LESSON 14: Box plots

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

In this lesson you will:

- Use box plots to compare distributions from different data sets.
- Use groupings of labeled data.
- Learn about median and the inter quartile range (IQR) as indicators of central tendency.

Contents

DATA FOR THIS LESSON

File	Description
	This data set contains the famous Fisher iris data set. The data set consists of measurements of 150 flower samples from each of three species of flowers: *Iris setosa*, *Iris virginica*, and *Iris versicolor*. The measurements are in mm. Four features were measured for each sample: ▪ The length of the flower sepal ▪ The width of the flower sepal ▪ The length of the flower petal ▪ The width of the flower petal All 150 samples from the Fisher iris data are stored in a single table called `meas`: ▪ The four columns correspond to the four types of measurements: sepal length, sepal width, petal length and petal width, respectively.

`fisheriris`	▪ The first 50 rows contain data for *Iris setosa* ▪ The second 50 rows contain data for *Iris virginica* ▪ The third 50 rows contain data for *Iris versicolor*. The species information is kept in a separate vector called `species`. The data is sometimes referred to as Anderson's Iris data in honor of Edgar Anderson, the biologist who collected the data. See http://en.wikipedia.org/wiki/Iris_flower_data_set for additional information. **Note:** This dataset comes with the MATLAB distribution so you don't have to download it separately.
DaphneIslandBeaks.txt SantaCruzIslandBeaks.txt	▪ The data set consists of measurements of beak sizes in mm of Darwin's ground finch (*Geospiza fortis*) taken at Daphne Island and at Santa Cruz Island in the Galápagos by Peter and Rosemary Grant. ▪ The populations of the two islands differ, although the islands are less than 10 km apart. ▪ The data was extracted from a data set distributed with the case study Natural Selection and Darwin's Finches by Martin Wikelski available on the web at http://wps.prenhall.com/esm_freeman_evol_3/0,8018,8412374-,00.html. ▪ The original data is summarized in the article: "The classical case of character release: Darwin's finches (Geospiza) on Isla Daphne Major, Galápagos" by P. T. Boag and P. R. Grant that appeared in *Biological Journal of the Linnean Society* 22:243-287 (11284). See http://en.wikipedia.org/wiki/Peter_and_Rosemary_Grant for additional information on the work of Peter and Rosemary Grant.

SETUP FOR LESSON 14

- Set the Current Directory to `Z:\working\MATLAB\Lesson14`. (You will need to make a new directory for `Lesson14`.)
- Download the data file to your `Lesson14` directory.
- Create a `lesson14Script` script file in your `Lesson14` directory.

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

Create a new cell in which you type and execute:

```
load fisheriris;
```

You should see the following 2 variables in your Workspace Browser:

- `meas` - an array in which each column corresponds to a particular type of measurement and each row corresponds to the 4 measurements for a particular speciman. The different species are combined into a single array.
- `species` - a cell column vector containing the species designation for the speciman given in the corresponding row of `meas`. Possible values are `'setosa'`, `'versicolor'`, and `'virginica'`.

EXERCISE 1: Diagramming an array
Draw a picture of the `meas` and `species` arrays and label their rows and columns. How are the rows and columns of these arrays related?

