Exploratory Data Analysis and Visualization

2. EDA

Li Xinke

DS

City University of Hong Kong

What are Data?

- For individuals (people, objects, etc.), measurements are taken for some variables, and the resulting measurements value are data.
- Variable is a characteristic of an individual, and it has two types:
  - **categorical** a.k.a **qualitative** (e.g., gender, blood type, disease status)
    - A categorical variable is called **ordinal** if its categories can be ordered (e.g., your letter grade of this course, covid-19 severity level)
  - **numerical** a.k.a **quantitative** (e.g., height, weight, age, income, blood pressure)
- Note: only numerical variables allow arithmetic operations, categorical variables does not.

Data Tables

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

variable type ? ? ? ?

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

$n = 5$

Example from Python

Auto.csv: information for cars

- mpg: miles per gallon
- cylinders: number of cylinders
- displacement: engine displacement
- horsepower: engine horsepower
- weight: vehicle weight
- acceleration: time to accelerate
- year: model year
- **origin: origin of car (1. American, 2. European, 3. Japanese)**
- name: vehicle name

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin	name
0	18.0	8	307.0	130	3504	12.0	70	1	chevrolet chevelle malibu
1	15.0	8	350.0	165	3693	11.5	70	1	buick skylark 320
2	18.0	8	318.0	150	3436	11.0	70	1	plymouth satellite
3	16.0	8	304.0	150	3433	12.0	70	1	amc rebel sst
4	17.0	8	302.0	140	3449	10.5	70	1	ford torino

Exploratory Data Analysis (EDA)

An **initial** examination of the data.

1. Examine each variable
2. Examine pair of variables to study their relationship

General EDA methods include two types,

- numerical summary (“calculate numbers”)
- graphical summary (“make plots”)

Univariate Analysis——Distribution

First few questions about examining one variable:

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

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin	name
0	18.0	8	307.0	130	3504	12.0	70	1	chevrolet chevelle malibu
1	15.0	8	350.0	165	3693	11.5	70	1	buick skylark 320
2	18.0	8	318.0	150	3436	11.0	70	1	plymouth satellite
3	16.0	8	304.0	150	3433	12.0	70	1	amc rebel sst
4	17.0	8	302.0	140	3449	10.5	70	1	ford torino

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 ...

Numerical Summary of a Categorical Variable

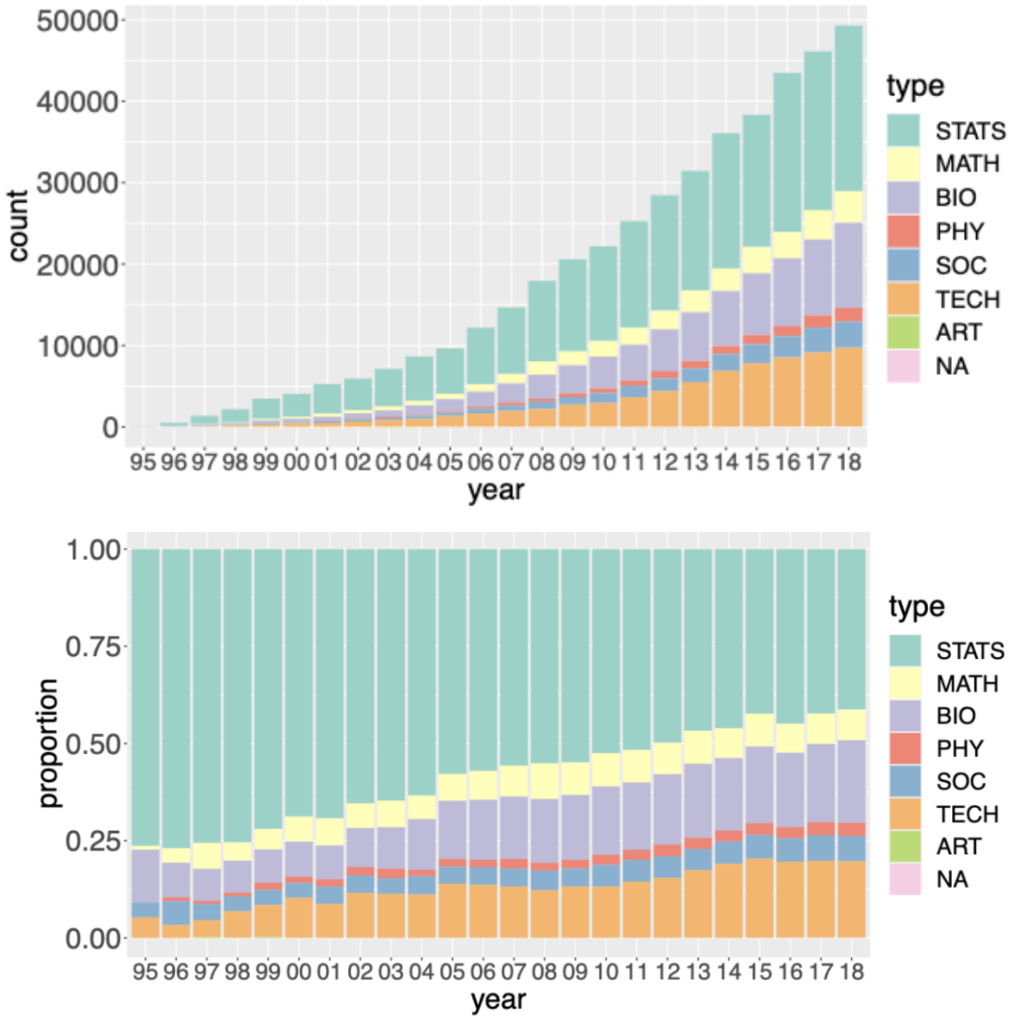
One can describe the distribution of a categorical variable is by using the **count** or the **percentage** of individuals who fall in each category.

- Example: a numerical summary of the education data

Education	**Count (millions)**	**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

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

Citation count vs citation percentage

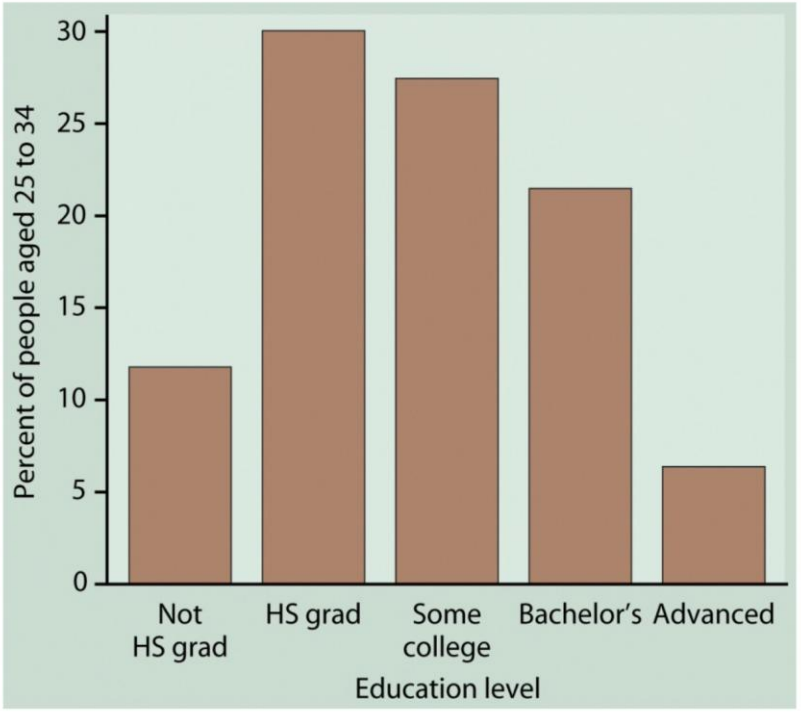


https://www.cell.com/patterns/pdfExtended/S2666-3899(22)00129-5

Graphical Summary of a Categorical Variable

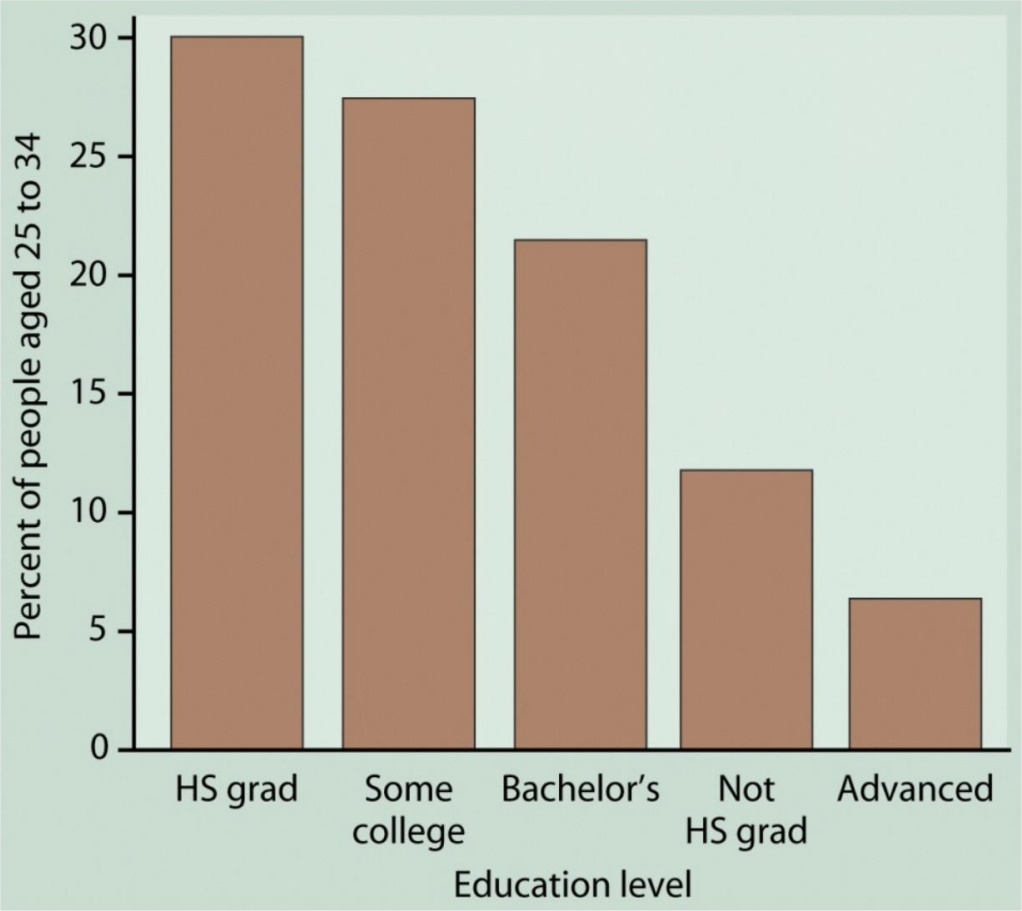
Education	Count (millions)	Percentage
Less than high school	4.7	12.3
High school graduate	11.8	30.7
Some college	10.9	28.3
Bachelor's degree	8.5	22.1
Advanced degree	2.5	6.6

