

# **MEDSTA 2: Regression models in medical research**

**15 January 2014**

Øystein Ariansen Haaland, PHD

Department of Global Public Health and Primary Care,  
University of Bergen

- Based on MEDSTA or HELSTA
- We use STATA
- 15-16 Jan
- 17-18 Feb
- 28-29 Apr
- Lectures from 9-12
- Groups from 12.30-15
- Study from 15-16

- Two home assignments
  - Due on 16 Feb and 27 Apr
- One home exam
  - Due on 18 May
- Oral presentation
  - 12 May
  - Riise et al. (2011)
  - Moster et al. (2010)
  - Skjærven et al. (2012)

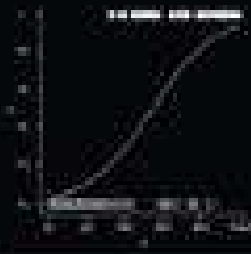
Marit B. Veierød  
Stian Lydersen  
Petter Laake (red.)

# MEDICAL STATISTICS

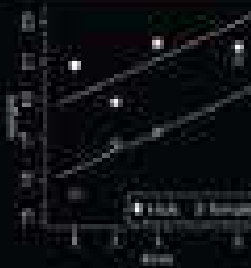
in clinical and  
epidemiological  
research

  
GYLDENDAL  
AKADEMISK

# REGRESSION MODELS AS A TOOL IN MEDICAL RESEARCH



	OR	95% CI	p-value
Gender	1.2	0.8-1.8	0.34
Age	1.05	1.0-1.1	0.001
Smoking	0.8	0.6-1.1	0.001



WERNER VACH

 CRC Press  
Taylor & Francis Group  
4 CRYSTAL DRIVE, HALEY, MD 21054

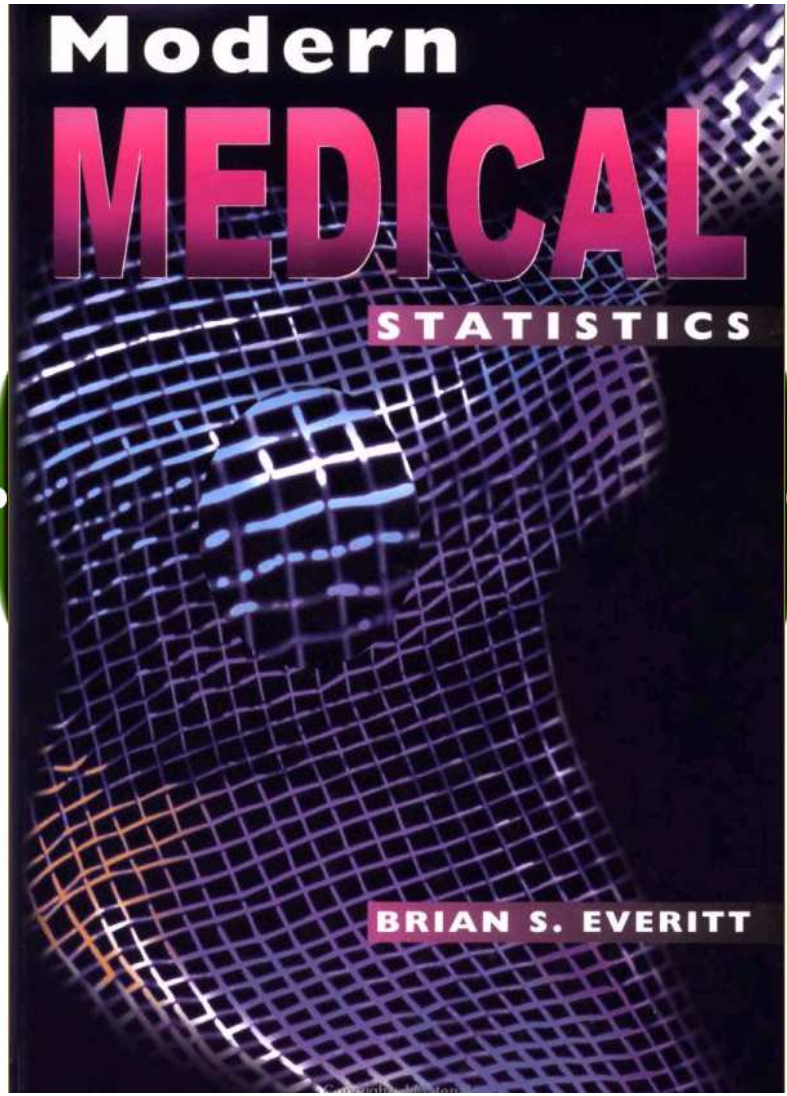
Petter Laake, Anette Hjartåker, Dag S. Thelle og Marit B. Veierød (red.)

# Epidemiologiske og kliniske forskningsmetoder



**Book in Norwegian**

**Gyldendal Akademisk 2007**



Statistical science plays an increasingly important role in medical research. Over the last few decades, many new statistical methods have been developed which have particular relevance for medical researchers and, with the appropriate software now easily available, these techniques can be used almost routinely to great effect. These innovative methods include survival analysis, generalized additive models and Bayesian methods.

**MODERN MEDICAL STATISTICS** covers these essential new techniques at an accessible technical level, its main focus being not on the theory but on the effective practical application of these methods in medical research.

- Exercises are provided at the end of each chapter
- Numerous practical examples throughout
- Up-to-date information on the software packages available that relate directly to the techniques covered in each chapter
- Glossary

**MODERN MEDICAL STATISTICS** is an indispensable practical guide for medical researchers and medical statisticians as well as an ideal text for advanced courses in medical statistics and public health.

