



Box-and-Whisker Plots with The SAS System®

David Shannon

Amadeus Software Limited

PSI Conference, 12th May 2003.



Topics to be Covered

- **Introduction**
- **How is the Boxplot Constructed?**
- **Boxplots in The SAS System**
- **The SAS System's Proc Boxplot**
 - Statistical Capabilities
 - Styles of Boxplots
 - Grouping within Blocks
- **Conclusions**

Introduction

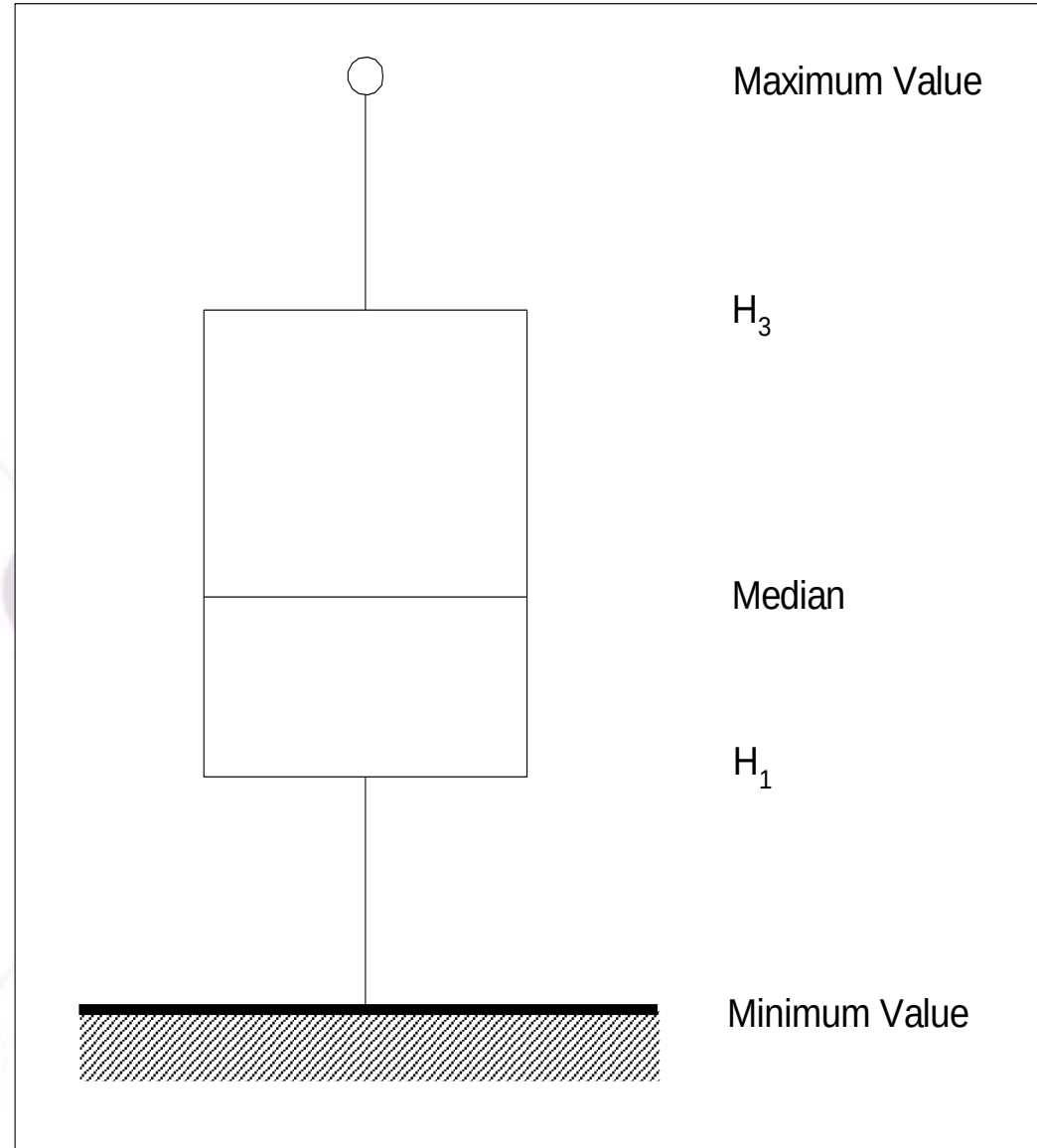
- **Boxplot first proposed by Tukey in 1977**
 - Several variations on the boxplot subsequently published
 - Notably in 1978 by McGill, Tukey and Larsen which introduced the notched boxplot
- **Graphical method of exploring distributions of qualitative data**
 - Used widely by many researchers – not just the pharmaceutical industry
 - Means of exploring and presenting data
- **One of the Statisticians most commonly available tools is The SAS System**

How is the Boxplot Constructed?

