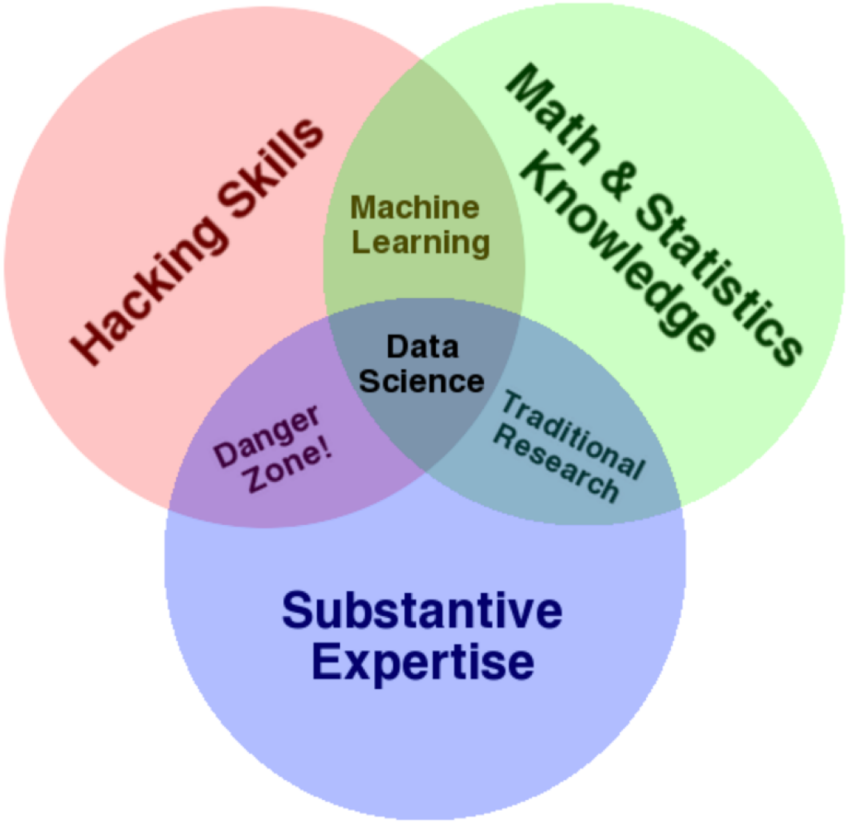
DATA ANALYTICS FUNDAMENTALS

DSA8001

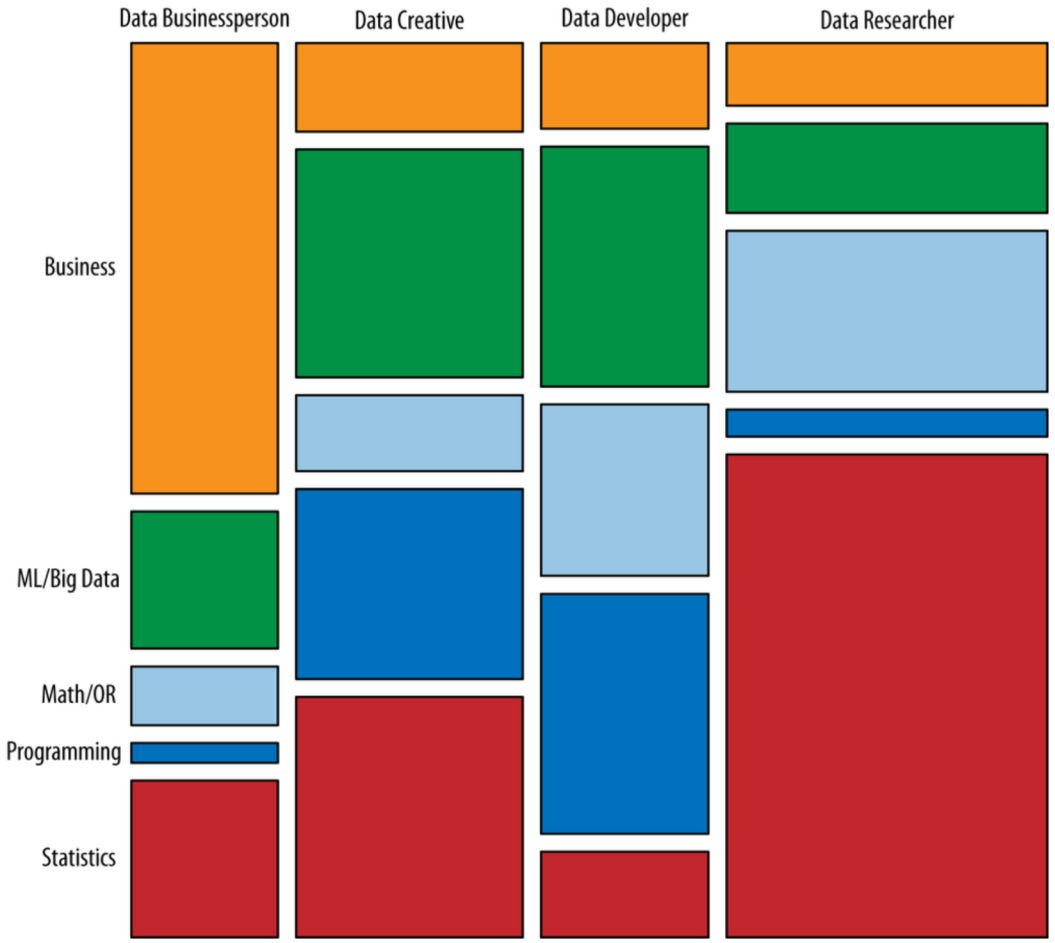
WHAT IS DATA SCIENCE?

- Data Science is an exciting discipline that allows you to turn raw data into understanding, insight and knowledge



Venn diagram of data science ([1], Figure 1-1)

A DATA SCIENCE PROFILE



An illustration of subfields of data science ([1], Figure 1-4)

KEY TERMS OF DATA ANALYTICS – VARIABLES

- **Variable** is a quantity, quality or property that you can measure, the values of which may vary from measurement to measurement.
  - *Synonyms: table column, field, attribute, property, feature, vector, dimension, etc.*
- There are two basic types of variable:
  - **Numeric variables** – variables whose values are are recorded as numbers (integer or real values)

    Example:

    wind_speed(mph) = {27.345, 31.528, 40.209, 50.123, …}

    age(years) = {15, 27, 28, 30, …}
  - **Categorical variables** – variables whose values are recorded as symbols.

    Example:

    gender = {Male, Female}

    country = {UK, USA, Australia, New Zealand, …}

KEY TERMS OF DATA ANALYTICS – VARIABLES

- **Numeric variables:**
  - **Discrete Variables (*synonyms: integer, count*)** – numeric variables that may take on only certain (distinct) numeric values (usually obtained by counting)

    Example: Number of persons attending a conference (you cannot have half a person)
  - **Continuous Variables (*synonyms: float, double, interval, numeric*)** – numeric variables that may take any real value in some interval

    Example: rainfall(mm) = {23, 15.2345, 3.567, …}
- **Categorical variables:**
  - **Nominal** – categorical variables whose values **CANNOT** be naturally ranked

    Example: Gender (Male, Female), Eyes Colour (Black, Brown, Green, …)
  - **Ordinal** – categorical variables whose values **CAN BE** naturally ranked

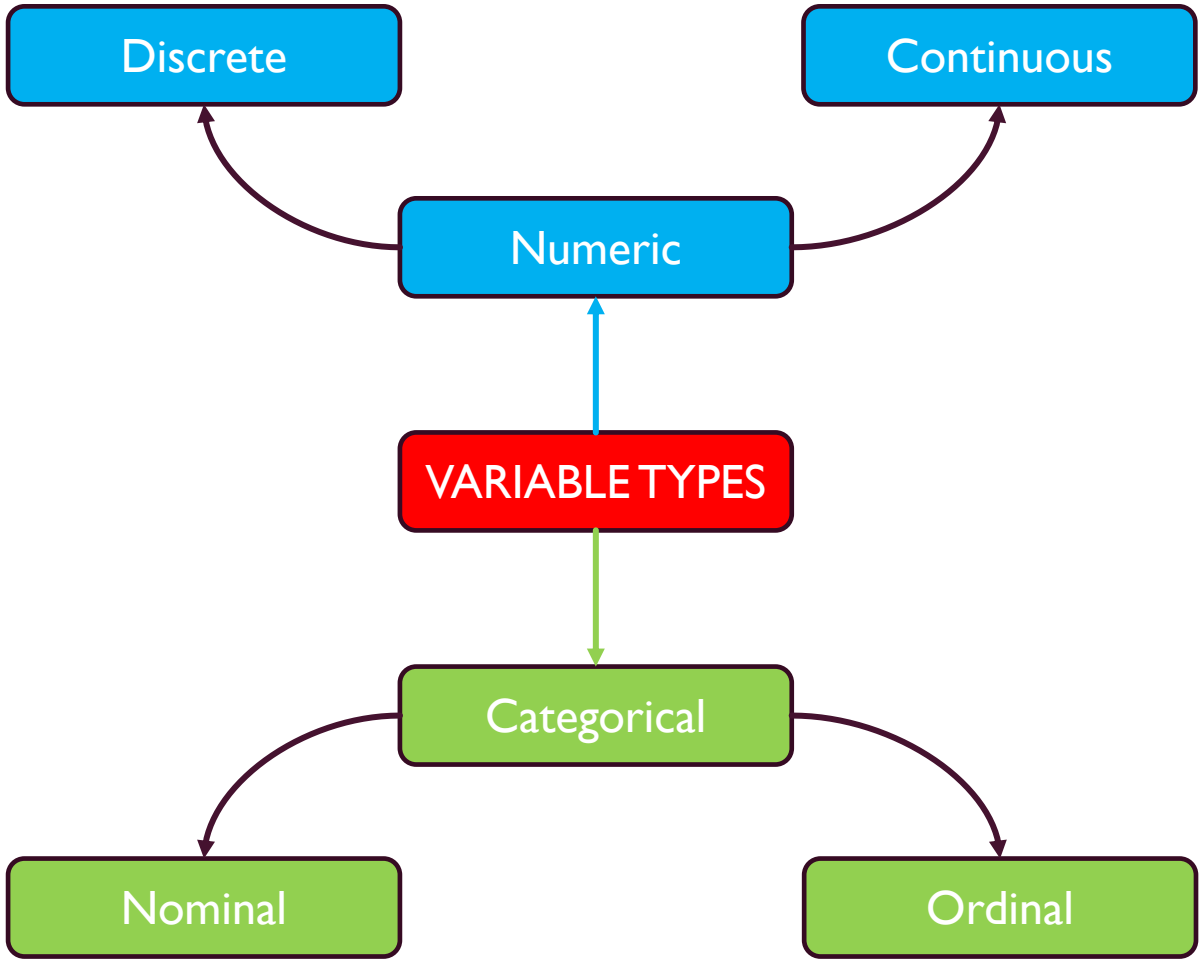
    Example: Driving Speed (low, medium, high; low < medium < high), Education (BSc, MSc, PhD; BSc < MSc < PhD)

NOTE: *Sometimes categorical variables can be presented as finite set of discrete numeric variables*

Example: gender = {0, 1} where 0 = Male, 1 = Female

daytime = {1, 2, 3, 4} where 1 = Early morning, 2 = Working hours, 3 = Evening, 4 = Late night

KEY TERMS OF DATA ANALYTICS – VARIABLES (SUMMARY)



KEY TERMS OF DATA ANALYTICS – UNIVARIATE AND MULTIVARIATE DATASETS

- **Univariate dataset** – dataset consisted of measurements that correspond to the single variable
  - Example: Dataset consisted of the values that correspond to the height of all students in the class
- **Multivariate dataset** - data set consisted of measurements that correspond to two or more variables
  - Most relevant when the individual components aren't as useful when considered on their own (in other words, as univariate quantities) in any given statistical analysis
  - Example: Spatial coordinates (a horizontal *x*-coordinate and a vertical *y*-coordinate)
- **Univariate data analysis** – the analysis performed on single variable
- **Multivariate data analysis** – the simultaneous analysis of two or more variables
- **Observations** are measurements made under similar conditions
  - **In univariate datasets a single observation consists of a single measurement (value, symbol, data point)**
  - **In multivariate datasets a single observation contains several measurements, each associated with a different variable**

AN EXAMPLE OF UNIVARIATE AND MULTIVARIATE DATASETS (TABULAR)

UNIVARIATE DATASET

COLOR
BLACK
WHITE
RED

OBSERVATIONS

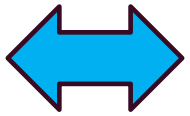
MULTIVARIATE DATASET

COLOR	HEX
BLACK	#000000
WHITE	#FFFFFF
RED	#FF0000

AN EXAMPLE OF UNIVARIATE AND MULTIVARIATE DATASETS (JSON)

UNIVARIATE DATASET

COLOR
BLACK
WHITE
RED



```
{
  color: [black,  white,  red]
}
```

OBSERVATIONS

```
{
  color:  [black,  white,  red],
  hex:    [#000000, #FFFFFF, #FF0000]
}
```

COLOR	HEX
BLACK	#000000
WHITE	#FFFFFF
RED	#FF0000

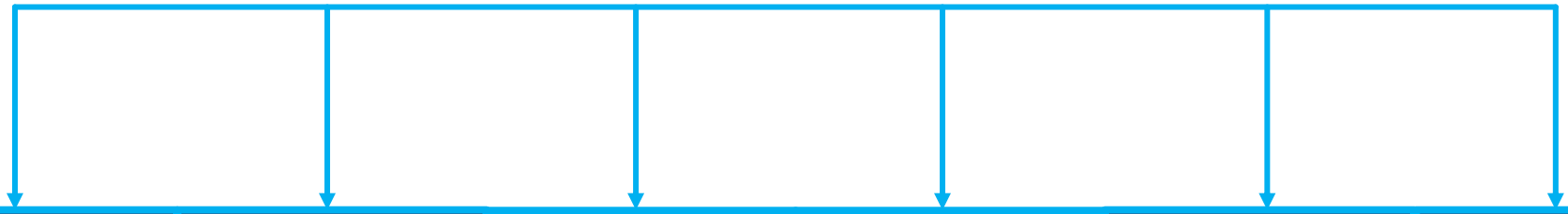
MULTIVARIATE DATASET

KEY TERMS OF DATA ANALYTICS – TABULAR DATASETS

- **Tabular dataset** is a set of values, each associated with a variable and an observation. Following rules apply:
  - ***Variables are table columns***
  - ***Observations are table rows***
  - Tabular data is ***tidy*** if each value is placed in its own "cell", each variable in its own column, and each observation in its own row.
  - ***The size*** of the data set is defined by the number of observations (rows) in the table
  - ***The dimensionality*** of the dataset is defined by the number of variables (columns) in the table

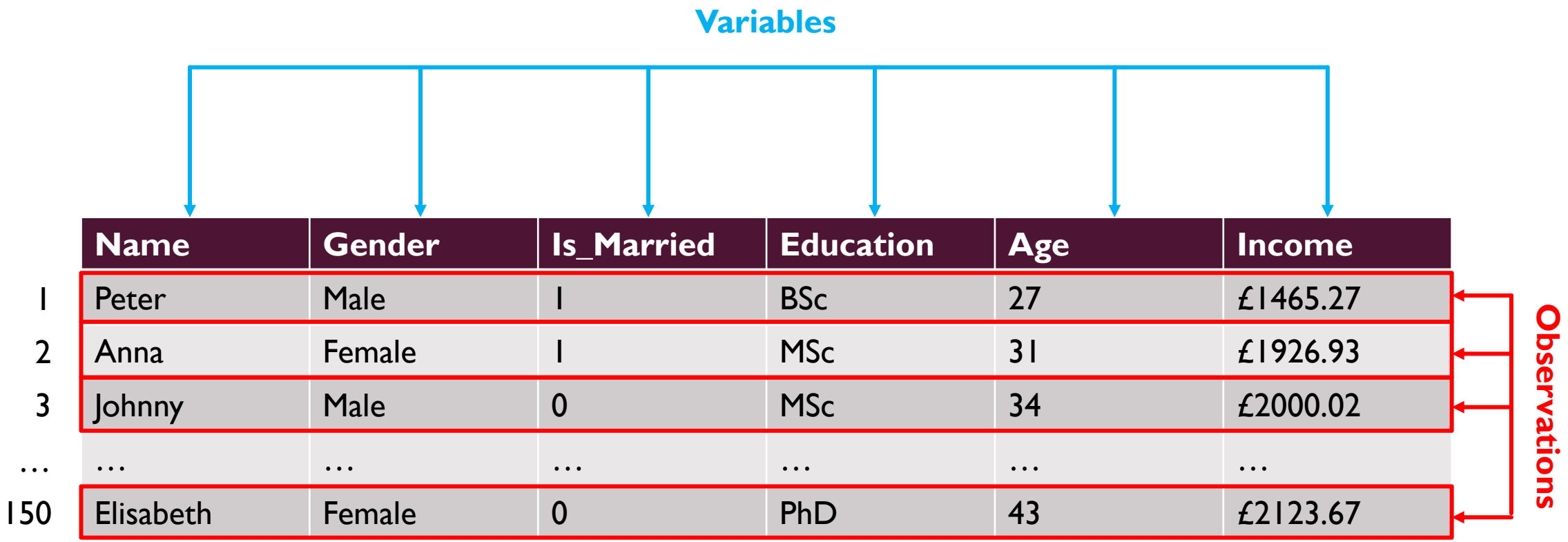
AN EXAMPLE OF TABULAR DATASET

Variables



	Name	Gender	Is_Married	Education	Age (years)	Income
1	Peter	Male	1	BSc	27	£1465.27
2	Anna	Female	1	MSc	31	£1926.93
3	Johnny	Male	0	MSc	34	£2000.02
…	…	…	…	…	…	…
150	Elisabeth	Female	0	PhD	43	£2123.67

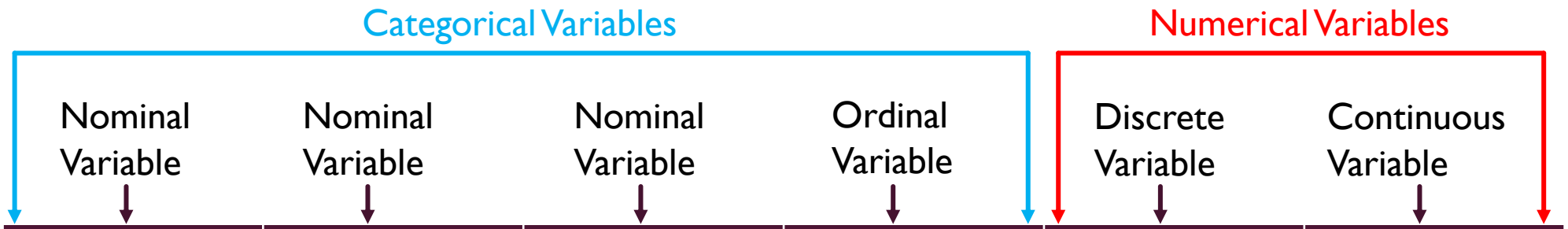
AN EXAMPLE OF TABULAR DATASET



	Name	Gender	Is_Married	Education	Age	Income
1	Peter	Male	1	BSc	27	£1465.27
2	Anna	Female	1	MSc	31	£1926.93
3	Johnny	Male	0	MSc	34	£2000.02
…	…	…	…	…	…	…
150	Elisabeth	Female	0	PhD	43	£2123.67

Dataset Size: 150
Dimensionality: 6

AN EXAMPLE OF TABULAR DATASET



	Name	Gender	Is_Married	Education	Age	Income
1	Peter	Male	1	BSc	27	£1465.27
2	Anna	Female	1	MSc	31	£1926.93
3	Johnny	Male	0	MSc	34	£2000.02
…	…	…	…	…	…	…
150	Elisabeth	Female	0	PhD	43	£2123.67

KEY TERMS OF DATA ANALYTICS – POPULATION AND SAMPLING

- **Population** is the (usually) large pool of observational units that we are interested in.
- **Sample** is a smaller collection of observational units that is selected from the population.
- **Sampling** refers to the process of selecting observations from a population. Four most commonly used sampling strategies are:
  - Simple random sampling
  - Stratified sampling
  - Cluster sampling
  - Multistage sampling

KEY TERMS OF DATA ANALYTICS – POPULATION AND SAMPLING

- **Representative sample** - A sample is said be a *representative sample* if the characteristics of observational units selected are a good approximation of the characteristics from the original population.
- **Bias** – *Bias* corresponds to a favoring of one group in a population over another group.
- **Generalizability -** *Generalizability* refers to the largest group in which it makes sense to make inferences about from the sample collected. This is directly related to how the sample was selected.
- **Parameter** is a calculation based on one or more variables measured in the population. Parameters are almost always denoted symbolically using Greek letters such as $\mu, \sigma, \rho, \beta, \ldots$
- **Statistic** is a calculation based on one or more variables measured in the sample. Statistics are usually denoted by lower case Arabic letters with other symbols added sometimes. These include $\bar{x}, s, p, , b$ ...

SAMPLING STRATEGIES – SIMPLE RANDOM SAMPLING

Simple random sampling is a sampling strategy where the individuals are selected from the list of units in the population, by means of some random process (e.g. random number tables or pseudo-random number generators), in such way that each individual has equal chance to be selected.

- Selection can be preformed sequentially, i.e. individuals can be selected from population one at the time without replacement so that at each stage each of the remaining individuals in the population has the same probability of being selected

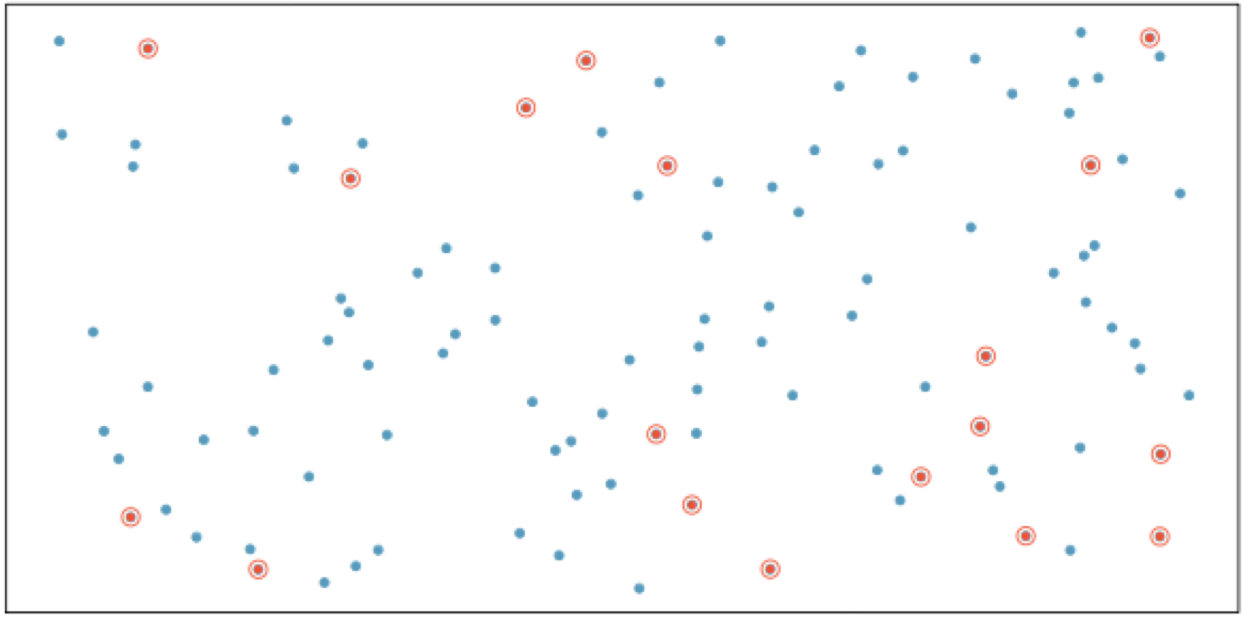


*Illustration of the **simple random sampling** where the 18 individuals (red dots) are being selected from the list of units in the population by means of some random process*

SAMPLING STRATEGIES – STRATIFIED SAMPLING

- ***Stratified sampling*** is a divide-and-conquer sampling strategy where the population is divided in groups called **strata**, and the sample of individuals is then drawn from each stratum using some other random sampling process, usually simple random sampling.
- In general, the strata are chosen so that units in each stratum are as alike as possible and units in different strata are as different as possible.
- Stratification has two purposes:
  1. To increase the accuracy and precision of the overall population estimates, and
  2. To ensure that domains of study are adequately represented
- This sampling strategy is used in cases when it is known that the population is heterogeneous with respect to one or more variables which may have a bearing on the factor being studied.

SAMPLING STRATEGIES – STRATIFIED SAMPLING (EXAMPLE)

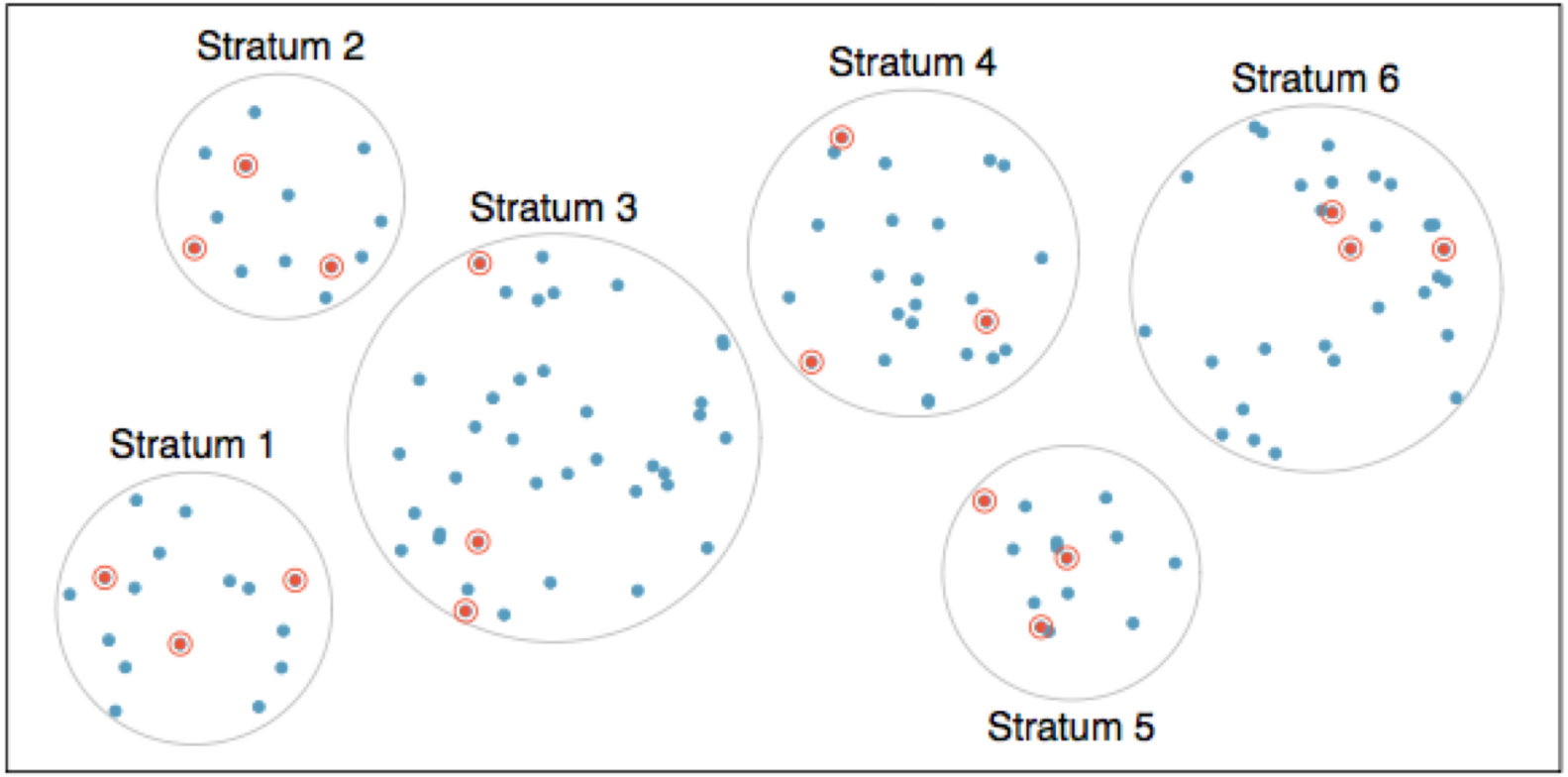


*Illustration of the **stratified sampling** where all the individuals from the populations were grouped into strata (6 stratums) and then the simple random sampling was employed within each stratum to select 3 individuals per stratum (18 individuals in total)*

SAMPLING STRATEGIES – CLUSTER AND MULTISTAGE SAMPLING

- ***Cluster sampling*** is a sampling strategy where the population is divided into many groups, called **clusters**, and then we sample a fixed number of clusters and include all observations from each of those clusters in the sample.
- ***Multistage sampling*** is a sampling process similar to the cluster sampling, but rather then keeping all observations in each cluster, we collect a random sample within each selected cluster.
- Sometimes cluster or multistage sampling can be more economical than the alternative sampling techniques.
- Unlike stratified sampling, cluster and multistage sampling strategies are most helpful when there is a lot of case-to-case variability within a cluster but the clusters themselves don't look very different from one another.
  - ***Example:*** *if neighbourhoods represented clusters, then cluster or multistage sampling work best when the neighbourhoods are very diverse*
- A downside of these methods is that more advanced analysis techniques are typically required

SAMPLING STRATEGIES – CLUSTER SAMPLING (EXAMPLE)

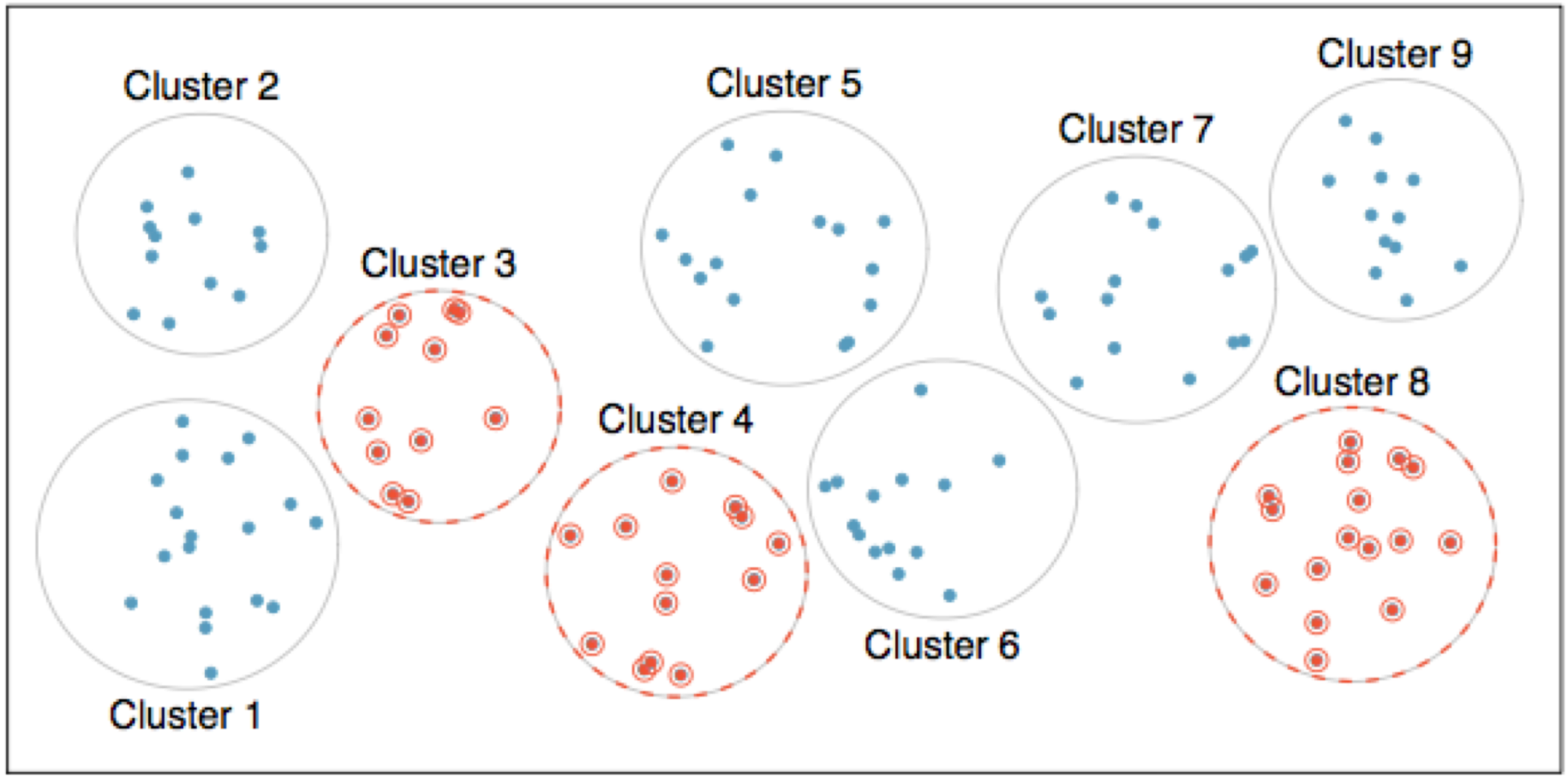


*Illustration of the **cluster sampling** where all the individuals from the populations were binned into 9 clusters, three of these clusters were sampled and all observations within these three cluster were included in the sample*

SAMPLING STRATEGIES – MULTISTAGE SAMPLING (EXAMPLE)

