Bootstrap Methods Short Course

By Michael Chernick

To the International Biometrics Society, German Region

Day 2

November 5, 2010

Bootstrap Topics Day 2

- **Examples of bootstrap applications: (1) P-value adjustment - consulting example, (2) Confidence Interval for Process Capability C_{pk}, (3) Bioequivalence - Efron's Patch Data example**
- **Examples where bootstrap is not consistent:**
  (1) infinite variance case for a population mean,
  (2) extreme values, (3) survey sampling
- **Available Software**
- **Efficient Algorithms in SAS**
- **Examples with Software Solutions**

Examples of bootstrap applications

- *P*-value adjustment - a consulting example
- Many problems in the course of a clinical trial involve multiple comparisons or repeated significance tests for a key endpoint at various follow-up times.
- In these cases, the individual test *p*-values are not appropriate and *p*-value adjustment is appropriate.
- Conservative estimates based on the Bonferroni inequality are often used but sometimes may be too conservative.

P-value Adjustment Application

- Westfall and Young (1993) have demonstrated useful bootstrap and permutation approaches which work in a wide variety of multiple testing situations.
- Their methods are implemented in the SAS software package (Version 6.12 or higher) through a procedure called PROC MULTTEST.
- Chernick has implemented this approach in a number of clinical trials.
- As a consultant on a particular clinical trial he employed *p*-value adjustment to determine if results differed significantly depending on the country where the patient was treated.

P-value Adjustment Application (continued)

- This example is presented in Section 8.5.3 of Chernick (2007).
- A company conducted a clinical trial for a medical treatment in one country but due to slow enrollment decided to extend the trial to other countries.
- The initial country we denote as country E.
- The other four countries are labeled A, B, C and D.

P-value Adjustment Application (continued)

- Fisher's exact test was used to compare failure rates for the treatment with failure rates for the control. The primary statistical analysis of the endpoint.
- In country E, the result showed that the treatment was superior to the control, but this was not the case in the other countries.
- The client wanted to show that there were differences among countries which made the poolability of the data questionable.

P-value Adjustment Application (continued)

- They wanted to claim that only the data in country E was relevant to the submission since they were seeking regulatory approval only in country E.
- This involved comparing treatment success in each country compared to country E.

P-value Adjustment Application (continued)

- There are 4 relevant pairwise comparisons of other countries with country E.
- Consequently, the raw *p*-values from the individual Fisher tests are not appropriate.
- The raw *p*-values were compared with the Bonferroni adjustment and the bootstrap adjustment.

P-value Adjustment Application (continued)

TABLE 8.1 from Chernick (2007) page 152

Comparison of Treatment Failure Rates

Country	failure rate (percentage)	failure rate (fraction)
– A	40%	(18/45)
– B	41%	(58/143)
– C	29%	(20/70)
– D	29%	(51/177)
– E	22%	(26/116)

TABLE 8.2 from Chernick (2007) page 153

Comparison of p-value adjustments

Countries	Raw p	Bonf. p	Boot. p
– E vs A	0.0307	0.1229	0.0855
– E vs B	0.0021	0.0085	0.0062
– E vs C	0.3826	1.0000	0.7654
– E vs D	0.2776	1.0000	0.6193

P-value Adjustment Application (continued)

- The raw *p*-values indicated that failure rate for E was statistically significantly different (lower) from A and B at the 5% level.
- But results are misleading since they ignore the multiple testing.
- The Bonferroni bound shows only E and B to be statistically significantly different at the 10% level.

P-value Adjustment Application (continued)

- But the Bonferroni bound is known to be excessively conservative in some situations.
- Bootstrap provides an appropriate answer.
- For the bootstrap estimate we again find that E and B are clearly different but now we find that the p-value for E and A is below 0.10 and so E is statistically significantly better than A at the 10% level.

References on *p*-value adjustment

(1) Chernick, M.R. (2007). *Bootstrap Methods: A Guide for Practitioners and Researchers, 2nd Edition.* Wiley, New York.

(2) Westfall, P. and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples of p-Value Adjustment.* Wiley, New York.

Confidence Intervals for C_{pk} Application

- Many manufacturing companies use process capability indices to measure how well the production process behaves relative to specification limits.
- These indices were used by the Japanese in the 1960s - 1980s as part of the quality movement.
- The movement in the US in the 1990s has also included the use of these measures.

Confidence Intervals for C_{pk} Application (continued)

- Definition of C_{pk}: Let m be the process mean and σ the process standard deviation.
- Let LSL represent the lower specification limit and USL the upper specification limit.
- Then C_{pk} is the minimum of (USL-m)/(3 σ) and (m-LSL)/(3 σ).
- In practice m and σ are usually estimated from the data to give an estimate of C_{pk}.
- For Gaussian distributions, confidence intervals can be generated and hypothesis tests performed.

Confidence Intervals for C_{pk} Application (continued)

- C_{pk} is a measure of how well the process stays within the specification limits.
- Common benchmark values are 1.0 and 1.33.
- A C_{pk} of 1.0 implies for Gaussian data that at least 99.73% of the cases would be within specifications.
- A C_{pk} of 1.33 implies for Gaussian data that at most 6 out 100,000 cases would be expected to fall outside the specification limits.

Confidence Intervals for C_{pk} - Application (continued)

- A C_{pk} of 2.67 would imply for Gaussian data that it would take about 306 million years to see one defect (a case outside the specification limits) when 10,000 parts are produced each day over 5 day work weeks.
- Obviously, the higher the C_{pk} is the better the situation is.
- However, the probability statements depend heavily on the Gaussian assumption.
- People tend to think of numbers like 1.0 and 1.33 as good and numbers under 1.0 as bad without regard to the distribution of the data.

Confidence Intervals for C_{pk} Application (continued)- Misuse of Gaussian Assumption

- If the data are skewed or have short tails, the probability associated with numbers like 1.0 are misleading.
- It is possible for the proportion of cases that are within the specification limits to be higher than 99.73% for C_{pk} of 1.0 when the data have shorter tails than the Gaussian.
- Even for Gaussian data, probability results are based on knowing C_{pk}.

Confidence Intervals for C_{pk} Application (continued)- Misuse of Gaussian Assumption

- In practice we estimate it from data.
- More appropriate analysis would be to determine confidence limits for C_{pk}.
- Confidence bounds for non-Gaussian data can be very different from those for Gaussian data.

