A generalized boxplot for skewed and heavy-tailed distributions implemented in Stata

Vincenzo Verardi
joint with C. Vermandele and C. Bruffaerts

UK Stata users meeting

September 2014

LA LIBERTÉ DE CHERCHER

CENTRE DE RECHERCHE EN ÉCONOMIE DU DÉVELOPPEMENT (CRED)

UNIVERSITÉ DE NAMUR

Roadmap

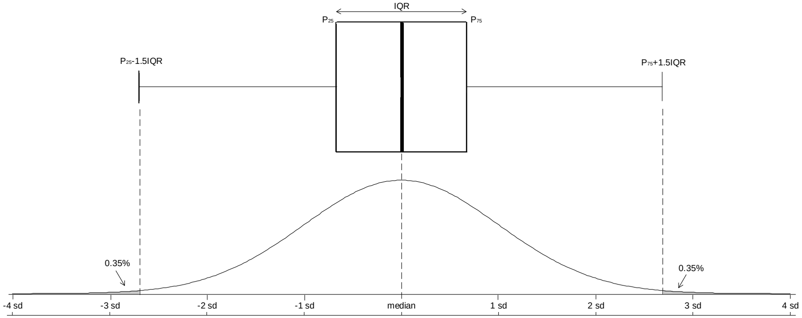
Structure of the presentation

- Introduction
- Preamble (Tukey g and h: $T_{g,h}$)
- A generalized boxplot
- Simulations
- Examples (Eartquakes in Latin America and Footballers' wages)
- Stata command
- Conclusion
- References

Univariate outliers identification

Standard Boxplot, Standard Normal distribution

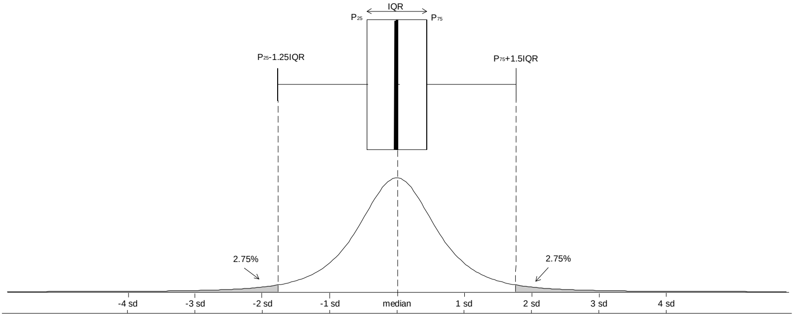
- X is the $(n \times 1)$ data vector (n individuals, 1 variable)



Univariate outliers identification

Standard Boxplot, heavy tailed t_2 distribution

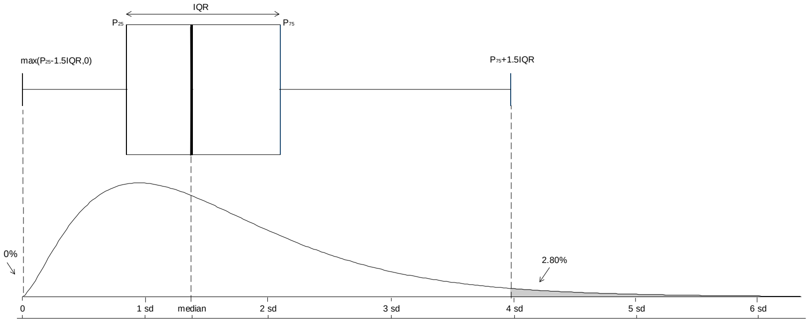
- X is the $(n \times 1)$ data vector (n individuals, 1 variable)



Univariate outliers identification

Standard Boxplot, skewed χ^2_5 distribution

- X is the $(n \times 1)$ data vector (n individuals, 1 variable)



Univariate outliers identification

Limitations of the boxplot

- Only suited for (almost) symmetric data and (approximately) mesokurtic distributions

