

## Topic 14 – Sample Quantiles and Box Plots

### Statistics for Managers

June 3, 1999

We shall start with a data set.

### Example 1. Service Times Data and Histogram

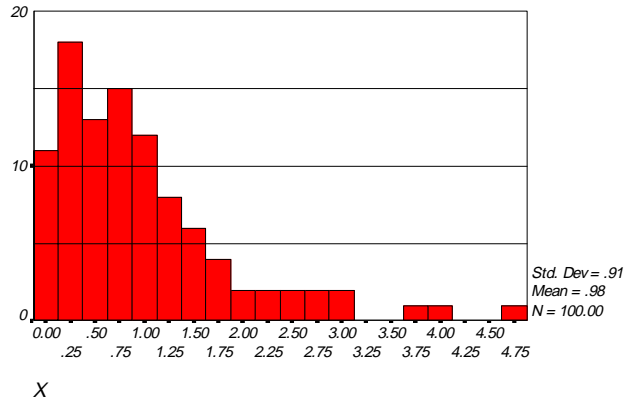
- Table 1 Service Times – Ascending Order n = 100 (minutes)

0.01	0.12	0.23	0.39	0.59	0.77	1.02	1.21	1.43	2.33
0.04	0.13	0.26	0.41	0.59	0.77	1.03	1.21	1.45	2.54
0.04	0.14	0.28	0.43	0.66	0.82	1.03	1.23	1.48	2.59
0.05	0.15	0.32	0.46	0.66	0.82	1.07	1.25	1.63	2.66
0.05	0.19	0.32	0.46	0.71	0.84	1.08	1.29	1.67	2.75
0.07	0.2	0.33	0.47	0.71	0.86	1.08	1.3	1.81	2.94
0.07	0.2	0.33	0.49	0.73	0.87	1.1	1.36	1.85	2.96
0.1	0.21	0.34	0.5	0.74	0.92	1.11	1.38	1.92	3.74
0.1	0.21	0.34	0.52	0.75	0.92	1.11	1.39	2.06	4.05
0.11	0.23	0.38	0.52	0.75	0.97	1.19	1.41	2.14	4.77

- Table 2 Summary Statistics

Statistic	Value
min	0.01
max	4.77
range	4.76
range/20	0.238

Since the data set is large,  $n = 100$ , 20 intervals may make a good-looking histogram. Class interval widths of 0.25 minutes will be easy to work with.



• Figure 1 Sample Histogram of Service Times

This sample histogram indicates that the normal probability-mass function would not be a very good model for the population of service times.

## Quantile

A fractile, or quantile, is a value at or below which a given fraction of the data must lie.

## Quantile

Quartiles are  $\frac{1}{4}$ ,  $\frac{1}{2}$ , and  $\frac{3}{4}$  quantiles. In other words, one-quarter of the data are less than or equal to the first-quartile; and, three-quarters of the data are less than or equal to the third-quartile

## Median

The median is the  $\frac{1}{2}$  quantile. In other words, half of the data are less than or equal to the median.

## Percentile

**Percentiles are 1/100, 2/100, 3/100, etc. quantiles. In most instances, the term “percentile” is substituted for “quantile.”**

- Table 3 Percentiles of Service Times

Fraction of Data	Percentile
0.00%	0.01
2.50%	0.04
25.00%	0.33
50.00%	0.76
75.00%	1.29
97.50%	3.37
100.00%	4.77

- Table 4 Excel Percentile Function

Fraction	Percentile Function
p	PERCENTILE(DataRange, p)
0.25	PERCENTILE(DataRange, 0.25)

