Statistics 120
Plots Based on Quantiles

Percentiles and Quantiles

The k-th *percentile* of a set of values divides them so that $k\%$ of the values lie below and $(100-k)\%$ of the values lie above.

- The 25th percentile is known as the *lower quartile*.
- The 50th percentile is known as the *median*.
- The 75th percentile is known as the *upper quartile*.

It is more common in statistics to refer to *quantiles*. These are the same as percentiles, but are indexed by sample fractions rather than by sample percentages.

Some Difficulties

The previous definition of quantiles and percentiles is not completely satisfactory. For example, consider the six values:

3.7 2.7 3.3 1.3 2.2 3.1

What is the lower quartile of these values?

There is no value which has 25% of these numbers below it and 75% above.

To overcome this difficulty we will use a definition of percentile which is in the spirit of the above statements, but which (necessarily) makes them hold only approximately.

Defining Quantiles

We define the quantiles for the set of values:

3.7 2.7 3.3 1.3 2.2 3.1

as follows.

First sort the values into order:

1.3 2.2 2.7 3.1 3.3 3.7

Associate the ordered values with sample fractions equally spaced from zero to one.

Sample fraction	0	.2	.4	.6	.8	1
Quantile	1.3	2.2	2.7	3.1	3.3	3.7

Defining Quantiles

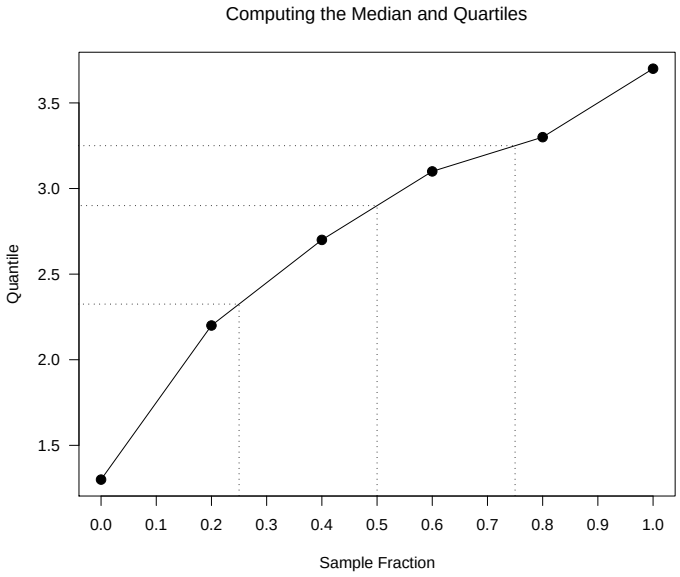
The other quantiles of

$$1.3 \quad 2.2 \quad 2.7 \quad 3.1 \quad 3.3 \quad 3.7$$

can be obtained by linear interpolation between the values of the table.

The median corresponds to a sample fraction of .5. This lies half way between 0.4 and 0.6. The median must thus be $.5 \times 2.7 + .5 \times 3.1 = 2.9$

The lower quartile corresponds to a sample fraction of .25. This lies one quarter of the way between .2 and .4. The lower quartile must then be $.75 \times 2.2 + .25 \times 2.7 = 2.325$.


Computing the Median and Quartiles
Quantile
3.5
3.0
2.5
2.0
1.5
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Sample Fraction

The General Case

Given a set of values $x_1, x_2, \ldots . x_n$ we can define the quantiles for any fraction p as follows.

Sort the values in order

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

The values $x_{(1)}, \ldots, x_{(n)}$ are called the *order statistics* of the original sample.

Take the order statistics to be the quantiles which correspond to the fractions:

$$p_i = \frac{i-1}{n-1}, \quad (i = 1, \ldots, n),$$

The Quantile Function

In general, to define the quantile which corresponds to the fraction p, use linear interpolation between the two nearest p_i.

If p lies a fraction f of the way from p_i to p_{i+1} define the pth quantile to be:

$$Q(p) = (1 - f)Q(p_i) + fQ(p_{i+1})$$

As special cases, define the median and quartiles by:

Median:	$Q(.5)$
Lower Quartile:	$Q(.25)$
Upper Quartile:	$Q(.75)$

The function Q defined in this way is called the *Quantile Function*.

