Effective Applications
Development, Maintenance
and Support Benchmarking

John Ogilvie
CEO ISBSG
7 March 2017
What is the ISBSG?

ISBSG (International Software Benchmarking Standards Group) is a not-for-profit organisation.

The ISBSG was founded in 1997 by a group of national software metrics associations. Their aim was to promote the use of IT industry data to improve software processes and products.

The ISBSG mission is to help YOU and your organisation improve the planning and management of your IT software projects. To achieve this:

- We have 2 repositories of IT software development / maintenance data. This data originates from trusted, international IT organisations. Our data can be used as a benchmark for your IT project.
- You will find valuable information on a wide variety of topics, in our many reports and books.

The ISBSG mission is supported by our partners, who represent IT and Metrics organisations and associations from around the world.

- Explore ISBSG Offerings at www.isbsg.org
Introduction

In this Workshop we will explore:

- Why organisations choose to Benchmark
- Potential benefits of Benchmarking
- Pitfalls of failing to appropriately plan
- Need to introduce rigor into the benchmarking process
Introduction

Å Benchmarking can provide significant insights into performance of IT Processes.
  ï Can identify where can make improvements most effectively

Å To be successful, the following must be fully determined prior to starting:
  ï Objective of Benchmark
  ï Benefits expected
  ï Criteria for selecting applications and processes
Why Benchmark?

Many IT organisations routinely collect cost, effort, defect and sometimes size data.

- However rarely go the extra step of turning raw data into information that would facilitate change.

Benchmarking is the activity that turns data into information by:

- Measuring current practices
- Comparing current performance to past performance or peer performance.
- Interpreting the results and recommending actions
Why Benchmark?

- Organisations usually start with *internal* benchmarking to identify areas of improvement.
- Next step is to compare with *external* business units or wider *industry* performance:
  - Assess performance against competitors and invest in new tools, techniques or technologies.
- Sometimes an industry comparison highlights the cost of poor IT processes motivates management to rethink AD/M strategy and improve internal measurement.
Example Measuring Competitiveness

- Senior management of a software company wondered how competitive they were when it comes to **productivity**.
- Many bids for projects were **lost** and they wished to improve, especially their Microsoft.Net department.
- Analysis of the bids by department showed the following:

<table>
<thead>
<tr>
<th></th>
<th>PDR (h/FP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>3,2</td>
</tr>
<tr>
<td>Percentile 10%</td>
<td>3,8</td>
</tr>
<tr>
<td><strong>Percentile 25%</strong></td>
<td>5,9</td>
</tr>
<tr>
<td>Median</td>
<td>7,6</td>
</tr>
<tr>
<td><strong>Percentile 75%</strong></td>
<td>12,9</td>
</tr>
<tr>
<td>Percentile 90%</td>
<td>18,9</td>
</tr>
<tr>
<td>Max.</td>
<td>34,2</td>
</tr>
<tr>
<td>N</td>
<td>35</td>
</tr>
</tbody>
</table>
Result

- This analysis data indicated that the bids were well outside best industry performance — between the 75% and 90% percentiles.

- This caused a review of the bid phase which showed a number of issues:
  - Estimates were extremely pessimistic due to severe penalties in case of overruns;
  - In a number of stages, risk surcharges were added;
  - They wished to work in fixed team of 6 fte, but ISBSG data shows that the project size was usually too small for this teams size to be efficient;

- As a result the bid process was redesigned, making the company more successful!
Benchmarking Risks

• In recent years benchmarking is being used more to assess and compare cost effectiveness of IT suppliers

• Most large fixed term outsourcing contracts financially reward or penalise the supplier based on performance against an industry benchmark or a client performance baseline.

  • Given the substantial financial implications for both client and supplier there is an absolute need for rigorous benchmarking activity

  • A case study will be presented later in the workshop
Establishing “Terms of Reference”

It is recommended that prior to commencing a benchmarking program a “Terms of Reference” be established with an agreed position on:

- Strategic Intent of the Benchmark
- Type of Benchmark
- Benchmark Performance Metrics
- Standards for Measures
- Scope of Benchmark
- Frequency of Benchmark
- Benchmark Peers
- Benchmarking Report
- Dispute Resolution Process
Strategic Intent of Benchmark

The objectives of the benchmarking activity:

- What are results required to demonstrate
  - Within what period and for what purpose
- Criteria for judging success of benchmark

Common reasons for benchmarking include monitoring:

- Process improvement initiatives
- Outsourcing contract performance against targets
- Consistency across organisational units
- Benefits from new initiatives against those claimed
- Performance against peers or whole industry
Type of Benchmark

Establish whether the organisation will benchmark:

- Internally to demonstrate improvement trends over time
- Externally to compare results with independent external organisational units, or industry as a whole.

