Analysis of Factors affecting Male Fertility

ISyE 6739

Anirudha Sundaresan

Sharbani Pandit

Mayuri Rajput

Sahana Ravishankar

INDEX

CONTENTS

INDEX OF FIGURES

INTRODUCTION

Infertility affects approximately **1 out of every 6 couples**. An infertility diagnosis is given to a couple who are unable to conceive over the course of one year. When the problem lies with the male partner it is referred to as **male infertility**. Male infertility factors contribute to **approximately 30% of all infertility cases**, and male infertility alone accounts for approximately one-fifth of all infertility cases.

This project attempts to identify the factors that have a more pronounced effect relative to other factors that are otherwise widely perceived to be the reason of low sperm count. The results however may seem contentious as the data size is small to make any definitive remarks about ruling some factors out.

DATASET DESCRIPTION

Semen samples have been collected from 100 volunteers according to the WHO criteria. Sperm concentration are related to socio-demographic data, environmental factors, health status, and life habits. The dataset comprises of following predictors:

- Season in which the analysis was performed. 1) winter, 2) spring, 3) Summer, 4) fall. The seasons in the original dataset are categorized as (-1,-0.33,0.33,1)
- Age at the time of analysis (18-36 years)
- Childhood diseases contraction. 1) yes 2) No
- Accident or serious trauma. 1) Yes 2)No
- Surgical Intervention 1) Yes 2)No
- High fevers in last year 1) More than 3 months ago 2) Less than 3 months ago 3) No
- Frequency of alcohol consumption 1) several times a day, 2) every day, 3) several times a week, 4) once a week, 5) hardly ever or never
- Smoking habit 1) never, 2) occasional 3) daily
- Number of hours sitting per day (0,1)
- Output: Diagnosis normal (N), altered (O)

Data Set Characteristics:	Multivariate	**Number of Instances:**	100	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	10	**Date Donated**	2013-01-17
Associated Tasks:	Classification, Regression	**Missing Values?**	N/A	**Number of Web Hits:**	116991

DATA ANALYSIS

Descriptive Statistics

Descriptive statistics for this data has limited options as the data is categorical in nature. Sunburst chart is used to understand what percentage of altered people had certain lifestyle habits or subjected to an environment.

It can be inferred from the diagram that following factors have most pronounced effect on fertility:

