EXAMINE

Univariate Statistics

Notation

The following notation is used throughout this chapter unless otherwise noted:

Let $y_1 < \ldots y_m$ be m distinct ordered observations for the sample and $c_1, \ldots, c_m$ be the corresponding caseweights. Then

$$cc_i = \sum_{k=1}^{i} c_k = \text{cumulative frequency up to and including } y_i$$

and

$$W = cc_m = \sum_{k=1}^{m} c_k = \text{total sum of weights.}$$

Descriptive Statistics

Minimum and Maximum

$$\text{min} = y_1, \quad \text{max} = y_m$$

Range

$$\text{range} = y_m - y_1$$

Mean $(\bar{y})$

$$\bar{y} = \frac{\sum_{i=1}^{m} c_i y_i}{W}$$

Confidence Interval for the Mean

$$\text{lower bound} = \bar{y} - t_{\alpha/2, W-1}\text{SE}$$
$$\text{upper bound} = \bar{y} + t_{\alpha/2, W-1}\text{SE}$$

where SE is the standard error.

Median

The median is the 50th percentile, which is calculated by the method requested. The default method is HAVERAGE.

Interquartile Range (IQR)

IQR = 75th percentile – 25th percentile, where the 75th and 25th percentiles are calculated by the method requested for percentiles.

Variance (s^2)

$$s^2 = \frac{1}{W-1} \sum_{i=1}^{m} c_i \left(y_i - \bar{y}\right)^2$$

Standard Deviation

$$s = \sqrt{s^2}$$

Standard Error

$$SE = \frac{s}{\sqrt{W}}$$

Skewness $(\mathbf{g_1})$ and SE of Skewness

$$g_1 = \frac{WM_3}{(W-1)(W-2)s^3}$$

$$SE(g_1) = \sqrt{\frac{6W(W-1)}{(W-2)(W+1)(W+3)}}$$

$$M_3 = \sum_{i=1}^{m} c_i\left(y_i - \bar{y}\right)^3$$

Kurtosis $(\mathbf{g_2})$ and SE of Kurtosis

$$g_2 = \frac{W(W+1)M_4 - 3M_2^2(W-1)}{(W-1)(W-2)(W-3)s^4}$$

$$M_2 = \sum_{i=1}^{m} c_i\left(y_i - \bar{y}\right)^2$$

$$M_4 = \sum_{i=1}^{m} c_i\left(y_i - \bar{y}\right)^4$$

$$SE(g_2) = \sqrt{\frac{4\left(W^2-1\right)SE^2(g_1)}{(W-3)(W+5)}}$$

5% Trimmed Mean $_{0.9}$

$$T_{0.9} = \frac{1}{0.9W}\left\{\left(cc_{k_1+1} - tc\right)y_{k_1+1} + \left(W - cc_{k_2-1} - tc\right)y_{k_2} + \sum_{i=k_1+2}^{k_2-1} c_i y_i\right\}$$

where k_1 and k_2 satisfy the following conditions

$$cc_{k_1} < tc \le cc_{k_1+1}, \quad W - cc_{k_2} < tc \le W - cc_{k_2-1}$$

and

$$tc = 0.05W$$

Note: If $k_1 + 1 = k_2$, then $T_{0.9} = y_{k_2}$

Percentiles

There are five methods for computation of percentiles. Let

$$tc_1 = Wp, \quad tc_2 = (W+1)p$$

where p is the requested percentile divided by 100, and k_1 and k_2 satisfy

$$cc_{k_1} \le tc_1 < cc_{k_1+1}$$
$$cc_{k_2} \le tc_2 < cc_{k_2+1}$$

Then,

$$g_1 = \frac{\left(tc_1 - cc_{k_1}\right)}{c_{k_1+1}}, \quad g_1^* = tc_1 - cc_{k_1}$$

$$g_2 = \frac{\left(tc_2 - cc_{k_2}\right)}{c_{k_2+1}}, \quad g_2^* = tc_2 - cc_{k_2}$$

Let x be the pth percentile; the five definitions are as follows:

Waverage (Weighted Average at y_{tc_1})

$$x = \begin{cases} y_{k_1+1} & \text{if } g_1^* \geq 1 \\ \left(1-g_1^*\right)y_{k_1} + g_1^* y_{k_1+1} & \text{if } g_1^* < 1 \text{ and } c_{k_1+1} \geq 1 \\ \left(1-g_1\right)y_{k_1} + g_1 y_{k_1+1} & \text{if } g_1^* < 1 \text{ and } c_{k_1+1} < 1 \end{cases}$$

Round (Observation Closest to tc_1)

If $c_{k_1+1} \geq 1$, then

$$x = \begin{cases} y_{k_1} & \text{if } g_1^* < \frac{1}{2} \\ y_{k_1+1} & \text{if } g_1^* \geq \frac{1}{2} \end{cases}$$

If $c_{k_1+1} < 1$, then

$$x = \begin{cases} y_{k_1} & \text{if } g_1 < \frac{1}{2} \\ y_{k_1+1} & \text{if } g_1 \geq \frac{1}{2} \end{cases}$$

Empirical (Empirical Distribution Function)

$$x = \begin{cases} y_{k_1} & \text{if } g_1^* = 0 \\ y_{k_1+1} & \text{if } g_1^* > 0 \end{cases}$$

Haverage (Weighted Average at y_{tc_2})

$$x = \begin{cases} y_{k_2+1} & \text{if } g_2^* \ge 1 \\ \left(1-g_2^*\right)y_{k_2} + g_2^* y_{k_2+1} & \text{if } g_2^* < 1 \text{ and } c_{k_2+1} \ge 1 \\ \left(1-g_2\right)y_{k_2} + g_2 y_{k_2+1} & \text{if } g_2^* < 1 \text{ and } c_{k_2+1} < 1 \end{cases}$$

Aempirical (Empirical Distribution Function with Averaging)

$$x = \begin{cases} \left(y_{k_1} + y_{k_1+1}\right)/2 & \text{if } g_1^* = 0 \\ y_{k_1+1} & \text{if } g_1^* > 0 \end{cases}$$

Note: If either the 25th, 50th, or 75th percentiles is request, Tukey Hinges will also be printed.

Tukey Hinges

Let Q_1, Q_2, and Q_3 be the 25th, 50th, and 75th percentiles. If $c^* \ge 1$, where $c^* = \min(c_1, \ldots, c_m)$, define

$$d = \frac{\text{greatest integer } \le \left(\left(W+3\right)/2\right)}{2}$$

$$L_1 = d$$

$$L_2 = W/2 + 1/2$$

$$L_3 = W + 1 - d$$

Otherwise

$$d = \frac{\text{greatest integer } \le \left(W/c^* + 3\right)/2}{2}$$

and

$$
\begin{aligned}
L_1 &= dc^* \\
L_2 &= W/2 + c^*/2 \\
L_3 &= W + c^* - dc^*
\end{aligned}
$$

Then for every i, $i = 1, 2, 3$, find h_i such that

$$cc_{h_i} \le L_i < cc_{h_i+1}$$

and

$$
Q_i = \begin{cases}
\left(1 - a_i^*\right) y_{h_i} + a_i^* y_{h_i+1} & \text{if } a_i^* < 1 \text{ and } c_{h_i+1} \ge 1 \\
\left(1 - a_i\right) y_{h_i} + a_i y_{h_i+1} & \text{if } a_i^* < 1 \text{ and } c_{h_i+1} < 1 \\
y_{h_i+1} & \text{if } a_i^* \ge 1
\end{cases}
$$

where

$$
\begin{aligned}
a_i^* &= L_i - cc_{h_i} \\
a_i &= \frac{a_i^*}{c_{h_i+1}}
\end{aligned}
$$

M-Estimation (Robust Location Estimation)

The M-estimator T of location is the solution of

$$\sum_{i=1}^{m} c_i \boldsymbol{\Psi}\left(\frac{y_i - T}{s}\right) = 0$$

where $\boldsymbol{\Psi}$ is an odd function and s is a measure of the spread.

