S.K Somaiya College of Arts, Science and Commerce
Somaiya University
Vidyavihar, Mumbai 400 077

Report Writing
On

Prediction of semen parameter from environmental factors and lifestyle using Statistical Analysis

MSc Statistics Part-2
SEM - III
Paper: - Statistical Analysis using SAS

Submitted by
Varsha Jadhav (31031820020)
Niki Mehta (31031820027)
Jyothi Puligilla (31031820038)
Vedang Sawant (31031820041)
Moheed Tai (31031820044)
Vishal Yadav (31031820046)

INDEX

ABSTRACT

Fertility rates have dramatically decreased in the last two decades, especially in men. It has been described that environmental factors as well as life habits may affect semen quality. In this report we use Statistical techniques in order to predict semen characteristics from environmental factors, life habits and health status, as a possible. Decision Support System that can help in the study of the male fertility potential. Data of One hundred young healthy volunteers were taken as sample which has been analysed by Statistical technique. Data was also recorded to fulfil a validated questionnaire about life habits and health status. Sperm concentration and percentage of motile sperm were related to socio-demographic data, environmental factors, health status, and life habits.

INTRODUCTION

In the last two decades there has been a notable decline in fertility rates. It was considered that this decline is due to changes in behaviour related to economic aspects, with the incorporation of women into labour and the consequent delay in the age at which you decide to have offspring, and the widespread use of contraceptives. Although it is clear that the social aspect has contributed significantly to this global decline in fertility some authors suggest that occurred synchronously with the deterioration of reproductive health caused by adverse biological factors Rates of demand for assisted reproduction treatment show an increase of situations where the male factor is altered. Over the past decades since the publication of a meta-analysis directed by Elisabeth Carlsen, remains a debate about the possibility of a decline in seminal quality. Numerous studies show a decrease in semen parameters of men, although there are studies that found no evidence of that decline. About the causes of this possible decline in semen quality that affect male fertility, several factors have been considered, from an increase in the incidence of male reproductive diseases, environmental or occupational factors, to a certain lifestyle. Semen analysis is the cornerstone of the male study. Although semen analysis alone cannot determine whether a male can have offspring, it is a good predictor of male fertility potential. Semen analysis is also necessary to evaluate candidates to become semen donors. In this paper, we study the male fertility, approaching the problem from the perspective of possible influence of environmental factors and life habits in semen quality. To do that we use Statistical techniques to produce a Decision Support Systems (DSS) that can help in the prediction of semen parameters.

LITERATURE REVIEW

David Gil, Jose Luis Girela, Joaquin De Juan, M Jose Gomez-Torres, and Magnus Johansson used three artificial intelligence techniques such as decision trees, Multilayer Perceptron and Support Vector Machines in order to compare and evaluate their performance in the prediction of the seminal quality from data of environmental factors and lifestyle.

Filippo Amato, Alberto Lopez, Eladia Maria Pena-Mendez published article regarding use of artificial neural network for medical diagnosis in Applied Biomedicine journal in which it is mentioned an extensive amount of information is currently available to clinical specialists, ranging from details of clinical symptoms to various types of biochemical data and outputs of imaging devices. Each type of data provides information that must be evaluated and assigned to a particular pathology during the diagnostic process. To streamline the diagnostic process in daily routine and avoid misdiagnosis, artificial intelligence methods (especially computer aided diagnosis and artificial neural networks) can be employed. These adaptive learning algorithms can handle diverse types of medical data and integrate them into categorized outputs.

OBJECTIVE

- To predict the fertility of semen
- To explore the data for understanding of semen parameter

DATA

Data contain sperm concentration of 100 individuals which were provided voluntarily. Semen sample were analysed according to the WHO 2010 criteria. Sperm concentration are related to socio-demographic data, environmental factors, health status, and life habits. Data on following variables were given viz. Season in which the analysis was performed. Age at the time of analysis, Childish Diseases, Accident or serious trauma, Surgical intervention, High fevers in the last year, Frequency of alcohol consumption, Smoking habit, Number of hours spent sitting per day and Diagnosis

