Farewell to Power Analysis: The Celebrated Kin to the Infamous Statistical Significance

Electronic Supplementary Material

Shinichi Nakagawa, Malgorzata Lagisz, Yefeng Yang & Szymon Drobniak

27 February 2022

Contents

Setups **2**

Aims of this Supporting Information **2**

Preambles **2**
Figure S1 . . . 4

(1) Estimating sample sizes in the 'fictitious' experiment with large effects **4**
Figure S2 . . . 6
Presumed effect sizes (large) . . . 6
Sample size calculation for diet effects . . . 6
Sample size calculation for sex difference (interaction) . . . 7
Figure S3 . . . 8

(2) Estimating sample sizes in the 'ficticious' experiment with realstic (small) effects **9**
Figure S4 . . . 10
Presumed effect sizes (small) . . . 10
Sample size calculation for diet effects . . . 10
Sample size calculation for sex difference (interaction) . . . 11

(3) Corrleated samples and statistical power **11**

R Session Information **12**

References **13**

```
################################################################################
# Supplementary Material for Farewell to Power Analysis
# Code written by:
#         Yefeng Yang (yefeng.yang1@unsw.edu.au); School of Biological,
#                                     Earth and Environmental Sciences,
#                     University of New South Wales, Sydney, Australia
#         Shinichi Nakagawa (s.nakagawa@unsw.edu.au); School of Biological,
#                                     Earth and Environmental Sciences,
#                     University of New South Wales, Sydney, Australia
################################################################################
```

Setups

Loading packages and custom functions. If your computer do not have the required packages, please install them via `install.packages("package.name")`

```
# load required packages
pacman::p_load(dplyr, magrittr, tidyr, stringr, ggplot2, cowplot,
    patchwork, tidyverse, here, readxl, retrodesign, pwr, Superpower,
    pander)

# custom function for approximate sample size for main
# effect and interactive effect
short_cut <- function(d, method = c("normal", "interaction")) {
    method <- match.arg(method)
    if (method == "normal") {
        size <- 16 * (1/d^2)
    } else {
        size <- 32 * (1/d^2)
    }
    size
}
```

Aims of this Supporting Information

In this document, we show how we got sample sizes presented in the main text in two scenarios under: 1) with relatively large effect sizes (1) and with small but realistic effect sizes (2). In addition, we provide an example of how correlated samples can increase statistical power (3).

Preambles

Statistical power are determined by the following three parameters:

(1) Type I error probability, α, also known as significance threshold, which is usually fixed at 0.05 (see Table I);

(2) sample size, n, that is the number of subjects required for an experiment

(3) standardized effect size, $E[\theta]/\sqrt{Var[\theta]}$, where θ is the effect size of interest, which is indicated by the real difference between two groups (in our case: obesogenic diet vs. control diet), $E[\theta]$ is the population average/expectation, and $Var[\theta]$ is the respective variance; note that standardized mean difference d is an example of a standardized effect size (for more on effect size, see also Fig 2 and Box 2).

```
# drawing Fig 1 Parameter 1: alpha level vs. power

#### set a range of alpha levels (0.01 to 0.1)
alpha_range <- seq(0.001, 1, by = 0.01)

#### calculate power at the set alpha levels using a medium
#### magnitude of standardized effect size 0.5 with a
#### standard deviation of 0.2
power_range <- retro_design(A = 0.5, s = 0.2, alpha = alpha_range)

#### create a data frame
```

```r
power_vs_alpha <- data.frame(alpha = alpha_range, power = power_range$power)

#### plot
power_vs_alpha_plot <- ggplot(power_vs_alpha) + geom_line(aes(x = alpha,
    y = power), show.legend = F) + scale_y_continuous(breaks = seq(0,
    1, 0.2), limits = c(0, 1)) + geom_hline(yintercept = 0.8,
    colour = "red") + labs(x = "Type 1 error (alpha)", y = "Statistical power",
    title = "(A) alpha level vs. power") + theme_bw()


### Parameter 2: n vs. power

#### set a range of n (2 to 100)
n_range <- seq(2, 100, by = 2)

## calculate power for a two-sample t test
## (two-independent-samples-design) using a medium
## magnitude of standardized effect size 0.5
power_range2 <- pwr.t.test(d = 0.5, n = n_range, sig.level = 0.05,
    type = "two.sample", alternative = "two.sided")

#### create a dataframe
power_vs_n <- data.frame(n = n_range, power = power_range2$power)

#### plot
power_vs_n_plot <- ggplot(power_vs_n) + geom_line(aes(x = n,
    y = power), show.legend = F) + scale_y_continuous(breaks = seq(0,
    1, 0.2), limits = c(0, 1)) + geom_hline(yintercept = 0.8,
    colour = "red") + labs(x = "Sample size (n)", y = "Statistical power",
    title = "(B) n vs. power") + theme_bw()

### Parameter 3: effect size vs. power

#### create a plausible range of standardized effect sizes
es_range <- seq(0.01, 1.01, by = 0.01)

#### calculate power with alpha 0.05
power_range3 <- retrodesign::retro_design(A = es_range, s = 0.2,
    alpha = 0.05)

#### create a dataframe
power_vs_es <- data.frame(es = es_range, power = power_range3$power,
    alpha = rep(c("0.05"), length(es_range)))

#### plot
power_vs_es_plot <- ggplot(power_vs_es) + geom_line(aes(x = es,
    y = power), show.legend = F) + scale_y_continuous(breaks = seq(0,
    1, 0.2), limits = c(0, 1)) + geom_hline(yintercept = 0.8,
    colour = "red") + labs(x = "Effect size (d)", y = "Statistical power",
    title = "(C) effect size vs. power") + theme_bw()


### put all figures together
```

```
power_plot <- power_vs_alpha_plot/power_vs_n_plot/power_vs_es_plot
power_plot
```

