

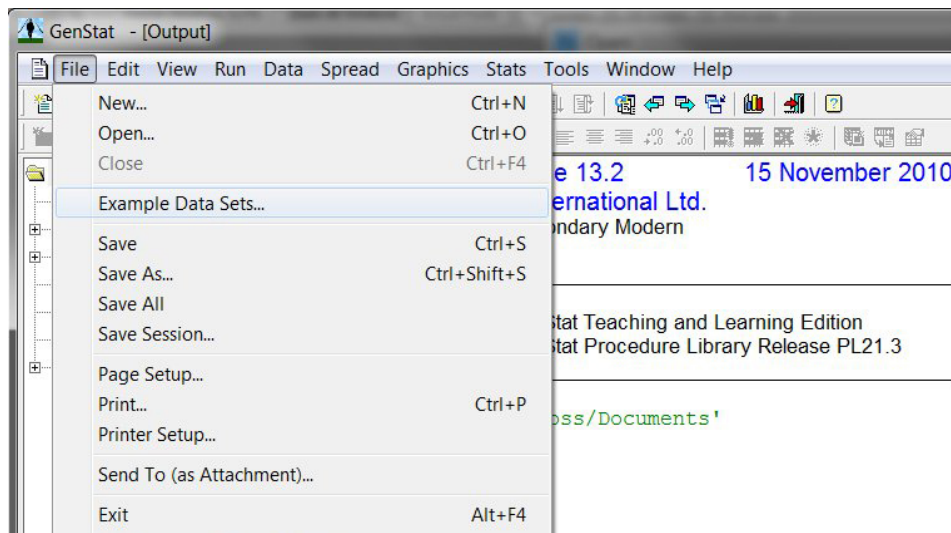


Iron Deficiency in Infants and Toddlers

In the late 1990s, researchers from the University of Otago carried out a study to assess the prevalence of iron deficiency in 6-24 month old children in the South Island of New Zealand. Iron deficiency is associated with detrimental effects on the health, growth and development of children, so they were also interested in exploring which factors have an association with iron deficiency status. To investigate these issues, the food intake of 323 children from Christchurch, Dunedin and Invercargill were monitored, and blood samples and other measurements were taken from the children.

This lesson investigates the prevalence of iron deficiency in the sample data, and the influence of some other factors like sex and age on this prevalence.

1. To open the data we click on **File>Example Data Sets**:



This brings up the Example Datasets dialog shown below. Click on the **Filter by topic** drop-down menu and select the **NZ Schools Example Data sets** option. Choose the file **Iron.gsh** and click on **Open data**.

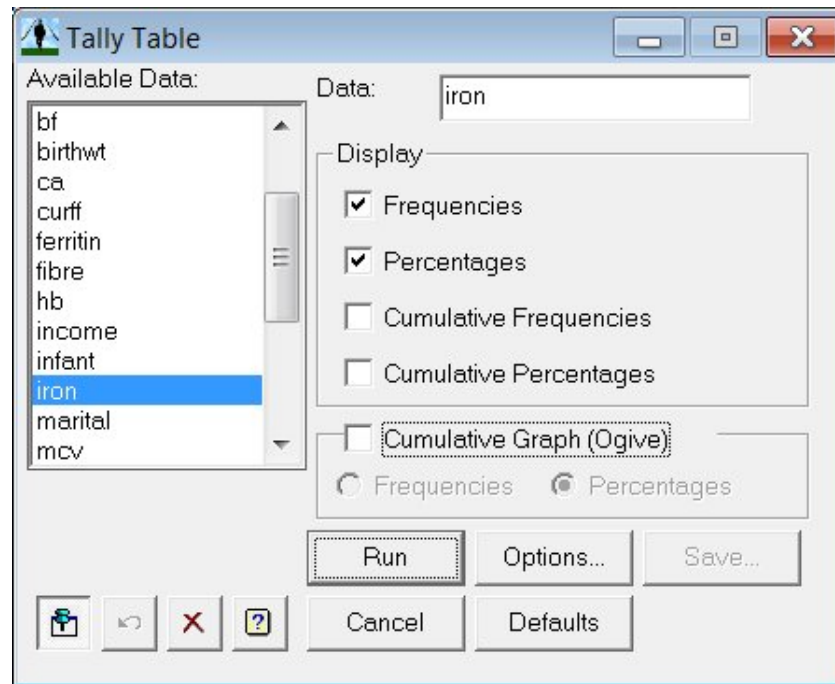
Row	ID	hb	mcv	zpp	ferritin	iron	ncrp10	age	infant	birthwt	bf
1	258	124	79	54	22.8	Normal	Normal	22.23	Toddler	2870	Otherwise
2	328	107	80	40	8	Depleted	Normal	24.43	Toddler	4500	Otherwise
3	349	110	75	*	*		Normal	24.93	Toddler	3020	Otherwise
4	362	115	81	50	14	Normal	Normal	21.9	Toddler	4410	Otherwise
5	390	110	73	48	6	Stage 2 Deficient	Normal	21.37	Toddler	4310	Breast feeding
6	413	*	*	*	*			22.4	Toddler	3620	Otherwise
7	444	99	76	33	16.6	Normal	Normal	20.07	Toddler	3665	Otherwise
8	455	101	81	45	16.8	Normal	Normal	14.53	Toddler	2970	Breast feeding
9	462	111	79	30	8.2	Depleted	Normal	18.23	Toddler	3321	Otherwise
10	496	112	82	33	7.7	Depleted	Normal	24.77	Toddler	3490	Otherwise
11	821	*	*	*	*			13.93	Toddler	3270	Otherwise
12	819	112	78	28	22.4	Normal	Normal	16	Toddler	3020	Otherwise
13	104	112	79	45	*		Normal	12.63	Toddler	3140	Breast feeding
14	214	*	*	*	*			13.43	Toddler	3300	Breast feeding
15	261	122	79	39	17.3	Normal	Normal	19.9	Toddler	2055.15	Otherwise
16	238	*	*	*	*			12.9	Toddler	3750	Breast feeding
17	800	*	*	*	*			13.7	Toddler	4812	Otherwise

