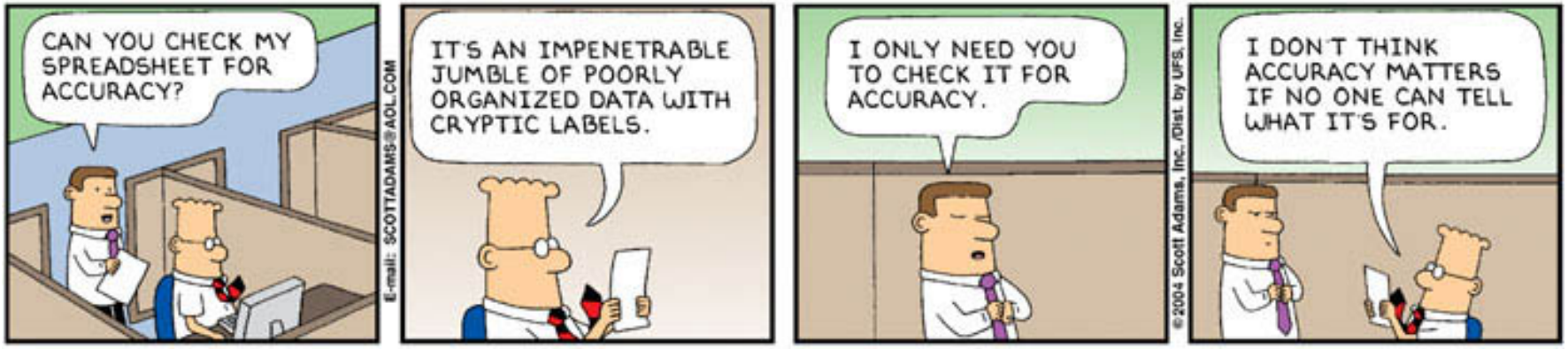
Underlying Theory & Basic Issues

Dewayne E Perry
ENS 623
Perry@ece.utexas.edu

All Too True



Validity

- In software engineering, we worry about various issues:
  - *E-Type systems*:
    - Usefulness – is it doing what is needed
    - Is it doing it in an acceptable or appropriate way
  - *S-Type programs*:
    - correctness of functionality – is it doing what it is supposed to do
    - Are the structures consistent with the way it should perform
- In empirical work, worried about similar kinds of things
  - Are we testing what we mean to test
  - Are the results due solely to our manipulations
  - Are our conclusions justified
  - What are the results applicable to
- The questions correspond to different *validity* concerns
- Concerned about the logic of evidence

Validity

- 4 primary types of validity
  - Construct Validity
  - Internal Validity
  - Statistical Conclusion
  - External Validity
- Comments
  - This organization differs somewhat from R&R
  - Each sequentially dependent on preceding

Construct Validity

- Are we measuring what we intend to measure
  - ★ Akin to the requirements problem: are we building the right system
  - ★ If we don't get this right, the rest doesn't matter
- Constructs: abstract concepts
  - ★ Theoretical constructions
  - ★ Must be operationalized in the experiment
- Necessary condition for successful experiment
- Divide construct validity into three parts:
  - ★ Intentional Validity
  - ★ Representation Validity
  - ★ Observation Validity

Construct Validity

- **Intentional Validity**
  - ★ Do the constructs we chose adequately represent what we intend to study
  - ★ Akin to the requirements problem where our intent is *fair scheduling* but out requirement is FIFO
  - ★ Are our constructs specific enough
  - ★ Do they focus in the right direction
  - ★ Eg, is it *intelligence* or *cunningness*

Construct Validity

- **Representation Validity**
  - ★ How well do the constructs or abstractions translate into observable measures
  - ★ Two primary questions:
    - ➢ Do the sub-constructs properly define the constructs
    - ➢ Do the observations properly interpret, measure or test the constructs
  - ★ 2 ways to argue for representation validity
    - ➢ Face validity
      - ✓ Claim: on the face of it, seems like a good translation
      - ✓ Very weak argument
      - ✓ Strengthened by consensus of experts
    - ➢ Content validity
      - ✓ Check the operationalization against the domain for the construct
      - ✓ The extent to which the tests measure the content of the domain being tested - ie, cover the domain
      - ✓ The more it covers the relevant areas, the more content valid
    - ➢ Both are nonquantitative judgments

