

1. Four blood measurements were used to categorise the children as either Stage 1, 2 or 3 iron deficient, or as having normal iron levels. Obtain summary statistics that tell us how many children are in each group.
2. Exclude the cases which are missing iron classifications from the analysis. **ncrp10** measures whether a child has elevated C-reactive protein levels, which indicates an infection. In such cases, iron status categorisation is unreliable. Obtain summary statistics to tell us how many children have elevated C-reactive protein levels, then remove the cases from the analysis.
3. Calculate 95% confidence intervals for the proportion of children with each iron status. What do you conclude from your confidence intervals?
4. Investigate whether the prevalence of iron deficiency varies with age, by generating counts of children with each iron deficiency status for both toddlers and infants. Display these results visually with a bar chart or histogram. What differences do you notice between the counts of children with each iron status, based on this graph?
5. Create a box-and-whiskers plot for each iron status using the continuous age variable. What differences do you notice between the counts of children with each iron status, based on this graph?
6. Calculate 95% confidence intervals for the difference in proportions between toddlers and infants, at each level of iron deficiency (depleted, Stage 2 and Stage 3). What do you conclude from your confidence intervals?
7. Investigate the effect of some of the other factors in the spreadsheet on iron deficiency status. For example, you could look at whether the sex of the child has an effect on their iron status, or whether prematurely born babies were more likely to be iron deficient.
8. The variables **hb**, **mcv**, **zpp** and **ferritin** are measurements taken from the blood samples of the children in the study, measuring haemoglobin levels, cell

volume, zinc levels and iron levels. Check the distributions of these readings for normality. Do the data appear to be normally distributed? If the data are not normally distributed, suggest a way of transforming the data that may remedy this problem, and investigate whether this transformation offers an improvement.

9. The variable **ferritin** measures iron stores in the blood. Investigate the effect of breast feeding, formula feeding, and high cow's milk intake (**bf**, **curff** and **milk500**) on this variable with boxplots, summary statistics and confidence intervals for differences in means. Does there appear to be a difference in iron stores for the breast fed children compared to the other children, for the formula fed children compared to the other children, and for the children who drink a lot of cow's milk when compared to other children?
9. Bootstrap the difference in mean iron stores for breast fed children and other children, and provide a bootstrap estimate of the mean difference along with a 95% confidence interval. What do you conclude from your bootstrap confidence interval for the difference in the means?
10. Examine the relationships between some of the continuous variables and the iron stores. Start with **nfeall**, the dietary iron level variable. From earlier, you should have found that the distribution of the ferritin data is skewed. Plot a histogram of the dietary iron data to determine whether a transformation is necessary for this data too. If it is, take the log of the data, and make sure to do a similar calculation for the **ferritin** variable. Then produce a scatter plot of the two variables, and discuss whether there appears to be a relationship between the two variables. Calculate the correlation between the two variables.