Carl Higgs
School of Population and Global Health
The University of Melbourne

Power to detect a difference in correlations in identical and non-identical twins

Simulation study and power calculator utility

Research project in partial fulfilment of the degree
Master of Biostatistics

Supervisors:
Dr Enes Makalic
Dr Katrina Scurrah

with additional advice and support from Dr Elasma Milanzi

June 2018

I gratefully acknowledge the advice and support from my supervisors Elasma, Enes and Katrina through the semester. In addition, the contribution of advice from my colleague Koen Simons on techniques for improving the efficiency of my code was much appreciated. Thank you Elasma, Enes, Katrina and Koen.

Supervisor statement

Mr Carl Higgs completed a Master of Biostatistics research project under the supervision of Dr Katrina Scurrah, Dr Elasma Milanzi and Dr Enes Makalic. The research project examined statistical approaches for testing the equality of two correlation coefficients in the context of twins studies. The project was largely empirical in nature and required strong computer science and statistics skills.

From the outset, we decided to schedule weekly one-hour supervisory meetings with Carl and these were held every Thursday. Carl regularly attended each meeting and was happy to tackle any challenge with a positive attitude. Throughout the course of the project, Carl was an enthusiastic learner, demonstrated hard work (e.g., he completed comprehensive power and sample size simulation studies) and produced quality results including a very flexible web-based software tool for power analysis. Carl showed a high level of initiative, demonstrated the ability to work independently and was excellent at communicating with the supervisors during all stages of the project. Carl's enthusiasm for the technical side of the project occasionally led to Carl being side tracked and losing focus on the main research task.

Overall, Carl has done some fantastic work in this research area and the outputs of his project will be a useful contribution to researchers working on the design and analysis of twins studies.

Dr Enes Makalic
Dr Katrina Scurrah

7 June 2018

Declaration

This is to certify that

1. this project is evidence of my own work, with direction and assistance provided by my project supervisor(s),
2. due acknowledgement has been made in the report to all other material used, and
3. this work has not been previously submitted for academic credit

Carl Higgs, July 2018

Abstract

Background

A power analysis for the detection of differences in correlations between identical and non-identical twin pairs was undertaken to support researchers in the early stages of a classic twin study.

Methods

Results

Conclusion

Contents

Acronyms

DZ Dizygotic twins
ICC Intra-class correlation coefficient
FZ Fisher's Z (test)
GV Generalised Variable (test)
MSLR Modified signed log-likelihood ratio (test)
MZ Monozygotic twins
PT Permutation test
SLR Signed log-likelihood ratio (test)

1. Introduction

The classic twin study exploits the differing degrees of genetic relatedness in identical (monozygotic) and non-identical (dizygotic) twins in order to draw inferences on the heritability of traits. In broad terms, heritability is the degree to which variation in a trait or phenotype, such as propensity to gain body weight, or become a centenarian, can be attributed to shared genetic effects. The calculation and comparison of Pearson correlations across the two twin groups is a routine preliminary step undertaken by researchers in this field. However, a range of factors can impact on researchers' ability to detect a true effect given data.

This thesis reports on a simulation-based power analysis for the detection of differences in correlations between monozygotic and dizygotic twin pairs under a range of scenarios, and the associated development of R functions and an applied interactive power calculator. These tools address an identified absence of tools for this fundamental step in the twin study context.

In this chapter we first define key genetic concepts, how these are exploited by the classic twin study, and some of the assumptions which are relied upon in order to make such inferences. Then, the concept of power analysis and related statistical concepts are defined in order to provide background, justification and notation for the following chapters.

1.1 The classic twin study and power

Identical twins (monozygotic, MZ) arise from the same zygote, or fertilised egg, and are genetically very similar. Non-identical twins (dizygotic, DZ) arise from fertilisation of two seperate eggs, and are as genetically alike as ordinary siblings. The comparison of phenotypic traits within identical and non-identical twin pair samples allows for partitioning the variance in traits into that attributable to shared environment, individual environment or to genetics [1]. Known as the classic twin study, this approach can help us to better understand the mechanics of health and disease processes so that we can develop intervention measures which appropriately target the hypothesised causal mechanisms.

Falconer, heritability and difference in MZ and DZ; Paraphrase Visscher re heritability and the various sense of this

Contemporary twin studies use mixed effects and structural equation modelling to evaluate differences in variance components for a particular trait between mono and dizygotic twins accounting for differential within pair similarities related to zygosity [2, 3]. Michael Neale and colleagues developed methods and software to facilitate the analysis of variance components in twin studies using structural equation modelling [2]. Brad Verhulst, building on the work of Peter Visscher, developed functions for power analyses in this variance component modelling context, for example detecting a difference in genetic correlations [4, 5, 6].

Visscher's work noted that the ratio of MZ to DZ twins is an important parameter to consider with regard to power to estimate variance components in twin studies, involving a trade off between power to detect additive genetic ($> 1:1$) or shared environmental effects ($< 1:1$) [6]. Such effects may be the main objects of interest in contemporary studies when planning a study, and power calculators such as that of Verhulst [4] specifically cater to these concerns.

However, a routine preliminary step undertaken by researchers in this field is the calculation and comparison of Pearson correlations across the two twin groups. This research

project focuses on the reported needs of researchers undertaking this early analysis step. Indeed, of 26 results returned from a 'Best Match' PubMed database search for "twin pearson difference", five published between 1994 and 2014 made comparisons between MZ and DZ twin groups based on differences in Pearson correlations; based on the stated correlations and group sample sizes, and using a standard formula based power test we implemented (described later), three of these articles reported differences in correlations based on tests with 40% or lower power (see Appendix B). While the use of simple statistical tests in the early stages of analysis is justifiable, the application of these and reporting of results as though constituting meaningful evidence may be misleading. The use of power analysis before a study, or even once recruiting has been completed, may better guide researchers to investigate problems which can more reasonably be solved given the samples and data available, and processes of interest.

In undertaking a twin study we make certain assumptions, understanding that these likely do not strictly hold in practice, but the key one for the purposes of this power analysis is that our data is approximately normally distributed.

1.2 Correlation: Historical and statistical background

1.2.1 Pearson product-moment correlation

The statistical treatment of correlation was popularised by Francis Galton. In context of broad social interest in eugenics, Galton described methods which could be used to describe the 'co-relatededness' of variables sourced from closely related family members [7, 8]. Karl Pearson was a keen follower of Galton's research, and writing in the context of inheritance and natural selection made use of 'Galton's function', describing it as a coefficient of correlation [9]. Known today as the product-moment or Pearson correlation coefficient, the function is used to describe the magnitude and direction of linear change in one random variable as another changes [10, 11, 12].

Inference based on the Pearson correlation coefficient assumes bivariate normality. That is, given two random variables (x_1, x_2), we assume they have jointly normal probability distribution [13]

$$p(x_1, x_2) = \frac{1}{2\pi\sigma_{x_1}\sigma_{x_2}\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left\{\frac{(x_1-\mu_{X_1})^2}{\sigma_{X_1}^2} - \frac{2\rho(x_1-\mu_{X_1})(x-\mu_{X_2})}{\sigma_{X_1}\sigma_{X_2}} + \frac{(x_2-\mu_{X_2})^2}{\sigma_{X_2}^2}\right\}}$$

The population correlation ρ (rho) is the covariance $\mathrm{Cov}(x_1, x_2) = \mathrm{E}[(x_1 - \mu_{X_1})(x_2 - \mu_{X_2})]$ divided by the product of the two variables' standard deviations $\sigma_X\sigma_Y$. The estimate of ρ based on a sample of observed data is denoted as r.

$$r = \frac{\sum_{i=1}^{n}(x_{1_i} - \bar{x}_1)(x_{2_i} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{1_i} - \bar{x}_1)^2 \sum_{i=1}^{n}(x_{2_i} - \bar{x}_2)^2}}$$

If the variables $\boldsymbol{x}$ are standardised using their respective sample means $\bar{x} = \frac{1}{n}\sum_{i=1} x_i$ and standard deviations $s = \frac{1}{\sqrt{n-1}}$ where n is the sample size, as u

$$u = \frac{1}{\sqrt{n-1}}(x_i - \bar{u}),$$

then the standardised regression coefficient b_1 arising from the simple linear regression of $u_1 = b_0 + b_1 u_2$ is equal to the Pearson correlation [14].

1.2.2 Spearman rank correlation

In the case of data with outliers which violate our assumptions of bivariate normality, the non-parametric Spearman's correlation r_s may be considered as an alternative to Pearson's: using the same formula for r, is is calculated using the rank order transformation of the variables rather than their raw values [15, 16].

1.2.3 Alternative approaches to correlation

The scope of the present study was restricted to Pearson and Spearman correlations, however alternative approaches to estimating correlation should be considered and for completeness reviews of the following are included in appendix A: Kendall's τ; partial correlations; and intra-class correlations. The latter two are of particular relevance to twin studies.

1.2.4 Inference on observed correlations

Ronald Fisher observed that a geometric transformation of the correlation coefficient using its inverse hyperbolic tangent could be used to approximate a normal distribution [17].

$$z = \operatorname{arctanh}(r) = \frac{1}{2}\log_e \frac{1+r}{1-r}$$

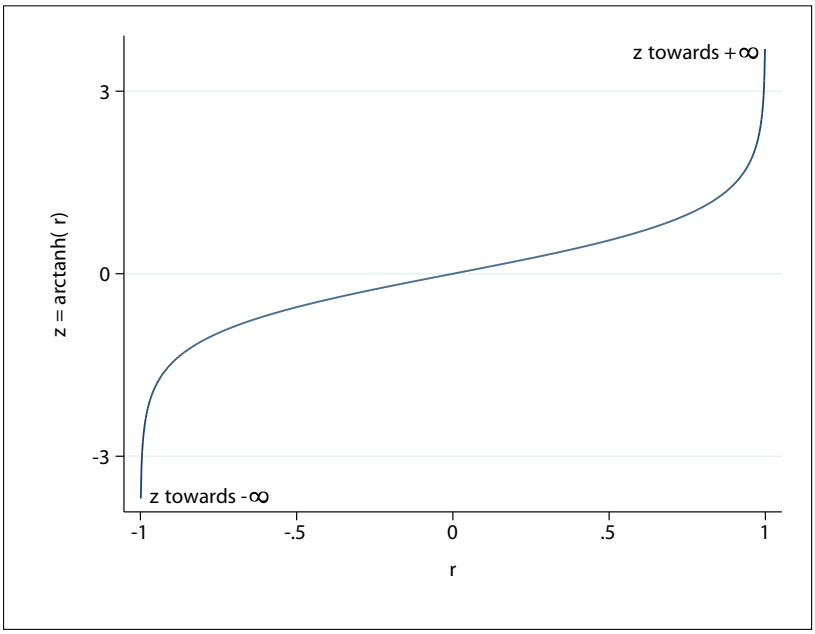

Fig. 1.1 Transformation of r to z using Fisher's method.

Fisher's Z transformation has the effect of mapping the distribution of correlation of coefficients from a domain of $(-1 < r < +1)$ to $(-\infty < z < +\infty)$, and is used for inferences on the Pearson correlation coefficient, as well as the rank-based and ICC

varieties described above and in Appendix A [10, 15, 18]. Through a comparison of the use of the z-transformation with exact and other methods used in hypothesis tests concerning r, David concluded that the z-transformation should be appropriate for most purposes [13]. Being a simple and accurate approximation of the normal distribution, this transformation known as Fisher's Z is ubiquitous in statistical treatment of correlation coefficients, for example when seeking to compare their differences and is employed in functions found in contemporary statistical software packages such as *R* and Stata [19, 16].

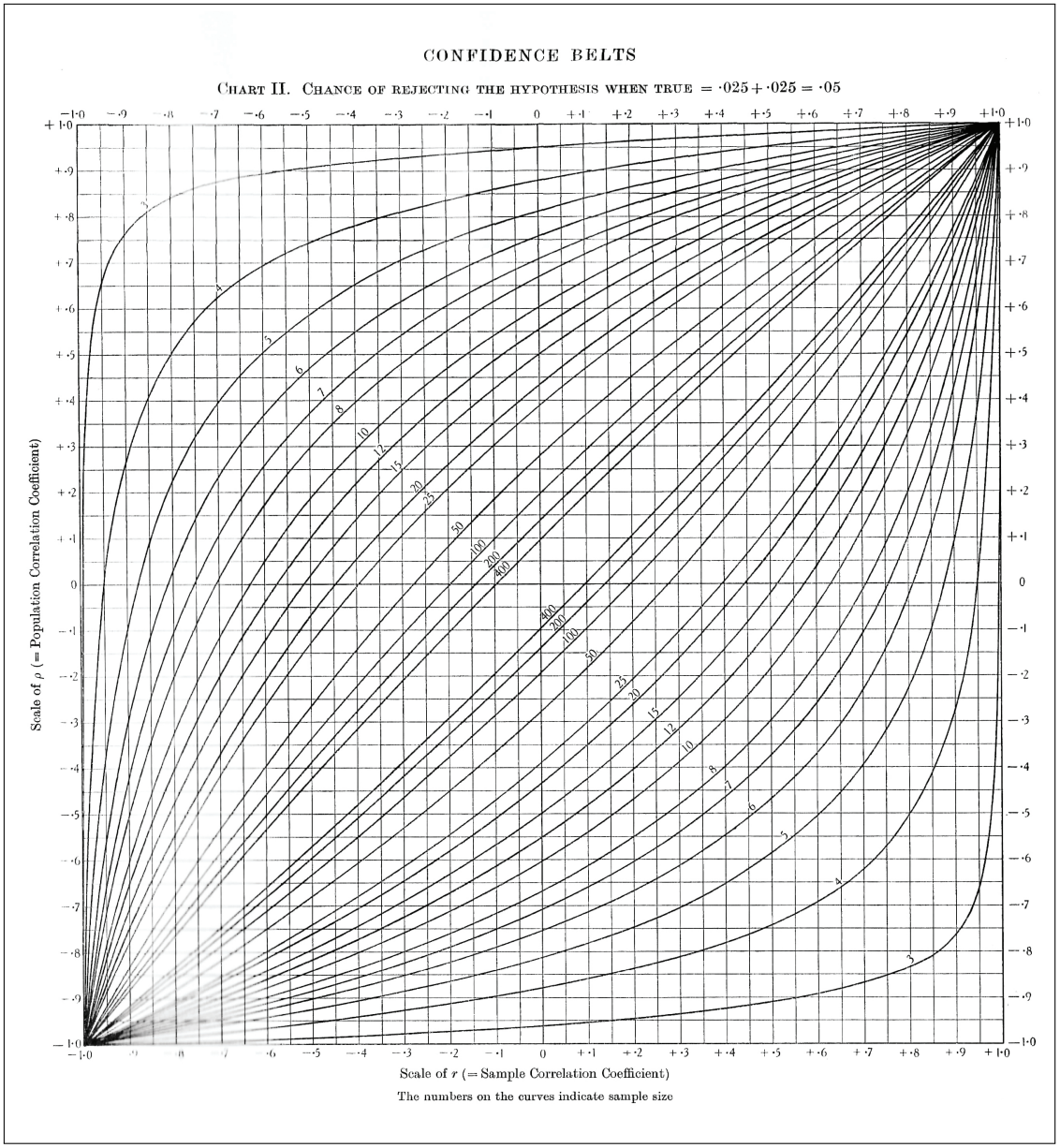

Fig. 1.2 This visualisation of the chance of rejecting the null hypothesis when true given alpha, rho and n was an early inspiration for visualising our simulation results [13].

