Choosing Statistical Software

Brent Logan, PhD, Professor
Mei-Jie Zhang, PhD, Professor
Qun Xiang, MS, Biostatistician
Medical College of Wisconsin, Division of Biostatistics

Friday, June 15, 2012
12:00-1:00 pm
Medical Education Building– M2050

The Medical College of Wisconsin is accredited by the Accreditation Council for Continuing Medical Education to provide continuing medical education for physicians.

The Medical College of Wisconsin designates this Live activity for a maximum of 1.0 *AMA PRA Category 1 Credit(s)* ™. Physicians should claim only the credit commensurate with the extent of their participation in the activity.

Hours of Participation for Allied Health Professionals

The Medical College of Wisconsin designates this activity for up to 1.0 hours of participation for continuing education for allied health professionals.

Financial Disclosure

- In accordance with the ACCME® standard for Commercial Support Number 6, all in control of content disclosed any relevant financial relationships. The following in control of content had no relevant financial relationships to disclose.

Name:	Role in Meeting:
Ruta Brazauskas, PhD	Planning Committee
Haley Montsma, BBA	Planning Committee
Brent Logan, PhD	Speaker
Mei-Jie Zhang, PhD	Speaker
Qun Xiang, MS	Speaker

Learning Objectives

- Familiarize the audience with basic statistical software options
- Discuss the pros and cons of different types of software
- Demonstrate the different types of software on a simple analysis

Evaluation Forms

Your opinion matters!

Help us plan future meetings, by completing and submitting your evaluation forms.

Thank you.

Statistical Analysis

- **Statistical Analysis:**

  Biomedical and Clinical Researchers commonly employ a variety of statistical analyses (Quantitative and Qualitative methods)

- **Statistical Software:**

  All of these statistical analyses are done by a specific statistical software

- **Available statistical software:**

  Many statistical software applications are available to biomedical researchers

  American Statistical Association website lists over 100 statistical packages

Choose Statistical Software

- **How to choose?**
  - My colleagues use…
  - In grad school I learned …
- In general it depends on:

  The cost of software

  The Platform of computer being used

  Ease of Use (technical level)

  Type of analysis

  Graphics capability

- There is no accepted norm regarding which software to use for biomedical studies

Statistics

- "***Statistical Software applications used in health services research (HSR): analysis of published studies in the U.S***.", By Dembe, et al., *BMC Health Services Research*, 2011, 11:252
- Reviewed total of 1139 articles published 2007 – 2009 in three US HSR journals.
- 535 articles mentioned software used were included in the study
- % of statistical software used:

STATA	46.0%
SAS	42.6%
SUDAAN	6.2%
SPSS	5.8%
Others	18.5%

Statistical Packages

- **General Statistical Packages** -- Statistical Software for general statistical analysis:

  SAS, STATA, SPSS, R (SPLUS), …
- **Specific Packages** -- Statistical software developed for a specific need:

  SUDAAN – complex survey, clustered and correlated data

  Sigmaplot – scientific graph

  Analyzing genomic and microarray data

  Analyzing econometrics data (Time Series)

  Analyzing spatial statistics data
- **High-Level software languages** for developing statistical software and new statistical procedure

  JAVA, C++, Fortran 90, ….

Common Statistical Package: **SAS**

- SAS Institute Started in 1976: more than 40,000 sites worldwide, including 90% of Fortune 500 companies
- Much more than a simple software system: integration of statistical methodologies, database technology and business applications has helped **SAS to become one of most commonly used commercial statistical software**
- SAS has been widely used in:

  Life sciences area

  Automotive industry

  Telecommunications industry

  Home land security

  Economic forecasting,

  Waste and Fraud detection area

Common Statistical Package: **SAS**

- SAS has become one of the "**Biggest players**" in the statistical software arena with various capabilities:
- **SAS/STAT package** – supports most statistical models and methods:

  ANOVA, Regression, Categorical Data Analysis, Multivariate Analysis, Survival Analysis, Psychometric Analysis, Cluster Analysis, Nonparametric analysis,

  Power Analysis, ….
- **SAS/INSIGHT package** -- Supports Exploratory Data Analysis:

  Linked across multiple windows and let user to uncover trends, spot outliers, and patterns of the data
- **SAS/IML package** – Allows Statistician to develop matrix-based

  analysis
- **SAS/GRAPH package** -- Delivers high level graph for

  Data analysis, Data visualization with maps, charts and plots

  Visualization presentation

Common Statistical Package: **SPSS**

- SPSS software system was developed by three Stanford Univ students in 1960s-now owned by IBM
- SPSS was originally designed for mainframe computer system and introduced SPSS/PC+ since 1980s
- **SPSS supports numerous add-on modules for**:

  Regression, advanced models, classification trees, exact tests, categorical analysis, trend analysis, complex sample analysis, ….
- **SPSS supports numerous stand-alone products**:

  SamplePower (sample size calculation package), ….
- SPSS is targeting similar application areas as SAS. The client base for SAS is much more extensive than that of SPSS
- **Commonly used by social scientists and psychologists**

Common Statistical Package: **STATA**

- STATA was developed in 1980s. It is used by many businesses and academic institutions for researchers in the fields of Biomedicine, epidemiology, sociology, economics, …
- **STATA has included a graphical user interface which uses menus and dialog boxes to give access to nearly all built-in commands**
- **STATA supports most statistical models and methods:**

  Data management, basic statistics, linear model/GLM, Longitudinal data/survival analysis,

  Nonparametric methods, exact test, ….

