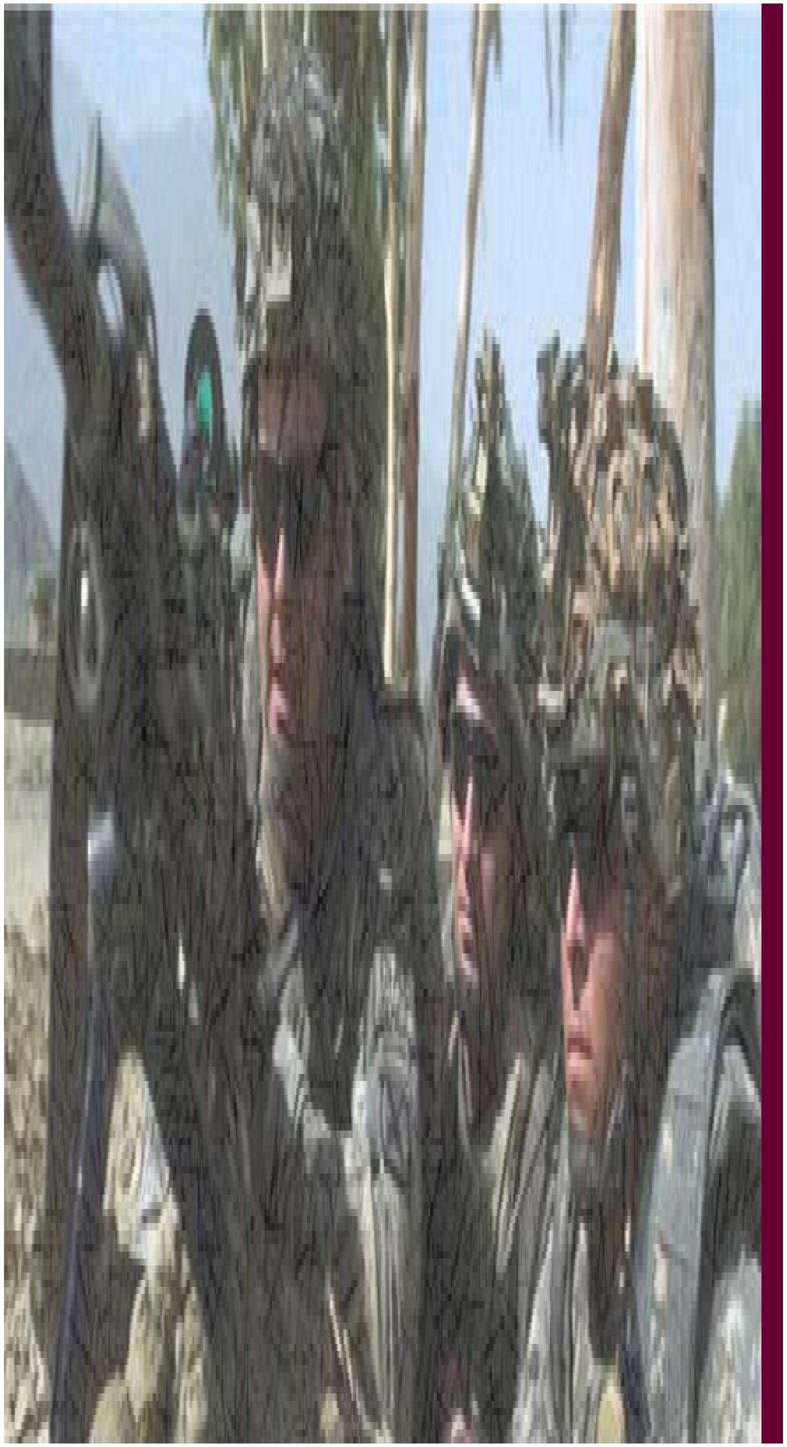


# Utilizing Statistical Inference to Guide Expectations and Test Structuring during Operational Testing and Evaluation

Joy Brathwaite
School of Aerospace Engineering
Georgia Institute of Technology

Dr. Alton Wallace
Dr. Robert Holcomb
Institute for Defense Analyses

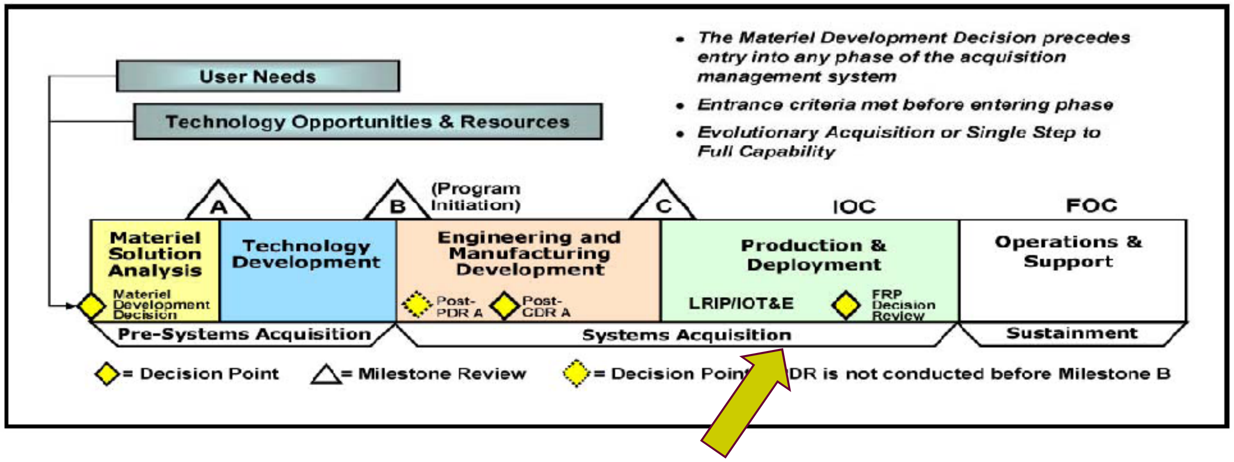
May 10-12, 2011

8th Annual Acquisition Research Symposium

# Outline of Presentation

- Motivation
- Statistical Inference and Operational Testing
- Evaluating Potential Test Results
- Application to Situational Awareness System
- Summary

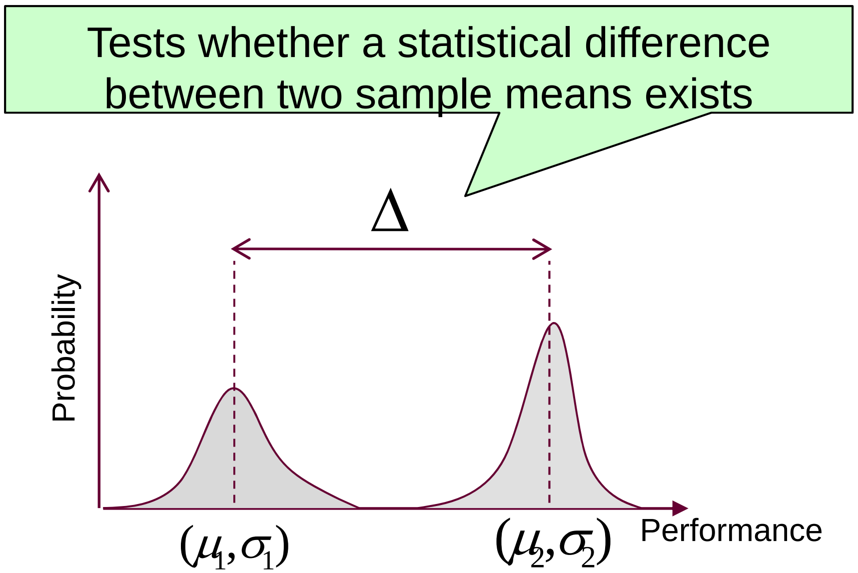
# Motivation




- Initial Operational Testing and Evaluation occurs during the Production & Deployment acquisition phase
- Congress requires testing of major weapons systems to be conducted under operationally realistic conditions to determine operational suitability
- Comparative tests are utilized during operational testing to baseline a system under test (SUT) through a series of tactical battles
  - Goal is to determine whether and by how much the unit's performance systematically improves with the SUT
  - Several approaches, both quantitative and qualitative, are used to assess a systematic improvement (e.g. statistical analysis and user evaluations)

