11-1-2022

Getting started Analyzing Data in SPSS

Kristi Thompson
Western University

Data analysis with *SPSS*

OR AS MUCH AS I CAN REASONABLY CRAM INTO A SHORT SESSION

About this session

Part 1: a quick and ***very*** dirty introduction to some basic, practical statistics for data analysis

Part 2: a demonstration of how to run some common procedures in SPSS using two sample datasets

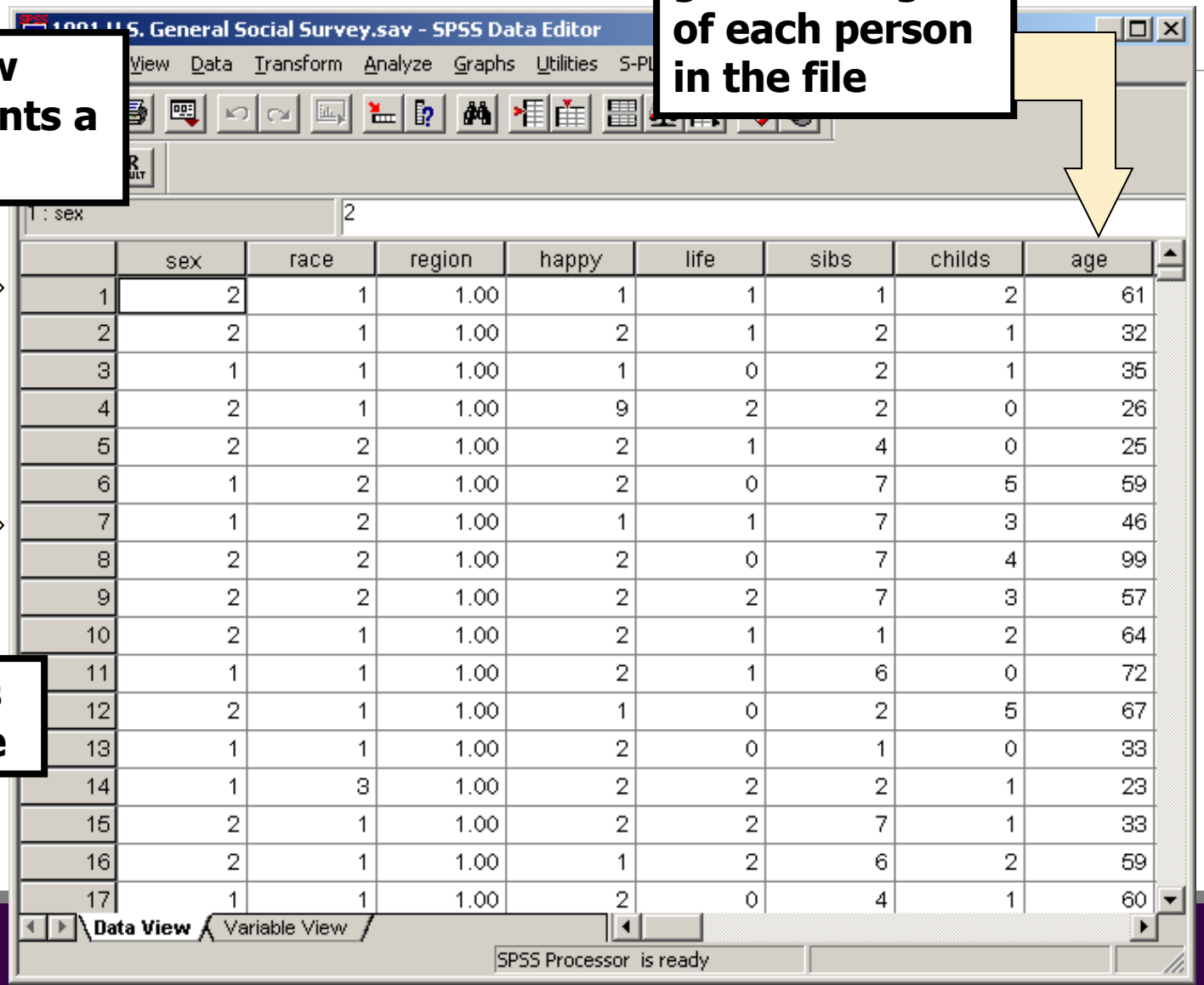
Part 3: I make the slides, data, a walkthrough, and a video recording of the session available for you to study at your leisure

