Psychometric Theory

Basic Concepts of Variance, Covariance and Correlation

Estimates of Central Tendency

- Consider a set of observations X = $\{x_1, x_2 \ldots x_n\}$
- What is the best way to characterize this set
  - Mode: most frequent observation
  - Median: middle of ranked observations

Mean:

$$\text{Arithmetic} = \overline{X} = \sum_{1}^{n} (Xi)/N$$

$$\text{Geometric} = \sqrt[n]{\prod_{1}^{n} (Xi)}$$

$$\text{Harmonic} = \frac{N}{\sum_{1}^{n} (1/Xi)}$$

Alternative expressions of mean

- Arithmetic mean = $\sum x_i/N$
- Alternatives are anti transformed means of transformed numbers
- Geometric mean = $\exp(\sum \ln(x_i)/N)$
  - (anti log of average log)
- Harmonic Mean = reciprocal of average reciprocal
  - $1/(\sum (1/x_i)/N)$

Why all the fuss?

- Consider 1,2,4,8,16,32,64
- Median = 8
- Arithmetic mean = 18.1
- Geometric = 8
- Harmonic = 3.5
- Which of these best captures the "average" value?

Consider two sets, which is more?

subject	Set 1	Set 2
1	1	10
2	2	11
3	4	12
4	8	13
5	16	14
6	32	15
7	64	16
median	8	13
arithmetic	18.1	13.0
geometric	8.0	12.8
harmonic	3.5	12.7

Estimating the mean time of therapy

- A therapist has 20 patients, 19 of whom have been in therapy for 26-104 weeks (median, 52 weeks), 1 of whom has just had their first appointment. What is the average time patients are in therapy.
- Is this the average for this therapist the same as the average for the patients seeking therapy?

Estimating the mean time of therapy

- 19 with average of 52 weeks, 1 for 1 week
  - Therapists average is (19*52+1*1)/20 = 49.5 weeks
  - Median is 52
- But therapist sees 19 for 52 weeks and 52 for one week so the average length is
  - ((19*52)+(52*1))/(19+52) = 14.6 weeks
  - Median is 1

Estimating Class size

•5 faculty members teach 20 courses with the following distribution: What is the average class size?

Faculty member	A	B	C	D	average
1	10	20	10	10	12.5
2	10	20	10	10	12.5
3	10	20	10	10	12.5
4	100	20	20	10	37.5
5	400	100	100	100	175
department	106	36	30	28	50

Estimating class size

- What is the average class size?
- If each student takes 2 courses, what is the average class size from the students' point of view?
- Department point of view: average is 50 students/class

N	Size
10	10
5	20
4	100
1	400

Estimating Class size

Faculty member	A	B	C	D	average
1	10	20	10	10	12.5
2	10	20	10	10	12.5
3	10	20	10	10	12.5
4	100	20	20	10	37.5
5	400	100	100	100	175
department	106	36	30	28	50

Estimating Class size (student weighted)

Faculty member	A	B	C	D	average
1	10	20	10	10	14
2	10	20	10	10	14
3	10	20	10	10	14
4	100	20	20	10	73
5	400	100	100	100	271
Student	321	64	71	74	203

Estimating class size

Department perspective:

20 courses, 1000 students => average = 50

Student perspective: 1000 students enroll in classes with an average size of 203!

Faculty perspective: chair tells perspective faculty members that median faculty course size is 12.5, tells the dean that the average is 50 and that most upper division courses are small.

Measures of dispersion

- Range (maximum - minimum)
- Interquartile range (75% - 25%)
- Deviation score $x_i = X_i$-Mean
- Median absolute deviation from median
- Variance = $\sum x_i^2/(N-1)$ = mean square
- Standard deviation sqrt (variance) =sqrt($\sum x_i^2/(N-1)$)

Raw scores, deviation scores and Standard Scores

- Raw score for i_{th} individual X_i
- Deviation score $x_i=X_i$-Mean X
- Standard score = x_i/s_x
- Variance of standard scores = 1
- Mean of standard scores = 0
- Standard scores are unit free index

Variance of Composite

	X	Y
X	Variance X	Covariance XY
Y	Covariance XY	Variance Y

Variance (X+Y) = Var X+ Var Y + 2 Cov XY

Variance of Composite

	X	Y
X	$\sum x_i^2/(N-1)$	$\sum x_i y_i/(N-1)$
Y	$\sum x_i y_i/(N-1)$	$\sum y_i^2/(N-1)$

