Question 1 (23 marks)

Part A (13 Marks):

a1. Write down the statistical hypotheses. (1 Mark)

Answer :

Null hypothesis: There exists no significant difference in the effectiveness of the two drug types.

Alternate hypothesis: There exists a significant difference in the effectiveness of the two drug types.

Now, you want to recruit people to conduct your experiment.

a2. What is the study population? (1 Mark)

Answer : The study population is the males in the age group between 40-50 years.

a3. What is the best (ideal) way of recruiting study participants to your experiment? (1 Mark)

Answer : The cheapest and easiest way of recruiting study participants is to gather a random sample.

a4. Write the SAS codes to enter raw data into SAS. (2 Marks)

Answer :

///*Q1 part A ***///**

```
data drug;

  length subjectid  10;
  input subjectid drugA drugB;
  datalines;
0023 25 23
0125 45 41
9834 35 35
2130 50 51
3243 24 20
1231 32 27
3541 37 34
5682 31 30
0012 35 32
0230 43 34
```

;

a5. Write the SAS codes to perform an appropriate statistical test/s to answer the research question, including assessing the test assumptions. Paste your SAS outputs. (3 marks: 2-SAS code, 1-outputs)

Answer :

SAS code:

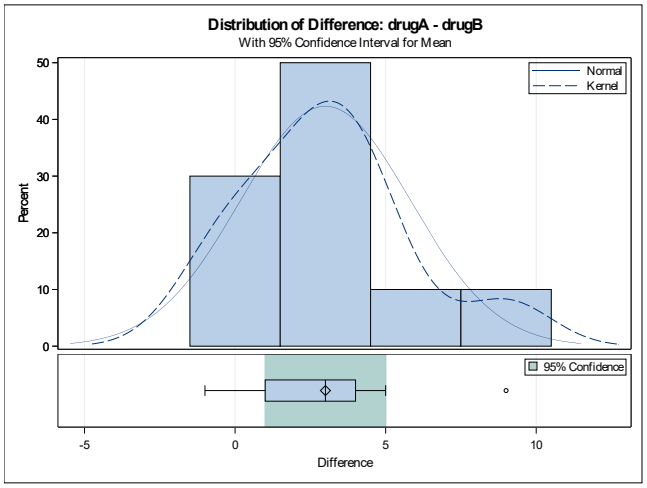
```
proc ttest data=drug;
     paired drugA*drugB;
  run;
```

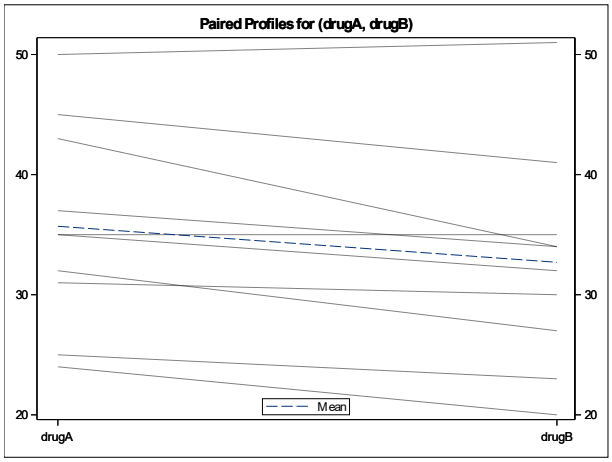
Output:

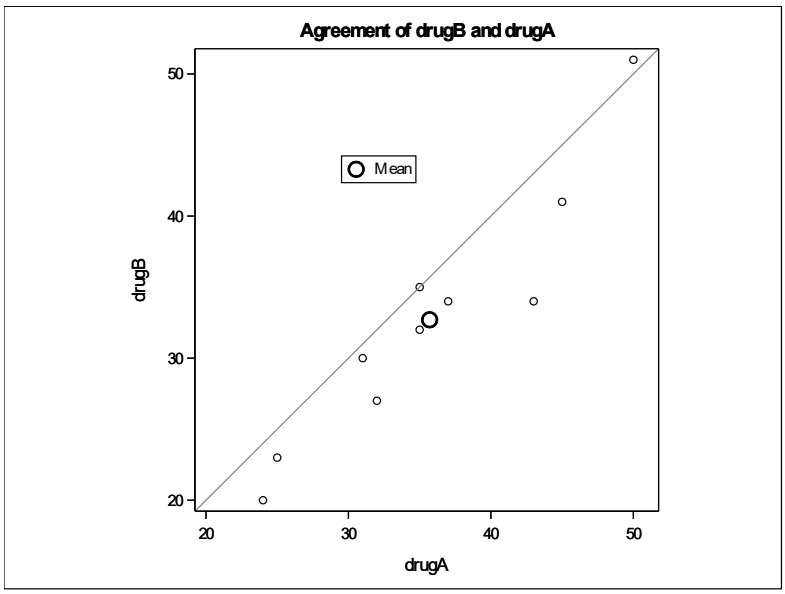
N	Mean	Std Dev	Std Err	Minimum	Maximum
10	3.0000	2.8284	0.8944	-1.0000	9.0000

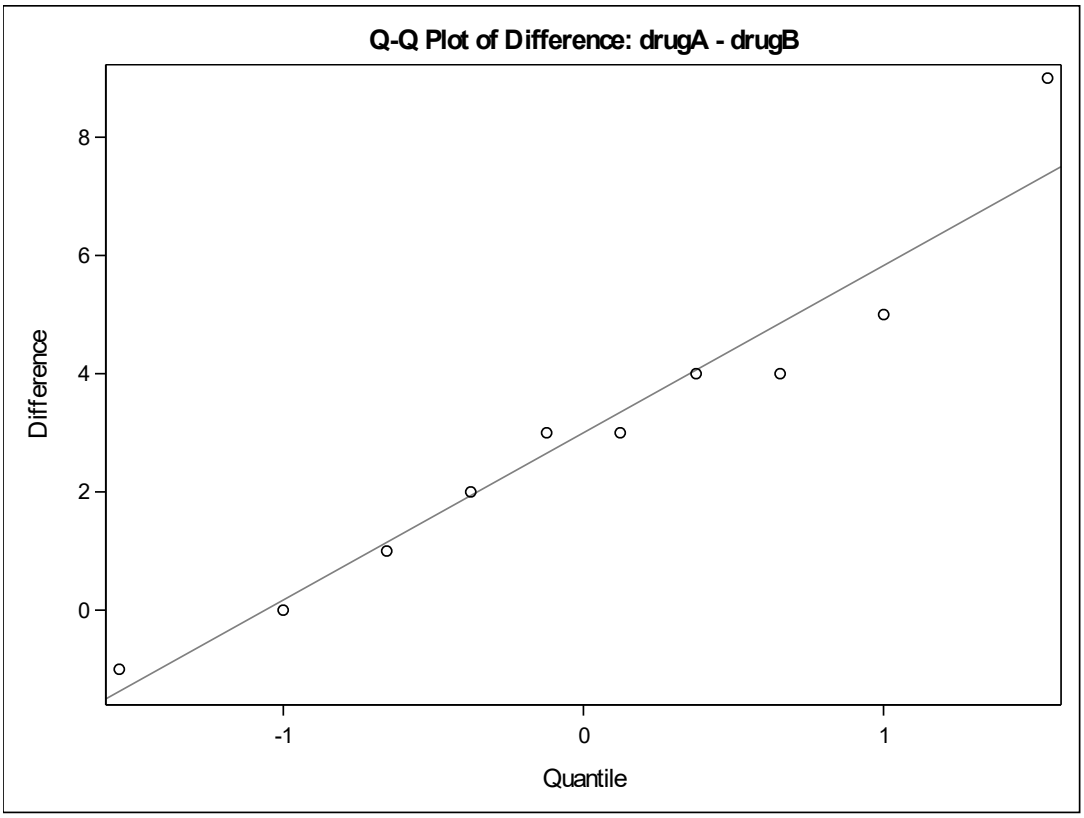
Mean	95% CL Mean		Std Dev	95% CL Std Dev	
3.0000	0.9767	5.0233	2.8284	1.9455	5.1636

DF	t Value	Pr > |t|
9	3.35	0.0085









a6. Using the SAS outputs obtained in part a5. answer the research question, providing appropriate statistics to support all your conclusions and interpretations. (4 Marks)

Answer: The p-value for the t-stat for the difference in the effectiveness of the two drugs is close to zero, concluding that there exists a statistically significant difference in the effectiveness of the two drug types. Furthermore, the results show that the mean of the difference is 3 which suggest that on an average the effectiveness of drug A is higher than drug B. The assumptions are tested as follows:

Assumption 1: The **dependent variable** is a **continuous** scale

Assumption 2: The **independent variable** consists of **categor**ical variable with two categories: drug A and drug B.

