Summary Statistics in SAS

Statistics 135

Autumn 2005

Summary Statistics in SAS

There are a number of approaches to calculating summary statistics in **SAS**. The most common three are

- `PROC MEANS`

  Provides data summarization tools to compute descriptive statistics for variables across all observations and within groups of observations.

- `PROC UNIVARIATE`

  Calculates many of the statistics that `PROC MEANS` plus some standard univariate graphical summaries, comparison of data to fixed distributions, and parameter estimation

- `PROC TABULATE`

  Displays descriptive statistics in tabular format, using some or all of the variables in a data set. You can create a variety of tables ranging from simple to highly customized.

`PROC TABULATE` computes many of the same statistics that are computed by other descriptive statistical procedures such as `PROC MEANS`, `PROC FREQ`, and `PROC REPORT`.

Example: Roofing Shingle Sales

Data on sales last year in 49 sales districts were collected for a maker of asphalt roofing shingles.

- Sales in 1000s of squares (sales)
- Promotional expenditures in 1000s of $ (promotion)
- Number of active accounts (accounts)
- Number of competing brands (brands)
- District potential (potential)

PROC MEANS

- Calculates descriptive statistics based on moments
- Estimates quantiles, which includes the median
- Calculates confidence limits for the mean
- Identifies extreme values
- Performs a t test.

```
PROC MEANS <option(s)> <statistic-keyword(s)>;
  BY <DESCENDING> variable-1 <...
     <DESCENDING> variable-n><NOTSORTED>;
  CLASS variable(s) </ option(s)>;
  FREQ variable;
  ID variable(s);
  OUTPUT <OUT=SAS-data-set>
         <output-statistic-specification(s)>
         <id-group-specification(s)>
         <maximum-id-specification(s)>
         <minimum-id-specification(s)>
         </ option(s)> ;
  TYPES request(s);
  VAR variable(s) < / WEIGHT=weight-variable>;
  WAYS list;
  WEIGHT variable;
```

There are a wide range of statistics calculated in this PROC. These include

- Descriptive statistics:

  `N, NMISS, MEAN, STDDEV|STD, VAR, MIN, MAX, RANGE, CV, SKEWNESS|SKEW, KURTOSIS|KURT, STDERR, CSS, SUM, SUMWGT, USS, CLM` (2-sided CI of μ), `LCLM, UCLM` (1-sided CI of μ)

  The default statistics are `N, MEAN, STD, MIN, MAX`

- Quantile statistics:

  `MEDIAN|P50, Q3|P75, P1, P90, P5, P95, P10, P99, Q1|P25, QRANGE`

- Hypothesis testing

  `PROBT, T`

There any many options available in this PROC. The most useful are

- DATA = SAS-data-set: Sets the data set for the PROC.
- ALPHA = α (default = 0.05): This sets confidence level to be $1 - \alpha$ for the confidence procedures.
- FW = field-width: Specifies the field width to display statistics in displayed output. Has no effect on values saved in an output data set.
- PRINT|NOPRINT (default = PRINT): Specifies whether output is to be printed.

```
PROC MEANS DATA = shingles;
  TITLE 'PROC MEANS Output of Roofing Shingle Sales';
  TITLE2 'Default Output';
  VAR sales promotion accounts brands potential;
```

PROC MEANS Output of Roofing Shingle Sales 2
Default Output 19:43 Sunday, November 27, 2005

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
sales	49	178.6183673	79.7929447	30.9000000	339.4000000
promotion	49	5.4938776	1.5544839	2.5000000	9.0000000
accounts	49	52.6938776	14.1276975	24.0000000	83.0000000
brands	49	8.9387755	2.3220695	4.0000000	14.0000000
potential	49	10.0000000	4.7609523	3.0000000	20.0000000

```
PROC MEANS DATA = shingles
  MEAN STD MIN Q1 MEDIAN Q3 MAX CLM PROBT T /* statistics */
  ALPHA = 0.01  FW = 8;   /* options */
  TITLE 'PROC MEANS Output of Roofing Shingle Sales';
  TITLE2 'Statistics Selected';
  VAR sales promotion accounts brands potential;
```

```
PROC MEANS Output of Roofing Shingle Sales                              3
Statistics Selected                      19:43 Sunday, November 27, 2005

The MEANS Procedure
```

