DSO 530 Week1 Technical Notes: review of basic statistics I

What are Data?

- For individuals (people, objects, etc.), measurements are taken for some **variables**, and the resulting measurement values are **data**.
- Variable is a characteristic of an individual, and it has two types:

  - **categorical** a.k.a. **qualitative** (e.g. gender)

    - A categorical variable is called **ordinal** if its categories can be ordered (e.g. your letter grade of this course)

  - **numerical** a.k.a. **quantitative** (e.g. height)

- Note: Only numerical variables allow arithmetic operations. Categorical variables don't.

Data Tables

- Columns correspond to Variables
- Rows correspond to individuals, often called **observations**
- The number of rows is traditionally denoted by n

variable type ? ? ? ?

$n = 5$

Song	Artist	Genre	Size (MB)	Length (sec)
My Friends	D. Williams	Alternative	3.83	247
Up the Road	E. Clapton	Rock	5.62	378
Jericho	k.d. lang	Folk	3.48	225
Dirty Blvd.	L. Reed	Rock	3.22	209
Nothingman	Pearl Jam	Rock	4.25	275

Exploratory Data Analysis (EDA)

An initial examination of the data.

1. Examine each variable

2. Examine each pair of variables to study their relationship

At 1 and 2,
- numerical summary (“to calculate numbers”)
- graphical summary (“to make plots”)

Distribution

Questions about examining one variable:

- What are the possible values this variable takes?
- How frequently does this variable take those values?

$\Rightarrow$ **distribution**:

frequencies of the possible values of a variable

How to describe and display the distribution of

- a categorical variable?
- a numerical variable?

Categorical (Qualitative) Variables

The values of a categorical variable are labels of categories.

- Example: education levels of 38.4 million young American adults from the 1999 Current Population Survey
- Variable: education level
- *n* = 38.4 million
- Five labels: “Less than high school”, “High school graduate”, “Some college”, “Bachelor’s degree”, “Advanced degree”
- The data have the format
  - “Some college”, “Less than high school”, “High school graduate”, “High school graduate”, “High school graduate”, “Some college”, “Bachelor’s degree”, “Advanced degree”, “Bachelor’s degree”, ...

Numerical Summary of a Categorical Variable

One can describe the distribution of a categorical variable is by using the **count** or the **percentage** of individuals who fall in each category.

- Example: a numerical summary of the education data

Education	Count (millions)	Percent
Less than high school	4.7	12.3
High school graduate	11.8	
Some college	10.9	
Bachelor's degree	8.5	
Advanced degree	2.5	

Numerical Summary of a Categorical Variable

One can describe the distribution of a categorical variable is by using the **count** or the **percentage** of individuals who fall in each category.

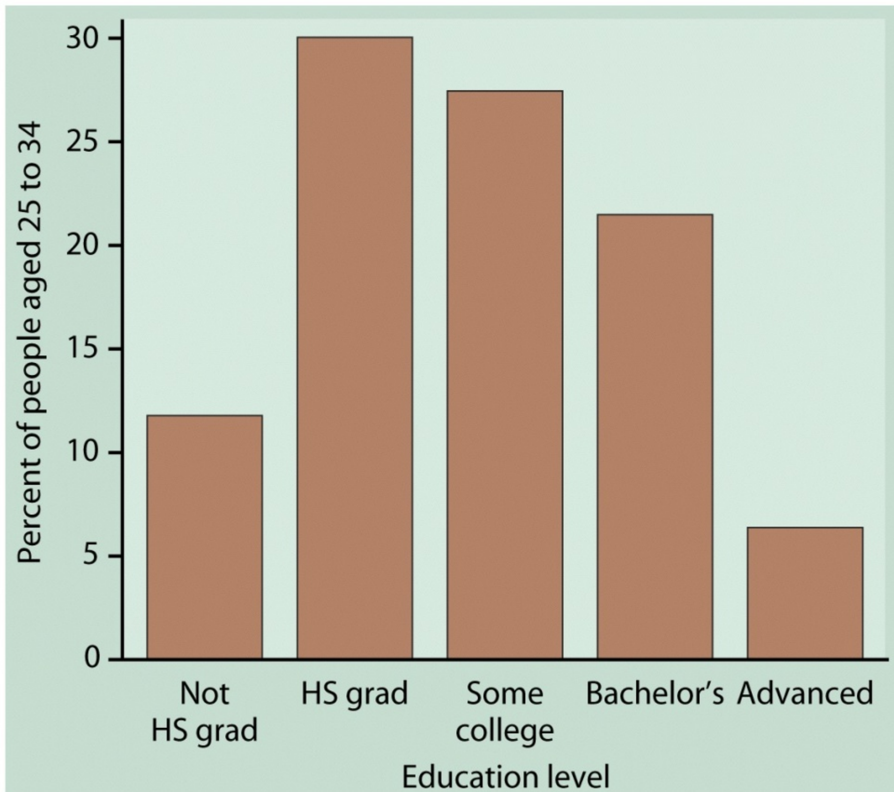
- Example: a numerical summary of the education data

- Question: do **counts** and **percentages** convey the same information?