Assumption 3: There are **no significant outliers** in the **differences** of the effectiveness, except for one outlier, as seen from the box plot presented in the output.

Assumption 4: The **distribution of the differences** in the **effect**iveness of the two drugs is close to normal as it can be seen from the histogram and the Q-Q plot.

a7. Increasing the sample size would definitely increase the power of a study. Suggest another improvement for this study to further assess the effectiveness of the two drugs. (1 Mark)

Answer: We may increase the alpha level to increase the power of the study.

Part B (10 Marks):

Research Question: Is there a significant difference between the two treatments?

b1. Write a SAS program to enter raw data. (2 Marks)

Answer:

```
///***Q1 part B ***///

data yield;
  length id  3;
  input id yield trt;
  datalines;
1 4.81 1
2 4.17 1
3 4.41 1
4 3.59 1
5 5.87 1
6 3.83 1
7 6.03 1
8 4.89 1
9 4.32 1
10 4.69 1
11 6.31 2
12 5.12 2
13 5.54 2
14 5.5 2
15 5.37 2
16 5.29 2
17 4.92 2
18 6.15 2
19 5.8 2
20 5.26 2
;
```

b2. Write down the statistical hypotheses. (1 Marks)

Answer:

Null hypothesis: There exists no significant difference between the two treatments.

Alternate hypothesis: There exists a significant difference between the two treatments.

b3. Write the SAS codes to conduct a statistical analysis to answer the research hypothesis. Your answer should include SAS codes to check for the test assumptions, if necessary. Paste all relevant SAS outputs. (3 marks: 2-SAS codes, 1-outputs)

Answer: SAS code:

```
proc ttest data= yield sides=2 alpha=0.05 h0=0;
        title "Two sample t-test example";
        class trt;
        var yield;
  run;
```

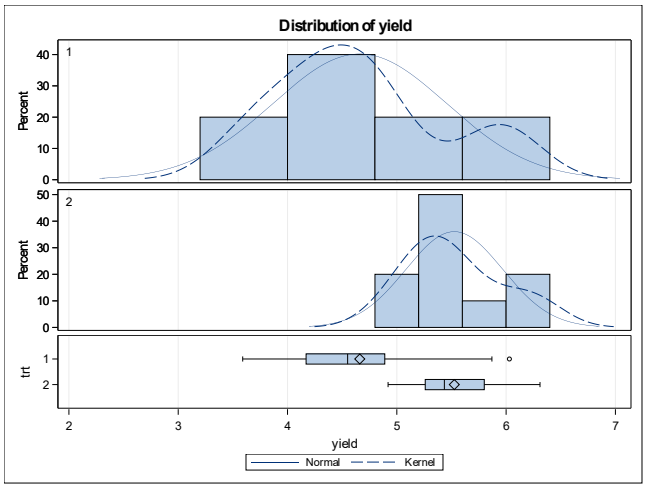
Output:

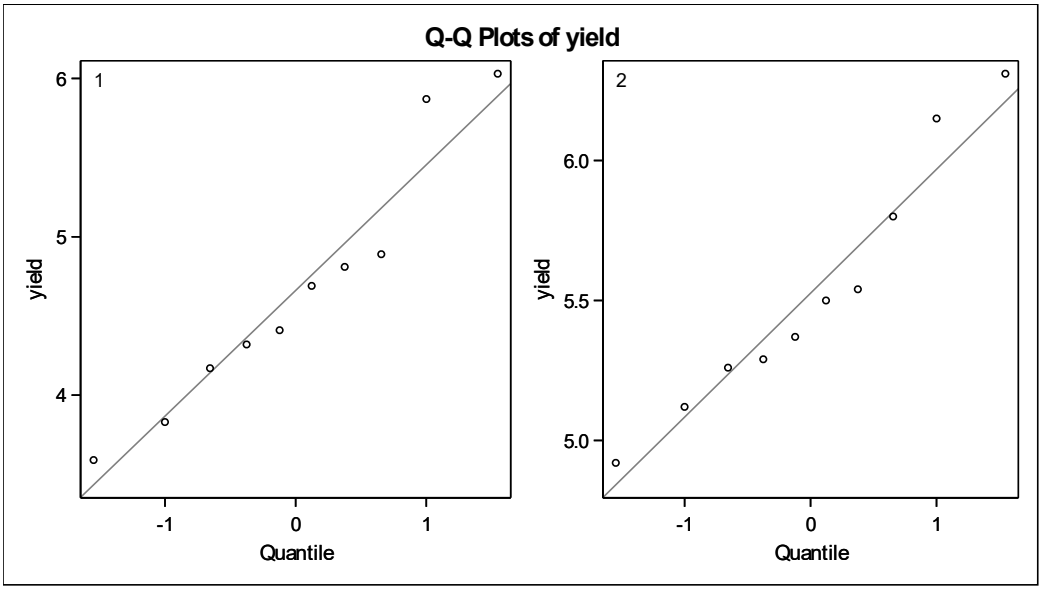
trt	N	Mean	Std Dev	Std Err	Minimum	Maximum
1	10	4.6610	0.7937	0.2510	3.5900	6.0300
2	10	5.5260	0.4426	0.1400	4.9200	6.3100
Diff (1-2)		-0.8650	0.6426	0.2874		

trt	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		4.6610	4.0932	5.2288	0.7937	0.5459	1.4489
2		5.5260	5.2094	5.8426	0.4426	0.3044	0.8080
Diff (1-2)	Pooled	-0.8650	-1.4687	-0.2613	0.6426	0.4855	0.9502
Diff (1-2)	Satterthwaite	-0.8650	-1.4809	-0.2491			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	18	-3.01	0.0075
Satterthwaite	Unequal	14.104	-3.01	0.0093

Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	9	9	3.22	0.0968





b4. Using the SAS outputs obtained in part b3. answer the research question, providing appropriate statistics to support all your conclusions and interpretations. (4 Marks)

Answer: One of the important assumptions of the t-test states that the variance of the two groups should be equal and Folded F test for which the p-value is larger than 0.05 at 5% alpha level

concluding that the variance of the two groups is equal and thus, we interpret the results from the pooled method.

The p-value for the t-stat for the pooled difference is close to zero, concluding that there exists a statistically significant difference between the two treatments. Furthermore, the results show that the average yield of the second treatment is higher than average yield of first treatment. The assumptions are tested as follows:

Assumption 1: The **dependent variable** is a **continuous** scale

Assumption 2: The **independent variable** consists of **categor**ical variable with two treatments.

Assumption 3: There are **no significant outliers** in the **average y**ields from two treatments except for one outlier in treatment 1, as seen from the box plot presented in the output.

Assumption 4: The **distribution for the y**ields for both the treatments is close to normal as it can be seen from the histogram and the Q-Q plot.

Question 2 (20 marks)

Research Question: Is there a significant difference among the three shoe brands index scores?

