Econometrics

	Course Calendar
Week	Main Content
Week 7	Extension of Simple Regression: Functional Forms I
Week 8	Extension of Simple Regression: Functional Forms II
Week 9	Extension of Simple Regression: Functional Forms III
Week 10	Multiple Regression
Week 11	Multiple Regression: Problem of Inference
Week 12	Multiple Regression: Functional Forms
Week 13	Introduction to Dummy Variables
Week 14	Introduction to Dummy Variables and Regression Methods
Week 15	Regression with Dummy Variables: Hands-on-Exercise
Week 16	Application of Regression

Econometrics

Lecture 14. Introduction to Dummy Variables & DV in Regression

Geetha Rani Prakasam, Ph.D

Professor,

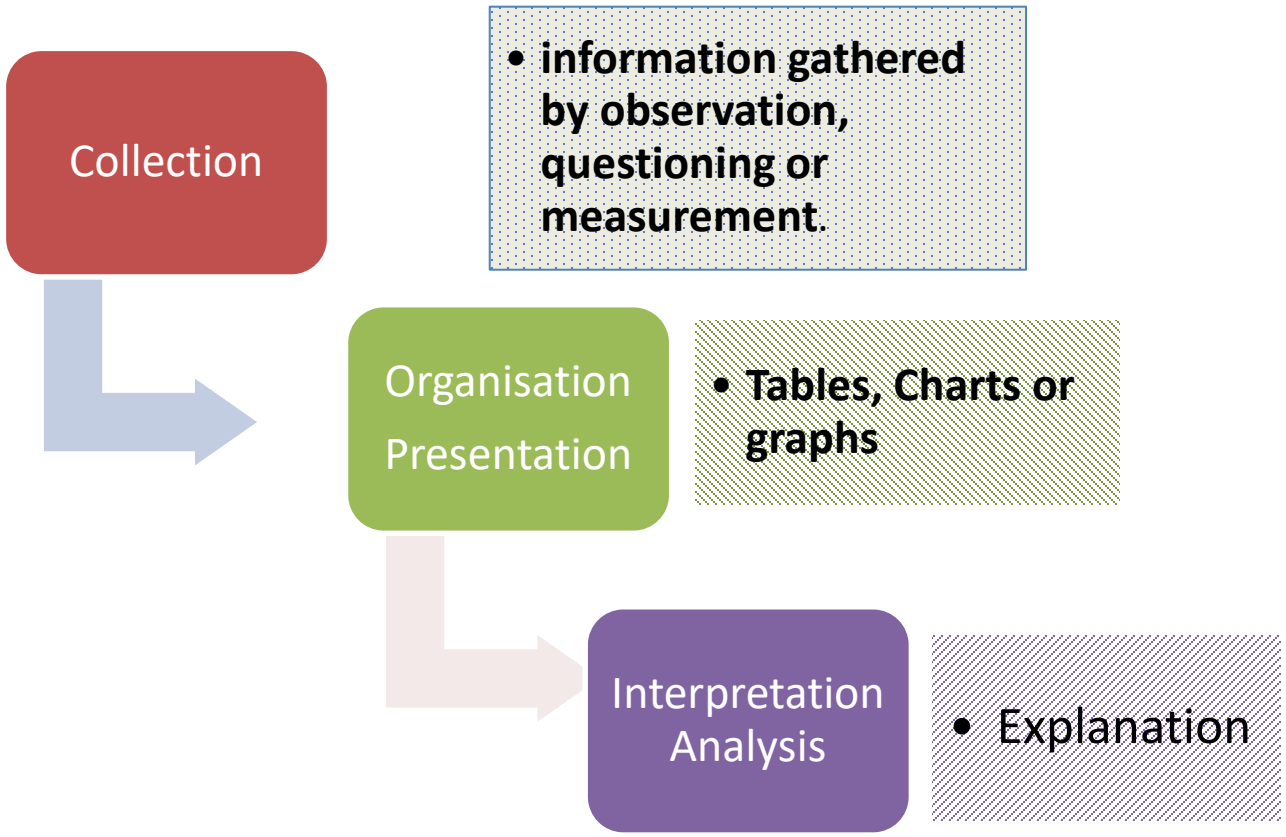
Recap

- Different functional forms in Multiple Regression
- Introduction to Dummy Variables

Outline

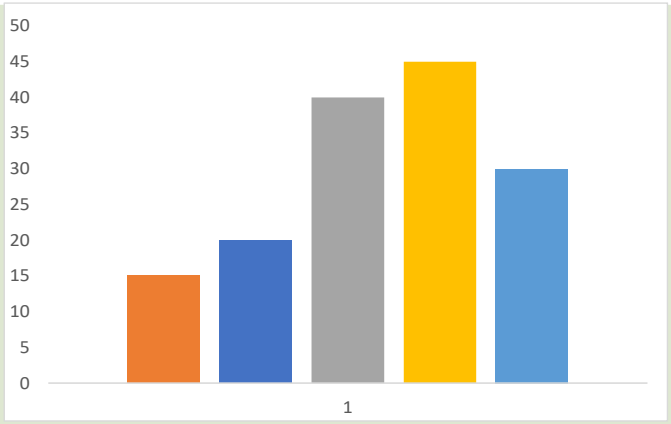
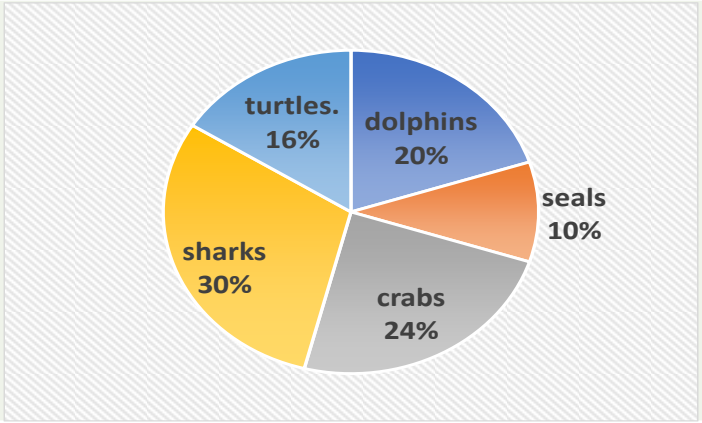
- What is data and data analysis?
- Type of Variables
- Dummy Variables
- Caution in the Use of Dummy Variables
- Dummy X variables in Multiple Regression: ANOVA Model

What is data and data analysis?



