

## Statistics 120 Plots Based on Quantiles II

### Comparison Using Densities

```
> ctrl = c(41.0, 38.4, 24.4, 25.9, 21.9, 18.3,
           13.1, 27.3, 28.5, -16.9, 26.0, 17.4,
           21.8, 15.4, 27.4, 19.2, 22.4, 17.7,
           26.0, 29.4, 21.4, 26.6, 22.7)

> ozone = c(10.1, 6.1, 20.4, 7.3, 14.3, 15.5,
            -9.9, 6.8, 28.2, 17.9, -9.0, -12.9,
            14.0, 6.6, 12.1, 15.7, 39.9, -15.9,
            54.6, -14.7, 44.1, -9.0)

> dens = dtrace(list(Control = ctrl, Ozone = ozone))

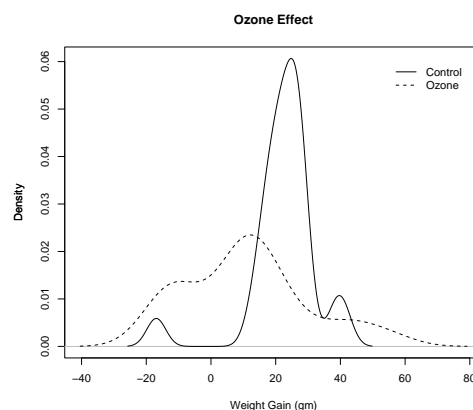
> plot(dens, main = "Ozone Effect",
       xlab = "Weight Gain (gm)")
```

### An Example – Rats and Ozone

A group of young rats was randomly split into two groups. One group was used as a control and the other raised in an ozone enriched environment

The following weight gains were observed:

Control	41.0	38.4	24.4	25.9	21.9
	18.3	13.1	27.3	28.5	-16.9
	26.0	17.4	21.8	15.4	27.4
	19.2	22.4	17.7	26.0	29.4
	21.4	26.6	22.7		
Ozone	10.1	6.1	20.4	7.3	14.3
	15.5	-9.9	6.8	28.2	17.9
	-9.0	-12.9	14.0	6.6	12.1
	15.7	39.9	-15.9	54.6	-14.7
	44.1	-9.0			



### A “Standard” Analysis

- A standard analysis would use a two-sample  $t$ -test to see whether ozone exposure has a significant effect on weight gain.
- The mean weight gains were:
 

<i>Control</i>	22.4
<i>Ozone</i>	11.0
- The  $p$ -value for a two-sided test is 0.02.
- This is weak evidence that ozone exposure decreases the growth rates of juvenile rats.

### What the Plot Shows

- The distributions of weight gains for the two groups are very different.
- The peak of the “ozone” group is shifted to the left relative to the control group.
- The ozone group is more spread out than the control.
- There is an isolated small peak in the control group to the left of zero.

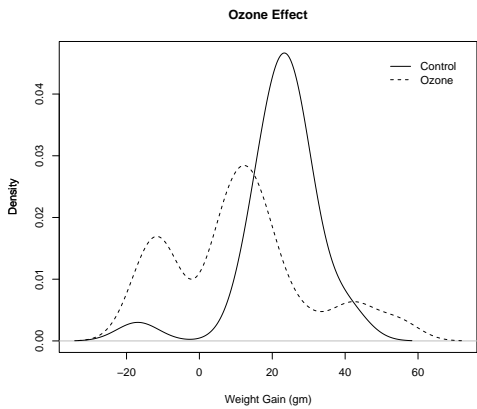
### A “Graphical” Analysis

- A  $t$ -test showed a difference in average weight gain, but there is rather more going on here.
- We can see this by comparing the full distribution of the values, rather than just the means.
- We have several ways of doing this:
  - Stem-and-Leaf plots
  - Histograms
  - Density Plots
  - Quantile-Quantile Plots

### Using Equal Bandwidths

- The default bandwidths used in producing the densities in the plot were quite different.
- A value of close to 10 was use for the control group and a value of close to 28 for the ozone group.
- As a compromise we can try using 20 for both groups to make the results directly comparable.
 

```
> dens = dtrace(list(Control = ctrl,
                     Ozone = ozone),
                 bw = 20)
```



### What the Plot Shows

- In the lower tails of the weight-gain distributions, the gains for the ozone group tend to be lower than those of the control group.
- The lowest weight gain values are negative.
- In the centre of the weight-gain distributions the weight gains for the ozone group are positive, but not as big as those of the control group.
- In the top tails of the weight-gain distributions, the gains for the ozone group are greater than those for the control group.