Answer:

The null hypothesis for the research question states that there exists no significant difference among the three shoe brands index scores. However, the alternate hypothesis for the research question states that there exists a significant difference in the shoe brands index scores for atleast one pair. The question has been tested using one-way ANOVA which is based on the assumption of normality and variance. There are no outliers present in the data for index scores for any of the three brands and follows normal distribution. The equality of variance is tested using Levene's test where the p-value of the F-stat is much larger than 0.05 concluding that the variance of the three groups is equal. Furthermore, the p-value of the F-stat for the research question is close to zero, concluding that there exists a statistically significant difference in the shoe brands index score for atleast two brands. The descriptive statistics show that the average index for the Brand 3 (Pum) has been largest (M=7.7, SD=1.397), followed by Brand 1 (Nik) (M=6.03, SD=0.512) and Brand 2 (Adi) (M=4.6, SD=0.535).

Question 3 (22 marks)

Answer: The preliminary analysis shows that the carbohydrates consumed have been 37.6% with standard deviation of 7.58% with minimum and maximum consumption of 24% and 51%. And the average age of the respondents is 46.15 with standard deviation of 12.77 with minimum and maximum ages being 23 and 64 years indicating that the data is quite diverse. The average

age of respondents in the sample is 110.7 with standard deviation of 16.63 and minimum and maximum weight being 85 kgs and 144 kgs indicating that the sample comprises of comparatively heavier people. The average protein intake of respondents is 15.9% with standard deviation very low at 2.22% and minimum and maximum protein intakes being 12% and 20%. The histogram for carbohydrates consumption is close to normal as the p-value for all the normality tests is larger than 0.05.

The p-value of the F-stat is less than 0.05 at 5% alpha level which concludes that atleast one of the independent variable included in the model have significant impact on the carbohydrates.

R-square (coefficient of determination) for the model is 48% which suggests that all the three variables included in the model explain about 48% variation in carbohydrates.

The analysis of the individual predictor variables shows that the p-values for the age is much larger than 0.05 at 5% alpha level, concluding that age do not significantly impact the carbohydrates. The p-value for the "weight" has been significant at 5% alpha level and for the variable "protein" the t-stat is significant at 1% alpha level. The coefficient of weight has been -0.228 which means that increase in the weight by 1kg results in 23% decrease in carbohydrates and vice versa. On the other hand, increase in protein intake by 1% results in 1.96% increase in carbohydrates and vice versa. The fitted model for carbohydrates can be written as:

$$\text{Carbohydrates} = 36.96 - 0.11368*\text{age} - 0.228*\text{weight} + 1.9577*\text{protein}$$

Model diagnostics:

The predicted value against the residual on the scatter plot shows that there is no evidence of heteroskedasticity. The Q-Q plot shows that the distribution of the error term seems to close to normal as the there are no deviations at any of the ends. The cook's D has been larger than the 0.2 for only one of the data sample. To avoid any multicollinearity issues the correlation matrix has been produced and shows that the correlations between various independent variables are less than 0.6 indicating no collinearity issues.

Improvements:

The R-square of the model has been of medium level and there is a need to add other variables that helps larger variation in the predictor variable.

Question 4 (30 marks)

The preliminary analysis shows that the average infections for the 20 respondents have been 1.39 with standard deviation of 2.34 with minimum and maximum ear-infections being 0 and 17. And the average age of the respondents is 20.77 with standard deviation of 4.32 with minimum and maximum ages being 15 and 29. Since three of the four predictor variables included in the model

are categorical, there should not be any concern for the multicollinearity. The histogram for ear-infections is close to normal as the p-value for all the normality tests is larger than 0.05.

The p-value of the F-stat is close to zero (0.0024) which concludes that atleast one of the independent variable included in the model have significant impact on the infections. However, the R-square for the model has been very low at 5.66% which suggests that there is a need to look for other variables which can explain the variation in infections.

The analysis of the individual predictor variables shows that the p-values for the age and gender have been much larger than 0.05 at 5% alpha level, concluding that these variables do not significantly impact the self-diagnosing the ear infections. The p-value for the "swimmer" has been significant at 1% alpha level and for the variable "location" the t-stat is significant at 5% alpha level. The coefficient of swimmer has been -0.822 which means that that self-diagnosis of the respondents who swim occasionally is larger than the swimmers who swim frequently by 0.822 on an average. On the other hand, on an average, the self diagnosis of number of ear infections for non-beach swimmers is significantly larger than beach swimmers by 0.66. The fitted model for number of ear infections can be written as:

$$\text{infections} = 2.858 - 0.035*\text{age} + 0.027*\text{gender} - 0.822*\text{swimmer} - 0.6688*\text{location}$$

The predicted value against the residual on the scatter plot shows that there is no evidence of heteroskedasticity. The Q-Q plot shows that the distribution of the error term is not close to normal as the there are deviations at both the ends. The values of cook's D have been quiet smaller. To avoid any multicollinearity issues the correlation matrix has been produced and shows that the correlations between various independent variables are less than 0.6 indicating no collinearity issues. The linearity assumption tested using scatter plots show that there is no violation of the assumption.

Appendix

SAS code

```
data drug;
length subjectid  4;
   input subjectid drugA drugB;
   datalines;
0023 25 23
0125 45 41
9834 35 35
2130 50 51
3243 24 20
1231 32 27
3541 37 34
5682 31 30
0012 35 32
0230 43 34
;

proc ttest data=drug;
      paired drugA*drugB;
   run;


data yield;
   length id  3;
   input id yield trt;
   datalines;
1 4.81 1
2 4.17 1
3 4.41 1
4 3.59 1
5 5.87 1
6 3.83 1
```

```
7 6.03 1
8 4.89 1
9 4.32 1
10 4.69 1
11 6.31 2
12 5.12 2
13 5.54 2
14 5.5 2
15 5.37 2
16 5.29 2
17 4.92 2
18 6.15 2
19 5.8 2
20 5.26 2
;

proc ttest data= yield sides=2 alpha=0.05 h0=0;
          title "Two sample t-test example";
          class trt;
          var yield;
  run;

DATA Q2;
 INFILE "/folders/myfolders/Sas_Ass/1523422766_Q2_data.txt" truncover;
 INPUT id $1-1 brand $2-2
rating_color $3-3 rating_work $4-4 pref $5-5 ;
RUN;

PROC FORMAT;
   VALUE brands
     1 = 'Nik'
     2 = 'Adi'
     3 = 'Pum';
   RUN;

   DATA Q2_code;
   SET Q2;
   FORMAT brand brands;
RUN;
```