Education	Count (millions)	Percentage
Less than high school	4.7	12.3
High school graduate	11.8	30.7
Some college	10.9	28.3
Bachelor’s degree	8.5	22.1
Advanced degree	2.5	6.6

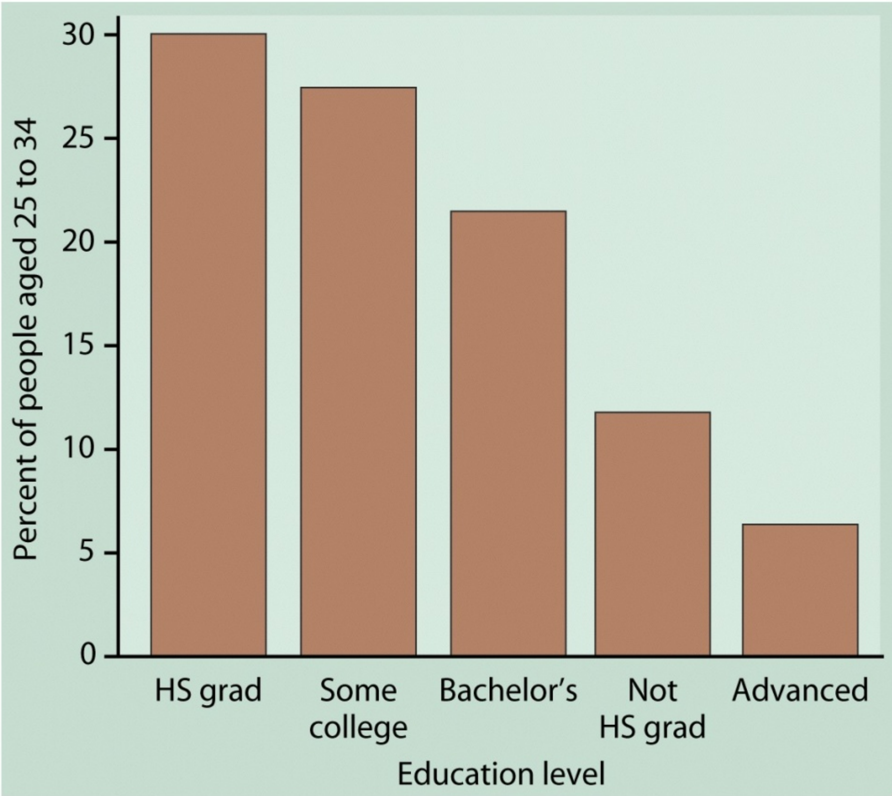
Graphical Summary of a Categorical Variable

- In a **bar chart** the height of each bar is proportional to the count (or percentage) of each category



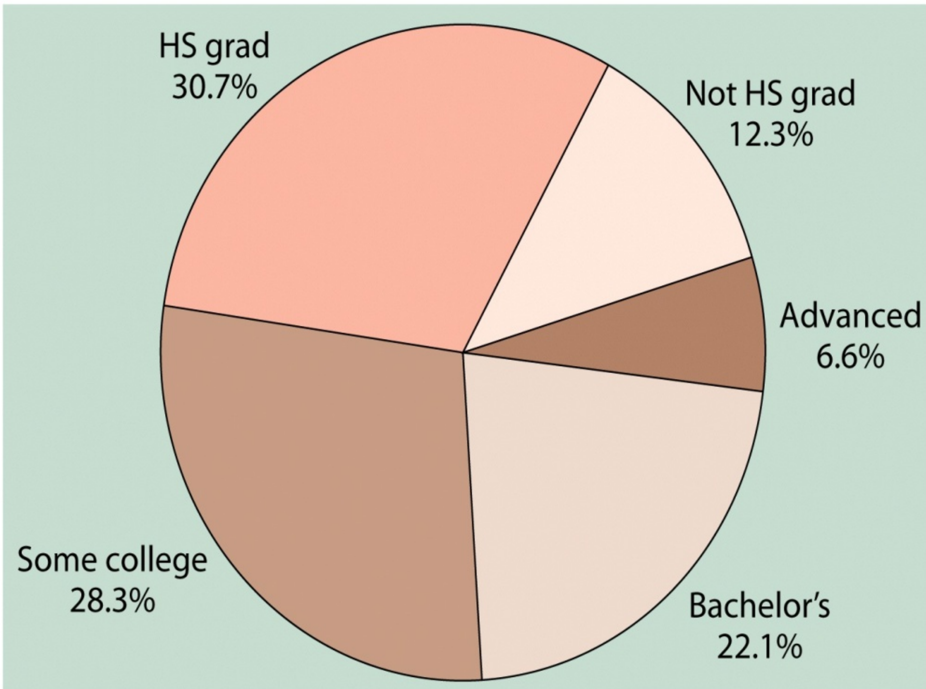
Graphical Summary of a Categorical Variable

- A bar chart is called a **Pareto chart** when the categories are sorted by frequency (popular in quality control)



Graphical Summary of a Categorical Variable

- In a **pie chart** the area of each piece is proportional to the count (or percentage) of each category



