Concepts in Testing and Measurement

Raymond Nelson, MA, NCC

Stimulus --->>> Response

- Emotional reaction
- Cognitive reaction (memory/attention/effort)
- Behaviorally conditioned reaction

Concepts

- Sensitivity (deception)
- Specificity (truth)
- False positive error
- False negative error
- Base-rate (prior probability)
- Screening test
- Diagnostic test
- Successive hurdles
- Reliability
- Validity
- Base-rate (prior)
- Mean (average)
- Standard Deviation
- Proxy data
- Normal distribution
- Normal range
- Event specific exam
  - Single issue
  - Multi-facet
- Multiple issue exam
- Alfa (cutscore)
- P-value (score)
- Significance (statistical)
- Parsimony
- Scientific method
- Hypothesis / null hypothesis

Sensitivity

- The ability to notice or detect the issue of concern
- In polygraph, sensitivity refers to the ability to determine when the examinee is lying about involvement in the behavioral issue under investigation

Specificity

- Refers to the ability to determine when the issue of concern is NOT present
- In polygraph, specificity refers to the ability to determine when the examinee is being truthful about non-involvement in the behavioral issue of concern
- Determines the ability to prevent false-negative errors
- Positive results should ideally result ONLY from the specific issue of concern
- No such thing as a perfect test

False Positive Error

- Type 1 error
  - Positive is sometimes bad
    - HIV
    - Cancer
  - Value judgments are situational
    - Pregnancy
  - Science is value-neutral (factual)
- In polygraph, type 1 errors occur when a truthful person produces a significant reaction indicative of deception

False Negative Error

- Type 2 error
- In polygraph, false negative errors occur when a deceptive persons produces a truthful test result

Reliability

- Interrater reliability
  - Most important for polygraph testing
- Test re-test reliability
  - Not important for all types of tests
  - Some conditions are expected to change

Validity

- ***Construct Validity***
  - Does the test measure what we say it measures?
- ***Criterion Validity***
  - Does the test put the cases into the correct category?
- ***Incremental Validity***
  - Professionals make better decisions when they use more information
- *Convergent Validity*
  - Information agreement from different threads of investigation
- *Ecological Validity*
  - Does the experimental condition resemble real-life?
- *Internal Validity*
  - Do we make correct assumptions about cause and effect?
- External Validity
  - Does the information generalize to other situations?

Test Accuracy

- Determined by three (3) factors
  - Sensitivity
    - Determined by the cutscore or alfa
  - Specificity
    - Determined by the construct validity of the test
  - Base-rate / prior probability
    - Unknown
    - Estimated from our information about the individual and the population (and sub-group)

Accuracy in Field Practice

- Accurate polygraphs give professional interrogator a great advantage
  - Greater certainty about whom to interrogate and whom NOT to interrogate
  - Increases productiveness of interrogation
  - Decreases ethical complaints

Prior Probability (Base-Rate)

- Refers to our estimation of the probability of involvement in the issue – prior to conducting the examination
- Incidence rate
- Prior probability
- Estimated from risk factors for the individual, population, and sub-group

PCSOT Conditional Probability

- Decision accuracy set hypothetically at .90 (N=1,000)

Hypothetical Base Rate = .50 (50% or 1 in 2 offenders)

	Deceptive	*Truthful*	*Total*	
Failed Test	*450 (TP)*	*50 (FP)*	*500*	*(FPI = .10)*
Passed Test	*50 (FN)*	*450(TN)*	*500*	*(NPV = .9)*
Total	*500*	*500*	*1,000*	
	(Sens = .9)	*(Spec = .9)*		

Hypothetical Base Rate = .10 (10% or 1 in 10 offenders)

	Deceptive	*Truthful*	*Total*	
Failed Test	*90 (TP)*	*90 (FP)*	*180*	*(FPI = .5)*
Passed Test	*10 (FN)*	*810 (TN)*	*820*	*(NPV = .99)*
Total	*100*	*900*	*1,000*	
	(Sens = .9)	*(Spec = .9)*		

Mean (Average)

- How are the scores of truthful people similar
- How are the scores of deceptive people similar

Standard Deviation

- How are the scores of deceptive people different
- How are the scores of truthful people different

Proxy Data

Tests are used to measure things that cannot be measured physically.

