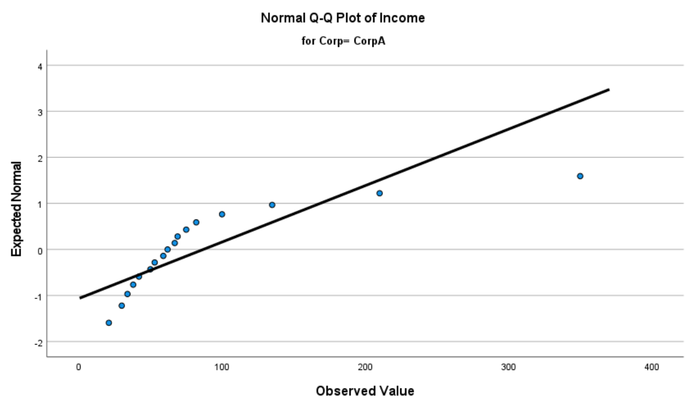
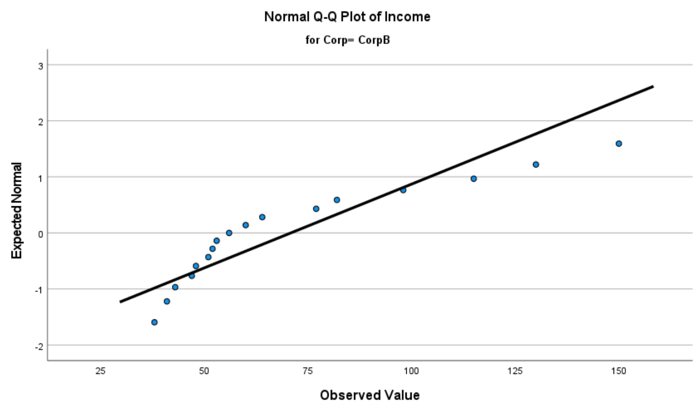
Practice Problem Topic 2-3: Transforming and Back Transforming

Do employees in two corporations have a difference in average annual income?

Comparing Mean Annual Income in Two Corporations
Random samples of 17 employees were selected from each of two corporations (Corp A and Corp B) and the annual income of each employee was recorded (in thousands of dollars), obtaining data as shown below.

Annual Income (in thousands of dollars)	
Corp A	**Corp B**
21	38
30	41
34	43
38	47
42	48
50	52
53	51
59	53
62	56
67	60
69	64
75	77
82	82
100	98
135	115
210	130
350	150





Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Corp	Statistic	df	Sig.	Statistic	df	Sig.
Income	CorpA	.289	17	<.001	.688	17	<.001
	CorpB	.229	17	.019	.837	17	.007

a. Lilliefors Significance Correction

Group Statistics					
	Corp	N	Mean	Std. Deviation	Std. Error Mean
Income	CorpA	17	86.88	81.562	19.782
	CorpB	17	70.88	33.449	8.113

Test of Homogeneity of Variance					
		Levene Statistic	df1	df2	Sig.
Income	Based on Mean	2.817	1	32	.103
	Based on Median	1.270	1	32	.268
	Based on Median and with adjusted df	1.270	1	20.578	.273
	Based on trimmed mean	1.663	1	32	.207

(a) What test would you choose to test if there is a difference in average annual income of employees between the two corporations. Do the data meet the required assumptions?

This should be tested with a two-sample t-test for independent samples because there is one categorical explanatory variable (corporation) with two levels and the response variable is a continuous, quantitative variable, that is, annual income. The purpose is to test if there is a difference between two population means.

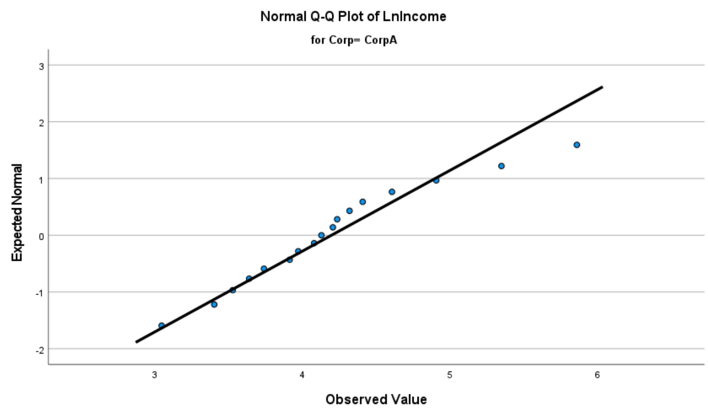
The assumptions:

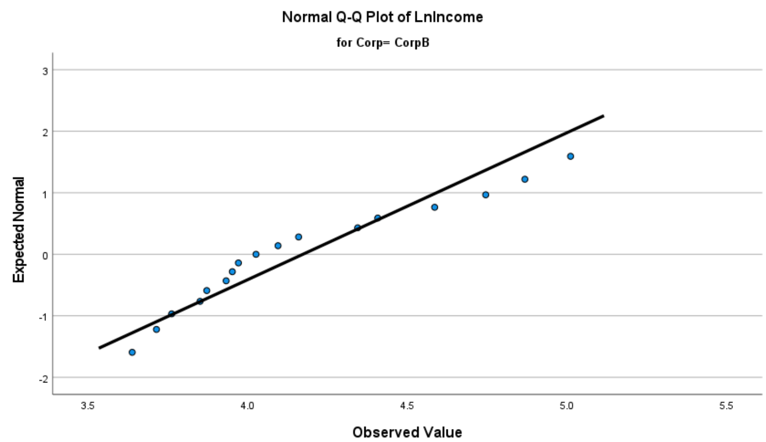
1. Random selection from population (which is met)
2. Two independent samples (which is met)
3. The Shapiro Wilk test results in both P-values being less than 0.05 (< 0.001 and 0.007); therefore, the assumption of normality is NOT met.
4. Based on means, Levene's test for equality of variances gives P = 0.103, which is greater than 0.05; therefore, the assumption of equal variances is met. [Note: Levene's test is more accurate than using the ratio of the standard deviations because that gives 81.562/33.449 = 2.44, which is greater than 2 and would indicate that the standard deviations are different. When you have the results for Levene's test, use that instead of the ratio of standard deviations.]

Therefore, the pooled two-sample t-test cannot be applied since the data do not come from normally distributed populations.

Let's try a Natural Log Transformation (Ln)
The natural log transformed data are shown below along with analysis

Annual Income (in thousands of dollars)	
Corp A	**Corp B**
3.045	3.638
3.401	3.714
3.526	3.761
3.638	3.850
3.738	3.871
3.912	3.951
3.970	3.932
4.078	3.970
4.127	4.025
4.205	4.094
4.234	4.159
4.317	4.344
4.407	4.407
4.605	4.585
4.905	4.745
5.347	4.868
5.858	5.011





Tests of Normality							
		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Corp	Statistic	df	Sig.	Statistic	df	Sig.
LnIncome	CorpA	.146	17	.200*	.956	17	.560
	CorpB	.166	17	.200*	.916	17	.125

*. This is a lower bound of the true significance.
a. Lilliefors Significance Correction

Test of Homogeneity of Variance					
		Levene Statistic	df1	df2	Sig.
LnIncome	Based on Mean	1.701	1	32	.201
	Based on Median	1.738	1	32	.197
	Based on Median and with adjusted df	1.738	1	26.545	.199
	Based on trimmed mean	1.720	1	32	.199

(b) Do the data meet the required assumptions of the selected test after log transformation?

Checking the assumptions based on the logged data:

1. Random selection from population (which is met)
2. Two independent samples (which is met)
3. The Shapiro Wilk test results in both P-values being greater than 0.05 (0.560 and 0.125); therefore, the assumption of normality is met.
4. Based on means, Levene's test for equality of variances gives P = 0.201, which is greater than 0.05; therefore, the assumption of equal variances is met.

Therefore, the pooled two-sample t-test can be performed on the logged data since the data fit all the assumptions.

Below is SPSS Output for the two-sample t-test based on log transformed data

Group Statistics

	Corp	N	Mean	Std. Deviation	Std. Error Mean
LnIncome	CorpA	17	4.194871	.7025184	.1703857
	CorpB	17	4.172014	.4186192	.1015301

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance				95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	Lower	Upper
LnIncome	Equal var. assumed	1.701	.201	.115	32	.454	.909	.0228573	.1983423	-.3811527	.4268672
	Equal var. not assumed			.115	26.090	.455	.909	.0228573	.1983423	-.3847724	.4304870

Although the SPPS output is shown, answer the questions below without using the numbers highlighted in yellow.

(c) At the 5% significance level, test whether there was a difference in average logged annual income of employees between the two corporations.

