Chapter 33
Box Plots and Mosaic Plots

Chapter Table of Contents

VARIABLES 478

METHOD 480

OUTPUT 481

Multiple Comparison Options 483

Multiple Comparison Circles 485

REFERENCES 486

Chapter 33
Box Plots and Mosaic Plots

Box plots are pictorial representations of the distribution of values of a variable. The central line in each box marks the median value and the edges of the box mark the first and third quartiles.

The *median* value of a distribution is the 50th *percentile*: It is the value less than and greater than 50% of the data. The first and third *quartiles* are the 25th and 75th percentiles. By combining these three values in a schematic diagram and plotting individual markers for extreme data values, the box plot provides a concise display of a distribution (Tukey 1977).

Mosaic plots are pictorial representations of frequency counts of a single nominal variable or cross-classified nominal variables. Because mosaic plots display the frequencies graphically, they are easier to understand than crosstabulations. You can select and brush mosaic plots to explore dependencies between variables.

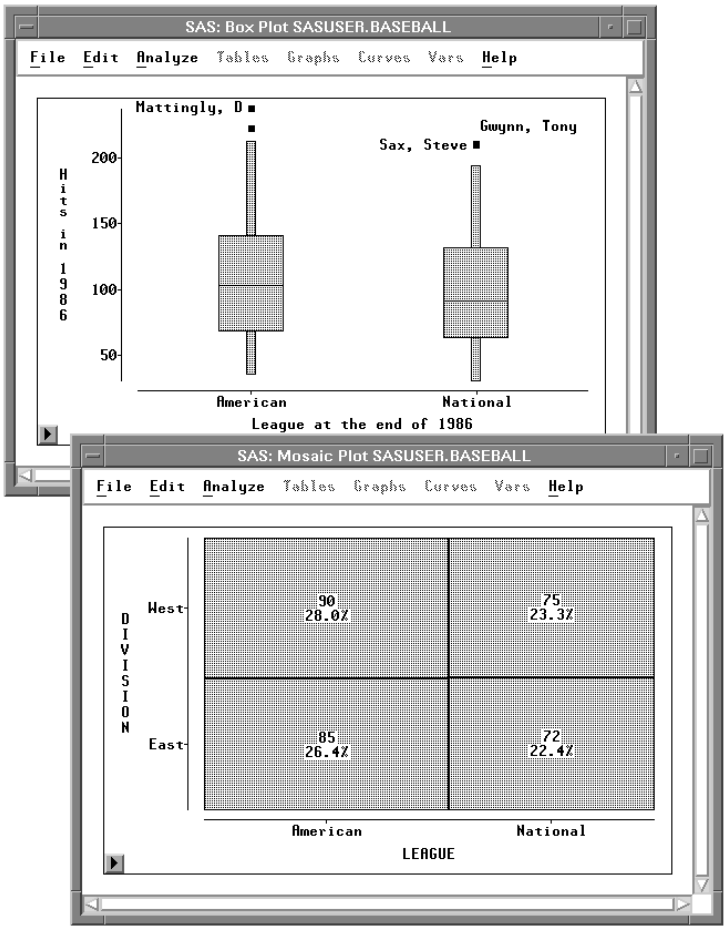

Figure 33.1. Box Plot and Mosaic Plot

Variables

To create a box plot or mosaic plot, choose **Analyze:Box Plot/Mosaic Plot (Y)**. If you have previously selected one or more variables, they are assigned the required **Y** variable role. A single plot is created containing a separate schematic diagram for each **Y** variable selected. For interval **Y** variables, box plots are created. For nominal **Y** variables, mosaic plots are created.

If you have not selected any variables, a variables dialog appears.

