Chapter 201

Descriptive Statistics – Summary Tables

Introduction

This procedure is used to summarize continuous data. Large volumes of such data may be easily summarized in statistical tables of means, counts, standard deviations, etc. Categorical group variables may be used to calculate summaries for individual groups. The tables are similar in structure to those produced by cross tabulation.

This procedure produces tables of the following summary statistics:

- **Count**
- **Missing Count**
- **Sum**
- **Mean**
- **Standard Deviation (Std Dev)**
- **Standard Error (Std Error)**
- **Lower 95% Confidence Limit for the Mean (95% LCL)**
- **Upper 95% Confidence Limit for the Mean (95% UCL)**
- **Median**
- **Minimum**
- **Maximum**
- **Range**
- **Interquartile Range (IQR)**
- **10th Percentile (10th Pctile)**
- **25th Percentile (25th Pctile)**
- **75th Percentile (75th Pctile)**
- **90th Percentile (90th Pctile)**
- **Variance**
- **Mean Absolute Deviation (MAD)**
- **Mean Absolute Deviation from the Median (MADM)**
- **Coefficient of Variation (COV)**
- **Coefficient of Dispersion (COD)**
- **Skewness**
- **Kurtosis**

Types of Categorical Variables

Note that we will refer to two types of categorical variables: *Group Variables* and *Break Variables*.

The values of a *Group Variable* are used to define the rows, sub rows, and columns of the summary table. Up to two Group Variables may be used per table. Group Variables are not required.

Break Variables are used to split a database into subgroups. A separate report is generated for each unique set of values of the break variables.

Data Structure

The data below are a subset of the Resale dataset provided with the software. This (computer simulated) data gives the selling price, the number of bedrooms, the total square footage (finished and unfinished), and the size of the lots for 150 residential properties sold during the last four months in two states. This data is representative of the type of data that may be analyzed with this procedure. Only the first 8 of the 150 observations are displayed.

Resale dataset (subset)

State	Price	Bedrooms	TotalSqft	LotSize
Nev	260000	2	2042	10173
Nev	66900	3	1392	13069
Vir	127900	2	1792	7065
Nev	181900	3	2645	8484
Nev	262100	2	2613	8355
Nev	147500	2	1935	7056
Nev	167200	2	1278	6116
Nev	395700	2	1455	14422

Missing Values

Observations with missing values in either the group variables or the continuous data variables are ignored. The procedure also allows you to specify up to 5 additional values to be considered as missing in categorical group variables.

Summary Statistics

The following sections outline the summary statistics that are available in this procedure.

Count

The number of non-missing data values, *n*. If no frequency variable was specified, this is the number of rows with non-missing values.

Missing Count

The number of missing data values. If no frequency variable was specified, this is the number of rows with missing values.

Sum

The sum (or total) of the data values.

$$Sum = \sum_{i=1}^{n} x_i$$

Mean

The average of the data values.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Variance

The sample variance, s^2, is a popular measure of dispersion. It is an average of the squared deviations from the mean.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Standard Deviation (Std Dev)

The sample standard deviation, *s*, is a popular measure of dispersion. It measures the average distance between a single observation and the mean. It is equal to the square root of the sample variance.

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Standard Error (Std Error)

The standard error of the mean, a measure of the variation of the sample mean about the population mean, is computed by dividing the sample standard deviation by the square root of the sample size.

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

95% Confidence Interval for the Mean (95% LCL & 95% UCL)

This is the upper and lower values of a 95% confidence interval estimate for the mean based on a *t* distribution with *n* – 1 degrees of freedom. This interval estimate assumes that the population standard deviation is not known and that the data for this variable are normally distributed.

$$95\%\ \mathrm{CI} = \bar{x} \pm t_{a/2,n-1} s_{\bar{x}}$$

Minimum

The smallest data value.

Maximum

The largest data value.

Range

The difference between the largest and smallest data values.

$$Range = Maximum - Minimum$$

Percentiles

The $100p^{th}$ percentile is the value below which $100p\%$ of data values may be found (and above which $100p\%$ of data values may be found).The $100p^{th}$ percentile is computed as

$$Z_{100p} = (1\text{-}g)X_{[k1]} + gX_{[k2]}$$

where $k1$ equals the integer part of $p(n+1)$, $k2=k1+1$, g is the fractional part of $p(n+1)$, and $X_{[k]}$ is the k^{th} observation when the data are sorted from lowest to highest.

