Determination of Unbiased Estimates of Characteristic Values

Goran Turk and Dejan Zupan

University of Ljubljana
Faculty of Civil and Geodetic Engineering
Statistical distributions

- Normal distribution: dead load, dimensions of the structure, strength of the concrete, etc.
- Lognormal distribution: snow load, strength of the steel, etc.
- Gamma distribution: sustained load, etc.
- Gumbel distribution: max. life load, max. wind load, max. ground acceleration, etc.
- Frechet distribution: max. sea wave height, etc.
Characteristic values from different distributions:

- Normal distribution:
  - $m_X = 15.0$
  - $\sigma_X = 3.75$
  - $x_\alpha = 8.83$

- Lognormal distribution:
  - $m_X = 15.0$
  - $\sigma_X = 3.75$
  - $x_\alpha = 9.71$

- Gamma distribution:
  - $m_X = 15.0$
  - $\sigma_X = 3.75$
  - $x_\alpha = 9.41$

- Gumbel distribution:
  - $m_X = 15.0$
  - $\sigma_X = 3.75$
  - $x_\alpha = 8.00$

- Frechet distribution:
  - $m_X = 15.0$
  - $\sigma_X = 3.75$
  - $x_\alpha = 11.07$
Normal distribution

- Characteristic value can be determined by

\[ U = \frac{X - m_X}{\sigma_X} \rightarrow \frac{x_\alpha - m_X}{\sigma_X} = F_U^{-1}(\alpha) \rightarrow x_\alpha = m_X + \sigma_X F_U^{-1}(\alpha). \]

- Mean \( m_X \) and variance \( \sigma^2_X \) are not known. Only estimates can be obtained from the sample.

\[
\hat{m}_X = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{and} \quad \hat{\sigma}^2_X = S^*_X = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{X})^2.
\]

- From each sample we get different estimates of the mean and variance \( \bar{X} \) in \( S^*_X \).

- The simplest (naive) estimate of characteristic value is:

\[
\hat{X}_\alpha = \bar{X} + S^*_X F_U^{-1}(\alpha).
\]
Normal distribution - cont.
Characteristic value determination

- The characteristic value is in general determined by
  \[ \hat{X}_\alpha = \bar{X} + \lambda S_X^* \].

- Since the mean and variance of \( X \) is not known, only the estimate of characteristic value \( \hat{X}_\alpha \) can be determined. The actual characteristic value \( x_\alpha \) is not known.

- The estimate \( \hat{X}_\alpha \) is a random variable, whose mean should coincide with the actual characteristic value \( x_\alpha \) (in the case of unbiased estimate).

- Instead of unbiased estimate one may prefer confidence interval with chosen probability \( \alpha_\lambda \) such that
  \[ P[\hat{X}_\alpha < x_\alpha] = 1 - \alpha_\lambda. \]

- In order to obtain the unbiased estimate or confidence interval the distribution of \( \hat{X}_\alpha \) has to be determined.
Distribution of characteristic value $\hat{X}_\alpha$

- The characteristic value $\hat{X}_\alpha = \bar{X} + \lambda S^*_X$ is distributed by non-central $t$ distribution only in the case of normally distributed $X$.

- After the transformation

$$P \left[ \hat{X}_\alpha < x_\alpha \right] = P \left[ \bar{X} + S^*_X F^{-1}_U (\alpha) < m_X + \sigma_X F^{-1}_U (\alpha) \right]$$

$$= P \left[ \frac{\bar{X} - m_X}{\sigma_X} + \frac{S^*_X F^{-1}_U (\alpha)}{\sigma_X} < F^{-1}_U (\alpha) \right]$$

the random variable $Z$ is defined as

$$Z = \frac{\bar{X} - m_X}{\sigma_X} + F^{-1}_U (\alpha) \frac{S^*_X}{\sigma_X} = \frac{U}{\sqrt{n}} + V,$$

where the distribution of $U$ is standardised normal, while the distribution of $V$ can be derived from $\chi^2$ distribution.

- The distribution of $Z$ is obtained by convolution of $U/\sqrt{n}$ and $V$.

$$f_Z(z) = \int_{-\infty}^{\infty} f_V(v) f_{U/\sqrt{n}}(z - v) \, dv.$$
How good the characteristic value $\hat{X}_\alpha$ is?

- Probability that $\hat{X}_\alpha$ is smaller than $x_\alpha$

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>10</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.415</td>
<td>0.432</td>
<td>0.442</td>
<td>0.449</td>
<td>0.453</td>
<td>0.457</td>
<td>0.462</td>
<td>0.489</td>
</tr>
<tr>
<td>0.05</td>
<td>0.402</td>
<td>0.421</td>
<td>0.433</td>
<td>0.440</td>
<td>0.446</td>
<td>0.450</td>
<td>0.456</td>
<td>0.487</td>
</tr>
<tr>
<td>0.01</td>
<td>0.387</td>
<td>0.409</td>
<td>0.422</td>
<td>0.431</td>
<td>0.437</td>
<td>0.442</td>
<td>0.449</td>
<td>0.485</td>
</tr>
</tbody>
</table>

