B.Tech. Project

Attaining maximum power under cost constraints in single gene experiments

P Jishnu Jaykumar (201352005)
Abhijeet Singh Panwar (201351005)

under the supervision of
Dr. Bhargab Chattopadhyay

Indian Institute of Information Technology, Vadodara

May 5, 2017

Introduction

- Single gene experiments are assays for querying ribonucleic acid species abundance in individual or pools of cells.

Introduction

- Single gene experiments are assays for querying ribonucleic acid species abundance in individual or pools of cells.
- They have enormous importance in varied fields like evolution, pathology and drug development.

Introduction

- Single gene experiments are assays for querying ribonucleic acid species abundance in individual or pools of cells.
- They have enormous importance in varied fields like evolution, pathology and drug development.
- However, these experiments are usually performed by pre-specifying the sample size during experimental design, which may lead to under powered or over powered hypothesis tests.

Introduction

- Single gene experiments are assays for querying ribonucleic acid species abundance in individual or pools of cells.
- They have enormous importance in varied fields like evolution, pathology and drug development.
- However, these experiments are usually performed by pre-specifying the sample size during experimental design, which may lead to under powered or over powered hypothesis tests.
- This resulted in a shift to experimental design paradigms with minimal number of replicates to maximize power.

Introduction

- Single gene experiments are assays for querying ribonucleic acid species abundance in individual or pools of cells.
- They have enormous importance in varied fields like evolution, pathology and drug development.
- However, these experiments are usually performed by pre-specifying the sample size during experimental design, which may lead to under powered or over powered hypothesis tests.
- This resulted in a shift to experimental design paradigms with minimal number of replicates to maximize power.
- Several studies have analyzed sample size required to maximize power in hypothesis testing in RNA abundance experiments.

Introduction - Contd.

- Recently, for the two-stage experimental design framework, algorithms like SCOTTY have been proposed.

Introduction - Contd.

- Recently, for the two-stage experimental design framework, algorithms like SCOTTY have been proposed.
- But SCOTTY fails, if there are multiple (> 2) cell types.

Introduction - Contd.

- Recently, for the two-stage experimental design framework, algorithms like SCOTTY have been proposed.
- But SCOTTY fails, if there are multiple (> 2) cell types.
- Also, two-stage procedures are imperfect if pilot sample is unrepresentative.

Introduction - Contd.

- Recently, for the two-stage experimental design framework, algorithms like SCOTTY have been proposed.
- But SCOTTY fails, if there are multiple (> 2) cell types.
- Also, two-stage procedures are imperfect if pilot sample is unrepresentative.
- Thus the focus is to maximize the power of the hypothesis test related to a single gene belonging to more than two cell types under a cost constraint.

Some facts

- In experimental research, limited funding is allotted beforehand to carry out the sampling process.
- Thus under a cost constraint we have to carry out the test of equivalence or inferiority (or superiority) with minimum errors.
- Note that we fix the probability of type-I error to α (0.05 in our case),
- So the idea is to minimize the probability of type-II error or equivalently maximizing the power under cost constraints.

Formulation

- Suppose there are K cell types belonging to a particular gene.

Formulation

- Suppose there are K cell types belonging to a particular gene.
- For the i^{th} cell type, suppose $X_{i1}, \ldots, X_{in_i}$ be iid random variables, not necessary normal, with means μ_i and variances σ_i^2.

Formulation

- Suppose there are K cell types belonging to a particular gene.
- For the i^{th} cell type, suppose $X_{i1}, \ldots, X_{in_i}$ be iid random variables, not necessary normal, with means μ_i and variances σ_i^2.
- Consider $\delta = \mathbf{c}'\boldsymbol{\mu}\left(\sum_{i=1}^{K} c_i\mu_i\right)$, $\mathbf{c} = (c_1, \ldots, c_K)$ is known.

