Sample Size Estimation

FGCZ Protein Informatics Training

Witold Wolski wew@fgcz.ethz.ch
01 March, 2021

Sample size estimation

Overview

- Workflow LFQ analysis at the FGCZ
- Kickoff meeting
- Sample size estimation (the `power.t.test` function)
- Sample size estimation example (hands on)

The LFQ workflow - bioinformatics part

- **kickoff meeting** - bioinformatics part
  - Agenda
- **QC experiment**
  - to determine within group variability
  - *Sample size QC*
  - Fix sample sizes
- **Main experiment**
  - Data Analysis and *result delivery* including linear modelling and ORA and GSEA analysis.
  - meeting to discuss results

Kickoff meeting - bioinformatics

- Aim
  - Improve quality of services
  - Get all the information we need to perform analysis
  - Give *statistical* guidance
  - Ensure reproducibility
- Use of an Agenda
  - helps to prepare for the meeting
  - streamlines discussion
  - focuses on points relevant to the data analysis
  - Keep meeting short (15-30 minutes)
  - Ensure reproducibility

https://wolski.github.io/prolfqua/articles/Auxiliary_ExDesignSurvey.html

Kickoff meeting - bioinformatics

Document

- which protein database to use
- which and how many samples to use for the QC
- collect all the parameters for sample size estimation
- specify the design of the main experiment
  and the hypothesis to be tested
- document all the names of the factors and factor levels

https://wolski.github.io/prolfqua/articles/Auxiliary_ExDesignSurvey.html

Sample size estimation

- Hand in 4 samples of the same condition (ideally of condition with highest variability).
- Estimate the variance of all measured protein.
- Compute sample sizes for main experiment given:
  - biologically relevant effect size
  - power of the test
  - observed variance
  - size of test

Types of error when testing hypothesis

A **type I error** (false positive) occurs when
the null hypothesis (H0) is true, but is rejected.
The *type I error rate* or **significance level** (p-Value)
is the probability of rejecting the
null hypothesis given that it is true.

A **type II error** (false negative) occurs when
the null hypothesis is false,
but erroneously fails to be rejected.
The *the type II error rate* is denoted by the Greek letter β
and is related to the **power of a test** (which equals $1 - \beta$).

For a given test, the only way to reduce both error rates
is to **increase the sample size**, and this may not be feasible.

		reality	
		H_0 = true	H_0 = false
conclusion	H_0 is not rejected	OK	type II error
	H_0 is rejected	type I error	OK

Sample size estimation

- Because of the equivalence between linear models and the t-test we use the sample size estimation method `power.t.test`.

Sample size estimation - esitmating difference d

What **biologically relevant difference** can be detected with a two-sample t-test using a significance level of 0.05 and a power of 0.8 given a standard deviation of 0.5 and group size of 10?

```
power.t.test(n = 10, sd = 0.5, power = 0.8, sig.level = 0.05)
```

```
##
##      Two-sample t test power calculation
##
##               n = 10
##           delta = 0.6624728
##              sd = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Sample size estimation - estimating power d

Power is the probability of rejecting the null hypothesis when, in fact, it is false.

What is **the power** of the two-sample t-test using a significance level of 0.05 if we want to detect a biologically relevant difference of 0.59 given a standard deviation of 0.5 and group size of 10 samples?

```
power.t.test(delta = 0.59, n = 10, sd = 0.5, sig.level = 0.05)
```

```
##
##      Two-sample t test power calculation
##
##               n = 10
##           delta = 0.59
##              sd = 0.5
##       sig.level = 0.05
##           power = 0.7039889
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Why `delta = 0.59`? It is log2(FC), and `2^(0.59) = 1.505247`.

