Funded by NHS National Institute for Health Research

University of BRISTOL

CTEU Bristol

Creating summary tables using the *sumtable* command

Lauren Scott and Chris Rogers

University of Bristol Clinical Trials and Evaluation Unit

2016 London Stata Users Group meeting

Scott LJ, Rogers CA. Creating summary tables using the sumtable command. Stata Journal. 2015;15(3):775-83

Introduction

- Summary tables are commonly used to describe characteristics within a population
- It is often of interest to compare characteristics of two or more groups
  - Treatment groups in a randomised controlled trial
  - Cohort groups in an observational study
- Summaries may include
  - Counts and percentages
  - Means and standard deviations (SDs)
  - Medians and interquartile ranges (IQRs)

Introduction

Table 1 Baseline characteristics

	Control group (n=121)	Intervention group (n=127)
DEMOGRAPHY		
Male gender (n, %)	75 (62%)	88 (69%)
Age (years; mean, SD)	59 (10.5)	58 (9.3)
Body mass index (mean, SD)	27 (4.9)	26 (4.0)
Smoking status (n, %)		
Non-smoker	80 (66%)	74 (58%)
Ex-smoker	12 (10%)	14 (11%)
Smoker	29 (24%)	39 (31%)
BASELINE SCORES		
Measurement score 1 (median, IQR)	12 (6, 17)	13 (7, 20)
Measurement score 2 (median, IQR)	10 (7, 13)	9 (6, 13)
….	…	…

Introduction

- These tables are typically created by summarising or tabulating data in the Stata output window and copying into reports/documents
- This method may be
  - Time consuming
  - Open to transposition error
  - Frustrating if data are updated

Introduction

- The Stata command *sumtable* can be used to summarise data such that the manual aspect is removed
- *Sumtable* enables the user to present a number of different summary types within one table
- The end result is an Excel spreadsheet that can easily be manipulated for reports or other documents

Introduction

- The resultant Excel spreadsheet contains
  - A label column to describe the variables
  - A level label column to describe categorical variables
  - Three columns of summary data for each specified group and overall

Introduction

- Details of the three summary columns

Variable type	Summary columns		
	First column	Second column	Third column
Binary	Count	Percentage	Missing count
Categorical	Counts	Percentages	Missing count
Continuous (symmetric data)	Mean	SD	Missing count
Continuous (skewed data)	Median	IQR	Missing count
Continuous (range required)	Median	Minimum and maximum	Missing count

The sumtable command

- Syntax:

sumtable *sumvar groupvar* , vartype(*vartype_options*) [vartext(*text*) dp1(#) dp2(#) first(1) last(1) exportname(*text*)]

  - *sumvar* is the variable to be summarised
  - *groupvar* is a group variable by which the *sumvar* data is summarised. The *groupvar* variable must be numeric.

- sumvar, groupvar and vartype are required
- vartext, dp1, dp2, first, last and exportname are optional

Sumtable options

- **vartype**(*name*) must be specified to identify the summary type
  - Binary
  - Binary2
  - Categorical
  - Categorical2
  - Contmean
  - Contmed
  - Contrange
  - Events
  - Headerrow

Sumtable options

- **vartext**(*text*) is the label specified to describe the variable that is being summarised. Defaults to Stata variable label if it exists or variable name if not.
- **dp1**(#) is the number of decimal places displayed for the first statistic in each group. Defaults to zero.
- **dp2**(#) is the number of decimal places displayed for the second statistic in each group. Defaults to one.

Sumtable options

- **first**(1) should be specified for the first row of a table (i.e. the first time any sumtable code is run for this summary table)
- **last**(1) should be specified for the last row of a table (i.e. the last time any sumtable code is run for this summary table).
- **exportname**(*text*) is the name or path name assigned to the Excel summary dataset that is produced from this set of commands. This defaults to ‘summarydatasetexcel’ and is stored in the users current directory.