**BRIAN S. EVERITT** is Professor of Behavioural Statistics and Head of the Biostatistics and Computing Department at the Institute of Psychiatry, King's College, London

**BY THE SAME AUTHOR**  
Applied Multivariate Data Analysis, 2nd edition  
0 340 74122 8  
Cluster Analysis, 4th edition  
0 340 76119 9



Distributed in the United States of America by  
Oxford University Press Inc., New York

[www.oxfordpublishers.com](http://www.oxfordpublishers.com)



# Intro

- Useful reminders
- Random sample
  - Subset of population
  - Subjects selected at random
  - No bias
- 1000 children selected at random from birth registry = random
- Only 800 participate in study = biased



# Intro

- Two main types of variables
- Continuous variables
  - Can take «all» values in some range (age, height, weight)
- Categorical variables
  - Fit into categories (city, education, older than 67)

# Intro

- Summary statistics for a sample
  - Mean: sum of all observations divided by the number of observations
  - $\bar{x} = \frac{1}{n} \sum_i x_i$
  - Median: order observations and pick the one in the middle
  - Mode: the most common observation

# Intro

- Variance (of a sample)
  - Mean square distance of each observation from the mean
  - $\text{Var}(X) = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$
- Standard deviation
  - Square root of variance
  - $\text{SD}(X) = \sqrt{\frac{1}{n-1} \sum_i (x_i - \bar{x})^2}$

# Intro

- Standard error (of a sample)
  - The standard deviation of the MEAN of the sample
  - $SE(X) = SD(\bar{X}) = \sqrt{\frac{\frac{1}{n-1} \sum_i (x_i - \bar{x})^2}{n}}$
  - SE(X) is small when n is large

# Intro

- Standard error (of a sample)
  - Consider 1000 classes at a university (statistics, biology, physics, etc.)
  - Study height of female students
  - SD corresponds to the variability between all students
  - SE corresponds to the variability between the mean heights of each class

# Intro

- Standard error (of a sample)
  - Large classes will have similar mean heights
  - The mean height of small courses will vary more (extra tall or short individuals will have a greater impact on the mean height)

# Intro

- Normal distribution

- $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$

- $\mu$ : expectation/mean

- $\sigma$ : standard deviation

- $X \sim N(\mu, \sigma^2)$ : « $X$  is normally distributed with mean equal to  $\mu$  and variance equal to  $\sigma^2$ .»

# Intro

- Central limit theorem
  - Means are normally distributed when we consider many observations (large  $n$ )
  - $\frac{\bar{X}-\mu}{SE(X)} \sim N(0,1)$  , or
  - $\sqrt{n} \frac{\bar{X}-\mu}{SD(X)} \sim N(0,1)$

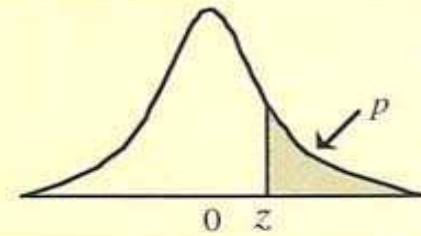


# Intro

- Z-test (significance level  $\alpha = 0.05$ )
  - $H_0: \mu = \mu_0$
  - Two-sided
    - $H_1: \mu \neq \mu_0$
  - One sided
    - $\mu < \mu_0$  or  $\mu > \mu_0$
  - $Z = \frac{\bar{X} - \mu}{SD(\bar{X})} \sim N(0,1)$
  - $H_0: Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$

# Intro

- Z-test (significance level  $\alpha = 0.05$ )
  - $H_0: Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$
  - Two-sided ( $H_1: \mu \neq \mu_0$ )
    - Reject  $H_0$  if  $Z < -1.96$  or  $Z > 1.96$
  - One-sided ( $H_1: \mu < \mu_0$ )
    - Reject  $H_0$  if  $Z < -1.645$
  - One-sided ( $H_1: \mu > \mu_0$ )
    - Reject  $H_0$  if  $Z > 1.645$



Second decimal place of z

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2297	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026

# Intro

- t-test (significance level  $\alpha = 0.05$ )

- $H_0: \mu = \mu_0$

- $H_1: \mu \neq \mu_0, \mu < \mu_0$  or  $\mu > \mu_0$

- $T = \frac{\bar{X} - \mu}{SE(X)} \sim t(n)$

- $H_0: T = \frac{\bar{X} - \mu_0}{\sqrt{\frac{1}{n-1} \frac{\sum_i (x_i - \bar{x})^2}{n}}}$

# Intro

- t-test (significance level  $\alpha = 0.05$ )

$$- H_0: T = \frac{\bar{X} - \mu_0}{\sqrt{\frac{1}{n-1} \frac{\sum_i (x_i - \bar{x})^2}{n}}}$$

– Two-sided ( $H_1: \mu \neq \mu_0$ )

- Reject  $H_0$  if  $T < -t_{0.025}(n)$  or  $T > t_{0.025}(n)$

– One-sided ( $H_1: \mu < \mu_0$  or  $H_1: \mu > \mu_0$ )

- Reject  $H_0$  if  $T < -t_{0.05}(n)$  or  $T > t_{0.05}(n)$

**Table A3 Percentage points of the *t* distribution.**

Adapted from Table 7 of White *et al.* (1979) with permission of the authors and publishers.