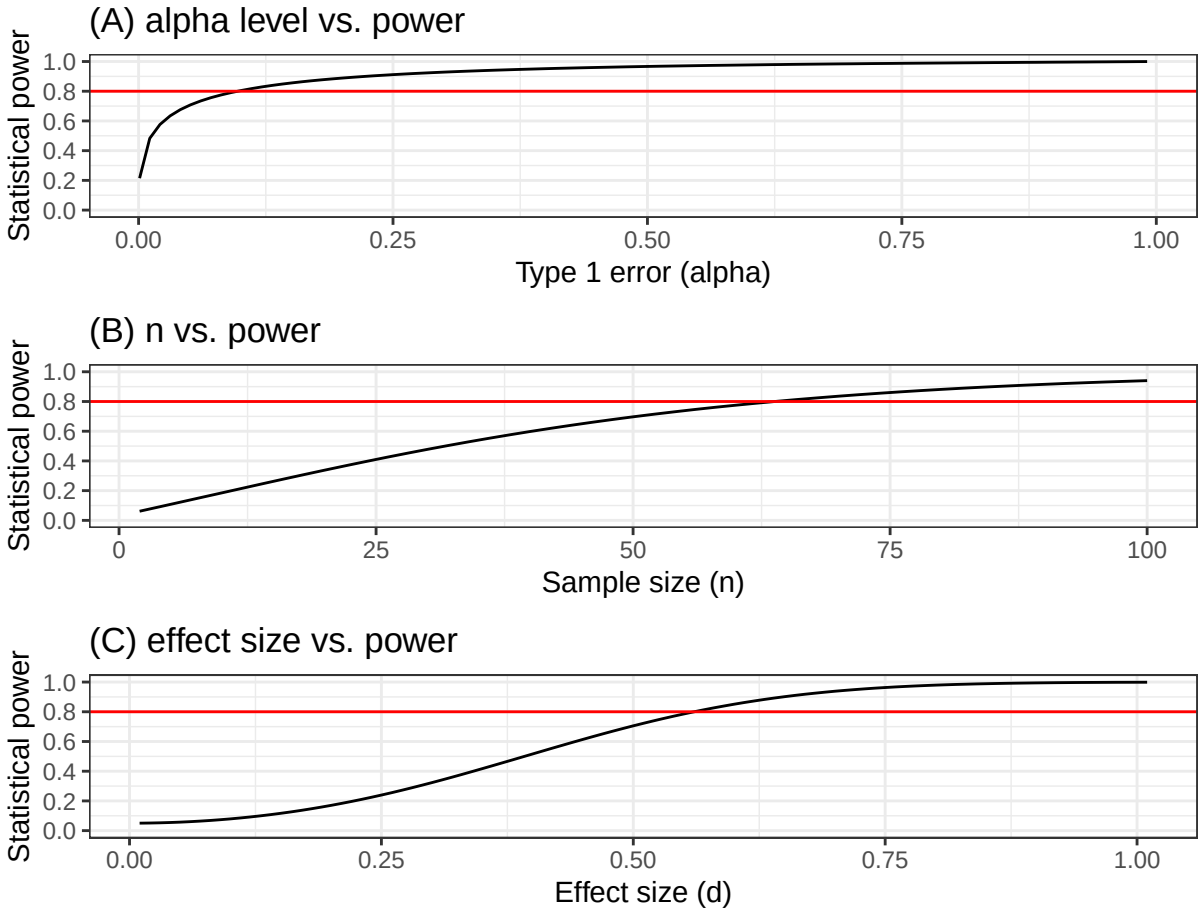

Figure S1

An example showing how the three parameters affect the statistical power: (A) Type I error (α), (B) Sample size (n), (C) Magnitude of the standardized effect size (d). These figures are simulated using `retro_design()` function in `retrodesign` package [1]. See the corresponding code chunk for detailed code.

