Use of STATA in Pediatric Research - An Indian Perspective
Who is a Pediatrician?

Dr. Bhavneet Bharti, PGIMER- Chandigarh
• Research is an important part of curriculum of pediatric medicine

• Research Project is necessary for Postgraduates

• In order to fulfill their MD/DM/MCH requirements thesis mandatory
Research Questions
Pediatrics—Endless queries?

• Which needle causes less pain in infants undergoing vaccination?

• Which drug is better for the treatment of Pediatric HIV, sepsis and many other diseases
Statistics and Pediatric Research

• For answering these queries - Statistics plays increasingly important role

• It is not possible, for example, to have a new drug treatment approved for use without solid, statistical evidence to support claims of efficacy and safety
Many new statistical methods have been developed with particular relevance for medical researchers. These methods can be applied routinely using statistical software packages.
Statistical softwares

• Statistical knowledge of most physicians may be best described as “limited”
Available Statistical Packages

Proprietary
- Excel
- SPSS
- STATA
- MINITAB
- SAS
- Comprehensive metanalysis

Free Software
- EpiInfo
- R
- Revman
- LibreOffice Calc
- PSPP
### Total Sales by Store

<table>
<thead>
<tr>
<th>Store</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store 1</td>
<td>$313,765</td>
</tr>
<tr>
<td>Store 2</td>
<td>$107,160</td>
</tr>
<tr>
<td>Store 3</td>
<td>$351,751</td>
</tr>
<tr>
<td>Store 4</td>
<td>$131,047</td>
</tr>
<tr>
<td>Store 5</td>
<td>$252,136</td>
</tr>
<tr>
<td>Store 6</td>
<td>$167,462</td>
</tr>
<tr>
<td>Store 7</td>
<td>$210,073</td>
</tr>
<tr>
<td>Store 8</td>
<td>$308,092</td>
</tr>
<tr>
<td>Store 9</td>
<td>$97,492</td>
</tr>
<tr>
<td>Store 10</td>
<td>$393,484</td>
</tr>
<tr>
<td>Store 11</td>
<td>$396,891</td>
</tr>
<tr>
<td>Store 12</td>
<td>$151,168</td>
</tr>
<tr>
<td>Store 13</td>
<td>$251,390</td>
</tr>
<tr>
<td>Store 14</td>
<td>$392,776</td>
</tr>
<tr>
<td>Store 15</td>
<td>$259,654</td>
</tr>
<tr>
<td>Store 16</td>
<td>$225,184</td>
</tr>
<tr>
<td>Store 17</td>
<td>$335,785</td>
</tr>
</tbody>
</table>

### Total Sales by Region

<table>
<thead>
<tr>
<th>Region</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>West</td>
<td>$1,718,258</td>
</tr>
<tr>
<td>South</td>
<td>$534,389</td>
</tr>
<tr>
<td>Midwest</td>
<td>$1,009,268</td>
</tr>
<tr>
<td>East</td>
<td>$900,431</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>$4,162,346</strong></td>
</tr>
</tbody>
</table>

### Total Sales by Category

<table>
<thead>
<tr>
<th>Category</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automotive</td>
<td>$86,285</td>
</tr>
<tr>
<td>Gardening</td>
<td>$52,048</td>
</tr>
<tr>
<td>Electronics</td>
<td>$83,026</td>
</tr>
<tr>
<td>Jewelry</td>
<td>$93,035</td>
</tr>
<tr>
<td>Sporting</td>
<td>$30,016</td>
</tr>
<tr>
<td>Housewares</td>
<td>$19,149</td>
</tr>
<tr>
<td>Books</td>
<td>$42,247</td>
</tr>
<tr>
<td>Games</td>
<td>$18,420</td>
</tr>
</tbody>
</table>

**Total Sales By Category**

- Automotive
- Gardening
- Electronics
- Jewelry
- Sporting
- Housewares
- Books
- Games
Microsoft Excel

**COST**
- Individual License for Microsoft Office Professional $350
- Volume Discounts available for large organizations and universities
- Free Starter Version available on new PCs

**PRO**
- Nearly ubiquitous and is often pre-installed on new computers
- User friendly
- Very good for basic descriptive statistics, charts and plots

**CON**
- Costs money
- Not sufficient for anything beyond the most basic statistical analysis
COST

- $1,395.00 per single user license

CON

- Costs Money
- Not suitable for very complicated statistical computation and analysis
- Not often used in academic research

