Some Problems and Solutions in the Experimental Science of Technology: The Proper Use and Reporting of Statistics in Computational Intelligence, with an Experimental Design from Computational Ethnomusicology

Mehmet Vurkaç
Portland State University

Follow this and additional works at: https://pdxscholar.library.pdx.edu/systems_science_seminar_series

Part of the Design of Experiments and Sample Surveys Commons, Statistical Methodology Commons, and the Statistical Models Commons

Let us know how access to this document benefits you.

Recommended Citation

This Book is brought to you for free and open access. It has been accepted for inclusion in Systems Science Friday Noon Seminar Series by an authorized administrator of PDXScholar. For more information, please contact pdxscholar@pdx.edu.
Some Problems & Solutions in the Experimental Science of Technology

The Proper Use and Reporting of Statistics in Computational Intelligence, with an experimental design from Computational Ethnomusicology

Systems Science Seminar Series
Feb. 25, 2011

Mehmet Vurkaç
PhD Candidate, Electrical & Computer Engineering, PSU
Assistant Professor, Electrical Engineering & Renewable Energy, OIT
Outline

• Statistics, the Scientific Method & Critical Thinking
• Misuse of Statistical Techniques
• Statistical Significance & Statistical Power
• Problems with Statistical Significance
• What To Do?
• Cross-Validation and Related Techniques
• Dissertation Research, Data & Experimental Design
Why is Statistics Important?

- The science of Science
- Critical thinking
- Social responsibility

“Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.” H. G. Wells

Mehmet Vurkaç
Critical Thinking

• Cognitive Psychology
• Philosophy
• Quantitative Literacy
• Information Literacy
• Cultural & Intercultural Competence
The Scientific Method

• Three components from ancient Greeks, Indians, Arabs, and late-Medieval/Renaissance Europe
  
  • Logic (resolution & composition)
  • Experimentation (measurement & repetition)
  • Theory (Greek & Arabian works)
The Scientific Method

• Early Version
  • Observation
  • Hypothesis
  • Testing
  • Reformulation or Conclusion
The Scientific Method

• The Modern Scientific Method
  • Accuracy
  • Objectivity
  • Skepticism
  • Open-mindedness
Parsimony (Skepticism) and Goodness-of-Fit

- Occam’s Razor
- Laplace’s principle of insufficient reason
- Einstein
- Newton’s position on hypotheses
- Lendaris/Stanley conjecture
The Scientific Method

- Additional Elements for Reliable Experimentation
  - Randomization & Blocking
  - Bootstrapping
  - Double-Blinding
  - Factorial Design
Misuse of Statistical Techniques in Science, Medicine and Technology

- Hastie/Tibshirani/Friedman (2011) ‘The Elements of Statistical Learning’
- Siegfried (2010) *Science News*
- Zucchini (2000) *Journal of Mathematical Psychology*
- Forster (2000) *Journal of Mathematical Psychology*
- Salzberg (1997) *Data Mining and Knowledge Discovery*
- Prechelt (1996) *Neural Networks*
- Flexer (1996) *Cybernetics and Systems*
Misuse of Statistical Techniques in Model Evaluation

- Salzberg (1997): nonexistent patterns
- Prechelt (1996)
  - 200 NN papers
  - 29% not on real-world problems
  - Only 8% with more than one alt. hypothesis
- Flexer (1996)
  - Only 3 out of 43 leading-journal NN papers used a holdout set.
- Hastie, Tibshirani, Friendman: cross-validation errors in top-rank journals
- Holte (1993): significance by accident (UCI repository)
- Ziliak (2008): 80% equate st.sig. with importance
Multipliciity/Bonferroni Example

**Design:**
- 14 algorithms on 11 data sets
- Those 154 combinations compared to a default classifier
- Two-tailed paired $t$ test with $p < 0.05$

**Problem:**
- At least 99.96% chance of incorrectly claiming statistical significance
Demo Example: the Math

154 chances to be significant.
Expected number of significant results = 154 * 0.05 = 7.7

\[ \text{Alpha}^* = \Pr(\text{finding at least one difference} \mid \text{there is no difference}) \]
\[ (1 - \text{Alpha}^*) = \Pr(\text{right conclusion per experiment}) \]
\[ (1 - \text{Alpha}^*)^n = \Pr(\text{making no mistakes}) \]