Confidence Intervals for C_{pk} Application (continued) - Gunter's Examples

- To understand how C_{pk} relates to tail probabilities for different distributions Gunter compared
- 1. highly skewed data - chi square distribution with 4.5 degrees of freedom to
- 2. A heavy-tailed symmetric distribution- student's t with 8 degrees of freedom and
- 3. A short tailed distribution - uniform.

Confidence Intervals for C_{pk} Application (continued) - Gunter's Examples

- Gunter computes the expected value out of one million parts that would fall outside of the limits set by 3 standard deviations.

Results:

- In case 1. 14,000 are outside the limits all above the upper limit.
- In case 2. 4,000 are outside the limits (2,000 below the lower limit and 2,000 above the upper limit)
- In case 3. None are outside the limits!
- In contrast for Gaussian data 2,700 would fall outside (1350 below the lower limit and 1350 above the upper limit.

Confidence Intervals for C_{pk} Application (continued) - Lesion Data

- At Biosense Webster, we generated lesions in beef heart to compare effectiveness of different catheter types or the Radiofrequency (RF) generators used with the catheters.
- Lesion depth is a common performance measure.
- Lesions should be deep enough for effective ablation to eliminate an arrhythmia but not so deep as to cause a complication such as cardiac tamponade (perforation of the heart).

Confidence Intervals for C_{pk} Application (continued) - Lesion Data

- Thirty sample lesions were generated in a beef heart to evaluate the performance of a particular catheter.
- Lower and upper specification limits could be set so that safe and effective lesions would be the ones that fall within the limits.
- Given these limits C_{pk} is defined and can be estimated by bootstrapping.

Confidence Intervals for C_{pk} Application (continued) Bootstrap Approach

- A bootstrap sample of size 30 is generated by sampling with replacement 30 times from the original sample of 30.
- A Monte Carlo approximation to the bootstrap distribution is gotten by generating say 10,000 bootstrap samples.
- A C_{pk} estimate is computed for each bootstrap sample.

Confidence Intervals for C_{pk} Application (continued) Bootstrap Approach

- The histogram of C_{pk} estimates is used to obtain bootstrap confidence intervals for C_{pk}.
- The simplest such confidence interval is Efron's percentile method which takes the interval from the 2.5 percentile of the bootstrap histogram to the 97.5 percentile of the bootstrap histogram as the 95% confidence interval.

Confidence Intervals for C_{pk} Application (continued) Summary Statistics for Lesion Depth

- For catheter number 12, thirty sample lesions were generated in a particular beef heart.
- The mean depth was 6.650 millimeters.
- The standard deviation of the depth was 0.852 millimeters.
- The minimum depth was 5.000 millimeters.
- The maximum depth was 9.000 millimeters.

Confidence Intervals for C_{pk} Application (continued) - Specification Limits

- Specification limits for lesion depth are not really known.
- For illustrative purposes, we hypothetically set some plausible values.
- LSL= 4.000 millimeters.
- USL= 10.000 millimeters.
- Target value (mean value) = 7.000 millimeters.

Confidence Intervals for C_{pk} Application (continued) - Specification Limits

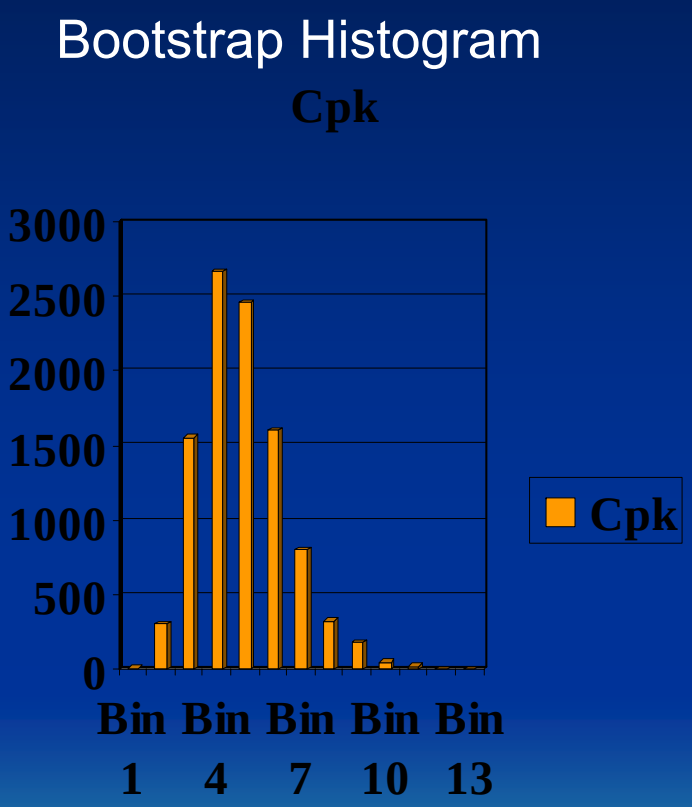
- Based on these limits we can compute C_{pk}.
- A common sample estimate is the plug-in method which replaces the mean m and the standard deviation σ with their sample estimates.
- The observed estimate is 1.036.

Confidence Intervals for C_{pk} Application (continued) Shapiro-Wilk Test for Normality of Lesion Depth Data

- We apply the Shapiro-Wilk test to the lesion depth data to see if Gaussian confidence limits can be applied.

- The *p*-value for the Shapiro-Wilk test is 0.0066, indicating non-normality. The Lesion Depth data appears highly skewed and this is apparent in the bootstrap histogram as well, explaining the failure of the test for normality.

Confidence Intervals for C_{pk} Application (continued) Bootstrap Histogram for 10,000 Bootstrap Samples



- Bin half width 0.05
- Bin1 center 0.7
- Bin 2 center 0.8
- Bin 3 center 0.9
- Bin 4 center 1.0
- Bin 5 center 1.1
- Bin 6 center 1.2
- Bin 7 center 1.3
- Bin 8 center 1.4
- Bin 9 center 1.5
- Bin 10 center 1.6
- Bin 11 center 1.7
- Bin 12 center 1.8
- Bin 13 center 1.9

Confidence Intervals for C_{pk} Application (continued) - Bootstrap Percentile Method for Confidence Intervals

- The Bootstrap Percentile Method (Efron's) gives the interval [0.8408, 1.4608] as a 95% confidence interval for this case.
- Heavlin's approximate 95% confidence interval for C_{pk} in the Gaussian case is [0.7084,1.3640].
- The large difference indicates the inappropriateness of using Gaussian methods with non-Gaussian data.
- Although higher order bootstrap confidence intervals are usually preferred, keep in mind the small sample size and heavy skewness of the lesion depth distribution which is even evident in the bootstrap histogram.
- So results from Chernick and LaBudde (2010) indicate that the Percentile Method can possibly give a more accurate confidence interval than BCa and other high-order bootstrap methods.