1.3 Power analysis and hypothesis testing

Power analysis involves a compromise between type 1 (α) and type 2 (β) error thresholds, respectively representing the expected proportion of null hypotheses to be rejected when true and not rejected when false [20]. These could be chosen to suit the requirements of a particular study, but for historical reasons the respective values chosen tend to be 5% ($\alpha = .05$) and 20% ($\beta = 0.2$) [21].

1.3.1 Significance testing and controversy

There are controversies around the choice of such values, relating to what it means to deem something as 'statistically significant', a synonym for $p < \alpha$. A recent commentary suggested redefining significance as $p < 0.005$ [22]; other authors writing in a biostatistics context have advised reporting of exact p-values to be prefered in general to dichotomisation at a somewhat arbitrary α level, and that there is justification for use in some contexts of confidence intervals based on $\alpha = 0.1$ [23]. These are epistemological questions, detailed discussion of which is beyond the scope of this research project. However it is important to acknowledge that the present study is concerned with an applied statistical question (power to detect differences in correlations in groups with distinct genetic relatedness) with strong historical links to the hypothesis testing paradigm.

1.3.2 Importance of power analysis

Regardless of debates surrounding the use of hypothesis testing and thresholding, the question of power is one of academic rigour with moral implications: if a study is to be conducted with support through significant contribution of public resources in the way of grant funding, there is an ethical imperative to consider *a priori* whether the experiment in question can anticipate production of meaningful evidence [20, 21]. Power analysis

aims to answer this question through consideration of the sample size required to detect an effect of such magnitude to be considered meaningful given the specified values of α.

1.3.3 Calculation of power

The probability of observing a true effect, the power of a statistical test, is equal to the proportion of null hypotheses rejected when false; that is, $1 - \beta$ [21]. Hence, the choice of β will indicate directly the intent for power: if in planning a twin study we agree an appropriate value for β is 0.2, then we are saying that our power to detect a difference of the required magnitude between identical and non-identical twins should be at least as high as 0.8, or 80%. The ability to achieve this power depends on sample size, the ratio of MZ to DZ twins, the magnitude of the respective correlations themselves, and the degree to which our assumptions of bivariate normality hold. The robustness of the choice of test method employed to departures from distributional assumptions is also an important consideration.

{The aim of this report This report describes methods and results of a simulation analysis comprised of more than half a million power estimates, detailing more than 100,000 distinct MZ and DZ twin comparison scenarios evaluated across five different approaches to detecting a difference in correlations. These can be accessed to answer specific questions. The task of this report is to evaluate subsets of these and demonstrate the value and validity of the functions giving rise to them, and provide options to researchers such as those of Twins Australia for their analyses which may be expanded upon.

2. Methods

The preceding chapter summarised a review of the literature relating to twins, correlations, power and simulations undertaken to inform our approach to analysis. Through this review we identified a series of suitable approaches and tests for evaluating differences in Pearson and Spearman correlations in two groups. Important considerations were the efficiency of our implemented simulation functions, and a well-designed data structure to support our planned as well as future outputs. The *R* programming environment was used for all analyses [19].

2.1 Hypothesis tests for difference in correlations

2.1.1 Fisher's Z test (analytical approach)

A test statistic for the difference in correlations can be calculated as the difference in Fisher's Z transformed values weighted by the approximate standard error of the difference [10, 13]

$$t_{\hat{\theta}} = \frac{\hat{\theta}}{se_{\hat{\theta}}} = \frac{z_{MZ} - z_{DZ}}{\sqrt{(n_{MZ} - 3)^{-1} + (n_{DZ} - 3)^{-1}}}$$

The type 2 error rate β is estimated through comparison of this test statistic $t_{\hat{\theta}}$ to a reference value on the standard normal distribution. Using the cumulative normal distribution function Φ (Phi), the reference score q is calculated as the normal probability quantile corresponding to our α level divided by the sidedness of our test.

$$q = \Phi^{-1}(\alpha/\text{sidedness})$$

Where we refer to sidedness, we mean whether we are concerned with a single- or two-tailed probability. Here, we are testing the hypothesis that $\rho_{MZ} = \rho_{DZ}$ using a two-tailed p-value, implying 'admissible alternatives' to be the case that ρ_{MZ} is greater than or less than ρ_{DZ}, that is, $|\rho_{MZ} - \rho_{DZ}| > 0$ [13]. One-tailed consideration, for example that $\rho_{MZ} > \rho_{DZ}$, is not considered in this report however, the functions developed may be parameterised in this way if desired.

Employing the concepts detailed above, β is calculated as

$$\beta = \Phi\big(q - t_\theta\big)$$

Our power estimate for the detection of difference in correlations is $\text{power}(\theta) = 1 - \beta$. Putting the above altogether, we calculate power using the Fisher's z test statistic as,

$$\text{power} = 1 - \Phi\left(\Phi^{-1}_{\alpha/2} - \text{abs}\left(\frac{\text{arctanh}(r_{MZ}) - \text{arctanh}(r_{DZ})}{\sqrt{(n_{MZ} - 3)^{-1} + (n_{DZ} - 3)^{-1}}}\right)\right)$$

The code we used to implement the analytic Fisher's Z test approach to power calculation in *R* is displayed in listing 2.1.

Listing 2.1 Fisher's Z test (analytic approach)

```
# Fishers Z test - no sim
fz_nosim <- function(r1,r2,n1,n2,
                     alpha = 0.05, sidedness=2,method = "pearson",
                     power = TRUE) {
  # Calculate Fisher's Z
  z1     <- atanh(r1)
  z2     <- atanh(r2)

  # Take difference
  zdiff  <- z1-z2

  # Calculate standard error and test statistic
  z_se   <- sqrt(1/(n1-3) + 1/(n2-3))
  z_test <- zdiff/z_se

  # Optionally return p-value for observing diff at least this large under H0
  z_p    <- sidedness*pnorm(-abs(z_test))
  if (power == FALSE) return("p" = z_p)
  z_ref   <- qnorm(1-alpha/sidedness)
  z_power <- 1-pnorm(z_ref - abs(z_test))
  return(z_power)
}
```

The above method is the de facto standard, as used for example in the Stata `power two correlations` [16]. However, other options for evaluating the difference in Pearson or Spearman correlations should be considered.

2.1.2 Fisher's Z test (simulation approach)

Using a simulation approach we take our hypothesis tests and apply them to draws from simulated data designed to mimic our samples through parameterisation using the hypothesised underlying bivariate population distributions.

So where in our formula we might plug in hypothesised sample coefficients of 0.2 and 0.5, in the simulation we use these values to represent the true correlations in the underlying population from which we draw our samples. Over a large number of simulations of bivariate twin data the proportion of hypothesis tests returning p-values lower than our type 1 error threshold is our power estimate.

The simulation-based Fisher's Z test function *R* code is displayed in listing 2.2.

Listing 2.2 Fisher's Z test (simulation approach)

```
# Fishers Z test
fz <- function(a,b,sidedness=2,method = "pearson") {
  # Two samples
  n1 <- nrow(a)
  n2 <- nrow(b)

  # Compute z-transformed sample correlation coefficients
  z1     <- atanh(cor(a,method = method)[2,1])
  z2     <- atanh(cor(b,method = method)[2,1])
  zdiff  <- z1-z2

  # calculate standard error and test statistic
  z_se   <- sqrt(1/(n1-3) + 1/(n2-3))
  z_test <- zdiff/z_se

  # return p-value
  z_p    <- sidedness*pnorm(-abs(z_test))
  return(z_p)
}
```

In addition to applying the Fisher Z test in a simulation context, alternate tests we identified and implemented for inclusion in our simulation study were as follows.

2.1.3 Zou's confidence interval

Zou's confidence interval is used to calculate a confidence interval for the difference in two correlations, and a hypothesis test employing this method is featured in the R package `cocor` [24, 12]. A hypothesis test using Zou's confidence interval evaluates whether zero lies within the lower and upper bounds of the interval estimate of the difference in correlations, returning 1 if so or otherwise zero. Over a run of simulations this would be expected to return identical results to the Fisher Z test, but may be more efficient.

Zou's approach expands on earlier work [25] to calculate a confidence interval for a difference in correlations using a so-called Simple Asymptotic approach, using what Zou refers to as a Modified Asymptotic method [24]. Both approaches draw heavily on Fisher's earlier work [10]. The modified asymptotic method of Zou consists of first calculating confidence intervals for the two respective z-transformed correlations (transformed as per Fisher's method, described above):

$$(l_{z_k}, u_{z_k}) = z_k \pm \sqrt{\frac{1}{n_k - 3}} \times \Phi^{-1}_{\alpha/2}, \text{ where } k \in \{1, 2\}$$

Then, the lower (L) and upper (U) bounds of the modified asymptotic confidence interval for the difference in correlations are calculated:

$$L = r_1 - r_2 - \sqrt{(r_1 - \tanh(l_{z_1}))^2 + (\tanh(u_{z_2}) - r_2)^2}$$

$$U = r_1 - r_2 + \sqrt{(\tanh(u_{z_1}) - r_1)^2 + (r_2 - \tanh(l_{z_2}))^2}$$

If zero is within the bounds of the confidence interval for the difference, the test returns as 1, and otherwise 0.

Our implemenation of the Zou's confidence interval test function is displayed in listing 2.3.

Listing 2.3 Zou's confidence interval

```
zou <- function(a,b,alpha = 0.05,sidedness=2,method = "pearson") {
  # From Zou (2007) and used in Cocor (note typo for U in paper; should be '+')
  #  However, really, this is equivalent to fz test for hypothesis testing purposes

  # compute z- transformed correlations and differences
  r  <- c(cor(a,method = method)[2,1], cor(b,method = method)[2,1])
  z  <- atanh(r)
  zdiff  <- z[1]-z[2]

  # calculate standard error for respective z scores
  n  <- c(nrow(a),nrow(b))
  z_se   <- sqrt(1/(n-3))

  # calculate reference threshold
  z_ref  <- qnorm(1-alpha/sidedness)

  # calculate respective confidence intervals
  ci_mat <- matrix(c(-1,-1,1,1),nrow = 2, ncol = 2, dimnames =list(c("Mz","Dz"),c("l","u")))
  z_ci   <- z + ci_mat * z_se * z_ref
  r_ci   <- tanh(z_ci)

  # calculate Zou's Modified Asymptoptic confidence interval for difference in correlations
  L      <- r[1]-r[2] - sqrt((r[1]     - r_ci[1,1])^2 + (r_ci[2,2] - r[2]     )^2)
  U      <- r[1]-r[2] + sqrt((r_ci[1,2] - r[1]     )^2 + (r[2]      - r_ci[2,1])^2)
  r_diff_ci <- c(L,U)

  # return test value (0 or 1, however, in the power context this resolves to same outcome as p)
  ci_test <- (L < 0) && (0 < U)
  return(c(ci_test,r_diff_ci))
}
```

2.1.4 Generalised Variable Test

The generalised variable (GV) test involves transformation of the simulated sample correlations into so-called pivotal quantities the difference of which is used to calculate a test statistic and p-value [26]. Synthesising two reported approaches [27, 28] this test was first implemented as an example by my supervisor Enes Makalic in a Matlab script, and subsequently adapted by myself as a function in R. A compiled version using RCPP to leverage C++ routines for random number draws was suggested by my colleague Koen Simons, and adopted to improve the function's run time. However, this later version was not compatible with the parallelised simulation approach, and in this context the non-RCCP 'GVT-r' version was used.

Given two bivariate normal samples $k \in \{1, 2\}$, the sample correlation coefficients r_k are used to estimate two respective quantities $r_k^* = \frac{r_k}{\sqrt{(1-r_k^2)}}$, and the generalised variables G_{ρ_k}:

$$G_{\rho_k} = \frac{r_k^* \sqrt{W_k} - U_k}{\sqrt{(r_k^* \sqrt{(W_k)} - U_k)^2 + V_k}}$$

where,

$$U_k \sim N(0,1),\ V_k \sim \chi^2_{n_k-1},\ \text{and}\ W_k \sim \chi^2_{n_k-2}$$

A p-value using the GV test is calculated as twice the value of the smaller of two quantities: the proportion of differences in G_{ρ_k} less than 0, and the proportion greater than 0.

