The locally most powerful rank test: A rank test tailored to microbiome data

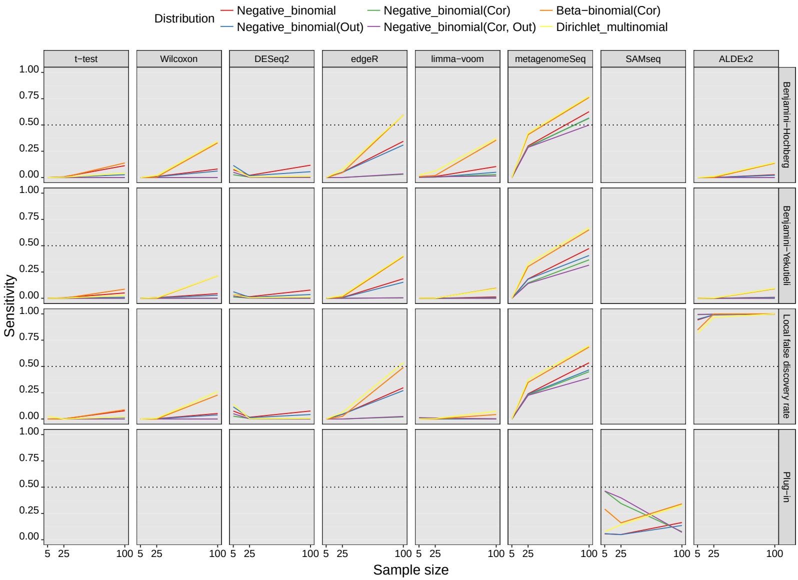
Stijn Hawinkel, Alemu Takele Assefa and Olivier Thas

January 15, 2017

Motivation

- **Differential abundance**: difference in mean taxon abundance between two groups
- Parametric methods (based on negative binomial and Gaussian distributions) have more power than the Wilcoxon rank sum test

Power to detect differential abundance



The Wilcoxon rank sum test

- a.k.a. the **Mann-Whitney U** or **Wilcoxon-Mann-Whitney (WMW)** test
- Given two sets of observations $\mathbf{Y_1}$ and $\mathbf{Y_2}$ with sample sizes n_1 and n_2, it tests the null-hypothesis that for randomly sampled observations

$$H_0 : P(Y_1 > Y_2) = 0.5$$

- The test statistic is of the form

$$\sum_{i=1}^{n1} \sum_{j=1}^{n} I(Y_{i1} \geq Y_{j,pooled})$$

The Wilcoxon rank sum test

- Note that

$$\sum_{j=1}^{n} I(Y_{i1} \geq Y_{j,pooled})$$

equals the rank R_i of Y_{i1} (observation i in group 1)

- The test statistic then becomes

$$T = \sum_{i=1}^{n1} R_i$$

hence the Wilcoxon *rank sum* test

Optimality of rank sum tests

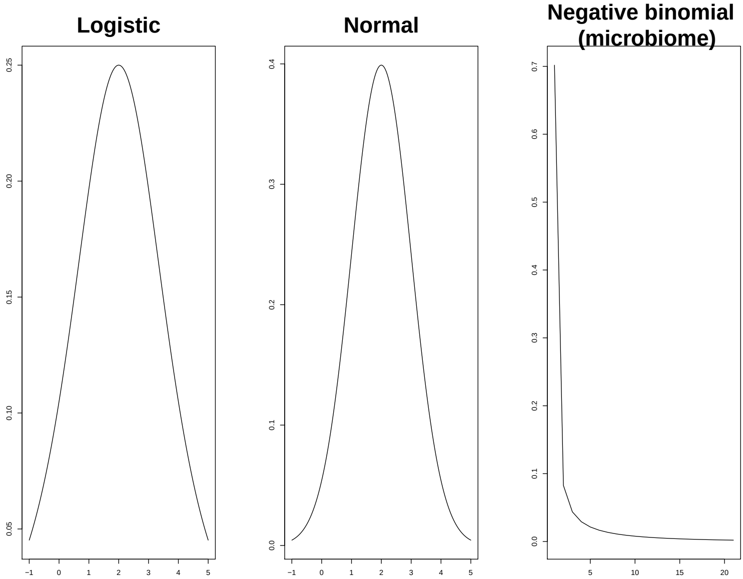
- More formally we can define the linear rank test statistics as

