

Chapter 9

Resampling and Bayesian Estimation

The best way to predict the future is to invent it.

—Alan Kay

Overview

- Resampling techniques
- Characteristics of bootstrapping
- Evaluation of bootstrapping
- Bayesian estimation
- Evaluation of Bayesian estimation
- Conclusion

Resampling techniques

- Techniques for *resampling*—also known as *computer-intensive methods*—are forms of internal replication
- Most recombine the cases in a data set in different ways to estimate statistical precision, with possibly fewer distributional assumptions compared with traditional statistical tests
- They are better known in the natural sciences, but they can be applied in the behavioral sciences, too

Resampling techniques

- Perhaps the best known resampling methods are the bootstrap, the jackknife, and randomization procedures
- Some of these methods work by instructing the computer to take large numbers of random samples (e.g., > 1,000) from a raw data set (e.g., nonparametric bootstrapping)
- Because of sampling with replacement:
 1. Each case can appear more than once in the same generated sample
 2. Case composition across the generated samples will vary

Resampling techniques

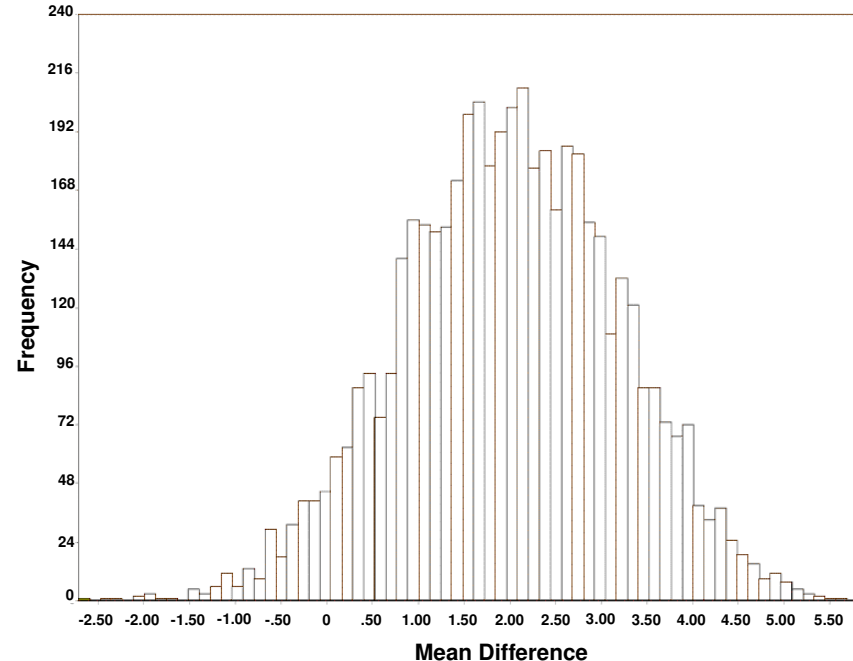
- The computer generates an estimator in each generated sample and constructs an empirical frequency distribution of that estimator across all generated samples
- The properties of the empirical frequency distribution, such as its central tendency and variability, can be used to estimate precision (e.g., construct a confidence interval)
- Recall that traditional methods for interval estimation may rely on estimates of standard error in large samples and certain distributional assumptions, such as normality

Characteristics of bootstrapping

- The bootstrap method was developed by B. Efron in the late 1970s (e.g., Diaconis & Efron, 1983) and is probably the best known and most flexible of resampling methods
- There are two general forms, parametric and nonparametric
- *Nonparametric bootstrapping* usually samples with replacement from a raw data file
- Depending on the estimator, it may make no assumptions other than that the sample distribution reflects the basic shape of the population distribution
- When repeated many times by the computer, nonparametric bootstrapping constructs an empirical sampling distribution

Characteristics of bootstrapping

- Example: Screenshot of an empirical sampling distribution of $M_1 - M_2$ across 5,000 generated samples selected at random from the data set in Table 9.1 constructed by the Bootstrap module of SimStat (<http://www.simstat.com/>)



- The standard deviation in the above distribution is 1.195, which is a bootstrapped estimate of the standard error of $M_1 - M_2$

Characteristics of bootstrapping

- *Parametric bootstrapping* allows specific assumptions about the parameters of population distributions
- Instead of sampling with replacement from an actual data set, bootstrapped samples of a specified size are drawn from a probability density function that reflects those parameters—example:

Parametric bootstrapping in Amos 5 (Arbuckle, 2003) for structural equation modeling generates random covariance matrices from the sample covariance matrix, assuming multivariate normality

