

Introduction to statistical programs:
What exists?
What can you do with it?

Geert
De Meyer



1

Conflict of interest – Geert De Meyer

- Ghent University employee involved in statistical consulting services
- No links with statistical software companies

2

Stat-Gent valorizes UGent statistical expertise

Flexible access in a stable and professional framework

- | | |
|--|--|
| <ul style="list-style-type: none">• Single point of contact• Guidance and supervision by UGent professors• Dedicated personnel• Professional infrastructure | <ul style="list-style-type: none">• Design, data analysis and prediction• Clinical trial design• Longitudinal data analysis• Survival analysis• Causal data analysis |
|--|--|

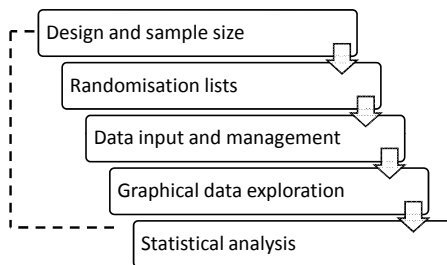
3

Selecting statistical software

The 'best' statistical software program does not exist. One needs to select that program that best fits one's personal needs.

I will introduce a limited number of programs to discuss important features to consider, not to compare these programs.

Clinical research and software

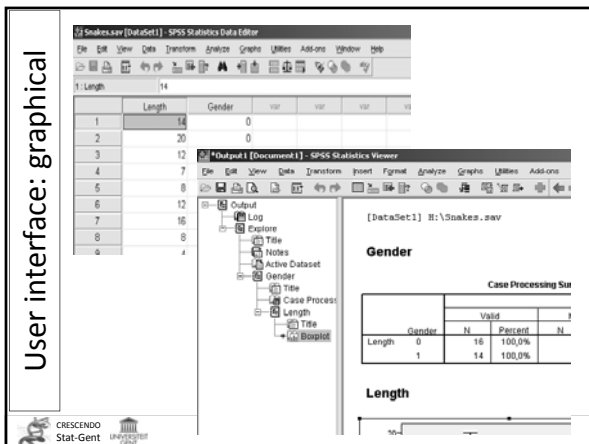


Considering interrelations improves quality and efficiency

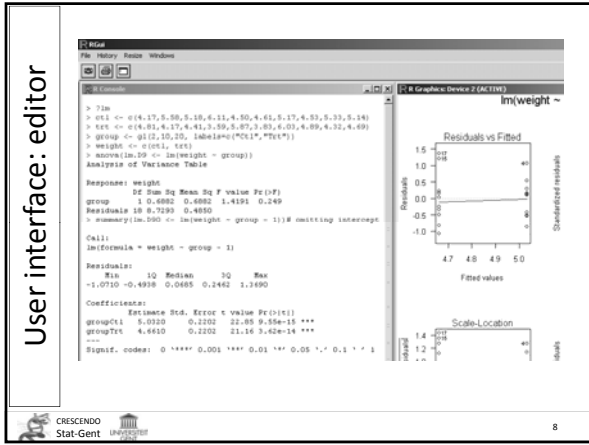
Considerations for software selection

Ease of use	Data input	Procedures	Other
<ul style="list-style-type: none">• User interface• Guided analysis• Manuals and user groups	<ul style="list-style-type: none">• Spreadsheet• Importing data files• Data management	<ul style="list-style-type: none">• Data analysis methods• Graphics• Sample size	<ul style="list-style-type: none">• Validation• Pricing