This opens a large spreadsheet containing blood measurement data from the children study, as well as additional information about the children and their parents. By switching to the Output Window, we can see a list of these variables, as well as a description of what they measure:

Data on Iron levels for 323 young children recorded in 25 variables:

- ID - subject ID number
- hb - haemoglobin (g/L)
- mcv - mean cell volume (fL)
- zpp - zinc protoporphyrin (umol/mol hb)
- ferritin - ferritin (ug/l)
- iron - iron deficiency status
- ncrp10 - to define children with infection (elevated C-reactive protein)
- age - age of child
- infant - infant = 5-11.9 months of age; toddler=12-24 months of age
- birthwt - infant birth weight
- bf - to define children who were currently breastfeeding
- premi - to define children who were born prematurely
- curff - to define children who were currently formula feeding
- sex - sex
- caucasian - ethnicity
- tertiary - maternal education
- income - household income level
- smokers - smoker in the household
- marital - marital status
- nkjall - the estimated total average energy intake per day (breast milk & food)
- nfeall - total average iron intake per day from food and breast milk
- fbre - total average fibre intake per day from food & breast milk
- ca - total average calcium intake per day from food & breast milk
- vtc - total average vitamin C intake per day from food & breast milk
- milk500 - to define children with a high milk intake (> 0.5 litre)

2. Four blood measurements were used to categorise the children as either Stage 1, 2 or 3 iron deficient, or as having normal iron levels. To obtain summary statistics which tell us how many children are in each group, click **Stats>Tally**, and choose the **iron** variable. Untick the **Cumulative Frequencies**, **Cumulative Percentages** and **Cumulative Graph** options, and click **Run**:

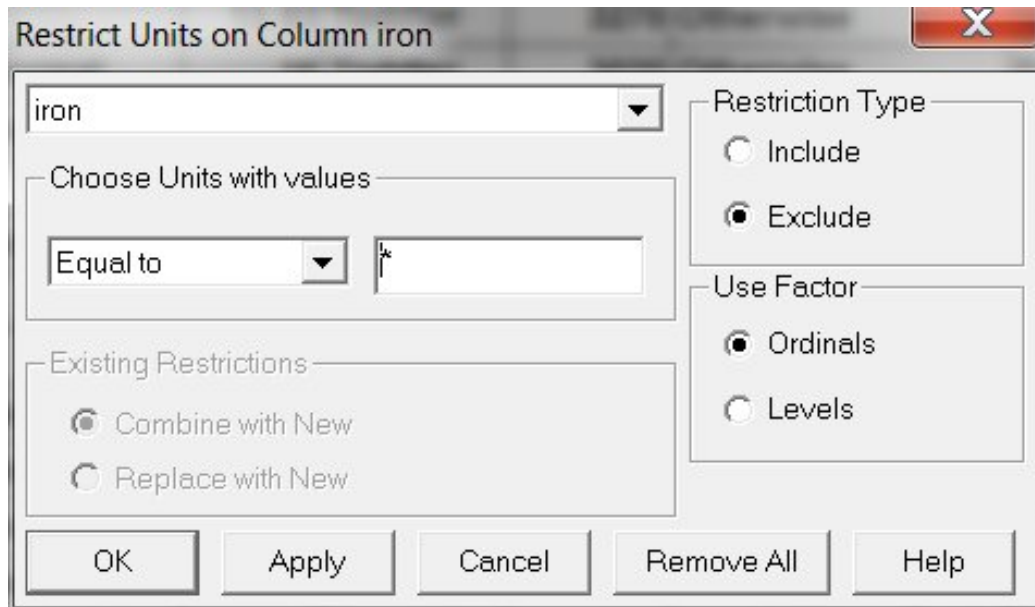


The resulting output is displayed in the Output Window:

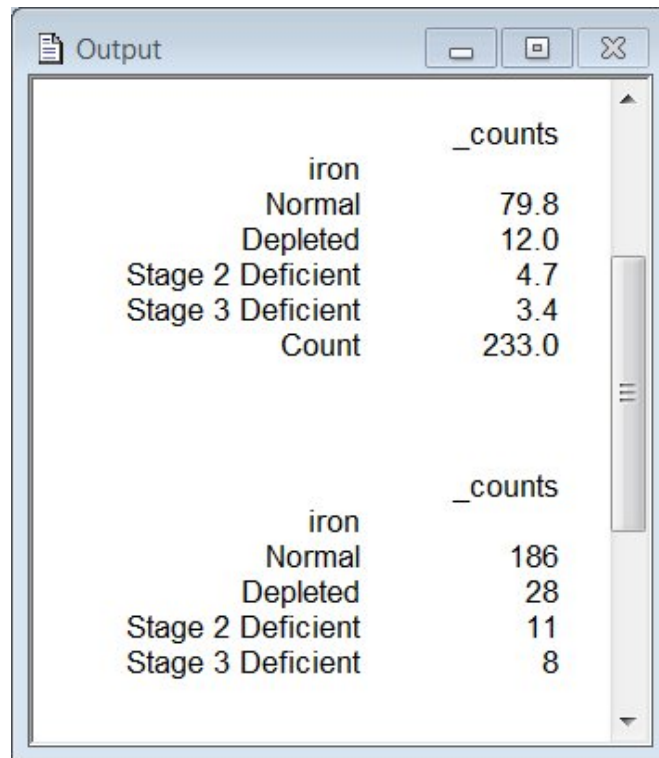
Value	Frequency	Percentage
Depleted	28	11.2
Normal	202	81.1
Stage 2 Deficient	11	4.4
Stage 3 Deficient	8	3.2
74 (22.9%) missing values		

We can see that 74 of the 323 children have not been categorised into one of these groups, because they had missing blood sample measurements. We therefore want to exclude these cases from the analysis.

