Inferência em Amostras Pequenas: Métodos Bootstrap

Augusto Sousa da Silva Filho

[1]Programa Permanente de Capacitação Docente
FACULDADES ANHANGUERA

Abstract. *A amostra original representa a população da qual foi extraída. Dessa maneira, as reamostras obtidas a partir dessa amostra representam o que obteríamos se retirássemos diversas amostras da população. A distribuição bootstrap de uma estatística, baseada em um grande número de reamostras, representa a distribuição da estatística, com base em um grande número de amostras.*

Keywords: Bootstrap, Intervalos de Confiança, Erros Padrões; Média e Mediana.

1. Introdução

Existem métodos de estimação (MV, MMG) e testes de significância (RV,W,ML) que produzem estimadores e testes estatísticos com propriedades desejáveis de amostras grandes. Em amostras pequenas, gostaríamos de conhecer a performance relativa dos estimadores ou testes estatísticos que têm a mesma distribuição assintótica, e de saber quão confiável é a inferência assintótica. Há alguns resultados amostrais finitos que podem ser obtidos pela expansão Edgeworth, mais isso é muito complicado e está além do escopo deste trabalho.

Neste trabalho, veremos alternativas de reamostragem - métodos baseados em retirar sucessivamente amostras repetidas. Será discutido o método bootstrap e haverá um comentário sobre o método de Monte Carlo.

Os testes RV, W e ML tem distribuições assintóticas normal ou χ^2. Na prática, porém, não sabemos como é a performance desses testes em amostras pequenas. Muitos apresentam distorções de tamanho substanciais, isto é, pode-se testar ao nível de 5% de significância usando-se as distribuições assintóticas normal ou χ^2 , sendo que o verdadeiro nível de significância é 25%. Além disso, as performances de dois estimadores que têm a mesma distribuição assintótica normal podem ser diferentes em amostras pequenas.

Para examinar esses problemas, discutiremos dois métodos de reamostragem, ou métodos que dependem da retirada de amostras repetidas:

1 . Métodos Monte Carlo;
2 . Métodos Bootstrap.

Eles resolvem diferentes aspectos de inferência em amostras pequenas. O método 1 é usado para escolher entre, digamos, dois estimadores ou dois testes estatísticos. O método 2 **é usado para se obter a distribuição amostral pequena do estimador escolhido ou do teste estatístico.**

2. Método Monte Carlo

Nos métodos Monte Carlo[1] é possível escolher o modelo, considerar um tamanho amostral N, fixar os parâmetros em certos valores e retirar amostrar repetidas da distribuição do termo do erro. Isso gera os dados do tamanho N, a partir dos quais as estimativas dos parâmetros são calculadas. Repeti-se isso M vezes. Então, a distribuição das M estimativas dos parâmetros dará a distribuição amostral de tamanho N. Podemos então comparar essa distribuição amostral com a distribuição assintótica. Ou podemos obter as distribuições amostrais dos dois estimadores e compara-las.

Considere, por exemplo, o modelo de regressão:

$$y_t = \beta x_t + \mu_t \sim IN(0;\sigma^2)$$

onde x_t são valores dados. Esse não é o caso em que se faria um estudo Monte Carlo, mas o consideramos aqui para ilustrar o procedimento.

Consideremos $\beta = 4$, $\sigma^2 = 1,0$, e o tamanho amostral 50. Retiramos números aleatórios da distribuição $N(0,1)$. Muitos programas de computador (SAS; R, Minitab) têm geradores de números aleatórios. Chame isso $(u_1, u_2, \ldots, u_{50})$. Usando-os junto com x_t e $\beta = 0$, geramos os valores amostrais de $y : (y_1, y_2, \ldots, y_{50})$. Podemos usar esses valores amostrais para obter uma estimativa $\hat{\beta}$ de β. Repetimos isso M vezes (digamos M=1000). Então a distribuição de $\hat{\beta}_j = (j = 1, 2, \ldots, 1000)$ produz a distribuição do estimador dos mínimos quadrados $\hat{\beta}$. Veremos que $\overline{\beta} = \frac{1}{1000}\sum_1^{1000}\hat{\beta}_j$ estaria muito perto de 4,0 nesse caso.

