

SCS 1307

Probability & Statistics



by
Dr Dilshani Tissera
Department of Statistics
University of Colombo

- Introduction to Statistics & Basic Concepts
- Data Types and Summary Statistics
- Introduction to Probability
- Conditional probability, Bayes' theorem, Independence
- Discrete Random variables
- Discrete Distributions
- Continuous Random variables
- Normal Distribution

Learning Outcomes

- LO1.** Calculate probabilities of events and expectations of random variables for elementary problems such as games of chance.
- LO2.** Differentiate between dependent and independent events.
- LO3.** Explain how events that are independent can be conditionally dependent (and vice-versa). Identify real- world examples of such cases.
- LO4.** Make a probabilistic inference in a real-world problem using Bayes' theorem to determine the probability of a hypothesis given evidence
- LO5.** Identify cases of Binomial Distribution, Poisson Distribution, Normal Distribution and compute probabilities related to these distributions. Apply the knowledge of these discrete & continuous distributions to solve real world problems.
- LO6:** Apply the tools of probability to solve problems

Introduction to Statistics

What is Statistics?

The science of **collecting** **organizing** **analyzing** **presenting** data,
and
finally drawing conclusions in the best possible way

Basic Terms

- **Population**
 - the complete collection of individuals or objects that are of interest to study.
- **Sample**
 - a subset of the population.
- **Variable**
 - observable, measurable or recordable characteristic.
 - basic unit of analysis.
- **Observation**
 - a value that a variable assumes for a single element

Basic Terms...

- **Data**
 - The set of observations collected for the variable
- **Parameter**
 - a numerical summary measure used to describe a characteristic of a population.
- **Statistic**
 - a numerical summary measure used to describe a characteristic of a sample.

Statistical Software

EXCEL, SPSS, MINITAB, SAS, S Plus, Matlab, R, Python, and more...

Why We Study Statistics?

- To evaluate published data
- To collect/summarize data meaningfully
- To present the findings (reports or verbally)
- To interpret the statistical results
- To employ statistical methods to make decisions
- To develop new techniques
- To generalize the decisions

Summarizing Data

Summarizing Data

- In practice we have huge amount of data.
- It is difficult to look at each observation and draw conclusions.
- Need ways to summarize the data to bring out their main features.

Types of Variables and Levels of Measurements

Variables

- Characteristics of interest, to study.
- Different statistical methods are appropriate for the different type of variables.
- There are two kinds of variables.
 - Numerical (Quantitative) Variables
 - Categorical (Qualitative) Variable

Categorical (Qualitative) Variable

- Variable having categories or classifications that are not numerical in nature.
 - Examples
 - Gender (**Male, Female**)
 - Social class of a person (**High, Medium, Low**)
 - Quality of computers (**Good, Bad**)
 - Your hair color
 - Your ethnicity
 - Did you pay income tax last tax year? (Yes/No)

