INDIAN INSTITUTE OF TECHNOLOGY MADRAS

सिद्धिर्भवति कर्मजा

IIT Madras

ONLINE DEGREE

Statistics for Data Science -1

Lecture 3.1: Describing Numerical Data- Frequency tables for numerical data

Usha Mohan

Indian Institute of Technology Madras

Review

Review

1. What is statistics?
    - Descriptive statistics, inferential statistics.
    - Distinguish between a sample and a population.

Review

1. What is statistics?
    - Descriptive statistics, inferential statistics.
    - Distinguish between a sample and a population.
2. Understand how data are collected.
    - Identify variables and cases (observations) in a data set

Review

1. What is statistics?
   - Descriptive statistics, inferential statistics.
   - Distinguish between a sample and a population.
2. Understand how data are collected.
   - Identify variables and cases (observations) in a data set
3. Types of data-
   - classify data as categorical(qualitative) or numerical(quantitative) data.
   - Understand cross-sectional versus time-series data.
   - Measurement scales-nominal, ordinal, interval and ratio.

Review

1. What is statistics?
   - Descriptive statistics, inferential statistics.
   - Distinguish between a sample and a population.
2. Understand how data are collected.
   - Identify variables and cases (observations) in a data set
3. Types of data-
   - classify data as categorical(qualitative) or numerical(quantitative) data.
   - Understand cross-sectional versus time-series data.
   - Measurement scales-nominal, ordinal, interval and ratio.
4. Describing categorical data

Review

1. What is statistics?
   - Descriptive statistics, inferential statistics.
   - Distinguish between a sample and a population.
2. Understand how data are collected.
   - Identify variables and cases (observations) in a data set
3. Types of data-
   - classify data as categorical(qualitative) or numerical(quantitative) data.
   - Understand cross-sectional versus time-series data.
   - Measurement scales-nominal, ordinal, interval and ratio.
4. Describing categorical data
   - Creating frequency tables, understanding relative frequency

Review

1. What is statistics?
   - Descriptive statistics, inferential statistics.
   - Distinguish between a sample and a population.
2. Understand how data are collected.
   - Identify variables and cases (observations) in a data set
3. Types of data-
   - classify data as categorical(qualitative) or numerical(quantitative) data.
   - Understand cross-sectional versus time-series data.
   - Measurement scales-nominal, ordinal, interval and ratio.
4. Describing categorical data
   - Creating frequency tables, understanding relative frequency
   - Creating pie charts and bar charts

Review

1. What is statistics?
   - Descriptive statistics, inferential statistics.
   - Distinguish between a sample and a population.
2. Understand how data are collected.
   - Identify variables and cases (observations) in a data set
3. Types of data-
   - classify data as categorical(qualitative) or numerical(quantitative) data.
   - Understand cross-sectional versus time-series data.
   - Measurement scales-nominal, ordinal, interval and ratio.
4. Describing categorical data
   - Creating frequency tables, understanding relative frequency
   - Creating pie charts and bar charts
   - Understanding violations

Review

1. What is statistics?
   - Descriptive statistics, inferential statistics.
   - Distinguish between a sample and a population.
2. Understand how data are collected.
   - Identify variables and cases (observations) in a data set
3. Types of data-
   - classify data as categorical(qualitative) or numerical(quantitative) data.
   - Understand cross-sectional versus time-series data.
   - Measurement scales-nominal, ordinal, interval and ratio.
4. Describing categorical data
   - Creating frequency tables, understanding relative frequency
   - Creating pie charts and bar charts
   - Understanding violations
   - Descriptive measures of Mode and Median

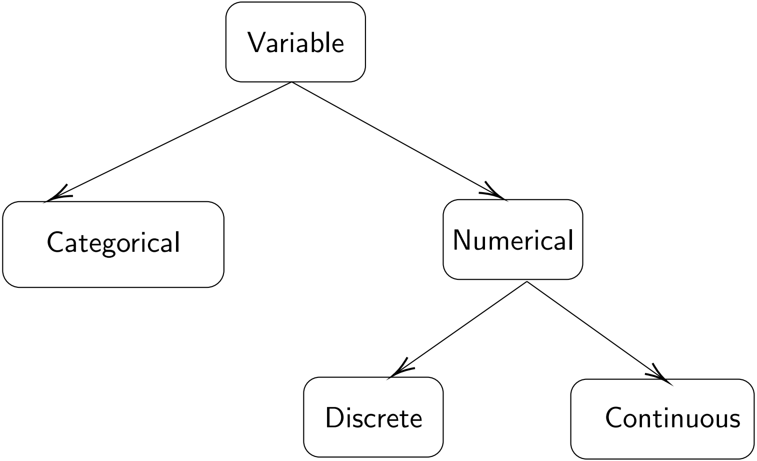
Frequency tables

- Organizing numerical data

Graphical summaries

- Histograms
- Stem-and-leaf diagram

Types of variables



Organizing numerical data

Organizing numerical data

- Recall, a discrete variable usually involves a count of something, whereas a continuous variable usually involves a measurement of something.