Esse é um exemplo muito simples para ilustrar o procedimento. Conhecemos aqui a distribuição amostral pequena de $\hat{\beta}$, de maneira que não precisamos do estudo Monte Carlo. Entretanto, consideremos o modelo:

$$\begin{aligned} y_t &= \beta x_t + u_t \\ u_t &= \rho u_t - 1 + e_t \qquad e_t \sim IN(0,\sigma^2) \end{aligned}$$

e usaremos o estimador Cochrane-Orcutt em duas etapas. Primeiro obtemos a estimativa MQO em $\hat{u}_{t-1}$ para obtermos $\hat{\rho}$. Usamos isso para transformar os dados e chegar à estimativa MQG β^* de β.

Não conhecemos a distribuição de pequena amostra de β^*. A distribuição assintótica de β^* é normal com média 0 e variância $\left(x'\sum^{-1}x\right)^{-1}$, onde $x' = (x_1, x_2, \ldots, x_N)$ e $\sum$ é a matriz de covariância dos erros u_t. O experimento Monte Carlo procede da seguinte forma. Suponha que consideramos uma amostra de tamanho 50 e nos é dado $(x_1, x_2, \ldots, x_{50})$, que permanecem os mesmos em todas as replicações. Fixamos o valor dos parâmetros $(e_1, e_2, \ldots, e_{50})$ da distribuição $N(0,\sigma^2)$. Para gerar u_t, precisamos

[1] Este trabalho é conciso. Mais detalhes podem ser encontrado em D.F. Hendry, "Monte Carlo Experimentation in Economic", Cap. 16 em Z, Griliches e M.D. intrilligator) (eds.), Handbook of Econometrics, Vol 2. (Amsterdã: North-Holland, 1984), e Seção 3.6 de D.F. Hendry, Dynamic Econometrics (Oxford: Oxford University Press, 1995).

de um valor inicial u_0. É costume tomar u_t, que é $N\left(0; \sigma^2/(1-\rho^2)\right)$. Usando o $\hat{u}_t$, geramos agora y_t. Temos então uma amostra de tamanho $50 : (y_t, x_t), t = 1, 2, \ldots, 50$. Usamos o procedimento Cochrane-Orcutt para obtermos a estimativa MQG $\hat{\beta}_{MQG}$. Repetimos isso, digamos, 1000 vezes. Obtemos assim 1000 valores de $\hat{\beta}_{MQG}$ e isso produz uma distribuição amostral pequena de $\hat{\beta}_{MQG}$.

2.1. Métodos Monte Carlo Mais Eficientes

O procedimento descrito pode ser melhorado. Hendry discute dois métodos: o método de variáveis antitéticas e o método de covariâncias. Não veremos aqui detalhes[2]. O método de variáveis antitéticas utiliza novamente as amostras aleatórias retiradas dos erros, por exemplo, se $(e_1, e_2, \ldots, e_{50})$ é uma amostra aleatória dos erros, o mesmo vale para $(-e_1, -e_2, \ldots, -e_{50})$.

2.2. Superfícies de Resposta

Os resultados de Monte Carlo que descremos referem-se a um conjunto de valores paramétricos (θ) e a um tamanho amostral N. Por exemplo, obtemos o desvio (digamos B) de um estimador MQG de um valor paramétrico θ, tamanho amostral N. Podemos obter resultados de vários outros experimentos Monte Carlo de diferentes valores de θ e N. Uma maneira fácil de resumir esses resultados é calcular uma superfície de resposta[3].

Trata-se de uma regressão da forma:

$$B_j = \alpha + \beta_1 \theta_j + \beta_2 N_j + u_j$$

onde K é o número de diferentes valores de θ e N para os quais não conduzimos o estudo Monte Carlo. O método da superfície de resposta é uma maneira conveniente e útil de sumarizar resultados de vários experimentos Monte Carlo[4]. O método Monte Carlo que descrevemos pode ser utilizado para comparar a performance em amostras pequenas de diferentes estimadores ou testes estatísticos. Calculamos os diferentes estimadores e testes estatísticos com cada uma das amostras geradas.

3. Métodos de Reamostragem: Bootstrap e Jackknife

Na seção anterior, descremos o método Monte Carlo, no qual amostras repetidas eram retiradas das distribuições de erros assumidas. Há vários outros métodos em que amostras são retiradas da amostra dada. Dois deles são o jackknife, introduzido por Quenouille[5], e bootstrap, apresentado por Efron[6]. Nesses métodos, a informação da amostra

[2]Ver Hendry (1984) e Hendry (1995, Seções 3.6.2 e 3.6.3).

