4
Reamostragem

O tipo de estatística não-paramétrica que foi ensinado no passado desempenhou um importante papel na análise de dados que não são contínuos e, portanto, não podem empregar a distribuição normal de probabilidade para fazer estimativas de parâmetros e de intervalo de confiança. Mas existe uma nova perspectiva sobre estimação não-paramétrica que também se relaciona com estimação de parâmetros e de intervalo de confiança para variáveis no mínimo em escala intervalar.

Com isso, não se tem que assumir que o intervalo de confiança para um parâmetro segue a distribuição normal. Pode-se até mesmo gerar intervalos de confiança para parâmetros como a mediana, o que geralmente é difícil de avaliar com as técnicas de inferência paramétrica tradicionais.

Essa abordagem não-paramétrica é conhecida como reamostragem e tem conquistado apoio como uma alternativa aos métodos clássicos de inferência paramétrica.

A reamostragem descarta a distribuição amostral assumida de uma estatística e calcula uma distribuição empírica – a real distribuição da estatística ao longo de centenas ou milhares de amostras.

Com a reamostragem, não se tem que confiar na distribuição assumida nem se tem que ser cuidadoso quanto à violação de uma das suposições inerentes. Pode-se calcular uma real distribuição de estatísticas da amostra e pode-se agora ver onde o 95 ou o 99 percentil estão realmente, acreditando-se que a mostra original seja confiável.

Mas de onde vêm as múltiplas amostras? É necessário reunir amostras separadas, aumentando sensivelmente o custo de coleta de dados? Ao longo dos anos estatísticos desenvolveram diversos procedimentos para criar as múltiplas amostras necessárias para a reamostragem *a partir da amostra original*.

Agora uma amostra pode gerar um grande número de outras amostras que podem ser empregadas para gerar a distribuição amostral empírica de uma estatística de interesse.

4.1
Conceitos Básicos em Reamostragem

Reamostragem, contudo, não usa a distribuição de probabilidades assumida, mas ao invés disso ela calcula uma distribuição empírica de estatísticas estimadas. Criando múltiplas amostras da amostra original, a reamostragem agora precisa apenas do poder computacional para estimar um valor de uma estatística para cada amostra. Logo que eles estejam todos calculados, pode-se realizar o teste de normalidade dos valores e até mesmo construir intervalos de confiança e realizar testes de hipóteses.

A reamostragem engloba diversos métodos. Para esta tese de doutorado, se estudará e aplicará as de *bootstrap* e *jackknife*.

4.2
Métodos de Reamostragem

Uma diferença chave entre os vários métodos de reamostragem é se as amostras são extraídas com ou sem reposição. A amostragem com reposição obtém uma observação a partir da amostra e então a coloca de volta na amostra para possivelmente ser usada novamente. A amostragem sem reposição obtém observações da amostra, mas uma vez obtidas eles não estão mais disponíveis.

O verdadeiro poder da reamostragem vem de amostragem **com reposição.** Pesquisas têm mostrado que esse método fornece estimativas diretas dos intervalos de confiança e valores-p, apesar de haver modificações nos métodos simples para obtenção dos intervalos de confiança.

4.3
Jackknife versus *Bootstrap*

Os métodos *jackknife* e *bootstrap* diferem na maneira como eles obtém a amostra.

O método *jackknife* computa **n** subconjuntos(**n**=tamanho da amostra) pela eliminação seqüencial de um caso de cada amostra. Assim cada amostra tem um tamanho de n – 1 e difere apenas pelo caso omitido em cada amostra.

Apesar de o método *jackknife* ter sido ultrapassado pelo *bootstrap* como um eficiente estimador de intervalos de confiança e cálculos de significâncias, ele continua como uma medida viável de observações influentes (uma observação que exerce uma influência desproporcional sobre um ou mais aspectos das estimativas e essa influência pode ser baseada em valores extremos das variáveis) e uma opção para muitos pacotes estatísticos.

O método *bootstrap* obtém sua amostra via amostragem com reposição da amostra original. A chave é a substituição das observações após a amostragem, o que permite ao pesquisador criar tantas amostras quanto necessárias e jamais se preocupar quanto à duplicação de amostras, exceto quando isso acontecer ao acaso. Cada amostra pode ser analisada independentemente e os resultados compilados ao longo da amostra. Por exemplo, a melhor estimativa da média é exatamente a média de todas as médias estimadas ao longo das amostras.

