Q1. What is the mean father's age?

The mean father's age is 28.90 years

Descriptives

			Statistic	Std. Error
Father's age	Mean		28.90	1.059
	95% Confidence Interval for Mean	Lower Bound	26.77	
		Upper Bound	31.04	
	5% Trimmed Mean		28.62	
	Median		29.50	
	Variance		47.113	
	Std. Deviation		6.864	
	Minimum		19	
	Maximum		46	
	Range		27	
	Interquartile Range		10	
	Skewness		.508	.365
	Kurtosis		-.562	.717

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

The mean father's age for low birthweight babies is 24.83 years

Low birthweight	Mean		24.83	2.600
	95% Confidence Interval for Mean	Lower Bound	18.15	
		Upper Bound	31.52	
	5% Trimmed Mean		24.43	
	Median		23.00	
	Variance		40.567	
	Std. Deviation		6.369	
	Minimum		20	
	Maximum		37	
	Range		17	
	Interquartile Range		9	
	Skewness		1.816	.845
	Kurtosis		3.545	1.741

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

The father's age is not normally distributed. According to the significance value in Shapiro Wilk (.039), rounded to 0.04. The data is not normally distributed because the significance in Shapiro Wilk is lower than 0.05

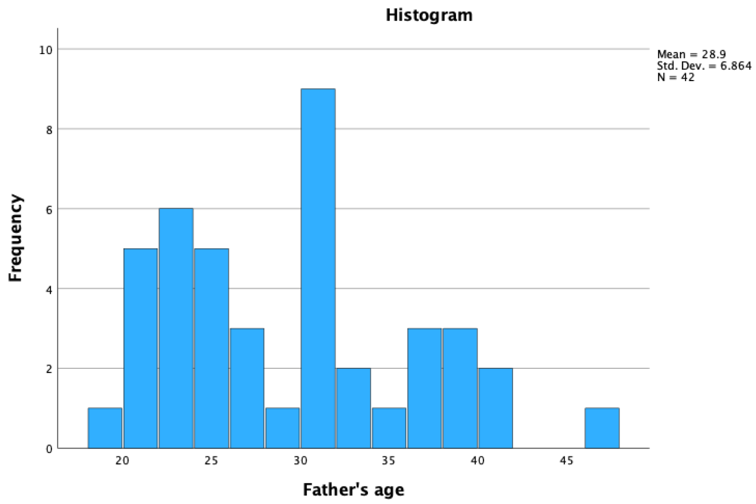
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Father's age	.120	42	.138	.944	42	.039

a. Lilliefors Significance Correction

Skewness	.508	.365
Kurtosis	-.562	.717

I also see that I have a bit of Skewness and Kurtosis, and the majority of the data is moved to the left-hand side, hence it’s positively skewed.



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

The mean score of the log transformed father's age is 1.44

fage_LOG	Mean		1.4493	.01571
	95% Confidence Interval for Mean	Lower Bound	1.4175	
		Upper Bound	1.4810	
	5% Trimmed Mean		1.4477	
	Median		1.4698	
	Variance		.010	
	Std. Deviation		.10178	
	Minimum		1.28	
	Maximum		1.66	
	Range		.38	
	Interquartile Range		.15	
	Skewness		.139	.365
	Kurtosis		-.973	.717

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

The real value (mean of father's age) is 28.90 years, and the value extracted from the log transformed father's age variable is 28.18 (10^{1.45}). Hence I can confirm that the above mean score is a good representation of the real value.

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

The new variable (log transform of father's age) is normally distributed because we obtain 0.13 significance value in Shapiro-Wilk. The significance value is larger than 0.05
The log transformation is the variable that works the best of the 3 variables (fage_SQRT, fage_INV, and fage_LOG), given the fact it has the highest significance value.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
fage_SQRT	.115	42	.184	.954	42	.087
fage_LOG	.108	42	.200*	.958	42	.129
fage_INV	.147	42	.024	.953	42	.084

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

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

The significance value in Shapiro Wilk is very low <.001, which means that the data is not normally distributed because the significance in Shapiro Wilk is lower than 0.05

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Years father was in education	.217	42	<.001	.844	42	<.001

a. Lilliefors Significance Correction

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

We reject the null hypothesis, meaning that we found enough evidence to confirm that the variable is not normally distributed, given the fact that the significance value is lower than 0.05.

