STAT 6337 Project 1

Lakshmipriya Narayanan

Problem 1

(a)

We perform chi-square test of independence on variables HYPERTENSION and CVD.

H_0 : Hypertension is not associated with CVD. VS.
H_A : Hypertension is associated (depends on) with CVD.

Test Statistic, $\chi^2_{obs} = 123.05$ and p-value $= < 0.0001$

Since the p-value is $< 0.0001 << \alpha = 0.05$, at the 5% level we reject the H_0 and conclude that hypertension depends on Cardiovascular disease over the age of 24.

(b)

Summary statistics for CIGS PER DAY:
From the histogram, we observe that it is clearly right-skewed implying, variable CIGS PER DAY is not normally distributed. Similarly, the boxplot is not symmetrical because there is only one whisker(for normality, the whiskers of the plot must be about the same size) and there are a few outliers indicating that the data in this variable is not normally distributed.

Parametric test: We perform two-sample t-test to check if the values of CIGS PER DAY are different in the two groups of CVD since CVD is categorical and CIGS PER DAY is numerical.

H_0 : Cigarettes smoked per day is not associated with CVD (Mean difference between the two groups of CVD= 0.) VS.
H_A : Cigarettes smoked per day is associated with CVD (Mean difference between two groups of CVD is non-zero).

First, we look at the F-test:
Consider $H_0 : \sigma_1^2 = \sigma_2^2$ VS $H_A : \sigma_1^2 \neq \sigma_2^2$. Then,
$F_{obs} = 1.15$ and p-value $= 0.003 < 0.05$. So we reject H_0 and conclude that variances are unequal and consider the Satterthwaite t-test for unequal variances.

Therefore, Test Statistic, $t_{obs} = -3.15$ and p-value $= 0.0017$.
Since the p-value is $<< \alpha = 0.05$, at the 5% level we reject the H_0 and conclude that the number of cigarettes smoked per day is associated with Cardiovascular disease over the age of 24.

Non-Parametric test: We perform Wilcoxon-Mann-Whitney Rank sum test and achieve the following results:
Test statistic, $W_{obs} = 2,619,021$ and p-value $= 0.0042$. Since the p-value is $< \alpha = 0.05$, we reject the H_0 at the 5% significance level and conclude that the number of cigarettes smoked per day is associated with CVD.

We notice that both parametric test and non-parametric tests have the same conclusion of showing association between CIGS PER DAY and CVD.

(c)

From the plot for BMI vs Glucose, we can see that there is an inverse relationship between them. When glucose levels are high, body mass index is less and vice-versa.

Whereas, for most people who have a cardiovasular disease, they have high glucose and low BMI. For most people who do not suffer from a cardiovascular disease, they have low glucose and high BMI. Nevertheless, there are people who have high BMI and high glucose and low BMI and low glucose. So, we can infer that BMI and glucose are possibly weakly associated and CVD has an effect on these variables.
(See output for 1c ttest procedure : "Effect on BMI and glucose by CVD.")

(d)

We perform one-sample t test.
$H_0 : \mu_{SBP} = 125\ mmHg$ vs $H_A : \mu_{SBP} > 125\ mmHg$
Test statistic, $t_{obs} = 23.48$ and p-value = $< 0.0001 << \alpha = 0.05$.
We reject H_0 and conclude that at the 5% level, we have significant evidence that the average systolic blood pressuse of this population is more than 125 mmHg.

(e)

SBP levels (in mmHg)	Group Nmber
< 117.5	1
117.5 - 128	2
129 - 143	3
> 144	4

We perform Chi-square test of independence.
H_0 : SBPgroup is not associated with Hypertension. VS.
H_A : SBPgroup is associated with hypertension.

Test statistic, $\chi^2_{obs} = 1120.1830$ and p-value = $< 0.0001 << \alpha = 0.05$.
Therefore, we reject H_0 at the 5% significance level and state that the new categorical variable (SBP groups based on quartiles) are associated with hypertension.

