STA 135

Nutrition Data Analysis

Final Project

Jared Yu
6-10-2019

1. Introduction

The goal of this analysis is to examine beef and vegetable products through their nutritional characteristics. Different multivariate approaches will be used and additionally there will be comparisons with univariate counterparts. The multivariate tools include: simultaneous confidence intervals based on T^2 and Bonferroni correction, two-sample Hotelling's T^2 test, principal component analysis, and linear discriminant analysis. The idea is to understand better these various new multivariate statistical methods that have been learned over the quarter in STA 135 using a dataset that has been chosen.

2. Summary

The dataset comes from the website https://data.world and has been collected from the United States Department of Agriculture. It originally contains 8,618 different observations, each of which are different foods that have been placed into groups such as: Beef Products, Vegetables and Vegetable Products, Snacks, Beverages, Sweets, etc. The two largest food groups, Beef Products and Vegetables and Vegetable Products have been chosen as the two populations to examine from the dataset. The columns for each of the different food observations contain nutritional information for 100 grams of each food. There are 45 total columns in the dataset, of which 23 will be utilized as nutritional descriptions. These nutritional descriptions contain information such as Protein (g), Fat (g), Vitamin B12 (mcg), etc. for each food.

The first test on the dataset was to check the assumption of multivariate normality for the data. Normality is an important assumption for the tests that will be performed on the dataset. Although no specific multivariate test has been covered in class, three different more well-known methods have been used. These include: Mardia's test, Henze-Zirkler's multivariate normality test, and Royston's multivariate test. All three tests were performed using the MVN package in R, and all the results came back saying that the dataset is not normally distributed. Therefore, this will be accounted for in the interpretation of the results for all the tests since the tests themselves require the normality assumption to hold.

Sample estimates of the two populations have also been taken. $Fig.$ 1 shows the sample mean of each column for both Beef and Vegetable data. The data consists of nutrient data which can lead to extremely small values and not all of them have the same measurement. The largest number belongs to the Energy column which is measured in kcal, while other nutrients may be measured in either grams or micrograms.

	Sample Mean	Nutrients
1	143.2666	Energy kcal
2	14.3084	Protein g
3	7.0516	Fat g
4	5.8901	Carb g
5	1.0835	Sugar g
6	1.2762	Fiber g
7	96.4087	VitA mcg
8	0.3123	VitB6 mg
9	1.6417	VitB12 mcg
10	12.2861	VitC mg
11	0.3173	VitE mg
12	26.0817	Folate mcg
13	3.1786	Niacin mg
14	0.1883	Riboflavin mg
15	0.0974	Thiamin mg
16	29.9464	Calcium mg
17	0.1560	Copper mcg
18	2.0013	Iron mg
19	25.6454	Magnesium mg
20	0.5252	Manganese mg
21	132.8230	Phosphorus mg
22	13.7814	Selenium mcg
23	3.1105	Zinc mg

Fig. 1 Combined sample mean of nutrients for both populations of beef and vegetables.

$Fig.$ 2 shows the sample mean split between the two populations. After splitting the data into the two distinct food groups, it is apparent that their characteristics are quite different. The beef food groups contains high amounts in nutrients that are relevant to beef and vice versa for vegetables.

	Beef	Vegetables
Energy kcal	212.7495	63.8816
Protein g	24.3697	2.8133
Fat g	12.3614	0.9852
Carb g	0.1085	12.4957
Sugar g	0	2.3214
Fiber g	0.0029	2.7309
VitA mcg	78.8816	116.4336
VitB6 mg	0.4348	0.1724
VitB12 mcg	3.0737	0.0057
VitC mg	0.381	25.8878
VitE mg	0.2144	0.4349
Folate mcg	7.074	47.7983
Niacin mg	4.8855	1.2285
Riboflavin mg	0.2326	0.1376
Thiamin mg	0.0748	0.1232
Calcium mg	12.8795	49.4457
Copper mcg	0.1543	0.158
Iron mg	2.522	1.4064
Magnesium mg	20.2135	31.8514
Manganese mg	0.6927	0.3338
Phosphorus mg	198.3795	57.9239
Selenium mcg	24.483	1.5547
Zinc mg	5.4206	0.4711

Fig. 2 Sample mean separated into two populations, beef and vegetable products.

Along with the sample mean and variance covariance, summary statistics containing information about quantiles, min, and max have been created for both Beef and Vegetable data ($Fig. 3a$ and $Fig. 3b$). The summary statistics show that not only are there rather low values for nutrients that are only relevant to one of the two food groups, it is likely that there are many which are largely zeroes. The sugar column for beef is an example where there are only zeroes and so the column will be removed later for the purpose of vital calculations such as the sample variance-covariance matrix. To match with the beef data, the sugar column will also be removed from the vegetable data.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Energy kcal	77	160.25	199	212.7495	246	854
Protein g	1.5	20.965	24.27	24.3697	28.1575	36.12
Fat g	1.98	6.16	9.46	12.3614	16.31	94
Carb g	0	0	0	0.1085	0	7.89
Sugar g	0	0	0	0	0	0
Fiber g	0	0	0	0.0029	0	1.4
VitA mcg	0	0	2	78.8816	4	28319
VitB6 mg	0	0.32	0.413	0.4348	0.5808	1.083
VitB12 mcg	0	1.7325	2.405	3.0737	3.15	96
VitC mg	0	0	0	0.381	0	50.3
VitE mg	0	0.0725	0.17	0.2144	0.31	2.09
Folate mcg	0	3	7	7.074	8	290
Niacin mg	0	3.74	4.7775	4.8855	5.9605	17.525
Riboflavin mg	0	0.15	0.194	0.2326	0.247	3.425
Thiamin mg	0	0.06	0.072	0.0748	0.085	0.559
Calcium mg	2	6	11	12.8795	16	485
Copper mcg	0	0.069	0.084	0.1543	0.105	14.588
Iron mg	0.17	1.91	2.35	2.522	2.83	44.55
Magnesium mg	0	18	22	20.2135	24	49
Manganese mg	0	0.005	0.011	0.6927	0.014	328
Phosphorus mg	0	177	204	198.3795	222	497
Selenium mcg	0	20.525	24.75	24.483	30.775	168
Zinc mg	0	4.01	5.11	5.4206	6.8475	12.28

Fig. 3a Summary statistics for the beef products.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Energy kcal	3	23	36	63.8816	82	372
Protein g	0	1.14	1.985	2.8133	3.0925	57.47
Fat g	0	0.16	0.3	0.9852	0.56	18.71
Carb g	0.38	4.35	7.3	12.4957	17.3175	85.51
Sugar g	0	0	1.155	2.3214	2.8825	43.9
Fiber g	0	1.2	2	2.7309	3.1	70.1
VitA mcg	0	0	10	116.4336	83.5	3863
VitB6 mg	0	0.069	0.116	0.1724	0.1885	4.228
VitB12 mcg	0	0	0	0.0057	0	2.25
VitC mg	0	4.8	11	25.8878	25.85	1900
VitE mg	0	0	0.07	0.4349	0.49	12.25
Folate mcg	0	12	23	47.7983	49	5881
Niacin mg	0	0.418	0.6915	1.2285	1.2	127.5
Riboflavin mg	0	0.036	0.064	0.1376	0.114	17.5
Thiamin mg	0	0.034	0.061	0.1232	0.1	23.375
Calcium mg	0	14	27	49.4457	52	813
Copper mcg	0	0.051	0.091	0.158	0.1572	6.1
Iron mg	0	0.46	0.76	1.4064	1.39	66.38
Magnesium mg	0	13	20	31.8514	33	770
Manganese mg	0	0.1288	0.212	0.3338	0.37	6.704
Phosphorus mg	0	27	43	57.9239	70	548
Selenium mcg	0	0.4	0.7	1.5547	0.9	46.1
Zinc mg	0	0.2	0.31	0.4711	0.52	7.66

Fig. 3b Summary statistics for the vegetable products.

