hospital level.
  - ✓ Predict AMR for the use of deployment of antimicrobial stewardship program (ASPs).

Correlation and Regression

- Is there any relationship between two variables?

  ✓ For example: what is the relationship between “antimicrobial consumption (AMC) (x)” and “antimicrobial resistance (AMR) (y)” in *E. coli* at a hospital level?

- What is the strength of relationship between (x) and y?

✓ Correlation

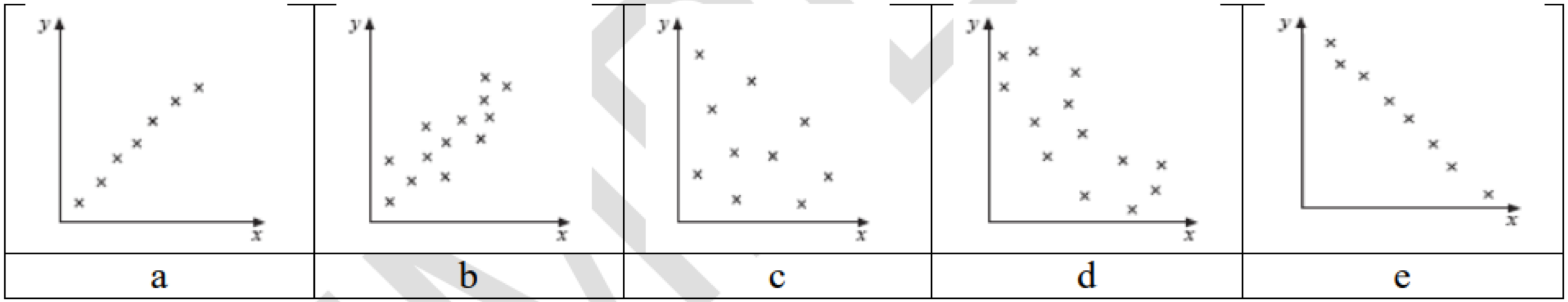
Correlation and Regression

- **Independent variable:** antimicrobial consumption (AMC) (x)
- **Dependent Variable:** antimicrobial resistance (AMR) (y)

- Can we describe this relationship and use this to predict “antimicrobial resistance (AMR) (y)” from “antimicrobial consumption (AMC) (x)?
  - ✓ Regression

