

# Chapter 6: RANDOM SAMPLING AND DATA DESCRIPTION

## Part 2: Comparative Box Plots

Time Sequence Plots

Probability Plots

Normal Probability Plots

Sections 6-4 to 6-6

Another use of boxplots...

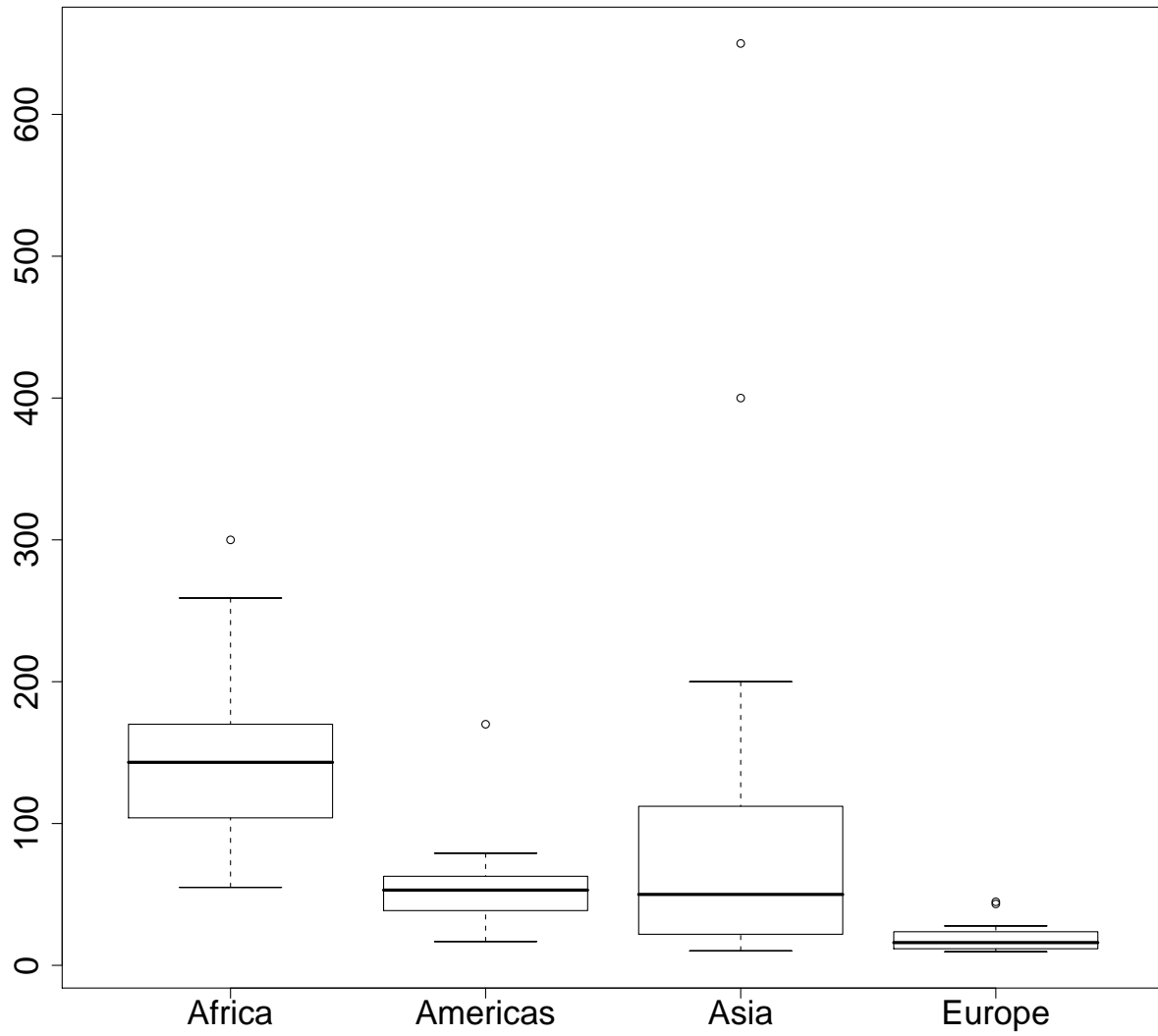
## Comparative boxplots

Data on Infant-Mortality. The observations are nations of the world around 1970

Variable:

*infant* Infant-mortality rate per 1000 live births.

