Chapter 209

Descriptive Statistics – Summary Lists

Introduction

This procedure is used to summarize continuous data. Large volumes of such data may be easily summarized in statistical lists of means, counts, standard deviations, etc. Up to 8 categorical group variables may be used to calculate summaries for individual group combinations. The summary lists may be output directly to a new dataset.

This procedure produces lists 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**

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

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.

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

Variance

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

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

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.

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

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

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

where *k*1 equals the integer part of *p*(*n*+1), *k*2=*k*1+1, *g* is the fractional part of *p*(*n*+1), and $X_{[k]}$ is the kth 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.

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

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.

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

Skewness

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

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

where

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

Kurtosis

Measures the heaviness of the tails in the data distribution.

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

where

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

Example 1 – Basic Summary List with 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 Lists procedure options
- Find and open the **Descriptive Statistics – Summary Lists** 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**

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

Summary List

Summary List ───

		Statistics			
Variable				**95% LCL**	**95% UCL**
	Count	**Mean**	**SD**	**Mean**	**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 data summary list is presented with each variable represented on a single row.

Plots of Each Statistic

Plots of each Statistic ───

Count by Variable

Mean by Variable

(More Plots Follow)
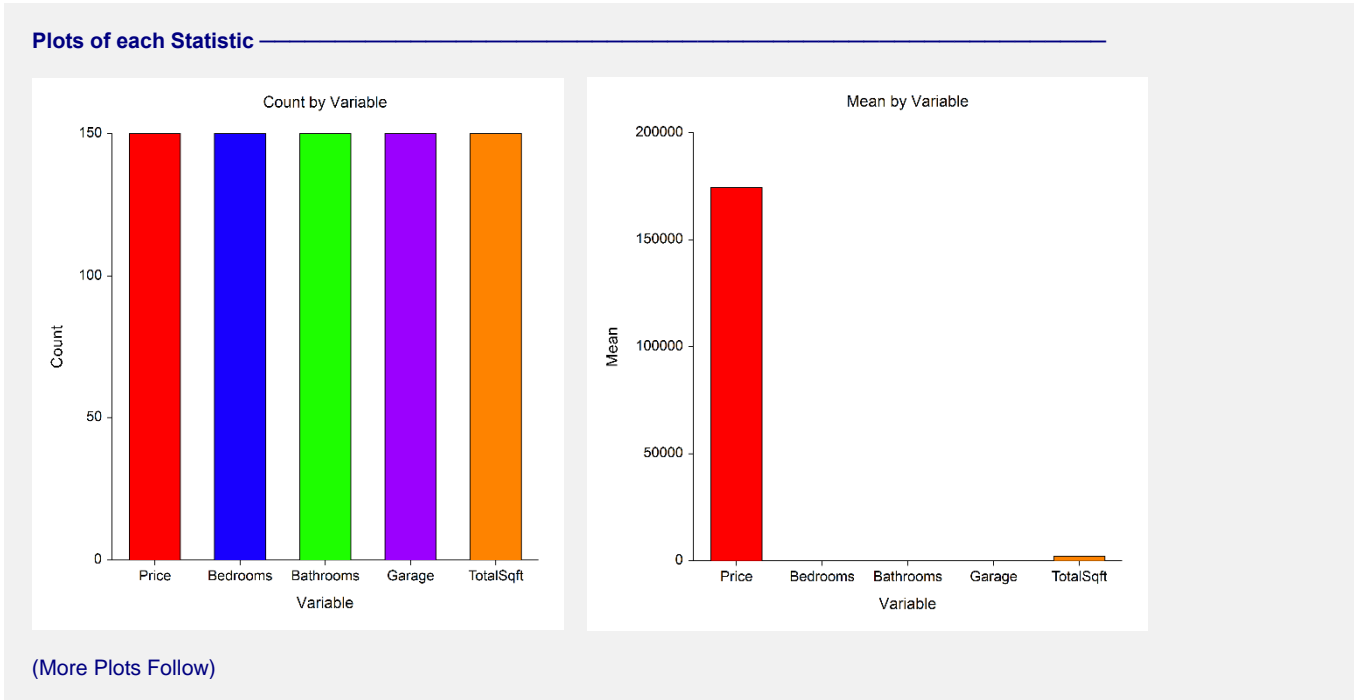
The plots are not very informative because the variables have vastly different scales.

