Lecture 2: Descriptive statistics, normalizations & testing

What do we need to know about a microbiome to understand it?

Scope of Bioinformatics/Omics data standards

Scope-General	Scope-Specific	Description
Experiment description	Reporting (Minimum information)	Documentation for publication or data deposition
	Data exchange & modeling	Communication between organizations and tools
	Terminology	Ontologies and CV's to describe experiments or data
Experiment execution	Physical standards	Reference materials, spike-in controls
	Data analysis & quality metrics	Analyze, compare, QA/QC experimental results

Existing reporting standards for Omics

Acronym	Full name	Domain	Organization
CIMR	Core Information for Metabolomics Reporting	Metabolomics	MSI
MIAME	Minimum Information about a Microarray Experiment	Transcriptomics	MGED
MIAPE	Minimum Information about a Proteomics Experiment	Proteomics	HUPO-PSI
MIGS-MIMS	Minimum Information about a Genome/Metagenome Sequence	Genomics	GSC
MIMIx	Minimum Information about a Molecular Interaction eXperiment	Proteomics	HUPO-PSI
MINIMESS	Minimal Metagenome Sequence Analysis Standard	Metagenomics	GSC
MINSEQE	Minimum Information about a high-throughput Nucleotide Sequencing Experiment	Genomics, Transcriptomics (UHTS)	MGED
MISFISHIE	Minimum Information Specification For In Situ Hybridization and Immunohistochemistry Experiments	Transcriptomics	MGED

MIxS

- The GSC family of minimum information standards (checklists) – Minimum Information about any (x) Sequence (MIxS)
- MIGS – genomes
- MIMS – metagenomes
- MIMARKS – marker genes
- 15 additional environmental package

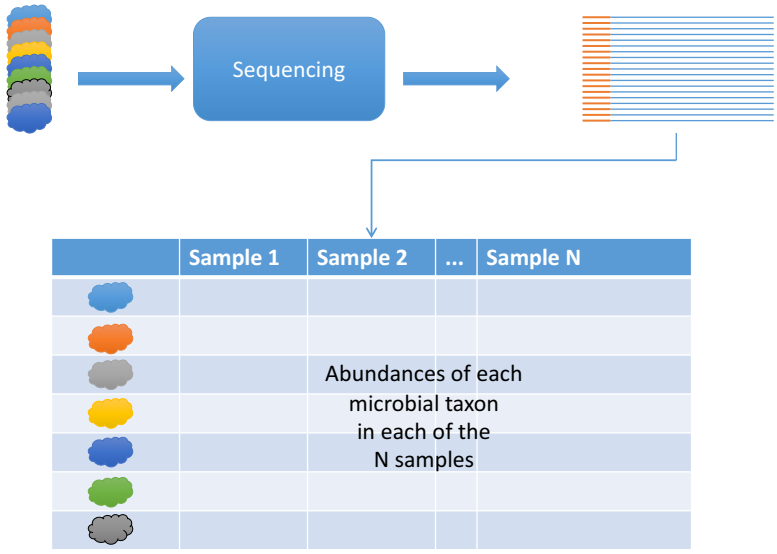
Specification projects	MIGS	MIMS	MIMARKS	New checklists
Checklists		metagenomes	survey	specimen	e.g. pan-genomes
Shared descriptors	collection date, environmental package, environment (biome), environment (feature), environment (material), geographic location (country and/or sea, region), geographic location (latitude and longitude), investigation type, project name, sequencing method, submitted to INSDC			
Checklist specific descriptors	assembly, estimated size, finishing strategy, isolation and growth condition, number of replicons, ploidy, propagation, reference for biomaterial		target gene	
Applicable environmental packages (measurements and observations)	Air Host-associated Human-associated Human-oral Human-gut Human-skin Human-vaginal		Microbial mat/biofilm Miscellaneous natural or artificial environment Plant-associated Sediment Soil Wastewater/sludge Water	

The minimum information about a genome sequence (MIGS-MIMS) specification

http://gensc.org

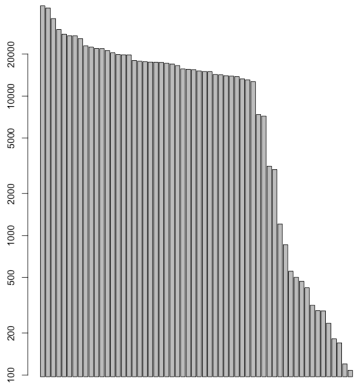
What does the number of sequences tell us about the physical characteristics of the microbiomes?