d.f.	One-sided <i>P</i> -value								
	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
d.f.	Two-sided <i>P</i> -value								
	0.5	0.2	0.1	0.05	0.02	0.01	0.005	0.002	0.001
1	1.00	3.08	6.31	12.71	31.82	63.66	127.32	318.31	636.62
2	0.82	1.89	2.92	4.30	6.96	9.92	14.09	22.33	31.60
3	0.76	1.64	2.35	3.18	4.54	5.84	7.45	10.21	12.92
4	0.74	1.53	2.13	2.78	3.75	4.60	5.60	7.17	8.61
5	0.73	1.48	2.02	2.57	3.36	4.03	4.77	5.89	6.87
6	0.72	1.44	1.94	2.45	3.14	3.71	4.32	5.21	5.96
7	0.71	1.42	1.90	2.36	3.00	3.50	4.03	4.78	5.41
8	0.71	1.40	1.86	2.31	2.90	3.36	3.83	4.50	5.04
9	0.70	1.38	1.83	2.26	2.82	3.25	3.69	4.30	4.78
10	0.70	1.37	1.81	2.23	2.76	3.17	3.58	4.14	4.59
11	0.70	1.36	1.80	2.20	2.72	3.11	3.50	4.02	4.44
12	0.70	1.36	1.78	2.18	2.68	3.06	3.43	3.93	4.32
13	0.69	1.35	1.77	2.16	2.65	3.01	3.37	3.85	4.22
14	0.69	1.34	1.76	2.14	2.62	2.98	3.33	3.79	4.14
15	0.69	1.34	1.75	2.13	2.60	2.95	3.29	3.73	4.07
16	0.69	1.34	1.75	2.12	2.58	2.92	3.25	3.69	4.02
17	0.69	1.33	1.74	2.11	2.57	2.90	3.22	3.65	3.96
18	0.69	1.33	1.73	2.10	2.55	2.88	3.20	3.61	3.92
19	0.69	1.33	1.73	2.09	2.54	2.86	3.17	3.58	3.88
20	0.69	1.32	1.72	2.09	2.53	2.84	3.15	3.55	3.85
21	0.69	1.32	1.72	2.08	2.52	2.83	3.14	3.53	3.82
22	0.69	1.32	1.72	2.07	2.51	2.82	3.12	3.50	3.79
23	0.68	1.32	1.71	2.07	2.50	2.81	3.10	3.48	3.77
24	0.68	1.32	1.71	2.06	2.49	2.80	3.09	3.47	3.74
25	0.68	1.32	1.71	2.06	2.48	2.79	3.08	3.45	3.72
26	0.68	1.32	1.71	2.06	2.48	2.78	3.07	3.44	3.71
27	0.68	1.31	1.70	2.05	2.47	2.77	3.06	3.42	3.69
28	0.68	1.31	1.70	2.05	2.47	2.76	3.05	3.41	3.67
29	0.68	1.31	1.70	2.04	2.46	2.76	3.04	3.40	3.66
30	0.68	1.31	1.70	2.04	2.46	2.75	3.03	3.38	3.65
40	0.68	1.30	1.68	2.02	2.42	2.70	2.97	3.31	3.55
60	0.68	1.30	1.67	2.00	2.39	2.66	2.92	3.23	3.46
120	0.68	1.29	1.66	1.98	2.36	2.62	2.86	3.16	3.37
∞	0.67	1.28	1.65	1.96	2.33	2.58	2.81	3.09	3.29

# Intro

- Both Z-test and t-test
  - 1 sample
    - $H_0: \mu = \mu_0$
  - 2 samples
    - $H_0: \mu_1 = \mu_2$
  - Paired test
    - Paired measurements  $(x_{1i}, x_{2i})$
    - $d_i = x_{1i} - x_{2i}$
    - $H_0: d = d_0$

# Intro

- Confidence intervals (CIs)
  - Assume that  $H_0$  is true
  - We now expect that there is a 95% chance that the true mean ( $\mu$ ) is contained in a 95% CI
  - Known variance (Z-test)
    - $(\bar{x} - 1.96 \times \sigma/\sqrt{n}, \bar{x} + 1.96 \times \sigma/\sqrt{n})$
  - Unknown variance (t-test)
    - $(\bar{x} - t_{0.025}(n) \times SE(X), \bar{x} + t_{0.025}(n) \times SE(X))$



# Intro

- Type I error
  - Falesly rejecting  $H_0$
  - Probability of Type I error is  $\alpha$ 
    - $P(\text{Reject } H_0 | H_0 \text{ true}) = \alpha$  is a mathematical way to say the same
  - Using  $\alpha = 0.05$  means that we will keep a true  $H_0$  19 out of 20 times

# Intro

- Type II error
  - Falesly keeping  $H_0$
  - Probability of Type II error is  $\beta$ 
    - Mathematical:  $P(\text{Keep } H_0 | H_0 \text{ false}) = \beta$
  - Increasing  $\alpha$  will cause  $\beta$  to decrease
    - Often not desirable
  - A high effect size will cause  $\beta$  to decrease
  - A large sample size will cause  $\beta$  to decrease

# Intro

- Statistical power
  - The probability of rejecting  $H_0$  if  $H_0$  is false
  - The power is 1-minus probability of Type II error,  $(1 - \beta)$
  - Mathematical:  
 $(1 - \beta) = P(\text{reject } H_0 | H_0 \text{ false})$

# Intro

- Statistical power
  - Depends on  $\alpha$ , effect size, and sample size
  - Difficult to change  $\alpha$
  - Effect size is fixed and often unknown
  - Sample size is the only parameter we truly control
  - We will talk more about power later