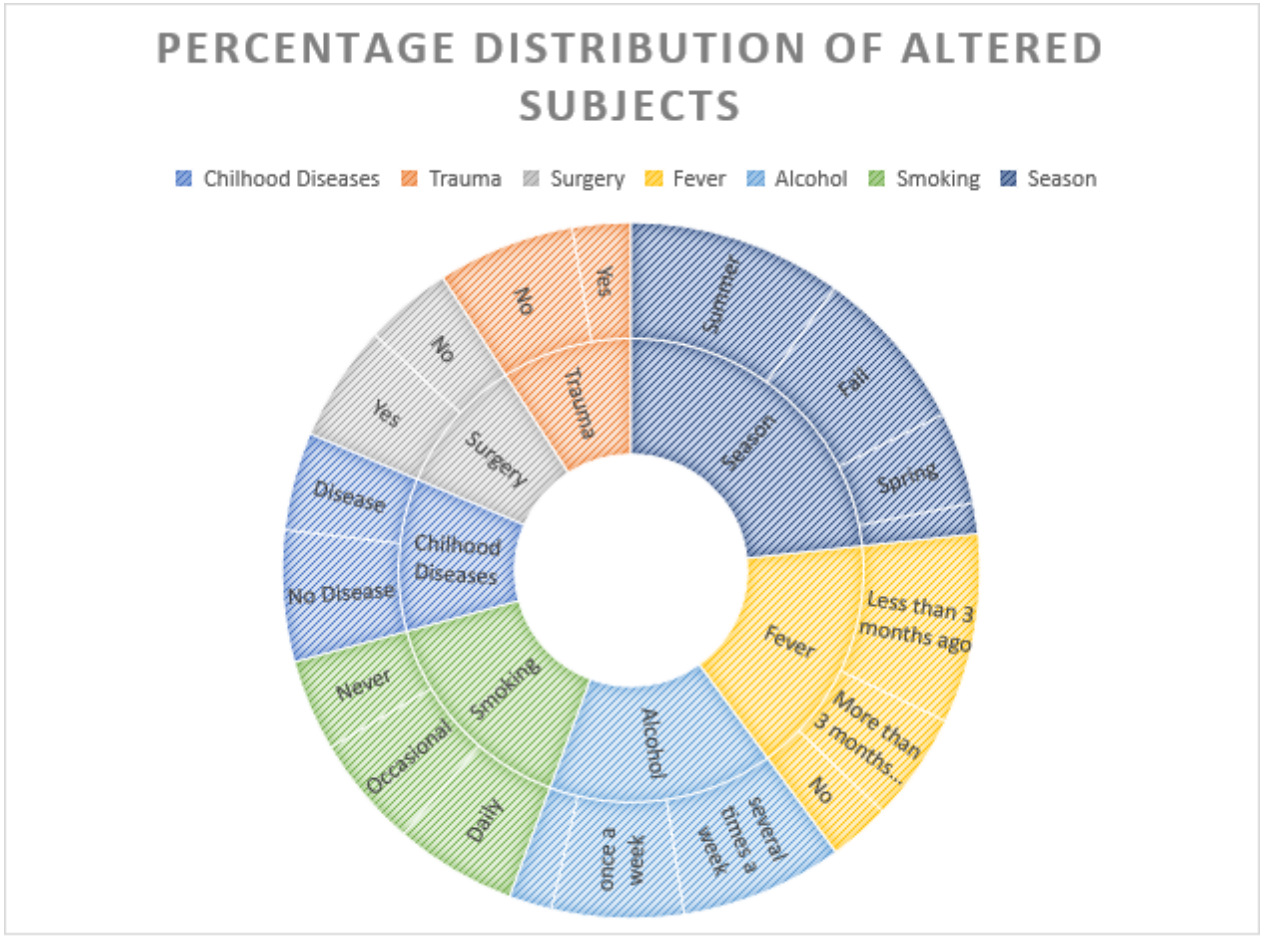
1. Season: Alteration is maximum during the summers and least during winters.
2. Fever: Those subjects who had fever in the last 3 months showed a remarkable decrease in their sperm count.
3. Alcohol: There was a stark negative effect of consumption of alcohol on male fertility.
4. Surgery: Subjects who had to undergo a major surgery in their lifetime had a significant effect on their reproducibility.
5. 

Figure 1. Sunburst diagram of population proportions

On the other hand, if we check the age of people who were altered vs who were normal, one can easily notice that the incidence of alteration increases with age. Most of the altered population lies in the age group of 30 and 36 years.

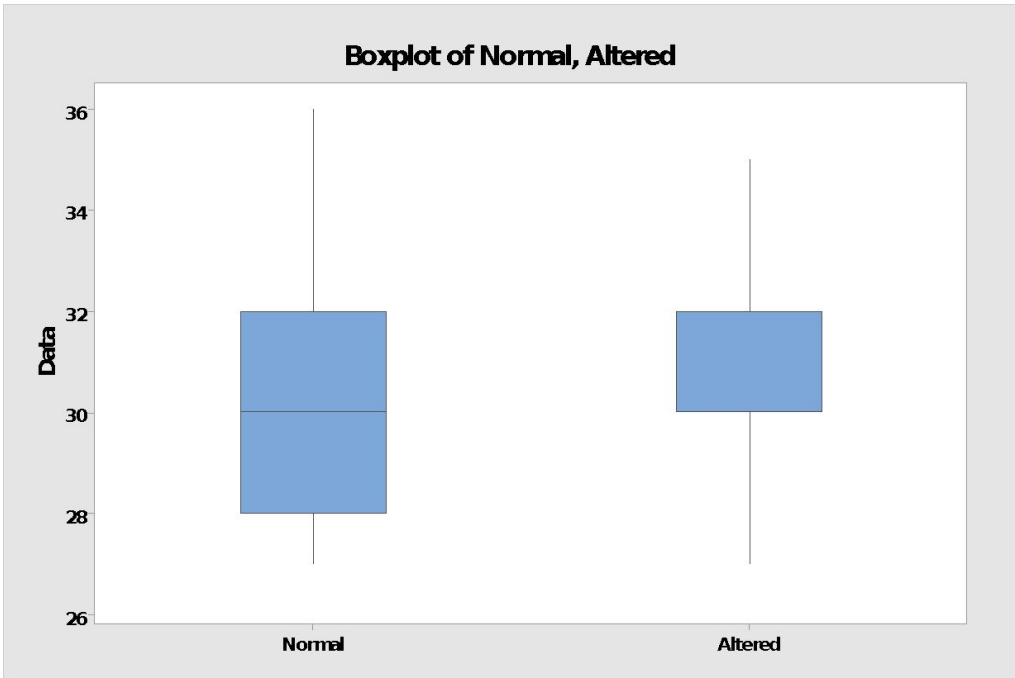


Figure 2. Boxplot of age in altered as well as normal population

The interval plot of sedentary habits illustrates that people who sat for longer number of hours had altered physiology.

