Assumptions of the SDir Method

Simulations Supported Evaluation SDir

Andrew D. Vigotsky Aaron R. Caldwell

Table of contents

Formalizing SD_{ir}

Let x_i be 0 when subject i receives the control intervention or 1 when they receive the experimental intervention. α_i is the effect of the control and β_i is the effect of the experimental intervention for subject i. In a trial, we are interested in estimating the experimental intervention's effect relative to the control's effect ($\gamma_i = \beta_i - \alpha_i$):

$$\begin{aligned} \text{post}_i &= \text{pre}_i + \alpha_i\,(1 - x_i) + \beta_i\,x_i \\ (\text{post}_i - \text{pre}_i) &= \alpha_i + (\beta_i - \alpha_i)\,x_i \\ \delta_i &= \alpha_i + \gamma_i\,x_i \end{aligned}$$

Nota bene the heterogeneity of treatment effects is $\text{Var}[\gamma]$.

$$\begin{aligned} \text{Var}[\delta \mid x = 0] &= \sigma_\alpha^2 \\ \text{Var}[\delta \mid x = 1] &= \sigma_\alpha^2 + \sigma_\gamma^2 + 2\,\rho_{\alpha,\gamma}\,\sigma_\alpha\,\sigma_\gamma \end{aligned}$$

The difference between these variances is $\text{Var}[\delta \mid x = 1] - \text{Var}[\delta \mid x = 0] = \sigma_\gamma^2 + 2\,\rho_{\alpha,\gamma}\,\sigma_\alpha\,\sigma_\gamma$. This implies that $\sigma_\beta^2 = \sigma_\alpha^2 + \sigma_\gamma^2$ when $\rho_{\alpha,\gamma} = 0$, which is SD_{ir}^2.

What does $\rho_{\alpha,\gamma} = 0$ imply about the correlation between counterfactual outcomes (α and β)?

$$\begin{aligned} \gamma_i &= \beta_i - \alpha_i \\ \implies \sigma_\gamma^2 &= \sigma_\beta^2 + \sigma_\alpha^2 - 2\,\rho_{\beta,\alpha}\,\sigma_\beta\,\sigma_\alpha \\ \implies \sigma_\beta^2 - \sigma_\alpha^2 &= \sigma_\beta^2 + \sigma_\alpha^2 - 2\,\rho_{\beta,\alpha}\,\sigma_\beta\,\sigma_\alpha \\ \implies \rho_{\beta,\alpha} &= \frac{\sigma_\alpha}{\sigma_\beta} \end{aligned}$$

This is a strong assumption. First, it implies that negative correlations between potential outcomes are $_{\text{impossible}}$, but there are reasonable cases when they can be expected (see below). Second, it implies that the treatment effect (γ) is orthogonal to the control effect (α), which may be a reasonable assumption in some but certainly not all experiments. Here, our goal is to dissect this assumption to provide a couple of examples of when it may and may not hold.

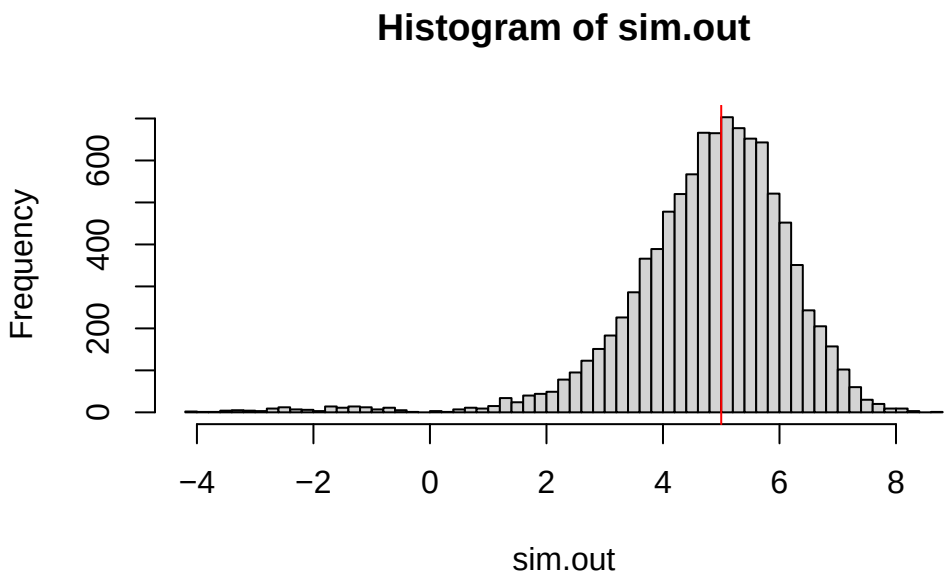
When are SD_{ir}'s assumptions reasonable?