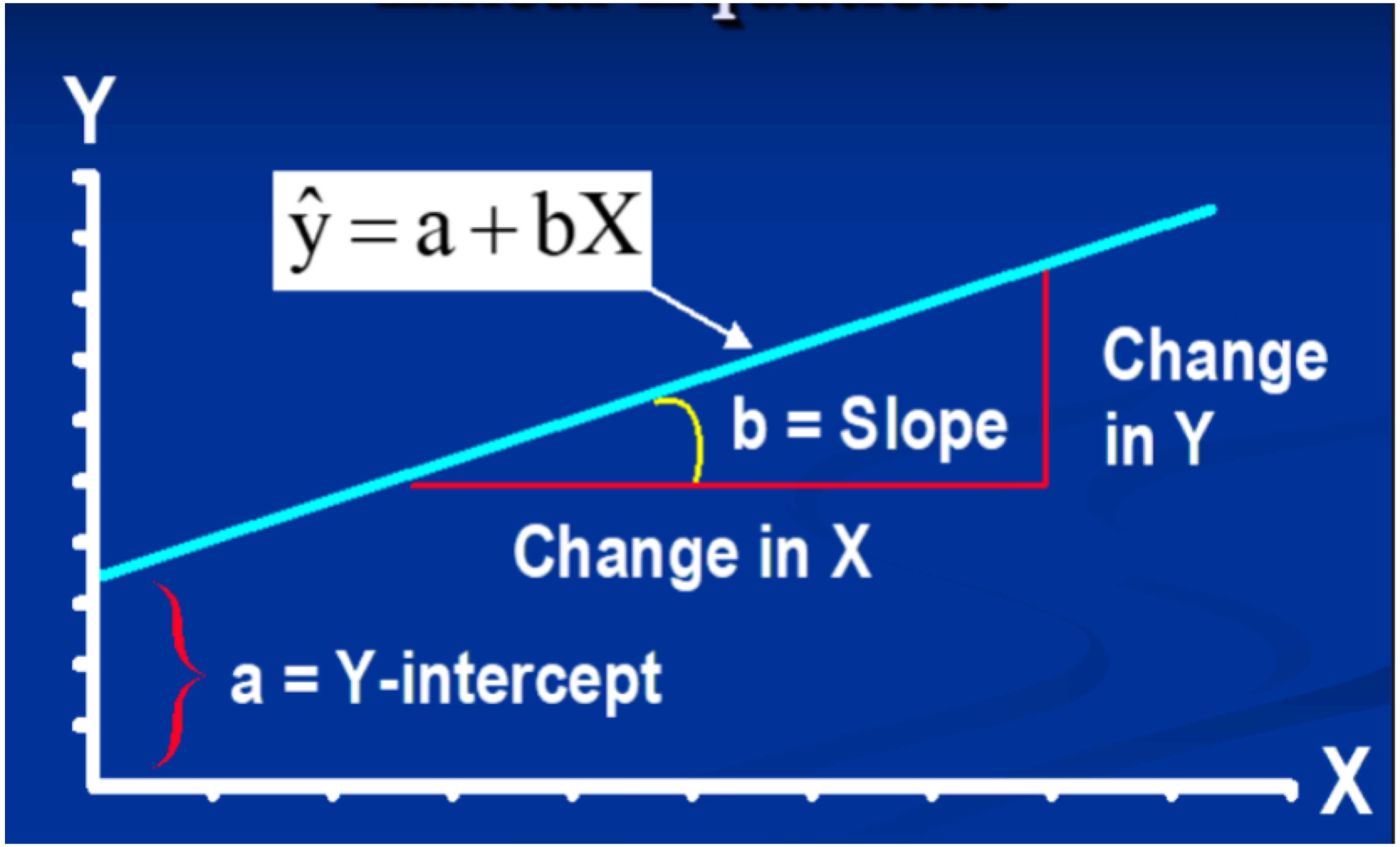
Correlation and Regression …

- In correlation we assess the strength of association between x and y. Sample correlation coefficient denoted by r.
- Correlation coefficient (r) takes values between -1 (perfect negative) to +1 (perfect positive) ($-1 \leq r \leq +1$). $r = 0$ indicates no linear association



Correlation and Regression …

- Regression tells us how values in y change as a function of changes in values of x.
- Use a variable x to predict some outcome variable y.



Regression ***

- **Linear regression**: Dependent variable (y) is continuous

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i \; = 1,2, \dots, n.$$

Dependent Variable (y_i)

Independent Variable (x_i)

  - If AMR depends on other factors such as age, gender etc. Then we use more independent variables

Regression ***

- **Logistic regression:** Dependent variable (y) is binary (two category – **Resistance and Susceptible**)

  ✓ Define $y_i = \begin{cases} 1, \text{Resistance} \\ 0, \text{Susceptible} \end{cases}$ and $\pi_i = P(y_i = 1)$, then

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i + \varepsilon_i, i \ = 1,2, \dots, n.$$

- **Multinomial regression:** Dependent variable (y) is multinomial (more than two category: AMR: **Resistance, Susceptible and Intermediate**)
- **Ordinal regression:** Dependent variable (y) has natural ordering

Inferential Statistics

- Inferential statistics divided into: Estimation and testing hypothesis.

- **Statistical Hypothesis:**
  - ✓ a Statistical Hypothesis is a statement about a population,
  - ✓ which we want to verify on the basis of sample information.

- For example: the prevalence of resistant isolates in the group which received antimicrobial treatments is the same as the prevalence in the group which did not.

Test of Hypothesis

- **Null Hypothesis** (H_0): Hypothesis that we want to tested.
- **Alternative Hypothesis** (H_1): Logical opposite (contradicts) of the null hypothesis.
- For example:
  - ✓ H_0: the prevalence of resistant isolates in the group which received antimicrobial treatments is the same as the prevalence in the group which did not
  - ✓ H_1: in this example that the prevalence of resistant isolates differs significantly between these two groups.
- In the process of accepting and rejecting a H_0, we consider two types of error: type I error and type II error.