Median

The median (or 50th percentile) is the “middle number” of the sorted data values.

$$Median = Z_{50}$$

Interquartile Range (IQR)

The difference between the 75th and 25th percentiles (the 3rd and 1st quartiles). This represents the range of the middle 50% of the data. It serves as a robust measure of the variation in the data.

$$IQR = Z_{75} - Z_{25}$$

Mean Absolute Deviation (MAD)

A measure of dispersion that is not affected by outliers as much as the standard deviation and variance. It measures the average absolute distance between a single observation and the mean.

$$MAD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

Mean Absolute Deviation from the Median (MADM)

A measure of dispersion that is even more robust to outliers than the mean absolute deviation (MAD) since the median is used as the center point of the distribution. It measures the average absolute distance between a single observation and the median.

$$MADM = \frac{\sum_{i=1}^{n} | x_i - Median |}{n}$$

Coefficient of Variation (COV)

A relative measure of dispersion used to compare the amount of variation in two samples. It is calculated by dividing the standard deviation by the mean. Sometimes it is referred to as COV or CV.

$$COV = \frac{s}{\overline{x}}$$

Coefficient of Dispersion (COD)

A robust, relative measure of dispersion. It is calculated by dividing the robust mean absolute deviation from the median (MADM) by the median. It is frequently used in real estate or tax assessment applications.

$$COD = \frac{MADM}{Median} = \frac{\left(\dfrac{\sum_{i=1}^{n} | x_i - Median |}{n} \right)}{Median}$$

Skewness

Measures the direction and degree of asymmetry in the data distribution.

$$Skewness = \frac{m_3}{m_2^{3/2}}$$

where

$$m_r = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^r}{n}$$

Kurtosis

Measures the heaviness of the tails in the data distribution.

$$Kurtosis = \frac{m_4}{m_2^2}$$

where

$$m_r = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^r}{n}$$

Example 1 – Basic Variable Summary Report (No Group Variables)

The data used in this example are in the Resale dataset.

Setup

To run this example, complete the following steps:

1 Open the Resale example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Resale** and click **OK**.

2 Specify the Descriptive Statistics – Summary Tables procedure options

- Find and open the **Descriptive Statistics – Summary Tables** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1a** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
Data Variable(s)	**Price, Bedrooms, Bathrooms, Garage, TotalSqft**
Statistics	**Count, Mean, Std Dev, 95% LCL, 95% UCL**
Report Options (*in the Toolbar*)	
Variable Labels	**Column Names**

3 Run the procedure

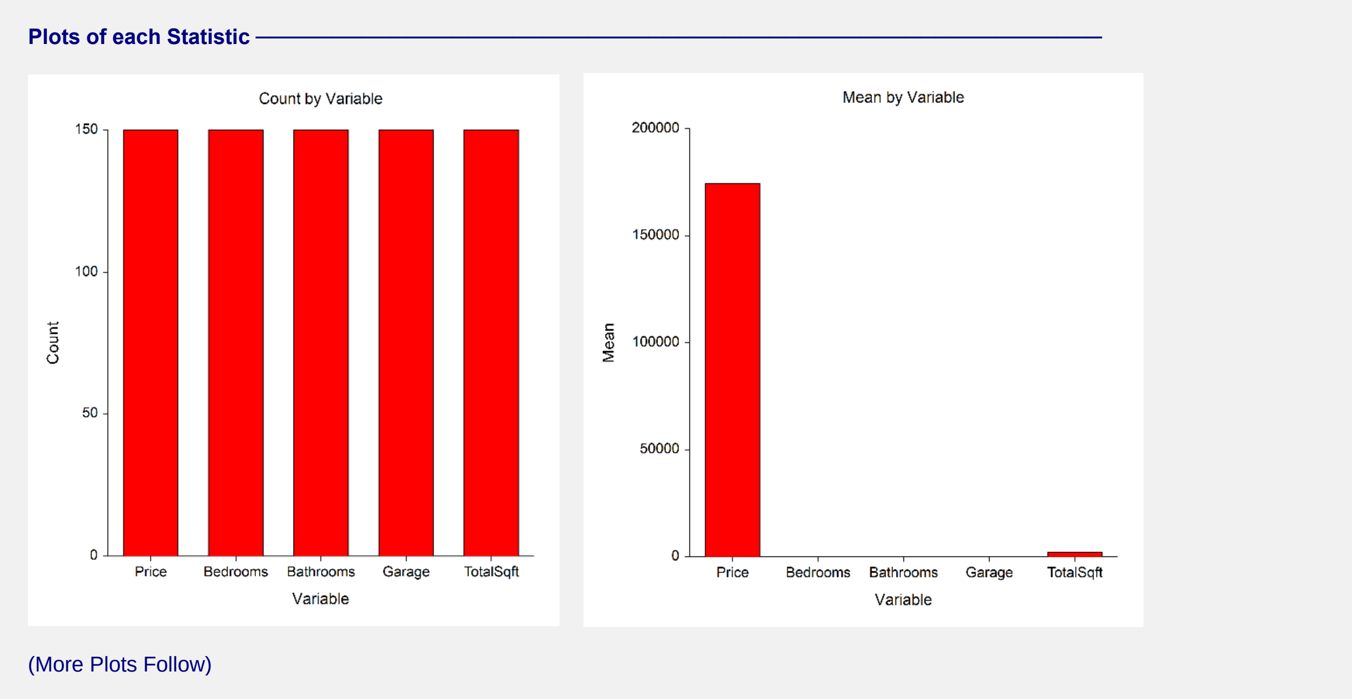
- Click the **Run** button to perform the calculations and generate the output.