# Intro

- Research design
  - Randomized controlled trial (RCT)
    - Gold standard
  - Cohort
    - Consider all subjects in a cohort
  - Case-control
    - Comparing cases with controls
  - Ecological
    - Consider whole populations

# **Introduction to Stata**

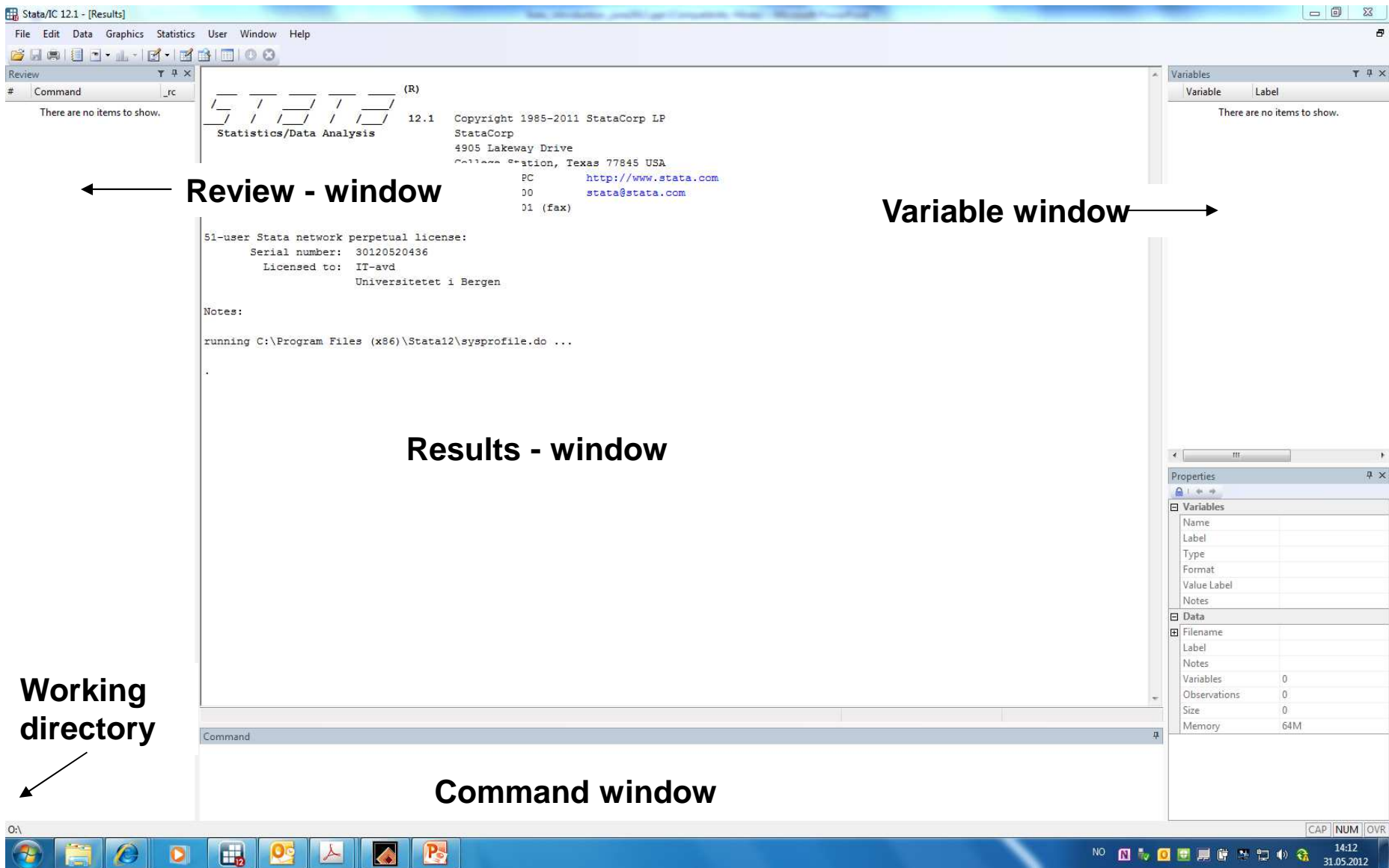
# STATA

- Stata/MP – fast version for dual core and multicore processor computers, 32767 variables)
- Stata/SE – large datasets, 32767 variables
- **Stata Intercooled – standard version, 2047 variables**
- Small Stata – 99 variables, 1000 observations

# Why use STATA?

- Cheap (~1000 NOK for university licence)
- Easy, flexible, publication-quality graphics
- Both point-and-click and commands
- Simple, compact, intuitive command syntax
- User-written methods available for download (one-click installation)
- Widely used in Biostatistics- and Epidemiology-courses and books.
- Can do everything SPSS can do!
- Can do things SPSS can **not** do!





