DESIGN AND ANALYSIS OF QUANTITATIVE PROTEOMIC EXPERIMENTS

Intro to statistical methods and examples using Skyline

Course organizers

Brendan MacLean  Senior Scientist, University of Washington
Lead developer of Skyline

Meena Choi          PhD Candidate, Purdue University
Lead developer of MSstats

Olga Vitek,             Associate Professor, Northeastern University
DESIGN AND ANALYSIS OF QUANTITATIVE PROTEOMIC EXPERIMENTS

Plan for the day

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00-10:30</td>
<td>Intro to statistical methods. Intro to Skyline.</td>
</tr>
<tr>
<td>10:30-10:45</td>
<td>Refreshments.</td>
</tr>
<tr>
<td>10:45-12:00</td>
<td>Case study in Skyline.</td>
</tr>
<tr>
<td>12:00-1:00</td>
<td>Lunch break on your own.</td>
</tr>
<tr>
<td>1:00-2:30</td>
<td>Case study in Skyline. Introduction to MSstats.</td>
</tr>
<tr>
<td>2:30-2:45</td>
<td>Refreshments.</td>
</tr>
<tr>
<td>2:45-4:00</td>
<td>Case study in MSstats, open time for questions</td>
</tr>
</tbody>
</table>
EXPERIMENTAL DESIGN
AND
BASIC STATISTICS

Olga Vitek

College of Science
College of Computer and Information Science
WHY STATISTICS?

• Variation and uncertainty are unavoidable
  • *Technical variation*: sampling handling, storage, processing
  • *Instrumental variation*: elution time, ion suppression
  • *Signal processing*: peak boundaries, identity, intensity
  • *Biological variation*: variation in protein abundance

"Statistics: a body of methods for making wise decisions in the face of uncertainty." (W. A. Wallis)
OUTLINE

• Translate scientific question into statistics
  • Statistical terms for ‘biomarker’ (or ‘signature’)

• Experimental design
  • Replication, randomization, blocking

• Basic data analysis
  • Simple summaries and models
STATISTICAL GOAL 1: CLASS DISCOVERY
Discover proteins or subjects with similar patterns

- No known class labels
  - E.g., no ‘healthy’ or ‘disease’
  - All variation treated equally
  - No error rates

- Can’t find something meaningful if unsure what we look for
  - Best used for visualization

Gehlenborg et al, Nature Methods, 2010
STATISTICAL GOAL 2: CLASS COMPARISON

Compare mean abundances in subject groups

- Known class labels
  - Compare group averages
  - Report p-values, posterior probabilities etc
- Useful when compare groups of subjects
  - Best used for basic biology
  - Initial (Tier III) biomarker discovery screen
DIFFERENTIALLY ABUNDANT PROTEINS ARE NOT ALWAYS BIOMARKERS

Single protein: Differentially abundant not predictive

Healthy population - mean feature abundance

Healthy individuals in study - observed mean

Disease population - mean feature abundance

Disease individuals in study - observed mean

Inference (conclusions regarding $y_H$ and $y_D$)

Statistical model (properties of $y_H$ and $y_D$)

Observed Systematic Random deviation

Feature intensity = mean signal + due to all sources of variation

$y_{ij} = \mu_i + \epsilon_{ij}$

Log(feature abundance)

Healthy

Disease

Differentially abundant and predictive

Differentially abundant and not predictive
BIOMARKER PROTEINS ARE NOT ALWAYS DIFFERENTIALLY ABUNDANT

Differentially abundant and predictive

Not differentially abundant but predictive

Single protein: Differentially abundant → Predictive
STATISTICAL GOAL 3: CLASS PREDICTION

Classify each subject into a known group

- Known class labels
  - Predict individual subjects
  - Report misclassification error (sensitivity, specificity, predictive value etc)
- Useful when focus on an individual
  - Tier I or Tier II biomarker discovery studies
Outlines

- Translate scientific question into statistics
  - Statistical terms for ‘biomarker’ (or ‘signature’)

- Experimental design
  - Replication, randomization, blocking

- Basic data analysis
  - Simple summaries and models
A STATISTICIAN’S VIEW OF THE EXPERIMENT

Large populations of individuals

Randomly selected individuals

Noisy measurements on selected individuals

Dangers:
Bias: conclusions systematically differ from truth
Inefficiency: unnecessary variation in the data
DEFINITION OF BIAS AND INEFFICIENCY

Bias: \( \bar{Y}_{1..} \) - \( \bar{Y}_{2..} \) systematically different from \( \mu_{1k} - \mu_{2k} \)

Inefficiency: Large \( Var(\bar{Y}_{1..} - \bar{Y}_{2..}) \)
DEFINITION OF BIAS AND INEFFICIENCY

