Statistics (2023)

SB, NIBMG

Statistics

- The science of collectiong, organizing, presenting, analyzing, and interpreting data to assist in making more effective decisions

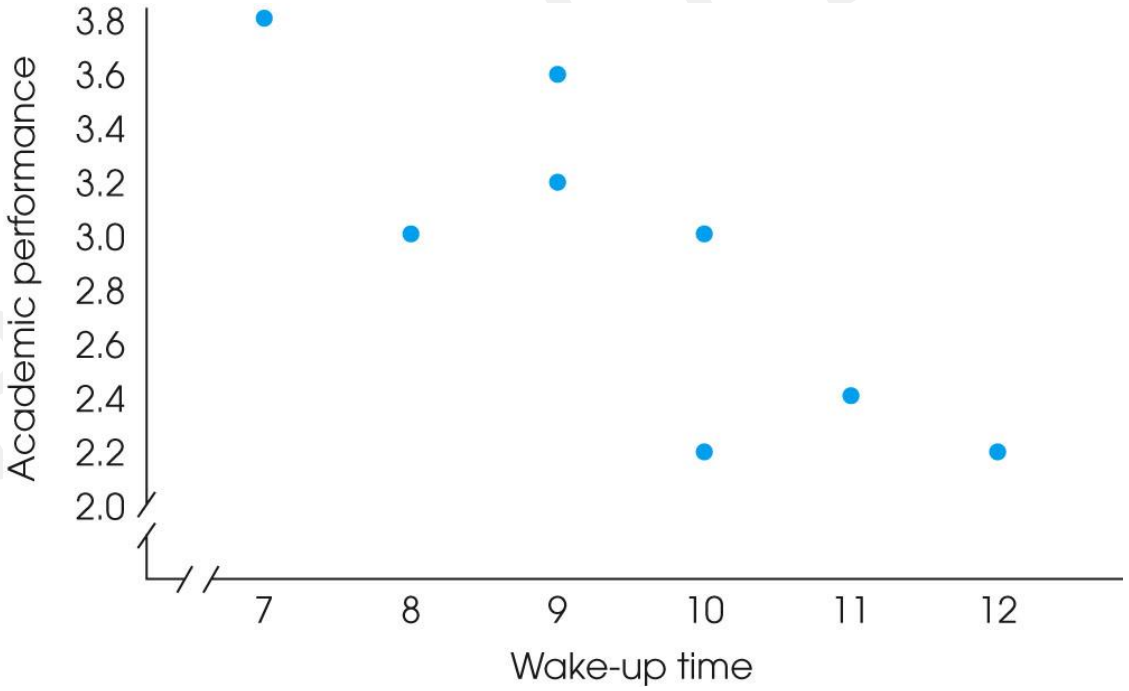
- Statistical analysis – procedures used to process, summarize, and investigate data to make effective decisions.

Correlational Studies

- The goal of a **correlational** study is to determine whether there is a relationship between two variables and to describe the relationship.

- A **correlational** study simply observes the two variables as they exist naturally.

Child	Wake-up Time	Academic Performance
A	11	2.4
B	9	3.6
C	9	3.2
D	12	2.2
E	7	3.8
F	10	2.2
G	10	3.0
H	8	3.0



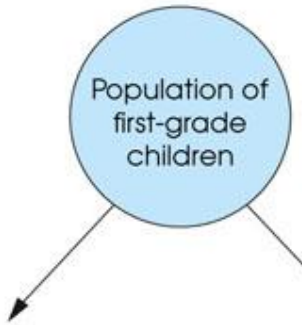
Experiments

- The goal of an **experiment** is to demonstrate a cause-and-effect relationship between two variables; that is, to show that changing the value of one variable causes changes to occur in a second variable.

Experiments (cont.)

- In an **experiment**, one variable is manipulated to create treatment conditions. A second variable is observed and measured to obtain scores for a group of individuals in each of the treatment conditions. The measurements are then compared to see if there are differences between treatment conditions. All other variables are controlled to prevent them from influencing the results.
- In an experiment, the manipulated variable is called the **independent variable** and the observed variable is the **dependent variable**.

Step 1
Experiment:
Compare two teaching methods



Data
Test scores for the students in each sample

Sample A
Taught by Method A

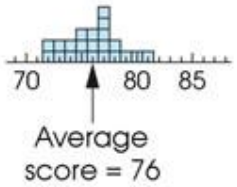
73	75	72	79
76	77	75	77
72	75	76	78
80	74	76	78
73	77	74	81
77	77		

Sample B
Taught by Method B

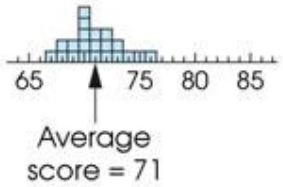
68	70	73	71
67	72	70	71
75	68	70	71
72	74	69	72
76	73	70	70
69			

Step 2
Descriptive statistics:
Organize and simplify

Sample A



Sample B



Step 3
Inferential statistics:
Interpret results

The sample data show a 5-point difference between the two teaching methods. However, there are two ways to interpret the results:

1. There actually is no difference between the two teaching methods, and the sample difference is due to chance (sampling error).
2. There really is a difference between the two methods, and the sample data accurately reflect this difference.

The goal of inferential statistics is to help researchers decide between the two interpretations.

Statistical data

- The collection of data that are relevant to the problem being studied is commonly the most difficult, expensive, and time-consuming part of the entire research project.
- Statistical data are usually obtained by counting or measuring items.
  - **Primary data** are collected specifically for the analysis desired
  - **Secondary data** have already been compiled and are available for statistical analysis
- A **variable** is an item of interest that can take on many different numerical values.
- A **constant** has a fixed numerical value.

Data

Statistical data are usually obtained by counting or measuring items. Most data can be put into the following categories:

- **Qualitative -** data are measurements that each fail into one of several categories. (hair color, ethnic groups and other attributes of the population)
- **quantitative** - data are observations that are measured on a numerical scale (distance traveled to college, number of children in a family, etc.)

Qualitative data

Qualitative data are generally described by words or letters. They are not as widely used as quantitative data because many numerical techniques do not apply to the qualitative data. For example, it does not make sense to find an average hair color or blood type.

Qualitative data can be separated into two subgroups:

- **dichotomous** (if it takes the form of a word with two options (gender - male or female)
- **polytomous** (if it takes the form of a word with more than two options (education - primary school, secondary school and university).

Quantitative data

Quantitative data are always numbers and are the

result of counting or measuring attributes of a population.

Quantitative data can be separated into two

subgroups:

- **discrete** (if it is the result of *counting* (the number of students of a given ethnic group in a class, the number of books on a shelf, ...)
- **continuous** (if it is the result of *measuring* (distance traveled, weight of luggage, …)

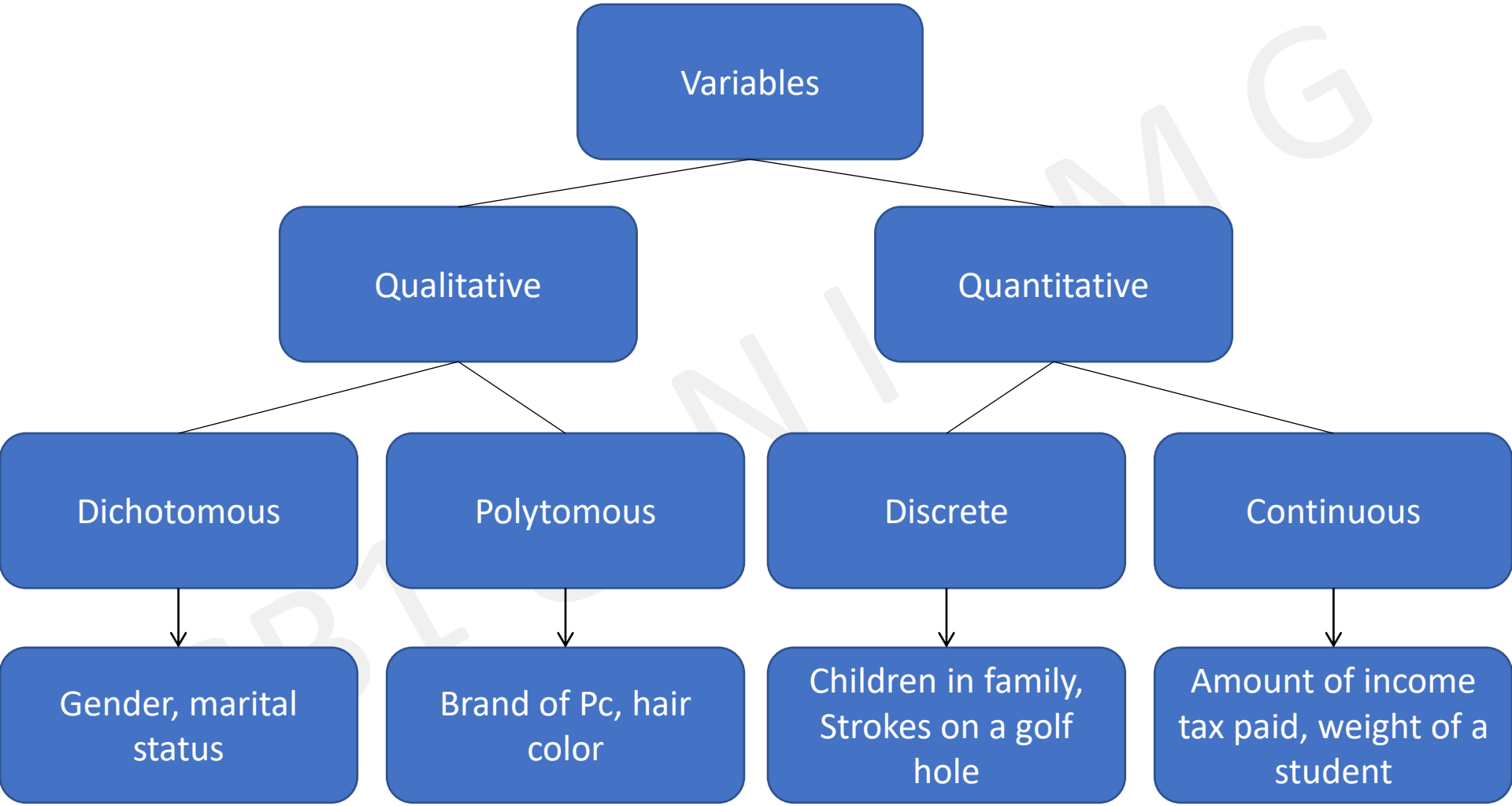
Variables

- A **variable** is a characteristic or condition that can change or take on different values.
- Most research begins with a general question about the relationship between two variables for a specific group of individuals.

Types of Variables

- Variables can be classified as discrete or continuous.
- **Discrete variables** (such as class size) consist of indivisible categories, and **continuous variables** (such as time or weight) are infinitely divisible into whatever units a researcher may choose. For example, time can be measured to the nearest minute, second, half-second, etc.

Types of variables



Measuring Variables

- To establish relationships between variables, researchers must observe the variables and record their observations. This requires that the variables be **measured**.
- The process of measuring a variable requires a set of categories called a **scale of measurement** and a process that classifies each individual into one category.

4 Types of Measurement Scales

1. A **nominal scale** is an unordered set of categories identified only by name. Nominal measurements only permit you to determine whether two individuals are the same or different.
2. An **ordinal scale** is an ordered set of categories. Ordinal measurements tell you the direction of difference between two individuals.

4 Types of Measurement Scales

3. An **interval scale** is an ordered series of equal-sized categories. Interval measurements identify the direction and magnitude of a difference. The zero point is located arbitrarily on an interval scale.
4. A **ratio scale** is an interval scale where a value of zero indicates none of the variable. Ratio measurements identify the direction and magnitude of differences and allow ratio comparisons of measurements.

Population

- The entire group of individuals is called the **population**.
- For example, a researcher may be interested in the relation between class size (variable 1) and academic performance (variable 2) for the **population of third-grade children**.

Sample

- Usually, populations are so large that a researcher cannot examine the entire group. Therefore, a **sample** is selected to represent the population in a research study. The goal is to use the results obtained from the sample to help answer questions about the population.

Types of statistics

- **Descriptive statistics** – Methods of organizing, summarizing, and presenting data in an informative way
- **Inferential statistics** – The methods used to determine something about a population on the basis of a sample
  - Population –The entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest
  - Sample – A portion, or part, of the population of interest

Descriptive Statistics

- **Descriptive statistics** deals with methods for organizing and summarizing data.
- For example, tables or graphs are used to organize data, and descriptive values such as the average score are used to summarize data.
- A descriptive value for a population is called a **parameter** and a descriptive value for a sample is called a **statistic**.

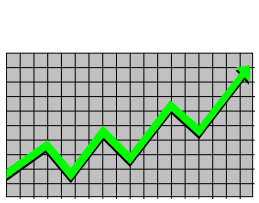
Inferential Statistics

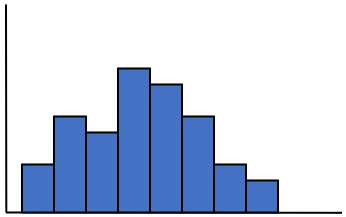
- **Inferential statistics** are methods for using sample data to make general conclusions (inferences) about populations.
- Because a sample is typically only a part of the whole population, sample data provide only limited information about the population. As a result, sample statistics are generally imperfect representatives of the corresponding population parameters.

Descriptive Statistics

- Collect data
  - e.g., Survey

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





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

1- Numerical presentation

Tabular presentation

Simple frequency distribution Table

Title

Name of variable (Units of variable)	Frequency	%
- - Categories -		
Total		

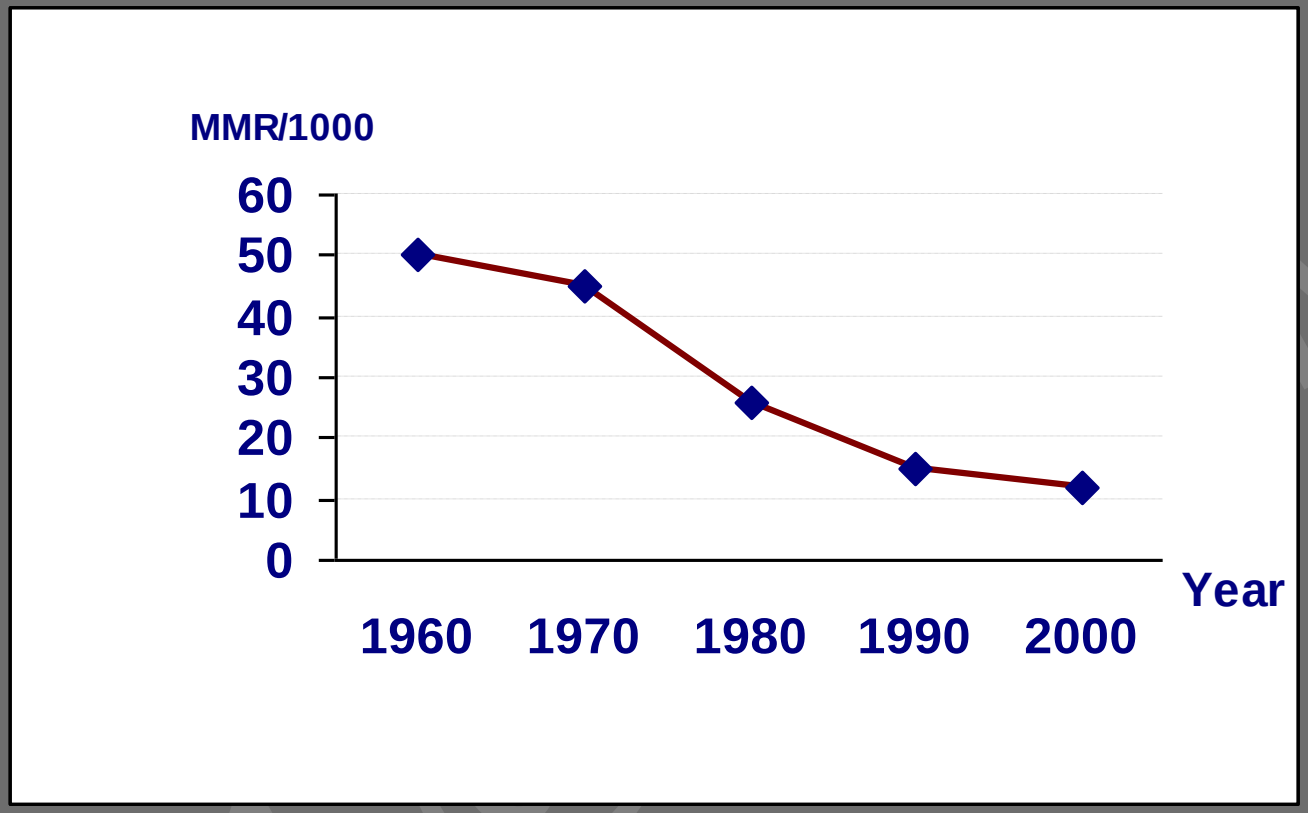
Table (I): Distribution of 50 patients at the surgical department of Alexandria hospital in May 2008 according to their ABO blood groups

Blood group	Frequency	%
A	12	24
B	18	36
AB	5	10
O	15	30
Total	50	100

Table (II): Distribution of 50 patients at the surgical department of Alexandria hospital in May 2008 according to their age

Age (years)	**Frequency**	**%**
20-<30	12	24
30-	18	36
40-	5	10
50+	15	30
Total	50	100

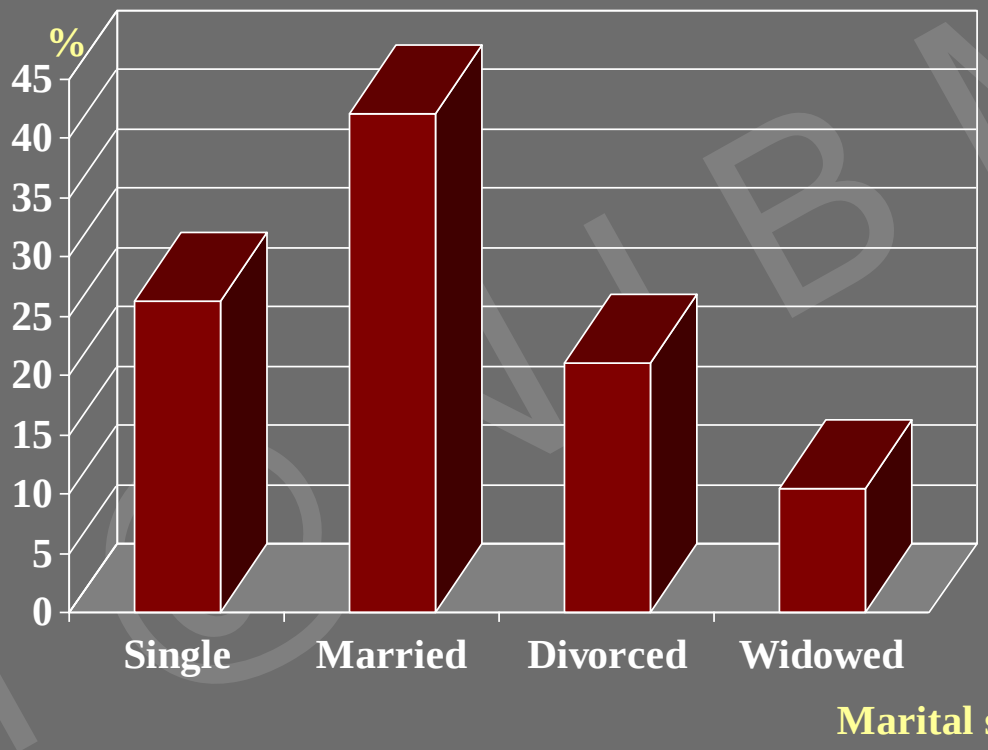
Line Graph (e.g. for time series data)



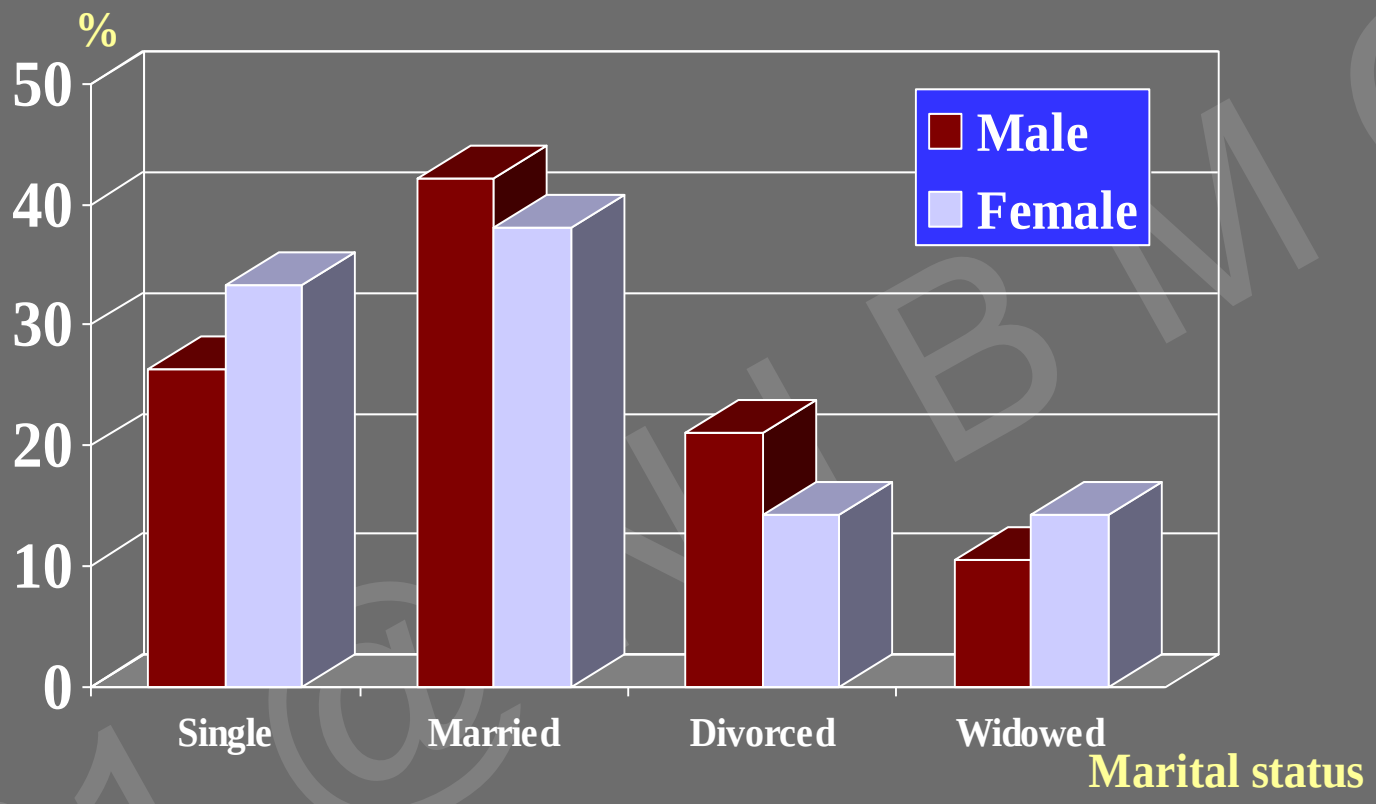
Year	MMR
1960	50
1970	45
1980	26
1990	15
2000	12

Figure (1): Maternal mortality rate of (country), 1960-2000

Bar chart (Column chart)



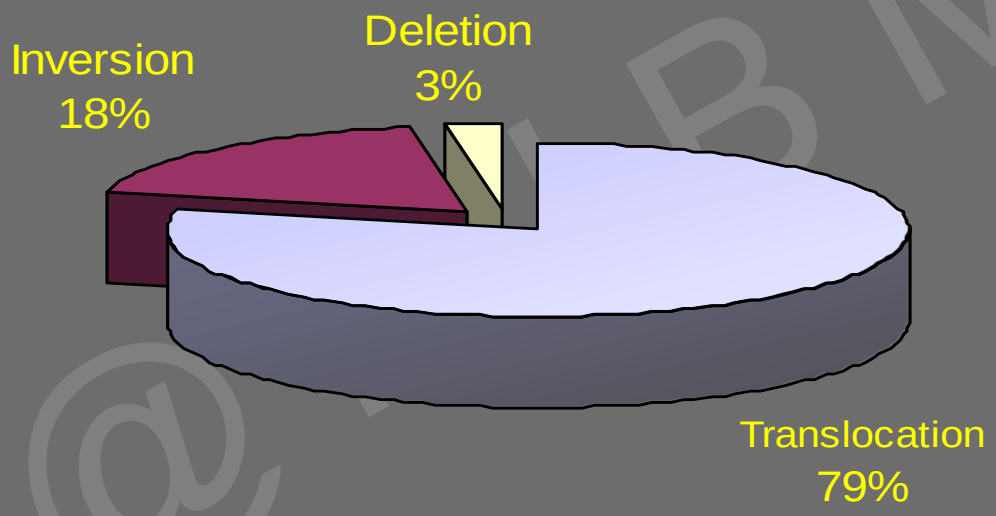
Grouped Bar chart



Bar chart

- Another common method for graphically presenting nominal and ordinal scaled data
- One bar is used to represent the frequency for each category
- The bars are usually positioned vertically with their bases located on the horizontal axis of the graph
  - One convention is to call vertical bar chart a ‘column chart’
- The bars are separated, and this is why such a graph is frequently used for nominal and ordinal data – the separation emphasize the plotting of frequencies for distinct categories

Pie chart



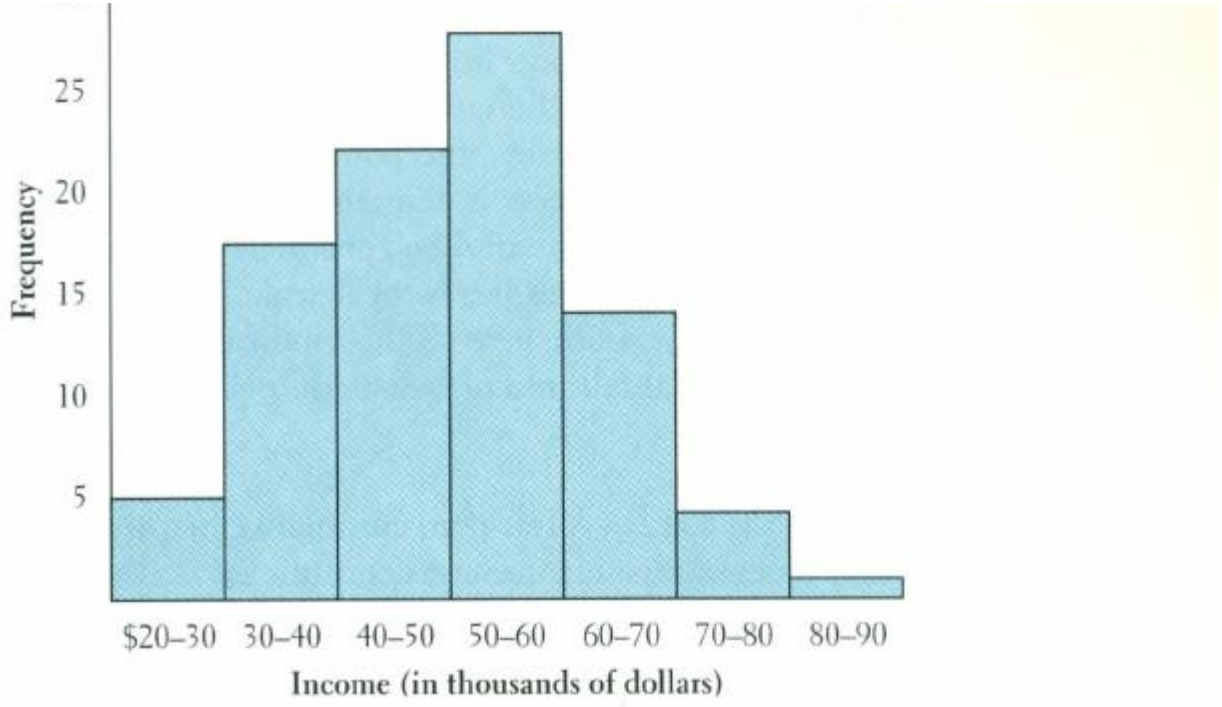
Pie Chart

- The pie chart is an effective way of displaying the percentage breakdown of data by category.
- Useful if the relative sizes of the data components are to be emphasized
- Pie charts also provide an effective way of presenting ratio- or interval-scaled data after they have been organized into categories

Histogram

- Frequently used to graphically present interval and ratio data
- Is often used for interval and ratio data
- The adjacent bars indicate that a numerical range is being summarized by indicating the frequencies in arbitrarily chosen classes

Histogram example



Time Series Graph

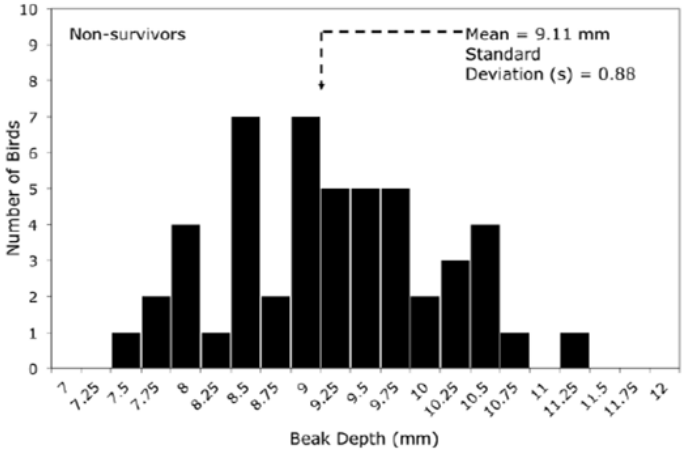
- The time series graph is a graph of data that have been measured over time.
- The horizontal axis of this graph represents time periods and the vertical axis shows the numerical values corresponding to these time periods

Time Series Example

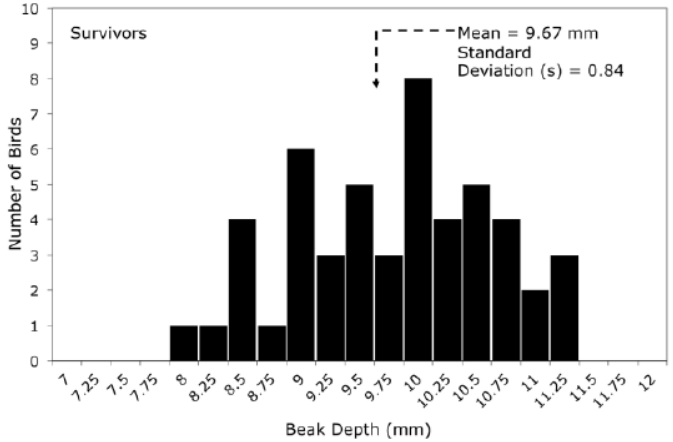
Measures of Average: Mean, Median, and Mode

In the two graphs below, the center and spread of each distribution is different. **The center of the distribution can be described by the mean, median, or mode.** ***These are referred to as measures of central tendency.***

Beak Depths of 50 Medium Ground Finches That Did Not Survive the Drought



Beak Depths of 50 Medium Ground Finches That Survived the Drought



Mean

Students in a biology class planted eight bean seeds in separate plastic cups and placed them under a bank of fluorescent lights. Fourteen days later, the students measured the height of the bean plants that grew from those seeds and recorded their results below:

Plant No.	1	2	3	4	5	6	7	8
Height (cm)	7.5	10.1	8.3	9.8	5.7	10.3	9.2	8.7

7.5 + 10.1 + 8.3 + 9.8 + 5.7 + 10.3 + 9.2 + 8.7 = **69.6 centimeters**

mean = **69.6 cm**/8 = **8.7 centimeters**

The mean for this sample of eight plants is 8.7 centimeters and serves as an *estimate for the true mean of the population* of bean plants growing under these conditions. In other words, if the students collected data from hundreds of plants and graphed the data, the center of the distribution **should be around 8.7 centimeters.**

Median

When the data are ordered from the largest to the smallest, the median is the midpoint of the data. *It is not distorted by extreme values, or even when the distribution is not normal.* For this reason, it may be more useful for you to use the median as the main descriptive statistic for a sample of data in which some of the measurements are extremely large or extremely small.

To determine the median of a set of values, you first arrange them in numerical order from lowest to highest. The middle value in the list is the median. If there is an even number of values in the list, then the median is the mean of the middle two values.

Median

A researcher studying mouse behavior recorded in Table 3 the time (in seconds) it took 13 different mice to locate food in a maze.

Length of Time for Mice to Locate Food in a Maze

Mouse No.	1	2	3	4	5	6	7	8	9	10	11	12	13
Time (sec.)	31	33	163	33	28	29	33	27	27	34	35	28	32

I. Arrange the time values in numerical order from lowest to highest:

27, 27, 28, 28, 29, 31, **32**, 33, 33, 33, 34, 35, 163

II. Find the middle value. This value is the median:

Median = 32 seconds

Median

A researcher studying mouse behavior recorded in Table 3 the time (in seconds) it took 13 different mice to locate food in a maze.

Length of Time for Mice to Locate Food in a Maze

Mouse No.	1	2	3	4	5	6	7	8	9	10	11	12	13
Time (sec.)	31	33	163	33	28	29	33	27	27	34	35	28	32

Median = 32 seconds

mean = **41 seconds**

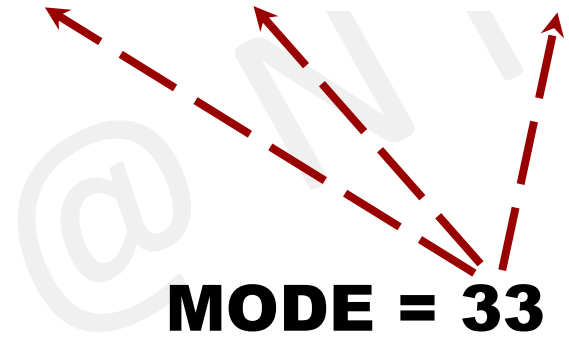
In this case, the **median is 32 seconds**, but the **mean is 41 seconds,** which is longer than all but one of the mice took to search for food. In this case, the mean would not be a good measure of central tendency unless the **really slow mouse** is excluded from the data set.

The mean is not the central tendency in this case because the values are not evenly distributed due to the large value in mouse number 3.

