Predicting Abnormal Fertility with a Mutual Information Approach

Jonah Chambers, Soufieh Hosseini & Manuel Nyagisere

May 19, 2021

1 SECTION 1

1.1 INTRODUCTION

Information theory is a study on how to quantify information, and is built around the concepts of mutual information and entropy. Entropy is defined as the rate at which information is generated by performing an experiment repeatedly. In other words, it is a measure of the randomness of a discrete random variable. Mutual information, in turn, is defined in terms of entropy and mutual entropy, and is the information that one random variable contains about another random variable.

For this report, our goal is to apply principles in mutual information as taught in class to reinforce our understanding on this subject. In particular, we measure the relationship between sperm concentration and health and lifestyles of adult males to uncover predictors of infertility in this population. We perform our analysis in Mathematica, and present our results and discussion here.

1.2 ORGANIZATION

This report is structured as follows: Section 1 introduces our application area. Section 2 describes the data set, our decision to apply mutual information and discusses our approach to the problem. Section 3 presents the results, and Section 4 analyzes and discusses these findings.

2 SECTION 2

2.1 DATA SET

The data set is retrieved from Kaggle, an online open data set repository (www.kaggle.com). It has 100 samples and 10 features with a diagnosis outcome variable – "normal or altered". The description of each variable is given in table 2.1.

Table 1: A Description of Variables in the Fertility Data Set

Season	Season analysis performed (Spring, Summer, Fall, Winter)
Age	Age at time of analysis
Childish disease	Yes/No (Chicken pox, Measles, Mumps, Polio)
Accident or Serious trauma	Yes/No
Surgical Intervention	Yes/No
Frequency of alcohol consumption	Several times a day, every day, several times a week, once a week, hardly ever or never
Smoking habit	Never, occasional, daily
Number of hours spent sitting per day	Hours
Diagnosis	Normal/Altered

2.2 SUMMARY STATISTICS

The fertility data set has 2 numerical variables, age and number of hours spent sitting per day. Table 2.2 gives the summary statistics for these features. None of these variables have missing values and there appears to be an extreme value in the latter variable.

Table 2: Summary Statistics of the Fertility Data Set

	Age	Hours Spent Sitting per Day
Min	27.0	1.0
1st Quartile	28.0	5.0
Median	30.0	7.0
Mean	30.11	10.8
3rd Quartile	32.0	9.0
Max	36	342

A complete matrix of correlations is also computed between each pair of variables to summarize the relationships between them. We see that most correlations are low and a good number are negative. A subset of the correlation matrix is given in figure 1.

3 SECTION 3

3.1 Method

The challenge now was how to gather information from the data set that can give insight into predicting abnormalities in male fertility. As previously mentioned the data set had 9 independent variables and one dependent variable which was

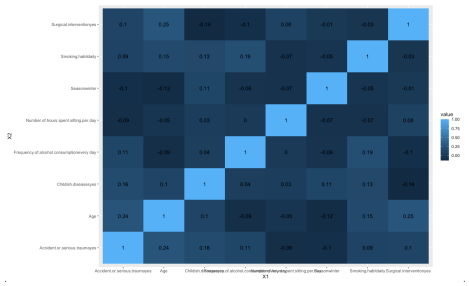

Figure 1: A Correlation Matrix for the Fertility Data Set

the diagnosis of the birth (see Table 1 for the complete list of variables). The goal of this research project was to apply an information theory topic to this challenge of predicting fertility to give insight into actions that can be taken as preventative health measures. To accomplish this task mutual information was used to assess the information of each variable. The independent variable with the highest amount of mutual information with the diagnosis variable would be the most correlated variable. That would mean that this variable would be the most useful variable to track for predicting a fathers chance at having altered fertility.

To accomplish this the previously mentioned dataset was downloaded from Kaggle. The first step in this data analysis was data preprocessing. Data preprocessing can involve removing entries without data, using averaged replacement methods or possibly having to convert strings to integers. In this data set there were no missing values, and all of the values were discrete variables. There was one value for hours spent sitting that had a value of 342 hours/day. Since this value is illogical it was replaced with the average value of the rest of the values, which was 8 hours/day. Next, the data file was uploaded into Mathematica where all of the data analysis and mutual information calculations were performed. Before performing any data analysis, the data should be visualized. Since there are 10 variables, 10 bar graphs would take up an unnecessary amount of space so data was visualized using 2 methods. Histograms were used for the integer valued data and tables were used for the non integer data. These can be seen in Figure 2 as well as Tables 3 and 4.