Types of summaries (vartype options)

- **binary** should be specified for numeric binary variables coded 0 and 1 where only the number of 1's is of interest (displays "n/N" and "%")
- **binary2** is the same as binary without denominators (displays "n and "%")
- **categorical** should be specified for multi-category variables (displays "n/N" and "%" for each category)
- **categorical2** is the same as categorical without denominators (displays "n" and "%" for each category)

Types of summaries (vartype options)

- **contmean** should be specified for continuous variables to be summarised by means and standard deviations (i.e. symmetrical data)
- **contmed** should be specified for continuous variables to be summarised by medians and inter-quartile ranges (i.e. skewed data)
- **contrange** should be specified for continuous variables to be summarised by medians and ranges

Types of summaries (vartype options)

- **events** should be specified for count variables where the total number of events and the number of subjects who experienced the event are of interest

Types of summaries (vartype options)

- **events** should be specified for count variables where the total number of events and the number of subjects who experienced the event are of interest

	Group 1 (n=50)		Group 2 (n=50)		Overall (n=100)	
RBC transfusion (events/patients, % patients)	29/15	30%	54/20	40%	83/35	35%

Types of summaries (vartype options)

- **headerrow** may be used to break up a summary table. It is not necessary, but may be useful to split the final summary table into sections. No summary variable should be specified with this option.

Example 1

- The Stata dataset nlsw88 contains details of 2246 people in America
- Amongst other things it contains data on their hourly wage ($) and whether or not they graduated college (1=graduate, 0=non-graduate)
- Suppose we are interested in summarising hourly wage by graduate status

Example 1

- sysuse nlsw88, clear
- sumtable wage collgrad, first(1) last(1) vartype(contmed)

label	stat1_grp0	stat2_grp0	stat1_grp1	stat2_grp1	stat1_all	stat2_all	miss_grp0	miss_grp1	miss_all
hourly wage	6	(4.0, 8.1)	10	(6.6, 12.4)	6	(4.3, 9.6)	0	0	0

Example 1

- sysuse nlsw88, clear
- sumtable wage collgrad, first(1) last(1) vartype(contmed) dp1(1)

label	stat1_grp0	stat2_grp0	stat1_grp1	stat2_grp1	stat1_all	stat2_all	miss_grp0	miss_grp1	miss_all
hourly wage	6	(4.0, 8.1)	10	(6.6, 12.4)	6	(4.3, 9.6)	0	0	0
hourly wage	5.6	(4.0, 8.1)	9.7	(6.6, 12.4)	6.3	(4.3, 9.6)	0	0	0

Example 1

- sysuse nlsw88, clear
- sumtable wage collgrad, first(1) last(1) vartype(contmed) dp1(2) dp2(2)

label	stat1_grp0	stat2_grp0	stat1_grp1	stat2_grp1	stat1_all	stat2_all	miss_grp0	miss_grp1	miss_all
hourly wage	6	(4.0, 8.1)	10	(6.6, 12.4)	6	(4.3, 9.6)	0	0	0
hourly wage	5.6	(4.0, 8.1)	9.7	(6.6, 12.4)	6.3	(4.3, 9.6)	0	0	0
hourly wage	5.64	(4.03, 8.05)	9.68	(6.63, 12.44)	6.27	(4.26, 9.60)	0	0	0

Example 1

- sysuse nlsw88, clear
- sumtable wage collgrad, first(1) last(1) vartype(contmed) dp1(2) dp2(2) vartext("Hourly wage ($)")

label	stat1_grp0	stat2_grp0	stat1_grp1	stat2_grp1	stat1_all	stat2_all	miss_grp0	miss_grp1	miss_all
hourly wage	6	(4.0, 8.1)	10	(6.6, 12.4)	6	(4.3, 9.6)	0	0	0
hourly wage	5.6	(4.0, 8.1)	9.7	(6.6, 12.4)	6.3	(4.3, 9.6)	0	0	0
hourly wage	5.64	(4.03, 8.05)	9.68	(6.63, 12.44)	6.27	(4.26, 9.60)	0	0	0
Hourly wage ($)	5.64	(4.03, 8.05)	9.68	(6.63, 12.44)	6.27	(4.26, 9.60)	0	0	0

Example 1

- sysuse nlsw88, clear
- sumtable wage collgrad, first(1) last(1) vartype(contmed) dp1(2) dp2(2) vartext("Hourly wage ($)") exportname("Wages by graduate status")

label	stat1_grp0	stat2_grp0	stat1_grp1	stat2_grp1	stat1_all	stat2_all	miss_grp0	miss_grp1	miss_all
hourly wage	6	(4.0, 8.1)	10	(6.6, 12.4)	6	(4.3, 9.6)	0	0	0
hourly wage	5.6	(4.0, 8.1)	9.7	(6.6, 12.4)	6.3	(4.3, 9.6)	0	0	0
hourly wage	5.64	(4.03, 8.05)	9.68	(6.63, 12.44)	6.27	(4.26, 9.60)	0	0	0
Hourly wage ($)	5.64	(4.03, 8.05)	9.68	(6.63, 12.44)	6.27	(4.26, 9.60)	0	0	0

