Displaying and Summarizing Data I

Sean Hellingman ©

Introduction to Statistical Data Analysis (ADSC1000)

shellingman@tru.ca

Fall 2024

THOMPSON RIVERS UNIVERSITY

Topics

Introduction

- **Statistics** are used, to some capacity, in almost every field.
- Statistics teaches us how to make intelligent judgments and informed decisions in the presence of uncertainty and variation.
- We use statistics in many fields:
    - Politics.
    - Medical sciences.
    - Biology and Ecology.
    - Business insights & decision-making.
    - Sports.

Populations

- Investigations usually focuses on a well-defined collection of objects defining a **population** of interest.
- Some examples of populations:
  - All voting eligible persons.
  - All persons who have contracted influenza in the past two years.
  - All cow moose on the Bonaparte Plateau.
  - All electric cars sold in 2023.
  - All left handed relief pitchers.
- When desired information is available for all objects in the population, we have what is called a **census**.

Samples

- Often very difficult and inefficient to conduct a census
- A **sample** is used to represent the population instead
  - Survey voters as they leave polling station.
  - Administrative data on influenza related hospitalizations.
  - Areal moose sighting data.
  - Targeted surveys with a potential reward for people who have purchased electric cars.
  - Samples of pitchers at different levels.
- There are different types of samples that will be discussed later on.

Goal of Samples

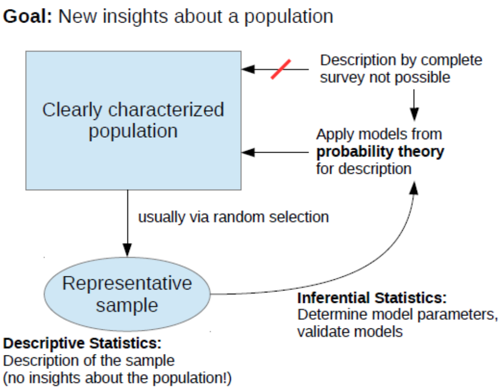


Figure: *source:* (1)

- The goal is to obtain some (new & important) insights about this population.

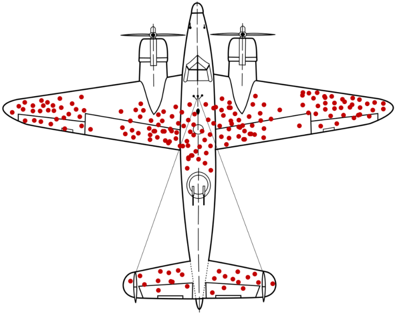
Descriptive Statistics

- The task of **descriptive statistics** is to characterise (describe) the sample
    - Not meant to gain any insights about the population.
    - Important to become acquainted with the data.
    - Examine data quality (very important for inferential statistics).

Inferential Statistics

- The objective of **inferential statistics** is to draw conclusions about the population from the sample
    - Estimate unknown parameters of assumed population distributions.
    - Test the validity of previously made assumptions about the population.

Example 1



- Wald concluded that the *not returning air planes were hit at very vulnerable locations and crashed*.

Scale of Measurement

- When designing a study one should select a variable with the **highest possible** scale of measurement.
  - Birth date is more informative than age.
  - Test scores are more informative than test letter grades.

- It is often not possible to avoid selecting less informative variables.

Notation

- Some important notation used in this course:

Notation	Meaning
N	Number of observations in the population
n	Number of observations in the sample
X (capital letter)	Random variable
x (lower case letter)	Value of a random variable

Stem-and-Leaf Displays

- Procedure:
  1. Select one or more leading digits for the stem values. The trailing digits become the leaves.
  2. List possible stem values in a vertical column.
  3. Record the leaf for every observation beside the corresponding stem value.
  4. Order the leaves from smallest to largest on each line.
  5. Indicate the units for stems and leaves someplace in the display.

- Usually, a display based on between 5 and 20 stems is recommended.

Stem-and-Leaf Displays in R

- The `stem()` function in R takes a *vector* as the input and generates a stem-and-leaf display.
  - The `scale =` argument will split or combine the stems.

- *You may need to adjust the* `scale =` *argument to get a nice number of stems.*

Example 2

- Assume we have a sample of seven ($n = 14$) test scores [0-100]: 93, 84, 86, 78, 95, 81, 72, 92, 87, 86, 79, 99, 81, 52.
  - Write out an appropriate stem-and-leaf plot.
  - Verify this plot using R.

Question

- What kind of information can a stem-and-leaf display give us?

Stem-and-Leaf Information

- Identification of a typical or representative value.
- Extent of spread about the typical value.
- Presence of any gaps in the data.
- Extent of symmetry in the distribution of values.
- Number and location of peaks.
- Presence of any outlying values.

