SOLUTIONS TO ASSIGNMENT 1

Question 1

Data were collected on 96 species of mammals from zoo records and the literature. These mammals were grouped on the basis of the observed average litter size. This is an observational study, the group status of the mammals is established based on observe characteristic, which is beyond the control of the investigator, of the mammals. A cause-and-effect inference could only be drawn from carefully designed randomized experiment. The finding is limited to the sample of mammals used for the study since the sample can by no means represent the population of mammal given that data were gathered from just available zoo records and the literature.

Question 2

To obtain the summary statistics, side-by-side boxplots, and normality plots for each group, click on *Descriptive Statistics* in *Analyze* menu. Selecting *Explore* invokes a dialog box with "*Both"* option (both Statistics and Plots) as the default. In order to obtain the normality plots for each group, make sure that this option is selected in *Plots.*

(a) The summary statistics for the two groups are displayed below (small litter size is coded as "0" and large litter size is coded as "1"):

Descriptives

	size			Statistic	Std. Error
brainsiz	0	Mean		6.8859	.76459
		95% Confidence Interval for Mean	Lower Bound	5.3501	
			Upper Bound	8.4216	
		5% Trimmed Mean		6.5247	
		Median		5.0000	
		Variance		29.815	
		Std. Deviation		5.46030	
		Minimum		.42	
		Maximum		20.00	
		Range		19.58	
		Interquartile Range		8.00	
		Skewness		.816	.333
		Kurtosis		-.355	.656
	1	Mean		10.9684	1.46640
		95% Confidence Interval for Mean	Lower Bound	8.0131	
			Upper Bound	13.9238	
		5% Trimmed Mean		10.1645	
		Median		7.9700	
		Variance		96.765	
		Std. Deviation		9.83692	
		Minimum		.94	
		Maximum		36.35	
		Range		35.41	
		Interquartile Range		15.37	
		Skewness		1.092	.354
		Kurtosis		.228	.695

The sample mean for the small litter size group is 6.89 while that of the large group is 10.97. Thus in this sample, mammals with large litter sizes has larger relative brain weight compared to mammals with small litter sizes. The standard deviation for small litter size group is 5.46, compared to 9.84 for the large litter size group.

The standard errors of the small litter size and large litter size groups are, respectively, .765 and 1.466. The standard error of the mean is a measure of how much the value of the mean may vary from sample to sample from the same distribution. In order words, it is a measure of the accuracy of the sample mean as an estimate of the population mean. The smaller the value of the standard error, the more reliable is the sample mean as an estimate of the population mean.

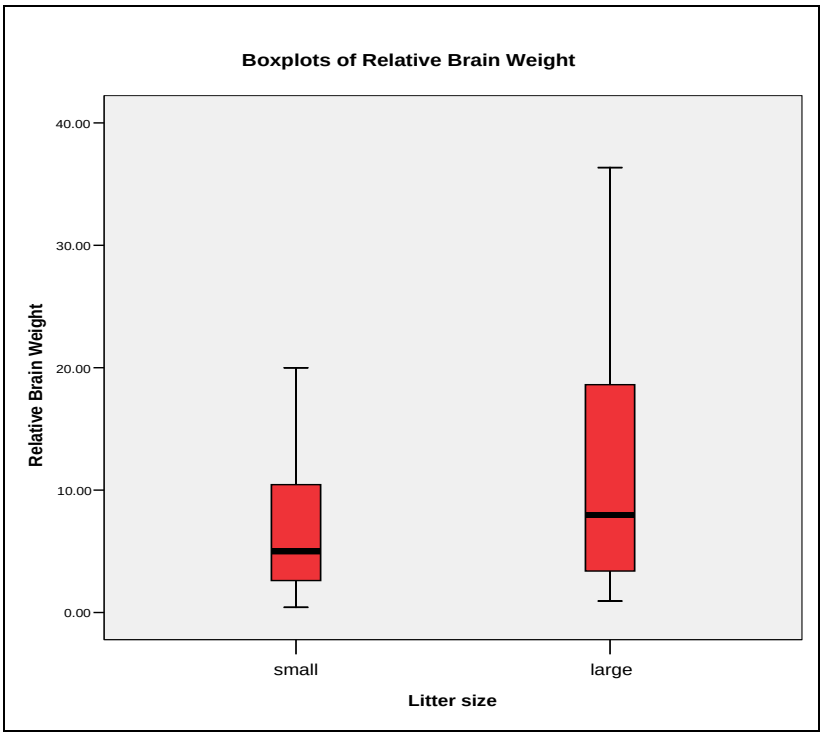
(b) The 95% confidence interval for mean relative brain weight is (5.350, 8.422) for small litter size group and (8.013, 13.924) for large litter size group. The 95% confidence interval for the large litter size group is wider. To explain why it is wider, consider the formula for confidence interval:

Sample mean ± t critical value*standard error.

Given the same critical values for both confidence interval (since both are 95%), the group with the larger standard error would have a wider confidence interval. Thus the shorter the confidence interval is the more precise the sample mean as an estimate of the population mean.

Yes, the two confidence intervals have a small overlap. If the overlap is substantial one might begin to think that the two groups may not have really distinct mean relative brain weight.

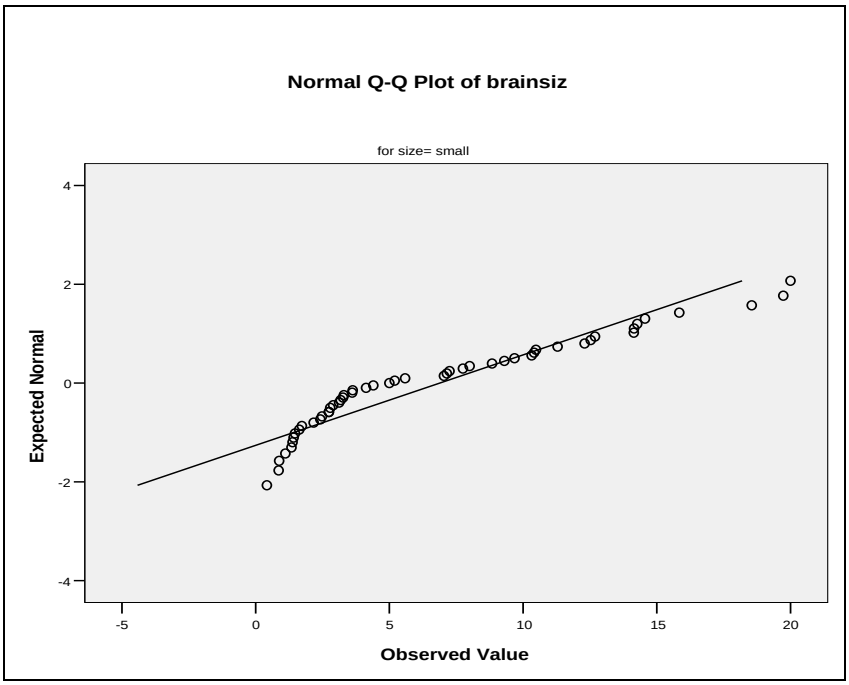
(c) The side-by-side boxplot of relative brain weight for the two groups is displayed below:



Clearly the median of the large litter size group has higher median than the small litter size group. Also, the spread of the two groups are very different - the interquartile range of the large litter size group (15.37) is wider than that of the small litter size group (IQR=8 for small litter size group).