Organizing numerical data

- Recall, a discrete variable usually involves a count of something, whereas a continuous variable usually involves a measurement of something.
- First group the observations into classes (also known as categories or bins) and then treat the classes as the distinct values of qualitative data.

Organizing numerical data

- Recall, a discrete variable usually involves a count of something, whereas a continuous variable usually involves a measurement of something.
- First group the observations into classes (also known as categories or bins) and then treat the classes as the distinct values of qualitative data.
- Once we group the quantitative data into classes, we can construct frequency and relative-frequency distributions of the data in exactly the same way as we did for categorical data.

Organizing discrete data (single value)

Organizing discrete data (single value)

- If the data set contains only a relatively small number of distinct, or different, values, it is convenient to represent it in a frequency table.

Organizing discrete data (single value)

- If the data set contains only a relatively small number of distinct, or different, values, it is convenient to represent it in a frequency table.
- Each class represents a distinct value (single value) along with its frequency of occurrence.

Example

- Suppose the dataset reports the number of people in a household. The following data is the response from 15 individuals.

Example

- Suppose the dataset reports the number of people in a household. The following data is the response from 15 individuals.
- 2,1,3,4,5,2,3,3,3,4,4,1,2,3,4

Example

- Suppose the dataset reports the number of people in a household. The following data is the response from 15 individuals.
- 2,1,3,4,5,2,3,3,3,4,4,1,2,3,4
- The distinct values the variable, number of people in each household, takes is 1,2,3,4,5.
- The frequency distribution table is

Value	Tally mark	Frequency	Relative frequency
1			
2			
3			
4			
5			
Total			

Graph

Organizing continuous data

Organize the data into a number of classes to make the data understandable. However, there are few guidelines that need to be followed. They are

Organizing continuous data

Organize the data into a number of classes to make the data understandable. However, there are few guidelines that need to be followed. They are

1. Number of classes: The appropriate number is a subjective choice, the rule of thumb is to have between 5 and 20 classes.

Organizing continuous data

Organize the data into a number of classes to make the data understandable. However, there are few guidelines that need to be followed. They are

1. Number of classes: The appropriate number is a subjective choice, the rule of thumb is to have between 5 and 20 classes.
2. Each observation should belong to some class and no observation should belong to more than one class.

Organizing continuous data

Organize the data into a number of classes to make the data understandable. However, there are few guidelines that need to be followed. They are

1. Number of classes: The appropriate number is a subjective choice, the rule of thumb is to have between 5 and 20 classes.
2. Each observation should belong to some class and no observation should belong to more than one class.
3. It is common, although not essential, to choose class intervals of equal length.

Some new terms

Some new terms

1. Lower class limit: The smallest value that could go in a class.

Some new terms

1. Lower class limit: The smallest value that could go in a class.
2. Upper class limit: The largest value that could go in a class.

Some new terms

1. Lower class limit: The smallest value that could go in a class.
2. Upper class limit: The largest value that could go in a class.
3. Class width: The difference between the lower limit of a class and the lower limit of the next-higher class.

Some new terms

1. Lower class limit: The smallest value that could go in a class.
2. Upper class limit: The largest value that could go in a class.
3. Class width: The difference between the lower limit of a class and the lower limit of the next-higher class.
4. Class mark: The average of the two class limits of a class.

Some new terms

1. Lower class limit: The smallest value that could go in a class.
2. Upper class limit: The largest value that could go in a class.
3. Class width: The difference between the lower limit of a class and the lower limit of the next-higher class.
4. Class mark: The average of the two class limits of a class.
5. A class interval contains its left-end but not its right-end boundary point.

Example

Example

- The marks obtained by 50 students in a particular course.

