

Study on Factors that Influence Academic Performance

STA304 - Fall 2023 -Assignment 1

Yi Kuang 1006983187

2023-11-19

Part 1: Designing a survey

GOAL:

The act of learning is intrinsic to human nature. Every day, through various interactions, we engage in the process of learning to refine our thoughts and behaviour. Apparently, this continuous cycle of learning is not exclusive to formal educational settings; it permeates every facet of our lives. Given its universality, learning extends its relevance beyond the confines of academia. It possesses the capability to resonate with and benefit a wide range of subjects. For example, educational institutions might want to tailor their teaching methodologies by incorporating the ideas from this project. Also, individuals, particularly students, might want to optimize their learning strategy to maximize their outcomes by referring to the results of this project. Teachers and parents when supporting children could also discover a better technique to aid in their studying process.

As I delved into the research, through recent years, increasing studies on effective learning focused on the environmental factors that predict the student's academic performance while leaving the student-related input, for example, their study time allocation, rarely mentioned. This led me to believe that environmental factors do significantly influence students' academic performance, and study time is a weak indicator of a student's academic success as a premise in many studies. Upon further investigation, there is one widely-cited article unfolding on this topic confirming my belief (Plant et al., 2005). However, drawn from my past experience, those who achieve impressive academic success are those who devote a lot of time to studying. Certainly, there are also scattered studies that suggest a direct positive relationship between study time and academic grade (Ali et al., 2013). For this project, I attempted to examine and validate the common ground recent studies on effective learning have agreed upon. I will carry out a survey aiming at gathering representative students' data on study habits and performance outcomes to have a better sense that helps us conduct simulation. The data, combined with insights gained from other external sources and logical reasoning, will enhance the accuracy of the simulation. This simulation is crucial as it provides a comprehensive view for subsequent statistical analysis on factors counting towards the studying outcomes. Through this endeavour, the project aims to shed light on the relationship between study time and academic performance, while considering other pertinent factors that might influence academic success. By conducting this project, I hope to provoke some insights into the field of education and individuals who are eager to elevate their learning experiences.

PROCEDURE:

We initiated the process by crafting the survey. The primary objective was to devise questions that deepen our understanding of the study time allocated in relation to academic performance. Other than two non-intrusive demographic questions, which might offer valuable insight but are presumed to be less, the other four questions are tied to the goal of interest. One of the questions will surely be a measure of academic

performance, and I believe cGPA is a beneficial indicator, even if not perfect, as it encapsulates one's overall academic progress. Two other questions will focus on one's study duration and one's study timeframe. The last one would be a question about challenges when studying. Crafting all these questions demands meticulous attention to wording, so the respondents will feel at ease and respond in the most accurate manner. Beyond the contents, the presentation of the questions is not limited to the traditional ways of multiple choice or text entry, I also incorporated sliders to be more engaging. The disclaimer reassures the respondents about the confidentiality of their responses, in addition to that they feel free to quit the survey and skip the questions.

Afterwards, the survey will be carried out on a survey platform called Qualtrics, which allows for the distribution of surveys and easy collection of responses, ensuring participants can easily engage in the survey and the researchers can easily access the data (Qualtrics, 2020).

As the survey was carefully designed, the survey will be distributed to 10,000 undergraduate students at the University of Toronto (UofT) across faculties using a simple random sampling method. For the survey, our primary target population is all the undergraduate students at the UofT across faculties and years of study.

Our frame population is a list of all the UofT undergraduates who have access to the UofT email system. Our sample includes all my frame population who indeed filled out my online survey. We aim to gather 1,000 units of sample. By projecting a 10% response rate, we applied simple random sampling across faculties to ensure students from every faculty have the chance to be surveyed and distributed the survey to 10,000 students. The entire procedure has its strengths but for sure has its limitations.

Drawbacks:

- Non-response bias: Not all students completed the survey when they saw it. This potentially skews the results as not all perspectives are included.
- Limited demographics: Since the focus population is all undergraduates at UofT, the survey does not capture the nuances of all other institutions. Thus, it results in a skewed sample of data.
- Access to digital platforms: There is certainly a population who do not have access to the internet. This exclusion of this population might result in a non-representative sample.
- Over-reliance on cGPA: This singular focus might lead to an incomplete understanding of one's academic performance, because it obscures one's specific strength, given, for example, there are activities counted towards academia but are not counted as cGPA. Thus, it might lead to a biased result.

Strengths:

- Careful design of the survey: The questions are carefully designed and there is a smooth transition between the questions. This helps the respondents be more engaged while completing the survey, so there is a greater likelihood that a complete observation will be gathered.
- Sampling method: The simple sampling method I implemented ensures students from all faculties have a chance to represent their characteristics. This allows me to have a relatively less biased sample.
- Cost-efficiency: Thanks to the cost-free nature of the platform, I don't have to worry about implementing the survey with concern about the budget. This enables me to collect a huge sample size with not much worry.
- Convenience: The entire process of the survey, ranging from the design to the implementation, is straightforward and convenient and the questions are easy to fill out for the respondents. This makes the survey easily reproducible and accessible.

SHOWCASING THE SURVEY

CLICK ME :)

Highlighted Three Questions:

Q1. We're interested in understanding study habits in relation to academic performance. Could you please share your current Current cGPA (cumulative grade point average) in 4.0? (Your response will remain confidential and is only used for this project)

(The answer is a slider with a lower bound of 0 and an upper bound of 4, with one decimal place of the answer, and the participant can also choose not to respond)

Reasons for selection:

cGPA is a direct quantitative representation of their academic performance, which sets as our primary objective to observe in relation to other factors.

Benefits:

1. The disclaimer mitigates the respondent's concern to some level.
2. The respondents find it easy and efficient to enter their input, thanks to the slider.
3. The slider might be more interactive and engaging compared to the traditional response buttons.
4. This question standardizes the form of the academic grade, in a total of four and the cumulative grade point average.

Drawbacks:

1. The respondent might not accurately report his or her current cGPA because of the recall bias. This question presumes the respondents, however, are aware of their current cGPA.
 - Unavoidability: Some respondents might not have the habits to keep track of their grades; some might check after every assignment. This variability is hard to account for in this single question.
2. The respondents might find it sensitive to reveal their grades. Regardless of the disclaimer, with a sense of insecurity, they might still not respond to the answer or adjust their grades.
 - Unavoidability: Due to the intrinsic nature of this question, it is impossible to eliminate entirely the sensitivity even with the promise of the disclaimer. Also, the nature of self-esteem, combined with the insecurity of anonymity, leads some responses to be biased.
3. The respondents might have a grade with two decimal places, which does not fit our one decimal place precision slider.
 - Unavoidability: It is a trade-off between the ease of response and the precision of the answer. In this case, I weigh more on the ease and have a precision of one decimal place.

Q2. During which time of the day do you mostly study?

Daytime (6 am - 6 pm, broadly)

Nighttime (6 pm - 6 am, broadly)

Reasons for selection:

Study time depends not only on the length of the time but when the study takes place. This categorical response is crucial to further understanding the factors correlated with academic performance.

Benefits:

1. This binary question easily divides the respondents into two broad categories, which later might give us more insights into this difference in relation to our goal.
2. The respondents might find it easy to enter their input, given the dichotomous nature of the question.

Drawback:

1. There is vagueness for those who have a mixed pattern in studying (i.e. people might study sometime in the morning, while sometime at night. In this case, people might misjudge the most study time and therefore enter the biased data)
 - Unavoidability: The term ‘mostly’ is subjective. This subjectivity makes it hard to standardize the answer. However, providing a strict definition of ‘mostly’ and the nature of this question might over-complicate the question.
2. There might be seasonal and regional differences, given time varies on people’s perception. People’s perception of understanding 6 pm in summer is different from people who understand 6 pm in winter if we take people in Toronto as an example.
 - Unavoidability: Standardizing the options is a plausible approach to improve this question in the aspect of clarity, but the variation in season and region in bettering this question might be overwhelming to the respondents.
3. There are only two slots to choose from, which has lower precision in the response.
 - Unavoidability: There is a trade-off between precision and ease of response. Adding more time slots might overwhelm the respondents.

Q3. On average, how many hours per day do you dedicate to academic studying outside of regular class hours?

(The answer is a slider with a lower bound of 0 and an upper bound of 24, with one decimal place of the answer, and the participant can also choose not to respond)

Reason for selection:

The amount of time devoted to studying has a possible correlation with the study outcomes. The response is in a quantitative form, which is easy for further data analysis and correlated with further questions.

Benefit:

1. With consideration of it as a continuous answer but responding in a discrete manner – since the respondents estimate the response, the response is designed in a slider form, holding one decimal place of the answer.
2. The slider might be more interactive and engaging compared to the traditional response buttons.
3. The respondents do not have to enter their answers in predefined categories or ranges. The slider adds more flexibility.
4. It specifies the kind of studying, which reduces confusion on what specific kind, extracurricular or academic.
5. ‘On average’ reduces some confusion in answering the question, where the respondents might turn in an estimation.

Drawbacks:

1. The definition of ‘studying’ still differs among students, where those academic activities which do not count towards the manifestation of the academic performance, in this case, cGPA, will influence our results. The variation adds to a bias of the result.
 - Unavoidability: A comprehensive definition might be helpful to mitigate the problem, but will surely overcomplicate the question and overwhelm the respondents.
2. This question allows for an accurate response. However, the respondents might not have the precise answer in mind when answering.
 - Unavoidability: Some people have the habit of keeping track of their study time, while some might not. In this case, the variation accounts for the bias. Yet, the question of asking the context might help, which, however, overcomplicates the question.
3. There might be answers to 24-hour studying time, which might be a skewed observation. We lack the context to analyze this data apparently.
 - Unavoidability: It is unlikely to study 24 hours, but if the context is valid, it might be sensible. Additional context supplements might be helpful but will overcomplicate the survey.

