Estimativas Bootstrap para o Enviezamento, Desvio-padrão e Intervalo de Confiança para a Taxa Média de Incidência de Dengue no Estado do Rio de Janeiro

Giovani Glaucio de Oliveira Costa
Universidade Federal Rural do Rio de Janeiro
Instituto Multidisciplinar
Rua Professor Paris S/N. Centro. Nova Iguaçu. Rio de Janeiro. Brasil
giovaniglaucio@ufrrj.br

Resumo:

A taxa de incidência é uma medida estatística que mede risco, no caso do presente trabalho, o risco de se adoecer por dengue no estado do Rio de Janeiro. A razão de se preferir utilizar a taxa e não o número absoluto, é que com a taxa podem-se fazer comparações entre populações de diferentes tamanhos (com números desiguais de pessoas sob risco). A dengue reemergiu no estado do Rio de Janeiro em 1986, e a partir deste ano a doença se tornou endêmica, apresentando em seguida característica de epidemia. Este trabalho objetivou, então, estimar a taxa média total de incidência da dengue no estado do Rio de Janeiro, através dos métodos ***CIS*** (Computer Intensive Statistics), computação intensiva, com a aplicação da técnica de reamostragem *Bootstrap*. Como resultado, tem-se uma chance de 95% do intervalo *[13,61% a 53,95%]* conter o percentual de risco de se adoecer por dengue no Estado do Rio de Janeiro. Uma estatística expressiva, alarmante e que se configura uma tendência de epidemia da doença no estado e que pode orientar as autoridades na intensidade em que promoverão medidas preventivas e de erradicação da doença no estado.

Palavras-chaves: taxas de incidência de dengue no Rio de Janeiro, estimação, bootstrap.

1-Introdução

A taxa de incidência é uma medida estatística que mede risco, no caso do presente trabalho, o risco de se adoecer por dengue no Estado do Rio de Janeiro. A razão de se preferir utilizar a taxa e não o número absoluto, é que com a taxa podem-se fazer comparações entre populações de diferentes tamanhos (com números desiguais de pessoas sob risco).

A dengue reemergiu no estado do Rio de Janeiro em 1986, e a partir deste ano, a doença se tornou endêmica apresentando em seguida anos epidêmicos. A média das taxas média de incidência em anos não epidêmicos era de 27 casos/100.000 hab., já a média dos anos epidêmicos é de 470 casos/100.000hab.

Dada a importância de se divulgar dados científicos sobre o panorama atual da epidemia no estado, este artigo tem o objetivo de apresentar uma estimativa intervalar mais atualizada da taxa média de dengue no estado do Rio de Janeiro tomando como amostra inicial os bairros do município do Rio de Janeiro, obtida através de métodos ***CIS***(Computer Intensive Statistics),computação intensiva, com a aplicação da técnica de reamostragem *Bootstrap*.

Estão em curso inúmeras investigações sobre a teoria *Bootstrap,* nomeadamente no que toca à validade assintótica e à aplicação na construção de intervalos de confiança. A importância do tema do estudo ora proposto pode avaliar-se pela quantidade de artigos que, nos últimos dez anos, estão aparecendo em todas as revistas da especialidade.

2-Base de Dados

A base de dados que servirá para realizar as simulações bootstrap foi retirada da "Tabela de Número de Casos e Taxa de Incidência de Dengue por Áreas de Planejamento, Regiões Administrativas e Bairros" da Coordenação de Programas de Epidemiologia do Município do Rio de Janeiro no ano de 2005.

A amostra é composta pelos 158 bairros do município do Rio de Janeiro e serve de base para a estimativa dos casos de dengue para o estado do Rio de Janeiro. A representatividade da amostra é discutível, mas pode-se notar que existem muitos

bairros do município do Rio de Janeiro com perfil de infra-estrutura e condições sócio-econômicos semelhantes à de bairros de outros municípios do estado.