- We can see larger difference from 0.5 (which would be a desired value) for smaller sample sizes.
Improved characteristic value $\hat{X}_{\alpha,\lambda}$

- Characteristic value $\hat{X}_{\alpha,\lambda}$ is determined so that the following equation holds
  \[ P \left[ \hat{X}_{\alpha,\lambda} < x_{\alpha} \right] = 1 - \alpha_{\lambda}. \]

- Characteristic value $\hat{X}_{\alpha,\lambda}$ is thus defined as
  \[ \hat{X}_{\alpha,\lambda} = \bar{X} + \lambda S_{\bar{X}}^*. \]

- Similarly as before a random variable $Z_{\lambda}$ is defined
  \[ Z_{\lambda} = \frac{\bar{X} - m_X}{\sigma_X} + \lambda \frac{S_{\bar{X}}^*}{\sigma_X} = \frac{U}{\sqrt{n}} + V_{\lambda}. \]

- $\lambda$ is to be found from equation (e.g. $\alpha_{\lambda} = 0.5$)
  \[ F_{Z_{\lambda}} \left( F_{U}^{-1}(\alpha) \right) = 1 - \alpha_{\lambda} = 0.5. \]

<table>
<thead>
<tr>
<th>$n$</th>
<th>$3$</th>
<th>$4$</th>
<th>$5$</th>
<th>$6$</th>
<th>$7$</th>
<th>$8$</th>
<th>$10$</th>
<th>$100$</th>
<th>$\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$</td>
<td>1.938</td>
<td>1.830</td>
<td>1.780</td>
<td>1.750</td>
<td>1.732</td>
<td>1.719</td>
<td>1.702</td>
<td>1.650</td>
<td>1.645</td>
</tr>
</tbody>
</table>
Unbiased characteristic value $\hat{X}_{\alpha, \varepsilon}$

- Expected value of the characteristic value $\hat{X}_{\alpha, \varepsilon}$ is

$$E[\hat{X}_{\alpha, \varepsilon}] = \int_{-\infty}^{\infty} \tilde{x} f_{\hat{X}_{\alpha, \varepsilon}}(\tilde{x}) d\tilde{x} = x_{\alpha} \rightarrow$$

$$E \left[ \bar{X} + \varepsilon S^*_X \right] = m_X + \sigma_X F_{U}^{-1}(\alpha) \rightarrow E \left[ \frac{\bar{X} - m_X}{\sigma_X} + \varepsilon \frac{S^*_X}{\sigma_X} \right] = F_{U}^{-1}(\alpha).$$

- The correction factor $\varepsilon$ is determined from the equation

$$Z_{\varepsilon} = \frac{\bar{X} - m_X}{\sigma_X} + \varepsilon \frac{S^*_X}{\sigma_X} = \frac{U}{\sqrt{n}} + V_{\varepsilon} \rightarrow E[Z_{\varepsilon}] = F_{U}^{-1}(\alpha).$$

<table>
<thead>
<tr>
<th>$n$</th>
<th>$3$</th>
<th>$4$</th>
<th>$5$</th>
<th>$6$</th>
<th>$7$</th>
<th>$8$</th>
<th>$10$</th>
<th>$100$</th>
<th>$\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$</td>
<td>1.938</td>
<td>1.830</td>
<td>1.780</td>
<td>1.750</td>
<td>1.732</td>
<td>1.719</td>
<td>1.702</td>
<td>1.650</td>
<td>1.645</td>
</tr>
<tr>
<td>$\varepsilon$</td>
<td>1.856</td>
<td>1.785</td>
<td>1.750</td>
<td>1.729</td>
<td>1.715</td>
<td>1.705</td>
<td>1.691</td>
<td>1.649</td>
<td>1.645</td>
</tr>
</tbody>
</table>
Comparison with some of European codes

- Different European codes propose different procedures for the characteristic value determination:

<table>
<thead>
<tr>
<th>$n$</th>
<th>5</th>
<th>6</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive estimate</td>
<td>$-1.645$</td>
<td>$-1.645$</td>
<td>$-1.645$</td>
<td>$-1.645$</td>
<td>$-1.645$</td>
<td>$-1.645$</td>
</tr>
<tr>
<td>ENV 1991-1*</td>
<td>$-1.80$</td>
<td>$-1.77$</td>
<td>$-1.72$</td>
<td>$-1.68$</td>
<td>$-1.67$</td>
<td></td>
</tr>
<tr>
<td>ENV 1991-1**</td>
<td>$-2.33$</td>
<td>$-2.18$</td>
<td>$-1.92$</td>
<td>$-1.76$</td>
<td>$-1.73$</td>
<td></td>
</tr>
<tr>
<td>ENV 206</td>
<td>$-1.87$</td>
<td>$-1.62$</td>
<td>$-1.48$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zupan, Turk*</td>
<td>$-1.645$</td>
<td>$-1.645$</td>
<td>$-1.645$</td>
<td>$-1.645$</td>
<td>$-1.645$</td>
<td>$-1.645$</td>
</tr>
<tr>
<td>Zupan, Turk**</td>
<td>$-1.750$</td>
<td>$-1.729$</td>
<td>$-1.691$</td>
<td>$-1.675$</td>
<td>$-1.667$</td>
<td>$-1.659$</td>
</tr>
</tbody>
</table>

