**Prerequisties:**

install.packages("tidyverse")

library(tidyverse)

dat = read.csv("dataset.csv", header = TRUE)

class(dat)

dat[3,1]

**Task 1:**

class(dat)

View(dat)

length(count.fields("dataset.csv",skip = 2))

sapply(dat,class)

lapply(dat, class)

difference = difftime(as.Date("2014-06-30"), as.Date("2013-07-01"), units = "weeks")

difference

**How many rows and columns are in the data?**

Number of Rows: 365

**How many hospitals are in the data?**

9

**What data types are in the data?**

Date object

Attendance int64

Admissions int64

Tri\_1 float64

Tri\_2 int64

...

Tri\_1.8

float64 Tri\_2.8

int64 Tri\_3.8

int64 Tri\_4.8

int64 Tri\_5.8

object Length: 64, dtype: object

There are all numeric and character data types in the data and the csv data is arranged as a Data frame.

**What time period does the data cover?**

0 1-Jul-13

1 2-Jul-13

2 3-Jul-13

3 4-Jul-13

4 5-Jul-13

...

360 26-Jun-14

361 27-Jun-14

362 28-Jun-14

363 29-Jun-14

364 30-Jun-14

364 days

**What’s the difference between “Attendance” and “Admissions”?**

0 99

1 97

2 84

3 106

4 96

...

360 112

361 112

362 95

363 87

364 113

Name: Admissions, Length: 365, dtype: int64

0 235

1 209

2 204

3 199

4 193

...

360 222

361 224

362 239

363 218

364 234

Name: Attendance, Length: 365, dtype: int64

the major difference between Attendances and admissions could be understood through their definitions i.e., attendance is the number of patients recorded as arriving at a public emergency department in our case Hospitals whereas Admissions are the number of patients who are admitted to the hospital for care and/or treatment subsequently

**What do the variables Tri\_1 , Tri\_2, ... represent?**

0 8.0

1 NaN

2 7.0

3 3.0

4 4.0

...

360 4.0

361 NaN

362 5.0

363 15.0

364 8.0

Name: Tri\_1, Length: 365, dtype: float64

So, there are a series of steps that happen Upon arrival in the ED, where people undergo a brief triage, or interview aka Triage’s, that helps to determine the nature and severity of their illness.

Triage categories are allocated to each patient based on an assessment of their presenting conditions, generally by the triage nurse, with triage 1 being the most urgent and triage 5 being the least urgent. (Triage 1: Resuscitation- immediate, within seconds; Triage 2: Emergency within 10 minutes; Triage 3: Urgent- within 30 minutes; Triage 4: Semi-urgent- within 60 minutes; Triage 5: Non-urgent - within 120 minutes). N/A - Values is less than 3 and has been suppressed.

Individuals with serious illnesses are then seen by a physician more rapidly than those with less severe symptoms or injuries

**Task 2:**

ed\_data\_link <- 'dataset.csv'

top\_row <- read\_csv(ed\_data\_link, col\_names = FALSE, n\_max = 1)

second\_row <- read\_csv(ed\_data\_link, n\_max = 1)

top\_row

i=2

for(i in range(0,64))

{

print(second\_row[i])

i= i + 2

}

column\_names <- second\_row %>%

unlist(., use.names=FALSE) %>%

make.unique(., sep = "\_\_") # double underscore

column\_names[2:8] <- str\_c(column\_names[2:8], '0', sep='\_\_')

column\_names

daily\_attendance <-

read\_csv(ed\_data\_link, skip = 2, col\_names = column\_names)

View(daily\_attendance)

class(daily\_attendance)

head(daily\_attendance)

sapply(daily\_attendance,class)

lapply(daily\_attendance, class)

j=c(2,9,16,23,30,37,44,51,58)

for(i in seq(2,64, by= 7)){

print(colnames(dat,do.NULL = TRUE, prefix = "col")[i])

}

**Task 2.2 Tidying data**

**1. Now we have a data frame. Answer the following questions for this data frame.**

**Does each variable have its own column? (1 point)**

Yes, each unique variable (data from different hospitals) has its own column

**Does each observation have its own row? (1 point)**

Yes, here each observation (dates in our case) has its own row.