The GV test function *R* code is displayed in listing 2.4.

Listing 2.4 GV test (R version)

```
gvt_r <- function(a,b,M=1e4,method = "pearson") {
  # Two samples
  n1 <- nrow(a)
  n2 <- nrow(b)

  # Compute sample correlation coefficients
  r1 <- cor(a,method = method)[2,1]
  r2 <- cor(b,method = method)[2,1]
  r  <- c(r1,r2)

  # Generate random numbers
  V2     <- matrix(data=0, nrow = M, ncol = 2)
  V2[,1] <- rchisq(M, df = n1-1, ncp = 1)
  V2[,2] <- rchisq(M, df = n2-1, ncp = 1)

  W2     <- matrix(data=0, nrow = M, ncol = 2)
  W2[,1] <- rchisq(M, df = n1-2, ncp = 1)
  W2[,2] <- rchisq(M, df = n2-2, ncp = 1)

  Z <-matrix(data = rnorm(2*M), nrow=M, ncol = 2)

  # Compute test statistic
  rstar <- r/sqrt(1-r^2)
  top   <- c(sqrt(W2[,1])*rstar[1],sqrt(W2[,2])*rstar[2]) - Z
  G     <- top / sqrt( top^2 + V2 )

  # Compute p value
  Grho <- G[,1] - G[,2];
  p    <- 2*min( mean(Grho<0), mean(Grho>0) );
  return(p)
}
```

2.1.5 Signed log-likelihood ratio test

The signed log likelihood ratio (SLR) test is formulated as the signed difference in sample correlation coefficients multiplied by the square root of the sum of respective coefficients' log-likelihoods. The test here is a partial implementation of a recently reported modified signed log-likelihood ratio (MSLR) test for differences in two bivariate normal correlations [28]. The SLR and MSLR tests are well established general hypothesis tests [29, 30, 31, 26], the novelty in Kazemi and Jafari's approach being the applied context of difference in correlations. However, we (myself, nor my supervisors) were unable to successfully replicate the 'modified' portion of Kazemi and Jafari's reported algorithm. Due to time constraints, and noting that the 'unmodified' SLR test appeared to return

p-values similar to the other hypothesis tests it was decided that inclusion of the SLR test would be a valid option to consider.

The SLR test function *R* code is displayed in listing 2.5.

Listing 2.5 Signed log-likelihood ratio test

```
slr <- function(a,b,M=1e4,sidedness=2,method = "pearson") {
  # Signed Log-likelihood Ratio test (an 'unmodified' version of test
  # described in Krishnamoorthy and Lee, Kazemi and Jafari , DiCiccio etc)
  # Two samples
  n  <- c(nrow(a),nrow(b))

  # Compute z-transformed sample correlation coefficients
  r  <- c(cor(a,method = method)[2,1], cor(b,method = method)[2,1])
  z  <- atanh(r)

  # Calculate average z as a plug in value
  rf <- tanh(mean(z))

  # calcaulte SLR
  slr <-sign(r[1]-r[2])*sqrt(sum(n*log(((1-rf*r)^2)/((1-r^2)*(1-rf^2)))))

  # return p-value
  p    <- 2 * (1 - pnorm(abs(slr)));
  return(p)
}
```

2.1.6 Permutation test

The permutation test is a non-parametric approach which compares the absolute difference of the Z-transformed sample correlations with those using correlations from a series of group membership permutations using the sample rank orders as values. Under a hypothesis of no difference in correlation, those differences arising from permutations would be assumed to be equally likely as those observed, or anticipated to be observed [32]. Across a series of M permutations (in this study, 10,000), a p-value is calculated as the proportion of permutation derived absolute differences ((abs($z^*_{MZ} - z^*_{DZ}$))) of greater magnitude than abs($z_{MZ} - z_{DZ}$).

The implementation of this permutation test in *R* is displayed in listing 2.6.

Listing 2.6 Permutation test

```
pt <- function(a,b,M=1e4,sidedness=2,method = "pearson") {
  # Based on Efron and Tibshirani, 1993
  # Store size, and calculate z-transformed correlations
  n  <- c(nrow(a),nrow(b))
  r  <- c(cor(a,method = method)[2,1], cor(b,method = method)[2,1])
  z  <- atanh(r)

  # Store rank-ordered vector representations, in one column
  v  <- cbind(rank(rbind(a[,1],b[,1]),ties.method = "random"),
              rank(rbind(a[,2],b[,2]),ties.method = "random"))
  # label rows
  rownames(v) <- c(rep("A",n[1]),rep("B",n[2]))

  # initial empty test vector
  rtest <- numeric(0)

  # run M permutations (default is 10,000),
  #  - returns test that absolute magnitude of difference
  #    is at least as great as that of the input z-transformed corr. diff.
  for (i in 1:M){
    permute <- cbind(v,rbinom(sum(n),1,0.5))
    rstar   <- c(cor(permute[permute[,3]==0,c(1,2)],method = method)[2,1],
                 cor(permute[permute[,3]==1,c(1,2)],method = method)[2,1])
    zstar   <- atanh(rstar)
    rtest   <- c(rtest,
                 abs(zstar[1]-zstar[2]) > abs(z[1]-z[2]))
    }

  # return p-value: proportion of test results at least as large as obs'd
  p <- mean(rtest)
  return(p)
}
```

2.2 Simulation

2.2.1 Approach for one simulation

A function `corr_diff_test()` was developed to undertake a single comparative simulation using any of the above tests (listing 2.7). Within a single simulation, each requested test is evaluated using samples drawn from the same two simulated bivariate populations (MZ and DZ), returning a *p*-value. The analytic Fisher Z test returns either a *p*-value or a power estimate based on the population parameters.

In the detailing of computational aspects of our methodology we use the term 'parameter' to refer to one of the options which may be specified within a function; an argument is the specific value which is passed to that parameter. For example, the parameter ρ for the respective MZ and DZ groups may be defined by specifying the argument `rho = c(-0.65,0.2)`. The main parameters which can be specified in the function call to `corr_diff_test()` are described with example arguments in Table 2.1. In the following text we refer to the set of parameters that give rise to a simulation of estimated

power for a particular test as a scenario. Power estimates returned across a series of tests for comparison purposes are refered to as a composite scenario.

Listing 2.7 Single run simulation code

```
corr_diff_test <- function(rho = c(.2,.5), n = c(30,90), distr = "normal",
                    param1a = c(0,0), param1b = c(0,0),param2a = c(1,1), param2b = c(1,1),
                    alpha = 0.05, sidedness = 2, test = c("fz","gtv","pt","slr","zou"),
                    method ="pearson", lower.tri = FALSE) {
  if(lower.tri==TRUE){
    # optionally, only calculate results for lower matrix half
    #   when comparing across all correlation combinations
    if(rho[1] < rho[2]) {
      return(NA)
    }
  }
  # initialise empty results vector
  results <- list()

  # if requested, process analytical Fisher's Z
  if ("fz_nosim" %in% test) {
    results[["fz_nosim"]] <- fz_ns_compiled(rho[1],rho[2],n[1],n[2],
                                            alpha = 0.05, sidedness = 2, method = method, power = FALSE)
    if(length(test)==1) return(results)
  }
  # process selected hypothesis tests, each using same draw of simulated data
  require("simstudy")
  a <- genCorGen(n[1], nvars = 2, params1 = param1a, params2 = param2a,
                 dist = distr, corMatrix = matrix(c(1, rho[1], rho[1], 1), ncol = 2),
                 wide = TRUE)[,2:3]
  b <- genCorGen(n[2], nvars = 2, params1 = param1b, params2 = param2b,
                 dist = distr, corMatrix = matrix(c(1, rho[2], rho[2], 1), ncol = 2),
                 wide = TRUE)[,2:3]
  if ("fz"       %in% test) results[["fz"]]       <- fz_compiled(a,b)
  if ("gtv"      %in% test) results[["gtv"]]      <- gtv(a,b) # uses rccp ; so elsewise compiled
  if ("gtvr"     %in% test) results[["gtvr"]]      <- gtv_compiled(a,b)
  if ("pt"       %in% test) results[["pt"]]       <- pt_compiled(a,b)
  if ("slr"      %in% test) results[["slr"]]      <- slr_compiled(a,b)
  if ("zou"      %in% test) results[["zou"]]      <- zou_compiled(a,b)[1]
  return(rbind(results[test]))
}
```

The function `corr_diff_test()` uses the *R* package `simstudy` function `genCorGen()` to generate bivariate correlated data for the simulated MZ (group `a` in the code above) and DZ (group `b`) twin pair samples [33]. The choice of available distributions and parameterisations is normal(μ, σ), binomial(probability p), Poisson(rate λ) gamma(μ, dispersion k), or uniform(min, max). In addition to specifying sample size, distribution and parameterisation, a correlation matrix may be specified; this was used to parameterise the underlying population correlations from which bivariate samples should be drawn. Three distinct distribution types were modelled in our simulation based power analysis: normal, 'mild' skew and 'extreme' skew. These are respectively explained in the captions of Figures 2.3, 2.4, and 2.5, which illustrate example sample draws from these distributions.

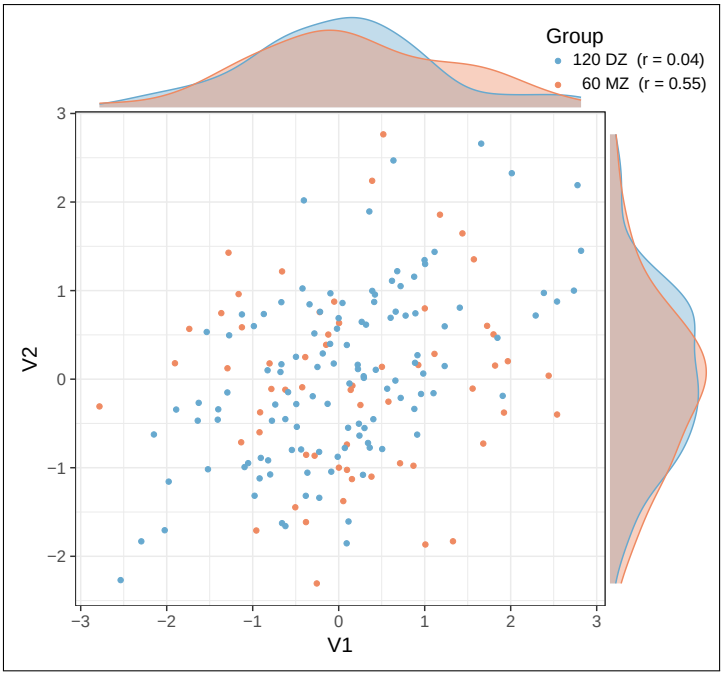

Fig. 2.3 An example of the bivariate normal scenario, with distributional assumptions asymptotically met. Both variables are standardised with mean $\mu = 0$ and standard deviation $\sigma = 1$.

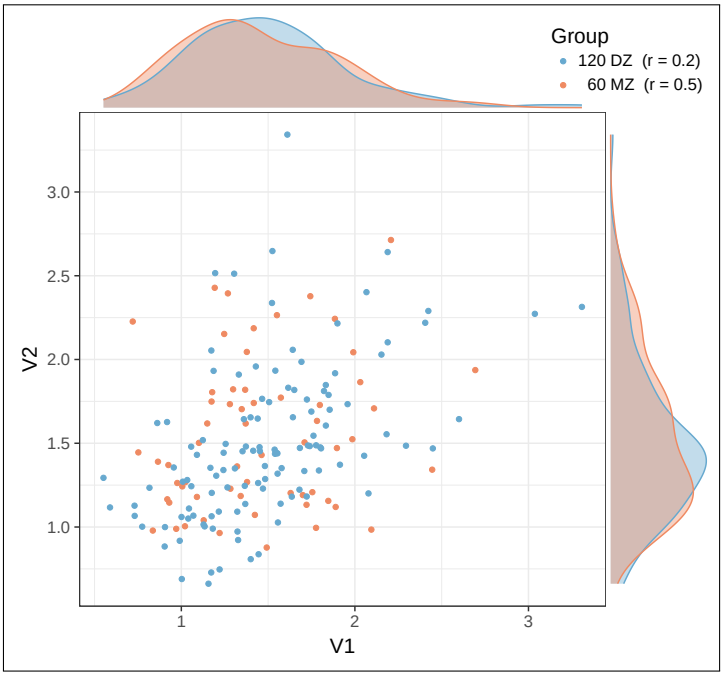

Fig. 2.4 A 'mild skew' scenario based on a gamma distribution with mean 1.5 and dispersion 0.09 (which the genCorGen function uses to inform shape and scale parameters for the distribution). This parameterisation was chosen through experimentation with the intent to represent a mild departure from an assumed normal population distribution, with a slight positive skew

2.2.2 Approach for multiple simulations

The `corr_diff_test()` function described above runs a single simulation of drawing from samples from two bivariate populations. However, for asymptotic normality to hold — the long run approximation of a normal distribution due to the Central Limit Theorem [11] — many more simulations must be run to achieve a fair assessment of the proportion

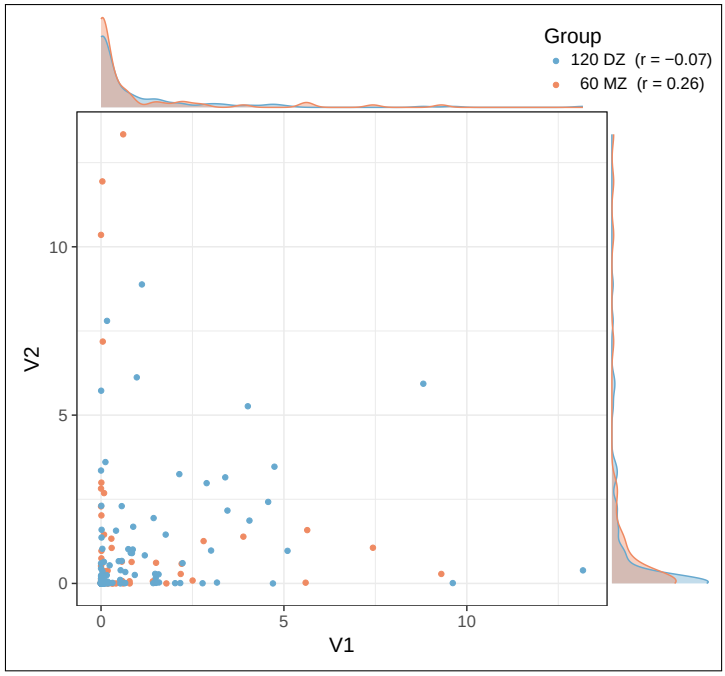

Fig. 2.5 An 'extreme skew' scenario based on a gamma distribution with mean 1 and dispersion 5. This results in an extreme positive skew to the distribution, analogous to that of biological processes where most observations will be clustered around a certain value, however some outliers may be extremely elevated.