Summary Table

```
Summary Table ─────────────────────────────────────────────────────────────

                                               Variable
Statistic
                        Price    Bedrooms    Bathrooms       Garage    TotalSqft
Count                     150         150          150          150          150
Mean                   174392        2.42          2.4     1.266667      1893.38
Standard Deviation   97656.81   0.8919476    0.8047677    0.5636252     754.2496
Lower 95% CL Mean      158636    2.276093     2.270158     1.175731     1771.689
Upper 95% CL Mean      190148    2.563908     2.529842     1.357602     2015.071
```

The table is created with the statistics as rows and the data variables as columns when the positions are both set to "Auto".

Plots of Each Statistic



Plots of each Statistic

(More Plots Follow)

The plots are not very informative because the variables have vastly different scales.

Example 1b – Adjust Item Table Positions (Data Variables in Rows and Statistics in Columns)

To rotate the table, all we have to do is change the position of one of the items. To do this, change **Data Variable(s) Position** to **Rows** and run the procedure again to get the results.

4 Modify the Data Variable(s) Position

- The settings for this section are listed below and are stored in the **Example 1b** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
Data Variable(s) Position	**Rows**

5 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

```
Summary Table ────────────────────────────────────────────────────────────────────────

                                        Statistic

Variable                                 Standard       Lower 95%      Upper 95%
                  Count         Mean    Deviation         CL Mean        CL Mean
Price               150       174392     97656.81          158636         190148
Bedrooms            150         2.42    0.8919476        2.276093       2.563908
Bathrooms           150          2.4    0.8047677        2.270158       2.529842
Garage              150     1.266667    0.5636252        1.175731       1.357602
TotalSqft           150      1893.38     754.2496        1771.689       2015.071
```

The table is now rotated with the data variables as rows and the statistics as columns. Notice that the actual summary statistic values are exactly the same.

Example 2 – Variable Summary Report (One Group Variable)

The data used in this example are in the Resale dataset.

Setup

To run this example, complete the following steps:

1 **Open the Resale example dataset**

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Resale** and click **OK**.

2 **Specify the Descriptive Statistics – Summary Tables procedure options**

- Find and open the **Descriptive Statistics – Summary Tables** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 2a** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
Data Variable(s)	**Price, TotalSqft, LotSize**
Statistics	**Count, Mean, Std Dev**
Include Group Variable 1	**Checked**
Variables	**State**
Report Options (*in the Toolbar*)	
Variable Labels	**Column Labels**
Data Labels	**Value Labels**

3 **Run the procedure**

- Click the **Run** button to perform the calculations and generate the output.

Summary Table

```
Summary Table ─────────────────────────────────────────────────────────

                                         Variable

State                             Sales     Total Area     Lot Size
         Statistic                Price         (Sqft)       (Sqft)
Nevada   Count                       88             88           88
         Mean                  170762.5        1881.33     8571.454
         Standard Deviation    98665.72        788.569      2419.88

Virginia Count                       62             62           62
         Mean                  179543.5       1910.484     8076.597
         Standard Deviation    96771.49       708.6572     2301.226


Total    Count                      150            150          150
         Mean                    174392        1893.38     8366.913
         Standard Deviation    97656.81       754.2496     2376.334
```

The table displays the group variable values as the rows, the statistics as the subrows, and the data variables as the columns. The plots are not shown because they are not very informative because the variables have vastly different scales. Totals are given for the group variable.

Example 2b – Adjust Item Table Positions (Data Variables in Rows, Statistics in Sub Rows, and Group Variable in Columns)

To rotate the table, all we have to do is change the position of one of the items. To do this, change **Data Variable(s) Position** to **Rows** and run the procedure again to get the results.