**Does each value have its own cell? (1 point)**

Yes, there lies each unique value corresponding to each variable and its observation(date) in a separate cell.

**2. Use spreading and/or gathering (or their pivot\_wider and pivot\_longer new equivalents) to transform the data frame into tidy data (6 points). The key is to put data from the same measurement source in a column and to put each observation in a row. Please answer the following questions.**

**How many spreading (or pivot\_wider) operations do you need? (1 point)**

We need (9[for no of hospitals]x5[for number of triads])45 operations in the data set

**How many gathering (or pivot\_longer) operations do you need? (1 point)**

We need (9[for no of hospitals]x5[for number of triads])45 operations in the data set.

**Explain the steps in detail. (3 points)**

So, for this we need to download the database: -

# The first step involves

install.packages("tidyverse")

#Or alternatively we can also install just tidyr:

install.packages("tidyr")

There are two fundamental verbs of data tidying:

gather() function that takes multiple columns and gathers them into key-value pairs. It is actually responsible for making the data wider and longer.

spread() function that takes two columns (key & value) and spreads them into multiple columns. This makes the data longer and, in that terms, wider.

So, we apply these two functions on our variables Tri1\_1, Tri1\_2, Tri1\_3, Tri1\_4 and Tri1\_5(for all hospitals) and generate an immersive and tidy data.

**3. Are the variables having the expected variable types in R? Clean up the data types. (3 points)**

No, not all the variables are having expected variable types in R as there are several variables such as Triage\_1(Tri\_1) and Traige\_2(Tri\_2) that should have “numeric” data types but due to several null/Not available(N/A) values it treats the variable type as “character”.

Cleaning up done in R code.

**4. Are there any missing values? Fix the missing data. Justify your actions. (2 points)**

Yes, there are several values in various columns that are “N/A” that are missing or not available.

We can fix this by removing the NAs as a missing value likely represents no data in that group and make them equivalent to zero count signifying that they don’t count to any value in our data

**Task 3**

data %>%

dplyr::select(Attendance\_\_0,Admissions\_\_0,Tri\_1\_\_0,Tri\_2\_\_0,Tri\_3\_\_0,Tri\_4\_\_0,Tri\_5\_\_0)

```{r General Summary}

#Task 3.1

data %>%

summarise(Column\_n = colnames(dat,do.NULL = TRUE, prefix = "col")[2],

sum\_Attendance = sum(daily\_attendance$Attendance\_\_0),

sum\_Admissions = sum(daily\_attendance$Admissions\_\_0))

```

```{r}

#Task 3.2

x= c(mean(as.numeric(data$Admissions\_\_0)),mean(as.numeric(data$Admissions\_\_1)),mean(as.numeric(data$Admissions\_\_2)),mean(as.numeric(data$Admissions\_\_3)),mean(as.numeric(data$Admissions\_\_4)),mean(as.numeric(data$Admissions\_\_5)),mean(as.numeric(data$Admissions\_\_6)),mean(as.numeric(data$Admissions\_\_7)),mean(as.numeric(data$Admissions\_\_8)))

```

ED Demands Per Year

#Barplot for represnting ED demands

arg\_list<-c("Royal.Perth.Hospital","Fremantle.Hospital", "Princess.Margaret.Hospital.For.Children","King.Edward.Memorial.Hospital.For.Women"

,"Sir.Charles.Gairdner.Hospital","Armadale.Kelmscott.District.Memorial.Hospital","Swan.District.Hospital","Rockingham.General.Hospital","Joondalup.Health.Campus")

lines(x,names.arg =arg\_list,las=2,main = "Barplot for Represnting ED demands per year")

plot(as.numeric(data$Admissions\_\_0),main = "Linechart for Represnting ED demands per year",type="o")#For Just One Hospital