← Review - window

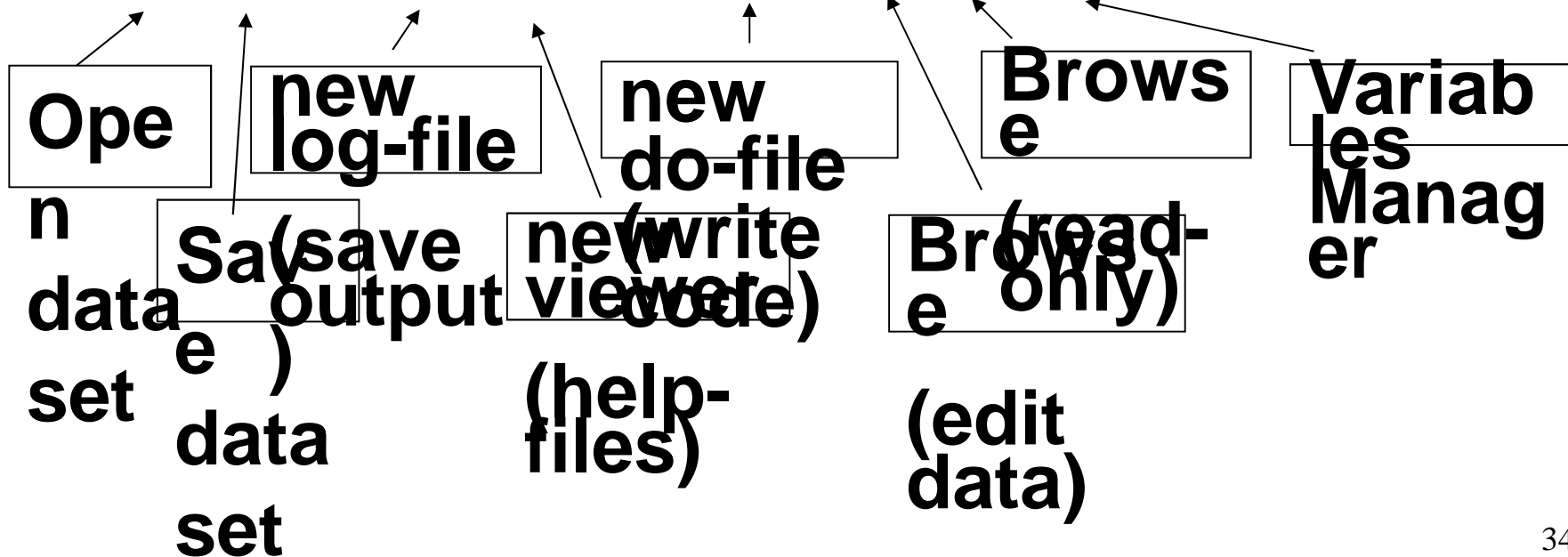
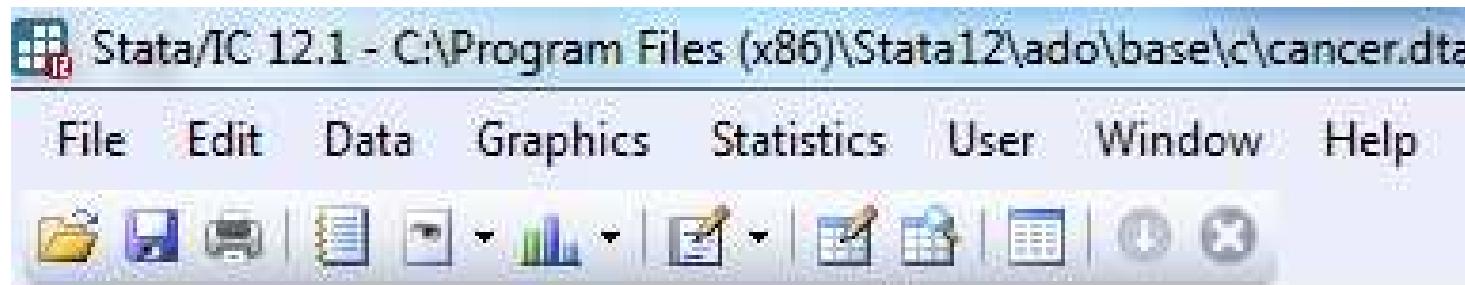
Variable window →

