STA13: Elementary Statistics

Lecture 2

Book Sections 2.1, 2.2, 2.10

Dmitriy Izyumin

January 10 2018

Review

- An experimental unit is an individual or object from which we record some information for the purpose of analysis.
- A variable is a characteristic of an experimental unit. A variable can assume multiple values. The format and number of possible values determines the type of variable.
- A measurement (or an observation) is the recorded value of a variable for one specific experimental unit.
- Data is a set of measurements collected for analysis.

Review

- A variable is a characteristic of an experimental unit. A variable can assume multiple values. The format and number of possible values determines the type of variable.
- Qualitative variables have values that are labels or categories.
- Quantitative variables have values that are numbers or amounts.
  - Discrete variables are quantitative variables that can take on countably many values.
  - Continuous variables are quantitative variables that can take on a continuum of values.

Review

- The population is the set of ALL subjects of interest in a study.
- A sample is a subgroup of the population.

We want to reach conclusions about the whole population, but don't have the resources to analyze the whole population directly. Instead we may select a sample to analyze, and then generalize the findings to the population.

Some things we'll consider:

- How to choose a sample appropriately?
- How accurately do our findings in the sample reflect the properties of the population?
- How to make inferences about the population based on the sample?

Descriptive Statistics

Before performing any statistical analysis, it is useful to explore the data and learn about its features.

Some questions we might consider:

- Are there any trends or patterns in the data?
- How spread out is the data? Is it clustered around certain values?
- If the data is multivariate, how are the variables related?
- Are there any observations that really stand out from the rest?

Descriptive Statistics

Descriptive statistics are tools and methods used for inspecting and summarizing data. The goal of descriptive statistics is to summarize the data in ways that are easy to interpret and showcase the important features of the data.

Descriptive statistics should be the first step of any analysis.

Who's your favorite Beatle?

The Beatles (left to right): Ringo, George, John, Paul

Who's your favorite Beatle?

Suppose I'm interested in learning how the different members rank according to UC Davis students.

I ask a sample of UC Davis students to name their favorite member of the Beatles.

Review:

- What is an experimental unit here?
- What is the variable of interest?
- What are the possible values of this variable?
- What type of variable is it?

How to Summarize the Data?

After surveying the sample I say: "50 people prefer Ringo."

Is this helpful in figuring out how the Beatles are ranked by the students?

There is a lot more information in the statement:
"I surveyed 300 people, and 50 prefer Ringo."

Frequency and Relative Frequency

NOTE: From now on we will denote the sample size by n.

Frequency is the number of observations in the sample that fall into a given category. (50 people prefer Ringo).

Relative Frequency is the fraction of observations that fall into a given category. It is found by dividing the frequency by n. ($\frac{50}{300} = \frac{1}{6} = 0.1667$ of the sample prefers Ringo).

Percentage is the percent of observations that fall into a given category. It is found by converting the relative frequency to a percent value. (16.67% of the sample prefers Ringo).