Part 2: Data Analysis

The survey constructs an outline of data of interest. In this section, the data was collected based on later described reasoning and will be computed in later simulation.

Data

Survey Implementation

The survey was indeed carried out in a sample size of 63 after distributing the survey to 250 email recipients, surpassing our anticipated 10% response rate. The data from this survey largely helped me better understand the possible distribution of the population of the parameters of interest, primarily the questions regarding study time and studying barriers.

After briefly analyzing these 63 observations, I found there are 28 people spent more time at night studying, while 35 people tended to spend more time in the day studying. I would like to use this finding as the respective probability for my later simulation.

When asked about the primary study barriers:

18 respondents pointed to physical and mental health

24 respondents saw time management as their barrier

3 respondents regarded the barrier as the environment

8 respondents claimed it was the instructional quality

2 marked ‘other’

(However, a closer examination revealed that there were three participants who originally miscategorized their barriers, so I cleaned the discrepancy accordingly)

Here, I also represented the data in probability for the later simulation.

External Sources

The data from two demographic questions are scratched from the official UofT data analytics documents (Graduation, Employment and Government Loan Default Rates 2018 | Data & Analytics, 2018). For the year of study question, I found the data on graduation statistics, where there is an average of 77% graduation rate in 2018. This rate was calculated on those who did not graduate within a seven-year timeframe and I would like to include this piece of information for my later simulation. However, given a typical 4 years as the university timeframe from logical reasoning, I would like to put 50% on four years, each with a quarter weight, and the rest 50% will be put on five years or more.

Another document indicated the enrollment condition for each field of study in 2020. (Facts & Figures | Data & Analytics, 2021) I did some configuration on the categorization of the subjects according to my survey options.

Social Sciences and Related takes up 21,008, for **Social Sciences** in my survey

Math & Physical Sciences takes up 8,777, for **Natural Sciences** in my survey

Humanities and Related and **Fine & Applied Arts** take up 1,862 and 7,296, for **Humanities & Arts** in my survey

Engineering & Applied Sciences take up 6,758, for **Engineering** in my survey

Biological Sciences and Health Professions take up 4819 and 8618 respectively for **Life Science** in my survey.

General Arts and Science and **Education & Physical Education** take up 14,038 and 1,213 respectively for **Other** in my survey.

The data here was computed in the form of proportion for later simulation.

In addition, the data for another two questions in my survey are collected from a study on the topic of time management conducted in UofT throughout the 2017-2018 academic year, in which the characteristics of data inspired me to include in my simulation (Oreopoulos et al., 2022). In addition, with its large sample size, 3,581 students from different UofT campuses and overall detailed and sensible procedure of experiment conduction, I believed in the validity of the data and decided to employ the characteristics of the data.

During the regular week, the respondents studied a mean of 15.595 hours, with a standard deviation of 13.135 hours (Oreopoulos et al., 2022). This increased to a mean of 24.779 hours per week with a standard deviation of 17.922 hours per week. To get a clearer daily perspective, I propose averaging the mean and the standard deviation from both time periods and dividing the values by seven, given they are the same sample and the data is measured weekly. This might provide a more accurate representation of a daily study duration.

Meanwhile, the academic performance was measured by the GPA of the respondents, in which, throughout 2017-2018, the average GPA was 2.507 and the standard deviation was .996 (Oreopoulos et al., 2022).

I will use these characteristics of data in my later simulation.

Consideration on Distribution

Having all the necessary data in hand, I am also interested in the distribution of these traits or variables in my later simulation.

For the variables **year of study**, **field of study**, **study time**, and **study barrier**, I will implement random sampling, mimicking the sampling nature of a survey, which significantly reduces bias and increases the change of obtaining a representative sample for further accurate inference.

For the variable **academic performance**, I will perform a normal distribution and will put a lower bound at 1.5 and an upper bound at 4.0, since students lower than 1.5 are suggested to be suspended and, realistically

speaking, one cannot achieve a grade higher than 4.0 in UofT(Academic Standing, 2019). The setting of a normal distribution might be related to most of the setting of grade distribution and the paper also suggested so (Curving Scores With a Normal Distribution, 2023; Oreopoulos et al., 2022)

For the variable **study duration**, I will also perform a normal distribution and will put a lower bound at 0 and an upper bound at 18, since students practically speaking can allocate 18 hours to studying on a daily basis. It is normally distributed because the paper displays the distribution of the time spent during the exam period and of the time spent during the regular week, they all show a normal distribution with, however, different mean and variance.

Simulation

After gathering all the essential data, I plan to implement a simulation on R to have a holistic view of my study (R Core Team, 2023). Also, the following implementation of code cannot execute without certain features from the packages, **truncnorm**, **dplyr**, and **tibble** (Muller, 2023; Olaf Mersmann, 2023; Wickham, 2023). At the very beginning, I would like to ensure the same sample is drawn every time for better reproducibility by the **set.seed()** function with input 1 as the argument. The traits of my observations will be cohesive with the characteristics of the data I found in the previous part, and the simulation part is as follows,

id: Assign unique IDs to the respondents using **sample()** function, without replacement, from 1 to 1,000.

year of study: Simulate using **sample()** function with the potential values of the year of study, “1 year”, “2 year”, “3 year”, “4 year”, “5+ years”, and the sample size of 1,000.

field of study: Simulate using the **sample()** function of a sample size of 1,000 with the field of study, “Social Science”, “Natural Science”, “Humanities”, “Engineering”, “Life Sciences”, and “Other” with replacement. In addition, attach the probability to the corresponding field of study, 21008/74385, 8777/74385, (1862+7296)/74385, 6758/74385, (4819+8618)/74385, (14038+1213)/74385, respectively.

study time: Simulate using **sample()** function of a sample size of 1,000 with the time of study, “night”, “morning” with associated probabilities 28/63 and 35/63 respectively.

study duration: Simulate using the **rtruncnorm()** function from the **truncnorm** library. This function helps us set the lower boundary of 0 and an upper boundary of 18, as described previously, with a mean of $(15.595 + 24.779)/(27)$ and a standard deviation of $(13.135 + 17.922)/(27)$ on a sample size of 1,000.

study barrier: Simulate using the **sample()** function on the study barrier, “health”, “time_management”, “environment”, “instructional_quality”, “external_responsibilities”, and “Other” with associated probabilities 18/63, 24/63, 3/63, 8/63, 8/63, 2/63 on a sample size of 1,000, with replacement.

academic performance: Again, simulate using the **rtruncnorm()** function from the **truncnorm** library on the academic performance of 1,000 respondents with a lower bound of 1.5 and an upper bound of 4. The mean here is 2.507 and the standard deviation is 0.996.

After simulating all the variables, we combine them in a single data frame using the function **tibble()** from the package **tibble**, setting the stage for further analysis.

I will display the snippet of the data in the appendix.

Table 1. Numeric variables from the simulation

variable name	mean	median	variance
cGPA	2.652	2.652	.415
Study Duration	3.3268	3.0221	3.751126

From Table 1, there are two variables fitted in.

cGPA: The mean and median are the same, 2.656, with a low variance, .415. This suggests a symmetric distribution of cGPA among the respondents and that the cGPA of respondents is closely clustered around the mean.

Study Duration: The mean is 3.3268 and the median is 3.0221, with a variance, 3.751126. With the difference in mean and median, there is a potential skew in the data. Given the mean is greater than the median, it might be a positive skew. In addition, the variance is relatively large, suggesting there is a diversity in study duration. This sheds light on an intriguing correlation with cGPA, aligning with our goal of study.

Table 2. Categorical variables from the simulation

variable name	realization	proportion
Field of Study	Social Science	28%
Field of Study	Natural Science	12.4%
Field of Study	Humanities & Arts	11.8%
Field of Study	Engineering	9.0%
Field of Study	Life Sciences	17.5%
Field of Study	Other	21.3%
Study Barrier	Time Management	39.4%
Study Barrier	Environment	4.6%
Study Barrier	Instructional Quality	12.5%
Study Barrier	External Responsibilities	12.8%
Study Barrier	Physical & Mental Health	26.9%
Study Barrier	Other	3.8%
Year of Study	First Year	12.6%
Year of Study	Second Year	14.4%
Year of Study	Third Year	11.0%
Year of Study	Fourth Year	10.5%
Year of Study	Five Years or More	51.5%
Study Time	Night	53.6%
Study Time	Day	46.4%

From Table 2, there are four variables fitted in.

Field of Study: The most popular field of study is **Social Science** with 28% of respondents, followed by **Other** with 21.3%. The third most popular one is **Life Sciences** with 17.5%. The remaining fields of study are **Natural Science** with 12.4%, **Humanities & Arts** with 11.8%, and **Engineering** with 9.0%.

Here, **Social Science** is the most popular one at UofT, suggesting a strong interest of students in this field or possibly greater availability. The traditional STEM programs, **Natural Science** and **Engineering** in contrast, have the least enrollment suggesting the selectivity or also the limited availability nature of the programs. It is also noteworthy that the category, **Other**, ranked second in terms of enrollment. This suggests further decomposition of this category since it encompasses a lot more subjects and more is to be discovered.

