ANALISIS FREKUENSI

Statistika dan Probabilitas
Mengacu kepada paparan pada 

- The Second ASEAN Natural Disaster Conference
  - University of Yangon, Myanmar
  - 29-30 September 2014
FREQUENCY ANALYSIS ON EXTREME HYDROLOGIC DATA

Istiarto
Master of Engineering in Natural Disaster Management
Universitas Gadjah Mada, INDONESIA
Engineers (hydraulics, hydrology) are frequently required to derive design values on hydrologic events, such as flow discharge or rainfall depth (precipitation).

- Statistical method applied to time series data is a common practice
  - Fit those data to theoretical probability distributions
  - Use cumulative distribution function of the selected probability distribution to predict magnitude of hydrologic events
  - Define the design value(s) based on the magnitude of hydrologic event having particular probability of occurrence

This method is known as frequency analysis.
Frequency Analysis

- Time series (annual or partial time series of hydrologic event data)
- Fitting of the time series to theoretical distributions
- Predicting probability of occurrence of the hydrologic events
- Define design values of the hydrologic events
Common probability distributions applicable to hydrologic events

Probability Distributions

Theoretical Probability Distributions

- Gumbel
- Log Normal
- Log Pearson Type III
- Normal
The theoretical probability distributions
- their pdf and cdf have complex expression so that they are not readily solvable
- graphical method is a practical solution to fit data to the theoretical distributions

Ven te Chow et al. (1988) developed a transformation coordinate such that the cdf of a theoretical distribution shows as a straight line

fitting data to theoretical distribution and predicting probability of occurrence can be easily carried out
Transformed probability coordinate

- Chow et al. (1988)

\[ y_T = \bar{y} + K_T \cdot s \]

- \( y_T \) is the hydrologic magnitude at T-year return period,
- \( \bar{y} \) is the average value of the data,
- \( K_T \) is a frequency factor, and
- \( s \) is the standard deviation of the data.
Chow et al. (1988)

\[ y_T = \bar{y} + K_T \cdot s \]

- \(
\bar{y}
\)
- \(s\) constant

Plot of \((y_T \text{ vs } K_T)\) is a straight line

\(K_T\) is function of \(T\) and distribution

\[
\text{prob}(Y < y_T) = 1 - \frac{1}{T} \iff T = \frac{1}{1 - \text{prob}(Y < y_T)}
\]
Normal Distribution

- $y_T$
- $T$ or $\text{prob}(Y < y_T)$
- $y_T$
- $K_T$
Normal Distribution

\[ K_T = z \]

\[ z = w - \frac{2.515517 + 0.802853w + 0.010328w^2}{1 + 1.432788w + 0.189269w^2 + 0.001308w^3} \]

\[ w = \begin{cases} \left[ \ln \left( \frac{1}{p^2} \right) \right]^{1/2} & 0 < p = 1/T < 0.5 \\ \left[ \ln \left( \frac{1}{(1-p)^2} \right) \right]^{1/2} & p = 1/T > 0.5 \end{cases} \]
In the past

- Probability papers were available in the market
  - Normal, Log Normal, Gumbel distributions
  - Not available for Log Pearson Type III since $K_T$ changes with data
    - Log Pearson Type III was plotted on Log Normal paper
    - It shows as a curve, not a straight line

- Data were manually plotted on the paper
  - Scatter data plot and theoretical line give good insight into the distribution of the data
  - Fitting the data to the theoretical line can easily be performed
Nowadays

- Frequency analysis can easily be carried out by spreadsheet computer application
  - Unfortunately, spreadsheet cannot provide data plot the way probability paper does
  - Most of frequency analysis application program do not provide data plot at all
  - Lost of insight into data pattern (distribution)
  - Lost of ‘physical’ meaning of the computed values
**JUDUL DATA :** Debit Maksimum DPS Citarum