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



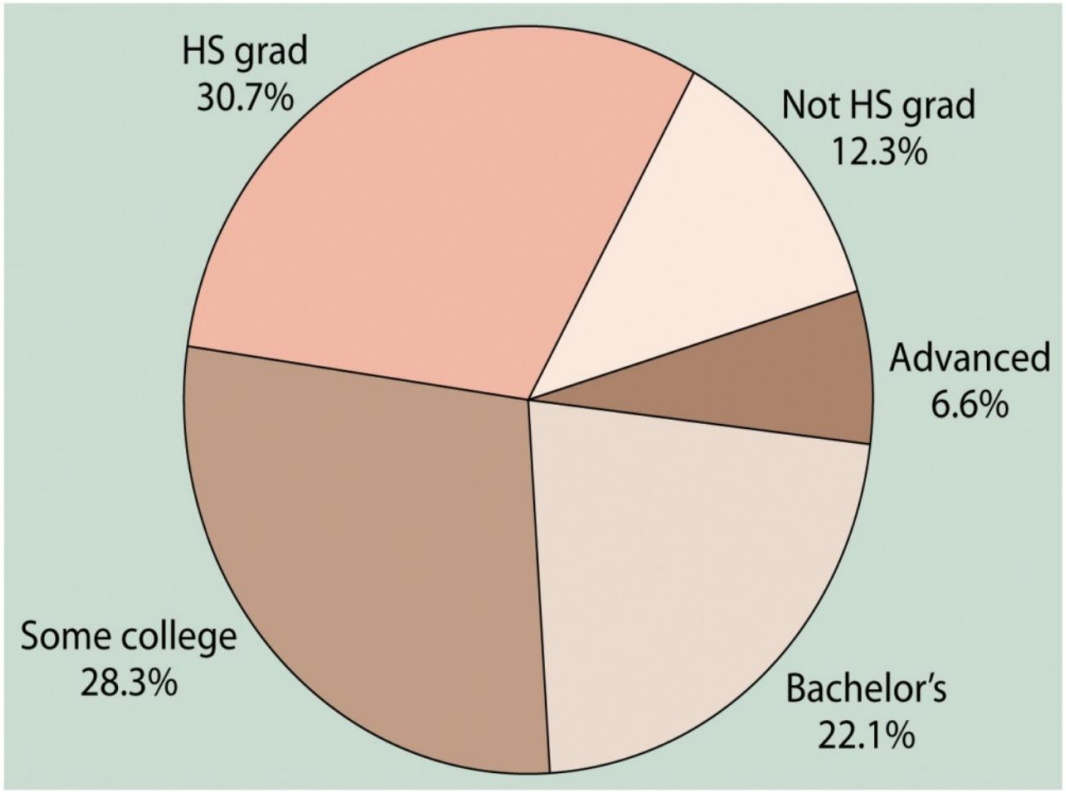
Graphical Summary of a Categorical Variable

- 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

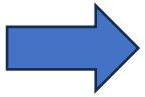
- Pie charts are less useful than bar charts if we want to compare actual counts (easier to compare bars than angles of wedges)

Univariate Analysis——Distribution

Distribution:

frequencies of the possible values of a variable.

How to describe and display the distribution of

- a categorical variable?
- a numerical variable? 

Methods:
- numerical summary
- graphical summary

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

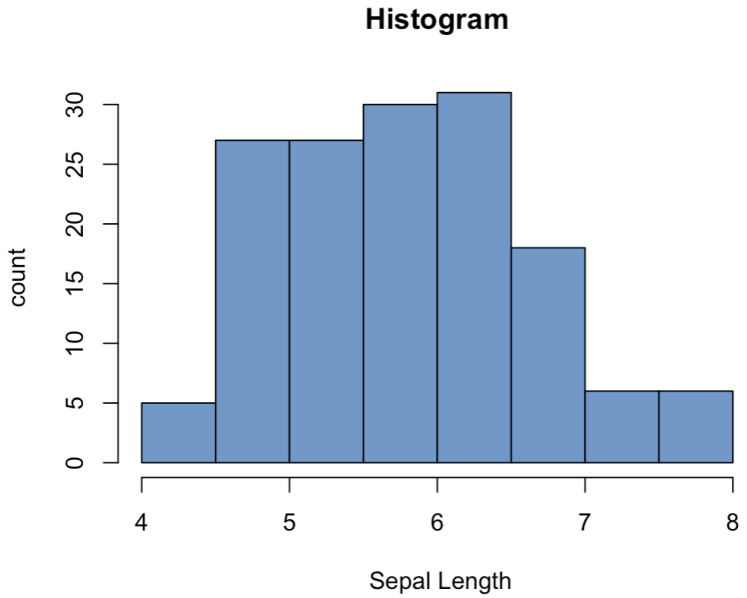
	A	B	C	D	E
1	sepal length	sepal width	petal length	petal width	iris
2	5.1	3.5	1.4	0.2	Iris-setosa
3	4.9	3	1.4	0.2	Iris-setosa
4	4.7	3.2			Iris-setosa
5	4.6	3.1			Iris-setosa
6	5	3.6			
7	5.4	3.9			
8	4.6	3.4			
9	5	3.4			
10	4.4	2.9			

Graphical Summary of a Numerical Variable

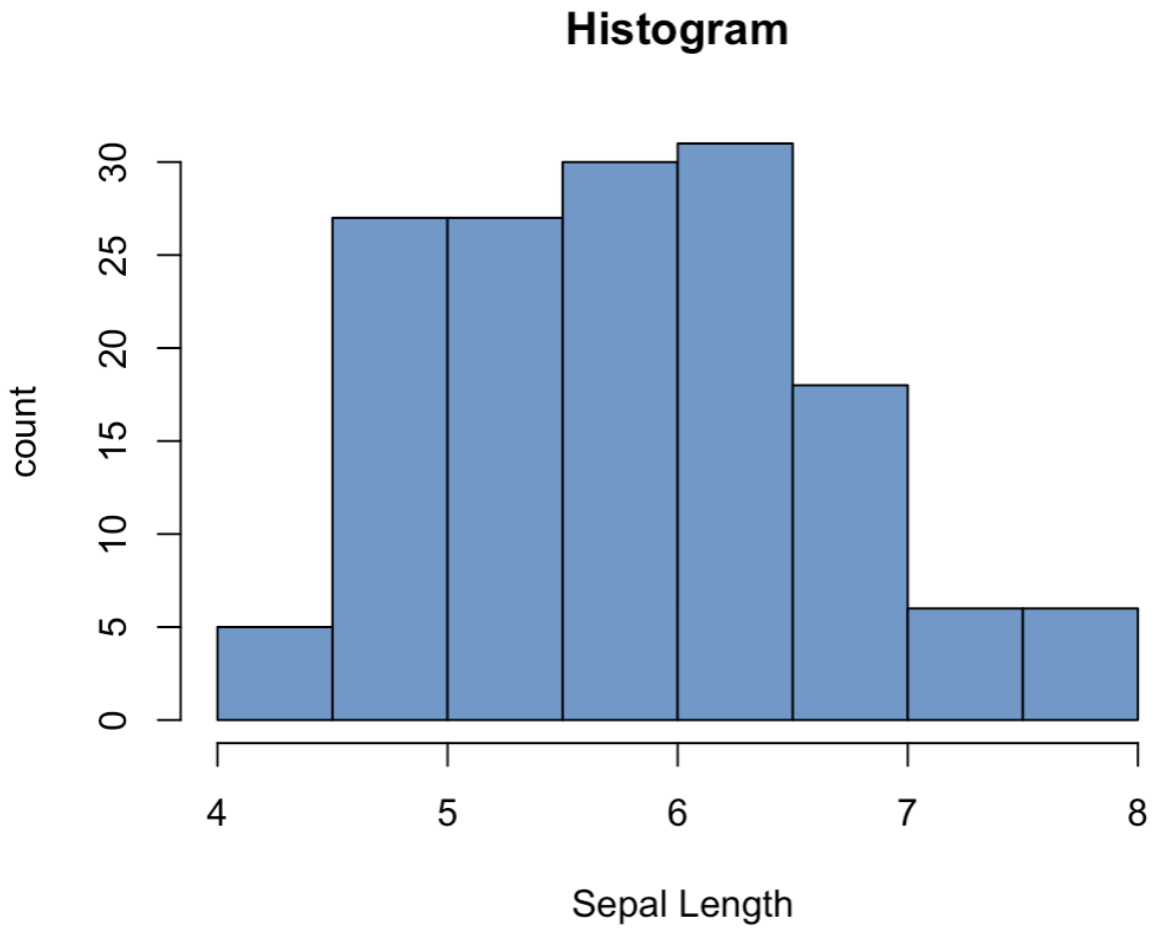
Histogram

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

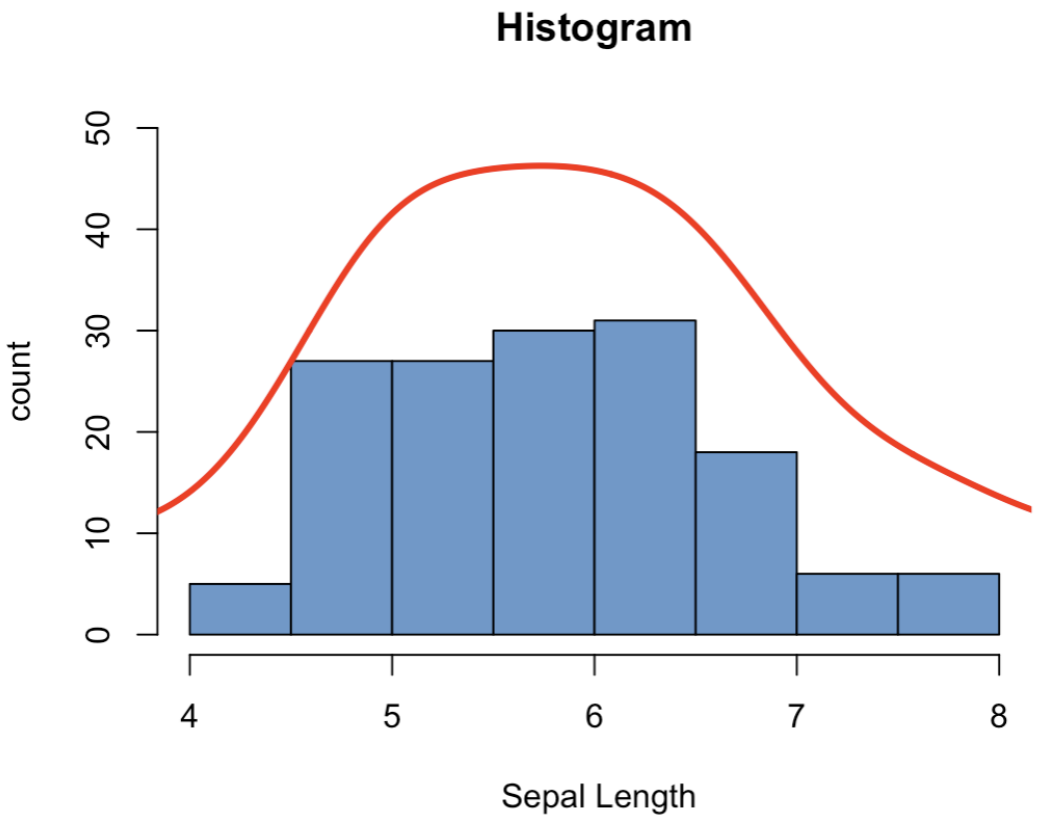



```
 Sepal.Length
Min.   :4.300
1st Qu.:5.100
Median :5.800
Mean   :5.843
3rd Qu.:6.400
Max.   :7.900
```