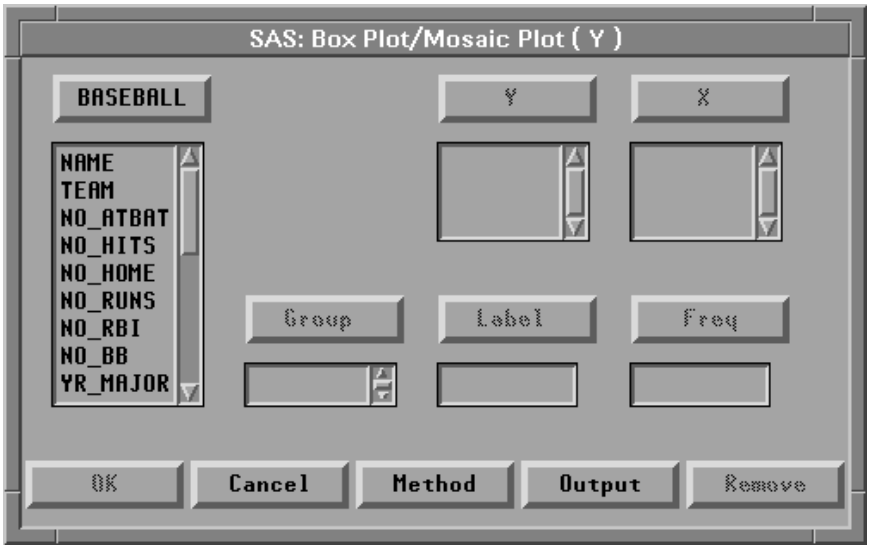

Figure 33.2. Box Plot/Mosaic Plot Variables Dialog

In the dialog, select at least one **Y** variable.

You can select one or more **X** variables to compare distributions. If you do not select **X** variables, you get one plot containing one schematic diagram for each **Y** variable. If you select **X** variables, you get one plot for each **Y** variable, and each plot contains one schematic diagram for each combination of **X** values. For example, Figure 33.3 shows the box plot created using the **BASEBALL** data set with **NO_HITS** as the **Y** variable and **LEAGUE** as the **X** variable.

You can select one or more **Group** variables if you have grouped data. This creates a separate box or mosaic plot for each group. For example, Figure 33.4 shows the box plots created using the **BASEBALL** data set with **NO_HITS** as the **Y** variable and **LEAGUE** as the **Group** variable.

You can select a **Label** variable to label extreme values in box plots.

If you select a **Freq** variable, each observation is assumed to represent *n* observations, where *n* is the value of the **Freq** variable.

You can identify extreme values in the box plot and display the *mean* or average value. You can also control the marker size of extreme values and the information shown in the box plot axes.