One may be reasonably comfortable with the assumptions made by the SD_{ir} method and its associated data-generating model in a few different experimental contexts. In their original papers, proponents of SD_{ir} described parallel group experiments in which no or negligible

changes were expected in the control group, while change was expected in the intervention group (Hopkins 2015; Atkinson and Batterham 2015). Such contexts are arguably cases when SD_{ir}'s assumptions would be reasonable. For instance, in exercise science, a common relevant study would be one where inactive, healthy adults are randomized to one of two groups: non-exercise control or a resistance training intervention, which aims to answer the question: How does resistance training in healthy, inactive adults affect isometric knee extension strength? Suppose that these inactive adults are in a relatively stable period where their strength levels can be assumed to be stationary on average, albeit with some random fluctuations. The investigator may be fine with the principal assumption of the SD_{ir} method: Effects of the intervention are independent of *changes* that would occur in the control condition.

We can simulate such a study and show that the SD_{ir} provides (relatively) unbiased estimates of treatment effect heterogeneity (NB, a bias exists insofar as standard deviation is biased due to Jensen's inequality).

```
nsim <- 1e4
sigma_meas <- 5
n <- 100
D <- 20
sdir <- 5

calc.sdir <- function(x,y) sign(var(y)-var(x))*sqrt(abs(var(y)-var(x)))

sim.out <- c()
for(i in 1:nsim) {
  control_deltas <- rnorm(n,0,5)
  exp_deltas <- control_deltas + rnorm(n,D,sdir)

  obs_control <- control_deltas + rnorm(n,0,sigma_meas)
  obs_exp <- exp_deltas + rnorm(n,0,sigma_meas)

  sim.out <- c(sim.out,
               calc.sdir(obs_control, obs_exp))
}

hist(sim.out,breaks="fd")
abline(v=sdir, col="red")
```




```
print(paste("True SD of treatment effects:",sdir))
```

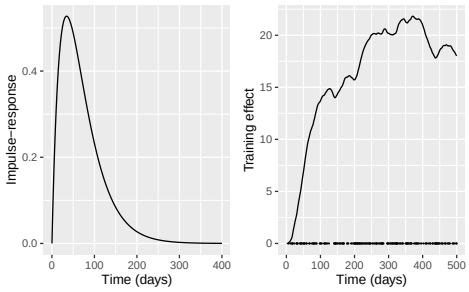
```
[1] "True SD of treatment effects: 5"
```

```
print(paste("SDir estimate:",mean(sim.out)))
```

```
[1] "SDir estimate: 4.78665655288726"
```

When are SD_{ir}'s assumptions unreasonable?

We'll now turn to an example in which the SD_{ir} assumptions do not hold. Suppose adaptations to a single training session are defined by the impulse-response function (left), and repeated exercise sessions result in cumulative adaptation (right).



We would like to study the effects of exercising approximately three times per week relative to no exercise. We will recruit from the general population, meaning that the individuals recruited will have (a) variable baselines, (b) variable training effects, and (c) varying levels of training experience. (a) and (b) will be modeled jointly using a multivariate log-normal distribution, which assumes that baseline levels (without any training) and adaptations are correlated. Finally, (c) training experience was modeled using the frequency of training, in which participant i's probability of training on a given day is $p_i \sim \text{Beta}(0.4, 0.8)$; the cumulative training effect from the 500 days before enrollment was then simulated and used as the pre-intervention score.

