Software Size Growth Study

June 2017

Presenter: Marc Russo,
Corinne Wallshein
Naval Center for Cost Analysis
marc.russo1@navy.mil
corinne.wallshein@navy.mil

Outline

- Abstract
- Study questions
- GAO recommendation on software growth
- Data
- Percent change overview
- Uncertainty overview
- Example problem
- Correlation and subsets
- Conclusion and future research

Abstract

Software cost estimating relationships often rely on software size growth percentages.

Actual delivered source lines of code (SLOC) may be predicted with categories of early code estimates such as new, modified, reuse, and auto-generated SLOC. Uncertainty distributions will be presented to represent growth by code category for use in cost modeling.

Uncertainty distributions will be based on the actual percentage growth for Department of Defense programs' computer software configuration items in selected data subsets.

Questions Answered by Study

- What is the growth or shrinkage for types of SLOC (New, Modified, Reused, Auto-Generated, and Total), requirements, peak staff, effort hours, and duration?
- What uncertainty should be associated with growth?
- Is requirements growth correlated to SLOC growth?
- What other areas can be explored?

GAO on Software Growth/Shrinkage

Per 2009 GAO Cost Estimating and Assessment Guide:

"It is extremely important to include the expected growth in software size from requirements growth or underestimation (that is, optimism). Adjusting the software size to reflect expected growth from requirements being refined, changed, or added or initial size estimates being too optimistic and less reuse than expected is a best practice. **This growth adjustment should be made before performing an uncertainty analysis** [on effort or cost CERs created from actual, final reports]. Understanding software will usually grow, and accounting for it by using historical data, will result in more accurate software sizing estimates."

Data

- Non-random sample of secondary data
- Projects reported at the CSCI level by Software Resource Data Reports on the OSD/CAPE website called Cost Assessment Data Enterprise
- Content
  - Allows for collection of project context, responsible company or government entity, certified maturity level, requirements count, product size, effort hours, and schedule

Description of Data Processing

Each program submitted:

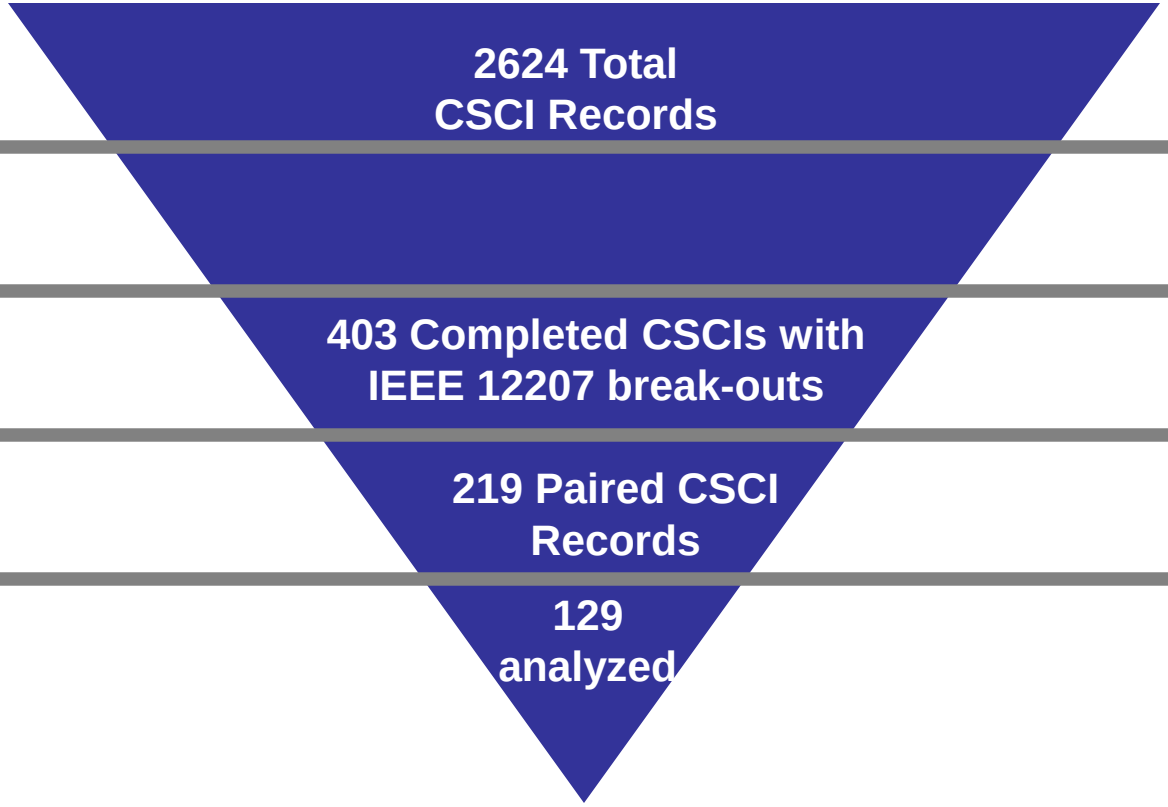
SRDR Initial Developer Report (Estimates)

&

SRDR Final Developer Report (Actuals)

- Analysis based on a subset of paired initial to final records from 2014 SRDR data set:
  - Requirements between 10 and 1000
  - Total SLOC between 100 and 1 Million
  - Effort Hours below 150,000

Data Analysis Pedigree



Since last ICEAA (2016)
Outliers and records outside analysis scope were excluded

Data Demographics (SLOC)