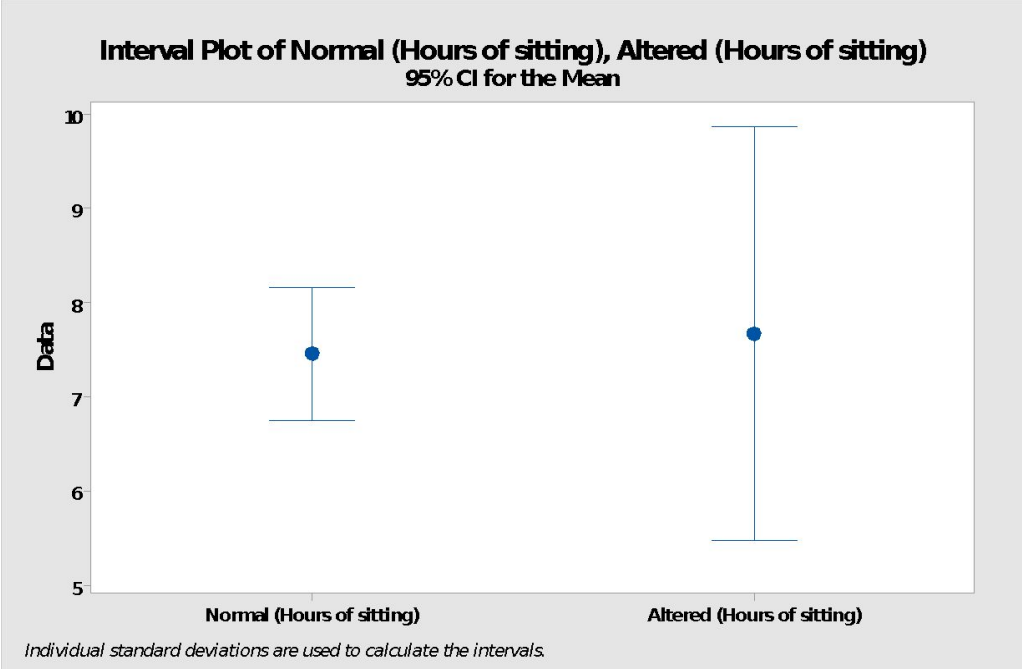


Figure 3. Interval of hours of sitting in altered and normal population

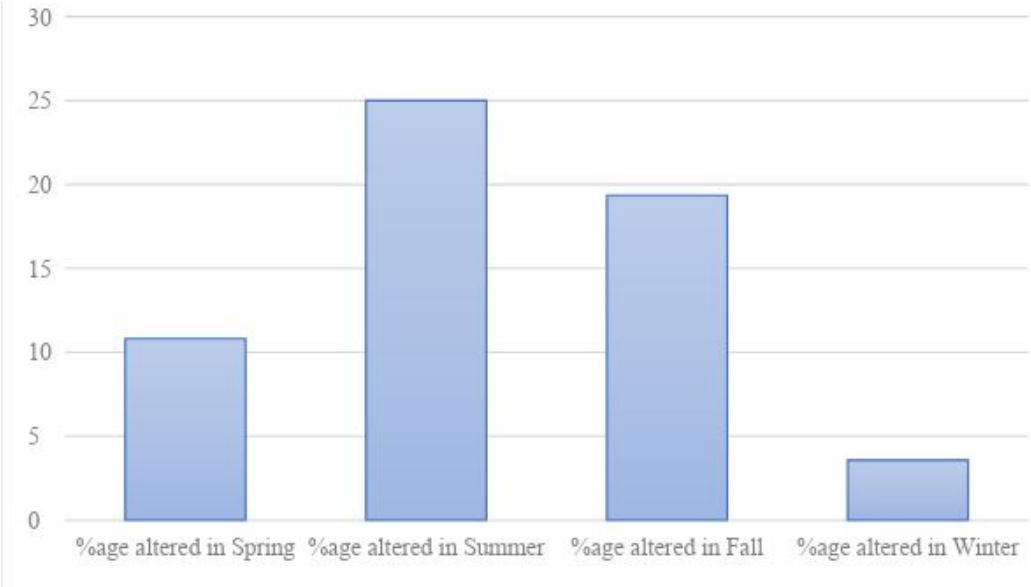


Figure 4. Seasonal variation of altered subjects

It is rather a stark finding that seasons have huge impact on male fertility. Effect was most pronounced in summer and fall seasons with winters having the least alteration.