# Statistical Inference

- Statistical inference noted as a best practice in system evaluation (CBASSE 1998)
- An applied statistical approach is often used to quantify and evaluate differences between treatment and control groups (Woolbridge 2003)
- In operational testing, statistical inference evaluates the performance difference between the SUT and the current status quo




| Interval/Ratio Data | |
|---|---|
| Two independent samples | **t-test**<br>z-test<br>single factor between subjects ANOVA |
| Two dependent samples | t-test<br>z test<br>single factor between subjects ANOVA |
| **Ordinal/Rank-Order Data** | |
| Two independent samples | Mann-Whitney U test<br>van der Waerden normal-scores test |
| Two dependent samples | Wilcoxon matched pairs signed-ranks test<br>**Binomial sign test** |
| **Categorical/Nominal Data** | |
| Two independent samples | Chi-square test<br>z-test |
| Two dependent samples | McNemar test<br>Gart test |

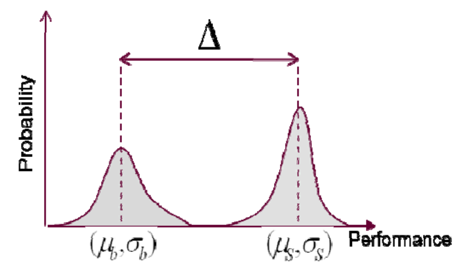
# Statistical Inference

1 **State Research Question**

Does use of the SUT improve the mean performance of a unit?

2 **Specify Null and Alternative Hypotheses**

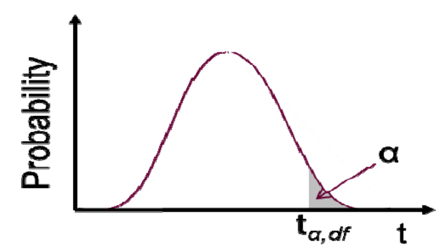
$$H_\phi : \mu_S = \mu_b$$
$$H_a : \mu_S > \mu_b$$




3 **Calculate Test Statistic**

$$t_{\alpha, df} = \frac{\overline{X}_S - \overline{X}_b}{\sqrt{\frac{s_S^2}{n_S} + \frac{s_b^2}{n_b}}}$$

4 **Compute Probability of Rejection**




5 **State Conclusions**

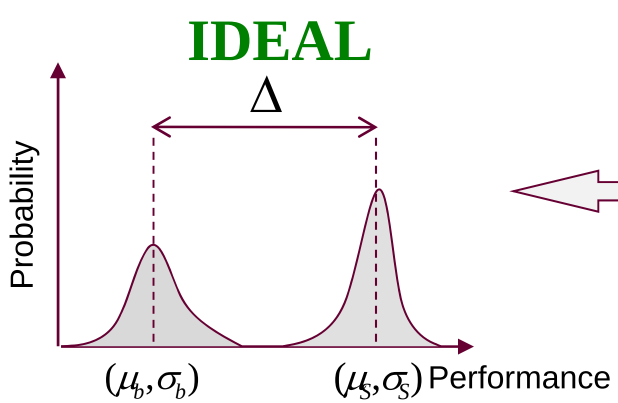
Did the SUT unit outperform the baseline unit statistically?