Computing Quantiles with R

The R function `quantile` can be used to compute the quantiles of a set of values.

```
> x = c(1.3, 2.2, 2.7, 3.1, 3.3, 3.7)
> quantile(x)
   0%   25%   50%   75%  100%
1.300 2.325 2.900 3.250 3.700
```

```
> quantile(x, seq(0, 1, by = 0.1))
  0%  10%  20%  30%  40%  50%  60%  70%
1.30 1.75 2.20 2.45 2.70 2.90 3.10 3.20
 80%  90% 100%
3.30 3.50 3.70
```

Graphs Based on Quantiles

- There are many types of graph which are based on the quantiles of a set of observations.
- Different sets of quantiles provide varying degrees of data summary.
- It is this varying level of summarisation makes the use of quantiles attractive.

Boxplots and Variations

- Real name *box-and-whisker plots*.
- Draw a box from the lower quartile to the upper quartile.
- Extend a whisker from the ends of the box to the furthest observation which is no more than 1.5 times inter-quartile range from the box.
- Mark any observations beyond this as “outliers.”

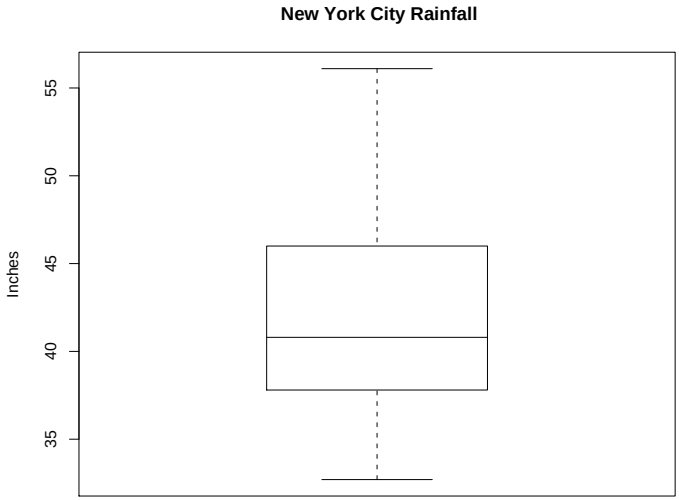
Producing Barplots With R

A single boxplot, vertically aligned.

```
> boxplot(rain.nyc,
          main = "New York City Rainfall",
          ylab = "Inches")
```

The basic call can be heavily customized.

```
> boxplot(rain.nyc,
          col = "lightgray",
          horizontal = TRUE,
          main = "New York City Rainfall",
          xlab = "Inches")
```


New York City Rainfall
Inches
35
40
45
50
55

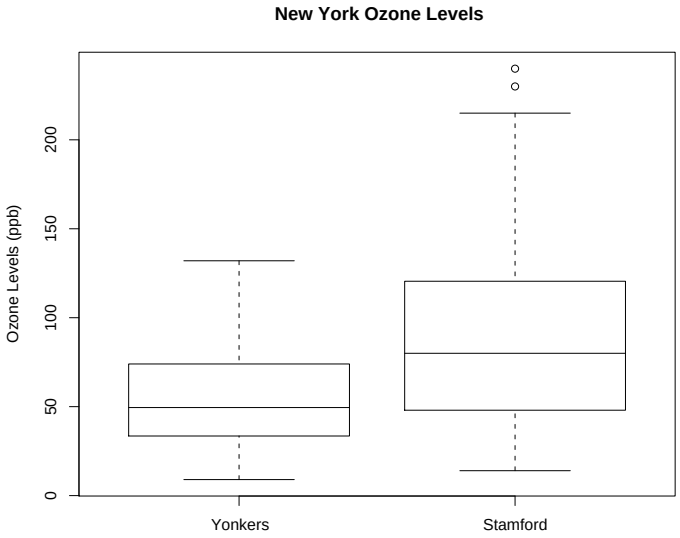
Comparing Samples with Boxplots

- It is possible to compare two or more samples with boxplots.
- By producing the plots on the same scale we are able to make direct comparisons of:
  - medians
  - quartiles
  - inter-quartile ranges
- The comparison of medians using boxplots can be regarded as the graphical equivalent of two-sample *t*-tests and one-way analysis of variance.

The New York Ozone Data

- The ozone levels in Yonkers and Stamford can be compared with boxplots.
- The two samples are passed to boxplot as separate arguments.
- Labels can be provided to label the two samples.

