HW-4

MS -Business Intelligence & Analytics
Spring 2016
BIA – 654 A

Feb 21, 2016

Mohit Ravi Ghatikar
CWID - 10405877

Ethics Statement

I pledge on my honor that I have not given or received any unauthorized assistance on this assignment/examination. I further pledge that I have not copied any material from a book, article, the Internet or any other source except where I have expressly cited the source.

Signature _Mohit Ravi Ghatikar_______________ Date: 02/21/2016___

HW-4

1)

We need to conduct a 2-sample t-test to check if the two means obtained from different populations are equal or not. To conduct a 2-sample t-test, the sample are independent and should come from a normal distribution and the population variances must be equal. These conditions are satisfied.

Point Estimate = X_1 bar – X_2 bar

= 75 – 86

= -11

S_1^2 = 120

S_2^2 = 100

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)}$$

S_p^2 = (12-1) * 120 + (12-1) * 100 / 22

= 110

The Test statistic is:

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

t = (-11) – 0 / (110 * 0.16)^2

= -2.56

Null Hypothesis: The means are not different. ($\mu_1 - \mu_2 = 0$)

Alternate Hypothesis: The means are different. ($\mu_1 - \mu_2 \neq 0$)

We calculated the t value to be -2.56

The p-value at 95% confidence level is for 22 degrees of freedom(n1 + n2 - 2) is: 0.017856

Since the p-value is less than the alpha value of 0.05, we can reject the null hypothesis.

Thus, the mean test scores of the 2 groups are significantly different from each other with 95% confidence interval.

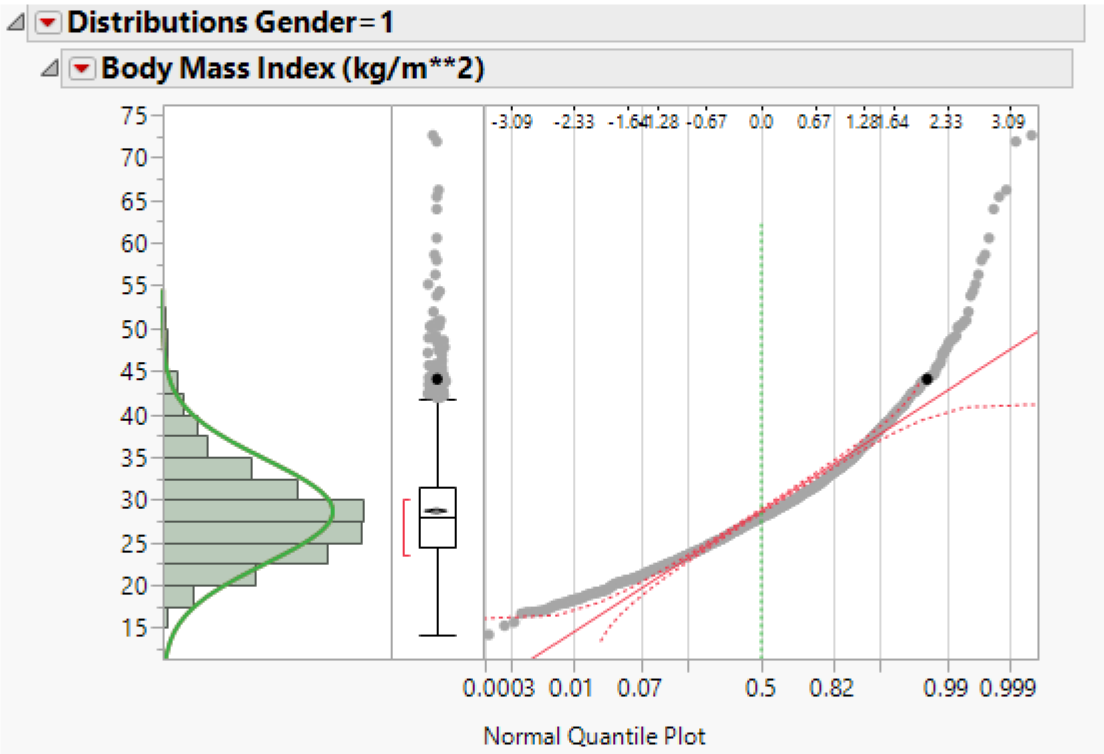
2)

We need to check if the Variances of BMI are equal for Men and Women.

