First Year University Students' GPA and their Gender

Wei-Han Wang [redacted]

April 16, 2021

Abstract

In the introduction, it will introduce the topic of interest of the report and provide reasons to the importance of the topic. After introduction, we will look at the process of data collection as well as data cleaning process. Data cleaning is important as it removes unnecessary variables within a data or observations that will affect the fairness of the results. In the methods section, we will introduce and explain the six main statistical methods that are used to analyze the data with the goal of obtaining an answer to our curiosity. The results section will lay out our results of different statistical methods using plots or tables. Each figure would be accompanied by explanations. At the end, we will end with conclusions to summarize the three sections before as well as suggesting any future work ideas and improvements. At the end of the report, there is an appendix with two sections that document the rigorous mathematical derivations of two of our statistical methods.

Introduction

We will be analyzing the relationship between female and male students' first year university GPA. Gender inequality has always been an underlying issue in education system where different sex have different expectations of them during schooling. According to edchange.org, female students' started to fall behind of male students after high school (Chapman). Apparently, female students received less attention from teachers and sometimes the attention they received were more negative than those given to male students.

It's important for people to recognize this plausible issue in the education curriculum and allow all students to perform their best without external factors altering their ability. Throughout the report, we will separate the data of first year university students' GPA based on sex groups and compare their difference. At the end of the report, we will make a conclusion statement on whether the sex gap in academic performance is true.

The first hypothesis is for the goodness of fit test where we assume our data distribution is a normal distribution. The seecond hypothesis is for hypothesis test where we assume the mean of female and male first year university students' GPA would be the same.

Data

This data is a collection of the first year university students' GPA. This dataset is a sample data from a data collected by a professor of an university in mid-America in 1996. The data contains 219 observations with 10 variables. The data was found through the database of R, located in `stat2Data` package. The ten variables in the data are:

GPA: first year GPA of students
HSGPA: students' high school GPA

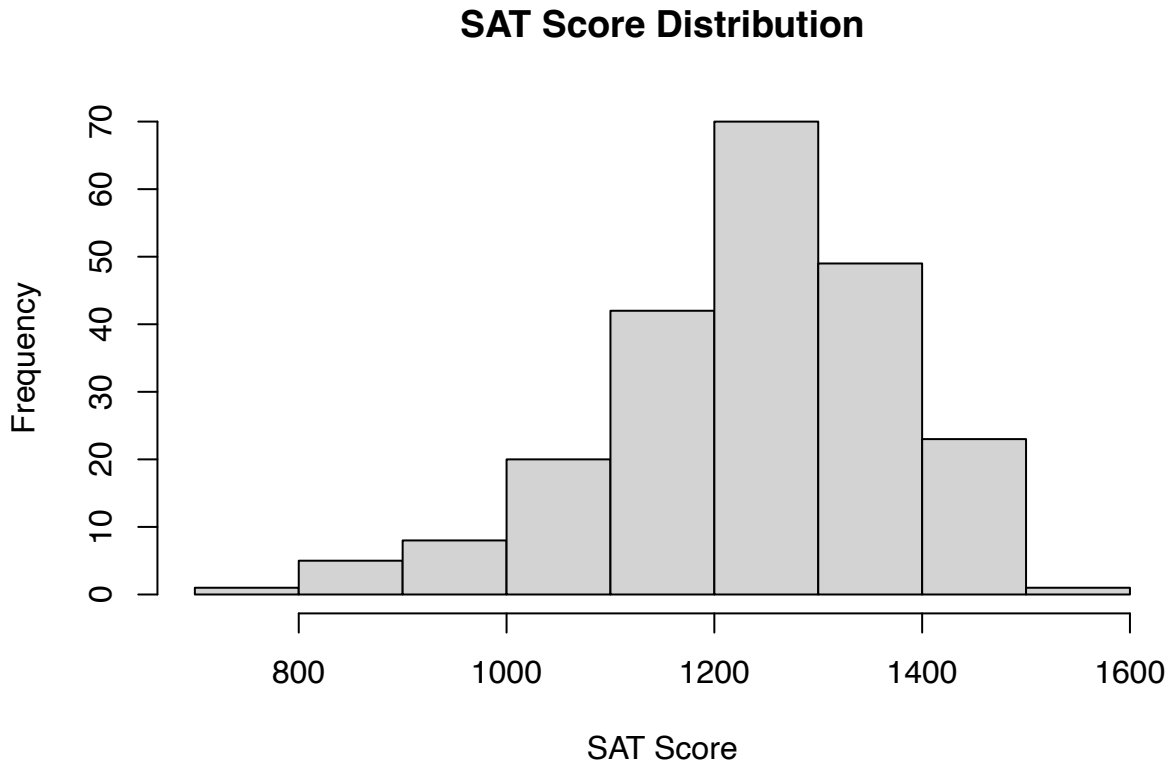
SATV: SAT score of students' reading section
SATM: SAT score of students' mathematics sectiion
Male: indicating the sex of students; 0 means female and 1 means male
HU: humanities credits students earned in high school
SS: social sciences credits students earned in high school
FirstGen: indicates if students are first generaiton university students in the family; 0 means the first gen, 1 means not the first gen
White: indicates if students are white
universitybound: 1 indicates students' high school had more than 50% of students planning on attending university; 0 means otherwise

However, since the focus of the report is on the first year university students' GPA, sex of the students, and their SAT score, we want to only have these variables in our data. In this data, it does not include the total SAT score, but only score of separate SAT sections. We simply add the two columns together to form a column of total SAT score. The important variables for the data set are GPA, Male, SATV, and SATM. After we've created a new variable, we use `select()` to select the three columns we need from the data.