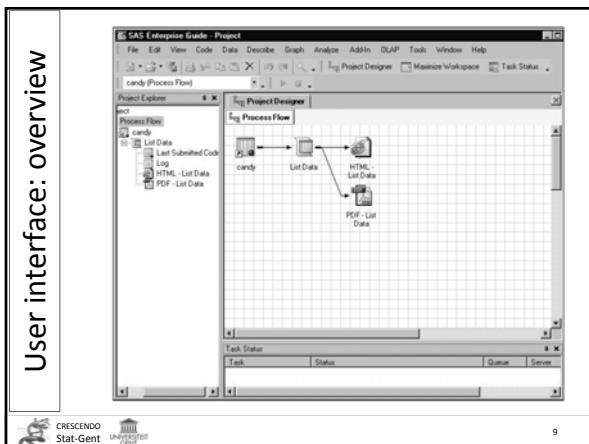
User interface: graphical



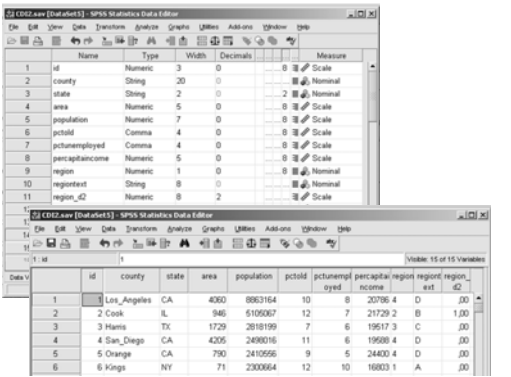
User interface: editor



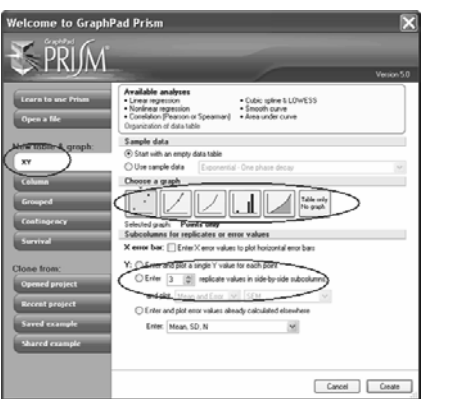
User interface: overview



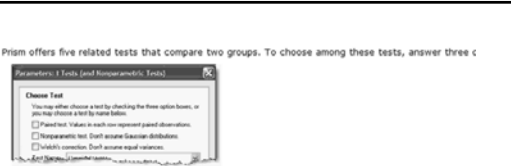
Guided analysis



Guided analysis



Guided analysis



Analysis checklist: Unpaired t test

- ✓ **Are the populations distributed according to a Gaussian distribution?**
The unpaired t test assumes that you have sampled your data from populations that follow a Gaussian distribution. [help you test this assumption.](#)
- ✓ **Do the two populations have the same variances?**
The unpaired t test assumes that the two populations have the same variances (and thus the same standard deviation). Prism tests for equality of variance with an F test. The P value from this test answers this question: If the two populations really have the same variance, what is the chance that you would randomly select samples whose ratio of variances is 1.0 (or further) as observed in your experiment? A small P value suggests that the variances are different. Don't base your conclusion solely on the F test. Also think about data from other similar experiments. If you have previous data that convinces you that the variances are really equal, ignore the F test (unless the P value is very small). In some contexts, finding that populations have different variances may be as important as finding differences in means.
- ✓ **Are the data unpaired?**
The unpaired t test works by comparing the difference between means with the standard error of the difference between means. If the data are paired or matched, then you should use a paired t test instead. If the pairing is effective in controlling for experimental variability, the paired t test will be more powerful than the unpaired test.

Interpreting results: Unpaired t

Confidence Interval

The unpaired t test compares the means of two groups. The most useful result is the confidence interval for the difference between the means. If the assumptions of the analysis are met, you can be 95% sure that the 95% confidence interval contains the true difference between the means. The point of the experiment was to see how far apart the two means are. The confidence interval tells you how precisely you know that difference.

For many purposes, this confidence interval is all you need.

