Statistics: Making Sense of Data
Assignment 2: Description of Data, Plots, and Summary Statistics

Welcome to your final assignment! This document provides all the information that you will need to answer the questions on Assignment 2. Below you will find a description of some data, and several tables and plots constructed from that data. Please make sure that you have all 8 pages of this document.

About the Data:

In the week 8 lectures, we looked at some data from the Canadian Trial of Carbohydrates in Diabetes (CCD). For the CCD study, the primary objective was to investigate how HbA1c, a measure of diabetic controlled, differed between subjects on a low glycemic index diet and subjects on a high glycemic index diet, where the subjects all had type 2 diabetes which was being treated with diet alone. Like most large studies, much more data were collected than what was needed to study the primary objective, and many secondary analyses were carried out. In this assignment, we will consider part of a secondary analysis.

For the CCD study, subjects were asked if they would allow some limited genetic testing to be done. The data considered in this assignment include subjects who agreed to this testing, from the low and high glycemic index diets, as well as subjects who were assigned to a high-fat diet that was also studied in parallel to the high carbohydrate diets.

In this assignment we will focus on one of the risk factors of heart disease that was not considered in the week 8 lectures. In particular, we'll look at HDL (high-density lipoprotein) cholesterol. Low levels of HDL are associated with increased risk of heart disease. In order to increase levels of HDL, dietary recommendations include modifying the type of fat consumed by reducing the intake of saturated fats and increasing the intake of polyunsaturated fats (PUFA).

The risk of heart disease is also increased by the presence of chemicals called inflammatory cytokines, such as tumour necrosis factor-alpha (TNF-alpha), which are produced in high amounts in people with type 2 diabetes. One of the ways TNF-alpha may affect heart disease risk is by modifying the way dietary fat is related to cholesterol, such as HDL. Of interest to us is two genetic polymorphisms associated with the production of TNF-alpha, at positions 238 and 308 in the DNA strand. We're interested in whether or not a subject has an A (adenine), rather than the more common G (guanine) in these positions. Based on previous research, position 308 is of particular interest.

The data we will look at in this assignment were collected on the CCD study subjects at baseline, that is before they started on the treatment diets that they were randomly assigned to as part of the study. In this assignment, we will be looking at data collected on 110 subjects.

The data include the following variables:

- sex - a categorical variable with categories male ('M') and female ('F')
- A308 - a categorical variable that is 'A' if the subject has the A genetic variant at position 308 and is 'noA' otherwise

- A238 - a categorical variable that is ‘A’ if the subject has the A genetic variant at position 238 and is ‘noA’ otherwise
- HDL - the level of HDL cholesterol
- PUFA - amount of polyunsaturated fat consumed in the subject’s regular diet, measured as the percentage of total calories consumed

Below are several tables of summary statistics and plots of the data. For this assignment, you are required to use these summary statistics and plots, and carry out tests and construct confidence intervals in order to understand the data.

Sex	Frequency	Relative Frequency
Female	58	0.527
Male	52	0.473

Table 1: Counts and relative counts of sex.

A238	Frequency	Relative Frequency
A	31	0.282
noA	79	0.718

Table 2: Counts and relative counts of A238.

A308	Frequency	Relative Frequency
A	33	0.300
noA	77	0.700

Table 3: Counts and relative counts of A308.

		Five number summary						
Variable	Number of observations	Minimum	First Quartile	Median	Third Quartile	Maximum	Mean	Standard deviation
HDL (all observations)	110	0.670	0.995	1.175	1.328	1.950	1.180	0.2599
HDL females	58	0.880	1.090	1.210	1.398	1.950	1.263	0.2494
HDL males	52	0.670	0.910	1.060	1.230	1.780	1.087	0.2411
HDL A308 is A	33	0.670	1.010	1.210	1.540	1.780	1.232	0.2998
HDL A308 is noA	77	0.680	0.990	1.140	1.240	1.950	1.158	0.2395
PUFA (all observations)	110	3.313	4.796	6.029	7.149	9.733	6.022	1.5700