Mode

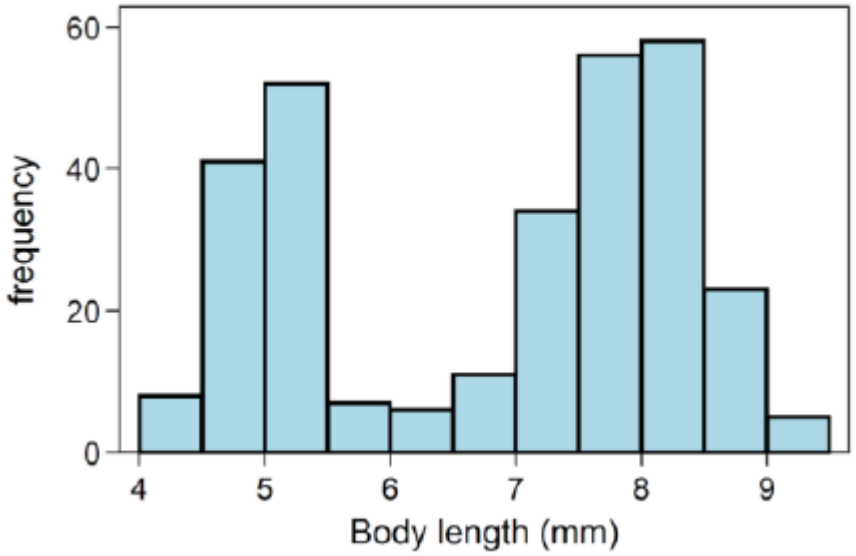
The mode is another measure of the average. It is the value that appears most often in a sample of data. In the example shown the mode is 33 seconds.

Mouse No.	1	2	3	4	5	6	7	8	9	10	11	12	13
Time (sec.)	31	33	163	33	28	29	33	27	27	34	35	28	32



Mode

The mode is not typically used as a measure of central tendency in biological research, but it can be useful in describing some distributions. Describing these data with a measure of central tendency like the mean or median would obscure this fact.



Clearly there is no central tendency here but in fact 2 central tendencies. The graph at the left shows a distribution of body lengths with two peaks, or modes—called a **bimodal distribution**.

Measures of Variability: Range, Standard Deviation, and Variance

Variability describes the extent to which numbers in a data set diverge from the central tendency. It is a measure of how "spread out" the data are from the mean.

A mean, mode, or median do not give a sense of how far apart the values are in the sample. The amount of variability (spread) is important to know because the amount of variability will let you analyze data more effectively by providing insight on the population that you gathered information on.

Range

The simplest measure of variability in a sample of normally distributed data is the range, which is the difference between the largest and smallest values in a set of data.

Students in a biology class measured the width in centimeters of eight leaves from eight different maple trees and recorded their results in the table below.

Width of Maple Tree Leaves

Plant No.	1	2	3	4	5	6	7	8
Width (cm)	7.5	10.1	8.3	9.8	5.7	10.3	9.2	8.7

I. Identify the largest and smallest values in the data set:

largest = 10.3 centimeters, smallest = 5.7 centimeters

II. To determine the range, subtract the smallest value from the largest value:

:

range = 10.3 centimeters – 5.7 centimeters = 4.6 centimeters

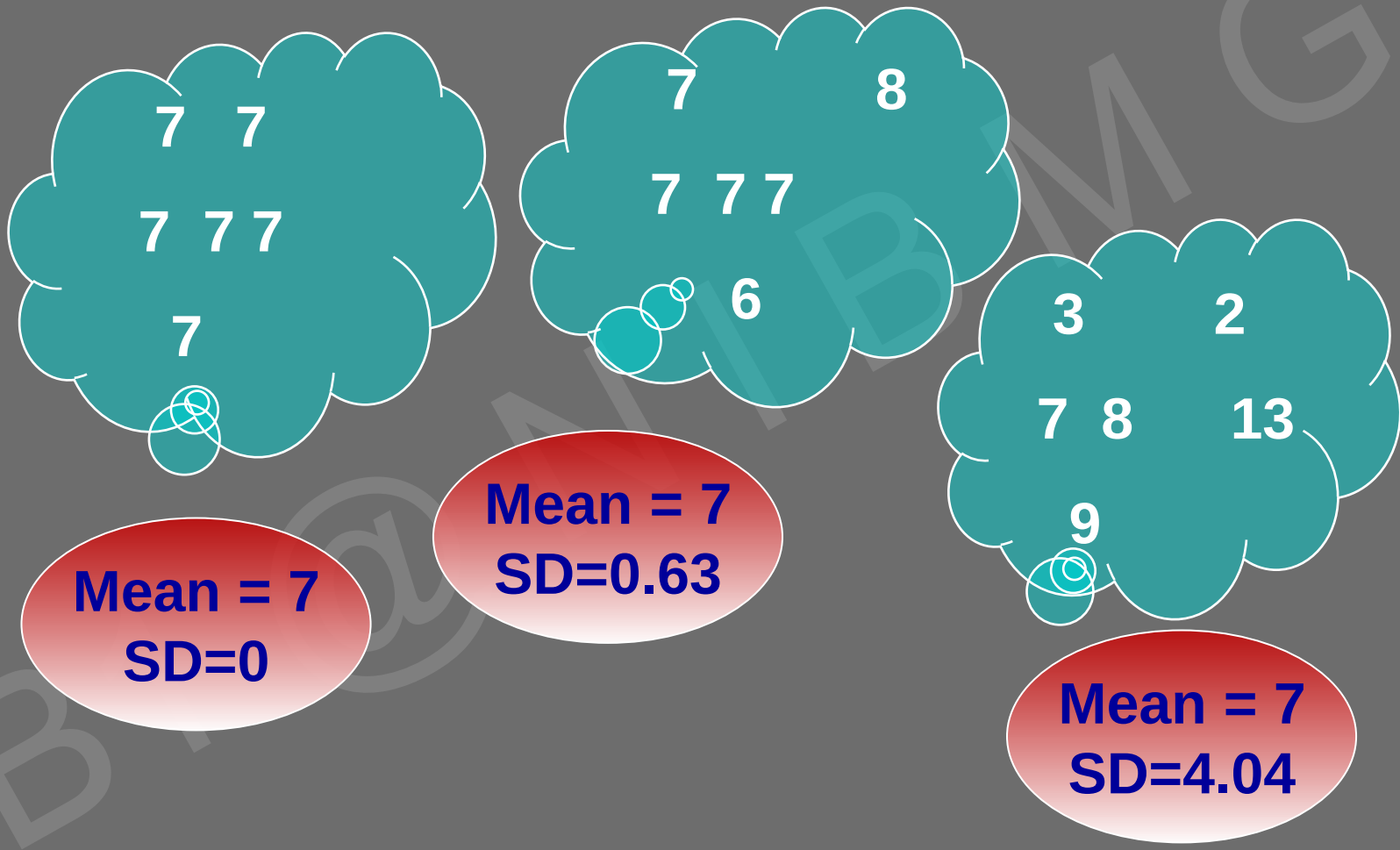
Range = 10.3 centimeters – 5.7 centimeters = **4.6 centimeters**

Width of Maple Tree Leaves

Plant No.	1	2	3	4	5	6	7	8
Width (cm)	7.5	10.1	8.3	9.8	5.7	10.3	9.2	8.7

A larger range value indicates a greater spread of the data—in other words, the larger the range, the greater the variability. *However, an extremely large or small value in the data set will make the variability appear high*. For example, if the maple leaf sample had not included the very small leaf number 5, the range would have been only 2.8 centimeters. **The variance and standard deviation** provides a more reliable measure of the "true" spread of the data.

Standard deviation SD



Variance (s^2)

Heights in Centimeters of Five Randomly Selected Pea Plants Grown at 8-10 °C

Plant	Height (cm)	Deviations from mean	Squares of deviation from mean
	(x_i)	$(x_i - x)$	$(x_i - x)^2$
A	10	2	4
B	7	-1	1
C	6	-2	4
D	8	0	0
E	9	1	1
	$\Sigma x_i = 40$	$\Sigma (x_i - x) = 0$	$\Sigma (x_i - x)^2 = 10$

X_i = score or value; X (bar) = mean; Σ = sum of

Notice that Deviations from the mean are just a subtraction from the mean (to give variance).

Then they are squared to make any negative values (below the mean) become positive values.

They are then added together.

Variance (s^2)

$$S^2 = \frac{\Sigma(X - \bar{X})^2}{n-1}$$

Heights in Centimeters of Five Randomly Selected Pea Plants Grown at 8-10 °C

Plant	Height (cm)	Deviations from mean	Squares of deviation from mean
	(x_i)	$(x_i - x)$	$(x_i - x)^2$
A	10	2	4
B	7	-1	1
C	6	-2	4
D	8	0	0
E	9	1	1
	$\Sigma x_i = 40$	$\Sigma (x_i - x) = 0$	$\Sigma (x_i - x)^2 = 10$

X_i = score or value; X (bar) = mean; Σ = sum of

There were 5 plants in the population thus n - 1 = 4

degrees of freedom

10/4 = **2.5 = Variance**

- the degree to which individual members within the sample vary from the mean

Mean, Var, SD

- **Average = Sample Mean(X):** $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

- **Sample Variance (X):** $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ = Average of Squared Deviations from Mean.

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$ = Average of Squares - Square of Average.

≥ 0 (Zero if and only if all are x_i are identical)

- Note a popular convention is to use the denominator (n-1) for sample variance, i.e.

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

- **Sample standard deviation of X** $= \sqrt{Sample\ Var(X)}$ has the same unit as X.

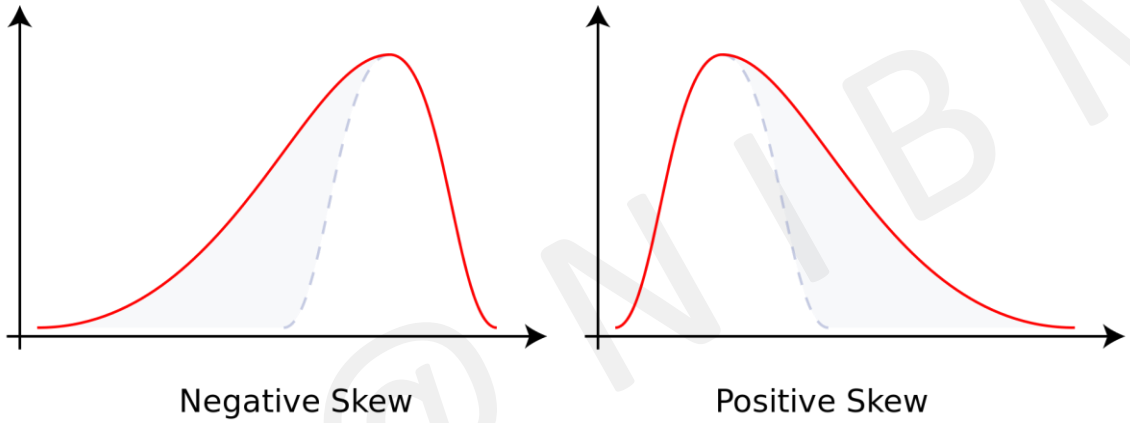
- Note variance does not change with "location shift." $Var(aX + b) = a^2 . Var(X)$

Quantiles

- **Quantiles:** α'th quantile of a data set or 100α 'th percentile is defined as any point with α proportion [or $(100 * \alpha)\ \%$] of the data below that. Example:
- 0'th quantile = Minimum of the data points.
- 0.25'th quantile = 1'st quartile = a point with ¼ th of the data below.
- 0.5'th quantile = Median = a point with half of the data below.
- 0.75'th quantile = 3'rd quartile = a point with ¾ th of the data below.
- 1'th quantile = Maximum of the data points.
- **Inter Quartile Range (IQR)** = Q3 – Q1 : a measure of spread.
- **Range:** Max – Min (Also a measure of spread).

Skewness

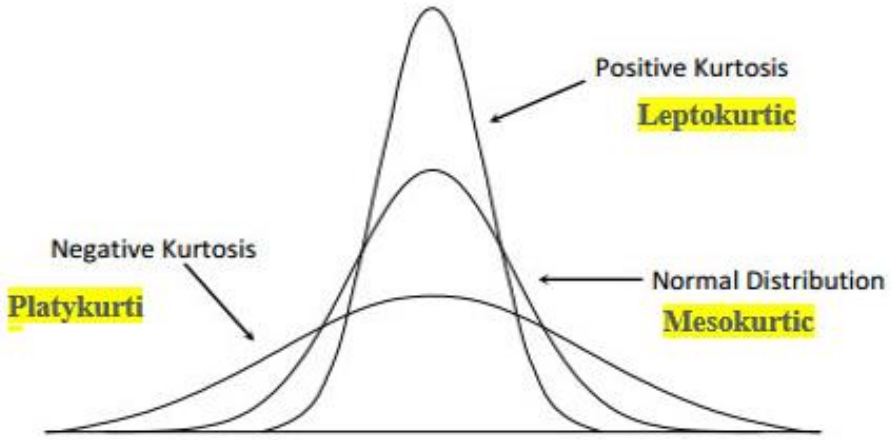
- **Skewness measures the asymmetry in a distribution.**
  - **Positive Skew (Right Skew): Smaller values are more likely. Tail on the right.**
  - **Negative Skew (Left Skew): Larger values are more likely. Tail on the left.**



- $Skewness = \frac{\mu_3}{\sigma^3}$, where $\mu_3 = Average\ of\ [X - \mu]^3$ and σ is the SD.

Skewness and Kurtosis

- **Kurtosis measures the 'peaked'ness vs 'flat'ness of a distribution.**



- $Kurtosis = \frac{\mu_4}{\sigma^4}$, where $\mu_4 = Average\ of\ [X - \mu]^4$ and σ is the SD.

- Excess Kurtosis = Kurtosis - 3 (where 3 is the kurtosis of normal distribution).

Bivariate Frequency Distribution Table

Table (IV): Distribution of 60 patients at the chest department of Alexandria hospital in May 2008 according to smoking & lung cancer

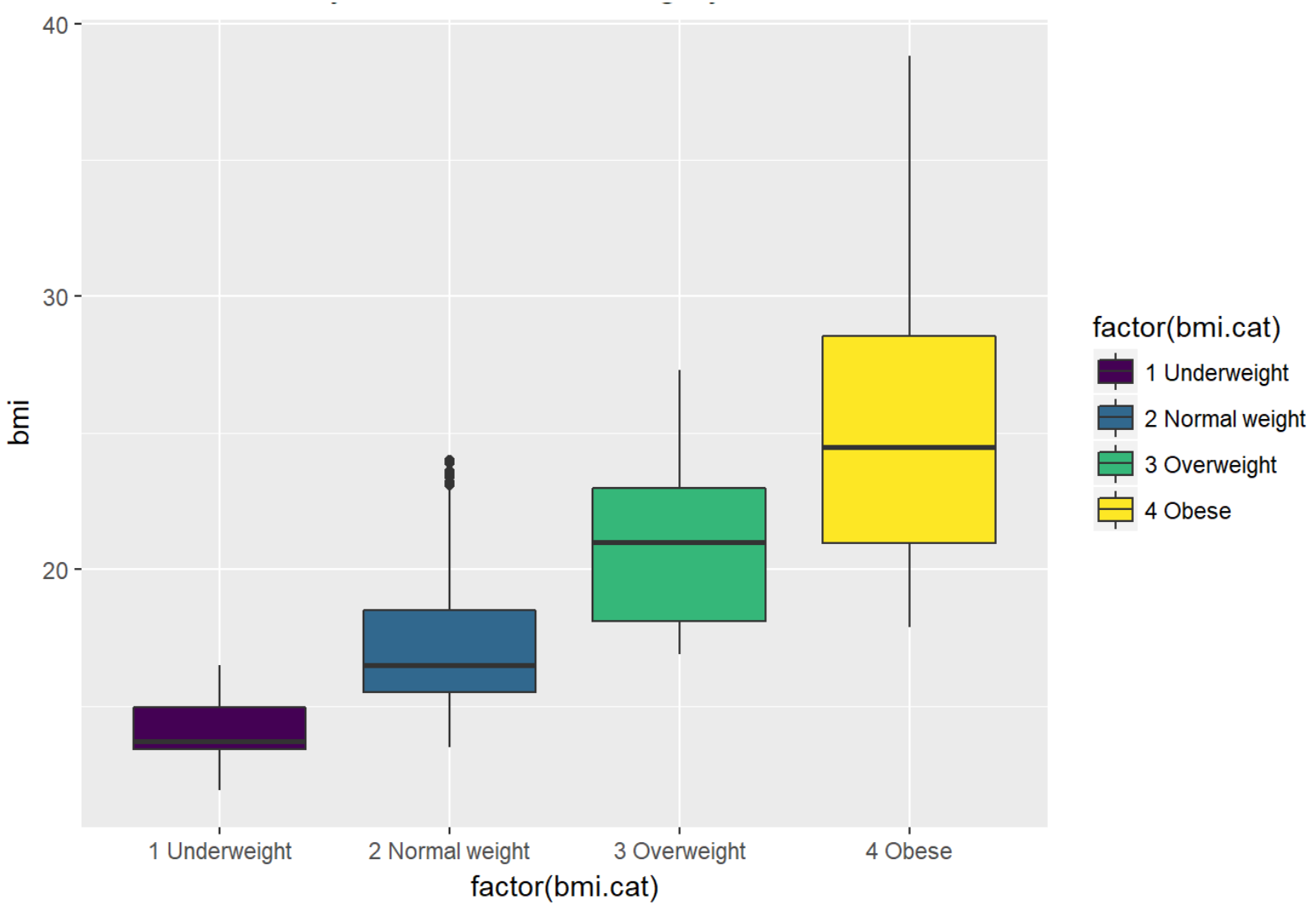
Smoking	Lung cancer				Total	
	positive		negative			
	No.	%	No.	%	No.	%
Smoker	15	65.2	8	34.8	23	100
Non smoker	5	13.5	32	86.5	37	100
Total	20	33.3	40	66.7	60	100

Bivariate Relationship (Contingency Table)

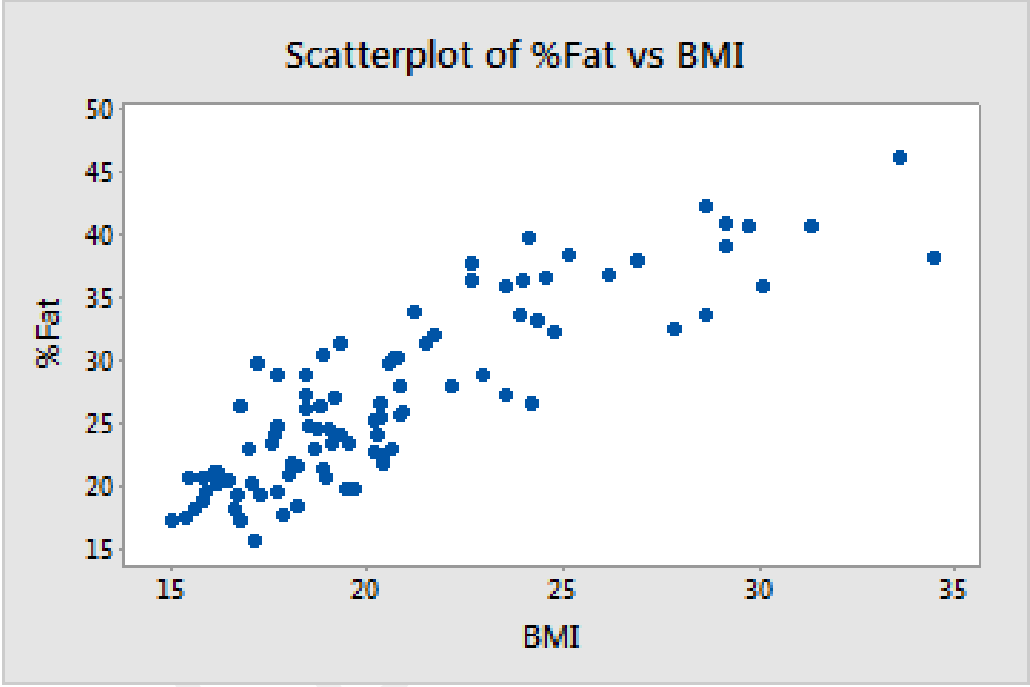
	Y=1	Y=0
X=1	a	b
X=0	c	D

- Strength of relationship between two dichotomous variables can be studied using
- Relative Risk or Risk Ratio =$\mathrm{RR} = \dfrac{\text{Risk of Y among X=1}}{\text{Risk of Y among X=0}} = \dfrac{a/(a+b)}{c/(c+d)}$
- Or
- Odds Ratio =$\mathrm{OR} = \dfrac{\text{Odds of Y among X=1}}{\text{Odds of Y among } X=0} = \dfrac{a/b}{c/d} = \dfrac{a*d}{b*c}$
- Note: Odds = Risk/(1-Risk) and Risk = Odds/(1+Odds).

Box Plot: Relationship between continuous and discrete variables



Bivariate Relationship (Continuous)



- Y= % Body Fat

  Outcome/response
- X = BMI

  Explanatory Variable

- Bivariate relationship can be visualized using a scatter plot.
- Strength of relationship and direction (increasing/decreasing) trend can be studied using covariance or 'correlation'.
- A trend line or curve can fitted to capture the pattern of change of Y with X.
- Fitting a trend line/curve helps to predict Y values for new observations for which X is known.

Covariance

- **Cov(X, Y)** $= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

  $=$ Average Of Product of Deviations from Mean

- Also **Cov(X, Y)** $= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$

  = Average of Products - Product of Averages.

- Note a population convention is to use the denominator (n-1) for covariance, i.e.

$$\textbf{Cov(X, Y)} \quad = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- This is also the default convention in R function cov() and excel function COVAR().
- Note covariance does not change with “location shift of x and/or y”
- $Cov(aX + b,\ cY + d) = a \,.\, c \,.\, Cov(X, Y)$

Pearson's Correlation Coefficient

- **Cor(X, Y)** $= \frac{Cov(X,Y)}{\sqrt{Var(X).Var(Y)}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{(x_i-\bar{x})^2}\sqrt{(y_i-\bar{y})^2}}$

- Note: Correlation does not change with "location shift of x and/or y" or even with "scaling of x and/or y". It is unit free.

- Correlation has the same sign as covariance. To obtain correlation, the covariance is scaled by the maximum possible magnitude of covariance,
  - $Cov(X,Y) \leq \sqrt{Var(X).Var(Y)}$

- *The maximum covariance* is achieved when Y and X are exactly linearly related.

- $-1 \leq Cor(X,Y) \leq 1$ [Correlation is 1 (or -1) if all points pass through a straight line with positive slope (or negative slope respectively).]

Best Fitting Line: Least Squares Methods

- Fit $y_i = a + bx_i + e_i$
- SSE = $\sum_i (y_i - a - bx_i)^2$
- By minimizing the sum of squared errors (SSE) we get **best fitting line with slope and intercept:** $\boldsymbol{b = \frac{Cov(X,Y)}{Var(X)}}$ **and** $\boldsymbol{a = \bar{y} - \hat{b}\bar{x}}$
- Note : $b = \frac{Cov(X,Y)}{Var(X)} = Cor(X,Y)\frac{SD(Y)}{SD(X)} = r\frac{s_Y}{s_X}$
- Therefore, slope of the best fitting line is related to the covariance/correlation (the direction is same). It is zero when the correlation or covariance is zero. Likewise correlation/covariance is zero whenever the best fitting line is horizontal (b=0).

Correlation: Linear Dependence

- Covariance and Correlation only capture linear dependence.
- If Y and X are related exactly through a sinusoidal (wave like) function or a circular pattern, we can have Cov=Cor=0.
- If Y is a strictly increasing function of X, e.g. Y=log(X) , correlation is not 1 (although the relation is perfect), because it is non-linear.
- One can use Rank Correlation to capture increasing/decreasing relationship (ignoring the actual gaps).

- **Spearman's Rank Correlation =** Pearson's Correlation Between Ranks of X and Y = $Cor(R, r)$.

$$R_i = Rank\ of\ x_i\ = position\ of\ x_i\ after\ sorting\ the\ x\ data\ in\ increasing\ order$$

$$r_i = Rank\ of\ y_i\ = position\ of\ y_i\ after\ sorting\ the\ y\ data\ in\ increasing\ order$$

- Spearman's rank correlation is 1 (or -1) when Y is strictly increasing (or decreasing) in X.

Inferential Statistics

- **Inferential statistics** are methods for using sample data to make general conclusions (inferences) about populations.
- Because a sample is typically only a part of the whole population, sample data provide only limited information about the population. As a result, sample statistics are generally imperfect representatives of the corresponding population parameters.

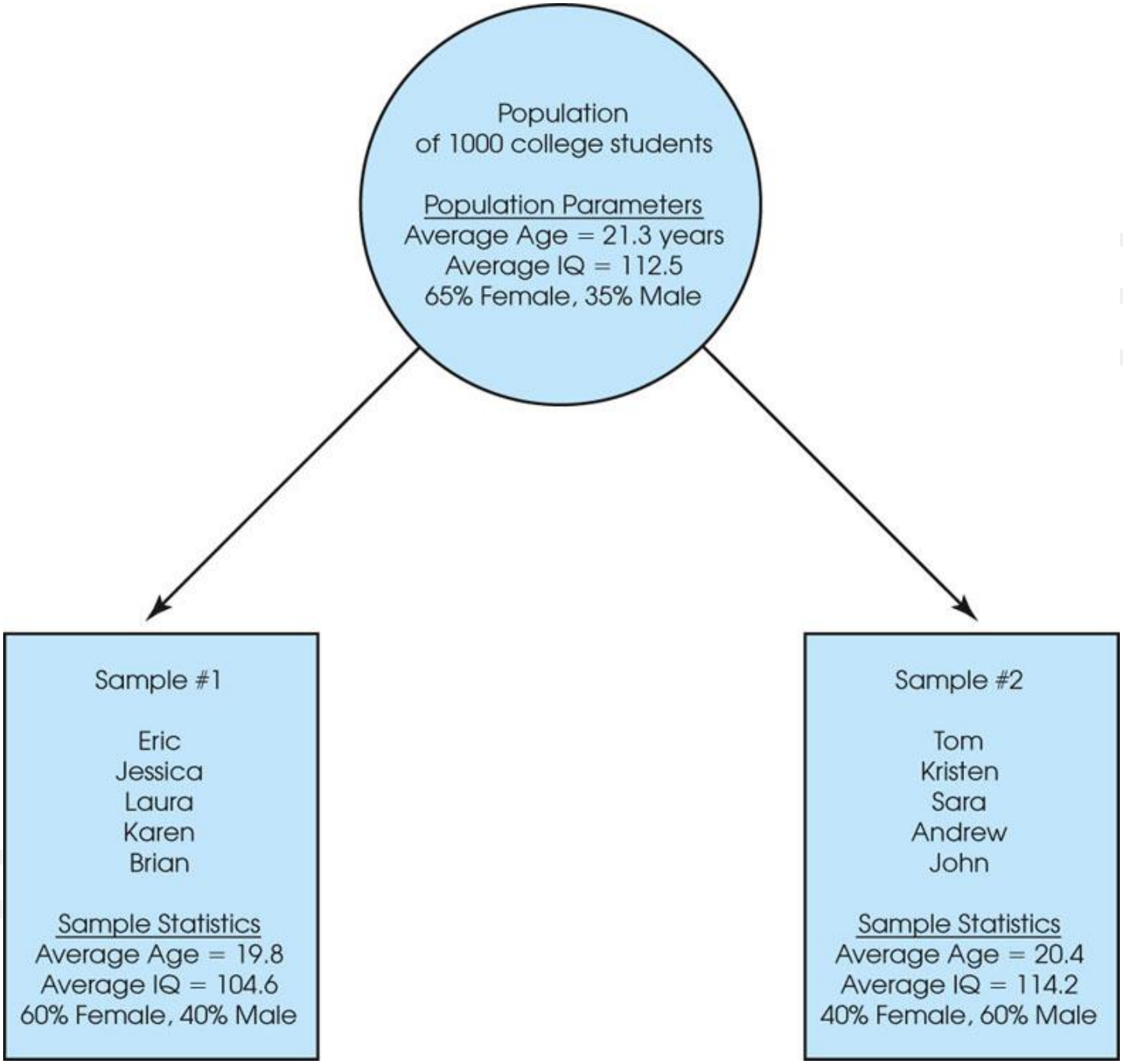
Inferential Statistics

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

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

Sampling Error

- The discrepancy between a sample statistic and its population parameter is called **sampling error**.
- Defining and measuring sampling error is a large part of inferential statistics.


Population
of 1000 college students
Population Parameters
Average Age = 21.3 years
Average IQ = 112.5
65% Female, 35% Male
Sample #1
Eric
Jessica
Laura
Karen
Brian
Sample Statistics
Average Age = 19.8
Average IQ = 104.6
60% Female, 40% Male
Sample #2
Tom
Kristen
Sara
Andrew
John
Sample Statistics
Average Age = 20.4
Average IQ = 114.2
40% Female, 60% Male

Probability Theory

Probability – Models for random phenomena

Random phenomena

- Unable to predict the outcomes, but in the long-run, the outcomes exhibit statistical regularity.

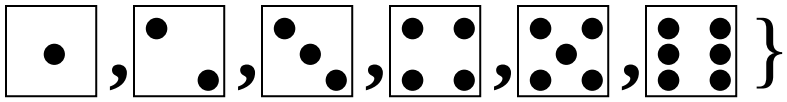
Examples

1. Tossing a coin – outcomes S ={**Head, Tail**}

   Unable to predict on each toss whether is Head or Tail.

   In the long run can predict that 50% of the time heads will occur and 50% of the time tails will occur

2. Rolling a die – outcomes

$S = \{$  $\}$

Unable to predict outcome but in the long run can one can determine that each outcome will occur 1/6 of the time.

Use symmetry. Each side is the same. One side should not occur more frequently than another side in the long run. If the die is not balanced this may not be true.

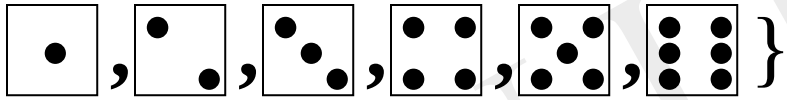
The sample Space, *S*

The **sample space**, *S*, for a random phenomena is the set of all possible outcomes.

Examples

1. Tossing a coin – outcomes S ={**Head, Tail**}

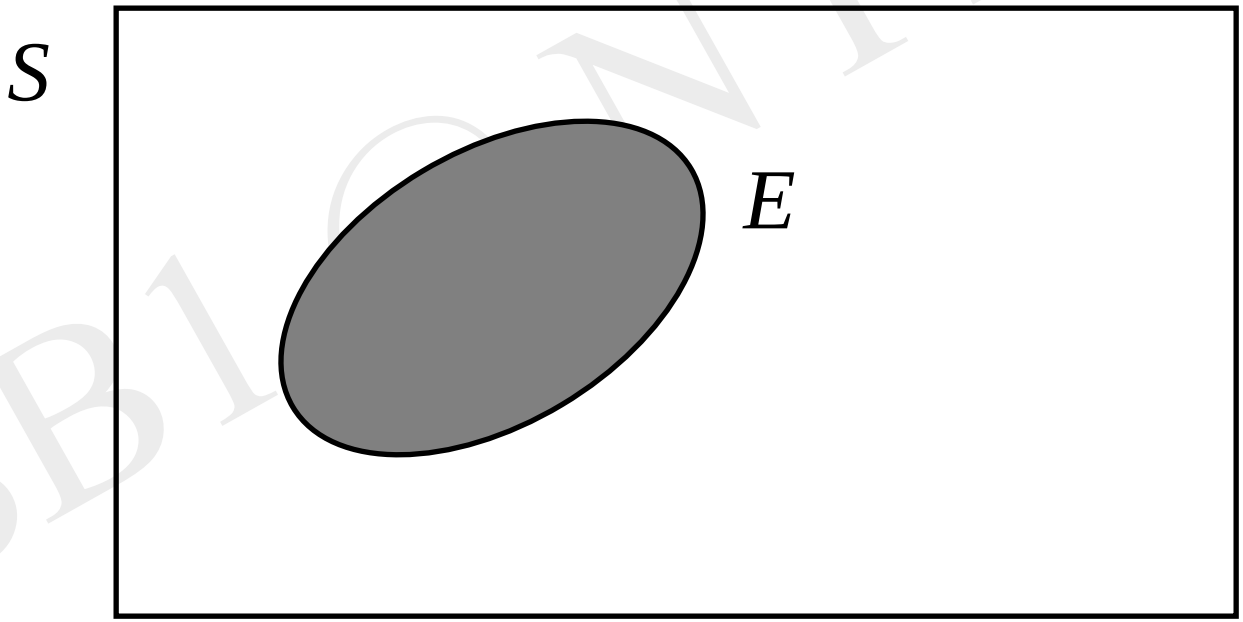
2. Rolling a die – outcomes

S ={  }

={1, 2, 3, 4, 5, 6}

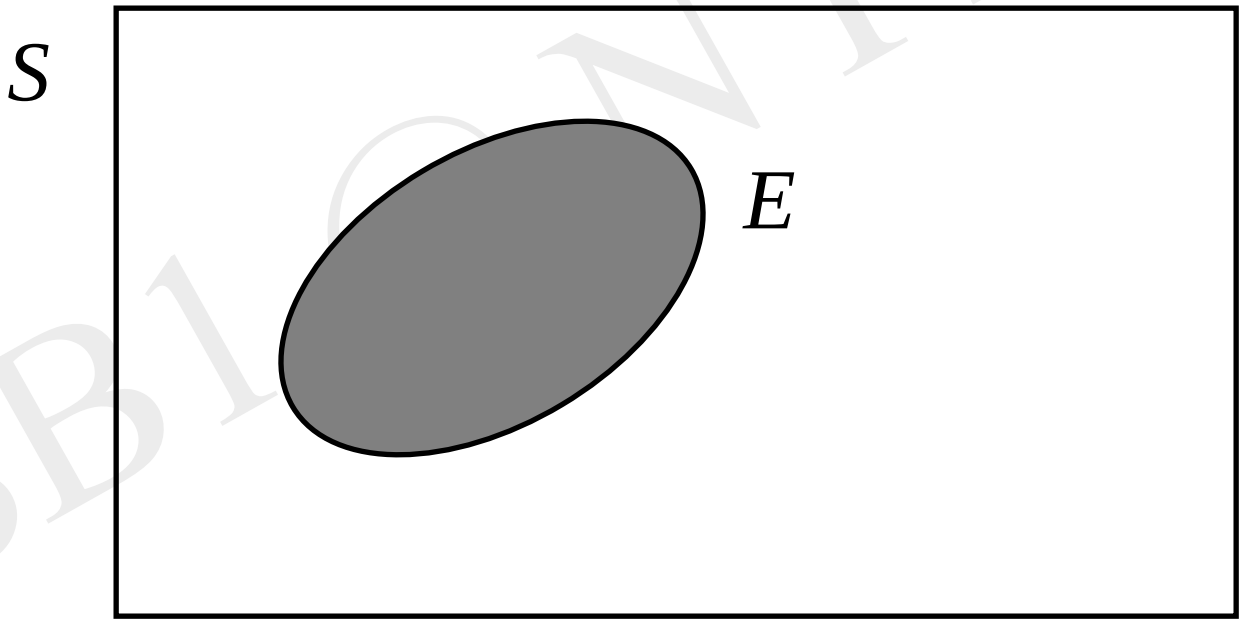
An Event , *E*

The **event**, *E*, is any subset of the **sample space**, *S*. i.e. any set of outcomes (not necessarily all outcomes) of the random phenomena



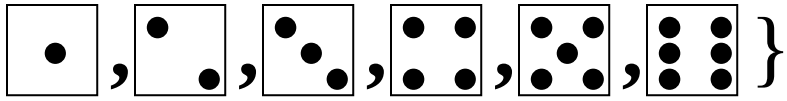
Venn diagram

The **event**, *E*, is said to **have occurred** if after the outcome has been observed the outcome lies in *E*.