- Data is a collection of information gathered by observation, questioning or measurement.
- It include facts, numbers or measurements, tables, Charts or graphs
- Interpretation of the Tables, Charts, graphs, etc

What is data and data analysis?

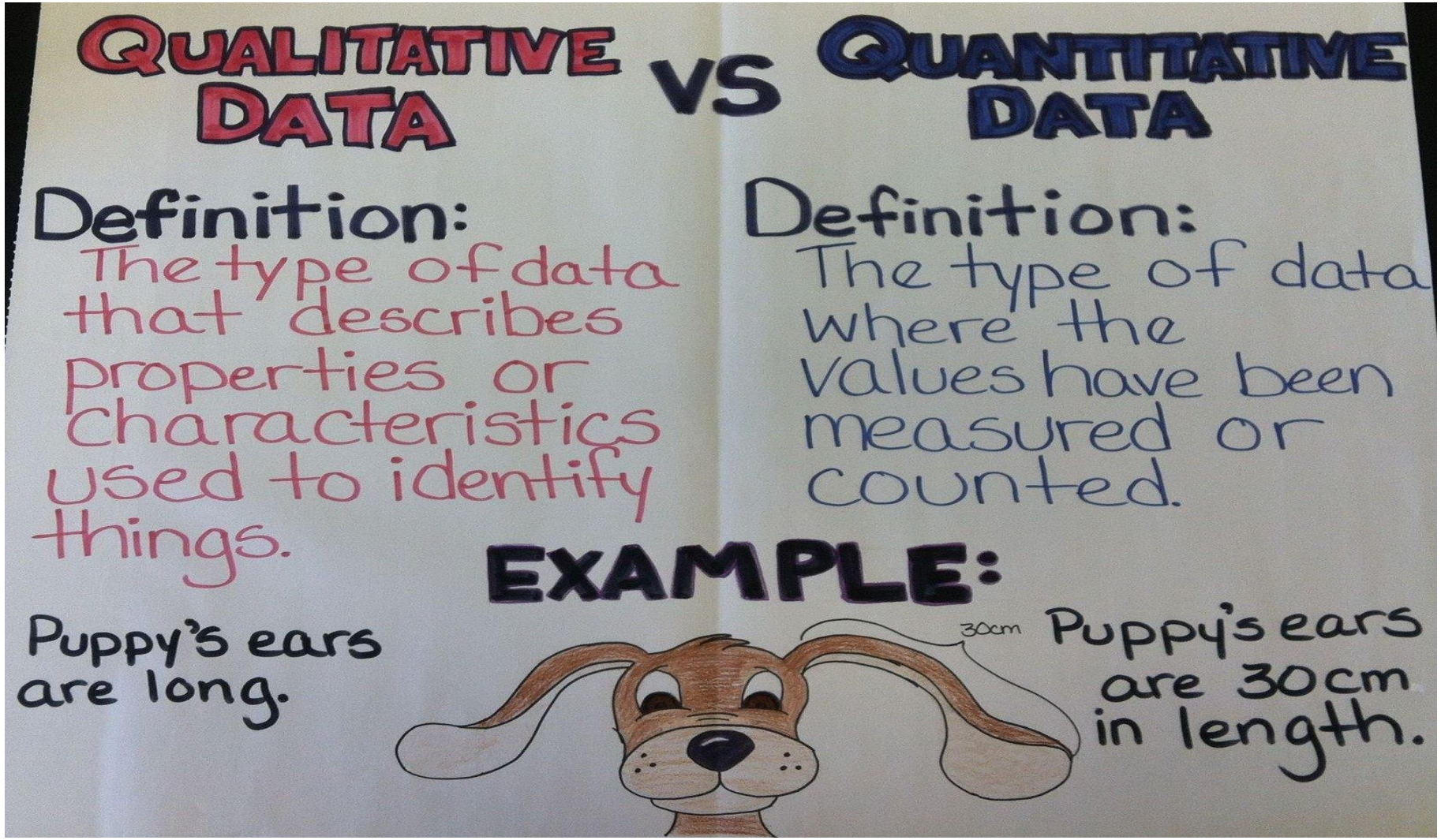
Collection	Organization	Interpretation
Annual Whale sighting reports for five years with 15, 20, 40, 45, and 30 sighting a year		There has been an average of 30 whale sighting a year for the last five years.
On an aquarium visit the class saw 10 dolphins, 5 seals, 12 crabs, 15 sharks and 8 turtles.		Sharks were seen the most and Seals were seen the least.