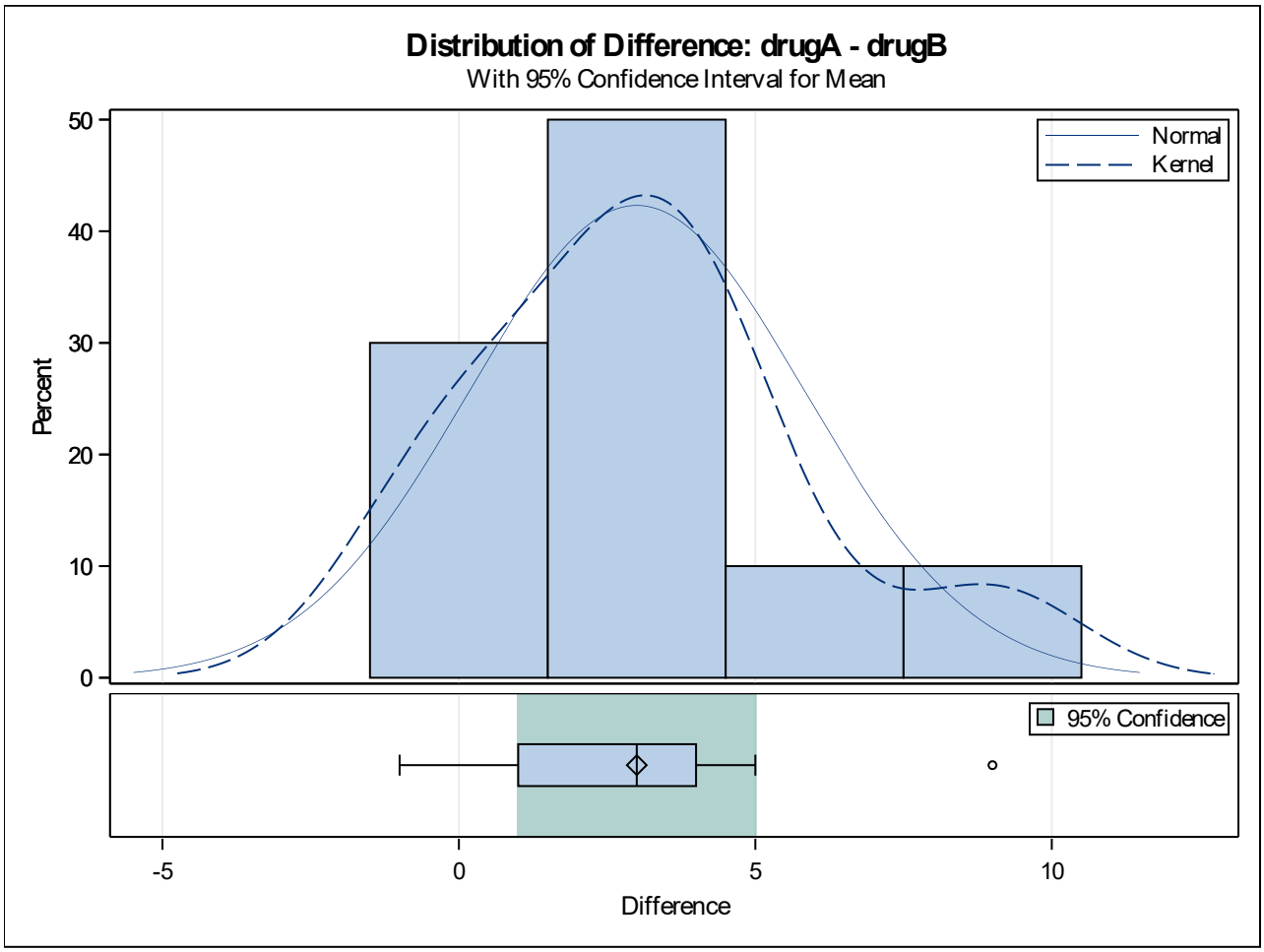
```
data Q2_new;
set Q2_code;
index=(2.5*pref + 1.5*rating_work + rating_color)/5;
run;

proc glm data = Q2_new;
      class brand;
      model index = brand;
      means brand / tukey;
   run;

DATA Q3;
  INFILE "/folders/myfolders/Sas_Ass/1523422766_Q3_data.txt" truncover FIRSTOBS=2;
  INPUT carb age weight protein;
RUN;

proc univariate normaltest data=Q3;
   histogram carb / normal(percents=20 40 60 80 midpercents);
   inset n normal(ksdpval) / pos = ne format = 6.3;
run;

proc reg data=Q3;
      model carb = age weight protein
      / stb clb corrb;
         run;

 proc import datafile="/folders/myfolders/Sas_Ass/Q4_data.csv"
     out=Q4
     dbms=csv
     replace;
     getnames=yes;
run;

          proc reg data=GLMDesign;
         model infections= col2 col3 col5 col7
/corrb;
   run;
```

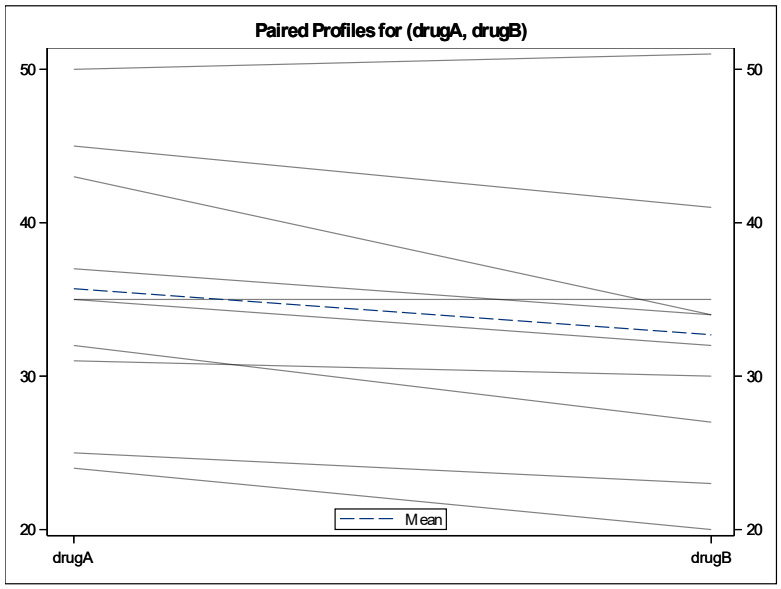
SAS Output

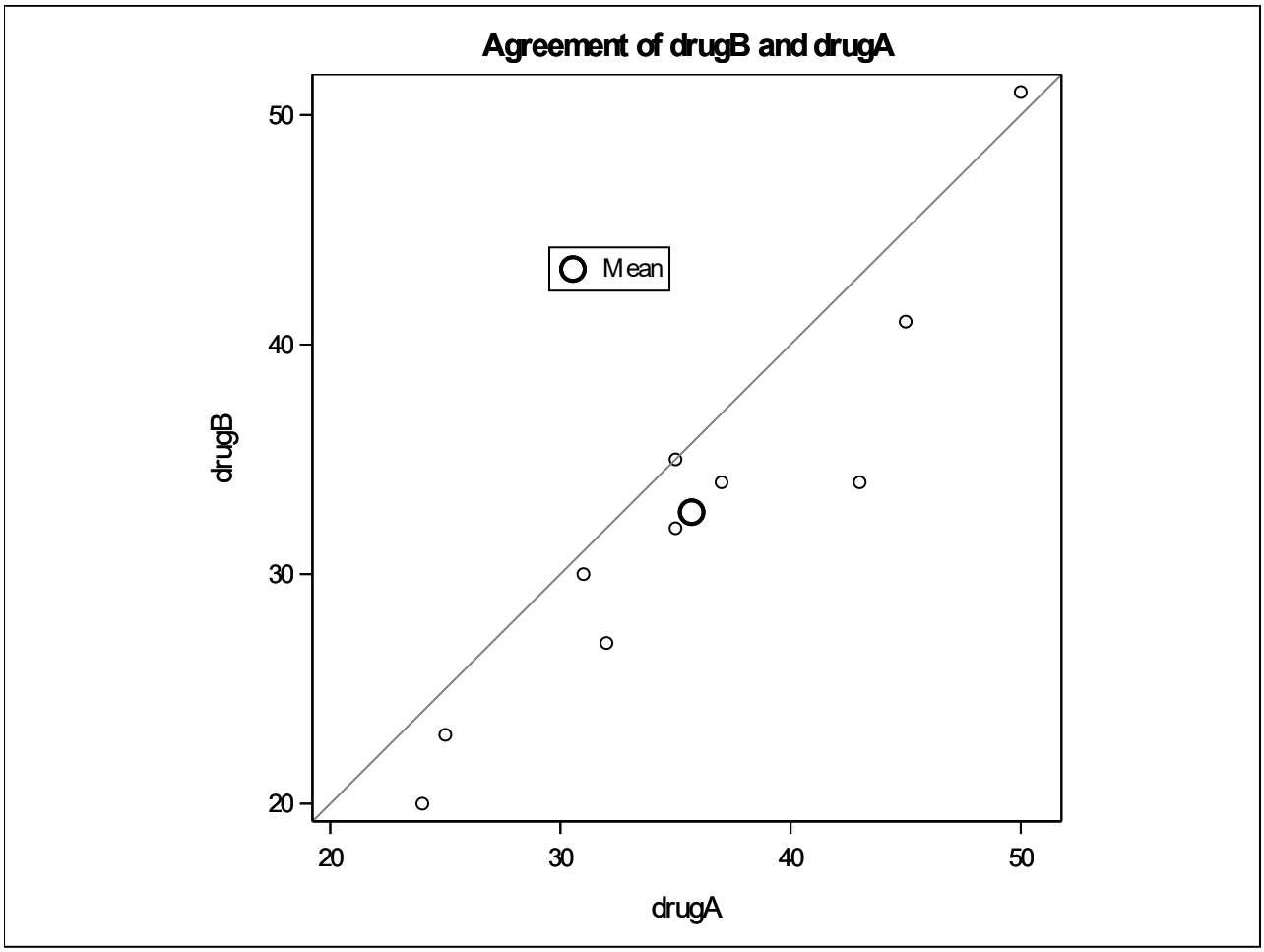
N	Mean	Std Dev	Std Err	Minimum	Maximum
10	3.0000	2.8284	0.8944	-1.0000	9.0000

