



We have data! Now what?

November 06, 2008

Norocel Popa
BMO IT Best Practices and QA
Norocel.Popa@bmonb.com

Creating prediction models, using
historical data, for SDLC sub-processes.



Christmas vacation



-3700\$
-3500\$
-5000\$
-4300\$



Agenda



- Peer Review estimation model
 - using box-plot charts to understand the distribution of the data
- Defect prediction model
 - using a Monte Carlo simulation to deal with a range of data
- Defect fix Rate model
 - using a math model to assess process performance
- Change Request estimation model
 - find correlations between two variables
- Tying them all together – Project effort estimation model
- Q&A

Is this a useful model?



- How much time do you need to review the HLRD document and to fix all the defects you will find?



Peer Review Estimation Model



What have we done in PCG ?

- collected data for peer reviews
- analyzed stats for different types of docs
- created a prediction model based on box-plot charts
- revisit the model every year based on newly acquired data

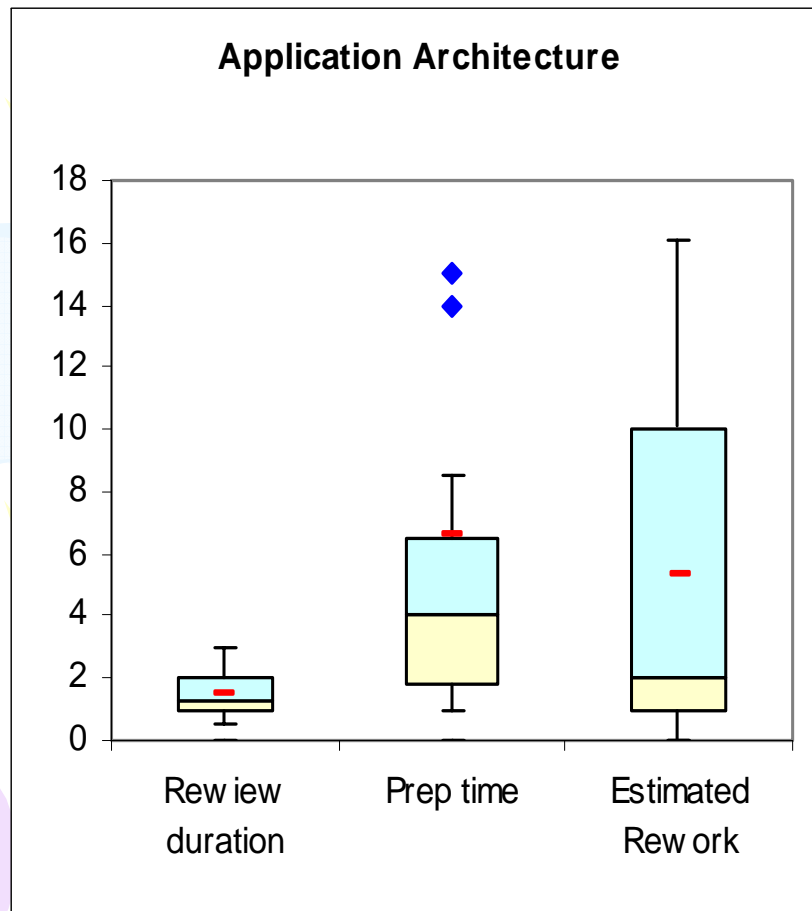
Why box-plots and not average?



- 100 docs reviewed: 99 of them in 1 hour, and 1 of them in 100 hours.
- Using average review time we will get close to 2 hours for every review.
- That would be wrong in 99% of the cases.