To do this, click **Spread>Restrict/Filter>By Value**, and choose the **iron** variable. Click the **Exclude** option, and from the drop-down box choose **Equal to** option. Finally, add ***** to the empty box, and click **OK**.



3. The factor **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 using **Spread>Restrict/Filter>To Groups** with the **Combine with New** option.
4. By hand, calculate 95% confidence intervals for the proportion of children with each iron status. The information necessary to calculate these intervals can be obtained by clicking on **Stats>Frequency Tables**, entering **Iron** into the **Groups** box, tick the **Display table as percentage of** box, and clicking **Run**. The desired information is now in the Output Window:



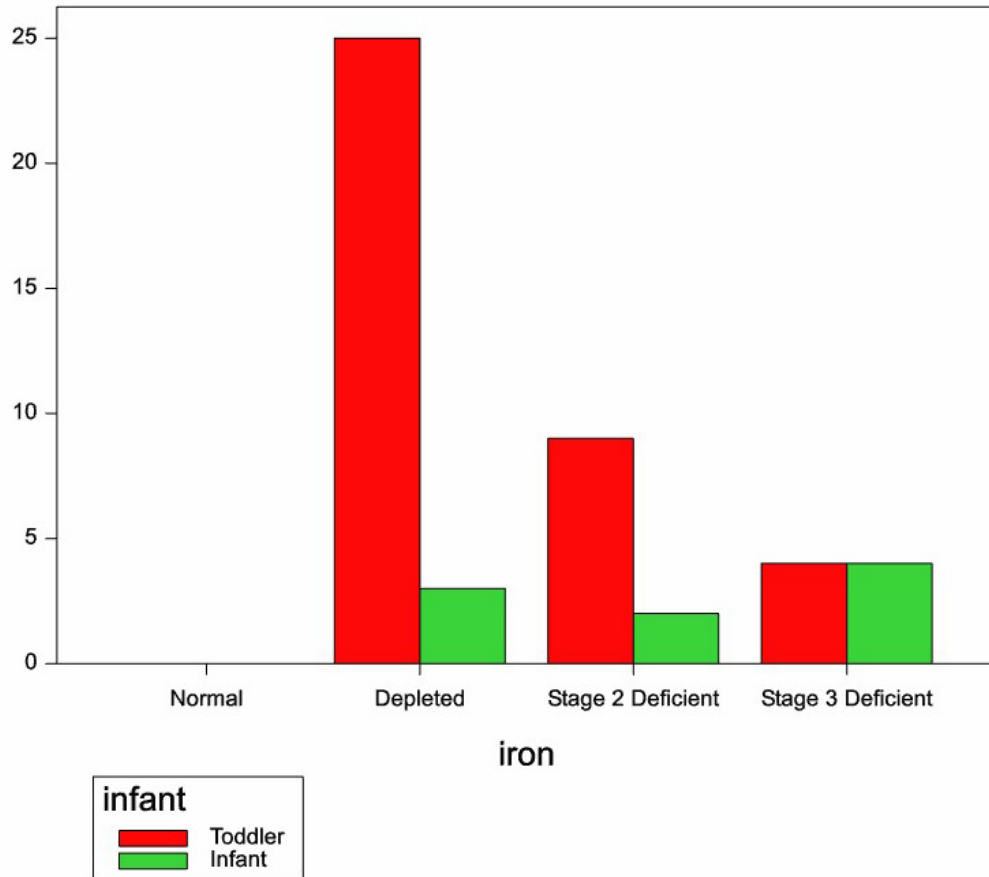
What do you conclude from your confidence intervals?

- To investigate whether the prevalence of iron deficiency varies with age, click **Stats>Frequency Tables**. Enter **Infant** and **Iron** into the **Groups** box, tick the **Display frequencies in spreadsheet** box, and click **Run**. The resulting spreadsheet shows the counts of children with each iron deficiency status for both age groups:

Row	<i>infant</i>	Normal	Depleted	Stage 2 Deficient	Stage 3 Deficient
1	Toddler	123	25	9	4
2	Infant	63	3	2	4