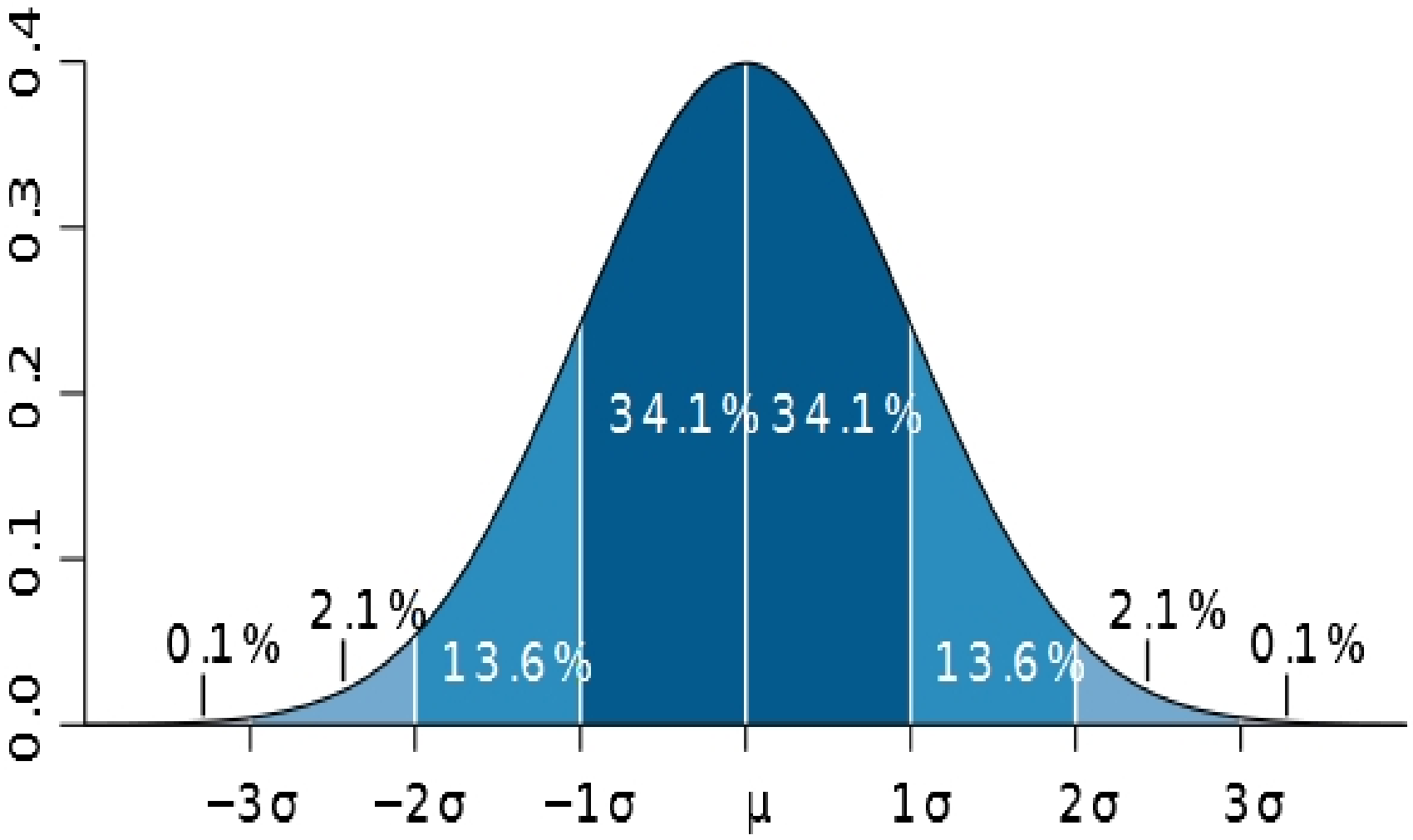
Tests work by evaluating proxy features that are correlated with the thing of interest at statisically significant levels.