Divide the interval [4, 8] into 8 equal intervals

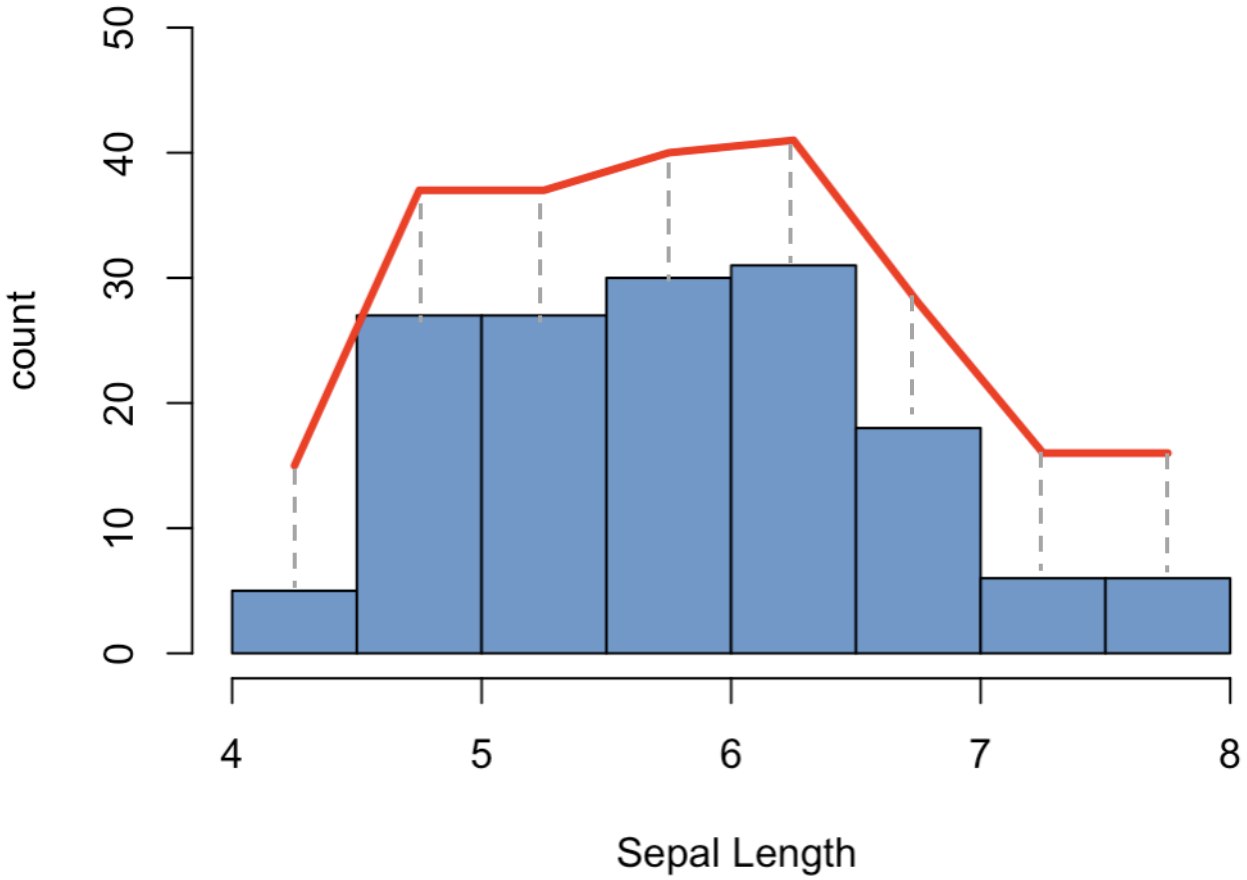
Graphical Summary of a Numerical Variable

Histogram and density curve



A density curve shows (not exactly) the proportion of values in each range.
A density plot is a smooth curve that shows the distribution of the data in a more continuous way.

Why not

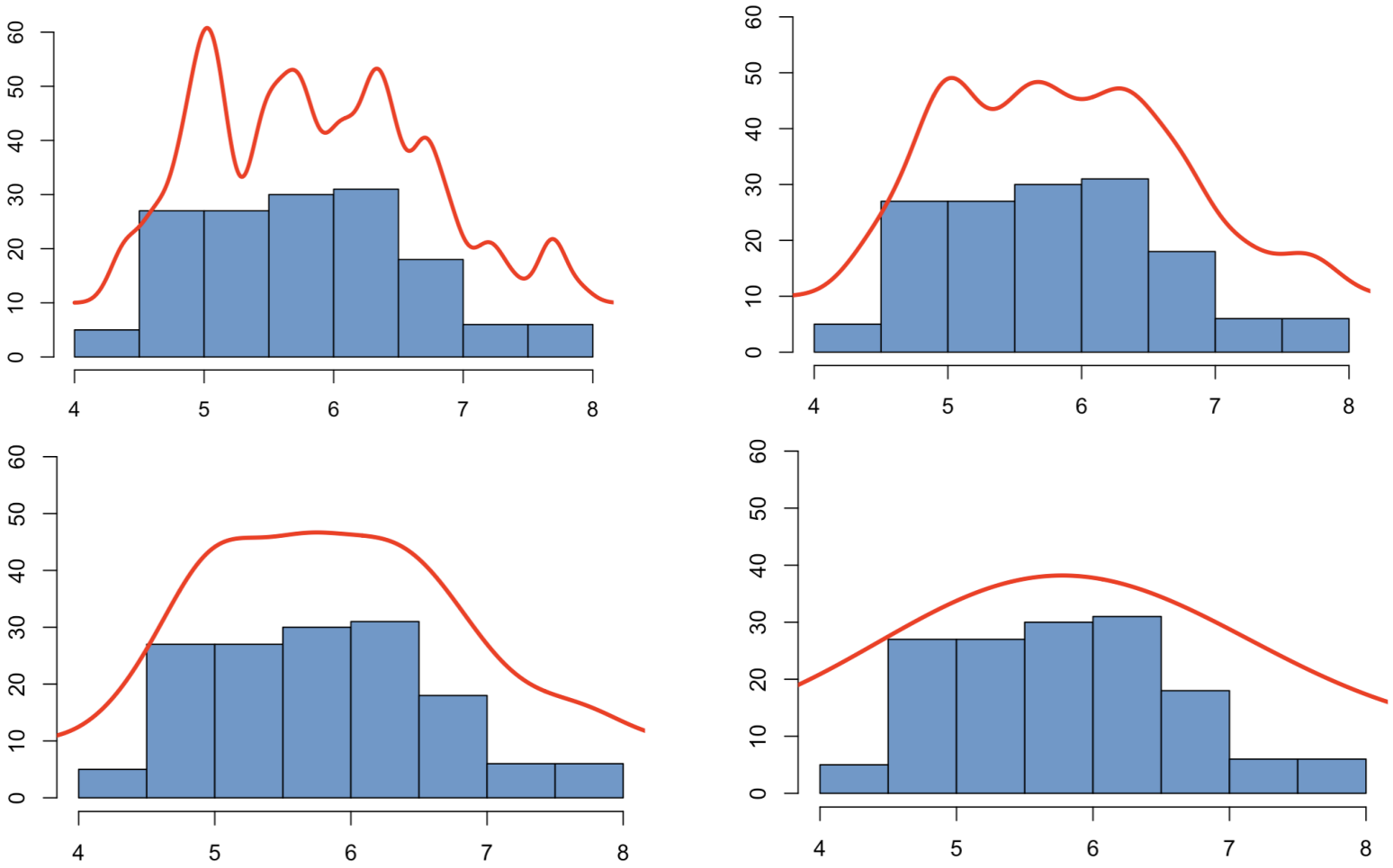


A density curve does not exactly mark the proportion of values in each range.

Graphical Summary of a Numerical Variable

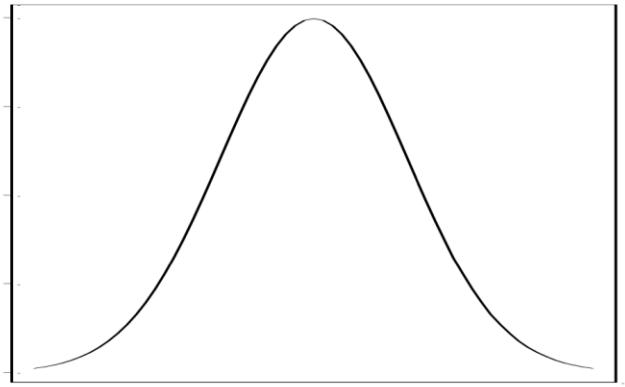
"Smoothing" is done by **kernel density estimation** methods.