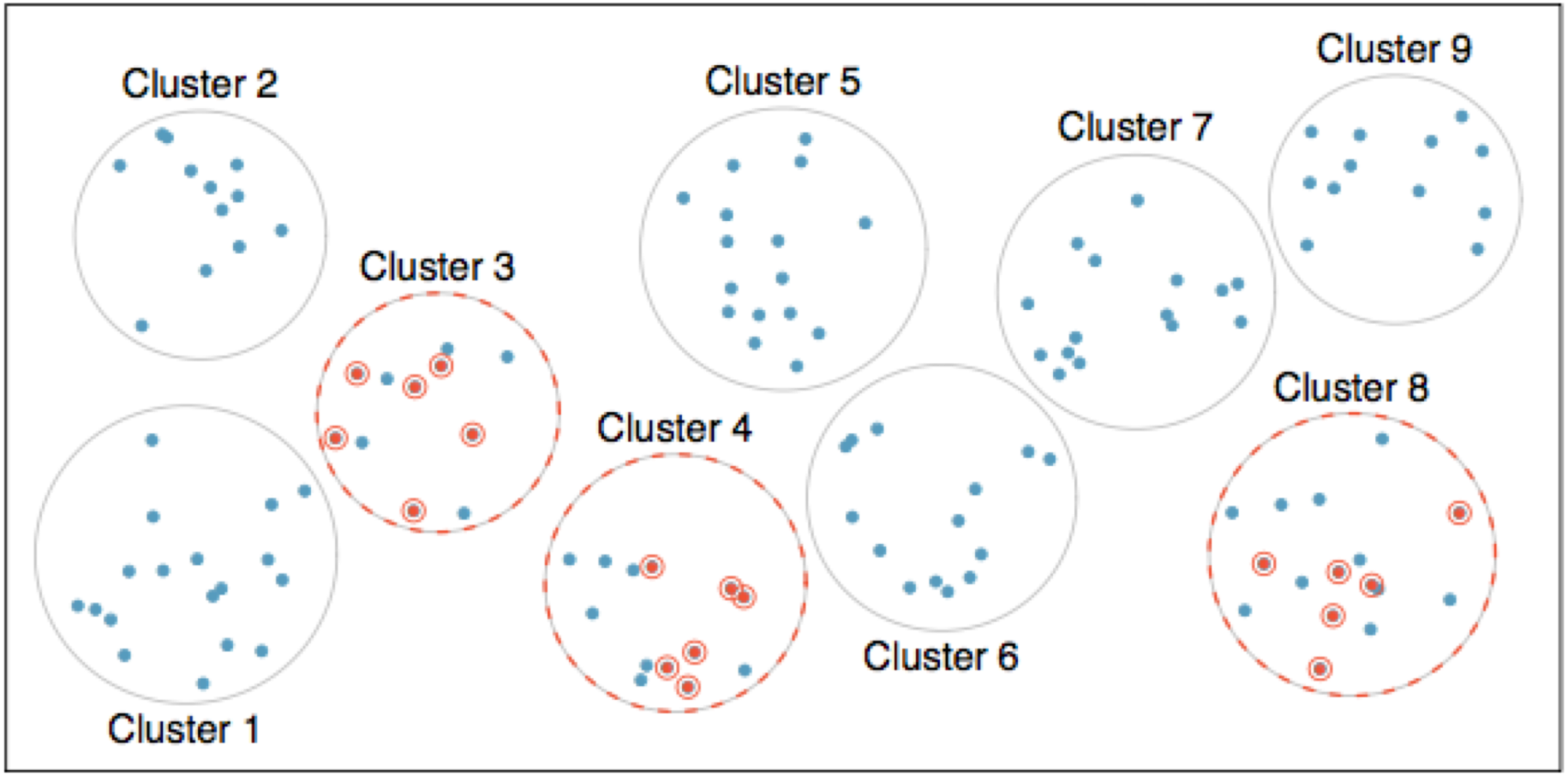


*Illustration of the **multistage sampling** where all the individuals from the populations were binned into 9 clusters, three of these clusters were sampled and randomly selected subset of each of these three clusters were included in the sample*

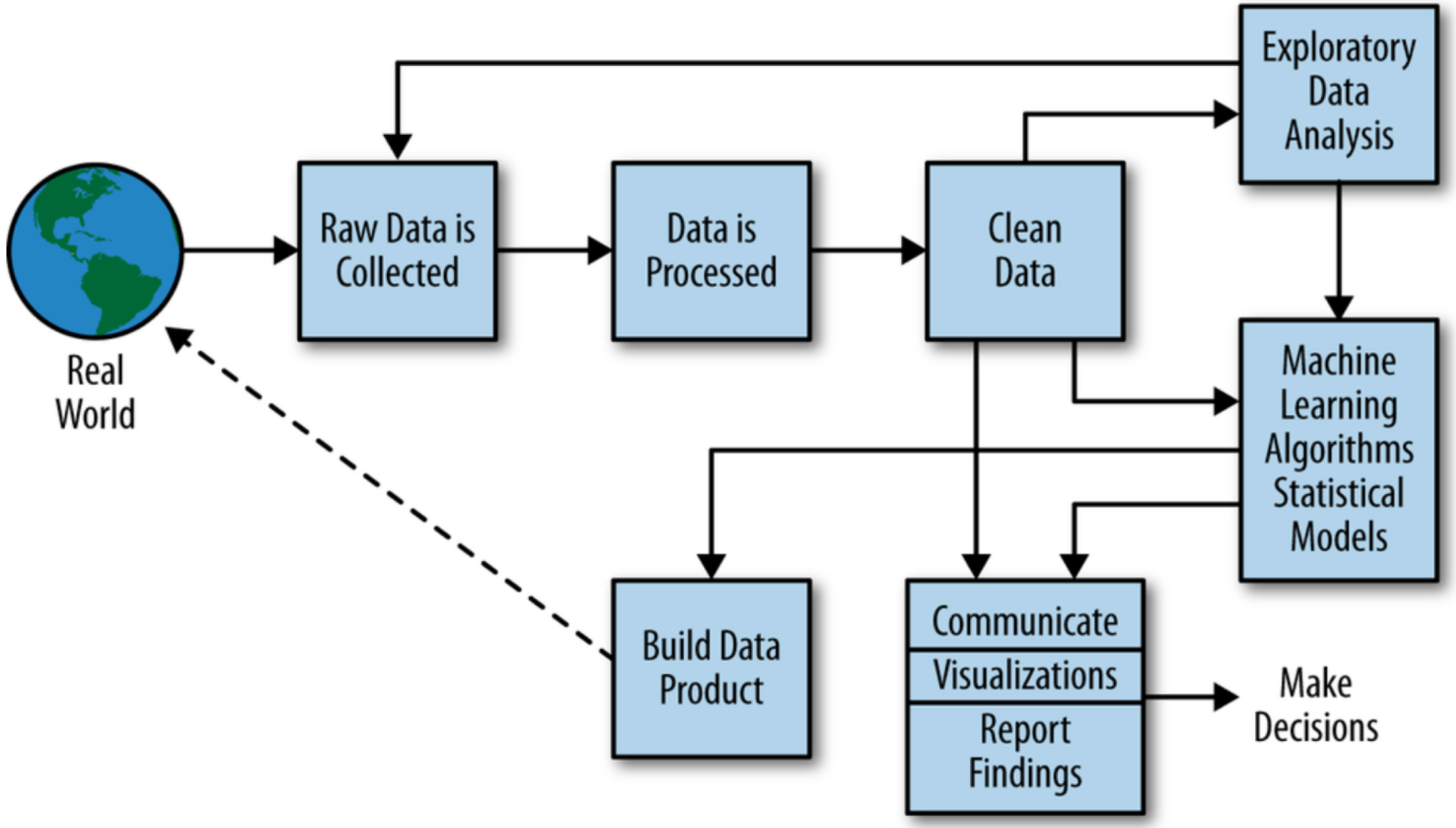
SAMPLING STRATEGIES - EXAMPLE

- Suppose we are interested in estimating the malaria rate in a densely tropical portion of rural Indonesia. We learn that there are 30 villages in that part of the Indonesian jungle, each more or less similar to the next. Our goal is to test 150 individuals for malaria. What sampling method should be employed?

SOLUTION

- ***Simple random sampling***
  - A simple random sample would likely draw individuals from all 30 villages, which could make data collection extremely expensive.
- ***Stratified sampling***
  - Stratified sampling would be a challenge since it is unclear how we would build strata of similar individuals.
- ***Cluster and Multistage sampling***
  - Cluster sampling or multistage sampling seem like very good ideas.
  - If we decided to use multistage sampling, we might randomly select half of the villages, then randomly select 10 people from each. This would probably reduce our data collection costs substantially in comparison to a simple random sample, and this approach would still give us reliable information.

DATA SCIENCE PROCESS



The Data Science Process ([1], Figure 2-2)

EXPLORATORY DATA ANALYSIS

- Exploratory Data Analytics (EDA) is creative process of exploring data sets for patterns and relationships.

EDA goals:

1. Develop understanding about data by formulating questions
2. Search for answers using visualisation techniques and summary statistics
    1. Visualise distributions of all variables (using boxplots and histograms)
    2. Visualise time series of data
    3. Investigate all pairwise relationships between variables using scatterplots
    4. Perform data cleaning and variable transformation
    5. Perform summary statistics (mean, median, the lower and upper quartiles, find minimum and maximum values, identify missing data, errors and outliers)
3. Use answers obtained from previous step to refine starting questions and/or generate new questions

GUIDELINES ABOUT FORMULATING QUESTIONS DURING EDA PHASE

- Use questions as tools to guide your investigation
- Asking question will focus your attention on specific part of your dataset and helps you decide which EDA techniques to use
- It is difficult to ask a revealing question at the start of your analysis as you do not know what insights are hidden in your dataset
  - There is no universal set of rules about which question you should ask to guide your research
- Useful starting questions for making discoveries about data:
  - *What type of variation occurs within my variables?*
  - *What is the relationship between variables?*

HOW TO ANSWER: *WHAT TYPE OF VARIATION OCCURS WITHIN MY VARIABLES?*

- **NUMERIC VARIABLES:**

  Summary Statistics

  - Measures of centrality (Mean, Median, Mode)
  - Measures of variability (Variance, Standard deviation, Range, Quantiles, Five number summary)

  Visualization Techniques

  - Histograms
  - Boxplots

NOTE:* *Summary Statistics *are statistics used to quantitatively describe a collection of measurements by summarizing them in a form of a single number*

HOW TO ANSWER: *WHAT TYPE OF VARIATION OCCURS WITHIN MY VARIABLES?*

- **CATEGORICAL VARIABLES:**

  Summary Statistics

  - Counts
  - Percentages
  - Proportions

  Visualization Techniques

  - Bar charts

HOW TO ANSWER: *WHAT IS RELATIONSHIP BETWEEN VARIABLES?*

- **SUMMARY STATISTICS**
  - Covariance and Correlation (Numeric – Numeric)
  - Contingency tables (Categorical – Categorical)
- **VISUALIZATION TECHNIQUES**
  - Scatterplots (Numeric – Numeric)
  - Paired Boxplots (Numeric - Categorical)
  - Paired Histograms (Numeric – Categorical)
  - Mosaic plots (Categorical - Categorical)

MEASURES OF CENTRALITY – MEAN, MEDIAN, MODE

- Measures of centrality are commonly used to investigate characteristics of numeric variable, usually by figuring out the location of:
  - ***the typical measurement*** of a collection of measurements ***(Mean)***,
  - ***the middle measurement*** of a collection of measurements ***(Median)***, or
  - ***the most common measurement*** of a collection of measurements ***(Mode)***.

MEASURES OF CENTRALITY – MEAN

- ***Synonyms: Arithmetic Mean, Average***
- The mean is considered to be the central (typical) measurement of a collection of observations,
- For a collection of numerical values $x = \{x_1, x_2, \dots, x_n\}$ the mean is calculated using the following formula:

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

NOTE: *The Mean is very sensitive to the presence of outliers (e.g. one or more extremely large or small values)*

MEASURES OF CENTRALITY – MEAN (EXAMPLE)

What is the average weekly rainfall?

DAY	RAINFALL (cm)
Monday	0
Tuesday	10
Wednesday	0
Thursday	0
Friday	1
Saturday	1
Sunday	2

MEASURES OF CENTRALITY – MEDIAN

- The median is considered to be the middle measurement of a collection of observations
- If we have a collection of the observations, we find the median by sorting all the observations in ascending order and pick up the middle value (if the number of observations is odd) or finding the mean of two middle values (if there is an even number of observations)
- For a collection of unordered numerical values $\boldsymbol{x} = \{x_1, x_2, \dots, x_n\}$ the median is calculated using the following approach:
  1. Sort values $\{x_1, x_2, \dots, x_n\}$ in ascending order (from smallest to largest)
  2. Find the median using the following formula:

$$m_x = \begin{cases} x_i, & if\ \boldsymbol{n}\ is\ odd \\ \dfrac{x_j + x_k}{2}, & if\ \boldsymbol{n}\ is\ even \end{cases}$$

where $i = \frac{n+1}{2}, j = \frac{n}{2}, k = \frac{n+2}{2}$

NOTE: *The median is very resistant to outliers*

MEASURES OF CENTRALITY – MEAN (EXAMPLE 1)

What is the median value of weekly rainfall?

DAY	RAINFALL (cm)
Monday	0
Tuesday	10
Wednesday	0
Thursday	0
Friday	1
Saturday	1
Sunday	2

MEASURES OF CENTRALITY – MEAN (EXAMPLE 2)

What is the median value of weekly rainfall?

DAY	RAINFALL (cm)
Monday	0
Tuesday	10
Wednesday	0
Thursday	0
Friday	NA
Saturday	1
Sunday	2

MEASURES OF CENTRALITY – MODE

- **The mode *(Synonym: modal value)*** is the observation that occurs most frequently in the dataset
- ***HINT**: As the median, the mode is very resistant to outliers*

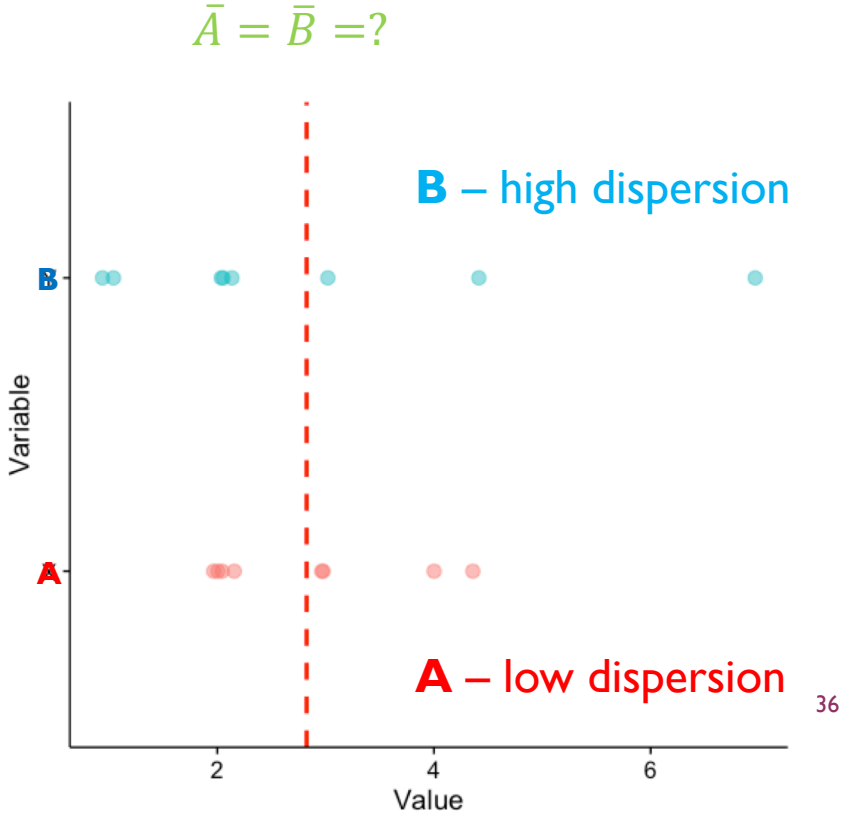
DAY	RAINFALL (cm)
Monday	0
Tuesday	10
Wednesday	0
Thursday	0
Friday	1
Saturday	1
Sunday	2

MEASURES OF VARIABILITY

- Measures of variability describe how dispersed observations in the univariate dataset are, i.e. they describe whether the observations are tightly clustered or spread out.

A	B
2	1
4.4	4.4
3	1
3	3
2	2
2.2	2.2
2	2
4	7

Variables **A** and **B**
have same mean
but different dispersion



MEASURES OF VARIABILITY – VARIANCE, STANDARD DEVIATION

- **Variance** – Average squared distance of each observation from the mean. For a given set of numerical variables $\boldsymbol{x = \{x_1, x_2, \dots, x_n\}}$, variance $(\boldsymbol{s_x^2})$ is calculated using following formula:

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

- **Standard Deviation** – The square root of the variance, and has the same scale as the observations in the numerical variable.

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

HINT 1: *The standard deviation is useful when considering how close the observations are to the mean. Focus on the conceptual meaning of the standard deviation as a descriptor of variability rather than the formulae.*

HINT 2: *Usually 68% of the data will be within one standard deviation of the mean and about 95% will be within two standard deviations.*

MEASURES OF VARIABILITY – VARIANCE, STANDARD DEVIATION

$\bar{A} = \bar{B} = ?$

A	B
2	1
4.4	4.4
3	1
3	3
2	2
2.2	2.2
2	2
4	7

$s_A^2 = ?$

$s_A = \sqrt{s_A^2} = ?$

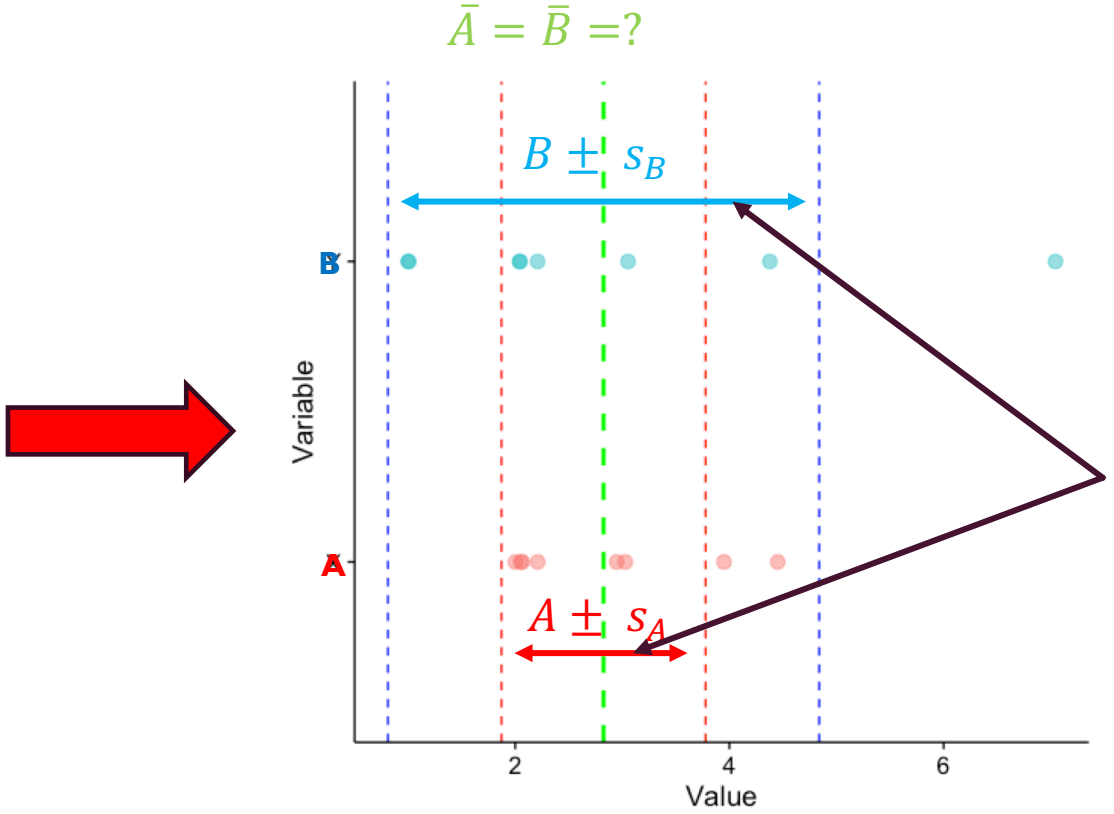
$s_B^2 = ?$

$s_B = \sqrt{s_B^2} = ?$

MEASURES OF VARIABILITY – VARIANCE, STANDARD DEVIATION

A	B
2	1
4.4	4.4
3	1
3	3
2	2
2.2	2.2
2	2
4	7

$s_A = ?$ $s_B = ?$



Usually 68% of observations lie within one st. dev. from the mean!

MEASURES OF VARIABILITY – ORDER STATISTICS

- **Order Statistics** – Statistics based on sorted (ranked) data
- **Quantile –** The value computed from a sorted collection of numerical measurements (ascending order) and indicates an observation's rank when compared to all the other present observations (it can take values between 0 and 1):
  - **Example:** Median is 0.5^{th} quantile (it gives you a value below which half of the measurements lie)
- **Percentiles –** identical to quantiles but expressed in "percent scale" of 0 to 100
  - P^{th} quantile is identical to 100x P^{th} percentile
  - ***P^{th} percentile is the cutoff point that indicates that at least P percent of the observation in the dataset take on this value or less***
  - **Example:** 80^{th} percentile is the cutoff point which indicates that 80% of observations in the dataset may be found below that point
  - **Example:** Median is 50^{th} percentile (i.e 0.5x100 = 50)
- **Quartiles** are three cut off point that divide the dataset into four equal groups.
  - ***First Quartile (Q_1 = 0.25th quantile = 25th percentile)*** *is defined as the middle value between smallest observation and median*
  - ***Second Quartile (Q_2 = 0.5th quantile = 50th percentile)*** *is the median of the dataset (splits dataset in half)*
  - ***Third Quartile (Q_3 = 0.75th quantile = 75th percentile)*** *is the middle value between the median and the highest observation in the dataset*

MEASURES OF VARIABILITY – ORDER STATISTICS

- **Range** – Difference between the smallest and the largest observations in a numerical variable.
  - ***HINT:*** *Extremely sensitive to outliers and therefore not very useful as a general measure of dispersion in the data*
- Five number summary – provides basic information about variability in the dataset and it is comprised of:
  - ***0th percentile (minimum)***
  - ***25th percentile (Q_1)***
  - ***50th percentile (Q_2)***
  - ***75th percentile (Q_3)***
  - ***100th percentile (maximum)***
- **Interquartile Range (IQR)** – Measures the width of the "middle 50 percent" of the data, that is, the range of the values that lie between 0.25th and 0.75th quantile.
  - HINT: IQR is very resistant to outliers

MEASURES OF VARIABILITY – ORDER STATISTICS

x
9
5
2
3
7
8
1
4
6

x_{sorted}
1
2
3
4
5
6
7
8
9

Five Number Summary

Min (0^{th} perc)

Q_1 (25^{th} perc)

Q_2 (median, 50^{th} perc)

Q_3 (75^{th} perc)

Max (100^{th} perc)

$IQR = Q_3 - Q_1 = A = \frac{7+8}{2} - \frac{3+4}{2} = 5$

$Range = 9 - 1 = 8$

RELATIONSHIPS BETWEEN NUMERICAL VARIABLES - COVARIANCE

- When analyzing data, it's often useful to be able to investigate the *relationship* between two numerical variables to assess trends
  - ***Example:*** *We might expect height and weight of people to have positive relationship (taller people tend to weigh more). However, we cannot say the same for relationship between hand span and the length of the hair.*
- Suppose for *n* individuals you have a sample of observations for two variables, labeled $\boldsymbol{x} = \{x_1, x_2, \dots, x_n\}$ and $\boldsymbol{y} = \{y_1, y_2, \dots, y_n\}$, where x_i corresponds to y_i where $i = 1, \dots, n$. Covariance r_{xy} is calculated using the following formula (where $\bar{x}$ and $\bar{y}$ represent respective sample means of both sets of observations):

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})$$

1. If $r_{xy} = 0 \Longrightarrow$ there is ***NO*** linear relationship between numerical variables $\boldsymbol{x}$ and $\boldsymbol{y}$
2. If $r_{xy} > 0 \Longrightarrow$ *there is a* ***POSITIVE*** *linear relationship between numerical variables x and y (as x increases, y increases, and vice versa)*
3. If $r_{xy} < 0 \Longrightarrow$ there is a **NEGATIVE** linear relationship between numerical variables x and y *(as x increases, y decreases, and vice versa)*

COVARIANCE - EXAMPLE

$\bar{A} = \bar{B} = ?$

A	B
2	1
4.4	4.4
3	1
3	3
2	2
2.2	2.2
2	2
4	7

$s_A = ?$ $s_B = ?$

$r_{AB} = ?$

HOW STRONG IS THIS RELATIONSHIP?

RELATIONSHIPS BETWEEN NUMERICAL VARIABLES – CORRELATION

- The problem with covariance is that by relying solely on covariance we cannot quantify strength of linear relationship between two variables (because there are no upper or lower limits which covariance coefficient can take)
- Correlation helps us to overcome this problem i.e. to interpret covariance by identifying both the direction and the strength of any association.
- There are different types of correlation coefficient's but the most common of these is ***Pearson's product-moment correlation coefficient*** (ρ_{xy}), which is calculated using the following formula (where s_x and s_y represent respective sample standard deviations of both variables):

$$\rho_{xy} = \frac{r_{xy}}{s_x s_y}; -1 \leq \rho_{xy} \leq 1$$

  - *If* $\rho_{xy} = 1 \Longrightarrow$*there is a* ***PERFECT POSITIVE*** *linear relationship between variables* x *and* y
  - *If* $0 < \rho_{xy} < 1 \Longrightarrow$*there is a* ***POSITIVE*** *linear relationship between variables* x *and* y *(closer to* 1 *stronger it is)*
  - *If* $\rho_{xy} = -1 \Longrightarrow$*there is a* ***PERFECT NEGATIVE*** *linear relationship between variables* x *and* y
  - *If* $-1 < \rho_{xy} < 0 \Longrightarrow$*there is a* ***NEGATIVE*** *linear relationship between variables* x *and* y *(closer to* -1 *stronger it is)*
  - *If* $\rho_{xy} = 0 \Longrightarrow$*there is* ***NO*** *linear relationship between variables* x *and* y

RELATIONSHIPS BETWEEN NUMERICAL VARIABLES – CORRELATION

- **Guidelines for interpreting the absolute strength of Pearson's product moment correlation coefficient (ρ_{xy})[10]:**
  - $|\rho_{xy}| = 0.0$ "**NO**" linear relationship
  - $0.0 < |\rho_{xy}| \leq 0.19$ "**very weak**" (positive or negative) linear relationship
  - $0.20 \leq |\rho_{xy}| \leq 0.39$ "**weak**" (positive or negative) linear relationship
  - $0.40 \leq |\rho_{xy}| \leq 0.59$ "**moderate**" (positive or negative) linear relationship
  - $0.60 \leq |\rho_{xy}| \leq 0.79$ "**strong**" (positive or negative) linear relationship
  - $0.80 \leq |\rho_{xy}| < 1.0$ "**very strong**" (positive or negative) linear relationship
  - $|\rho_{xy}| = 1.0$ "**PERFECT**" (positive or negative) linear relationship

- ***Example:***
  - $\rho_{xy} = -0.54$ *There is a "**moderate negative linear relationship**" between variables x and y*
  - $\rho_{xy} = 0.34$ *There is a "**weak positive linear relationship**" between variables x and y*

CORRELATION - EXAMPLE

$\bar{A} = \bar{B} = ?$

A	B
2	1
4.4	4.4
3	1
3	3
2	2
2.2	2.2
2	2
4	7

$s_A = ?$ $s_B = ?$

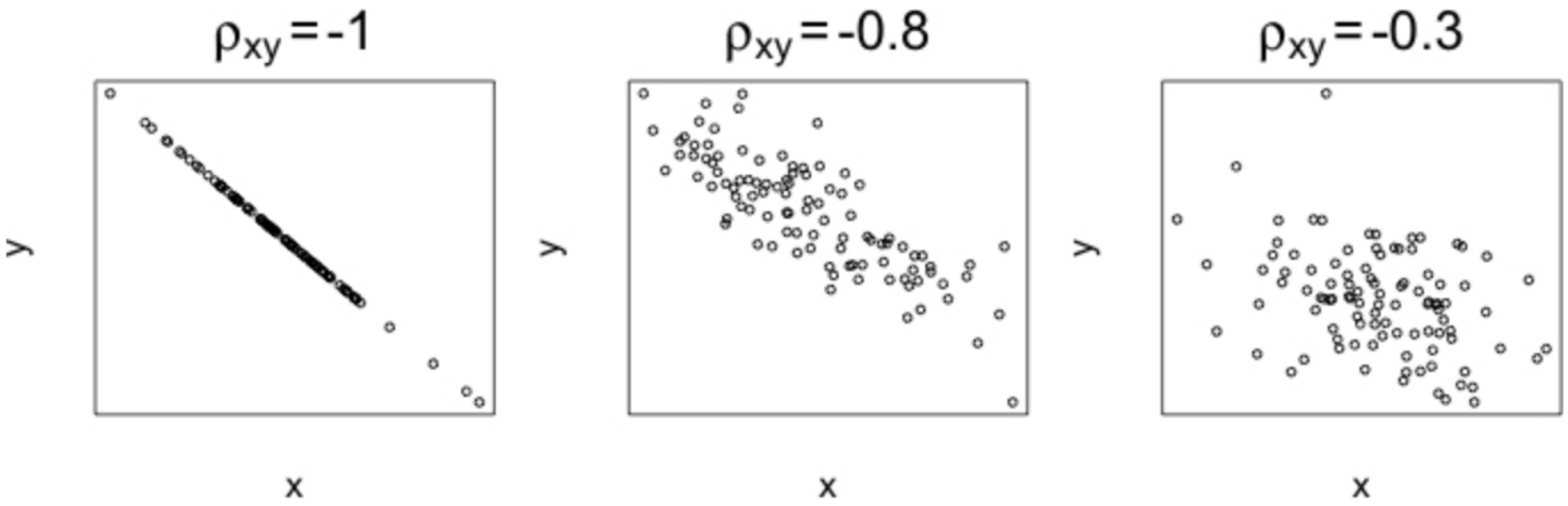
$\rho_{AB} = ?$

$$\rho_{AB} = \frac{r_{AB}}{s_A s_B} = ?$$

CORRELATION – AN EXAMPLE OF NEGATIVE LINEAR RELATIONSHIP

PERFECT NEGATIVE
LINEAR RELATIONSHIP

WEAK NEGATIVE
LINEAR RELATIONSHIP

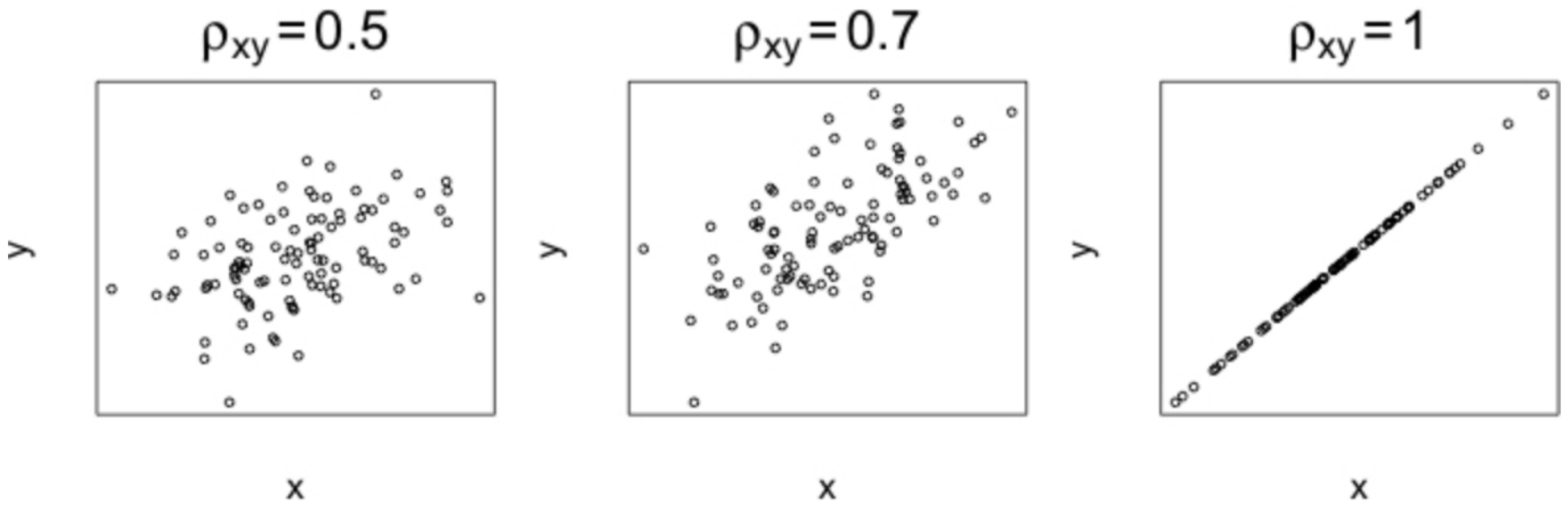


VERY STRONG NEGATIVE
LINEAR RELATIONSHIP

CORRELATION – AN EXAMPLE OF POSITIVE LINEAR RELATIONSHIP

MODERATE POSITIVE
LINEAR RELATIONSHIP

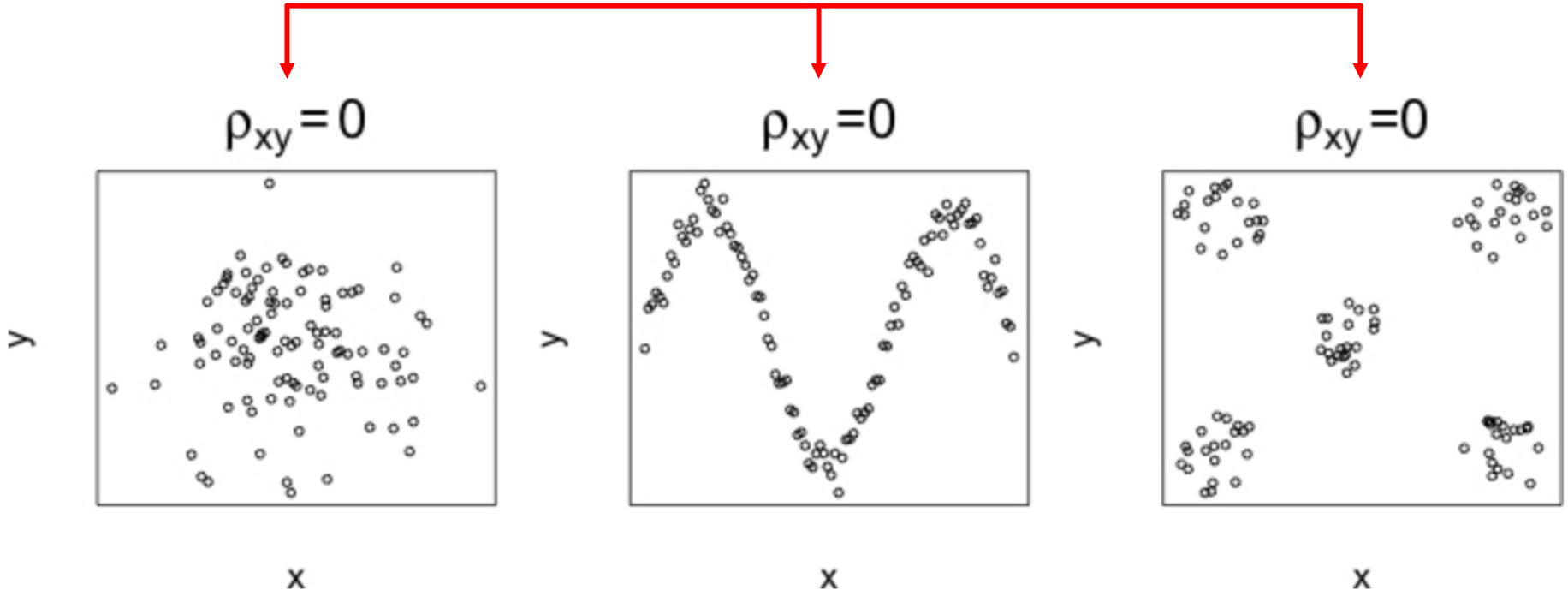
PERFECT POSITIVE
LINEAR RELATIONSHIP



STRONG POSITIVE
LINEAR RELATIONSHIP

CORRELATION – AN EXAMPLE OF NO LINEAR RELATIONSHIP

THERE IS NO LINEAR RELATIONSHIP BETWEEN VARIABLES X & Y



SUMMARY STATISTICS FOR CATEGORICAL DATA – FREQUENCY TABLES

- Frequency Tables – Statistical technique that we use when we want to get more insight into properties of categorical variables
- Frequency tables consist of three columns:
  1. **Frequency Column (F)** – number of occurrences of each symbol that is stored inside the categorical variable
  2. **Relative Frequency (RF)** – proportion of occurrences of each symbol inside the categorical variable.
     - ***HINT:*** *for any frequency table the sum of all relative frequencies when written as proportions must be equal to 1*
  3. **Percentages (P)** – Proportions multiplied by 100
     - ***HINT:*** *for any frequency table the sum of all percentages must be equal to 100*

FREQUENCY TABLES - EXAMPLE

During one day 25 people donated blood. The blood type of each donor is given in the table below:

0	0	A	0	0
A	A	B	A	0
0	0	0	AB	0
A	0	A	0	A
B	AB	B	A	0

Blood Type	F	RF	P(%)
0	12	12/25=0.48	0.48x100=48
A	?	?	?
B	?	?	?
AB	?	?	?
Σ	25	1	100

What is the rarest blood type?