*region* Levels: Africa, Americas, Asia,  
Europe.



Compare spreads.  
Compare centers.

# Time Sequence Plots

## Section 6-5

When data is collected over time, it can be informative to plot the data in sequence.

Time sequence plot can show trends and cycles.

The compressive strength data we previously looked at has a time component to it...

Consider the following set of  $n = 80$  data points we saw earlier.

105	97	245	163	207	134	218	199	160	196
221	154	228	131	180	178	157	151	175	201
183	153	174	154	190	76	101	142	149	200
186	174	199	115	193	167	171	163	87	176
121	120	181	160	194	184	165	145	160	150
181	168	158	208	133	135	172	171	237	170
180	167	176	158	156	229	158	148	150	118
143	141	110	133	123	146	169	158	135	149

We didn't consider the time component previously, but we can look at it as time sequence plot...

## Compressive strengths with time component included

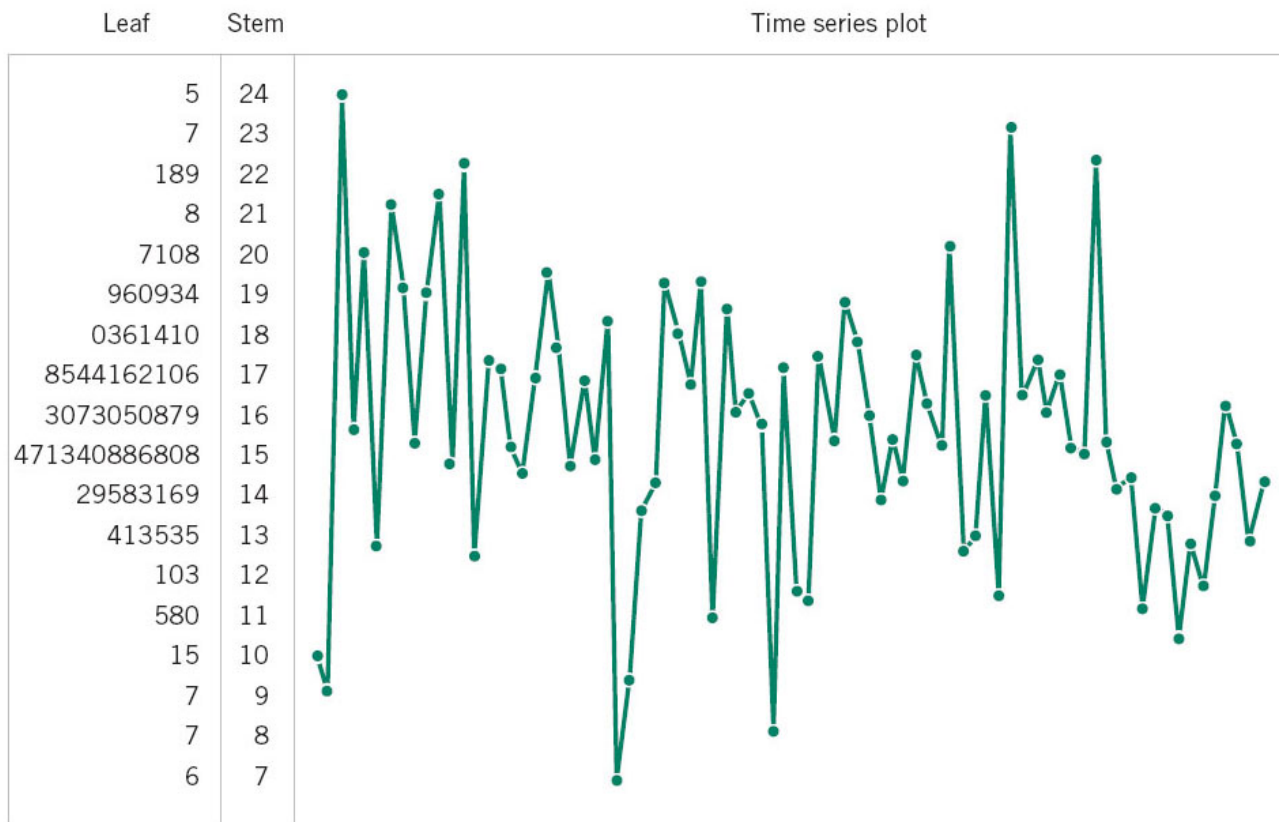
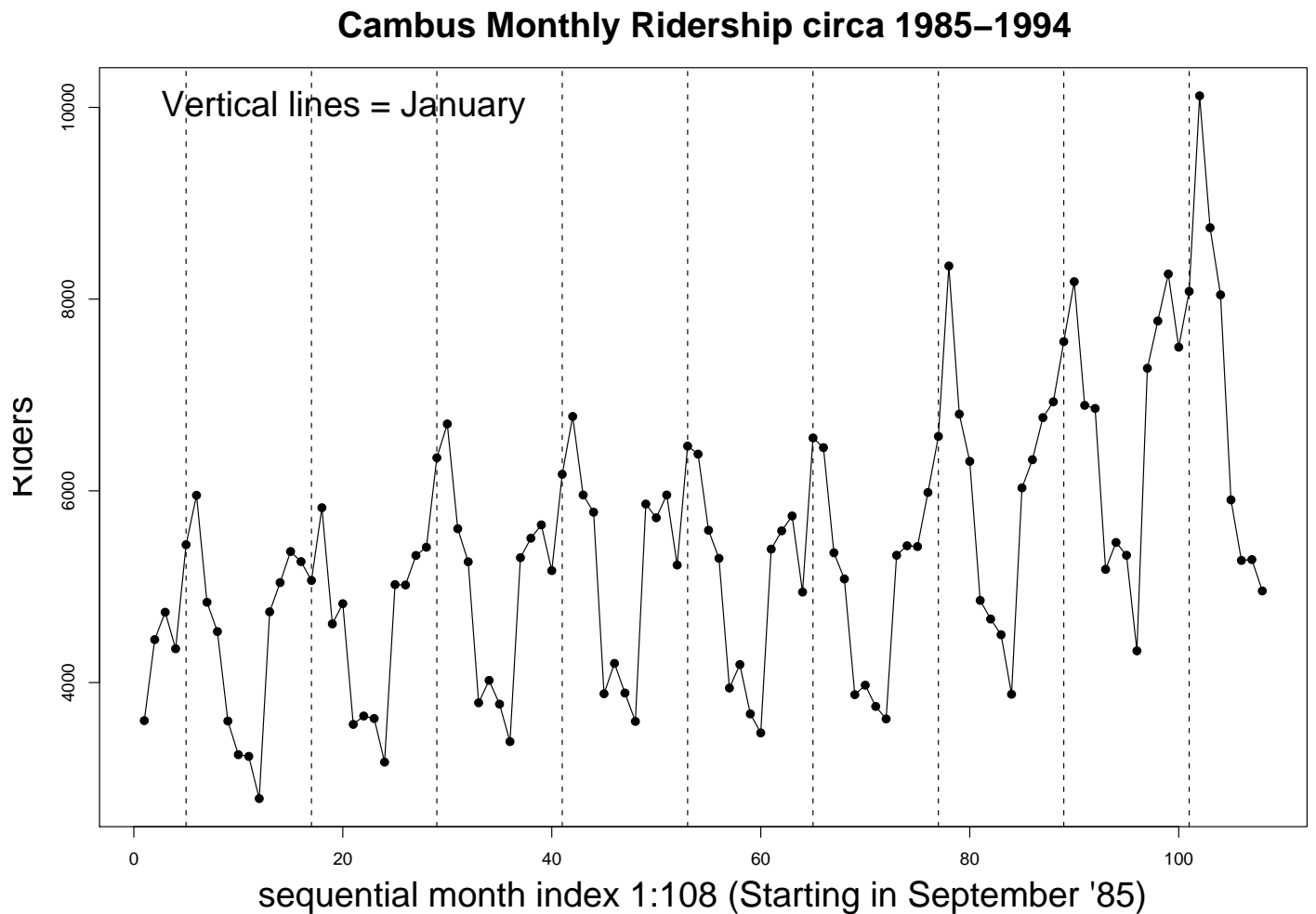


Figure 6-17 A digidot plot of the compressive strength data in Table 6-2.