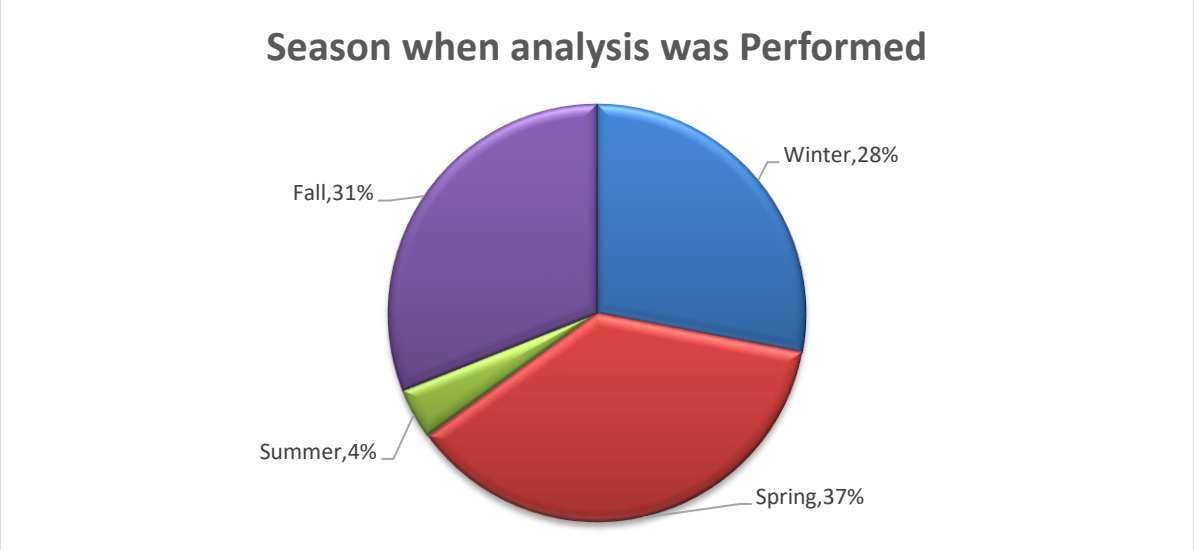

Fig -1

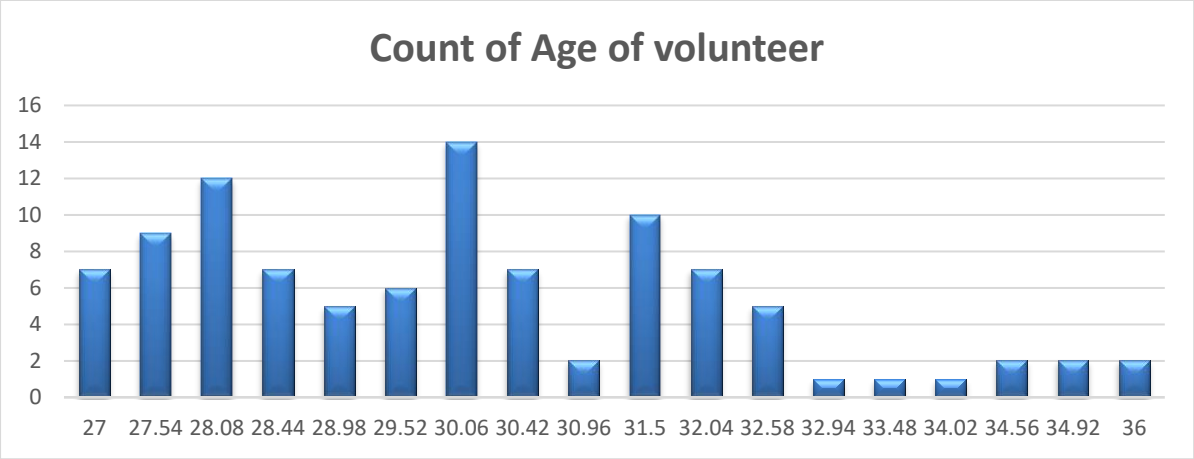

Fig - 2

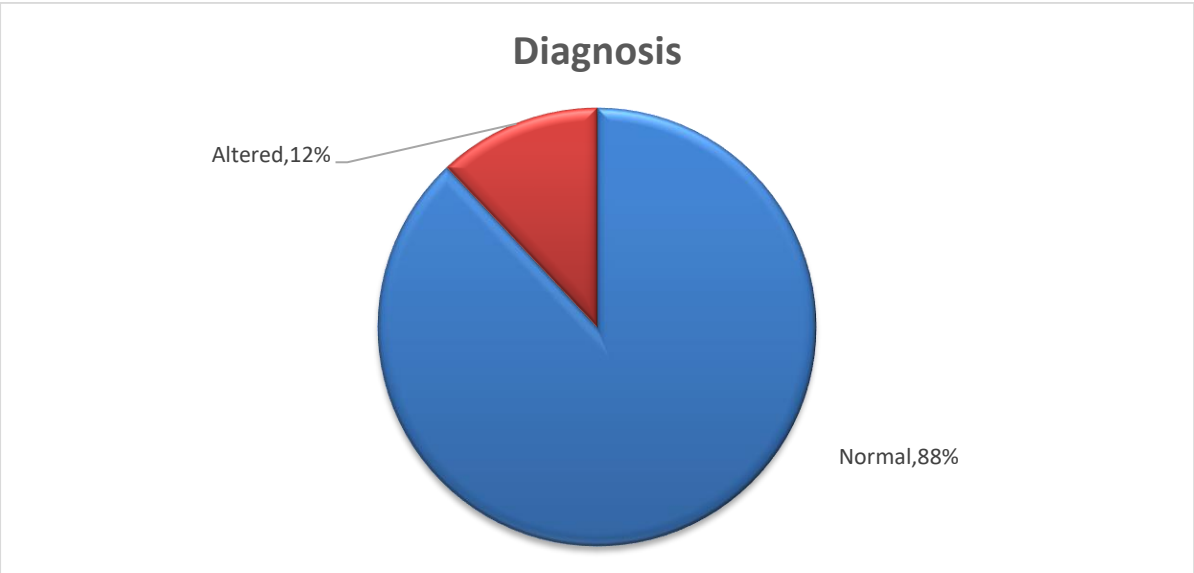

Fig – 3

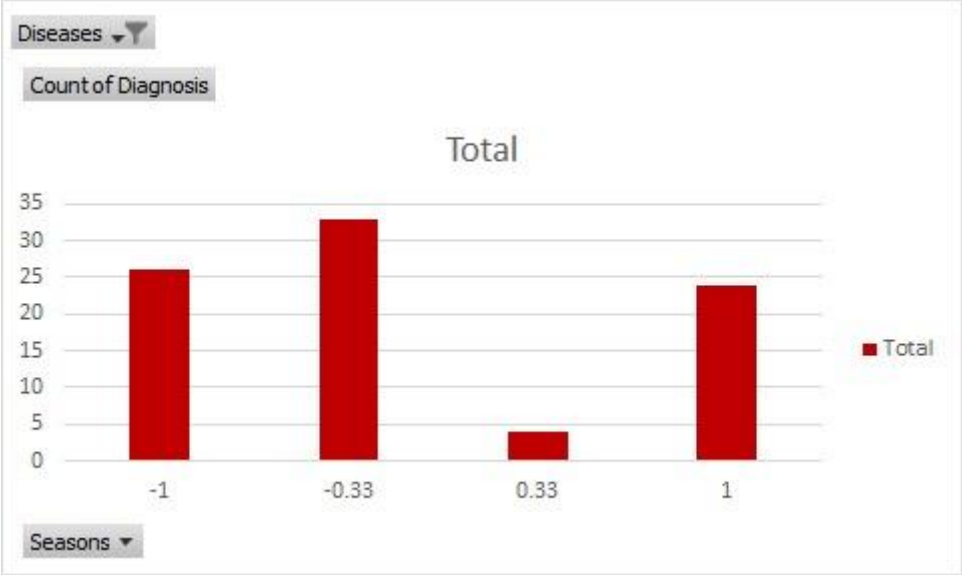
Diseases