**Task 3.1 Select a hospital**

**Select a hospital and create a dataset for only the selected hospital. (1 point)**

Selecting First Hospital i.e. Royal Melbourne Hospitals Dataset

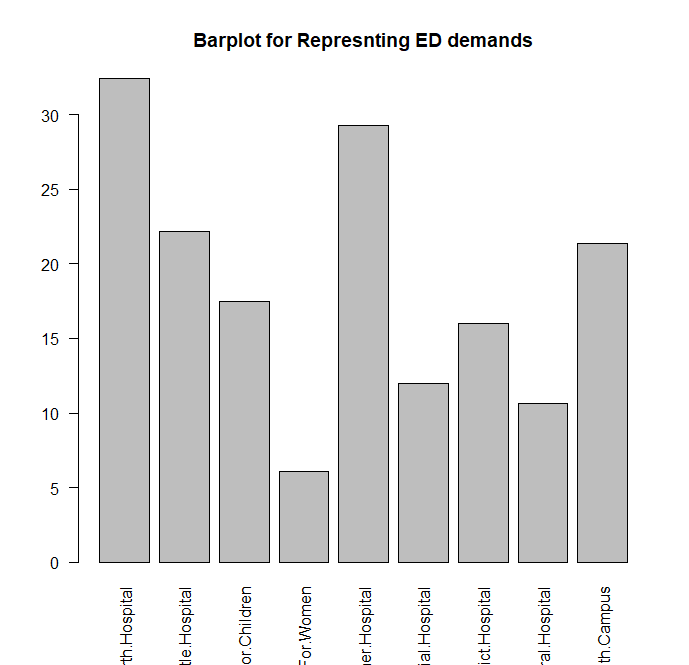
**Task 3.2 For the hospital selected, if we want to compare the volume of ED demands across the year, which plot can we use? Show your plot and explain what the plot shows. (Hint: Which variable measures the ED demands?) (3 points)**

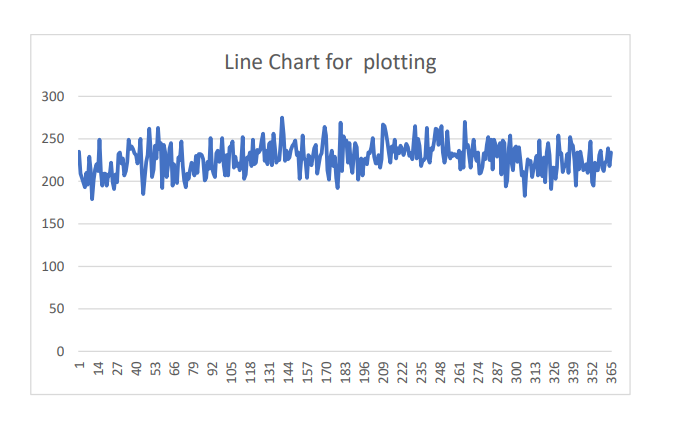
If we want to compare the volume of ED Demands over the year, we can measure it using Attendance and Admissions variable.

The plot we could use to efficiently differentiate between the ED demand over a year is a line chart as it shows day by day estimate of variable progression for a variable of a particular hospital.

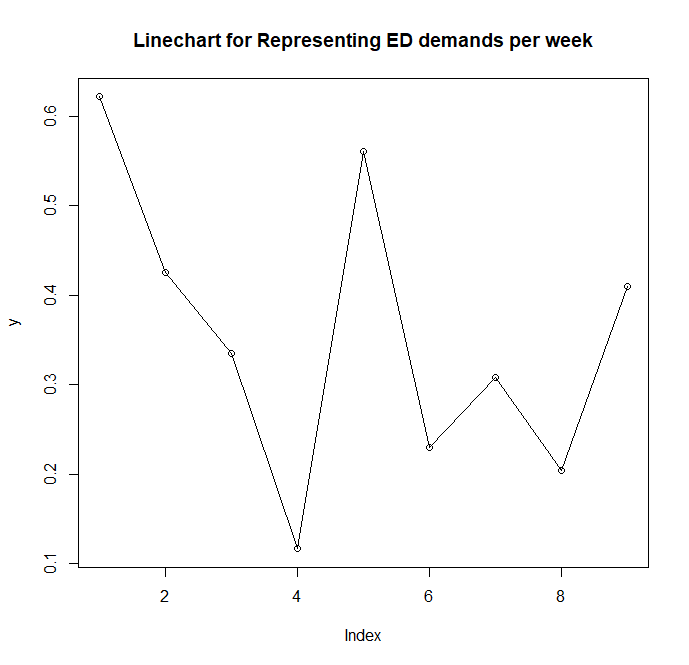
The plot we could use to efficiently differentiate between the ED demand over a year between different hospitals is a Bar plot because as from the diagram we could see that each different value of means of Admissions of various hospitals given can be observed and analyzed distinctively and uniquely.