Notes on Bar Charts and Pie Charts

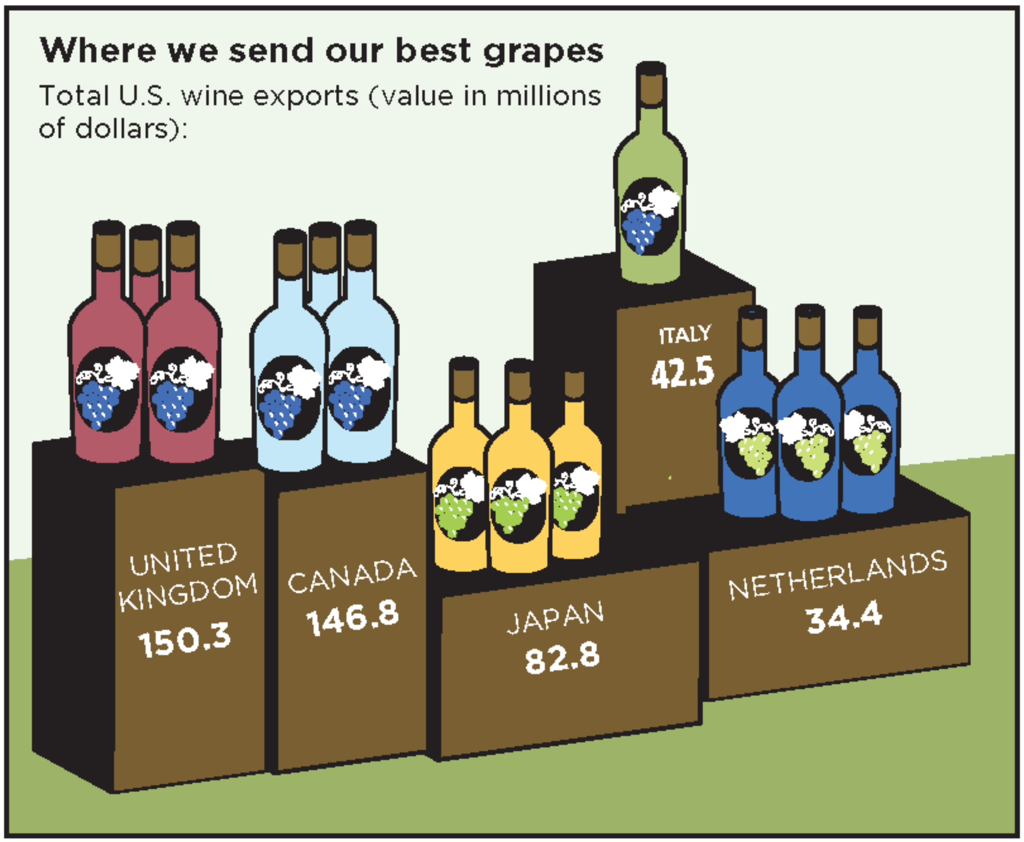
Pie charts are commonly chosen to illustrate market shares or sources of revenue for a company

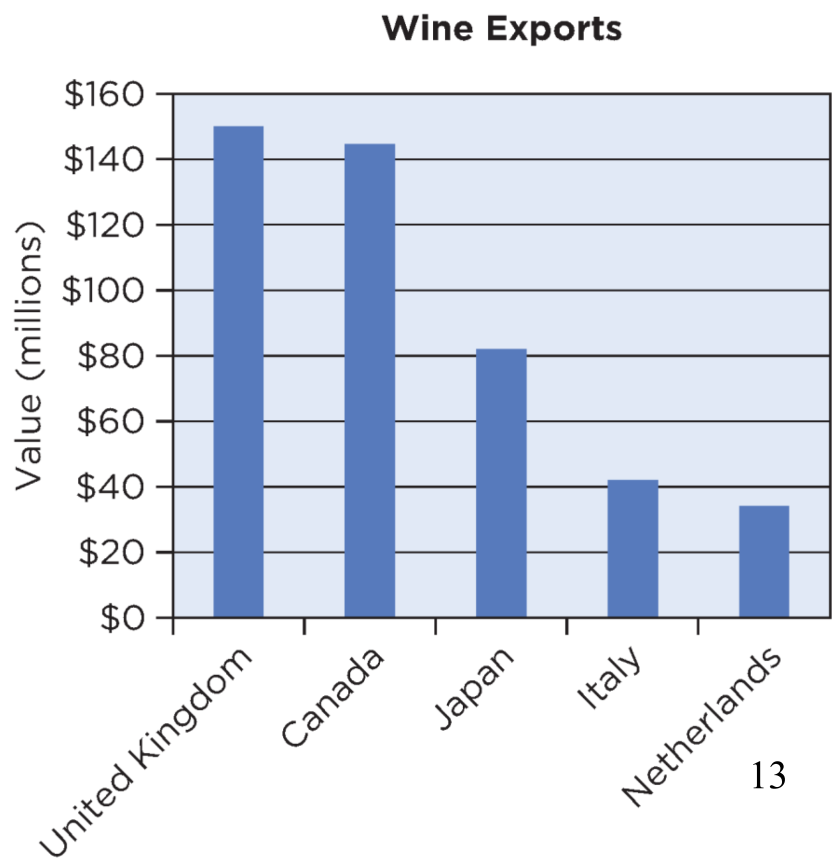
Pie charts are less useful than bar charts if we want to compare actual counts (easier to compare bars than angles of wedges)

The Area Principle

In a graph/chart, an area representing data should be proportional to the amount of the data.