From sequences to OTU/ASV table



8

Number of reads per sample




```
barplot(sort(sample_sums(mb), decreasing = T),
        names.arg = NA, log="y")
```

9

Normalizing OTU/ASV tables for sequencing effort

Raw Counts

	Sample 1	...	Sample N
	n_{11}		n_{1N}
	n_{21}		n_{2N}
	n_{31}		n_{3N}
	n_{41}		n_{4N}
	n_{51}		n_{5N}
	n_{61}		n_{6N}
	n_{71}		n_{7N}
	$n_{\cdot 1}$		$n_{\cdot 1}$

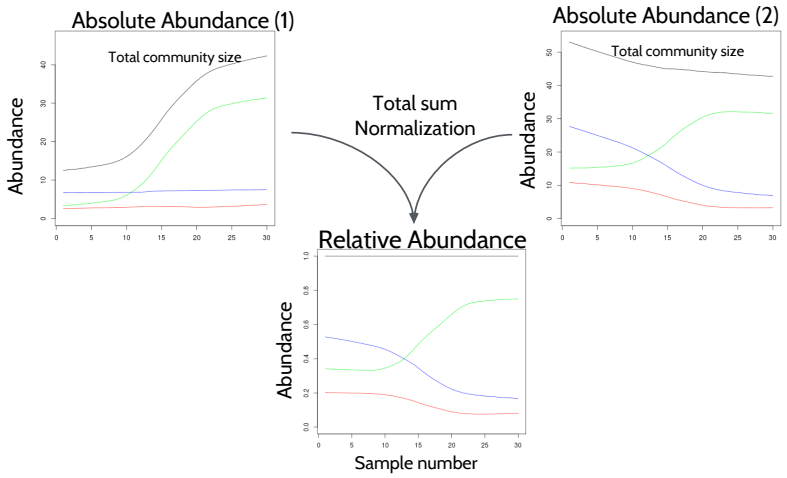
Proportions

	Sample 1	...	Sample N
	p_{11}		p_{1N}
	p_{21}		p_{2N}
	p_{31}		p_{3N}
	p_{41}		p_{4N}
	p_{51}		p_{5N}
	p_{61}		p_{6N}
	p_{71}		p_{7N}
	1		1

$p_{ij} = n_{ij}/n_{\cdot j}$

10

Potential problem with relative abundance



11

Negative correlation of the relative abundances

- The proportions are negatively correlated by design.
- If one (or more) OTUs/ASVs were to increase in absolute abundance, the relative abundances of all others will decrease to accommodate the additive constraint.

12

Compositional data analysis: log ratios

- Main idea: ratios of absolute and compositional data are preserved
- $\log\frac{x_i}{x_j} = \log\frac{\omega_i/M}{\omega_j/M} = \log\frac{\omega_i}{\omega_j}$,
- Where
  - $M = total\ community\ size$
  - *i, j* = microbe
- More details is Aitchison, J. (1986). The statistical analysis of compositional data.

13

Other normalizations

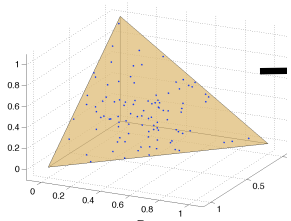
- Normalized by 1 component, n_d
  - $y_{ij} = \log\left(\frac{n_{ij}}{n_{dj}}\right) = \log(n_{ij}) - \log(n_{dj})$
  - $n_{dj} > 0$ for all d
  - Assuming the true abundance of d is the same across all samples
- Normalized by geometric mean (centered)
  - $y_{ij} = \log(\frac{n_{ij}}{g(n_{1j},\ldots,n_{Tj})} = \log(n_{ij}) - \log(g(n_{1j},\ldots,n_{Tj}))$
  - $g(n_{1j},\ldots,n_{Tj}) = \left(\prod_{i=1}^{T} n_{ij}\right)^{1/T}$
- Note: log[0] -> -∞; so often we add '*pseudo-counts*' before these transformations.