Study Barrier: The most responded is **Time Management** with 39.4%, followed by **Health** with 26.9%. **Instructional Quality** and **External Responsibilities** were similar in response rate, with 12.5% and 12.8% respectively. Finally, the last one is **Other**, with a response rate of 3.8%.

Here, time management is a huge concern for most of the respondents, suggesting that UofT students might find the rigour of the academic courses. Followed by **Physical & Mental Health**, with 26.9%, which suggests UofT should invest more in medical and mental services on campus. The rest of the concerns suggest UofT could invest more in courses teaching students to balance out campus life and other commitments.

Year of Study: The respondents who are in their **First Year**, **Second Year**, **Third Year**, **Fourth Year**, comprise a similar proportion of the sample, with 12.6%, 14.4%, 11.0%, and 10.5%, and the respondents from **Five Years or More** of the academic study is the majority with 51.5%.

Here, the majority of the sample respondents are from the **Five Years or More** category, suggesting that most of the students would extend their academic progress and they might find academic courses at UofT challenging in a sense. This, compared to a similar pattern in the academic enrollment of other years, leads to interesting questions about why students extend their academic years.

Study Time: Students who prefer to study at night take up 53.6% and the students who study during the day take up 46.4%.

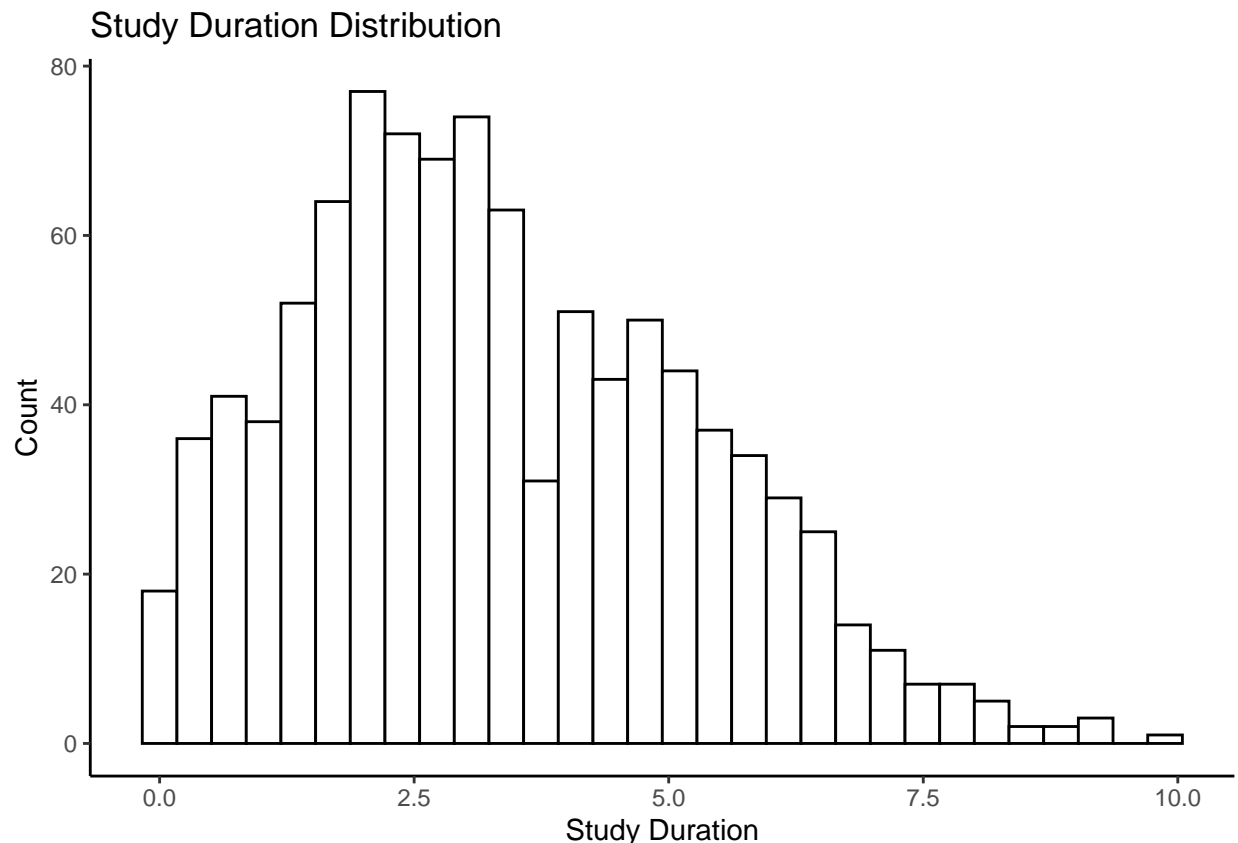
Here, the slight preference for nighttime study suggests that a lot of the students might be night owls. This piece of information helps the library schedule in a more thorough manner.

Plot

At the beginning, I would like to introduce the library `ggplot2` has the function `ggplot` that allows me to visualize the following data (Wickham, 2016).

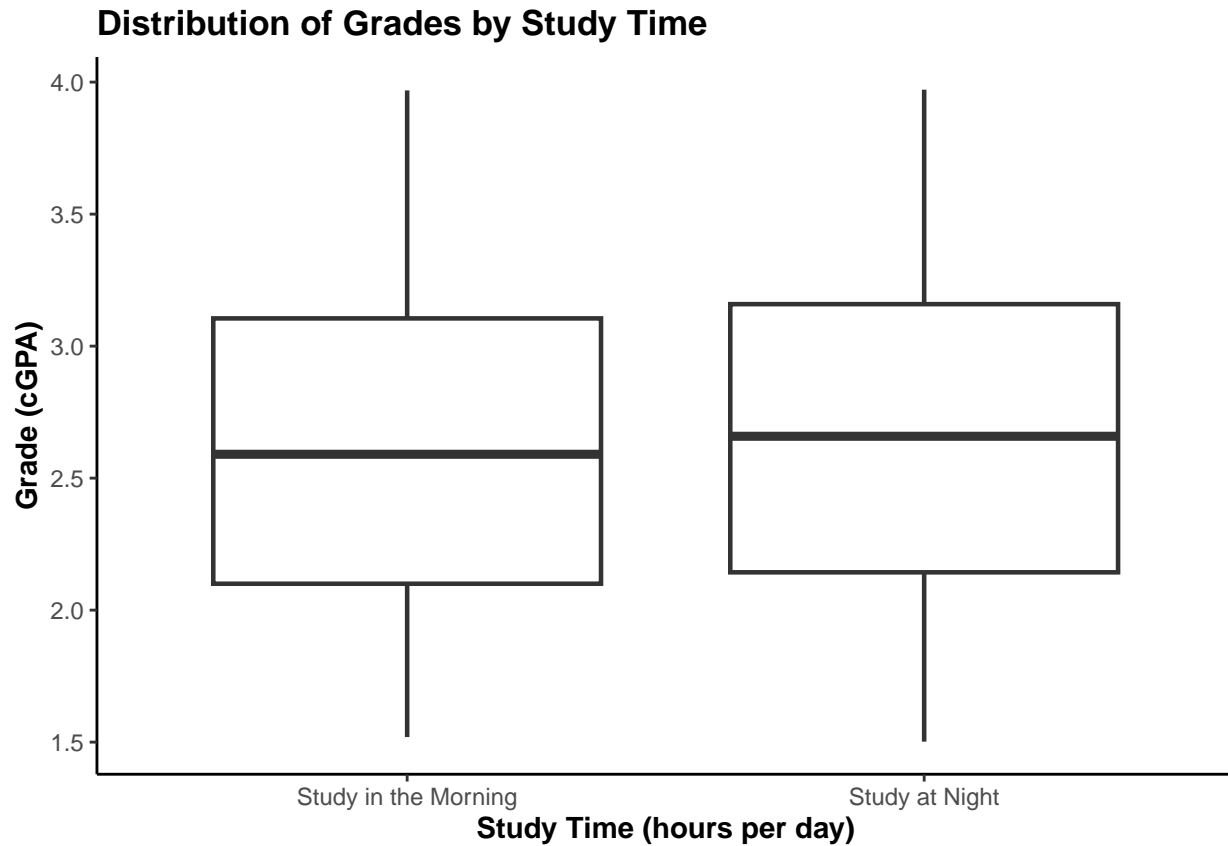
From the summary table, I predicted a possibility of the distribution of the variable, **study duration** and I would like to have a plot to verify my assumption.

Figure 1. A Histogram of the Distribution of the Sample Study Duration



This plot, which directly shows the distribution of the data, verifies my assumption on the trait of the distribution, where most of the data is clustered at 2.5 and it is right-skewed suggesting that most of the respondents have a lower study duration while a few respondents have a higher duration. Also, the spread is relatively wide suggesting that, as previously mentioned, it is worth exploring the potential of the variability.