Frequency Tables

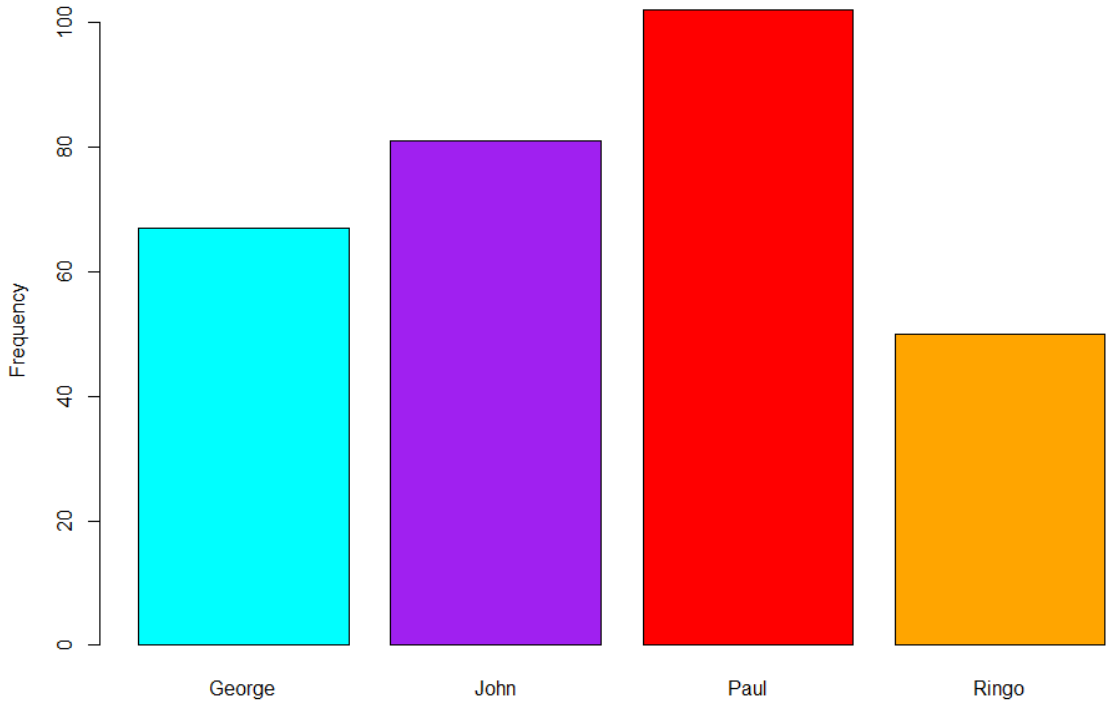
Name	Frequency	Relative Frequency
George	67	$\frac{67}{300} = 0.2233$
John	81	$\frac{81}{300} = 0.27$
Paul	102	$\frac{102}{300} = 0.34$
Ringo	50	$\frac{50}{300} = 0.1667$

Checking that you did it right:

- Frequency column has to add up to n
- Relative frequency column has to add up to 1

Bar Charts

The relative frequencies summarize the data pretty well, but sometimes it's better to have a visual representation, like a bar chart:



Bar Chart Details

- A Pareto diagram is a bar chart with columns arranged from tallest to shortest.
- You may use frequencies instead of relative frequencies (as I did). This will affect the scale of the y-axis, but not the shape of the plot.
- If using frequency, then the heights of the bars will add up to n.
- If using relative frequency, then the heights of the bars will add up to 1.

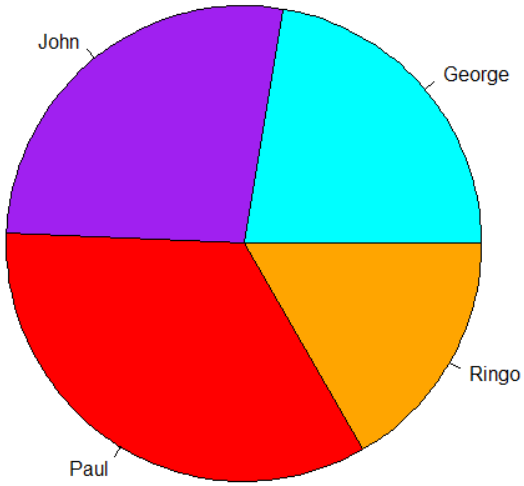
Making a Bar Chart

Steps for making a bar chart:

1. Calculate the (relative) frequencies.
2. Make notches on the x-axis to mark where the bars will go. The bars should have equal width.
3. Draw the bars. The heights are equal to the (relative) frequencies.
4. Label the bars.
5. Color. (optional)

Pie Charts

We could also summarize the same data with a pie chart:



Here the central angles (or the sector areas) reflect the relative frequencies.

Pie Charts

Steps for making a pie chart:

1. Calculate the relative frequencies
2. Calculate the central angles (multiply relative frequencies by 360).
3. Draw the sectors.
4. Label the categories.
5. Color. (optional)

Which is Better?

- Pie charts have no advantage.
- Bar charts are easier to read when there are many categories.
- Bar charts are easier to read when two or more categories have similar frequencies.

Misleading Graphs - What Not to Do

The following makes a difference between categories in a bar chart seem more drastic than it really is:

- Changing the width of the bars proportionally to height
- Starting the y-axis at a number other than 0
- "Stretching" the y-axis

Misleading Graphs - What Not to Do

Quantitative variables - What is Different?

