## Literature review

### Background

Given a brief to summarise existing methods for estimating power to detect a difference in correlations between identical (monozygotic) and non-identical (dizygotic) twins, with a view to comparing and perhaps improving upon this, this review sets out to develop the background knowledge to understand the research question and the assumptions underpinning it.

To insert here: a brief background of genetics context giving rise to historical interest in correlation;

### Brief history of correlation coefficients and their usage

Much has been written on the history of correlation: Florence Nightingale David provided a concise summary of history of correlation up to 1938, including formulas and notes on various approximations and methods for detecting difference, and a proof of the distribution of for any n and ; employed to produce statistical tables and a graphical visualisation of ‘chance of rejecting the hypothesis when true’ for different values of r and (1). More recently, other reviews (2, 3) have described the different approaches available to estimate correlation in different contexts. The following paragraphs provide a brief review with key formulae, setting the context for the usage of the term correlation in the current project.

#### Product moment correlation coefficient (r)

The (Pearson) product-moment correlation coefficient is used to describe strength and direction of linear relationship between two normally distributed random variables. The population product-moment correlation is estimated by as the covariance of x and y, divided by the sum of their standard deviations; however, this formula can be arranged in a number of ways (see appendix). The correlation coefficient is interpreted as the average observed change in one variable given a unit increase in the other, when both variables have been standardised with means of (0,0) and standard deviation of (1,1); the resulting estimand ranges from (perfect negative linear relationship), through 0 (no apparent association), to (perfect positive linear relationship). The formulation below aims to transparently display the calculation of , using matrix multiplication in the final equation

As noted above, a bivariate normal distribution is assumed. There are non-parametric approaches to estimating correlation, which draw upon the rank order transformation of the source data to estimate the degree of bivariate monotonicity.

#### Spearman correlation coefficient ()

Spearman’s correlation coefficient is the equivalent of Pearson’s, but performed using the variables’ rank order (4);

#### Kendall’s tau ()

Kendall’s (tau) is another non-parametric approach, providing a summary measure of correlation based on concordance of trend across the sample (4). Pairs are concordant if the product of consecutive rank pair differences is , and discordant if this product is . The number of concordant () and discordant () pairs are tallied, and the difference () is the score . Kendall's is calculated as , where is the total number of pairs calculated as . Kendall's includes further adjustment to account for ties in pairs of x and y variables on the rank data to consider neighbouring pair concordance with regard to sign, and its three variations approach the question of tied rank in different ways (see appendix).

#### Intra-class correlation coefficient (ICC)

In the context of paired data (such as twin sibling pairs), the calculation of an intra-class correlation coefficient (ICC) is likely preferable (5-8). These may be estimated using linear mixed-effects modelling allowing for random effect at the within pair cluster level. There are different kinds of ICCs which may be calculated, depending on the modelling purpose: these may reflect consistency or absolute agreement. In the classic twin study context, the ICC used measures consistency of similarity across twin pairs and resolves to the ratio of additive variance to total variance (additive plus residual variances):

This formula expresses the proportion of variance attributable to between pair differences, and as such unlike the previously detailed correlation coefficient formulations (Pearson, Spearman, and Kendall’s) which have range (), the ICC has range from 0 to 1.

This consideration of genetic variance partitioned to sources (e.g. additive, common and environmental) is known as variance component modelling. This is what is undertaken through the classical twin study design, in which the ICC is understood as an estimate of heritability ; the smaller the residual (common) variance, the weaker the role of environment and the greater the expected role of the genetic additive contribution. Further, heritability is also theoretically estimated as twice the difference in the within pair ICCs for identical (monozygotic, or MZ) and non-identical (fraternal, dizygotic or DZ) twins (6, 7, 9).

Considered through this lens, it is assumed that the common environmental effects shared by each kind of twin pair will be similar, such that between group differences may be attributable to genetics.

Calculation of the ICC in the context of a mixed effect model by way of the ratio of between-pairs to total variance allows for considerably increased nuance: if the estimated model adjusts for covariates (say sex, or body mass index) as fixed effects, then the ICC estimates for model cluster variables for which random effects are estimated will be conditional on these adjustments. This potentially multivariate context further distinguishes the ICC from its bivariate peers and .