# Statistical Inference in OT&E

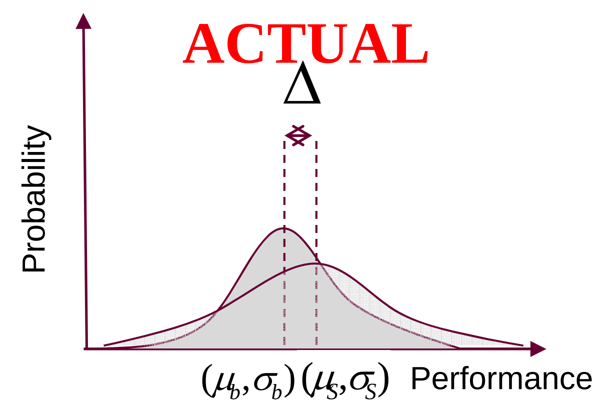
- Evaluated a Situational Awareness System as an effective tool against fratricide in 2001 (Edwards 2001)
  - **System Confidence Demonstration (SCD)**
    - No significant statistical difference between SUT and non-SUT units
    - Nearly impossible for SUT crew to statistically outperform baseline as baseline did so well
  - **Virtual Integration Exercise (VIE)**
    - Overall, no significant difference occurred in fratricide rates between baseline and SUT

Assessing the difference in performance mean between two independent samples

$$t_{\alpha,\nu} = \frac{\overline{X}_S - \overline{X}_b}{\sqrt{\frac{s_S^2}{n_S} + \frac{s_b^2}{n_b}}}$$




Small sample sizes and large variability lead to inconclusive results



# Evaluating Potential Test Results

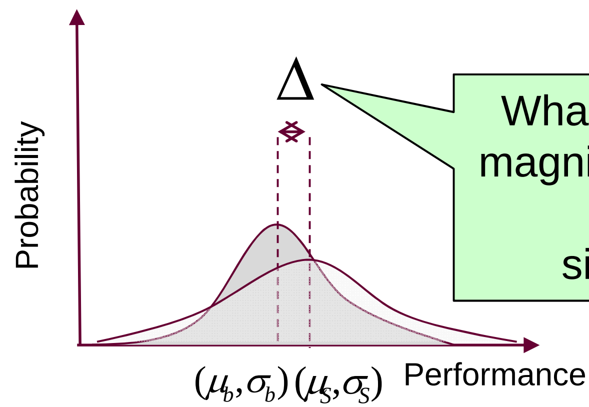
- Comparative tests are costly to administer and difficult to repeat
- Understand potential results a priori to guide expectations, test structuring and enable a more effective utilization of resources

1. What improvement in the mean performance is needed over the baseline to confidently assess whether there is a statistical difference?

2. Is the required performance of the unit needed to show a statistical difference reasonable?

$$\overline{X}_b = \overline{X}_S - t_{\alpha,\nu}\left(\sqrt{\frac{s_S^2}{n_S} + \frac{s_b^2}{n_b}}\right)$$

**ACTUAL**




What is the needed magnitude of delta for statistical significance?

# Guiding Expectations and Test Structuring

## Analysis of Systematic Difference

- Several approaches, both quantitative and qualitative, are used to assess a systematic improvement (e.g. statistical analysis and user evaluations)
- Statistical inference noted as a best practice in system evaluation (CBASSE 1998)

## Problems Experience in Previous Tests

- Evaluated a Situational Awareness System as an effective tool against fratricide in 2001 (Edwards 2001)
  - **System Confidence Demonstration (SCD)**
    - No significant statistical difference between SUT and non-SUT units
    - Nearly impossible for SUT crew to statistically outperform baseline as baseline did so well
  - **Virtual Integration Exercise (VIE)**
    - Overall, no significant difference occurred in fratricide rates between baseline and SUT

## Potential results of test a priori may:

- Provide guidance on the potential benefits of conducting test
- Provide guidance on structuring the test
- Lead to a more cost-effective test execution
- Provide maximal information given resources expended

# Outline of Presentation

- Motivation
- Statistical Inference and Operational Testing
- Evaluating Potential Test Results
- Application to Situational Awareness System
- Summary

