Statistics applied to Management.

Franck JAOTOMBO

Outline & Program

- This is a 30h00 course over several seminar days
  - Some sessions on basics – theory and exercises
  - Some sessions on excel – hands on practice and complements
- The topic is statistics applied to management
  - It is not mathematical statistics
  - It is applied statistics therefore there will be *some* applied maths
- You will be evaluated on the following:
  - 50% on a group project and/or individual homework
  - 50% on a final written exam
- Final exam:
  - Written on paper, calculators authorized (and one or two recto-verso memory aid)
  - No phone, no computer, no note books

Teaching / learning materials

- You must take notes during the sessions
  - The slides are useful & necessary but not sufficient
- Learn to read text books and complete your notes through your readings
  - Statistics for Business and Economics – 12th / 13th edition Mac Clave, Benson and Sincich, PEARSON
  - Méthodes statistiques appliquées au management 2e édition – Hahn & Macé, PEARSON
- There are usually more exercises on the intranet platform than we can cover during the sessions
  - I strongly recommend that you pick some of those and work on them in your spare times - if you want to maximize your chance of success at your exams.
- CAREFUL : Plagiarism
  - Disciplinary board + 0/20

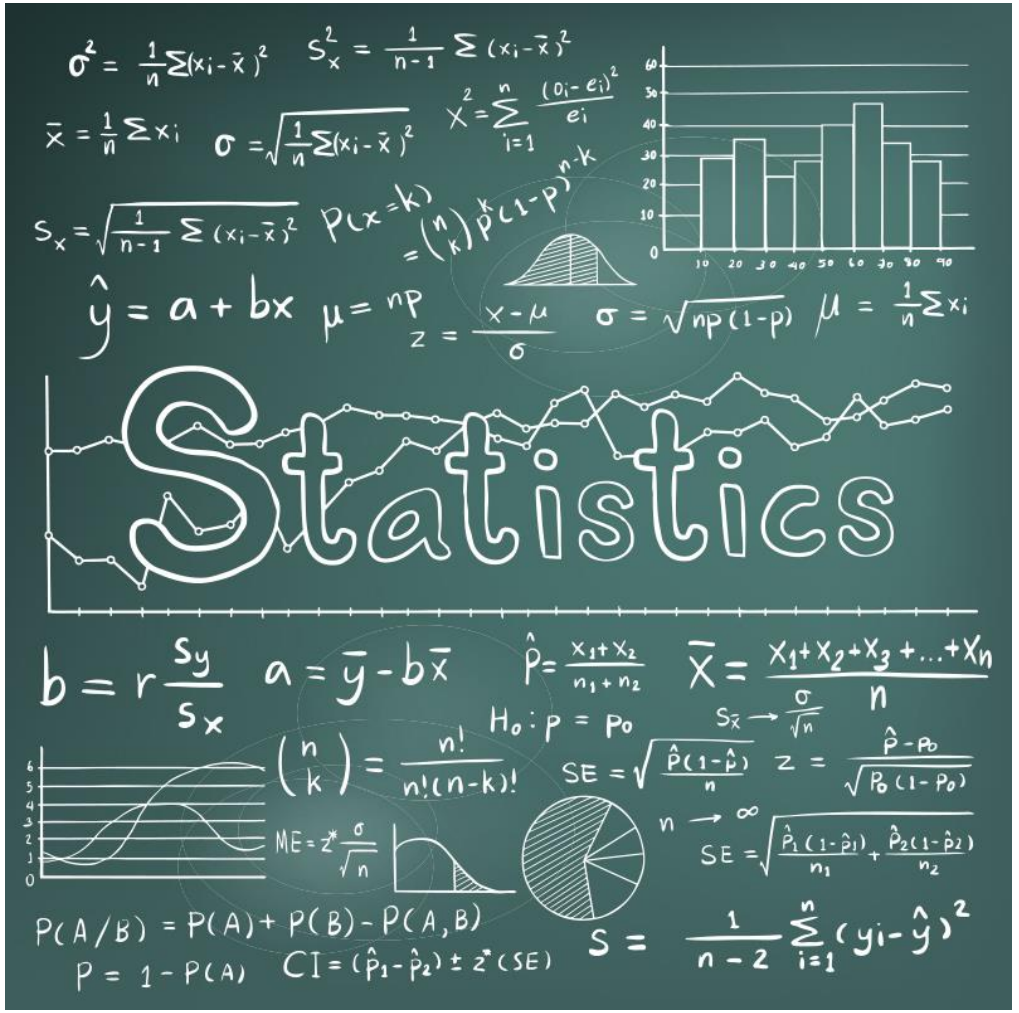
General Introduction

Classification of variables

WHY DO WE NEED STATISTICS ?

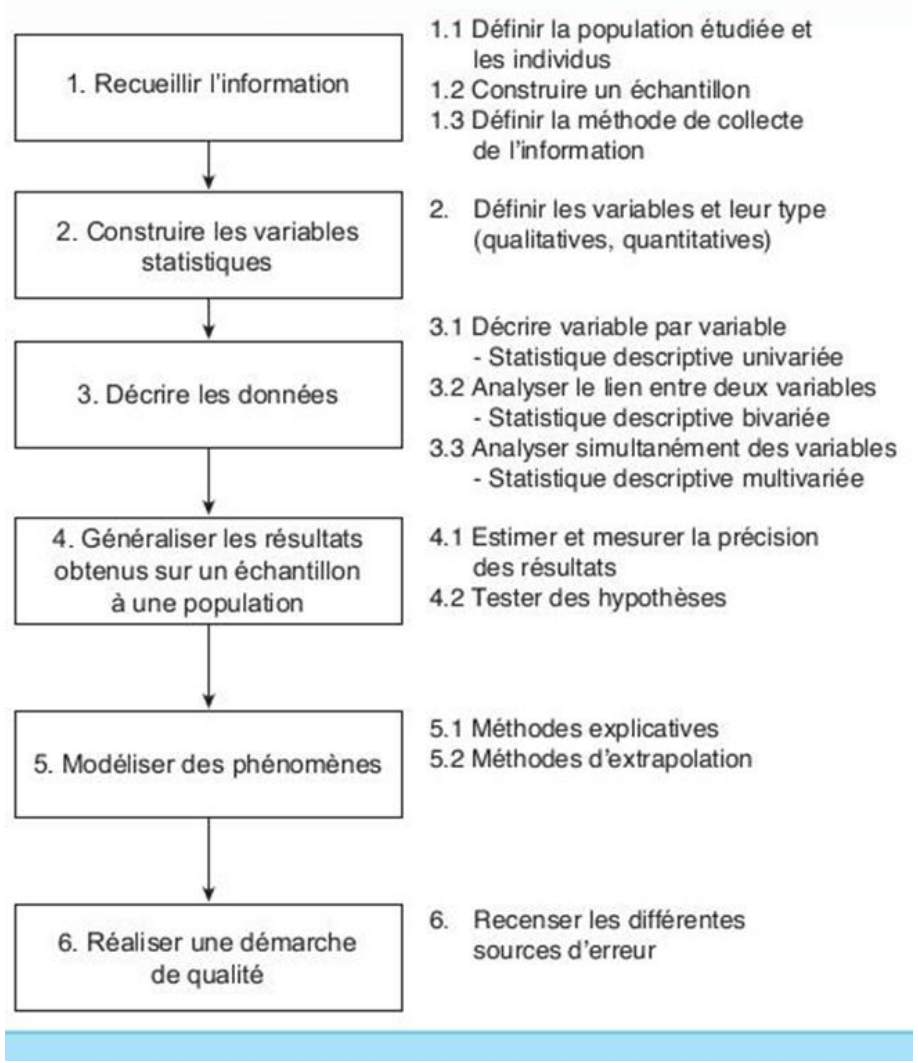
Making decision in an uncertain environment

- Running an organization (leading, managing, organizing…) is mostly about making decisions
  - Should we launch this new product on this market ?
- To make informed (wise) decisions, we need reliable information
  - Information is encapsulated within all sorts of data
  - Statistics is a tool to help process, summarize, analyze, and interpret data
- In sum : **statistics facilitates decision making**
- Another reason : this is the age of Artificial Intelligence
  - AI is largely based on statistics



Example : Application in Marketing

- Your company has recently created a new product
  - You are the marketing manager
  - Your boss wants to know if it is a good idea to sell the product in Europe or in South America
  - How would you use this process in order to answer his question ?



Descriptive and Inferential Statistics

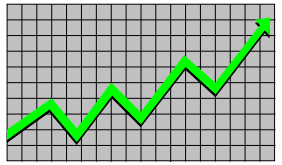
Two branches of statistics:

- Descriptive statistics
  - Graphical and numerical procedures to summarize and process data
- Inferential statistics
  - Using data to make predictions, forecasts, and estimates to assist decision making

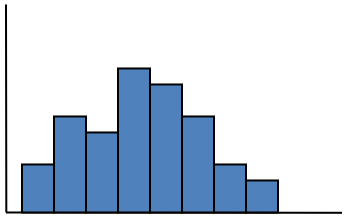
Descriptive Statistics

- Collect data
  - e.g., Survey

- Present data
  - e.g., Tables and graphs



- Summarize data
  - e.g., Sample mean = $\frac{\sum X_i}{n}$



Inferential Statistics

- Estimation
  - e.g., Estimate the population mean weight using the sample mean weight
- Hypothesis testing
  - e.g., Test the claim that the population mean weight is 70 kgs

Inference is the process of drawing conclusions or making decisions about a population based on sample results

Example : market study

- Your company has recently created a new product
  - You are the marketing manager
  - Your boss wants to know if it is a good idea to sell the product in South America
  - How would you proceed to answer his question ?

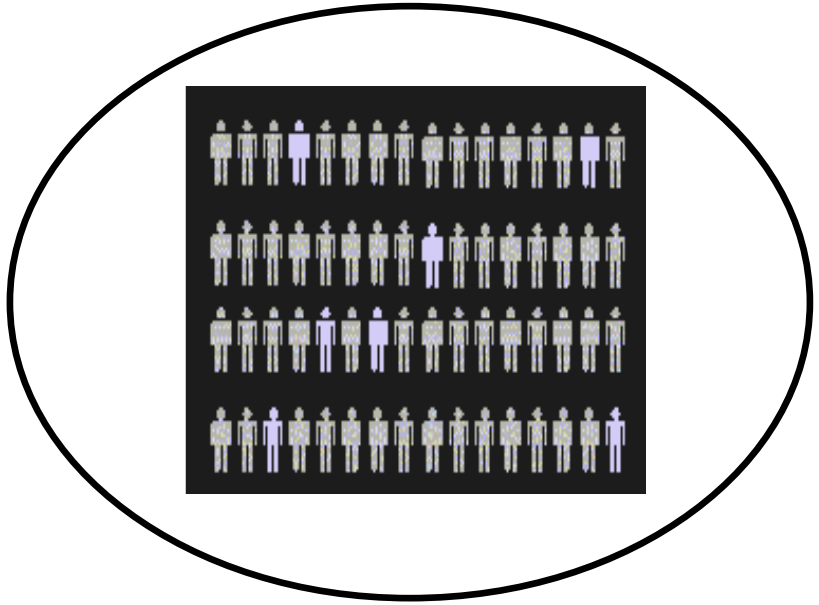
- Determine the target prospect (Population)
  - Select an appropriate subset (Sample)
  - Select the characteristics of interest (Variables)
    - Age, Gender, Income, Occupation, Purchasing Intention…
- Describe these characteristics
  - Descriptive Statistics
    - What is the mean age of the target ?
    - Can we visualize the income distribution of the respondents?
    - Is there a link between income and purchasing intention ?
- Explore models explaining most accurately the relationships between these characteristics
  - Inferential Statistics
    - Which characteristics explain the best the purchasing intentions of the target ?
    - Can I accurately predict the purchasing intention of each prospect (statistical unit / observation) from her profile ?

Basic Vocabulary of Statistics

- VARIABLES
  - Variables are characteristics of an item or individual; they are what you analyze when you use a statistical method.
- DATA
  - Data are the different values associated with a variable.
- OPERATIONAL DEFINITIONS
  - Data values are meaningless unless their variables have operational definitions, universally accepted meanings that are clear to all associated with an analysis.
- POPULATION
  - A population consists of all the items or individuals about which you want to draw a conclusion. The population is the "large group"
- SAMPLE
  - A sample is the portion of a population selected for analysis. The sample is the "small group"
- PARAMETER
  - A parameter is a numerical measure that describes a characteristic of a population.
- STATISTIC
  - A statistic is a numerical measure that describes a characteristic of a sample.

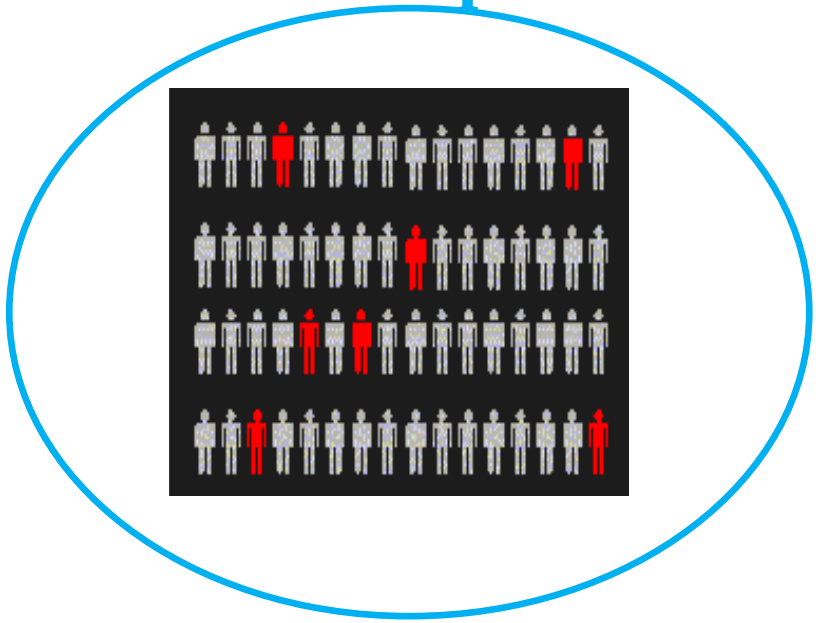
Population vs. Sample

Population



Measures used to describe the population are called **parameters**

Sample



Measures used to describe the sample are called **statistics**

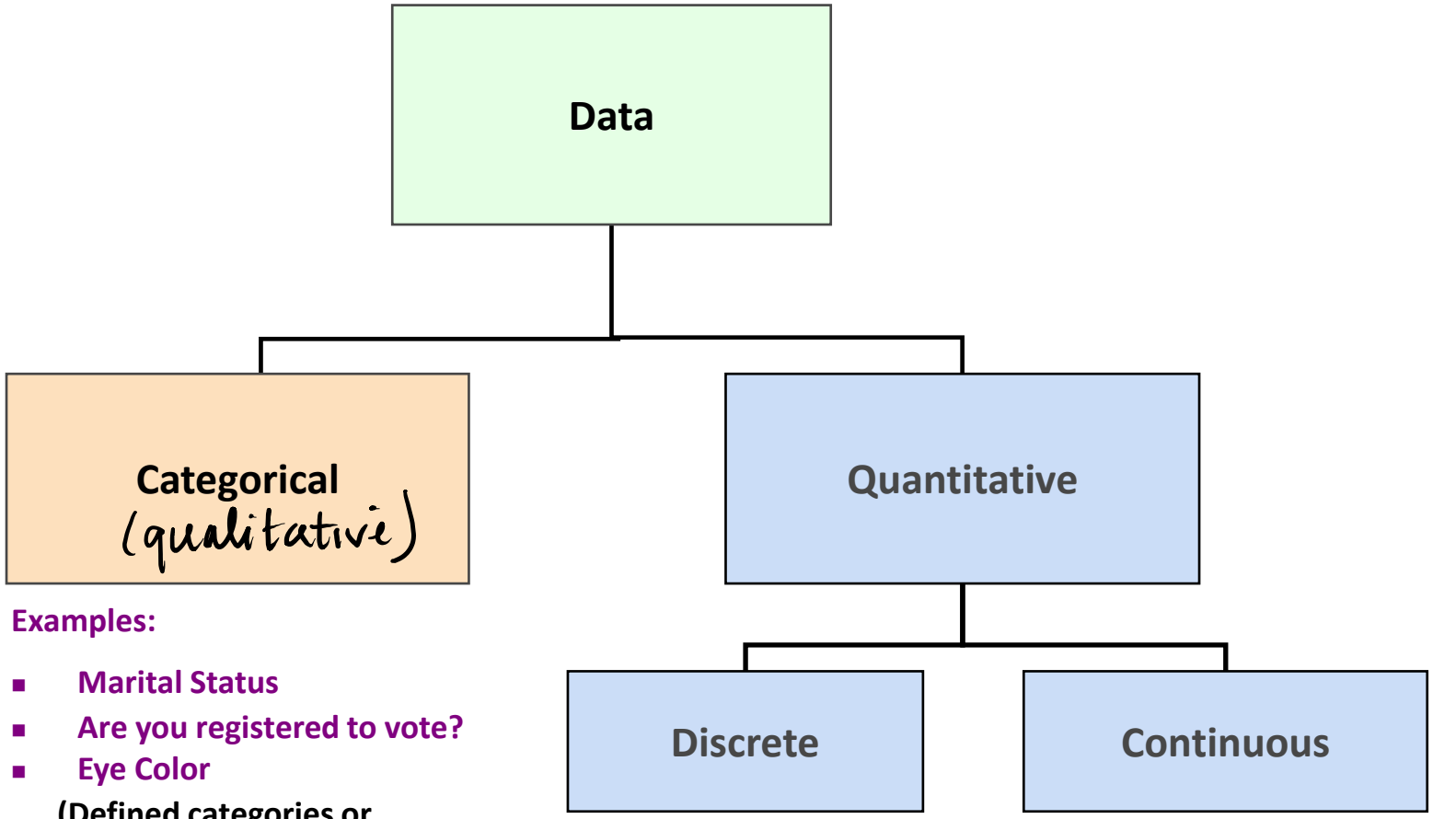
Types of Data Analysis

- Quantitative Methods (structured data)
  - Testing theories using numbers
- Qualitative Methods (unstructured data)
  - Testing theories using language
    - Magazine articles/Interviews
    - Conversations
    - Newspapers
    - Media broadcasts
- In this course we will address only quantitative methods.

What makes a data set ?

- The population
  - The set of (all) the individual units under study
- A statistical unit (or individual)
  - A unit of the population
  - Usually a line (a row) in the data set
- Sample
  - A subset of the population (preferably representative)
- Variables
  - What is being measured
  - Usually a column in the data base

Classification of Variables



Examples:

- Marital Status
- Are you registered to vote?
- Eye Color

(Defined categories or groups)

Examples:

- Number of Children
- Defects per hour

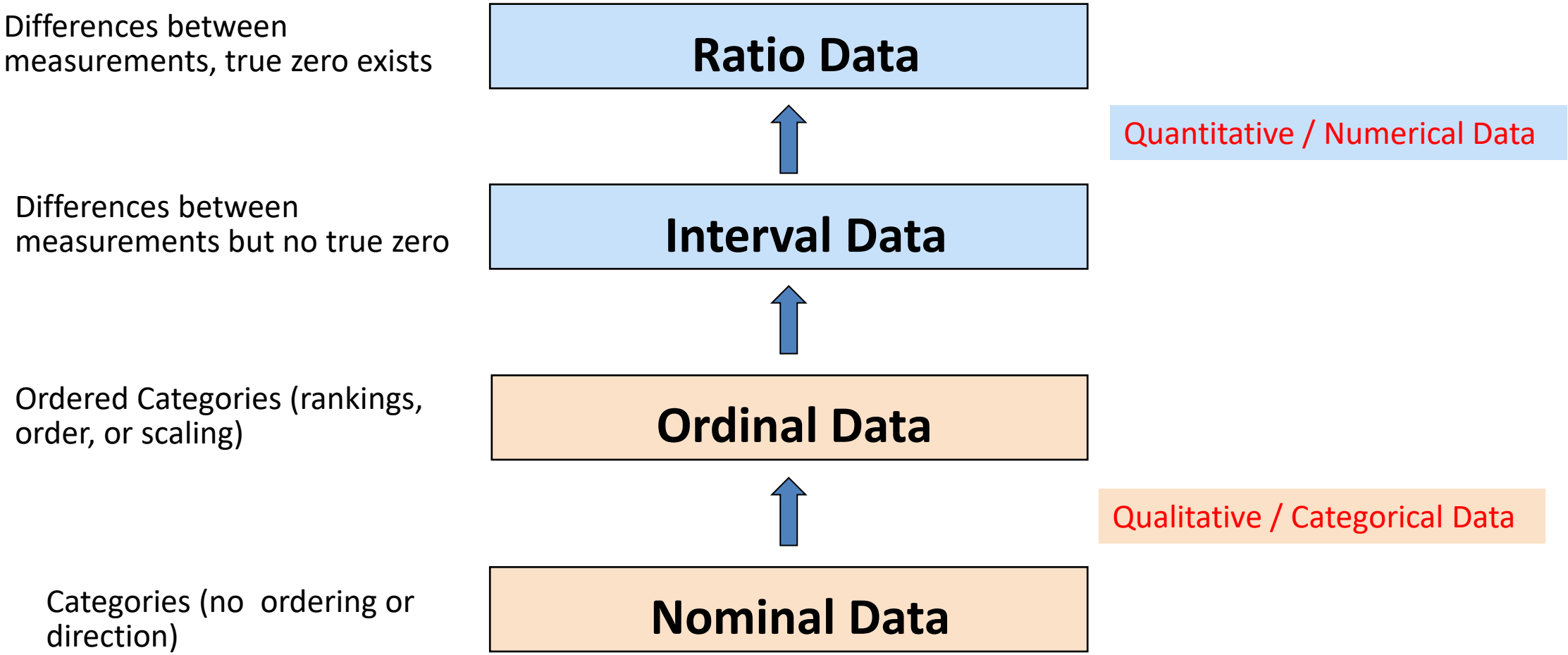
(Counted items)

Examples:

- Weight
- Voltage

(Measured characteristics)

Measurement Levels



Levels of Measurement

- Categorical (entities are divided into distinct categories):
  - Binary variable: There are only two categories
    - e.g. dead or alive.
  - Nominal variable: There are more than two categories
    - e.g. whether someone is an omnivore, vegetarian, vegan, or fruitarian.
  - Ordinal variable: The same as a nominal variable but the categories have a logical order
    - e.g. whether people got a fail, a pass, a merit or a distinction in their exam.
- Numerical (entities get a distinct score):
  - Interval variable: Equal intervals on the variable represent equal differences in the property being measured
    - e.g. the difference between 6 and 8 is equivalent to the difference between 13 and 15.
    - Zero value is relative
  - Ratio variable: The same as an interval variable, but the ratios of scores on the scale must also make sense
    - e.g. a score of 16 on an anxiety scale means that the person is, in reality, twice as anxious as someone scoring 8.
    - Zero value is absolute

Table 1.1 ♦ Levels of Measurement, Arithmetic Operations, and Types of Statistics

Stevens's Levels of Measurement	*Logical and Arithmetic Operations That Can Be Applied (According to Stevens)*	*Traditional or Conservative Recommendation*	*Simpler Distinction Between Two Types of Variables*[a]
Nominal	=, ≠	Only nonparametric statistics	Categorical
Ordinal	=, ≠, <, >	Only nonparametric statistics	Quantitative
Interval[b]	=, ≠, <, >, +, −	Parametric statistics	Quantitative
Ratio	=, ≠, <, >, +, −, ×, ÷	Parametric statistics	Quantitative

Exercise 1-1

- What are the variables and measurement levels for the following tables ?

	No Disorder	Disorder	Total
Selected	3	9	12
Rejected	6805	845	7650
Total	6808	854	7662

Table 1

AL	35.5	HI	43	MA	45.5	NM	35.7	SD	37.9
AK	55.6	ID	42.3	MI	42.1	NY	39.8	TN	36.3
AZ	42.5	IL	40.1	MN	43	NC	35.3	TX	38.3
AR	26.5	IN	39.4	MS	26.3	ND	37.7	UT	48.4
CA	46.7	IA	39	MO	42.5	OH	40.7	VT	46.7
CO	51.8	KS	43.9	MT	40.6	OK	34.3	VA	44.3
CT	51.2	KY	36.6	NE	37	OR	50.8	WA	49.7
DE	50.7	LA	30.2	NV	41	PA	40.1	WV	34.3
FL	43.2	ME	42.6	NH	56	RI	38.8	WI	40.6
GA	38.3	MD	43.8	NJ	47.8	SC	32	WY	44.1

Table 2

Exercise 1-2

- What are the level of measurement of the following variables
  - Personal Computer Ownership (Yes / No)
  - Student class designation (Freshman, Sophomore, Junior, Senior)
  - Product satisfaction (Satisfied, Neutral, Unsatisfied)
  - Type of Stocks Owned (Growth / Value / Other)
  - Internet Provider (Microsoft Network / AOL / Other)
  - Standard & Poor's bond ratings (AAA, AA, A, BBB, BB, B, CCC, CC, C, DDD, DD, D)
  - The number of students in each class of a school
  - The number of stars in the sky
  - The distance between the planets
  - The income of the French households
  - Each passing year

Exercise 1-3

- Chemical and manufacturing plants sometimes discharge toxic-waste materials such as DDT into nearby rivers and streams. These toxins can adversely affect the plants and animals inhabiting the river and the riverbank. The U.S. Army Corps of Engineers conducted a study of fish in the Tennessee River (in Alabama) and its three tributary creeks: Flint Creek, Limestone Creek, and Spring Creek. A total of 144 fish were captured, and the following variables were measured for each:
- Classify each of the five variables measured as quantitative or qualitative.
  - 1. River/creek where each fish was captured
  - 2. Species (channel catfish, largemouth bass, or smallmouth buffalo fish)
  - 3. Length (centimeters)
  - 4. Weight (grams)
  - 5. DDT concentration (parts per million)
  - 6. Mile (The distance where the fish was caught)

RIVER	MILE	SPECIES	LENGTH	WEIGHT	DDT
FCM	5	CCATFISH	42.5	732	10
FCM	5	CCATFISH	44	795	16
FCM	5	CCATFISH	41.5	547	23
FCM	5	CCATFISH	39	465	21
FCM	5	CCATFISH	50.5	1252	50
FCM	5	CCATFISH	52	1255	150
LCM	3	CCATFISH	40.5	741	28
LCM	3	CCATFISH	48	1151	7.7
LCM	3	CCATFISH	48	1186	2

Tables and Frequency Distributions

REPRESENTING DATA WITH TABLES

Organizing Qualitative Data: Summary Table

- A **summary table** indicates the frequency (amount) or percentage of items in a set of categories so that differences between categories can be seen.

Summary Table From A Survey of 1000 Banking Customers

Banking Preference?	Percent
ATM	16%
Automated or live telephone	2%
Drive-through service at branch	17%
In person at branch	41%
Internet	24%

Exercise 2-1 (PBL data)

Interviewee	Position	Organization	Experience
1	VicePresident	Commercial	30
2	Postproduction	Government	15
3	Analyst	Commercial	10
4	SeniorManager	Government	30
5	SupportChief	Government	30
6	Specialist	Government	25
7	SeniorAnalyst	Commercial	9
8	DivisionChief	Government	6
9	ItemManager	Government	3
10	SeniorManager	Government	20
11	MROManager	Government	25
12	LogisticsMgr.	Government	30
13	MROManager	Commercial	10
14	MROManager	Commercial	5
15	MROManager	Commercial	10
16	Specialist	Government	20
17	Chief	Government	25

- What types of data are there in this data set ?
- Tally the categorical variables and generate a summary table with their frequencies and their relative frequency
- Do this manually and check with MS Excel

Summarizing 2 categorical variables : Contingency Table

- A random sample of 400 invoices is drawn.
- Each invoice is categorized as a small, medium or large amount.
- Each invoice is also examined to identify if there are any errors.
- This data are then organized in the contingency table to the right.

Contingency Table Showing Frequency of Invoices Categorized By Size and The Presence Of Errors

	No Errors	Errors	Total
Small Amount	170	20	190
Medium Amount	100	40	140
Large Amount	65	5	70
Total	335	65	400

Exercise 2-2

- From the previous invoice contingency table :
  - Generate a percentage contingency table based on total score
  - Generate a percentage contingency table based on row (line) total score
  - Generate a percentage contingency table based on column total score

Contingency Table Based On Percentage Of Overall Total

	No Errors	Errors	Total
Small Amount	170	20	190
Medium Amount	100	40	140
Large Amount	65	5	70
Total	335	65	400

42.50% = 170 / 400
25.00% = 100 / 400
16.25% = 65 / 400

	No Errors	Errors	Total
Small Amount	42.50%	5.00%	47.50%
Medium Amount	25.00%	10.00%	35.00%
Large Amount	16.25%	1.25%	17.50%
Total	83.75%	16.25%	100.0%

83.75% of sampled invoices have no errors and 47.50% of sampled invoices are for small amounts.