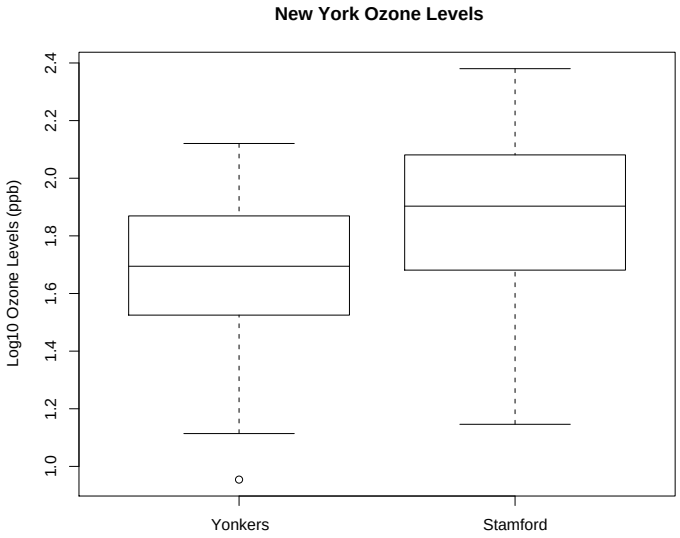
```
> boxplot(yonkers, stamford,
          names = c("Yonkers", "Stamford"),
          main = "New York Ozone Levels",
          ylab = "Ozone Levels (ppb)")
```


New York Ozone Levels
Ozone Levels (ppb)
0
50
100
150
200
Yonkers
Stamford

Transformations

- The median ozone level in Stamford is higher than that in Yonkers.
- The spread of the values in Stamford is also larger than the spread in Yonkers.
- When there is a difference in data spreads it is common to transform the values so that the speads are equal.
- This makes it possible to compare the medians in the absence of any other differences.

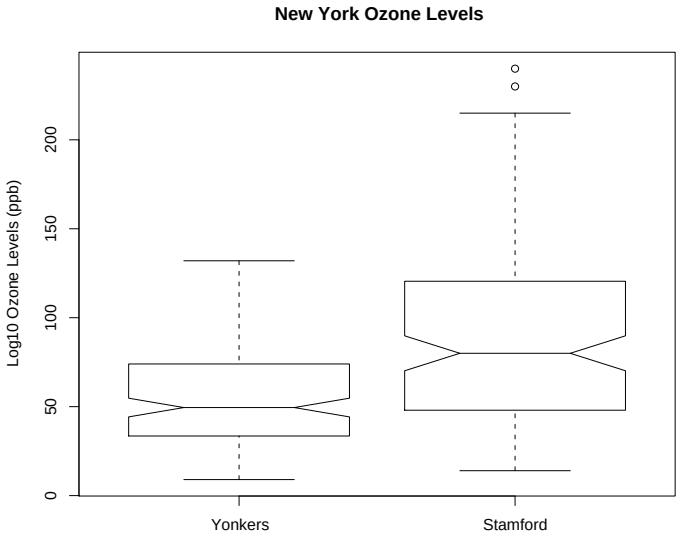
```
> boxplot(log10(yonkers), log10(stamford),
          names = c("Yonkers", "Stamford"),
          main = "New York Ozone Levels",
          ylab = "Log10 Ozone Levels (ppb)")
```


New York Ozone Levels
Log10 Ozone Levels (ppb)
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
Yonkers
Stamford

Significance Testing

- The boxplot function can be used to carry out a formal significance test of whether there is difference between the median levels of the underlying populations.
- This is done by specifying `notch=TRUE` as an argument.
- If the resulting notches in the sides of the boxplots do not overlap then there is a significant difference between the mediands of the underlying values.

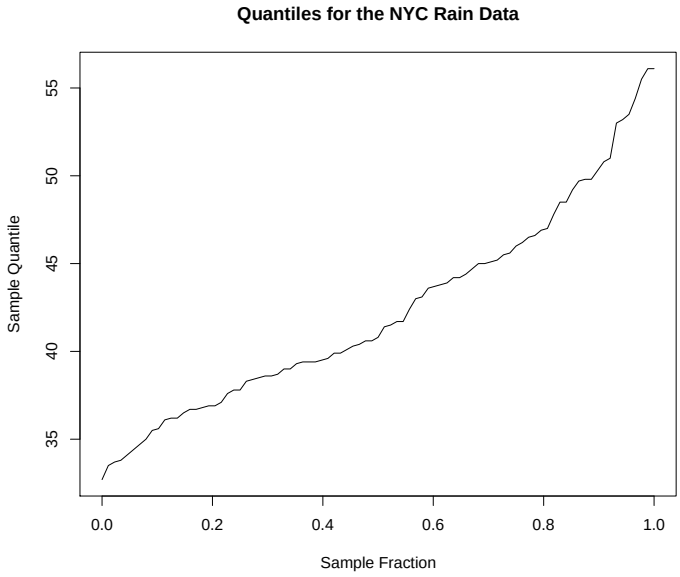
```
> boxplot(yonkers, stamford,
          names = c("Yonkers", "Stamford"),
          main = "New York Ozone Levels",
          ylab = "Log10 Ozone Levels (ppb)",
          notch = TRUE)
```