What is the is the most common blood type?

What proportion of donor were type 0?

RELATIONSHIP BETWEEN CATEGORICAL VARIABLES – CONTINGENCY TABLES

- Contingency table – a table that summarizes data for two categorical variables (table of counts by category)
  - ***HINT:*** *Each value in the table represents the number of times a particular combination of variable outcomes occurred*

	spam	num_char	line_breaks	format	number
1	no	21,705	551	html	small
2	no	7,011	183	html	big
3	yes	631	28	text	none
	⋮	⋮	⋮	⋮	⋮
3921	no	15,829	242	html	small

What proportion of spam emails contains text without numbers?

Solution: ?

CONTINGENCY TABLE

NUMBER

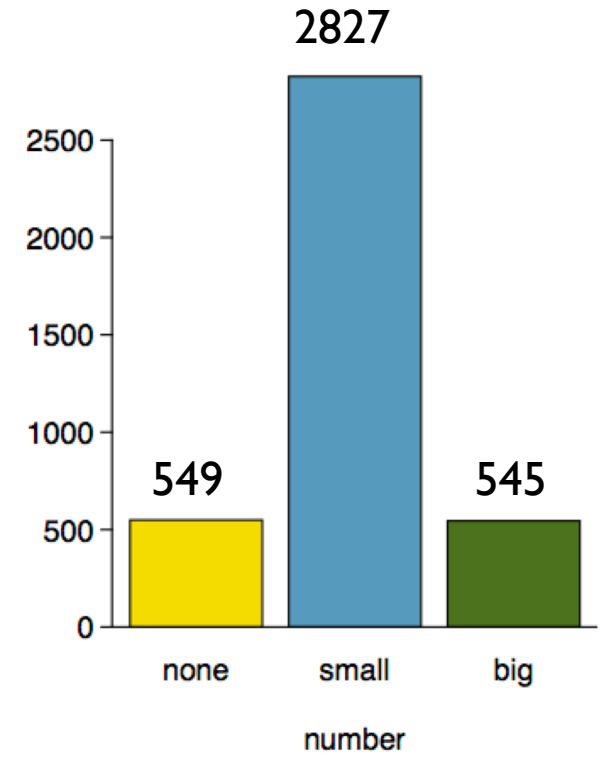
IS_SPAM	none	small	big	Σ
spam	149	168	50	367
not_spam	400	2659	495	3554
Σ	549	2827	545	3921

EXPLORING THE DATA DISTRIBUTION – BAR CHARTS

- **Bar charts** are common visual tools for exploring single categorical variables
  - **X-axis** represent different symbols (categories) of a categorical variable
  - **Y-axis** represents frequency or proportion of occurrence of each category

NUMBER

IS_SPAM	none	small	big	Σ
spam	149	168	50	367
not_spam	400	2659	495	3554
Σ	549	2827	545	3921



*Bar chart of **number** column for the **email** dataset*

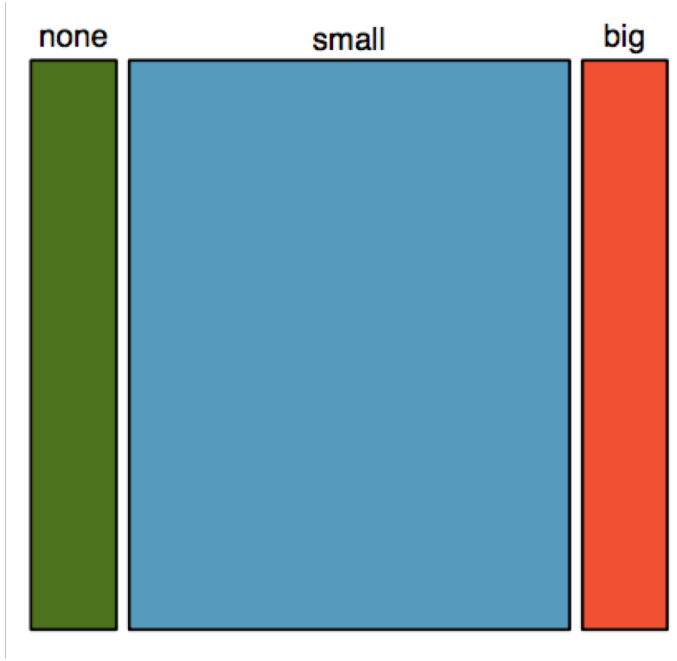
EXPLORING THE DATA DISTRIBUTION – MOSAIC PLOTS

- **Mosaic Plot** is a graphical representation of contingency table information and it is similar to a bar plot.
  - It can be used to visualize one or two categorical variables from contingency table.
- ***HINT:*** *Mosaic plots use box areas to represent to represent the number of observations that box represents*

NUMBER

IS_SPAM

	none	small	big	Σ
spam	149	168	50	367
not_spam	400	2659	495	3554
Σ	549	2827	545	3921



*The one-variable mosaic plot for **number** column*

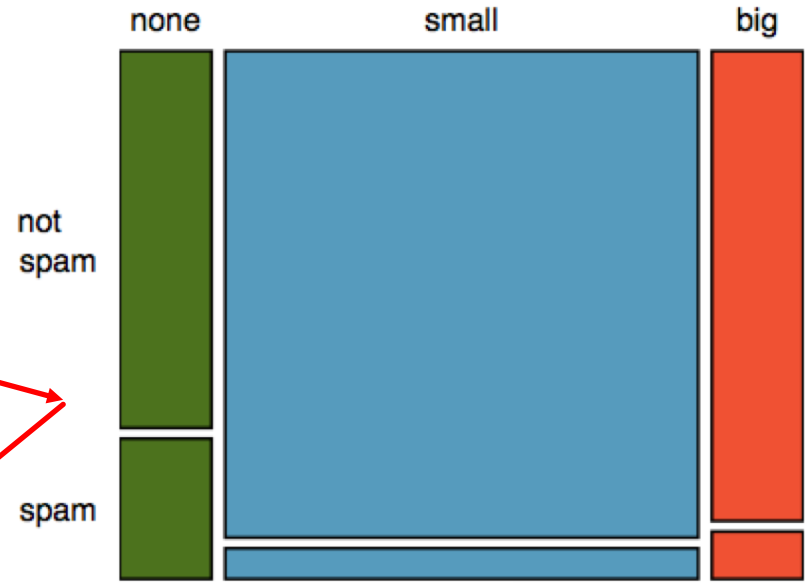
EXPLORING THE DATA DISTRIBUTION – MOSAIC PLOTS

- **Mosaic Plot** is a graphical representation of contingency table information and it is similar to a bar plot.
  - It can be used to visualize one or two categorical variables from contingency table.
- ***HINT:*** *Mosaic plots use box areas to represent to represent the number of observations that box represents*

NUMBER

IS_SPAM

	none	small	big	Σ
spam	149	168	50	367
not_spam	400	2659	495	3554
Σ	549	2827	545	3921



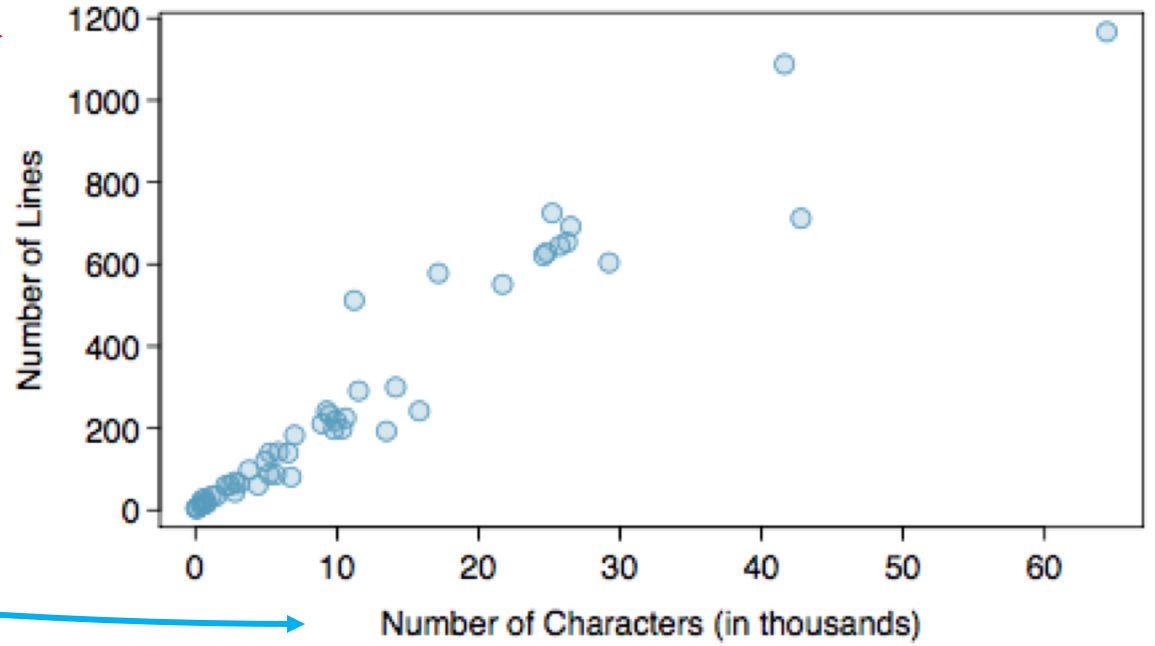
*The two-variable mosaic plot for **number** & **spam** columns*

HINT: *Each category of **number** column is split proportionally according to the fraction of emails that were **spam***

EXPLORING THE DATA DISTRIBUTION – SCATTERPLOTS

- Scatterplots are plots that provide a case-by-case view of data for two numerical variables
- ***HINT:*** *Scatterplots are helpful in* ***quickly spotting associations between two numerical variables****, whether those associations come in the form of simple trends or whether those relationships are more complex*

spam	num_char	line_breaks	format	number
no	21,705	551	html	small
no	7,011	183	html	big
yes	631	28	text	none
⋮	⋮	⋮	⋮	⋮
no	15,829	242	html	small

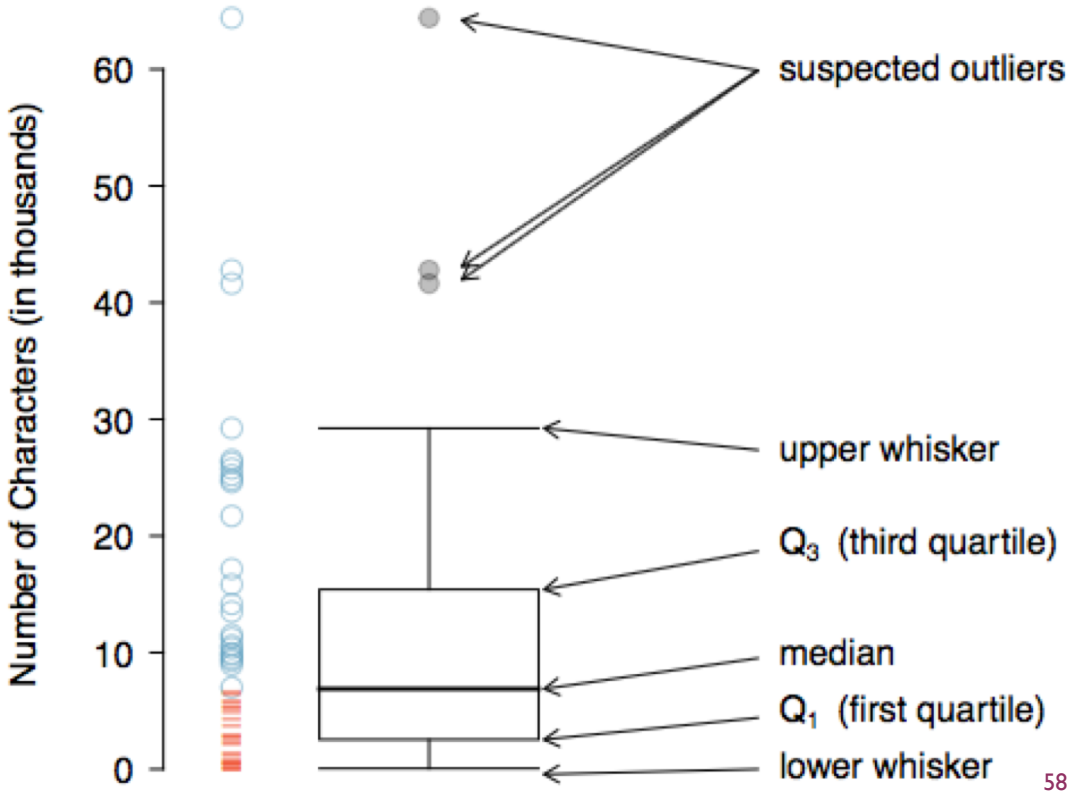


Q: *Do* ***num_char*** *&* ***line_breaks*** *columns appear to be* ***associated*** *or* ***independent****?*

EXPLORING THE DATA DISTRIBUTION – BOXPLOTS

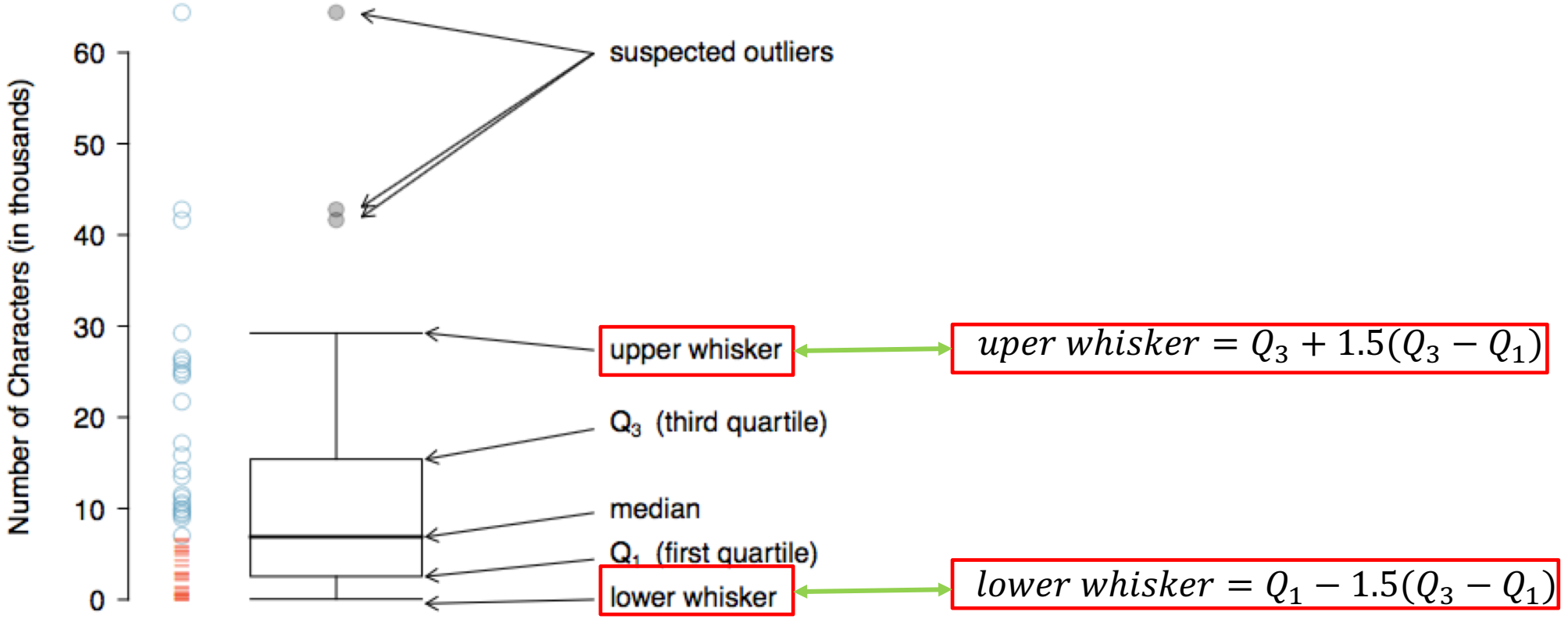
- **Boxplot** is a visualization technique used for explaining important features of the distribution of the target numerical variable by providing insight into: ***centrality, spread, skewness***, *and* ***possible outliers***.

spam	num_char	line_breaks	format	number
no	21,705	551	html	small
no	7,011	183	html	big
yes	631	28	text	none
⋮	⋮	⋮	⋮	⋮
no	15,829	242	html	small

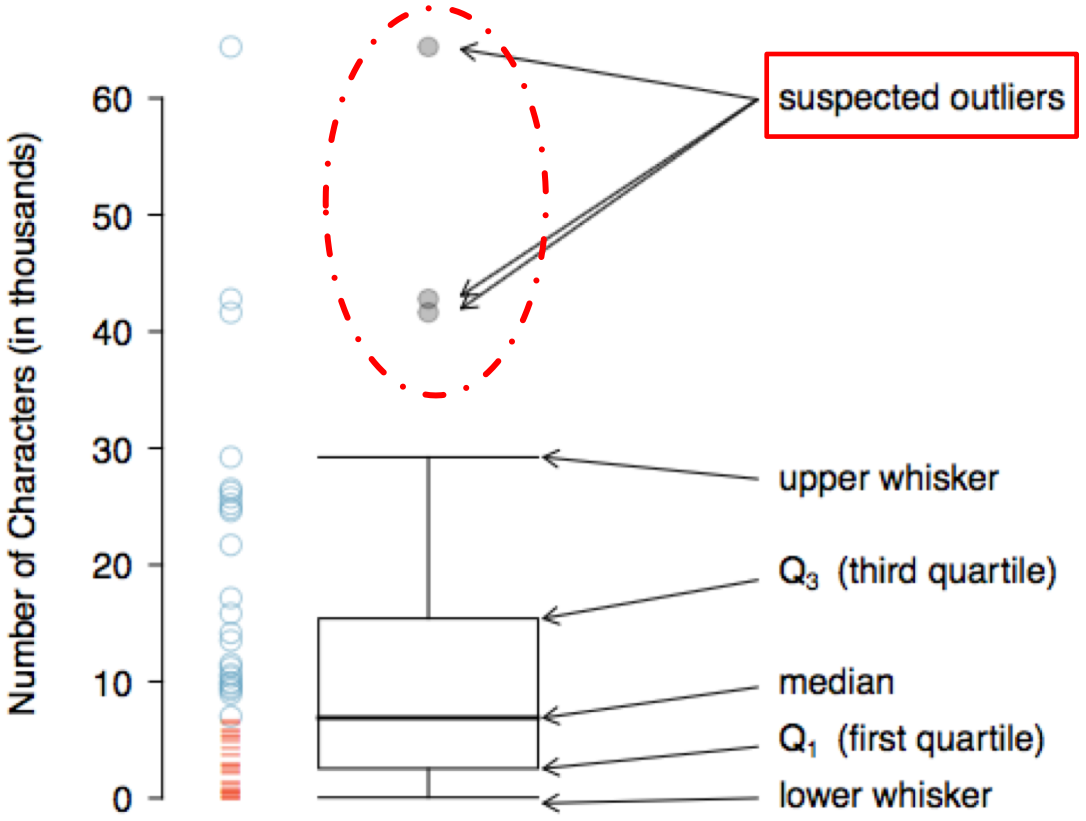


EXPLORING THE DATA DISTRIBUTION – BOXPLOTS

- **Boxplot** is a visualization technique used for explaining important features of the distribution of the target numerical variable by providing insight into: ***centrality, spread, skewness***, *and* ***possible outliers***.



EXPLORING THE DATA DISTRIBUTION – BOXPLOTS



Suspected outliers are the observation beyond the maximum reach of the whiskers.

HINT: *An outlier is an observation that appears extreme relative to the rest of the data*

HINT: *Why is it important to look for outliers?*

1. *To identify strong skew in the distribution.*
2. *To identify data collection or entry errors.*
3. *To get an insight into interesting properties of the data.*

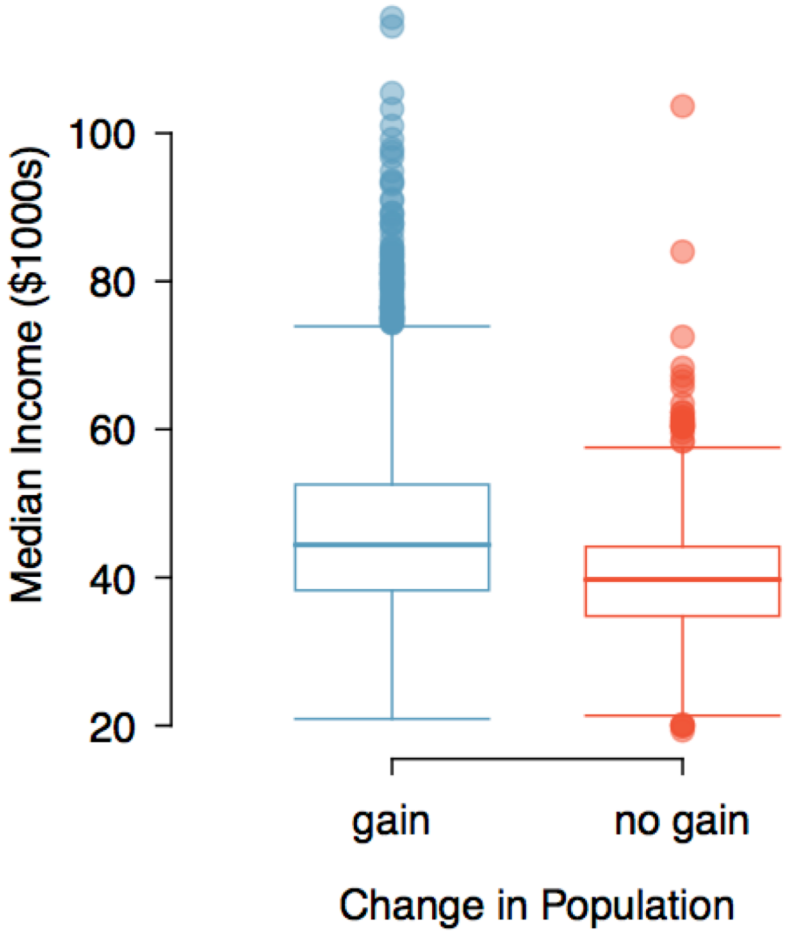
EXPLORING THE DATA DISTRIBUTION – SIDE-BY-SIDE BOXPLOT

- **Side-by-Side Boxplots** is a traditional tool for comparing numerical observations across categories (particularly useful for comparing **centrality** and **spread** of numerical observations between categories)

population gain					
41.2	33.1	30.4	37.3	79.1	34.5
22.9	39.9	31.4	45.1	50.6	59.4
47.9	36.4	42.2	43.2	31.8	36.9
50.1	27.3	37.5	53.5	26.1	57.2
57.4	42.6	40.6	48.8	28.1	29.4
43.8	26	33.8	35.7	38.5	42.3
41.3	40.5	68.3	31	46.7	30.5
68.3	48.3	38.7	62	37.6	32.2
42.6	53.6	50.7	35.1	30.6	56.8
66.4	41.4	34.3	38.9	37.3	41.7
51.9	83.3	46.3	48.4	40.8	42.6
44.5	34	48.7	45.2	34.7	32.2
39.4	38.6	40	57.3	45.2	33.1
43.8	71.7	45.1	32.2	63.3	54.7
71.3	36.3	36.4	41	37	66.7
50.2	45.8	45.7	60.2	53.1	
35.8	40.4	51.5	66.4	36.1	

no gain		
40.3	33.5	34.8
29.5	31.8	41.3
28	39.1	42.8
38.1	39.5	22.3
43.3	37.5	47.1
43.7	36.7	36
35.8	38.7	39.8
46	42.3	48.2
38.6	31.9	31.1
37.6	29.3	30.1
57.5	32.6	31.1
46.2	26.5	40.1
38.4	46.7	25.9
36.4	41.5	45.7
39.7	37	37.7
21.4	29.3	50.1
43.6	39.8	

TABLE: *Median household income from a random sample of 100 counties that had population gain are shown on the left. Median incomes from a random sample of 50 counties that had no population gain are shown on the right.*



EXPLORING THE DATA DISTRIBUTION – SIDE-BY-SIDE BOXPLOT

- **Side-by-Side Boxplots** is a traditional tool for comparing numerical observations across categories (particularly useful for comparing **centrality** and **spread** of numerical observations between categories)

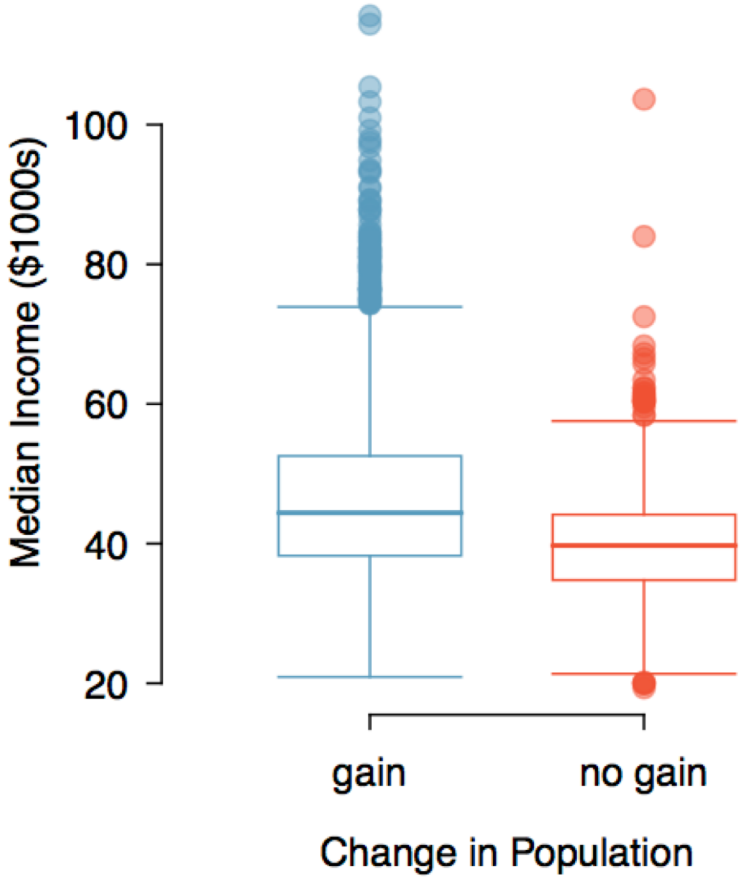
TASK: *Compare the incomes for counties across the two groups.*

Q: *What do you notice about the approximate centre of each group?*

A: ?

Q: *What do you notice about the variability between groups?*

A: ?



EXPLORING THE DATA DISTRIBUTION - HISTOGRAMS

- **Histograms** are plots that are used for **describing the shape of the data distribution** of the target numerical variable
- **Histograms** also provide a view of **data density** of the target numerical variable
  - Higher bars represent where the data are relatively more common
- **Histograms** visually look like **bar charts**, however there are some subtle differences:
  1. Histograms are used for displaying distributions of numerical variables, while bar charts are used for categorical variables
  2. Similar to bar charts histograms also measure frequencies, however it is first necessary to "bin" the observations of the target numerical variable, meaning to define intervals and then count the number of observations that fall within each one (***the chosen bin width can alter the story that histogram is telling***).

