Data Processing Techniques

Normal Distribution
Binomial Distribution

• Remember the selection of a program out of four proposed programs (A,B,C,D) to receive funding
  • selection is carried out every year in a 5-year period
  • each proposal is given an equal chance to be selected
  • estimating the probabilities of A to get funding (the number of successes, i.e. the number of times that A is selected to receive funding

http://istiarto.staff.ugm.ac.id
Binomial Distribution

• In every selection
  • $\text{prob(As)} = \text{probability of A being selected} = \frac{1}{4} = 0.25 = p$
  • $\text{prob(Ag)} = \text{probability of A being not selected} = 1 - 0.25 = 0.75 = q$

• Within five selections (five years)
  • probability of A being selected $x$ times is

\[
f_x(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}
\]

\[
f_x(x; 5, 0.25) = \binom{5}{x} 0.25^x (1-0.25)^{5-x}
\]
## Binomial Distribution

Number of successes (number of times A gets fundings) in five year period

<table>
<thead>
<tr>
<th>Number of successes</th>
<th>Number of events</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1</td>
<td>0.237</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>0.396</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>0.264</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>0.088</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>0.015</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Σ 1.000
Probability distribution of Program A be funded

Number of repetitions = 5

Probability distribution of Program A be funded

Number of times A be funded

Probability

0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45
0 0.237 0.264 0.396 0.088 0.015 0.001

Number of times A be funded

0 1 2 3 4 5
Binomial Distribution

• If the selection is carried out in more number of years (more repetitions)
  • 10 years
  • 20 years
  • $n$ years
    • there are $n+1$ possible outcomes
    • the number of times that A may get funding is either $n$ times, $n-1$ times, ...
      , 1 time, 0 time
Probability distribution of Program A receives fund

Number of repetitions = 10
Probability distribution of Program A receives fund

Number of repetitions = 20

Number of times A receives fund

Probability

0 0.05 0.1 0.15 0.2 0.25
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Normal Curve

• If the selection is repeated \( n \) times and \( n \) is sufficiently large
  • the histogram of the probability distribution has small intervals
  • line connecting the bars’ crest forms a curve having particular characteristics \( \rightarrow \) Normal Curve
Probability distribution of Program A receives fund

Number of repetitions = 20
Normal Curve

- Normal curve
  - has a bell-shape form and particular characteristics,
  - not every bell-shape curve is a normal curve.
- Normal curve represents a probability distribution, known as normal distribution.
- Therefore, binomial distribution problems can be solved by normal distribution approach.
- Normal distribution is easily computed
  - normal distribution table,
  - function in statistical softwares.
Normal Distribution

• Characteristics of a normal distribution
  • continuous distribution,
  • symmetrical about its mean value,
  • scores tend to group close to its mean value,
  • scores extend to infinity but only small number of them lie beyond 3 times standard deviation from the mean.
Normal Distribution

μ-4σ  μ-3σ  μ-2σ  μ-1σ  μ  μ+1σ  μ+2σ  μ+3σ  μ+4
Normal Distribution

Area = 0.9973

Area = 0.00135

Area = 0.00135

http://istiarto.staff.ugm.ac.id
Normal Distribution

The probability density function (pdf) of a normal distribution is given by:

\[ p_X(x) = \left(\frac{2\pi\sigma^2}{2}\right)^{\frac{1}{2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \]

for \( N(\mu,\sigma^2) \).
\(\mu_1 = \mu_2 = \mu_3\)
\(\sigma_1^2 < \sigma_2^2 < \sigma_3^2\)
\[ \mu_1 < \mu_2 < \mu_3 \]
\[ \sigma_1^2 = \sigma_2^2 = \sigma_3^2 \]
Normal Distribution

- If $X$ is normally distributed, $N(\mu, \sigma^2)$, the probability of $X$ being less than $x$ ($X < x$) can be found by:

$$\text{prob}(X < x) = P_X(x) = \int_{-\infty}^{x} p_X(t) \, dt = \int_{-\infty}^{x} \left(2\pi\sigma^2\right)^{-1/2} e^{-\frac{1}{2}(t-\mu)/\sigma^2} \, dt$$

area under pdf curve $\rightarrow$ cumulative distribution function, cdf
pdf-cdf

$p_X(x)$

$P_X(x)$

$\mu$

pdf

cdf

$-\infty$

$+\infty$
Normal Distribution

- Area under the pdf curve
  - is the probability of event
  - is the percentile rank
  - \( \text{prob}(X < x) = \text{prob}(-\infty < X < x) \)
    - = area under the pdf curve from \(-\infty\) to \(x\)
  - \( \text{prob}(-\infty < X < +\infty) = 1 \)
  - \( \text{prob}(X > x) = \text{prob}(x < X < +\infty) \)
    - = area under the pdf curve from \(x\) to \(+\infty\)
    - = \(1 - \text{prob}(X < x)\)
Normal Distribution