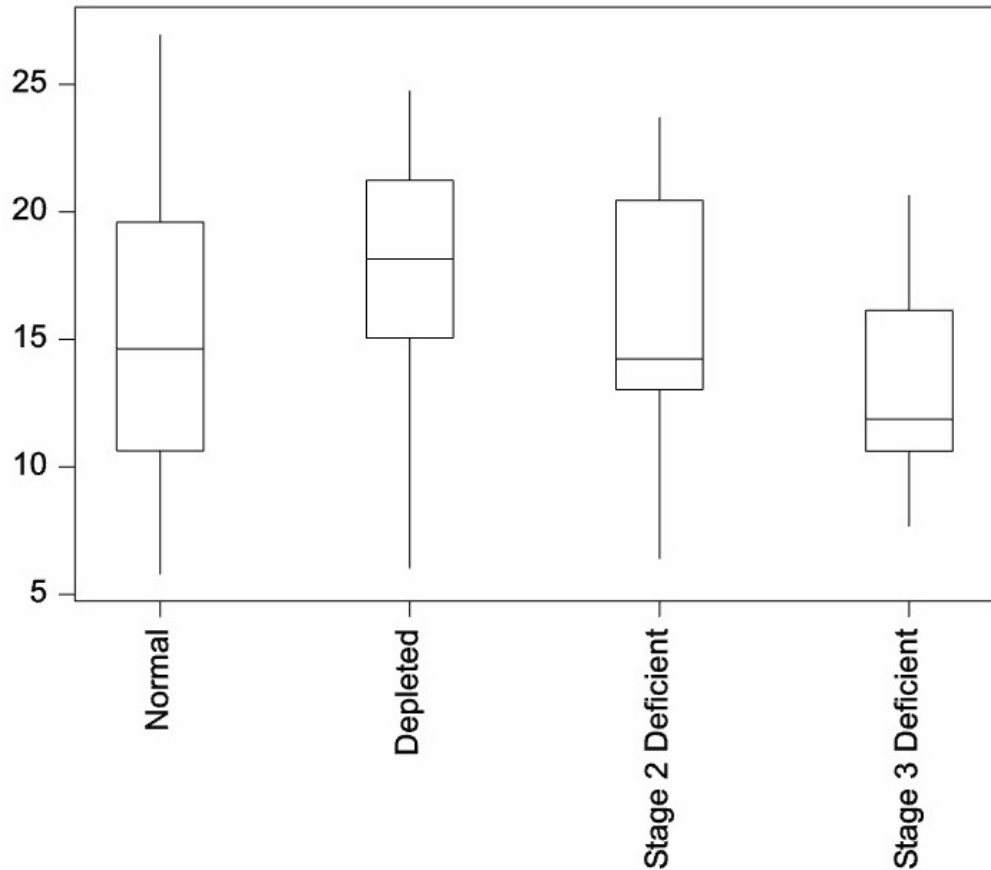
To display these results visually, we can create a bar chart. First, remove the children with normal iron status from the analysis using **Spread>Restrict/Filter>To Groups**. Then, click on **Graphics>Bar Charts**, and choose the **Summary tables** option. Click on **Create a summary table**, and choose the **Forming a summary table using a grouping factor** option. Click **OK**. Choose the

Two way table option, and enter **infant** and **iron** as grouping factors. Click **OK**, then **Run** to produce the following graph:



What differences do you notice between the counts of children with each iron status, based on this graph?

Alternatively, we can create a box-and-whiskers plot using the continuous age variable **age**. Reintroduce the children with normal iron status (remove all the filters and then reintroduce the ones that exclude infected children and those with missing iron status). Then click on **Graphics>Boxplot**, enter **age** as the **Data variate**, and **iron** as the **Grouping factor**. Click on **Run** to produce the following graph:



To display a summary table of iron status by age using percentages instead of counts, we can proceed as before using **Stats>Frequency Tables**, but also ticking the box **Display table as percentage of** and choosing **Iron** from the drop-down box:

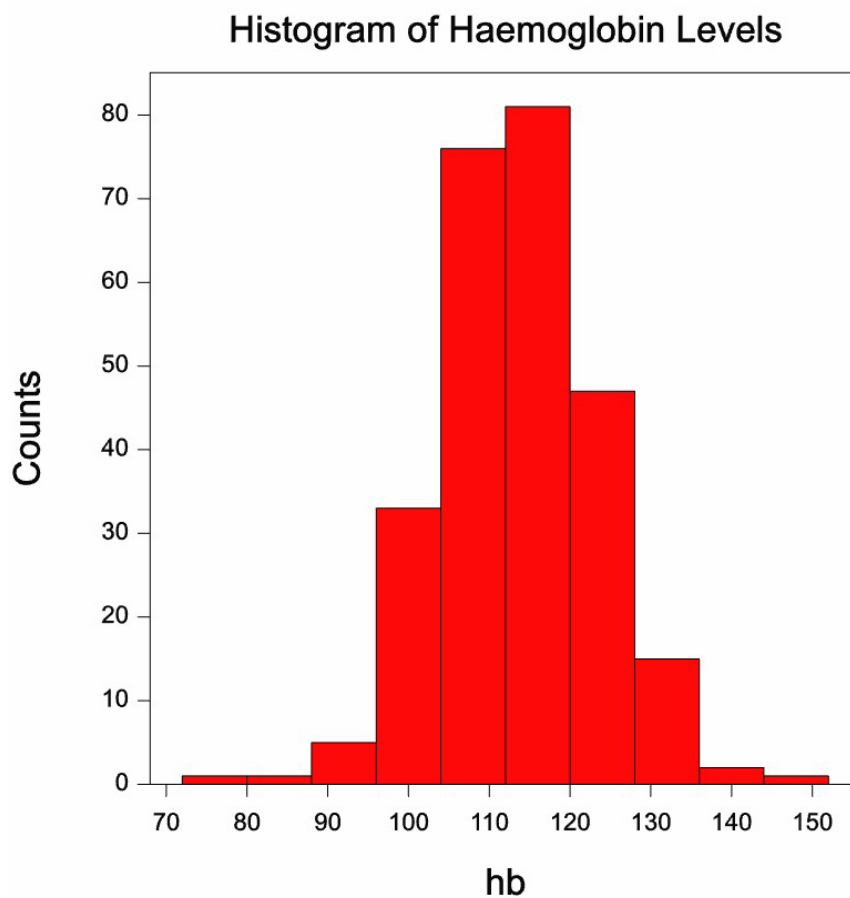
Row	infant	Normal	Depleted	Stage 2 Deficient	Stage 3 Deficient	Margin
1	Toddler	66.1	89.3	81.8	50.0	69.1
2	Infant	33.9	10.7	18.2	50.0	30.9
3	Margin	186.0	28.0	11.0	8.0	233.0

This tells us that, for example, 89.3% of the 28 children with depleted levels of iron are toddlers (ages 12 to 24 months), not infants (ages 6 to 11.9 months).

Using this summary table, calculate 95% confidence intervals (by hand) 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?