Additionally, a bar plot ($Fig.4$) and distribution table ($Fig.5$) have been created to show the count and population of several food groups to gain a better understanding of the dataset. It is

apparent that the dataset consists largely of beef, vegetables, and baked goods. Other foods are considerably less well represented in terms of proportion of the dataset.

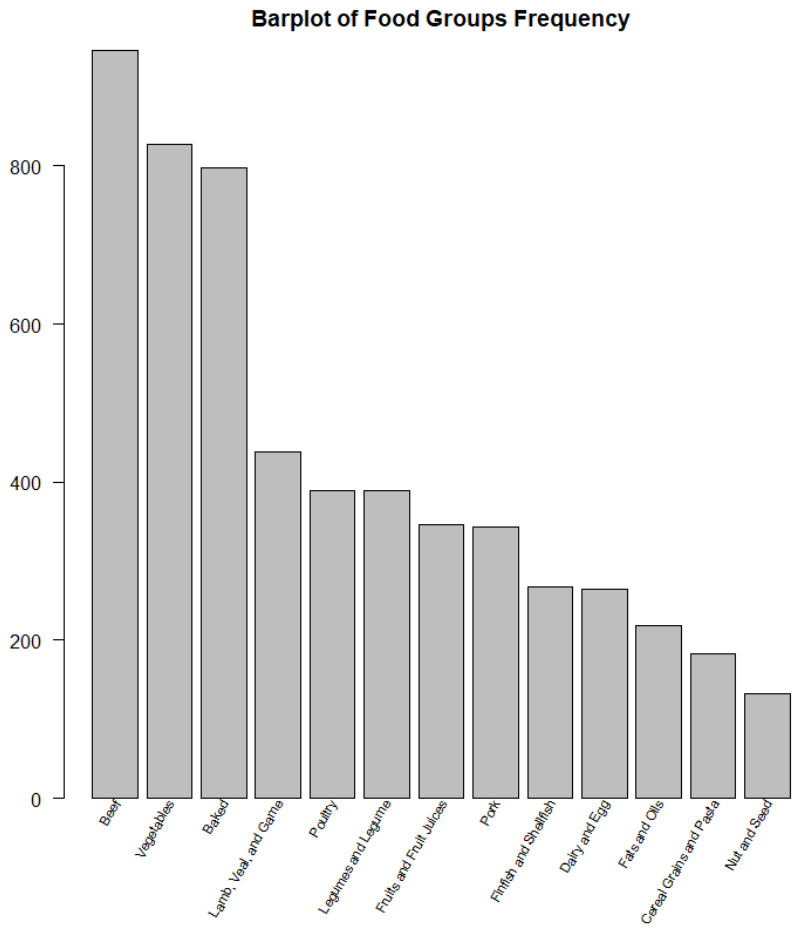

Fig. 4 Bar plot of the frequency of some of the food groups in the dataset.

Percentage and Count of Food Groups

	Food Groups	Percentage (%)	Count
1	Beef Products	17.07	946
2	Vegetables and Vegetable Products	14.94	828
3	Baked Products	14.38	797
4	Lamb, Veal, and Game Products	7.90	438
5	Poultry Products	7.04	390
6	Legumes and Legume Products	7.02	389
7	Fruits and Fruit Juices	6.24	346
8	Pork Products	6.19	343
9	Finfish and Shellfish Products	4.82	267
10	Dairy and Egg Products	4.76	264
11	Fats and Oils	3.95	219
12	Cereal Grains and Pasta	3.30	183
13	Nut and Seed Products	2.40	133

Fig. 5 Distribution table showing the count and percentage of some of the food groups within the dataset.

A set of paired boxplots have been created for each column, comparing the distribution of each nutrient for both populations (*Fig*. 6). Through the boxplot, outliers that were far beyond any other groups of data were removed. The set of box plots shows how there are times where one food groups contains a considerable amount of a nutrient while the other food groups lacks it. At other times they both contain relatively close amounts of the nutrient.

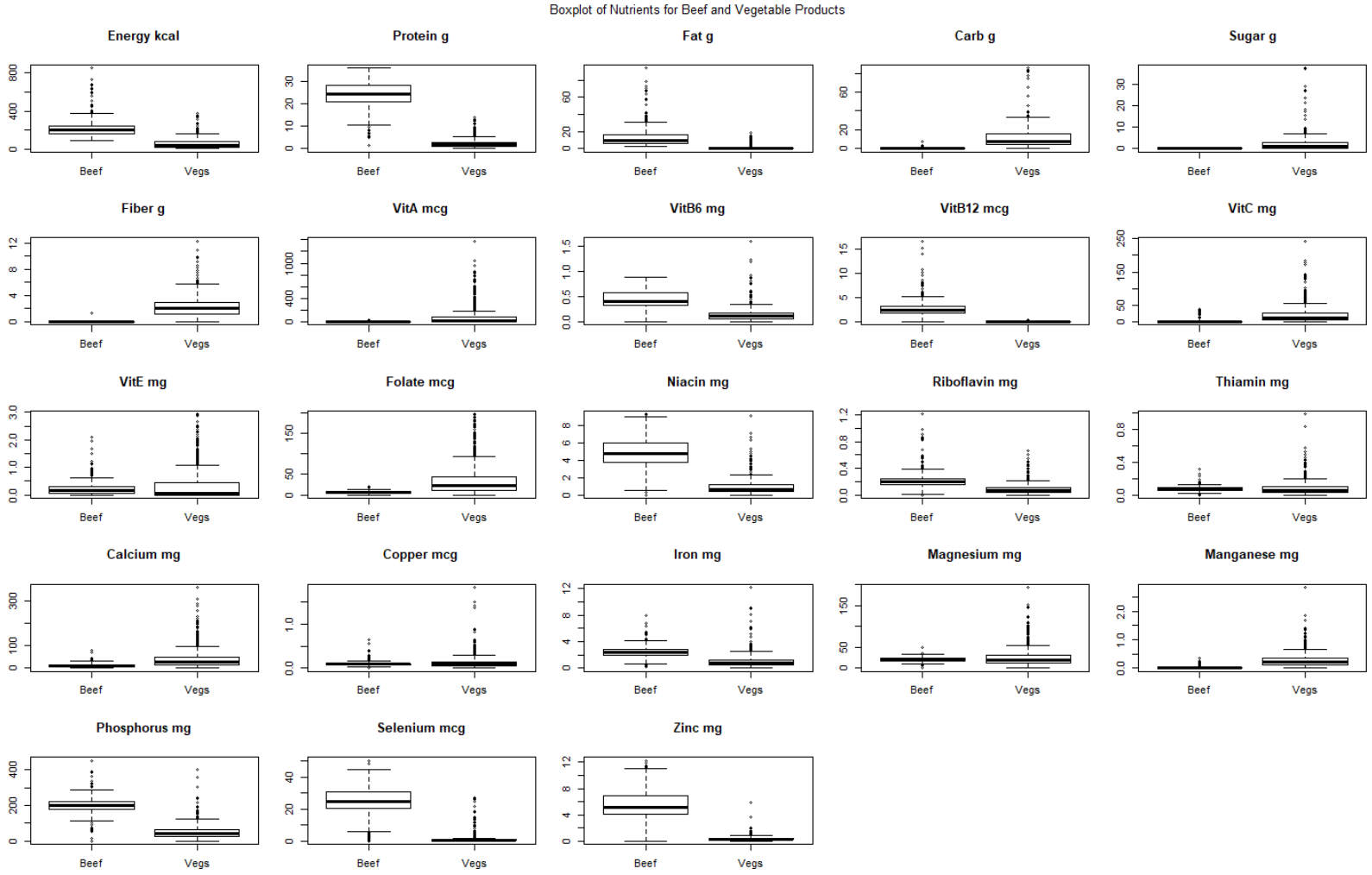

Fig. 6 Combined box plot comparing the nutrients for both beef and vegetable populations. Outliers have been removed.

