Box Plots and Transformations—CensusAtSchool

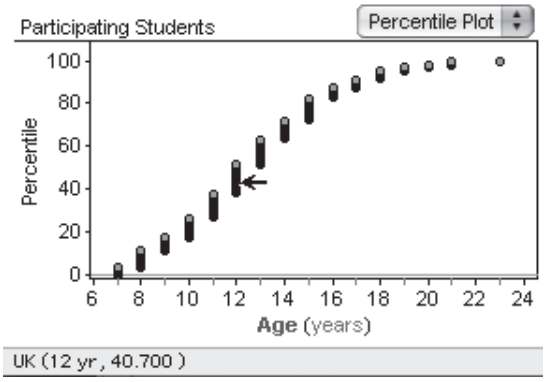
You will need
- **CensusAt School.ftm**

CensusAtSchool is an international project that collects and disseminates data about students from the United Kingdom, South Africa, Australia, and New Zealand. In this activity you'll use Fathom to learn about a random sample of students from this project.

EXAMINE DATA

1. Open the Fathom document **CensusAtSchool.ftm.** In this document you will find a collection of 500 students from the UK, South Africa, and Queensland, Australia, who were randomly selected from the CensusAtSchool project.

2. Make a graph with *Age* on the horizontal axis. Choose **Percentile Plot** from the pop-up menu. Your plot should look like a dot plot, but with the dot positions stretched out vertically. The cases are sorted and then plotted in order. The vertical axis shows the *percentile* of the points. That is, if a point is at 40.7, then 40.7% of the cases have an age less than or equal to this case.



3. Place the cursor on a point. Don't click. When the arrow points straight left, you'll see the coordinates of the point in the status bar at the bottom of the Fathom window. Find the point as close to 50% as you can.

Q1 What is the age at the 50th percentile? What is another name for the 50th percentile?

Q2 What are the ages at the 25th and 75th percentiles? What is another name for these percentiles?

Q3 How old does a student have to be to be in the oldest 10% of the population? Describe how you used the percentile plot to find out. How young does a student have to be to be in the youngest 10%?

Q4 The percentile plot is steep at the beginning, a little less steep in the middle, and shallow at the end. What does that mean for the distribution of students in your sample?

4. Split the graph by dragging *Place* to the vertical axis.

Q5 Use the percentile plots for the three countries to compare their *Age* distributions.

© 2007 Key Curriculum Press

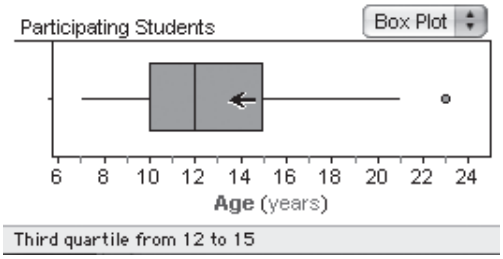
Q6 Choose another numeric attribute and make a new percentile plot for that attribute. Sketch the plot and explain what its shape means for the attribute's distribution.

INVESTIGATE

Box Plots

A box plot is also useful in showing the shape of a distribution and whether or not there are outliers.

5. Select the graph of *Age* and *Place* that you made in step 4 and choose **Graph | Remove Y Attribute: Place.**
6. Choose **Box Plot** from the pop-up menu. A box extends from the 25th percentile (or Q_1) to the 75th percentile (or Q_3) and is cut by a line at the median.
7. To see exactly the median and Q_3, hold the cursor over the right side of the box and read the status bar in the lower-left corner of the Fathom window as shown.



Q7 Using the box plot, find the median, Q_1, and Q_3. Compare your answers to your answers for Q1 and Q2. Were your estimates using the percentile plots fairly accurate?

Q8 The box plot shows an outlier. What is the value of that outlier? Verify that it is an outlier using the $1.5 \cdot IQR$ rule.

8. Split the graph again by dragging *Place* to the vertical axis.

Q9 Using the box plots, describe each country's distribution of *Age* in terms of shape, center, and spread.

Q10 The combined box plot was skewed right a bit and had an outlier, but the box plots from step 8 are all approximately normal with no outliers. Explain why this is so.

9. Add a filter to the graph to examine the distribution of *Age* for the three countries for students less than 15 years old.