Contingency Table Based On Percentage of Row Totals

	No Errors	Errors	Total
Small Amount	170	20	190
Medium Amount	100	40	140
Large Amount	65	5	70
Total	335	65	400

89.47% = 170 / 190
71.43% = 100 / 140
92.86% = 65 / 70

	No Errors	Errors	Total
Small Amount	89.47%	10.53%	100.0%
Medium Amount	71.43%	28.57%	100.0%
Large Amount	92.86%	7.14%	100.0%
Total	83.75%	16.25%	100.0%

Medium invoices have a larger chance (28.57%) of having errors than small (10.53%) or large (7.14%) invoices.

Contingency Table Based On Percentage Of Column Total

	No Errors	Errors	Total
Small Amount	170	20	190
Medium Amount	100	40	140
Large Amount	65	5	70
Total	335	65	400

50.75% = 170 / 335
30.77% = 20 / 65

	No Errors	Errors	Total
Small Amount	50.75%	30.77%	47.50%
Medium Amount	29.85%	61.54%	35.00%
Large Amount	19.40%	7.69%	17.50%
Total	100.0%	100.0%	100.0%

61.54% of invoices with errors are of medium size.

Exercise 2-3

- Chemical and manufacturing plants sometimes discharge toxic-waste materials such as DDT into nearby rivers and streams. These toxins can adversely affect the plants and animals inhabiting the river and the riverbank. The U.S. Army Corps of Engineers conducted a study of fish in the Tennessee River (in Alabama) and its three tributary creeks: Flint Creek, Limestone Creek, and Spring Creek. A total of 144 fish were captured, and the following variables were measured for each:
  - 1. River/creek where each fish was captured
  - 2. Species (channel catfish, largemouth bass, or smallmouth buffalo fish)
  - 3. Length (centimeters)
  - 4. Weight (grams)
  - 5. DDT concentration (parts per million)
  - 6. Distance from the confluent where the fish was caught
- These data are saved in the **DDT** file

- Generate the summary table of each categorical variable of the DDT file with MS Excel
- Generate the contingency table of the categorical variables in the DDT data, with MS Excel

Organizing Quantitative Data: Frequency Distribution

- The **frequency distribution** is a summary table in which the data are arranged into numerically ordered classes.

- You must pay attention to selecting the appropriate number of **class groupings** for the table, determining a suitable width of a class grouping, and establishing the boundaries of each class grouping to avoid overlapping.

- The number of classes depends on the number of values in the data. With a larger number of values, typically there are more classes. In general, a frequency distribution should have at least 5 but no more than 15 classes.

- To determine the **width of a class interval**, you divide the **range** (Highest value– Lowest value) of the data by the number of class groupings desired.

Construction of a Frequency Distribution

Rule 1: Determine k, the number of classes.

Rule 2: Classes should be the same width, w; the width is determined by the following:

$$w = \text{Class Width} = \frac{\text{Largest Observation} - \text{Smallest Observation}}{\text{Number of Classes}} \quad (1.1)$$

Always round class width, w, upward.

Rule 3: Classes must be inclusive and nonoverlapping.

Rule 1: Number of Classes

The number of classes used in a frequency distribution is decided in a somewhat arbitrary manner.

Rule 2: Class Width

After choosing the number of classes, the next step is to choose the class width:

$$w = \text{Class Width} = \frac{\text{Largest Observation} - \text{Smallest Observation}}{\text{Number of Classes}}$$

The class width must always be rounded upward in order that all observations are included in the frequency distribution table.

Rule 3: Inclusive and Nonoverlapping Classes

Classes must be inclusive and nonoverlapping. Each observation must belong to one and only one class. Consider a frequency distribution for the ages (rounded to the nearest year) of a particular group of people. If the frequency distribution contains the classes "age 20 to age 30" and "age 30 to age 40," to which of these two classes would a person age 30 belong?

Quick Guide to Approximate Number of Classes for a Frequency Distribution

Sample Size	*Number of Classes*
Fewer than 50	5–7
50 to 100	7–8
101 to 500	8–10
501 to 1,000	10–11
1,001 to 5,000	11–14
More than 5,000	14–20

Practice and experience provide the best guidelines. Larger data sets require more classes; smaller data sets require fewer classes. If we select too few classes, the patterns and various characteristics of the data may be hidden. If we select too many classes, we will discover that some of our intervals may contain no observations or have a very small frequency.

Exercise 2-4 & example

- A manufacturer of insulation randomly selects 20 winter days and records the daily high temperature

24, 35, 17, 21, 24, 37, 26, 46, 58, 30, 32, 13, 12, 38, 41, 43, 44, 27, 53, 27

  - Sort raw data in ascending order
  - Find range
  - Select number of classes (recommended: 5)
  - Compute class interval (width)
  - Determine class boundaries (limits)
  - Compute class midpoints
  - Count observations & assign to classes

Solutions

- Sort raw data in ascending order:
  - 12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58
- Find range: **58 - 12 = 46**
- Select number of classes: **5 (usually between 5 and 15)**
- Compute class interval (width): **10 (46/5 then round up)**
- Determine class boundaries (limits):
  - Class 1: 10 to less than 20
  - Class 2: 20 to less than 30
  - Class 3: 30 to less than 40
  - Class 4: 40 to less than 50
  - Class 5: 50 to less than 60
- Compute class midpoints: **15, 25, 35, 45, 55**
- Count observations & assign to classes

Organizing Quantitative Data: Frequency Distribution Example

Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Class	Midpoints	Frequency
10 but less than 20	15	3
20 but less than 30	25	6
30 but less than 40	35	5
40 but less than 50	45	4
50 but less than 60	55	2
Total		20

Relative & Percent Frequency Distribution Example

Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Class	Frequency	Relative Frequency	Percentage
10 but less than 20	3	.15	15
20 but less than 30	6	.30	30
30 but less than 40	5	.25	25
40 but less than 50	4	.20	20
50 but less than 60	2	.10	10
Total	20	1.00	100

Cumulative Frequency Distribution Example

Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Class	Frequency	Percentage	Cumulative Frequency	Cumulative Percentage
10 but less than 20	3	15%	3	15%
20 but less than 30	6	30%	9	45%
30 but less than 40	5	25%	14	70%
40 but less than 50	4	20%	18	90%
50 but less than 60	2	10%	20	100%
Total	20	100	20	100%

Exercise 2-5

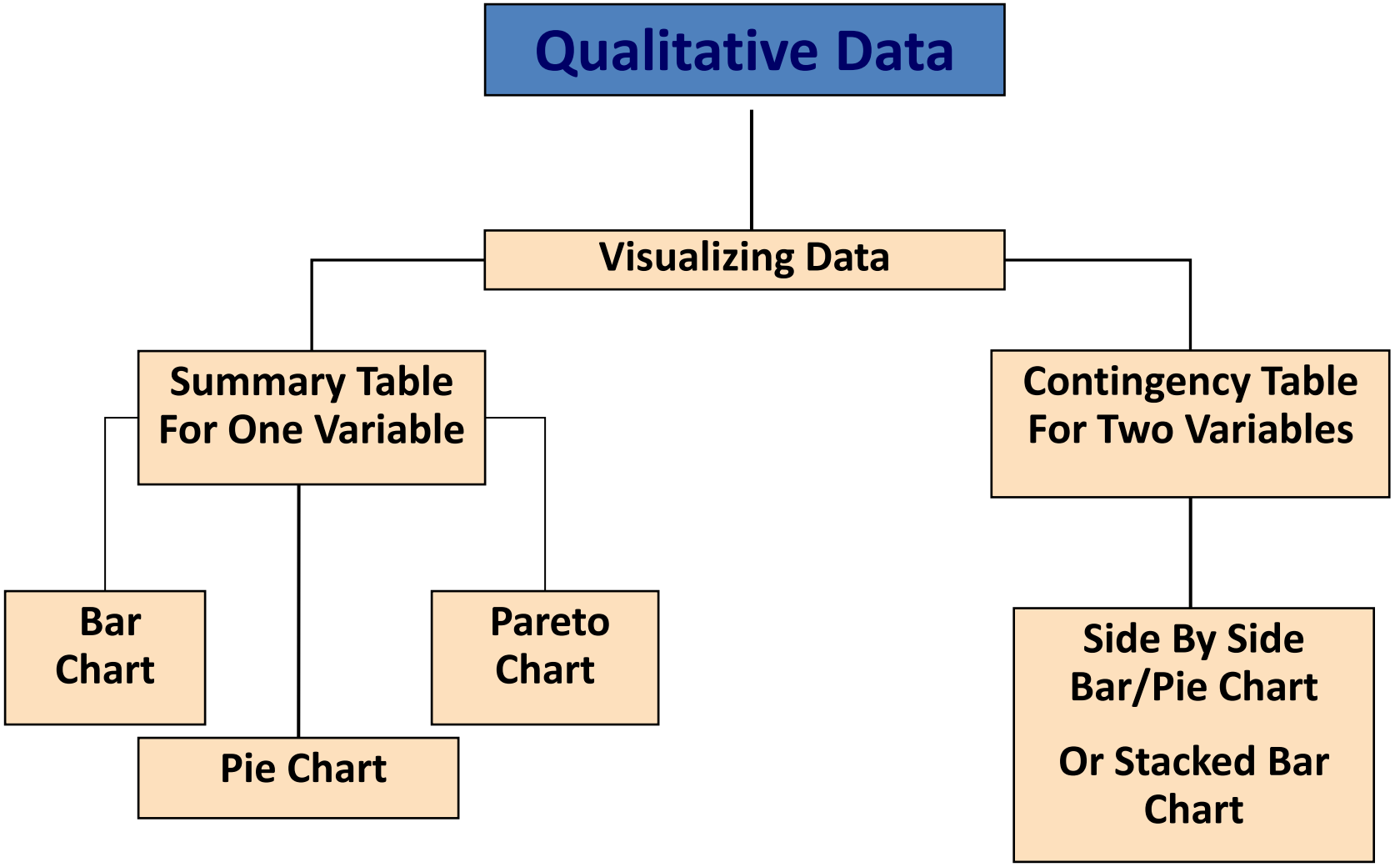
- Construct a full (frequency + relative + cumulative) distribution for the following data set

17 62 15 65 28 51 24 65 39 41 35 15 39 32 36 37

Charts, Bars, Pies, Boxplots

REPRESENTING DATA GRAPHICALLY

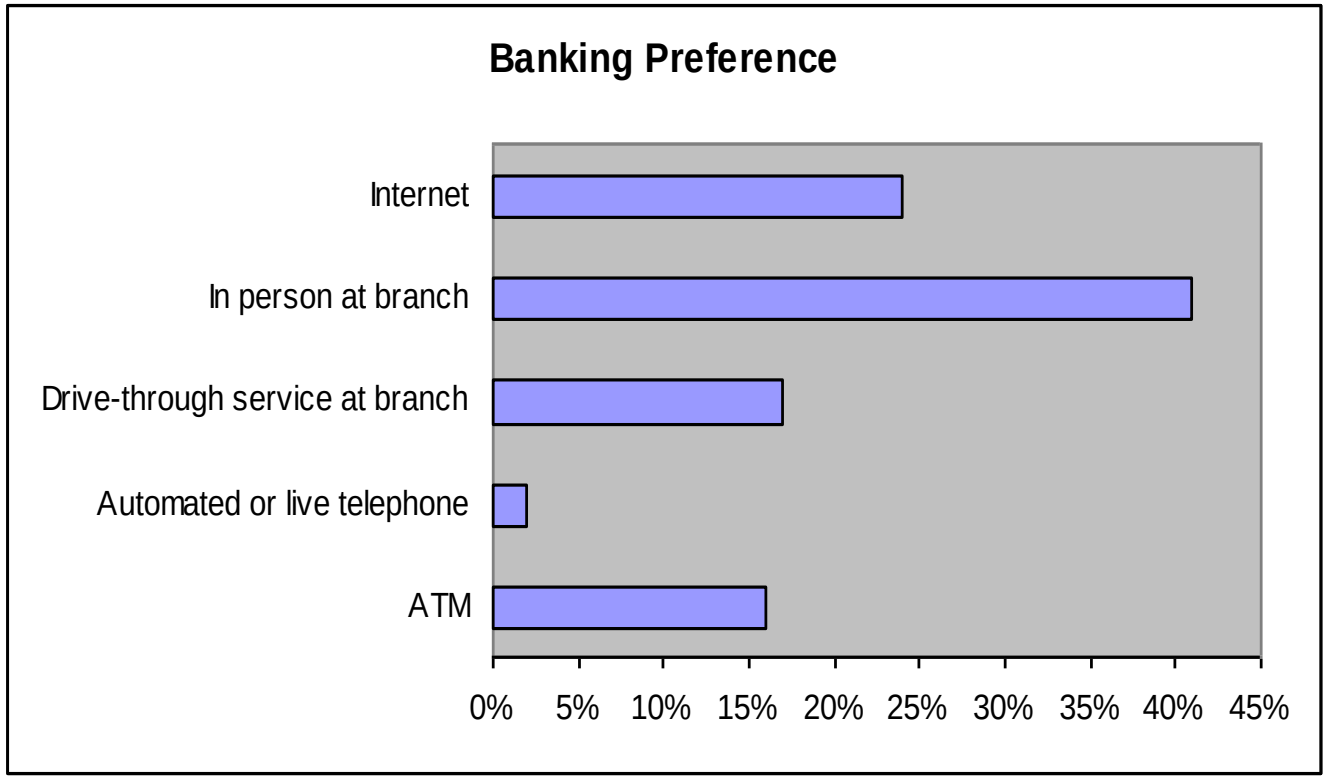
Visualizing Qualitative Data Through Graphical Displays



Visualizing qualitative Data: The Bar Chart

- In a **bar chart,** each bar shows a category of the variable, the length of which represents the frequency (amount) or percentage of values falling into the category (based on the summary table of the variable).

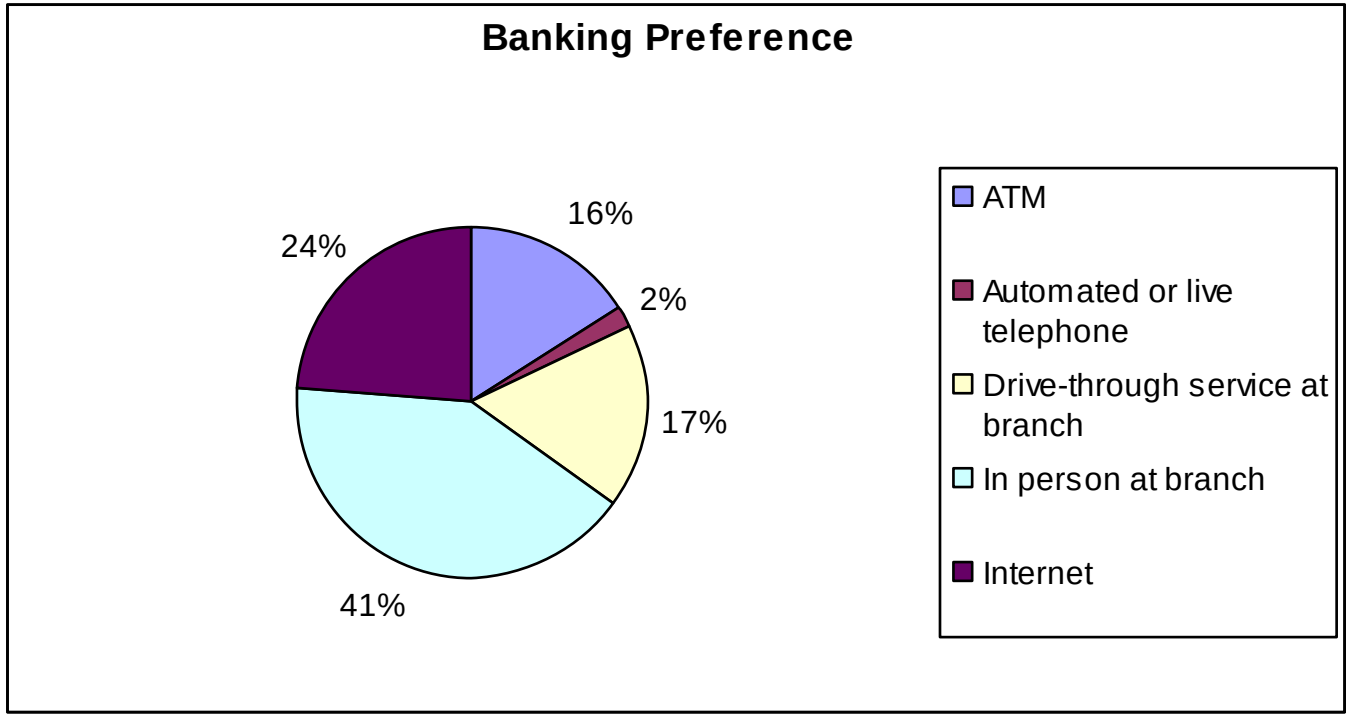
Banking Preference?	**%**
ATM	16%
Automated or live telephone	2%
Drive-through service at branch	17%
In person at branch	41%
Internet	24%



Visualizing qualitative Data: The Pie Chart

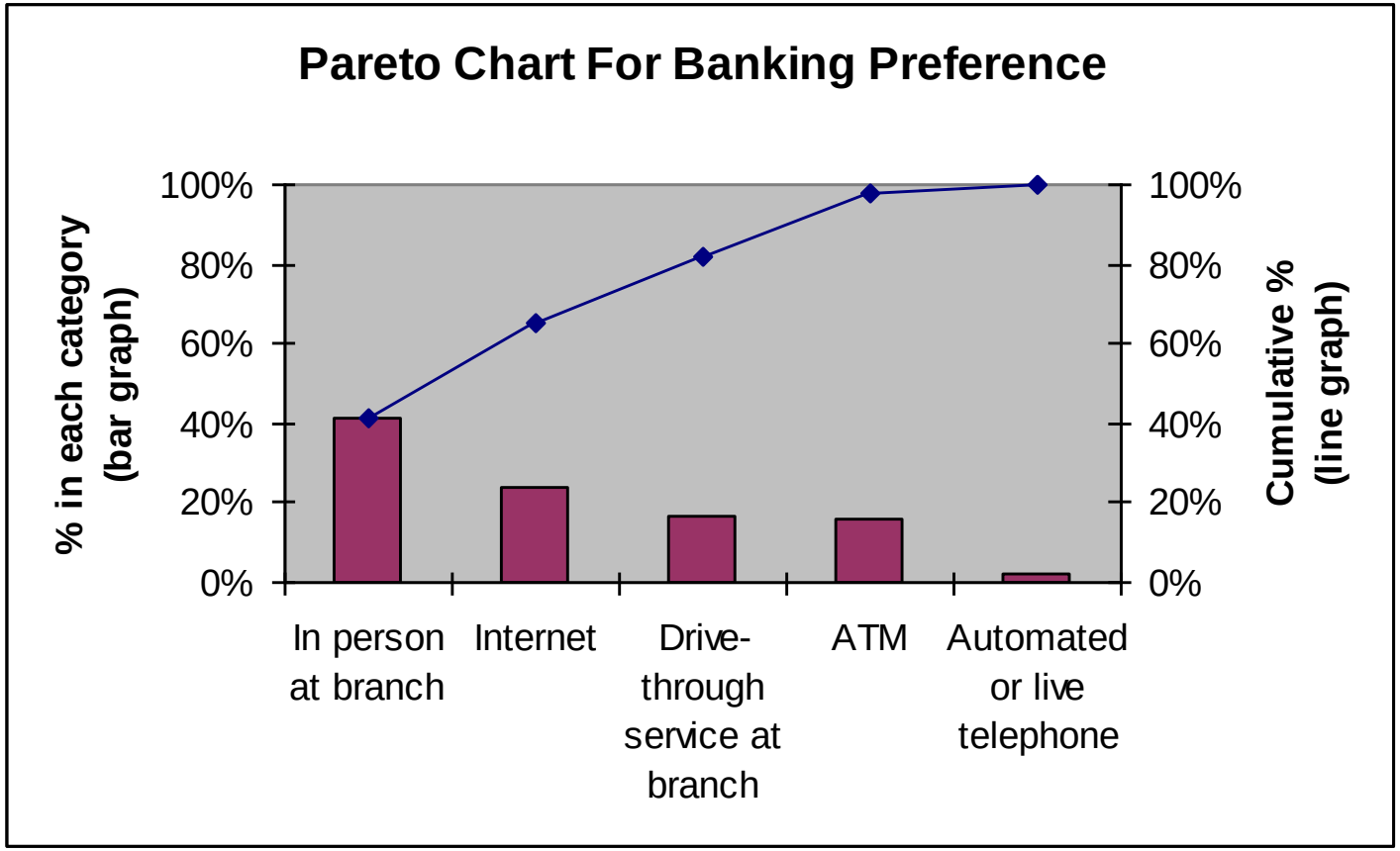
- The **pie chart** is a circle broken up into slices that represent categories of the variable. The size of each slice is proportional to the percentage of each category.

Banking Preference?	**%**
ATM	16%
Automated or live telephone	2%
Drive-through service at branch	17%
In person at branch	41%
Internet	24%



Visualizing qualitative Data: The Pareto Chart

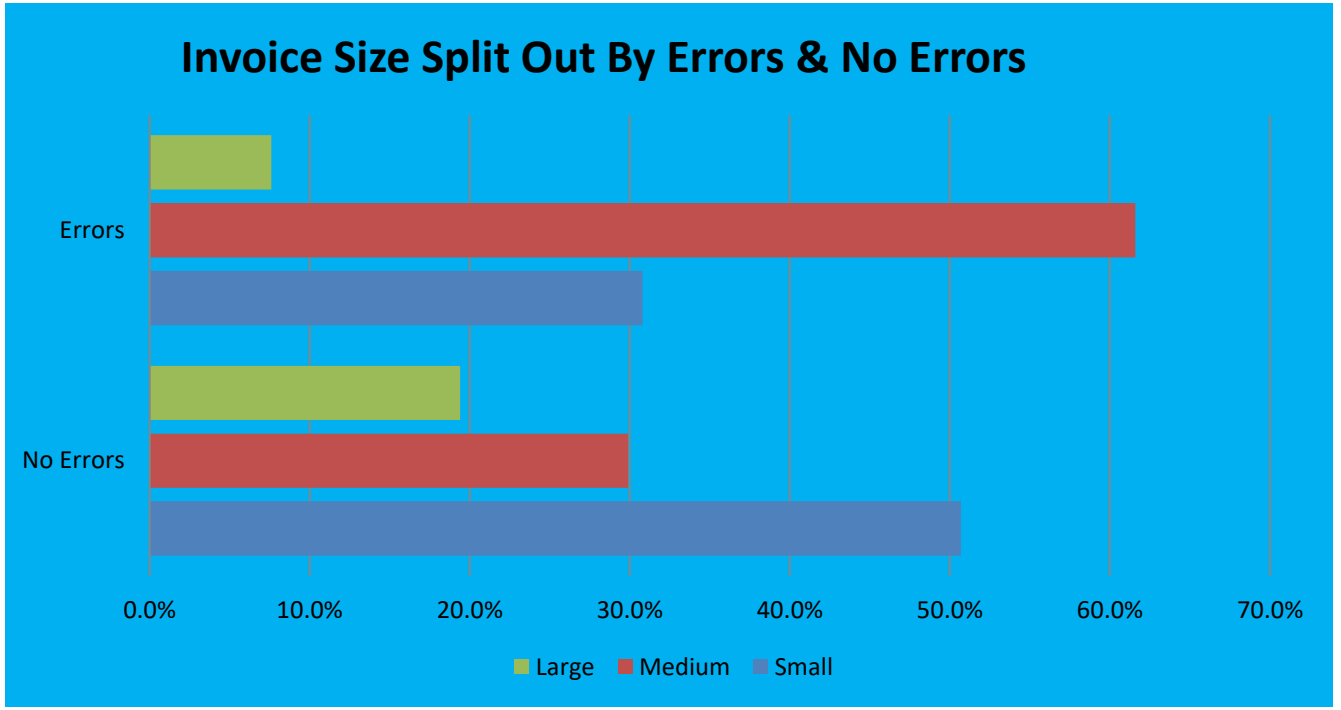
- Used to portray qualitative data
- A vertical bar chart, where categories are shown in descending order of frequency
- A cumulative polygon is shown in the same graph



Visualizing qualitative Data: Side By Side Bar Charts

- The **side by side bar chart** represents the data from a contingency table.

	No Errors	Errors	Total
Small Amount	50.75%	30.77%	47.50%
Medium Amount	29.85%	61.54%	35.00%
Large Amount	19.40%	7.69%	17.50%
Total	100.0%	100.0%	100.0%

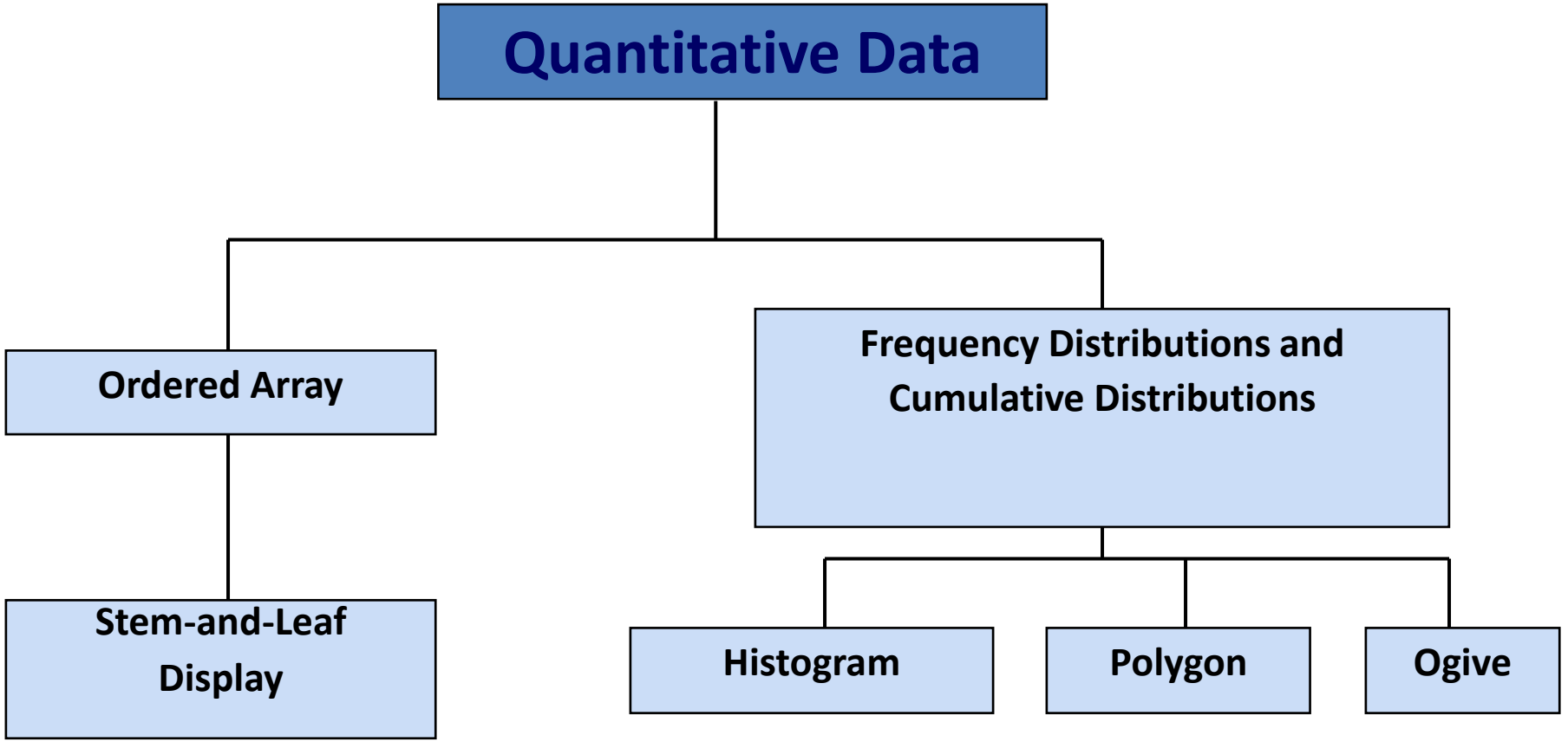


Invoices with errors are much more likely to be of medium size (61.54% vs 30.77% and 7.69%)

Exercise 3-1

- Plot several graphic displays of all the categorical variables in the DDT data file
  - Bar charts
  - Pie charts
  - Pareto charts
  - Side by side bar charts

Visualizing Quantitative Data By Using Graphical Displays



Organizing Quantitative Data: Stem and Leaf Display

- A **stem-and-leaf display** organizes data into groups (called stems) so that the values within each group (the leaves) branch out to the right on each row.