Table 3: Frequency of Yes/No Variables

Variable	Yes	No
Childish Disease	87	13
Accident	44	56
Surgical Intervention	51	49

After the data was visualized and the distribution of each variable was assessed the next step was to calculate the mutual information between each independent variable and the diagnosis which can be written as **I(X;Diagnosis)**.

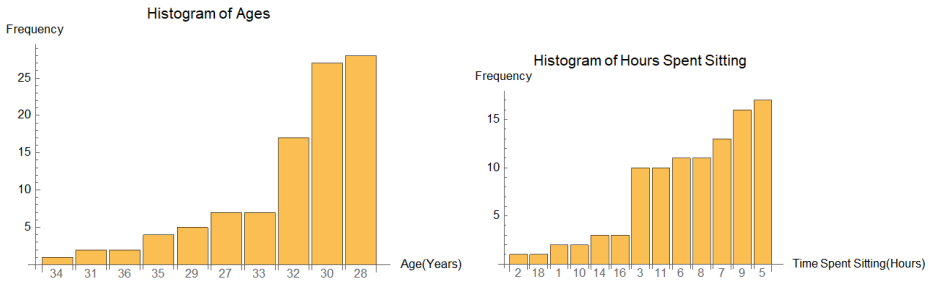

Figure 2: Histogram Age (Left) of Hours each Participant Spends Sitting on Average (Right)

Table 4: Frequency of Categorical Variables

Season	Summer 4	Winter 28
	Fall 31	Spring 37
Fever	More than 3 months 63	Less than 3 months 9
	No 28	
Alcohol Consumption	Several Times a Day 1	Every Day 1
	Once a Week 39	Hardly Ever/Never 40
	Several Times a Week 19	
Smoking habit	Daily 21	Occasional 23
	Never 56	

To calculate mutual information we need to know 3 values. The first is H(X), which is the entropy of the independent variable. This is done by first calculating the probability of each possible value, then summing the product of the individual probabilities times the log base 2 of the probability. This is better represented by the formula in Equation 1. This same process is repeated to calculate the second value we need to calculate mutual information which is H(Y), which is the H (Diagnosis in this project).

$$H(X) = -\Sigma_i p_i log(p_i) \qquad \text{Entropy of X} \qquad (1)$$

The final variable needed to calculate mutual information was H(XY) which is the joint entropy of the two variables combined. Joint entropy is better defined as the entropy of the product probability space, and can be calculated using equation 2.

$$H(XY) = -\Sigma_X \Sigma_Y p(X,Y) log p(X,Y) \qquad \text{Joint entropy of X and Y} \qquad (2)$$

All of these values were calculated in Mathematica using the code in the supplementary information. Once these values were calculated the mutual information could be calculated using Equation 3.

$$I(Y;X) = H(Y) - H(Y|X) \qquad (3)$$

$$= H(Y) + H(X) - H(YX) \qquad \text{Joint entropy of X and Y} \qquad (4)$$

This process was repeated for all 9 independent variables. Once the mutual information was calculated the final step was to assess which variables had the highest and lowest amounts of mutual information, as these would be the most informative variables for predicting fertility abnormalities.

4 SECTION 4

4.1 Results

After performing the mutual information analysis using Mathematica, we were able to output 9 results for mutual information of each independent attribute and the male fertility results. The results of the mutual information analysis are shown in table 3.1 below and are visualized in figure 4 below:

Table 5: Results from Mutual Information Analysis

Attribute	Mutual Information
Age	0.11828
Season	0.0320832
Childish Diseases	0.00109878
Accident/serious trauma	0.0151874
Surgical intervention	0.00212719
High fevers in the last year	0.0106989
Frequency of alcohol consumption	0.0321704
Smoking habit	0.00153191
Hours spent sitting/day	0.090645
Age and hours spent sitting/day	0.815208

