Unit -II

Testing and Data Modeling

Sampling of continuous distributions

- 5.1: The Sampling Distribution of a Sample Mean
- 5.2: Sampling Distributions for Counts and Proportions

Sampling Distributions

5.1 The Sampling Distribution of a Sample Mean

Objectives

5.1 Sampling distribution of a sample mean

- The mean and standard deviation of $\bar{x}$
- For normally distributed populations

Reminder: the two types of data

- ***Quantitative***
  - Something that can be counted or measured and then averaged across individuals in the population (e.g., your height, your age, your IQ score)

- ***Qualitative***
  - Something that falls into one of several categories. What can be counted is the proportion of individuals in each category (e.g., your gender, your hair color, your blood type—A, B, AB, O).

How do you figure it out? Ask:

- What are the *n* individuals/units in the sample (of size "*n*")?
- What is being recorded about those *n* individuals/units?
- Is that a number (quantitative) or a statement (categorical)?

Reminder: What is a sampling distribution?

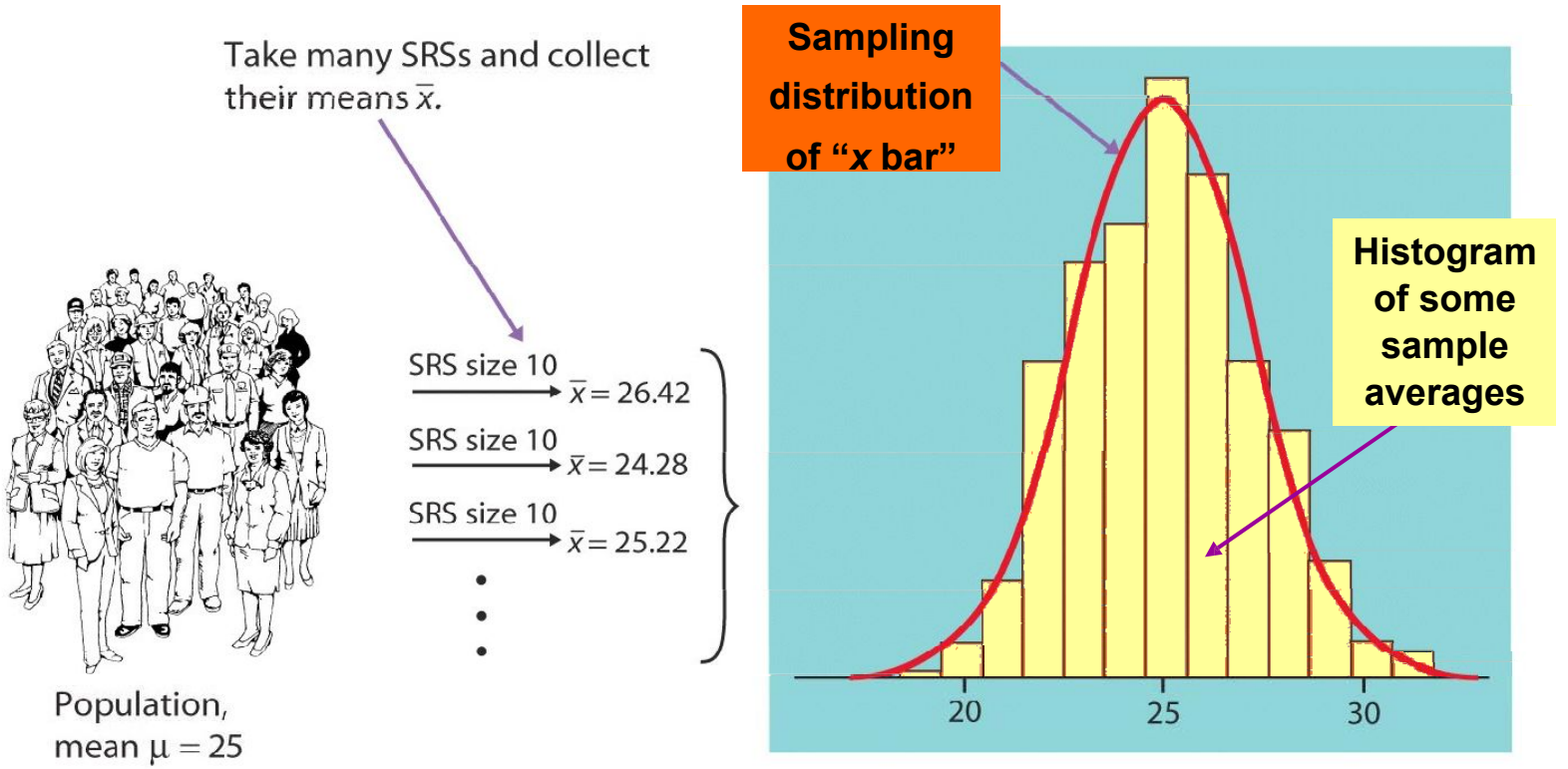
The **sampling distribution of a statistic** is the distribution of all possible values taken by the statistic when all possible samples of a fixed size *n* are taken from the population. It is a theoretical idea—we do not actually build it.

The sampling distribution of a statistic is the **probability distribution** of that statistic.

Sampling distribution of the sample mean

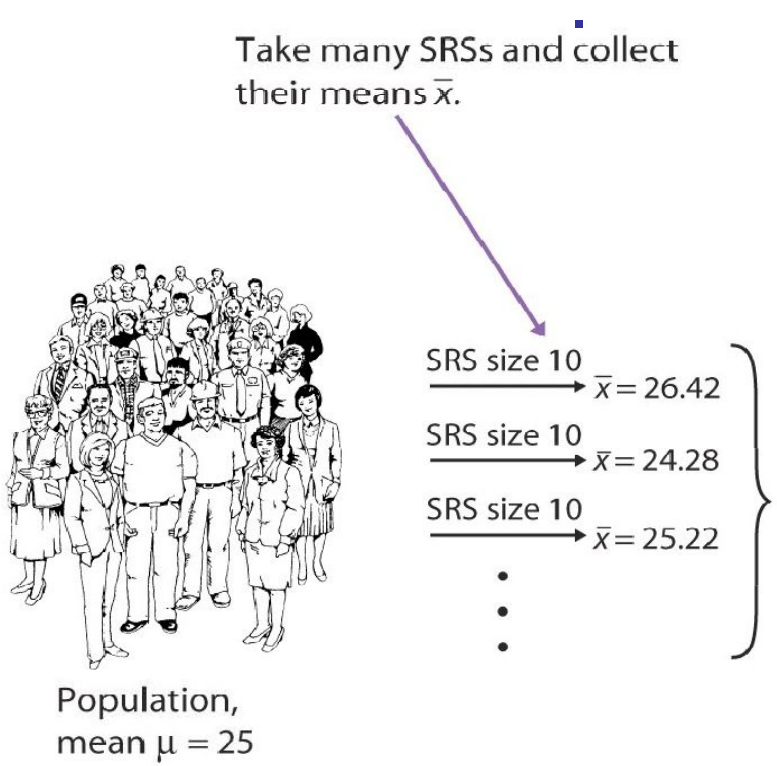
We take many random samples of a given size *n* from a population with mean μ and standard deviation σ.

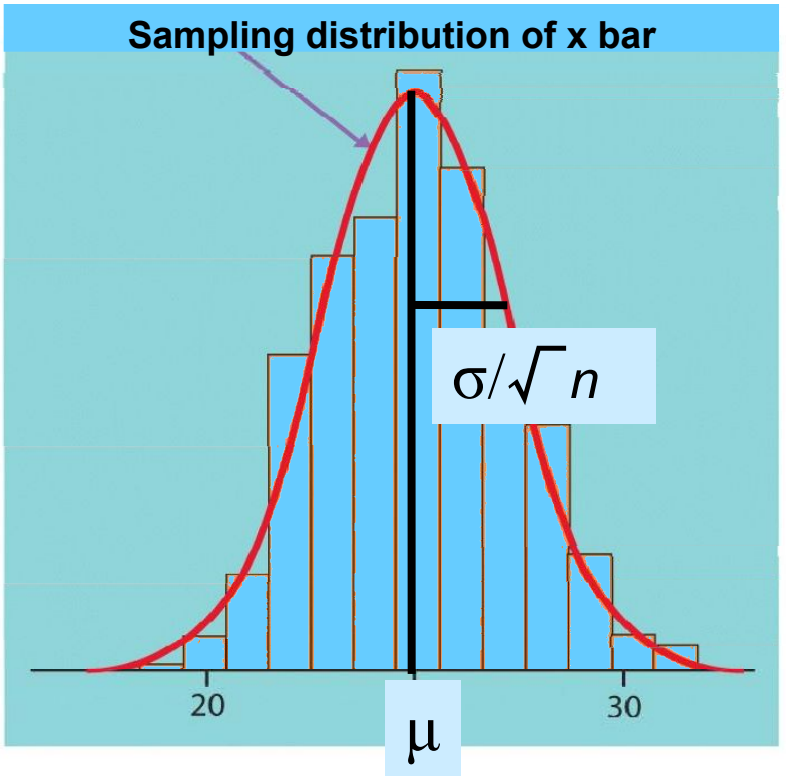
Some sample means will be above the population mean μ and some will be below, making up the sampling distribution.



For any population with mean μ and standard deviation σ:

- The **mean**, or center of the sampling distribution of $\bar{x}$, is equal to the population mean μ: $\mu_{\bar{x}} = \mu$.
- The **standard deviation** of the sampling distribution is $\sigma/\sqrt{n}$, where n is the sample size: $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.





- Mean of a sampling distribution of $\bar{x}$

  There is no tendency for a sample mean to fall systematically above or below μ, even if the distribution of the raw data is skewed. Thus, the mean of the sampling distribution is an **unbiased estimate** of the population mean μ — it will be "correct on average" in many samples.