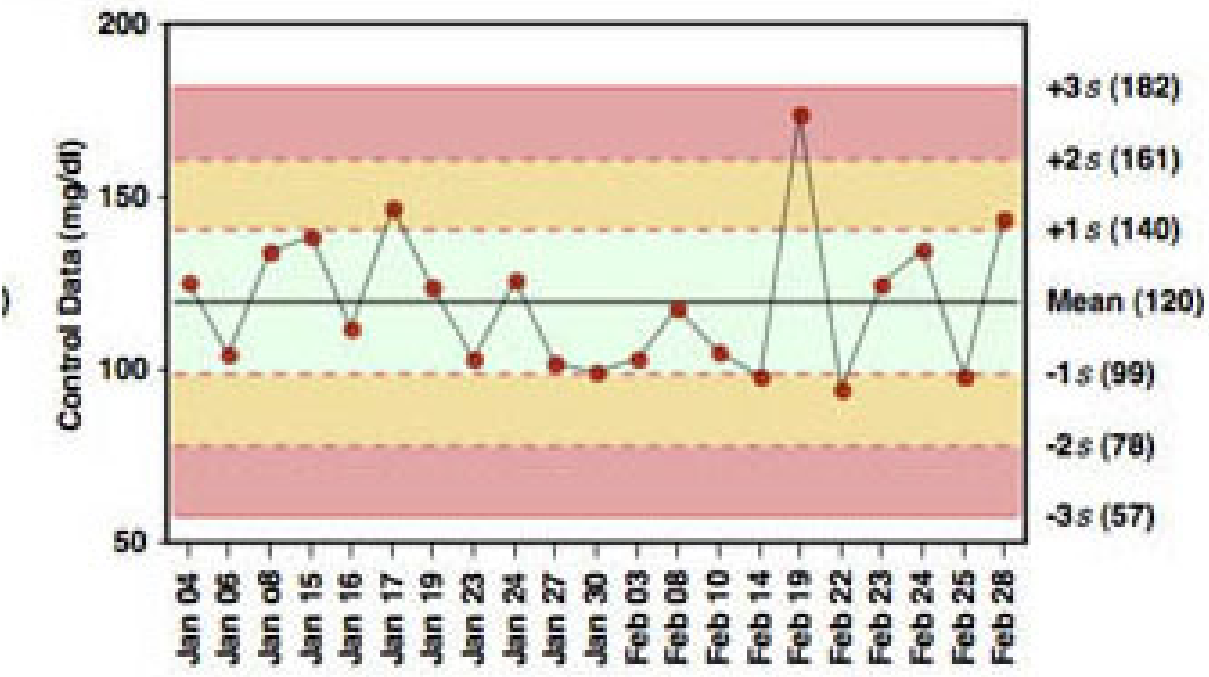
Time sequence plots can help you recognize trends.



Look at the popularity (i.e. rank) of your name as years go by.

<http://www.babynamewizard.com/namevoyager/lnv0105.html>

# Quality control charts



To improve productivity.

To prevent defects.

To provide information about process.

# Probability Plots

## Section 6-6

Let's return to the stem-n-leaf diagram for the compressive strength data.

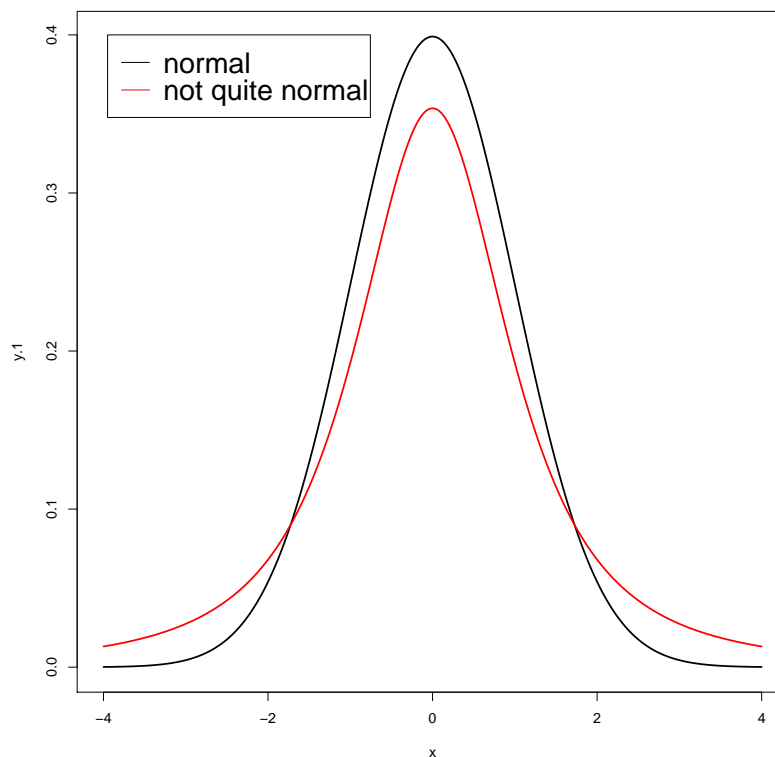
The decimal point is 1 digit(s)