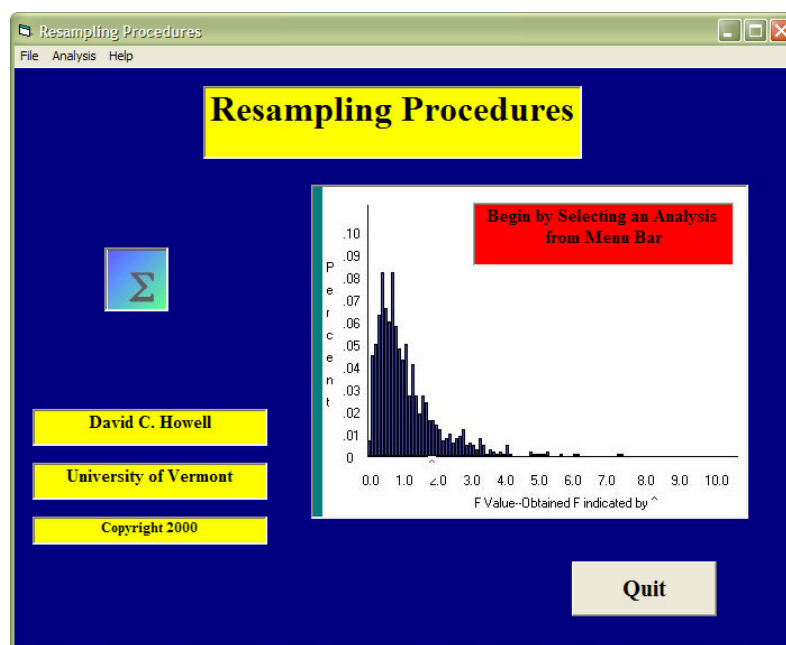
Characteristics of bootstrapping

- However, it is nonparametric bootstrapping that is better known in the behavioral sciences
- The bootstrap method can be implemented with just about any statistical method, including traditional tests such as t or F
- Bootstrapped test statistics generally have the same distributional assumptions as their traditional counterparts and are subject to the same general limitations (chap. 3)
- That is, bootstrapping does not “cure” statistical tests of their shortcomings

Evaluation of bootstrapping

- There are relatively few computer tools for bootstrapping in the social sciences, but this situation is slowly changing
- The freely available program Resampling Statistics by D. Howell is good for learning more about these methods:

<http://www.uvm.edu/~dhowell/StatPages/Resampling/Resampling.html>



Evaluation of bootstrapping

- Bootstrapping may be especially useful for estimating standard errors for statistics with complex distributions, for which there is no practical approximate method
- A potential application is to estimate standard errors of effect sizes, but computer tools for bootstrapping do not at present generally calculate effect sizes other than correlations
- Some limitations include the following:
 - ✓ Bootstrapping (and resampling in general) is not a substitute for external replication
 - ✓ Bootstrapping is not without distributional assumptions, albeit possibly fewer compared with more traditional methods

Evaluation of bootstrapping

- Some limitations include:
 - ✓ The “population” in nonparametric bootstrapping is merely the researcher’s sample
 - ✓ If the researcher’s sample is small, unrepresentative, or the observations are not independent, resampling from it can magnify the effects of these features (see Rodgers, 1999)
 - ✓ Bootstrap analyses are probably biased in small samples (just as they are in other methods)—that is, bootstrapping is not a “cure” for small sample sizes

Evaluation of bootstrapping

- See Efron and Tibshirani (1993) for more information about the bootstrap method
- Lunneborg (2001) and B. Thompson (1993) describe the use of the bootstrap method to estimate standard errors and confidence intervals in behavioral studies

Bayesian estimation

- Traditional statistical tests estimate the probability of the data under a point null hypothesis, or $p(D | H_0)$
- The percentage associated with a traditional confidence interval (e.g., 95%) is not generally interpretable as the chance that the interval contains the corresponding population parameter
- This is not what researchers really want to know—for example:
 - ✓ What is the probability of the *hypothesis*, given the data?
 - ✓ What is the probability that the true value of the parameter falls within the confidence interval?
- These kinds of questions can be addressed in a Bayesian approach

Bayesian estimation

- Bayesian methods have been around for decades, but few researchers or students in the behavioral sciences hear much about them
- Part of the reason is the excessive focus on statistical tests in education about data analysis methods
- Another is the paucity of computer tools for Bayesian analyses of behavioral data, but this is slowly changing

Bayesian estimation

- Bayesian methods have been widely used in many disciplines, such as computer science (e.g., managing computer networks, artificial intelligence) and medicine (e.g., evaluation of screening test results)—see Gatsonis et al. 2001)
- Edwards, Lindman, & Savage (1963) introduced Bayesian statistics to the behavioral sciences in the 1960s
- Some later introductions include Iversen (1984) and Pitz (1982)