Organisations aware of own limitations will recognise the need to improve before conducting an external comparison

- It is recommended to start by internally benchmarking and establish measurement processes, then do some external, industry comparisons.
Benchmark Performance Metrics

IT Organisations typically focus initial benchmarking activities on areas that impact IT costs.

Measure cost effectiveness and quality of IT Processes and Products by optimising the following Key Result Areas (KRA)s:

- **Cost-effectiveness of the process**: delivering “value” for money invested.
- **Efficiency of the process**: how “productively” is software being developed or maintained.
- **Speed of delivery**: how quickly can product be delivered or problem solved.
Benchmark Performance Metrics

Å Key Result Areas (Continued)

- Quality of the product - how "good" is the product delivered
- Quality of the process - how much time and money was wasted in rework.
- Customer Satisfaction - how well does the product delivered meet or exceed customer expectations

Å Benchmarking is not a "one size fits all" activity

- Need to decide which of the KRA's above are most critical to the success of the organisation
Standards for Measures

When comparing between projects or organisations need to ensure measurement units are equivalent.

Most organisations chose to measure cost effectiveness by measuring effort input rather than cost input.

- Comparing cost this year against previous years, especially globally, is impacted by currency exchange fluctuations, inflation rates, salary increases.

- Project Productivity Rate (function points delivered per person day) ensures “apples to apples” comparisons.
Standards for Measures

To ensure valid comparisons the following must be verified as consistent:

- Type of Function Points recorded
- Type of Day Recorded
- Accuracy of the Measures
- Scope of the Effort Measures
- Scope of the Size Measures
- Scope of the Project Life Cycle Activities

To simplify the process and facilitate external industry benchmarking use of the ISO Standard is recommended. (ISO/IEC 29155-1)
Scope of Benchmark

In order to decide if into the benchmarking set or if they need to be grouped and compared separately, the following characteristics need to be considered:

- Different Delivery Options
- Different Types of Requirements
- Different Resourcing Profiles
- Different Technology Profiles
- Different Size Profiles
- Different Functional Domains (Refer ISO/IEC 14143)
- Diverse and/or Large User Base
- Different Project Classifications
Scope of Benchmark

Å It is recommended that prior to selecting projects or applications:
   ï Group into like projects
   ï Classify using categories listed previously

Å Review to ensure that each benchmarking set contains a mix of all types consistent with the organisation’s portfolio or

Å Grouped into like categories for comparison exclusively within those categories.
Frequency of Benchmark

Å Need to consider:

- Allow sufficient time to demonstrate benefits of implementing new tools or technologies
- Projects with long durations
- If benchmarking periods set too widely apart the usefulness of the result may have diminished

Å Outsourcing contracts typically require quarterly or annual benchmarking
Projects that display high productivity, (low hours per FP), may reflect a lower risk profile. For example:

- Experienced and stable team
- Experienced Project Manager
- Known technology
- Stable requirements
- Small project (manageable size)
- Good customer commitment
- Single site delivery
- Fewer document deliverables
ISBSG Analysis Report “What Project Attributes Influence Productivity”

Projects displaying low productivity, (ie high hours per FP), may reflect a high risk profile or, of course, a higher level of deliverables, testing and governance.

- Inexperienced and/or unstable team
- Inexperienced Project Manager
- New technology platform
- Low customer commitment
- Unstable requirements
- Multiple site delivery
- High level of deliverable documentation
- Higher level of testing due to system criticality
Statistical Validity of the Result

Å Beyond the Statistical Average

- In my experience, the major reason for failure of Benchmarking and Baselining programs is unwillingness to consider how to ensure the statistical validity of the results.