- **Each boxplot is constructed from raw data**
 - Mean
 - Median
 - Quartiles
 - Outlying observations
- **Two common styles of boxplot used in statistics today**
 - Skeletal boxplot
 - Schematic boxplot

Tukey's 1977 Boxplot

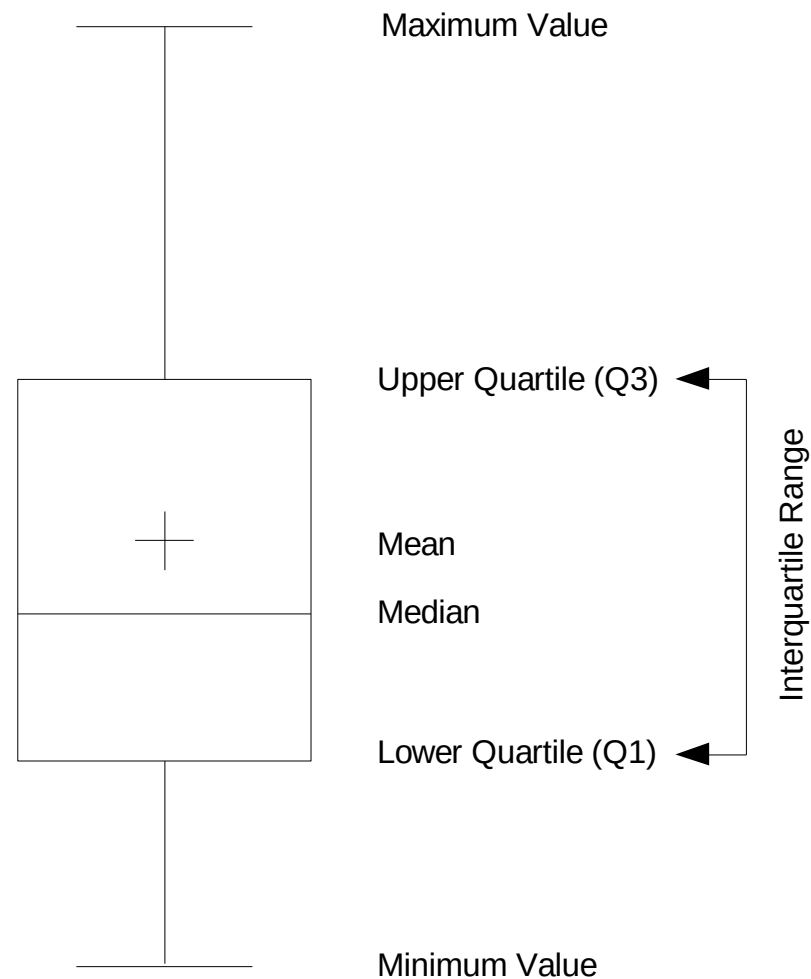
- **Whiskers extended to the extremes of the data**
- **Minimum value had hatched line**
- **Hinges were used to define boundaries of the box**



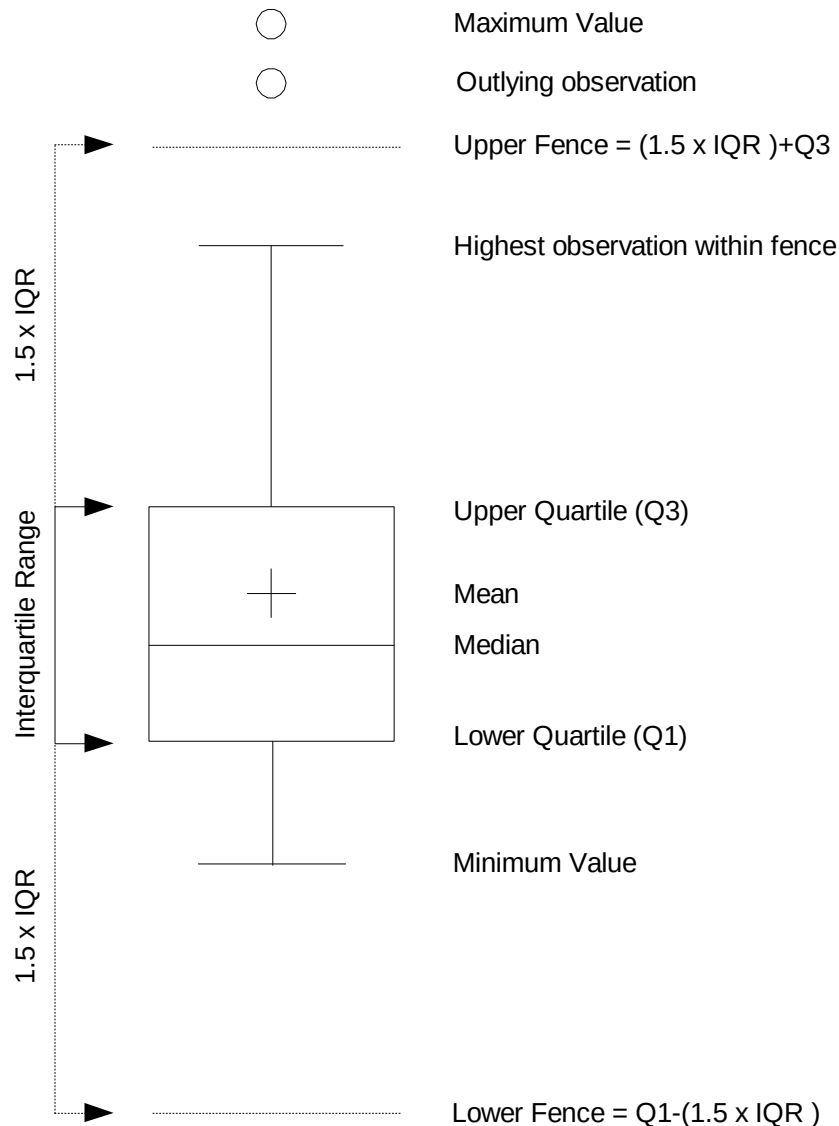
How is the Boxplot Constructed?