14

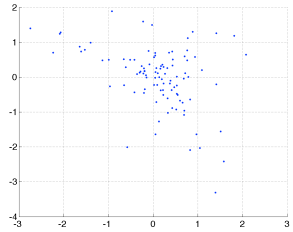
CLR: Centered Log-ratio transformation

- $clr(x) = \log \frac{x}{g(x)}$
- $g(x) = \sqrt[N]{x_1 x_2 \ldots x_N}$
- Transformed data are unconstrained in N-1 dimensions

Compositional data



CLR transformed data



15

Why is microbiome diversity important?

Describing microbiome community is alike to taking a demographic census

	Town1	...	TownN
carpenter	p_{11}		p_{1N}
banker	p_{21}		p_{2N}
student	p_{31}		p_{3N}
teacher	p_{41}		p_{4N}
doctor	p_{51}		p_{5N}
police	p_{61}		p_{6N}
chef	p_{71}		p_{7N}
	1		1

- How many professions are represented?
- How well represented are the different professions?
- Are some professions more popular?

17

Alpha diversity definition(s)

- Alpha diversity describes the diversity of a single community (specimen).
- In statistical terms, it is a scalar statistic computed for a single observation (column) that represents the diversity of that observation.
- There are many statistics that can describe diversity: e.g. taxonomical richness, evenness, dominance, etc.

18

Species richness

- Suppose we observe a community that can contain up to k 'species'.
- The relative proportions of the species are P = $\{p_1, \ldots, p_k\}$.
- Richness is computed as

  $R = \mathbf{1}(p_1) + \mathbf{1}(p_2) + \ldots + \mathbf{1}(p_k)$,

  where $\mathbf{1}(.)$ is an indicator function, i.e. $\mathbf{1}(x) = 1$ if $p_i \neq 0$, and 0 otherwise.
- Higher R means greater diversity
- Very dependent upon depth of sampling and sensitive to presence of rare species

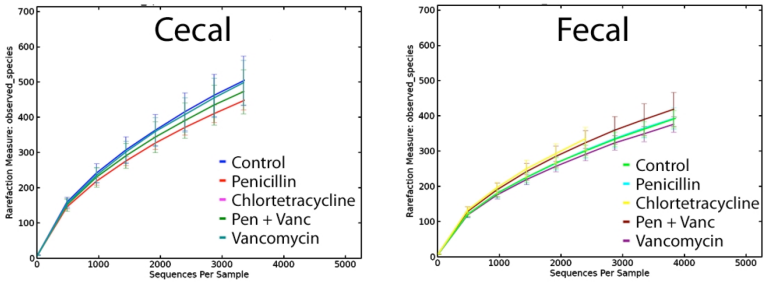
19

Rarefaction curves

- Note: rarefication as a means for normalization is from statistical standpoint a bad idea. Don't throw away information!
- Rarefaction curves are not the same!
- Useful to assess sensitivity of sample size to observed alpha-diversity estimates.
- Idea:
  - Let $N_1, \ldots, N_K$ be a set of numbers $N_i < N_{i+1}$;
  - Let $n'_{ij}{}^{(k)}$ be abundance of taxon i in sample j subsampled to N_k total counts per sample;
  - Estimate average alpha diversity for each N_k over a several repeated subsamplings;
  - Plot the average alpha diversity as a function of sample size.

20

Rarefactions



Supplementary Figure 6. Rarefaction curves measuring alpha diversity in fecal and cecal communities. The vertical axis shows the number of OTUs observed after sampling the number of tags or sequences shown on the horizontal axis. Curvature toward horizontal indicates that increased sequencing effort is required to observe novel OTUs, when only rare OTUs remain to be discovered. Rarefaction curves were based on the V3 16S rRNA sequences and analyzed at OTU-level phylotypes, defined by ≥97% identity. Values represent the Mean ± 95% confidence interval.

Cho, I., Meth, BA., Nondorf, L., Li, K., Alekseyenko, AV., Blaser, MJ. "Subtherapeutic antibiotics alter the murine colonic microbiome and early life adiposity", Nature 488, 621 -- 626 (30 August 2012).

21

Chao1 index

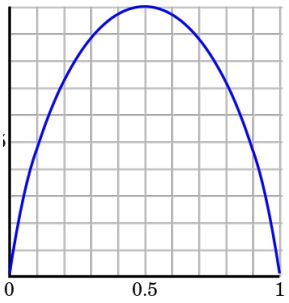
- Species richness index is often too sensitive to depth of sampling,
- Chao1 index overcomes this problem by applying a correction
- $R_C = S_{obs} + (\frac{f_1^2}{2f_2})$,
- Where f_1 is the number of taxa with a single observation (singletons), f_2 is the number of taxa with exactly two observations.
- If a sample contains a lot of singleton taxa, then there is a greater chance that this sample is undersampled.