$$T = \sum_{i=1}^{n} c_i a(R_i)$$

with $c_i = 1$ for group 1 and $c_i = 0$ for group 2 and $a()$ the *score function* of the ranks

- For the Wilcoxon rank sum test, $a(R_i) = R_i$
- This choice leads to the best power when the data follow the logistic distribution

Distributions



Logistic

Normal

Negative binomial (microbiome)

LMPRT

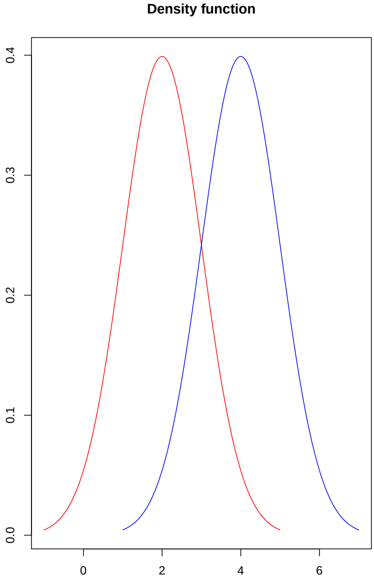
- An old theory provides us with the optimal scores for any distribution

$$a(R_i) = E_f\Big(\frac{\partial}{\partial\Delta} log(f(Y_{(i)};\Delta))|_{\Delta=0}\Big)$$

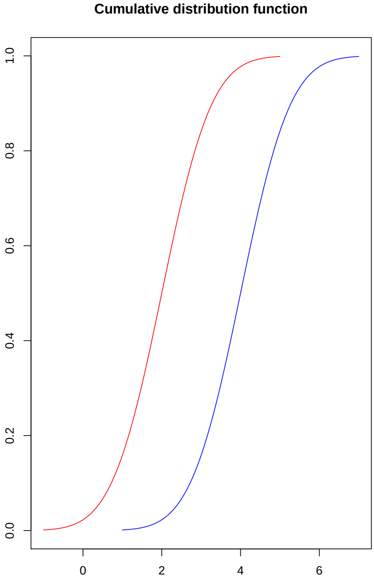
- $Y_{(i)}$ the i-th order statistic, i.e. the i-th smallest observation
- $\Delta = \mu_2 - \mu_1$
- This leads to the *l*ocally *m*ost *p*owerful *r*ank *t*est (LMPRT)
- **Assumption**: Location shift, same shape of distribution

$$f_1(y) = f_2(y - \Delta)$$

Location shift assumption



Density function



Cumulative distribution function

LMPRT

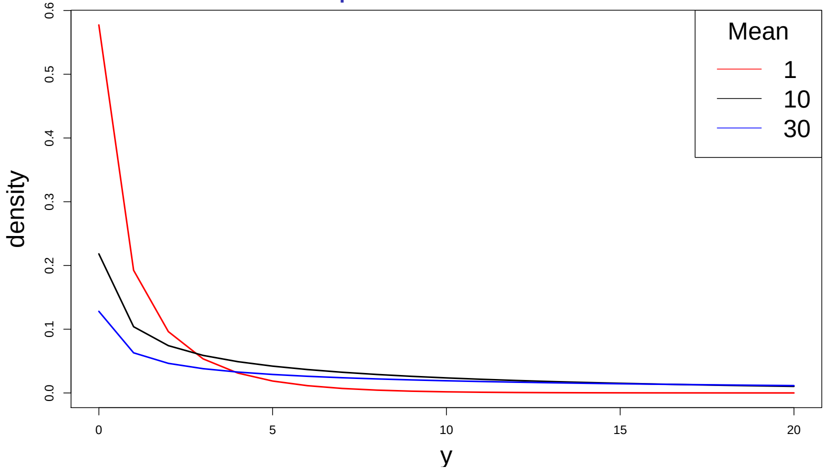
$$a(R_i) = E_f\Big(\frac{\partial}{\partial\Delta} log(f(Y_{(i)};\Delta))|_{\Delta=0}\Big)$$