Box-plot chart



- Identifies outliers
- Shows the spread of data
- Shows median
- Shows average
- Show where each of the 4 quartiles are

PCG Model for Formal Reviews



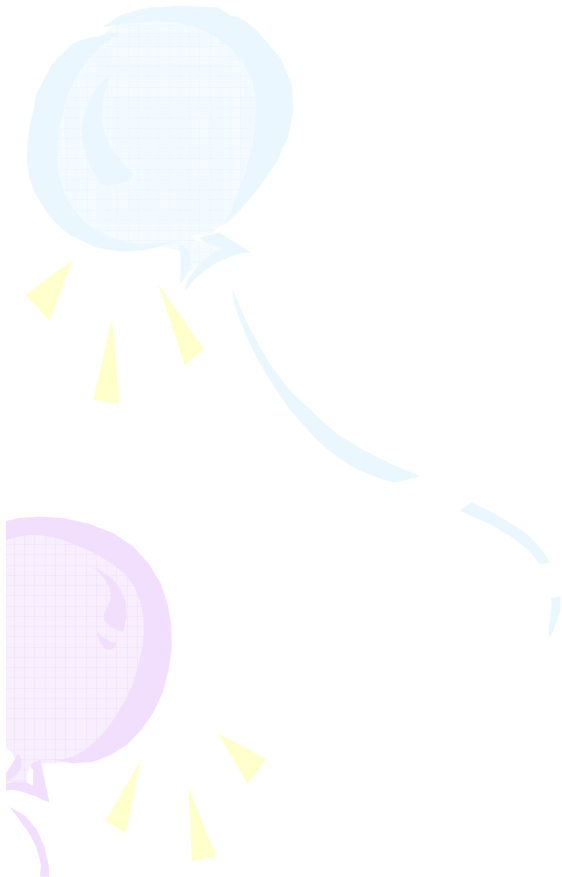
- Formal Review estimation model Excel spreadsheet.



Defect Prediction Model

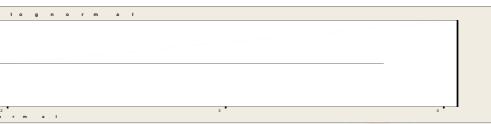
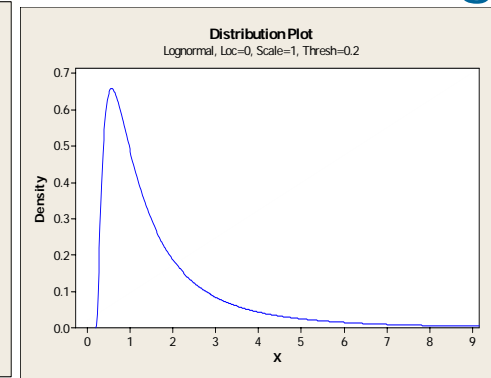
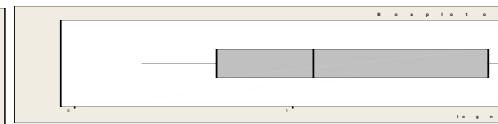
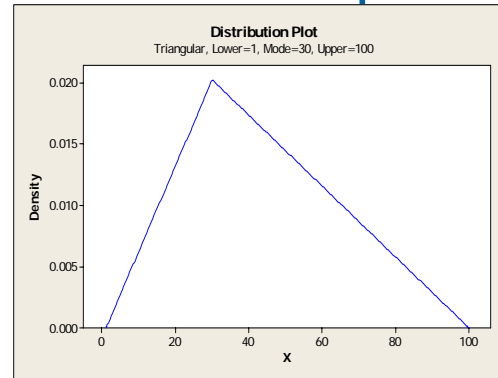
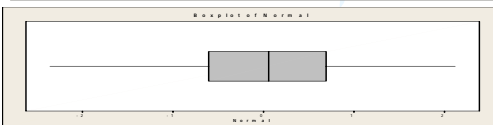
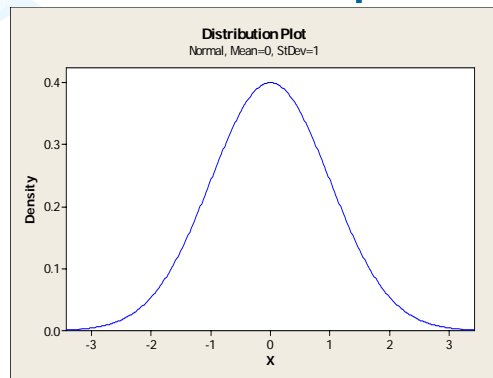


- Defect Prediction estimation model
Excel spreadsheet.