Examples

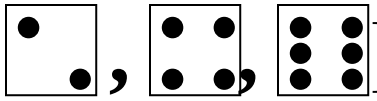
1. Rolling a die – outcomes

S = {  }

= {1, 2, 3, 4, 5, 6}

E = the event that an even number is rolled

= {2, 4, 6}

= {  }

Special Events

The Null Event, The empty event - ϕ

ϕ = { } = the event that contains no outcomes

The Entire Event, The Sample Space - S

S = the event that contains all outcomes

The empty event, ϕ , never occurs.

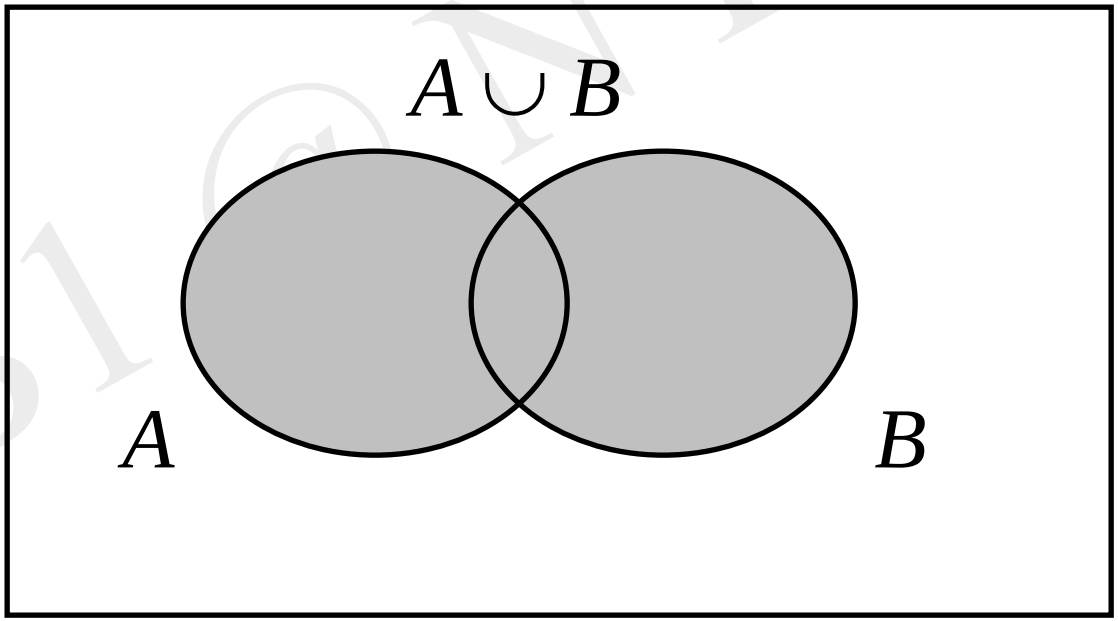
The entire event, S, always occurs.

Set operations on Events

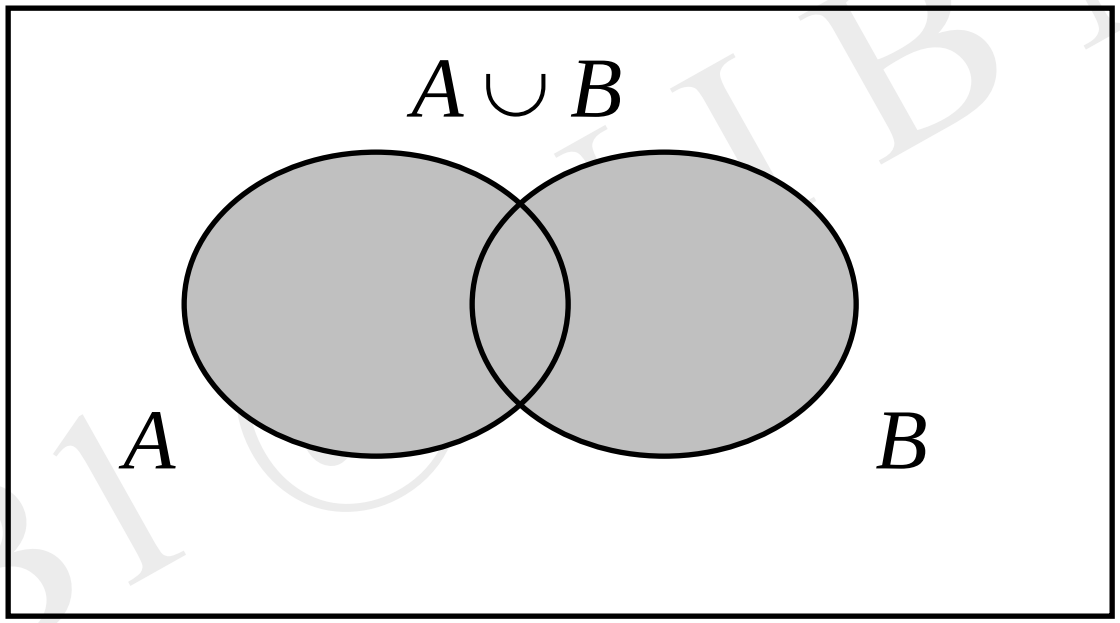
Union

Let A and B be two events, then the **union** of A and B is the event (denoted by $A \cup B$) defined by:

$$A \cup B = \{e |\ e \text{ belongs to } A \textbf{ or } e \text{ belongs to } B\}$$



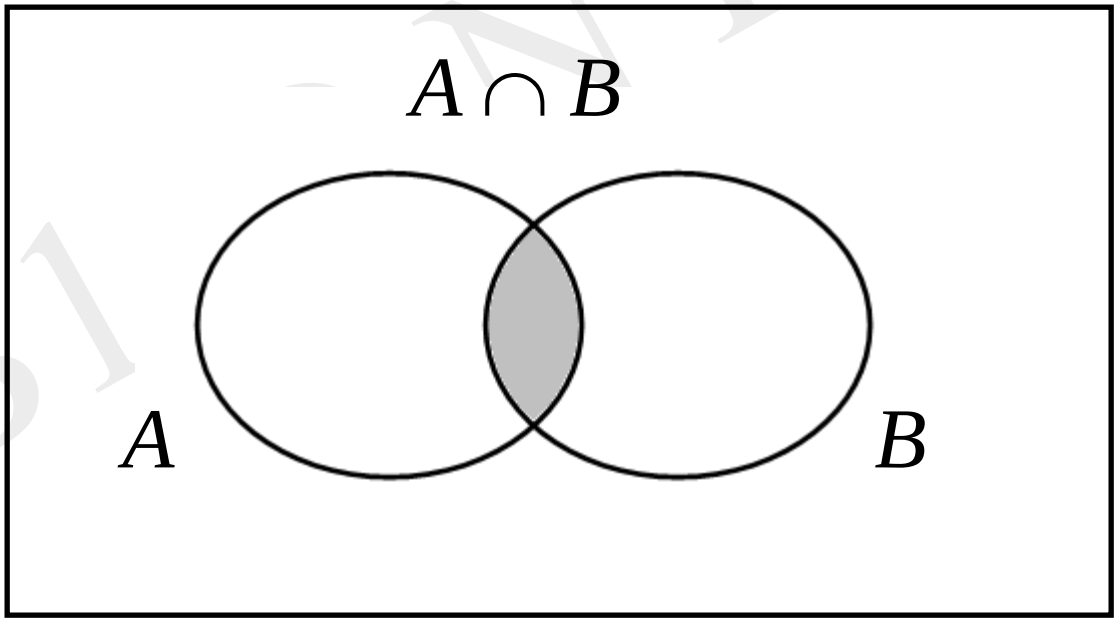
The event $A \cup B$ **occurs** if the event A **occurs or** the event and B **occurs .**



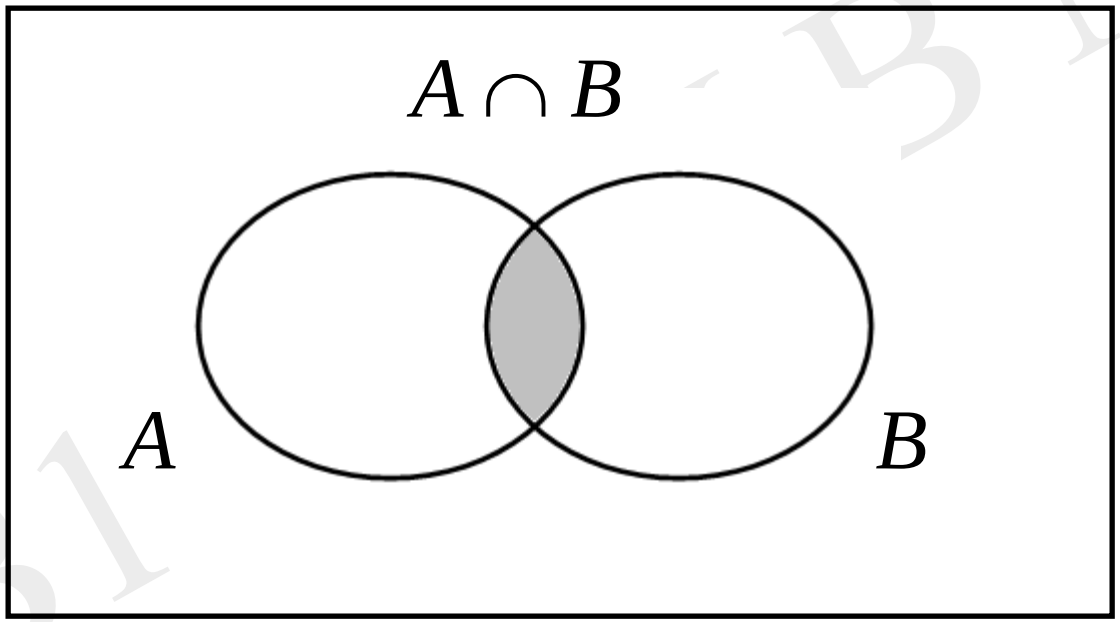
Intersection

Let A and B be two events, then the **intersection** of A and B is the event (denoted by $A \cap B$) defined by:

$$A \cap B = \{e |\ e \text{ belongs to } A \textbf{ and } e \text{ belongs to } B\}$$



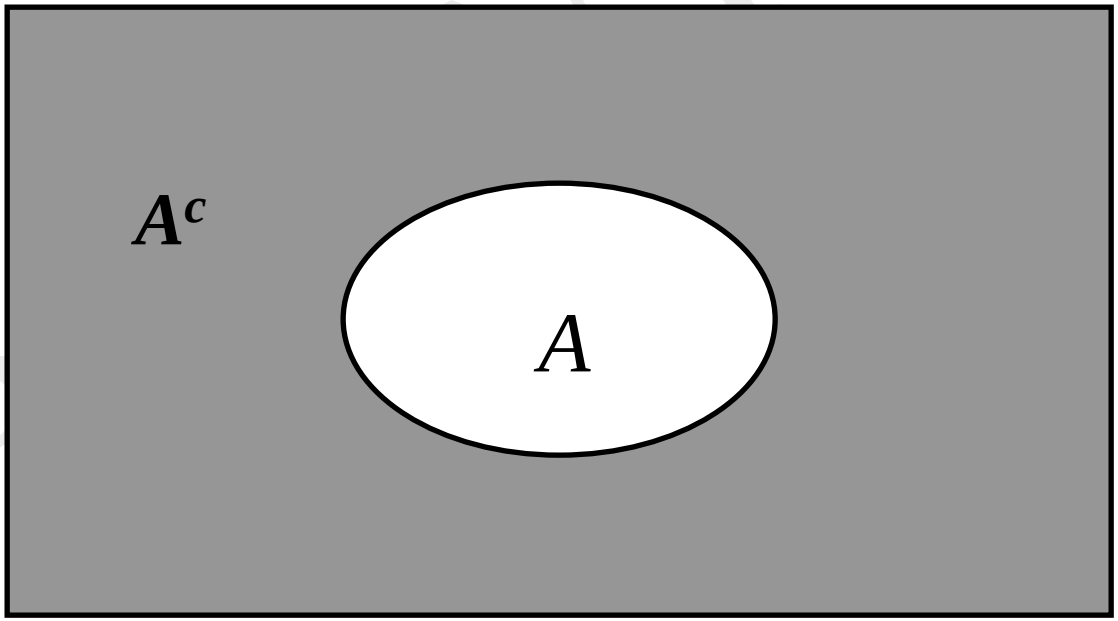
The event $A \cap B$ **occurs** if the event A **occurs and** the event and B **occurs** .



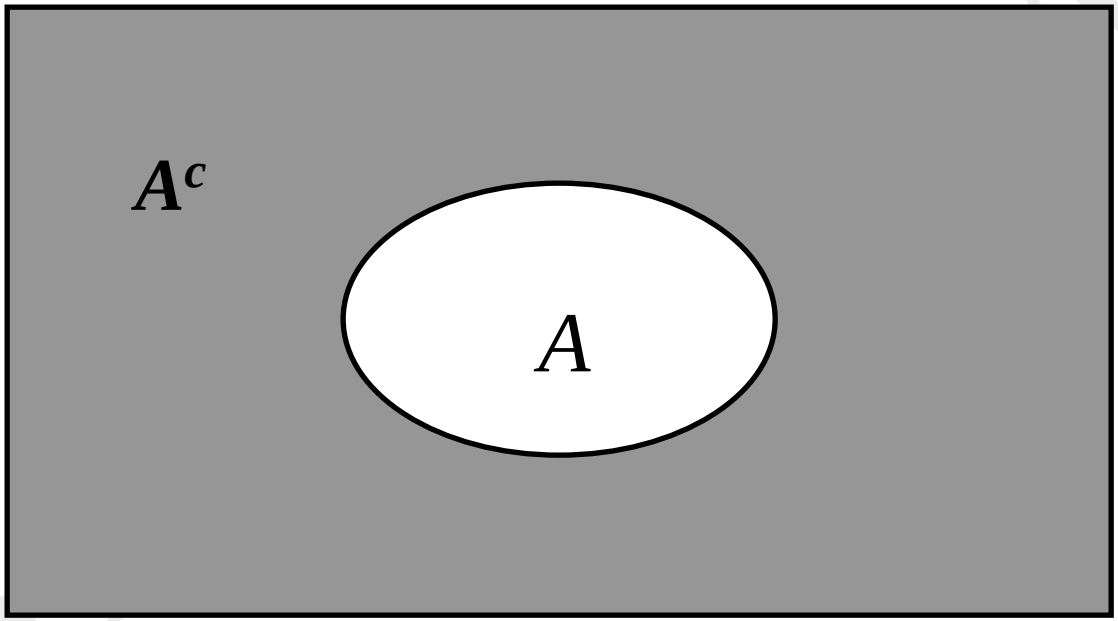
Complement

Let A be any event, then the **complement** of A (denoted by $\bar{A}$ or A^c) defined by:

$$A^c = \{e | \ e \textbf{ does not belong to } A\}$$



The event A^c **occurs** if the event A **does not occur**



In problems you will recognize that you are working with:

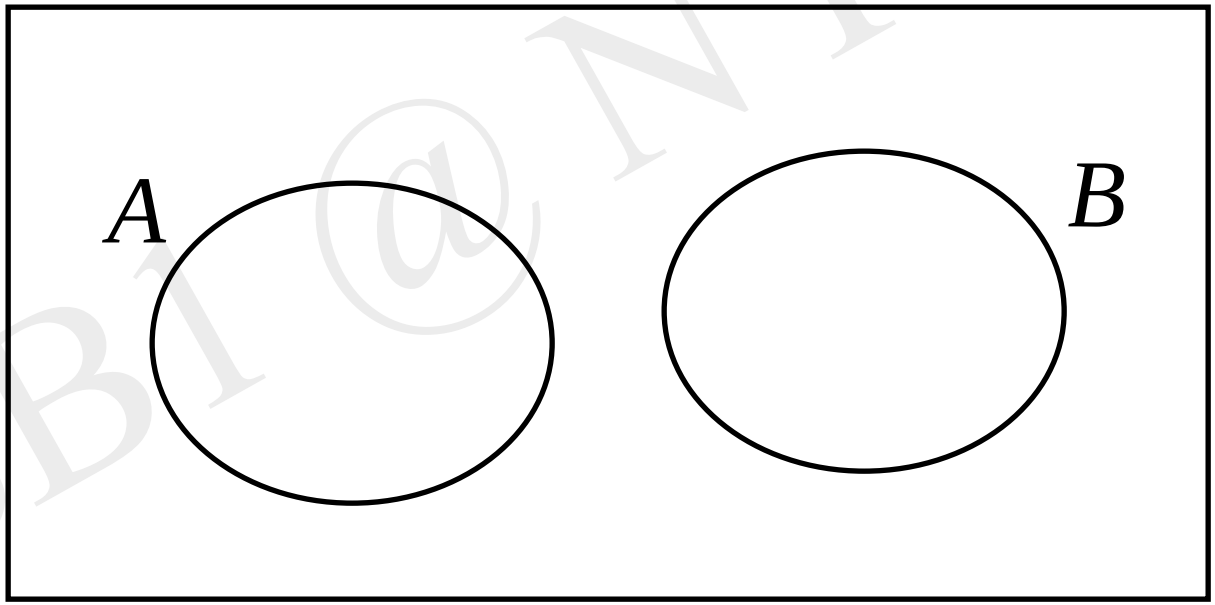
1. **Union** if you see the word **or**,
2. **Intersection** if you see the word **and**,
3. **Complement** if you see the word **not**.

*** Note:** The complement of "at least one" is "none" and complement of "at least two" is "less than 2" and so on. This is often useful to calculate probabilities through the complementary event.

Definition: mutually exclusive

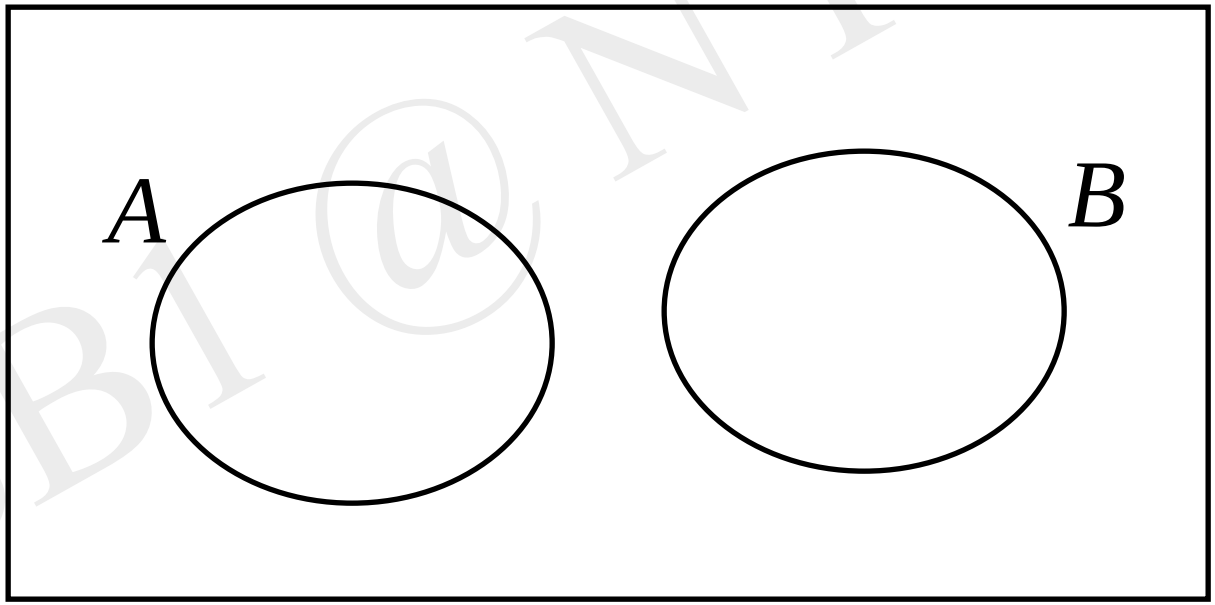
Two events A and B are called **mutually exclusive** or **disjoint** if:

$$A \cap B = \phi$$



If two events *A* and *B* are are **mutually exclusive** then:

1. They have no outcomes in common.
   They can't occur at the same time. The outcome of the random experiment can not belong to both *A* and *B*.



Probability

Definition: probability of an Event *E*.

Suppose that the sample space $S = \{o_1, o_2, o_3, \ldots o_N\}$ has a finite number, N, of oucomes.

Also, each of the outcomes is equally likely (because of symmetry).

Then for any event E

$$P[E] = \frac{n(E)}{n(S)} = \frac{n(E)}{N} = \frac{\text{no. of outcomes in } E}{\text{total no. of outcomes}}$$

Note: the symbol $n(A) =$ no. of elements of A

Techniques for counting

Rule 1

Suppose we carry out have a sets A_1, A_2, A_3, … and that any pair are mutually exclusive

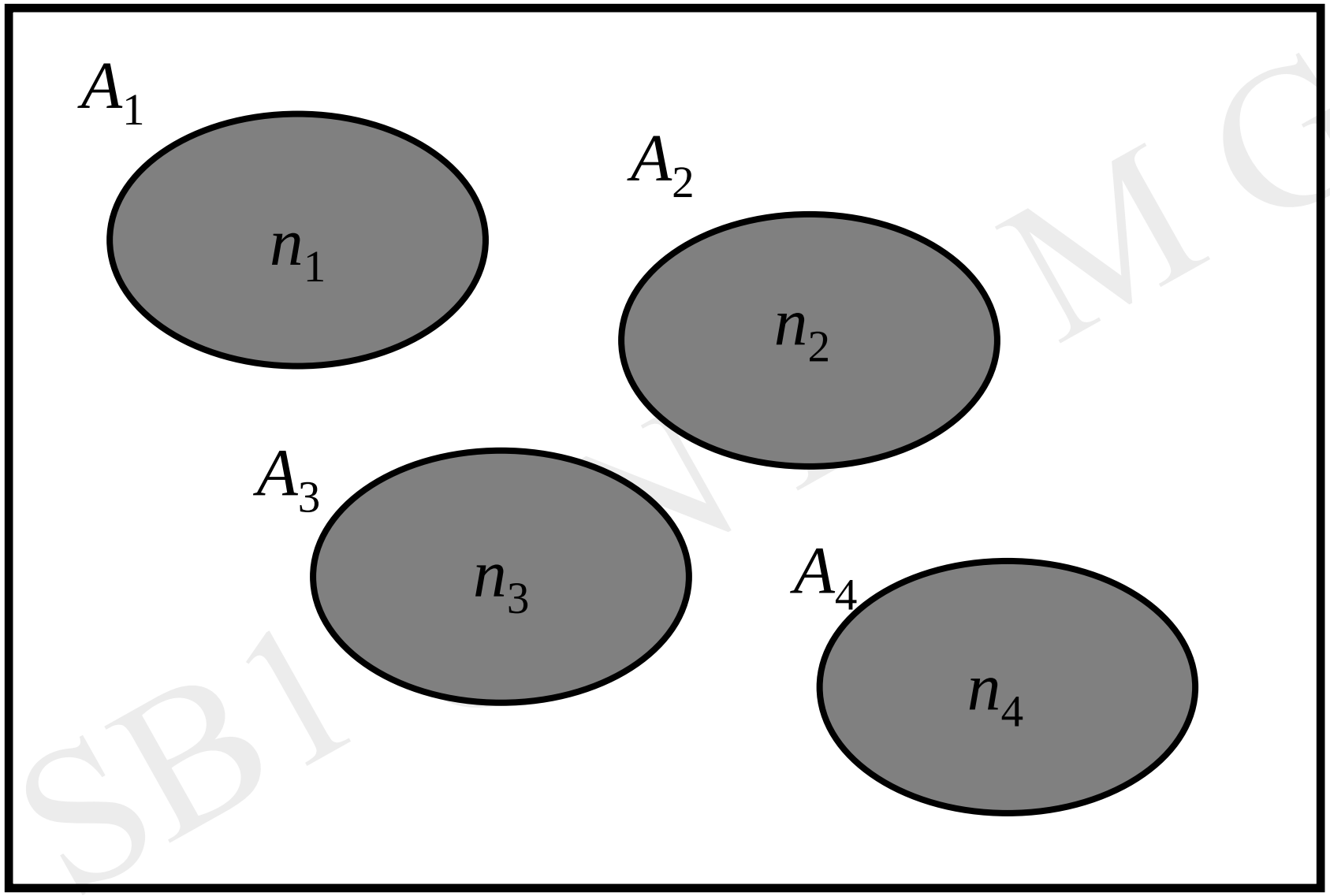
(i.e. $A_1 \cap A_2 = \phi$) Let

$n_i = n\,(A_i)$ = the number of elements in A_i.

Let $A = A_1 \cup A_2 \cup A_3 \cup \ldots$.

Then $N = n(\,A\,)$ = the number of elements in A

$= n_1 + n_2 + n_3 + \ldots$


A_1
n_1
A_2
n_2
A_3
n_3
A_4
n_4

Rule 2

Suppose we carry out two operations in sequence

Let

n_1 = the number of ways the first operation can be performed

n_2 = the number of ways the second operation can be performed once the first operation has been completed.

Then $N = n_1 \, n_2$ = the number of ways the two operations can be performed in sequence.

The Multiplicative Rule of Counting

Suppose we carry out k operations in sequence

Let

n_1 = the number of ways the first operation can be performed

n_i = the number of ways the i^{th} operation can be performed once the first $(i - 1)$ operations have been completed. $i = 2, 3, \ldots, k$

Then $N = n_1 n_2 \ldots n_k$ = the number of ways the k operations can be performed in sequence.

Examples

1. **Permutations:** How many ways can you order n objects

Solution:

Ordering n objects is equivalent to performing n operations in sequence.

1. Choosing the first object in the sequence ($n_1 = n$)
2. Choosing the 2^{nd} object in the sequence ($n_2 = n - 1$).

   …

k. Choosing the k^{th} object in the sequence ($n_k = n - k + 1$)

   …

n. Choosing the n^{th} object in the sequence ($n_n = 1$)

The total number of ways this can be done is:

$$N = n(n-1)\ldots(n-k+1)\ldots(3)(2)(1) = n!$$

Example How many ways can you order the 4 objects $\{A, B, C, D\}$

Solution:

$N = 4! = 4(3)(2)(1) = 24$

Here are the orderings.

ABCD	*ABDC*	*ACBD*	*ACDB*	*ADBC*	*ADCB*
BACD	*BADC*	*BCAD*	*BCDA*	*BDAC*	*BDCA*
CABD	*CADB*	*CBAD*	*CBDA*	*CDAB*	*CDBA*
DABC	*DACB*	*DBAC*	*DBCA*	*DCAB*	*DCBA*

Examples - continued

2. Permutations of size *k* (< *n*): How many ways can you choose *k* objects from *n* objects in a specific order

Solution: This operation is equivalent to performing *k* operations in sequence.

1. Choosing the first object in the sequence ($n_1 = n$)
2. Choosing the 2^{nd} object in the sequence ($n_2 = n$ -1).

   …

k. Choosing the k^{th} object in the sequence ($n_k = n - k + 1$)

The total number of ways this can be done is:

$$N = n(n-1)\ldots(n-k+1) = n!/\,(n-k)!$$

This number is denoted by the symbol

$$ {}_nP_k = n(n-1)\ldots(n-k+1) = \frac{n!}{(n-k)!} $$

Definition: $0! = 1$

This definition is consistent with

$$ {}_nP_k = n(n-1)\ldots(n-k+1) = \frac{n!}{(n-k)!} $$

for $k = n$

$$ {}_nP_n = \frac{n!}{0!} = \frac{n!}{1} = n! $$

Example How many permutations of size 3 can be found in the group of 5 objects {*A*, *B*, *C*, *D*, *E*}

Solution: $${}_5P_3 = \frac{5!}{(5-3)!} = 5(4)(3) = 60$$

ABC	*ABD*	*ABE*	*ACD*	*ACE*	*ADE*	*BCD*	*BCE*	*BDE*	*CDE*
ACB	*ADB*	*AEB*	*ADC*	*AEC*	*AED*	*BDC*	*BEC*	*BED*	*CED*
BAC	*BAD*	*BAE*	*CAD*	*CAE*	*DAE*	*CBD*	*CBE*	*DBE*	*DCE*
BCA	*BDA*	*BEA*	*CDA*	*CEA*	*DEA*	*CDB*	*CEB*	*DEB*	*DEC*
CAB	*DAB*	*EAB*	*DAC*	*EAC*	*EAD*	*DBC*	*EBC*	*EBD*	*ECD*
CAB	*DBA*	*EBA*	*DCA*	*ECA*	*EDA*	*DCB*	*ECB*	*EDB*	*EDC*

Example We have a committee of $n = 10$ people and we want to choose a **chairperson**, **a vice-chairperson** and a **treasurer**. Suppose that 6 of the members of the committee are male and 4 of the members are female. What is the probability that the three executives selected are all male?

Solution: Again, we want to select 3 persons from the committee of 10 in a specific order. (Permutations of size 3 from a group of 10).The total number of ways that this can be done is:

$$ {}_{10}P_3 = \frac{10!}{(10-3)!} = \frac{10!}{7!} = 10(9)(8) = 720 $$

This is the size, $N = n(S)$, of the sample space S. Assume all outcomes in the sample space are equally likely.

Let E be the event that all three executives are male

$$ n(E) = {}_6P_3 = \frac{6!}{(6-3)!} = \frac{6!}{3!} = 6(5)(4) = 120 $$

Hence

$$P[E] = \frac{n(E)}{n(S)} = \frac{120}{720} = \frac{1}{6}$$

Thus if all candidates are equally likely to be selected to any position on the executive then the probability of selecting an all male executive is:

$$\frac{1}{6}$$

Examples - continued

3. **Combinations of size k ($\leq n$):** A combination of size k chosen from n objects is a subset of size k where the order of selection is irrelevant. How many ways can you choose a combination of size k objects from n objects (order of selection is irrelevant)

Here are the combinations of size 3 selected from the 5 objects $\{A, B, C, D, E\}$

$\{A,B,C\}$	$\{A,B,D\}$	$\{A,B,E\}$	$\{A,C,D\}$	$\{A,C,E\}$
$\{A,D,E\}$	$\{B,C,D\}$	$\{B,C,E\}$	$\{B,D,E\}$	$\{C,D,E\}$

Important Notes

1. In **combinations** ordering is **irrelevant**. Different orderings result in the same combination.
2. In **permutations** order is **relevant**. Different orderings result in the different permutations.

How many ways can you choose a combination of size k objects from n objects (order of selection is irrelevant)

Solution: Let n_1 denote the number of combinations of size k. One can construct a permutation of size k by:

1. Choosing a combination of size k (n_1 = unknown)
2. Ordering the elements of the combination to form a permutation ($n_2 = k!$)

Thus $${}_nP_k = \frac{n!}{(n-k)!} = n_1 k!$$

and $$n_1 = \frac{{}_nP_k}{k!} = \frac{n!}{(n-k)!k!} =$$ the # of combinations of size k.

The number:

$$n_1 = \frac{{}_nP_k}{k!} = \frac{n!}{(n-k)!k!} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k(k-1)(k-2)\cdots(1)}$$

is denoted by the symbol

$${}_nC_k \quad \text{or} \quad \binom{n}{k} \quad \text{read “}n\text{ choose }k\text{”}$$

It is the number of ways of choosing k objects from n objects (order of selection irrelevant).

The above definition of $P[E]$, i.e.

$$P[E] = \frac{n(E)}{n(S)} = \frac{n(E)}{N} = \frac{\text{no. of outcomes in } E}{\text{total no. of outcomes}}$$

Applies only to the special case when

1. The sample space has a finite no.of outcomes, and
2. Each outcome is equi-probable
3. Probability calculation reduces to counting the number of outcomes satisfying the event.

If this is not true a more general definition of probability is required.

Rules of Probability

Basic Rules

$$P(\phi) = P(Null\ Event) = 0$$

$$\mathrm{P(S)=P}(Sure\ Event) = 1$$

$$0 \leq P(E) \leq 1\ , for\ all\ E \subset S$$

Rule The additive rule (Mutually exclusive events)

$$P[A \cup B] = P[A] + P[B]$$

i.e.

