Air Force Flight Test Center

War-Winning Capabilities … On Time, On Cost

Power and Confidence

The Overarching Question in All T&E: An Analogy from the Mathematical Court of Law – Innocent Until Proven Guilty

May 2012

Mr. Todd Remund
Dr. William Kitto
812 TSS/EN
Edwards AFB, CA 93524
todd.remund@edwards.af.mil

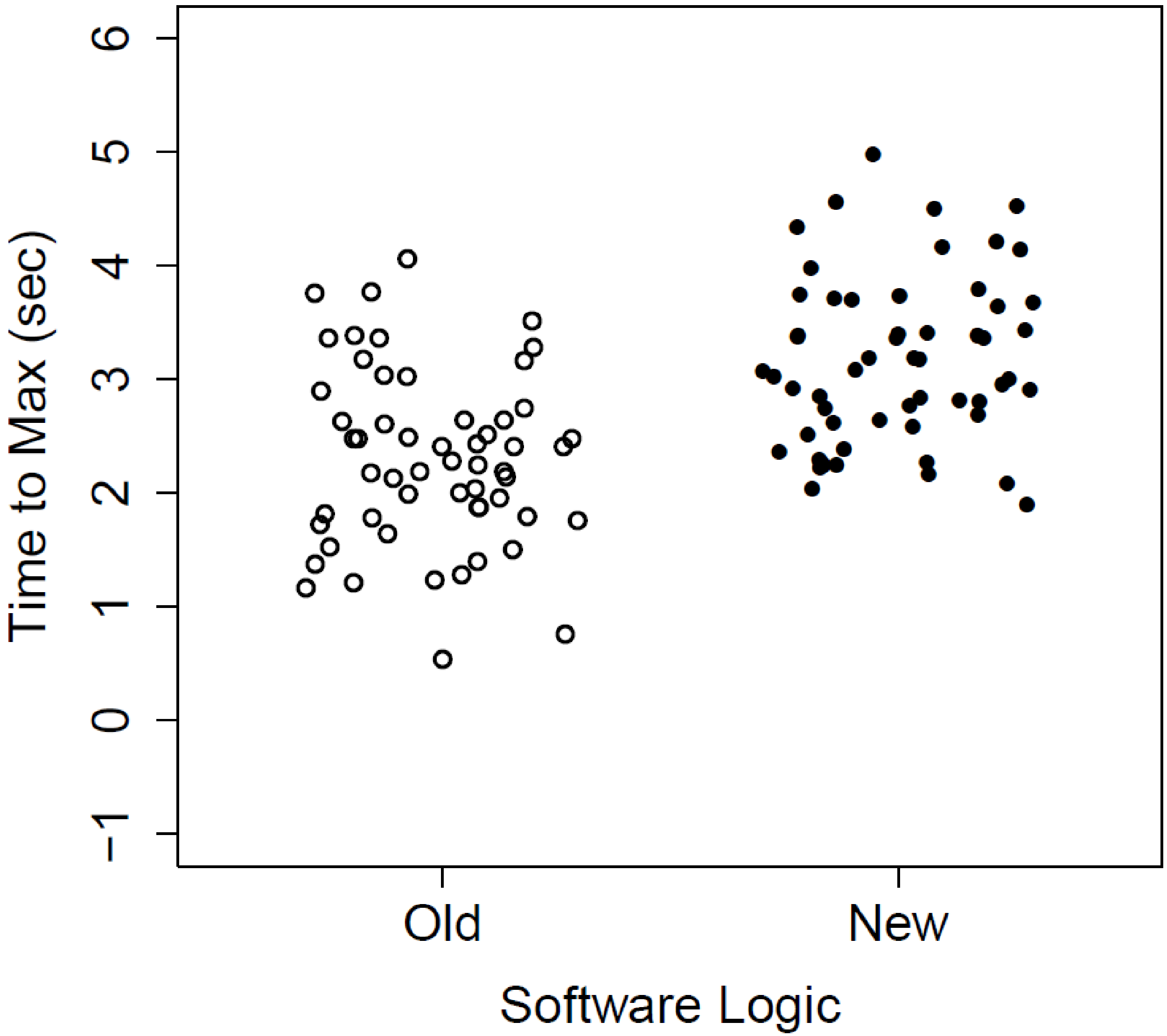
Approved for public release; distribution is unlimited.
AFFTC-PA No.: 12283

