LESSON 14: Box plots questions

FOCUS QUESTION: How can I compare the distributions for data sets that have outliers?

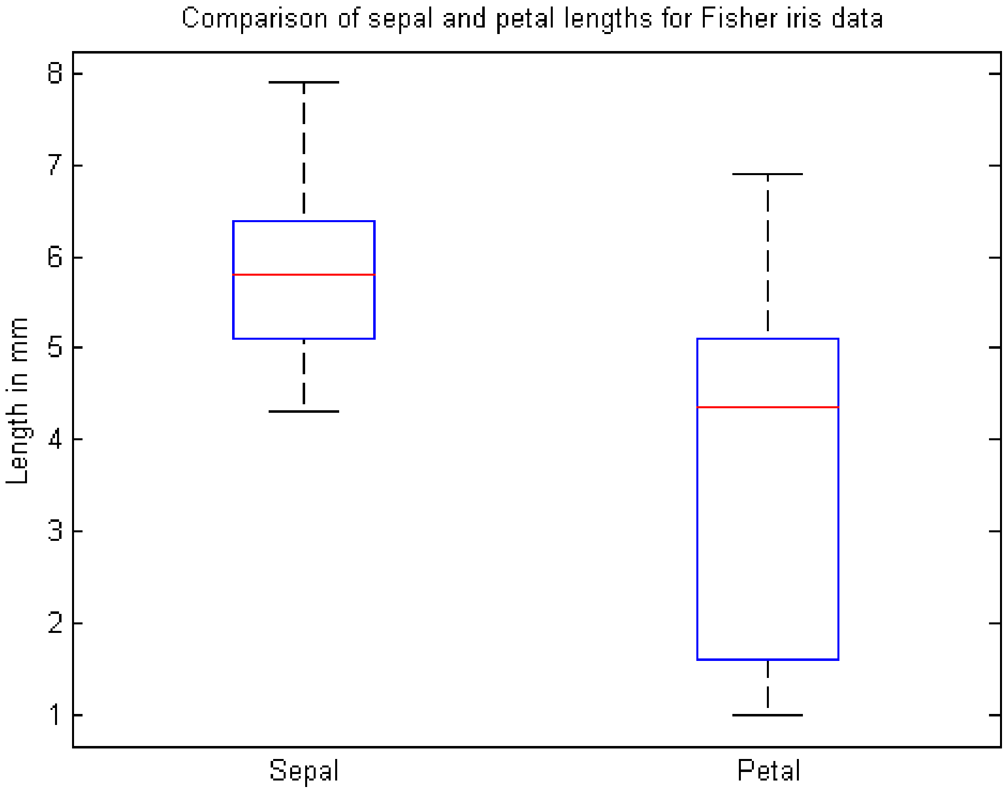
Contents

EXAMPLE 1: Load the Fisher iris data (comes with MATLAB)

```
load fisheriris;
```

EXAMPLE 2: Compare the distributions of sepal and petal lengths using box plots

```
flowerLens = meas(:, [1, 3]);
figure
boxplot(flowerLens, 'Label', {'Sepal', 'Petal'})
ylabel('Length in mm')
title('Comparison of sepal and petal lengths for Fisher iris data')
```