- **Skeletal Plot**

- Whiskers extend from the minimum to the maximum data values
- Box represents the interquartile range – divided by the median
- Arithmetic mean is presented by a symbol



How is the Boxplot Constructed?



- **Schematic Plot**

- **Box also represents the IQR, mean and median**
- **Fences are derived as**
Lower = $Q1 - (1.5 \times \text{IQR})$
Upper = $Q3 + (1.5 \times \text{IQR})$
- **Whiskers drawn to the minimum and maximum observations inside fences**
- **Observations outside fences are plotted as outliers**

Boxplots in The SAS System

- **Prior to SAS V8 boxplots can be produced by:**
 - **PROC UNIVARIATE** for line plots of boxplots
 - A Base SAS procedure
 - Produced via the PLOT option
 - **PROC GPLOT**
 - A SAS/GRAPH procedure
 - Produced via the SYMBOL statements' interpolation option
 - **PROC SHEWHART**
 - A SAS/QC procedure
 - Produced via the BOXCHART statement
 - Very closely resembles the BOXPLOT procedure

Boxplots in The SAS System

- **Prior to SAS V8 boxplots can be produced by:**
 - **SAS/GRAPH ANNOTATE code**
 - Potential for complexity
 - Requires expertise for maintenance
- **PROC BOXPLOT introduced in V7/V8**
 - Introduced in the SAS/STAT module
 - Also requires SAS/GRAPH as it produces high resolution graphics
 - Enables side-by-side boxplots to be produced on groups of data

The SAS System's Proc Boxplot

- Minimum syntax is:

```
PROC BOXPLOT;
```

```
    PLOT analysis-var * group-var / options;
```

```
RUN;
```

- **Where:**

- *Analysis-var* is the measurement of interest
- *Group-var* is plotted on the x-axis
- *Options* can be any of 89 methodology and appearance options

The SAS System's Proc Boxplot

- Minimum syntax is:

```
PROC BOXPLOT;
```

```
    PLOT analysis-var * group-var / options;
```

```
RUN;
```

- Input data set must be sorted on the **GROUP** variable, otherwise:
 - An error is returned when *group-var* is numeric
 - Character variables are treated similarly to **CLASS** variables that have not been consolidated into unique groups

The SAS System's Proc Boxplot

- **Statistical Capabilities**

- **Procedure calculates all statistics directly from data**

- No method of saving these statistics in V8 – but is introduced in the next version of SAS (V9)

- **Two options are available to provide information directly in the plot:**

- Specify the NLEGEND option to print the sample size of each group in the legend
 - When box widths are drawn in proportion to the group size (requested via BOXWIDTHSCALE=), specify the BWSLEGEND option to print the function used to scale the box widths in the legend

