Chapter 2 Notes - Statistics

Matt Warner

1 Types of Data

Terminology

- If we measure or record
  - one oberservation from each individual or object we have **univariate data**
  - two oberservations from each individual or object we have **bivariate data**
  - more than two oberservations from each individual or object we have **multivariate data** (e.g. height, weight, gender, blood presure, and cholesterol level of each person)
- Types of Data
  - **Categorical or qualitative** - non-numerical oberservations that may be placed in categories.
  - **Numerical or quantitative** - observations that are numbers.
- A numerical data set is
  - **Discrete** if its set of possible values is a finite set or a countable infinite set (i.e. an infinite sequence with a first value, second value, etc.)
  - **Continuous** if its set of possible values is an infinite set that forms an interval on the number line (usually associated with measuring)

Discrete vs Continuous

Continuous Data:	**Discrete Data:**
It can take on any value in an interval	It can only have specific values.
Data that can be measured (i.e. speed of a car)	Data that can be counted (i.e number of books)

2 Bar Charts and Pie Charts

Frequency distribution

Defintion

The natural summary measures for a categorical data set are the number of times each category occurred and the proportion of times each category occured. These values are usually displayed in a table as in Table 2.1

A **frequency distrobution** for categorical data is a summary table that presents categories, counts, and proportions.

1. Each unique value in a categorical data set is a label, or class. In Table 2.1 the classes are mammals, birds, reptiles, etc.
2. The frequency is the count for each class. In Table 2.1, the frequency for the mammals class is 202 (i.e. 202 mammals were on the critically endangered species list.)

3. The relative frequency, or sample proportion, for each class is the frequency of the class divided by the total number of observations. In Table 2.1, the relative frequency for the amphibians class is $\frac{552}{2000} = 0.276$

Class	Frequency	Relative frequency
Mammals	202	0.101
Birds	222	0.111
Reptiles	266	0.133
Amphibians	552	0.276
Fishes	468	0.234
Insect	290	0.145
Total	2000	1.000

Figure 1: Table 2.1

Bar Charts

A **Bar chart** is a graphical representation of a frequency distribution for categorical data. An example of a bar chart is shown in Figure 2.5.

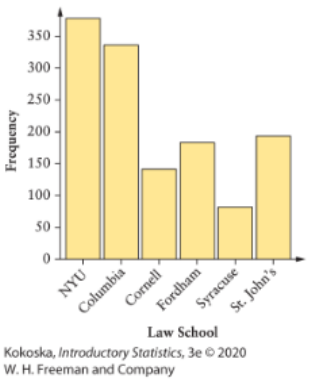

Figure 2.5 Bar chart showing the number of first-time candidates in July 2017 for the bar exam from certain law schools.

Pie Charts

A **pie chart** is another graphical representation of a frequency distribution for categorical data. An example of a pie chart is shown in Figure 2.8

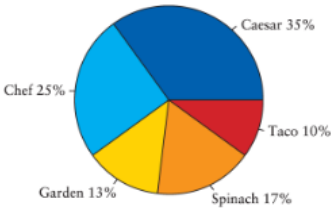

Figure 2.8 Type of salad ordered at the Spice Symphony restaurant.

3 Frequency Distributions and Histograms

Defintion

A **frequency distribution** for numerical data is a summary table that displays classes, frequencies, relative frequencies, and cummulative relative frequencies.

How to Construct a Frequency Distribution for Numerical Data

1. Choose a range of values that captures all the data. Divide it into nonoverlapping (usually equal) intervals. Each interval is called a class, or class interval. The endpoints of each class are the class boundaries.

2. We use the left-endpoint convention. An observation equal to an endpoint is allocated to the class with that value as its lower endpoint. Hence, the lower class boundary is always included in the interval, and the upper class boundary is never included. This ensures that each observation falls into exactly one interval.

3. In practice, there should be 5-20 intervals. Use friendly numbers, for example, $10 - 20$ and $20 - 30$, rather than more complicated categories, such as $15.376 - 18.457$ and $18.457 - 21.538$.

4. Count the number of observations in each class interval. This count is called the class frequency or simply the frequency.

5. Compute the proportion of observations in each class. This ratio, the class frequency divided by the total number of observations, is called the relative frequency.