Variable	Mean	Std Dev	Minimum	Lower Quartile	Median	Upper Quartile
sales	178.6	79.7929	30.9000	116.7	168.0	236.5
promotion	5.4939	1.5545	2.5000	4.5000	5.5000	6.5000
accounts	52.6939	14.1277	24.0000	44.0000	52.0000	62.0000
brands	8.9388	2.3221	4.0000	8.0000	9.0000	11.0000
potential	10.0000	4.7610	3.0000	6.0000	9.0000	13.0000

Variable	Maximum	Lower 99% CL for Mean	Upper 99% CL for Mean	Pr > |t|	t Value
sales	339.4	148.0	209.2	<.0001	15.67
promotion	9.0000	4.8982	6.0895	<.0001	24.74
accounts	83.0000	47.2805	58.1072	<.0001	26.11
brands	14.0000	8.0490	9.8285	<.0001	26.95
potential	20.0000	8.1757	11.8243	<.0001	14.70

PROC UNIVARIATE

- descriptive statistics based on moments (including skewness and kurtosis), quantiles or percentiles (such as the median), frequency tables, and extreme values

- histograms and comparative histograms. Optionally, these can be fitted with probability density curves for various distributions and with kernel density estimates.

- quantile-quantile plots (Q-Q plots) and probability plots. These plots facilitate the comparison of a data distribution with various theoretical distributions.

- goodness-of-fit tests for a variety of distributions including the normal

- the ability to inset summary statistics on plots produced on a graphics device

- the ability to analyze data sets with a frequency variable

- the ability to create output data sets containing summary statistics, histogram intervals, and parameters of fitted curves

```
PROC UNIVARIATE < options > ;
  BY variables ;
  CLASS variable-1 <(v-options)> < variable-2 <(v-options)> >
       < / KEYLEVEL= value1 | ( value1 value2 ) >;
  FREQ variable ;
  HISTOGRAM < variables > < / options > ;
  ID variables ;
  INSET keyword-list < / options > ;
  OUTPUT < OUT=SAS-data-set > < keyword1=names...keywordk=names >
         < percentile-options >;
  PROBPLOT < variables > < / options > ;
  QQPLOT < variables > < / options > ;
  VAR variables ;
  WEIGHT variable ;
```

This PROC generates a very large amount of output by default, and other options will increase it. Some useful ones are

- ALPHA = α (default $= 0.05$): This sets default confidence level to be $1 - \alpha$ for the confidence procedures. Can be overridden for specific intervals

- CIBASIC <(<TYPE = keyword> <ALPHA = α)>: Gives confidence intervals for μ, σ, and σ^2 assuming the data is normally distributed. TYPE specifies whether the interval is TWOSIDED (default), LOWER, or UPPER.

- CIPCTLDF <(<TYPE = keyword> <ALPHA = α)>
  CIQUANTDF <(<TYPE = keyword> <ALPHA = α)> :

  Calculates confidence intervals for quantiles by a distribution-free method based on ranks. TYPE takes the keywords LOWER, UPPER, SYMMETRIC (default), and ASYMMETRIC.

- `CIPCTLNORMAL <(<TYPE = keyword> <ALPHA =` α`)>`
  `CIQUANTNORMAL <(<TYPE = keyword> <ALPHA =` α`)>`:

  Calculates confidence intervals for quantiles assuming normally distributed data. The options are the same as those for `CIBASIC`.

- `MU0 =` μ_0: Sets the null hypothesis for the location parameter for tests of location. If you specify one value, it is used for all variables. If you specify more than one, you must specify the variables with a `VAR` statement. The default value is 0.