From Figure S1A we can see that when an experiment commits a higher Type 1 error (which we do not want), it is easier to achieve a desired statistical power (i.e., Cohen's recommendation: 80% power). Increasing sample size (n) and magnitude of standardized effect size are effective ways to increase the statistical power of a given experiment (Figure S1B and S1C).

(1) Estimating sample sizes in the 'fictitious' experiment with large effects

When designing an 'fictitious' diet experiment in your grant proposal, You choose a common significance threshold, $\alpha = 0.05$ and the nominal power level of 80%. Based on your pilot or external information (e.g., relevant studies or a meta-analysis on maternal effect), you assume maternal obesogenic diet will lead to a 30% increase in the number of mistakes in a memory task in males, 20% in females and 10% difference

between the sexes. To quantify the diet effect using a standardized effect size (i.e., d). We assumed the followings:

control group (both male and female) - mean = 100 (arbitrary unit) and standard deviation (sd) = 30;

male treatment group - mean = 130 and sd = 30;

female treatment group - mean = 120 and sd = 30.

Note that we assume the homogeneity of variances among groups (i.e. sd = 30).

```
## scenario 1: large effects

### set up an independent design
design <- ANOVA_design(
  design = "2b*2b", # independent design, which means no correlation
  n =290, # the sample size in each group for testing sex difference
  mu = c(130, 120, 100, 100),
  sd = 30,
  labelnames = c("diet", "obesogenic ", "control ", "sex", "male", "female"),
  plot = FALSE)

meanplot_largeES <- design$meansplot + labs(x = "Groups", y = "Mean")
Figure_S2 <- meanplot_largeES
Figure_S2
```