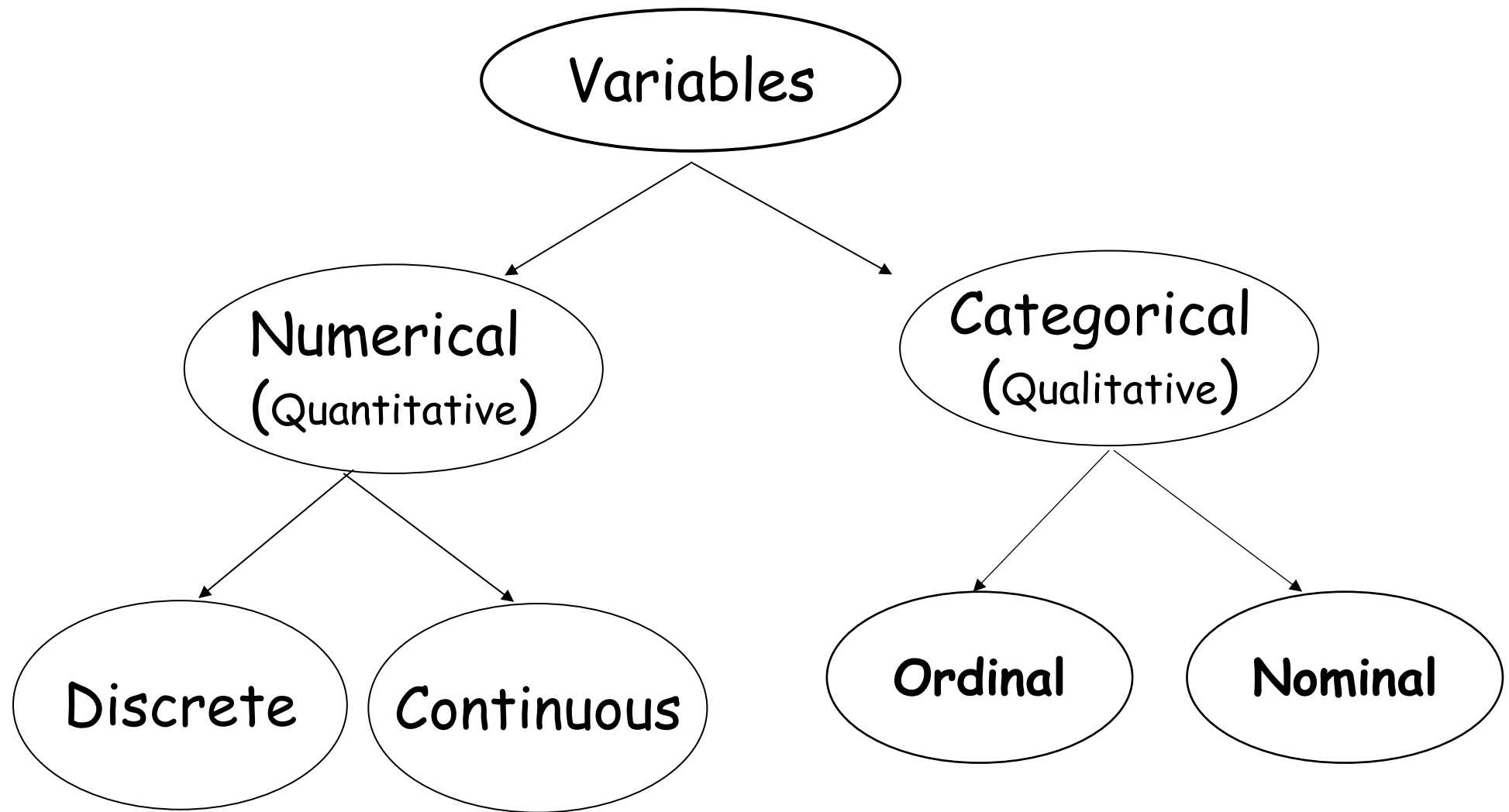
Numerical (Quantitative) Variable

- Variable whose values are numerical in nature.
 - Weight, Height, Volume,...
 - Exam marks
 - Number of defective parts in a computer
 - Age of a person
 - Blood cholesterol level a person
- Quantitative variable can be further sub divided as, **Discrete** Variable and **Continuous** Variable

Discrete and Continuous Variables...

- **Discrete variable**
 - There is a gap between two possible values
 - take only countable or finite values.
 - Number of customers arriving at a supermarket.
 - Number of defective parts in a computer.
- **Continuous variable**
 - *Theoretically*, no gap between two possible values
 - take uncountable number of values or any real values.
 - Amount of rainfall.
 - Time taken to complete a computer job.

Types of Variables



Measures Of Locations

Measures of Locations

- Minimum
 - The first observation of the ordered data set
- Maximum
 - The last observation of the ordered data set
- Measure of Center
 - A quantity locates a center of the data set.
 - The most common measures of center are,
 - 1. Mean**
 - 2. Median**
 - 3. Mode**
- Quartiles
 - A quantities that divide the data set into four equal parts.
- Percentiles
 - A quantities that divide the data set into hundred equal parts.