Taxas de Incidência por 100.000 Habitantes de Casos de Dengue dos Bairros do Município do Rio de Janeiro-2005

183.0	0.0	29.4	14.6	22.5	7.7	0.0	14.0	14.2
0.0	15.8	19.8	9.0	36.3	0.0	3.0	9.4	11.7
41.6	0.0	14.8	39.1	21.3	49.0	0.0	0.0	0.0
11.3	0.0	21.2	0.0	25.6	30.1	8.1	4.6	2.4
20.4	12.2	10.9	7.9	13.8	17.2	13.9	0.0	12.8
15.5	15.0	15.0	0.0	5.9	0.0	0.0	21.0	13.8
17.2	39.6	19.3	7.3	11.8	0.0	1.5	22.9	0.0
0.0	4.3	0.0	9.8	0.0	0.0	4.4	0.0	9.1
14.5	23.0	10.2	15.6	4.3	21.8	6.0	9.0	8.1
13.0	13.8	22.9	13.1	5.1	6.7	0.0	0.0	10.4

11.0	0.0	17.3	44.6	76.3	0.0	1.7
5.7	8.9	18.6	97.4	0.0	0.0	0.0
0.0	9.7	3.7	31.1	36.7	0.0	1.7
24.7	0.0	11.0	531.4	46.6	17.8	0.0
12.0	6.2	6.2	33.3	1526.7	8.0	5.7
6.9	7.3	4.1	44.3	407.4	4.0	0.0
11.3	4.5	0.0	73.6	107.5	3.1	10.3
31.9	13.9	11.6	40.0	55.9	7.8	1.2
5.8	4.3	3.3	26.8	0.0	4.5	1.0
6.3	6.1	0.0	22.1	0.0	11.7	-

3-Metodologia da Pesquisa

Como comentado na introdução, este trabalho objetiva propor um processo inferencial para a taxa média de incidência de dengue do estado do Rio de Janeiro.

A idéia é utilizar técnicas *CIS* (Computer *Intensive Statistics*), que cogitam o modelo de densidade de probabilidade e que explica o comportamento aleatório da estatística observada e seus parâmetros característicos.

As técnicas ***CIS*** dispõem-se principalmente de dois métodos que serão empregados no estudo referido: o *bootstrap* e o *jackknife*.Este trabalho trata da estimação *Bootstrap.*

Através deste artigo é especificado o viés, erro-padrão e o intervalo de confiança para a taxa média de dengue no estado para o procedimento *bootstrap*.Com estes resultados, pode-se obter um procedimento computacional, um algoritmo, para a construção de intervalos de confiança e testes de hipóteses para as estimativas obtidas.

As taxas de incidência de dengue para os bairros do estado e do município do Rio de Janeiro são nomeados e classificados com grande freqüência, mas sempre a nível descritivo, já que o desconhecimento da distribuição por amostragem exata da variável aleatória taxa de incidência de dengue torna inviável fazer acompanhar as estimativas do respectivo erro padrão, para não falar na construção de intervalo de confiança ou na realização de testes de significância.

A opção de se usar as metodologias ***CIS*** surge quando não se conhece o viés e/ou o desvio padrão teórico das estimativas e/ou quando o modelo de distribuição de probabilidade destas estimativas não se adere à curva normal de probabilidades , o que acontece em algumas das estatísticas paramétricas das ciências biológicas.

Nestes casos, com a aplicação do *bootstrap* é possível obter, de forma expedita, através das computações "pesada", estimativas do desvio padrão e do viés da estatística em causa em substituição análise teórica. Com o *bootstrap*, por exemplo, é possível determinar a distribuição por amostragem da estatística e seus parâmetros característicos.O método *bootstrap* permite ladear a insuficiência da teoria da amostragem que se faz sentir em diversos estudos de estimação .

4-Resumo Teórico de Reamostragem

O tipo de estatística não-paramétrica que foi ensinado no passado desempenhou um importante papel na análise de dados que não são variáveis contínuas, em escala nominal ou ordinal, 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 amostra 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.

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 este trabalho, se estudará e aplicará o B*ootstrap*.

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, apesar de ter havido avanços nos métodos simples para obtenção dos intervalos de confiança.

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.

