Module: MATPMDB
Assessment: Statistical Inference Portfolio
Due Date/Time: 10 Nov at 23:59
AIAS Levels Allowed: Level 3

	Please tick the boxes/include appropriate information below
Student ID Number	3542490
Word Count (penalties apply for exceeding the stated limit)	35767
I have read and understand the severity of academic misconduct – see link here.	☑
I give consent for my work to be used as an exemplar to future students.	☑
I have checked my submitted document to ensure it complies with module requirements.	☑
Link to version-controlled file (i.e. on OneDrive, Google Docs, Github, or other) which contain evidence of the process I undertook to complete this assignment. Information on how to create a Microsoft 365 OneDrive folder is available HERE. *Please see notes below	https://github.com/sezfabian/Statistical-Inference-Portfolio
I understand that if there is a concern about potential academic misconduct, including inappropriate use of AI tools, then I could be asked to provide evidence of my drafting process during an academic integrity meeting if I have not done so using the link above. Not providing evidence of my drafting process could prejudice the outcome of academic misconduct cases.	☑
Tailored feedback. If you would like tailored feedback on a specific aspect (or aspects) of your work (e.g., referencing, writing style, grammar), then please give details here.	
If you used AI at (or below) the level allowed, please explain briefly which AI, how you used it, and for what purpose.	I used AI to learn r programming concepts faster and understand complex functions from the matpmdb manual.

**This may include (but is not limited to) drafts, versions of the finished document, notes, references, AI output, and AI prompts. These materials are not marked or graded, but they are simply a way to demonstrate how your work was created and to confirm that any AI use in your final submission is within the permitted AIAS scale for your assessment. Providing this helps safeguard you, showing your authentic process, and protecting you should any academic integrity questions arise.*
For more information, please see the Coversheet FAQ.

Statistical Inference Portfolio

Fabian Cheruiyot. 3542490

2025-11-05

Introduction

To track my personal weight loss journey, I collected a dataset using my smart watch, a kitchen scale, a bathroom scale and the Google fit app that contains hours of sleep, calories consumed, calories used, and my body weight in pounds, with each row representing a day of the month. Additionally, I recorded the type of day as busy/relaxed based on whether I attended any classes during the day. I this portfolio I applied the statistical methods; t-test, Linear regression, Fisher's exact test and the A log-likelihood/support function to analyze and draw supported conclusions from my data.

Health data tracking

```
october_data <- read.table("october_data.txt", header=TRUE)
head(october_data, 5)
```

```
##      Date Sleep.Hours. Calories_Consumed Calories_Used Caloric_Difference Weight.lbs.
Type_of_Day
## 1 1/10/2025         5.33              1120          2635               1515       233.2
Busy
## 2 2/10/2025         8.33               548          3032               2484       232.8
Busy
## 3 3/10/2025         8.38              1450          2493               1043       231.8
Relaxed
## 4 4/10/2025         6.45              1250          2652               1402       231.4
Relaxed
## 5 5/10/2025         8.23              1040          2471               1431       230.8
Relaxed
```

1. Student T-test

a). Weight Change over the month of October (One-sided t-test)

Here I will examine whether my body weight significantly changed between the first half and second half of the month of October.

- ***Null Hypothesis(H_0)***: There is no difference in my body weight between the first and second half of the month of October.
  Alternative Hypothesis(H_1): My mean body weight is lower in the second half of the month compared to the first half of the month.

```
# Student t-test comparing mean of body weight recorded in the second
half of the month against the first half of the month.
# Covert 'Date' from string to Date format
october_data$Date <- as.Date(october_data$Date, format = "%d/%m/%Y")
# Split data between to two halves by date
second_half <- october_data$Date >= "2025-10-16"
# Student t-test
ans <-
t.test(october_data$Weight.lbs.[second_half],october_data$Weight.lbs.[!
second_half], alternative = "less")
# Mean Values
ans$estimate

## mean of x mean of y
##  225.9938  230.3867

# Print P-value
cat("  P-value:", ans$p.value)

##   P-value: 1.779958e-08
```

- ***One-sided test Justification***: Based on general knowledge and observations I expect my body weight to reduce from maintaining a calorie deficit over a sustained period of time.