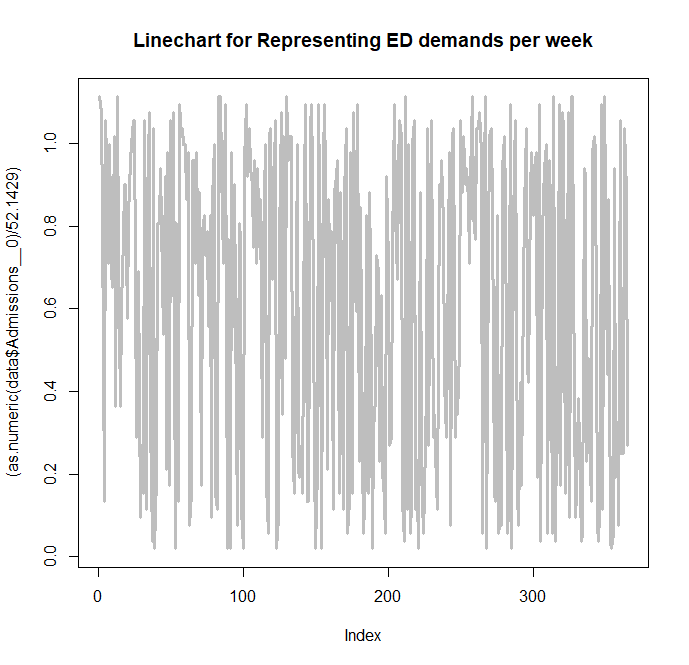
Here, we took mean for the consideration of the values for each hospital as it would be the perfect statistical measure that would give us an average number of people Admitting in the hospitals for the year; in result giving us a comparative distribution for comparing demand of ED across the year.





**Task 3.3 How do the ED demands change during a week? Show it visually using** **violin plots (2 points), describe the results (2 points) and provide your interpretation (2 points).**





Here is the change among all the hospitals observed over the ED’s demand over the week as we can see it fluctuates and goes down nearly at the 4 th day (which is the lowest) and is at peak during the start of the week and after 4 th day it goes on peaking high and low simultaneously.

**Task 3.4 Use skimr and fitdistrplus libraries to answer the following questions. Which distributions are appropriate for modelling the ED demand? (1 point)**

From the implementation On R file, we can conclude that weather Poisson Distribution or Lognormal distribution is appropriate for modelling ED Demand as here, an event (Admittance in hospitals as we’ve taken this variable for the evaluation) can occur any number of times during a time period plus events occur independently(assumption). In other words, if an event occurs, it does not affect the probability of another event occurring in the same time period. So, satisfying all the conditions for Poisson Distribution; I believe it is appropriate for ED Demand.

**Which variables meet the assumptions for the Poisson distribution and why? (2 points)**

The variable that meets the assumptions of Poisson Distribution are Attendance and Admissions as they are independent and one's occurrence does not affect the probability of another event. [which is not true in terms of Triads variable as if one triad (or tri\_5) does show 0 value then there is no scope for further testing as patient is highly unlikely to have any value for Tri\_1 as he/she’s not critically emergent.

**To reduce the dependence between consecutive days, randomly sample 150 records out of the whole dataset (all records for the selected hospital) for modelling (2 points).**

The rate of occurrence is constant; that is, the rate does not change based on time.