Sample size estimation

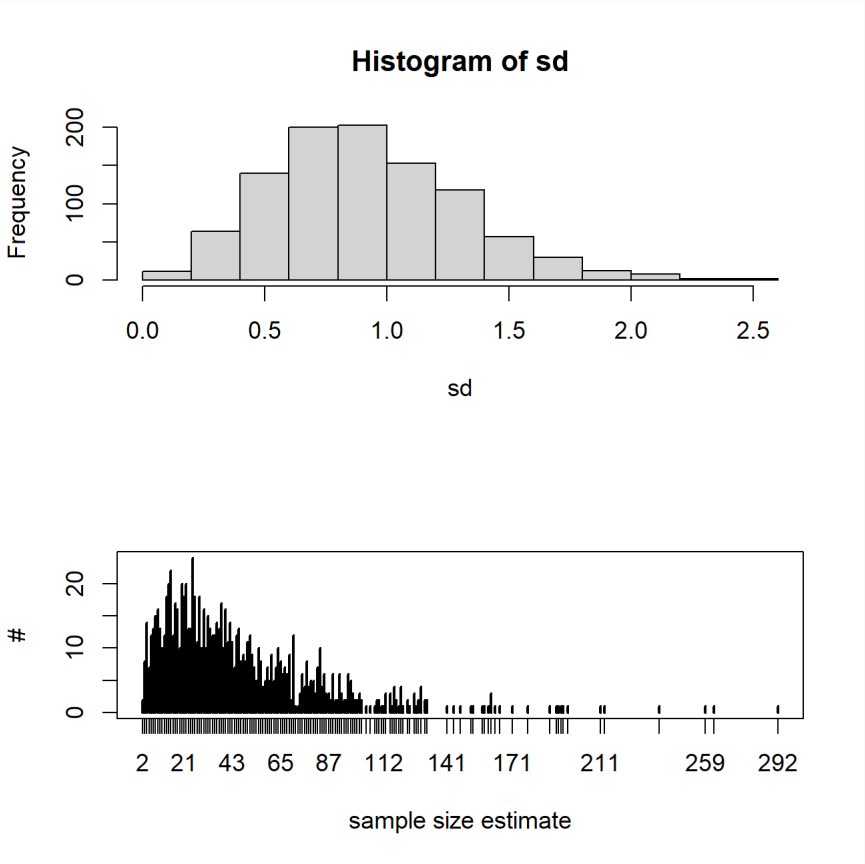
What is **the sample size** of the two-sample t-test using a significance level of 0.05 if we want to detect a biologically relevant difference of 0.59 given a standard deviation of 0.5 and a power of 0.8?

```
power.t.test(delta = 0.59, sd = 0.5, sig.level = 0.05, power = 0.8)
```

```
##
##      Two-sample t test power calculation
##
##               n = 12.31238
##           delta = 0.59
##              sd = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Sample size estimation - error of

```
sd ← sapply(1:1000,
            function(x){
              sd(rnorm(4,0,1))
              })
tmp ← sapply(sd,
             function(x){
               power.t.test(d = 0.59,
                            sd = x,
                            sig.level = 0.05,
                            power = 0.8)$n})
par(mfrow = c(2,1))
hist(sd)
plot(table(ceiling(tmp)),
     ylab="#", xlab="sample size estimate")
```



We simulate 1000 samples of size 4 from normal distribution and determine the standard deviation of each sample. For each standard deviation we computed the sample size.

Sample size estimation - error of

For a protein your uncertainty of the standard deviation and the required sample size is large when you measure 4 samples to estimate the sd.

In an LFQ experiment you measure thousands of proteins and determine their variances. The error of the median standard deviation will be small and therefore the sample size estimate should be OK.

```
power.t.test(sd = median(sd), d = 0.59, sig.level = 0.05, power = 0.8)$n
```

```
## [1] 36.47463
```

We assume that the true variance of the proteins is the same.
Proteins will have **different** true but unknown underlying variances.

Calculating Observed Power

You run a test and you have the **estimates** of the sd, $p.value$ and fold change $delta$.
We could compute the power by taking the *estimates* obtained by the test.