### What the Plot Shows

- The control group forms a single group (with a single outlier).
- There is some evidence that the ozone group consists of three clusters of rats.
- Some rats in the top cluster of the ozone group appear to have greater weight gains than any of the control group rats.

### Interpretation

- The results seem to suggest that most rats are harmed by ozone exposure and that some benefit.
- The effect is probably a result of the way the experiment was run.
- The rats in each group were housed together.
- The ozone probably had a detrimental effect on all the rats, but those most effected were put off their food (hence the weight loss).
- This left a surplus of food for the least affected rats and so they were able to put on a lot of weight.

### Comparison Using Quantile Quantile Plots

- Because the relationship between the two set of weight gains is complex, it is useful to produce a Q-Q plot to get more detail on how the groups line up.
- Producing the plot is easy.

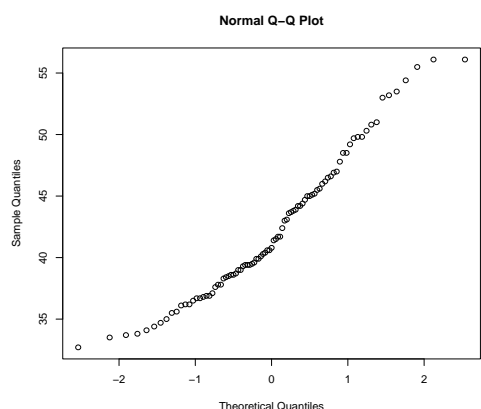
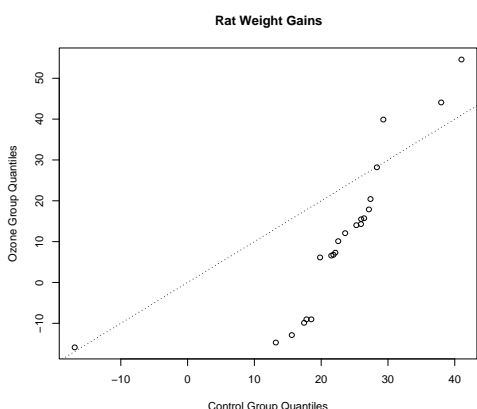
```
> qqplot(ctrl, ozone,
  main = "Rat Weight Gains",
  xlab = "Control Group Quantiles",
  ylab = "Ozone Group Quantiles")
> abline(0, 1, lty="dotted")
```

### Theoretical Quantile Quantile Plots

- Quantile-quantile plots can be used to compare the distributions of two sets of numbers.
- They can also be used to compare the distributions of one set of values with some theoretical distribution.
- Most commonly, the yardstick distribution is the standard normal distribution:

$$P[X \leq x] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$$

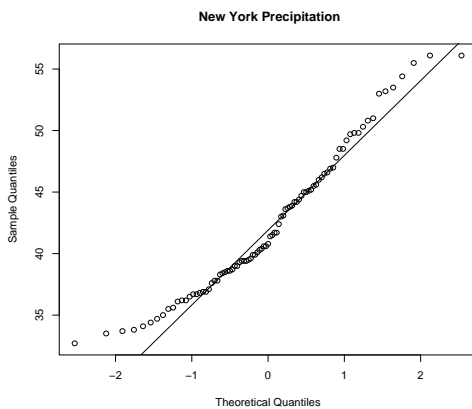
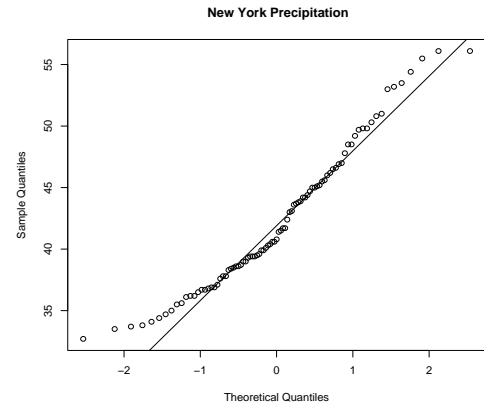
- If the values being plotted resemble a sample from a normal distribution, they will lie on a straight line with intercept equal to the mean of the values and slope equal to the standard deviation.