In popular media, charts decorated to attract attention often violate the area principle.





Numerical (Quantitative) Variables

The values of a numerical variable are numbers allowing arithmetic operations.

- Example: the "sepal length" variable from the Iris data (http://www.saedsayad.com/datasets/iris.txt)
- n = 150
- The data have the format

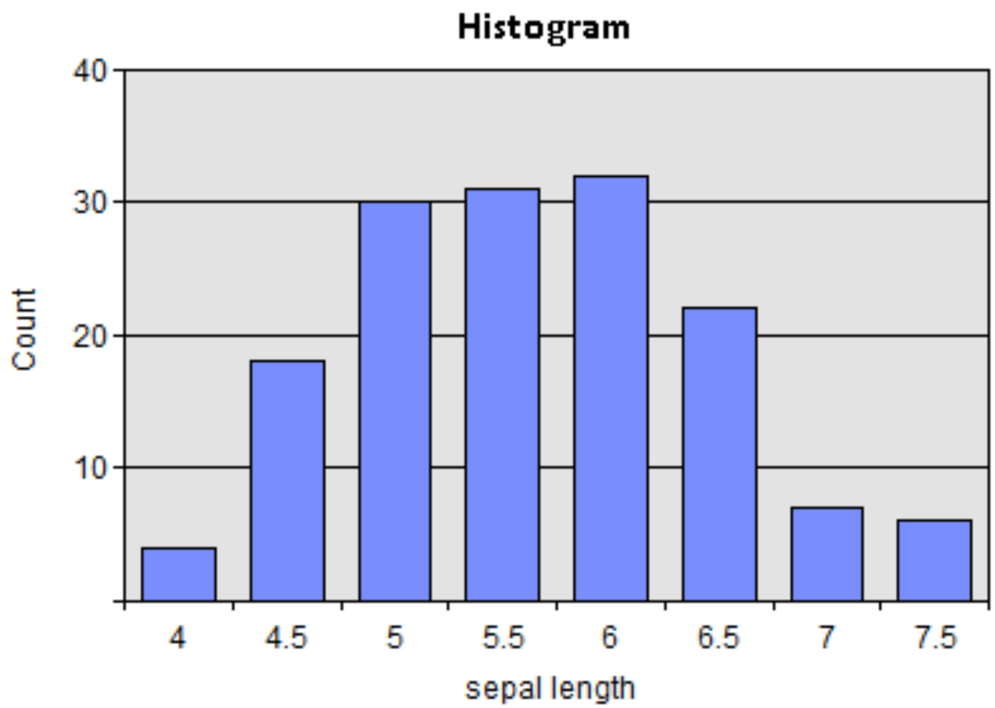
	A	B	C	D	E
1	sepal length	sepal width	petal length	petal width	iris
2	5.1	3.5	1.4	0.2	Iris-setosa
3	4.9	3	1.4	0.2	Iris-setosa
4	4.7	3.2	[illegible]	[illegible]	Iris-setosa
5	4.6	3.1	[illegible]	[illegible]	Iris-setosa
6	5	3.6	[illegible]	[illegible]	Iris-setosa
7	5.4	3.9	[illegible]	[illegible]	Iris-setosa
8	4.6	3.4	[illegible]	[illegible]	Iris-setosa
9	5	3.4	[illegible]	[illegible]	Iris-setosa
10	4.4	2.9	[illegible]	[illegible]	Iris-setosa
11	4.9	3.1	1.5	0.1	Iris-setosa

Graphical Summary of a Numerical Variable

Histogram

To make a histogram:

1. Divide the range of the possible values into equal intervals.
2. For each interval draw a rectangle whose base is the interval and whose height is proportional to the number of observations falling into the interval.