- if we use the density of the logistic distribution we find $a(R_i) = c_1 * R_i + c_2$, i.e. the WMW
- We could use e.g. the density of the negative binomial distribution BUT
  - This assumption can be wrong
  - We want to be distribution free (develop a **rank** test)
  - Violates the location shift assumption
  - We have plenty of data! $=>$ Let's estimate f from the data

Uninformative scores

- This is a two step approach:
  - Estimate scores $a(R_i)$
  - Use these scores for hypothesis testing
- The score estimation must not be related to the hypothesis of interest!
- We estimate scores based on observations of only **one** of both groups

The location shift assumption



- **Solution**: divide taxa in groups with homogeneous variance where location-shift does hold approximately
- Scores are calculated *conditional* on the variance (or the zero frequency)

Differences in mean

- Taxa are further subdivided into groups of rather homogeneous means
- Δ is then the difference between these means

groups	Variance group 1	$\dots$	Variance group v
Mean group 1	$Y_{11(1)}, Y_{11(2)}, \dots, Y_{11(l)}$	$\dots$	$Y_{v1(1)}, Y_{v1(2)}, \dots, Y_{v1(l)}$
Mean group 2	$Y_{12(1)}, Y_{12(2)}, \dots, Y_{12(l)}$	$\dots$	$Y_{v2(1)}, Y_{v2(2)}, \dots, Y_{v2(l)}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
Mean group m	$Y_{1m(1)}, Y_{1m(2)}, \dots, Y_{1m(l)}$	$\dots$	$Y_{vm(1)}, Y_{vm(2)}, \dots, Y_{vm(l)}$

Outline

$$a(R_i) = E_f\left(\frac{\partial}{\partial\Delta} log(f(Y_{(i)};\Delta))|_{\Delta=0}\right)$$

1. Estimate f non-parametrcially
2. Approximate the derivative to Δ numerically
3. Find the value of $\frac{\partial}{\partial\Delta}$ at $\Delta = 0$ through linear regression
4. Approximate the expectation E_f as an average through bootstrapping

1) Estimate the density

groups	Variance group 1	...	Variance group v
Mean group 1	$Y_{11(1)}, Y_{11(2)}, \dots, Y_{11(l)}$	...	$Y_{v1(1)}, Y_{v1(2)}, \dots, Y_{v1(l)}$
Mean group 2	$Y_{12(1)}, Y_{12(2)}, \dots, Y_{12(l)}$	...	$Y_{v2(1)}, Y_{v2(2)}, \dots, Y_{v2(l)}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
Mean group m	$Y_{1m(1)}, Y_{1m(2)}, \dots, Y_{1m(l)}$	...	$Y_{vm(1)}, Y_{vm(2)}, \dots, Y_{vm(l)}$