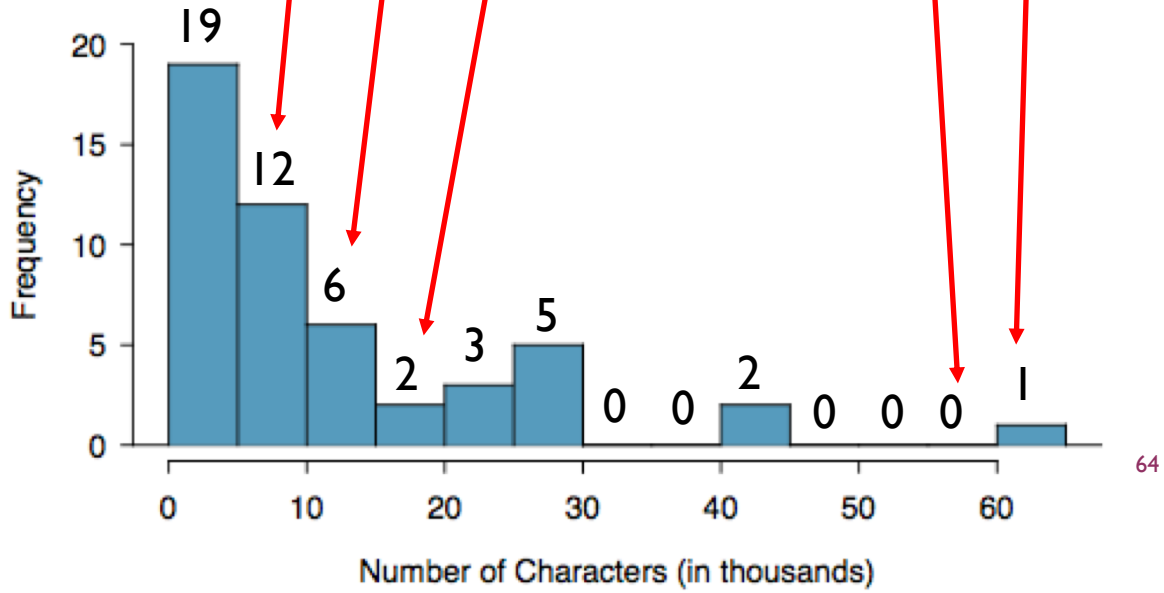
EXPLORING THE DATA DISTRIBUTION - HISTOGRAMS

HOW HISTOGRAMS ARE CONSTRUCTED?

spam	num_char	line_breaks	format	number
no	21,705	551	html	small
no	7,011	183	html	big
yes	631	28	text	none
⋮	⋮	⋮	⋮	⋮
no	15,829	242	html	small

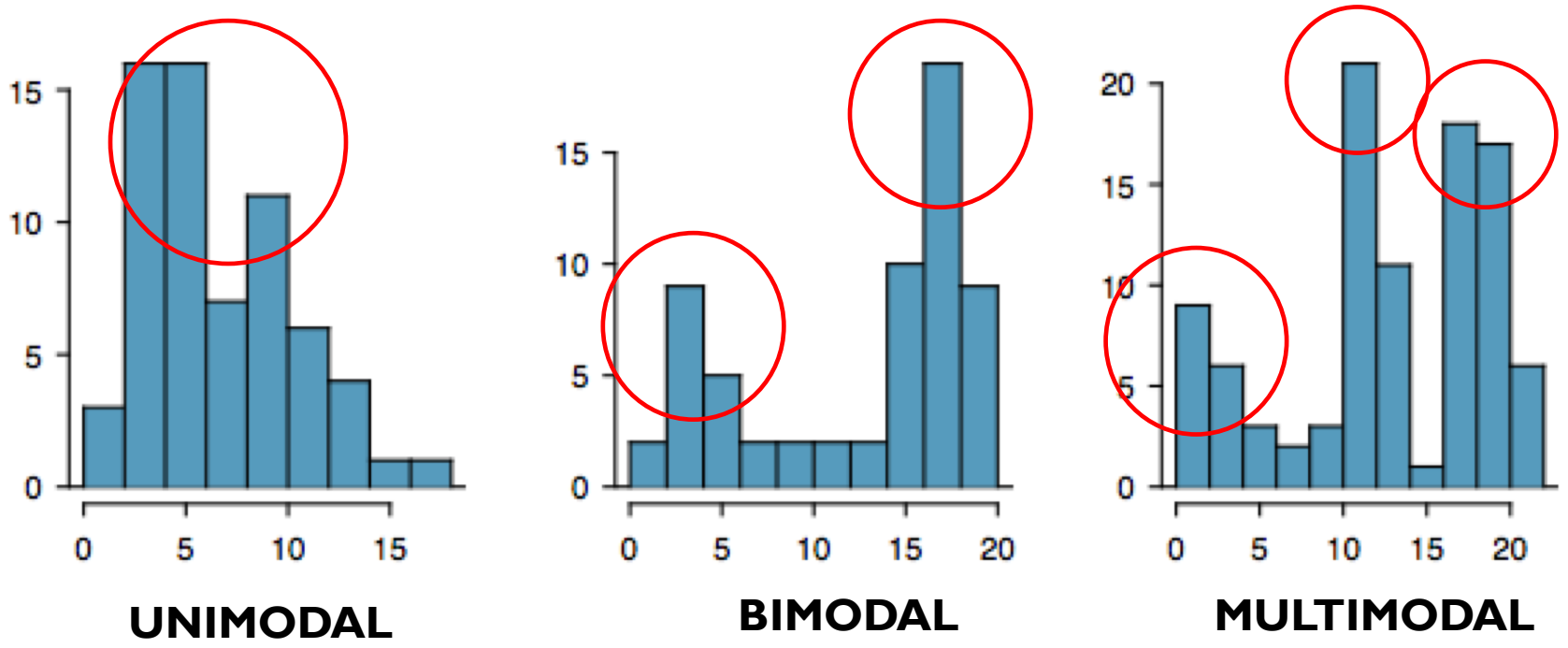
TABLE OF COUNTS FOR THE BINNED NUM_CHAR OBSERVATIONS

Characters (in thousands)	0-5	5-10	10-15	15-20	20-25	25-30	···	55-60	60-65
Count	19	12	6	2	3	5	···	0	1



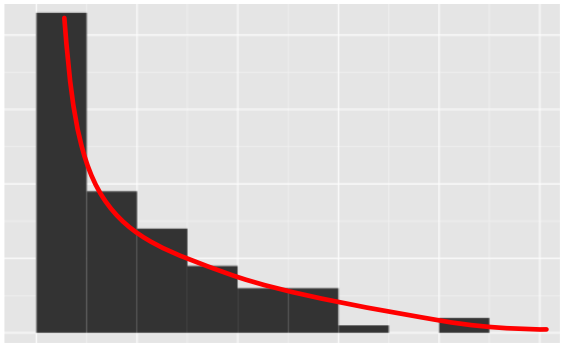
EXPLORING THE DATA DISTRIBUTION - HISTOGRAMS

- **Histograms** are plots that are used for **describing the shape of the data distribution** of the target numerical variable:
  - ***The mode*** *is represented by a prominent peak in the distribution*

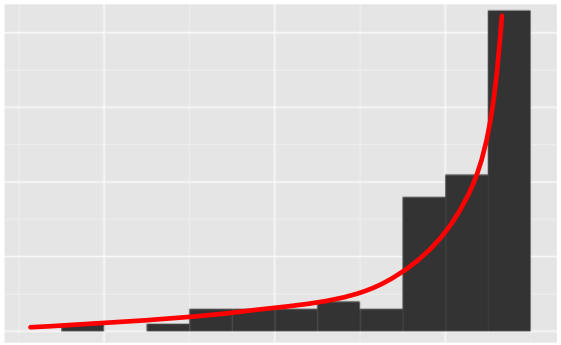


EXPLORING THE DATA DISTRIBUTION - HISTOGRAMS

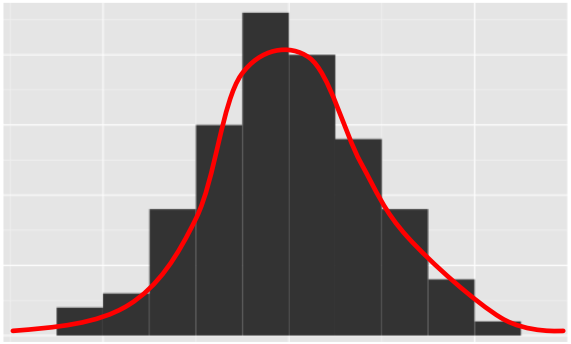
- **Histograms** are plots that are used for **describing the shape of the data distribution** of the target numerical variable:
  - When data trail off to the right i.e. have a **longer right tail**, the shape is said to be ***right skewed.***
  - Data sets with reverse characteristic – a **long tail to the left**, the shape is said to be ***left skewed.***
  - Data sets that show **roughly equal trailing off in both directions** are called ***symmetric***.



RIGHT SKEWED



LEFT SKEWED



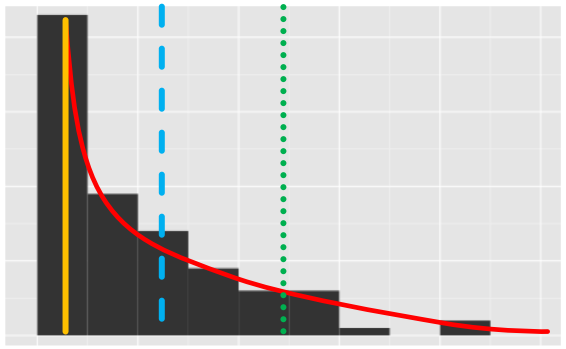
SYMETRIC

UNIFORM

EXPLORING THE DATA DISTRIBUTION - HISTOGRAMS

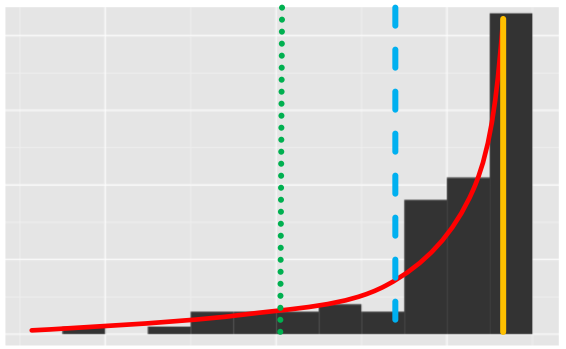
- If the distribution is SYMMETRIC, it is often more helpful to use the MEAN and SD to describe the center and spread
  - The following rule applies: $mean \approx median \approx mode$
- If the distribution is skewed or has extreme outliers, it is often more helpful to use the MEDIAN and IQR to describe the center and spread
  - If the distribution is RIGHT SKEWED, the following rule applies: $mode < median < mean$
  - If the distribution is LEFT SKEWED, the following rule applies: $mean < median < mode$

RIGHT SKEWED



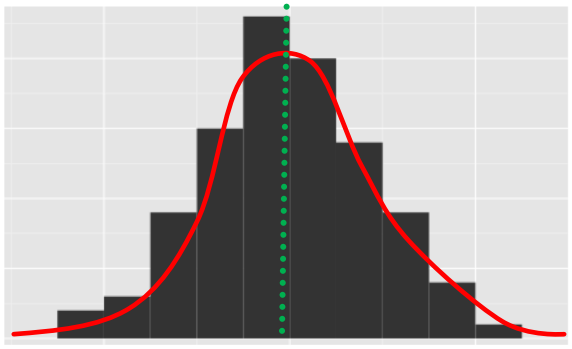
$mode < median < mean$

LEFT SKEWED



$mean < median < mode$

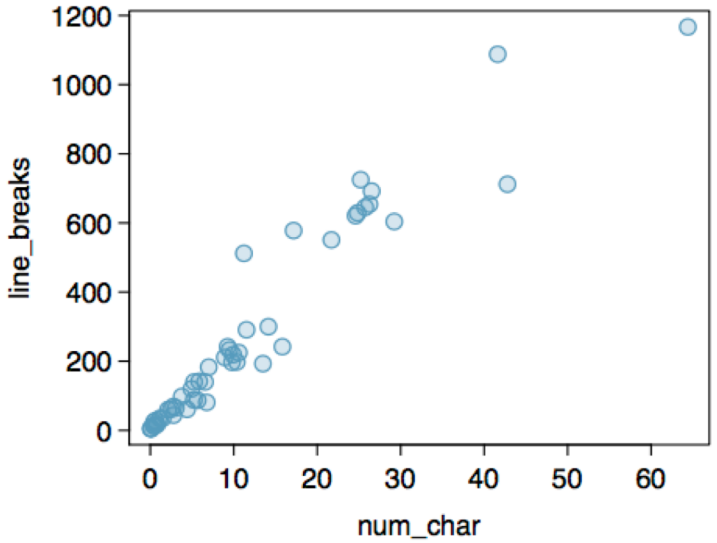
SYMETRIC

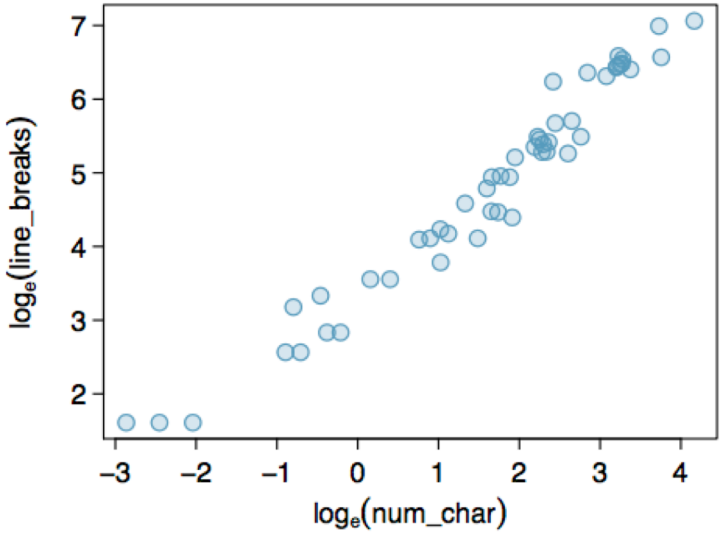


$mean \approx median \approx mode$

EXPLORING THE DATA DISTRIBUTION - HISTOGRAMS

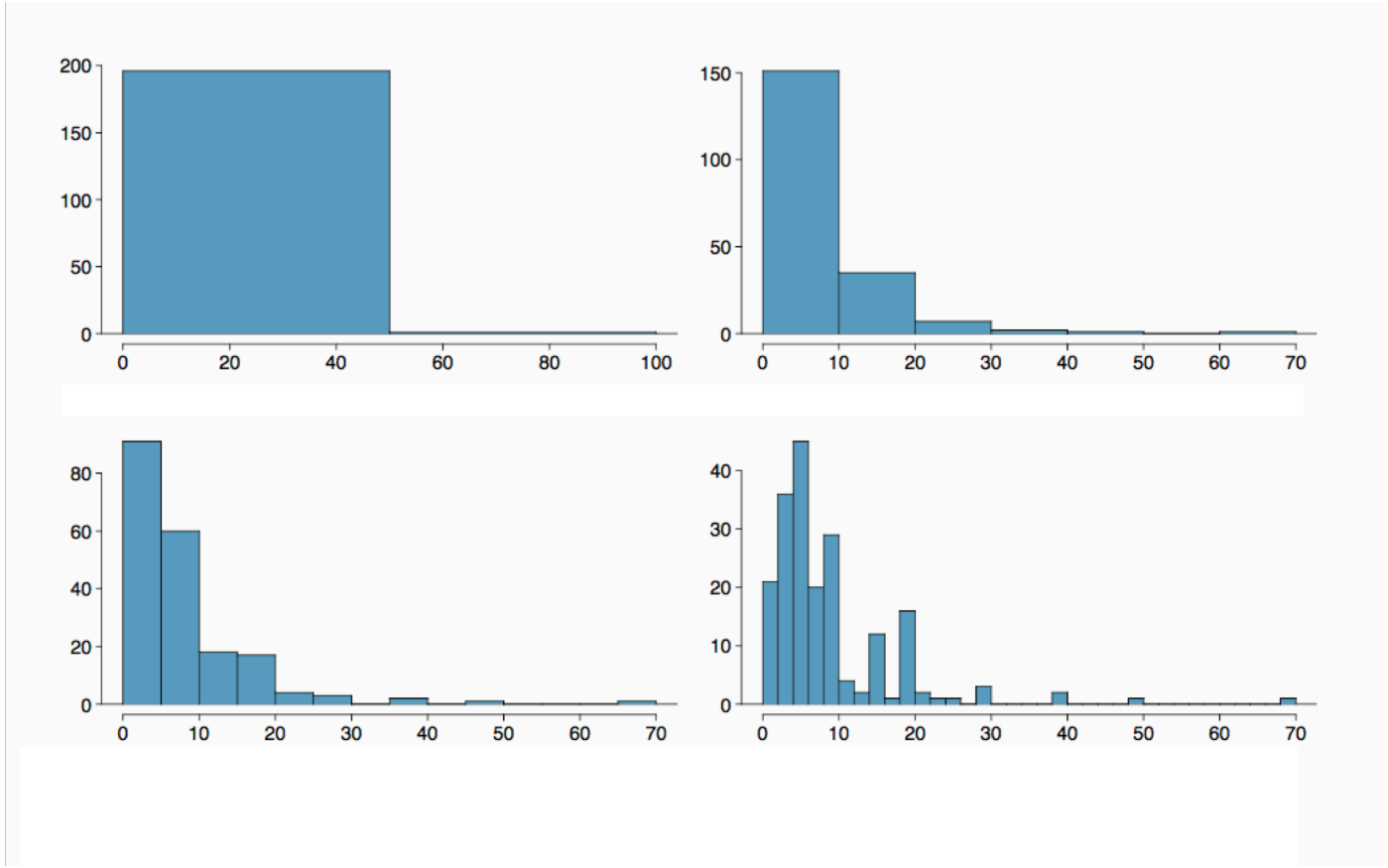
- Many statistical analysis techniques require data which are approximately symmetrical or not too skewed.
- **If data are skewed to the right**, transformations such as $y = \sqrt{x}$, $y = ln(x)$ and $y = -\frac{1}{x}$ (in increasing order of skewness severity) give new samples which are less skewed
- **If data are skewed to the left**, transformations such as $y = x^2$ and $y = x^3$ (in increasing order of skewness severity) give new samples which are less skewed



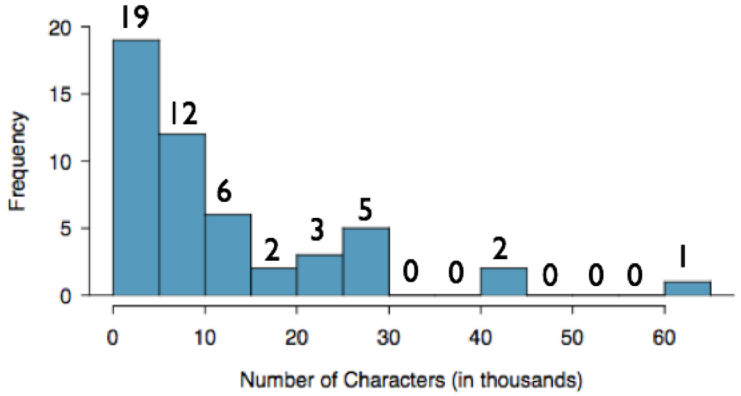


EXPLORING THE DATA DISTRIBUTION - HISTOGRAMS

- ***The chosen bin width can alter the story that histogram is telling***



ABSOLUTE FREQUENCY HISTOGRAM VS RELATIVE FREQUENCY HISTOGRAM

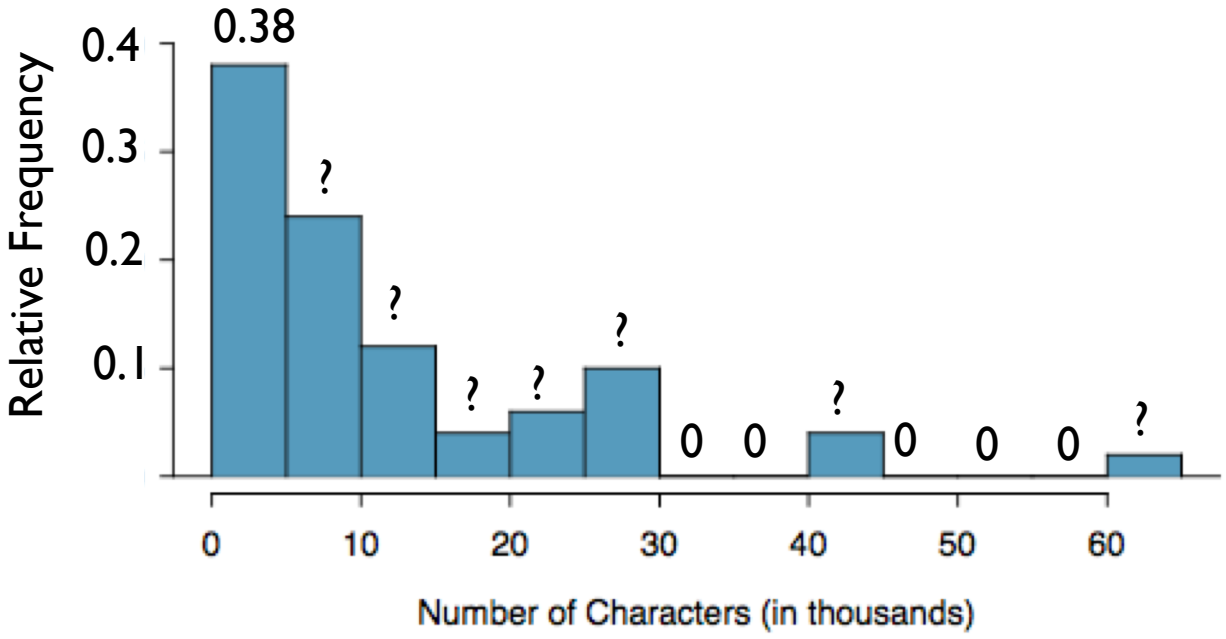


Absolute Frequency Histogram

$$total_{char} = 19 + 12 + 6 + 2 + 3 + 5 + 2 + 1 = 50 \text{ (in thousands)}$$

$$total_{RF} = \frac{19}{50} + \frac{?}{50} + \frac{?}{50} + \frac{?}{50} + \frac{?}{50} + \frac{?}{50} + \frac{?}{50} + \frac{?}{50}$$
$$= 0.38 + ? + ? + ? + ? + ? + ? + ? = 1$$

- Relative frequency histograms have the same shape as corresponding absolute frequency histograms
- Total area of all bars in RF histogram is equal to 1
- **Use the Relative Frequency histogram when you want to investigate whether the proportion is less than or greater than a certain value**

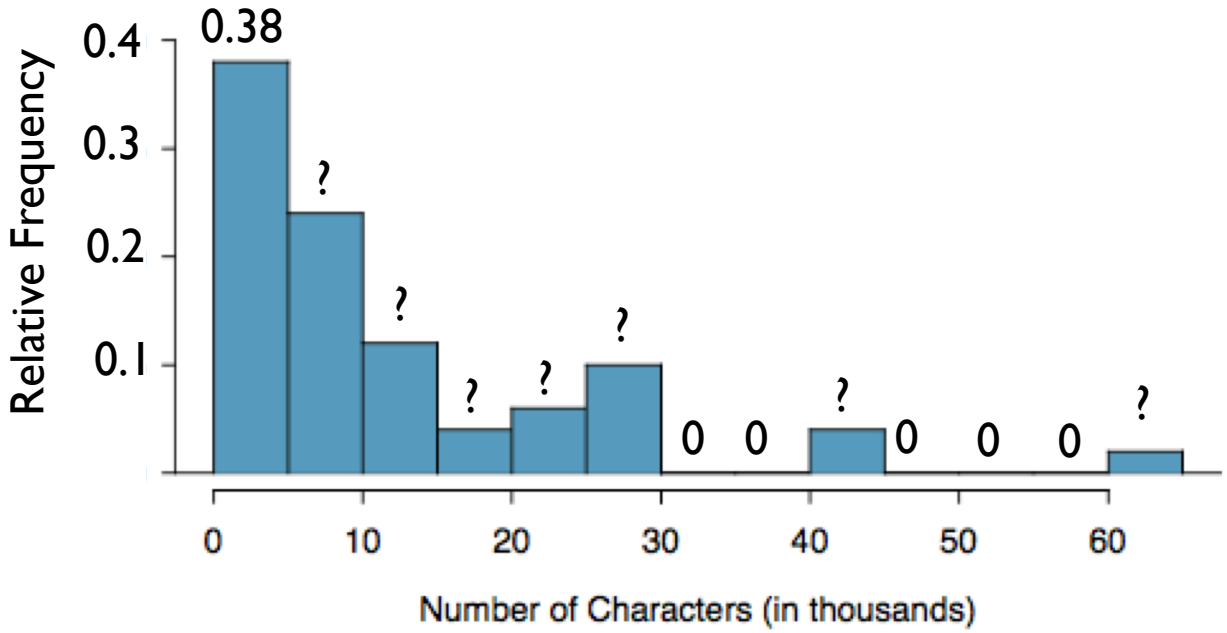


Relative Frequency Histogram

RELATIVE FREQUENCY HISTOGRAMS (EXAMPLE)

- ***Q1:*** *What percent of emails have between 10 and 25 thousands of characters?*
- ***A1:?***

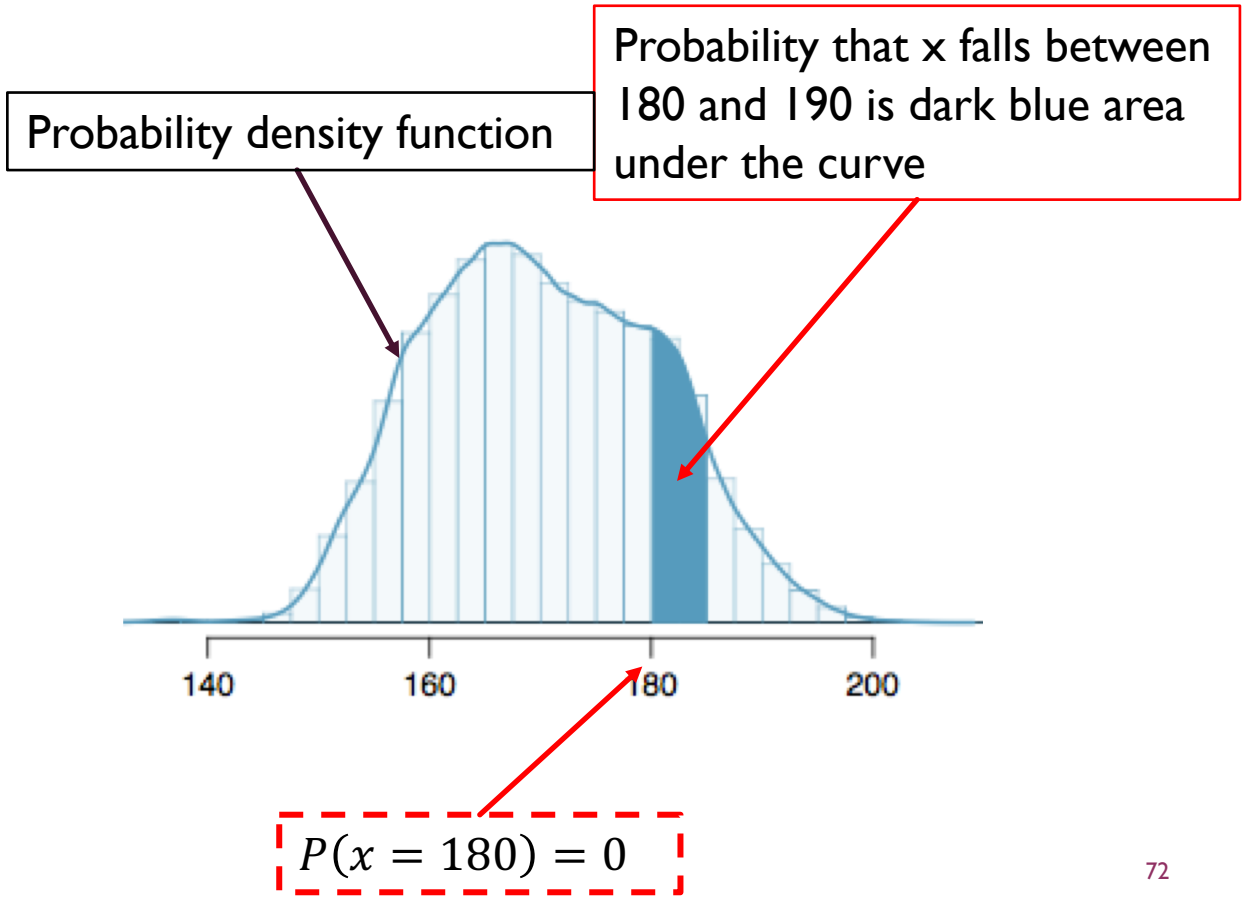
- ***Q2:*** *What percent of emails have between 12.5 and 22.5 thousand characters?*
- ***A2:*** *?*



Relative Frequency Histogram

FROM HISTOGRAMS TO CONTINUOUS DISTRIBUTIONS – DENSITY CURVES

- A density curve is a smoothed version of relative frequency histogram, and it is used for the visualization of continuous variables or very large populations.
  - This smooth curve represents a **probability density function** (also called **density** or **distribution**)
- Total area under the curve is equal to 1 (which is analogue to the total area of all bars in relative frequency histograms)
  - Measuring areas under a density curve corresponds to measuring probabilities
- ***NOTE:*** *Probability that x is equal to some value from continuous distribution is ALWAYS equal to 0. This happens because a single point on density curve diagram has no width and therefore area under the curve is equal to 0*



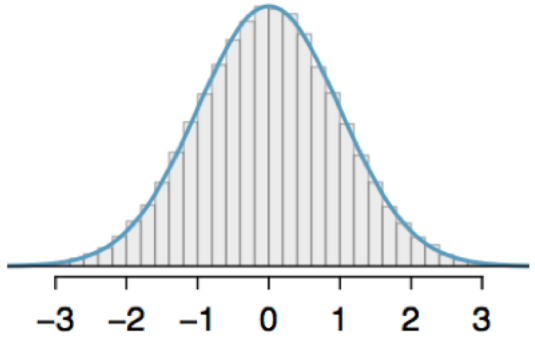
EXPLORING THE DATA DISTRIBUTION – NORMAL DISTRIBUTIONS

- Among all the distributions we see in practice, one is overwhelmingly the most common. Indeed it is so common, that people often know it as the **normal curve** or **normal distribution**
- Properties of the normal distribution are:
  - It is unimodal and symmetric around its mean bell shaped curve
  - Mean, Median and Mode are equal
  - It is determined by two parameters: mean (μ) and standard deviation (σ), and it is usually denoted as $N(\mu, \sigma)$.
  - **The area under the normal curve is 1**
- ***NOTE: Normal distribution $N(\mu = 0, \sigma = 1)$, with mean $\mu = 0$ and standard deviation $\sigma = 1$, is called STANDARD NORMAL DISTIBUTION***

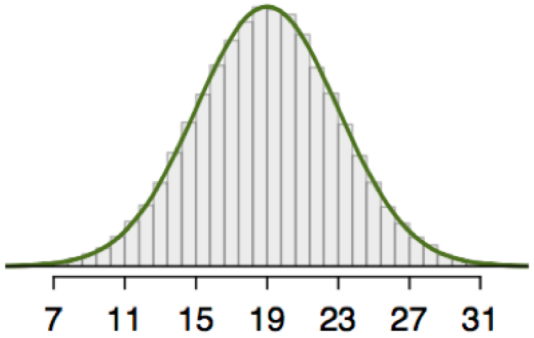
EXPLORING THE DATA DISTRIBUTION – NORMAL DISTRIBUTIONS

- Changing the mean shifts the bell curve to the left or right, while changing the standard deviation stretches or constricts the curve.

$N(\mu = 0, \sigma = 1)$

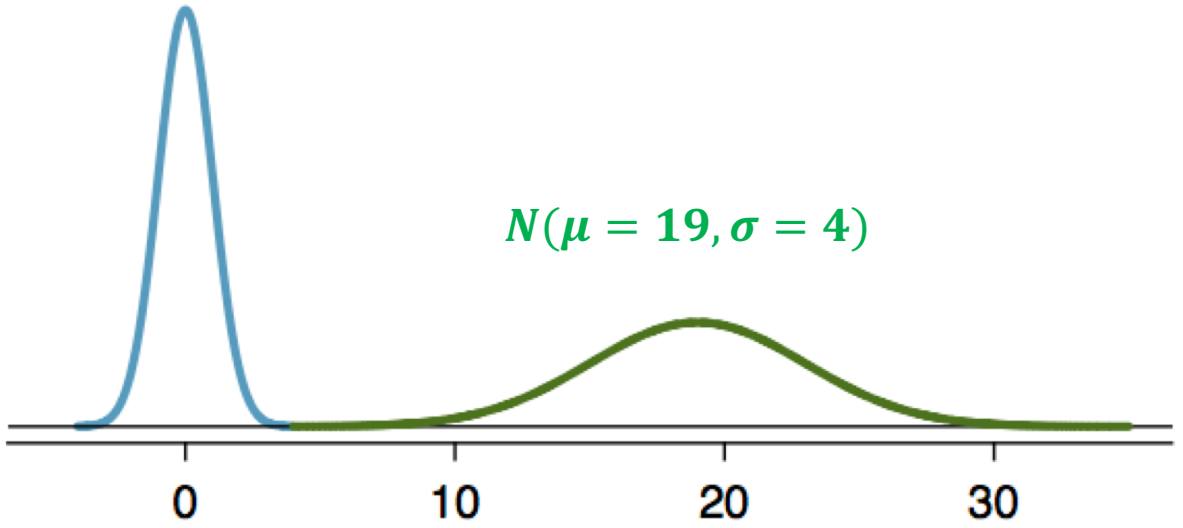


$N(\mu = 19, \sigma = 4)$



$N(\mu = 0, \sigma = 1)$

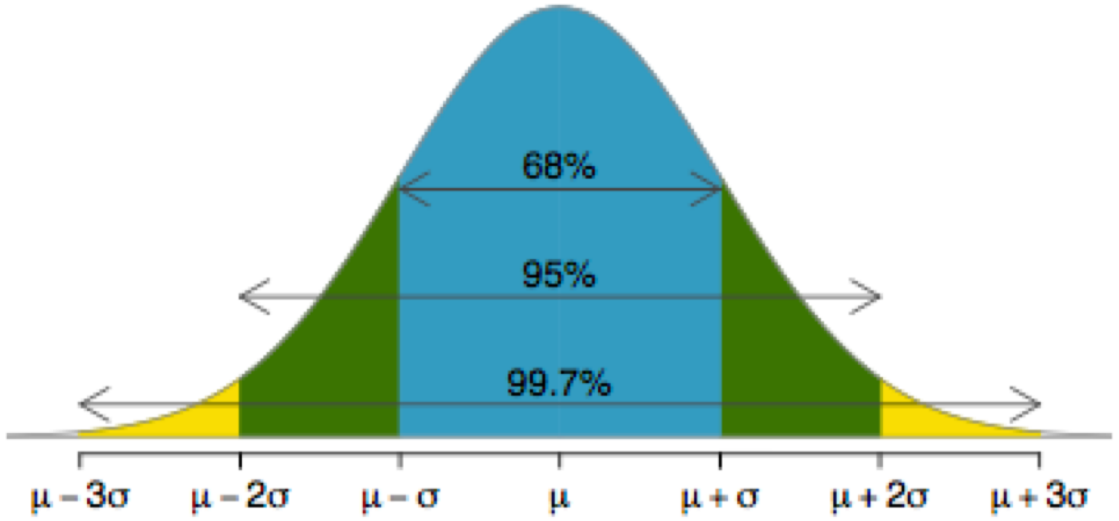
$N(\mu = 19, \sigma = 4)$



The normal models $N(\mu = 0, \sigma = 1)$ and $N(\mu = 0, \sigma = 1)$ plotted on the same scale

THE KEY FACTS ABOUT THE NORMAL DISTRIBUTION – 68-95-99.7 RULE

- Approximately 68% of observations lie within 1 standard deviations from the mean in the normal distribution
- Approximately 95% of observations lie within 2 standard deviations from the mean in the normal distribution
- Approximately 99.7% of observations lie within 3 standard deviations from the mean in the normal distribution



ANALYZING NORMALLY DISTRIBUTED DATA – Z-SCORES

- In order to analyze normally distributed data we should convert available observations into the standard deviation units and measure their distances from the mean.
  - To perform this type of conversion we use the standardization technique called Z-score.
- The Z-score of the observation is the number of standard deviations it falls above or below the mean. For an observation x that follows the normal distribution $N(\mu, \sigma)$, we calculate the Z-score using the following formula:

$$z = \frac{x - \mu}{\sigma}$$

  - ***Example 1:*** *if the observation is one standard deviation above the mean, its Z-score is 1, i.e.* $z = 1$
  - ***Example 2:*** *if the observation is one 1.5 standard deviation below the mean, its Z-score is -1.5, i.e.* $z = -1.5$
- ***NOTE:*** *We can use Z-scores to roughly identify which observations are more unusual than others.*
  - One observation x_1 is said to be more unusual than another observation x_2 if the absolute value of its Z-score is larger than the absolute value of the other observation's Z- score, i.e $|z_1| > |z_2|$

Z-SCORES EXAMPLE

- Head lengths of brushtail possums follow a nearly normal distribution with mean 92.6 mm and standard deviation 3.6 mm.
  a) Compute the Z- scores for possums with head lengths of 95.4 mm and 85.8 mm.
  b) Use calculated Z-scores to determine how many standard deviations above or below the mean measured head lengths of these two possums fall
  c) Head length of which possum is more unusual?