7		6	to the right of the
8		7	
9		7	
10		15	
11		058	
12		013	
13		133455	
14		12356899	
15		001344678888	
16		0003357789	
17		0112445668	
18		0011346	
19		034699	
20		0178	
21		8	
22		189	
23		7	
24		5	

It looks 'normally' distributed, but is it?

Having the correct general shape is a start, but there are specific probabilities that coincide with the normal distribution.

For example...



For the red probability distribution, less than 95% is between -2 and 2 because there is more 'left' in the tails than in the normal distribution.

Scaling the distribution won't get you the normal distribution.

The previous example shows a distribution that is *nearly* normal, which will often be close enough to the normal for our specific needs.

But, in general, we we want to be able to detect non-normality, or when a distribution is not normal.

We can use a **Normal Probability Plot** for this goal.

Suppose we have  $n = 10$  data points and we want to know if they were likely to have been generated from a normal distribution.

0.693	-1.236	-1.141	1.366	-0.756
-1.321	0.665	-0.627	1.635	-1.242

We can compare these 10 values to 10 ‘likely’ values from a known normal distribution.

What are the 10 ‘likely’ values from a normal?

We use the *percentiles* of the standard normal to get these ‘likely’ values.

Take one from...

the bottom 10% of the distribution,

the next 10% of the distribution,

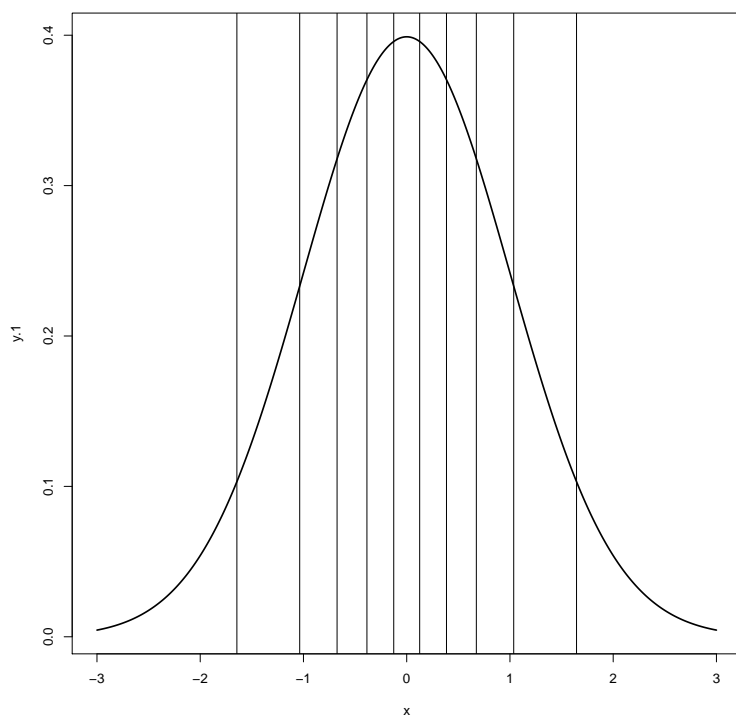
...

the 10% of the distribution below the center,

the 10% of the distribution above the center,

...

the top 10% of the distribution.

