BIG DATA STATISTICS FOR BUSINESS

AY 2023-24

Giovanni De Luca
Parthenope University of Naples

1

Summary measures for categorical variables

- The most important summary measure of a categorical variable is the **mode**.
- The **mode** is the most frequent category.
- To find the mode, we first have to organize the data in a table.

2

2

Data visualization

- For a categorical variable, we consider

1. Bar plot
2. Pie chart

3

3

Graphical representation for categorical variables

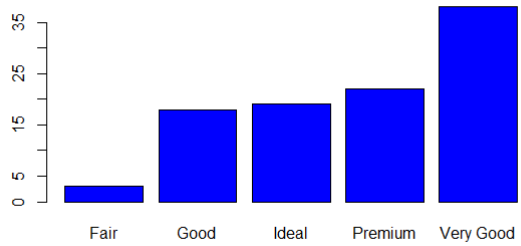
- The **bar plot** is the graphical representation used for categorical variable.
- In the dataset *Diamonds* diamonds are classified by the categorical variables *Cut* with categories «Fair», «Good», «Very Good», «Premium», «Ideal» and *Color*, with categories «*D*» (best), «*E*», «*F*», «*G*», «*H*», «*I*», «*J*» (worst).

4

4

Graphical representation for categorical variables

- **Bar plot** for the categorical variable *Cut*.

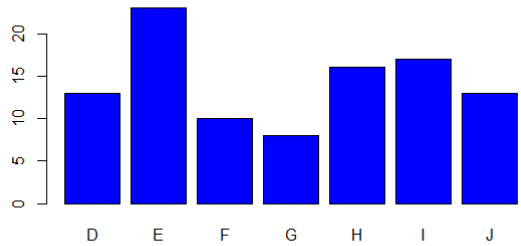


5

5

Graphical representation for categorical variables

- **Bar plot** for the categorical variable *Color*.



6

6

Graphical representation for categorical variables

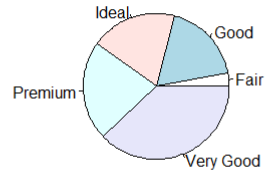
- The **pie chart** is a circle having a slice of pie for each category. The size of the slice corresponds to the percentage of observation in the category.

7

7

Graphical representation for categorical variables

- **Pie chart** for the categorical variable *Cut*.

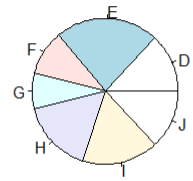


8

8

Graphical representation for categorical variables

- **Pie chart** for the categorical variable *Color*.



9

9

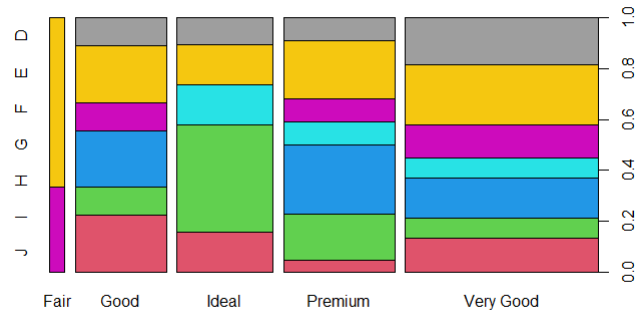
Association between two categorical variables

- The association between two categorical variables can be visualized using a **bar plot** showing the percentages of the categories of a variable in correspondence of the categories of the other variable.

10

10

Association between variables *cut* and *color*



11

11

Association between two categorical variables

- The association between two categorical variables is measured through the Chi-squared (χ^2) statistics.
- $\chi^2 = 0$ implies independence (no association)
- A low value of χ^2 suggests a weak association.
- The higher the χ^2, the stronger the association.
- The hypothesis testing considers

$$H_0: \chi^2 = 0$$

- We reject the H_0 if the *p*-value of the test is less than 0.05,

$$p\text{-value} < 0.05$$

12

12

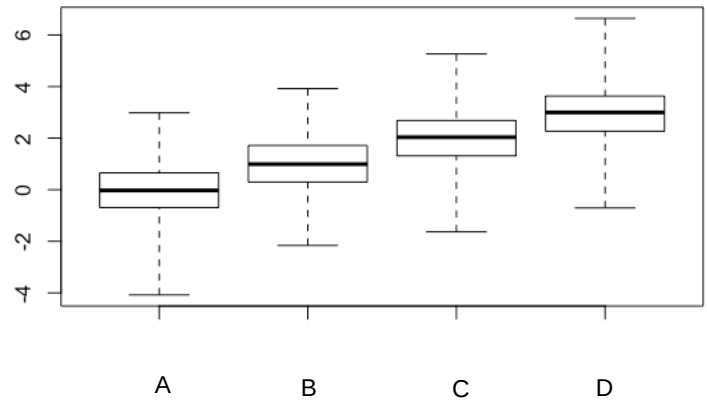
Association between a categorical variable and a numerical variable

- The association between a categorical variable and a numerical variable can be enlighted using a sequence of box plots.
- The number of box plots is equal to the number of categories.

13

13

Association between a categorical variable and a numerical variable



14

14

Mean dependence

- The association between a numerical and a categorical variable is measured through the mean dependence.
- We want to test if the mean of the numerical variable changes in the subsets defined by the categories.

15