We assume for Gender is male if it is equal to 1 and female if it is equal to 2.

We also check if the samples are coming from a normal distribution.

For Males:



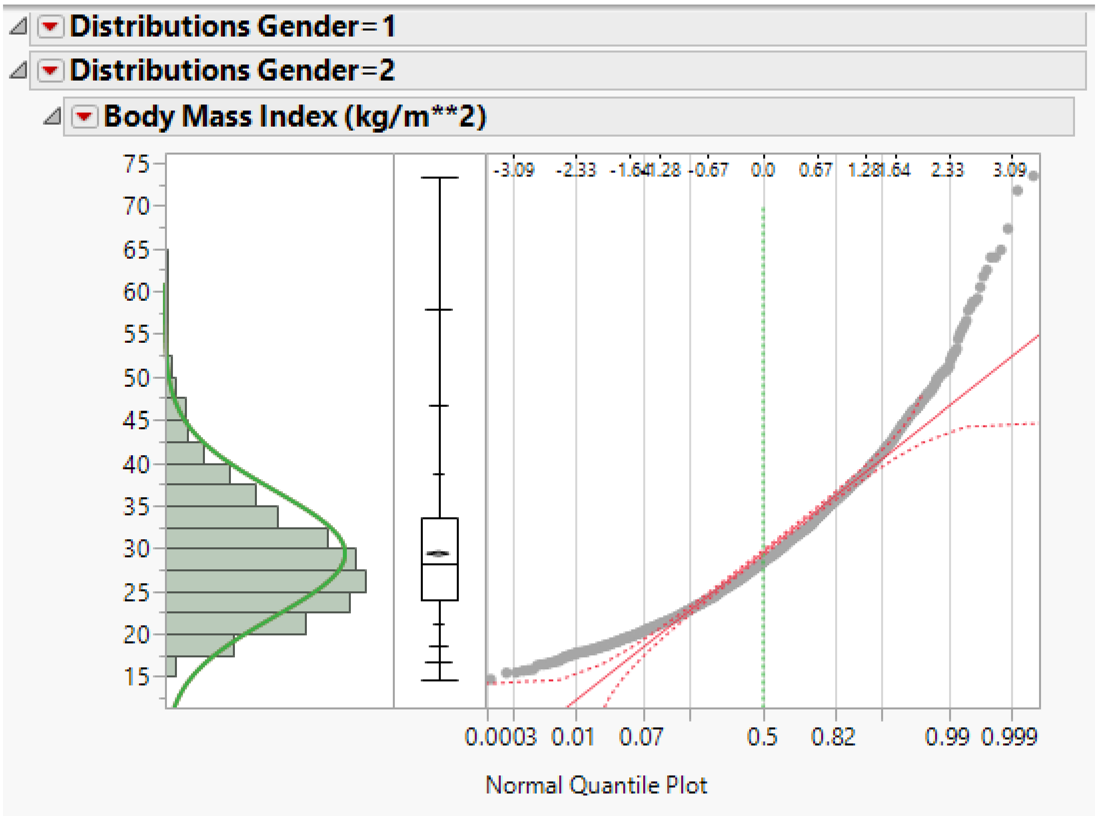
Goodness-of-Fit Test

KSL Test

D		Prob>D
0.076766	<	0.0100*

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

For Females:



Goodness-of-Fit Test

KSL Test

D		Prob>D
0.065034	<	0.0100*

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

As we can see the samples are not normally distributed. We cannot use F-test if the normality assumption is not met irrespective of the sample size.

Therefore we use Levene's test (or Brown-Forsyth test.)