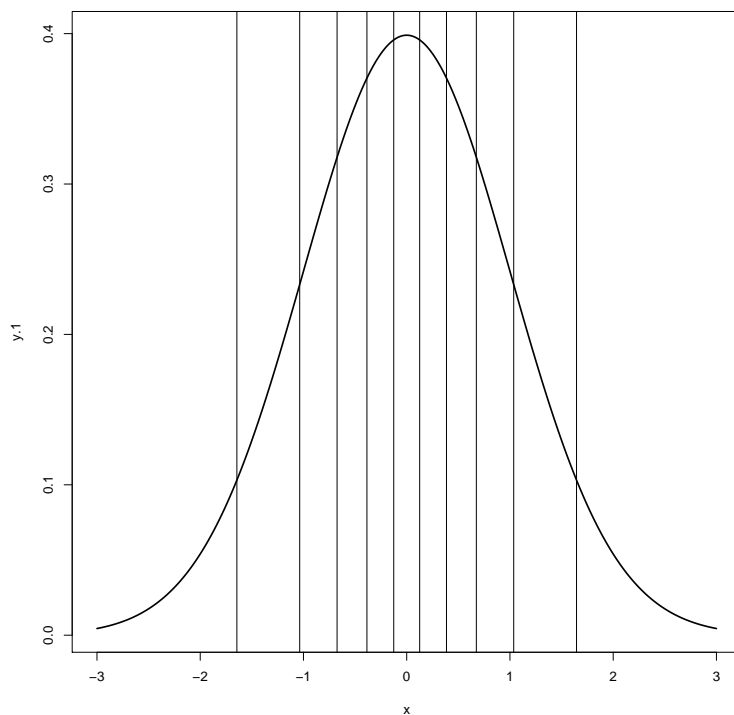


We split the difference for picking a point in the lowest 10% and find the z-value at which 5% of the probability is below it...

$$p(Z < -1.64) = 0.05$$

-1.64 is our representative value from the bottom 10%.

The other 9 are found similarly.



percentile	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
z-value	-1.64	-1.04	-0.67	-0.39	-0.13	0.13	0.39	0.67	1.04	1.64

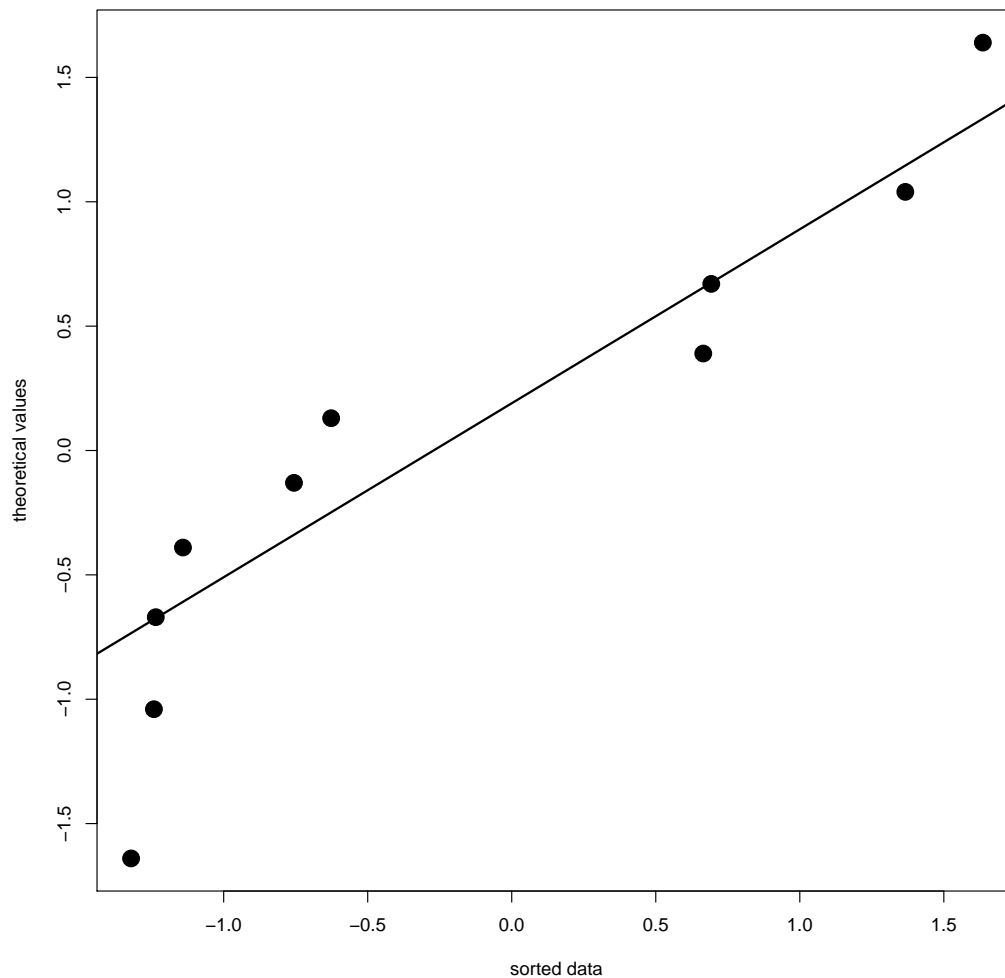
Notice that we get more representative points from around zero than in the tails (the vertical lines above are the most dense near zero).