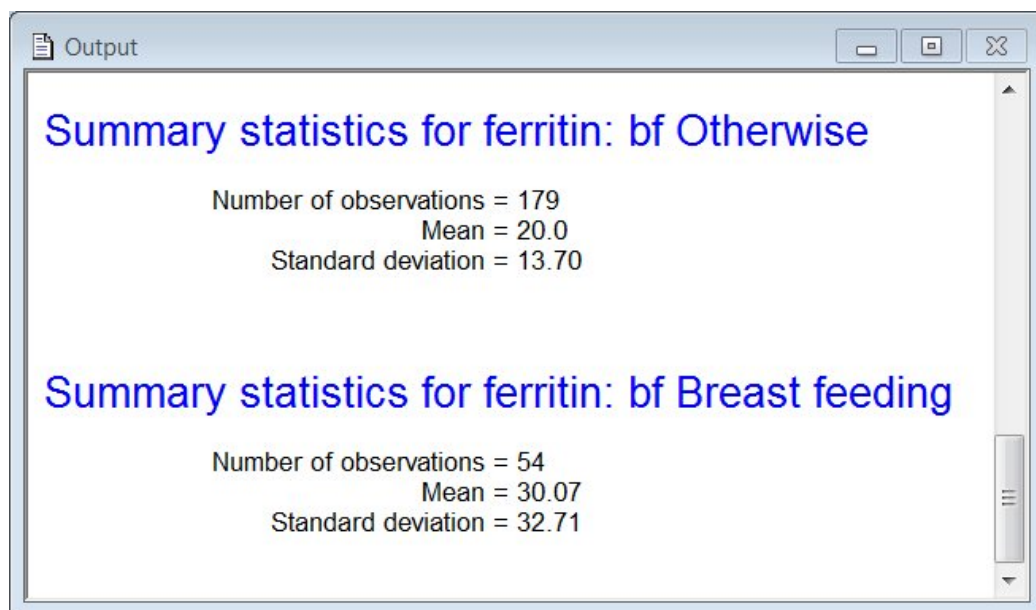
- Investigate the effect of some of the other factors in the spreadsheet on iron deficiency status, using techniques presented above. For example, you could look at whether the sex of the child has an effect on their iron intake, or whether prematurely born babies were more likely to be iron deficient.
- 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 by clicking **Graphics>Histogram** and selecting one of these variables. The resulting graph for the haemoglobin levels is as follows:

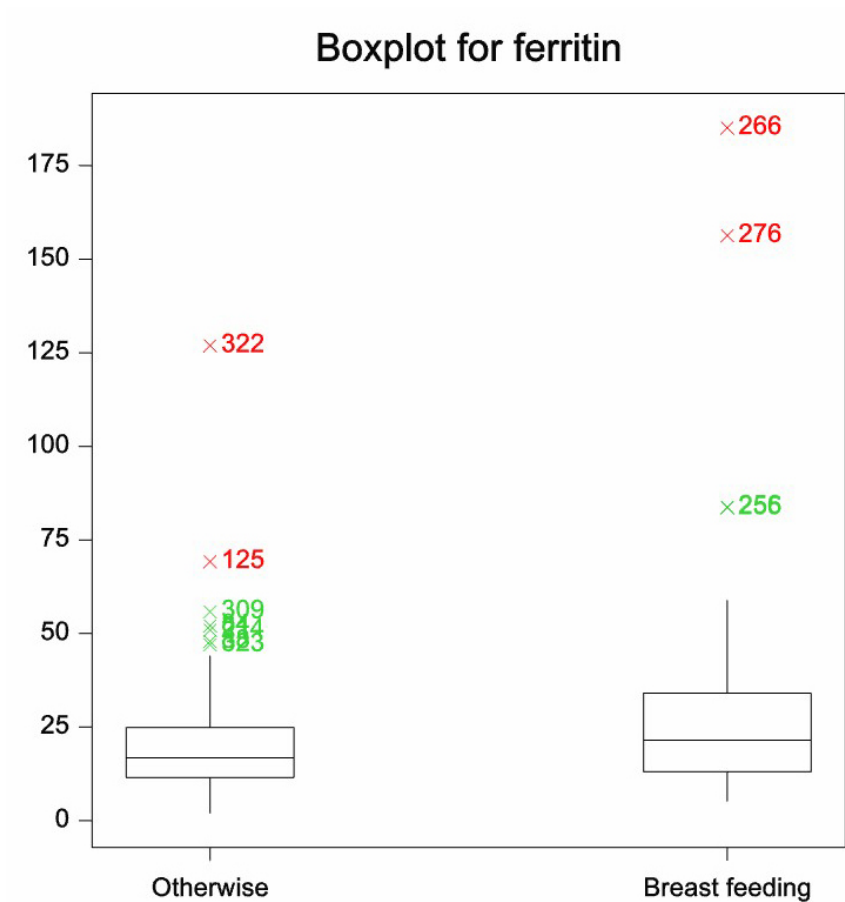


Do the data appear to be normally distributed?

Plot the other three histograms, and comment on the normality of the data. 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.

