Practical 3. Hypothesis testing: Analysis of Variance ANOVA – Answers

Question 1: How many cases and how many variables does the dataset has?

cases	45
variables	two

Question 1.1 : Specify the type of variable and the scale of measurement for each, indicating which one is the explanatory or independent variable and which is the response or dependent variable.

Distance	Categorical, ordinal with three labels: 20m, 50m and 150m. These are also called treatments. Independent or explanatory variable
PAH concentration	Continuous, ratio scale. Response or dependent variable

Question 1.2: Which one is the factor and how many levels does it has?

The factor is distance to the oil platform with three levels: 20m, 50m and 150m.

Question 1.3: Describe the data observed

The concentration of PAH in water samples ranged in between 21.80 mg/L and 150.67 mg/L, with a higher average concentration near the oil platform (Table 1).

TABLE 1. Summary statistics for the concentration of polycyclic aromatic hydrocarbons (PAH) in water samples taken at 20m, 50m and 150m from an oil platform in Maracaibo lake, South America.

Distance	minimum	maximum	mean	Standard deviation	median	IQR
20m	105.11	150.67	122.09	12.41	119.59	15.92
50m	34.78	57.39	45.78	5.80	45.13	5.83
150m	21.79	30.65	25.87	2.45	25.63	2.71

Question 1.4: Write the logical hypothesis and its null equivalent (logical null hypothesis) in an IF-THEN statement.

Logical Hypothesis	**If** the oil platform is a source of pollution, **then** we would expect the concentration of PAH in water **to be different** among the distances, increasing near to the oil platform.
Logical Null Hypothesis	**If** the oil platform is a source of pollution, **then** we would expect the concentration of PAH in water **not to be different** among the distances.

Question 1.5: Write the null statistical hypothesis

Ho : **μ 20m = μ 50m = μ 150m**

Hi: μ 20m ≠ μ 50m ≠ μ 150m

Question 2.1: Is the data from both groups normally distributed? Write your results in narrative form. Remember to include the value of the test statistic and the p-value.

The Shapiro-Wilks test indicated the PAH concentration from water samples taken at 20m (W = 0.95164, p-value = 0.5507), 50m (W = 0.96937, p-value = 0.8486) and 150m from the oil platform (W = 0.97669, p-value = 0.9418) come from a population with normal distribution (Fig. 1)