3. Analysis

Next the goal is to do a one-sample inference based on only the filtered beef data. The column for sugar consists of all zeroes and so it was removed. The reason is that this would lead to errors during the process of calculating the sample variance-covariance matrix.

3.1.1 Univariate One-at-a-time testing

The first of the one-sample tests are the univariate one-at-a-time hypothesis test. The following is the hypothesis for the univariate case:

$$H_0: \mu = \mu_0 \; vs. \; H_1: \mu \neq \mu_0.$$

In this case, the μ_0 that has been chosen is the sample mean rounded to the nearest whole number. The reason is that the nutrients themselves don't have a specific null mean that would be sensible to have as a comparison. The p-values of the tests can be seen in $Fig. 7a$.

	Nutrient	p-value
1	Energy kcal	0.8804
2	Protein g	0.0080
3	Fat g	0.1772
4	Carb g	0.0000
5	Fiber g	0.1577
6	VitA mcg	0.0080
7	VitB6 mg	0.0000
8	VitB12 mcg	0.0000
9	VitC mg	0.0022
10	VitE mg	0.0000
11	Folate mcg	0.2777
12	Niacin mg	0.0038
13	Riboflavin mg	0.0000
14	Thiamin mg	0.0000
15	Calcium mg	0.4591
16	Copper mcg	0.0000
17	Iron mg	0.0000
18	Magnesium mg	0.1961
19	Manganese mg	0.0000
20	Phosphorus mg	0.9078
21	Selenium mcg	0.9475
22	Zinc mg	0.0000

Fig. 7a p-values of univariate confidence intervals.

Only eight of the nutrients had a p-value greater than 0.05. The eight nutrients are: Energy, Fat, Fiber, Folate, Calcium, Magnesium, Phosphorus, and Selenium. So, only these variables are insignificant at 5% significance level. The confidence intervals of the nutrients have also been created; they can be seen in $Fig. 7b$.

	Lower Bound	Upper Bound
Energy.kcal	208.4009	218.8012
Protein.g	24.1156	24.7664
Fat.g	11.8042	13.0599
Carb.g	0.0597	0.1139
Fiber.g	-0.0011	0.0069
VitA.mcg	2.4074	2.9109
VitB6.mg	0.4241	0.4472
VitB12.mcg	2.4711	2.6612
VitC.mg	0.0973	0.4431
VitE.mg	0.1941	0.2216
Folate.mcg	5.9002	6.3471
Niacin.mg	4.7169	4.9453
Riboflavin.mg	0.2002	0.2146
Thiamin.mg	0.0708	0.0745
Calcium.mg	11.7162	12.6279
Copper.mcg	0.0862	0.092
Iron.mg	2.3555	2.4459
Magnesium.mg	19.8738	20.6144
Manganese.mg	0.0156	0.0197
Phosphorus.mg	192.796	199.6061
Selenium.mcg	23.3505	24.6073
Zinc.mg	5.3209	5.6108

Fig. 7b Confidence intervals for univariate tests.

3.1.2 Simultaneous testing using Hotelling's T^2 and Bonferroni's corrected method

Then the same process was repeated using both Hotelling's T^2 and Bonferroni to do simultaneous confidence intervals. The Hotelling's T^2 and Bonferroni methods test the following hypothesis test:

$$H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0 \; vs. \; H_1: \boldsymbol{\mu} \neq \boldsymbol{\mu}_0,$$

where $\boldsymbol{\mu}$ is a vector of the population means for each of the nutrients and $\boldsymbol{\mu}_0$ is a vector of the same null mean used in the univariate test. The null hypothesis for the Hotelling's T^2 is rejected if the test statistic T^2 has the following result,

$$T^2 = n(\bar{x} - \mu_0)' S^{-1} (\bar{x} - \mu_0) > \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha).$$

The value for T^2 is 71,181.91 while the critical value is approximately 34.98, therefore we reject the null hypothesis at significance level 5%. The confidence interval for each of the nutrients under the Hotelling's T^2 can be seen in $Fig. 8a$.

	Lower Bound	Upper Bound
Energy kcal	197.9306	229.2716
Protein g	23.4604	25.4216
Fat g	10.5401	14.324
Carb g	0.0052	0.1685
Fiber g	-0.0092	0.015
VitA mcg	1.9006	3.4177
VitB6 mg	0.4008	0.4704
VitB12 mcg	2.2797	2.8526
VitC mg	-0.2509	0.7913
VitE mg	0.1663	0.2493
Folate mcg	5.4504	6.7969
Niacin mg	4.487	5.1753
Riboflavin mg	0.1858	0.229
Thiamin mg	0.067	0.0782
Calcium mg	10.7983	13.5457
Copper mcg	0.0803	0.0979
Iron mg	2.2644	2.537
Magnesium mg	19.1283	21.3599
Manganese mg	0.0115	0.0238
Phosphorus mg	185.9401	206.462
Selenium mcg	22.0853	25.8726
Zinc mg	5.0291	5.9027

Fig. 8a Confidence interval using Hotelling's T^2.

Similarly, the null hypothesis for the Bonferroni is rejected if,

$$\max_{1\leq j\leq p} \frac{\left|\bar{X}_j - \mu_j\right|}{\frac{S_j}{\sqrt{n}}} \geq t_{n-1}\left(\frac{\alpha}{2p}\right),$$

where p is the number of variables, S is the sample variance-covariance matrix and $\bar{X}$ is the vector of sample means. The p-value of this test is 0 and so we reject the null hypothesis under the Bonferroni corrected test. The confidence interval for each of the nutrients under the Bonferroni's method can be seen in $Fig. 8b$.

	Lower Bound	Upper Bound
Energy.kcal	205.4914	221.7107
Protein.g	23.9335	24.9484
Fat.g	11.453	13.4112
Carb.g	0.0446	0.1291
Fiber.g	-0.0034	0.0092
VitA.mcg	2.2666	3.0517
VitB6.mg	0.4176	0.4536
VitB12.mcg	2.418	2.7144
VitC.mg	5e-04	0.5399
VitE.mg	0.1864	0.2293
Folate.mcg	5.7752	6.4721
Niacin.mg	4.653	5.0092
Riboflavin.mg	0.1962	0.2186
Thiamin.mg	0.0697	0.0755
Calcium.mg	11.4611	12.8829
Copper.mcg	0.0845	0.0936
Iron.mg	2.3301	2.4713
Magnesium.mg	19.6667	20.8215
Manganese.mg	0.0144	0.0208
Phosphorus.mg	190.8909	201.5112
Selenium.mcg	22.9989	24.9589
Zinc.mg	5.2398	5.6919

Fig. 8b Confidence interval using Bonferroni's corrected method.