Table 2.1 Description of parameter options for single simulation

Parameter	Description	Example arguments
`method`	Correlation method to use for testing difference	`'pearson'`
`rho`	ρ for each group's bivariate distribution	`c(-0.21,0.59)`
`n`	Sample size for groups 1 (MZ) and 2 (DZ)	`c(30,60)`
`dist`	Distribution to be used for both groups' bivariate distribution	`'normal'`
`param1a`	Distribution parameter 1 for respective samples in group 1	`c(0,0)`
`param1b`	Distribution parameter 1 for respective samples in group 2	`c(0,0)`
`param2a`	Distribution parameter 2 for respective samples in group 1	`c(1,1)`
`param2b`	Distribution parameter 2 for respective samples in group 2	`c(1,1)`
`test`	Tests to be evaluated	`c("fz_nosim","fz","gtv")`
`alpha`	α value to use for hypothesis tests	`0.05`
`sidedness`	Sidedness for hypothesis tests	`2`

of null hypotheses rejected when false for a particular scenario.

A wrapper function `corr_power()` allows for the `corr_diff_test()` function to be called M times, returning a power estimate for each test specified in the function call. These power estimates are derived from the series of simulated p-values for each test considered, or directly in the case of the analytic Fisher's Z test.

2.3 Scenario combinations

The flexibility of the simulation commands we developed allows for an very broad array of scenarios to be considered. A researcher could use these tools as is to aid in the development of a statistical analysis plan for their study. For this report we had to decide on a limited subset of scenarios for consideration. Two important considerations guiding this decision were the minimum number of scenario combinations required to facilitate useful comparisons of scenarios, and the time taken to run each of these. Our initial plan for the series of scenarios is reported in table 2.2.

Table 2.2 Description of parameter options for single simulation

Parameter	**Initial plan**	
	Argument resolution	**Combinations**
`method`	Pearson and Spearman correlations	2
`rho`	ρ combinations -0.95 through 0.95 at 0.5 resolution	$39^2 = 1521$
`n`	group combinations 15, 30, 60, 120, 240, 480, 960	$7^2 = 49$
`dist`	normal, 'mild skew', 'extreme skew'	3
`param1a`	dictated by distribution choice, above	-
`param1b`	dictated by distribution choice, above (equal to param1a)	-
`param2a`	dictated by distribution choice, above	-
`param2b`	dictated by distribution choice, above (equal to param2a)	-
`test`	Fisher's Z (analytic and sim), Zou's CI, GVT, SLR, PT	6
`alpha`	.05	1
`sidedness`	2	1
Total scenarios		2,683,044
Total simulations	1,000 simulations for each scenario	2,683,044,000
Total simulations	100, 1,000 and 10,000 simulations for each scenario (unrealistic!)	29,781,788,400

We decided *a prior* to set some parameters as fixed, : we only conducted two-sided tests with α of 0.05; having decided to limit ourselves to three basic distributional forms (normal, and two kinds of non-normal using distinct gamma distribution parameterisations) the distribution parameters were fixed to achieve these forms; the respective simulations of MZ and DZ twin pair samples each use the same distributional form, although the population correlation and sample size may vary (e.g. we don't compare bivariate normal MZ with a gamma skewed DZ); the simulated bivariate data for each twin group shares the same parameterisation (e.g. for normal data both MZ or DZ variables simulated based on

μ 0 and σ 1). Nevertheless, we were aware that our initial plan would be over-ambitious: were each scenario to be processed consecutively, each taking 1 second to process at an optimistically steady rate with no computer crashes we might expect 1,000 simulations of each scenario in Table 2.2 to take $2,683,044,000/60/60/24/365 \approx 85$ years! Under advice from my colleague Koen Simons, I undertook time tests of 1,000 iterations of each function and their byte code compiled versions, employing only the most efficient versions. The permutation test implementation was particularly inefficient, and given time constraints for improving the code's efficiency this was abandoned. Noting that the resolution of correlation combinations was a major contributor to anticipated length of processing time, this was reduced to comparison of correlations from -0.9 through 0.9 at a 0.1 resolution resulting in 361 instead of 1,521 correlation combinations. Based on the preliminary time tests, the initial and revised time estimates are displayed along with function time results in Table 2.3. These estimates suggested an anticipated running time of 16 days, based on my personal two core i7-processor laptop with 8gb RAM.

Table 2.3 Description of parameter options for single simulation

Test	**Time/1000 runs (secs)**	
	as is	**compiled**
Fisher's z (no sim)	0.03	0.02
Fisher's Z	0.56	0.36
GTV (r)	18.65	16.55
GTV (R C++)		12.08
Permutation (PT)	2473.18	2341.62
SLR	0.49	0.33
Zou's CI	0.38	0.28
Total 1 (PT, R-compiled GTV)	2493.29	2359.16
Total 2 (No PT, RCCP-compiled GTV)		13.07
Total 3 (No PT, R-compiled GTV)	20.11	17.54
Initial scenario: Total 1 $\times 1521 \times 49 \times 3 \times 2/60/60/24/365$	35 years	33 years
Best scenario: Total 2 $\times 361 \times 49 \times 3 \times 2/60/60/24$	25 days	16 days
Next best scenario: Total 3 $\times 361 \times 49 \times 3 \times 2/60/60/24$		21 days

The processing workflow used the R package data.table, a reported fast and memory efficient object for reading, writing and otherwise processing results **??**. The data.table was set up in wide format with rows per composite scenario (106,134 rows) and

columns per parameter with five additional columns to contain power estimates per test (13 columns, once processed). In long form, this corresponds to $5 \times 106,134 = 530,670$ *power estimate scenario rows. Each wide-form row corresponds to a call to the* `corr_power()` *wise using the* `data.table` *objects optimised processing syntax. A subsequent revision to the approach w*

The full set of composite scenarios were processed using 100 *simulations, and select scenarios using* 1,000

Scenario subset 1: How does sample size impact power estimates?

- Approach:
  - hold constant `(rho1==0.2)(rho2==0.5)`
  - process and report results for 100, 1000 and 10,000 simulations
- Considerations:
- number of simulations (validity)
- use of test
- use of Pearson or Spearman's correlation
- use of Distribution

Scenario subset 2: How does MZ:DZ ratio impact power estimates?

- Approach:
  - hold constant `n1 == 60 and n2 == 120`; ie. n = 180, MZ:DZ ratio 0.5:1
  - hold constant `n1 == 90 and n2 == 90`; ie. n = 180, MZ:DZ ratio 1:1
  - process and report results for 1000 simulations
- Considerations:
- use of test
- use of Pearson or Spearman's correlation
- use of Distribution
- magnitude and difference of correlations

Scenario A1 nrow(dtk[(n1==60)(n2==120),) 2166 rows

The 100 simulation set of results was used as initial proof of concept for the approach; the scenarios using 1,000 results provide the main results for the inferences made

in this report; the 10,000 simulations scenarios were produced in order to evaluate sensitivity of simulated power estimates to the number of simulation runs, and in particular establish validity of the choice of 1,000 simulations for main results.

2.4 Evaluating power

Using results of scenarios based on 10,000 simulations each, power estimates for detection of a difference in correlations from MZ twins with ρ 0.2 and DZ twins with ρ of 0.5 across a range of sample sizes and MZ:DZ ratios of .5:1 and 1:1 were compared using each of 5 tests, across the three distributions and measured with Pearson and Spearman correlations.

To evaluate the validity of using 1,000 instead of 10,000 simulations to estimate power, results from the same set of parameters giving rise to the above scenarios were considered in graphic form comparing each of 100, 1000 and 10,000 simulations. Holding all parameters except for sample size constant, a monotonic spline function based on power and $\log(n)$ is used to interpolate the power by sample size for each test. This in turn is used to estimate for each test the required sample size to achieve 80% power for these 6 composite scenarios (3 distributions, two MZ:DZ sample size combinations).

Impact of magnitude of correlations on power to detect a difference in correlations is evaluated using fitted contour plots for each test, again by the three distribution types and two MZ:DZ ratios using sample size of 180, each evaluated using 1,000 simulations. These 6 plots, evaluated across all correlation combinations for each test, are based on $6 \times 5 \times 361 = 10,830$ scenarios.

The same six composite scenarios as the power plots are used to evaluate

etc. I have put more old style contour plots in appendix, which may be referred to Ů

specific examples for GTV and SLR

results in format of images

contour plot

interpolation using monotonic increasing spline function

Required sample size to achieve 80

Difference to achieve 80

3. Results

Power estimates were calculated for a series of 530,670 parameter combination scenarios. Example draws from the three considered distributions of normal, mild skew, and extreme skew are displayed respectively in Figures 2.3, 2.4 and 2.5.

Mean power estimates by test were.... (note will point out in discussion why this metric is not useful)

Spline interpolated plots for required sample size to achieve 80% power were produced for the series of tests (see Figure 3.9).

The impact of group size ratio using a fixed N

Contour plots comparing by power given rho by test for MZ:DZ ratios of 60:120 and 90:90 (to discuss - hints of bias in SLR implementation viz. ratio)

Spline interpolated plots for required sample size given difference by test for MZ:DZ ratios of 60:120 and 90:90 (to discuss, why overlooking magnitude of respective correlations is problematic; hence, value in power calculator to address specific scenarios)

Appendix plots - test specific contour plots (don't show in results)

3.1 section title

Main results relate to images contained here. These will be introduced and explained, and involve the key important comparisons vis impact on power and required sample size to achieve this: number of simulations (justifies 1000 simulation approach); choice of ratio; choice of test method; impact of non-normality; magnitude of correlations; etc. I have put more old style contour plots in appendix, which may be referred to — specific examples for GTV and SLR

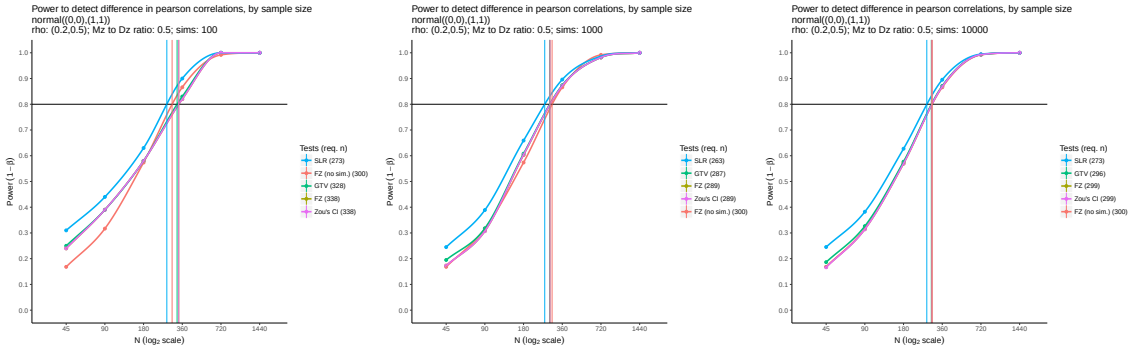

Fig. 3.6 Left to right, above: Sample size estimates to achieve 80% power using (a) 100, (b) 1,000, and (c) 10,000 simulations for bivariate normal distributions with respective ρ of 0.2 and 0.5 and an MZ:DZ ratio of 0.5:1.

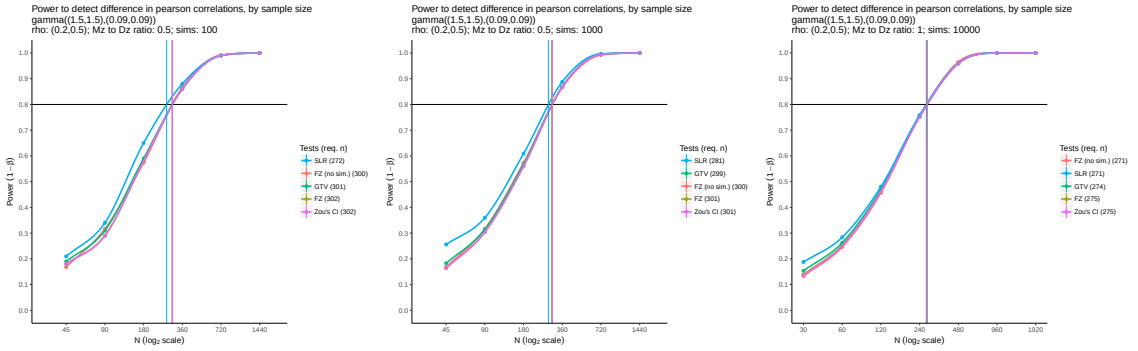

Fig. 3.7 Left to right, above: Sample size estimates to achieve 80% power using (a) 100, (b) 1,000, and (c) 10,000 simulations for bivariate "mild skew" gamma distributions with respective ρ of 0.2 and 0.5 and an MZ:DZ ratio of 0.5:1.

