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
Roadmap

Structure of the presentation

- Introduction
- Preamble (Tukey $g$ and $h$: $T_{g,h}$)
- A generalized boxplot
- Simulations
- Examples (Earthquakes 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 $\chi^2_5$ 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 ($-1 \leq MC \leq 1$):
    $$\begin{cases} 
    \left[ Q_{0.25} - 1.5e^{-4MC} IQR ; Q_{0.75} + 1.5e^{3MC} IQR \right] & \text{if } MC \geq 0 \\
    \left[ Q_{0.25} - 1.5e^{-3MC} IQR ; Q_{0.75} + 1.5e^{4MC} IQR \right] & \text{if } MC < 0,
    \end{cases}$$
  - Copes well with asymmetry ($MC \leq 0.6$) but does not take (explicitly) into account heaviness 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)
Preamble: Tukey g and h distribution

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

$$Y = \frac{1}{g} \left[ \exp (gZ) - 1 \right] \exp \left( hZ^2 / 2 \right)$$
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 $Y = \frac{1}{g} \left[ \exp (gZ) - 1 \right] \exp \left( hZ^2 / 2 \right)$
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 $Y = \frac{1}{g} \left[ \exp(gZ) - 1 \right] \exp \left( hZ^2 / 2 \right)$
Univariate outliers identification

**Standard Boxplot**

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

**Theoretical detection rate $\alpha$**

- More generally, a theoretical detection rate equal to $\alpha$ is given by $[Q_{0.25} - c(\alpha) \text{IQR}; Q_{0.75} + c(\alpha) \text{IQR}]$ 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 heaviness.
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 \( s_0 = \text{IQR}(\{x_j\}) \) and \( m_0 = \text{Q}_0.5(\{x_j\}) \)
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 \( s_0 = \text{IQR}(\{x_j\}) \) and \( m_0 = \text{Q}_{0.5}(\{x_j\}) \)

2. Shift the dataset to obtain only strictly positive values:
   \( r_i = x_i^* - \min(\{x_j^*\}) + 0.1 \)
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 \( s_0 = \text{IQR}(\{x_j\}) \) and \( m_0 = Q_{0.5}(\{x_j\}) \)

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 \((0, 1)\):
   \[
   \tilde{r}_i = \frac{r_i}{\min(\{r_j\}) + \max(\{r_j\})}
   \]
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 
   \[ s_0 = \text{IQR}(\{x_j\}) \]
   \[ m_0 = Q_{0.5}(\{x_j\}) \]

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 \((0, 1)\):
   \[ \tilde{r}_i = \frac{r_i}{\min(\{r_j\}) + \max(\{r_j\})} \]

4. Consider the inverse normal (also called probit) transformation
   \[ w_i = \Phi^{-1}(\tilde{r}_i) \]
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 \( s_0 = \text{IQR}(\{x_j\}) \) and \( m_0 = Q_{0.5}(\{x_j\}) \)

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 \((0, 1)\):
   \( \tilde{r}_i = \frac{r_i}{\min(\{r_j\}) + \max(\{r_j\})} \)

4. Consider the inverse normal (also called probit) transformation
   \( w_i = \Phi^{-1}(\tilde{r}_i) \)

5. Center and reduce the values \( w_i \):
   \( w_i^* = \frac{w_i - Q_{0.5}(\{w_j\})}{\text{IQR}(\{w_j\})/1.3426} \)
Procedure

Transformation

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

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

### Transformation

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

\[
\hat{g} = \frac{1}{z_{0.9}} \ln \left( - \frac{P_{0.9}(\{w_j^*\})}{P_{0.1}(\{w_j^*\})} \right), \quad \hat{h} = \frac{2 \ln \left( -\hat{g} \frac{P_{0.9}(\{w_j^*\})}{P_{0.9}(\{w_j^*\}) + P_{0.1}(\{w_j^*\})} \right)}{z_{0.9}^2}
\]

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 $T_{\hat{g}^*, \hat{h}^*}$ distribution:

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

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

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

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

Considered distributions

- **Standard Normal**
  - $sk = 0, ex.kurt = 0$

- **Student(2)**
  - $sk = 0, ex.kurt = \infty$

- **Exponential(1)**
  - $sk = 2, ex.kurt = 6$

- **Fréchet (2)**
  - $sk = \infty, ex.kurt = \infty$

- **Triangular(0, 0.1, 1)**
  - $sk = 0.6, ex.kurt = 2.88$

- **Beta(2, 5)**
  - $sk = 0.54, ex.kurt = -0.6$
Numerical example

Quality of fit of transformed variable

Tukey g and h Kernel

Vincenzo Verardi (UK Stata users meeting)
Standard, Adjusted and Generalized boxplots

1. X~Standard Normal
2. X~Student(2)
3. X~Exponential(1)
4. X~Fréchet(2)
5. X~Triangular(0.0.1,1)
6. X~Beta(2,5)
## Sensitivity and Specificity

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

<table>
<thead>
<tr>
<th>Outliers: U(4.9,5.1)</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\epsilon$ n=100</td>
<td>n=1000 n=100 n=1000</td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>100.00% 100.00% 99.06% 99.32%</td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>100.00% 100.00% 99.19% 99.58%</td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>98.10% 100.00% 97.81% 99.12%</td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>92.40% 100.00% 97.71% 98.95%</td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>100.00% 100.00% 96.82% 98.91%</td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>98.30% 100.00% 97.95% 99.45%</td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>100.00% 100.00% 91.96% 91.98%</td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>100.00% 100.00% 92.95% 92.83%</td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>100.00% 100.00% 90.93% 91.70%</td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>100.00% 100.00% 91.05% 91.54%</td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>100.00% 100.00% 96.68% 98.47%</td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>100.00% 100.00% 97.71% 99.15%</td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>100.00% 100.00% 94.93% 95.47%</td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>100.00% 100.00% 96.29% 96.72%</td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>99.70% 100.00% 98.48% 99.47%</td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>98.80% 100.00% 98.54% 99.90%</td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>100.00% 100.00% 96.55% 99.35%</td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>99.50% 100.00% 98.42% 99.95%</td>
<td></td>
</tr>
</tbody>
</table>
### Sensitivity and Specificity

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

<table>
<thead>
<tr>
<th>Fréchet(2)</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Boxplot</td>
<td>ε=1% 100.00% 100.00% 91.96% 92.00%</td>
<td>ε=5% 100.00% 100.00% 93.31% 93.41%</td>
</tr>
<tr>
<td>Adjusted Boxplot</td>
<td>ε=1% 100.00% 100.00% 94.18% 95.26%</td>
<td>ε=5% 100.00% 100.00% 93.38% 94.54%</td>
</tr>
<tr>
<td>Generalized Boxplot</td>
<td>ε=1% 100.00% 100.00% 96.74% 98.96%</td>
<td>ε=5% 100.00% 100.00% 98.08% 99.57%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Triangular(0, 0.1, 1)</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Boxplot</td>
<td>ε=1% 100.00% 100.00% 99.75% 99.83%</td>
<td>ε=5% 99.50% 100.00% 99.90% 99.95%</td>
</tr>
<tr>
<td>Adjusted Boxplot</td>
<td>ε=1% 53.20% 56.40% 99.39% 99.99%</td>
<td>ε=5% 34.40% 8.20% 99.23% 100.00%</td>
</tr>
<tr>
<td>Generalized Boxplot</td>
<td>ε=1% 98.90% 99.90% 96.71% 99.26%</td>
<td>ε=5% 93.80% 97.70% 97.67% 99.86%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Beta(2, 5)</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Boxplot</td>
<td>ε=1% 100.00% 100.00% 98.81% 99.42%</td>
<td>ε=5% 76.40% 98.70% 99.07% 99.97%</td>
</tr>
<tr>
<td>Adjusted Boxplot</td>
<td>ε=1% 76.40% 98.70% 99.07% 99.97%</td>
<td>ε=5% 54.60% 71.10% 98.62% 99.94%</td>
</tr>
<tr>
<td>Generalized Boxplot</td>
<td>ε=1% 99.60% 100.00% 97.26% 99.36%</td>
<td>ε=5% 98.48% 99.60% 98.48% 99.79%</td>
</tr>
</tbody>
</table>
Example 1: Daily earnings of 50 top football players

Estimated medcouple: 0.12
Example 2: 200 earthquakes in Latin America (2013)

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