Formulation

- Suppose there are K cell types belonging to a particular gene.
- For the i^{th} cell type, suppose $X_{i1}, \ldots, X_{in_i}$ be iid random variables, not necessary normal, with means μ_i and variances σ_i^2.
- Consider $\delta = \mathbf{c}'\boldsymbol{\mu}\left(\sum_{i=1}^{K} c_i\mu_i\right)$, $\mathbf{c} = (c_1, \ldots, c_K)$ is known.
- Estimator of δ is $\widehat{\delta}_n = \sum_{i=1}^{K} c_i\bar{X}_{n_i}$.

Formulation Contd.

- γ_i's: sample size allocation ratios (unknown), $n_i = \gamma_i n_1$, $\gamma_1 = 1$.

Formulation Contd.

- γ_i's: sample size allocation ratios (unknown), $n_i = \gamma_i n_1$, $\gamma_1 = 1$.
- $Var(\widehat{\delta}_n) = \sum_{i=1}^{K} c_i^2 \sigma_i^2 / n_i = \frac{1}{n_1} \sum_{i=1}^{K} c_i^2 \sigma_i^2 / \gamma_i$.

Formulation Contd.

- γ_i's: sample size allocation ratios (unknown), $n_i = \gamma_i n_1$, $\gamma_1 = 1$.
- $Var(\widehat{\delta}_n) = \sum_{i=1}^{K} c_i^2 \sigma_i^2 / n_i = \frac{1}{n_1} \sum_{i=1}^{K} c_i^2 \sigma_i^2 / \gamma_i$.
- Using CLT,

$$\frac{\sqrt{n_1}(\widehat{\delta}_n - \delta)}{\sqrt{\sum_{i=1}^{K} c_i^2 \sigma_i^2 / \gamma_i}} \xrightarrow[\mathcal{D}]{} N(0,1)$$

Test for Equivalence

- Compare δ, to two equivalence margins with $\delta_1 < \delta < \delta_2$

$$H_0 : \delta \leq \delta_1 \text{ or } \delta \geq \delta_2 \text{ against } H_A : \delta_1 < \delta < \delta_2$$

Test for Equivalence

- Compare δ, to two equivalence margins with $\delta_1 < \delta < \delta_2$

$$H_0 : \delta \leq \delta_1 \text{ or } \delta \geq \delta_2 \text{ against } H_A : \delta_1 < \delta < \delta_2$$

- Consider the test for equivalence. Then the null hypothesis H_0 is rejected at $\alpha\%$ if

$$\delta_1 + z_\alpha \sqrt{Var(\widehat{\delta}_n)} < \widehat{\delta}_n < \delta_2 + z_\alpha \sqrt{Var(\widehat{\delta}_n)}$$

Test for Equivalence

- Compare δ, to two equivalence margins with $\delta_1 < \delta < \delta_2$

$$H_0 : \delta \leq \delta_1 \text{ or } \delta \geq \delta_2 \text{ against } H_A : \delta_1 < \delta < \delta_2$$

- Consider the test for equivalence. Then the null hypothesis H_0 is rejected at $\alpha\%$ if

$$\delta_1 + z_\alpha \sqrt{Var(\widehat{\delta}_n)} < \widehat{\delta}_n < \delta_2 + z_\alpha \sqrt{Var(\widehat{\delta}_n)}$$

- The approximate power function is

$$P_{TE}(\delta) = \Phi\left(-z_\alpha + \frac{\delta_2 - \delta}{\sqrt{Var(\hat{\delta}_n)}}\right) - \Phi\left(z_\alpha - \frac{\delta - \delta_1}{\sqrt{Var(\hat{\delta}_n)}}\right)$$