*Other notes to incorporate*:

* Probably more scope to detail foundations of ICC with Fisher and ANOVA
* Bias corrected estimator for the ICC in the context of a one way analysis of variance (ANOVA) model, evaluated using a simulation study (10).
* Chen et al (11) compare the estimation of ICCs using various mixed effects modelling approaches rather than ANOVA

### Differences in correlations

Much work has focused on inference using the product moment correlation (12-17). In the early twentieth century a useful geometry inspired transformation for approximating a normal distribution was proposed using the inverse hyperbolic tangent of the vector , known as Fisher’s (1, 5, 18). Algebraically,

scales from a range of to with a range of . Fisher’s is approximately normally distributed with mean and standard deviation , which may be approximated as

such that an approximate confidence interval for with significance level of is

where (Phi) is the inverse normal distribution indexed at (1). A slightly more refined approximation for is ; both of the formulas provided are approximations of a method of moments based derivation of Fisher's transformation (1).

David (1) notes that an hypothesis of no difference between sample correlation coefficients for two groups (which for relevance sake we will denote as and ), may be tested by evaluating cross-over of z-transformation derived confidence intervals (L for lower and U for upper)

or , equivalently,

This useful principle of logic may be applied to other empirical methods of confidence interval calculation, such as using bootstrap estimates which will be detailed later.

#### Power

* Neyman and ES Pearson (19)
* Cohen (20)
* Given the potential increased complexity of estimation of ICCs by way of mixed effects models, any calculation will be much more complex than the simple examples using the product moment correlation coefficient! Perhaps this is a good argument for Enes’ suggestion to employ simulation approaches to derive empirical estimates of the confidence intervals and power this way…

#### Empirical methods of distribution approximation

* Bootstrap (Efron and Tibshirani), simulation and permutation - (Enes’ references and like articles)

## References

1. David FN. Tables of the ordinates and probability integral of the distribution of the correlation in small samples. London: Biometrika; 1938.

2. Hauke J, Kossowski T. Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data. Quaestiones Geographicae2011. p. 87.

3. Liu J, Tang W, Chen G, Lu Y, Feng C, Tu XM. Correlation and agreement: overview and clarification of competing concepts and measures. Shanghai Archives of Psychiatry. 2016;28(2):115-20.

4. StataCorp. Stata: Release 13. College Station, TX: StataCorp LP; 2013.

5. Fisher RA, Sir, -. Statistical methods, experimental design and scientific inference / R.A. Fisher ; edited by J.H. Bennett ; with a foreword by F. Yates. Bennett JH, editor. Oxford [England] ; New York :: Oxford University Press; 1990.

6. Castillo-Fernandez JE, Spector TD, Bell JT. Epigenetics of discordant monozygotic twins: implications for disease. Genome Medicine. 2014;6(7):60.

7. Falconer DS. Introduction to quantitative genetics / D.S. Falconer. Edinburgh: Oliver & Boyd; 1960.

8. Rabe-Hesketh S, Skrondal A. Multilevel and longitudinal modeling using Stata. 3rd ed. College Station, Tex.: Stata Press Publication; 2012.

9. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era — concepts and misconceptions. Nature Reviews Genetics. 2008;9:255.

10. Atenafu EG, Hamid JS, To T, Willan AR, M Feldman B, Beyene J. Bias-corrected estimator for intraclass correlation coefficient in the balanced one-way random effects model. BMC Medical Research Methodology. 2012;12:126-.

11. Chen G, Taylor PA, Haller SP, Kircanski K, Stoddard J, Pine DS, et al. Intraclass correlation: Improved modeling approaches and applications for neuroimaging. Human Brain Mapping. 2018;39(3):1187-206.

12. Galton F. Kinship and Correlation. The North American Review. 1890;150(401):419-31.

13. Galton F. Co-Relations and Their Measurement, Chiefly from Anthropometric Data. Proceedings of the Royal Society of London. 1888;45:135-45.

14. Francis Galton. 1822–1922. A Centenary Appreciation. By Pearson Karl , F.R.S. With frontispiece drawing of Francis Galton. [Pp. 23. London: Cambridge University Press. Price 2s. net.]. Journal of the Institute of Actuaries. 2016;53(3):311-2.