Example: Thrust Response

- Problem with engine stalls
- Software is modified to fix the problem

- Does this modification alter performance parameters?
  - Thrust Response: How long does it take for the speed to stabilize after a throttle input?
  - Compare the old mod to the new mod
  - Innocence is assumed: New performs at least as good as the old, less or equal time to max speed.
  - If guilty: New mod performs worse, more time to max speed.

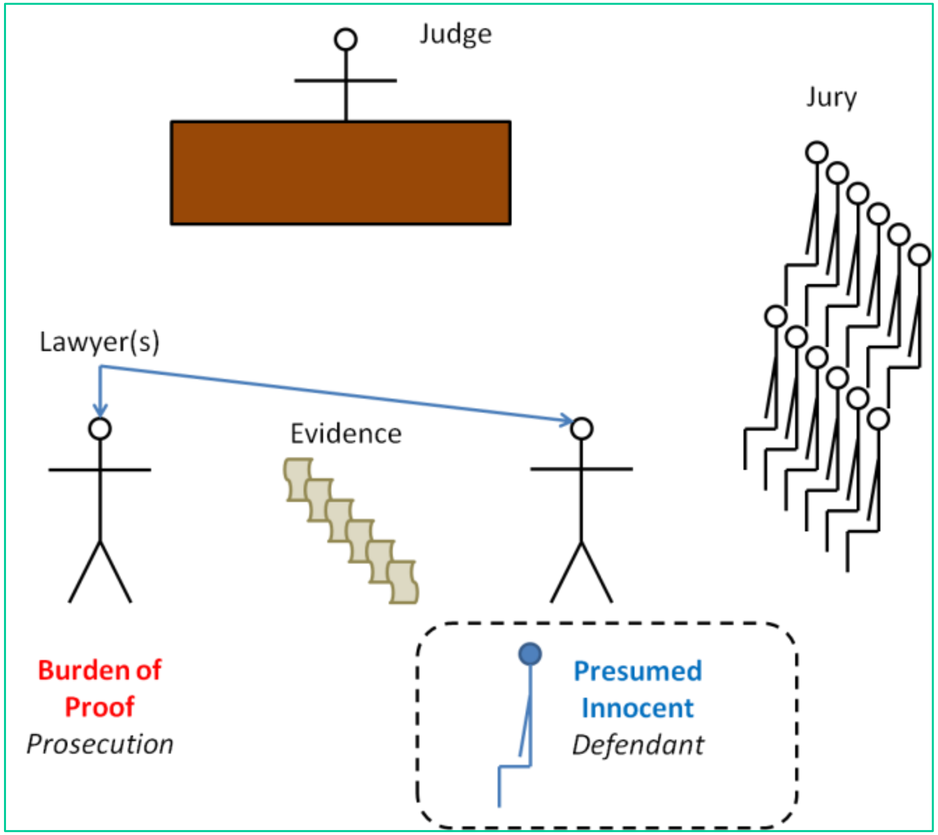
The Example



Legal and Mathematical Courts

INNOCENT UNTIL PROVEN GUILTY

Legal Court of Law



Judge

Jury

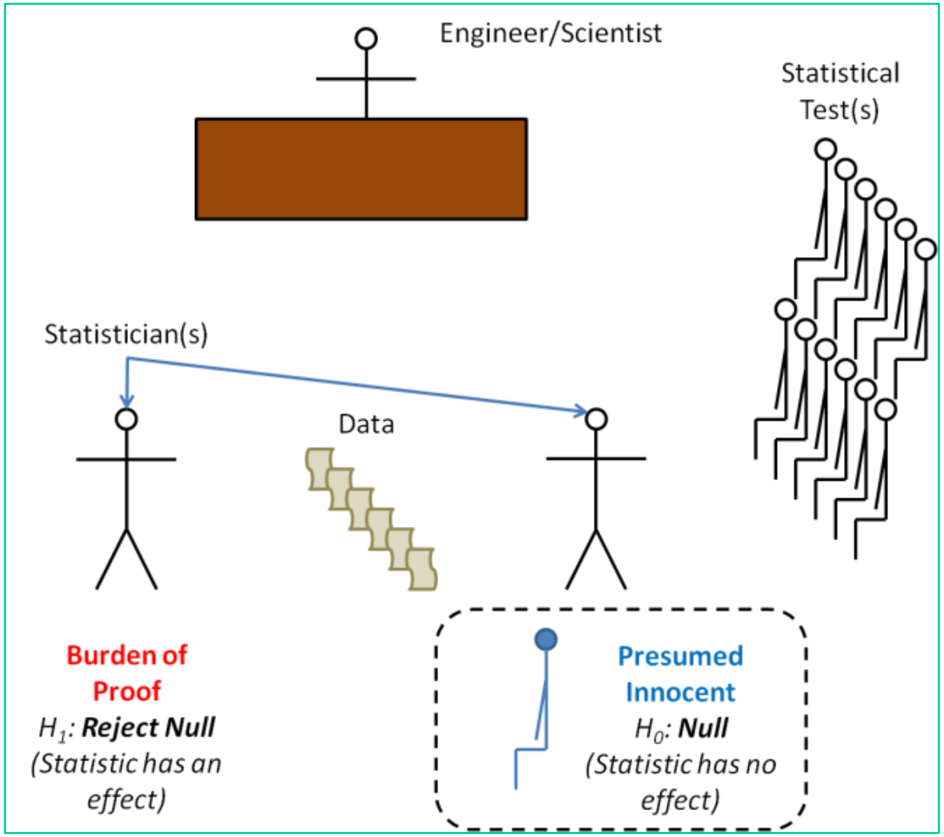
Lawyer(s)

Evidence

Burden of Proof
Prosecution

Presumed Innocent
Defendant

Mathematical Court of Law



Engineer/Scientist

Statistical Test(s)

Statistician(s)

Data

Burden of Proof
*H_1: **Reject Null***
(Statistic has an effect)

Presumed Innocent
*H_0: **Null***
(Statistic has no effect)

Risks in Judgment
(Legal Court)

- **There are two verdicts:**
  - **Not Guilty or Guilty**
- **Four possible outcomes:**