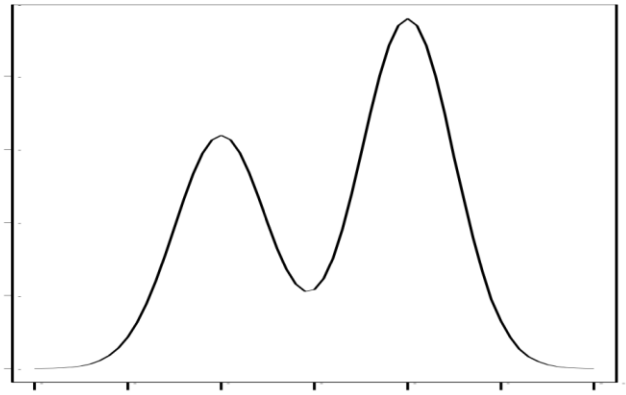
https://en.wikipedia.org/wiki/Kernel_density_estimation

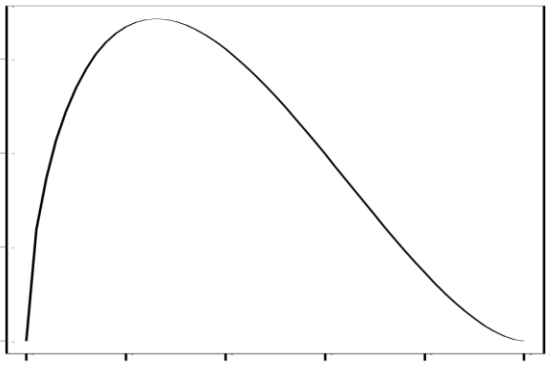


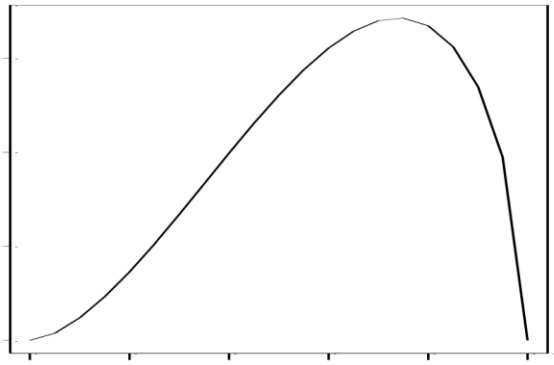
Gaussian kernel with different bandwidth

Shapes of a Distribution





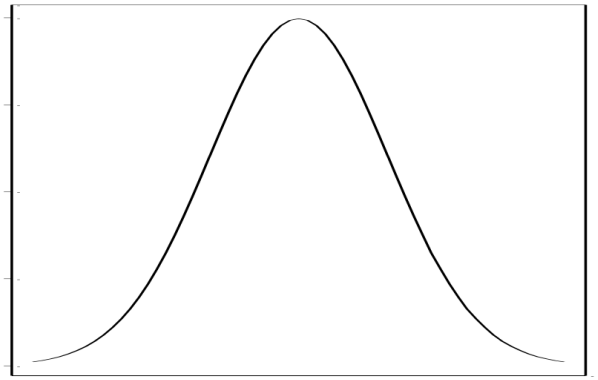




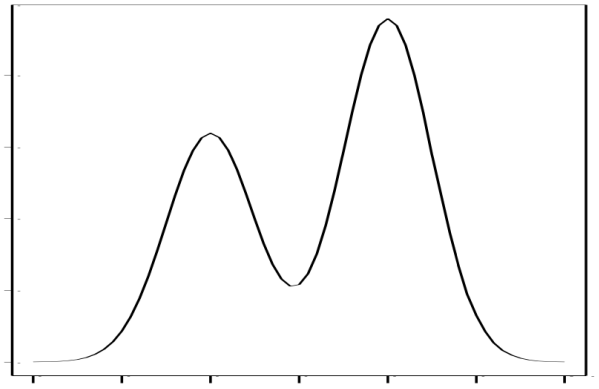
Words that Describe Distributions

- **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：** when the left tail is much longer than the right tail

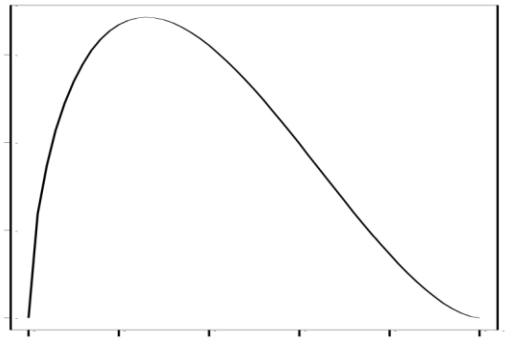
Shapes of a Distribution



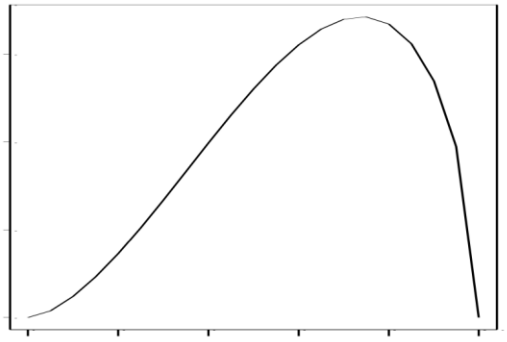
unimodal



bimodal



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 details

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

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

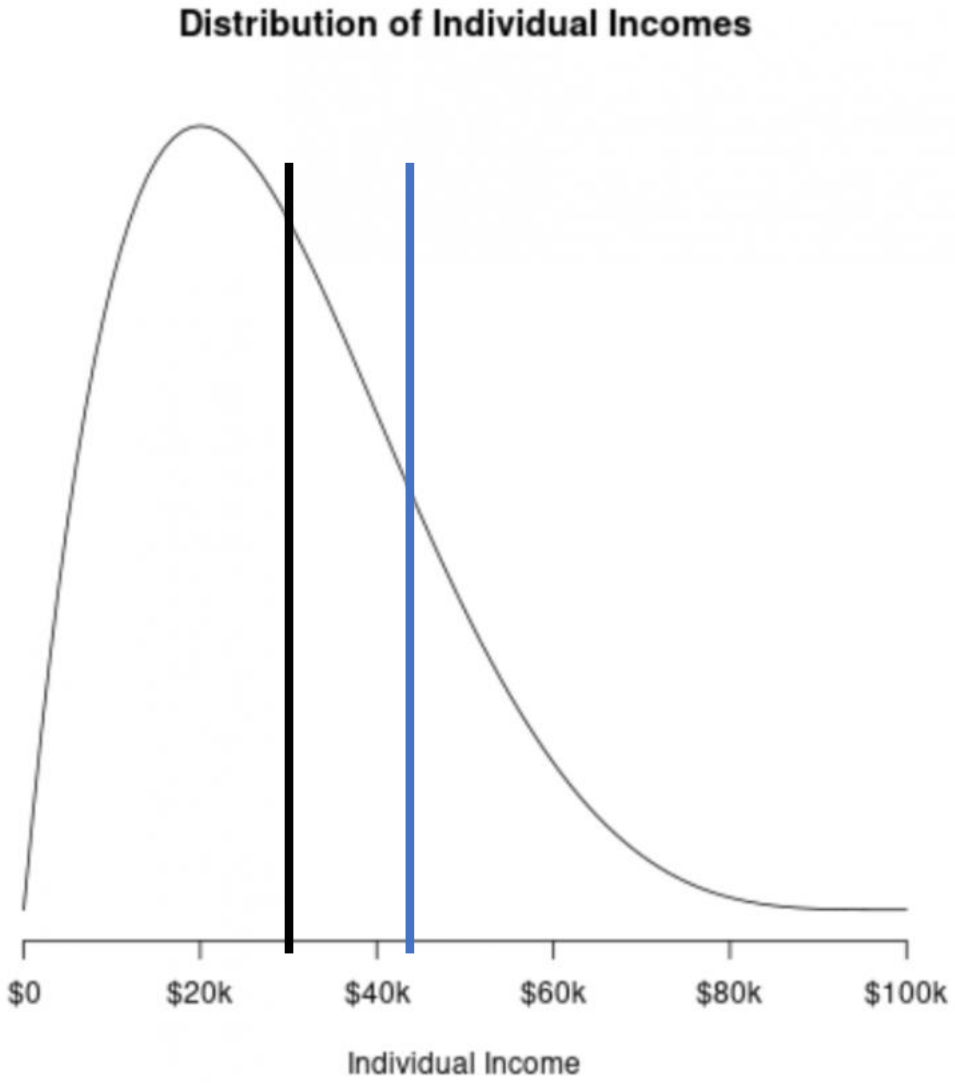
“Robustness” (resistant to outlier)

- 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?



More about Numerical Summary: Mode

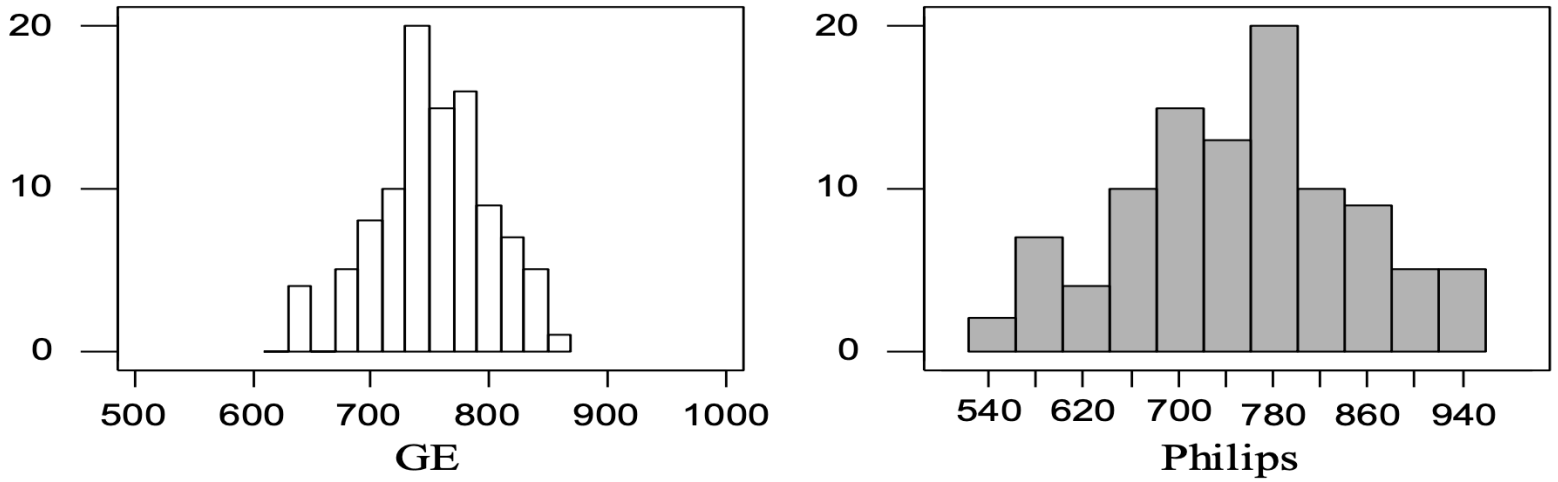
- Categorical variable: the category with the highest frequency

- Numerical variable: location of a major peak of the distribution

Numerical Summary of a Numerical Variable

Some measure of spread is needed.

"GE" and "Philips" Lightbulb Lifetimes (in hours)



- "Philips" has more **fluctuation** although *average* is about same as "GE".
- We say that "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 & SD

Variance: The "average" of the squared deviations of all the measurements from the mean.

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

- Based on the **deviations:**

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

- Variance:

$$\sigma^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{N}$$

"average" squared deviation

Why Taking Squares?

- Sum of deviations (not squared) is just 0.
- Squaring the deviations converts the negative deviations to positive numbers.
- So Variance is strictly positive as long as the variable does not take a single value only.

Variance and Standard Deviation (SD)

Standard Deviation: The square root of the variance

- SD is the most common measure of spread or variability

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

- Notation:

  Variance = s^2, SD = s

Why divide by n-1 and not n?

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{N}$$

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

- It's unimportant if *n* is large.

- Dividing by *n-1* gives an unbiased estimate of variance.

What happens when $n = 1$??