Example 1

- sysuse nlsw88, clear
- sumtable wage collgrad, first(1) vartype(contmed) dp1(1)
- sumtable married collgrad, last(1) vartype(categorical)

label	levellab	stat1_grp0	stat2_grp0	stat1_grp1	stat2_grp1	stat1_all	stat2_all	miss_grp0	miss_grp1	miss_all
hourly wage		5.6	(4.0, 8.1)	9.7	(6.6, 12.4)	6.3	(4.3, 9.6)	0	0	0
married	single	616/1714	35.9%	188/532	35.3%	804/2246	35.8%	0	0	0
	married	1098/1714	64.1%	344/532	64.7%	1442/2246	64.2%	0	0	0

Example 1

- sysuse nlsw88, clear
- sumtable wage collgrad, first(1) vartype(contmed) dp1(1)
- sumtable married collgrad, last(1) vartype(binary)

label	stat1_grp0	stat2_grp0	stat1_grp1	stat2_grp1	stat1_all	stat2_all	miss_grp0	miss_grp1	miss_all
hourly wage	5.6	(4.0, 8.1)	9.7	(6.6, 12.4)	6.3	(4.3, 9.6)	0	0	0
married	1098/1714	64.1%	344/532	64.7%	1442/2246	64.2%	0	0	0

Example 2

- The Stata dataset ‘auto’ contains details about 74 cars including price, weight, etc
- It also contains a variable called ‘foreign’ which identifies whether the car is foreign or domestic (1=Foreign and 0=Domestic)
- Suppose we are interested in comparing foreign and domestic cars

Example 2

- Sysuse auto, clear
- cd “*Desktop\Pathname*”
- sumtable foreign, first(1) vartype(headerrow) vartext("CAR DETAILS")
- sumtable price foreign, vartype(contmed) dp1(0) dp2(0)
- sumtable mpg foreign, vartype(contrange) dp1(0) dp2(0)
- sumtable weight foreign, vartype(contmean) dp1(0) dp2(0)
- sumtable length foreign, vartype(contmean) dp1(1) dp2(1)
- sumtable rep78 foreign, last(1) vartype(categorical) vartext("Repairs since 1978") dp1(0) dp2(1) exportname(“Details by car groups”)

Example 2

label	levellab	stat1_grp0	stat2_grp0	stat1_grp1	stat2_grp1	stat1_all	stat2_all	miss_grp0	miss_grp1	miss_all	rowcount
CAR DETAILS											1
Price		4783	(4184, 6234)	5759	(4499, 7140)	5007	(4195, 6342)	0	0	0	2
Mileage (mpg)		19	(12, 34)	25	(14, 41)	20	(12, 41)	0	0	0	3
Weight (lbs.)		3317	695	2316	433	3019	777	0	0	0	4
Length (in.)		196.1	20.0	168.5	13.7	187.9	22.3	0	0	0	5
Repairs since 1978	1	2/48	4.2%	0/21	0.0%	2/69	2.9%	4	1	5	6
	2	8/48	16.7%	0/21	0.0%	8/69	11.6%	4	1	5	7
	3	27/48	56.3%	3/21	14.3%	30/69	43.5%	4	1	5	8
	4	9/48	18.8%	9/21	42.9%	18/69	26.1%	4	1	5	9
	5	2/48	4.2%	9/21	42.9%	11/69	15.9%	4	1	5	10

Example 2

- Sysuse auto, clear
- cd “*Desktop\Pathname*”
- sumtable foreign, first(1) vartype(headerrow) vartext("CAR DETAILS")
- sumtable price foreign, vartype(contmed) dp1(0) dp2(0)
- sumtable mpg foreign, vartype(contrange) dp1(0) dp2(0)
- sumtable weight foreign, vartype(contmean) dp1(0) dp2(0)
- sumtable length foreign, vartype(contmean) dp1(1) dp2(1)
- sumtable rep78 foreign, last(1) vartype(categorical2) vartext("Repairs since 1978") dp1(0) dp2(1) exportname(“Details by car groups”)