- The following Case Study was a situation in which a multi-million dollar, multi-country outsourcing contract almost resulted in litigation, millions of dollars in penalties and contract cancellation.
  - After the first year, client claimed millions of dollars in penalties
  - Resolution finally reached by redoing baseline
Case Study Scenario

- ABC company is outsourcing their application development and maintenance.
- They wish to establish a set of targets for annual improvements in productivity based on what they were achieving internally prior to outsourcing.
- The contract with the vendor specified shared risk/reward
  - bonus/penalty payments for over/under achievement against targets.
Case Study Scenario

Å ABC had not implemented a productivity measurement program but prior to the contract they retrospectively calculated Functional Size and gathered the effort for 180 projects across multiple countries and technologies.

Å For the contract they required 28 productivity targets across a size, technology and country matrix

Å If at least 5 data points were available, the average was used as the baseline
Case Study Scenario

At the end of each Quarter, average performance of completed projects was compared to the targets and bonuses or penalties calculated based on what cost would have been if target met.

- No minimum number of projects was specified to use in the calculation

The following charts use data extracted from the ISBSG Repository to demonstrate how a baseline or benchmark value should be established and used.
Data Used in This Case Study

For the purposes of this presentation, data was extracted from the ISBSG Development and Enhancement Repository.

- Data Quality Rating: A or B
- Development Type: Enhancement
- Count Method: IFPUG
- Application Group: Business Application
- Development Platform: Mainframe/Midrange/Multi

Analyses and tables where produced using Minitab Statistical Software
What Questions arise?

What are the characteristics of the data we have?
- Shape of distribution
- Handling of Outliers

Baseline:
- How much data is required
- Do performance segments make sense

Measurement:
- How do we determine how productivity has changed
- How much measurement data is required
## ISBSG Relative Size

Å Categorises the Functional Size into groups, similar to clothing sizes, as follows:

<table>
<thead>
<tr>
<th>Relative Size</th>
<th>Functional Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. XXS Extra-extra-small</td>
<td>=&gt; 0 and &lt;10</td>
</tr>
<tr>
<td>2. XS Extra-small</td>
<td>=&gt; 10 and &lt;30</td>
</tr>
<tr>
<td>3. S Small</td>
<td>=&gt; 30 and &lt;100</td>
</tr>
<tr>
<td>4. M1 Medium1</td>
<td>=&gt; 100 and &lt;300</td>
</tr>
<tr>
<td>5. M2 Medium2</td>
<td>=&gt; 300 and &lt;1000</td>
</tr>
<tr>
<td>6. L Large</td>
<td>=&gt; 1,000 and &lt; 3,000</td>
</tr>
<tr>
<td>7. XL Extra-large</td>
<td>=&gt; 3,000 and &lt; 9,000</td>
</tr>
<tr>
<td>8. XXL Extra-extra-large</td>
<td>=&gt; 9,000 and &lt; 18,000</td>
</tr>
<tr>
<td>9. XXXL Extra-extra-extra-large</td>
<td>=&gt; 18,000</td>
</tr>
</tbody>
</table>

[www.isbsg.org](http://www.isbsg.org)
## Examine the Data: Descriptive Statistics

<table>
<thead>
<tr>
<th>Size</th>
<th>Count</th>
<th>Mean</th>
<th>TrMean</th>
<th>StDev</th>
<th>Minimum</th>
<th>Median</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>XXS</td>
<td>43</td>
<td>27.01</td>
<td>19.84</td>
<td>45.27</td>
<td>1.40</td>
<td>7.90</td>
<td>236.30</td>
</tr>
<tr>
<td>XS</td>
<td>157</td>
<td>19.06</td>
<td>14.25</td>
<td>29.83</td>
<td>0.90</td>
<td>10.90</td>
<td>271.60</td>
</tr>
<tr>
<td>S</td>
<td>424</td>
<td>16.34</td>
<td>12.23</td>
<td>30.27</td>
<td>0.40</td>
<td>10.00</td>
<td>424.90</td>
</tr>
<tr>
<td>M1</td>
<td>470</td>
<td>13.19</td>
<td>11.47</td>
<td>12.67</td>
<td>0.90</td>
<td>9.60</td>
<td>97.90</td>
</tr>
<tr>
<td>M2</td>
<td>187</td>
<td>14.52</td>
<td>13.10</td>
<td>12.68</td>
<td>0.80</td>
<td>11.10</td>
<td>80.70</td>
</tr>
<tr>
<td>L</td>
<td>31</td>
<td>11.69</td>
<td>10.25</td>
<td>11.63</td>
<td>1.00</td>
<td>9.10</td>
<td>42.90</td>
</tr>
<tr>
<td>XL</td>
<td>4</td>
<td>1.50</td>
<td>*</td>
<td>1.34</td>
<td>0.10</td>
<td>1.30</td>
<td>3.30</td>
</tr>
<tr>
<td>XXL</td>
<td>2</td>
<td>0.35</td>
<td>*</td>
<td>0.21</td>
<td>0.20</td>
<td>0.35</td>
<td>0.50</td>
</tr>
</tbody>
</table>
Examine the Data