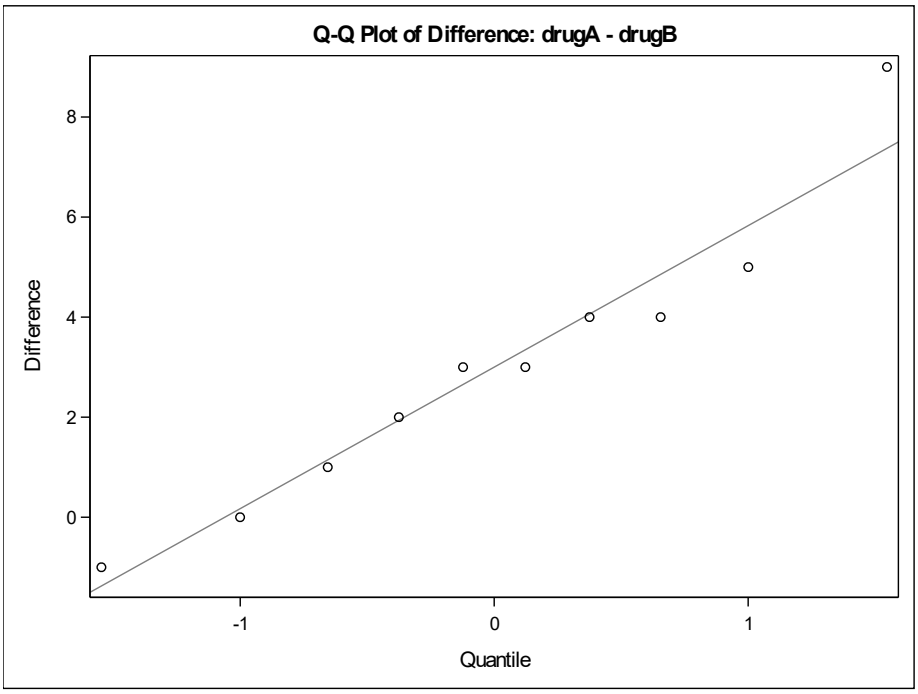
Mean	95% CL Mean		Std Dev	95% CL Std Dev	
3.0000	0.9767	5.0233	2.8284	1.9455	5.1636

DF	t Value	Pr > |t|
9	3.35	0.0085









Two sample t-test example

The TTEST Procedure

Variable: yield

trt	N	Mean	Std Dev	Std Err	Minimum	Maximum
1	10	4.6610	0.7937	0.2510	3.5900	6.0300
2	10	5.5260	0.4426	0.1400	4.9200	6.3100
Diff (1-2)		-0.8650	0.6426	0.2874		

trt	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		4.6610	4.0932	5.2288	0.7937	0.5459	1.4489
2		5.5260	5.2094	5.8426	0.4426	0.3044	0.8080
Diff (1-2)	**Pooled**	-0.8650	-1.4687	-0.2613	0.6426	0.4855	0.9502
Diff (1-2)	**Satterthwaite**	-0.8650	-1.4809	-0.2491			

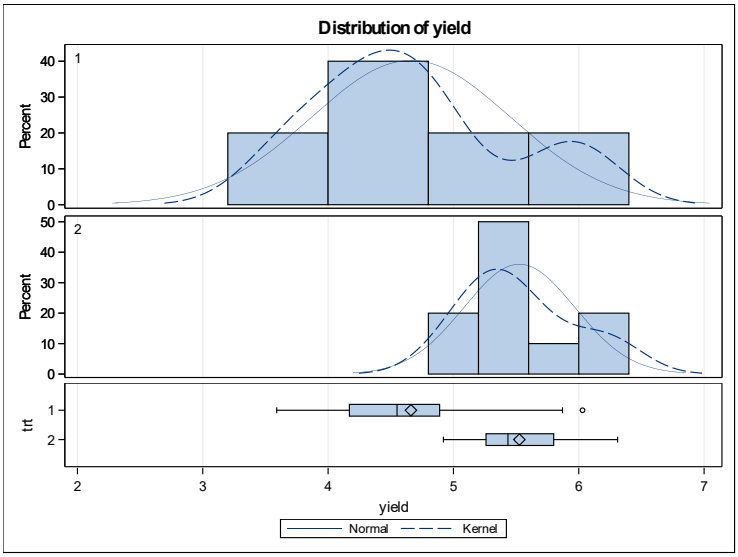
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	18	-3.01	0.0075
Satterthwaite	Unequal	14.104	-3.01	0.0093

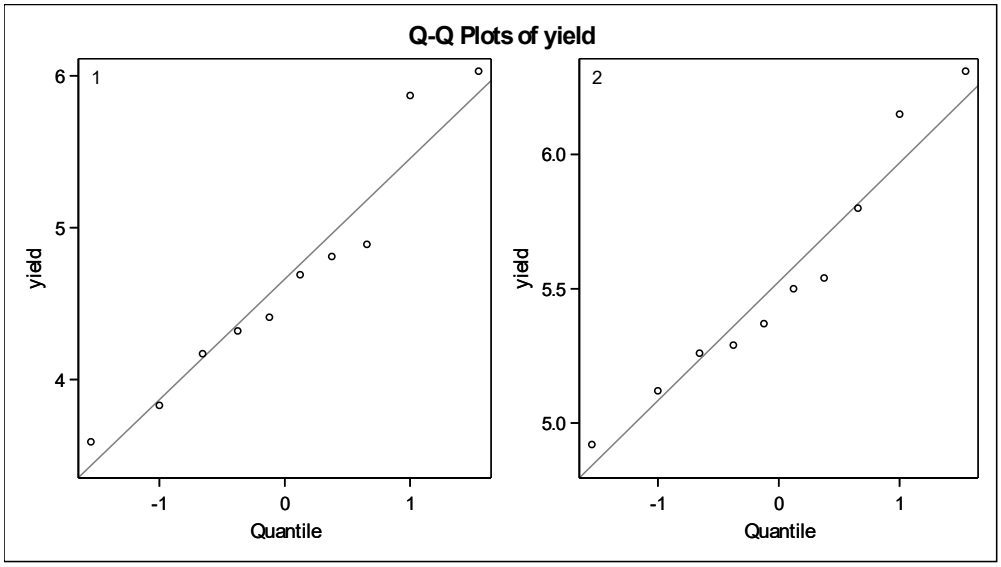
Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	9	9	3.22	0.0968

Two sample t-test example

The TTEST Procedure

Variable: yield





Two sample t-test example

The GLM Procedure

Class Level Information		
Class	**Levels**	**Values**
brand	3	1 2 3

Number of Observations Read	12
Number of Observations Used	12

Two sample t-test example

The GLM Procedure

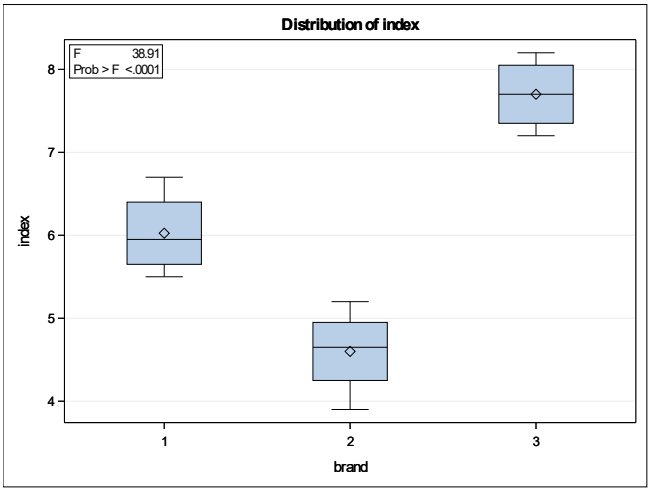
Dependent Variable: index

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	19.26166667	9.63083333	38.91	<.0001
Error	9	2.22750000	0.24750000		
Corrected Total	11	21.48916667			