Data Analysis procedure and steps

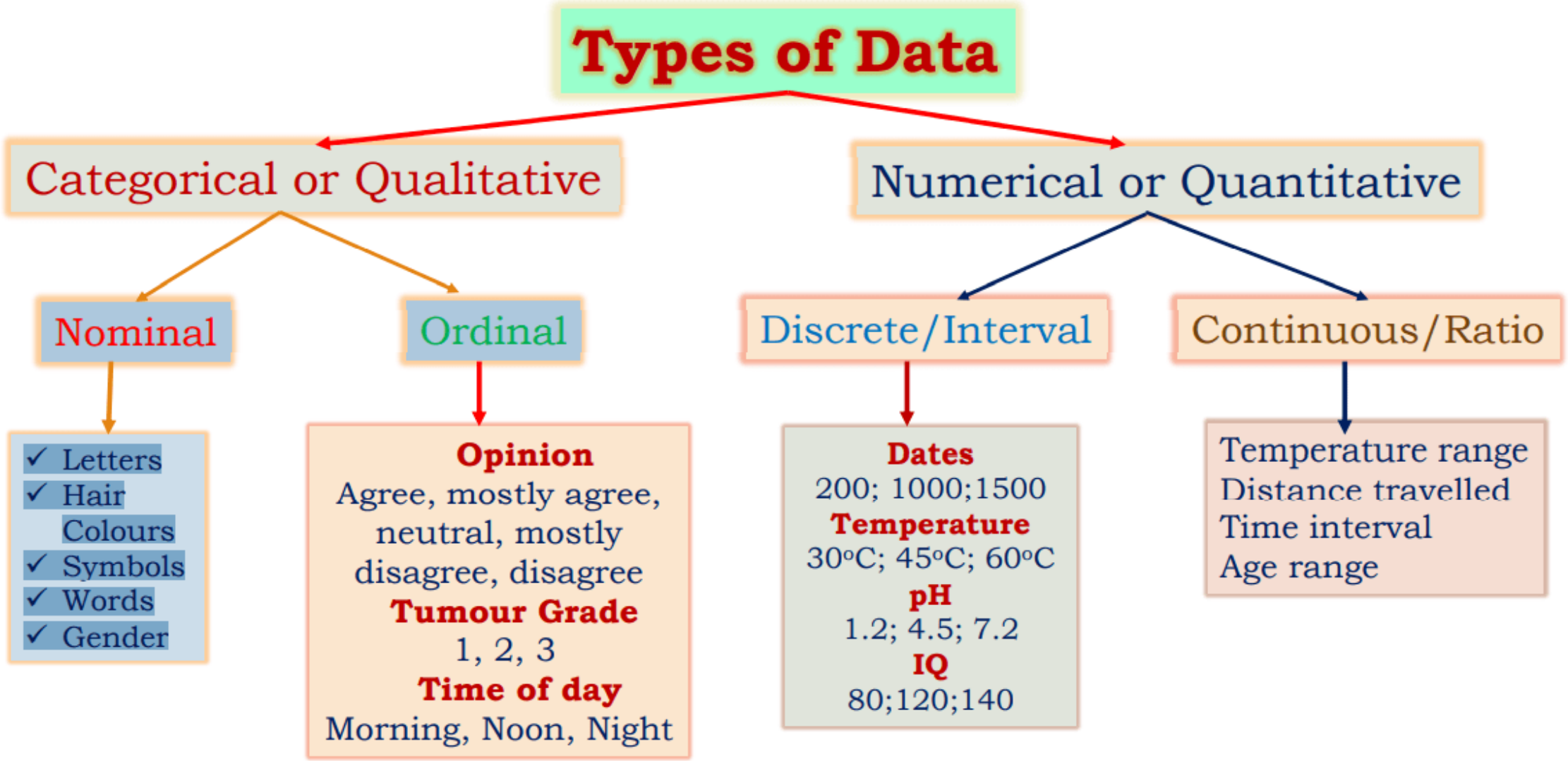
- **The initial step is to decide the information prerequisites or how the information is gathered.** Information might be isolated by age, gender, or income. Data values might be mathematical or be isolated by class.
- **The second step in data analytics is the way toward gathering it.** This should be possible through an assortment of sources, for example, computers, online sources, cameras, or through the workforce.
- **When the information is gathered, it should be coordinated so it tends to be examined.** Association may occur on an accounting page or other type of programming that can take statistical data.
- **The data is then cleaned up before the examination.** This implies it is scoured and checked to guarantee there is no duplication, and that it isn't deficient. This progression remedies any mistakes before it goes on to an information expert to be dissected.

Types of Data

- There are only two classes of data in statistics, that is **Qualitative and Quantitative data.**
- But, after that, there is a subdivision and it breaks into **4 types of data**.
- Data types are like a guide for doing the whole study of statistics correctly ☺!



https://quizizz.com/media/resource/gs/quizizz-media/quizzes/27d415e6-766d-4e53-93da-3cda83bfd395