O intervalo de confiança também pode ser diretamente calculado. As duas abordagens mais simples:

1. Calculam o erro padrão simplesmente como o desvio padrão das estimativas estimadas;

2. Literalmente ordenam as estimativas e definem os valores que contém os 5% extremos (ou 1%) dos valores estimados.

4.4
Limitações

Apesar de procedimentos de reamostragem não serem restritos por quaisquer suposições paramétricas, eles ainda têm certas limitações:

1. A amostra deve ser grande o bastante e obtida (a princípio aleatoriamente) de forma a ser representativa da população completa. Técnicas de reamostragem não podem conter quaisquer enviezamentos que traga como conseqüência uma amostra não representativa;
2. Métodos paramétricos são melhores em muitos casos para fazer estimativas pontuais. Os procedimentos de reamostragem podem completar as estimativas pontuais de métodos paramétricos fornecendo as estimativas de intervalos de confiança;
3. As técnicas de reamostragem não são adequadas para identificar parâmetros que têm um domínio amostral muito estreito, como os valores mínimos e máximos. A reamostragem funciona melhor quando a distribuição inteira é considerada para obter o parâmetro em análise.

Os procedimentos de reamostragem que foram discutidos neste capítulo fornecem uma perspectiva alternativa sobre uma das avaliações-chaves feitas na análise de dados: a variabilidade da estatística estudada. Essa é a base do teste de hipótese e da avaliação de significância estatística. Técnicas de reamostragem aumentam a habilidade do pesquisador para examinar a real distribuição dos estimadores tratados, ao invés de confiar plenamente na distribuição assumida. Técnicas como essas fornecem um modo direto para "conhecer seus dados" e evitar a armadilha muito comum de se tornar muito confiante em técnicas estatísticas ao invés de raciocinar sobre o que se sabe a respeito dos dados.

Nas próximas seções, se abordarão com mais detalhes as duas técnicas de reamostragem exemplificadas neste capítulo.

4.5

O Método *Jackknife*

O *jackknife* é um método não paramétrico destinado a estimar o enviezamento e, portanto reduzi-lo, e a variância de estimadores em condições teoricamente complexas ou em que não se tem confiança no modelo especificado.

Foi introduzido por Quenouille em 1949, retomado por Tukey em 1958, e desenvolvido na última década. Tal como o *bootstrap* é um método de reamostragem, pois se baseia na construção de subamostras da amostra original.

4.5.1

O Procedimento para Obtenção da Amostra *Jackknife*

1º) Seleciona-se uma amostra original de tamanho "***n***":

$$x = \{x_1, x_2, x_3, \ldots, x_{n-1}, x_n\}$$

2º) Define-se a estatística de interesse:

$$\hat{\theta} = F(x)$$

3º) Gera-se a amostra *jackknife* 1:

$$x^{(1)} = \{x_2, x_3, \ldots, x_{n-1}, x_n\}$$

$$\hat{\theta}_{(1)} = F[x^{(1)}]$$

4º) Gera-se a amostra *jackknife* 2:

$$x^{(2)} = \{x_1, x_3, \ldots, x_{n-1}, x_n\}$$

$$\hat{\theta}_{(2)} = F[x^{(2)}]$$

e assim por diante...

5º) Gera-se a amostra *jackknife* n-1:

$$x^{(n-1)}=\{x_1,x_2,...,x_{n-2},x_n\}$$

$$\hat{\theta}_{(n-1)} = F[x^{(n-1)}]$$

6º) Gera-se a amostra *jackknife* n:

$$x^{(n)}=\{x_1,x_2,...,x_{n-1}\}$$

$$\hat{\theta}_{(n)} = F[x^{(n)}]$$

7º) Estima-se o erro padrão da estatística do 2º passo através da expressão:

$$\hat{S}_{Jack} = \frac{n-1}{n} \sum_{i=1}^{n} \left\{ [\hat{\theta}_{(i)} - \hat{\theta}_{(.)}]^2 \right\}^{1/2}, \quad \text{sendo } \hat{\theta}(.) = \sum_{i=1}^{n} \hat{\theta}_{(i)}/n$$

Onde **θ** é o valor que a estatística ser assume na amostra original.

Este processo multiplica a informação inicial, pois para além das estimativas de **θ**, obtêm-se mais ***n*** valores para a mesma estatística. É a partir destes valores que serão determinadas às estimativas *jackknife*.

É de notar que, para que se preserve uma certa qualidade das estimativas, o *Jackknife* só deve ser aplicado à estatística funcional. Uma estatística funcional é aquela obtida junto a estimadores que são funções dos elementos da amostra.Contudo, também se obtém bons resultados na estimativa do erro padrão para estimativas não funcionais, desde que sejam definidas simetricamente nos seus ***n*** argumentos (invariantes a qualquer permutação das variáveis).

As propriedades do enviezamento *jackknife*, tal como do seu erro padrão, dependem da estatística considerada. Sabe-se que quando a estatística é uma

funcional quadrática, a estimativa do enviezamento funcional é uma estimativa centrada do verdadeiro enviezamento.