Results - window

Working directory

Command window

# Stata Toolbar



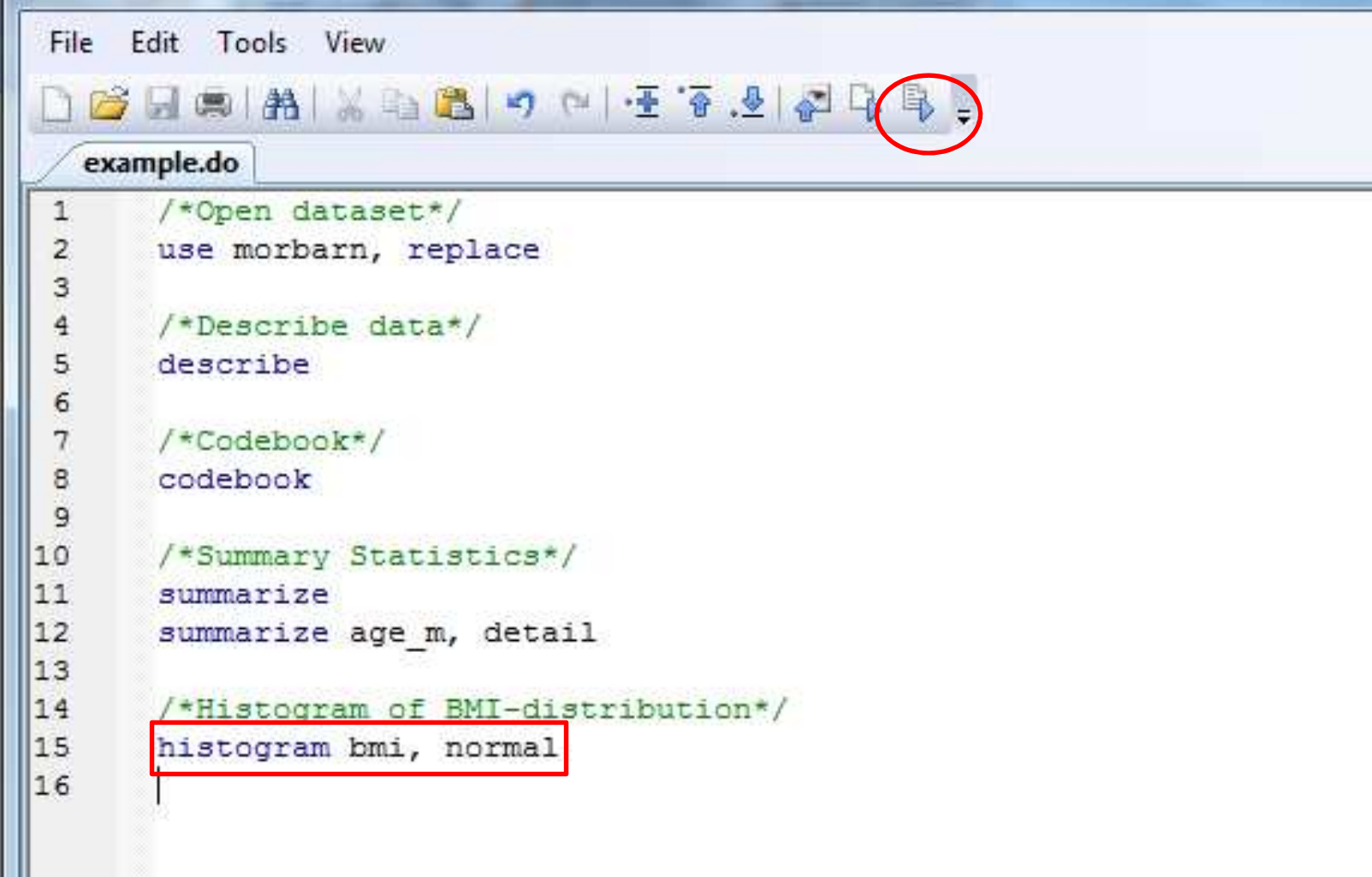
# Main filetypes

- Stata dataset (.dta)
- Do-file(.do): Collection of Stata-code
- Log-file(.smcl/.log) : Code and results stored together, except for graphics
- Graphics file (.gph)

# Ways of working

- Only point-and-click on the menus (not recommended)
- Write commands on the command line
- Write commands in do-file and submit
- Use the menus to generate commands and paste into do-file

# Do-file example



```
File Edit Tools View
example.do
1 /*Open dataset*/
2 use morbarn, replace
3
4 /*Describe data*/
5 describe
6
7 /*Codebook*/
8 codebook
9
10 /*Summary Statistics*/
11 summarize
12 summarize age_m, detail
13
14 /*Histogram of BMI-distribution*/
15 histogram bmi, normal
16
```

# Stata command syntax

- `[prefix:] command [varlist] [if] [in] [weight] [,options]`
- Example: `summarize age_m, detail`
- Options are always separated from the command by a comma
- Carriage Return is command delimiter (End of line = end of command)
- Case sensitive (lowercase letters)
- Type `help commandname` on the command line to see the syntax for a specific command
- Possible abbreviations are underlined in help-file:
  - describe

# Regression Syntax

- General syntax:
  - regression-command outcome exposure, options
- Linear regression:
  - regress sys\_bp height age sex
  - reg sys\_bp height age sex
- Logistic regression:
  - logistic high\_bp height age sex
- Cox regression:
  - stset follow\_up\_time, failure(death)  
stcox sys\_bp age sex

# Categorical exposure variables in Stata

- Example with two categorical exposure variables:
  - agegroup: 1=0-19, 2=20-39, 3=40-59
  - sex: 0=women, 1=men
- Use lowest value as reference:
  - regress sys\_bp i.agegroup i.sex
- Select agegroup=2 and sex=1 as reference:
  - regress sys\_bp b2.agegroup b1.sex



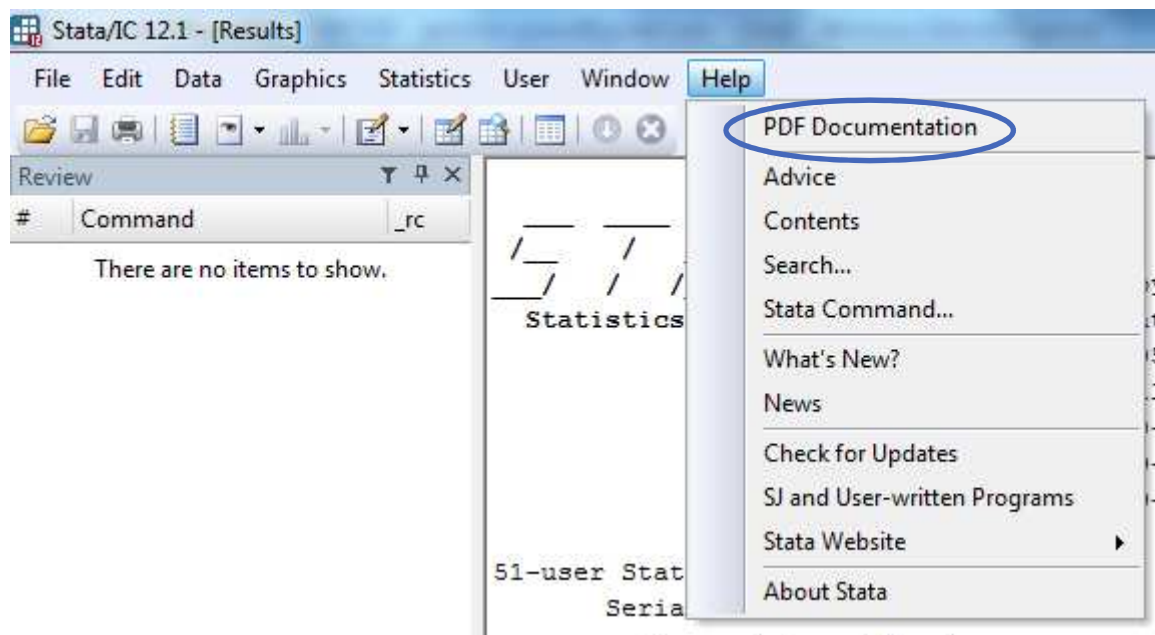
# **Command to make Stata display reference category in regression output**