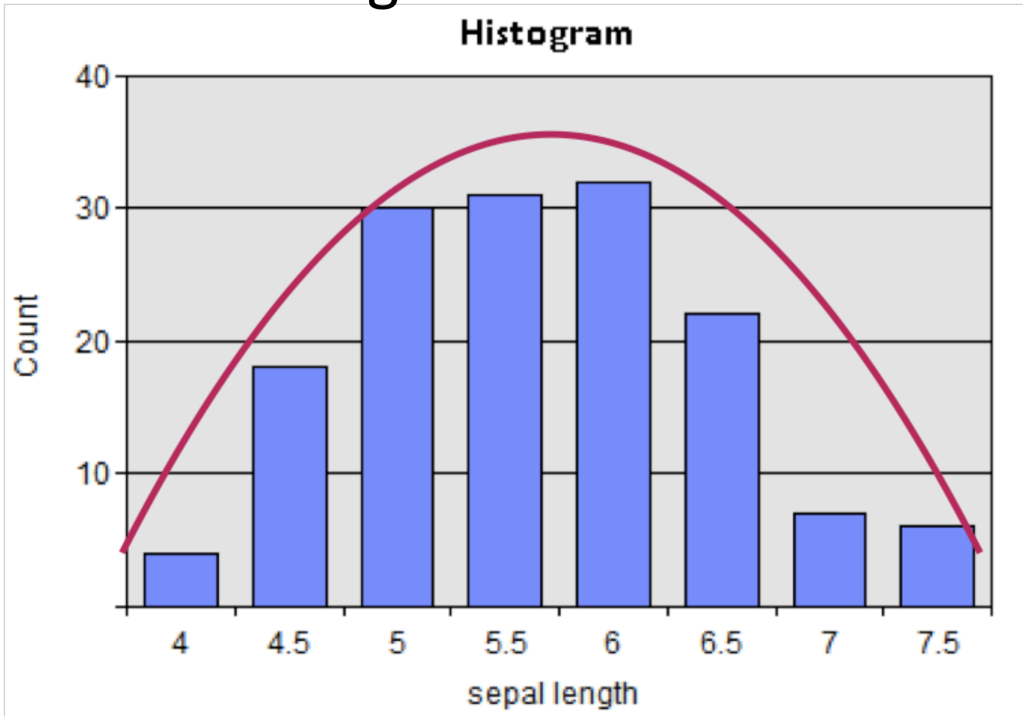


Graphical Summary of a Numerical Variable

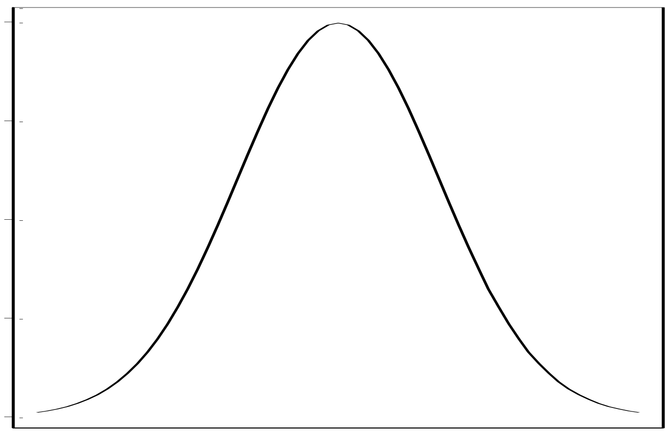
Histogram

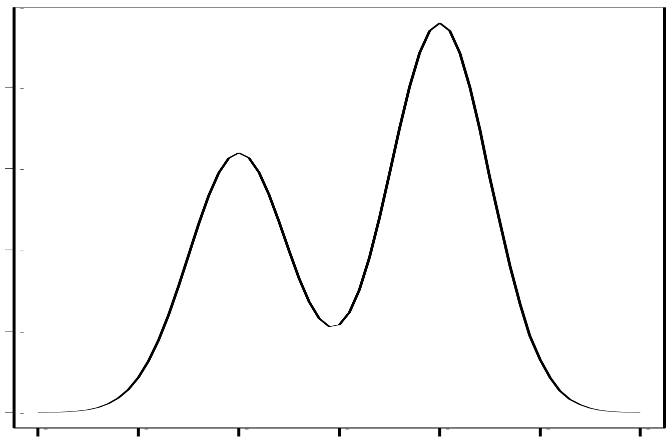
To make a histogram:

1. Divide the range of the possible values into equal intervals.
2. For each interval draw a rectangle whose base is the interval and whose height is proportional to the number of observations falling into the interval.



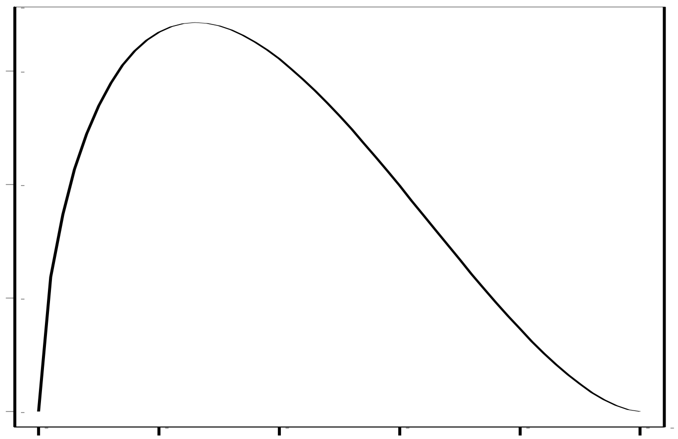
Shapes of a Distribution

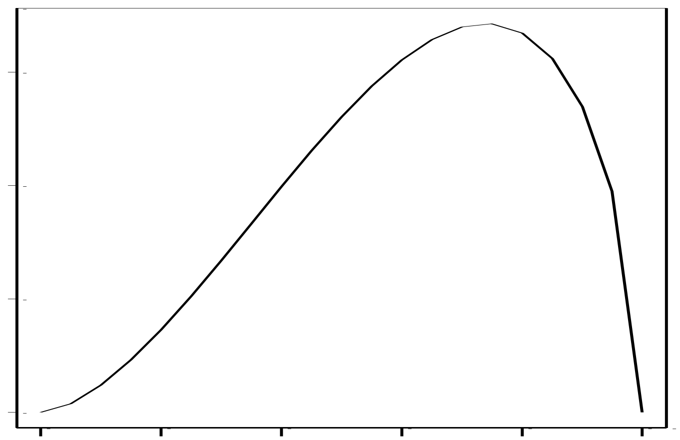




Words to describe the two shapes?