Example

- The marks obtained by 50 students in a particular course.
- 68, 79, 38, 68, 35, 70, 61, 47, 58, 66, 60, 45, 61, 60, 59, 45, 39, 80, 59, 62, 49, 76, 54, 60, 53, 55, 62, 58, 67, 55, 86, 56, 63, 64, 67, 50, 51, 78, 56, 62, 57, 69, 58, 52, 42, 66, 42, 56, 58.

Example

- The marks obtained by 50 students in a particular course.
- 68, 79, 38, 68, 35, 70, 61, 47, 58, 66, 60, 45, 61, 60, 59, 45, 39, 80, 59, 62, 49, 76, 54, 60, 53, 55, 62, 58, 67, 55, 86, 56, 63, 64, 67, 50, 51, 78, 56, 62, 57, 69, 58, 52, 42, 66, 42, 56, 58.

Class interval	Tally mark	Frequency	Relative frequency
30-40			
40-50			
50-60			
60-70			
70-80			
80-90			
Total			

Frequency table

68, 79, 38, 68, 35, 70, 61, 47, 58, 66, 60, 45, 61, 60, 59, 45, 39, 80, 59, 62, 49, 76, 54, 60, 53, 55, 62, 58, 67, 55, 86, 56, 63, 64, 67, 50, 51, 78, 56, 62, 57, 69, 58, 52, 42, 66, 42, 56, 58.

Class interval	Tally mark	Frequency	Relative frequency
30-40	𝍷𝍷𝍷	3	0.06
40-50	𝍸𝍷	6	0.12
50-60	𝍸𝍸𝍸𝍷𝍷𝍷	18	0.36
60-70	𝍸𝍸𝍸𝍷𝍷	17	0.34
70-80	𝍷𝍷𝍷𝍷	4	0.08
80-90	𝍷𝍷	2	0.04
Total		50	1

Section summary

1. Frequency table for discrete single value data.
2. Frequency table for continuous data using class intervals.

Steps to construct a histogram

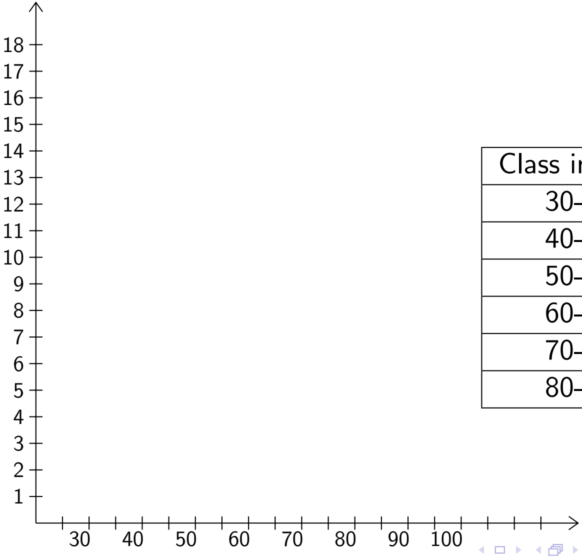
Step 1 Obtain a frequency (relative-frequency) distribution of the data.

Step 2 Draw a horizontal axis on which to place the classes and a vertical axis on which to display the frequencies (relative frequencies).

Step 3 For each class, construct a vertical bar whose height equals the frequency (relative frequency) of that class.

Step 4 Label the bars with the classes, the horizontal axis with the name of the variable, and the vertical axis with "Frequency" ("Relative frequency").

Histogram



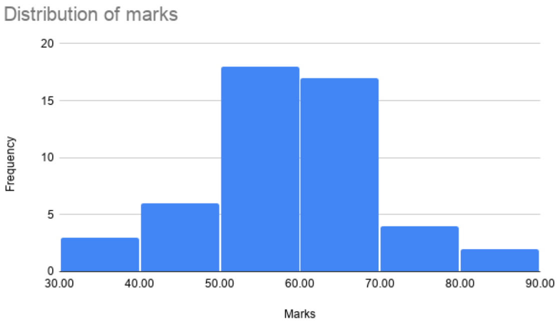
Class interval	frequency
30-40	3
40-50	6
50-60	18
60-70	17
70-80	4
80-90	2

Histogram

`https://docs.google.com/spreadsheets/d/109W3ga8TZG3pWJwofG4h0yE7xvoGOK_kCvmm0e9w0kQ/edit?usp=sharing`

Histogram

`https://docs.google.com/spreadsheets/d/109W3ga8TZG3pWJwofG4h0yE7xvoGOK_kCvmmOe9wOkQ/edit?usp=sharing`



Stem-and-leaf diagram

Definition

In a stem-and-leaf diagram (or stemplot)[1] *, each observation is separated into two parts, namely, a stem-consisting of all but the rightmost digit-and a leaf, the rightmost digit.*

[1] Weiss, Neil A. Introductory Statistics: Pearson New International Edition. Pearson Education Limited, 2014.