SPSS data view

This row represents a person

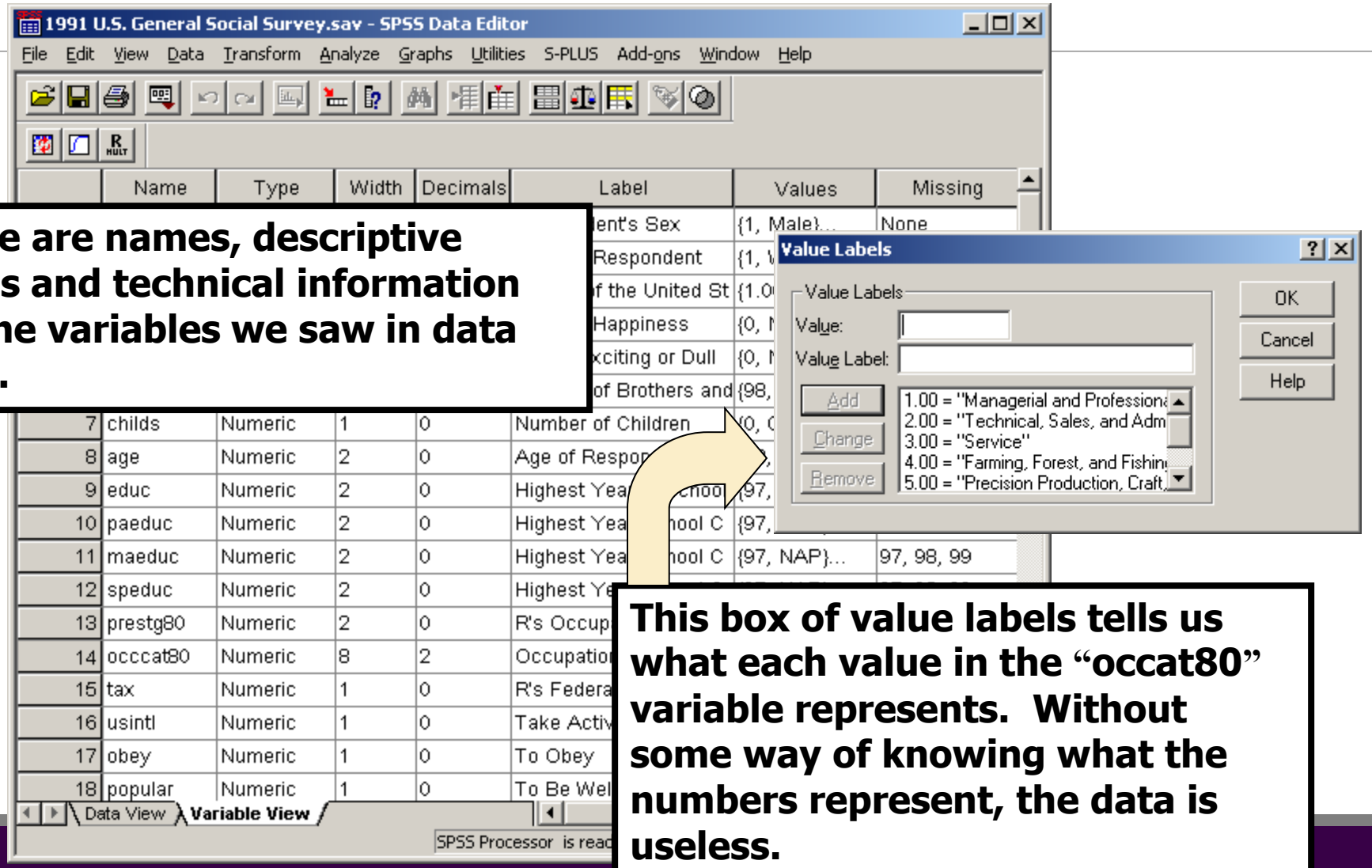
So does this one

This variable gives the age of each person in the file



	sex	race	region	happy	life	sibs	childs	age
1	2	1	1.00	1	1	1	2	61
2	2	1	1.00	2	1	2	1	32
3	1	1	1.00	1	0	2	1	35
4	2	1	1.00	9	2	2	0	26
5	2	2	1.00	2	1	4	0	25
6	1	2	1.00	2	0	7	5	59
7	1	2	1.00	1	1	7	3	46
8	2	2	1.00	2	0	7	4	99
9	2	2	1.00	2	2	7	3	57
10	2	1	1.00	2	1	1	2	64
11	1	1	1.00	2	1	6	0	72
12	2	1	1.00	1	0	2	5	67
13	1	1	1.00	2	0	1	0	33
14	1	3	1.00	2	2	2	1	23
15	2	1	1.00	2	2	7	1	33
16	2	1	1.00	1	2	6	2	59
17	1	1	1.00	2	0	4	1	60

SPSS variable view



These are names, descriptive labels and technical information for the variables we saw in data view.

This box of value labels tells us what each value in the "occat80" variable represents. Without some way of knowing what the numbers represent, the data is useless.

How statistics is often taught

$$s = \sqrt{\frac{\sum_{i=1}^{n} \ln(x_i)^2 - \left[\sum_{i=1}^{n} \ln(x_i)\right]^2 / n}{n-1}}$$

$$\hat{\psi} = \exp\left(\frac{O-E}{V}\right)$$

$$O = a$$

$$E = (a+b)(a+c)/n$$

$$V = \frac{(a+b)(c+d)(a+c)(b+d)}{n^2(n-1)}$$

$$z_p = \frac{O-E}{\sqrt{V}}$$

$$CI = \exp\left(\frac{(O-E) \pm z_{\alpha/2}\sqrt{V}}{V}\right)$$

Take Σx where xi = ai+bi+ci+di…

What you really need to know

Even if you are running your own analysis, the equations don't really matter: the computer will do the math for you

For most of your career, what you really need is to know how to read statistics well enough to extract information from articles

It all comes down to two numbers, numbers to help you answer two questions:

- Is there an effect?
- How big is it?

Zero or Not Zero – the Null Hypothesis

Statistics is generally about trying to prove that something (an independent variable) has an effect on another thing (a dependent or outcome variable). To do that, you assume that the effect is zero (the null hypothesis) then try to prove yourself wrong.

- The independent variable is innocent (of having an effect) until proven guilty, beyond a reasonable doubt!