Age of Surveyed College Students						
	Day Students					
	16	17	17	18	18	18
	19	19	20	20	21	22
	22	25	27	32	38	42
	Night Students					
	18	18	19	19	20	21
	23	28	32	33	41	45

Age of College Students

Day Students

Stem	Leaf
1	67788899
2	0201257
3	28
4	2

Night Students

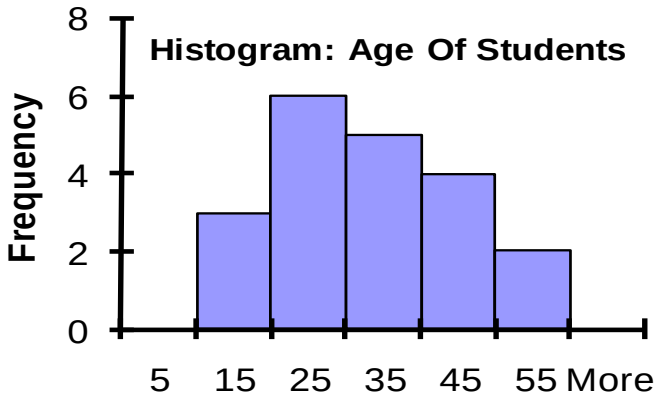
Stem	Leaf
1	8899
2	0138
3	23
4	15

Visualizing Quantitative Data: The Histogram

- A vertical bar chart of the data in a frequency distribution is called a **histogram**.
- In a histogram there are no gaps between adjacent bars.
- The **class boundaries** (or **class midpoints**) are shown on the horizontal axis.
- The vertical axis is either **frequency, relative frequency or percentage**.
- The height of the bars represent the frequency, relative frequency or percentage.

Class	Frequency	Relative Frequency	Percentage
10 but less than 20	3	.15	15
20 but less than 30	6	.30	30
30 but less than 40	5	.25	25
40 but less than 50	4	.20	20
50 but less than 60	2	.10	10
Total	20	1.00	100

(In a percentage histogram the vertical axis would be defined to show the percentage of observations per class)



Exercise 3-2

- Generate a histogram of the numerical variables in the DDT data file with MS Excel
- Describe the distribution of grade point averages contained in the data file **Grade Point Averages**.
- Describe the following random sample of 10 final exam grades for an introductory accounting class with a stem-and-leaf display.

  88 51 63 85 79 65 79 70 73 77

  (Recommendation: use Stem and Leaf Display)

Visualizing Quantitative Data: The Polygon (line chart) & Ogive

- **A frequency (or percentage) polygon** is formed by having the midpoint of each class represent the data in that class and then connecting the sequence of midpoints at their respective class frequencies (or percentages).

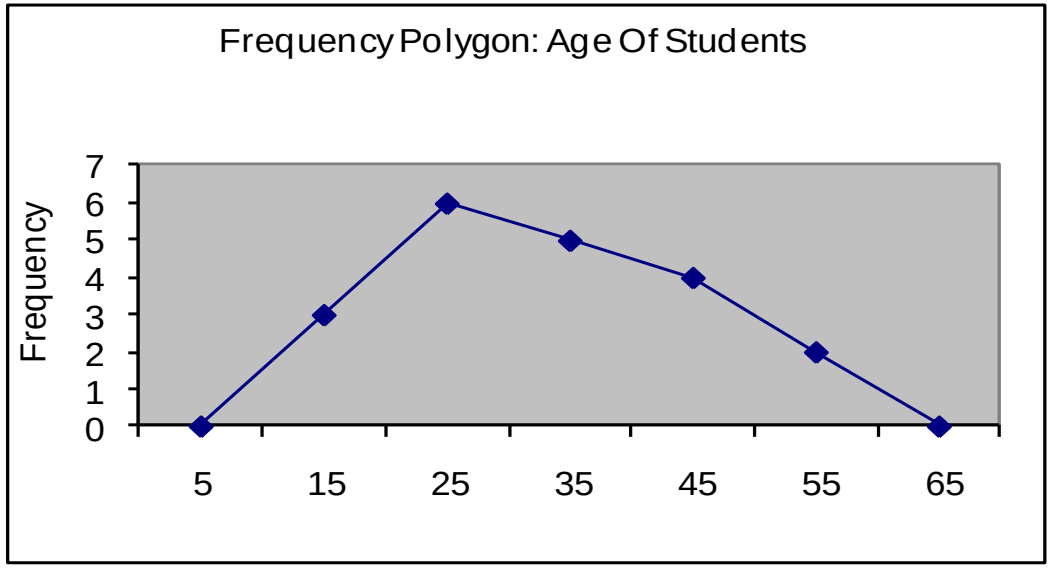
- The **cumulative frequency (or percentage) polygon**, called ogive, displays the variable of interest along the X axis, and the cumulative frequencies (or percentages) along the Y axis.

- Useful when there are two or more groups to compare.

Visualizing Quantitative Data: The Frequency Polygon

Class	Class Midpoint	Frequency
10 but less than 20	15	3
20 but less than 30	25	6
30 but less than 40	35	5
40 but less than 50	45	4
50 but less than 60	55	2

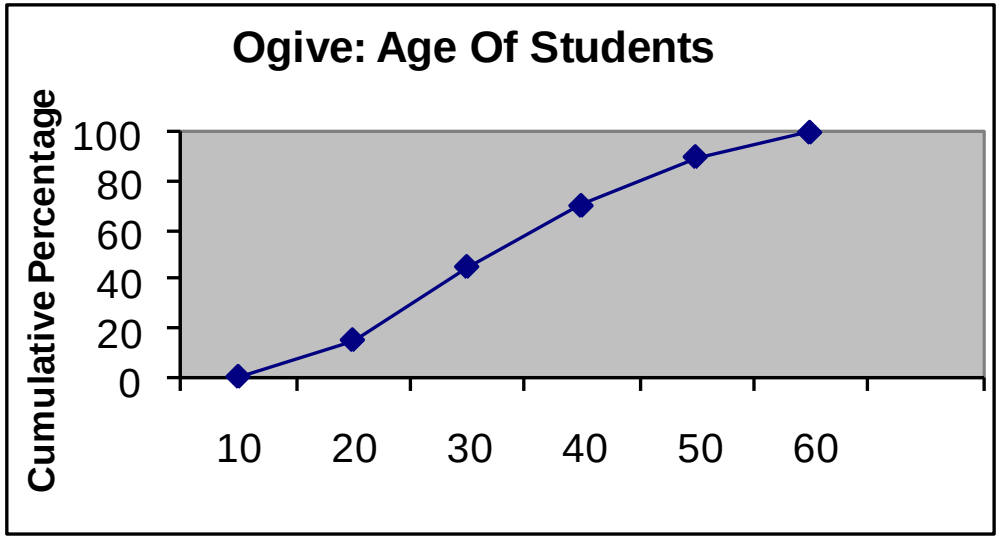
(In a percentage polygon the vertical axis would be defined to show the percentage of observations per class)



Visualizing Quantitative Data: The Ogive (Cumulative % Polygon)

Class	Lower class boundary	% less than lower boundary
10 but less than 20	10	15
20 but less than 30	20	45
30 but less than 40	30	70
40 but less than 50	40	90
50 but less than 60	50	100

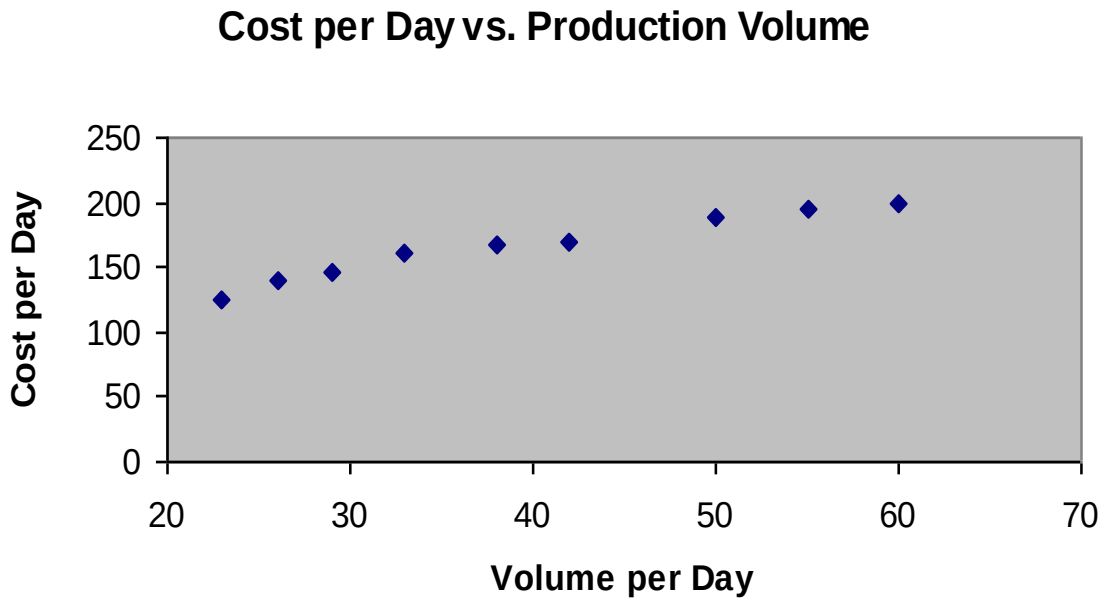
(In an ogive the percentage of the observations less than each lower class boundary are plotted versus the lower class boundaries.



Visualizing Two Quantitative Variables: The Scatter Plot

- **Scatter plots** are used for data consisting of paired observations taken from two quantitative variables
- One variable is measured on the vertical axis and the other variable is measured on the horizontal axis
- Scatter plots are used to examine possible relationships between two quantitative variables

Volume per day	Cost per day
23	125
26	140
29	146
33	160
38	167
42	170
50	188
55	195
60	200

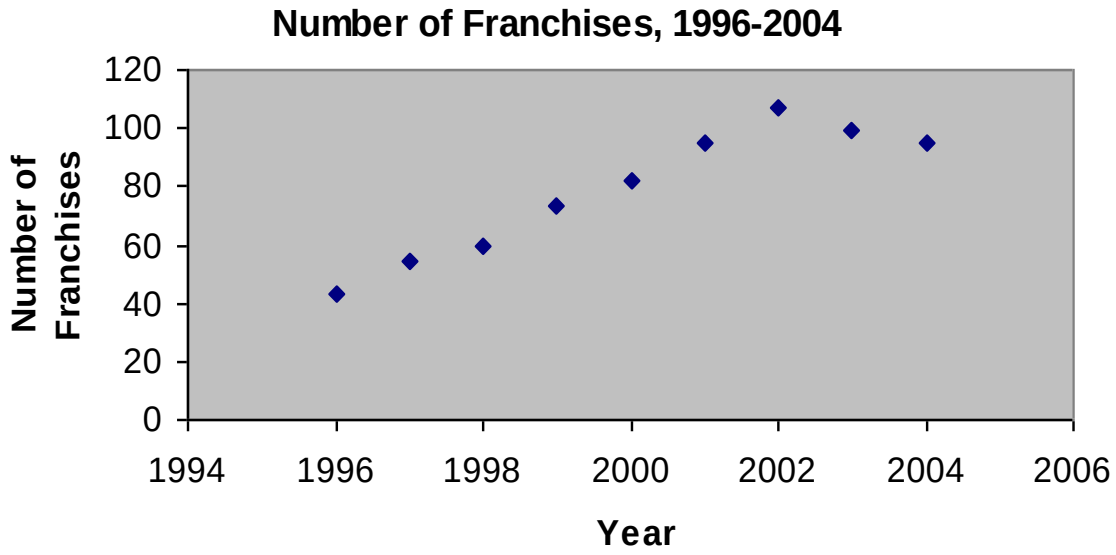


Visualizing Two Quantitative Variables: The Time Series Plot

- A **Time Series Plot** is used to study patterns in the evolution of values of a quantitative variable over time

- The Time Series Plot:
  - Numeric variable is measured on the vertical axis and the time period is measured on the horizontal axis

Year	Number of Franchises
1996	43
1997	54
1998	60
1999	73
2000	82
2001	95
2002	107
2003	99
2004	95



Principles of Excellent Graphs

- The graph should not distort the data.
- The graph should not contain unnecessary adornments (sometimes referred to as chart junk).
- The scale on the vertical axis should begin at zero.
- All axes should be properly labeled.
- The graph should have a title.
- The simplest possible graph should be used for a given set of data.

Representing data with numbers and graphs

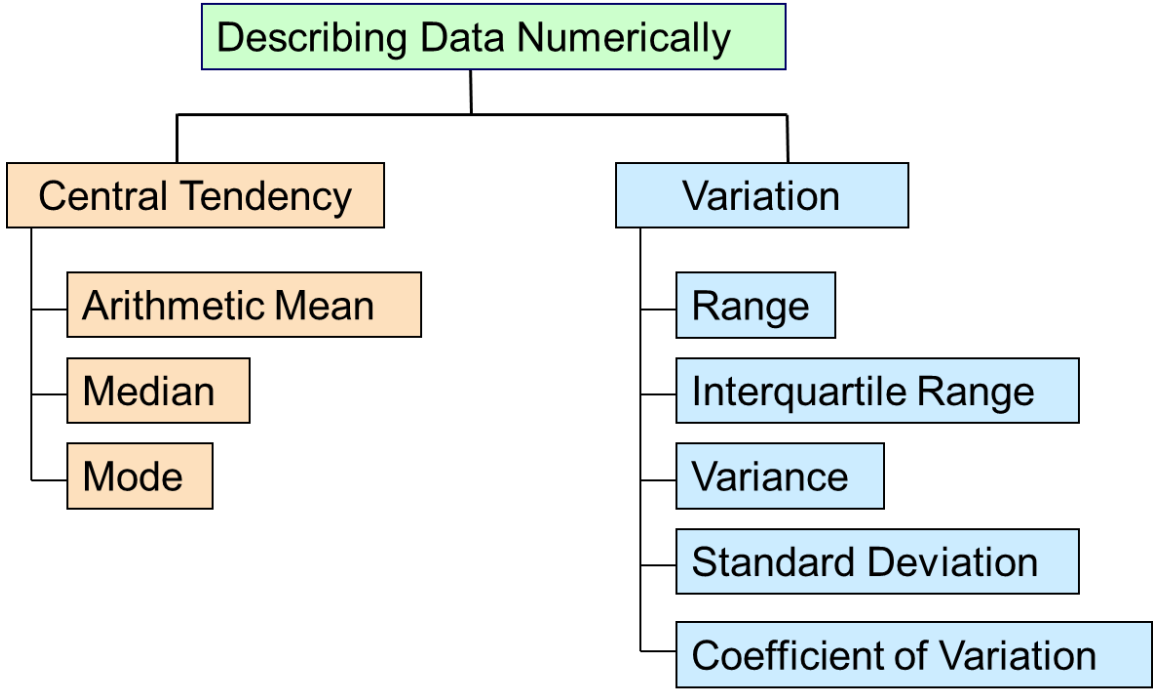
EXERCISE : EXPLORE ON YOUR OWN !

Winds Data

8.9	12.4	8.6	11.3	9.2	8.8	35.1	6.2	7
7.1	11.8	10.7	7.6	9.1	9.2	8.2	9	8.7
9.1	10.9	10.3	9.6	7.8	11.5	9.3	7.9	8.8
8.8	12.7	8.4	7.8	5.7	10.5	10.5	9.6	8.9
10.2	10.3	7.7	10.6	8.3	8.8	9.5	8.8	9.4

Weight Loss Data

		Results			
		None	Small	Average	Large
Diet	A	15	21	45	13
	B	26	31	34	5
	C	33	17	49	20



Central Tendency

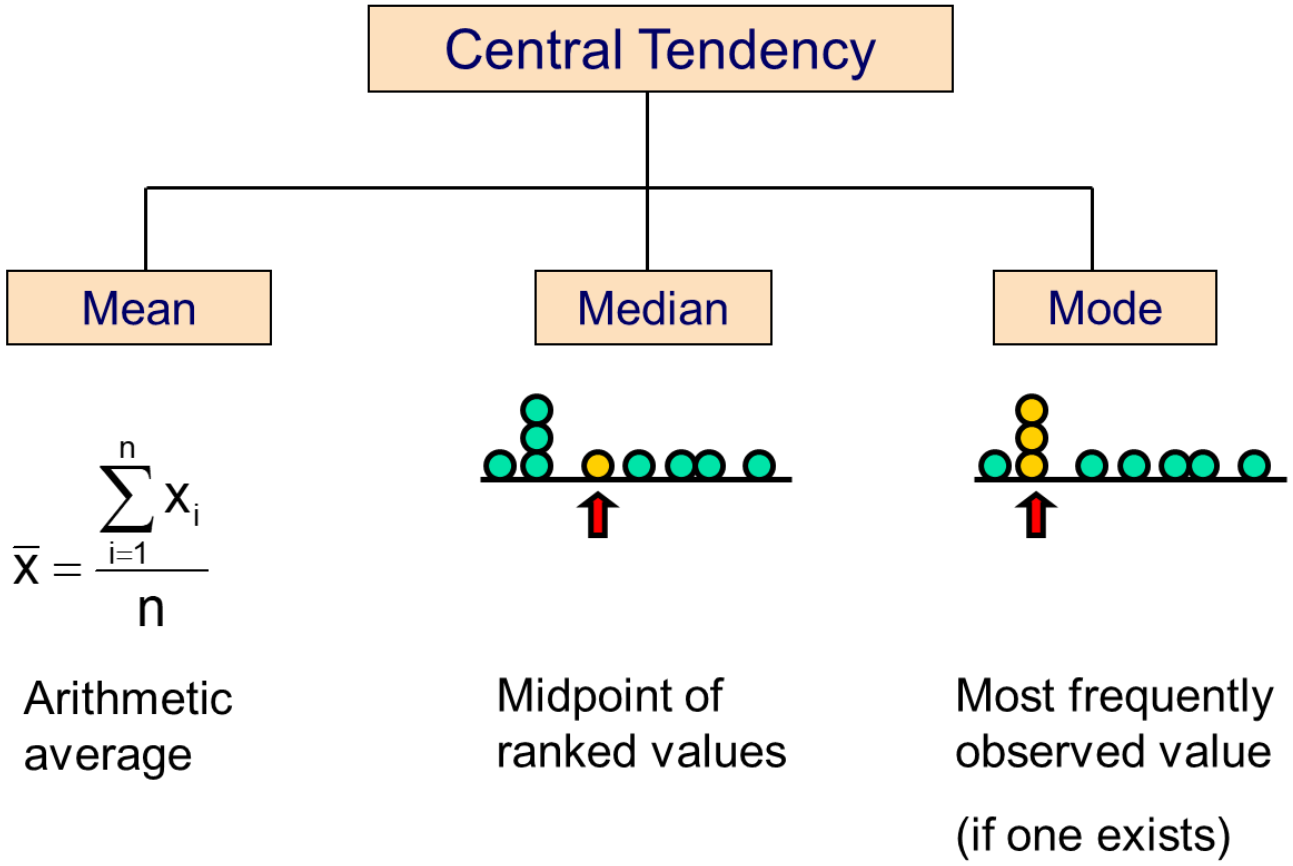
Shape

Spread / Dispersion

DESCRIBING DATA NUMERICALLY

Measures of Central Tendency

- The **central tendency** is the extent to which all the data values group around a typical or central value.
- The **variation (spread / dispersion)** is the amount of dispersion or scattering of values
- The **shape** is the pattern of the distribution of values from the lowest value to the highest value.



Arithmetic Mean

- The arithmetic mean (mean) is the most common measure of central tendency
  - For a population of N values:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Population values

Population size

  - For a sample of size n:

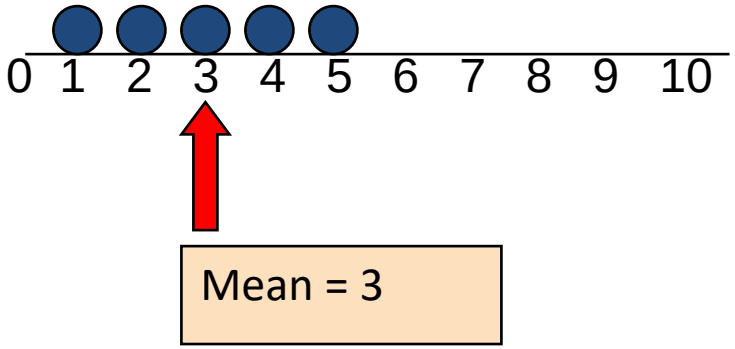
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Observed values

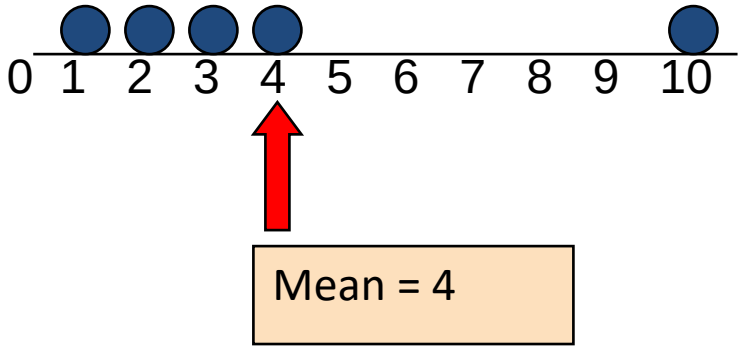
Sample size

Arithmetic Mean : example

- Mean = sum of values divided by the number of values
  - Affected by extreme values (outliers)



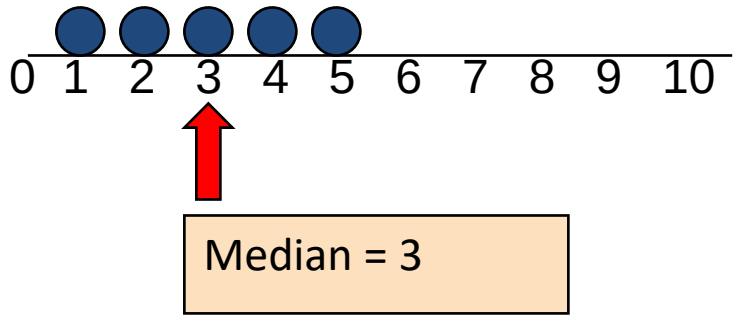
$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

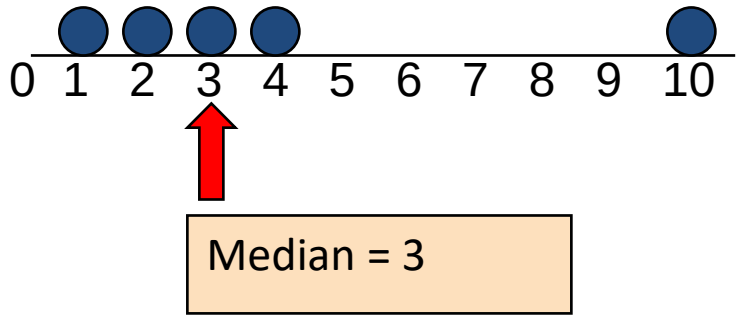


$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

Median

- In an ordered list, the median is the “middle” number (50% above, 50% below)





  - Not affected by extreme values

Finding the Median

- The location of the median:

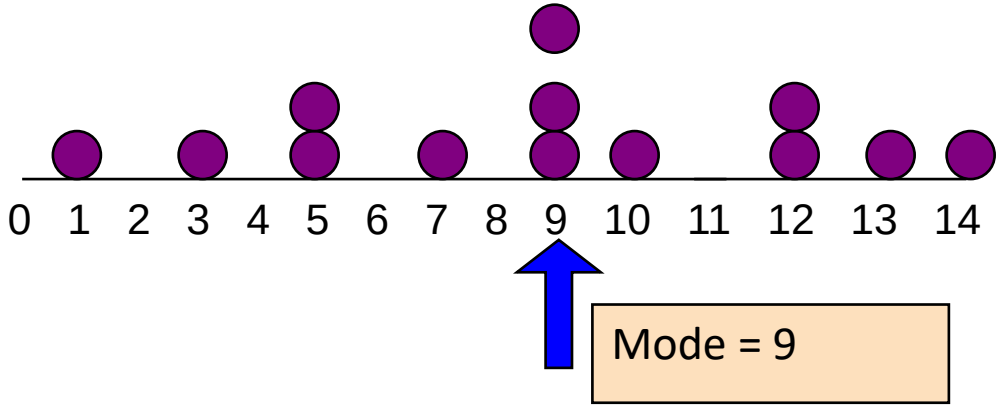
$$\text{Median position} = \left(\frac{n+1}{2}\right)^{th} \text{ position in the ordered data}$$

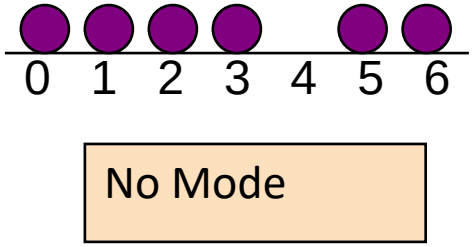
  - If the number of values is odd, the median is the middle number
  - If the number of values is even, the median is the average of the two middle numbers

- Note that $\frac{n+1}{2}$ is not the *value* of the median, only the *position* of the median in the ranked data

Mode

- A measure of central tendency
- Value that occurs most often
- Not affected by extreme values
- Used for either numerical or categorical data
- There may be no mode
- There may be several modes





Example

House Prices:

$2,000,000
500,000
300,000
100,000
100,000

Sum 3,000,000

- **Mean:** ($3,000,000/5)
  = **$600,000**
- **Median:** middle value of ranked data
  = **$300,000**
- **Mode:** most frequent value
  = **$100,000**

Exercise

- Compute the Arithmetic mean, the median and the mode for the following set of numbers

  60 84 65 67 75 72 80 85 63 82 70 75

- Which measure of location (central tendency) is the “best”?

Geometric Mean

- Geometric mean
  - Used to measure the rate of change of a variable over time

$$\overline{X}_G = (X_1 \times X_2 \times \cdots \times X_n)^{1/n}$$

- Geometric mean rate of return
  - Measures the status of an investment over time

$$\overline{R}_G = [(1+R_1)\times(1+R_2)\times\cdots\times(1+R_n)]^{1/n} - 1$$

  - Where R_i is the rate of return in time period i

Geometric mean : example

An investment of \$100,000 declined to \$50,000 at the end of year one and rebounded to \$100,000 at end of year two:

$X_1 = \$100{,}000 \quad X_2 = \$50{,}000 \quad X_3 = \$100{,}000$

50% decrease 100% increase

The overall two-year return is zero, since it started and ended at the same level.

Use the 1-year returns to compute the arithmetic mean and the geometric mean:

Arithmetic mean rate of return:

$$\overline{X} = \frac{(-.5)+(1)}{2} = .25 = 25\%$$

Misleading result

Geometric mean rate of return:

$$\overline{R}_G = [(1+R_1)\times(1+R_2)\times\cdots\times(1+R_n)]^{1/n} - 1$$
$$= [(1+(-.5))\times(1+(1))]^{1/2} - 1$$
$$= [(.50)\times(2)]^{1/2} - 1 = 1^{1/2} - 1 = 0\%$$

More representative result

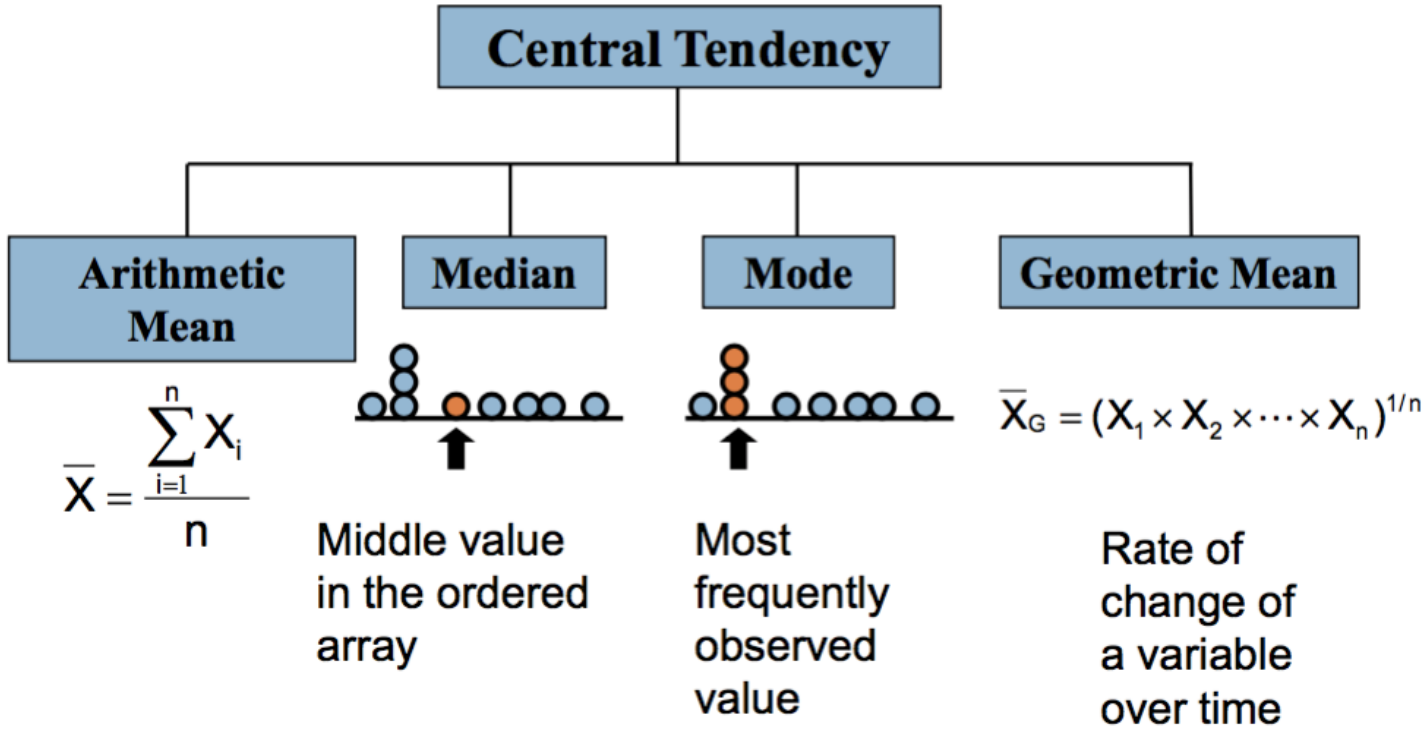
Geometric mean : exercises

- Exercise 1
  - An investment of $100,000 rose to $150,000 at the end of year one and increased to $180,000 at end of year two:
  - What is the mean percentage return over time?