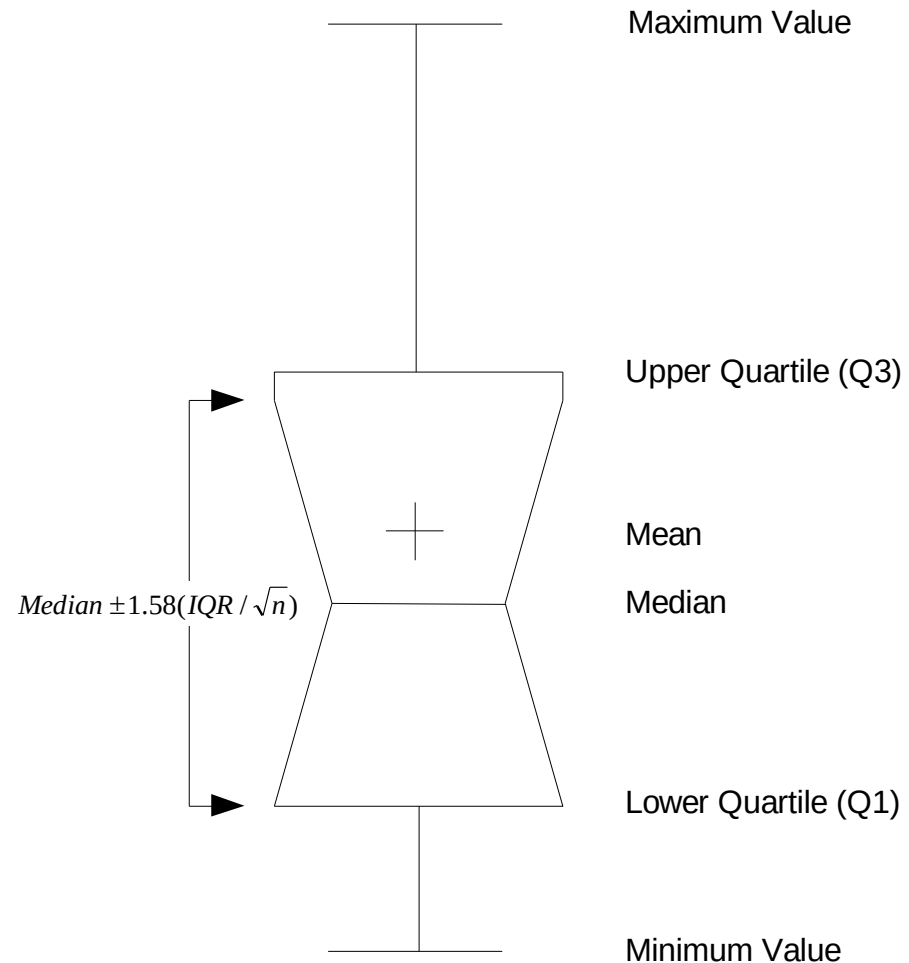
The SAS System's Proc Boxplot

- **Statistical Capabilities**

- **Approximate 95% confidence interval of the median via NOTCHES option**

- **Notches are calculated by:**

$$\text{Median} \pm 1.58 \left(\frac{\text{IQR}}{\sqrt{n}} \right)$$

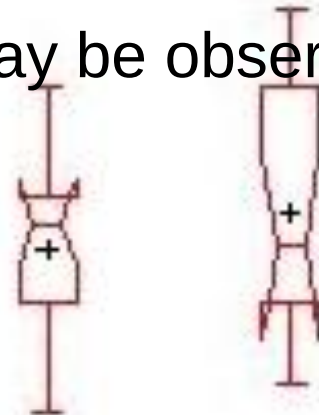


The SAS System's Proc Boxplot

- **Statistical Capabilities**

- Approximate 95% confidence interval of the median via NOTCHES option

- When like groups of data are plotted side-by-side and notch boundaries do not overlap, it can be concluded there is a difference between the medians at around the 5% level of statistical significance
 - In small sample sizes a notch may be observed as folding back on itself



The SAS System's Proc Boxplot

- **Statistical Capabilities**
 - **Five methodologies available to calculate quantile statistics**
 - Weighted average X_{np}
 - Observation number closest to np
 - Empirical distribution function
 - Weighted average $X_{p(n+1)}$
 - Empirical distribution function with averaging (the default)
 - The same option as in all Base SAS procedures which calculate quantiles is used to determine which method is used:

PCTLDEF= *index*

The SAS System's Proc Boxplot

- **Producing a SKELETAL boxplot**

- This is the default style of plot
- Example: Plot white blood cell counts for each of three treatment groups on a given study day

* Data must be sorted by the grouping variable;