Example 1b – Storing the Summary List in a New NCSS Data File

To store the data summary list in a new data file, check **Store the Summary List in a New NCSS Data File** and run the procedure again to get the results.

Be careful because any unsaved data will be lost!

4   Specify the data storage

- 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	
Store the Summary List in a New NCSS Data File	**Checked**
Output File Name...	**%mydocs_NCSS%\Data\ResaleSummary.NCSS**
Automatically Reopen the Current Dataset after.........	**Unchecked** (so that we can review the summary
the Save Operation Completes	data file that has been created.)

5   Run the procedure

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

Descriptive Statistics – Summary Lists

Summary List Storage Information ───

Output Data File Name: {NCSS Documents Folder}\Data\ResaleSummary.NCSS
Original Raw Data File: {Example Data Folder}\Resale.NCSS

Data Variable(s):   (5) Price, Bedrooms, Bathrooms, Garage, TotalSqft
Group Variable(s):   (0)

Summary Statistic(s):   (5) Count, Mean, SD, 95% LCL Mean, 95% UCL Mean

(Summary List Report and Plots Follow)

The data summary list output data file is described in this report. Go to the data table to see the new summary data file that has been created.

Example 2 – Summary List with One Group Variable

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

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 Lists procedure options
- Find and open the **Descriptive Statistics – Summary Lists** 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)......................................	**Pain**
Group Variable(s)	**Drug**
Statistics	**Mean, Minimum, 25th Pctile, Median, 75th Pctile, Maximum**

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

Summary List

Summary List of Pain ───

Statistics for Pain

Drug	Mean	Minimum	25th Pctile	Median	75th Pctile	Maximum
Kerlosin	57.78571	33	44.5	60	69.5	83
Laposec	62	46	54.75	63.5	68.5	79
Placebo	71.33334	60	66	70	76	87

The summary list is presented with each level of Drug represented on a single row. The various statistics are listed in columns.