(f)

Using the summary statistics for TOTAL CHOL, we can see that skewness is $0.852 > 0$ is positive indicating slight right-skewdness. The kurtosis is 3.919 which indicates that the tails are slightly thicker than what normal distribution is usually supposed to have. Hence, we can say that this variable may follow normal distribution based on the summary statistics observations. From the histogram for total cholestrol, we can observe that it is right-skewed. But, from the boxplot we can observe that it is symmetric but with more than a few outliers indicating normality. The Q-Q plot helps us infer normality assumption better because the points are all on the 45^0 straight line.
From the χ^2 goodness of fit test, the p-value obtained was < 0.0001 and the test statistic, $\chi^2 obs = 3895.3848$ is very large. Therefore, we can conclude that the variable total cholestrol is not normally distributed because we assume H_0 : TOTAL CHOL follows normal distribution and H_A : TOTAL CHOL does not follow normal distribution.
Clearly, the histogram and the goodness of fit test indicate that this variable is not normally distributed.

Problem 2:

Parametric tests:
We perform paired t-test because these samples are not independent.
For this dataset our hypotheses are:
H_0 : Mean Difference between treatments Difference, $D = T_i - T_j$ for $i, j = 1, 2, 3, 4$ is $\mu_D = 0$ VS $H_A : \mu_D \neq 0$

Results table			
Mean Difference	Test statistic, t_{obs}	p-value	Conclusion
μ_{T2-T1}	2.63	$0.017 > \alpha = 0.004$	Fail to Reject H_0
μ_{T3-T1}	7.87	$< 0.0001 < \alpha$	Reject H_0
μ_{T4-T1}	10.53	$< 0.0001 < \alpha$	Reject H_0
μ_{T3-T2}	5.02	$< 0.0001 < \alpha$	Reject H_0
μ_{T4-T2}	6.24	$< 0.0001 < \alpha$	Reject H_0
μ_{T4-T3}	1.97	$0.0643 > \alpha$	Fail to Reject H_0

From the table, we can conclude that Treatments 1 and 2 (high CO2 pressure without H and low CO2 pressure without H) and Treatments 3 and 4 (high CO2 pressure with H and low CO2 pressure with H) do not differ from each other whereas, the rest of the combinations of the treatments differ from each other.

Non-Parametric tests:
We perform Sign test because these samples are conducted on the same dog and hence not independent.
H_0 : Median Difference between treatments Difference, $D = T_i - T_j$ for $i, j = 1, 2, 3, 4$ is $\nu_D = 0$ VS $H_A : \nu_D \neq 0$

Results table			
Median Difference	Test statistic, C_{obs}	p-value	Conclusion
ν_{T2-T1}	3.5	$0.1435 > \alpha = 0.004$	Fail to Reject H_0
ν_{T3-T1}	8.5	$< 0.0001 < \alpha$	Reject H_0
ν_{T4-T1}	8.5	$< 0.0001 < \alpha$	Reject H_0
ν_{T3-T2}	6.5	$0.0044 > \alpha$	Fail to Reject H_0
ν_{T4-T2}	6.5	$0.0044 > \alpha$	Fail to Reject H_0
ν_{T4-T3}	6	$0.0075 > \alpha$	Fail to Reject H_0

From the table, we notice that both parametric and non-parametric tests differ in conclusions *i.e.*, Treatments 1 and 3 (high CO_2 pressure without H and high CO_2 pressure with H) and Treatments 1 and 4 (high CO_2 pressure without H and low CO_2 pressure with H) do differ from each other whereas, the rest of the combinations of the treatments do not differ from each other.

Therefore, we can infer that the conclusions of the two tests contradict!

We use significance level, $\alpha = 0.0004$ to reduce Type I error rate. This is because we have performed multi-variate tests and the probability of incorrectly rejecting a null hypothesis is high. So, we reduce our α from 0.05 to 0.0004.

Relevant SAS outputs:

1(a): Association of Hypertension with CVD

The FREQ Procedure

Frequency Percent Row Pct Col Pct

Table of HYPERTENSION by CVD

	CVD		
HYPERTENSION	0	1	Total
0	1017 22.94 86.04 31.03	165 3.72 13.96 14.26	1182 26.66
1	2260 50.97 69.50 68.97	992 22.37 30.50 85.74	3252 73.34
Total	3277 73.91	1157 26.09	4434 100.00

Statistics for Table of HYPERTENSION by CVD

Statistic	DF	Value	Prob
Chi-Square	1	123.0504	<.0001
Likelihood Ratio Chi-Square	1	134.5014	<.0001
Continuity Adj. Chi-Square	1	122.1940	<.0001
Mantel-Haenszel Chi-Square	1	123.0226	<.0001
Phi Coefficient		0.1666	
Contingency Coefficient		0.1643	
Cramer's V		0.1666	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	1017
Left-sided Pr <= F	1.0000
Right-sided Pr >= F	<.0001
Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Sample Size = 4434