**Cara Urut Data :** b

Ket. : B = urutan debit besar ke kecil ; K = kecil ke besar

**UJI CHI-SQUARE**

**Jumlah kelas :** 5

Ket. : Jumlah kelas yang dikehendaki untuk uji Chi-Kuadrat

**Confidence Interval :** 0.05

Ket. : Derajat Ketidakpercayaan yang diinginkan

**KALA-ULANG**

**Jumlah kasus :** 8

Ket. : Jumlah probabilitas yang dikehendaki

<table>
<thead>
<tr>
<th>Tahun</th>
<th>Debit (m³/dt)</th>
<th>Probabilitas</th>
</tr>
</thead>
<tbody>
<tr>
<td>1918</td>
<td>244.00</td>
<td>.900</td>
</tr>
<tr>
<td>1919</td>
<td>217.00</td>
<td>.500</td>
</tr>
<tr>
<td>1920</td>
<td>285.00</td>
<td>.200</td>
</tr>
<tr>
<td>1921</td>
<td>261.00</td>
<td>.100</td>
</tr>
<tr>
<td>1922</td>
<td>295.00</td>
<td>.050</td>
</tr>
<tr>
<td>1923</td>
<td>252.00</td>
<td>.020</td>
</tr>
<tr>
<td>1924</td>
<td>275.00</td>
<td>.010</td>
</tr>
<tr>
<td>1925</td>
<td>204.00</td>
<td>.001</td>
</tr>
<tr>
<td>1926</td>
<td>208.00</td>
<td></td>
</tr>
<tr>
<td>1927</td>
<td>194.00</td>
<td></td>
</tr>
<tr>
<td>1928</td>
<td>256.00</td>
<td></td>
</tr>
<tr>
<td>1929</td>
<td>207.00</td>
<td></td>
</tr>
<tr>
<td>1930</td>
<td>354.00</td>
<td></td>
</tr>
<tr>
<td>1931</td>
<td>445.00</td>
<td></td>
</tr>
<tr>
<td>1932</td>
<td>350.00</td>
<td></td>
</tr>
<tr>
<td>1933</td>
<td>336.00</td>
<td></td>
</tr>
<tr>
<td>1934</td>
<td>328.00</td>
<td></td>
</tr>
<tr>
<td>1973</td>
<td>269.00</td>
<td></td>
</tr>
<tr>
<td>1974</td>
<td>323.00</td>
<td></td>
</tr>
</tbody>
</table>

**PROSES**

Example of frequency analysis performed by using spreadsheet application/computer program
**KALA-ULANG Debit Maksimum DPS Citarum**

<table>
<thead>
<tr>
<th>$P(x \geq X_m)$ Probabilitas</th>
<th>$T$ Kala-Ulang</th>
<th>Karakteristik Debit ($m^3/dt$) Menurut Probabilitasnya</th>
<th><strong>NORMAL</strong></th>
<th><strong>LOG-NORMAL</strong></th>
<th><strong>GUMBEL</strong></th>
<th><strong>LOG-PEARSON III</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$X_T$</td>
<td>$K_T$</td>
<td>$X_T$</td>
<td>$K_T$</td>
<td>$X_T$</td>
</tr>
<tr>
<td>0.9</td>
<td>1.1</td>
<td>214.997</td>
<td>-1.282</td>
<td>219.417</td>
<td>-1.202</td>
<td>225.065</td>
</tr>
<tr>
<td>0.5</td>
<td>2</td>
<td>286.200</td>
<td>0.000</td>
<td>281.092</td>
<td>-0.092</td>
<td>277.073</td>
</tr>
<tr>
<td>0.2</td>
<td>5</td>
<td>332.961</td>
<td>0.842</td>
<td>330.748</td>
<td>0.802</td>
<td>326.173</td>
</tr>
<tr>
<td>0.1</td>
<td>10</td>
<td>357.403</td>
<td>1.282</td>
<td>360.103</td>
<td>1.330</td>
<td>358.682</td>
</tr>
<tr>
<td>0.05</td>
<td>20</td>
<td>377.588</td>
<td>1.645</td>
<td>386.299</td>
<td>1.802</td>
<td>389.865</td>
</tr>
<tr>
<td>0.02</td>
<td>50</td>
<td>400.306</td>
<td>2.054</td>
<td>418.069</td>
<td>2.373</td>
<td>430.228</td>
</tr>
<tr>
<td>0.01</td>
<td>100</td>
<td>415.452</td>
<td>2.326</td>
<td>440.687</td>
<td>2.781</td>
<td>460.475</td>
</tr>
<tr>
<td>0.001</td>
<td>1,000</td>
<td>457.894</td>
<td>3.090</td>
<td>510.803</td>
<td>4.043</td>
<td>560.420</td>
</tr>
</tbody>
</table>