4 Modify the Data Variable(s) Position

- The settings for this section are listed below and are stored in the **Example 2b** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
Data Variable(s) Position	**Rows**

5 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

```
Summary Table ─────────────────────────────────────────────────────────

                                            State
Variable
                  Statistic              Nevada     Virginia       Total
Sales Price       Count                      88           62         150
                  Mean                 170762.5     179543.5      174392
                  Standard Deviation   98665.72     96771.49    97656.81

Total Area (Sqft) Count                      88           62         150
                  Mean                  1881.33     1910.484     1893.38
                  Standard Deviation    788.569     708.6572    754.2496

Lot Size (Sqft)   Count                      88           62         150
                  Mean                 8571.454     8076.597    8366.913
                  Standard Deviation    2419.88     2301.226    2376.334
```

The table is now rotated with the data variables as rows and the group variable values as columns. Notice that the actual summary statistic values are exactly the same.

Example 2c – Adjust Item Table Positions (Data Variables in Rows, Group Variable in Sub Rows, and Statistics in Columns)

To change the table so that statistics are presented as columns with the group variable as subrows and the data variables as rows, change the position of **Statistics** to **Columns** with the position for Data Variable(s) still set to Rows and run the procedure again to get the results.

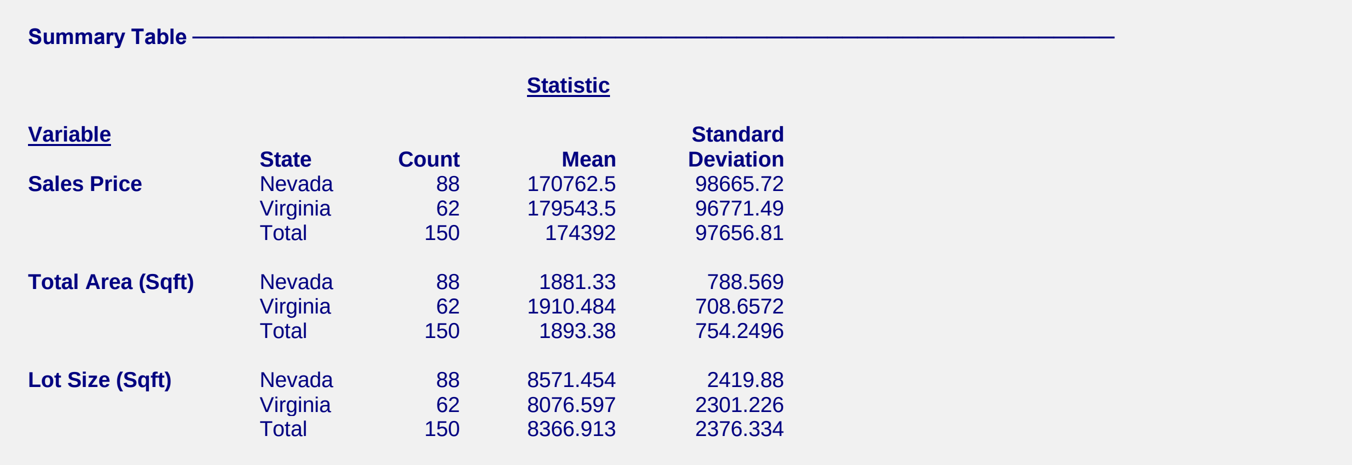
6 Modify the Statistics Position

- The settings for this section are listed below and are stored in the **Example 2b** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
Statistics Position....................................	**Columns**

7 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.



```
Summary Table ─────────────────────────────────────────────────────────

                                        Statistic

Variable                                                   Standard
                        State      Count        Mean      Deviation
Sales Price             Nevada        88    170762.5       98665.72
                        Virginia      62    179543.5       96771.49
                        Total        150      174392       97656.81

Total Area (Sqft)       Nevada        88     1881.33        788.569
                        Virginia      62    1910.484       708.6572
                        Total        150     1893.38       754.2496

Lot Size (Sqft)         Nevada        88    8571.454        2419.88
                        Virginia      62    8076.597       2301.226
                        Total        150    8366.913       2376.334
```

The table now has the data variables as rows and the group variable values as subrows with the statistics as columns.

Example 3 – Variable Summary Report (Two Group Variables)

The data used in this example are in the Pain dataset. In this example we will show you how to make even more customizations to adjust the appearance of the tables and plots and how easy it is to make position adjustments.

Setup

To run this example, complete the following steps:

1 **Open the Pain example dataset**

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Pain** and click **OK**.

