


[Chapter Contents](#)
[Previous](#)
[Next](#)

PROC CAPABILITY and General Statements

Robust Estimators

The CAPABILITY procedure provides several methods for computing robust estimates of location and scale, which are insensitive to outliers in the data.

Winsorized Means

The k -times Winsorized mean is a robust estimator of location which is computed as

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

where n is the number of observations, and $x_{(i)}$ is the i th order statistic when the observations are arranged in increasing order:

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

The Winsorized mean is the mean computed after replacing the k smallest observations with the $(k+1)$ st smallest observation, and the k largest observations with the $(k+1)$ st largest observation.

For data from a symmetric distribution, the Winsorized mean is an unbiased estimate of the population mean. However, the Winsorized mean does not have a normal distribution even if the data are normally distributed.

The Winsorized sum of squared deviations is defined as

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

A Winsorized t test is given by

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

where the standard error of the Winsorized mean is

$$\text{STDERR}(\bar{x}_{wk}) = \frac{n-1}{n-2k-1} \frac{s_{wk}}{\sqrt{n(n-1)}}$$

When the data are from a symmetric distribution, the distribution of t_{wk} is approximated by a Student's t distribution with $n-2k-1$ degrees of freedom. Refer to Tukey and McLaughlin (1963) and Dixon and Tukey (1968).

A $100(1 - \alpha)\%$ Winsorized confidence interval for the mean has upper and lower limits

$$\bar{x}_{wk} \pm t_{1-\alpha/2} \text{STDERR}(\bar{x}_{wk})$$

where $t_{1-\alpha/2}$ is the $(1 - \alpha)/2$ 100 th percentile of the Student's t distribution with $n-2k-1$ degrees of freedom.

Trimmed Means

The k -times trimmed mean is a robust estimator of location which is computed as

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

where n is the number of observations, and $x_{(i)}$ is the i th order statistic when the observations are arranged in increasing order:

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

The trimmed mean is the mean computed after the k smallest observations and the k largest observations in the sample are deleted.

For data from a symmetric distribution, the trimmed mean is an unbiased estimate of the population mean. However, the trimmed mean does not have a normal distribution even if the data are normally distributed.

A robust estimate of the variance of the trimmed mean t_{tk} can be obtained from the Winsorized sum of squared deviations; refer to Tukey and McLaughlin (1963). the corresponding trimmed t test is given by

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

where the standard error of the trimmed mean is

$$\text{STDERR}(\bar{x}_{tk}) = \frac{s_{tk}}{\sqrt{(n-2k)(n-2k-1)}}$$

and s_{tk} is the square root of the Winsorized sum of squared deviations.

When the data are from a symmetric distribution, the distribution of t_{tk} is approximated by a Student's t distribution with $n-2k-1$ degrees of freedom. Refer to Tukey and McLaughlin (1963) and Dixon and Tukey (1968).

A $100(1 - \alpha)\%$ trimmed confidence interval for the mean has upper and lower limits

$$\bar{x}_{tk} \pm t_{1-\alpha/2} \text{STDERR}(\bar{x}_{tk})$$

where $t_{1-\alpha/2}$ is the $(1 - \alpha)/2$ 100 th percentile of the Student's t distribution with $n-2k-1$ degrees of freedom.

Robust Estimates of Scale

The sample standard deviation, which is the most commonly used estimator of scale, is sensitive to outliers. Robust scale estimators, on the other hand, remain bounded when a single data value is replaced by an arbitrarily large or small value. The CAPABILITY procedure computes several robust measures of scale, including the interquartile range Gini's mean difference G , the median absolute deviation about the median (MAD), Q_n , and S_n . In addition, the procedure computes estimates of the normal standard deviation σ derived from each of these measures.

The interquartile range (IQR) is simply the difference between the upper and lower quartiles. For a normal population, σ can be estimated as $\text{IQR}/1.34898$.

Gini's mean difference is computed as

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

For a normal population, the expected value of G is $2\sigma/\sqrt{\pi}$. Thus $G\sqrt{\pi}/2$ is a robust estimator of σ when the data are from a normal sample. For the normal distribution, this estimator has high efficiency relative to the usual sample standard deviation, and it is also less sensitive to the presence of outliers.

A very robust scale estimator is the MAD, the median absolute deviation from the median (Hampel, 1974), which is computed as

$$\mathbf{MAD} = \text{med}_i(|x_i - \text{med}_j(x_j)|)$$

where the inner median, $\text{med}_j(x_j)$, is the median of the n observations, and the outer median (taken over i) is the median of the n absolute values of the deviations about the inner median. For a normal population, 1.4826MAD is an estimator of σ .

The MAD has low efficiency for normal distributions, and it may not always be appropriate for symmetric distributions. Rousseeuw and Croux (1993) proposed two statistics as alternatives to the MAD. The first is

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

where the outer median (taken over i) is the median of the n medians of $|x_i - x_j|, j = 1, 2, \dots, n$. To reduce

small-sample bias, $c_{sn}S_n$ is used to estimate σ , where c_{sn} is a correction factor; refer to Croux and Rousseeuw (1992).

The second statistic is

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

where

$$k = \binom{h}{2}$$

and $h = [n/2] + 1$. In other words, Q_n is 2.219 times the k th order statistic of the $\binom{n}{2}$ distances between

the data points. The bias-corrected statistic $c_{qn}Q_n$ is used to estimate σ , where c_{qn} is a correction factor; refer to Croux and Rousseeuw (1992).



[Copyright © 1999 by SAS Institute Inc., Cary, NC, USA. All rights reserved.](#)