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

The mean score for the variable "years father was in education" after the Box-Cox transformation is 13.71

Descriptives

			Statistic	Std. Error
fedyrs_BC	Mean		13.7141	.28104
	95% Confidence Interval for Mean	Lower Bound	13.1465	
		Upper Bound	14.2817	
	5% Trimmed Mean		13.7688	
	Median		13.7990	
	Variance		3.317	
	Std. Deviation		1.82137	
	Minimum		10.68	
	Maximum		15.76	
	Range		5.08	
	Interquartile Range		3.39	
	Skewness		-.267	.365
	Kurtosis		-1.176	.717

Q10. Is this new variable normally distributed? Explain.

The new variable is not normally distributed, the significance value in Shapiro Wilk is <.001, value lower than 0.05
The variable hasn't been normalized with the Box-Cox transformation.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
fedyrs_BC1	.226	42	<.001	.849	42	<.001

a. Lilliefors Significance Correction

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

The mean score for the B-C transformed fathers' years in education, for mothers aged under 35, is 13.55

fedyrs_BC1	Aged < 35	Mean		13.5503	.29509
		95% Confidence Interval for Mean	Lower Bound	12.9524	
			Upper Bound	14.1483	
		5% Trimmed Mean		13.5869	
		Median		13.7990	
		Variance		3.309	
		Std. Deviation		1.81908	
		Minimum		10.68	
		Maximum		15.76	
		Range		5.08	
		Interquartile Range		3.39	
		Skewness		-.134	.383
		Kurtosis		-1.195	.750

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

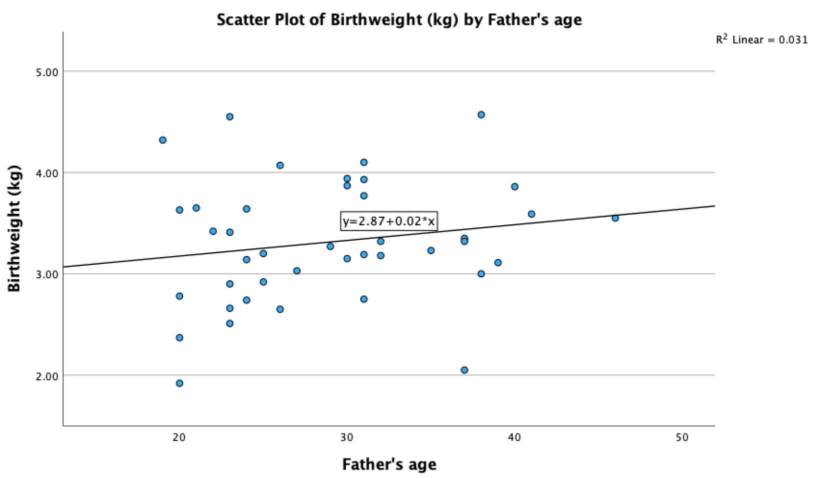
- **Pearson product-moment correlation**
- **Spearman's Rank order correlation**
- **Point-Biserial correlation**
- **Phi-Coefficient**

I use Spearman's Rank order correlation. We can use this correlation with ordinal or continuous data. Birth weight and father's age are continuous variables.

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

I choose Spearman because we can use this correlation with continuous data and because I cannot find 2 of the 3 essential criteria for Pearson correlation.

I don't see a linear relationship between birth weight and father's age



Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Birthweight (kg)	.059	42	.200*	.990	42	.968
Father's age	.120	42	.138	.944	42	.039

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

The data for the variable father's age is not normally distributed, the significance value in Shapiro Wilk is .039 (lower value than 0.05).

