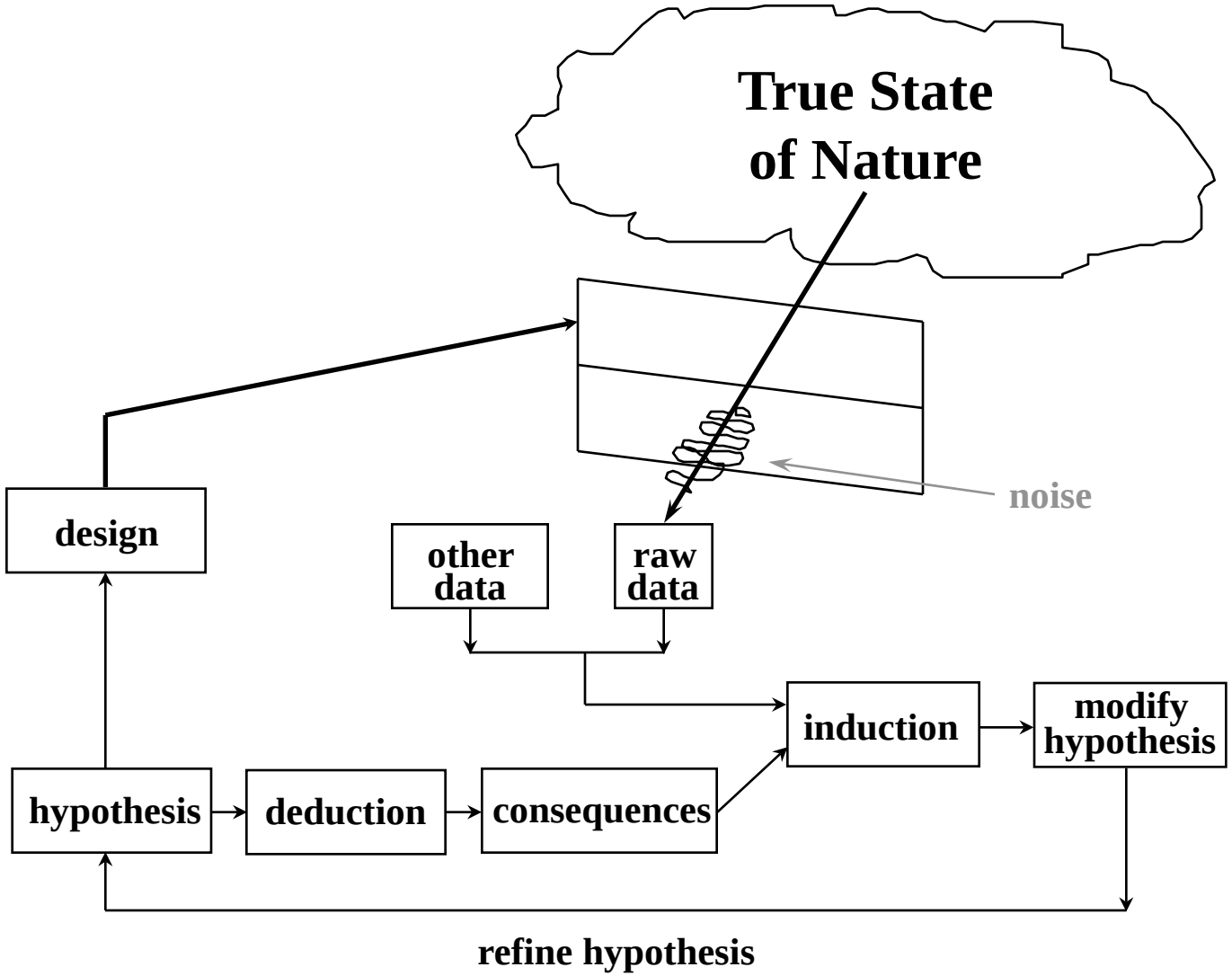
Empirical Software Engineering Studies

- Individual programmer studies have credibility due to well understood techniques from psychology and statistics.
- Large software development studies with the addition of large population social factors are not well established or credible.
- Establish a spectrum of empirical techniques that are robust to large variances from social factors present.

Reconciling Theory with Reality



Definitions

- An empirical study is a study reconciling theory and reality.
- Anecdotal and case studies are empirical studies that investigate phenomena in the context of a current theory in its real-life context.
- An experiment is an empirical study that shows a mechanism by directly manipulating the independent factors to elicit a dependent factors' predicted (from theory) responses.

Validity

- In empirical work, worried about similar kinds of evaluations that we use on our products
  - 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 with the logic of demonstrating causal connections, about the logic of evidence
- 4 primary types of validity
  - Construct Validity
  - Internal Validity
  - Statistical Conclusion
  - External Validity

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
  - 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

Construct Validity

- Concurrent Validity
  - The observed measure correlates highly with an established set of measures
  - Eg, shorter forms of tests against longer forms
- 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

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

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

Other Considerations

- Ethics
  - Typically about privacy
  - Good news: nothing life threatening
- Retrospective versus Prospective
  - Archival versus gathering data
  - Archival: no control of the quantity or quality of data
  - Gathering: various kinds of problems
- In Vivo versus In Vitro
  - In a real context versus in the lab
  - Lab conditions hard to make realistic
    - less expensive
    - Students freely available
  - Research preference for professional developers
    - Difficult to get