New York Ozone Levels
Log10 Ozone Levels (ppb)
0
50
100
150
200
Yonkers
Stamford

Quantile Plots

- Quantile plots directly display the quantiles of a set of values.
- The sample quantiles are plotted against the fraction of the sample they correspond to.
- There is no built-in quantile plot in R, but it is relatively simple to produce one.

```
> x = rain.nyc
> n = length(x)
> plot((1:n - 1)/(n - 1), sort(x), type="l",
       main = "Quantiles for the NYC Rain Data",
       xlab = "Sample Fraction",
       ylab = "Sample Quantile")
```


Quantiles for the NYC Rain Data
Sample Quantile
35
40
45
50
55
0.0
0.2
0.4
0.6
0.8
1.0
Sample Fraction

Quantile-Quantile Plots

- Quantile-quantile plots allow us to compare the quantiles of two sets of numbers.
- This kind of comparison is much more detailed than a simple comparison of means or medians.
- There is a cost associated with this extra detail. We need more observations than for simple comparisons.

Quantile-Quantile Plots, How to...

- Obtain the sample fraction values for the larger batch of values.
- For each batch, compute the quantiles corresponding to the computed fractions.
- Plot the computed quantiles against each other.

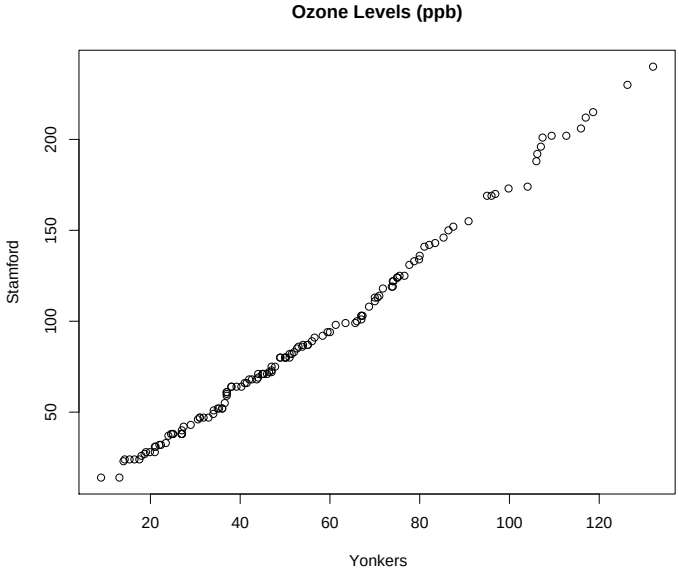
```
> n = max(length(x), length(y)) > p =
(1:n - 1)/(n - 1) > qx = quantile(x, p) > qy = quantile(
p) > plot(qx, qy)
```

But there is an easier way . . .

The R Quantile-Quantile Plot Function

- Q-Q plots are an important tool in statistics and there is an R function which implements them.
- The function is called `qqplot`.
- The first two arguments to qqplot are the samples of values to be compared.

```
> qqplot(yonkers, stamford,
         xlab = "Yonkers",
         ylab = "Stamford",
         main = "Ozone Levels (ppb)")
```


Ozone Levels (ppb)
Stamford
Yonkers
50
100
150
200
20
40
60
80
100
120

Modelling The Relationship

The points in the plot fall close to a straight line. This suggests that the the quantiles of the two samples satisfy:

$$\text{stamford} = a + b \times \text{yonkers}$$

or

$$\text{stamford} = b \times \text{yonkers}$$

One way to check which of these situations applies is to add some straight lines to the plot.

We also need to expand the limits on the graph, because we want to check whether the points lie on a line through and origin.

An Improved Plot

First we get the ranges of the two samples. There are NA values present and we have to make sure to omit these.