When participant i is randomized, they will either exercise every other day (intervention) or not at all (control) for 60 days. Thus, those who were well-trained prior to enrollment will experience little-to-no change (intervention) or detraining effects (control). Conversely, those who were untrained prior to enrollment will experience training effects (intervention) or little-to-no change (control).

```
sims <- pbmcapply::pbmclapply(1:500, function(x) {
  nsub <- 500
  probs <- rbeta(nsub, 0.4, 0.8)
  chars <- exp(MASS::mvrnorm(nsub,c(2,1),matrix(c(.25,0.125,0.125,0.25),2,2)))
  df <- c()
  for(i in 1:nsub) {
    baseline <- chars[i,1]
    a <- chars[i,2]

    trained <- rbinom(500,1,probs[i])

    trained0 <- c(trained,rep(0,60))
    trained1 <- c(trained,rep(c(0,1),length.out=60))

    ts <- 1:length(trained0)
    N <- length(ts)

    convolved0 <- pracma::conv(
      irf(ts, a, 30, 40),
      trained0
    )[1:N] + baseline

    convolved1 <- pracma::conv(
      irf(ts, a, 30, 40),
      trained1
    )[1:N] + baseline
```

```
    df <- rbind(df, data.frame(id = i,
                               group = as.numeric(i > nsub/2),
                               pre = convolved0[500],
                               post0 = convolved0[N],
                               post1 = convolved1[N]))
  }

  df$post <- ifelse(df$group,df$post1,df$post0)
  df$delta <- df$post - df$pre

  control <- subset(df, group == 0)
  exper <- subset(df, group == 1)

  data.frame(
    var_ir = var(exper$delta) - var(control$delta),
    varD = var( (df$post1-df$pre) - (df$post0-df$pre) ),
    rho = cor( df$post0-df$pre , df$post1-df$pre ),
    rho1 = cor( df$post0-df$pre , df$post1-df$post0 )
    )
}, mc.cores = 8)

colMeans( do.call(rbind, sims) )
```

```
    var_ir       varD        rho       rho1
-5.9385712 17.7849303  0.7710205 -0.4387016
```

