2-25-2011

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

Recommended Citation

Vurkaç, Mehmet, "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" (2011). *Systems Science Friday Noon Seminar Series*. 59.
https://pdxscholar.library.pdx.edu/systems_science_seminar_series/59

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

Portland State UNIVERSITY

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*
- Ziliak/McCloskey (2008) ‘The Cult of Statistical Significance’
- Ioannidis (2005) *PLoS Medicine*
- Miller (2004) *The Journal of Systems and Software*
- 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*
- Holte (1993) *Machine Learning*

Misuse of Statistical Techniques in Model Evaluation

- Miller (2004), Zucchini (2000) & Salzberg (1997): multiplicity effect
- 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

Multiplicity/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

Alpha* = P(finding at least one difference | there is no difference)
(1 – Alpha*) = P(right conclusion per experiment)
(1 – Alpha*)^n = P(making no mistakes)

Real alpha = 1–[(1 – 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)
  - α threshold must be set in advance!

Statistical Power

- Type-2 Error: difference, but not observed.
- P(Type-2 Error) $\equiv \beta$
- 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 ~ 15% of necessary power.

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, α and effect size.
- Find sample size, given desired power , α and effect size.

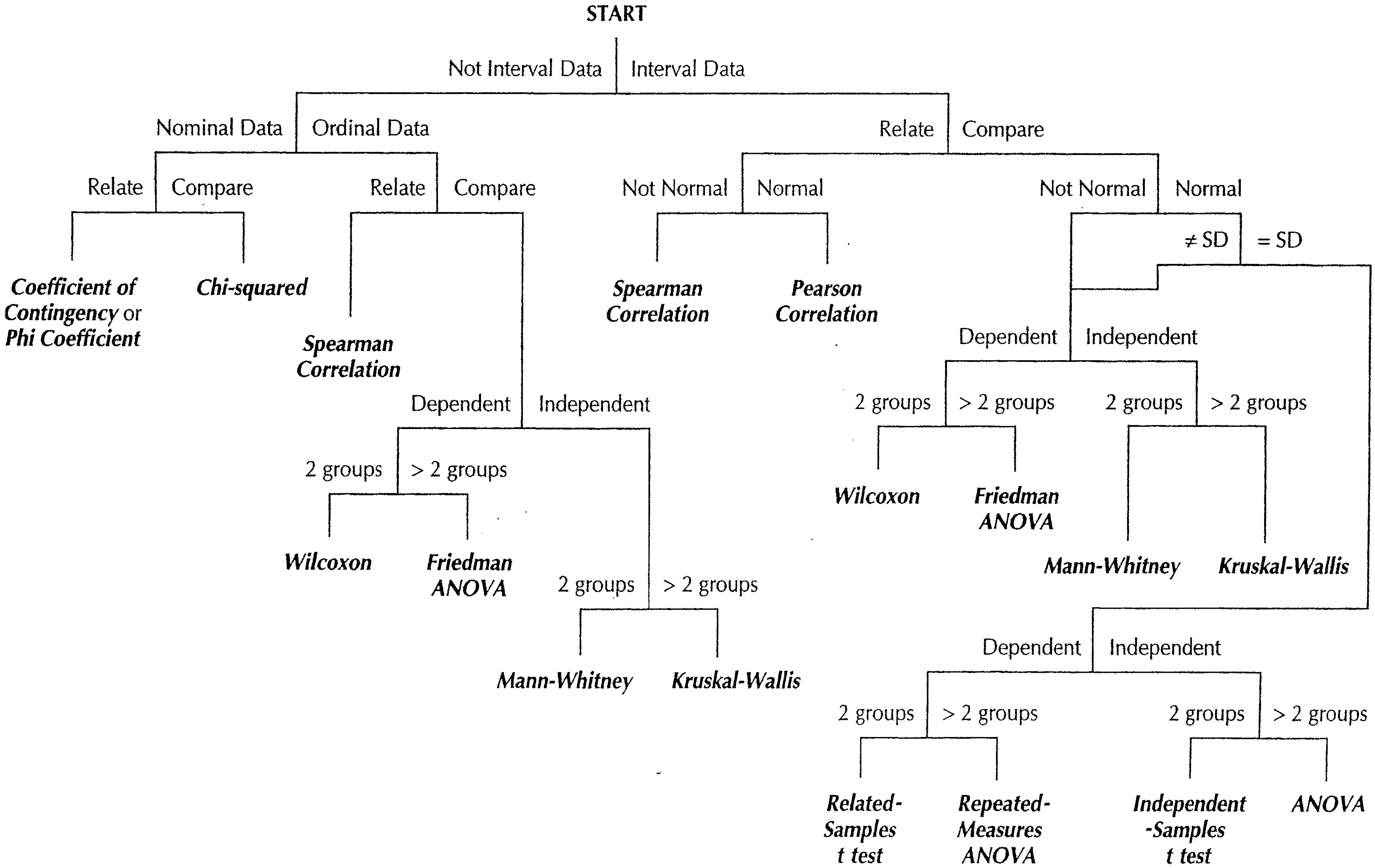
Meta-Significance & Other Problems

- 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
  - Nominal Data
    - Relate
      - ***Coefficient of Contingency*** or ***Phi Coefficient***
    - Compare
      - ***Chi-squared***
  - Ordinal Data
    - Relate
      - ***Spearman Correlation***
    - Compare
      - Dependent
        - 2 groups
          - ***Wilcoxon***
        - > 2 groups
          - ***Friedman ANOVA***
      - Independent
        - 2 groups
          - ***Mann-Whitney***
        - > 2 groups
          - ***Kruskal-Wallis***
- Interval Data
  - Relate
    - Not Normal
      - ***Spearman Correlation***
    - Normal
      - ***Pearson Correlation***
  - Compare
    - Not Normal
      - Dependent
        - 2 groups