R-Square	Coeff Var	Root MSE	index Mean
0.896343	8.144508	0.497494	6.108333

Source	DF	Type I SS	Mean Square	F Value	Pr > F
brand	2	19.26166667	9.63083333	38.91	<.0001

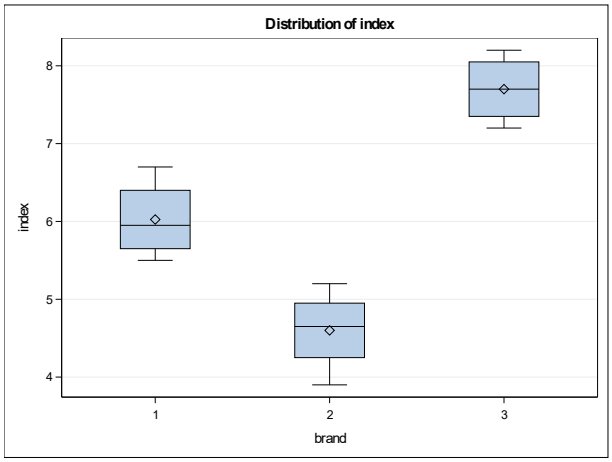
Source	DF	Type III SS	Mean Square	F Value	Pr > F
brand	2	19.26166667	9.63083333	38.91	<.0001



Two sample t-test example

The GLM Procedure

Tukey's Studentized Range (HSD) Test for index



Note This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than
: REGWQ.

Alpha	0.05
Error Degrees of Freedom	9
Error Mean Square	0.2475
Critical Value of Studentized Range	3.94840
Minimum Significant Difference	0.9822

Means with the same letter are not significantly different.			
Tukey Grouping	**Mean**	**N**	**brand**
A	7.7000	4	3
B	6.0250	4	1
C	4.6000	4	2

Two sample t-test example

The UNIVARIATE Procedure
Variable: carb

Moments			
N	20	**Sum Weights**	20
Mean	37.6	**Sum Observations**	752
Std Deviation	7.5839165	**Variance**	57.5157895
Skewness	0.1523864	**Kurtosis**	-0.6683669
Uncorrected SS	29368	**Corrected SS**	1092.8
Coeff Variation	20.1699907	**Std Error Mean**	1.69581528

Basic Statistical Measures			
Location		**Variability**	
Mean	37.60000	**Std Deviation**	7.58392
Median	37.00000	**Variance**	57.51579
Mode	30.00000	**Range**	27.00000
		Interquartile Range	11.00000

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	22.17223	**Pr > |t|**	<.0001
Sign	**M**	10	**Pr >= |M|**	<.0001
Signed Rank	**S**	105	**Pr >= |S|**	<.0001

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.976197	**Pr < W**	0.8762
Kolmogorov-Smirnov	**D**	0.091857	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.020803	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.164436	**Pr > A-Sq**	>0.2500

Two sample t-test example

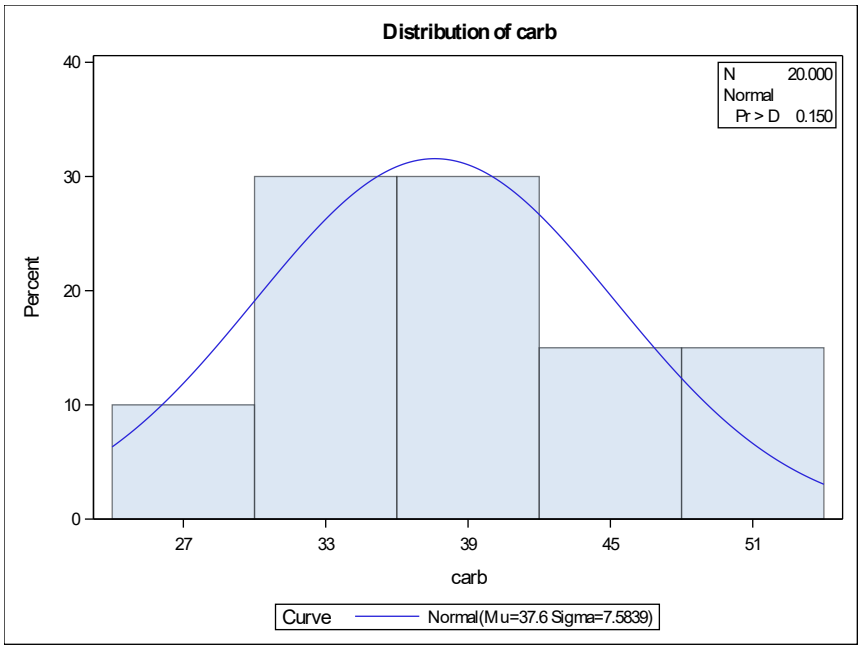
The UNIVARIATE Procedure
Variable: carb

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	51.0
99%	51.0
95%	50.5
90%	49.0
75% Q3	42.5
50% Median	37.0
25% Q1	31.5
10%	28.5
5%	25.5
1%	24.0
0% Min	24.0

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
24	18	43	6
27	4	46	17
30	13	48	8
30	9	50	11
30	5	51	12

Two sample t-test example

The UNIVARIATE Procedure
Fitted Normal Distribution for carb



Parameters for Normal Distribution		
Parameter	**Symbol**	**Estimate**
Mean	Mu	37.6
Std Dev	Sigma	7.583916

Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.09185736	**Pr > D**	>0.150
Cramer-von Mises	**W-Sq**	0.02080339	**Pr > W-Sq**	>0.250
Anderson-Darling	**A-Sq**	0.16443617	**Pr > A-Sq**	>0.250

Two sample t-test example

The UNIVARIATE Procedure
Fitted Normal Distribution for carb

Histogram Bin Percents for Normal Distribution		
Bin Midpoint	Percent	
	Observed	Estimated
27	10.000	12.168
33	30.000	25.831
39	30.000	30.265
45	15.000	19.576
51	15.000	6.985

Quantiles for Normal Distribution		
Percent	Quantile	
	Observed	Estimated
20.0	30.0000	31.2172
40.0	35.5000	35.6786
60.0	39.0000	39.5214
80.0	44.5000	43.9828

Two sample t-test example

The REG Procedure
Model: MODEL1
Dependent Variable: carb

Number of Observations Read	20
Number of Observations Used	20

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	525.13714	175.04571	4.93	0.0130
Error	16	567.66286	35.47893		
Corrected Total	19	1092.80000			

Root MSE	5.95642	**R-Square**	0.4805
Dependent Mean	37.60000	**Adj R-Sq**	0.3831
Coeff Var	15.84154		

Parameter Estimates								
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Standardized Estimate**	**95% Confidence Limits**	
Intercept	1	36.96006	13.07128	2.83	0.0121	0	9.25017	64.66994
age	1	-0.11368	0.10933	-1.04	0.3139	-0.19148	-0.34544	0.11808
weight	1	-0.22802	0.08329	-2.74	0.0146	-0.50007	-0.40458	-0.05145
protein	1	1.95771	0.63489	3.08	0.0071	0.57356	0.61180	3.30363