Shapes of a Distribution



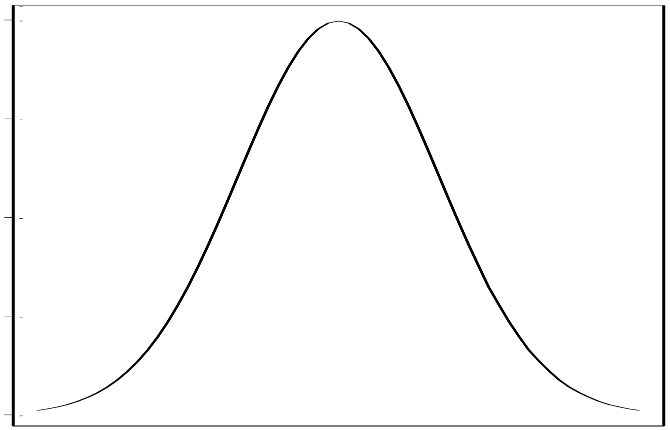


Words to describe the two shapes?

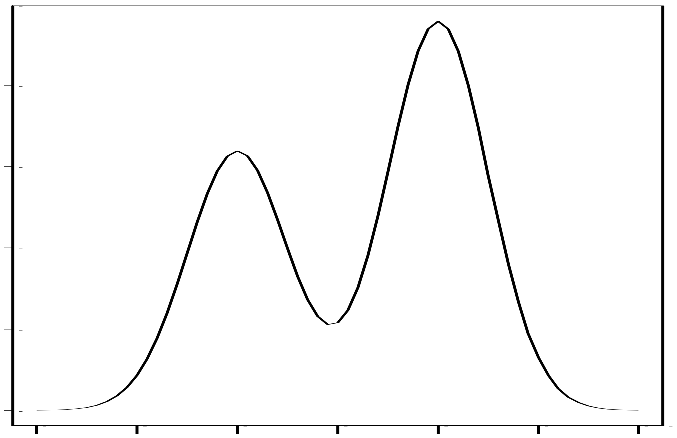
Words that Describe Distributions

- **Unimodal** : has one major peak
- **Bimodal**: has two major peaks
- **Symmetric**: there is a symmetry with respect to the middle point
- **Skewed to the right**: when the right tail (larger values) is much longer than the left tail (smaller values)
- **Skewed to the left**

Shapes of a Distribution

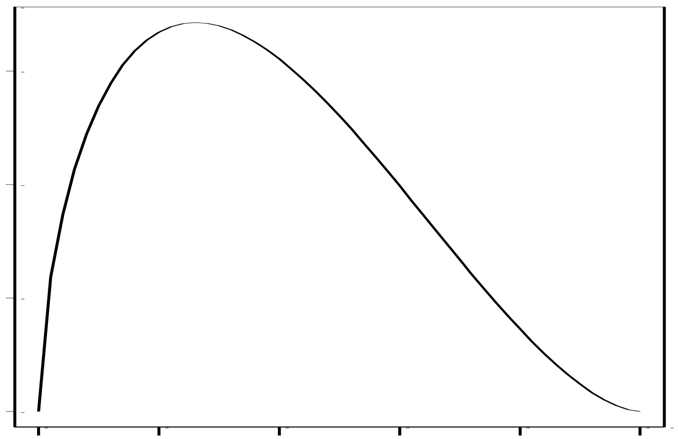


unimodal

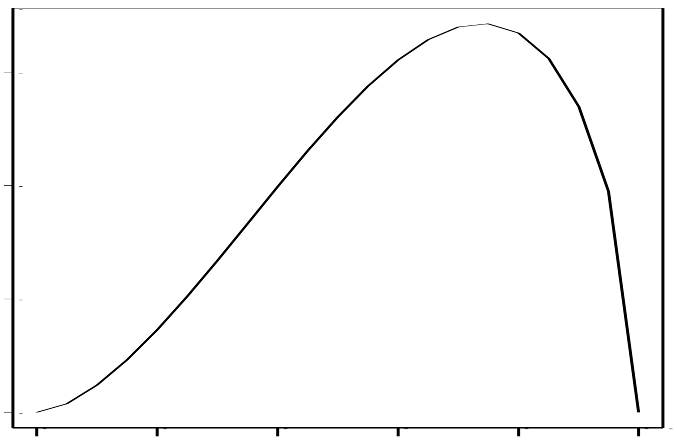


bimodal

Shapes of a Distribution



skew to the right



skew to the left

Numerical Summary of a Numerical Variable

- Different ways of getting at the idea of a “center” of a distribution:

  - Mean = average
  - Median = 50th percentile

More detail

E.g. if data is 6, 9, 8, 3, 3, 1

$$\text{Mean} = \frac{6+9+8+3+3+1}{6} = 5$$

For a variable *x* with *n* observed values $x_1, x_2, \ldots, x_n$

the mean of *x* is

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

Median = 50th percentile

Arrange data in order.

