

LI Analysis Training Series

Tukey Box Plots

(Last Revised: 4/23/2007 (6/24/1999))

Melissa Ives, Rodney Funk & Michael Dennis
 Chestnut Health Systems
 Bloomington IL 61701
 309-827-6026
www.chestnut.org

Acknowledgement: This document was developed under contract #270-2003-00006 from the Substance Abuse and Mental Health Services Administration (SAMHSA's) Center for Substance Abuse Treatment (CSAT). Any opinions about this data are those of the authors and do not represent official positions of the government or individual grantees.

Purpose: The primary purpose of Tukey Boxplots (also known as Box and Whisker plots) is to create a graphic image of the several key measures of distribution including the minimum, maximum, median and 25th and 75th percentile (the middle 50%). They are particularly useful in showing the distribution of a single variable across several groups or populations. Boxplots can be created in either SPSS or in Microsoft Excel. They are a simpler and more intuitive depiction of the information that may also be depicted in a histogram.

Procedure:

Tukey box plots are created using the 'EXAMINE' command with the PLOT=BOXPLOT subcommand. Selecting the menu item Graph → Boxplot will run the Examine command. Boxplots can also be created using percentiles and plugging the data into the excel file "Tukey boxplot template.xls" as noted below.

Example 1

```
EXPLORE
  VARIABLES=spsm_0 spsm_3 SPSm_dif /COMPARE VARIABLE/PLOT=BOXPLOT
  /STATISTICS=NONE/NOTOTAL/ID=XSITE
  /MISSING=LISTWISE
```