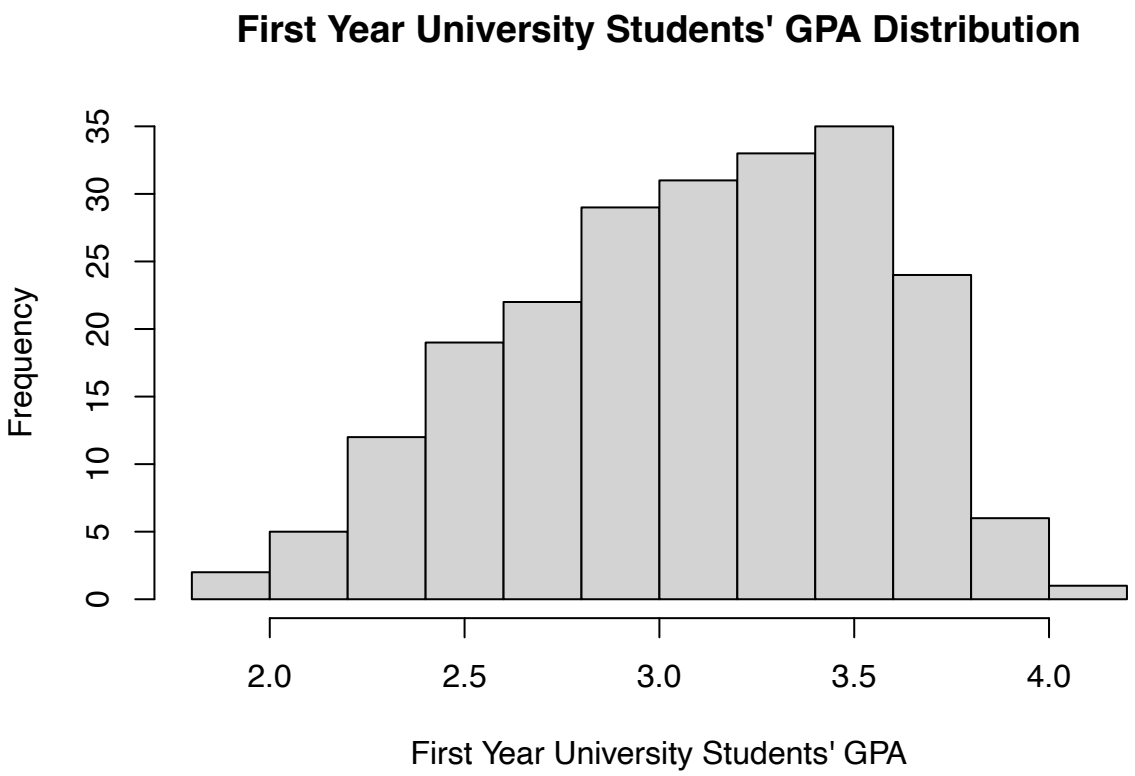
The following is a composed table of basic numeric summary of first year university students' GPA.

Numerical Summary of First Year university Students' GPA				
Avg GPA	Min GPA	Max GPA	Standard Deviation GPA	GPA Variance
3.10	1.93	4.15	0.47	0.22

After calculation, we know the average first year university students' GPA of the 219 students is 3.096. The minimum and maximum GPA is 1.93 and 4.15, respectively. Standard deviation of GPA is 0.47 and variance is 0.22.



The distribution of SAT score seems to be a normal distribution. The histogram does not have any visible outliers. The graph has one peak at around 1200 of SAT score.



The distribution of GPA among first year university students seem to also be a normal distritbuion with a slight skewed to the left. There's no visible outliers in this graph. The graph has a peak around 3.5 of GPA.

We will perform bootstrap on the mean of female and male students in the data. We will use `filter()` to filter out rows of female and male students and store them each under a new variable. Bootstrapping helps us to create a lot of simulations to obtain a more credible result. Using bootstrapping, we can calculate the confidence interval of female and male students' GPA and we can compare if there's any difference between two sexes.

Numerical Summary of First Year University Student's GPA based on Sex			
Sex	Variance	Mean	Standard Deviation
Female	0.2714	3.0732	0.4663
Male	0.2166	3.1226	0.4655

After calculation, the average first year female university students' GPA is 3.0732. Standard deviation of GPA is 0.4663 and variance is 0.2714. The average first year male university students' GPA is 3.1226. Standard deviation of male students' GPA is 0.4655 and variance is 0.2166.

All analysis for this report was programmed using `R version 4.0.4`.

Methods

To evaluate the sex inequality issue, we would be using different statistical methods to analyze and observe the properties of the data.

The following six statistical methods are the ones we will be doing in the report, chronologically: goodness of fit test, maximum likelihood estimator, confidence interval, bayesian credible interval, hypothesis test, and linear regression.

Goodness of Fit Test

A goodness of fit test allows statistician to understand if a data is representative of the population (Stephanie). From the GPA and SAT score distribution histograms in previous section, we will first assume the data is a normal distribution. However, it is possible the data we collected is an outlier that doesn't correspond with the general population it represents. Thus, we want to use goodness of fit test to know if the data is representative of the population.

Our null hypothesis is the data distribution fits a standard normal distribution. The alternative hypothesis is the data distribution does not fit a standard normal distribution.

We can verify whether to reject null hypothesis or not by comparing each break of a standard normal distribution bell curve with our distribution curve break. This would calculate the test statistic of our data.

After we obtain the test statistic of our data, we choose a 5% signficance level for our chi square value and compare it with test statistic. When we have the chi-square value, we would also know the degrees of freedom. If our test statistic is higher than chi-square value, we would reject the null hypothesis, and vice versa.

For this test, we calculate female and male students' data separately and obtain conclusions on whether both female and male students' GPA data fit a normal distribution.

Maximum Likelihood Estimator

Maxmimum likelihood estimation is a method where statisticians focus on parameters of the distribution for such parameters would produce the maximized likelihood of the process that the data we observed (Eppes). For our normal distribution, our parameters of interests are mean, μ, and variance, σ^2.

Assume our data of first year university students' GPA is a random sample of normal random variables with mean μ of 3.1 and fixed variance of σ^2 of 0.22. We'll use the maximum likelhood estimator (MLE) approach to estimate the mean μ. For MLE of μ, it is $\hat{\mu}$. For MLE of σ^2, it is $\hat{\sigma^2}$. All derivations regarding the MLE can be found in Section 1 of the Appendix.

Confidence Interval

A confidence interval (CI) is a plausible range of values that true population mean could be based on bootstrapping. If our sample mean is not within the range, then we can say our sample mean is not a population mean. With repeated sampling (ie. bootstrapping), we are likely to include the true population mean within the interval.

We will be using a 95% confidence interval. The reason we pick this is it's not too wide where we would get signifcant results most of the times like a 99% CI, but it's also not as narrow as a 90% interval where conclusions could be misled due to a smaller range.

We will repeat 1000 times with 219 sample size each time from our GPA data to calculate the average GPA of female and male first year university students separately. After we've obtained the data, we make two histograms of both female and male students and use `quantile()` to compute the 95% confidence interval of population mean.

Bayesian Credible Interval

Bayesian credible interval uses the Bayes' Rule to calculate prior and posterior distribution of data. A prior distribution is a distribution and assumption we make without analyzing the data. A posterior distribution is the distribution we derive after knowing the data.

Assume we are interested in finding a 95% credible interval of the parameter mean of female and male students GPA, μ. Suppose our data is a random sample of normal random variables with mean, μ and fixed variance, $\sigma^2 = 0.2174$ for female students and $\sigma^2 = 0.2166$ for male students. We use 'var()' to find individual variance for female and male.