[3]Mais detalhes sobre a metodologia de superfície de resposta podem ser encontrados em Hendry (1984), citado anteriormente.

[4]Esse método foi utilizado por MacKinnon para resumir os resultados de vários experimentos Monte Carlo e obter valores críticos de testes de cointegracao. Ver J.G. MacKinnon, “Critical Values for Cointegration Tests”, em R.F. Engle e C.W.J. Granger (eds.), Long Run Economic Relationships (Oxford: Oxford University Press, 1991), pp. 267-276.

[5]M.Quenouille, “Notes on Bias in Estimation”, Biometrika, Vol 43, 1956, pp. 353-360.

[6]B. Efron, “Bootstrap Methods: Another Look at the Jackknife”, Annals of Statistics, Vol. 7,1979, pp. 1-26

é "reciclada" para se chegar às distribuições amostrais das estatísticas de interesse. O jackknife exclui um número de observações a cada rodada de cálculo, e o bootstrap faz a amostragem aleatória das observações amostrais a cada rodada de cálculo. O jackknife mais simples, que "exclui um", procede da seguinte forma: seja θ o estimador de $y_1, y_2, \ldots, y_n$, o conjunto de observações. Os passos da estimação jackknife são:

(1) Calcule $\hat{\theta}_i$ excluindo y_i do conjunto de observações;
(2) Calcule $p_i = n\hat{\theta} - (n-1)\hat{\theta}_i$;
(3) O estimador jackknife de θ é $\theta^* = \sum P_i/n$;
(4) O estimador da variância jackknife é $V^* = \sum(p_i - \theta^*)^2/n(n-1)$.

Métodos jacknife mais detalhados são discutidos em Wu[7]. O método bootstrap é outro método de reamostragem para o mesmo propósito que o jackknife: reduzir desvios e prover desvios padrão mais confiáveis. O método bootstrap funciona da seguinte maneira: seja $(y_1, y_2, \ldots, y_n)$ a amostra dada. Retira-se dessa amostra uma amostra de tamanho n com reposição. Chamaremos essa amostra de $B_j = (y_1^*, y_2^*, \ldots, y_n^*)$. Essa é amostra bootstrap. Cada y_i^* é uma escolha aleatória de $(y_1, y_2, \ldots, y_n)$. Fazemos isso para $j = 1, 2, \ldots, m$ e calculamos $\hat{\theta}_j$ de cada uma das amostras bootstrap β_j distribuição de $\hat{\theta}_j$ é a distribuição bootstrap do estimador θ. As estimativas bootstrap do desvio e da variância de θ são derivadas dessa distribuição bootstrap.

Observe que a reamostragem não adiciona nenhuma informação nova à amostra original. A vantagem dos métodos como o bootstrap é o resultado da maneira pela qual a informação amostra é processada. No caso da distribuição normal, toda informação sobre a média amostral é resumida na média amostral e na variância amostral. Logo, outras maneiras de processar a informação amostral não produzem melhores resultados nesse caso. São nos casos em que não há distribuição amostral finita das estatísticas prontamente disponível que o bootstrap é útil.

A distribuição bootstrap pode ser freqüentemente assimétrica. Nesse caso, não é suficiente olhar para a variância bootstrap. Varias aplicações antigas da econometria usaram o método bootstrap para obter a variância das estatísticas amostrais. Mesmo se os erros padrão assintóticos e bootstrap forem os mesmos em qualquer exemplo, os intervalos de confiança poderiam ser diferentes se a distribuição bootstrap fosse assimétrica. Para o modelo auto-regressivo $y_t = y_{t-1} + e_t$, baseado nos números sumspot de 1770-1889, Efron e Tibshirani[8] obtiveram $\hat{\rho} = 0,815$ com erro padrão assintótico 0,053. O desvio padrão bootstrap baseado em 1000 amostras bootstrap foi 0,055, concordando satisfatoriamente com o resultado assintótico. Porém, a distribuição bootstrap foi assimétrica para a esquerda, o que significa que os intervalos de confiança são diferentes.