- ***Statistical interpretation***: P-value < 0.05, hence we reject the null hypothesis. This affirms the alternate hypothesis that my body weight decreased.

- **Contextual interpretation**: Over the course of the month of October, my body weight shows a statistically significant decrease that aligns with my weight loss goals that can be attributed to having maintained a calorie deficit on most days of the month.

b). Impact of good sleep on daily activity and calories consumed (Two-sided t-test)

i. **Daily Activity (Calories Used)**

Here I will examine whether I was more or less active(used more/less calories) on days that I had good(enough) sleep where good sleep is defined by the adult standard of 8-hours sleep.

Null Hypothesis(H_0): Getting good sleep has no impact on whether I am more or less active during the day.

Alternative Hypothesis(H_1): Good sleep impacts how active I am during the day.

```
# Split data into two based on number of hours of sleep
good_sleep <- october_data$Sleep.Hours >= 8
# Student t-test
ans <-
t.test(october_data$Calories_Used[good_sleep],october_data$Calories_Use
d[!good_sleep])
# Mean Values
ans$estimate

## mean of x mean of y
##  2947.500  3045.095

# Print P-value
cat("  P-value:", ans$p.value)

##   P-value: 0.675311
```

- ***Two-sided test Justification***: Daily activity can increase or decrease with good sleep as sleeping more could mean having a less active day as more time is spent sleeping or having enough rest to be more active during the day.

- ***Statistical interpretation***: **P-value > 0.05**, hence we fail to reject the null hypothesis, thus: Getting good sleep has no impact on whether I am more or less active during the day.

- **Contextual interpretation**: Despite a difference in the calculated mean of calories used on days with good sleep and days without good sleep, the t-test analysis indicates that there is no statistically significant difference in my daily activity, between days with good sleep and days without. This implies that daily activity is not associated with sleep or the sample size is not sufficient.

ii. **Food intake (Calories Consumed)**

Here I will examine whether I consumed more or less calories on days that I had good(enough) sleep, where good sleep is defined by the adult standard of 8-hours sleep.

Null Hypothesis(H_0): Getting good sleep has no impact on the amount of calories consumed during the day.
Alternative Hypothesis(H_1): Good sleep impacts how much food(calories) I consumed during the day.

```r
# Student t-test
ans <-
t.test(october_data$Calories_Consumed[good_sleep],october_data$Calories
_Consumed[!good_sleep])
# Mean Values
ans$estimate

## mean of x mean of y
##  1714.200  1767.571

# Print P-value
cat("  P-value:", ans$p.value)

##   P-value: 0.8676897
```

- ***Two-sided test Justification***: Sleeping 8 or more hours could affect the calories I consume in both directions. Sleep may or may not affect both the amount of food and the type of food I consume.

- ***Statistical interpretation***: P-value > 0.05, hence we fail to reject the null hypothesis, thus: Getting good sleep has no impact on the amount of calories consumed during the day.

- **Contextual interpretation**: Despite a difference in the calculated mean of calories consumed on days with good sleep and days without good sleep, the t-test analysis indicates that there is no significant difference in the mean. We observe that my eating habits are independent of whether I had enough sleep or not.