SOLUTION:

a)

$$x_1 = 95.4 \; and \; x_1 \sim N(\mu = 92.6, \sigma = 3.6)$$

$$x_2 = 85.8 \; and \; x_2 \sim N(\mu = 92.6, \sigma = 3.6)$$

$$z_1 = \frac{x_1 - \mu}{\sigma} = \; ?$$

$$z_2 = \frac{x_2 - \mu}{\sigma} = \; ?$$

b)

A possum with the head length of 95.4 is ? standard deviations ABOVE the mean

A possum with the head length of 85.8 is ? standard deviations BELOW the mean

c) Because ?, opossum with the head length of ? mm is more unusual than opossum with the head length of ? mm

ANALYZING NORMALLY DISTRIBUTED DATA – Z – SCORES

NOTE:

- ***If a random variable $X \sim N(\mu, \sigma)$, then the random variable $Z = \frac{x-\mu}{\sigma} \sim N(0, 1)$***
  - This implies that if your original data (x) is (approximately) normally distributed, their z scores are distributed (approximately) $N(0,1)$
- By calculating a Z-score we "convert" the data value from its normal distribution $N(\mu, \sigma)$ to a value from the standard normal distribution $N(0,1)$ in such a way it maintains all the properties of the original dataset.
  - This means that in order to calculate percentiles for any $N(\mu, \sigma)$ distribution, all we need is $N(0,1)$ percentiles which are listed in normal probability table and calculated from the corresponding z scores.

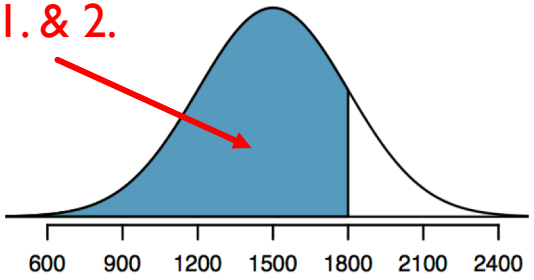
AN ALGORITHM FOR SOLVING NORMAL PROBABILITY PROBLEMS

- Steps to get through many different calculations involving percentiles in normal distributions, when you are given a cutoff value and asked to calculate a probability:
  1. Draw and label a picture of the normal distribution (doesn't need to be exact)
  2. Shade in the region of interest
  3. Calculate the z-score of the cutoff value
  4. Look up the percentile for the z-score in the normal probability table
  5. Do you need to subtract from 1?
- Always verify that the final answer makes sense with the picture you drew

NORMAL PROBABILITY EXAMPLES – FACEBOOK EXAMPLE 1

- Suppose the average number of Facebook friends is approximated well by the normal model $N(\mu = 1500, \sigma = 300)$. Randomly selected person Julie has 1800 friends. She would like to know what percentile she falls among other Facebook users? What is the percentage of people that have more friends than Julie?

1. & 2.



3. $z = \frac{x-\mu}{\sigma} = \frac{1800-1500}{300} = 1$

4.

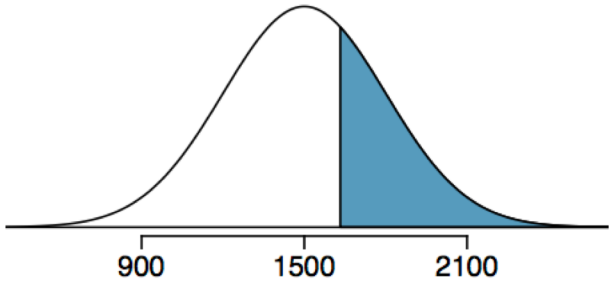
	Second decimal place of Z									
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015

Julie is 84.13th percentile

If 84.13% have less facebook friends than Julie, than the proportion of people that have more friends is 15.87%

NORMAL PROBABILITY EXAMPLES – FACEBOOK EXAMPLE 2

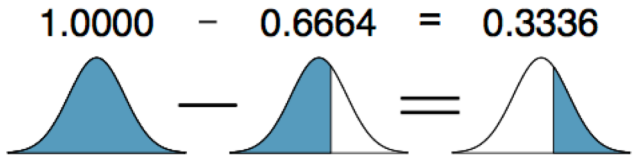
- Suppose the average number of Facebook friends is approximated well by the normal model $N(\mu = 1500, \sigma = 300)$. What is the probability that a randomly selected person has **AT LEAST** 1630 friends on Facebook?



$$z = \frac{x-\mu}{\sigma} = \frac{1630-1500}{300} = 0.43$$

Z	0.00	0.01	0.02	*0.03*
0.0	0.5000	0.5040	0.5080	0.5120
0.1	0.5398	0.5438	0.5478	0.5517
0.2	0.5793	0.5832	0.5871	0.5910
0.3	0.6179	0.6217	0.6255	0.6293
0.4	0.6554	0.6591	0.6628	*0.6664*
0.5	0.6915	0.6950	0.6985	0.7019
0.6	0.7257	0.7291	0.7324	0.7357

To find the area above Z = 0.43, we compute one minus the area of the lower tail



The probability that randomly selected person has at least 1630 friends on Facebook is **0.3336**

From the normal probability table we can see that ***0.6664 is the probability that the randomly selected person has Z score LESS than 0.43 (lower tail),*** i.e. less than 1630 friend.

NORMAL PROBABILITY EXAMPLES – FACEBOOK EXAMPLE 3

- Suppose the average number of Facebook friends is approximated well by the normal model $N(\mu = 1500, \sigma = 300)$. A randomly selected person is at the 79.95th percentile. How many Facebook friends does this person have?

Z	0.00	0.01	0.02	*0.03*	0.04
0.0	0.5000	0.5040	0.5080	0.5120	0.5160
0.1	0.5398	0.5438	0.5478	0.5517	0.5557
0.2	0.5793	0.5832	0.5871	0.5910	0.5948
0.3	0.6179	0.6217	0.6255	0.6293	0.6331
0.4	0.6554	0.6591	0.6628	*0.6664*	0.6700
0.5	0.6915	0.6950	0.6985	0.7019	0.7054
0.6	0.7257	0.7291	0.7324	0.7357	0.7389
0.7	0.7580	0.7611	0.7642	0.7673	0.7704
0.8	0.7881	0.7910	0.7939	0.7967	0.7995

$$z = 0.84 = \frac{x_{friends} - 1500}{300}$$

$$x_{friends} = 1500 + 0.84 \times 300 = 1752$$

Randomly selected person, which is at the 79.95th percentile, has 1752 friends on Facebook

NORMAL PROBABILITY EXAMPLES – QUALITY CONTROL EXAMPLE

- At Heinz factory the amounts which go into bottles of ketchup are supposed to be normally distributed with mean 36 oz. standard deviation 0.11 oz. Once every 30 minutes a bottle is selected from the production line, and its contents are noted precisely. If the amount of ketchup in the bottle is below 35.8 oz. or above 36.2 oz., then the bottle fails the quality control inspection. What percentage of bottles have less than 35.8 ounces of ketchup?

NORMAL PROBABILITY EXAMPLES – QUALITY CONTROL EXAMPLE

- **What percentage of bottles pass the quality control inspection?**

(a) 1.82%
(b) 3.44%
(c) 6.88%
(d) 93.12%
(e) 96.56%

NORMAL PROBABILITY EXAMPLES – HEALTHCARE EXAMPLE

- **Body temperatures of healthy humans are distributed nearly normally with mean $98.2°F$ and standard deviation $0.73°F$. What is the cutoff for the lowest 3% of human body temperatures?**

NORMAL PROBABILITY EXAMPLES – HEALTHCARE EXAMPLE

- **Body temperatures of healthy humans are distributed nearly normally with mean $98.2°F$ and standard deviation $0.73°F$. What is the cutoff for the highest 10% of human body temperatures?**

(a) $97.3°F$
(b) $99.1°F$
(c) $99.4°F$
(d) $99.6°F$

EVALUATING NORMAL DISTRIBUTION – STATISTICAL TESTS

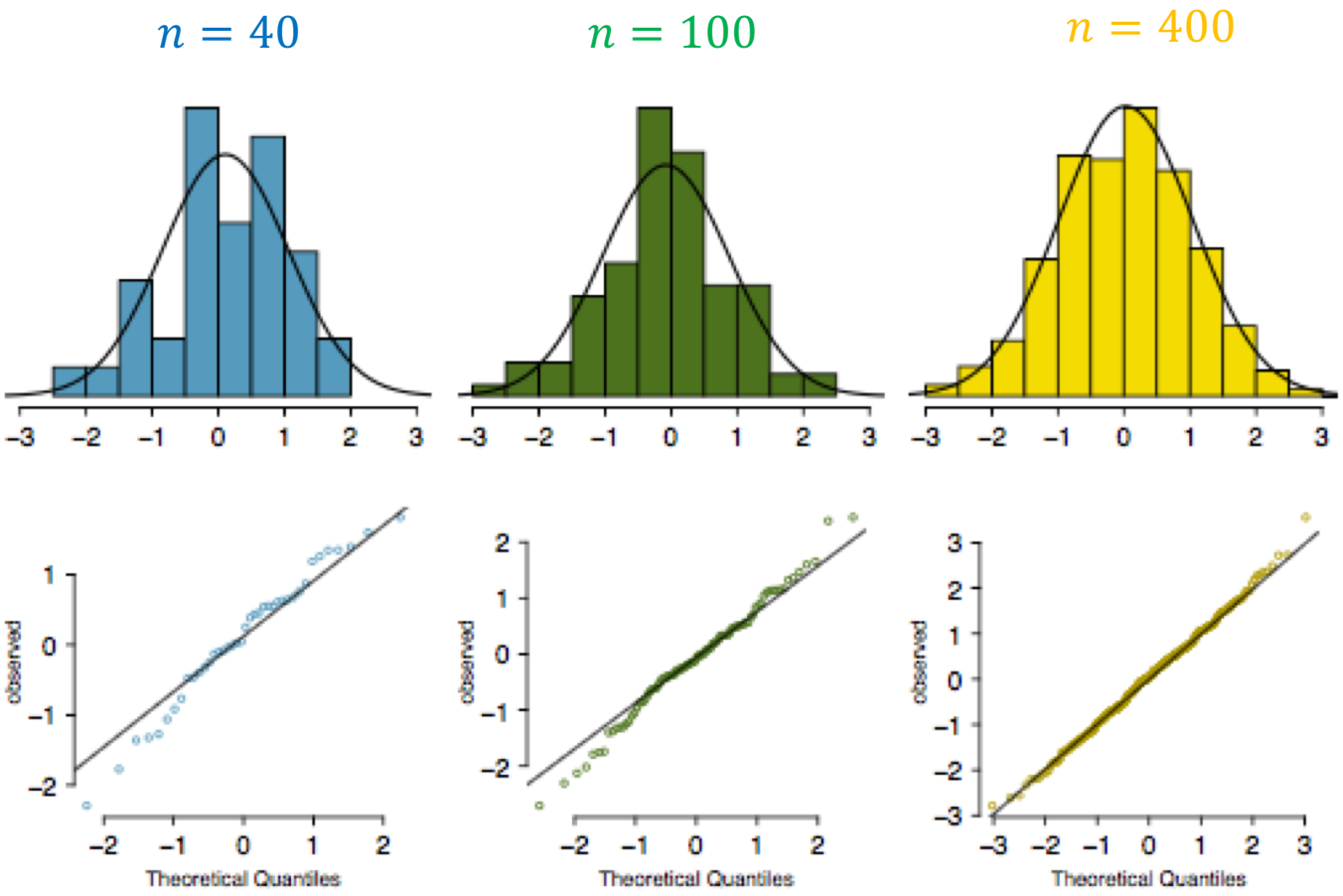
- Two of the most popular approaches to evaluating whether sample data follow the normal distribution or not, is by using **statistical tests** or by using **visualization techniques**.
- Statistical approaches of evaluating whether a given sample of data follow the normal distribution:
  - Shapiro-Wilk test,
  - Kolmogorov – Smirnov test,
  - Anderson – Darling test, etc.
- A main drawback of using statistical tests for evaluating if data follow the normal distribution is that they are very sensitive to the presence of outliers, i.e. if a certain number of outliers is present in a normally distributed dataset statistical tests would report that the data set is not drawn from a normal distribution (this problem can be overcome using visualization techniques).

EVALUATING NORMAL DISTRIBUTION – VISUALIZATION TECHNIQUES

Two visualization techniques that can be used for normality assessment are:

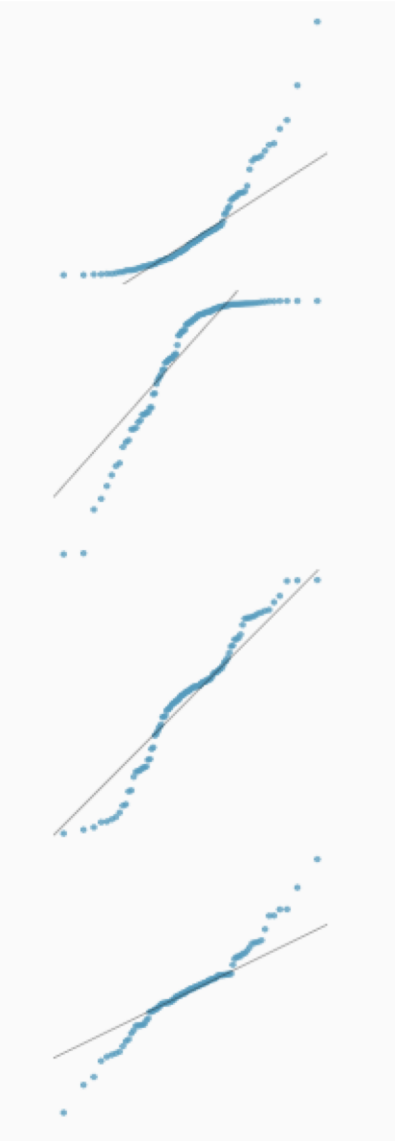
- Histogram with the best fitting normal curve overlaid on the plot.
  - The sample mean $\bar{x}$ and sample standard deviation σ are used as the parameters of the best fitting normal curve
  - The closer this curve fits the histogram, the more reasonable the normal model assumption is
- The normal probability plot (synonyms: quantile-quantile plot or QQ plot)
  - Data are plotted on the y-axis of a normal probability plot, and theoretical quantiles (following a normal distribution) on the x-axis.
  - The closer the points are to a perfect straight line, the more confident we can be that the data follow the normal model

EVALUATING NORMAL DISTRIBUTION – VISUALIZATION TECHNIQUES



An illustration of histograms and normal probability plots for three simulated normal data sets

EVALUATING NORMAL DISTRIBUTION – VISUALIZATION TECHNIQUES



Right skew – Points bend up and to the left of the line

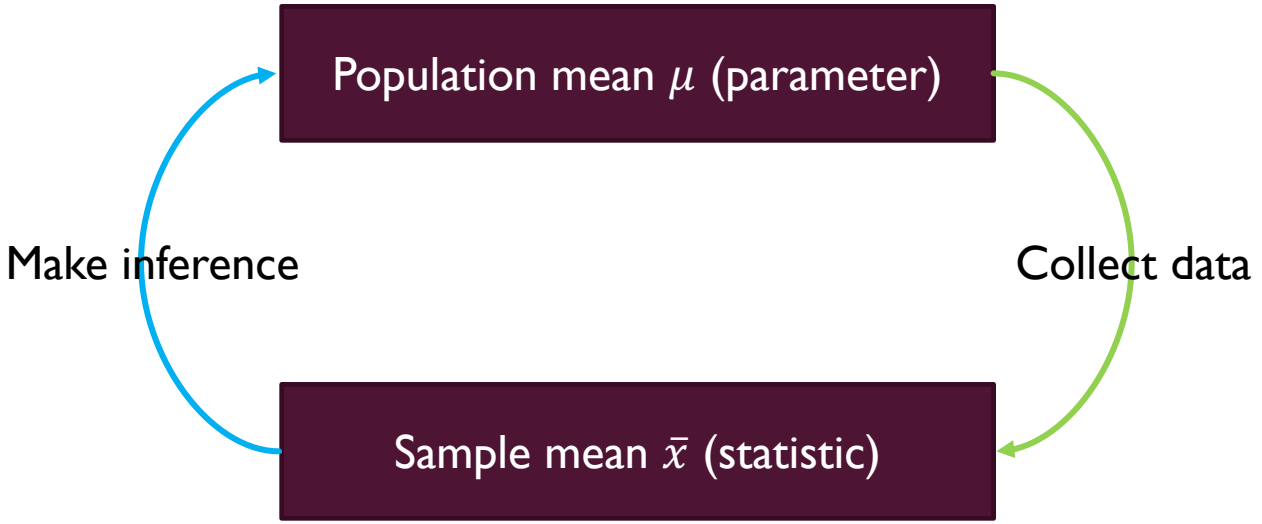
Left skew – Points bend down and to the right of the line

Short tails *(narrower than the normal distribution)* – Points follow an S shaped-curve

Long tails *(wider than the normal distribution)* – Points start below the line, bend to follow it, and end above it

SIGNIFICANCE TESTS

- The purpose of **statistical inference** is to allow us to draw conclusions about and assess population parameters for a specific population based on a sample of data taken from that population.
- Since it is very difficult (or impossible) to collect data from the complete population, we use **sample statistics** *(such as the mean, proportions, etc.)* as **point estimates** for the **unknown population parameters** of interest.



A conceptualization of statistical practice to illustrate the definitions of statistical inference.
*We use **sample mean ($\bar{x}$)** as a **point estimate** of the **population mean (μ)***

SAMPLING DISTRIBUTION

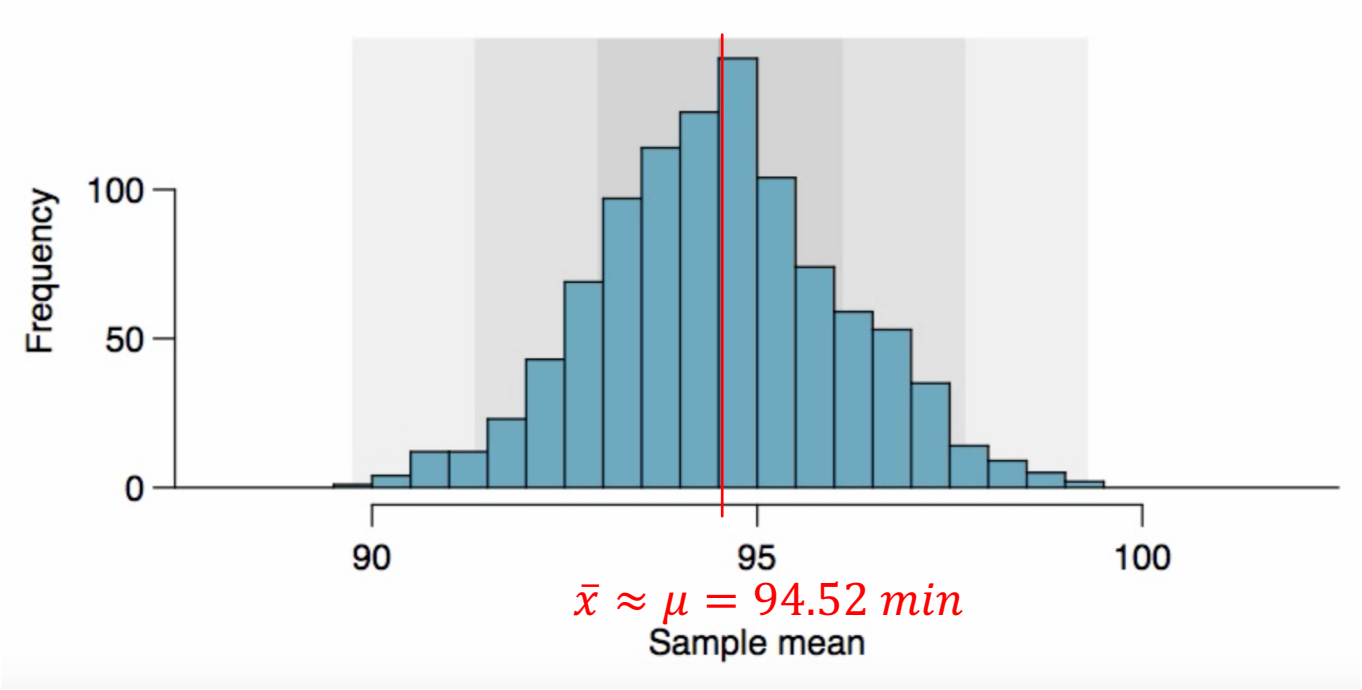
- Point estimates generally vary from one sample to another, and this sampling variation suggests our estimate may be close, but it will not be exactly equal to the true population parameter.
- The **sampling distribution** represents the distribution of the point estimates based on samples of a fixed size from a certain population (It is useful to think of a particular point estimate as being drawn from such a distribution)
- Being able to understand and model this natural variability inherent in estimated sample statistics (using relevant sampling distributions) is a key part of many statistical analyses.
- Like any other probability distribution, the central "balance" point of a sampling distribution is its **mean**, but the standard deviation of a sampling distribution is referred to as a ***standard error***.
  - ***NOTE:*** *The slight change in terminology reflects the fact that the probabilities of interest are no longer tied to raw measurements or observations per se, but rather to a quantity calculated from a sample of such observations.*
- The standard error of an estimate describes how far the point estimate is from the the true population parameter
- ***NOTE: Do not confuse standard deviation (which measures the variability of individual data points inside the sample) with the standard error (which measures how far the point estimate is from the the true population parameter).***

SAMPLING DISTRIBUTION - EXAMPLE

- Suppose we want to make conclusions about the average running time of the 10.000 people that attended Belfast marathon and lets assume that we have an access to the phone numbers of all contestants.
- We called the first 100 people at random and asked them about their running time. We use this sample mean as the point estimate of the mean of the population.
- Based on our first sample of 100 randomly selected people, we calculate that the mean of their running was 95.61 min.
- We decided to take a second sample of another 100 randomly selected people (with replacement, i.e. it is allowed to contact same people again) and we recorded that their average running time was 93.43 mins. Suppose we take another sample (95.30 min) and another (94.60 min), and so on.
- If we do this many many times – which we can do only because we have access to the phone numbers of all contestants – we can build up a **sampling distribution** for the sample mean when the sample size is 100 .

SAMPLING DISTRIBUTION – EXAMPLE (CONTINUED)

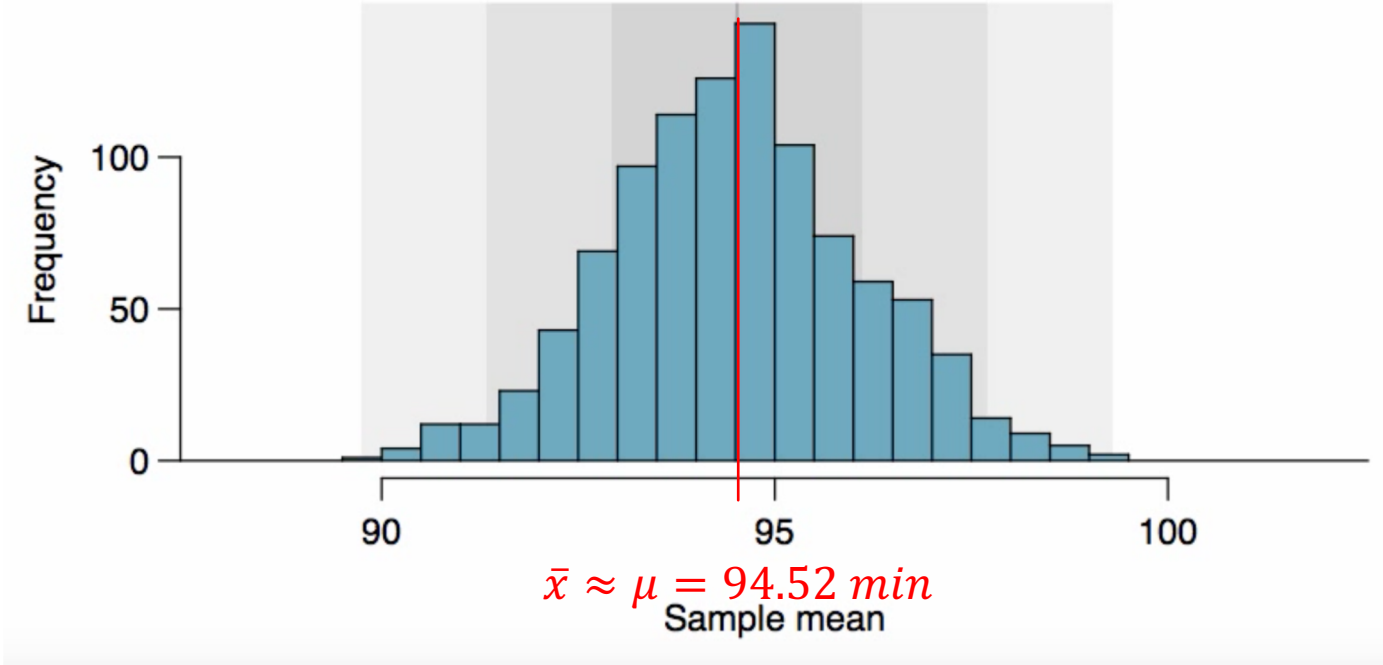
- For the sake of argument, suppose we knew that the average running (true population mean) time of all 10.000 contestants was $\mu = 94.52\ min$
- If we plot a histogram of collected sample means, we will get **the approximately normal distribution** which is centered exactly around the true population mean of 94.52 min. Intuitively, this makes sense, as the sample means should tend to "fall around" the true population mean



A histogram of 1000 sample means for average running time of Belfast marathon, where the samples are of size n = 100.

SAMPLING DISTRIBUTION – EXAMPLE (CONTINUED)

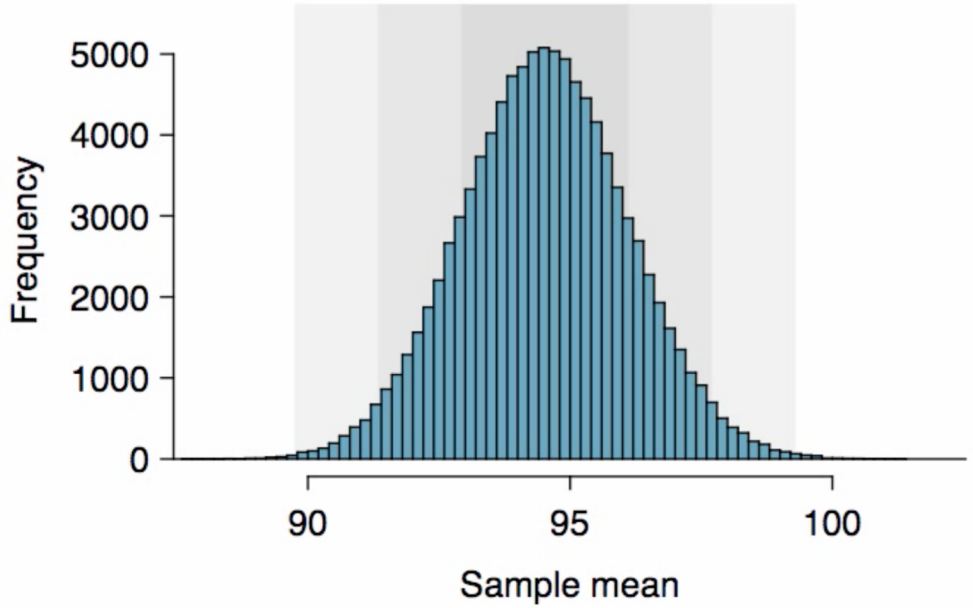
- We can see that the sample mean has some variability around the population mean, which can be quantified using the standard deviation of this distribution of sample means: $\sigma_{\bar{x}} = 1.92\ min$.
- The standard deviation of the sample mean $\sigma_{\bar{x}}$ tells us **how far the typical estimate is away from the actual population mean**, 1.92 min. It also describes the typical error of the point estimate, and for this reason we usually call this standard deviation the standard error (SE) of the estimate.

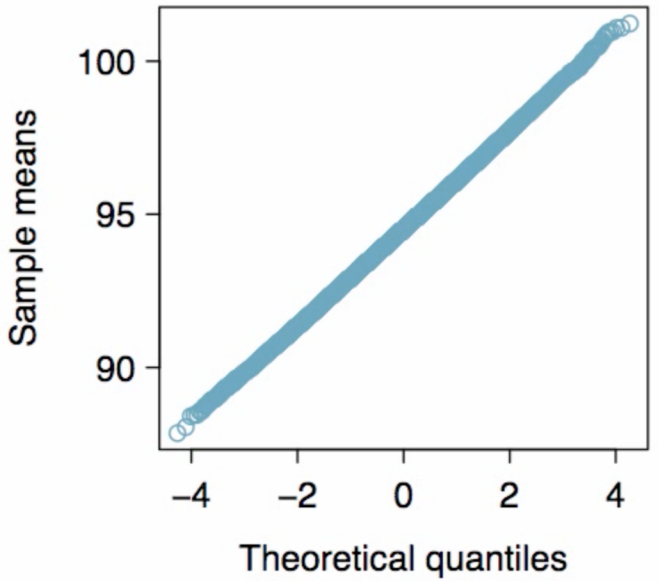


A histogram of 1000 sample means for average running time of Belfast marathon, where the samples are of size n = 100.

SAMPLING DISTRIBUTION – EXAMPLE (CONTINUED)

- Instead of collecting 1000 samples, lets collect 100.000 samples of size 100, calculate the mean of each and plot them in the histogram to get a very accurate depiction of the sampling distribution.
- As you can see in the histogram below (left figure) the distribution of sample means closely resembles the normal distribution.
- Because all of the points closely fall around a straight line in the normal probability plot (right figure), we can conclude the distribution of sample means is nearly normal.
- This result can be explained by the **Central Limit Theorem.**





CENTRAL LIMIT THEOREM

CENTRAL LIMIT THEOREM (CLT) informal description – if a sample consists of at least 30 independent observations and the data are not strongly skewed then the distribution of the sample mean is approximated well by the normal distribution

CENTRAL LIMIT THEOREM FOR THE SAMPLE MEAN

- The distribution of the sample mean is approximated well by a normal model:

$$\bar{x} \sim N(mean = \mu, SE = \frac{\sigma}{\sqrt{n}})$$

- Where:
  - **SE** represents **standard error of the mean**, which is defined as the standard deviation of the sampling distribution,
  - σ is the standard deviation of the population, and
  - n is the sample size

- ***NOTE: If the standard deviation of the population (σ) is unknown, use the standard deviation of the sample (s) to approximate the standard error of the mean:***

$$\bar{x} \sim N(mean = \mu, SE = \frac{s}{\sqrt{n}})$$

CENTRAL LIMIT THEOREM FOR THE SAMPLE MEAN

- The distribution of the sample mean is approximated well by the normal model:

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

The following conditions need to be met for the CLT to apply:

1. ***Independence:*** Sampled observations must be independent. This is difficult to verify, but is more likely if:
   - random sampling/assignment is used, and
   - if sampling without replacement, n < 10% of the population.
2. ***Sample size/skew:*** Either the population distribution is normal, or if the population distribution is skewed, the sample size is large.
   - the more skewed the population distribution, the larger sample size we need for the CLT to apply
   - for moderately skewed distributions $n > 30$ is a widely used rule of thumb

   This is also difficult to verify for the population, but we can check it using the sample data, and assume that the sample mirrors the population.