Null Hypothesis: $Sigma_1^2 = Sigma_2^2$

Alternate Hypothesis: $Sigma_1^2 \neq Sigma_2^2$

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	38.3136	1	5386	<.0001*
Brown-Forsythe	92.7899	1	5386	<.0001*
Levene	102.6607	1	5386	<.0001*
Bartlett	100.8123	1	.	<.0001*
F Test 2-sided	1.4748	2724	2662	<.0001*

Since the p-value of <0.0001 is less than alpha of 0.05, we reject the null hypothesis. There is sufficient evidence to prove that the variances are not equal with a 95% confidence interval.

Next, we conduct a 2-sample t-test assuming unequal variances.

Null Hypothesis: The means are not different. ($\mu_1 - \mu_2 = 0$)

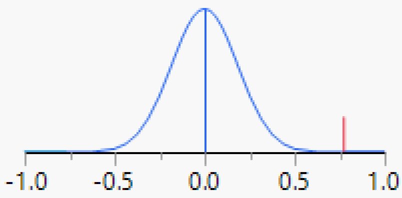
Alternate Hypothesis: The means are different. ($\mu_1 - \mu_2 \neq 0$)

t Test

2-1

Assuming unequal variances

Difference	0.76907	t Ratio	4.17737
Std Err Dif	0.18410	DF	5236.02
Upper CL Dif	1.12999	Prob > |t|	<.0001*
Lower CL Dif	0.40815	Prob > t	<.0001*
Confidence	0.95	Prob < t	1.0000



Again, the p-value of 0.0001 is less than alpha of 0.05. Therefore we can reject the null hypothesis and conclude that the means of BMI are different for both the genders at 95% confidence interval.

3)

We need to test whether or not the mean lengths of cuckoo eggs found in the nests of the three foster-parent species are the same. We can use ANOVA to carry out the analysis.

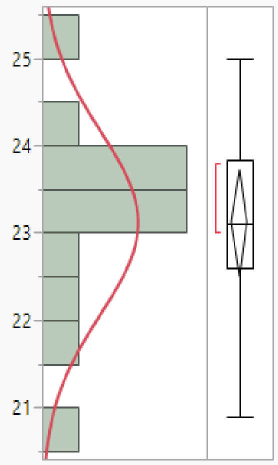
The assumptions for ANOVA are:

- Populations are normally distributed.
- Populations have equal variances.
- Samples are randomly and independently drawn.

Normality assumption:

Distributions Nest of=Hedge Sparrow

Cuckoo eggs length (mm)



Normal(23.1143,1.04944)

Quantiles

Summary Statistics

Fitted Normal

Parameter Estimates

Goodness-of-Fit Test

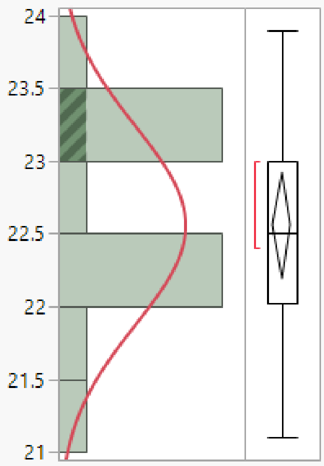
Shapiro-Wilk W Test

W	Prob<W
0.955599	0.6505

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

Distributions Nest of=Robin

Cuckoo eggs length (mm)



Normal(22.5563,0.68212)

Quantiles

Summary Statistics

Fitted Normal

Parameter Estimates

Goodness-of-Fit Test

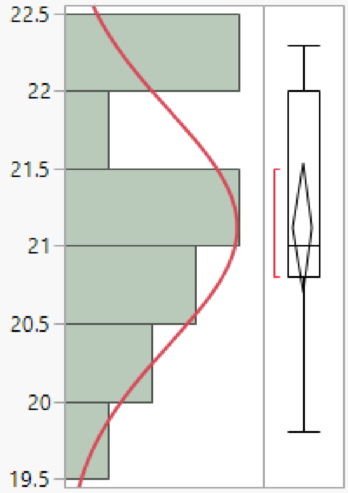
Shapiro-Wilk W Test

W	Prob<W
0.965255	0.7571

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

Distributions Nest of=Wren

Cuckoo eggs length (mm)



Normal(21.12,0.75423)

Quantiles

Summary Statistics

Fitted Normal

Parameter Estimates

Goodness-of-Fit Test

Shapiro-Wilk W Test

W	Prob<W
0.941847	0.4062

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

By running the Shapiro-wilk test on all three variables, we conclude that they come from a normal distribution since the p-values are greater than the alpha value of 0.05

Variance equality assumption:

ANOVA assumes that the samples have equal variances. To check this assumption this we use Leven's test.