Ket: 1. $X_T = \mu + K_T \cdot \sigma$
2. Menurut Uji Chi-Kuadrat, yang terbaik menggunakan distribusi NORMAL
3. Sedangkan menurut Uji Smirnov-Kolmogorov, yang terbaik menggunakan distribusi NORMAL

**Example of frequency analysis performed by using spreadsheet application/computer program**
Nowadays

There is no (or at least difficult to find) computer application on frequency analysis having data plot capability

- Commercial application program may exist (?)
- Practising engineers and students might be not able to afford for commercial application program

Need for frequency analysis application program, which is a free-ware and capable to produce graphical plot of the data on probability paper
AProb_4E → versi terkini AProb_4.1

- A computer application on frequency analysis that has data plotting capability
- Can be freely downloaded from website http://istiarto.staff.ugm.ac.id/
Analisis Frekuensi

Debit Maksimum Sungai Titaraya di Tanggi 1945-2010

jumlah data = 66
rata-rata = 659
simpangan baku = 212
kurtosis = 2.77
excess kurtosis = -0.23
skewness = -0.17
Distribusi Gumbel
rentang keyakinan (1-α) = 0.90
Hidrologi

1. Analisis Frekuensi
   - Frequency Analysis on Hydrologic Data (AProb_4E), in English
   - Analisis Data Hidrologi Ekstrem (AProb_3, AProb_31)
   - Analisis Data Hidrologi Ekstrem (AProb_2, AProb_21)
   - Analisis Frekuensi Data Hidrologi Ekstrem

2. Memplotkan Data pada Kertas Probabilitas
   - Plot Data pada Kertas Probabilitas (PProb_4)
Plotting position of data values on probability paper

\[ \text{prob}(Y < y_m) = \frac{m}{n+1} \]

- \( m \) is the rank of the data being sorted in ascending order
- \( n \) is the number of data

This is known as Weibull formula of plotting time series data
Confidence interval \((1 - \alpha)\)

- An interval within which the true value (which is unknown) can reasonably be expected to lie.
- The size of the interval depends on the confidence level \((1 - \alpha)\)

The estimate of event value for a particular return period, \(y_T\)

- An upper and lower limits are specified by adjustment of \(K_T\)

\[
U_{T,\alpha} = \bar{Y} + s_y \cdot K^U_{T,\alpha}
\]

\[
L_{T,\alpha} = \bar{Y} + s_y \cdot K^L_{T,\alpha}
\]

upper and lower confidence limit factors for normally distributed data; these are determined using the noncentral \(t\) distribution
Goodness of fit test

- Smirnov-Kolmogorov test
- Chi-square test

Both tests are applied with confidence level of \((1 - \alpha) = 0.90\)
Goodness of Fit Test

- Smirnov-Kolmogorov Test

\[ \Delta_{\text{max}} = \max \left| \text{prob}(Y < y) - \text{prob}(Y < y) \right| \]

\[ \rightarrow \text{rejected if} \quad \Delta_{\text{max}} < D_c \]

- \( D_c \) is critical value according to the Smirnov-Kolmogorov table
- according to the observed data
- according to the distribution being tested
Goodness of Fit Test

- Chi-square Test

\[ \chi^2_c = \sum_{i=1}^{k} \left( \frac{O_i - E_i}{E_i} \right) \]

\[ \Rightarrow \text{rejected if } \chi^2_c > \chi^2_{1-\alpha/2,k-p-1} \]

- \( p \) is the number of parameters estimated from the data
- \( O_i \) the observed relative frequency in the \( i \)th class interval
- \( E_i \) the expected relative frequency, according to the distribution being tested, in the \( i \)th class interval
- \( k \) the number of class intervals
Statistics of data
--> number of data : 66
--> minimum : 115
--> maximum : 1120
--> mean : 659.409091
--> standard dev : 212.143274
--> kurtosis : 2.772395
--> excess kurtosis : -0.227605
--> skewness : -0.167274