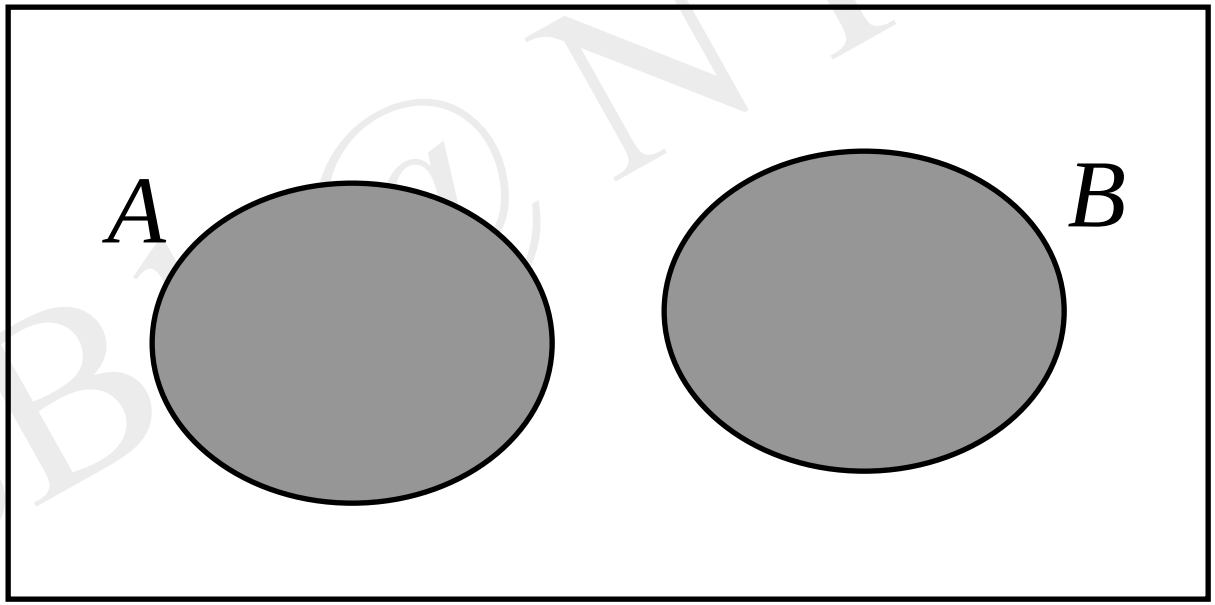
$$P[A \text{ or } B] = P[A] + P[B]$$

if $A \cap B = \phi$

(*A* and *B* mutually exclusive)

If two events *A* and *B* are are **mutually exclusive** then:

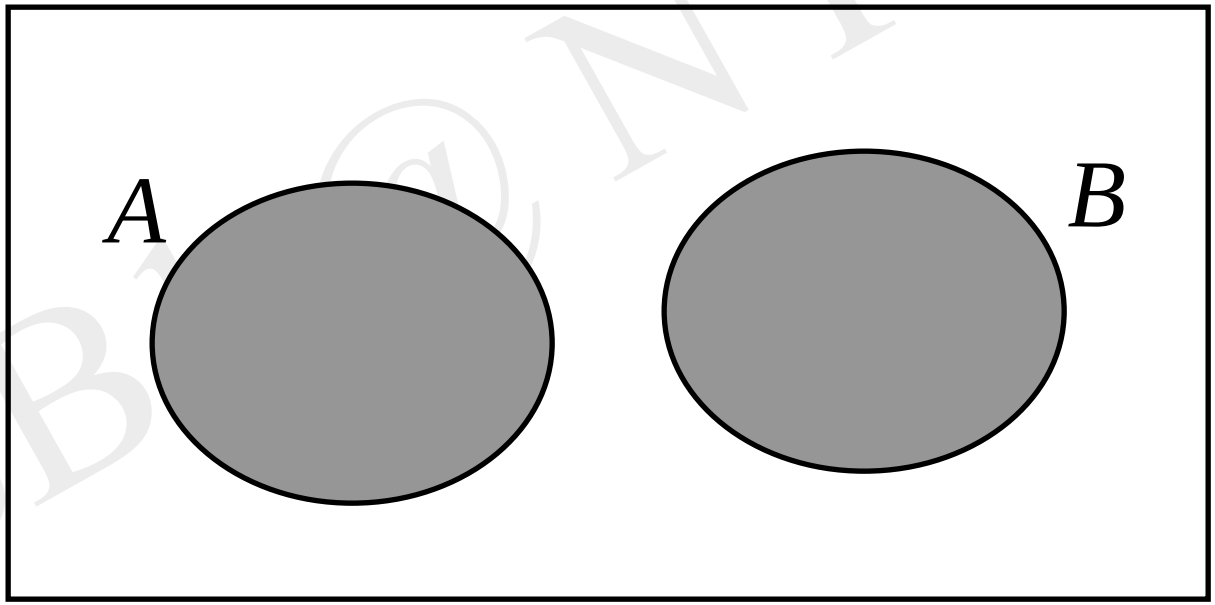
1. They have no outcomes in common.
   They can't occur at the same time. The outcome of the random experiment can not belong to both *A* and *B*.



$$P[A \cup B] = P[A] + P[B]$$

i.e.

$$P[A \text{ or } B] = P[A] + P[B]$$



Rule The additive rule
(In general)

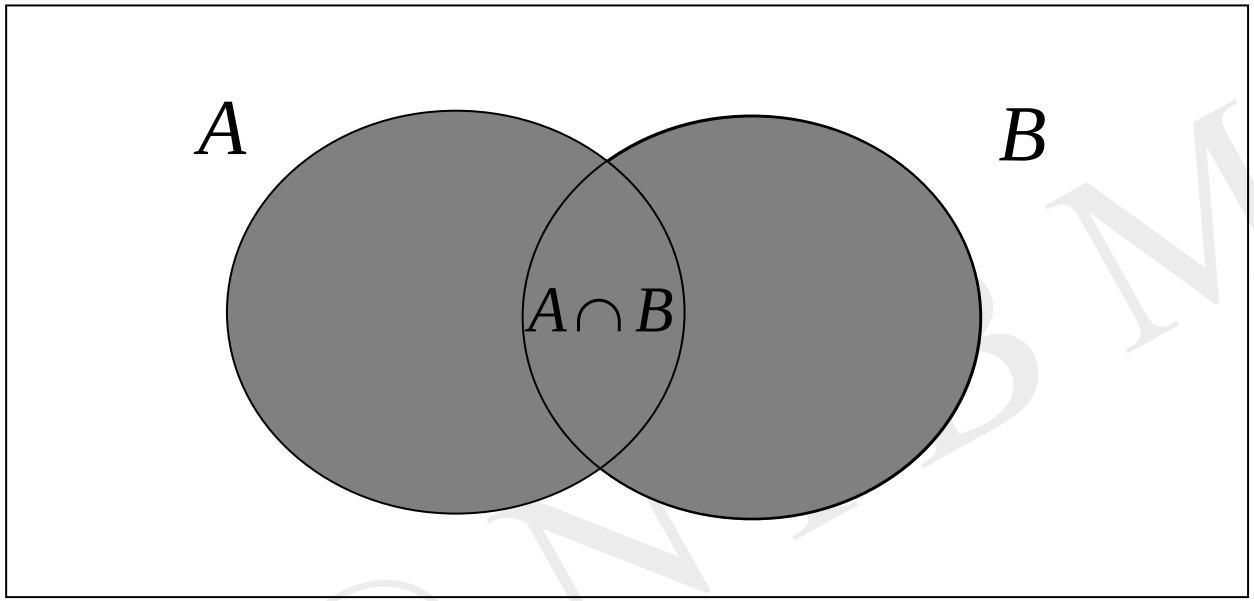
$$P[A \cup B] = P[A] + P[B] - P[A \cap B]$$

or

$$P[A \text{ or } B] = P[A] + P[B] - P[A \text{ and } B]$$

Logic

$A \cup B$



When $P[A]$ is added to $P[B]$ the outcome in $A \cap B$ are counted twice

hence

$$P[A \cup B] = P[A] + P[B] - P[A \cap B]$$

Rule for complements

2. $P\left[\bar{A}\right] = 1 - P\left[A\right]$

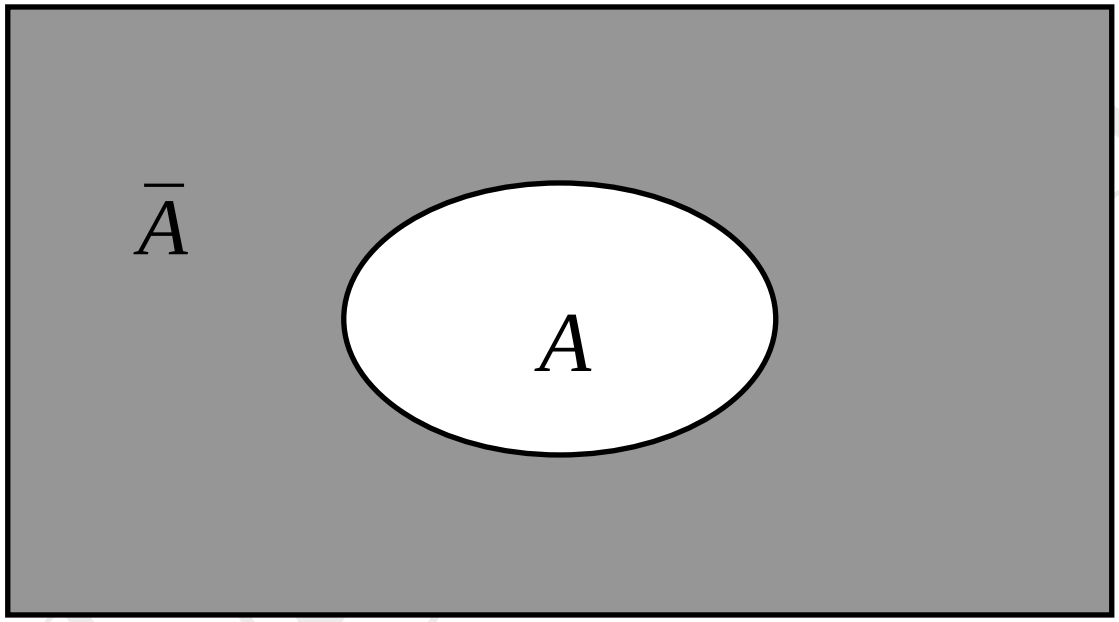
or

$$P\left[\text{not } A\right] = 1 - P\left[A\right]$$

Complementation Rule:

$\bar{A}$ and A are **mutually exclusive**.

and $S = A \cup \bar{A}$



thus $1 = P[S] = P[A] + P[\bar{A}]$

and $P[\bar{A}] = 1 - P[A]$

Conditional Probability

Conditional Probability

- Frequently before observing the outcome of a random experiment you are given information regarding the outcome
- How should this information be used in prediction of the outcome.
- Namely, how should probabilities be adjusted to take into account this information
- Usually the information is given in the following form: You are told that the outcome belongs to a given event. (i.e. you are told that a certain event has occurred)

Definition

Suppose that we are interested in computing the probability of event *A* and we have been told event *B* has occurred.

Then the conditional probability of *A* given *B* is defined to be:

$$P\left[A|B\right] = \frac{P\left[A \cap B\right]}{P\left[B\right]} \quad \text{if} \quad P\left[B\right] \neq 0$$

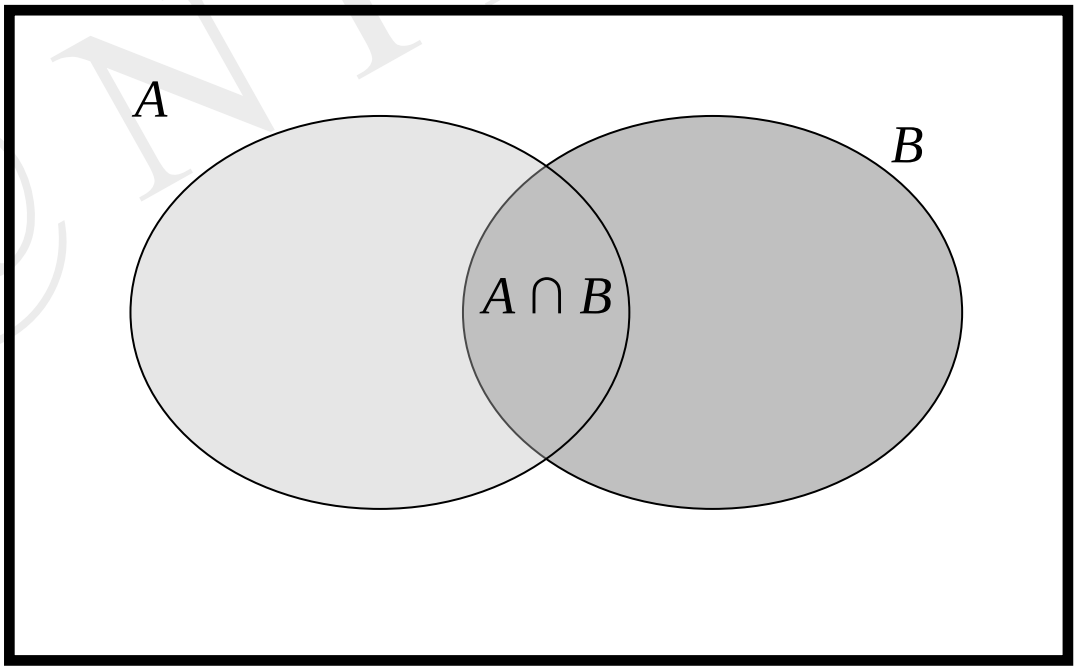
Rationale:

If we're told that event B has occurred then the sample space is restricted to B.

The probability within B has to be normalized, This is achieved by dividing by $P[B]$

The event A can now only occur if the outcome is inside of $A \cap B$. Hence the new probability of A is:

$$P\left[A|B\right] = \frac{P\left[A \cap B\right]}{P\left[B\right]}$$



An Example

The academy awards is soon to be shown.

For a specific married couple the probability that the husband watches the show is 80%, the probability that his wife watches the show is 65%, while the probability that they both watch the show is 60%.

If the husband is watching the show, what is the probability that his wife is also watching the show

Solution:

The academy awards is soon to be shown.

Let B = the event that the husband watches the show

$P[B]= 0.80$

Let A = the event that his wife watches the show

$P[A]= 0.65$ and $P[A \cap B]= 0.60$

$$P\left[A|B\right] = \frac{P\left[A \cap B\right]}{P\left[B\right]} = \frac{0.60}{0.80} = 0.75$$

Independence

Independence Definition

- Two events A and B are said to be independent if the conditional probability of an event is not affected by the knowledge of the other event.
  - i.e. P(A|B) = P(A)
  - **or equivalently,** P(B|A) = P(B)

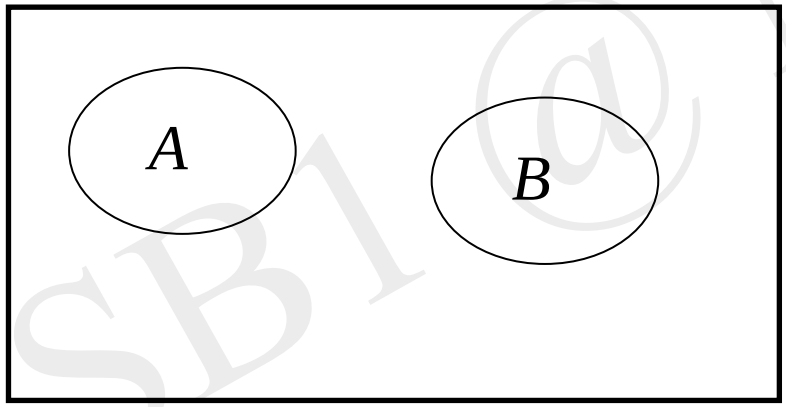
- Another **equivalent** condition for independence is:
  - $P(A \cap B) = P(A) \times P(B)$

Difference between **independence** and **mutually exclusive**

mutually exclusive

Two mutually exclusive events are independent only in the special case where

$$P[A] = 0 \text{ and } P[B] = 0. \text{ (also } P[A \cap B] = 0$$

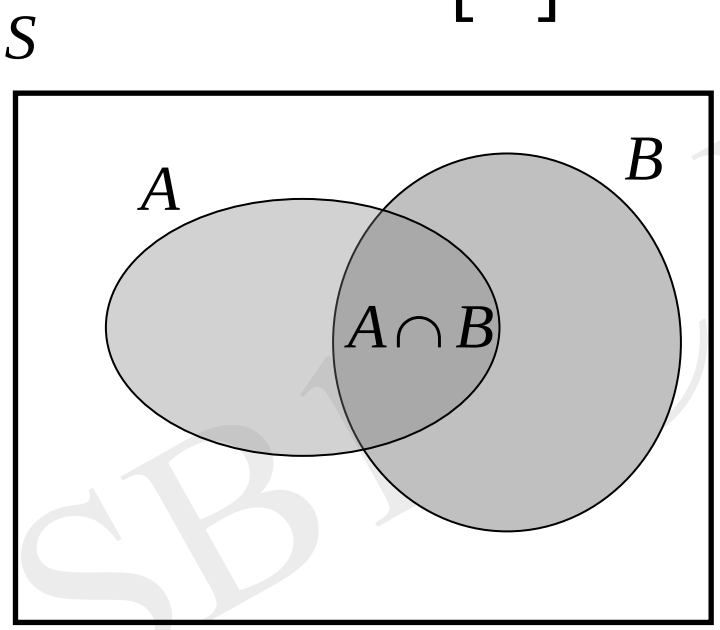


Mutually exclusive events are highly dependent otherwise. *A* and *B* **cannot** occur simultaneously. If one event occurs the other event does not occur.

Independent events

$$P[A \cap B] = P[A]P[B]$$

or $$\frac{P[A \cap B]}{P[B]} = P[A] = \frac{P[A]}{P[S]}$$



The ratio of the probability of the set *A* within *B* is the same as the ratio of the probability of the set *A* within the entire sample *S*.

The multiplicative rule of probability

$$P\left[A \cap B\right] = \begin{cases} P\left[A\right]P\left[B\middle|A\right] & \text{if } P\left[A\right] \neq 0 \\ P\left[B\right]P\left[A\middle|B\right] & \text{if } P\left[B\right] \neq 0 \end{cases}$$

and

$$P\left[A \cap B\right] = P\left[A\right]P\left[B\right]$$

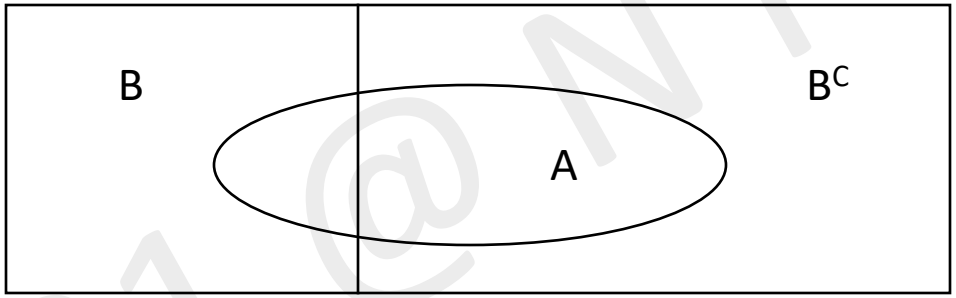
if A and B are **independent**.

Breaking a probability down by conditioning

- P(A) can be written by "conditioning on B" when P(A) is hard but P(A | B) is easier.
- Combination of the addition rule and the multiplication rule for independent probabilities.
- Useful trick for calculating hard probabilities.

Breaking a probability down by conditioning (cont.)

- $P(A) = P(A \text{ and } B) + P(A \text{ and } B^C)$
  $= P(B)\,P(A \mid B) + P(B^C)\,P(A \mid B^C)$



- You can also divide into more than two categories.

Breaking a probability down by conditioning - genetics example

- A woman has a 60% chance of getting breast cancer (BC) if she carries a BRCA1 mutation (m).
- She has a 10% chance of getting BC if she does not have the mutation m.
- The frequency of the mutation is about 2%?

- What is her overall risk?

- $P(BC) = P(m)\,P(BC \mid m) + P(\text{no m})\,P(BC \mid \text{no m})$
  $= (.02)(.60) + (.98)(.10) = .11$

Bayes' rule

- Multiplication rule (non-independent) combined with conditioning.

- Use it to "turn a conditional probability around," i.e. if P(A | B) is hard but P(B | A) is easy.

Bayes' rule (cont.)

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B \mid A)}{P(B)}$$

$$= \frac{P(A)P(B \mid A)}{P(A)P(B \mid A) + P(A^C)P(B \mid A^C)}$$

Bayes' rule - breast cancer example

- Chance that she has a mutation given that she has breast cancer is
- $P(m \mid BC) = P(m \text{ and } BC) / P(BC)$

  $= P(m)\, P(BC \mid m) / P(BC)$

  $= (.02)\,(.60) / [(.02)\,(.60) + (.98)\,(.10)]$

  $= (.02)\,(.60) / (.11) = .11$

Random variables and probability distributions

Random variables

- A random variable is a numerical measurement corresponding to each outcome of an experiment.

- Example 1: X = value obtained from a die roll

- Example 2: Y = genotype of random person (coded as 0 for ‘aa’, 1 for ‘Aa’ and 2 for ‘Aa’)

Events about random variables

- Events can be constructed based on value of a random variable.

- X = value obtained from a die roll

  A = (X = 1) , B = (X is odd)

- Y = genotype of random person (0 for ‘aa’, 1 for ‘Aa’ and 2 for ‘Aa’).

  C = (The person is homozygous) = ($Y \neq 1$)

Probability distributions

- A probability distribution describes the likelihood of each possible outcome for a random variable.

- X = value obtained from a die roll

  $P(X = 1) = 1/6$ $P(X = 2) = 1/6$ etc.

- Y = genotype of random person

  $P(Y = aa) = .3$ $P(Y = Aa) = .2$ $P(Y = AA) = .5$

Addition rule - examples

- X = value obtained from a die roll

  A = (X = 1) B = (X is even)

  P(A U B) = 1/6 + 3/6 = 4/6

- Y = genotype of random person (0 for ‘aa’, 1 for ‘Aa’ and 2 for ‘Aa’).

  C = (Y = 0) D = (Y = 2)

  P(homozygote) = P(C U D) = .3 +.5 = .8

Complementation rule - examples

- X = Value obtained for die roll

  A = (X > 1)

  P(A) = 1 - P(X = 1) = 1 - 1/6 = 5/6

- Y = genotype of random person

  C = (genotype contains at least one A)=($Y \leq 1$)

  P(C) = 1 - P(genotype contains no A's)

  = 1 – P(Y=2) = 1 - .3 = .7

Independence Examples

- X = outcome of die roll

  Y = outcome of another die roll

  Any event involving X and any event involving Y are independent.

- Dependent:

  (X = 2) and (X = 3) are **not independent** as they are mutually exclusive.

Multiplication rule for independent probabilities - examples

- X = outcome of die roll
  Y = outcome of another die roll
  A = (X > 1) B = (Y = 2)
  P(A and B) = 5/6 x 1/6 = 5/36

- Y = genotype of random person
  Assume random mating, which means maternal and paternal alleles are random draws from the population.
  P(AA) = P(Y = 0) = p x p = p^2

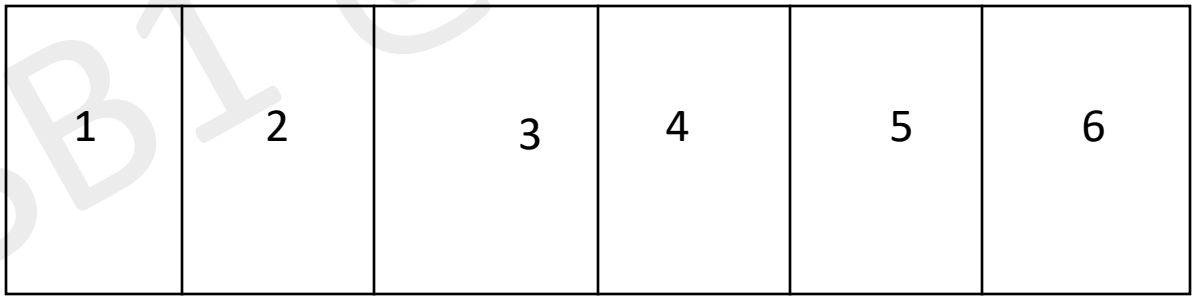
Conditional probability - examples

- X = outcome of die roll

  A = (X > 1) B = (X = 2)

  P(B | A) = P(X = 2 | X > 1) = 1/5

  P(A | B) = P(X > 1 | X = 2) = 1



Conditional probability - examples

- X = outcome of die roll
  - A = (X > 1), B = (X = 2)

$$P(B \mid A) = P(X = 2 \text{ and } X > 1) / P(X > 1)$$
$$= (1/6) / (5/6) = 1/5$$

$$P(A \mid B) = P(X = 2 \text{ and } X > 1) / P(X = 2)$$
$$= (1/6) / (1/6) = 1$$

Discrete Random Variable & pmf

- Any random variable with a finite (or countable) number of possible values.
- Specifying the distribution of a discrete random variable requires the probability for every value in its support i.e.

$$P(X = k)\ for\ each\ k \in Support\ of\ X.$$

- It is also called p.m.f (probability mass function) of X. **Values of the p.m.f , i.e. $\boldsymbol{P(X = k)}$ must always add up to 1.**
- Using the pmf, any probability can be calculated.

For example:

- $P(4 < X < 9) = \sum_{k=5}^{8} P(X = k) = p_5 + p_6 + p_7 + p_8$

Expectation

- In a dataset, if values x_i are repeated f_i times. Then,

$$\text{Sample Average (mean): } \bar{X} = \frac{\sum_i x_i f_i}{n}$$

- The expectation of a random variable is the long-run average. If the sample size were very large the relative frequencies f_i/n could be replaced by the probabilities.
- Let random variable X take values x_1, x_2, ...
- $E(X) = \sum_i x_i \, P(X = x_i) = \sum_{i=1}^{k} x_i \, p_i$

Expectation - example

- X = outcome of die roll
- X takes values 1, 2, 3, 4, 5, 6
- Equally Likely Case:
  - E(X) = 1(1/6) + 2(1/6) + 3(1/6) + 4(1/6) + 5(1/6) + 6(1/6)
    = 3.5
- In general:
  - E(X) = $1 * p_1 + 2 * p_2 + 3 * p_3 + 4 * p_4 + 5 * p_5 + 6 * p_6$

Variance and standard deviation

- The variance of a random variable tells you how much variability there is around the expected value.
  - $V(X) = E[X - EX]^2 = E(X^2) - E^2(X)$
    $= \sum_i x_i^2 p_i - \{\sum_i x_i\, p_i\}^2$

- The standard deviation is the square root of the variance.

- $Var(X) \geq 0$ and is equal to zero only when X takes a single value with probability 1. Example $P(X = 3) = 1$, which also means $P(X = k) = 0$, for all $k \neq 3$. Such a random variable is called a "degenerate" of "constant" random variable.

Binomial distribution

- Start with an experiment that has only two outcomes (yes/no, heads/tails, affected/unaffected).
- Do the experiment *n* independent times (e.g., toss *n* coins).
- X is the number of “successes” out of *n* trials (success defined arbitrarily).
- The distribution of X is binomial(*n*, *p*), where *p* is the probability of “success” in each experiment.

Bernoulli & Binomial distributions

- Bernoulli (p) : X can take values 0 and 1.

$$P(X = 1) = p$$
$$P(X = 0) = 1 - p = q$$

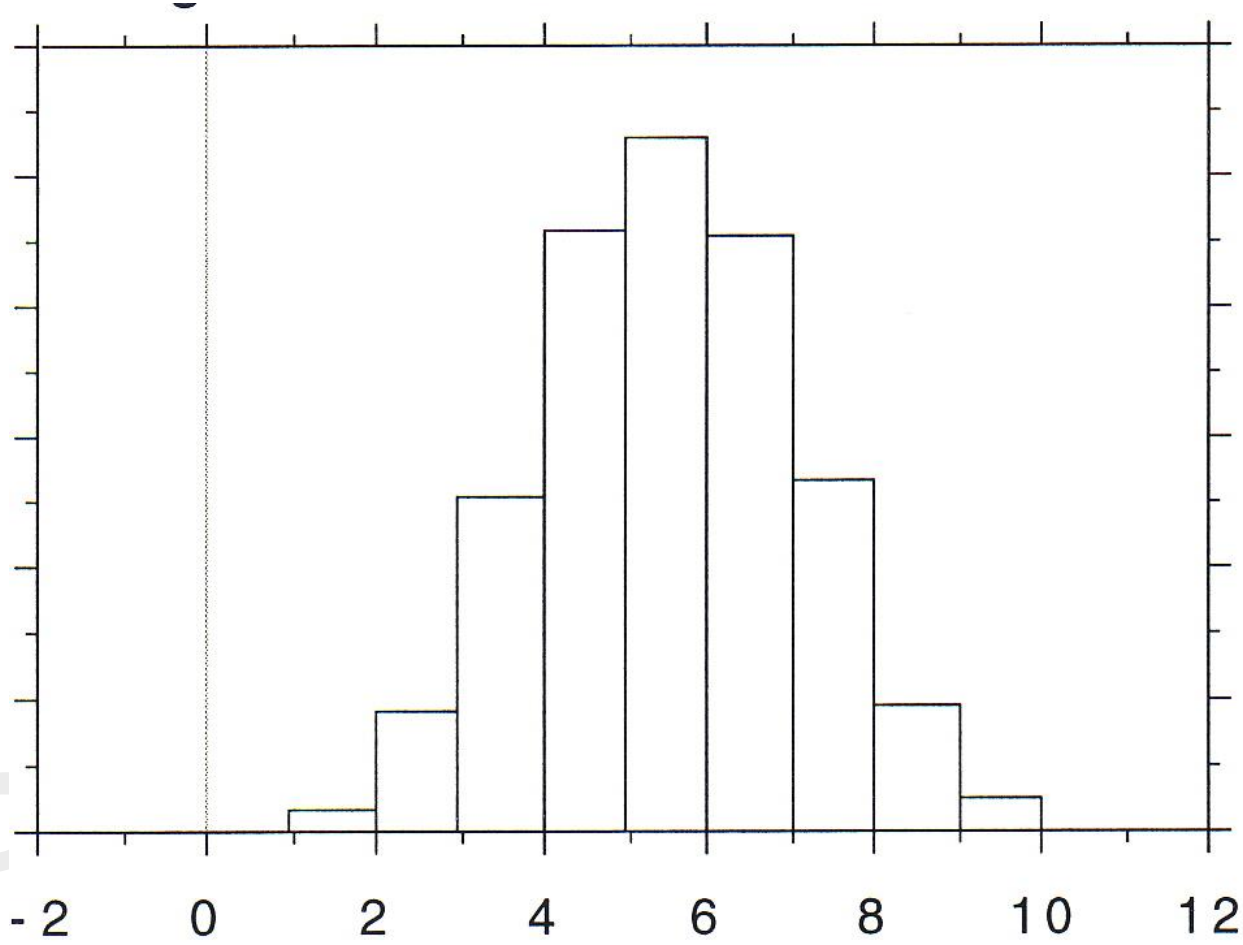
- Binomial (n, p): X can take values from 0 to *n*.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

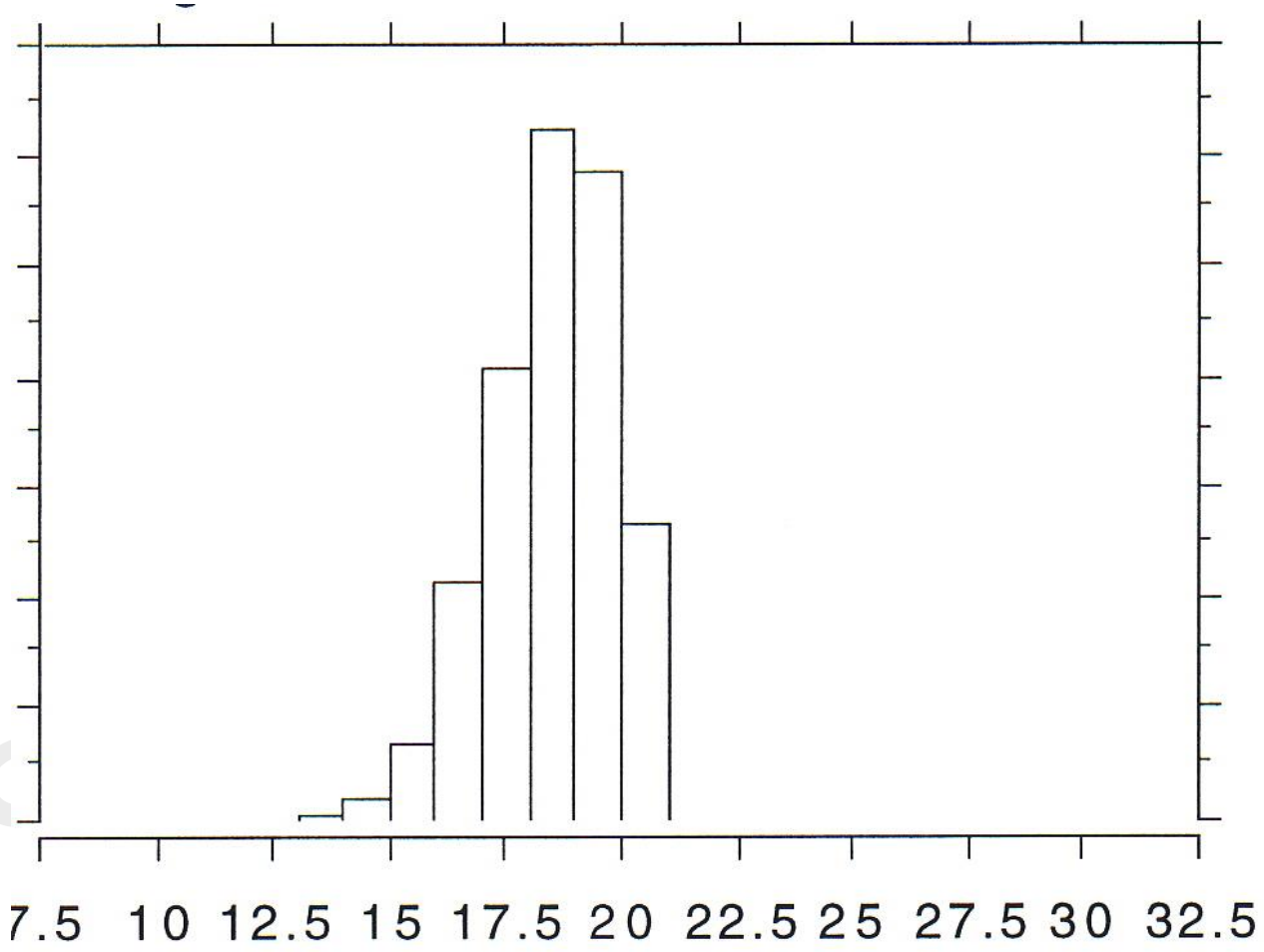
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

$$n! = n(n-1)(n-2)\dots(1)$$
$$0! \equiv 1$$

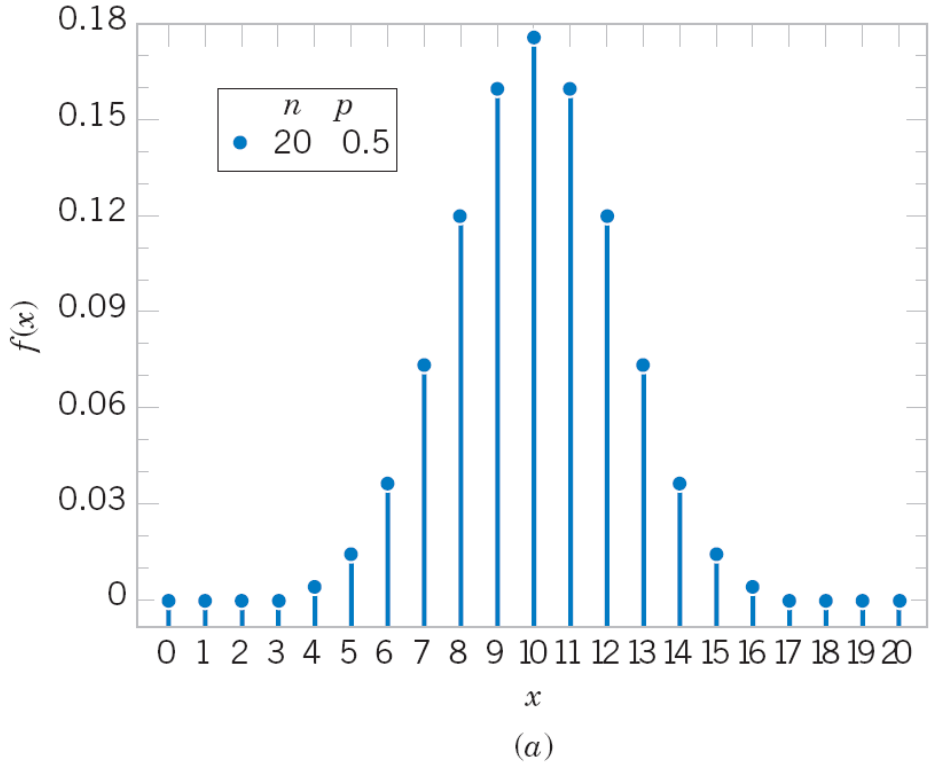
Binomial (10, 0.5)

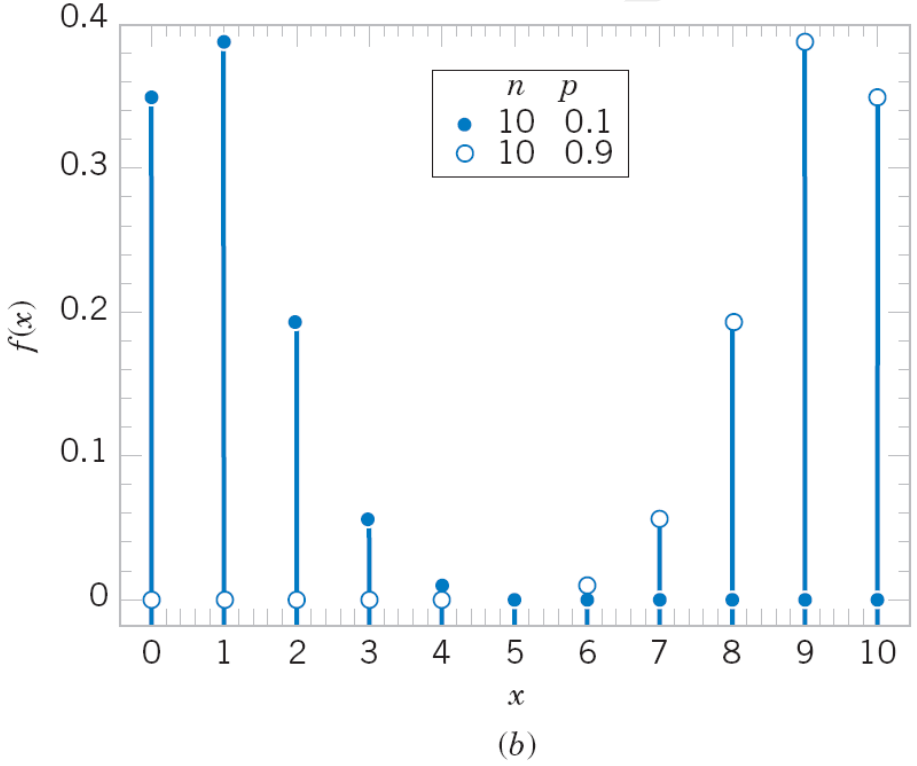


Binomial (20, 0.9)



Binomial Distribution (pmf)





Binomial distribution for selected values of n and p.