Variable	Quantiles			Moments						
	Max	Median	Min	Mean	Std Dev	SE Mean	N	Skewness	Kurtosis	CV
Initial New LS	192000	12028	120	25858	36928.37	3251.36	129	2.70	7.81	142.81
Final New LS	268800	18644	500	37370	49402.21	4349.62	129	2.25	5.47	132.20
Initial Modified LS	158718	2000	0	10548	25628.81	2256.49	129	4.33	20.47	242.97
Final Modified LS	196168	640	0	9463	25359.16	2232.75	129	4.99	29.23	267.99
Initial Reused LS	514800	7900	0	44556	94915.04	8356.80	129	3.41	12.18	213.03
Final Reused LS	617008	6000	0	55031	111247.56	9794.80	129	2.89	8.83	202.15
Initial Auto-Generated LS	16490	0	0	293	1940.40	170.84	129	6.94	49.39	661.68
Final Auto-Generated LS	213650	0	0	3247	20735.86	1825.69	129	8.97	86.71	638.53
Initial SLOC LS	614111	48237	904	81256	107902.35	9500.27	129	2.74	8.72	132.79
Final SLOC LS	818071	46200	1169	105111	141337.50	12444.07	129	2.63	8.41	134.47

- All data either reported in Logical Statements (LS) count or converted using the following:
  - Logical Statements (LS) = 0.66 x Non-Commented Source Statements (NCSS)
  - LS = 0.33 x Physical Source Lines of Code (SLOC)

Data Demographics (Other Variables)

Variable	Quantiles			Moments						
	Max	Median	Min	Mean	Std Dev	SE Mean	N	Skewness	Kurtosis	CV
Initial Effort Hours	133855	18643	575	31122.61	32456.77	2857.66	129	1.58	1.85	104.29
Final Effort Hours	139786	27265	1486	37799.27	35288.98	3107.02	129	1.27	0.78	93.36
Initial Requirements	990	184	10	274.19	260.11	22.90	129	1.14	0.30	94.86
Final Requirements	965	208	11	275.53	246.38	21.69	129	1.18	0.65	89.42
Initial Duration (Months)	100.11	20.02	0.23	20.59	19.67	1.73	129	1.11	1.64	95.57
Final Duration (Months)	109.09	21.01	0.36	21.48	20.40	1.80	129	1.39	3.56	94.99
Initial Peak Staff	71	8	1	11.84	12.57	1.11	129	2.27	5.68	106.19
Final Peak Staff	69	9	1	12.24	11.70	1.03	129	2.05	5.14	95.61

- All data either reported in Logical Statements (LS) count or converted using the following:
  - Logical Statements (LS) = 0.66 x Non-Commented Source Statements (NCSS)
  - LS = 0.33 x Physical Source Lines of Code (SLOC)

Process Overview

- From the data set have the ability to calculate percent change from initial to final using this formula:

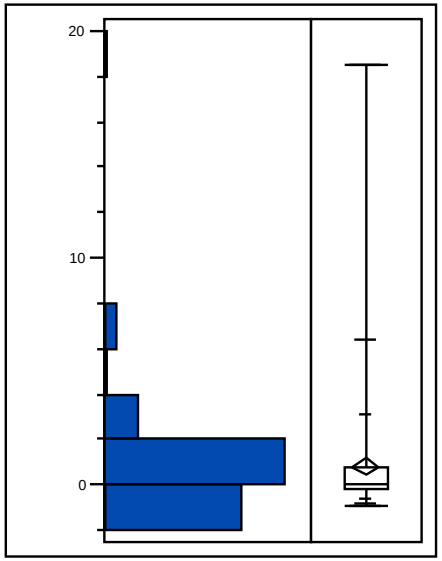
$$\boldsymbol{Percent\ Change = \frac{(Final - Initial)}{Initial}}$$

- Calculations were performed on all code types, requirement counts, duration in months, effort hours, and peak staff

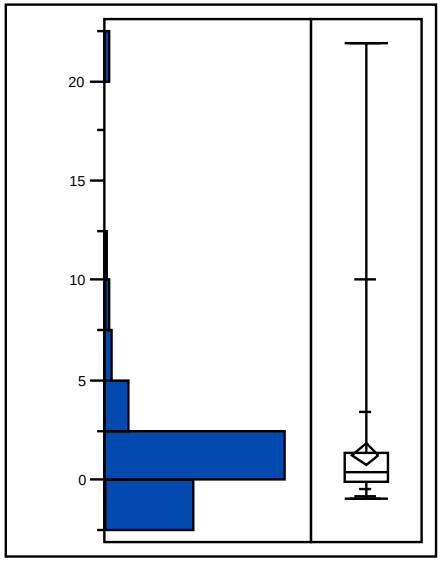
- Crystal Ball batch fit capability used to determine best fit for percent change uncertainty

Percent Change (PC) Summary SLOC (Logical Statements [LS])