Test of Hypothesis …

Decision	**H_0 is true**	**H_0 is false**
Reject H_0	Type I error (False Positive) $\alpha = P(\text{Type I error})$	Correct decision $1 - \beta = P(\text{Correct decision})$
Accept H_0	Correct decision $1 - \alpha = P(\text{Correct decision})$	Type II error (False Negative) $\beta = P(\text{Type II error})$

- **Level of significance** if a test: $\alpha = P(\text{Type I error}) = P(Reject\ H_0\ when\ H_0\ is\ true)$
- **Power** of test: $1 - \beta = 1 - P(\text{Type II error}) = P(rejecting\ H_0\ when\ H_0\ is\ false)$
- **Decision Rule:** H_0 is Rejected if p-value $< \alpha$ otherwise Accept H_0
- **Def of p-value:** p-value of a test is the smallest value of for which H_0 can be rejected.

Test of Hypothesis: commonly used test

1. Testing the significance of
   - single mean ($H_0: \mu = \mu_0$ vs. $H_1: \mu \neq \mu_0$)
   - single proportion ($H_0: p = p_0$ vs. $H_1: p \neq p_0$)
2. Testing the equality of
   - two mean ($H_0: \mu_M = \mu_F$ vs. $H_1: \mu_M \neq \mu_F$)
   - two proportion ($H_0: p_M = p_F$ vs. $H_1: p_M \neq p_F$)

❑Test of Significance

✓The normal test ($n \geq 30$) ✓The t-test ($n < 30$)

Test of Hypothesis: commonly used test

3. Testing the equality of several mean $(H_0: \mu_1 = \mu_2 = \cdots = \mu_p)$
   - ✓ Test of Significance: The F-test or ANOVA test
4. Testing the independence of variable (H_0: There is no association between smoking and lung cancer)
   - ✓ Test of Significance: The chi-square (χ^2) test
   - ✓ For example, *Shigella* species isolates from urban Dhaka and rural Matlab were tested for resistance to all clinically relevant antibiotics in Bangladesh.
5. Testing regression coefficient: $H_0: \beta_1 = 0$
   - ✓ Test of Significance: t-test

Confidence Interval

- Confidence intervals are often interpreted as the range of values within which we expect the population parameter to lie within a certain probability
- For example, the best estimate of the percentage of community isolates resistant to ampicillin in 2009 is 39.6%, but the 95% CI is 36.3%–43.1%.

Test of Hypothesis: commonly used test

Parametric or non-parametric?	Outcome variable	Number of groups[1]	Statistical test	Key assumptions
Parametric	Categorical: nominal with two levels (dichotomous)	Two or more	Chi-squared test	Expected frequency in any cell of a contingency table is not <5 or no more than 80% of cells have a value of <5
Non-parametric	Categorical: ordinal, or numeric when assumptions for a t-test are not met	Two groups	Mann-Whitney U test (Wilcoxon rank-sum test)	• Row and column totals are fixed • Outcome can be ranked
Non-parametric	Categorical: ordinal, or numeric when ANOVA test assumptions are not met	Three or more groups	Kruskal-Wallis test	Outcome can be ranked
Parametric	Numeric	Two groups	Student's t-test	• Normal distribution of outcome variable • Residuals have normal distribution • Variance is the same in both groups (otherwise use modified t-test)
Parametric	Numeric	Two or more groups	One-way ANOVA	• Normal distribution of outcome variable • Variance is the same in all groups
Parametric	Numeric	Two or more groups	Simple linear regression with one exposure variable	• Normal distribution of outcome variable for a given exposure value • Linear relationship (roughly) between exposure and outcome (check with scatterplot) • Homoscedasticity: the variance of residuals is the same for any value of the exposure variable
Parametric	Categorical: nominal with two levels (dichotomous)	Two groups	Binomial logistic regression	Linear relationship between the exposure and log odds

Software

- SPSS
- STATA
- R
- Python
- WHONET *** (no need any coding)

JAETECH GLOBAL

THANK YOU

If you have any questions and queries, I will be happy to answer them during the QA session.