Plots of Each Statistic

Plots of each Statistic for Pain ――

(More Plots Follow)
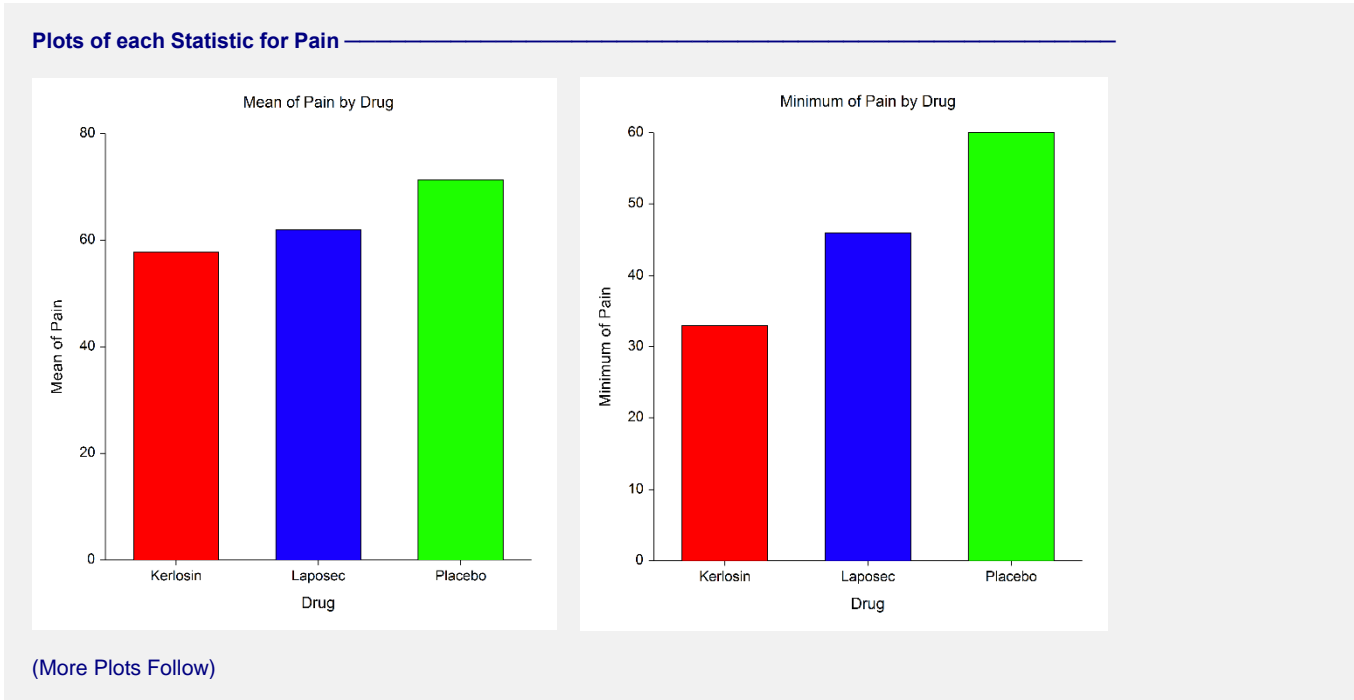
Individual plots for each statistic are created with the Group Variable, "Drug," on the group (X) axis. These plots are very useful for seeing overall trends. From the plots shown here, it is apparent that the average and minimum pain response is lower for both drugs than for placebo. Kerlosin appears to control pain the best from these results. Statistical tests would need to be performed, however, to assert statistical significance in the differences.

Example 2b – Storing the Summary List in a New NCSS Data File

To store the data summary list in a new data file, check **Store the Summary List in a New NCSS Data File** and run the procedure again to get the results.

Be careful because any unsaved data will be lost!

4   Specify the data storage
- 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	
Store the Summary List in a New NCSS Data File	**Checked**
Output File Name...	**%mydocs_NCSS%\Data\ PainDrugSummary.NCSS**
Automatically Reopen the Current Dataset after.......... the Save Operation Completes	**Unchecked** (so that we can review the summary data file that has been created.)

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

Summary List Storage Information ———

Output Data File Name: {NCSS Documents Folder}\Data\PainDrugSummary.NCSS
Original Raw Data File: {Example Data Folder}\Pain.NCSS

Data Variable(s):   (1) Pain
Group Variable(s):   (1) Drug

Summary Statistic(s):   (6) Mean, Minimum, 25th Pctile, Median, 75th Pctile, Maximum

(Summary List Report and Plots Follow)

The data summary list output data file is described in this report. Go to the data table to see the new summary data file that has been created.

Example 3 – Summary List with Two Group Variables

The data used in this example are in the Pain dataset. In this example we'll show you how to make even more customizations to adjust the appearance of the tables and plots.

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 Lists procedure options
- Find and open the **Descriptive Statistics – Summary Lists** 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.

<u>Option</u>   <u>Value</u>

Variables Tab
Data Variable(s)..................................... **Cov, Pain**
Group Variable(s) .... **Drug, Time**
Statistics .... **Count, Mean, Std Dev**

Report Options Tab
Use Short Statistical Names on.............. **Unchecked**
Reports and Plots
Sum, Mean, CI Limits **2**
SD, SE, Var, MAD, MADM **2**

Plots Tab
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.................................. **90**
    Margin Above the Text..................... **10**