Binomial distribution - example

- Sample 10 people with BRCA1 mutation (Chance of cancer= 0.6) .
- What is the probability that none of them will get breast cancer?

$$P(X=k)=\binom{n}{k}p^{k}(1-p)^{n-k}$$

$$P(X=0)=\binom{10}{0}(.6)^{0}(.4)^{10}=\frac{10!}{(0!)(10!)}(1)(.4)^{10}$$

$$=(1)(1)(.4)^{10}=.0001$$

Mean and Variance

- Bernoulli (p):
  - E(X) = 1*p + 0 * (1 – p) = p
  - V(X) = $EX - EX^2 = p - p^2 = p(1-p) = pq$

- Binomial (n, p) :

Sum of n independent Bernoulli trials.

  - E(X) = n * p
  - V(X) = n * p * q

Relative Frequency Interpretation of Probability

- **Probability:** If an experiment is repeated N times, then Proportion of times an event A will happen will converge to $P(A)$ as N $\rightarrow \infty$.

- **Distribution of a Random Variable:** Proportion of times a random variable X takes a certain value k will converge to $P(X = k)$, as the number of repetitions $N \rightarrow \infty$.

- **Mean and Variance:** If a sample of size N is drawn from a distribution e.g. $X \sim Bin(n, p)$, the sample looks like $x_1, x_2, \dots, x_N$. The sample average (mean) and sample variance of these observations will approximate the theoretical expectation (in this case np) and theoretical variance (in this case npq) , and the approximation becomes accurate as $N \rightarrow \infty$.

- $\frac{1}{N}\sum_{i=1}^{N} x_i \approx E(X)$ **and** $\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \approx Var(X)$

Continuous distribution

- So far, we've only talked about random variables with discrete outcomes.

- Suppose our random variable is "height of random person from population." It is continuous random variable.

- Height may have a "normal" or "Gaussian" distribution (named after mathematician Gauss). It is an example of a continuous distribution.

- A continuous distribution does not put any 'mass' anywhere. The probability of an exact value of height is 0. Example:

$$P(H = 5.436789\ldots.) = 0$$

- However, intervals have a positive probability i.e.

$$P(5.43 \leq H \leq 5.44) > 0$$

Density Function

- The shape of histogram of a continuous distribution is specified by a curve $f(x)$ called pdf (probability density function) of X. **Just like probabilities add up to 1, total area under the density curve =**

  $\int_a^b f(x)dx = P(-\infty < X < \infty) =$ **1.**

- Probability for any interval is the area under the density curve within that interval:

$$P(a < X < b) = \int_a^b f(x)dx\,.$$

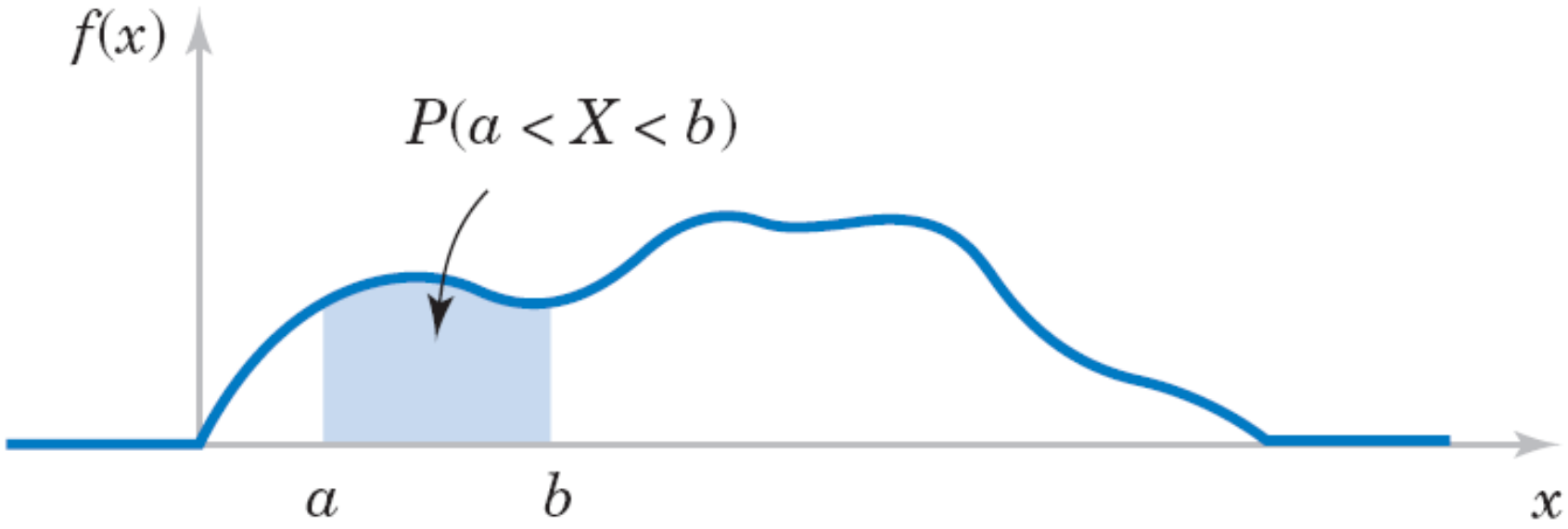
- $P(Small\ \ Interval\ near\ x) \approx f(x) * dx$

$$\approx histogram\ height\ * width\ of\ bin$$

- Thus, density $f(x)$ can be interpreted as the 'instantaneous rate of accumulation of probability near x'.

PDF and Probabilities (Continuous)

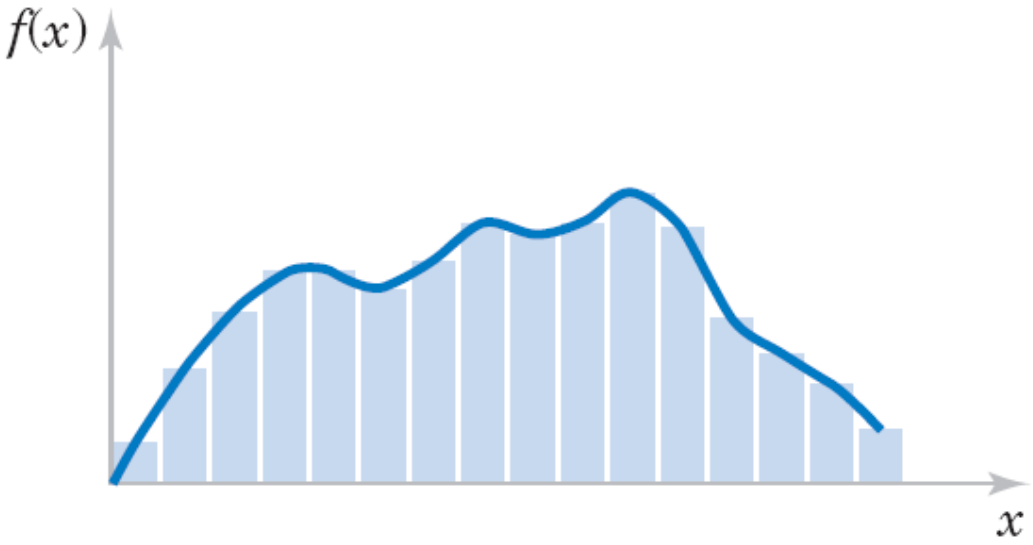
Probability Density Function



Probability determined from the area under $f(x)$.

PDF: Continuous Random Variables

Probability Density Function



A histogram approximates a probability density function. The area of each bar equals the relative frequency of the interval. The area under $f(x)$ over any interval equals the probability of the interval.

Expectation and variance

- The mean of a continuous random variable is similar to discrete, only the summation needs to be replaced by an integral.
- $E(X) = \int xf(x)dx$

$$\approx Sum\ over\ small\ intervals\ with\ mid - value\ x_i :$$

$$\sum_i (x_i) \times [f(x_i) * dx] = \sum_i (x_i) \times prob\ of\ interval$$

- $V(X) = E[(X - EX)^2] = EX^2 - E^2X$

$$= \int x^2 f(x)dx - \left\{\int xf(x)dx\right\}^2$$

Standard Normal distribution

Standard normal distribution is a very commonly used distribution in statistics. It has a bell-shaped symmetric density centered at 0.

Suppose Z has standard normal distribution:

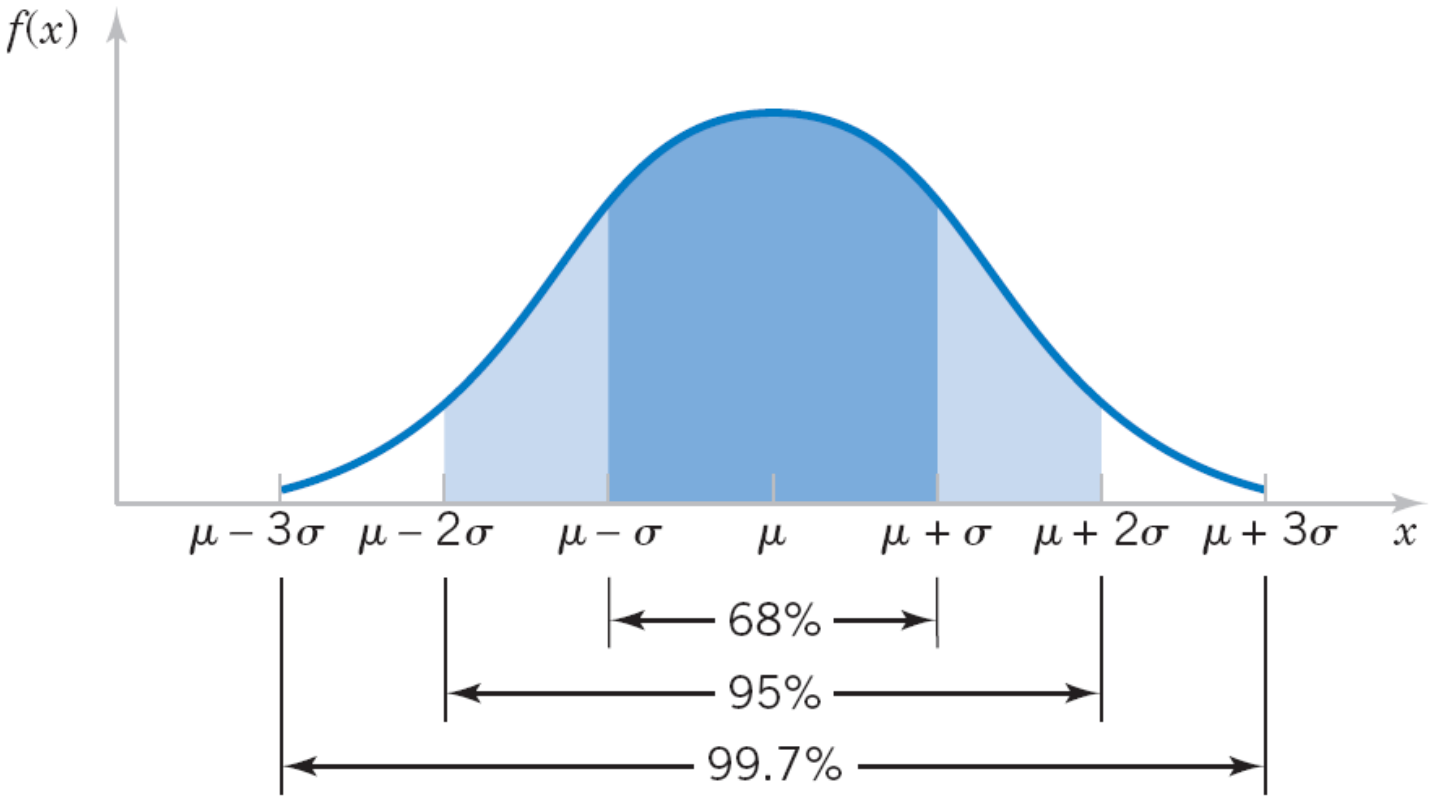
- Mean of Z is 0.
- Variance of Z is 1.
- We write $Z \sim Normal(0,1)$.
- Density: $f(z) = \frac{1}{\sqrt{2*pi}} e^{-\frac{z^2}{2}}, for -\infty < x < \infty$ (R function dnorm(), Excel function DNORM)

Normal distribution - mean and variance

- $X = \sigma Z + \mu$ is a normal random variable.
- Mean of X is μ.
- Variance of X is σ^2.
- We write $X \sim Normal(\mu, \sigma^2)$.
- Density: $f(x) = \frac{1}{\sqrt{2*pi}} e^{-\frac{(x-\mu)\text{^}2}{2\sigma^2}}, for -\infty < x < \infty$ (R function dnorm(), Excel function DNORM)

Important Continuous Distributions

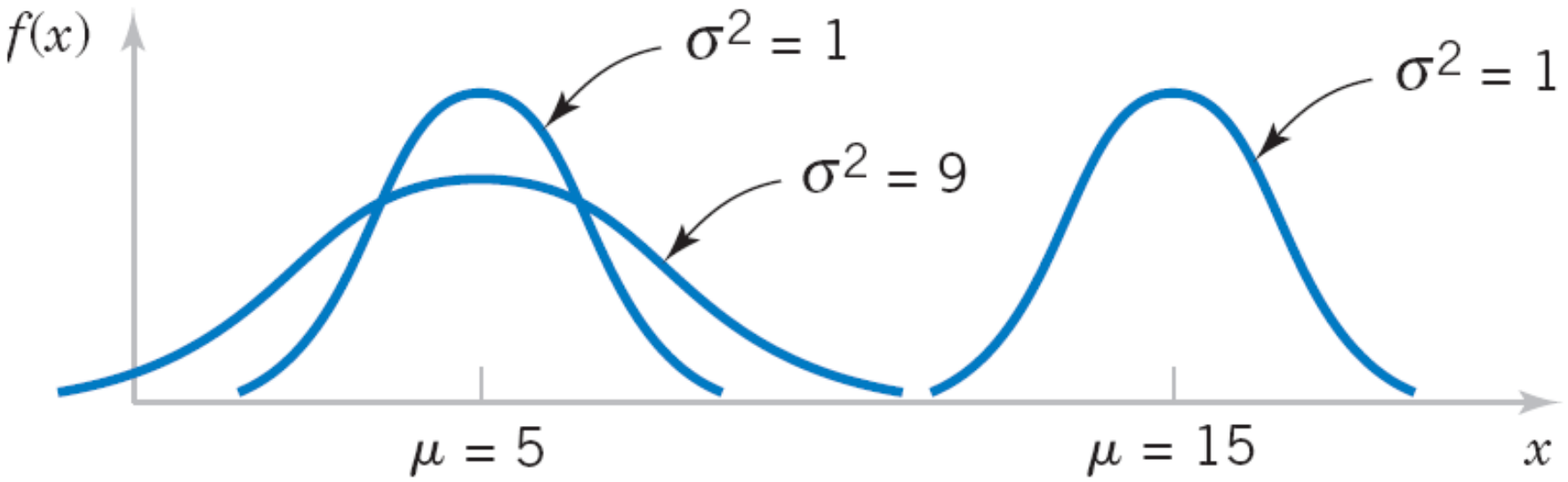
Normal Distribution



Probabilities associated with a normal distribution.

Important Continuous Distributions

Normal Distribution



Normal probability density functions for selected values of the parameters μ and σ^2.

Uniform Distribution on [0, 1]

- U ~ Uniform[0,1]
- All values in [0,1] are equally likely.
- Density: $f(u) = 1, \text{for } 0 \leq \text{u} \leq 1$
- Mean: $E(U) = \int_0^1 uf(u)du = \frac{1}{2}$
- Variance: $V(U) = E(U^2) - E^2(U) = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$
- Uniform on any interval: X ~ Uniform[A, B].
- Density: $f(x) = \frac{1}{B-A}\ for\ A \leq x \leq B$

Cumulative Distribution Function (CDF)

- Consider the following frequency table for a dataset (sample) $x_1, x_2, \ldots x_n$ of a discrete variable having k unique values denoted as $t_1, t_2, \ldots t_k$ with t_j repeated f_j times.

Data Value	t_1	t_2	t_3	…	t_{k-1}	t_k	
Frequency	f_1	f_2	f_3	…	f_{k-1}	f_k	Total=n
Cumulative Frequency	f_1	$f_1 + f_2$	$f_1 + f_2 + f_3$	…	$\sum_{j=1}^{k-1} f_i$	$\sum_{j=1}^{k} f_j = \mathrm{n}$	
Cumulative Relative Frequency	$\frac{f_1}{n}$	$\frac{f_1}{n} + \frac{f_2}{n}$	$\frac{f_1}{n} + \frac{f_2}{n} + \frac{f_3}{n}$	…	$\sum_{j=1}^{k-1} \frac{f_i}{n}$	$\sum_{j=1}^{n} \frac{f_j}{n} = 1$	

- Note that the last row basically gives the **proportion of points** below a certain t_j. This is known as the sample Cumulative Distribution Function $F_n(t)$.

- Note that these probabilities can also be obtained directly from the raw data without creating a frequency table as follows $\mathrm{F_n(t)} = \{ \#i\colon\ x_i \leq t\}/\ n$.

CDF and Quantiles: Discrete Case

- Similar concept also exists for the population or theoretical distribution. Consider the following pmf of the discrete distribution of X:

Possible Values of Random Variable	t_1	t_2	t_3	…	t_{k-1}	t_k	
Probability (PMF)	π_1	π_2	π_3	…	π_{k-1}	π_k	Total=1
Cumulative Probability (CDF)	π_1	$\pi_1 + \pi_2$	$\pi_1 + \pi_2 + \pi_3$	…	$\sum_{j=1}^{k-1} \pi_j$	$\sum_{j=1}^{k} \pi_j = 1$	

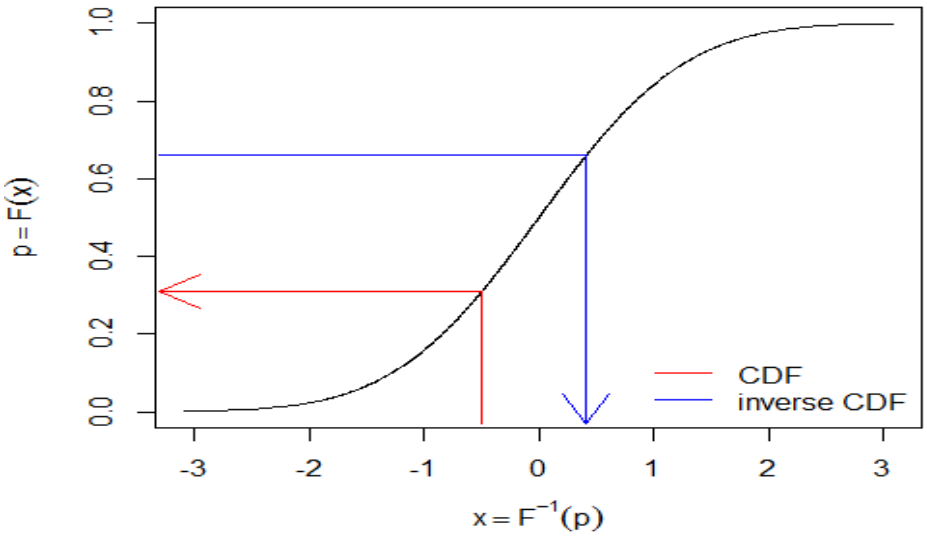
- Note that the last row basically gives the **probability of points** below a certain t_k. This is known as the Cumulative Distribution Function: $\mathrm{F}(t) = P(X \leq t)$.

- Note that if t_m denotes the sample median, we should have $F_n(t_m) = 0.5$
- The population median is the point t_M such that $F(t_M) = P(X \leq t_M) = 0.5$
- In fact, any quantile of the data and the distribution can be defined through the CDF.
- The $q'th$ quantile or $100 * q$ 'th percentile of the sample and population are defined as solutions of $F_n(x) = q$ and $F(x) = q$ respectively.

CDF & Quantiles: Continuous Case

- For a continuous distribution, the CDF at a point in the distribution can be obtained by integrating the area under the density curve from $-\infty$ upto that point.
  - $F(t) = P(X \leq t) = \int_{t=-\infty}^{t} f(x)\, dx$ [conversely, $F(t) = \frac{d}{dx} f(x) \Big| x = t$]
- Note that CDF (both discrete and continuous case) either remains constant or increases as we move to the right (as cumulative sum or cumulative area can only increase or stay the same).
- It remains constant in regions where the distribution puts 0 mass or density. Thus, discrete distribution CDF usually looks like a step function, increasing in jumps. Continuous CDF increases smoothly.
- In either case, the quantiles (median, Q1, Q3 etc.) can be obtained by locating 0.5, 0.25, 0.75 etc. on the y-axis, drawing a horizontal line, seeing where it cuts the CDF and drawing a vertical line to the x-axis. In mathematical notation:
  - The q^{th} quantile is $x = F^{-1}(q)$, since $F(x) = q$. [Quantile Function and CDF are inverses of each other].

CDF & Quantiles in R



- In R, pnorm() denotes CDF of normal distribution and qnorm() its quantiles. Similarly, pchisq() denotes CDF of chi-square distribution, and qchisq() its quantiles etc.
- Example: Try these 2 commands in R

  > qnorm(c(0.05, 0.5, 0.95))

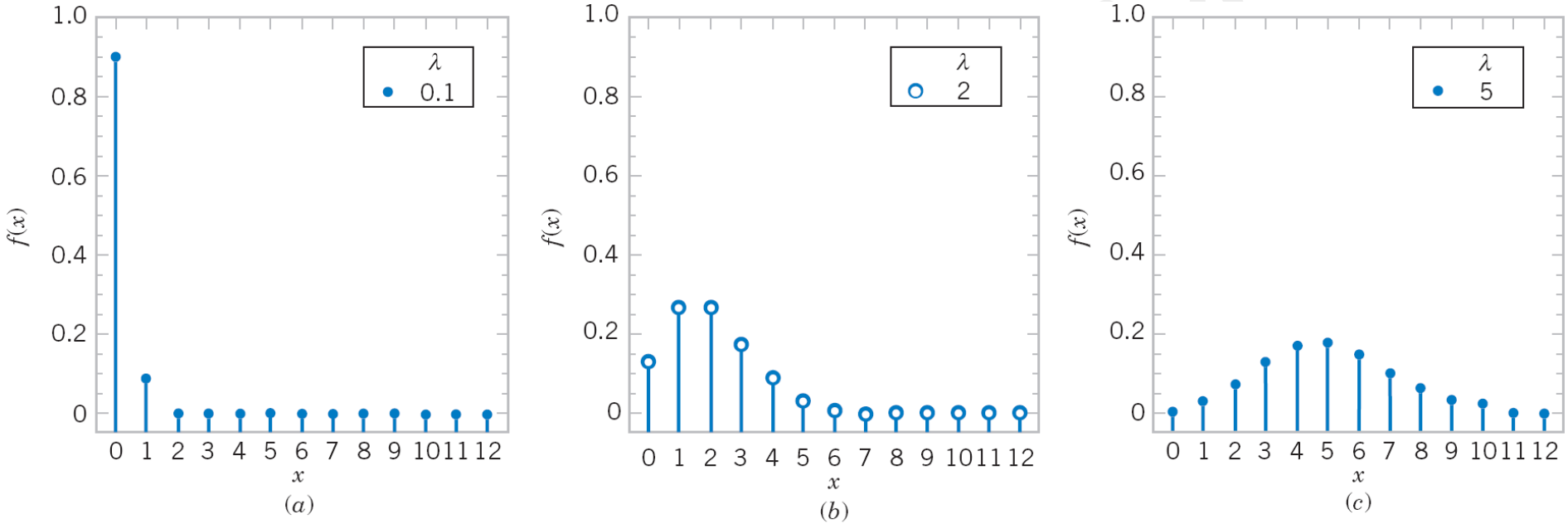
  > pnorm(c(-1.644854, 0, 1.644854))
- The value 1.64 is often used it statistical tests at level of significance 5%.

Some other common distributions

- **Poisson (discrete):** We say the count variable X has Poisson distribution (with rate or mean=λ), i.e., $X \sim Poisson(\lambda)$ if the pmf is
- $P(X = k) = f(k) = \frac{\lambda^k}{k!} e^{-\lambda k}$, $k = 0,1, \dots, \infty$
- The mean and variance are equal: $E(X) = \lambda$ and $V(X) = \lambda$.
- Poisson is used to model count data for rare events: Example number of road accidents on a particular day in a city. We can also model this as $Bin(N, p)$, where N is the number of streets (very large) and p is a very small 'accident probability' in a street (and accidents happen at all streets independently).
- Interestingly, the two models are equivalent. It can be shown that Poisson is just the limit of Binomial when $N \to \infty$, $p \to 0$ and $Np \to \lambda$.
  - Recall, Binomal mean and variance are Np and Npq , the limits of both of which are λ .
  - The advantage of using Poisson over Binomial in such cases is we can only deal with one parameter λ (mean number of accidents), which can be easily estimated from the data observed daily.

Poisson Distribution

Shape of Poisson pmf

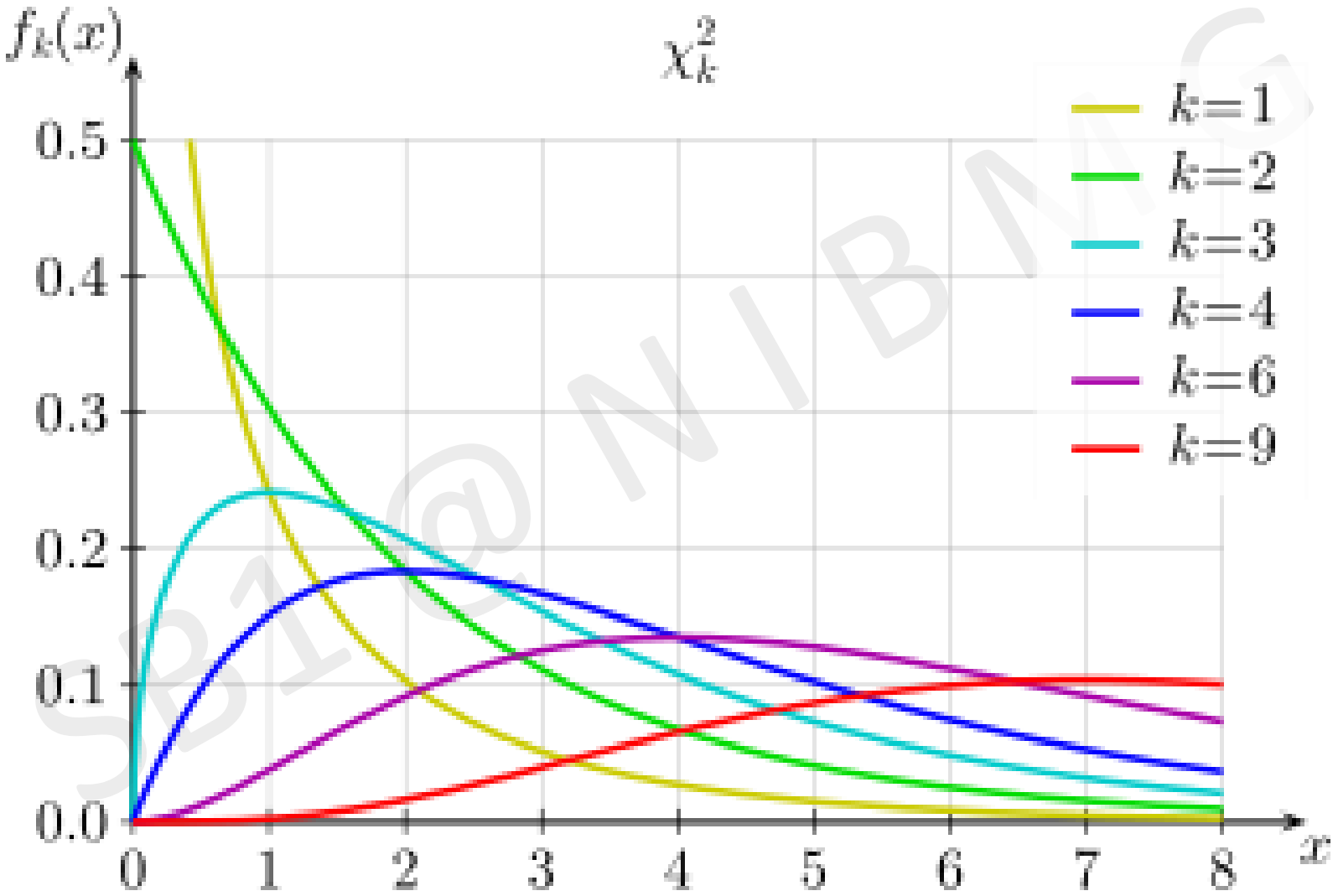


Poisson distribution for selected values of the parameter λ.

Some other common distributions

- **Chi-square (Continuous):** We say that the positive variable X has Chi-squared distribution (with k degrees of freedom), i.e., $X \sim \chi^2(k)$ if X is a sum of squares of independent standard normal variables:
- $X = Z_1^2 + Z_2^2 + \cdots Z_k^2$, for $k = 1,2, \ldots . < \infty$
- The mean and variance: $E(X) = k$ and $V(X) = 2k$.
- Unlike normal, it is skewed. The skewness and kurtosis gradually decrease with df (degrees of freedom).
- It is often useful in statistics for hypothesis testing (e.g., test of independence in a contingency table).
- If data are normally distributed, the ratio of the sample variance and population variance (scaled appropriately) has a chi-square distribution.
- $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(k = n-1)$, where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

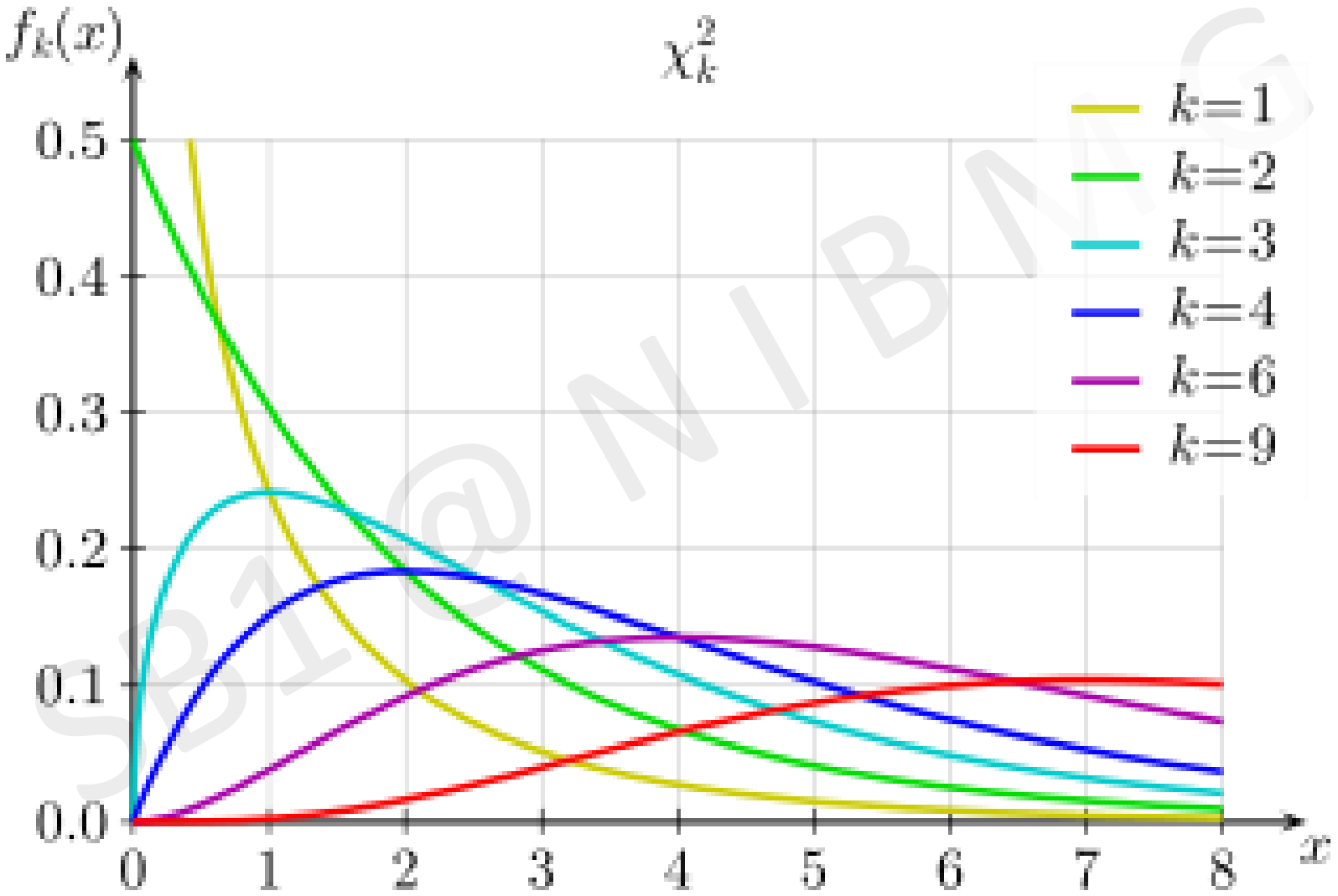
Chi-squared Distribution



Exponential Distribution

- This distribution can be used to model lifetime of bulb or other equipment assuming random shocks without any wear and tear of the equipment.
- It has a memoryless property. Meaning, the distribution of X conditional on $X > t$, is the same as the unconditional distribution of X, regardless of t.
- It has one parameter (rate $\lambda > 0$). We say $X \sim Exp(\lambda)$. The pdf and cdf are:
  - $f(x) = \lambda e^{-\lambda x}$ and $F(x) = 1 - e^{-\lambda x}$ [Note: $\frac{dF}{dx} = f$]
- The mean and variance are: $E(X) = \frac{1}{\lambda}$ and $V(X) = \frac{1}{\lambda^2}$.