# Examination of the Force Effectiveness

- Operational needs statements from theater called for ground and aerial robotic capability to enable better situational awareness
- Evaluation of a SUT to improve the unit situational awareness on the battlefield
- Data on SUT performance gathered from its LUT 09
- Operational performance evaluation of a battalion with and without the SUT systems

| Mission | Mission Type | Success | BLUFOR Starting Strength | BLUFOR Casualties | OPFOR Starting Strength | OPFOR Casualties |
|---|---|---|---|---|---|---|
| 1 | Raid | yes | 130 | 10 | 50 | 26 |
| 2 | Raid | yes | 130 | 7 | 50 | 25 |
| 3 | Defend | yes | 130 | 25 | 50 | 0 |
| 4 | Attack | yes | 130 | 15 | 50 | 10 |
| 5 | Attack | yes | 130 | 25 | 50 | 8 |
| 6 | Cordon and Search | yes | 130 | 8 | 50 | 7 |
| 7 | Defend | yes | 130 | 16 | 50 | 15 |
| 8 | Cordon and Search | yes | 130 | 12 | 50 | 6 |
| 9 | Raid | partially | 130 | 7 | 50 | 3 |
| 10 | Cordon and Search | yes | 130 | 20 | 50 | 8 |
| 11 | Attack | no | 130 | 14 | 50 | 10 |
| 12 | Stability Operations | yes | 130 | 2 | 50 | 5 |
| 13 | Raid | yes | 130 | 10 | 50 | 22 |

# Performance Metrics of Interest

1. **Missions Not Accomplished** — Considered missions which had a conclusive result