Count of Diagnosis
Total
35
30
25
20
15
10
5
0
-1
-0.33
0.33
1
Total
Seasons

Fig-4

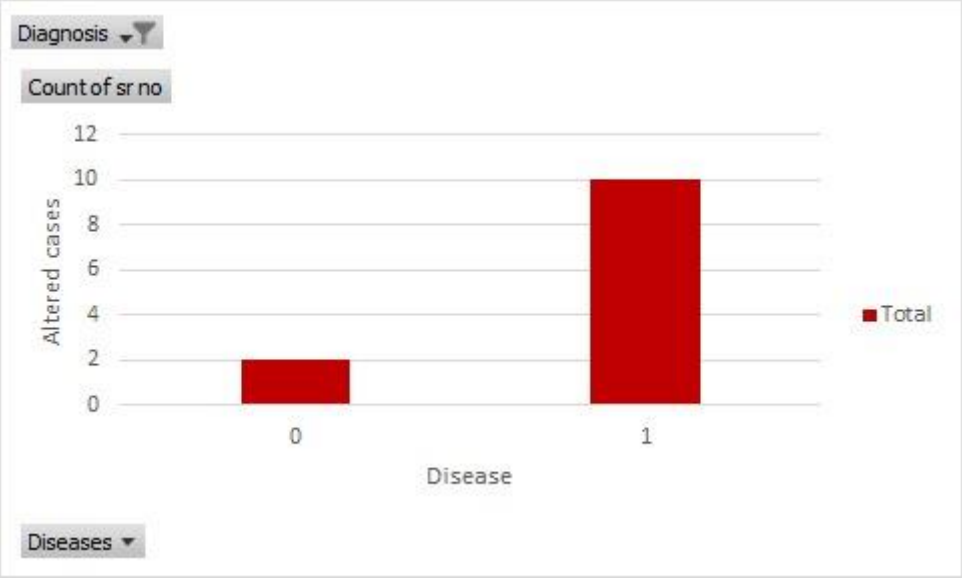
Diagnosis
Count of sr no
12
10
8
6
4
2
0
Altered cases
0
1
Disease
Total
Diseases