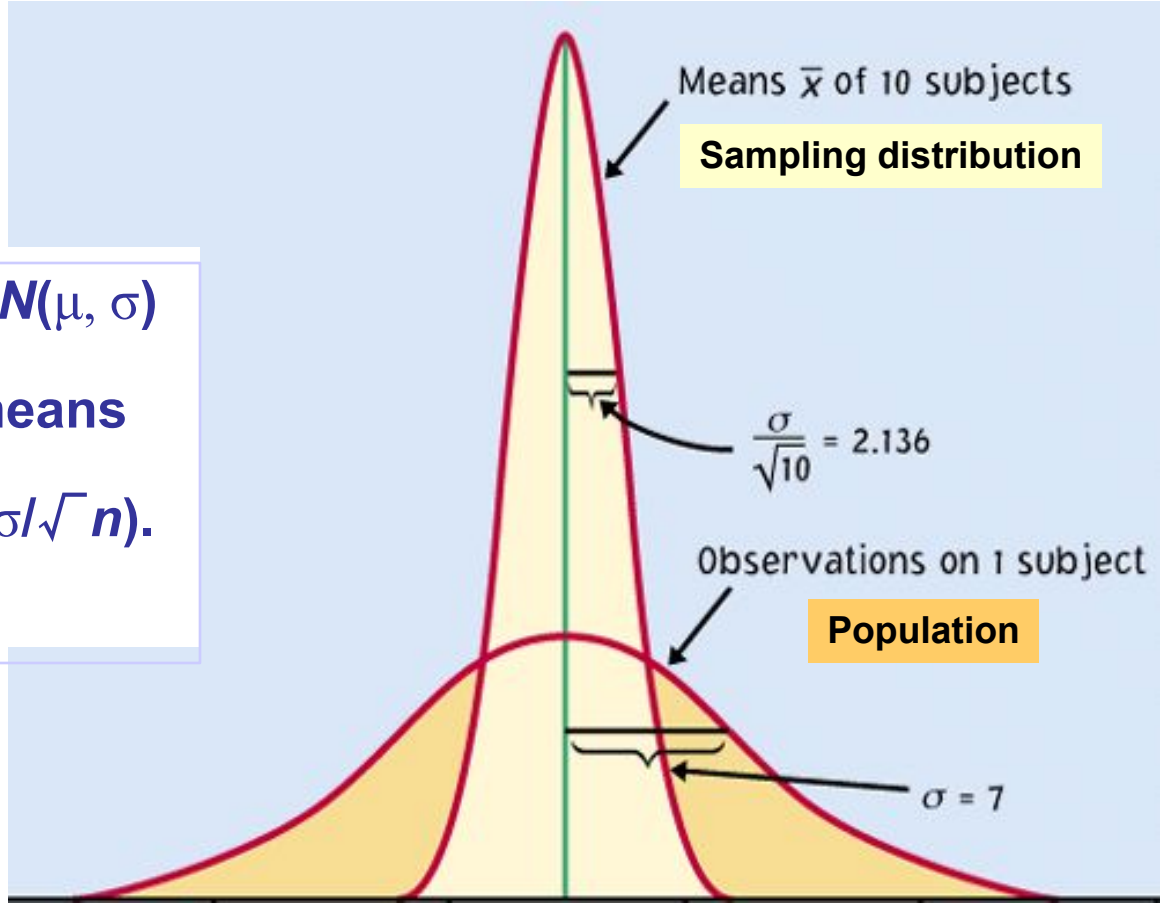
- Standard deviation of a sampling distribution of $\bar{x}$

  The standard deviation of the sampling distribution measures how much the sample statistic varies from sample to sample. It is smaller than the standard deviation of the population by a factor of $\sqrt{n}$. **Averages are less variable than individual observations**.

For normally distributed populations

When a variable in a population is normally distributed, the sampling distribution of $\bar{x}$ for all possible samples of size *n* is also normally distributed.

If the population is *N*(μ, σ) then the sample means distribution is *N*(μ, σ/√*n*).



population vs. sample

In a large population of adults, the mean IQ is 112 with standard deviation 20. Suppose 200 adults are randomly selected for a market research campaign.

- The distribution of the sample mean IQ is:

A) Exactly normal, mean 112, standard deviation 20

B) Approximately normal, mean 112, standard deviation 20

C) Approximately normal, mean 112 , standard deviation 1.414

D) Approximately normal, mean 112, standard deviation 0.1

C) Approximately normal, mean 112 , standard deviation 1.414

Population distribution : $N(\mu = 112; \sigma = 20)$

Sampling distribution for $n = 200$ is $N(\mu = 112; \sigma/\sqrt{n} = 1.414)$

Application

Hypokalemia is diagnosed when blood potassium levels are below 3.5mEq/dl. Let's assume that we know a patient whose measured potassium levels vary daily according to a normal distribution $N(\mu = 3.8, \sigma = 0.2)$.

If only one measurement is made, what is the probability that this patient will be misdiagnosed with Hypokalemia?

$$z = \frac{(x - \mu)}{\sigma} = \frac{3.5 - 3.8}{0.2} = -1.5, \quad P(z < -1.5) = 0.0668 \approx 7\%$$

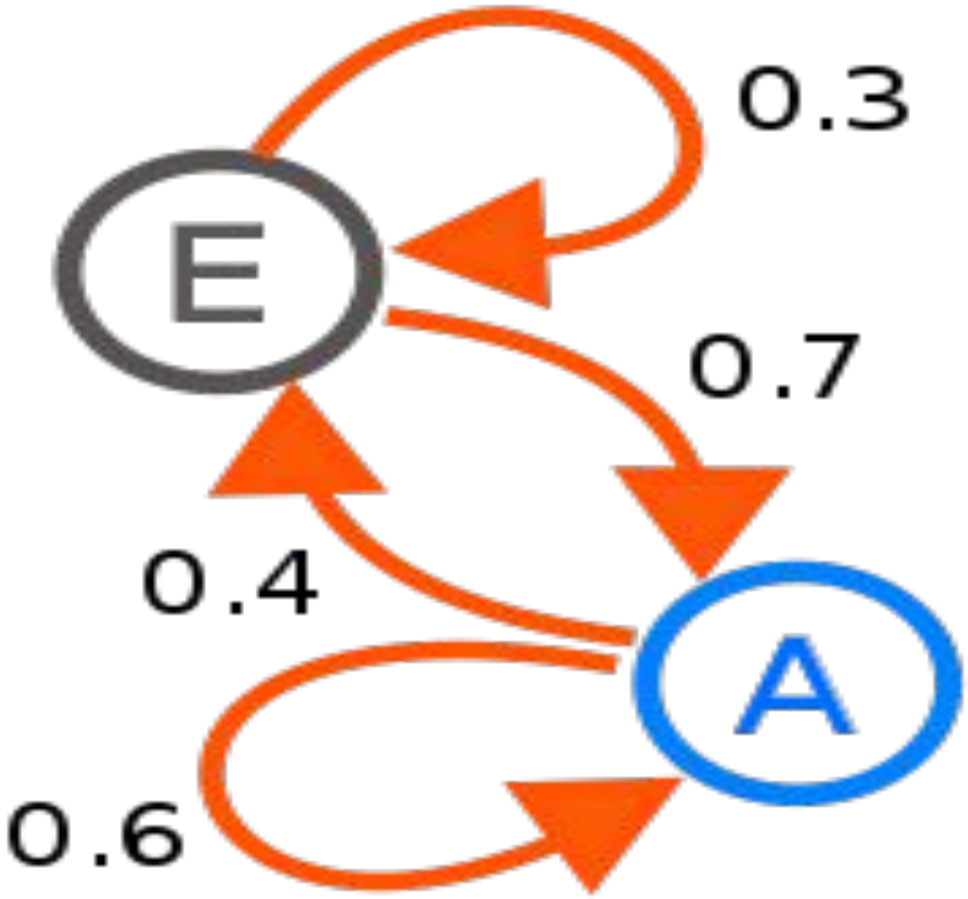
Instead, if measurements are taken on 4 separate days, what is the probability of a misdiagnosis?

$$z = \frac{(\bar{x} - \mu)}{\sigma/\sqrt{n}} = \frac{3.5 - 3.8}{0.2/\sqrt{4}} = -3, \quad P(z < -3) = 0.0013 \approx 0.1\%$$

Note: Make sure to standardize (z) using the standard deviation for the sampling distribution.

Monte Carlo method

- **Monte Carlo methods**, or **Monte Carlo experiments**, are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results.
- The underlying concept is to use randomness to solve problems that might be deterministic in principle.
- They are often used in physical and mathematical problems and are most useful when it is difficult or impossible to use other approaches.
- Monte Carlo methods are mainly used in three problem classes: optimization, numerical integration, and generating draws from a probability distribution.
- A Markov process is a stochastic process that satisfies the Markov property (sometimes characterized as "memorylessness"). In simpler terms, it is a process for which predictions can be made regarding future outcomes based solely on its present state and—most importantly—such predictions are just as good as the ones that could be made knowing the process's full history.
- In other words, conditional on the present state of the system, its future and past states are independent.



A diagram representing a two-state Markov process. The numbers are the probability of changing from one state to another state.

- In principle, Monte Carlo methods can be used to solve any problem having a probabilistic interpretation.
- By the law of large numbers, integrals described by the expected value of some random variable can be approximated by taking the empirical mean (a.k.a. the 'sample mean') of independent samples of the variable.
- When the probability distribution of the variable is parameterized, mathematicians often use a Markov chain Monte Carlo (MCMC) sampler.
- Other examples include modeling phenomena with significant uncertainty in inputs such as the calculation of risk in business and, in mathematics, evaluation of multidimensional definite integrals with complicated boundary conditions.
- In application to systems engineering problems (space, oil exploration, aircraft design, etc.), Monte Carlo–based predictions of failure, cost overruns and schedule overruns are routinely better than human intuition or alternative "soft" method.

