ISBSG
IT Confidence 2014

2°International Conference on
IT Data collection, Analysis and Benchmarking
Tokyo (Japan) - October 22, 2014

Beyond The Statistical Average

The KISIS Principle "Keeping it Simple is Stupid"

John Ogilvie

CEO, ISBSG

Goals of the presentation

G1. Understand the characteristics of the data available to establish a productivity baseline

G2. Statistical considerations in establishing a productivity baseline

G3. Statistical considerations in measuring actual performance against a baseline

Case Study

- 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

- For each of New Development and Application Enhancement projects ABC required 28 (4 technologies X 7 FP size bands) performance baselines in Hours/Function Point
- Annual % improvement targets were specified
- ABC had data from 128 internal projects. If there were at least 5 in a particular segment, the baseline was set as the average.
- Otherwise an industry data average was used.

Case Study

- At the end of each quarter, the average actual performance in each segment was calculated and the bonus/penalty rules applied
  - No minimum number of data points for the calculation was specified
  - In many cases only 1 or 2 actual projects in each category

After 12 months there was considerable conflict between ABC and vendor with ABC threatening legal action and vendor threatening to exit contract .

- Both over and under achievements were challenged

What Questions arise in Case Study

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

Data Used in This Presentation

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

ISBSG Relative Sizes

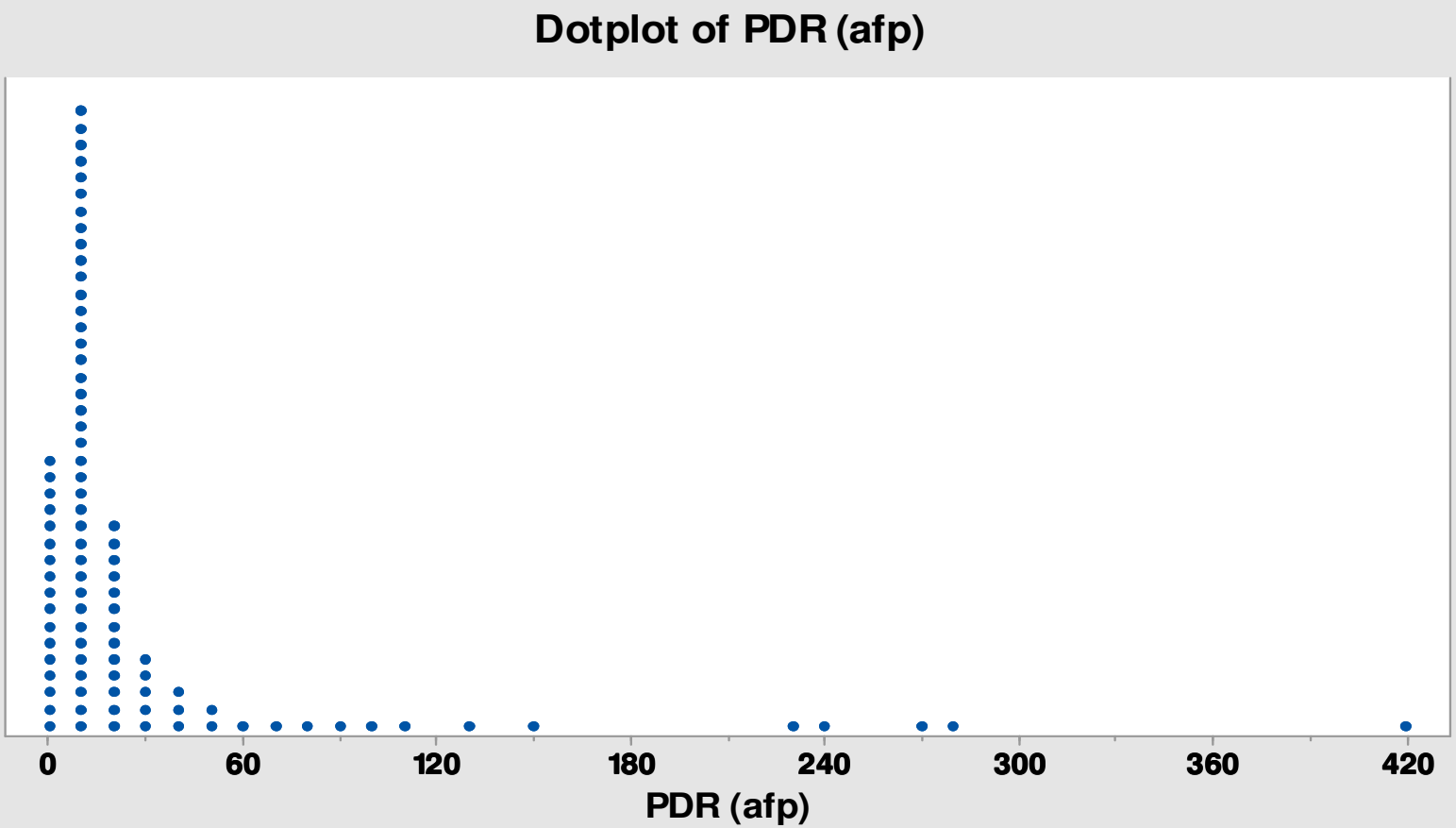
Categorises the Functional Size by relative sizes as follows:

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

Examine the Data: Descriptive Statistics

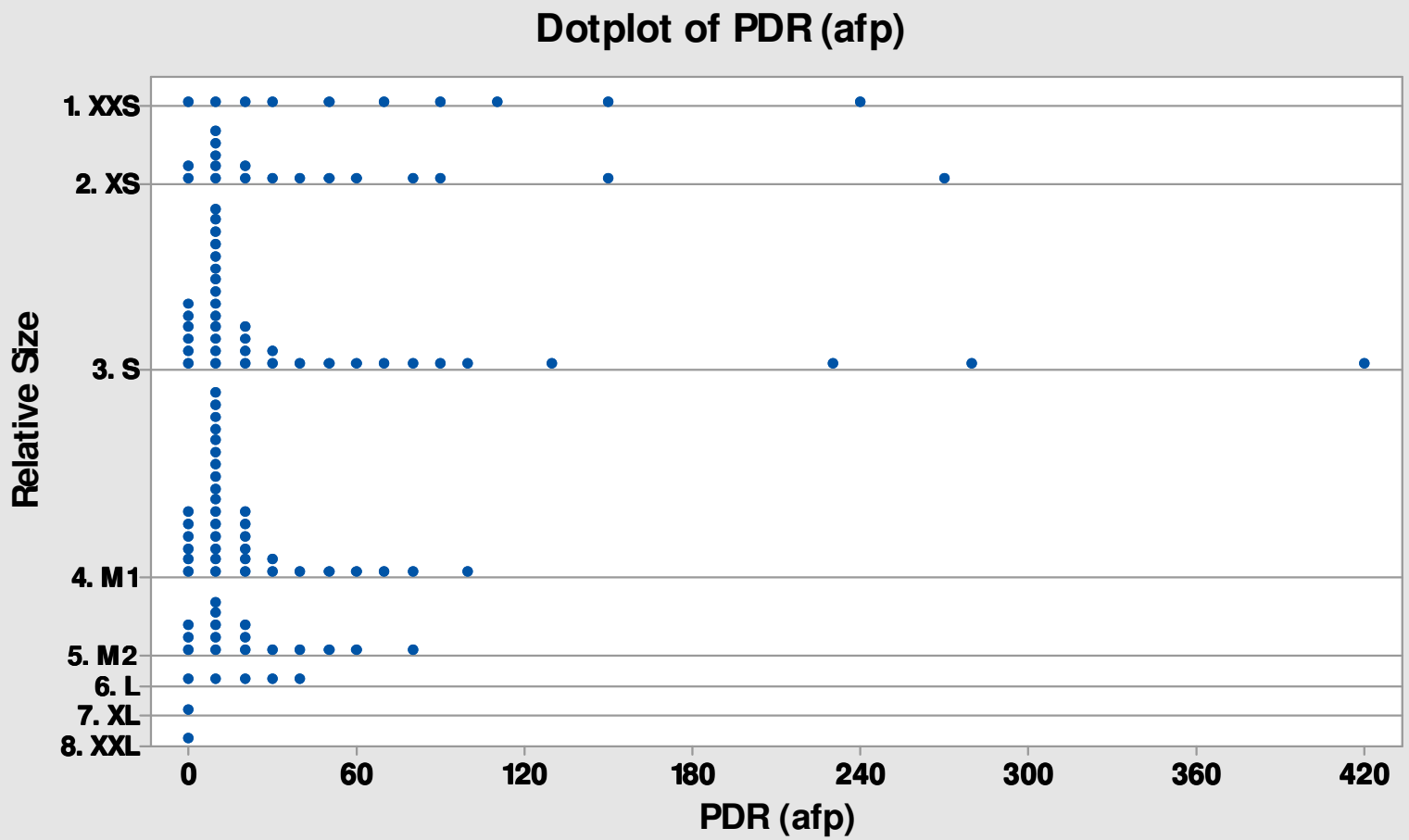
Relative Size	Total Count	Mean	TrMean	StDev	Minimum	Median	Maximum
1. XXS	43	27.01	19.84	45.27	1.40	7.90	236.30
2. XS	157	19.06	14.25	29.83	0.90	10.90	271.60
3. S	424	16.34	12.23	30.27	0.40	10.00	424.90
4. M1	470	13.19	11.47	12.67	0.90	9.60	97.90
5. M2	187	14.52	13.10	12.68	0.80	11.10	80.70
6. L	31	11.69	10.25	11.63	1.00	9.10	42.90
7. XL	4	1.50	*	1.34	0.10	1.30	3.30
8. XXL	2	0.35	*	0.21	0.20	0.35	0.50

