Simultaneous confidence intervals for comparing biodiversity indices estimated from metagenomic trials

Ralph Scherer

26. September 2013

Next generation sequencing Experiments

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)

Medizinische Hochschule Hannover

Example: Human gut microbiome trial

- 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 $\rightarrow$ 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 α-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

Medizinische Hochschule Hannover

Human gut microbiome trial

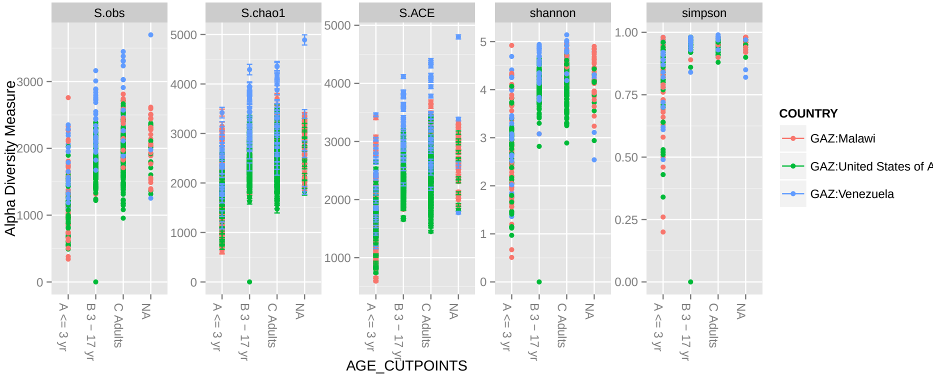

Figure : Different α-diversity measures separated by age and geography

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 $\hat{\pi}_{\mathbf{i}} = \hat{\pi}_{i1}, ..., \hat{\pi}_{iS}$,
- $\hat{\pi}_{\mathbf{i}}$ represents the estimated probability of occurring for every species s, $s = 1, ..., S$ in sample i, $i = 1, ..., k$
- The corresponding variance estimators $\widehat{Var}(\hat{\varphi}^{(D)})$ and $\widehat{Var}(\hat{\varphi}^{(H)})$ mainly depend on the probabilities $\hat{\pi}_{\mathbf{i}}$ and number of species n_i
- According to Rogers and Hsu (2001), one can not assume equal variances across the samples

α-diversity measures and related issues

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

MHH Medizinische Hochschule Hannover

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

$$\widehat{\varphi}_i^{(D)} - \widehat{\varphi}_{i'}^{(D)} \pm q_{2,1-\alpha;M,R}\sqrt{\widehat{Var}(\hat{\varphi}_i^{(D)}) + \widehat{Var}(\hat{\varphi}_{i'}^{(D)})} \qquad (1)$$

  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, $\widehat{\varphi}^{(D)}$ is replaced with $\widehat{\varphi}^{(H)}$ and $\widehat{Var}(\hat{\varphi}^{(D)})$ with $\widehat{Var}(\hat{\varphi}^{(H)})$

Medizinische Hochschule Hannover

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 $\pi_{\boldsymbol{i}}$ is the same for every replicate $j, \;\; j = 1, ..., r$
- 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

MHH Medizinische Hochschule Hannover

Two ways to calculate the diversity index

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

Replicate j	Species $s = 1$	...	Species $s = S$	Index	Param. of interest
1	y_{i11}	...	y_{i1S}	$\hat{\theta}_{i1}$	
2	y_{i21}	...	y_{i2S}	$\hat{\theta}_{i2}$	
3	y_{i31}	...	y_{i3S}	$\hat{\theta}_{i3}$	
r	y_{ir1}	...	y_{irS}	$\hat{\theta}_{ir}$	
ANOVA model estimator					$\bar{\theta}_i$

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

Replicate j	Species $s = 1$	...	Species $s = S$	Param. of interest
1	y_{i11}	...	y_{i1S}	
2	y_{i21}	...	y_{i2S}	
3	y_{i31}	...	y_{i3S}	
r	y_{ir1}	...	y_{irS}	
$\sum_{j=1}^{r}$	$y_{i.1}$	...	$y_{i.S}$	$\hat{\theta}_{i.}$

Asymptotic gaussian SCIs based on an ANOVA model (**AG**)

- In case of replicated counts, $\bar{\theta}_i$ may 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

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

with variance

$$\hat{\sigma}^2 = (\sum_{i=1}^{k} \sum_{j=1}^{r_i})(\hat{\varepsilon}_{ij} - \bar{\varepsilon}_i^2)/(\sum_{i=1}^{k} r_i - k)) \tag{3}$$

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.

t_{max} SCIs based on an ANOVA model (**WY**)

- Following method (a) compute the parameter of interest $\hat{\theta}_{ij}$, i.e. Simpson's φ measure, for every replication j, $j = 1, ..., r$, 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, $b = 1, ..., B$ build the test statistic

$$t^*_{ii'} = \frac{\bar{\varepsilon}^*_i - \bar{\varepsilon}^*_{i'}}{\sqrt{((\hat{\sigma}^2_{i\hat{\varepsilon}})^*/n_i + (\hat{\sigma}^2_{i'\hat{\varepsilon}})^*/n_{i'})}}. \tag{4}$$