P value

The P value is used to ask whether the difference between the mean of two groups is likely to be due to chance. It answers this question:

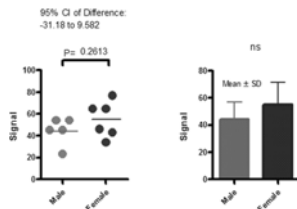
If the two populations really had the same mean, what is the chance that random sampling would result in means as far apart (or more so) than observed in this experiment?

It is traditional, but not necessary and often not useful, to use the P value to make a simple statement about whether or not the difference is "statistically significant".

You will interpret the results differently depending on whether the P value is *small* or *large*.

Graphing tips: Unpaired t

Points or bars?



The graphs above plot the sample data for an unpaired t test. We prefer the graph on the left, which shows each individual data point. This shows more detail, and is easier to interpret, than the bar graph on the right.

Basic Statistics

- Descriptive statistics
 - "True" Mean and Confidence Interval
 - Shape of the Distribution, Normality
- Correlations
 - Purpose (What is Correlation?)
 - Simple Linear Correlation (Pearson r)
 - How to Interpret the Values of Correlations
 - Significance of Correlations
 - Outliers
 - Quantitative Approach to Outliers
 - Correlations in Non-homogeneous Groups
 - Nonlinear Relations between Variables
 - Measuring Nonlinear Relations
 - Exploratory Examination of Correlation Matrices
 - Censoring vs. Pairwise Deletion of Missing Data
 - How to Identify Biases Caused by the Bias due to Pairwise Deletion of Missing Data
 - Pairwise Deletion of Missing Data vs. Mean Substitution
 - Spurious Correlations
 - Are correlation coefficients "additive"?
 - How to Determine Whether Two Correlation Coefficients are Significant
- t-test for independent samples
 - Purpose, Assumptions
 - Arrangement of Data
 - t-test graphs
 - More Complex Group Comparisons
- t-test for dependent samples
 - Within-group Variation
 - Purpose

t-Test for Independent Samples

Purpose, Assumptions. The t-test is the most commonly used method to evaluate the differences in means between two groups. For example, the t-test can be used to test for a difference in test scores between a group of patients who were given a drug and a control group who received a placebo. Theoretically, the t-test can be used even if the sample sizes are very small (e.g., as small as 10, some researchers claim that even smaller n's are possible), as long as the variables are normally distributed within each group and the variance of scores in the two groups is not reliably different (see also Elementary Concepts). As mentioned before, the normality assumption can be evaluated by looking at the distribution of the data (via histograms or by performing a normality test). The equality of variances assumption can be verified with the F-test, or you can use the more robust Levene's test. If these conditions are not met, then you can evaluate the differences in means between two groups using one of the nonparametric alternatives to the t-test (see *Nonparametrics and Distribution Fitting*).

The p-level reported with a t-test represents the probability of error involved in accepting our research hypothesis about the existence of a difference. Technically speaking, this is the probability of error associated with rejecting the hypothesis of no difference between the two categories of observations (corresponding to the groups) in the population when, in fact, the hypothesis is true. Some researchers suggest that if the difference is in the predicted direction, you can consider only one half (one "half") of the probability distribution and thus divide the standard p-level reported with a t-test in "two-tailed" probability by two. Others, however, suggest that you should always report the standard, two-tailed t-test probability.

See also, Student's t Distribution.

Arrangement of Data. In order to perform the t-test for independent samples, one independent (grouping) variable (e.g., Gender male/female) and at least one dependent variable (e.g., a test score) are required. The means of the dependent variable will be compared between selected groups based on the specified values (e.g., male and female) of the independent variable. The following data set can be analyzed with a t-test comparing the average WCC score in males and females.