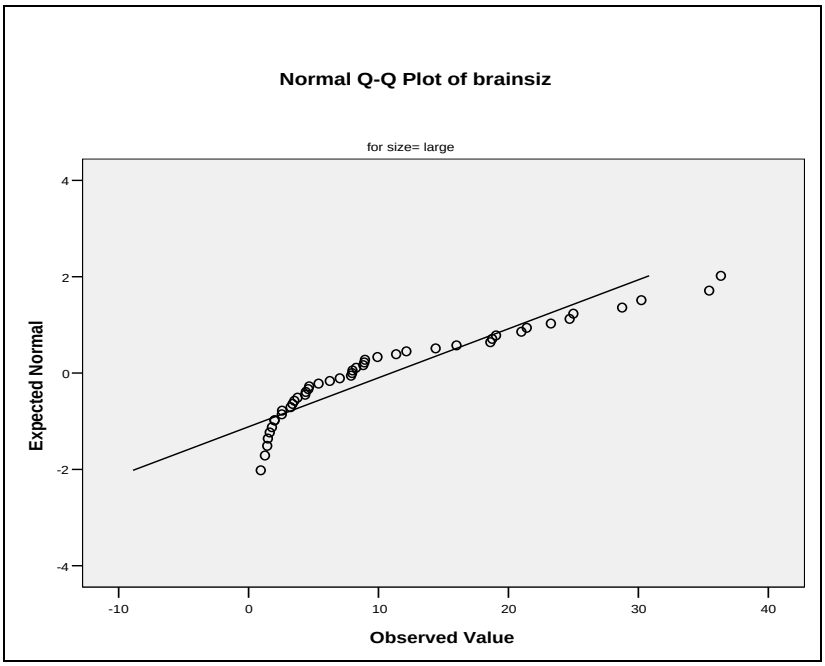
The position of the medians and the whiskers indicate that both distributions may be skewed to the right. Given the skewness of both distributions, the median is recommended as a measure of central tendency since it is not quickly influenced by extreme observations. There are clearly no outliers in both distributions.

(d) The normal quantile plot for the small litter size group is displayed below:



There are systematic departures from a straight-line pattern in the above plot. The plot shows that the distribution of relative brain weight for the small litter size group is skewed to the right.

The normal quantile plot for the large litter size group is displayed below:



There are also systematic departures from a straight-line pattern in this plot. The plot shows that the distribution of relative brain weight for the large litter size group is skewed to the right. The assumption of normality necessary to apply the t-tools may be seriously violated in either case.

(e) The assumptions necessary for the application of t-tools are:

- The samples are independently selected from two populations,
- The populations are normally distributed.

We do not have enough information to assess independence. Usually the validity of independence assumption depends on how the data was collected. This is a critical assumption and its violation may warrant a different statistical tool other than t-tool. The normality assumption seems to be violated in this data.

Question 3

(a) The summary statistics for the log-transformed relative brain weight is displayed below:

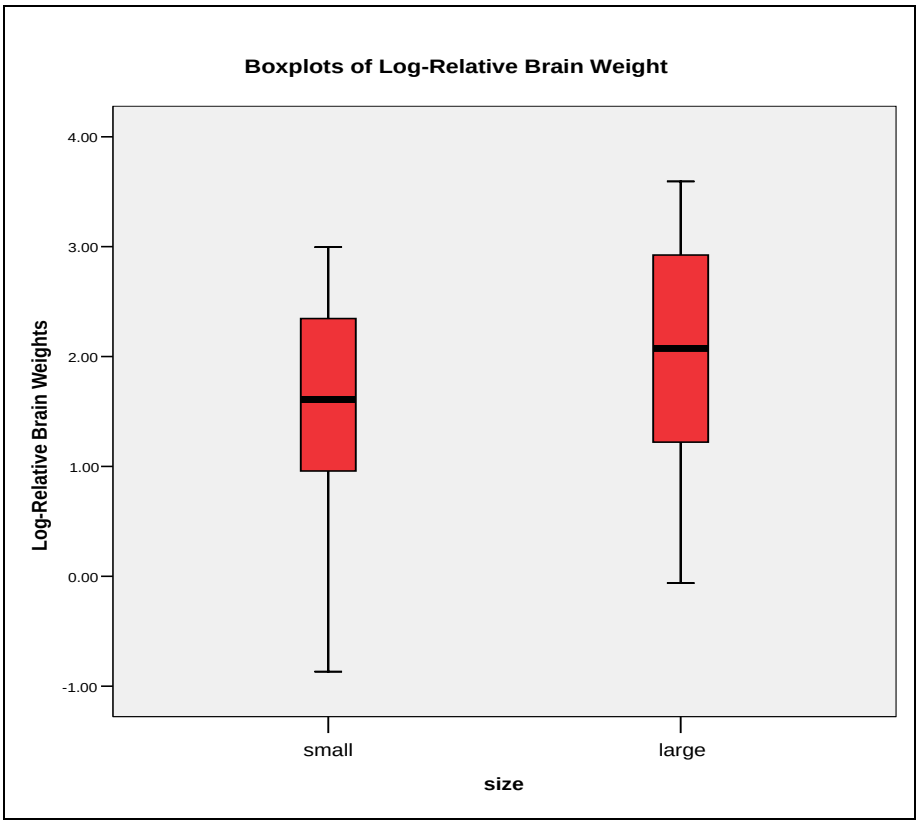
Descriptives

	size			Statistic	Std. Error
logbrain	small	Mean		1.5525	.13334
		95% Confidence Interval for Mean	Lower Bound	1.2846	
			Upper Bound	1.8203	
		5% Trimmed Mean		1.5834	
		Median		1.6094	
		Variance		.907	
		Std. Deviation		.95223	
		Minimum		-.87	
		Maximum		3.00	
		Range		3.86	
		Interquartile Range		1.44	
		Skewness		-.390	.333
		Kurtosis		-.645	.656
	large	Mean		1.9494	.15150
		95% Confidence Interval for Mean	Lower Bound	1.6441	
			Upper Bound	2.2548	
		5% Trimmed Mean		1.9617	
		Median		2.0757	
		Variance		1.033	
		Std. Deviation		1.01629	
		Minimum		-.06	
		Maximum		3.59	
		Range		3.66	
		Interquartile Range		1.73	
		Skewness		-.162	.354
		Kurtosis		-1.022	.695

The mean and standard deviation of the log-transformed relative brain weight are 1.55 and .95, respectively, for small litter size group; and 1.94 and 1.02 for large litter size group. On the log-scale the mean of the large litter size group is still higher than the small litter size group.

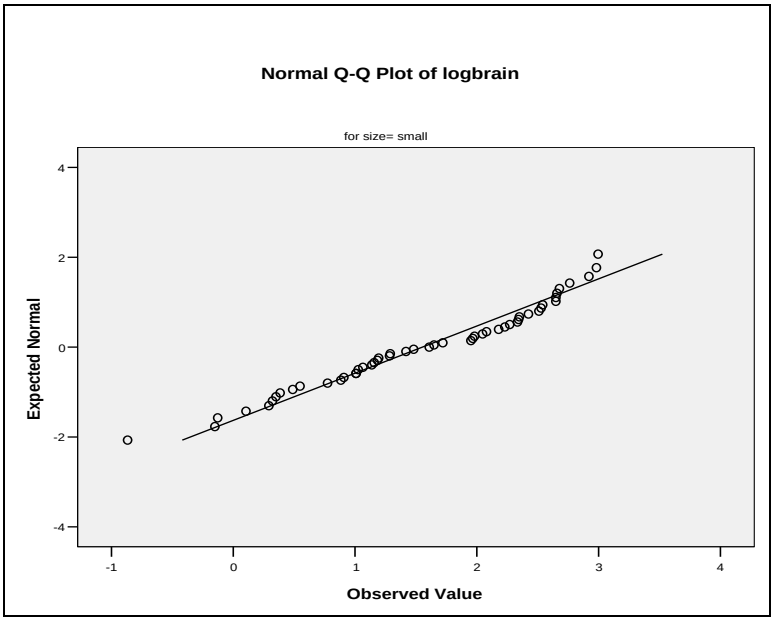
The standard error of the mean, on the log-scale, for the small litter size group (.133) is less than that of the large litter size group (.152).

(b) The side-by-side boxplot for the two groups are:



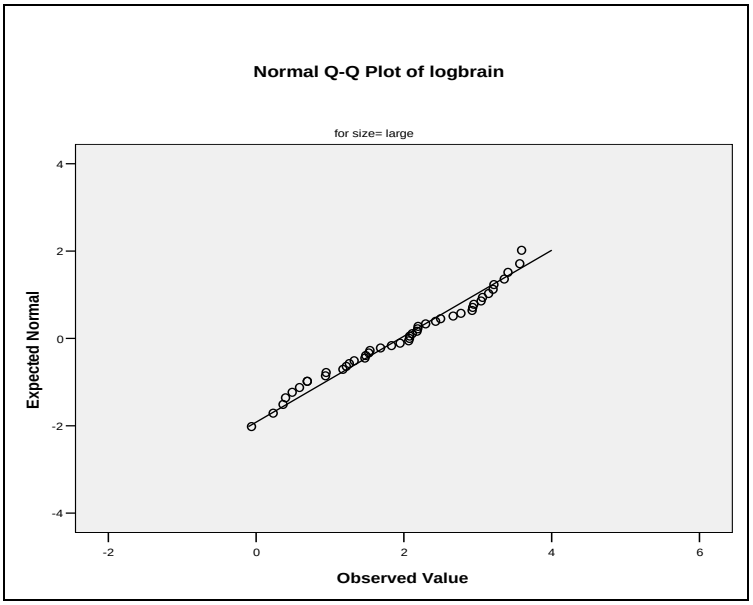
The median of the (log-) relative brain weight of the large litter size group is (2.08) higher than (1.61) that of the small litter size group. The two distributions appear to be symmetric; neither of the two is skewed. The spread of the two distributions are about the same.