An alternative form of M-estimation is

$$\sum_{i=1}^{m} c_i \left(\frac{y_i - T}{s}\right) \omega\left(\frac{y_i - T}{s}\right) = 0$$

where

$$\omega(u) = \frac{\boldsymbol{\Psi}(u)}{u}$$

After rearranging the above equation, we get

$$T = \frac{\sum_{i=1}^{m} c_i y_i \omega\left(\frac{y_i - T}{s}\right)}{\sum_{i=1}^{m} c_i \omega\left(\frac{y_i - T}{s}\right)}$$

Therefore, the algorithm to find M-estimators is defined iteratively by

$$T_{k+1} = \frac{\sum_{i=1}^{m} c_i y_i \omega\left(\frac{y_i - T_k}{s}\right)}{\sum_{i=1}^{m} c_i \omega\left(\frac{y_i - T_k}{s}\right)}$$

The algorithm stops when either

$|T_{k+1} - T_k| \le \varepsilon\left[(T_{k+1} + T_k)/2\right]$, where $\varepsilon = 0.005$

or the number of iterations exceeds 30.

M-Estimators

Four M-estimators (Huber, Hampel, Andrew, and Tukey) are available. Let

$$u_i = \frac{y_i - T}{s}$$

where

$s =$ median of $\tilde{y}_1, \ldots, \tilde{y}_m$ with caseweights $c_1, \ldots, c_m$

and

$\tilde{y}_i = \left| y_i - \tilde{y} \right|$, where $\tilde{y}$ is the median.

Huber (k), $k > 0$

$$\omega(u_i) = \begin{cases} 1 & \text{if } \left|u_i\right| \le k \\ \dfrac{k}{u_i}\operatorname{sgn}(u_i) & \text{if } \left|u_i\right| > k \end{cases}$$

The default value of $k = 1.339$

Hampel (a, b, c), $0 < a \le b \le c$

$$\omega(u_i) = \begin{cases} 1 & \text{if } \left|u_i\right| \le a \\ \dfrac{a}{u_i}\operatorname{sgn}(u_i) & \text{if } a < \left|u_i\right| \le b \\ \dfrac{a}{u_i}\dfrac{c - \left|u_i\right|}{c - b}\operatorname{sgn}(u_i) & \text{if } b < \left|u_i\right| \le c \\ 0 & \text{if } \left|u_i\right| > c \end{cases}$$

By default, $a = 1.7$, $b = 3.4$ and $c = 8.5$.

Andrew's Wave (*c*), *c* > 0

$$\omega(u_i) = \begin{cases} \dfrac{c}{\pi u_i}\sin\left(\dfrac{\pi u_i}{c}\right) & \text{if } |u_i| \le c \\ 0 & \text{if } |u_i| > c \end{cases}$$

By default, $c = 1.34\pi$

Tukey's Biweight (*c*)

$$\omega(u_i) = \begin{cases} \left(1 - \dfrac{u_i^2}{c^2}\right)^2 & \text{if } |u_i| \le c \\ 0 & \text{if } |u_i| > c \end{cases}$$

By default, c = 4.685.

Tests of Normality

Shapiro-Wilk Statistic (*W*)

Since the W statistic is based on the order statistics of the sample, the caseweights have to be restricted to integers. Hence, before W is calculated, all the caseweights are rounded to the closest integer and the series is expanded. Let c_i^* be the closest integer to c_i; then

$$cc_i^* = \sum_{k=1}^{i} c_k^*, \quad W_s = cc_m^* = \sum_{k=1}^{m} c_k^*$$

The original series $y = \{y_1, \ldots, y_m\}$ is expanded to

$$x = \{x_1, \ldots, x_{W_s}\}$$

where

$$x_{cc_{i-1}^*+1} = \ldots = x_{cc_i^*} = y_i, \quad i = 1, \ldots, m$$

Then the W statistic is defined as

$$W = \frac{\left(\sum_{i=1}^{W_s} a_i x_i\right)^2}{\sum_{i=1}^{W_s} \left(x_i - \bar{x}\right)^2}$$

