Simultaneous confidence intervals for comparing biodiversity indices estimated from metagenomic trials

Ralph Scherer

26. September 2013
Background

- Human body: More bacterial cells inside \((10^{14})\) than our own cells \((10^{13})\)
- A fact is: The key to understand the human condition lies in understanding the human genome
- But this may be insufficient
  \(\rightarrow\) Sequencing the genomes of our own microbes is necessary too
- Both together can give more information than each alone
- **Metagenomics:** Obtain genomic information directly from microbial communities in their natural habitats
- See “A primer on metagenomics” (Wooley et al., 2010)
Yatsunenko et al. (2012) studied gut microbiomes of 531 individuals. The cohort were healthy children and adults from the Amazonas of Venezuela, rural Malawi and US metropolitan areas. The main interest was to find out if there are differences between age categories or between geographical areas. The data were pre-processed with qiime software. After the quality steps 1,093,740,274 Illumina reads remained. These resulted after the otu-picking script and taxonomic assignment in an OTU table with 11905 different taxa and corresponding counts for the 531 individuals. Mean Count per replicate is 1,935,000. But: There is one replicate with a row sum of 1 → deleted in the following analysis.
Comparison of diversity

- There are several ways to identify possible differences between age groups or geographical areas.
- One solution may be the comparison of the diversity (here: Degree of variation of bacterial species within human gut) between defined groups.
- This can be done using $\alpha$-diversity measures like Shannon or Simpson index.
- Due to the multiple sample design (three geographical areas), simultaneous confidence intervals or multiplicity adjusted $p$-values for the differences between the diversity measures are needed.
Human gut microbiome trial

Figure: Different $\alpha$-diversity measures separated by age and geography
α-diversity measures and related issues

Unequal variances

- The Simpson index $\varphi_i^{(D)} = \sum_{s=1}^{S} \pi_{is}^2$, as well as the Shannon index $\varphi_i^{(H)} = -\sum_{s=1}^{S} \pi_{is} \log(\pi_{is})$ depend on the probability vectors $\hat{\pi}_i = \hat{\pi}_{i1}, ..., \hat{\pi}_{iS}$.

- $\hat{\pi}_i$ represents the estimated probability of occurring for every species $s$, $s = 1, ..., S$ in sample $i$, $i = 1, ..., k$.

- The corresponding variance estimators $\hat{\text{Var}}(\varphi_i^{(D)})$ and $\hat{\text{Var}}(\varphi_i^{(H)})$ mainly depend on the probabilities $\hat{\pi}_i$ and number of species $n_i$.

- According to Rogers and Hsu (2001), one can not assume equal variances across the samples.
Over-dispersion

- Species counts usually show over-dispersion
- Over-dispersion occurs, if the observed variance exceeds the nominal variance of the postulated distribution
- Typically, species counts exhibit a high variation across replicates and a high number of zero counts
- This indicates an over-dispersed distribution
- Idea: Nonparametric bootstrap methods
  - Only based on observed data
  - Take the over-dispersion into account
Asymptotic SCIs (AM)

- Rogers and Hsu (2001) and Fritsch and Hsu (1999) constructed SCIs for the Shannon and Simpson index considering heterogeneous variances.

- Tukey-type SCIs for the Simpson index are constructed in the following way:

\[
\hat{\varphi}_i^{(D)} - \hat{\varphi}_{i}^{(D)} \pm q_{2,1-\alpha;M,R} \sqrt{\hat{\text{Var}}(\hat{\varphi}_i^{(D)}) + \hat{\text{Var}}(\hat{\varphi}_{i}^{(D)})} \tag{1}
\]

with \(q_{2,1-\alpha;M,R}\) being a two-sided quantile from an \(M\)-variate normal distribution with correlation matrix \(R\).