Common Statistical Package: **R (S-PLUS)**

- **S-PLUS** is extension of the statistical analysis language S developed at AT&T Lab
- **R** was developed based on S-like syntax by professors at the University of Auckland, NZ
- **R is an open source software** and allows users to modify the R software and fall within the open source software agreement
- Led to **rapid development of the R software system**
- **R possesses an extensive statistical capability**:

  All basic statistical models and methods and most newly developed models and methods
- **R graphic package** produces journal quality plots
- R-Spin-Off projects:

  “Bioconductor” for gene expression analysis

  “R spatial projects” for spatial statistical analysis, ….

Common Statistical Package: **Sigmaplot**

- **SigmaPlot** is a proprietary software package for scientific graphing and data analysis. It runs on MS Windows.
- It was merged into SPSS in 1996 and currently it is owned and maintained by SYSTAT.
- SigmaPlot has excellent ternary plot features and one of the best customized editing features.
- **SigmaPlot is easy to use and generates high-quality graphics quickly.**
- **Now, SigmaPlot Has Extensive Statistical Analysis Features:**

  Basic statistical testing, Regression, Repeated Measure, Survival Analysis, ….

Criteria for selection

- Cost
  - Annual vs. one time fee
  - Availability of site license
- Platform
  - PC vs. Unix vs. Mac vs. Linux
- Ease of Use
  - Easy, point and click interface
  - Syntax and programming statements
  - Advantage/burden of syntax based analysis
  - Spreadsheet view of data
- Type of analysis
  - Specialized techniques vs. general purpose
- Graphics
  - Interactive, Customizable

Cost

- Annual cost
  - SAS $218/yr (through MCW Biostats)
  - SPSS $326/yr faculty license
  - Stata $295/yr
- One time purchase
  - Sigmaplot $549
  - Stata $595
- Free
  - R
  - Excel Statistics Add in

Platform

Software	WIndows	Mac	Unix	Linux
SAS	X		X	X
SPSS	X	X		X
Sigmaplot	X			
Stata	X	X	X	
Excel Stats	X	X		X
R	X	X	X	X

Ease of Use

- Easy, point and click interface, spreadsheet view of data
  - Sigmaplot – links with Excel
  - SPSS (Syntax also available)
  - Excel Stats
  - Stata (Syntax also available)
- Primarily syntax based
  - SAS (also has some menu drived functionality, though not as easy to use)
  - R
- SAS and R do not typically have easy views of the data in a spreadsheet format (print out specific contents when needed)

Type of analysis

- Most of the larger packages (SAS, SPSS, Stata, R) can handle most common analyses
- SPSS used historically in the Social sciences
- Stata useful for handling weights with survey sampling
- SAS has good data handling/manipulation in addition to comprehensive analysis tools
- R is very flexible, people write their own routines to do different types of analysis
  - Quality control limited – need to know what you are doing
  - Limited manuals
- Excel can only do a very limited number of simple analyses
- Specialized analysis: SUDAAN, etc.

Graphics

- SAS and R
  - Customizable, but primarily syntax driven
- SPSS, SigmaPlot, Stata
  - More interactive

Examples of using SAS-dataset

```
data dat1;
   input cond $ test msat;
   label cond = 'Experimental condition';
   label test = 'Fraction correct on post-test';
  label msat = 'Math SAT score';
  datalines;
   A     0.71   650
  A     0.82   710
  A     0.82   510
  A     0.76   590
  A     0.76   500
  A     0.71   730
  A     0.71   570
  A     0.68    490
  A     0.85   530
  A     0.87   620
  A     0.82   780
  B     0.65   690
  B     0.53   710
  B     0.88   780
  B     0.59   690
  B     0.76   730
  B     0.59   700
  B     0.65   740
   B    0.63   750
  ;
 run;
```

reference - http://facweb.cs.depaul.edu/cmiller/it223/ttest.html

Examples of Using SAS Frequency table

- proc freq data=dat1;

  table cond; run

Cond	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A	11	57.89	11	57.89
B	8	42.11	19	100.00

Examples of Using SAS--t test

```
proc ttest data=dat1;
var test;   class cond;   run;
```