Exponential Distribution



Joint, Marginal, Conditional Distributions (Discrete)

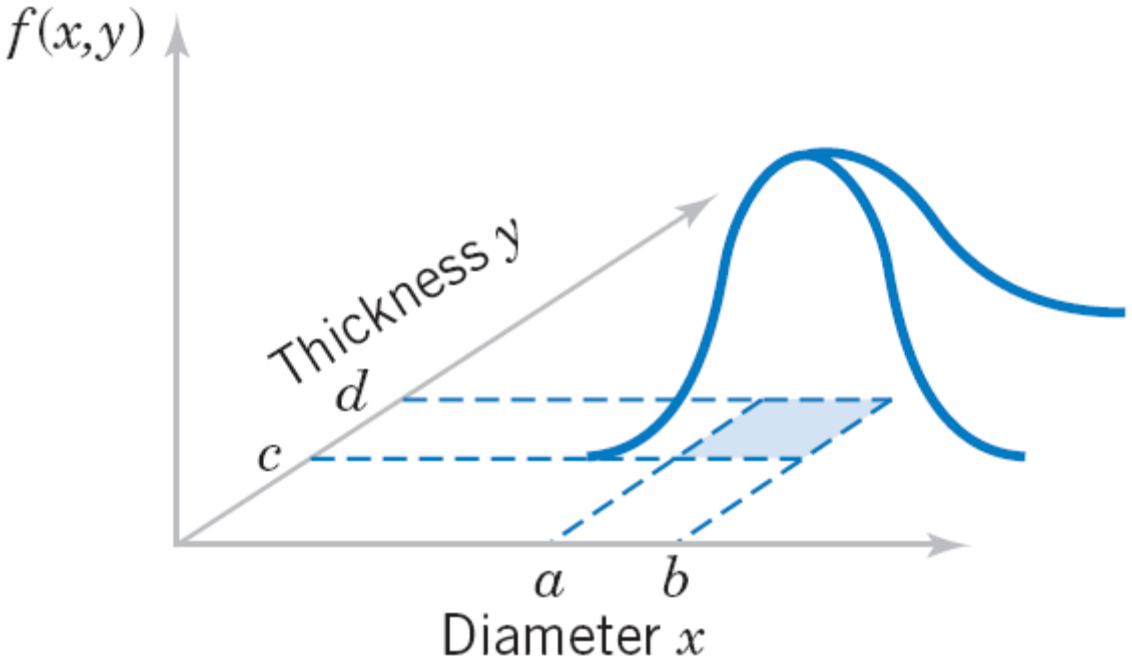
- Two (discrete) random variables X, Y together have a joint distribution, from which marginal and conditional distributions can be calculated:
  - **Joint:** $\mathrm{P}(X = x_j, Y = y_k) = p_{jk}$
  - **Marginals:** $p_{j+} = \sum_k p_{jk}$, i.e., $P(X = x_j) = \sum_k P(X = x_j, Y = y_k)$ (By theorem of total probability)
  - **Conditionals:** $p_{j|k} = \frac{p_{jk}}{p_k}$, i.e., $P(X = x_j | Y = y_k) = \frac{P(X=x_j, Y=y_k)}{P(X=x_j)}$
  - When X and Y are independent:
    - $p_{jk} = p_{j+} \times p_{+k}$, i.e., $\mathrm{P}(X = x_j, Y = y_k) = P(X = x_j) \times P(X = x_j)$
    - $p_{j|k} = p_j$ and $p_{k|j} = p_j$
- **Covariance:** $E(XY) - E(X)E(Y)$, where
  - $E(XY) = \sum_j \sum_k x_j y_k p_{jk}$, $E(X) = \sum_j x_j \, p_{j+}$, etc.
- **Conditional Means:** $E(X|Y = y_k) = \sum_j x_j p_{j|k}$
- **Conditional Variances:**
  - $V(X|Y = y_k) = E(X^2|Y = y_k) - E^2(X|Y = y_k)$

Joint, Marginal, Conditional Distributions (Continuous)

- Two (continuous) random variables X, Y together have a joint distribution, from which marginal and conditional distributions can be calculated:
  - **Joint Density:** $f(x, y)$
  - **Marginals:** $f(x) = \int_y f(x, y)dy$ (By theorem of total probability)
  - **Conditionals:** $f(x|y) = \frac{f(x,y)}{f(y)}$
  - When X and Y are independent:
    - $\mathrm{f}(x, y) = f(x) \times f(y)$
    - $f(x|y) = f(x)$ and $f(y|x) = f(y)$
- **Covariance:** $E(XY) - E(X)E(Y)$,
  - $E(XY) = \int_x \int_y x\ y\ f(x, y)\ dx\ dy$, $E(X) = \int_x xf(x)dx$, etc.
- **Conditional Means:** $E(X|Y = y) = \int_x x\ f(x|y)\ dx$
- **Conditional Variances:**
  - $V(X|Y = y) = E(X^2|Y = y)$ - $E^2(X|Y = y)$

More Than One Random Variable

Joint Distributions



Probability of a region is the volume enclosed by $f(x, y)$ over the region.

Joint Distribution from Joint Density

Joint distribution can be used to calculate probability of any region in the 2D space of X and Y.

$$P(a < X < b, c < Y < d) = \int_a^b \int_c^d f(x, y)\, dy\, dx$$

More Than One Random Variable and Independence

Independence

The random variables $X_1, X_2, \ldots, X_n$ are **independent** if

$$P(X_1 \in E_1, X_2 \in E_2, \ldots, X_n \in E_n) = P(X_1 \in E_1)P(X_2 \in E_2)\cdots P(X_n \in E_n)$$

for *any* sets $E_1, E_2, \ldots, E_n$.

Functions of Random Variables

$$Y = X + c$$

$$E(Y) = E(X) + c = \mu + c \tag{3-23}$$

$$V(Y) = V(X) + 0 = \sigma^2 \tag{3-24}$$

$$Y = cX$$

$$E(Y) = E(cX) = cE(X) = c\mu \tag{3-25}$$

$$V(Y) = V(cX) = c^2V(X) = c^2\sigma^2 \tag{3-26}$$

Functions of Random Variables

Linear Combinations of Independent Random Variables

The mean and variance of the linear function of **independent** random variables are

$$Y = c_0 + c_1X_1 + c_2X_2 + \cdots + c_nX_n$$

$$E(Y) = c_0 + c_1\mu_1 + c_2\mu_2 + \cdots + c_n\mu_n$$

and

$$V(Y) = c_1^2\sigma_1^2 + c_2^2\sigma_2^2 + \cdots + c_n^2\sigma_n^2$$

Functions of Random Variables

Linear Combinations of Independent Random Variables

Let $X_1, X_2, \ldots, X_n$ be independent, normally distributed random variables with means $E(X_i) = \mu_i$, $i = 1, 2, \ldots, n$ and variances $V(X_i) = \sigma_i^2$, $i = 1, 2, \ldots, n$. Then the linear function

$$Y = c_0 + c_1X_1 + c_2X_2 + \cdots + c_nX_n$$

is normally distributed with mean

$$E(Y) = c_0 + c_1\mu_1 + c_2\mu_2 + \cdots + c_n\mu_n$$

and variance

$$V(Y) = c_1^2\sigma_1^2 + c_2^2\sigma_2^2 + \cdots + c_n^2\sigma_n^2$$

Multivariate Discrete: Multinomial

- Consider n independent trials, where each trial has k possible outcomes (called categories) labelled as $c_1, c_2, \dots, c_J$, with probabilities $\pi_1, \pi_2, \dots \pi_J$, with $\sum_{j=1}^{J} \pi_j = 1$.
  - For example: Genotype counts (#AA, #Aa, #aa) or blood group counts in a sample or number of times we get 1,2,3,4,5,6 in 100 rolls of a dice.
  - Let the # times, we observe category c_j be X_j .
  - Note that $X_1 + X_2 + \cdots + X_J = n$ and $0 \leq X_j \leq n$
  - We say that jointly the frequencies of all the J categories has multinomial, i.e., $(X_1, X_2, \dots X_j, \dots X_J) \sim Multinomial\left(n, (\pi_1, \pi_2, \dots \pi_J)\right)$

- This is just a multivariate generalization of the binomial.
- If $X \sim Bin(n, p)$, and $Y = \#Failures = (n - X)$, we can say that $(X, Y) \sim Multinomial\big(n, (p, 1-p)\big)$.

- The pmf looks similar to binomial:

$$P\big(X_1 = k_1, X_2 = k_2, \dots, X_J = k_J\big) = \frac{n!}{k_1! k_2! \dots k_J!} \pi_1^{k_1} \pi_2^{k_2} \dots \pi_J^{k_J}$$

Multivariate Discrete: Multinomial

- The constant factor can be obtained as follows. Consider the sequence of outcomes $(c_1, c_1, \dots k_1 times, c_2, c_2, \dots k_2\ times, \dots c_J, c_J, \dots k_J\ times)$. All ways of shuffling this sequence is $n!$. Let m be the number of distinguishable permutations. Then for each of these m ways, we can shuffle the category-1 trials among themselves in $k_1!$ ways, category-2 labels in $k_2!$ ways and so on.
  - $m \times (k_1! \times k_2\ ! \times \cdots \times k_J!) = n!$

- Multinomial random vector can be expressed as a sum of independent categorical random vectors. Let $\mathrm{Z} = (Z_{i1}, Z_{i2}, \dots Z_{iJ})$ be the categorical random vector for the $i'th$ trial, i.e., $Z_{ij} = 1$ if the i'th trial gave category c_j and $Z_{ij} = 0$ for all the other categories. Then, it is clear that X (category counts) is just the column-wise sum of Z , $X_j = \sum_i Z_{ij}$. Just like, Binomial is sum of Bernoulli random variables. Using this representation, it is easy to show that:

- $E(X_j) = n\, E(Z_{1j}) = \pi_j$, $V(X_j) = n\, V(Z_{1j}) = n\, \pi_j\, (1 - \pi_j)$. In fact, $X_j \sim Bin(n, \pi_j)$
- $Cov(X_j, X_k) = n\, Cov(Z_{1j}, Z_{1k}) = -\, n\, \pi_j\, \pi_k$, which is negative (as total count is fixed, $=n$).

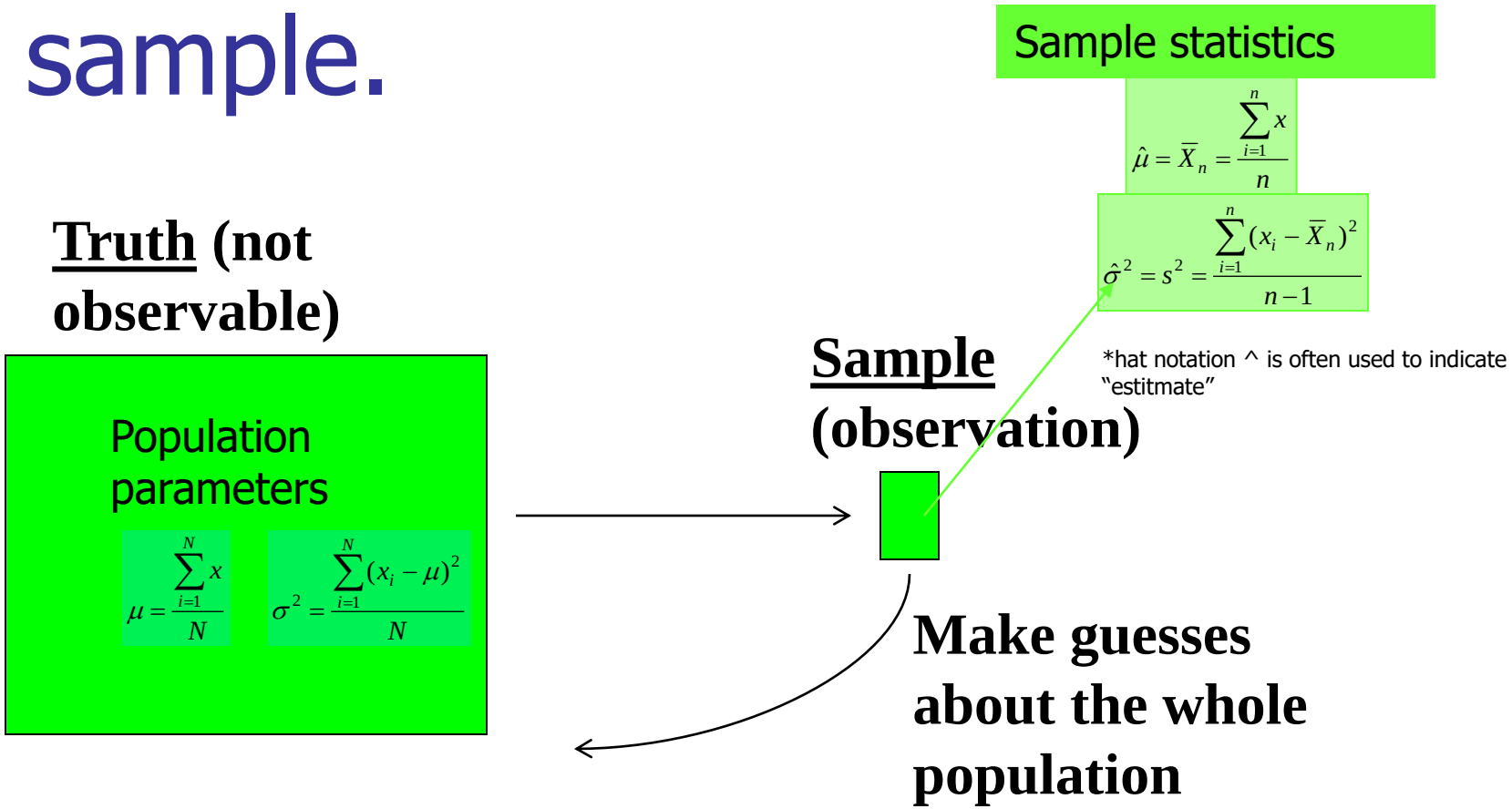
- Also, if we merge some categories: we still get a multinomial:
  $$(X_1 + X_2 + X_3, X_4, \dots X_J) \sim Mult(n, (\pi_1 + \pi_2 + \pi_3, \pi_4, \dots, \pi_J))$$

- Finally, conditional on total of some categories each category total has multinomial.
  $$(X_1, X_2, X_3 \mid X_1 + X_2 + X_3 = m) \sim Mult\left(m, \left(\frac{\pi_1}{\pi_1+\pi_2+\pi_3}, \frac{\pi_2}{\pi_1+\pi_2+\pi_3}, \frac{\pi_3}{\pi_1+\pi_2+\pi_3}\right)\right)$$

INFERENTIAL STATISTICS

Statistical Inference

The process of making guesses about the truth from a sample.



Sampling Schemes

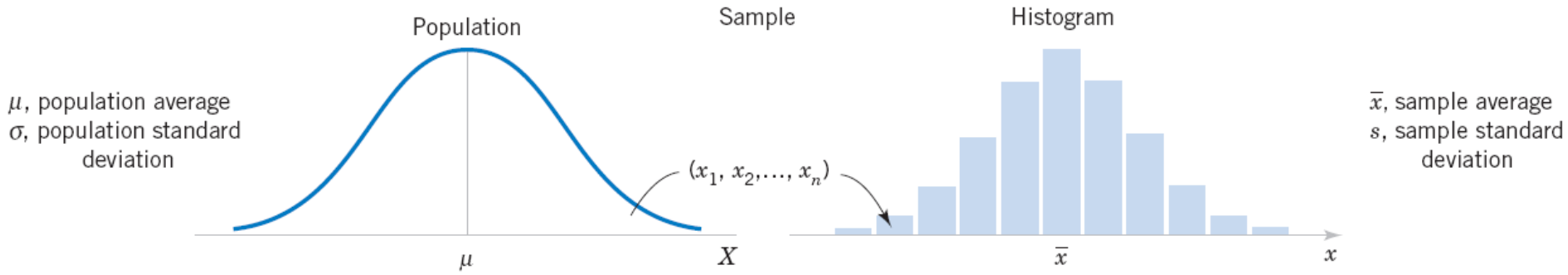
Field surveys use several kinds of sampling schemes to choose a representative sample from a large population.

- **Simple Random Sampling:** There are several ways of sampling from a finite population. A 'simple random sample' (SRS) gives equal 'probability of being sampled' to each unit in the population.
  - SRSWR samples 'with replacement' meaning each population unit can be picked more than once (it is not removed once it is picked).
  - SRSWOR samples 'without replacement' (a sampled unit gets removed from the population, before next unit is picked).

- The following sampling schemes can be used to select a representative small sample from a large population, as a simple random sample from the whole population may sometimes be biased (non-representative) due to chance.
  - **Clustered Sampling:** First divides the population into clusters (example districts) and then takes a random sample from each cluster.

  - **Systematic Sampling:** Here we select units from the population (ordered in some way) in a regular interval, example every 10^{th} unit (e.g., every 10^{th} household in a ward from a listing). If the original ordering is random (with respect to variables of interest), this mimics a random sample from the population.

Statistical Inference



Relationship between a population and a sample.

Statistics vs. Parameters

- **Sample Statistic** – any summary measure calculated from data; e.g., could be a mean, a difference in means or proportions, an odds ratio, or a correlation coefficient
  - E.g., the mean vitamin D level in a sample of 100 men is 63 nmol/L
  - E.g., the correlation coefficient between vitamin D and cognitive function in the sample of 100 men is 0.15

- **Population parameter** – the true value/true effect in the entire population of interest
  - E.g., the true mean vitamin D in all middle-aged and older European men is 62 nmol/L
  - E.g., the true correlation between vitamin D and cognitive function in all middle-aged and older European men is 0.15

Examples of Sample Statistics:

Single sample mean
Single sample proportion
Difference in means (ttest)
Difference in proportions (Z-test)
Odds ratio/risk ratio
Correlation coefficient
Regression coefficient
…

Statistical Models

- Although, in most cases, the populations in question are finite, they are very large. Hence, a good approximation to the large population histogram / frequency table is a theoretical probability distribution. Thus, we assume that the observations/ measurements obtained in our study are random (independent) samples drawn from a theoretical probability distribution.

- For example, we may assume that $X \sim Bin(n, p)$, in which case we might estimate the unknown (true) success probability from our sample using the point estimator $\hat{p} = X/n$.

- Similarly, although there may be finitely many height values in a population $H_1, H_2, \ldots, H_N$, it is large. Hence instead of thinking of the histogram of these N values as our population, we assume that the sample values $X_1, \mathrm{X}_2, \ldots, \mathrm{X}_\mathrm{n}$ are directly drawn from a $Normal(\mu, \sigma^2)$. $\hat{\mu} = \bar{X}$ and $\widehat{\sigma^2} = S^2$ give possible estimators for μ and σ^2.

- Such models are also called parametric models of data. Data is assumed to come from a known theoretical distribution with some unknown parameter values. The goal of inference is to draw conclusions about these unknown parameters.

Point Estimation

- Having chosen a sample from the population, and having collected required data from the sample, we are interested in estimating (guessing) about unknown quantities in the population (average, odds ratio, correlation, hazard ratio, mortality rate etc. in the population).

- When a single value from the sample is used to represent an unknown population parameter, we call it '**point estimation**'.

- **Estimator and estimate:** Some sample statistic may be used as an 'estimator' of a population parameter. Example: $\overline{X}, S^2$, etc. An estimator is a theoretical construct giving a formula or procedure to derive a value given a sample $(X_1, X_2, \dots, X_n)$, where the X_i's are random variables denoting 'sample observations' whose values may be yet unknown.

- Once a sample is collected, the observed data is represented as $(x_1, \dots x_n)$.The realized (observed/calculated) value of the estimator in the current sample is called the 'estimate' $\bar{x} = -2.36, s^2 = 1.56$.

Sampling Variation

- Drawing a sample many times from the population (e.g., by different investigators) will lead to different values of the statistic/estimate.
- If all possible samples could be drawn (e.g. $\mathrm{M} = \binom{N}{n}$ SRSWOR samples), then we would get M different values of our estimate (example sample mean or SD). The histogram of these M values ($\bar{X}_1^*, \bar{X}_2^*, \ldots \bar{X}_M^*$) would represent the sampling distribution of the estimator.
- The **Bias (or sampling bias)** of an estimator is the deviation of the mean of its sampling distribution from the true value of the unknown parameter, i.e., in this case
  - .
- The **Sampling Variance** is a measure of variability of the sampling distribution of the statistic around its sampling mean, not around the true population value μ. The **Standard Error** of an estimator is just the SD of its sampling distribution (i.e., square root of sampling variance).
  - Standard Error of $\bar{X}$ = $\sqrt{Sampling\ Variance\ of\ \bar{\bar{X}}} = \sqrt{Var(\bar{X})} = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}\left(\bar{X}_m^* - \bar{\bar{X}}\right)^2}$
- Note that as the sample size increases , the standard error decreases (more and more overlap between the samples and hence similarity of the estimators) and as $n \to N$, standard error goes to 0 (all samples are same as the population and are identical).

The Central Limit Theorem!

If all possible random samples, each of size *n*, are taken from any population with a mean μ and a standard deviation σ, the sampling distribution of the sample means (averages) will:

1. have mean: $\mu_{\bar{x}} = \mu$

2. have standard deviation: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

3. be approximately normally distributed regardless of the shape of the parent population (normality improves with larger *n*). It all comes back to Z!

Point Estimation

A **point estimate** of some population parameter θ is a single numerical value $\hat{\theta}$ of a statistic $\hat{\Theta}$.

Unknown Parameter θ	Statistic $\hat{\Theta}$	Point Estimate $\hat{\theta}$
μ	$\bar{X} = \frac{\Sigma X_i}{n}$	$\bar{x}$
σ^2	$S^2 = \frac{\Sigma(X_i - \bar{X})^2}{n - 1}$	s^2
p	$\hat{P} = \frac{X}{n}$	$\hat{p}$
$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2 = \frac{\Sigma X_{1i}}{n_1} - \frac{\Sigma X_{2i}}{n_2}$	$\bar{x}_1 - \bar{x}_2$
$p_1 - p_2$	$\hat{P}_1 - \hat{P}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$	$\hat{p}_1 - \hat{p}_2$

Unbiased Estimators

- In a parametric model it is generally easy to calculate the (Sampling) Bias of an estimator.

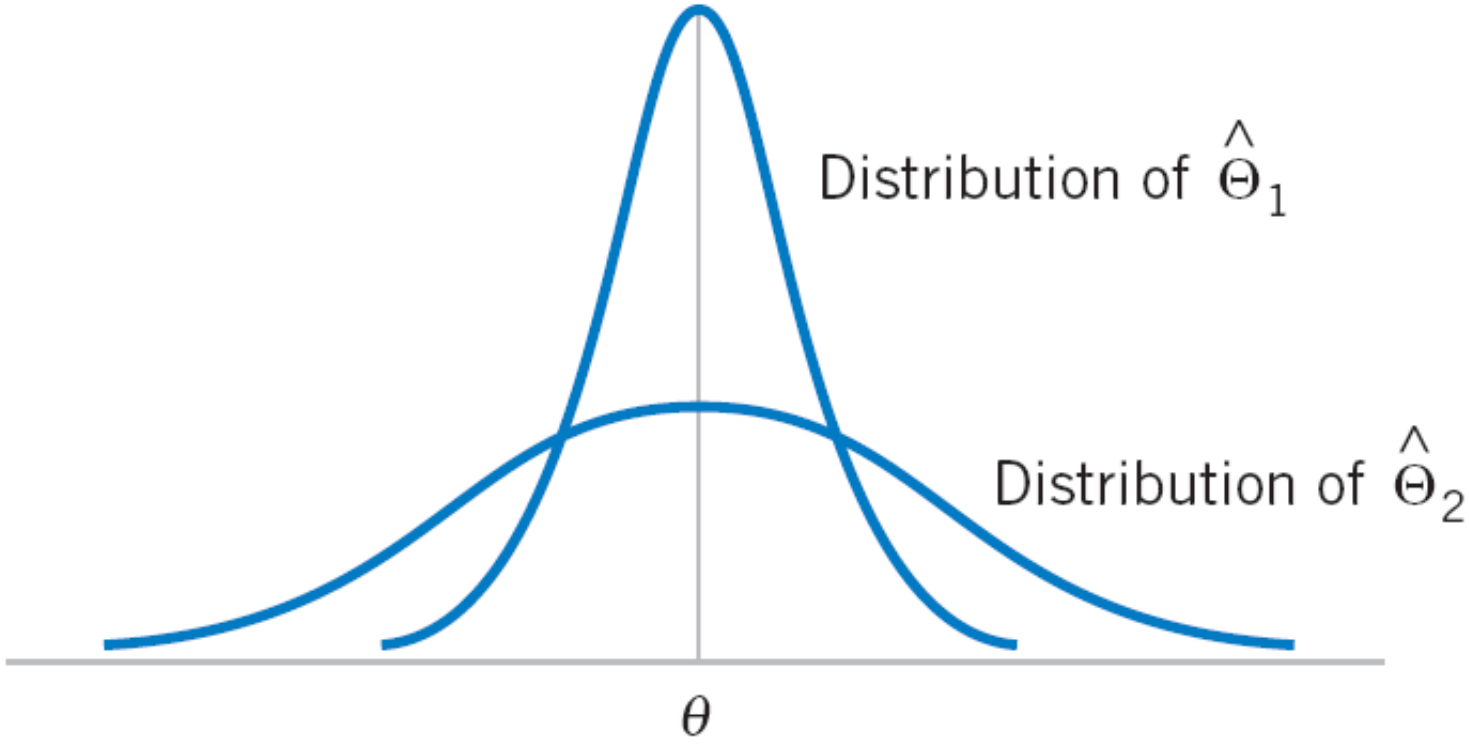
- Assuming $X \sim Bin(n,p)$, $E(\hat{p}) = E(X/n) = \frac{np}{n} = p$. Hence Bias = $E(\hat{p}) - p = 0$. Hence sample proportion is unbiased estimator for population mean.

- Assuming $X_1, \ldots X_n \sim N(\mu, \sigma^2)$, $\mathrm{E}(\hat{\mu}) = E(\bar{X}) = \frac{n\mu}{n} = \mu$. Hence, sample mean is unbiased estimator for population mean.

- In the case above, as shown in the class, if $S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$, the $E(S_0^2) = \frac{n-1}{n}\sigma^2$. Hence $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is an unbiased estimate of σ^2.

- When comparing two unbiased estimators, the one with smaller sampling variance (standard error) should be preferred.

Point Estimation



The sampling distributions of two unbiased estimators $\hat{\Theta}_1$ and $\hat{\Theta}_2$.

Interval Estimation

- In 'Interval Estimation', we want to convey our degree of our belief in our estimate. Hence, we report an interval $[L(x_1, .. x_n), U(x_1, \dots x_n)]$ as our interval estimate of unknown parameter θ.

- We say we have 95% confidence in this interval, if 95% of the intervals generated by repeated sampling from the population will contain the unknown true value θ.

- Mathematically:
  - $P_\theta[L(X_1, .. X_n) \leq \theta \leq L(X_1, \dots X_n)] = 0.95$, for any θ

- Solving the above equation to derive the interval leads to 95 % CI. In general, 0.95 is replaced by $(1 - \alpha)$ to derive a $100\ (1 - \alpha)\%$ CI of θ.

Interval Estimation: Normal Mean

- Assume a random sample $(X_1, .. X_n)$ from a normally distributed population, i.e., $X_i \sim N(\mu, \sigma^2)$, with $\sigma^2 = \sigma_0^2$ being known. The sampling distribution of $\bar{X}$ is $N(\mu, \frac{\sigma_0^2}{n})$.

- To give a $100\ (1-\alpha)\%$ CI around the sample mean, we solve for a half-width w such that

$$P_\mu(\bar{X} - w \leq \mu \leq \bar{X} + w) = (1-\alpha)$$

$$P_\mu\left(-\frac{w}{\sigma_0/\sqrt{n}} \leq \frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} \leq \frac{w}{\sigma_0/\sqrt{n}}\right) = (1-\alpha)$$

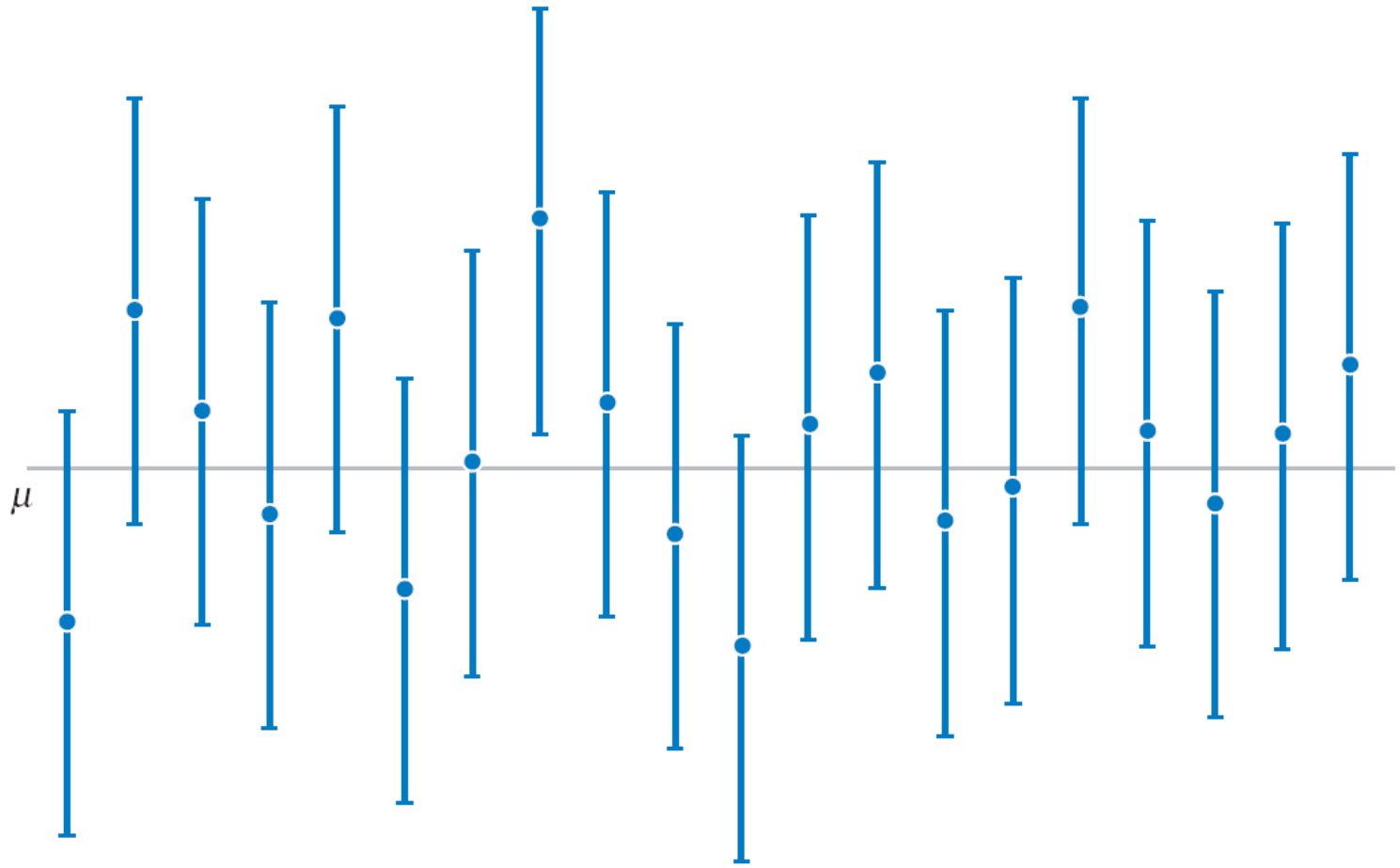
- Noting that $Z = \frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} \sim N(0,1)$, we choose $\frac{w}{\sigma_0/\sqrt{n}}$ to be $z_{\alpha/2}$ the $(1-\alpha/2)$'th quantile of the standard normal distribution.