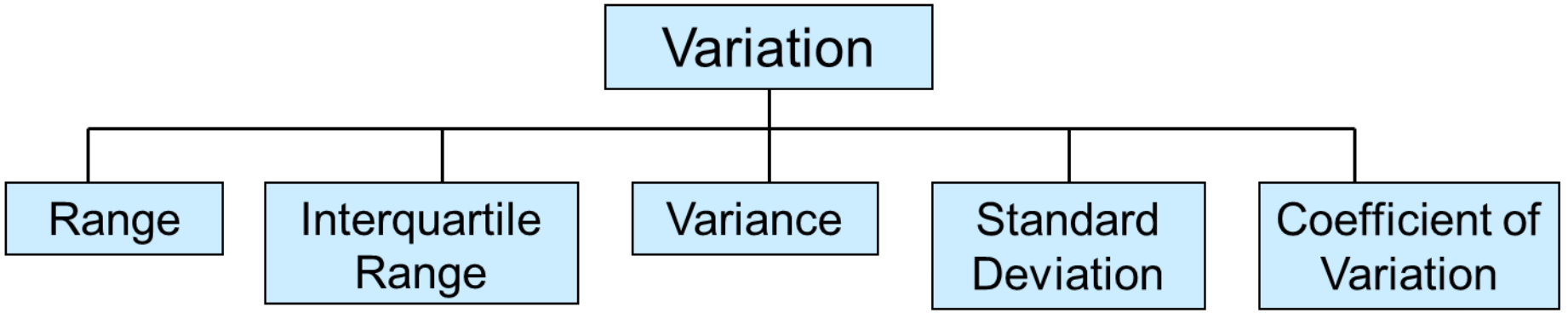
- Exercise 2
  - Find the annual growth rate if sales have grown 25% over 5 years.

Summary : Central Tendency

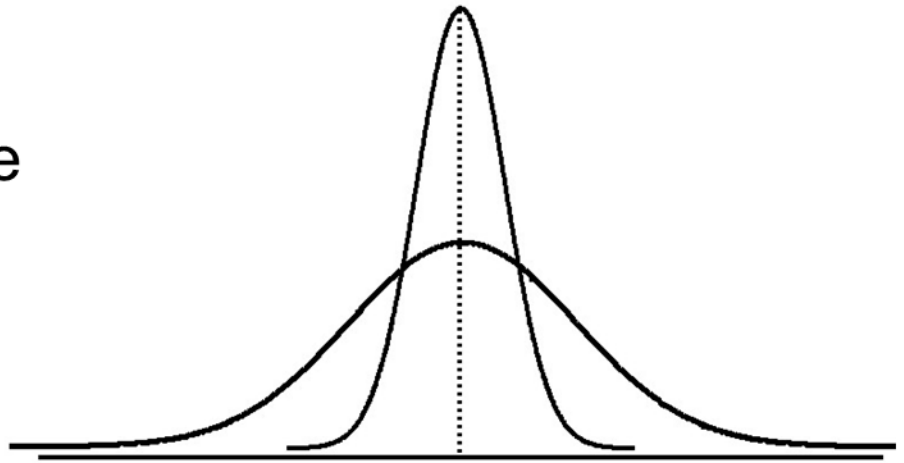
- The **mean** is generally used, unless extreme values (outliers) exist.
- The **median** is often used, since the median is not sensitive to extreme values. For example, median home prices may be reported for a region; it is less sensitive to outliers.
- In some situations it makes sense to report both the **mean** and the **median**.
- **Mode** in the only option for categorical data



Measures of Variability / Spread / Dispersion



- Measures of variation give information on the spread or variability of the data values.



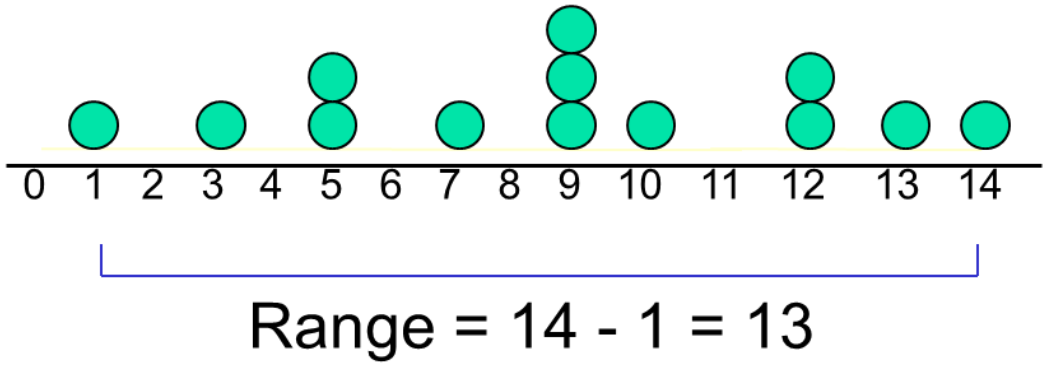
Same center, different variation

Range

- Simplest measure of variation
- Difference between the largest and the smallest observations:

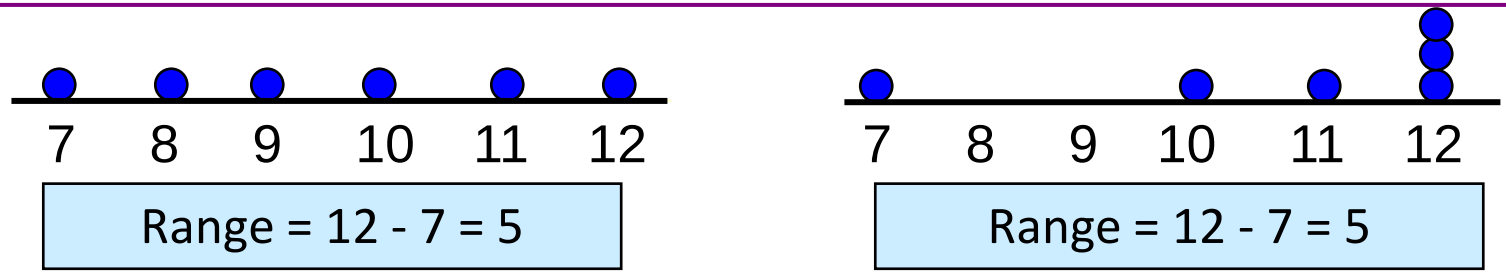
$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$

Example:



Disadvantages of the Range

- Ignores the way in which data are distributed



- Sensitive to outliers

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,5

Range = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120

Range = 120 - 1 = 119

Interquartile Range

- Can eliminate some outlier problems by using the interquartile range

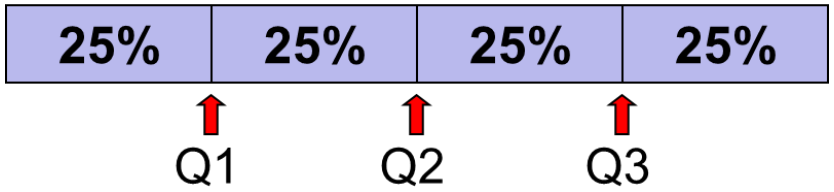
- Eliminate high- and low-valued observations and calculate the range of the middle 50% of the data

- Interquartile range = 3rd quartile – 1st quartile

  $IQR = Q_3 - Q_1$

Quartile Measures

- Quartiles split the ranked data into 4 segments with an equal number of values per segment



- The first quartile, Q1, is the value for which 25% of the observations are smaller and 75% are larger
- Q2 is the same as the median (50% of the observations are smaller and 50% are larger)
- The third quartile, Q3, is the value for which 75% of the observations are smaller and 25% are larger

- Find a quartile by determining the value in the appropriate position in the ranked data, where
  - First quartile position: $Q_1 = (n+1)/4$ ranked value
  - Second quartile position: $Q_2 = (n+1)/2$ ranked value
  - Third quartile position: $Q_3 = 3(n+1)/4$ ranked value

where **n** is the number of observed values

Quartiles : example

Sample Data in Ordered Array: 11 12 13 16 16 17 18 21 22

n = 9

Q_1 is in the (9+1)/4 = 2.5 position of the ranked data,

so **Q_1 = position#2 + 0.5*(position#3 – position#2) = 12 + 0.5*(13-12) = 12.5**

Q_2 is in the (9+1)/2 = 5^{th} position of the ranked data,

so **Q_2 = median = 16**

Q_3 is in the 3(9+1)/4 = 7.5 position of the ranked data,

so **Q_3 = position#7 + 0.5*(position#8 – position#7) = 18 + 0.5*(21-18) = 19.5**

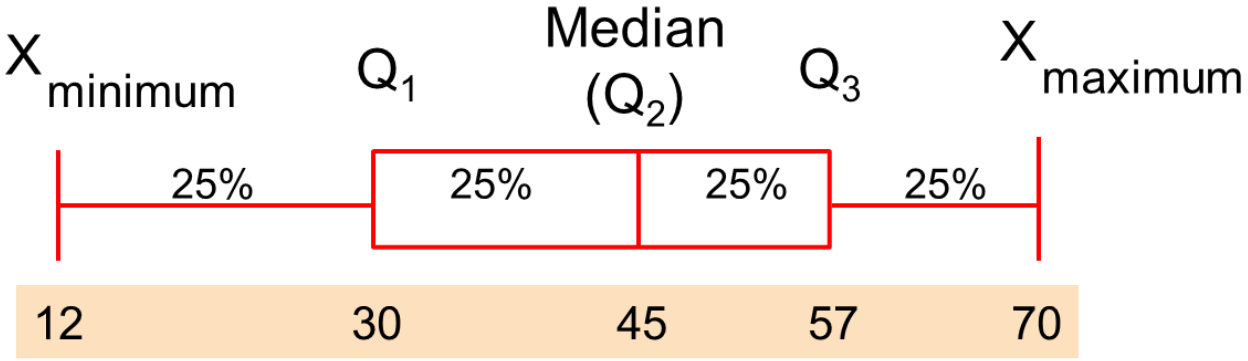
Q_1 and Q_3 are measures of non-central location
Q_2 = median, is a measure of central tendency

Introduction to Box-and-Whisker Plot

- A box-and-whisker plot is a graph that describes the shape of a distribution
- Created from **the five-number summary**: the minimum value, Q_1, the median, Q_3, and the maximum
- The inner box shows the range from Q_1 to Q_3, with a line drawn at the median
- Two "whiskers" extend from the box. One whisker is the line from Q_1 to the minimum, the other is the line from Q_3 to the maximum value

The plot can be oriented horizontally or vertically

Example:



Constructing Full Boxplots

- Draw a single vertical (or horizontal) axis spanning the range of the data. Draw short horizontal lines at the lower and upper quartiles and at the median. Then connect them with vertical lines to form a box.

- Erect "fences" around the main part of the data.
  - The upper fence is 1.5IQRs above the upper quartile.
  - The lower fence is 1.5IQRs below the lower quartile.
  - Note: the fences only help with constructing the boxplot and should not appear in the final display.

Use the fences to grow "whiskers."

  - Draw lines from the ends of the box up and down to the most extreme data values found within the fences.
  - If a data value falls outside one of the fences, we do not connect it with a whisker.

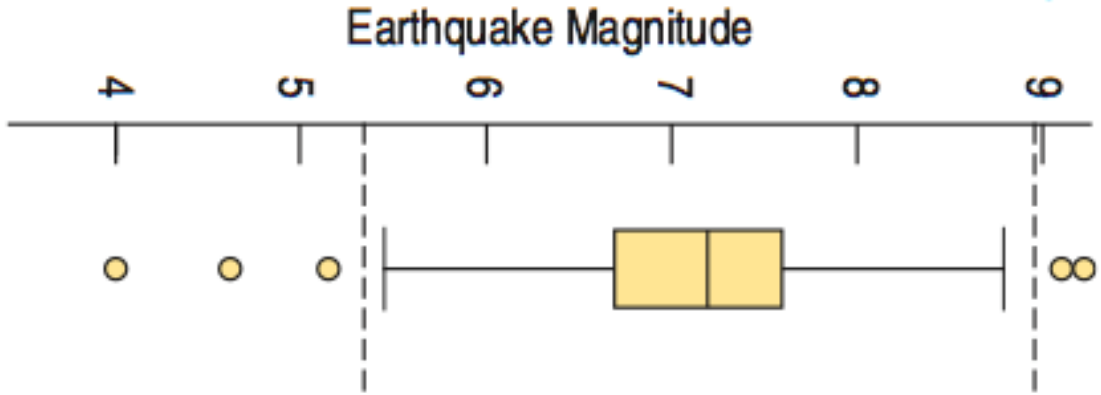
- Add the outliers by displaying any data values beyond the fences with special symbols.
  - We often use a different symbol for "far outliers" that are farther than 3IQRs from the quartiles.

Boxplots : illustration

- The smallest tsunami-causing earthquake had magnitude 4.0.
- The largest tsunami-causing earthquake had magnitude 9.1.
- The middle half of tsunami-causing earthquakes is between 6.7 and 7.6.
- Half of tsunami-causing earthquakes have magnitudes below 7.2 and half are above 7.2.
- A tsunami-causing earthquake less than 6.7 is small.
- A tsunami-causing earthquake more than 7.6 is big.

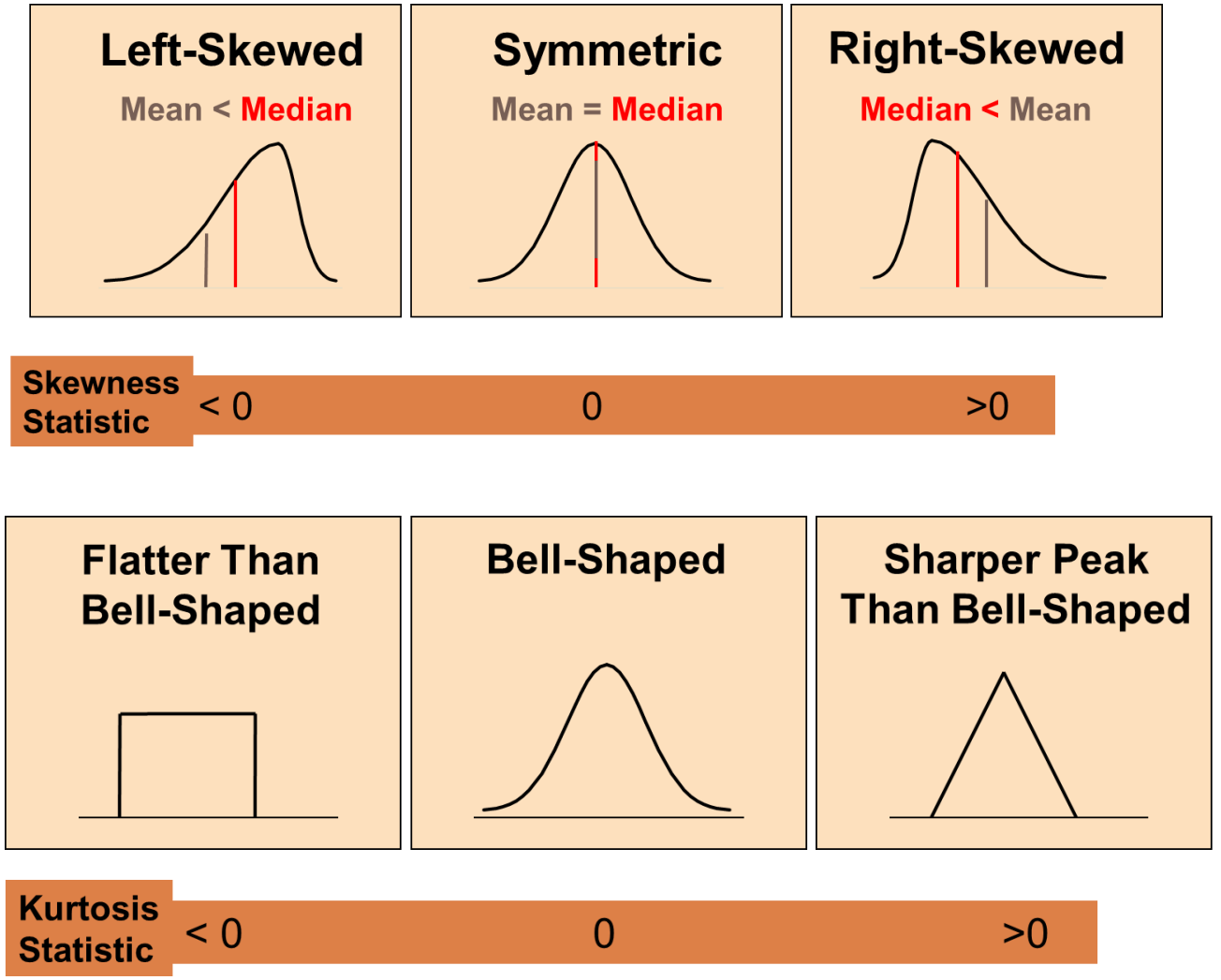
- Q1 = 6.7, Q3 = 7.6 so IQR = 7.6 – 6.7 = 0.9
- Lower Fence = 6.7 – 1.5 × 0.9 = 5.35
- Upper Fence = 7.6 + 1.5 × 0.9 = 8.95

Max	9.1
Q3	7.6
Median	7.2
Q1	6.7
Min	4.0



Shape of a distribution

- Describes how data are distributed
- Two useful shape related statistics are:
  - Skewness
    - Measures the amount of asymmetry in a distribution
  - Kurtosis
    - Measures the relative concentration of values in the center of a distribution as compared with the tails



Exercise

For the following set of numbers :

60 63 65 67 70 72 75 75 80 82 84 85

- Give the 5 numbers summary
- Compute the range and the IQR
- Draw the **Full Box Plot**
- Is the corresponding distribution symmetric ?

Left-Skewed	Symmetric	Right-Skewed
$Median - X_{smallest}$	$Median - X_{smallest}$	$Median - X_{smallest}$
$>$	$\approx$	$<$
$X_{largest} - Median$	$X_{largest} - Median$	$X_{largest} - Median$
$Q_1 - X_{smallest}$	$Q_1 - X_{smallest}$	$Q_1 - X_{smallest}$
$>$	$\approx$	$<$
$X_{largest} - Q_3$	$X_{largest} - Q_3$	$X_{largest} - Q_3$
$Median - Q_1$	$Median - Q_1$	$Median - Q_1$
$>$	$\approx$	$<$
$Q_3 - Median$	$Q_3 - Median$	$Q_3 - Median$

Variance

Population Variance

- Average of squared deviations of values from the mean
  - Population variance:

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

Where μ = population mean

N = population size

x_i = i^{th} value of the variable x

Sample Variance

- Average (approximately) of squared deviations of values from the mean
  - Sample variance:

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

Where $\bar{X}$ = arithmetic mean

n = sample size

X_i = i^{th} value of the variable X

Standard Deviation

Population Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data

- Population standard deviation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Sample Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data

- Sample standard deviation:

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

Example: Sample Standard Deviation

Sample Data (x_i) : 10 12 14 15 17 18 18 24

n = 8 Mean = $\bar{x}$ = 16

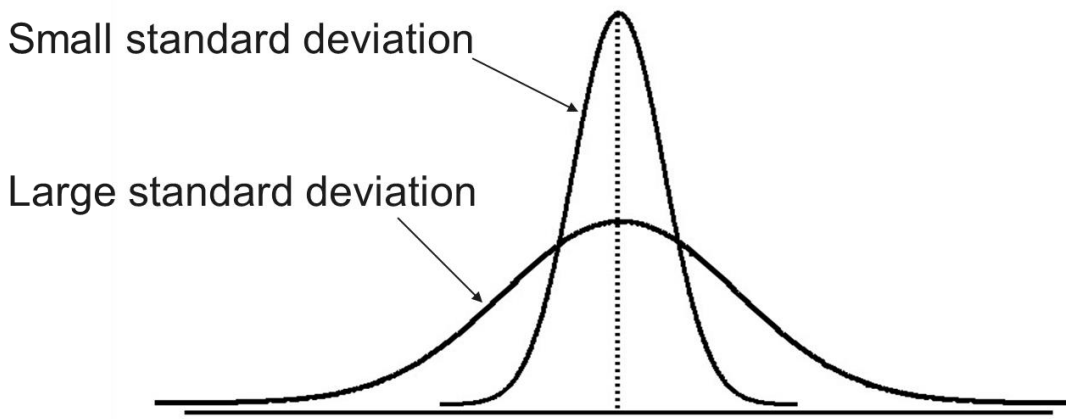
$$s = \sqrt{\frac{(10 - \bar{X})^2 + (12 - \bar{x})^2 + (14 - \bar{x})^2 + \cdots + (24 - \bar{x})^2}{n - 1}}$$

$$= \sqrt{\frac{(10 - 16)^2 + (12 - 16)^2 + (14 - 16)^2 + \cdots + (24 - 16)^2}{8 - 1}}$$

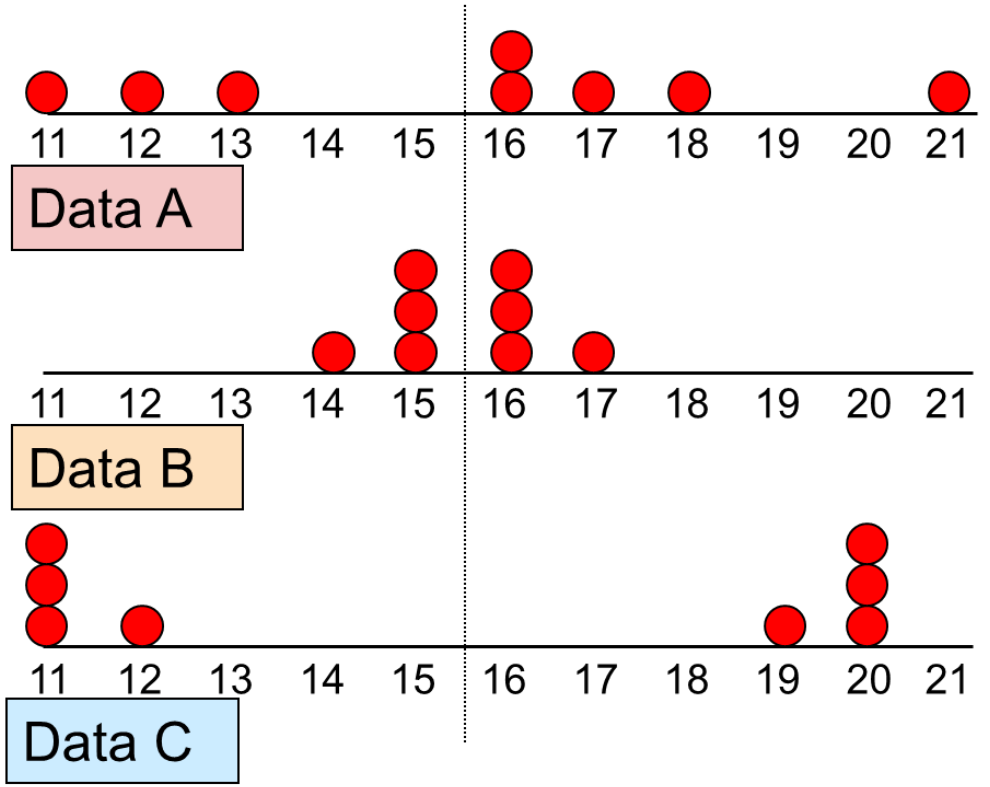
$$= \sqrt{\frac{130}{7}} = 4.3095$$

→ A measure of the "average" scatter around the mean

Comparing standard deviation



Mean = 15.5 for each data set



s = 3.338
(compare to the two cases below)

s = 0.926
(values are concentrated near the mean)

s = 4.570
(values are dispersed far from the mean)

Exercises

1.11 The following results on summations will help us in calculating the sample variance s^2. For any constant c,

a $\sum_{i=1}^{n} c = nc.$

b $\sum_{i=1}^{n} cy_i = c\sum_{i=1}^{n} y_i.$

c $\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i.$

Use (a), (b), and (c) to show that

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right].$$

1.12 Use the result of Exercise 1.11 to calculate s for the $n = 6$ sample measurements 1, 4, 2, 1, 3, and 3.

- Compute the variance and standard deviation of the following set of values
  - 6 8 10 12 14 9 11 7 13 11

Variability : summary

- The more the data are spread out, the greater the range, variance and standard deviation.
- The more the data are concentrated, the smaller the range, variance and standard deviation.
- If the values are all the same (no variation), all these measures will be zero.
- None of these measures are ever negative.

Comparing variation : Coefficient of Variation

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean
- Can be used to compare two or more sets of data measured in different units

Population coefficient of variation:

$$CV = \left(\frac{\sigma}{\mu}\right) \cdot 100\%$$

Sample coefficient of variation:

$$CV = \left(\frac{s}{\bar{x}}\right) \cdot 100\%$$

- Stock A:
  - Average price last year = \$50
  - Standard deviation = \$5

$$CV_A = \left(\frac{s}{\bar{x}}\right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

- Stock B:
  - Average price last year = \$100
  - Standard deviation = \$5

$$CV_B = \left(\frac{s}{\bar{x}}\right) \cdot 100\% = \frac{\$5}{\$100} \cdot 100\% = 5\%$$

Both stocks have the same standard deviation, but stock B is less variable relative to its price

Chebychev's Theorem

- For any population with mean μ and standard deviation σ, and $k > 1$, the percentage of observations that fall within the interval

$$[\mu \pm k\sigma]$$

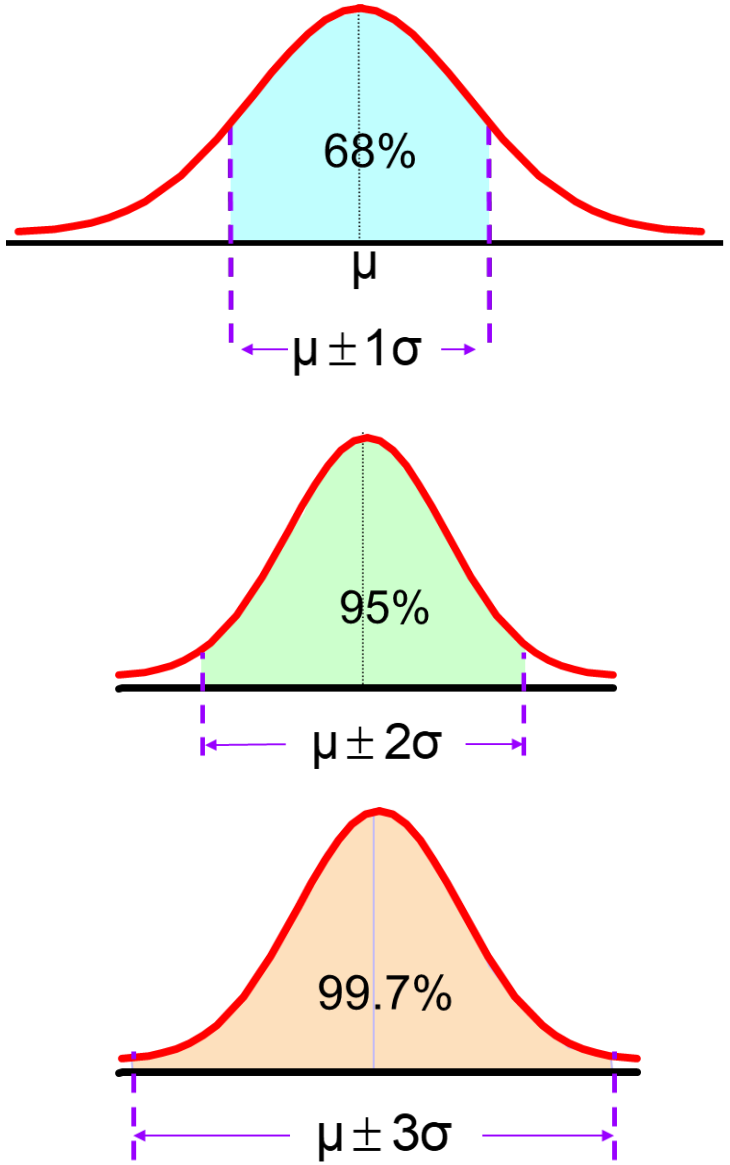
Is *at least*

- Regardless of how the data are distributed, at least $(1 - 1/k^2)$ of the values will fall within k standard deviations of the mean (for $k > 1$)
  - Examples:

At least		within
$(1 - 1/1.5^2) = 55.6\%$	..... $k = 1.5$	$(\mu \pm 1.5\sigma)$
$(1 - 1/2^2) = 75\%$	..... $k = 2$	$(\mu \pm 2\sigma)$
$(1 - 1/3^2) = 89\%$	..... $k = 3$	$(\mu \pm 3\sigma)$

The Empirical Rule

- If the data distribution is **bell-shaped**, then the intervals
  - $\boldsymbol{\mu \pm 1\sigma}$ contains about **68%** of the values in the population or the sample
  - $\boldsymbol{\mu \pm 2\sigma}$ contains about **95%** of the values in the population or the sample
  - $\boldsymbol{\mu \pm 3\sigma}$ contains almost all (about **99.7%**) of the values in the population or the sample



Exercises

- Suppose that the variable Math SAT scores is bell-shaped with a mean of 500 and a standard deviation of 90. Then,
  - 68% of all test takers scored between
    - ?
  - 95% of all test takers scored between
    - ?
  - 99.7% of all test takers scored between
    - ?