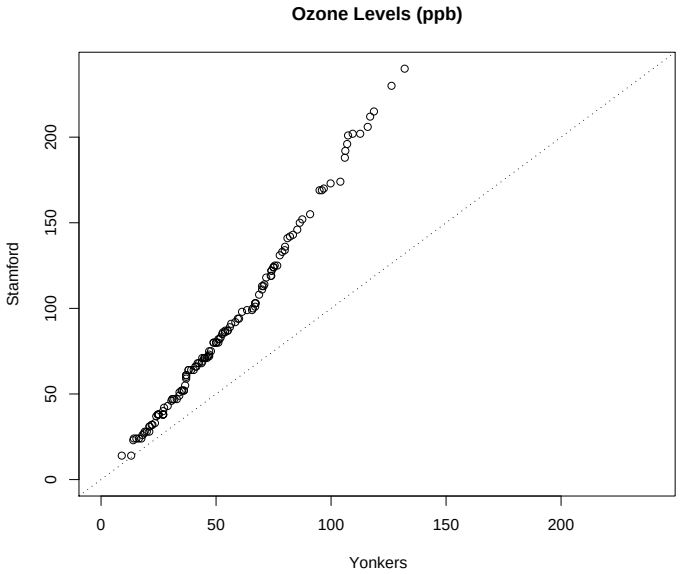
```
> xlim = range(0, yonkers, na.rm = TRUE)
> ylim = range(0, stamford, na.rm = TRUE)
```

Now we produce a Q-Q plot with the expanded limits.

```
> qqplot(yonkers, stamford,
         xlim = ylim, ylim = ylim,
         xlab = "Yonkers",
         ylab = "Stamford",
         main = "Ozone Levels (ppb)")
```

And mark in the line $y = x$.

```
> abline(a=0, b=1, lty="dotted")
```


Ozone Levels (ppb)
Stamford
Yonkers
0
50
100
150
200

A Multiplicative Relationship

The model "stamford $= b \times$ yonkers" seems to explain the relationship. What is the value of b?

We can find out by averaging the ratios of the quantiles.

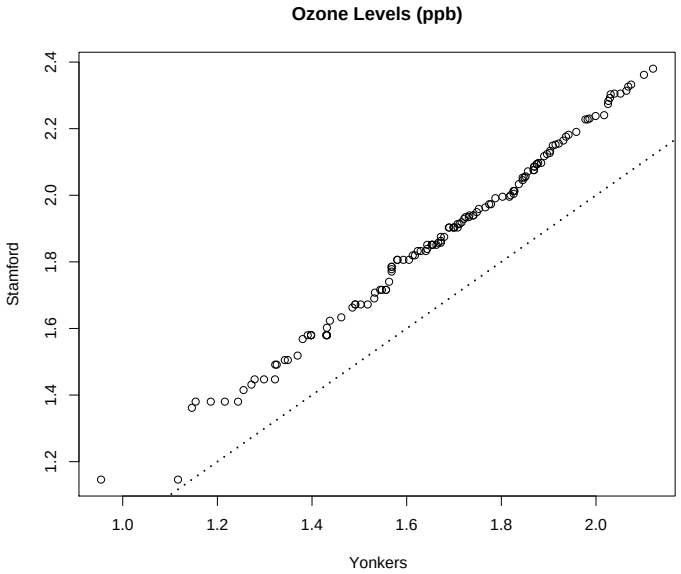
```
> y = na.omit(yonkers)
> s = na.omit(stamford)
> n = max(length(y), length(s))
> mean(quantile(s, seq(0, 1, length=n))/
       quantile(y, seq(0, 1, length=n)))
[1] 1.589835
```

The values in Stamford are about 1.6 times those in Yonkers.

An Improved Plot

A alternative way to proceed here is to work with the logs of the data values. This has the advantage of turning multiplication into addition.

```
> qqplot(log10(yonkers), log10(stamford),
         xlab = "Yonkers",
         ylab = "Stamford",
         main = "Ozone Levels (ppb)")
> abline(0, 1, lty = "dotted", lwd = 2)
```


Ozone Levels (ppb)
Stamford
Yonkers
1.2
1.4
1.6
1.8
2.0
2.2
2.4
1.0
1.2
1.4
1.6
1.8
2.0

Investigating Further . . .

The distance between the two lines is about .2. Since

$$10^{0.2} = 1.584893$$

we come to same conclusion this way.