Bias: \( \bar{Y}_{1..} - \bar{Y}_{2..} \) systematically different from \( \mu_{1k} - \mu_{2k} \)

Inefficiency: Large \( \text{Var}(\bar{Y}_{1..} - \bar{Y}_{2..}) \)

Can be prevented by 3 principles of experimental design

Statistical inference: conclusions on \( \mu_1 - \mu_2 \)

Healthy population

Healthy subjects

Spectra

Subject selection

Loop:

Description of Probabilistic systematic variation

Observed Systematic Random deviation

\[ \text{DEFINITION OF BIAS AND INEFFICIENCY} \]
Two levels of randomness imply two types of replication:

- **Biological replicates**: selecting multiple subjects from the population
- **Technical replicates**: multiple runs per subject

Oberg and Vitek, *J. Proteome Research*, 8, 2009
PRINCIPLE 2: RANDOMIZATION
Prevents bias

Two levels of randomness imply two types of randomization:

- **Biological replicates**: random selection of subjects from the population
- **Technical replicates**: random allocation of samples to all processing steps

Oberg and Vitek, *J. Proteome Research*, 8, 2009
EXAMPLE: LACK OF RANDOMIZATION
Hu, Coombes, Morris, Baggerly, Briefings in Functional Genomics, 2005

- Serum samples with five types of cancer
- SELDI-TOF MS
  - normalized, peak picked

Hierarchical clustering of samples

Cancer subtype confounded with time

Same time-based clustering on the QC samples!
PRINCIPLE 3: BLOCKING
Helps reduce both bias and inefficiency

(b) Complete randomization

Complete randomization = inflated variance

Block-randomization = restriction on randomization = systematic allocation

Oberg and Vitek, J. Proteome Research, 8, 2009

Two levels of randomness imply two types of blocks:

- **Biological replicates:** subjects having similar characteristics (e.g. age)
- **Technical replicates:** samples processed together (e.g. in a same day)
EXAMPLE: LACK OF BLOCKING
Hu, Coombes, Morris, Baggerly, Briefings in Functional Genomics, 2005

- Serum samples with two types of cancer
- SELDI-TOF MS, 3 fractions
  - normalized, peak picked

Hierarchical clustering of samples

Protocol change
MATCHING
Blocking with respect to biological risk factors

Complete randomization
= inflated variance

Block-randomization
= restriction on randomization
= systematic allocation

## EXAMPLE

**Block-randomized selection of subjects from repository**

<table>
<thead>
<tr>
<th>Stratification</th>
<th>Disease group</th>
<th>Control</th>
<th>Stable angina</th>
<th>Unstable angina</th>
<th>NSTEMI</th>
<th>STEMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥ 58 y.o; Female</td>
<td></td>
<td>354</td>
<td>300</td>
<td>49</td>
<td>39</td>
<td>29</td>
</tr>
<tr>
<td>≥ 58 y.o; Male</td>
<td></td>
<td>701</td>
<td>843</td>
<td>143</td>
<td>86</td>
<td>54</td>
</tr>
<tr>
<td>&lt; 58 y.o; Female</td>
<td></td>
<td>80</td>
<td>56</td>
<td>5</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>&lt; 58 y.o; Male</td>
<td></td>
<td>264</td>
<td>190</td>
<td>34</td>
<td>23</td>
<td>27</td>
</tr>
</tbody>
</table>

### Counts in the initial repository of samples

<table>
<thead>
<tr>
<th>Stratification</th>
<th>Disease group</th>
<th>Control</th>
<th>Stable angina</th>
<th>Unstable angina</th>
<th>NSTEMI</th>
<th>STEMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥ 58 y.o; Female</td>
<td></td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>≥ 58 y.o; Male</td>
<td></td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>&lt; 58 y.o; Female</td>
<td></td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>&lt; 58 y.o; Male</td>
<td></td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

### Counts of subjects included in the study

*Mass spectra acquired without technical replication*
MULTIPLEXING
Blocking with respect to mass spectrometry run

Multiplexing reduces both bias and variance
(assuming that extra sample handling does not introduce extra variation)
OUTLINE

• Translate scientific question into statistics
  • Statistical terms for ‘biomarker’ (or ‘signature’)

• Experimental design
  • Replication, randomization, blocking

• Basic data analysis
  • Simple summaries and models
TWO-SAMPLE T-TEST
Simple example: label-free experiment, one feature/protein

\[ H_0: \text{'status quo', no change in abundance, } \hat{G}_1 - \hat{G}_0 = 0 \]
\[ Ha: \text{change in abundance, } \hat{G}_1 - \hat{G}_0 \neq 0 \]

observed \( t = \frac{\hat{G}_1 - \hat{G}_0}{\sqrt{\text{Estimate of variation}}} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \)

\[ s_1^2 = \frac{1}{n_1-1} \sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_1)^2 \]