Example 2

```
EXAMINE
  VARIABLES=SPSM_0 SPSM_3 SPSM_DIF BY A4D_C
  /PLOT BOXPLOT
  /COMPARE GROUP
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```


Example 2
Explore

[DataSet1] G:\DataRequests\ReferralSource\RefSrc0_3.sav

a4d_c

Case Processing Summary

| a4d_c | | Cases | | | | | |
|---|--------------------------------------|-------|---------|---------|---------|-------|---------|
| | | Valid | | Missing | | Total | |
| | | N | Percent | N | Percent | N | Percent |
| spsm_0 spsm_0: Substance Problem Scale (Past Month) | 1.00 Juvenile Justice/Corrections | 802 | 56.8% | 610 | 43.2% | 1412 | 100.0% |
| | 2.00 School | 250 | 73.7% | 89 | 26.3% | 339 | 100.0% |
| | 3.00 Self/Family | 439 | 68.9% | 198 | 31.1% | 637 | 100.0% |
| | 4.00 Other | 515 | 63.3% | 298 | 36.7% | 813 | 100.0% |
| spsm_3 spsm_3: Substance Problem Scale (Past Month) | 1.00 Juvenile Justice/Corrections | 802 | 56.8% | 610 | 43.2% | 1412 | 100.0% |
| | 2.00 School | 250 | 73.7% | 89 | 26.3% | 339 | 100.0% |
| | 3.00 Self/Family | 439 | 68.9% | 198 | 31.1% | 637 | 100.0% |
| | 4.00 Other | 515 | 63.3% | 298 | 36.7% | 813 | 100.0% |
| SPSm_dif | 1.00 Juvenile Justice/Corrections | 802 | 56.8% | 610 | 43.2% | 1412 | 100.0% |
| | 2.00 School | 250 | 73.7% | 89 | 26.3% | 339 | 100.0% |
| | 3.00 Self/Family | 439 | 68.9% | 198 | 31.1% | 637 | 100.0% |
| | 4.00 Other | 515 | 63.3% | 298 | 36.7% | 813 | 100.0% |

Descriptives

| a4d_c | | Statistic | | Std. Error |
|--|--------------------------------------|-------------------------------------|--|------------|
| spsm_0 spsm_0: Substance Problem Scale (Past Month) | 1.00 Juvenile Justice/Corrections | Mean | 2.2893 | .10920 |
| | | 95% Confidence Interval for Mean | Lower Bound 2.0749 Upper Bound 2.5036 | |
| | 5% Trimmed Mean | 1.9217 | | |
| | Median | 1.0000 | | |
| | Variance | 9.564 | | |
| | Std. Deviation | 3.09260 | | |
| | Minimum | .00 | | |
| | Maximum | 15.00 | | |
| | Range | 15.00 | | |
| | Interquartile Range | 4.00 | | |
| | Skewness | 1.584 | .086 | |
| | Kurtosis | 2.106 | .172 | |
| | 2.00 School | Mean | 2.6760 | .20690 |
| | | 95% Confidence Interval for Mean | Lower Bound 2.2685 Upper Bound 3.0835 | |