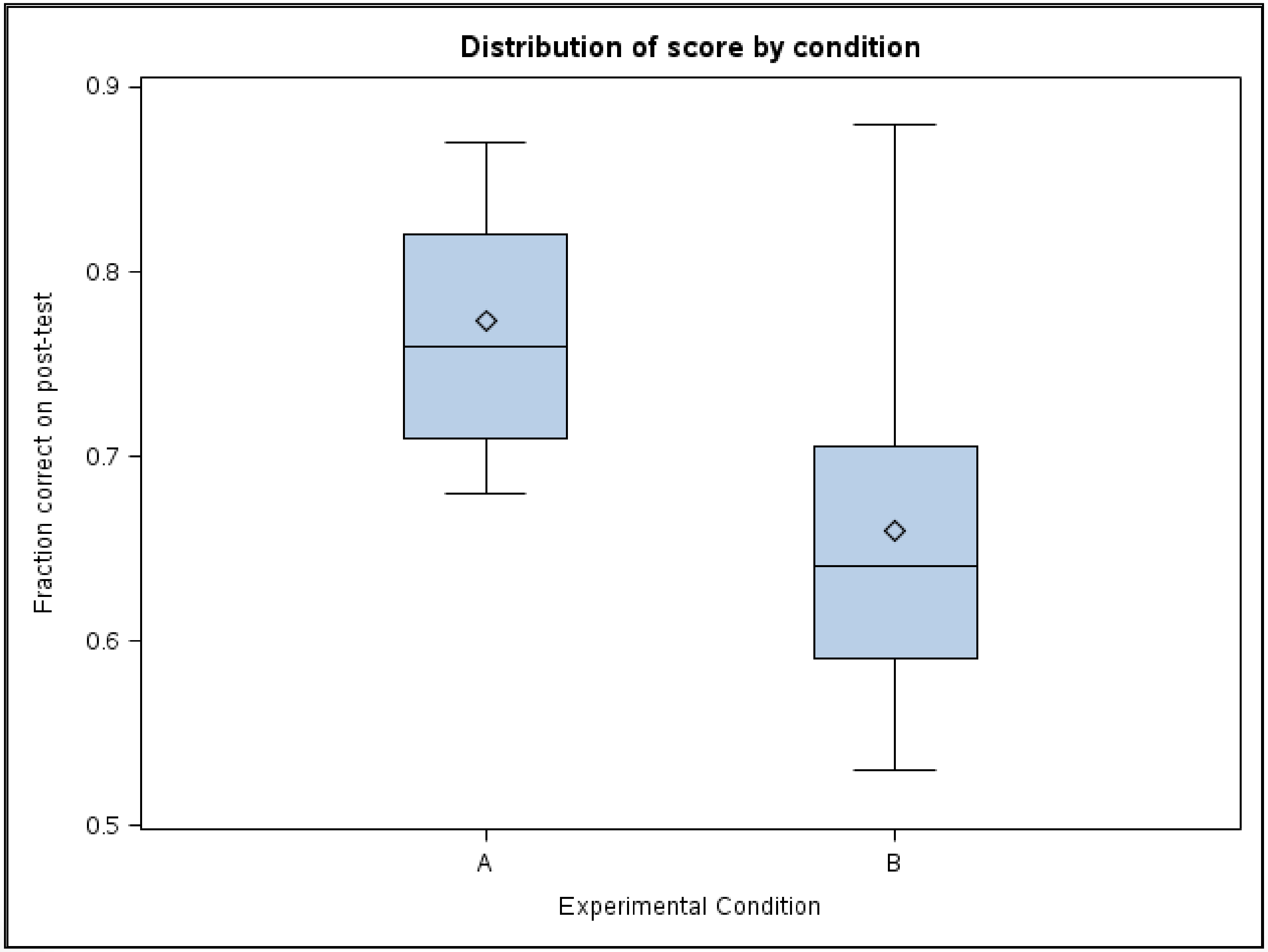
condition	N	Mean	Std Dev	Std Err	Minimum	Maximum
A	11	0.7736	0.0655	0.0197	0.6800	0.8700
B	8	0.6600	0.1110	0.0392	0.5300	0.8800
Diff (1-2)		0.1136	0.0871	0.0405		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	17	2.81	0.0121
Satterthwaite	Unequal	10.52	2.59	0.0261

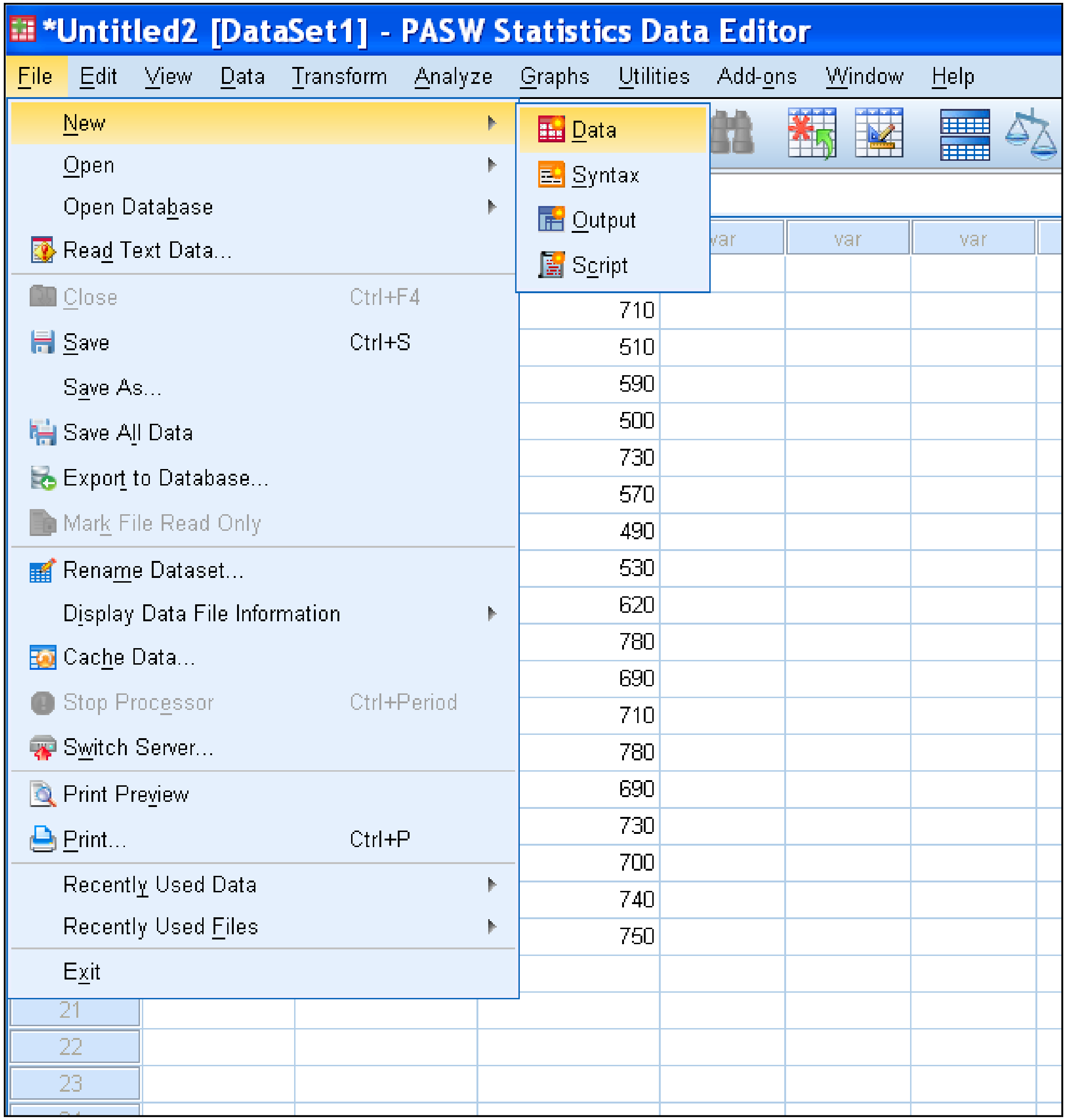
Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	7	10	2.87	0.1274