Fig-5

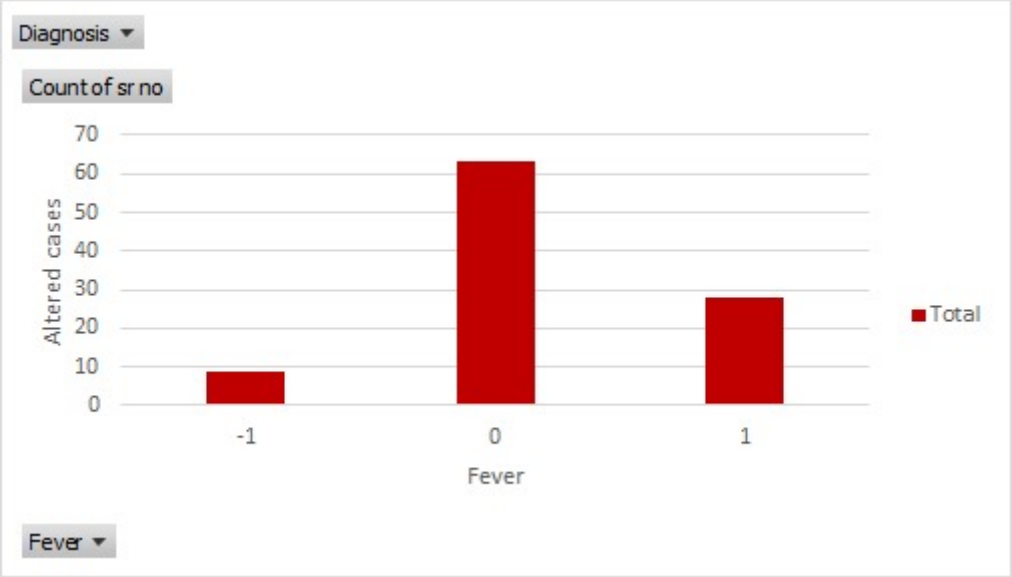

Fig-6

METHODOLOGY

Initially, data of 100 semen observations with 10 attributes was collected by WHO in 2010 was extracted from the UCI machine learning repository. The extracted data was multivariate with no missing observation. Total 9 input features were present in data such as Season, Age, Childish disease, Accidental or serious trauma, surgical intervention, high fevers, frequency of alcohol consumption, smoking habit, number of hours spent sitting per day, and one target feature i.e., diagnosis. The data was imbalanced since the diagnosis variable was having 88 normal observations and 12 altered observations. To make data balanced, altered observations were oversampled to 88 observations. Due to oversampling, the data size was increased to 176. To standardize data, min-max scaling was used. After pre-processing data was imported into SAS for further analysis. The imported data was divided into train and test set with 140 and 36 observations respectively. Since the outcome variable was having binary output i.e., normal and altered logistic regression model was fitted on the training set, and to make a prediction as well as to check the validity of the model, the testing set was used.

RESULTS

Response Profile		
Ordered Value	Diagnosis	Total Frequency
1	0	70
2	1	10

Probability modeled is Diagnosis='0'.

Fig - 1

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	62.283	66.759
SC	64.665	90.579
-2 Log L	60.283	46.759

Fig – 2

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	13.5241	9	0.1403
Score	11.4636	9	0.2453
Wald	8.7370	9	0.4619

Fig – 3

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.8680	0.6413	19.9975	<.0001
Seasons	1	-0.6459	0.4228	2.3344	0.1265
Age	1	-0.6817	0.4725	2.0812	0.1491
Disease	1	-0.1372	0.3501	0.1537	0.6951
Trauma	1	0.8642	0.4958	3.0385	0.0813
Surgical_interventio	1	0.0224	0.4278	0.0027	0.9583
Smoking	1	0.8018	0.4336	3.4196	0.0644
Fever	1	0.7389	0.4005	3.4047	0.0650
alcohol_consumption	1	-0.2120	0.4270	0.2466	0.6194
Sitting	1	-0.7127	0.4652	2.3468	0.1255