Construct Validity

- Observation Validity
  - How good are the measures themselves
  - Different aspects illuminated by
    - Predictive validity
    - Criterion validity
    - Concurrent validity
    - Convergent validity
    - Discriminant validity

Construct Validity

- ★ Predictive Validity
  - Observed measure predicts what it should predict and nothing else
  - Eg, college aptitude tests are assessed for their ability to predict success in college
- ★ Criterion Validity
  - Degree to which the results of a measure agree with those of an independent standard
  - Eg, for college aptitude, GPA or successful first year
- ★ Concurrent Validity
  - The observed measure correlates highly with an established set of measures
  - Eg, shorter forms of tests against longer forms

Construct Validity

- ★ Convergent Validity
  - ➢ Observed measure correlates highly with other observable measures for the same construct
  - ➢ Utility is not that it duplicates a measure but is a new way of distinguishing a particular trait while correlating with similar measures
- ★ Discriminant Validity
  - ➢ The observable measure distinguishes between two groups that differ on the trait in question
  - ➢ Lack of divergence argues for poor discriminant validity
- ★ R&R discuss various interesting correlations on convergent and discriminant validity among various psychological tests
  - ➢ In terms of validity, reliability and stability

Internal Validity

- Are the values of the dependent variables solely the result of the manipulations of the independent variables
- Have we ruled out rival hypotheses
- Have we eliminated confounding variables
  - Participant variables
  - Experimenter variables
  - Stimulus, procedural and situational variables
  - Instrumentation
  - Nuisance variables

Internal Validity

- Never completely satisfied (Systems never error free either)
- Campbell, Stanley and Cook
  - Standardize choice of control groups
  - Try to isolate potential invalidity sources
- Confounding effects
  - Treatment effect and some other effect cannot be separated
- Confounding sources of internal invalidity
  - H: *History*
    - takes place between pre and post test
    - May contaminate post test results
  - M: *Maturation*
    - older/wiser/better between pre/post
  - I: *Instrumentation*
    - change due to test instrument
  - S: *Selection*
    - nature of participants
    - Control over assignment may have effects

Statistical Conclusion Validity

- Are the presumed causal variable X and its effect Y statistically related
  - Ie, do they covary
  - If unrelated then the one cannot be the cause of the other
- 3 questions (sequentially dependent)
  - Is the study sufficiently sensitive
  - What is the evidence that they covary
  - How strongly do they covary

External Validity

- Two positions
  - ★ The generalizability of the causal relationship beyond that studied/observed
    - ➢ Eg, do studies of very large reliable real-time systems generalize to small .COM companies
  - ★ The extent to which the results support the claims of generalizability
    - ➢ Eg, do the studies of 5ESS support the claim that they are representative of real-time ultra reliable systems

External Validity (EV)

- SWE: lab studies tend to be with students and that restricts EV and generalization
- Ecological validity: representativeness of the real world
- Efficacy vs effectiveness studies
  - ★ Former very rigorous, latter more open
  - ★ Former needs to be more concerned with EV
  - ★ Latter with internal validity
- EV: the demonstrated validity of the generalizations that the researcher intended the research to make at the outset and the validity of the generalized inferences that the researcher offers at the end

Generalizability

- Generalizability considerations
  - People
  - Researchers
  - Places, environments, settings
  - Time
  - Treatments, levels of treatments
  - Procedures, conditions and measurements
  - Technology
- Reproducibility
  - Key to generalizability is whether the study can be reproduced
  - Replication is an exact as possible repeat
    - Same procedure, different sample
    - Look for congruent results
  - Replications on a broader set of subjects under additional conditions further strengthens generalizability

Causation

- **Philosophical issues**
  - ★ Aristotle: formal, material, efficient, final causes
  - ★ Alternatives: concomitant variation, invariable sequence
  - ★ Issue of power, causal efficacy -> invariably joined
  - ★ Necessity versus invariable sequence
  - ★ Sufficiency
  - ★ Temporal precedence

Causation

- Plurality of causes
  - ★ In behavioral sciences more in terms of one of the causes than a single cause
  - ★ Experiment while holding everything else constant
  - ★ Causality reserved for experimental results
  - ★ Working definition of cause
    - ➢ Forget infinite regressive trail of reason
    - ➢ Non-philosophical working definition:
      - ✓ *A proximal antecedent agent or agency that initiates a sequence of events that re necessary and sufficient in bringing about the observed effects*
    - ➢ Proximal since it is occurs at a time near the result
    - ➢ Antecedent as it clearly precedes the effect
    - ➢ Agent is set up intentionally
    - ➢ Experimenter exercises the control lever
    - ➢ Sufficient: effect not seen in absence of treatment