3.1. Por que o bootstrap funciona?

Pode parecer que o bootstrap crie dados a partir do nada. Isso parece suspeito. Entretanto, não estamos utilizando as observações das reamostras como se elas fosse dados reais - o bootstrap não é um substituto para o acréscimo de dados com vistas ao aumento

[7]C.F. J. Wu, "Jackknife, Bootstrap and Other Re-sampling Methods in Regression Analysis", Annals of Statistics, Vol.14, 1986, pp. 1261-1295.

[8]B.Efron e R. Tibshirani, "Bootstrap Methods for Standard Errosrs, Confidence Intervals and Other Measures of Statistician Accuracy", Statistical Science, Vol. 1,1986, pp.54-77

da precisão. Em vez disso, a idéia do bootstrap é a de se empregarem as medias das reamostras para se estimar como a média amostral de uma amostra de tamanho N, extraída dessa população, varia em decorrência da amostragem aleatória.

A utilização dos dados duas vezes - uma vez para se estimar a média populacional e outra, para se estimar a variação das medias amostrais - é um procedimento perfeitamente legitimo. Já fizemos isso muitas vezes antes: por exemplo, quando calculamos tanto $\overline{x}$ quanto $s/\sqrt{n}$ a partir dos mesmos dados. O que temos de diferente agora é que:

(1) Calculamos um erro padrão utilizando a reamostragem, em vez da fórmula $s/\sqrt{n}$;
(2) Utilizamos a distribuição bootstrap para ver se a distribuição amostral é, ou não, aproximadamente Normal, em vez de simplesmente esperarmos que nossa amostra seja grande o suficiente para que o teorema central do limite se aplique;

A idéia do bootstrap também é válida para outras estatísticas além das medias amostrais. Para utilizarmos o bootstrap de maneira mais geral, lançamos mão de um outro principio - um que já aplicamos diversas vezes sem pensarmos muito sobre ele. Ele é conhecido como o **principio do Plug-in**. Este principio consiste em estimar um parâmetro, uma quantidade que descreve a população, utilizando a estatística que é a quantidade correspondente para a amostra.

O principio do plug-in sugere que a média populacional μ seja estimada por meio da média amostral $\overline{x}$, e que o desvio padrão populacional σ seja estimado pelo desvio padrão amostral s. Da mesma forma pode-se estimar a mediana populacional pela mediana amostral. Para estimarmos o desvio padrão $\sigma/\sqrt{n}$ da média amostral para uma AAS, aplicamos o principio do plug-in, empregando s na fórmula para obter . A idéia do bootstrap em si mesma é uma forma do principio do plug-in: substitua a distribuição populacional pela distribuição dos dados e, então extraia amostras (ou reamostras) que imitem o processo de construção de uma distribuição amostral.

4. Distribuição Amostral e distribuição bootstrap

Os intervalos de confiança, os testes de hipóteses e os erros padrões baseaim-se todos na idéia da *distribuição amostral* de uma estatística - a distribuição dos valores que essa estatística pode assumir em todas as amostras possíveis de mesmo tamanho extraídas da mesma população. Na prática, não podemos tomar um número muito grande de amostras aleatórias para construir a distribuição amostral. Em vez disso, utilizamos um atalho: se já começamos com um modelo para a distribuição populacional, as leis da probabilidade nos dizem (em algumas situações) qual é a distribuição amostral. Se a população tem uma distribuição Normal, então a distribuição amostral de $\overline{x}$ também é Normal.

Em diversas situações, não dispomos de qualquer modelo para a população. Nesses casos, não é possível apelarmos para a teoria da probabilidade, e não temos, também condições de extrair uma quantidade muito grande de amostral. Numa situação como essa, o bootstrap vem em nosso auxílio. Usamos a única de que dispomos como se fosse a população e dela extraímos diversas reamostras, para construirmos a distribuição bootstrap. Usa-se a distribuição bootstrap no lugar da distribuição amostral.

Na prática, não costuma ser exeqüível extraírem-se todas as reamostras possíveis. Realizamos o bootstrap utilizando cerca de 1000 reamostras escolhidas aleatoriamente.

