Topic #1. Frequency Tables and Graphs

Cheng Peng

Contents

1 Basic Statistical Terminologies **1**
1.1 What is statistics? . . . 1
1.2 Population vs Sample . . . 1
1.3 Types Statistics and Data . . . 2

2 Summarizing Qualitative Data **2**
2.1 Frequency Tables . . . 2
2.2 Charting Categorical Data . . . 3
2.2.1 Bar Chart . . . 3
2.2.2 Pie Chart . . . 4

3 Summary of Numerical Data **4**
3.1 Frequency Table . . . 5
3.2 Histogram . . . 6

4 Exercises **6**

5 Use of Technology **7**

1 Basic Statistical Terminologies

In this note, we introduce basic terminology of statistics and methods for summarizing a given data set.

1.1 What is statistics?

Statistics is the science of collecting, organizing, visualizing, analyzing, and interpreting data in order to make decision.

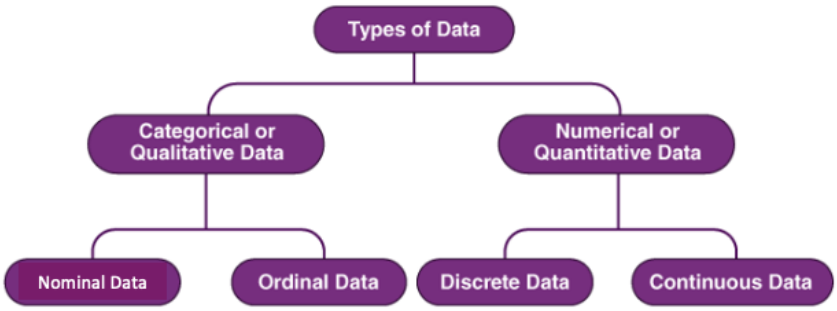
1.2 Population vs Sample

- **Population**: The collection of **all** outcomes, responses, measurements, or counts that are of interest (the right group in the following figure).
- **Statistics**: A subset of the population (the left group in the following figure).

- **Parameter**: the numeric characteristic of the population. For example, the average height of **all** students at WCU. Here `all WCU students` is a population.
- **Statistic**: the numeric characteristic of the sample (i.e., a subset of the population). For example, the average height of **subset of** students at WCU. Here `the subset of WCU students` is a sample taken from the population of all WCU students.

1.3 Types Statistics and Data

- **Descriptive Statistics** involves organizing, summarizing, and displaying data. For example, we can use tables, charts, averages, etc. All topics in this note and next note will focus on descriptive statistics.
- **Inferential Statistics** uses the sample data to make inferences about the underlying population. For example, all topics from week #3 are inferential statistics.
- **Data Types**: There different ways for classifying data in statistics. The following diagram given one of the simple bu widely used methods.



- **Data Types** examples
  - Nominal Data (also called unordered data): the place of birth, major, eye color, etc.
  - Ordinal Data: Military Rank (private, corporal, etc.), Course Grade (A, B, C, D, F), etc.
  - Discrete (a subset of which is "counting"): Number of children in a family, Shoe Size, etc.
  - Continuous: Weight, Height, temperature, income, GPA, etc.

2 Summarizing Qualitative Data

For a given categorical data, we can use frequency tables and charts to summarize the distribution of the data. Note that the given data set could either be a population or a sample.

2.1 Frequency Tables

Since each distinct data value represents a category, the number of values in each category is the frequency of the category. `An ordinary frequency table` is a two-column table in which the left column lists the category labels and the right column lists the corresponding frequencies.

There are four types of frequency tables. The other three frequency tables are relative frequency table, cumulative frequency table, cumulative relative frequency table.

`a relative frequency = ordinary frequency / total`

`a cumulative frequeny table` is constructed based on the cumulatively combined categories. See the following example for more detail.

Example 1: In a class of 20 students, 3 students received a grade of "A", 6 students received a "B", 7 students received a "C", 3 students received a "D", and 1 student received an "F". These results are summarized, in a variety of ways, in the following table: (Note that for the ordered categorical variable "Grade" we also create the discrete quantitative variable "Grade-point".)