Solution 1

Modify the whiskers of the boxplot to deal with asymmetry

- **Adjusted Boxplot** (Hubert and Vandervieren, 2008).
  - The whiskers of the boxplot are moved according to a robust measure of asymmetry, the medcouple ($-1 \leq MC \leq 1$):

$$\begin{cases} \left[Q_{0.25} - 1.5e^{-4\,\mathrm{MC}}\,\mathrm{IQR}\,;\,Q_{0.75} + 1.5e^{3\,\mathrm{MC}}\,\mathrm{IQR}\right] & \text{if MC} \geq 0 \\ \left[Q_{0.25} - 1.5e^{-3\,\mathrm{MC}}\,\mathrm{IQR}\,;\,Q_{0.75} + 1.5e^{4\,\mathrm{MC}}\,\mathrm{IQR}\right] & \text{if MC} < 0, \end{cases}$$

  - Copes well with asymmetry ($MC \leq 0.6$) but does not take (explicitly) into accout heavyness of tails
  - Rejection rate set to 0.7%
  - Rule based on simulations
  - Computational complexity $O(n \log n)$ (see Gelade et *al.*, 2014).

Univariate outliers identification

Limitations of the boxplot

- Only suited for (almost) symmetric data and (approximately) mesokurtic distributions

Solution 2

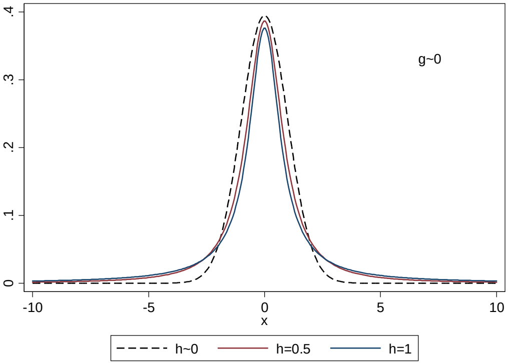
Modify the whiskers of the boxplot to deal with asymmetry and tail heavyness

- **Generalized Boxplot**
  - Do a rank preserving transformation of the data to end-up with a known distribution
  - Use the theoretical quantiles of the latter to set whiskers (after applying an inverse transformation)
  - Cope with both the skewness and tail heavyness
  - Set the desired rejection rate to any chosen level
  - Computational complexity $O(n)$ (as the standard boxplot)

Heavy-tailed distributions

Definition

If $Z \sim N(0,1)$, $g \neq 0$ and $h \in \mathbb{R}$, the random variable Y is said to be $T_{g,h}$ distributed if $Y = \frac{1}{g}\left[\exp\left(gZ\right) - 1\right] \exp\left(hZ^2/2\right)$

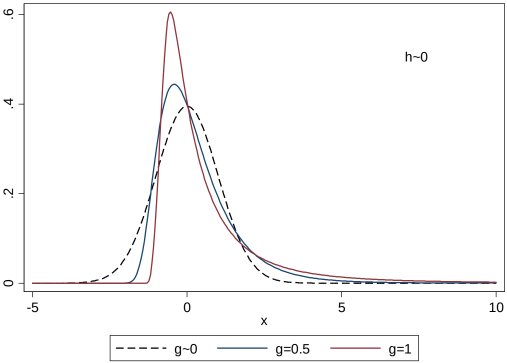


Preamble: Tukey g and h distribution

Asymmetrical distributions

Definition

If $Z \sim N(0,1)$, $g \neq 0$ and $h \in \mathbb{R}$, the random variable Y is said to be $T_{g,h}$ distributed if $Y = \frac{1}{g}\left[\exp\left(gZ\right)-1\right]\exp\left(hZ^2/2\right)$

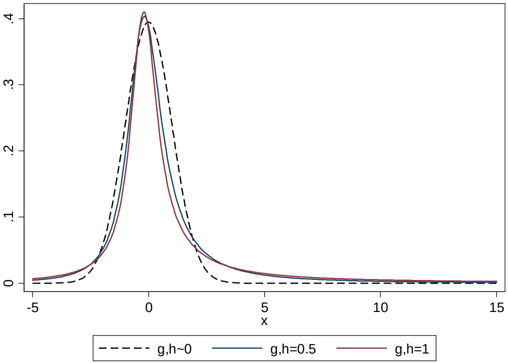


Preamble: Tukey g and h distribution

Asymmetrical and heavy-tailed distributions

Definition

If $Z \sim N(0,1)$, $g \neq 0$ and $h \in \mathbb{R}$, the random variable Y is said to be $T_{g,h}$ distributed if $Y = \frac{1}{g}\left[\exp\left(gZ\right) - 1\right] \exp\left(hZ^2/2\right)$



