THE UNIVERSITY of EDINBURGH

SDA: SURVEY DOCUMENT AND ANALYSIS

DO MEN OR WOMEN CONSIDER IT MORE MANLY TO EAT INSTANT NOODLES?*

Teach statistics, don't teach software; with SDA you can!

What Windows Media Player does for sound and video files, SDA does for numeric microdata, although it can also handle aggregate and time-series data.

*Answer: Women do. Can you find the table that tells you so?

SDA allows you to 'read/play', manipulate, analyse and visualise the content of numeric data, especially microdata.

SDA is a server-side application, accessed through any forms-capable web browser (Internet Explorer, Firefox, Chrome etc).

SDA is best at :

- taking a raw data file
- merging it with metadata
- providing variable-level documentation bundled with the data
- doing common types of statistical analysis (descriptive and inferential)
- downloading data and syntax files for further analysis in other statistical software

enjoy5th how much enjoyed 5th year

Description of the Variable

"On the whole, would you say you enjoyed your fifth year at school?"

This question was only asked of sixth year leavers in four versions of the D questionnaire: two of these also included ENJOY4TH.

Only respondents who were sent Levels C and D, versions 3, 4, 5, 6, 7 and 8, have this variable.

Percent	N	Value	Label
82.6	2,005	1	yes
17.4	421	2	no
	1,278	96	left before fifth year [left before S5]
	3,676	98	not asked in questionnaire received [not asked]
	11	99	not answered
	15,767	.	(No Data)
100.0	23,158		Total

Properties

Data type:	numeric
Missing-data codes:	96,98,99
Mean:	1.17
Std Dev:	.38
Record/columns:	1/2340-2342

Source: Scottish school leavers (1980) survey, 1981

WITH SDA, YOU CAN....

teach numeracy, eg how to read tables:
column % versus row %

Frequency Distribution

Cells contain: -**Column percent** -N of cases

		gendercode 1 male	2 female	ROW TOTAL
InstantNoodlesMasculinity	1: very masculine	7.8 / 5	15.6 / 10	11.7 / 15
	2: moderately masculine	32.8 / 21	31.2 / 20	32.0 / 41
	3: neither masculine nor feminine	56.2 / 36	50.0 / 32	53.1 / 68
	4: moderately feminine	3.1 / 2	3.1 / 2	3.1 / 4
	COL TOTAL	100.0 / 64	100.0 / 64	100.0 / 128

Frequency Distribution

Cells contain: -**Row percent** -N of cases

		gendercode 1 male	2 female	ROW TOTAL
InstantNoodlesMasculinity	1: very masculine	33.3 / 5	66.7 / 10	100.0 / 15
	2: moderately masculine	51.2 / 21	48.8 / 20	100.0 / 41
	3: neither masculine nor feminine	52.9 / 36	47.1 / 32	100.0 / 68
	4: moderately feminine	50.0 / 2	50.0 / 2	100.0 / 4
	COL TOTAL	50.0 / 64	50.0 / 64	100.0 / 128

Source: Student ratings of food masculinity, healthiness and palatability, 2015

generate **univariate descriptive statistics** (=information), with or without statistical measures

Variables

Role	Name	Label	Range	MD	Dataset
Row	dam_t_parva_pp	dam t. parva antibody titre	.0-120.5		1

Summary Statistics

Mean =	35.21	Std Dev =	23.70	Coef var =	.67
Median =	31.00	Variance =	561.78	Min =	.00
Mode =	18.50	Skewness =	.74	Max =	120.50
Sum =	195,524.50	Kurtosis =	-.22	Range =	120.50

Source: Infectious diseases of East African livestock (IDEAL) project dataset: Theileria Parva infection outcome data

as well as **multivariate descriptive statistics**, with measures of significance

Variables

Role	Name	Label	Range	MD	Dataset
Row	q4	Carry laptop?	1-2		1
Column	college	College Membership	1-3		1

Frequency Distribution

Cells contain: -**Column percent** -N of cases

		college 1 Humanities & Social Science	2 Science & Engineering	3 Medicine & Veterinary Medicine	ROW TOTAL
q4	1: yes	40.0 / 44	25.9 / 28	21.8 / 67	26.5 / 139
	2: no	60.0 / 66	74.1 / 80	78.2 / 240	73.5 / 386
	COL TOTAL	100.0 / 110	100.0 / 108	100.0 / 307	100.0 / 525
	Means	1.60	1.74	1.78	1.74
	Std Devs	.49	.44	.41	.44

Color coding:	<-2.0	<-1.0	<0.0	>0.0	>1.0	>2.0	Z
N in each cell:	Smaller than expected			Larger than expected			

Summary Statistics

Eta* =	.16	Gamma =	.28	Chisq-P(2) =	13.76 (p= 0.00)
R =	.16	Tau-b =	.14	Chisq-LR(2) =	13.05 (p= 0.00)
Somers' d* =	.12	Tau-c =	.13		

*Row variable treated as the dependent variable.