Statistics of log data
--> number of data : 66
--> minimum : 2.060698
--> maximum : 3.049218
--> mean : 2.790866
--> standard dev : 0.172613
--> kurtosis : 7.239843
--> excess kurtosis : 4.239843
--> skewness : -1.587294

Goodness of fit test, $\alpha = 0.10$ (confidence level $1-\alpha$) = 0.90

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Gumbel</th>
<th>Log Normal</th>
<th>Log Pearson III</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smirnov-Kolmogorov</td>
<td>OK</td>
<td>OK</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>Discrepancy (max)</td>
<td>0.118</td>
<td>0.114</td>
<td>0.047</td>
<td>0.049</td>
</tr>
<tr>
<td>Chi-square</td>
<td>OK</td>
<td>fail</td>
<td>fail</td>
<td>OK</td>
</tr>
<tr>
<td>Chi-square (max)</td>
<td>6.545</td>
<td>13.758</td>
<td>13.758</td>
<td>3.152</td>
</tr>
</tbody>
</table>

Estimates on values according to various return periods [years]

<table>
<thead>
<tr>
<th>Return period</th>
<th>Gumbel</th>
<th>Log Normal</th>
<th>Log Pearson III</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>625</td>
<td>618</td>
<td>581</td>
<td>659</td>
</tr>
<tr>
<td>5</td>
<td>812</td>
<td>863</td>
<td>854</td>
<td>838</td>
</tr>
<tr>
<td>10</td>
<td>936</td>
<td>1030</td>
<td>919</td>
<td>931</td>
</tr>
<tr>
<td>20</td>
<td>1060</td>
<td>1190</td>
<td>959</td>
<td>1010</td>
</tr>
<tr>
<td>50</td>
<td>1210</td>
<td>1400</td>
<td>990</td>
<td>1100</td>
</tr>
<tr>
<td>100</td>
<td>1320</td>
<td>1560</td>
<td>1000</td>
<td>1150</td>
</tr>
<tr>
<td>200</td>
<td>1440</td>
<td>1720</td>
<td>1010</td>
<td>1210</td>
</tr>
<tr>
<td>500</td>
<td>1590</td>
<td>1940</td>
<td>1020</td>
<td>1270</td>
</tr>
</tbody>
</table>
Gumbel

1-prob [%]:

Jumlah data = 66
Rata-rata = 659
Simpangan baku = 212
Kurtosis = 2.77
Excess kurtosis = -0.23
Skewness = -0.17

Distribusi Gumbel
Rentang keyakinan (1-\(\alpha\)) = 0.90
Debit Maksimum Sungai Tirtaraya di Tanggi 1945-2010
jumlah data = 66
rata-rata log-data = 2.7909
simpangan baku log-data = 0.1726
kurtosis log-data = 7.24
excess kurtosis log-data = 4.24
skewness log-data = -1.59
Distribusi Log Normal
rentang keyakinan (1-\(\alpha\)) = 0.90
Debit Maksimum Sungai Tirtaraya di Tanggi 1945-2010
jumlah data = 66
rata-rata log-data = 2.7909
simpangan baku log-data = 0.1726
kurtosis log-data = 7.24
excess kurtosis log-data = 4.24
skewness log-data = 0.15
Distribusi Log Pearson III
rentang keyakinan (1-\(\alpha\)) = 0.90
Debit Maksimum Sungai Tirtaraya di Tanggi 1945-2010
jumlah data = 66
rata-rata = 659
simpangan baku = 212
kurtosis = 2.77
excess kurtosis = -0.23
skewness = -0.17
Distribusi Normal
rentang keyakinan (1-α) = 0.90

Debit [m³/s]
Probabilitas [%]
Improvement

- Feature that will be added: test for outliers

\[ y_H = \bar{Y} + K_n \cdot s_Y \]
\[ y_L = \bar{Y} - K_n \cdot s_Y \]

in log unit

high and low outlier thresholds

if logarithms of the data are greater than \( y_H \)
or less than \( y_L \) \( \Rightarrow \) they are considered as outliers
References

Thank You