where

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

$$a_1^2 = a_{W_s}^2 = \begin{cases} \dfrac{\Gamma\left(W_s/2\right)}{\sqrt{2}\Gamma\left(\left(W_s+1\right)/2\right)} & \text{if } 5 \le W_s \le 20 \\[2ex] \dfrac{\Gamma\left(\left(W_s+1\right)/2\right)}{\sqrt{2}\Gamma\left(W_s/2+1\right)} & \text{if } W_s > 20 \end{cases}$$

$$a_1 = -\sqrt{a_1^2}\,, \quad a_{W_s} = \sqrt{a_{W_s}^2}$$

$$a_i = \left(2/c\right)m_i, \quad i = 2,\ldots,W_s - 1$$

$$m_i = \Psi^{-1}\left(\frac{i-\alpha}{W_s - 2\alpha + 1}\right), \text{where } \Psi \text{ is the c.d.f. of a standard normal distribution}$$

$$\alpha = 0.314195 + 0.063336\beta - 0.010895\beta^2$$

$$\beta = \log_{10} W_s$$

$$c^2 = 4\sum_{i=1}^{W_s-1} \frac{m_i^2}{\left(1-2a_i^2\right)}$$

Based on the computed W statistic, the significance is calculated by linearly interpolating within the range of simulated critical values given in Shapiro and Wilk (1965).

If non-integer weights are specified, the Shapiro-Wilk's statistic is calculated when the weighted sample size lies between 3 and 50. For no weights or integer weights, the statistic is calculated when the weighted sample size lies between 3 and 5000.

If $W > w_{0.99}$, the critical value of 99th percentile, the significance is reported as >0.99. Similarly, if $W < w_{0.01}$, the critical value of first percentile, the significance is reported as <0.01.

Kolmogorov-Smirnov Statistic with Lilliefors' Significance

Lilliefors (1967) presented a table for testing normality using the Kolmogorov-Smirnov statistic when the mean and variance of the population are unknown. This statistic is[1]

$$D_a = \max\{D_+, D_-\}$$

where

$$D_+ = \max_i \left\{\hat{F}(y_i) - F(y_i)\right\}$$
$$D_- = \max_i \left\{F(y_i) - \hat{F}(y_{i-1})\right\}$$

where $\hat{F}(x)$ is the sample cumulative distribution and $F(x)$ is the cumulative normal distribution whose mean and variance are estimated from the sample.

Dallal and Wilkinson (1986) corrected the critical values for testing normality reported by Lilliefors. With the corrected table they derived an analytic approximation to the upper tail probabilities of D_a for probabilities less than 0.1.

[1] This algorithm applies to SPSS 7.0 and later releases.

The following formula is used to estimate the critical value D_c for probability 0.1.

$$D_c = \frac{\left(-b - \sqrt{b^2 - 4ac}\right)}{2a}$$

where, if $W \le 100$,

$$a = -7.01256(W + 2.78019)$$

$$b = 2.99587\sqrt{W + 2.78019}$$

$$c = 2.1804661 + \frac{0.974598}{\sqrt{W}} + \frac{1.67997}{W}$$

If[2] $W > 100$

$$a = -7.90289126054 * W^{0.98}$$
$$b = 3.180370175721 * W^{0.49}$$
$$c = 2.2947256$$

The Lilliefors significance *p* is calculated as follows:

If $D_a = D_c$, $p = 0.1$.

If $D_a > D_c$, $p = \exp\left\{aD_a^2 + bD_a + c - 2.3025851\right\}$.

If $D_{0.2} \le D_a < D_c$, linear interpolation between $D_{0.2}$ and D_c where $D_{0.2}$ is the critical value for probability 0.2 is done.

If $D_a > D_{0.2}$, p is reported as > 0.2 .

[2] This algorithm applies to SPSS 7.0 and later releases. To learn about algorithms for previous releases, call SPSS Technical Support.