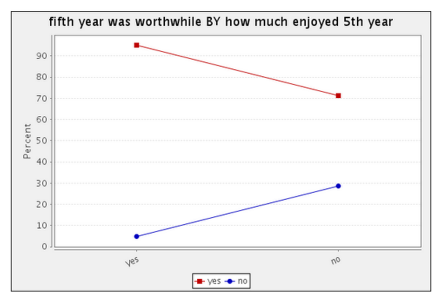
Source: Survey of undergraduate technology use and attitudes, 2012

Summary Statistics for sex = 1(male)

Eta* =	.31	Gamma =	.77	Rao-Scott-P: F(1,10327) =	48.49 (p= 0.00)
R =	.31	Tau-b =	.31	Rao-Scott-LR: F(1,10327) =	36.66 (p= 0.00)
Somers' d* =	.24	Tau-c =	.13	Chisq-P(1) =	52.15
				Chisq-LR(1) =	39.43

*Row variable treated as the dependent variable.

Chart for sex = 1(male)



Source: Scottish school leavers (1980) survey, 1981

analyse with **control and/or filter variables**

compute 90%, 95%, and 99% **confidence intervals**

Variables

Role	Name	Label	Range	MD	Dataset
Row	q4	Carry laptop?	1-2		1
Column	college	College Membership	1-3		1

Frequency Distribution

Cells contain: -**Column percent** -Confidence intervals (95 percent) -N of cases

		college 1 Humanities & Social Science	2 Science & Engineering	3 Medicine & Veterinary Medicine	ROW TOTAL
q4	1: yes	40.0 (31.2-49.6) 44	25.9 (18.4-35.1) 28	21.8 (17.5-26.8) 67	26.5 (22.9-30.4) 139
	2: no	60.0 (50.4-68.8) 66	74.1 (64.9-81.6) 80	78.2 (73.2-82.5) 240	73.5 (69.6-77.1) 386
	COL TOTAL	100.0 --- 110	100.0 --- 108	100.0 --- 307	100.0 --- 525

Color coding:	<-2.0	<-1.0	<0.0	>0.0	>1.0	>2.0	Z
N in each cell:	Smaller than expected			Larger than expected			

Source: Survey of undergraduate technology use and attitudes, 2012

YOU CAN ALSO....

compute **design effects** for complex sample surveys

Frequency Distribution

Cells contain: -**Column percent** -Design effect (DEFT) -Weighted N

		sex 1 male	2 female	ROW TOTAL
stayon	1: yes	27.7 / 1.003 / 292.2	39.2 / 1.018 / 383.6	33.2 / 1.009 / 675.8
	2: no	72.3 / 1.003 / 764.3	60.8 / 1.018 / 594.4	66.8 / 1.009 / 1,358.7
	COL TOTAL	100.0 / 1.000 / 1,056.6	100.0 / 1.000 / 978.0	100.0 / 1.000 / 2,034.5
	Means	1.72	1.61	1.67
	Std Devs	.45	.49	.47
	Unweighted N	1,022	1,036	2,058

Color coding:	<-2.0	<-1.0	<0.0	>0.0	>1.0	>2.0	Z
N in each cell:	Smaller than expected			Larger than expected			

Summary Statistics

Eta* =	.12	Gamma =	-.26	Rao-Scott-P: F(1,9668) =	30.20 (p= 0.00)
R =	-.12	Tau-b =	-.12	Rao-Scott-LR: F(1,9668) =	30.25 (p= 0.00)
Somers' d* =	-.12	Tau-c =	-.12	Chisq-P(1) =	30.97
				Chisq-LR(1) =	31.02

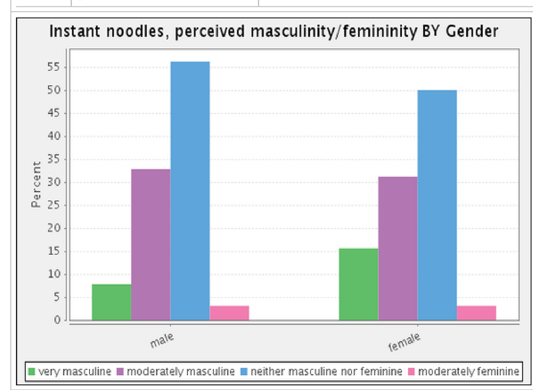
*Row variable treated as the dependent variable.