Examples of Using SAS-Graphs

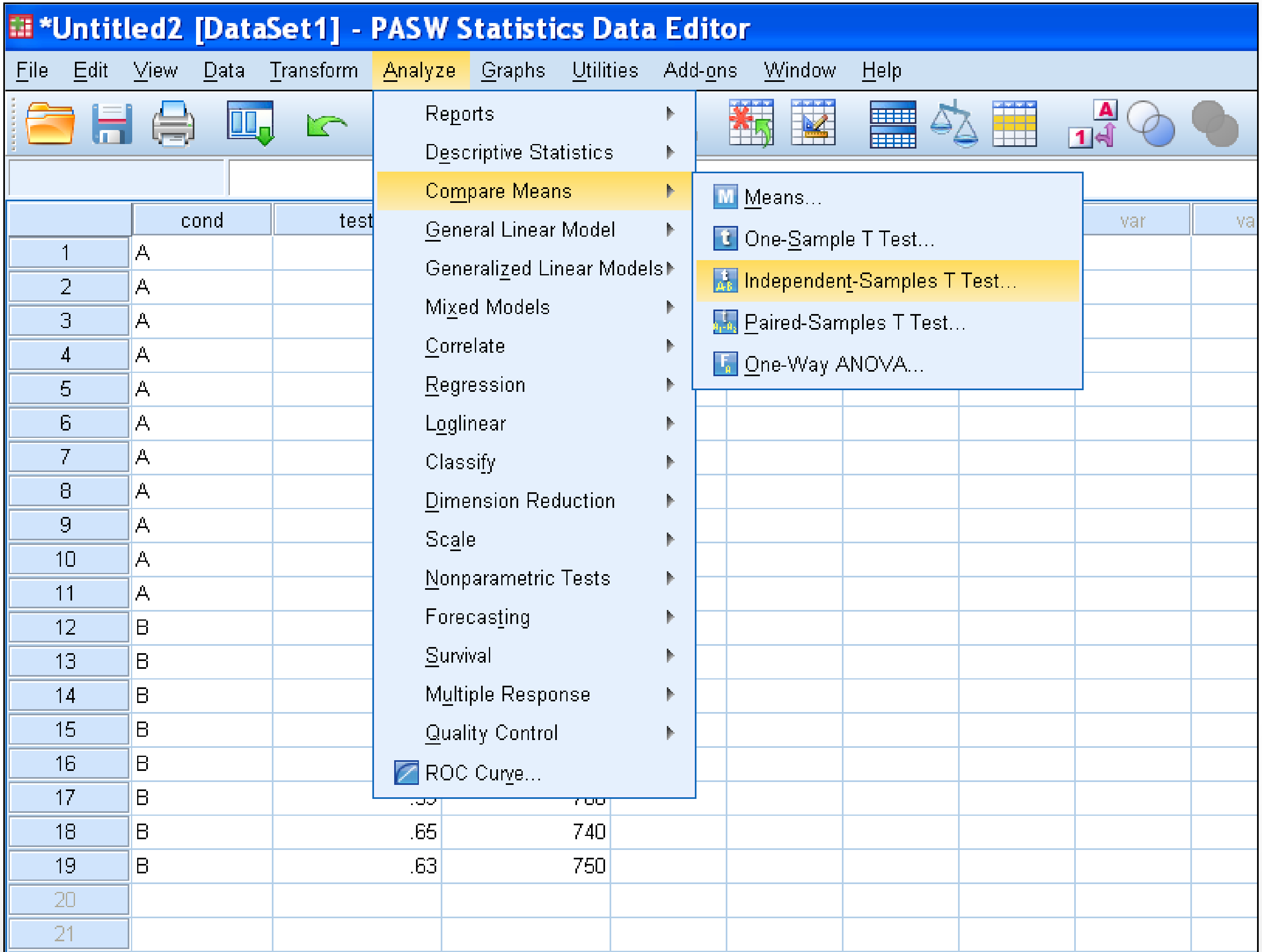
```
proc boxplot data=dat1;
Plot test*cond;  run;
```