Mean

- The average value of the observation

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

- Very sensitive with extreme values.
-
- There are different version of the mean
 - Arithmetic mean (ordinary mean)
 - Weighted Mean
 - When we wish to give greater emphasis on some values
 - The weighted mean with weights $w_1, w_2, w_3, \dots, w_n$ is:

$$\bar{X}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

The Median (M)

- The value which divide the ordered data set into two equal parts
- Half of the observations are smaller and other half are larger than the M.
- Steps to find the median
 1. Arrange all observations in order (the smallest to the largest)
 2. If the number of observation (n) is;
 - odd, then the center value:

$$M = \frac{(n+1)^{th}}{2} \text{ value}$$

- even, then mean of the two middle values in the ordered list:

$$M = \frac{\left(\frac{n}{2}\right)^{th} \text{ value} + \left(\frac{n+1}{2}\right)^{th} \text{ value}}{2}$$

Mean Versus Median

- Mean is the informal midpoint of the data set.
 - The data set may/may not be divided into two equal parts.
- Median is the formal midpoint of the data set.
 - Exactly at the middle, data set is divided into two equal parts.
- The mean and the median are the same only if the distribution is symmetrical.
- The median is a measure of center that is resistant to skew and outliers. The mean is not.

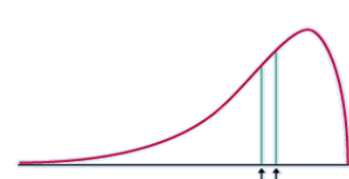
Mean and median for a symmetric distribution



Mean=Median

Symmetric

Mean and median for skewed distributions



Mean<Median

Left skewed



Median<Mean

Right skewed

The Mode

- Most frequent observation.
- There may be more than one mode

Value	20	30	40	50	60
Count	5	10	2	10	10

- There may be a data set without a mode

Value	20	30	40	50	60
Count	5	5	5	5	5

- Less frequently used than the mean or the median
- The only measure used for both categorical and numerical data

Blood Group	O+	O-	B+	B-	A+	A-
Count	5	10	12	2	11	8

Resistant Measure

- Does not influence by extreme values.
- Does not respond strongly to changes in a few observations, no matter how large these changes.
- Mean is not a resistant measure
- Median and mode are resistant measures

Quartiles

- Divide the ordered data into four equal subsets.
 - First quartile; $Q_1 = 1(n+1)/4^{\text{th}}$ value
 - Second quartile; $Q_2 = 2(n+1)/4^{\text{th}}$ value
 - Third quartile; $Q_3 = 3(n+1)/4^{\text{th}}$ value

Percentiles

- Divide the ranked data into 100 equal subsets.
 - The k -th percentile (P_k):
 - $k\%$ of the data are smaller and $(100-k)\%$ of the data are larger.

Notes:

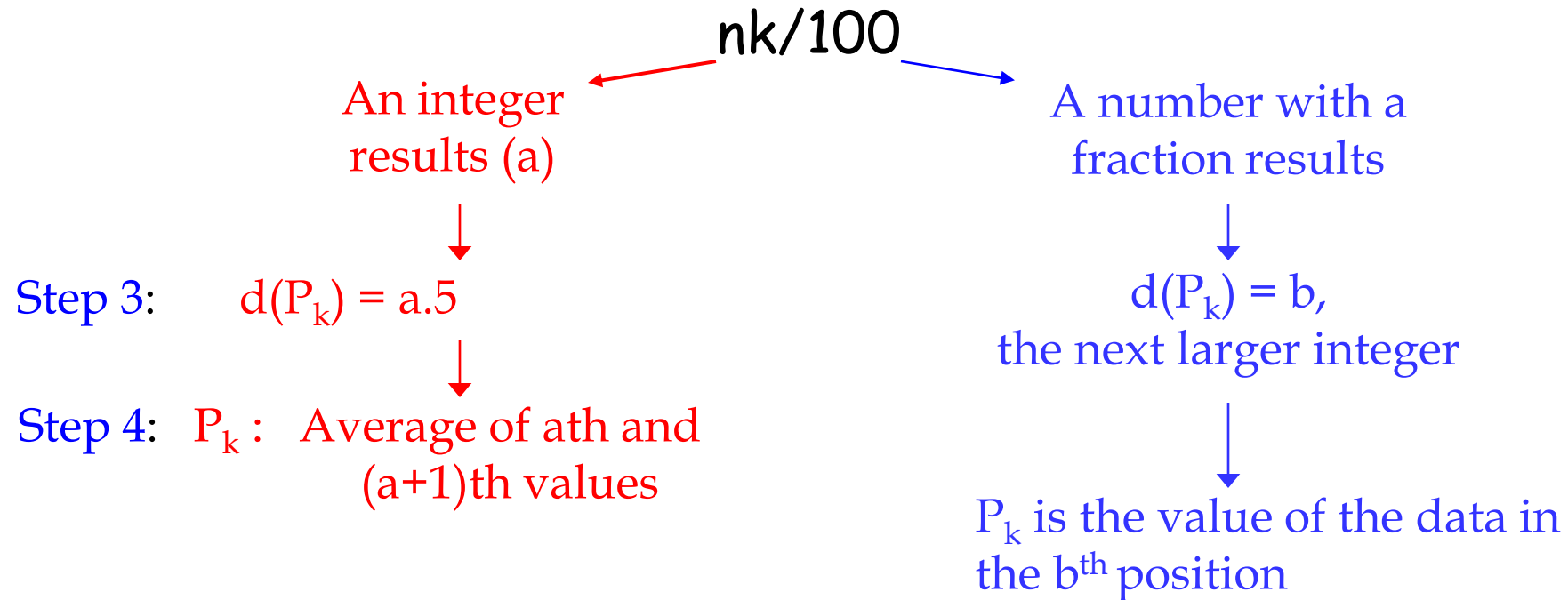
1. $Q_1 = P_{25}$,
2. $Q_3 = P_{75}$
3. Median = $P_{50} = Q_2$