Figure 2. A Side-by-side Boxplot of Those Who Prefer to Study during the Day and at Night in term of Academic Performance

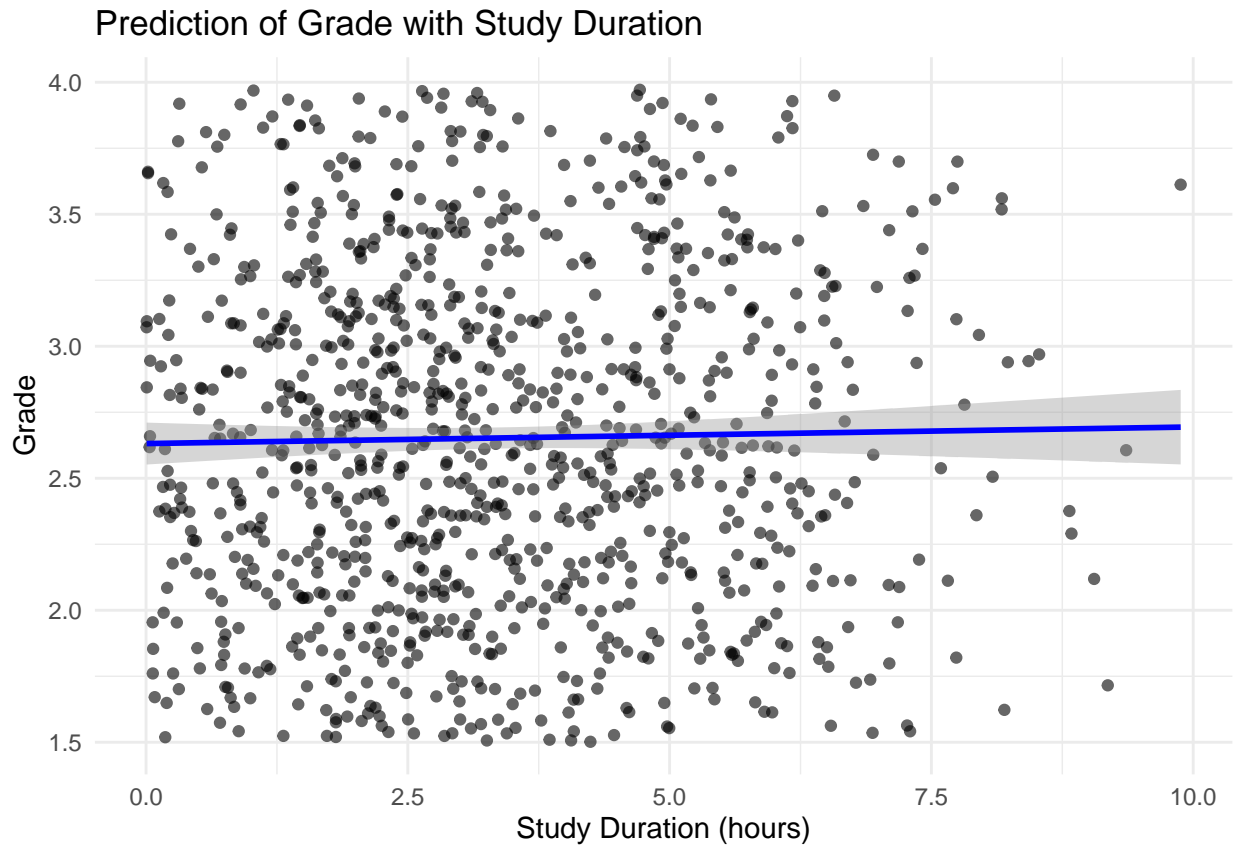


As suggested in the last figure, Figure 2 provides a direct visual comparison between two groups of respondents in terms of grade. The generic traits are quite similar. The box's height or the interquartile suggests similar data dispersion. Also, the whiskers, the line extending from the box, suggest similar data dispersion. However, it shows a slight difference in terms of the median, as indicated by the clear middle line, and possibly the mean.

Therefore, given the similarity in many traits, I would like to delve deeper. I propose to calculate the confidence interval on the academic performance of these two groups of people to have a clear sense of their academic performance.

It is also of interest to examine if there is a correlation between the study duration and academic performance or even if the study duration could possibly predict academic performance, as the goal of this study suggests. Therefore, a scatter plot and a linear model are depicted to check if there is a certain pattern on these two variables.

Figure 3. Scatter plot of Study Duration and Grade with a Fitted Linear Line



Each point represents a student's academic performance with respect to their study duration. Given the unclear nature of the distribution of the dots, since they all clustered evenly across the plot, we might take a reference on the linear model. However, the fitted linear regression line which directly indicates a linear relationship between the two variables suggests only a slight positive slope. This slightness triggers further statistical analysis on the significance in the difference. Therefore, later, a hypothesis test would be carried out to verify if there is any relationship between these two variables.

Method

At the very beginning, I wanted to acknowledge the methodologies later deployed, linear regression hypothesis test and student t confidence interval, from the course I took from STA238 at UofT (STA238H1 S LEC0101 20221:Probability, Statistics and Data Analysis II, n.d.).

Before executing our plan, it is crucial to understand the basic logic of hypothesis testing and confidence interval. They are both inferential statistics that help us gain a deeper understanding of the entire population based on a given sample.

Hypothesis Testing: This is to infer certain conditions for the entire population by validating the plausibility of an assumption based on the evidence we gained from a sample.

For this method, we first have to state our hypotheses, usually two. A null hypothesis (H_0) states there is no effect and an alternative hypothesis (H_1) which we want to prove.

Second, we calculate our test statistics, which is a standardized measure from our sample.

Third, we want to calculate the probability, which is the p-value, of our test statistics that possibly occur under the null hypothesis. A small p-value indicates that we have strong statistical evidence against the

null hypothesis, whereas a big one suggests we don't have enough evidence against the null hypothesis. To calculate the p-value, we might want to consider students' t-distribution. It is a distribution primarily used when we have a sample approximately normal and a small sample size or the population variance is unclear. Normally, since we don't have direct access to the population variance, sample variance has its own variability which is captured by the student's t distribution. Given its uncertainty in variance estimation, it is a more conservative approach compared to a normal distribution. Also, given its uncertainty in variance estimation, the shape of the distribution is determined by the degree of freedom, which represents how many values can vary independently in the calculations. The tail of the shape becomes heavier when the degree of freedom is small, and the shape approximates a normal curve when the degree of freedom becomes big.

Usually, the significance level is .05, meaning there is only to 5% chance that the test statistic in the given distribution, say in a student's t distribution, would occur, and, based on the result of our p-value against the null hypothesis, we want to make a decision to accept or reject the alternative hypothesis we want to prove.

This whole process enables us to justify if a test statistic we calculated from the sample drawn is statistically strong enough to prove against our null hypothesis. If the p-value of our test statistic is small, where the test statistic is unlikely to happen, we have enough evidence to reject the null hypothesis.

Confidence Interval: This is to infer, with a certain level of confidence, a probable range of a population parameter (a trait of the population which our sample is drawn from) that is expected to fall.

For this method, we are interested in the mean of the sample, the sample size, and the standard deviation, as we can have a range centred on the mean of the sample and the range is extended to a length described by margin error, which is determined by $t_{999,.95} \times \frac{sd}{\sqrt{n}}$. Here, t refers to a value from the student's t distribution with a 95% confidence level and our sample size. Because of the uncertainty in the variance estimate – where we don't know the population variance – the student's t distribution is more suitable compared to a normal distribution when we calculate the margin error. The margin error is the range in which we have 95% confidence in our true estimate of the population mean.

Task 1: To decipher if study duration could predict academic success

From the previous scatter plot, it is unclear if there is a linear relationship between these two variables. Therefore, by testing the coefficient of the predictor variable, **study duration** on a fitted linear model in relation to the response variable **grade**, we might have a clearer sense.

Firstly, we might want to fit a linear model on the two variables, **study duration** and **grade**, using `lm()` function with **study duration** as the predictor variable and **grade** as the response variable, so that we have an equation $\hat{Y} = \hat{\beta}_{studyduration} X_{studyduration} + \hat{\beta}_0$.

The null hypothesis (H_0) is that $\hat{\beta}_{studyduration} = 0$, suggesting it has no linear relationship to the response variable, and the alternative hypothesis (H_1) is that $\hat{\beta}_{studyduration} \neq 0$, suggesting it has a linear relationship to the response variable.

Later, a test statistic is calculated by $\hat{\beta}_{studyduration} - 0 / SE_b$, where $\hat{\beta}_{studyduration}$ is the coefficient estimate, 0 is the hypothesized value according to our null hypothesis and SE_b is the standard error of the coefficient estimate. The obtained ratio indicates how many standard error units away the coefficient is from zero and indicates how likely the population mean is from a sample mean (Bhandari, 2020).

The degree of freedom is our sample size minus the number of our parameters, which is one for the parameter in this case, given there are $\beta_{studyduration}$ and β_0 in our equation with a level of confidence of 95%. The corresponding p-value will be calculated with the previously calculated degree of freedom. By later comparing against our test statistics, we might have a final decision.

Task 2: To examine the confidence interval of the academic performance between the daytime studiers and the nighttime studiers

Since the sample size is large enough (greater than 30) and the observations are independent from each other, according to central limit theory, the shape of the distribution of academic performance of each group

is approximately normal and a t distribution, which will be utilized in constructing the confidence interval, is valid to apply. Also, with some inputs luckily calculated from the previous part and $t_{999,0.025}$ is easy to be computed in R, the access to the confidence interval, $\bar{x} \pm t_{999,0.025} \frac{s}{\sqrt{n}}$, becomes much easier, where \bar{x} is the mean of the variable, s is the sample standard deviation, and n is the sample size.