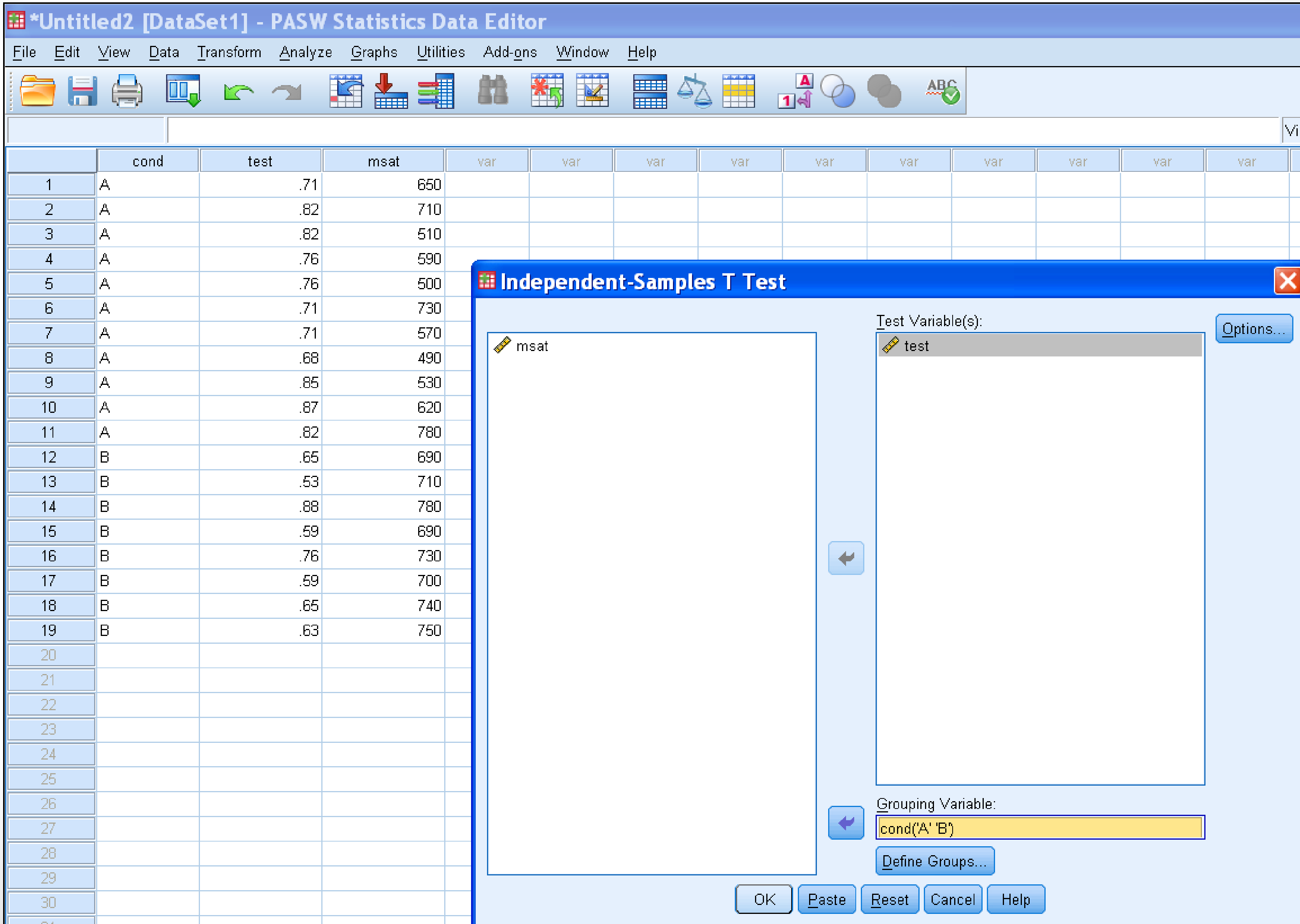
Examples of Using SPSS-Data



Examples of Using SPSS-t test (1)



Examples of Using SPSS-t test (2)



Examples of Using SPSS-Output

*Output1 [Document1] - PASW Statistics Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Add-ons Window Help

- Output
  - T-Test
    - Title
    - Notes
    - Active Dataset
    - Group Statistic
    - Independent S

T-Test

[DataSet1]

