Two-Stage Designs for proteomic and gene expression studies applying methods differing in costs

Alexandra Goll Peter Bauer

Section of Medical Statistics
Medical University of Vienna

supported by FWF-Fund Nr. P18698-n15

Multiple Comparison Procedures
Vienna 2007

Motivation

Single-stage designs

Single-stage designs have low power to detect existing effects when a large number of hypotheses is tested.

Two-stage designs

Two-stage designs are a good option to improve the power:

- Pilot Design: The final test decision is only based on the second stage data.
- Integrated Designs: The final test decision is based on the pooled observations over both stages.

Two-Stage-Designs

Screen for promising hypotheses in the first stage which are further investigated in the second stage:

- Limit of resources: total costs C are fixed
- A fraction r of the resources C is used in the first stage for screening
- The remaining resources $(1 - r)C$ are used for second stage

In Genomic or Proteomic studies ...

Scenario 1: Different costs

Different costs per observation may arise at both stages:

- costs per observation in the first stage set to $c_1 = 1$
- cost ratio between stages $c_2 > 1$

Scenario 2: Different costs and effect sizes

There is an increasing focus on using a less accurate assay in early stages and a more accurate one in later stages:

- cost ratio $c_2 > 1$
- effect size ratio between stages $k > 1$

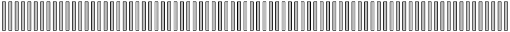
The Two-Stage Design: Scenario 1

m_1 hypotheses

effect size

Δ

$n_1 = \frac{rC}{m_1}$



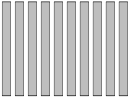
Select H_i: $p_{i1} < \gamma_1$

m_2

effect size

Δ

$n_2 = \frac{(1-r)C}{c_2 m_2}$

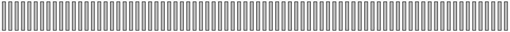


Test decision: Reject H_i: $p_{i2} < \gamma_2$

The Two-Stage Design: Scenario 2

m_1 hypotheses

$n_1 = \frac{rC}{m_1}$

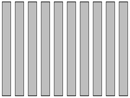


effect size

Δ

Select H_i: $p_{i1} < \gamma_1$

m_2



effect size

$k\Delta$

$n_2 = \frac{(1-r)C}{c_2 m_2}$

Test decision: Reject H_i: $p_{i2} < \gamma_2$

Test Problem:

Consider

m_1 hypotheses for the mean of independent normally distributed observations μ_i with known variance σ^2

$$H_{0i}: \ \mu_i = 0 \quad \text{versus} \quad H_{1i}: \ \mu_i > 0 \quad i = 1, \ldots, m_1$$

assuming independence of observations across hypotheses

Scenario 1: Example

Consider an experiment with

- $m_1 = 1000$ ... number of hypotheses tests
- $C = 20000$ ... fixed total costs (limit of resources)

Given

- $\pi_0 = 0.99$ ... proportion of true null hypotheses among all m_1 hypotheses
- $\Delta = 0.75$ ... effect size

Control level $\alpha = 0.05$:

- FWE: probability of at least one Type I Error
  $\rightarrow$ Bonferroni Adjustment
- FDR: expected proportion of Type I Errors among the rejected hypotheses $\rightarrow$ Storey's procedure

Scenario 1: Example

Single-stage designs

Distribute total costs equally among the hypotheses:
$20000/1000 = 20$ observations per hypothesis test

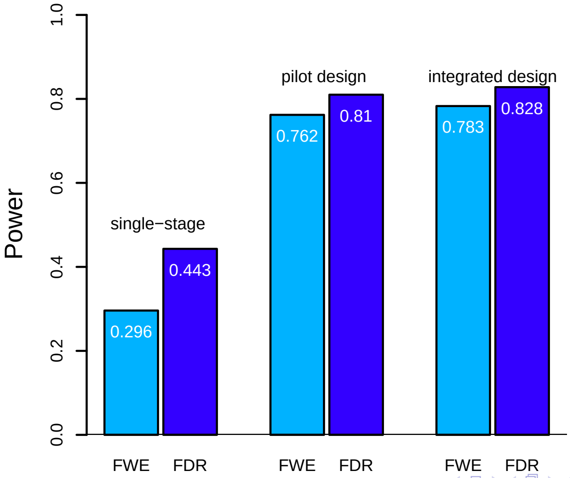
Two-stage designs

The power (expected fraction of correctly rejected alternatives) is optimized with respect to:

- r ..fraction of total costs used in the first stage
- γ_1 ..selection boundary after first stage