Bayesian estimation

- These principles are all supported in a Bayesian approach:
 1. Not all hypotheses are equally plausible before there is evidence
 2. Not all researchers will see the same hypothesis as equally plausible prior to data collection
 3. Imprecise data have less sway on subsequent hypothesis plausibility than precise data
 4. Impact of initial differences in estimates of hypothesis plausibility become less important as there are more results
- In contrast, p values from statistical tests do not take account of H_0 plausibility (and nil hypotheses may be especially implausible)

Bayesian estimation

- In the broadest sense, Bayesian statistics can be seen as a set of methods for the orderly expression and revision of belief as new evidence is gathered (Edwards et al., 1963)
- These methods are based on Bayes's theorem:

$$p(H | D) = \frac{p(H) p(D | H)}{p(D)}$$

$p(H | D)$ is the posterior probability of the hypothesis, given the data

$p(H)$ is the prior probability of the hypothesis irrespective of the data (i.e., its plausibility before data collection)

$p(D | H)$ is the *likelihood*, the conditional probability of the data, given the hypothesis (not necessarily a traditional H_0)

$p(D)$ is the prior probability of the data irrespective of the truth of the hypothesis

Bayesian estimation

- That is, Bayes's theorem takes an initial belief about the hypothesis, $p(H)$, and combines it with information from the sample to generate an updated belief, $p(H | D)$
- It also shows us that the only way to estimate the probability of some hypothesis in light of the data is through estimation of the prior probability of each and the likelihood of the data
- Recall the *inverse probability error*—the interpretation of p values from statistical tests as though their form were $p(H_0 | D)$ instead of $p(D | H_0)$
- This is why Gigerenzer (1993) refers to this false belief as the Bayesian Id's wishful thinking error

Bayesian estimation

- How do we estimate the prior probability of some hypothesis?
- In the absence of any information, equal probabilities are assigned to all competing hypotheses
- This reflects the *principle of indifference*, also known as *agnostic priors* or *uninformative priors* (Harsanyi, 1983)

Bayesian estimation

- However, it is rare that we really have absolutely no information
- One possibility is to survey experts in the field, in this case researchers
- There are methods from cognitive psychology and computer science for obtaining consistent prior probabilities (e.g., Anderson, 1998)
- These methods are not subjective, and on the whole they are probably more objective than the logic of statistical tests (e.g., Matthews, 2000)

Bayesian estimation

- Perhaps one of the most potentially useful applications of Bayesian statistics is interval estimation in the context of collecting more and more data about a population parameter
- In traditional statistical methods, parameters are constants
- This is why in APA style parameters are usually presented in a plain font (i.e., no italics; e.g., μ , σ)

Bayesian estimation

- In a Bayesian approach, parameters are seen as random variables with their own distributions
- Variables are usually represented with italics (e.g., μ , σ)
- The distribution for a parameter summarizes the current state of knowledge about it
- The central tendency is the best single guess about the true value of the parameter, and the variability reflects the amount of uncertainty

Bayesian estimation

- The inverse of the conditional variance (i.e., the squared standard error) of the distribution for a parameter is a measure of precision
- Meta-analysis uses the same principle for an observed effect size: The estimate of precision is weighted by a function of the inverse of the conditional variance, which in part reflects sample size (chap. 8)
- Prior and posterior distributions for a parameter are typically described by a mathematical function

Bayesian estimation

- This mathematical function may correspond to a known probability density function, such as a normal distribution, a central t distribution, or a multivariate normal distribution
- Selecting the proper distribution is a matter of statistical modeling based on principles similar to those in other techniques, such as structural equation modeling
- This is not to say that this part of a Bayesian analysis is easy—it can be difficult to select the proper distribution(s)

Bayesian estimation

- As new results (i.e., parameter estimates in samples) are collected, they are weighted by their precision and combined with the prior distribution
- That is, the mean and variance of the distribution for the random parameter are updated based on new information
- Results that are more precise have a greater potential impact on the posterior distribution (new state of knowledge)
- Through this process the update of knowledge in some areas can be tracked systematically (e.g., Table 9.2)

Bayesian estimation

- A Bayesian confidence interval—also called a Bayesian *credible interval* or *highest density region*—is estimated within the posterior distribution for a parameter
- The percentage associated with a Bayesian confidence interval (e.g., 95%) is interpreted as the probability that the true value of the parameter falls within the interval
- Recall that traditional confidence intervals are not generally interpreted this way
- An exception is when there is *no* information about the parameter (i.e., all values are equally likely), but this situation is rare (see Reichardt & Gollob, 1997)