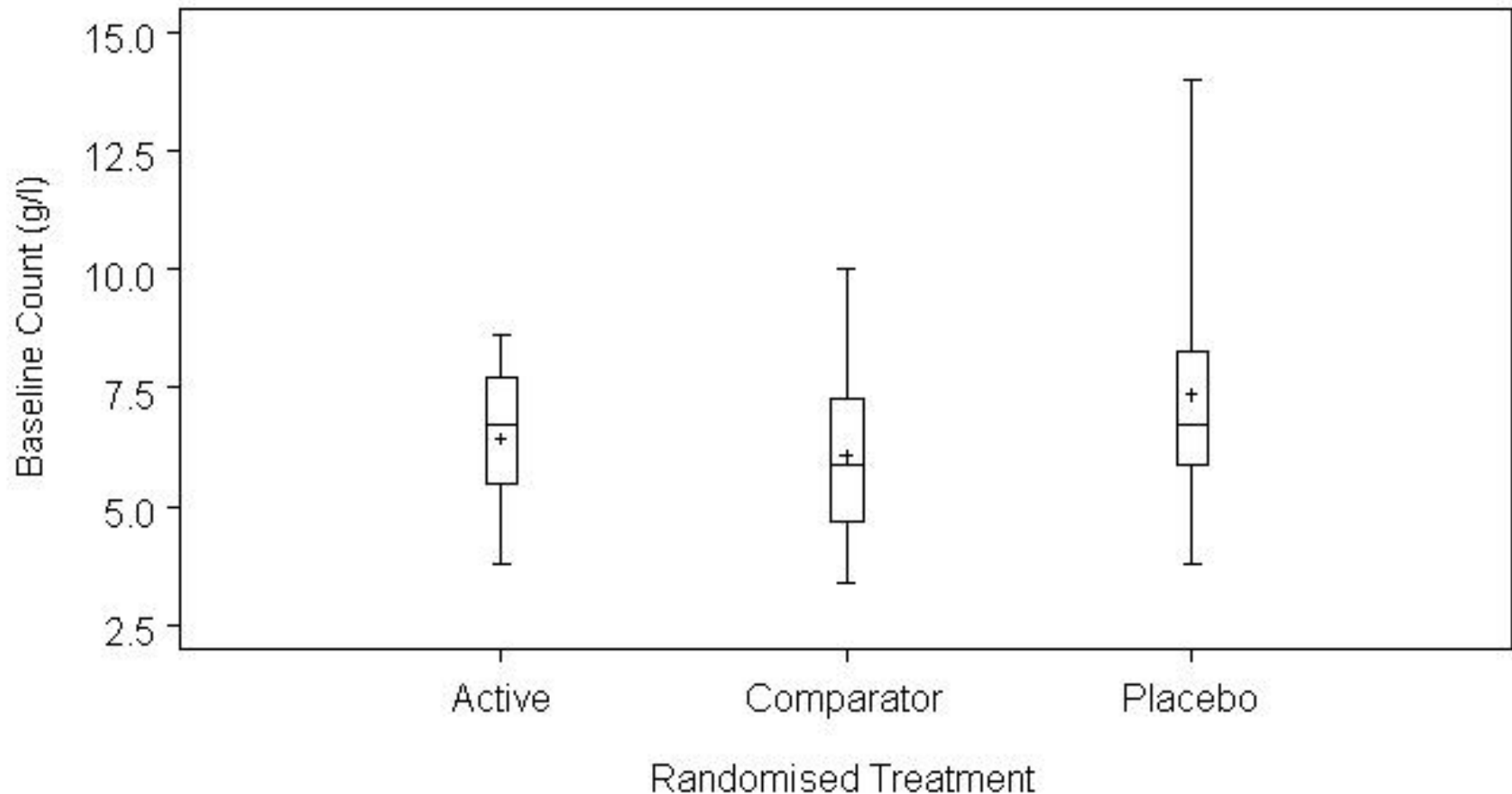
```
proc sort data=boxplot.wbc out=work.skeletal;  
  by trtcd;  
run;
```

* Generate skeletal plot for the BASELINE analysis variable;

```
proc boxplot data=work.skeletal (where=(day=1)) ;  
  plot baseline*trtcd;  
run;
```

The SAS System's Proc Boxplot

White Blood Cell Count at Day 1 (g/l)
Skeletal Box-and-Whisker Plot



The SAS System's Proc Boxplot

- Producing a SCHEMATIC boxplot

- Also:

- Identify outlying observations
 - Use the subject number to identify outlying observations
 - Draw box widths in proportion with sample size

```
proc boxplot data=work.schematic(where=(trtcd=1)) ;  
  plot supdia*day / boxstyle=schematicid  
                    boxwidthscale=0.5;  
  id subjid;  
run;
```