https://pharmacygyan.com/wp-content/uploads/2021/10/types-of-data-min.png

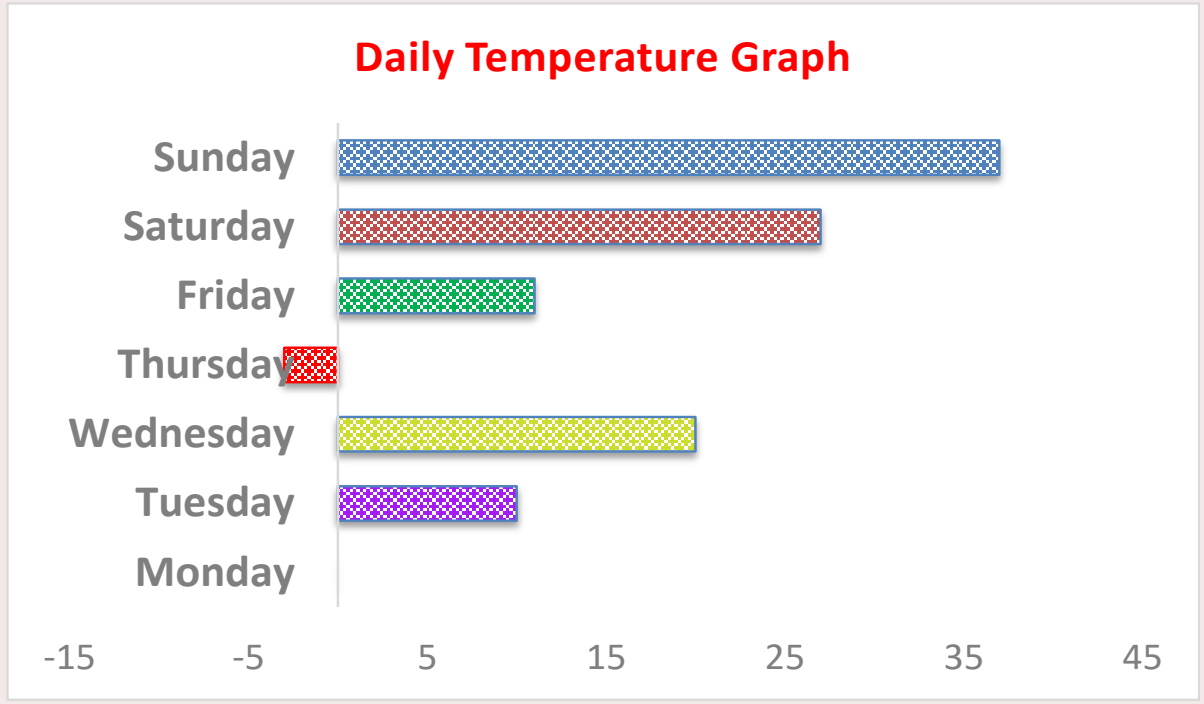
Quantitative Data

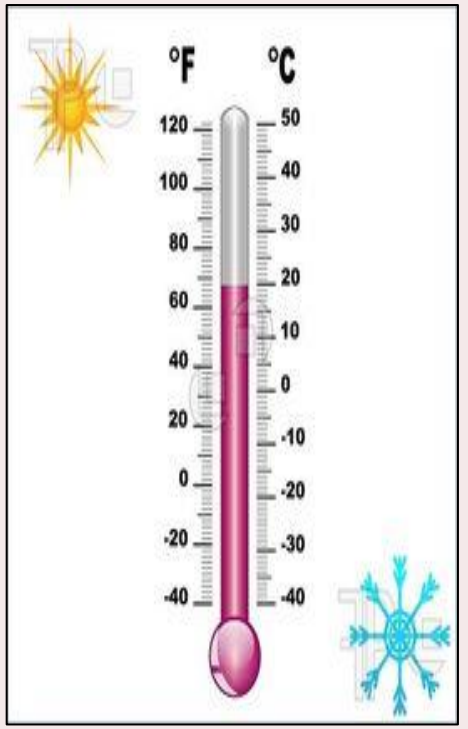
- **Continuous variable**
- Continuous data is quantitative data that can be measured.
  it has an infinite number of possible values within a selected range; e.g. temperature range, income, expenditure, age, height, weight, hours of work, distance to workplace in kms, etc.
- **Discrete variable**
- discrete data is quantitative data that can be counted and has a finite number of possible values; e.g. days of the week, chairs in a classroom, teachers in a university, number of books across subjects in a library, number of computers, number of mobile phones, number of pizzas, number of burgers, etc.

Continuous Data

quantitative data that can be measured

Discrete Data quantitative





Continuous Data, quantitative (in Celsius degree)

In this graph, the days of the week are discrete data but the temperature is continuous data.

Continuous data – infinite values

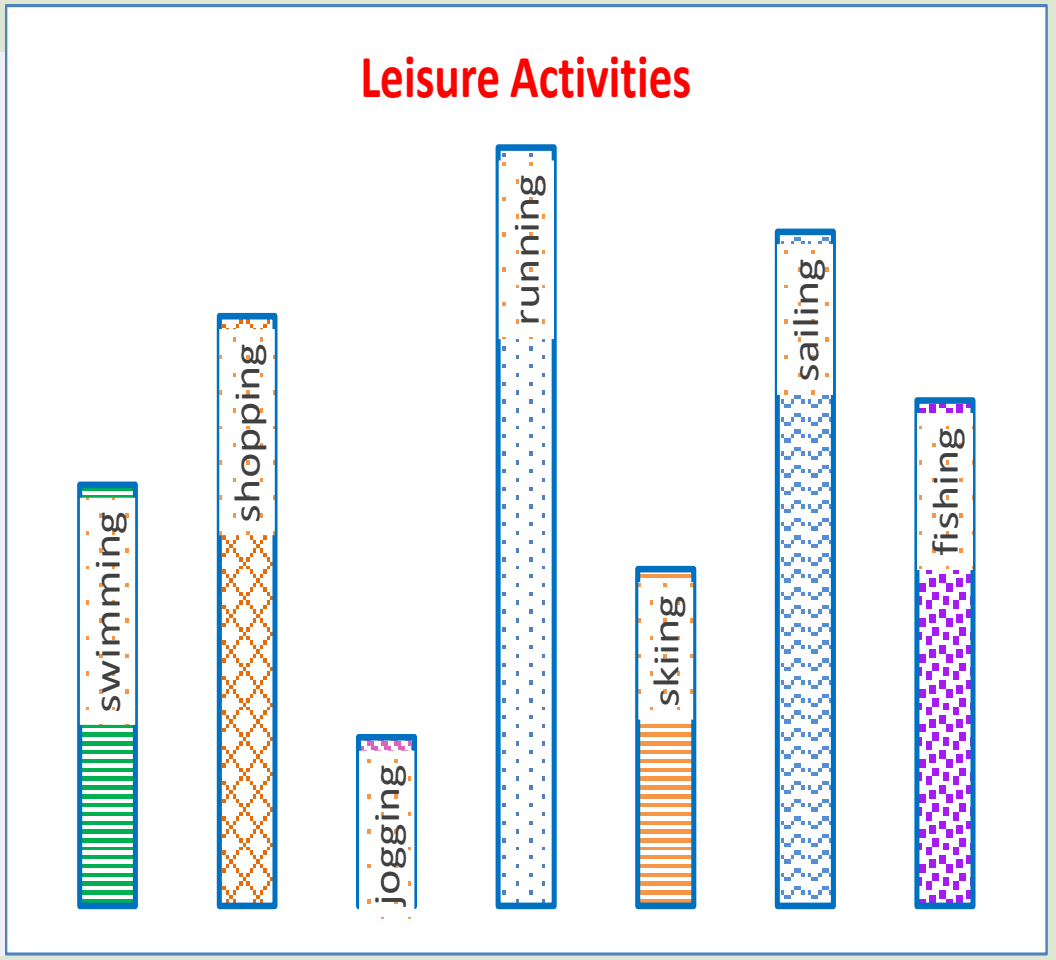
Discrete data – finite values

https://d1whtlypfis84e.cloudfront.net/guides/wp-content/uploads/2018/02/13180950/5762835.jpg