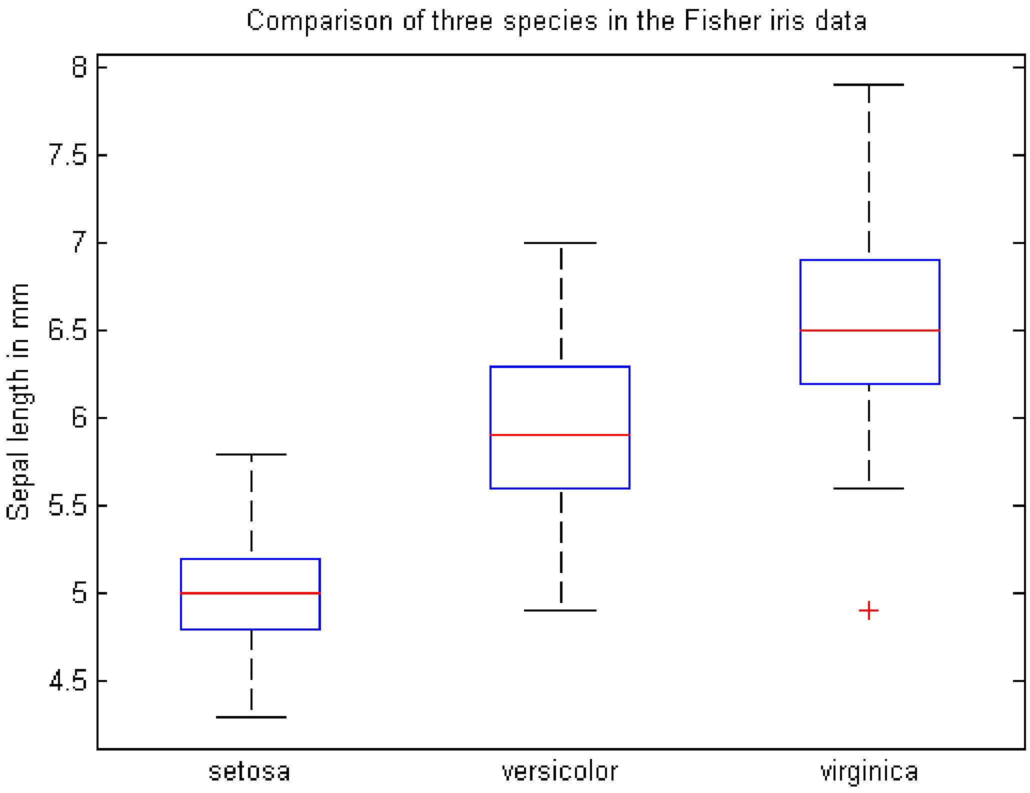


Questions	Answers
How many rows and columns does `lengths` have?	The `lengths` variable holds an array with 150 rows and 2 columns.
What do the curly brackets (`{}`) represent?	The curly brackets designate a cell array.
What is a cell array?	A cell array is a MATLAB data structure designed to hold a collection of elements of different sizes.
Do I need a cell array?	Since `'Sepal'` and `'Petal'` happen to be strings of the same length, we could have used character arrays in this case. Usually it is easier to use cell arrays.
What does each vertical diagram in the figure represent?	Each diagram is a box plot, representing the distribution of of one of the columns. The two columns in this case hold sepal lengths and petal lengths, respectively.
Why are box plots useful?	Box plots show many details of a distribution that are not visible in the histogram. They are particularly good for distributions that have outliers.
What does the line across the central region of each box represent?	The central line marks the data set median.
What does the bottom of each box represent?	The bottom of the box marks the 25th percentile for the data set.
What is the 25th percentile?	The approximate definition of 25th percentile is the value such that 25% of the data set values are below and 75% of the data set values are above. Unfortunately there is no universally accepted formula for computing percentiles and so the results you get from different software will vary, especially when the data sets are small.

What does the top of each box represent?	The top of the box marks the 75th percentile for the data set.
What does the height of each box represent?	The height of the box represents the inter quartile range (IQR) of the data set.
Is the median line always in the center of the box?	No, the distribution of the values on either side of the median doesn't have to be symmetric, so the median line could fall anywhere in the box depending on the distribution.
What do the ends of the whiskers (lines extending from the boxes) represent?	These lines mark the highest and lowest values of the data set that are within 1.5 times the inter quartile range of the box edges (the fence).
Is it possible to make the whiskers represent a different interval?	Yes, this is done with the `'whisker'` parameter of `boxplot`. For example `boxplot(sepalLengths, 'whisker', 2.0)` creates a box plot in which the whiskers represent the highest and lowest values within two IQR's of the box edges.
What do the plus signs (+) represent?	The plus signs mark individual values outside the range of the whiskers

