JMP GENOMICS VALIDATION USING COVERING ARRAYS AND EQUIVALENCE PARTITIONING

WENJUN BAO AND JOSEPH MORGAN
JMP DIVISION, SAS INC.
THE SOFTWARE TESTING DILEMMA

Defect density\(^1\) for released software:

- **Industry**: \(\approx 1 - 25\) defects/1000 lines of code.
- **Microsoft Applications**: \(\approx 0.5\) defects/1000 lines of code.
- **Space Shuttle**: \(\approx 0\) defects/1000 lines of code.

\(^1\)S. McConnell, "Code Complete: A Practical Handbook of Software Construction," 2004
# THE SOFTWARE TESTING DILEMMA

Downtime hourly cost$1:  

- Brokerage operations $6,450,000  
- Credit card authorization $2,600,000  
- Ebay $225,000  
- Amazon $180,000  
- Home shopping channel $113,000  
- Airline reservation center $89,000  
- Cellular service activation $41,000  

Debugging, testing, and verification activities in typical commercial software development organizations range from 50% to 75% of total software development costs.

A recent National Institute of Standards and Technology (NIST) report\(^1\) investigating the economic impact of inadequate software testing methods and tools found:

“...the national annual costs of an inadequate infrastructure for software testing is estimated to range from $22.2 billion to $59.5 billion”

THE SOFTWARE TESTING DILEMMA

- Testing is necessary.
- Testing is important.
- Testing is expensive.
- Poor testing is even more expensive!

“... TOO LITTLE TESTING MAY BE INITIALLY CHEAPER BUT NOT IN THE LONG RUN.”

A statistical discovery software tool for biological data including different big data (“Omics”).

A JMP/SAS Integration product.

More than two hundred analytical procedures (AP), with a rich set of options for each procedure.
ANOVA LAUNCH DIALOG
ANOVA LAUNCH DIALOG VALIDATION

- Twenty one distinct GUI controls.
- Each control provides between two and fifteen choices.

1. Distribution 15 choices
2. Filter zero/missing data 2 choices
3. Separate results 2 choices
4. LSMeans diff. set 4 choices
5. Multiple testing method 14 choices
6. Alpha 3 choices
7. Multiple testing adj. 2 choices
8. Component fixed-eff. tests 2 choices
9. Plot std. residuals 3 choices
10. Filtration method 9 choices
11. LSMeans methods 15 choices
12. Cluster LSMean profiles 2 choices
13. Unbounded variance 2 choices
14. Kenward-Roger df method 2 choices
15. U diffs. 2 choices
16. Include p-values 2 choices
17. Include adj. p-values 2 choices
18. Scale Volcano y-axis 2 choices
19. Journal results by chrom. 2 choices
20. Primary annotation 2 choices
21. Secondary annotation 2 choices
ANOVA LAUNCH DIALOG VALIDATION

- Input space:
  \[15 \times 2^2 \times 4 \times 14 \times 3 \times 2^2 \times 3 \times 9 \times 15 \times 2^{10} = 16,721,510,400\] inputs

- Exhaustive testing:
  - \(\approx 17\) billion test cases.
  - Assuming 5 seconds would be needed to execute and verify each case, then the ANOVA AP would require:

  **2,651 YEARS TO COMPLETE!**
THE DESIGN OF EXPERIMENTS INSPIRATION

One way of addressing the dilemma is to recognize that it is somewhat similar to a statistical design of experiments (DOE) problem. There have been several key milestones in the application of DOE ideas to software testing.

- **Orthogonal Latin squares:** In 1985, R. Mandl (Commun. ACM, v28.10) proposed using orthogonal Latin squares for testing compilers.

- **Orthogonal arrays:** In 1992, R. Brownlie (AT&T Tech. Journal, v71.3) proposed orthogonal arrays to generate test suites for application testing.