The male students' prior distribution of μ is assumed to be Normal($\mu_0 = 0$, $\sigma_0^2 = 10$) in hopes of yielding a neutral/non-informative prior. The male students' posterior distribution of μ is Normal with mean $\frac{0.2166n\hat{\mu_{MLE}}}{\mu+0.2166n}$, and variance $\frac{2.166\mu}{0.2166+10n}$.

The female students' prior distribution of μ is assumed to be Normal($\mu_0 = 0$, $\sigma_0^2 = 10$) in hopes of yielding a neutral/non-informative prior. The female students' posterior distribution of μ is Normal with mean $\frac{0.2174n\hat{\mu_{MLE}}}{\mu+0.2174n}$, and variance $\frac{2.174\mu}{0.2174+10n}$.

We can use the first 2.5th percentile and 97.5th percentiles of both female and male distribution to derive a range of values for μ, in which μ_{male} and μ_{female} have 95% probability of falling into. We choose 95% for this method because we want to obtain a consistent significance α level through our report.

All derivations regarding the posterior distribution can be found in Section 2 of the Appendix.

Note that both of the sex are of normal distribution as we have tested in goodness of fit test section. Therefore, both of their derivations would be the same since the prior distribution is the same.

Hypothesis Test

To conduct hypothesis test for difference of two means, we are going to use a two-sample t test. A two-sample t test is similar to a one-sample t test but with two tails, one on each side of the bell curve ("Hypothesis Test: Difference between Means."). The two-sample t test will help us understand whether our results is based on chance or if it's significant. H_0: μ_1 - $\mu_2 = 0$, μ_1 is mean of male students' GPA that follows a normal distribution $N(\mu, \sigma^2)$ and μ_2 is mean of female students' GPA that follows a normal distribution $N(\mu, \sigma^2)$. H_0: $\mu_1 - \mu_2 \neq 0$, μ_1 is mean of male students' GPA that follows a normal distribution $N(\mu, \sigma^2)$ and μ_2 is mean of female students' GPA that follows a normal distribution $N(\mu, \sigma^2)$

We will use `t.test()` to calculate p value and T calculated value of the data. After, we will use `qt()` to find the T critical value of the data using $\alpha = 0.05$ and degrees of freedom from the `t.test()` output. If T critical value is larger than T calculated value, we don't have sufficient evidence to reject the null hypothesis. Thus we fail to reject H_0.

Linear Regression

A linear regression is used to show relationship between two discrete variables through plotting each observation onto a scatterplot. Using a linear regression plot, we can identify whether the correlation between the dependent and independent variables are strong, medium, or weak. To better visualize the relationship between the two variables, sometimes we also plot a best-fit line. A best-fit line is a line where it shows the smallest prediction error of each data point. The formula for best-fit line is the formula of linear regression without the ϵ_i that indicates the prediction error. The following is the linear regression formula:

$$Y = \beta_0 + \beta_1 X_i + \epsilon_i$$

X_i is the indepedent variable of the data and it represents the predicted observation value. β_0 represents the y-intercept when $X_i = 0$. β_1 is the slope of the linear regression plot. The epsilon ϵ_i represents prediction

error, which is the difference between each data point and the best-fit line. It is also the flunctuation to the best-fit line. Y represents the dots on the linear regression plot, which are the dependent variable of this model.

We will plot linear regression between SAT score and GPA based on female and male students. Previous sections we focused on the difference and the distribution of female and male students' GPA to understand if there's any gender inequality. Since SAT scores is another academic testing measure that's done at the end of high school, we'd like to observe if there's any significance difference between the two sexes.

Results

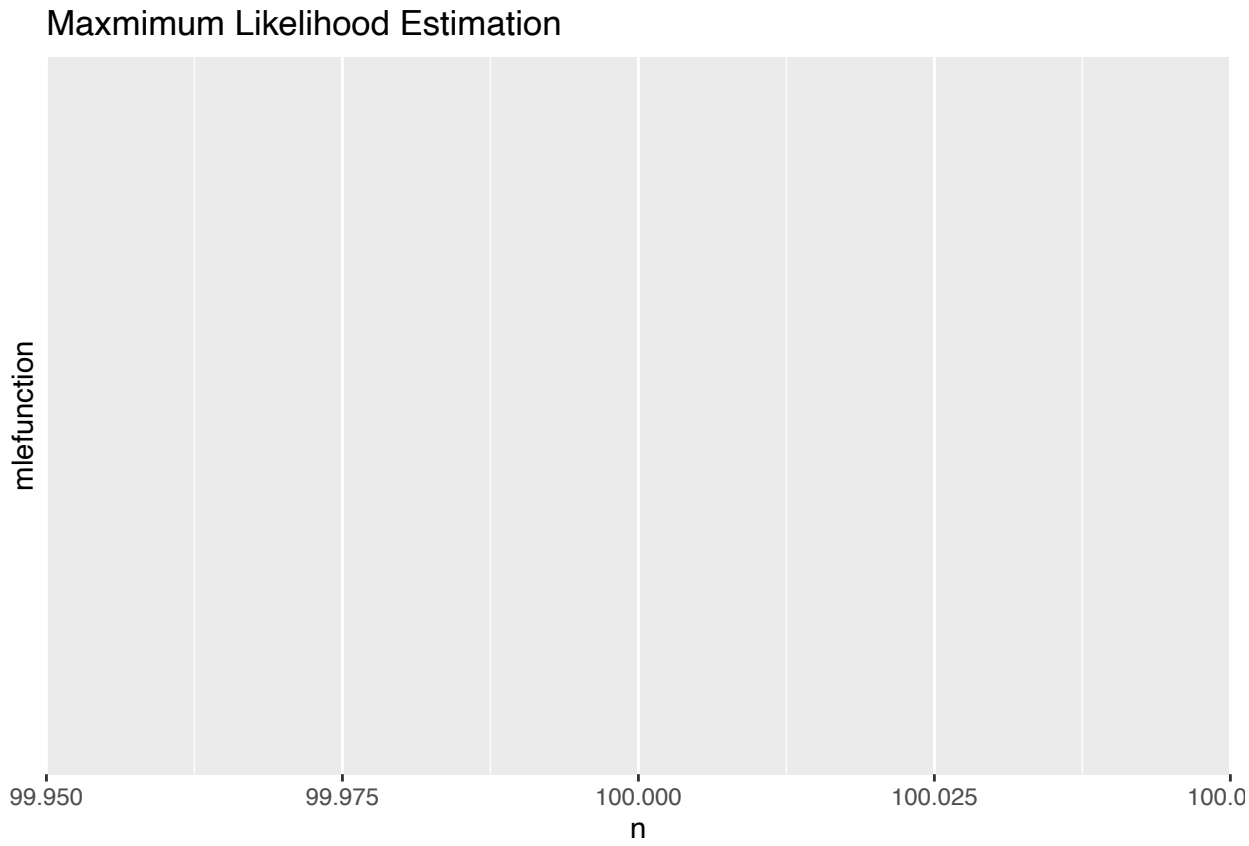
The following six sub sections will introduce the results of our simulation and tests.