## Empirical Cumulative Distribution Function and Ogive

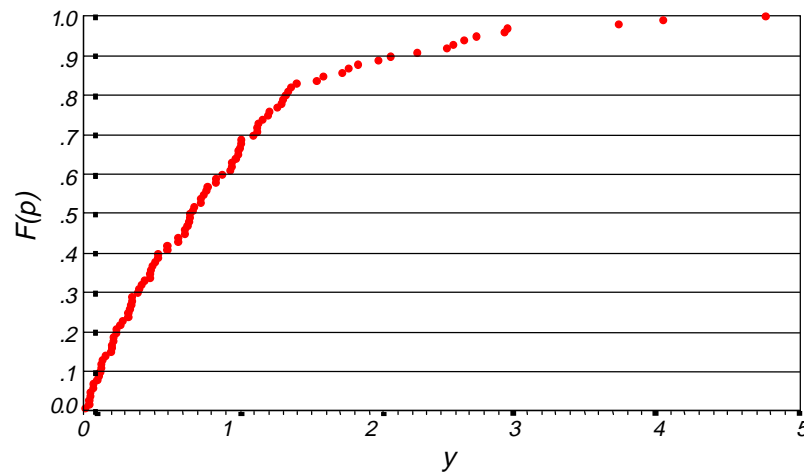
The empirical cumulative distribution function  $\hat{F}(y)$  is the fraction of data less than or equal to  $y$ :

$$\hat{F}(y) = \text{Fraction of data } \leq y$$

In other words, if  $y_p$  is the  $p^{\text{th}}$  quantile, then

$$\hat{F}(y_p) = p$$

The graph of a cumulative distribution function is called the ogive.



• Figure 2 Ogive of Service Time Data

## Normal Probability Plot

A normal probability plot is a plot of the  $i^{\text{th}}$  data point with rank  $R_i$  versus the quantile of the standard normal distribution corresponding to the fraction

$$p = \frac{R_i - 3/8}{n + 1/4}$$

The special formula for the fraction is needed for accuracy in tails of the normal distribution. A normal probability plot is also called a “Q-Q Plot”, since it is plotting Quantile versus Quantile.

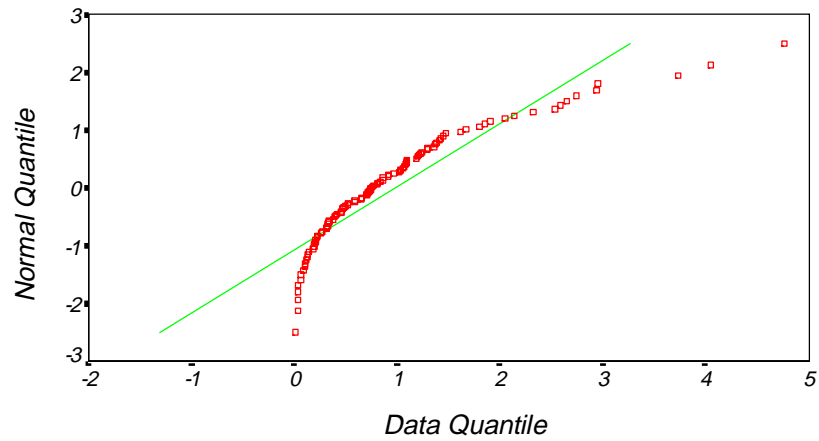
### Example 1. Normal Probability Plot

- Table 5 First 19 Data and Normal Quantile for Service Time Data

Obs.	R	p	Normal Quantile
0.01	1	0.01	-2.50
0.04	2	0.02	-2.14
0.04	3	0.03	-1.94
0.05	4	0.04	-1.80
0.05	5	0.05	-1.68
0.07	6	0.06	-1.59
0.07	7	0.07	-1.51
0.10	8	0.08	-1.43
0.10	9	0.09	-1.37
0.11	10	0.10	-1.30
0.12	11	0.11	-1.25
0.13	12	0.12	-1.20
0.14	13	0.13	-1.15
0.15	14	0.14	-1.10
0.19	15	0.15	-1.05
0.20	16	0.16	-1.01
0.20	17	0.17	-0.97
0.21	18	0.18	-0.93
0.21	19	0.19	-0.89