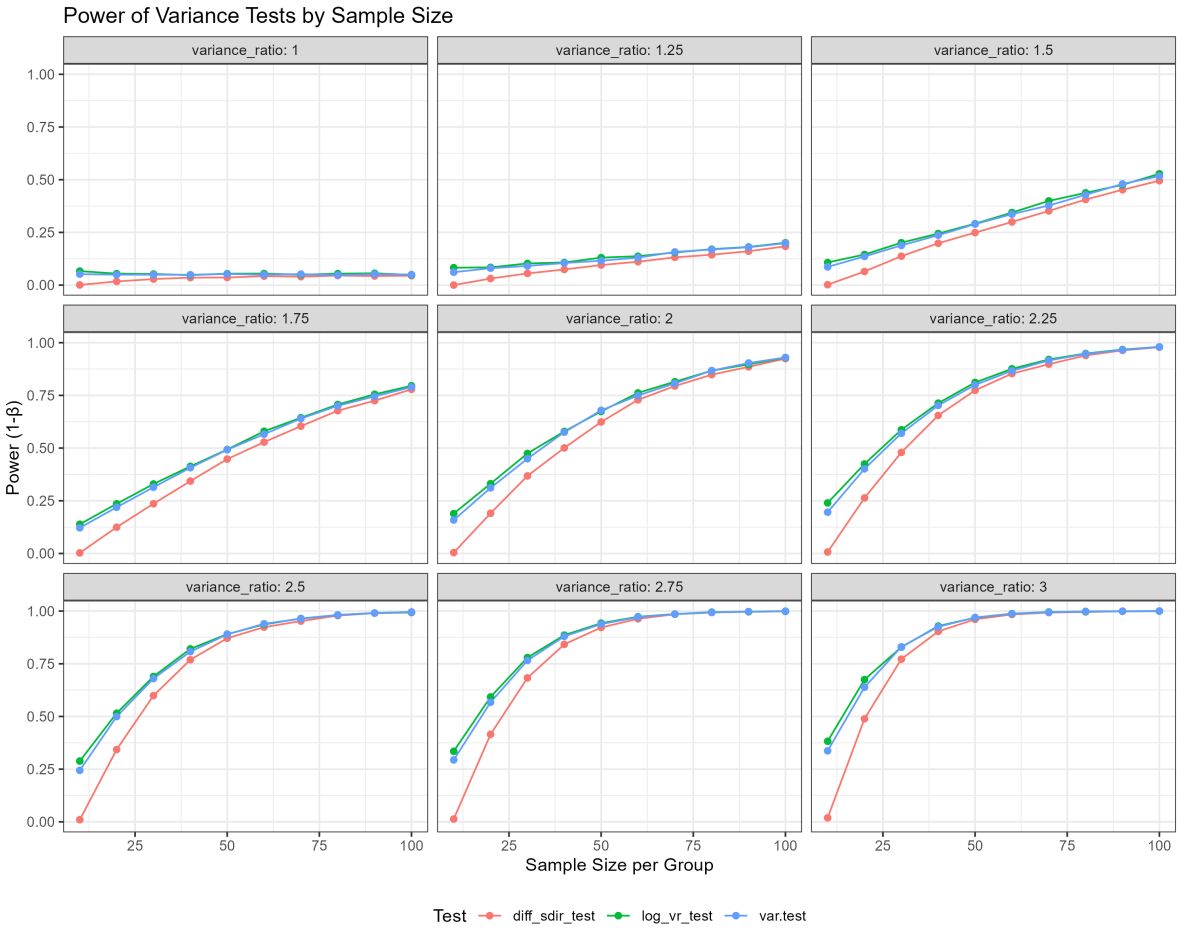
Consideration of Statistical Power

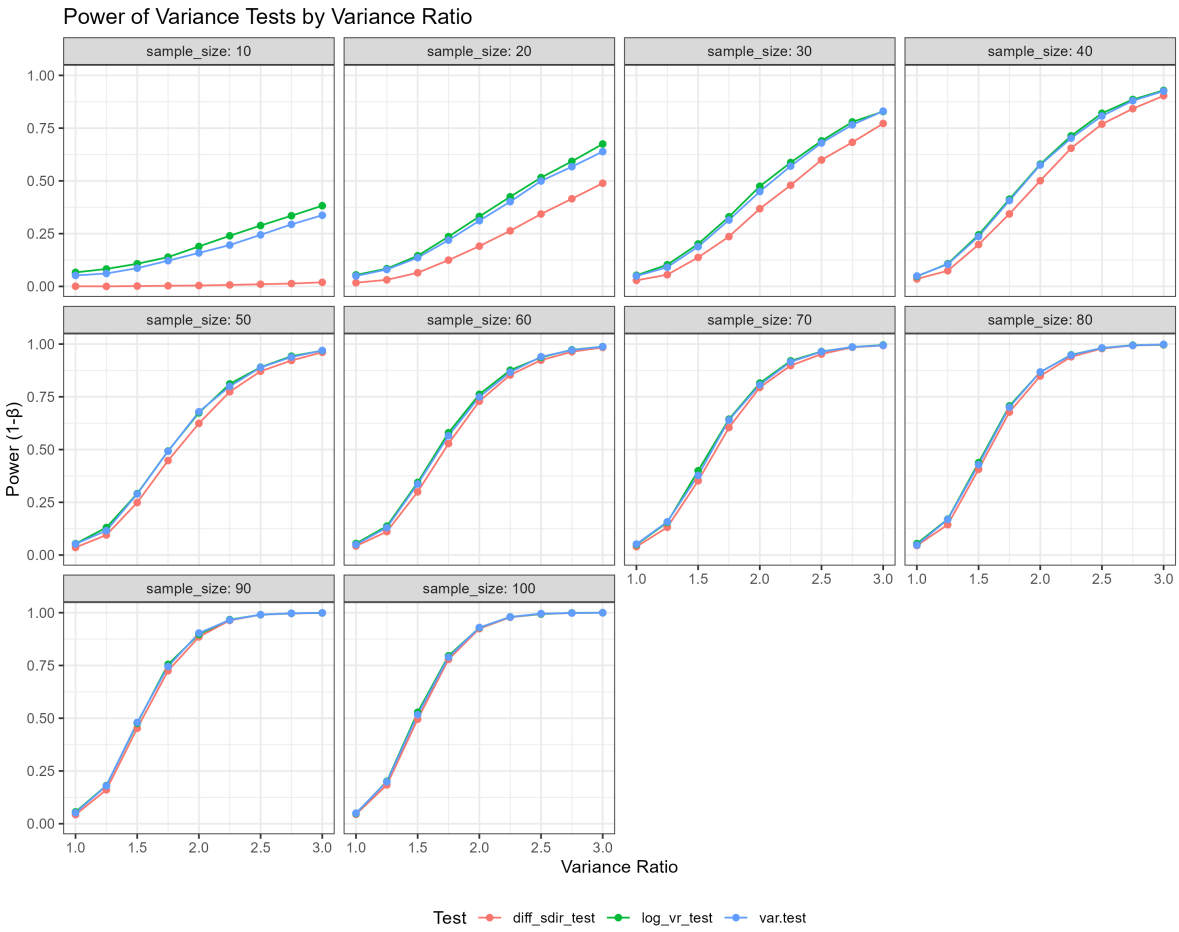
These simulations compares the statistical power of three different approaches for detecting treatment response heterogeneity:

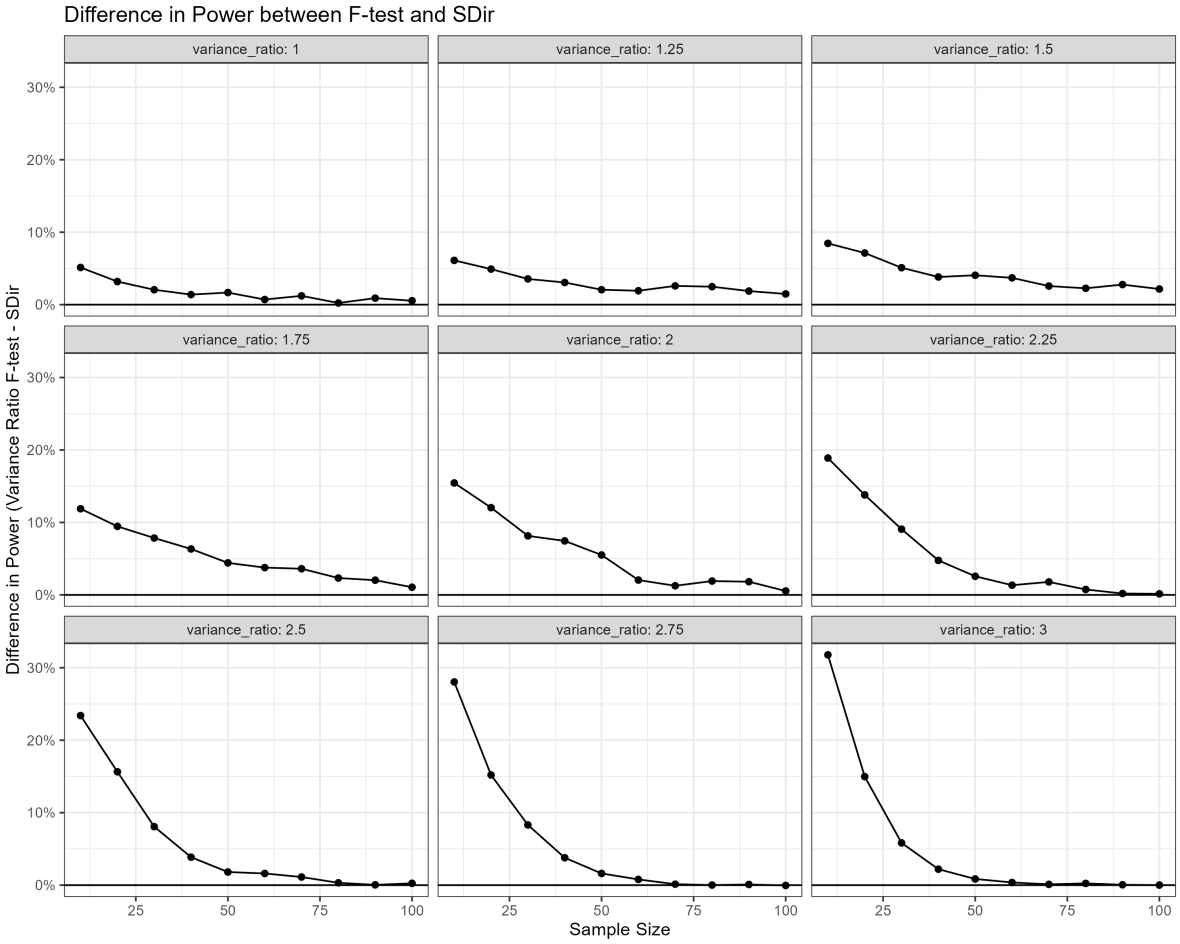
- `diff_sdir_test`: Standard deviation of individual responses (SD_{ir}) test
- `log_vr_test`: Log variance ratio test
- `var.test`: Traditional F-test for comparing variances (variance ratio test)