Test for Inferiority

- Compare δ, to δ_0, we have,

$$H_0 : \delta \leq -|\delta_0| \text{ against } H_A : \delta > -|\delta_0|$$

Test for Inferiority

- Compare δ, to δ_0, we have,

$$H_0 : \delta \leq -|\delta_0| \text{ against } H_A : \delta > -|\delta_0|$$

- Consider the inferiority test. Then the null hypothesis H_0 is rejected at $\alpha\%$ if

$$\widehat{\delta}_n > -|\delta_0| + z_\alpha \sqrt{Var(\widehat{\delta}_n)}$$

Test for Inferiority

- Compare δ, to δ_0, we have,

$$H_0 : \delta \leq -|\delta_0| \text{ against } H_A : \delta > -|\delta_0|$$

- Consider the inferiority test. Then the null hypothesis H_0 is rejected at $\alpha\%$ if

$$\widehat{\delta}_n > -|\delta_0| + z_\alpha \sqrt{Var(\widehat{\delta}_n)}$$

- The approximate power function

$$P_{TI}(\delta) = \Phi\left(-z_\alpha + \frac{\delta + |\,\delta_0\,|}{\sqrt{Var(\hat{\delta}_n)}}\right)$$

Maximizing Power Under Cost Constraints

- Suppose, we have a certain amount of money to carry out sampling, say ₹A_0.

Maximizing Power Under Cost Constraints

- Suppose, we have a certain amount of money to carry out sampling, say ₹A_0.
- If it costs ₹a_i to sample each observation from i^{th} cell type, then $A_0 = \sum_{i=1}^{K} a_i n_i \implies n_1 = \frac{A_0}{\sum_{i=1}^{K} a_i \gamma_i}$.

Maximizing Power Under Cost Constraints

- Suppose, we have a certain amount of money to carry out sampling, say ₹A_0.
- If it costs ₹a_i to sample each observation from i^{th} cell type, then $A_0 = \sum_{i=1}^{K} a_i n_i \implies n_1 = \frac{A_0}{\sum_{i=1}^{K} a_i \gamma_i}$.
- In order to maximize power, the optimal allocation ratio

$$\gamma_i = \frac{\mid c_i \mid \sigma_i}{\mid c_1 \mid \sigma_1} \sqrt{\frac{a_1}{a_i}} \tag{1}$$

Maximizing Power Under Cost Constraints

- Suppose, we have a certain amount of money to carry out sampling, say ₹A_0.
- If it costs ₹a_i to sample each observation from i^{th} cell type, then $A_0 = \sum_{i=1}^{K} a_i n_i \implies n_1 = \frac{A_0}{\sum_{i=1}^{K} a_i \gamma_i}$.
- In order to maximize power, the optimal allocation ratio

$$\gamma_i = \frac{\mid c_i \mid \sigma_i}{\mid c_1 \mid \sigma_1} \sqrt{\frac{a_1}{a_i}} \tag{1}$$

- The optimal sample size for i^{th} cell type is,

$$n_{io} = \frac{A_0 \mid c_i \mid \sigma_i}{\sum_{i=1}^{K} \mid c_l \mid \sigma_l \sqrt{a_l a_i}} \tag{2}$$

Sequential Procedure

- **Step 1**: First obtain $N_{io} = m(\geq 2)$ observations from each of the cell types and for i^{th} cell type ,check

$$m \geq \frac{A_0 \mid C_i \mid S_{iN_{io}}}{\sum_{i=1}^{K} \mid c_l \mid s_{lN_{lo}} \sqrt{a_l a_i}}$$

- If for i^{th} cell type, the above condition gets satisfied, no further observations should be obtained from that cell type. Else go to step 2 after carrying out step 1 for all cell types.

Sequential Procedure Contd.

- **Step 2**: Add $m^{'}$ observations corresponding to the i^{th} cell type and set $N_{io} = m + m^{'}$ and check,

$$m + m' \geq \frac{A_0 \mid C_i \mid S_{iN_{io}}}{\sum_{i=1}^{K} \mid c_l \mid s_{lN_{lo}} \sqrt{a_l a_i}}$$

- If for i^{th} cell type, the above condition gets satisfied, no further sampling needs to be done for that cell type. Else repeat step 2 and continue this until for all cell types, the condition gets satisfied.