- `set showbaselevels on, permanently`
- Since the option «permanently» is included you only have to do this once, unless you have to reinstall Stata.
- You can always get back to the default setting :
  - `set showbaselevels off, permanently`

# New in version 11/12

- Variables manager - Add labels, change format etc.
- Command highlighted with colours – easier to read and write code.
- Easier to specify categorical variables and interactions in regression models (you don't need the prefix xi anymore)
- You can have the Data Editor open while you run do-files, use dialog boxes, edit graphs, etc.
- Can choose between several fonts in graphics
- No need to prespecify memory location

# Getting help

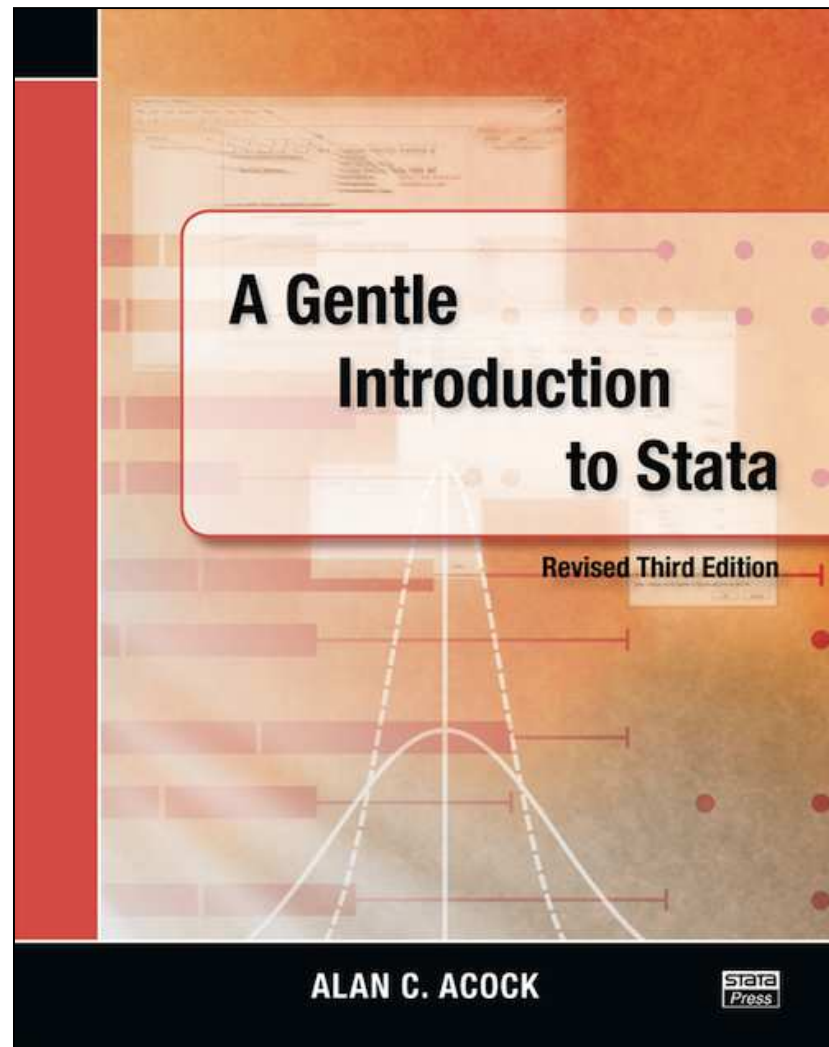
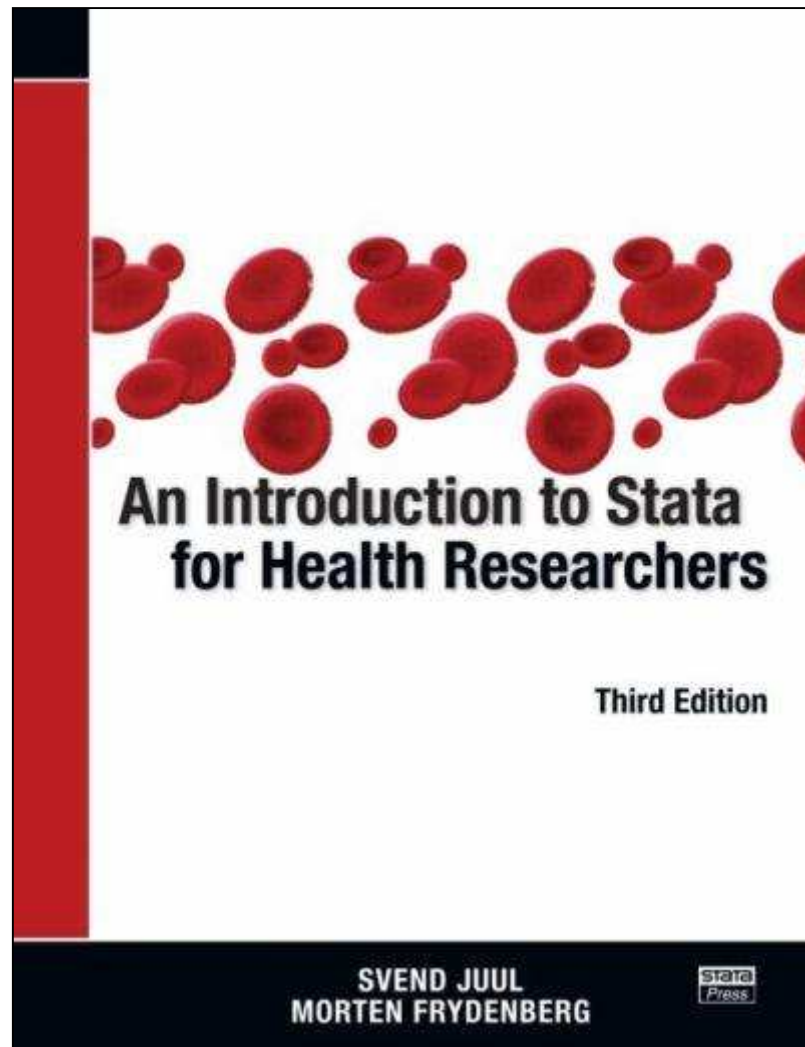