Result

After computation, the mentioned results described previously are as follows:

The Summary Tables of the Statistics of Interest 1

statistic	value
$\hat{\beta}_{studyduration}$	0.006294
<i>Standard Error</i>	0.010526
<i>Test Statistic</i>	0.598
<i>P-value</i>	0.55

The order of the statistics calculated reflects the order of the calculation process. It is clear that, from our analysis, $\hat{\beta}_{studyduration} = 0.006294$, which suggests that on average while keeping other variables constant, one more hour invested in the study will promote 0.006294 on cGPA. The validity of this explanation should be statistically validated by the hypothesis test. As can be seen, the p-value calculated is much greater than 0.05, suggesting that we do not have strong evidence against the $H_0 : \hat{\beta}_{studyduration} = 0$. Therefore, we do not have strong evidence to suggest the statistical significance of explaining the academic performance by study duration. While this might be the case for our study, it does not necessarily suggest studying more does not lead to an improvement in grades. It is highly plausible that other factors also play a role in academic performance and this should be noted on further analysis of academic performance.

The Summary Tables of the Statistics of Interest 2

statistic	value
<i>mean cGPA for the daytime studiers</i>	2.626846
<i>standard deviation of cGPA for the daytime studiers</i>	0.638204
<i>amount of daytime studier</i>	536
<i>mean cGPA for the nighttime studiers</i>	2.681824
<i>standard deviation of cGPA for the nighttime studiers</i>	0.650393
<i>amount of nighttime studier</i>	464
<i>t statistic for daytime studier</i>	1.964408
<i>t statistic for nighttime studier</i>	1.965101
<i>lower bound CI for daytime studier</i>	2.572694
<i>upper bound CI for daytime studier</i>	2.680997
<i>lower bound CI for nighttime studier</i>	2.622490
<i>upper bound CI for nighttime studier</i>	2.741157

The order of the statistic calculated also reflects the order of the calculation process.

The mean cGPA for those who prefer to study during the day is approximately 2.6268 with a standard deviation of 0.638204, while for the counterpart, the mean is roughly 2.6818 with a standard deviation 0.6504. The 95% confidence interval for the mean cGPA for those who prefer to study during the day ranges from 2.5727 to 2.6810, which means that we are 95% confident that the true mean cGPA for the daytime studiers at UofT falls within this range. Likewise, the 95% confidence interval for the mean cGPA for those who prefer to study at night ranges from 2.6225 to 2.7412. This means that we are 95% confident that the true mean cGPA for the nighttime studiers at UofT falls within this interval.

Upon closer examination, we found there is a slight overlap, suggesting the cGPA for the two groups do not have that much difference. Also, on average, the cGPA might be marginally higher for those who study at night, the difference to the counterpart is also not significant.

It should be noted that our sample comes with its inherent limitations, though I do my best to ensure its representativeness. For example, non-response bias, while I indeed distributed the survey through email, those who hypothetically provided more values into our study did not respond, leading to potential bias in our results.

In conclusion, the primary objective of this project was to assess the significance of the study time with respect to academic achievement, focusing on two variables: study duration and study period. However, these two aspects do not strongly suggest a relationship with academic performance. Though our study does not suggest so, it does not mean there is indeed no relationship. There are surely limitations to my study. If the condition permits, further research might investigate even deeper into the nuances of study habits in relation to academic progress. Certainly, more factors will be taken into consideration, such as physical and mental health, instructional qualities, and the environment as indicated by the response from the survey. As suggested by the respondents, these factors do have an impact on their quality of learning. Hopefully, in the future, subsequent studies will shed more light on this topic.

Part 3: Referencing

Generative AI Statement

I used ChatGPT Version available on Oct 5, 2023 (ChatGPT, n.d.). I used the tool primarily for my second part, with some paraphrasing concerns and inspirations on other parts as well. Given the long and complex nature of my interaction with ChatGPT on this project, I aim to provide a comprehensive overview of the interaction.

Though it seems the first part involves tons of help from AI, it was my time commitment instead. There are some phrases, however, that indeed get helped by ChatGPT for phrasing insights, however. In the second part, I devoted a lot of time to designing and troubleshooting the survey, so I prompted many questions related to suggestions on bettering the design and format of the survey. I revised my survey multiple times and ChatGPT only assisted with suggestions on better the words and ideas on making the delivery more engaging. Later, in the project, ChatGPT helped me with some ideas on the benefits and drawbacks of the questions in my survey. It helps me quite a lot on this. Moving on, in part two, ChatGPT helped me refresh my understanding of some statistical concepts, including students' t distribution, hypothesis testing and confidence interval. There are parts of the paragraphs rephrased or inspired by ChatGPT as well. For those who are interested in a deeper dive, I attached the entire conversation log in the appendix. Please feel free to check it out.

Bibliography

1. Academic Standing. (2019, January 28). Faculty of Arts & Science. <https://www.artsci.utoronto.ca/current/faculty-registrar/academic-standing>
2. Ali, S., Haider, Z., Munir, F., Khan, H., & Ahmed, A. (2013). Factors contributing to the students academic performance: A case study of Islamia University Sub-Campus. *American Journal of Educational Research*, 1(8), 283–289.
3. Bhandari, P. (2020, December 11). What Is Standard Error? | How to Calculate (Guide with Examples). Scribbr. <https://www.scribbr.com/statistics/standard-error/>
4. ChatGPT. (n.d.). Retrieved October 5, 2023, from <https://chat.openai.com>
5. Curving Scores With a Normal Distribution. (2023). <https://michaelminn.net/tutorials/normal-curve-grading/>
6. Facts & Figures | Data & Analytics. (2020). <https://data.utoronto.ca/facts-and-figures/>
7. Graduation, Employment and Government Loan Default Rates 2018 | Data & Analytics. (2018). <https://data.utoronto.ca/graduation-employment-and-government-loan-default-rates-2018/>
8. Muller, K. (2023). tibble: Simple Data Frame (3.2.1) [Computer software]. <https://tibble.tidyverse.org/>
9. Olaf Mersmann. (2023). Truncated Normal Distribution (1.0-9) [Computer software]. <https://github.com/olafmersmann/truncnorm>
10. Oreopoulos, P., Patterson, R. W., Petronijevic, U., & Pope, N. G. (2022). Low-Touch Attempts to Improve Time Management among Traditional and Online College Students. *Journal of Human Resources*, 57(1), 1–43. <https://doi.org/10.3368/jhr.57.1.0919-10426R1>
11. Plant, E. A., Ericsson, K. A., Hill, L., & Asberg, K. (2005). Why study time does not predict grade point average across college students: Implications of deliberate practice for academic performance. *Contemporary Educational Psychology*, 30(1), 96–116.
12. Qualtrics. (2020). Qualtrics XM - Experience Management Software (Oct, 2023) [Computer software]. Qualtrics. <https://www.qualtrics.com/>
13. R Core Team. (2023). R: A Language and Environment for Statistical Computing (Version 2021) [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
14. STA238H1 S LEC0101 20221:Probability, Statistics and Data Analysis II. (n.d.). Retrieved October 11, 2023, from <https://q.utoronto.ca/courses/253127>
15. Wickham, H. (2016). ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics (3.4.3) [Computer software]. <https://ggplot2.tidyverse.org>
16. Wickham, H. (2023). dplyr: A Grammar of Data Manipulation [Computer software]. <https://dplyr.tidyverse.org>

Appendix

The glimpse of the data simulated:

```
## Rows: 1,000
## Columns: 7
## $ id      <int> 836, 679, 129, 930, 509, 471, 299, 270, 978, 187, 307, ~
## $ year    <chr> "2 year", "4 year", "5+ years", "5+ years", "5+ years", ~
## $ program <chr> "Natural Science", "Humanities", "Other", "Humanities", ~
## $ study_duration <dbl> 0.1803289, 0.7578351, 1.6508998, 7.5909357, 2.4854624, ~
## $ study_time <chr> "morning", "night", "night", "night", "morning", "morni~
## $ study_barrier <chr> "time_management", "instructional_quality", "time_manag~
## $ grade     <dbl> 2.610532, 1.709624, 3.825409, 2.538039, 1.686681, 1.643~
```

The Full Interaction with ChatGPT for this Project:

User

five names that capture all academic disciplines

ChatGPT

Capturing all academic disciplines under just five names is a challenging task, as there are numerous f

1. ****Natural Sciences:**** This can include disciplines like physics, chemistry, biology, earth sciences,
2. ****Social Sciences & Humanities:**** This broad category can encompass disciplines like psychology, soc
3. ****Engineering & Technology:**** This would cover all engineering disciplines (mechanical, electrical, c
4. ****Life & Health Sciences:**** This would include medicine, nursing, pharmacy, dentistry, veterinary sc
5. ****Business, Economics & Law:**** This category would cover disciplines like business administration, f

It's important to note that these categories are quite broad and might still miss out on some niche dis

User

Six names

ChatGPT

Certainly! Let's attempt to categorize academic disciplines into six broad umbrella terms:

1. ****Natural & Physical Sciences:**** This can include disciplines like physics, chemistry, biology, earth
2. ****Social Sciences:**** This would cover disciplines such as psychology, sociology, anthropology, polit
3. ****Humanities & Arts:**** This category can encompass history, literature, philosophy, languages, visual
4. ****Engineering & Technology:**** This would cover all engineering disciplines (mechanical, electrical, c
5. ****Life & Health Sciences:**** This would include medicine, nursing, pharmacy, dentistry, veterinary sc
6. ****Business, Economics & Law:**** This category would cover disciplines like business administration, f

Again, these categories are broad and might still miss out on some niche disciplines, but they can serve

User

better formulate What aspects of life do you think largely influence your learning efficiency? in a sur

ChatGPT

Certainly! To make the question more specific and easier for respondents to answer in a survey, you might