- “Beyond a reasonable doubt” is calculated mathematically – that is all statistical significance is.

Precision and Significance

All research has some uncertainty associated with it

Quantitative statistical techniques let you be mathematically precise about how uncertain you are – let you calculate the error

- What this is in fact doing is using the random variation in your data to calculate whether the effect you are seeing is likely to in fact be a result of random variation in your population
- Technically, if the true effect in the general population is zero, how likely are you to have drawn a random sample showing an effect as big as the one you are seeing?
- But thinking of the significance value as your “likelihood of being wrong about there being an effect” is not terrible

A result will be considered statistically significant if the calculated effect size is large compared to the calculated error. This is represented by the **P value,** which is just a ratio of effect size to error expressed as a percentage.

Significance: Statistical and Practical

A result is usually considered statistically significant if it is calculated that the likelihood of the effect size being due to random variation in the population is less than 5%, or the p value is < .05

- Smaller is better. You'll see people talking about highly or very highly significant results – p values of less than .01 or .001

The calculated effect size is your best estimate, based on this sample of data, of the true effect in the population. You can also calculate a range for the effect – "correct within 3 percentage points, 19 times out of 20" is how this is often reported in political polls

A result can be very highly statistically significant and yet have no practical significance whatsoever. Again, all statistical significance indicates is that the theoretical "true value" is ***probably not zero***. The larger your sample, the more precisely calculated the error. With a large enough sample, even tiny effects can be significant.