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

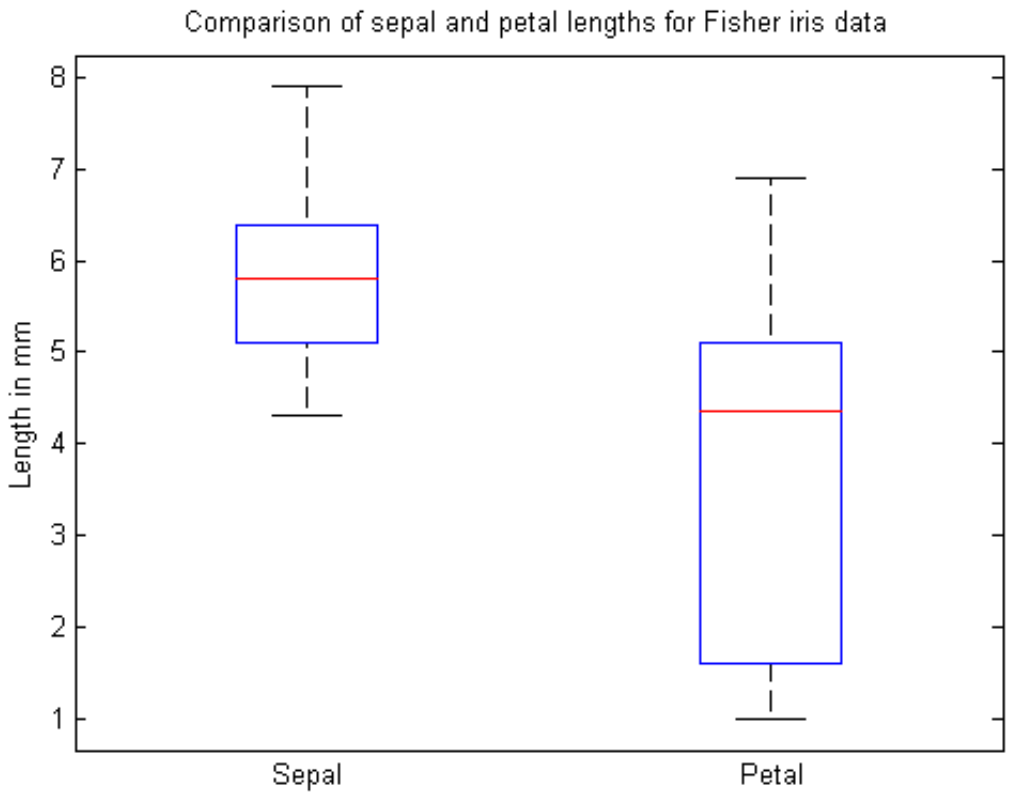
Create a new cell in which you type and execute:

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

You should see the following variable in your Workspace Browser:

- `flowerLens` - an array containing 2 columns corresponding to the sepal and petal lengths, respectively.

You should also a Figure Window with a labeled box plot:



EXERCISE 2: Create a three-column disease array
Load the NYC diesases data sets (`NYCDiseases.mat`). Create variable called `diseases` that holds a three-column array. The first column is the monthly counts of measles, the second column is the monthly counts of mumps, and the third column is the monthly counts of chicken pox.

EXERCISE 3: Display and label box plots of NYC diseases
Create a box plot similar to that of EXAMPLE 2 for the `diseases` array.

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

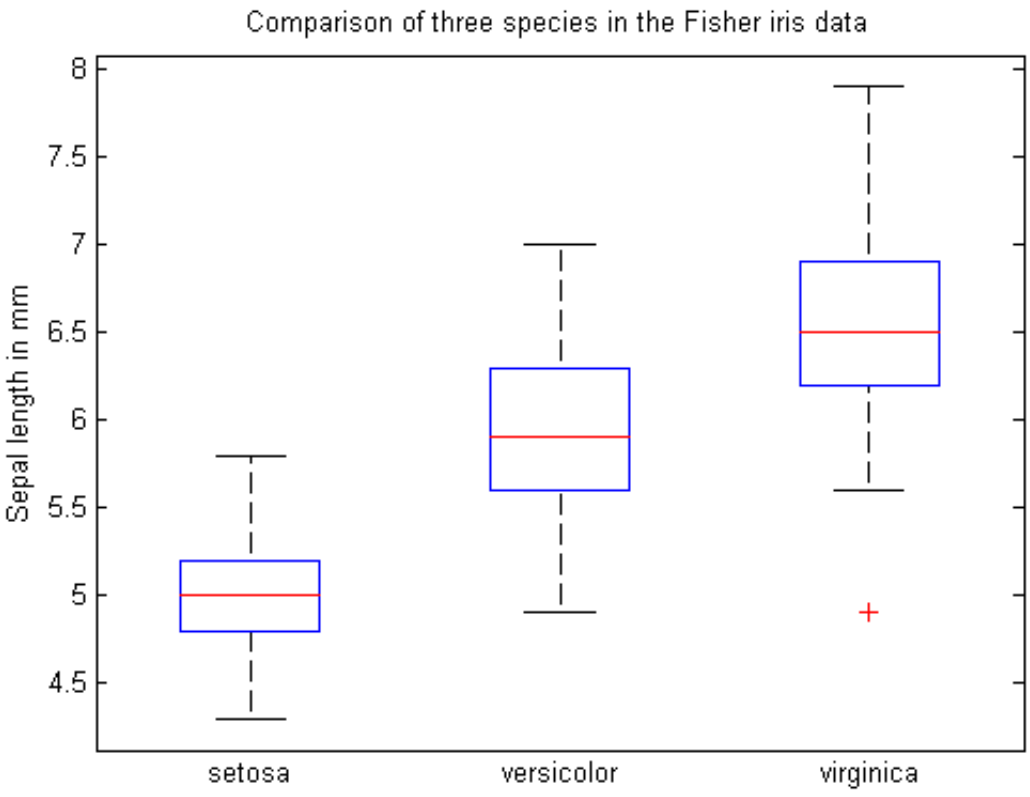
Create a new cell in which you type and execute:

```
sepalLens = meas(:, 1);        % Define a variable for the sepal length
figure
boxplot(sepalLens, species)    % The species vector specifies the group
ylabel('Sepal length in mm')
title('Comparison of three species in the Fisher iris data')
```