3.2 Confidence Region of Protein and Fat using Hotelling's T^2 and Bonferroni's method

Additionally, the simultaneous confidence region was done using the Protein and Fat columns from the filtered beef data. The plot in $Fig. 8$ shows the confidence region of the two variables using both Hotelling's T^2 and Bonferroni's corrected method.

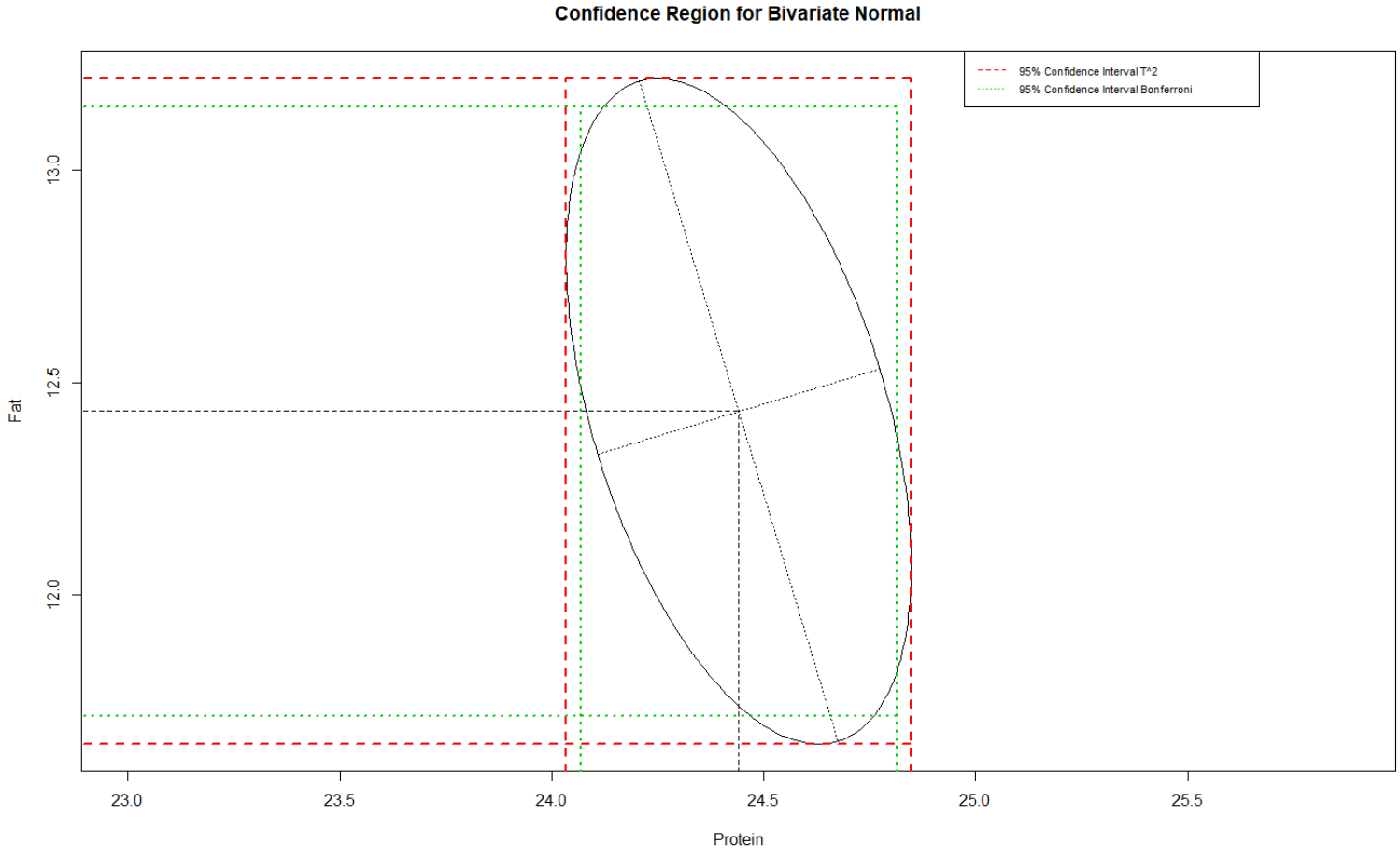

Fig. 8 Confidence region plot using both Hotelling's T^2 and Bonferroni's corrected method.

The plot shows the 95% probability that both variables' true means are contained in the confidence region. As expected, the Bonferroni method creates a tighter region than Hotelling's T^2 method.

3.3 Two-sample Hotelling's T^2 test

Afterwards a two-sample Hotelling's T^2 using both the beef and vegetable data was done. The following is the hypothesis test for the Hotelling's T^2,

$$H_0: \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \vec{0},$$

where $\boldsymbol{\mu}_1$ is the population mean for beef nutrients and $\boldsymbol{\mu}_2$ is the population mean for vegetable nutrients. The T^2 test statistic is 1,821.6 and a p-value of 2.2×10^{-16}. Therefore, the null hypothesis that the two samples are the same is rejected. Since the null hypothesis was rejected, simultaneous confidence intervals using both Hotelling's T^2 and Bonferroni's methods were created to check the significant components. The two sets of confidence intervals can be seen in $Fig. 9a$ and $Fig. 9b$.

	Lower Bound	Upper Bound
Energy kcal	135.8509	175.6762
Protein g	20.8508	23.0999
Fat g	9.4478	13.6002
Carb g	-13.2328	-9.098
Fiber g	-2.5563	-1.9175
VitA mcg	-131.2567	-58.9868
VitB6 mg	0.2445	0.3373
VitB12 mcg	2.2559	2.8716
VitC mg	-24.9345	-14.6943
VitE mg	-0.2617	-0.021
Folate mcg	-37.3971	-22.6176
Niacin mg	3.4753	4.303
Riboflavin mg	0.0896	0.1473
Thiamin mg	-0.028	0.0083
Calcium mg	-40.922	-21.5349
Copper mcg	-0.0689	-0.0051
Iron mg	1.0866	1.6002
Magnesium mg	-10.3713	-1.3597
Manganese mg	-0.3138	-0.2105
Phosphorus mg	130.3396	157.5461
Selenium mcg	20.627	24.8508
Zinc mg	4.5926	5.5428

Fig. 9a Simultaneous confidence interval using Hotelling's T^2 for testing two-samples.

	Lower Bound	Upper Bound
Energy kcal	145.3986	166.1285
Protein g	21.39	22.5607
Fat g	10.4433	12.6047
Carb g	-12.2416	-10.0893
Fiber g	-2.4031	-2.0707
VitA mcg	-113.9307	-76.3128
VitB6 mg	0.2667	0.3151
VitB12 mcg	2.4035	2.724
VitC mg	-22.4795	-17.1493
VitE mg	-0.204	-0.0787
Folate mcg	-33.8538	-26.1608
Niacin mg	3.6738	4.1045
Riboflavin mg	0.1034	0.1335
Thiamin mg	-0.0193	-4e-04
Calcium mg	-36.2742	-26.1828
Copper mcg	-0.0536	-0.0204
Iron mg	1.2097	1.477
Magnesium mg	-8.2108	-3.5201
Manganese mg	-0.289	-0.2352
Phosphorus mg	136.8621	151.0236
Selenium mcg	21.6396	23.8381
Zinc mg	4.8204	5.315

Fig. 9b Simultaneous confidence interval using Bonferroni's for testing two-samples.

