Preliminary literature review (just notes at this stage)

The study of heredity arguably gave rise to modern methods for correlation

Define heritability –

- Falconer (1) p163 – Identical Twins ; maybe also see Visscher (2)
    - Explains the relevance of the comparison between identical twins and non-identical twins: 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.
    - Assuming that
        - non-additive genetic variance is not a factor at play, and
        - that some other confounding has not occurred resulting in dizygotic twins being more similar than would otherwise be expected (example given is blood sharing in fraternal cattle twins in birthing complications --- not sure of human analogy??)
        - then,

          'The difference between the correlation coefficients of identicals and fraternals … could be taken as an estimate of half the heritability'

$$\frac{h^2}{2} = r_{MZ} - r_{DZ}, \qquad \text{so}$$

$$h^2 = 2(r_{MZ} - r_{DZ})$$

History of estimation of correlation in modern statistics and genetic analysis

- Gregor Mendel (sweet pea experiments > segregation > heritability)
- Francis Galton's articulation of Mendelian genetics in formal statistical treatment (3, 4) >> also sweet peas and humans (height, bones) >> regression / correlation
- Karl Pearson, extension of Galton's work on correlation ("Galton's function" >> bivariate correlation), and drive for numerical accuracy (down to FN David in 1938) (5, 6)
- Student, and 'probable error of a correlation coefficient' (7)
- Soper, and approximating the frequency distribution of ρ with r (8)
- Ronald Fisher, and approximation of distribution of population correlation coefficient using inverse hyperbolic tangent transformation (Fisher's z) (9)
- Neyman and Egan Pearson (10) (confidence intervals; and NHST --- relates to questions of power)
- Florence Nightingale David (11) – 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 r for any n and ρ; and graphical visualisation of 'chance of rejecting the hypothesis when true' for different values of r and ρ. Might be inspiration for a contemporary visualisation of our approach.

The above history largely traces back over the references given by David (Student to Neyman & Pearson); however, a useful summary table of historical landmarks is provided in Hauke and Kossowski (12) , as preface to their comparison of Pearson and Spearman (non-parametric) correlation coefficients – which may be relevant later on once I start differentiating these. Darwin connection not a surprise given Galton relationship; but Gauss & JS Mill --- so extending beyond the sphere of genetics there

Table 1. Landmarks In the History of Correlation and Regression.

Date	Person	Event
1823	Carl Friedrich Gauss, German mathematician	Developed the normal surface of N correlated variates.
1843	John Stuart Mill, British philosopher	Proposed four canons of induction, including concomitant variation.
1846	Augusts Bravais. French naval officer and astronomer	Referred to "une correlation", worked on bivariate normal distribution.
1868	Charles Darwin, Galton's cousin, British natural philosopher	"All parts of the organisation are ... connected or correlated."
1877	Sir Francis Gallon, British, the first biometrician	First discussed "reversion", the predecessor of regression.
1885	Sir Francis Gallon	First referred to "regression". Published bivariate scatterplot with normal isodensity lines, the first graph of correlation. "Completed the theory of bi-variate normal correlation." (Pearson 1920)
1888	Sir Francis Gallon	Defined *r* conceptually, specified its upper bound.
1895	Karl Pearson, British statistician	Defined the (Galton-) Pearson product-moment correlation coefficient.
1920	Kart Pearson	Wrote "Notes on the History of Correlation".
1985		Centennial of regression and correlation.

Figure 1. Snapshot from Hauke article - just as a note of the earlier history, which I probably won't trace back myself!

Random effects & Interclass correlation coefficient (ICC) – correlations in context of clustering (like twin pair)

- Fisher and Analysis of Variance
    - From Wikipedia (not a reputable reference I know!; preliminary): "in the ICC, the data are centered and scaled using a pooled mean and standard deviation"
- I think ICC will be apt approach – e.g. used in Castillo-Fernandez et al (13) considering epigenetics of discordant Mz (and Dz) twins, with reference to 'intra-pair correlation coefficients' (ICC) in place of r in the formula for heritability sourced from Falconer (1)

Non parametric approaches to correlation

- Spearman (14)
    - See review and real data comparison with Pearsons r by Hauke and Kossowski (12)
- Kendall's tau A and B

Power

- Neyman and ES Pearson (10)
- Cohen (15)

Methods of estimation

- Approximations
    - Analytic / numerical
        - Fisher's z
        - others
    - Empirical distribution
        - Bootstrap
        - Simulation

- Permutation

Variance component models for Mz Dz twins (perhaps first?) --- see Falconer (1)

Ratio of Mz/Dz and variance as positive constrained; Mixture of chi(0) and chi(1) distributions - see Visscher 2008 (16)

References

1. Falconer DS. Introduction to quantitative genetics / D.S. Falconer. Edinburgh: Oliver & Boyd; 1960.
2. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era — concepts and misconceptions. Nature Reviews Genetics. 2008;9:255.
3. Galton F. Kinship and Correlation. The North American Review. 1890;150(401):419-31.
4. Galton F. Co-Relations and Their Measurement, Chiefly from Anthropometric Data. Proceedings of the Royal Society of London. 1888;45:135-45.
5. 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.
6. 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.
7. "Student". Probable Error of a Correlation Coefficient. Biometrika. 1908;6(2-3):302-10.
8. Soper HE. On the Probable Error of the Correlation Coefficient to a Second Approximation*. Biometrika. 1913;9(1-2):91-115.
9. Fisher RA. Frequency Distribution Of The Values Of The Correlation Coeffients In Samples From An Indefinitely Large Popultion. Biometrika. 1915;10(4):507-21.
10. Neyman J, Pearson ES. Contributions to the theory of testing statistical hypotheses. Statistical Research Memoirs. 1936;1:1-37.
11. David FN. Tables of the ordinates and probability integral of the distribution of the correlation in small samples. London: Biometrika; 1938.
12. 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.
13. Castillo-Fernandez JE, Spector TD, Bell JT. Epigenetics of discordant monozygotic twins: implications for disease. Genome Medicine. 2014;6(7):60.
14. Spearman C. The Proof and Measurement of Association between Two Things. The American Journal of Psychology. 1904;15:72-101.

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

16. Visscher PM, Gordon S, Neale MC. Power of the classical twin design revisited: II detection of common environmental variance. Twin research and human genetics : the official journal of the International Society for Twin Studies. 2008;11(1):48-54.