**Hands-on-Training in Data Analysis by Kreativstorm. Homework-W2**

Q1. What is the mean father’s age?

28.90

Q2. What is the mean father’s age for low birthweight babies?

24.83

Q3. Is the father's age normally distributed? Justify your answer.

No, it is not normally distributed because the p-value (Shapiro-Wilk) is 0.039 which is less than the significance level of 0.05. (0.039< 0.05).

Q4. If you apply the log transformation to the father's age, what is the mean score of the transformed variable?

1.4493

Q5. Is the above mean score a good representation of the real value? Justify your answer.

Yes, because the original data of the variable was positively skewed, applying the log transformation helped to normalize the original data by reducing the skewness to a normal distribution which is more suitable for statistical analysis. In this case, however, the transformed mean would represent the central tendency of the data under the assumption of the transformation.

Q6. Is the new variable (log transform of father’s age) normally distributed? Justify your answer.

Yes, it is normally distributed because the p-value 0.129 in the normality test (Shapiro-Wilk) is greater than significance level of 0.05.(0.129>0.05)

Q7. Is the variable “years father was in education” normally distributed?

No, it is not normally distributed.

Q8. Mentioning the null and alternative hypotheses, explain the above answer.

Below is the hypothesis testing and conclusion on the above statement:

* H0: The variable “Years father was in education” follow a normal distribution.
* Ha: The variable “Years father was in education” does not follow a normal distribution

Since the p-value (0.0001) is less than the significance level of 5%, we reject the null hypothesis in favor of the alternative hypothesis. We therefore conclude that the variable “Years father was in education” does not follow a normal distribution.

Q9. What is the mean score for the variable “years father was in education” after you apply the Box-Cox transformation?

0.51

Q10. Is this new variable normally distributed? Explain.

No, it is not normally distributed.

Q11. What is the mean score for this new variable (B-C transformed fathers’ years in education) for mothers aged under 35?

0.486

Q12. Which test would you use to investigate the relationship between birth weight and father's age? Pearson product-moment correlation

* Pearson product-moment correlation
* Spearman’s Rank order correlation
* Point-Biserial correlation
* Phi-Coefficient

Q13. Justify the above choice in terms of the distribution of data and the nature of the test.

* Null Hypothesis (H0): There is no significant relationship between birth weight and father's age.
* Alternative Hypothesis (Ha): There is a significant relationship between birth weight and father's age.

P-value = 0.26

R = 0.176

0.26 > 0.05

Conclusion: We fail to reject the null hypothesis. Based on the analysis, a p-value of 0.26 suggests that there is no strong statistical evidence to support the presence of a significant relationship between birth weight and father's age.

Furthermore, the choice of Pearson product-moment correlation for the test was choose based on the following checks below:

* There is a Normality test in the distribution.
* The variables have a normal distribution.
* Both variables have continuous data
* Does not have monotonic relationship.

Q14. What is the direction of that relationship? Zero relationship

Q15. What is the form of that relationship? Non-linear relationship

Q16. What is the degree of that relationship? Weak relationship (Negligible correlation)

Q17. What test would you use to investigate the relationship between smoking and birth weight? **Pearson product-moment correlation**

* Pearson product-moment correlation
* Spearman’s Rank order correlation
* Point-Biserial correlation
* Phi-Coefficient

Q18. Report on the above results including information about direction/form/degree of the relationship.

* Null Hypothesis (H0): Smoking during pregnancy has no significant effect on birth weight.
* Alternative Hypothesis (Ha): Smoking during pregnancy is associated with a significant change in birth weight.

p-value = 0.043

r = -0.314

0.043< 0.05

Interpretation**:** Since the p-value (0.043) is less than the chosen significance level of 0.05, we reject the null hypothesis in favor of the alternate hypothesis, which suggests that there is statistically significant evidence supporting the alternative hypothesis statement.

Q19. If you wanted to see the effect of the length of a baby on birthweight, what would your independent variable be? **Length of baby**

* Length of baby
* Birthweight

Q20. In statistics, when creating a scatterplot, it is a common practice to put the independent variable on the x-axis and the dependent variable on the y-axis. With this in mind, create a scatterplot for the above case and provide the regression line. For homework submitted using MS Word, insert a picture of the scatterplot.

A graph with dots and numbers

Description automatically generated

Q21. Is the relationship between the length of a baby and birthweight linear? Yes

* Yes
* No

Q22. Justify the above choice.

The direction of the relationship has a moderate positive correlation. Moreso, the increasement is consistent along the same axis hence it has a linear relationship.

Q23. Is there any evidence to suggest that the birth weight, length of baby, and head circumference are related? Yes

* Yes
* No

Q24. Justify the above choice.

* Null Hypothesis (H0): There is no significant relationship between birth weight, length, and head circumference.
* Alternative Hypothesis (Ha): There is a significant relationship between at least one pair of variables.

From the correlation analysis,

p = 0.0001

0.0001< 0.05.

We reject the null hypothesis in favour of the alternate hypothesis. This suggest that there is a significant relationship between at least one pair of variables.

Q25. Describe the above relationship in your own words and provide evidence for your claims.

The relationship between the variables has a moderate positive correlation (0.50 – 0.70)

See extract of the correlation analysis below:

A screenshot of a data sheet

Description automatically generated