Solution: The raw data might be in the following form:

`A, C, B, F, D, B, C, B, C, C, A, D, C, B, C, B, A, C, B, D`

The resulting frequency tables is given by

Grade	Freq	Rel. Freq	Cum. Freq	Cum. Rel. Freq
F	1	.05	1	.05
D	3	.15	4	.20
C	7	.35	11	.55
B	6	.30	17	.85
A	3	.15	20	1.00
Total	**20**	**1.00**		

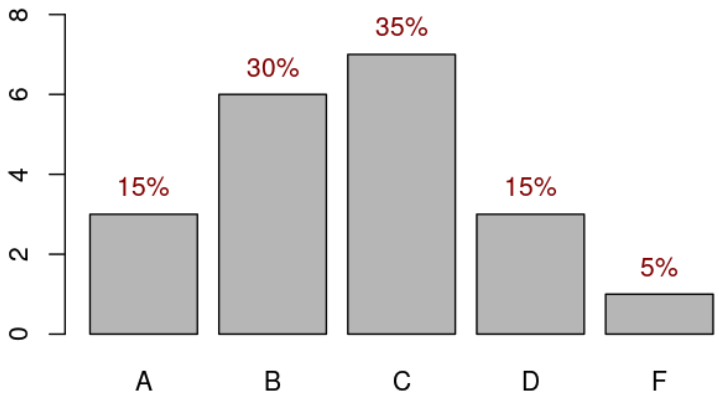
Remarks: (1). The cumulative categories are defined to be F, D or below, C or below, etc. (2). For a nominal data, the cumulative frequency table may not be practically meaningful because the combined categories may not make practical sense.

2.2 Charting Categorical Data

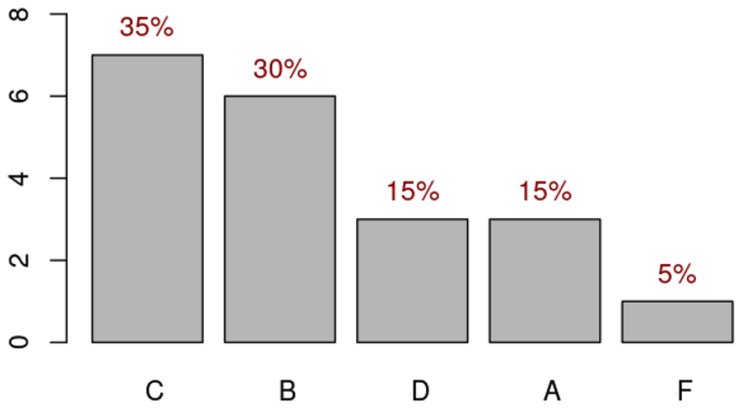
Two major charts are used to characterize the distribution of a given categorical data set: bar chart and pie chart. Both chart are geometric representations of the frequency table discussed earlier.

2.2.1 Bar Chart

Example 2: We convert the frequency table of the course grade data in the following



Remark: We can rearrange the vertical bars in ascending or descending order to get pare-to chart.



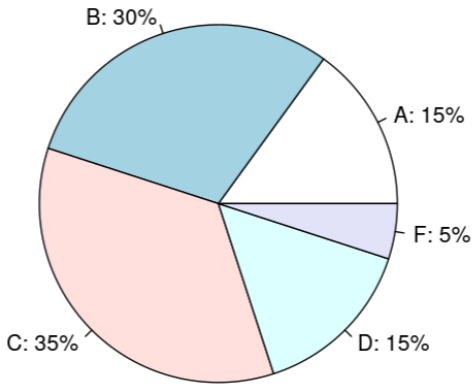
2.2.2 Pie Chart

To construct a pie chart manually, we need to calculate the degrees of the central angle of the circle and then slice it based on the degrees of the central angle.

Example 3: we still use the course grade frequency table (relative frequency) to to calculate the degrees of the corresponding central angle in the following table.

Grade	Relative Freq	Pie Chart Angle
F	.05	$0.05 \times 360^{o} = 18^{o}$
D	.15	$0.15 \times 360^{o} = 54^{o}$
C	.35	126^{o}
B	.30	108^{o}
A	.15	54^{o}
Total	1.00	360^{o}

The corresponding pie chart is given by



3 Summary of Numerical Data

There are primarily two methods that are commonly used to summarize a given numerical data: Frequency tables and histograms.