(c) The normal quantile plot for log-relative brain weight for the small litter size group is



The log-transformed data follows approximately a normal distribution.

Also the normal quantile plot for the large litter size group shown below also indicates that the assumption of normality is feasible for the log-transformed data.



(d) The log-transformation appears to be very successful in eliminating the skewness present in the data on the original scale of measurement. Also, the assumption of normality is feasible for the log-transformed data.

Question 4

(a) Let $\bar{z}_L$ and $\bar{z}_S$ represent the averages of the logged values for the large litter size sample and small litter size sample, respectively.

Using the SPSS output in part (a) of Question 3, the difference in averages for the log transformed data is

$\bar{z}_L - \bar{z}_S = 1.9494 - 1.5525 = .3969$

The antilogarithm is = $\exp(\bar{z}_L - \bar{z}_S) = 1.49$.

For symmetric distributions, the mean and the median are approximately equal. As the log-transformed data have approximately symmetric distributions, the following relationships hold:

$$\bar{z}_L - \bar{z}_S \approx \text{Median}[\log(y_L)] - \text{Median}[\log(y_S)] = \log[\text{Median}(y_L)] - \log[\text{Median}(y_S)]$$

The last equality follows from the fact that log preserves ordering.

It follows that,

$$\bar{z}_L - \bar{z}_S \text{ estimates } \log\left[\frac{\text{Median}(y_L)}{\text{Median}(y_S)}\right]$$

and, therefore

$$\exp(\bar{z}_L - \bar{z}_S) \text{ estimates } \left[\frac{\text{Median}(y_L)}{\text{Median}(y_S)}\right].$$

Therefore the ratio

$$\left[\frac{\text{Median}(y_L)}{\text{Median}(y_S)}\right] \text{ can be estimated by 1.49.}$$

In other words, the median relative brain weight for large litter size mammals is 1.49 times as large as the median relative brain weight for small litter size mammals. Equivalently, the median relative brain weight for large litter size mammals is estimated to be 49% more than the median relative brain weight for small litter size mammals.

Note to grader: Please note that this solution is a lot detailed than the answers that most student will provide. If a student interprets simply as "the relative brain weight for large litter size mammals is 1.49 times as large as the relative brain weight for small litter size mammals", you **may** still award full mark.

(b) The related SPSS output is displayed below:

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
logbrain	Equal variances assumed	.167	.684	-1.975	94	.051	-.39697	.20099	-.79605	.00211
	Equal variances not assumed			-1.967	90.684	.052	-.39697	.20182	-.79788	.00394

The null and alternative hypotheses are as follows:

H_0 : there is no difference between groups on log relative brain weight,
H_a : there is difference between groups on log relative brain weight.

The side-by-side boxplot for the log-transformed data in Question 3 exhibited approximately the same spread for the two groups. Moreover, the outcome of the Levene's test in the above output confirms that there is no evidence against equality of variances assumption. Thus, the relevant test in the output above is the one that assume equal variances.

Under the null hypothesis, the test statistic t follows a t distribution with 94 degrees of freedom. The value of the test statistic is -1.975, and p-value is .051. This indicates that there is little or no evidence against the null hypothesis of no difference between groups on log relative brain weight.

(c) The 95% confidence interval for the difference in log-relative brain weight is (-.796, .002). To obtain the 95% confidence interval on the original scale we take the antilogarithm of the endpoints. Thus, the 95% confidence interval for difference in relative brain weight is (.451, 1.002).