Poderíamos estimar diretamente a distribuição amostral escolhendo aleatoriamente 1000 amostras de mesmo tamanho a partir da população original. Entretanto, é muito mais rápido e barato fazer o computador obter as reamostras a partir da amostra original, do que se selecionar diversas amostras da população. Mesmo se dispuséssemos de um grande orçamento, preferiríamos gastá-lo na obtenção de uma única amostra maior do que em diversas amostras menores. Uma amostra maior fornece uma estimativa mais precisa.

Na maioria dos casos, a distribuição bootstrap tem aproximadamente a mesma forma e dispersão da distribuição amostral, porém está centrada no valor da estatística original, e não no valor do parâmetro de interesse. O bootstrap permite-nos calcular os erros padrões originais das estatísticas para as quais não dispomos de fórmulas, bem como chegar a Normalidade para estatísticas que não podem ser manipuladas facilmente pela teoria.

5. Intervalos de confiança Bootstrap

Recordemos o já familiar intervalo de confiança t de uma amostra para a média de uma população Normal,

$$\overline{x} \pm t^* \frac{s}{\sqrt{n}}$$

Este intervalo se baseia na distribuição Normal da média amostral $\overline{x}$ e na fórmula $s/\sqrt{n}$ para o erro padrão de $\overline{x}$.

Quando uma distribuição bootstrap é aproximadamente Normal e tem um viés pequeno, podemos utilizar, essencialmente, os mesmos procedimentos com o erro padrão bootstrap a fim de obtermos um intervalo de confiança para qualquer parâmetro.

5.1. O intervalo de confiança $bootstrap - t^*$

Suponha que a distribuição bootstrap de uma estatística extraída de uma AAS de tamanho n seja aproximadamente Normal e que o viés seja pequeno. Um intervalo de confiança de nível C aproximado para o parâmetro que correspondente a essa estatística, segundo o princípio do plug-in, é

$$\text{estatistica} \pm t^* EP_{\text{boot, estatística}}$$

onde t^* é o valor crítico da distribuição $t(n-1)$ com área C entre $-t^*$ e t^*.

É possível calcular-se um intervalo de confiança bootstrap-t para qualquer parâmetro, fazendo-se o bootstrap da estatística correspondente (conforme o princípio do plug-in). Não precisamos de condições sobre a população, nem de conhecimento especial sobre a distribuição amostral da estatística. A natureza flexível e quase automática dos intervalos bootstrap-t é algo maravilhoso - entretanto, existe uma armadilha. Esses intervalos funcionam bem somente quando a distribuição bootstrap nos informa que a distribuição amostral é aproximadamente Normal e possui um viés pequeno. Como podemos saber se essas condições estão sendo suficientemente satisfeitas para que possamos confiar no intervalo de confiança?

Para isso, devemos averiguar por meio dos percentis do bootstrap. Os intervalos de confiança baseiam-se na distribuição amostral de uma estatística. Um intervalo de confiança de 95% começa delimitando os 95% centrais da distribuição amostral. Nos intervalos de confiança t, os valores t críticos representam um atalho para a delimitação desses 95% centrais, de maneira que nem sempre é adequado o emprego dos intervalos t. Uma maneira de se verificar se esses intervalos t (usando-se o bootstrap ou a fórmula para os erros padrões) são razoáveis e, portanto, compará-los com os 95% centrais da distribuição bootstrap, que são delimitados pelos percentis de ordem 2,5 e 97,5. Na distribuição bootstrap, o intervalo entre esses dois valores é freqüentemente utilizado como um intervalo de confiança por si só, sendo conhecido como um *intervalo de confiança percentil do bootstrap*.

6. Aplicação

Examplo 1 *Será mostrado a utilização do Método Bootstrap através de exemplos práticos. Utilizou-se o programa Minitab.*

A base de dados utilizada neste exemplo faz parte do programa S-Plus e está disponível em www.insightful.com/Hesterberg/bootstrap.

A Verizon é a empresa responsável (o termo legal é Distribuidora Titular de Telefonia Local, DTTL) por uma grande área da região leste dos Estados Unidos. Como tal, cabe a ela fazer o serviço de reparos para os clientes das demais companhias telefônicas dessa região (conhecidas como Distribuidoras Concorrentes de Telefonia Local, DCTL**). A Verizon estará sujeita a multas caso os tempos de reparo (tempo para resolver problemas nas linhas telefônicas) para os clientes das empresas concorrentes forem substancialmente maiores que os tempos para os seus próprios clientes. Isso é determinado por meio de testes de hipóteses, negociados junto à Comissão de Serviços Públicos (CSP).*

Comecemos nossa análise observando os próprios clientes da Verizon. A figura 1 mostra a distribuição de uma amostra aleatória de 1664 tempos de reparo, constantes do arquivo $eg18_001.txt$. Uma rápida olhada na distribuição revela que os dados estão longe de ter uma distribuição Normal. A distribuição tem uma longa cauda à direta (assimetria à direita).

