Box Plots and Transformations—CensusAtSchool

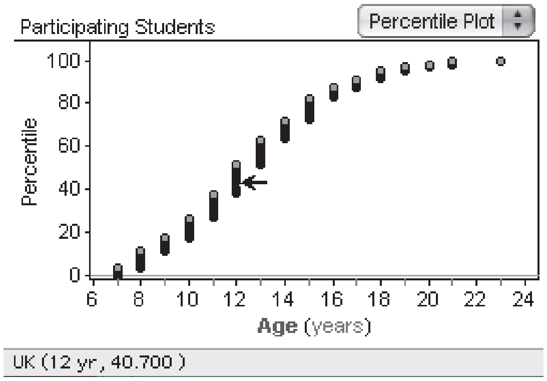
You will need
- **CensusAt School.ftm**

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.

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

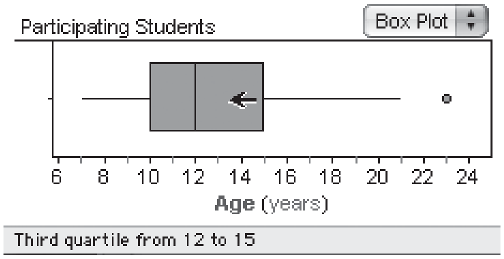
Box Plots

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.



Third quartile from 12 to 15

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.

Sliders and Transformations

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

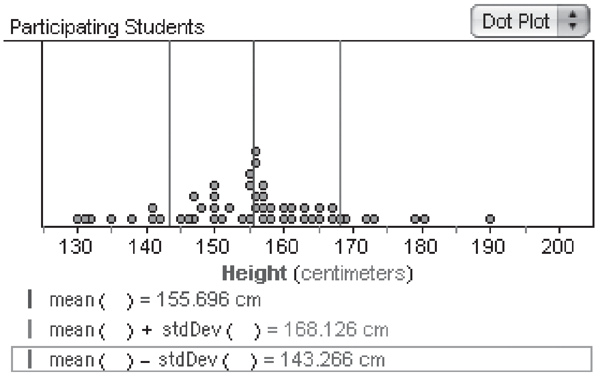
10. Make a new graph and plot *Height* on the horizontal axis.

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().

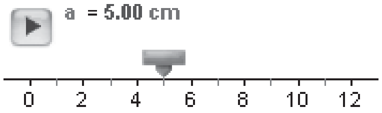


```
| mean( ) = 155.696 cm
| mean( ) + stdDev( ) = 168.126 cm
| mean( ) - stdDev( ) = 143.266 cm
```

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.

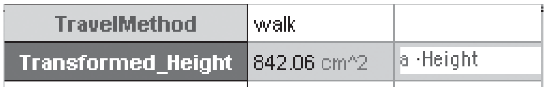
Q14 Compare the dot plot you made in step 12 to the ones you made in step 15 as you changed the value of the constant *a*. Does the center of your data change? The spread? If so, by how much do the values change?

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.

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?

Box Plots and Transformations—CensusAtSchool Activity Notes

Objectives

- Understanding that a percentile is a measure of position and that the quartiles and median are also percentiles
- Learning how to describe or compare distributions, with either box plots or percentile plots
- Finding rules for how summary measures are affected when adding or multiplying each data value by a constant

Activity Time: 40–50 minutes

Setting: Paired/Individual Activity or Whole-Class Presentation (use **CensusAtSchool.ftm** for either)

Statistics Prerequisites

- Comparing distributions
- Familiarity with quartiles
- Familiarity with the mean, median, and spread (standard deviation) in a general sense

Statistics Skills

- Comparing distributions using percentile plots and box plots
- Transformations and their effects on summary measures
- Percentiles and using a percentile plot
- Box plots and the five-number summary
- Finding outliers using the $1.5 \cdot IQR$ rule
- Comparing types of plots

AP Course Topic Outline: Part I A, B (2–5), C, E

Fathom Prerequisites: Students should be able to make and split graphs and plot values.

Fathom Skills: Students make box plots and modified box plots, calculate five-number summaries, use sliders to transform attributes, work with percentiles and reading percentile plots, create formulas, split graphs using categorical variables, and use filters.

General Notes: This activity uses the CensusAtSchool data to introduce students to Fathom. It requires little to no knowledge of Fathom or statistics. This activity focuses on two different topics: measures of position, including percentile plots and box plots, and the effects of transformations on summary measures. Fathom makes it easy for students to work with a large data set and quickly compare distributions. Additionally, students can dynamically change transformation parameters and watch how the distributions change.