It's possible to calculate how large a sample needs to be in order to detect an effect of a given size (Cochran's formula).

Some SPSS output

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.277 [a]	0.077	0.074	2.173

[a] Predictors: (Constant), EDU: Education YRS, Q1: Gender, AGE: Age (logical)

ANOVA [b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	388.576	3	129.525	27.425	0.000 [a]
	Residual	4680.415	991	4.723		
	Total	5068.991	994			

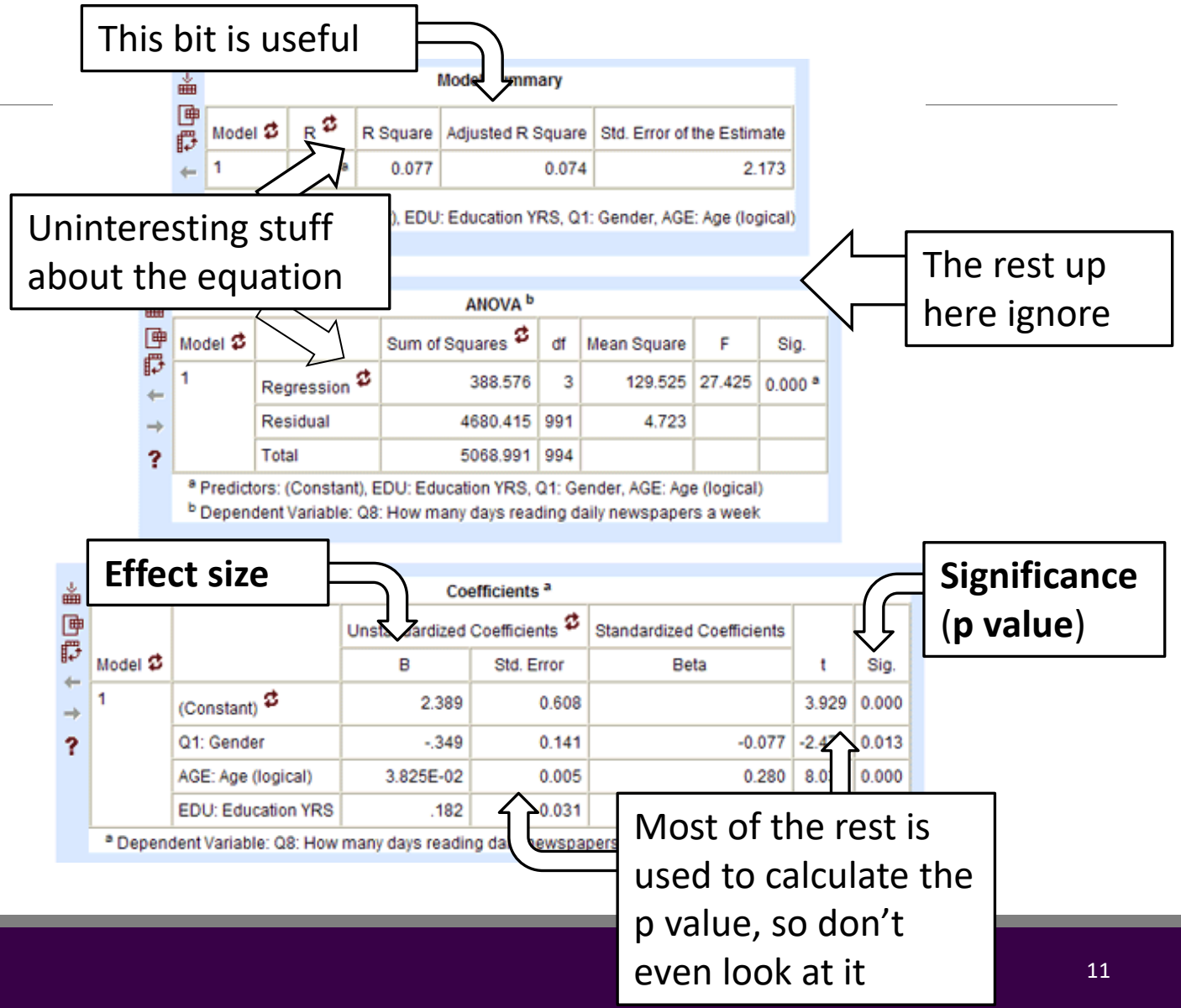
[a] Predictors: (Constant), EDU: Education YRS, Q1: Gender, AGE: Age (logical)
[b] Dependent Variable: Q8: How many days reading daily newspapers a week

