Human Cognitive Performance Data Analysis

Cachary Tolentino

Stockton University

April 12, 2025

# Abstract

Human cognition is a crucial portion of a well-functioning individual. Many factors play a role in the deterioration or nourishment of cognitive performance. Thus, it is important to understand and research what these factors are. In the growing age of technology, cognitive performance has been dramatically affected positively or negatively by constant screen time and media consumption (Stenberg et al., 2013). This research focuses on uncovering how habitual activities such as diet, exercising, and sleep may affect cognitive function. Furthermore, to analyze these issues, we approach the problem via performing statistical analysis on a real-world dataset, called “Human Cognitive Performance Analysis” by Samx\_sam from Kaggle, containing various information of individuals and their health statistics. Through this experiment, we can analyze and predict the possibility of cognitive performance as well as demonstrating the usage of statistical methods to learn about the dataset.

# Introduction

Human Cognitive Performance is a metric in evaluating the ability to process, understand, and respond to information. This metric varies from individual to individual primarily affected by their lifestyle. Lifestyles can greatly influence cognitive performance positively when monitored and balanced. But it can also deteriorate this performance.

A common problem in the modern age workforce is the damage in circadian rhythms due to overnight work (Chellappa et al., 2018). For example, according to Chellappa et al., “a cognitive slowing was observed under circadian misalignment with median reaction times of ~300 ms when assessed 11h after scheduled awakening” (2018). This shows that even during work, especially for some, cognitive performance can greatly influence daily life. This is crucial, especially for growing children, as media consumption and screen time have overall increased which can affect their learning and attention. Lifestyles are only one of the few factors that can deteriorate cognitive performance, many other factors play a role such as eating habits, exercise, and sleep.

Through this study, we will analyze various health information of multiple individuals collected as a large dataset. We will perform multiple statistical methods such as: histograms to visualize an overall distribution of health information, probability of having certain cognitive performance score given a condition, and much more. This study’s goal is to perform as many statistical methods as possible in various question formats to understand human cognitive performance and the effects of certain factors in determining their scores.

# Methods

[1.2] To understand the dataset, it is important to visualize what the data looks like. To do so constructing a relative frequency histogram allows us to quickly view how the data is distributed. For the dataset being used, we specifically focused on the percentage of occurrences for each specific column of data. This is possible via graphing the data in Excel using a pivot table and adding the selected column to the rows and values. Values will be change to “Count”. Once done, we can construct a pivot chart for our data. Finally, we can change how the value is presented as “% grand total”. This was done for all the 12 columns (ignoring the unique ID as it serves only as an identifier).

[1.3] Once the histograms for each column of the dataset were found, it was important to further analyze what these graphs are informing us. To do so we needed to find the mean median, mode, variance, and standard distribution for each graph. We did so via the built-in Excel functions (AVERAGE, MEDIAN, MODE. SNGL, VAR.P, and STDEV.P), categorical data will be ignored as Excel limits these functions to numerical data only.

[1.4] One area we focused on to understand how cognitive performance is affected is through an individual’s sleep duration. In the following experiment, we tried to predict the chances of an individual being sleep deprived. Sleep duration for individuals is approximately distributed with mean = 7.01 and standard deviation = 3.01. What fraction of all individuals would have sleep deprivation in the following intervals: 4.00 to 10.02 hours? 0.99 to 13.03 hours? 4.00 to 16.04 hours? Less than -2.02 or more than 16.04 hours? We can find these answers via simply finding the empirical rule and discerning the common standard deviations.

[2.1, 2.2, 2.3] Another way of constructing our understanding of the distribution of the data is through set notation. An individual has noted their gender, diet type, and exercise frequency. Let F denote female and M for male. We can denote NVG for non-vegetarians, VGN for vegetarians, and VEN for vegans. Then we can also denote L for low exercise frequency, M for medium, and H for high. Construct a space S. Find A, the following subset of possibilities containing no vegetarians, B the subset containing a female, and C, the subset containing a vegetarian. List the element of A, B, C, A B, A B, A C, A C, B C, B C, and C . We can find these answers by simply finding every combination of our data. Then performing the set operation given.