Understanding variance of data

- Populations of data are characterized by their distribution
- Minimum, median and maximum of a population
- Resulted predictions are presented as a range



Monte Carlo simulation



- We will use a Monte Carlo simulation that approximates a triangular distribution and takes 3 inputs:

- minimum
- median
- maximum

of our predicted data population.

- The simulation will provide a prediction and a confidence factor associated with that prediction



Confidence factor

- Confidence factors:
 - 99.99%
 - 55%
 - 33%

- Industry:
 - pharmaceutical drug research
 - banking
 - software development

PCG model

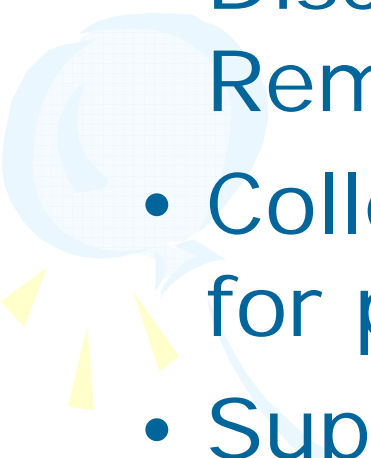



- Monte Carlo simulation with numbers from slide 9.

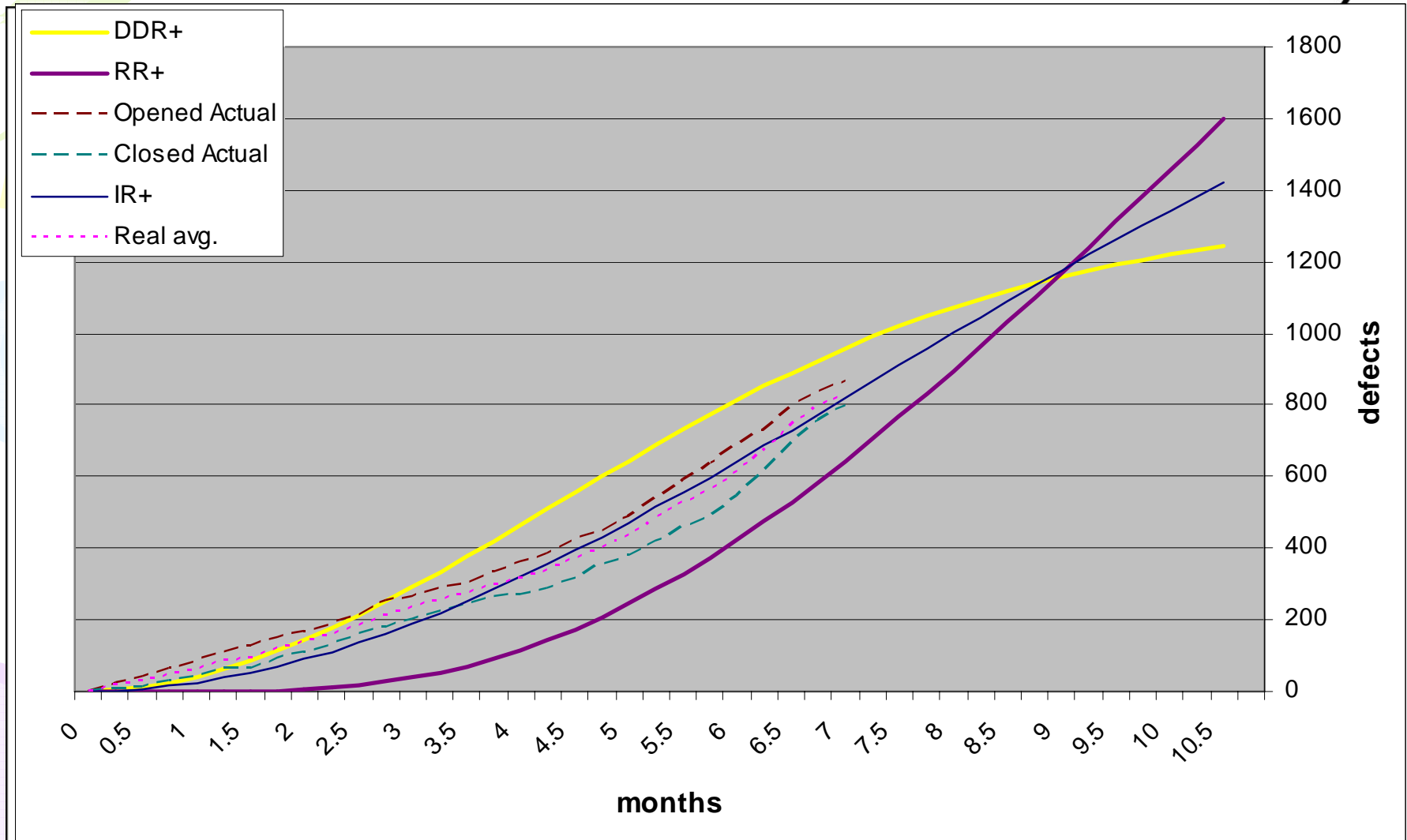


Defect Fix Rate model

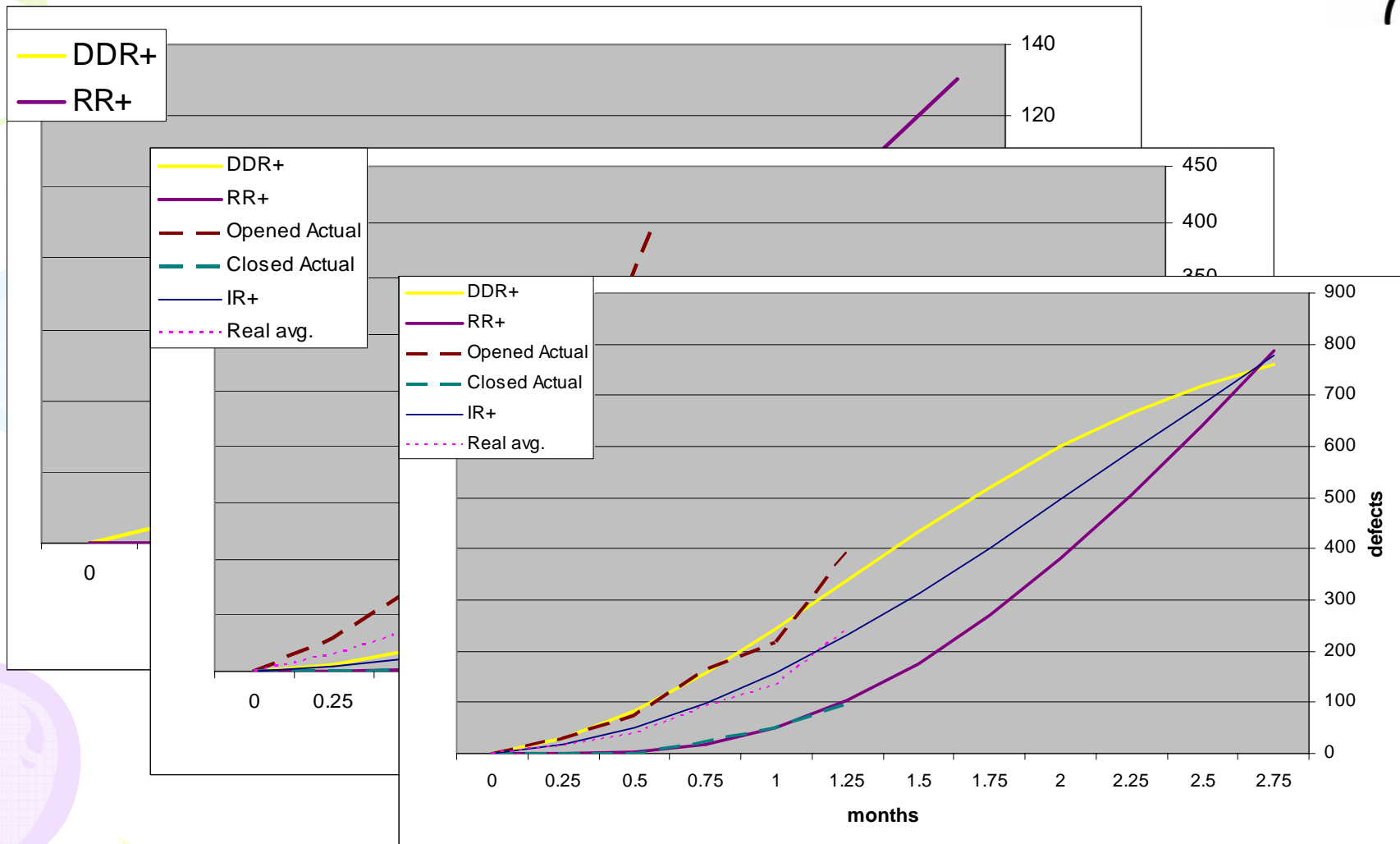


- Build a math model for Defect Discovery Rate and Defect Removal rate
 - Collect real-time data for defects for projects
 - Super-impose the 2 charts
 - Calculate new timelines of defect Removal rates
- 
- 

