Basics of data mining and data-driven decision making in biology

With a demo using XLSTAT

Jean-Paul Maalouf
Webinar plan

• Statistics in biology: exploratory or decisional
• A closer look at exploratory statistics (data mining)
  > A basic example: the scatter plot
  > Exploring a dataset with PCA
  > XLSTAT: PCA demo
• Moving to decision making
  > Choosing a hypothesis or two
  > Setting up an experimental design
  > Taking a decision from the results we get
  > XLSTAT: t-test demo
• Conclusion
• Survey
Two big worlds of statistical tools

**Exploratory (data mining)**

- No prior hypotheses
- Fishing for hypotheses in a big dataset
- Examples: scatter plots, correlation, PCA, clustering…
Two big worlds of statistical tools

**Exploratory (data mining)**
- No prior hypotheses
- Fishing for hypotheses in a big dataset
- Examples: scatter plots, correlation, PCA, clustering…

**Decisional**
- We have a hypothesis and we want to test it
- Experimental designs
- Examples: t-test, ANOVA, regression
Two big worlds of statistical tools

**Exploratory (data mining)**
- No prior hypotheses
- Fishing for hypotheses in a big dataset
- Examples: scatter plots, correlation, PCA, clustering…

**Decisional**
- We have a hypothesis and we want to test it
- Experimental designs
- Examples: t-test, ANOVA, regression

*In practice, we often combine both approaches*
A closer look at exploratory statistics (data mining)

Exploratory (data mining) – an example dataset

<table>
<thead>
<tr>
<th>Plant</th>
<th>Variety</th>
<th>Leaf area</th>
<th>Leaf thickness</th>
<th>Fruit size</th>
<th>Soil moisture</th>
<th>Soil fertility</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>Variety1</td>
<td>11,4</td>
<td>2,457</td>
<td>11,33</td>
<td>0,456</td>
<td>1,9</td>
</tr>
<tr>
<td>P2</td>
<td>Variety2</td>
<td>6,8</td>
<td>3,367</td>
<td>13,05</td>
<td>0,417</td>
<td>1,53</td>
</tr>
<tr>
<td>P3</td>
<td>Variety2</td>
<td>10,2</td>
<td>2,536</td>
<td>1,93</td>
<td>0,381</td>
<td>1,62</td>
</tr>
<tr>
<td>P4</td>
<td>Variety1</td>
<td>8,9</td>
<td>3,003</td>
<td>17,73</td>
<td>0,541</td>
<td>1,73</td>
</tr>
<tr>
<td>P5</td>
<td>Variety2</td>
<td>6,5</td>
<td>3,62</td>
<td>18,05</td>
<td>0,281</td>
<td>1,19</td>
</tr>
<tr>
<td>P6</td>
<td>Variety2</td>
<td>8,2</td>
<td>2,607</td>
<td>1,11</td>
<td>0,638</td>
<td>1,99</td>
</tr>
<tr>
<td>P7</td>
<td>Variety3</td>
<td>5,1</td>
<td>3,136</td>
<td>8,5</td>
<td>0,33</td>
<td>1,18</td>
</tr>
<tr>
<td>P8</td>
<td>Variety2</td>
<td>6,7</td>
<td>2,903</td>
<td>2,76</td>
<td>0,497</td>
<td>1,64</td>
</tr>
<tr>
<td>P9</td>
<td>Variety1</td>
<td>10,2</td>
<td>2,851</td>
<td>16,21</td>
<td>0,558</td>
<td>1,62</td>
</tr>
<tr>
<td>P10</td>
<td>Variety2</td>
<td>11,4</td>
<td>3,234</td>
<td>13,38</td>
<td>0,468</td>
<td>1,96</td>
</tr>
<tr>
<td>P11</td>
<td>Variety1</td>
<td>10,4</td>
<td>3,201</td>
<td>6,7</td>
<td>0,57</td>
<td>1,61</td>
</tr>
<tr>
<td>P12</td>
<td>Variety3</td>
<td>8,1</td>
<td>3,379</td>
<td>12,26</td>
<td>0,413</td>
<td>1,74</td>
</tr>
<tr>
<td>P13</td>
<td>Variety2</td>
<td>9,6</td>
<td>2,71</td>
<td>12,77</td>
<td>0,477</td>
<td>1,32</td>
</tr>
</tbody>
</table>
A closer look at exploratory statistics (data mining)

Exploratory (data mining) – scatter plot

[Scatter plot showing the relationship between soil moisture and leaf area for different varieties.]
A closer look at exploratory statistics (data mining)

Data mining – how to summarize a big part of what the dataset has to offer in one single plot?