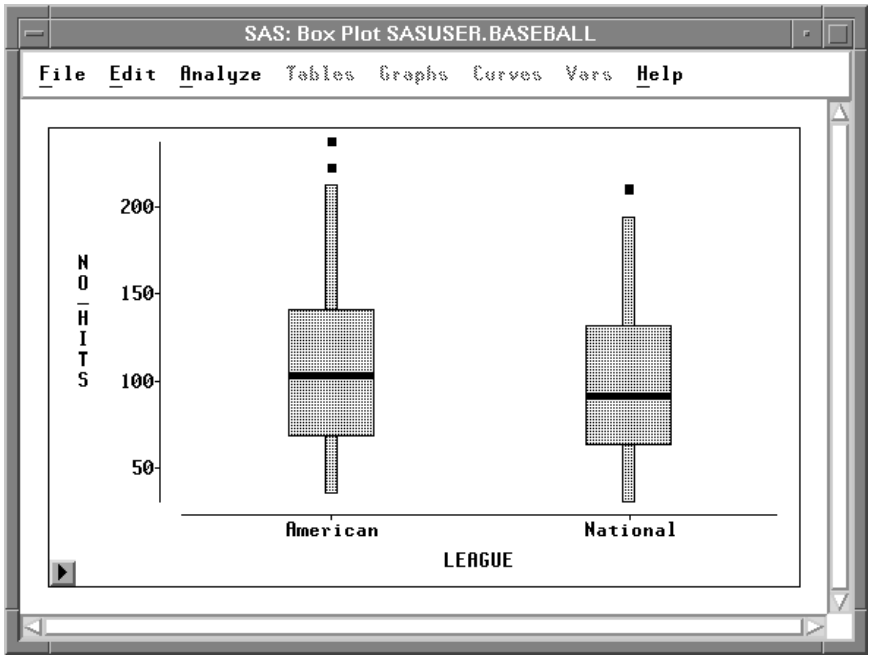

Figure 33.3. Box Plot Using **X** Variable

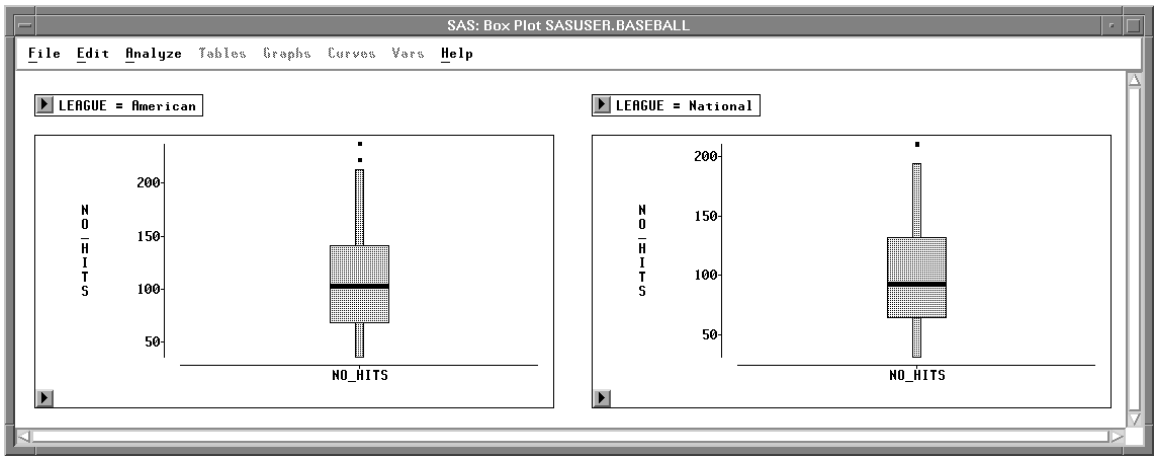

Figure 33.4. Box Plot Using **Group** Variable

Method

Observations with missing values for **Y** variables are not used. Observations with **Freq** values that are missing or that are less than or equal to 0 are not used. Only the integer part of **Freq** values is used.

The following method is used to compute the median and quartiles. Let

n be the number of data values

$y_1, y_2, \ldots, y_n$ be the data values listed in increasing order

p be the desired percentile (25, 50, or 75)

i be the integer part, and f the fractional part, of the ordinal of the desired percentile:

$$i + f = n * p/100$$

Then the value of the desired percentile is

$$\begin{array}{ll} (y_i + y_{i+1})/2 & \text{if } f = 0 \\ y_{i+1} & \text{if } f > 0 \end{array}$$

You can adjust three calculation methods by clicking on the **Method** button in the variables dialog. This displays the method options dialog.

Figure 33.5. Box Plot/Mosaic Plot Method Options Dialog

By default, *whiskers* on the box plot are drawn from the quartiles to the farthest observation not farther than 1.5 times the distance between the quartiles. Type your preferred whisker length factor in the entry field. The figures in this chapter were created using whisker lengths that were 1.0 times the distance between the quartiles; this results in more observations being classified as outliers.

By default, for variables in mosaic plots, values that represent less than 4% of the total frequency are grouped together in an "**Other**" category. The Method dialog enables you to change the threshold at which values are grouped in the **Other** category.

By default, **X** variable values are sorted by their formatted value. Turn off the **Sort X Formatted** check box to sort **X** variable values by their unformatted value.