[2.4] An important aspect in statistical analysis would be learning what the probability of occurrences would be. In our dataset we can learn, for example, what the probability of an individual’s caffeine intake would be like. The proportion of coffee intake, 0-99, 100-199, 200-299, 300-399, and 400-499, in the population are approximately 20.231%, 19.903%, 19.871%, 20.000%, and 19.995%, respectively. A single individual is chosen at random from the population. List the sample space for this experiment. Assign probabilities for each of the simple events (ranges). What is the probability that the person chosen at random has either 0-99 or 200-299 mg of caffeine intake? The sample space will be the events/ranges that are given. Assuming each event has an equal probability to happen, we can assume that their number out of 100 will be their probability. Lastly, we can find the probability of events by adding their probabilities.

[2.5] In terms of probability, we can find certain information that could be useful for understanding the demographic that our population consists of. For example, a group of individuals contains six people. Two of the six are to be randomly selected to be classified as cognitively healthy or unhealthy. If two of the people are unhealthy, find the probability that at least one of the two people checked is unhealthy. Find the probability that both are unhealthy. If four of the people are unhealthy, find the probabilities indicated from before. To approach this problem, it would be best to find the sample space of all possibilities, then add the probabilities based on the criteria given.

[2.6] As the research in cognitive performance become more well known. More and more people are beginning to volunteer for research. Due to the influx of volunteers for the research, a raffle with 8000 tickets was handed, one per volunteer. There are only three positions needed to be filled. If four of the researchers were also given one ticket each, what is the probability that the four organizers will win all the prizes? Exactly two of the prizes? Exactly one of the prizes? None of the prizes? This problem can be solved via using the formula for combinations. Each combination changes based on the desired results.

[2.7] Although our population consists of primarily any individual older than a teenager, it is important to learn about how screen time have affected cognitive performance. For example, qn individuals are randomly selected from the population of 8000. If the first two individuals picked have a screen time duration of 7-9, what is the probability that the next three individuals will also have the same screen time? If the first three individuals have a screen time duration of 7-9, what is the probability of the next two people having the same screen time? If the first four people have a screen time duration of 7-9, what is the probability that the next person will also have the same screen time? This problem can be solved by using a mixture of combinations and conditional probability.

[2.8] To further learn more about the ages of volunteers in our population, we can construct a problem that concerns a mixture of probability and independence of events. For example, If A, being the age between 18 to 19, and B, being the age between 20 to 21, are such that . Find the following: , , and . We can solve this problem by using our understanding of additive and multiplicative law of probability.

[2.9, 2.10] An interesting information we can learn from the data is to compare how cognitive performances differ depending on the gender. We can construct a problem to learn more about this. For example, male (and others) and female were observed to have varying cognitive performance scores. It shows that 70% of females had high cognitive scores, 40% of males and others had high cognitive scores. A group of 20 people, 15 female, and 5 males, were subjected to a test to check their cognitive performance. A response picked at random from the 20 was negative. What is the probability that it was that of a male or other? This problem can be solved by using Baye’s theorem and Theorem of Total Probability.

[3.1, 3.2] A problem in the cognitive test was given to the volunteers. The task was to match three pictures of animals to the word identifying that animal. If a participant assigns the three words at random to the three pictures, find the probability distribution for Y, the number of correct matches. This problem can be solved using the probability distribution for a discrete random variable.

[3.3] As a follow up from the previous experiment, another test was conducted with more pictures to be matched. In the following information that was gathered, a problem can be asked to find the expected and variance of the number of correct matches. Let Y , the number of correct matches, be a random variable p(y), the probability of it occurring, given in the accompanying table. Find .