Shannon index

- Suppose we observe a community that can contain up to k 'species'.
- The relative proportions of the species are P = {p_1, …, p_k}.
- Shannon index is related to the notion of information content from information theory. It roughly represents the amount of information that is available for the distribution of P.
- When $p_i = p_j$, for all i and j, then we have no information about which species a random draw will result in. As the inequality becomes more pronounced, we gain more information about the possible outcome of the draw. The Shannon index captures this property of the distribution.
- Shannon index is computed as

  $S_k = -p_1\log_2 p_1 - p_2\log_2 p_2 - \ldots - p_k\log_2 p_k$

  Note as $p_i \rightarrow 0$, $\log_2 p_i \rightarrow -\infty$, we therefore define $p_i\log_2 p_i = 0$.
- Higher S_k means higher diversity



23

From Shannon to Evenness

- Shannon index for a community of k species has a maximum at $\log_2 k$
- We can make different communities more comparable if we normalize by the maximum
- Evenness index is computed as

  $E_k = S_k/\log_2 k$
- E_k=1 means total evenness

24

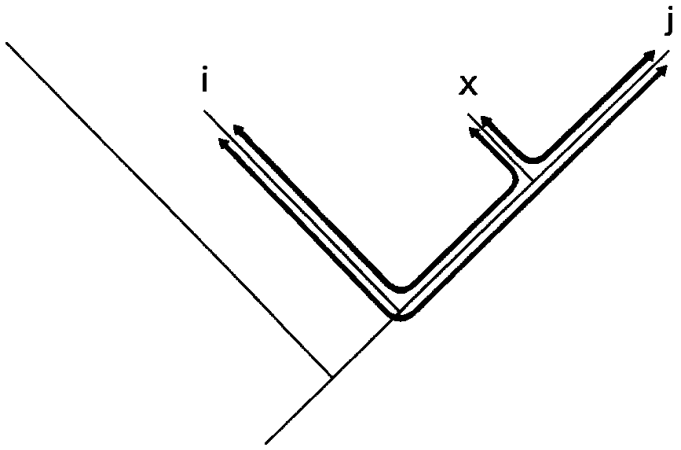
Simpson index

- Suppose we observe a community that can contain up to k 'species'.
- The relative proportions of the species are P = $\{p_1, \ldots, p_k\}$.
- Simpson index is the probability of resampling the same specie on two consecutive draws with replacement.
- Suppose on the first draw we picked specie i, this event has probability p_i, hence the probability of drawing that species twice is $p_i * p_i$.
- Simpson index is thus computed as
  $D=1 - (p_1^2 + p_2^2 + \ldots + p_k^2)$
- D = 0 means no diversity (1 species is completely dominant)
- D = 1 means complete diversity

25

Phylogenetic Diversity (Faith's D)

- Faith (Biological Conservation 1992, 61, 1-10) considered the problem of selecting species for conservation so as to preserve diversity.
- Faith defines PD (phylogenetic diversity) as the sum of all the branch lengths. PD is analogous to total information in the tree.
- The marginal contribution of a tip x is then $\min_{i,j}(D_{x,i} + D_{x,j} - D_{i,j})$. Higher value suggest a greater impact on conservation.



26

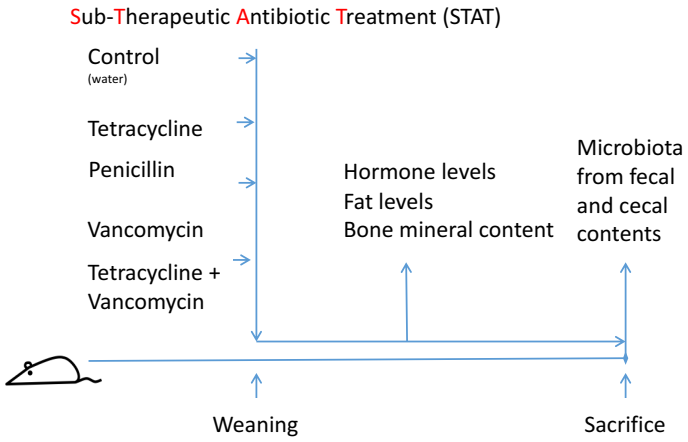
Numbers equivalent diversity

- Often it is convenient to talk about alpha diversity in terms of equivalent units:
  - How many equally abundant taxa will it take to get the same diversity as we see in a given community?
- For richness there is no difference in statistic
- For Shannon, remember that $\log_2 k$ is the maximum which is attained when all species are equally represented. Hence the diversity in equivalent units is 2^{Sk}
- For Simpson the equivalent units measure of diversity is 1/(1-D)

27

How to compare microbiomes?