Table 4: Summary statistics for the quantitative variables.

```
   Cell Contents                                           Total Observations in Table:  110
|-------------------------|
|                       N | (Frequency)
|           N / Row Total | (Row proportion)
|           N / Col Total | (Column proportion)
|         N / Table Total | (Proportion of total number of observations)
|-------------------------|


             | sex
        A308 |         F |         M | Row Total |
-------------|-----------|-----------|-----------|
           A |        18 |        15 |        33 |
             |     0.545 |     0.455 |     0.300 |
             |     0.310 |     0.288 |           |
             |     0.164 |     0.136 |           |
-------------|-----------|-----------|-----------|
         noA |        40 |        37 |        77 |
             |     0.519 |     0.481 |     0.700 |
             |     0.690 |     0.712 |           |
             |     0.364 |     0.336 |           |
-------------|-----------|-----------|-----------|
Column Total |        58 |        52 |       110 |
             |     0.527 |     0.473 |           |
-------------|-----------|-----------|-----------|
```

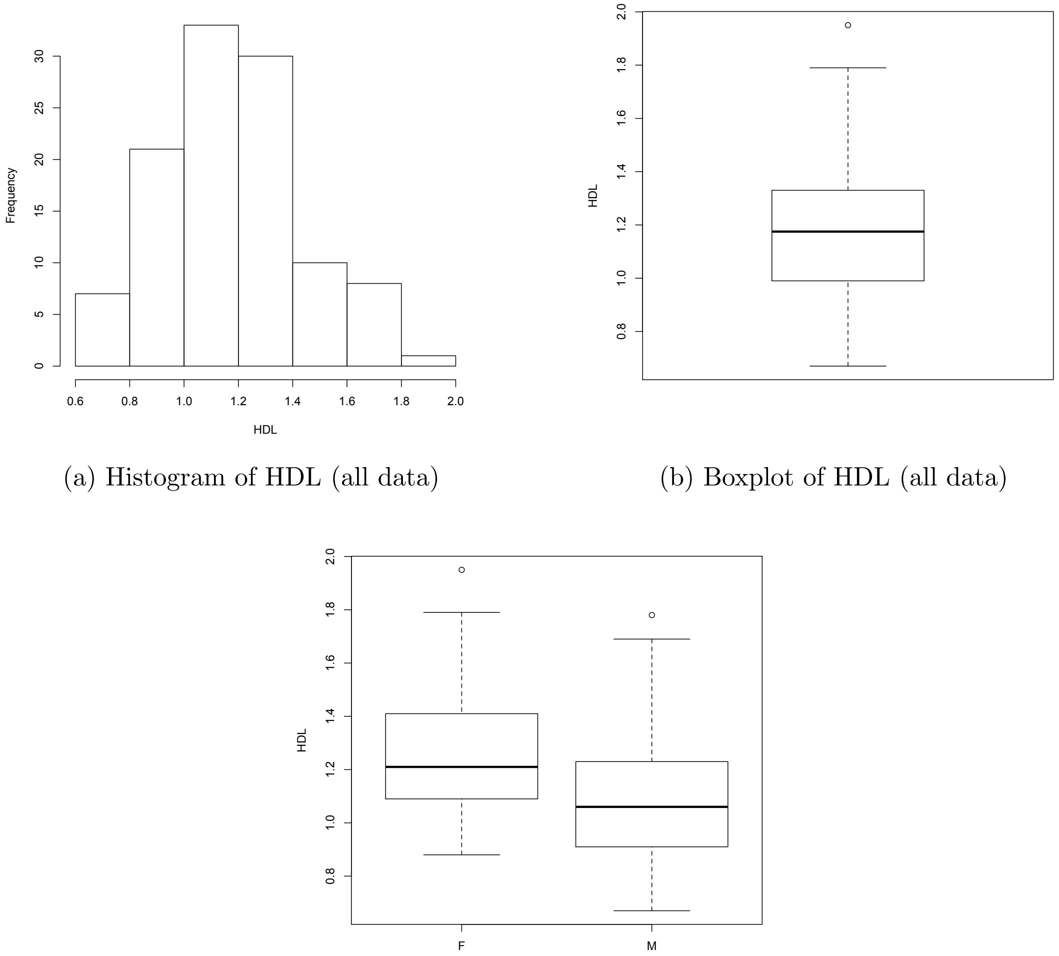
Table 5: Counts and proportions of A308, separately by sex. Note that each cell (box) contains four different numbers, giving the frequency and row proportion and column proportion and total proportion respectively (see legend at top of page); you might not need all of these numbers, but it's up to you to decide which you need.