****Which of the following aspects of your life do you believe most significantly influence your learning**

- [] ****Physical Health:**** Regular exercise, diet, sleep quality, etc.
- [] ****Mental Well-being:**** Stress levels, mental health conditions, etc.
- [] ****Study Environment:**** Quietness, comfort, availability of resources, etc.
- [] ****Time Management:**** Ability to prioritize tasks, allocate study time, etc.
- [] ****Social Interactions:**** Time spent with friends, family, or in extracurricular activities.
- [] ****Motivation & Interest:**** Personal interest in the subject matter or motivation to achieve certain goals.
- [] ****External Responsibilities:**** Work commitments, family duties, etc.
- [] ****Other:**** Please specify: _____

This format not only makes the question clearer but also provides respondents with a structured way to answer. User

capture all the factors in six terms

ChatGPT

Certainly! Here's an attempt to capture the factors influencing learning efficiency in six broad terms:

****Which of the following factors do you believe most significantly influence your learning efficiency?**

1. ****Physical & Mental Well-being:****
 - Includes aspects like regular exercise, diet, sleep quality, stress levels, and mental health conditions.
2. ****Study Environment & Resources:****
 - Encompasses the quietness, comfort, and availability of study materials and resources.
3. ****Time Management & Organization:****
 - Pertains to the ability to prioritize tasks, allocate study time, and manage one's schedule.
4. ****Social & External Commitments:****
 - Covers time spent with friends, family, extracurricular activities, work commitments, and family duties.
5. ****Motivation & Personal Interest:****
 - Relates to personal interest in the subject matter and motivation to achieve academic or personal goals.
6. ****Instructional Quality:****
 - Refers to the quality of teaching, clarity of instructions, and the effectiveness of the curriculum.

This format provides a concise yet comprehensive overview of the factors that might influence a person's learning efficiency. User

an approachable to formulate a question regarding cgpa in a survey

ChatGPT

Certainly! When asking about CGPA (Cumulative Grade Point Average) in a survey, it's essential to be clear

****We're interested in understanding study habits in relation to academic performance. Could you please**

- [] ****Below 2.0****
- [] ****2.0 - 2.49****
- [] ****2.5 - 2.99****
- [] ****3.0 - 3.49****
- [] ****3.5 - 4.0****
- [] ****Prefer not to say****

This format provides respondents with a range option, which might feel less invasive than asking for a
User

design me a warning before the survey question

ChatGPT

Certainly! Here's a warning or disclaimer that can be placed before the survey question about CGPA:

****Disclaimer:****

The upcoming question seeks information about your academic performance in the form of CGPA. Please know

This warning is designed to be respectful and reassuring, emphasizing the confidentiality of the response
User

paraphrase: This survey will take less than a minute.

Disclaimer: The upcoming question seeks information about your academic performance in the form of CGPA
ChatGPT

Certainly! Here's a paraphrased version:

This survey can be completed in under a minute.

Note: The next question will ask about your CGPA as a measure of academic achievement. Rest assured, your

User

paraphrase: The survey was carried out on a survey platform called Qualtrics, which allows for easy dis

For the survey, our primary target population is all the undergraduate students at the University of Tor

Our frame population is all the UofT undergraduates who either have access to the online platforms (Mes

Our sample population includes all my frame population who actually filled out my online survey, where

Drawbacks:

Selection Bias: although I did my best to hand out the survey as randomly as possible without any consc

Peer influence: the results might be influenced by peers, as in the library setting, there are people c

Non-response bias: Not all students completed the survey when they saw it. In this case, I might miss th

Time bias: it is obvious that there is a limited amount of time allowing me to collect every possible s

Strength:

Mixed Distribution Method: I introduced my surveys through both offline and online methods, which increa

Broad Target: I allowed every undergraduate student at UofT to complete my survey which captured the fu

Convenience and Cost-efficiency: I collected my 49 samples within a day and the entire process was conv
ChatGPT

Certainly! Here's a paraphrased version:

The survey was conducted using the Qualtrics platform, known for its user-friendly interface for both d

Our main focus for the survey was the entire undergraduate student body at the University of Toronto (U

The broader group we aimed to reach consisted of UofT undergraduates who either were active on online p

From this group, our actual respondents, or sample population, were those who took the time to complete

****Potential Shortcomings**:**

- ****Selection Bias**:** Despite efforts to distribute the survey impartially, there's potential for bias,
- ****Influence from Peers**:** Given that some participants in the library were in groups and might have c
- ****Non-response Bias**:** The fact that not every student who saw the survey completed it means some per
- ****Temporal Bias**:** The short window for data collection in the library might have favored early riser

****Strengths**:**

- ****Diverse Distribution Channels**:** By leveraging both digital and in-person methods, the chances of o
- ****Inclusive Approach**:** By inviting all UofT undergraduates to participate, the survey aimed for a co
- ****Efficiency and Cost-effectiveness**:** Gathering 49 responses in a single day was both time-efficient

User

some additional benefits?

ChatGPT

Of course! Let's explore some additional benefits of the second question:

****Question**:**

****During which time of the day do you usually study? Daytime (6am - 6pm) Nighttime (6pm - 6am)****

****Additional Benefits**:**

- **Simplicity**: The dichotomous nature of the question (Daytime vs. Nighttime) makes it quick and easy.
- **Broad Insights**: By categorizing study times into two main periods, the question can provide a general overview.
- **Flexibility for Analysis**: The binary response format allows for various analytical approaches, from simple percentages to more complex statistical models.
- **Potential for Further Research**: The responses can serve as a foundation for more detailed follow-up studies.
- **Adaptability**: The question's format can be easily adapted or expanded for different populations or contexts.

These additional benefits highlight the strengths of the question in terms of its design, potential for user engagement, and the insights it can provide.

why the drawbacks mentioned for the three questions are unavoidable
ChatGPT

Certainly! Let's delve into why the drawbacks for each of the three questions might be unavoidable:

1. Question on cGPA

- **Sensitivity**: Some students might be uncomfortable sharing their cGPA.
- **Unavoidability**: It's inherent in the nature of the question. Academic performance is personal, and students may not feel comfortable disclosing it.
- **Recall Bias**: Some students might not remember their exact cGPA.
- **Unavoidability**: While you could provide a range instead of asking for an exact number, this might not be as precise.

2. Question on Study Time

- **Ambiguity in "Usually"**: The term "usually" can be subjective.
- **Unavoidability**: Providing a strict definition might overcomplicate the question. Keeping it broad allows for more natural responses.
- **Lack of Specific Time Slots**: The question broadly categorizes the day.
- **Unavoidability**: Adding more time slots could overwhelm respondents and make the question more complex.
- **Variability in Study Sessions**: Some students might study during both daytime and nighttime.
- **Unavoidability**: Including too many options (e.g., "Both daytime and nighttime", "Varies", etc.) might lead to inconsistent responses.
- **Cultural or Seasonal Differences**: The definition of "Daytime" and "Nighttime" might vary.
- **Unavoidability**: Standardizing the times is necessary for consistency. Accommodating every cultural or seasonal variation would make the question too cumbersome.

3. Question on Study Duration

- **Varying Definitions of "Studying"**: What constitutes "studying" might differ among students.
- **Unavoidability**: Providing a detailed definition might overcomplicate the question. Keeping it general allows for more natural responses.
- **Weekend Study Habits**: The question focuses on weekdays.
- **Unavoidability**: Including weekends might introduce more variability due to differing weekend routines.

In essence, while these drawbacks exist, they arise from the need to balance precision, simplicity, and user engagement. The goal is to create a question that provides useful insights while being easy and comfortable for students to answer.

what if the drawback for the first question is The respondents might find it sensitive to reveal their grades?
ChatGPT

Certainly! Let's address the drawback you've mentioned for the first question and explain why it might be a concern.

****1. Question on cGPA**:**

- ****Sensitivity and Potential Inaccuracy**:** Respondents might find it sensitive to reveal their grades.
- ****Unavoidability**:**
 - ****Inherent Sensitivity**:** Academic performance is inherently a sensitive and personal topic for many students.
 - ****Trust in Anonymity**:** While the disclaimer emphasizes confidentiality, building complete trust in the survey process can be challenging.
 - ****Trade-off Between Honesty and Participation**:** If the question were made optional to address sensitivity, it might lead to less accurate data.
 - ****Alternative Approaches and Their Limitations**:** Using grade ranges (e.g., 3.0-3.5) instead of exact cGPA values might be less sensitive but could also reduce the precision of the data.

The drawback you've highlighted is a valid concern in survey research, especially when dealing with personal information like grades.
User

what if the drawback is that it's hard to recall the cGPA?

ChatGPT

If the drawback for the first question is that respondents might find it challenging to recall their exact cGPA, here's how to address it:

****1. Question on cGPA**:**

- ****Difficulty in Recalling Exact cGPA**:** Some students might not remember their precise cGPA, especially if it's been a while since they checked.
- ****Unavoidability**:**
 - ****Trade-off Between Precision and Ease**:** Asking for an exact cGPA provides more precise data, but it might be more burdensome for respondents.
 - ****Variability in Grade Checking Habits**:** Students have different habits when it comes to checking their grades, leading to inconsistent recall.
 - ****Potential for Estimation**:** Even if students can't recall their exact cGPA, they might provide an estimated value, which could affect the accuracy of the data.
 - ****Alternative Approaches and Their Limitations**:** Providing a "Not Sure" option might accommodate those who can't recall, but it would also result in missing data for that specific question.