15. Pearson K. Note on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London. 1895;58(347-352):240-2.

16. "Student". Probable Error of a Correlation Coefficient. Biometrika. 1908;6(2-3):302-10.

17. Soper HE. On the Probable Error of the Correlation Coefficient to a Second Approximation\*. Biometrika. 1913;9(1-2):91-115.

18. Fisher RA. Frequency Distribution Of The Values Of The Correlation Coeffients In Samples From An Indefinitely Large Popultion. Biometrika. 1915;10(4):507-21.

19. Neyman J, Pearson ES. Contributions to the theory of testing statistical hypotheses. Statistical Research Memoirs. 1936;1:1-37.

20. Cohen J. Statistical Power Analysis for the Behavioral Sciences. New York: Laurence Erlbaum Associates; 1988.

## Appendix

Rough sketches – in Stata using mata for now, as that is where I am probably most fluent. Note, I realised too late on the weekend that I couldn’t open the Stata data file provided by Katrina last week as it is newer than my version of Stata (13). I made do with other datasets. I’ll convert the supplied twins data tomorrow at work and run through the tutorial then ---- I realise that this covers a lot of the aspects of mixed modelling and the ICC that I’m still getting to grips with.

Part of the point below is to figure out how to replicate commands from scratch – that’s how I remember how things are calculated / what they represent, and also if we are creating a tool later on it will be useful to be able to as far as possible not rely on existing commands.

\* Project: BCA\_rp2

\* Purpose: Notes on correlation coefficients and their calculation

\* Date: 20180319

\* Author: Carl Higgs

/\*

Suppose

Correlation is one method used to describe the relationship between two variables X and Y.

There are different approaches (kinds of correlation coefficient).

Pearson product-moment correlation coefficient may be the most common: supposing X and Y were standardised to share a common scale,

the product moment correlation coefficient describes the direction and proportional change in Y for a given a unit increase in X.

Expectation: E(Y) = \mu

Variance: Var(Y) = E(Y^2)-\mu^2

Standard deviation: sigma(Y) = sqrt(Var(Y))

\*/

// for convenience, using car data

sysuse auto.dta, clear

gen id = \_n

gen group = length > 200

mata

data = st\_data(.,"id price mpg")

n = rows(data)

id = data[.,1]

y = data[.,2]

x = data[.,3]

ydiff = y:-mean(y)

xdiff = x:-mean(x)

sd\_y = sqrt(1/(n-1) \* sum((ydiff):^2))

sd\_x = sqrt(1/(n-1) \* sum((xdiff):^2))

r = (sum(x:\*y)- n\*mean(x)\*mean(y))/((sd\_x\*sd\_y)\*(n-1))

r

// or equiv:

r

r = 1/(n-1)\*((sum(x:\*y)- n\*mean(x)\*mean(y))/((sd\_x\*sd\_y)))

r

or equiv:

r = sum(xdiff:\*ydiff) / sqrt(sum(xdiff:^2)\* sum(ydiff:^2))

r

// or perhaps -following Fisher - apparently equal

r = 1/(sd\_x\*sd\_y) \* sum(xdiff:\*ydiff)/(n-1)

r

r = 1/(n-1)\*((sum(xdiff:\*ydiff))/(sd\_x\*sd\_y))

r

/\* -.4685966882 \*/

end

\* crosscheck

corr price mpg

/\*

(obs=74)

| price mpg

-------------+------------------

price | 1.0000

mpg | -0.4686 1.0000

\*/

\* following statistical thinking in epidemiology (yu), we can also approach this as

foreach var in price mpg {

su `var'

capture gen `var'\_std = (`var'-r(mean))/r(sd)

}

mata

test = st\_data(.,"price\_std mpg\_std")

ty = test[.,1]

tx = test[.,2]

(ty'tx)/(n-1)

/\* -.4685966854 Perhaps the disjunct is due to rounding in Stata? \*/

// and applying this as per our formulas above

r = ((ydiff:/sd\_y)'(xdiff:/sd\_x))/(n-1)

/\* -.4685966882 equal to other mata results \*/

end

\* note this differs in the 9th decimal place

// confidence interval

// following Nick Cox http://www.stata-journal.com/sjpdf.html?articlenum=pr0041

// but also David 1938, who notes that the Fisher-transformed value z' ~N(mean \zeta, sd 1/\sqrt(n-3)), such that ci for z' is zeta

\pm \psi\_0 / sqrt(n-3). See p.xxx and moreso xxvi, where in dagger footnote an approximation for stddev is derived as \sigma\_z'^2 =

(3n^2+8)/(x(n-1)^2) . The formula this approximates itself is derived using the method of moments on Fisher's transformation.