The result is that only Thiamin from the Hotelling's T^2 confidence interval contains zero. This implies that the beef and vegetable populations are significantly different.

3.4 Principal component analysis (PCA)

Principal component analysis was also done on just the beef population. The sample has twenty-two variables and so twenty-two principal components were created. A scree plot showing the size of each component has been plotted alongside a cumulative proportion plot ($Fig.10$). On the right-hand side of the same plot, a proportion of variance plot shows the cumulative proportion of variance accounted for by each additional eigenvalue that is included. The scree plot shows that the elbow could be around component 3, but the cumulative proportion plot shows that the 80% mark would be around principal component 7. It can be said that a good cutoff for the first two principal components would be around 80%, however with twenty-two different variables, having already over 40% of the variation being accounted for is quite good for the first two principal components.

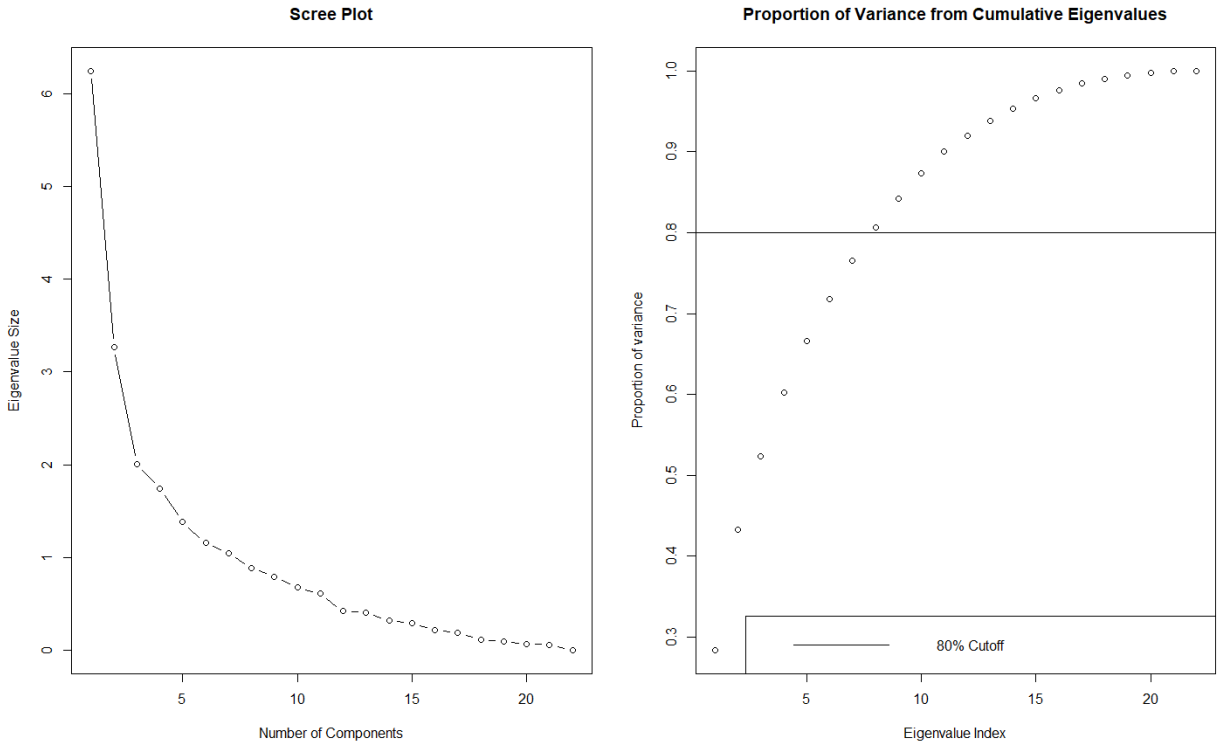

Fig. 10 Scree plot and proportion of variance from cumulative eigenvalues.

In $Fig.11a$, the loadings on the first three principal components are shown. The loadings below indicate how much of each variable contribute towards the determination of each principal component. The loadings of nutrients which beef is known to contain have higher magnitude than other nutrients which are less well-known to be in beef.

	Comp.1	Comp.2	Comp.3
Energy kcal	0.1832	0.2232	0.0897
Protein g	-0.2544	-0.1051	-0.2125
Fat g	0.231	0.2303	0.1218
Carb g	0.0137	0.2329	0.4341
Fiber g	-0.033	0.1993	0.5089
VitA mcg	0.1157	0.2055	-0.031
VitB6 mg	-0.2366	-0.2975	0.1684
VitB12 mcg	-0.228	0.2867	-0.1313
VitC mg	-0.0011	0.1194	-0.0356
VitE mg	-0.0065	-0.0957	-0.0788
Folate mcg	-0.2169	-0.1601	0.1764
Niacin mg	-0.2535	-0.3042	0.1733
Riboflavin mg	-0.257	0.2686	-0.1262
Thiamin mg	-0.2502	0.2069	0.2113
Calcium mg	-0.0511	-0.1948	0.3781
Copper mcg	-0.2445	0.2981	-0.1132
Iron mg	-0.2395	0.2708	-0.268
Magnesium mg	-0.3027	0.0388	0.1104
Manganese mg	-0.0985	0.31	0.2448
Phosphorus mg	-0.3287	-0.0894	-0.004
Selenium mcg	-0.3164	-0.0838	0.0312
Zinc mg	-0.2377	0.0996	-0.0897

Fig. 11a Loadings on the first three principal components of the beef data.

Below the first biplot shows the first and second principal components for the beef data ($Fig. 11b$). There is a main group of points which show that most of the scores of the data are quite similar in terms of the first and second principal components. There are however several groups of outliers on the left and right side of the main group. On the right side, there are two groups which are mainly Australian Wagyu beef. In the left there are mainly boneless plate steak. On the extreme at the top left there are beef patties which have a nutrient content different from any specific cut of beef. Therefore, it makes sense that their content is more extreme in terms of its nutritional difference than the other types of beef. The main group can be further sub-divided as it is evident that there are further groups, but it has not been done for the purpose of the analysis.