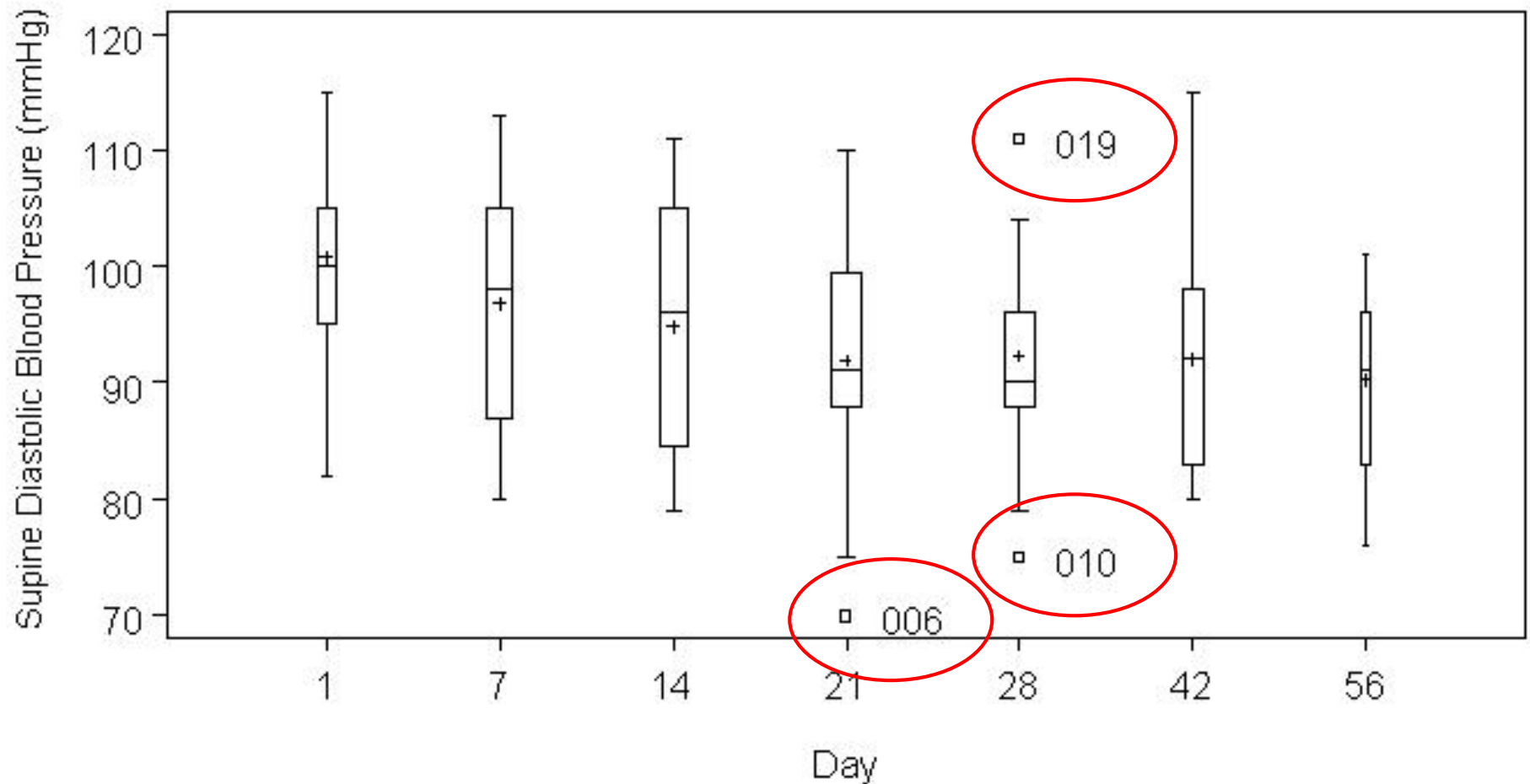
Specifies the ID
data

Draws box widths
proportional to square
root of sample size

Requests
SCHEMATIC style
with ID's

The SAS System's Proc Boxplot

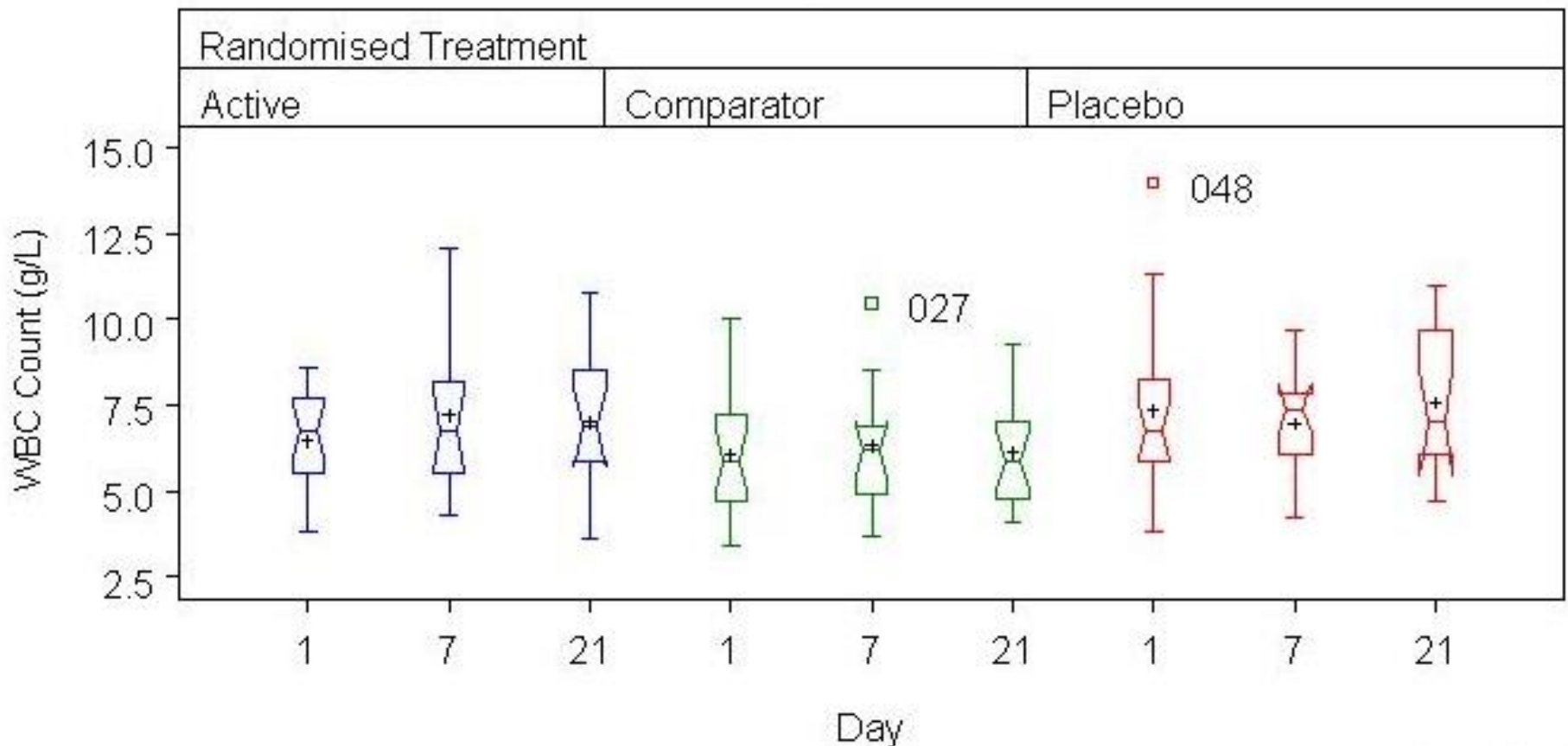
Supine Diastolic Blood Pressure for Active Group (mmHg)
Schematic Box-and-Whisker Plot



