Microbiome Discovery

Alpha Diversity

Alpha diversity: 单个样本/环境存在哪些species, 有多少种species存在

Beta diversity: 样本间species的相似性或差异性

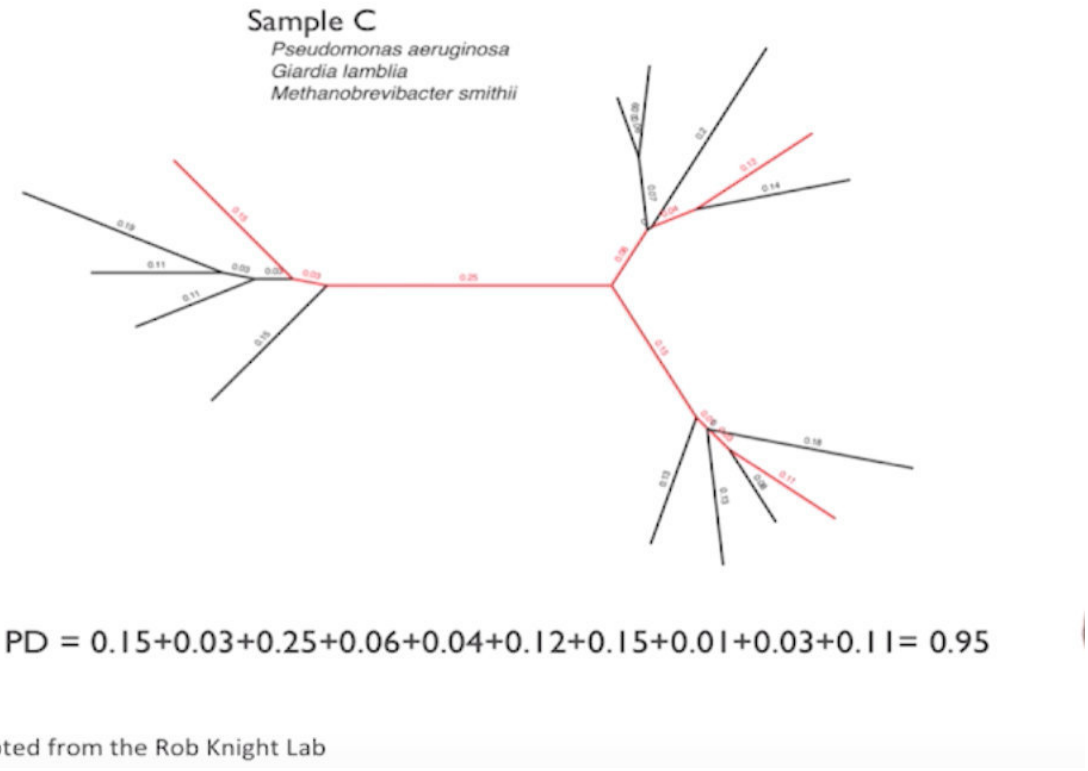
Measure alpha diversity:

Method 1: Species count, counts how many species there; we often use 97% OTUs to represent specie(使用97%OTU代表物种计算物种数量从而表示多样性).

Simple species count ignores that It doesn't take into account how closely related different species are to each other(忽略了样本中物种间的相似性).

Method 2: Phylogenetic diversity, takes into account the level of evolutionary diversity or melecular diversity within a sample(使用物种进化多样性得到多样性).

Add up the total brance length that occupied of the sample and that gives you measure(累加系统发育树中单个样本的分支长度得到其多样性).



Method 3: Chao1 Estimator

Try to predict how many species actually are in a sample/community given that you have a finit sample of members of the community(尝试得到一个sample/community实际含有多少物种)

$$S_1 = S_{obs} + \frac{F_1^2}{2F_2}$$

Number of singletons (only saw one example)

Number of doubletons (saw exactly two examples)

$$var(S_1) = F_2\left[\left(\frac{(F_1/F_2)}{4}\right)^4 + (F_1/F_2)^3 + \left(\frac{F_1/F_2}{2}\right)^2\right]$$

No one uses the variance.
No one has tested this empirically in microbiome data.

