Chapter 2: Tools for Exploring Univariate Data

Section 2.1: Introduction

What is Data?

- Data are numerical facts.
- A set of data contains information about *individuals*.
- Information is organized in *variables*.
- A variable is a property of an individual (e.g., age, gender, ...).

Definitions

- **Individuals** are the objects described by a set of data. Individuals may be people, but they may also be animals or things.
- A **variable** is any characteristic of an individual. A variable can take different values for different individuals.

Next we want to distinguish between the different types of variables - different types of variables provide different kinds of information and the type will guide what kinds of summaries (graphs/numerical) are appropriate.

Think about it:

- Could I compute the "AVERAGE AGE" for all UCLA students? _____
- Could I compute the "AVERAGE GENDER" for all UCLA students? _____

Gender is said to be a _____ variable,

Age is a _____ variable.

Types of Variables

- **Quantitative** variables are measurements and counts.
  - Variables with few repeated values are treated as *continuous*.
  - Variables with many repeated values are treated as *discrete*.
- **Qualitative** variables describe group membership.

  - *Categorical* variables are _____
  - *Ordinal* variables are _____

(See Figure 2.1.1 on page 42.)

Qualitative or Quantitative?

- Hair Color: _____
- Salary: _____
- Exam grade (0–100): _____
- Letter grade (A, B, C, etc.): _____
- Weight: _____
- Color preference (1=Red, 2=Blue, 3=Green): _____

Section 2.2: Presentation of Data

Some useful roles for presenting data efficiently.

- Make it simple for presentation, e.g., rounding numbers.
- Maintain complete accuracy in calculation to avoid rounding errors.
- Keep full detail for reference.

Section 2.3: Simple Plots for Continuous Variables

- Dot plots
- Stem-and-leaf plots
- Histograms

A good picture is worth more than a thousand words!

What is distribution?

- The **distribution** of a variable tells us what values it takes and how often it takes these values.

Dot plots plot a batch of numbers on a scale. Good for small to moderate batches of data.

Example: make a dot plot of the 15 numbers:

54 59 35 41 46 25 47 60 54 46 49 46 41 34 22

Stem-and-leaf plots (also called stem plots) show a quick picture of the shape of a distribution. Good for plotting 15–150 data points.

Example: Make a stem plot of the 15 numbers:

54 59 35 41 46 25 47 60 54 46 49 46 41 34 22

Separate each observation into a **stem** and a **leaf**. A leaf often is a single digit.

For example, for the first observation, 54, the stem is _____ , and the leaf is _____ .

Procedure: (a) Write the stems (b) Write the leaves (c) Sort the leaves (d) Write the units.

Comments on Stemplot:

- It is possible to re-construct the original data set.
- It works well for small numbers of observations, but not so well for large data sets.
- Software can make complicated stem-and-leave plots.
- A histogram does a similar job and is preferred for large data sets.

Histogram is a good graph for quantitative variables with a large number of observations.

Example: The data show the money (in dollars) that 50 shoppers spent at a supermarket. The data are sorted.

```
 3.11,  8.88,  9.26, 10.81, 12.69, 13.78, 15.23, 15.62, 17.00, 17.39,
18.36, 18.43, 19.27, 19.50, 19.54, 20.16, 20.59, 22.22, 23.04, 24.47,
24.58, 25.13, 26.24, 26.26, 27.65, 28.06, 28.08, 28.38, 32.03, 34.98,
36.37, 38.64, 39.16, 41.02, 42.97, 44.08, 44.67, 45.40, 46.69, 48.65,
50.39, 52.75, 54.80, 59.07, 61.22, 70.32, 82.70, 85.76, 86.37, 93.34.
```

To make a histogram of the distribution, proceed as follows:

1. Find the range of the values: min=_____ , max=_____

2. Divide the range of the data into classes (i.e., intervals) of equal width. What is reasonable here?

   _____ classes: _____ .

3. Count the number of observations in each class. These counts are called **frequencies**.

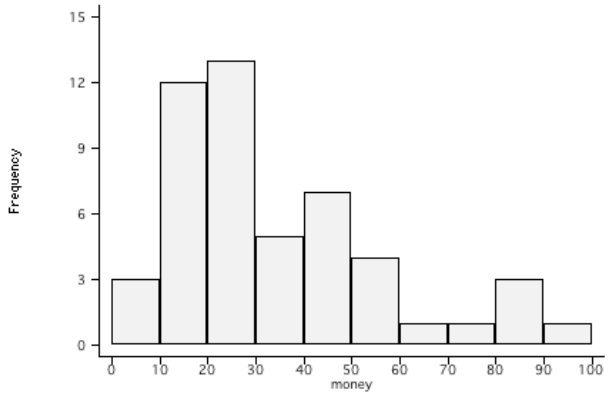
A Frequency Table

Class	Frequency (Count)	Relative Frequency	Percent
$[0, 10)$	3	$3/50 = .06$	6
$[10, 20)$	12	$12/50 = .24$	24
$[20, 30)$	13	.26	26
$[30, 40)$	5	.10	10
$[40, 50)$	7	.14	14
$[50, 60)$	4	.08	8
$[60, 70)$	1	.02	2
$[70, 80)$	1	.02	2
$[80, 90)$			
$[90, 100)$	1	.02	2
Total	50	1.00	100

   Note: Relative Frequency = Frequency/Total number of observations.