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)

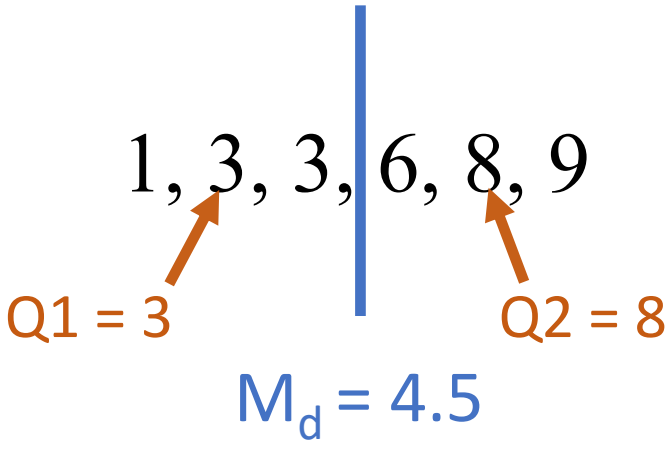
Quartile? Percentile? Quantile?

Quartiles

- Define **first quartile (Q1)** to be the median of the observations no greater than the median
- Define **third quartile (Q3)** to be the median of the observations no less than the median
- **Interquartile range (IQR) = Q3 - Q1**

Find the quartiles for 1, 3, 3, 6, 8, 9

Quartiles



The **interquartile range IQR** is Q3 - Q1 = 5

What is quantile?

Quantiles are values that split **sorted data** into equal parts. In general terms, a *q*-quantile divides sorted data into *q* parts.

- **Quartiles (4-quantiles)**: Three quartiles split the data into four parts.
- **Deciles (10-quantiles)**: Nine deciles split the data into 10 parts.
- **Percentiles (100-quantiles):** 99 percentiles split the data into 100 parts.

Example: customer satisfaction ratings

20 measurements on the 10 point scale:

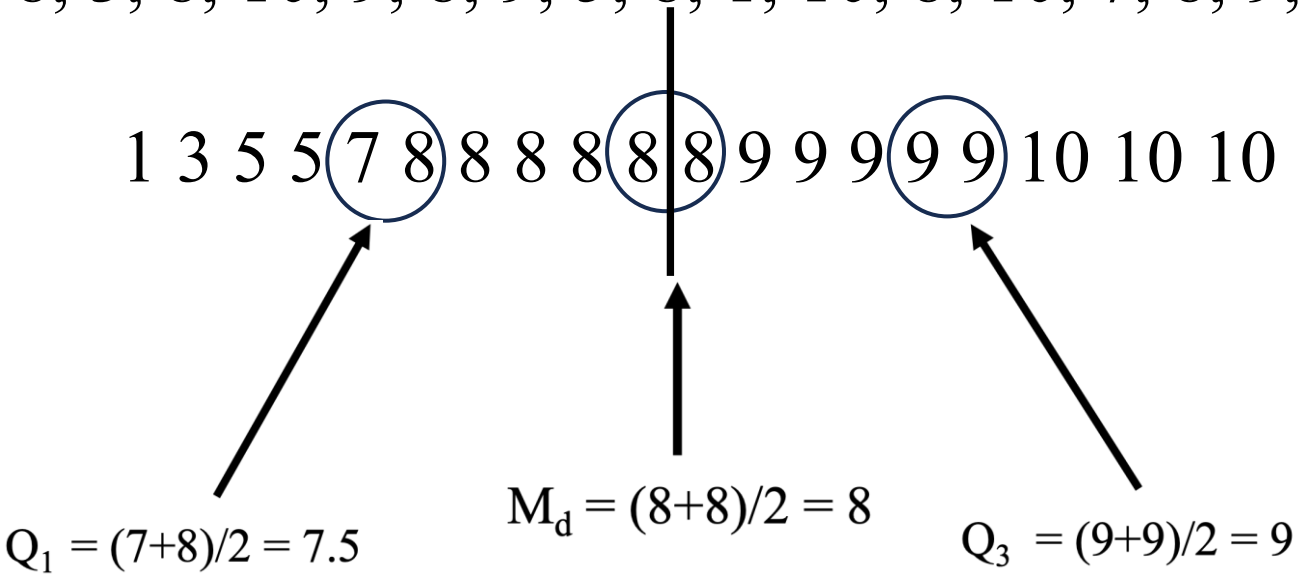
9, 8, 3, 8, 10, 9, 8, 9, 5, 8, 1, 10, 8, 10, 7, 8, 9, 10, 5, 9

- Question: Calculate the IQR for this dataset.

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



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

IQR and SD(s)

IQR

- Robust measure (same as the median)
- Has the same units as the observations

SD (s)

- NOT a robust measure (same as the mean)
- Has the same units as the observations
- s = 0 if and only if all the observations are equal

When IQR = 0?

How did John Tukey present a distribution?

- Hinges and 5-number summaries.

Example: 13 ordered numbers,

1, 1, 2, 2, 3, 6, 8, 11, 11, 13, 14, 16, 26

- Median: 8
- First quartile: 2
- Third quartile: 13

} hinges

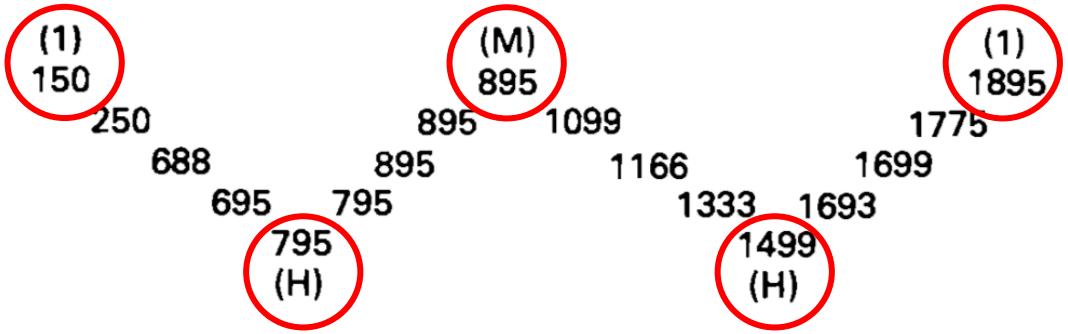
- Minimum: 1
- Maximum: 26

Hinges and 5-number summaries

- Tukey's down-up-down-up pattern.

Hinges illustrated

A) The 17 AUTO PRICES of EXHIBIT 1--in folded form

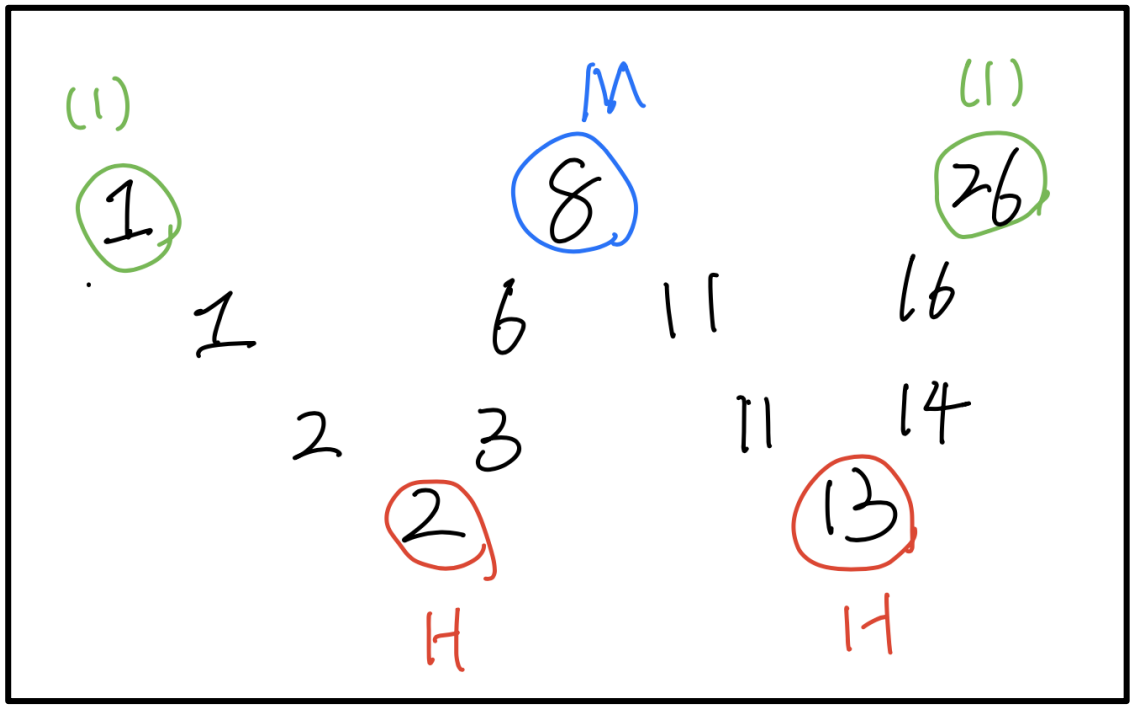


[17 prices, 1HMH1: 150, 795, 895, 1499, 1895 dollars]

5-number summaries

- Our example:

1, 1, 2, 2, 3, 6, 8, 11, 11, 13, 14, 16, 26



5-number summaries

- Tukey's table to recording numeric summaries

#13

M7	1.5	
H4	0.1	3.0
1	−3.2	9.8

- Our example:

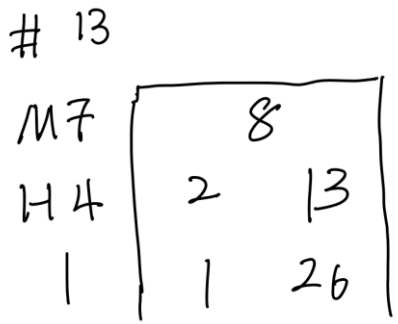
1, 1, 2, 2, 3, 6, 8, 11, 11, 13, 14, 16,26

13

M7	8	
H4	2	13
1	1	26

Today's 5(or 6)-number summaries

- John Tukey style:

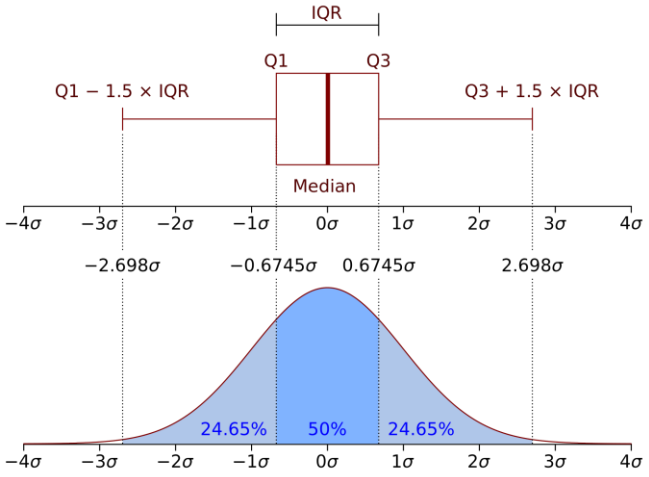


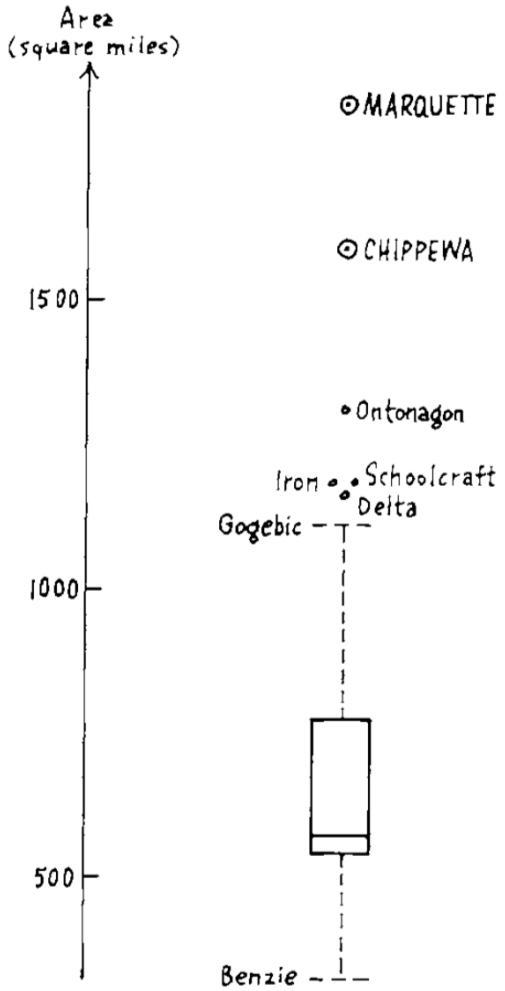
- Today (R code version):

```
> summary(c(1,1, 2, 2, 3, 6, 8, 11, 11, 13, 14,16, 26))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   8.000   8.769  13.000  26.000
```

Box-and-whisker display (boxplot)

- A box-plot usually includes two parts, a box and a set of whiskers, thus, the plot is also called the **box-and-whisker plot**.
- Invented by John Tukey in 1969.



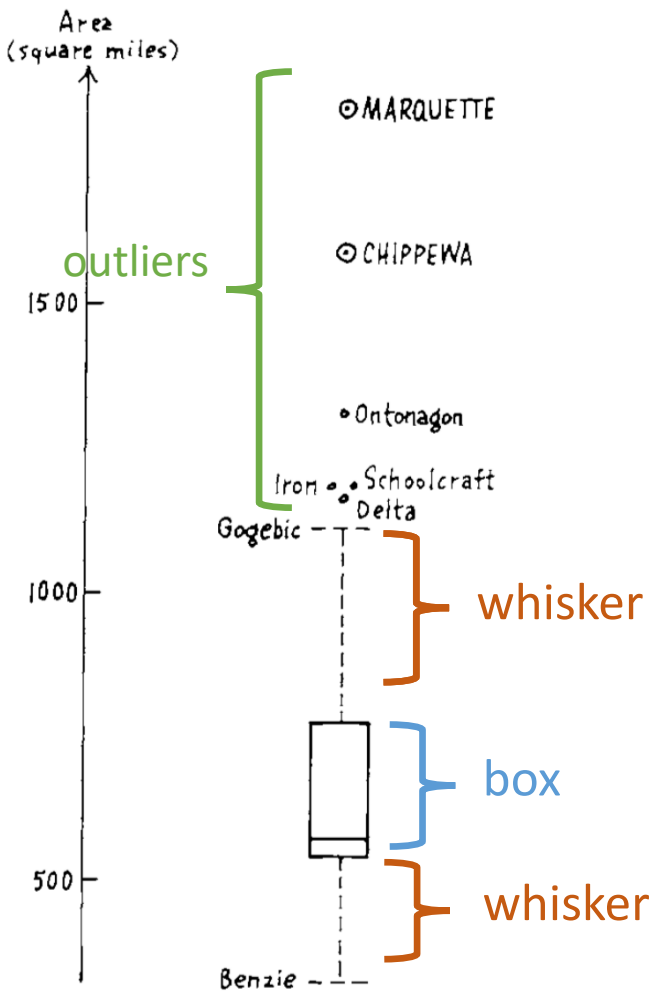


Box-and-whisker display (boxplot)

1. Box
   - Lower : first quartile (Q_1)
   - Middle: Median (M_d)
   - Upper: third quartile (Q_3)

2. Whisker
   - Lower: the larger value between Q_1 – 1.5*IQR and minimum
   - Upper: the smaller value between Q_3 + 1.5*IQR and maximum

3. Outliners: values not covered by whisker



Box-and-whisker display (boxplot)

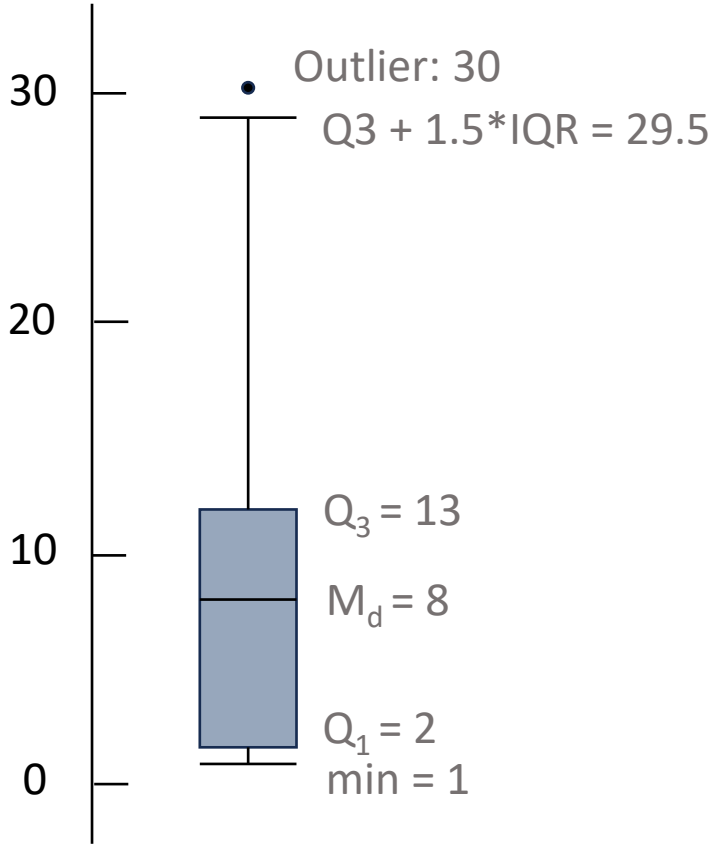
Example: 1, 1, 2, 2, 3, 6, 8, 11, 11, 13, 14, 16, 30

Use the previous table

# 13		
M 7	8	
H 4	2	13
1	1	30

Compute IQR = 13 – 2 = 11

We can get the following plot



Choosing a Summary

- For a skewed distribution or a distribution with strong outliers the five number summary is usually better than just mean and SD
- Use SD for the spread when you use the mean for the center

WARNING: Do not use only boxplots and numerical summaries to describe the shape of a distribution. Add a histogram. Why? (e.g., bimodal distribution)

Relationship between two variables ——**multivariate analysis**

- So far we have focused on individual variables

- Now we will study relationship between two variables
  - Two categorical variables
  - Two numerical variables

multivariate analysis for categorical variables

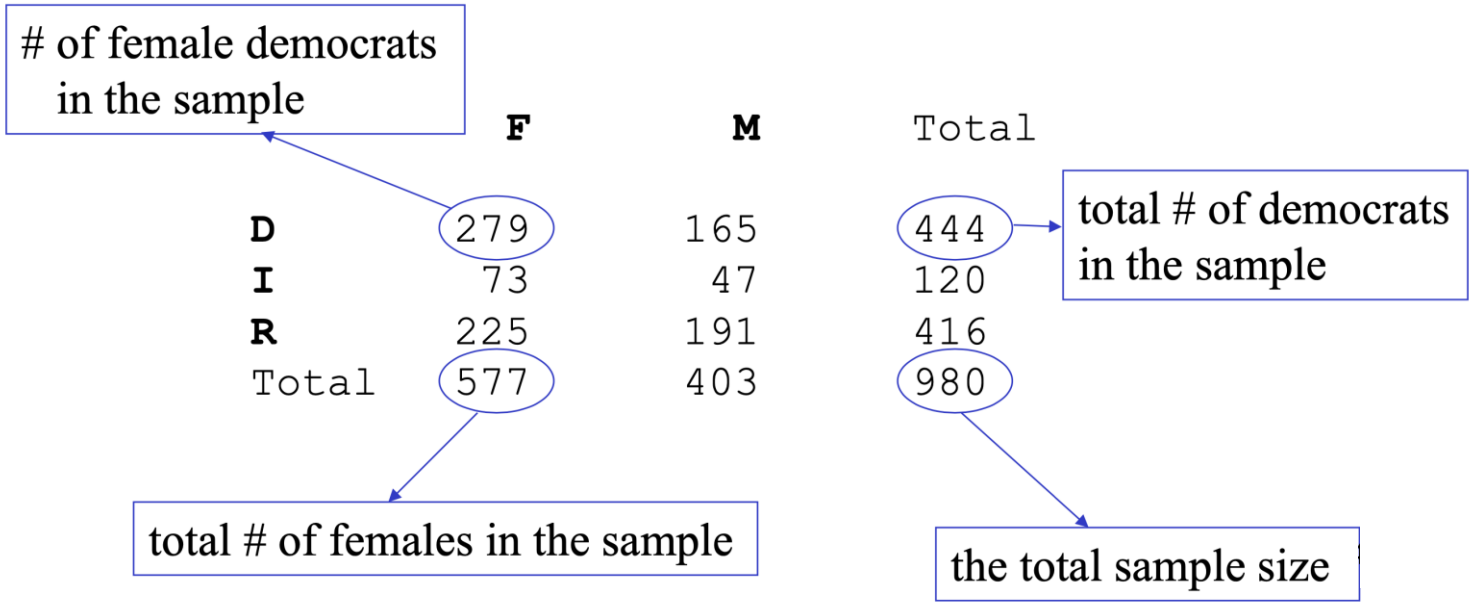
——Two-way (i.e., contingency) tables

- **Contingency Table** are used to describe the relationship between two categorical variables. The tables contain counts or proportions (percentages).

Two-way tables

- E.g., Cross-classification of a sample of 980 Americans by gender and party identification
- Variables:

rows: Party (D,I,R) *columns: Gender (F,M)*



Notation

	F	M	Total
D	279	165	444
I	73	47	120
R	225	191	416
Total	577	403	980

Party = **row variable**
Gender = **column variable**

Each combination of values of the two variables = **cell**

What is the total # of cells in the above table?