<u>Spearman's Rank order correlation</u>

I obtain an r value of 0.18 (negligible correlation) and a p value of 0.25 (higher value than 0.05), we don't reject the null hypothesis, and we confirm that there is no statistically significant correlation between birth weight and the father's age.

Correlations

			Birthweight (kg)	Father's age
Spearman's rho	Birthweight (kg)	Correlation Coefficient	1.000	.178
		Sig. (2-tailed)	.	.259
		N	42	42
	Father's age	Correlation Coefficient	.178	1.000
		Sig. (2-tailed)	.259	.
		N	42	42

<u>Pearson product-moment correlation</u>

I also analyze the values in the Pearson correlation, and I obtain similar values.
The r value is 0.18 (negligible correlation). The significance is, p = 0.27 (a higher value than 0.05). Hence we don't reject the null hypothesis, and we confirm that there is no statistically significant correlation between birth weight and the father's age.

Correlations

		Birthweight (kg)	Father's age
Birthweight (kg)	Pearson Correlation	1	.176
	Sig. (2-tailed)		.266
	N	42	42
Father's age	Pearson Correlation	.176	1
	Sig. (2-tailed)	.266	
	N	42	42

Q14. What is the direction of that relationship?

Positive direction

Q15. What is the form of that relationship?

Linear relationship

Q16. What is the degree of that relationship?

Weak positive correlation and negligible correlation

Q17. What test would you use to investigate the relationship between smoking and birth weight?

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

Point-Biserial correlation, because smoking is a dichotomous variable and birth weight is a continuous variable. I find statistically sig. negative correlation between smoking and birth weight, r = -0.314 and p= 0.043
The p value is below 0.05; hence we reject the null hypothesis. We confirm that there is a negative correlation between birth weight and smoking. The birth weight is lower when smoking is 1 (smoking mother).

Correlations

		Birthweight (kg)	smoker
Birthweight (kg)	Pearson Correlation	1	-.314*
	Sig. (2-tailed)		.043
	N	42	42
smoker	Pearson Correlation	-.314*	1
	Sig. (2-tailed)	.043	
	N	42	42

*. Correlation is significant at the 0.05 level (2-tailed).

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

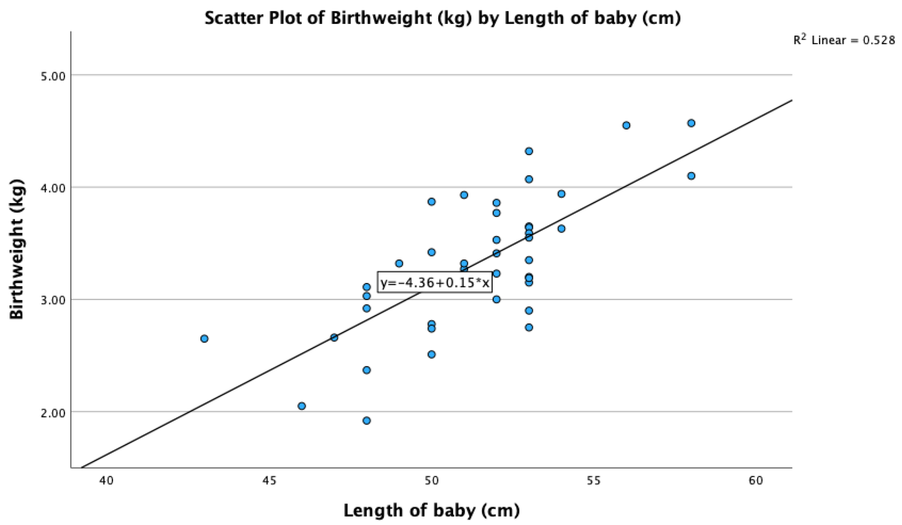
Direction: negative correlation
Form: linear
Degree of relationship: low negative correlation (weak correlation)

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**
- **Birthweight**

Length of the baby would be the independent variable because we are aiming to see the changes in the birth weight depending on the values in the variable length of the baby.

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.