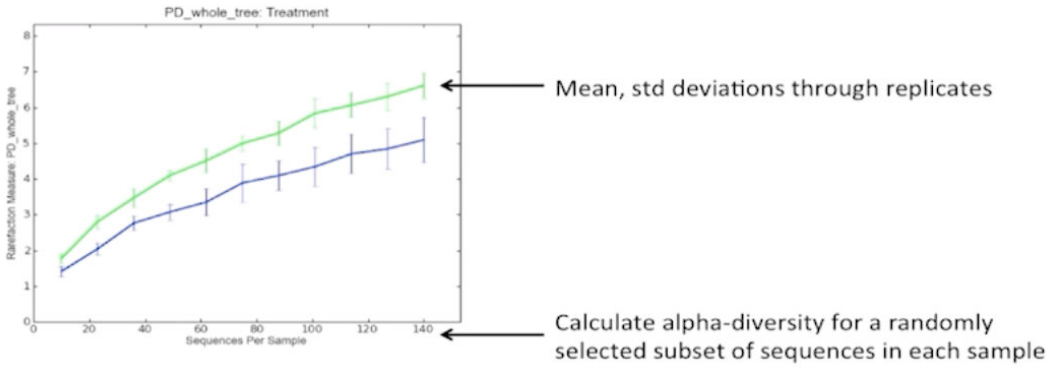
Can test using the Yatsunenko et al. data...

http://palaeo-electronica.org/2011_1/238/estimate.htm

Rarefaction: Did you go deep enough?

Calculate alpha-diverisyt for a randomly selected subset of sequences in each sample(随机挑选部分序列计算物种多样性, 10% M次/M 得到平均值...; 20%..... 若曲线持续上升, 则证明获得数据量不足以代表实际物种多样性)

Rarefaction: Did you go deep enough?



Summary:

Alpha diversity measures diversity $within$ communities

Beta diversity measures diversity $between$ communities

Phylogenetic diversity(PD) is worth trying

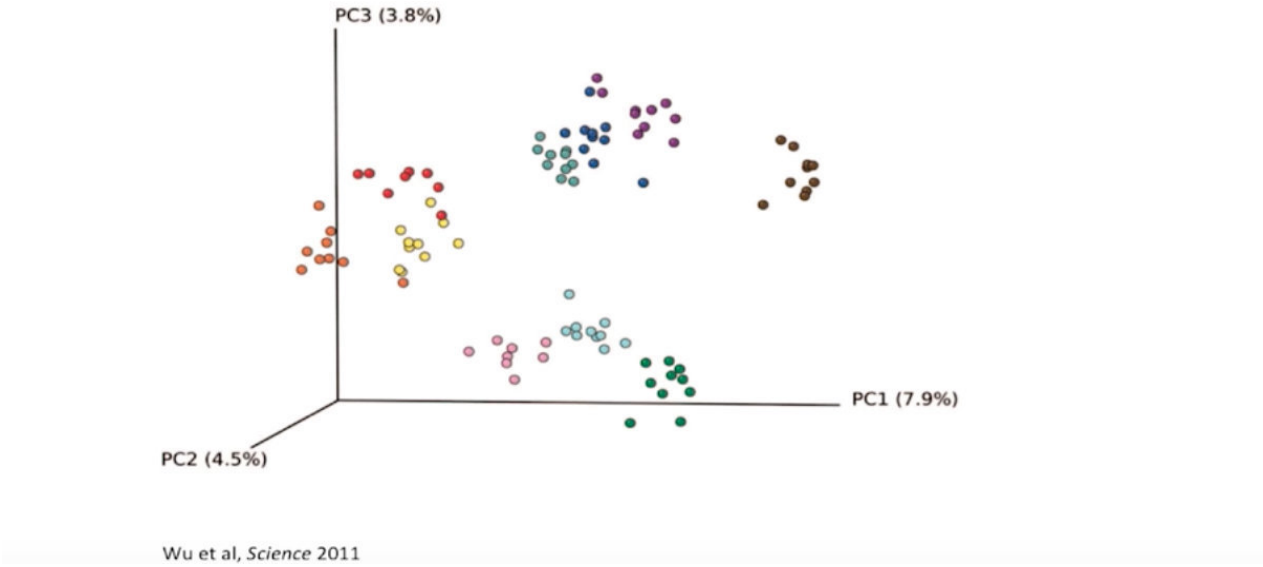
Most people use PD, Chao1, Shannon and OTU count, and there is prefect correlation between each other

Rarefaction determines saturation

There is room for experimental validataion(to compare your data to existed deeply sequenced data)

