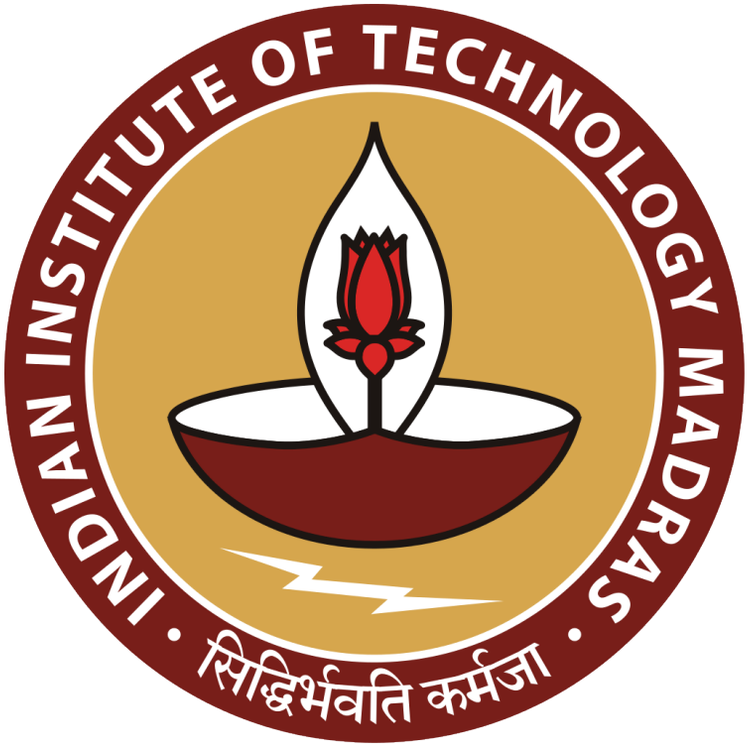

IIT Madras

ONLINE DEGREE

Statistics for Data Science -1

Lecture 4.9: Association between categorical and numerical variables

Usha Mohan

Indian Institute of Technology Madras

Introduction

- Understand the association between a categorical variable and numerical variable.
- Assume the categorical variable has two categories (dichotomous)

Example 1: Gender versus marks

A teacher was interested in knowing if female students performed better than male students in her class. She collected data from twenty students and the marks they obtained on 100 in the subject.

Example 1: Gender versus marks-Data

	Gender	Marks
1	F	71
2	F	67
3	F	65
4	M	69
5	M	75
6	M	83
7	F	91
8	F	85
9	F	69
10	F	75
11	M	92
12	F	79
13	M	71
14	M	94
15	F	86
16	F	75
17	F	90
18	M	84
19	F	91
20	M	90

Example 1: Scatter plot

Example 1: Scatter plot

Point Bi-serial Correlation Coefficient

- Let X be a numerical variable and Y be a categorical variable with two categories (a dichotomous variable).
- The following steps are used for calculating the Point Bi-serial correlation between these two variables:

Step 1 Group the data into two sets based on the value of the dichotomous variable Y. That is, assume that the value of Y is either 0 or 1.

Step 2 Calculate the mean values of two groups: Let $\bar{Y}_0$ and $\bar{Y}_1$ be the mean values of groups with $Y = 0$, and $Y = 1$, respectively.

Step 3 Let p_0 and p_1 be the proportion of observations in a group with $Y = 0$ and $Y = 1$, respectively, and s_X be the standard deviation of the random variable X.

The correlation coefficient

$$r_{pb} = \left(\frac{\bar{Y}_0 - \bar{Y}_1}{s_x} \right) \sqrt{p_0 p_1}$$