- Probabilities
  - \( \text{prob}(X = x) = \text{area under the pdf curve from } x \text{ to } x \)  
    = 0
  - \( \text{prob}(X \leq x) = \text{prob}(X < x) \)
  - \( \text{prob}(X \geq x) = \text{prob}(X > x) \)
  - \( \text{prob}(xa \leq X \leq xb) = \text{prob}(xa < X < xb) \)
  - \( \text{prob}(X < \mu) = 0.5 \)
  - \( \text{prob}(X > \mu) = 0.5 \)
  - \( \text{prob}(\mu-x < X < \mu) = \text{prob}(\mu < X < \mu+x) \)
  - \( \text{prob}(\mu < X < \mu-x) = \text{prob}(X > \mu+x) \)
Standard Normal Distribution

• Normal distribution is normally expressed in standard normal distribution.

  • $z$ score
    \[ z_x = \frac{x - \mu}{\sigma} \]

    $Z$ is normally distributed with $\mu = 0$ and $\sigma^2 = 1$, $N(0,1)$ → standard normal distribution

  • pdf and cdf
    \[
    p_z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < +\infty
    \]
    \[
    \text{prob}(Z < z) = P_z(z) = \int_{-\infty}^{z} p_z(z) \, dt = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \, dt
    \]
Standard Normal Distribution

\[
Z = \frac{X - \mu}{\sigma}
\]

Where:
- \(X\) is the value from the distribution
- \(\mu\) is the mean
- \(\sigma\) is the standard deviation

Z scores:
- \(Z = -3\) corresponds to \(X = \mu - 3\sigma\)
- \(Z = 1\) corresponds to \(X = \mu + 1\sigma\)

Graph showing the distribution with key points:
- \(\mu - 4\sigma\) to \(\mu + 4\sigma\)
Standard Normal Distribution

- Tables for pdf and cdf of standard normal distribution are available.
- We can make ourselves these tables by using statistical softwares or spreadsheet softwares.
  - See also functions available in MS Excel.
    - =NORM.DIST(...)
    - =NORM.S.DIST(...)
    - =NORM.INV(...)
    - =NORM.S.INV(...)
- Try to construct these tables using MS Excel.
NORMAL DISTRIBUTION

Observed pdf vs Theoretical Normal Curve
## Annual Peak Discharge of XYZ River

<table>
<thead>
<tr>
<th>Year</th>
<th>Discharge (m³/s)</th>
<th>Year</th>
<th>Discharge (m³/s)</th>
<th>Year</th>
<th>Discharge (m³/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>473</td>
<td>23</td>
<td>1110</td>
<td>45</td>
<td>843</td>
</tr>
<tr>
<td>2</td>
<td>544</td>
<td>24</td>
<td>717</td>
<td>46</td>
<td>450</td>
</tr>
<tr>
<td>3</td>
<td>872</td>
<td>25</td>
<td>961</td>
<td>47</td>
<td>284</td>
</tr>
<tr>
<td>4</td>
<td>657</td>
<td>26</td>
<td>925</td>
<td>48</td>
<td>460</td>
</tr>
<tr>
<td>5</td>
<td>915</td>
<td>27</td>
<td>341</td>
<td>49</td>
<td>804</td>
</tr>
<tr>
<td>6</td>
<td>535</td>
<td>28</td>
<td>690</td>
<td>50</td>
<td>550</td>
</tr>
<tr>
<td>7</td>
<td>678</td>
<td>29</td>
<td>734</td>
<td>51</td>
<td>279</td>
</tr>
<tr>
<td>8</td>
<td>700</td>
<td>30</td>
<td>991</td>
<td>52</td>
<td>712</td>
</tr>
<tr>
<td>9</td>
<td>669</td>
<td>31</td>
<td>792</td>
<td>53</td>
<td>468</td>
</tr>
<tr>
<td>10</td>
<td>347</td>
<td>32</td>
<td>626</td>
<td>54</td>
<td>841</td>
</tr>
<tr>
<td>11</td>
<td>580</td>
<td>33</td>
<td>937</td>
<td>55</td>
<td>613</td>
</tr>
<tr>
<td>12</td>
<td>470</td>
<td>34</td>
<td>687</td>
<td>56</td>
<td>871</td>
</tr>
<tr>
<td>13</td>
<td>663</td>
<td>35</td>
<td>801</td>
<td>57</td>
<td>705</td>
</tr>
<tr>
<td>14</td>
<td>809</td>
<td>36</td>
<td>323</td>
<td>58</td>
<td>777</td>
</tr>
<tr>
<td>15</td>
<td>800</td>
<td>37</td>
<td>431</td>
<td>59</td>
<td>442</td>
</tr>
<tr>
<td>16</td>
<td>523</td>
<td>38</td>
<td>770</td>
<td>60</td>
<td>206</td>
</tr>
<tr>
<td>17</td>
<td>580</td>
<td>39</td>
<td>536</td>
<td>61</td>
<td>850</td>
</tr>
<tr>
<td>18</td>
<td>672</td>
<td>40</td>
<td>708</td>
<td>62</td>
<td>829</td>
</tr>
<tr>
<td>19</td>
<td>115</td>
<td>41</td>
<td>894</td>
<td>63</td>
<td>887</td>
</tr>
<tr>
<td>20</td>
<td>461</td>
<td>42</td>
<td>626</td>
<td>64</td>
<td>602</td>
</tr>
<tr>
<td>21</td>
<td>524</td>
<td>43</td>
<td>1120</td>
<td>65</td>
<td>403</td>
</tr>
<tr>
<td>22</td>
<td>943</td>
<td>44</td>
<td>440</td>
<td>66</td>
<td>505</td>
</tr>
</tbody>
</table>
Annual Peak Discharge of XYZ River
## Annual Peak Discharge of XYZ River