- `NEXTROBS =` n: Specifies the number of extreme observations (n smallest and n largest) to be displayed for each variable.

- `NORMAL`: Generates 4 tests of normality - Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling, and Cramer-von Mises. I suspect, but can't confirm that the Kolmogorov-Smirnov test is actually the Lilliefors test as you don't want to specify a mean and variance of the normal for the test, which would be required for the strict use of the Kolmogorov-Smirnov test.

- `PLOT`: Produces stem-and-leaf, box plot, and normal probability plot in line-printer output. If a `BY` statement is used, side-by-side box plots are generated.

- `ROBUSTSCALE`: Generates a table of robust estimates of scale. These include the interquartile range, Gini's mean difference, median absolute deviation around the median (MAD), plus a couple more due to Rousseeuw and Croux (1993).

- `TRIMMED=values <(<TYPE = keyword> <ALPHA =` α`)>`
  `TRIM=values <(<TYPE = keyword> <ALPHA =` α`)>`: Generates a table of trimmed means where `value` specifies the number or proportion of observations trimmed.

- `WINSORIZED=values <(<TYPE = keyword> <ALPHA =` α`)>`
  `WINSOR=values <(<TYPE = keyword> <ALPHA =` α`)>`: Generates a table of Winsorized means, a robust measure of location. The options work the same as for `TRIMMED`.

- `VARDEF=divisor`: Specifies the divisor to use in calculating variances. There are 4 choices

Value	Divisor	Formula for Divisor
`DF`	Degrees of freedom	$n-1$
`N`	Number of observations	n
`WDF`	Sum of Weights minus one	$(\sum_i w_i) - 1$
`WEIGHT|WGT`	Sum of Weights	$\sum_i w_i$

Lets now look at the various statements that can be included in a `PROC UNIVARIATE` block

- `VAR`: Specifies the analysis variables and there order in the results. If omitted, all variables will be analyzed. If you are going to store results from the analysis, this is required.

- `BY`: Generates separate analyses for each combination of the variables given. The default is to expect the data set to be sorted by the `BY` variables. This can be overridden by the `NOTSORTED` option.

- `CLASS`: Specifies one or two variables that the procedure uses to group the data into classification levels. An option to `BY` that doesn't require sorting your data. However it is restricted to at most 2 variables where `BY` can have more.

- `FREQ`: Allows specification of a numeric variable whose value represents the frequency of the observation.

- `WEIGHT`: Specifies numerical weights for analysis variables in the calculations. This is similar to `FREQ`, but allows for non-integer weights. The main use of this is to assume that the variance of observation i satisfies

$$\mathrm{Var}(X_i) = \frac{\sigma^2}{w_i}$$

  When calculating summary moments, the weighted versions look like

$$\bar{x}_w = \frac{\sum_i w_i x_i}{\sum_i w_i} \qquad s_w^2 = \frac{1}{d}\sum_i w_i (x_i - \bar{x}_w)^2$$

where d is taken from the VARDEF option.

- ID: Specifies one or more variables to include in the table of extreme observations.

- HISTOGRAM: Creates histograms and optionally superimposes estimated parametric and non-parametric density curves. The parametric distributions that can be fit are Beta, Exponential, Gamma, Lognormal, Normal, and Weibull. (Will discuss more later when discussing graphics).

- PROBPLOT: Creates a probability plot, which compares the ordered variable values with the percentiles of a specified theoretical distribution (default = NORMAL). The distributions available are the beta, exponential, gamma, lognormal, normal, two-parameter Weibull, and three-parameter Weibull.

- QQPLOT: Creates quantile-quantile plots (Q-Q plots) using high-resolution graphics and compares ordered variable values with quantiles of a specified theoretical distribution.

Q-Q plots are preferable for graphical estimation of distribution parameters, whereas probability plots are preferable for graphical estimation of percentiles. (Will look at the differences later between the two.)

- `INSET`: Places a box or table of summary statistics in a high-resolution `HISTOGRAM`, `PROBPLOT`, or `QQPLOT`.