PRO

- Easy to learn and use
- Often taught in schools in introductory statistics courses
- Widely used in engineering for process improvement
<table>
<thead>
<tr>
<th>patid</th>
<th>gender</th>
<th>active</th>
<th>obesity</th>
<th>diabetes</th>
<th>bp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Female</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Hypotension</td>
</tr>
<tr>
<td>2</td>
<td>Male</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Hypertension</td>
</tr>
<tr>
<td>3</td>
<td>Female</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Hypertension</td>
</tr>
<tr>
<td>4</td>
<td>Male</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Normal</td>
</tr>
<tr>
<td>5</td>
<td>Female</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Hypertension</td>
</tr>
<tr>
<td>6</td>
<td>Male</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Normal</td>
</tr>
<tr>
<td>7</td>
<td>Male</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Hypertension</td>
</tr>
<tr>
<td>8</td>
<td>Male</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Hypertension</td>
</tr>
<tr>
<td>9</td>
<td>Male</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Hypertension</td>
</tr>
<tr>
<td>10</td>
<td>Male</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Hypertension</td>
</tr>
<tr>
<td>11</td>
<td>Male</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Hypertension</td>
</tr>
<tr>
<td>12</td>
<td>Male</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Hypertension</td>
</tr>
<tr>
<td>13</td>
<td>Male</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Hypertension</td>
</tr>
<tr>
<td>14</td>
<td>Male</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Hypertension</td>
</tr>
<tr>
<td>15</td>
<td>Male</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Hypertension</td>
</tr>
<tr>
<td>16</td>
<td>Male</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Hypertension</td>
</tr>
<tr>
<td>17</td>
<td>Male</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Hypertension</td>
</tr>
</tbody>
</table>

**Regression options:**
- Automatic Linear Modeling...
- Linear...
- Curve Estimation...
- Partial Least Squares...
- Binary Logistic...
- Multinomial Logistic...
- Ordinal...
- Probit...
- Nonlinear...
- Weight Estimation...
- 2-Stage Least Squares...
- Optimal Scaling (CATREG)...

**Missing values:**
- Missing Value Analysis...
- Multiple Imputation
- Complex Samples
- Quality Control

**ROC Curve:**
- ROC Curve...
**COST**
- From $1000 to $12000 per license depending on license type.

**CON**
- Very expensive
- Not adequate for modeling and cutting edge statistical analysis

**PRO**
- Easy to learn and use
- More powerful than Minitab
- One of the most widely used statistical packages in academia and industry
- Has a command line interface in addition to menu driven user interface
- One of the most powerful statistical package that is also easy to use.
**PRO**

- Consists of multiple modules to accomplish various tasks beyond just statistical analysis.
- Ability to rapidly develop a questionnaire
- Customize the data entry process
- Quickly enter data into that questionnaire
- Analyze the data

**COST**

- Free

**CON**

- Not a dedicated statistical package
- Not as powerful as commercial alternative for performing advanced analysis and modeling
proc tabulate data=yr2008 nospsses;
  var volnew high low close;
  table date=' ', (high='Weekly High' low='Weekly Low'
    volnew='Volume(100,000)') cl
    * mean='' * f=comma15. / rts=15;
    class date;
run;

proc gchart data=work.sectors;
  pie Sector / suvar=Percentage descending detail=Issuer de
  value=none other=5 otherlabel='Combined'
  noheading legend html=htmlvar name='figure_
run;
quit;
ods html close;
SAS

**COST**
- Complicated pricing model
- $8,500 first year license fee

**CON**
- Very very expensive
- Not user friendly
- Steep learning curve
- Relatively poor graphics capabilities

**PRO**
- Widely accepted as the leader in statistical analysis and modeling
- Widely used in the industry and academia
- Very flexible and very powerful.
Math can be beautiful ...

\[ \cos(r^2) e^{-r^2} \]
PRO

- Widely used and accepted in industry and academia
- Very powerful and flexible
- Very large user base
- Lots of books and manuals
- Several User Interface Shells available

COST

- Free / Open Source

CON

- Not user friendly
- Requires steep learning curve
• R is free but more difficult for those who are not into the world of math and pure statistics.
• There is R Commander package that can ease your comprehension
• SAS is more visual than R that makes its use simpler for those who are not familiar with programming languages.
STATA is one thing which has united us today

“Diversity is the one thing we all have in common. Celebrate it every day.”