Univariate outliers identification

Standard Boxplot

- An outlier is defined as any observation lying outside the fence defined by whiskers $P_{25} - 1.5\,\mathrm{IQR}$ and $P_{75} + 1.5\,\mathrm{IQR}$

Theoretical detection rate α

- More generally, a theoretical detection rate equal to α is given by $[Q_{0.25} - c(\alpha)\,\mathrm{IQR}\,;\, Q_{0.75} + c(\alpha)\,\mathrm{IQR}]$ with $c(\alpha) = \frac{z_{1-\alpha/2} - z_{0.75}}{z_{0.75} - z_{0.25}}$ where z_p denotes the quantile of order p of the standard normal distribution.

Limitations of the boxplot

- Only suited for (almost) symmetric data and (approximately) mesokurtic distributions

Solution

- Modify the boxplot to deal with asymmetry and tail heavyness.

Procedure

Transformation

For an initial dataset $\{x_1, \ldots, x_n\}$, the guidelines of the new method are the following:

1. Center and reduce the data: $x_i^* = \frac{x_i - m^0}{s_0}$ where $s_0 = \mathrm{IQR}(\{x_j\})$ and $m_0 = \mathrm{Q}_{0.5}(\{x_j\})$

Procedure

Transformation

For an initial dataset $\{x_1, \ldots, x_n\}$, the guidelines of the new method are the following:

1. Center and reduce the data: $x_i^* = \frac{x_i - m^0}{s_0}$where $s_0 = \mathrm{IQR}(\{x_j\})$ and $m_0 = \mathrm{Q}_{0.5}(\{x_j\})$
2. Shift the dataset to obtain only strictly positive values: $r_i = x_i^* - \min(\{x_j^*\}) + 0.1$

Transformation

For an initial dataset $\{x_1, \ldots, x_n\}$, the guidelines of the new method are the following:

1. Center and reduce the data: $x_i^* = \frac{x_i - m^0}{s_0}$where $s_0 = \mathrm{IQR}(\{x_j\})$ and $m_0 = \mathrm{Q}_{0.5}(\{x_j\})$
2. Shift the dataset to obtain only strictly positive values: $r_i = x_i^* - \min(\{x_j^*\}) + 0.1$
3. Standardize r_i to map x_i on the open interval $(0, 1)$: $\widetilde{r}_i = \frac{r_i}{\min(\{r_j\}) + \max(\{r_j\})}$

Procedure

Transformation

For an initial dataset $\{x_1, \ldots, x_n\}$, the guidelines of the new method are the following:

1. Center and reduce the data: $x_i^* = \frac{x_i - m^0}{s_0}$ where $s_0 = \mathrm{IQR}(\{x_j\})$ and $m_0 = \mathrm{Q}_{0.5}(\{x_j\})$
2. Shift the dataset to obtain only strictly positive values: $r_i = x_i^* - \min(\{x_j^*\}) + 0.1$
3. Standardize r_i to map x_i on the open interval $(0, 1)$: $\widetilde{r}_i = \frac{r_i}{\min(\{r_j\}) + \max(\{r_j\})}$
4. Consider the inverse normal (also called probit) transformation $w_i = \Phi^{-1}(\widetilde{r}_i)$

Procedure

Transformation

For an initial dataset $\{x_1, \ldots, x_n\}$, the guidelines of the new method are the following:

1. Center and reduce the data: $x_i^* = \frac{x_i - m^0}{s_0}$ where $s_0 = \mathrm{IQR}(\{x_j\})$ and $m_0 = \mathrm{Q}_{0.5}(\{x_j\})$
2. Shift the dataset to obtain only strictly positive values: $r_i = x_i^* - \min(\{x_j^*\}) + 0.1$
3. Standardize r_i to map x_i on the open interval $(0, 1)$: $\widetilde{r}_i = \frac{r_i}{\min(\{r_j\}) + \max(\{r_j\})}$
4. Consider the inverse normal (also called probit) transformation $w_i = \Phi^{-1}(\widetilde{r}_i)$
5. Center and reduce the values w_i: $w_i^* = \frac{w_i - Q_{0.5}(\{w_j\})}{\mathrm{IQR}(\{w_j\})/1.3426}$