1(b): Summary statistics for CIGS PER DAY, boxplot and histogram

The UNIVARIATE Procedure
Variable: CIGSPERDAY

Moments			
N	4402	Sum Weights	4402
Mean	8.96637892	Sum Observations	39470
Std Deviation	11.9317058	Variance	142.365604
Skewness	1.26231055	Kurtosis	1.07611289
Uncorrected SS	980454	Corrected SS	626551.024
Coeff Variation	133.071622	Std Error Mean	0.17983637

Basic Statistical Measures			
Location		Variability	
Mean	8.966379	Std Deviation	11.93171
Median	0.000000	Variance	142.36560
Mode	0.000000	Range	70.00000
		Interquartile Range	20.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	49.85854	Pr > |t|	<.0001
Sign	M	1074.5	Pr >= |M|	<.0001
Signed Rank	S	1155088	Pr >= |S|	<.0001

Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.285629	Pr > D	<0.0100
Cramer-von Mises	W-Sq	76.0177	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	422.2777	Pr > A-Sq	<0.0050

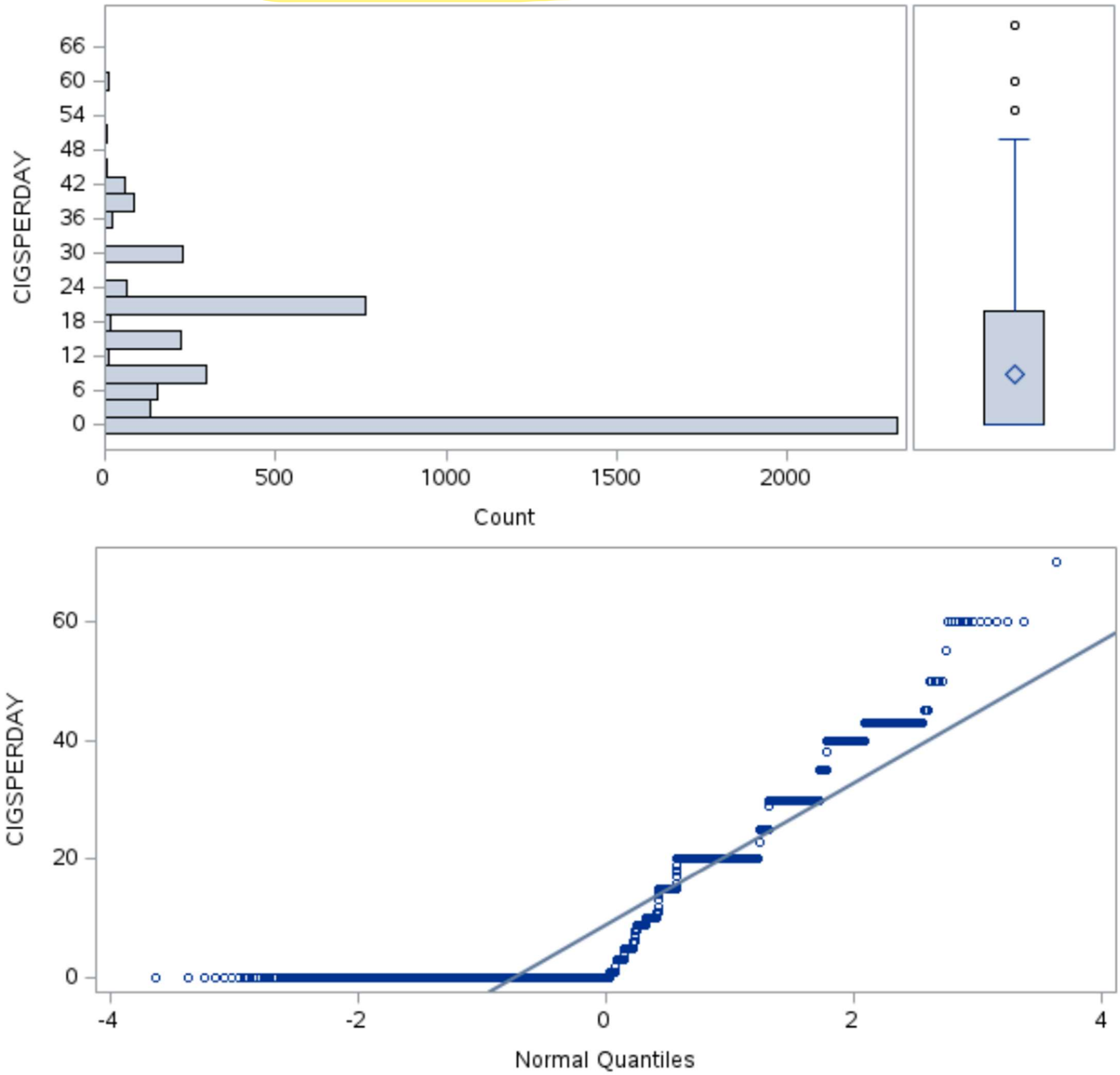
Quantiles (Definition 5)	
Level	Quantile
100% Max	70
99%	43
95%	30
90%	25
75% Q3	20
50% Median	0