Examine the Data



Each symbol represents up to 17 observations.

Examine the Data



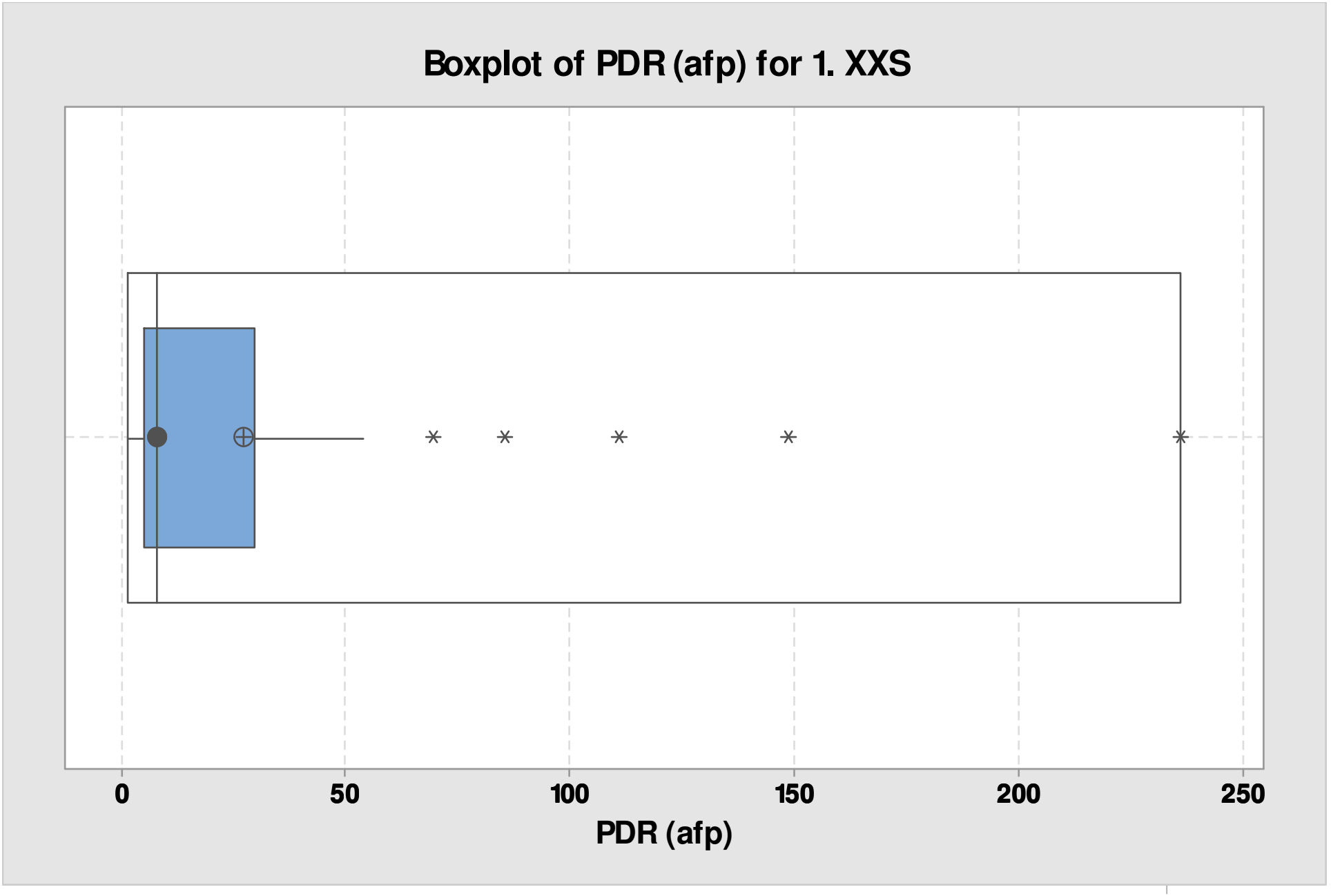
Each symbol represents up to 16 observations.