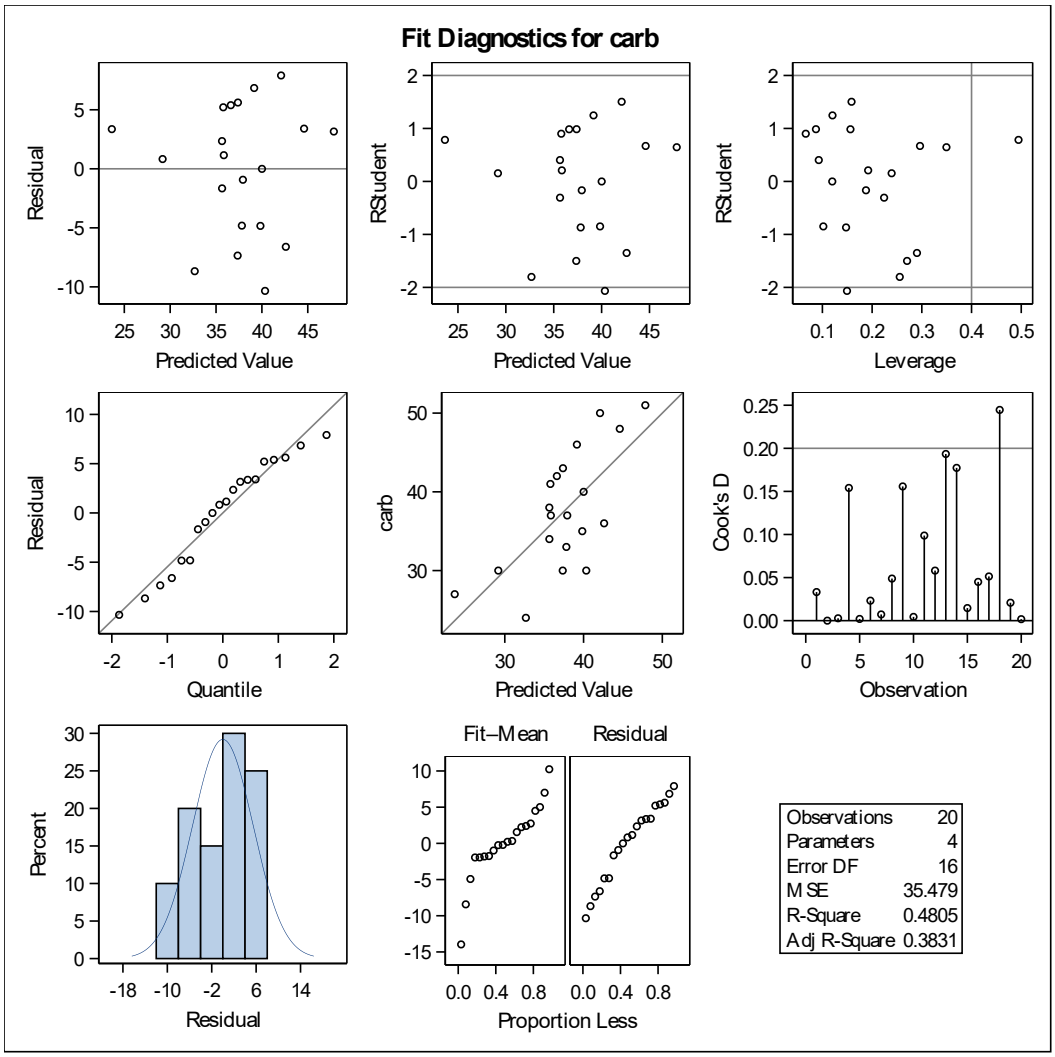
Two sample t-test example

The REG Procedure
Model: MODEL1
Dependent Variable: carb

Correlation of Estimates				
Variable	**Intercept**	**age**	**weight**	**protein**
Intercept	1.0000	-0.2818	-0.6112	-0.5823
age	-0.2818	1.0000	0.0735	-0.2021
weight	-0.6112	0.0735	1.0000	-0.1587
protein	-0.5823	-0.2021	-0.1587	1.0000

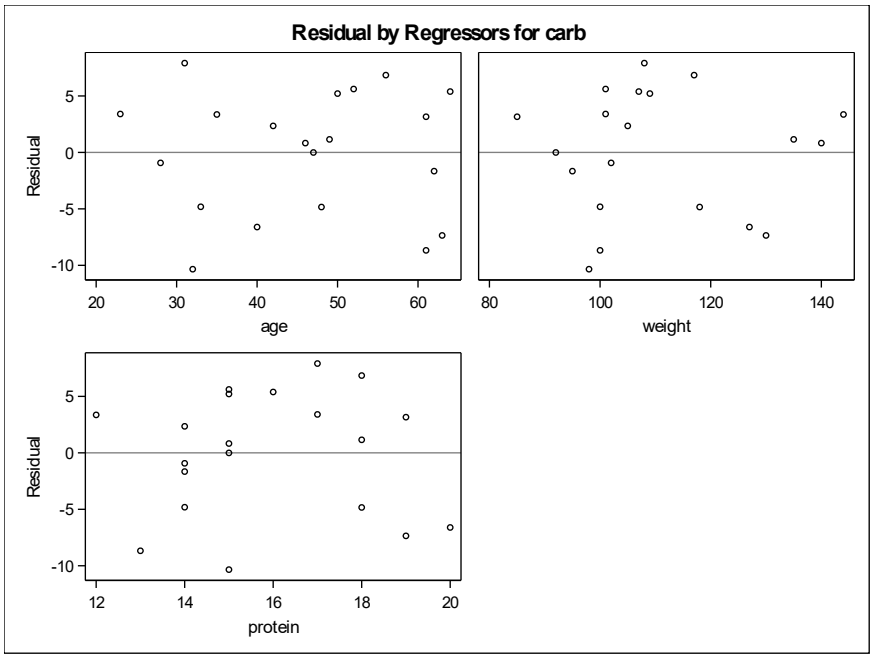
Two sample t-test example

The UNIVARIATE Procedure
Variable: INFECTIONS



Two sample t-test example

The UNIVARIATE Procedure
Variable: INFECTIONS



Moments			
N	287	**Sum Weights**	287
Mean	1.38675958	**Sum Observations**	398
Std Deviation	2.33854124	**Variance**	5.46877513
Skewness	3.20185866	**Kurtosis**	14.180533
Uncorrected SS	2116	**Corrected SS**	1564.06969
Coeff Variation	168.633501	**Std Error Mean**	0.13803972

Basic Statistical Measures			
Location		**Variability**	
Mean	1.386760	**Std Deviation**	2.33854
Median	0.000000	**Variance**	5.46878
Mode	0.000000	**Range**	17.00000
		Interquartile Range	2.00000

Two sample t-test example

The UNIVARIATE Procedure
Variable: INFECTIONS

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	10.04609	**Pr > |t|**	<.0001
Sign	**M**	68	**Pr >= |M|**	<.0001
Signed Rank	**S**	4658	**Pr >= |S|**	<.0001

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.624089	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.27659	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	5.609732	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	30.88889	**Pr > A-Sq**	<0.0050

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	17
99%	11
95%	5
90%	4
75% Q3	2
50% Median	0
25% Q1	0
10%	0
5%	0
1%	0
0% Min	0

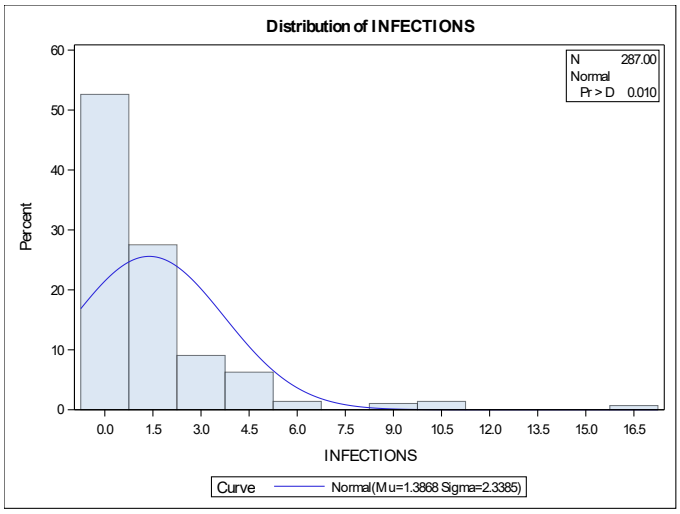
Two sample t-test example

The UNIVARIATE Procedure
Variable: INFECTIONS

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	284	10	65
0	283	10	249
0	282	11	30
0	278	16	31
0	277	17	47