Stem-and-Leaf Information

```
The decimal point is 1 digit(s) to the right of the |

5 | 2
6 |
7 | 289
8 | 114667
9 | 2359
```

- Identification of a typical or representative value: 81 - 87.
- Extent of spread about the typical value: 52 - 99.
- Presence of any gaps in the data: 53 - 72.
- Extent of symmetry in the distribution of values: Negative skew.
- Number and location of peaks: One peak.
- Presence of any outlying values: 52 (maybe).

Histograms

- A **histogram** is an approximate representation of the distribution of numerical data.
- Useful for visualising the number of times an outcome (or range of outcomes) occurs.
- Obtained by splitting the range of a metric variable in consecutive intervals.
- Algorithms automatically select number of equal length intervals.
  - Can often be useful to hand select the interval lengths (bins).
  - Can add empirical densities to the plots.

Histograms in R

- Base R:

```
hist(vector, xlab = "Variable.name", main = "Histogram
of ...")
```

- ggplot:

```
ggplot(data.frame, aes(x=Variable)) + geom_histogram(bins
= 10)+
   xlab("Variable.name")+
   ylab("Frequency")+
   ggtitle("Histogram of ...")
```

- Note: *See example code for other options and specifications.*

Example 3

- Assume we have a sample of fifteen ($n = 15$) test scores [0-100]: 93, 84, 86, 78, 95, 81, 72, 93, 84, 78, 45, 71, 78, 95, 88.
  - Create a histogram for the test scores in R.
  - Adjust the intervals to better understand the *central* part of the data.
  - Add an estimated density to one of the histograms **(remember to add the `prob = TRUE` argument)**.

Histogram Shapes

- A **unimodal** histogram is one that rises to a single peak and then declines.
- A **bimodal** histogram has two different peaks.
- A unimodal histogram is **positively skewed** if the right or upper tail is stretched out compared with the left or lower tail
- A unimodal histogram is **negatively skewed** if the stretching is to the left.

Histogram Shapes

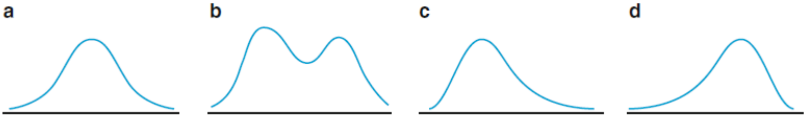


Figure: *source:* (2)

- Smoothed histograms: (a) symmetric unimodal; (b) bimodal; (c) positively skewed; and (d) negatively skewed.

Qualitative Data

- **Qualitative Data**: non-numerical (descriptive) data such as *favourite colour* or *place of birth*.
  - **Bar graphs** (histograms) can also be used to examine qualitative (categorical) data.
  - Sometimes categorical data will have a natural order
    - Highest degree obtained.
  - Other cases it will be arbitrary.
    - Favourite colour.
  - The intervals of the graphs should be of equal lengths.

Bar Graph in R

- Base R:

```
barplot(table(vector),xlab = "Category", ylab =
"Frequency", main = "Barplot for ...")
```

- ggplot:

```
ggplot(data.frame, aes(x=Variable))+
   xlab("Category")+
   ylab("Frequency")+
   ggtitle("Barplot of ...")+
   geom_bar()
```

Example 4

- Assume we have a sample of the favourite colour of ten ($n = 10$) students: red, red, blue, green, green, orange, red, orange, yellow, green.
    - Create a bar graph for the frequency of favourite colours in R.

Boxplots

- A **boxplot** is visual summary that is *resistant* to the to the presence of a few outliers.

- Boxplots may be used to describe the most prominent features of a dataset:
  - The center $\tilde{x}$ (median).
  - The spread, or variability within the data.
  - The extent and nature of any departure from symmetry (skew).
  - Identification of possible outliers.

Boxplot Construction

- Algorithm:
  1. Order the n observations from smallest to largest.
  2. Separate the smallest half from the largest half.
     (If n is odd the median $\tilde{x}$ is included in both halves).
  3. Each of the smallest half and largest halves are split in half again.
     (The upper fourth is the median of the largest half).
  4. The plot is generated from these points.

Boxplots in R

- Base R:

```
boxplot(vector, ylab = "Variable", main = "Boxplot of
...")
```

- ggplot:

```
ggplot(data.frame, aes(y=Variable))+
   ylab("Variable")+
   ggtitle("Boxplot of ...")+
   geom_boxplot()
```

Example 5

- Assume we have the same sample of fifteen ($n = 15$) test scores [0-100]: 93, 84, 86, 78, 95, 81, 72, 93, 84, 78, 45, 71, 78, 95, 88.
  - Create a boxplot for the test scores in R.