Output

To view or modify output options associated with your plot, click on the **Output** button of the variables dialog. This displays the output options dialog.

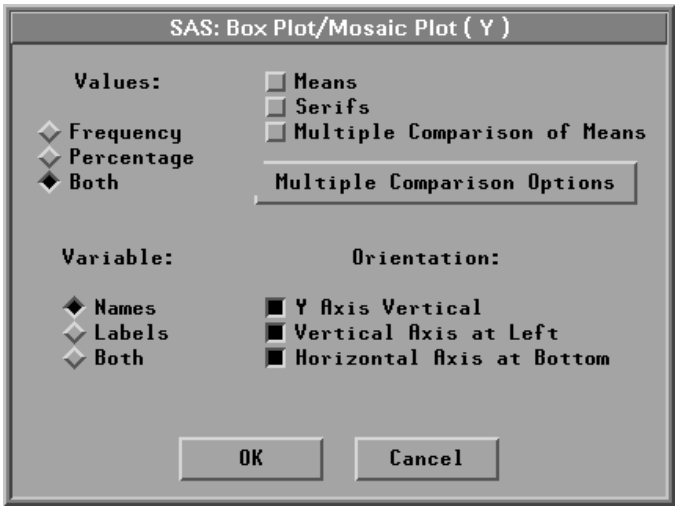

Figure 33.6. Box Plot/Mosaic Plot Output Options Dialog

Values:Frequency labels mosaic boxes with the frequency of observations represented in each box.

Values:Percentage labels mosaic boxes with the percentage of observations represented in each box.

Values:Both labels mosaic boxes with both frequency and percentage.

Means displays mean diamonds on box plots. The central line in the diamond marks the mean. The size of the diamond is two standard deviations, one on either side of the mean.

Serifs displays serifs at the ends of box plot whiskers.

Multiple Comparison of Means displays a *comparison circle* (Sall 1992) for each box. The center of each circle marks the mean of each box. The color and line style of each circle indicates how the mean value of one box compares with the means of other boxes. A selected circle is highlighted and is drawn in red on color monitors. Circles corresponding to categories whose mean values are significantly different from a selected group are drawn in cyan on color monitors. Circles corresponding to categories whose mean values are not different are drawn with a dashed line and are red on color monitors. See the section “Multiple Comparison Circles” later in this chapter.

Multiple Comparison Options	displays the Multiple Comparison Options dialog window.
Variable:Names	labels the axes with variable names.
Variable:Labels	labels the axes with variable labels.
Variable:Both	labels the axes with both names and labels.
Orientation: Y Axis Vertical	draws the axis for the **Y** variable vertically. If this option is off, the **Y** axis is horizontal.
Orientation: Vertical Axis at Left	places the vertical axis at the left side of the plot. If this option is off, the vertical axis is at the right side.
Orientation: Horizontal Axis at Bottom	places the horizontal axis at the bottom of the plot. If this option is off, the horizontal axis is at the top.

You can modify other aspects of box and mosaic plots with the pop-up menu.