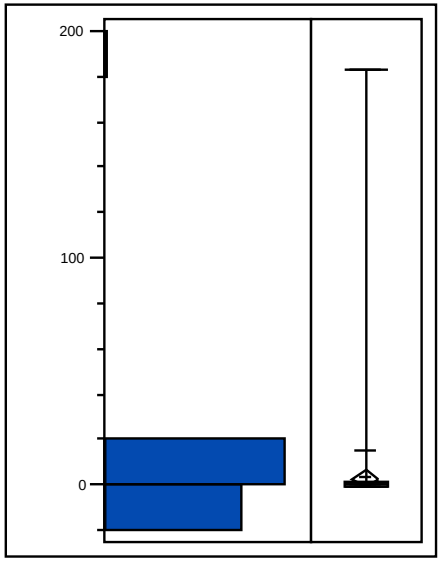
Total SLOC in LS



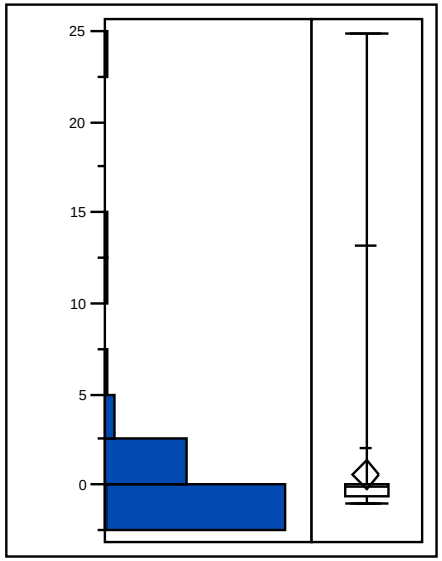
New LS



Mod LS



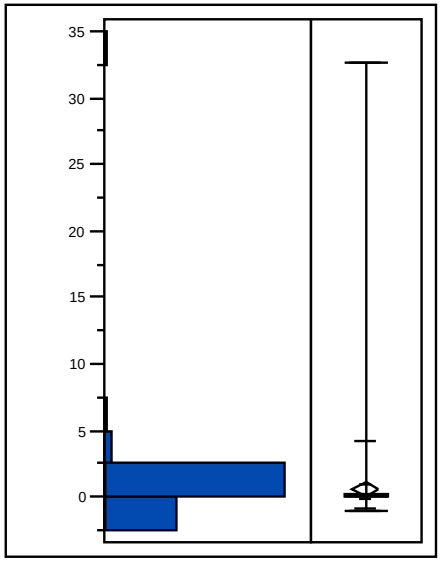
Reused LS



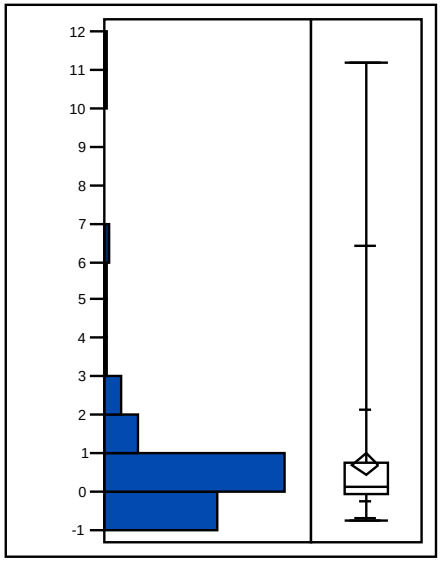
Variable	Quantiles			Moments						
	Max	Median	Min	Mean	Std Dev	SE Mean	N	Skewness	Kurtosis	CV
PC for New LS	21.90	0.37	-0.94	1.26	3.16	0.28	129	4.57	25.27	251.23
PC for Modified LS	182.73	0.01	-1.00	2.65	19.26	2.02	91	9.28	87.58	726.13
PC for Reused LS	24.88	-0.11	-1.00	0.55	3.49	0.38	83	5.23	31.30	634.92
PC for Auto-Generated LS	1.01	-0.78	-1.00	-0.39	0.94	0.47	4	1.89	3.61	-242.09
PC for Total SLOC in LS	18.55	0.05	-0.93	0.78	2.20	0.19	129	4.86	33.58	281.32

Percent Change (PC) Summary Other Variables

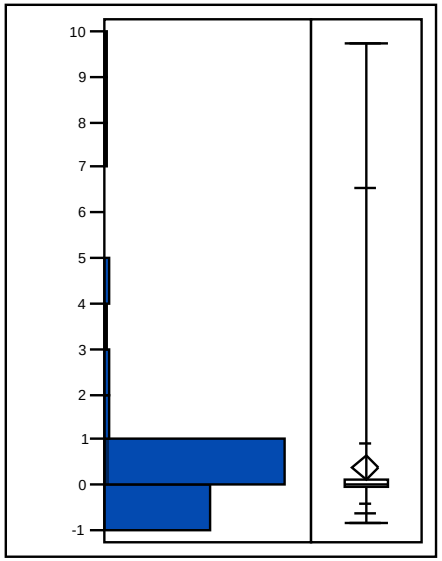
Duration (Months)



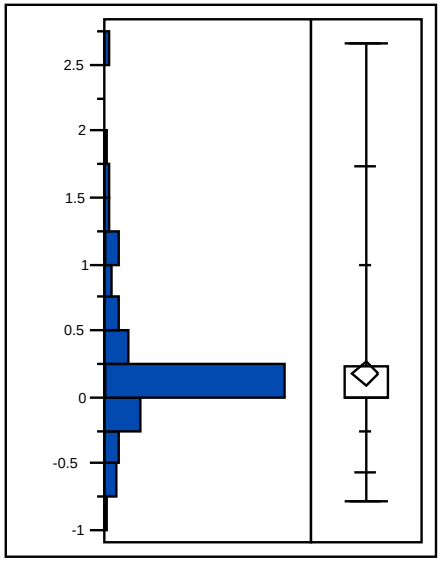
Effort Hours



Requirements



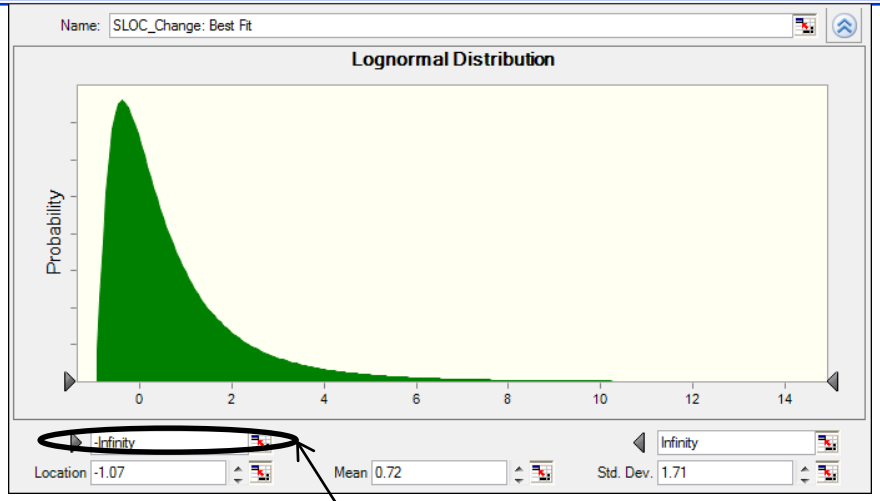
Peak Staff



Variable	Quantiles			Moments						
	Max	Median	Min	Mean	Std Dev	SE Mean	N	Skewness	Kurtosis	CV
PC in Duration (Months)	32.63	0.01	-0.98	0.53	2.99	0.26	129	9.89	105.71	567.83
PC in Effort Hours	11.20	0.14	-0.78	0.72	1.75	0.15	129	3.94	18.47	243.93
PC in Requirements	9.71	0.00	-0.83	0.36	1.51	0.13	129	4.38	21.08	415.21
PC in Peak Staff	2.67	0.00	-0.79	0.17	0.54	0.05	129	2.11	6.09	308.01