Bayesian estimation

- As mentioned, both standard meta-analysis and Bayesian statistics are methods for research synthesis
- A sensitivity analysis can be performed in either method to evaluate robustness against violation of certain assumptions
- A basic question of a sensitivity analysis in Bayesian statistics is whether the posterior results change appreciably when other reasonable probability models are specified (Gelman, Carlin, Stern, & Rubin, 1995)

Bayesian estimation

- These kinds of questions cannot generally be answered in a standard meta-analysis, but they pose no special problems in Bayesian analysis:
 - ✓ What is the probability that a treatment is beneficial?
 - ✓ How does hypothesis plausibility affect the accumulated results?
- There are ways to incorporate Bayesian methods in a meta-analysis to take account of hypothesis plausibility (e.g., Cornell & Mulrow, 1999)
- See Howard, Maxwell, and Fleming (2000) for a comparison of standard meta-analysis and Bayesian statistics for research synthesis

Evaluation of Bayesian estimation

- There are some significant hurdles to the wider use of Bayesian methods in the behavioral sciences
- The relative lack of computer tools for behavioral scientists is a problem
- Another is that many reference works for Bayesian statistics are quite technical and require familiarity with integral notation for probability distributions and estimation techniques for the parameters of different kinds of distributions

Evaluation of Bayesian estimation

- However, researchers comfortable with structural equation modeling or related types of model-fitting techniques should be able to manage the basics of Bayesian estimation
- The flexibility afforded by Bayesian statistics is worth the time to learn more about them

Conclusion

- There are many other alternatives to traditional statistical tests, including robust statistics (e.g., Wilcox, 1998) and exploratory data analysis (e.g., Tukey, 1977)
- The title of the article by Wilcox (1998) asks the question, “How Many Discoveries Have Been Lost by Ignoring Modern Statistical Methods?”
- We could also add the question, How much time and research effort has been wasted?

Conclusion

- The answer to both questions is probably “many, a lot,” given all the potential problems of using statistical tests as basically the only way to test hypotheses in the behavioral sciences
- The point is that there is no shortage of alternatives to traditional statistical tests
- Now is also a good time to invent the future

References

- Anderson, J. L. (1998, June). Embracing uncertainty: The interface of Bayesian statistics and cognitive psychology. *Conservation Ecology*, 2(1). Retrieved July 2, 2001, from <http://www.consecol.org/vol2/iss1/art2/index.html>
- Arbuckle, J. L. (2003). Amos 5 [Computer software]. Chicago: Smallwaters.
- Cornell, J., & Mulrow, C. (1999). Meta-analysis. In H. J. Adler & G. J. Mellenbergh (Eds.), *Research methodology in the social, behavioral, and life sciences* (pp. 285-323). Thousand Oaks, CA: Sage.
- Diaconis, P. & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American*, 248(5), 116-130.
- Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193-242.
- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.
- Gatsonis, C., Kass, R. E., Carling, B., Carriquiry, A., Gelman, A., Verdinelli, I., et al. (Eds.). (2001). *Case studies on Bayesian statistics* (Vol. 5). New York: Springer-Verlag.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman & Hall.
- Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Vol. 1. Methodological issues* (pp. 311-339). Hillsdale, NJ: Lawrence Erlbaum.
- Harsanyi, J. C. (1983). Bayesian decision theory, subjective and objective probabilities, and acceptance of empirical hypotheses. *Synthese*, 57, 341-365.
- Howard, G. S., Maxwell, S. E., & Fleming, K. J. (2000). The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychological Methods*, 5, 315-332.
- Iversen, G. R. (1984). *Bayesian statistical inference*. Newbury Park, CA: Sage.

- Lunneborg, C. E. (2001). Random assignment of available cases: Bootstrap standard errors and confidence intervals. *Psychological Methods*, 6, 402-412.
- Matthews, R. A. J. (2000). Facts versus factions: The use and abuse of subjectivity in scientific research. In J. Morris (Ed.), *Rethinking risk and the precautionary principle* (pp. 247-282). Woburn, MA: Butterworth-Heinemann.
- Pitz, G. F. (1982). Applications of Bayesian statistics in psychology research. In G. Keren (Ed.), *Statistical and methodological issues in psychology and social sciences research* (pp. 245-281). Hillsdale, NJ: Erlbaum.
- Reichardt, C. S., & Gollob, H. F. (1997). When confidence intervals should be used instead of statistical tests, and vice versa. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37-64). Mahwah, NJ: Lawrence Erlbaum.
- Rodgers, J. L. (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. *Multivariate Behavioral Research*, 34, 441-456.
- Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education*, 61, 361-377.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
- Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist*, 53, 300-314.