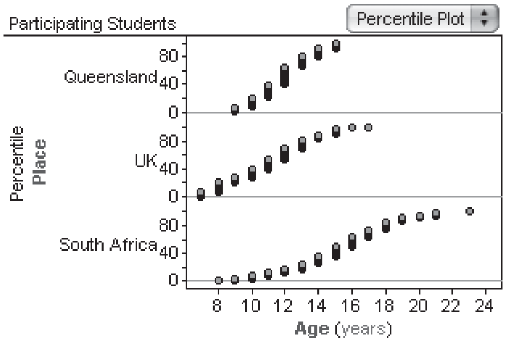
Procedure: Try working with the document **CensusAtSchool.ftm** before giving it to your students. This file has 500 cases. If it's too slow on your computers, select about half the cases and delete them. To select many cases, click on the first row number in the case table. Then scroll halfway down and Shift-click on another row number. All the cases in between will be selected. Choose **Edit | Delete Cases.**

EXAMINE DATA

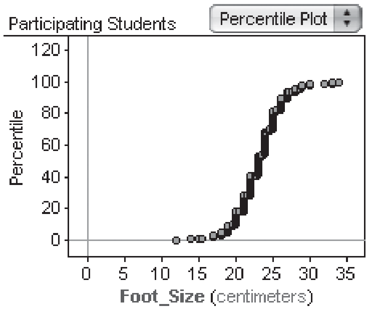
Q1–Q3 The 25th and 75th percentiles are Q_1 and Q_3 respectively. The median is the 50th percentile. The age at the 25th percentile is 10 yr. The age at the 50th percentile is 12 yr. The age at the 75th percentile is 15 yr. The oldest 10% of the population is in the 90th percentile. The age at the 90th percentile is 17 yr, and the age at the 10th percentile is 8 yr. To find these, find points as close as possible to 90 and 10.

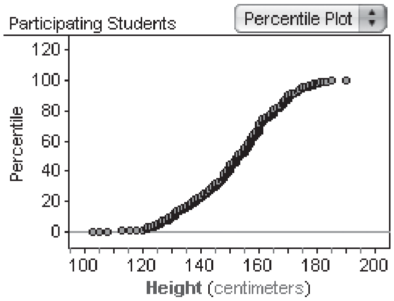
Q4 When the percentile plot is steep over an interval, that means there is a concentration of data values in that interval, whereas when it is shallow, that means there are fewer values in that interval. The percentile plot is steep at first and then gradually becomes shallower, so the age distribution is skewed right.

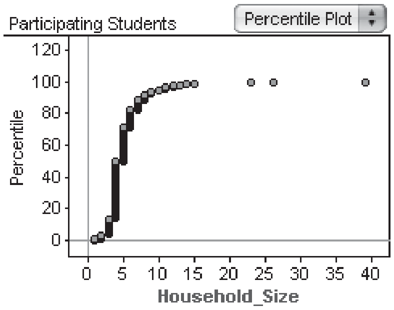
Q5 The split percentile graph shows that the narrowest range of ages is for the Queensland students; the South African students span the greatest range and have all the oldest students; and the youngest students are from the United Kingdom. The percentile plot for South Africa starts out shallow and gets steeper in the middle and then levels out at about 20 years old. The plot for the UK is, from the beginning, fairly steep and looks almost linear, then at the very last two dots it levels out. The plot for Queensland is mostly steeper still but has a small shallow part around 15.



Q6 The percentile plot for *AgeMonths* is naturally similar to the plot for *Age*, but it shows a more detailed distribution. The plot for *Foot_Size* is steep in the center with a shallow part at either end, meaning that most of the data are concentrated between about 19 cm and 25 cm. The plot for *Height* doesn't have such a steep middle section, indicating that the heights are more spread out. The plots for *Household_Size, SchoolAgedMales,* and *SchoolAgedFemales* are all quite steep at the beginning, with long shallow ends, meaning that household sizes tend to be small. Similarly, the plot for *Time_to_Travel* has a steep early part and a long shallow end.







INVESTIGATE

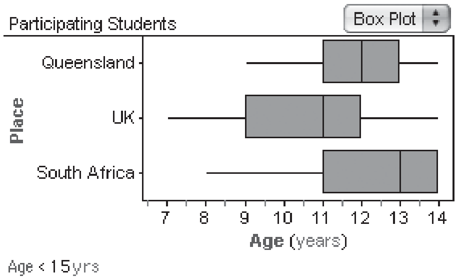
Q7 Median = 12, $Q_1 = 10$, and $Q_3 = 15$. These values should be similar to, if not exactly the same as, the estimates in Q_1 and Q_2.