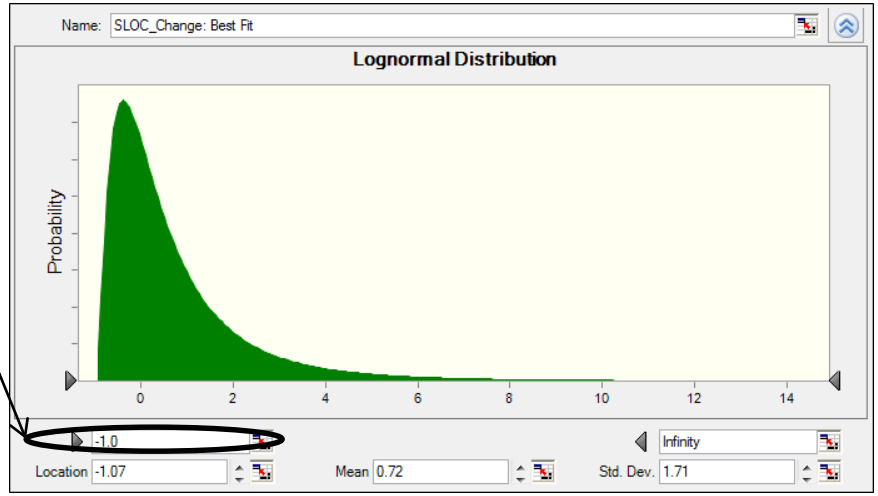
Uncertainty Overview

Uncertainty Distributions
SLOC Percent Change (Example)

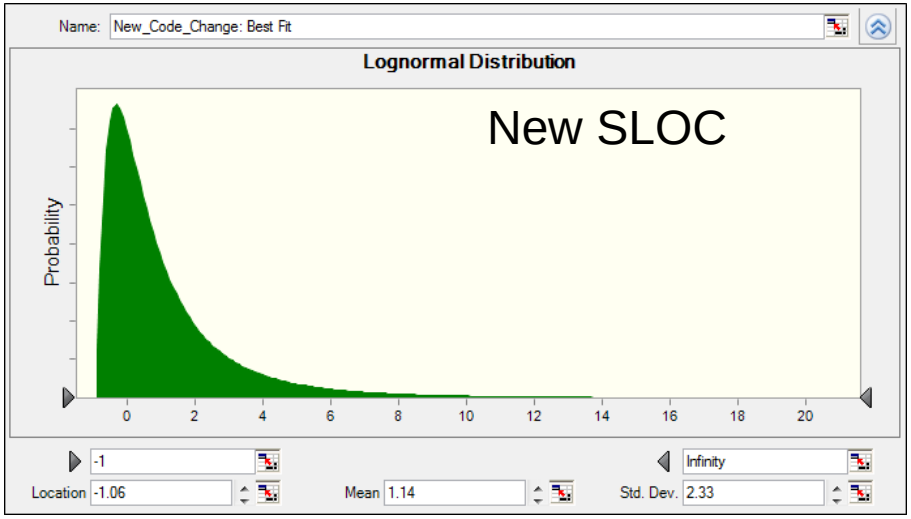


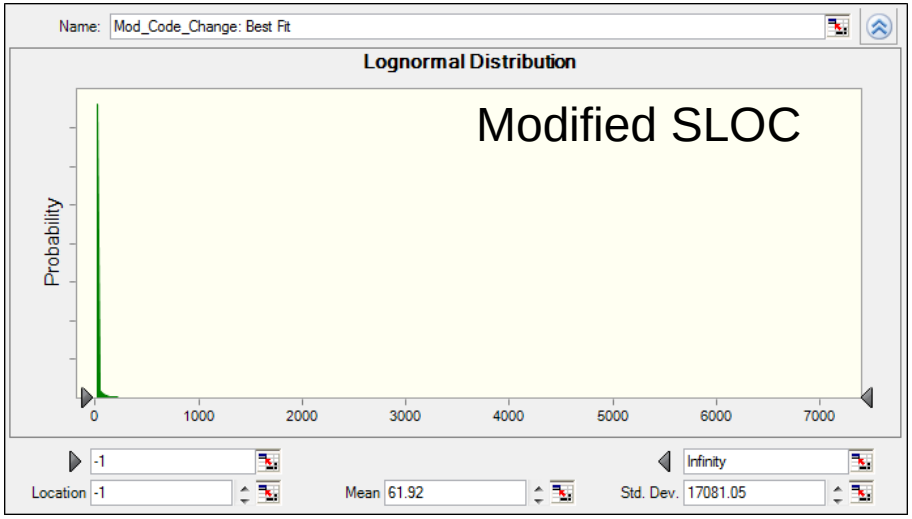
Distribution	A-D	A-D P-Value	Parameters
Lognormal	1.765	0.000	Mean=0.724, Std. Dev.=1.712, Location=-1.073
Gamma	3.576	0.000	Location=-0.940, Scale=1.421, Shape=1.212
Max Extreme	5.498	0.000	Likeliest=0.102, Scale=0.946
Weibull	8.428	0.000	Location=-0.935, Scale=1.508, Shape=0.791
Logistic	8.895	0.000	Mean=0.385, Scale=0.881
Normal	15.207	0.000	Mean=0.782, Std. Dev.=2.20
Student's t	15.866	---	Midpoint=0.782, Scale=0.781, Deg. Freedom=1.057
Min Extreme	27.727	0.000	Likeliest=2.239, Scale=4.740
BetaPERT	35.720	---	Minimum=-1.01, Likeliest=-0.935, Maximum=20.194
Beta	96.275	---	Min=-0.426, Max=403.425, Alpha=0.3, Beta=100
Triangular	114.089	---	Minimum=-1.01, Likeliest=-0.935, Maximum=20.194
Uniform	172.307	0.000	Minimum=-1.084, Maximum=18.698