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.389	0.608		3.929	0.000
	Q1: Gender	-.349	0.141	-0.077	-2.478	0.013
	AGE: Age (logical)	3.825E-02	0.005	0.280	8.030	0.000
	EDU: Education YRS	.182	0.031	0.211	5.946	0.000

[a] Dependent Variable: Q8: How many days reading daily newspapers a week

How I look at **SPSS** output



This is the information that gets published in research summaries. Read up on significance and effect size

Controlling Vio[illegible]o the Community: An Evaluation of the Boston

"The log-rank statistic co[illegible]med that postrelease time-to-arrest for BRI subjects was significa[illegible] different from the control subjects [illegible]mes (p = .0039) and for violent crimes (p = .0309). A[illegible]ear postrelease, 36.1 percent of BRI participants had been arrested for a new crime, while 5[illegible].1 per[illegible]nt of control group subjects had been arrested for a new crime (Figure 1)."

Anthony A. Braga, Anne M. Piehl, and David Hureau. Journal of Research in Crime and Delinquen[illegible] 46(4)

Highly Significant:

Less than .01

Significant:

Less than .05

Effect size:

15% reduction

Variables: Levels of Measurement

Categorical (aka Nominal): numbers in dataset are labels for categories

- E.g. first language spoken. 1=English, 2=French, 3=Spanish, 4=Ojibway etc.

Ordinal: numbers label ordered categories that go from smallest to largest

- E.g. educational attainment: 1=Less than H.S., 2=H.S. Diploma, 3=Community College Diploma, 4=University Degree
- E.g. how did the food taste ranging from 1=terrible to 5 = great. (Scale could have gone from 1 to 10. Actual numbers are arbitrary, order is not.)

Interval: number is an actual number that counts or measures something

- E.g. income, age, amount of time wasted on Wordle…

Why does it matter? Choice of statistical technique depends primarily on two things: number of variables being analyzed, and whether the variables are continuous or categorical.

Basically… don't try to calculate the mean of "language spoken".

Univariate statistics

Categorical

- Average: mode – the most common value.

Ordinal

- Average: median, the number with half the values lower and half higher. (Mode works, but usually less informative.)

Interval / continuous

- Central tendency: mean. Median can also be useful.
- Dispersion: standard deviation.

Bivariate Statistics

Two Categorical Variables: the Cross-tab

Chief justice numeric code * direction of the decision Crosstabulation

			direction of the decision		
			conservative	liberal	Total
Chief justice numeric code	Burger	Count	1510	1243	2753
		% within Chief justice numeric code	54.8%	45.2%	100.0%
	Rehnquist	Count	1106	896	2002
		% within Chief justice numeric code	55.2%	44.8%	100.0%
	Roberts	Count	453	408	861
		% within Chief justice numeric code	52.6%	47.4%	100.0%
	Vinson	Count	386	390	776
		% within Chief justice numeric code	49.7%	50.3%	100.0%
	Warren	Count	713	1460	2173
		% within Chief justice numeric code	32.8%	67.2%	100.0%
Total		Count	4168	4397	8565
		% within Chief justice numeric code	48.7%	51.3%	100.0%