Quantiles (Definition 5)	
Level	Quantile
25% Q1	0
10%	0
5%	0
1%	0
0% Min	0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	4433	60	2838
0	4432	60	2839
0	4427	60	3843
0	4426	60	4107
0	4423	70	3142

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	32	0.72	100.00

Distribution and Probability Plot for CIGSPERDAY



Parametric test for association of CVD with CIGS PER DAY:

The TTEST Procedure

Variable: CIGSPERDAY

CVD	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		3257	8.6193	11.6886	0.2048	0	70.0000
1		1145	9.9537	12.5504	0.3709	0	60.0000
Diff (1-2)	Pooled		-1.3344	11.9187	0.4095		
Diff (1-2)	Satterthwaite		-1.3344		0.4237		

CVD	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		8.6193	8.2177	9.0209	11.6886	11.4115	11.9796
1		9.9537	9.2260	10.6814	12.5504	12.0566	13.0867
Diff (1-2)	Pooled	-1.3344	-2.1372	-0.5316	11.9187	11.6748	12.1730
Diff (1-2)	Satterthwaite	-1.3344	-2.1654	-0.5035			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	4400	-3.26	0.0011
Satterthwaite	Unequal	1886.4	-3.15	0.0017

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	1144	3256	1.15	0.0030

Non- Parametric test for association of CVD with CIGS PER DAY:

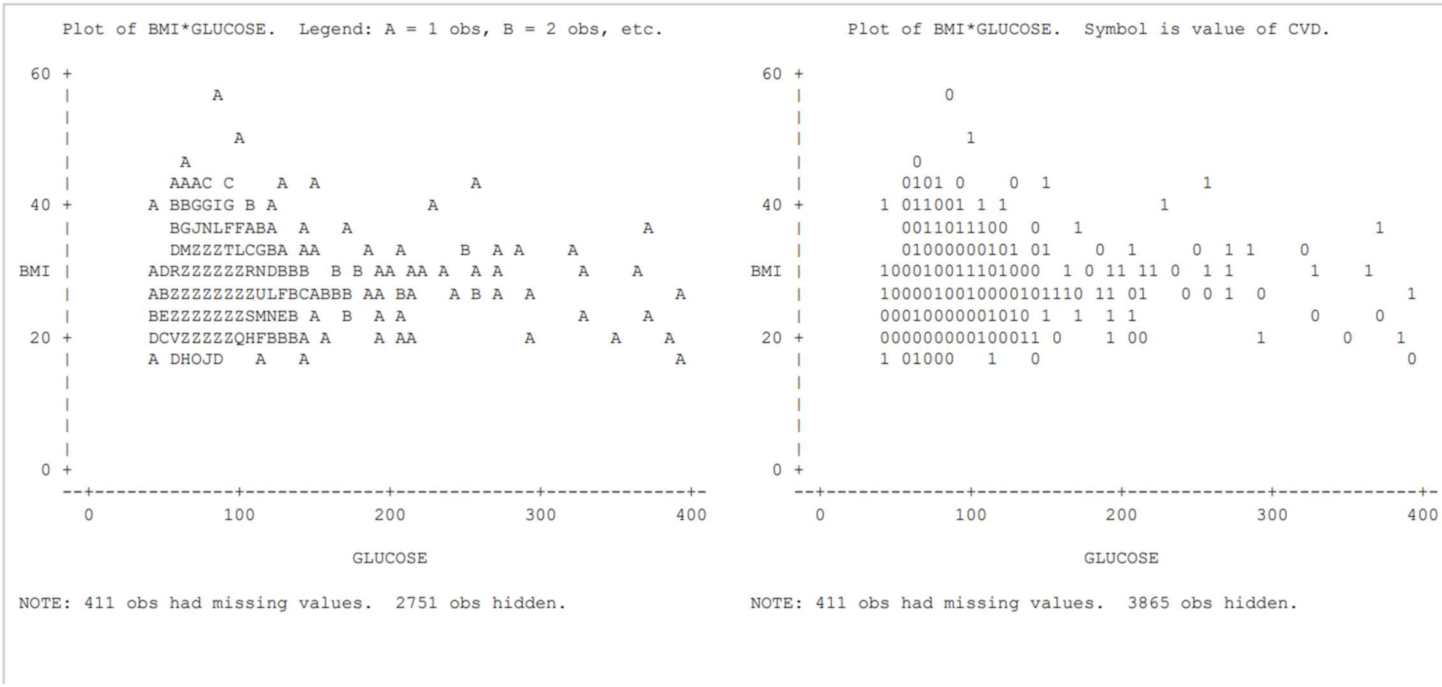
The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable CIGSPERDAY Classified by Variable CVD					
CVD	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	1145	2619020.50	2520717.50	34310.4052	2287.35415
0	3257	7071982.50	7170285.50	34310.4052	2171.31793
Average scores were used for ties.					