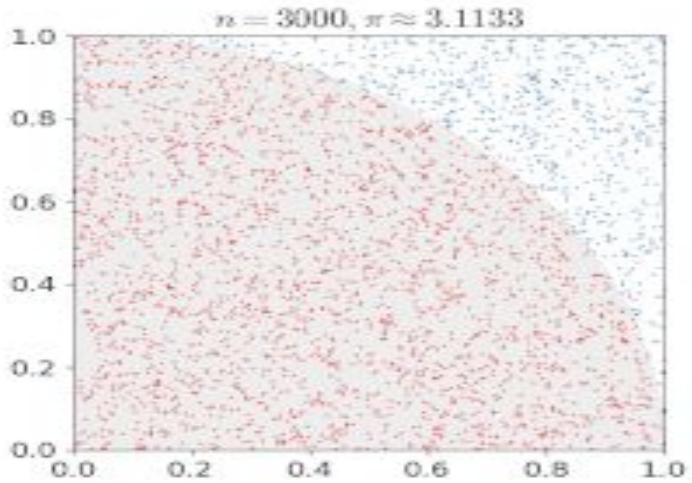
Monte Carlo methods vary, but tend to follow a particular pattern:

- Define a domain of possible inputs
- Generate inputs randomly from a probability distribution over the domain
- Perform a deterministic computation on the inputs
- Aggregate the results

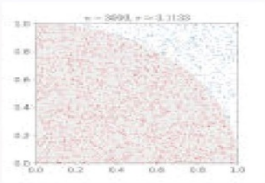
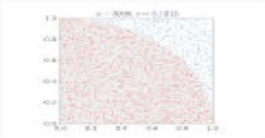
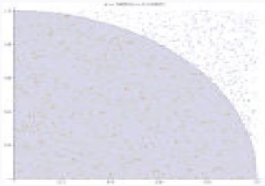
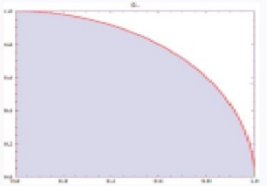
Example

- For example, consider a quadrant (circular sector) inscribed in a unit square. Given that the ratio of their areas is $\pi/4$, the value of π can be approximated using a Monte Carlo method:

1. Draw a square, then inscribe a quadrant within it
2. Uniformly scatter a given number of points over the square
3. Count the number of points inside the quadrant, i.e. having a distance from the origin of less than 1
4. The ratio of the inside-count and the total-sample-count is an estimate of the ratio of the two areas, $\pi/4$. Multiply the result by 4 to estimate π.



- Click on a date/time to view the file as it appeared at that time.

	Date/Time	Thumbnail	Dimensions	User	Comment
current	**16:00, 16 February 2017**		500 × 500 (476 KB)	Nicoguaro	Make the plot square and increase gif delay.
	15:38, 16 February 2017		640 × 480 (476 KB)	Nicoguaro	Bigger text in the axes, and colors from ColorBrewer. Code in Python.
	18:29, 7 November 2011		500 × 500 (373 KB)	Rayhem	Slowed animation to avoid looking like a blinky page element, improved resolution, added counter for number of points, shaded points inside/outside the circle. ==Mathematica 7.0 Source== <pre> tinyColor[color_, point_] := {PointSize[Small], color, Point[
	23:12, 14 March 2011		360 × 369 (363 KB)	CaitlinJo	{{Information |Description ={{en|1=As points are randomly scattered inside the unit square, some fall within the unit circle. The fraction of points inside the circle over all points approaches pi as the number of points goes toward infinity. This ani

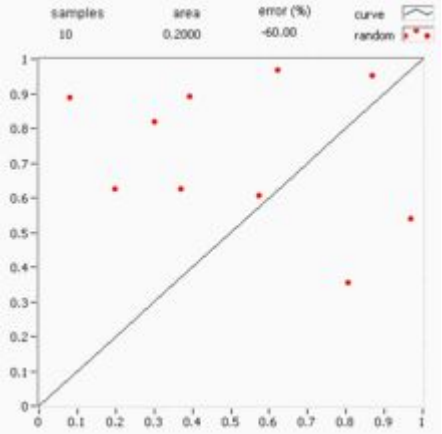
Applications

- **Physical science**
- **Engineering**
- **Climate change and radiative forcing**
- **Computational biology**
- **Computer graphics**
- **Applied statistics**
- **Artificial intelligence for games**
- **Design and visuals**
- **Search and rescue**
- **Finance and business**
- **Library science**

Use in mathematics

- **Integration:**

Deterministic numerical integration algorithms work well in a small number of dimensions, but encounter two problems when the functions have many variables. First, the number of function evaluations needed increases rapidly with the number of dimensions.



- **Simulation and optimization**
- **Inverse problem**

What is Hypothesis Testing

- Hypothesis testing is a statistical method that is used in making a statistical decision using experimental data. Hypothesis testing is basically an assumption that we make about a population parameter.
- It evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data.
- **Example:** You say an average student in the class is 30 or a boy is taller than a girl. All of these is an assumption that we are assuming and we need some statistical way to prove these. We need some mathematical conclusion whatever we are assuming is true.

Need for Hypothesis Testing

- Hypothesis testing is an important procedure in statistics. Hypothesis testing evaluates two mutually exclusive population statements to determine which statement is most supported by sample data. When we say that the findings are statistically significant, it is thanks to hypothesis testing.

Parameters of hypothesis testing

- **Null hypothesis(H0):** In statistics, the null hypothesis is a general given statement or default position that there is no relationship between two measured cases or no relationship among groups. In other words, it is a basic assumption or made based on the problem knowledge.

  Example: A company production is = 50 units/per day etc.
- **Alternative hypothesis(H1):** The alternative hypothesis is the hypothesis used in hypothesis testing that is contrary to the null hypothesis.

  Example: A company's production is not equal to 50 units/per day etc.
- **Level of significance** It refers to the degree of significance in which we accept or reject the null hypothesis. 100% accuracy is not possible for accepting a hypothesis, so we, therefore, select a level of significance that is usually 5%. This is normally denoted with and generally, it is 0.05 or 5%, which means your output should be 95% confident to give a similar kind of result in each sample.
- **P-value** The P value, or calculated probability, is the probability of finding the observed/extreme results when the null hypothesis(H0) of a study-given problem is true. If your P-value is less than the chosen significance level then you reject the null hypothesis i.e. accept that your sample claims to support the alternative hypothesis.

Steps in Hypothesis Testing

- **Step 1**– We first identify the problem about which we want to make an assumption keeping in mind that our assumption should be contradictory to one another
- **Step 2** – We consider statical assumption such that the data is normal or not, statical independence between the data.
- **Step 3** – We decide our test data on which we will check our hypothesis
- **Step 4** – The data for the tests are evaluated in this step we look for various scores in this step like z-score and mean values.
- **Step 5 –** In this stage, we decide where we should accept the null hypothesis or reject the null hypothesis

Types of Hypothesis Testing

- Z Test
- To determine whether a discovery or relationship is statistically significant, hypothesis testing uses a z-test.
- It usually checks to see if two means are the same (the null hypothesis). Only when the population standard deviation is known and the sample size is 30 data points or more, can a z-test be applied.
- T Test
- A statistical test called a t-test is employed to compare the means of two groups. To determine whether two groups differ or if a procedure or treatment affects the population of interest, it is frequently used in hypothesis testing.
- Chi-Square
- You utilize a Chi-square test for hypothesis testing concerning whether your data is as predicted.
- To determine if the expected and observed results are well-fitted, the Chi-square test analyzes the differences between categorical variables from a random sample. The test's fundamental premise is that the observed values in your data should be compared to the predicted values that would be present if the null hypothesis were true.

Example:

- Given a coin and it is not known whether that is fair or tricky so let's decide the null and alternate hypothesis
- Null Hypothesis(H0): a coin is a fair coin.
- Alternative Hypothesis(H1): a coin is a tricky coin.
- Toss a coin 1st time and assume that the result is head- P-value = (as head and tail have equal probability)
- Toss a coin 2nd time and assume that result again is head, now p-value

Formula For Hypothesis Testing

- To validate our hypothesis about a population parameter we use statistical functions. we use the z-score, p-value, and, level of significance(alpha) to make evidence for our hypothesis.

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

where,

is the sample mean,
μ represents the population mean,
σ is the standard deviation and
n is the size of the sample

Type I Error

- A type I error appears when the null hypothesis (H_0) of an experiment is true, but still, it is rejected.
- It is stating something which is not present or a false hit. A type I error is often called a false positive (an event that shows that a given condition is present when it is absent).
- In words of community tales, a person may see the bear when there is none (raising a false alarm) where the null hypothesis (H_0) contains the statement: “There is no bear”.
- The type I error significance level or rate level is the probability of refusing the null hypothesis given that it is true.
- It is represented by Greek letter α (alpha) and is also known as alpha level.
- Usually, the significance level or the probability of type i error is set to 0.05 (5%), assuming that it is satisfactory to have a 5% probability of inaccurately rejecting the null hypothesis.

Type II Error

- A type II error appears when the null hypothesis is false but mistakenly fails to be refused.
- It is losing to state what is present and a miss. A type II error is also known as false negative (where a real hit was rejected by the test and is observed as a miss), in an experiment checking for a condition with a final outcome of true or false.
- A type II error is assigned when a true alternative hypothesis is not acknowledged.
- In other words, an examiner may miss discovering the bear when in fact a bear is present (hence fails in raising the alarm).
- Again, H0, the null hypothesis, consists of the statement that, “There is no bear”, wherein, if a wolf is indeed present, is a type II error on the part of the investigator.
- Here, the bear either exists or does not exist within given circumstances, the question arises here is if it is correctly identified or not, either missing detecting it when it is present, or identifying it when it is not present.
- The rate level of the type II error is represented by the Greek letter β (beta) and linked to the power of a test (which equals 1−β).

Table of Type I and Type II Error

- The relationship between truth or false of the null hypothesis and outcomes or result of the test is given in the tabular form:

Error Types	**When H_0 is True**	**When H_0 is False**
Don't Reject	Correct Decision (True negative) Probability = $1 - \alpha$	Type II Error (False negative) Probability = β
Reject	Type II Error (False Positive) Probability = α	Correct Decision (True Positive) Probability = $1 - \beta$

Type I and Type II Errors Example

- **Example 1**: Let us consider a null hypothesis – A man is not guilty of a crime.
- Then in this case:

Type I error (False Positive)	Type II error (False Negative)
He is condemned to crime, though he is not guilty or committed the crime.	He is condemned not guilty when the court actually does commit the crime by letting the guilty one go free.