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

Output

Summary List of Cov ───

Statistics for Cov

Drug	Time	Count	Mean	Standard Deviation
Kerlosin	0.5	7	179.00	45.41
Kerlosin	1	7	168.71	32.39
Kerlosin	1.5	7	157.14	45.13
Kerlosin	2	7	184.14	27.89
Kerlosin	2.5	7	140.43	21.42
Kerlosin	3	7	150.14	34.16
Laposec	0.5	7	170.43	44.29
Laposec	1	7	172.86	14.70
Laposec	1.5	7	160.71	40.22
Laposec	2	7	160.29	33.33
Laposec	2.5	7	148.57	33.04
Laposec	3	7	154.00	25.65
Placebo	0.5	7	162.14	37.65
Placebo	1	7	144.14	40.58
Placebo	1.5	7	156.29	34.93
Placebo	2	7	150.14	23.96
Placebo	2.5	7	183.14	36.72
Placebo	3	7	156.86	44.89

Plots of each Statistic for Cov ───
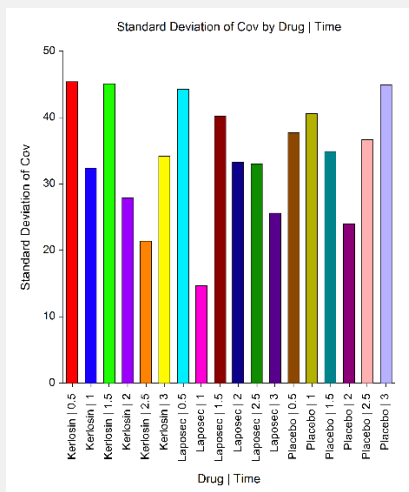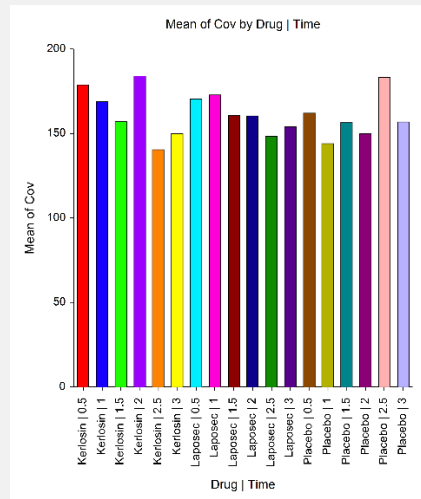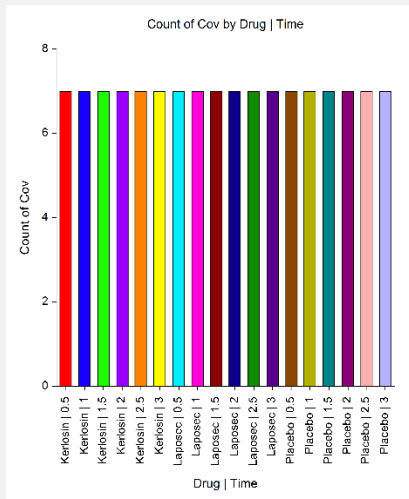
Descriptive Statistics – Summary Lists

Summary List of Pain ───

Statistics for Pain

Drug | Time	Count	Mean	Standard Deviation
Kerlosin | 0.5	7	76.00	5.26
Kerlosin | 1	7	69.86	6.62
Kerlosin | 1.5	7	65.00	4.65
Kerlosin | 2	7	54.00	6.38
Kerlosin | 2.5	7	41.71	5.06
Kerlosin | 3	7	40.14	4.71
Laposec | 0.5	7	72.29	5.22
Laposec | 1	7	71.00	4.93
Laposec | 1.5	7	64.00	5.07
Laposec | 2	7	60.57	4.50
Laposec | 2.5	7	52.43	4.65
Laposec | 3	7	51.71	3.95
Placebo | 0.5	7	78.14	7.22
Placebo | 1	7	76.29	6.60
Placebo | 1.5	7	70.86	6.69
Placebo | 2	7	72.14	2.85
Placebo | 2.5	7	65.14	2.79
Placebo | 3	7	65.43	3.74