Goodness of Fit Test

Goodness of Fit Test Results with 5 percent α level		
Sex	Female	Male
P-value	0.3141	0.7187
χ^2	12.689	7.0714
Degrees of Freedom	11	10
χ^2 critical value	19.6751	18.3070

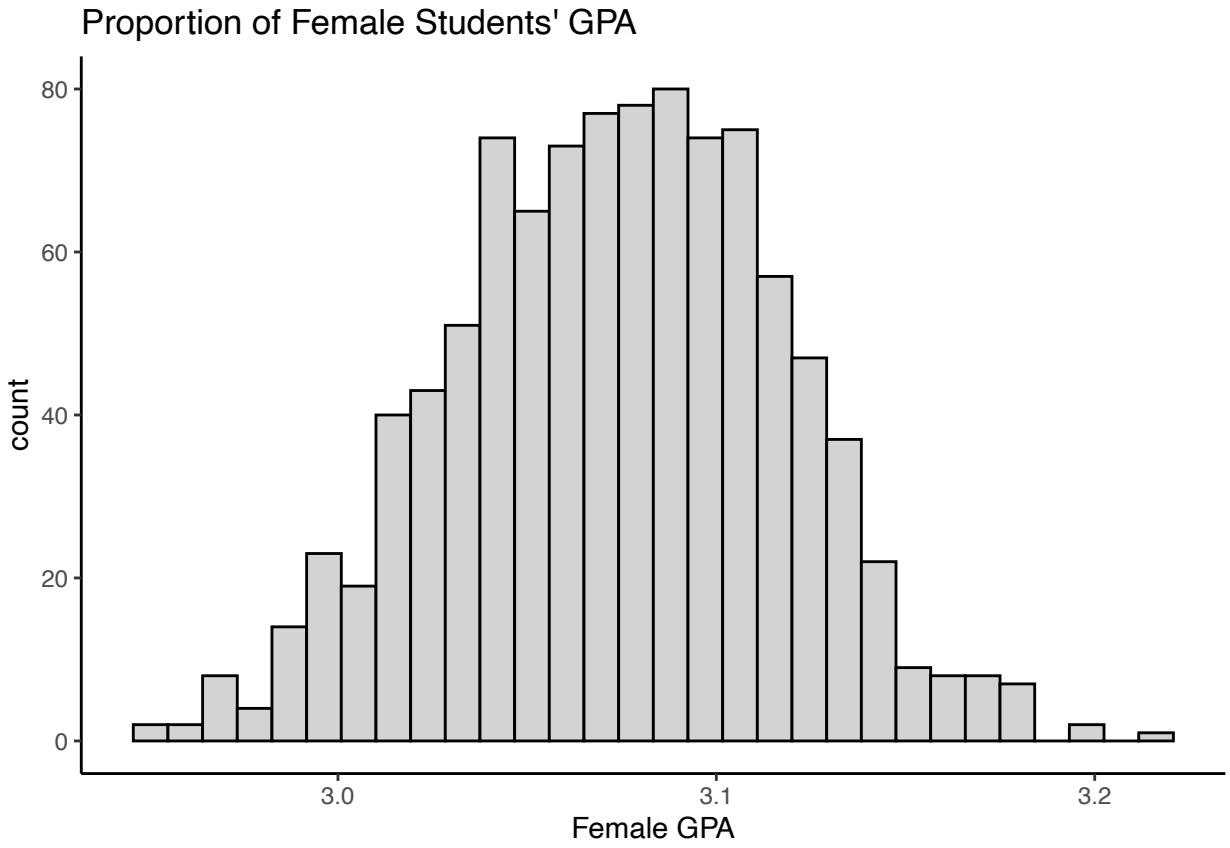
We see that chi square critical value is larger than chi square calculated value for both female and male results so we FAIL TO REJECT null hypothesis that says both data is normal distribution. Therefore, we can say that female and male students' GPA follow a normal distribution.