Histograms

A histogram is a graphical representation of a frequency distrobution, a plot of frequency versus class interval. Given a frequency distrobution, here is a procedure for constructing a histogram.

1. Draw a horizontal line (measurement) axis and place tick marks corresponding to the class boundaries.

2. Draw a vertical axis and place tick marks corresponding to frequency. Label each axis.

3. Draw a rectangle above each class with height equal to frequency.

Note:-

The shape of a distribution may be symmetric or skewed. The center of a distribution referes to the position of the majority of the data, and measures of variability indicate the spread of the data. The ***variability*** (or dispersion) of a distrobution describes how much the measurements vary, as well as how compact or how spread out the data are.

Density Histograms

If the class widths are unequal in a frequency distribution, then neither the frequency nor the relative frequency should be used on the vertical axis of the corresponding histogram. To account for the unequal class widths, set the area of each rectangle equal to the relative frequency. In this case, the height of each rectangle is called the *density*, and it is equal to the relative frequency divided by the class width.

How to find the Density

To find the density for each class:

1. Set the area of each rectangle equal to the relative frequency.

   The area of each rectangle is height times width

$$\begin{aligned}\text{Area of rectangle} &= \text{Relative frequency} \\ &= (\text{ Height }) \times (\text{ Class width })\end{aligned}$$

2. Solve for height.

$$\text{Density} = \text{Height} = (\text{ Relative frequency })/(\text{ Class width })$$

Class	Frequency	Relative frequency	Width	Density	
13–16	407	0.0129	3	0.0043	$(= 0.0129/3)$
16–20	2413	0.0765	4	0.0191	$(= 0.0765/4)$
20–25	4379	0.1389	5	0.0278	$(= 0.1389/5)$
25–30	3789	0.1202	5	0.0240	$(= 0.1202/5)$
30–40	5667	0.1798	10	0.0180	$(= 0.1798/10)$
40–50	4883	0.1549	10	0.0155	$(= 0.1549/10)$
50–60	5628	0.1785	10	0.0179	$(= 0.1785/10)$
60–70	4361	0.1383	10	0.0138	$(= 0.1383/10)$
Total	31527	1.0000			

Figure 2: Table showing density

Shape of Distribution

Because the relative frequency is equal to the area of each rectangle in a density histogram, the sum of the areas of all the rectangles is 1 . This is an important concept as we begin to associate area with probability.

The shape of a distribution, represented in a histogram, is an important characteristic. To help describe the various shapes, we draw a smooth curve along the tops of the rectangles that captures the general nature of the distribution (as shown in Figure 2.25). To help identify and describe distributions quickly, a smoothed histogram is often drawn on a graph without a vertical axis, without any tick marks on the measurement axis, and without any rectangles (as shown in Figure 2.26).

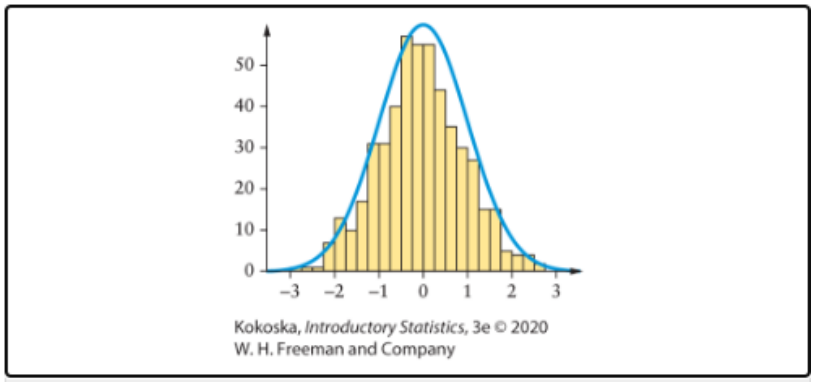

Figure 2.25 Smooth curve that captures the general shape of the distribution.

Figure 2.26 Typical smoothed histogram.

Describing Peaks

1. A **unimodal** distribution has *one peak*. This is very common, as almost all distributions have a single peak.
2. A **Bimodal** distribution has *two* peaks. This shape is not very common but may occur if data from two different populations are accidentally mixed.
3. A **Multimodal** distribution has more than *one* peak. A distribution with more than two distinct peaks is very rare.

Further classification

1. A unimodal distribution is **symmetric** if there is a vertical line of symmetry in the distribution.