Supplementary file 2: Sample size calculation / power analysis

Introduction

This R Markdown file documents our Sample size calculation / power analysis.

We need to estimate how many observations we need to develop a casemix system.

We use the published results of *Parsons et al* (see reference below) for our power analysis.

Table 4 of *Parsons et al* contains the group means, standard errors and sample sizes for their complex needs case-mix groups. From the two case-mix algorithms in their paper, the complex needs clients are thought to be most similar to the home care clients in the Netherlands.

Approach

We take a coarse approach towards power analysis, as we are developing case-mix groups that do not yet exist. Our approach is to assume a scientifically reasonable effect size (see Gelman, reference below) and use this for our power calculation.

For a case-mix classification system, this requires information on:

- differences in mean hours between case-mix groups,
- the variation (Standard deviation) in hours within case-mix groups and
- the relative distribution of clients over the various case-mix groups.

By focussing our power analysis on the most difficult case-mix groups regarding these three dimensions, we simplify the analysis considerably. We assume normal distributions within case-mix groups.

Assumed effect size

We look for differences between case-mix groups where the case-mix groups have a low share of clients, where the difference in means is small, and the variation within the groups is large. In Table 4 of *Parsons et al* we identify two case-mix group pairings that we expect will determine the lower bound on the required sample size:

Cluster 2 and cluster 5 both have a low share of clients (5% and 6% of all clients for cluster 2 and cluster 5 respectively) and a small (but meaningful) difference in mean hours (2.4 hours/week vs 3.0 hours per week), with standard deviations of 1.2 and 1.4 respectively.

Cluster 7 and cluster 8 both have high standard deviations (3.4 and 4.8 for cluster 7 and cluster 8 respectively) but a larger difference in mean hours (4.4 hours/week vs 6.0 hours per week). Cluster 8 has the lowest share of clients (4%), whereas cluster 7 has a relatively large share (17%).

```r
n_total <- 1007
# cluster parameters 2 en 5
group1_mean <- 2.35
group1_size <- 51
group1_se <- 0.17

group2_mean <- 2.98
group2_size <- 60
group2_se <- 0.18
```


Elissen AMJ, *et al. BMJ Open* 2020; 10:e035683. doi: 10.1136/bmjopen-2019-035683

```r
n_total <- 1007
# cluster parameters 7 en 8
group1_mean <- 4.42
group1_size <- 171
group1_se <- 0.26

group2_mean <- 6.07
group2_size <- 44
group2_se <- 0.73
```

It turns out that the power analysis gives very similar results for both pairs. We present our power analysis with the pair that requires the biggest sample size, clusters 7 and 8.

Assumed effect size

This translates into an assumed effect size (Cohen's *d*) of about 0.4 - 0.5.

```r
# sd
sd1 <- sqrt(group1_size) * group1_se

sd2 <- sqrt(group2_size) * group2_se

# cohen's d (https://en.wikipedia.org/wiki/Effect_size#Cohen's_d)

mean_diff <- group2_mean - group1_mean
pooled_sd_n <- ((group1_size - 1)*(sd1^2)) + ((group2_size - 1)*(sd2^2))

pooled_sd_d <- ((group1_size + group2_size) - 2)

pooled_sd <- sqrt(pooled_sd_n/pooled_sd_d)

cohens_d <- mean_diff / pooled_sd

cohens_d
```

```
## [1] 0.4416199
```

Power analysis

Our power analysis consists of estimating the probability of detecting a statistically significant difference (the power) with the significance threshold set at p = 0.05. We do this by repeatedly drawing a set of clients from two normal distributions (thus allowing for negative values) with the reported means and SD's and calculating the p-value for non-zero difference. Based on our experimental design, we expect a minimum of 1500 observations.

```r
set.seed(123)

required_sample_size <- 1500

# create exp sample

calculate_p_values <- function(n_samples) {
  p_values <- c()
  estimates <- c()
```


Elissen AMJ, *et al. BMJ Open* 2020; 10:e035683. doi: 10.1136/bmjopen-2019-035683

```
  for(i in 1:1000){
    df <- rbind( data.frame(y = rnorm(round((group1_size/n_total) * n_samples, 0),
                                      group1_mean, sd1),
                            x = "group 1"),
             data.frame(y = rnorm(round((group2_size/n_total) * n_samples, 0),
                                  group2_mean, sd2),
                        x = "group 2"))

    lmfit <- lm(y  ~ factor(x), data = df)

    # extract p value from lmfit
    p_values[i] <- coef(summary(lmfit))[2,4]
    # extract estimate from lmfit
    estimates[i] <- coef(summary(lmfit))[2,1]

  }
    return(list(p_values, estimates))
}
res <- calculate_p_values(required_sample_size)
```