Step 1: Already selected the pooled two-sample t-test and checked the assumptions.
Step 2:
H_0: $\mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$ (There is no difference in average logged annual income of employees between the two corporations.)
Ha: $\mu_1 \neq \mu_2$ or $\mu_1 - \mu_2 \neq 0$ (There is a difference in average logged annual income of employees between the two corporations.)
Parameter: $\mu_1 - \mu_2 = \mu_{CorpA} - \mu_{CorpB}$

>>>>>>>>>>
Step 3:
Estimate of the difference between means = $\bar{y}_1 - \bar{y}_2 = 4.194871 - 4.172014 = 0.022857$
(in thousands of dollars)

Estimate of the pooled population standard deviation:

$$s_p = \sqrt{\frac{(n_1-1)s_1{}^2 + (n_2-1)s_2{}^2}{n_1+n_2-2}}$$
$$= \sqrt{\frac{(17-1)(0.7025184)^2 + (17-1)(0.4186192)^2}{17+17-2}} = 0.578262$$

Standard error of the estimate of the difference between means:
$$SE(\bar{y}_{CorpA} - \bar{y}_{CorpB}) = s_p\sqrt{(1/n_1)+(1/n_2)}$$
$$= 0.578262\sqrt{(1/17)+(1/17)} = 0.198342$$

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s_p\sqrt{(1/n_1)+(1/n_2)}} = \frac{\bar{y}_1 - \bar{y}_2}{SE(\bar{y}_1 - \bar{y}_2)} = \frac{0.022857}{0.198342}$$
$$= 0.115$$

Step 4: $df = n_1 + n_2 - 2 = 17 + 17 - 2 = 32 \approx 30$
P-value: (P > 0.25) x 2 = P > 0.50 [SSPS: P-value = 0.909]
There is weak evidence against H_0 because P-value is greater than 10% (Guidelines)
P > α (0.05), therefore do not reject H_0.

Step 5: At the 5% significance level, the data do not provide sufficient evidence to conclude that there is a difference in average logged annual income of employees between the two corporations.

>>>>>>>>>>
(d) Determine a 95% confidence interval for the difference in average logged annual income of employees between the two corporations.
>>>>>>>>>>

Step 1: Critical value is:
For a 95% confidence interval, α = 1 – 0.95 = 0.05.
At $df = n_1 + n_2 - 2 = 17 + 17 - 2 = 32 \approx 30$, $t_{\alpha/2} = t_{0.05/2} = t_{0.025} = 2.042$

Step 2:
Parameter: $\mu_1 - \mu_2 = \mu_{CorpA} - \mu_{CorpB}$
Estimate = $\bar{y}_1 - \bar{y}_2 = 4.194871 - 4.172014 = 0.022857$ (in thousands of dollars)
Standard error of the estimate: $SE(\bar{y}_{CorpA} - \bar{y}_{CorpB}) = 0.198342$

Calculation of the confidence interval:

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2} \times SE(\bar{y}_1 - \bar{y}_2)$$

$$0.022857 \pm 2.042 \times 0.198342$$

0.022857 ± 0.40501

(-0.382, 0.428) thousand dollars

Step 3: We can be 95% confident that the difference in average logged annual income of employees between the two corporations is between -0.382 and 0.428 thousand dollars.

>>>>>>>>>>

>>>>>>>>>>

(e) Back transform the confidence interval you found in part (d) to the original scale and interpret the meaning of this confidence interval on the original scale.

Taking the antilog of both endpoints we get the confidence interval on the original scale is:

$$(e^{-0.382}, e^{0.428}) = (0.6825, 1.5342)$$

Interpretation:

A 95% confidence interval for the ratio of the medians on the original scale $\left[\frac{Med(CorpA)}{Med(CorpB)}\right]$ is

$(0.6825, 1.5342)$

OR

It is estimated with 95% confidence that the median annual income of Corporation A is between 0.6825 and 1.5342 times the median annual income of Corporation B (in thousands of dollars).

(f) Based on the confidence interval you found in part (e) after back transformation, would you conclude that there is a difference in median annual income between Corporation A and Corporation B (with 95% confidence)? Explain the logic of your answer.

Since the confidence interval after back transformation, $(0.6825, 1.5342)$, contains 1, at the 95% confidence level, there is insufficient evidence that there is a difference in median annual income between Corporation A and Corporation B.

>>>>>>>>>>