With the graph selected, choose **Object | Add Filter.** Type Age<15yr and click **OK.**

Q11 Compare the three distributions of *Age* for students less than 15 years old.

Q12 Contrast the information you can learn from a box plot with that from a histogram. List the advantages and disadvantages of each.

© 2007 Key Curriculum Press

Sliders and Transformations

To experiment with transformations, you'll look at a new distribution: the heights of students from Queensland, Australia.

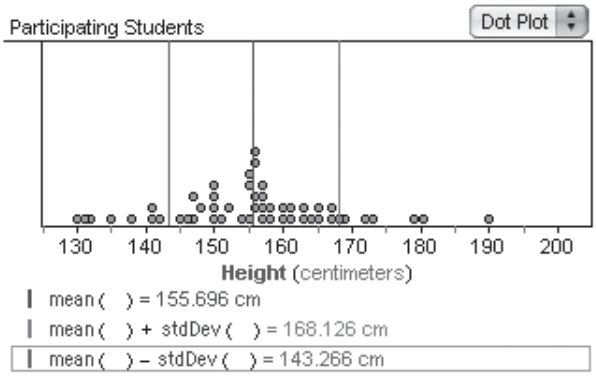
10. Make a new graph and plot *Height* on the horizontal axis.

Place = "Queensland"

11. Filter the *collection* to show only Queensland students.

Choose **Graph | Plot Value** and type mean(). Click **OK.**

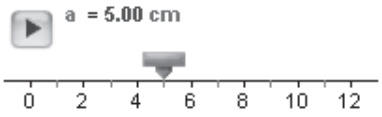
12. On your new plot, plot the mean. Then plot the values mean()+stdDev() and mean()−stdDev().



Q13 Do the two values mean()+stdDev() and mean()−stdDev() capture the majority of the dots?

How would the distribution of heights change if you added a constant value to every height measurement?

13. Drag a slider from the shelf. It will be named *V1*. Rename the slider *a*. Click once to the right of 5.00 and type cm. Press Enter.



In the inspector, click once on the word <*new*>. Type in the name of the new attribute and press Enter. Double-click in the attribute's formula cell and type a+Height. Click **OK.**

14. In the collection's inspector, create an attribute named *Transformed_Height*. Define *Transformed_Height* with the formula *a* + *Height*.

15. Duplicate the *Height* graph by choosing **Object | Duplicate Graph** and replace *Height* with *Transformed_Height*. Use the slider to change the value of the constant *a* and observe the effects.

Q14 Compare the dot plot you made in step 12 to the ones you made in step 15 as you changed the value of the constant *a*. Does the center of your data change? The spread? If so, by how much do the values change?

Q15 Slowly drag the slider thumb to the right, then back to the left. Describe what happens to the distribution of transformed heights. Create formulas for the transformed mean and spread.

© 2007 Key Curriculum Press

Q16 When you add a constant to every value, does the range change? Explain.

Next you'll rescale the data by multiplying each height measurement by a constant.

16. Double-click in the formula cell for *Transformed_Height* and change the formula as shown.

TravelMethod	walk	
Transformed_Height	842.06 cm^2	a ·Height

You will need to change the scale on your axis to see the data. Experiment with different ranges for the horizontal axis.

17. Observe the effects on the plot, mean, and spread as you change the slider this time.

Q17 Compare the original dot plot you made in step 12 to the ones you made in step 17 as you change the value of the constant *a*. Does the center of your data change? The spread? If so, by how much do the values change?

Try some negative values of *a* as well.

Q18 Slowly drag the slider thumb to the right, then back to the left. Describe what happens to the distribution of transformed heights. Create formulas for the transformed mean and spread.

Q19 When you multiply every value by a constant, does the range change?

EXPLORE MORE

Using the techniques you've learned, explore the data to answer these questions or make up your own.

1. Redo the Sliders and Transformations section of this activity using the median as your measure of center and the interquartile range as your measure of spread. For each transformation, answer these questions: Does the center of your data change? The spread? If so, create formulas for the transformed center and spread.
2. Using box plots, compare the distributions of *Age* split by *TravelMethod*. Then use a filter and find out if there are any differences in method of travel between older students and younger students.
3. Using box plots, compare the distributions of *Time_to_Travel* split by *Place*. Which country's students have the longest travel times to get to school? Is there any explanation for the outliers?

© 2007 Key Curriculum Press