Group Statistics

Assume that there are $k(k \geq 2)$ combinations of grouping factors. For every combination i, $i = 1, 2, \ldots, k$, let $\{y_{i1}, \ldots, y_{im_i}\}$ be the sample observations with the corresponding caseweights $\{c_{i1}, \ldots, c_{im_i}\}$.

Spread versus Level

If a transformation value, a, is given, the spread(s) and level(l) are defined based on the transformed data. Let x be the transformed value of y; for every $i = 1, \ldots, k$, $j = 1, \ldots, m_i$

$$x_{ij} = \begin{cases} \ln y_{ij} & \text{if } a = 0 \\ y_{ij}^{a} & \text{otherwise} \end{cases}$$

Then the spread (s_i) and the level (l_i) are respectively defined as the Interquartile Range and the median of $\{x_{i1}, \ldots, x_{im_i}\}$ with corresponding caseweights $\{c_{i1}, \ldots, c_{im_i}\}$. However, if a is not specified, the spread and the level are natural logarithms of the Interquartile Range and of the median of the original data.

Finally, the slope is the regression coefficient of s on l, which is defined as

$$\frac{\sum_{i=1}^{k}\left(l_i - \bar{l}\right)\left(s_i - \bar{s}\right)}{\sum_{i=1}^{k}\left(l_i - \bar{l}\right)^2}$$

In some situations, the transformations cannot be done. The spread-versus-level plot and Levene statistic will not be produced if:

- a is a negative integer and at least one of the data is 0

- a is a negative non-integer and at least one of the data is less than or equal to 0
- a is a positive non-integer and at least one of the data is less than 0
- a is not specified and the median or the spread is less than or equal to 0

Levene Test of Homogeneity of Variances

The Levene test statistic is based on the transformed data and is defined by

$$L_a = \left(\frac{W-k}{k-1}\right)\frac{\sum_{i=1}^{k} w_i\left(\bar{z}_i - \bar{z}\right)^2}{\sum_{i=1}^{k}\sum_{l=1}^{m_i} c_{il}\left(z_{il} - \bar{z}_i\right)^2}$$

where

$$w_i = \sum_{l=1}^{m_i} c_{il}$$

$$\bar{x}_i = \frac{\sum_{l=1}^{m_i} c_{il} x_{il}}{w_i}$$

$$z_{il} = \left|x_{il} - \bar{x}_i\right|$$

$$\bar{z}_i = \sum_{l=1}^{m_i} \frac{c_{il} z_{il}}{w_i}$$

$$\bar{z} = \sum_{i=1}^{k} \frac{w_i \bar{z}_i}{W}$$

The significance of L_a is calculated from the F distribution with degrees of freedom $k-1$ and $W-k$.

Groups with zero variance are included in the test.

Robust Levene's Test of Homogeneity of Variances

With the current version of Levene's test L_a, the followings can be considered as options in order to obtain robust Levene's tests:

- Levene's test L_b based on $z_{li}^{(b)} = |x_{il} - \tilde{x}_i|$ where $\tilde{x}_i$ is the median of x_{il}'s for group i.

  Median calculation is done by the method requested. The default method is HAVERAGE. Once the $\tilde{x}_i$'s and hence $z_{li}^{(b)}$'s are calculated, apply the formula for L_a, shown in the section above, to obtain L_b by replacing z_{il}, $\bar{z}_i$ and $\bar{z}$ with $z_{li}^{(b)}$, $\bar{z}_i^{(b)}$ and $\bar{z}^{(b)}$ respectively.