Var (X+Y) = $\sum(x+y)^2/(N-1) = \sum x_i^2/(N-1) + \sum y_i^2/(N-1) + 2\sum x_i y_i/(N-1)$

Variance of composite of n variables

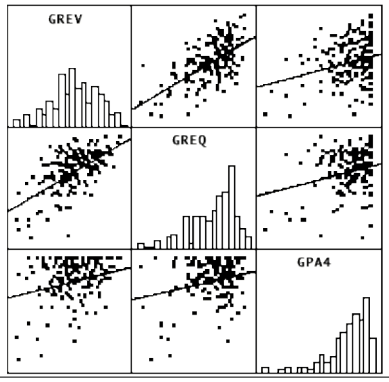
	X_1	X_2	…	X_i	X_j	…	X_n
X_1	Vx_1						
x_2	Cx_1x_2	Vx2					
…			…				
X_i	Cx_1x_i	Cx_2x_i		Vx_i			
X_j	Cx_1x_j	Cx_2x_j		Cx_ix_i	Vx_j		
…						…	
X_n	Cx_1x_n	Cx_2x_n		Cx_ix_n	Cx_jx_n		Vxn

Variance of composite of n items has n variances and n*(n-1) covariances

Measures of relationship

- Regression $y = bx + c$
  - $b_{y.x} = Cov_{xy}/Var_x$
- Correlation
  - $r_{xy} = Cov_{xy}/sqrt(V_x * V_y)$
  - Pearson Product moment correlation
    - Spearman (ppmc on ranks)
    - Point biserial (x is dichotomous, y continuous)
    - Phi (x, y both dichotomous)

SPLOM of GRE V, Q, GPA



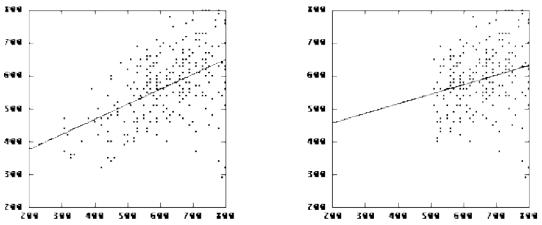
Correlation Matrix: GRE V, Q, GPA

PEARSON CORRELATION MATRIX

	GREV	GREQ	GPA4
GREV	1.00		
GREQ	0.61	1.00	
GPA4	0.27	0.25	1.00

NUMBER OF OBSERVATIONS: 163

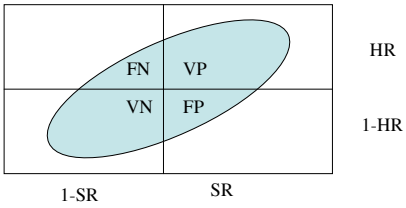
The effect of restriction of range



Phi coefficient of correlation

Hit Rate = Valid Positive + False Negative

Selection Ratio = Valid Positive + False Positive



Phi =(VP - HR*SR) /sqrt(HR*(1-HR)*(SR)*(1-SR)

Correlation size ≠ causal importance

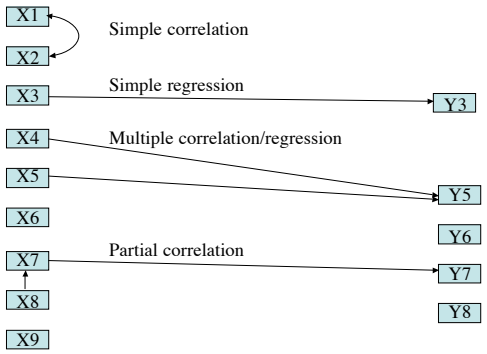
	Pregnant	Not Pregnant	Total
Intercourse	2	1,041	1,043
No intercourse	0	6,257	6,257
Total	2	7,298	7,300

Correlation size ≠ causal importance

	Pregnant	Not Pregnant	Total
Intercourse	.0003	.1426	.1429
No intercourse	.0000	.8571	.8571
Total	.0003	.9997	1.0000

Phi =(VP - HR*SR) /sqrt(HR*(1-HR)*(SR)*(1-SR)= .04

Variance, Covariance, and Correlation

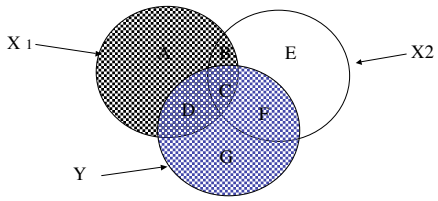


Measures of relationships with more than 2 variables

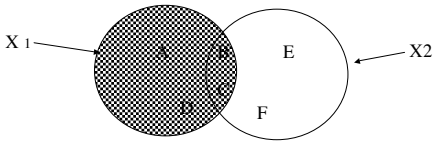
- Partial correlation
  - The relationship between x and y with z held constant (z removed)
- Multiple correlation
  - The relationship of x1 + x2 with y
  - Weight each variable by its independent contribution