- A company produces lightbulbs with a mean lifetime of 1,200 hours and a standard deviation of 50 hours.
  - a. Describe the distribution of lifetimes if the shape of the population is unknown.
  - b. Describe the distribution of lifetimes if the shape of the distribution is known to be bell-shaped.

Locating Extreme Outliers: Z-Score

- To compute the **Z-score** of a data value, subtract the mean and divide by the standard deviation.
- The Z-score is the number of standard deviations a data value is from the mean.
- A data value is considered an extreme outlier if its Z-score is less than -3.0 or greater than +3.0.
- The larger the absolute value of the Z-score, the farther the data value is from the mean.

$$Z = \frac{X - \bar{X}}{s}$$

Where :

- X represents the data value
- X is the sample mean
- s is the sample standard deviation

Example :

- Suppose the mean math SAT score is 490, with a standard deviation of 100.
- Compute the Z-score for a test score of 620.

Exercise

- Consider the company which produces lightbulbs with a mean lifetime of 1,200 hours and a standard deviation of 50 hours.
  - a. Find the z-score for a lightbulb that lasts only 1,120 hours.
  - b. Find the z-score for a lightbulb that lasts 1,300 hours.
- Consider a very large number of students taking a college entrance exam such as the SAT. And suppose the mean score on the mathematics section of the SAT is 570 with a standard deviation of 40.
  - a. Find the z-score for a student who scored 600.
  - b. A student is told that his z-score on this test is -1.5. What was his actual SAT math score?

Exercise

- Open de DDT dataset in Excel
  - For the quantitative variables :
    - Compute all the central tendency indicators and measures
    - Compute all the measures of spread (or variations)
    - Compute all the measures of shape
  - Use either direct excel functions or the analysis toolpalk add-on

Weighted Mean and Measures of Grouped Data

- The weighted mean of a set of data is

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{n} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{n}$$

  - Where w_i is the weight of the i^{th} observation and $n = \sum w_i$

- Use when data is already grouped into n classes, with w_i values in the i^{th} class

Suppose data are grouped into K classes, with frequencies $f_1, f_2, \ldots, f_K$, and the midpoints of the classes are $m_1, m_2, \ldots, m_K$

- For a sample of n observations, the mean is

$$\bar{x} = \frac{\sum_{i=1}^{K} f_i m_i}{n} \quad \text{where} \quad n = \sum_{i=1}^{K} f_i$$

Suppose data are grouped into K classes, with frequencies $f_1, f_2, \ldots, f_K$, and the midpoints of the classes are $m_1, m_2, \ldots, m_K$

- For a sample of n observations, the variance is

$$s^2 = \frac{\sum_{i=1}^{K} f_i (m_i - \bar{x})^2}{n-1}$$

Weighted mean : exercise

- Suppose that a student who completed 15 credit hours during his first semester of college received one A, one B, one C, and one D. Suppose that a value of 4 is used for an A, 3 for a B, 2 for a C, 1 for a D, and 0 for an F.
  - Calculate the student's semester GPA.

- Zack's Investment Research is a leading investment research firm. Zack's will make one of the following recommendations with corresponding weights for a given stock: Strong Buy (1), Moderate Buy (2), Hold (3), Moderate Sell (4), or Strong Sell (5). Suppose that on a particular day, 10 analysts recommend Strong Buy, 3 analysts recommend Moderate Buy, and 6 analysts recommend Hold for a particular stock.
  - Based on Zack's weights, find the mean recommendation.

Mean and variance for groups : exercise

Coffee shop customers were randomly surveyed and asked to select a category that described the cost of their recent purchase. The results were as follows:

Cost (in USD)	$0 < 2$	$2 < 4$	$4 < 6$	$6 < 8$	$8 < 10$
Number of Customers	2	3	6	5	4

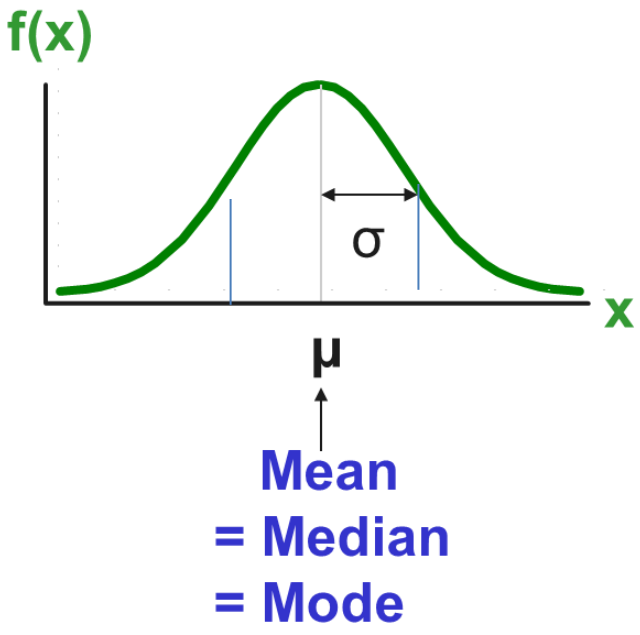
Find the sample mean and standard deviation of these costs.

One of the most frequent probability distribution in Nature

NORMAL PROBABILITY DISTRIBUTION

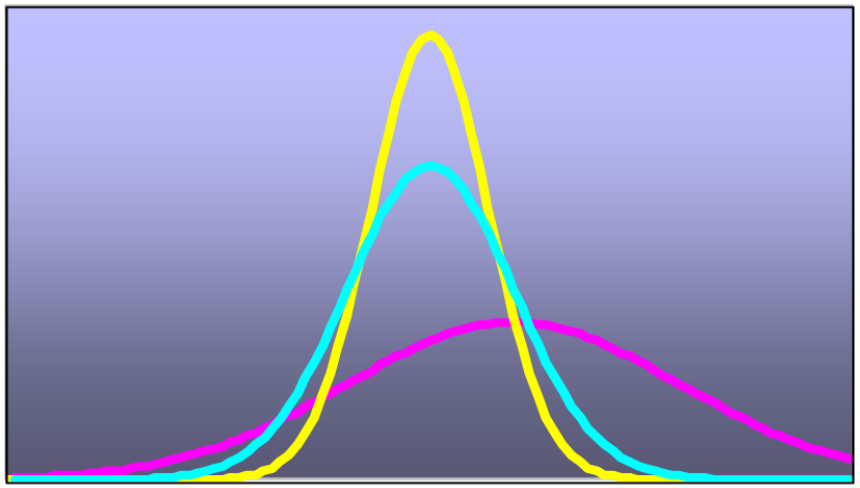
The Normal Distribution

- Bell Shaped
- Symmetrical
- Mean, Median and Mode are Equal
  - Location is determined by the mean, μ
  - Spread is determined by the standard deviation, σ
  - The random variable has an infinite theoretical range:
  - $+\infty$ to $-\infty$

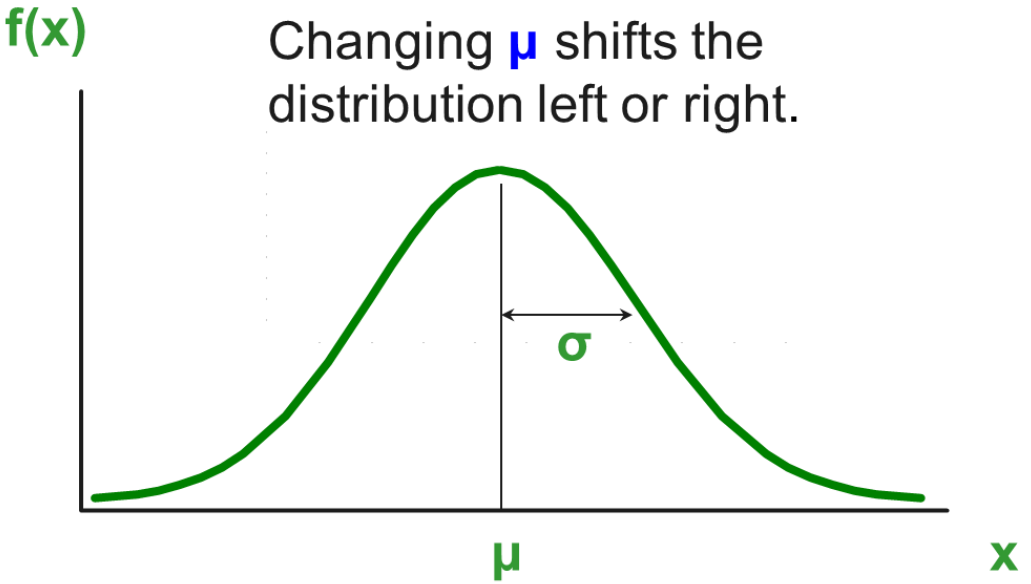


- The normal distribution closely approximates the probability distributions of a wide range of random variables
- Distributions of sample means approach a normal distribution given a "large" sample size
- Computations of probabilities are direct and elegant
- The normal probability distribution has led to good business decisions for a number of applications

Normal Distributions



By varying the parameters μ and σ, we obtain different normal distributions



Given the mean μ and variance σ^2 we define the normal distribution using the notation

$$X \sim N(\mu, \sigma^2)$$

The Normal Probability Density Function

- The formula for the normal probability density function is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

Where e = the mathematical constant approximated by 2.71828

π = the mathematical constant approximated by 3.14159

μ = the population mean

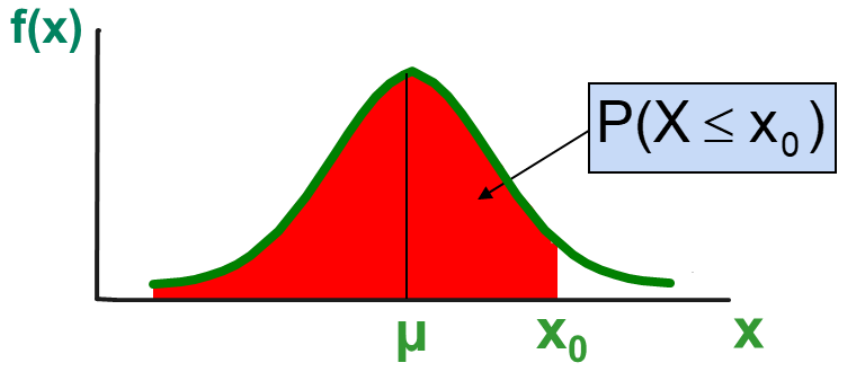
σ^2 = the population variance

x = any value of the continuous variable, $-\infty < x < \infty$

Cumulative Normal Distribution

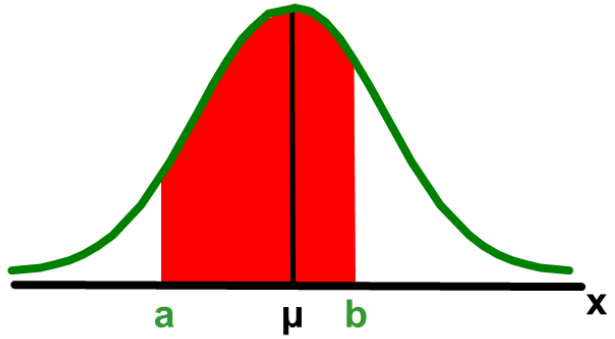
- For a normal random variable X with mean μ and variance σ^2 , i.e., $X \sim N(\mu, \sigma^2)$, the cumulative distribution function is

$$F(x_0) = P(X \le x_0)$$



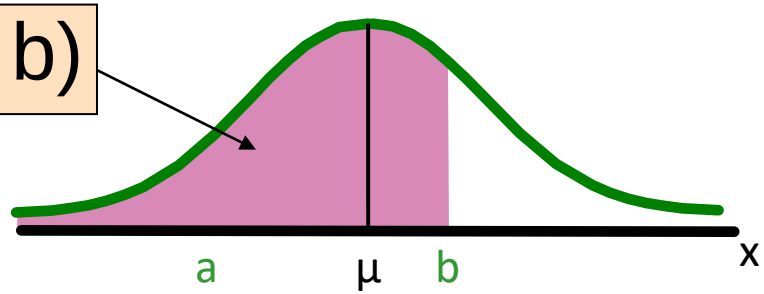
The probability for a range of values is measured by the area under the curve

$$P(a < X < b) = F(b) - F(a)$$

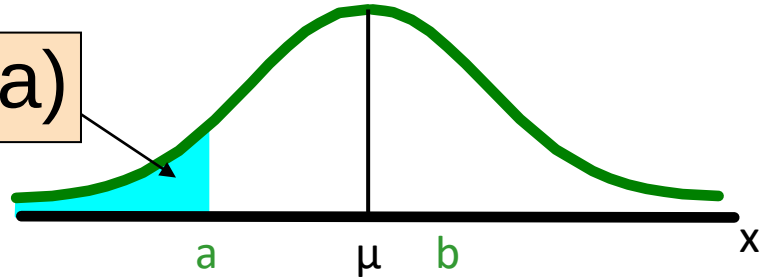


Finding Normal Probabilities

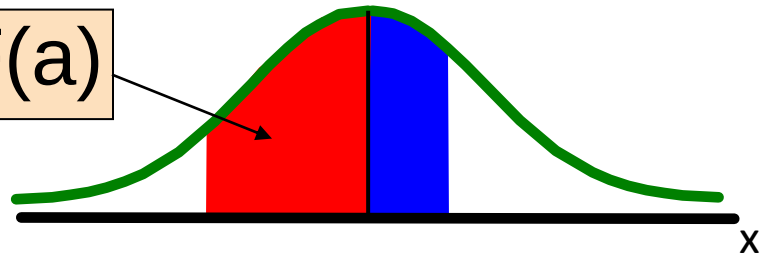
$$F(b) = P(X < b)$$



$$F(a) = P(X < a)$$



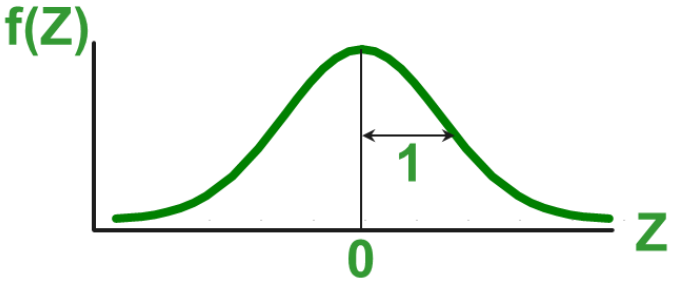
$$P(a < X < b) = F(b) - F(a)$$



The Standard Normal Distribution

- Any normal distribution (with any mean and variance combination) can be transformed into the standardized normal distribution (Z), with mean 0 and variance 1

$$Z \sim N(0,1)$$



- Need to transform X units into Z units by subtracting the mean of X and dividing by its standard deviation

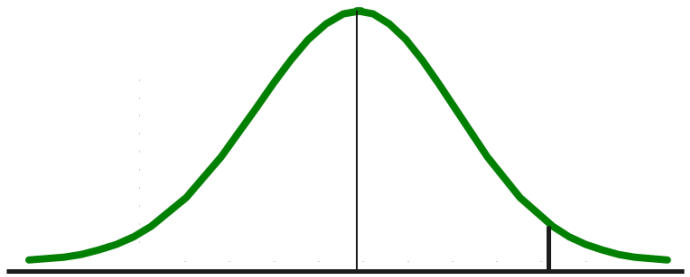
$$Z = \frac{X - \mu}{\sigma}$$

- If X is distributed normally with mean of 100 and standard deviation of 50, the Z value for X = 200 is

$$Z = \frac{X - \mu}{\sigma} = \frac{200 - 100}{50} = 2.0$$

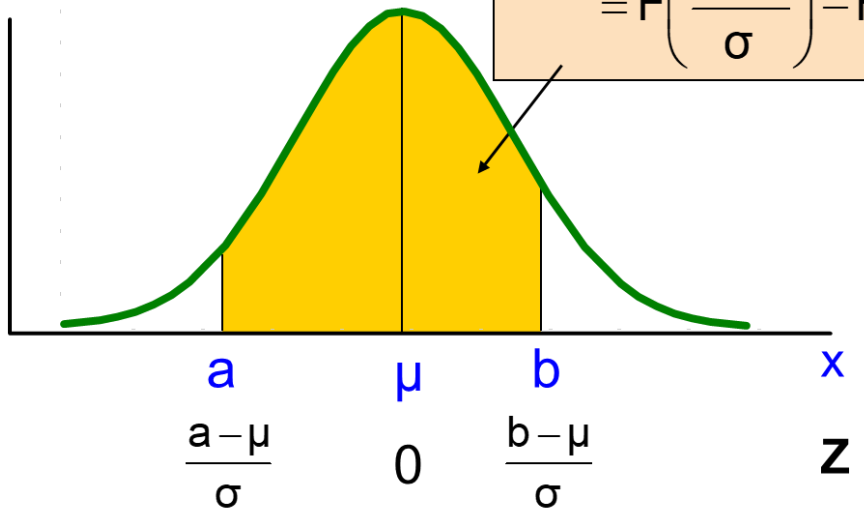
- This says that X = 200 is two standard deviations (2 increments of 50 units) above the mean of 100.

Comparing X and Z units



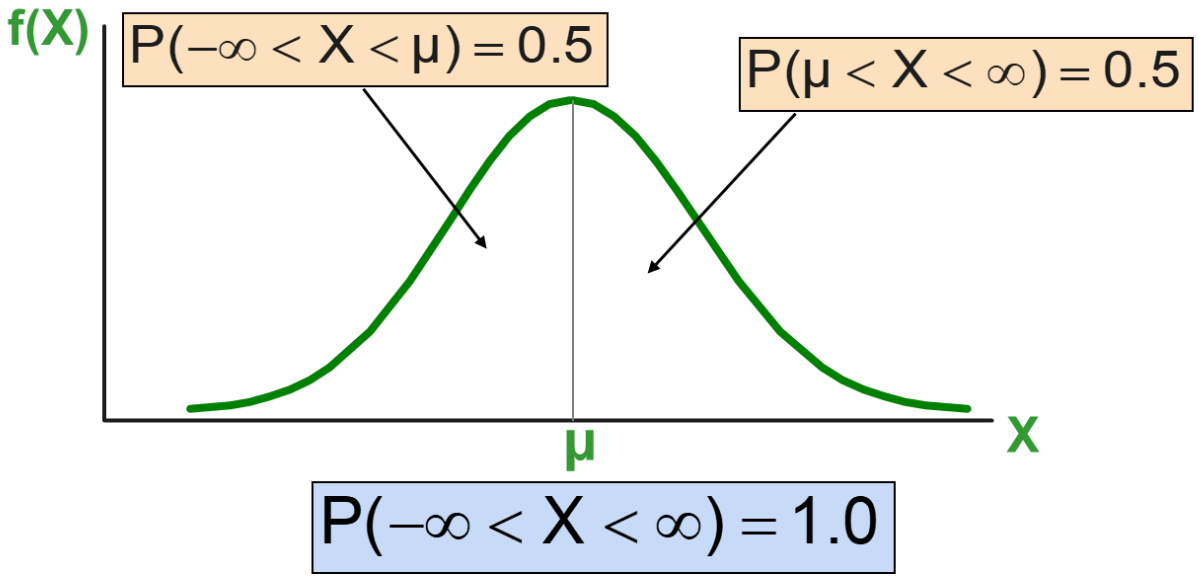
Note that the distribution is the same, only the scale has changed. We can express the problem in original units (X) or in standardized units (Z)

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$
$$= F\left(\frac{b-\mu}{\sigma}\right) - F\left(\frac{a-\mu}{\sigma}\right)$$



Probability as Area Under the Curve

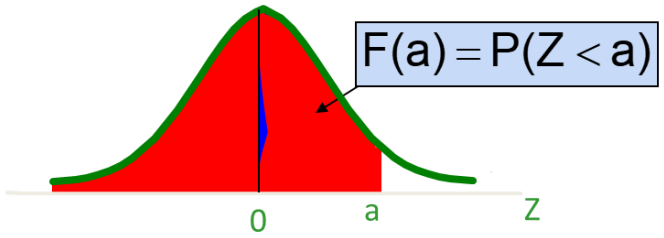
The total area under the curve is 1.0, and the curve is symmetric, so half is above the mean, half is below



- The Standard Normal Distribution table shows values of the cumulative normal distribution function

- For a given Z-value a , the table shows F(a)

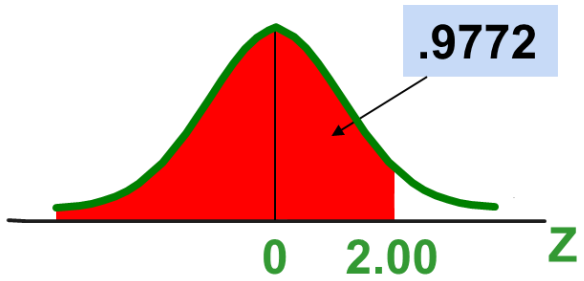
  (the area under the curve from negative infinity to a)



The Standard Normal Table

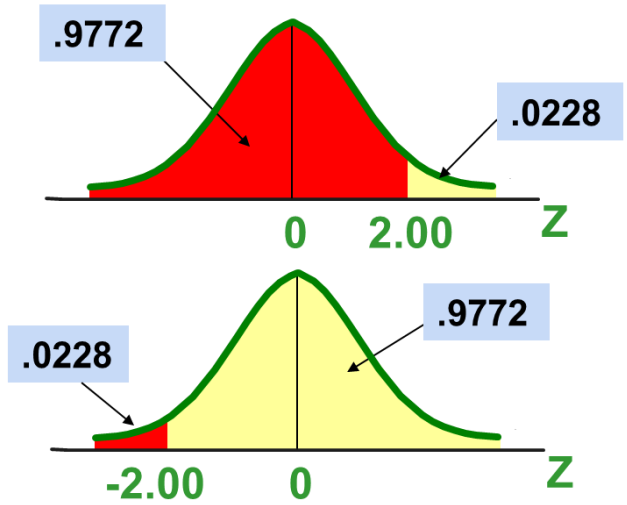
- The table gives the probability F(a) for any value a

Example:
P(Z < 2.00) = .9772



- For negative Z-values, use the fact that the distribution is symmetric to find the needed probability:

Example:
P(Z < -2.00) = 1 – 0.9772
= 0.0228



General Procedure for Finding Probabilities

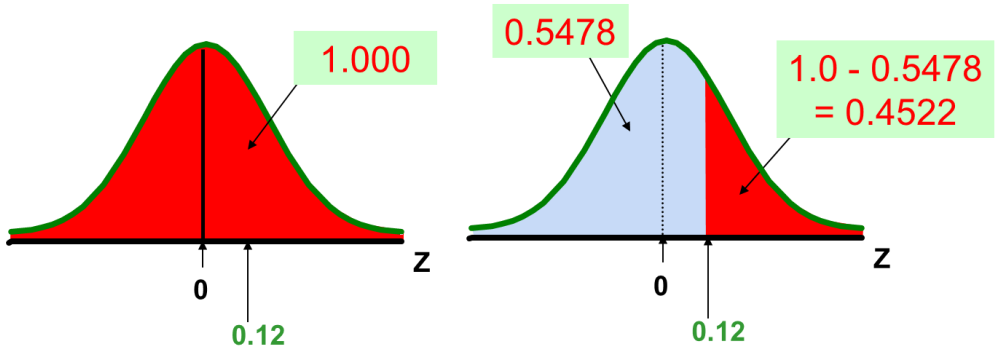
To find P(a < X < b) when X is distributed normally:

- Draw the normal curve for the problem in terms of X
- Translate X-values to Z-values
- Use the Cumulative Normal Table

- Now Find P(X > 8.6)…

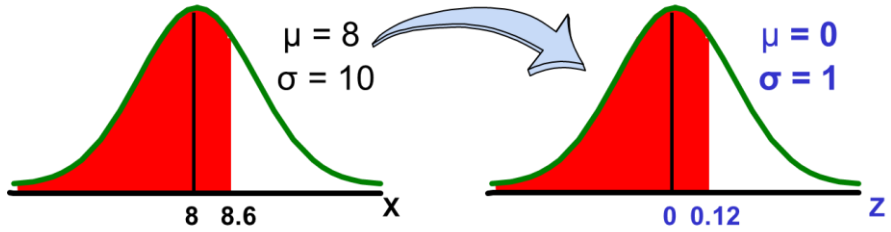
P(X > 8.6) = P(Z > 0.12) = 1.0 - P(Z ≤ 0.12)

= 1.0 - 0.5478 = 0.4522



- Suppose X is normal with mean 8.0 and standard deviation 5.0. Find P(X < 8.6)

$$Z = \frac{X - \mu}{\sigma} = \frac{8.6 - 8.0}{5.0} = 0.12$$

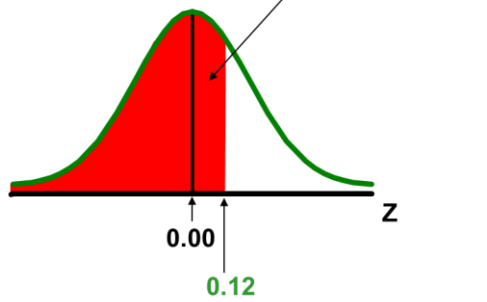


P(X < 8.6) P(Z < 0.12)

Standardized Normal Probability Table (Portion)

z	F(z)
.10	.5398
.11	.5438
.12	.5478
.13	.5517

P(X < 8.6)
= P(Z < 0.12)
F(0.12) = 0.5478



Exercises

- Suppose X is normal with mean 18.0 and standard deviation 5.0.
  - Find P(18 < X < 18.6)
  - Now Find P(17.4 < X < 18)
- A client has an investment portfolio whose mean value is equal to $1,000,000 with a standard deviation of $30,000. He has asked you to determine the probability that the value of his portfolio is between $970,000 and $1,060,000.
  - Determine the Z score of the portfolio values
  - Using your Z score table, give the solution of the question

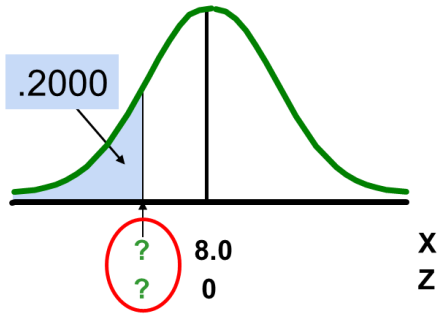
Finding the X value for a Known Probability

- Steps to find the X value for a known probability:
  1. Find the Z value for the known probability
  2. Convert to X units using the formula:

$$X = \mu + Z\sigma$$

Example:

- Suppose X is normal with mean 8.0 and standard deviation 5.0.
- Now find the X value so that only 20% of all values are below this X

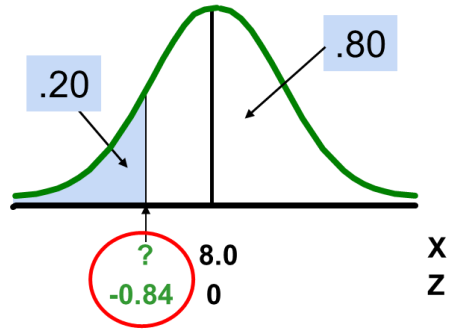


1. Find the Z value for the known probability

Standardized Normal Probability Table (Portion)

z	F(z)
.82	.7939
.83	.7967
.84	.7995
.85	.8023

- 20% area in the lower tail is consistent with a Z value of -0.84



2. Convert to X units using the formula:

$$X = \mu + Z\sigma$$
$$= 8.0 + (-0.84)5.0$$
$$= 3.80$$

So 20% of the values from a distribution with mean 8.0 and standard deviation 5.0 are less than 3.80

Evaluating Normality

Comparing data characteristics to theoretical properties

- Construct charts or graphs
  - For small- or moderate-sized data sets, construct a stem-and-leaf display or a boxplot to check for symmetry
  - For large data sets, check if the histogram or polygon appears to be bell-shaped

- Compute descriptive summary measures
  - Do the mean, median and mode have similar (almost equal) values?
  - Is the interquartile range approximately 1.33 σ?
  - Is the range approximately 6 σ?
  - Are the skewness and kurtosis values between -1 (or -2) and +1 (or +2) ?

- Observe the distribution of the data set
  - Do approximately 2/3 (68%) of the observations lie within mean ±1 standard deviation?
  - Do approximately 95% of the observations lie within mean ±2 standard deviations?

- Evaluate normal probability plot
  - Is the normal probability plot approximately linear (i.e. a straight line) with positive slope?
  - Arrange data into ordered array
  - Find corresponding standardized normal quantile values (Z)
  - Plot the pairs of points with observed data values (X) on the vertical axis and the standardized normal quantile values (Z) on the horizontal axis
  - Evaluate the plot for evidence of linearity

How do we know the extent to which the statistic from our sample, estimating a parameter from our population is reliable ?