EXAMPLE 3: Draw a box plot of the sepal lengths by species

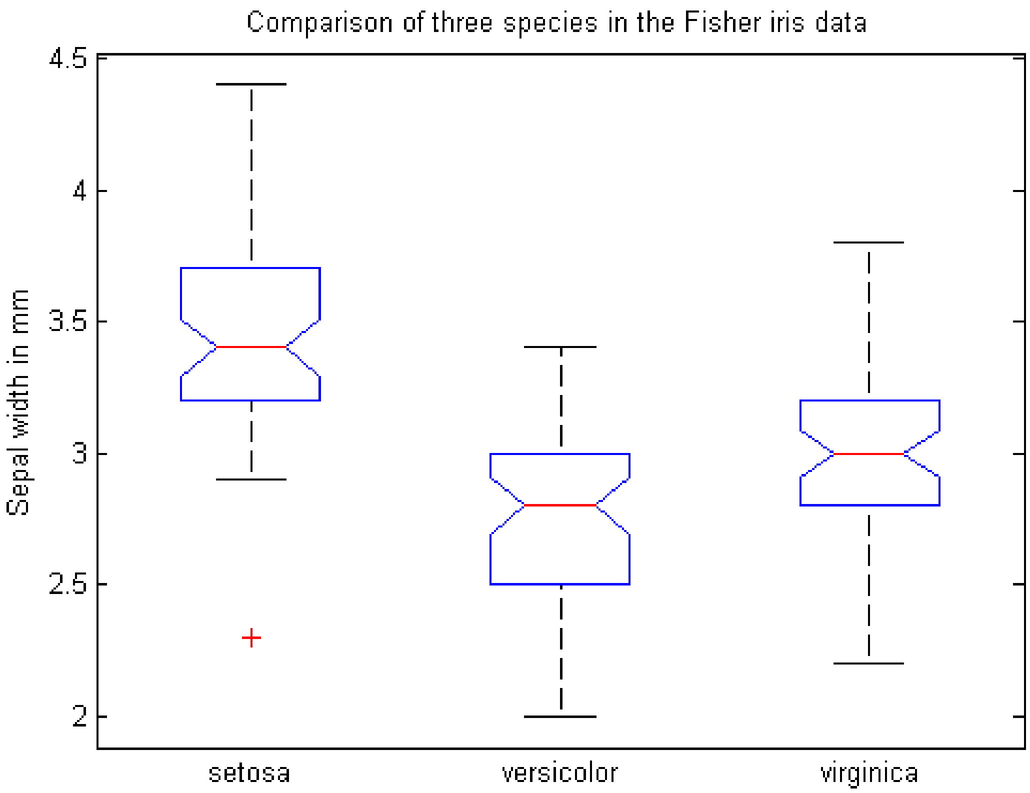
```
    sepalLens = meas(:, 1);
    figure
    boxplot(sepalLens, species)
    ylabel('Sepal length in mm')
    title('Comparison of three species in the Fisher iris data')
```



MATLAB syntax	Description
How is the grouping of data into boxes determined in this example?	Normally, MATLAB draws a box plot for each column. However, another option is to provide a single column of data along with another vector of the same size giving the labels for each data point. The data are then grouped into boxes by label.
How many boxes will there be when I use grouping?	MATLAB creates a separate box for each unique label in the grouping vector, `species`.

EXAMPLE 4: Draw a notched box plot of the sepal widths

```
sepalWidths = meas(:, 2);
figure
boxplot(sepalWidths, species, 'notch', 'on')
ylabel('Sepal width in mm')
title('Comparison of three species in the Fisher iris data')
```



MATLAB syntax	Description
Why notch the box plot?	The notch marks the 95% confidence interval for the medians. We are 95% certain that the actual median for the underlying population actually is within the interval marked by the notches. If the notches from two box plots don't overlap, we can assume at the (0.05 significance level) that the medians are different. Note: the actual definition of the 95% confidence interval is that it is an estimate of an interval that the median will be in for 95% of the samples.

Why should I care about confidence intervals and significance?	Confidence intervals provide a statistical measure for how reliably the sample data represents an underlying population. (After all you can't measure all of the Iris plants of a certain type, so you never can know for certain.) If we make more measurements, the confidence interval will shrink --- that is you can narrow down the possible range of the median. Confidence intervals also give us an indicator of how many plants you need to measure to get a reliable estimate of the median.