References for C_{pk} Example

(1) Chernick, M.R. (1999). *Bootstrap Methods: A Practitioner's Guide.* Wiley, New York.

(2) Franklin, L.A. and Wasserman, G.S. (1991). *Bootstrap confidence interval estimates of C_{pk}: An Introduction.* Communications in Statistics - Simulation and Computation **20,** 231-242.

(3) Gunter , B.H. (1989a). *The use and abuse of C_{pk}.* Quality Progress **22 (3),** 108-109.

(4) Gunter, B.H. (1989b). *The use and abuse of C_{pk}.* Quality Progress **22 (5),** 79-80.

(5) Heavlin, W.D. (1988). *Statistical properties of capability indices.* Technical Report # 320, Technical Library, Advanced Micro Devices Inc., Sunnyvale.

References for C_{pk} Example (continued)

(6) Kane,V.E. (1986). *Process Capability Indices.* Journal of Quality Technology **24,** 41-52.

(7) Kotz, S. and Johnson, N.L. (1993). *Process Capability Indices.* Chapman and Hall, London.

(8) Price, B. and Price, K. (1992). *Sampling variability of capability indices.* Technical Report, Wayne State University, Detroit, Michigan.

(9) Resampling Stats, Inc. (1997). *Resampling Stats User's Guide.* (Peter Bruce, Julian Simon and Terry Oswald, authors).

References for C_{pk} Example (continued)

(10) Ryan, T.P. (1989). *Statistical Methods for Quality Improvement.* John Wiley and Sons, Inc., New York.

(11) Chernick, M. R. and LaBudde, R. (2008). More Qualms About Bootstrap Confidence Intervals. *Unpublished Manuscript to be submitted to the Journal of Nonparametric Statistics.*

Bioequivalence Application - Efron's Patch Data Example - Bias of Ratio estimates

- If X is an unbiased estimate of θ and Y is an unbiased estimate of a parameter μ that is independent of X then the estimator X/Y has $E(X/Y) \geq \theta/\mu$.
- The proof involves Jensen's inequality. Since X and Y are statistically independent

  $E(X/Y) = E(X)\,E(1/Y) = \theta\,E(1/Y)$.
- Now the reciprocal function $f(z)=1/z$ is a convex function and therefore by Jensen's inequality

  $f(E(Y)) \leq E(f(Y)) = E(1/Y)$, so $f(\mu) = 1/\mu \leq E(f(Y)) = E(1/Y)$.
- Consequently $E(X/Y) = \theta\,E(1/Y) \geq \theta/\mu$.
- In fact the inequality is strict unless f is a linear function of z.

Bioequivalence Application - Efron's Patch Data Example - Bias of Ratio estimates (continued)

- B=bias of X/Y= E(X/Y)- θ /μ.
- B is positive.
- Ratio estimates are common in survey sampling.
- See Cochran (1977) for examples.
- The Patch Data example for bioequivalence comes from Efron and Tibshirani (1993) pp. 126-133 and is also presented in Chernick (1999) pp. 41- 43 or Chernick (2007) pp. 44 - 46.

Patch Data Example

- This was a small clinical trial used to show the FDA that a product produced at a new plant is “equivalent” to the product produced at the old plant.
- The experimental setup is as follows:

(1) The patch is manufactured at the old site with the hormone denoted old, (2) the patch is manufactured at the new site with the hormone denoted new, and (3) the patch is manufactured at the new site with no hormone denoted placebo.

Patch Data Example (continued)

- It is a cross-over trial with each of eight patients receiving the new patch, old patch, and placebo in a “random” order.
- FDA defines equivalence through the parameter

  θ / μ where θ = E(new) - E(old) and μ = E(old) - E(placebo).

- Equivalence means $| \theta / \mu | \leq 0.20$. Since we must estimate both θ and μ this is translated to mean that the upper 95% confidence bound for

  $| \theta / \mu |$ is less than or equal to 0.20.

Patch Data Example (continued)

- The natural estimate for θ is the average of the new patch hormone level in patients minus the average of the old patch hormone level in patients.
- Similarly, the natural estimate of μ is the average of the old patch hormone level in patients minus the average of the placebo level in patients.
- The plug-in estimate, which is the ratio of these two estimates, is a possible estimate of the ratio.

Patch Data Example (continued)

- In the simplified case where the two estimates are independent we have seen that such an estimate is biased.
- In this case the estimates are correlated since the new ratio average appears in both.
- Still this plug-in estimate is biased.

Patch Data Example (continued)
Bootstrap Approach

- The bootstrap approach to this problem is to estimate the bias of the plug-in estimator by bootstrapping and then subtract the bootstrap estimate of the bias from the plug-in estimator to get the bootstrap estimate of the ratio.

- The results from Efron and Tibshirani (1993) are summarized in the following table taken from their book.

Patch Data Summary Table

Subject	Old-Placebo	New-Old
– 1	8406	-1200
– 2	2342	2601
– 3	8187	-2705
– 4	8459	1982
– 5	4795	-1290
– 6	3516	351
– 7	4796	-638
– 8	10238	-2719
– Avg.	6342	-452.3

Patch Data Example - Conclusion

- Based on the data the plug-in estimate for θ/μ is -0.0713 which is considerably less in absolute value than 0.20.
- But this estimate has a potentially large bias. Efron and Tibshirani (1993) generated 400 bootstrap samples and found the bootstrap estimate of bias to be only 0.0043.
- After applying this bias adjustment to our estimate we still get an estimate that is considerably less than 0.20 in absolute value and hence we are safe in accepting equivalence since we expect the confidence bound to be below 0.20 also.

Patch Data Example Why Ratios?