- In other words, $100\ (1-\alpha)\%$ CI of μ is

$$[\bar{X} - z_{\alpha/2} \times\ \sigma_0/\sqrt{n}, \bar{X} + z_{\alpha/2} \times\ \sigma_0/\sqrt{n}]$$

Inference on the Mean of a Population, Variance Known

95% Confidence Interval for the Mean



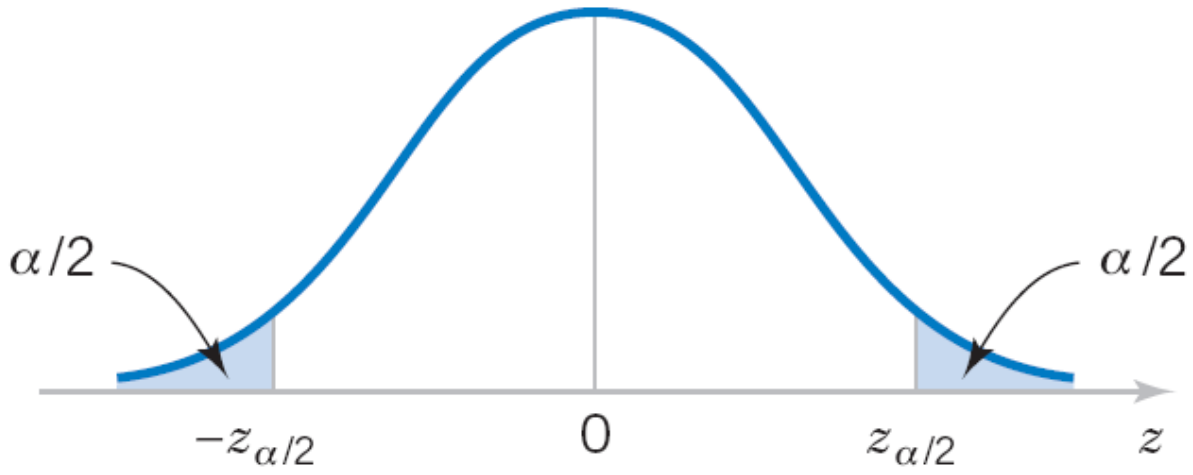
Repeated construction of a confidence interval for μ.

Inference on the Mean of a Population, Variance Known

Confidence Interval on the Mean

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

$$P\{-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\} = 1 - \alpha$$



The distribution of Z.

Interval Estimation

- The above CI is based on exact normality of $\bar{X}$. In general, for any estimator that looks like an average, e.g., Binomial proportion $\hat{p}$, the 'central limit theorem' says that the estimator is approximately normally distributed for large sample, meaning:
- $\hat{\theta} \sim N\left(\theta, SE\left(\hat{\theta}\right)\right)$
- Hence, approximate $100\ (1-\alpha)\%$ CI of θ is

$$[\hat{\theta} - z_{\alpha/2} \times SE\left(\hat{\theta}\right),\ \hat{\theta} + z_{\alpha/2}\ \times SE(\hat{\theta})]$$

- In most cases, the formula of SE contains unknown parameters, e.g. $SE(\hat{p}) = \sqrt{Var(X/n)} = \sqrt{p\ q/n}$. In such cases, a point estimate from the sample is plugged-in to get an $\widehat{SE}$ and hence an approximate CI. Example 95 % CI for p :

$$[\hat{p} - 1.96 \times \sqrt{\hat{p}\ \hat{q}/n}\ , \hat{p} +\ 1.96\ \times \sqrt{\hat{p}\ \hat{q}/n}]$$

Hypothesis Testing

- **Null Hypothesis:** The current or default belief about the population parameter(s). E.g., 1) H_0: $\rho = 0$ (no correlation) or 2) H_0: $OR = 1$ (no association) or 3) H_0: $\mu_1 = \mu_2$ (no difference in means), etc.
- **Alternative Hypothesis:** The investigators belief, which is to be tested. Example H_1: $\rho < 0$ (negative correlation between HDL and LDL) or H_0: $OR \neq 1$ (presence of association between a SNP and a disease) or H_0: $\mu_1 > \mu_2$ (higher mean expression i.e., upregulation)

- **Test Statistic:** Any 'statistic' $T(X)$ which is expected to have small (close to zero) values if H_0 were true and large (away from 0) values if H_1 were true. E.g. 1) $T(X)$ =sample correlation r or 2) sample odds-ratio $\mathrm{T(X)} = \widehat{OR}$ or 3) $\mathrm{T}(X) = \bar{X}_1 - \bar{X}_2$
- **Standardized Test Statistic:** For convenience, sometimes the test statistic may be shifted and scaled so that it has a known distribution (under the null hypothesis) that does not depend on any parameters. This is usually done as follows:

$$Z = \frac{[T - E(T \mid \theta_0)]}{SE(T \mid \theta_0)} = \frac{T - m_0}{s_0}$$

- **Decision Rule/Rejection Region:** We reject H_0 when the value of the test statistic is larger (or smaller or larger in magnitude), than a threshold c , depending on whether the test is 1) lower-tailed (left tailed) $T < c_1$ or $Z < c_2$) Upper tailed (right tailed) $T > c_1$ or $Z > c_2$ or 3) two-tailed | T | $> c_1$ or | Z | $> c_2$. Note that the decision rule (rejection region) will remain the same whether we use a standardized or a non-standardized test statistic, the cutoff value for T and Z will be different related by $c_2 = \frac{c_1 - m_0}{s_0}$

Hypothesis Testing

Statistical Hypotheses

Two-sided Alternative Hypothesis

$$H_0: \mu = 50 \text{ cm/s}$$
$$H_1: \mu \neq 50 \text{ cm/s}$$

One-sided Alternative Hypotheses

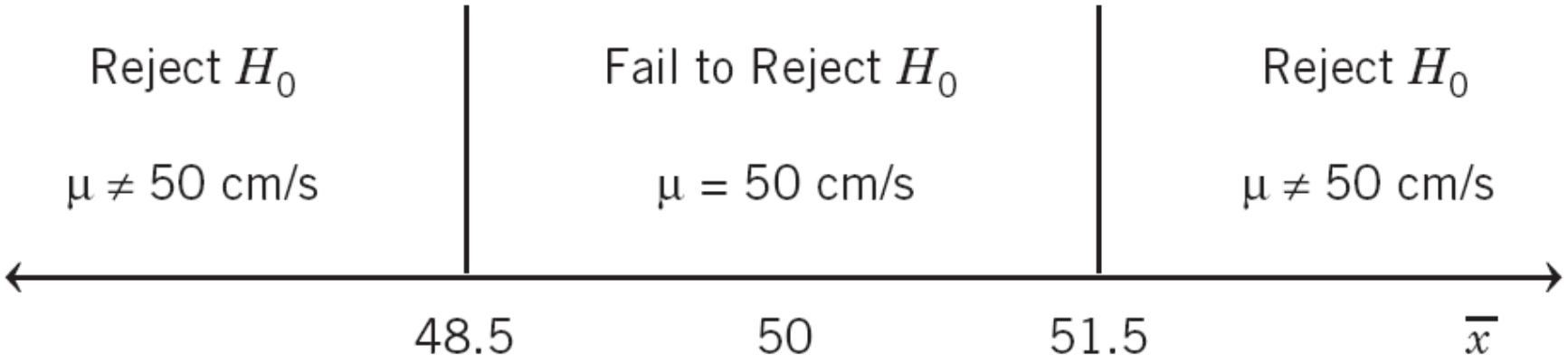
$H_0: \mu = 50$ cm/s $\quad H_1: \mu < 50$ cm/s $\quad$ or $\quad H_0: \mu = 50$ cm/s $\quad H_1: \mu > 50$ cm/s

Hypothesis Testing

Testing Statistical Hypotheses

$$H_0: \mu = 50 \text{ cm/s}$$

$$H_1: \mu \neq 50 \text{ cm/s}$$



Decision criteria for testing $H_0: \mu = 50$ cm/s versus $H_1: \mu \neq 50$ cm/s.

Hypothesis Testing

Testing Statistical Hypotheses

Rejecting the null hypothesis H_0 when it is true is defined as a **type I error.**

Failing to reject the null hypothesis when it is false is defined as a **type II error.**

Hypothesis Testing

Testing Statistical Hypotheses

Decisions in Hypothesis Testing

Decision	H_0 Is True	H_0 Is False
Fail to reject H_0	No error	Type II error
Reject H_0	Type I error	No error

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true})$$

The type I error probability is also called the **significance level**, of the test.

Hypothesis Testing

- It is not possible to reduce both Type-1 Error and Type-2 Error simultaneously. If c increases, Type-1 Error decreases (test becomes stringent, False Positives reduce), but Type-2 Error increases (False Negatives Increase). If c decreases, the reverse happens.

- Since preventing False Positives (Type-1 Error) is generally more critical, we determine cutoff-s by limiting the False Positive Rate (Probability of Type-1 Error).

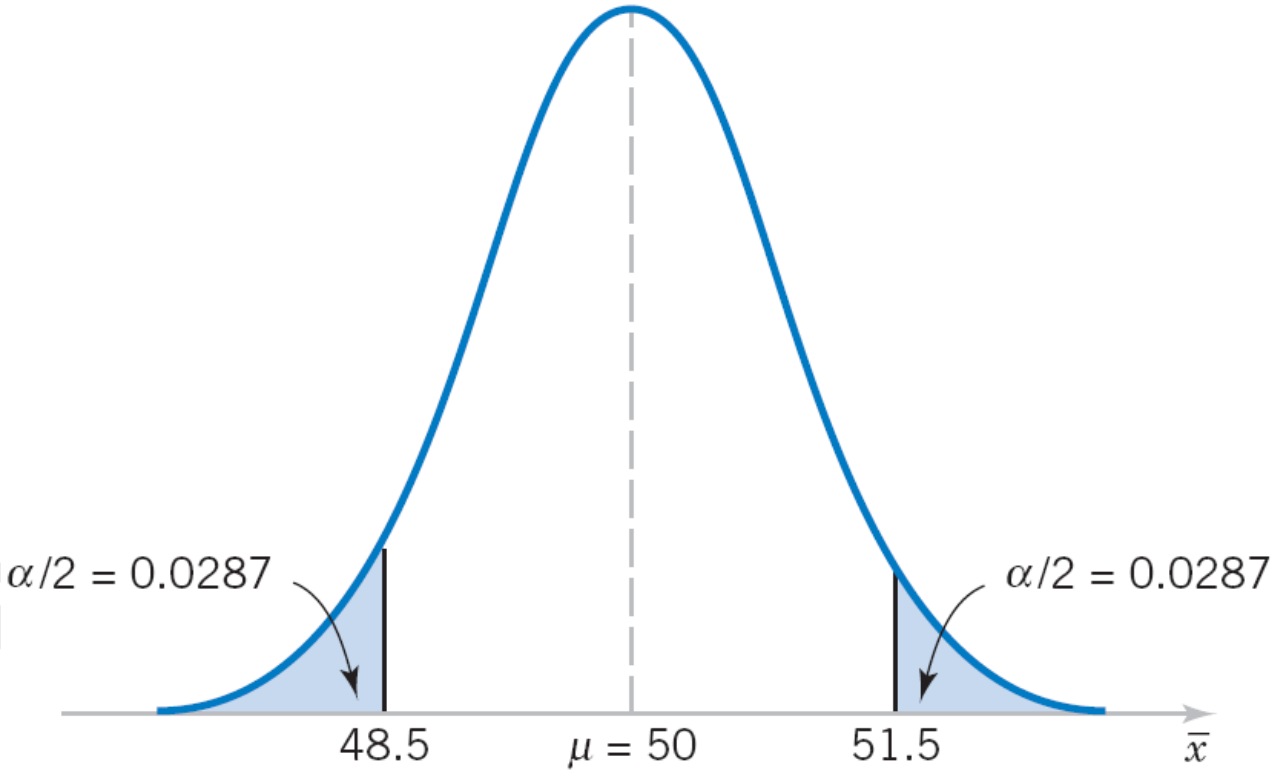
- **Level of Significance** (α): The target (maximum allowed) type-1 error probability.

- **Critical Value** (c_α) is the cutoff for the test-statistic which gives probability of type-1 error equal to α.

- **Null Distribution:** The null hypothesis distribution of the test-statistic which is required to derive the critical value (rejection region).

Hypothesis Testing

Testing Statistical Hypotheses



The critical region for H_0: $\mu = 50$ versus H_1: $\mu \neq 50$ and $n = 10$.

Hypothesis Testing: Critical Value Approach

- $X_1, \ldots . X_n \sim N(\mu, \sigma^2)$, with $\sigma^2 = \sigma_0^2$ known. We want to test:
- One-tailed (left): $H_0: \mu = \mu_0$ vs $H_1: \mu < \mu_0$
  - Test statistic: $T = \bar{X} - \mu_0$
  - Standardized Test Statistic: $Z = \frac{\bar{X} - \mu_0}{\sigma_0 / \sqrt{n}}$
  - Null Distribution: $Z \sim N(0,1)$ under H_0
  - Critical Value at level α : c_α such that $P(N(0,1) < c_\alpha) = \alpha$, hence $c_\alpha = z_\alpha$ (the $\alpha' th$ quantile of standard normal).
  - Rejection Region: $Z < c_\alpha$

- One-tailed (right): $H_0: \mu = \mu_0$ vs $H_1: \mu > \mu_0$
  - Critical Value at level α : $c_\alpha = z_{1-\alpha}$
  - Rejection Region: $Z > c_\alpha$

- Two-tailed: $H_0: \mu = \mu_0$ vs $H_1: \mu \neq \mu_0$
  - Critical value at level α : $c_\alpha = z_{1-\alpha/2}$ (Equal prob of type-1 error on both sides =$\alpha/2$)
  - Rejection Region: $|Z| > c_\alpha$ i.e. $\{Z > c_\alpha \; or Z < -c_\alpha\}$

- Example decision statement : **Rejection of H_0: <u>We reject the null hypothesis</u>** that 'population mean of X is μ_0' at 5% level of significance. **Acceptance of H_0: <u>We are unable to reject the null hypothesis</u>** that 'population mean of X is μ_0' at 5% level of significance.

Inference on the Mean of a Population, Variance Known

Hypothesis Testing on the Mean

We wish to test:

$$H_0: \mu = \mu_0$$
$$H_1: \mu \neq \mu_0$$

The **test statistic** is:

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \qquad (4\text{-}13)$$

Inference on the Mean of a Population, Variance Known

Hypothesis Testing on the Mean

Reject H_0 if the observed value of the test statistic z_0 is either:

$$z_0 > z_{\alpha/2} \qquad z_0 < -z_{\alpha/2}$$

Fail to reject H_0 if

$$-z_{\alpha/2} \leq z_0 \leq z_{\alpha/2}$$

Hypothesis Testing: P-Value Approach

The Critical value needs to be recalculated if target level of significance changes.

- **P-value:** An alternative approach is to report a p-value defined as '**the smallest level α for which the test will be rejected**'. So, p-value=0.0001 means, that the test (H_0) will be rejected for all levels $\alpha \geq 0.0001$. But if $\alpha = 0.000099$, we will be unable to reject the test.

- **Rejection Rule:** From the above definition, it is clear that once the p-value is calculated and reported, the test can be rejected whenever p-value $< \alpha$, where α is any arbitrary target level of significance.

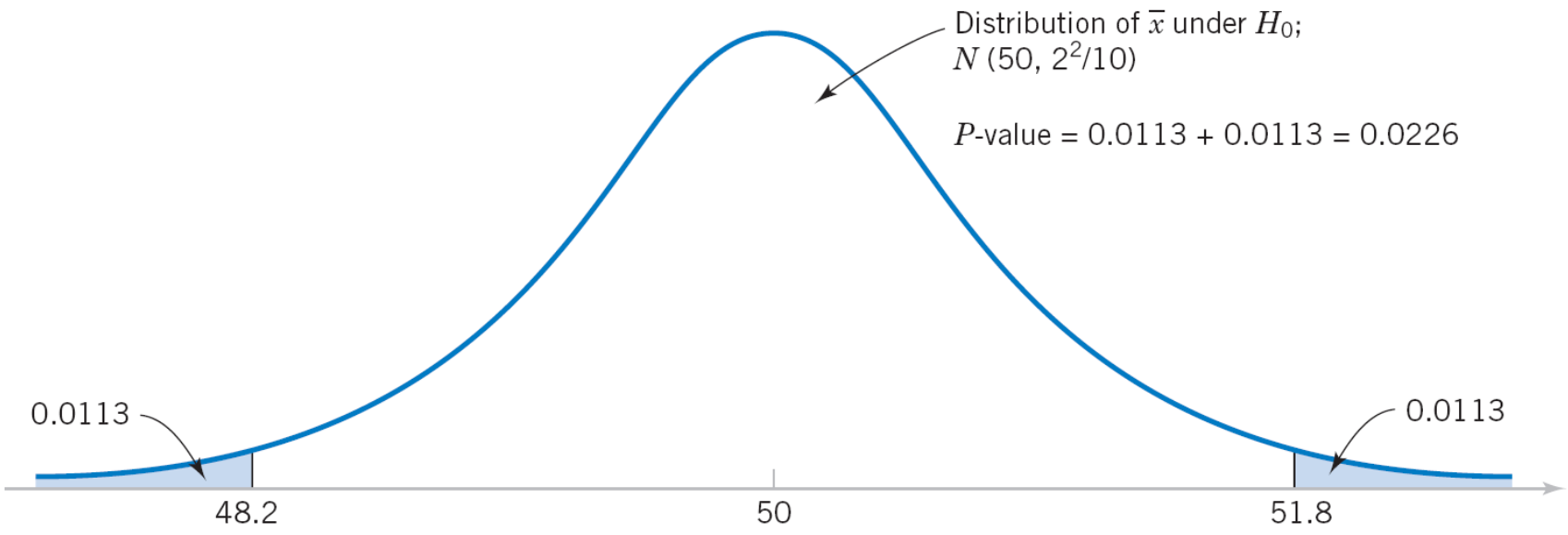
- To calculate the P-value, note that as we decrease the level of significance, the c_α increases, until the c_α coincides with t_{obs} the observed value of the test statistic. Decreasing α below this, will lead to $c_\alpha > t_{obs}$, hence H_0 will be accepted. Thus, p-value can be obtained by solving $c_{\alpha_{\min}} = t_{obs}$. From the definition of c_α (for a right-tailed test), this is same as solving: $P(T > t_{obs}) = \alpha_{min} = p - value$.

- Therefore, we can also say **that p-value = Chance of observing a more extreme T value than the observed value (under repeated sampling assuming the null hypothesis).** We can calculate p-values from:
  - One tailed (left): P-value = $P(T < t_{obs}$ | Null Distribution)
  - One tailed (right): P-value = $P(T > t_{obs}$ | Null Distribution)
  - Two tailed: P-value = $P(|T| > |t_{obs}|$ | Null Distribution) = $2 * P(T > |t_{obs}|$ | Null Distribution)

Hypothesis Testing

P-Values in Hypothesis Testing

The ***P*-value** is the smallest level of significance that would lead to rejection of the null hypothesis H_0.



Inference on the Mean of a Population, Variance Known

Hypothesis Testing on the Mean

Testing Hypotheses on the Mean, Variance Known

Null hypothesis: $H_0: \mu = \mu_0$

Test statistic: $Z_0 = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$

Alternative Hypotheses	*P*-Value	Rejection Criterion for Fixed-Level Tests
$H_1: \mu \neq \mu_0$	Probability above z_0 and probability below $-z_0$, $P = 2[1-\Phi(\lvert z_0 \rvert)]$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_1: \mu > \mu_0$	Probability above z_0, $P = 1-\Phi(z_0)$	$z_0 > z_\alpha$
$H_1: \mu < \mu_0$	Probability below z_0, $P = \Phi(z_0)$	$z_0 < -z_\alpha$

Here $\Phi()$ denotes the CDF of N(0,1). Same as pnorm() in R.

Inference on the Mean of a Population, Variance Known

Confidence Interval on the Mean

Relationship between Tests of Hypotheses and Confidence Intervals

If $[l,u]$ is a $100(1 - \alpha)$ percent confidence interval for the parameter, then the test of significance level α of the hypothesis

$$H_0: \theta = \theta_0$$

$$H_1: \theta \neq \theta_0$$

will lead to rejection of H_0 if and only if the hypothesized value is not in the $100(1 - \alpha)$ percent confidence interval $[l, u]$.

This is because with simple calculation it can be shown that , $\theta_0 \in [\widehat{\theta} - w(\alpha), \widehat{\theta} + w(\alpha)]$ is same as saying $| \widehat{\theta} - \theta_0 | < c_\alpha$ (acceptance region of test). Thus, 2-sided $100(1-\alpha)\%$ CI can give the hypothesis testing decision for a 2-tailed test at level α.

Unknown Variance

- Recall that the Chi-square Distribution with k df is a sum of squares of k independent standard normal variables.
- If mean is known, $\mu = \mu_0$
  $X_1, X_2, \ldots X_n \sim N(\mu_0, \sigma^2), \sum_i \left(\frac{(X_i-\mu}{\sigma}\right)^2 = \frac{nS_0^2}{\sigma^2} \sim \chi^2(df = n)$

- If mean is estimated from the data as $\bar{X}$, the n terms are not independent, in fact $\sum(X_i - \bar{X})$ = 0 (total deviation is 0).

- In this case it can be shown that $\sum_i \left(\frac{(X_i-\bar{X}}{\sigma}\right)^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(df = n-1)$

- Also, each deviation $X_i - \bar{X}$ and hence sample variance S^2 is independent of the mean $\bar{X}$.

- **Student's t-distribution** is defined as the distribution of the variable
  $t = \frac{Z}{\sqrt{Y/k}}$ where $Z \sim N(0,1)$ and $Y \sim \chi^2(k)$ with $Z \; and \; Y$ being independent.

- The t-distribution is a symmetric around 0 (and flatter than the standard normal). As the df k increases to infinity, t gradually goes to Z (standard normal).
  - Note that $E\left(\frac{Y}{k}\right) = 1$ and $Var\left(\frac{Y}{k}\right) = \frac{2}{k} \to 0$. Therefore, $\frac{Y}{k} \to 1$, so $t \to Z$.

Inference on the Mean of a Population, Variance Unknown

4-5.1 Hypothesis Testing on the Mean

Let $X_1, X_2, \ldots, X_n$ be a random sample for a normal distribution with unknown mean μ and unknown variance σ^2. The quantity

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a t distribution with $n - 1$ degrees of freedom.

One-Sample t-test

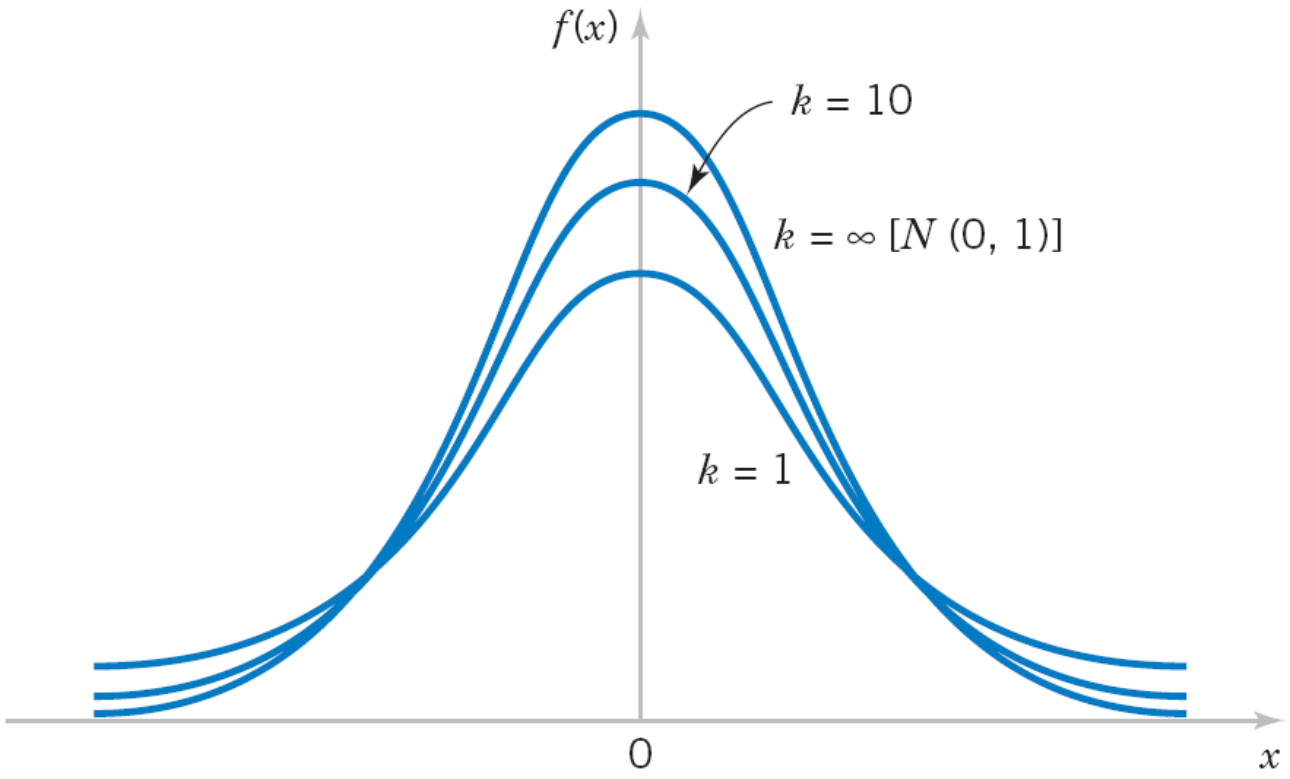
- We assume $X_1, X_2, \ldots X_n \sim N(\mu, \sigma^2)$, with unknown variance. We want to test $H_0: \mu = \mu_0$ *vs.* $H_1: \mu > \mu_0$
- We modify the previous test statistic by plugging in S in place of σ.
- Student's one-sample t-statistic, $T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$
- T has Student's t-distribution with $n - 1$ df, because T can be re-written as

$$T = \frac{\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}}{S/\sigma} = \frac{\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}}{\sqrt{\frac{\frac{(n-1)S^2}{\sigma^2}}{(n-1)}}} \sim \frac{Z}{\sqrt{Y/k}} \sim t_{n-1}$$, with $Y \sim \chi^2(n-1)$ and $k = (n-1)$.

- Thus critical values and p-value of the statistic T can be obtained from Student's t-distribution with $(n-1)$ degrees of freedom. For large samples, the df becomes large, hence critical values and p-values from standard normal give a good approximation. Another way to see why this happens is that the ratio $\frac{S}{\sigma} \to 1$. So, the known variance and unknown variance cases coincide, $S \approx \sigma$ (the unknown SD).

Inference on the Mean of a Population, Variance Unknown

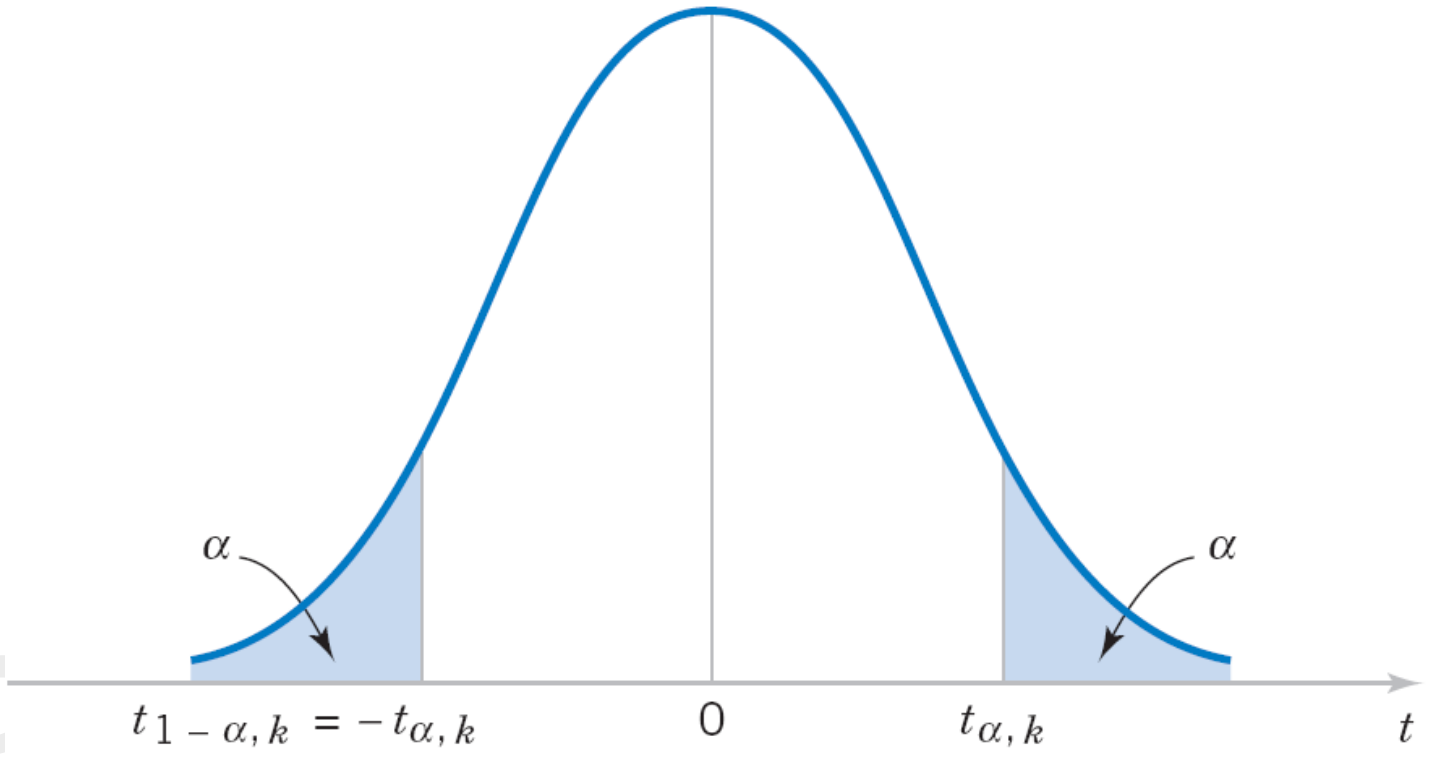
Hypothesis Testing on the Mean



Probability density functions of several t distributions.

Inference on the Mean of a Population, Variance Unknown

Hypothesis Testing on the Mean

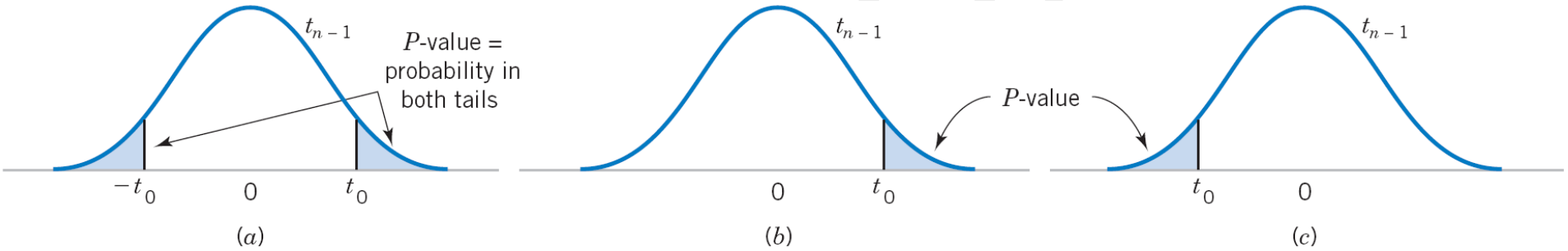


Percentage points of the t distribution.

Inference on the Mean of a Population, Variance Unknown

Hypothesis Testing on the Mean

Calculating the P-value



Calculating the P-value for a t-test: (a) H_1: $\mu \neq \mu_0$; (b) H_1: $\mu > \mu_0$; (c) H_1: $\mu < \mu_0$.

Inference on the Mean of a Population, Variance Unknown

4-5.1 Hypothesis Testing on the Mean

Testing Hypotheses on the Mean of a Normal Distribution, Variance Unknown

Null hypothesis: $H_0: \mu = \mu_0$

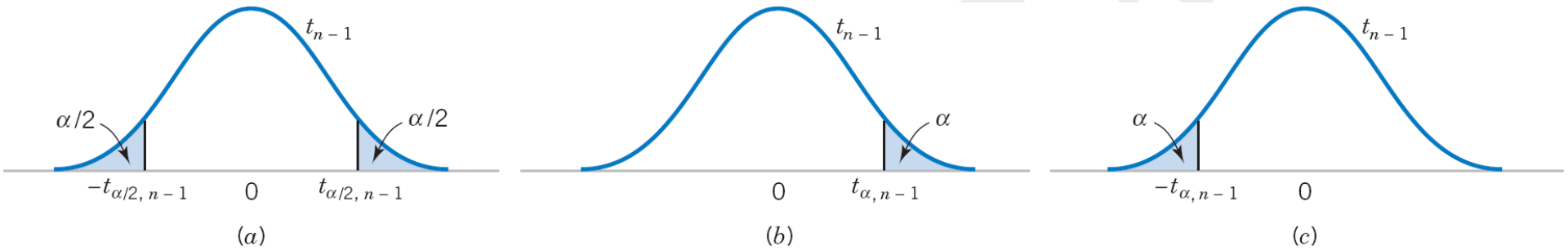
Test statistic: $T_0 = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$

Alternative Hypotheses	*P*-Value	Rejection Criterion for Fixed-Level Tests
$H_1: \mu \neq \mu_0$	Sum of the probability above t_0 and the probability below $-t_0$	$t_0 > t_{\alpha/2,n-1}$ or $t_0 < -t_{\alpha/2,n-1}$
$H_1: \mu > \mu_0$	Probability above t_0	$t_0 > t_{\alpha,n-1}$
$H_1: \mu < \mu_0$	Probability below t_0	$t_0 < -t_{\alpha,n-1}$

The locations of the critical regions for these situations are shown in Fig. 4-19*a*, *b*, and *c*, respectively.

Inference on the Mean of a Population, Variance Unknown

Hypothesis Testing on the Mean



The distribution of T_0 when H_0: $\mu = \mu_0$ is true, with critical region for (a) H_1: $\mu \neq \mu_0$, (b) H_1: $\mu > \mu_0$, and (c) H_1: $\mu < \mu_0$.

The two-sample T-test

- Is the difference in means that we observe between two groups more than we'd expect to see based on chance alone?

Two-Sample t-test

- We assume $X_1, X_2, \ldots X_m \sim N(\mu_1, \sigma^2)$, $Y_1, Y_2, \ldots Y_n \sim N(\mu_2, \sigma^2)$ with **unknown (but equal) variance**. We want to test $H_0: \mu_1 = \mu_2$ *vs.* $H_1: \mu_1 > \mu_2$

- Note that $Var(\bar{X} - \bar{Y}) = \sigma^2 \left(\frac{1}{m} + \frac{1}{n}\right)$
- Student's two-sample t-statistic, $T = \frac{\bar{X} - \bar{Y}}{S\sqrt{\frac{1}{m}+\frac{1}{n}}}$, where $S = S_{pooled} = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{(m+n-2)}$

- T has Student's t-distribution with $\mathrm{m} + n - 2$ df, because T can be re-written as

$$T = \frac{\frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{m}+\frac{1}{n}}}}{S/\sigma} = \frac{\frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{m}+\frac{1}{n}}}}{\sqrt{\frac{\frac{(m+n-2)S^2}{\sigma^2}}{(m+n-2)}}} \sim \frac{Z}{\sqrt{Y/k}} \sim t_{m+n-2}$$, with $Y \sim \chi^2(m+n-2)$ and $k = (m+n-2)$.