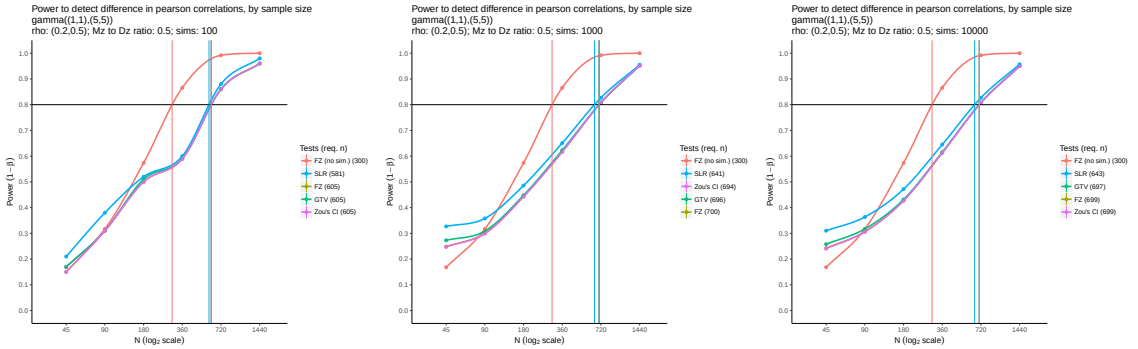

Fig. 3.8 Left to right, above: Sample size estimates to achieve 80% power using (a) 100, (b) 1,000, and (c) 10,000 simulations for bivariate "extreme skew" gamma distributions with respective ρ of 0.2 and 0.5 and an MZ:DZ ratio of 0.5:1.

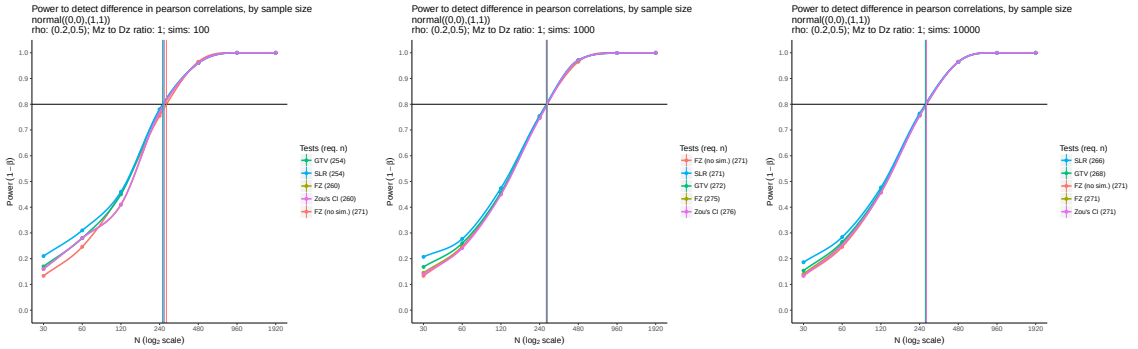

Fig. 3.9 Left to right, above: Sample size estimates to achieve 80% power using (a) 100, (b) 1,000, and (c) 10,000 simulations for bivariate normal distributions with respective ρ of 0.2 and 0.5 and an MZ:DZ ratio of 1:1.

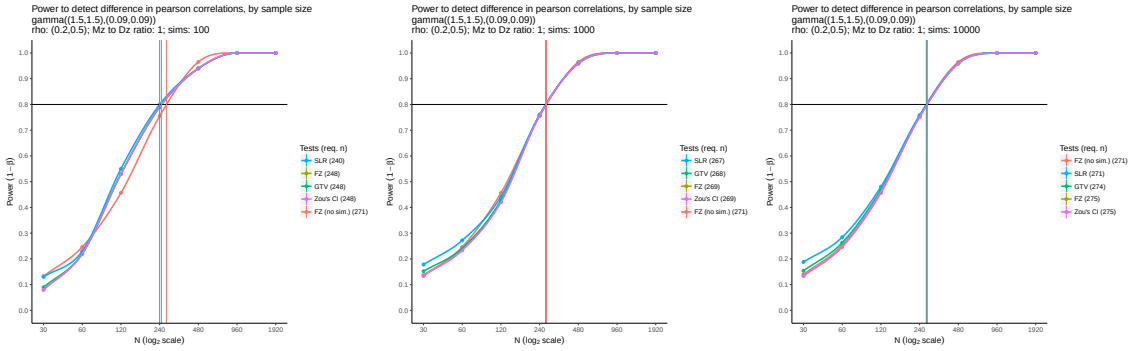

Fig. 3.10 Left to right, above: Sample size estimates to achieve 80% power using (a) 100, (b) 1,000, and (c) 10,000 simulations for bivariate "mild skew" gamma distributions with respective ρ of 0.2 and 0.5 and an MZ:DZ ratio of 1:1.

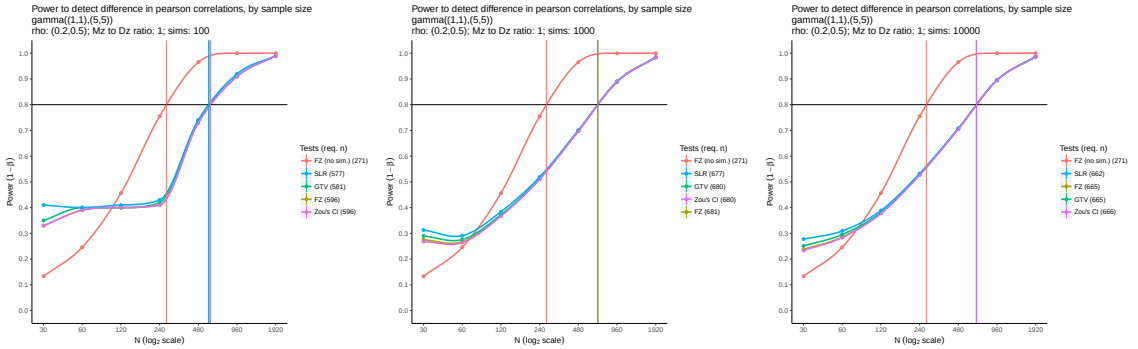

Fig. 3.11 Left to right, above: Sample size estimates to achieve 80% power using (a) 100, (b) 1,000, and (c) 10,000 simulations for bivariate "extreme skew" gamma distributions with respective ρ of 0.2 and 0.5 and an MZ:DZ ratio of 1:1.

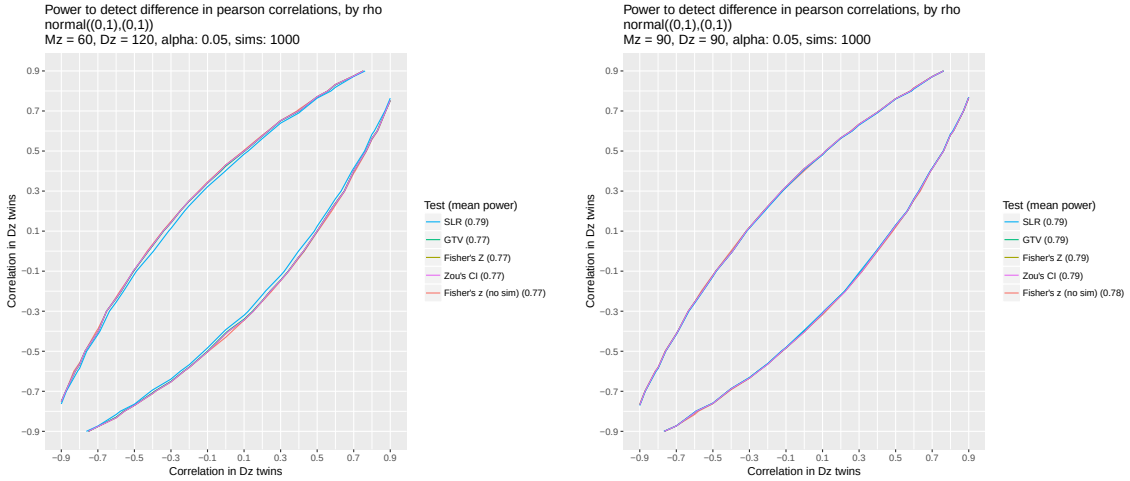

Fig. 3.12 Left to right, above: Contour plots of power given rho by test for bivariate normal distribution, using MZ:DZ group size ratio of (a) 60:120, and (b) 90:90

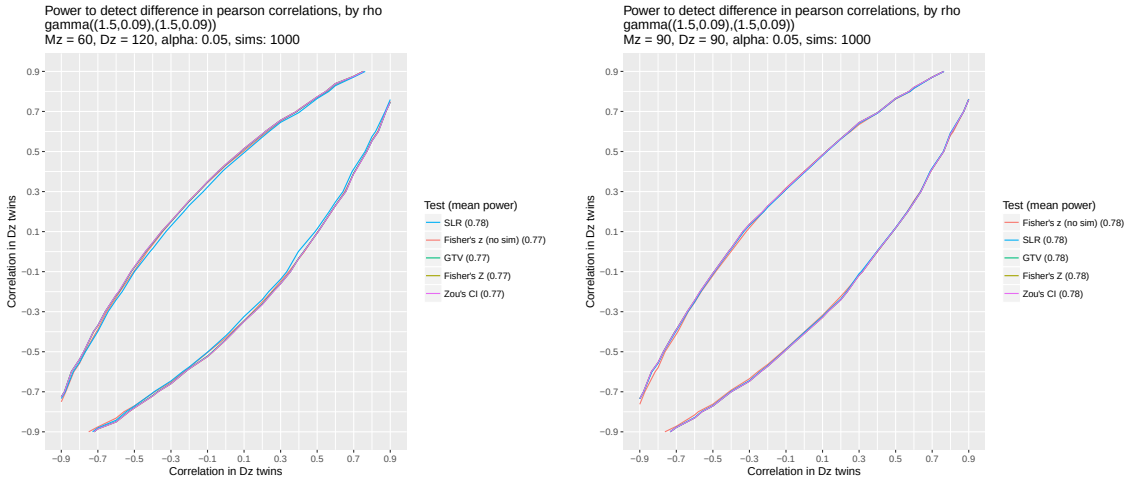

Fig. 3.13 Left to right, above: Contour plots of power given rho by test for bivariate "mild skew" gamma distribution, using MZ:DZ group size ratio of (a) 60:120, and (b) 90:90

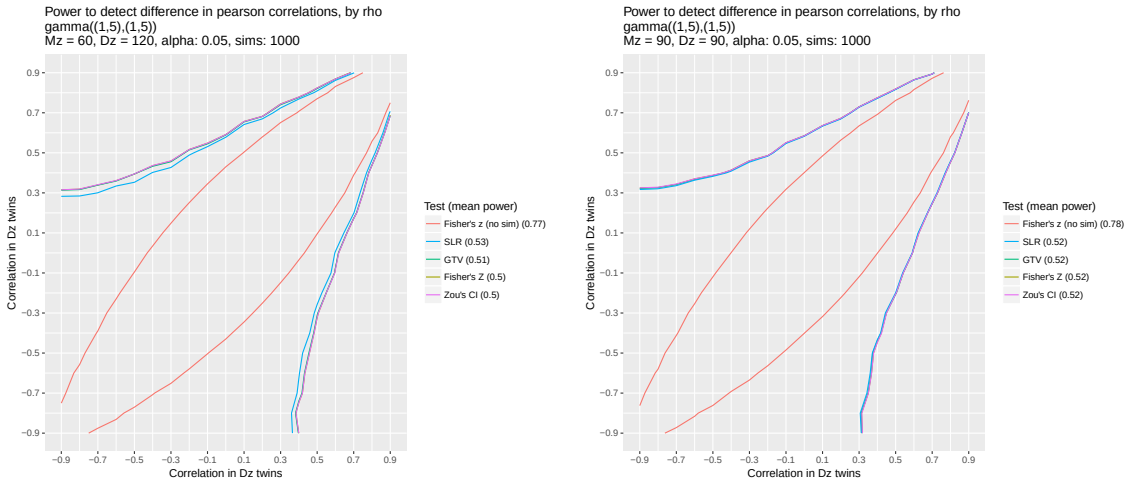

Fig. 3.14 Left to right, above: Contour plots of power given rho by test for bivariate "extreme skew" gamma distribution, using MZ:DZ group size ratio of (a) 60:120, and (b) 90:90

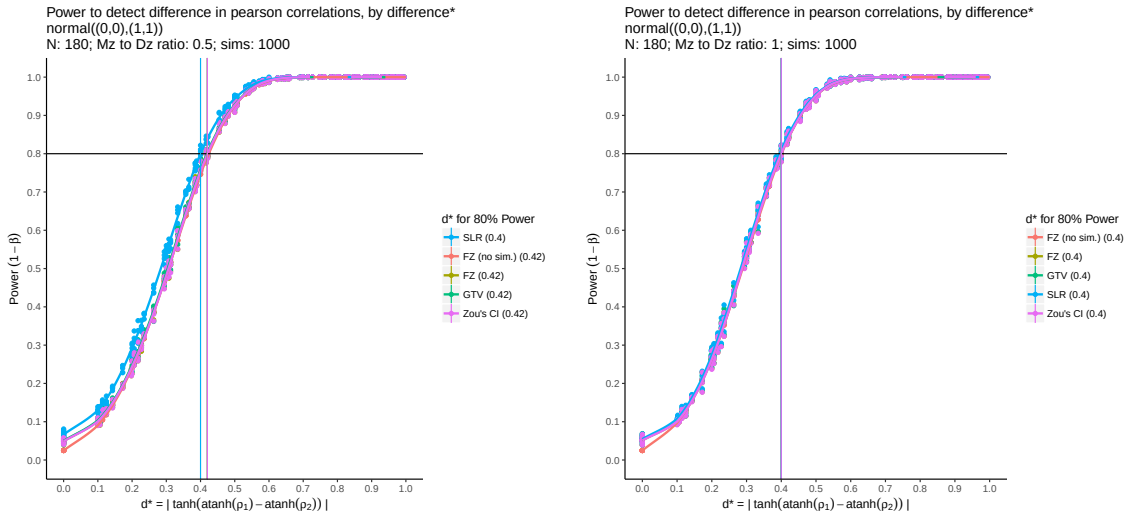

Fig. 3.15 Left to right, above: Power estimates by test given difference in rho for bivariate normal distribution, using MZ:DZ group size ratio of (a) 60:120, and (b) 90:90