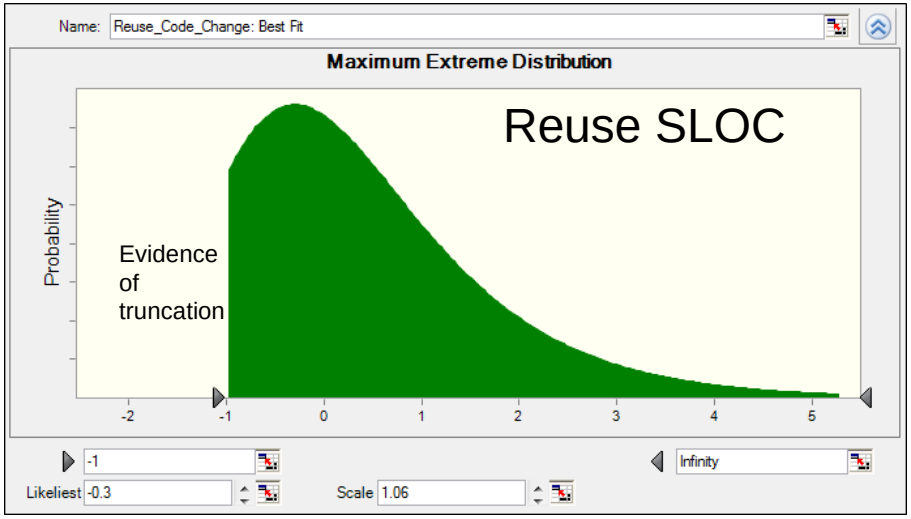
To ensure that uncertainty range does not provide a negative value (for Total SLOC) each distribution needs to be truncated at -1