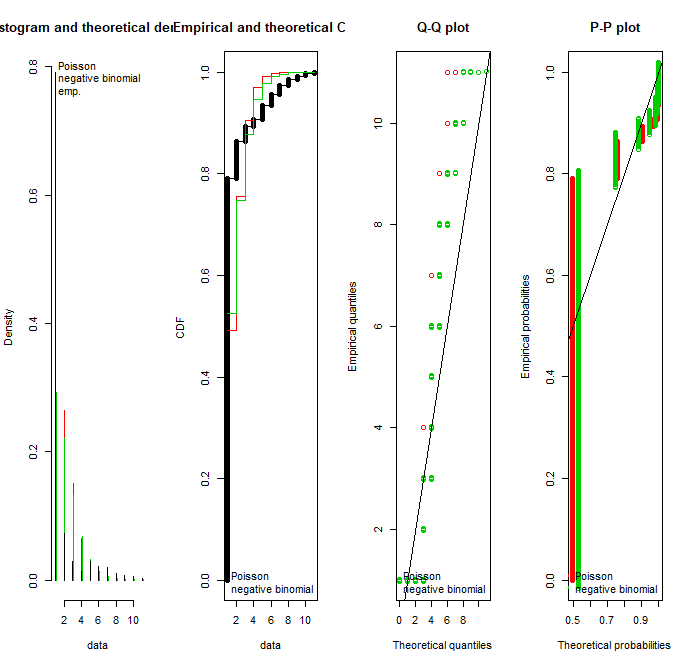
**Task 4.1: Fitting distributions (4 points)**

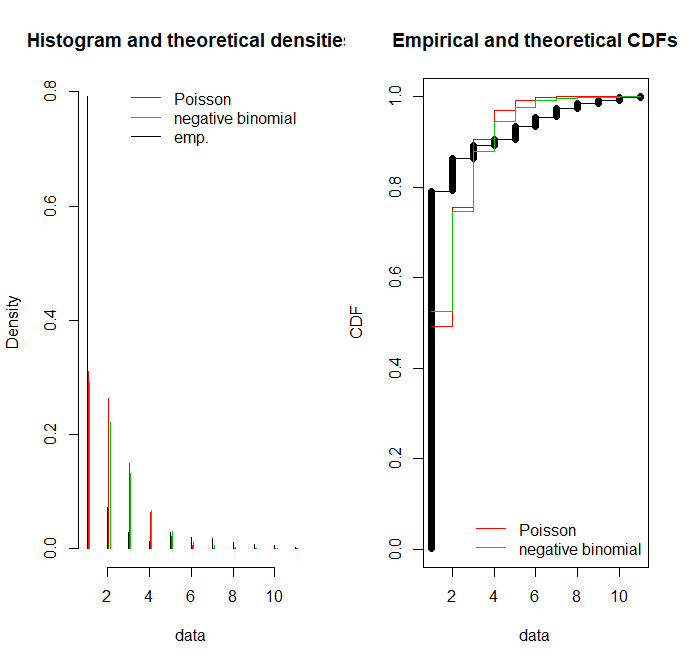
**Fit a Poisson distribution and a negative binomial distribution on**

**provided by the package fitdistrplus .**

**Task 4.2: Compare distributions (6 points)**

**Compare the log likelihood of two fitted distributions**





As we can see Negative Binomial is a better fit for the given data as it fits passably than Poisson distribution as seen from the graph plots generated above.

Poisson does somewhat fir the data but more accurately we could see Negative Binomial dist. Fits it more closely (all the green points in the data graph).

**Task 4.3: Try other distributions (research question 1) (10 points)**

**Find which distributions R stats library includes. Try to fit some of them to different Triage variables. Analyse and explain the results. Write a short report (200 words).**

Mathematical R software has already stood out because of its flexibility as a functional language that builds a bridge between software development and data analysis. For example, a single R power is an easy way one can develop and quickly adapt to meet the various needs arising from data management and social analysis while simultaneously using other languages ​​to deliver computer-based solutions. This book aims to introduce a framework that is accessible to statistical planning and software development using a variety of tools made available for R, from specific packages to genre control programs.

In most cases, when dealing with data we are actually dealing with data sets (see Chapter 03) where different environmental variables are aligned together (usually in columns). For datasets there is an easy way to get simple summary statistics that include adding a summary of the function to the database itself (instead of adding values as seen earlier).