Real alpha = 1 – \[(1 – \text{Alpha}^*)^n \] = 0.0003
What’s Involved and What Can Be Done

• Hypothesis Testing?
• Statistical Significance, Statistical Power & Conf. Int.
• Meta-Significance?
• Cross-Validation, the Jackknife & the Bootstrap
• AIC, BIC, TIC, NIC, etc. (information criteria)
• Minimum Description Length (MDL)
• The Bayesian framework
Statistical Significance

• Hypothesis Testing
  • Do different treatments produce different outcomes?
  • Not feasible to study entire populations.
  • Sampling introduces uncertainty.
  • Need a measure of how much to trust results.

• Type-1 Error: no underlying difference, but observed
  • The likelihood of type-I errors is the p value. (reported)
  • $\alpha$ threshold must be set in advance!
Statistical Power

• Type-2 Error: difference, but not observed.

• P(Type-2 Error) ≡ β

• Typically, $\beta \leq 20$, an 80% chance of detecting a stated magnitude of difference (effect size).

• $(1 - \beta)$ is called statistical power. (controllable)

• Out of 86 clinical studies
  • 5 described power/sample size
  • 59 reported not-significant results
  • 21 of those lacked power to detect even large effects
  • In 57 studies, sample sizes $\sim 15\%$ of necessary power.

Mehmet Vurkaç
Significance & Power

• Ideal statement of the type:

  “There is at least an 80% likelihood that, had there been a 30% difference between groups, we would have found that difference with a value of \( p \) of less than 0.05.”

• Online and other-software calculators exist.

• Find power, given sample size, \( \alpha \) and effect size.

• Find sample size, given desired power, \( \alpha \) and effect size.
Meta-Significance & Other Problems

• Is statistical significance itself statistically significant?
• The standard 0.05 and 0.01 thresholds are arbitrary.
• Not the same as practical significance.
• Publication bias
• Encourages dismissal of observed differences in favor of the null.
• Regression to the mean (Tversky/Kahneman, 1971)
• Using a single $p$ value from a single study is irrational.
• If not significant, maybe study wasn’t powerful enough to find a small effect.
What To Do?

Even more important to understand what we’re doing and what it means.

• Correct methodology
  • Choice of Tests: ANOVA, Wilcoxon, …
  • Design: Cross-Validation, Bootstrap, …
  • Selection Criteria: penalty schemes (AIC, BIC, …)

• Sufficient data

• Checking assumptions against requirements

• Careful interpretation

• Suspension of judgment when appropriate
START

Not Interval Data | Interval Data

Nominal Data | Ordinal Data

Relate | Compare

Coefficient of Contingency or Phi Coefficient | Chi-squared

Spearman Correlation | Spearman Correlation

Dependent | Independent

2 groups | > 2 groups

Wilcoxon | Friedman ANOVA

2 groups | > 2 groups

Mann-Whitney | Kruskal-Wallis

Dependent | Independent

2 groups | > 2 groups

Related-Samples t test | Repeated-Measures ANOVA

Independent-Samples t test | ANOVA

Not Normal | Normal

Compare

Not Normal | Normal

≠ SD | = SD

Dependent | Independent

2 groups | > 2 groups

Mann-Whitney | Kruskal-Wallis

2 groups | > 2 groups

Related-Samples t test | Repeated-Measures ANOVA

Independent-Samples t test | ANOVA
Cross-Validation

• What is it?
• What types are there?
• What are related techniques and equivalencies?
• What are the alternatives?
Cross-Validation: What is it?

The use of separate data sets for training, tuning and assessment.
Cross-Validation: What types are there?

• Holdout (basic)
• Multifold \((k\)-fold, Geisser, 1975)\)
• Leave-One-Out (LOO)
Cross-Validation: Related techniques

• The Bootstrap
• The Jackknife
Cross-Validation: Equivalences & Performance

• Holdout → unbiased estimate of generalization performance.

• AIC, LOO & Bootstrap → asymptotically equivalent, except
  • LOO degrades as $n$ increases.
  • LOO overfits in model selection.