Hypothesis Testing

Hypothesis testing is a method of statistics used for testing an assumption regarding a population parameter. It is a statistical inference tool for analysing the hypothesis performed on a sample data from a larger population.

The following two-sided hypothesis tests (unequal variances) were performed:

1. Comparison of proportion of smokers and alcoholics who have altered fertility rates

Using two population proportion test in R, we found that the null hypothesis can't be rejected. This means that there is no statistical significance as to which of the two factors - Smoking and Alcohol consumption have a higher effect on fertility rates (being diagnosed with altered fertility rates).

Smoking vs. Alcohol
$\hat{p}$ = 6/100 # smoking (0,1)
$\hat{q}$ =11/100 # alcohol (0.4, 0.6, 0.8, 1)
n_1 = 100
n_2= 100
$\hat{pq}$ = 17/200 # x_1+x_2 / n_1+n_2
$z = (\hat{p} - \hat{q})/\sqrt{((\hat{pq})*(1-\hat{pq})*((1/n_1)+(1/n_2))}$
print(z)
z = -1.267754
z_α = 1.645
Null hypothesis is not rejected.

2. Comparison of mean age of normal people and affected people

By performing t-test with the help of Minitab we found that there was no significant difference in the mean ages of people with normal fertility rates vs those with altered fertility rates.

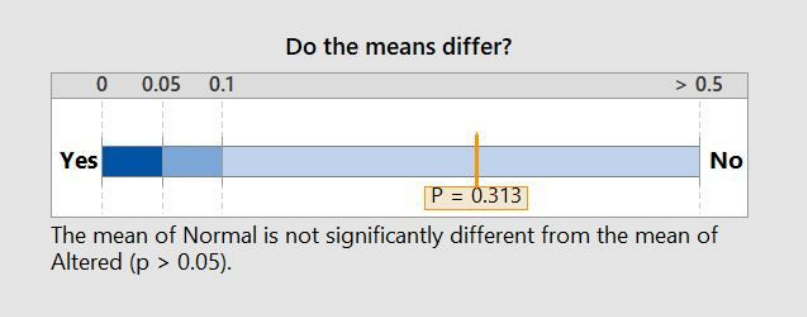


Figure 5. Hypothesis testing via the p-value approach

Individual Samples

Statistics	Normal	Altered
Sample size	88	12
Mean	30.034	30.667
95% CI	(29.55, 30.52)	(29.445, 31.888)
Standard deviation	2.2866	1.9228

Difference Between Samples

Statistics	*Difference
Difference	-0.63258
95% CI	(-1.9247, 0.65954)

*Difference = Normal - Altered

Figure 6. Hypothesis testing via confidence interval (CI) approach

3. Comparison of sedentary behavior of normal vs diagnosed people

Using Minitab, we found that there is not enough evidence to show that the mean number of sitting hours of normal vs diagnosed people differ.

