IIT Madras

ONLINE DEGREE

Statistics for Data Science -1

Describing Categorical Data- Single Variable

Usha Mohan

Indian Institute of Technology Madras

Summarizing categorical data

- Graphical summaries of categorical data: bar chart and pie chart.

Summarizing categorical data

- Graphical summaries of categorical data: bar chart and pie chart.
- Need for a compact measure.

Summarizing categorical data

- Graphical summaries of categorical data: bar chart and pie chart.
- Need for a compact measure.
- Numbers that are used to describe data sets are called descriptive measures.

Summarizing categorical data

- Graphical summaries of categorical data: bar chart and pie chart.
- Need for a compact measure.
- Numbers that are used to describe data sets are called descriptive measures.
- Descriptive measures that indicate where the center or most typical value of a data set lies are called measures of central tendency.

Mode

Definition

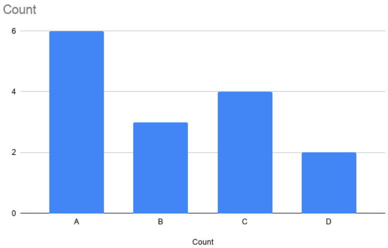
The mode of a categorical variable is the most common category, the category with the highest frequency

The mode labels

- The longest bar in a bar chart
- The widest slice in a pie chart.
- In a Pareto chart, the mode is the first category shown.

Example

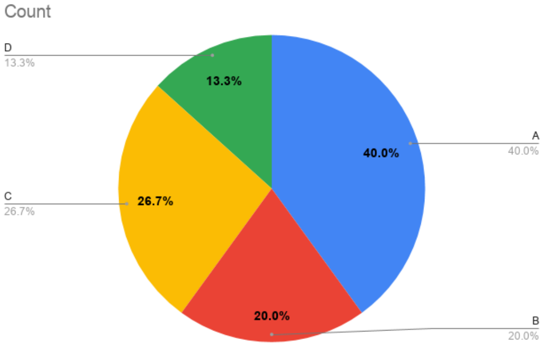
- Let consider the example A,A,B,C,A,D,A,B,C,C, A,B,C,D,A
- The longest bar in a bar chart



The most common category is "A"

Example

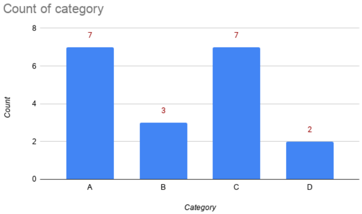
- Let consider the example A,A,B,C,A,D,A,B,C,C, A,B,C,D,A
- The widest slice in a pie chart.



The most common category is "A"

Bimodal and multimodal data

- If two or more categories tie for the highest frequency, the data are said to be bimodal (in the case of two) or multimodal (more than two).
- Let consider the example A,A,B,C,A,C,A,B,C,C, A,C,C,D,A,A,C,D,B



- Both category "A" and "C" have highest frequency.

Median

- Ordinal data offer another summary, the median, that is not available unless the data can be put into order.

Median

- Ordinal data offer another summary, the median, that is not available unless the data can be put into order.

Definition

The median of an ordinal variable is the category of the middle observation of the sorted values.

- If there are an even number of observations, choose the category on either side of the middle of the sorted list as the median.

Example

- Consider the grades of 15 students which is listed as A,B,B,C,A,D,B,B,A,C, B,B,C,D,A

Example

- Consider the grades of 15 students which is listed as A,B,B,C,A,D,B,B,A,C, B,B,C,D,A
  - The ordered data is A,A,A,A,B,B,B,B,B,B,C,C,C,D,D

Example

- Consider the grades of 15 students which is listed as A,B,B,C,A,D,B,B,A,C, B,B,C,D,A
  - The ordered data is A,A,A,A,B,B,B,B,B,B,C,C,C,D,D
  - The median grade is the category associated with the 8 observation which is "B".
- Consider the grades of 14 students which is listed as A,B,B,C,A,D,B,B,A,C, B,B,C,D

Example

- Consider the grades of 15 students which is listed as A,B,B,C,A,D,B,B,A,C, B,B,C,D,A
  - The ordered data is A,A,A,A,B,B,B,B,B,B,C,C,C,D,D
  - The median grade is the category associated with the 8 observation which is "B".
- Consider the grades of 14 students which is listed as A,B,B,C,A,D,B,B,A,C, B,B,C,D
  - The ordered data is A,A,A,B,B,B,B,B,B,C,C,C,D,D