Sequential Procedure and Characteristics

- One can find optimal sample size using the stopping rule. N_{io}, is the smallest integer, such that

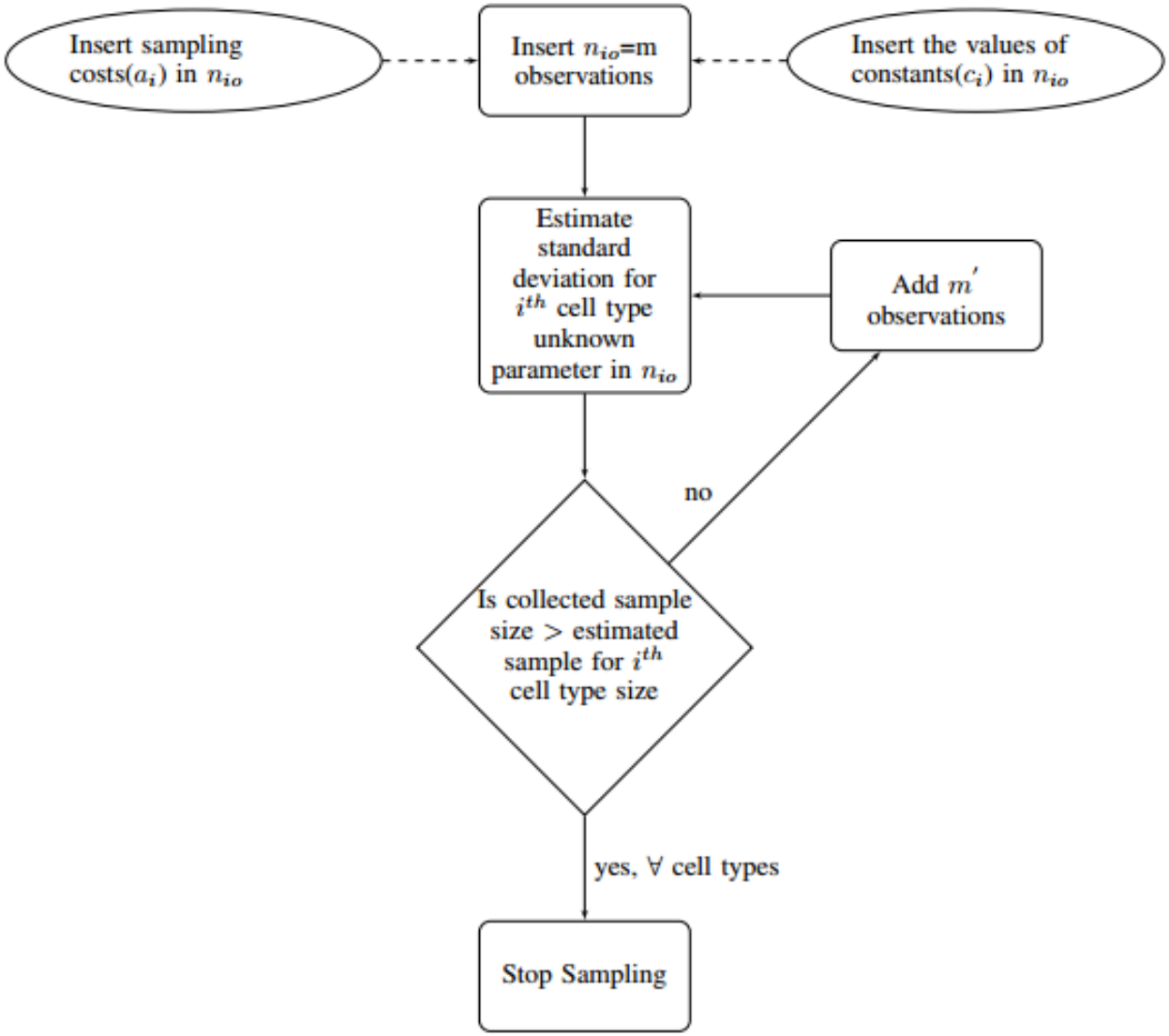
$$n_{io} \geq \frac{A_0 \mid C_i \mid S_{in_{io}}}{\sum_{i=1}^{K} \mid c_l \mid s_{ln_{lo}} \sqrt{a_l a_i}} \tag{3}$$