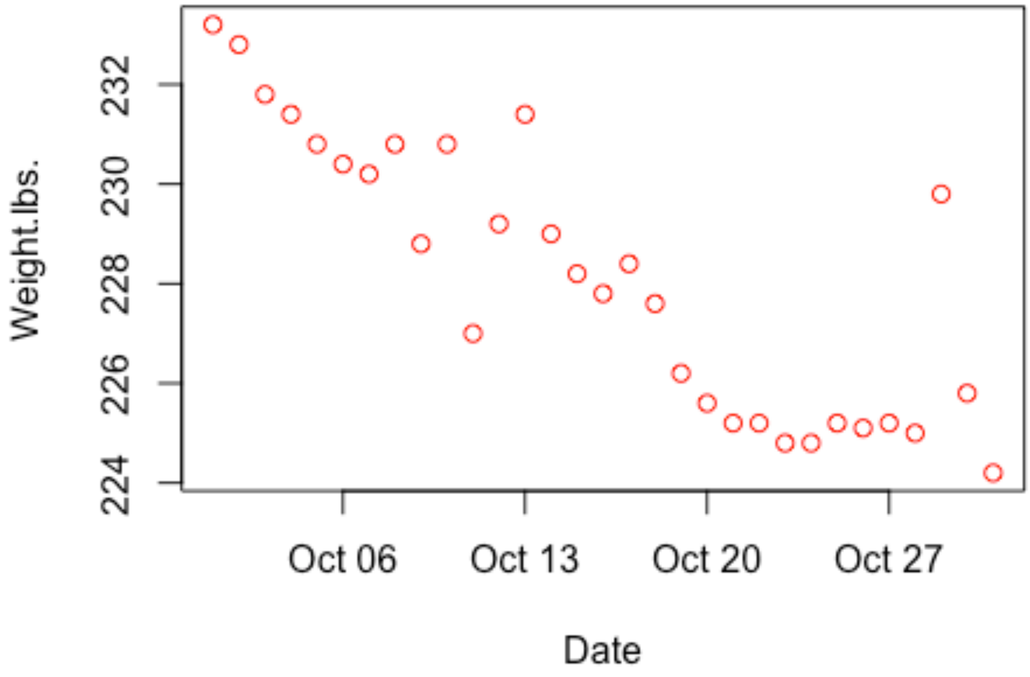
2. Linear regression

My dataset includes 4 key independent variables (sleep (hrs), calories_consumed, calories_used, and weight(lbs)). The collected data is noisy as I did not use neither exact methods nor perfect equipment for measurement/collection of the data.
In this section I will analyze the variables to map relationships and check if I can use the modeled relationships to influence my weight loss journey.

a). Weight loss trend analysis.

- Here I am examining whether my body weight decreases with liner trend through out the month of October

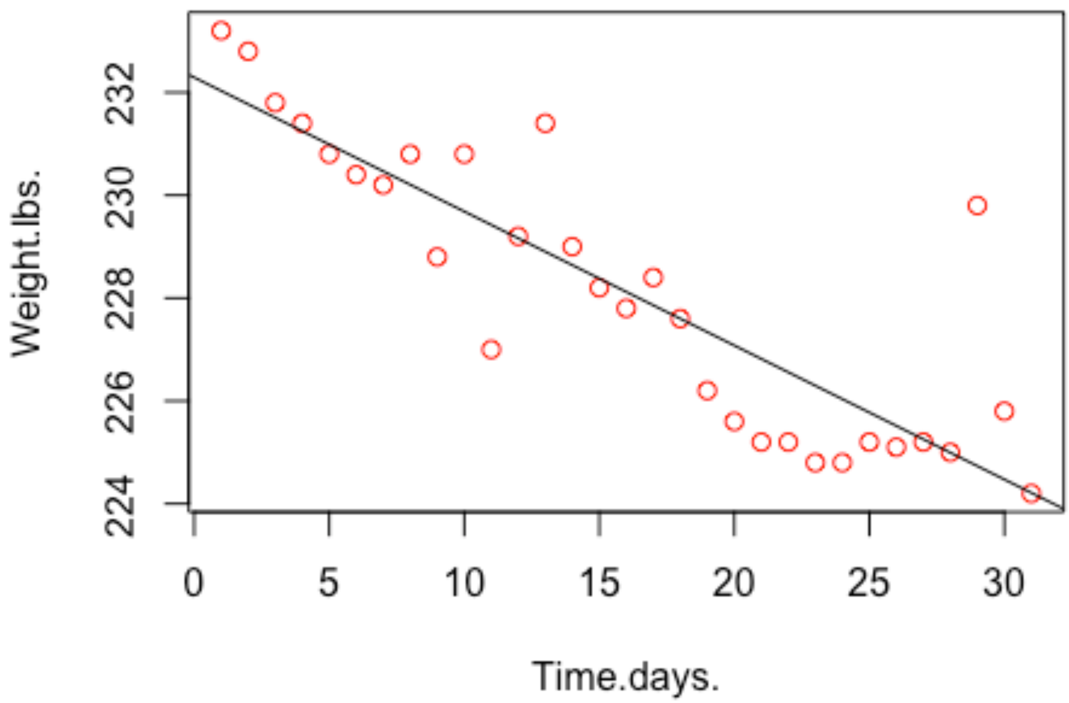
```r
# Scatter plot of weight against date
plot(Weight.lbs.~Date, data=october_data, col="red" )
```