4. $q_{1-\alpha}$ is the $1-\alpha$ empirical quantile of the B values $\max(t^*_{ii'})$.
5. The resulting simultaneous confidence intervals are constructed in the following way

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

where $\hat{\sigma}^2_i$ is the residual mean square for the ith treatment in the ANOVA model

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

$$t^*_{ii'} = \frac{(\hat{\theta}^*_{i.} - \hat{\theta}^*_{i'.}) - (\hat{\theta}_{i.} - \hat{\theta}_{i'.})}{\sqrt{((\hat{\sigma}^2_{\hat{\theta}_{i.}})^* + (\hat{\sigma}^2_{\hat{\theta}_{i'.}})^*)}} \tag{6}$$

   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

$$\hat{\theta}_{i.} - \hat{\theta}_{i'.} \pm q_{1-\alpha} \sqrt{(\hat{\sigma}^2_{\hat{\theta}_{i.}} + \hat{\sigma}^2_{\hat{\theta}_{i'.}})}, \tag{7}$$

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 δ_m for all bootstrap samples
- Construct SCIs according to Besag et al. (1995)
    1. Rank the differences seperately
    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 δ_m by taking $\left[\delta_m^{[B+1-t^*]}; \delta_m^{[t^*]}\right]$, i.e. the $B+1-t^*$th and t^*th value from the ordered sample of the joint empirical distribution obtained for δ_m.

Simulation results

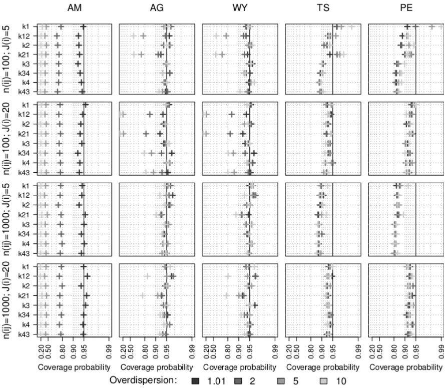

Figure : Simulation results for the Shannon index

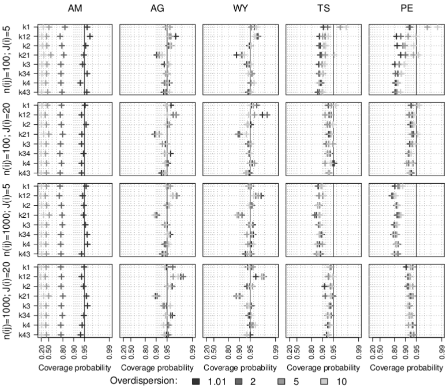

Figure : Simulation results for the Simpson index

Analysed example data set

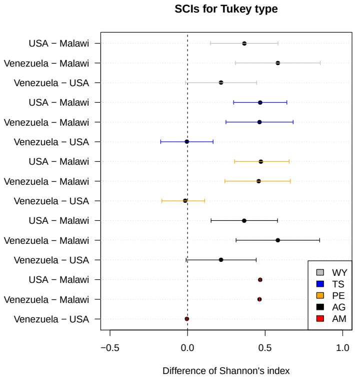

Figure : Example data results for the Shannon index

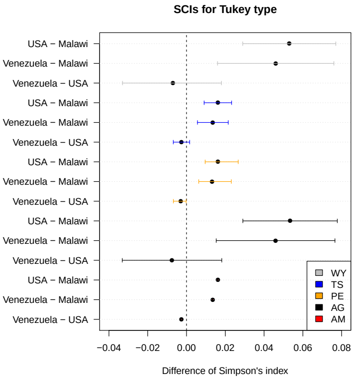

Figure : Example data results for the Simpson index

MHH Medizinische Hochschule Hannover

Software implementation

- 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

Medizinische Hochschule Hannover

Literature I

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian Computation and Stochastic-Systems. *Statistical Science*, 10(1):3–41.

Fritsch, K. S. and Hsu, J. C. (1999). Multiple comparison of entropies with application to dinosaur biodiversity. *Biometrics*, 55(4):1300–1305.

Hothorn, T., Bretz, F., and Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3):346–63.

Rogers, J. A. and Hsu, J. C. (2001). Multiple comparisons of biodiversity. *BIOMETRICAL JOURNAL*, 43(5):617–625.

Scherer, R. and Schaarschmidt, F. (2013). Simultaneous confidence intervals for comparing biodiversity indices estimated from overdispersed count data. *Biometrical . . .* , 55:246–263.

Tukey, J. W. (1953). The problem of multiple comparisons.

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing*. John Wiley & Sons.

Wooley, J. C., Godzik, A., and Friedberg, I. (2010). A primer on metagenomics. *PLoS computational biology*, 6(2):e1000667.

Yatsunenko, T., Rey, F. E., Manary, M. M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., Heath, A. C., Warner, B., Reeder, J., Kuczynski, J., Caporaso, J. G., Lozupone, C. a., Lauber, C., Clemente, J. C., Knights, D., Knight, R., and Gordon, J. I. (2012). Human gut microbiome viewed across age and geography. *Nature*, 486(lvic):222–7.