4. Draw the histogram.

Stata Histogram



Comments on Histogram:

- The vertical axis can be frequency (count) or relative frequency or percent.
- Each bar represents a class, the base of the bar covers the class.
- No horizontal space between bars.
- Be careful about the boundaries: inclusive or exclusive?
- How many classes? Use your judgment _____
- Software has defaults. Stata uses 5 classes by default.

Interpreting stem-and-leaf plots and histograms

- Look for the **overall pattern** and for striking **deviations** from that pattern.
- Describe the overall pattern of a distribution by its **shape**, **center**, and **spread**.
- A striking deviation is an **outlier**, an individual value that falls outside the overall pattern.

We will learn how to describe center and spread numerically in Section 2.4.

Some things to look for in describing shape are:

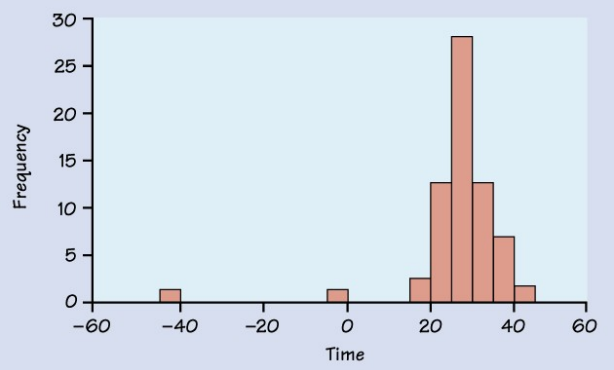
- Does the distribution have one or several major peaks, called **modes**? A distribution with one major peak is called **unimodal**.
- Is it *approximately* **symmetric** or is it **skewed** in one direction? A distribution is **positively skewed** if the right tail (larger values) is much longer than the left tail (smaller values).

Now, how to interpret the histogram of the money spent by 50 shoppers?

- Look for the **overall pattern**:

- Look for striking **deviations** from the overall pattern:

Example: Histogram of Simon Newcomb's 66 measurements of the passage time of light. The values are the deviations from 24,800 nanoseconds (a nanosecond = 10^{-9} seconds).



- Look for the **overall pattern**:

- Look for striking **deviations** from the overall pattern:

Section 2.4: Numerical Summaries for Continuous Variables

In this section we will focus on numerical summaries of the center and the spread of the distribution (appropriate for quantitative data only!).

Two Measures of Center

- sample mean $\bar{x}$ — the average value
- sample median M or Med — the middle value

Measuring Center: The Sample Mean

Notation

- n= sample size (i.e., # of observations)
- observations: $x_1, x_2, ..., x_n$

The **sample mean** $\bar{x}$ is the average value.

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

or, in more compact notation,

$$\bar{x} = \frac{1}{n}\sum x_i$$

Example: Golf scores of 12 members of a women's golf team in tournament play.

89 90 87 95 86 81 102 97 83 88 91 79

The mean is _____

Measuring Center: The Sample Median

The **sample median** M is the middle value of a distribution, the number such that half the observations are smaller and the other half are larger.

Steps to find the sample median M

1. Arrange all observations in increasing order.
2. If n is odd, M is the center observation in the ordered list.
3. If n is even, M is the average of the two center observations in the ordered list.

Compute the median for the Golf scores.

Ordered data: 79 81 83 86 87 88 89 90 91 95 97 102

The median M= _____ .

Mean versus Median

Example Golf Scores Data: 89 90 87 95 86 81 102 97 83 88 91 79

What if the worst player's score was incorrectly entered as 201 instead of 102?

The mean would be _____

The median would be _____

Comments on Mean and Median

1. The **mean** is _____ to extreme observations. The **median** is _____ to extreme observations.
2. The mean and median of a symmetric distribution are _____ together.
3. If a distribution is negatively skewed (long tail on the left), the mean is _____ than the median.

Center does not tell the entire story. Most useful description of data is a measure of center and also the **spread**.

Measures of Spread

1. Range
2. Quartiles
3. Interquartile Range
4. Standard Deviation

Note:

- The **Range** = Maximum Value - Minimum Value. It is _____ to extreme observations.