Joint distribution

- A two-way table with proportions (or percentages) describes the ***joint distribution*** of the two variables.
- Each cell gives the proportion of the total sample size

	F	**M**	Total
D	279	165	444
I	73	47	120
R	225	191	416
Total	577	403	980

	F	**M**	Total
D	28.5%	16.8%	45.3%
I	7.4%	4.8%	12.2%
R	23.0%	19.5%	42.5%
Total	58.9%	41.1%	100.0%

Marginal distribution

- Distribution of a single variable in a two-way table = ***marginal distribution***

rows: "Party"	columns: "Gender"		
	F	**M**	Total
D	28.5%	16.8%	45.3%
I	7.4%	4.8%	12.2%
R	23.0%	19.5%	42.5%
Total	58.9%	41.1%	100.0%

	"Party"	"Gender"	
D	45.3%	**F**	58.9%
I	12.2%	**M**	41.1%
R	42.5%		

marginal distribution of "**Party**"

Conditional distribution

distribution of one variable after we condition on (i.e., restrict our attention to) the value of the other variable = ***conditional distribution***

E.g., What is the distribution (in our sample of 980) of party identification conditional on Gender = F **?**

	F	**M**	Total
D	279	165	444
I	73	47	120
R	225	191	416
Total	577	403	980

Conditional distribution

- What is the distribution (in our sample of 980) of party identification conditional on Gender = F **?**

	F	**M**	Total
D	279	165	444
I	73	47	120
R	225	191	416
Total	577	403	980

D	279/577	**D**	48.4%
I	73/577	**I**	12.6%
R	225/577	**R**	39.0%

Another Example

Two-way table:

	Hospital A	Hospital B
Died	300	50
Survived	2700	950

Death status distributions are specified by the % died:

Hospital A: 300/3000=10%
Hospital B: 50/1000 = 5%

Is Hospital A worse?

Another Example

	Hospital A	Hospital B
Died	300	50
Survived	2700	950
Died:	10%	5%

In fact, patients have two condition types.

Good condition

	Hospital A	Hospital B
Died	5	10
Survived	995	800
Died:	0.5%	1.2%

Bad condition

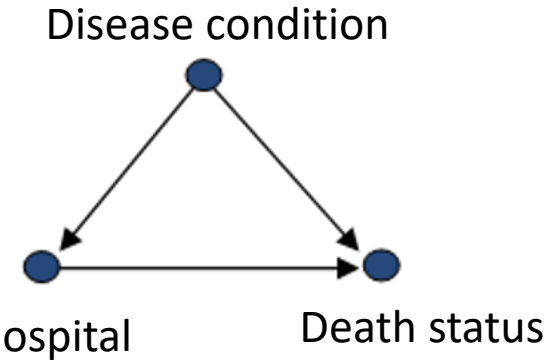
	Hospital A	Hospital B
Died	295	40
Survived	1705	150
Died:	14.8%	21.1%

Simpson's paradox

Association between two variables has a different direction from the association conditional on a third variable (**lurking variable (hidden) or confounding variable (considered)**)

What is the lurking variable in our example?

	Hospital A	Hospital B
Died	300	50
Survived	2700	950
Died:	10%	5%



Good condition

	Hospital A	Hospital B
Died	5	10
Survived	995	800
Died:	0.5%	1.2%

Bad condition

	Hospital A	Hospital B
Died	295	40
Survived	1705	150
Died:	14.8%	21.1%

Simpson's Paradox

- A change in the direction of association between two variables when data are separated into groups defined by a third variable
- Berkeley Sex discrimination case: https://homepage.stat.uiowa.edu/~mbognar/1030/Bickel-Berkeley.pdf

Department	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	933	64%	**825**	62%	108	82%
B	585	63%	**560**	63%	25	68%
C	918	35%	325	37%	**593**	34%
D	792	34%	417	33%	375	35%
E	584	25%	191	28%	**393**	24%
F	714	6%	373	6%	341	7%
Total	**4526**	**39%**	**2691**	**45%**	**1835**	**30%**

Legend:

greater percentage of successful applicants than the other gender

greater number of applicants than the other gender

bold - the two 'most applied for' departments for each gender

multivariate analysis for categorical variables

——Scatterplot

- Shows the relationship between **two numerical variables** measured on the same individuals
- The values of one variable -> horizontal axis
- The values of the other variable -> vertical axis
- Each individual (observation) appears as a point in the plot

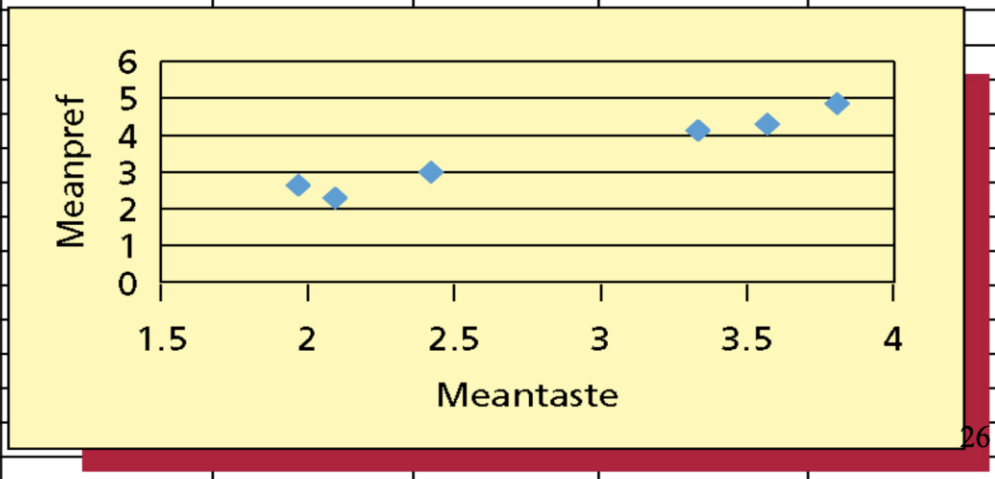
To add a categorical variable to the scatterplot, you can use a different color or symbol for each category

https://seaborn.pydata.org/tutorial/categorical.html

Scatter plot

Restaurant Ratings: Mean Preference vs Mean Taste

	A	B	C	D	E	F
1	Restaurant	Meantaste	Meanconv	Meanfam	Meanprice	Meanpref
2	Borden Burger	3.5659	2.7005	2.5282	2.9372	4.2552
3	Hardee's	3.329	3.3483	2.7345	2.7513	4.0911
4	Burger King	2.4231	2.7377	2.3368	3.0761	3.0052
5	McDonald's	2.0895	1.938	1.4619	2.4884	2.2429
6	Wendy's	1.9661	2.892	2.3376	4.0814	2.5351
7	White Castle	3.8061	3.7242	2.6515	1.708	4.7812
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						



Scatterplot

What can we see from a scatterplot?

- **form:** clusters, linear association, etc.
- **direction:** positive association, negative association.
- **strength:** how close the data points follow the form.

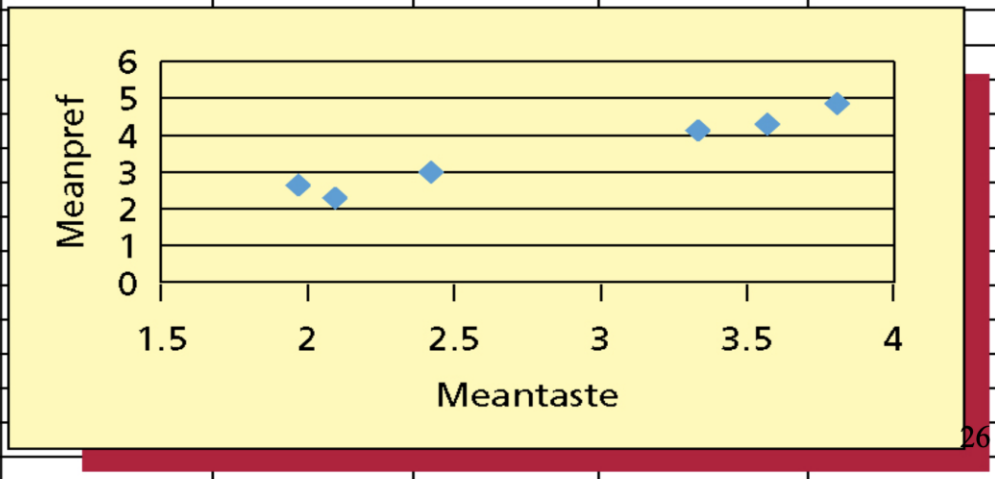
- Positive association: above-average values of one variable accompany above-average values of the other, and below-average values also tend to occur together.

- Negative association: above-average values of one variable accompany below-average values of the other and vice versa

Scatter plot

Restaurant Ratings: Mean Preference vs Mean Taste

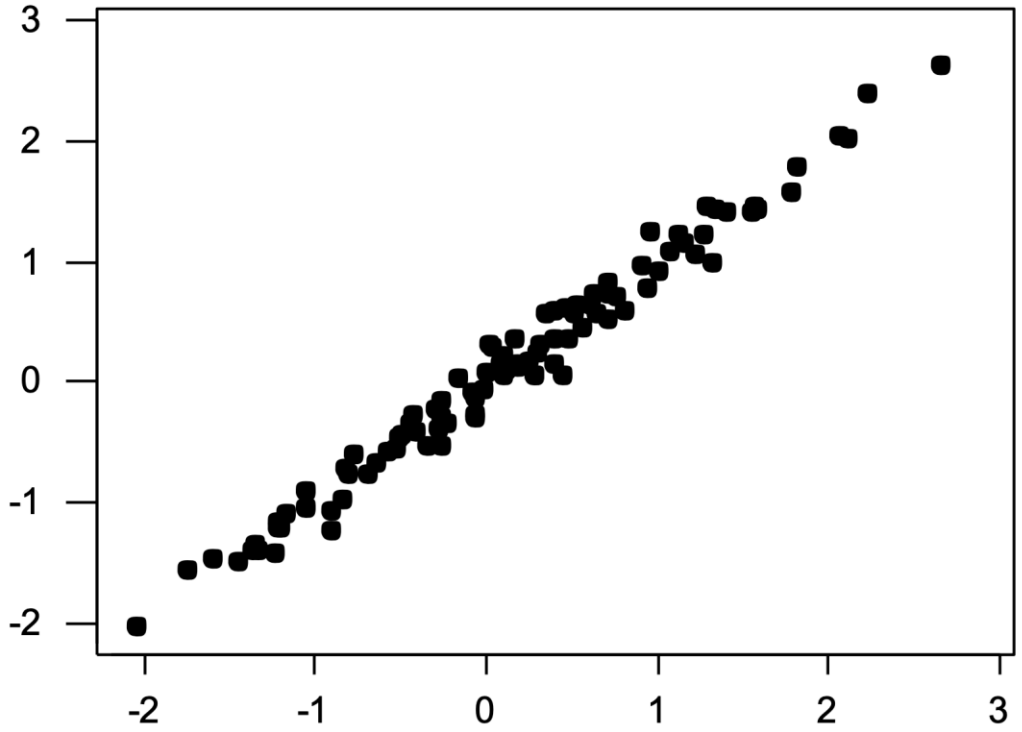
	A	B	C	D	E	F
1	Restaurant	Meantaste	Meanconv	Meanfam	Meanprice	Meanpref
2	Borden Burger	3.5659	2.7005	2.5282	2.9372	4.2552
3	Hardee's	3.329	3.3483	2.7345	2.7513	4.0911
4	Burger King	2.4231	2.7377	2.3368	3.0761	3.0052
5	McDonald's	2.0895	1.938	1.4619	2.4884	2.2429
6	Wendy's	1.9661	2.892	2.3376	4.0814	2.5351
7	White Castle	3.8061	3.7242	2.6515	1.708	4.7812
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						