CONFIDENCE INTERVAL

Sample Mean

- Let $X_1, X_2, \ldots, X_n$ represent a random sample from a population
- The **sample mean** value of these observations is defined as :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

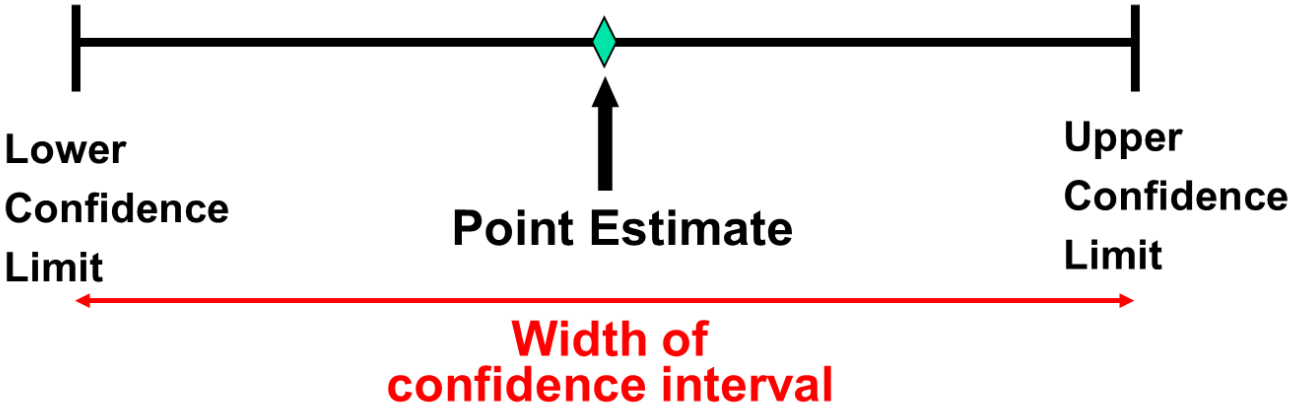
- Different samples of the same size from the same population will yield different sample means
- A measure of the variability in the mean from sample to sample is given by the **Standard Error of the Mean**:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

- Note that the standard error of the mean decreases as the sample size increases

Point and Interval Estimates

- A point estimate is a single number,
- a confidence interval provides additional information about variability



We can estimate a Population Parameter …		with a Sample Statistic (a Point Estimate)
Mean	μ	$\bar{x}$
Proportion	P	$\hat{p}$

- How much uncertainty is associated with a point estimate of a population parameter?
- An interval estimate provides more information about a population characteristic than does a point estimate
- Such interval estimates are called confidence interval estimates
- An interval gives a range of values:
  - Takes into consideration variation in sample statistics from sample to sample
  - Based on observation from 1 sample
  - Gives information about closeness to unknown population parameters
  - Stated in terms of level of confidence
    - Can never be 100% confident

Confidence Interval and Confidence Level

- If $P(a < \theta < b) = 1 - \alpha$ then the interval from a to b is called a $100(1 - \alpha)\%$ confidence interval of θ.
- The quantity $100(1 - \alpha)\%$ is called the confidence level of the interval
  - α is between 0 and 1
  - In repeated samples of the population, the true value of the parameter θ would be contained in $100(1 - \alpha)\%$ of intervals calculated this way.
  - The confidence interval calculated in this manner is written as $a < \theta < b$ with $100(1 - \alpha)\%$ confidence
- Suppose confidence level = 95%
- Also written $(1 - \alpha) = 0.95$
- A relative frequency interpretation:
  - From repeated samples, 95% of all the confidence intervals that can be constructed of size n will contain the unknown true parameter
- A specific interval either will contain or will not contain the true parameter
  - No probability involved in a specific interval

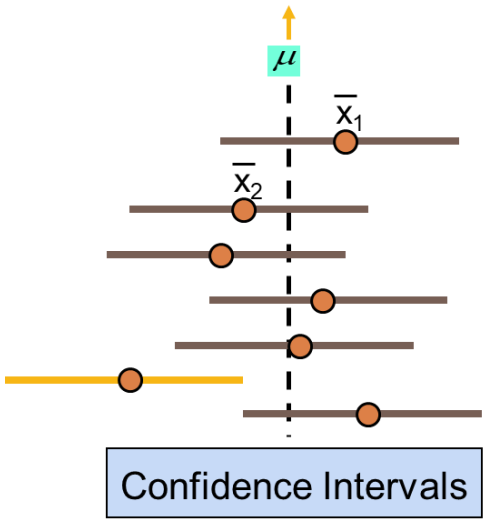
Confidence Interval Example

Cereal fill example

- Population has μ = 368 and σ = 15.
- If you take a sample of size n = 25 you know
  - 368 ± 1.96 * 15 /$\sqrt{25}$ = (362.12, 373.88) contains 95% of the sample means
  - When you don't know μ, you use $\bar{X}$ to estimate μ
    - If $\bar{X}$ = 362.3 the interval is 362.3 ± 1.96 * 15 /$\sqrt{25}$ = (356.42, 368.18)
    - Since 356.42 ≤ μ ≤ 368.18 the interval based on this sample makes a correct statement about μ.

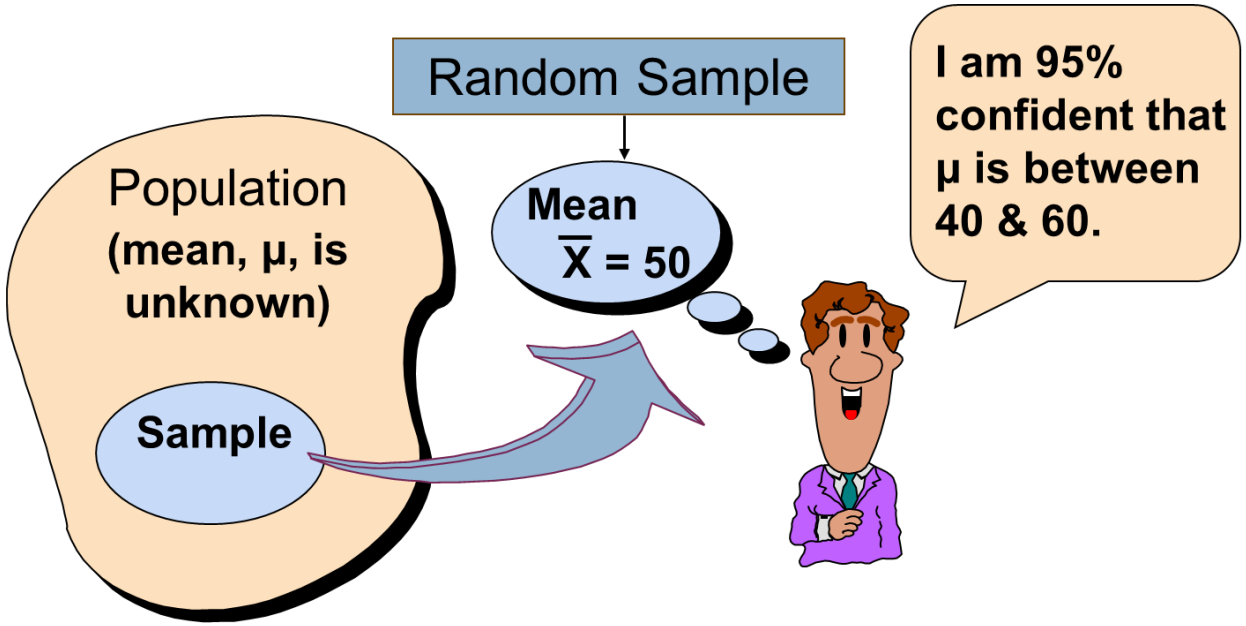
But what about the intervals from other possible samples of size 25?

Sample #	$\bar{X}$	Lower Limit	Upper Limit	Contain μ?
1	362.30	356.42	368.18	Yes
2	369.50	363.62	375.38	Yes
3	360.00	354.12	365.88	No
4	362.12	356.24	368.00	Yes
5	373.88	368.00	379.76	Yes



Estimation Process

- In practice you only take one sample of size n
- In practice you do not know μ so you do not know if the interval actually contains μ
- However you do know that 95% of the intervals formed in this manner will contain μ
- Thus, based on the one sample, you actually selected you can be 95% confident your interval will contain μ (this is a 95% confidence interval)



Note: 95% confidence is based on the fact that we used Z = 1.96.

General Formula

- The general formula for all confidence intervals is:

Point Estimate ± (Critical Value)(Standard Error)

- Where:
  - Point Estimate is the sample statistic estimating the population parameter of interest
  - Critical Value is a table value based on the sampling distribution of the point estimate and the desired confidence level
  - Standard Error is the standard deviation of the point estimate

- Suppose confidence level = 95%
- Also written $(1 - \alpha) = 0.95$
- A relative frequency interpretation:
  - From repeated samples, 95% of all the confidence intervals that can be constructed of size n will contain the unknown true parameter
- A specific interval either will contain or will not contain the true parameter
  - No probability involved in a specific interval

Confidence Interval for μ (σ Known)

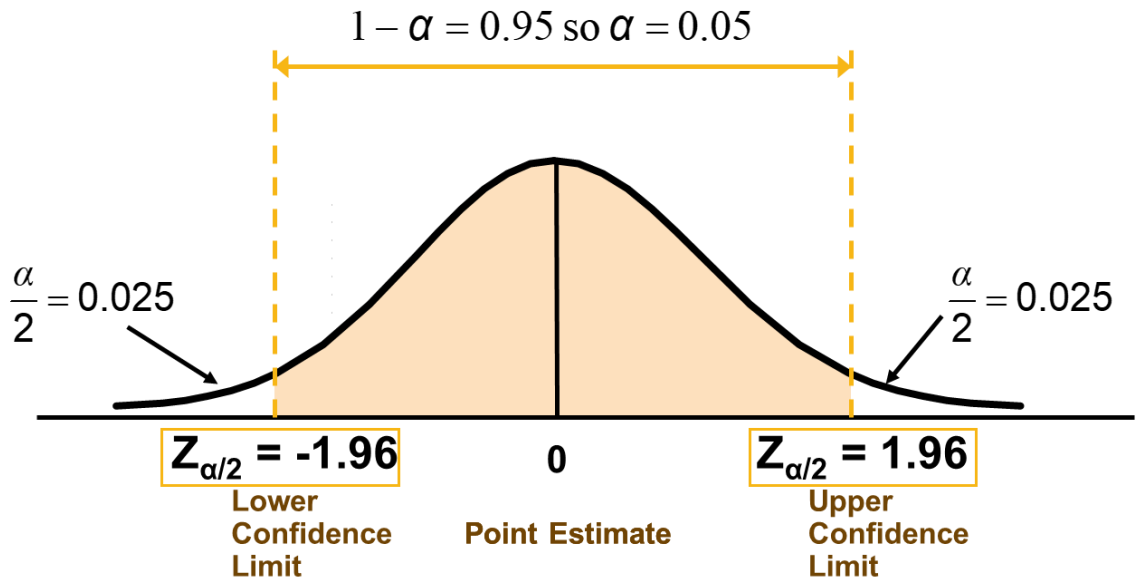
- Assumptions
  - Population variance σ^2 is known
  - Population is normally distributed
  - If population is not normal, use large sample
- Confidence interval estimate:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

(where $z_{\alpha/2}$ is the normal distribution value for a probability of $\alpha/2$ in each tail)

$$Z_{\alpha/2} = \pm 1.96$$

- Consider a 95% confidence interval:



Margin of Error

- The confidence interval,

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- Can also be written as $\bar{x} \pm ME$ where ME is called the margin of error

$$ME = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- The interval width, w, is equal to twice the margin of error

$$ME = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

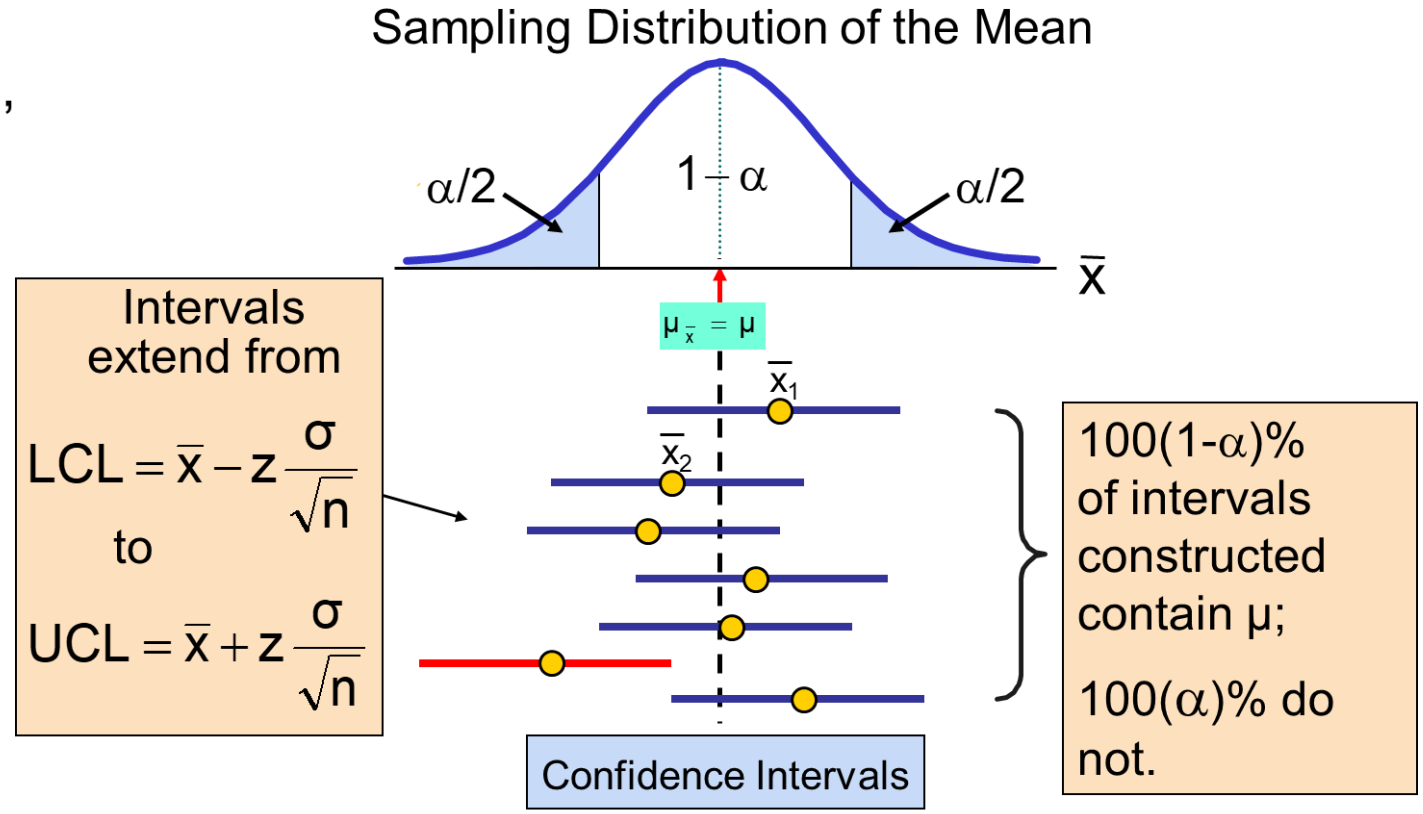
The margin of error can be reduced if

- the population standard deviation can be reduced ($\sigma\downarrow$)
- The sample size is increased ($n\uparrow$)
- The confidence level is decreased, $(1 - \alpha) \downarrow$

Intervals and Level of Confidence (σ Known)

- Commonly used confidence levels are 90%, 95%, 98%, and 99%

Confidence Level	*Confidence Coefficient,* $1-\alpha$	*$Z_{\alpha/2}$ value*
80%	.80	1.28
90%	.90	1.645
95%	.95	1.96
98%	.98	2.33
99%	.99	2.58
99.8%	.998	3.08
99.9%	.999	3.27



Intervals extend from

$$LCL = \bar{x} - z\frac{\sigma}{\sqrt{n}}$$

to

$$UCL = \bar{x} + z\frac{\sigma}{\sqrt{n}}$$

100(1-α)% of intervals constructed contain μ;

100(α)% do not.

Example

- A sample of 11 circuits from a large normal population has a mean resistance of 2.20 ohms. We know from past testing that the population standard deviation is 0.35 ohms.

- Determine a 95% confidence interval for the true mean resistance of the population.

- Solution:

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$= 2.20 \pm 1.96\,(0.35/\sqrt{11})$$

$$= 2.20 \pm 0.2068$$

$$= [1.9932; 2.4068]$$

- We are 95% confident that the true mean resistance is between 1.9932 and 2.4068 ohms

- Although the true mean may or may not be in this interval, 95% of intervals formed in this manner will contain the true mean

Exercise

- Suppose that shopping times for customers at a local mall are normally distributed with known population standard deviation of 20 minutes. A random sample of 64 shoppers in the local grocery store had a mean time of 75 minutes.
  - Find the standard error, margin of error, and the upper and lower confidence limits of a 95% confidence interval for the population mean, μ.
  - Give a clear interpretation of your results
- Repeat the exercise for à 90% and a 99% confidence interval
  - What happens to the confidence interval width when we increase confidence ?
  - What happens to the confidence interval width when we increase the risk level?

Confidence Interval for μ (σ Unknown)

- Do You Ever Truly Know σ ?
  - Probably not!
  - In virtually all real world business situations, σ is not known.
  - If there is a situation where σ is known then μ is also known (since to calculate σ you need to know μ.)
  - If you truly know μ there would be no need to gather a sample to estimate it.
- If the population standard deviation σ is unknown, we can substitute the sample standard deviation, S
  - This introduces extra uncertainty, since S is variable from sample to sample
  - So we **use the t distribution** instead of the normal distribution

- Assumptions
  - Population standard deviation is unknown
  - Population is normally distributed
  - If population is not normal, use large sample
- Use Student's t Distribution
- Confidence Interval Estimate:

$$\overline{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}}$$

(where $t_{\alpha/2}$ is the critical value of the t distribution with n -1 degrees of freedom and an area of α/2 in each tail)

Student's t distribution

- Consider a random sample of n observations
  - with mean $\bar{x}$ and standard deviation s
  - from a normally distributed population with mean μ

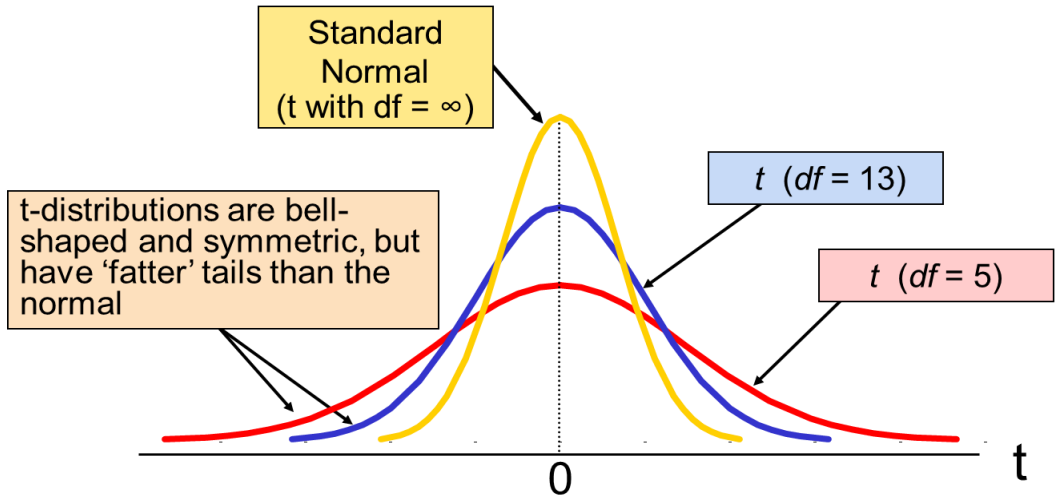
- Then the variable $$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

follows the Student's t distribution with (n-1) degrees of freedom

- The t is a family of distributions
- The t value depends on degrees of freedom (d.f.)
  - Number of observations that are free to vary after sample mean has been calculated

$$d.f. = n - 1$$

Note: t ⟶ Z as n increases



Example of t distribution confidence interval

A random sample of n = 25 has $\overline{X} = 50$ and S = 8. Form a 95% confidence interval for μ

- d.f. = n – 1 = 24, so $t_{\alpha/2} = t_{0.025} = 2.0639$

The confidence interval is

$$\overline{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}}$$

$$= 50 \pm (2.0639) \frac{8}{\sqrt{25}}$$

$$= [46.698; 53.302]$$

- Interpreting this interval requires the assumption that the population you are sampling from is approximately a normal distribution (especially since n is only 25).

- This condition can be checked by:
  - Checking if skewness and kurtosis is close to zero - ideally in [-1;1]
  - creating a Normal probability plot
  - creating a Boxplot
  - checking the numerical descriptive measures' rules

Exercise

- Recently gasoline prices rose drastically. Suppose that a study was conducted using truck drivers with equivalent years of experience to test run 24 trucks of a particular model over the same highway.
  - Estimate the population mean fuel consumption for this truck model with 90% confidence if the fuel consumption, in miles per gallon, for these 24 trucks was as follows:
- 15.5 21.0 18.5 19.3 19.7 16.9 20.2 14.5 16.5 19.2 18.7 18.2 18.0 17.5 18.5 20.5 18.6 19.1 19.8 18.0 19.8 18.2 20.3 21.8
- The data are stored in the data file **Trucks**.

Confidence Interval for Proportions

- An interval estimate for the population proportion (π) can be calculated by adding an allowance for uncertainty to the sample proportion (p)

- The distribution of the sample proportion is approximately normal if the sample size is large, with standard deviation

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

- We will estimate this with the sample data:

$$s_p = \sqrt{\frac{p(1-p)}{n}}$$

- Upper and lower confidence limits for the population proportion are calculated with the formula

$$p \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

- where
  - $Z_{\alpha/2}$ is the standard normal value for the level of confidence desired
  - p is the sample proportion
  - n is the sample size

- Note: we must have $np > 5$ and $n(1-p) > 5$

Exercise

- Management wants an estimate of the proportion of the corporation's employees who favor a modified bonus plan. From a random sample of 344 employees, it was found that 261 were in favor of this particular plan.
  - Find a 90% confidence interval estimate of the true population proportion that favors this modified bonus plan.

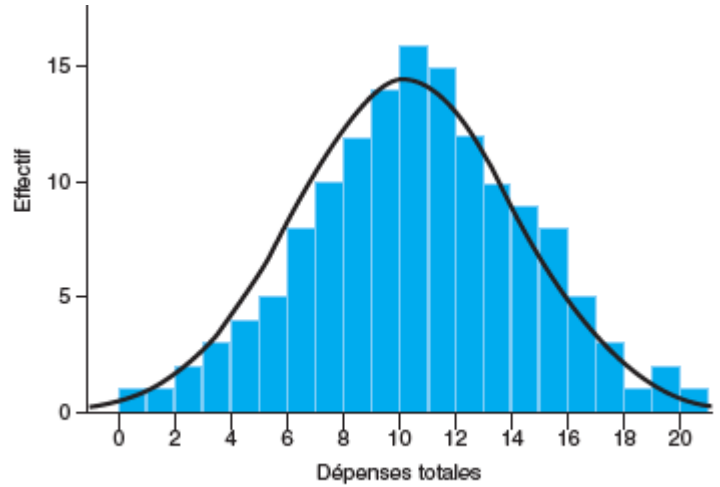
How do we know if the value of a (population) parameter is significantly different from another value ?

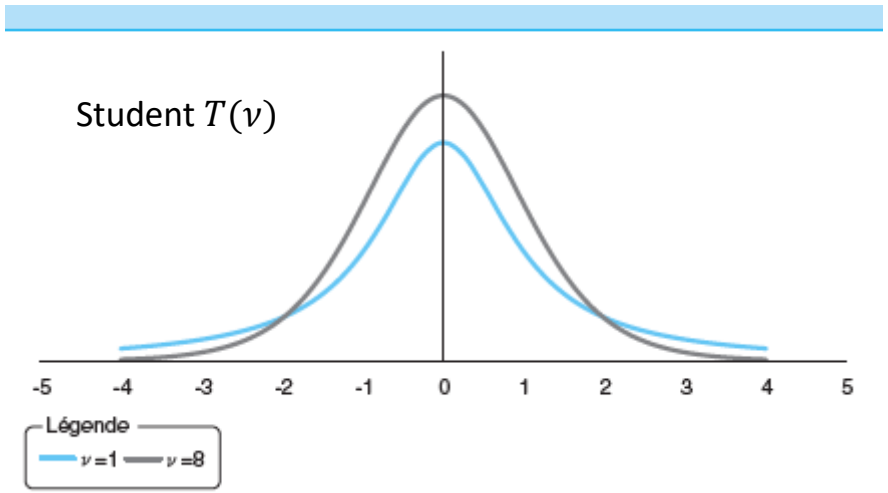
HYPOTHESIS TESTING

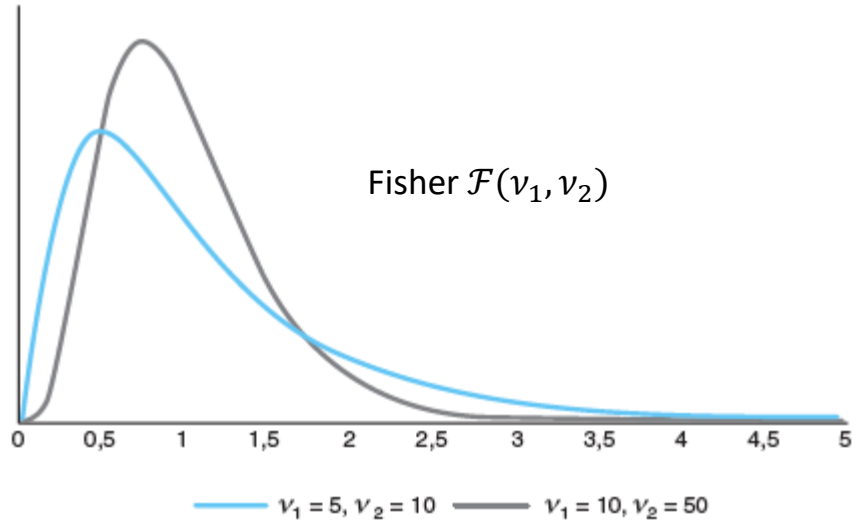
Probability distributions : summary

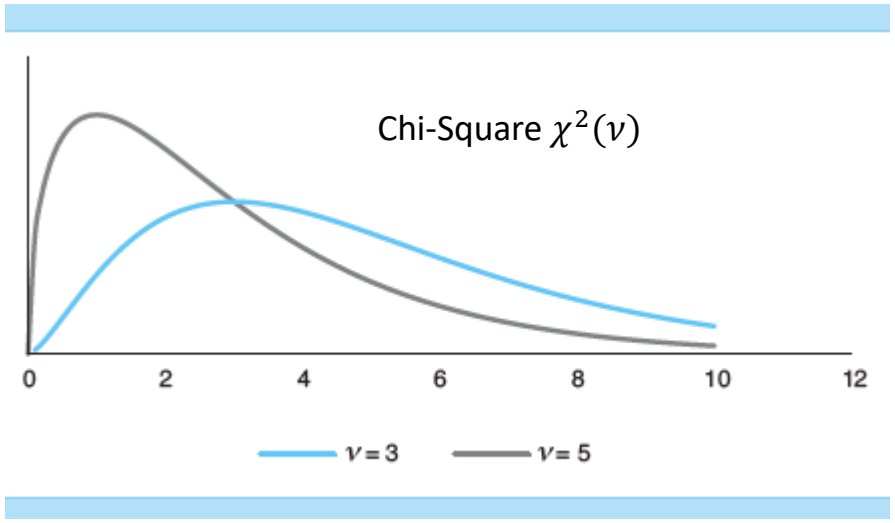
- The probability distribution of a (random) variable X gives us information about how likely is one to draw a given value of X
  - The area under the curve measures the probability of drawing the values in the delimited area
  - For example: if X follows a normal distribution N(μ,σ) :
    - 68% of the values of X range from μ-σ to μ+σ
    - 95% of the values of X range from μ-1.96*σ to μ+1.96*σ
      - The probability that a value X lies out of the interval [μ-1.96*σ ; μ+1.96*σ] is less than 5%
      - 1.96 is called critical (threshold) z score value for the normal distribution
- The critical (threshold) values change according to the distribution of the variable

Examples of distributions









Concepts of Hypothesis Testing

- A hypothesis is a claim (assumption) about a population parameter:

  - population mean

Example: The mean monthly cell phone bill of this city is μ = $52

  - population proportion

Example: The proportion of adults in this city with cell phones is P = .88

The Null Hypothesis, H_0

- States the assumption (numerical) to be tested

Example: The average number of TV sets in U.S. Homes is equal to three ($H_0 : \mu = 3$)

- Is always about a population parameter, not about a sample statistic

$H_0 : \mu = 3$

$H_0 : \bar{x} = 3$

Null and Alternate Hypotheses

The Null Hypothesis, H_0

- Begin with the assumption that the null hypothesis is true
  - Similar to the notion of innocent until proven guilty
- Refers to the status quo
- Always contains "=" , "≤" or "≥" sign
- May or may not be rejected

The Alternative Hypothesis, H_1

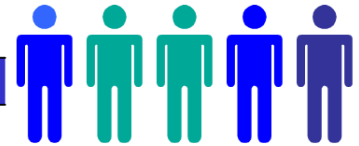
- Is the opposite of the null hypothesis
  - e.g., The average number of TV sets in U.S. homes is not equal to 3 (H_1: $\mu \neq 3$)
- Challenges the status quo
- Never contains the "=" , "≤" or "≥" sign
- May or may not be supported
- Is generally the hypothesis that the researcher is trying to support

Hypothesis Testing Process

Claim: the population mean age is 50. (Null Hypothesis: H_0: μ = 50)

Population

Now select a random sample



Sample

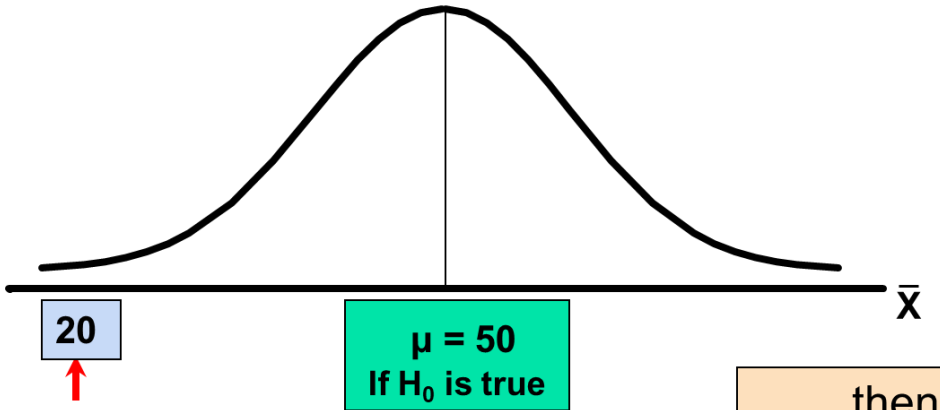
Suppose the sample mean age is 20: $\bar{x}$ = 20

Is $\bar{x}$=20 likely if μ = 50?