- Estimate f as $\hat{f}$ in every mean-variance group with a kernel smoother

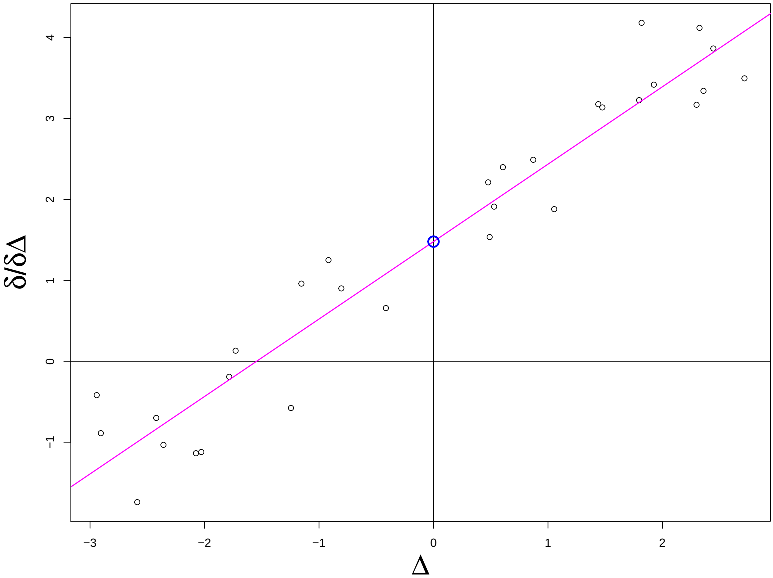
2) Approximate the derivative to Δ numerically

$$\frac{\partial}{\partial \Delta} log(f(Y_{(i)};\Delta)) \approx \frac{log(f_{12}(Y_{12(i)}) - log(f_{11}(Y_{11(i)})}{\mu_2 - \mu_1}$$

for $\Delta_{12} = \mu_2 - \mu_1$

groups	Variance group 1	...	Variance group v
Mean group 1 (μ_1)	$Y_{11(1)}, Y_{11(2)}, \ldots, Y_{11(l)}$	...	$Y_{v1(1)}, Y_{v1(2)}, \ldots, Y_{v1(l)}$
Mean group 2 (μ_2)	$Y_{12(1)}, Y_{12(2)}, \ldots, Y_{12(l)}$	...	$Y_{v2(1)}, Y_{v2(2)}, \ldots, Y_{v2(l)}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
Mean group m (μ_m)	$Y_{1m(1)}, Y_{1m(2)}, \ldots, Y_{1m(l)}$	...	$Y_{vm(1)}, Y_{vm(2)}, \ldots, Y_{vm(l)}$

3) Find the value of $\frac{\partial}{\partial\Delta}$ at $\Delta = 0$ through linear regression



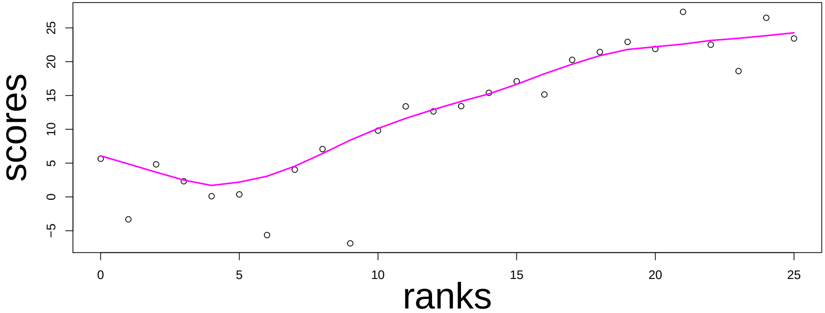
4) Approximate the expectation E_f as an average through bootstrapping

- Stratified bootstrap by mean-variance subgroups
- Sample $n = n_1 + n_2$ observations with replacement (n = total number of samples)