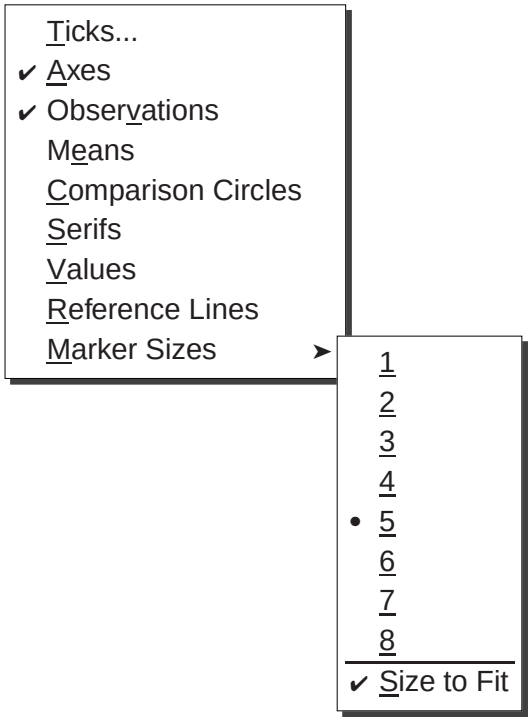

Figure 33.7. Box Plot/Mosaic Plot Pop-up Menu

Ticks...	specifies tick labels on the **Y** axis.
Axes	toggles the display of axes.
Observations	toggles the display of observations (boxes and extreme values). When this menu is toggled off, observations are displayed only if selected.

Means toggles the display of mean diamonds in box plots.

Comparison Circles toggles the display of comparison circles in box plots.

Serifs toggles the display of serifs at the ends of box plot whiskers.

Values toggles the display of values for means, medians, quartiles, and ends of whiskers in box plots. Toggles the display of frequency and percentage counts in mosaic plots.

Reference Lines toggles the display of lines that indicate the position of major ticks on the **Y** axis. This option is not available unless the axes are visible.

Marker Sizes sets the size of markers that display extreme values in box plots.

Multiple Comparison Options

Box plots enable you to examine means in different groups. Statistical questions you might have about the group means include

- Which underlying group means are likely to be different?
- Which group means are better than the mean of a standard group?
- Which group means are statistically indistinguishable from the best?

From the **Multiple Comparison Options** dialog, you can select a multiple comparison of means test and a confidence level for the test. Multiple comparison tests enable you to infer differences between means and also to construct simultaneous confidence intervals for these differences.

All of the tests implemented in SAS/INSIGHT software are constructed assuming that the displayed variables are independent and normally distributed with identical variance. For details, refer to Hsu (1996).

Each of the tests available in SAS/INSIGHT software is described below. In the descriptions that follow, k is the number of categories (that is, the number of boxes in the box plot), n_i is the number of observations for the ith category, μ_i is the true mean for the ith category, $\hat{\mu}_i$ is the sample mean for the ith category, $\nu = \sum_{i=1}^{k}(n_i - 1)$ is the total degrees of freedom, and $\hat{\sigma}$ is the root mean square error, also known as the pooled standard deviation. Each test creates a table showing $100(1-\alpha)\%$ confidence intervals for the difference $\hat{\mu}_i - \hat{\mu}_j, i \neq j, i = 1 \ldots k$.