Matematicamente a obtenção da amostra bootstrap e suas estimativa do erro padrão é obtida da seguinte maneira:

Seja uma amostra original e a estatística de interesse abaixo:

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

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

(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:

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

(3º) Calcula-se o erro padrão *bootstrap*, S_{boot}, dado por:

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

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

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.

5-Estudo de Caso

Como ilustração da performance do *Bootstrap,* elaborou-se um exemplo numérico onde se aplica esta técnica à taxa média de incidência de dengue para o estado do Rio de Janeiro, contando com uma amostra original de 158 bairros do município. O cálculo de estimativas do desvio-padrão e do viés desta estatística, bem como do intervalo de confiança, assim como a determinação da sua distribuição por amostragem, só foi possível com o método *Bootstrap*, dado que o desconhecimento das respectivas expressões teóricas e seu modelo de probabilidade invalida a aplicação da estimação tradicional.

A estatística em foco é a taxa total média da incidência dos casos de dengue nos bairros do município do Rio de Janeiro por 100.000 habitantes.

A aplicação do Bootstrap foi feita de acordo com as etapas descritas na seção 4 para obtenção da amostra *Bootstrap*.No *Bootstrap,* utilizou-se o procedimento da amostragem com reposição descrito no texto, considerando 1000 réplicas de cada amostra de 158 bairros,isto é, B=1000 e n=158.

A computação das estimativas nas 1000 subamostras foi realizada através do pacote estatístico *Stata Versão 8.0.*

Tabela 1

Estimativas do viés e do Erro-padrão Bootstrap:

Variável	**Réplicas**	**Média da Taxa Média *Método Tradicional***	**Média da Taxa Média Total *Bootstrap***	**Viés**	**Erro-padrão**
Taxa Média Total de Incidência de Dengue	1000	29,95	29,99	0,04	10,84

Tabela 2

Estimativas do Intervalo de Confiança de 95% Bootstrap:

Intervalos de Confiança	Limites de Confiança	
	Limite Inferior	Limite Superior
Percentílico	13,61	53,94
Normal	8,67	51,22

A estimativa pontual bootstrap para média da taxa média total de incidência de dengue no estado do Rio de Janeiro é de 29,99%, mas esta estimativa não permite estabelecer a precisão do processo inferencial. A tendência da estimação foi de 0,04, relativamente baixa e a precisão pode ser medida pelo erro-padrão da taxa média total , que resultou em 10,84, relativamente alto, o que denota instabilidade, grande variação das estimativas nas 1000 simulações. Têm-se duas opções para o intervalo de confiança para a taxa média total para a incidência de dengue no estado: o percentílico, que não se baseia em nenhum modelo de probabilidade teórica para a distribuição de amostragem e o baseado na hipótese da normalidade das taxas médias totais estimadas das 1000 réplicas obtidas.Pela observação do histograma abaixo, percebe-se claramente a assimetria da distribuição por amostragem das estimativas em foco, não semelhante com a curva normal,o que indica assumir o intervalo de confiança percentílico para a estimação da taxa média total de incidência da de dengue no estado do Rio de Janeiro.