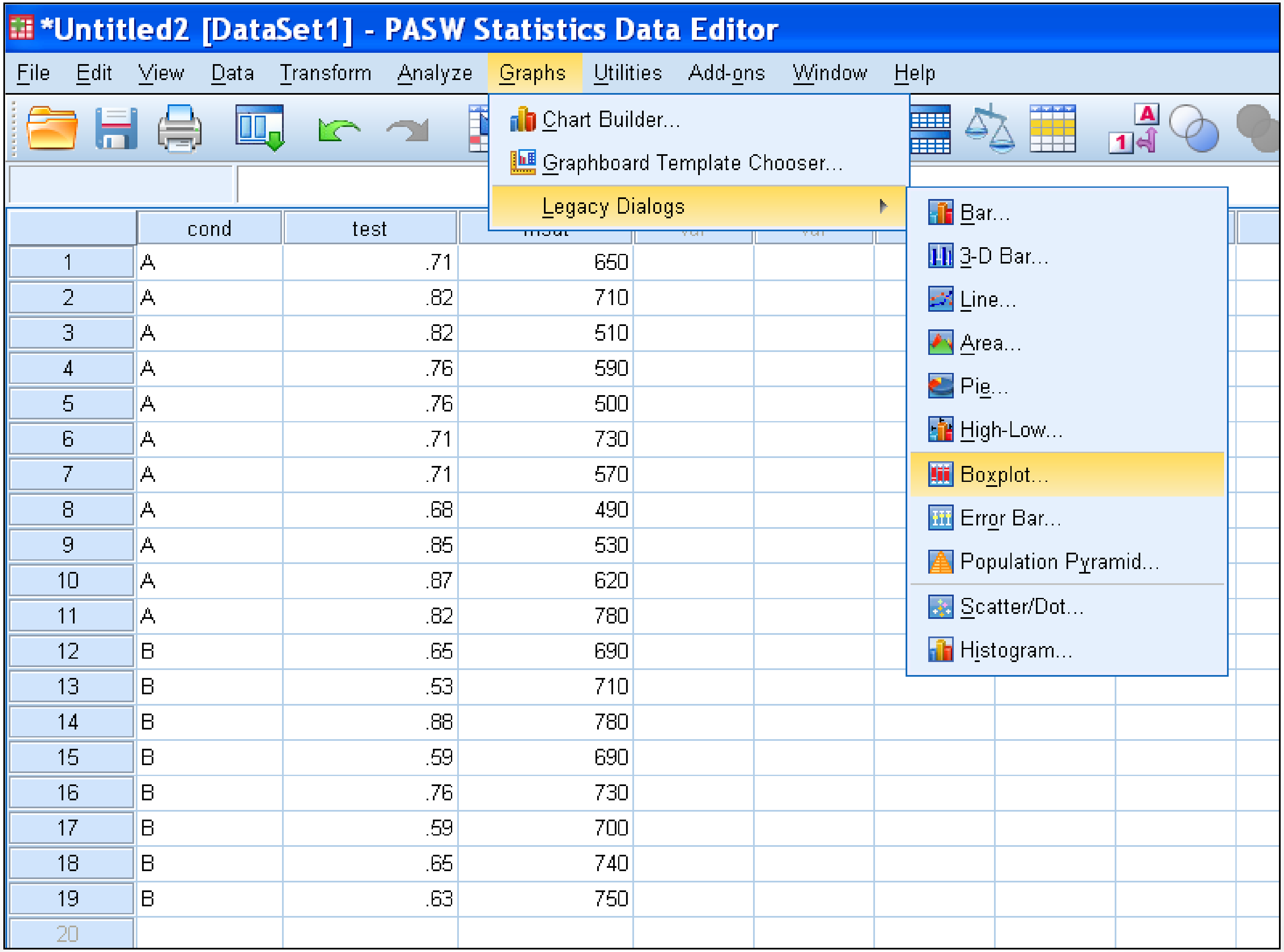
Group Statistics

	cond	N	Mean	Std. Deviation	Std. Error Mean
test	A	11	.7736	.06546	.01974
	B	8	.6600	.11097	.03923

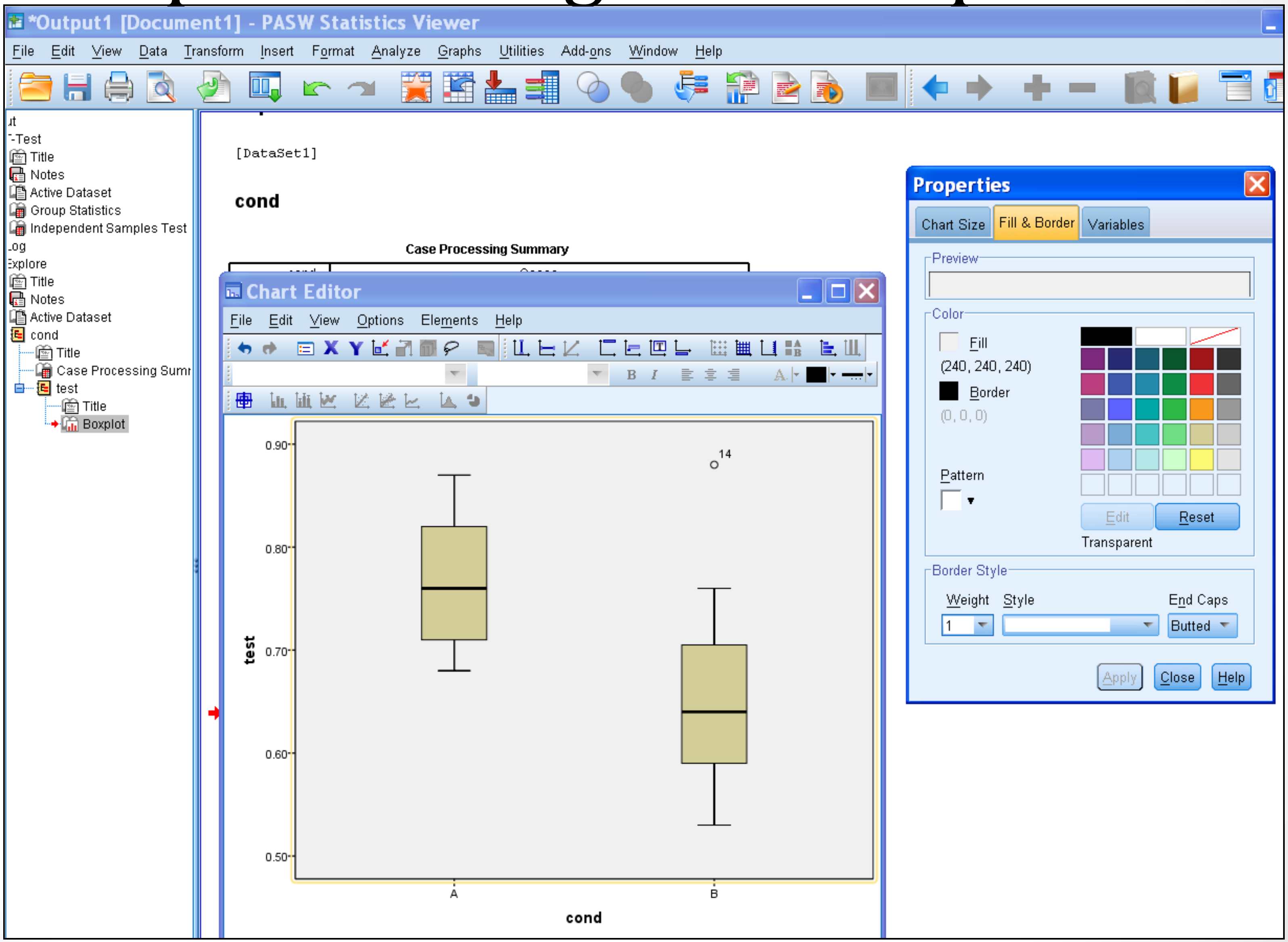
Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means					95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
test	Equal variances assumed	1.006	.330	2.807	17	.012	.11364	.04049	.02822	.19905
	Equal variances not assumed			2.587	10.520	.026	.11364	.04392	.01643	.21084