As an example, let’s take a look at the Iris flower database contained in a built-in data package R. The data set contains 50 samples from each of the three types of Iris (Setosa, Virginica and Versicolor). Four dimensions were measured from each sample containing the length and width (in inches) of both sepals and petals. This database is widely used as a model as it was used by Fisher to develop a direct discrimination model based on where he intended to distinguish these three species from each other using a combination of these four factors.

Using this database, let's use the summary function to subtract the minimum, first quartile and quidile of thrid, median, mean and maximum statistics (in numerical variation in the database) and counting frequencies (input items).

**Task 5: Research question 2 (15 points)**

**There are more than one way to fit a distribution to a set of numbers. Produce a short literature review on different distribution fitting methods, showing the pros and cons of each method. 5 points will be given to relevance of the literature. 7 points will be given for the quality of comparative analysis of distribution fitting methods. 3 points will be given for the quality of presentation.**

The important distributions are: -

• The Bernoulli distribution, which takes value 1 with probability p and value 0 with probability q = 1 − p.

The Bernoulli distribution is a special case of the binomial distribution with n = 1. The kurtosis goes to infinity for high and low values of p, but for p = 1 / 2 the two-point distributions including the Bernoulli distribution have a lower excess kurtosis than any other probability distribution, namely −2.

The Bernoulli distributions for 0 ≤ p ≤ 1 forms an exponential family.

The maximum likelihood estimator of p based on a random sample is the sample mean.

• The binomial distribution, which describes the number of successes in a series of independent Yes/No experiments all with the same probability of success.

The binomial distribution is frequently used to model the number of successes in a sample of size n drawn with replacement from a population of size N. If the sampling is carried out without replacement, the draws are not independent and so the resulting distribution is a hypergeometric distribution, not a binomial one. However, for N much larger than n, the binomial distribution remains a good approximation, and is widely used.

• The Poisson binomial distribution, which describes the number of successes in a series of independent Yes/No experiments with different success probabilities.

The ordinary binomial distribution is a special case of the Poisson binomial distribution, when all success probabilities are the same, that is p 1 = p 2 = ⋯ = p n.

• The negative binomial distribution or Pascal distribution, a generalization of the geometric distribution to the nth success.

The Pascal distribution (after Blaise Pascal) and Polya distribution (for George Pólya) are special cases of the negative binomial distribution. A convention among engineers, climatologists, and others is to use "negative binomial" or "Pascal" for the case of an integer-valued stopping-time parameter r, and use "Polya" for the real-valued case.

For occurrences of "contagious" discrete events, like tornado outbreaks, the Polya distributions can be used to give more accurate models than the Poisson distribution by allowing the mean and variance to be different, unlike the Poisson. "Contagious" events have positively correlated occurrences causing a larger variance than if the occurrences were independent, due to a positive covariance term.

• The Poisson distribution, which describes a very large number of individually unlikely events that happen in a certain time, interval. Related to this distribution are a number of other distributions: the displaced Poisson, the hyper-Poisson, the general Poisson binomial and the

Poisson type distributions.

The Poisson distribution is an appropriate model if the following assumptions are true. k is the number of times an event occurs in an interval and k can take values 0, 1, 2, .... The occurrence of one event does not affect the probability that a second event will occur. That is, events occur independently. The rate at which events occur is constant. The rate cannot be higher in some intervals and lower in other intervals. Two events cannot occur at exactly the same instant; instead, at each very small sub-interval exactly one event either occurs or does not occur. Or the actual probability distribution is given by a binomial distribution and the number of trials is sufficiently bigger than the number of successes one is asking about. If these conditions are true, then k is a Poisson random variable, and the distribution of k is a Poisson distribution.

The Beta distribution on [0,1], a family of two-parameter distributions with one mode, of which the uniform distribution is a special case, and which is useful in estimating success probabilities. The beta distribution has been applied to model the behaviour of random variables limited to intervals of finite length in a wide variety of disciplines.

In Bayesian inference, the beta distribution is the conjugate prior probability distribution for the Bernoulli, binomial, negative binomial and geometric distributions. For example, the beta distribution can be used in Bayesian analysis to describe initial knowledge concerning probability of success such as the probability that a space vehicle will successfully complete a specified mission. The beta distribution is a suitable model for the random behaviour of percentages and proportions. The usual formulation of the beta distribution is also known as the beta distribution of the first kind, whereas beta distribution of the second kind is an alternative name for the beta prime distribution.