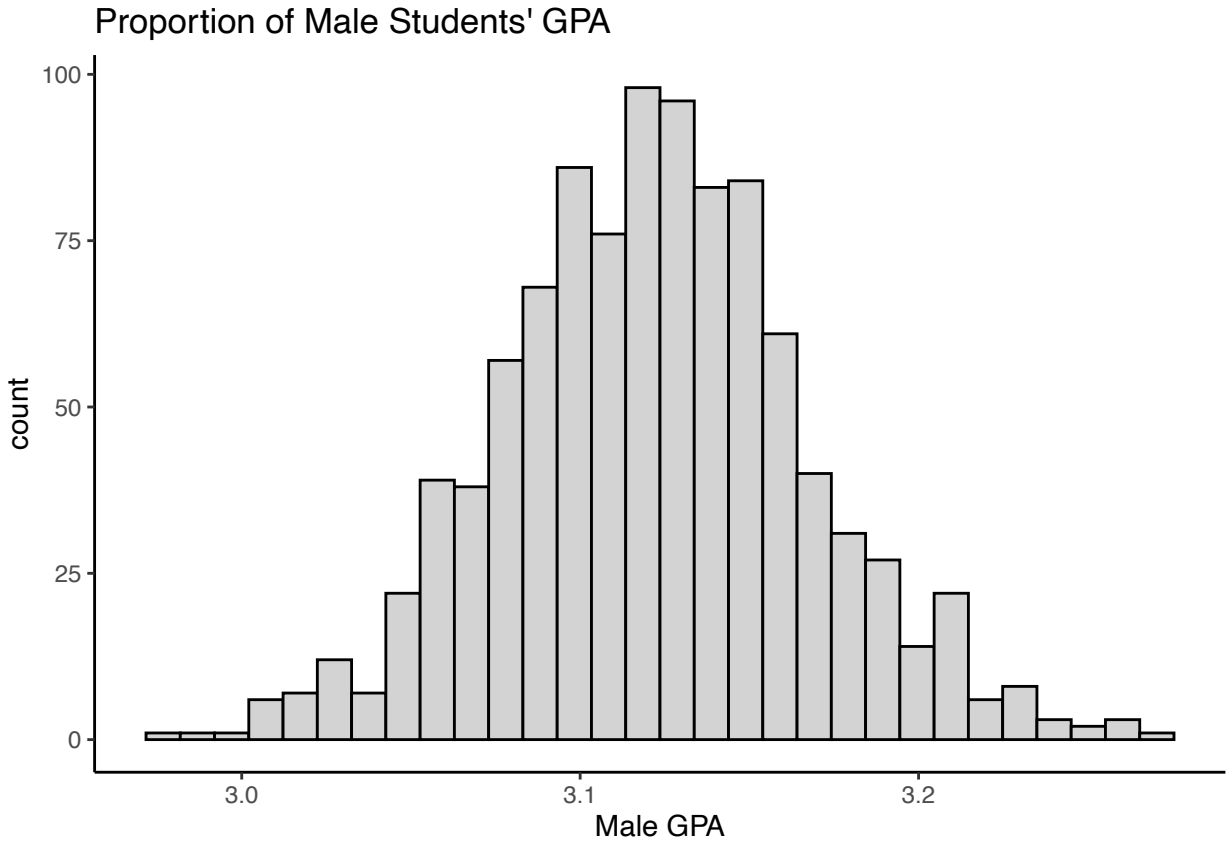
Maximum Likelihood Estimator



From our derivations of maximum likelihood estimator for our normally distributed data (see Appendix 1), we find out that the $\hat{\mu}$ and $\hat{\sigma^2}$ of MLE are the same as a normal distribution mean μ and variance σ^2.

Confidence Interval





The histogram of proportion of female students' GPA is unimodel with a peak around GPA of 3.08. There are possible outliers on the histogram where the bins disconnected. The histogram of proportions of male students' GPA is also unimodel with a peak around GPA of 3.11. There are no visible outliers on the histogram.

After computing the 95% confidence interval based on sex, the results are organized into a table.

95 Confidence Interval based on Sex		
Sex	2.5 percent	97.5 percent
Female	2.99	3.16
Male	3.03	3.21

From the table, we can say that we are 95% confident that the GPA interval of female first year university students is from 2.99 to 3.16 and the GPA interval of male first year university students is from 3.03 to 3.21.

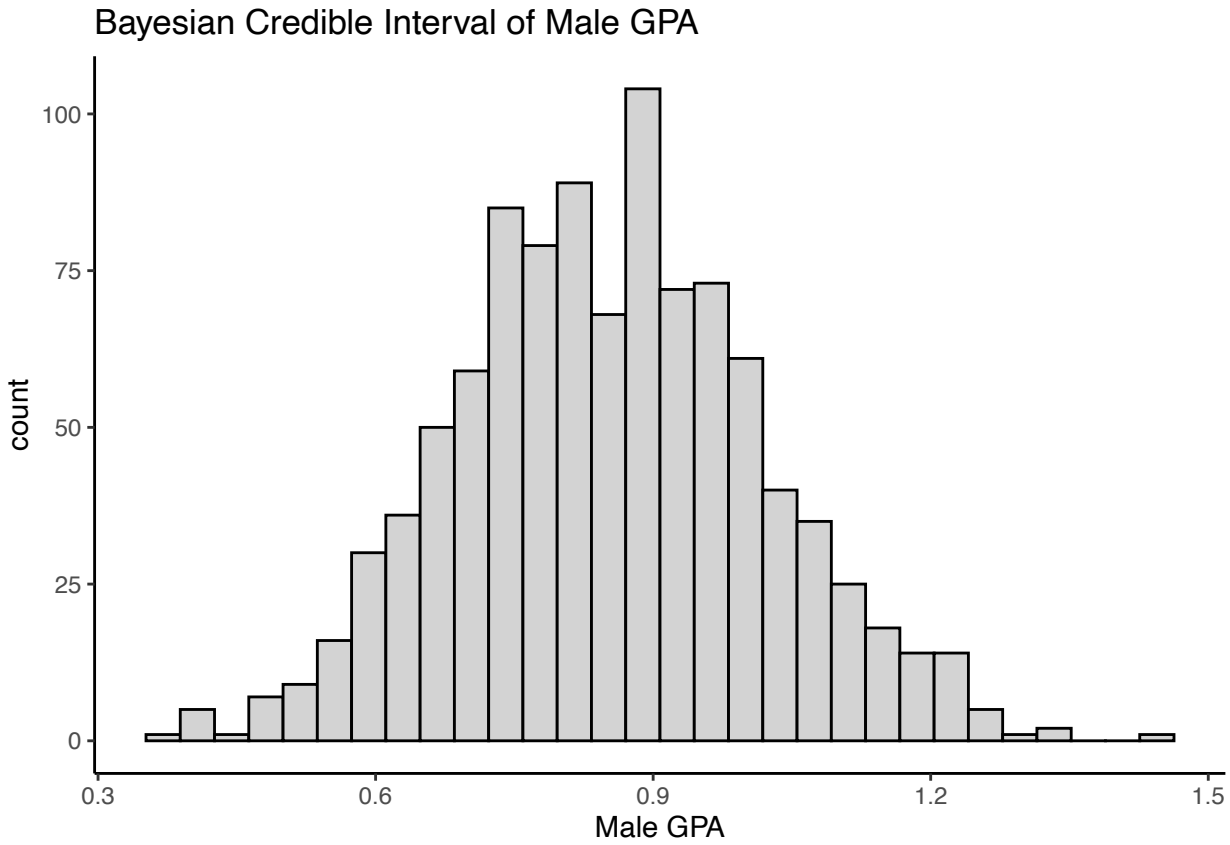
The confidence interval is valid because from our data section, we've known the mean of female students' GPA is 3.07. And we can see 3.07 is within the range of 2.99 to 3.16. The mean of male students' GPA is 3.12 which is within the interval of 3.03 and 3.21.