4.6
O Método *Bootstrap*

O *bootstrap*, introduzido por Efron no final dos anos 70, vem historicamente na linha do *jackknife* e pode-se dizer que é uma técnica não paramétrica que procura substituir a análise estatística teórica (insuficiente em muitos casos como se exemplifica na utilização da análise fatorial) pela força bruta da computação, cada vez mais acessível e menos dispendiosa.

A terminologia, introduzida por Efron (1979), é basicamente uma técnica de reamostragem, que permite aproximar a distribuição de uma função das observações pela distribuição empírica dos dados baseada em uma amostra de tamanho finita. A amostragem é feita, com reposição, da distribuição da qual os dados são obtidos, se esta é conhecida (*bootstrap* Paramétrico) ou da amostra original (*bootstrap* não-paramétrico). Neste último caso supõe-se que as observações são obtidas da função de distribuição empírica F(x), que designa uma massa de probabilidade igual *1\n* para cada ponto amostral.

O *bootstrap* aborda o cálculo do intervalo de confiança de parâmetros e cálculos de valores-p, em circunstâncias em que outras técnicas não são aplicáveis, em particular no caso em que o número de amostras é reduzido.

Esta técnica foi extrapolada para resolução de muitos problemas de difícil resolução através de técnicas de análise estatística tradicionais, baseadas na hipótese de um elevado número de amostras.

A técnica *bootstrap* tenta realizar o que seria desejável na prática, se tal fosse possível: ***repetir a experiência***.

As observações são escolhidas de forma aleatória e as estimativas re-calculadas.

A idéia básica da técnica *bootstrap* é: uma vez que não se dispõe de toda a população de amostras (observações) faça-se o melhor com o que se dispõe, que é o conjunto de amostrado: $x = (x_1,..., x_n)$.

A técnica *bootstrap* trata a amostra original como se esta representasse exatamente toda a população (conjunto de experiências, realizações).

Segundo Isabel Proença (1988), até agora o *bootstrap* chega aos mesmos resultados que o processo tradicional, baseado na máxima verossimilhança. A sua grande virtude consiste em apresentar solução para casos em que a dedução da precisão da estimativa, de seu viés e do EMQ aparenta ser impossível ou mesmo demasiado complexa. É especialmente para estes casos que o *bootstrap* se vocaciona.

4.6.1
O Procedimento para Obtenção da Amostra *Bootstrap*

Seja uma amostra original e a estatística de interesse abaixo:

$x = \{x_1, x_2, x_3,..., x_{n-1}, x_n\}$.

$\hat{\theta} = F(x)$

1º) Geram-se as amostras *bootstrap* $x_{(1)}, x_{(2)}, x_{(3)},... , x_{(n^*)}$ com reposição de x.

2º) Calculam-se as estimativas da estatística de interesse:

$\hat{\theta}_{(b)} = F[x_{(b)}]$, b=1, ..., B

3º) Calcula-se o erro padrão *bootstrap*, $\hat{S}_{boot}$, dado por:

$$\hat{S}_{boot} = \frac{1}{B-1} \left\{ . \sum_{b=1}^{B} [\hat{\theta}_b - \hat{\theta}_{(*)}]^2 \right\}^{1/2}, \text{ sendo}$$

$$\hat{\theta}_{(*)} = \frac{\sum_{b=1}^{B} \theta_{(b)}}{B}$$

O procedimento acima se aplica ao caso do *bootstrap* não-paramétrico. Opta-se por utilizar o *bootstrap* paramétrico, procede-se da mesma forma, com a única diferença de que cada amostra *bootstrap* é obtida da distribuição paramétrica que originou os dados que se tem em mãos, ao invés de reamostrar-se as observações disponíveis.

Desde o aparecimento do *bootstrap*, vários autores vêm tentando estabelecer confirmação empírica ou teórica da sua validade. Devido ao fato desta técnica atuar como um método de aproximação de distribuições, todas as provas de consistência e precisão dos estimadores são resultados assintóticos, mas sua validade pode ser estendida para espaços amostrais finitos. Para um estudo mais detalhado dos teoremas e demonstrações pode-se consultar Shao e Tu (1995), que oferecem um apanhado geral dos trabalhos que surgiram sobre o assunto. As conclusões a que chegaram estes autores são que a aproximação *bootstrap* é válida para a maioria das estatísticas de interesse e que seus estimadores são consistentes.

Se $B \rightarrow \infty$, então as estimativas do erro-padrão, do enviesamento e do EMQ se igualam às estimativas de máxima verossimilhança (Efron, 1982). Para o cálculo das estimativas *bootstrap* geralmente é suficiente um valor de B=100. Contudo, para se determinar à distribuição por amostragem com precisão deve considerar-se um valor para B substancionalmente mais elevado. Segundo Efron (1982), geralmente B = 1000 proporciona bons resultados. E em ambos os casos, convém ensaiar diferentes valores para B até se verificar a convergência dos resultados.