A probability function is an element of a probability space that shows the possible outcomes of a given set of values. It is also used to measure the probability of a given subset of values. The probability function P can accept subsets of the sample space's data, such as the heads and tails. However, since the space is full of random variables, its representation is usually transformed into a set of numbers. This type of probability distribution is more common to study. The probability function is a set of possible outcomes that refers to the probability space displaystyle XX, where X is the set of possible results. The same concept holds for displaystyle E subset X Esubset X, where the probability can be measured.

A probability distribution is divided into two classes: the discrete probability distribution and the mass function. The former refers to the list of probabilities encoded by the discrete list of possible outcomes, while the latter refers to the probability distribution. The concept of continuous probability distributions is applicable to situations where the outcome can be computed in a continuous range. For instance, real numbers can be used to determine the temperature of a given day. The existence of probabilities is generally described as a probability density function. The distribution is also referred to as the integral of the probability function.

A probability distribution that has a one-dimensional sample space is called uni variate, while one that has a vector space of dimension 2 is called multivariate. Both distributions give the probabilities of a random variable and a list of two or more randomly-variable values. The various components of the probability function, as well as the various probability mass functions and the characteristic function, help to identify a definite probability distribution.

A probability distribution is a description of the probabilities of an event or set of events that can be derived from the sample space. It can be derived by dividing the probability of an event by its corresponding number or set of random digits.

To define the probability of random distributions, it is usually necessary to distinguish between continuous and discrete random variables. In the case of continuous random variables, the probability is assigned to the number of possible outcomes. When a variable takes values from a linear continuum, any individual outcome has zero probability. In other words, events that include infinitely many possible outcomes can have positive probability. A continuous probability distribution can be described using the probability density function or the cumulative distribution function. The former refers to the infinitesimal likelihood that a given value will be met in a given interval.

The number of students who enter the student union per minute will not follow a Poisson distribution due to the nature of the classes and the arrival of individual students. The number of earthquakes per year may not follow a random distribution if the probability of aftershocks increases due to a large earthquake. Examples of events that are guaranteed but not distributed are not distributed and may be models using a Zero-inflated Poisson distribution.

We have seen that time domain samples of a speech signal are very well described by an LD. This gives a first-order pdf characterization for speech. In order to obtain a higher order statistical description, we first transform the speech signal into uncorrelated components. We then show that these components themselves are distributed like Laplacian random variables with different parameters.

The speech signal is regarded as a nonstationary process, and yet we will see that the shapes of the pdf of its components are almost time-invariant and are LDs with slowly time-varying parameters. A KLT is used here to decorrelate the data vector and to separate the speech components. It is assumed that successive speech samples could be accurately modeled by the pdf of the KLT components that are uncorrelated random variables. This modeling is described by the KLT and the energy (or mean absolute value) of the KLT components. Speech is often viewed as a nonstationary process; therefore, this transformation varies slowly with time and should be updated with the arrival of new samples. In this letter, the KLT is adaptively estimated by the algorithm proposed in [9]. We consider the three most dominant components of a speech signal in the KLT domain obtained by the algorithm. The pdfs of these components (that are uncorrelated) for a 10-s speech sample are approximated very well by Laplacian pdfs with different parameters. Note that minor components that have very small energies are neglected in practice.

To further investigate the behavior of speech components, the test is similarly applied to these three components. The results are depicted, where the parameter for each KLT component is also computed over every 200-ms time frame with 100-ms overlaps between frames. Again, we see that the K coefficients are better approximated by the LD rather than by the others. Hence, the joint pdf of samples of a speech signal could be approximated by assuming that speech in the KLT domain is a multivariate random vector with uncorrelated LD elements and different, slowly time-varying parameter

In the concept of probability, the function of probability (PDF), or the density of random random variables, is a function that has a value in any sample (or point) in the sample space (a set of values ​​that can be taken at random variables) is interpreted as providing opportunities related to random random variables. In other words, while the absolute probability of random random variance taking any given value is 0 (because there is an unlimited set of possible values ​​to start with), a PDF value in two different samples can be used to include, in any particular drawing of random variables, how much that random variables can be equal to one sample compared to another sample.