3.1 Frequency Table

A histogram displays numerical data by grouping data into "bins" of equal width. Each bin is plotted as a bar whose height corresponds to how many data points are in that bin. Bins are also sometimes called "intervals", "classes", or "buckets".

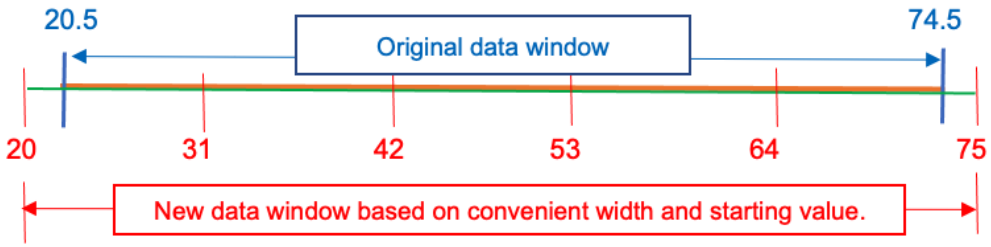
There are several steps to follow when creating bins (with equal width):

- Determine the number of bins
- Extend data window (from the smallest to the largest data values) if necessary to get "convenient" end values of the extended window. Caution: never shrink the data window because we must include all data values in one and only one of the bins!
- Find the boundary values (cut-offs) so that all bins have equal width which is equal to **data-window-width/number-of-bins**.
- The number of data values in each bin is the frequency of the bin.

Example 4 - Length of CD: Listed below are the lengths (in minutes) of randomly selected CDs of country, rock, and movie soundtracks.

```
20.5, 29, 32, 32, 32, 33, 36, 37, 38, 39, 39, 43, 47, 48, 49, 49, 49,
50, 50, 51, 51, 52, 52, 52, 53, 54, 54, 54, 56, 56, 57, 58, 60, 61, 62,
62, 69, 73, 74, 74.5
```

Solution: we follow the above suggested steps to define bins illustrated in the following figure.



- The number of bins chosen for this frequency table is 5.
- The original data window is '[20.5, 74.5]'. The two end values are decimals. We extended the data window on both sides and get a an extended window [20, 75].
- The bin width = (75-20)/5 = 11.
- The boundaries of the 5 bins are: 20, 31, 42, 53, 64, 75.
- the five bins are: [20, 31], (31, 42], (42, 53], (53, 64], (64, 75]. Note that the boundary values must be included in one and only one bins. We use `"["` and `")"` to denote inclusion and exclusion respectively. For example, in the second bin [31, 42), 31 is included in [31, 42) but 42 is NOT included in [31, 42). It is included in [42, 53).
- with the above defined bins, the resulting frequency tables are given by

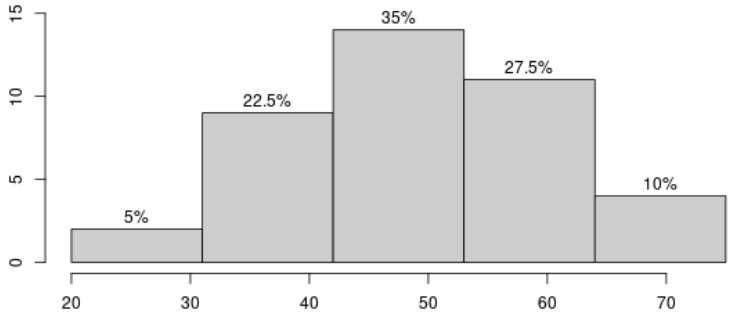
Classes	Freq	Cum. Freq	Rel. Freq	Cum. Rel. Freq.
[20, 31]	2	2	0.05	0.05
(31, 42]	9	11	0.225	0.275
(42, 53]	14	25	0.35	0.625
(53, 64]	11	36	0.275	0.9
(64, 75]	4	40	0.1	1.0

3.2 Histogram

Similar to the bar chart and pie chat, a histogram is also a geometric representation of the frequency table constructed above. Since the bins are defined based on the numerical boundaries, they must be placed on the correct scales when constructing a histogram. This also means that the histogram is different from the bar chart in different perspectives:

- There is no gaps between the adjacent vertical bars because the horizontal axis is a numerical axis.
- We cannot rearrange the vertical bars as we did in a bar chart to make a pareto chart since we cannot shuffle the boundaries on the numerical axis.