```
power.t.test(delta = estimated_fc, n = nr_samples_per_group,
             sd = esitmated_sd,  sig.level = p.value)
```

Some proteomics software is doing it (e.g. Progenesis).
How good might this power estimate be?
How *good* was the sample size estimate given an sd estimate of a single protein?
Calculating Observed Power Is Just Transforming Noise
How you report the power of your test?
State the parameters of the sample size estimation for your experiment (which includes power).

Sample size estimation in practice

- Data preprocessing
- Checking assumptions
- Sample size estimation using `prolfqua`

https://wolski.github.io/prolfqua/articles/QualityControlAndSampleSizeEstimation.html

Data preprocessing - Normalizing Intensities

- log2 transform the data
- apply robust z-transformation

$$I_i^n = \frac{I_i^t - med(I_i^t)}{mad(I_i^t)} \cdot \sum_{i=1}^{N} mad(I_i^t)/N$$

All samples have the same deviation

$$mad(I_1^t) = mad(I_2^t)$$

equal to the average original variance of all samples.

More methods fo data normalization are discussed here: Elo *et al.*

Data preprocessing - Protein Intensities

Since we want the sample size estimates on *protein level* we need to infer the protein intensity from peptide intensities.

There are many methods to infer the protein intensity

- aggregate all peptides (e.g. MaxQuant proteinGroups.txt)
- aggregate not more than the top n peptides (e.g. top 3).
- tukey median polish - a two-way decomposition of a 2D matrix.

$$y_{ij} = b_0 + c_i + d_j + \epsilon_{ij}$$

where b_0 - global intercept, c_i, d_j denote the column and row effects, ϵ_{ij} - error.

It is also used in the MSstats package.

Data preprocessing - Protein Intensities

```
peptides_hemo
```

```
##                  S1 S2 S3
## MVLSPADK         14 15 14
## TNVK              7  4  7
## AAWGK             8  2 10
## VGAHAGEYGAEALER  15  9 10
## MFLSFPTTK         0  2  0
```

```
med.d ← medpolish(peptides_hemo,
                  trace.iter = FALSE)
med.d$overall + med.d$col
```

```
## S1 S2 S3
##  8  7  8
```

```
as.matrix(med.d$overall + med.d$row)
```

```
##                  [,1]
## MVLSPADK           14
## TNVK                7
## AAWGK               8
## VGAHAGEYGAEALER    10
## MFLSFPTTK           0
```

Tukey median polish example

Sample size estimation report

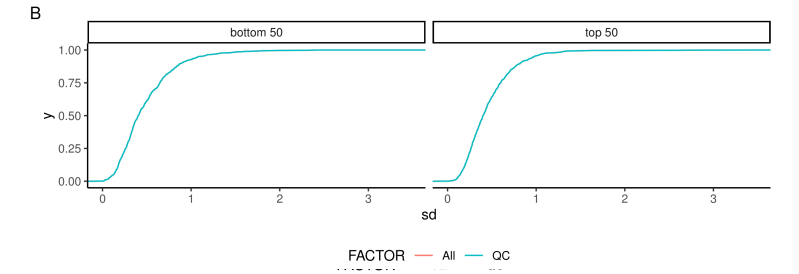
- check if there are no technical problems
- check if assumptions for data normalization are met.

Visualize the distribution of standard deviation
of all proteins using the density function and
the empirical cumulative density function (ecdf)

Sample size estimates for parameters
specified in protocol

- *power* 0.8
- *significance level* 0.05
- *delta* $0.59, 1, 2$
- and standard deviation taken from proteins

quantile	sd	FC=1.51	FC=2	FC=4
50%	0.40	9	4	3
60%	0.47	12	5	3
70%	0.57	16	7	3
80%	0.68	22	9	4
90%	0.85	34	13	5



Conclusion

- This sample size calculation ignores multiple testing problem
- Works if *biological variability* **>>** *biochemical + technical variability*
- Works if your observations are independent and identically distributed (and normally in addition).
- To understand sources of variance measure technical and biochemical and biological replicates.
  - technical coefficient of variation (CV)
  - biochemical CV
  - biological CV
- Sample size calculation is based on the standard deviation estimate (sd)
  - sd estimate for single protein have a large error
  - small error for the median (sd) of all proteins
- Calculating Observed Power Is Just Transforming Noise