Motivating example

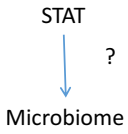


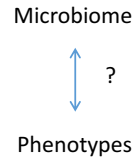
Cho, I., Meth, BA., Nondorf, L., Li, K., Alekseyenko, AV., Blaser, MJ. "Subtherapeutic antibiotics alter the colonic microbiome and early life adiposity in mice". Nature. 2012 Aug 30;488(7413):621-6.

29

Questions

- Are there any specific taxa, which are associated with antibiotic treatment?
  - By presence/absence patterns
  - By relative abundance
- Is there correlation between abundance of any taxa and metabolic phenotypes (hormone levels, fat, bone)?





30

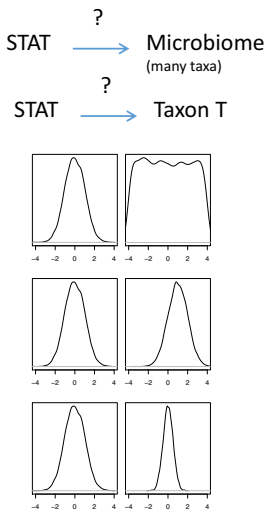
Hypotheses

- Are **precise** statements that are amenable to being proven false using data.
- *Null hypothesis*: a proposition that corresponds to default position. ("Nothing special is happening")
- *Alternative hypothesis*: a proposition that describes a non default outcome ("Something *interesting* is going on")
- The inference is obtained by rejecting the Null hypothesis. Null hypothesis can never be confirmed by the data, nor does it have to be!

31

Example of hypotheses

- General question: Are any taxa associated with antibiotic treatment?
- Univariate hypothesis question: Is taxon T associated with antibiotic treatment?
- Null hypothesis: abundance of taxon T follow the same distribution in treated and control groups.
- Alternative hypothesis 1: abundance of taxon T follow distribution of different *form* in the two groups.
- Alternative hypothesis 2: abundance of taxon T follow the same form of distribution but with different *mean/median* between groups.
- Alternative hypothesis 3: abundance of taxon T follow the same form of distribution but with different *variance* between groups.



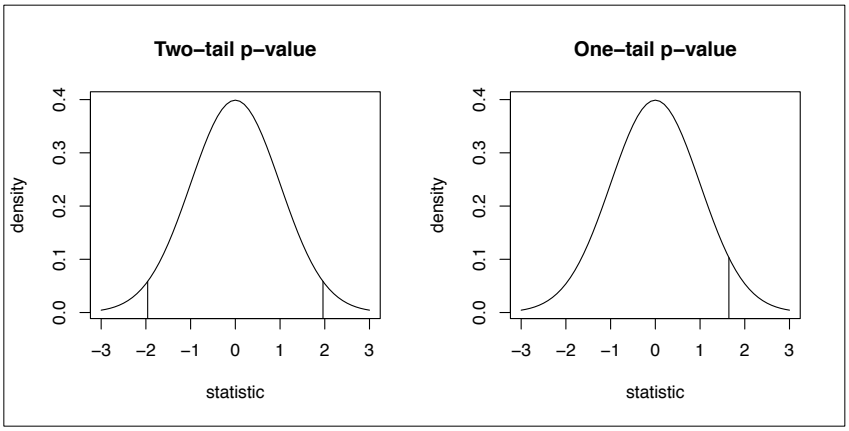
32

Flow of statistical inference under hypothesis testing

- Define a test statistic that evaluates evidence against the Null Hypothesis;
  - What is a good statistic to compare the averages of two samples: x1, …, xN and y1, …, yM? What is the null hypothesis here?
- Determine the distribution of the test statistic under the Null Hypothesis;
  - Options here:
    - Asymptotic properties of the statistic;
    - Monte Carlo simulations: bootstrap, permutation, …
  - How would the distribution of statistic above look like under the Null?
- Calculate the test statistic value in the observed data;
- Compare the observed test statistic to the distribution of the statistic, when the null hypothesis is true.
  - If the probability of observing a statistic as extreme or more is small enough (P<0.05?), reject the null hypothesis.

