STAT 515

Statistical Methods I

Sections 2.6-2.7

Percentiles and Boxplots

Brian Habing
Department of Statistics
University of South Carolina

Outline

- Percentiles
- Five Number Summary
- Outliers

Percentiles

The *p*th percentile of a set of measurements has at least *p*% of the measurements at or below it and at least (100-*p*)% of the measurements at or above it.

Example

Consider the data set 2, 3, 3, 4, 8

Percentiles

SAS has five ways of calculating percentiles!

http://support.sas.com/documentation/cdl/en/procstat/63104/HTML/default/viewer.htm#procstat_univariate_sect028.htm

You can specify one of five definitions for computing the percentiles with the PCTLDEF= option. Let n be the number of nonmissing values for a variable, and let $x_1, x_2, \ldots, x_n$ represent the ordered values of the variable. Let the rth percentile be y, set $p = \frac{r}{100}$, and let

$$\begin{aligned} np &= j + g \quad \text{when PCTLDEF=1, 2, 3, or 5} \\ (n+1)p &= j + g \quad \text{when PCTLDEF=4} \end{aligned}$$

where j is the integer part of ***np***, and g is the fractional part of ***np***. Then the PCTLDEF= option defines the rth percentile, y, as described in the following table.

PCTLDEF	Description	Formula
1	weighted average at x_{np}	$y = (1-g)x_j + gx_{j+1}$ where x_0 is taken to be x_1
2	observation numbered closest to ***np***	$y = x_j$ if $g < \frac{1}{2}$ $y = x_j$ if $g = \frac{1}{2}$ and j is even $y = x_{j+1}$ if $g = \frac{1}{2}$ and j is odd $y = x_{j+1}$ if $g > \frac{1}{2}$
3	empirical distribution function	$y = x_j$ if $g = 0$ $y = x_{j+1}$ if $g > 0$
4	weighted average aimed at $x_{(n+1)p}$	$y = (1-g)x_j + gx_{j+1}$ where x_{n+1} is taken to be x_n
5	empirical distribution function with averaging	$y = \frac{1}{2}(x_j + x_{j+1})$ if $g = 0$ $y = x_{j+1}$ if $g > 0$

Quartiles

Consider the data set 2, 3, 3, 4, 8

Inter-quartile Range

IQR = 75th percentile – 25th percentile

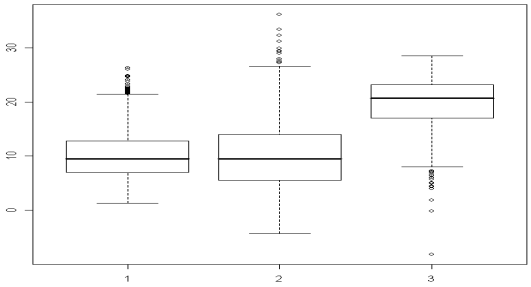
"Five number summary" is the minimum, 25th percentile, median, 75th percentile, and the maximum.

Which To Use?

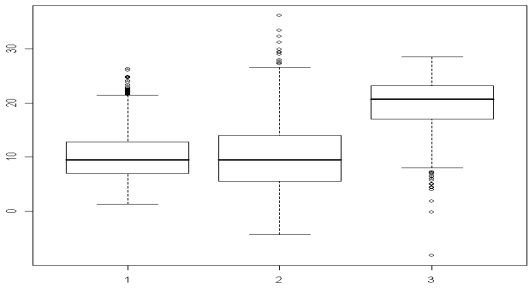
Data is approximately symmetric and unimodal – use the mean and standard deviation

Otherwise – use the five number summary

Box Plot



Box Plot



Example revisited

Consider the data set 2, 3, 3, 4, 8

Outliers

Outliers are values that are unusual in the context of the data set. If the data consists of one variable they are usually the values that are unusually large or unusually small.

Common explanations of outliers are:

1) Error in observing or recording the value
2) Comes from a different population
3) A rare event

Outliers

Outliers can only be removed in the case where it is clearly an error in observation or recording… not just because you think it was an error.

One option is to report the results both with and without the outlier(s).