          - ***Wilcoxon***
        - > 2 groups
          - ***Friedman ANOVA***
      - Independent
        - 2 groups
          - ***Mann-Whitney***
        - > 2 groups
          - ***Kruskal-Wallis***
    - Normal
      - ≠ SD
      - = SD
        - Dependent
          - 2 groups
            - ***Related-Samples t test***
          - > 2 groups
            - ***Repeated-Measures ANOVA***
        - Independent
          - 2 groups
            - ***Independent-Samples t test***
          - > 2 groups
            - ***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

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

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

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):
  - $N = O(W/\varepsilon)$
    - 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

Hastie, T., Tibshirani, R., and Friedman, J. H., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York, NY: Springer Series in Statistics, 2009.

McClave, J. T., and Sincich, T., *Statistics*, Upper Saddle River, NJ: Prentice Hall, 2003.

Rice, E. F., Jr., and Grafton, A., *The Foundations of Early Modern Europe, 1460–1559*, New York, NY: W. W. Norton & Company, 1994.

Baron, R. A., and Kalsher, M. J., *Essentials of Psychology*, Needham, MA: Allyn & Bacon, A Pearson Education Company, 2002.

Box, G. E. P., Hunter, W. P., and Hunter, J. S., *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*, New York, NY: John Wiley & Sons, 1978.

Kapur, J. N., and Kesavan, H. K., *Entropy Optimization Principles with Applications*, Boston, MA: Academic Press, 1992.

Lendaris, G. G., and Stanley, G. L., "Self-Organization: Meaning and Means," *Inf. Syst. Sci.; Proc. 2nd Congress on Inf. Syst. Sci.*, Baltimore, MD: Spartan Books, 1965.

Lincoln, Y. S., and Guba, E. G., *Naturalistic Inquiry*, Beverly Hills, CA: Sage Publications, Inc., 1985.

Forster, M. R., "Key Concepts in Model Selection: Performance and Generalizability," *Journal of Mathematical Psychology*, 44, pp. 205–231, 2000.

Flexer, A., "Statistical Evaluation of Neural Network Experiments: Minimum Requirements and Current Practice," In R. Trappl, ed., *Cybernetics and Systems '96: Proc. 13th European Meeting on Cybernetics and Systems Res.*, pp. 1005–1008, Austrian Society for Cybernetic Studies, 1996.