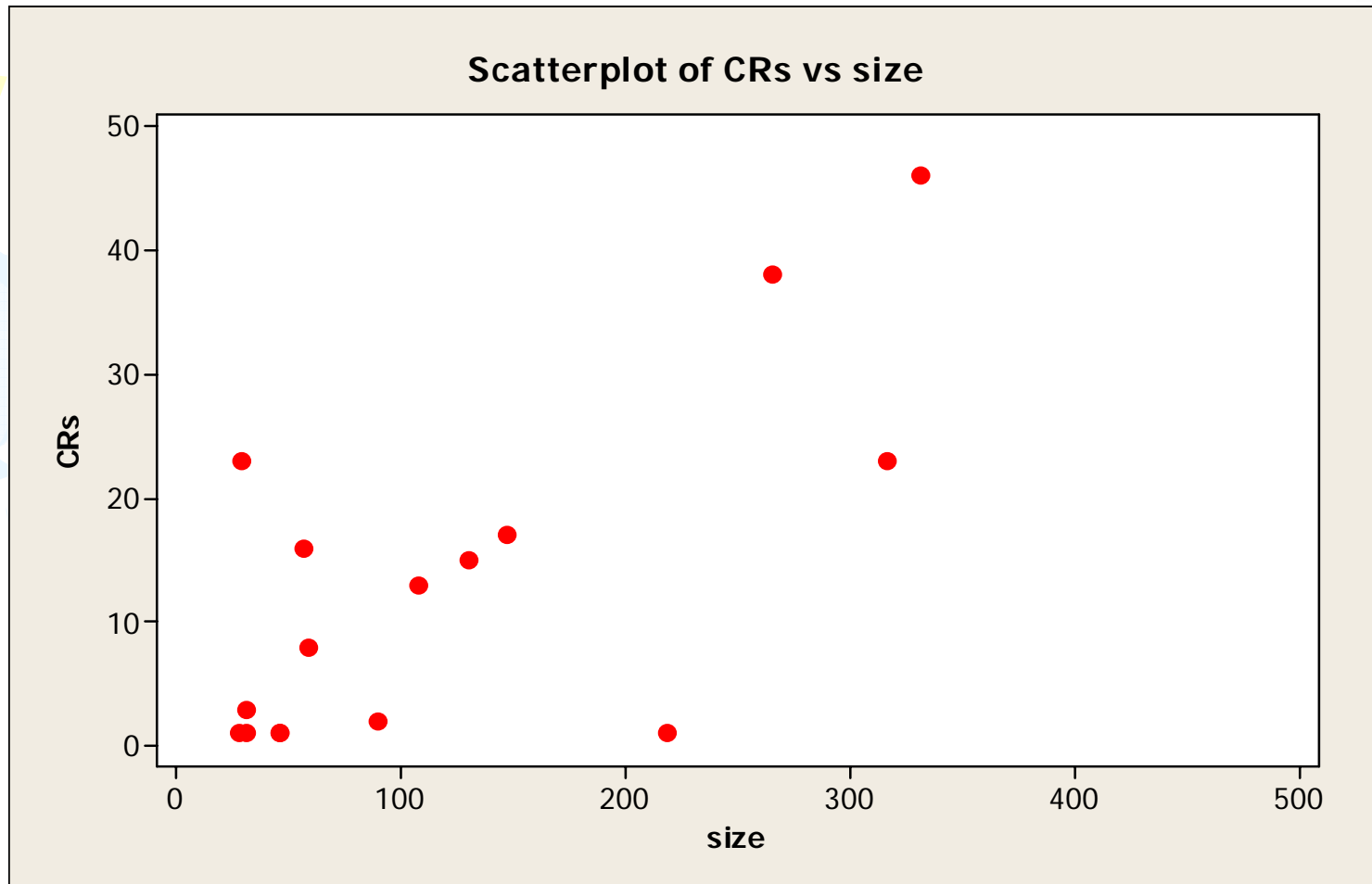
Real data vs. "theoretical" curve



Measurement Activities in Projects – Data analysis



Finding a correlation.



Use a statistical analysis tool to analyze the data



- Find if there is a correlation between the 2 sets of data
- If there is a correlation, how strong that is
- Identify “unusual” data-points
- What is the formula



Analysis results

- **Regression Analysis: CRs versus Size**

The regression equation is

$$\text{CRs} = 1.584 + 0.09530 \text{ Size}$$

$$S = 9.97948 \quad R\text{-Sq} = 52.1\% \quad R\text{-Sq}(\text{adj}) = 48.6\%$$