<table>
<thead>
<tr>
<th>Plant</th>
<th>Variety</th>
<th>Leaf area</th>
<th>Leaf thickness</th>
<th>Fruit size</th>
<th>Soil moisture</th>
<th>Soil fertility</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>Variety1</td>
<td>11,4</td>
<td>2,457</td>
<td>11,33</td>
<td>0.456</td>
<td>1.9</td>
</tr>
<tr>
<td>P2</td>
<td>Variety2</td>
<td>6,8</td>
<td>3,367</td>
<td>13,05</td>
<td>0.417</td>
<td>1.53</td>
</tr>
<tr>
<td>P3</td>
<td>Variety2</td>
<td>10,2</td>
<td>2,536</td>
<td>1,93</td>
<td>0.381</td>
<td>1.62</td>
</tr>
<tr>
<td>P4</td>
<td>Variety1</td>
<td>8,9</td>
<td>3,003</td>
<td>17,73</td>
<td>0.541</td>
<td>1.73</td>
</tr>
<tr>
<td>P5</td>
<td>Variety2</td>
<td>6,5</td>
<td>3,62</td>
<td>18,05</td>
<td>0.281</td>
<td>1.19</td>
</tr>
<tr>
<td>P6</td>
<td>Variety2</td>
<td>8,2</td>
<td>2,607</td>
<td>1,11</td>
<td>0.638</td>
<td>1.99</td>
</tr>
<tr>
<td>P7</td>
<td>Variety3</td>
<td>5,1</td>
<td>3,136</td>
<td>8,5</td>
<td>0,33</td>
<td>1,18</td>
</tr>
<tr>
<td>P8</td>
<td>Variety2</td>
<td>6,7</td>
<td>2,903</td>
<td>2,76</td>
<td>0,497</td>
<td>1,64</td>
</tr>
<tr>
<td>P9</td>
<td>Variety1</td>
<td>10,2</td>
<td>2,851</td>
<td>16,21</td>
<td>0,558</td>
<td>1,62</td>
</tr>
<tr>
<td>P10</td>
<td>Variety2</td>
<td>11,4</td>
<td>3,234</td>
<td>13,38</td>
<td>0,468</td>
<td>1,96</td>
</tr>
<tr>
<td>P11</td>
<td>Variety1</td>
<td>10,4</td>
<td>3,201</td>
<td>6,7</td>
<td>0,57</td>
<td>1,61</td>
</tr>
<tr>
<td>P12</td>
<td>Variety3</td>
<td>8,1</td>
<td>3,379</td>
<td>12,26</td>
<td>0,413</td>
<td>1,74</td>
</tr>
<tr>
<td>P13</td>
<td>Variety2</td>
<td>8,6</td>
<td>2,71</td>
<td>4,27</td>
<td>0,127</td>
<td>1,25</td>
</tr>
</tbody>
</table>
A closer look at exploratory statistics (data mining)

Data mining – how to summarize a big part of what the dataset has to offer in one single plot?

Principal Component Analysis

- Concentrates plenty of information on a 2-dimensional plot (or a bit more)
- Allows to detect variables that look like one another
- Allows to detect individuals (plants) that look like one another
Data mining – how to summarize a big part of what the dataset has to offer in one single plot?

Principal Component Analysis

- Concentrates plenty of information on a 2-dimensional plot (or a bit more)
- Allows to detect variables that look like one another
- Allows to detect individuals (plants) that look like one another

*Fishing for hypotheses made super-easy. Let’s check it out in XLSTAT*
PCA: XLSTAT demo
Here are a couple of hypotheses we proposed thanks to the PCA

- Soil fertility induces an increase of leaf area
- Soil moisture induces an increase of leaf area
- Soil fertility induces a decrease of leaf thickness
- Leaf area induces a decrease of soil moisture
- Variety 3 is more adapted to humid and fertile environments
- …
Moving to decision making

Here are a couple of hypotheses we proposed thanks to the PCA

• Soil fertility induces an increase of leaf area
• Soil moisture induces an increase of leaf area
• Soil fertility induces a decrease of leaf thickness
• Leaf area induces a decrease of soil moisture
• Variety 3 is more adapted to humid and fertile environments
• ...