- Ratios were used in the Patch Data Example because the definition of equivalence involved ratios.
- Sometimes ratios are easier to think about than mean differences. They are dimensionless quantities (percentage differences).
- Had we looked at mean differences, we would not have had a bias problem but we would have had to determine a meaningful difference and estimate a standard error.

References for Patch Data Example

(1) Cochran, W. (1977). *Sampling Techniques*. **3rd ed.,** Wiley, New York.

(2) Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap.* Chapman & Hall, New York.

(3) Chernick, M. R. (1999). *Bootstrap Methods: A Practitioner's Guide.* Wiley, New York.

(4) Chernick, M. R. (2007). *Bootstrap Methods: A Guide for Practitioners and Researchers.* Wiley, New York.

Individual Bioequivalence

- Bioequivalence is a level of performance expect for a new formulation of an approved drug or for approval of a generic drug.
- The FDA has a Guidance document on how to conduct bioequivalence (bioavailability) trials.
- Three types of bioequivalence have been defined (1) average bioequivalence, (2) population bioequivalence and (3) individual bioequivalence.
- Currently the FDA only requires average bioequivalence be shown (a change over past policy).
- Bootstrap solutions useful in determining individual bioequivalence and population bioequivalence have been devised and shown to be consistent.
- We shall look only at individual bioequivalence as an example.
- In the model we consider crossing over twice with the sequence TRR meaning new treatment first and then the reference treatment 2 time and RTR, reference first followed by new treatment and then the reference again.

Individual Bioequivalence: Model

Consider the following model for pharmacokinetic response in a 2 treatment crossover design using only sequences RTR and TRR randomized 1:1.

$$Y_{ijkl} = \mu + F_l + P_j + Q_k + W_{ijk} + S_{ikl} + \varepsilon_{ijkl}$$, where μ is the overall mean,

P_j is the fixed effect for the jth period with the constraint $\sum P_j = 0$, Q_k is the fixed effect for the k^{th} sequence with $\sum Q_k = 0$, F_l is the fixed effect for the l^{th} drug. For these trials, we only have two drugs the new and old formulations denoted T for the new treatment and R for the reference formulation. We also have the constraint that $F_T + F_R = 0$.

Now S_{ikl} is a random effect of the i^{th} subject in the k^{th} sequence with the l^{th} treatment, W_{ijk} is the fixed interaction between treatment , sequence and period and ε_{ijkl} is a random noise (error) component with mean 0 independent and identically distributed and independent of all the fixed and random effects.

Individual Bioequivalence: Definition

- Under the linear model given on the previous slide individual bioequivalence is accepted if after testing

H_0: $\Delta_{PB} \leq \Delta$ versus H_1: $\Delta_{PB} > \Delta$, where $\Delta_{PB} = P_{TR} - P_{RR}$ with $P_{TR} = \text{prob}(|Y_T - Y_R| \leq r)$ and $P_{RR} = \text{prob}(|Y_T - Y'_R| \leq r)$ where Δ and r are determined fixed constants and Y'_R is the observed response the second time the reference treatment is given.

Bootstrap Results for this Trial

- See Schall and Luus (1993) for a description of a bootstrap hypothesis test for this problem.
- Pigeot (2001) in a survey article describes the Schall and Luus method in detail, shows that their method is not consistent and modifies it by constructing a bootstrap percentile method confidence interval to use in the test.
- In an earlier work Shao, Kübler and Pigeot (2000) prove that the bootstrap method Pigeot describes in Pigeot (2001) is consistent.

Use of Bootstrap in Clinical Trials

- Although we have seen the bootstrap used in medical examples and for individual bioequivalence in clinical trials, a case can be made to say that it is underutilized in clinical trials
- Sauerbrei and Royston (2007) make a compelling case for this with the hope that their paper will lead to greater use.
  - They illustrate through examples (1) that the bootstrap can be used as an aid in the design of a trial (2) the bootstrap can be used in conjunction with data-dependent modeling in clinical trials (3) the bootstrap can help find factors that interact with the treatment in affecting patient response. [See also Gunter, Zhu and Murphy (2010)]

References for Individual Bioequivalence and Clinical Trials

(1) Chernick, M. R. (2007). *Bootstrap Methods: A Guide for Practitioners and Researchers.* Wiley, New York.

(2) Gunter, L., Zhu, J and Murphy, S. A. (2010) Variable selection for qualitative interactions. Statistical Methodology journal homepage: www.elsevier.com/locate/stamet

(3) Pigeot, I. (2001). The jackknife and bootstrap in biomedical research – Common principles and possible pitfalls. *Drug Information J.* 35, 1431-1443.

(4) Sauerbrei, W. and Royston, P. (2007). Modelling to extract information from clinical trials data: On some roles for bootstrap. *Statist. Med.* 26, 4989-5001.

(5) Schall, R., and Luus, H. G. (1993). On population and individual bioequivalence. *Statist. Med.* 12, 1109-1124.

(6) Shao, J., Kübler, and Pigeot, I. (2000). Consistency of the bootstrap procedure in individual bioequivalence. *Biometrika* 87, 573-585.

Examples where the bootstrap fails

- The conference on bootstrap in Ann Arbor, Michigan in 1990 and published in LePage and Billard (editors) (1992) explores the application and limitations of the bootstrap.
- Mammen (1992) provides mathematical conditions for the bootstrap to work.
- Bickel and Freedman (1981) and Knight (1989) provide examples where the bootstrap fails to be consistent.

Examples where the bootstrap fails (continued)

- Athreya (1987) shows that the bootstrap estimate of the sample mean is inconsistent when the population distribution has an infinite variance.

- Angus (1993) provides similar inconsistency results for the maximum and minimum of a sequence of independent identically distributed observations.

Examples where the bootstrap fails (continued)

We shall describe the inconsistency of the bootstrap in three cases and then provide remedies

(1) sample mean with infinite population variance,

(2) maximum term in an i.i.d sequence of observations and

(3) estimate of a mean in a survey sample.

In (3) a simple adjustment to the sampling that adjusts for the finite population size can be used to obtain a consistent bootstrap estimate.

Example where the bootstrap fails - Sample Mean with Infinite Population Variance

- Singh (1981) and Bickel and Freedman (1981) showed that in the case of estimating the mean from an i.i.d. sample with a finite population variance the bootstrap procedure is consistent.