Median M_d = 50th percentile = "middle observation"

[if number of observations is even, average the middle two.]

E.g. for data 1, 3, 3, 6, 8

$M_d = 3$

E.g. for data 1, 3, 3, 6, 8, 9

$M_d = (3 + 6) / 2 = 4.5$

"Robustness" (resistant to outlier)

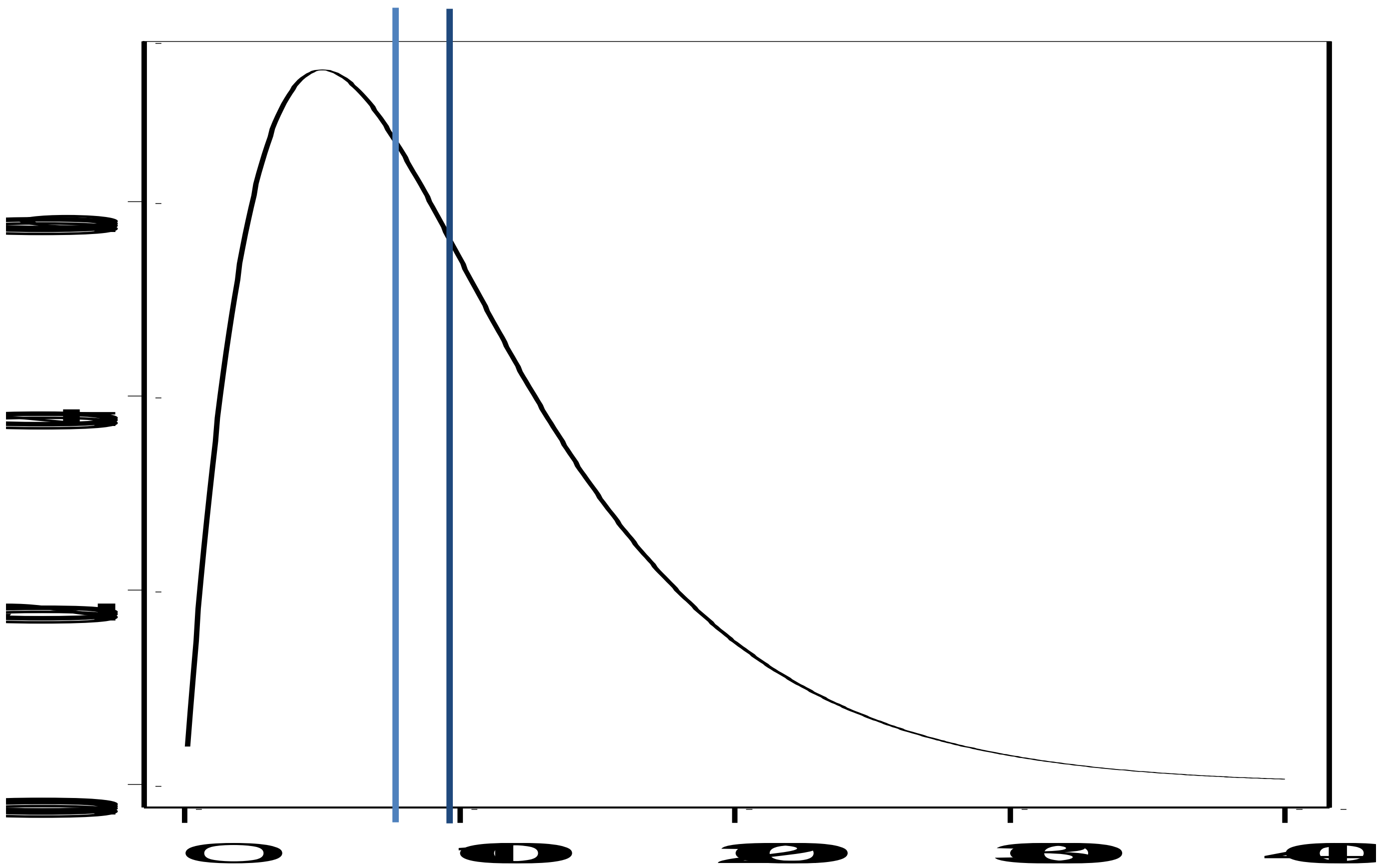
Robust = insensitive to a few extreme observations

Which is more robust: mean or median ?