Example 5 - Length of CD (cont'd): The histogram based on the frequency table is given in the following.



4 Exercises

Summarize the following data sets by using frequency tables (relative frequency, cumulative frequency, etc.) and histogram. You can use `IntroStatsApps` to check your work.

Exercise 1. Weights of 18- to 24- Year- Old Males. The U. S. National Center for Health Statistics publishes data on weights and heights by age and sex in the document Vital and Health Statistics. The weights shown in the following, given to the nearest tenth of a pound, were obtained from a sample of 18- to 24- year- old males. Use the cut-point grouping to organize these data into frequency and relative- frequency distributions. Use a class width of 20 and a first cut-point of 120.

```
129.2, 132.1, 136.7, 142.8, 145.6, 146.4, 149.9, 150.7, 151.3, 155.2, 158.5, 158.6, 161.0, 161.7, 165.0
```

Exercise 2. The following are the miles per gallon.

```
22.8, 22.9, 23.3, 23.4, 23.6, 23.7, 23.8, 23.9, 23.9, 24.1, 24.1, 24.2, 24.3, 24.4, 24.5, 24.5, 24.6, 2
```

Exercise 3. Following are 80 measurements of the iron-solution index of tin-plate specimens, designed to measure the corrosion resistance of tin-plated steel. The original data set has been sorted in an ascending order as:

```
14, 26, 28, 28, 28, 28, 30, 32, 34, 35, 36, 36, 37, 37, 40, 40, 40, 41, 41, 41, 42,
```

Exercises 4. From the 140 children whose urinary concentration of lead were investigated 40 were chosen who were aged at least 1 year but under 5 years. The following concentrations of copper were found.

```
0.70, 0.45, 0.72, 0.30, 1.16, 0.69, 0.83, 0.74, 1.24, 0.77, 0.65, 0.76, 0.42, 0.94,
0.36, 0.98, 0.64, 0.90, 0.63, 0.55, 0.78, 0.10, 0.52, 0.42, 0.58, 0.62, 1.12, 0.86,
0.74, 1.04, 0.65, 0.66, 0.81, 0.48, 0.85, 0.75, 0.73, 0.50, 0.34, 0.88
```

Exercise 5 The following data set represents the shoe sizes of 100 random selected students from a large university.

```
8.0, 13.0, 8.5, 9.0, 11.0, 9.5, 10.0, 8.0, 11.0, 8.0, 10.0, 11.0, 10.0, 11.0, 6.0,
9.0, 8.0, 8.0, 12.0, 10.5, 9.5, 11.0, 6.0, 8.0, 10.0, 11.5, 11.0, 7.0, 10.5, 15.0,
12.0, 8.5, 8.0, 10.0, 8.0, 7.0, 10.5, 10.0, 5.0, 7.0, 10.0, 14.0, 14.0, 8.5, 8.0,
13.0, 11.0, 6.0, 8.0, 11.5, 8.5, 7.0, 12.5, 8.5, 15.0, 10.0, 6.0, 11.0, 11.0, 10.0,
10.5, 11.0, 7.5, 7.0, 7.5, 10.5, 10.0, 11.0, 9.5, 11.0, 9.5, 10.5, 7.5, 11.0, 13.0,
10.0, 9.0, 12.0, 8.0, 8.0, 9.0, 12.0, 8.5, 8.0, 11.0, 9.0, 9.0, 7.0, 9.0, 12.0, 5.5,
9.5, 8.0, 9.0, 12.0, 9.5, 9.0, 11.0, 13.0, 7.5
```

`Caution`: The data values are given in the numeric form, but they are labels shoe sizes. Therefore, this is a categorical data set.

5 Use of Technology

In this class, a piece of software created by myself will be used to check your work. The following screenshot gives the list of the available apps that cover about 90% topics in this class.

The following video demonstrates how to use the apps to find the solutions to the above examples based on the **course grades** and **length of CD** data sets.