Measuring Spread: the Quartiles

Percentiles: The p-th percentile is the value such that p% of the observations fall at or below that value.

Some Common percentiles:

- Median M = 50-th percentile ($p = 50$).
- First quartile Q_1 = 25-th percentile ($p = 25$).

- Third quartile Q_3 = 75-th percentile ($p = 75$).

NOTE

1. First quartile Q_1 = the median of observations to the _____ of the overall median M.
2. Third quartile Q_3 = the median of observations to the _____ of the overall median M.

Example Golf Score Data

Ordered data: 79 81 83 86 87 88 89 90 91 95 97 102

The median M = _____ , the first quartile Q_1 = _____ , the third quartile Q_3 = _____

If we have only 11 observations:

79 81 83 86 87 88 89 90 91 95 97

The median M = _____ , the first quartile Q_1 = _____ , the third quartile Q_3 = _____

NOTE:

1. Many other textbooks/packages used different formula for Q_1 and Q_3.
2. For other p-th percentiles, just compute the position at $\frac{n*p}{100}$ and round it.
3. Applying above calculation for $p = 25, 75$ may yield slightly different answer.
4. If you use software, just use the values that your software gives you.

Measuring the Spread: the Interquartile Range

The **Interquartile Range**, IQR, is the distance between the first and third quartiles,

$$IQR = Q_3 - Q_1$$

The $1.5 \times IQR$ Criterion for Outliers:

- An observation is a suspected outlier if it falls more than $1.5 \times IQR$ above the third quartile or below the first quartile.

Mathematically, x is a suspected outlier if

$$x < Q_1 - 1.5 * IQR \text{ or } x > Q_3 + 1.5 * IQR.$$

The five-number summary and boxplots

The five-number summary is

Minimum, Q_1, M, Q_3, Maximum

Try it! Golf Score Data

Ordered data: 79 81 83 86 87 88 89 90 91 95 97 102

The five-number summary is

Boxplots

A boxplot is a graphical representation of the five-number summary.

Steps:

- Make a box with ends at the quartiles Q_1 and Q_3
- Draw a line in the box at the median
- Check for possible outliers using the $1.5 * IQR$ rule and if any, plot them individually
- Extend lines from end of box to smallest and largest observations that are not possible outliers

Try it! Golf Score Data

Ordered data: 79 81 83 86 87 88 89 90 91 95 97 102

Recall the **five-number summary** is:

$IQR =$

$1.5 * IQR =$

$Q_1 - 1.5 * IQR =$

$Q_3 + 1.5 * IQR =$

Sketch the basic boxplot:

Suppose the worst golf score of 102 was really 107.

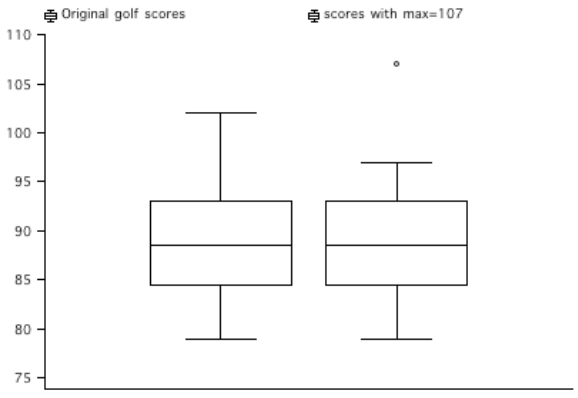
Ordered data: 79 81 83 86 87 88 89 90 91 95 97 107

Then the five number summary would be:

The IQR and 1.5*IQR would be the same, so the "boundaries" for checking for possible outliers are again _____ and _____ . Now we would have one potential high outlier, the maximum value of 107.

Sketch the modified boxplot if we have this one outlier.

Here are the boxplots for the golf score data generated by Stata.



Notes on Boxplots:

- Side-by-side boxplots are good for _____
- Watch out – points plotted individually are _____
- Can't comment on the _____ of the distribution (see Figure 2.4.6 on page 74).

Measuring the spread: the standard deviation

When the mean is used to measure center, the most common measure of spread is the standard deviation. The standard deviation is a **measure of the spread of the observations from the mean**. We will refer to it as a kind of "average distance" of the observations from the mean. We like to **think of the standard deviation as "roughly, the average distance of the observations from the mean."**

Definitions:

- **variance** s^2

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$

- **standard deviation** s is the square root of the variance s^2,

$$s = \sqrt{s^2} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}}$$

Equivalently,

$$s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$$

$$s = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$$

Try It! Golf Score Data:

89 90 87 95 86 81 102 97 83 88 91 79

The mean was computed earlier to be $\bar{x} = 89$. **Find the standard deviation** for this data.

$s =$ _____

Not much fun to do it by hand. Is ok for small number of observations, but in general, we would have a calculator or computer do it for us.