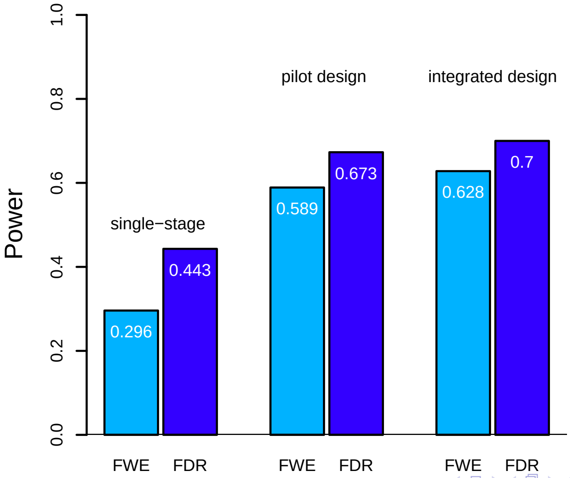
Asymptotic optimal designs

$C = 20000$, $m_1 = 1000$, $\Delta = 0.75$, $\alpha = 0.05$, $\pi_0 = 0.99$, $c_2 = 5$



Asymptotic optimal designs

$C = 20000$, $m_1 = 1000$, $\Delta = 0.75$, $\alpha = 0.05$, $\pi_0 = 0.99$, $c_2 = 15$



Break Even Point in Cost Ratio

If the cost ratio c_2 increases, the power of the two-stage design decreases.

Question

Is there a cost ratio c_2^*, where it does not make sense to apply a two-stage design as compared to the single-stage design?

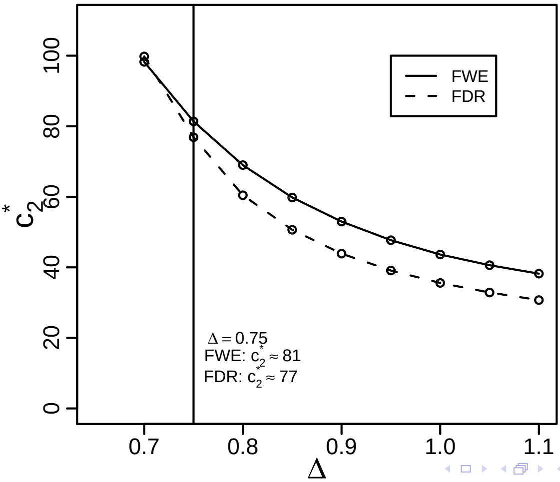
Integrated design

c_2^* does not exist:
The power of the asymptotic optimal integrated design

- is always larger than of the corresponding single-stage design.
- converges to the power of the single-stage design ($\lim\limits_{c_2 \to \infty} r = 1$)

Break Even Point: Pilot Design

$C = 20000$, $m_1 = 1000$, $\pi_0 = 0.99$, $\alpha = 0.05$



Impact of Design Misspecifications

Whereas costs are usually known a priori the optimal designs depend on the unknown parameters π_0 and Δ.

Is there an amount of misspecification where it would have been better to use a single-stage design?

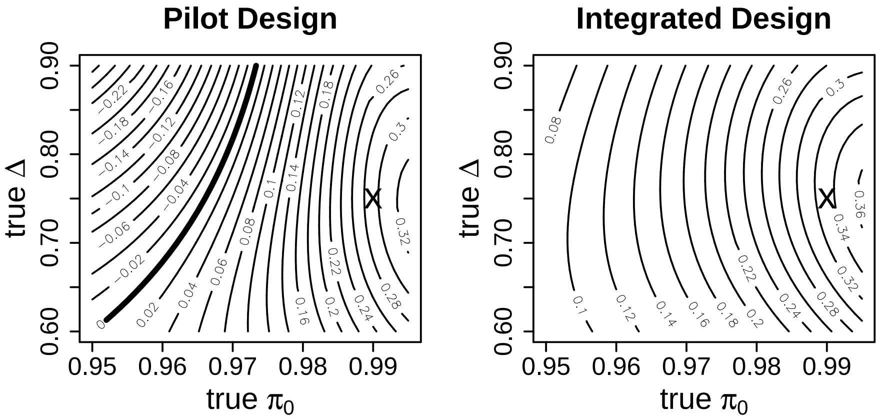
Example

- $C = 20000$
- $m_1 = 1000$
- $c_2 = 15$

r and γ_1 planned for the situation:

- $\pi_0 = 0.99$
- $\Delta = 0.75$

Difference of power values between the two-stage designs and the corresponding single-stage designs (Control of FWE:)