	GENDER	WCC
case 1	male	111
case 2	male	110
case 3	male	100
case 4	female	102
case 5	female	104
	mean WCC in males =	110
	mean WCC in females =	103

Chapter 3

The FREQ Procedure

Contents

- Overview: FREQ Procedure 44
- Getting Started: FREQ Procedure 46
 - Frequency Tables and Statistics 66
 - Agreement Study 73
- Syntax: FREQ Procedure 75
 - PROC FREQ Statement 76
 - BY Statement 78
 - EXACT Statement 79
 - OUTPUT Statement 82
 - TABLES Statement 85
 - TEST Statement 110
 - WEIGHT Statement 111
- Details: FREQ Procedure 112
 - Inputting Frequency Counts 112
 - Grouping with Formats 113
 - Missing Values 114
 - In-Database Computation 116
 - Statistical Computations 118

R Documentation

Fitting Linear Models

Description

`lm` is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (although `glm` may provide a more convenient interface for these).

Usage

```
lm(formula, data, subset, weights, na.action,
  method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
  singular.ok = TRUE, contrasts = NULL, offset, ...)
```

Arguments


formula an object of class "formula" (or one that can be coerced to that class): a symbolic description of the model to be fitted. The details of model specification are given under "Details".


data an optional data frame, list or environment (or object coercible by `as.data.frame` to a data frame) containing the variables in the model. If not found in `data`, the variables are taken from `environment(formula)`, typically the environment from which `lm` is called.

subset an optional vector specifying a subset of observations to be used in the fitting process.

weights an optional vector of weights to be used in the fitting process. Should be `NULL` or a numeric vector. If non-`NULL`, weighted least squares is used with weights `wweights` (that is, minimizing $\sum(w_i * e_i^2)$), otherwise ordinary least squares is used.

na.action a function which indicates what should happen when the data contain NAs. The default is set by the `na.action` setting of `options`, and is `na.fail` if that is unset. The "factory-fresh" default is `na.omit`. Another possible value is `NULL`, no action. Value `na.exclude` can be useful.

 19



www.listserv.uga.edu

The University of Georgia

Home | Browse | Manage | Request | Manuals | Register


Date: Fri, 22 Jan 2010 18:48:40 -0500
Reply-To: ewilpelt@spoo7FWAGL.COM
Sender: "SAS(r) Discussion" <SAS-L@LISTSERV.UGA.EDU>
From: Ken Borowiak <ewilpelt@spoo7FWAGL.COM>
Subject: Re: CDISC format of XML
Comments-To: jsagomir@SASPOS@OPALMA.COM
In-Reply-To: <201001222127.00MSp10032705@aliba.cc.uga.edu>
Content-Type: text/plain; charset="us-ascii"

Jeff,

If you are interested in creating ODM XML then take a look at:
<http://support.sas.com/rnd/base/alexengine/proccdisc/78774.pdf>

PROC CDISC requires you to create auxiliary fields so it can create the XML file properly.