Salzberg, S. L., “On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach,” *Data Mining and Knowledge Discovery*, 1, pp. 317–327, 1997.

Holte, R., “Very simple classification rules perform well on most commonly used datasets,” *Machine Learning*, Vol. 11, No. 1, pp. 63–90, 1993.

Prechelt, L., “A quantitative study of experimental evaluations of neural network algorithms: Current research practice,” *Neural Networks*, 9, 1996.

Miller, J., “Statistical significance testing—a panacea for software technology experiments?” *The Journal of systems and Software*, 73, pp. 183–192, 2004.

Zucchini, W., “An introduction to model selection,” *Journal of Mathematical Psychology*, 44, pp. 41–61, 2000.

Akaike, H., “Information theory and an extension of the maximum likelihood principle,” B. N. Petrov and F. Csaki (eds.), *2nd International Symposium on Information Theory*, Budapest: Akademiai Kiado, pp. 267–281, 1973.

Schwarz, G., “Estimating the dimension of a model,” *Annals of Statistics*, 6, pp. 461–465, 1978.

Busemeyer, J. R., and Wang, Y.-M., “Model comparisons and model selections based on generalization test methodology,” *Journal of Mathematical Psychology*, 44, pp. 171–189, 2000.

Golden, R. M., “Statistical tests for comparing possibly misspecified and non-nested models,” *Journal of Mathematical Psychology*, 44, pp. 153–170, 2000.

Tversky, A., and Kahneman, D., “Belief in the law of small numbers,” In D. Kahneman, P. Slovic, and A. Tversky (eds.), *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, pp. 3–20, 1971.

Shao, J., “Model Selection by Cross-Validation” *Journal of the American Statistical Association*, Vol. 88, No. 422, pp. 486–494, 1993.

Kohavi, R., “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” *IJCAI 1995 (14th International Joint Conference on Artificial Intelligence*), Vol. 14, pp. 1137–1143, 1995.

Blum, A., Kalai, A., and Langford, J., “Beating the Holdout: Bounds for K-fold and Progressive Cross-Validation, ” *Proceedings of the 10th Annual Conference on Computational Learning Theory (COLT)*, 1999.

Efron, B., “Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation,” *Journal of the American Statistical Association*, Vol. 78, No. 382, pp. 316–331, 1983.

Ziliak, S. T., and McCloskey, D. N., *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, Ann Arbor, MI: The University of Michigan Press, 2008.

Zhang, P., “Model Selection Via Multifold Cross-Validation,” *The Annals of Statistics*, Vol. 21, No. 1, pp. 299–313, 1993.

Zhang, P., “On the distributional properties of model selection criteria,” *Journal of the American Statistical Association*, Vol. 87, No. 419, pp. 732–737, 1992.

Geisser, S., “The Predictive Sample Reuse Method With Applications,” *Journal of the American Statistical Association*, Vol. 70, pp. 320–328, 1975.

Stone, M., “Cross-Validation Choice and Assessment of Statistical Predictions,” *Journal of the Royal Statistical Society*, Ser. B, 36, pp. 111–147, 1974.

Stone, M., “An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike’s Criterion,” *Journal of the Royal Statistical Society*, Ser. B, 39, pp. 44–47, 1977.

Cohen, P. R. , and Jensen, D., “Overfitting Explained,” *Preliminary Papers, Sixth International Workshop on Artificial Intelligence and Statistics*, pp. 115–122, 1997.

S. Haykin, *Neural Networks – A Comprehensive Foundation (Second Edition/Low-Price Edition)*, Delhi, India: Pearson Education, Inc. (Singapore), 2004.

Cohen, P. R., *Empirical Methods for Artificial Intelligence*, Cambridge, MA: A Bradford Book, MIT Press, 1995.

Siegfried, T., “Odds Are. It’s Wrong: Science fails to face the shortcomings of statistics,” *Science News – The Magazine of the Society for Science & the Public*, Vol. 177, No. 7, p. 26, 2010.

http://www.ted.com/talks/lee_smolin_on_science_and_democracy.html

Questions & Discussion