In a more accurate sense, PDF is used to determine the probability of a random variance that falls within a given range of values, as opposed to taking a single value. This opportunity is provided by the compilation of this flexible PDF at that distance - that is, it is given less space for congestion but above the horizontal axis and between the lowest values ​​in the range. The task of consolidation of opportunity is not contradictory everywhere, and its universal integration is equal to 1.

The terms "possible distribution function" and "job opportunities" are also sometimes used to denote job-strengthening work. However, this use is not uncommon among probabilities and mathematicians. In some sources, the "opportunity distribution function" may be used when the distribution of opportunity is defined as activity in excess of normal sets of values ​​or it may refer to accumulated distribution activity, or it may be a massive energy activity (PMF) rather than congestion. The “overcrowding” itself is used for mass work opportunities, leading to further confusion. Typically, PMF is used in the context of random random variables (random variables that take values ​​from the calculated set), while PDF is used in the context of continuous random variables.

Suppose some germs last four to six hours. The chance that the bacterium will stay upright for 5 hours is equal to zero. Most bacteria live for about 5 hours, but there is no chance that any given bacteria will die in about 5.00 ... hours. However, it is possible that the bacterium dies within 5 hours and 5.01 hours beyond measure. Suppose the answer is 0.02 (e.g., 2%). After that, the probability that the bacterium dies between 5 and 5 hours of 5,001 should be about 0.002, because this time period is one-tenth longer than before. The probability of a bacterium dying within 5 hours and 5.0001 hours should be about 0.0002, and so on.

In this example, the average (probability of death at rest) / (interval) is approximately equal, and is equal to 2 per hour (or 2 - 1 hour). For example, there is a 0.02 chance of death in a 0.01 hour period between 5 and 5.01 hours, and (0.02 chances / 0.01 hours) = 2 - 1 hour. This 2 - 1 hour number is called with a chance of dying in about 5 hours. Therefore, the chances of the bacterium dying within 5 hours can be recorded as (2 hours - 1) dt. It is possible that the bacterium dies within a five-hour window, when the dt time for this window. For example, the chance that it lives longer than 5 hours, but is shorter than (5 hours + 1 nanosecond), is (2 hours - 1) × (1 nanosecond) ≈ 6 × 10-13 (using unit conversion 3.6 × 1012 nanoseconds = 1 hour).

**Task 6: Ethics question (7 points)**

**During your work, have you identified any issues that have ethical implications? (2 points) Does it concern** **security or privacy? (2 points) How was the risk mitigated? (3 points)**

Yes, there are some issues that actually have a major implication ethically; one of them being the “Global Warming” leading to extremely heating conditions around places. With the inevitable rise of extreme weather events, it is crucial that we better understand its potential impact on our everyday life.

Yes, it concerns majorly on our security as more and more people face to be the victims of such crisis caused as a result of degradation and harmful activities caused by humans to the nature.

In order to mitigate the risk, we could take precautionary measure in order to safeguard the environment we life in and sort of protect it sustainably to maintain harmony and balance in the nature.

**Task 7: Reflection (2 points)**

**Answer the following questions:**

**1. What help did you receive from other students? What did you learn from them? (1 point)**

I received motivation from other students to work quickly and perfectly towards the assignment task given and explore more content on what needs to be done and moreover, learn additional content as well from extra materials widening my scope in the field.

2. Please estimate the mark that you will receive for assignment 1. Please provide both a point estimate and an interval estimate (a confidence interval). You don’t need to provide a mathematical model, but please explain how you use conditional information to reach the estimates. Based on the conditional information, explain what you would have done differently to improve that mark?

I should receive 90-95% of the marks in this assignment as I believe I have tried my level best to deliver the best quality work through research, learning a completely new language, excelling at it by working in real term data and models.

My confidence level is 95% and my point estimate would be 89.I have used the information provided and gone and delivered results specific to the questions asked and answered them correctly to reach solutions so that forms my basis of my estimates above.

I would have started learning about the subject in inter trimester breaks and started developing small app and codes just for practice and getting good grip on the content.