Compare 1, 3, 3, 6, 8

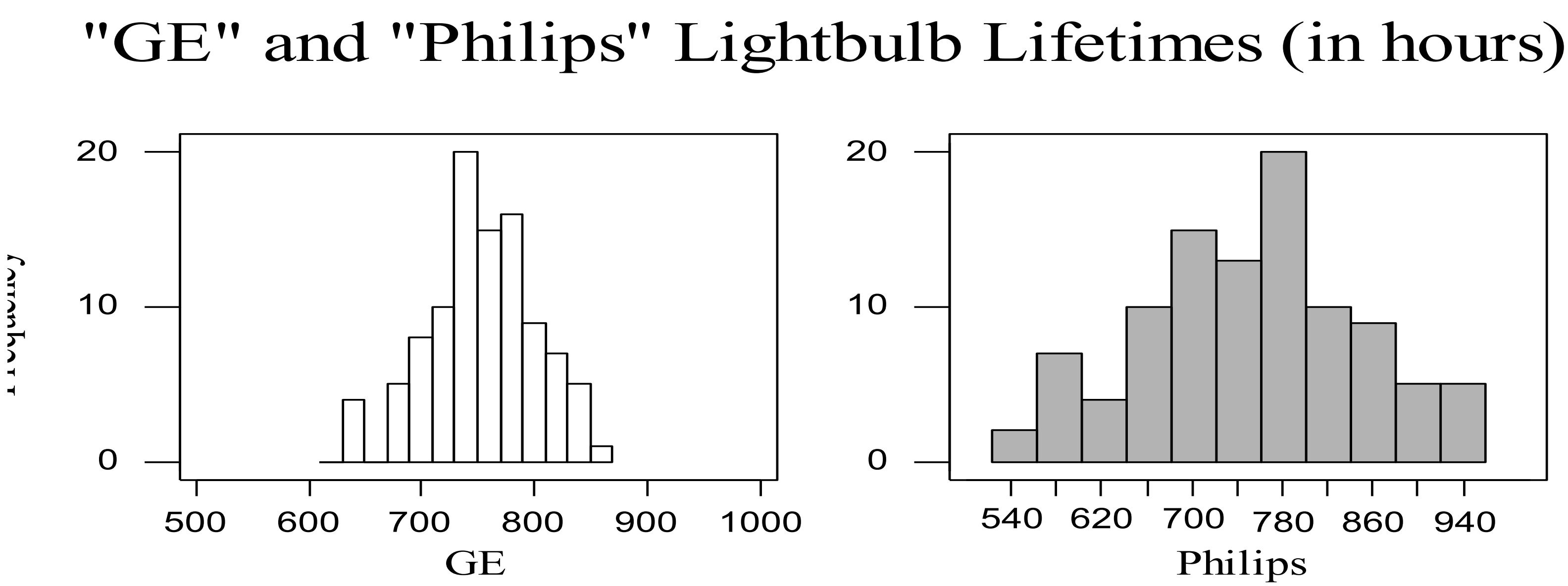
to 1, 3, 3, 6, 8000000

Which is which?



Numerical Summary of a Numerical Variable (cont'd)

Some measure of spread is needed.



"Philips" has more fluctuation although *average* is about same as "GE"

"GE" exhibits better quality control: not much variation.

Mean and *median* do not completely summarize a data set.
Need to know amount of fluctuation!

Common measures of variability

Variance:

The "average" of the squared deviations of all the measurements from the mean (*details to follow*)

Standard Deviation:

The square root of the variance

Variance and Standard Deviation (SD)

- SD is the most common measure of spread or variability

  Relationship: $\text{SD} = \sqrt{\text{Variance}}$

Notation:

$$\text{Variance} = s^2, \quad \text{SD} = s$$

Idea of variance and SD

- How far away are the observations, on average, from the mean?

- Based on the **deviations:**

$x_i - \bar{x}$ = deviation from the mean for the *i*-th observation

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

"average" squared deviation

Example: car mileage case

Gas mileages of a new midsize model (five randomly selected cars):

$$x_1 = 30.8, \quad x_2 = 31.7, \quad x_3 = 30.1, \quad x_4 = 31.6, \quad x_5 = 32.1$$

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = 31.26$$

$$s = \ldots$$

Calculating Variance and SD

x_i	$\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
30.8	31.26	-.46	.20
31.7	31.26	.44	.19
30.1	31.26	-1.16	1.35
31.6	31.26	.34	.12
32.1	31.26	.84	.71
		0	2.57

$n = 5$

$n - 1 = 4$

$$s^2 = \frac{2.57}{4}$$

$$= .64$$

$$s = \sqrt{.64} = .80$$

Other Measures of Spread

- Range = max – min

- Interquartile range (IQR)

Quartiles

- Define **first quartile** to be the median of the observations below the median
- Define **third quartile** to be the median of the observations above the median

1, 3, 3, 6, 8, 9

M=4.5

Q1=3 Q3=8

The **interquartile range IQR** is Q3 - Q1

Example: customer satisfaction ratings

20 measurements on the 10 point scale:

9, 8, 3, 8, 10, 9, 8, 9, 5, 8, 1, 10, 8, 10, 7, 8, 9, 10, 5, 9

1 3 5 5 7 8 8 8 8 8 8 9 9 9 9 9 10 10 10 10

$M_d = (8+8)/2 = 8$

$Q_1 = (7+8)/2 = 7.5$

$Q_3 = (9+9)/2 = 9$

$IQR = Q_3 - Q_1 = 9 - 7.5 = 1.5$

More about Numerical Summary: Mode

Categorical variable: the category with the highest frequency

Numerical variable: location of a major peak of the distribution