- Anonymous
• It is a general-purpose statistical software package used by people of different backgrounds and professional disciplines

• Most Stata users are non-physicians
• Although Stata has simple commands, these may be difficult for non-programmers to use.

• Generally, physicians are familiar with “clicking on” rather than writing commands
No matter which book you choose or which course you attend.

Since Stata is used by people in many fields, most training programs offered are geared toward programmers and non-physicians.
Use of STATA in Pediatric research

12 - 24%
Pediatric Journals (US vs Indian)
Use of STATA and other software in Pediatrics research

Pediatric 2012

Pediatric 2013
Use of STATA and other softwares in Pediatric research

J of Pediatrics 2013

Ind Pediatrics 2013
Attitudes of canadian researchers toward the return to participants of incidental and targeted genomic findings obtained in a pediatric research setting
CV Fernandez, C Strahlendorf, D Averd... - Genetics in ... , 2013 - nature.com
... Data were analyzed using SPSS version 14 (IBM, Somers, NY). ... CPGCC, Canadian Pediatric Cancer Genome Consortium; FORGE, Finding of Rare Genes Canada Consortium. ... results that should be offered to participants as indicated by Canadian Pediatrics Cancer Genome ...
Cited by 2 Related articles All 4 versions Cite

Effect of child health status on parents' allowing children to participate in pediatric research
J Vanhelst, L Hardy, D Bert, S Duhem... - BMC medical ..., 2013 - biomedcentral.com
... using the Statistical Package for the Social Sciences, version 11.5 for Windows (SPSS Inc., Chicago ... time in discussions with investigators before enrolling their child in a pediatric clinical research ... Ramet J, Van den Anker J: The European Academy of pediatrics (EAP/CESP) and ... Related articles All 13 versions Cite More▼
No of Google scholar hits for various softwares
## Popularity of analytic softwares

<table>
<thead>
<tr>
<th>Year</th>
<th>R</th>
<th>SAS</th>
<th>SPSS</th>
<th>Stata</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>7</td>
<td>9120</td>
<td>7310</td>
<td>24</td>
</tr>
<tr>
<td>1996</td>
<td>4</td>
<td>9130</td>
<td>8560</td>
<td>92</td>
</tr>
<tr>
<td>1997</td>
<td>9</td>
<td>10600</td>
<td>11400</td>
<td>214</td>
</tr>
<tr>
<td>1998</td>
<td>16</td>
<td>11400</td>
<td>17900</td>
<td>333</td>
</tr>
<tr>
<td>1999</td>
<td>25</td>
<td>13100</td>
<td>29000</td>
<td>512</td>
</tr>
<tr>
<td>2000</td>
<td>51</td>
<td>17300</td>
<td>50500</td>
<td>785</td>
</tr>
<tr>
<td>2001</td>
<td>155</td>
<td>20900</td>
<td>78300</td>
<td>969</td>
</tr>
<tr>
<td>2002</td>
<td>286</td>
<td>26400</td>
<td>66200</td>
<td>1260</td>
</tr>
<tr>
<td>2003</td>
<td>639</td>
<td>36300</td>
<td>43500</td>
<td>1720</td>
</tr>
<tr>
<td>2004</td>
<td>1220</td>
<td>45700</td>
<td>56000</td>
<td>2350</td>
</tr>
<tr>
<td>2005</td>
<td>2210</td>
<td>55100</td>
<td>71000</td>
<td>2980</td>
</tr>
<tr>
<td>2006</td>
<td>3420</td>
<td>60400</td>
<td>69000</td>
<td>3940</td>
</tr>
<tr>
<td>2007</td>
<td>5070</td>
<td>61900</td>
<td>167000</td>
<td>4900</td>
</tr>
<tr>
<td>2008</td>
<td>7000</td>
<td>63100</td>
<td>155000</td>
<td>6150</td>
</tr>
<tr>
<td>2009</td>
<td>9320</td>
<td>60400</td>
<td>136000</td>
<td>7530</td>
</tr>
<tr>
<td>2010</td>
<td>11500</td>
<td>52000</td>
<td>109000</td>
<td>8890</td>
</tr>
<tr>
<td>2011</td>
<td>13600</td>
<td>44800</td>
<td>74900</td>
<td>10900</td>
</tr>
<tr>
<td>2012</td>
<td>17000</td>
<td>33500</td>
<td>49400</td>
<td>14700</td>
</tr>
</tbody>
</table>
Traffic on Email discussion lists
Number of posts
Number of people registered in the main discussion group for each software-2013
## Comparison using no of Blogs