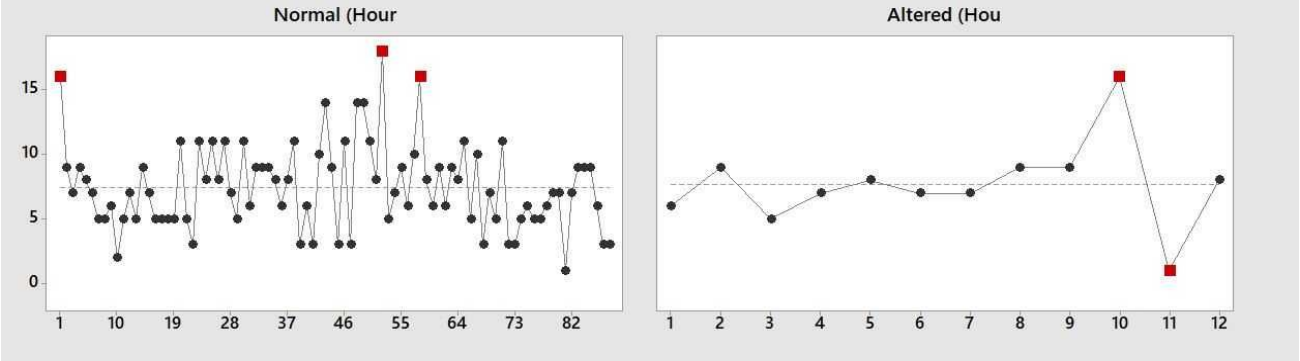


Figure 7. Plot of number of sitting hours of normal and diagnosed people

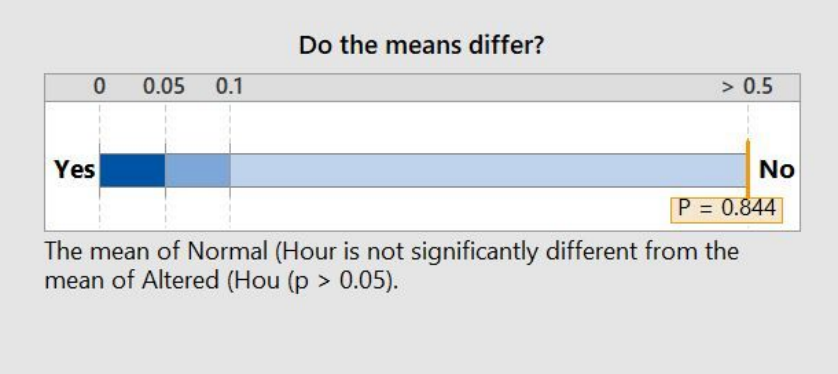


Figure 8. Hypothesis testing via p-value approach

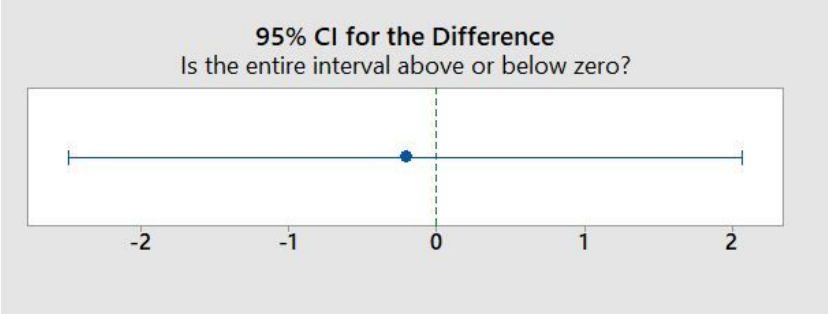


Figure 9. Hypothesis testing via CI approach

Individual Samples

Statistics	Normal (Hour	Altered (Hou
Sample size	88	12
Mean	7.4545	7.6667
95% CI	(6.753, 8.156)	(5.4768, 9.8565)
Standard deviation	3.3111	3.4466

Difference Between Samples

Statistics	*Difference
Difference	-0.21212
95% CI	(-2.4928, 2.0686)

Figure 10. Summary of CI approach

Logistic Regression

Conditions for Logistic Regression:

Logistic regression does not require a linear relationship between the dependent and independent variables. The error terms (residuals) do not need to be normally distributed. Homoscedasticity (where all random variables have the same finite variance) is not required. Finally, the dependent variable in logistic regression is not measured on an interval or ratio scale.

Logistic regression requires the observations to be independent of each other. The observations should not come from repeated measurements or matched data. It also requires there to be little or no multicollinearity among the independent variables. This means that the independent variables should not be too highly correlated with each other. Moreover, it also assumes linearity of independent variables and log odds.

Since our data has a mixture of categorical values and continuous values, logistic regression is the preferred model. The output in our dataset can take two values, 'N' if the diagnosis is normal and 'O', if altered. This model is used to estimate the probability of a binary response based on many predictors. With logistic regression, we can analyze the effect of each factor on the output.

In our dataset, we have some categorical variables that have more than two levels, and thus we would need to use dummy variables. This is done to prevent multicollinearity caused by including a dummy variable for every single category. We also have to explicitly state the intercepts to be used for our analysis.