- **Example 2:** Null hypothesis- A patient's signs after treatment A, are the same from a placebo.

Type I error (False Positive)	Type II error (False Negative)
Treatment A is more efficient than the placebo	Treatment A is more powerful than placebo even though it truly is more efficient.

Example of a Type II Error

- Assume a biotechnology company wants to compare how effective two of its drugs are for treating diabetes. The null hypothesis states the two medications are equally effective. A null hypothesis, H_0, is the claim that the company hopes to reject using the one-tailed test. The alternative hypothesis, H_a, states the two drugs are not equally effective. The alternative hypothesis, H_a, is the state of nature that is supported by rejecting the null hypothesis.
- The biotech company implements a large clinical trial of 3,000 patients with diabetes to compare the treatments. The company randomly divides the 3,000 patients into two equally sized groups, giving one group one of the treatments and the other group the other treatment. It selects a significance level of 0.05, which indicates it is willing to accept a 5% chance it may reject the null hypothesis when it is true or a 5% chance of committing a type I error.
- Assume the beta is calculated to be 0.025, or 2.5%. Therefore, the probability of committing a type II error is 97.5%. If the two medications are not equal, the null hypothesis should be rejected. However, if the biotech company does not reject the null hypothesis when the drugs are not equally effective, a type II error occurs.

Z-TEST

- In a z-test, we assume the sample is normally distributed. A z-score is calculated with population parameters such as population mean and population standard deviation.
- We use this test to validate a hypothesis that states the sample belongs to the same population.

•**Null**: Sample mean is same as the population mean.

•**Alternate**: Sample mean is not same as the population mean.

The statistic used for this hypothesis testing is called z-statistic,
the score for which we calculate as:

$z = (x — \mu) / (\sigma / \sqrt{n})$, where

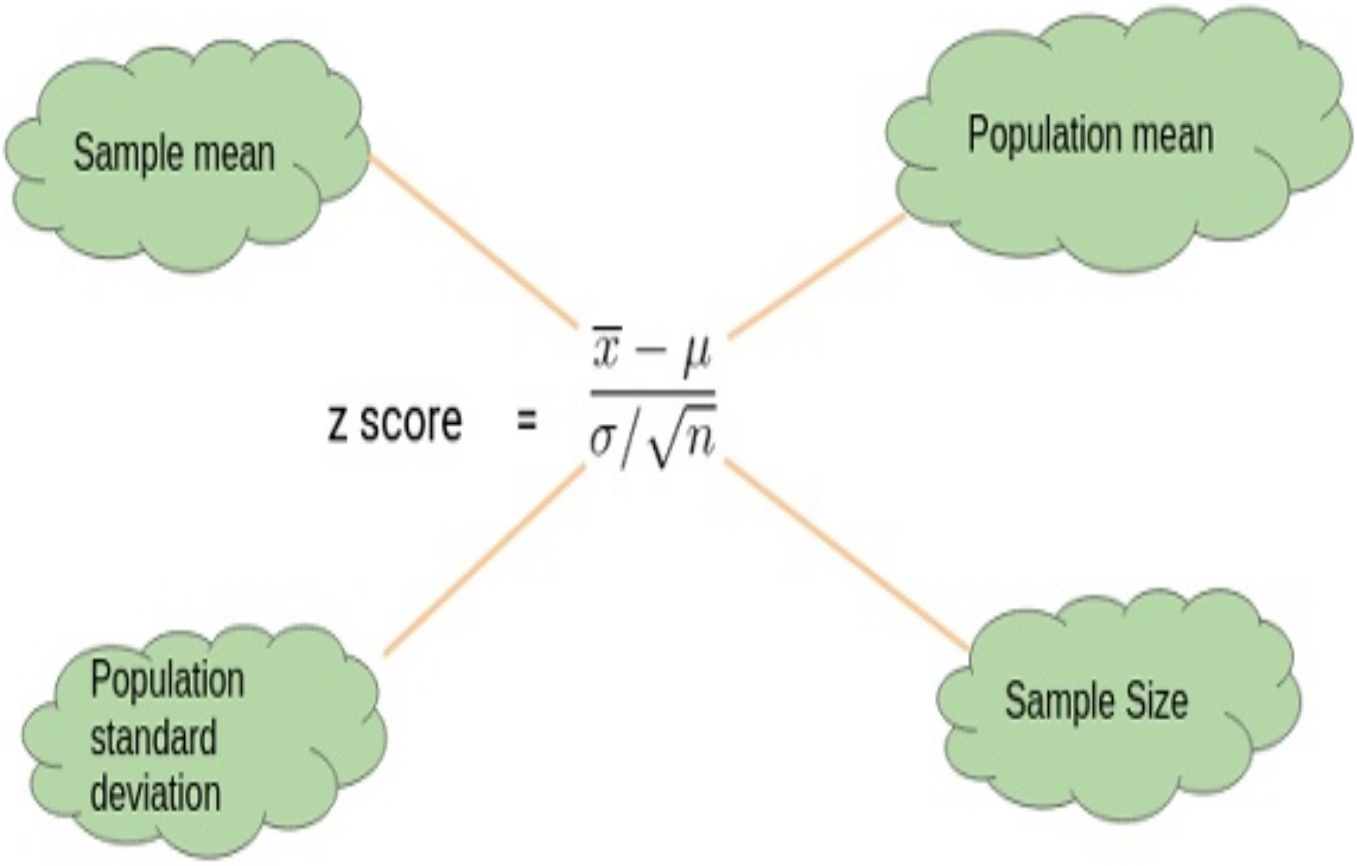
x=sample mean

μ=population mean

$\sigma / \sqrt{n}$= population standard deviation

If the test statistic is lower than the critical value, accept the hypothesis.

- z tests are a statistical way of testing a Null Hypothesis when either:
- We know the population variance, or
- We do not know the population variance, but our sample size is large $n \geq 30$
- If we have a sample size of less than 30 and do not know the population variance, we must use a t-test.
- This is how we judge when to use the z-test vs the t-test. Further, it is assumed that the z-statistic follows a standard normal distribution. In contrast, the t-statistics follows the t-distribution with a degree of freedom equal to n-1, where n is the sample size.
- It must be noted that the samples used for z-test or t-test must be independent sample, and also must have a distribution identical to the population distribution.
- This makes sure that the sample is not “biased” to/against the Null Hypothesis which we want to validate/invalidate.


Sample mean
Population mean
z score = $\frac{\overline{x} - \mu}{\sigma / \sqrt{n}}$
Population standard deviation
Sample Size

Here's an Example to Understand a One Sample z-Test

- Let's say we need to determine if girls on average score higher than 600 in the exam.
- We have the information that the standard deviation for girls' scores is 100. So, we collect the data of 20 girls by using random samples and record their marks.
- Finally, we also set our α value (significance level) to be 0.05.

In this example:

- Mean Score for Girls is 641
- The number of data points in the sample is 20
- The population mean is 600
- Standard Deviation for Population is 100

Score
650
730
510
670
480
800
690
530
590
620
710
670
640
780
650
490
800
600
510
700

$$\text{z score} = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$

$$= \frac{641 - 600}{100/\sqrt{20}}$$

$$= 1.8336$$

p value = .033357.

Critical Value = 1.645

Z score > Critical Value

P value < 0.05

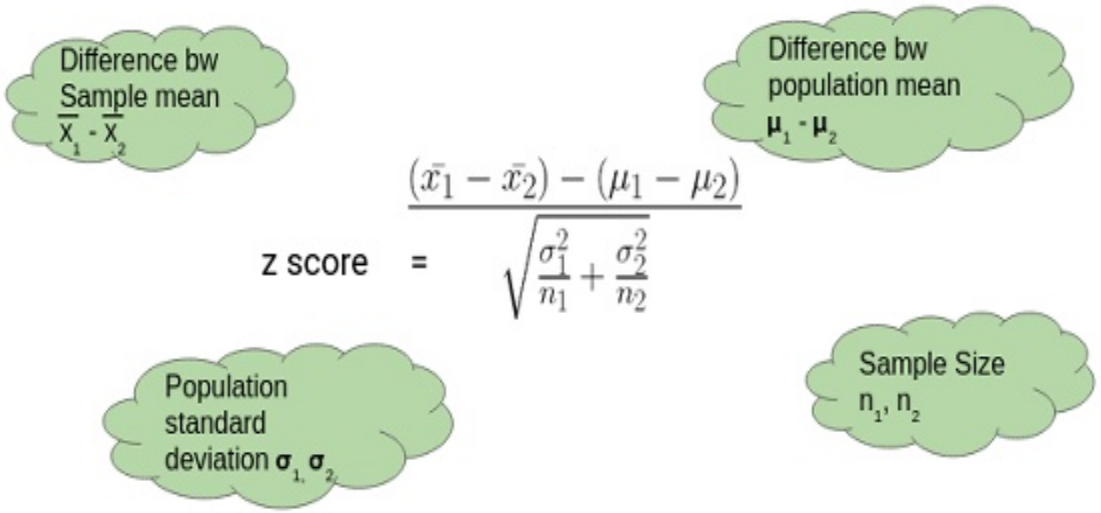
$H_0 : \mu \leq 600$

$H_1 : \mu > 600$

Since the P-value is less than 0.05, we can reject the null hypothesis and conclude based on our result that Girls on average scored higher than 600.

Two-Sample Z-Test

- We perform a Two Sample z-test when we want to compare **the mean of two samples.**



Here's an Example to Understand a Two Sample Z-Test

- Here, let's say we want to know if Girls on an average score 10 marks more than the boys. We have the information that the standard deviation for girls' Score is 100 and for boys' score is 90. Then we collect the data of 20 girls and 20 boys by using random samples and record their marks. Finally, we also set our α value (significance level) to be 0.05.