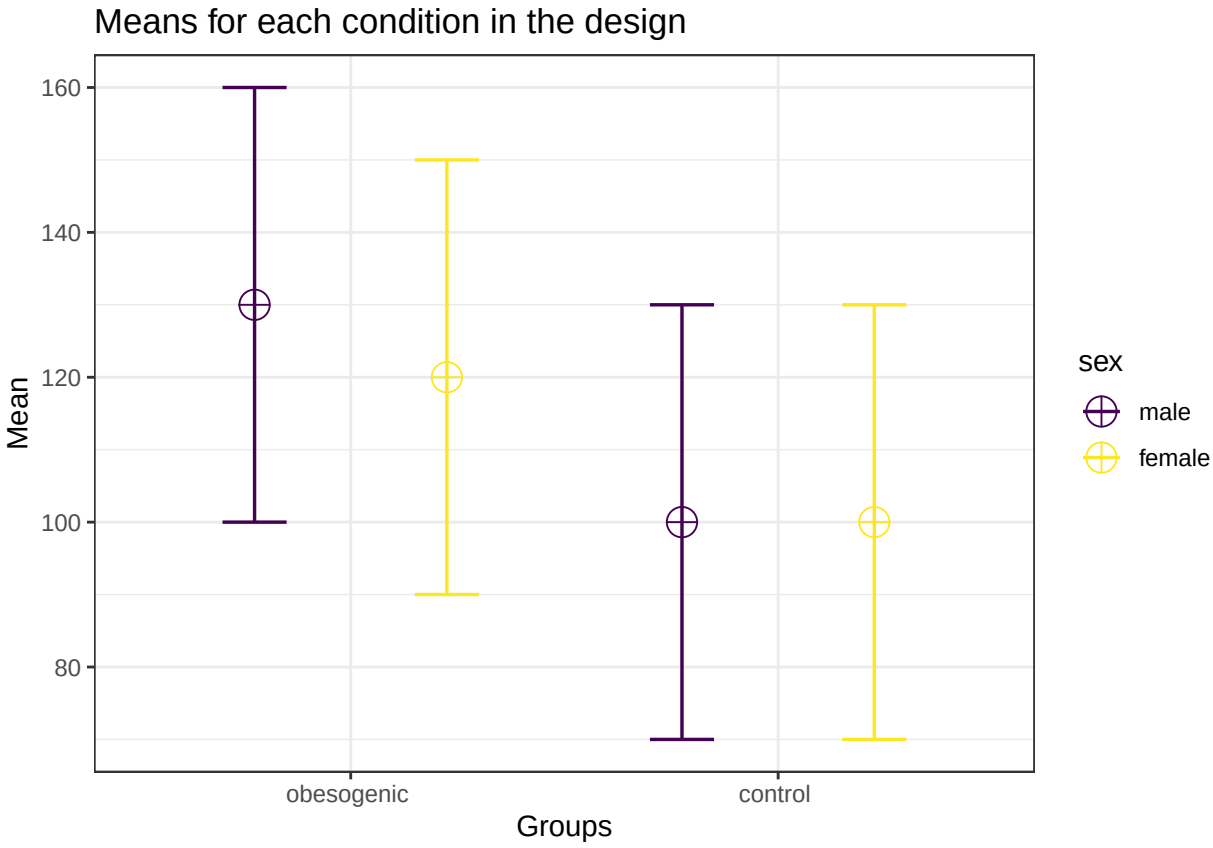

Figure S2

Visualization of the assumed means and standard deviation (sd) of each group, using the package `Superpower`. Error bars represent sd. Here we assume each group is independent (note that this is not quite true if we take one male and one female from one mother but for convenience, let's assume this)

Presumed effect sizes (large)

Using these means and sds, we have the following standardized mean difference d, corresponding % differences:

(1) 1.0, corresponding to a 30% increase in the number of mistakes in a memory task in males after a diet intervention;

(2) 0.67, corresponding to a 20% increase in the number of mistakes in a memory task in females after a diet intervention;

(3) 0.33, corresponding to a 10% sex difference or interaction between diet and sex.

You are planning to use a typical two-sample t-test to examine the statistical significance of the diet effect in males and females. Then you can approximate sample size required for each group using the following formula [2]:

$$n = 16\frac{Var[\theta]}{E[\theta]^2} = 16\frac{1}{d^2}$$

However, for the last effect (interaction effect), this requires comparing 4 groups so that this formula does not work. Yet you could still use this formula by replacing 16 with 32 as interaction involves four groups rather than two.

Sample size calculation for diet effects

In the following section, we use our custom function (based on the above formula) and one existing R package `pwr` to estimate the sample size used in your proposed experiment with different scenarios (sample size mentioned in the fictitious story in the main text).

```
# Cohen's d male
(130 - 100)/30
```

```
## [1] 1
```

```
# Cohen's d female
(120 - 100)/30
```

```
## [1] 0.6666667
```

```
# the first set (surprising large effects)

## male treatment effect
pwr.t.test(d = 1, sig.level = 0.05, power = 0.8, type = "two.sample",
    alternative = "two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 16.71472
##               d = 1
##       sig.level = 0.05
```

```
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
# our result from our custom function is very close
pwr_independent_m_d1 <- short_cut(d = 1, method = "normal")

## female treatment effect
pwr.t.test(d = 0.67, sig.level = 0.05, power = 0.8, type = "two.sample",
    alternative = "two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 35.95537
##               d = 0.67
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
pwr_independent_f_d0.67 <- short_cut(d = 0.67, method = "normal")
```

Sample size calculation for sex difference (interaction)

We cannot use `pwr.t.test` for getting sample size for the sex difference. So first we can use our formula and then, we use the package `Superpower` to obtain a simulation based sample size.

```
# Cohen's d sex difference (interaction)
((130 - 100) - (120 - 100))/30
```

```
## [1] 0.3333333
```

```
pwr_independent_i_d0.33 <- short_cut(d = 0.33, method = "interaction")
pwr_independent_i_d0.33
```

```
## [1] 293.8476
```

Main results of the outputs of the `ANOVA_exact` when detecting large effects:

```
## scenario 1: large effects perform ANOVA and calculate
## power for in dependent design
power_results <- ANOVA_exact(design, alpha_level = 0.05, verbose = FALSE)

power_results$main_results
```

```
##              power partial_eta_squared    cohen_f non_centrality
## diet     100.00000          0.14836492 0.41738692     201.388889
## sex       80.94609          0.00692025 0.08347738       8.055556
## diet:sex  80.94609          0.00692025 0.08347738       8.055556
```

Simulation (using the `ANOVA_exact` function) shows that collecting data from $n = 294$ F1 mice (per group) has 81% power for the interaction or sex difference (see code chunk for R syntax).

We also plot a power curve over a range of sample sizes (Figure S4), from which you can visually explore whether the expected power is achieved for the interaction (bottom panel), and if so, at which sample size.

```
plot.power <- plot_power(design, min_n = 5, max_n = 300, desired_power = 80,
    plot = FALSE)
```

```
## Achieved Power and Sample Size for ANOVA-level effects
##   variable                  label   n achieved_power desired_power
## 1     diet Desired Power Achieved  12          80.60            80
## 2      sex Desired Power Achieved 284          80.13            80
## 3 diet:sex Desired Power Achieved 284          80.13            80
```

```
plot.power$plot_ANOVA + labs(x = "Sample size per group")
```