Regards,
 Ken Borowiak

 20

help archive by thread - Windows Internet Explorer

<http://listserv.uga.edu.au/cgi-bin/jsp1001/01/01/01/01/index.html#and>

- Re: [R] Error in plot.new() Barry Rowlingson (Sun 24 Jan 2010 - 21:25:53 GMT)
- Re: [R] Error in plot.new() Sean Lee Picard (Sun 24 Jan 2010 - 21:40:40 GMT)
- Re: [R] Error in plot.new() Barry Rowlingson (Sun 24 Jan 2010 - 22:38:16 GMT)
- Re: [R] Error in plot.new() Sean Lee Picard (Mon 25 Jan 2010 - 20:00:42 GMT)
- [R] How to define degree() in macro: Heleen Lee (Sun 24 Jan 2010 - 21:28:03 GMT)
- Re: [R] How to define degree() in macro: Gavin Simpson (Mon 25 Jan 2010 - 12:05:46 GMT)
- [R] Catastrophic data reported on time analysis: Marvato Lata (Sun 24 Jan 2010 - 22:26:24 GMT)
- Re: [R] Categorical data reported on time analysis: Dennis Murphy (Sun 24 Jan 2010 - 22:42:24 GMT)
- [R] Creating directories & folders: Steven King (Sun 24 Jan 2010 - 23:01:53 GMT)
 - Re: [R] Creating directories & folders: Caedrick W. Johnson (Sun 24 Jan 2010 - 23:22:49 GMT)
 - Re: [R] Creating directories & folders: Steven King (Sun 24 Jan 2010 - 23:34:53 GMT)
 - Re: [R] Creating directories & folders: Caedrick W. Johnson (Sun 24 Jan 2010 - 23:42:34 GMT)
- Re: [R] Creating directories & folders: Henrique Dallazuanna (Mon 25 Jan 2010 - 00:43:50 GMT)
- [R] problem with the precision of numbers: Amy (Sun 24 Jan 2010 - 22:24:10 GMT)
- Re: [R] problem with the precision of numbers: Ted Harding (Sun 24 Jan 2010 - 23:54:56 GMT)
- Re: [R] problem with the precision of numbers: William Dunlap (Mon 25 Jan 2010 - 00:12:29 GMT)
- Re: [R] problem with the precision of numbers: Ted Harding (Mon 25 Jan 2010 - 00:48:00 GMT)
- Re: [R] problem with the precision of numbers: Gabor Grothowalski (Mon 25 Jan 2010 - 01:04:30 GMT)
- Re: [R] problem with the precision of numbers: Amy (Mon 25 Jan 2010 - 16:04:15 GMT)
- Re: [R] problem with the precision of numbers: William Dunlap (Mon 25 Jan 2010 - 18:26:51 GMT)
- Re: [R] problem with the precision of numbers: Hans W. Borchers (Mon 25 Jan 2010 - 19:18:59 GMT)
- [R] Functional data analysis - problem with functional linear regression: Benjamin Cwick (Mon 25 Jan 2010 - 01:00:46 GMT)

 21

Statistics with R (part 1: vector arithmetics tutorial)

```

> x * y
[1] 7 9 11 13 15
> length(x)
[1] 5
> |
  
```

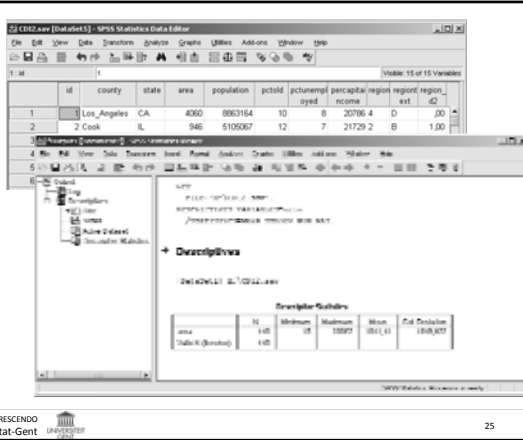
16 ratings 19,119 views

X	Y
12	77
17	99
23	107
34	108
45	144
51	130
60	199

	df	SS	MS	F	Significance F
Regression	1	7770.91	7770.91	23.31081	0.004764306
Residual	5	1665.804	333.1608		
Total	6	9437.714			

	Coefficient	Standard Error	t Stat	P-value
Intercept	54.90003	15.78232	3.478379	0.017886
X	1.982231	0.420559	4.828127	0.004764

Spreadsheet for data only



Data input : basics

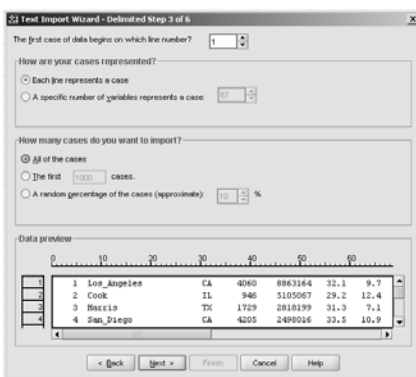
Manual

- Possible in most programs
- Data validation (only defined values possible)
- Procedures for comparing double data entry

Copy paste

- Generally not supported
- Watch out with data formats and regional settings
 - Dates
 - Comma vs point

Data input : file import



Data input : file import

Data Input

Description

Reads a file in table format and creates a data frame from it, with cases corresponding to lines and variables to fields in the file.