2. **Mission Success Rate**

$$MSR = \frac{Number\ of\ Missions\ Accomplished}{Total\ Missions\ Conducted}$$

3. **BLUFOR Casualty Rate**

$$B\ Casualty\ Rate = \frac{BLUFOR\ Losses}{BLUFOR\ Starting\ Strength}$$

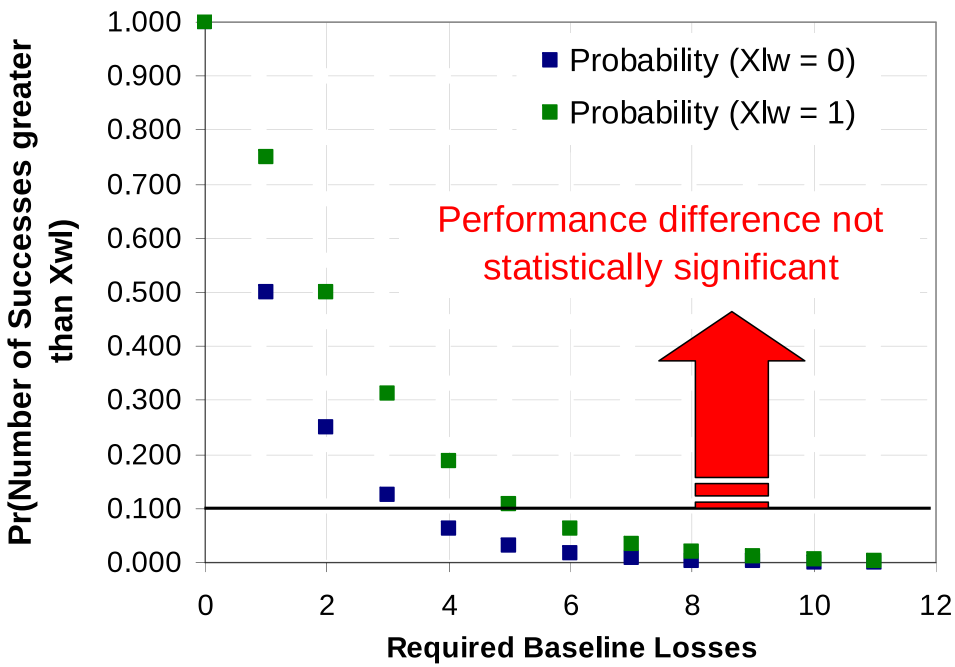
4. **OPFOR Casualty Rate**

$$O\ Casualty\ Rate = \frac{OPFOR\ Losses}{OPFOR\ Starting\ Strength}$$

5. **BLUFOR Fratricide Rate**

$$Fratricide\ Rate = \frac{BLUFOR\ Fratricides}{BLUFOR\ Losses}$$

# Missions Not Accomplished




Xwl – Number of missions accomplished by SUT unit but not baseline unit

Xlw – Number of missions accomplished by baseline unit but not SUT unit

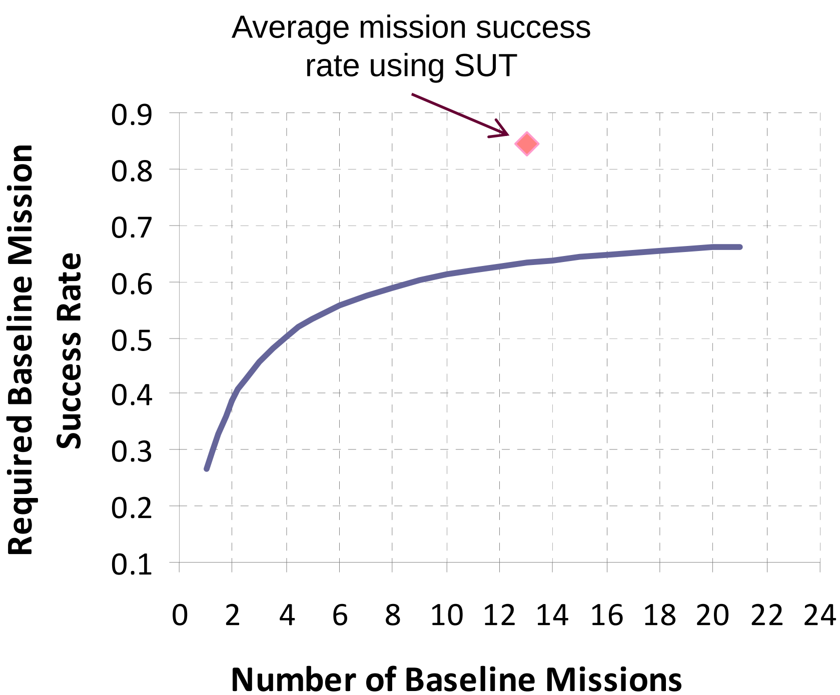
- Comparative evaluation using binomial sign test at 90% confidence level (Sheskin 2004)
- Given the results of the LUT 09, the baseline unit would have to lose 4 or more missions to statistically underperform the SUT unit

- Given the starting strength ratio of 2:1, it is unlikely the baseline unit will lose 4 missions
- Modify test structure to use a lower starting strength ratio

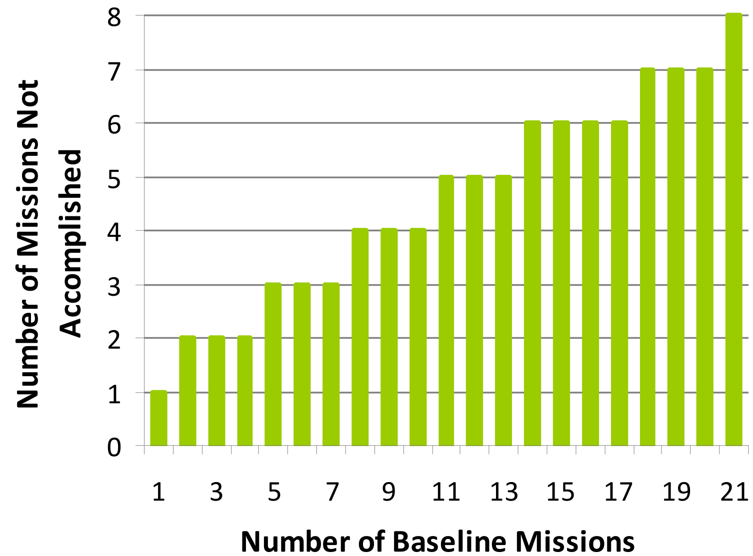
# Mission Success Rate

- Comparative evaluation using two proportion z-test at 90% confidence level

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$




- Given an expected 13 baseline missions to be conducted, the required performance of the baseline unit is a maximum mission success rate of 63%