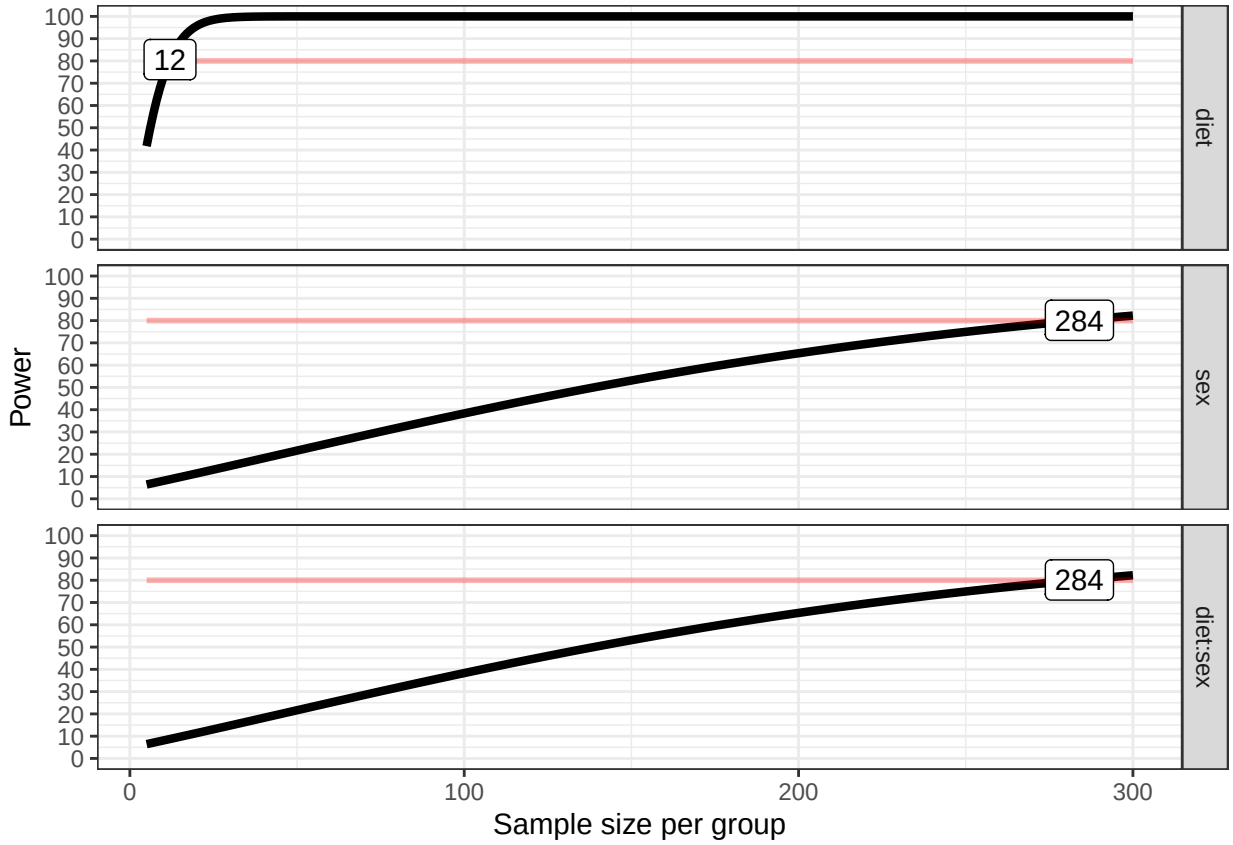

Figure S3

Power curves for large inter-generational effect in a dependent design (diet and sex are manipulated between animals). Top panel = the main effect - diet; Middle panel = the main effect - sex; Bottom panel = the interactive effect, sex difference. The orange horizontal lines denote the expected statistical power (80%). Note that the main diet effect is an average effect over the two sexes.

As you see, the simulation-based method suggests we need 284 subjects to reach 80% to detect the interaction effect, which confirm what we obtained form the formula was close enough.

(2) Estimating sample sizes in the 'ficticious' experiment with realstic (small) effects

This time, we assume maternal obesogenic diet will have more realistic effects on pups' memory: a 5% increase in males, 3% in females and thus 1% difference in the diet effect between the sexes (interaction).

