Power analysis for two-sample parallel design

Parameters	**Notes**		
ℓ	Missing rate, $\ell = \frac{exp\,(a+b\mu)}{1+exp\,(a+b\mu)}$, should be learned from data.		
m	The number of proteins		
α	Type I error rate, often $\alpha = 0.05$.		
β	Type II error rate, often $\beta = 0.2$, and power is $1-\beta$.		
σ_{cs}	Standard deviation for cases	σ_{cl}	Standard deviation for controls
n_{cs}	Actual sample size for cases	n_{cl}	Actual sample size for controls
μ_{cs}	Mean for cases after taking Log_2 scale.	μ_{cl}	Mean for controls after taking Log_2 scale.

Given the empirical data, as summarized at http://rpubs.com/gc5k/proQC
We can reasonably have assumptions below

- we always assume $\sigma_{cs} \approx \sigma_{cl}$ as observed in PPP1 data.
- ℓ is between 0~1, and $\ell \propto \mu$. When μ is close to 13, $\ell \approx 0.5$; As ℓ is upon the expression level, so it is possibly $\ell_{cs} \neq \ell_{cl}$.
- $\sigma_{.}$ is close to 0.75 empirically.
- we are interested in two-tailed test.

Linear model analysis

$$y = a + bx + e$$

in which x is coded 1 and 0 for cases and controls.
Then we have $cov(x,y) = E(xy) - E(x)E(y) = \frac{n_{cs}n_{cl}}{(n_{cs}+n_{cl})^2}(\mu_{cs}-\mu_{cl})$, and $var(x) = \frac{n_{cs}n_{cl}}{(n_{cs}+n_{cl})^2}$. So, $E(b) = \frac{cov(x,y)}{var(x)} = \mu_{cs} - \mu_{cl}$.

The standard error of the regression coefficient is $\sigma_b = \sqrt{\frac{\sigma_y^2 - b^2\sigma_x^2}{(n_{cs}+n_{cl})\sigma_x^2}} = \sqrt{\frac{\sigma_{cl}^2}{(n_{cs}+n_{cl})\sigma_x^2}}$, and the corresponding t-test statistic can be constructed as

$$t = \frac{\mu_{cs}-\mu_{cl}}{\sqrt{\frac{\sigma_{cl}^2}{(n_{cs}+n_{cl})\sigma_x^2}}} = \sqrt{n_{cs}+n_{cl}}\frac{(\mu_{cs}-\mu_{cl})}{\sigma_{cl}}\sigma_x = \frac{(\mu_{cs}-\mu_{cl})}{\sigma_{cl}}\sqrt{\frac{n_{cs}n_{cl}}{n_{cs}+n_{cl}}}$$

After convert it to Chisquare test, its NCP is $\frac{(\mu_{cs}-\mu_{cl})^2}{\sigma_{cl}^2}\frac{n_{cs}n_{cl}}{n_{cs}+n_{cl}} = n\frac{(\mu_{cs}-\mu_{cl})^2}{\sigma_{cl}^2}\frac{r_{cs}r_{cl}}{r_{cs}+r_{cl}}$.

Two-sample parallel design

Assuming $\kappa = \frac{n_{cs}}{n_{cl}}$, the ratio between cases and controls,

The sample size required for n_{cl} can be written as

$$n_{cl} = \frac{\left(z\alpha/_{2} + z_{\beta}\right)^2 \sigma^2}{(u_{cs}-u_{cl})^2}$$

and $n_{cs} = \kappa n_{cs}$

R code

```
m=1000
mu=14.5
mu0=14
sd=0.75
```

```
kappa=1

alpha=0.05/m
beta=0.20
ncl=(1+1/kappa)*(sd*(qnorm(1-alpha/2)+qnorm(1-beta))/(mu-mu0))^2
ceiling(ncl)# 32
NCP=(mu-mu0)^2/sd^2*(ncl*ncl*kappa)/(ncl+kappa*ncl)
Power=length(which(pchisq(rchisq(1000,1,ncp=NCP),1,lower.tail=F) < alpha))
```

Power calculation for ProBatch

The sample size needed is

$$n = \frac{\sigma^2}{(\mu_{cs} - \mu_{cl})^2}\left(z_{\left(1-\frac{\alpha}{2m}\right)} + z_{(1-\beta)}\right)^2$$

Input parameters:

1) m: the number of proteins observed
2) μ_{cs}, μ_{cl}: the mean of protein expression
3) $z_{\left(1-\frac{\alpha}{2m}\right)}$: the cutoff value for z-score
4) $z_{(1-\beta)}$: the cutoff value for z-score

In real data analysis, m can be estimated from a small-scale pilot experiment. Assuming $m = 1000$, $\sigma \approx 0.75$. $z_{\left(1-\frac{\alpha}{2m}\right)} = 4.06$ and $z_{(1-\beta)} = 0.86$.

For a particular protein, say $\mu_{cs} = 14$, and $\mu_{cl} = 14.5$. then $n \approx 220$, so it needs 220 cases and 220 controls.

Missing data leads to reduced power

Due to missing data, in particular for proteins of expression level merely above the threshold, missing value is extremely high, the statistical power is jeopardized