You should see the following variable in your Workspace Browser:

- `sepalLens` - a vector containing the sepal lengths of all 150 specimans

You should also a Figure Window with a labeled box plot:



EXERCISE 4: Display box plots of petal lengths

EXERCISE 5: Load the `diaries.mat` data of Lesson 10.
Define a variable called `totalAlarm` that holds the total number of times each subject in the cohort used the alarm.

EXERCISE 6: Show boxplots of total alarm use by gender
Display figure similiar to EXAMPLE 3 with box plots of total alarm use broken down by gender.

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

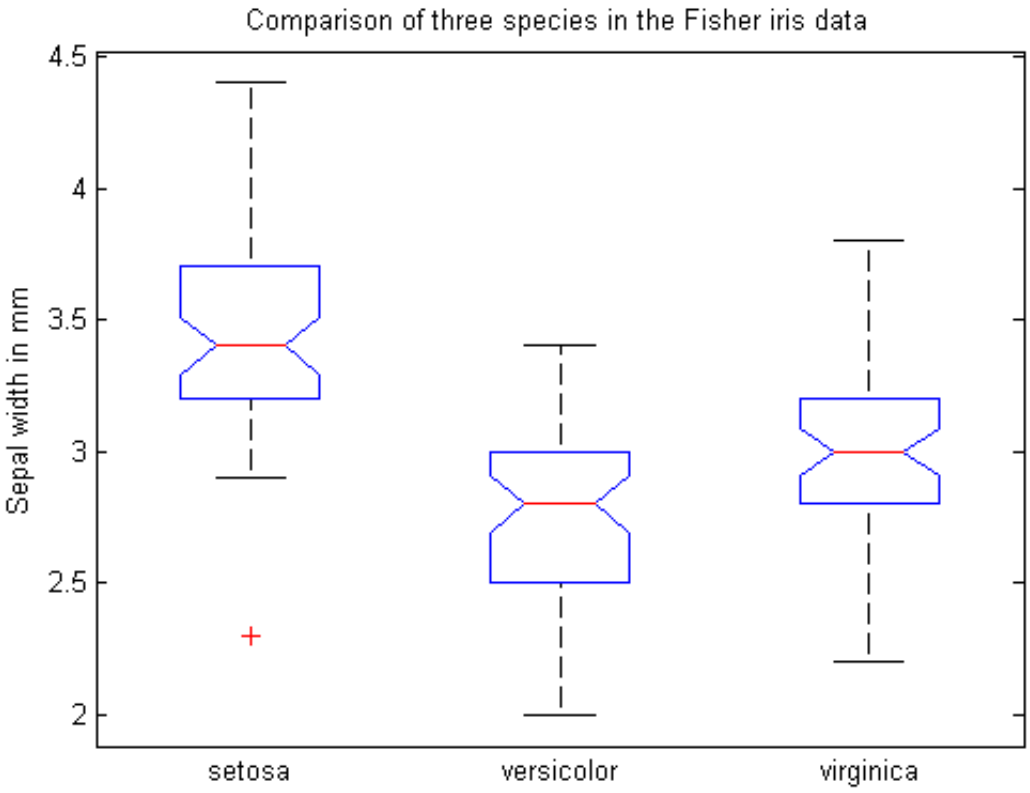
Create a new cell in which you type and execute:

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

You should see the following variable in your Workspace Browser:

- `sepalWidths` - a vector containing the sepal widths of all 150 specimans

You should also a Figure Window with a labeled box plot:



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

Create a new cell in which you type and execute:

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

You should see the following 2 variables in your Workspace Browser:

- `Daphne` - a column vector with beak sizes of the Daphne Island finches
- `SantaCruz` - a column vector with the beak sizes of the Santa Cruz Island finches

