Ch. 1 – Descriptive Statistics and Overview

Syllabus

- Course structure
- Office hours
- Materials and canvas
- Lab requirement, **you need to pass both for credit**
- Grading, homework, quizzes, exams
- Calendar

Lecture Style

- Defintion(s)/Method/Explanation(s) $\rightarrow$ example(s) $\rightarrow$ comments *and* questions
- **Definitions** are in bold.
- *Math jargon* is italicized. You don't need to know these terms in this , but they frequently show up in other math work.
- Color coding will be used to highlight different ingredients of a method. Will be dropped after the first one or two examples.
- Sections from the book will be written in the title of the first relevant slide.

Lecture Outline, 8/12

- 1.1 Statistical nomenclature
- 1.1 Sampling methods
- 1.2 Pictorial and tabular methods in descriptive statistics

Statistical nomenclature, Section 1.1

- **Population**: group of interest.
- **Data**: collection of facts, usually about the population.
- **Quantitative data**: data which are numbers. Example: length, weight, volume and etc.
- **Quantitative data**: data which are categories. Example: open or close; color; dead, alive, dead, or undead; and etc.
- **Census**: Data of the entire population
- **Sample**: Data of a subset of the population. Most data is sample data. Sampling means to generate a sample.
- Sample data is compromised of observations consisting of variables, quantities/characteristics of interest.
- **Univariate** (data): Observations measure one variable
- **Multivariate** (data): Observations measure more than one variable(s)
- **Bivariate** (data): Observations measure exactly two variables

Statistical nomenclature

- **Enumerative Studies**: Studies of an existing fixed, finite population. Example, meauring plant heights or coin flips.
- **Analytic Studies**: Studies of a process which may not exist. Example, developing methods for measuring variance of plant heights or coin flips.

Sampling Methods

- **Bias** (statistical): Source of incorrect measurement of population characteristics. Example, sampling only NBA players to determine the average height of a US citizen.
- **Simple Random Sampling** (SRS): samlping method were each member of the population of interest is eligible to be randomly selected to be included in the sample.
- **Strata**: Partitions of a population based on similar traits. Example, grouping dogs based on fur color; hair color is the strata.
- **Stratified Sampling**: sampling method were the population is divided into observable strata. Used to avoid under-representation.
- **Convenience Sampling**: Sampling which to prioritizes convenience, usually not random and has bias. Example, poll only our classmates when the population of interest is all students at the U.

Recurring Definitions

- Quantitative data can be categorized into discrete and continuous data.
- **discrete**: Number of observable or possible values of data are countable or finite. Example, number of days in a month, number of times a coin lands on heads in 50 flips, heights of students rounded to the nearest inch, any integer and etc.
- **continuous**: Observable or possible values consist of entire intervals. Example, heights unrounded, weights of shipping containers, any number between 0 and 1 and etc. On a computer all data is discrete, but if each number appears once, the data should be treated as continuous.

Recurring Definitions

- **sample size**: number of observations in a sample.
- population, sample, sample size, data, categorical and quantitative variables (discrete and continuous), observations will appear throughout.
- **Probability**: The field of mathematics that describes the behavior of objects in the presence (or lack of) uncertainty or randomness.
- **Distribution**: What values a variable takes and how frequently it takes them. Can express the distribution of data in plots or model the dsitribution of probabilistic objects as a *generalized* function, called a distribution.

Pictorial and Tabular Methods in Descriptive Statistics, Section 1.2

Given quantitative data (should be discrete), the distribution can be described using the following:

- stem-and-leaf plot
- dot plot
- frequency distribution
- histogram

Stem-and-Leaf Plot, Method

Construction

1. Select the leading digits to be stem digits and remianing digits to be leaf digits.
2. Draw a vertical line with the stem digits on the left.
3. Record the corresponding leaf values to right of the line.
4. Indicate the units of the stem and leaf digits.

Stem-and-Leaf Plot, Example

Data: number of heads in 100 coin flips

$$63, 47, 52, 54, 50, 57, 39, 41, 49, 49, 50, 52, 55, 51, 56, 53, 47, 49, 50, 62$$

Sample size $= 20$
discrete quantitative data

Stem-and-Leaf Plot, Example

Step 1: Select the leading digits to be stem digits and remianing digits to be leaf digits.