Quantitative Data Analysis

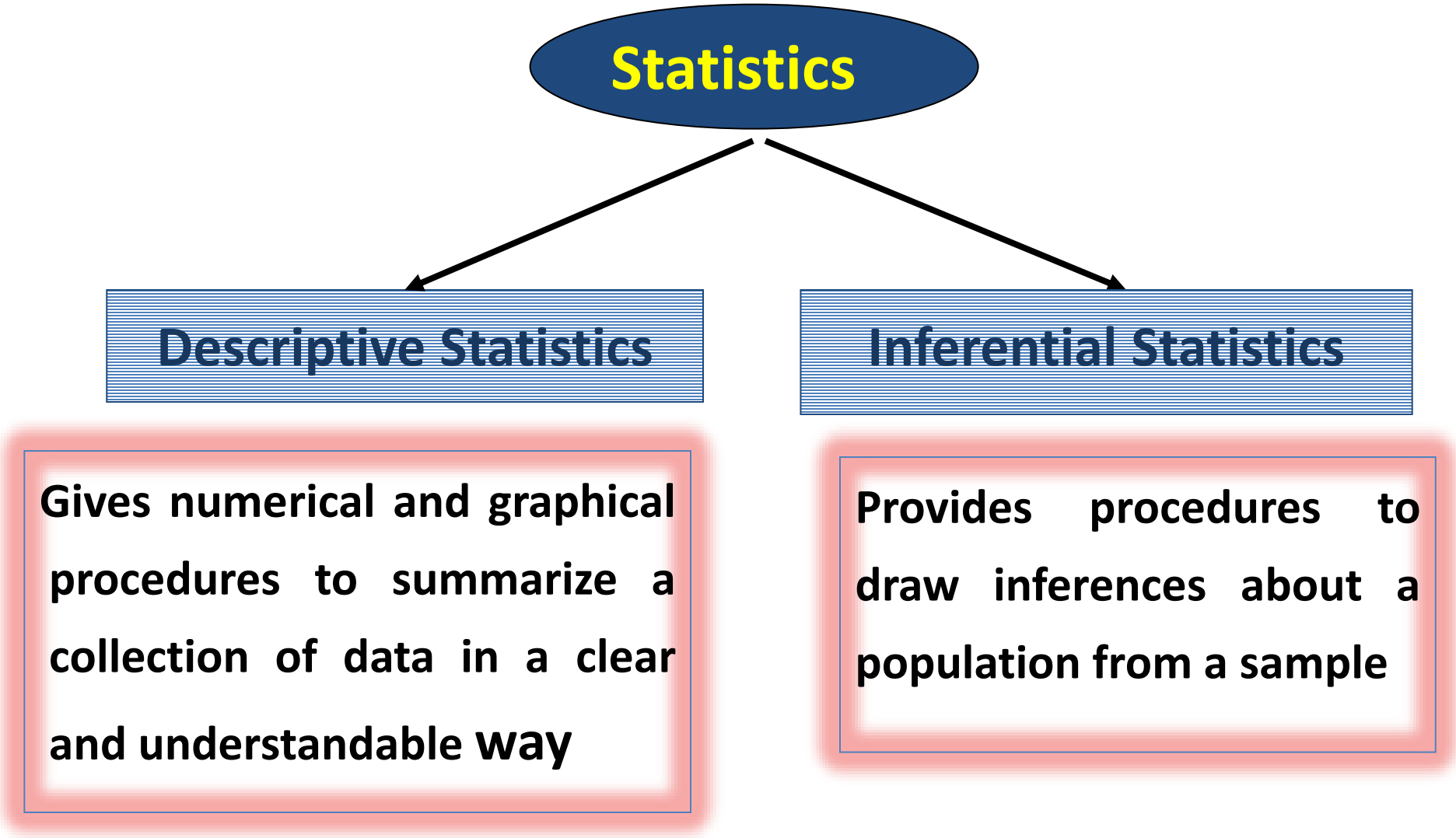
- Identify the data is a quantitative or qualitative data
- If quantitative data: we can carry out the descriptive statistics
  - Tabulation of Quantitative Data
  - Analysis of Quantitative Data
  - Graphical Representation
  - Measures of Central Tendency: Mean, Median and Mode
  - Measures of Variability: Range, Variance, SD
  - Measures of Relationship: Correlation and Regression