- Table 6 Excel Normal Quantile Function

Fraction	Normal Quantile
p	NORMINV(p,0,1)
0.95	1.65



- Figure 3 Normal Probability Plot of Service Time Data

**If the data were normal, we would expect to see the data points fall along the straight line.**

**Depending on the particular statistical software, the plot may have the data on the X-axis or the Y-axis.**

## Box-Whisker Plot

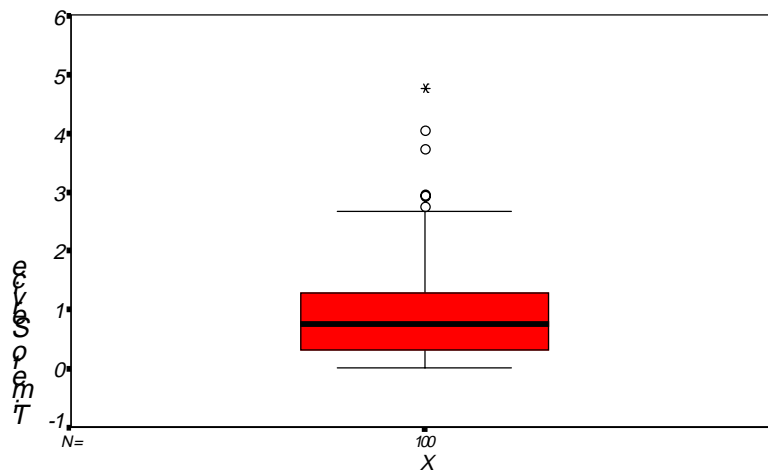
A box-whisker plot is a special graph of the 0.025<sup>th</sup> percentile, first quartile, median, third quartile, and 0.975<sup>th</sup> percentile.

The lowest and highest horizontal lines are the “whiskers” (see Figure), and denote the y-axis positions of the 2.5<sup>th</sup> and 97.5<sup>th</sup> percentiles, respectively. Therefore, the vertical interval spanned by the whiskers encompasses 95% of the data.

The lowest and highest horizontal edges of the box denote the y-axis positions of the first and third quartiles. Therefore, the vertical interval spanned by the box encompasses 50% of the data.

The heavy horizontal line inside the box denotes the y-axis position of the median of the data.

### Example 2. Box-Whisker Plot



• Figure 4 Box-Whisker Plot of Service Time Data

## Outlier and Extreme Data Points

A data point is an outlier if it is larger or equal to

Third-quartile +  $1.5 \times$  box-width;

or is smaller or equal to

First-quartile -  $1.5 \times$  box-width.

In addition, the data point is called an extreme point if it is larger or equal to

Third-quartile +  $3.0 \times$  box-width;

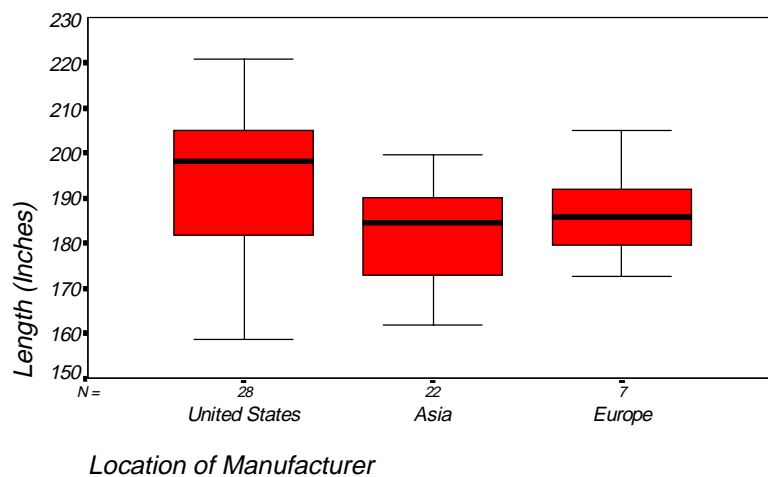
or is smaller or equal to

First-quartile -  $3.0 \times$  box-width.

In the box-whisker plot outliers are denote with an “O” symbol; an extreme point is denote with an “\*” symbol.

## Box-Whisker Plots of Grouped Data

Box-whisker plots are especially useful for comparing data over several groups. Using the 92 auto data, we see clear differences between car lengths for the different countries of origin.



- Figure 5 Box-Whisker Plots of Car Length for Different Counties of Origin