Suppose I record the heights (in inches) of everyone in this class.

Review:

- What is an experimental unit here?
- What is the variable of interest?
- What are the possible values of this variable?
- What type of variable is it?

Can I make a bar chart to summarize the data? What would the x-axis labels be?

Histogram

In the case of qualitative data, we had classes (categories) already available. In the case of quantitative data, we just have a spectrum of numbers.

We can handle this by creating our own classes.

We can divide the observed range of data into chunks, treat these chunks as classes, and make a relative frequency bar chart from these classes. This plot is called a histogram.

Making a Histogram

Steps for making a histogram:

1. Calculate the range = max-min
2. Decide on the number of classes
3. Calculate the class width $= \frac{\text{range}}{\#\text{ of classes}}$
4. If the class width is inconvenient, choose a convenient class width close to what you obtained in step 3
5. Construct a frequency table using the left inclusion principle (this means that each class is of the form $a \leq x < b$ for some a, b)
6. Make a relative frequency bar chart

Classes

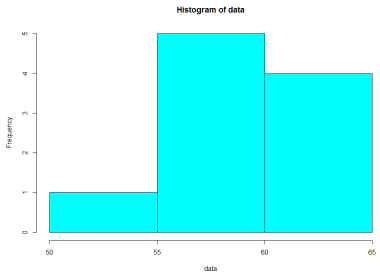
Rules for choosing classes:

- all classes have the same width
- the bars have to touch (not true for bar charts)
- each observation has to fall into a class
- the first and last classes can't be empty
- left inclusion principle - each class includes the left boundary, but not the right

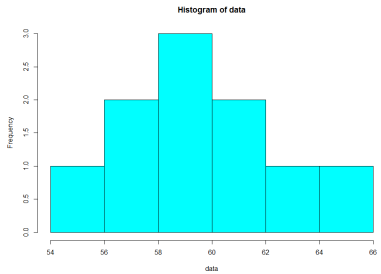
Classes

Data: 57 60 61 62 65 59 64 58 54 59

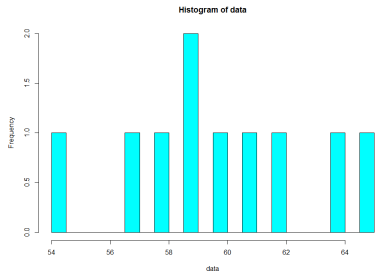
3 classes



6 classes



21 classes



Here 3 classes are too few, and 21 are too many. The middle picture gives a better overall summary of the data.

Reading a Histogram

In a frequency histogram the height of a bar equals the number of observations falling into that class.

In a relative frequency histogram, the height of a bar equals

- the proportion of observations falling into that class.
- the probability that a randomly selected measurement from the dataset will fall in that class.

The shape of the histogram is important as well.

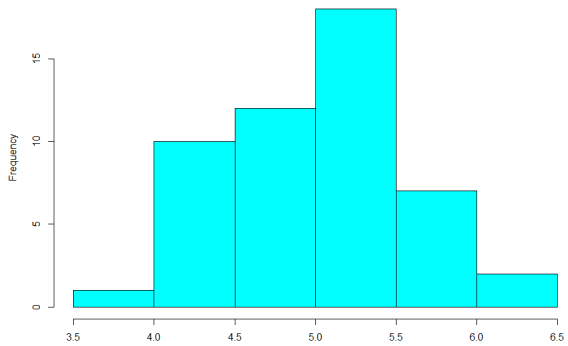
Modes

Modes are noticeable peaks in the distribution of the observations. They tell us where the observations are concentrated.

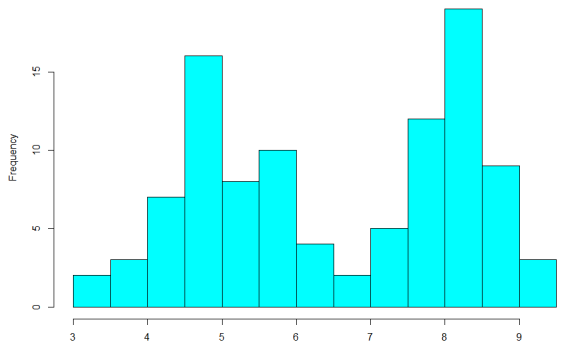
Unimodal data has one distinct peak.