- `OUTPUT < OUT=SAS-data-set >`
  `< keyword1=names...keywordk=names >`
  `< percentile-options >`:

  Allows for summary statistics to be stored in a **SAS** dataset.

```
PROC UNIVARIATE DATA = shingles
  NORMAL  CIBASIC  PLOTS  ALPHA = 0.01;
  VAR sales;
  TITLE 'Roofing Shingle Sales';
```


The UNIVARIATE Procedure
Variable: sales

Moments

N	49	Sum Weights	49
Mean	178.618367	Sum Observations	8752.3
Std Deviation	79.7929447	Variance	6366.91403
Skewness	0.15086445	Kurtosis	-0.8142449
Uncorrected SS	1868933.41	Corrected SS	305611.873
Coeff Variation	44.6723066	Std Error Mean	11.3989921

Basic Statistical Measures

Location		Variability	
Mean	178.6184	Std Deviation	79.79294
Median	168.0000	Variance	6367
Mode	200.1000	Range	308.50000
		Interquartile Range	119.80000

Basic Confidence Limits Assuming Normality

Parameter	Estimate	99% Confidence Limits	
Mean	178.61837	148.04394	209.19279
Std Deviation	79.79294	63.01266	107.36813
Variance	6367	3971	11528

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	15.66966	Pr > |t|	<.0001
Sign	M	24.5	Pr >= |M|	<.0001
Signed Rank	S	612.5	Pr >= |S|	<.0001

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.975674	Pr < W	0.4002
Kolmogorov-Smirnov	D	0.075675	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.040111	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.307989	Pr > A-Sq	>0.2500

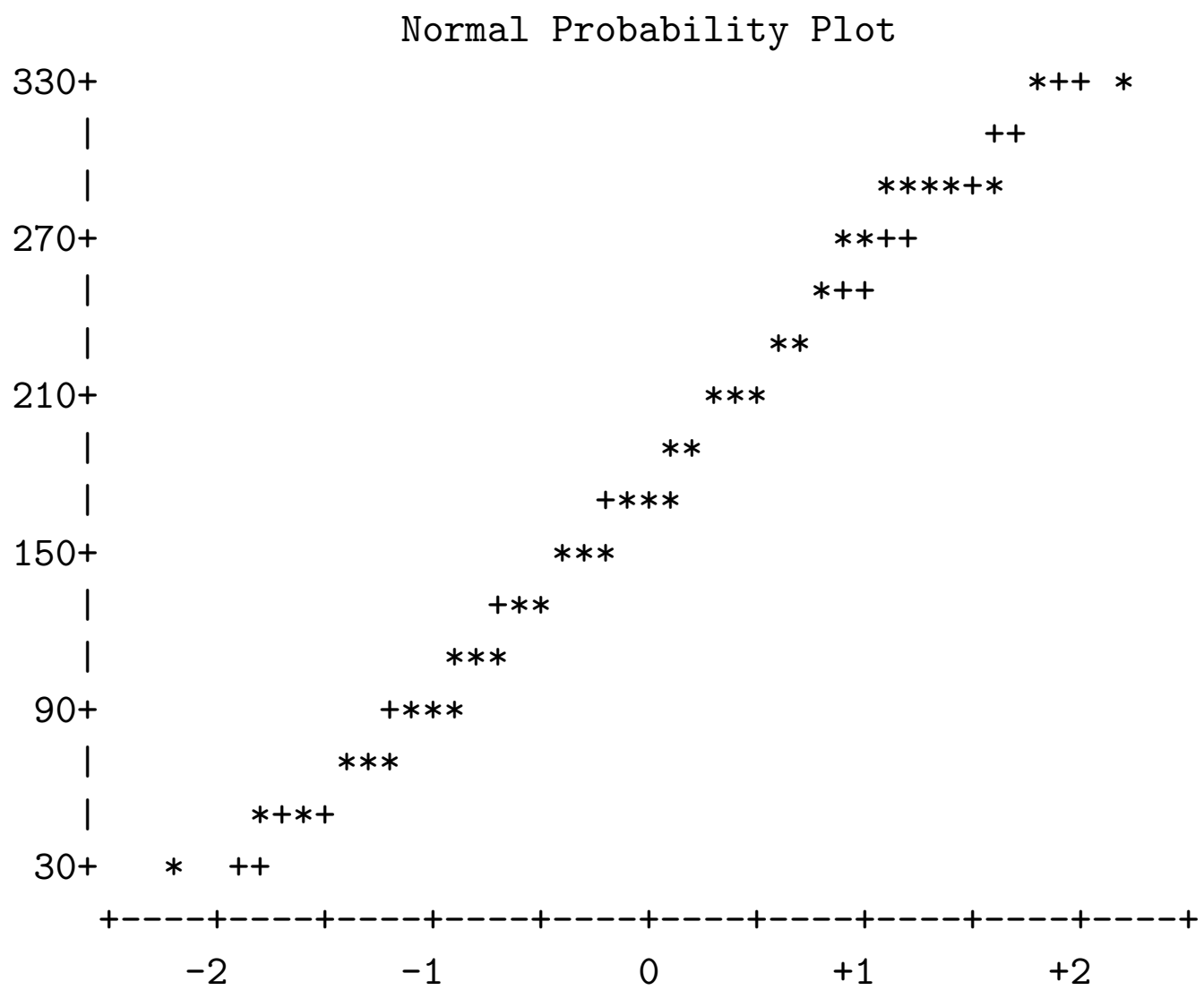
Quantiles (Definition 5)