Correlation (r)

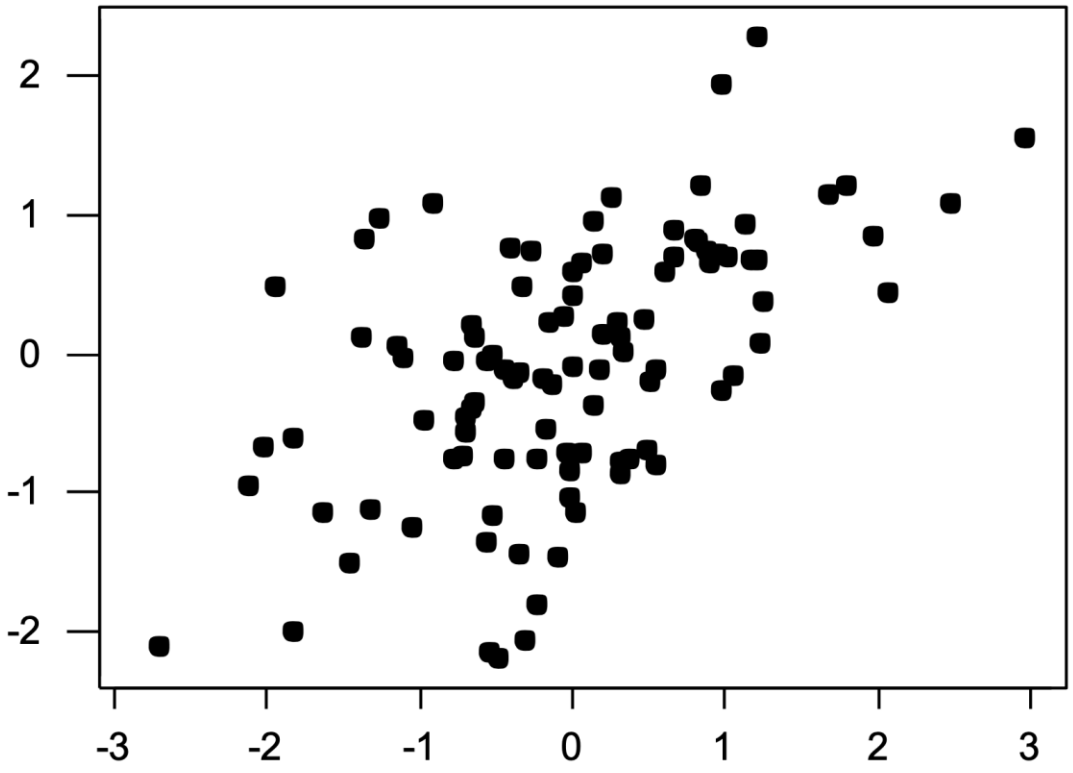
- Measures the **direction** and **strength** of the **linear relationship** between two numerical variables
- Is always between –1 and 1
- The strength increases as you move away from 0 to either –1 or 1

Highly correlated variables



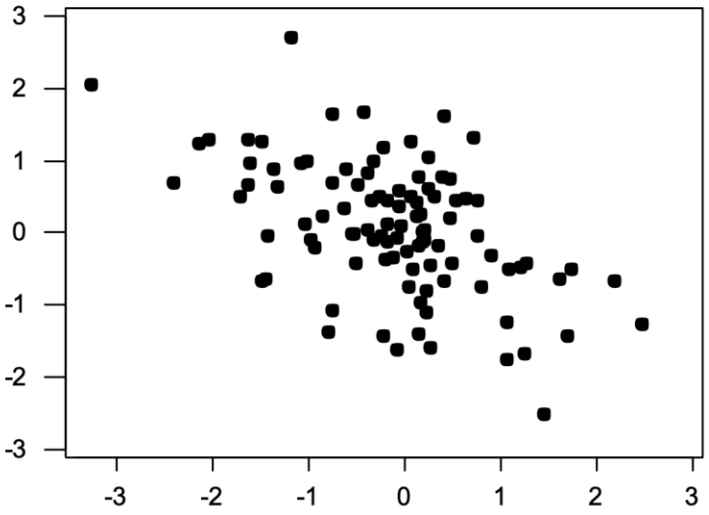
$r = 0.99$

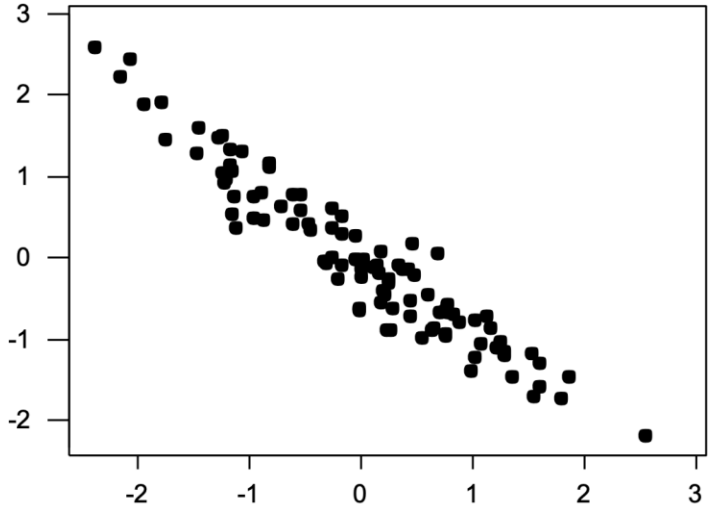
Moderate correlation



$r = 0.55$

Negative correlation

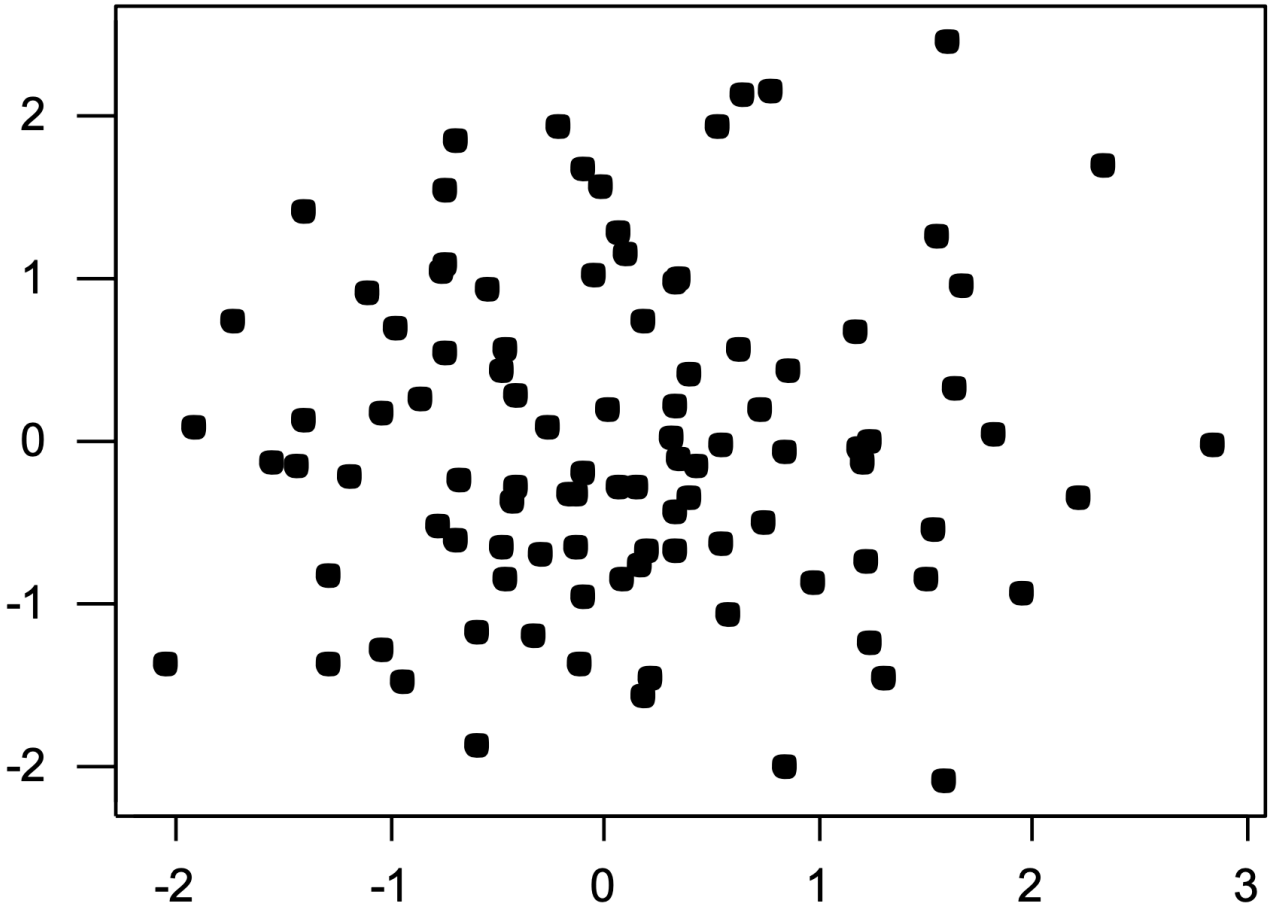




$r = -0.52$

$r = -0.96$

Zero correlation



Notes about correlation

- Both variables have to be numerical
- r has no units of measurement
- r does not change if you change the units of measurement of the data (e.g., from *lbs* to *kg*)

Notes about correlation

- $r > 0$ indicates positive association between the variables, $r < 0$ indicates negative association

- Extremes $r = -1$ and $r = 1$occur if and only if the points on a scatterplot lie exactly along a straight line

- r measures the strength of only the linear relationship, it does not describe curved relationships

- r is not robust

Correlation does not imply causation

1. What is data
    - Observations
    - Categorical Variables
    - Numerical Variables

2. EDA

	Categorical Variables	**Numerical Variables**
Univariate Analysis	Count/Percentage	Mean, Median, Mode, Quantiles, Variance, SD
	Bar chart/Pie chart	Histogram & Density Curve, Box plot
Multivariate Analysis	Two-way tables	Correlation
		Scatter Plot