Plots of each Statistic for Pain ─────────────────────────────────

Count of Pain by Drug | Time

Mean of Pain by Drug | Time

Standard Deviation of Pain by Drug | Time
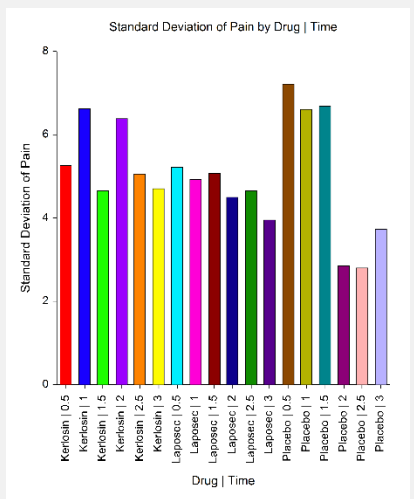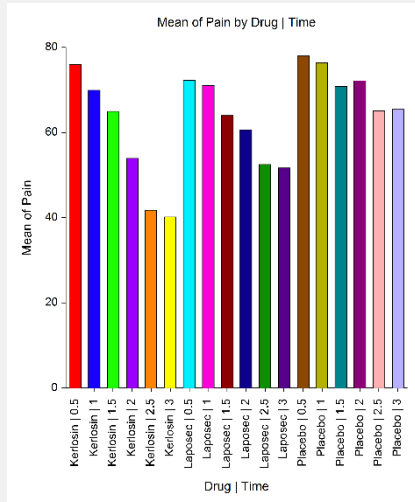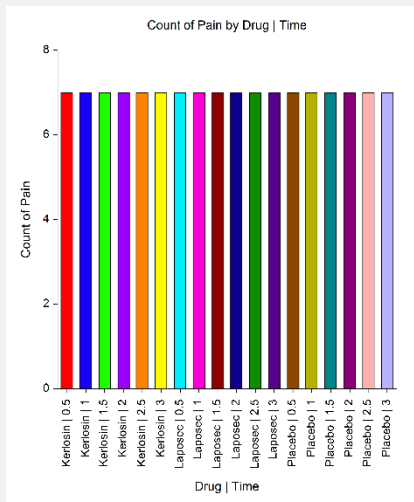
Summary list tables are presented separately for the data variables, Cov and Pain, with each combination of Drug and Time represented on a single row. The various statistics are listed in columns. From the plots shown here, it is apparent that the average and minimum pain response is lower for both drugs 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.

Example 3b – Storing the Summary List in a New NCSS Data File (Row-by-Row)

To store the data summary list in a new data file row-by-row, check **Store the Summary List in a New NCSS Data File** and run the procedure again to get the results.

<u>**Be careful because any unsaved data will be lost!**</u>

4   **Specify the data storage**

- 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.

<u>Option</u>	<u>Value</u>
Variables Tab	
Store the Summary List in a New NCSS Data File	**Checked**
Output File Name..	**%mydocs_NCSS%\Data\ PainDrugTimeSummary1.NCSS**
Data Variable Storage	**Store as Rows**
Automatically Reopen the Current Dataset after.......... the Save Operation Completes	**Unchecked** (so that we can review the summary data file that has been created.)

5   **Run the procedure**

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

Summary List Storage Information ——

Output Data File Name:	{NCSS Documents Folder}\PainDrugTimeSummary1.NCSS
Original Raw Data File:	{Example Data Folder}\Pain.NCSS
Data Variable(s):	(2) Cov, Pain
Group Variable(s):	(2) Drug, Time
Summary Statistic(s):	(3) Count, Mean, SD

(Summary List Report and Plots Follow)

The data summary list output data file is described in this report. Go to the data table to see the new summary data file that has been created.