<table>
<thead>
<tr>
<th>Discharge (m$^3$/s)</th>
<th>Frequency</th>
<th>Relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 - 200</td>
<td>150</td>
<td>1</td>
</tr>
<tr>
<td>200 - 300</td>
<td>250</td>
<td>2</td>
</tr>
<tr>
<td>300 - 400</td>
<td>350</td>
<td>3</td>
</tr>
<tr>
<td>400 - 500</td>
<td>450</td>
<td>10</td>
</tr>
<tr>
<td>500 - 600</td>
<td>550</td>
<td>9</td>
</tr>
<tr>
<td>600 - 700</td>
<td>650</td>
<td>12</td>
</tr>
<tr>
<td>700 - 800</td>
<td>750</td>
<td>10</td>
</tr>
<tr>
<td>800 - 900</td>
<td>850</td>
<td>11</td>
</tr>
<tr>
<td>900 - 1000</td>
<td>950</td>
<td>6</td>
</tr>
<tr>
<td>1000 - 1100</td>
<td>1050</td>
<td>0</td>
</tr>
<tr>
<td>1100 - 1200</td>
<td>1150</td>
<td>2</td>
</tr>
</tbody>
</table>

| Total                | 66        | 1.00               |
Annual Peak Discharge of XYZ River
Annual Peak Discharge of XYZ River

Relative frequency

Discharge (m³/s)

theoretical distribution
observed data

https://istiarto.staff.ugm.ac.id
## Annual Peak Discharge of XYZ River

<table>
<thead>
<tr>
<th>Discharge (m³/s)</th>
<th>$F_Q(q_{upper})$</th>
<th>$F_Q(q_{lower})$</th>
<th>Relative freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 – 200</td>
<td>0.0142</td>
<td>0.0038</td>
<td>0.0104</td>
</tr>
<tr>
<td>200 – 300</td>
<td>0.0432</td>
<td>0.0142</td>
<td>0.0290</td>
</tr>
<tr>
<td>300 – 400</td>
<td>0.1078</td>
<td>0.0432</td>
<td>0.0646</td>
</tr>
<tr>
<td>400 – 500</td>
<td>0.2231</td>
<td>0.1078</td>
<td>0.1152</td>
</tr>
<tr>
<td>500 – 600</td>
<td>0.3875</td>
<td>0.2231</td>
<td>0.1645</td>
</tr>
<tr>
<td>600 – 700</td>
<td>0.5755</td>
<td>0.3875</td>
<td>0.1880</td>
</tr>
<tr>
<td>700 – 800</td>
<td>0.7475</td>
<td>0.5755</td>
<td>0.1720</td>
</tr>
<tr>
<td>800 – 900</td>
<td>0.8735</td>
<td>0.7475</td>
<td>0.1259</td>
</tr>
<tr>
<td>900 – 1000</td>
<td>0.9473</td>
<td>0.8735</td>
<td>0.0738</td>
</tr>
<tr>
<td>1000 – 1100</td>
<td>0.9819</td>
<td>0.9473</td>
<td>0.0346</td>
</tr>
<tr>
<td>1100 – 1200</td>
<td>0.9949</td>
<td>0.9819</td>
<td>0.0130</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.9911</td>
</tr>
</tbody>
</table>
Observed vs Theoretical Curves