In this example:

- Mean Score for Girls (Sample Mean) is 641
- Mean Score for Boys (Sample Mean) is 613.3
- Standard Deviation for the Population of Girls' is 100
- Standard deviation for the Population of Boys' is 90
- Sample Size is 20 for both Girls and Boys
- Difference between Mean of Population is 10

Score
650
730
510
670
480
800
690
530
590
620
710
670
640
780
650
490
800
600
510
700

Score
630
720
462
631
440
783
673
519
543
579
677
649
632
768
615
463
781
563
488
650

$$\text{z score} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$= \frac{(641 - 613.3) - (10)}{\sqrt{\frac{100^2}{20} + \frac{90^2}{20}}}$$

$$= 0.588$$

P value = 0.278

Critical Value = 1.645

Z score < Critical Value

P value > 0.05

$H_0 : \mu_1 - \mu_2 \leq 10$

$H_1 : \mu_1 - \mu_2 > 10$

Thus, we can **conclude based on the p-value that we fail to reject the Null Hypothesis**. We don't have enough evidence to conclude that girls on average score of 10 marks more than the boys. Pretty simple, right?

T-TEST

- We use a t-test to compare the mean of two given samples.
- Like a z-test, a t-test also assumes a normal distribution of the sample.
- When we don't know the population parameters (mean and standard deviation), we use t-test.

THE THREE VERSIONS OF A T-TEST

1. Independent sample t-test: compares mean for two groups
2. Paired sample t-test: compares means from the same group at different times
3. One sample t-test: tests the mean of a single group against a known mean

The statistic for this hypothesis testing is called t-statistic, the score for which we calculate as:

$t=(x1—x2) / (\sigma / \sqrt{n1} + \sigma / \sqrt{n2})$, where

$x1$=mean of sample 1

$x2$=mean of sample 2

$n1$=sample size 1

$n2$=sample size 2

There are multiple variations of the t-test.

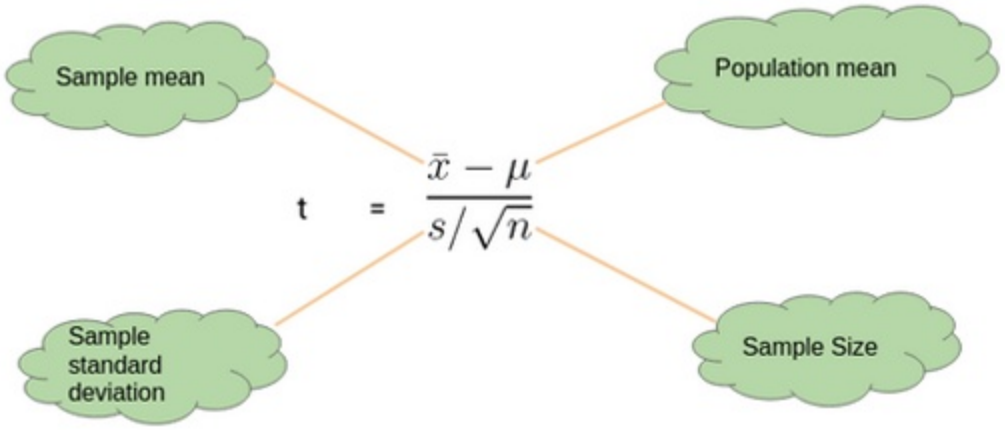
T-tests are a statistical way of testing a hypothesis when:

- We do not know the population variance
- Our sample size is small, $n < 30$

Examples of T Test

One-Sample T-Test

- We perform a One-Sample t-test when we want to **compare a sample mean with the population mean**. The difference from the z-Test is that we do **not have the information on Population Variance** here. We use the **sample standard deviation** instead of population standard deviation in this case.



Here's an Example to Understand a One Sample T-Test

- Let's say we want to determine if on average girls score more than 600 in the exam. We do not have the information related to variance (or standard deviation) for girls' scores. To a perform t-test, we randomly collect the data of 10 girls with their marks and choose our α value (significance level) to be 0.05 for Hypothesis Testing.

In this example:

- Mean Score for Girls is 606.8
- The size of the sample is 10
- The population mean is 600
- Standard Deviation for the sample is 13.14

Girls_Score
587
602
627
610
619
622
605
608
596
592

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

$$= \frac{606.8 - 600}{13.14/\sqrt{10}}$$

$$= 1.64$$

$H_0 : \mu \leq 600$

$H_1 : \mu > 600$

Critical Value = 1.833

t score < Critical Value

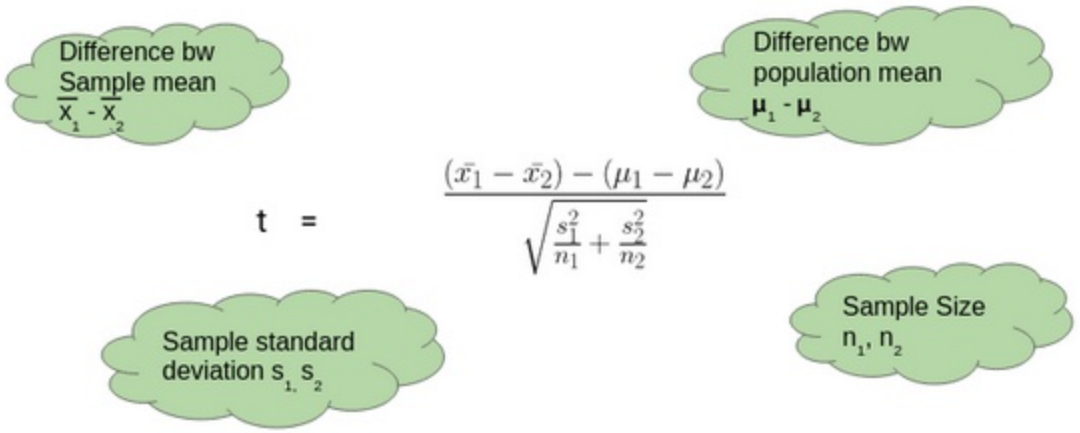
P value = 0.0678

P value > 0.05

Our **p-value is greater than 0.05 thus we fail to reject the null hypothesis** and don't have enough evidence to support the hypothesis that on average, girls score more than 600 in the exam.

Two-Sample T-Test

- We perform a Two-Sample t-test when we want to compare the mean of two samples.



Here's an Example to Understand a Two-Sample T-Test

- Here, let's say we want to determine if on average, boys score 15 marks more than girls in the exam.
- We do not have the information related to variance (or standard deviation) for girls' scores or boys' scores.
- To perform a t-test. we randomly collect the data of 10 girls and boys with their marks.
- We choose our α value (significance level) to be 0.05 as the criteria for Hypothesis Testing.

In this example:

- Mean Score for Boys is 630.1
- Mean Score for Girls is 606.8
- Difference between Population Mean 15
- Standard Deviation for Boys' score is 13.42
- Standard Deviation for Girls' score is 13.14

Girls_Score
587
602
627
610
619
622
605
608
596
592

Boys_Score
626
643
647
634
630
649
625
623
617
607

$$t = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$\frac{(630.1 - 606.8) - (15)}{\sqrt{\frac{(13.42)^2}{10} + \frac{(13.14)^2}{10}}}$$

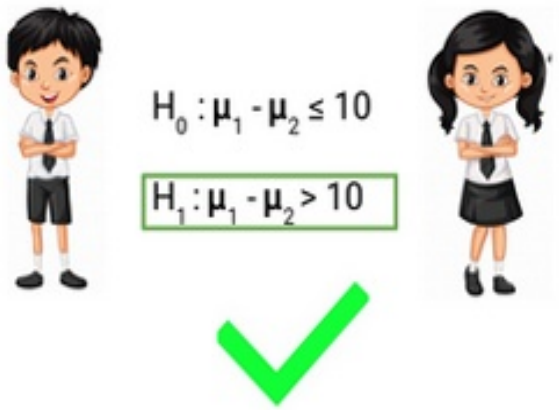
Critical Value = 1.833

t = 2.23

P value = 0.019

Critical Value > t score

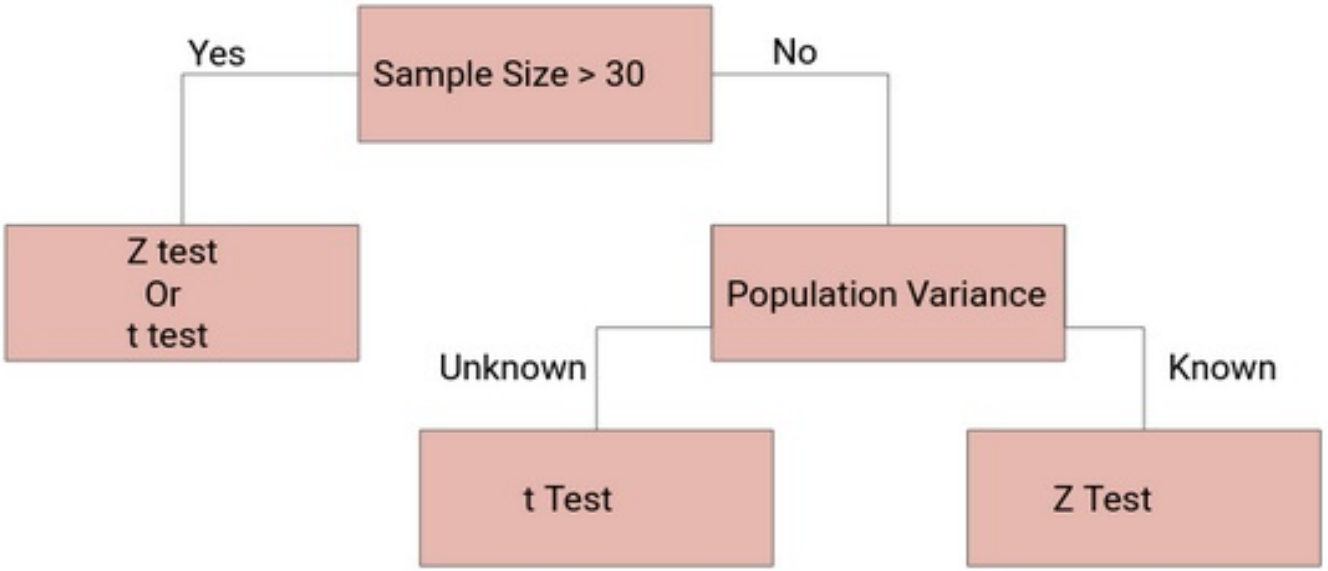
P value < 0.05



Thus, **p-value is less than 0.05 so we can reject the null hypothesis** and conclude that on average boys score 15 marks more than girls in the exam.