Fisher references the sd of z as 1/\sqrt(n-3) in section 35 on transformed coefficients in the chapter on correlation coefficients

in Statistical Methods for Research Workers (14th ed. 1990, p200).

mata

alpha = 0.05

df = n-3

r\_ci = r,tanh(atanh(r) :+ invnormal(1-alpha/2) \* (-1,1)/sqrt(df))

-1-

C:\Users\Carl\OneDrive\Research\2 - BCA\Research project\bca\_rp2\scripts\stata\_notes.do.txt Tuesday, 20 March 2018 02:18

r\_ci

end

ci2 price mpg, corr

/\* Incidentally, David's formula for probability distribution of the correlation coefficient r for any n and rho

p(r) = (1-\rho^2)^((n-1)/2) /(\pi \times (n-3)) \times (1-r^2)^((n-4)/2)\times d^(n-2)/d(r\rho)^(n-2)) \times (arccos(-\rho r) /

\sqrt(1-\rho^2 r^2))

\*/

/\*

For a sample drawn from normal distribution, using summation notation:

\mu is estimated by x\bar, and \sigma is estimated by s; r is the estimator for the population correlation coefficient \rho.

x\bar = 1/n \Sum\_(i=1)^n x

s = \sqrt(((x-x\bar)^2)/(n-1))

r = (1/(n-1))\Sum\_(i=1)^n(((x\_i-x\bar)/s\_x)((y\_i-y\bar)/s\_x))

= (1/(n-1))\Sum\_(i=1)^n((x\_i-x\bar)(y\_i-y\bar)/(s\_x s\_y))

= (XtX YtY)/XtY

Note comment on mix distribution here:

https://math.stackexchange.com/questions/14630/generating-random-values-from-non-normal-and-correlated-distributions

https://stackoverflow.com/questions/4454513/combining-two-normal-random-variables/4454941#4454941

Is it product-moment correlation coefficient due to the 2\*2 determinant calc.?

ie. for 2x2 matrix inverse is

det = ad-bc

(2x2 matrix)^-1 = 1/det[flipped and negative 2x2 matrix with positive diagonal]

hmmm - tatachoric correlation

https://books.google.com.au/books?id=gbrFCgAAQBAJ&pg=PA31&lpg=PA31&dq=ad-bc+product+moment+odds+ratio&source=bl&ots=D6myJbq2yY&sig=tIvP

tIQrSkE1WrpDPc461CMwes0&hl=en&sa=X&ved=0ahUKEwj5ssfUkPfZAhVFi7wKHVt6A50Q6AEIKTAA#v=onepage&q=ad-bc%20product%20moment%20odds%20ratio&f=

false

The Pearson product-moment correlation coefficient may be considered a measure of interclass correlation, in that it is a global

measure. An alternate approach which may have relevance in the case of twin studies is intra-class correlation, which reflects the

average correlation in context of paired measurements. In the case of twin sibling pairs for example, we may be interested in

whether values within pairs tend to be more similar than those between pairs.

see David Duffy https://genepi.qimr.edu.au/staff/davidD/asthma11.html for discussion in twin context, including note citing Falconer

re heritability that

H=2(ICC\_MZ-ICC\_DZ) =(VA+1.5\*VD)/VP

That is heritability - which may be is the population variance standardised sum of additive plus 1.5 times dominance variances.

also provides a heuristic icc interpretation table "heuristics for intraclass correlations for a single variable (Table)":

Table 2. Genetic hypothesis testing for a single continuous trait in the classical twin design. The MZ and DZ intraclass correlations

are rMZ and rDZ respectively.