Transformation

6. Adjust the distribution of the values w_i^* $(i = 1, \ldots, n)$ by the Tukey $T_{\widehat{g}^*,\widehat{h}^*}$ distribution:

$$\widehat{g} = \frac{1}{z_{0.9}} \ln\left(-\frac{P_{0.9}\left(\{w_j^*\}\right)}{P_{0.1}\left(\{w_j^*\}\right)}\right), \qquad \widehat{h} = \frac{2\ln\left(-\widehat{g}\,\frac{P_{0.9}\left(\{w_j^*\}\right)P_{0.1}\left(\{w_j^*\}\right)}{P_{0.9}\left(\{w_j^*\}\right)+P_{0.1}\left(\{w_j^*\}\right)}\right)}{z_{0.9}^2}$$

Transformation

6. Adjust the distribution of the values w_i^* ($i = 1, \ldots, n$) by the Tukey $T_{\widehat{g}^*, \widehat{h}^*}$ distribution:

$$\widehat{g} = \frac{1}{z_{0.9}} \ln\left(-\frac{P_{0.9}\left(\{w_j^*\}\right)}{P_{0.1}\left(\{w_j^*\}\right)}\right), \qquad \widehat{h} = \frac{2\ln\left(-\widehat{g}\,\frac{P_{0.9}\left(\{w_j^*\}\right)P_{0.1}\left(\{w_j^*\}\right)}{P_{0.9}\left(\{w_j^*\}\right)+P_{0.1}\left(\{w_j^*\}\right)}\right)}{z_{0.9}^2}$$

7. Select the rejection bounds (L_-^*, L_+^*) using specific quantiles of the adjusted distribution (here $P_{0.35}$ and $P_{99.65}$)

Procedure

Transformation

6. Adjust the distribution of the values w_i^* $(i = 1, \ldots, n)$ by the Tukey $T_{\widehat{g}^*, \widehat{h}^*}$ distribution:

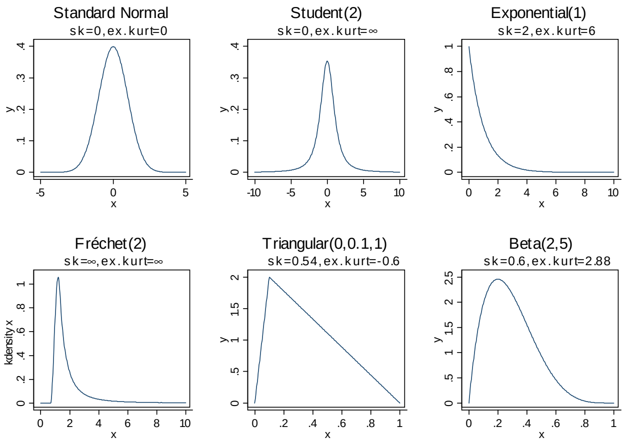
$$\widehat{g} = \frac{1}{z_{0.9}} \ln\left(-\frac{P_{0.9}\left(\{w_j^*\}\right)}{P_{0.1}\left(\{w_j^*\}\right)}\right), \qquad \widehat{h} = \frac{2\ln\left(-\widehat{g}\,\frac{P_{0.9}\left(\{w_j^*\}\right)P_{0.1}\left(\{w_j^*\}\right)}{P_{0.9}\left(\{w_j^*\}\right)+P_{0.1}\left(\{w_j^*\}\right)}\right)}{z_{0.9}^2}$$

7. Select the rejection bounds (L_-^*, L_+^*) using specific quantiles of the adjusted distribution (here $P_{0.35}$ and $P_{99.65}$)
8. Build the detection bounds B_-^* and B_+^* (whiskers) for the original dataset applying the complete inverse transformation

$$f(L_\pm^*) = \Phi\left(Q_{0.5}\left(\{w_j\}\right) + \frac{\mathrm{IQR}(\{w_j\})}{1.3426} L_\pm^*\right)$$