Biology and statistics
Feb 19, 2015
www.xlstat.com
Moving to decision making

Here are a couple of hypotheses we proposed thanks to the PCA:

- Soil fertility induces an increase of leaf area
- **Soil moisture induces an increase of leaf area**
- Soil fertility induces a decrease of leaf thickness
- Leaf area induces a decrease of soil moisture
- Variety 3 is more adapted to humid and fertile environments
- …

*Let’s figure out a study to decide if this hypothesis is valid or not*
Moving to decision making: experimental design

Soil moisture induces an increase of leaf area

30 generously watered pots
30 poorly watered pots
Soil moisture induces an increase of leaf area

30 generously watered pots

30 poorly watered pots

Leaf area is measured when the plants become mature
Taking a decision out of the results we get

Classically, we formulate two hypotheses

Null hypothesis: H0

\[ \text{area (high watering)} = \text{area (low watering)} \]
Taking a decision out of the results we get

Classically, we formulate two hypotheses

Null hypothesis: H0

\[ \text{area (high watering)} = \text{area (low watering)} \]

Alternative hypothesis: Ha

\[ \text{area (high watering)} \neq \text{area (low watering)} \]
Taking a decision out of the results we get

Classically, we formulate two hypotheses

\( H_0 \): area (high watering) = area (low watering)

\( H_a \): area (high watering) ≠ area (low watering)
Taking a decision out of the results we get

Classically, we formulate two hypotheses

\[ H_0: \text{area (high watering)} = \text{area (low watering)} \]

\[ H_a: \text{area (high watering)} \neq \text{area (low watering)} \]

We define a risk threshold \textbf{alpha} ( = 0.05, commonly)
Taking a decision out of the results we get

Classically, we formulate two hypotheses

\[ H_0: \text{area (high watering)} = \text{area (low watering)} \]

\[ H_a: \text{area (high watering)} \neq \text{area (low watering)} \]

We define a risk threshold \( \alpha \) (\( = 0.05 \), commonly)

We run a statistical test (t-test for example) and obtain a \( p \)-value
Taking a decision out of the results we get

Classically, we formulate two hypotheses

$H_0$: area (high watering) = area (low watering)

$H_a$: area (high watering) \neq area (low watering)

We define a risk threshold $\alpha$ ( = 0.05, commonly)

We run a statistical test (t-test for example) and obtain a p-value

- If p-value < 0.05: We reject $H_0$ and accept $H_a$
Classically, we formulate two hypotheses

\( H_0 \): area (high watering) = area (low watering)

\( H_a \): area (high watering) \( \neq \) area (low watering)

We define a risk threshold \textbf{alpha} ( = 0.05, commonly)

We run a statistical test (t-test for example) and obtain a \textbf{p-value}

- If \textbf{p-value} < 0.05: We reject \( H_0 \) and accept \( H_a \), with a risk proportional to p-value of being wrong.
Taking a decision out of the results we get

Classically, we formulate two hypotheses

\( H_0 \): area (high watering) = area (low watering)

\( H_a \): area (high watering) \( \neq \) area (low watering)

We define a risk threshold \( \alpha \) ( = 0.05, commonly)

We run a statistical test (t-test for example) and obtain a \( p\)-value

- If \( p\)-value < 0.05: We reject \( H_0 \) and accept \( H_a \), with a risk proportional to \( p\)-value of being wrong. The lower the \( p\)-value, the lower the risk of being wrong when stating that there is a difference between the two water treatments.
Taking a decision out of the results we get

Classically, we formulate two hypotheses

\[ H_0: \text{area (high watering)} = \text{area (low watering)} \]
\[ H_a: \text{area (high watering)} \neq \text{area (low watering)} \]

We define a risk threshold \textbf{alpha} ( = 0.05, commonly)

We run a statistical test (t-test for example) and obtain a \textbf{p-value}

- If \textbf{p-value} < 0.05: We reject H0 and accept Ha, with a risk proportional to p-value of being wrong.
- If \textbf{p-value} > 0.05: We do not reject H0 (but that does not mean we should necessarily accept it)
Comparing 2 averages (t-test): XLSTAT demo
Statistics in biology – a conclusion

- Data mining tools (scatter plots, PCA) allow to look for hypothesis
- Decisional tools (t-test) allow to test hypotheses
- We showed how both approaches may be complementary
- Many other tools are out there waiting for you to be discovered!
Thank you for your attention!
Thank you for your attention!

Survey time…