The summary data values are stored in **PainDrugTimeSummary1.NCSS** as follows

Variable	Drug	Time	Count	Mean	SD
Cov	Kerlosin	0.5	7	179	45.4129203347828
Cov	Kerlosin	1	7	168.714285714286	32.3867991920283
Cov	Kerlosin	1.5	7	157.142857142857	45.1347189771118
Cov	Kerlosin	2	7	184.142857142857	27.8892367018089
Cov	Kerlosin	2.5	7	140.428571428571	21.4231739234032
Cov	Kerlosin	3	7	150.142857142857	34.1585937036278
Cov	Laposec	0.5	7	170.428571428571	44.2939316492946
Cov	Laposec	1	7	172.857142857143	14.7017977520729
Cov	Laposec	1.5	7	160.714285714286	40.2231868690779
Cov	Laposec	2	7	160.285714285714	33.3302379515173
Cov	Laposec	2.5	7	148.571428571429	33.0396587091793
Cov	Laposec	3	7	154	25.6450125105578
Cov	Placebo	0.5	7	162.142857142857	37.6538115442451
Cov	Placebo	1	7	144.142857142857	40.5767937431753
Cov	Placebo	1.5	7	156.285714285714	34.9319065502886
Cov	Placebo	2	7	150.142857142857	23.9612782869123
Cov	Placebo	2.5	7	183.142857142857	36.7216038657925
Cov	Placebo	3	7	156.857142857143	44.8940551796804
Pain	Kerlosin	0.5	7	76	5.25991127935317
Pain	Kerlosin	1	7	69.8571428571429	6.61887632529297
Pain	Kerlosin	1.5	7	65	4.65474668125631
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

Example 3c – Storing the Summary List in a New NCSS Data File (Column-by-Column)

To store the data summary list in a new data file column-by-column, check **Store the Summary List in a New NCSS Data File** and run the procedure again to get the results.

<u>Be careful because any unsaved data will be lost!</u>

6   Specify the data storage

- 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	
Store the Summary List in a New NCSS Data File	**Checked**
Output File Name...	**%mydocs_NCSS%\Data\ PainDrugTimeSummary2.NCSS**
Data Variable Storage	**Store as Columns**
Automatically Reopen the Current Dataset after .... the Save Operation Completes	**Unchecked** (so that we can review the summary data file that has been created.)

7   Run the procedure

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

Descriptive Statistics – Summary Lists

Summary List Storage Information ——————————————————————————————————————

Output Data File Name:	{NCSS Documents Folder}\PainDrugTimeSummary2.NCSS
Original Raw Data File:	{Example Data Folder}\Pain.NCSS
Data Variable(s):	(2) Cov, Pain
Group Variable(s):	(2) Drug, Time
Summary Statistic(s):	(3) Count, Mean, SD

(Summary List Report and Plots Follow)

The data summary list output data file is described in this report. Go to the data table to see the new summary data file that has been created.

The summary data values are stored in **PainDrugTimeSummary2.NCSS** as follows

Drug	Time	Cov_Count	Cov_Mean	Cov_SD	Pain_Count	…
Kerlosin	0.5	7	179	45.4129203347828	7	.
Kerlosin	1	7	168.714285714286	32.3867991920283	7	.
Kerlosin	1.5	7	157.142857142857	45.1347189771118	7	.
Kerlosin	2	7	184.142857142857	27.8892367018089	7	.
Kerlosin	2.5	7	140.428571428571	21.4231739234032	7	.
Kerlosin	3	7	150.142857142857	34.1585937036278	7	.
Laposec	0.5	7	170.428571428571	44.2939316492946	7	.
Laposec	1	7	172.857142857143	14.7017977520729	7	.
Laposec	1.5	7	160.714285714286	40.2231868690779	7	.
Laposec	2	7	160.285714285714	33.3302379515173	7	.
Laposec	2.5	7	148.571428571429	33.0396587091793	7	.
Laposec	3	7	154	25.6450125105578	7	.
Placebo	0.5	7	162.142857142857	37.6538115442451	7	.
Placebo	1	7	144.142857142857	40.5767937431753	7	.
Placebo	1.5	7	156.285714285714	34.9319065502886	7	.
Placebo	2	7	150.142857142857	23.9612782869123	7	.
Placebo	2.5	7	183.142857142857	36.7216038657925	7	.
Placebo	3	7	156.857142857143	44.8940551796804	7	.