- From the plot I can observe that there exists some type of linear relationship between weight and time in days. I will hence use linear regression to determine if I can define mapping relationship of body weight to time(days) given my activity and eating habits remain consistent with the collected data.
  Null Hypothesis(H_0): My Body weight does not have a linear trend over time, in the month of October(β = 0).

Alternative Hypothesis(H_1): My Body weight decreased linearly over time in the month of October, (β > 0)

```r
october_data$Time.days. <- seq(from=1, to=31)
# Scatter plot of weight against date
plot(Weight.lbs.~Time.days., data=october_data, col="red" )
# Liner model fit
line_fit <- lm(Weight.lbs.~Time.days., data=october_data)
# Plot linear model
abline(line_fit)
```



```r
# Summary Linear model
summary(line_fit)

##
## Call:
## lm(formula = Weight.lbs. ~ Time.days., data = october_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4230 -0.8550 -0.0514  0.4503  5.0701
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 232.29097    0.52047 446.310  < 2e-16 ***
## Time.days.   -0.26073    0.02839  -9.182 4.41e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.414 on 29 degrees of freedom
## Multiple R-squared:  0.7441,    Adjusted R-squared:  0.7353
## F-statistic: 84.32 on 1 and 29 DF,  p-value: 4.405e-10
```

- ***One-sided test Justification***: From the collected data, we have a sustained caloric difference over the month. It is hence expected for weight to decrease over time.

- ***Statistical interpretation***: The y-intercept(α) estimate is **232.29097**, and the slope(β) estimate is -**0.26073** hence the fitted model is:

  $$\boldsymbol{y_i = 232.29097 - 0.26073x_i + \epsilon_i}$$

  The correlation coefficient R-squared = 0.7441. This is closer to one, indicating that the model is a strong fit.

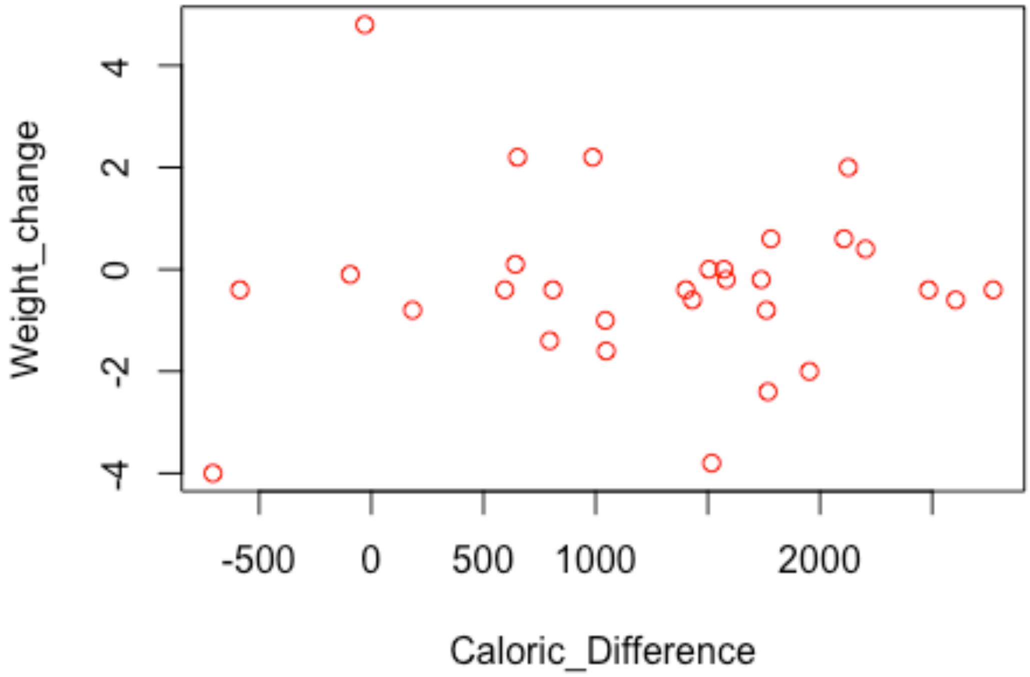
  We observe that two-sided **P-value < 0.05**, our one-sided P-value is **P-value/2** which is extremely small i.e <<< 0.05, thus the fitted model is highly significant. **We fail to reject the null hypothesis**, hence affirm the alternative, that my Body weight decreased by approximately 0.26 pounds per day over the month of October.