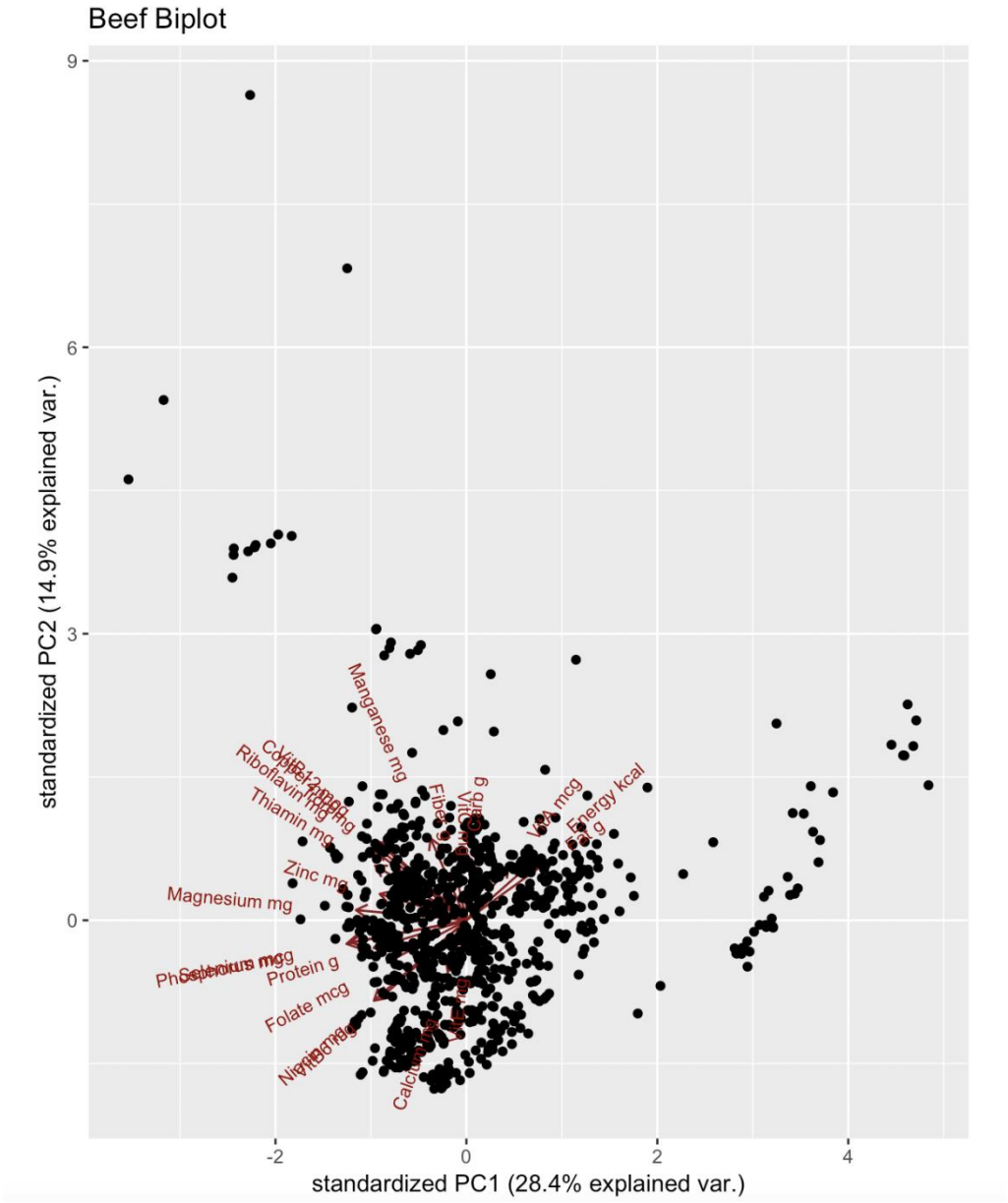

Fig. 11b Biplot of the beef data.

Next is another biplot showing the scores of the beef and vegetable data for the first two principal components ($Fig. 11c$). Certain extreme outliers have been removed so that the plot is simpler to view. Also, the principal component code was re-run using the subset of the data. The data seems to form a rather apparent V-shape, with vegetables taking up the bottom and left portions of the plot, with beef making up the top right. There seems to be some slight overlap, but the ellipse for the groups do not overlap.

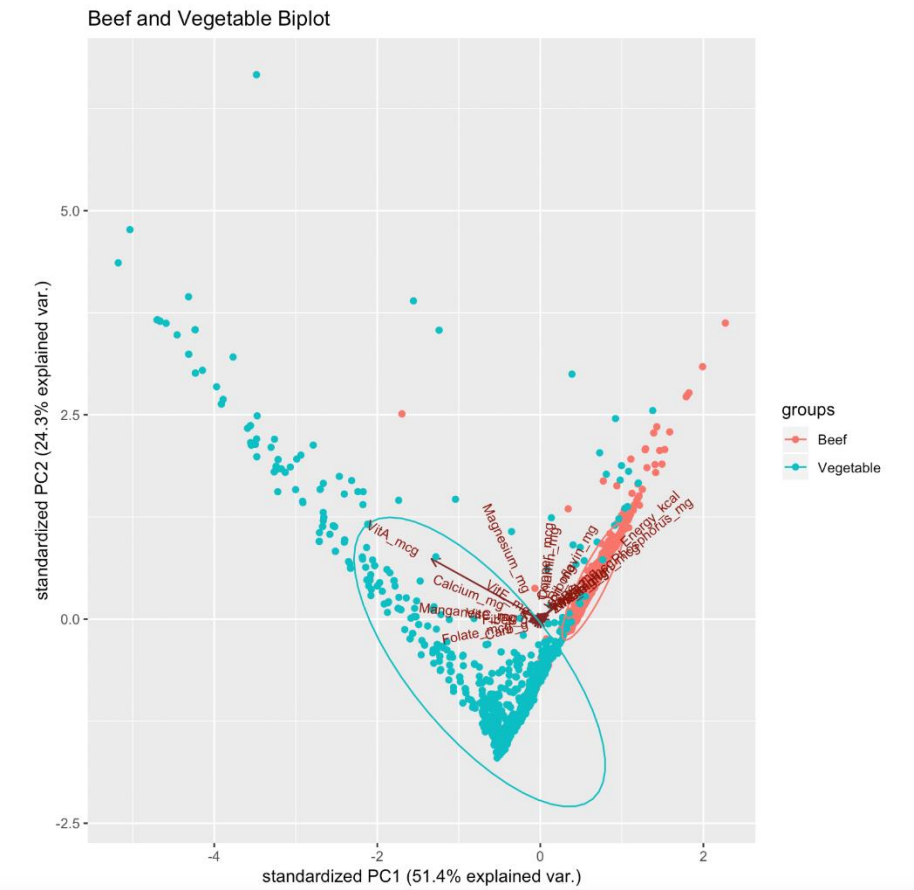

Fig. 11c Biplot of beef and vegetable data.

3.5 Linear Discriminant Analysis (LDA)

Linear discriminant analysis will also be used on the data to try and see how the different food groups can be classified into their respective populations. The two variables that were chosen are magnesium and energy. The reason is that both beef and vegetables have a decent amount of both nutrients and therefore it would be useful to use them both in determining the food group based on these two variables.