Combine the proxies in an optimal statistical or structural model (recipe)

Normal Range

- Area defined by 2 standard deviations above and below the mean (average) in a normal distribution of data

Gaussian (normal) Distribution



Gaussian-Gaussian Model
Equivariance Gaussian Model
Barland (1985)

Diagnostic Exams

- Any test conducted in response to a known problem
  - Known incident
  - Known allegation
- ***Always*** a single issue exam
- Event-specific
  - Single issue
  - Multi-facet (single issue)
- Fail one = fail all

Screening Exams

- Any test conducted in the absence of a known issue, known problem, or known allegation
- Often multiple issues (mixed issues)
  - Conceivable that an examinee could be involved in one or more issues an not others
  - Fail one/fail all rule does not apply
  - No split calls
    - NO/INC if anything is SR
- Sometimes single issue
  - “Screening” is not defined by the number of issues

Successive Hurdles

- "Medical Model"
  - Medicine
  - Psychology / neuropsychology
  - Polygraph
- Strategic use of Screening and Diagnostic Tests
  - Screening tests optimized for sensitivity
  - Diagnostic (single issue) exams optimized for specificity

Multi-facet Exams

- Known incident/allegation (single issue)
- Multiple questions that describe several possible levels or roles of involvement in a single known incident/allegation
- MGQT/Investigative techniques
- Sensitivity to deception is equivalent to that of the ZCT
  - No evidence of superior sensitivity
- Specificity to truthfulness is significantly weaker than the ZCT
  - False positive error rate is much higher than the ZCT
  - Inconclusive rate is much higher than the ZCT
    - Especially for truthful persons

Alpha

- Tolerance for error
- Established before conducting the test
  - How sure will you need to be about the decision?
- A matter of both ***science*** and ***policy***
- Expressed in the form of a decimal value
  - .05
  - .01
  - .1

P-value

- Probability value
- Probability of error
- Probability of a false positive error
  - Probability that the test score and result was produced by a truthful person
- Probability of a false negative error
  - Probability that the test score and result was produced by a deceptive person
- Expressed as a decimal

Significance

- In science, "***significant***" always means "***statistically significant***"
- Expressed as a probability of error
  - P-value
- A result is significant (statistically significant) when the p-value (probability of error) is less than alpha (tolerance for error)
  - p <= a

Parsimony

- A theory should account for the greatest range of phenomena with the simplest explanation
- When one or more explanations are satisfactory
  - The simplest explanation is the best
- Explanations and hypothesis require proof in the form of evidence from scientific experiments
  - More complex explanations require more evidence
  - Expert opinion alone is not satisfactory

Scientific Method

- Hypothesis testing
  - Design an experiment to dis-prove the null-hypothesis
- Calculate the probability of error
- Publish
  - Invite criticism
  - Replication
- Standards and controls for operation
- General acceptance by the scientific community

Hypothesis Testing

- Hypothesis
  - There is a difference (statistically significant difference) in the scores of truthful and deceptive persons
- Null-hypothesis
  - There is no difference
- Design an experiment to demonstrate that there is no difference
  - Discard the hypothesis when no difference is found
- Publish evidence in support of the hypothesis if a difference is found
  - File-drawer problem

- “The six most questionable word used to formulate the justification for a conclusion by any forensic analyst are 'BASED ON MY TRAINING AND EXPERIENCE...' Training and experience in the absence of demonstrative evidence mean little to me. A reputable examiner should be able to show the decision maker – the prosecutor, the defense attorney, the judge and the jury – the basis for a conclusion which is understandable and can be justified by data or images. If the examiner resort [only] to the 'trust me, I know what I am doing logic,' a red flag should immediately go up: DON'T TRUST HIM!”

- Joseph Bono, MA
- President,
- American Academy of Forensic Science
- 2010, Presidents Message
- Academy News – September 201010issue 5