Analysis Methods: Quantitative vs. Qualitative

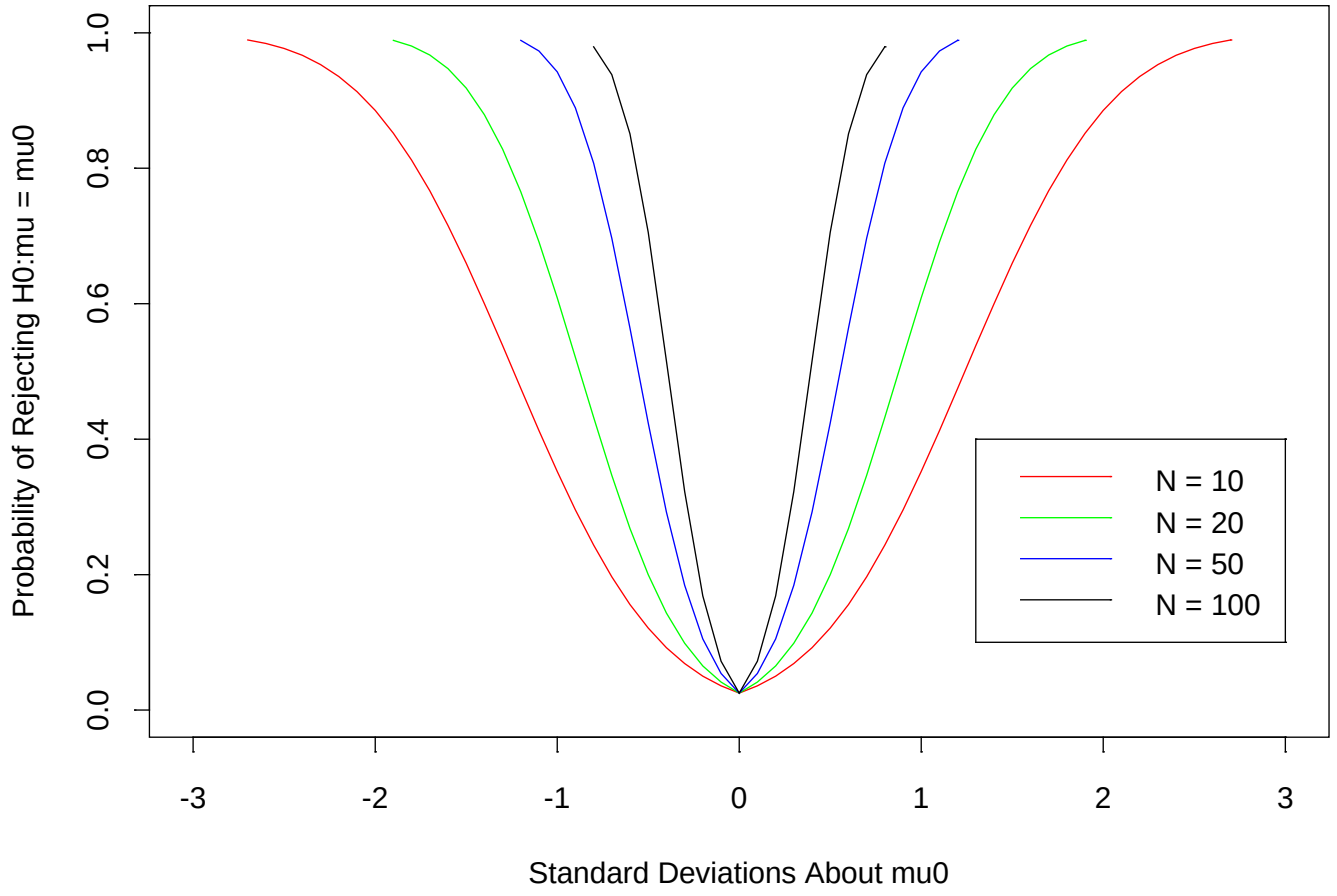
"In many instances, both forms of data are necessary--not quantitative used to test qualitative, but both used as supplements, as mutual verification and, most important for us, as different forms of data on the same subject, ..."

From Glasser & Strauss' the "Discovery of Grounded Theory: strategies for qualitative research", p. 18.

Significance & Hypothesis Testing

- Neyman-Pearson Hypothesis Testing Theory
- State H0 and H1.
- Set level of significance, a.
  - Determine which observations are consistent with H0.
  - Calculate a probability measure to reflect this set.
- Use observations to accept or reject H0.
- Errors
  - Type 1: rejecting H0 when H0 is true.
  - Type 2: failing to reject H0 when H0 is false.

Power of an Experiment



Grounded Theory (Qualitative Analysis)

- Grounded theory is a set of methods to generate theories from systematically obtained and analyzed data.
- Process iterates between collecting and analyzing data.
  - Comparative analysis
  - Theoretical sampling
  - Constructing formal theory
  - Clarifying and assessing comparative studies

Drawing Conclusions

- Fundamentals
  - credible interpretation
  - repeatability
  - understand validity limits
  - identify underlying mechanisms
  - practical significance
- Non-fundamentals
  - Quantitative Analysis
  - Qualitative Analysis
  - Identical Results
  - Correlation Studies
  - Opportunistic Studies

How do we make progress?

- Better empirical studies
  - Answers an important question
  - Establishes principles
  - Enables generating and refining hypotheses
  - Cost effective
  - Repeatable
- Credible interpretations
  - Construct, internal, and external validity
  - Test hypotheses
  - Removal of alternative explanations
  - Adequate precision
  - Available to public

NOTE: use this template in reading the papers and evaluating them for the next class