Examples of Using SPSS-Graphs



Examples of Using SPSS-Boxplot



Example of Using R

- **Open and Running R**
- **Read In data:**

```
> cond <- c(rep("A",11),rep("B",8))
> test <- c(0.71,0.82,0.82,0.76,0.76,0.71,0.71,0.68,0.85,0.87,0.82,0.65,0.53,
+ 0.88,0.59,0.76,0.59,0.65,0.63)
```

- **Compute Frequency Table for “cond”:**

```
> table(cond)
```

- **R-output:**

```
cond
 A  B
11  8
```

Example of Using R

- **T-test of testing equal mean of test score between condition A and B:**

> *t.test(test~cond, var.equal = T)* [var.equal=F for unequal variance T-test]

Two Sample t-test

data: test by cond

t = 2.8069, df = 17, p-value = 0.01213

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.02821979 0.19905294

sample estimates:

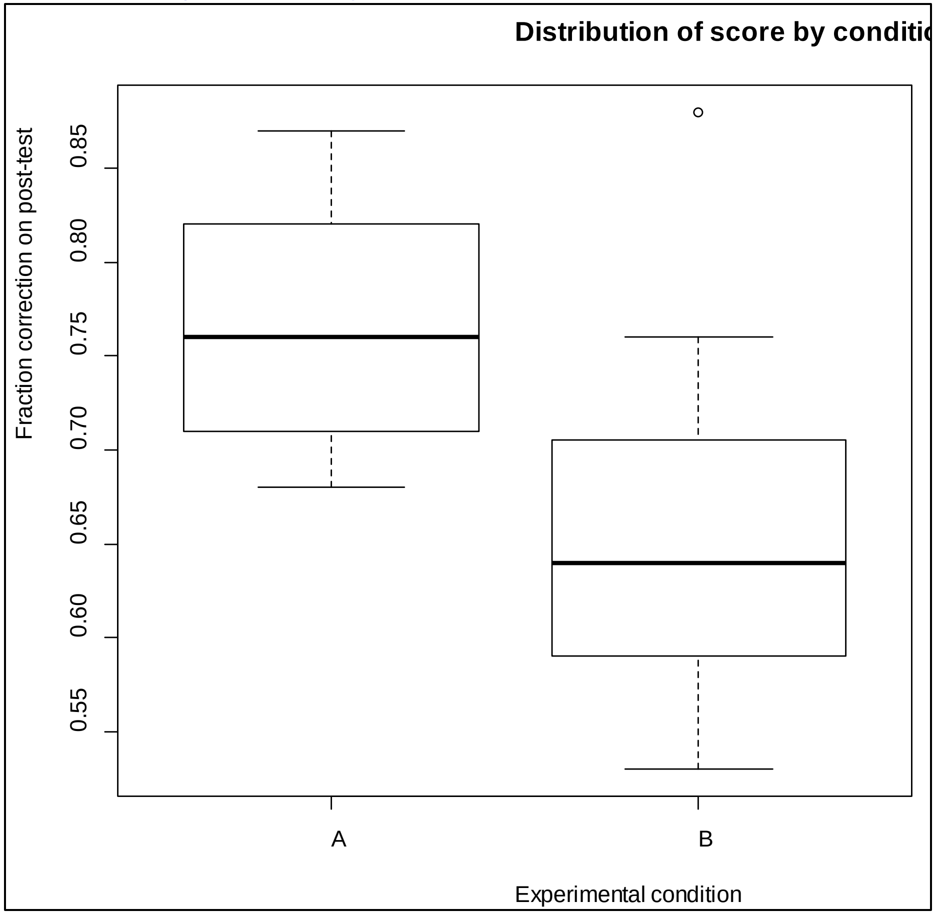
mean in group A mean in group B

0.7736364 0.6600000

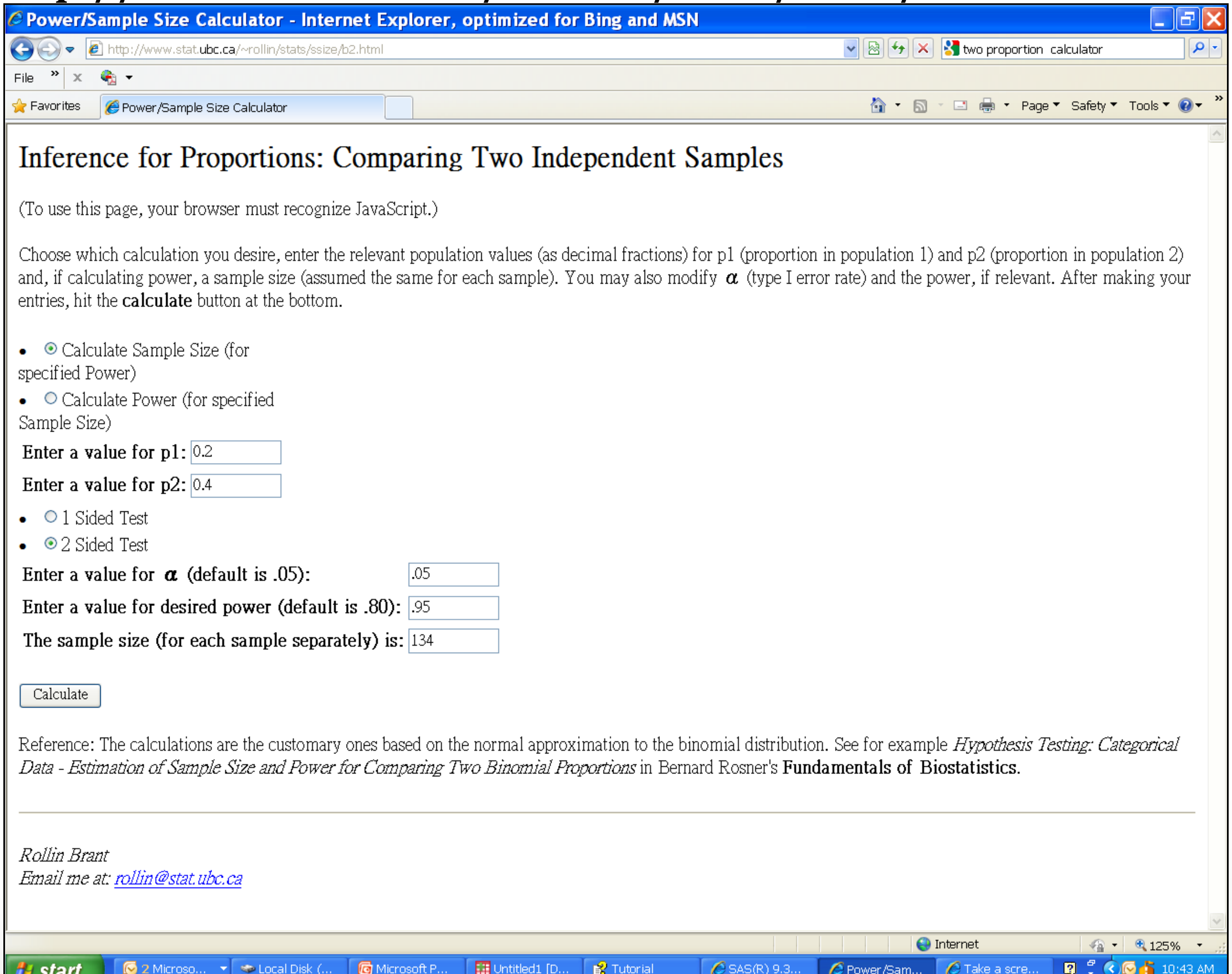
Example of Using R

- Make BOX plot for test score by condition A and B:

```
> boxplot(test~cond, xlab="Experimental condition",
+       ylab="Fraction correction on post-test",
+       main="Distribution of score by condition")
```



Online Calculators-Sample size and power calculation http://www.stat.ubc.ca/~rollin/stats/ssize/b2.html



Inference for Proportions: Comparing Two Independent Samples

(To use this page, your browser must recognize JavaScript.)

Choose which calculation you desire, enter the relevant population values (as decimal fractions) for p1 (proportion in population 1) and p2 (proportion in population 2) and, if calculating power, a sample size (assumed the same for each sample). You may also modify α (type I error rate) and the power, if relevant. After making your entries, hit the **calculate** button at the bottom.