The SAS System's Proc Boxplot

- Grouping Boxplots within Blocks**

White Blood Cell Count Over Time, Blocked by Treatment
Schematic ID Style with Notches



The SAS System's Proc Boxplot

- **Grouping Boxplots within Blocks**

- An additional blocking variable can be specified
- Allows data to be presented side-by-side in groups
- Syntax for blocking is as follows:

PLOT *analysis-variable* * *group-variable(block-variable)* / options ;

- To produce the previous figure the code required is expected to be:

```
PROC BOXPLOT data=...;  
  PLOT result * day (trtcd) / options ;  
  ID subjid;  
RUN;
```

**However when DAY and TRTCD
are numeric this produces...**



The SAS System's Proc Boxplot

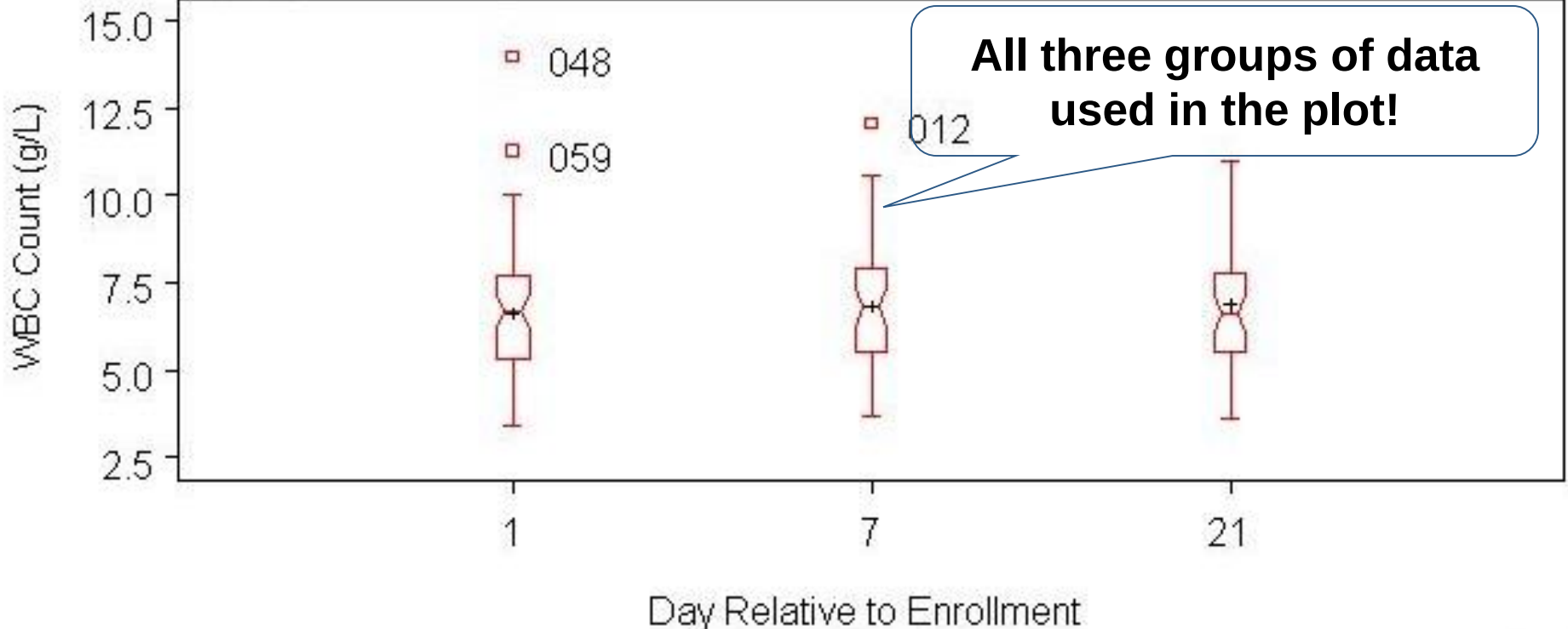
**Only one block
group shown**

White Blood Cell Count Over Time, Blocked by Treatment
Schematic ID Style with Notches

Randomised Treatment

Active

**All three groups of data
used in the plot!**



The SAS System's Proc Boxplot

- To avoid misrepresenting data we must understand how to construct the input data set:
 - Convert the GROUP variable to character – avoiding the requirement to sort on this variable
 - The BLOCK variable is promoted to the first sort variable on the BY statement of PROC SORT
 - To ensure the GROUP variables' data are correctly consolidated in the plot – sort on the character GROUP variable within the BLOCK variable