- **Covering arrays:** In 1998, C. Mallows (Technometrics, v40.3) proposed covering arrays to generate test suites for application testing.
Definition 1

A covering array $\text{CA}(N; t, k, v)$ is an $N \times k$ array such that the $i$-th column contains $v$ distinct symbols. If a $\text{CA}(N; t, k, v)$ has the property that for any $t$ coordinate projection, all $v^t$ combinations of symbols exist at least once, then it is a $t$-covering array (or strength $t$ covering array). A $t$-covering array is optimal if $N$ is minimal for fixed $t$, $k$, and $v$. 
**COVERING ARRAYS**

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CA(6; 2, 5, 2)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

**Are these optimal?**

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CA(12; 3, 5, 2)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
MIXED LEVEL COVERING ARRAYS

Definition 2

A mixed level covering array $\text{MCA}(N; t, (v_1 \cdot v_2 \cdot \ldots \cdot v_k))$ is an $N \times k$ array such that the $i$-th column contains $v_i$ distinct symbols. If for any $t$ coordinate projection, all $\prod v_i$ combinations of symbols exist, then it is a $t$-covering array and is optimal if $N$ is minimal for fixed $t$, $k$, and $(v_1 \cdot v_2 \cdot \ldots \cdot v_k)$.

An optimal $\text{MCA}(9; 2, (3^2 \cdot 2^3))$. 

\[
\begin{array}{cccccc}
1 & 1 & 2 & 2 & 2 \\
1 & 2 & 2 & 2 & 1 \\
1 & 3 & 1 & 1 & 1 \\
2 & 1 & 1 & 2 & 1 \\
2 & 2 & 1 & 1 & 2 \\
2 & 3 & 2 & 1 & 1 \\
3 & 1 & 1 & 1 & 2 \\
3 & 2 & 2 & 1 & 1 \\
3 & 3 & 2 & 2 & 2 \\
\end{array}
\]
ANOVA - Mixed Level Covering Array:

\[ \text{MCA}( N; 2, (15^2 \cdot 14 \cdot 9 \cdot 4 \cdot 3^2 \cdot 2^{14}) ) \]

What is \( N \)?

\[ N = 225 \]

If we need 5 seconds for each case, then the ANOVA AP would require:

18.75 MINUTES TO COMPLETE!
## COVERING ARRAYS AND GENOMICS TESTING

### Design

<table>
<thead>
<tr>
<th>Run</th>
<th>Distribution</th>
<th>FilterData</th>
<th>SeparateResults</th>
<th>LSMeansDiffSet</th>
<th>MultipleTestingMethod</th>
<th>Alpha</th>
<th>MultipleTestAdj</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Lognormal</td>
<td>yes</td>
<td>no</td>
<td>Simple Differences</td>
<td>Hochberg</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>2</td>
<td>Generalized Poisson</td>
<td>yes</td>
<td>Differences with a Control</td>
<td></td>
<td>Hommel</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>3</td>
<td>Poisson</td>
<td>no</td>
<td>yes</td>
<td>Simple Differences</td>
<td>None</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>4</td>
<td>t</td>
<td>no</td>
<td>yes</td>
<td>Simple Differences</td>
<td>Bonferroni</td>
<td>0.05</td>
<td>yes</td>
</tr>
<tr>
<td>5</td>
<td>Gamma</td>
<td>no</td>
<td>yes</td>
<td>Differences with a Control</td>
<td>DependentFDR</td>
<td>0.05</td>
<td>yes</td>
</tr>
<tr>
<td>6</td>
<td>Exponential</td>
<td>no</td>
<td>yes</td>
<td>All Pairwise Differences</td>
<td>Holm</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>7</td>
<td>Negative Binomial</td>
<td>no</td>
<td>Differences with a Control</td>
<td></td>
<td>pFDR</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>8</td>
<td>Multinomial</td>
<td>no</td>
<td>yes</td>
<td>All Pairwise Differences</td>
<td>Sidak</td>
<td>0.05</td>
<td>yes</td>
</tr>
<tr>
<td>9</td>
<td>t</td>
<td>yes</td>
<td>no</td>
<td>Simple Differences</td>
<td>FDR</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>10</td>
<td>Beta</td>
<td>no</td>
<td>no</td>
<td>All Pairwise Differences</td>
<td>AdaptiveHochberg</td>
<td>0.05</td>
<td>yes</td>
</tr>
</tbody>
</table>

### Metrics

<table>
<thead>
<tr>
<th>t</th>
<th>Coverage</th>
<th>Diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>100.00</td>
<td>7.90</td>
</tr>
<tr>
<td>3</td>
<td>95.16</td>
<td>21.06</td>
</tr>
<tr>
<td>4</td>
<td>80.72</td>
<td>41.01</td>
</tr>
<tr>
<td>5</td>
<td>58.32</td>
<td>62.35</td>
</tr>
<tr>
<td>6</td>
<td>35.59</td>
<td>79.68</td>
</tr>
</tbody>
</table>

Optimize Maximum iterations: 250
Can we do better?