Deciding Between Z-Test and T-Test

- If the sample size is large enough, then the z-Test and t-Test will conclude with the same results. For a **large sample size**, **Sample Variance will be a better estimate** of Population variance, so even if population variance is unknown we can use the z-test using sample variance



Z Test vs T Test

	Z Test	T Test
Assumption	Population standard deviation is known	Population standard deviation is unknown
Sample Size	Large sample size (n > 30)	Small sample size (n < 30)
Distribution	Z-distribution	T-distribution
Test Statistic	(Sample mean – Population mean) / (Population SD / √n)	(Sample mean – Population mean) / (Sample SD / √n)
Hypothesis Testing	Test for a population mean or proportion	Test for a population mean
Degrees of Freedom	Not applicable	n – 1
Application	Used when the population standard deviation is known and the sample size is large	Used when the population standard deviation is unknown or the sample size is small
Example	Testing whether the average height of male adults is significantly different from a known value	Testing whether a new teaching method improves student test scores compared to the old method

CHI-SQUARE TEST

We use the chi-square test to compare categorical variables.

THE TWO TYPES OF CHI-SQUARE TEST

1. Goodness of fit test: determines if a sample matches the population
2. A chi-square fit test for two independent variables:
   used to compare two variables in a contingency table to check if the data fits
   A small chi-square value means that data fits.
   A large chi-square value means that data doesn't fit.
   The hypothesis we're testing is:
   - **Null**: Variable A and Variable B are independent.
   - **Alternate**: Variable A and Variable B are not independent.
   The statistic used to measure significance, in this case, is called chi-square statistic.
   The formula we use to calculate the statistic is:
   $X2 = \Sigma [(Or,c — Er,c)2 / Er,c]$ where
   Or,c=observed frequency count at level r of Variable A and level c of Variable B
   Er,c=expected frequency count at level r of Variable A and level c of Variable B

- A chi-square test is a statistical test that is used to compare observed and expected results.
- The goal of this test is to identify whether a disparity between actual and predicted data is due to chance or to a link between the variables under consideration.
- As a result, the chi-square test is an ideal choice for aiding in our understanding and interpretation of the connection between our two categorical variables.
- A chi-square test or comparable nonparametric test is required to test a hypothesis regarding the distribution of a categorical variable.
- Categorical variables, which indicate categories such as animals or countries, can be nominal or ordinal.
- They cannot have a normal distribution since they can only have a few particular values.

For example, a meal delivery firm in India wants to investigate the link between gender, geography, and people's food preferences.

- It is used to calculate the difference between two categorical variables, which are:
- As a result of chance or
- Because of the relationship

Formula For Chi-Square Test

$$x_c^2 = \frac{\Sigma\,(O_i - E_i)^2}{E_i}$$

Where
c = Degrees of freedom
O = Observed Value
E = Expected Value

Bayesian test

9.1.8 Bayesian Hypothesis Testing

Suppose that we need to decide between two hypotheses H_0 and H_1. In the Bayesian setting, we assume that we know prior probabilities of H_0 and H_1. That is, we know $P(H_0) = p_0$ and $P(H_1) = p_1$, where $p_0 + p_1 = 1$. We observe the random variable (or the random vector) Y. We know the distribution of Y under the two hypotheses, i.e, we know

$$f_Y(y|H_0), \quad \text{and} \quad f_Y(y|H_1).$$

Using Bayes' rule, we can obtain the posterior probabilities of H_0 and H_1:

$$P(H_0|Y=y) = \frac{f_Y(y|H_0)P(H_0)}{f_Y(y)},$$

$$P(H_1|Y=y) = \frac{f_Y(y|H_1)P(H_1)}{f_Y(y)}.$$

One way to decide between H_0 and H_1 is to compare $P(H_0|Y=y)$ and $P(H_1|Y=y)$, and accept the hypothesis with the higher posterior probability. This is the idea behind the *maximum a posteriori (MAP) test*. Here, since we are choosing the hypothesis with the highest probability, it is relatively easy to show that the error probability is minimized.
To be more specific, according to the MAP test, we choose H_0 if and only if

$$P(H_0|Y=y) \geq P(H_1|Y=y).$$

In other words, we choose H_0 if and only if

$$f_Y(y|H_0)P(H_0) \geq f_Y(y|H_1)P(H_1).$$

Note that as always, we use the PMF instead of the PDF if Y is a discrete random variable. We can generalize the MAP test to the case where you have more than two hypotheses. In that case, again we choose the hypothesis with the highest posterior probability.

MAP Hypothesis Test

Choose the hypothesis with the highest posterior probability, $P(H_i|Y=y)$.
Equivalently, choose hypothesis H_i with the highest $f_Y(y|H_i)P(H_i)$.

Stochastic Processes and Data Modeling: Markov process in Data modeling

- In the context of data modeling, Markov processes can be used to represent and analyze patterns in sequential data.
- Sequential data refers to data that is ordered or occurs over time, such as text, time-series data, sensor readings, and more.
- Markov processes offer a way to capture dependencies and transitions between states in sequential data, making them a valuable tool for data modeling and analysis.
- $P(Xn+1=x \mid Xn=xn, Xn-1=xn-1, \ldots, X0=x0) = P(Xn+1=x \mid Xn=xn)$

Here's how Markov processes can be applied to data modeling:

- **Text Generation and Natural Language Processing (NLP):** Markov processes can be used to model the generation of text sequences. Each state represents a word or a sequence of words, and the transition probabilities between states capture the likelihood of certain words following others. This approach can be used for text generation, autocomplete suggestions, and even machine translation.
- **Time-Series Analysis:** Markov processes are often used to model time-series data, where the state at each time step represents the value of the time series at that point. Transition probabilities can be derived from historical data to predict future values or identify anomalies.
- **Hidden Markov Models (HMMs):** Hidden Markov Models are a type of Markov process where the states are not directly observable but emit observable symbols. HMMs are widely used for tasks like speech recognition, part-of-speech tagging, and bioinformatics.
- **Recommendation Systems:** Markov processes can be employed to model user behavior over time in recommendation systems. States could represent different user preferences or states of interaction, and transitions can indicate how users navigate between those states.

- **Stock Market Modeling:** In finance, Markov processes can be used to model stock price movements. The current price would be the state, and transition probabilities could be derived from historical price data to make short-term predictions.
- **Web Page Navigation and Clickstream Analysis:** Markov models can be used to analyze user navigation patterns on websites. Each page or state represents a web page, and transitions reflect the likelihood of moving from one page to another. This information can help optimize website layouts and improve user experience.
- **Healthcare and Epidemiology:** Markov models can be applied to model disease progression, treatment outcomes, and other health-related processes. States could represent different health states, and transitions could represent transitions between health states based on treatment outcomes or natural progression.

There are several types of Markov processes, including:

- **Homogeneous Markov Chains:** The transition probabilities do not change over time. The process is "memoryless" in the sense that the future behavior is determined solely by the current state.
- **Time-Inhomogeneous Markov Chains:** The transition probabilities can change over time. In this case, the memoryless property still holds, but the probabilities can vary at different time steps.
- **Continuous-Time Markov Chains:** These are similar to discrete-time Markov chains, but the state transitions occur in continuous time rather than discrete time steps. They are often used to model systems where events happen in a continuous manner, such as in queuing systems.

Markov processes find numerous applications, such as:

- **Random Walks:** Modeling the movement of particles or individuals in a space.
- **Queueing Systems:** Studying waiting times and service processes in systems like customer service centers.
- **Markov Chain Monte Carlo (MCMC) Methods:** Used for simulation and statistical inference in complex models.
- **Natural Language Processing:** Modeling language generation and understanding tasks.
- **Finance and Economics:** Modeling stock prices, economic variables, and more.

- When applying Markov processes in data modeling, one key step is to estimate the transition probabilities from the available data.
- This often involves counting the occurrences of transitions in the data and normalizing to obtain probabilities.
- Depending on the complexity of the model and the available data, various techniques like Maximum Likelihood Estimation (MLE) or Bayesian methods can be used.

- Overall, Markov processes provide a powerful framework for modeling sequential data, capturing dependencies between states, and making predictions about future states based on historical information.

Hidden Markov Models

- Hidden Markov Models (HMMs) are a type of probabilistic model used in data modeling, particularly in situations where you have sequential data and want to capture both observed and hidden (latent) states that generate the data.
- HMMs have applications in various fields, including natural language processing, speech recognition, bioinformatics, finance, and more.
- They are particularly useful when dealing with data that has a sequential or temporal nature.

Here's how Hidden Markov Models work in data modeling:

- **Basic Components:** An HMM consists of two main components:

  Hidden States: These are the underlying states that generate the observed data. Hidden states are not directly observable; instead, they emit observations.

  Observations: These are the data points that we observe. Each observation is generated by one of the hidden states.
- **State Transitions:** Hidden states transition from one to another based on certain probabilities. These transition probabilities dictate how likely it is for the model to move from one state to another at each time step.
- **Emission Probabilities:** Each hidden state has associated emission probabilities that describe the likelihood of generating specific observations. These probabilities help link the hidden states to the observed data.
- **Learning and Inference:** Given a sequence of observations, the goal is to estimate the parameters of the HMM, including the transition probabilities and emission probabilities. This involves techniques such as the Expectation-Maximization (EM) algorithm or the Baum-Welch algorithm. Once the model is trained, it can be used for inference tasks like predicting hidden states based on observations or predicting future observations given a sequence of past observations.