- In the case of an infinite variance, the population distribution might have a distribution, F(x) satisfying
  $1 - F(x) \sim cx^{-\alpha} L(x)$ where L is a slowly varying function as $x \to \infty$, c is a nonnegative constant and $0 < \alpha \leq 2$.

- Under these conditions, the sample mean appropriately normalized, converges to a stable distribution.

Example where the bootstrap fails - Sample Mean with Infinite Population Variance (continued)

- For $\alpha = 2$ the variance of F is finite and the central limit applies. For $\alpha < 2$ the population variance is infinite.
- Theorem 1 of Athreya (1987) proves the inconsistency of the bootstrap for the case where $1 < \alpha < 2$.
- The result tells us that when we appropriately normalize the sample mean and apply the bootstrap substitutions the bootstrap version of the normalized mean converges to a **random probability distribution** and **not** to the corresponding **fixed stable distribution** that the sample mean converges to.

Example where the bootstrap fails - Estimating extreme values

- For i.i.d. random variables Gnedenko's theorem usually applies to the maximum or minimum values.
- Gnedenko's theorem states that when appropriately normalized the minimum value and the maximum value converge to one of three extreme value distribution families.
- The appropriate family depends on the tail behavior of the population distribution.

Example where the bootstrap fails - Estimating extreme values (continued)

- Angus (1993) showed that using the appropriate normalization and the bootstrap substitution, the maximum and minimum converge to a **random probability distribution** and **not the fixed extreme value distribution** from Gnedenko's theorem that the sample extremes converge to.

Bootstrap Remedies

- In the past decade many of the problems where the bootstrap is inconsistent remedies have been found by researchers to give good modified bootstrap solutions that are consistent.
- For both problems describe thus far a simple procedure called the *m-out-n* bootstrap has been shown to lead to consistent estimates .
- Zelterman (1993) uses semi-parametric bootstrap methods to get around the problem

The *m-out-of-n* Bootstrap

- This idea was proposed by Bickel and Ren (1996) for handling doubly censored data.
- Instead of sampling n times with replacement from a sample of size n they suggest to do it only m times where m is much less than n.
- To get the consistency results both m and n need to get large but at different rates. We need m = o(n).

  That is $m/n \to 0$ as m and n both $\to \infty$.
- This method leads to consistent bootstrap estimates in many cases where the ordinary bootstrap has problems, particularly (1) mean with infinite variance and (2) extreme value distributions.

Example where the bootstrap fails - Survey Sampling

- In survey sampling, we generally have a finite population with N values but we can only afford to take a sample of size n from the population of size N.
- Generally, to draw inferences about population parameters like the mean, the population total or the population variance, the sample is drawn with a random mechanism such as simple random sampling or stratified random sampling and estimates are obtained from the sample.

Example where the bootstrap fails - Survey Sampling (continued)

- Under simple random sampling, a random sample of size n has mean μ where μ is the population mean but the variance of the sample mean is $(1 - f)\ \sigma^2 / n$ where $f = n/N$ rather than σ^2 / n (which is the result for infinite populations).
- The factor f is called the finite population correction.
- See Cochran (1977) for more details.

Example where the bootstrap fails - Survey Sampling (continued)

- Applying the ordinary bootstrap to the sample mean of a sample of size n from a finite population will lead to a bootstrap distribution for the sample mean with variance σ^2 / n rather than $(1 - f) \sigma^2 / n$, the correct variance.
- So the ordinary bootstrap is **not** consistent.
- The bootstrap can be made consistent by choosing an appropriate bootstrap sample size m which is smaller than the original sample size n.
- See Chernick (1999, 2007) Section 9.4 for details and additional references.

References on when bootstrap fails

(1) Angus, J. E. (1993). Asymptotic theory for bootstrapping the extremes. *Communs. Statist. Theory and Methods* **22,** 15-30.

(2) Athreya, K. B. (1987). Bootstrap estimation of the mean in the infinite variance case. *Ann. Statist.* **15,** 724 - 731.

(3) Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9,** 1196 - 1217.

(4) Chernick, M.R. (1999). *Bootstrap Methods: A Practitioner's Guide.* Wiley, New York.

References on when bootstrap fails (continued)

(5) Chernick, M.R. (2007). *Bootstrap Methods: A Guide for Practitioners and Researchers, 2nd Edition.* Wiley, New York.

(6) Cochran, W. (1977). *Sampling Techniques*. **3rd ed.,** Wiley, New York

(7) Knight, K. (1989). On the bootstrap of the sample mean in the infinite variance case. *Ann. Statist.* **17,** 1168-1175.

References on when bootstrap fails (continued)

(8) LePage, R., and Billard, L. (editors). (1992). *Exploring the Limit of Bootstrap.* Wiley, New York.

(9) Mammen, E. (1992). *When Does the Bootstrap Work? Asymptotic Results and Simulations* Springer-Verlag, Heidelberg.

(10) Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.* **9,** 1187-1195.

Available Software

- Resampling Stats from Resampling Stats Inc. (provides basic bootstrap tools in easy to use software and is good as an elementary teaching tool).
- SPlus from Insightful Corporation (good for advanced bootstrap techniques such as BCa, easy to use in new Windows based version). The current module Resample is what we are using for the course.
- S functions provided by Tibshirani (see Appendix in Efron and Tibshirani text or visit Rob Tibshirani's web site http:/www.stat-stanford.edu/~tibs)
- Stata has a bootstrap algorithm available that some users rave about.

Available Software (continued)

- Mathworks and other examples (see Susan Holmes web page: http:/www-stat.stanford.edu/~susan) or contact her by email (may be outdated)
- SAS macros are available and Proc MULTTEST and Proc SURVEY SELECT do bootstrap sampling (version 9.2).
- R CRAN libraries

Efficiency of SAS Algorithms

Paper by Opdyke published online for InterStat in October 2010 compares speed of 7 SAS algorithms to generate bootstrap samples.

