Notes on methods for the political survey

Chris Lightfoot, ⟨chris@ex-parrot.com⟩

July 16, 2003

Abstract

Some brief notes summarising the statistics and method in use.

1 Definitions

x_i	ith variate drawn from distribution X
N	number of observations x_i
$\mathbf{s}_x$	sum of observations of x_i, $\sum_i^N x_i$
$\mathbf{s}_{xx}$	sum of observations of x_i^2, $\sum_i^N x_i^2$
$\bar{x}$	mean of the x_i, $\mathbf{s}_x/N$
$\mathrm{cov}(x, y)$	covariance between X and Y, $\overline{xy} - \bar{x}\bar{y} = \mathbf{s}_{xy}/N - \mathbf{s}_x\mathbf{s}_y/N^2$
$\mathbf{ss}_x$	$\mathbf{s}_x - \bar{x}$
$a + bx$	best fit line for y as a function of x; $b = \mathbf{ss}_{xy}/\mathbf{ss}_{xx}$, $a = \bar{y} - b\bar{x}$
r^2	correlation coefficient for linear fit; $r^2 = \mathbf{ss}_{xy}^2/\mathbf{ss}_{xx}\mathbf{ss}_{yy}$
C_{mn}	covariance between variables m and n; element in covariance matrix C

2 Statements and answers

Each *statement* or proposition has a 'normal' and a 'converse' form. These are supposed to be antonyms, for instance

normal	Family is more important than society.
converse	Society is more important than family.

... though many are more ambiguous than that.

Each *answer* is an integer between -2 and $+2$ inclusive; we assign these the following labels:

-2	disagree strongly
-1	disagree
0	no opinion
$+1$	agree
$+2$	agree strongly

Each respondent is presented with a random mixture of normal and converse statements; the statements are presented in random order. A few statements are repeated in both forms.

When we put the same statement to a respondent in both normal and converse forms, we record both answers, x and y. We then compute a best-fit line $y = f(x) = a + bx$ between each pair of answers for that statement, and measure the goodness-of-fit r^2. For the remaining analysis we use either answers to the statement in its normal form or, if they are not available, answers in the converse form mapped to the normal form through $f(x)$. We use r^2 to verify that the two forms of the statement are fair antonyms.

3 Principal component analysis

Once we have a large number of responses to all the statements, we can compute the covariance between the mth and nth statements, $\mathrm{cov}(x_m, x_n) = C_{mn}$ for each m, n. C is called the covariance matrix. Its eigenvectors $\mathbf{e}_k$ define *principal axes*, linear combinations of the various statements which describe the directions of maximum variation in the data. The corresponding eigenvalues λ_k tell us how significant each axis of variation is, the eigenvectors with the largest λ_k being the most significant.

4 References

- Covariance, `http://mathworld.wolfram.com/Covariance.html`
- Linear least squares fitting, `http://mathworld.wolfram.com/LeastSquaresFitting.html`
- Principal component analysis, `http://www.cis.hut.fi/~jhollmen/dippa/node30.html`