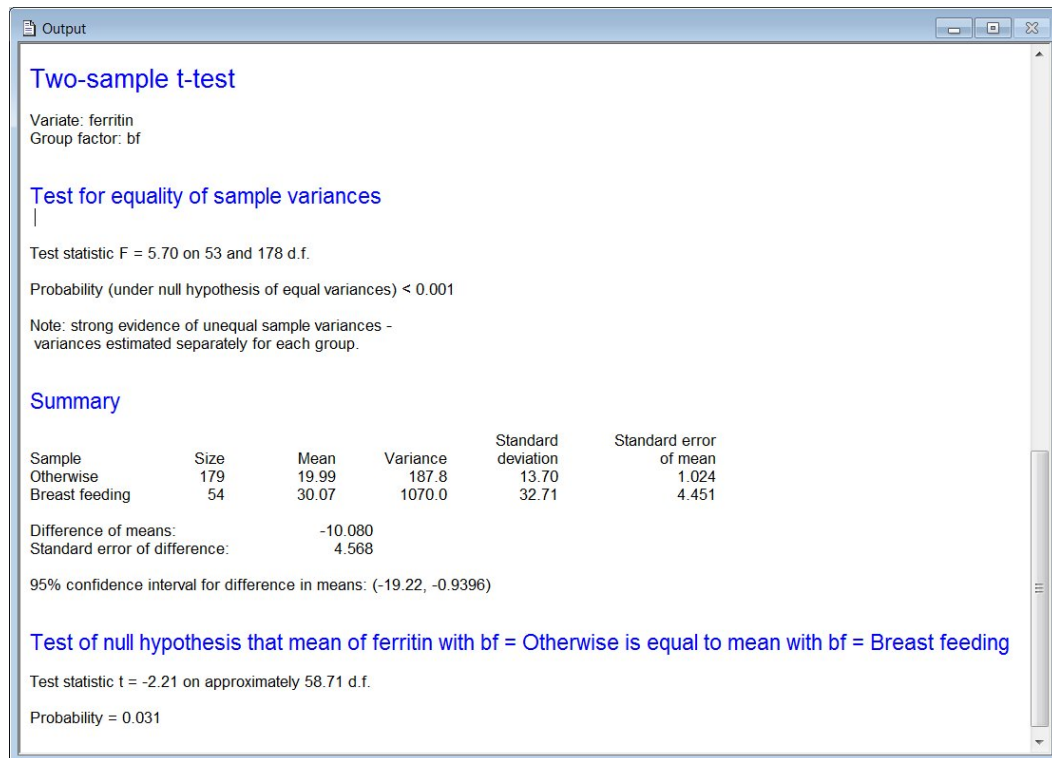
8. The variable **ferritin** measures iron stores in the blood. Investigate the effect of breast feeding, formula feeding, and high cow's milk intake (**bf**, **curff** and **milk500**) on this variable by clicking **Stats>Summary Statistics**, entering **ferritin** into the **Variates** box, and one of the categorical predictors into the **By Groups** box. Tick only **No. of non-missing values**, **Arithmetic Mean**, and **Standard Deviation**, and also tick the **Boxplot** option. Click **Run**. For the breast feeding predictor, the summary output in the Output Window and the boxplot are as follows:





Based on these results, does there appear to be a difference in iron stores for the breast fed children compared to the other children?

To examine whether there is a difference in means for the two groups, perform a t-test using **Stats>t-tests**. Change the type of test to a **two-sample** test, and the data arrangement to **Group factor with variate**. Enter **ferritin** as the data variate and **bf** as the Group factor. Click **Run** to obtain the following results:



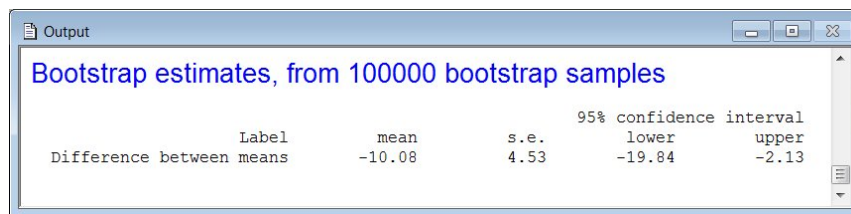
The t-test is equivalent to considering whether zero is in the confidence interval for the difference in means, and a 'large' test statistic suggests that this difference is not equal to zero. To assess whether this is 'large' enough, we look at the probability value, and if it is less than 0.05 (for a test equivalent to using a 95% confidence interval), then we conclude that there is indeed a difference in means. Is this the case here?

Repeat this process with the other two predictor variables and report your conclusions.

- Estimate the distribution of the mean difference in iron stores for breast fed children and other children, and using this distribution provide a new estimate of the mean difference along with a 95% confidence interval it.

To do this, we use the bootstrap technique. This involves randomly sampling a new dataset from the original sampled data. By doing this several times, we create a large number of datasets that we might have seen had we performed the experiment multiple times. By computing the mean for each of these datasets, we get an estimate of the distribution of the statistic.

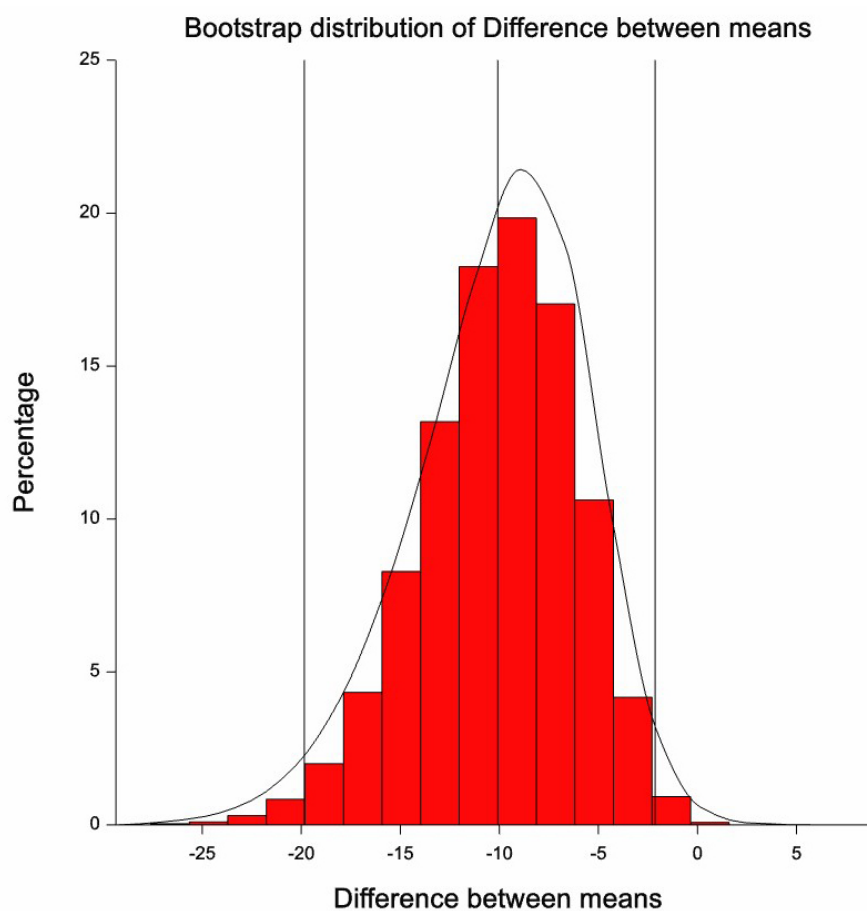
In GenStat we can perform the bootstrap using **Stats>Bootstrap**, and choosing the **Difference between means of two samples** option. Set the data arrangement to **Group factor with variate**, enter **ferritin** as the data variate and **bf** as the Group factor, set the number of bootstrap samples to a large number (e.g. 10,000 or 100,000) and click **Run**. The results obtained should look similar to the following:



Output

Bootstrap estimates, from 100000 bootstrap samples

Label	mean	s.e.	95% confidence interval	
			lower	upper
Difference between means	-10.08	4.53	-19.84	-2.13

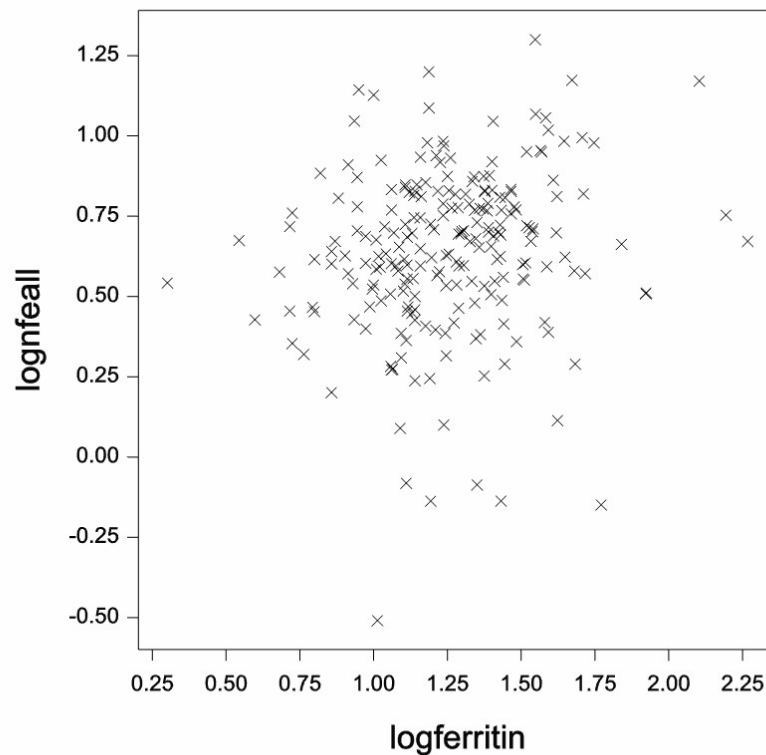


What do you conclude from your bootstrap confidence interval for the difference in the means?

Perform a bootstrap for the mean difference in iron stores for formula fed children and other children, and report your findings.

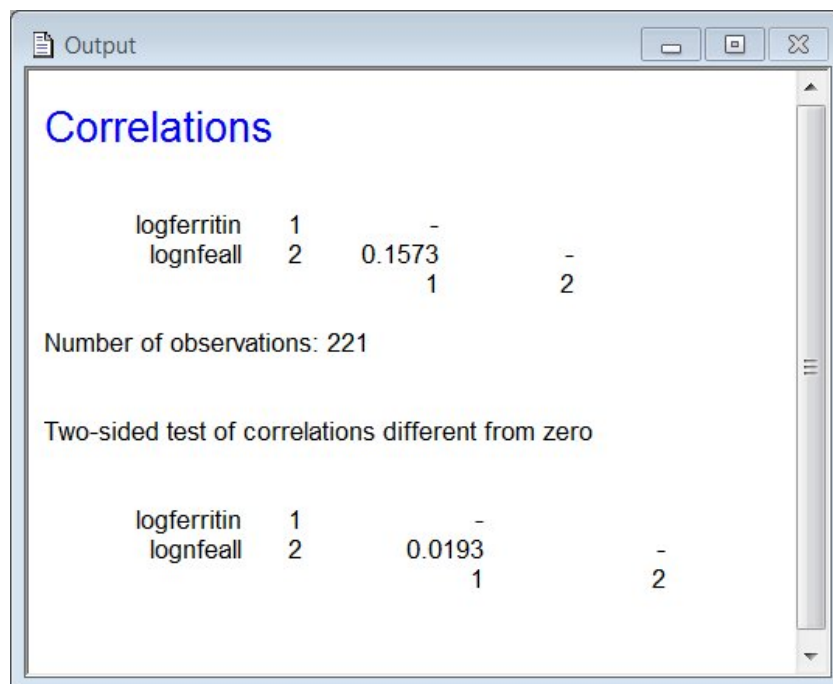
10. Now use correlations to examine the relationships between some of the continuous variables and the iron stores. Start with **nfeall**, the dietary iron level variable. From earlier, you should have found that the distribution of the ferritin data is skewed. Plot a histogram of the dietary iron data to determine whether a transformation is necessary for this data too. If it is, click on **Data>Calculations**, and enter **LOG10(nfeall)** into the box at the top, saving the result to a new variable **lognfeall**. Make sure to do a similar calculation for the **ferritin** variable.