- On the command line:
  - `findit searchword`
  - `search searchword`
  - `help command`

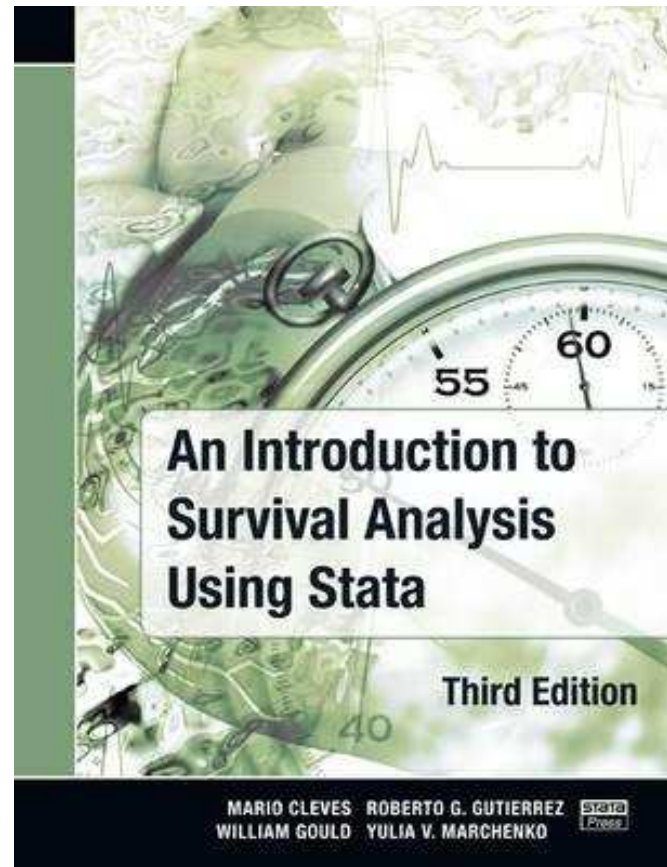
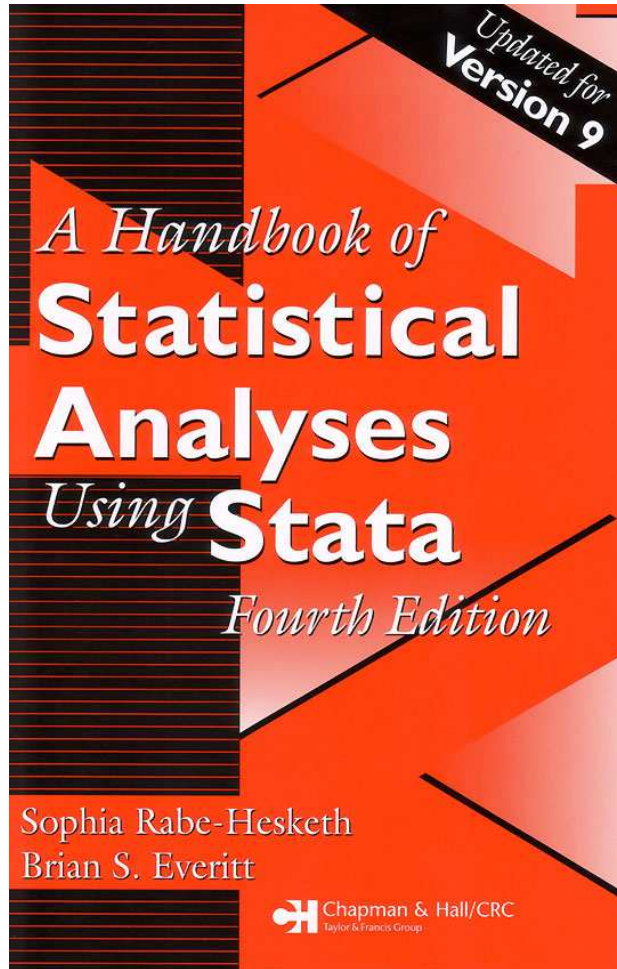
# Web resources

- [www.stata.com/support](http://www.stata.com/support)
- Statalist:  
<http://www.stata.com/statalist/>
- Stata Journal: <http://www.stata-journal.com/>
- UCLA tutorial:  
<http://www.ats.ucla.edu/stat/stata/>

# Good introduction books!



# More books!



# Book on graphics