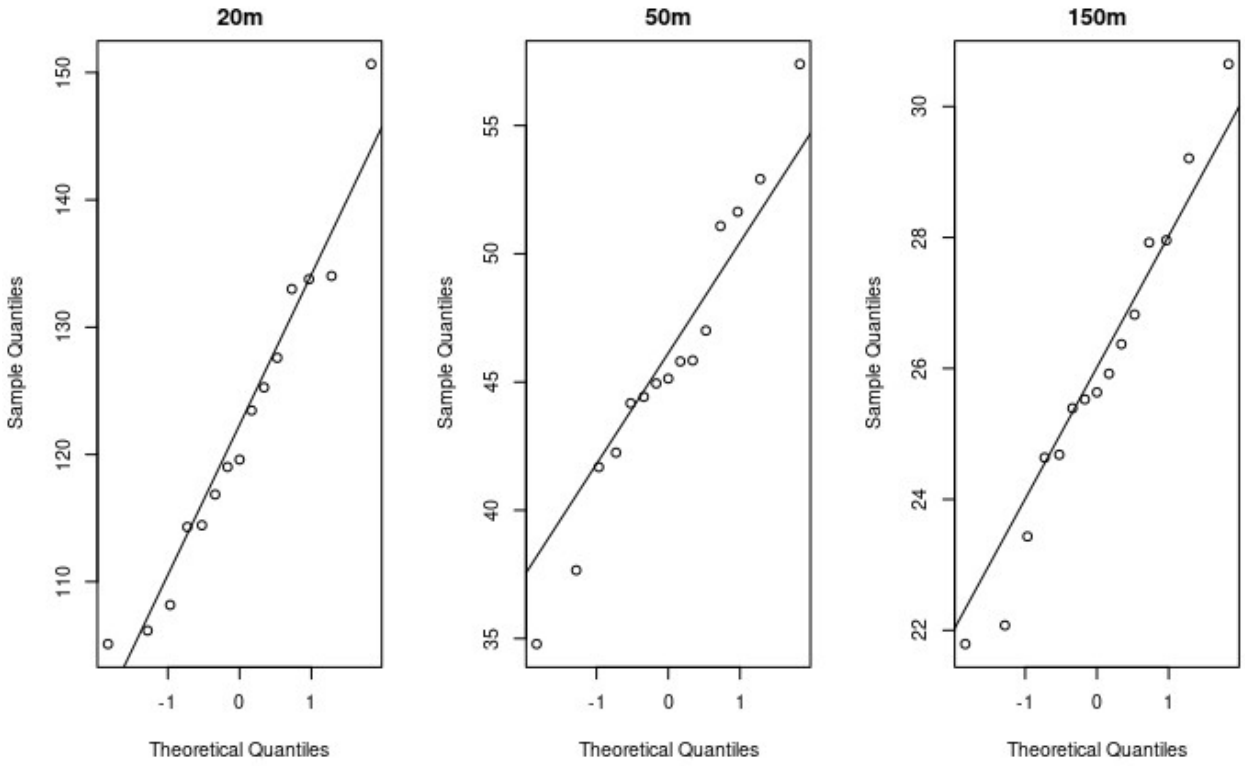

Figure 1. Quantile-Quantile (Q-Q) plot for the concentration of polycyclic aromatic hydrocarbons (PAH) in water samples taken at 20m, 50m and 150 m from an oil platform.

Question 2.2: Are the variances homogeneous? Write your results in narrative form. Remember to include the value of the test statistic, degrees of freedom and the p-value.

The variances of PAH concentration from water samples taken at 20m, 50m and 150m from an oil platform were not homogeneous (Bartlett test, K^2=29.26; df=2;p-value=$4.43*10^{-7}$). Therefore, data were transformed using logarithm base 10 and homogeneity of variances checked again (K^2=1.46; df=2;p-value=0.4803).

Question 3.1: Do you failed to accept or to reject your Null Hypothesis? Write the results of your one-way ANOVA in a narrative form and remember to include the value of the test statistic, degrees of freedom and p-value.

I failed to accept Ho.

The ANOVA showed that the concentration of PAH significantly differed among water samples taken at 20m, 50m and 150 m from the oil platform (Table 2).

TABLE 2. Analysis of Variance for the log transformed concentration of polycyclic hydrocarbon (PAH) in water samples taken at 20m, 50m and 150m from an oild platform in Maracaibo lake, South America.

Source of variation	df	Sums of Square	Mean Square	F	p-value
Distance	2	18.482	9.241	783.5	$0.002*10^{-12}$
Residuals	42	0.495	0.012		
Total	44	18.977			

Question 4.1: Write the results of your multiple comparison test in a narrative form and remember to include the value of the differences and p-value.

A posteriori pair-wise comparisons showed that the concentration of PAH in water from samples taken at 20m from the oil platform were significantly higher than those taken at 50m (diff=0.103; p-value <0.0001) and 150m (diff=35.48; p-value <0.0001). Similarly, the concentration of PAH in water from samples taken at 50m from the oil platform were also significantly higher than those taken at 150m (diff=3.69; p-value <0.0001).

Question 4.2: Why did we use the log-transformed PAH concentration if we are using non-parametric tests?

Because the non-parametric Kruskal-Wallis test and the post-hoc Dunn test assume homoscedasticity and the original variable had heterogeneous variances among the distances or groups.

Question 5.1: How many independent or explanatory variables do you have now? Identify the factors and their levels

Now I have 2 independent or explanatory variables	
Factor 1: Distance	3 Levels: 20m, 50m and 150m
Factor 2: Location	2 Levels: Water column and Sediment