Null Hypothesis: $Sigma_1^2 = Sigma_2^2 = Sigma_3^2$

Alternate Hypothesis: At least one of the variance is not equal to the other.

Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
Hedge Sparrow	14	1.049437	0.7591837	0.7571429
Robin	16	0.682123	0.5437500	0.5437500
Wren	15	0.754226	0.5840000	0.5600000

Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	1.4334	2	42	0.2499
Brown-Forsythe	0.7124	2	42	0.4963
Levene	0.7087	2	42	0.4981
Bartlett	1.4243	2	.	0.2407

Since the p-value for Leven's test is 0.4981 which is greater than 0.05, we fail to reject the null hypothesis at 95% confidence interval. Thus the variances are equal.

We can now proceed with the ANOVA calculations:

Null Hypothesis: $\mu_1 = \mu_2 = \mu_3$

Alternate Hypothesis: Not all population means are the same

Oneway Anova

Summary of Fit

Rsquare	0.515333
Adj Rsquare	0.492254
Root Mean Square Error	0.834673
Mean of Response	22.25111
Observations (or Sum Wgts)	45

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Nest of	2	31.111927	15.5560	22.3287	<.0001*
Error	42	29.260518	0.6967		
C. Total	44	60.372444			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Hedge Sparrow	14	23.1143	0.22308	22.664	23.564
Robin	16	22.5563	0.20867	22.135	22.977
Wren	15	21.1200	0.21551	20.685	21.555

Std Error uses a pooled estimate of error variance

The F-ratio is 22.3287 with the p-value of 0.0001. Since the p-value is less than 0.05, we reject the Null hypothesis. Thus we conclude that the mean lengths of cuckoo eggs for Hedge sparrow, Robin and Wren are not the same with 95% confidence interval.

4)

Null Hypothesis: $\mu_1 = \mu_2 = \mu_3$

Alternate Hypothesis: Not all population means are the same

a)

1	2	3
9.5	8.5	7.7
3.2	9	11.3
4.7	7.9	9.7
7.5	5	11.5
8.3	3.2	12.4

X1(bar) = 6.64, X2(bar) = 6.72, X3(bar) = 10.52

X(bar) = Average of (6.64,6.72,10.52) = 7.96

n1 = n2 = n3 = 5

n = 15

c=3

SSB = $5(6.64-7.96)^2 + 5(6.72 - 7.96)^2 + 5(10.52 - 7.96)^2 = 49.168$

SSW = $(9.5 - 6.64)^2 + (3.2 - 6.64)^2 + ...+ (12.4 - 10.52)^2 = 66.108$

SST = SSB+SSW = 115.276

MSB = SSB / (c-1) = 49.168 / 2 = 24.584

MSW = SSW / (n-c) = 66.108 / 12 = 5.509

$F_{STAT} = MSB/MSW = 4.4625$

The p-value is 0.0356. Since the p-value is less than 0.05, we can reject the null hypothesis. Thus there is sufficient evidence to conclude that not all the population means are the same with 95% confidence interval.

b) Using SAS JMP:

Oneway Anova

Summary of Fit

Rsquare	0.426524
Adj Rsquare	0.330945
Root Mean Square Error	2.347126
Mean of Response	7.96
Observations (or Sum Wgts)	15

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Store	2	49.16800	24.5840	4.4625	0.0356*
Error	12	66.10800	5.5090		
C. Total	14	115.27600			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
1	5	6.6400	1.0497	4.3530	8.927
2	5	6.7200	1.0497	4.4330	9.007
3	5	10.5200	1.0497	8.2330	12.807

Std Error uses a pooled estimate of error variance

We get the same F-ratio as calculated previously. Therefore we reject the Null Hypothesis and conclude that not all population means are the same.