2 **Specify the Descriptive Statistics – Summary Tables procedure options**

- Find and open the **Descriptive Statistics – Summary Tables** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 3a** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
Data Variable(s)	**Pain**
Statistics	**Mean, Minimum, 25th Pctile, Median, 75th Pctile, Maximum**
Include Group Variable 1	**Checked**
Variables	**Drug**
Include Group Variable 2	**Checked**
Variables	**Time**
Report Options Tab	
Display Group Variable Marginal Totals on the Summary Tables	**Unchecked**
Use Short Statistical Names on Reports and Plots	**Checked**
Sum, Mean, CI Limits	**2**
Plots Tab	
Separate Plots	
Show Bar Charts	**Checked**
Show Line Charts	**Unchecked**
Bar Chart Format (*Click the Button*)	
Numeric Axis Tab	
Max (Boundaries)	**100**
Combined Plots	
Show Bar Charts	**Checked**
Show Line Charts	**Unchecked**

Bar Chart Format (*Click the Button*)

Group Axis Tab

Lower Axis Tick Label Layout (*Click the Button*)
Alignment .... **Right**
Rotation Angle.................................. **45**
Margin Above the Text..................... **10**

3 Run the procedure

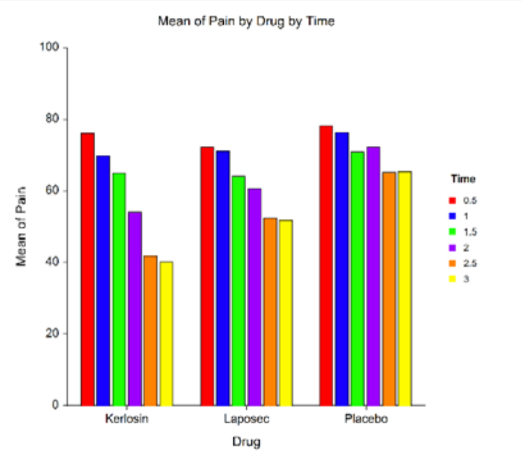
- Click the **Run** button to perform the calculations and generate the output.

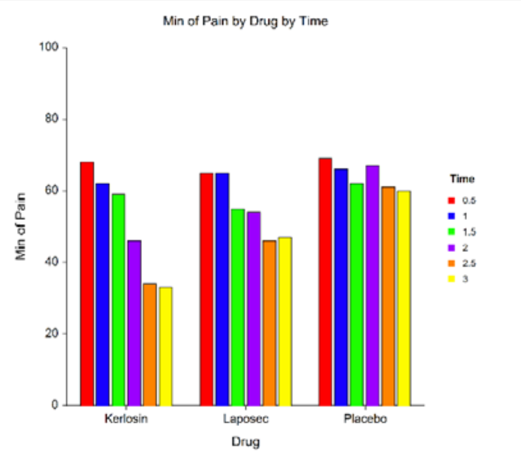
Output

```
Summary Table of Pain ─────────────────────────────────────────────────

                                           Time
Drug
             Statistic     0.5        1      1.5        2      2.5        3
Kerlosin     Mean        76.00    69.86    65.00    54.00    41.71    40.14
             Min            68       62       59       46       34       33
             25th Pctile    71       64       61       47       36       37
             Median         77       68       65       56       43       41
             75th Pctile    81       77       69       57       46       45
             Max            83       80       72       64       48       46

Laposec      Mean        72.29    71.00    64.00    60.57    52.43    51.71
             Min            65       65       55       54       46       47
             25th Pctile    68       67       60       56       47       49
             Median         71       71       65       61       53       50
             75th Pctile    78       77       68       65       56       56
             Max            79       78       70       66       58       58

Placebo      Mean        78.14    76.29    70.86    72.14    65.14    65.43
             Min            69       66       62       67       61       60
             25th Pctile    71       69       64       70       63       62
             Median         79       78       70       73       65       67
             75th Pctile    87       82       76       73       68       68
             Max            87       84       81       76       69       70
```

The table is displays Group Variable 1 (Drug) values as the rows, the statistics as the subrows, and Group Variable 2 (Time) values as the columns.