Stem-and-leaf diagram

Definition

In a stem-and-leaf diagram (or stemplot)[1] *, each observation is separated into two parts, namely, a stem-consisting of all but the rightmost digit-and a leaf, the rightmost digit.*

- For example, if the data are all two-digit numbers, then we could let the stem of a data value be the tens digit and the leaf be the ones digit.

[1] Weiss, Neil A. Introductory Statistics: Pearson New International Edition. Pearson Education Limited, 2014.

Stem-and-leaf diagram

Definition

In a stem-and-leaf diagram (or stemplot)[1] *, each observation is separated into two parts, namely, a stem-consisting of all but the rightmost digit-and a leaf, the rightmost digit.*

- For example, if the data are all two-digit numbers, then we could let the stem of a data value be the tens digit and the leaf be the ones digit.
  - The value 75 is expressed as

[1] Weiss, Neil A. Introductory Statistics: Pearson New International Edition. Pearson Education Limited, 2014.

Stem-and-leaf diagram

Definition

In a stem-and-leaf diagram (or stemplot)[1] *, each observation is separated into two parts, namely, a stem-consisting of all but the rightmost digit-and a leaf, the rightmost digit.*

- For example, if the data are all two-digit numbers, then we could let the stem of a data value be the tens digit and the leaf be the ones digit.
  - The value 75 is expressed as

Stem	Leaf
7 |	5

[1] Weiss, Neil A. Introductory Statistics: Pearson New International Edition. Pearson Education Limited, 2014.

Stem-and-leaf diagram

Definition

In a stem-and-leaf diagram (or stemplot)[1] *, each observation is separated into two parts, namely, a stem-consisting of all but the rightmost digit-and a leaf, the rightmost digit.*

- For example, if the data are all two-digit numbers, then we could let the stem of a data value be the tens digit and the leaf be the ones digit.
  - The value 75 is expressed as

Stem	Leaf
7	5

  - The two values 75, 78 is expressed as

[1] Weiss, Neil A. Introductory Statistics: Pearson New International Edition. Pearson Education Limited, 2014.

Stem-and-leaf diagram

Definition

In a stem-and-leaf diagram (or stemplot)[1] *, each observation is separated into two parts, namely, a stem-consisting of all but the rightmost digit-and a leaf, the rightmost digit.*

- For example, if the data are all two-digit numbers, then we could let the stem of a data value be the tens digit and the leaf be the ones digit.
  - The value 75 is expressed as

Stem	Leaf
7	5

  - The two values 75, 78 is expressed as

Stem	Leaf
7	5,8

[1] Weiss, Neil A. Introductory Statistics: Pearson New International Edition. Pearson Education Limited, 2014.

Steps to construct a stemplot

Step 1 Think of each observation as a stem—consisting of all but the rightmost digit—and a leaf, the rightmost digit.

Step 2 Write the stems from smallest to largest in a vertical column to the left of a vertical rule.

Step 3 Write each leaf to the right of the vertical rule in the row that contains the appropriate stem.

Step 4 Arrange the leaves in each row in ascending order.

Example

- The following are the ages, to the nearest year, of 11 patients admitted in a certain hospital: 15, 22, 29, 36, 31, 23, 45, 10, 25, 28, 48

Example

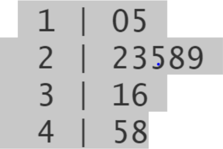
- The following are the ages, to the nearest year, of 11 patients admitted in a certain hospital: 15, 22, 29, 36, 31, 23, 45, 10, 25, 28, 48
- Draw a stem-and-leaf plot for this data set.

Example

- The following are the ages, to the nearest year, of 11 patients admitted in a certain hospital: 15, 22, 29, 36, 31, 23, 45, 10, 25, 28, 48
- Draw a stem-and-leaf plot for this data set.

Example

- The following are the ages, to the nearest year, of 11 patients admitted in a certain hospital: 15, 22, 29, 36, 31, 23, 45, 10, 25, 28, 48
- Draw a stem-and-leaf plot for this data set.



1	|	05
2	|	23589
3	|	16
4	|	58

Section summary

1. Construct a histogram for grouped data.
2. Construct a stemplot to describe numerical data.