Wilcoxon Two-Sample Test					
				t Approximation	
Statistic	Z	Pr > Z	Pr > |Z|	Pr > Z	Pr > |Z|
2619021	2.8651	0.0021	0.0042	0.0021	0.0042
Z includes a continuity correction of 0.5.					

Kruskal-Wallis Test		
Chi-Square	DF	Pr > ChiSq
8.2088	1	0.0042

1(c): Plot of BMI vs Glucose



Effect on BMI and Glucose by CVD:

The TTEST Procedure

Difference: BMI - GLUCOSE

N	Mean	Std Dev	Std Err	Minimum	Maximum
4023	-56.3046	24.1661	0.3810	-376.8	-1.1200

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
-56.3046	-57.0516	-55.5577	24.1661	23.6494	24.7061

DF	t Value	Pr > |t|
4022	-147.78	<.0001

1(d): Test for average SBP > 125 mmHg

The TTEST Procedure

Variable: SBP

N	Mean	Std Dev	Std Err	Minimum	Maximum
4434	132.9	22.4216	0.3367	83.5000	295.0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
132.9	132.4	Infty	22.4216	21.9645	22.8983

DF	t Value	Pr > t
4433	23.48	<.0001

1(e): Summary statistics of SBP to create a categorical SBP

The UNIVARIATE Procedure
Variable: SBP

Moments			
N	4434	Sum Weights	4434
Mean	132.907758	Sum Observations	589313
Std Deviation	22.421597	Variance	502.728011
Skewness	1.14790056	Kurtosis	2.08648398
Uncorrected SS	80552863	Corrected SS	2228593.27
Coeff Variation	16.8700438	Std Error Mean	0.33671983

Basic Statistical Measures			
Location		Variability	
Mean	132.9078	Std Deviation	22.42160
Median	129.0000	Variance	502.72801
Mode	120.0000	Range	211.50000
		Interquartile Range	26.50000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	394.7132	Pr > |t|	<.0001
Sign	M	2217	Pr >= |M|	<.0001
Signed Rank	S	4916198	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	295.0
99%	202.0
95%	176.5
90%	163.0
75% Q3	144.0
50% Median	129.0
25% Q1	117.5
10%	109.0
5%	104.5
1%	97.0
0% Min	83.5

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
83.5	3647	242	965
83.5	2793	243	1240
85.0	3740	244	894
85.5	2093	248	3649
90.0	3027	295	501