Relationship Interpretation

rMZ > 4rDZ Epistasis

rMZ > 2rDZ Genetic dominance (or epistasis; shared environment small)

2rDZ > rMZ > rDZ Additive genes and shared environment (genetic dominance small)

rMZ = 2rDZ Additive genetic effect - either monogenic or polygenic

rMZ = rDZ > 0 No genetic contribution - effects of family environment

rMZ = rDZ = 0 No familial aggregation

\*/

// in stata / mata, using adoption data; see stata manual

use http://www.stata-press.com/data/r13/adoption, clear

describe

icc iq3 family mc, mixed

/\*

Two-way mixed-effects model

Consistency of agreement

Random effects: family Number of targets = 10

Fixed effects: mc Number of raters = 2

--------------------------------------------------------------

iq3 | ICC [95% Conf. Interval]

-----------------------+--------------------------------------

Individual | .7142152 .1967504 .920474

Average | .8332853 .3288078 .9585904

--------------------------------------------------------------

F test that

ICC=0.00: F(9.0, 9.0) = 6.00 Prob > F = 0.007

\*/

/\* different types of ICCs: consistency of agreement (CA-ICC) and absolute agreement (AA-ICC); "We want to compare individual CA-ICC

with individual AA-ICC for each of the three IQ variables." \*/

mixed iq3 || family: , reml var

estat icc

/\*

Intraclass correlation

------------------------------------------------------------------------------

Level | ICC Std. Err. [95% Conf. Interval]

-----------------------------+------------------------------------------------

-2-

C:\Users\Carl\OneDrive\Research\2 - BCA\Research project\bca\_rp2\scripts\stata\_notes.do.txt Tuesday, 20 March 2018 02:18

family | .7216082 .1557153 .3619825 .9221322

------------------------------------------------------------------------------

\*/

// note that the icc presented here is equal to

// var(\_cons) / (var(\_cons) + var(Residual))

// However, it is also not the value we calculated earlier. Perhpas because we include mc as a fixed effect??

mixed iq3 mc || family: , reml var

estat icc

/\* This is our magic number!! But what if we don't want mc as fixed effect??

Residual intraclass correlation

------------------------------------------------------------------------------

Level | ICC Std. Err. [95% Conf. Interval]

-----------------------------+------------------------------------------------

family | .7142153 .1632988 .342513 .923013

------------------------------------------------------------------------------

\*/

mixed iq3 || family: || family: mc, reml

/\* close

Conditional intraclass correlation

------------------------------------------------------------------------------

Level | ICC Std. Err. [95% Conf. Interval]

-----------------------------+------------------------------------------------

family | .7215698 .1557673 .3618508 .9221456

------------------------------------------------------------------------------

Note: ICC is conditional on zero values of random-effects covariates.

\*/

// Some explanation here: https://www.stata.com/statalist/archive/2012-05/msg01132.html

// trying my own thing (which i establish above, is equiv to

// mixed iq3 mc || family: , reml var

// estat icc

// Which is not quite what we want - we don't want mc -- trial again below

use http://www.stata-press.com/data/r13/adoption, clear

reshape wide iq3 iq9 iq15, i(family) j(mc)

mata

data\_f = st\_data(.,"family iq31 iq32")

mother = data\_f[.,2]

child = data\_f[.,3]

// following fisher

xbar = 1/(2\*rows(data\_f)) \* sum(mother:+child)

mdiff = mother:-mean(mother)

cdiff = child:-mean(child)

var = 1/(2\*rows(data\_f)) \* (sum(mdiff:^2)+sum(cdiff:^2))

sd = sqrt(var)

// the below corresponds to consistency of agreement

r\_icc\_ca = 1/(rows(data\_f)\*var) \* sum(mdiff:\*cdiff)

end

// ALTERNATE APPROACH USING MEAN SQUARES ---

// NOT WORKING, but may be required for CIs

// formulated for wide data with paired data (cluster size 2)

mata

data\_w = mother,child

k = cols(data\_w) // cluster size

n = rows(data\_w) // observations

// preliminary holders for estimates - summed in iteration

WMS = 0 // within pair mean squares

BMS = 0 // between pair mean squares

mean\_total = mean(mean(data\_w)')

rater\_means = colsum(data\_w):/rows(data\_w)

// between rater mean squares (joint??)