One Categorical and One Continuous Variable: Compare Means / T-test

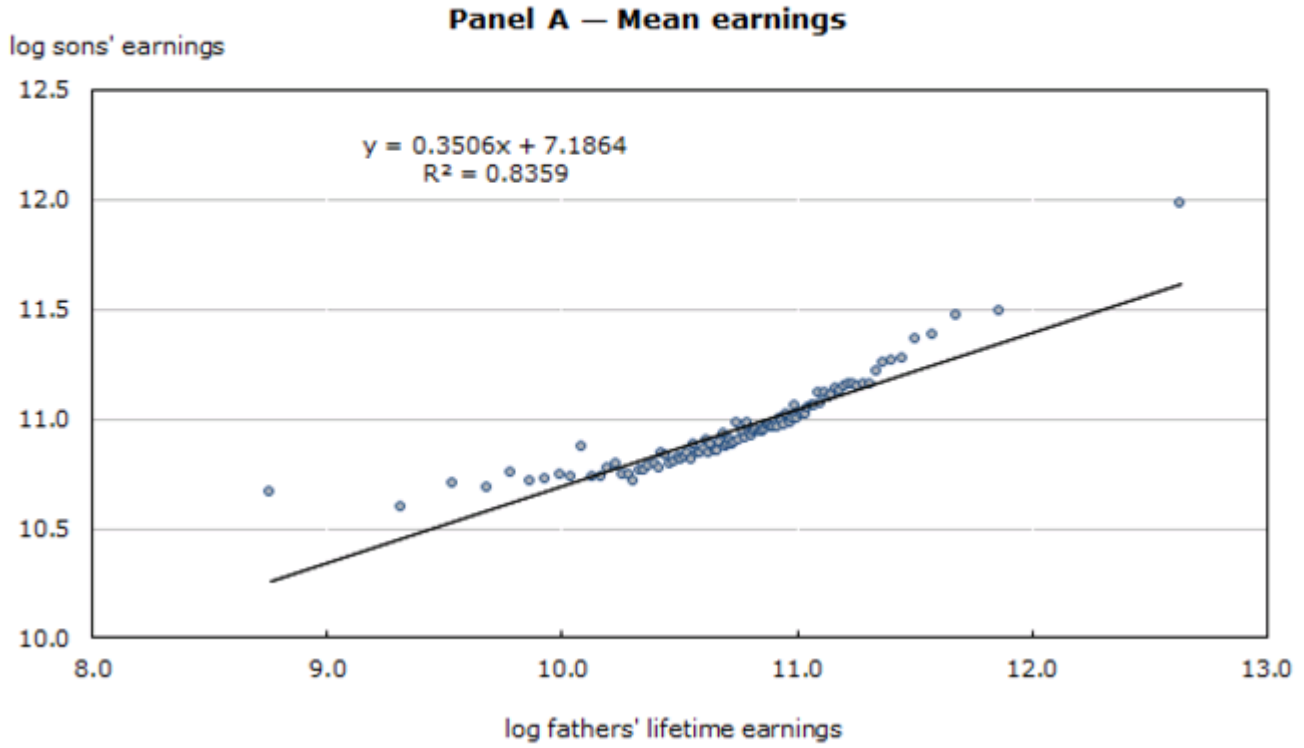
Data Source: U.S. Educational Longitudinal Survey 2002

Group Statistics

	Uses computers in math class	N	Mean	Std. Deviation	Std. Error Mean
Math test standardized score	No	11510	50.6461	9.73192	.09071
	Yes	2958	48.0220	10.99147	.20211

Two Continuous Variables: Correlations

Chart 5
The relationship between fathers' and sons' earnings



Source: Statistics Canada, authors' calculations based on data from the Intergenerational Income Database.

Multivariate Statistics

IT'S ALL DEPENDENT ON…

The dependent variable

Remember:

- Dependent or outcome – thing you think is being influenced by the independent variable