Note to grader: The interval (-.002, .796) is equally valid for the log-scale and corresponding antilogarithm (.998, 2.217), for the original scale.

The 95% confidence interval for difference in relative brain weight contains 1. Meaning that 1 is a plausible value for the ratio of the median relative brain weight. Thus, equality is plausible.

The conclusion is thus consistent with that of the hypothesis test.

Question 5

The pooled estimate of the standard deviation is:

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

$$s_p = \sqrt{\frac{(51-1)*.9522^2 + (45-1)*1.0163^2}{(51+45-2)}} = .9827$$

and the standard error for the difference in sample averages is:

$$\text{SE}(\bar{z}_L - \bar{z}_S) = s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = .9827*\sqrt{\frac{1}{51}+\frac{1}{45}} = .20099.$$

The t-statistic is:

$$t = \frac{\bar{z}_L - \bar{z}_S}{\text{SE}(\bar{z}_L - \bar{z}_S)} = \frac{.39697}{.20099} = 1.975.$$

The calculation is consistent with the SPSS output.

Question 6

(a) The related SPSS output is displayed below:

Ranks

	size	N	Mean Rank	Sum of Ranks
brainsiz	small	51	43.57	2222.00
	large	45	54.09	2434.00
	Total	96		

Test Statistics[a]

	brainsiz
Mann-Whitney U	896.000
Wilcoxon W	2222.000
Z	-1.847
Asymp. Sig. (2-tailed)	.065

a. Grouping Variable: size

The null and alternative hypotheses are as follows:

H_0 : there is no difference in relative brain sizes between the two groups,

H_a : there is difference in relative brain sizes between the two groups.

The Mann-Whitney U statistic = 896 with a z approximation of -1.847 and p-value = .065 (U can be approximated by a normal distribution and standardized with the standard normal Z).

(b) The p-value indicates that there is no difference in relative brain weight between the two groups. Thus the outcome of the test is consistent with the test in Question 4.

Question 7

Other factors that might be considered include gestation length, habitat (whether mammals live on trees, holes or surface ground), eating habit (whether mammals are omnivores, carnivores, or herbivores etc).

LAB ASSIGNMENT 1 MARKING SCHEMA

Proper Header and appearance: 10 points

Question 1

Description of study: 2 points
Study design: 1 point
Type of inference: 1 point
Scope of inference: 1 point

Question 2

(a) Descriptive statistics: 4 points (2 points for each group)
Comparison of mean and standard deviations: 2 points
Standard errors and meaning: 4 points
(b) Confidence intervals: 4 points (2 points each)
Wider interval and explanation: 3 points
Overlap: 1 point
(c) Boxplots: 4 points
Comparisons of median and interquartile range: 2 points
Shape of distribution: 2 points
Measure of central tendency and justification: 3 points
Outliers: 1
(d) Normal probability plot: 4 points
Description of pattern: 2 points
(e) T-tool assumptions: 2 points
Violation: 1 point

Question 3

(a) Descriptive statistics: 4 points (2 points per group)
Comparison of mean and standard deviations: 2 points
Comparison of standard errors: 2 points
(b) Boxplots: 4 points
Comparison of median: 2 points
Shape and spread: 2 points
(c) Normal probability plots: 2 points
Description of pattern: 2 points
(d) Effect of log-transformation: 2 points

Question 4

(a) Difference of averages: 2 points
Antilog of difference: 2 points
Interpretation of antilog of difference: 4 points
(b) T-test: 6 points itemized as follows
Hypotheses: 2 points
Test-statistic: 1 point
P-value: 1 point
Null distribution: 1 point
Conclusion: 1 point
(c) Confidence intervals: 4 points
Conclusion: 2 points
Consistency: 1 point

Question 5

Standard error calculation: 4 points
Test-statistic: 2 points
Consistency: 1 point

Question 6

(a) Mann-Whitney test: 5 points itemized as follows
    Hypotheses: 2 points
    P-value: 1 point
(d) Conclusion and consistency: 2 points

Question 7

Suggested factors: 2 points

TOTAL = 111

Solutions developed by Adeniyi Adewale
Grant: G227150109 – Acc1Time Science
Revised by Henryk Kolacz