CENTRAL LIMIT THEOREM FOR THE SINGLE PROPORTION

- Sample proportions will be nearly normally distributed with mean equal to the population mean, p, and standard error equal to $\sqrt{\frac{p(1-p)}{n}}$

$$\hat{p} \sim N\left(mean = p, SE = \sqrt{\frac{p(1-p)}{n}}\right)$$

The following conditions need to be met for the CLT to apply:

1. ***Independence:*** Sampled observations must be independent. This is difficult to verify, but is more likely if:
   - random sampling/assignment is used, and
   - if sampling without replacement, n < 10% of the population.
2. ***Sample size/skew:*** At least 10 success and 10 failure observations

CONFIDENCE INTERVALS – GENERAL FORMULA

- By reporting a single point estimate, like the ***sample mean ($\bar{x}$)***, it is very likely that we will not capture the exact population parameter, e.g. the true ***population mean (μ)***
- On the other hand, if we report a range of the plausible values we have a good chance to capture a true population parameter.
- A plausible range of values for the population parameter is called a **confidence interval**.
- Confidence intervals may be constructed in different ways, depending on the type of statistic and therefore the shape of the corresponding sampling distribution. For symmetrically distributed sample statistics, like those involving means and proportions, a general formula of confidence interval is:

$$ci = [l, u] = [sample\ statistic - z^* \times SE, sample\ statistic + z^* \times SE]$$

where:

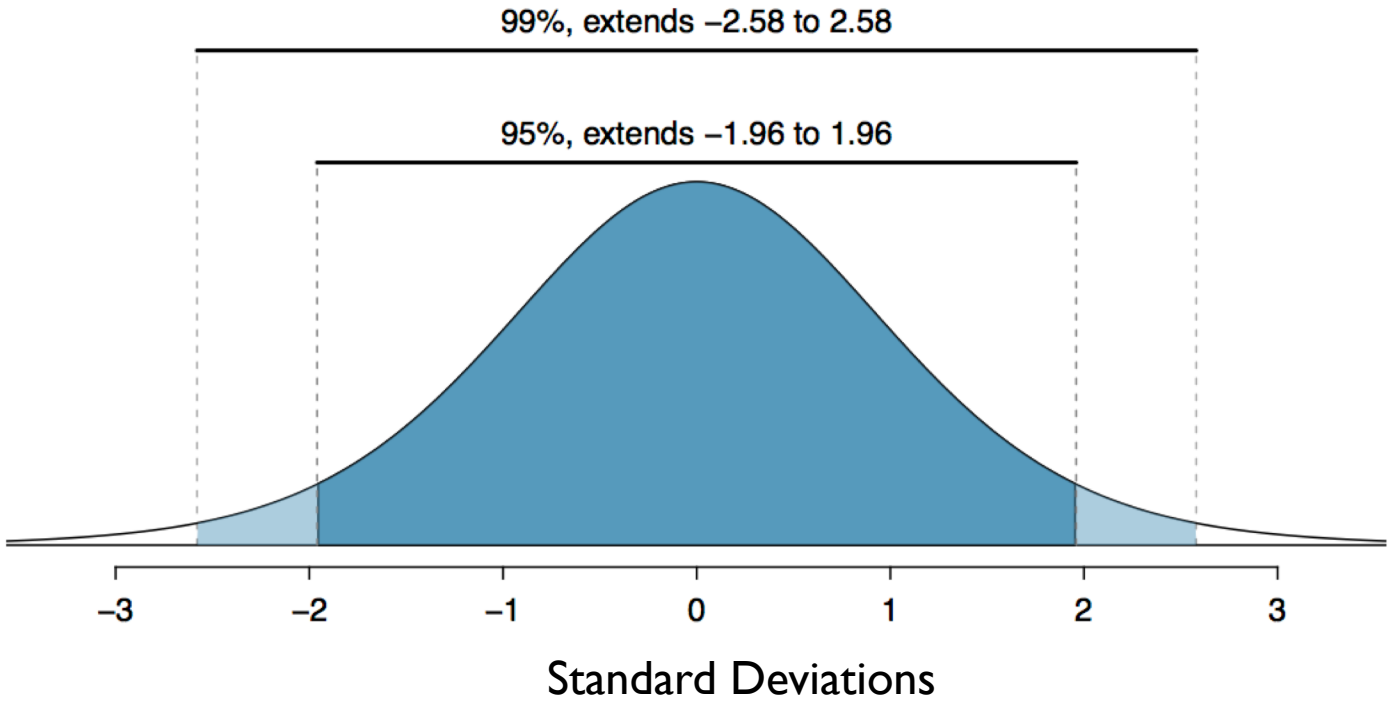
- z^* is critical value and can have a different values depending on the confidence level.
- SE is the standard error of an estimate

CONFIDENCE INTERVALS

$$ci = [l, u] = [sample\ statistic - z^* \times SE, sample\ statistic + z^* \times SE$$

- In a confidence interval, $z^* \times SE$ is called the ***margin of error***, and for a given sample the margin of error changes as the confidence level changes.
- In order to change the confidence level we need to adjust z^* in the above formula.
- Two most commonly used confidence intervals in practice are 95% and 99% confidence intervals
  - For 95% confidence interval $z^* = 1.96$
  - For 99% confidence interval $z^* = 2.58$
  - To find z^* value for other confidence intervals use standard normal Z-table

CONFIDENCE INTERVALS

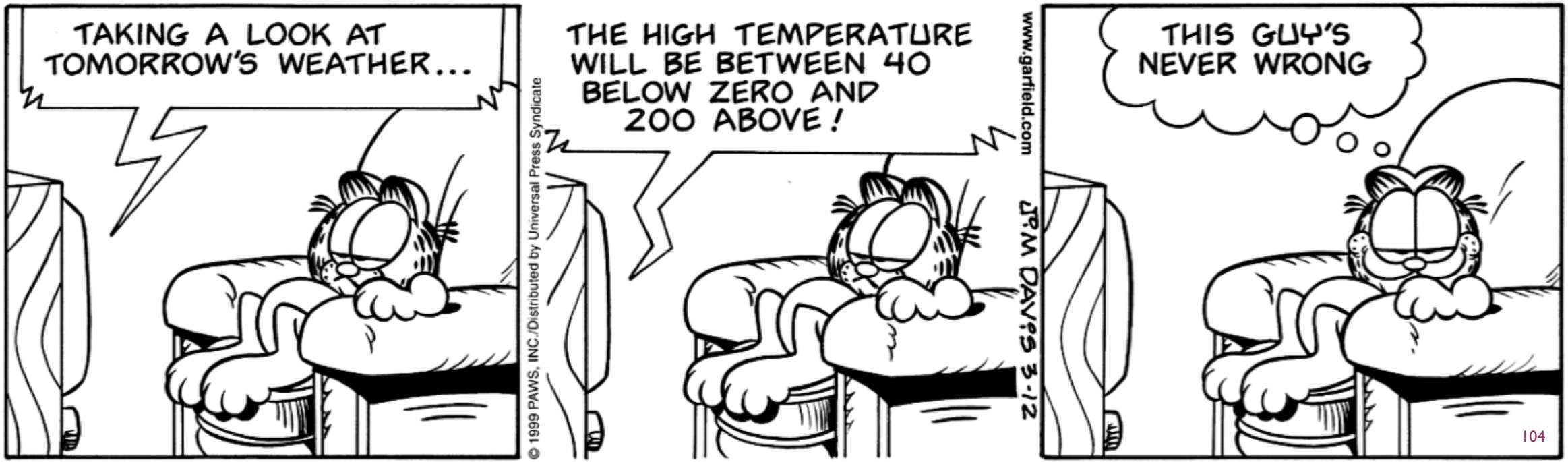


An illustration of 95% and 99% confidence intervals

CONFIDENCE INTERVALS

A wider interval.

- **Q:** *If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, is it better to use a smaller or wider confidence interval?*
- **A:** *A wider confidence interval. However be careful as too wide a confidence intervals may not be very informative*



INTERPRETING CONFIDENCE INTERVAL

- ***We are XY% confident that the true population parameter is between the lower bound (l) and the upper bound (u) of our confidence interval***
  - It is incorrect to interpret the confidence interval as capturing the population parameter with a certain probability
- ***Confidence intervals only try to capture the population parameter***
  - Confidence intervals say nothing about the confidence of capturing individual observations, a proportion of observations or about capturing point estimates

SIGNIFICANCE TESTS

The five steps in statistical inference are:

1. Formulation of the practical problem in terms of statistical hypotheses
2. Construction of a test statistic
3. Description of a critical region and/or the calculation of the *p*-value
4. Significance level or size of the test
5. Further assessment

1. FORMULATION OF THE PRACTICAL PROBLEM IN TERMS OF STATISTICAL HYPOTHESES

- ***Null hypothesis*** (H_0)
  - Represents what we currently hold as true
  - ***In a one sample situation*** it often expresses the maintenance of the status quo, i.e. no difference from our' previous knowledge.
  - ***In a two-sample situation*** it is usually the hypothesis of no difference between the populations being compared.
  - H_0 needs to be relatively simple in form (see 2. Construction of a test statistic)
- ***Alternative Hypothesis*** (H_A)
  - Represents what we want to test
  - It expresses the range of situations that we wish the test to be able to diagnose. Depending upon the outcome of the test we may take action.

I. FORMULATION OF THE PRACTICAL PROBLEM IN TERMS OF STATISTICAL HYPOTHESES

NOTE:

- H_0 is basically a standard with which the evidence for H_A can be compared.
- If the test provides a result which ***appears unlikely assuming H_0 to be true*** and ***is reasonable under the assumption of H_A*** our conclusion is that :

  "***H_0 is rejected in favour of H_A***".
- If the test result supports H_0 rather than H_A we do not say:
  "***H_0*** is accepted as true" or "***H_0*** is accepted";

  Instead, we should say:

  "***There is no evidence to reject H_0 in favour of H_A***"

2. CONSTRUCTION OF A TEST STATISTIC

- The majority of tests involve the calculation of a **test statistic** t – *a function of the data plus the information in the hypothesis* H_0
- ***A test statistic should satisfy two properties:***
  1. Its probability distribution must be calculable (at least approximately) under assumption that H_0 is true
  2. it should behave differently when H_0 is true from when H_A is true
- We can often construct several different test statistics for testing the same H_0 and H_A.

3. DESCRIPTION OF CRITICAL REGION AND CALCULATION OF THE P-VALUE

CRITICAL REGION

- A region of values of the test statistic t which support our preference for H_A rather than H_0 is called a ***critical region***
  - If our calculated value of t (calculated under the assumption that H_0 is true) falls in a suitable critical region ***we reject H_0 in favour of H_A***
  - Otherwise, ***we are unable to reject H_0 in favour of H_A***
- Tests are constructed so that the lack information, particularly too little data, tends to result in non-critical values of the test statistic.
  - ***Hence, it is UNWISE TO TALK POSITIVELY ABOUT "ACCEPTING H_0"***

NOTE: *Lack of strong evidence to reject H_0 in favour of H_A may indicate that we have not collected enough data to reject it*

3. DESCRIPTION OF CRITICAL REGION AND CALCULATION OF THE P-VALUE

P-VALUE

- The p-value quantifies the strength of the evidence against the null hypothesis H_0 and in favour of the alternative hypothesis H_A
- We usually use a summary statistic of the data, like the sample mean, to help compute the p-value and evaluate the hypotheses.
- A computed "small" p-value would result if either:
  - H_0 is true and an improbable event has occurred, or
  - H_A is true
- ***In practice:***
  - If p-value is small, H_0 is rejected in favour of H_A
  - If p-value is not "small", the evidence does not support the rejection of H_0 in favour of H_A

4. SIGNIFICANCE LEVEL OR THE SIZE OF THE TEST

- Whichever the approach used in the previous step (very small p-values or calculated t-statics falls outside critical region), it is possible to reject H_0 when in fact it is true.
- In making a test of H_0 against H_A we can make two kinds of error
  - ***Type I Error*** - H_0 is rejected when in fact it is true
  - ***Type II Error*** - H_0 is not rejected when in fact it is false

	Do not reject H_0	Reject H_0
H_0 is true	✓	***Type I Error***
H_0 is false	***Type II Error***	✓

4. SIGNIFICANCE LEVEL OR THE SIZE OF THE TEST

Type I Error (False Positive)

Type II Error (False Negative)

Images taken from: http://www.statisticssolutions.com/to-err-is-human-what-are-type-i-and-ii-errors

4. SIGNIFICANCE LEVEL OR THE SIZE OF THE TEST

- ***Significance level*** $\boldsymbol{\alpha}$ is the probability of rejecting $\boldsymbol{H_0}$ when in fact it is true, i.e. probability of committing **Type I Error**
- The choice of the value $\boldsymbol{\alpha}$ depends on the particular problem and how serious it is if a true $\boldsymbol{H_0}$ is rejected.
  - Usually $\boldsymbol{\alpha}$ lies between 0.01 and 0.10, and is most often 0.05
  - ***Example:*** *if we choose significance level* $\boldsymbol{\alpha = 0.05}$ *this means that we will allow 5 incorrect rejections of* $\boldsymbol{H_0}$ *from every 100 we make.*
- The significance level $\boldsymbol{\alpha}$ should be chosen prior to the experiment or at least before the analysis of the data.
- We often express the significance level and p-value as percentages
  - ***Example:*** *For significance level of 5% (*$\alpha = 0.05$*) there is 5% chance that the result is due to chance*

5. FURTHER ASSESSMENT

- If we present our conclusion simply as "reject H_0 in favour of H_A at a given significance level" or "not reject H_0 in favour of H_A at a given significance level', it is often wasteful of the information in the data.
- We can use the p-value to provide additional assessment to the above statements.
- For example, working with a significance level of 5%, the basic conclusions are:

IF: $p \leq 5\%, \ (p \leq 0.05)$	The **test is significant at 5% level** and H_0 **is rejected** in favour of H_A
IF: $p > 5\%, \ (p > 0.05)$	The **test is not significant at 5% level** and H_0 **is not rejected** in favour of H_A
FURTHER ASSESSMENT CAN BE MADE ALONG THE FOLLOWING LINES	
$p > 10\%$	There is no (or very little) evidence for rejecting H_0 in favour of H_A
$5\% < p \leq 10\%$	On the available evidence we cannot reject H_0 in favour of H_A but we have some suspicion ***(i.e. we would like to obtain more evidence)***
$1\% < p \leq 5\%$	Significant at 5% level and H_0 is rejected in favour of H_A. If the decision to change is important, we should probably seek further evidence
$0.1\% < p \leq 1\%$	Highly significant at 5% level. There is considerable evidence for rejection H_0 in favour H_A.
$p \leq 0.1\%$	Very highly significant at 5% level. We are very confident that H_A is to be preffered to H_0.

TEST OF SINGLE MEAN μ USING THE Z STATISTIC

- The Z statistic can be used to hypothesize about the population mean (μ) given a sample of data from the population.
- Suppose we have a random sample of size n from the normal distribution N(μ, σ^2) where μ is unknown but the variance σ^2 is known. We wish to test
  - H_0: μ=μ_0 (specified) against
  - H_A: $\mu < or \neq or > \mu_0$ at a given significance level α.
- The test statistic is:

$$Z = \frac{\bar{x} - \mu}{SE} = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

- A confidence interval for the single mean is calculated as follows:

$$ci = [\bar{x} - z^* \times \sqrt{\frac{\sigma^2}{n}}, \bar{x} + z^* \times \sqrt{\frac{\sigma^2}{n}}]$$

TEST OF THE COMPARISON OF TWO MEANS USING THE Z STATISTIC

- Suppose we have random samples of size n_1 and n_2 from independent normal populations with unknown means μ_1 and μ_2 and known variances σ_1^2 and σ_2^2.
- We wish to test
  - H_0:$\mu_1 - \mu_2 = \theta_0$, a specified value (mostly θ_0=0, that is μ_1=μ_2)

  against
  - H_A: $\mu_1 - \mu_2 \neq \theta_0$, or a one-sided alternative
  - H_A: $\mu_1 > \mu_2$; H_A: $\mu_1 < \mu_2$ at a given significance level α.
- The test statistic is:

$$Z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1).$$

- A confidence interval for the comparison of two means is calculated as follows:

$$ci = [(\bar{x}_1 - \bar{x}_2) - z^* \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\bar{x}_1 - \bar{x}_2) + z^* \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}]$$

TEST OF THE COMPARISON OF TWO MEANS USING THE Z STATISTIC - EXAMPLE

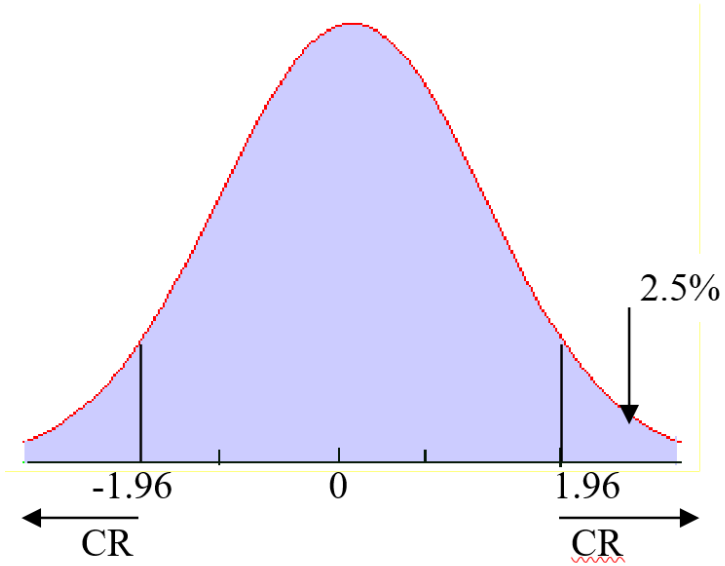
- Suppose we wish to determine if there is a difference in mean weight between the two sexes in a particular bird species. The following data were obtained:
  - Male sample size n_1=125, mean weight $\overline{x_1} = 92.31g, s_1^2 = 56.22g^2$.
  - Female sample size n_2=85, mean weight $\overline{x_2} = 88.84g, s_2^2 = 65.41g^2$s

  - Test $H_0: \mu_1 = \mu_2$ i.e. $\mu_1 - \mu_2 = 0$
  - Against $H_A: \mu_1 \neq \mu_2$ i.e. $\mu_1 - \mu_2 \neq 0$

    at a 5% significance level. If significant, give a 95% c.i. for $\mu_1 - \mu_2$

  - Under H_0, test statistic: $$Z = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim \text{approx. } N(0,1)$$
  - Test statistic value: $$z = \frac{92.31 - 88.84}{\sqrt{\frac{56.22}{125} + \frac{65.41}{85}}} = 3.41.$$

TEST OF THE COMPARISON OF TWO MEANS USING THE Z STATISTIC - EXAMPLE



- Test is highly significant at 5% level since $P(|Z| > 3.14) < 0.01$. Hence we are confident that the mean weights are different, in particular with male mean weight greater than female mean weight.

  How different?

TEST OF THE COMPARISON OF TWO MEANS USING THE Z STATISTIC - EXAMPLE

- From: $P(-1.96 \leq \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \leq 1.96) \approx 0.95$

- Approximate 95% confidence limits for $\mu_1 - \mu_2$ and $(\bar{x}_1 - \bar{x}_2) \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

- That is, approximately 95% confidence interval is [1.31, 5.63].
- Note Since $\mu_1 - \mu_2 = 0$ does <u>not</u> lie in this interval, the test is significant at 5% level.

TESTS OF PROPORTIONS(S)

- **Single Proportion**
  - Suppose we have a random sample of n units from a large population, a proportion p (unknown) of which possess a certain attribute. Let ***x*** units in the sample possess the attribute.
  - Then the sample proportion $\hat{p} = \frac{x}{n}$ estimates p with test statistic:

$$Z = \frac{\frac{x}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \text{approx. } N(0,1).$$

  - A confidence interval for single proportion: $ci = [\frac{x}{n} - z^* \times \sqrt{\frac{p(1-p)}{n}}, \frac{x}{n} + z^* \times \sqrt{\frac{p(1-p)}{n}}]$
- **Comparison of Two Proportions**
  - Suppose we have two large samples of sizes of n_1 and n_2 from two populations where proportions p_1 and p_2 respectively have an attribute.
  - The test statistic is:

$$Z = \frac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim \text{approx. } N(0,1).$$ ***NOTE***: $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$

  - A confidence interval for two proportions: $ci = [(\frac{X_1}{n_1} - \frac{X_2}{n_2}) - z^* \times \sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}, (\frac{X_1}{n_1} - \frac{X_2}{n_2}) + z^* \times \sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}]$
  - ***NOTE***: $\widehat{p_1} = \frac{X_1}{n_1}, \widehat{p_2} = \frac{X_2}{n_2}$

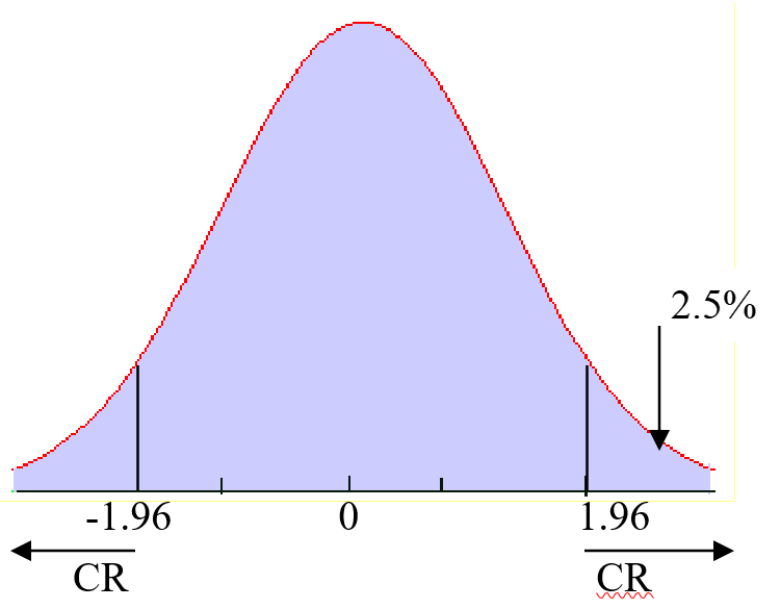
TEST OF SINGLE PROPORTION - EXAMPLE

- In a random sample of 120 graduates, 78 spent 3 years at university and 42 more than 3 years. Test the hypothesis that 70% of graduates obtain degrees in 3 years.

- Let p=P(graduate in 3 years) (unknown)
  - H_0: $p = 0.7$
  - H_1: $p \neq 0.7$

TEST OF SINGLE PROPORTION - EXAMPLE

- Sample proportion $\hat{p} = \frac{78}{120} = 0.65.$
- Test statistic $z = \frac{\frac{78}{120} - 0.7}{\sqrt{\frac{0.7 \times 0.3}{120}}} = -1.2.$



- Test not significant at 5% level. We have insufficient evidence for rejecting H_0.

TEST OF TWO PROPORTIONS - EXAMPLE

- We wish to compare the germination rates of spinach seeds for two different methods of preparation.
  - Method A 80 seeds sown, 65 germinate
  - Method B 90 seeds sown, 80 germinate.

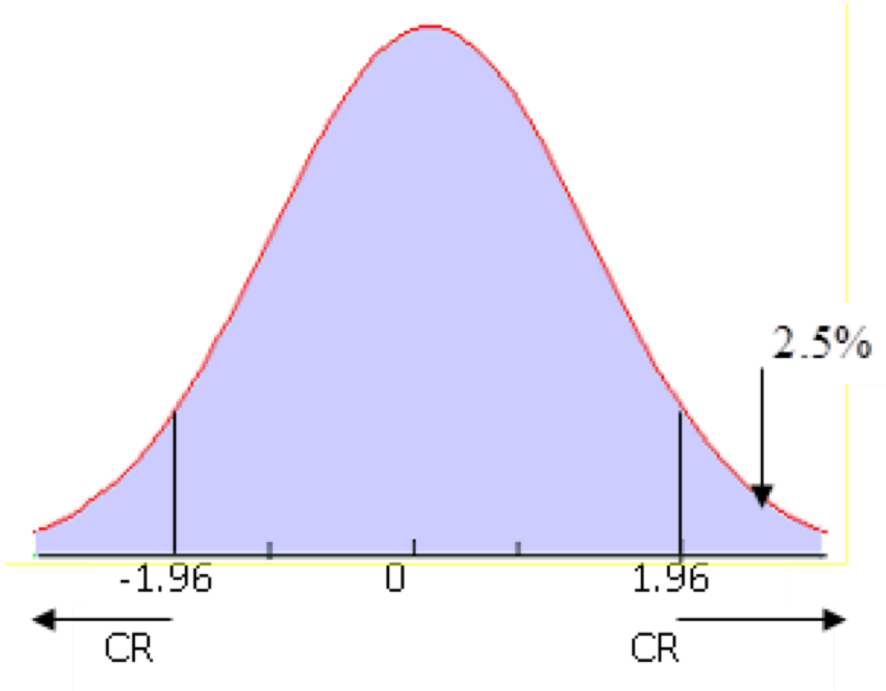
- Let proportions germinating be p_1 and p_2.
  - H_0: $p_1 = p_2 = p$ (unknown), against: H_1: $p_1 \neq p_2$.

- Estimate p by: $\hat{p} = \frac{65+80}{80+90} = 0.853.$

$$\hat{p}_1 = \frac{65}{80} = 0.8125, \quad \hat{p}_2 = \frac{80}{90} = 0.889.$$

TEST OF TWO PROPORTIONS - EXAMPLE

- Under H_0, $$z = \frac{0.8125 - 0.889}{\sqrt{(0.853)(0.147)\left(\frac{1}{80} + \frac{1}{90}\right)}} = -1.4$$

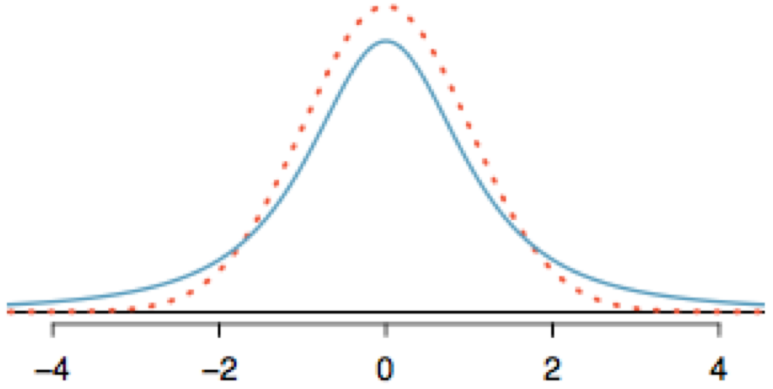


- Test not significant at 5% level. We have no evidence for supposing the germination rates to be different.
- (Hypotheses involving proportions can also be tested using the χ^2 distribution. See chapter χ^2 **Test**).

TESTS BASED ON THE (STUDENT'S) T-DISTRIBUTION

The T Distribution

- When the population standard deviation (σ) is unknown (which happens almost always), and when we have a small data sample ($n < 30$) the uncertainty of the standard error estimate is addressed by using a new distribution: ***the t distribution***
- This distribution also has a bell shape, but its tails are thicker than the normal model's.
  - Therefore observations are more likely to fall beyond two SDs from the mean than under the normal distribution.
- These extra thick tails are helpful for resolving our problem with a less reliable estimate the standard error (since *n* is small)

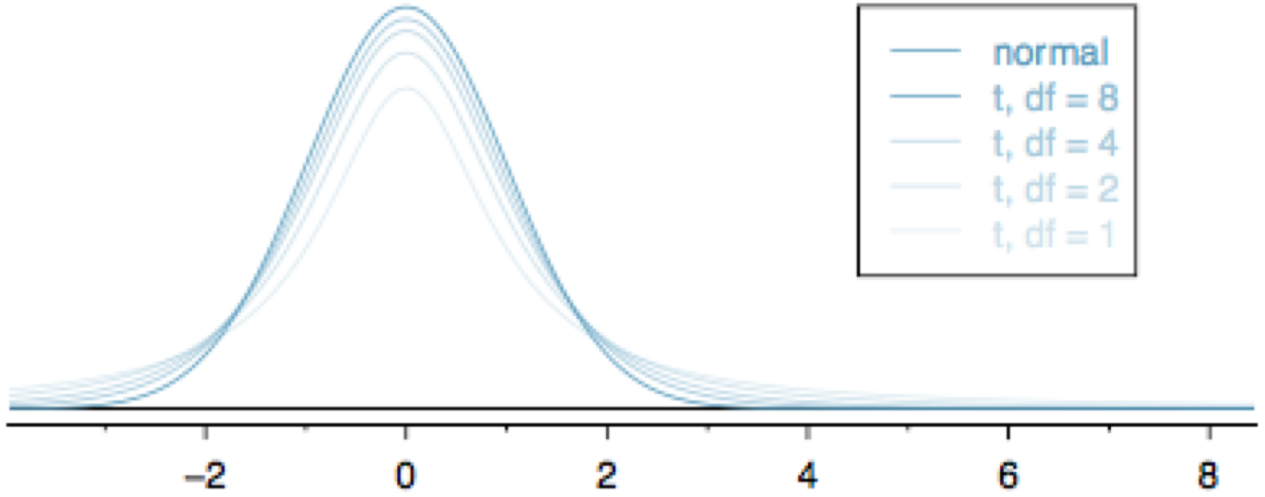


Comparison of a t-distribution (solid blue line) and a normal distribution (dotted red line)

TESTS BASED ON THE (STUDENT'S) T-DISTRIBUTION

The T Distribution

- Always centered at zero, like the standard normal distribution, and has a single parameter ***degrees of freedom*** (***df***) which describes the precise form of the bell-shaped t-distribution.



NOTE: *The larger the degrees of freedom, the more closely the t - distribution resembles the standard normal model. When* $df \geq 30$, the t-distribution is nearly indistinguishable from the normal distribution.

TESTS BASED ON THE (STUDENT'S) T-DISTRIBUTION

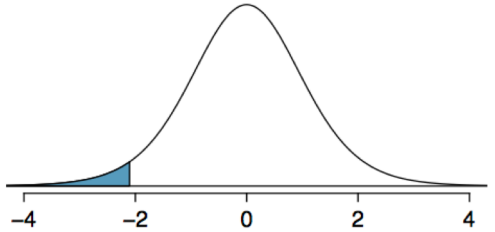
- **The T - Table**
- When we perform tests based on the t-distribution, we use a ***t-table*** in place of the normal probability table.
  - Each row in the t-table represents a t-distribution with different degrees of freedom.
  - The columns correspond to tail probabilities.

one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	⋮	⋮	⋮	⋮	⋮	

An abbreviated look at the t-table

TESTS BASED ON THE (STUDENT'S) T-DISTRIBUTION

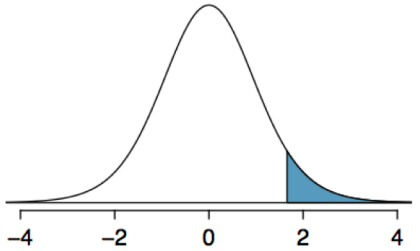
- **The T – Table (Example 1)**
- What proportion of the t-distribution with 18 degrees of freedom falls below -2.10?
  - To find this area, we identify the appropriate row: df = 18.
  - Then we identify the column containing the absolute value of -2.10; it is the third column.
  - Because we are looking for just one tail, we examine the top line of the table, which shows that a one tail area for a value in the third row corresponds to 0.025. Therefore, about 2.5% of the distribution falls below -2.10



one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	⋮	⋮	⋮	⋮	⋮	
	17	1.33	1.74	2.11	2.57	2.90
	18	**1.33**	**1.73**	**2.10**	**2.55**	**2.88**
	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85

TESTS BASED ON THE (STUDENT'S) T-DISTRIBUTION

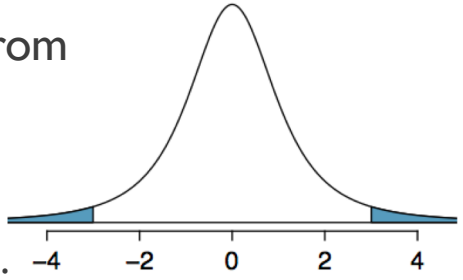
- **The T – Table (Example 2)**
- What proportion of the t-distribution with 20 degrees of freedom falls above 1.65?
  - We identify the row in the t-table using the degrees of freedom: df = 20.
  - Then we look for 1.65; it is not listed. It falls between the first and second columns (1.33<1.65.<1.72). Since these values bound 1.65, their tail areas will bound the tail area corresponding to 1.65.
  - We identify the one tail area of the first and second columns, 0.050 and 0.10, and we conclude that between 5% and 10% of the distribution is more than 1.65 standard deviations above the mean.



one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05
13	1.35	1.77	2.16	2.65	3.01
14	1.35	1.76	2.14	2.62	2.98
15	1.34	1.75	2.13	2.60	2.95
16	1.34	1.75	2.12	2.58	2.92
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85

TESTS BASED ON THE (STUDENT'S) T-DISTRIBUTION

- **The T – Table (Example 2)**
- What proportion of the t-distribution with 2 degrees of freedom falls 3 standard deviations from the mean (above or below)?
  - We identify the row in the t-table using the degrees of freedom: df = 20
  - Next, find the columns that capture 3; because 2.92 < 3 < 4.30, we use the second and third columns.
  - Finally, we find bounds for the tail areas by looking at the two tail values: 0.05 and 0.10. We use the two tail values because we are looking for two (symmetric) tails.



	one tail	0.100	0.050	0.025	0.010	0.005
	two tails	0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	⋮	⋮	⋮	⋮	⋮	⋮

TESTS BASED ON THE (STUDENT'S) T-DISTRIBUTION

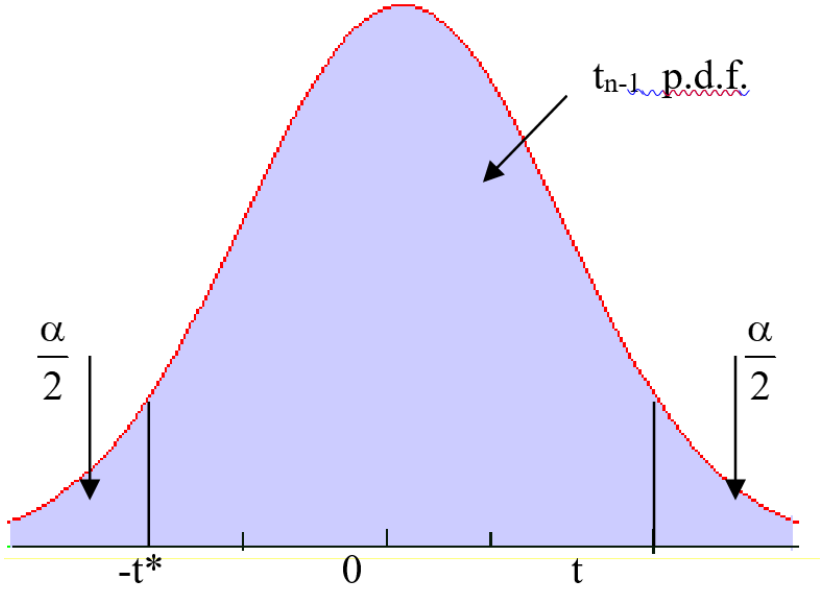
Single Mean μ

- Suppose we have a random sample of size n (n < 30) from N(μ, σ^2), where μ and σ^2 are both unknown. We wish to test:
  - $H_0: \mu = \mu_0$ (specified)
  - $H_A: \mu \neq \mu_0$ (or a one-sided alternative).
- The test statistic $t = \frac{\bar{x}-\mu_0}{\sqrt{\frac{s^2}{n}}}$ is an observation from t_{n-1}.
- A confidence interval for the single mean is calculated as follows:

$$ci = [\bar{x} - t^*_{n-1} \times \sqrt{\frac{s^2}{n}}, \bar{x} + t^*_{n-1} \times \sqrt{\frac{s^2}{n}}]$$

T-TEST (SINGLE MEAN) – EXAMPLE

- The temperature of warm water springs in a basin is reported to have a mean of 38°C. A sample of 12 springs from the west end of the basin had mean temperature 39.4 and variance 1.92. Have springs at the west end a different mean temperature? Give a 95% c.i. for the mean temperature.

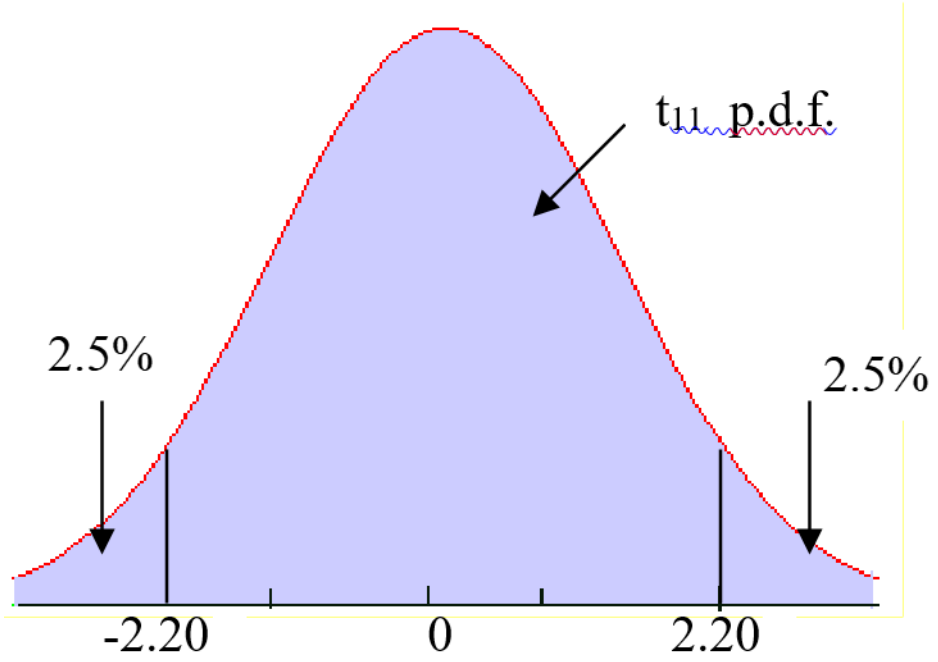


- Denote west end spring temperature by X where X has mean μ and variance σ^2. We estimate σ^2 by $s^2=1.92$ with 11 degrees of freedom.