Partial and Multiple Correlation

The conceptual problem



Variance, Covariance and Correlation



V1 = A + B + C + D C12 = B + C V2 = E + B + C + F

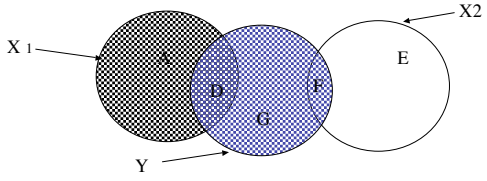
r = C12/sqrt(V1V2)

V1.2 = A + D V2.1 = E + F

V1.2 = V1(1-r^2) V2.1 = V2(1-r^2)

Multiple Correlation

Independent Predictors



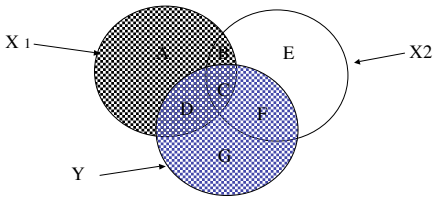
V1 = A + *B* + *C* + D C12 = *B* + *C* C1Y.2 = D

V2 = E + *B* + *C* + F C1Y = *C* + D C2Y.1 = F

VY = D + *C* + F + G C2Y = *C* + F C(12)Y3 = D + *C* + F

V1.2 = A + D V2.1 = E + F

Partial and Multiple Correlation



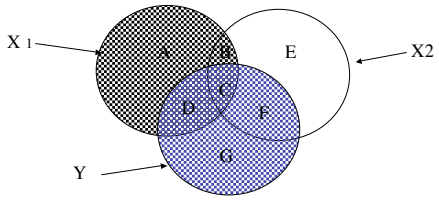
V1 = A + B + C + D C12 = B + C C1Y.2 = D

V2 = E + B + C + F C1Y = C + D C2Y.1 = F

VY = D + C + F + G C2Y = C + F C(12)Y3 = D + C + F

V1.2 = A + D V2.1 = E + F

Partial and Multiple Correlation: Partial Correlations



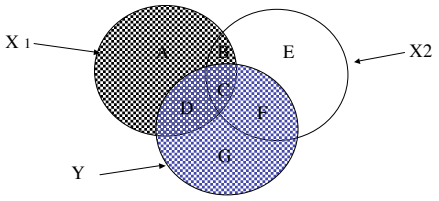
$V1 = A + B + C + D$ $C12 = B + C$ $C1Y.2 = D$ $C2Y.1 = F$

$V2 = E + B + C + F$ $C1Y = C + D$

$VY = D + C + F + G$ $C2Y = C + F$ $r_{1Y.2} = \frac{(r_{1y} - r_{12} * r_{2Y})}{sqrt((1-r_{12}^2)*(1- r_{y2}^2))}$

$V1.2 = A + D$ $V2.1 = E + F$

Partial and Multiple Correlation: Multiple Correlation-correlated predictors



$Y = b_1X_1 + b_2X_2$ $b_1 = (r_{x1y} - r_{12} * r_{2y})/(1-r_{12}^2)$

$b_2 = (r_{x2y} - r_{12} * r_{1y})/(1-r_{12}^2)$

$R^2 = b_1r_1 + b_2r_2$

Multiple Correlation as an unweighted composite

	X_1	X_2	Y
X_1	Vx_1	Cx_1x_2	Cx_1y
X_2	Cx_1x_2	Vx_2	Cx_2y
Y	Cx_1y	Cx_2y	Vy

$Vx_1x_2 = Vx_1 + Vx_2 + 2Cx_1x_2$ $R(x_1x_2)y = \frac{C(x_1x_2)y}{Sqrt(Vx_1x_2)*V_y}$

$C(x_1x_2)y = Cx_1y + Cx_2y$

Multiple Correlation as a weighted composite

	b_1X_1	b_2X_2	Y
b_1X_1	$b_1^2Vx_1$	$b_1b_2Cx_1x_2$	b_1Cx_1y
b_2X_2	$b_1b_2Cx_1x_2$	$b_2^2Vx_2$	b_2Cx_2y
Y	b_1Cx_1y	b_2Cx_2y	Vy