- When estimating the simultaneous confidence intervals for the Shannon index, \(\hat{\varphi}^{(D)}\) is replaced with \(\hat{\varphi}^{(H)}\) and \(\hat{\text{Var}}(\hat{\varphi}^{(D)})\) with \(\hat{\text{Var}}(\hat{\varphi}^{(H)})\).
Disadvantages of the asymptotic SCIs

- Rogers and Hsu (2001) and Fritsch and Hsu (1999) constructed intervals under the assumption of multinomial distributed counts without replicates.
- The probability vector $\pi_i$ is the same for every replicate $j$, $j = 1, \ldots, r$.
- If the data has replicates, the counts may be summed up for every species inside every sample and the indices can then be calculated on the resulting vectors.
- This may lead to an underestimation of the variance.
- Over-dispersion is not considered adequately.
Two ways to calculate the diversity index

(a) Diversity estimation with an ANOVA model, treatment $i$

<table>
<thead>
<tr>
<th>Replicate $j$</th>
<th>Species</th>
<th>...</th>
<th>Species</th>
<th>Index</th>
<th>Param. of interest</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s = 1$</td>
<td>$y_{i11}$</td>
<td>...</td>
<td>$y_{i1s}$</td>
<td>$\hat{\theta}_{i1}$</td>
<td></td>
</tr>
<tr>
<td>$s = S$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$s = S$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$s = S$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ANOVA model estimator $\bar{\theta}_i$

(b) Diversity estimation on summend up counts, treatment $i$

<table>
<thead>
<tr>
<th>Replicate $j$</th>
<th>Species</th>
<th>...</th>
<th>Species</th>
<th>Param. of interest</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s = 1$</td>
<td>$y_{i11}$</td>
<td>...</td>
<td>$y_{i1s}$</td>
<td></td>
</tr>
<tr>
<td>$s = S$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$s = S$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$s = S$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$\sum_{j=1}^{r} y_{i1} ... y_{iS} \hat{\theta}_i$. 
Simultaneous confidence intervals
Asymptotic SCIs with variance estimators considering heterogeneity

Asymptotic gaussian SCIs based on an ANOVA model (AG)

- In case of replicated counts, $\bar{\theta}_i$ may be estimated from an ANOVA model according to method method (a).
- With $\bar{\theta}_i$ and the residuals $\hat{\varepsilon}_{ij} = \hat{\theta}_{ij} - \bar{\theta}_i$, the well-known Tukey-type intervals (Tukey, 1953; Hothorn et al., 2008) can be constructed:

$$
\bar{\theta}_i - \bar{\theta}_{i'} \pm t_{2,1-\alpha;M,R,df=\sum r_i-k} \hat{\sigma} \sqrt{\frac{1}{r_i} + \frac{1}{r_{i'}}}
$$

with variance

$$
\hat{\sigma}^2 = \left( \sum_{i=1}^{k} \sum_{j=1}^{r_i} (\hat{\varepsilon}_{ij} - \bar{\varepsilon}_i^2) / (\sum_{i=1}^{k} r_i - k) \right)
$$

and $t_{2,1-\alpha;M,R,df=\sum r_i-k}$ being a two-sided quantile from an $M$-variate $t$-distribution with correlation matrix $R$. 

R. Scherer (Inst. of Biometry, MHH)
ICSI 2013, Hannover
26. September 2013 11 / 18
Simultaneous confidence intervals
Asymptotic SCIs with variance estimators considering heterogeneity

$t_{\text{max}}$ SCIs based on an ANOVA model (WY)

- Following method (a) compute the parameter of interest $\hat{\theta}_{ij}$, i.e. Simpson’s $\phi$ measure, for every replication $j$, $j = 1, \ldots, r$, separately.
- Bootstrap the estimated indices directly according to Westfall and Young (1993)
  1. Fit a linear model to the estimated indices $\hat{\theta}_{ij}$ resulting in $\hat{\theta}_i$
  2. Bootstrap the residuals $\hat{\varepsilon}_{ij}$ unstratified
  3. For every bootstrap step $b$, $b = 1, \ldots, B$ build the test statistic

$$t_{ij}^* = \frac{\bar{\varepsilon}_i^* - \bar{\varepsilon}_{i'}^*}{\sqrt{(\hat{\sigma}_{i\varepsilon}^2/n_i + (\hat{\sigma}_{i'i\varepsilon}^2)/n_{i'}')}}.$$ (4)