Once we fit the model to our data using Python (StatsModels), we can get an overview of the model coefficients, how well the coefficients fit, and other statistical measures. Using this, we can also isolate and inspect parts of the model output. The confidence interval measures gives us an idea of how robust the model coefficients are. Further analysis using the odds ratio, tells us how a 1 unit increase or decrease in a variable affects the odds of being diagnosed with altered fertility.

Logistic regression does not have a direct equivalent to the R2 that is found in ordinary least squares (OLS) regression that represents the proportion of variance explained by the predictors. However, it is possible to use an analog, so-called a pseudo-R2, to mimic the OLS-R2 in evaluating the goodness-of-fit and to explain the variability. The pseudo-R2 simply compares the log-likelihood from the null model (only an intercept) to the log-likelihood from the full model (all covariates included).

From this perspective, the definition of McFadden R2 seems quite appropriate - the gold standard value of 1 corresponds to a situation where we can predict whether a given subject will be diagnosed as normal or altered with 100% accuracy. As an error metric, we use the pseudo-R2 value, which evaluates to 0.2487 for our model.

The table below shows the estimated coefficients for each independent variable (including the dummy variables) and their corresponding confidence intervals.

Logit Regression Results

Dep. Variable:	Output	No. Observations:	100
Model:	Logit	Df Residuals:	83
Method:	MLE	Df Model:	16
Date:	Sat, 21 Apr 2018	Pseudo R-squ.:	0.2487
Time:	15:25:02	Log-Likelihood:	-27.567
converged:	False	LL-Null:	-36.692
		LLR p-value:	0.3094

	coef	std err	z	P>|z|	[0.025	0.975]
Age	6.6181	3.891	1.701	0.089	-1.009	14.245
Diseases	0.5696	1.044	0.546	0.585	-1.477	2.616
Accident	-1.6410	0.857	-1.914	0.056	-3.321	0.039
Surgical	0.3057	0.798	0.383	0.702	-1.258	1.870
Sedentary	3.2801	2.417	1.357	0.175	-1.457	8.017
Season_-0.33	1.1442	1.401	0.817	0.414	-1.601	3.889
Season_0.33	2.4082	2.037	1.182	0.237	-1.584	6.401
Season_1.0	2.2449	1.375	1.632	0.103	-0.450	4.940
High Fever_0	-1.3633	1.113	-1.225	0.221	-3.544	0.818
High Fever_1	-1.8579	1.352	-1.374	0.170	-4.509	0.793
Alcohol_0.4	-3.0270	9.88e+05	-3.06e-06	1.000	-1.94e+06	1.94e+06
Alcohol_0.6	22.0911	6.67e+04	0.000	1.000	-1.31e+05	1.31e+05
Alcohol_0.8	21.7080	6.67e+04	0.000	1.000	-1.31e+05	1.31e+05
Alcohol_1.0	19.8649	6.67e+04	0.000	1.000	-1.31e+05	1.31e+05
Smoking Habit_0	-0.3467	1.006	-0.345	0.730	-2.319	1.625
Smoking Habit_1	0.3162	0.952	0.332	0.740	-1.550	2.183
intercept	-29.3731	6.67e+04	-0.000	1.000	-1.31e+05	1.31e+05

Figure 11. Coefficients of predictor variables (inclusive of their dummies)

Alcohol_0.6	3.926852e+09
Alcohol_0.8	2.677002e+09
Alcohol_1.0	4.238537e+08
Age	7.485254e+02
Sedentary	2.657767e+01
Season_0.33	1.111447e+01
Season_1.0	9.439345e+00
Season_-0.33	3.140021e+00
Diseases	1.767513e+00
Smoking Habit_1	1.371853e+00
Surgical	1.357511e+00
Smoking Habit_0	7.070439e-01
High Fever_0	2.558142e-01
Accident	1.937932e-01
High Fever_1	1.560045e-01
Alcohol_0.4	4.846101e-02
intercept	1.751509e-13

Figure 12. Odds Ratio of predictors

This table of odds ratios shows that the most decisive factor for deciding the diagnosis result is the **alcohol consumption rate**. This is **followed by the age factor**. It is seen that the fertility is

altered more for people in the aged group. This is expected, but this must be taken cautiously as we have seen in the introduction that the age bracket of the random sample under consideration has been skewed towards the aged.