I obtain a linear positive correlation, and the values are normally distributed because the significance in Shapiro Wilk (for both variables) is larger than 0.05

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Birthweight (kg)	.059	42	.200*	.990	42	.968
Length of baby (cm)	.166	42	.005	.948	42	.054

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

I find statistically sig. positive correlation between birth weight and length of baby, r = 0.72 and p= <.001. It's a high positive correlation. Hence, I reject the null hypothesis.

Correlations

		Birthweight (kg)	Length of baby (cm)
Birthweight (kg)	Pearson Correlation	1	.727**
	Sig. (2-tailed)		<.001
	N	42	42
Length of baby (cm)	Pearson Correlation	.727**	1
	Sig. (2-tailed)	<.001	
	N	42	42

**. Correlation is significant at the 0.01 level (2-tailed).

I use Pearson Correlation because we have 2 continuous variables
(according to Pearson Correlation, variables should be measured in interval or ratio scale).
The relationship is linear, and both variables are normally distributed.

Q21. Is the relationship between the length of baby and birthweight linear?
- **Yes**
- **No**

Yes, it's a linear positive correlation.

Q22. Justify the above choice.

I obtain an r = 0.72 and p= <.001, which means that is a high positive correlation. Hence I find consistent enough evidence to claim that we find correlation, because the significance is lower than 0.05 (p= <.001). This means that I can reject the null hypothesis and confirm that there is correlation between the length of the baby and the birth weight.

Q23. Is there any evidence to suggest that the birth weight, length of baby, and head circumference are related?
- **Yes**
- **No**

Yes, the 3 variables measure different aspects of the baby and the 3 variables are continuous.

Q24. Justify the above choice.

In order to confirm that there is a correlation between the 3 variables, I need to check the 3 parameters for Pearson, before using the Pearson Correlation.

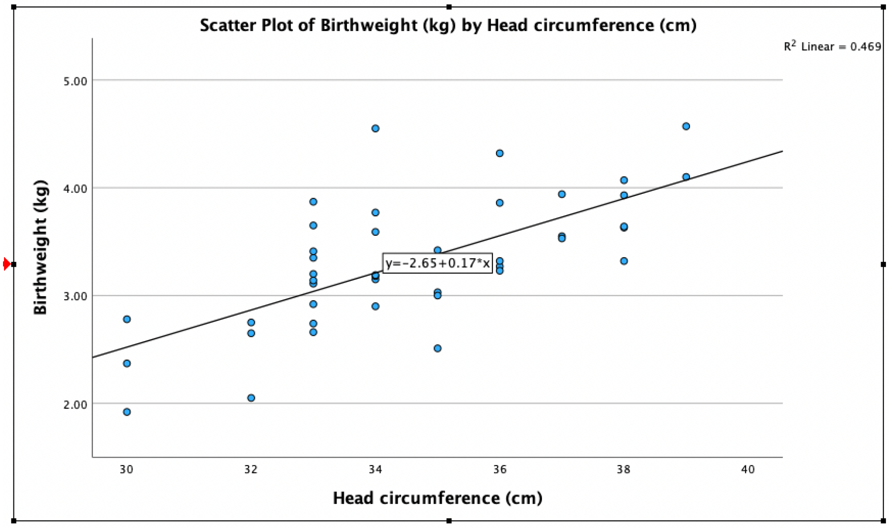
1. The variables are continuous
2. The values are normally distributed, because the significance in Shapiro Wilk (for the 3 variables) is larger than 0.05
   significance value for birth weight .96
   significance value for length of baby .054
   significance value for head circumference .067