Qualitative Data: Also known as categorical data

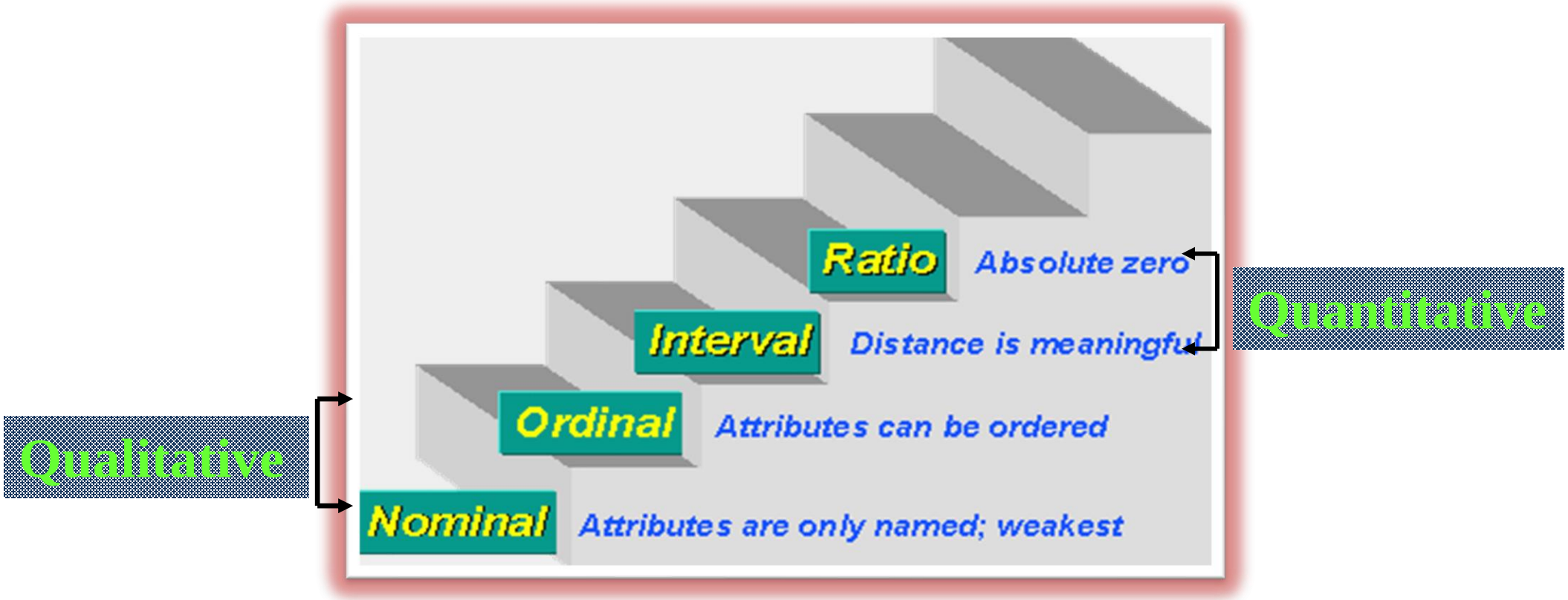


https://i.ytimg.com/vi/RI50I0GV3gE/maxresdefault.jpg

In this graph, the leisure activities preferences are shown


Statistics
Descriptive Statistics
Inferential Statistics
Gives numerical and graphical procedures to summarize a collection of data in a clear and understandable way
Provides procedures to draw inferences about a population from a sample

Levels of Measurements



https://conjointly.com/img/kb/Assets/images/measlev2.gif

Nominal Variables

- Nominal variables allow for only qualitative classification.
- They can be measured only in terms of whether the individual items belong to certain distinct categories, but we cannot quantify or even rank /order the categories.
- Nominal data has no order, and the assignment of numbers to categories is purely arbitrary.
- Because of lack of order or equal intervals, one cannot perform arithmetic (+, -, /, *) or logical operations (>, <, =) on the nominal data.

Examples of Nominal Variables

- *Gender*
  - Male =1
  - Female =2
- *Marital Status*
  - Unmarried = 1
  - Married = 2
  - Divorcee = 3
  - Widower = 4

Ordinal Variables

- An ordinal variable is similar to a nominal or categorical variable.
- The difference between the two is that there is a clear ordering of the categories.
- For example, suppose we have a variable, economic status, with three categories:- low, medium and high.
- In addition to being able to classify people into these three categories, we can rank or order the categories as low, medium and high.

Examples: Ordinal Variables

- 1 = Very low or nil
  2 = Low
  3 = Medium
  4 = Great
  5 = Very great
- Here, the variable 'Time Involvement' is an ordinal variable with 5 states.
- Ordinal variables often cause confusion in data analysis
- Some treat them as nominal variables.
- Others treat them as interval scale variables, assuming that the underlying scale is continuous, but because of the lack of a sophisticated instrument, they could not be measured on an interval scale.

Interval Data

- Interval Data are measured and ordered with the nearest items but have no meaningful zero.
- The central point of an Interval scale is that the word 'Interval' signifies 'space in between', which is the significant thing to recall, interval scales not only educate us about the order but additionally about the value between every item.
- *Interval data can be negative, though ratio data can't.*
- Even though interval data can show up fundamentally the same as ratio data, the thing that matters is in their characterized zero-points.

Interval Data

- If the zero-point of the scale has been picked subjectively, at that point the data can't be ratio data and should be interval data.
- Hence, with interval data we can easily correlate the degrees of the data and also we can add or subtract the values.
- There are some descriptive statistics that we can calculate for interval data are central point (mean, median, mode), range (minimum, maximum), and spread (percentiles, interquartile range, and standard deviation).

Examples of Interval data

- Temperature (°C or F, but not Kelvin)
- Dates (1066, 1492, 1776, etc.)
- Time interval on a 12-hour clock (6 am, 6 pm)

Ratio Data

- Ratio Data are measured and ordered with equidistant items and a meaningful zero and never be negative like interval data.
- One good example of ratio data is the measurement of heights. It could be measured in centimetres, inches, meters, or feet and it is not practicable to have a negative height.
- Ratio data enlightens us regarding the order for variables, the contrasts among them, and they have absolutely zero. It permits a wide range of estimations and surmisings to be performed and drawn.

Ratio Data

- **Ratio data is fundamentally the same as interval data, aside from zero means none.**
- The difference between interval and ratio data is that **Ratio data has a defined zero point**.
- Descriptive statistics which we can calculate for ratio data are the same as interval data which are central point (mean, median, mode), range (minimum, maximum), and spread (percentiles, interquartile range, and standard deviation).