(1) One-Pass, Duplicates-Yes (OPDY)
(2) Proc Survey Select (PSS)
(3) Hash Table, Proc Summary (HTPS)
(4) Hash Table, Hash Iterator (HTHI)
(5) Direct Access (DA)
(6) Output – Sort – Merge (Out-SM)
(7) Algorithm 4.8 (A4.8)

More on Opdyke Paper

- Algorithms 2-6 are commonly used by SAS users according to Opdyke
- Algorithms 1 and 7 are not common but 1 is faster in almost all scenarios based on his comparisons of real run times all on the same machine
- He also presents CPU time and other statistics related to the execution of the software but the emphasis is on real run time and not on computational complexity
- OPDY only requires Base SAS while others may require SAS/STAT

Sample Comparisons from Table 1
Run Times Relative to OPDY (algorithm time/OPDY time)

N per * Stratum	# of Strata	n*	m*	PSS	A4.8	HTPS	HTIT	DA	Out-SM
10,000	2	500	500	5.3	10.2	9.0	4.5	8.8	7.5
100,000	2	500	500	23.3	83.3	6.2	4.6	7.3	9.7
1,000,000	2	500	500	31.1	121.2	2.5	1.9	1.4	4.2
10,000,000	2	500	500	42.2	164.9	10.0	2.7	0.7	7.6

* N is the number of Monte Carlo replications in each stratum
n is size of original sample, m is the size of bootstrap sample

Opdyke goal and conclusions

- Goal – Develop a non-resource intensive algorithm for random sampling with replacement to execute bootstraps faster than competititors on the SAS platform.
- Conclusions
  - OPDY achieves this objective
  - OPDY is better than PSS which was developed by the SAS institute with a similar objective
  - Results depend on the computer system being used and hashing algorithms may be more competitive on systems with very fast I/O speeds

Opdyke Reference

- Opdyke, J.D., “Much Faster Bootstraps Using SAS®,” ***InterStat***, October 2010. (has code for 5 SAS routines used in the comparison but does not contain the two hashing algorithms)

Example 1: Guinea Pig Survival Times

- *Survival in Day ordered smallest to largest for 72 guinea pig in a medical experiment*
- *43 45 53 56 56 57 58 66 67 73 74 79 80 80 81 81 81 82 83 83 84 88 89 91 91 92 92 97 99 99 100 100 101 102 102 102 103 104 107 108109 113 114 118 121 123 126 128 137 138 139 144 145 147 156 162 174 178 179 184 191 198 211 214 243 249 329 380 403 511 522 598*

Example 1: Guinea Pig Survival Times

- *Assignment 1*
  - *a. Make a histogram of the survival times. The distribution is strongly skewed.*
  - *b. The central limit theorem says that the sampling distribution of the sample mean x-bar becomes Normal as the sample size increases. Is the sampling distribution roughly Normal for n = 72? To find out, bootstrap these data and inspect the bootstrap distribution of the mean (use a Normal quantile plot). How does the distribution differ from Normality? Is the bootstrap distribution more or less skewed than the data distribution?*

Example 1: Guinea Pig Survival Times

- *Assignment 1*
  - *Using the Histogram tool in the Resampling Stats toolbar with 8 bins:*
  - Bin MidPt Counts % Total Cu. Freq.

Bin MidPt	Counts	% Total	Cu. Freq.
82.643	45	62.5	62.5
161.929	17	23.611	86.111
241.214	4	5.556	91.667
320.5	1	1.389	93.056
399.786	2	2.778	95.833
479.071	1	1.389	97.222
558.357	1	1.389	98.611
637.643	1	1.389	100

The distribution is very skewed to the right (note that survival times cannot be negative).

Example 1: Guinea Pig Survival Times

- *Assignment 1*
- *Now we generate a bootstrap resample using the "R" tool, and add a cell to calculate the average of the resample. Then we use the "RS" tool to generate 1,000 bootstrap resamples, scored by the average and placed in the "Results" worksheet. To generate the normal Q-Q plot, we sort the resample averages in column A from low to high, add column B with a sequence number 1 .. 1,000. Then we add another column (C) with a formula, e.g., "=(B1-0.5)/1000" to calculate the probabilities of the ECDF from the sequence number. Finally, we add a column that gives the standard normal quantile z corresponding the ECDF value in column C (e.g., "=norminv(c1, 0,1)"). After this, we generate a scatter plot of the observed resample averages vs. the standard normal quantiles. For comparison, the Resampling Stats Histogram graph and table are also shown on the handout.*
- *The resampling averages do indeed look much more normal. There is still, however, a ittle skewness evident to the right and left. Note that the histogram is useful in judging asymmetry.*

Example 1: Guinea Pig Survival Times

- ***SPLUS SOLUTION:***

> #Problem 18.4

> pigs = c(43, 45, 53, 56, 56, 57, 58, 66, 67, 73, 74, 79, 80, 80, 81, 81, 81, 82, 83, 83, 84, 88, 89, 91, 91, 92, 92, 97, 99, 99,100, 100, 101, 102, 102, 102, 103, 104, 107, 108, 109, 113, 114, 118, 121, 123, 126, 128, 137, 138, 139, 144, 145, 147, 156, 162, 174, 178, 179, 184, 191, 198, 211, 214, 243, 249, 329, 380, 403, 511, 522, 598)

> length(pigs) #show count [1] 72

- **> summary(pigs)**
- **Min. 1st Qu. Median Mean 3rd Qu. Max.**
- **43.00 82.75 102.50 141.80 149.30 598.00**

Skewness is evident from difference between median and mean and between median and two extrema.

- ***The histogram on the handout shows very strong skew to the right. Survival times are an extreme value distribution, so this is understandable.***
- *The normal quantile plot (see handout) shows complete non-normality. The quadratic shape is typicalof extreme value distributions.*

Example 1: Guinea Pig Survival Times

- ***SPLUS CODE AND OUTPUT***

```
> boot = bootstrap(pigs, mean)
Forming replications 1 to 100 Forming replications 101 to 200
… Forming replications 901 to 1000
> summary(boot)
Call: bootstrap(data = pigs, statistic = mean)
Number of Replications: 1000
Summary Statistics:
Statistic    Observed   Bias      Mean    SE
mean         141.8      -0.8457   141     12.67
Empirical Percentiles:   2.5%   5%     95%     97.5%
For mean                 117.9  120.6  162.8   166.6
BCa Confidence Limits:   2.5%   5%     95%     97.5%
For mean                 121    125    167.4   174
> plot(boot) 100
```