EXAMPLE 5: Load the Daphne and Santa Cruz beak size data

```
 Daphne = load('DaphneBeaks.txt');
 SantaCruz = load('SantaCruzBeaks.txt');
```

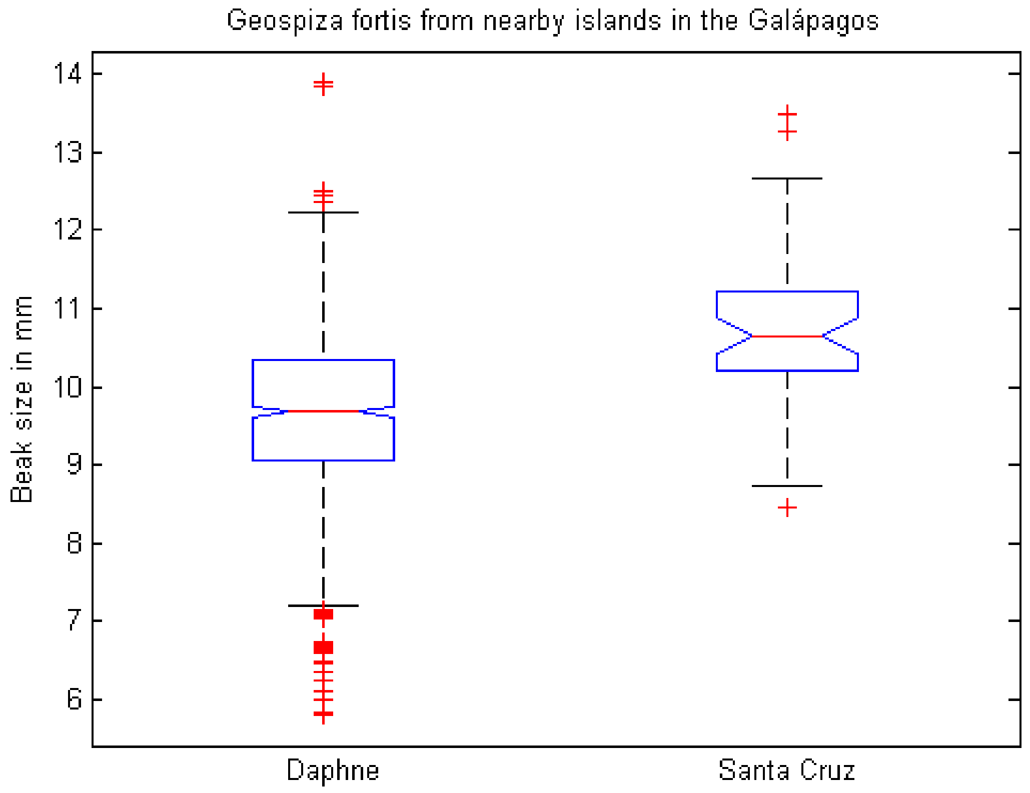
EXAMPLE 6: Create a labeled vector of beak sizes for plotting

```
beakSizes = [Daphne; SantaCruz];
islands = [repmat('  Daphne  ', size(Daphne)); ...
           repmat('Santa Cruz', size(SantaCruz))];
```

MATLAB syntax	**Description**
What does the variable `islands` hold?	The `islands` variable holds an array of strings in a single column. The first 751 elements of `island` hold `'  Daphne  '` and the remaining 43 elements hold `'Santa Cruz'`
Why were there blanks around the word Daphne?	The forming of an array using vertical concatenation with the `semicolon` requires that the respective components be of the same width.

EXAMPLE 7: Create a box plot of unequal length data sets using labeled data

```
figure
boxplot(beakSizes, islands, 'notch', 'on')
ylabel('Beak size in mm')
title('Geospiza fortis from nearby islands in the Galápagos');
```



MATLAB syntax	Description
Why is it necessary to use labels in this example?	The Daphne and Santa Cruz data sets have different numbers of elements so it would not be possible to arrange them as columns in the same array.
How does this example use the grouping?	The labeled groups allow us to by-pass the problem of unequal size data sets.

This lesson was written by Kay A. Robbins of the University of Texas at San Antonio and last modified on 08-Nov-2012. Please contact krobbins@cs.utsa.edu with comments or suggestions. The photo was taken by Danielle Langlois in July 2005 and is available under public license at http://commons.wikimedia.org/wiki/File:Iris_versicolor_3.jpg.

Published with MATLAB® 8.3