Plots of each Statistic for Pain

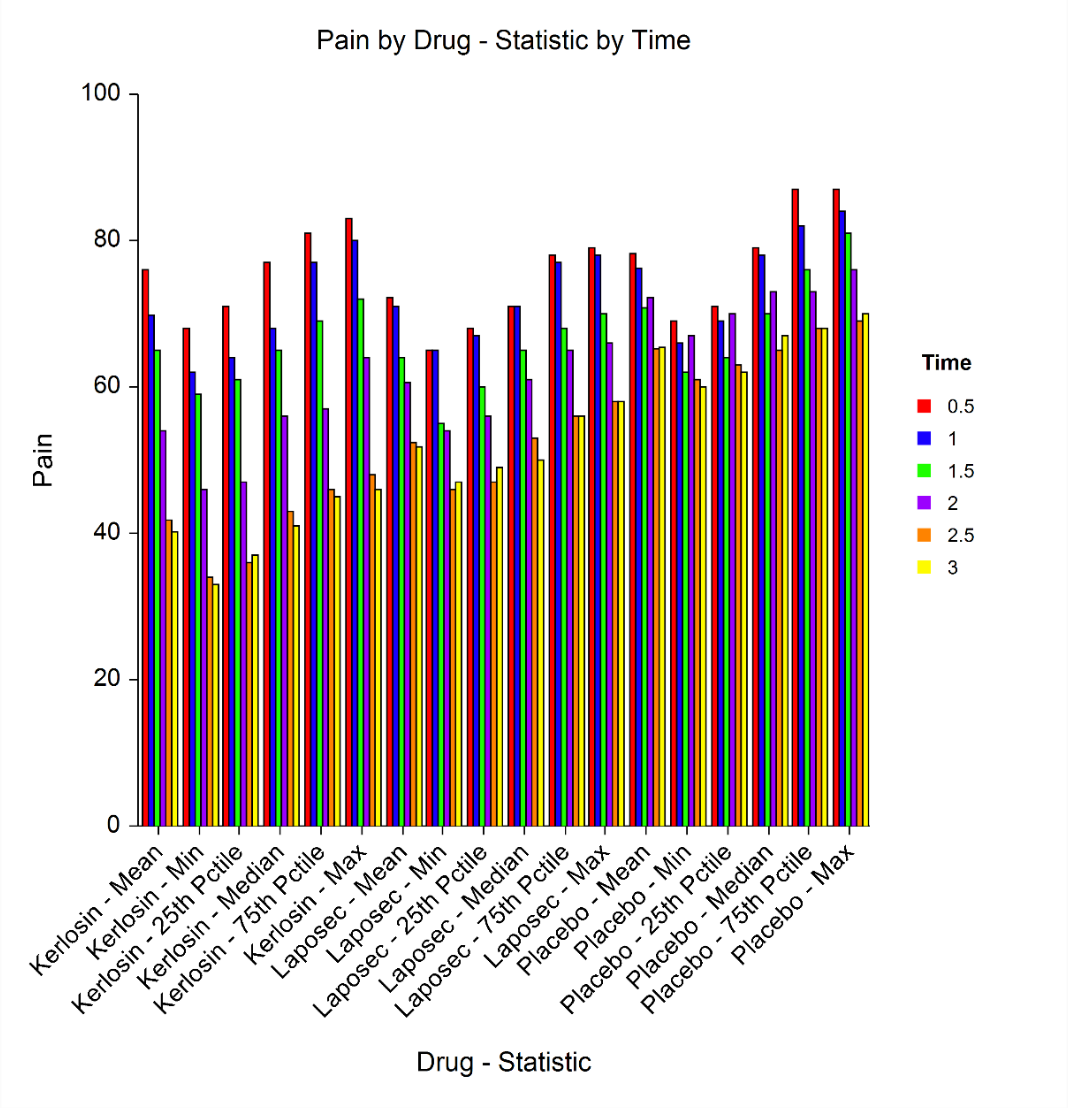




(More Plots Follow)

Individual plots are created with the table row item (Group Variable 1 --- “Drug”) on the group (X) axis and the table column item (Group Variable 2 --- “Time”) as the legend variable. A separate plot is created for each statistic. These plots are very useful for seeing overall trends. From the plots shown here, it is apparent that the average and minimum pain response is lower for both drugs than for placebo and that the pain control is better over time. Kerlosin appears to control pain the best from these results. Statistical tests would need to be performed, however, to assert statistical significance in the differences.

Combined Plot of Pain



The combined plot displays all of the information in the table. We rotated the group axis labels so they would not overlap and be readable. The table row item (Group Variable 1 --- “Drug”) and table sub row item (Statistic) are combined on the group (X) axis. The table column item (Group Variable 2 --- “Time”) is the legend variable.

Example 3b – Adjust Item Table Positions (Group 2 Variable in Rows, Group 1 Variable in Sub Rows, and Statistics in Columns)

To change the orientation on the tables and plots, simply change the position the items. We will display the **Statistics as the columns** and **Time as the rows**. This will put **Drug as the sub row**.