Uncertainty Distributions SLOC Percent Change

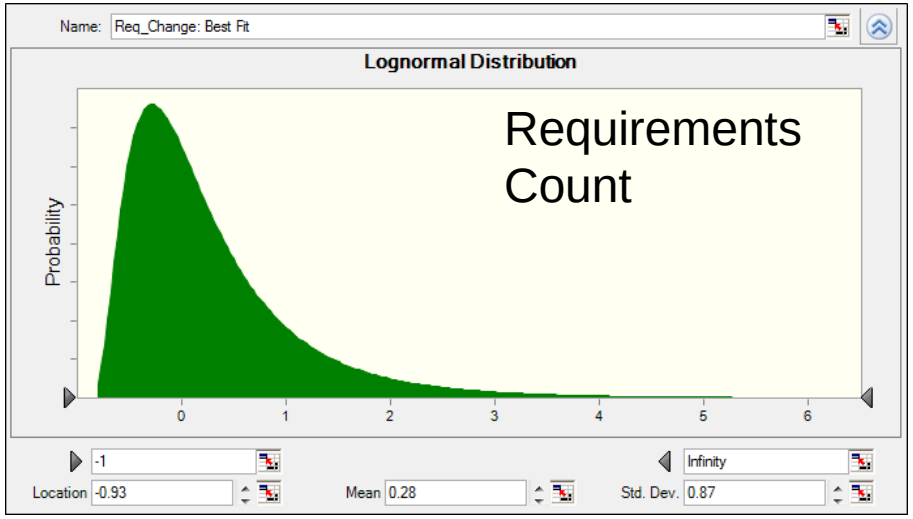


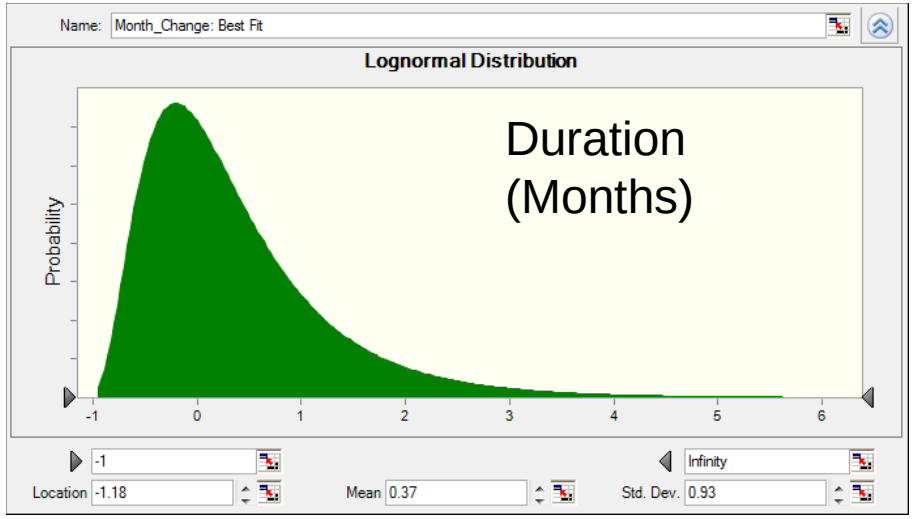


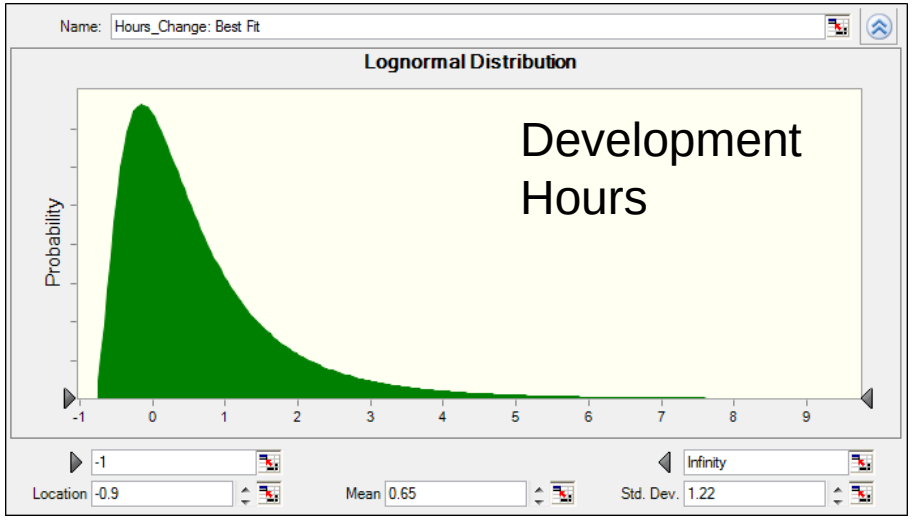


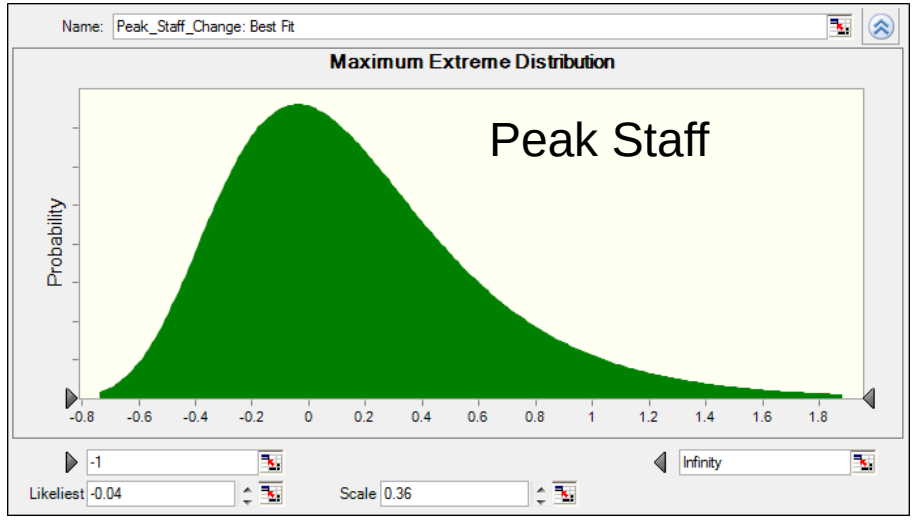
- Auto-generated distribution not available due to Crystal Ball Batch Fit requiring 15 data points
  - See Data Demographic chart

Uncertainty Distributions
Other Variables Percent Change









Example

- Program is able to provide SLOC, in logical statements, by initial New, Modified, Reuse, and Auto-Generated
- To estimate final data sizes, apply growth factors to initial data sizes
- Program Data:

CSCI	New (Initial)	Mod (Initial)	Reuse (Initial)	Auto (Initial)
1	200	4,699	31,144	16,490
2	200	2,236	22,803	340
3	3,354	1,147	67,083	25,660
4	10,000	15,000	275,000	1,100

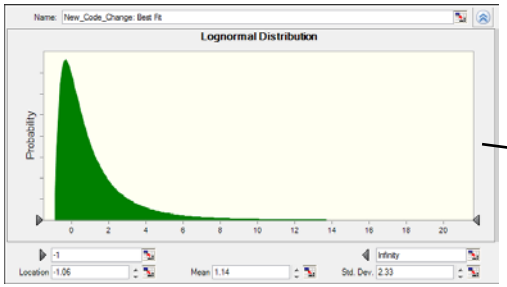
Example cont.

- Apply formula to initial variables

  Final = Initial * (1 + Percent Change)

CSCI	New (Initial)	1+ New PC	Mod (Initial)	1+ Mod PC	Reuse (Initial)	1+ Reuse PC	Auto (Initial)	1 + Auto PC
1	200	1+ 1.26	4,699	1 + 2.65	31,144	1 + .55	16,490	1 - .39
2	200	1+ 1.26	2,236	1 + 2.65	22,803	1 + .55	340	1 - .39
3	3,354	1+ 1.26	1,147	1 + 2.65	67,083	1 + .55	25,660	1 - .39
4	10,000	1+ 1.26	15,000	1 + 2.65	275,000	1 + .55	1,100	1 - .39

- Apply uncertainty (example)



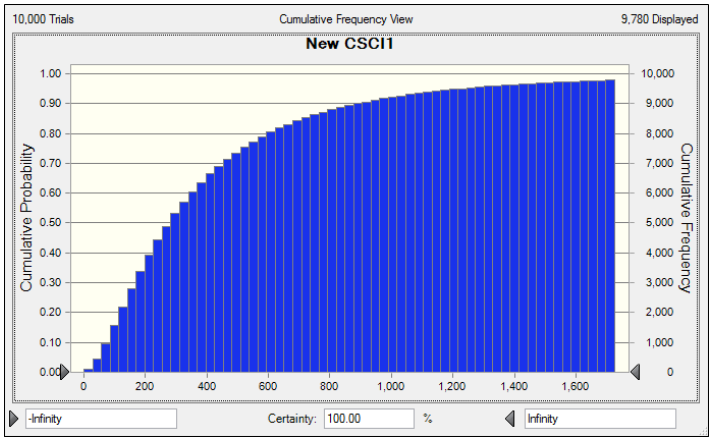
Example

- Results

CSCI	New (Initial)	1+ New PC	New (Final)	Mod (Initial)	1+ Mod PC	Mod (Final)	Reuse (Initial)	1+ Reuse PC	Reuse (Final)	Auto (Initial)	1 + Auto PC	Auto (Final)
1	200	1+ 1.26	451	4,699	1 + 2.65	17,166	31,144	1 + .55	48,284	16,490	1 - .39	10,082
2	200	1+ 1.26	451	2,236	1 + 2.65	8,168	22,803	1 + .55	48,284	340	1 - .39	208
3	3,354	1+ 1.26	7,571	1,147	1 + 2.65	4,190	67,083	1 + .55	48,284	25,660	1 - .39	15,689
4	10,000	1+ 1.26	22,573	15,000	1 + 2.65	54,795	275,000	1 + .55	48,284	1,100	1 - .39	673

- Uncertainty
  - As an example the uncertainty distribution and analysis is provided for **CSCI 1** New