JMS = rowsum(((rater\_means:-mean\_total):^2):/(k-1))

// iterate to build estimatess

for(i=1; i<=n;i++){

ybar\_i = rowsum(data\_w[i,.])/k

BMS = BMS + (ybar\_i-mean\_total)^2 /(n-1)

for(j=1; j<=k;j++){

WMS = WMS + (data\_w[i,j]-ybar\_i)^2 / (n\*(k-1))

}

}

// Residual mean square error

RMSE = sum(((data\_w:-mean\_total):^2):-((k-1)\*JMS-(n-1)\*BMS))/((n-1)\*(k-1))

// ALTERNATE APPROACH - revisit first, but attempt to note adjust for mc

mata

data\_f = st\_data(.,"family iq31 iq32")

mother = data\_f[.,2]

child = data\_f[.,3]

// following fisher

xbar = 1/(2\*rows(data\_f)) \* sum(mother:+child)

mdiff = mother:-mean(mother)

cdiff = child:-mean(child)

-3-

C:\Users\Carl\OneDrive\Research\2 - BCA\Research project\bca\_rp2\scripts\stata\_notes.do.txt Tuesday, 20 March 2018 02:18

var = 1/(2\*rows(data\_f)) \* (sum(mdiff:^2)+sum(cdiff:^2))

sd = sqrt(var)

// the below corresponds to consistency of agreement

r\_icc\_ca = 1/(rows(data\_f)\*var) \* sum(mdiff:\*cdiff)

end

end

// Spearman correlation coefficient provides a non-parametric approach to considering correlation: instead of making distributional

assumptions, it considers the degree to which the relationship two variables can be considered to display a monotonic trend.

// (according to wikipedia): defined as the pearson correlation coefficient of the two variables using their rank order

// In stata / mata:

sysuse auto.dta

gen id = \_n

// append rank order of the respective variables

foreach var in price mpg {

egen `var'\_rank = rank(`var')

}

mata

data\_r = st\_data(.,"id price\_rank mpg\_rank")

y\_r = data\_r[.,2]

x\_r = data\_r[.,3]

yr\_diff = y\_r:-mean(y\_r)

xr\_diff = x\_r:-mean(x\_r)

sd\_yr = sqrt(1/(n-1) \* sum((yr\_diff):^2))

sd\_xr = sqrt(1/(n-1) \* sum((xr\_diff):^2))

r\_s = 1/(n-1)\*((sum(xr\_diff:\*yr\_diff))/(sd\_xr\*sd\_yr))

end

spearman price mpg

// Kendall's tau is another non-parametric approach calculated and is a summary measure of correlation based on concordancy of trend

across the sample. Pairs are concordant if the product of consecutive rank pair differences is > 0, and discordant if this product is

< 0. The number of concordant (C) and discordant (D) pairs are tallied, and the difference C - D is the score S. Kendall's Tau A

is calculated as S / N, where N is the total number of pairs calculated as n(n-1)/2. Kendall's b includes further adjustment to

account for ties in pairs of x and y variables.

// note that the kendall's tau b formula may not be technically correct in calculation of ties, but it gives the right answer - I

simplified so it works!

mata

N = n\*(n-1)/2

C = J(N,1,.)

D = J(N,1,.)

t\_x = J(N,1,.)

t\_y = J(N,1,.)

k = 1

for(i=2; i<=rows(data\_r);i++){

for(j=1; j<=i-1;j++){

// calculate rank differences

x\_ij\_diff = (x\_r[i]-x\_r[j])

y\_ij\_diff = (y\_r[i]-y\_r[j])

// calculate ties

t\_x[k] = x\_ij\_diff == 0

t\_y[k] = y\_ij\_diff == 0

// calculate concordancies

concordancy = x\_ij\_diff\*y\_ij\_diff

C[k] = concordancy > 0

D[k] = concordancy < 0

// increment pair index

k = k+1

}

}

// calculate Score

S = sum(C) - sum(D)

// Kendall's tau a

k\_tau\_a = S/N

k\_tau\_a

// Kendall's tau b

N\_tx = sum(t\_x)

N\_ty = sum(t\_y)

k\_tau\_b = S/sqrt((N-N\_tx)\*(N-N\_ty))

k\_tau\_b

end

-4-