P-values

- If the Null Hypothesis was in fact true a *statistic,* used to perform the test, would follow a certain distribution: the *null distribution*.
- P-value is the tail probability under the null distribution.



34

Distribution of OTU/ASV abundance data

- *Justifiable* distribution assumptions often allow for better statistical tests.
- Properties of OTU/ASV abundance data:
  - Correlated: Sums to 1, hence to increase something, something else has to decrease
  - Variable across subjects
- Can *possibly* be modeled through compound Dirichlet-Multinomial distribution (we will talk about this distribution later in the course). Marginal univariate (Beta-binomial) tests have to be described.
- When distribution specific tests are not available, we have to rely on non-parametric (distribution free) tests, possibly at the cost of decreasing the power of the tests.

35

Some practical statistical tests to try with your microbiomes…

Chi Squared test for taxon incidence

- Raw Counts

	Sample 1	...	Sample N
	n_{11}		n_{1N}
	n_{21}		n_{2N}
	n_{31}		n_{3N}
	n_{41}		n_{4N}
	n_{51}		n_{5N}
	n_{61}		n_{6N}
	n_{71}		n_{7N}
	$n_{\cdot 1}$		$n_{\cdot 1}$

- Incidence table

	Sample 1	...	Sample N
	1_{11}		1_{1N}
	1_{21}		1_{2N}
	1_{31}		1_{3N}
	1_{41}		1_{4N}
	1_{51}		1_{5N}
	1_{61}		1_{6N}
	1_{71}		1_{7N}

$$1_{ij} = \begin{cases} 1, \text{ if } n_{ij} > 0 \\ 0, \text{ otherwise} \end{cases}$$

37

Chi Squared test for taxon incidence

- We focus on a single taxon
- Suppose the observations of the taxon come from two groups (e.g. control vs. STAT)
- Question: Is the frequency of occurrence of this taxon in two groups different?
- Null hypothesis: the frequency is the same.
- Significant Chi Square test indicates a difference in the *rate* of occurrence of the taxon.
- In R: `chisq.test`

Tax	Lab
1 or 0	Control
...	...
1 or 0	STAT

	Control	STAT	
Present	n_{11}	n_{21}	$n_{2\cdot}$
Absent	n_{12}	n_{22}	$n_{1\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	N

38

Mann-Whitney U or Wilcoxon rank-sum two-sample test

- Assumptions:
  - Independent observations
  - Observations can be ordered with respect to each other
- Null hypothesis: The distribution in two samples is the same. If one randomly draws one observation from each sample X, Y; then Pr(X>Y) = Pr(Y>X)
- Two-sided alternative hypothesis: Pr(X>Y) ≠ Pr(Y>X)
- Interpretation: for continuous observations, significant tests indicate change in the median
- Example: Is the abundance of a taxon different between STAT and Control?
- In R: `wilcox.test`

39

Connection with predictivity

- Mann-Whitney U-statistic calculation:
  - Convert the observations to ranks
  - Compute the sum of ranks in each sample, R_1 and R_2
  - $U_1 = R_1 - n_1(n_1 + 1)/2$
  - $U_2 = R_2 - n_2(n_2 + 1)/2$
  - $U = \min(U_1, U_2)$
- One can show that U statistic is equivalent to AUC. $AUC = U/(n_1\, n_2)$
- AUC, area under receiver operator characteristic (ROC) curve, measures how well we can distinguish one sample from another. AUC = 0.5 means predictivity no better than random, AUC = 1.0 perfect predictivity.