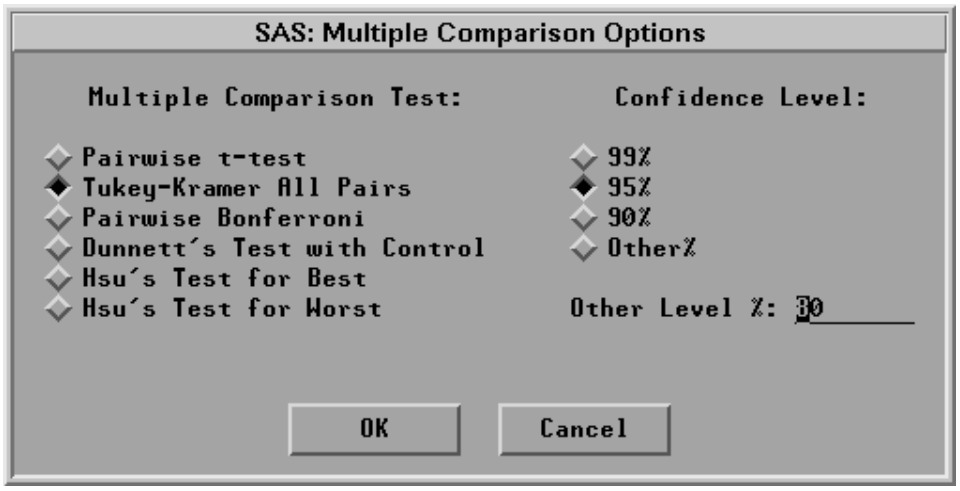

Figure 33.8. Multiple Comparison Options

The **Pairwise t-test** is not a true simultaneous comparison test, but rather uses a pairwise t test to provide confidence intervals about the difference between two means. These intervals have a half-width equal to $t_{\alpha/2,\nu}\hat{\sigma}\sqrt{n_i^{-1}+n_j^{-1}}$. Although each confidence interval was computed at the $100(1-\alpha)\%$ level, the probability that all of your confidence intervals are correct *simultaneously* is less than $100(1-\alpha)\%$. The actual simultaneous confidence for the t-based intervals is approximately $100(1-k\alpha)\%$. For example, for five groups the actual simultaneous confidence for the t-based intervals is approximately only 75%.

The **Tukey-Kramer** method is a true "multiple comparison" test, appropriate when all pairwise comparisons are of interest; it is the default test used. The test is an exact α-level test if the sample sizes are the same, and it is slightly conservative for unequal sample sizes. The confidence interval around the point-estimate $\hat{\mu}_i - \hat{\mu}_j$ has half-width $q^*\hat{\sigma}\sqrt{n_i^{-1}+n_j^{-1}}$. It is a common convention to report the quantity $\sqrt{2}q^*$ as the Tukey-Kramer quantile, rather than just q^*.

The **Pairwise Bonferroni** method is also appropriate when all pairwise comparisons are of interest. It is conservative; that is, Bonferroni tests performed at a nominal significance level of α actually have a somewhat greater level of significance. The Bonferroni method uses the t distribution, like the pairwise t test, but returns smaller intervals with half-width $t_{\alpha/(k(k-1)),\nu}\hat{\sigma}\sqrt{n_i^{-1}+n_j^{-1}}$. Note that the t probability ($\alpha/2$, since this is a two-sided test) is divided by the total number of pairwise comparisons ($k(k-1)/2$). The Bonferroni test produces wider confidence intervals than the Tukey-Kramer test.

Dunnett's Test with Control is a two-sided multiple comparison method used to compare a set of categories to a control group. The quantile that scales the confidence interval is usually denoted $|d|$. If the ith confidence interval does not include zero, you may infer that the ith group is significantly different from the control. A control group may be a placebo or null treatment, or it may be a standard treatment. While the interactive nature of SAS/INSIGHT enables you to select any category to use as the basis of comparison in Dunnett's test, you should select a category only if it truly

is the control group. To select a category, click on the corresponding comparison circle.

Hsu's Test for Best can be used to screen out group means that are statistically less than the (unknown) largest true mean. It forms *nonsymmetric* confidence intervals around the difference between the largest sample mean and each of the others. If an interval does not properly contain zero in its interior, then you may infer that the associated group is not among the best.

Similary, **Hsu's Test for Worst** can be used to screen out group means that are statistically greater than the (unknown) smallest true mean. If an interval does not properly contain zero in its interior, then you may infer that the true mean of that group is not equal to the (unknown) smallest true mean.

Multiple Comparison Circles

In addition to a table that summarizes the statistics for simultaneous multiple comparison of means, SAS/INSIGHT software provides a graphical technique to help visualize which groups are significantly different from a selected group. Each test is accompanied by a *comparison circles* plot that graphically illustrates the comparisons (Sall 1992).

There is a circle next to the box plot and centered at each category's sample mean. The radius of the ith circle is $q\hat{\sigma}/\sqrt{n_i}$, where q is a quantile used to scale the circles according to the test being used. For details on how each quantile is computed, see refer to Hsu (1996).