If not likely,

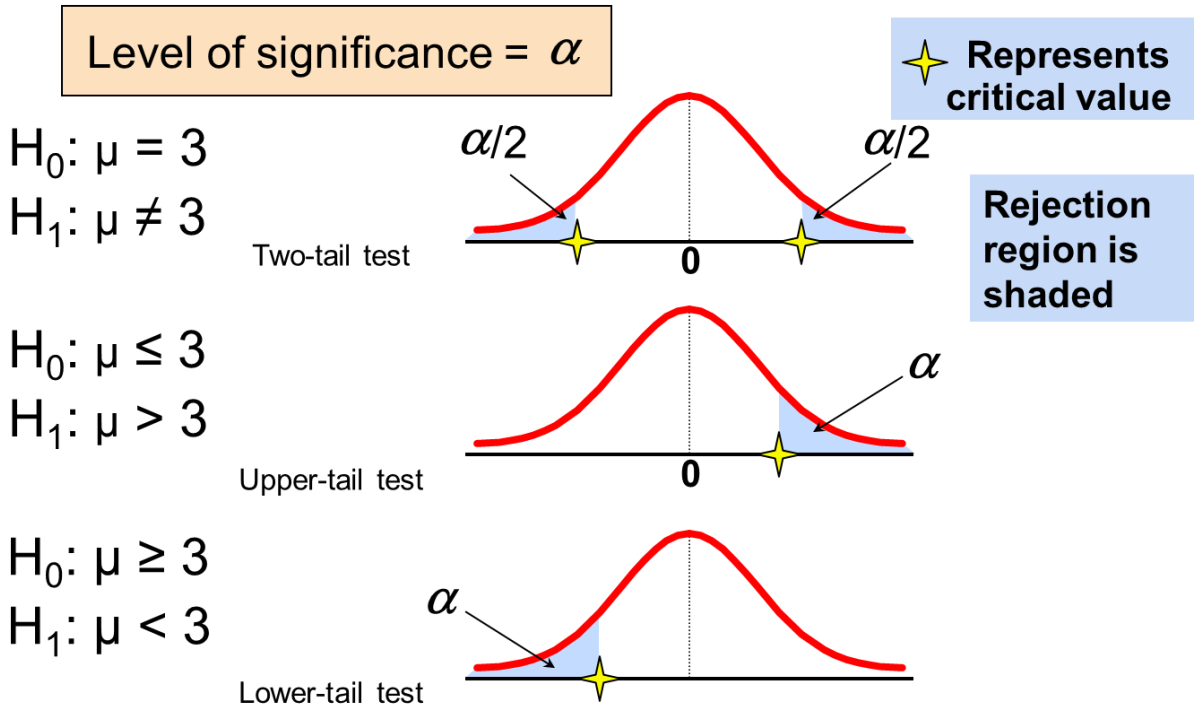
REJECT Null Hypothesis

Sampling Distribution of $\bar{X}$



Level of Significance and Rejection Region

- Level of Significance
  - Defines the unlikely values of the sample statistic if the null hypothesis is true
    - Defines rejection region of the sampling distribution
  - Is designated by α , (level of significance)
    - Typical values are 0.01, 0.05, or 0.10
  - Is selected by the researcher at the beginning
  - Provides the critical value(s) of the test



Two ways to be wrong

- We cannot be sure of obtaining the correct result for our hypothesis test and there are two ways of getting it wrong:
  - Type I Error: the null hypothesis is true but we reject it
  - Type II Error: the null hypothesis is false but we fail to reject it
- The analogy with the jury trial (where the initial assumption is innocence) is
  - convicting an innocent person
  - letting a guilty person go free.
- We tend to think that in this context, I is worse.

Possible Hypothesis Test Outcomes

	Actual Situation	
Decision	H_0 True	H_0 False
Fail to Reject **H_0**	**Correct Decision** **$(1 - \alpha)$**	**Type II Error** **(β)**
Reject **H_0**	**Type I Error** **(α)**	**Correct Decision** **$(1 - \beta)$**

Key:
Outcome
(Probability)

$(1 - \beta)$ is called the power of the test

A Trial as a Hypothesis Test

- Think about the logic of jury trials:
  - To prove someone is guilty, we start by assuming they are innocent.
  - We retain the assumption of innocence, until the facts make it unlikely beyond "a reasonable doubt"
  - Then, and only then, we reject the hypothesis of innocence and declare the person guilty.

- If the evidence is not strong enough to reject the presumption of innocence, the jury returns with a verdict of "not guilty."
  - The jury does not say that the defendant is innocent.
  - All it says is that there is not enough evidence to convict, to reject innocence.
  - The defendant may be innocent or may be guilty, but the jury is not sure (confident) of guilt.

- Expressed statistically, we will fail to reject the null hypothesis.
  - We never declare the null hypothesis to be true, because we simply do not know whether it's true or not.
  - Sometimes in this case we say that the null hypothesis has been retained

Hypothesis Testing for the Mean (using critical values)

- If σ is **known**
- Convert sample statistic $\bar{X}$ to Z_{stat} test statistic where :

$$Z_{stat} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- Z_{stat} follows a N(0,1) distribution
- Determine the critical Z values for a specified level of significance $\boldsymbol{\alpha}$ from the Z-table or using a computer
- Decision Rule: If the test statistic falls in the rejection region, reject H_0; otherwise do not reject H_0

- If σ is **unknown,** use the sample standard deviation s
- Convert sample statistic $\bar{X}$ to t_{stat} test statistic where :

$$t_{stat} = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

- t_{stat} follows a student t(n-1) distribution
- Determine the critical t values for a specified level of significance $\boldsymbol{\alpha}$ from the t-table or using a computer
- Decision Rule: If the test statistic falls in the rejection region, reject H_0; otherwise do not reject H_0

Examples (two tailed)

- **Test the claim that the true mean diameter of a manufactured bolt is 30mm (Assume σ = 0.8)**
  - The following are given : $\bar{X}$ =29.84, n=100

1. Choose a risk level (α)
2. State H_0 and H_1 for the case
3. Determine the appropriate Statistic
4. Find the critical values
5. Make a final decision

- The average cost of a hotel room in New York is said to be $168 per night. To determine if this is true, a random sample of 25 hotels is taken and resulted in an $\bar{X}$ of $172.50 and an **s** of $15.40.
- Test the appropriate hypotheses at α=0.05 and α=0.10

Examples (one tailed)

- The production manager of Northern Windows, Inc., has asked you to evaluate a proposed new procedure for producing its Regal line of double-hung windows. The present process has a mean production of 80 units per hour with a population standard deviation of σ = 8. The manager does not want to change to a new procedure unless there is strong evidence that the mean production level is higher with the new process. An estimation on sample of 25 units yields a mean value of 83 units per hour.

1. Choose a risk level (α)
2. State H_0 and H_1 for the case
3. Determine the appropriate Statistic
4. Find the critical values
5. Make a final decision

- The production manager of Twin Forks Ball Bearing, Inc., has asked your assistance in evaluating a modified ball bearing production process. When the process is operating properly, the process produces ball bearings whose weights are normally distributed with a population mean of 5 ounces and a population standard deviation of 0.1 ounce. A new raw-material supplier was used for a recent production run, and the manager wants to know if that change has resulted in a lowering of the mean weight of the ball bearings. There is no reason to suspect a problem with the new supplier, and the manager will continue to use the new supplier unless there is strong evidence that underweight ball bearings are being produced. An estimation on a sample of 16 yields a mean weight of 4.962
  - What is the decision you recommend ?

Exercises : use critical values

- The production manager of Circuits Unlimited has asked for your assistance in analyzing a production process. This process involves drilling holes whose diameters are normally distributed with a population mean of 2 inches and a population standard deviation of 0.06 inch. A random sample of nine measurements had a sample mean of 1.95 inches. Use a significance level of a = 0.05 to determine if the observed sample mean is unusual and, therefore, that the drilling machine should be adjusted.

- Grand Junction Vegetables is a producer of a wide variety of frozen vegetables. The company president has asked you to determine if the weekly sales of 16-ounce packages of frozen broccoli has increased. The mean weekly number of sales per store has been 2,400 packages over the past 6 months. You have obtained a random sample of sales data from 134 stores for your study

Descriptive Statistics Broccoli:

Variable	N	Mean	Standard Error Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Broccoli	134	3593	425	4919	156	707	2181	2300	27254

Hypothesis testing with p-value

- Example: How likely is it to see a sample mean of 2.84 (or something further from the mean, in either direction) if the true mean is $\mu = 3.0$?

assume $\sigma=0.8$ et $n=100$

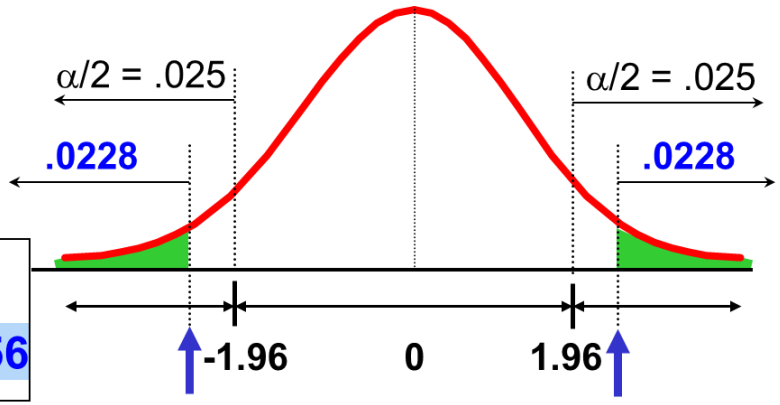
$\bar{x} = 2.84$ is translated to a z score of z = -2.0

$P(z < -2.0) = .0228$

$P(z > 2.0) = .0228$

p-value

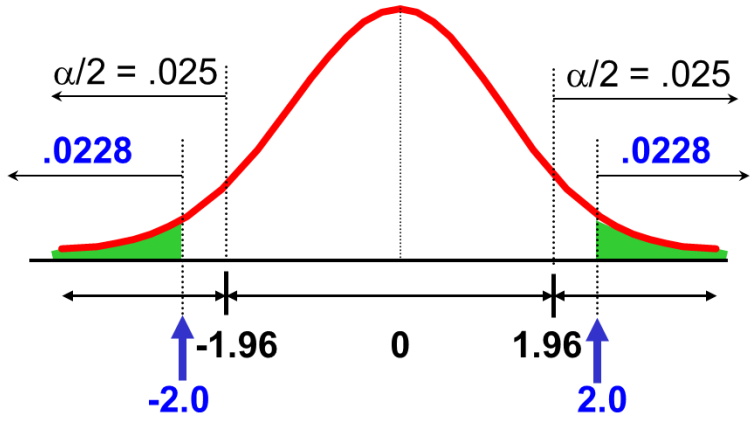
= .0228 + .0228 = .0456



- Compare the p-value with α
  - If p-value < α , reject H_0
  - If p-value ≥ α , do not reject H_0

Here: p-value = .0456
α = .05

Since .0456 < .05, we reject the null hypothesis



Exercises : use p-values

- The production manager of Circuits Unlimited has asked for your assistance in analyzing a production process. This process involves drilling holes whose diameters are normally distributed with a population mean of 2 inches and a population standard deviation of 0.06 inch. A random sample of nine measurements had a sample mean of 1.95 inches. Use a significance level of a = 0.05 to determine if the observed sample mean is unusual and, therefore, that the drilling machine should be adjusted.

- Grand Junction Vegetables is a producer of a wide variety of frozen vegetables. The company president has asked you to determine if the weekly sales of 16-ounce packages of frozen broccoli has increased. The mean weekly number of sales per store has been 2,400 packages over the past 6 months. You have obtained a random sample of sales data from 134 stores for your study

Descriptive Statistics Broccoli:									
Variable	N	Mean	Standard Error Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Broccoli	134	3593	425	4919	156	707	2181	2300	27254

Examples

- **Test the claim that the true mean diameter of a manufactured bolt is 30mm (Assume σ = 0.8)**
  - The following are given : $\bar{X}$ =29.84, n=100

1. Choose a risk level (α)
2. State H_0 and H_1 for the case
3. Determine the appropriate Statistic
4. Find the p-values
5. Make a final decision

- The average cost of a hotel room in New York is said to be $168 per night. To determine if this is true, a random sample of 25 hotels is taken and resulted in an $\bar{X}$ of $172.50 and an **s** of $15.40.
- Test the appropriate hypotheses at α=0.05 and α=0.10

How do we ascertain whether two variables are significantly linked?

SIMPLE BIVARIATE ANALYSIS

Types of associations

- Between two categorical variables
  - Contingency table
  - H_0 : the two variables are independent
    - Chi square (χ^2) statistics to ascertain probability of association
    - Cramer's V to measure effect size

- Between two continuous variables
  - Regular « individual vs variable » table
  - H_0 : the two variables are not correlated
    - T Student / F Fisher statistics to ascertain probability of association
    - Coefficient r to measure effect size

Independence Chi Square (χ^2) test

- You want to study the link (dependence) between 2 categorical variables
  - H_0: the two variables are independent
- You perform a Chi Square (χ^2) test
  - You determine the probability value of the computed Chi Square of your contingency table
  - If that value is **greater** than α (often = 0.05) then you **cannot** reject H_0
    - i.e the two variables ARE NOT significantly linked
  - If that value is **smaller** than α (often = 0.05) then you **can** reject H_0
    - i.e the two variables ARE significantly linked

The principle

- Consider a contingency table
  - n_{ij} indicates the actual (observed) frequency of the cell on row (line) i and column j
  - $E(n_{ij})$ indicates the expected value under assumption of independence (H_0).
  - $X^2 = \sum_{i,j} \frac{[n_{ij} - E(n_{ij})]^2}{E(n_{ij})}$ follows a Chi Square distribution (if $E(n_{ij}) > 5$)

	X_1	X_j	X_c	$Total$
Y_1	n_{11}		n_{1c}	l_1
Y_i		n_{ij}		l_i
Y_l	n_{l1}		n_{lc}	l_l
$Total$	c_1	c_j	c_c	n

The formula

- If the row and column variables are independent: $E(n_{ij}) = \frac{l_i c_j}{n}$

$$X^2 = \sum_{i,j} \frac{\left[n_{ij} - \frac{l_i c_j}{n}\right]^2}{\frac{l_i c_j}{n}}$$

	X_1	X_j	X_c	$Total$
Y_1	n_{11}		n_{1c}	l_1
Y_i		n_{ij}		l_i
Y_l	n_{l1}		n_{lc}	l_l
$Total$	c_1	c_j	c_c	n

  - l_i and c_j indicate respectively the total (marginal) frequency of row i and column j
  - X^2 follows a Chi Square distribution with a degree of freedom = $(l-1)*(c-1)$
    - where $\boldsymbol{l}$ = number of modalities on line and $\boldsymbol{c}$= number of modalities on column
    - We need only to compare X^2 with the threshold values of the Chi Square (χ^2) distribution

- Effect size : $\phi = \sqrt{\frac{X^2}{n}}$ and $V_{Cramer} = \sqrt{\frac{X^2}{n.\min[(l-1),(c-1)]}}$

Example

- We want to test the relationship between smoking and lung cancer
  - Our variables are :
    - **Smoking** which has two modalities : Yes/No
    - **Lung cancer** which has also two modalities : Yes/No
  - The questions are
    - Is there a relationship between contracting lung cancer and smoking ?
    - If so which modalities contribute significantly to the relationship ?
  - The contigency table is given as following :

Observed Values		Lung Cancer		
		Yes	No	Total
Smoking	Yes	50	33	83
	No	26	39	65
	Total	76	72	148

Exercise

- We want to evaluate the effects of 3 weight loss diets
- There are 4 types of effects:
  - No result (no observed loss of weight)
  - Small result (weight loss < 20% of the objective)
  - Average result (weight loss between 20% and 50% of the objective)
  - High result (weight loss > 50% of the objective)
- Is there any significant link between the diet type and the weight loss results variables ?
  - Data : Diet_Raw_Data.xlsx

		Results			
		None	Small	Average	Large
Diet	A	15	21	45	13
	B	26	31	34	5
	C	33	17	49	20

Exercise Part 1

- Read the data <Diet_Raw_Data.xlsx>
- Using Excel's cross (pivot) table function « tableau croisé dynamique) », generate a contingency table
  - Make a 'copy as value' of the table and add the marginal (total values) : this is the table of the observed values
  - Compute the table of expected values : $E(n_{ij}) = \frac{l_i c_j}{n}$
  - Compute the residuals table $r_{ij} = \frac{\left[n_{ij} - \frac{l_i c_j}{n}\right]}{\sqrt{\frac{l_i c_j}{n}}}$ which follows a standard normal N(0,1) distribution for n>30
  - Compute the Chi-square table where the cells are $\chi^2{}_{ij} = \frac{\left[n_{ij} - \frac{l_i c_j}{n}\right]^2}{\frac{l_i c_j}{n}}$
    - l_i = sum of the row values (row or line marginals)
    - c_j = sum of the column value (column marginals)
  - Compute $X^2 = \sum_{cells} \frac{\left[n_{ij} - \frac{l_i c_j}{n}\right]^2}{\frac{l_i c_j}{n}}$
  - Which modalities contribute the most to the X^2 (the r_{ij} greater than 1.96 or smaller than -1.96) ?

Exercise Part 2

- Compare X^2 with the threshold (5%) value of a Chi Square distribution with (L-1)*(C-1) degrees of freedom
  - L = number of modalities for the row (line) variable
  - C = number of modalities for the column variable
- 2 Methods
  - **Method 1** : compute the threshold (critical) value of a chi square distribution where the (right side) probability = α (0.05)
    - If the value of $X^2 > \chi^2$(threshold) then reject $H_0 \Rightarrow$ the two variables are linked (there is dependence)
    - If the value of $X^2 \leq \chi^2$(threshold) then DO NOT reject $H_0 \Rightarrow$ the two variables are independent (not linked)
  - **Method 2** : compute the p-value of X^2 using a χ^2 distribution with (L-1)*(C-1) degrees of freedom
    - If p-value < α (0.05) then reject $H_0 \Rightarrow$ the two variables are linked (there is dependence)
    - If p-value ≥ α (0.05) then DO NOT reject $H_0 \Rightarrow$ the two variables are independent (not linked)

Exercise Part 3

- We repeat the previous exercise with a different data format
  - We want to evaluate the effects of 3 weight loss diets
  - There are 4 types of effects:
    - 0 = No result (no observed loss of weight)
    - 1 = Small result (weight loss < 20% of the objective)
    - 2 = Average results (weight loss between 20% and 50% of the objective)
    - 3 = High results (weight loss > 50% of the objective)
  - Is there any significant link between the diet type and the weight loss results variables ?
    - Data : Diet_Raw_Data_Numbers.xlsx
  - Compute the Effect Size
- Make a summary (synthesis) of your final conclusion(s)

The fast way using Excel functions

- TEST.KHIDEUX(plage_réelle;plage_attendue)
  - Plage réel : observed table
  - Plage attendu : expected table
- Example (diet vs result) : =TEST.KHIDEUX(B10:E12;B18:E20)
  - P value of our data : 0.003873389 < 0.05
- Chi square value:
  - Degree freedom (ligne-1)*(colonne-1) = 6
  - Compute Chi Square =KHIDEUX.INVERSE(probabilité ; degré de liberté) = 19.17797217

Linear Correlation Test (Pearson)

- We want to test if there is a significant association between **two continuous** variables
- The covariance between two variables X and Y indicates if there is an association between the variation of the two variables around their respective means

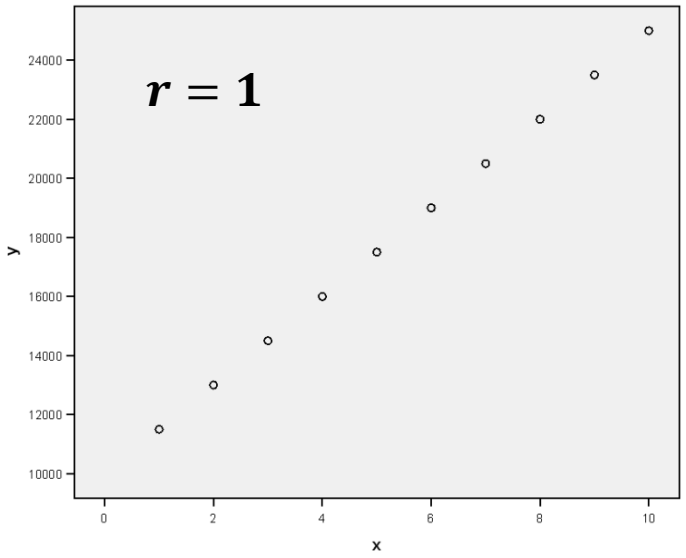
$$cov(X,Y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

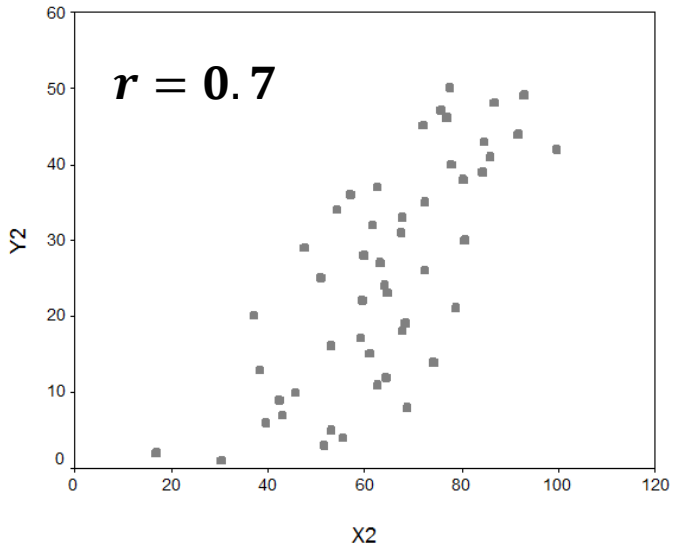
- Correlation is a standardized measure of the covariance

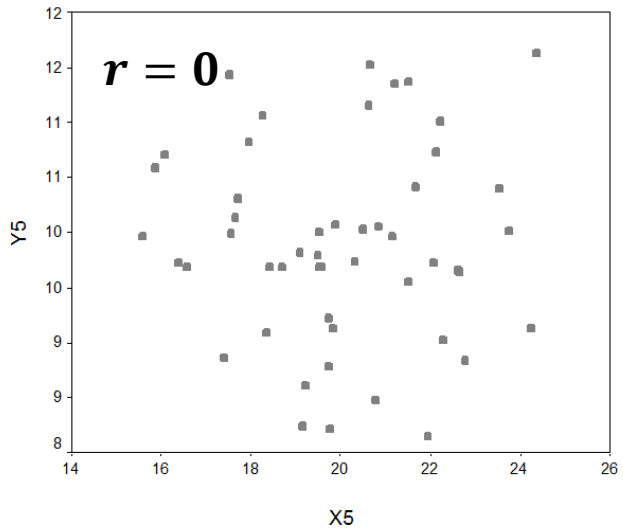
$$r = \frac{cov(X,Y)}{s_X s_Y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

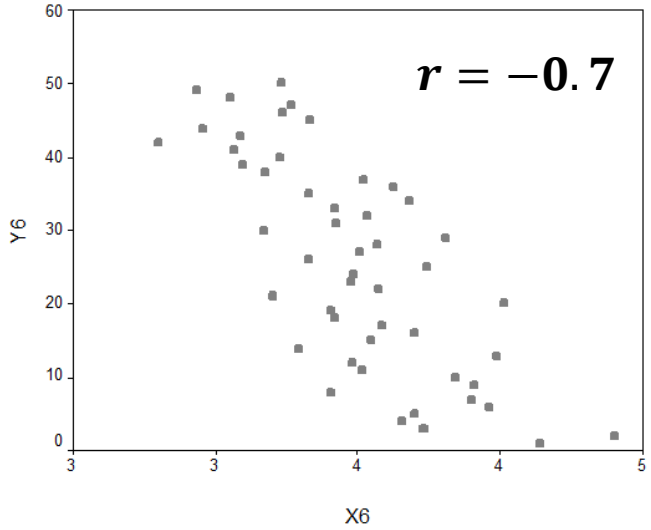
s_X and s_Y are the respective standard deviations of X and Y

Examples (scatter plots)









Significance Testing

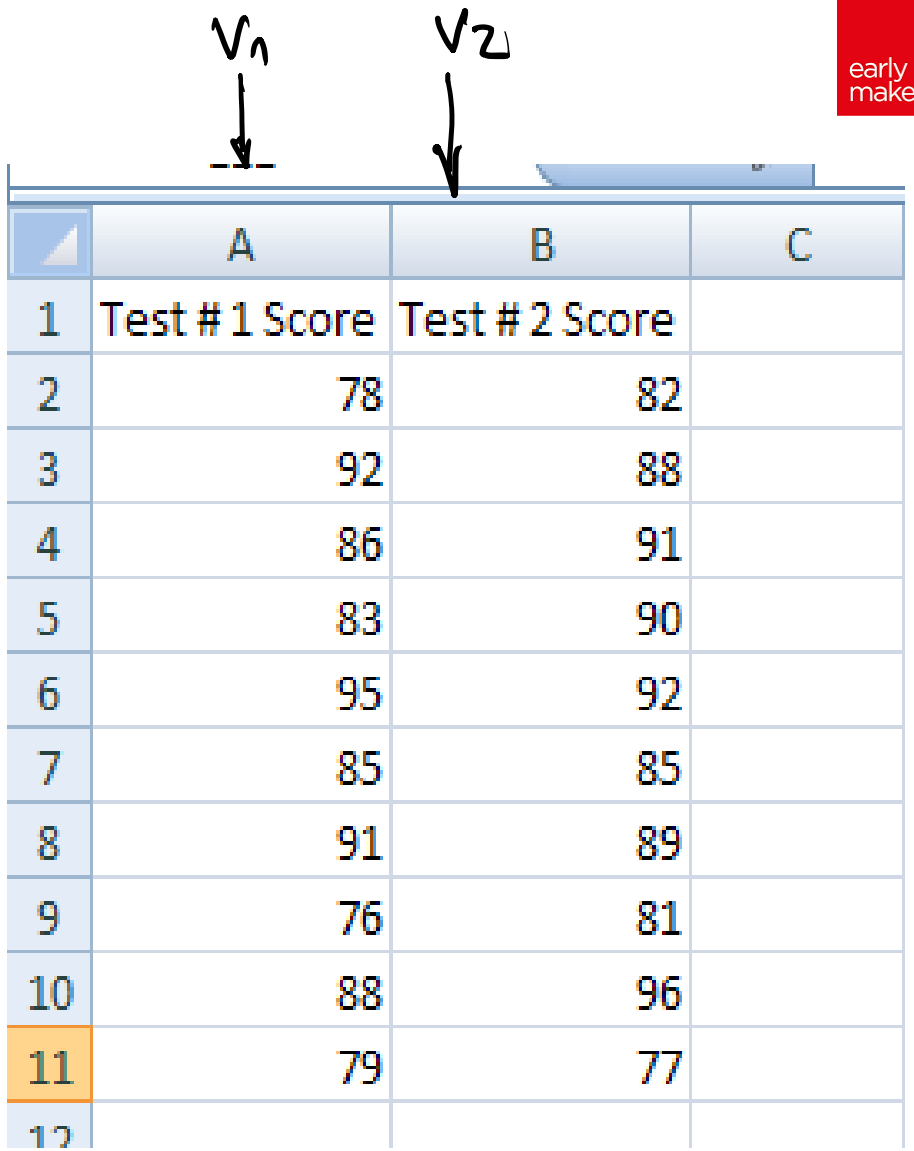
- $H_0: \rho = 0 = cor(X, Y)$
- Estimator of ρ = r
- Statistics :

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$ follows a Student distribution T(n-2)

- We need to compare the actual t values computed from our data to the corresponding critical (threshold) values of Student distribution with n-2 degrees of freedom for a chosen α (0.05) value.
- Rule of thumb : ρ is significant if $|r| \geq \frac{2}{\sqrt{n}}$

Example

- Compute the correlation between these two test scores
  - What is the Null and Alternative Hypotheses for a correlation test ?
  - Choose a significance risk level
  - Choose a statistic and compute its value
  - Check with the rule of thumb
  - State your conclusion



	A	B	C
1	Test # 1 Score	Test # 2 Score	
2	78	82	
3	92	88	
4	86	91	
5	83	90	
6	95	92	
7	85	85	
8	91	89	
9	76	81	
10	88	96	
11	79	77	
12			

Exercise Part 1

- Data set : « Exam Anxiety.dat »
  - Import the data with MS Excel
    - Code : Id Variable
    - Revise : Time spent in revising
    - Exam : Exam performance (%)
    - Anxiety : Exam anxiety level
    - Gender (Sex)
  - Save the data in Excel format (.xlsx)
  - Select the appropriate variables for a correlation study and copy them in a new excel sheet
    - Provide all the relevant information on univariate descriptive statistics
  - Draw the scatter plots of the variables (two by two)

Exercise Part 2

- Using the formulae in the previous slides, compute the covariance between the different variables
    - Derive the correlations from your results
- Check your results, using the embedded Excel functions
- Compute the t value and the F value for each of your correlation
  - Give the p-value of each of your correlations tests
  - Write your conclusions, including effects and significance.