4 Modify the Statistics and Group Variable 2 Positions

- The settings for this section are listed below and are stored in the **Example 3b** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
Statistics Position....................................	**Columns**
Group Variable 2 Position.......................	**Rows**

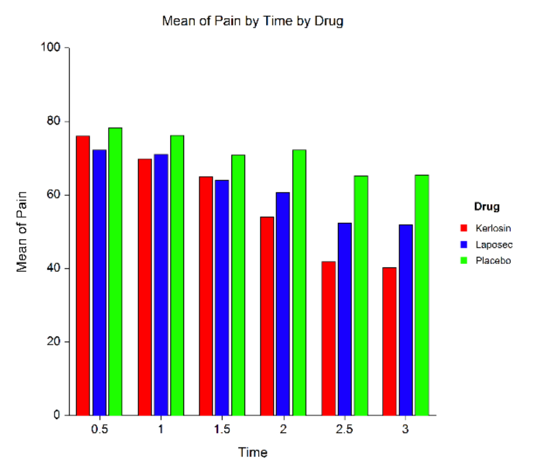
5 Run the procedure

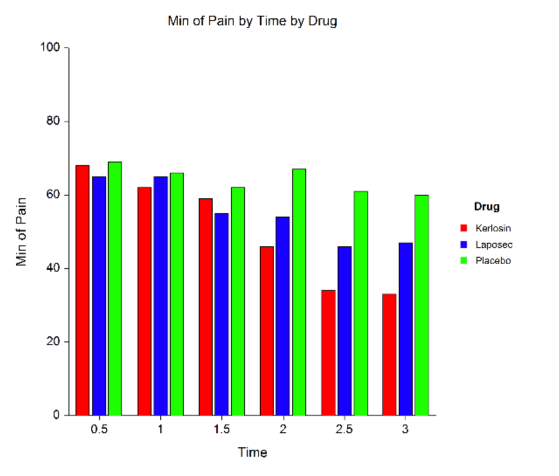
- Click the **Run** button to perform the calculations and generate the output.

```
Summary Table of Pain ──────────────────────────────────────────────────

                                   Statistic

Time                                   25th                75th
        Drug        Mean      Min     Pctile     Median    Pctile     Max
0.5     Kerlosin    76.00      68        71         77        81       83
        Laposec     72.29      65        68         71        78       79
        Placebo     78.14      69        71         79        87       87

1       Kerlosin    69.86      62        64         68        77       80
        Laposec     71.00      65        67         71        77       78
        Placebo     76.29      66        69         78        82       84

1.5     Kerlosin    65.00      59        61         65        69       72
        Laposec     64.00      55        60         65        68       70
        Placebo     70.86      62        64         70        76       81

2       Kerlosin    54.00      46        47         56        57       64
        Laposec     60.57      54        56         61        65       66
        Placebo     72.14      67        70         73        73       76

2.5     Kerlosin    41.71      34        36         43        46       48
        Laposec     52.43      46        47         53        56       58
        Placebo     65.14      61        63         65        68       69

3       Kerlosin    40.14      33        37         41        45       46
        Laposec     51.71      47        49         50        56       58
        Placebo     65.43      60        62         67        68       70
```

The table is displays Group Variable 2 (Time) values as the rows, Group Variable 1 (Drug) values as the subrows, and the statistics as the columns.

Plots of each Statistic for Pain

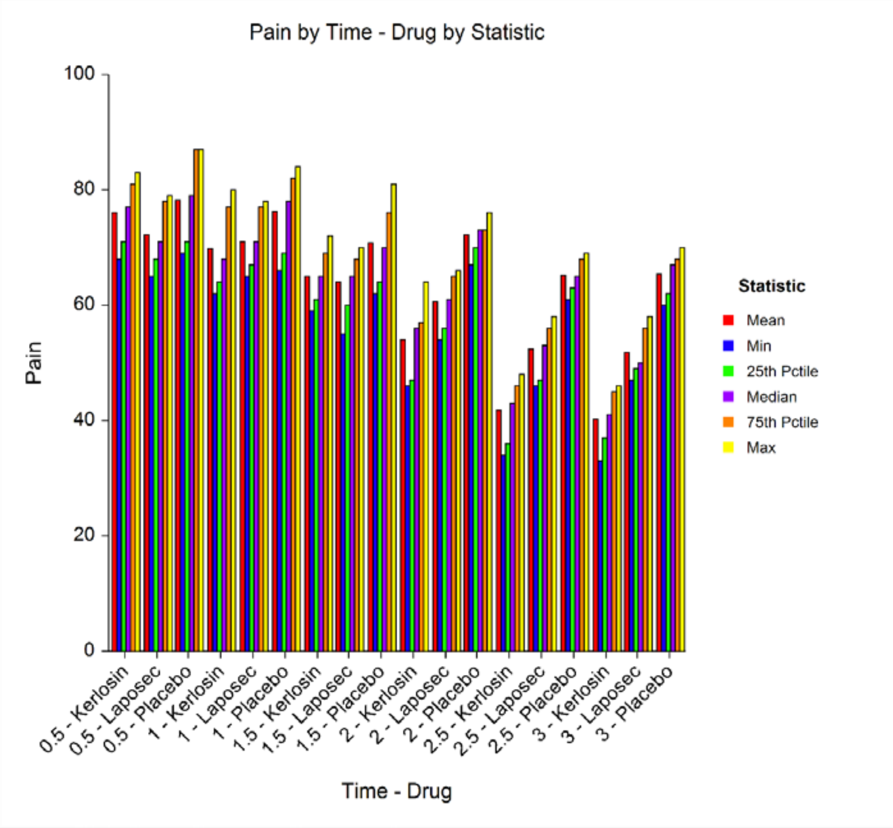




(More Statistic Plots Follow)