If the jth group is selected (by clicking on its circle), then its circle is highlighted. This circle is red on color monitors. You can determine whether another group is significantly different than the selected group based on how much their corresponding circles overlap. If their circles are nested or nearly overlap so that the external angle of intersection is greater than 90 degrees, then you cannot claim that the means of the two groups are different. If, however, the two circles are disjoint or just barely overlap so that their external angle of intersection is less than 90 degrees, then you can conclude that the means of the two groups are significantly different at the given confidence level.

Circles corresponding to categories that are significantly different from the selected group are drawn in cyan on color monitors. Circles corresponding to categories that are not different are drawn with a dashed line and are red on color monitors.

The geometry behind comparison circles is based on the Pythagorean Theorem: since the radius of the ith circle is $r_i = q\hat{\sigma}/\sqrt{n_i}$, and since the circle is centered at $\hat{\mu}_i$, then if the two circles meet at right angles, the distance between centers is the hypotenuse of the right triangle formed by the circles' radii. Therefore, when the circles meet at right angles, $|\hat{\mu}_i - \hat{\mu}_j| = q\hat{\sigma}\sqrt{n_i^{-1} + n_j^{-1}}$. Statistically, this geometry corresponds to the critical case in which zero happens to fall on the boundary of the confidence interval about $\hat{\mu}_i - \hat{\mu}_j$. If $|\hat{\mu}_i - \hat{\mu}_j| > q\hat{\sigma}\sqrt{n_i^{-1} + n_j^{-1}}$, then the external intersection of the circles is less than 90 degrees, and zero is not contained in the confidence interval about $\hat{\mu}_i - \hat{\mu}_j$. Thus the circles are significantly different.

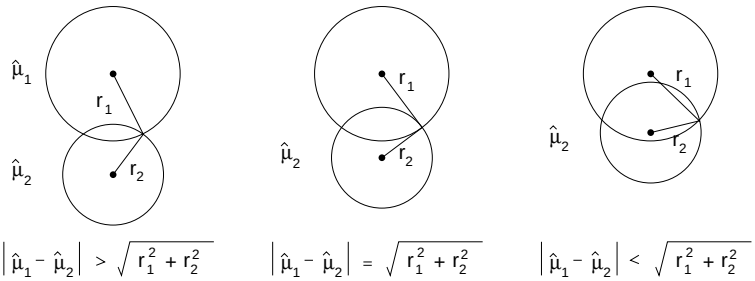

Figure 33.9. The Geometry of Multiple Comparison Circles

The statistics for Hsu's Test for Best and Hsu's Test for Worst are computed differently from the other tests. First, the comparison circles are not selectable. The Test for Best automatically selects the category with the largest sample mean; the Test for Worst selects the category with the smallest sample mean. Second, the quantile used to scale the comparison circles is the maximum of the quantiles computed by running Dunnett's one-sided test $k - 1$ times, with each "non-best" (or "non-worst") group serving in turn as the "control" for Dunnett's test.

Because Hsu's Test for Best does not provide symmetric intervals about $\hat{\mu}_i - \hat{\mu}_j$, the comparison circle technique must be modified. While the statistical table reports exactly which groups can be inferred not to be the best, the comparison circles are more conservative because the quantile used to scale the circle radii is the maximum of all quantiles encountered during Hsu's test. The same is true for Hsu's Test for Worst.

⊕ **Related Reading:** Box Plots, Chapter 4.

⊕ **Related Reading:** Mosaic Plots, Chapter 5.

⊕ **Related Reading:** Distributions, Chapter 12.

References

Hartigan, J.A. and Kleiner, B. (1984), "A Mosaic of Television Ratings," *The American Statistician*, 38, 32–35.

Hsu, J.C. (1996), *Multiple Comparisons: Theory and Methods*, London: Chapman & Hall.

Sall, J. (1992), "Graphical Comparison of Means," *Statistical Computing and Statistical Graphics Newsletter*, 3, 27–32.

Tukey, J.W. (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/INSIGHT User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999. 752 pp.

SAS/INSIGHT User's Guide, Version 8

ISBN 1–58025–490–X

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, October 1999