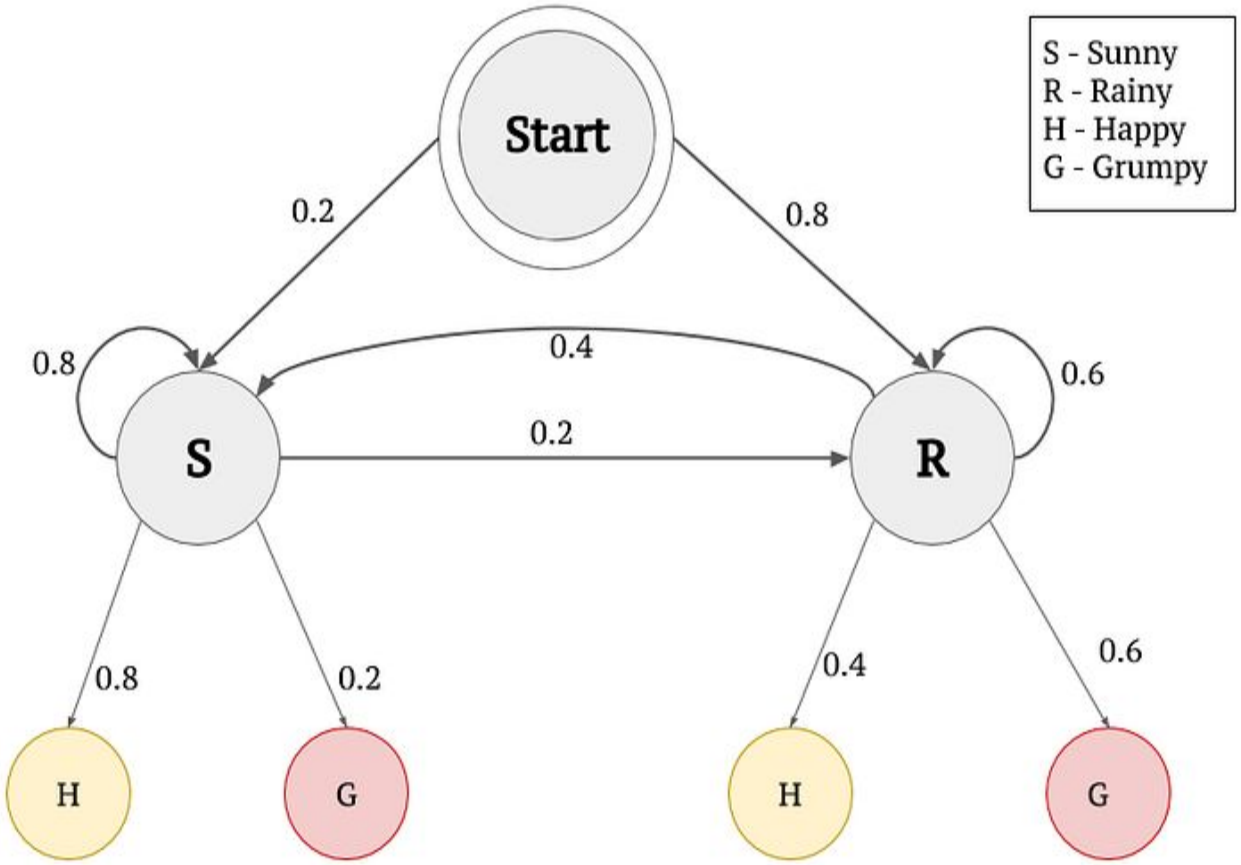
Applications:

HMMs have a wide range of applications:

- **Speech Recognition:** The observed data is the audio signal, and the hidden states correspond to phonemes or words.
- **Natural Language Processing:** HMMs can be used for part-of-speech tagging, named entity recognition, and machine translation.
- **Bioinformatics:** Modeling DNA sequences and protein sequences, identifying genes, and predicting secondary structures.
- **Finance:** Modeling financial time series data for forecasting and risk analysis.
- **Gesture Recognition:** Recognizing gestures from sensor data, such as accelerometer readings from wearable devices.
- **Robotics:** Mapping environments and localizing robots using sensor data.

Hidden Markov Model as a finite state machine

- Consider the example given below in Fig. which elaborates how a person feels on different climates.



- Set of states (S) = {Happy, Grumpy}
- Set of hidden states (Q) = {Sunny , Rainy}
- State series over time = z$\in$ S_T
- Observed States for four day = {z1=Happy, z2= Grumpy, z3=Grumpy, z4=Happy}

❖ The feeling that you understand from a person emoting is called the observations since you observe them.

❖ The weather that influences the feeling of a person is called the hidden state since you can't observe it.

Poisson Process

- The Poisson process is a mathematical concept used to model events that occur randomly over time or space.
- It is widely applied in various fields, including data modeling, to describe the occurrence of rare or random events.
- In the context of data modeling, the Poisson process can be particularly useful for modeling the distribution of events such as customer arrivals, website visits, accidents, machine failures, and more.

Here are the key components and characteristics of a Poisson process in data modeling:

- **Event Occurrence:** The Poisson process models the occurrence of discrete events in continuous time. These events could be anything from customer purchases to earthquakes, depending on the context.
- **Independence:** The events in a Poisson process are assumed to be independent of each other. This means that the occurrence of one event does not affect the probability of another event occurring.
- **Constant Rate:** The rate of event occurrences is assumed to be constant over time. This rate is denoted by λ (lambda) and represents the average number of events that occur per unit of time.
- **Memorylessness:** The probability of an event occurring in a small time interval is proportional to the length of the interval and is not affected by previous events. In other words, the process has no memory of the past.

- The Poisson process is characterized by the Poisson distribution, which describes the probability of a certain number of events occurring in a fixed interval of time or space, given the average rate of events.
- The probability mass function of the Poisson distribution is given by:

$P(X = k) = (e^{(-\lambda)} * \lambda^k) / k!$

Where:

- $P(X = k)$ is the probability of observing k events in the interval.
- e is the base of the natural logarithm (approximately 2.71828).
- λ is the average rate of events.
- k is the number of events.

- In data modeling, the Poisson process can be used to create models that predict the likelihood of a certain number of events happening within a given time period.
- It's important to note that while the Poisson process assumes a constant rate and independence, real-world data might not always conform perfectly to these assumptions.
- Variations and extensions of the Poisson process, such as the compound Poisson process and the non-homogeneous Poisson process, can be used to model more complex scenarios.
- Overall, the Poisson process is a valuable tool for modeling rare or random events in various fields, including data science, operations research, and reliability engineering.

Gaussian Processes

- Gaussian Processes (GPs) are a powerful and flexible technique for modeling complex data relationships, especially in the context of machine learning and data modeling.
- They are commonly used for regression, classification, and uncertainty estimation tasks.
- GPs provide a non-parametric and probabilistic approach to modeling data, making them particularly useful when dealing with limited data or uncertain environments.

Here are the key concepts and characteristics of Gaussian Processes in data modeling:

- **Probabilistic Modeling:** GPs provide a probabilistic framework for modeling data, which means that they not only predict a single output value but also provide a probability distribution over possible output values. This distribution captures the uncertainty associated with predictions.
- **Non-parametric Approach:** Unlike many traditional machine learning models that have fixed numbers of parameters (weights), GPs are non-parametric. They do not assume a predefined functional form for the underlying relationship between inputs and outputs. Instead, they adapt their complexity based on the observed data.
- **Flexibility:** GPs are versatile and can be used for various types of data, including continuous, discrete, or mixed data. They can also handle different types of input spaces, such as scalar inputs, multi-dimensional inputs, and even structured inputs like graphs.
- **Kernel Functions:** GPs rely on kernel functions (also known as covariance functions) to define the similarity or correlation between input data points. The choice of kernel function shapes the behavior of the GP model and determines how it captures dependencies in the data.

- **Regression:** In Gaussian Process regression, the goal is to predict a continuous output variable. GPs not only provide predictions but also estimate the uncertainty associated with those predictions. This uncertainty information is particularly valuable in decision-making processes.
- **Classification:** Gaussian Process classification extends GPs to handle discrete output variables, typically in binary classification problems. It estimates class probabilities and captures the uncertainty in these estimates.
- **Hyper parameters:** GPs have hyperparameters that influence their behavior, such as the parameters of the kernel function and noise level. These hyperparameters are typically learned from data through optimization methods.
- **Interpolation and Extrapolation:** GPs are well-suited for interpolation (predicting values within the range of observed data) and extrapolation (predicting values beyond the range of observed data). However, care must be taken when extrapolating, as GPs can become uncertain and less reliable in regions with little or no observed data.

Gaussian Processes are widely used in various fields, including:

- **Regression and Prediction:** When you have limited data and want to make predictions along with estimating prediction uncertainty.
- **Bayesian Optimization:** Optimizing expensive, black-box functions where each evaluation is costly. GPs guide the search by predicting the function's behavior and uncertainty.
- **Surrogate Modeling:** Replacing computationally expensive simulations with GP models to speed up optimization or sensitivity analysis.
- **Spatial Data Analysis:** Modeling spatial relationships between data points, such as in geostatistics.
- **Time Series Modeling:** Capturing complex temporal dependencies in data.
- **Anomaly Detection:** Identifying abnormal patterns in data by modeling normal behavior.

Auto-Regressive and Moving average processes,

- AutoRegressive (AR) and Moving Average (MA) processes are essential components in time series data modeling.
- They are commonly used to model and understand the underlying patterns, trends, and dependencies present in sequential data.

AutoRegressive (AR) Process:

In an AutoRegressive process, the value of a time series at a given time step is modeled as a linear combination of its past values.

- An AR(p) process of order p considers the previous p values in order to predict the current value.

Mathematically, an AR(p) process can be expressed as:

$$Xt=c+\phi 1Xt-1+\phi 2Xt-2+\ldots+\phi pXt-p+\varepsilon t$$

Where:

- Xt is the value of the time series at time
- c is a constant term.
- $\phi 1,\phi 2,\ldots,\phi p$ are the autoregressive coefficients.
- $Xt-1,Xt-2,\ldots,Xt-p$ are the past values of the time series.
- εt is a white noise error term at time t.