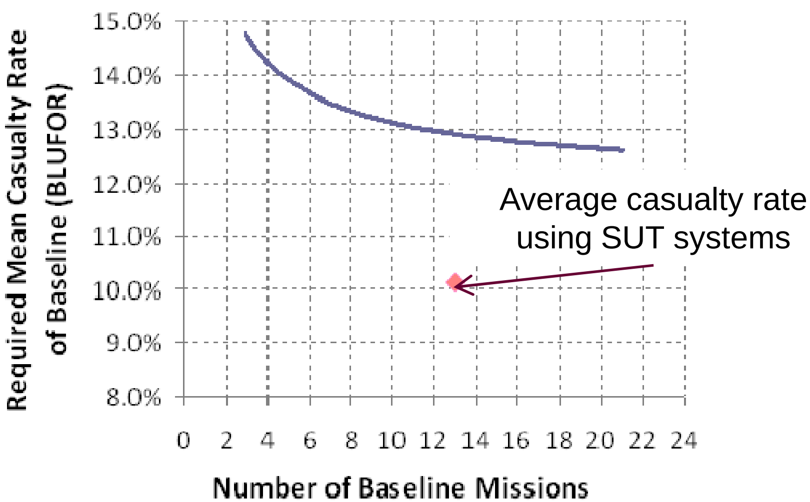

- Given the starting strength ratio of 2:1, it is unlikely that a 63% mission success rate will be observed
- Modify test structure to use a lower starting strength ratio

# Casualty Rates

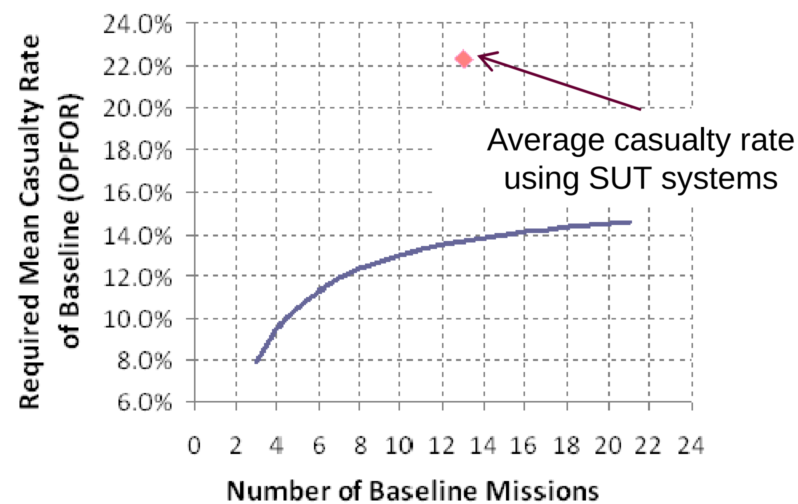
- Comparative evaluation using t-test at 90% confidence level (Sheskin 2004)

$$t_{obt} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

- Assume variability is the same for both baseline and SUT unit




- Given an expected 13 baseline missions to be conducted:
  - Minimum required BLUFOR rate is 12.9%
  - Maximum required OPFOR rate is 13.7%




- Typical observed BLUFOR and OPFOR rates are around 10% and 25% respectively
- Possible to observe positive impact of SUT on BLUFOR rate, but highly unlikely for OPFOR rate