Interpretation: These golf scores are *roughly* _____ away from their mean score of _____ *on average.*

The 68% – 95% Rule

For many unimodal and bell-shaped distributions, approximately

- 68% of the data lie within 1 standard deviation of the sample mean $\bar{x}$
- 95% of the data lie within 2 standard deviation of the sample mean $\bar{x}$

Notes about the standard deviation:

- The standard deviation s is in the **original units** while the variance s^2 is not.
- s= 0 means _____
- Like the mean, standard deviation s is _____
- So use the mean and standard deviation for _____
- and the five-number summary is better for _____
- Watch out if you use a calculator _____
- The number _____ is called **degrees of freedom** of the standard deviation.

Changing the unit of measurement

A linear transformation

$$x_{new} = a + bx$$

Effect of linear transformation

- measures of center: mean ($\bar{x}$) and median (M)

- measures of spread: standard deviation (s) and interquartile range (IQR)

Note: linear transformation do NOT change the shape of a distribution.

Example: Seven day temperature forecast for L.A.

Degrees in Fahrenheit	70	72	74	75	76	76	77
Degrees in Celsius	21.1	22.2	23.3	23.9	24.4	24.4	?

Note: C = $\frac{5}{9}$ (F - 32)

(a) What are the mean, median, IQR and s.d. in Degrees in Fahrenheit?

(b) What are the mean, median, IQR and s.d. in Degrees in Celsius?

Can you tell the relationship of the answers in (a) and (b)?

Section 2.5 Repeated and Grouped Data

Frequency tables and bar graphs (see the next section) are useful in describing

- repeated data (i.e., discrete variable)
- grouped data (continuous variable)

Section 2.6 Qualitative Variables

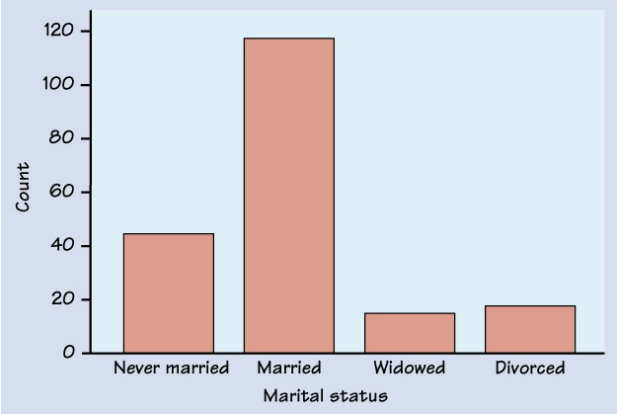
The distribution of a categorical variable lists the categories and the count or proportion or percent of items that fall into each category.

Example: The distribution of marital status for all Americans age 18 and over in 1995 is

Marital status	Count(millions)	Percent
Never married	43.9	22.9
Married	116.7	60.9
Widowed	13.4	7.0
Divorced	17.6	9.2
Total	191.6	100

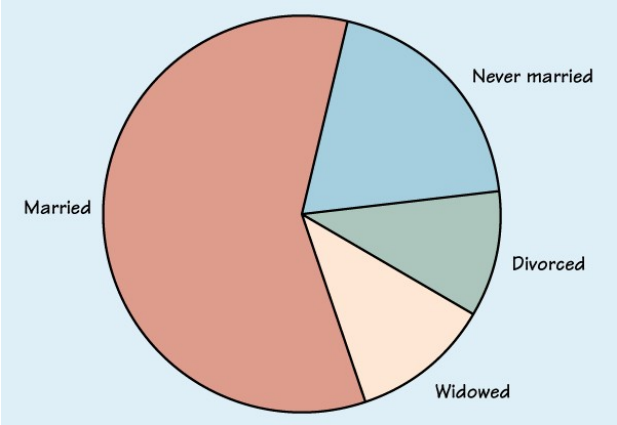
A graphical display of this distribution can be a bar graph or a pie chart.

Bar Graph helps us quickly compare the sizes of the different groups.



Note: The vertical axis of a bar graph can be either the count (frequency) or the percent (relative frequency).

Pie Chart helps us see what part of the whole each group forms.



Note: It is not as easy to draw by hand. Not as easy to compare sizes of pie pieces versus comparing heights of bars. Thus:

- Don't Use Pie Charts — harder for people to distinguish between the sizes of pieces of the pie — better frame of reference is height in bar graphs.
- No GOOD reason to use a pie chart — a bar graph does a good job.

Recall: We have been doing *Exploratory Data Analysis* - using graphs and numerical summaries to describe the variables in a data set.

1. Dot plot, stem plot and histogram for quantitative variables.
2. Bar graph and pie chart for qualitative variables.

Section 2.7 Summary. Read it!