Observations included	Correlation
All ($n = 110$)	0.030
A308 is A ($n = 33$)	-0.310
A308 is noA ($n = 77$)	0.156

Table 6: Correlations between HDL and PUFA.

	Data		
Estimate	All data	A308 is A only	A308 is noA only
n	110	33	77
R^2	0.001	0.096	0.024
b_0	1.150	1.675	1.025
S.E. of b_0	0.099	0.250	0.101
b_1	0.005	-0.072	0.022
S.E. of b_1	0.016	0.039	0.016
p-value	0.755	0.079	0.175

Table 7: Results of linear regressions of HDL on PUFA. p-values are for the test of $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$.

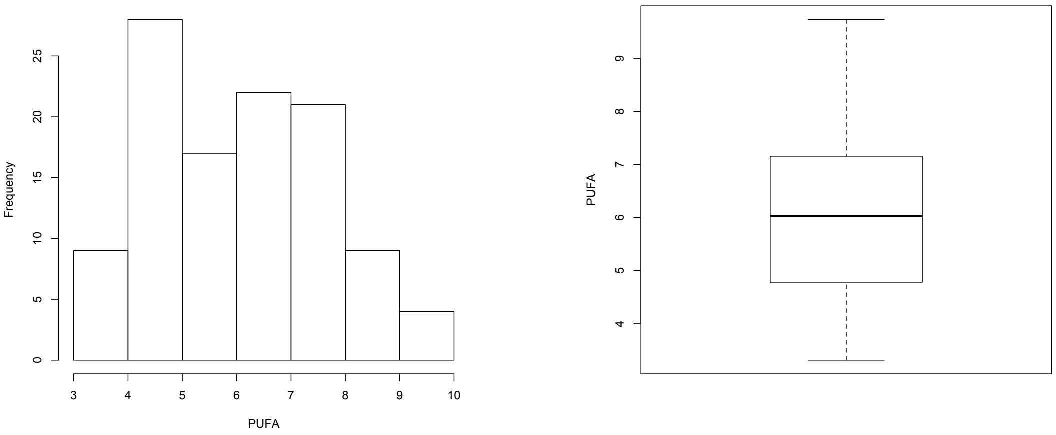


(a) Histogram of HDL (all data)

(b) Boxplot of HDL (all data)

(c) Boxplots of HDL for females and males

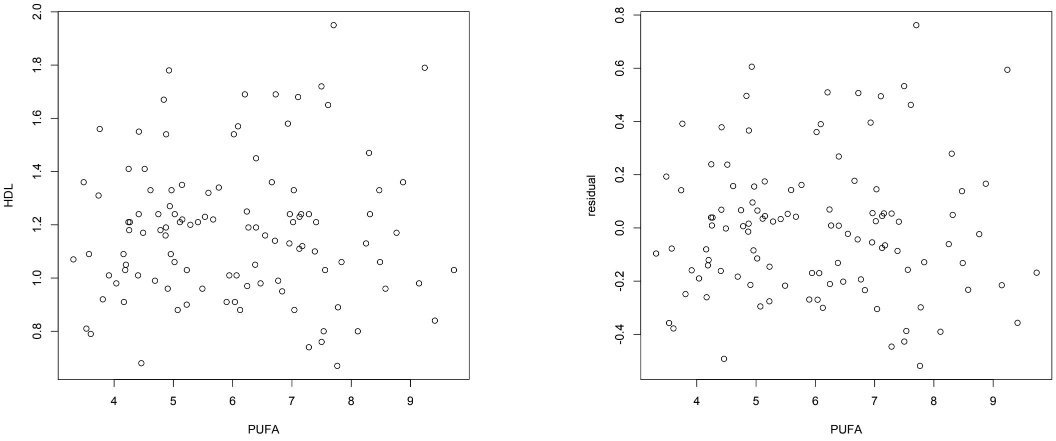
Figure 1: Plots of HDL.