Para os clientes da Verizon desta amostra, o tempo médio dos reparos foi $\overline{x} = 8,41$ horas. Essa é uma estatística extraída da uma única amostra aleatória (embora bastante grande). Se tomarmos mais amostras, a estatística $\overline{x}$ irá variar, e sua confiabilidade como estimador da média populacional μ depende de quanto ela varia de amostra para amostra.

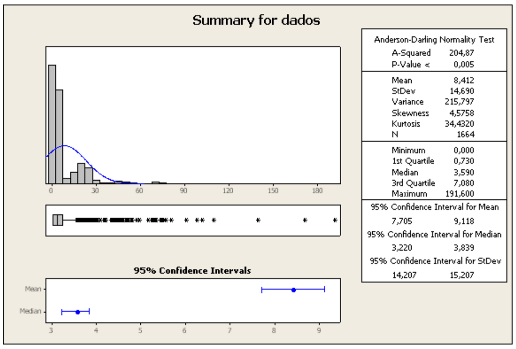

Figura 1. Histograma

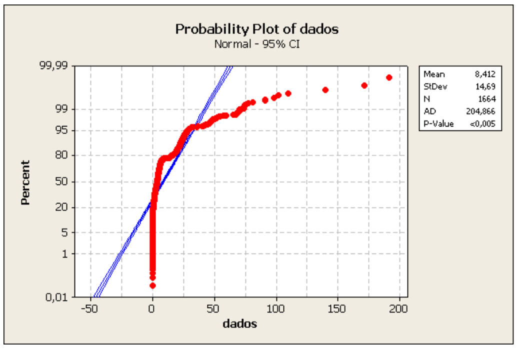

Figura 2. Probability Plot

O gráfico acima mostra a fuga clara à normalidade, apresentando uma grande assimetria.

O gráfico a seguir mostra a distribuição de 1000 médias de reamostras para os dados dos tempos de reparo da Verizon, utilizando um histograma e uma curva de densidade.

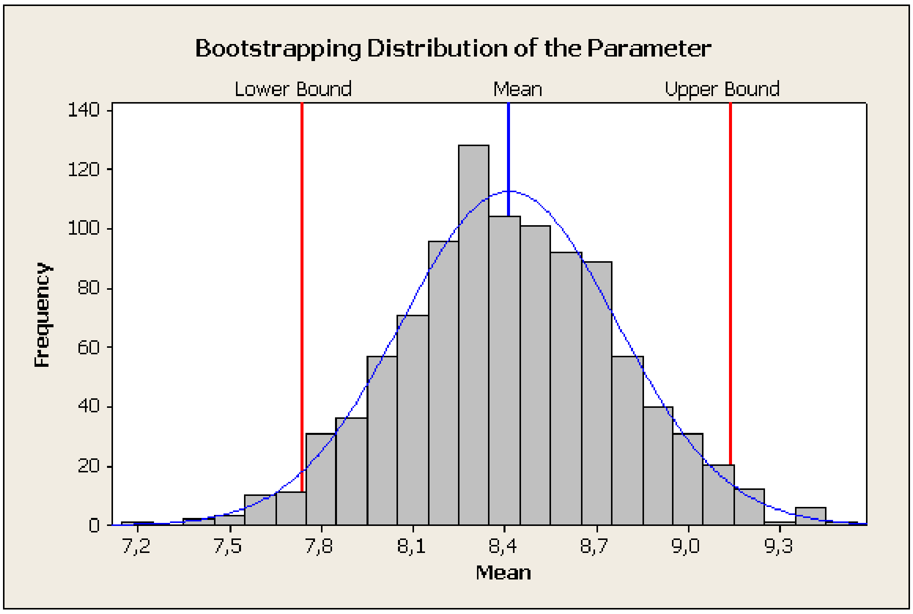

Figura 3. Simulação bootstrap de 1000 médias

Vemos que a distribuição bootstrap é aproximadamente Normal. O teorema central do limite diz que a distribuição amostral da média $\overline{x}$ é aproximadamente Normal se n for grande. Assim, a forma da distribuição bootstrap está próxima daquela que esperamos que a distribuição amostral tenha.