Association of categorical SBP with Hypertension

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of HYPERTENSION by SBPGRP

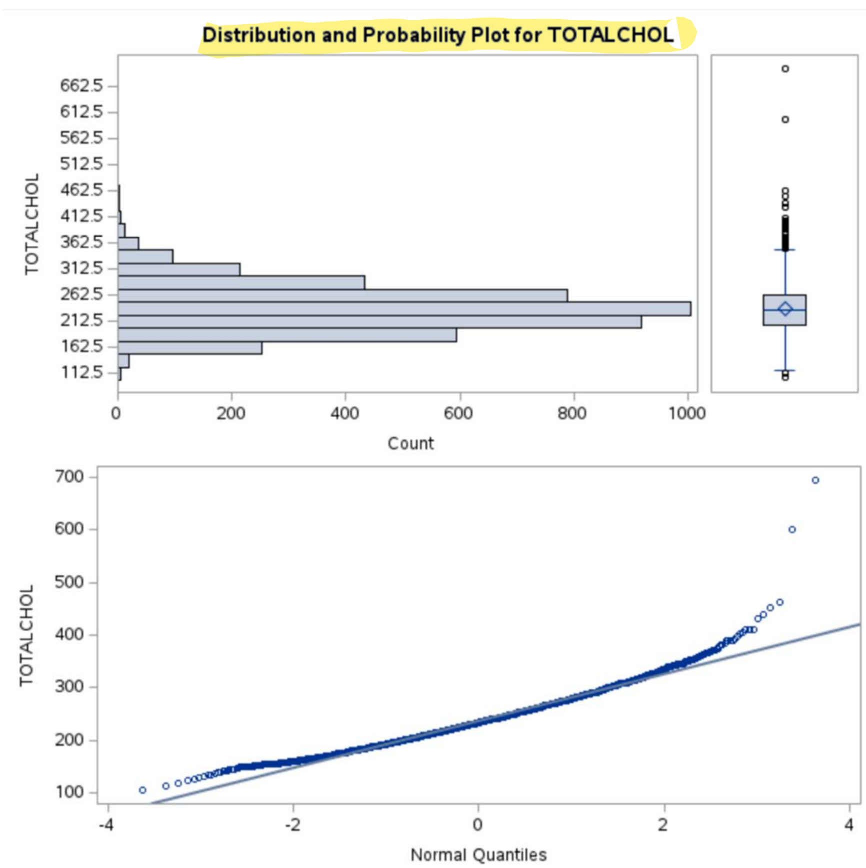
HYPERTENSION	SBPGRP 1	2	3	4	Total
0	676 15.25 57.19 58.89	369 8.32 31.22 32.65	128 2.89 10.83 12.09	9 0.20 0.76 0.82	1182 26.66
1	472 10.65 14.51 41.11	761 17.16 23.40 67.35	931 21.00 28.63 87.91	1088 24.54 33.46 99.18	3252 73.34
Total	1148 25.89	1130 25.48	1059 23.88	1097 24.74	4434 100.00

Statistics for Table of HYPERTENSION by SBPGRP

Statistic	DF	Value	Prob
Chi-Square	3	1120.1830	<.0001
Likelihood Ratio Chi-Square	3	1274.0100	<.0001
Mantel-Haenszel Chi-Square	1	1088.0201	<.0001
Phi Coefficient		0.5026	
Contingency Coefficient		0.4491	
Cramer's V		0.5026	

Sample Size = 4434

1(f): Normality test for TOTAL CHOL using plots



Normality test for TOTAL CHOL using summary statistics

The UNIVARIATE Procedure
Variable: TOTALCHOL

Moments			
N	4382	Sum Weights	4382
Mean	236.984254	Sum Observations	1038465
Std Deviation	44.6510984	Variance	1993.72059
Skewness	0.85247219	Kurtosis	3.91993985
Uncorrected SS	254834343	Corrected SS	8734489.91
Coeff Variation	18.8413777	Std Error Mean	0.67452175

Basic Statistical Measures			
Location		Variability	
Mean	236.9843	Std Deviation	44.65110
Median	234.0000	Variance	1994
Mode	240.0000	Range	589.00000
		Interquartile Range	58.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	351.3367	Pr > |t|	<.0001
Sign	M	2191	Pr >= |M|	<.0001
Signed Rank	S	4801577	Pr >= |S|	<.0001

Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.042222	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1.730376	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	11.53559	Pr > A-Sq	<0.0050

Quantiles (Definition 5)	
Level	Quantile
100% Max	696
99%	355
95%	313
90%	293
75% Q3	264
50% Median	234

Quantiles (Definition 5)	
Level	Quantile
25% Q1	206
10%	183
5%	170
1%	154
0% Min	107

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
107	1694	439	564
113	2649	453	3632
119	4249	464	203
124	2563	600	1157
126	1957	696	3299

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	52	1.17	100.00

Normality test for TOTAL CHOL using goodness of fit test

Chi-Square Test for Equal Proportions	
Chi-Square	3895.3848
DF	249
Pr > ChiSq	<.0001