Example of Ratio data:

- Age (from 0 years to 100+)
- Temperature (in Kelvin, but not °C or F)
- Distance (measured with a ruler or any other assessing device)
- Time interval (measured with a stop-watch or similar)
- Income, height, weight, annual sales, market share, product defect rates, time to repurchase, unemployment rate, and crime rate are examples of ratio data.

To sum up:

- Nominal data and ordinal data are the types of qualitative data or categorical data.
- Interval data and ratio data are the types of quantitative data which are also known as numerical data.
- Nominal Data are not measured but observed and they are unordered, non-equidistant, and also have no meaningful zero.
- Ordinal Data is also not measured but observed and they are ordered however non-equidistant and have no meaningful zero.
- Interval Data are measured and ordered with equidistant items yet have no meaningful zero.
- Ratio Data are also measured and ordered with equidistant items and a meaningful zero.

- Dummy Variables

What is a dummy variable?

- A dummy variable denotes whether something is true, which is 1, or false, which is 0.
- Dummy variables are also called indicator variables.
- Nominal
- Binary
- Categorical
- Ordinal

Constructing a dummy variable..

A dummy variable takes the value of one (1) for some observations to indicate the presence of an effect/group and zero (0) for the remaining observations.

Ex:

(1) Mgt = 1 if the student goes to private school;

= 0 if the student goes to govt. School.

(2) gender = 1 if the respondent is a male;

= 0 if the respondent is a female.

(3) Participation =1 if the child goes to school;

=0 if not going to school

Dummy X variables

- Y is influenced not only by some X variables that are readily quantifiable but also by some other X variables that are not readily quantifiable.
- Other names of DV - Indicator Variables, Binary Variables, Categorical Variables, Dichotomous Variables, Qualitative Variables
- For instance: school performance affected by management type or ownership of schools
- DVs can be incorporated in regression models just as easily as quantitative variables.
- A regression model may contain regressors that are all exclusively dummy, or qualitative, in nature. Such models are called **Analysis of Variance (ANOVA) models.**

EX. 9.1: Public School Teachers' Salaries by Geographical Region

- Table 9.1 gives data on average salary (in dollars) of public school teachers in 50 states and the District of Columbia for the academic year 2005–2006.
- These 51 areas are classified into three geographical regions: (1) Northeast and North Central (21 states in all), (2) South (17 states in all), and (3) West (13 states in all).
- Suppose we want to find out if the average annual salary of public school teachers differs among the three geographical regions of the country.
- If we take the simple arithmetic average of the average salaries of the teachers in the three regions, we will find that these averages for the three regions are:

Table 9.1 Public School Teachers' Salaries by Geographical Region

Provinces	Salary	Spending	D2	D3
Connecticut	60,822	12,436	1	0
Illinois	58,246	9,275	1	0
Indiana	47,831	8,935	1	0
Iowa	43,130	7,807	1	0
Kansas	43,334	8,373	1	0
Maine	41,596	11,285	1	0
Massachusetts	58,624	12,596	1	0
Michigan	54,895	9,880	1	0
Minnesota	49,634	9,675	1	0
Missouri	41,839	7,840	1	0
Nebraska	42,044	7,900	1	0
New Hampshire	46,527	10,206	1	0
New Jersey	59,920	13,781	1	0
New York	58,537	13,551	1	0
North Dakota	38,822	7,807	1	0

Table 9.1 Public School Teachers' Salaries by Geographical Region

Provinces	Salary	Spending	D2	D3
Ohio	51,937	10,034	1	0
Pennsylvania	54,970	10,711	1	0
Rhode	55,956	11,089	1	0
South Dakota	35,378	7,911	1	0
Vermont	48,370	12,475	1	0
Wisconsin	47,901	9,965	1	0
Alabama	43,389	7,706	0	1
Arkansas	44,245	8,402	0	1
Delaware	54,680	12,036	0	1
Columbia	59,000	15,508	0	1
Florida	45,308	7,762	0	1
Georgia	49,905	8,534	0	1
Kentucky	43,646	8,300	0	1
Louisiana	42,816	8,519	0	1

Table 9.1 Public School Teachers' Salaries by Geographical Region

Provinces	Salary	Spending	D2	D3
Maryland	56,927	9,771	0	1
Mississippi	40,182	7,215	0	1
North Carolina	46,410	7,675	0	1
Oklahoma	42,379	6,944	0	1
South Carolina	44,133	8,377	0	1
Tennessee	43,816	6,979	0	1
Texas	44,897	7,547	0	1
Virginia	44,727	9,275	0	1
West Virginia	40,531	9,886	0	1
Alaska	54,658	10,171	0	0
Arizona	45,941	5,585	0	0
California	63,640	8,486	0	0

Table 9.1 Public School Teachers' Salaries by Geographical Region

Provinces	Salary	Spending	D2	D3
Colorado	45,833	8,861	0	0
Hawaii	51,922	9,879	0	0
Idaho	42,798	7,042	0	0
Montana	41,225	8,361	0	0
Nevada	45,342	6,755	0	0
New Mexico	42,780	8,622	0	0
Oregon	50,911	8,649	0	0
Utah	40,566	5,347	0	0
Washington,	47,882	7,958	0	0
Wyoming	50,692	11,596	0	0

Note: D2 = 1 for states in the Northeast and North Central; 0 otherwise.
D3 = 1 for states in the South; 0 otherwise.
Source: National Educational Association, as reported in 2007.

EX. 9.1: Public School Teachers' Salaries by Geographical Region

- $49,538.71 (Northeast and North Central),
- $46,293.59 (South), and
- $48,104.62 (West).
- These numbers look different, but are they statistically different from one another?
- There are various statistical techniques to compare two or more mean values, which generally go by the name of analysis of variance.
- But the same objective can be accomplished within the framework of regression analysis.