Lack of Causation

- Correlation and causation
  - Correlations show a relationship
  - With path analysis, multiple regression analysis, one can begin to make causal inferences, or build causal models
  - But demonstration of causality is a logical and experimental, rather than a statistical problem
- Enabling versus causing
  - Permits but does not cause
  - Eg: marriage is primary cause of divorce - NOT
- Other
  - Not sufficient
  - May not be necessary (eg, high IQ and good living)
  - May be probabilistic
- Some unsubstantiated causal claims:
  - Drinking wine prevents arteriosclerosis
  - Eating broccoli prevents colon cancer
  - Post hoc ergo propter hoc – necessity not demonstrated

Hume's Classical Rules

- C&E must be contiguous in space and time
- C must be prior to E
- A constant union between C and E
- The same C always produces the same E and the same E never arises but from the same C
- Like Es imply like Cs
- Like Cs produce like Es
- Cs may have multiple components, ie subCs
- Some Cs are not complete in themselves

Distillation

- *Co-variation Rule* : cause is positively correlated with effect
- *Temporal Precedence Rule* : causes must precede effects
- *Internal Validity Rule* :all plausible alternative explanations must be ruled out
- *Reality* : must settle for best available evidence

Control

- Constancy of conditions
  - Maintaining extraneous conditions that might affect variables
  - Calibrating various elements in an experiment
- Control series
  - Expectation control
  - Behavior control
- Control condition
  - Of primary concern to us
  - Control + treatment

Mill's Method

- Method of agreement
  - If X then Y
    - If several instances of this and only X present
    - Then X is a sufficient condition for Y
- Method of Difference
  - If ~X then ~Y
    - IfY does not occur when X is absent
    - Then X is a necessary condition of Y
- Joint Method
  - Should lead to better, more highly justified conclusions than either method separately
- Method of Concomitant Variation
  - Relates changes in the amount or degree of change
  - Y = f(x)
    - Y is functionally related to variations in X
  - Eg, stronger treatments show larger effects

Practical Application

- Practical problems:
  - Rare to find perfect covariance, r = 1.0
  - Rare to rule out every plausible alternative hypothesis
  - Experience needed to determine adequate control conditions
- Two group design
  - Treatment: if X then Y
  - Control: if ~X then ~Y

Randomness

- Randomness is fundamental in experiments
  - ★ Quasi-experimental otherwise
  - ★ Must compensate for it otherwise
- Must have a clear view of its role
  - ★ It is "the reasoned basis for inference" in experiments
  - ★ The basis for statistical methods
- RA Fischer, The Design of Experiments, Edinburgh: Oliver & Boyd, 1935, 1949
  - ★ Credited with the invention of randomization
  - ★ Introduced the formal properties of randomization

The Lady Tasting Tea

- "A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup."
- Experiment
  - ★ 8 cups of tea
  - ★ 4 made each way
  - ★ Presented in random order
    - ➢ often determined by a random number table
  - ★ Subject knows the experimental design
  - ★ Her task is to determine two sets of 4
    - ➢ Agreeing if possible with the treatments received

The Lady Tasting Tea

- What would be expected if the Lady was "without any faculty of discrimination"
  - ★ Ie, if she made no changes in her judgments in response to changes in the order of presentation
  - ★ There are 70 possible divisions of 8 into 4
- Randomization has insured that all orderings are equally probably
- The chance of accidentally choosing the right ordering is 1/70
  - ★ Ie, probability of random ordering agreeing with the Lady's fixed judgments is 1/70 or $\binom{8}{4} = 70$
  - ★ 0.014 is the significance level for testing the null hypothesis of having no judgment

The Lady Tasting Tea

- Example serves well
  - ★ The Lady is not a sample from a population of ladies – concerns her alone
  - ★ Her eight judgments are not independent observations (rule of 4 each)
  - ★ Later cups differ from earlier ones
- Inferences are justified because the only probability distribution used in the inference is the one created by the experimenter

Fischer's argument

- Experiments do not require
  - ★ That experimental units be homogeneous
  - ★ That experimental units be a random sample from a population of unit
- It is sufficient to require that treatments be allocated at random to experimental units for valid inferences
- Probability enters the experiment only thru the random assignment of treatments