Usage

```
read.table(file, header = FALSE, sep = "", quote = "\"",
  as.is = !stringsAsFactors,
  na.strings = "NA", colClasses = NA, rows = -1,
  skip = 0, check.names = TRUE, fill = !blank.lines.skip,
  strip.white = FALSE, blank.lines.skip = TRUE,
  comment.char = "#",
  allowEscapes = FALSE, flush = FALSE,
  stringsAsFactors = default.stringsAsFactors(),
  encoding = "unknown")

read.csv(file, header = TRUE, sep = ",", quote="\"", dec=".",
  fill = TRUE, comment.char="", ...)

read.csv2(file, header = TRUE, sep = ";", quote="\"", dec=";",
  fill = TRUE, comment.char="", ...)
```

Horizontal lines for notes.

Data mngt: transform



Often transformed variable is added to the dataset leading to a messy data set. Best practice is to lock the original raw data in one dataset, and to work on the transformed data in another working data set.

Horizontal lines for notes.

Data mngt: combine datasets

```
Code:
proc sort data=Toy; by CompanyCode; run;
proc sort data=Company; by CompanyCode; run;
data Merged_ToyCompany;
merge Toy Company;
by CompanyCode;
run;

Log Results:
NOTE: There were 7 observations read from the data set WORK.TOY.
NOTE: There were 2 observations read from the data set WORK.COMPANY.
NOTE: The data set WORK.MERGED_TOYCOMPANY has 7 observations and 4 variables.
```

Generally only included in high-level software programs. Differences in transparency (flowchart) and error or warning messages.

Horizontal lines for notes.

Data analysis methods

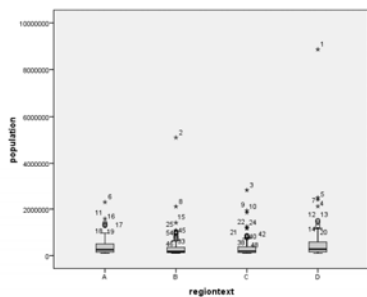
	Graphpad Prism	Excel	Statistica	SPSS	Medcalc	SAS	R	Spotfire St
T-test	+	+	+	+	+	+	+	+
ANOVA	+	+	+	+	+	+	+	+
Nonparametrics	+	-	+	+	+	+	+	+
Linear regression	+	+	+	+	+	+	+	+
Fisher exact	+	-	+	+	+	+	+	+
Logistic regression	-	-	+	+	+	+	+	+
Contingency tables	-	-	+	+	+	+	+	+
Kaplan Meier/logrank	+	-	+	+	+	+	+	+
Cox regression	-	-	+	+	+	+	+	+
Many more	-	-	+	+	-	+	+	+