Compare our data points (sorted) to the above:

our data	-1.32	-1.24	-1.24	-1.14	-0.76	-0.63	0.66	0.69	1.37	1.64
----------	-------	-------	-------	-------	-------	-------	------	------	------	------

They're easier to compare if we plot them against each other...

## Normal Probability Plot



If our data points were likely to have been generated from a normal, the plotted points will fall approximately along a straight line.

We often plot a line that connects the 25<sup>th</sup> and the 75<sup>th</sup> percentile points for ease of judgement.

NOTE:\* These 10 data points were randomly generated from a normal, though they don't fall *exactly* along the line.

In general, if you have  $n$  data points, you want to choose a representative value in every consecutive  $\frac{1}{n}$  probability interval (but we'll split the difference and pick the midpoint of the interval).

1) Sort your data points and denote them as

$$x_{(1)}, x_{(2)}, \dots, x_{(n)}.$$

$x_{(1)}$  is the minimum value in the data set.

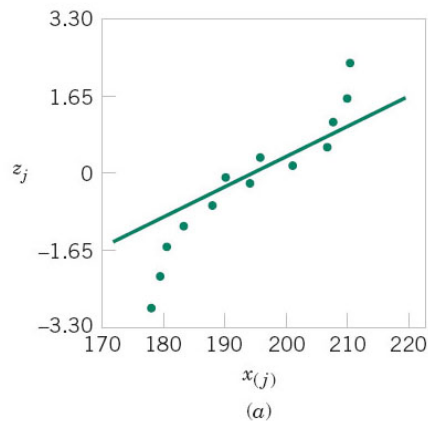
$x_{(n)}$  is the maximum value in the data set.

2) For each  $x_{(j)}$  get the associated representative value as  $z_j$  where

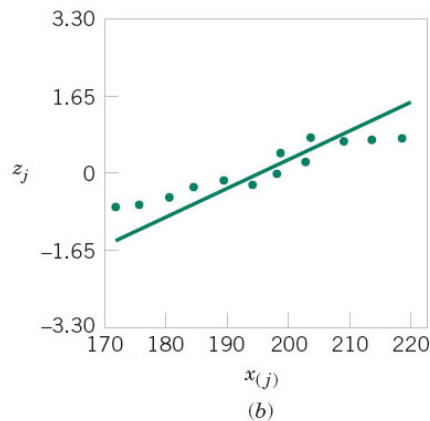
$$\frac{j - 0.5}{n} = P(Z \leq z_j) = \Phi(z_j)$$

3) Plot all  $n$  points as  $(x_{(j)}, z_j)$

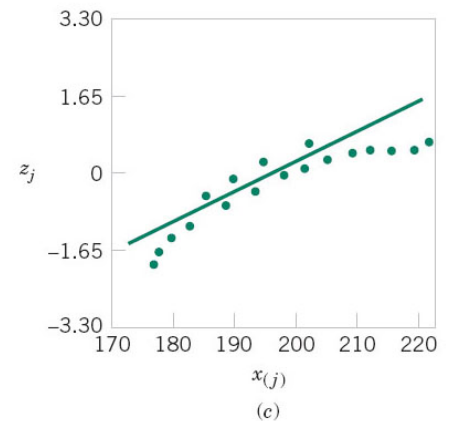
Things to look for in your normal probability plot...



Light-tails  
compared to  
normal



Heavy tails  
compared to  
normal



Right - skew

All these are signs of non-normality.

---

A few notes...

There are other ways of getting the 'representative' theoretical value, but this is the one we'll use.

You can also do a probability plot on probability paper which has special axes, but we'll use the regular axes and plot  $(x_{(j)}, z_j)$  as in these examples.

Other books (and software) often show this plot with  $x_{(j)}$  on the vertical axis, and  $z_j$  on the horizontal axis (a reflection over  $y=x$  line). The general concept holds, but the descriptions at the top of this page will switch.