New (Final)
451



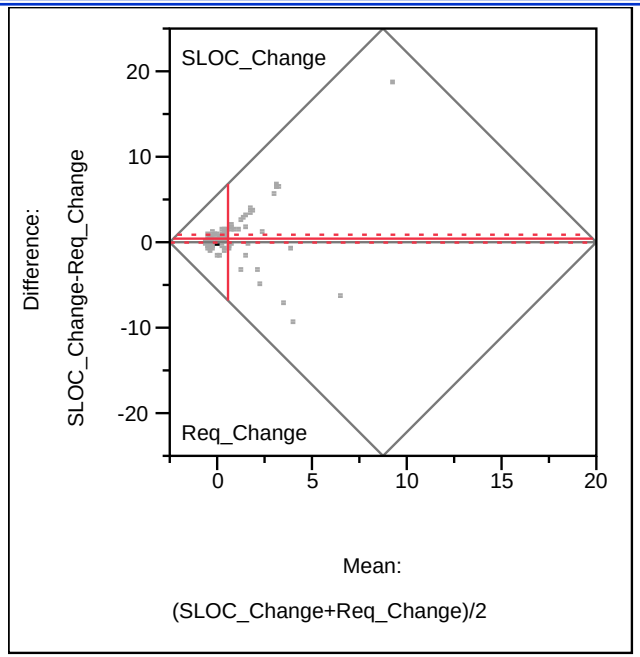
Percentile	New (Final)
10th	88
Mean	429
90th	907

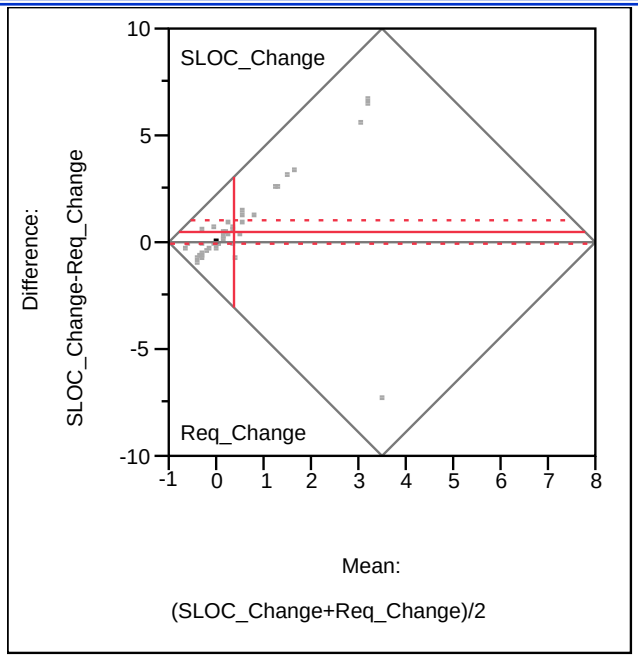
Uncertainty in growth levels should be applied to all CSCI factors

Additional Explorations

Requirements and SLOC

- Are Requirements and SLOC correlated?
- The data set shows no correlation between total SLOC change and requirements change though they both increase
- A second look, removing items with requirements count over 200, shows similar trend



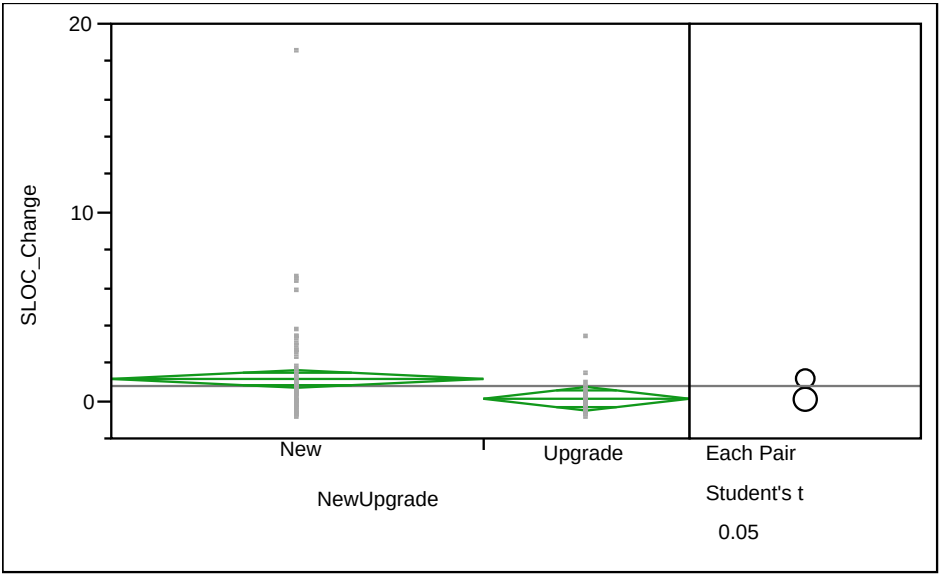


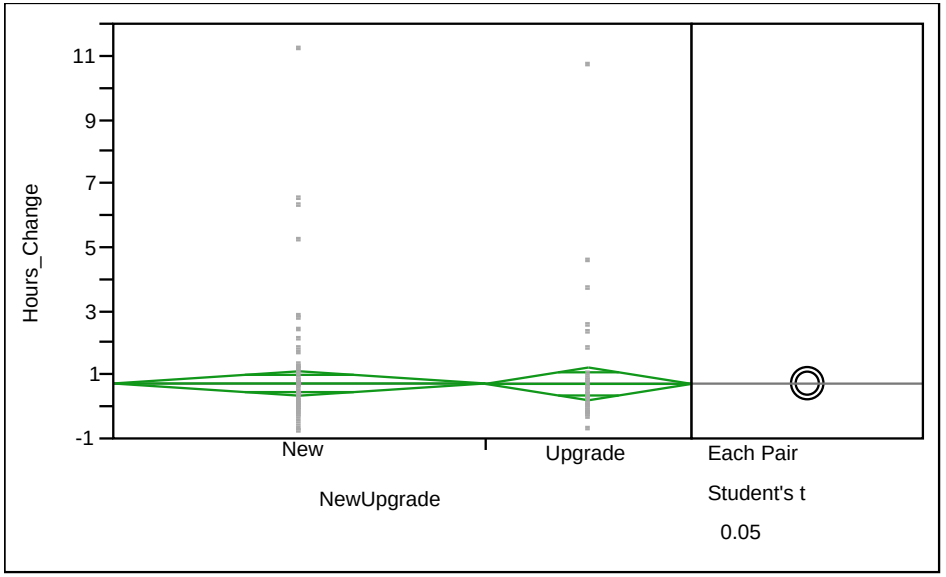
SLOC_Change	0.78
Req_Change	0.36
Mean Difference	0.42
Std Error	0.23
N	129
Correlation	0.025