Sample 1	Sample 2	Ranks 1	Ranks 2
0.135	2.680	8	1
-0.907	1.078	18	2
-0.801	0.080	16	9
0.452	0.493	6	5
-0.523	0.010	15	11
0.075	-0.322	10	13
1.038	-0.370	3	14
-1.140	0.633	19	4
-2.308	-0.020	20	12
-0.808	0.368	17	7
	Rank Sums	132	78
	U	77	23
	U statistic	23	
	AUC	0.77	0.23

40

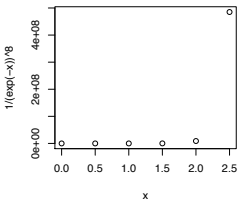
Kruskal-Wallis one-way analysis of variance (more than two samples/groups)

- Assumptions:
  - Independent observations that follow distribution with the same shape and scale
  - Observations can be ordered with respect to each other
- Null hypothesis: The location (median) of all the groups is the same.
- Alternative hypothesis: Location for at least one group is different from location of at least one other group
- Example: Is the abundance of a taxon different in STAT/control over 3 sampled time points?
- In R: `kruskal.test`

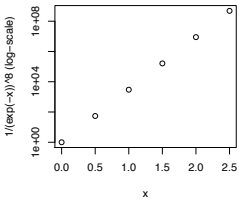
41

Correlation coefficients, rank correlations

- Linear correlation coefficient (Pearson) assumes linear dependence between two variables
- Rank correlation coefficient measure the extent of monotonicity between two variables
- Null hypothesis for correlation testing: correlation coefficient is equal to 0.



Pearson correlation coefficient: 0.66 (not significant, p=0.15)



Diaconis, P. (1988), Group Representations in Probability and Statistics, Lecture Notes-Monograph Series, Hayward, CA: Institute of Mathematical Statistics, ISBN 0-940600-14-5

42

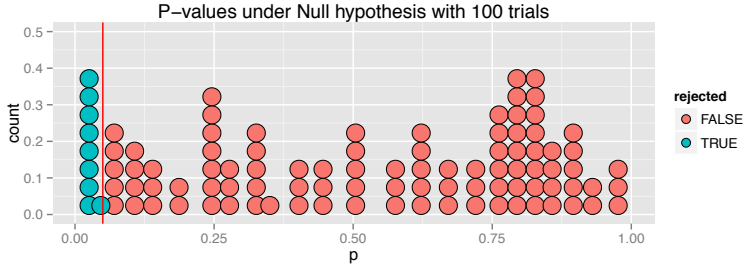
Rank correlation coefficients

- Spearman's ρ: Rank correlation measure defined as the Pearson correlation of the two variables after conversion to ranks
- Kendall's τ: Rank correlation measure based on counting concordant pairs. [(x_1,y_1) and (x_2,y_2) are concordant if $x_1>x_2$ when $y_1>y_2$]
- Example: Is there correlation between any given two taxa? Is there correlation between a given metabolic variable and a given taxon?
- In R:
  - `cor.test(x, y, method='spearman')`
  - `cor.test(x, y, method='kendall')`