| 5% Trimmed Mean | | 2.3067 | | |

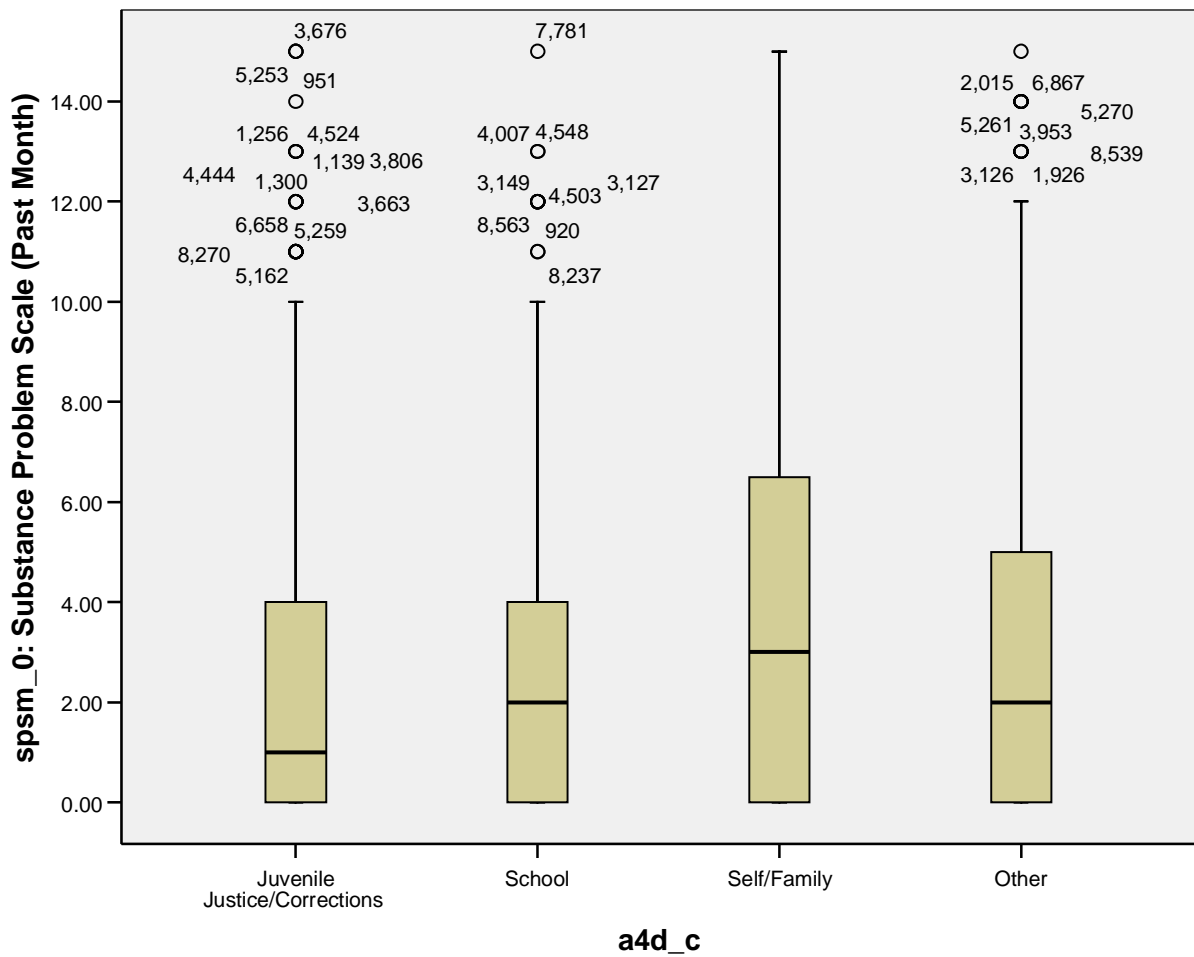
| | | | | | |
|--------|--|-----------------------------------|----------------------------------|---------|--------|
| | | Median | | 2.0000 | |
| | | Variance | | 10.702 | |
| | | Std. Deviation | | 3.27137 | |
| | | Minimum | | .00 | |
| | | Maximum | | 15.00 | |
| | | Range | | 15.00 | |
| | | Interquartile Range | | 4.00 | |
| | | Skewness | | 1.520 | .154 |
| | | Kurtosis | | 1.825 | .307 |
| | 3.00 Self/Family | Mean | | 4.1390 | .19509 |
| | | 95% Confidence Interval for Mean | Lower Bound | 3.7555 | |
| | | | Upper Bound | 4.5224 | |
| | | 5% Trimmed Mean | | 3.8461 | |
| | | Median | | 3.0000 | |
| | | Variance | | 16.709 | |
| | | Std. Deviation | | 4.08766 | |
| | | Minimum | | .00 | |
| | | Maximum | | 15.00 | |
| | | Range | | 15.00 | |
| | | Interquartile Range | | 7.00 | |
| | | Skewness | | .850 | .117 |
| | | Kurtosis | | -.321 | .233 |
| | 4.00 Other | Mean | | 3.1515 | .15769 |
| | | 95% Confidence Interval for Mean | Lower Bound | 2.8417 | |
| | | | Upper Bound | 3.4612 | |
| | | 5% Trimmed Mean | | 2.7972 | |
| | | Median | | 2.0000 | |
| | | Variance | | 12.806 | |
| | | Std. Deviation | | 3.57852 | |
| | | Minimum | | .00 | |
| | | Maximum | | 15.00 | |
| | | Range | | 15.00 | |
| | | Interquartile Range | | 5.00 | |
| | | Skewness | | 1.194 | .108 |
| | | Kurtosis | | .710 | .215 |
| spsm_3 | spsm_3: Substance Problem Scale (Past Month) | 1.00 Juvenile Justice/Corrections | Mean | 1.3005 | .09237 |
| | | | 95% Confidence Interval for Mean | 1.1192 | |
| | | | Lower Bound | 1.4818 | |
| | | | Upper Bound | | |
| | | 5% Trimmed Mean | | .8857 | |
| | | Median | | .0000 | |
| | | Variance | | 6.842 | |
| | | Std. Deviation | | 2.61575 | |
| | | Minimum | | .00 | |

| | | | | |
|------------------|----------------------------------|-------------|---------|--------|
| | Maximum | | 16.00 | |
| | Range | | 16.00 | |
| | Interquartile Range | | 1.00 | |
| | Skewness | | 2.560 | .086 |
| | Kurtosis | | 6.703 | .172 |
| 2.00 School | Mean | | 1.7320 | .19778 |
| | 95% Confidence Interval for Mean | Lower Bound | 1.3425 | |