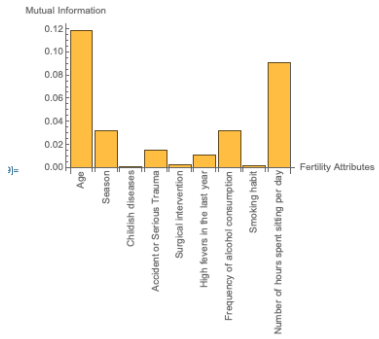

Figure 3: Visualization of Mutual information Results

From this analysis, it is easy to identify that age grants us the least level of uncertainty and the number of hours sitting per day comes close behind that. The lowest level of mutual information was from Childish Diseases, indicating

that there are the highest levels of uncertainty when it comes to how much this attribute could tell us about male fertility. The results of this analysis align well with the reality of male fertility. Based on existing research, it is known that age is a driving factor in male fertility: "Researchers believe the increased risk of health conditions might be due to random genetic mutations in sperm that occur more commonly in older men than in younger men.[1] It is also known that "A sedentary job doubled the risk of high levels of sperm DNA damage."[2] and could result in abnormal sperm. What is interesting from this analysis though is that in searching "male fertility risk factors" on google, you obtain countless articles such as the first result which is a Mayo Clinic article on "Male Infertility" in which smoking (which provided us with one of the least levels of uncertainty) was mentioned frequently but age and sitting were mentioned only 1 time [3]. This is a recurring trend among most of the results, indicating there could be specific emphasis placed on certain attributes, such as smoking, while the most significant attributes, such as sitting and age, are not highlighted as clearly. The results of this analysis ignite further probing questions and interests as well. For instance, probing deeper into the attribute of number of hours sitting, we can see that the actual act of sitting isn't directly resulting in abnormal sperm, but really the increase in temperature rise in the scrotum as a result of sitting is the driving factor that impacts sperm production[3] . We also checked the mutual information between age and number of hours spent sitting per day, which ended up being a high level of mutual information and thus a very low level of uncertainty between the two attributes.

4.2 Other Use Cases

There are many possible use cases for the results of this mutual information analysis. The most basic utilization of this analysis is in healthcare, where it is critical to understand the habits and attributes of both the mother and father to determine fertility and possible risks. Understanding that a potential father could be at higher risk if he is older or works at a sedentary life is critical to family planning.

Health insurance companies could also utilize this information to optimize the care and coverage of their customers. Some health insurance companies, in order to drive patient centric care, will track habits such as daily activity, to ensure customers are living healthy lifestyles. This same information could be used in conjunction with member age to determine those who are at increased risk of fertility abnormalities and could allow health insurance providers to social engineer changing customer habits. Another use case for this would be health insurance companies using this information to determine premium costs or plan selection. While it is illegal to change premium costs based on current health or medical history (due to the affordable care act), it is possible to change costs based on tobacco usage and age [4] , 2 of the 10 factors we analyzed for. Health insurance companies could use this level of decreased uncertainty with smoking habits as well as if they are in the typical age of starting a family to determine their plan policies and premiums.

4.3 Conclusion

The analysis of 10 male fertility attributes was performed using a mutual information approach and a data set from kaggle.com where 100 volunteers provide a semen sample analyzed according to the WHO 2010 criteria. This dataset listed 9 attributes and the final fertility of "Normal" or "Abnormal" of the semen. The attributes that provided us with the least level of uncertainty was age and number of hours sitting per day. This analysis and its results are critical in the healthcare industry to better understand fertility, in health insurance to provide patient centric care and optimize plan policies, and countless use cases we may not have even discovered yet.

5 References

[1] "Paternal age: How does it affect a baby? - Mayo Clinic." https://www.mayoclinic.org/healthy-lifestyle/getting-pregnant/expert-answers/paternal-age/faq-20057873. Accessed 10 May. 2021.

[2] "The impact of sedentary work on sperm nuclear DNA integrity." https://pubmed.ncbi.nlm.nih.gov/30869 Accessed 10 May. 2021.

[3] "Male infertility - Symptoms and causes - Mayo Clinic." https://www.mayoclinic.org/diseases-conditions/male-infertility/symptoms-causes/syc-20374773. Accessed 10 May. 2021.

[4] "How Health Insurance Marketplace® Plans Set Your Premiums ...." https://www.healthcare.gov/how-plans-set-your-premiums/. Accessed 10 May. 2021.