To calculate d, you assume (Figure S8):

control group (both male and female) - mean = 100 (arbitrary unit) and sd = 30;

male treatment group - mean = 105 and sd = 30;

female treatment group - mean = 103 and sd = 30.

```
## scenario 2: small effects
### set up an independent design
design2 <- ANOVA_design(
  design = "2b*2b", # independent design
  n =7128, # the sample size used for testing sex difference
  mu = c(105, 103, 100, 100),
  sd = 30,
  labelnames = c("diet", "obesogenic ", "control ", "sex", "male", "female"),
  plot = FALSE)

meanplot_smallES <- design2$meansplot +
  labs(x = "Groups", y = "Mean", title = "Realistic small effect")
meanplot_smallES
```

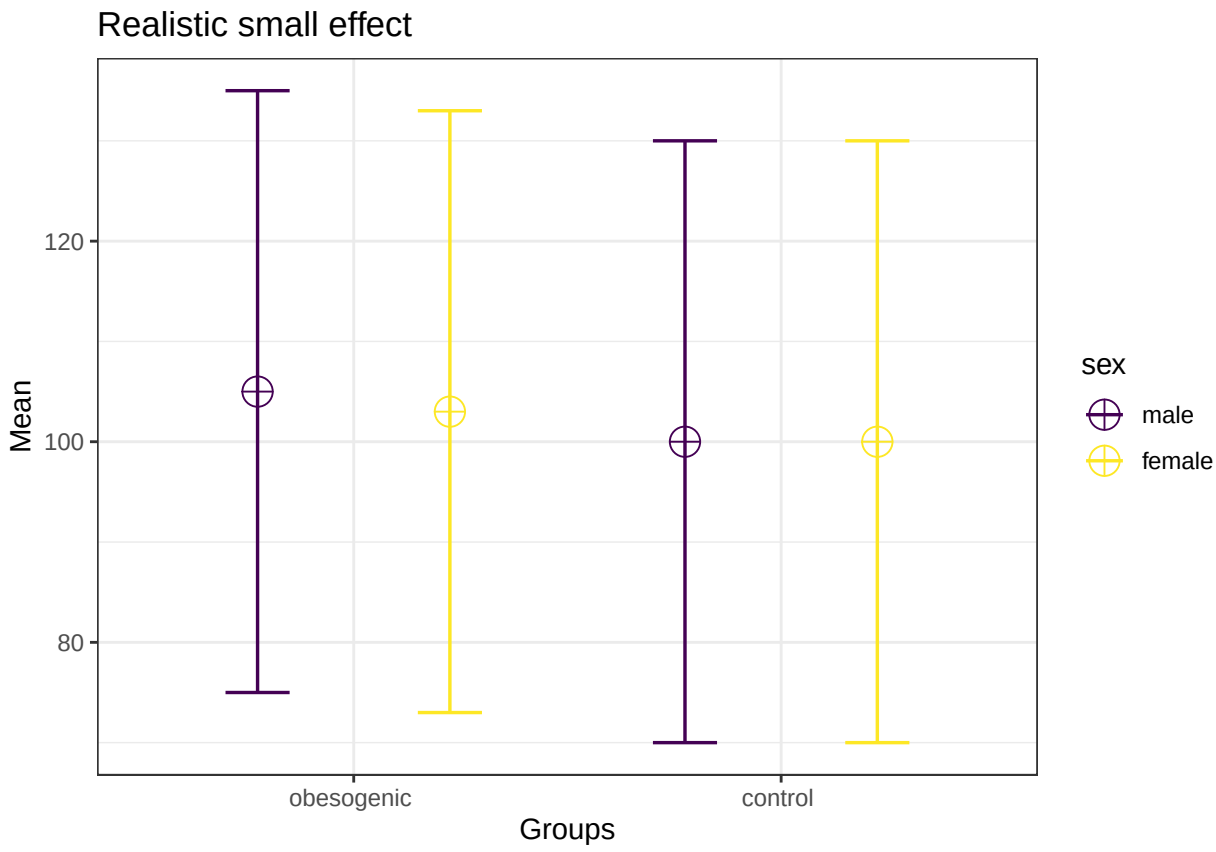

Figure S4

Visualization of the expected means and standard deviation (sd) of each group using the package `Superpower` under more realistic scenarios. Error bars represent sd. Here we assume each group is independent (note that this is not quite true if we take one male and one female from one mother but for convenience, let's assume this).

Presumed effect sizes (small)

As with the above, we assumed all groups share a common sd = 30 (population standard deviation). Then you can obtain the following standardized mean difference d:

(1) 0.16, corresponding to 5% increase in the number of mistakes in a memory task in males after a diet intervention;

(2) 0.1, corresponding to 3% increase in the number of mistakes in a memory task in females after a diet intervention;

(3) 0.06, corresponding to 2% sex difference or interaction between diet and sex.