- **Contextual interpretation**: Maintaining a calorie deficit over the month of October resulted in an almost directly related decrease in body weight despite variance from that could stem from ignored factors like food mass, waste and water consumption.

b). Caloric deficit against Weight Change.

- According to the first law of thermodynamics, a difference between calories consumed and calories used should result in a proportional change in body mass. However, the collected data does not account for all variables that impact body mass like water and waste mass.
  Here, I am examining whether we can predict daily body weight change using the calorie deficit.

```r
# Calculate change in body weight
october_data$Weight_change <- c(NA, diff(october_data$Weight.lbs.))
# Plot change in body against calorie deficit
plot(Weight_change~Caloric_Difference, data=october_data, col="red" )
```



- From the scatter plot we cannot establish any kind of relationship between the days calculated change in body weight and the recorded calorie deficit. To test this, I am using liner regression to determine if we can predict change in body weight on a daily basis from calorie deficit.
  Null Hypothesis(H_0): There is no linear relationship between caloric deficit in a day and the change in body weight over the same period of time (β = 0).
  Alternative Hypothesis(H_1): Body weight changes proportionally to caloric deficit on a daily basis (β > 0)

```r
# Linear model fit
line_fit <- lm(Weight_change~Caloric_Difference, data=october_data)
# Model summary
summary(line_fit)

##
## Call:
## lm(formula = Weight_change ~ Caloric_Difference, data =
october_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7211 -0.5073 -0.0926  0.3709  5.0862
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -2.865e-01  5.474e-01  -0.523    0.605
## Caloric_Difference -1.076e-05  3.565e-04  -0.030    0.976
##
## Residual standard error: 1.729 on 28 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  3.252e-05, Adjusted R-squared:  -0.03568
## F-statistic: 0.0009106 on 1 and 28 DF,  p-value: 0.9761
```

- ***One-sided test Justification***: Based on the principles of thermodynamics, we expect a higher change in body weight for higher calorie deficit.

- ***Statistical interpretation***: One-sided P-value = 0.98/2 = 0.49 > 0.05, hence we fail to reject the null hypothesis. There is no linear relationship between caloric deficit in a day and the change in body weight over the same period of time.
  The correlation coefficient R-squared = 3.252e-05. This is very close to zero, indicating that the model is not strong fit.

- **Contextual interpretation**: Caloric deficit in a day does not directly correspond to body weight change during on a daily basis. This means that the data sample size may be insufficient or we cannot infer calorie difference over short time intervals(24hrs) from the weight change or the vice versa.

3. Fisher’s exact test

To meet meet my weight loss goals I had set a daily calorie deficit target of 1000 cals per day at the beginning of the month of October. Additionally I aim to have a consistent sleep schedule of about 8 hours of sleep per day.
In this section I will be using Fishers exact test to understand relationships between discrete variables; type_of_day(busy - day with classes, relaxed - no classes), sleep (good_sleep - more than 8hrs, bad_sleep - less than 8 hours) and, whether I hit my calorie deficit target.