| | | Upper Bound | 2.1215 | |
| | 5% Trimmed Mean | | 1.2644 | |
| | Median | | .0000 | |
| | Variance | | 9.779 | |
| | Std. Deviation | | 3.12719 | |
| | Minimum | | .00 | |
| | Maximum | | 16.00 | |
| | Range | | 16.00 | |
| | Interquartile Range | | 2.00 | |
| | Skewness | | 2.217 | .154 |
| | Kurtosis | | 4.634 | .307 |
| 3.00 Self/Family | Mean | | 2.1412 | .16554 |
| | 95% Confidence Interval for Mean | Lower Bound | 1.8159 | |
| | | Upper Bound | 2.4666 | |
| | 5% Trimmed Mean | | 1.6945 | |
| | Median | | .0000 | |
| | Variance | | 12.030 | |
| | Std. Deviation | | 3.46846 | |
| | Minimum | | .00 | |
| | Maximum | | 16.00 | |
| | Range | | 16.00 | |
| | Interquartile Range | | 3.00 | |
| | Skewness | | 1.782 | .117 |
| | Kurtosis | | 2.355 | .233 |
| 4.00 Other | Mean | | 1.5010 | .12600 |
| | 95% Confidence Interval for Mean | Lower Bound | 1.2534 | |
| | | Upper Bound | 1.7485 | |
| | 5% Trimmed Mean | | 1.0529 | |
| | Median | | .0000 | |
| | Variance | | 8.177 | |
| | Std. Deviation | | 2.85947 | |
| | Minimum | | .00 | |
| | Maximum | | 15.00 | |
| | Range | | 15.00 | |
| | Interquartile Range | | 2.00 | |
| | Skewness | | 2.429 | .108 |

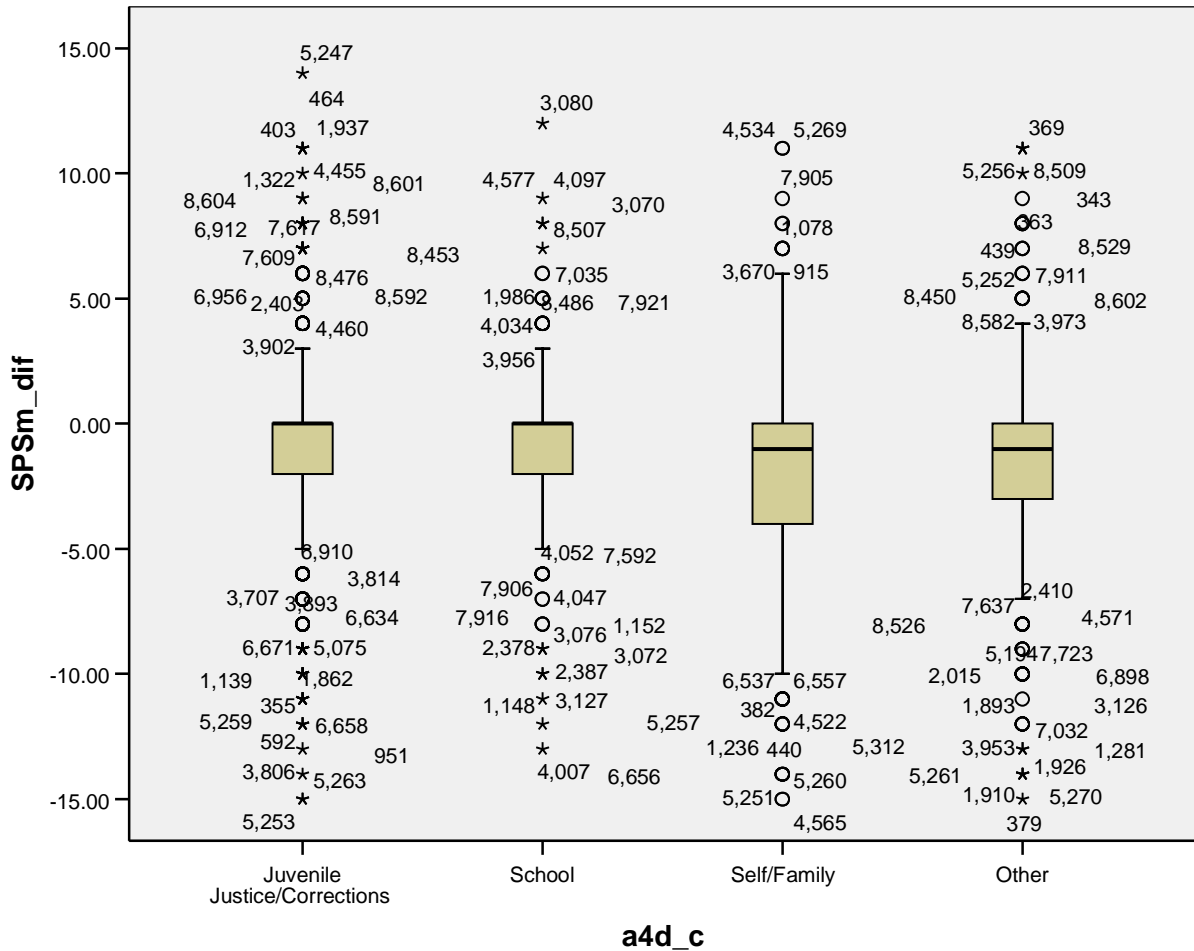
| SPSm_dif | Category | Statistic | Value | Significance | |
|---------------------|--------------------------------------|----------------------------------|----------------------------|-------------------|--|
| SPSm_dif | 1.00 Juvenile Justice/Corrections | Kurtosis | 5.869 | .215 | |
| | | Mean | - .9888 | .12599 | |
| | | 95% Confidence Interval for Mean | Lower Bound Upper Bound | -1.2361 -.7415 | |
| | | 5% Trimmed Mean | - .9569 | | |
| | | Median | .0000 | | |
| | | Variance | 12.730 | | |
| | | Std. Deviation | 3.56794 | | |
| | | Minimum | -15.00 | | |
| | | Maximum | 14.00 | | |
| | | Range | 29.00 | | |
| | Interquartile Range | 2.00 | | | |
| | Skewness | -.314 | .086 | | |
| | 2.00 School | Kurtosis | 2.350 | .172 | |
| | | Mean | -.9440 | .20777 | |
| | | 95% Confidence Interval for Mean | Lower Bound Upper Bound | -1.3532 -.5348 | |
| | | 5% Trimmed Mean | -.9022 | | |
| | | Median | .0000 | | |
| | | Variance | 10.792 | | |