Two sample t-test example

The UNIVARIATE Procedure
Fitted Normal Distribution for INFECTIONS



Parameters for Normal Distribution		
Parameter	**Symbol**	**Estimate**
Mean	Mu	1.38676
Std Dev	Sigma	2.338541

Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.2765899	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	5.6097318	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	30.8888909	**Pr > A-Sq**	<0.005

Two sample t-test example

The UNIVARIATE Procedure
Fitted Normal Distribution for INFECTIONS

Histogram Bin Percents for Normal Distribution		
Bin Midpoint	Percent	
	Observed	Estimated
0.0	52.613	21.227
1.5	27.526	25.129
3.0	9.059	19.990
4.5	6.272	10.684
6.0	1.394	3.836
7.5	0.000	0.924
9.0	1.045	0.149
10.5	1.394	0.016
12.0	0.000	0.001
13.5	0.000	0.000
15.0	0.000	0.000
16.5	0.697	0.000

Quantiles for Normal Distribution		
Percent	Quantile	
	Observed	Estimated
20.0	0.00000	-0.58141
40.0	0.00000	0.79430
60.0	1.00000	1.97922
80.0	2.00000	3.35493

Two sample t-test example

The GLMMOD Procedure

Class Level Information		
Class	**Levels**	**Values**
GENDER	2	Fema Male
SWIMMER	2	Frequent Occasional
LOCATION	2	Beach NonBeach

Number of Observations Read	287
Number of Observations Used	287

Two sample t-test example

The GLMMOD Procedure

Parameter Definitions				
Column Number	Name of Associated Effect	CLASS Variable Values		
		GENDER	SWIMMER	LOCATION
1	Intercept			
2	AGE			
3	GENDER	Fema		
4	GENDER	Male		
5	SWIMMER		Frequent	
6	SWIMMER		Occasional	
7	LOCATION			Beach
8	LOCATION			NonBeach

Two sample t-test example

The REG Procedure
Model: MODEL1
Dependent Variable: INFECTIONS

Number of Observations Read	287
Number of Observations Used	287

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	4	88.54514	22.13628	4.23	0.0024
Error	282	1475.52455	5.23236		
Corrected Total	286	1564.06969			

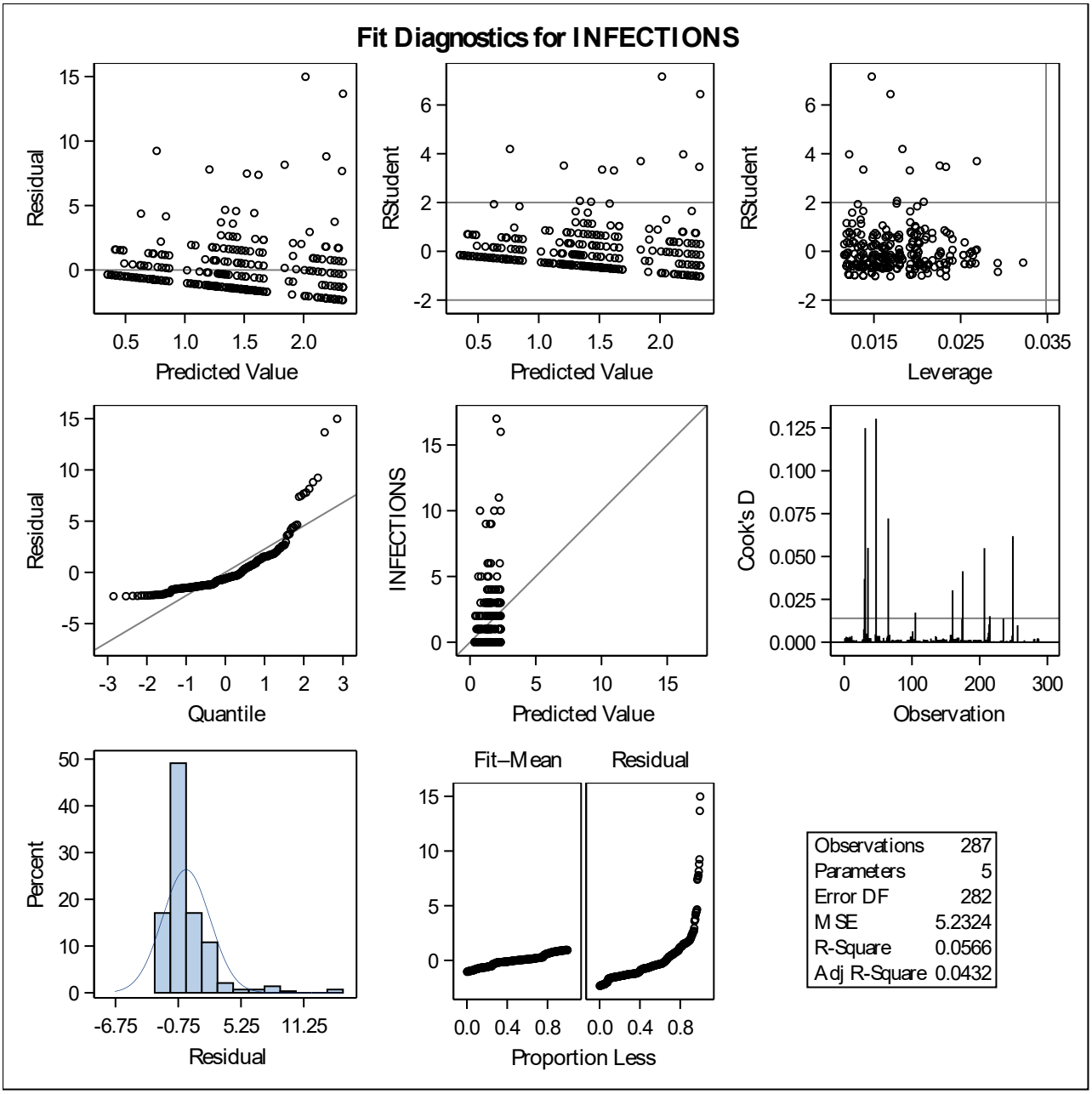
Root MSE	2.28743	**R-Square**	0.0566
Dependent Mean	1.38676	**Adj R-Sq**	0.0432
Coeff Var	164.94817		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	2.85815	0.68223	4.19	<.0001
Col2	AGE	1	-0.03508	0.03154	-1.11	0.2670
Col3	GENDER Fema	1	0.02706	0.28860	0.09	0.9254
Col5	SWIMMER Frequent	1	-0.82209	0.27014	-3.04	0.0026
Col7	LOCATION Beach	1	-0.66884	0.27509	-2.43	0.0157

Two sample t-test example

The REG Procedure
Model: MODEL1
Dependent Variable: INFECTIONS

Correlation of Estimates						
Variable	**Label**	**Intercept**	**Col2**	**Col3**	**Col5**	**Col7**
Intercept	Intercept	1.0000	-0.9307	-0.0580	-0.1949	-0.0987
Col2	AGE	-0.9307	1.0000	-0.0610	-0.0083	-0.0913
Col3	GENDER Fema	-0.0580	-0.0610	1.0000	0.0173	-0.1587
Col5	SWIMMER Frequent	-0.1949	-0.0083	0.0173	1.0000	0.0150
Col7	LOCATION Beach	-0.0987	-0.0913	-0.1587	0.0150	1.0000


Fit Diagnostics for INFECTIONS
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
INFECTIONS
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 287
Parameters 5
Error DF 282
MSE 5.2324
R-Square 0.0566
Adj R-Square 0.0432

The REG Procedure
Model: MODEL1
Dependent Variable: INFECTIONS