Example 1: Guinea Pig Survival Times

- ***SPLUS SCRIPT***

```
pigs = c( 43, 45, 53, 56, 56, 57, 58, 66, 67, 73, 74, 79, 80, 80, 81, 81,
    81, 82, 83, 83, 84, 88, 89, 91, 91, 92, 92, 97, 99, 99,100, 100, 101,
    102,102, 102, 103, 104, 107, 108, 109, 113, 114, 118, 121, 123,
    126, 128, 137,138, 139, 144, 145, 147, 156, 162, 174, 178, 179,
    184, 191, 198, 211, 214, 243, 249, 329, 380, 403, 511, 522, 598)
length(pigs) #show count
summary(pigs)
hist(pigs)
qqnorm(pigs)
boot = bootstrap(pigs, mean)
summary(boot)
plot(boot)
qqnorm(boot)
```

Example 1: Guinea Pig Survival Times

- ***R SCRIPT***

```
pigs = c( 43, 45, 53, 56, 56, 57, 58, 66, 67, 73, 74, 79, 80, 80, 81, 81,
    81, 82, 83, 83, 84, 88, 89, 91, 91, 92, 92, 97, 99, 99,100, 100, 101,
    102,102, 102, 103, 104, 107, 108, 109, 113, 114, 118, 121, 123,
    126, 128, 137,138, 139, 144, 145, 147, 156, 162, 174, 178, 179,
    184, 191, 198, 211, 214, 243, 249, 329, 380, 403, 511, 522, 598)
length(pigs) #show count
summary(pigs)
hist(pigs, breaks=20)
qqnorm(pigs)
boot = NULL
for (k in 1:1000) {
s = sample(pigs, size = length(pigs), replace = TRUE)
boot[k] = mean(s)
}
summary(boot)
hist(boot, breaks=20)
qqnorm(boot)
```

Example 2: Percentiles as an aid in detecting non-Normality.

It is difficult to see any significant asymmetry in the bootstrap distribution of the correlation of Salaries vs Batting Averages Compare the percentiles and the t interval; does the difference between these suggest any skewness?

RESAMPLING STATS SOLUTION:

Salary	Average	Salary	Average	
9500000	0.269	3000000	0.250	
8000000	0.282	500000	0.214	
7333333	0.327	675000	0.234	
7250000	0.259	630000	0.324	Correlation in data 0.10676
7166667	0.240	7166667	0.240	Correlation in resample 0.06863
7086668	0.270	3150000	0.273	
6375000	0.253	5625000	0.238	
6250000	0.238	7333333	0.327	

Now use the "RS" tool to generate 1,000 resamples, each scored on the correlation based on the formula in the cell created. This puts 1,000 correlation estimates in the "Results" worksheet. Use the Histogram tool to show the resampling distribution, which is somewhat bell-shaped, but still possibly skewed and flat at the peak:

Example 2: Percentiles as an aid in detecting non-Normality.

Now compute the t-based and EP ("Efron or Empirical Percentile") 95% confidence intervals using the standard error of the resampling estimates plus the original sample correlation for the t-based interval, and the Excel "percentile" function for the EP interval:

Original correlation estimate 0.1068
Average resampling correlation 0.0965
Bias estimate 0.0102
Standard error of resamples 0.1236
Degrees of freedom 48
2-tail critical t value 2.011
t-based 95% LCL -0.1417 t-based 95% UCL 0.3552
EP 95% LCL -0.1365 EP 95% UCL 0.3343

Note that the EP interval will be biased by about the same amount as the Resampling average differs from the sample correlation. This bias is principally the result ofskewness in the resampling distribution. Adjusting the EP limits by +0.0102 makes them fairly close to the t-based interval, i.e., (-0.147, 0.344) vs. (-0.142, 0.355). Otherwise the EP limits are both lower.

Example 2: Percentiles as an aid in detecting non-Normality.

S-PLUS SCRIPT:

```
#Bootstrap Methods, Assignment 2, Problem 1
#Hesterberg Exer. 18-031, based on the data of Example 18-10
ex18.031 = read.table("ex18_031.txt", header=TRUE) #read file and parse
head(ex18.031) #show first few records
summary(ex18.031) #some quantiles
corrsamp = cor(ex18.031$Salary, ex18.031$Average) #Pearson correlation for sample
Corrsamp #bootstrap the correlation between salary and batting average
b = bootstrap(ex18.031, cor(Salary, Average)) #bootstrap the correlation
summary(b)
#compute t-based C.I.
df = nrow(ex18.031)-2 #degrees of freedom for correlation
df
tcrit = qt(.975, df) #critical t value for 95% C.I.
tcrit
corrsamp - tcrit*b$estimate[3] #compute t-based LCL
corrsamp + tcrit*b$estimate[3] #compute t-based UCL
par(mfrow=c(1,2))
plot(b)
qqnorm(b)
```

Example 2: Percentiles as an aid in detecting non-Normality.

- **R SCRIPT:**

```
#Bootstrap Methods, Assignment 2, Problem 1
#Hesterberg Exer. 18-031, based on the data of Example 18-10
ex18.031 = read.table("ex18_031.txt", header=TRUE) #read file and parse
head(ex18.031) #show first few records
summary(ex18.031) #some quantiles
corrsamp = cor(ex18.031$Salary, ex18.031$Average) #Pearson correlation for sample
corrsamp
#bootstrap the correlation between salary and batting average (use trick in help file!)
library('bootstrap')
theta = function(x){ cor(ex18.031$Salary[x],ex18.031$Average[x]) } #helper function
b = bootstrap(1:nrow(ex18.031),1000,theta) #bootstrap the correlation
bmean = mean(b$thetastar) #mean of bootstrap correlations
bmean
bbias = bmean - corrsamp #bias
bbias
bse = sd(b$thetastar) #standard error of bootstrap resampling distribution
bse
#compute t-based C.I.
df = nrow(ex18.031)-2 #degrees of freedom for correlation
df
```

Example 2: Percentiles as an aid in detecting non-Normality.

- **R SCRIPT CONTINUED:**

```
tcrit = qt(.975, df) #critical t value for 95% C.I.
tcrit
corrsamp - tcrit*bse #compute t-based LCL
corrsamp + tcrit*bse #compute t-based UCL
#Now compute the empirical percentile 95% confidence interval
quantile(b$thetastar, probs=c(.025,.975))
par(mfrow=c(1,2))
hist(b$thetastar)
lines(density(b$thetastar))
qqnorm(b$thetastar)
```

Example 3: Comparing various Confidence Intervals.