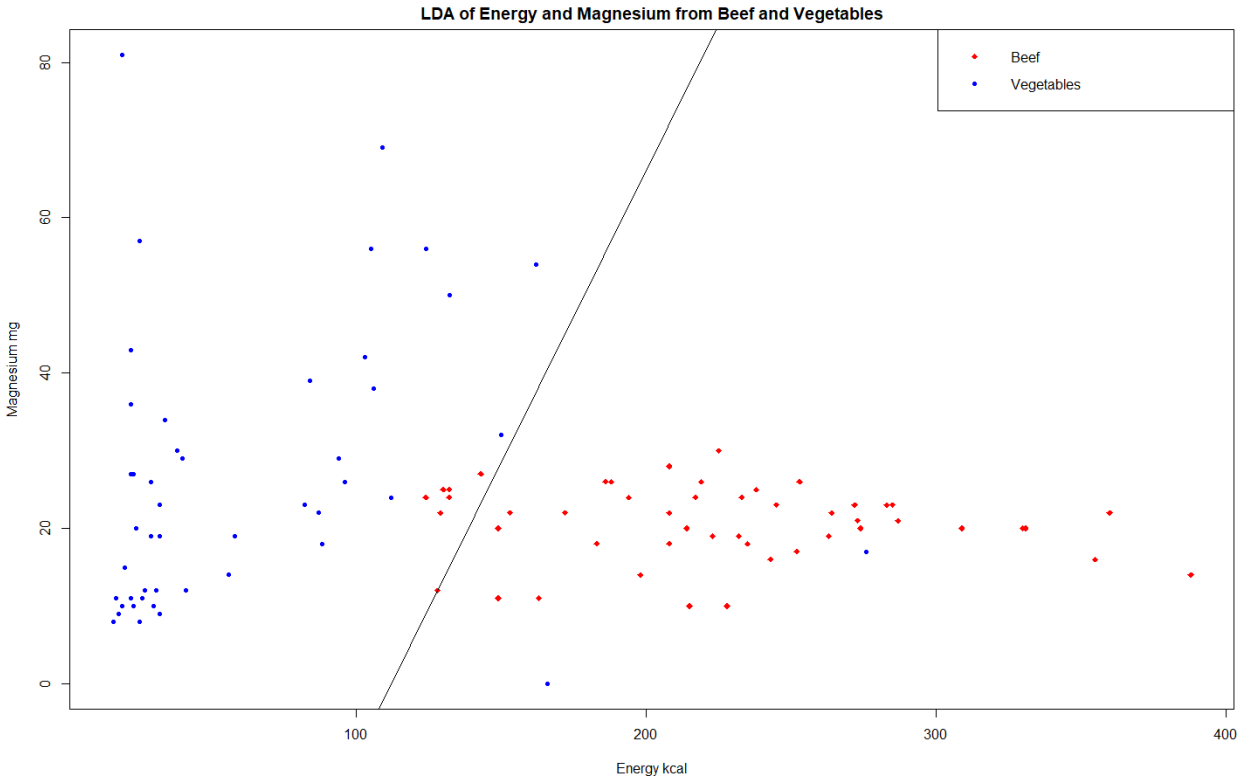

Fig. 12 LDA of Energy and Magnesium from beef and vegetable food groups (randomly sampled).

Below is a stacked histogram of the LDA values for the beef and vegetable groups ($Fig.$ 13). They show that the beef group seems to be more spread out while the vegetables are more concentrated in one area. This could explain why the vegetables have a better accuracy in comparison to the beef classification. They also have a different mean for the LDA values, making it so that LDA is effective in determining the difference between one observation and another.

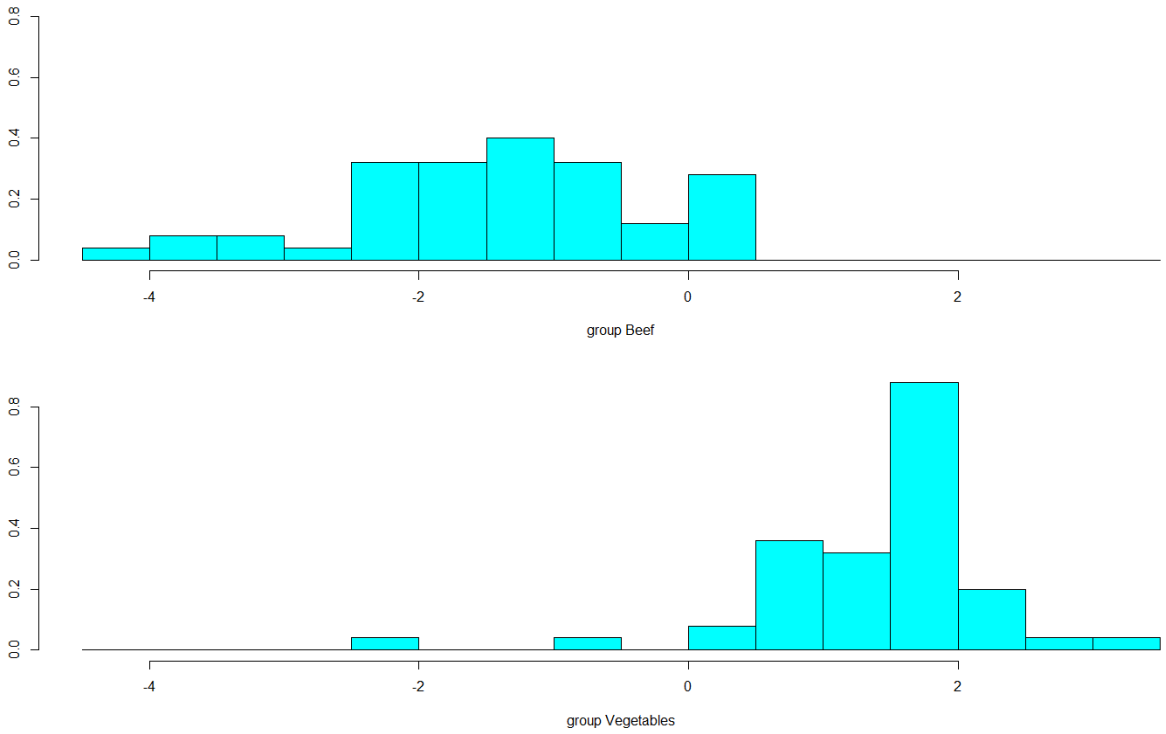

Fig. 13 Stacked histogram of the LDA values for the beef and vegetable groups.

Lastly is a confusion matrix of the results from using LDA on the sample of the data. It gives an overall accuracy of 91% from the sample of 100 observations. It shows that beef was misclassified more often than vegetables, where there were 7 errors (14%) with beef and 2 errors (4%) with vegetables. Looking at the stacked histogram of the LDA values, it does seem to make sense since the beef data has much more spread out values in comparison to vegetables.

	Beef	Vegetables
Beef	43	7
Vegetables	2	48

4. Interpretation

The analysis was done on two food groups, beef and vegetables, out of 13 original food groups. Using the 2 largest food groups, it was apparent that there could be some errors with the interpretation, since the tests (e.g., Hotelling's T^2 and Bonferroni's corrected) require the normality assumption. Therefore, this fault with the data must be considered when analyzing the results.

The first of the series of tests that were performed are the confidence intervals for each of the variables in the beef and vegetable groups. Doing the univariate one-at-a-time test for each of the variables showed that only eight of the variables were rejected under the null hypothesis that the true mean was equal to a null mean derived from rounding the sample mean (section 3.1.1). The purpose of the one-at-a-time hypothesis test was mainly to analyze the difference between its results and those of the simultaneous confidence intervals.

After the one-at-a-time test was done, the first of the simultaneous confidence intervals was done using Hotelling's T^2 (section 3.1.2). The simultaneous confidence intervals are wider for all the variables in comparison to the intervals in the one-at-a-time case. This is expected as the interval should be larger for each individual variable to create an overall higher probability that all the variables will contain each of their respective true means at the same time.

Next the same type of simultaneous confidence interval was done using Bonferroni's corrected confidence intervals (section 3.1.2). Bonferroni's corrected method utilizes the same technique of enlarging each of the individual confidence intervals so that overall, they can better contain the Type I error. The difference between the Hotelling's T^2 and Bonferroni's corrected simultaneous intervals is that Bonferroni's creates a tighter bound, leading to a less conservative bound than Hotelling's T^2. The goal however is the same in which both are looking to control the Type I error rate.