# BLUFOR Fratricide Rate

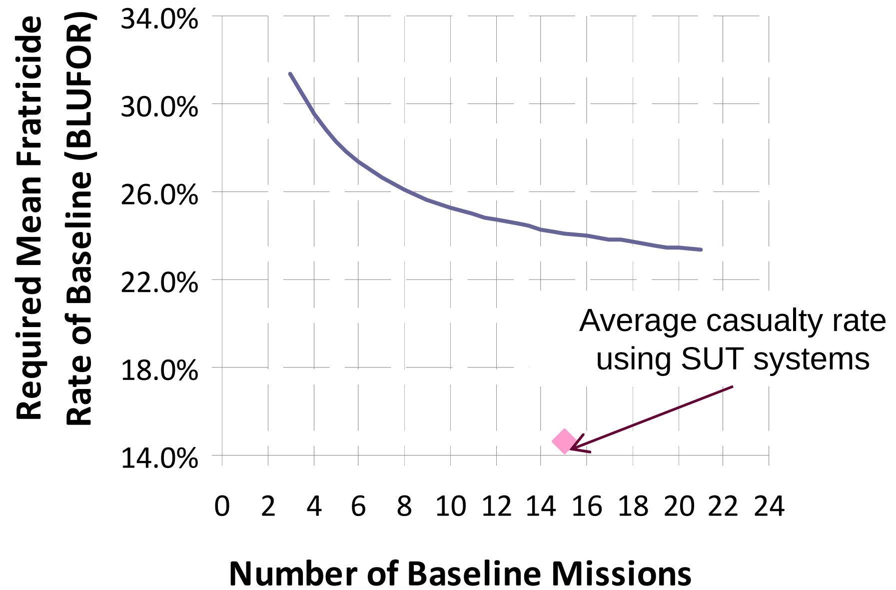
- Comparative evaluation using t-test at 90% confidence level (Sheskin 2004)

$$t_{df} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- Assume variability is the same for both baseline and SUT unit

- Observed BLUFOR fratricides rates are around 13% (Gadsden & Outteridge 2006)
- Highly unlikely to observe significant performance difference between the two units

- Given an expected 13 baseline missions to be conducted, minimum required BLUFOR fratricide rate is 25%



# Sensitivity Analysis

- Analysis predicated on a number of assumptions
  - Variability in performance measures is identical for the SUT and baseline unit
  - 90% confidence interval is the more appropriate confidence interval for the analysis
  - Performance of SUT unit in LUT 09 is representative of future performance in subsequent OT&E

Required casualty rates and mission success metrics are consistent with observed values

Required improved performance of SUT raised concerns about being able to provide conclusive results in a comparative test

| | | Required Values for Statistical Significance in IOT&E | | | |
|---|---|---|---|---|---|
| **Metrics** | **Observed LUT 09** | **Initial Results** | **50% Variability Reduction** | **Confidence Level = 80%** | **SUT** |
| **Missions Not Accomplished** | 1 | 4-6 | N/A | 4 | -- |
| **Mission Success Rate** | 0.85 | 63.2% | N/A | 71.1% | 98.2% |
| **BLUFOR Casualty Rate** | 10.1% | 12.9% | 12.1% | 11.9% | 4.7% |
| **OPFOR Casualty Rate** | 22.3% | 13.7% | 16.2% | 16.7% | 31.0% |
| **BLUFOR Fratricide Rate** | 14.6% | 24.5% | 21.6% | 21.2% | 7.3% |

Required fratricide rate remains high

Required OPFOR casualty rate remains low

# Summary

- Using statistical inference insight may be gained about possible outcomes of comparative tests
  - Guide expectations
  - Point to areas where test may need restructuring
  - Enable a more effective utilization of resources

- For case study, it is likely that a comparative evaluation of these quantitative metrics will lead to statistically inconclusive results as performance requirements are high
  - Possible restructuring of test needed
  - Given current performance of SUT, a comparative test may not be an effective utilization of limited resources

- Extend analysis to qualitative measures of operational effectiveness which are gathered from surveys and interviews

# Utilizing Statistical Inference to Guide Expectations and Test Structuring during Operational Testing and Evaluation

Joy Brathwaite
School of Aerospace Engineering
Georgia Institute of Technology

Contact: joy.brathwaite@gatech.edu

Dr. Alton Wallace
Dr. Robert Holcomb
Institute for Defense Analyses

May 10-12, 2011

8th Annual Acquisition Research Symposium