Source: Scottish school leavers (1980) survey, 1981

produce **data visualisations**, eg histograms

Variables

Role	Name	Label	Range	MD	Dataset
Row	InstantNoodlesMasculinity	Instant noodles, perceived masculinity/femininity	1-4		1
Column	gendercode	Gender	1-2		1



Source: Student ratings of food masculinity, healthiness and palatability, 2015

predict the values of a variable based on the values of other variables, ie **inferential statistics**

Question: Is there a relationship between social class and gender on examination attainment at the end of compulsory schooling/age 16? – a regression analysis

Variables

Role	Name	Label	Range	MD	Dataset
Dependent	totoac	total o grades obtained at a-c	0-12		1
Independent	sex(d:1) - male	sex	0-1		1
Independent	dadclass(m:50)	social class of fathers job	0-1		1
Weight	reweight	weighting factor	.0000000000000000-3.76029968261718750		1

Regression Coefficients / **Test That Each Coefficient = 0**

	B	SE(B)	Beta	SE(Beta)	T-statistic	Probability
dadclass(d:10) - professional	4.602	.130	.299	.008	35.325	.000
dadclass(d:20) - intermediate	3.398	.102	.421	.013	33.375	.000
dadclass(d:31) - skilled(non-manual)	2.542	.115	.216	.010	22.138	.000
dadclass(d:32) - skilled(manual)	.900	.096	.144	.015	9.359	.000
dadclass(d:40) - partly skilled	.650	.103	.076	.012	6.287	.000
dadclass(d:60) - no occupation	.487	.103	.058	.012	4.732	.000
sex(d:1) - male	-.235	.037	-.039	.006	-6.286	.000
Constant	1.474	.093			15.818	.000

Color coding:	<-2.0	<-1.0	<0.0	>0.0	>1.0	>2.0	T
Effect of each variable:	Negative			Positive			

Multiple R = .415 R-Squared = .173 Adjusted R-Squared = .172 SE of Estimate (Root MSE) = 2.756

Global Tests for Groups of Variables

Group	Wald Chi-sq	df Numerator	df Denominator	Adjusted Wald F	P
All independent variables	4,537.200	7	21753	647.993	.000
dadclass(m:50)	4,497.434	6	21754	749.400	.000

Source: Scottish school leavers (1980) survey, 1981

SUMMARY

For the **teacher/researcher,** SDA provides

- a source of data for exercises/assignments
- a facility with which to teach basic numeracy without having to teach software
- teach introductory and intermediate level statistics courses without having to teach software
- perform simple recodes, collapse categories, etc
- define dummy variables and interactions on-the-fly
- perform more complex variable recodes and computes and share them with others
- statistical functions such as (correlations, regressions (multiple, logit/probit), with robust statistical measures, such as chi-square, R^2, etc
- teach difference between simple random sample and complex sample designs and how they affect measures (ie design effects).
- distance education, with no software licencing issues
- share your own data with other researchers or with a class
- for sensitive data, SDA is FISMA-moderate compliant: we can embargo individual variables, variable combinations, minimum cell counts, etc

For the **user,** it provides

- easy access to microdata, without a steep learning curve
- search for variables, within and among datasets in the collection
- standard descriptive statistics, univariate and multivariate
- standard statistical procedures for categorical and continuous variables: cross-tabulations, comparison of means, correlations, regressions
- context sensitive help
- output weighted by default (where applicable)
- download data files, including recoded/computed variables, for use in other software, eg SAS, SPSS, Stata, R, Excel
- basic data visualisation capabilities, can be saved as html
- a transferrable skill: SDA is also used as a data interface by other institutions such as ICPSR, the Roper Center, IPUMS, etc.

SDA is available from the Data Library web page
http://www.ed.ac.uk/is/data-library
or directly at http://stats.datalib.edina.ac.uk/sda/

To have your data files loaded into SDA, contact the Data Library datalib@ed.ac.uk

---

University of Edinburgh. Data Library and consultancy
datalib@ed.ac.uk