Sample size calculation for diet effects

Following similar procedures in estimating sample sizes for large effects (see above), you can obtain sample sizes with these new presumed effect sizes

```
# Cohen's d male
(105 - 100)/30
```

```
## [1] 0.1666667
```

```
# Cohen's d female
(103 - 100)/30
```

```
## [1] 0.1
```

```
# the first set (realistic small effects)

## male treatment effect
pwr.t.test(d = 0.167, sig.level = 0.05, power = 0.8, type = "two.sample",
    alternative = "two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 563.8262
##               d = 0.167
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
pwr_independent_m_d0.167 <- short_cut(d = 0.167, method = "normal")

## female treatment effect
pwr.t.test(d = 0.1, sig.level = 0.05, power = 0.8, type = "two.sample",
    alternative = "two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 1570.733
##               d = 0.1
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
pwr_independent_f_d0.1 <- short_cut(d = 0.1, method = "normal")
```

Sample size calculation for sex difference (interaction)

We can also use our formula to estimate (assuming independence of all groups)

```
# Cohen's d sex difference (interaction)
((105 - 100) - (102 - 100))/30
```

```
## [1] 0.1
```

```
# sex difference
pwr_independent_i_d0.067 <- short_cut(d = 0.067, method = "interaction")
pwr_independent_i_d0.067
```

```
## [1] 7128.536
```

We use a similar simulation-based approach to empirically calculate power for the interaction effect using `Superpower`. Main results of the outputs of the `ANOVA_exact` when assuming small effects:

```
## scenario 2: small effects perform ANOVA and calculate
## power for in dependent design
power_results2 <- ANOVA_exact(design2, alpha_level = 0.05, verbose = FALSE)
power_results2$main_results
```

```
##              power partial_eta_squared    cohen_f non_centrality
## diet     100.00000        0.0044253969 0.06667134         126.72
## sex       80.35012        0.0002777396 0.01666784           7.92
## diet:sex  80.35012        0.0002777396 0.01666784           7.92
```

Simulations (using the ANOVA_exact function) show that collecting data from $n = 7129$ F1 mice (per group) has 80.35% power for the interaction or sex difference. So the simulation result seems to catch with the sample size estimated by the formula.

(3) Corrleated samples and statistical power

As mentioned, correlated samples can increase the statistical power of an experiment so that we require fewer samples. Here, we assume that sibling traits are correlated ($r = 0.5$) regardless of sex. We find $n = 3564$ can reach the expected statistical power (80%) for interaction (i.e., sex difference). This number (3567) corresponds to

```
## scenario 1: small effects
### perform ANOVA and calculate power for in independent design
### assuming the siblings are very similar to each other - r = 0.5
design3 <- ANOVA_design(
```

```
  design = "2w*2w", # dependent design
  n =3564,
  r = 0.5,
  mu = c(105, 103, 100, 100),
  sd = 30,
  labelnames = c("diet", "obesogenic", "control", "sex", "male", "female"),
  plot = FALSE)

power_results3 <- ANOVA_exact(design3, alpha_level = 0.05, verbose = FALSE)
power_results3$main_results
```

```
##              power partial_eta_squared    cohen_f non_centrality
## diet     100.00000         0.034344069 0.18858827         126.72
## sex       80.33173         0.002217916 0.04714707           7.92
## diet:sex  80.33173         0.002217916 0.04714707           7.92
```

This number (3567) corresponds to the value calculated from the following formula:

$$n_{interaction} = \frac{32}{d^2}(1 - r)$$

Using this formula, we can assume a lower correlation ($r = 0.25$) and then, we find $n = 5346$. As before, we can verify this, using `ANOVA_design`:

```
## scenario 2: small effects
### perform ANOVA and calculate power for in independent design
# assuming the siblings are very similar to each other - r = 0.25
design4 <- ANOVA_design(
  design = "2w*2w",   # dependent design
  n =5346,
  r = 0.25,
  mu = c(105, 103, 100, 100),
  sd = 30,
  labelnames = c("diet", "obesogenic", "control", "sex", "male", "female"),
  plot = FALSE)

power_results4 <- ANOVA_exact(design4, alpha_level = 0.05, verbose = FALSE)
power_results4$main_results
```

```
##              power partial_eta_squared    cohen_f non_centrality
## diet     100.00000         0.023159080 0.15397447         126.72
## sex       80.33874         0.001479566 0.03849362           7.92
## diet:sex  80.33874         0.001479566 0.03849362           7.92
```

As you see, with $n = 5346$, we have ~80% power. We note that, as mention in the text, for more complex designs (e.g. including different strains of mice), we cannot use the formula or the functions form `Superpower`. We need to use other software packages which could accommodate such design features.