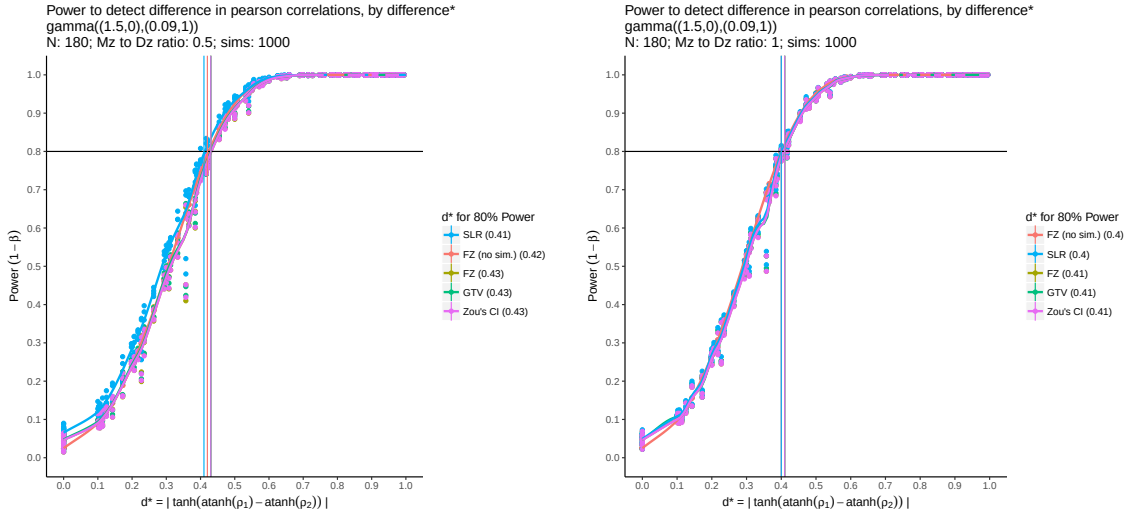

Fig. 3.16 Left to right, above: Power estimates by test given difference in rho for bivariate "mild skew" distribution, using MZ:DZ group size ratio of (a) 60:120, and (b) 90:90

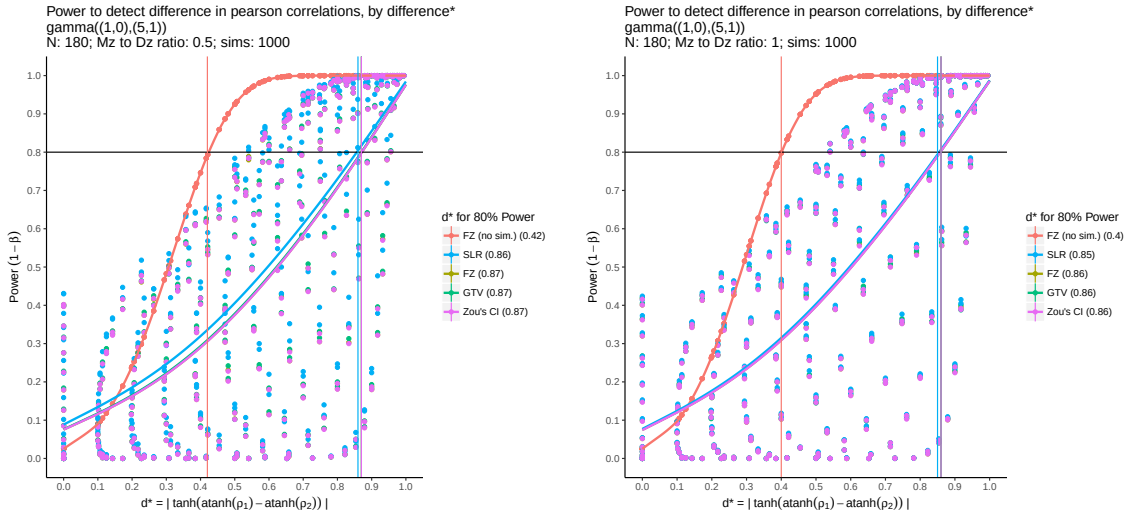

Fig. 3.17 Left to right, above: Power estimates by test given difference in rho for bivariate "extreme skew" distribution, using MZ:DZ group size ratio of (a) 60:120, and (b) 90:90

4. Discussion

4.1 section title

for now, this is plonked from presentation. I will discuss implications of results and relate to Visscher article, as well as those few results included in appendix - Twin Studies using Pearson correlations (some of whom's power seemed to post hoc check out; others were clearly underpowered). I know this is really the most important part. I'll get it in shape.

Efficiency (PT test); Number of simulations; ratio; problem with formula approach with assumption violation; problem with bias in methods (SLR); Strengths; Limitations; Next

Despite the reduced parameter set, our results are comprised of more than half a million power estimates which can be accessed to answer specific questions.

While the main use of such results are for considering particular scenarios, we can average over these for marginal estimates; under approximate normality the SLR test was estimated to have 82

The fitted contour plot on the left here compares power to detect a difference using all correlation combinations given groups sizes, in this example, under a bivariate normal distribution. The blue line indicates the 80

On the right sample size estimates to achieve 80

We can see here the extreme under estimation in sample size required to achieve 80

The SLR test appears a strong performer here - but its systematic elevation at tails of distribution is concerning – does this reflect a systematic bias, and perhaps innacurate rather than truly improved performance?

When no difference in correlations is present under bivariate normal distribution we

would expect power equivalent to our alpha level of 0.05, which, approximately, the other tests have; however, the SLR test has approximately 20%. This suggests its estimates here are upwardly biased.

After the SLR test the GTV was the next most powerful, but only marginally moreso than the Fisher's Z based tests.

We developed a flexible and extensible architecture geared to solving future problems.

discuss nested simulations - GVT and PT

However, the programming and analysis took longer than anticipated. The results I have shown today were processed using 100 simulations per parameter combination. The 1000 simulation results should be completed to update my results in a day or so.

There is more we can do to make this of more particular use in the twin context, for example by evaluating correlations in simulations using multivariable regression methods we could account for partial correlations; if we did this using mixed effects methods we could also consider power for difference in intra-class correlations.

Improve efficiency to allow higher resolution estimation

?? Cite Kaplen article on heritability meta-analysis — could lead into this by discussing the question of relavence of finding average power overlooking the particularities

Pre- and post-hoc analysis discussion; later and confidence intervals more use

keep emphasis on assumptions, hence need to consider a variety of scenarios

Difference plot is problematic an approach using the most pessimistic scenario (maximum required distance?) would arguably be more useful than average difference. Nevertheless, visualising the underlying scatter plot of estimates was useful to evalute impact of changing parameters such as the MZ:DZ ratio on power give difference.

4.2 Strengths

, and there are several benefits of this approach

The use of this approach particular scenario were recorded to ensure accurate scenario retrieval and ability to interpolate power estimates based on a subset of scenarios.

4.3 Computational considerations

It can be anticipated that simulation results based on 100 simulations would be unstable, however comparison with results produced using 1,000 and 10,000 simulations each allows for evaluation of sensitivity of power estimates to choice of simulation runs; in particular, to evaluate whether our goal of 1,000 simulations per scenario is sufficiently valid. It was originally intended to run all scenarios at 1,000 simulations each, however following a fortnight of processing on an 8 core workstation once results finished it was realised that (despite test runs) a syntax error in specifying the vector of additional columns to be created to contain results meant only the analytical result (which didn't require simulation) was retained.

Explain why not: Marginal estimates for power of each test across the range of correlations considered will be given by distribution and application of Pearson or Spearman test method — should be applied, not marginal

5. Conclusions

To sum up,

tests were overall quite similar

simulation important for power analysis where notable violation of assumptions of bivariate normality is anticipated (naive application of analytical Fisher's Z formula can be extremely over-optimistic in required sample estimates)

SLR implementation is biased highlights importance of critical consideration power of methods, as higher power may reflect bias. For this particular SLR test, power estimates in the case of unequal group ratios were upwardly biased.

Power should be considered contextualised using planned conditions for study and subject matter knowledge, and critical consideration of impact of methods employed.

We have created both the architecture for a process, as well as a database of simulation scenarios that can be interrogated. Both can be expanded as required. I have trialled an interactive power calculator web app, and it is planned incorporate the pre-processed database into this to allow on the fly estimates informed by our pre-processed results.

Glossary

2-tuple An *n*-tuple refers to a set of values. Due to the bivariate focus of our simulations, our simulation functions include a number of parameters which take 2-tuples of values. For example, for sample size `n = c(15,60` specifies an MZ group size of 15 and a DZ group size of 60; this in turn implies an MZ:DZ ratio of 1:4.

correlation A measure of the magnitude and direction of a linear relationship shared by two variables.

dizygotic Non-identical twins arising from fertilisation of two seperate fertilised eggs, and as genetically alike as ordinary siblings.

heritability The degree to which variation in a trait or phenotype, such as propensity to gain body weight, or become a centenarian, can be attributed to shared genetic effects.

monozygotic Identical twins, developing from the same fertilised egg (zygote) and genetically very similar.

r Sample estimate of the Pearson correlation coefficient.

ρ (rho) The Pearson correlation coefficient in the population.

References

[1] Teare MD (2011) Genetic epidemiology. Methods in molecular biology,, Humana Press, New York, NY

[2] Neale MC, Cardon LR (1992) Methodology for genetic studies of twins and families. NATO ASI series Series D, Behavioural and social sciences, Kluwer, Dordrecht

[3] Carlin JB, Gurrin LC, Sterne JA, Morley R, Dwyer T (2005) Regression models for twin studies: a critical review. Int J Epidemiol 34(5):1089–99, doi: 10.1093/ije/dyi153

[4] Verhulst B (2017) A Power Calculator for the Classical Twin Design. Behav Genet 47(2):255–261, doi: 10.1007/s10519-016-9828-9

[5] Visscher PM (2004) Power of the classical twin design revisited. Twin Res 7(5):505–12, doi: 10.1375/1369052042335250

[6] Visscher PM, Gordon S, Neale MC (2008) 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 11(1):48–54, doi: 10.1375/twin.11.1.48

[7] Galton F (1888) Co-Relations and Their Measurement, Chiefly from Anthropometric Data. Proceedings of the Royal Society of London 45:135–145

[8] Galton F (1890) Kinship and Correlation. The North American Review 150(401):419–431

[9] Pearson K (1895) Note on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London 58(347-352):240–242, doi: 10.1098/rspl.1895.0041

[10] Fisher S Ronald Aylmer (1990) Statistical methods, experimental design and scientific inference / R.A. Fisher ; edited by J.H. Bennett ; with a foreword by F. Yates. Oxford University Press, Oxford [England] ; New York :, ISBN 0198522290 :

[11] Casella G, Berger RL (2002) Statistical inference. Duxbury, Pacific Grove, CA, 2nd edn.

[12] Diedenhofen B, Musch J (2015) cocor: a comprehensive solution for the statistical comparison of correlations. PLoS One 10(3):e0121945, doi: 10.1371/journal.pone.0121945

[13] David F (1938) Tables of the ordinates and probability integral of the distribution of the correlation in small samples. Biometrika, London

[14] Tu YK, Gilthorpe MS (2012) Statistical thinking in epidemiology. CRC Press, Boca Raton, Fla.