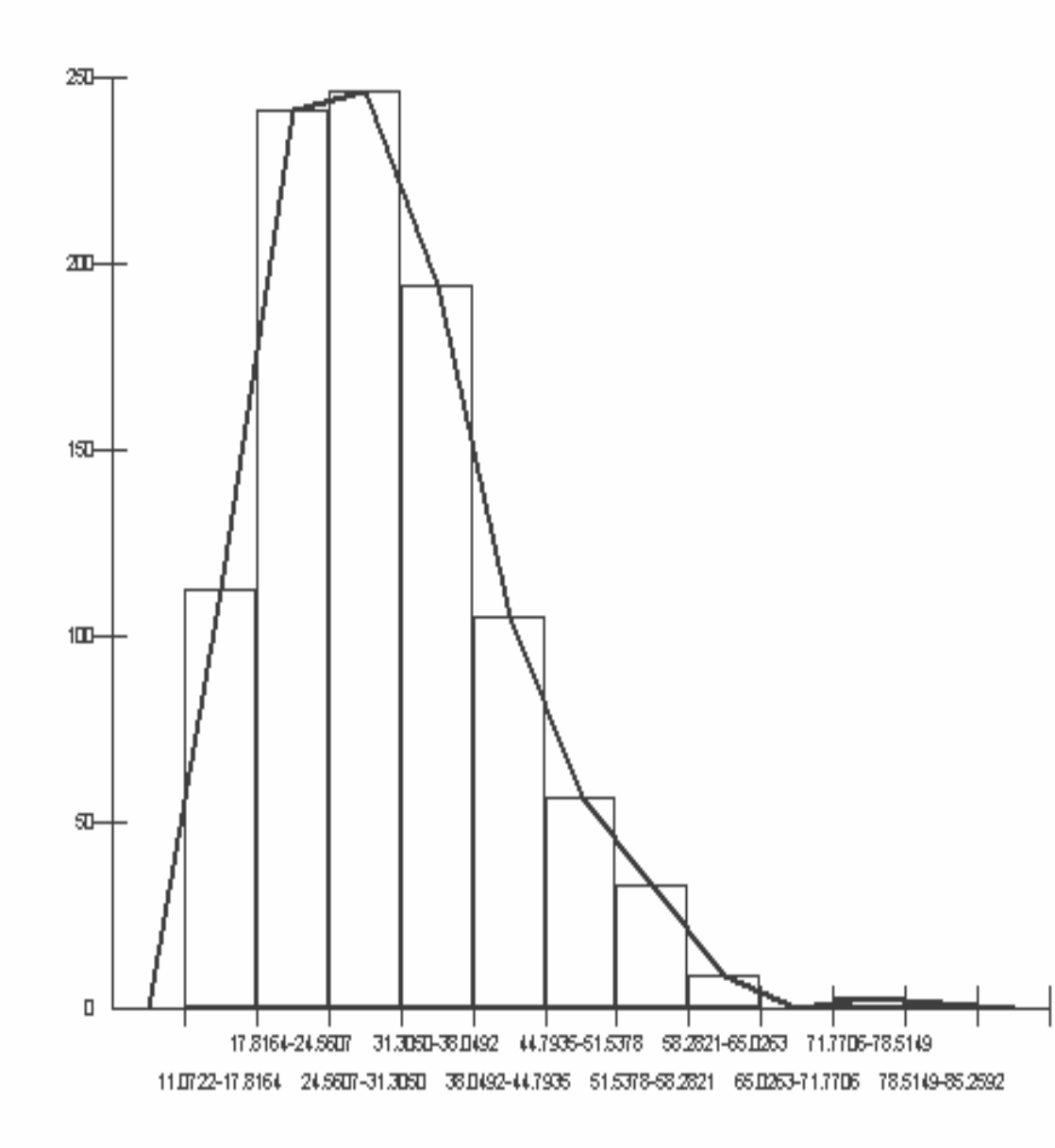
Pela análise do intervalo de confiança , existe uma probabilidade de 95% do intervalo *[13,61% a 53,95%]* conter o percentual de risco de se adoecer por dengue no Estado do Rio de Janeiro . Uma estatística expressiva, alarmante e que se configura uma tendência de epidemia da doença no estado.

Convém colocar que esta é uma forma "preliminar" de construir intervalos de confiança não paramétricos.Neste assunto, o *Bootstrap* tem mostrado recentes avanços, apresentado técnicas que permitem a obtenção de intervalos de confiança mais precisos.

A qualidade da estimação da taxa média de incidência de dengue pode ser melhorada ao considerar-se uma amostra com bairros além dos município do Rio de Janeiro.

Gráfico 1

Histograma das Estimativas das Taxas Médias do Total de Dengue nas 1000 Simulações



6-Conclusão

Este trabalho objetivou a estimação da taxa média total de incidência de casos de dengue no estado do Rio de Janeiro, que pode municiar as autoridades do estado e do municio com informações estatísticas sobre a epidemia e auxiliar no dimensionamento e da proporção exata da abrangência da enfermidade. Que pode motivar campanhas mais emergenciais e intensivas para o combate à epidemia e para o desenvolvimento de campanhas de esclarecimento à população, com foco na prevenção.