AutoRegressive (AR) Process Example:

- Suppose we have a monthly sales time series data for a retail store. We want to model this data using an AR(2) process, which means we'll consider the two previous values to predict the current value.
- The process can be described as:
- $Xt=c+\phi 1Xt-1+\phi 2Xt-2+\varepsilon t$
  Let's assume:
- $c=50$ (constant term)
- $\phi 1=0.6$ (coefficient for $Xt-1$)
- $\phi 2=0.3$ (coefficient for $Xt-2$)
- $X1=70$ (initial value)
- $X2=80$ (value at time $t=2$)

- Now, we can use the AR(2) process to generate values for the next time steps.
- $X3=50+0.6 \cdot 80+0.3 \cdot 70+\varepsilon 3$
- Here, $\varepsilon 3$ is a random error term.

Moving Average (MA) Process:

- In a Moving Average process, the value of a time series at a given time step is modeled as a linear combination of its past error terms. An MA(q) process of order q uses the previous q error terms to predict the current value. Mathematically, an MA(q) process can be expressed as:

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}$$

Where:

- X_t is the value of the time series at time t.
- μ is the mean of the time series.
- ε_t is the error term at time t.
- $\theta_1, \theta_2, \ldots, \theta_q$ are the moving average coefficients.
- $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_{t-q}$ are the past error terms.

AutoRegressive Moving Average (ARMA) Process:

An ARMA process combines both the autoregressive and moving average components to model time series data with both past values and past errors. An ARMA(p, q) process is given by the equation:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}$$

AR, MA, and ARMA processes are building blocks for more advanced time series models, such as the AutoRegressive Integrated Moving Average (ARIMA) and Seasonal ARIMA (SARIMA) models, which also incorporate differencing to make the time series stationary and account for seasonal patterns.

These processes are used in various fields such as finance, economics, meteorology, and more, where understanding and predicting patterns in sequential data are crucial for decision-making and analysis.

Moving Average (MA) Process Example:

Consider a daily temperature time series data for a city. We want to model this data using an MA(1) process, which means we'll consider the previous error term to predict the current value. The process can be described as:

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

Let's assume:

- $\mu = 25$ (mean temperature)
- $\varepsilon_{t-1} = 2$ (error term at time $t-1$)
- $\theta_1 = 0.7$ (coefficient for ε_{t-1})

Now, we can use the MA(1) process to generate temperature values for the next time steps.

$$X_{t+1} = 25 + \varepsilon_{t+1} + 0.7 \cdot 2$$

Here, ε_{t+1} is a random error term.

Bayesian Network

- Bayesian Networks (BNs) are a probabilistic graphical model used for representing and reasoning about uncertainty and dependencies among variables. They are particularly powerful for data modeling, as they provide a structured way to capture complex relationships in data, make predictions, and handle uncertainty.

Here's how Bayesian Networks are used in data modeling:

1. **Representing Dependencies:** Bayesian Networks use a directed acyclic graph (DAG) to represent the relationships among variables. Nodes in the graph represent variables, and directed edges represent conditional dependencies. This graphical representation makes it easy to visualize how variables influence each other.
2. **Probabilistic Modeling:** Each node in a Bayesian Network is associated with a probability distribution that describes the relationship between that node and its parents (nodes that have direct edges to it). This allows for probabilistic modeling, which is particularly useful when dealing with uncertain or incomplete data.
3. **Inference and Prediction:** Bayesian Networks can be used to perform various types of inference, such as computing posterior probabilities, making predictions, and estimating missing values. Given observed evidence (values of certain variables), BNs can propagate this information through the graph to compute probabilities of other variables.
4. **Causal Inference:** Bayesian Networks can help infer causal relationships between variables. By analyzing the structure of the network and the direction of edges, you can identify potential cause-and-effect relationships, aiding in understanding the underlying mechanisms in your data.

5.Decision Making and Risk Analysis: Bayesian Networks are used for decision analysis and risk assessment. By extending the basic structure with decision nodes and utility nodes, you can model decisions, outcomes, and their associated utilities. This enables optimal decision-making under uncertainty.

6. Anomaly Detection and Diagnosis: Bayesian Networks can be used for anomaly detection by modeling the normal behavior of a system and flagging instances that deviate significantly from the expected patterns. They are also used for diagnosing problems in complex systems by tracing the probabilistic relationships among variables.

7. Learning from Data: Bayesian Networks can be learned from data using algorithms that analyze the correlations and dependencies in the dataset. Learning the structure and parameters of a BN from data is valuable when dealing with large or complex datasets where manual specification of relationships may be impractical.

8. Feature Selection and Dimensionality Reduction: Bayesian Networks can help identify relevant features or variables that contribute most to the outcome of interest. This can be valuable in scenarios with high-dimensional data where feature selection or dimensionality reduction is necessary.

9. Text Mining and Natural Language Processing: Bayesian Networks can be used for tasks such as text classification, sentiment analysis, and information extraction in natural language processing. They can capture dependencies between words, topics, or sentiments in textual data.

- Overall, Bayesian Networks provide a flexible and intuitive way to model and reason about data.
- They are used in various domains, including healthcare, finance, manufacturing, environmental science, and more, where understanding complex relationships, handling uncertainty, and making informed decisions are crucial.

Regression

- Regression is a fundamental statistical technique used in data modeling to quantify the relationship between one or more independent variables (also known as predictors or features) and a dependent variable (also known as the target or response).
- The goal of regression analysis is to build a predictive model that can estimate the values of the dependent variable based on the values of the independent variables.
- There are several types of regression techniques, each designed for different types of data and modeling scenarios. Here are some common types of regression used in data modeling:

1. Linear Regression: Linear regression is the simplest form of regression. It assumes a linear relationship between the independent variables and the dependent variable. The model tries to find the best-fit line that minimizes the sum of squared differences between the predicted and actual values. Linear regression can be either simple (with one independent variable) or multiple (with multiple independent variables).

2. Polynomial Regression: Polynomial regression extends linear regression by introducing polynomial terms of the independent variables. This allows the model to capture more complex relationships that can't be adequately represented by a straight line.

3. Ridge Regression (L2 Regularization): Ridge regression is a regularized form of linear regression that adds a penalty term to the least squares objective function. This penalty term discourages the model from fitting large coefficients, which helps prevent overfitting and improves generalization.

4. Lasso Regression (L1 Regularization): Lasso regression is another regularized linear regression technique that uses a different penalty term. It not only discourages large coefficients but also encourages sparsity by pushing some coefficients to exactly zero. This can help with feature selection by automatically excluding less relevant variables.

5. Elastic Net Regression: Elastic Net combines both Ridge and Lasso regularization to mitigate their limitations. It provides a balance between selecting important features (like Lasso) and handling multicollinearity (like Ridge).

6. Logistic Regression: Despite its name, logistic regression is used for binary classification problems rather than regression. It models the probability of a binary outcome based on one or more independent variables. Logistic regression uses a logistic function to transform a linear combination of predictors into a value between 0 and 1, representing the probability of belonging to a particular class.

7. Time Series Regression: In time series data modeling, regression can be applied to predict future values of a time-dependent variable based on its past values and potentially other predictor variables.

8. Nonlinear Regression: When the relationship between the independent and dependent variables is nonlinear, various nonlinear regression techniques, such as exponential, logarithmic, or sigmoidal regression, can be used.

- Regression models are widely used in various fields such as economics, finance, biology, social sciences, and engineering.
- The choice of regression technique depends on the nature of the data, the underlying relationships, and the goals of the modeling task.
- Proper evaluation and validation techniques are crucial to ensure that the chosen regression model performs well on unseen data.

Queuing systems

- Queuing systems, also known as queueing theory or waiting line systems, are a mathematical framework used to model and analyze situations where entities (such as customers, tasks, or requests) arrive at a service facility, wait in line if necessary, and then receive service from one or more servers.
- Queuing systems are widely used in data modeling to understand and optimize various real-world scenarios involving waiting times, resource allocation, and system performance.

Here's how queuing systems are applied in data modeling:

1. Arrival Process: In a queuing system, entities arrive according to a certain pattern or distribution, known as the arrival process. This could be modeled using various distribution functions, such as exponential, Poisson, or even empirical distributions derived from historical data.

2. Service Process: Entities in a queuing system require service, which is provided by one or more servers. The service process defines how long it takes to serve an entity once it reaches the front of the queue. The service time can also follow various distribution functions.

3. Queueing Models: Different queuing models exist based on variations in arrival patterns, service times, the number of servers, and queue discipline (how entities are prioritized). Common queueing models include the M/M/1 (Poisson arrival, exponential service, single server) and M/M/c (Poisson arrival, exponential service, multiple servers) models.

4. Performance Metrics: Queuing systems are used to analyze performance metrics such as average waiting time, queue length, utilization of servers, and system throughput. These metrics help in understanding how well a system is functioning and in identifying potential bottlenecks.

- **5. Optimization:** Queuing theory can be used to optimize resource allocation. For example, a business might use queuing models to determine the optimal number of customer service representatives needed to minimize waiting times while controlling costs.
- **6. Simulation and Prediction:** Simulation is often used to model queuing systems in more complex scenarios, where the arrival and service processes may be more intricate. This can help predict system behavior and assess various "what-if" scenarios.
- **7. Network Queues:** In computer networking and telecommunications, queuing theory is used to model the behavior of data packets as they move through networks. This helps in designing and optimizing network architectures to ensure efficient data flow and minimal delays.
- **8. Application Examples:** Queuing systems are applied in various fields, such as retail (modeling customer checkout lines), healthcare (patient flow in hospitals), transportation (traffic congestion modeling), manufacturing (production line optimization), and more.

- By using queuing systems in data modeling, you can gain insights into how entities move through a system, predict waiting times, allocate resources effectively, and optimize system performance.
- This is particularly valuable for improving customer experience, resource utilization, and overall efficiency in a wide range of real-world scenarios.

THANK YOU