First produce a scatter plot of the two variables using **Graphics>2D Scatter Plot**, entering **logferritin** and **lognfeall** as the variates. The resulting graph is as follows:



Based on the graph, does there appear to be a relationship between the two variables? If so, what is the relationship?

We can test to see if the relationship is significant using the correlation between the two variables. Click **Statistics>Correlations>Correlation Coefficient**, enter **logferritin** and **lognfeall** as the variates, and tick **Test correlations against 0** with a **Two-sided** test. Click **Run** and switch to the Output Window:



If the test value is less than 0.05, there is a significant relationship between the two variables. Is that the case here?

Explore the relationship between (log) iron stores and some of the other continuous variables, e.g. dietary fibre, using log-transforms where necessary. Report your findings.

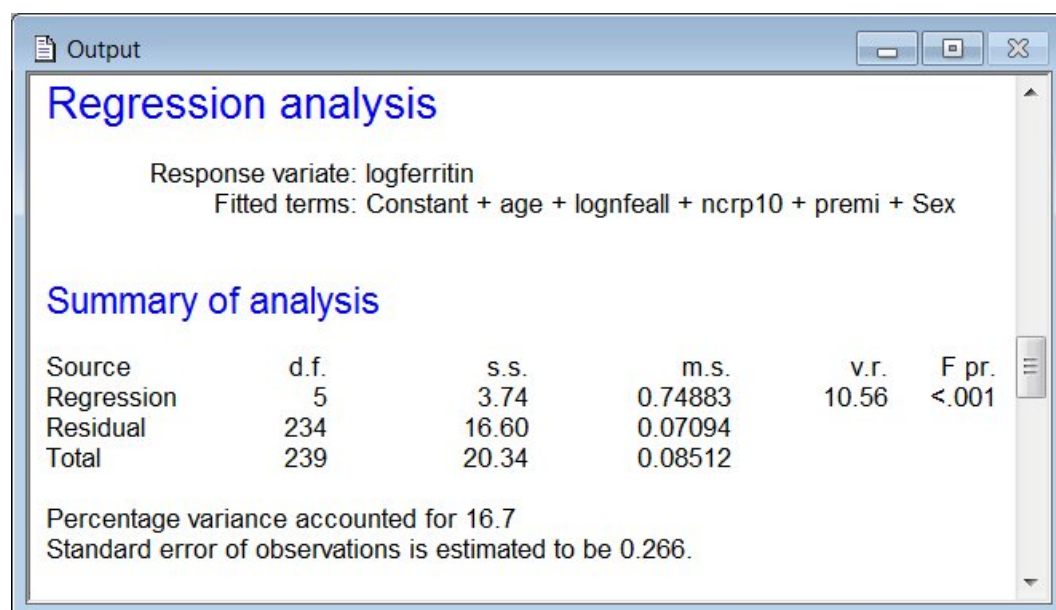
11. A limitation of this approach is that it considers only one variable at a time. A method that allows us to consider several of these variables simultaneously is multiple regression, which is beyond the high school syllabus.

To fit a multiple regression, switch to the Undergraduate version of GenStat by clicking **Tools>Options** and selecting the **Undergraduate Edition**.

Restart GenStat, and open the iron dataset again. Log-transform the dietary iron and iron stores variables as before, and in the spreadsheet, right click on the **Sex** title and choose **Convert to Variate**. This changes **Sex** to a variable coded as 1 for a girl, and 0 for a boy. Also convert **premi** to a variate (the coding is 1 for a prematurely born baby, and 0 otherwise). Double click on the **ncrp10** header in the spreadsheet, and click on the **Levels & Labels** button. Change the **Level** for **Elevated** to 1 and the **Level** for **Normal** to 0. Click **OK** in both boxes, then convert **ncrp10** to a variate.

Now, click on **Stats>Regression Analysis>Linear Models**. Choose **Multiple Linear Regression**, enter **logferritin** as the **Response Variate**, and then **age+lognfeall+premi+ncrp10+Sex** into the **Explanatory variates** box, and click **Run**.

The results in the Output Window should be as follows:



The screenshot shows the 'Output' window in GenStat. The title is 'Regression analysis'. Below it, it states 'Response variate: logferritin' and 'Fitted terms: Constant + age + lognfeall + ncrp10 + premi + Sex'. A section titled 'Summary of analysis' contains a table with the following data:

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	5	3.74	0.74883	10.56	<.001
Residual	234	16.60	0.07094		
Total	239	20.34	0.08512		

Below the table, it states: 'Percentage variance accounted for 16.7' and 'Standard error of observations is estimated to be 0.266.'

Estimates of parameters

Parameter	estimate	s.e.	t(234)	t pr.
Constant	1.693	0.105	16.08	<.001
age	-0.01323	0.00328	-4.03	<.001
lognfeall	0.1479	0.0676	2.19	0.030
ncrp10 Normal	-0.3449	0.0719	-4.80	<.001
premi	-0.0895	0.0489	-1.83	0.069
Sex	0.0677	0.0349	1.94	0.053

Parameters for factors are differences compared with the reference level:

Factor	Reference level
ncrp10	Elevated

Based on the ANOVA, at least one of the explanatory variables is significantly related to the iron store levels ($p < 0.001$), and the parameter estimates table tells us which of these variables (when the effect of all of the others is taken into account) are significantly related.

Use the **t-pr** column to decide which of the variables are significantly related to iron store levels (if the value is less than 0.05, it is considered significant).

Interpret the parameter estimates of the significant variables.