Utilizou-se a computação estatística pesada para se obter o intervalo de estimação, além do viés e do erro-padrão da estimativa.

As expectativas para trabalhos futuros constituem aumentar o número de bairros, além dos do município do Estado do Rio de Janeiro para estimar a taxa média de incidência de dengue no estado para se confirmar à regularidade mais precisa do comportamento da estatística investigada. O estudo de um modelo específico de probabilidade para a taxa média de incidência de dengue para a estatística investigada é também possibilidade de estudos futuros.

A investigação ora proposta pode trazer uma contribuição a respeito da análise efetuada. Ela representa uma aplicação direta de estimação utilizando métodos não paramétricos e através da estatística computacional aplicada O conhecimento do erro padrão e da distribuição por amostragem empírica permitiram construir intervalos de confiança e sair, consequentemente, do terreno puramente descritivo do problema.

Espera-se que com o sucesso na estimação da taxa média de incidência de casos de dengue no estado, as autoridades possam ter um dado importante que fundamente o combate mais intenso e sistemático a este grave estado de saúde do nosso estado.

7-Bibliografia

[1]Afifi, A. A. e **Clark**, V. (1984).Computer – Aided Multivariate Analysis. Lifetime Learning Publications. Belm. California.

[2]Anderson, T.W. (1984).An Introduction to Multivariate Statistical Analysis. 2ed.New York : John Wiley & Sons.

[3]Cazar, R. A. (2003).An Exercise on Chemometrics for a Quantitative Analysis Course.Madison: Journal of Chemical Education.

[4]Chatfield, C. e **Collins,** A. J. (1980).Introduction to Multivariate Analysis. Chapman and Hall. New York.

[5]Cliff, N., e **Hamburge**, C. D. (1967).The Study of Sampling Errors in Factor Analysis by Means of Artificial Experiments. Psychological Bulletin 68: 430-45.

[6] Costa, Giovani Glaucio de O. (2003). Busca de Fatores Associados à Prática de Atos Infracionais por Parte de Adolescentes no Estado do Rio de Janeiro: Um Estudo Preliminar, Estudo Orientado, PUC-RIO.

[7]David, A.Aaker , **Kumar**, V;**George**,S. Day.(1984).Marketing Research.

Dillon, W. R. e **Goldstein**, M.(1984).Multivariate Analysis : Methods and Applications . New York : John Wiley & Sons.

[8]Efron,B(1979).Bootstrap Methods:Another Look at the Jackknife, The Annals of Statistic,7,1-26.

[9]Efron,B.(1980).Computer Intensive Methods in Statistics" in Some Recent Advance in Statistic, Ed. J. Tiago de Oliveira e B.Epstein , Academia das Ciências de Lisboa, Lisboa.

[10]Efron,B.(1982).The Jackknife, the Bootstrap , and other Resampling Methods, CBNS 38,SIAM-NSF

[11] Ferreira, D. F. Análise Multivariada. Minas Gerais : Universidade Federal de Lavras.

[12]Hair, J. F. Jr. ; **Anderson,** R.E. ; **Tathan**, R. L. e **Black**, W. C.(2005). Trad. **Sant´Anna**,Adonai Schlup ; **Neto**, Anselmo Chaves. Análise Multivariada de Dados. 5. ed. Porto Alegre : Bookman.

[13]Hair, J. F. Jr. ; **Anderson,** R.E. ; **Tathan**, R. L. e **Black**, W. C.(1998). Multivariate Data Analysis. 5th ed. Upper Saddle River : Prentice Hall.

[14]Harman, Harry H.(1967).Modern Factor Analysis . 2 ed. Chicago : University of Chicago .

[15]Hawkins, D.M., Topics in Multivariate Analysis.Cambridge University Press: Cambridge.

[16]Johnson, D. E. (1998).Applied Multivariate Methods for Data Analysis. Pacific Grove: Duxbury Press.

[17]Johnson, R. A .e **Wichern** , D.W .(1998).Applied Multivariate Statistical Analysis. 4ed. Upper Saddle River: Prentice Hall.