• $k$-fold Cross-Validation superior to Holdout & LOO.
• 10-fold is better than any Bootstrap, but Stratified is best.
• Use lower $k$ with plentiful data; higher $k$ with few data.
• BIC $>$ AIC for model selection when data plentiful.
Alternatives: Penalty Schemes

• AIC (an information criterion, or Akaike inf. criterion)
• BIC
• others (Takeuchi’s TIC, etc.)
Alternatives: Penalty Schemes

• AIC
• BIC (Bayes information criterion, or Schwartz inf. criterion)
• others (Takeuchi, et al.)

The Bayesian Framework is not discussed here due to time constraints.
My Research

• Fields:
  • Computational Intelligence (Neural Networks)
  • Information Theory (RA)
  • Music Information Research (Computational Ethnomusicology)

• Populations:
  • 65536 binary attack-point rhythm vectors
  • The space of all MLPs and all prestructured MLPs
  • All RA-derived mathematical models partido-alto clave direction

• Variables of Interest:
  • Generalization performance on holdout data as measured by GGR
  • Explanatory power of RA models tempered by penalty factors
  • A random selection of vectors for representativeness & stat. power
  • Selection of human experts and non-experts

Mehmet Vurkaç
My Research

• Description of Samples:
  • One-hidden-layer fully connected MLPs
  • One-hidden-layer prestructured MLPs, selected according to OCCAM3 searches
  • Models of rhythm data selected according to heuristics and RA decision criteria (AIC, BIC, etc.)
  • RNG-ordering of vectors (traditional patterns added if missing)
  • 4 out of 7 local mid-level human experts on partido-alto clave direction for the “ceiling” benchmark
  • Self-selected convenience sample of available “clueless” human testers for the “floor” benchmark

• Description of Inference(s):
  Based on factorial design, with batches of different random-number seeds:
  • Generalization performance of fully connected neural networks
  • Generalization performance of prestructured neural networks
  • Generalization performance of RA models
  • Generalization performance of mid-level experts (as guideline)
  • Generalization performance of clueless testers (as guideline)
Factors in Neural-Net Experimentation

• Output encoding

• Training/Test regimes

• Network-design parameters
  • Learning rate (step size) & momentum
  • Epoch size
  • Derivative offset
  • Number of hidden layers
  • Number of processing elements per hidden layer
  • Learning schedules

• Spatial Crosstalk (separate concepts in one network)

• Decision-making instruments

• Early-stopping

• Bumping and jogging network weights

Mehmet Vurkaç
My Data

• $2^{16} = 65536$ possible input patterns (idealized rhythms)

• Three musical-teaching contexts (teacher types) for classification
  • Lenient
  • Firm
  • Strict

• Four output classes
  • Incoherent
  • Forward
  • Reverse
  • Neutral

• Three membership degrees in each output class
  • Strong
  • Average
  • Weak

Mehmet Vurkaç
My Data

• “Firm” teacher context selected.

• Data stabilized July 4, 2010, with 10,811 vectors.

• Two types of holdout sets created
  
  • Standard (random) holdout
    
    • Design data: 8651
      
      • Strong: 4745
      
      • Average: 2010
      
      • Weak: 1896
    
    • Holdout data: 2160
  
  • Weak holdout
    
    • Design data: 8442
      
      • Strong: 5931
      
      • Average: 2511
    
    • Holdout data: 2369
Experimental Design

• Five-fold stratified cross-validation is the best approach to performance estimation.

• Minimum training-set size for good generalization (Haykin):
  • \( N = O\left(\frac{W}{\varepsilon}\right) \)
  • 20 hidden elements → \( O(4413) \) examples
  • 40 hidden elements → \( O(8813) \) examples
  • These numbers are beyond the notion of parsimony.

• NeuralWare NeuralWorks manual gives higher numbers for my set size.
Eight Choices or Actions

• Output encoding
• Training classes
• Testing classes
• Training membership degrees
• Testing membership degrees
• Thresholding (NN) or Fitting (RA)
• Controls
  • Randomization testing for NNs
  • Random “structure” for RA models
  • Human floor and ceiling
• Random-Number Initialization
References


http://www.ted.com/talks/lee_smolin_on_science_and_democracy.html
Questions & Discussion