SLOC_Change	0.62
Req_Change	0.14
Mean Difference	0.48
Std Error	0.27
N	53
Correlation	-0.049

Continued analysis into how requirements growth is related to SLOC should be conducted

New/Upgrade Percent Change ANOVA Analysis





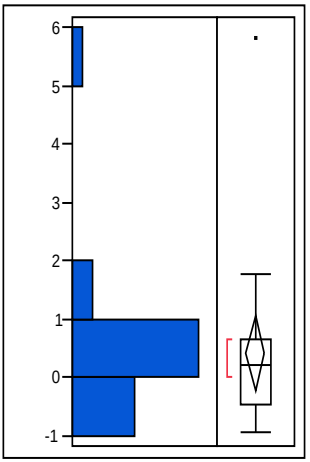
Oneway Anova Summary of Fit	
R^2	0.053
Adjusted R^2	0.045
Root Mean Square Error	2.149
Mean of Response	0.782
Observations (or Sum Wgts)	129

Oneway Anova Summary of Fit	
R^2	1.29e-5
Adjusted R^2	-0.008
Root Mean Square Error	1.760
Mean of Response	0.719
Observations (or Sum Wgts)	129

Mean difference for SLOC percent change for New versus Upgrade is pronounced
Means for Effort Hours percent change for New versus Upgrade are similar

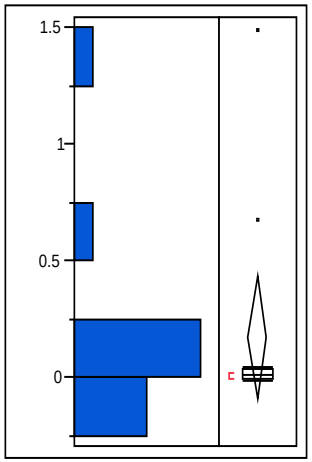
Program Type Percent Change SLOC Total

Aircraft- Fixed Wing



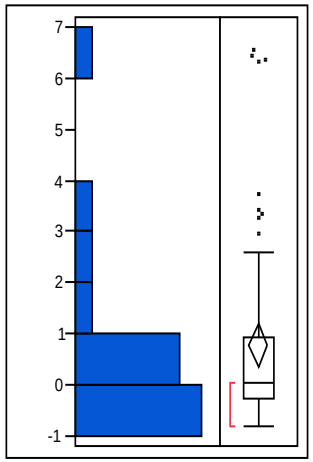
Mean	0.426
Std Dev	1.429
Std Err Mean	0.312
Upper 95% Mean	1.076
Lower 95% Mean	-0.225
N	21

Aircraft- Rotary Wing



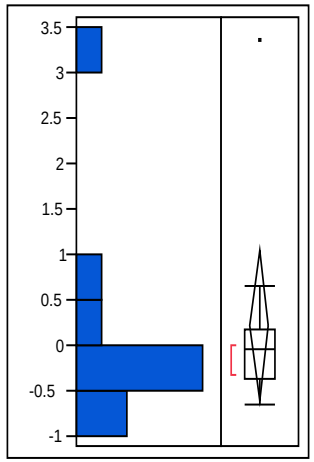
Mean	0.173
Std Dev	0.436
Std Err Mean	0.121
Upper 95% Mean	0.436
Lower 95% Mean	-0.090
N	13

C2-4I & Other



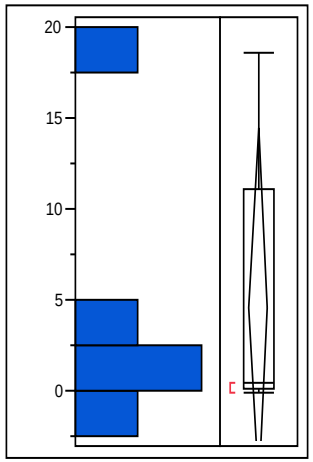
Mean	0.790
Std Dev	1.78
Std Err Mean	0.215
Upper 95% Mean	1.218
Lower 95% Mean	0.361
N	69

Missiles



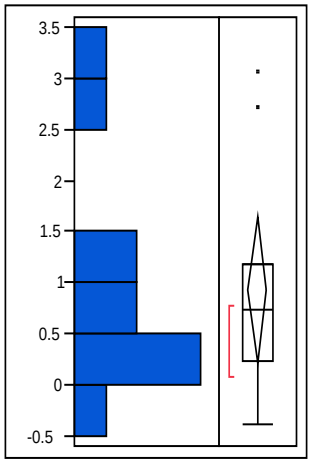
Mean	0.222
Std Dev	1.158
Std Err Mean	0.366
Upper 95% Mean	1.050
Lower 95% Mean	-0.606
N	10

Radar



Mean	4.563
Std Dev	7.970
Std Err Mean	3.564
Upper 95% Mean	14.459
Lower 95% Mean	-5.333
N	5

Ships



Mean	0.924
Std Dev	1.066
Std Err Mean	0.321
Upper 95% Mean	1.640
Lower 95% Mean	0.208
N	11

Mean Total SLOC percent change for all programs was 0.78

Conclusion and Future Research

- From this analysis, Percent Change averages and uncertainties are available to estimate growth and cross check software cost estimates
- Based on the 129 data points, requirements growth is not directly correlated to Total SLOC growth
  - Mean percent change for both requirements and Total SLOC grows
- Percent change analysis should be updated and improved as more data becomes available
- Analysis on software size growth will be continued

Questions?