• Expected relative frequency (according to normal distribution)
  • for the 2nd class

\[ f_Q(q = 250) = \int_{200}^{300} \left( \frac{2\pi s_Q^2}{2} \right)^{-\frac{1}{2}} e^{-\frac{1}{2}(q-q)^2/s_Q^2} dq \]

\[ = \int_{200}^{300} \left( 2\pi \cdot 210^2 \right)^{-\frac{1}{2}} e^{-\frac{1}{2}(q-2100)^2/210^2} dq \]

\[ = F_Q(300) - F_Q(200) \]

\[ = F_Z \left( \frac{300 - 660}{210} \right) - F_Z \left( \frac{200 - 660}{210} \right) \]

\[ = F_Z(-1.7143) - F_Z(-2.1905) \]

\[ = 0.0432 - 0.0142 = 0.0290 \]

\[ Q = 660 \text{ m}^3/\text{s} \]
\[ s_Q = 210 \text{ m}^3/\text{s} \]
Observed vs Theoretical Curves

- Another method to find expected relative frequency
  - for the 2nd class

\[
\begin{align*}
  f_Q(q_i) &= \Delta q_i \cdot p_Q(q_i) \\
  p_Q(q_i) &= p_z(z_i) \cdot \left| \frac{dz}{dq} \right| = \frac{p_z(z_i)}{s_Q} \\
  f_Q(q_i) &= \Delta q_i \frac{p_z(z_i)}{s_Q}
\end{align*}
\]
Observed vs Theoretical Curves

• Another method to find expected relative frequency
  • for the 2nd class

\[
i = 2: \\
\Delta q_i = 100 \text{ m}^3/\text{s} \\
q_i = 250 \text{ m}^3/\text{s} \rightarrow z_i = \frac{250 - 660}{210} = -1.95 \\
p_z(z_i) = 0.0593 \\
f_Q(q_i) = 100 \left( \frac{0.0593}{210} \right) = 0.028
\]
# Annual Peak Discharge of XYZ River

<table>
<thead>
<tr>
<th>Discharge (m³/s)</th>
<th>$p_q(q)$</th>
<th>Relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 − 200</td>
<td>9.95E-05</td>
<td>0.0100</td>
</tr>
<tr>
<td>200 − 300</td>
<td>2.82E-04</td>
<td>0.0282</td>
</tr>
<tr>
<td>300 − 400</td>
<td>6.39E-04</td>
<td>0.0639</td>
</tr>
<tr>
<td>400 − 500</td>
<td>1.15E-03</td>
<td>0.1152</td>
</tr>
<tr>
<td>500 − 600</td>
<td>1.66E-03</td>
<td>0.1656</td>
</tr>
<tr>
<td>600 − 700</td>
<td>1.90E-03</td>
<td>0.1898</td>
</tr>
<tr>
<td>700 − 800</td>
<td>1.73E-03</td>
<td>0.1733</td>
</tr>
<tr>
<td>800 − 900</td>
<td>1.26E-03</td>
<td>0.1262</td>
</tr>
<tr>
<td>900 − 1000</td>
<td>7.32E-04</td>
<td>0.0732</td>
</tr>
<tr>
<td>1000 − 1100</td>
<td>3.39E-04</td>
<td>0.0339</td>
</tr>
<tr>
<td>1100 − 1200</td>
<td>1.25E-04</td>
<td>0.0125</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.9917</td>
</tr>
</tbody>
</table>
Annual Peak Discharge of XYZ River

• When we use 100 m$^3$/s class interval, there is one class interval that has zero frequency → increase the class interval to 150 m$^3$/s.
## Annual Peak Discharge of XYZ River

<table>
<thead>
<tr>
<th>Discharge (m$^3$/s)</th>
<th>Observed data</th>
<th>Normal Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – 150</td>
<td>1</td>
<td>0.02</td>
</tr>
<tr>
<td>150 – 300</td>
<td>2</td>
<td>0.03</td>
</tr>
<tr>
<td>300 – 450</td>
<td>8</td>
<td>0.12</td>
</tr>
<tr>
<td>450 – 600</td>
<td>14</td>
<td>0.21</td>
</tr>
<tr>
<td>600 – 750</td>
<td>18</td>
<td>0.27</td>
</tr>
<tr>
<td>750 – 900</td>
<td>15</td>
<td>0.23</td>
</tr>
<tr>
<td>900 – 1050</td>
<td>6</td>
<td>0.09</td>
</tr>
<tr>
<td>1050 – 1200</td>
<td>2</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>66</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Annual Peak Discharge of XYZ River

Relative frequency

Discharge (m$^3$/s)

theoretical distribution

observed data

http://istiarto.staff.ugm.ac.id

38
Terima kasih