Unusual Observations

Obs	Size	CRs	Fit	SE Fit	Residual	St Resid
16	218	1.00	22.34	3.45	-21.34	-2.28R

Pearson correlation of Size and CRs = 0.722

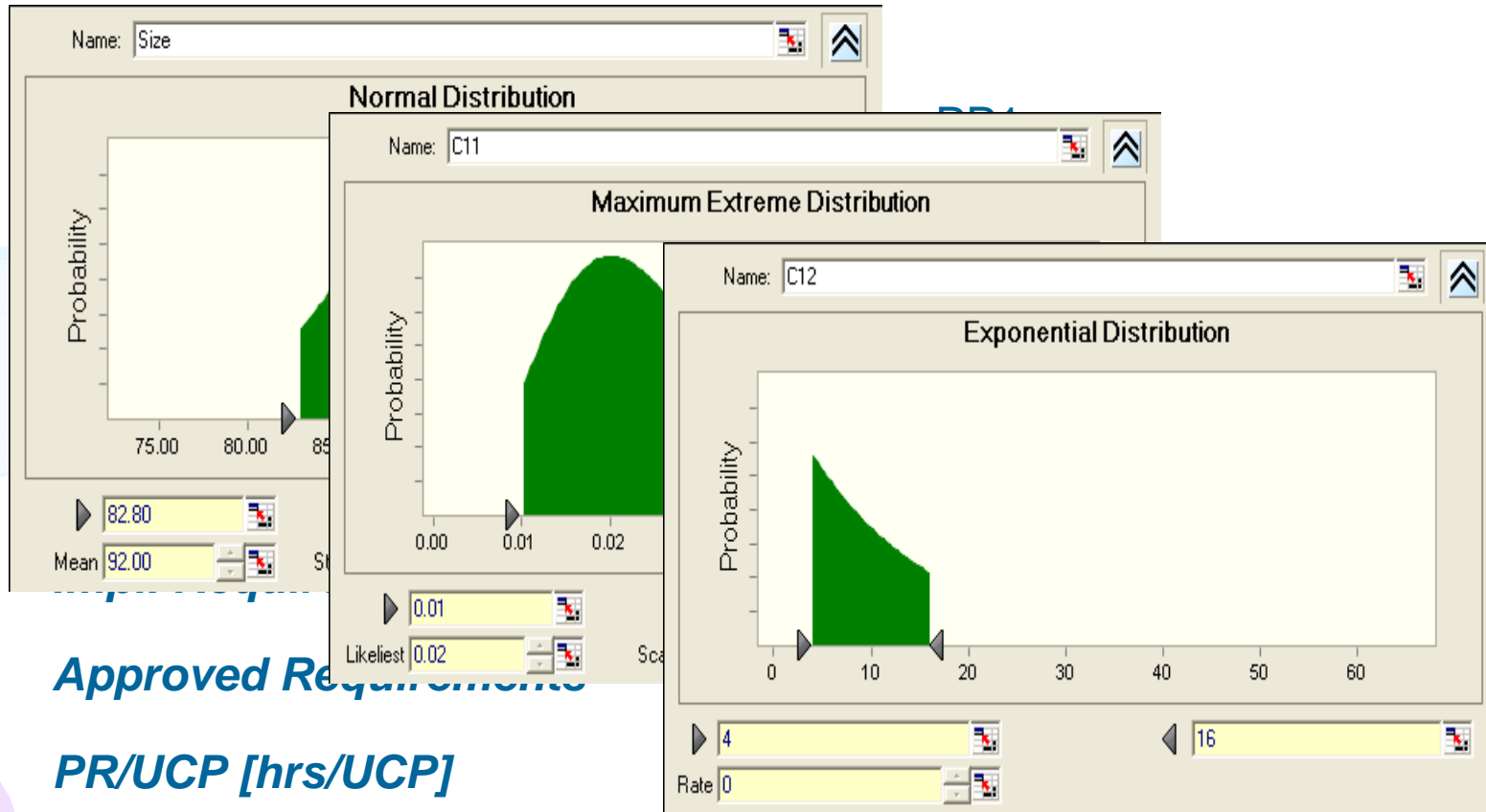
P-Value = 0.002

Project effort estimation model



- Gathered actual data for 150 projects
- Statistical analysis on these projects to identify consistency “buckets” — this could be very tedious and time consuming
- Multiple linear regression analysis — to identify meaningful factors and relationship between effort and those factors
- Statistical analysis on the delivery rate of each “bucket”

Project effort estimation model (continued)