Sample means in each group
Number of replicates
Sample variance
TWO-SAMPLE T-TEST
Simple example: label-free experiment, one feature/protein

H0: ‘status quo’, no change in abundance, \( \hat{G}_1 - \hat{G}_0 = 0 \)
Ha: change in abundance, \( \hat{G}_1 - \hat{G}_0 \neq 0 \)

observed \( t = \frac{\hat{G}_1 - \hat{G}_0}{\sqrt{\text{Estimate of variation}}} \)
\( = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \)

\( s_1^2 = \frac{1}{n_1-1} \sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_1)^2 \)
\( s_2^2 \)

Properties of the means

\( \frac{s_1^2}{n_1} \)

Variance of the sampling distribution of first mean

\( \sqrt{\frac{s_1^2}{n_1}} \)

Standard error of the first mean
As $n$ increases, the mean is less variable and more Normal

This is the Central Limit Theorem

**ASSUMPTION: NORMAL DISTRIBUTION**

As $n$ increases, the mean is less variable and more Normal

This is the Central Limit Theorem

Simulated example

Krzywinski and Altman, Points of Significance Collection, *Nature Methods*
EFFECT OF SAMPLE SIZE
As n increases, the estimates stabilize

Probability distribution of the data

Simulated example
Krzywinski and Altman, Points of Significance Collection, Nature Methods
FINDING DIFFERENTIALLY ABUNDANT PROTEINS

False positive rate

**H0**: ‘status quo’, no change in abundance, \( \hat{G}_1 - \hat{G}_0 = 0 \)

**Ha**: change in abundance, \( \hat{G}_1 - \hat{G}_0 \neq 0 \)

\[
\text{observed } t = \frac{\hat{G}_1 - \hat{G}_0}{\sqrt{\text{Estimate of variation}}}
\]

no difference \( \sim \) Student distribution

Distribution of the score if H0 is true

\( \alpha = \text{False Positive Rate} \)
FINDING DIFFERENTIALLY ABUNDANT PROTEINS

P-value

\( H_0: \) ‘status quo’, no change in abundance, \( \hat{G}_1 - \hat{G}_0 = 0 \)

\( Ha: \) change in abundance, \( \hat{G}_1 - \hat{G}_0 \neq 0 \)

\[
\text{observed } t = \frac{\hat{G}_1 - \hat{G}_0}{\sqrt{\text{Estimate of variation}}}
\]

no difference \( \sim \) Student distribution

Distribution of the score if H0 is true

\( p = p\text{-value} \)

\( p/2 \)

expected \( t \)

-observed \( t \)

observed \( t \)
WITH SMALL SAMPLE SIZE, P-VALUES ARE UNSTABLE

- Repeatedly sampling data leads to different results
- The problem worsens when testing many proteins
- Solutions:
  - Larger sample size
  - Adjustment for multiple testing

Simulated example
Halsey, Curran-Everett, Volwer and Drummond, Nature Methods, 2015
MULTIPLE TESTING
Control False Positive Rate for two proteins

For each protein:
H0: ‘status quo’, no change in abundance,  \( \hat{G}_1 - \hat{G}_0 = 0 \)
Ha: change in abundance,  \( \hat{G}_1 - \hat{G}_0 \neq 0 \)

\[
\text{observed } t = \frac{\hat{G}_1 - \hat{G}_0}{\sqrt{\text{Estimate of variation}}}
\]

no difference \( \sim \) Student distribution

The area = FPR = \( \alpha \)
MULTIPLE TESTING
Control False Positive Rate for two proteins

For each protein:
H0: ‘status quo’, no change in abundance, \( \hat{G}_1 - \hat{G}_0 = 0 \)
Ha: change in abundance, \( \hat{G}_1 - \hat{G}_0 \neq 0 \)

observed \( t = \frac{\hat{G}_1 - \hat{G}_0}{\sqrt{\text{Estimate of variation}}} \)

no difference \( \sim \) Student distribution

The area = FPR = \( \alpha \)
MULTIPLE TESTING
Control False Positive Rate for two proteins

For each protein:

H0: ‘status quo’, no change in abundance, \( \hat{G}_1 - \hat{G}_0 = 0 \)
Ha: change in abundance, \( \hat{G}_1 - \hat{G}_0 \neq 0 \)

\[
\text{observed } t = \frac{\hat{G}_1 - \hat{G}_0}{\sqrt{\text{Estimate of variation}}}
\]

- \( \text{no difference } \sim \text{Student distribution} \)