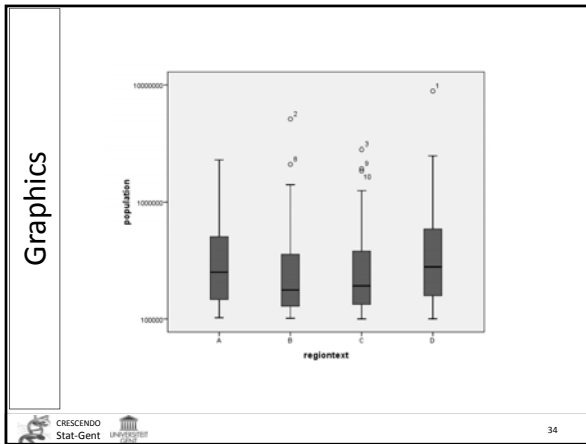
Graphics

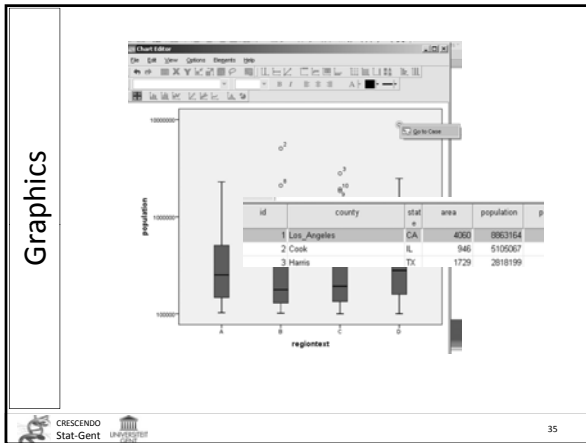
Graphical data exploration is key and supports the quality of the statistical analysis to a large extent. Therefore adequate graphical tools are essential, particularly for the less experienced statistician

- Most packages two-step approach
 - Standard templates
 - Manual editing and polishing
- Some have interactive browsing
- SAS and R have no manual editing

Graphics







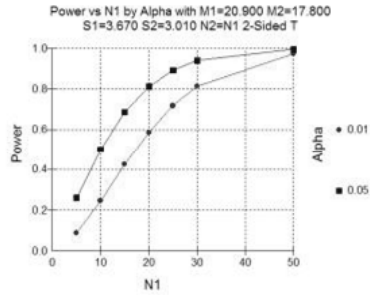
Sample size

Most packages do not provide sample size calculations or only consider a limited number of designs / tests

- Use online calculators
 - <http://www.cs.uiowa.edu/~rlenth/Power/>
- Specialized packages
 - PASS
 - nQUERY
- ... or consult a statistician

CRESCENDO Stat-Gent 36

Sample size



Sample size



Sample size

Select the analysis to be used in your study:

- CI for one proportion
- Test of one proportion
- Test comparing two proportions
- CI for one mean
- One-sample t test (paired or Satterthwaite)
- Two-sample t test (pooled or Satterthwaite)
- Linear regression
- Balanced ANOVA (any model)
- Two variances (F test)
- R-square (multiple correlation)
- Generic chi-square test
- Generic Poisson test
- Online tables of common distributions
- Post study

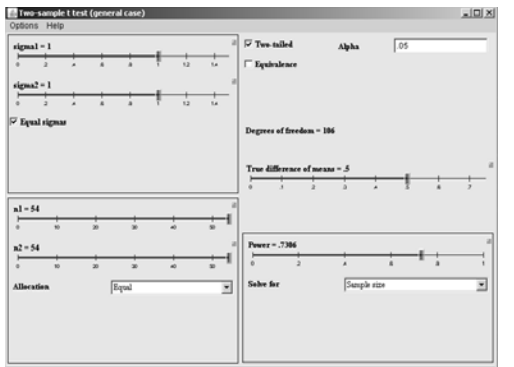
Run selection

This software is intended to be useful in planning statistical studies. Each selection provides a graphical interface for studying the p-varying parameters, and a simple provision for graphing one variable. Each dialog window also offers a Help menu. Please read the "Balanced ANOVA" selection provides another dialog with model.

Note: The dialogs open in separate windows. If you're running it for example, you'll have two "Help" menus there!

You may also download this software to run it on your own PC.

Sample size



Validation

21 CFR Part 11 details the FDA regulatory requirements for processes and controls that must be applied to electronic records. This code applies to drug registration processes and therefore has repercussions for data analysis software and the context in which it is used.

An important part of the validity of statistical software implies that the calculations are correct. It is demonstrated by specific testing and summarized in documents. SAS is considered as a gold standard, but other programs might be just as appropriate.

Pricing

- Extreme dynamic ranging from freeware to several thousands € per year
- As the technically best and most extensive program R is free you pay for interface, user friendliness, manuals, support, performance, etc.
- Most packages have a free trial period ... but it often takes longer to get to know the program