- The total cost of sampling $\sum_{i=1}^{K} N_{io}$ observations, N_{io} being the estimated final optimal sample size for i^{th} cell type computed using Equation 4, is ₹A_0 .

Contd.

The optimal allocation ratio as defined in Equation 2 to derive minimum variance depends on population variances of the K cell types. In practice, the value of the population variance of each cell type is unknown, therefore, the sample size required to obtain maximum power of the test cannot be computed.

Graphical Representation



Flowchart that describes the sequential procedure developed.

One of the simulation results

Cell Type	$\lvert c_i \rvert$	a_i	$\overline{N}_{io}$	$s(\overline{N}_{io})$	n_{io}	$\overline{N}_{io}/n_{io}$
1 N(μ=1, σ=2)	3	4	467.01	0.23	469	0.99
2 N(μ=2, σ=2)	3	4	233.49	0.78	235	0.99
3 N(μ=3, σ=4)	3	4	1168.9	0.48	1171	0.998
4 N(μ=4, σ=3)	1	2	440.3	0.70	442	0.996
5 N(μ=5, σ=2)	2	3	1080.92	0.80	1082	0.998
6 N(μ=6, σ=4)	3	4	1873.8	0.66	1873	1
7 N(μ=7, σ=5)	4	4	2187.02	0.72	2185	1
8 N(μ=8, σ=6)	5	3	1351.93	1.04	1352	1
9 N(μ=9, σ=3)	6	2	5968.11	1.09	5959	1.001
10 N(μ=10, σ=5)	7	2	3092.26	0.45	3090	1

Final Sample Sizes for Normal Distribution, simulated 5000 times when A_0=₹50000

One of the simulation results

Cell Type	$\lvert c_i \rvert$	a_i	$\overline{N}_{io}$	$s(\overline{N}_{io})$	n_{io}	$\overline{N}_{io}/n_{io}$
1 .95N(1,2)+.05N(20,4)	3	4	59	0.141	60	0.98
2 .95N(2,2)+.05N(20,4)	3	4	58	0.140	60	0.96
3 .95N(3,4)+.05N(20,4)	3	4	118	0.210	120	0.98
4 .95N(4,3)+.05N(20,4)	1	2	41	0.122	43	0.953
5 .95N(5,2)+.05N(20,4)	2	3	45	0.123	47	0.957
6 .95N(6,4)+.05N(20,4)	3	4	118	0.209	120	0.98
7 .95N(7,5)+.05N(20,4)	4	4	200	0.271	200	1
8 .95N(8,6)+.05N(20,4)	5	3	338	0.331	347	0.974
9 .95N(9,3)+.05N(20,4)	6	2	259	0.303	255	1.01
10 .95N(10,5)+.05N(20,4)	7	2	493	0.448	495	0.995

Final Sample Sizes for Mixture-Normal Distribution, simulated 5000 times Distribution when A_0=₹50000