Quantile	Estimate
100% Max	339.4
99%	339.4
95%	295.8
90%	291.5
75% Q3	236.5
50% Median	168.0
25% Q1	116.7
10%	73.4
5%	48.0
1%	30.9
0% Min	30.9

Extreme Observations

----Lowest----		----Highest----	
Value	Obs	Value	Obs
30.9	7	291.5	27
47.7	22	291.9	8
48.0	29	295.8	34
64.7	42	331.2	26
73.4	21	339.4	10

```
Stem Leaf                     #  Boxplot
  32 19                       2     |
  30                          0     |
  28 71226                    5     |
  26 938                      3     |
  24 9                        1     |
  22 0368                     4  +-----+
  20 00238                    5  |     |
  18 005                      3  |     |
  16 00388                    5  *--+--*
  14 16055                    5  |     |
  12 856                      3  |     |
  10 0767                     4  +-----+
   8 614                      3     |
   6 539                      3     |
   4 88                       2     |
   2 1                        1     |
     ----+----+----+----+
 Multiply Stem.Leaf by 10**+1
```


Normal Probability Plot
330+
270+
210+
150+
90+
30+
-2
-1
0
+1
+2

Now lets look at what happens with BY and CLASS statements

```
PROC SORT DATA = shingles2;
  BY potentcat;

PROC UNIVARIATE DATA = shingles2;
  VAR promotion
  BY potentcat;           /* sorted data */
```

potentcat=High

The UNIVARIATE Procedure
Variable: promotion

Moments

N	9	Sum Weights	9
Mean	5.01111111	Sum Observations	45.1
Std Deviation	1.55920208	Variance	2.43111111
Skewness	-0.2993867	Kurtosis	-1.7660273
Uncorrected SS	245.45	Corrected SS	19.4488889

Coeff Variation	31.1148973	Std Error Mean	0.51973403

skip a bunch of output

potentcat=Low

The UNIVARIATE Procedure
Variable: promotion

Moments

N	9	Sum Weights	9
Mean	5.03333333	Sum Observations	45.3
Std Deviation	1.39731886	Variance	1.9525
Skewness	0.69844229	Kurtosis	-0.6049314
Uncorrected SS	243.63	Corrected SS	15.62
Coeff Variation	27.7613019	Std Error Mean	0.46577295

```
PROC UNIVARIATE DATA = shingles;
  VAR accounts;
  CLASS potentcat;    /* unsorted data */
```

```
The UNIVARIATE Procedure
Variable:  promotion
potentcat = High
```