$H_0: \mu = 38$ $\quad$ $H_1: \mu \neq 38$

T-TEST (SINGLE MEAN) – EXAMPLE

- Under H_0, $t = \frac{39.4-38}{\sqrt{\frac{1.92}{12}}} = 3.5$ is an observation from t_{11}.



- Upper 2.5% point of t_{11} is 2.20. Test is significant at 5% level and we conclude that west springs do have a different temperature.

T-TEST (SINGLE MEAN) – EXAMPLE

- (Alternatively (using the t-Probability table, df=11, two tails test), p-value=P(|*T*|> 3.5)<0.01 or less than 1%).

$$P\left(-2.201 \le \frac{\overline{X}-\mu}{\sqrt{\frac{1.92}{12}}} \le 2.201\right) = 0.95$$

$$P(\overline{X} - 2.201 \times 0.4 \le \mu \le X + 2.201 \times 0.4) = 0.95$$

- Therefore, 95% c.i. for is $\text{ci} = [39.4 - 2.20 \times 0.4, 39.4 + 2.20 \times 0.4] = [38.5, 40.3]$

TESTS BASED ON THE (STUDENT'S) T-DISTRIBUTION

Paired Comparison Test

- In this case we are interested in the difference between two methods or properties where the observations occur naturally in pairs and taking the difference of the paired observations is valid.
  - ***It is not possible to pair arbitrarily.***
- We wish to test:
  - $H_0: \mu_{diff} = \theta_0$ (most often $\theta_0 = 0$)
  - $H_A: \mu_{diff} \neq \theta_0$ (or a one-sided alternative).
- The test statistic $t = \frac{\bar{x}_{diff} - \theta_0}{\sqrt{\frac{s^2_{diff}}{n}}}$ is an observation from t_{n-1}.
- A confidence interval for the paired differences is calculated as follows:

$$ci = [\bar{x}_{diff} - t^*_{n-1} \times \sqrt{\frac{s^2_{diff}}{n}}, \bar{x}_{diff} + t^*_{n-1} \times \sqrt{\frac{s^2_{diff}}{n}}]$$

T-TEST (PAIRED COMPARISON TEST) - EXAMPLE

- Consider an experiment to compare the effects of two sleeping drugs A and B. There are 10 subjects and each subject receives treatment with each of the two drugs (the order of treatment being randomised). The number of hours slept by each subject is recorded. Is there any difference between the effects of the two drugs?

Subject	1	2	3	4	5	6	7	8	9	10
Hours slept using A	9.9	8.8	9.1	8.1	7.9	12.4	13.5	9.6	12.6	11.4
Hours slept using B	8.7	6.4	7.8	6.8	7.9	11.4	11.7	8.8	8.0	10.0
Difference (A-B) x	1.2	2.4	1.3	1.3	0.0	1.0	1.8	0.8	4.6	1.4

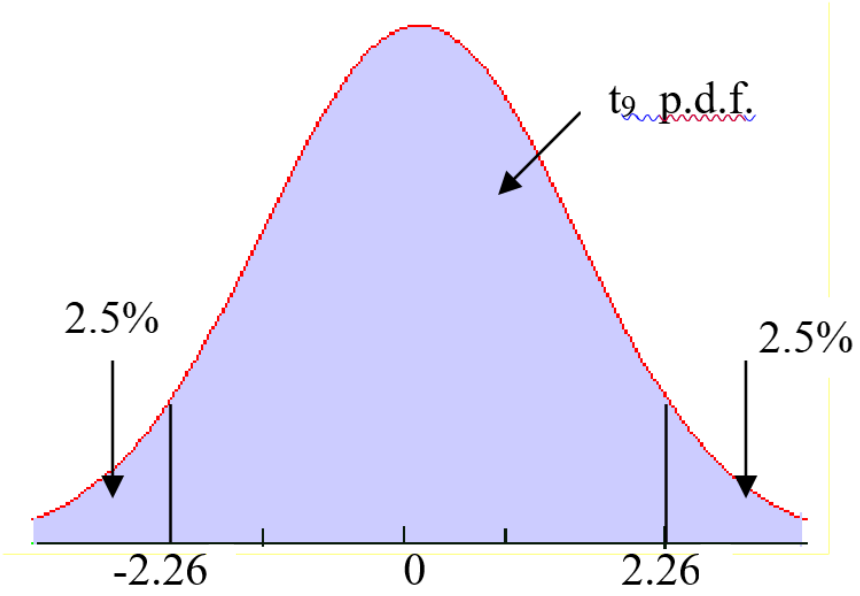
- The paired sample data have been reduced to a single sample of differences. This will tend to cancel out any subject effect assuming that the effect of the drug is additive.
- Assume x values to be normally distributed with mean μ.

$$\sum x_i = 15.8, \quad \bar{x} = 1.58, \quad \sum x_i^2 = 38.58, \quad s^2 = 1.513.$$

$H_0: \mu = 0$ $H_1: \mu \neq 0$

T-TEST (PAIRED COMPARISON TEST) - EXAMPLE

- Under H_0 , $t = \frac{1.58-0}{\sqrt{\frac{1.513}{10}}} = 4.06$ is an observation from the t distribution with 9 degrees of freedom.



- From t - Table we have P(|*T*| > 2.26)=0.05.
- The test is significant at a 5% level.

T-TEST (PAIRED COMPARISON TEST) - EXAMPLE

- (Alternatively (using the t-Probability table, df=9, two tails test), p-value = P(|*T*| > 4.06) < 0.01 or less than 1%).
- As we can see, the test is highly significant at the 5% level. We are thus confident that there is a difference between the drugs, in particular that drug A induces more sleep than drug B on average.

- A 95% confidence interval for the unknown mean difference μ is:

$$\left[\bar{x} - 2.262\sqrt{\frac{1.513}{10}},\ \ \bar{x} + 2.262\sqrt{\frac{1.513}{10}}\right]$$

- That is, [0.70, 2.46].

TESTS BASED ON THE (STUDENT'S) T-DISTRIBUTION

Comparison of Two Means – independent samples

- Random sample of size n_1 (n_1 small, <30) with sample mean $\overline{x_1}$, sample variance s_1^2 from a normal or approximately normal distribution with unknown mean μ_1 and unknown variance σ^2.
  - Random sample of size n_2(n_2 small, <30): $\overline{x_2}$, s_2^2, μ_2 and σ_2.
  - Note: The unknown population variances are equal.
- We wish to test:
  - $H_0: \mu_1 - \mu_2 = \theta_0$ (specified) (most often $\theta_0 = 0$)
  - $H_A: \mu_1 - \mu_2 \neq \theta_0$ at a given level. (or a one-sided alternative).
- The test statistic $t = \frac{(\overline{x_1}-\overline{x_2})-\theta_0}{\sqrt{s_{pooled}^2(\frac{1}{n_1}+\frac{1}{n_2})}}$ is an observation from $t_{n_1+n_2-2}$, where $s_{pooled}^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$
- A confidence interval for the paired differences is calculated as follows:

$$ci = [(\overline{x_1} - \overline{x_2}) - t^*_{n_1+n_2-2} \times \sqrt{s_{pooled}^2(\frac{1}{n_1}+\frac{1}{n_2})}, (\overline{x_1} - \overline{x_2}) + t^*_{n_1+n_2-2} \times \sqrt{s_{pooled}^2(\frac{1}{n_1}+\frac{1}{n_2})}]$$

T-TEST INDEPENDENT SAMPLES - EXAMPLE

- Two methods of oxidation care are used in an industrial process. Repeated measurements of the oxidation time are made to test the hypothesis that the oxidation time of method 1 is different than that of method 2 on average.

	Sample size	Sample mean	Sample variance
Method 1	9	41.3	20.7
Method 2	8	48.9	34.2

- We wish to test:

$$H_0: \mu_1 = \mu_2 \text{ that is, } \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 \neq \mu_2 \text{ that is, } \mu_1 - \mu_2 \neq 0$$

- We shall assume that the unknown population variances are equal. (*NOTE: This can be tested using an F -Test*)

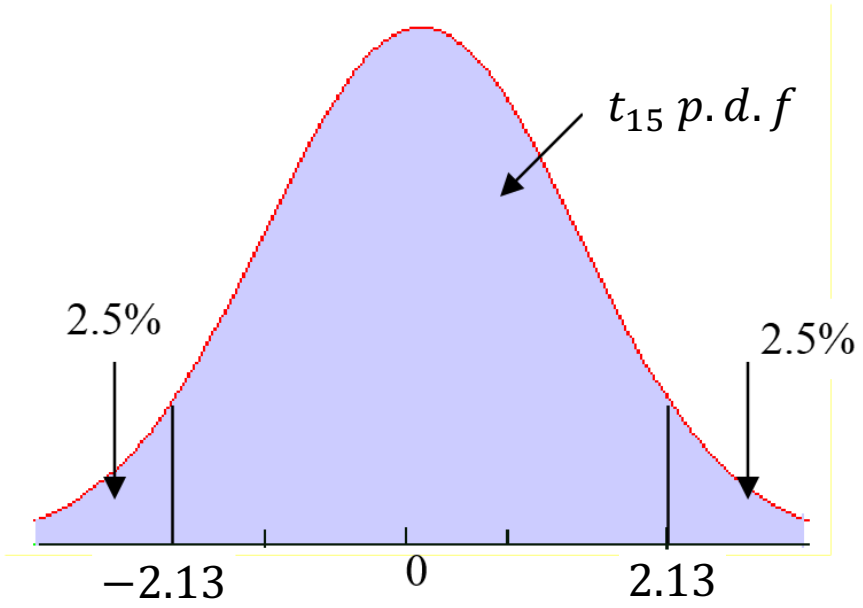
$$df = n_1 + n_2 - 2 = 9 + 8 - 2 = 15$$

$$s^2 = \frac{8(20.7) + 7(34.2)}{8+7} = 27 \qquad \text{wtith 15 df.}$$

T-TEST INDEPENDENT SAMPLES - EXAMPLE

- Under H_0, $t = \frac{41.3-48.9}{\sqrt{27\left(\frac{1}{9}+\frac{1}{8}\right)}} = -3.01$ is also an observation from t_{15}.

- Test significant at 5% level.



T-TEST INDEPENDENT SAMPLES - EXAMPLE

- Since p-value = P(|*T*|>3.01) < 0.01, test is highly significant at 5% level.

  We are confident that the oxidation time of method 1 is different than the oxidation time of method 2 on average.

- 95% c.i. for $(\mu_1 - \mu_2)$ is:

$$ci = [(41.3 - 48.9) - 2.13 \times \sqrt{27\left(\frac{1}{9} + \frac{1}{8}\right)}, (41.3 - 48.9) + 2.13 \times \sqrt{27\left(\frac{1}{9} + \frac{1}{8}\right)}]$$

- So 95% c.i. for $(\mu_1 - \mu_2)$ is:

$$ci = [-12.965, -2.235]$$

TESTS BASED ON THE χ^2 DISTRIBUTION

Goodness-of-fit Test for Classified Data

- Suppose that a sample of n observations is classified into k mutually exclusive and exhaustive classes, that is, each observation belongs to one and only one class.
- Let O_i be the observed frequency in the i^{th} class, $\sum_{i=1}^{k} O_i = n$.
- Consider a null hypothesis H_0 which specifies the probabilities of belonging to the k classes. Under H_0, let E_i be the expected frequency in the i^{th} class, $\sum_{i=1}^{k} E_i = n$.
- Under H_0, the goodness-of-fit test statistic

$$\phi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

is approximately distributed χ^2_{df} where $df = k - 1$ is number of independent parameters estimated from the data, i.e. degrees of freedom.

TESTS BASED ON THE χ^2 DISTRIBUTION

Goodness-of-fit Test for Classified Data

- The critical region lies in the right hand tail only of the χ^2 distribution since if H_0 is not true we would expect the E_i's to be quite different from the O_i's resulting in a larger than expected value of ϕ^2. (Small ϕ^2 results when E_i's and O_i's are in good agreement - certainly not a reason to reject H_0).

NOTE:

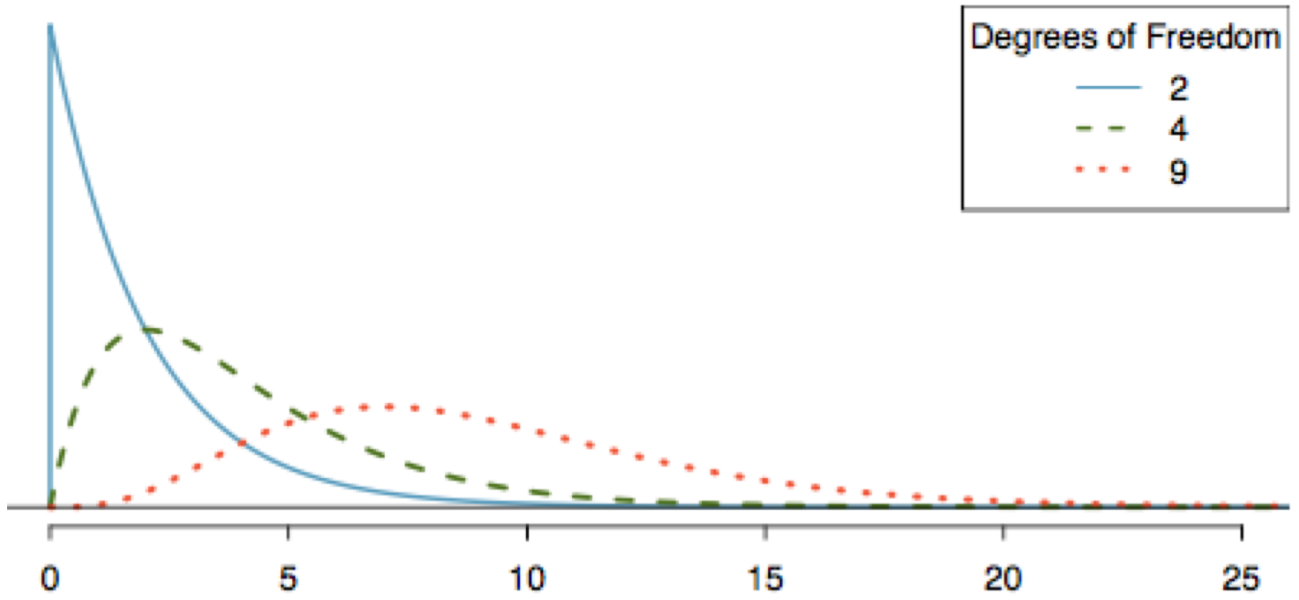
1. ***The exact distribution of ϕ^2 is discrete and is approximated by the continuous χ^2 distribution. For this approximation to be reasonable, E_i should be > 5 for each class. If not, combine adjacent classes with resultant loss of one or more degrees of freedom.***
2. ***In tests with only 1 degree of freedom, a better approximation is obtained by including Yates' continuity correction:***

$$\phi^2 = \sum_{i=1}^{k} \frac{(|O_i - E_i| - \frac{1}{2})^2}{E_i} \sim approx\ \chi_1^2$$

TESTS BASED ON THE χ^2 DISTRIBUTION

The χ^2 Distribution

- The χ^2 distribution has just one parameter called ***degrees of freedom*** (***df***), which influences the shape, center, and spread of the distribution.



An example of three χ^2 distributions with varying degrees of freedom

TESTS BASED ON THE χ^2 DISTRIBUTION

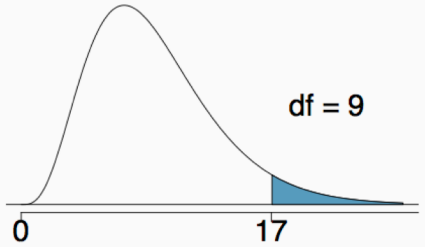
The χ^2 Table

- Our principal interest in the chi-square distribution is the calculation of p-values, which (as we have seen before) is related to finding the relevant area in the tail of a distribution.
- To do so, we use the chi-square table, which is very similar to the t-table:
  - We examine a particular row for distributions with different degrees of freedom, and we identify a range for the area.
  - One important difference from the t-table is that the chi-square table only provides upper tail values.

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46

TESTS BASED ON THE χ^2 DISTRIBUTION

The χ^2 Table (Example 1)



- What proportion of the χ^2 -distribution with 9 degrees of freedom falls above 17?
  - The cut off 17 falls between the fourth and fifth columns in the 9 degrees of freedom row.
  - Because these columns correspond to tail areas of 0.05 and 0.02, we can be certain that between 2% and 5% of χ^2_9 distribution falls above 17.

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32
	8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12
	9	10.66	12.24	14.68	16.92	19.68	21.67	23.59	27.88
	10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	29.59
	11	12.90	14.63	17.28	19.68	22.62	24.72	26.76	31.26

χ^2 TEST - EXAMPLE

- The geneticist Mendel evolved the theory that for a certain type of pea, the characteristics Round and Yellow, R and Green, Angular and Y, A and G occurred in the ratio 9:3:3:1. He classified 556 seeds and the observed frequencies were 315, 108, 101 and 32. Test Mendel's theory on the basis of these data.

$$H_0: \quad p_1 = \frac{9}{16}, \quad p_2 = \frac{3}{16}, \quad p_3 = \frac{3}{16}, \quad p_4 = \frac{1}{16}.$$

$$H_1: \quad \text{probabilities not as in } H_0.$$

Seed	O_i	P_i	$E_i = 556\ P_i$	$\frac{(O_i - E_i)^2}{E_i}$
R +Y	315	$\frac{9}{16}$	312.75	0.016
R +G	108	$\frac{3}{16}$	104.25	0.135
A+Y	101	$\frac{3}{16}$	104.25	0.101
A+G	32	$\frac{1}{16}$	34.75	0.218
	556 ($= n$)		556.00	0.470 ($=\phi^2$)

Under $H_0, \phi^2 = 0.47$ is an observation from χ^2_3. The test is not significant at the 5% level, that is, no evidence for supporting the rejection of H_0. (That is, data in agreement with theory).

χ^2 TEST - EXAMPLE

- In a random sample of 120 graduates, 78 spent 3 years at University and 42 more than 3 years. Test hypothesis that 70% obtain degree in 3 years. (See 10.1.3.1 - test of proportion using the normal distribution test).

H_0: P(degree in 3 years) = p = 0.7

H_A: $p \neq 0.7$

	O_i	E_i
Degree in 3 years	78	84
More than 3 years	42	36
	120	120

- Degrees of freedom = 2 - 1 = 1. Therefore use ½ correction.

$$\phi^2 = \frac{(|78-84|-\frac{1}{2})^2}{84} + \frac{(|42-36|-\frac{1}{2})^2}{36} = 1.2$$

- Test not significant at 5% level. No evidence to support the rejection of H_0.

COMPARING MEANS WITH ANOVA

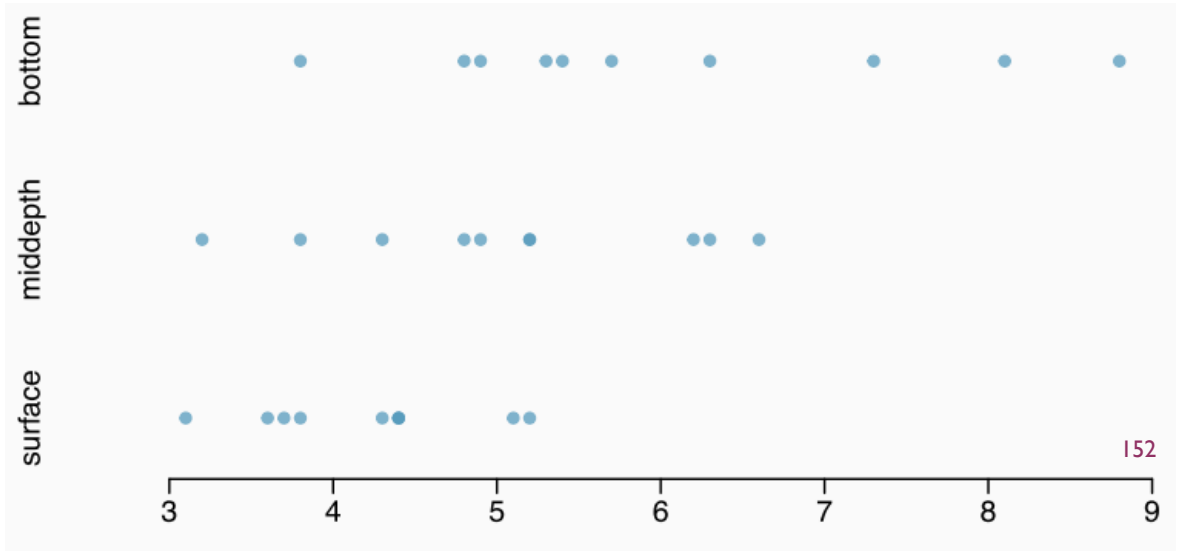
- The Wolf River in Tennessee flows past an abandoned site once used by the pesticide industry for dumping waste, including chlordane (pesticide), Aldrin and dieldrin (both insecticides).
- These highly toxic organic compounds can cause various cancers and birth defects.
- The standard methods to test whether these substances are present in a river is to take samples at six-tenths depth.
- But since these compounds are denser than water and their molecules tend to stick to particles of sediment, they are more likely to be found in higher concentrations near the bottom then near the middle.

DATA & RESEARCH QUESTION

- Is there a difference between the mean aldrin concentrations among the three levels? Data summary is provided in the table below:

	aldrin	depth
1	3.80	bottom
2	4.80	bottom
...		
10	8.80	bottom
11	3.20	middepth
12	3.80	middepth
...		
20	6.60	middepth
21	3.10	surface
22	3.60	surface
...		
30	5.20	surface

	n	mean	sd
bottom	10	6.04	1.58
middepth	10	5.05	1.10
surface	10	4.20	0.66
overall	30	5.1 0	1.37



ANOVA

- To compare means of 2 groups we use a Z or a T statistic.
- To compare the means for 3+ groups we use a new test called the Analysis of Variance (***ANOVA)*** and a new statistic called ***F***.
- ANOVA is used to assess whether the mean of the outcome variable is different for different levels of a categorical variable.
  - H_0: The mean outcome is the same across all categories,

    $$\mu_1 = \mu_2 = \cdots = \mu_k,$$

    where μ_i represents the main of the outcome for observations in category i.
  - H_A: At least one mean is different than others.

ANOVA CONDITIONS

1. The observations should be independent within and between groups:
    - If the data are a simple random from less than 10% of the population, the condition is satisfied.
    - Carefully consider whether the data may be independent (e.g. no pairing).
    - Always important, but sometimes difficult to check.
2. The observations within each group should be nearly normal.
    - Especially important when the sample sizes are small.
3. The variability across the groups should be about equal.
    - Especially important when the sample sizes differ between groups.

Z/T TEST VS ANOVA

Z Test ($n \geq 30$) / T Test ($n < 30$)

- Compare means from ***two*** groups to see whether they are so far apart that the observed difference cannot reasonably be attributed to sampling variability.
- $H_0: \mu_1 = \mu_2$
- $H_A: \mu_1 <, \neq, > \mu_2$
- Test Statistic:

$$Z/T = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{SE(\overline{x_1} - \overline{x_2})}$$

ANOVA

- Compare the means form ***two or more*** groups to see whether they are so far apart that the observed differences cannot all reasonably be attributed to sampling variability.
- $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$
- H_A: At least one mean is different than others.
- Test statistic:

$$F = \frac{\text{variability bet. groups}}{\text{variability } \frac{\text{w}}{\text{in}} \text{ groups}}$$

NOTE:

- *With only two groups t-test and ANOVA are equivalent, but only if we use a pooled standard variance in the denominator of the test statistic.*
- *With more than two groups, ANOVA compares the sample means to an overall* ***grand mean.***

HYPOTHESES

- **What are the correct hypotheses for testing for a difference between the mean aldrin concentrations among the three levels?**

(a) $H_0: \mu_B = \mu_M = \mu_S$

$H_A: \mu_B \neq \mu_M \neq \mu_S$

(b) $H_0: \mu_B \neq \mu_M \neq \mu_S$

$H_A: \mu_B = \mu_M = \mu_S$

(c) $H_0: \mu_B = \mu_M = \mu_S$

H_A: At least one mean is different.

(d) $H_0: \mu_B = \mu_M = \mu_S = 0$

H_A: At least one mean is different.

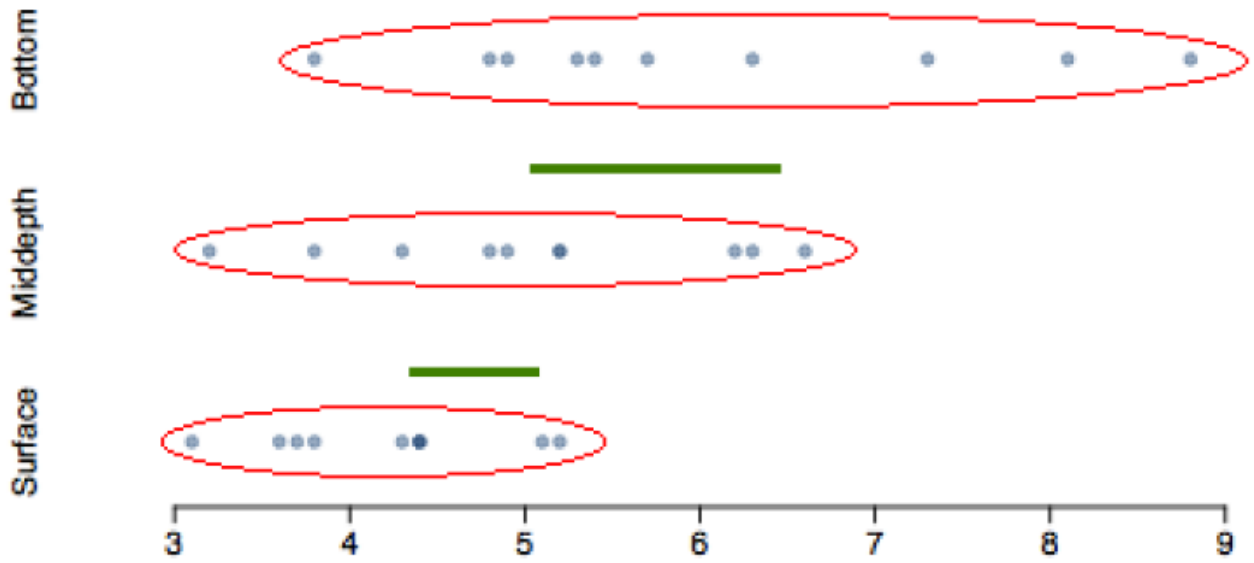
(e) $H_0: \mu_B = \mu_M = \mu_S$

$H_A: \mu_B > \mu_M > \mu_S$

TEST STATISTIC

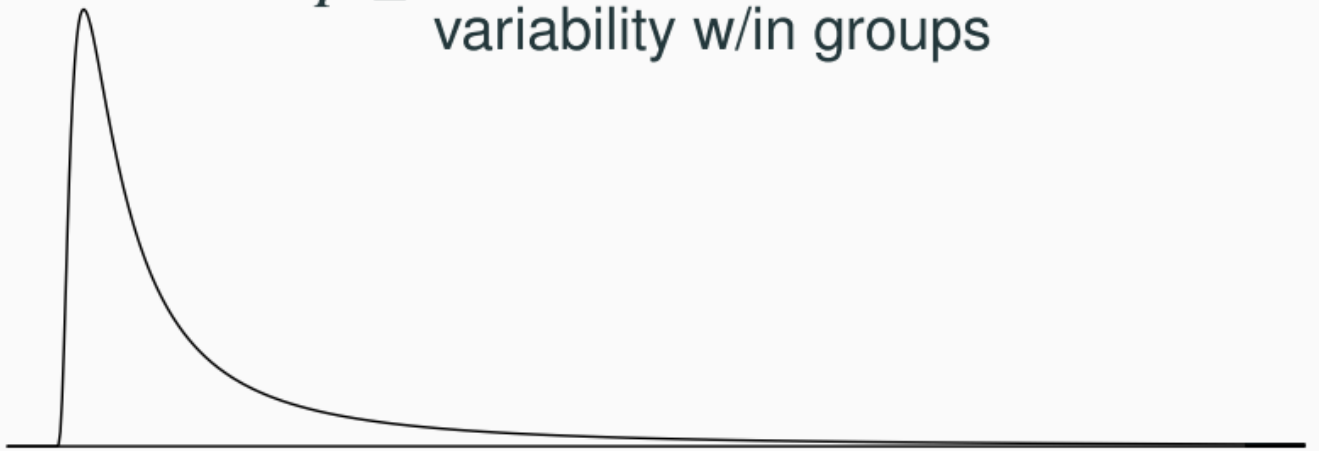
- **Does there appear to be a lot of variability within groups? How about between groups?**

$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$



F DISTRIBUTION AND P-VALUE

$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$



- In order to be able to reject H_0, we need a small p-value, which requires a large F statistic.
- In order to obtain a large F statistic, variability between sample means needs to be greater than variability within sample means.

F DISTRIBUTION AND P-VALUE

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(*G*roup)	depth	2	16.96	8.48	6.13	0.0063
(*E*rror)	Residuals	27	37.33	1.38		
	*T*otal	29	54.29			

- **Degrees of freedom (*df*) associated with ANOVA**
  - groups: $df_G = k - 1$, where k is the number of groups
  - total: $df_T = n - 1$, where n is the total sample size
  - error: $df_E = df_T - df_G$
  - $df_G = k - 1 = 3 - 1 = 2$
  - $df_T = n - 1 = 30 - 1 = 29$
  - $df_E = 29 - 2 = 27$

F DISTRIBUTION AND P-VALUE

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

- **Sum of squares between groups, SSG**
  - In some texts SSG will be referred to as SST (Sum of Squares between Treatments)
  - ***Measures the variability between groups***:

$$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2,$$

where n_i is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean.

	n	mean
bottom	10	6.04
middepth	10	5.05
surface	10	4.2
overall	30	5.1

$$SSG = (10\times(6.04 - 5.1)^2) + (10\times(5.05 - 5.1)^2) + (10\times(4.2 - 5.1)^2) = 16.96$$

F DISTRIBUTION AND P-VALUE

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(*G*roup)	depth	2	16.96	8.48	6.13	0.0063
(*E*rror)	Residuals	27	37.33	1.38		
	*T*otal	29	*54.29*			

- **Sum of squares total, SST**
  - In some texts SST will be referred to as SS_TOT
  - ***Measures the total variability in the data:***

$$SST = \sum_{i=1}^{k} (x_i - \bar{x})^2,$$

where x_i represent each observation in the dataset.

$$SST = (3.8 - 5.1)^2 + (4.8 - 5.1)^2 + (4.9 - 5.1)^2 + \cdots + (5.2 - 5.1)^2$$
$$= (-1.3)^2 + (-0.3)^2 + (-0.2)^2 + \cdots + (0.1)^2$$
$$= 1.69 + 0.09 + 0.04 + \cdots + 0.01$$
$$= 54.29$$

F DISTRIBUTION AND P-VALUE

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(*G*roup)	depth	2	16.96	8.48	6.13	0.0063
(*E*rror)	Residuals	27	*37.33*	1.38		
	*T*otal	29	54.29			

- **Sum of squares error, SSE**
  - Measures the variability within groups:

$$SSE = SST - SSG$$

$$SSE = 54.29 - 16.96 = 37.33$$

F DISTRIBUTION AND P-VALUE

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(*G*roup)	depth	2	16.96	*8.48*	6.13	0.0063
(*E*rror)	Residuals	27	37.33	*1.38*		
	*T*otal	29	54.29			

- **Mean square error**
  - Mean square error is calculated as sum of squares divided by the degrees of freedom.

$$MSG = 16.96 \div 2 = 8.48$$

$$MSE = 37.33 \div 27 = 1.38$$

F DISTRIBUTION AND P-VALUE

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.14	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

- **Test statistic, F value**
  - As we discussed before, the F statistic is the ratio of the between group and within variability.

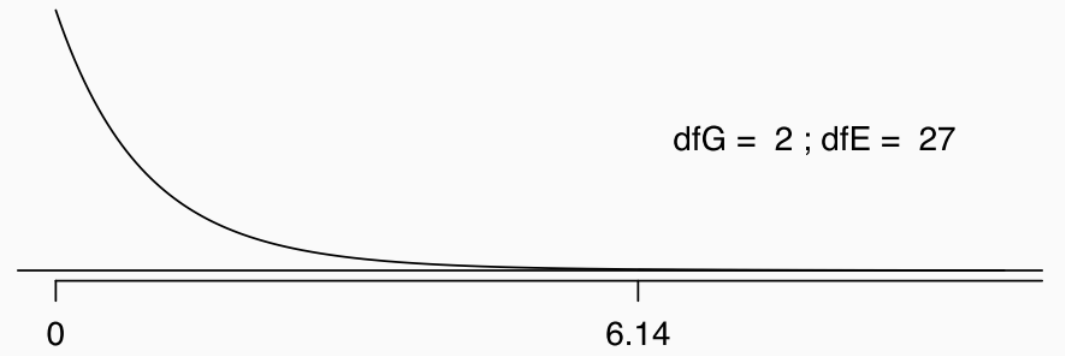
$$F = \frac{MSG}{MSE}$$

$$F = \frac{8.48}{1.38} = 6.14$$

F DISTRIBUTION AND P-VALUE

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(*G*roup)	depth	2	16.96	8.48	*6.14*	0.0063
(*E*rror)	Residuals	27	37.33	1.38		
	*T*otal	29	54.29			

- **p – value**
  - P – value is the probability of at least as large a ratio between the "between group" and "within group" variability, if in fact the means of all groups are equal. It is calculated as the area under the F curve, with degrees of freedom df_G and df_E, above the observed F statistic.