- Thus, critical values and p-value of the statistic T can be obtained from Student's t-distribution with $(m+n-2)$ degrees of freedom. When **either $\boldsymbol{m}$ or $\boldsymbol{n}$ or both are large**, the df becomes large, hence critical values and p-values from standard normal give a good approximation.

ttest, pooled variances

$$T = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{s_p^2}{n} + \frac{s_p^2}{m}}} \sim t_{n+m-2}$$

$$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$

Pooled variance

If you assume that the standard deviation of the characteristic (e.g., IQ) is the same in both groups, you can pool all the data to estimate a common standard deviation. This maximizes your degrees of freedom (and thus your power).

pooling variances:

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}{n-1} \quad \text{and} \quad (n-1)s_x^2 = \sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

$$s_y^2 = \frac{\sum_{i=1}^{m}(y_i - \bar{y}_m)^2}{m-1} \quad \text{and} \quad (m-1)s_y^2 = \sum_{i=1}^{m}(y_i - \bar{y}_m)^2$$

$$\therefore s_p^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 + \sum_{i=1}^{m}(y_i - \bar{y}_m)^2}{n+m-2}$$

$$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$

Degrees of Freedom!

Paired t-test

- Suppose we have paired data on the same units e.g., before and after treatment: $(X_i, Y_i), i = 1, \dots, n$
- Assume $X_i \sim N(\mu_1, \sigma_1^2)$ and $Y_i \sim N(\mu_2, \sigma_2^2)$. We want to test $H_0: \mu_1 = \mu_2$. If we assume the correlation is ρ ,
- $D_i = X_i - Y_i \sim N(\delta = \mu_1 - \mu_2, \sigma^2), i = 1, \dots n$
- Where $\sigma^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$
- Therefore, testing $\mu_1 = \mu_2$ is same as testing $\delta = 0$ with the difference data $D_1, \mathrm{D}_2, \dots, \mathrm{D}_\mathrm{n}$
- This can be done using a 1-sample t-test:

$$T = \frac{\overline{D} - 0}{S_D/\sqrt{n}} = \frac{\bar{X} - \bar{Y}}{S_D/\sqrt{n}} \sim t_{df=n-1}$$

- The above test is known as a paired t-test.

The multi-group F-test

- Is the differences in means that we observe between two or more groups more than we'd expect to see based on chance alone?

ANOVA- One way

- Model: We assume a normally distributed random variable X whose means are possibly different across I groups, but variances are equal.

$$X_{ji} = \mu_j + \epsilon_{ji}, i = 1,2 \dots, n_j\ , j = 1, \dots I\ , \sum_{j=1}^{I} n_j = N$$
$$\epsilon_{ji} \sim N(0, \sigma^2)$$

- We want to test H_0: $\mu_1 = \mu_2 = \cdots = \mu_I$ vs H_1: At least two of the means are unequal.

- Test Statistic is $F = \frac{Between\ Group\ Variance}{Pooled\ Within\ Group\ Variance}$

- F has an F-distribution with df1=$(I-1)$ and df2=$(N-I)$

- The F-distribution is defined as: $F = \frac{Y_1/k_1}{Y_2/k_2}$, where Y_1 and Y_2 are independent χ^2 distributed variables with df-s k_1 and k_2.

- The critical values and p-values of the test can be calculated using a right-tailed rejection region $F > c$.

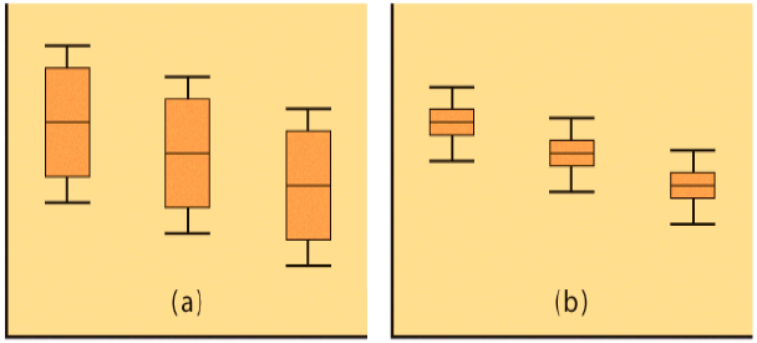
- When N is large enough the denominator is a very accurate estimator for the true variance σ^2 and hence, F can be approximated by $\chi^2(df = I-1)$.

ANOVA- One way

The**ANOVA F-statistic**compares variation due to specific sources (levels of the factor) with variation among individuals who should be similar (individuals in the same sample).

$$F = \frac{\text{variation among sample means}}{\text{variation among individuals in same sample}}$$

Difference in means small relative to overall variability



Difference in means large relative to overall variability

➔ F tends to be small ➔ F tends to be large

Larger F-values typically yield more significant results. How large depends on the degrees of freedom ($I-1$ and $N-I$).

One-way ANOVA F-statistic

- An F random variable with (df1=1, df2=k) is simply the square of a t distributed random variable with df=k.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$F = \frac{\left[n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \cdots + n_I(\bar{x}_I - \bar{x})^2\right]/(I-1)}{\left[(n_1-1)s_1^2 + (n_2-1)s_2^2 + \cdots + (n_I-1)s_I^2\right]/(N-I)}$$

- **Relationship with t-test:** It can be shown that for 2-groups: $F = t^2$, where F is the ANOVA statistic and t is the 2-sample t-statistic. So, these two tests are equivalent. But for two groups, t is more flexible as it can be used to do one-tailed tests and there are modifications available for unequal variance.

Proportion Tests

- Are the differences in proportions that we observe between one or more groups more than we'd expect to see based on chance alone?
  - One sample (current sample versus previously known proportion)
  - Two sample (compare two groups)
  - Chi-square Test (compare proportions for multiple groups and/or multiple categories)

Inference on Population Proportion

Hypothesis Testing on a Binomial Proportion

We will consider testing:

$$H_0: p = p_0$$
$$H_1: p \neq p_0$$

Let X be the number of observations in a random sample of size n that belongs to the class associated with p. Then the quantity

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \tag{4-64}$$

has approximately a standard normal distribution, $N(0, 1)$.

Inference on Population Proportion

Hypothesis Testing on a Binomial Proportion

Testing Hypotheses on a Binomial Proportion

Null hypotheses: $H_0: p = p_0$

Test statistic: $Z_0 = \dfrac{X - np_0}{\sqrt{np_0(1 - p_0)}}$

Alternative Hypotheses	*P*-Value	Rejection Criterion for Fixed-Level Tests
$H_1: p \neq p_0$	Probability above z_0 and probability below $-z_0$, $P = 2[1 - \Phi(\lvert z_0 \rvert)]$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_1: p > p_0$	Probability above z_0, $P = 1 - \Phi(z_0)$	$z_0 > z_\alpha$
$H_1: p < p_0$	Probability below z_0, $P = \Phi(z_0)$	$z_0 < -z_\alpha$

Two-group Binomial Proportion Test

- We have data $X \sim Bin(m, p_1)$ and $Y \sim Bin(n, p_2)$. We want to test, $H_0 : p_1 = p_2$ vs. say $H_1 : p_1 > p_2$.

- Test Statistic: $Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\,\hat{q}\left(\frac{1}{m}+\frac{1}{n}\right)}}$, where $\hat{p} = \frac{X+Y}{m+n}$, the pooled estimator

  the true success probability under H_0 is used to estimate the SE of the difference in proportions.

- The critical values or p-value can be determined from standard normal distribution for sufficiently large sample sizes.

- For both 1-sample and 2-sample test of proportions, when n is small, exact-test based on binomial CDF needs to be done, as the normal approximation may not work.

Test of Independence in a Contingency Table

Y \ X	1	2	… j …	J	Totals
1	n_{11}	n_{12}	… n_{1j} …	n_{1J}	n_{1+}
2	n_{21}	n_{21}	… n_{2j} …	n_{2J}	n_{2+}
: i :	: n_{i1} :	: n_{i2} :	: … n_{ij} … :	: n_{iJ} :	: n_{i+} :
I	n_{I1}	n_{I2}	… n_{Ij} …	n_{IJ}	n_{I+}
Totals	$\boldsymbol{n_{+1}}$	$\boldsymbol{n_{+2}}$	… $\boldsymbol{n_{+j}}$ …	$\boldsymbol{n_{+J}}$	$\boldsymbol{n_{++}}$=Grand Total

- For two categorical variables, bivariate joint distribution can be summarized into a contingency table ($I \times J$).
- Under H_0: X and Y are independent. Hence $\pi_{ij} = \pi_i \pi_j$, or$E(n_{ij}) = \mathrm{e_{ij}} = \frac{n_{i+} \times n_{+j}}{n_{++}}$.
- So, an expected table is given by the formula : E(Count)=Row Total × ColTotal / Grand Total
- Chi-squared statistic $T = \sum_i \sum_j (n_{ij} - e_{ij})^2 / e_{ij} = \sum_{cells} \frac{(Obs\ Count - Exp\ Count)^2}{Exp\ Count}$
- Provided all cell values are large enough, $T \sim \chi^2$ with $df = (I-1)(J-1)$. P-value or critical values can be calculated based on upper tail of this distribution. In R this test can be done using chisq.test() or prop.test().
- When some cell values are small, Fisher's Exact test can be done which is highly computationally intensive for large tables.

Multiple Testing Correction

- Conventionally a type-1 error rate of 5% i.e $\alpha = 0.05$ is used for single hypothesis tests.

- However, if an investigator is testing multiple hypotheses (e.g., A) testing all 16,000 genes for differential expression or B) even testing 5 candidate genes), using the same criterion for each hypothesis test will lead to a higher probability of type-1 error in the study. To see this note that $\#False\ Positives \sim Bin(Number\ of\ Tests, \alpha)$. So, higher the number of tests, greater the chances of reporting several false positives.

  - To maintain the same global type-1 error rate of $\alpha = 0.05$ for the whole study (defined as 'chance of reporting even one false positive result'), we need to allow smaller type-1 error for each test.
  - For example, if we allow $\alpha_j' = \alpha/M$ for each of the $j = 1, \ldots M$ tests, this strategy guarantees that the global type-1 error probability will be less than α.
  - The above 'Multiple Testing correction' strategy is called 'Bonferroni' correction, which is a stringent method. There are other more refined multiple testing procedures.
  - Bonferroni correction implies testing each gene at $\alpha' = \frac{0.05}{16000} = 3.1e^{-6}$ for case (A) and $\alpha' = \frac{0.05}{5} = 0.01$ for case (B).

Multiple Testing Correction

- Multiple testing correction leads to stringent critical value (rejection region), hence unless the observed test statistic is very large (or equivalently p-value very small), we will be unable to reject the test (i.e., discover the DE gene).

- Statistical replication on independent dataset is important whether or not multiple testing is involved. On replication on an independent dataset, probability of type-1 error becomes $0.05^2 = 0.0025$. It is particularly important to do a replication study when weaker multiple testing criteria are used to increase the number of discoveries. If multiple tests are done in the replication phase, a Bonferroni correction is again required.

Power

- Power is defined as the probability of detecting an association, when it is truly there.
- Pr(Reject Test | H_1) = 1 – Pr(Type II Error)
- Thus, power is simply the area under the alternative distribution of the test statistic of the 'rejection region'.
- What does power depend on? Recall that for 1 sample Z-test for normality $E(Z \mid \mu) = \frac{\mu - \mu_0}{\sigma/\sqrt{n}} = \frac{\delta}{\sqrt{n}} = \frac{Effect\ SIze}{\sqrt{Sample\ Size}}$, and Power =$P(Z > z_\alpha \mid Mean = \mu)$
  - Sample size : As $n \uparrow$, alternative distribution shifts to the right, power ↑
  - Effect size : As $\delta \uparrow$, alternative distribution shifts to the right, power ↑
  - Level of significance: As $\alpha \uparrow$, the cutoff z_α shifts to the left, power ↑
- Similar reasoning as above can be given for the unstandardized test statistic T: $T \sim N(\mu - \mu_0, \sigma^2/n)$
  - Sample size : As $n \uparrow$, both the null and the alternative distributions become less spread (less overlap), power ↑
  - Effect size : As $\delta \uparrow$, alternative distribution shifts to the right or overlap decreases, power ↑
  - Level of significance: As $\alpha \uparrow$, the cutoff c_α shifts to the left, power ↑

Power and Sample Size

For any statistical test, Power is some function of

1. Sample size (n).

2. Level of significance (α).
   a) Usually, α=0.05
   b) Sometimes α = 0.05/#Tests (Bonferroni Correction)

3. Effect Size (δ): How far is the true alternative distribution expected to be from the null.
   a) Almost always unknown.
   b) Power can be calculated for a few "plausible realistic values" or a range of such values.
   c) For a t-test , effect size can be measured by $\delta = \frac{\mu_1 - \mu_2}{\sigma}$. It can be shown that the power depends on this quantity. Meaning if means move further apart but noise increases keeping this ratio the same, the power remains the same.

Power and Sample Size

- **Power Calculation**: Given all 3 quantitites, i.e. n, α and δ the power can be easily calculated.

- **Sample Size Calculation:** Given the "target power" and any two of these, the third can be calculated. For example, given "target power", "level of significance" and "effect size", the "sample size required" can be determined.

- **Power/Sample size/Effect size:** Given any two, the third can be calculated using functions such as power.t.test(), power.prop.test() in R.

Multiple Testing and Power

- As mentioned before, multiple testing leads to loss of power as the per-test level of significance becomes more and more stringent.
- Thus, a candidate gene is always more powerful for the specific gene than an omics study. However, the 'power' of the candidate gene study to detect all other true genes is 0 (since those genes are unmeasured in the candidate study).
- If there are multiple truths among the genes tested, the overall chance of making a discovery, i.e., rejecting at least one test ('global' power) is higher than the per-test power.
- There is a tradeoff when testing many genes. Testing all or too many genes (indiscriminately) will add to the multiple testing penalty, reducing per-test power and hence also global power. However, if the gene set to be tested has a high-proportion of truths (example only cancer-related genes or only immune genes etc.), overall power may be gained in-spite of multiple testing.
- The caveat to choosing a smaller gene set based on knowledge is that if too much knowledge is used, many true genes may be thrown out; once again reducing power of discovery.

Principal Component Analysis

- Suppose we have multivariate data $X_{n\times p}$ with p variables observed on n individuals. Statistical methods such as regression analysis fail when $p \approx n$ or $p > n$. It is also difficult to visualize/summarize the data.

- In such cases, Dimension Reduction techniques (such as PCA) help us to reduce the data to $V_{n\times k}$ where $k \ll p$ (much less than p). After dimension reduction, we can work with a smaller set of variables $V_1, V_2, \ldots V_k$ instead of the original $X_1, X_2, \ldots X_p$, allowing easier visualization, analysis etc.

- In PCA , the new variables V_j are weighted linear combinations of the original variables. Hence, as such they lack interpretation (this is a drawback). However, often the top PC variables capture variability due to underlying hidden (latent) factors, such as disease status, gender, ancestry, batch effects etc. This happens in an 'unsupervised' manner without any knowledge of those factors. Hence if scatter plots of top PC-s show clusters, investigator needs to guess and confirm the sources of those variations.

- Construction: V_1, V_2, … are constructed so that
  - $V_1 = l_{11}X_1 + l_{12}X_2 + \cdots + l_{1p}X_p$ has the maximum variance among all linear combinations subject to $\sum_j l_{1j}^2 = 1$.
  - $V_2 = l_{21}X_1 + l_{22}X_2 + \cdots + l_{2p}X_p$ has the maximum variance among all linear combinations subject to $\sum_j l_{2j}^2 = 1$ and $Cor(V_1, V_2) = 0$ and so on,
  - $:$
  - $V_k = l_{k1}X_1 + l_{k2}X_2 + \cdots + l_{kp}X_p$ has the maximum variance among all linear combinations subject to $\sum_j l_{kj}^2 = 1$ and $Cor(V_j, V_k) = 0, for\ all\ j < k.$

Principal Component Analysis

- **PCA Algorithm:** Each column of X is centered by subtracting the mean (and optionally scaled by dividing by SD). The $p \times p$ covariance (or correlation matrix) $\frac{1}{n} X^T X$ is calculated. The Eigen Value Decomposition of this matrix gives "eigen values" (λ_j) denoting variance of each PC, and corresponding "eigen vectors" (l_j) representing loadings (weights) for that PC.

- The PCA algorithm returns the loading matrix L, columns of which contain the loadings l_j used to 'rotate' the original correlated X variables to make uncorrelated variables V. It also returns the new variables V $(= XL)$, the columns of which are the PC_1 scores, PC_2 scores, etc., for each individual.

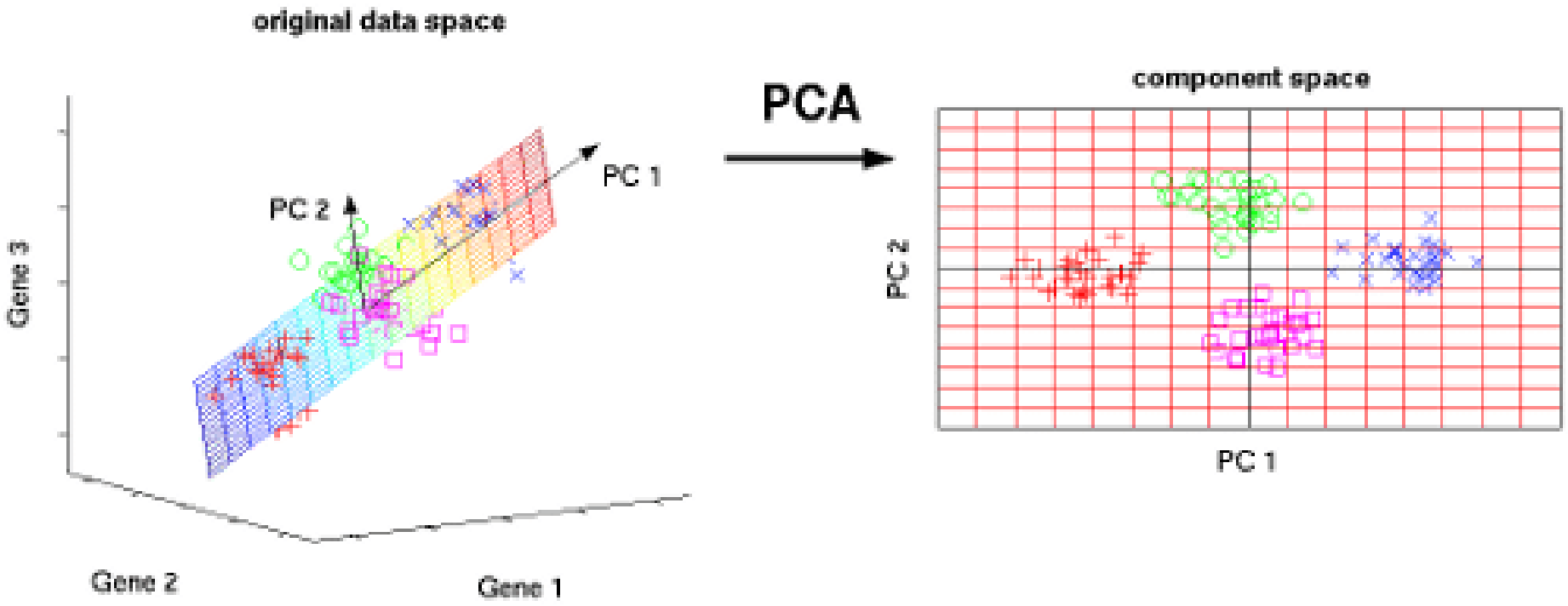
- The PC variances are in decreasing order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > 0$, $where\ \lambda_j = Var(V_j)$. The number of non-zero PC-s is $\min(n, p)$ or could be even lower if there are perfectly correlated variables in the data.

- A thumb rule to decide 'number of PC-s to retain' is based on cumulative proportion of variance explained. Example: $\frac{\lambda_1+\lambda_2}{\lambda_1+\lambda_2+\cdots+\lambda_k} < cutoff\ \ but\ \frac{\lambda_1+\lambda_2+\lambda_3}{\lambda_1+\lambda_2+\cdots+\lambda_k} \geq cutoff$, then we stop at 3 PC-s. A cutoff of 75% (0.75) or 80% may be used in some applications or even higher , e.g., 90% depending on the purpose.

- Another stopping rule is to visually check for random scatter among pairwise scatter plots of the PC-scores (V_j, V_{j+1}) . PC-s beyond this may have information for analysis but do not pick latent subgroups or major effects such as batch effects.

- PCA can be done using R functions prcomp() or princomp().

Principal Component Analysis



Hierarchical Clustering

- Cluster Analysis is an 'unsupervised classification' technique. Where data observations are grouped into classes based on the feature data alone without any prior data on group labels.

- **Hierarchical clustering** is a popular clustering algorithm. It can be 'agglomerative' or 'divisive'. Agglomerative hierarchical clustering starts from 0 clusters, clubs the two closest samples into a cluster and then gradually builds more and more clusters. Finally, all the samples/units are clustered together. Divisive proceeds in the reverse order starting with all units clustered together and then breaks into 2 clusters and so on until finally all units are in their own clusters.

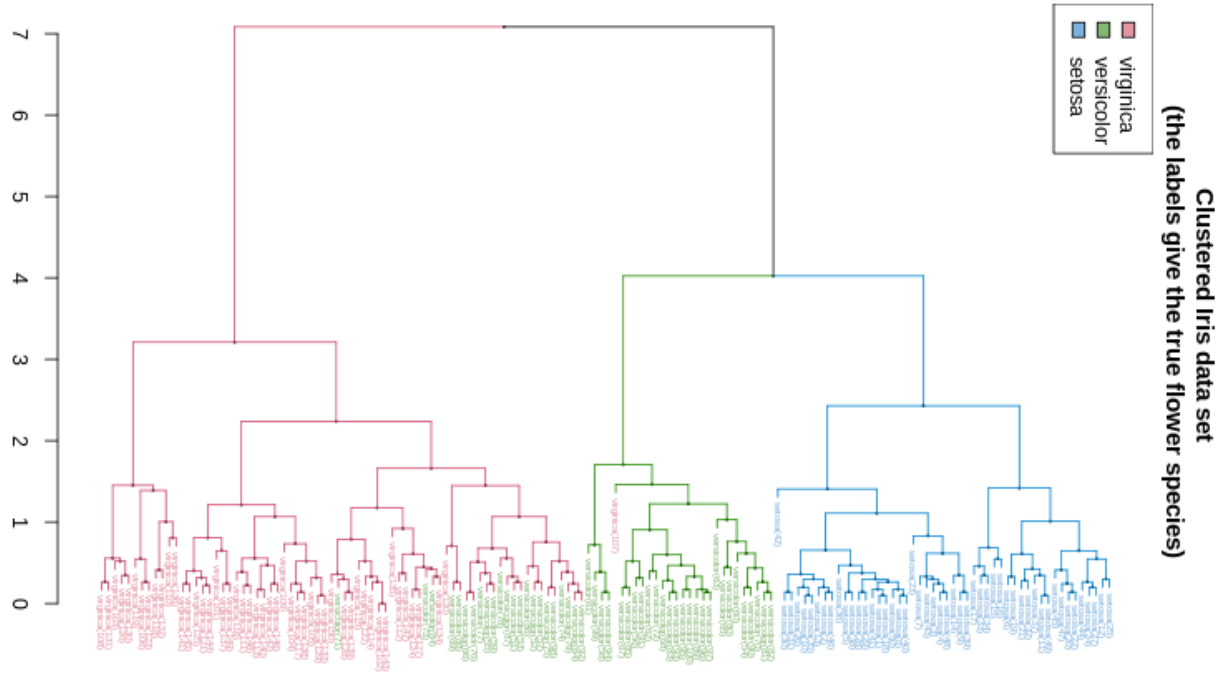
- **Dendrogram:** A plot showing the results of hierarchical clustering as a tree structure.

- At every step distance between two units (i, j) is measure by some appropriate distance metric such as Euclidean Distance in k-dimensional space $\mathrm{D_{ij}} = \sqrt{(X_{i1} - X_{j1})^2 + \cdots + (X_{ik} - X_{jk})^2}$

- Distance between two clusters A and B can be defined as single linkage (minimum pairwise distance $\min D_{ij}\ for\ i \in A, j \in B$, average linkage ($\overline{D_{ij}}$) or complete linkage ($\max D_{ij}$) .

- In R, the function hc=hclust(dist(X)) can be used to perform hierarchical clustering and plot(hc) to view a dendrogram. The function cutree(hc) can be used to cut the dendrogram at a certain height or a certain number of clusters and generate cluster labels accordingly.

Dendrogram of Iris Data (Hierarchical Clustering with Complete Linkage)



https://cran.r-project.org/web/packages/dendextend/vignettes/Cluster_Analysis.html#the-3-clusters-from-the-complete-method-vs-the-real-species-category

Supervised Learning (Classification/Prediction)

- **DataSet1 (Labelled Training Data):** Suppose, we are given a labelled dataset $(Y_i, X_{i1}, X_{i2}, \ldots, X_{ip})$ on i $= 1, \ldots, n$ individuals and 1 response variable (phenotypic outcome) and p features (e.g., genes).

- **DataSet2 (Unlabelled):** Further, we are given an unlabelled dataset $(X_{i1}, X_{i2}, \ldots, X_{ip})$ on m new individuals $i = n+1, \ldots . n+m$, for whom the response variable is unknown.

- The goal is to accurately **classify** the m individuals according to their most likely Y values (if Y is binary or categorical) or **predict** the most likely Y value, $\hat{Y}_i$.
- Generally, a model $Y = f(X) + Error$ is fitted to the n labelled data points to get an $\hat{f}$ (the trained/learned model), which is then used to predict the unlabelled observations as $\hat{Y}_i = \hat{f}(X_i), i = n+1, \ldots n+m$
  - f is chosen among a large class of functions by minimizing some measure of prediction error.
  - f may use only a few relevant features among many available features (a process known as feature selection)

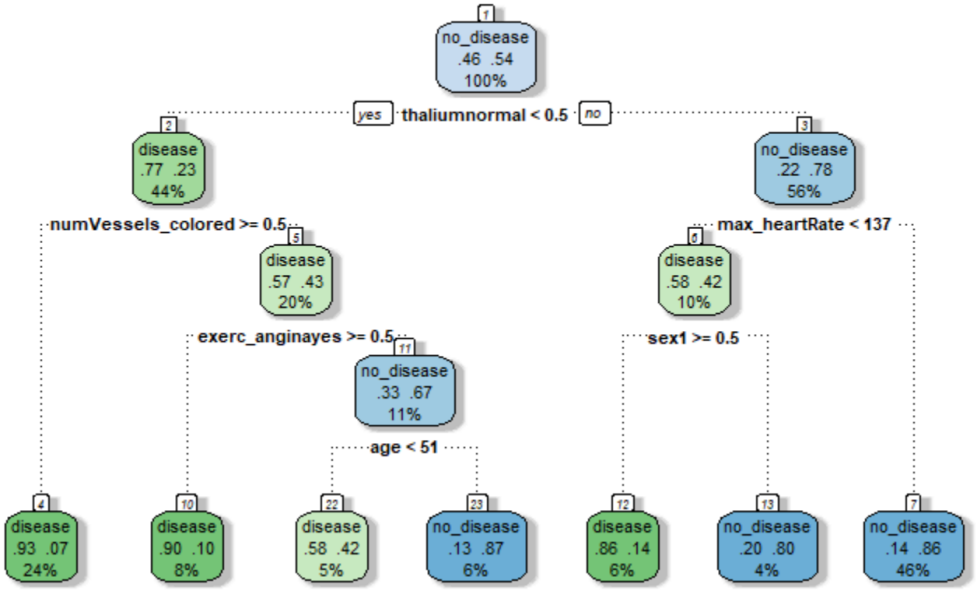
Common Supervised Learning Models

- The function f can be chosen among linear or non-linear models with different structures.
- Some commonly used ML models:
  - Linear Regression
  - Logistic Regression
  - Linear Discriminant Analysis
  - Support Vector Machines
  - Decision Tree (also known as CART)
  - Random Forest
  - Gradient Boosting
  - K-Nearest Neighbour (KNN)
  - Artificial Neural Networks

Decision Tree Model (CART: Classification and Regression Tree).

Classification: Y binary
Regression: Y continuous

Right: An example Classification Tree



- At the root node and every internal node, the observations in that node are split into two parts based on a feature. The feature to be used in the split and the cutoff-value are selected by minimizing an 'impurity' measure (e.g., Gini Index/Information Gain). Each leaf node (terminal node) gives a decision (yes/no).
- The full tree is grown and then branches are pruned back to prevent overfitting.
- Decision Trees give interpretable models. In many contexts interpretation may not be important. A black box model such as neural networks can give a better prediction error, which is what matters in many applications (e.g., predicting disease or prognosis). If biology is important, or if we are trying to develop a biomarker signature that uses only a few genes (that will be cost effective at large scale), interpretation of the model may be relevant.
- Even complex black box ML models can give some measure of variable importance for each of the original features, which gives some interpretation.

Model Complexity & Overfitting

- **Model Complexity:** Complexity of the model is high when many features are used, e.g., a complex linear $Y = \alpha + \sum_{j=1}^{10000} \beta_j X_j + error$ or when many parameters are are used for a few features. For example, even using a single feature, one can fit a polynomial of degree 100 : $Y = \alpha + \sum_{j=1}^{100} \beta_j X_1^j + error$

- A natural model to fit is the regression model (linear or logistic) depending on whether Y is continuous or binary. The linear regression model:
  - $Y = \alpha + \sum_{j=1}^{p} \beta_j X_j + e$
  - Fits a hyperplane to the data points on p -dimensional space.
  - Given two points in 2-dimensional space, a straight line exists that passes through both points.
  - Similarly, if the total number of variables in the linear model is $\geq n$, the hyperplane passes through all the points (i.e., it is a perfect fit in the training set).

- However, this may not be the best fitting line that would have been obtained if many more data points were available. Hence on independent external datasets, the model performs badly.

Overfitting and Cross-Validation

- The labelled data is generally split into two parts (A) **'training set'** for choosing the best predictive model and (B) **'testing set'** for assessing the predictive performance of the model.

- **Overfitting:** A complex model often overfits the training data. As model complexity increases, the training error decreases (ultimately becoming a perfect fit) but testing error starts increasing after some complexity.

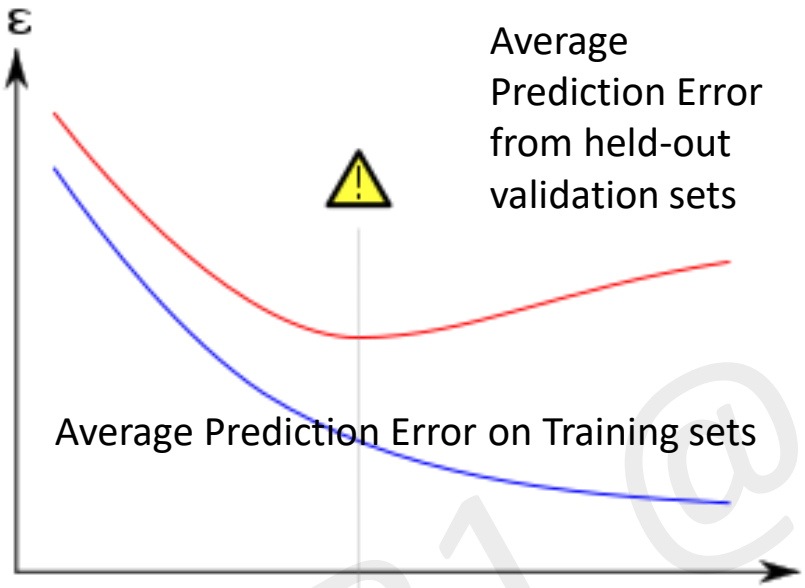
- **K-Fold Cross-Validation (k-fold CV)** is a method to prevent overfitting and choose a model which will have low prediction error on independent data (generalizable model). It works by partitioning the training data randomly into k equal parts. For any candidate value of a complexity parameter (example number of variables in regression or depth of a decision tree), the model is fit using $(k-1)$ parts (training set) and then the prediction error is estimated from the held-out k'th part (validation set). This is repeated for each of the k folds, and the average estimated prediction error (from these k error values) is plotted against the complexity parameter.

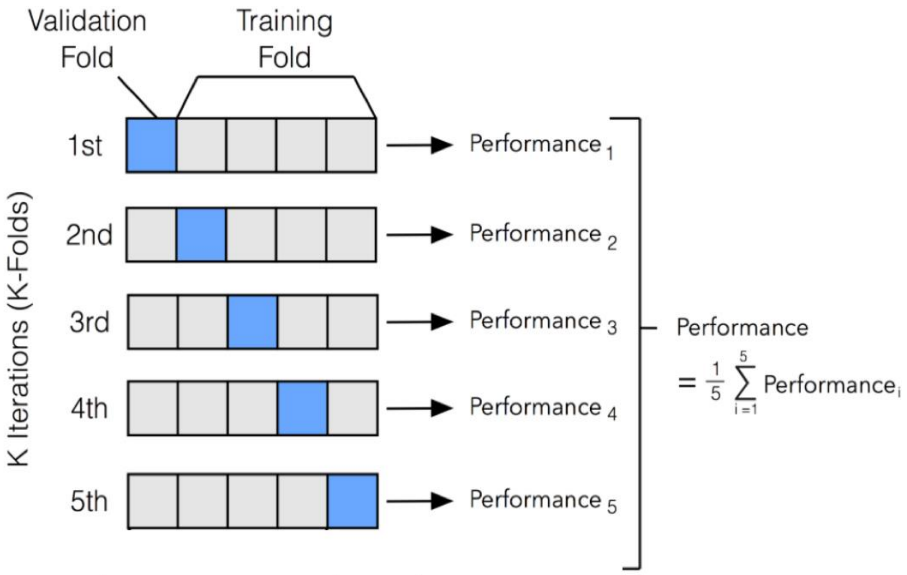
- The optimal point (model complexity value) for which the average CV prediction error is minimized, is chosen as the stopping point. At this point, the model may be refit on the whole training set (using the chosen optimal complexity).

- Finally, the held-out testing set, can give an estimate of the accuracy of trained model. It can then be applied on unlabelled data for prediction.

- Note that often a different convention/terminology is used (where the 'testing set' and 'validation set' terms are interchanged: (https://en.wikipedia.org/wiki/Training,_validation,_and_test_data_sets).

Overfitting and Cross-validation





References

- Many of the slides have been reproduced or modified from slides and other teaching material available on the internet.
  - **Contact me if you need reference for specific slides.**

- These slides are meant for your learning. Do not share indiscriminately or present publicly without appropriate citations.