- ⦿ Calculate Sample Size (for specified Power)
- ○ Calculate Power (for specified Sample Size)

Enter a value for p1: 0.2

Enter a value for p2: 0.4

- ○ 1 Sided Test
- ⦿ 2 Sided Test

Enter a value for α (default is .05): .05

Enter a value for desired power (default is .80): .95

The sample size (for each sample separately) is: 134

Calculate

Reference: The calculations are the customary ones based on the normal approximation to the binomial distribution. See for example *Hypothesis Testing: Categorical Data - Estimation of Sample Size and Power for Comparing Two Binomial Proportions* in Bernard Rosner's **Fundamentals of Biostatistics**.

Rollin Brant
Email me at: rollin@stat.ubc.ca

Free Biostatistics Drop-in Service

- **Medical College of Wisconsin**:
  Tuesdays and Thursdays
  Time: 1:00 PM—3:00 PM
  Building: Health Research Center
  Room: H2400 Biostatistics

- **MCW Cancer Center**
  Wednesdays 10:00 AM—12:00 PM
  Fridays 1:00 PM—3:00 PM
  Building: MCW Clinical Cancer Center
  Room: Clinical Trials Support Room
  CLCC: 3236 (Enter through C3233)

- **Froedtert Pavilion**:
  Mondays & Wednesdays
  Time: 1:00 PM—3:00 PM
  Building: Froedtert Pavilion
  Room: L772A- TRU Offices (Lower Level)

- **Clement J. Zablocki VA Medical Center:**
  1st & 3rd Monday of the month
  Time: 9:00 AM—11:00 AM
  Building: 111, 5th Floor B-wing
  Room: 5423

- **Marquette University:**
  Every Tuesday
  Time: 8:30 AM—10:30 AM
  Building: School of Nursing—Clark Hall
  Room: Office of Research and Scholarship: 112D
  Contact: Jessica Pruszynski, PhD to make an appointment
  Please note: Priority given to MU Nursing and Dental School personnel

Questions?