EX. 9.1: Public School Teachers' Salaries by Geographical Region

- To see this, consider the following model:
- $Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i$ **(9.2.1)**
  where Yi = (average) salary of public school teacher in state i
- D_{2i} = 1 if the state is in the Northeast or North Central = 0 otherwise (i.e., in other regions of the country)
- D_{3i} = 1 if the state is in the South = 0 otherwise (i.e., in other regions of the country)
- Note that Eq. (9.2.1) is like any multiple regression model considered previously, except that, instead of quantitative regressors, we have only qualitative, or dummy, regressors,

EX. 9.1: Public School Teachers' Salaries by Geographical Region

- Taking the value of 1 if the observation belongs to a particular category and 0 if it does not belong to that category or group.
- Hereafter, we shall designate all dummy variables by the letter D. Table 9.1 shows the dummy variables thus constructed.
- What does the model (9.2.1) tell us?
- Assuming that the error term satisfies the usual OLS assumptions, on taking expectation of Eq. (9.2.1) on both sides, we obtain:
- Mean salary of public school teachers in the Northeast and North Central:

$$E(Y_i \mid D_{2i} = 1, D_{3i} = 0) = \beta_1 + \beta_2 \qquad \textbf{(9.2.2)}$$

EX. 9.1: Public School Teachers' Salaries by Geographical Region

- Mean salary of public school teachers in the South:
  $E(Y_i \mid D_{2i} = 0, D_{3i} = 1) = \beta_1 + \beta_3$ **(9.2.3)**
- Now, how did we find out the mean salary of teachers in the West.
- If we guessed that this is equal to β_1, we would be absolutely right, for
- Mean salary of public school teachers in the West:
- $E(Y_i \mid D_{2i} = 0, D_{3i} = 0) = \beta_1$ **(9.2.4)**
- In other words, the mean salary of public school teachers in the West is given by the intercept, β_1, *in the multiple regression (9.2.1), and the "slope" coefficients* β_2 *and* β_3 *tell* by how much the mean salaries of teachers in the Northeast and North Central and in the South differ from the mean salary of teachers in the West.

EX. 9.1: Public School Teachers' Salaries by Geographical Region

- But how do we know if these differences are statistically significant?
- Before we answer this question, let us present the results based on the regression (9.2.1).
- Using the data given in Table 9.1, we obtain the following results:
- $\hat{Y}i = 48{,}014.615 + 1{,}524.099D_{2i} - 1{,}721.027D_{3i}$

  se = (1857.204) (2363.139) (2467.151)

  t = (25.853) (0.645) (−0.698) **(9.2.5)**

  (0.0000)* (0.5220)* (0.4888)* $R^2 = 0.0440$

  where * indicates the *p values*

Caution in the Use of Dummy Variables

- Our two dummy variables would sum to one (1) at every observations.
- That is, $D_1 + D_2 = 1$ for all observations; **perfect collinearity,** i.e, exact linear relationships among the variables.
- If we include a constant term, we face the problem of perfect multi-collinearity problem (i.e., linear dependence exists among columns of X Matrix.
- This is known as **dummy variable trap.**
- To avoid the dummy variable trap, we can either drop the dummy for one category as in the earlier case or we can include dummies for all categories without intercept term.

Caution in the Use of Dummy Variables

- **Rule: If a qualitative variable has m categories, introduce only (m – 1) dummy variables.**
- 2. The category for which no dummy variable is assigned is known as the **base, benchmark, control, comparison, reference, or omitted category. And all comparisons** are made in relation to the benchmark category.
- 3. The intercept value (β_1) *represents the mean value of the benchmark category.*
- *In* Example 9.1, the benchmark category is the Western region.
- Hence, in the regression (9.2.5) the intercept value of about 48,015 represents the mean salary of teachers in the Western states.

Caution in the Use of Dummy Variables

- 4. The coefficients attached to the dummy variables in Eq. (9.2.1) are known as the **differential intercept coefficients because they tell by how much the value of the category** that receives the value of 1 differs from the intercept coefficient of the benchmark category.
- For ex., in Eq. (9.2.5), the value of about 1,524 tells us that the mean salary of teachers in the Northeast or North Central is larger by about $1,524 than the mean salary of about $48,015 for the benchmark category, the West.

Caution in the Use of Dummy Variables

- 5. Dummy variable trap: to circumvent this trap by introducing as many dummy variables as the number of categories of that variable, provided we do not introduce the intercept in such a model.
- Thus, if we drop the intercept term from Eq. (9.2.6), and consider the following model,
- $Y_i = \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{2i} + u_i$ **(9.2.7)**
- we do not fall into the dummy variable trap, as there is no longer perfect collinearity.
- But make sure that when we run this regression, we use the no-intercept option in our regression package.

Summary and Conclusion

- Data and data analysis
- Variables and variable types and their respective statistical tests
- Dummy variables and their types
- Dummy variables in Multiple Regression
- Dummy variables, taking values of 1 and zero (or their linear transforms), are a means of introducing qualitative regressors in regression models.

Summary and Conclusions

- Dummy variables are a data-classifying device in that they divide a sample into various subgroups based on qualities or attributes (gender, marital status, race, religion, etc.) and *implicitly allow one to run individual regressions for each subgroup.*
- *If there are* differences in the response of the regressand to the variation in the qualitative variables in the various subgroups, they will be reflected in the differences in the intercepts or slope coefficients, or both, of the various subgroup regressions.
- DV technique needs to be handled carefully. First, if the regression contains a constant term, the number of DVs must be m-1.

Summary and Conclusions

- The coefficient attached to the dummy variables must always be interpreted in relation to the base, or reference, group—i.e, the group that receives the value of zero.
- The base chosen will depend on the purpose at hand.
- Finally, if a model has several qualitative variables with several classes, introduction of DV can consume a large number of df.

Reference

Chapter 9: Dummy Variable Regression Models, in Basic Econometrics by Damodar Gujarati.

What Next?

- Dummy variables in Multiple Regression – several interesting economic phenomenon