The individual plots are different now with the table row item (Group Variable 2 --- “Time”) on the group (X) axis and the table column item (Group Variable 1 --- “Drug”) as the legend variable. A separate plot is created for each statistic. These plots are again useful for seeing overall trends. There is a very distinct reduction in pain over time.

Combined Plot of Pain



Again, the combined plot displays all of the information in the table. The table row item (Group Variable 2 --- “Time”) and table sub row item (Group Variable 1 --- “Drug”) are combined on the group (X) axis. The table column item (Statistic) is the legend variable.

Example 3c – Adjust Item Table Positions (Creating a Separate Table for each Data Variable and Statistic Combination)

It is easy to create a separate table for each data variable and statistic combination (this can only be done when there is at least one group variable). We will display a separate table for each statistic with **Time as the rows** and **Drug as the columns**. There will be no sub row item.

6 Modify the Data Variable(s), Statistics, and Group Variable 2 Positions

- The settings for this section are listed below and are stored in the **Example 3c** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
Data Variable(s) Position	**Tables**
Statistics Position	**Tables**
Group Variable 2 Position	**Rows**

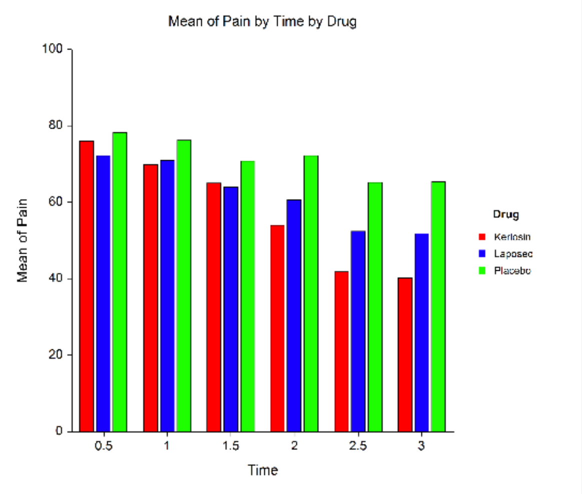
7 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Summary Table of Mean of Pain

Time	Drug Kerlosin	Laposec	Placebo
0.5	76.00	72.29	78.14
1	69.86	71.00	76.29
1.5	65.00	64.00	70.86
2	54.00	60.57	72.14
2.5	41.71	52.43	65.14
3	40.14	51.71	65.43

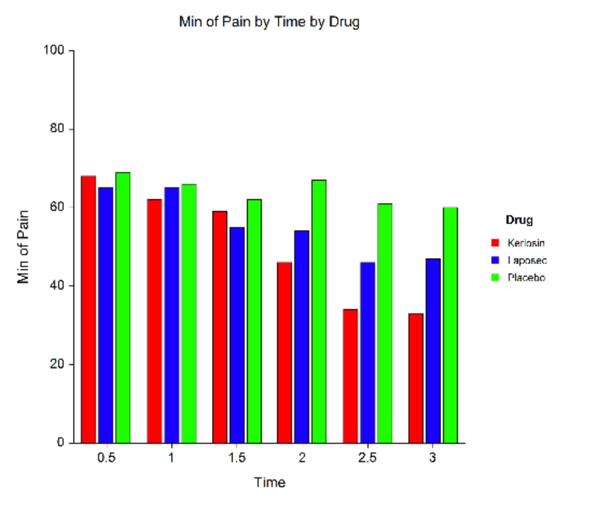
Plot of Mean of Pain



Summary Table of Min of Pain

Time	Drug		
	Kerlosin	Laposec	Placebo
0.5	68	65	69
1	62	65	66
1.5	59	55	62
2	46	54	67
2.5	34	46	61
3	33	47	60

Plot of Min of Pain



(Report continues with table and plot for each Data Variable/Statistic combination)

A separate table is created for each Data Variable/Statistic combination. If more than one data variable were entered, the report would be even longer. There is no combined plot in the output because the combined plot is the same as the individual plot in this case.