Verdict ***Truth***

- **Convict the Guilty** →
- **Release the Innocent** →
- **Convict the Innocent** → ***Type 1 Risk***

$$\Pr(T1) = \alpha$$

- **Release the Guilty** → ***Type 2 Risk***

$$\Pr(T2) = \beta$$

RISKS

M: Murder Trial (Criminal Court)
12/12 Convict (more evidence)
0/12 Support Not Guilty
E: Embezzlement (Civil Court)
6/8 Convict (less evidence)
2/8 Support Not Guilty

Decision Rule

M: Death Row for Serial Killer
E: 10 years for Embezzler

M: Innocent walks free
E: Same here

M: Death row for the innocent
E: 10 yrs for the innocent
M: Dead innocent man
E: Innocent imprisoned

M: Serial killer walks free
E: Embezzler walks free
M: Danger to public
E: More stolen $

Risks in Judgment
(Math Court)

- **There are two hypotheses:**

T: Thrust Response Trial
(x)/10 Reject H_0
(10-x)/10 Support H_0
You choose ***x***…

Decision Rule

Not Guilty $H_0 : \mu_{new} - \mu_{old} \le 0 \quad OR$

Guilty $H_1 : \mu_{new} - \mu_{old} > 0$

- **Four possible outcomes:**
  - **Declare diff > 0 when diff > 0** → ☺ T: Detect an operationally significant difference.