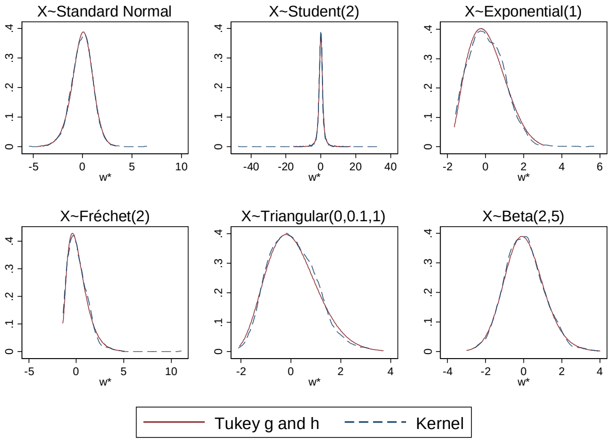
$$B_\pm^* = \left(f(L_\pm^*)\left[\min(\{r_j\}) + \max(\{r_j\})\right] + \min(\{x_j^*\}) - 0.1\right) s_0 + m_0$$

Considered distributions



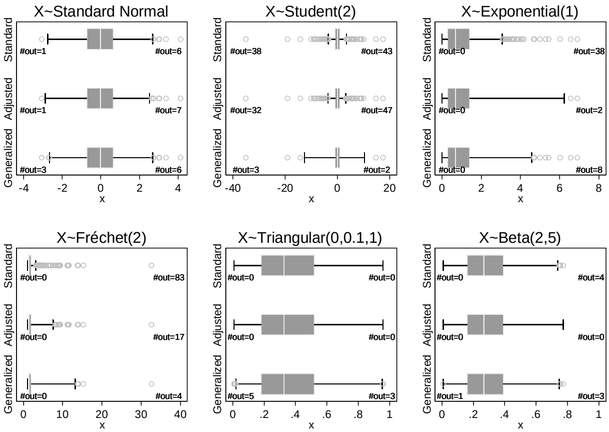
Numerical example

Quality of fit of transformed variable



Boxplot

Standard, Adjusted and Generalized boxplots



Outliers∼U(4.9,5.1) on the scale of the Normal (1000 replications)

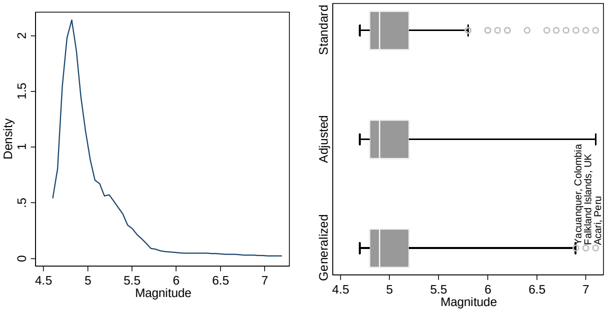
			Outliers: U(4.9,5.1)			
			Sensitivity		Specificity	
		ε	n=100	n=1000	n=100	n=1000
N(0,1)	Standard Boxplot	1%	100.00%	100.00%	99.06%	99.32%
		5%	100.00%	100.00%	99.19%	99.58%
	Adjusted Boxplot	1%	98.10%	100.00%	97.81%	99.12%
		5%	92.40%	100.00%	97.71%	98.95%
	Generalized Boxplot	1%	100.00%	100.00%	96.82%	98.91%
		5%	98.30%	100.00%	97.95%	99.45%
t2	Standard Boxplot	1%	100.00%	100.00%	91.96%	91.98%
		5%	100.00%	100.00%	92.95%	92.83%
	Adjusted Boxplot	1%	100.00%	100.00%	90.93%	91.70%
		5%	100.00%	100.00%	91.05%	91.54%
	Generalized Boxplot	1%	100.00%	100.00%	96.68%	98.47%
		5%	100.00%	100.00%	97.71%	99.15%
Exp(1)	Standard Boxplot	1%	100.00%	100.00%	94.93%	95.47%
		5%	100.00%	100.00%	96.29%	96.72%
	Adjusted Boxplot	1%	99.70%	100.00%	98.48%	99.47%
		5%	98.80%	100.00%	98.54%	99.90%
	Generalized Boxplot	1%	100.00%	100.00%	96.55%	99.35%
		5%	99.50%	100.00%	98.42%	99.95%

Sensitivity and Specificity

Outliers$\sim$U(4.9,5.1) on the scale of the Normal (1000 replications)