Fig - 4

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Seasons	0.858	0.475	1.552
Age	0.434	0.232	0.809
Disease	0.601	0.384	0.941
Trauma	3.509	1.999	6.162
Surgical_interventio	1.200	0.716	2.011
Smoking	1.687	1.005	2.831
Fever	1.898	1.187	3.036
alcohol_consumption	0.855	0.550	1.329
Sitting	0.589	0.347	1.000

Fig - 5

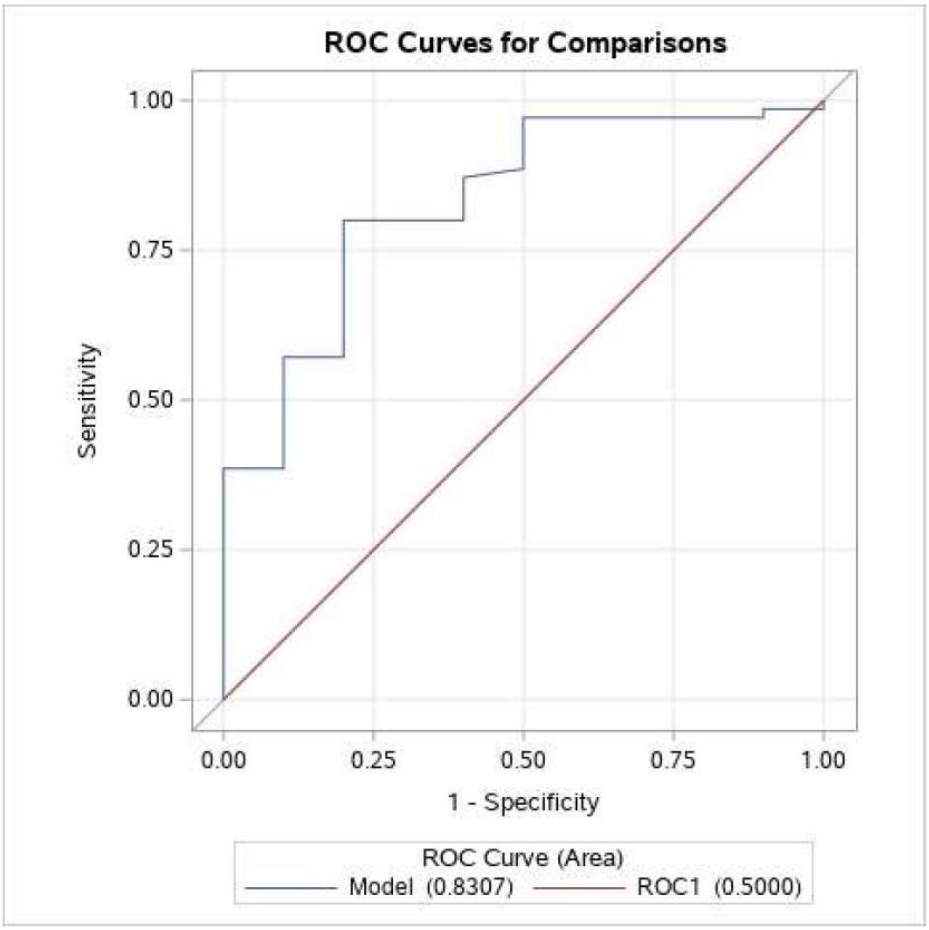

Fig - 6

ROC Association Statistics							
	Mann-Whitney						
ROC Model	Area	Standard Error	95% Wald Confidence Limits		Somers' D	Gamma	Tau-a
Model	0.8307	0.0691	0.6952	0.9662	0.6614	0.6624	0.1465
ROC1	0.5000	0	0.5000	0.5000	0	.	0

Fig – 7

CONCLUSIONS

- Maximum samples were collected in spring, winter and fall season whereas in summer only 4% samples were collected by researcher.
- The concentration of patients under study with age 30 & less is more.
- The above findings shows that 88% of the results were normal.
- Samples collected in winter and spring season were having maximum number of altered fertility individual.
- Diseased individuals were showing maximum altered cases
- Individuals who were gone through more than 3 months of fever were showing maximum number of altered cases
- The accuracy of the logistic model fitted is 83.%.

REFERENCES

- https://archive.ics.uci.edu/ml/datasets/Fertility
- Software used: SAS,Excel