a). Sleep vs Type of day

- Here I will examine whether having a busy or relaxed day determined if I got enough sleep or not.
  Null Hypothesis(H_0): Type of day(busy/relaxed) does not have an impact on whether I get enough sleep or not.
  Alternative Hypothesis(H_1): Type of day(busy/relaxed) determines whether I either get enough sleep or not.

```
# Add column Type_of_Sleep to the data
october_data$Type_of_Sleep <- october_data$Sleep.Hours >= 8
# Define table consisting of Type_of_Day and Type_of_Sleep
day_sleep_table <- table(october_data$Type_of_Day,
october_data$Type_of_Sleep)
# Print table
day_sleep_table

##
##           FALSE TRUE
##   Busy       11    3
##   Relaxed    10    7

cat()
# Fisher's exact test
ans <- fisher.test(day_sleep_table, alternative = "two.sided")
# Print P-value
cat("  P-value:", ans$p.value)

##   P-value: 0.280218
```

- ***Two-sided test Justification***: I could have either slept earlier when I knew I had classes during the day and possibly got enough sleep, or had less sleep as I had to wake earlier to attend classes.

- ***Statistical interpretation***: P-value = 0.28 > 0.05, hence, we fail to reject the null hypothesis. Therefore, the type of day did not impact whether I got >= 8 hours of sleep.

- **Contextual interpretation**: The results shows that from the data, there is no statistically significant evidence that I would get enough sleep during days without classes, compared to day with classes. This implies that my sleeping hours are not affected by whether I have classes to attend or the collected data sample is not sufficient.

b). Sleep vs Target Hit

- Here will examine if I hit my calorie deficit target(1000 cals) more or less frequently on busy days(had classes), compared to relaxed days(no classes).
  Null Hypothesis(H_0): Whether I hit my calorie deficit target or not, has no relationship to whether I had a busy or relaxed day.
  Alternative Hypothesis(H_1): I was more likely to hit or miss my target on busy days.

```
# Add column Target_hit to the data
october_data$Target_hit <- october_data$Caloric_Difference >= 1000
# Define table consisting of Type_of_Day and Target_hit
day_target_table <- table(october_data$Type_of_Day,
october_data$Target_hit)
# Print table
day_target_table

##
##             FALSE  TRUE
##    Busy         4    10
##    Relaxed      7    10

cat()
# Fisher's exact test
ans <- fisher.test(day_target_table, alternative = "two.sided")
# Print P-value
cat("  P-value:", ans$p.value)

##    P-value: 0.7073807
```

- ***Two-sided test Justification***: Busy days could mean more daily activity as I commute to campus, also it could mean I have less time to exercise/hit the gym. Having a busy day could thus either increase or decrease the likelihood that I hit my daily target.

- ***Statistical interpretation***: P-value = 0.71 > 0.05, hence we fail to reject the null Hypothesis. Hence, there is no significant statistical evidence that I was likely to hit or miss my Target on busy days compared to relaxed days.

- **Contextual interpretation**: From the recorded data, we cannot define a relation between having a busy day and whether I hit my target. Despite my expectation that there should be a relation between type of day and whether I hit my target, the statistical analysis indicates that either the data sample size is insufficient to determine a relationship or there exists no such relationship.

4. Likelyhood

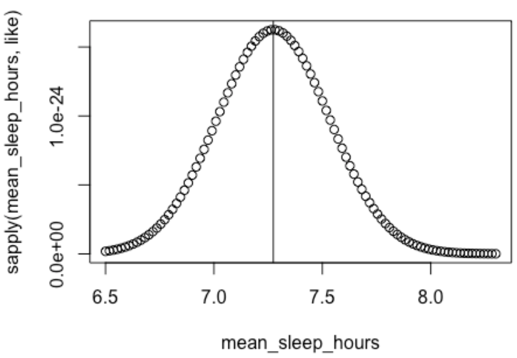
Using the collected data over the month of October, I am able to determine mean of the amount of Sleep in hours and the mean of Calorie difference on the first month of my weight loss journey.