$63, 47, 52, 54, 50, 57, 39, 41, 49, 49, 50, 52, 55, 51, 56, 53, 54, 49, 50, 62,$

Stem-and-Leaf Plot, Example

Step 2: Record the corresponding leaf values to right of the line.

$63, 47, 52, 54, 50, 57, 39, 41, 49, 49, 50, 52, 55, 51, 56, 53, 54, 49, 50, 62$

6	
5	
4	
3	

Stem-and-Leaf Plot, Example

Step 3: Draw a vertical line with the stem digits on the left.

$63, 47, 52, 54, 50, 57, 39, 41, 49, 49, 50, 52, 55, 51, 56, 53, 54, 49, 50, 62$

6	23
5	000122344567
4	17999
3	9

Stem-and-Leaf Plot, Example

Step 4: Indicate the units of the stem and leaf digits.

$63, 47, 52, 54, 50, 57, 39, 41, 49, 49, 50, 52, 55, 51, 56, 53, 54, 49, 50, 62$

6	23
5	000122344567
4	17999
3	9

stem units: tens

leaf units: ones

Stem-and-Leaf Plot, Example

Step 4: Indicate the units of the stem and leaf digits.

$63, 47, 52, 54, 50, 57, 39, 41, 49, 49, 50, 52, 55, 51, 56, 53, 54, 49, 50, 62$

6	23
5	000122344567
4	17999
3	9

stem units: tens

leaf units: ones

Stem-and-Leaf Plot, Second Example

Data: proportion of water in different drinks, rounded to nearest hundredths.

$$0.73, 0.79, 0.61, 0.06, 0.41, 0.26, 0.90, 0.07, 0.52, 0.93, 0.39, 0.72$$

0.9	0 3
0.8	
0.7	2 3 9
0.6	1
0.5	2
0.4	1
0.3	9
0.2	6
0.1	
0.0	6 7

stem units: tenths

leaf units: hundredths

Stem-and-Leaf Plot, Comments/Questions

- Shows visual shape of of the distribution (of data).
- Displays observed values
- Usually better for discrete data. Truly continuous will usually display one observation at each value.

Questions?

Stem-and-Leaf Plot, Comments and Questions

- Shows visual shape of of the distribution
- Displays observed values
- Usually better for discrete data. Truly continuous will usually display one observation at each value.

Questions?

Stem-and-Leaf Plot, Comments/Questions

- Shows visual shape of of the distribution
- Displays observed values
- Usually better for discrete data. Truly continuous will usually display one observation at each value.

Questions?

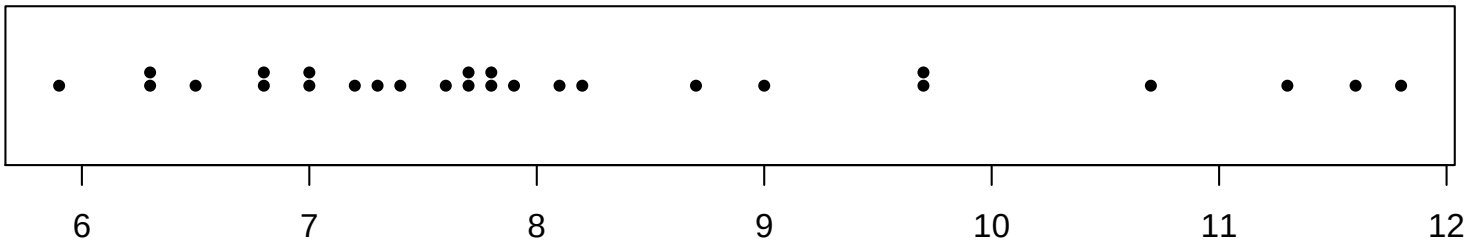
Dot Plot, Explanation and Example

Dot plots are simliar to stem-and-leaf plots, except they have a hortizonal bar and there are dots instead of leaf digits.