Hypothesis Test

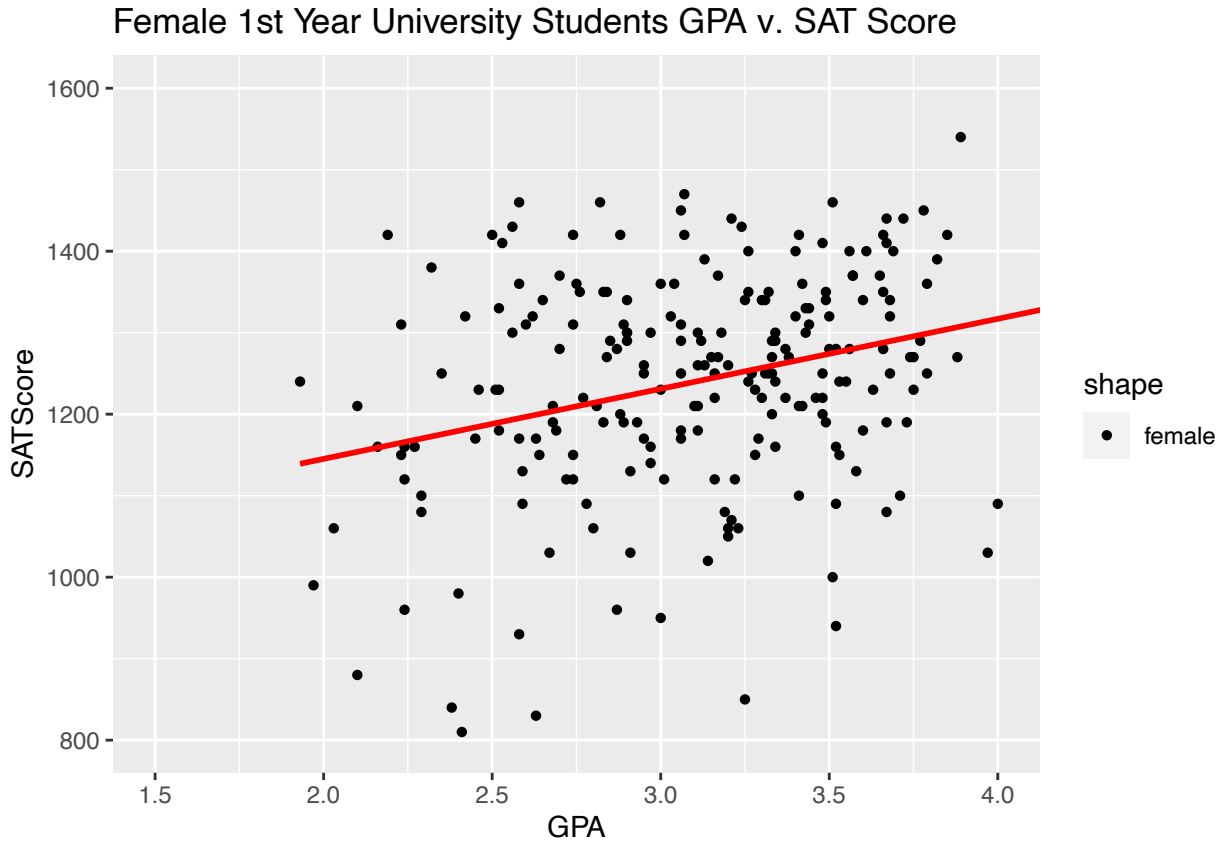
Hypothesis Test Summary		
P value	T Calculated Value	T Critical Value
0.4364	0.7797	1.9710

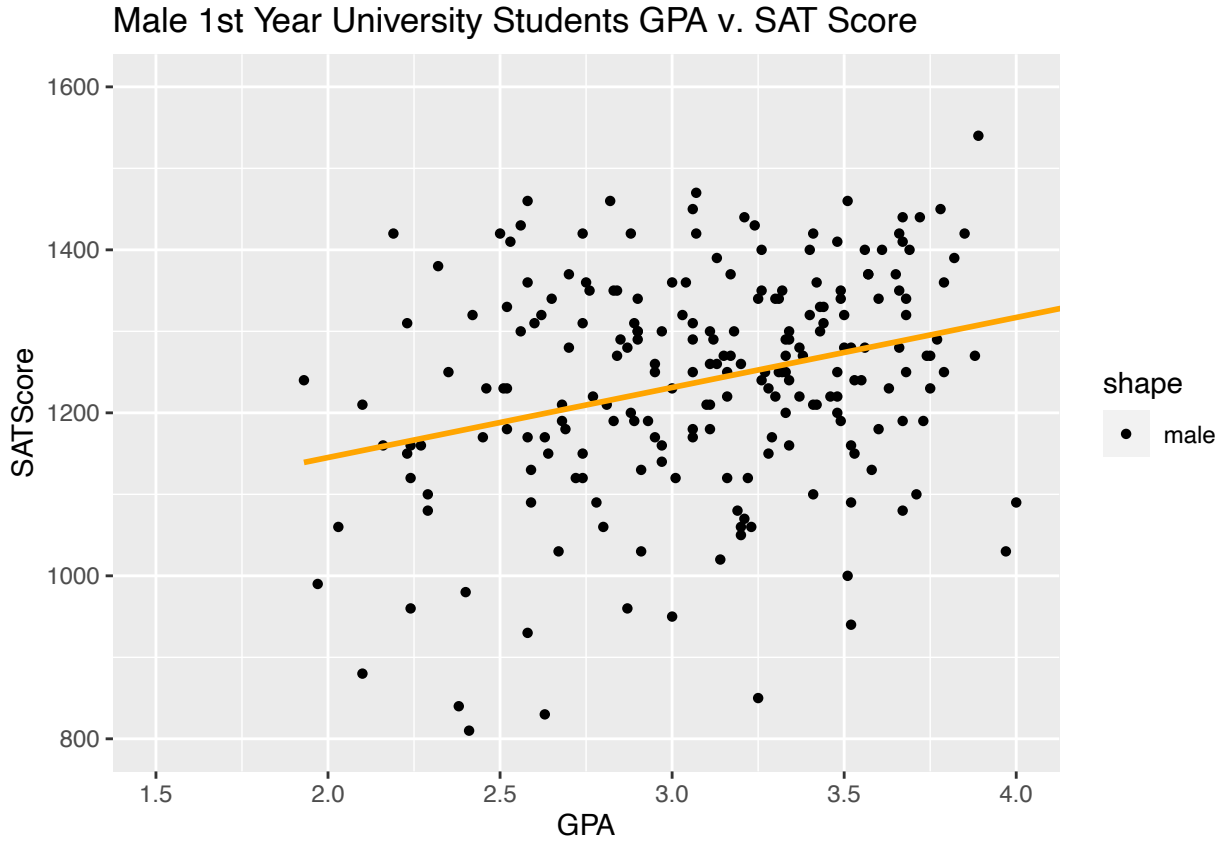
Since the p-value is greater than the significance level of 5% and T critical value is larger than calculated value, we don't have sufficient evidence to reject the null hypothesis of female and male students having same GPA average.

Bayesian Credible Interval



Linear regression





Scatterplot for female students' GPA and SAT score does not seem to have strong correlation between the two variables. The dots are very spread out with no clear slope. We graphed a linear regression line on the graph and we can further confirm our assumptiion that there's no strong correlation between two variables because the gap between each plot and the line is wide.

Male students' GPA and SAT score scatterplot seems to have a slightly weaker correlation between two variables than female students' scatterplot. The correlation of male students' GPA and SAT score doesn't seem to be strong. We also graphed a linear regression line on the graph and we can see that male students do have a weaker correlation because the residuals are comparatively larger than female students'.