$Vb_1x_1b_2x_2 = b_1^2Vx_1 + b_2^2Vx_2 + 2C\ b_1b_2Cx_1x_2$ $R(b_1x_1b_2x_2)y = \frac{C(b_1x_1b_2x_2)}{Sqrt(Vb_1x_1b_2x_2)*V_y}$

$C(b_1x_1b_2x_2)y = b_1Cx_1y + b_2Cx_2y$

Multiple Correlation as a weighted composite

	b_1X_1	b_2X_2	Y
b_1X_1	$b_1^2Vx_1$	$b_1b_2Cx_1x_2$	b_1Cx_1y
b_2X_2	$b_1b_2Cx_1x_2$	$b_2^2Vx_2$	b_2Cx_2y
Y	b_1Cx_1y	b_2Cx_2y	Vy

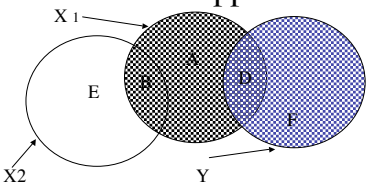
$R(b_1x_1b_2x_2)y = \frac{C(b_1x_1b_2x_2)}{Sqrt(Vb_1x_1b_2x_2)*V_y}$

$b_1 = (r_{x1y} - r_{12} * r_{2y})/(1-r_{12}^2)$

Problem: Find b1, b2 to maximize R

$b_2 = (r_{x2y} - r_{12} * r_{1y})/(1-r_{12}^2)$

Multiple Correlation: Suppression



$V1 = A + B + C + D$ $C12 = B + C$ $C1Y.2 = D$ $C2Y.1 = F$

$V2 = E + B + C + F$ $C1Y = C + D$

$VY = D + C + F + G$ $C2Y = C + F$ $C(12)Y3 = D + C + F$

$V1.2 = A + D$ $V2.1 = E + F$

Problems with correlations

- Simpson's paradox and the problem of aggregating groups
  - Within group relationships are not the same as between group or pooled relationships
- Phi coefficients and the problem of unequal marginal distributions
- Alternative interpretations of partial correlations

Sex discrimination?

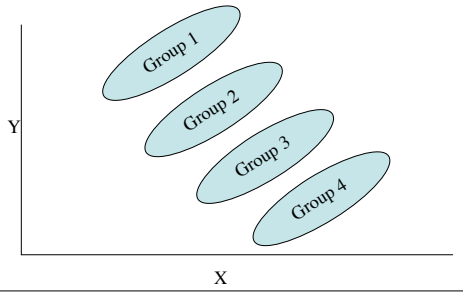
	Admit	Reject	Total
Male	40	10	50
Female	10	40	50
Total	50	50	100

Phi =(VP - HR*SR) /sqrt(HR*(1-HR)*(SR)*(1-SR)= -.80

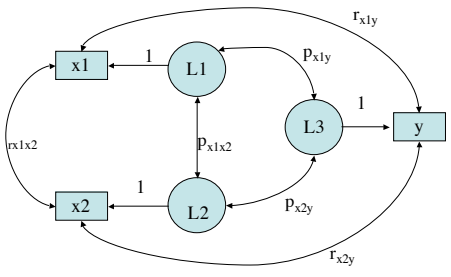
Sex discrimination?

	Department 1			Department 2		
	Admit	Reject	Total	Admit	Reject	Total
Male	40	5	45	0	5	5
Female	5	0	5	5	40	45
Total	45	5	50	5	45	50
Phi	.11			.11		
Pooled phi			-.8			

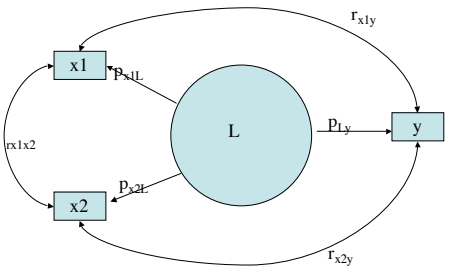
Within group vs Between Group correlation



Partial correlation: conventional model



Partial correlation: Alternative model



Partial Correlation: classical model

	X_1	X_2	Y
X_1	1.00		
X_2	.72	1.00	
Y	.63	.56	1.00

Partial r = $(r_{x1y}-r_{x1x2}*r_{x2y})/\text{sqrt}((1-r_{x1x2}^2)*(1-r_{x2y}^2))$

Rx1y.x2 = .33 (traditional model) but = 0 with structural model