The simulation explored various sample sizes (10-100 per group) and variance ratios (1.0-3.0) to determine the conditions under which each test performs optimally.

Visualizations of Simulation Results




Power of Variance Tests by Variance Ratio
sample_size: 10
sample_size: 20
sample_size: 30
sample_size: 40
sample_size: 50
sample_size: 60
sample_size: 70
sample_size: 80
sample_size: 90
sample_size: 100
Power (1-β)
Variance Ratio
Test
diff_sdir_test
log_vr_test
var.test



Key Findings

Influence of Sample Size

- Small samples (n $<$ 30): All tests show limited power, with the traditional F-test (var.test) and log variance ratio test performing slightly better than the SDir approach
- Medium samples (n = 30-60): Power increases substantially for all tests as sample size increases
- Large samples (n $>$ 60): All three tests converge in performance, particularly at higher variance ratios

Influence of Variance Ratio

- Small variance ratios ($<$ 1.5): All tests have limited power regardless of sample size

- Moderate variance ratios (1.5-2.0): Power increases significantly as sample size increases
- Large variance ratios (> 2.0): High power achieved with moderate to large sample sizes for all tests

Differences Between Tests

- The traditional F-test (var.test) generally shows marginally higher power than both the SDir test and log variance ratio test
- The power advantage of the F-test over the SDir test is most pronounced with smaller sample sizes and larger variance ratios
- As sample size increases, the performance gap between tests narrows substantially
- At sample sizes of 80-100, all three tests perform nearly identically for variance ratios above 2.0

Practical Implications

1. Sample size considerations: For reliable detection of treatment response heterogeneity, sample sizes of at least 30-40 per group are recommended, with 60+ per group being ideal.
2. Variance ratio detection thresholds:

- Variance ratios of 1.5 or lower are difficult to detect reliably unless sample sizes are large
- Variance ratios of 2.0 or higher can be detected with good power (>80%) with sample sizes of 50+ per group

3. Test selection: While the traditional F-test shows slight advantages in some scenarios, its power disadvantage diminishes with increasing sample size.
4. Planning studies: Researchers interested in treatment response heterogeneity should consider these power analyses when planning study sample sizes, especially if moderate heterogeneity effects are anticipated. Studies intending to detect TRH *should be substantially larger than the average exercise physiology study.*

Atkinson, Greg, and Alan M. Batterham. 2015. "True and False Interindividual Differences in the Physiological Response to an Intervention." *Experimental Physiology* 100 (6): 577–88. https://doi.org/10.1113/ep085070.

Hopkins, Will G. 2015. "Individual Responses Made Easy." *Journal of Applied Physiology* 118 (12): 1444–46. https://doi.org/10.1152/japplphysiol.00098.2015.