Comparing intervals. The bootstrap distribution of the 25% trimmed mean for the Seattle real estate sales is not strongly skewed. We were willing to give the 95% bootstrap t confidence interval for the trimmed mean of the population.Was that wise? Bootstrap the trimmed mean and give all of the bootstrap 95% confidence intervals: t, percentile and BCa. Make a picture that compares these Intervals by drawing a vertical line at x-bar25% and placing the intervals one above the other on this line. Describe how the intervals compare. Is the t Interval reasonably accurate?

Sample 25% trimmed mean 244.00 Resample 25% trimmed mean 261.43

Bin MidPt	Counts	% Total	Cu. Freq.
178.571	33	70.213	70.213
435.714	8	17.021	87.234
692.857	3	6.383	93.617
950	1	2.128	95.745
1207.143	0	0.000	95.745
1464.286	1	2.128	97.872
1721.429	0	0.000	97.872
1978.571	1	2.128	100

- *Note that the original data are very strongly skewed to the right.*

Example 3: Comparing various Confidence Intervals.

See handout for solution in SPLUS and R and script in SPLUS

R SCRIPT:

```
#BOOTSTRAP METHODS: ASSIGNMENT 2, PROBLEM 2
#Hesterberg Exercise 18.34
seattle = read.table("ex18_034.txt", header=FALSE, col.names=c('Price'))
summary(seattle)
sd(seattle$Price) #show original data standard deviation
trimdata = mean(seattle$Price, trim=0.25) #compute 25% trimmed mean of data
trimdata #display  #now preform bootstrap of trimmed mean
library(boot) #use Davison & Hinkley package for a change
helpf = function (x, i) { mean(x[i], trim = .25) }
b = boot(seattle$Price, helpf, 1000)
b #show results
mean(b$t) #bootstrap mean
bse = sd(b$t) #bootstrap SE
bse
bCI = boot.ci(b) #get set of confidence intervals
bCI
tcrit = qt(.975, nrow(seattle)-1) #critical t value for 95% confidence
tCI = c(trimdata - tcrit*bse, trimdata + tcrit*bse) #t-based C.I.
tCI
```

- `hist(b$t, breaks=20, freq=FALSE, main='Seattle RE Prices', xlab='Price',`
- `xlim=c(150,350), ylim=c(0,.030), col='blue') #show histogram`
- `lines(density(b$t)) #smooth curve`
- `segments(x0=trimdata, y0=0, x1=trimdata, y1=.025) #actual trimmed mean for sample`

Example 3: Comparing various Confidence Intervals.

See handout for solution in SPLUS and R and script in SPLUS

R SCRIPT CONTINUED:

```
hist(b$t, breaks=20, freq=FALSE, main='Seattle RE Prices', xlab='Price',
xlim=c(150,350), ylim=c(0,.030), col='blue') #show histogram
lines(density(b$t)) #smooth curve
segments(x0=trimdata, y0=0, x1=trimdata, y1=.025) #actual trimmed mean for sample
segments(x0 = tCI[1], y0=.025, x1 = tCI[2], y1 = .025)
text(320, .025, "t-based")
segments(x0 = bCI$percent[4], y0=.023, x1 = bCI$percent[5], y1 = .023)
text(320, .023, "percentile")
segments(x0 = bCI$bca[4], y0=.021, x1 = bCI$bca[5], y1 = .021)
text(320, .021, "BCa")
segments(x0 = bCI$normal[2], y0=.019, x1 = bCI$normal[3], y1 = .019)
text(320, .019, "Normal")
segments(x0 = bCI$basic[4], y0=.017, x1 = bCI$basic[5], y1 = .017)
text(320, .017, "Basic")
```

Example 4: Iowa Housing Prices

Bootstrap the correlation between selling price and square footage in the Ames, Iowa, housing data

a. Describe the bootstrap distribution.

b. Give a 95% confidence interval that is appropriate for these data. Explain yourchoice of interval.

c. State your conclusions from your analysis

See solutions and scripts on the handout

a. The observed value of the correlation was 0.8345, the bootstrap mean value was 0.8297,with a corresponding difference (bias) of –0.0049. This is less than 1% of the observed value. The bootstrap estimate of the standard error of estimated correlation was 0.0507, obviously larger than the bias.

The t-based 95% confidence interval using the estimated standard error is (0.733, 0.936), the same as that obtained by the limits.t() function.

The bootstrap percentile confidence interval is (0.704, 0.908).

The BCa interval is (0.703, 0.905).

The "tilting-t" interval from limits.tilt() is (0.711, 0.902).

Obviously the severe skewness and non-normality renders the simple t-based Interval untrustworthy.

Example 4: Iowa Housing Prices

The percentile and BCa intervals are quite consistent with each other, despite the expectation that the severe skewness might adversely affect the percentile interval. The "tilting-t"interval is a little different on the low limit. Part(b): All things considered, I'd go with the BCa interval of (0.703, 0.905) because it is typically accurate, agrees with the percentile method and is adjusts for skewness.

Example 5: Bootstrapping Variances

- *Problem 1: Pick a simple data set that you have used that involves a single set of independent identically distributed observations (that at least can be assumed to be such).*

a. Compute the sample variance.

b. Generate a bootstrap estimate of the variance using 1000 bootstrap replications. Compare these two estimates. Is the bootstrap estimate close to the sample estimate?

c. Now generate bootstrap confidence intervals [for the variance]. Do percentile method and BCa and if you know how to try the tilted interval as well. Do these intervals look different or similar?

See Handout for Results

Example 5: Bootstrapping Variances

- *Problem 2: Now do a simulation with a uniform distribution from [0,5]. The mean of the distribution is 2.5.*

a. Determine the variance of the distribution.

b. Generate 25 samples from the uniform distribution. Compute the sample variance and a bootstrap variance based on 1000 replications. Comparethese estimates to the true variance (which you computed for the Uniform distribution).

c. Now repeat the simulation 100 times [i.e., different size 25 samples each time]. Each time calcuate the bootstrap and sample estimates of the variance. Generate the percentile and the BCa confidence intervals two-sided 95% [for each sample]. For each type confidence interval you generate count how many times the intervals contain the true variance. Are they all close to 95%?

Look at the handout.

End of Course