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| y | 1 | 2 | 3 | 4 |
| P(y) | .4 | .3 | .2 | .1 |

[3.4] In an experiment with the volunteers, coffee was studied to see any correlation with cognitive performance and caffeine intake. Two types of coffee were presented: formula A (with a much lower caffeine concentration), and formula B (with a higher caffeine concentration). Four participants were selected, each given three cups of coffee in random order. Two contains formula A and the other contains formula B. Each participant was asked which of the cups made them focus more. Suppose that the two formulas are equally performant. Let Y be the number of participants stating a preference for formula B. Find the probability distribution function for Y. What is the probability that at least three of the four participants will state a preference for formula B? Find the expected value of Y. Find the variance of Y. This problem can be solved using our knowledge of Binomial probability distribution.

[3.5] Another factor that may play into a deteriorate cognitive performance is through higher levels of stress. Stress is known for its adverse effects on general health. Here, we can learn about how stress may affect cognitive performance. For example, of the population of participants, 60% have high stress levels (8+). If a group of randomly selected participants is asked, what is the probability that exactly five people must be interviewed to encounter the first participant who has a high stress level? At least five people? This problem can be solved by using Geometric probability distribution.

[3.6] Alongside testing for cognitive function, memory plays just as big of a role in keeping high cognition. During analysis, it was found that 25% of participants had a memory score from 400 to 500. If you were to randomly ask a participant, what is the probability that they would have more or less of a memory score on the first try? The second try? The third try? If you wanted to speed up the analysis, you have yourself and your coworker ask a participant, what is the probability that a total of four tries will be necessary for the both of you to find someone not having a 400-500 memory score? This problem can be solved using Negative Binomial probability distribution.

[3.7] During the early phase of testing, data was recorded. At one point, ten participants were fully examined. Four have been found to be non-vegetarian. The researchers select five participant records. What is the probability that all five participants were vegetarian or vegan? This problem can be solved using Hypergeometric probability distribution.

[3.8] While interviewing the volunteers, it was found that the amount of sleep a participants get has a Poisson distribution with an average of seven hours per day. If more than seven hours were slept in a day, the participant scores a higher cognitive performance. What is the probability that a randomly selected participant will not score a higher cognitive performance? This problem can be solved using Poisson probability distribution.

[3.11] Amongst the 8000 participants, the average cognitive score was 60-70 with a standard deviation of 10. Using Tchebysheff’s theorem, find a lower bound for the number of participants from a 500-sample expected to have a cognitive score between 50 to 80.

[4.2] similarly to a previous experiment, Let Y, the number of correct matches, be a random variable p(y), the probability of it occurring, given in the accompanying table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| y | 1 | 2 | 3 | 4 |
| P(y) | .4 | .3 | .2 | .1 |

We would like to know the distribution function and the graph of this function.

[4.3] Following the previous example, if Y has a density function

We would like to find the mean and variance of Y.

[4.4] Luckily, the company was able to procure an AI algorithm that can predict cognitive performance scores given the data that was available. The AI predicted scores is uniformly distributed over the interval 70 to 80 points. What is the probability that the score exceeds 75 points if it is known that scores exceed 72 points? This can be solved by using the Uniform probability distribution.

[4.6] It is known that caffeine intake has some form of relationship with cognitive performance. A barista at a local store that many of the volunteers go to observed that during the mid-day hours there is an approximately exponential distribution with a mean 300 mg per coffee cup sold. We want to know what the probability of the demand for more caffeine will exceed 400 mg per coffee cup on any random day. In addition, what would be the amount of caffeine per cup should the barista maintain during any random day so that the demand will exceed the capacity is only .05? This problem can be solved using the Gamma probability distribution alongside the exponential distribution of a gamma function.

# Results

According to the constructed relative frequency histograms (found in the Figures section), we can see a diverse distribution of each category. In Figure 1, we can conclude that there is an almost even distribution of Men and Women who volunteer for this dataset. This gives us confidence that derived information will be applicable for all kinds of individuals. Even more so as Figure 4 shows us multiple participants across all age ranges from 18 to 59.

Some other notable figures to mention are Figures 2 to 8. These figures exhibit the distribution of other factors that may affect cognitive performance. Notably, with Figure 7 having a distinctive concentration of having a screen time of 9 to 11 hours. This may or may not have adverse effects on overall cognitive performance.