CONCLUSION – IN CONTEXT

- **What is the conclusion of the hypothesis test?**
  - The data provide convincing evidence that the average aldrin concentration
    (a) is different for all groups.
    (b) on the surface is lower than the other levels.
    (c) is different for at least one group.
    (d) is the same for all groups.

WHICH MEANS DIFFER?

- We concluded that at least one pair of means differ. The natural question that follows is "which ones"?
- Use t-tests comparing each pair of means to each other:
  - with a common variance (MSE from the ANOVA table) instead of each group's variances in the calculation of the standard error,
  - and with a common degrees of freedom (df_E from the ANOVA table)
- We can do two sample t tests for differences in each possible pair of groups,

Can you see any pitfalls with this approach?

- When we run too many tests, the Type I Error rate increases.
- This issue is resolved by using a modified significance level.

MULTIPLE COMPARISONS

- The scenario of testing many pairs of groups is called **multiple comparisons**.
- The **Bonferroni correction** suggests that a more stringent significance level is more appropriate for these tests:

$$\alpha^* = \alpha/K \ ,$$

where K is the number of comparisons being considered.

- If there are k groups, then usually all possible pairs are compared and $K = \frac{k(k-1)}{2}$.

DETERMINING THE MODIFIED α

- **In the aldrin data set depth has 3 levels: bottom, mid-depth and surface. If $\alpha = 0.05$, what should be the modified significance level for two sample t tests for determining which pairs of groups have significantly different means?**

$$\mathrm{k} = 3 \rightarrow K = \frac{k(k-1)}{2} = \frac{3(3-1)}{2} = 3$$

(a) $\alpha^* = 0.05$

(b) $\alpha^* = \frac{0.05}{2} = 0.025$

(c) $\alpha^* = \frac{0.05}{3} = 0.0167$

(d) $\alpha^* = \frac{0.05}{6} = 0.0083$

WHICH MEANS DIFFER?

- If the ANOVA assumption of equal variability across groups is satisfied, we can use the data from all groups to estimate variability:
  - Estimate any within-group standard deviation with $\sqrt{MSE}$, which is s_{pooled}
  - Use the error degrees of freedom, df_E, for $t-$distributions.

Difference in two means: after ANOVA

$$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}$$

WHICH MEANS DIFFER? (CONT.)

- **Is there a difference between the average aldrin concentration at the bottom at mid-depth?**

	n	mean	sd
bottom	10	*6.04*	1.58
middepth	10	*5.05*	1.10
surface	10	4.2	0.66
overall	30	5.1	1.37

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
depth	2	16.96	8.48	6.13	0.0063
Residuals	*27*	37.33	*1.38*		
Total	29	54.29			

$$T_{df_E} = \frac{(\bar{x}_{bottom} - \bar{x}_{middepth})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{middepth}}}}$$

$$T_{27} = \frac{(6.04 - 5.05)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{0.99}{0.53} = 1.87$$

$$0.05 < p - value < 0.10 \qquad \text{(two – sided)}$$

$$\alpha^* = \frac{0.05}{3} = 0.0167$$

- Fail to reject H_0 the data do not provide convincing evidence of a difference between the average aldrin concentrations at bottom and mid-depth.

WHICH MEANS DIFFER? (CONT.)

- **Pairwise comparisons**
  - Is there a difference between the average aldrin concentration at the bottom and at surface?

$$T_{df_E} = \frac{(\bar{x}_{bottom} - \bar{x}_{surface})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{surface}}}}$$

$$T_{27} = \frac{(6.04 - 4.02)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{2.02}{0.53} = 3.81$$

$$p - value < 0.01 \qquad \text{(two – sided)}$$

$$\alpha^* = \frac{0.05}{3} = 0.0167$$

  - Reject H_0, data provide convincing evidence of a difference between the average aldrin concentrations at bottom and surface.

NOTE: Is there a difference between the average aldrin concentration at the mid-depth and at surface? (homework)

LINEAR REGRESSION

- **Linear regression** is a statistical technique that can be used for prediction and evaluating whether there is linear relationship between two numerical variables x and y.
- Linear regression assumes that the relationship between two variables $x = \{x_1, x_2, \dots, x_n\}$ and $y = \{y_1, y_2, \dots, y_n\}$ can be modeled by a ***straight line***:

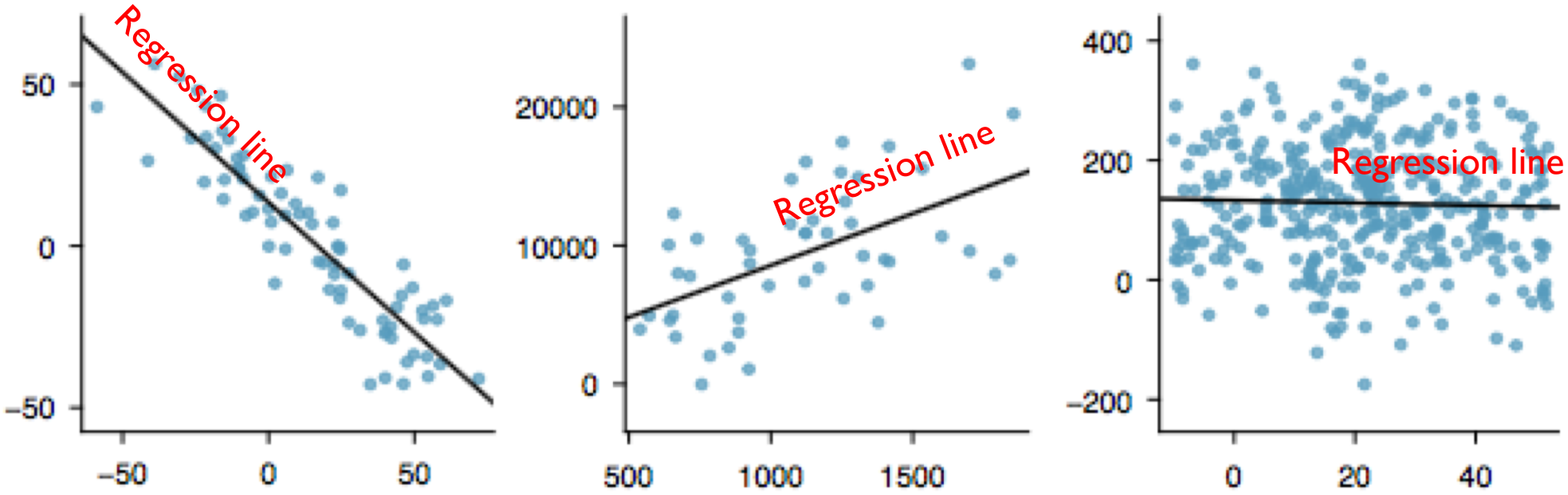
$$y = \beta_0 + \beta_1 x \qquad (1)$$

Where:

- x - **predictor variable** (synonyms: explanatory variable, independent variable)
- y – **response variable** (synonym: dependent variable)
- β_0 - ***intercept*** *(It is interpreted as the expected value of the response variable when the predictor is zero)*
- β_1 - ***slope parameter*** *(It is interpreted as the change in the mean response for each one-unit increase in the predictor).*

HINT: When the slope is zero, this implies that the predictor x has no effect on the value of the response y!

LINEAR REGRESSION - EXAMPLE



Three data sets where a linear model may be useful even though the data do not all fall exactly on the line.

LINEAR REGRESSION

- *Parameters $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ from equation (1) can be estimated using data and we write their point estimates as $\boldsymbol{b_0}$ and $\boldsymbol{b_1}$ respectively.*
- Therefore, linear regression model (1) based on point estimates has the following form:

$$\hat{y} = b_0 + b_1 x \quad (2)$$

where:

- b_0 is point estimate of the intercept β_0
- b_1 is point estimate of the slope parameter β_1
- we use $\hat{y} = \{\widehat{y_1}, \widehat{y_2}, \dots, \widehat{y_n}\}$ to indicate that this is a collection of estimated (predicted) observations of observed variable $y = \{y_1, y_2, \dots, y_n\}$, based on input collection of predictor observations $x = \{x_1, x_2, \dots, x_n\}$.

- *The differences between observed and estimated values are called **residuals (ε)**:*

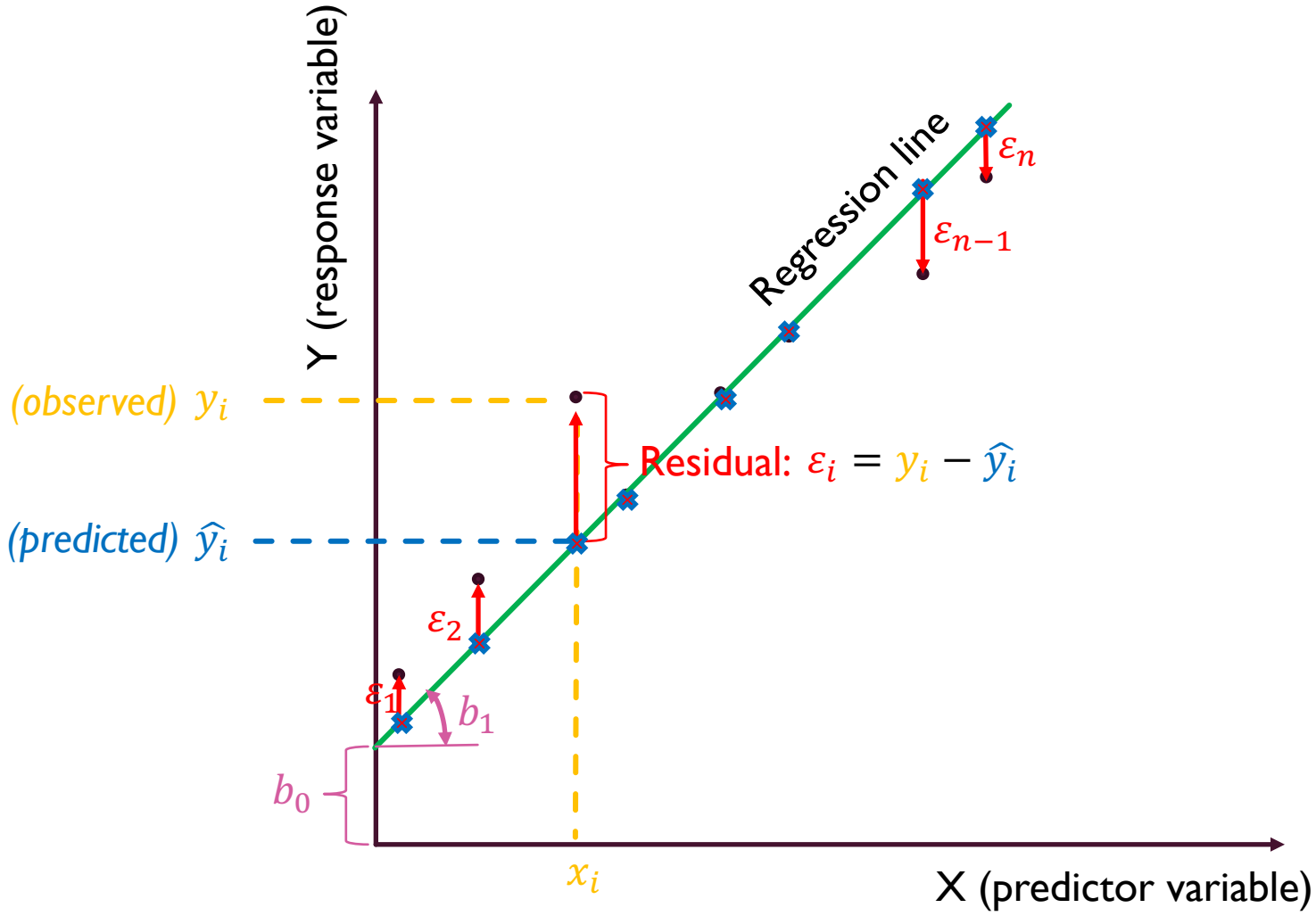
$$\varepsilon = y - \hat{y} = \{y_1 - \widehat{y_1}, y_2 - \widehat{y_2}, \dots, y_n - \widehat{y_n}\} = \{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n\}$$

- The residual of the i-th observation (x_i, y_i) is the difference of the observed response (y_i) and the response we would predict based on the model fit $(\widehat{y_i})$:

$$\varepsilon_i = y_i - \widehat{y_i}$$

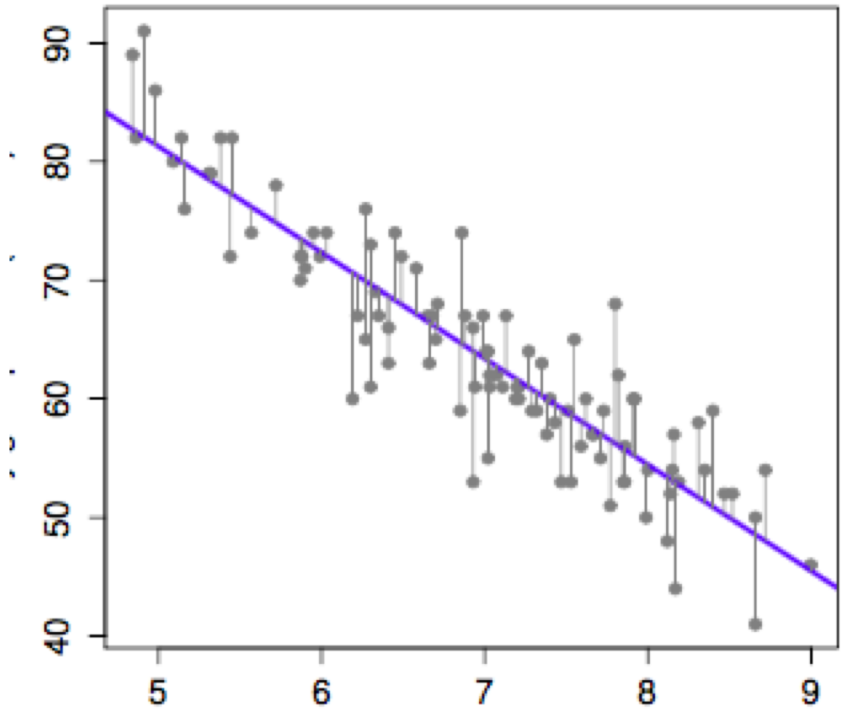
We typically identify $\widehat{y_i}$ by plugging x_i into the linear regression model (2)

LINEAR REGRESSION - EXAMPLE

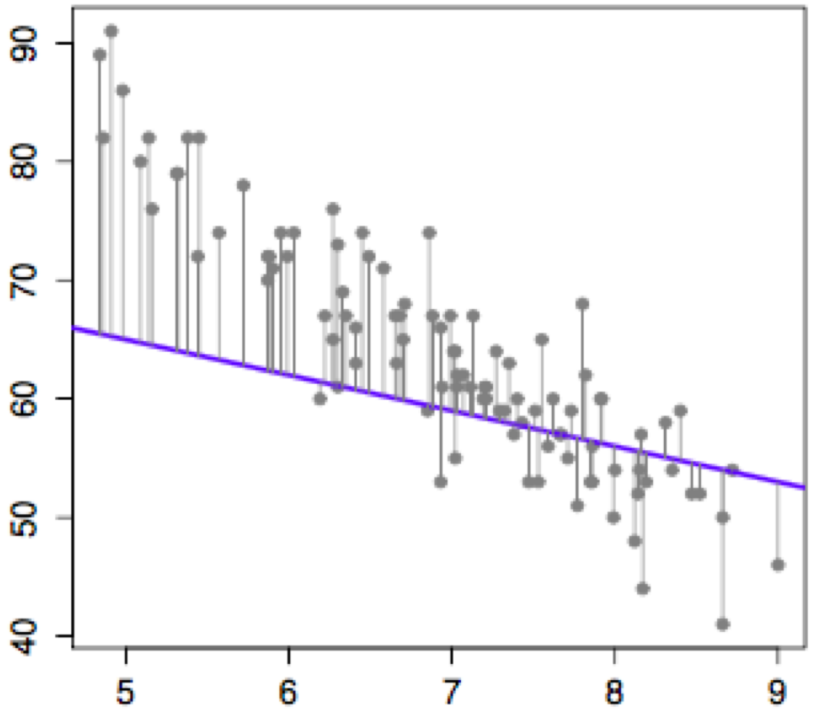


X	Y
x_1	y_1
x_2	y_2
…	…
x_i	y_i
…	…
x_n	y_n

- When the regression line represents a good approximation of our data set, all residuals look pretty small.



The best fitting regression line
(line that has the smallest possible residuals)



A poor fitting regression line
(large residuals)

FITTING A LINE BY ORDINARY LEAST SQUARES REGRESSION

- One of the most common approaches of finding the line with the smallest possible residuals is by using ***ordinary least squares regression (OLS).***
- The goal of the OLS is to find the line that minimizes the ***least square criterion***, i.e. minimizes the sum of squared residuals:

$$\sum_{i=1}^{n} {\varepsilon_i}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- The line that minimizes this least squares criterion is usually called the ***least squares line***.

FITTING A LINE BY ORDINARY LEAST SQUARES REGRESSION

- An algorithm of finding the least squares line:
  - Estimate the slope parameter b_1 using the following equation:

$$b_1 = \frac{s_y}{s_x}\rho_{xy}$$

where ρ_{xy} is the correlation between the two variables, and s_x and s_y are the sample standard deviations of the explanatory variable x and response variable y, respectively

  - Estimate the intercept parameter b_0 using the following equation:

$$b_0 = \bar{y} - b_1\bar{x} = \bar{y} - \frac{s_y}{s_x}\rho_{xy}\bar{x}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of the explanatory variable x and response variable y, respectively

  - When we put back estimated regression parameters b_0 and $b_1$into equation (2) we will get *the* ***formula of the least square line***:

$$\hat{y} = b_0 + b_1 x = \bar{y} - \frac{s_y}{s_x}\rho_{xy}\bar{x} + \frac{s_y}{s_x}\rho_{xy}x$$

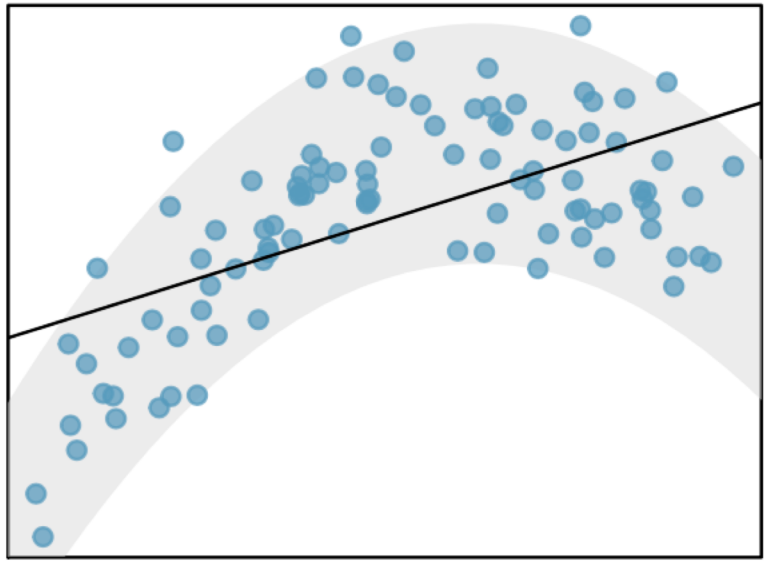
HINT: Once we have a formula of the least squares line, we can use input values of x to get predicted values $\hat{y}$

CONDITIONS FOR THE LEAST SQUARES LINE

- With a fitted simple linear model you're able to calculate a point estimate $\widehat{y_i}$ of the *mean response* value y_i.
  - To do this, you simply plug in (to the fitted model equation) the value of x_i you're interested in.
- When fitting a least squares line, we generally require the following conditions to be met:
  - **Linearity**
  - **Nearly normal residuals**
  - **Constant variability**

CONDITIONS FOR THE LEAST SQUARES LINE - LINEARITY

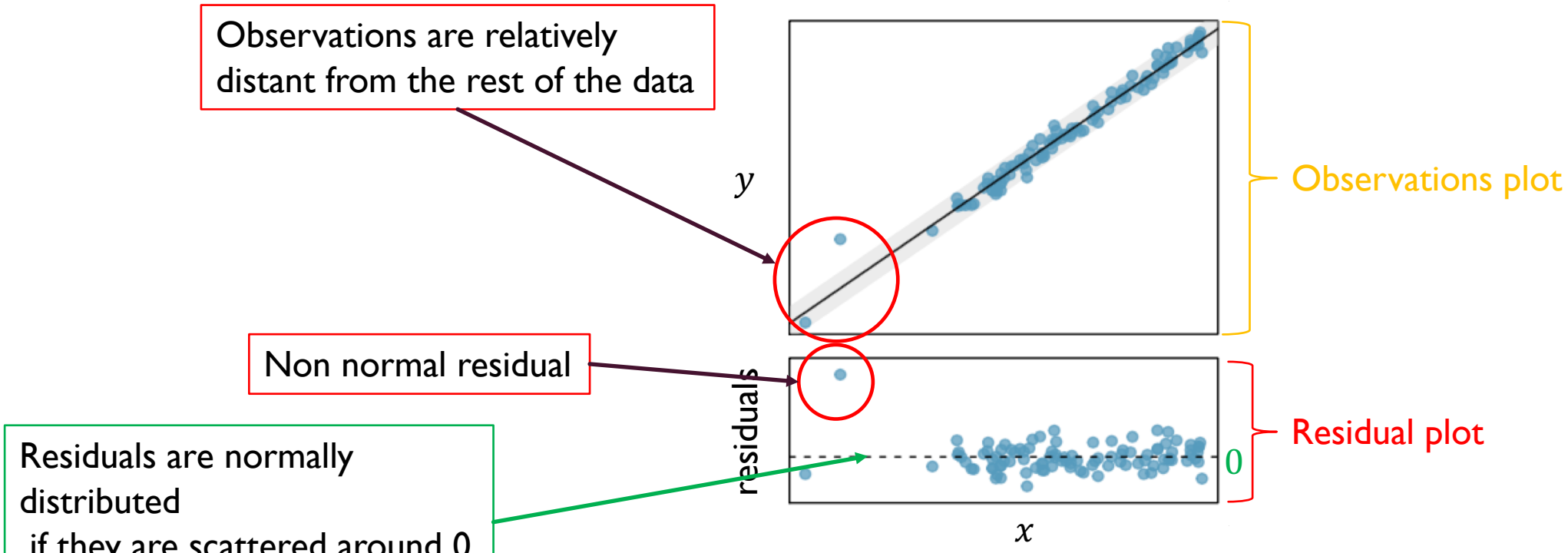
- **Linearity** – The data should show a linear trend. If there is a nonlinear trend (e.g. left panel of Figure 7.13), an advanced regression method should be applied.



An example when the linearity condition is not met
(We have non linear data which cannot be fitted by the straight line)

CONDITIONS FOR THE LEAST SQUARES LINE - NEARLY NORMAL RESIDUALS

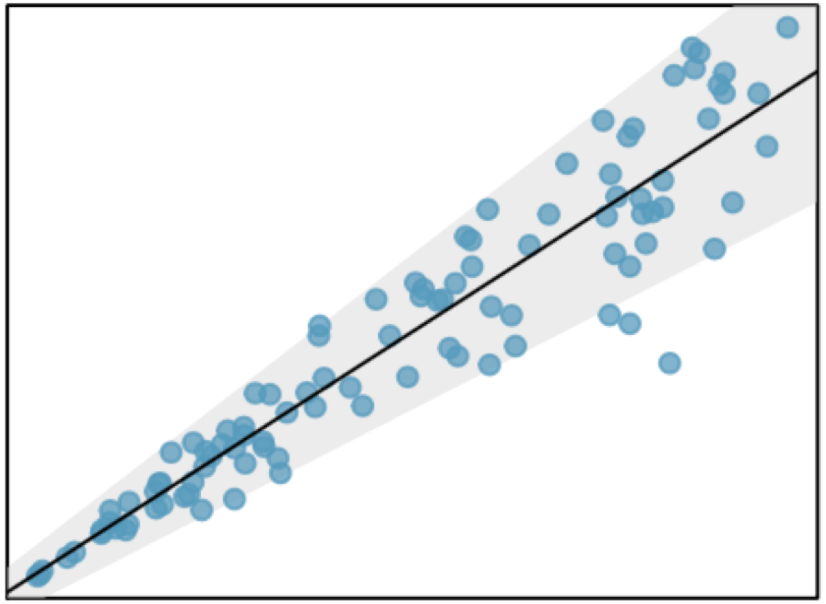
- **Nearly normal residuals** – Generally the residuals must be nearly normal. When this condition is found to be unreasonable, it is usually because of outliers or concerns about influential points.



An example when the condition for nearly normal residuals is not met

CONDITIONS FOR THE LEAST SQUARES LINE – CONSTANT VARIABILITY

- **Constant variability** – The variability of points around the least squares line remains roughly constant.



An example when the condition of constant variability is not met
(the variability of the data around the line increases with larger values of x)

FITTING THE LEAST SQUARES LINE - EXAMPLE

- Summary statistics for family income and gift aid data from a random sample of 50 students in the 2011 freshman class of Elmhurst College in Illinois is given in the following table:

	Family income, in $1000s ("X")	Gift aid, in $1000s ("Y")
MEAN	$\bar{x} = 101.8$	$\bar{y} = 19.94$
SD	$s_x = 63.2$	$s_y = 5.46$
		$\rho_{xy} = -0.499$

Equation of the least square regression line that could be used for predicting gift aid based on student's family income:

$$\widehat{aid} = b_0 + b_1 \times family\_income$$

Estimated coefficients b_0 and b_1 :

$$b_1 = \frac{s_y}{s_x}\rho_{xy} = \frac{5.46}{63.2} \times (-0.499) = -0.0431$$

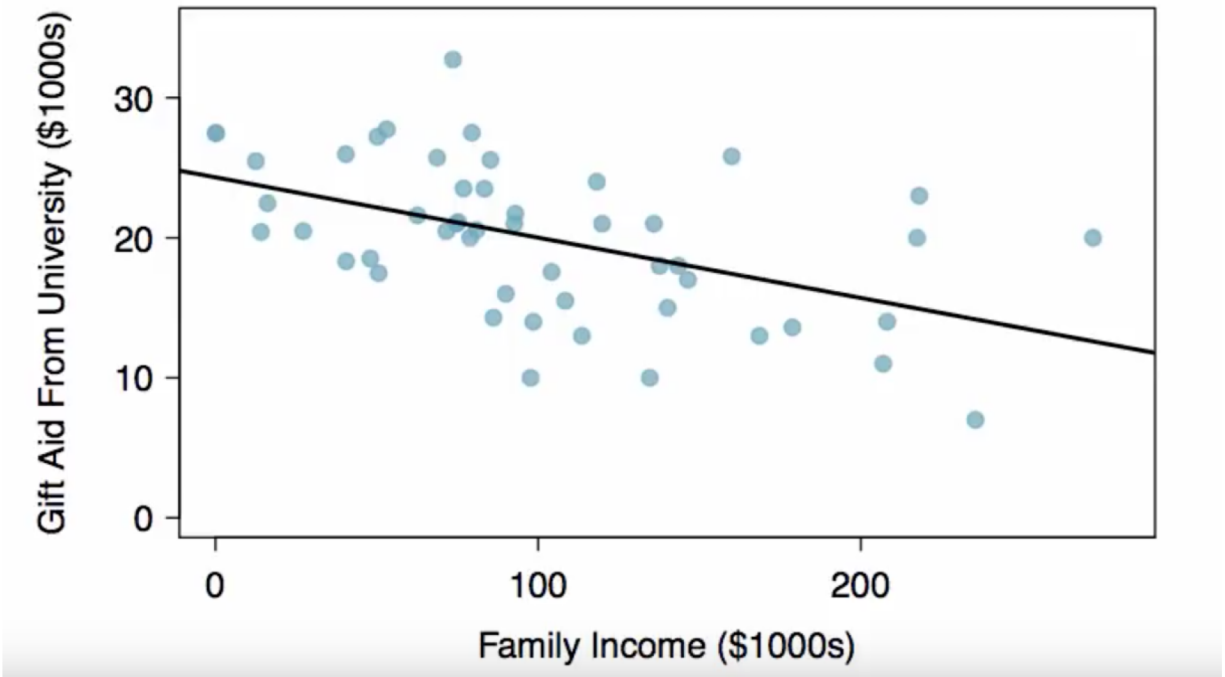
$$b_0 = \bar{y} - b_1\bar{x} = 19.94 - (-0.0431)101.8 = 24.3$$

Estimated the least square regression line that could be used for predicting gift aid based on student's family income:

$$\widehat{aid} = 24.3 - 0.0431 \times family\_income$$

FITTING THE LEAST SQUARES LINE – EXAMPLE (MODEL INTERPRETATION)

- How to interpret values of intercept and slope parameters?
  - ***THE ESTIMATED INTERCEPT*** ($b_0 = 24.3$ in \$1000s) – describes the average aid if a student's family had no income, i.e. if student's family doesn't have any income, student can expect gift aid of 24.3x\$1000=\$24300
  - ***THE ESTIMATED SLOPE*** ($b_1 = -0.0431$ in \$1000s) – For each additional \$1000 of family income we would expect a student to receive -0.0431x\$1000=-\$43.1 in aid on average, i.e. \$43 less in aid on average.



Estimated the least square regression line: $\widehat{aid} = 24.3 - 0.0431 \times family\_income$

QUANTIFYING THE STRENGTH OF A FIT OF THE REGRESSION MODEL

- *When we have estimated regression coefficients b_0 and b_1, how do we determine how good our model is?*
- One of the approaches to explain the strength of a linear fit is by using the **coefficient of determination** (R^2)
- The R^2 describes the proportion of the variation in the response that can be attributed to the predictor, i.e. that is explained by the least squares line. We calculate R^2 using following equation:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Where:

$SS_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ – ***sum of squared residuals***,

$SS_{tot} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ – ***total variability in outcome variable***

FITTING THE LEAST SQUARES LINE – EXAMPLE (STRENGTH OF A FIT)

- *Estimated the least square regression line:* $\widehat{aid} = 24.3 - 0.0431 \times family\_income$

	Family income, in $1000s ("X")	Gift aid, in $1000s ("Y")
MEAN	$\bar{x} = 101.8$	$\bar{y} = 19.94$
SD	$s_x = 63.2$	$s_y = 5.46$
		$\rho_{xy} = -0.499$

$SS_{tot} = 29.8$

$SS_{res} = 22.4$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{22.4}{29.8} = 0.25$$

The fact that we have obtained $R^2 = 0.25$ means that the $family\_income$ explains 25% of the variance in the outcome $\widehat{aid}$

INFERENCE FOR LINEAR REGRESSION

- In simple linear regression, there is a natural question that should always be asked:
  - ***Is there statistical evidence to support the presence of a relationship between the predictor x and the response y?***
  - To put this question in another way, we want to investigate if the slope $\boldsymbol{b_1}$ is significantly different from zero.
- To answer this question we set up our null and alternative hypotheses as follows:

  $\boldsymbol{H_0}: \boldsymbol{b_1 = 0}$ *(i.e. there **IS NOT** enough evidence to support the presence of a relationship between the predictor and the response)*

  $\boldsymbol{H_{A1}}: \boldsymbol{b_1 \neq 0}$ *(i.e. there **IS** enough evidence to support the presence of a relationship between the predictor and the response)*

  The test statistic for this test is:

$$t = \frac{b_1}{SE_{b1}}$$

  Where:

$SE_{b1} = \sqrt{\frac{\frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{n-1}}{\sum_{i=1}^{n}(x_i - \bar{x})}}$ – standard error of the slope estimate b_1

HINT: *Test statistic t follows t −distribution with $n - 2$ degrees of freedom, where the n is total number of observations in the dataset.* *We accept the null hypotheses if* $-t_{\frac{\alpha}{2},n-2} < t < t_{\frac{\alpha}{2},n-2}$

INFERENCE FOR LINEAR REGRESSION– EXAMPLE

Estimated the least square regression line: $\widehat{aid} = 24.3 - 0.0431 \times family\_income$

If we use R to fit the least squares line, as the result we will get following table

		Estimate	Std. Error	t value	Pr(>|t|)
b_0 →	(Intercept)	24.3193	1.2915	18.83	0.0000
b_1 →	family_income	-0.0431	0.0108	-3.98	0.0002

Test statistic t

R always shows P-value for two sided hypotheses test

HINT: if your test is one sided and the point estimate is in direction of H_A, then you can halve this p-value to get the one-tail area

Assuming that the desired significance level is 0.05,
with the small P-value (0.0002 < 0.05) attached to this test statistic null hypotheses $\boldsymbol{H_0: b_1 = 0}$ is rejected indicating that a relation exists between predictor variable $family\_income$ and response variable *aid.*

REFERENCES

These lecture notes are heavily influenced by following literature:

1. O'Neil C., Schutt R. Doing Data Science. *O'Reilly Media*, 2013.
2. Grolemund G., Wickham H. R for Data Science. *O'Reilly Media*, 2017.
3. Bruce A, Bruce P. Practical Statistics for Data Scientists. *O'Reilly Media*, 2017.
4. Ismay C, Kim A Y. An Introduction to Statistical and Data Sciences via R. Retreived September 5, 2017, from http://moderndive.com
5. Diez D M, Barr C D, Rundel M C. Open Intro Statistics (Third Edition). *OpenIntro, Inc.*,2015.
6. Larson G. Sta 101 – Data Analysis and Statistical Inference. *Duke University,* Summer 2015 Term 2, 2015.
7. Davies T M. The Book of R. *No Starch Press*, 2016.
8. Wikipedia contributors. Quartile. *Wikipedia, The Free Encyclopedia*, Retrieved September 5, 2017, from https://en.wikipedia.org/wiki/Quartile
9. Evans J D. Straightforward statistics for the behavioral sciences. *Brooks/Cole Publishing,* Pacific Grove, CA, 1996.