[15] Fieller EC, Hartley HO, Pearson ES TESTS FOR RANK CORRELATION COEFFICIENTS. I. Biometrika 44(3-4):470–481, ISSN 0006-3444

[16] StataCorp (2013) Stata Statistical Software: Release 13. Tech. rep., StataCorp LP

[17] Fisher RA (1915) Frequency Distribution Of The Values Of The Correlation Coeffients In Samples From An Indefinitely Large Popultion. Biometrika 10(4):507–521, doi: 10.1093/biomet/10.4.507

[18] Barrett JH (2008) Measuring the Effects of Genes and Environment on Complex Traits, 55–69. Humana Press, Totowa, NJ, doi: 10.1007/978-1-60327-148-6$_4$

[19] R Development Core Team (2018) R: A language and environment for statistical computing. Version 3.5.0 (Joy in Planning). Tech. rep., R Foundation for Statistical Computing

[20] Freiman JA, Chalmers TC, Smith H, Kuebler RR (1978) The Importance of Beta, the Type II Error and Sample Size in the Design and Interpretation of the Randomized Control Trial. New England Journal of Medicine 299(13):690–694, doi: 10.1056/nejm197809282991304

[21] Cohen J (1988) Statistical Power Analysis for the Behavioral Sciences. Laurence Erlbaum Associates, New York

[22] Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, Bollen KA, Brembs B, Brown L, Camerer C, Cesarini D, Chambers CD, Clyde M, Cook TD, De Boeck P, Dienes Z, Dreber A, Easwaran K, Efferson C, Fehr E, Fidler F, Field AP, Forster M, George EI, Gonzalez R, Goodman S, Green E, Green DP, Greenwald AG, Hadfield JD, Hedges LV, Held L, Hua Ho T, Hoijtink H, Hruschka DJ, Imai K, Imbens G, Ioannidis JPA, Jeon M, Jones JH, Kirchler M, Laibson D, List J, Little R, Lupia A, Machery E, Maxwell SE, McCarthy M, Moore DA, Morgan SL, MunafÃş M, Nakagawa S, Nyhan B, Parker TH, Pericchi L, Perugini M, Rouder J, Rousseau J, Savalei V, SchÃűnbrodt FD, Sellke T, Sinclair B, Tingley D, Van Zandt T, Vazire S, Watts DJ, Winship C, Wolpert RL, Xie Y, Young C, Zinman J, Johnson VE (2018) Redefine statistical significance. Nature Human Behaviour 2(1):6–10, doi: 10.1038/s41562-017-0189-z

[23] Clayton D, Hills M (1993) Statistical models in epidemiology. Oxford University Press, Oxford

[24] Zou GY (2007) Toward using confidence intervals to compare correlations. Psychol Methods 12(4):399–413, doi: 10.1037/1082-989x.12.4.399

[25] Olkin I, Finn JD (1995) Correlations redux. Psychological Bulletin 118(1):155–164, doi: 10.1037/0033-2909.118.1.155

[26] Krishnamoorthy K, Lee M (2014) Improved tests for the equality of normal coefficients of variation. Computational Statistics 29(1):215–232, doi: 10.1007/s00180-013-0445-2

[27] Krishnamoorthy K, Xia Y (2007) Inferences on correlation coefficients: One-sample, independent and correlated cases. Journal of Statistical Planning and Inference 137(7):2362–2379, doi: https://doi.org/10.1016/j.jspi.2006.08.002

[28] Kazemi MR, Jafari AA (2016) Modified Signed Log-Likelihood Ratio Test for Comparing the Correlation Coefficients of Two Independent Bivariate Normal Distributions. ArXiv e-prints

[29] Barndorff-Nielsen OE (1986) Inference on Full or Partial Parameters Based on the Standardized Signed Log Likelihood Ratio. Biometrika 73(2):307–322, doi: 10.2307/2336207

[30] Barndorff-Nielsen OE (1991) Modified signed log likelihood ratio. Biometrika 78(3):557–563, doi: 10.1093/biomet/78.3.557

[31] Diciccio TJ, Martin MA, Stern SE (2001) Simple and accurate one-sided inference from signed roots of likelihood ratios. Canadian Journal of Statistics 29(1):67–76, doi: doi:10.2307/3316051

[32] Efron B, Tibshirani R (1993) An introduction to the bootstrap. Monographs on statistics and applied probability, Chapman Hall, New York

[33] Goldfeld K (2018) simstudy: Simulation of Study Data. R package version 0.1.8

Appendices

Appendix A

Alternative approaches to correlation

The scope of the present study was restricted to Pearson and Spearman correlations, however alternative approaches to estimating correlation should be considered and for completeness the following are reviewed below: Kendall's τ; partial correlations; and intra-class correlations. The latter two are of particular relevance to twin studies, however were beyond scope for inclusion for this research project.

A.1 Kendall's τ

An alternate non-parametric option is Kendall's τ (tau) which provides a summary measure of correlation based on concordance 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 (CD) is the score S; Kendall's τ_a is calculated as $\frac{S}{n(n-1)/2}$, while other variant formulas include further adjustment to account for ties [15, 16].

A.2 Partial correlation

Adjustment for additional covariates, say x_3, may be introduced resulting in what is known as a partial correlation representing the linear relationship between x_1 and x_2 adjusting for the effects of x_3 [10]. An elegant computational approach to calculating partial correlations drawing on the output of multiple regression analysis using standard

software packages uses the test statistic of interest $t = \frac{b}{se}$, the sample size n, and number of covariates k [16]. Using the example where we are interested in the partial correlation of x_1 (dependent variable in our regression model) and x_2 adjusting for the effects of u_3, or $\rho_{x_1 x_2 \cdot x_3}$, where $k = 2$:

$$\rho_{x_1 x_2 \cdot x_3} = \frac{t_{x_2}}{\sqrt{t_{x_2}^2 + n - k}}$$

A.3 Intra-class correlations

The correlation approaches described above may be considered to be inter-class correlations: in the twin context, this amounts to looking at variable(s) pertaining to a set of (arbitrary) first members of MZ twins, and comparing with the remaining MZ twins. This approach overlooks the paired nature of twin data. In contrast, intra-class correlations (ICCs) draw upon within-pair pooled mean and standard deviation [10]. If we refer to twin membership using the subscript j, and let n be number of twin pairs in our sample:

$$\begin{aligned}
\bar{x} &= \frac{1}{2n} \sum_{i=1}^{n} x_{ij} + x_{ij+1} \\
s^2 &= \frac{1}{2n} \sum_{i=1}^{n} (x_{ij} - \bar{x})^2 + (x_{ij+1} - \bar{x})^2 \\
r_{\text{ICC}} &= \frac{1}{ns^2} \sum_{i=1}^{n} (x_{ij} - \bar{x})(x_{ij+1} - \bar{x})
\end{aligned}$$

The above is considered a more accurate approach to correlation in contexts such as those of twin pairs, which do not have a natural ordering [10]. A number of approaches to statistical modelling can be taken to account for the paired twin data structure [3]; a frequently used approach involves the calculation of ICCs for both MZ and DZ twin pairs through use of mixed effects modelling. Calculated in this way, r_{ICC} represents the estimated ratio of between pair variation in a phenotype to total variation; the degree

to which $r_{\text{ICC:MZ}}$ is larger than $r_{\text{ICC:DZ}}$ can be inferred to relate to the shared genetic basis for variation in the phenotype [18].

Appendix B

Annotated PubMed searches

B.1 "twin pearson difference" - 15 May 2018

#	Article	PMID	Relevant? (additional info)	notes	Power checks out in theory? (using fz_nosim R function)
1	Grassi MA, Rao V, Winkler KP, Zhang W, Bogaard JD, Chen S, LaCroix B, Lenkala D, Rehman J, Malik AB, Cox NJ, Huang RS. Genetic variation is the major determinant of individual differences in leukocyte endothelial adhesion. PLoS One. 2014 Feb 10;9(2):e87883. doi: 10.1371/journal.pone.0087883. eCollection 2014.	24520339	y? (abstract states use of Pearson, article states use of icc)	Endothelial leukocyte adhesion in 23 MZ (0.6) and 23 sibling (0.25)	fz_nosim(0.6,0.25,23,23) is approx 28% power
2	LÃ³pez-SolÃ C, Fontenelle LF, Alonso P, Cuadras D, Foley DL, Pantelis C, Pujol J, YÃ¼cel M, Cardoner N, Soriano-Mas C, MenchÃ³n JM, Harrison BJ. Prevalence and heritability of obsessive-compulsive spectrum and anxiety disorder symptoms: A survey of the Australian Twin Registry. Am J Med Genet B Neuropsychiatr Genet. 2014 Jun;165B(4):314-25. doi: 10.1002/ajmg.b.32233. Epub 2014 Apr 23.	24756981	y	Correlations in symptom domains of OCD were approx double in MZ (503) vs DZ (445), including when conditioning on sex	All check out > 80%
3	Davis OS, Band G, Pirinen M, Haworth CM, Meaburn EL, Kovas Y, Harlaar N, Docherty SJ, Hanscombe KB, Trzaskowski M, Curtis CJ, Strange A, Freeman C, Bellenguez C, Su Z, Pearson R, Vukcevic D, Langford C, Deloukas P, Hunt S, Gray E, Dronov S, et al. The correlation between reading and mathematics ability at age twelve has a substantial genetic component. Nat Commun. 2014 Jul 8;5:4204. doi: 10.1038/ncomms5204.	25003214	n (author pearson)		
4	Trivedi R, Bhattacharya A, Mehta F, Patel D, Parekh H, Gandhi V. Cephalometric study to test the reliability of anteroposterior skeletal discrepancy indicators using the twin block appliance. Prog Orthod. 2015 Feb 25;16:3. doi: 10.1186/s40510-015-0073-1.	25769138	n (twin block design)		
5	Pearson LJ, Williamson JH, Turner SA, Lacy-Hulbert SJ, Hillerton JE. Peripartum infection with Streptococcus uberis but not coagulase-negative staphylococci reduced milk production in primiparous cows. J Dairy Sci. 2013 Jan;96(1):158-64. doi: 10.3168/jds.2012-5508. Epub 2012 Nov 8.	23141827	n (author pearson; cows)		
6	Melo P, GonÃ§alves B, Menezes A, Azevedo C. Socially adjusted synchrony in the activity profiles of common marmosets in light-dark conditions. Chronobiol Int. 2013 Jul;30(6):818-27. doi: 10.3109/07420528.2013.767823. Epub 2013 Jun 14.	23767997	n (marmosets; but Pearson correlation differences were considered as precursor to mixed model)		
7	Baschat AA, Gungor S, Glosemeyer P, Huber A, Hecher K. Changes in umbilical venous volume flow after fetoscopic laser occlusion of placental vascular anastomoses in twin-to-twin transfusion syndrome. Am J Obstet Gynecol. 2010 Nov;203(5):479.e1-6. doi: 10.1016/j.ajog.2009.11.013. Epub 2010 Sep 22.	20864074	n (neonates; no report on zygosity; not comparing differences)		
8	Defranco EA, Chang JJ, Macones GA, Muglia LJ. The correlation in birth timing between singleton and twin gestations in the same mother. J Matern Fetal Neonatal Med. 2009 Feb;22(2):106-10. doi: 10.1080/14767050802488204.	19089775	n (not comparing differences - also twins vs singletons)		

#	Article	PMID	Relevant? (additional info)	notes	Power checks out in theory? (using fz_nosim R function)
9	Blickstein I, Goldman RD, Mazkereth R. Maternal age and birth weight characteristics of twins born to nulliparous mothers: a population study. Twin Res. 2001 Feb;4(1):1-3.	11665318	n (but no full text available to check fully)		
10	Zarzycki P, Kasprzak M, Rzedzicki Z, Sobota A, Wirkijowska A, Sykut-DomaÅ„ska E. Effect of blend moisture and extrusion temperature on physical properties of everlasting pea-wheat extrudates. J Food Sci Technol. 2015 Oct;52(10):6663-70. doi: 10.1007/s13197-015-1754-y. Epub 2015 Feb 9.	26396414	n (wheat, not twins; not comparing differences)		
11	Hennequin Y, Rorive S, Vermeylen D, Pardou A. [Twins: interpretation of height-weight curves at birth]. Rev Med Brux. 1999 Apr;20(2):81-5. French.	10335101	n (curves, not correlations)		
12	Timmons BA. Twins' reactions to delayed auditory feedback. Percept Mot Skills. 1985 Oct;61(2):559-65.	4069922	m (abstract reported significance, not magnitude/direction; full text not accessible. Very small sample (10,10))		
13	Cho H, Guo G, Iritani BJ, Hallfors DD. Genetic contribution to suicidal behaviors and associated risk factors among adolescents in the U.S. Prev Sci. 2006 Sep;7(3):303-11.	16775759	y	(Sex stratified: F (mz 141, dz114), M (mz 141, dz 131); follow up with ACE model	main results depression (fem 0.54,0.21) checks out 80%
14	McGregor B, Pfitzner J, Zhu G, Grace M, Eldridge A, Pearson J, Mayne C, Aitken JF, Green AC, Martin NG. Genetic and environmental contributions to size, color, shape, and other characteristics of melanocytic naevi in a sample of adolescent twins. Genet Epidemiol. 1999;16(1):40-53.	9915566	n (author pearson; no full text though)		
15	Wee LY, Sebire NJ, Bhundia J, Sullivan M, Fisk NM. Histomorphometric characterisation of shared and non-shared cotyledonary villus territories of monochorionic placentae in relation to pregnancy complications. Placenta. 2006 Apr-May;27(4-5):475-82. Epub 2005 Jul 14.	16023205	n (not mzdz; not comparing differences)		
16	Falciglia GA, Norton PA. Evidence for a genetic influence on preference for some foods. J Am Diet Assoc. 1994 Feb;94(2):154-8.	8300990	y	(14 Mz and 21 Dz; consider diff using Fisher Z, only if sig in Mz not Dz - not a great approach); regardless, infer that a genetic contribution underlies preference for foods including broccoli	fz_nosim(0.59,0.3,14,21) is approx 14% power for difference in apple; reported '< .05' result had 40% power fz_nosim(0.55,-0.05,14,21)
17	Lau EY, Sampson WJ, Townsend GC, Hughes T. An evaluation of maxillary and mandibular rotational responses with the Clark twin block appliance. Aust Orthod J. 2009 May;25(1):48-58.	19634464	n (twin block appliance)		
18	Weissman A, Matanes E, Drugan A. Accuracy of ultrasound in estimating fetal weight and growth discordancy in triplet pregnancies. J Perinat Med. 2016 Mar;44(2):223-7. doi: 10.1515/jpm-2014-0187.	25478731	n		