2: Parametric test for difference between Treatments

The TTEST Procedure

Variable: DIFFT1T2

N	Mean	Std Dev	Std Err	Minimum	Maximum
19	36.4211	60.3787	13.8518	-25.0000	183.0

Mean	99.6% CL Mean		Std Dev	99.6% CL Std Dev	
36.4211	-9.2606	82.1027	60.3787	40.4346	109.9

DF	t Value	Pr > |t|
18	2.63	0.0170

Fig 1:Treatments 1 and 2

Variable: DIFFT1T3

N	Mean	Std Dev	Std Err	Minimum	Maximum
19	111.1	61.5110	14.1116	-10.0000	248.0

Mean	99.6% CL Mean		Std Dev	99.6% CL Std Dev	
111.1	64.5143	157.6	61.5110	41.1928	111.9

DF	t Value	Pr > |t|
18	7.87	<.0001

Fig 2: Treatments 1 and 3

Variable: DIFFT1T4

N	Mean	Std Dev	Std Err	Minimum	Maximum
19	134.7	55.7455	12.7889	-2.0000	228.0

Mean	99.6% CL Mean		Std Dev	99.6% CL Std Dev	
134.7	92.5080	176.9	55.7455	37.3318	101.4

DF	t Value	Pr > |t|
18	10.53	<.0001

Fig 3: Treatments 1 and 4

Variable: DIFFT2T3

N	Mean	Std Dev	Std Err	Minimum	Maximum
19	74.6316	64.8573	14.8793	-84.0000	179.0

Mean	99.6% CL Mean		Std Dev	99.6% CL Std Dev	
74.6316	25.5615	123.7	64.8573	43.4338	118.0

DF	t Value	Pr > |t|
18	5.02	<.0001

Fig 4: Treatments 2 and 3

Variable: DIFFT2T4

N	Mean	Std Dev	Std Err	Minimum	Maximum
19	98.2632	68.6382	15.7467	-76.0000	178.0

Mean	99.6% CL Mean		Std Dev	99.6% CL Std Dev	
98.2632	46.3325	150.2	68.6382	45.9658	124.9

DF	t Value	Pr > |t|
18	6.24	<.0001

Fig 5: Treatments 2 and 4

Variable: DIFFT3T4

N	Mean	Std Dev	Std Err	Minimum	Maximum
19	23.6316	52.2592	11.9891	-113.0	154.0

Mean	99.6% CL Mean		Std Dev	99.6% CL Std Dev	
23.6316	-15.9069	63.1701	52.2592	34.9970	95.0949

DF	t Value	Pr > |t|
18	1.97	0.0643

Fig 6: Treatments 3 and 4

2: Non - Parametric test for difference between Treatments

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	2.629331	Pr > |t|	0.0170
Sign	M	3.5	Pr >= |M|	0.1435
Signed Rank	S	51	Pr >= |S|	0.0133

Fig 1:Treatments 1 and 2

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	7.869599	Pr > |t|	<.0001
Sign	M	8.5	Pr >= |M|	<.0001
Signed Rank	S	94	Pr >= |S|	<.0001

Fig 2: Treatments 1 and 3

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	10.53134	Pr > |t|	<.0001
Sign	M	8.5	Pr >= |M|	<.0001
Signed Rank	S	94	Pr >= |S|	<.0001

Fig 3: Treatments 1 and 4

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	5.015805	Pr > |t|	<.0001
Sign	M	6.5	Pr >= |M|	0.0044
Signed Rank	S	83	Pr >= |S|	0.0003

Fig 4: Treatments 2 and 3

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	6.240243	Pr > |t|	<.0001
Sign	M	6.5	Pr >= |M|	0.0044
Signed Rank	S	87	Pr >= |S|	<.0001

Fig 5: Treatments 2 and 4

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	1.971092	Pr > |t|	0.0643
Sign	M	6	Pr >= |M|	0.0075
Signed Rank	S	53.5	Pr >= |S|	0.0182