A graph of a graph

AI-generated content may be incorrect.

Figure 7: Screen Time Relative Frequency Histogram

To learn more about the dataset, we required to calculate the central tendency values of the numerical data. According to Figure 13, we can see these values such as mean, median, mode, variance, and standard deviation.

A table with numbers and a few black text

AI-generated content may be incorrect.

Figure 13: Central Tendency Data of Numerical Data

This data shows that from the dataset, the average participant was about 38 and a half years old. While the most common age being 40. Sleep duration shows that the average amount of hours slept is about 7 hours, while the most common being 3 hours. Stress levels average around 6.5 while the most common being 3. Daily screen time averages about 6.5 hours per day while 7.7 hours per day being the most common. Caffeine Intake averages 249 mg per day, while 76 mg per day being the most common. Reaction time averages 400 ms while 260 ms being the most common. Test scores, specifically, Memory scores show that the average score was 69.6, while 57 being the most common. Cognitive scores are shown to have the average score being 58.2 and 100 being the most common. Lastly, an AI model’s average prediction for cognitive scores of each individual shows to be 58.1 whole the most common being 100.

Although Figure 8 does not illustrate an approximately normal (bell shaped) distribution, to answer this problem it is still possible to use the empirical rule. Based on the empirical rule, one standard deviation would contain 68% of the measurements, two standard deviations would contain 95% of the measurements, and three standard deviations would contain 99.7% of the measurements. Therefore, we can calculate each standard deviation:

We can then find that the participants that have sleep deprivation between 4.00 to 10.02 hours is (4.00, 10.02) or one standard deviation, therefore 68% of the participants were sleep deprived. Similarly, sleep deprived participants between 0.99 to 13.03 hours is (0.99, 13.03), therefore 95% of participants are sleep deprived. For hours between 4.00 to 16.04 hours, we have 4.00 for the lower bound, while for the upper bound we have 16.04. This means that 81.5% of participants were sleep deprived. Finally, we can find the fraction of sleep deprived participants in the range of -2.02 to 16.04, which is three standard deviations, therefore 99.7% are sleep deprived. But since we are looking for those that are less than -2.02 or more than 16.04, we can perform the following:

This is possible since the entire distribution should be added to 100%, and we only desire those outside of 99.7%. Therefore, the fraction of sleep deprived participants with less than -2.02 or more than 16.04 hours of sleep is 0.003 or 0.3%.

Through set notation, we can group multiple characteristics of everyone. This can help in finding groups that may or may not fulfill certain conditions. To do so, we construct space s, which contains all the combinations of each characteristic.

Based on the given description, we can construct each set of notations that fulfill it. Therefore, A, containing no VGN will be:

Following a similar pattern, B containing a female will be:

Then we can also find C as:

Now that we have found A, B, and C, we can continue to create the sets for the following descriptions: A, B, C, A B, A B, A C, A C, B C, B C, and C .

fdbfdbdfbdbf

# Discussion

Note: Here is where you explain all the answers to from the above section. Make sure you tie in to the main point of the paper, maybe some limitation (if any) of the results.

Based on the results…

Note: [1.4] graph does not cover

# Conclusion

Note: Here is kinda just use it as a reflection of some kind. Like what did you learn overall, etc…

In conclusion….

# References

Chellappa, S. L., Morris, C. J., & Scheer, F. A. J. L. (2018). Daily circadian misalignment

impairs human cognitive performance task-dependently. *Scientific Reports,* 8(1), 3041–

11. <https://doi.org/10.1038/s41598-018-20707-4>

Samx\_sam. (2025). Human Cognitive Performance Analysis. *Kaggle.com.*

<https://www.kaggle.com/datasets/samxsam/human-cognitive-performance-analysis>

Sternberg, D. A., Ballard, K., Hardy, J. L., Katz, B., Doraiswamy, P. M., & Scanlon, M. (2013).

The largest human cognitive performance dataset reveals insights into the effects of

lifestyle factors and aging.*Frontiers in Human Neuroscience,*7.