Graph

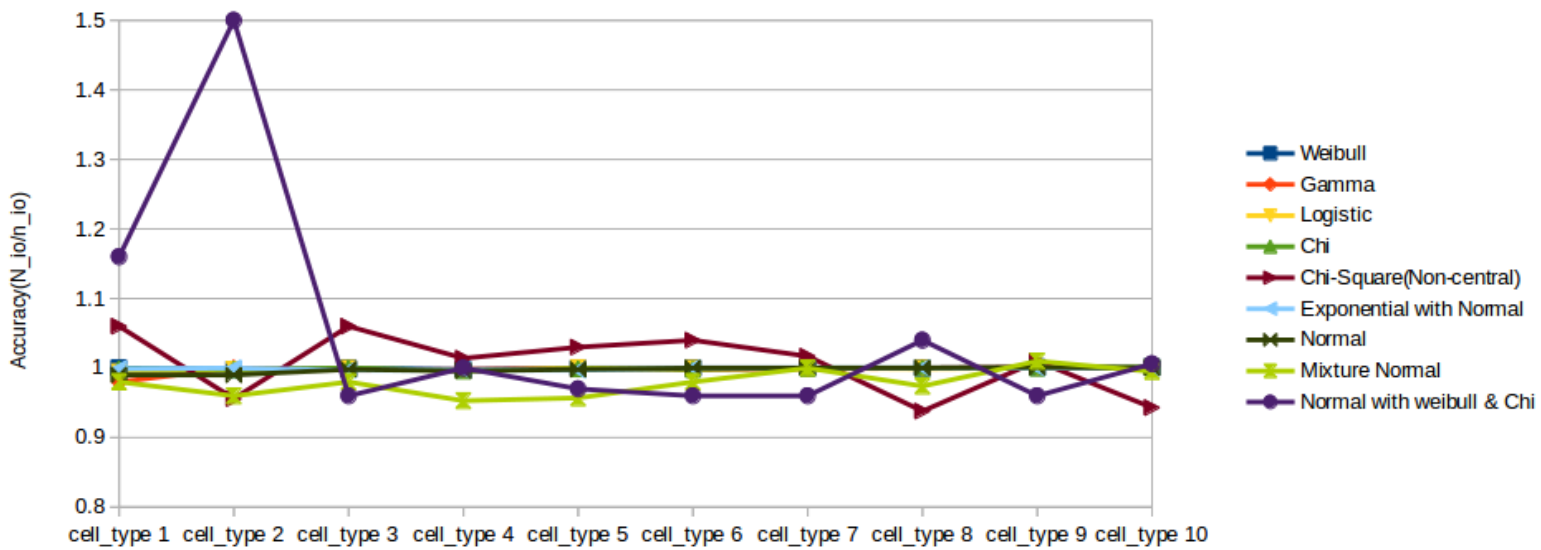


Figure: Accuracy w.r.t. to all the distributions taken into account for the simulation study.

About the toolkit

- A toolkit (web-application) for the sequential procedure was also developed.
- Input: A xlsx file containing data in a predefined format.
- Output: Optimal sample sizes for each cell type specified in the input file.
- Intermediate stages of the procedure can also be viewed.
- It is temporarily hosted at `http://139.59.74.69/alpha_version/`

Conclusions

- Using the sequential procedure we found the optimal sample sizes of the cell types, which are required to obtain maximum power under cost constraints.

Conclusions

- Using the sequential procedure we found the optimal sample sizes of the cell types, which are required to obtain maximum power under cost constraints.
- The simulation study showed that on an average, using our procedure, the estimated optimal sample size and theoretical sample size are close for most of the situations, except for the mixture of normal, weibull and chi distribution.

Conclusions

- Using the sequential procedure we found the optimal sample sizes of the cell types, which are required to obtain maximum power under cost constraints.
- The simulation study showed that on an average, using our procedure, the estimated optimal sample size and theoretical sample size are close for most of the situations, except for the mixture of normal, weibull and chi distribution.
- The procedure developed is independent of distributions.

Conclusions Contd.

- SCOTTY is limited to normal distribution and hence can't handle the datasets following other distributions.

Conclusions Contd.

- SCOTTY is limited to normal distribution and hence can't handle the datasets following other distributions.
- Due to this factor, our procedure gains an advantage over SCOTTY.