**YES!**

With Equivalence Partitioning, our Mixed Level Covering Array becomes:

\[ \text{MCA}(72; 2, (9 \cdot 8 \cdot 6 \cdot 4 \cdot 3^{2} \cdot 2^{14}) ) \]

If we need 5 seconds for each case, then the ANOVA AP would require:

**6 MINUTES TO COMPLETE!**
## Design

<table>
<thead>
<tr>
<th>Run</th>
<th>DistributionGroup</th>
<th>FilterData</th>
<th>SeparateResults</th>
<th>LSMMeansDiffSet</th>
<th>MultipleTestingMethodGroup</th>
<th>Alpha</th>
<th>MultipleTestAdj</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>discrete</td>
<td>no</td>
<td>yes</td>
<td>Simple Differences</td>
<td>None</td>
<td>0.05</td>
<td>no</td>
</tr>
<tr>
<td>2</td>
<td>discrete</td>
<td>no</td>
<td>yes</td>
<td>Simple Differences</td>
<td>None</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>3</td>
<td>continuous</td>
<td>no</td>
<td>yes</td>
<td>Simple Differences</td>
<td>FDR</td>
<td>0.05</td>
<td>yes</td>
</tr>
<tr>
<td>4</td>
<td>continuous</td>
<td>yes</td>
<td>yes</td>
<td>Differences with a Control</td>
<td>Bonferroni</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>5</td>
<td>discrete</td>
<td>no</td>
<td>no</td>
<td>All Pairwise Differences</td>
<td>AdaptiveFDR</td>
<td>0.05</td>
<td>no</td>
</tr>
<tr>
<td>6</td>
<td>discrete</td>
<td>no</td>
<td>yes</td>
<td>All Pairwise Differences</td>
<td>pFDR</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>7</td>
<td>continuous</td>
<td>yes</td>
<td>yes</td>
<td>Simple Differences</td>
<td>AdaptiveHolm</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>8</td>
<td>discrete</td>
<td>yes</td>
<td>yes</td>
<td>Differences with a Control</td>
<td>None</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>9</td>
<td>discrete</td>
<td>yes</td>
<td>yes</td>
<td>None</td>
<td>pFDR</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>10</td>
<td>discrete</td>
<td>yes</td>
<td>yes</td>
<td>Simple Differences</td>
<td>AdaptiveHolm</td>
<td>0.05</td>
<td>yes</td>
</tr>
</tbody>
</table>

### Metrics

<table>
<thead>
<tr>
<th></th>
<th>t-Coverage</th>
<th>t-Diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>100.00</td>
<td>12.22</td>
</tr>
<tr>
<td>3</td>
<td>94.44</td>
<td>28.76</td>
</tr>
<tr>
<td>4</td>
<td>76.78</td>
<td>51.24</td>
</tr>
<tr>
<td>5</td>
<td>50.74</td>
<td>72.22</td>
</tr>
<tr>
<td>6</td>
<td>27.46</td>
<td>86.43</td>
</tr>
</tbody>
</table>

Optimize Maximum iterations: 250
ANOVA VALIDATION

Exhaustive Testing

Covering Array Testing

Covering Array & Equivalence Partitioning Testing

17 Billion cases
2651 years (5/sec)

225 cases
19 minutes (5/sec)

72 cases
6 minutes (5/sec)
JMP GENOMICS VALIDATION

Validation Procedure

- Generate test cases
- Validate test cases
- Process test cases

Covering Array

Automated

Compare actual & expected results
CONCLUSION

- Test cases derived from covering arrays can exercise all \( t \)-way combinations of input parameter values at very little cost.

- Compared to exhaustive testing, the cost savings is dramatic with potentially very little loss in quality (i.e. undiscovered faults).

- Increasingly, software development organizations are embracing this technology.

  *Note*: The JMP development team started using covering arrays in 2006.

- Empirical research suggests that most software failures are due to faults resulting from the interaction of few parameters (typically \( \leq 6 \)).
REFERENCES

Thank You!!