#	Article	PMID	Relevant? (additional info)	notes	Power checks out in theory? (using fz_nosim R function)
19	Hoekstra RA, Bartels M, Verweij CJ, Boomsma DI. Heritability of autistic traits in the general population. Arch Pediatr Adolesc Med. 2007 Apr;161(4):372-7.	17404134	y	Compare correlations of MZ, DZ and twin siblings for 'first impression' of heritability of autism traits	Example is fz_nosim(0.59,0.36,33,31) ie. approx 20% power
20	Dammann KW, Smith C. Food-related environmental, behavioral, and personal factors associated with body mass index among urban, low-income African-American, American Indian, and Caucasian women. Am J Health Promot. 2011 Jul-Aug;25(6):e1-e10. doi: 10.4278/ajhp.091222-QUAN-397.	21721954	n ('Twin Cities' study area)		
21	Harlaar N, Meaburn EL, Hayiou-Thomas ME; Wellcome Trust Case Control Consortium., Davis OS, Docherty S, Hanscombe KB, Haworth CM, Price TS, Trzaskowski M, Dale PS, Plomin R. Genome-wide association study of receptive language ability of 12-year-olds. J Speech Lang Hear Res. 2014 Feb;57(1):96-105. doi: 10.1044/1092-4388(2013/12-0303).	24687471	n		
22	Ross JR, Bedi A, Clohisy JC, Gagnier JJ; ANCHOR Study Group., Larson CM. Surgeon Willingness to Participate in Randomized Controlled Trials for the Treatment of Femoroacetabular Impingement. Arthroscopy. 2016 Jan;32(1):20-4.e23. doi: 10.1016/j.arthro.2015.07.003. Epub 2015 Sep 26.	26395410	n		
23	Kotler M, Barak P, Cohen H, Averbuch IE, Grinshpoon A, Gritsenko I, Nemanov L, Ebstein RP. Homicidal behavior in schizophrenia associated with a genetic polymorphism determining low catechol O-methyltransferase (COMT) activity. Am J Med Genet. 1999 Dec 15;88(6):628-33.	10581481	n		
24	Loder RT. The sagittal profile of the cervical and lumbosacral spine in Scheuermann thoracic kyphosis. J Spinal Disord. 2001 Jun;14(3):226-31.	11389373	n		
25	Dale N, Pearson T, Frenguelli BG. Direct measurement of adenosine release during hypoxia in the CA1 region of the rat hippocampal slice. J Physiol. 2000 Jul 1;526 Pt 1:143-55.	10878107	n		
26	Son Y, Kang HJ, Song YM, Hwang JH. Relationships Between Self-awareness and Clinical Diagnostic Findings of Abnormal Foot Arch Height in Koreans. Ann Rehabil Med. 2017 Dec;41(6):1013-1018. doi: 10.5535/arm.2017.41.6.1013. Epub 2017 Dec 28.	29354578	n		

Appendix C

GTV test, varying ratio

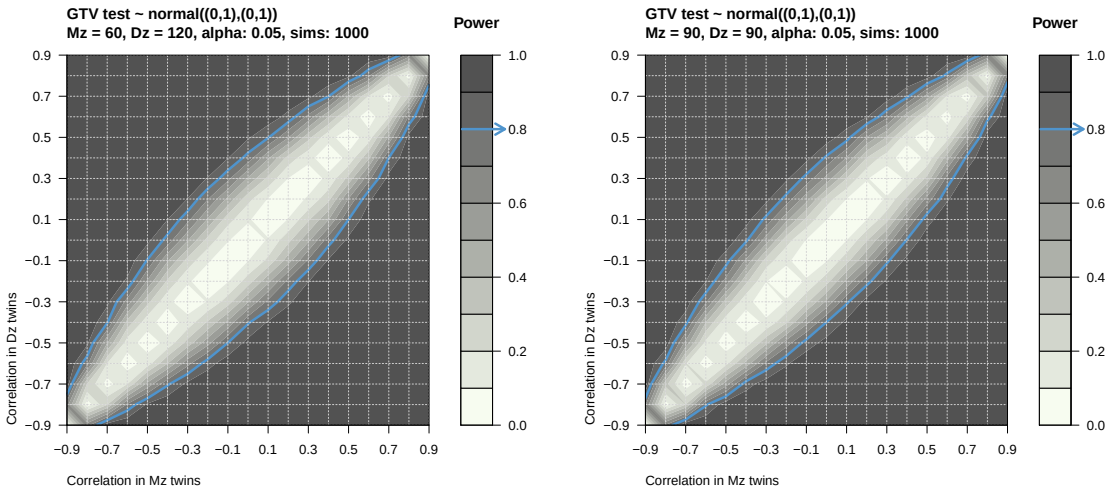

Fig. C.1 Left to right, above: GTV test power given rho contour plots for bivariate normal distribution, using MZ:DZ group size ratio of (a) 60:120, and (b) 90:90

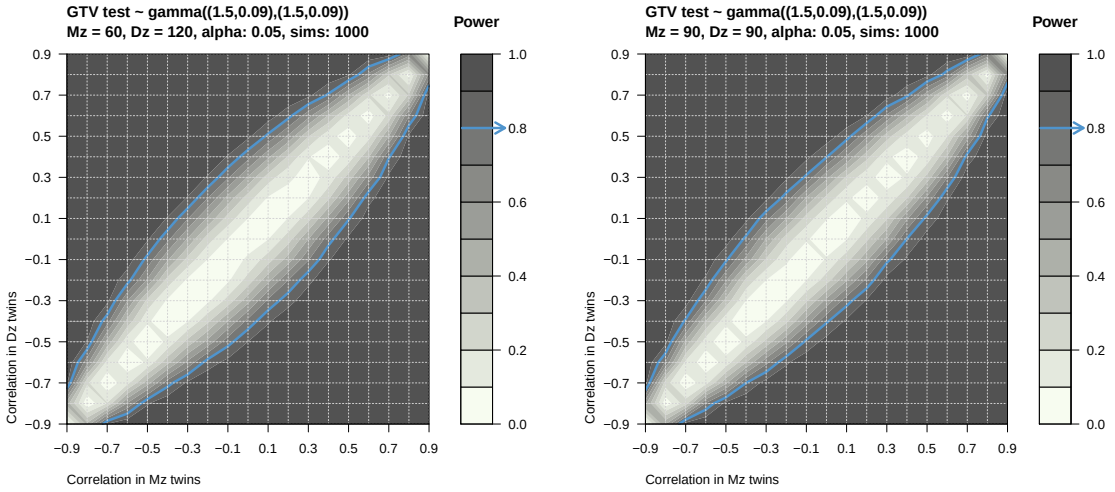

Fig. C.2 Left to right, above: GTV test power given rho contour plots for bivariate "mild skew" gamma distribution, using MZ:DZ group size ratio of (a) 60:120, and (b) 90:90

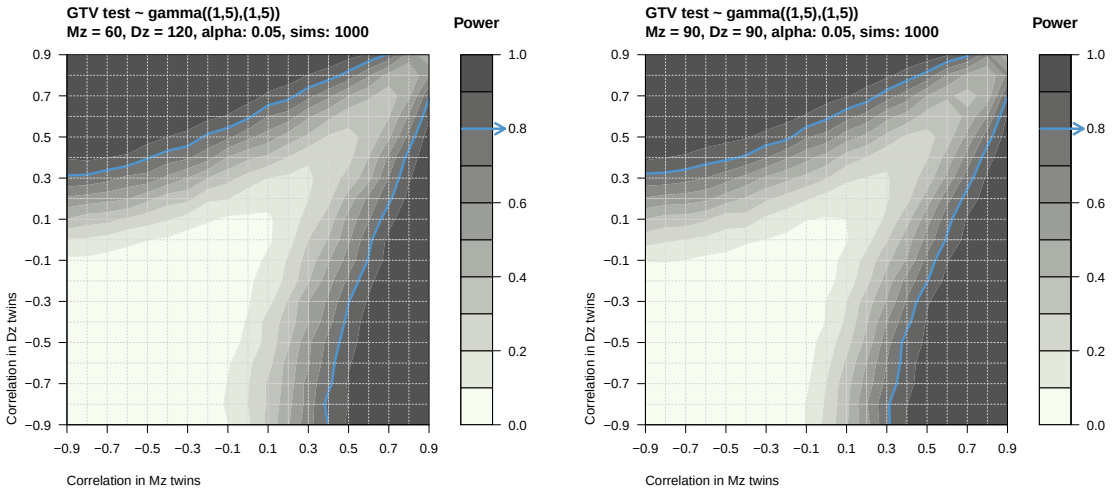

Fig. C.3 Left to right, above: GTV test power given rho contour plots for bivariate "extreme skew" gamma distribution, using MZ:DZ group size ratio of (a) 60:120, and (b) 90:90

Appendix D

SLR contour plot, varying ratio

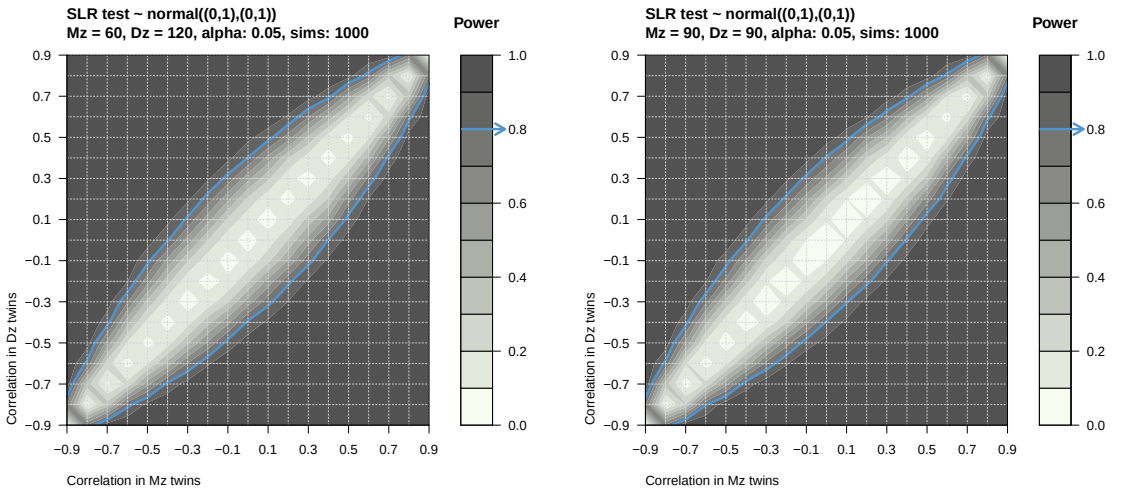

Fig. D.1 Left to right, above: SLR test power given rho contour plots for bivariate normal distribution, using MZ:DZ group size ratio of (a) 60:120, and (b) 90:90

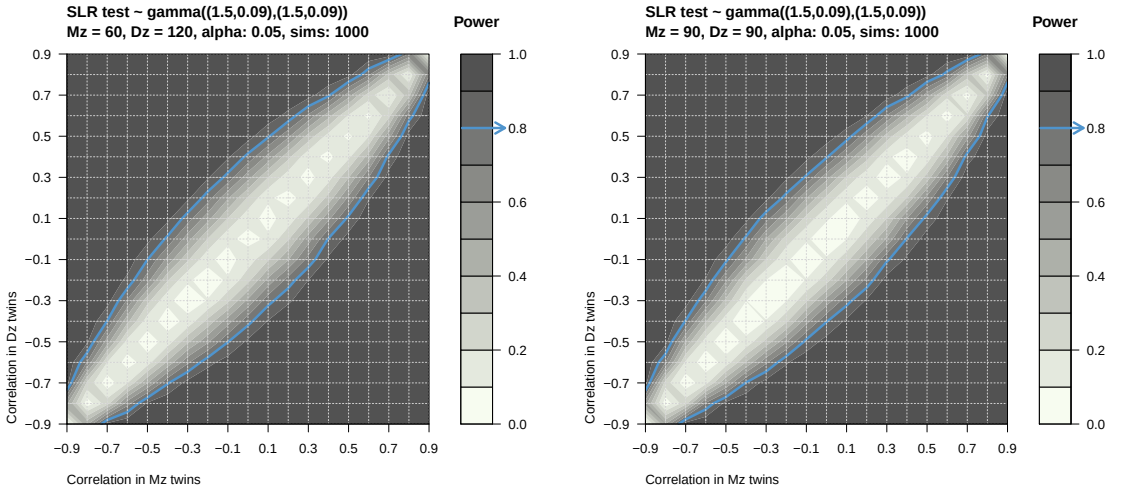

Fig. D.2 Left to right, above: SLR test power given rho contour plots for bivariate "mild skew" gamma distribution, using MZ:DZ group size ratio of (a) 60:120, and (b) 90:90

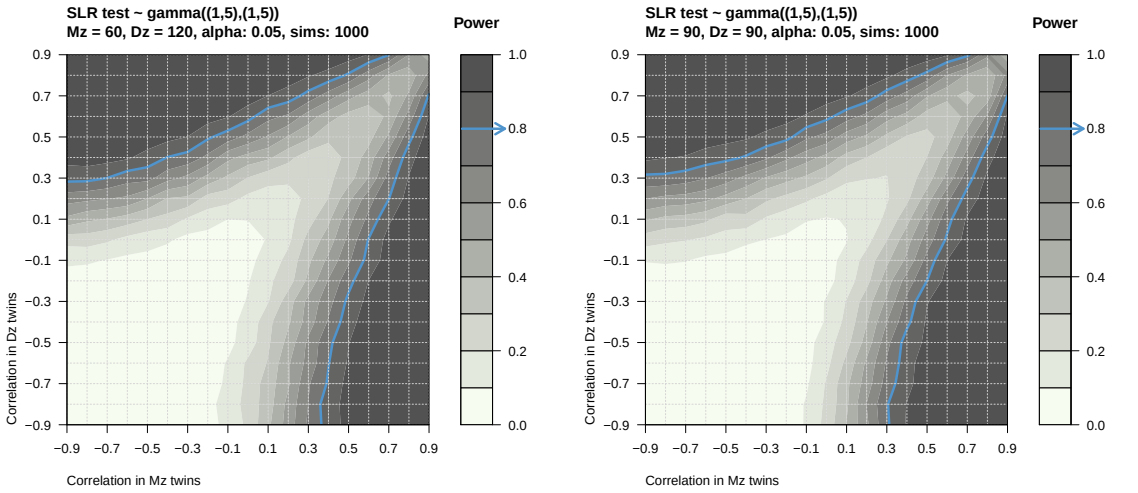

Fig. D.3 Left to right, above: SLR test power given rho contour plots for bivariate "extreme skew" gamma distribution, using MZ:DZ group size ratio of (a) 60:120, and (b) 90:90

Appendix E

R scripts