A distribuição bootstrap está centrada próximo à média da amostra original. Ou seja, como estimador da média da amostra original, a média da distribuição bootstrap apresenta um viés pequeno. Sabemos que a distribuição amostral de $overlinex$ está centrada na média populacional μ, ou seja, que $\overline{x}$ é um estimador não-viciado de μ. Dessa forma, a distribuição das reamostras de novo se comporta (a partir da amostral original) como esperaríamos que a distribuição amostral se comportasse (a partir da população).

Encontraremos também o intervalo de confiança para a média e para a mediana para as 1000 média reamostradas.

```
MTB > %intervalo.mac c1 1000 1 0,05
Executing from file: intervalo.mac

Bootstrap Confidence Interval

     The 95% Bootstrap Confidence Interval (Percentile Method)

     Mean      Lower Bound  Upper Bound
     8,41480   7,73949      9,13879
```

Para encontrarmos tal intervalo, utilizou-se uma macro que necessita de três informações: (b, est, alfa). Supondo que o conjunto de valores de interesse se encontra na célula C1 do aplicativo, temos que entrar com as seguintes informações: (b= número de interações). Neste exemplo utilizou-se um total de 1000 interações. A seguir, temos (est). O valor (1) representa que foi solicitado um intervalo de confiança para a média e o valor (2) indica a solicitação de um intervalo de confiança para a mediana. E o último valor de entrada é o nível de significância do teste. Neste exemplo, procuramos um intervalo ao nível de 95% de confiança.

A macro(Confidence Intervals for the Mean or Median using Bootstrap Methods) se encontra disponível na internet em: (http://www.minitab.com.au/support/macros/default.asp^x?action=code&id=108)

A seguir, encontramos o intervalo de confiança bootstrap para a mediana pelo método dos percentis.

```
MTB > %intervalo.mac c1 1000 2 0,05
Executing from file: intervalo.mac

Bootstrap Confidence Interval

     The 95% Bootstrap Confidence Interval (Percentile Method)

     Median   Lower Bound  Upper Bound
     3,6      3,22         3,82
```

O intervalo de confiança para o valor mediano é de 3,22 a 3,82. Foi utilizado o método dos percentis, com um nível de confiança de 95%.

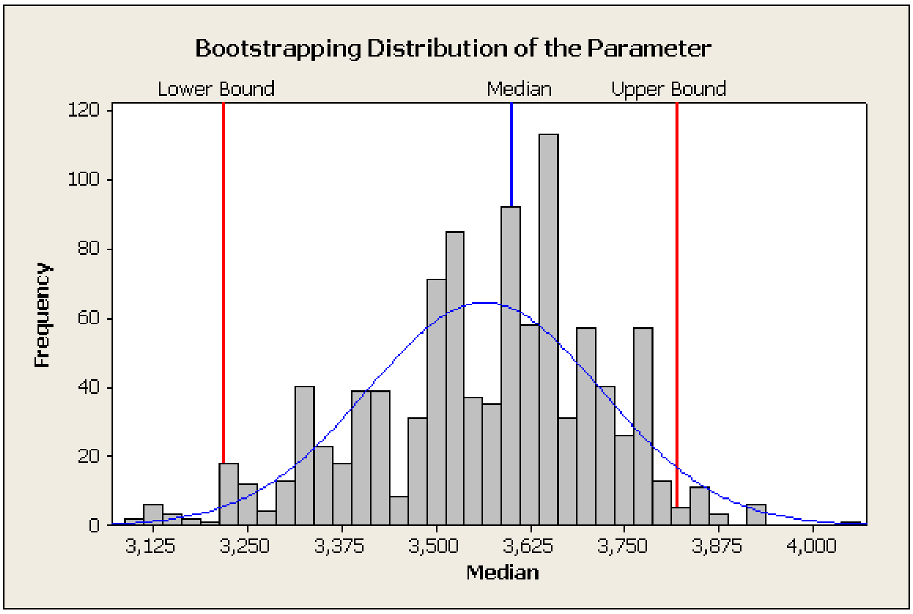

Figura 4. Simulação bootstrap

Examplo 2 *Considere uma amostra qualquer (por exemplo, a amostra MEDIDAS.MTW, disponível em ftp://ftp.est.ufmg.br/pub/fcruz/pacotes/MEDIDAS.MTW.*

```
original    media_o reamostra   media_r media
9,2980
9,3938
11,3871
9,4259
10,8253
.
.
.
```