<https://doi.org/10.3389/fnhum.2013.00292>

IMRAD (Introduction, Methods, Results and Discussion) | Science Writing Resources (new).

(n.d.). *Scwrl.ubc.ca.*<https://scwrl.ubc.ca/stem-writing-resources/features-of-academic-stem-research-writing/imrad/>

(2019). *Thevisualcommunicationguy.com*. [https://thevisualcommunicationguy.com/writing/how-](https://thevisualcommunicationguy.com/writing/how-to-organize-a-paper/how-to-organize-a-paper-the-imrad-format/)

[to-organize-a-paper/how-to-organize-a-paper-the-imrad-format/](https://thevisualcommunicationguy.com/writing/how-to-organize-a-paper/how-to-organize-a-paper-the-imrad-format/)

# Figures

A graph of a number of people

AI-generated content may be incorrect.

Figure 1: Gender Relative Frequency Histogram

A screenshot of a graph

AI-generated content may be incorrect.

Figure 2: Exercise Frequency Relative Frequency Histogram

A graph of a diet type distribution

AI-generated content may be incorrect.

Figure 3: Diet Type Relative Frequency Histogram

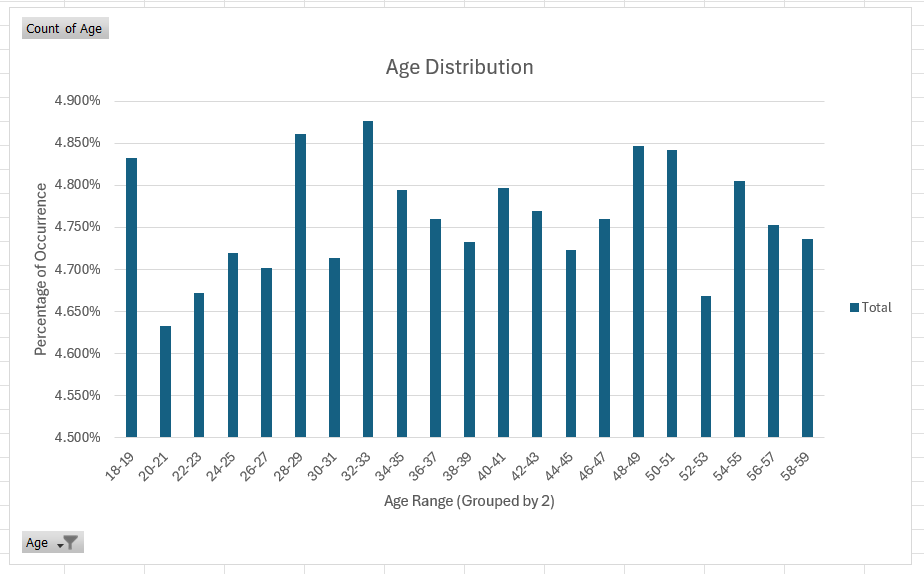


Figure 4: Age Relative Frequency Histogram

A graph of a sleep duration distribution

AI-generated content may be incorrect.

Figure 5: Sleep Duration Relative Frequency Histogram

A graph of a stress level distribution

AI-generated content may be incorrect.

Figure 6: Stress Level Relative Frequency Histogram

A graph of a graph

AI-generated content may be incorrect.

Figure 7: Screen Time Relative Frequency Histogram

A screenshot of a graph

AI-generated content may be incorrect.

Figure 8: Caffeine Intake Relative Frequency Histogram

A screenshot of a graph

AI-generated content may be incorrect.

Figure 9: Reaction Time Relative Frequency Histogram

A graph of a memory

AI-generated content may be incorrect.

Figure 10: memory Score Relative Frequency Histogram

A graph of a graph

AI-generated content may be incorrect.

Figure 11: Cognitive Score Relative Frequency Histogram

A graph of a bar chart

AI-generated content may be incorrect.

Figure 12: AI Prediction Relative Frequency Histogram

A table of numbers and a number

AI-generated content may be incorrect.

Figure 13: Central Tendency Data of Numerical Data