			Outliers: U(4.9,5.1)			
			Sensitivity		Specificity	
		ε	n=100	n=1000	n=100	n=1000
Fréchet(2)	Standard Boxplot	1%	✓ 100.00%	✓ 100.00%	✗ 91.96%	✗ 92.00%
		5%	✓ 100.00%	✓ 100.00%	✗ 93.31%	✗ 93.41%
	Adjusted Boxplot	1%	✓ 100.00%	✓ 100.00%	✗ 94.18%	! 95.26%
		5%	✓ 100.00%	✓ 100.00%	✗ 93.38%	✗ 94.54%
	Generalized Boxplot	1%	✓ 100.00%	✓ 100.00%	! 96.74%	! 98.96%
		5%	✓ 100.00%	✓ 100.00%	! 98.08%	✓ 99.57%
Triangular(0,0.1,1)	Standard Boxplot	1%	✓ 100.00%	✓ 100.00%	✓ 99.75%	✓ 99.83%
		5%	✓ 99.50%	✓ 100.00%	✓ 99.90%	✓ 99.95%
	Adjusted Boxplot	1%	✗ 53.20%	✗ 56.40%	✓ 99.39%	✓ 99.99%
		5%	✗ 34.40%	✗ 8.20%	✓ 99.23%	✓ 100.00%
	Generalized Boxplot	1%	! 98.90%	✓ 99.90%	! 96.71%	✓ 99.26%
		5%	✗ 93.80%	! 97.70%	! 97.67%	✓ 99.86%
Beta(2,5)	Standard Boxplot	1%	✓ 100.00%	✓ 100.00%	! 98.81%	✓ 99.42%
		5%	✓ 100.00%	✓ 100.00%	✓ 99.15%	✓ 99.72%
	Adjusted Boxplot	1%	✗ 76.40%	! 98.70%	✓ 99.07%	✓ 99.97%
		5%	✗ 54.60%	✗ 71.10%	! 98.62%	✓ 99.94%
	Generalized Boxplot	1%	✓ 99.60%	✓ 100.00%	! 97.26%	✓ 99.36%
		5%	! 98.48%	✓ 99.60%	! 98.48%	✓ 99.79%

Estimated medcouple: 0.43



Stata command

Syntax

`box_out` *varname* `[if] [in] [,out(`*varname*`) bdp(#) perc(#) nograph]`

Options

- *out:* Identifies the new variable to be created to identify individuals outside the fence defined by the whiskers
- *bdp:* Sets the desired Break-down point (in %). It is 10% by default
- *perc:* Sets the desired percentage of points outside the whiskers in case of uncontaminated data. It is set to 0.7% by default
- *nograph:* Suppresses the graph

Saved results and output

- *e(g), e(h):* Estimated skewness and elongation parameters of the underlying Tukey g and h distribution
- *e(lowerW), e(upperW):* Value of the lower and upper whiskers
- A basic boxplot is created but we recommend to refer to N. J. Cox, S.J. (2009) for better output

Conclusion

Generalized boxplot

We propose a very simple generalized boxplot that

- is suited for skewed and/or heavy-tailed distributions
- allows for setting the desired detection rate of atypical observation
- has a computational complexity of $O(n)$

In Stata

We provide a simple command that

- estimates the whiskers of the generalized boxplot
- creates a simple boxplot.
- we however refer to Cox (2009) and Cox(2013) for more complete graphs.

Complementary results

In multivariate analysis we have a projection based estimator

- to create a bagplot in 2D
- identify outliers for multivariate skewed and heavy-tailed distributions

References

- Bruffaerts, C., Verardi, V. and Vermandele, C., (2014). "A generalized boxplot for skewed and heavy-tailed distributions". *Statistics and Probability Letters* (forthcoming)
- Cox, N.J. (2013)."Speaking Stata: Creating and varying box plots: Correction". *Stata Journal* 13(2). pp. 398-400.
- Cox, N.J. (2009)."Speaking Stata: Creating and varying box plots". *Stata Journal* 9(3). pp. 478-496.
- Gelade, W., Verardi, V. and Vermandele, C. (2014). "Time efficient algorithms for robust estimators of location, scale, symmetry and tail heaviness". *Stata Journal* (forthcoming).
- Hubert, M. and Vendervieren, E. (2008). "An adjusted boxplot for skewed distributions". *Computational Statistics & Data Analysis* 52(12). pp 5186-5201