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

Create a new cell in which you type and execute:

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

You should see the following 2 variables in your Workspace Browser:

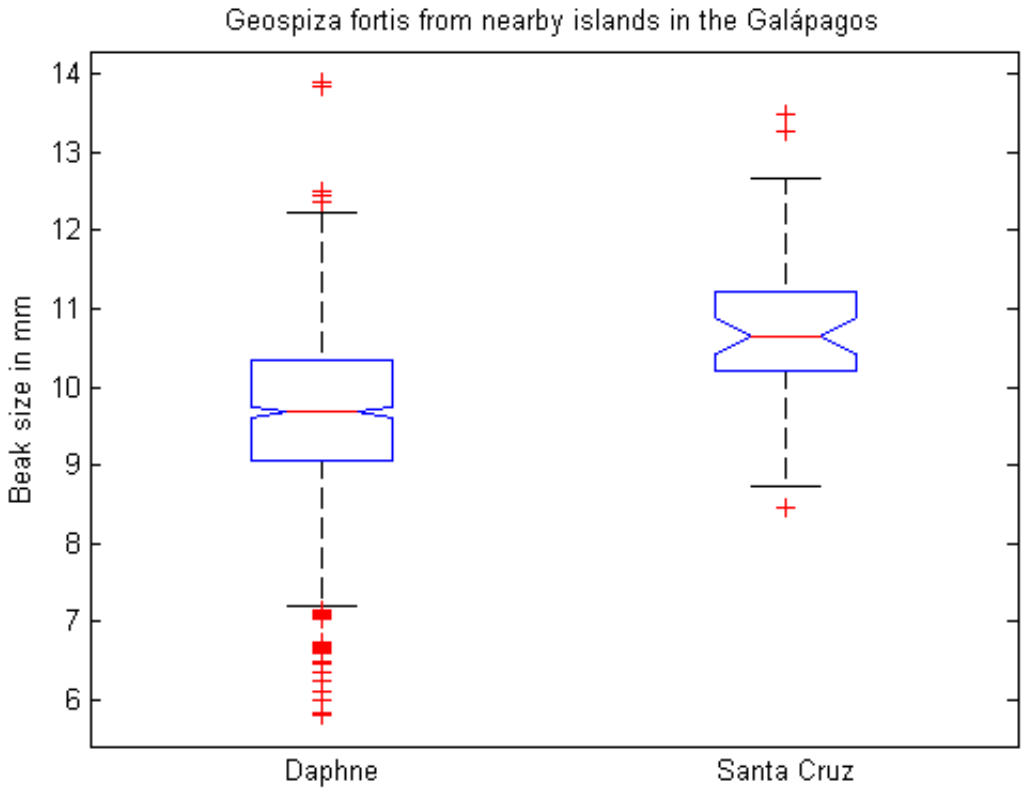
- `beakSizes` - a vector containing the beak sizes for all of the birds
- `islands` - a cell vector with the island designations corresponding to the values in `beakSizes`

EXAMPLE 7: Display box plots of the beak sizes

Create a new cell in which you type and execute:

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

You should see a Figure Window with a labeled box plot:



SUMMARY OF SYNTAX

MATLAB syntax	Description
boxplot(X)	Creates a box plot of the values in the array `X`. Each column of `X` is treated as a distinct data set and gets its own box. The `boxplot` function has a large number of optional parameters. We used the following options: ▪ The `'Label', labels` option provides string labels for the individual boxes(EXAMPLE 2). ▪ The `'notch', 'on'` (EXAMPLE 4) creates boxes that have V-shaped notches. The notches mark the 95% confidence intervals for the median. Also of interest are the `'orientation', 'horizontal'` and `'plotstyle', 'compact'` options, which are useful for display a large number of box plots in the same figure.
boxplot(X, labels)	Creates a box plot for each of the unique values in `labels`. The `boxplot` command uses the `labels` vector as an index vector to separate the values in `X` into different boxes. The `X` and `labels` vector must be of the same length.
repmat(X, n, m)	creates a new array by tiling the array `X` in a pattern with `n` rows and `m` columns.
repmat(X, size(A))	creates a new array by tiling the array `X` in a pattern whose size is the same size as the array `A` (i.e., the pattern has the same number of rows and columns as `A` does).

This lesson was written by Kay A. Robbins of the University of Texas at San Antonio and last modified on April 12, 2015. 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