Q8 The outlier is at 23 years. The *IQR* is 5, so $1.5 \cdot IQR = 7.5$. Then $Q_3 + 1.5 \cdot IQR = 22.5$, so 23 is definitely an outlier.

Q9 The plot again shows that the narrowest range of ages is for the Queensland students; the South African students span the greatest range and have all the oldest students; and the youngest students are from the United Kingdom. The box plots clearly show, however, that all three are fairly mound-shaped and that the students from South Africa are definitely older—75% of the South African students are 14 or older, whereas for both the UK and Queensland, 75% of the students are younger than 14. The spreads (*IQR*) for both South Africa and the UK are the same (4 yr), while Queensland has a spread (*IQR*) of 2 yr. The box plot tells a great deal more about the shape, center, and spread than both the histogram and the dot plot in this case.

Q10 The outlier (23) is from South Africa. When all of the ages are lumped together into one distribution, 23 years old is quite a bit older than everyone else; however, compared to only the other South African students, the outlier is not that much older. For the whole group, $Q_3 = 15$, but for South Africa, $Q_3 = 18$. So, $Q_3 + 1.5 \cdot IQR = 18 + 1.5(4) = 24$ instead of 22.5.

Q11 The distributions of *Age* for both South Africa and the UK are skewed toward the younger values. The middle half of the ages of the younger Queensland students are between 11 and 13. The spreads for the younger UK and South African students are a bit larger: The middle half of the ages are between 9 and 12 for the UK and between 11 and 14 for South Africa. Half of the younger Queensland students have ages above 12 and half 12 or below. Half of the younger South African students have ages above 13 and half 13 or below, and half of the younger UK students have ages above 11 and half 11 or below. Note that for South Africa there is no upper whisker and 50% of the students are between 13 and 14.



Q12 From a box plot, you can see the five-number summary exactly and outliers are clearly marked. These must be estimated from a histogram, which can be difficult. From a histogram, you can estimate the mean by estimating a balance point for the distribution. You cannot do this with a box plot. A histogram will reveal the frequency of the data within an interval. You do not know the exact values but you know how many are within the given boundaries. You know a lower bound and an upper bound, but not necessarily the exact least and greatest values. You know where there are clusters of data and where there are gaps. With a box plot, you get a sense of the basic shape of the distribution but you cannot see clusters or gaps. You cannot see the frequency but you can see the proportions.

10. This group of students was chosen because there are relatively few cases and it is easier to work with the transformed data than with the whole group.

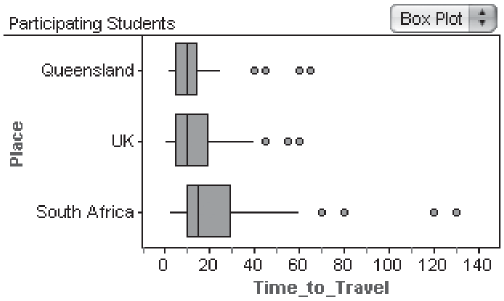
Q13 The majority of dots are captured between the two values—38 out of 52 are captured.

Q14–Q16 When adding the constant *a* to each height, the whole distribution shifts by that constant. So the range and spread stay the same because the same constant is added to all values in the range (including the minimum and maximum values). The mean changes to *original mean* + *a*.

Q17–Q19 When multiplying height by the constant *a*, all three values change. Students will need to resize their plots to see their data. The new mean will be *original mean* · *a*. The new range and the new spread will be the original values multiplied by $|a|$.

EXPLORE MORE

1. When adding the constant *a* to each height, the whole distribution shifts by that constant. So the *IQR* stays the same but the median changes to *original median* + *a*. When multiplying height by the constant *a*, both values change. Students will need to resize their plots to see their data. The new median will be *original median* · *a*. The new *IQR* will be *original IQR* · $|a|$.
2. Taxis have the largest median age associated with them at 16; cars and cycles tie for the lowest median age at 10; walkers have the largest *IQR* at 5 yr. The largest range of older students take taxis, walk, or ride the bus, while the largest range of younger students walk, ride in a car or the bus, or take an unlisted mode of transportation.
3. It appears that the South African students have the longest travel time: 50% have a travel time of at least 15 min. The UK and Queensland students both have median travel times of 10 min. This could be due to the large number of students who walk to school in South Africa. The outliers could represent students that live in very rural areas served by regional schools or students on the last stop on a bus route.


Participating Students
Box Plot
Place
Queensland
UK
South Africa
Time_to_Travel