The challenge of recalling exact figures, especially for something that might change periodically like cGPA, is a common issue in survey research.
User

what if I formulate one question as the following, and what are the drawbacks and unavoidability of the question?
ChatGPT

Of course! Please provide the formulation of the question you'd like to be analyzed, and I'll identify the potential drawbacks and unavoidability.
User

On average, how many hours per day do you dedicate to studying outside of regular class hours?

Less than 1 hour

1-2 hours

2-3 hours

3-4 hours

5-6 hours

6-7 hours

8+ hours

ChatGPT

Certainly! Let's analyze the provided question:

****Question**:**

On average, how many hours per day do you dedicate to studying outside of regular class hours?

Options:

- Less than 1 hour
- 1-2 hours
- 2-3 hours
- 3-4 hours
- 5-6 hours
- 6-7 hours
- 8+ hours

****Drawbacks**:**

1. ****Overlapping Timeframes**:** The options "3-4 hours" and "5-6 hours" skip the "4-5 hours" range. Responses in this range would be lost.
2. ****Varying Definitions of "Studying"**:** What constitutes "studying" might differ among students. For some, it might include all academic work, while for others, it might only refer to formal study sessions.
3. ****Recall and Estimation Bias**:** Some students might not track their study hours closely and could provide inaccurate estimates.
4. ****Potential for Extreme Responses**:** The "8+ hours" option is open-ended and can encompass a wide range of study durations, potentially skewing the average.
5. ****Exclusion of Weekend Study Habits**:** The question focuses on "per day" but doesn't specify whether it includes weekends.

****Unavoidability of the Drawbacks**:**

1. ****Overlapping Timeframes**:** While the omission of "4-5 hours" seems like an oversight, sometimes surveys use non-overlapping ranges to avoid confusion.
2. ****Varying Definitions of "Studying"**:** Providing a detailed definition might overcomplicate the question and lead to inconsistent interpretations.
3. ****Recall and Estimation Bias**:** It's a common challenge with self-reported data, especially when asking for specific time durations.
4. ****Potential for Extreme Responses**:** While "8+ hours" is a catch-all for longer study durations, breaking it down further might be too granular.
5. ****Exclusion of Weekend Study Habits**:** Specifying "weekdays" or "including weekends" could clarify the scope of the question.

In essence, while the question has its drawbacks, many of them arise from the inherent challenges of self-reported data collection. User

On average, how many hours per day do you dedicate to studying outside of regular class hours?

(The answer is a slider with a lower bound of 0 and an upper bound of 24, with one decimal place of precision.) ChatGPT

Alright, let's analyze the reformulated question with the slider response format:

****Question**:**

On average, how many hours per day do you dedicate to studying outside of regular class hours?

Response Format:

A slider ranging from 0 to 24 hours, allowing for one decimal place precision. Participants also have the option to select "less than 1 hour" or "more than 24 hours".

****Drawbacks**:**

1. ****Precision vs. Accuracy**:** While the slider allows for precise responses (up to one decimal place),
2. ****Potential for Misuse**:** With a slider, there's a possibility that respondents might accidentally s
3. ****Varying Definitions of "Studying"**:** As with the previous format, the term "studying" might be int
4. ****Overwhelming Range**:** The full 0-24 hour range might be overwhelming for some respondents, especial
5. ****Lack of Context for Extreme Values**:** If someone selects 24 hours, it's unclear whether they're in

****Unavoidability of the Drawbacks**:**

1. ****Precision vs. Accuracy**:** The slider format inherently offers precision, but it's challenging to e
2. ****Potential for Misuse**:** Any interactive survey element, like a slider, introduces the risk of misu
3. ****Varying Definitions of "Studying"**:** This drawback remains consistent across formats. A comprehens
4. ****Overwhelming Range**:** The range is necessary to capture all potential study habits, from those who
5. ****Lack of Context for Extreme Values**:** While additional questions or clarifications could provide c

In summary, the slider format offers a flexible and detailed response mechanism, but it also introduces User

what are some benefits

ChatGPT

Certainly! Let's explore the benefits of using a slider response format for the question:

****Question**:**

On average, how many hours per day do you dedicate to studying outside of regular class hours?

Response Format:

A slider ranging from 0 to 24 hours, allowing for one decimal place precision.

****Benefits**:**

1. ****Granular Data Collection**:** The slider allows for more detailed and nuanced responses, capturing v
2. ****User Engagement**:** Sliders can be more engaging and interactive than traditional dropdowns or radi
3. ****Flexibility**:** Respondents aren't confined to predefined categories or ranges. They can select any
4. ****Visual Representation**:** The slider provides a visual representation of the entire range of possib
5. ****Reduction in Bias**:** Unlike predefined categories, which might inadvertently influence a responden
6. ****Accommodates All Respondents**:** With the full 0-24 hour range, the question ensures that all poten
7. ****Option to Opt-Out**:** By allowing participants the choice not to respond, the format respects respon
8. ****Efficiency**:** Sliders can be quicker to use than dropdown menus, especially for questions where re

Incorporating interactive elements like sliders can enhance the survey-taking experience and provide re User

why select this question in my survey

ChatGPT

Selecting the question "On average, how many hours per day do you dedicate to studying outside of regular

1. ****Core Research Objective****: If the primary aim of your research is to understand the relationship between study habits and academic performance, this question provides a direct measure of study time.
2. ****Quantifiable Data****: The question provides quantitative data, which is easier to analyze statistically.
3. ****Understanding Study Habits****: The question offers insights into students' dedication and commitment to their studies.
4. ****Correlation with Performance****: By understanding how much time students dedicate to studying, you can explore its correlation with academic performance.
5. ****Basis for Further Questions****: Responses to this question can be a foundation for more detailed follow-up questions.
6. ****Informing Educational Strategies****: For educators or institutions, understanding how much students study can help tailor educational strategies.
7. ****Generalizability****: The question is broad enough to be applicable to a wide range of students, regardless of their major or year of study.
8. ****Comparative Analysis****: If similar studies have been conducted in the past, having this question allows for comparative analysis.

Including this question in your survey ensures that you capture essential data related to students' study

User

why i select the question to ask their cgpa

ChatGPT

Selecting the question that asks respondents about their cGPA (cumulative grade point average) can be justified for several reasons:

1. ****Direct Measure of Academic Performance****: cGPA is a widely recognized and standardized metric that directly measures a student's academic achievement.
2. ****Core to Research Objective****: If your research aims to understand the relationship between study habits and academic performance, cGPA is a core variable.
3. ****Quantitative and Objective Data****: cGPA provides quantitative data that can be statistically analyzed.
4. ****Basis for Correlation****: By collecting data on cGPA, you can correlate it with other variables in your survey.
5. ****Benchmarking and Grouping****: cGPA can be used to segment respondents into performance groups (e.g., high achievers, average, and struggling).
6. ****Comparative Analysis****: Asking about cGPA allows for comparisons across different demographics, majors, or years of study.
7. ****Consistency with Previous Research****: Many academic studies that delve into factors affecting student performance include cGPA as a key variable.
8. ****Informing Interventions****: For educational institutions or policymakers, understanding the distribution of cGPA can help identify areas for intervention.
9. ****Validation of Responses****: If other questions in your survey relate to study habits, resources used, or time management, cGPA can serve as a validation point.

Including a question about cGPA in your survey ensures you capture a direct, objective, and quantifiable

User

paraphrase: We began by designing the survey. The very first step was to formulate the questions that would be asked. Other than two insensitive demographic questions which might yield significant insight but we assume not. In addition to the contents, the presentation of the questions is not limited to the traditional ways of asking. Afterwards, the survey was carried out on a survey platform called Qualtrics, which allows for easy distribution. As the survey was carefully designed, the survey will be distributed to 10,000 undergraduate students at the University of Toronto. For the survey, our primary target population is all the undergraduate students at the UofT across all faculties. Our frame population is a list of all the UofT undergraduates who have access to the UofT email system.

Our sample population includes all my frame population who actually filled out my online survey. We aim
ChatGPT

Certainly! Here's a paraphrased version:

We initiated the process by crafting our survey. The primary objective was to devise questions that del
Apart from two non-intrusive demographic queries, which might offer valuable insights but are presumed
Beyond the content, we've diversified the response formats, incorporating not just traditional multiple
The survey was then hosted on Qualtrics, a platform known for its user-friendly interface for both dist
Given the survey's meticulous design, we plan to reach out to 10,000 undergraduate students from the Un
Our primary focus group is the entire undergraduate cohort at UofT, encompassing all faculties and acad
The frame population consists of UofT undergraduates with access to the university's email system.
From this group, our goal is to collect data from 1,000 respondents. Anticipating a 10% participation r

User

what are some drawbacks and strengths

ChatGPT

Certainly! Based on the provided information about your survey procedure, here are some potential drawbacks

Drawbacks:

1. **Response Bias**: Given the voluntary nature of participation, students with strong feelings or opin
2. **Recall Bias**: Questions that require students to remember and report their average study times or
3. **Limited Demographic**: While the survey focuses on the University of Toronto's undergraduate studen
4. **Over-reliance on cGPA**: While cGPA is a widely accepted measure of academic performance, it might
5. **Digital Divide**: Relying solely on an online platform like Qualtrics might exclude potential resp

Strengths:

1. **Comprehensive Question Design**: The survey covers various aspects related to study habits, from th
2. **Engaging Response Formats**: Incorporating diverse response formats, including sliders, can enhanc
3. **Confidentiality Assurance**: The inclusion of a disclaimer emphasizing confidentiality and the vol
4. **Stratified Sampling**: By ensuring representation from all faculties, the survey can draw more acc
5. **Wide Reach**: Distributing the survey to 10,000 students increases the likelihood of obtaining a d
6. **User-Friendly Platform**: Using Qualtrics, a platform known for its intuitive interface, ensures th
7. **Clear Target and Frame Population**: Having a well-defined target, frame, and sample population en

By understanding both the strengths and drawbacks of the survey procedure, you can better interpret the
User

Can you pick three among the strength and drawbacks and elaborate in detail

ChatGPT

Of course! Let's delve deeper into three selected points: two from the strengths and one from the drawbacks.