Fig 6: Treatments 3 and 4

```
* Create a pointer named FHS to the data file;
filename FHS "/home/u63986019/FramHeartStudy_data.csv";

DATA c; /* Assign name c to data */
INFILE FHS DSD FIRSTOBS = 2; /* Since the data is a CSV, use DSD FIRSTOBS = 2*/
INPUT AGE TOTALCHOL SBP DBP BMI CIGSPERDAY GLUCOSE HEARTRATE CVD HYPERTENSION; /*Input names of columns*/

/*Converting numerical variable SBP to categorical by grouping according to quartiles of SBP*/
IF SBP LE 117.5 then SBPGRP = 1;
IF SBP > 117.5 AND SBP LE 129 then SBPGRP = 2;
IF SBP > 129 AND SBP LE 144 then SBPGRP = 3;
IF SBP > 144 then SBPGRP = 4;
RUN;

/* 1a: Is hypertension associated with CVD? Perform Chi-square test of independence */
proc freq data=c;
tables HYPERTENSION*CVD / chisq;  /*contingency table of hypertension by cvd */
RUN;

/*1b: Summary statistics for CIGS PER DAY using UNIVARIATE gives histogram and boxplot */
proc univariate NORMAL PLOT;
var CIGSPERDAY;
RUN;

/* Association of CIGS PER DAY with CVD*/

/*PARAMETRIC TEST: ttest b/c one variable is categorical (CVD) and one variable is numerical (CIGS PER DAY) */
proc ttest data=c;
class CVD;
var CIGSPERDAY;
RUN;

/*NON PARAMETRIC TEST: Wilcoxon Mann- Whitney Rank Sum test b/c we're comparing a categorical variable with a numerical one *,
proc npar1way data = c wilcoxon;
class CVD;
var CIGSPERDAY;
RUN;

/*1c: Plot of BMI vs GLUCOSE with different plotting symbols for the two CVD groups (Yes or No) */
proc plot HPERCENT=50 VPERCENT=50; *Reduce the size of plots to 50%;
plot BMI*GLUCOSE;
plot BMI*GLUCOSE = CVD; *uses different plotting symbols for (Yes) CVD and (No) CVD;
RUN;

/* Paired t test to check if CVD affects BMI and Glucose b/c BMI and Glucose are dependent*/
proc ttest data = c;
paired BMI*GLUCOSE;
RUN;

/*1d: Is average SBP more than 125. We perform one sample right sided t test*/
proc ttest data = c
H0 = 125
sides = U; *Specify null hypotheis and right sided t test;
var SBP;
RUN;

/*1e: Quartiles of SBP to make it categorical */
proc univariate;
var SBP;
RUN;

/*Chi-Sqaure test to see if the categorical SBP and Hypertension are dependent on each other */
proc freq data = c;
tables HYPERTENSION*SBPGRP / chisq;
RUN;

/*1f: Normality of TOTAL CHOL: UNIVARIATE NORMAL PLOT gives graphs and summaries */
proc univariate NORMAL PLOT;
var TOTALCHOL;
RUN;

/*Goodness of fit test to check if TOTAL CHOL follows normal distribution */
proc freq data = c;
tables TOTALCHOL / chisq;
RUN;
```

```sas
* Create a pointer named Dogs to the data file;
filename Dogs "/home/u63986019/Dogs.dat";

DATA d; /* Assign name d to data */
INFILE Dogs; *reads the file whose pointer is Dogs;

/*Specifying the length of variables in Dogs dataset. Allot 8 bytes and names of treatments according to dataset as variables
LENGTH T1HighNoH T2LowNoH T3HighYesH T4LowYesH 8;

INPUT T1HighNoH T2LowNoH T3HighYesH T4LowYesH; /* Input the variable names */

/* Compute differences of each treatment with another and store it in a new column named after treatment differences */
DIFFT1T2 = T2LowNoH - T1HighNoH;
DIFFT1T3 = T3HighYesH - T1HighNoH;
DIFFT1T4 = T4LowYesH - T1HighNoH;
DIFFT2T3 = T3HighYesH - T2LowNoH;
DIFFT2T4 = T4LowYesH - T2LowNoH;
DIFFT3T4 = T4LowYesH - T3HighYesH;
RUN;


/*PARAMETRIC TEST : ttest with null hypothesis of differences = 0 and specified alpha level = 0.004 */
proc ttest data = d
H0 = 0
alpha = 0.004;
var DIFFT1T2 DIFFT1T3 DIFFT1T4 DIFFT2T3 DIFFT2T4 DIFFT3T4; *Run this test for all differences in treatments;
RUN;


/*NON PARAMETRIC TEST: Since data is dependent (treatments are done on the same dogs), we perform sign test. The UNIVARIATE
function gives the results of non-parametric sign test under null hypothesis for median differences = 0 */
proc univariate data = d;
var DIFFT1T2 DIFFT1T3 DIFFT1T4 DIFFT2T3 DIFFT2T4 DIFFT3T4; *Run this test for all differences in treatments;
RUN;
```