## Other distributions of interest

You can use the same tactic to compare your data to any hypothetical distribution.

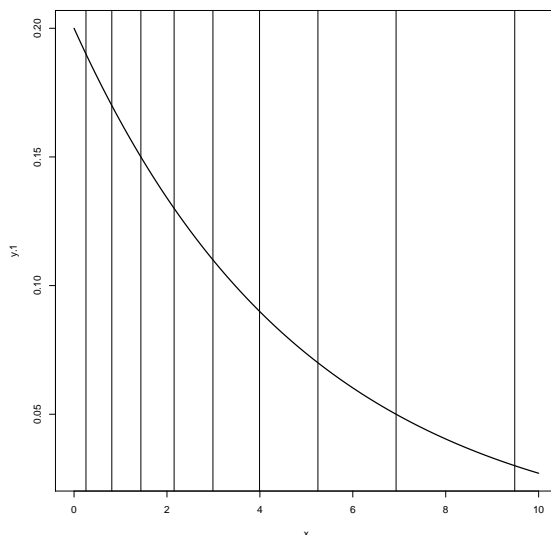
If you're interested in seeing if your data came from an exponential distribution with rate parameter  $\lambda = 0.2$ .

Observed values:

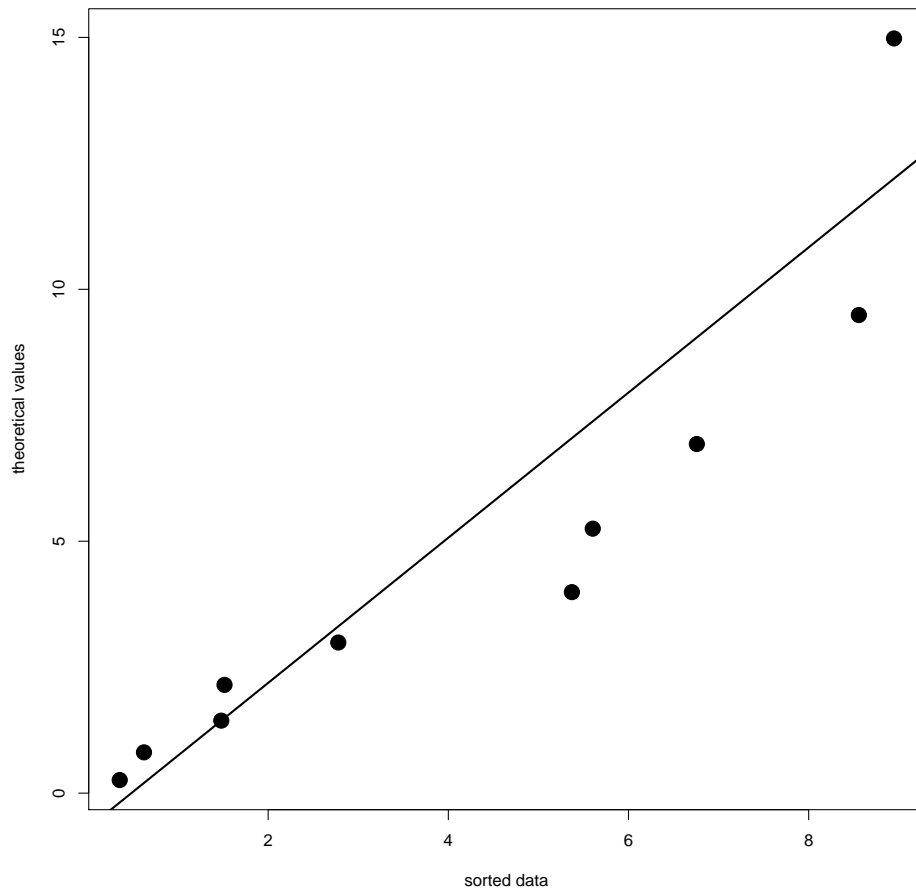
0.352 0.622 1.480 1.516 2.779 5.372 5.603 6.759  
8.557 8.947

Representative theoretical values:

0.26 0.81 1.44 2.15 2.99 3.99 5.25 6.93 9.49 14.98



The probability plot:



Again, if the data matches the theoretical distribution, then the points will fall approximately on a line.

(These data points were generated from an exponential with  $\lambda = 0.2$ .)