Dotplot of PDR (afp)

Each symbol represents up to 17 observations.
Examine the Data

Dotplot of PDR (afp)

Each symbol represents up to 16 observations.
Handling Outliers

An outlier is an unusually large or small observation. Outliers can have a disproportionate influence on statistical results, such as the mean, which can result in misleading interpretations.

A variety of techniques can be used:

- Trim the data by removing the top and bottom 5% - simple to do.
- Remove data more than 2 standard deviations from the mean (simple to do but assumes data has normal distribution).
- Statistical test that all values in the sample are from the same, normally distributed population.
- Graphically using a Boxplot.
Identifying Outliers with Boxplot

Boxplot of PDR (afp) for XXS
Boxplot Explanation

- "Box" shows values in from Quartile 1(Q1) to Quartile 3(Q3)
  - Inter Quartile Range (IQR) is from Q1 to Q3.
    - Value is Q3 – Q1
- Mean and Median are shown
- "Whiskers" go to 1.5*IQR above and below the box
- An outlier is taken to be any value beyond the Whiskers
- Applying this to each of the size groups and removing sizes 7&8 reduced the number of data points by 106 from 1,319 to 1,231
## Descriptive Statistics after Outliers Removed

<table>
<thead>
<tr>
<th>Relative Size</th>
<th>Count</th>
<th>Mean</th>
<th>TrMean</th>
<th>StDev</th>
<th>Minimum</th>
<th>Median</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>XXS</td>
<td>38</td>
<td>13.44</td>
<td>11.84</td>
<td>14.92</td>
<td>1.40</td>
<td>6.00</td>
<td>53.80</td>
</tr>
<tr>
<td>XS</td>
<td>147</td>
<td>13.08</td>
<td>12.31</td>
<td>10.12</td>
<td>0.90</td>
<td>10.30</td>
<td>44.80</td>
</tr>
<tr>
<td>S</td>
<td>388</td>
<td>10.62</td>
<td>10.13</td>
<td>7.24</td>
<td>0.40</td>
<td>8.60</td>
<td>32.70</td>
</tr>
<tr>
<td>M1</td>
<td>433</td>
<td>10.20</td>
<td>9.79</td>
<td>6.38</td>
<td>0.90</td>
<td>8.80</td>
<td>30.30</td>
</tr>
<tr>
<td>M2</td>
<td>178</td>
<td>12.50</td>
<td>11.89</td>
<td>8.78</td>
<td>0.80</td>
<td>10.15</td>
<td>38.90</td>
</tr>
<tr>
<td>L</td>
<td>28</td>
<td>8.64</td>
<td>8.20</td>
<td>7.04</td>
<td>1.00</td>
<td>6.30</td>
<td>27.70</td>
</tr>
</tbody>
</table>
Data Distribution after Outliers Removed

Dotplot of PDR (afp)

Each symbol represents up to 3 observations.
How much data is required for Baseline & Performance Measurement?

• The need in a baseline is to have sufficient data points (n) such that their average will closely estimate the population average.
• One approach, based on the Central Limit Theorem in statistical theory indicates:
  Â In general try to have n>30
  Â If data is highly skewed, ideally more data points
  Â If data is symmetric, less than 30 may suffice
• 5 data points was insufficient for establishing a baseline in the case study
In the Case Study, in addition to setting a baseline, ABC wanted to determine if target productivity was being met.