X ... point for which two-stage design parameters are optimized

- The integrated design is more robust against design misspecifications
- if the planned π_0 is larger than the true one:
  loss of power as compared to the single-stage design
- if the planned π_0 is smaller than the true one:
  increase of power as compared to the single-stage design

Scenario 2

The experimenter has two different candidate methods for the measurements from the very beginning:

- low-cost standard method: effect size$= \Delta$
  costs per observation = 1
- high-cost improved method: effect size$= k\Delta$
  costs per observation > 1
- cost ratio between methods $c_2 > 1$
- effect size ratio between methods $k > 1$

Two-Stage Procedures

- first stage: low second stage: low
- first stage: low second stage: high
- first stage: high second stage: high

Pilot Design controlling the FWE

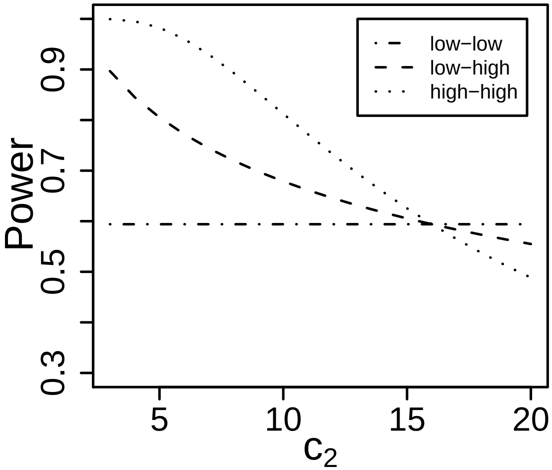
Consider an experiment with

- $\alpha = 0.05$
- $C = 20000$, $m_1 = 1000$
- low-cost method: effect size $\Delta = 0.5$
- high-cost method: $k = 4$ effect size: $0.5 * 4 = 2$

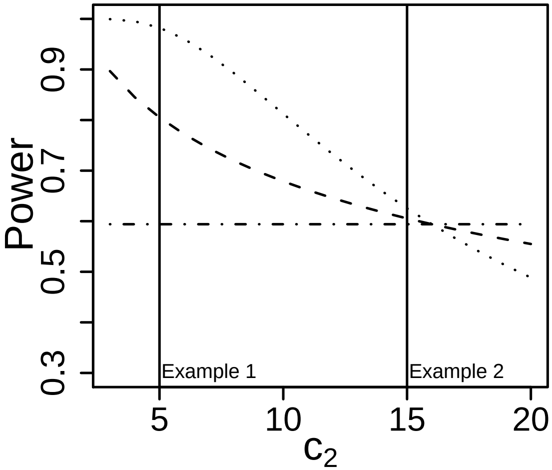
Question

Looking at different values for c_2:
Which of the 3 procedures has the maximal Power?

Common crossing point at $c_2 = k^2 = 16$



Common crossing point at $c_2 = k^2 = 16$



Examples:

Procedure		Example 1: $c_2 = 5$	Example 2: $c_2 = 15$
low-low	Power	0.594	0.594
	n_1	≈ 13	≈ 13
low-high	Power	0.805	0.605
	n_1	≈ 10	≈ 13
high-high	Power	0.983	0.625
	n_1	≈ 3	< 1

Conclusions

Different costs in both stages

Two-stage designs are a good option to improve the power even if the cost ratio between stages c_2 is fairly high.

Misspecification

- The integrated design is more robust against misspecification than the pilot design.
- Optimism in the planning phase with regard to the number of true alternatives may help to avoid loss of power.

If two different methods are available

- Depending on c_2 and k it is preferable to run two-stage designs which apply either the low-cost or the high-cost method at both stages.
- Switching from the low-cost to the high-cost method may only be advisable if there is lack of finance so that n_1 for the high-cost method is to small.

Selected References

- Benjamini,Y, Hochberg,Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing.
  J. R. Statist. Soc. B, 57, 289-300.
- Goll,A, Bauer,P (2007) Two-stage designs applying methods differing in costs.
  Bioinformatics, to appear.
- Satagopan,JM et al. (2004) Two-stage designs for gene-disease association studies. *Biometrics*, 58, 163-170.
- Storey,JD (2002) A direct approach to false discovery rates.
  J. R. Statist. Soc. B, 64, 479-498.
- Storey,JD, Taylor,JE, Siegmund,D (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Statist. Soc. B*, 66, 187-205.
- Wang,H et al. (2006) Optimal two-stage genotyping designs for genome-wide association scans, *Genetic Epidemiology*, 30, 356-368.
- Zehetmayer,S, Bauer,P, Posch,M (2005) Two-stage designs for experiments with a large number of hypotheses.
  Bioinformatics, 21, 3771-3777.