Example 2

label	levellab	stat1_grp0	stat2_grp0	stat1_grp1	stat2_grp1	stat1_all	stat2_all	miss_grp0	miss_grp1	miss_all	rowcount
CAR DETAILS											1
Price		4783	(4184, 6234)	5759	(4499, 7140)	5007	(4195, 6342)	0	0	0	2
Mileage (mpg)		19	(12, 34)	25	(14, 41)	20	(12, 41)	0	0	0	3
Weight (lbs.)		3317	695	2316	433	3019	777	0	0	0	4
Length (in.)		196.1	20.0	168.5	13.7	187.9	22.3	0	0	0	5
Repairs since 1978	1	2	4.2%	0	0.0%	2	2.9%	4	1	5	11
	2	8	16.7%	0	0.0%	8	11.6%	4	1	5	12
	3	27	56.3%	3	14.3%	30	43.5%	4	1	5	13
	4	9	18.8%	9	42.9%	18	26.1%	4	1	5	14
	5	2	4.2%	9	42.9%	11	15.9%	4	1	5	15
											11*Data missing for 5 patients (4, 1).

Example 2

- Sysuse auto, clear
- Label define rep78_label 1 “1 repair” 2 “2 repairs” 3 “3 repairs” 4 “4 repairs” 5 “5 repairs”
- Label values rep78 rep78_label
- cd “*Desktop\Pathname*”
- sumtable foreign, first(1) vartype(headerrow) vartext("CAR DETAILS")
- sumtable price foreign, vartype(contmed) dp1(0) dp2(0)
- sumtable mpg foreign, vartype(contrange) dp1(0) dp2(0)
- sumtable weight foreign, vartype(contmean) dp1(0) dp2(0)
- sumtable length foreign, vartype(contmean) dp1(1) dp2(1)
- sumtable rep78 foreign, last(1) vartype(categorical2) vartext("Repairs since 1978") dp1(0) dp2(1) exportname(“Details by car groups”)

Example 2

label	levellab	stat1_grp0	stat2_grp0	stat1_grp1	stat2_grp1	stat1_all	stat2_all	miss_grp0	miss_grp1	miss_all	rowcount
CAR DETAILS											1
Price		4783	(4184, 6234)	5759	(4499, 7140)	5007	(4195, 6342)	0	0	0	2
Mileage (mpg)		19	(12, 34)	25	(14, 41)	20	(12, 41)	0	0	0	3
Weight (lbs.)		3317	695	2316	433	3019	777	0	0	0	4
Length (in.)		196.1	20.0	168.5	13.7	187.9	22.3	0	0	0	5
Repairs since 1978	1 repair	2	4.2%	0	0.0%	2	2.9%	4	1	5	11
	2 repairs	8	16.7%	0	0.0%	8	11.6%	4	1	5	12
	3 repairs	27	56.3%	3	14.3%	30	43.5%	4	1	5	13
	4 repairs	9	18.8%	9	42.9%	18	26.1%	4	1	5	14
	5 repairs	2	4.2%	9	42.9%	11	15.9%	4	1	5	15
											11*Data missing for 5 patients (4, 1).

Example 2

		Domestic (n=52)		Foreign (n=22)		Overall (n=74)	
CAR DETAILS							
Price (median, IQR)		4783	(4184, 6234)	5759	(4499, 7140)	5007	(4195, 6342)
Weight (lbs.) (mean, SD)		3317	695	2316	433	3019	777
Length (in.) (mean, SD)		196.1	20.0	168.5	13.7	187.9	22.3
Repairs since 1978 (n, %)*							
	1 repair	2	4.2%	0	0.0%	2	2.9%
	2 repairs	8	16.7%	0	0.0%	8	11.6%
	3 repairs	27	56.3%	3	14.3%	30	43.5%
	4 repairs	9	18.8%	9	42.9%	18	26.1%
	5 repairs	2	4.2%	9	42.9%	11	15.9%

* Data missing for 5 cars (4 domestic, 1 foreign)

Conclusions

- Flexible command
- Transposition errors are eliminated
- Regular reporting is faster and more efficient
- Reports are easily replicable
- If data are changed or updated, programs can simply be re-run

Acknowledgements

- This work was supported by
  - The NIHR Bristol Biomedical Research Unit in Cardiovascular Disease
  - University of Bristol
  - Bristol Clinical Trials and Evaluation Unit

Thank you for listening