- $q_{1-\alpha}$ is the $1 - \alpha$ empirical quantile of the $B$ values $\max(t_{ij}^*)$.
- The resulting simultaneous confidence intervals are constructed in the following way

$$\bar{\theta}_i - \bar{\theta}_{i'}^* \pm q_{1-\alpha} \sqrt{(\hat{\sigma}_{i}^2/n_i + \hat{\sigma}_{i'}^2/n_{i'})},$$ (5)

where $\hat{\sigma}_i^2$ is the residual mean square for the $i$th treatment in the ANOVA model.
Simultaneous confidence intervals

Asymptotic SCIs with variance estimators considering heterogeneity

$t_{max}$ SCIs based on summed up counts (TS)

1. Bootstrap the original data set in a row, stratified by the $k$ levels of treatments.

2. Estimate the group wise index of interest $\hat{\theta}_i^*$ according to method (b) for every bootstrap sample.

3. In every bootstrap sample, calculate the test statistic

$$t_{ii'} = \frac{(\hat{\theta}_i^* - \hat{\theta}_{i'}^*) - (\hat{\theta}_i - \hat{\theta}_{i'})(\hat{\sigma}_i^2 + \hat{\sigma}_{i'}^2)^*}{\sqrt{((\hat{\sigma}_i^2)^* + (\hat{\sigma}_{i'}^2)^*)}}$$

with the variance estimators based on multinomial assumptions

4. $q_{1-\alpha}$ is the $1 - \alpha$ empirical quantile of the $B$ values $\max(t_{ii'}^*)$.

5. The resulting simultaneous confidence intervals are then

$$\hat{\theta}_i - \hat{\theta}_{i'} \pm q_{1-\alpha} \sqrt{(\hat{\sigma}_i^2 + \hat{\sigma}_{i'}^2)}$$
Simultaneous confidence intervals
Asymptotic SCIs with variance estimators considering heterogeneity

rank-perc SCIs based on summed up counts (PE)

- Bootstrap the original data set in a row, stratified by the $k$ levels of treatments.
- Estimate the group wise index of interest $\hat{\theta}_i^*$ according to method (b) for each bootstrap sample.
- Build differences of interest $\delta_m$ for all bootstrap samples
- Construct SCIs according to Besag et al. (1995)
  1. Rank the differences separately
  2. Compute and store maximum of ranks for each bootstrap sample
  3. Compute the $1 - \alpha$ quantile $t^*$ of the maximum ranks
  4. Finally, the confidence limits are constructed for each elementary parameter $\delta_m$ by taking $[\delta_m^{[B+1-t^*]}; \delta_m^{[t^*]}]$, i.e. the $B+1-t^*$th and $t^*$th value from the ordered sample of the joint empirical distribution obtained for $\delta_m$. 

R. Scherer (Inst. of Biometry, MHH)
ICSI 2013, Hannover
26. September 2013 14 / 18
Simulation results

Figure: Simulation results for the Shannon index

Figure: Simulation results for the Simpson index
analysed example data set

figure: example data results for the shannon index

figure: example data results for the simpson index
The publication corresponding to today’s talk is Scherer and Schaarschmidt (2013).

All methods except for the asymptotic methods based on the linear model are implemented in the R-package **simboot**.

The asymptotic method is implemented in the R-package **multcomp**.

The bioconductor package **phyloseq** was used to import the otu-table from qiime.

**simboot** is on github for bug reporting: https://github.com/shearer/simboot.

A github homepage http://shearer.github.io/simboot/ with a tutorial for sequence data is under development.
Literature I