Conclusions Contd.

- SCOTTY is limited to normal distribution and hence can't handle the datasets following other distributions.
- Due to this factor, our procedure gains an advantage over SCOTTY.
- We have developed a toolkit for estimating optimal sample sizes of different cell types corresponding to a single gene and for making an inference about hypothesis test.

Conclusions Contd.

- SCOTTY is limited to normal distribution and hence can't handle the datasets following other distributions.
- Due to this factor, our procedure gains an advantage over SCOTTY.
- We have developed a toolkit for estimating optimal sample sizes of different cell types corresponding to a single gene and for making an inference about hypothesis test.

Future Work

- The project consisted of simulation study and toolkit development for the sequential procedure with a **single gene** having multiple cell types.

Future Work

- The project consisted of simulation study and toolkit development for the sequential procedure with a **single gene** having multiple cell types.
- The future activities include developing a procedure for a **set of genes** with multiple cell types.

Future Work

- The project consisted of simulation study and toolkit development for the sequential procedure with a **single gene** having multiple cell types.
- The future activities include developing a procedure for a **set of genes** with multiple cell types.
- In addition to this, the toolkit will be also added up with extra features and would be made to run for set of genes scenario as well.

Future Work

- The project consisted of simulation study and toolkit development for the sequential procedure with a **single gene** having multiple cell types.
- The future activities include developing a procedure for a **set of genes** with multiple cell types.
- In addition to this, the toolkit will be also added up with extra features and would be made to run for set of genes scenario as well.
- The final outcome would be a rich web and desktop application.

Contributions

Abhijeet

- Writing the R code for the sequential procedure.
- Simulation study using
  - Gamma Distribution
  - Weibull Distribution
  - Normal-Exponential Distribution
  - Chi Distribution
  - Chi Square (Non-Central) Distribution
  - Logistic Distribution

Jishnu

- R code reviewing and bug fixing.
- Transforming the R code to Python code.
- Simulation study using
  - Normal distribution
  - Mixture-normal distributions
  - Mixture of Normal + Weibull + Chi Distribution
- Toolkit development.

References I

[BSM+13] Michele A Busby, Chip Stewart, Chase A Miller, Krzysztof R Grzeda, and Gabor T Marth, *Scotty: a web tool for designing rna-seq experiments to measure differential gene expression*, Bioinformatics **29** (2013), no. 5, 656–657.

[GCL11] Jiin-Huarng Guo, Hubert J Chen, and Wei-Ming Luh, *Sample size planning with the cost constraint for testing superiority and equivalence of two independent groups*, British Journal of Mathematical and Statistical Psychology **64** (2011), no. 3, 439–461.

References II

[GS91] Bhaskar Kumar Ghosh and Pranab Kumar Sen, *Handbook of sequential analysis*, CRC Press, 1991.

[JT99] Christopher Jennison and Bruce W Turnbull, *Group sequential methods with applications to clinical trials*, CRC Press, 1999.

[LG16] Wei-Ming Luh and Jiin-Huarng Guo, *Sample size planning for the noninferiority or equivalence of a linear contrast with cost considerations.*, Psychological methods **21** (2016), no. 1, 13.

[Lip90] Mark W Lipsey, *Design sensitivity: Statistical power for experimental research*, vol. 19, Sage, 1990.

References III

[LR06] Erich L Lehmann and Joseph P Romano, *Testing statistical hypotheses*, Springer Science & Business Media, 2006.

[MDS08] Nitis Mukhopadhyay and Basil M De Silva, *Sequential methods and their applications*, CRC press, 2008.

[SS81] Pranab Kumar Sen and Pranab K Sen, *Sequential nonparametrics: invariance principles and statistical inference*, Wiley New York, 1981.

Any Questions?

Statistics, likelihoods, and probabilities mean everything to men, nothing to God.

- Richelle E. Goodrich

Thank You.