The SAS System's Proc Boxplot

- Correct Code for Blocking

```
proc sort data=boxplot.wbc out=work.wbc;  
  by trtcd dayc ;  
run;
```

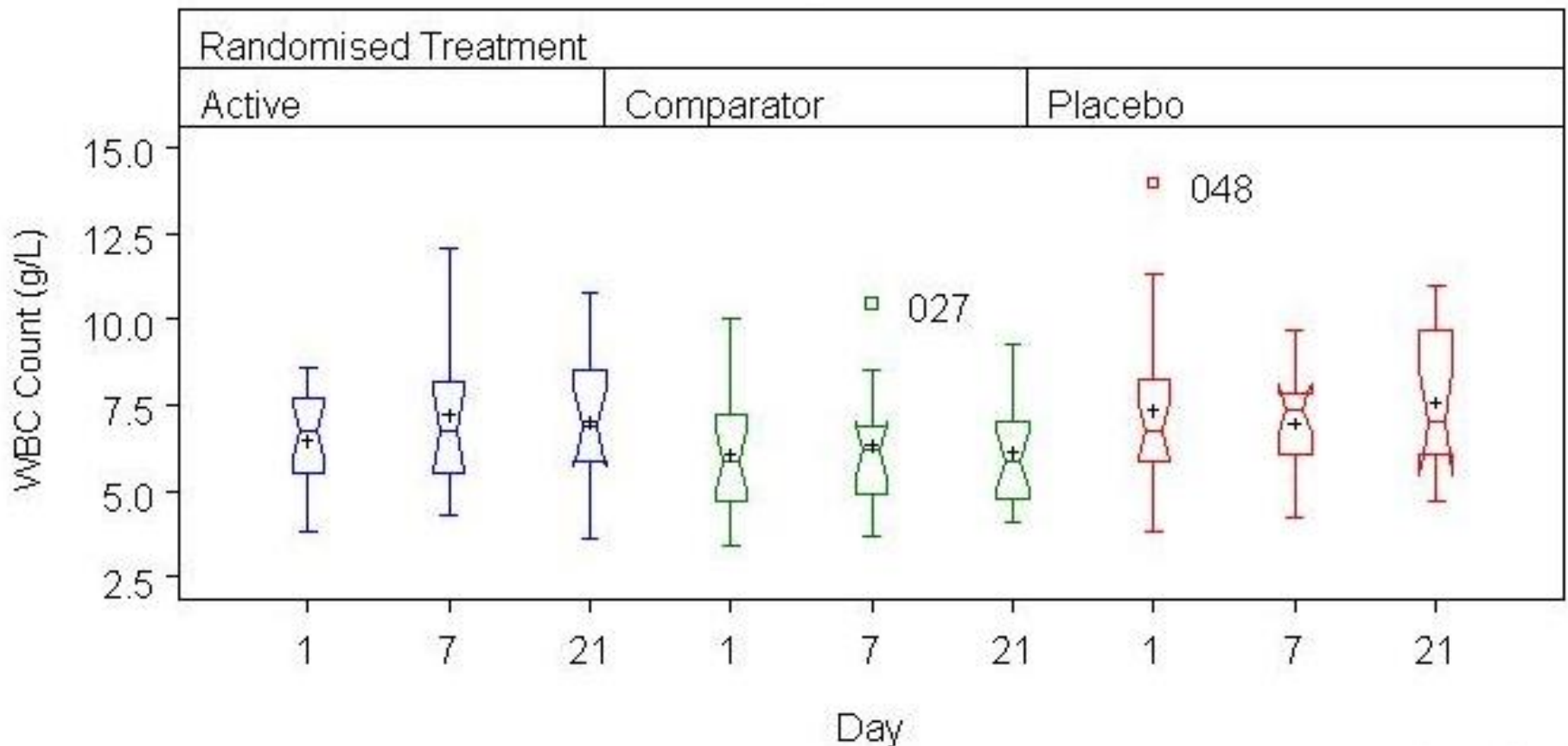
```
proc boxplot data=work.wbc;  
  plot result * dayc(trtcd) / options;  
run;
```

Must be character

The SAS System's Proc Boxplot

- Grouping Boxplots within Blocks**

White Blood Cell Count Over Time, Blocked by Treatment
Schematic ID Style with Notches



Conclusions

- **Boxplots are a useful means of exploring and presenting data**
- **The BOXPLOT procedure in SAS provides the tools needed to create high resolution box-and-whisker plots through standard SAS code**
 - **As all SAS GRAPH statements are honoured by the BOXPLOT procedure – the presentation can be extensively customised**
- **Users must be aware of the implications of not sorting the input data set – particularly when blocking data**



Questions on Box-and-Whisker Plots with The SAS System?

David Shannon

Senior Consultant

Amadeus Software Limited

Email: david.shannon@amadeus.co.uk

Web: www.amadeus.co.uk