Effect size

- Measure the strength of the relationship between two variables
  - For categorical variables ϕ *or* $Cramer's\ V$
  - For quantitative variables : $|r|$
- The effect size of these indices are given as following (rule of thumb) :
  - Under 0.1 : not significant
  - Between 0.1 and 0.3 : small (if significant)
  - Between 0.3 and 0.5 : moderate (if significant)
  - Between 0.5 and 0.7 : strong
  - Between 0.7 and 0.9 : very strong
  - Beyond 0.9 : colinearity or identity
    - One of the two variables should be removed from the analyses

Linear Regression Modeling

General Principles & Applications

Some review and reminding

- We want to model the observed data (ALL QUANTITATIVE VARIABLES)
  - Y is our outcome variable – the variable we want to predict in our data
  - And $X^1, X^2, \dots, X^p$ are the predictors – the variable to predict Y from our data
  - There are n individuals, each identified with index i
    - Y_i is the observed value of the outcome variable Y for individual unit i
    - X_i^j is the observed value of the predictor variable X^j for individual unit i
    - $\widehat{Y}_i$ is the value predicted by our model for Y_i
    - ε_i is the prediction error (in the population) of our model for individual unit i. It is a random variable following a normal distribution with mean 0 and standard deviation $\sigma \Leftrightarrow \boldsymbol{\varepsilon_i \sim \mathcal{N}(0, \sigma^2)}$
    - e_i is the prediction error (in the sample) of our model pour individual unit i : **it is called the residual**
- Then, we have :
  - $Y_i = \beta_0 + \beta_1 X_i^1 \dots + \beta_p X_i^p + \varepsilon_i$ (population theoretical model)
  - $Y_i = \widehat{Y}_i + e_i$

  with $\widehat{Y}_i = b_0 + b_1 X_i^1 \dots + b_p X_i^p$ (estimated model from sample)
  - In short : Data = Model + Error

Model parameters

- Choosing a $\hat{Y}$ model amounts to finding :
  - the right set of $(p+1)$ parameters $\beta_0, \beta_1, \ldots, \beta_p$
    - Estimated by $b_0, b_1, \ldots, b_p$ from the observed data
  - the right value of σ^2, the variance of the error distribution
    - Estimated by $\hat{\sigma}^2$
- For that, we need to measure the total (agregated) error of the model and find the appropriate parameters such that :
  - The total error is minimized
  - The number of non redundant parameters minimizes the total error
    - For example, when there is no predictor ($p = 0$) we predict all our data with one single value
    - When $p + 1 = n$, we predict our data set with itself – the model is useless
    - When $p = 1$, we have one predictor : a simple regression model
    - $n - (p+1) = n - p - 1$ is the degree of freedom

Math review / complements

X is a matrix with n lines (rows) and p+1 columns

X' is a matrix with p+1 lines (rows) and n columns

$(X'X)^{-1}$ is a square matrix with (p+1) lines and columns

$$X = \begin{bmatrix} 1\ X_1^1 & \cdots X_1^j \cdots & X_1^p \\ \vdots & \vdots & \vdots \\ 1\ X_i^1 & X_i^j & X_i^p \\ \vdots & \vdots & \vdots \\ 1\ X_n^1 & \cdots X_n^j \cdots & X_n^p \end{bmatrix}$$

X is the Matrix of the values the predictor variables.

$$X' = \begin{bmatrix} 1 & \cdots 1 \cdots & 1 \\ \vdots & \vdots & \vdots \\ X_j^1 & X_j^i & X_j^n \\ \vdots & \vdots & \vdots \\ X_p^1 & \cdots X_p^i \cdots & X_p^n \end{bmatrix}$$

X' is the Matrix transpose of X.
The lines of X are the column of X'.
The column of X are the lines of X'.
X' is sometimes noted ${}^tX$ ou X^T

- One can easily compute the matrix product of X and X' with Excel
- Then one can easily compute the inverse of that product matrix

Parameter Estimates

- The estimates of the regression coefficients β_j are given by :

$$\hat{\beta}_j = b_j = [(X'X)^{-1}X'Y]_j$$

  - Y is the (vector) column of the outcome (explained) variable
  - X is the matrix of the predictor variables
  - X' is the transpose of X
  - The estimate $\hat{\beta}_j = b_j$ is the j-th line of the matrix product $[(X'X)^{-1}X'Y]$
    - Y is matrix with n lines and 1 column
    - X' is a matrix with p+1 lines and n columns
    - X is a matrix with n lines and p+1 columns
    - $[(X'X)^{-1}X'Y]$ is a matrix with p+1 lines and 1 column
- The estimate of σ^2 is given by :

$$\hat{\sigma}^2 = \text{MSE} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n - p - 1}$$

Comparing Models

- The worst model (the baseline or reference model) is the one parameter model ($p = 0$)

$$\widehat{Y}_i = b_0$$

- FIRST : We want to test that any model having at least one predictor is better than the one parameter model

$$H_0: p = 0 \iff \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

(H_0: all the regression coefficients are zero)

$$H_1: p \neq 0 \iff \beta_1 \neq 0 \; ou \; \beta_2 \neq 0 \ldots ou \; \beta_p \neq 0$$

(H_1: at least one regression coefficient is NOT zero)

- SECOND : We want to test that each predictor has a meaningful contribution

$$H_0: \beta_j = 0 \iff X^j \text{ IS NOT a meaningful predictor}$$

$$H_1: \beta_j \neq 0 \iff X^j \text{ IS a meaningful predictor}$$

- THIRD : We want to evaluate the overall effect size of our model

Percentage of variance explained by the model : $R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{Y}_i - \bar{Y}_i)^2}{\sum(Y_i - \bar{Y}_i)^2}$

Regression Test Statistics

- FIRST: pertinence of the regression modeling
  - $F = \frac{MSR}{MSE} \sim \mathcal{F}(p, n-p-1)$
  - $MSR = \frac{SSR}{p} = \frac{\sum(\hat{Y}_i - \bar{Y}_i)^2}{p}$
  - $MSE = \frac{SSE}{n-p-1} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-p-1}$
  - $\mathcal{F}(p, n-p-1)$: Fisher distribution with p degrees of freedom in the numerator and n-p-1 degrees of freedom in the denominator
- SECOND: pertinence of predictor number j (X^j)

$$F_j = \frac{b_j^2}{MSE(X'X)_{jj}^{-1}} \sim \mathcal{F}(1, n-p-1) \text{ of Fisher}$$

$$t_j = \frac{b_j}{\sqrt{MSE(X'X)_{jj}^{-1}}} \sim T(n-p-1) \text{ of Student}$$

  - b_j: the estimated coefficient of predictor X^j
  - X : the matrix made of the observed variables
  - $(X'X)_{jj}^{-1}$: the j-th diagonal of the inverse of the matrix product between X and its transpose

Exercise – computing regression coefficients

- We want to predict Exam performance from the time spent in revising and level of exam anxiety
  - What is your outcome (explained) variable?
  - What are your predictor (explaining) variables?
- Upload (open) your <Exam_Anxiety.xlsx> Excel file
  - Copy your variables in a new Excel sheet
    - The predictor variables must be regrouped together (side by side)
    - Complete the following : $n = ?$ $p = ?$
- Generate your X Matrix
  - Add an extra column containing only « 1 » at the left of your predictor variable columns
- Generate your X' Matrix
  - Use <copier + transposer> from Excel
- Generate your X'X Matrix
  - Select the size of your matrix ($p + 1$ lines) x ($p + 1$ $colums$)
  - Use <PRODUITMAT + ctrl + shift (maj) + enter >
    - With X' and X as arguments – in that order
- Generate your $(X'X)^{-1}$ Matrix
  - Select the size of your Matrix
  - Use <INVERSEMAT + ctrl + shift (maj) + enter >
- Compute your X'Y Matrix
- Compute your $[(X'X)^{-1}X'Y]$ Matrix
- Give the value of your estimated coefficients. $\hat{\beta}_j = b_j$

Exercise : testing significance

- Compute the sum of squares
  - Generate columns with
    - $Y_i - \hat{Y}_i$
    - $\hat{Y}_i - \bar{Y}_i$
  - Generate a table :
    - On lines : SSR, SSE and SST
      - Check that SSR+SSE=SST
      - Add $R^2 = \frac{SSR}{SST}$ the determination coefficient and compute R the multiple correlation coefficient
    - On columns : sums of squared, degrees of freedom, means of squares (MSR, MSE, MST)
      - Check that MST = Variance (Y)
      - Add $F = \frac{MSR}{MSE}$ and compute the p-value and the critical value
- Go to Excel tab < données + utilitaire d'analyse + regression linéaire>
  - Check your results
- Conclusion
  - Pertinence of your regression model ?
  - Pertinence of your predictors ?
  - What is your final model ?

Confidence intervals for regression coefficients

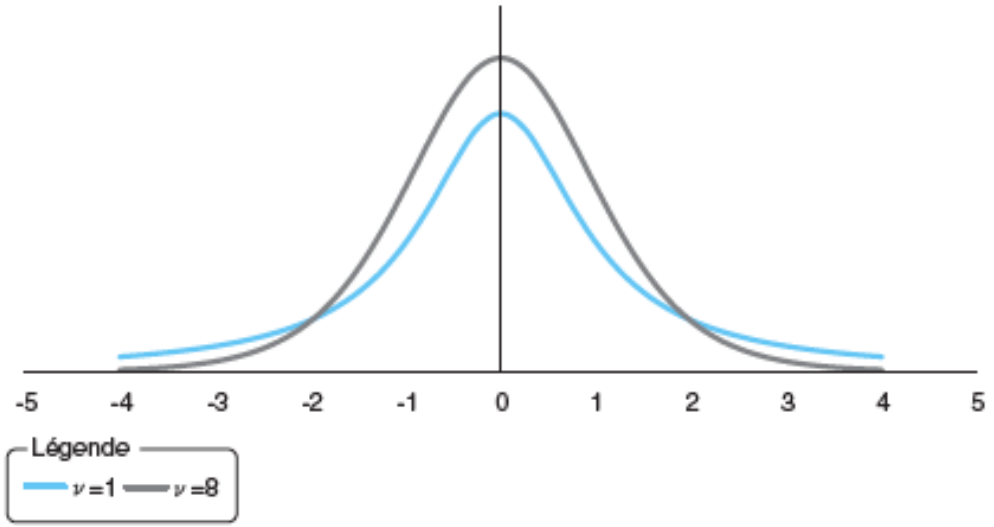
- We know that the test statistic is given by
  - $t_j = \frac{b_j}{\sqrt{MSE(X'X)^{-1}_{jj}}} \sim T(n-p-1)$ of Student

(Two tailed test)

OR

  - $F_j = \frac{b_j^2}{MSE(X'X)^{-1}_{jj}} \sim \mathcal{F}(1, n-p-1)$ of Fisher

(One tailed test)



$$H_0 : \beta_j = b_j^{obs}$$

$$-\tau_{\alpha/2} \leq \frac{b_j - b_j^{obs}}{\sqrt{MSE(X'X)^{-1}_{jj}}} \leq \tau_{\alpha/2}$$

$$\boldsymbol{b_j^{obs} - \tau_{\alpha/2}\sqrt{MSE(X'X)^{-1}_{jj}} \leq b_j \leq b_j^{obs} + \tau_{\alpha/2}\sqrt{MSE(X'X)^{-1}_{jj}}}$$

TWO VARIABLES REGRESSION MODELING

Least square regression line

- An equation can be fit to show the best linear relationship between two variables:

$$Y = \beta_0 + \beta_1 X$$

Where Y is the dependent variable and
X is the independent variable
β_0 is the Y-intercept
β_1 is the slope

- Estimates for coefficients β_0 and β_1 are found using a Least Squares Regression technique
- The least-squares regression line, based on sample data, is

$$\hat{y} = b_0 + b_1 x$$

- Where b_1 is the slope of the line and b_0 is the y-intercept:

$$b_1 = \frac{Cov(x,y)}{s_x^2} = r\left(\frac{s_y}{s_x}\right) \qquad b_0 = \bar{y} - b_1\bar{x}$$

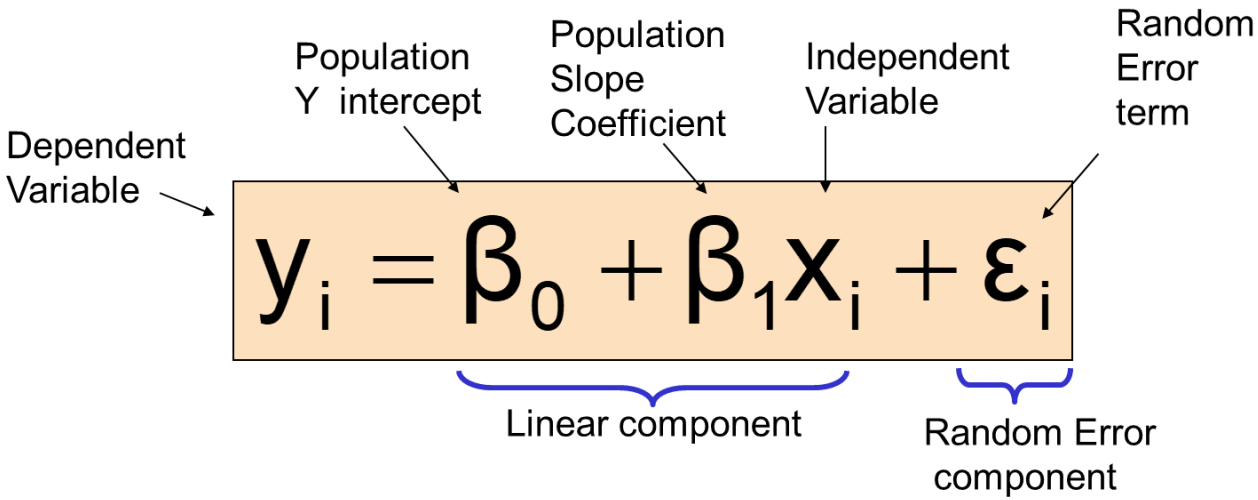
Simple linear regression model

- The relationship between X and Y is described by a linear function
- Changes in Y are assumed to be influenced by changes in X
- Linear regression population equation model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Where β_0 and β_1 are the population model coefficients and ε is a random error term.

The population regression model:



Assumptions and representation

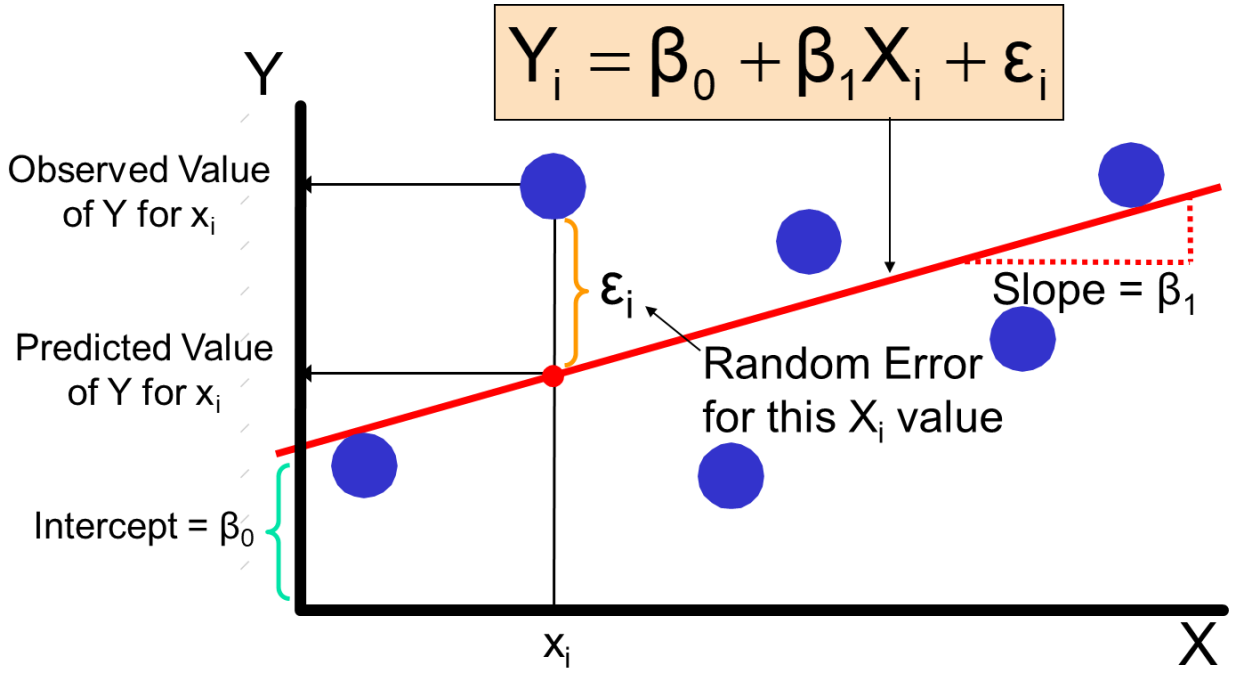
- The true relationship form is linear (Y is a linear function of X, plus random error)
- The error terms, ε_i are independent of the x values
- The error terms are random variables with mean 0 and constant variance, σ^2

  (the uniform variance property is called homoscedasticity)

$$E[\varepsilon_i] = 0 \quad \text{and} \quad E[\varepsilon_i^2] = \sigma^2 \quad \text{for } (i = 1, \ldots, n)$$

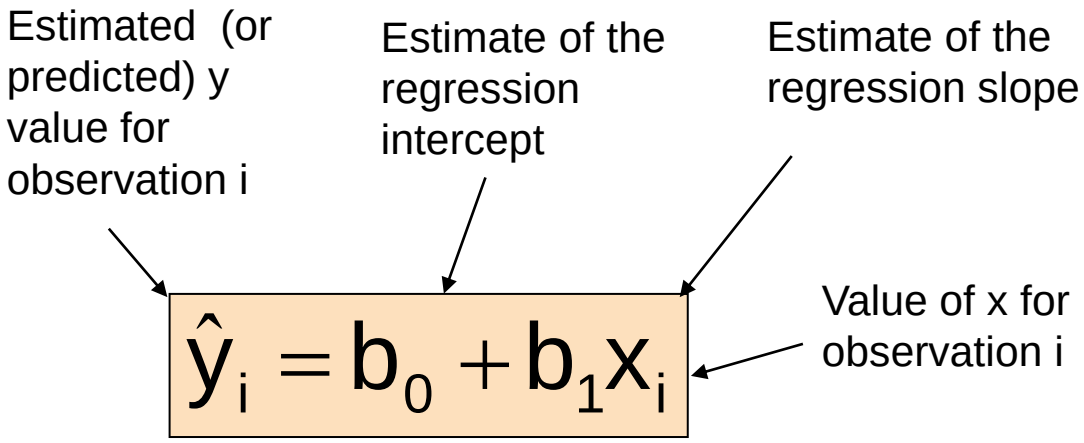
- The random error terms, ε_i, are not correlated with one another, so that

$$E[\varepsilon_i \varepsilon_j] = 0 \quad \text{for all } i \neq j$$



Parameter estimations

The simple linear regression equation provides an estimate of the population regression line



The individual random error terms e_i have a mean of zero

$$e_i = (y_i - \hat{y}_i) = y_i - (b_0 + b_1 x_i)$$

- b_0 and b_1 are obtained by finding the values of b_0 and b_1 that minimize the sum of the squared residuals (errors), SSE:

$$\min \text{ SSE} = \min \sum_{i=1}^{n} e_i^2$$

$$= \min \sum (y_i - \hat{y}_i)^2$$

$$= \min \sum [y_i - (b_0 + b_1 x_i)]^2$$

Differential calculus is used to obtain the coefficient estimators b_0 and b_1 that minimize SSE

Formula and example

- The slope coefficient estimator is

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cov(x,y)}{s_x^2} = r\frac{s_y}{s_x}$$

- And the constant or y-intercept is

$$b_0 = \bar{y} - b_1\bar{x}$$

- The regression line always goes through the mean $\bar{x}$, $\bar{y}$

A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)

House Price in $1000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Analysis of Variance

- Total variation is made up of two parts:

$$SST = SSR + SSE$$

Total Sum of Squares	Regression Sum of Squares	Error (residual) Sum of Squares

$$SST = \sum (y_i - \bar{y})^2$$

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

$$SSE = \sum (y_i - \hat{y}_i)^2$$

where:

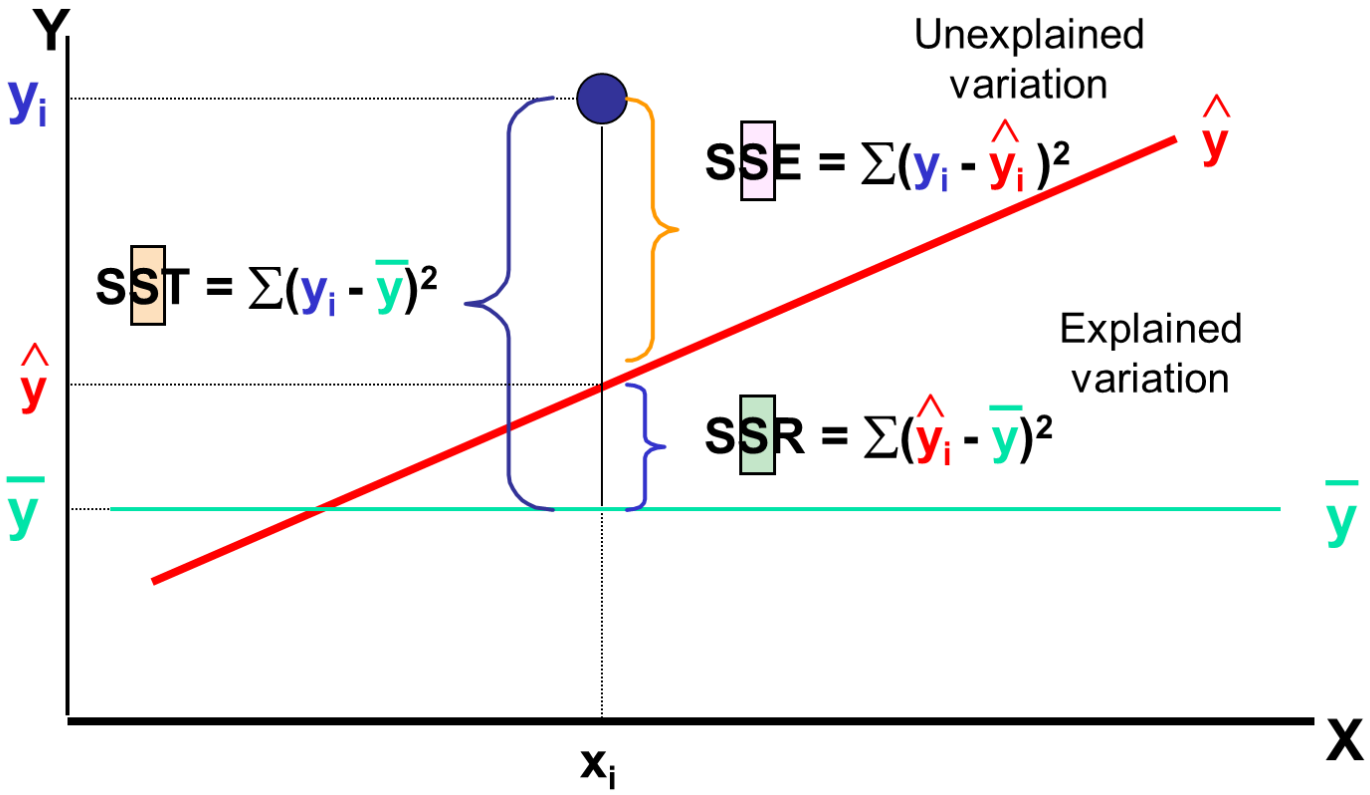
$\bar{y}$ = Average value of the dependent variable

y_i = Observed values of the dependent variable

$\hat{y}_i$ = Predicted value of y for the given x_i value

- SST = total sum of squares
  - Measures the variation of the y_i values around their mean, $\bar{y}$
- SSR = regression sum of squares
  - Explained variation attributable to the linear relationship between x and y
- SSE = error sum of squares
  - Variation attributable to factors other than the linear relationship between x and y

Coefficient of Determination



- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called R-squared and is denoted as R^2

$$R^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

note: $0 \le R^2 \le 1$

Variance of the population error

- An estimator for the variance of the population model error is

$$\hat{\sigma}^2 = s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SSE}{n-2}$$

- Division by n – 2 instead of n – 1 is because the simple regression model uses two estimated parameters, b_0 and b_1, instead of one

$s_e = \sqrt{s_e^2}$ is called the standard error of the estimate

Regression Statisticsa	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$s_e = 41.33032$

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Comparing group means

Independent samples T-test

Conditions

- We are interested in comparing quantitative variable X between two groups
  - Group #1
    - Sample size n_1
    - mean μ_1 estimated with $\bar{X}_1 = m_1$
    - variance $(\sigma_1)^2$estimated with $(s_1)^2$
  - Group #2
    - Sample size n_2
    - mean μ_2 estimated with $\bar{X}_2 = m_2$
    - variance $(\sigma_2)^2$estimated with $(s_2)^2$
- We assume
  - $(\sigma_1)^2 \backsimeq (\sigma_2)^2$(homogeneity of variance)
  - Scores in each group are independent (coming from different individuals)
  - The variable $D = \mu_1 - \mu_2$ follows an approximately normal distribution through multiple samplings

Statistics and Test

- $H_0: D = \mu_1 - \mu_2 = D_0$

$$t = \frac{\bar{D} - D_0}{s_D \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim T(n_1 + n_2 - 2) \quad \text{ou} \quad \frac{(\bar{D} - D_0)^2}{\frac{{s_D}^2}{n_1} + \frac{{s_D}^2}{n_2}} \sim \mathcal{F}(1, n_1 + n_2 - 2)$$

$$s_D = \sqrt{\frac{(n_1 - 1){s_1}^2 + (n_2 - 1){s_2}^2}{n_1 + n_2 - 2}} \quad \text{ou} \quad {s_D}^2 = \frac{(n_1 - 1){s_1}^2 + (n_2 - 1){s_2}^2}{n_1 + n_2 - 2}$$

- In the case of comparing means (independent t test), $D_0 = 0$

$$F = \frac{(\bar{X}_1 - \bar{X}_2)^2}{\frac{{s_D}^2}{n_1} + \frac{{s_D}^2}{n_2}} \sim \mathcal{F}(1, n_1 + n_2 - 2) \Leftrightarrow F = t^2 \text{ where } t \sim T(n_1 + n_2 - 2)$$

Exercise : comparing male and female students

- Back to <Exam_Anxiety.xlsx> data
  - Sort males and females into two groups
  - Compute the means for each group
    - $\bar{X}_1 = m_1$ and $\bar{X}_2 = m_2$
  - Compute the variance for each group
    - $(s_1)^2$ and $(s_2)^2$
  - Compute the pooled variance
    - $s_D{}^2 = \frac{(n_1-1)s_1{}^2+(n_2-1)s_2{}^2}{n_1+n_2-2}$
  - Compute the F and T values
  - Give the p-value
- Write your conclusions !

- Go to Excel tab < données + utilitaire d'analyse + test d'égalité des espérances >
  - Check your results

You have now completed « Statistics Applied to Management » basic course.

CONGRATULATIONS !