* $V_X = \frac{\sigma_X}{m_X}$ known

** $V_X = \frac{\sigma_X}{m_X}$ not known
Lognormal distribution

- Lognormal distribution is derived from normal distribution
  \[ Y = e^X \quad \text{or} \quad X = \ln Y. \]

- Same transformation is used for characteristic value estimate \( \hat{Y}_{\alpha,\lambda} \)
  \[ \hat{Y}_{\alpha,\lambda} = e^{\hat{X}_{\alpha,\lambda}} = e^{\bar{X} + \lambda S^*_X} = e^{\bar{X}} e^{\lambda S^*_X}. \]

- The relations between the moments of both distributions are
  \[ \bar{X} = \ln \left( \frac{\bar{Y}^2}{\sqrt{(S^*_Y)^2 + \bar{Y}^2}} \right) \quad \text{and} \quad S^*_X = \ln \left( \frac{(S^*_Y)^2}{\bar{Y}^2} + 1 \right). \]

- Finally we obtain the equation for the characteristic value
  \[ \hat{Y}_{\alpha,\lambda} = \frac{\bar{Y}^2}{\sqrt{(S^*_Y)^2 + \bar{Y}^2}} e^{\lambda \sqrt{\ln \left( \frac{(S^*_Y)^2}{\bar{Y}^2} + 1 \right)}}. \]

- Use the same values of \( \lambda \) (or \( \varepsilon \)) as for normally distributed parameter.
Simulations

- The correctness of the characteristic value determination was tested by Monte Carlo simulations.
- In each simulation the sample is generated \((m_X = 30, \sigma_X = 6)\).
- In each simulation the estimates \(\bar{X}\) in \(S_X^*\) and the characteristic value estimate \(\hat{X}_\alpha\) according to different methods are evaluated.

**Normal distribution:** The probability \(P[\hat{X}_{\alpha,\lambda} < x_{\alpha}]\) for \(\alpha \lambda = 0.5\)

<table>
<thead>
<tr>
<th></th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>(P[\hat{X}<em>{\alpha,\lambda} &lt; x</em>{\alpha}])</td>
<td>0.500282</td>
<td>0.500002</td>
<td>0.500133</td>
<td>0.499598</td>
<td>0.500515</td>
</tr>
</tbody>
</table>

**Lognormal distribution:** The probability \(P[\hat{Y}_{\alpha,\lambda} < y_{\alpha}]\) for \(\alpha \lambda = 0.5\)

<table>
<thead>
<tr>
<th></th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>(P[\hat{Y}<em>{\alpha,\lambda} &lt; y</em>{\alpha}])</td>
<td>0.505971</td>
<td>0.506795</td>
<td>0.507904</td>
<td>0.508235</td>
<td>0.50825</td>
</tr>
</tbody>
</table>
Comparison with some of European codes

- The average of characteristic values is computed and compared with the actual value $x_\alpha = 20.13$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>3</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>naive estimate</td>
<td>21.26</td>
<td>20.61</td>
<td>20.48</td>
<td>20.40</td>
<td>20.31</td>
</tr>
<tr>
<td>rel. error</td>
<td>5.6%</td>
<td>3.8%</td>
<td>1.7%</td>
<td>1.3%</td>
<td>0.9%</td>
</tr>
<tr>
<td>ENV 1991</td>
<td>12.09</td>
<td>17.55</td>
<td>18.41</td>
<td>18.79</td>
<td></td>
</tr>
<tr>
<td>rel. error</td>
<td>-39.9%</td>
<td>-12.8%</td>
<td>-8.5%</td>
<td>-6.7%</td>
<td></td>
</tr>
<tr>
<td>ENV 206</td>
<td>19.32</td>
<td>20.04</td>
<td>20.54</td>
<td>21.28</td>
<td></td>
</tr>
<tr>
<td>rel. error</td>
<td>-4.0%</td>
<td>-0.4%</td>
<td>2.0%</td>
<td>5.7%</td>
<td></td>
</tr>
<tr>
<td>rel. error</td>
<td>0.04%</td>
<td>-0.01%</td>
<td>-0.002%</td>
<td>-0.02%</td>
<td>0.01%</td>
</tr>
</tbody>
</table>
Conclusions

- The parameter $\lambda$ used in estimation of characteristic values depends on sample size.
- Values $\lambda$ are given by the codes, but are not unified (see e.g. ENV 1991-1, ENV 1993-1-1, ENV 1995-1-1, ENV 206).
- By the analytical derivation and numerical verification the correct values of $\lambda$ are obtained.
- Similarly, the unbiased estimates of the characteristic values can be obtained for other distributions of random variable $X$.
- The meaning of the characteristic value estimate should be clearly defined.
- The method of characteristic value estimation should be unified for all European standards and given for all distributions in use.