## R Functions

- The function `qqnorm` produces a basic Q-Q plot comparing a set of values with the normal distribution.
- The function `qqline` adds a straight line to the plot. The line passes through the point defined by the lower quartiles and the point defined by the upper quartiles.

```
> qqnorm(rain.nyc,
         main = "New York Precipitation")
> qqline(rain.nyc)
```



## Distribution Symmetry

- Suppose we have a collection of values  $x_1, \dots, x_n$ . We will say that the values are symmetrically distributed if their quantile function satisfies:

$$Q(.5) - Q(p) = Q(1-p) - Q(0.5), \quad \text{for } 0 < p < .5.$$

- This says that the  $p$ th quantile is the same distance below the median as the  $(1-p)$ th quantile is above it.
- When a set of values is "close" to normally distributed, a normal Q-Q plot can help to detect departures from symmetry,

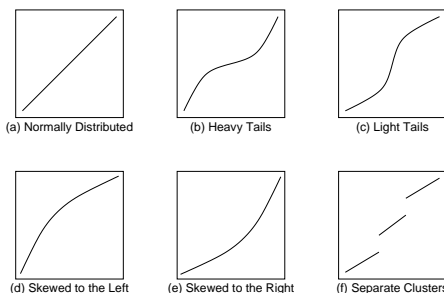
## Deviations From Normality

- The NYC rainfall plot shows a systematic deviation from normality.
- Detecting such deviations is important because many statistical techniques depend on the data they are applied to having an approximately normal distribution.
- Note: The importance of normality is often overstated in elementary statistics courses. The NYC rainfall would be fine to use for most normally based statistical techniques.

## A Symmetry Plot

- The obvious way to check the symmetry of a set of numbers is to plot the values  $Q(1-p_1), \dots, Q(1-p_{n/2})$  against the values of  $Q(p_1), \dots, Q(p_{n/2})$ .
- If the plotted points fall on the line  $y = x$ , then  $x_1, \dots, x_n$  are symmetrically distributed.
- There is no built-in R function which produces symmetry plots, but it is very easy to create such a plot.

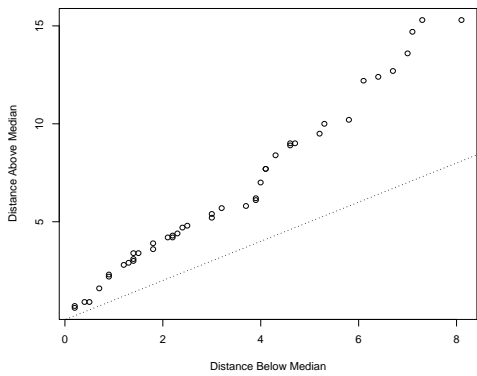
## Some Departures from Normality



## R Code

```
> symplot =
function(x)
{
  n = length(x)
  n2 = n %% 2
  sx = sort(x)
  mx = median(x)
  plot(mx - sx[1:n2], rev(sx)[1:n2] - mx,
       xlab = "Distance Below Median",
       ylab = "Distance Above Median")
  abline(a = 0, b = 1, lty = "dotted")
}

> symplot(rain.nyc)
```



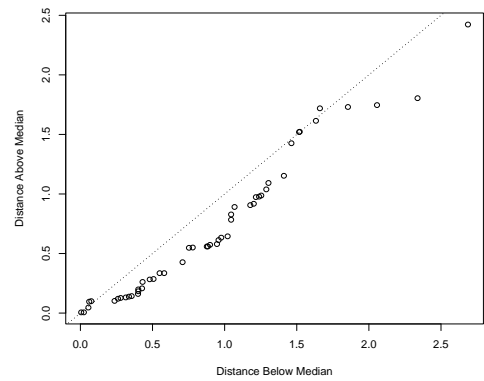
## Sample Size Considerations

- Both normal Q-Q plots and symmetry plots require large sample sizes to reliably represent the population being sampled.
- This is especially true for symmetry plots.
- Sample sizes of at least 1000 are desirable, although the plots do tend to get used on much smaller sample sizes.
- Running the command below repeatedly can show just how unstable the plots are with smaller sample sizes.

```
> symplot(rnorm(100))
```

## Transforming to Symmetry

- There appears to be evidence of lack of symmetry in the symmetry plot.
- The upper quantiles of the distribution are further from the median than the corresponding lower quartiles.
- This indicates that the distribution of values is skewed to the right.
- It can sometimes be useful to transform skewed distributions to more symmetric ones. Transformations which can be used to do this are: square roots, cube and other roots, logarithms and reciprocals.



## Transforming to Symmetry

- In the case of the rainfall data, it is hard to find a transformation which makes the distribution more symmetric.
- This is because of the internal clustering present in the values.
- Negative reciprocals do a fairly good job.

```
> symplot(-1/rain.nyc)
```