Example Data: Concrete flexural strength (in MPa) rounded to nearest tenth. Discrete quantitative data.

$$5.9, 6.3, 6.3, 6.5, 6.8, 6.8, 7.0, 7.0, 7.2, 7.3, 7.4, 7.6, 7.7, 7.7,$$

$$7.8, 7.8, 7.9, 8.1, 8.2, 8.7, 9.0, 9.7, 9.7, 10.7, 11.3, 11.6, 11.8$$



Dot Plot, Comments and Questions

- Only shows the visual shape of the distribution (of data), values not quickly readable.
- Should be used with discrete data.
- Less informative than stem-and-leaf plots, but more visually pleasing.

Questions?

Frequency Distribution, Definitions

- **Frequency**: Number of times a value is observed in a data set. Should be used with discrete data; continuous data usually reports each number once.
- **Relative Frequency**: Frequency of a value divide by the sample size.
- **Frequency Distribution**: Table of frequencies or relative frequencies at each observed value.

Frequency Distribution, Example

Data: Value of a dice roll, 20 rolls. Quantitative, discrete data.

$$4, 4, 3, 3, 2, 4, 6, 3, 6, 6$$

$$1, 2, 6, 3, 1, 3, 5, 5, 2, 2$$

$$\text{frequency of } 4 = 3 \qquad \text{sample size} = 20$$

$$\text{relative frequency of } 4 = \frac{\text{frequency of } 4}{\text{sample size}} = \frac{3}{20} = 0.15$$

dice roll	1	2	3	4	5	6
frequency	2	4	5	3	2	4
relative frequency	0.10	0.20	0.25	0.15	0.10	0.20

Frequency Distribution, Comments and Questions

Works well with discrete data.

Questions?

Histogram, Method

Histograms are appropriately visually quantitative, continuous data. Moreover, they can be used with discrete data as well.

bins: Intervals, usually equal length and consecutive. Most commonly, the left endpoint is excluded and the right endpoint is included.

1. On a a segment of the number line, draw bins so that your each datum lies within a bins.
2. Within each bin draw a bar extending to the number of data observed in the bin, or the frequency of obervations in the bin.

Histogram, Example

Data: Energy consumption of homes (BTU). Quantitative, continuous data. (Technically, the data is rounded, but each number only appears roughly once so it can be thought of as continuous.)

2.97, 4.00, 5.20, 5.56, 5.94, 5.98, 6.35, 6.62, 6.72, 6.78, 6.80, 6.85, 6.94, 7.15, 7.16, 7.23, 7.39, 7.62, 7.62, 7.69, 7.73, 7.87, 7.93, 8.00, 8.26, 8.29, 8.37, 8.47, 8.56, 8.58, 8.61, 8.67, 8.69, 8.81, 9.07, 9.27, 9.37, 9.43, 9.52, 9.58, 9.60, 9.76, 9.82, 9.83, 9.83, 9.84, 9.96, 10.04, 10.21, 10.28, 10.28, 10.30, 10.35, 10.36, 10.40, 10.49, 10.50, 10.64, 10.95, 11.09, 11.12, 11.21, 11.29, 11.43, 11.62, 11.70, 11.70, 12.16, 12.19, 12.28, 12.31, 12.62, 12.69, 12.71, 12.91, 12.92, 13.11, 13.38, 13.42, 13.43, 13.47, 13.60, 13.96, 14.24, 14.35, 15.12, 15.24, 16.06, 16.90, 18.26