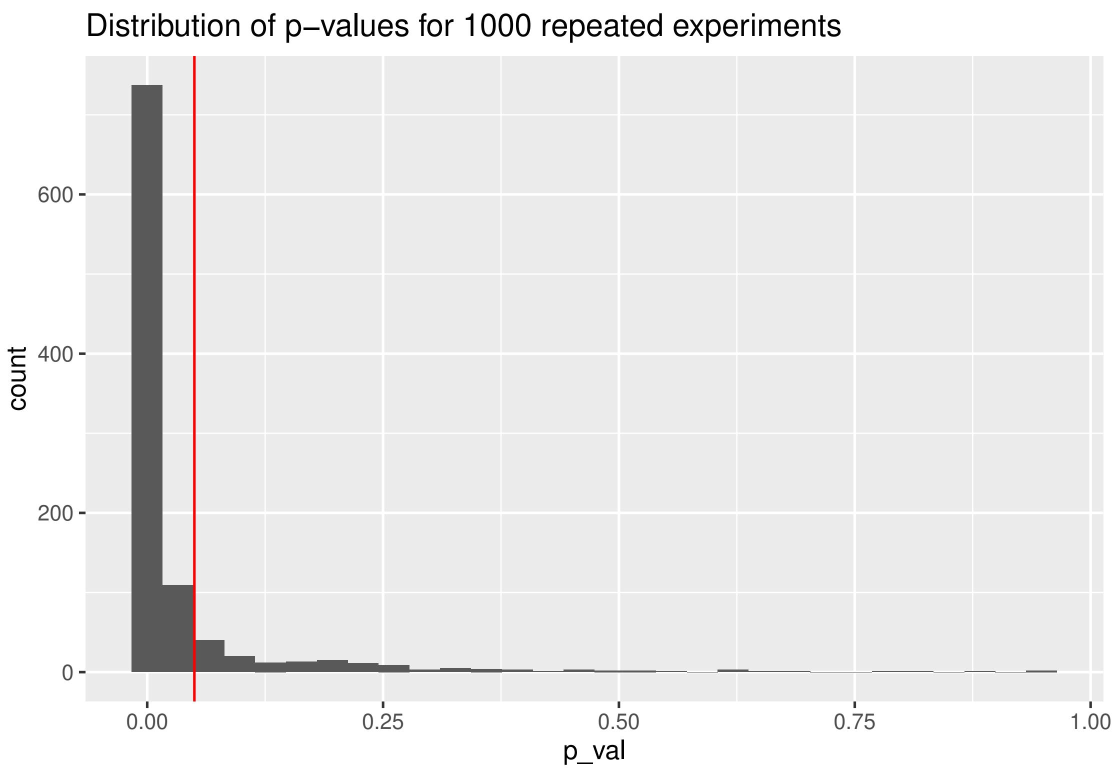
Result

The percentage of experiments where $p < 0.05$ is 84.7 %. The percentage of experiments where the estimated difference has the wrong sign is 0.7 %.

```
library(ggplot2)

ggplot(data.frame(p_val = res[[1]]), aes(x = p_val)) +
  geom_histogram() +
  geom_vline(xintercept = 0.05, col = "red") +
  ggtitle("Distribution of p-values for 1000 repeated experiments")
```


Elissen AMJ, *et al. BMJ Open* 2020; 10:e035683. doi: 10.1136/bmjopen-2019-035683



Thus, we expect that a sample size of 1500 is sufficient to have a high probability of detecting our assumed effect size.

```
# power
mean(res[[1]] <= 0.05)
```

```
## [1] 0.847
```

```
# mean difference between the two casemix groups
mean(res[[2]])
```

```
## [1] 1.680325
```

Verification of results by comparing with online sample size calculator

For clusters 2 and 5, we have verified this simulation using an online sample size calculator (https://clincalc.com/stats/samplesize.aspx) with the following parameters:

- Group means 2.4 and 3.0
- Common standard deviation of 1.3
- Required power 0.85 and therefore beta = 0.15
- Significance level alpha = 0.05

This results in a required sample size of 168 (84 in each group). Given that only 11% of all observations (5% and 6%) are clients in these two case-mix groups, total number of observations needed is 168 * (100/11) = 1527 observations.

Elissen AMJ, *et al. BMJ Open* 2020; 10:e035683. doi: 10.1136/bmjopen-2019-035683

We could not use the online sample size calculator for clusters 7 and 8, since the assumption of a common standard deviation is clearly violated for these case-mix groups.

References

Parsons M, Rouse P, Sajtos L, Harrison J, Parsons J, Gestro L. Developing and utilising a new funding model for home-care services in New Zealand. Health & Social Care in the Community. 2018;26(3):345-55.

Gelman A., https://statmodeling.stat.columbia.edu/2018/09/24/dont-calculate-post-hoc-power- using-observed-estimate-effect-size/

Elissen AMJ, *et al. BMJ Open* 2020; 10:e035683. doi: 10.1136/bmjopen-2019-035683