  Two significances of L_b are given. One is calculated from a F-distribution with degrees of freedom k - 1 and W - k. Another is calculated from a F-distribution with degrees of freedom k - 1 and v. The value of v is given by:

$$\boldsymbol{v} = \frac{\left(\sum_{i=1}^{k} u_i\right)^2}{\left(\sum_{i=1}^{k} \frac{u_i^2}{v_i}\right)}$$

where

$$\boldsymbol{u_i} = \sum_{l=1}^{m_i} c_{il}\,(z_{il}^{(b)} - \bar{z}_i^{(b)})^2$$

in which

$$\bar{z}_i^{(b)} = \sum_{l=1}^{m_i} \frac{c_{il}\, z_{il}^{(b)}}{w_i}$$

and

$v_i = w_i - 1.$

- Levene's test L_c based on $z_{il}^{(c)} = |x_{il} - T_{i,0.9}|$ where $T_{i,0.9}$ is the 5% trimmed mean of x_{il}'s for group i.

  Once the $T_{i,0.9}$'s and hence $z_{il}^{(c)}$'s are calculated, apply the formula of L_a to obtain L_c by replacing z_{il}, $\bar{z}_i$ and $\bar{z}$ with $z_{li}^{(c)}$, $\bar{z}_i^{(c)}$ and $\bar{z}^{(c)}$ respectively.

  The significance of L_c is calculated from a F-distribution with degrees of freedom k - 1 and W - k.

Plots

Normal Probability Plot (NPPLOT)

For every distinct observation y_i, R_i is the rank (the mean of ranks is assigned to ties). The normal score NS_i is calculated by

$$NS_i = \mathbf{\Psi}^{-1}\left(\frac{R_i}{W+1}\right)$$

where $\mathbf{\Psi}^{-1}$ is the inverse of the standard normal cumulative distribution function. The NPPLOT is the plot of $(y_1, NS_1), \ldots, (y_m, NS_m)$.

Detrended Normal Plot

The detrended normal plot is the scatterplot of $(y_1, D_1), \ldots, (y_m, D_m)$, where D_i is the difference between the Z-score and normal score, which is defined by

$$D_i = Z_i - NS_i$$

and

$$Z_i = \frac{y_i - \bar{y}}{s}$$

where $\bar{y}$ is the average and s is the standard deviation.

Boxplot

The boundaries of the box are Tukey's hinges. The length of the box is the interquartile range based on Tukey's hinges. That is,

$$IQR = Q_3 - Q_1$$

Define

$$\boldsymbol{STEP = 1.5\ IQR}$$

A case is an outlier if

$$Q_3 + STEP \le y_i < Q_3 + 2STEP$$
or
$$Q_1 - 2STEP < y_i \le Q_1 - STEP$$

A case is an extreme if

$$y_i \ge Q_3 + 2STEP$$

or

$$y_i \le Q_1 - 2STEP$$

References

Brown, M. B., and Forsythe, A. B. (1974). Robust Tests for the Equality of Variances. *Journal of the American Statistical Association*, 69, p364-367.

Dallal, G. E., and Wilkinson, L. 1986. An analytic approximation to the distribution of Lilliefor's test statistic for normality. *The American Statistician*, **40**(4): 294–296 (Correction: **41**: 248).

Frigge, M., Hoaglin, D. C., and Iglewicz, B. 1987. Some implementations for the boxplot. In: *Computer Science and Statistics Proceedings of the* 19*th Symposium on the Interface*, R. M. Heiberger and M. Martin, eds. Alexandria, Va.: American Statistical Association.

Glaser, R. E. (1983). Levene's Robust Test of Homogeneity of Variances. *Encyclopedia of Statistical Sciences 4*. NY: Wiley, p608-610.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W. 1983. *Understanding robust and exploratory data analysis*. New York: John Wiley & Sons, Inc.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W. 1985. *Exploring data tables, trends, and shapes*. New York: John Wiley & Sons, Inc.

Lilliefors, H. W. 1967. On the Kolmogorov-Smirnov tests for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62: 399–402.

Loh, W. Y. (1987). Some Modifications of Levene's Test of Variance Homogeneity. *Journal of Statistical Computation and Simulation*, 28, p213-226.

Tukey, J. W. 1977. *Exploratory data analysis*. Reading, Mass.: Addison-Wesley.

Velleman, P. F., and Hoaglin, D. C. 1981. *Applications*, *basics*, *and computing of exploratory data analysis*. Boston: Duxbury Press.