Q-STEP STATA 'HOW TO' GUIDES:

BIVARIATE ASSOCIATIONS (CROSS TABS, CORRELATION, AND MEANS COMPARISONS) IN STATA

Creator: Joshua Townsley

One of the first steps in the process of quantitative analysis is to test for simple associations between two variables (hence, 'bivariate'). The type of bivariate analysis we run depends on the level of measurement of the variables. This guide will outline how to run common methods of bivariate analysis – cross tabulation, correlation, and means comparisons – and how to test for statistical significance.

Remember first to clean your variables first – i.e. check that missing values are indeed set as missing values. This can be done by examining the coding of the variable first (**codebook profile_gross_household, tab(100)**) then setting desired values as missing (**mvdecode profile_gross_household, mv(16 17)**).

Q-Step
A step-change in quantitative social science skills
Funded by the Nuffield Foundation, ESRC and HEFCE
Nuffield Foundation
Sheffield Methods Institute.
The University Of Sheffield.

Cross Tabulation

Cross tabulation (or 'cross tabs') are appropriate when the two variables you want to analyse are categorical (i.e. ordinal or nominal).

We have a variable measuring age groups ("agecat2") and one measuring education level ("education"). We want to find out if there is a bivariate association between these two variables using the CROSSTAB (or tab) function, stipulating that we want column percentages as well as the raw numbers ("col").

The variable that goes first will be entered along the rows, with the second variable along the columns.

tab education agecat2, col

As we can see, there is a relationship between the two variables. Generally speaking, younger people are more likely to have a degree, while older people are more likely to have no qualifications and GCSE/Equivalent.

RECODE of profile_education_level (Education qualification (highest attained))	RECODE of age (Age)				
	Under 30	30-44	45-59	60+	Total
No Quals	11	17	43	153	224
	1.92	2.31	4.34	10.34	5.93
GCSE/Equiv	32	99	197	227	555
	5.57	13.45	19.88	15.35	14.68
A Level/Equiv	237	116	117	123	593
	41.29	15.76	11.81	8.32	15.69
University	240	345	321	444	1,350
	41.81	46.88	32.39	30.02	35.71
Other Qual	54	159	313	532	1,058
	9.41	21.60	31.58	35.97	27.99
Total	574	736	991	1,479	3,780
	100.00	100.00	100.00	100.00	100.00

To test for statistical significance, we can run a chi-square test and obtain a p-value for the test.

tab education agecat2, col chi2

The test produces a p-value of '0.000' (which we interpret as '<0.001'). This means that we reject the Null Hypothesis (stipulating that there is no statistically significant relationship between the variables in the population) at the 99.9% confidence level.

Correlation

Correlation is an appropriate test to run when we have two interval level variables. likeCon measures how much a respondent likes or dislikes the Conservative Party (on a scale of 0-10). likeLab measures how much a respondent likes or dislikes the Labour Party (on a scale of 0-10).

To run a correlation between these variables, we can use the correlation function (or "corr").

corr likeCon likeCon

	likeCon	likeLab
likeCon	1.0000	
likeLab	-0.4579	1.0000

As we can see, there is a weak negative correlation between these two variables. What do you think this means?

To test for statistical significance, we can stipulate that we want a significance test ("sig") using the pwcorr function.

pwcorr likeCon likeLab, sig

	likeCon	likeLab
likeCon	1.0000	
likeLab	-0.4579 0.0000	1.0000

The significance test produces a p-value for this correlation of '0.000' (which we interpret as '<0.001'). This means that we reject the Null Hypothesis (stipulating that there is no statistically significant relationship between the variables in the population) at the 99.9% confidence level.

Means Comparison

Means comparisons are an appropriate test to run when we have categorical and interval level variables. likeCon measures how much a respondent likes or dislikes the Conservative Party (on a scale of 0-10). profile_gender measures a respondent's gender (Male/Female).

The aim here is to compare the mean likeCon score between men and women (i.e. profile_gender).

To compare the means we can use the ttest function:

ttest likeCon, by(profile_gender)

This will produce a table with various information about the relationship:

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
Male	1,511	4.508934	.0831204	3.23102	4.345891	4.671978
Female	1,519	4.119157	.0792868	3.090151	3.963634	4.274681
combined	3,030	4.313531	.0575282	3.166667	4.200733	4.42633
diff		.3897771	.1148577		.1645702	.6149841

diff = mean(Male) - mean(Female) t = 3.3936
Ho: diff = 0 degrees of freedom = 3028

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0
Pr(T < t) = 0.9997 Pr(|T| > |t|) = 0.0007 Pr(T > t) = 0.0003

The table shows the mean likeCon score between Male and Female (4.5 vs 4.1), and the difference between them (0.4). At the bottom, the table shows the p-value associated with the difference being less than 0, not equal to =, and more than 0.

In this case, the probability that there is a difference between Male and Female (i.e. the probability that the difference is not 0) is less than .001.

Tasks:

1. Run a cross tab and chi-square test between levels of political attention (polatt) and gender (profile_gender). How do we interpret the results?

2. Run a correlation between likeLD (how much someone likes the Lib Dems) and likeUKIP (how much someone likes UKIP). How do we interpret the results?

3. Is there a relationship between gender and how much respondents like UKIP?