Strength: Comprehensive Question Design

****Elaboration**:**

The survey's design, which encompasses various facets related to study habits, is one of its most significant strengths.

Strength: Stratified Sampling

****Elaboration**:**

Stratified sampling is a methodological choice that ensures representation from all faculties within the institution.

Drawback: Over-reliance on cGPA

****Elaboration**:**

While cGPA is a standardized and widely accepted measure of academic performance, relying solely on it as a metric for student success is a drawback.

By elaborating on these points, we gain a deeper understanding of the methodological choices made in the survey design.

User
rephrase it: Convenience: The entire process of the survey ranging from the design to the implementation is user-friendly.

ChatGPT
Efficiency: From its inception to execution, the survey process is streamlined and straightforward. The data collection and analysis are efficient.

User

how should i rename the variable `study time` that consider those who study at night and during the day?
ChatGPT

To more accurately represent the variable that differentiates between those who study during the day and those who study at night, you could consider the following options:

1. ****Study Period****: This name indicates the specific time frame or period during which students typically study.
2. ****Preferred Study Timeframe****: This emphasizes the time frame that the student prefers or chooses most for studying.
3. ****Study Session Timing****: This suggests the timing of the student's primary study sessions.
4. ****Primary Study Window****: This indicates the main time window during which a student tends to study.
5. ****Dominant Study Hours****: This highlights the hours during which a student predominantly studies.

Any of these names would provide a clear understanding that the variable differentiates between daytime and nighttime study.

User

give me an example on conducting a CI and hypothesis test
ChatGPT

Certainly! Let's consider a hypothetical scenario.

****Scenario****

Suppose we want to determine if students who study more than 4 hours a day on average have a mean GPA significantly higher than those who study 4 hours or less.

- Sample mean GPA (\bar{x}) = 3.2

- Sample standard deviation (s) = 0.5

We want to conduct a hypothesis test and also compute a 95% confidence interval for the mean GPA of all

Hypothesis Test:

1. **Set up the hypotheses:**

- Null hypothesis (H_0): $\mu = 3.0$ (The mean GPA is 3.0)
- Alternative hypothesis (H_a): $\mu \neq 3.0$ (The mean GPA is not 3.0)

2. **Choose a significance level (α):**

- Let's choose $\alpha = 0.05$.

3. **Calculate the test statistic:**

Using the formula for a one-sample t-test:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Where:

μ_0 is the population mean under the null hypothesis (3.0 in this case),
 n is the sample size.

Plugging in our values:

$$t = \frac{3.2 - 3.0}{0.5/\sqrt{100}} = 4$$

4. **Determine the critical value and make a decision:**

For a two-tailed test with $\alpha = 0.05$ and 99 degrees of freedom (since $n = 100$), the

Conclusion: There is enough evidence at the 0.05 significance level to conclude that the mean GPA of

95% Confidence Interval for the Mean:

Using the formula for a confidence interval for a population mean with a known sample standard deviation

$$\bar{x} \pm t^* \left(\frac{s}{\sqrt{n}} \right)$$

Where:

t^* is the t-value that captures the desired level of confidence (for 95% confidence with 99 de

Plugging in our values:

$$3.2 \pm 1.984 \left(\frac{0.5}{\sqrt{100}} \right)$$

$$3.2 \pm 0.0992$$

Confidence Interval: (3.1008, 3.2992)

Interpretation: We are 95% confident that the true mean GPA of all students who study more than 4 h

Note: This is a simplified example, and in real research, additional considerations might be needed.

User

paraphrase: Because of the uncertainty in the variance estimate - where we don't know the population variance.
ChatGPT

Due to the unpredictability in the variance estimate, when the population variance is unknown, the student

User

paraphrase: The resulting ratio tells us how many standard-error units the coefficient is away from zero.
ChatGPT

The obtained ratio indicates how many standard error units away the coefficient is from zero.

User

students t distribution why p value needs degree of freedom and even minus one in a linear regression model

ChatGPT

The Student's t-distribution and the concept of degrees of freedom (often denoted as (df)) play crucial roles in statistical analysis, particularly in hypothesis testing and confidence interval estimation.

1. **Nature of the t-distribution**: The t-distribution is a family of distributions. Unlike the standard normal distribution, it is wider and flatter, especially for small sample sizes. As the degrees of freedom increase, the t-distribution approaches the standard normal distribution.
2. **Uncertainty and Small Sample Sizes**: The reason we use the t-distribution, especially in the context of small sample sizes, is to account for the additional uncertainty introduced by estimating the population standard deviation from the sample data.
3. **Degrees of Freedom in Linear Regression**: In linear regression, degrees of freedom are related to the number of independent variables in the model. For a simple linear regression with one predictor variable, the degrees of freedom are $n - 2$, where n is the sample size.
4. **Calculating the p-value**: The p-value is calculated based on the t-distribution and tells us the probability of observing a test statistic as extreme as the one calculated from the sample data, assuming the null hypothesis is true.

In essence, degrees of freedom reflect the amount of independent information in the data. By adjusting the degrees of freedom, we can account for the uncertainty introduced by small sample sizes and ensure that our statistical inferences are valid.

User

what is the degree of freedom of a linear model on one predictor variable

ChatGPT

In a simple linear regression model with one predictor variable, the concept of degrees of freedom is used to determine the number of independent pieces of information that go into estimating the parameters of the model.

The linear model can be represented as:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Where:

- y is the dependent variable
- x is the predictor (independent variable)
- β_0 is the y-intercept
- β_1 is the slope of the line
- ϵ represents the error terms

Given this model, the breakdown of degrees of freedom is as follows:

1. **Total Degrees of Freedom**: This is equal to the total number of observations minus 1, represented as $n - 1$.
2. **Model (or Regression) Degrees of Freedom**: Since we're estimating two parameters (the intercept β_0 and the slope β_1), the degrees of freedom for the model is 2.
3. **Error (or Residual) Degrees of Freedom**: This is the difference between the total degrees of freedom and the model degrees of freedom, represented as $n - 2$.

In summary, for a simple linear regression with one predictor variable:

- Model Degrees of Freedom = 2
- Residual Degrees of Freedom = $(n - 2)$
- Total Degrees of Freedom = $(n - 1)$

User

why we interested in the degree of freedom in p value and the sample size in a linear model. students t distribution

ChatGPT

Degrees of freedom (df) and sample size are both fundamental concepts in statistical analysis, especially in the context of hypothesis testing and confidence interval estimation.

1. **Student's t-distribution & Degrees of Freedom (df)**:
 - The t-distribution is a family of probability distributions. The shape of the t-distribution changes depending on the degrees of freedom. For smaller sample sizes, the distribution is wider and flatter, reflecting greater uncertainty.
 - For smaller sample sizes, when estimating population parameters, there's more uncertainty. The t-distribution accounts for this by having a heavier right tail.
 - When computing the p-value using a t-test, we refer to a specific t-distribution determined by its degrees of freedom.
2. **Degrees of Freedom in Linear Models**:
 - In the context of linear regression, df is related to the amount of independent information in the data. For a simple linear regression with one predictor variable, the degrees of freedom are $n - 2$.

- For a simple linear regression (one predictor), the error df is $(n - 2)$ (sample size minus two)
- The concept extends to multiple regression. With more predictors, we estimate more parameters, and

3. **Sample Size**:

- Sample size directly influences the degrees of freedom (larger sample size generally means more df)
- Sample size affects the precision of estimates. Larger samples provide more information, which often leads to more precise estimates.
- However, sample size also affects the df and the shape of the t-distribution. As sample size increases, the t-distribution approaches the normal distribution.

In essence, when evaluating the significance of predictors in a linear model (especially with small sample sizes), the t-distribution is preferred over the normal distribution.

User

ChatGPT

The Student's t-distribution is often preferred over the normal (or z) distribution in certain statistical tests.

1. **Estimation of Population Variance**:

- In many statistical tests, especially with small sample sizes, we don't have access to the true population variance.
- The estimated sample variance has its own variability, and this added uncertainty translates into a wider distribution.

2. **Small Sample Sizes**:

- The central limit theorem tells us that, given a large enough sample size, the sampling distribution of the sample mean will be approximately normal.
- However, for small sample sizes, this approximation might not hold, especially if the underlying distribution is skewed or has heavy tails.
- The t-distribution provides a better model for the sampling distribution of the sample mean for small sample sizes.

3. **Degrees of Freedom and Flexibility**:

- The t-distribution isn't just a single distribution. It's a family of distributions that change shape based on the degrees of freedom.
- As the degrees of freedom increase (which usually corresponds with larger sample sizes), the t-distribution becomes more like the normal distribution.

4. **Conservative Approach**:

- Given the added uncertainty in variance estimation, using the t-distribution tends to be a more conservative approach.
- This conservatism helps control the Type I error rate, ensuring that we don't mistakenly reject a null hypothesis too often.

In summary, while the normal distribution is very useful and applicable in many scenarios (especially with large sample sizes), the t-distribution is often preferred for small sample sizes or when the population variance is unknown.