Example

- Consider the grades of 15 students which is listed as A,B,B,C,A,D,B,B,A,C, B,B,C,D,A
  - The ordered data is A,A,A,A,B,B,B,B,B,B,C,C,C,D,D
  - The median grade is the category associated with the 8 observation which is "B".
- Consider the grades of 14 students which is listed as A,B,B,C,A,D,B,B,A,C, B,B,C,D
  - The ordered data is A,A,A,B,B,B,B,B,B,C,C,C,D,D
  - The median grade is the category associated with the 7 or 8 observation which is "B".

Example

- Consider the grades of 15 students which is listed as A,B,B,C,A,D,B,B,A,C, B,B,C,D,A

Example

- Consider the grades of 15 students which is listed as A,B,B,C,A,D,B,B,A,C, B,B,C,D,A
- The ordered data is A,A,A,A,B,B,B,B,B,B,C,C,C,D,D

Example

- Consider the grades of 15 students which is listed as A,B,B,C,A,D,B,B,A,C, B,B,C,D,A
- The ordered data is A,A,A,A,B,B,B,B,B,B,C,C,C,D,D
- The median grade is the category associated with the 8 observation which is "B".

Example

- Consider the grades of 15 students which is listed as A,B,B,C,A,D,B,B,A,C, B,B,C,D,A
- The ordered data is A,A,A,A,B,B,B,B,B,B,C,C,C,D,D
- The median grade is the category associated with the 8 observation which is "B".
- The most common grade is "B", hence mode is "B"

Example

- Consider the grades of 15 students which is listed as A,B,B,C,A,D,B,B,A,C, B,B,C,D,A
- The ordered data is A,A,A,A,B,B,B,B,B,B,C,C,C,D,D
- The median grade is the category associated with the 8 observation which is "B".
- The most common grade is "B", hence mode is "B"
- In this example both mode and median are the same.

Example

- Consider the grades of 15 students which is listed as A,B,B,C,A,D,A,B,A,C,B,A,C,D,A

Example

- Consider the grades of 15 students which is listed as A,B,B,C,A,D,A,B,A,C,B,A,C,D,A
- The ordered data is A,A,A,A,A,A,B,B,B,B,C,C,C,D,D

Example

- Consider the grades of 15 students which is listed as A,B,B,C,A,D,A,B,A,C,B,A,C,D,A
- The ordered data is A,A,A,A,A,A,B,B,B,B,C,C,C,D,D

Example

- Consider the grades of 15 students which is listed as A,B,B,C,A,D,A,B,A,C,B,A,C,D,A
- The ordered data is A,A,A,A,A,A,B,B,B,B,C,C,C,D,D
- The median grade is the category associated with the 8 observation which is "B".

Example

- Consider the grades of 15 students which is listed as A,B,B,C,A,D,A,B,A,C,B,A,C,D,A
- The ordered data is A,A,A,A,A,A,B,B,B,B,C,C,C,D,D
- The median grade is the category associated with the 8 observation which is "B".
- The most common grade is "A", hence mode is "A"

Example

- Consider the grades of 15 students which is listed as A,B,B,C,A,D,A,B,A,C,B,A,C,D,A
- The ordered data is A,A,A,A,A,A,B,B,B,B,C,C,C,D,D
- The median grade is the category associated with the 8 observation which is "B".
- The most common grade is "A", hence mode is "A"
- In this example both mode and median are the different.

Sectional summary

- The mode of a categorical variable is the most common category.
- The median of an ordinal variable is the category of the middle observation of the sorted values.

Summary

1. Tabulate data: frequency and relative frequency.
2. Charts of categorical data
   2.1 Pie charts
   2.2 Bar charts and Pareto charts
3. Best practices and misleading graphs
   3.1 Label your data.
   3.2 Dealing with multiple categories.
   3.3 Area principle
   3.4 Misleading graphs
      3.4.1 Decorated graphs
      3.4.2 Truncated graphs.
      3.4.3 Round-off errors.
4. Descriptive measures
   4.1 Mode.
   4.2 Median for ordinal data.