Bimodal data has two distinct peaks.

Unimodal



Bimodal



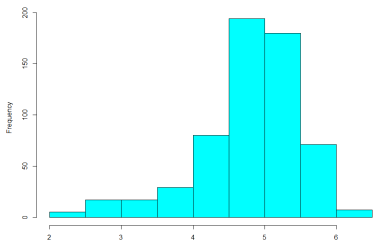
Symmetry and Skew

We say the histogram is symmetric if the right and left sides roughly mirror each other.

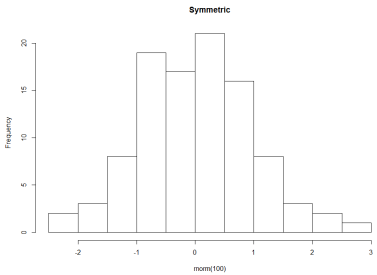
We say the histogram is skewed left if the left tail is noticeably longer than the right. Most of the observations are concentrated to the right, but there are a few unusually small values.

We say the histogram is skewed right if the right tail is noticeably longer than the left. Most of the observations are concentrated to the left, but there are a few unusually large values.

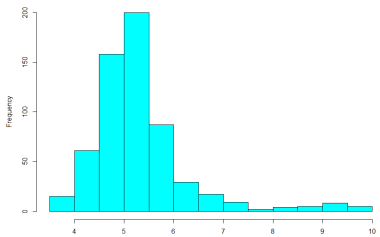
Skewed Left



Symmetric

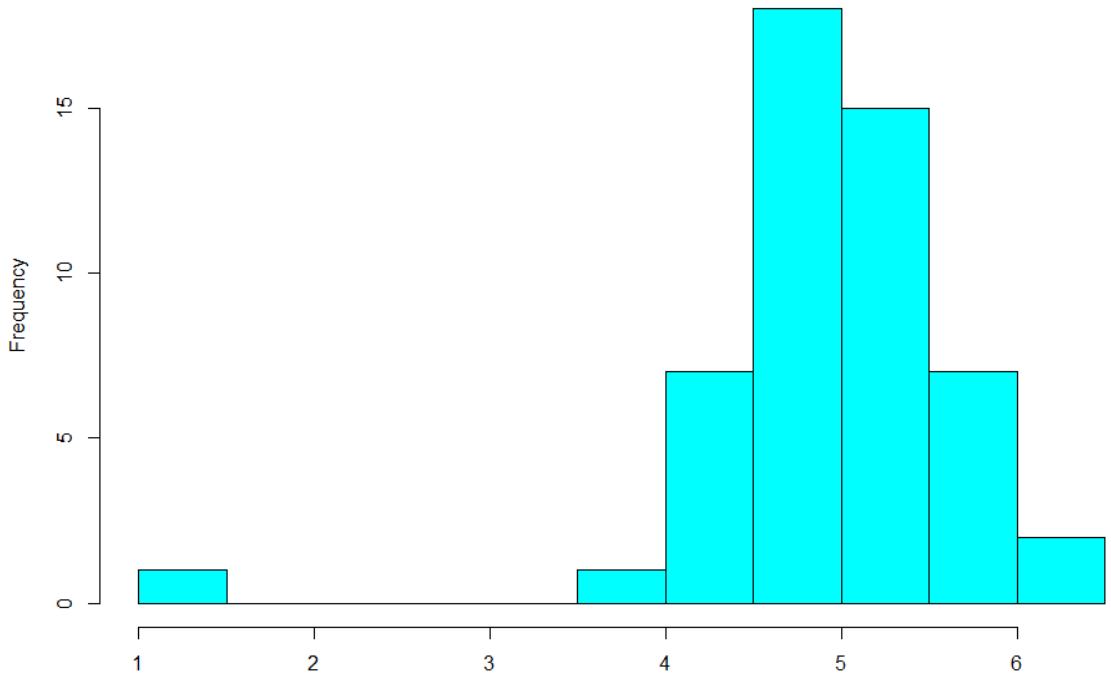


Skewed Right



Outliers

Outliers are strange or extreme measurements that stand out from the rest of the data set.



We will talk more about outliers later.