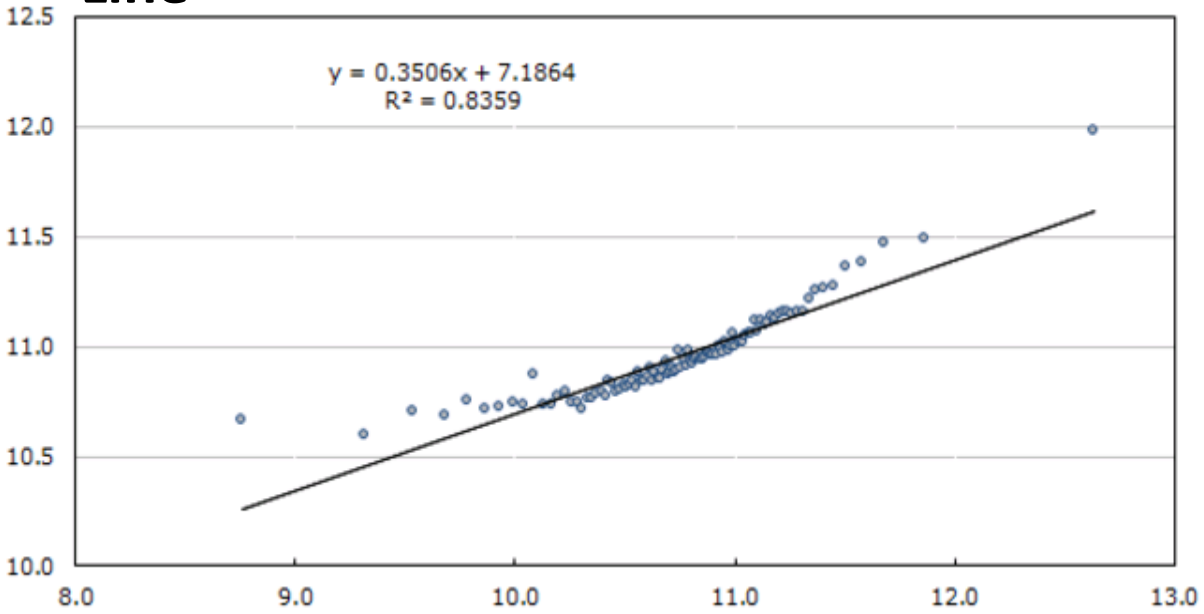
e.g. influence of bad weather weather on food prices. Weather damaging crops may have an effect on food prices; unlikely that food prices will affect the weather.

Choice of statistical procedure depends on the dependent variable. If it's continuous (income, age, test score) use **regression** (basically the multivariate form of a correlation) or **ANOVA** (similar but easier to use with categorical independent variables)

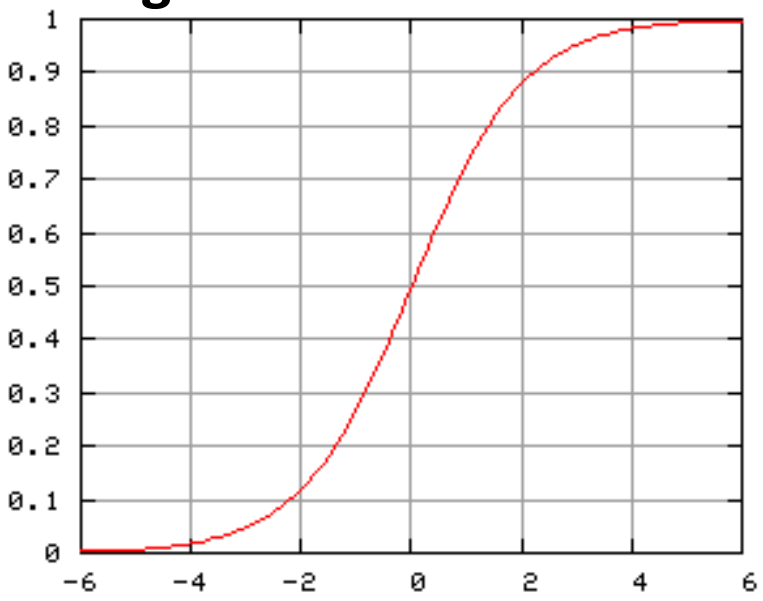
If the dependent is categorical or binary… life is being difficult. This happens a lot.

Logistic regression

Line



Logit curve



Resources

UCLA: Annotated SPSS Output

UCLA: Choosing the Correct Statistical Test

GNU PSPP – free alternative to / clone of SPSS

Basic SPSS Commands at U of Guelph

Statistical Methods for Practice and Research: A Guide to Data Analysis Using SPSS – ebook available through the library