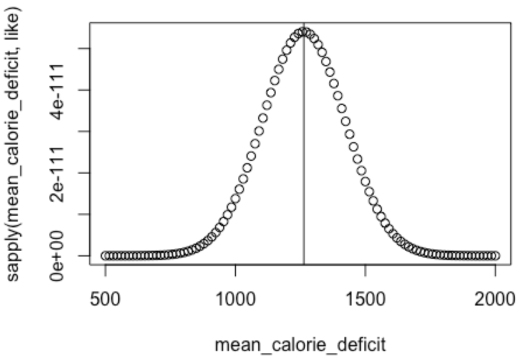
- Given that I maintain the same routine I can use the Likelihood to evaluate estimates of the mean hours of sleep and mean calorie deficit I can achieve over the rest of the months on my weight loss journey.
  In this section, I will plot the Likelihood functions, Support functions and define Credible intervals for the mean estimates.

a). Likelihood functions

- The collected data contains 31 rows, which is greater than 30 hence I have confidence that the calculated mean is close to the true mean.
  Using the mean and standard deviation, I am able to plot the likelihood functions:

```r
# Calculate mean and sd of Sleep(Hours)
sleep_data <- october_data$Sleep.Hours.
sleep_mean <- mean(sleep_data)
sleep_sd <- sd(sleep_data)
cat("Hours of sleep; mean:", sleep_mean, "  sd:", sleep_sd)

## Hours of sleep; mean: 7.273871   sd: 1.439314

# Calculate mean and sd of Caloric_Difference
calorie_deficit_data <- october_data$Caloric_Difference
calorie_deficit_mean <- mean(calorie_deficit_data)
calorie_deficit_sd <- sd(calorie_deficit_data)
cat("Caloric_Difference mean:", calorie_deficit_mean, "  sd:",
calorie_deficit_sd)

## Caloric_Difference mean: 1263.258   sd: 886.6851

# Plot Likelihood function for the mean of Sleep(Hours)
like <- function(sleep_mean){prod(dnorm(sleep_data,mean=sleep_mean,
sd=sleep_sd))}
mean_sleep_hours <- seq(from=6.5,to=8.3,len=100)
plot(mean_sleep_hours,sapply(mean_sleep_hours,like))
abline(v=mean(sleep_data))

# Plot Likelihood function for the mean of Caloric_Difference
like <-
function(calorie_deficit_mean){prod(dnorm(calorie_deficit_data,mean=calorie_deficit_mean, sd=calorie_deficit_sd))}
mean_calorie_deficit <- seq(from=500,to=2000,len=100)
plot(mean_calorie_deficit,sapply(mean_calorie_deficit,like))
abline(v=calorie_deficit_mean)
```