The age factor is followed by the the number of hours spent sitting per day. As expected, the more the number of hours spent sitting, more is the altered fertility. Another main effect which we expected to contribute more towards fertility alteration is the smoking factor. Our results indicate that smoking has negligible effect on the fertility.
Caveat: with a dataset of just 100 points, it might not be correct to blindly follow the results shown. We would definitely need a bigger population for understanding the true factors.

Contingency Table:

Since we have identified alcohol as the primary cause for altered fertility, let us consider how the variation of levels of alcohol consumed play a role in fertility alteration. For this, we employ the use of contingency tables.

A contingency table is a multi-way table that describes a data set in which each observation belongs to one category for each of several variables. For example, if there are two variables, one with r levels and one with c levels, then we have a $r \times c$ contingency table.

The table below shows the contingency table for alcohol. Alcohol consumption has 5 levels and the output has 2 levels - 0 for normal and 1 for altered.

Alcohol	0.2	0.4	0.6	0.8	1.0
Output					
0	1	1	15	33	38
1	0	0	4	6	2

From this table, it is clear that if alcohol is consumed daily/ several times a week, it is not good for fertility, as expected.

CONCLUSION

Section 1 of the data analysis describes the use of descriptive statistics to do an initial assessment of the factors that are important in altering male physiology in terms of reproducibility. It shows that factors like trauma and childhood diseases has minimal or no effect. Whereas factors like alcohol consumption, seasons (indicating towards a direct relationship with temperature), age and sedentary lifestyle had a pertinent altering effect.

Hypothesis testing on the other hand had a contradicting evidence suggesting that mean age of people for altered as well as normal physiology was the same. This may be due to the fact that the samples of altered physiology were fewer and towards a higher age bracket (refer Figure 2). Hypothesis testing also refuted any evidence that sedentary lifestyle may affect the sperm count in subjects.

In order to obtain a conclusive argument, logistic regression was performed in section 3 of Data analysis. **The odds ratios confirm that alcohol consumption, age and sedentary lifestyle are the leading factors of alteration.**

REFERENCES

1. http://archive.ics.uci.edu/ml/datasets/Fertility?ref=datanews.io (Dataset link)
2. Priya, N. "Improve machine learning results for semen analysis using ensemble meta classification", *International Journal* 8.8 (2017).
3. Mendoza-Palechor, Fabio E., et al. "Fertility Analysis Method Based on Supervised and Unsupervised Data Mining Techniques." (2016).
4. Gil, David, et al. "Predicting seminal quality with artificial intelligence methods." *Expert Systems with Applications* 39.16 (2012): 12564-12573.
5. http://blog.yhat.com/posts/logistic-regression-and-python.html
6. https://www.nytimes.com/2009/08/17/health/research/17hepatitis.html?_r=1&scp=1&sq=Hepatitis%20C&st=cse
7. http://www.statisticssolutions.com/assumptions-of-logistic-regression/
8. http://www.b-g.k12.ky.us/userfiles/1525/Classes/33386/Conditions%20of%20Tests.08-09%20rev%201.pdf
9. http://www.jerrydallal.com/lhsp/ctab.htm
10. http://www.statsmodels.org/dev/contingency_tables.html
11. http://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/

APPENDIX

We include the Python codes that we have used for Logistic Regression and for generating the contingency table.

Code:

```
# import all relevant libraries
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
from sklearn.cross_validation import train_test_split
import statsmodels.api as sm
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# Setting up the dataframe for statistical inferences

df = pd.read_csv('fertility_Diagnosis.txt', header=None) # Defining the dataframe.
df.columns = ["Season", "Age", "Diseases", "Accident", "Surgical", "High Fever", "Alcohol", "Smoking Habit", "Sedentary", "Output"]
df['Output'] = np.where(df['Output'] == 'N',0,1)
y_df = df['Output']
y_df=y_df.to_frame()
del df['Output']
data2 = pd.get_dummies(df, columns =["Season", "High Fever", "Alcohol", "Smoking Habit"], drop_first = True) # to avoid collinearity while using dummy variables.
data2['intercept']=1.0 # Defining the intercept manually.

# Defining the logit model
logit=sm.Logit(y_df.astype(float),data2.astype(float))
result=logit.fit()
print(result.summary())

# Odds ratio.
print (np.exp(result.params))

# Contingency Table for alcohol levels.
tab = pd.crosstab(y_df['Output'], df['Alcohol'])
tab
```

We also used Minitab to generate the Hypothesis testing results.