(a) Histogram

(b) Boxplot

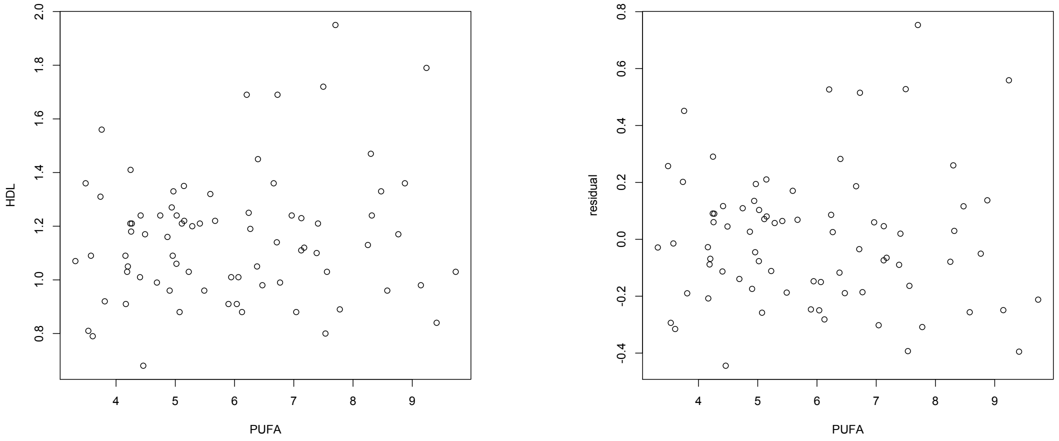
Figure 2: Plots of PUFA (all data).



(a) Scatterplot of HDL versus PUFA

(b) Residuals versus PUFA for regression of HDL on PUFA

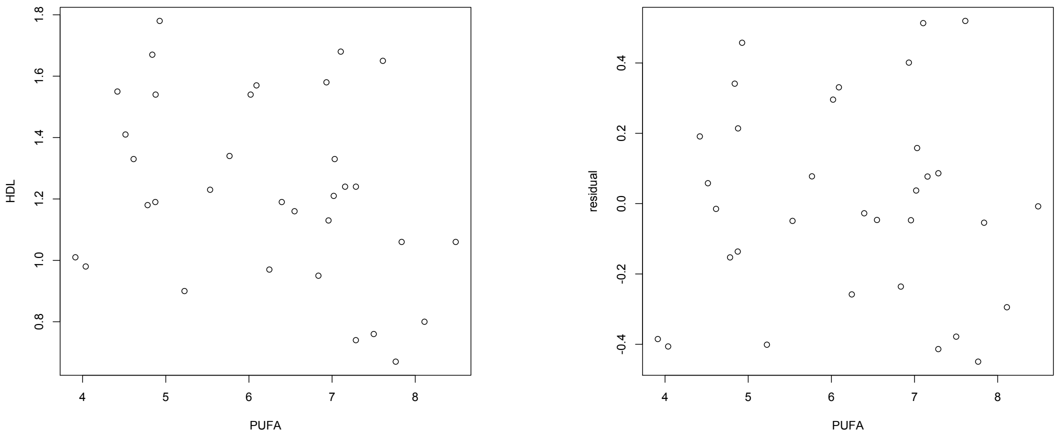
Figure 3: Plots related to the relationship between HDL and PUFA for all of the data.



(a) Scatterplot of HDL versus PUFA for subjects without the A variant at 308

(b) Residuals versus PUFA for regression of HDL on PUFA for subjects without the variant at 308

Figure 4: Plots related to the relationship between HDL and PUFA for subjects without the A variant at 308 (A308=noA).



(a) Scatterplot of HDL versus PUFA for subjects with A variant at 308

(b) Residuals versus PUFA for regression of HDL on PUFA for subjects with A variant at 308

Figure 5: Plots related to the relationship between HDL and PUFA for subjects with A variant at 308 (A308=A).