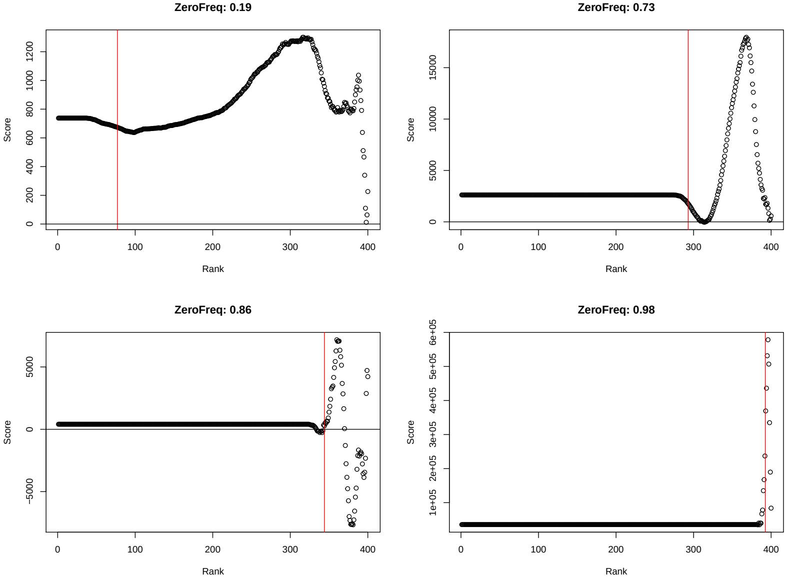
groups	Variance group 1	$\dots$	Variance group v
Mean group 1 (μ_1)	$Y_{11(1)}, Y_{11(2)}, \dots, Y_{11(l)}$	$\dots$	$Y_{v1(1)}, Y_{v1(2)}, \dots, Y_{v1(l)}$
Mean group 2 (μ_2)	$Y_{12(1)}, Y_{12(2)}, \dots, Y_{12(l)}$	$\dots$	$Y_{v2(1)}, Y_{v2(2)}, \dots, Y_{v2(l)}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
Mean group m (μ_m)	$Y_{1m(1)}, Y_{1m(2)}, \dots, Y_{1m(l)}$	$\dots$	$Y_{vm(1)}, Y_{vm(2)}, \dots, Y_{vm(l)}$

Scores

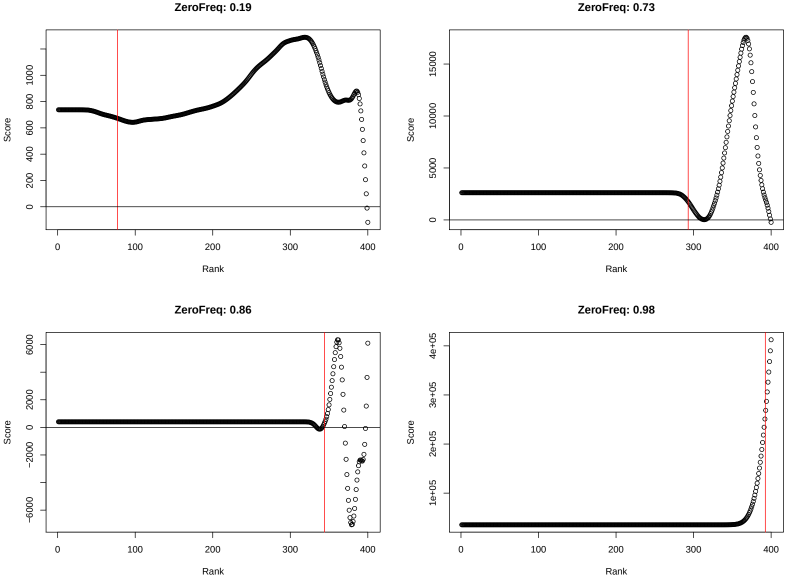
- We now have *n* scores for every variance group
- Smooth the $a(R_i)$ vs. R_i relationship



Real data example (American gut data): Raw scores



Real data example (American gut data): Smooth scores



Hypothesis testing with the LMPRT

- For every taxon: calculate variance
- Use scores from the corresponding variance group to calculate the test statistic for taxon j

$$T_j = \sum_{i=1}^{n} c_i a_j(R_{ij})$$

- P-values can easily be calculated by permuting the group labels c_i
- We only use the ranks R_{ij} in the final test!

Normalization

- For comparability with the WMW, we divide all the counts by their library sizes
- The whole algorithm works with *relative abundances*
- Later we may improve on this

Prospects

- Implement and optimize algorithm
- Test performance in simulation studies