Linear Regression Table			
Sex	$\hat{\beta}_0$	$\hat{\beta}_1$	Correlation Degree
Male	1067.26	68.82	0.26
Female	917.14	92.71	0.31

After computing the intercepts and slopes of both scatterplot, we also calculaated the correlation degree of two scatterplots and organize these values into a table. A correlatiion degree indicaates how strongly two numerical variables correlate with each other. Positive number means a positive correlation and negative means a negative correlation. The closer the degree value to zero, the weaker the correlation.

We see the correlation degree of male students is 0.26 and female students is 0.31. This proves our assumption from observing the scatterplot that the correlation of male students' GPA and SAT score is weaker than female students' GPA and SAT score. Our intercept for female students is 917.14 and slope is 92.71. This means for every 1.0 increase in GPA there's a 92.71 increase in SAT score. And when the GPA is zero, female students' SAT score would intercept y-axis at 917.14. Our intercept for male students is 1067.26 and slope is 68.82. It means for every 1.0 increase in GPA, there would be 68.82 increase in SAT score. When GPA of male students is zero, its SAT score would intercept y-axis at 1067.2,

Conclusions

Using goodness of fit test, we confirmed the distribution of data to be normal. We set the null hypothesis to be the data is a normal distribution and the alternate hypothesis is the data is not a normal distribution. We used chi-square test to find the chi-square critical value to be larger than calculated value. Thus we conclude we don't have sufficient evidence to reject the null hypothesis.

Using maximum likelihood estimator, we confirmed the parameters of our data that would maximized our results is the same as the parameters we have from the beginning which is $\mu = 3.1$ and $\sigma^2 = 0.22$.

From calculating 95% confidence interval of mean of female and male students' GPA, we understand the plausible range of values for the mean of the two sexes. After calculation, we see the two sample mean we calculated in the data section to be within the interval of both female and male. The two intervals are not significantly different so we can't conclude there's any apparent gender inequality.

From conducting hypothesis test of mean on both female and male students, we concluded the mean GPA of female and male are the same. In the test, we wrote the null hypothesis as there is no difference between two sexes. The alternate hypothesis to be there is a difference between the mean of female and male students. We set the significance level to 5% for the test and our p-value is greater than the significance level after calculation. Since p-value is greater than significance level, we don't reject the null hypothesis of female and male students having same GPA average.

From our linear regression plot, we showed there's no correlation between SAT score and first year university GPA for female and male students. The intercepts and slope of female and male students' linear regression plot are not significantly different so it shows female and male perform similarly on SAT score and their first year university GPA.

Weaknesses

One of the major weaknesses of the report is the main data we are working with is from a few decades ago. Therefore, the credibility of the data is debatable. Another weakness of the data is it's a sample from a bigger statistical data collection from a professor so this data could be the outlier among the data inputs of the original data.

One big weakness of the entire report is when we calculate the maximum likelihood estimator, we are making assumptions we know the distribution of the data through a goodness of fit test and visualizations. We also don't have enough prior distribution information for calculating the Bayesian approach section. There's no easy way to determine the distribution of the data besides visualizing the histogram plots. The estimations using goodness of fit test and visualizing histogram plots could be incorrect.

Next Steps

For future work, we can use a data that is more recent so we can obtain a better credibility and voice with our conclusions and findings. In addition, it is more ideal if we could find a larger data set that includes students from different backgrounds. A large sample size indicates the data is less likely to be biased or different from the general population.

We could look for different factors that could influence first year university students' GPA. Although sex could be one of the reasons, family backgrounds, education experience, and personal talent, and much more can factored into the difference or indifference of first year university students' GPA based on sex.

Discussion

After utilizing six differnet statistical methods to understand our analysis question, we don't have strong evidence to say there is a gender inequality between female and male students. Although female students' GPA confidence interval and linear regression model is slightly lower than male students' GPA, the difference within 0.05 which doesn't serve as a significant difference.

Even though with this data we couldn't find strong evidence to support the issue of gender inequality in education curriculum, it does not mean there's no gender inequality in school. It is possible the location, the environment, the culture and many more factors of the data that affect the results of our report.

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)

2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.

3. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)

4. (2021, April 19) *"Data Transformation Cheatsheet."* RStudio. [http://https://github.com/rstudio/cheatsheets/raw/master/data-transformation.pdf] (https://github.com/rstudio/cheatsheets/raw/master/data-transformation.pdf) (Last Accessed: April 16, 2021)

5. Chapman, Amanda. *"Gender Bias in Education."* EdChange Consulting and Workshops on Multicultural Education, Diversity, Equity, Social Justice. http://edchange.org/multicultural/papers/genderbias.html. (Last Accessed: April 19, 2021)

6. (2021, April 19) *"List of LaTeX Symbols."* LaTeX Wiki, http://latex.wikia.org/wiki/List_of_LaTeX_symbols#Hats.2C_bars.2C_and_accents. (Last Accessed: April 19, 2021)

7. (2021, April 19) *Bayesian Update of a Normal Prior Distribution* https://www.mhnederlof.nl/bayesnormalupdate.html. (Last Accessed: April 18, 2021)

8. Eppes, Marissa. (21, Sept. 2019) *"Maximum Likelihood Estimation Explained - Normal Distribution."* Medium, Towards Data Science. http://towardsdatascience.com/maximum-likelihood-estimation-explained-normal-distribution-6207b322e47f. (Last Accessed: April 18, 2021)

9. *Maximum Likelihood Estimation of Gaussian Parameters.* (2017, 18 Aug)http://jrmeyer.github.io/machinelearning/2017/08/18/mle.html. (Last Accessed: April 15, 2021)

10. Stephanie. *"Goodness of Fit Test: What Is It?"* Statistics How To. (16 Sept. 2020) https://www.statisticshowto.com/goodness-of-fit-test/ (Last Accessed: April 17, 2021)

11. *"Aligning Equations with Amsmath." Overleaf.* Online LaTeX Editor. [http://overleaf.com/learn/latex/aligning_equations_with_amsmath] (https://www.overleaf.com/learn/latex/aligning_equations_with_amsmath). (Last Accessed: April 16, 2021)

12. *"Tables."* Overleaf, Online LaTeX Editor. https://www.overleaf.com/learn/latex/tables#Reference_guide. (Last Accessed: April 15, 2021)

13. *"Hypothesis Test: Difference between Means."* Stat Trek. https://stattrek.com/hypothesis-test/difference-in-means.aspx#:~:text=This%20lesson%20explains%20how%20to,the%20following%20conditions%20are%20met%3A&text=Each%20population%20is%20at%20least%2020%20times%20larger%20than%20its%20respective%20sample. (Last Accessed: April 18, 2021)