Histogram, Example

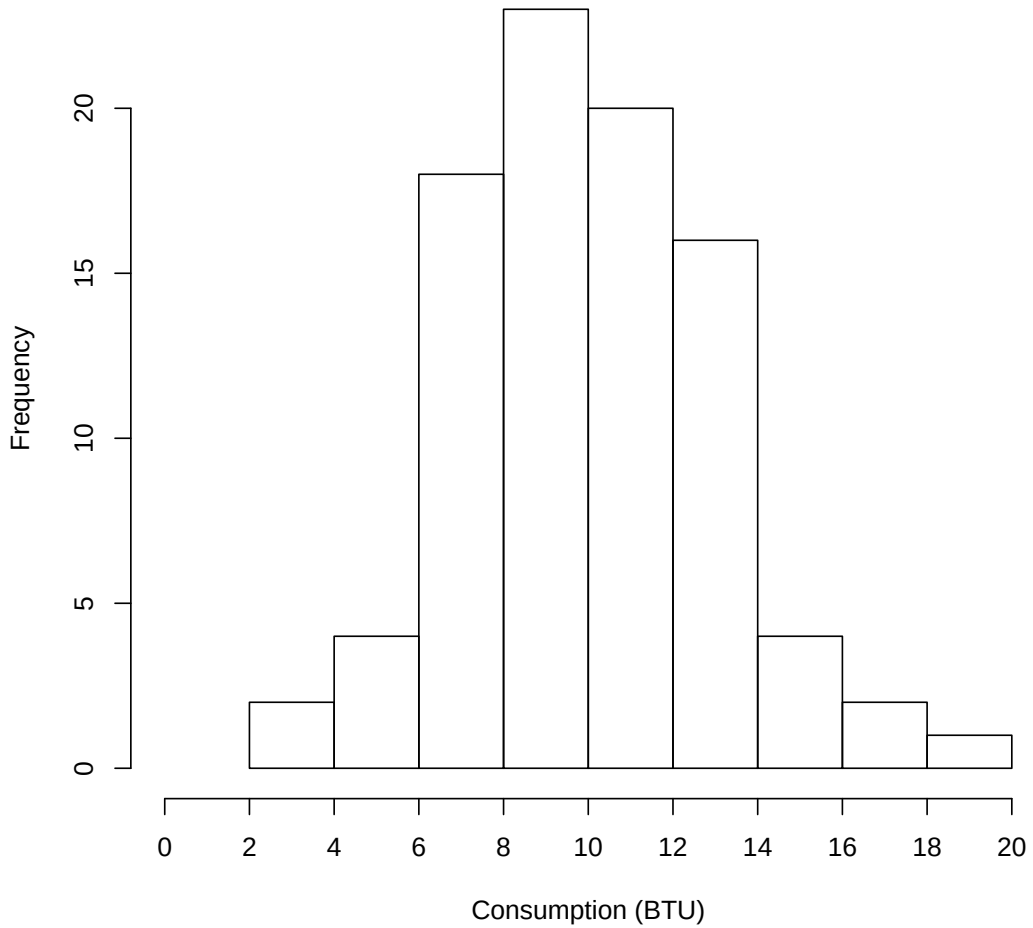
Step 1: On a a segment of the number line, draw bins so that your each datum lies within a bin.

The following interval contain all following bins:

$(0, 2]$, $(2, 4]$, $(4, 6]$, $(6, 8]$, $(8, 10]$, $(10, 12]$, $(12, 14]$, $(14, 16]$, $(16, 18]$, $(18, 20]$

Histogram, Example

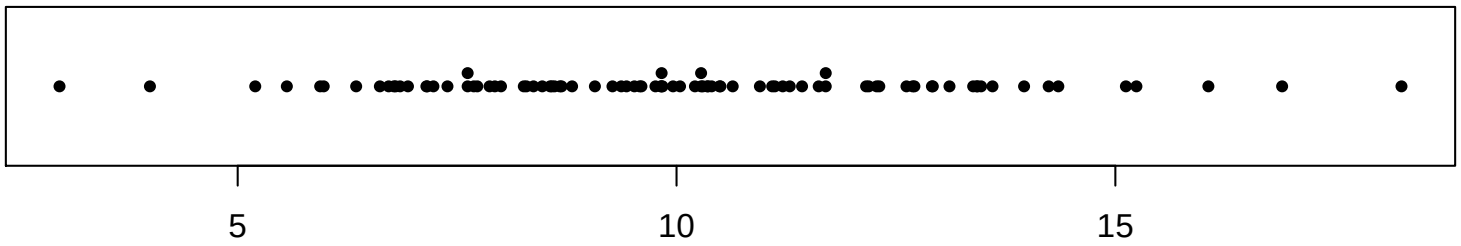
Step 2: Within each bin draw a bar extending to the number of data observed in the bin, or the frequency of obervations in the bin.



Histogram, Comments and Questions

- Can be used with continuous and discrete quantitative data.
- Shows shape of distribution.

What if we used a dot plot?



Hard to see the shape of the data, no bins.

Questions?

Visualizing Categorical Data

- bar plots, bars show frequency of observations
- frequency or relative frequency tables