- \( \text{P(at least one incorrect decision) } > \alpha \)
- The univariate FPR does not hold for the list
- Need to define a multivariate error rate

\[\text{The combined area is } > \alpha \!\]
MULTIPLE TESTING
Control False Positive Rate for two proteins

For each protein:

- H0: 'status quo', no change in abundance, $\hat{G}_1 - \hat{G}_0 = 0$
- Ha: change in abundance, $\hat{G}_1 - \hat{G}_0 \neq 0$

For each protein:

- $t = \frac{\hat{G}_1 - \hat{G}_0}{\sqrt{\text{Estimate of variation}}}$
- no difference $\sim$ Student distribution

- $P(\text{at least one incorrect decision}) > \alpha$
- The univariate FPR does not hold for the list
- Need to define a multivariate error rate

The combined area is > $\alpha$!
TESTING M PROTEINS
Change criteria from False Positive Rate to False Discovery Rate

<table>
<thead>
<tr>
<th></th>
<th># of proteins with no detected difference</th>
<th># of proteins with detected difference</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td># true non-diff. proteins</td>
<td>U</td>
<td>V</td>
<td>m₀</td>
</tr>
<tr>
<td># true diff. proteins</td>
<td>T</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>m − R</td>
<td>R</td>
<td>m</td>
</tr>
</tbody>
</table>

- **False discovery rate (FDR)**
  - An infinite number of measurements on same proteins
  - FDR: the *average* proportion of false discoveries
  
  \[
  \text{FDR} = \mathbb{E}\left[\frac{V}{\max(R, 1)}\right]
  \]

Bonferroni approach controls family-wise error rate = \( P(V > 0) \)
ALTERNATIVE TO TESTING: CONFIDENCE INTERVALS
Not all error bars are made the same

A 95% CI means that if we repeatedly collect data and draw confidence intervals, then 95% of them will contain the true mean.

CI are wider than bars indicating standard error of the mean!

Width of the intervals depends on the sample size.

Simulated example
Krzywinski and Altman, Points of Significance Collection, Nature Methods
ERROR BARS PROVIDE DIFFERENT INSIGHT

Absence of overlap does not always mean stat. significance

**Simulated example**

Krzywinski and Altman, Points of Significance Collection, *Nature Methods*
A common misconception about CIs is an expectation that a CI captures the mean of a second sample drawn from the same population with a CI% chance. Because CI position and size vary with each sample, this chance is actually lower. This variety in bars can be overwhelming, and visually relating their relative position to a measure of significance is challenging. We provide a reference of error bar spacing for common P values in Figure 3. Notice that P = 0.05 is not reached until s.e.m. bars are separated by about 1 s.e.m, whereas 95% CI bars are more generous and can overlap by as much as 50% and still indicate a significant difference. If 95% CI bars just touch, the result is highly significant (P = 0.005). All the figures can be reproduced using the spreadsheet available in Supplementary Table 1, with which you can explore the relationship between error bar size, gap and P value.

Be wary of error bars for small sample sizes—they are not robust, as illustrated by the sharp decrease in size of CI bars in that regime (Fig. 2b). In these cases (e.g., n = 3), it is better to show individual data values. Furthermore, when dealing with samples that are related (e.g., paired, such as before and after treatment), other types of error bars are needed, which we will discuss in a future column.

It would seem, therefore, that none of the error bar types is intuitive. An alternative is to select a value of CI% for which the bars touch at a desired P value (e.g., 83% CI bars touch at P = 0.05). Unfortunately, owing to the weight of existing convention, all three types of bars will continue to be used. With our tips, we hope you’ll be more confident in interpreting them.

Martin Krzywinski & Naomi Altman

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nmeth.2659).

We refer to Table 1 for examples of error bars of common P values, along with the relative gap size and deviation width needed to achieve those P values, for 95% CI and for s.e.m. values. The equations are provided in Table S1 in Supplementary File S1. Notice that 95% CI bars are more generous than s.e.m. bars, touching at P = 0.05. Sample means of 0 and 1 are used in these examples, giving a clear demonstration of the difference in behavior between s.e.m. and 95% CI bars. For the same reason, the same gap size will be needed for multiple comparisons, as the overlap indicates a significant difference in all cases. This is different from the behavior of s.e.m. bars, which will overlap for different sample means. For example, if you have a sample mean of 0.5 of 10 samples with a standard error of 0.1, the bars will still overlap, indicating no significant difference between the groups. However, if you have a sample mean of 1.0 of 10 samples with a standard error of 0.1, the bars will also overlap, indicating no significant difference between the groups. This is because the sample means are the same, and the overlap is due to the variability in the samples. It is therefore important to use the correct measure of variability when interpreting the results.