Outliers in Boxplots

- By default, potential outliers are shown as dots in boxplots generated by R.
- The width (interquartile range) of the *box* is defined as f_s.
- Any observation that is greater than $1.5f_s$ from the nearest quarter is considered an outlier.
- Any observation that is greater than $3f_s$ from the nearest quarter is considered an extreme outlier.

Question

- When do you think it would be useful to include side-by-side boxplots?

Comparative Boxplots

- An effective way of comparing two or more data sets consisting of observations on the same variable.
- Example:
  - Want to compare data science test scores between mathematics and computer science students

Comparative Boxplots in R

- Base R:

```
boxplot(Variable ~ Group, ylab = "Variable", xlab =
"Group", main = "Comparative Boxplot of ...")
```

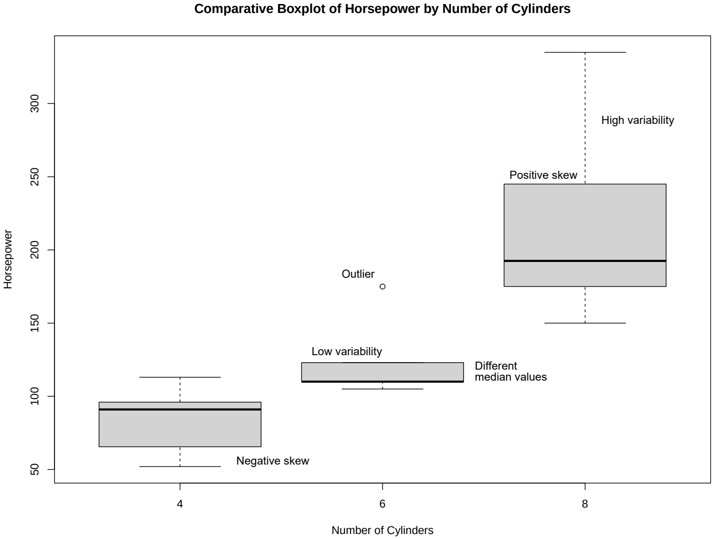
- ggplot:

```
ggplot(data.frame, aes(x = Group, y=Variable))+
   ylab("Variable")+
   xlab("Group")+
   ggtitle("Comparative Boxplot of ...")+
   geom_boxplot()
```

Example 6

- Using the *mtcars* dataset in R:
  - Create a comparitive boxplot for horsepower (hp) separated by the number of cylinders (cyl).
  - What do you notice?

Boxplot Example



Exercise 1

- The 2022/23 Premier League points totals $n = 20$ are 67, 62, 61, 60, 59, 89, 84, 75, 71, 36, 34, 31, 25, 52, 45, 44, 41, 40, 39, 38.
- Generate an appropriate stem-and-leaf plot for this data.
- Repeat this step using R.
- Think about the shape of the distribution of points.

Exercise 2

- Load the *iris* dataset in base R using `data("iris")`.
- Use `?iris` to familiarise yourself with the data.

Exercise 3

- Using the *iris* dataset in R:
    - Generate a histogram for the *Sepal.Length*.
    - Generate a histogram for the *Petal.Length*.
    - Comment on the shapes of the histograms.
    - Make any necessary adjustments to the intervals.

Exercise 4

- Using the *iris* dataset in R:
    - Generate a histogram including the density for the *Sepal.Width*.
    - Comment on the shape of this histogram.

Exercise 5

- Using the *iris* dataset in R:
    - Generate a bar graph for the frequencies of the *Species* variable.
    - What do you think this means about the sample?

Exercise 6

- Using the *iris* dataset in R:
  - Generate a boxplot for the *Sepal.Length*.
  - Generate a boxplot for the *Petal.Length*.
  - Are you able to learn anything different from these plots than the plots you generated in Exercise 3?

Exercise 7

- Using the *iris* dataset in R:
  - Generate a comparative boxplot for the *Sepal.Width* separated by *Species*.
  - Are there any noticeable differences across species?

Exercise 8

- An alternative plot to a boxplot is a violin plot.
- A violin plot combines a boxplot with an estimated density.
- Check out:Violin Plots and play with some violin plots.

References & Resources

1. Kohl, K., (2022). *Introduction to statistical data analysis with R (Second Edition)* Retrieved from https://github.com/stamats/ISDR/blob/main/IntroductionToStatisticalDataAnalysisWithR_ed2.pdf.

2. Devore, J. L., Berk, K. N., & Carlton, M. A. (2012). *Modern mathematical statistics with applications (Second Edition).* New York: Springer.

- Histograms
- Boxplots
- Violin Plots