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. (Need a tool and assumes data has normal distribution)
- Graphically using a Boxplot


Boxplot of PDR (afp) for 1. XXS
0
50
100
150
200
250
PDR (afp)

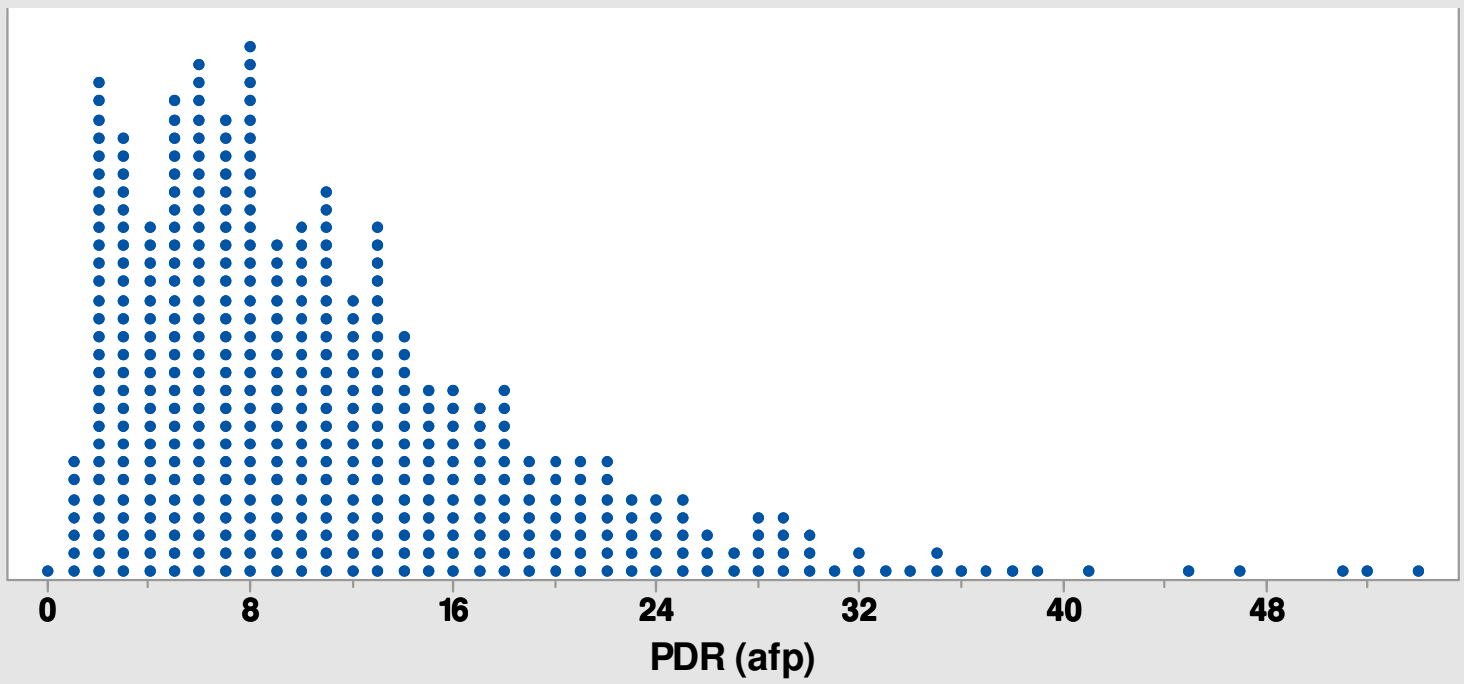
Boxplot

- "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

Relative Size	Total Count	Mean	TrMean	StDev	Minimum	Median	Maximum
1. XXS	38	13.44	11.84	14.92	1.40	6.00	53.80
2. XS	147	13.08	12.31	10.12	0.90	10.30	44.80
3. S	388	10.62	10.13	7.24	0.40	8.60	32.70
4. M1	433	10.20	9.79	6.38	0.90	8.80	30.30
5. M2	178	12.50	11.89	8.78	0.80	10.15	38.90
6. L	28	8.64	8.20	7.04	1.00	6.30	27.70

Data Distribution after Outliers Removed



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:

- In general try to have n>30
- If data is highly skewed, ideally more data points
- If data is symmetric , less than 30 may suffice

5 data points was insufficient for establishing a baseline in the case study

How much data required for Baseline & Performance Measurement

- 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:

  " We are 95% certain that the true mean is contained in the interval: $CI=S\pm1.96\sigma/\sqrt{n}$ "

  where:

  S=sample mean, σ=sample standard deviation, n=sample size

Required Sample Sizes

Standard Deviation	Confidence Interval	Sample Size @ Confidence Level	
Size M1=6.38		95%	90%
	± 1.0	159	112
	± 1.5	72	51
	± 2.0	42	30
	± 2.5	28	20
	± 3.0	20	15
	± 3.5	16	11
	± 4.0	13	9

For example, if we have 15 data points for size M1 projects, with average S, we can be confident at a level of 95% that the true average of M1 projects would be in the range of S±3

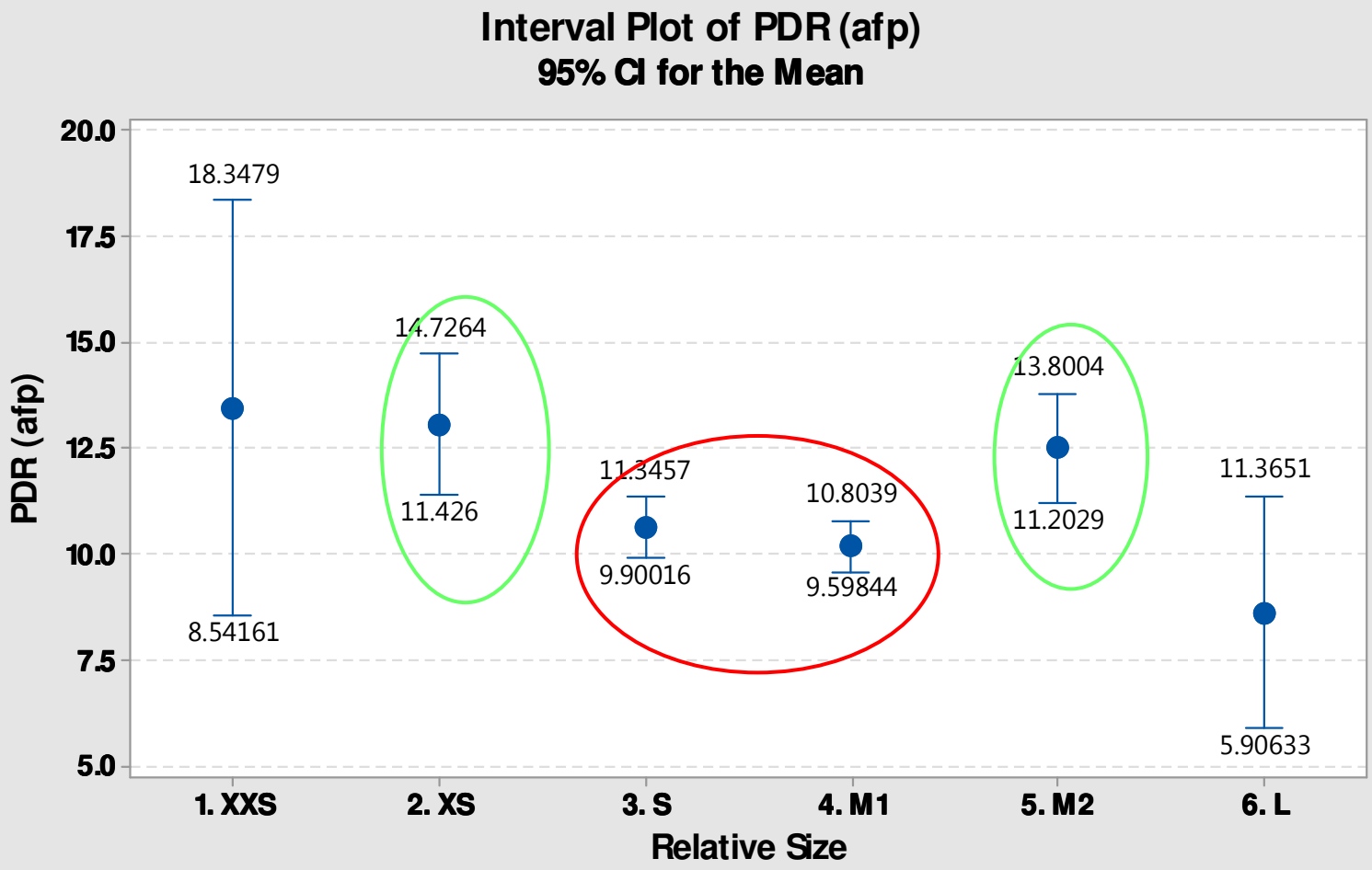
Therefore it is this range, not just the value of S which needs to be considered If the productivity target is in the range then it has been achieved.

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



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)

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.

Measure what you do and improve