Statistically, the 95% Confidence Interval for the true average of our performance is expressed as:

\[ \text{CI} = S \pm 1.96 \frac{\sigma}{\sqrt{n}} \]

where:

- \( S \) = sample mean,
- \( \bar{u} \) = sample standard deviation,
- \( n \) = sample size
Required Sample Sizes

<table>
<thead>
<tr>
<th>Standard Deviation</th>
<th>Confidence Interval</th>
<th>Sample Size @ Confidence Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size M1=6.38</td>
<td>95%</td>
<td>90%</td>
</tr>
<tr>
<td>± 1.0</td>
<td>159</td>
<td>112</td>
</tr>
<tr>
<td>± 1.5</td>
<td>72</td>
<td>51</td>
</tr>
<tr>
<td>± 2.0</td>
<td>42</td>
<td>30</td>
</tr>
<tr>
<td>± 2.5</td>
<td>28</td>
<td>20</td>
</tr>
<tr>
<td>± 3.0</td>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td>± 3.5</td>
<td>16</td>
<td>11</td>
</tr>
<tr>
<td>± 4.0</td>
<td>13</td>
<td>9</td>
</tr>
</tbody>
</table>

Å For example, if we have 15 data points for size M1 projects, with average S, we can be confident at a level of 90% that the true average of M1 projects would be in the range of S±3

Å Therefore it is this range, not just the value of S which needs to be considered

If the productivity target is in the range then it has been achieved.
Baseline – Do segments make sense

In deciding what segmentation should be used in establishing the baseline and subsequent performance management, the question is whether there is sufficient evidence that performance is different in each segment.

Too much segmentation reduces the number of data points in each segment which impacts the Confidence Interval of the measurement, as described earlier.
How to Determine if Segments Differ

Interval Plot of PDR (afp)
95% CI for the Mean

Individual standard deviations were used to calculate the intervals.
How to Determine if Segments Differ

- The Interval Plot indicates that S & M1 could be combined
- The fact that XS and M2 are similar is unexpected
  - Possibly need to add further attributes to data selection criteria
- The CI for XXS and L is too large to be useful due to small numbers of data points (38 and 28 respectively)
Baseline Recommendations

- Beware of basing conclusions on small numbers of data points.
- Check data for outliers.
  - Try and determine reason for outlier and do not remove if likely to occur in your own data.
- Do not segment data unless you are confident there is a real difference between segments.
- Your own data is always best. Industry data is a valuable benchmark reference and can provide data until you build up your own repository.

www.isbsg.org
Sources of Data

Å Your own data is always best provided you are satisfied as to its quality and there are sufficient data points for a statistically valid calculation.

Å ISBSG maintains a repository of Development and Enhancement which can be purchased.
  ï Currently contains over 7,000 project records
  ï Is supplied as an Excel spreadsheet to enable user to perform any analysis they require

Å ISBSG has released an on-line Productivity Query Tool (PDQ) to allow direct access to data
  ï Available by subscription monthly/annual
# Productivity Data Query - PDQ

<table>
<thead>
<tr>
<th>Project attributes</th>
<th>Project Delivery Rate</th>
<th>Speed of delivery</th>
<th>Defect Density</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Matches</td>
<td>1st Quartile</td>
<td>Median</td>
</tr>
<tr>
<td>Primary programming language</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organisation Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Application Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum team size</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Web Development</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intended Market</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Count Approach</td>
<td>FISMA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Development Platform</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Perform Query
Cautions about External Data

When selecting a benchmarking set from external sources consider the following factors:

- Organisational type
  - eg. small, large, government, financial
- Different User Priorities
- Quality of dataset
  - Engage benchmarking provider with own dataset
  - Purchase ISBSG data, an "open repository"
Prior to commencing the benchmarking program, sponsors and key stakeholders should agree:

- Structure and content
- Level of granularity
- Presentation technology
- Confidentiality and audience
- Review process and acceptance criteria
- Feedback process
It is recommended that if benchmarking is incorporated into contractual performance requirements then a formal dispute resolution process be also included in contract.
Summary

Â Despite the preceding warnings, do not feel that comparative benchmarking is impossible to achieve

Â However we are constantly confronted by contracts with performance targets against a single number to be derived from a large heterogeneous data set

Â Only some of the variables will apply in your unique situation.

ï If they do not apply chose to accommodate or ignore them from an informed position

ï Fail to consider them at your own risk
Acknowledgements

This presentation includes material from the paper “Effective Applications Development and Maintenance and Support Benchmarking”

Pam Morris (B.Sc., Grad. Dip. Computing, Dip Ed, CFPS, CSMS (Level 3))
Pam was the CEO for Total Metrics (Australia) and a member of the ISBSG Board