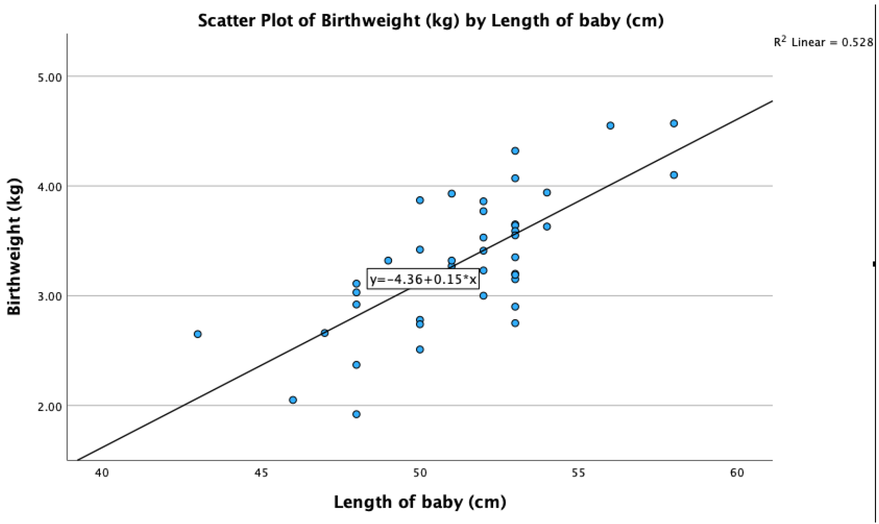
Tests of Normality

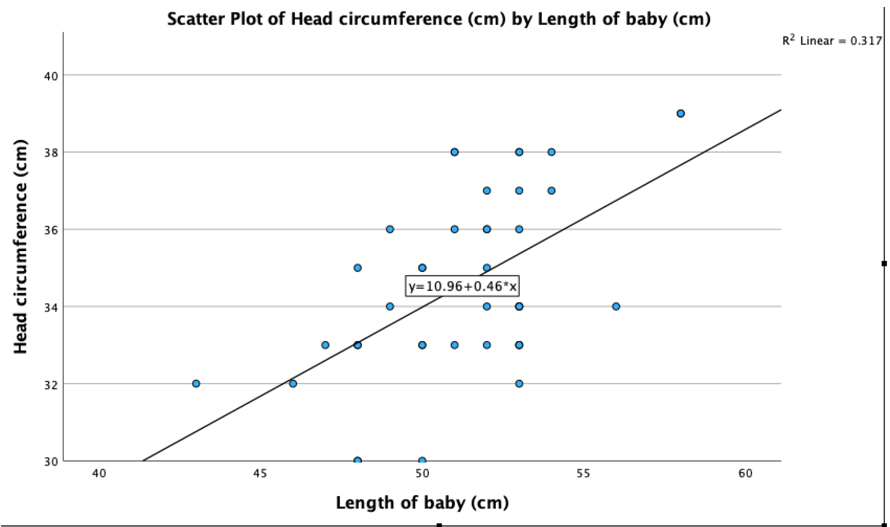
	Kolmogorov–Smirnov[a]			Shapiro–Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Birthweight (kg)	.059	42	.200*	.990	42	.968
Length of baby (cm)	.166	42	.005	.948	42	.054
Head circumference (cm)	.146	42	.026	.950	42	.067

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

3. The relationship between the variables is linear







4. I calculate the Pearson Correlation for the 3 variables

Correlations

		Birthweight (kg)	Length of baby (cm)	Head circumference (cm)
Birthweight (kg)	Pearson Correlation	1	.727**	.685**
	Sig. (2-tailed)		<.001	<.001
	N	42	42	42
Length of baby (cm)	Pearson Correlation	.727**	1	.563**
	Sig. (2-tailed)	<.001		<.001
	N	42	42	42
Head circumference (cm)	Pearson Correlation	.685**	.563**	1
	Sig. (2-tailed)	<.001	<.001	
	N	42	42	42

**. Correlation is significant at the 0.01 level (2-tailed).

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

I obtain a very low significance level for the 3 correlations (p = <.001), I can reject the null hypothesis and confirm that there is correlation between the 3 variables (Birth weight - Length of baby, Birth weight - Head circumference, and Length of baby - Head circumference). The 3 correlations are positive and linear.

<u>High positive correlation</u>
Birth weight - Length of baby, r = .72, p= <.001

<u>Moderate positive correlation</u>
Birth weight - Head circumference, r = .68, p = <.001

<u>Moderate positive correlation</u>
Length of baby - Head circumference, r = .56, p= <.001