| | | Std. Deviation | 3.28512 | | |
| | | Minimum | -13.00 | | |
| Maximum | | 12.00 | | | |
| Range | | 25.00 | | | |
| Interquartile Range | 2.00 | | | | |
| Skewness | -.236 | .154 | | | |
| 3.00 Self/Family | Kurtosis | 2.544 | .307 | | |
| | Mean | -1.9977 | .20363 | | |
| | 95% Confidence Interval for Mean | Lower Bound Upper Bound | -2.3979 -1.5975 | | |
| | 5% Trimmed Mean | -1.9141 | | | |
| | Median | -1.0000 | | | |
| | Variance | 18.203 | | | |
| | Std. Deviation | 4.26652 | | | |
| | Minimum | -15.00 | | | |
| | Maximum | 11.00 | | | |
| | Range | 26.00 | | | |
| Interquartile Range | 4.00 | | | | |
| Skewness | -.480 | .117 | | | |
| 4.00 Other | Kurtosis | .859 | .233 | | |
| | Mean | -1.6505 | .16753 | | |
| | 95% Confidence Interval for Mean | Lower Bound | -1.9796 | | |

| | | |
|---------------------|-------------|---------|
| Mean | Upper Bound | -1.3214 |
| 5% Trimmed Mean | | -1.5965 |
| Median | | -1.0000 |
| Variance | | 14.453 |
| Std. Deviation | | 3.80177 |
| Minimum | | -15.00 |
| Maximum | | 11.00 |
| Range | | 26.00 |
| Interquartile Range | | 3.00 |
| Skewness | | -.364 |
| Kurtosis | | 1.776 |
| | | .108 |
| | | .215 |

spsm_0



SPSm_dif



Comments:

In the Tukey boxplot template.xls (Excel) file, the following information is requested, where rows, Quarter 0 through Quarter 4, reflect the comparison groups (in this case waves of data at intake and 4 follow-ups) and the columns reflect percentiles and ranges.

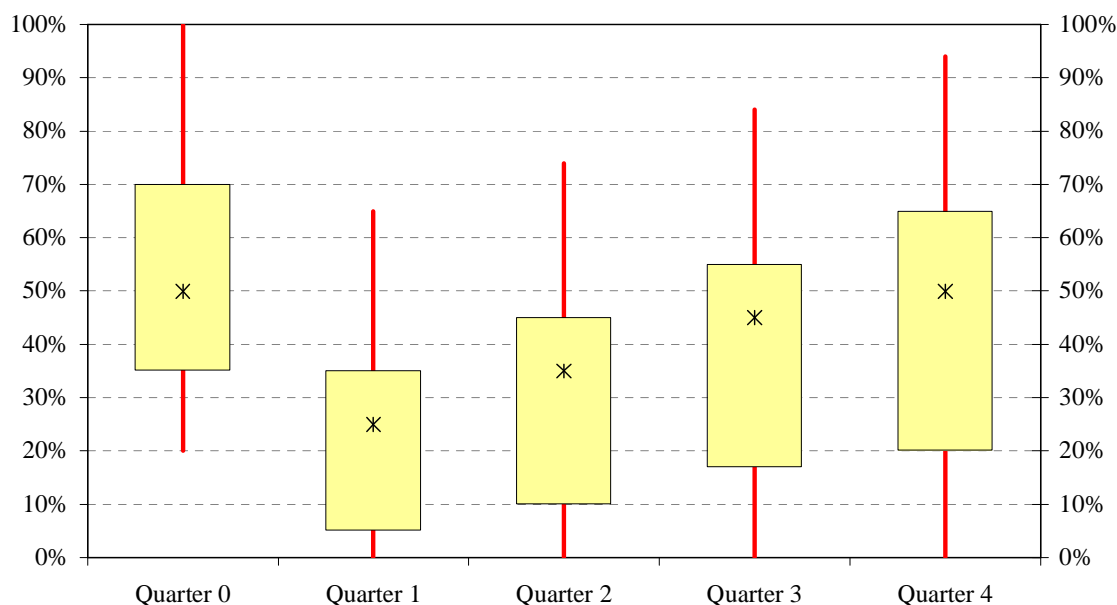
| A | B | C | D | E | F |
|-----------|------|-----|------|------|-----|
| Median | 25th | Max | Min | 75th | |
| Quarter 0 | 50% | 35% | 100% | 20% | 70% |
| Quarter 1 | 25% | 5% | 65% | 0% | 35% |
| Quarter 2 | 35% | 10% | 74% | 0% | 45% |
| Quarter 3 | 45% | 17% | 84% | 0% | 55% |
| Quarter 4 | 50% | 20% | 94% | 0% | 65% |

These data will produce the chart shown below, where the yellow boxes represent the Interquartile range (25th to 75th percentiles), the “*” represents the median, and the red

bars indicate the minimum and maximum values. The percentiles to fill in the above can be obtained by adding the following sub-command to the above syntax examples:

```
/PERCENTILES(25,50,75) HAVERAGE.
```

The resulting graph is shown below. The asterisk represents the median value from column B of the table above. The yellow boxes are defined by the 25th and 75th percentiles from columns C and F respectively. The red lines above and below the boxes are defined by the maximum value (column D) and the minimum value (column E).



In a normally distributed scale the red line below the box will be about the same length as the red line above the box, with a box that is typically smaller than the lines. Skewed scores typically have one line that is markedly longer than the other. Uniform scores typically have a large box size compared to the size of the associated lines. Small boxes and long lines indicate less variance (a narrow clustering of values with outliers or peaked distribution), while large boxes and short lines indicate greater variance (a wide spread of values or flat distribution). In comparing groups (e.g. quarters), if the entire box for one group does not overlap that of another, then the difference between those two groups is significant. If the median (asterisk) is not in the middle of the box, then the data are skewed.

Describing These Procedures. The Tukey boxplot procedures are normally used as a tool to help check the distributions of continuous variables by a categorical variable to help check the distribution of variables and to look for any outliers before statistical

analysis are performed. They are also useful for showing distributions in presentations, but are little used in journal articles.

Bibliography

Becker, L. A., *Explore: Statistics and Charts*, revised 9/24/1999

<http://web.uccs.edu/lbecker/SPSS/explore1.htm> (accessed 6/23/06).

Statistical Program for the Social Sciences version 14.0.2 (SPSS 2006). Chicago, IL:

Author (www.spss.com).

John W. Tukey. "Exploratory Data Analysis". Addison-Wesley, Reading, MA. 1977.