R Session Information

```
sessionInfo() %>%
    pander()
```

R version 4.1.2 (2021-11-01)

Platform: x86_64-apple-darwin17.0 (64-bit)

locale: en_AU.UTF-8||en_AU.UTF-8||en_AU.UTF-8||C||en_AU.UTF-8||en_AU.UTF-8

attached base packages: *stats*, *graphics*, *grDevices*, *utils*, *datasets*, *methods* and *base*

other attached packages: *pander(v.0.6.4)*, *Superpower(v.0.1.2)*, *pwr(v.1.3-0)*, *retrodesign(v.0.1.0)*, *readxl(v.1.3.1)*, *here(v.1.0.1)*, *forcats(v.0.5.1)*, *purrr(v.0.3.4)*, *readr(v.2.1.2)*, *tibble(v.3.1.6)*, *tidyverse(v.1.3.1)*, *patchwork(v.1.1.1)*, *cowplot(v.1.1.1)*, *ggplot2(v.3.3.5)*, *stringr(v.1.4.0)*, *tidyr(v.1.2.0)*, *magrittr(v.2.0.2)* and *dplyr(v.1.0.8)*

loaded via a namespace (and not attached): *TH.data(v.1.1-0)*, *minqa(v.1.2.4)*, *colorspace(v.2.0-2)*, *ellipsis(v.0.3.2)*, *rprojroot(v.2.0.2)*, *estimability(v.1.3)*, *htmlTable(v.2.4.0)*, *base64enc(v.0.1-3)*, *fs(v.1.5.2)*, *rstudioapi(v.0.13)*, *farver(v.2.1.0)*, *fansi(v.1.0.2)*, *mvtnorm(v.1.1-3)*, *lubridate(v.1.8.0)*, *xml2(v.1.3.3)*, *codetools(v.0.2-18)*, *splines(v.4.1.2)*, *knitr(v.1.37)*, *afex(v.1.0-1)*, *Formula(v.1.2-4)*, *jsonlite(v.1.7.3)*, *nloptr(v.2.0.0)*, *broom(v.0.7.12)*, *cluster(v.2.1.2)*, *dbplyr(v.2.1.1)*, *png(v.0.1-7)*, *compiler(v.4.1.2)*, *httr(v.1.4.2)*, *emmeans(v.1.7.2-9000003)*, *backports(v.1.4.1)*, *assertthat(v.0.2.1)*, *Matrix(v.1.4-0)*, *fastmap(v.1.1.0)*, *cli(v.3.2.0)*, *formatR(v.1.11)*, *htmltools(v.0.5.2)*, *tools(v.4.1.2)*, *lmerTest(v.3.1-3)*, *coda(v.0.19-4)*, *gtable(v.0.3.0)*, *glue(v.1.6.1)*, *reshape2(v.1.4.4)*, *Rcpp(v.1.0.8)*, *carData(v.3.0-4)*, *cellranger(v.1.1.0)*, *vctrs(v.0.3.8)*, *nlme(v.3.1-155)*, *xfun(v.0.29)*, *lme4(v.1.1-28)*, *rvest(v.1.0.2)*, *lifecycle(v.1.0.1)*, *pacman(v.0.5.1)*, *MASS(v.7.3-54)*, *zoo(v.1.8-9)*, *scales(v.1.1.1)*, *hms(v.1.1.1)*, *parallel(v.4.1.2)*, *sandwich(v.3.0-1)*, *RColorBrewer(v.1.1-2)*, *yaml(v.2.2.2)*, *gridExtra(v.2.3)*, *rpart(v.4.1.16)*, *latticeExtra(v.0.6-29)*, *stringi(v.1.7.6)*, *highr(v.0.9)*, *checkmate(v.2.0.0)*, *boot(v.1.3-28)*, *rlang(v.1.0.1)*, *pkgconfig(v.2.0.3)*, *evaluate(v.0.15)*, *lattice(v.0.20-45)*, *labeling(v.0.4.2)*, *htmlwidgets(v.1.5.4)*, *tidyselect(v.1.1.1)*, *plyr(v.1.8.6)*, *R6(v.2.5.1)*, *generics(v.0.1.2)*, *Hmisc(v.4.6-0)*, *multcomp(v.1.4-18)*, *DBI(v.1.1.2)*, *pillar(v.1.7.0)*, *haven(v.2.4.3)*, *foreign(v.0.8-82)*, *withr(v.2.4.3)*, *survival(v.3.2-13)*, *abind(v.1.4-5)*, *nnet(v.7.3-17)*, *modelr(v.0.1.8)*, *crayon(v.1.5.0)*, *car(v.3.0-12)*, *utf8(v.1.2.2)*, *tzdb(v.0.2.0)*, *rmarkdown(v.2.11)*, *jpeg(v.0.1-9)*, *grid(v.4.1.2)*, *data.table(v.1.14.2)*, *reprex(v.2.0.1)*, *digest(v.0.6.29)*, *xtable(v.1.8-4)*, *numDeriv(v.2016.8-1.1)*, *munsell(v.0.5.0)* and *viridisLite(v.0.4.0)*

References

1. Gelman A, Carlin J. Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. Perspectives on Psychological Science. 2014;9: 641–651.
2. Lehr R. Sixteen s-squared over d-squared: A relation for crude sample size estimates. Statistics in medicine. 1992;11: 1099–1102.