14. Zach. *"Welch's t-Test: When to Use It + Examples."* Statology. (20 Dec. 2020). https://www.statology.org/welchs-t-test/. (Last Accessed: April 17, 2021)

Appendix

Section 1

Probability density function of normal distribution:

$$f_X(x_i) = (2\pi\sigma^2)^{\frac{-1}{2}} e^{\frac{-1}{2}\frac{(x_i-\mu)^2}{\sigma^2}}$$

$$\begin{aligned} L(\mu, \sigma^2; x) &= f(x|\mu, \sigma^2) \\ &= (2\pi\sigma^2)^{\frac{-n}{2}} * e^{(-\frac{1}{2\sigma^2} * \sum_{i=1}^{n}(x_i-\mu)^2)} \end{aligned}$$

Take the log of the likelihood function above.

$$\begin{aligned} LL(\mu, \sigma^2; x) &= ln[f(x_1|\mu, \sigma^2) \cdot f(x_2|\mu, \sigma^2)...f(x_n|\mu, \sigma^2)] \\ &= ln[(2\pi\sigma^2)^{\frac{-n}{2}} * e^{(-\frac{1}{2\sigma^2} * \sum_{i=1}^{n}(x_i-\mu)^2)}] \\ &= -\frac{1}{2}nln(2\pi\sigma^2) - \frac{1}{2\sigma^2} * \sum_{i=1}^{n}(x_i-\mu)^2 \\ &= -\frac{1}{2}nln(2\pi) - \frac{1}{2}nln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2) \end{aligned}$$

Take partial derivative of log likelihood function with respect to μ:

$$\frac{d}{d\mu} -\frac{1}{2}nln(2\pi) - \frac{1}{2}nln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2) = 0$$

$$\begin{aligned} \frac{d}{d\mu}LL(\mu, \sigma^2; x_n) &= \frac{d}{d\mu} -\frac{1}{2}nln(2\pi) - \frac{1}{2}nln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2) \\ &= 0 - 0 + \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-n\mu) \end{aligned}$$

$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0$ when $\sum_{i=1}^{n}(x_i-\mu) = 0$. Using this, we solve for $\hat{\mu}$.

$$\begin{aligned} 0 &= \sum_{i=1}^{n}(x_i-\mu) \\ 0 &= \sum_{i=1}^{n}(x_i) - \sum_{i=1}^{n}(\mu) \\ 0 &= \sum_{i=1}^{n}(x_i) - n\mu \\ n\mu &= \sum_{i=1}^{n}(x_i) \\ \mu &= \frac{1}{n}\sum_{i=1}^{n}(x_i) \end{aligned}$$

Take partial derivative of log likelihood function with respect to variance, σ^2:

$$\begin{aligned}\frac{d}{d\sigma^2}LL(\mu,\sigma^2;x_n) &= \frac{d}{d\sigma^2}-\frac{1}{2}nln(2\pi)-\frac{1}{2}nln(\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2) \\ &= 0-\frac{n}{2\sigma^2}+[\frac{1}{2}\sum_{i=1}^{n}(x_i-n\mu)^2]\frac{1}{\sigma^2}^2 \\ &= -\frac{n}{2\sigma^2}+[\frac{1}{2}\sum_{i=1}^{n}(x_i-n\mu)^2]\frac{1}{\sigma^4} \\ &= \frac{n}{2\sigma^2}-n+\frac{1}{sigma^2}\sum_{i=1}^{n}(x_i-n\mu)^2\end{aligned}$$

$\frac{n}{2\sigma^2}-n+\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-n\mu)^2=0$ to solve for $\hat{\sigma^2}$

$$\begin{aligned}0 &= \frac{n}{2\sigma^2}-n+\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-n\mu)^2 \\ 0 &= -n+\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-n\mu)^2 \\ n &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-n\mu)^2 \\ n\sigma^2 &= \sum_{i=1}^{n}(x_i-n\mu)^2 \\ \sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\end{aligned}$$

Final results:

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}(x_i)$$

Section 2

Posterior distribution equation: $P(\mu,\sigma^2|data)=\frac{P(data|\mu,\sigma^2)P(\mu,\sigma^2)}{P(data)}$

Assume $X_1,...,X_n \overset{iid}{\sim} \mathcal{N}(\mu,\sigma^2)$ and $\mu \backsim \mathcal{N}(\mu_0,\sigma_0^2)$

$\frac{P(data|\mu,\sigma^2)P(\mu,\sigma^2)}{P(data)} \propto P(data|\mu)\cdot P(\mu)$
$= e^{\frac{-1}{2\sigma_0^2}(\mu-\mu_0)^2}\prod_{i=1}^{n}e^{\frac{-1}{2\sigma^2}(x_i-\mu)^2}$
$= e^{\frac{-1}{2\sigma_0^2}(\mu-\mu_0)^2-\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$

$\frac{-1}{2\sigma_0^2}(\mu-\mu_0)^2-\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$
$= \frac{-1}{2\sigma_0^2}(\mu^2-2\mu\mu_0+\mu_0^2)-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i^2-2x_i\mu+\mu^2)$
$= (\frac{-1}{2\sigma_0^2}-\frac{n}{2\sigma_0^2})\mu^2+(\frac{\mu_0}{2\sigma_0^2}+\frac{1}{2\sigma^2}\sum x_i)\mu$ (complete the square)

$= (\frac{-1}{2\sigma_0^2} - \frac{n}{2\sigma_0^2})(\mu^2 - 2\frac{\frac{\mu_0}{\sigma_0^2} + \frac{\sum x_i}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}})\mu + \mu_n$

$= \frac{-1}{2}(\frac{-1}{\sigma_0^2} - \frac{n}{\sigma_0^2})(\mu^2 - 2\frac{\frac{\mu_0}{\sigma_0^2} + \frac{\sum x_i}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}})\mu + \mu_n$

$= \frac{-1}{2}(\frac{-1}{\sigma_0^2} - \frac{n}{\sigma_0^2})(\mu - \mu_n)^2$

$= \frac{-1}{2\sigma_n^2}(\mu - \mu_n)^2$

$\because P(\mu|data) \propto e^{\frac{-1}{2\sigma_n^2(\mu-\mu_n)^2}}$

$\sigma_n^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}$

$\mu_n = \sigma_n^2(\frac{\mu_0}{\sigma_0^2} + \frac{\sum x_i}{\sigma^2})$

$\mu_n = \frac{\sigma_n^2}{\sigma_0^2}\mu_0 + \frac{\sigma_n^2}{\sigma^2}n\hat{\mu}_{MLE}$

posterior: $\mu|data = \mathcal{N}(\mu_n, \sigma_n^2)$