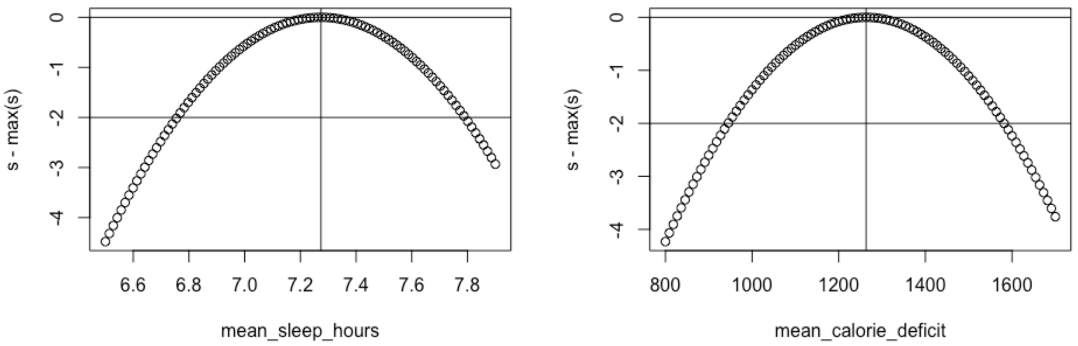




b). Support functions

- I can observe that the likelihood functions are maximized at the calculated means. I can thus plot support functions offset by the maximum values such that support=0 at max:

```
# Plot Support function for mean of Sleep(hours)
support_sleep <-
function(sleep_mean){sum(dnorm(sleep_data,mean=sleep_mean, sd=sleep_sd,
log=TRUE))}
mean_sleep_hours <- seq(from=6.5,to=7.9,len=100)
s <- sapply(mean_sleep_hours, support_sleep)
plot(mean_sleep_hours,s-max(s))
abline(v=mean(sleep_data))
abline(h=0)
abline(h=-2)

# Plot Support function for mean of Calorie deficit
support_calorie_diff <-
function(calorie_deficit_mean){sum(dnorm(calorie_deficit_data,mean=calorie_deficit_mean, sd=calorie_deficit_sd, log=TRUE))}
mean_calorie_deficit <- seq(from=800,to=1700,len=100)
s <- sapply(mean_calorie_deficit,support_calorie_diff)
plot(mean_calorie_deficit,s-max(s))
abline(v=calorie_deficit_mean)
abline(h=0)
abline(h=-2)
```



c). Credible Intervals

- In the above plots, I have added two horizontal lines at h = 0 and h = -2 that correspond to the max likelihood estimates for the means and two units of support for the mean values giving me their credible intervals.
  I can thus define the credible interval(95% likelihood interval) for the estimate of the mean of Sleep(hours) by off-setting the support graph by +2 such that the support line $\mathcal{S} = -2$ is now $\mathcal{S} = 0$ and resolving the roots as:

```
# Offset support function by +2
s <- sapply(mean_sleep_hours, support_sleep)
root_function <- function(x){support_sleep(x) - max(s) + 2}
# Calculate roots
lower_root <- uniroot( f = root_function, interval = c(6.6, 7))
upper_root <- uniroot( f = root_function, interval = c(7.6, 7.9))
cat("The credible interval for the mean of Sleep(hours) is:",
lower_root$root, " to", upper_root$root)

## The credible interval for the mean of Sleep(hours) is: 6.756844  to
7.790911
```

- Similarly the credible interval for the mean of the Calorie Difference can be calculated as follows:

```
# Offset support function by +2
s <- sapply(mean_calorie_deficit,support_calorie_diff)
root_function <- function(x){support_calorie_diff(x) - max(s) + 2}
# Calculate roots
lower_root <- uniroot( f = root_function, interval = c(800, 1000))
upper_root <- uniroot( f = root_function, interval = c(1500, 1700))
cat("The credible interval for the mean of Caloric_Difference is:",
lower_root$root, " to", upper_root$root)

## The credible interval for the mean of Caloric_Difference is:
944.7512  to 1581.765
```

Conclusion

From the statistical analysis of my fitness data over the month of October I was able to learn the following from my data:

a). Scale of time interval matters significantly.

- Daily measurements of weight showed no correlation between caloric deficit and the change in body weight, however, monthly analysis shows a strong linear trend. To properly track my weight large time intervals, i.e. bi-weekly analysis is necessary to define trends.

b). My habits are more consistent than I predicted.

- My sleeping habits and target achievement did not vary based on whether I had busy or relaxed days. Similarly, my calorie intake and activity did not differ by sleep quality, indicating that my habits are generally consistent despite changes in external circumstances.

c). Using Calorie Defficit tracking is an effective approach to Weight Loss management despite overestimation.

- Daily measurement of weight change revealed high variance which cannot be used to atomically define weight loss over short time interval. However, the fitted linear model predicts a change of -0.26lbs per day, which evaluates to 8.06lbs in 31 days. Actual loss was 9lbs -The general rule is 1lb = 3500cal thus from this rule we can calculate mean calorie deficit as (9 * 3500)/31 = 1016.13 cals.
  -The mean calorie difference from my recorded data is 1263.2 revealing that there is an overestimation of approximately 24.31%. This shows that the tracking methods were not perfect but the the fundamental strategy is validated.

d). Non-Significant results(P-value > 0.05) are still Informative

- Despite my initial concern that most of the statistical analysis reveal non-significant results, there are important inferences from this results that provide deeper understanding/lessons from the data.

The statistical analysis of my fitness data revealed a number of surprising patterns. More,importantly I was able to learn and use different methods in statistical analysis with Hypothesis testing at the core. Additionally, I was able to grasp the concept of one-sided vs two-side testing, and its significance in statistical analysis.