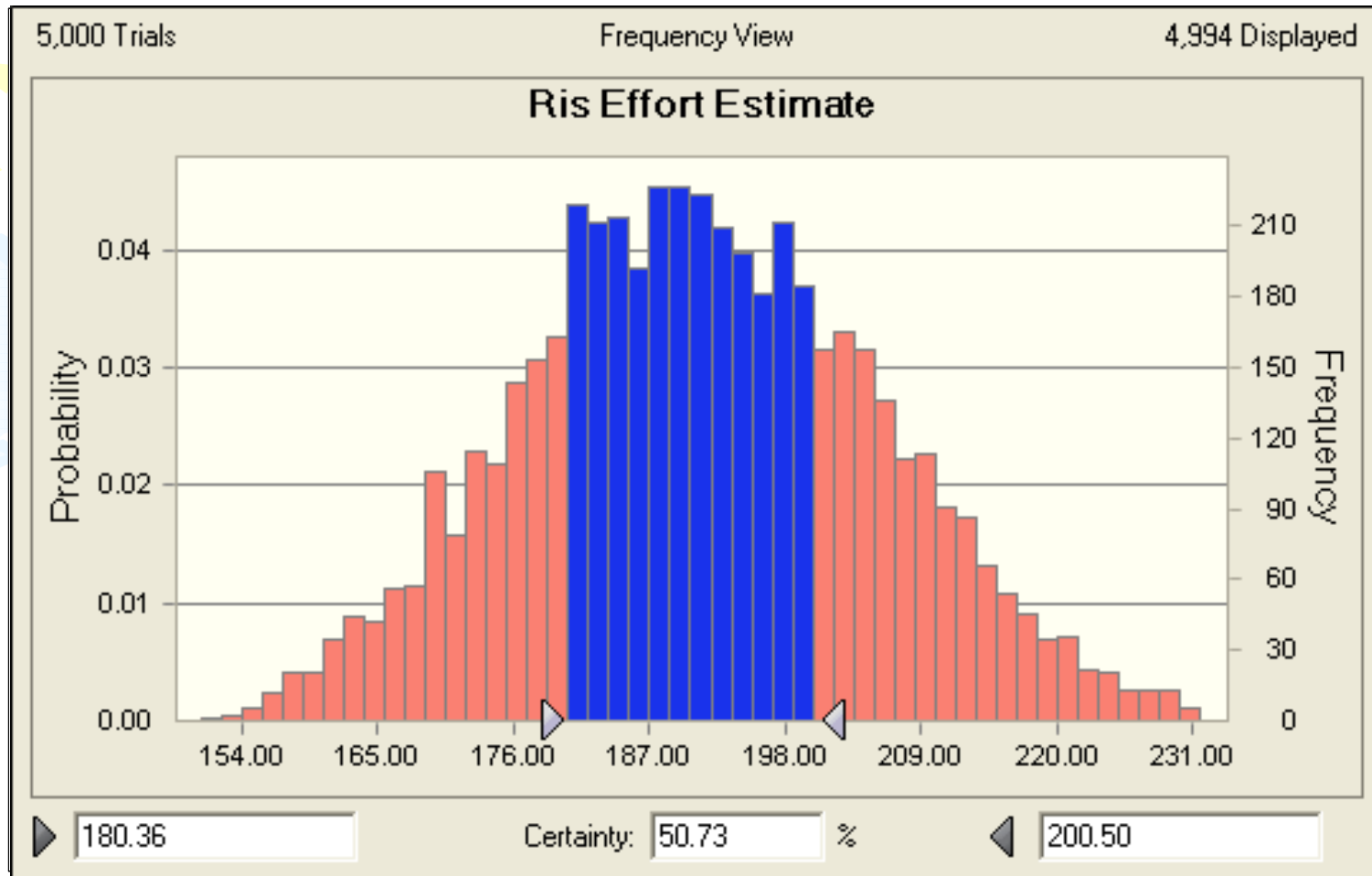


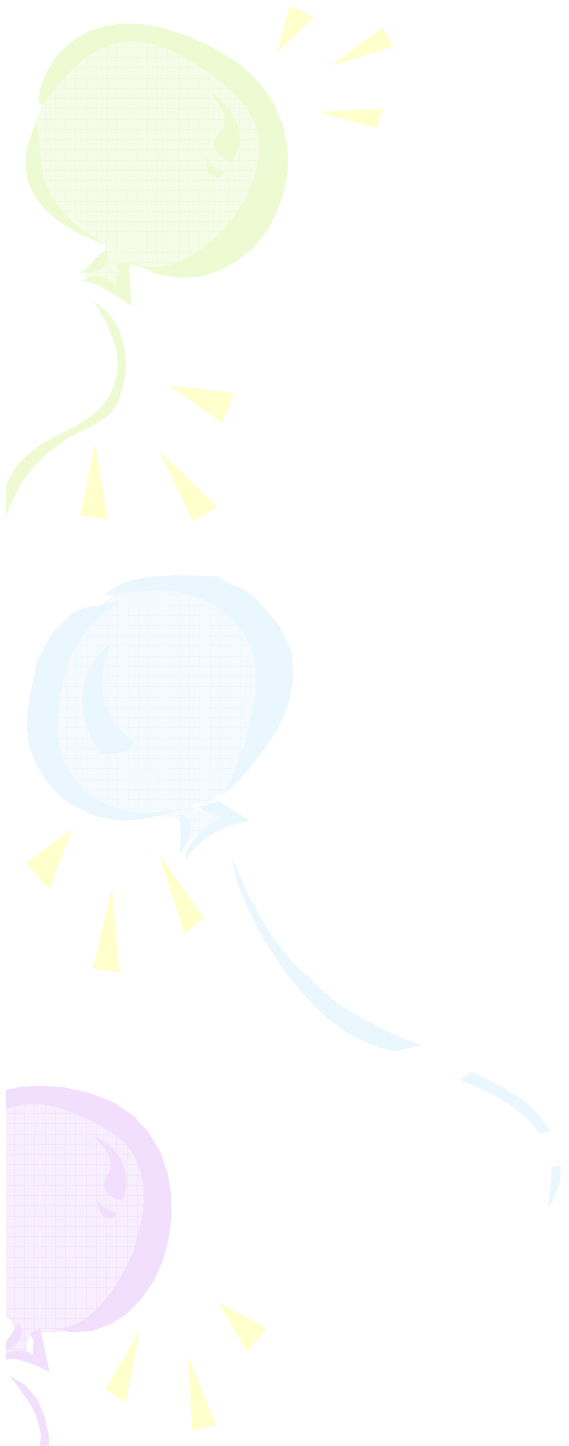
Approved Requirements

PR/UCP [hrs/UCP]

CR Total

Project effort estimation model (continued)





Q&A