Moments

N	9	Sum Weights	9
Mean	5.01111111	Sum Observations	45.1
Std Deviation	1.55920208	Variance	2.43111111
Skewness	-0.2993867	Kurtosis	-1.7660273
Uncorrected SS	245.45	Corrected SS	19.4488889
Coeff Variation	31.1148973	Std Error Mean	0.51973403

```
skip a whole bunch
```

Robust Measures

As noted earlier, **SAS** will generate robust measures of location and scale that will often work better in the presence of outliers.

Measures of locations include the median, the trimmed mean, and the Winsorized mean.

Measures of scale include the interquartile range, Gini's mean difference, median absolute deviation from the median (MAD), Q_n, and S_n. The last two measures were developed by Rousseeuw and Croux.

The trimmed and Winsorized means are a modification of the sample mean by dealing with the k smallest and k largest observations in a different way. Assume that the ordered observations are

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

Then these estimates of location are

- k-times trimmed mean $\bar{x}_{tk}$

$$\bar{x}_{tk} = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)}$$

i.e. the average of the middle $n-2k$ observations

If the distribution the observations are sampled from is symmetric, $\bar{x}_{tk}$ is an unbiased estimate of μ.

In this situation, inference can be performed on μ. This is based on

$$t_{tk} = \frac{\bar{x}_{tk} - \mu}{SE(\bar{x}_{tk})}$$

having an approximate t_{n-2k-1} distribution. The standard error satisfies

$$SE(\bar{x}_{tk}) = \frac{S_{wk}}{\sqrt{(n-2k)(n-2k-1)}}$$

where S^2_{wk} is the Winsorized sum of squared deviations (coming in a minute). This can be used to calculate confidence intervals

$$\bar{x}_{tk} \pm t_{1-\frac{\alpha}{2},n-2k-1} SE(\bar{x}_{tk})$$

and a test statistic

$$t_{tk} = \frac{\bar{x}_{tk} - \mu_0}{SE(\bar{x}_{tk})}$$

where μ_0 is the null hypothesis mean value.

- k-times Winsorized mean $\bar{x}_{wk}$

$$\bar{x}_{tk} = \frac{1}{n}\left((k+1)x_{(k+1)} + \sum_{i=k+2}^{n-k-1} x_{(i)} + (k+1)x_{(k-n)}\right)$$

With this estimate, the k smallest observations are replaced by $x_{(k+1)}$ and the k largest observations are replaced by $x_{(n-k)}$.

Like the trimmed mean, if the distribution the observations are sampled from is symmetric, $\bar{x}_{wk}$ is an unbiased estimate of μ.

Similarly, inference can be performed based on $\bar{x}_{wk}$. This is based on

$$t_{wk} = \frac{\bar{x}_{wk} - \mu}{SE(\bar{x}_{wk})}$$

having an approximate t_{n-2k-1} distribution. The standard error satisfies

$$SE(\bar{x}_{tk}) = \frac{n-1}{n-2k-1}\frac{S_{wk}}{\sqrt{n(n-1)}}$$

where S^2_{wk} is the Winsorized sum of squared deviations

$$S^2_{wk} = (k+1)(x_{(k+1)}-\bar{x}_{wk})^2+\sum_{i=k+2}^{n-k-1}(x_{(i)}-\bar{x}_{wk})^2+(k+1)(x_{(k-n)}-\bar{x}_{wk})^2$$

This can be used to calculate confidence intervals

$$\bar{x}_{wk} \pm t_{1-\frac{\alpha}{2},n-2k-1} SE(\bar{x}_{wk})$$

and a test statistic

$$t_{wk} = \frac{\bar{x}_{wk} - \mu_0}{SE(\bar{x}_{wk})}$$

where μ_0 is the null hypothesis mean value.