Suponha que você queira fazer inferência sobre a média da população correspondente, mas como a amostra é muito pequena, decide usar a técnica de bootstrap (uma técnica de reamostragem), para melhorar a sua estimativa.

Iremos construir uma macro que (i) extraia 5 amostras desta amostra (na prática são necessárias umas 200), de igual tamanho, com reposição, (ii) calcule a media de cada uma destas 5 amostras e (iii) disponibilize a diferença entre duas vezes a média da amostra original e média das médias das reamostras (a estimativa melhorada).

Suponha que você queira fazer inferência sobre a média da população correspondente, mas como a amostra é muito pequena, decide usar a técnica de bootstrap (reamostragem), para melhorar a sua estimativa.

Construa uma macro que (i) extraia 5 amostras desta amostra, de igual tamanho, com reposição, (ii) calcule a media destas 5 amostras e (iii) disponibilize a média destas médias (a estimativa melhorada), $x_b = 2\overline{x} - \overline{x}^*$.

```
MTB > base 1000
MTB > %frederico.mac
Executing from file: frederico.mac

Data Display

media
   9,85763
```

No Minitab, obteve-se os seguintes resultados.

```
original     media_o      reamostra    media_r      media
9,2980       9,90392      11,3871      9,9908       9,85763
9,3938                    9,3925       10,0543
11,3871                   11,3871      9,6418
9,4259                    9,2980       10,1064
10,8253                   9,3938       9,9578
.
.
.
```

Logo, temos a média da amostra original $\overline{x} = 9,903293$ e a média das médias das reamostras 9,85763.

7. Conclusão

Neste trabalho verificou-se que para se fazer o bootstrap para uma estatística (por exemplo a média amostral), deve-se retirar centenas de reamostras com reposição a partir da amostra original e calcular a estatística em questão para cada reamostra e inspecionar a distribuição bootstrap das estatísticas dessas reamostras.

Procuramos aplicar a metodologia bootstrap a exemplos práticos e observamos que a distribuição bootstrap aproxima-se da distribuição amostral da estatística. Isso é um exemplo do princípio do plug-in. Em geral, as distribuições bootstrap possuem aproximadamente a mesma forma e dispersão da distribuição amostral, porém está centrada na estatística (dos dados originais), ao passo que a distribuição amostral está centrada no parâmetro da população.

Na análise do exemplo Verizon, constatou-se que o bootstrap não é um substituto para o acréscimo de dados com vistas ao aumento da precisão. Em vez disso, a idéia do bootstrap é a de se empregar as médias das reamostras para se estimar como a média amostral de uma amostral de tamanho 1664, extraída dessa população, varia em decorrência da amostra aleatória.

A técnica de bootstrap tenta realizar o que seria desejável realizar na prática, se tal fosse possível: repetir a experiência. As observações são escolhidas de forma aleatória e as estimativas re-calculadas.

Referências

[1] FERREIRA, Daniel Furtado. (2005). *ESTATÍSTICA BÁSICA.* , 1 ed. Editora LAVRAS, 2005.

[2] MOORE, David S (2006). *THE PRACTICE OF BUSINESS STATISTICS: Using data for decisions.*, 1a. ed. LTC, Rio de Janeiro, 2006.

[4] J.L. Horowitz. (1997). *"Bootstrap Methods in Econometrics: Theory and Performance"; em D.M Kreps e K.F. Wallis (eds.), Advances in Economics and Econometrics: Theory and Applications.*,7ed. Iowa State University Press, Amer. Iowa.

[4] J. Jeong e G.S. Maddala, *"A Perspective on Application of Bootstrap Methods in Econometrics"*, em G. S. Maddala, C.R. Rao e H. D. Vinod (eds.), Handbook of Statistics, Vol. 11 (Amsterdã: Elsevier Science, 1993), pp. 573-610.