    $power = 1 - \beta$
  - **Don't see diff > 0 when diff < 0** → ☺ T: Do not detect a difference…none exists.

    $confidence = 1 - \alpha$

RISKS

  - **Declare diff > 0 when diff < 0** → ☹ T: Detect a difference that doesn't exist.
    Risk: Dump a good SW mod.

    Type 1 Risk
    $\Pr(T1) = \alpha$
  - **Don't see diff > 0 when diff > 0** → ☹ T: Fail to detect an operationally significant difference.
    Risk: Use a degraded SW mod.

    Type 2 Risk
    $\Pr(T2) = \beta$

Risk Probabilities

- **Assuming innocence to start, how many ‘jurors’ are necessary to be confident in rejecting innocence/no difference?**
  - **X out of 10 are necessary.**
- **There still is the chance the defendant is innocent though.**
  - **10-x out of 10 gives a probability measure**
  - **So if innocence is true, we are supposedly willing to convict them with probability of 1-x/10.**

$$\alpha = \frac{10 - x}{10} = 1 - \frac{x}{10}$$

  - **X=9, α=0.1**

Risk Probabilities

- IF the defendant is guilty…what then?
  - Under this scenario it is reasoned that 1 out of 10 operationally significant differences can slip through the court unnoticed.

$$\beta = 0.1$$

  - An operationally significant difference is determined to be at least as small as 0.5 seconds.
  - The power of seeing an operationally significant difference is