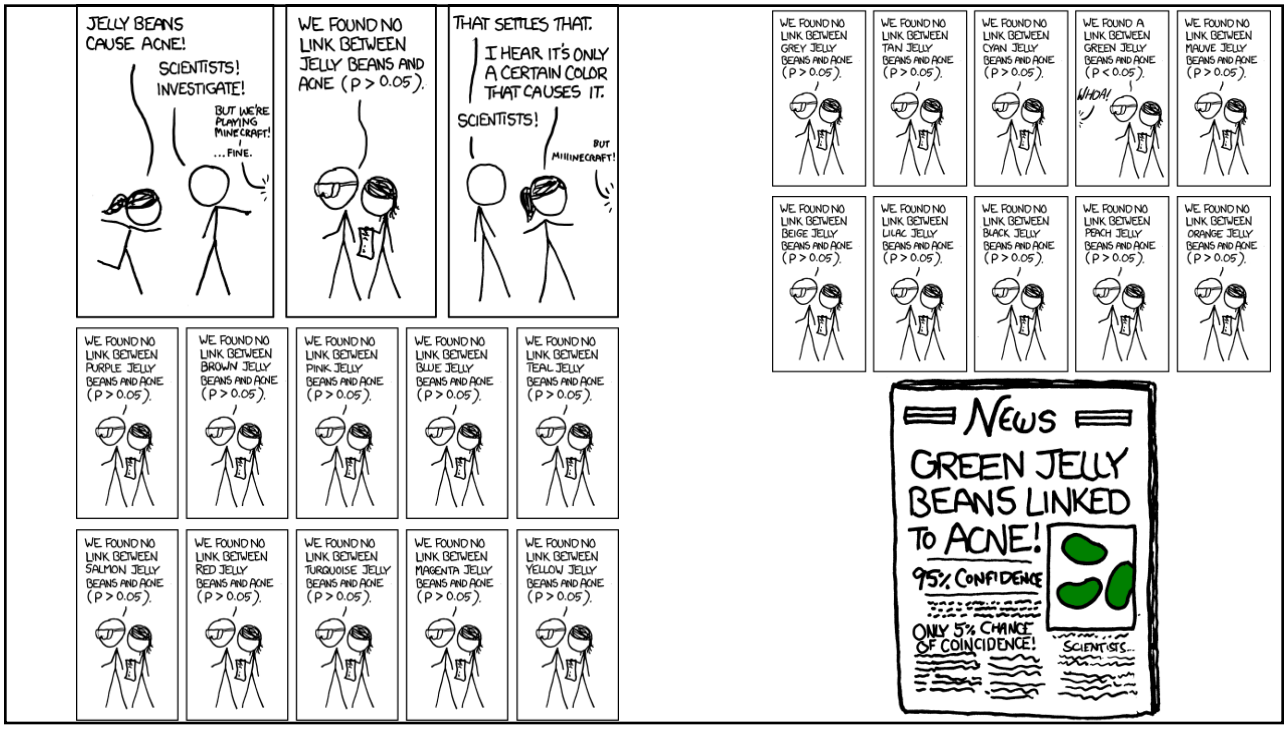
43

Problems with testing many hypotheses simultaneously

- We have many OTUs/ASVs/taxa that we would like to apply the test to.
- If the test is applied at specified significance level (probability of falsely rejecting the null, when it is true), we cannot guarantee that combined result is at the significance level originally specified.
- Since p-values are distributed uniformly if the null hypothesis is true, the expected number of rejections by mere chance m*α
- How do we control significance for multiple tests?



44



FWER: Family-wise error rate

	# not-rejected	# rejected	Total
# true null hypotheses	U	V	m_0
# non-true null hypotheses	T	S	$m-m_0$
Total	m-R	R	m

FWER control methods adjust the significance of each individual test to ensure overall significance at given α.
FWER result in more stringent tests.

- Suppose we perform m tests (e.g. m taxa are tested for association with antibiotic treatment)
- The number of true null hypotheses is unknown m_0
- V is false positive rate (Type I error)
- T is false negative rate (Type II error)
- We observe R, but S, T, U, V are unobserved
- FWER = Pr(V≥1)

46

Example: Bonferroni correction

- To ensure overall significance at a given α, one performs each individual test at α' = α/m
- Very stringent, results in loss of power (increase in Type II error)

47

FDR: false discovery rate

- Modifies the idea of controlling Type I error, to instead control the rate at which type I errors do occur
- FDR is the expected value of V/R
- Methods for FDR control
  - Benjamini–Hochberg
    - Assumes tests are independent
  - Benjamini–Hochberg–Yekutieli
    - Assumes that tests are uniformly correlated:
      - Positively correlated: if one test has low p-value, other tests are *more* likely to also be significant
      - Negatively correlated: if one test has low p-value, other tests are *less* likely to be significant

	# not-rejected	# rejected	Total
# true null hypotheses	U	V	m_0
# non-true null hypotheses	T	S	$m-m_0$
Total	m-R	R	m

48

FDR in R

- FDR is implemented in R as a p-value adjustment procedure.
- Input: p-values for a set of univariate tests
- Output: p-values that are adjusted to FDR
- E.g. 0.05 adjusted p-value means that expected rate of false positives is 0.05 for tests significant at that adjusted level
- `p.adjust`
  - Methods:
    - `method = 'fdr':` Benjamini-Hochberg
    - `method = 'BY':` Benjamini-Hochberg-Yekutieli

49

Filtering: reducing the number of tests

- We can improve the overall power of the tests by performing less simultaneous tests.
- Eliminate "uninteresting" taxa, e.g. a taxon does not have deep taxonomic resolution.
- Eliminate taxa that show low variability. These are not changing much overall thus are not likely to be different across factor levels.
- Eliminate taxa with low abundance. These are usually not measured very well and are likely to have little biological significance anyway.
- Note: A care needs to be taken with filtering procedures so as not to introduce selection bias, which will invalidate multiple comparison assumptions. A safe practice is for filtering to be blind towards the factor you would like to test.

50