<table>
<thead>
<tr>
<th>Software</th>
<th>Number of Blogs</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>452</td>
</tr>
<tr>
<td>SAS</td>
<td>40</td>
</tr>
<tr>
<td>Stata</td>
<td>8</td>
</tr>
<tr>
<td>Others</td>
<td>0-3</td>
</tr>
</tbody>
</table>
Comparison no of books written
Statistical software applications used in health services research: analysis of published studies in the U.S

Allard E Dembe1*, Jamie S Partridge2 and Laurel C Geist3

Abstract

Background: This study aims to identify the statistical software applications most commonly employed for data analysis in health services research (HSR) studies in the U.S. The study also examines the extent to which information describing the specific analytical software utilized is provided in published articles reporting on HSR studies.

Methods: Data were extracted from a sample of 1,139 articles (including 877 original research articles) published between 2007 and 2009 in three U.S. HSR journals, that were considered to be representative of the field based upon a set of selection criteria. Descriptive analyses were conducted to categorize patterns in statistical software usage in those articles. The data were stratified by calendar year to detect trends in software use over time.

Results: Only 61.0% of original research articles in prominent U.S. HSR journals identified the particular type of statistical software application used for data analysis. Stata and SAS were overwhelmingly the most commonly used software applications employed (in 46.0% and 42.6% of articles respectively). However, SAS use grew considerably during the study period compared to other applications. Stratification of the data revealed that the type of statistical software used varied considerably by whether authors were from the U.S. or from other countries.

Conclusions: The findings highlight a need for HSR investigators to identify more consistently the specific analytical software used in their studies. Knowing that information can be important, because different software packages might produce varying results, owing to differences in the software's underlying estimation methods.

Keywords: Statistical software, data analysis, SAS, Stata
Forecast
Growth of R - Why?

- The continued rapid growth in add-on packages
- The attraction of R’s powerful language
- The near monopoly R has on the latest analytic methods
- Its free price
- The freedom to teach with real-world examples from outside organizations, which is forbidden to academics by SAS and SPSS licenses (IBM is loosening up on this a bit)
What will slow R’s growth

• is its lack of a graphical user interface that:
  – Is powerful
  – Is easy to use
  – Provides direct cut/paste access to journal style output in word processor format
  – Is standard, i.e. widely accepted as *The One to Use*
  – Is open source
Although Stata is currently the fastest growing package,

it’s growth will slow in 2013 and level off by 2015 at around 23,000 articles, leaving it in fourth place.
The main cause of this will be inertia of users of the established leaders, SPSS and SAS, as well as the competition from all the other packages, most notably R.

R and Stata share many strengths and with R being free, there is doubt Stata will be able to beat R in the long run.
Why the difference?

• Learning to use a data analysis tool well takes significant effort, so people tend to continue using the tool they learned in college for much of their careers.

• As a result, the software used by professors and their students is likely to predict what the next generation of analysts will use for years to come.
How to decide?

• Does it run natively on your computer?
• Does the software provide all the methods you use? If not, how extensible is it?
• Does that extensibility use its own language, or an external one (e.g. Python, R, SQL) that is commonly accessible from many packages?
• Does it fully support the style (programming vs. point-and-click) that you like?
How to decide?

• Are its visualization options (e.g. static vs. interactive) adequate for your problems?
• Does it provide output the form you prefer (e.g. cut & paste vs. LaTeX integration)?
• Does it handle large enough data sets?
• Do your colleagues use it so you can easily share data and programs?
• Can you afford it?
How to decide doctors?

- Keep it simple
- Utilize pull-down menus (rather than commands)
  No discussion on commands
- Provision of cut and paste
  Scientific methods to disseminate information
There are five primary learning styles:

1. visual (picture),
2. visual (text),
3. auditory,
4. verbal,
5. kinesthetic.
Types of Learners

- Visual learners
- Auditory learners
- Kinesthetic learners
PEOPLE GENERALLY REMEMBER:

- 10% of what they read
- 20% of what they hear
- 30% of what they see
- 50% of what they hear and see
- 70% of what they say and write
- 90% of what they say, discuss, and do

PEOPLE ARE ABLE TO:

- Define
- List
- Describe
- Explain
- Demonstrate
- Apply
- Practice
- Analyze
- Design
- Create
- Evaluate
- Design
- Perform

Dale’s Cone of Experience

Source: Computer Strategies, LLC, 1998
THANK YOU