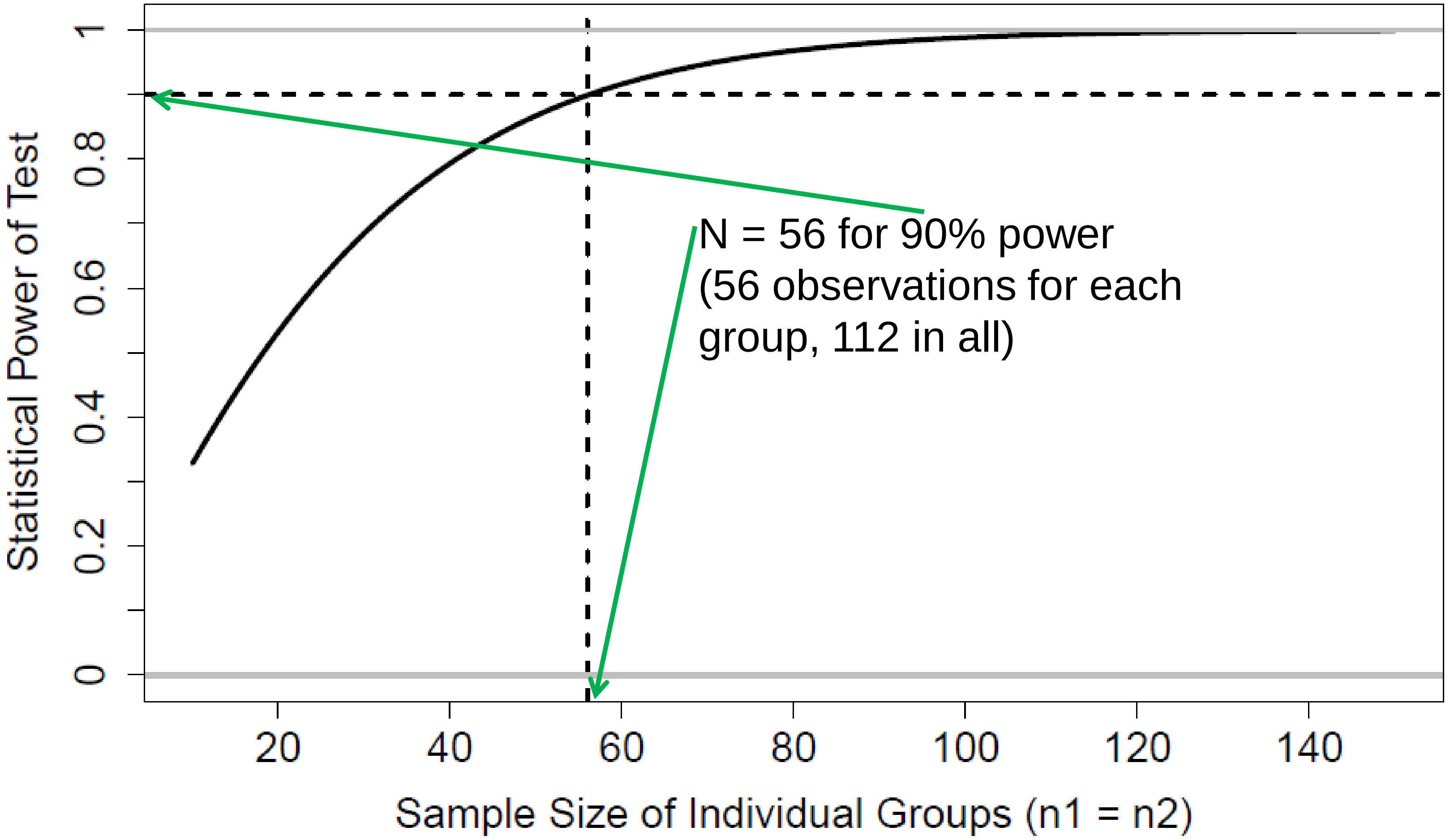
$$power = 1 - \beta = 1 - 0.1 = 0.9$$

  - Previous test data indicate that a good estimate of uncertainty measured as standard deviations is

$$\sigma = 0.9 \sec$$

How much data/evidence?

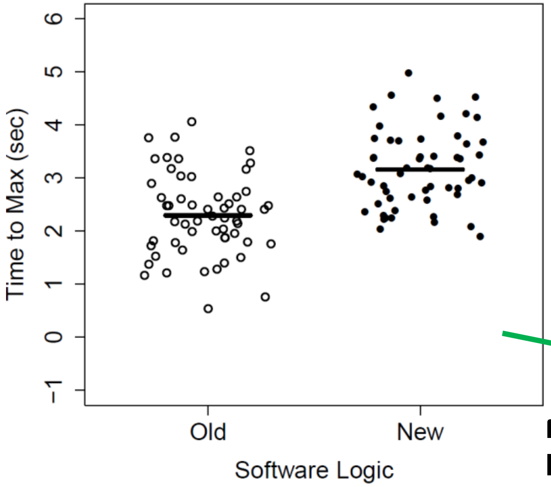
Time passes…data is gathered…the statistician now presents the data to the jury.



N = 56 for 90% power (56 observations for each group, 112 in all)

The Jury (t-test) Deliberates…

The plea is not guilty



Means are marked with black line. The overlap and spread in the data give opportunity to enter a plea of 'not guilty'. The jury is the 2-sample t-test procedure that will provide a number of jurors that vote not guilty.

It was determined that if no more than 1 'juror', out of 10, still hold to the not guilty state, then a conviction is in order. (P-value ≤ α)

"In light of the uncertainty or variance inherent in the samples, is there a significant difference between the two datasets?"

2-sample t-test:

Est. Diff. = 0.863 sec

95% CI = (0.582, 1.144)

SE($\text{Diff}_{\text{means}}$) = 0.142 sec

P-value = $3.77 * 10^{-13}$

4 out of a hypothetical 10 trillion jurors still believe the difference is zero. This equates to far less than 1 in 10 – a verdict of 'guilty' is delivered.

Based on the verdict, the engineer/judge decides to sentence the new software mod to life in prison. The new mod will not be used. Back to the drawing board…mod number 3.