Percentiles...

- A simple procedure to obtain P_k

Step 1 : Rank n data values, from the smallest to the largest

Step 2 : Calculate $nk/100$



Example

- Exam marks of 50 students are listed below. Find:
 - The first quartile,
 - The 58th percentile
 - The third quartile.

60	47	82	95	88	72	67	66	68	98
90	77	86	58	64	95	74	72	88	74
77	39	90	63	68	97	70	64	70	70
58	78	89	44	55	85	82	83	72	77
72	86	50	94	92	80	91	75	76	78

Example...

Find Q_1 first.

Step 1 : Rank the data

39 44 47 50 55 58 58 60 63 64 64 66 67 68 68 70 70 70 72 72 72 72 74 74 75 76
77 77 77 78 78 80 82 82 83 85 86 86 88 88 89 90 90 91 92 94 95 95 97 98

Step 2: Find $nk/100 = 50*25/100 = 12.5$

Step 3: $d(Q_1) = 13$

Step 4 : Q_1 is the 13th value from the lowest value. $Q_1 = 67$

Verify that $P_{58} = 77.5$ and $Q_3 = 86$.

Measures Of Dispersion

Measure of Spread (Dispersion/ Variability)

- Measures how far each value is away from measure of center
- Without measures of dispersed, measure of center may be misleading.
- For example

	median	spread of the data
2, 5, 6, 8, 9 :	6	$9-2=7$
5, 5, 6, 6, 6 :	6	$6-5=1$
6, 6, 6, 6, 6 :	6	$6-6=0$

- Measure the **variability** should be
 - a real number
 - zero if all the data are identical
 - increases as the data becomes more diverse
 - cannot be less than zero

Range (R) and Inter Quartile Range (IQR)

- Range : **R = largest value - smallest value**
 - The simplest measure of dispersion
 - Gives an indication of the absolute spread of the distribution
 - Depends only on the two extreme values
 - Not very informative and affected by outliers
- Inter Quartile Range (IQR): **$IQR = Q_3 - Q_1$**
 - **Range of the middle 50%** of the values
 - Upper 25% and lower 25% observations are discarded
 - Not affected by outliers
 - Not centered on the middle unless the distribution is symmetric
 - Often used with median
- Inter-Quartile Range (IQR)
 - range of the middle 50% of the values

$$IQR = Q_3 - Q_1$$

Variance

- Numerically, the variance equals the average of the several squared deviations from the mean.
- Variance for the sample of size n is given by

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

- A data set which is highly variable will have a larger variance compared to a data set which is relatively homogeneous.

Standard Deviation

- **Positive value of the square root of the variance.**
- If each number is increased by a constant c
 - The mean is increased by c
 - The standard deviation remains unaltered
- If each number is multiply by a constant k
 - The mean is multiplied by k
 - The standard deviation is multiplied by k

Skewness

- One measure of skewness is the Pearson's coefficient of skewness

$$\begin{aligned} \text{Pearson's Coefficient of Skewness} &= \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}} \\ &\approx \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}} \end{aligned}$$

- Generally skewness can take any value between +3 and -3