To examine more closely the simultaneous confidence region, two variables were chosen from the beef food group (section 3.2). The two variables are fat and protein, nutrients which can be found easily within foods that contain beef. A simultaneous confidence region was plotted where both the Hotelling's T^2 and Bonferroni's corrected confidence intervals are shown. The ellipse represents the hypothesized region where both populations means of fat and protein can be located based on the sample data. The similar results as before are seen here, where the confidence interval of Hotelling's T^2 is wider and thus more conservative than Bonferroni's corrected method.

A two-sample test using Hotelling's T^2 was also done to examine whether the beef or vegetable populations could be considered the same based on their nutrient contents (section 3.3). The results of the hypothesis test came back with a p-value of essentially 0, something expected as the nutrient content of beef and vegetables are considerably different. If the comparison were for something such as pork and lamb, it is likely that the p-value would've been relatively much larger. Since the hypothesis test was rejected, it was worth looking at each of the confidence intervals independently for each of the variables. If the confidence interval contained 0, this would indicate that the variable for the two populations were roughly equal. However, the simultaneous confidence intervals of both Hotelling's T^2 and Bonferroni's corrected method showed that only Thiamin in Hotelling's T^2 confidence interval contained 0. Looking back at the set of box plots, it seems that Thiamin is quite close for both food groups, despite there being many outliers for vegetables.

Then principal component analysis was used initially on only the beef population (section 3.4). The method turns the data into principal components where the first few principal

components can be kept while the less explanatory principal components can be ignored (the first few principal components contain the largest amount of variation). This process of seeing the variation that is explained by each cumulative principal component can be seen in the scree plot and plot showing the proportion of variation attributed to each cumulative eigenvalue. It was noted also in the analysis that despite the elbow break of 3 principal components only contributing to around 40% of the total variation, with 22 principal components that is already a positive sign.

In the plot showing the loadings of the principal components, the nutrients which beef is known to have contain loadings with larger magnitude. For example, nutrients such as: iron, Vitamin B12, protein, fat, etc. have larger values than other nutrients which beef is not known to have high amounts of. An example is Vitamin C, where the loadings for this nutrient are much closer to 0.

The biplot of the beef data shows that most of the data can be fit into a main super group that most of the different beef foods fall under. There are however certain mini groups of outliers that can be identified as being similar in terms of their description. This shows that much of the beef has similar nutritional properties, apart from certain types of special beef such as Wagyu or beef patties which have extreme nutritional characteristics that are different from any other typical cut of beef.

The biplot of both the beef and vegetable data shows that the vegetables express a much more varied set of scores for principal components 1 and 2. This makes sense, considering that there are many different types of vegetables, while beef is restricted to only cow meat. It is also apparent that despite the two groups being close to each other, they remain in relatively distinct groups that don't heavily overlap. This makes sense as beef and vegetables do have rather different nutritional characteristics and therefore their principal components capture different variations in the nutritional content to characterize the food as either beef or vegetable.

Lastly, linear discriminant analysis (LDA) was used on the two food groups (section 3.5). The purpose of using linear discriminant analysis is to classify observations into different groups based on their variables. Using the magnesium and energy variables from the dataset, LDA was able to do a good job of classifying foods into either their beef or food groups. It is apparent that the ability for linear discriminant analysis to be effective is dependent on certain choices, such as which food groups and which variables to represent are selected. In this case, the balance of the LDA values for beef and vegetables were balanced enough so that the accuracy was above 90%. It is possible that with different nutrients the accuracy could be improved, as it is apparent that the accuracy for beef is not that great. It is possible that magnesium or energy, despite their abundance in beef does not create the concentrated spread of LDA values that exists for vegetables.

5. Conclusion

The goal of the analysis has been to better understand the different tools and techniques that have been learned throughout the class, using the nutrition dataset as an example. The beef and vegetable were chosen from the rest of the food groups as the two populations to do analysis

with. Although it is quite apparent that beef and vegetables are different, it is not obvious nutritionally to what extent, as they still share many similar nutrients as well as many different nutrient contents. The numerous nutrients per food observation made it difficult to develop tests with simple and easy to interpret results. In many cases, the data had to be filtered down to fewer observations and extreme values removed for analysis to be done properly.

It was apparent that with the non-normal data that trying to test the hypothesis that the means were simultaneously equal to the rounded sample means would not be effective. With fewer variables involved, it could have possible to get different results. Another possibility is to round them to more decimal places to get different results. However, without enough domain knowledge about nutrition it was not possible to develop a more adequate null hypothesis for each of the 22 different variables.

Towards the beginning it was interesting to see how from the data would follow the pattern expected from using either Hotelling's T^2 or Bonferroni's corrected method in the simultaneous testing. The former method was more conservative in the confidence intervals. Despite this, there would be no difference in the results for either method as the tests came back quite close to 0 in terms of their p-values. If different food groups such as lamb and pork for instance were chosen with more selective parameters, it is possible that there could've been a difference in the results depending on which test methods were chosen.

Using principal component analysis produced rather effective results in grouping the different foods into their correct food groups. In this case, it may be assumed that beef and vegetables are quite different, but with 22 nutrients there was more similarity than previously assumed. The number of principal components was able to be reduced from 22 to 3, making it quite useful in terms of dimension reduction. Also, despite the reduction of the first 3 only being 40%, with the large number of parameters this is already considered quite effective.

Keeping all 22 variables made it possible during the beef biplot to distinguish several mini groups that existed apart from the super group of beef observations. By narrowing down the standardized scores to examine these mini groups, it was possible to see that they were all quite similar and the reason for their nutritional divergence was apparent in the descriptions of each mini cluster of beef.

Also, when using the biplot on both food groups together, it was apparent that, vegetables contain a much more diverse spread of nutritional content in comparison to just beef. It is likely that if other meats were grouped together with the beef, such as lamb, chicken, and pork, that the spread of principal component scores would increase as well. This would be associated with the more rich and diverse nutrient content of the newer larger meat food group. Something interesting was that the vectors for the nutrients that are associated with beef such as energy and protein were pointed to the right-side of the V-shape where the beef is clustered, while the nutrient associated with vegetables such as Vitamin A and Calcium were pointed towards the left on the left-side of the V-shape. It is not certain still why the V-shape occurred; it is likely that with more food groups that the shape would change as the foods would become more spread out into overlapping areas.

Using linear discriminant analysis on the chosen variables was quite effective on the different food groups. It is apparent that the choice of variables will influence determining which food group a food observation belongs to. In this case, the balance of the concentrated Magnesium and spread out Energy was useful in classifying a food as one group or the other. In the case of different food groups, it would be necessary first to choose the correct set of variables in doing the classification. If the LDA values do not have distinct histograms, it could make it difficult to obtain a decent accuracy during the classification process. However, with the chosen variables and food groups, an accuracy of over 90% is already quite good.

After doing nutrient analysis on beef in comparison to vegetables, it is apparent that the difference between the two food groups is not as extreme as previously imagined. The two food groups share many overlaps in terms of their nutritional content, and it is the few divergences from similarity that makes it possible to classify them as either being from the beef or vegetable food group. For further research into the nutritional similarities between food groups, it would be interesting to look at comparisons between more similar food groups, such as: fruit and vegetables, fish and shellfish, chicken and pork, etc. An issue however is the assumption of normality. When the beef and vegetables lacked normally distributed variables, it is unlikely that other food groups have normality for all their parameters. Looking at the box plots, it’s apparent that there’s a strong skew when it comes to the distribution of the data. It is possible that with transformations that the data would become more normal and that the results of the above tests could’ve been quite different.