The measures of scale are

- Interquartile Range

$$IQR = Q3 - Q1$$

If the data is normally distributed, σ can be estimated by

$$s_{IQR} = \frac{IQR}{1.34898} = \frac{IQR}{(\Phi^{-1}(0.75) - \Phi^{-1}(0.25))}$$

- Gini's mean difference

$$G = \frac{1}{\binom{n}{2}} \sum_{i<j} |x_i - x_j|$$

If the data is normally distributed,

$$E[G] = \sigma \frac{2}{\sqrt{\pi}}$$

thus σ can be unbiasedly estimated by

$$s_G = G \frac{\sqrt{\pi}}{2}$$

In addition, for normally distributed data, s_G has a high efficiency relative to s and is less sensitive to the presence of outliers.

- MAD

$$MAD = \text{med}_i(|x_i - \text{Med}|)$$

where Med is the median of the data. An estimate of σ for normally distributed data is

$$s_{MAD} = 1.4826 MAD$$

For normally distributed data this has a low efficiency and may not always be appropriate for symmetric distributions (not sure why). To deal with these problems the following two statistics have been proposed to the MAD

- S_n:

$$S_n = 1.1926 \text{med}_i(\text{med}_j|x_i - x_j|)$$

where the outer median (over i) is the median of n medians of $|x_i - x_j|$.

- Q_n:

$$Q_n = 2.219 \left\{|x_i - x_j|; i < j\right\}_{(k)}$$

where

$$k = \binom{\lfloor \frac{n}{2} \rfloor + 1}{2}$$

```
PROC UNIVARIATE DATA = shingles
  TRIM = 5 WINSOR = 5 ROBUSTSCALE;
  VAR sales;
```

Trimmed Means

Percent Trimmed in Tail	Number Trimmed in Tail	Trimmed Mean	Std Error Trimmed Mean	95% Confidence Limits		DF
10.20	5	177.8923	12.98790	151.5997	204.1849	38

Trimmed Means

Percent Trimmed in Tail	t for H0: Mu0=0.00	Pr > |t|
10.20	13.69678	<.0001

Winsorized Means

Percent Winsorized in Tail	Number Winsorized in Tail	Winsorized Mean	Std Error Winsorized Mean	95% Confidence Limits		DF
10.20	5	179.4143	13.02272	153.0512	205.7774	38

Winsorized Means

Percent Winsorized in Tail	t for H0: Mu0=0.00	Pr > |t|
10.20	13.77702	<.0001

Robust Measures of Scale

Measure	Value	Estimate of Sigma
Interquartile Range	119.8000	88.80784
Gini's Mean Difference	92.3745	81.86476
MAD	55.3000	81.98778
Sn	83.0050	84.55807
Qn	89.7648	87.27129

By changing k, we get

```
PROC UNIVARIATE DATA = shingles
  TRIM = 10 WINSOR = 10;
  VAR sales;
```

Trimmed Means

Percent Trimmed in Tail	Number Trimmed in Tail	Trimmed Mean	Std Error Trimmed Mean	95% Confidence Limits		DF
20.41	10	175.4724	14.11918	146.5506	204.3942	28

Trimmed Means

Percent Trimmed in Tail	t for H0: Mu0=0.00	Pr > |t|
20.41	12.42795	<.0001

Winsorized Means

Percent Winsorized in Tail	Number Winsorized in Tail	Winsorized Mean	Std Error Winsorized Mean	95% Confidence Limits		DF
20.41	10	178.5857	14.22171	149.4539	207.7176	28

Winsorized Means

Percent Winsorized in Tail	t for H0: Mu0=0.00	Pr > |t|
20.41	12.55726	<.0001

Summary of Trimmed and Winsorized Means

k	$\bar{x}_{tk}$	$\mathrm{SE}(\bar{x}_{tk})$	$\bar{x}_{wk}$	$\mathrm{SE}(\bar{x}_{wk})$
5	177.8923	12.98790	179.4143	13.02272
10	175.4725	14.11918	178.5857	14.22171