Beta diversity

Beta-diversity: measure of overall change



Euclidean: the most 'dangerous' distance(just the actual distance in space between two samples, 为空间内两点多实际距离)

- Euclidean distance

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

PcoA: pricipal corrodinatest analysis

The repository in GitHub: http://metagenome.cs.umn.edu/microbiomecodebrowser/doc/index.html

Chi-square: wroks great for gradients(很适合梯度数据)

- Chi-square distance

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

$$d(p,q) = \sum_{i=1}^{n} \frac{(p_i - \mu_i)^2}{\mu_i} + \sum_{i=1}^{n} \frac{(q_i - \mu_i)^2}{\mu_i}$$

Observed

Sample1	Sample2	
10	20	30
50	40	90
60	60	

Expected

Sample1	Sample2
15	15
45	45

Chi-squared

Sample1	Sample2
-5	5
5	-5

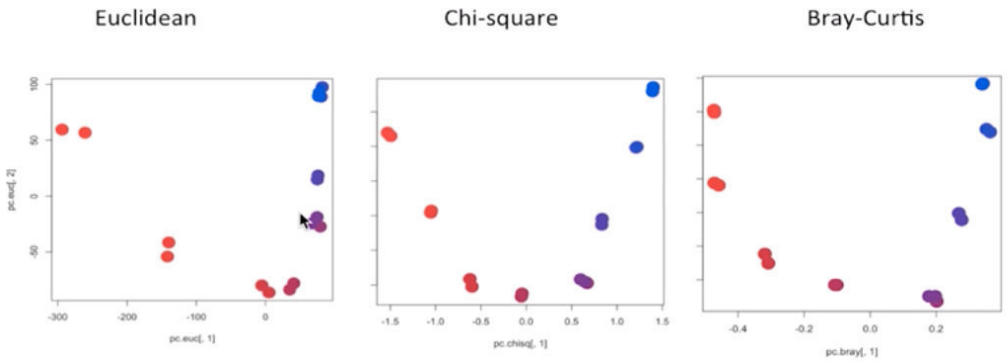
Chi2 = 100

Bray-Curtis: undershoot, 脱靶; It is a nice ecological distance

- For each species, find which sample has the lower value
- Take the sum of these lower values
- Divide by the total counts for all species
- i.e. how much does one sample “undershoot” the other

$$BC_{ij} = \frac{2C_{ij}}{S_i + S_j}$$

Comparison: Guerrero Negro



Summary:

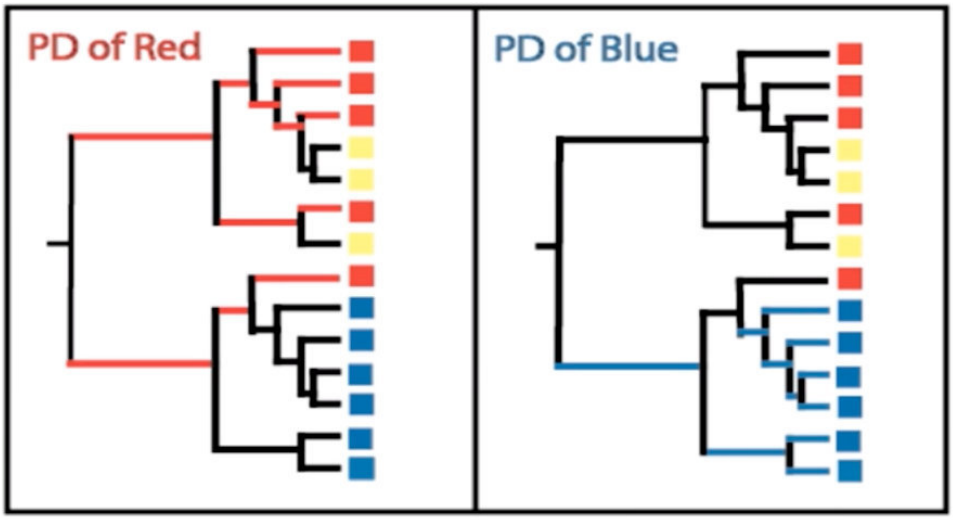
Beta diversity measures diversity $between$ communities

Most people use Bray Curties or UniFrac

UniFrac

Euclidean, Chi-square, Bray-Curtis don't use phylogenetics

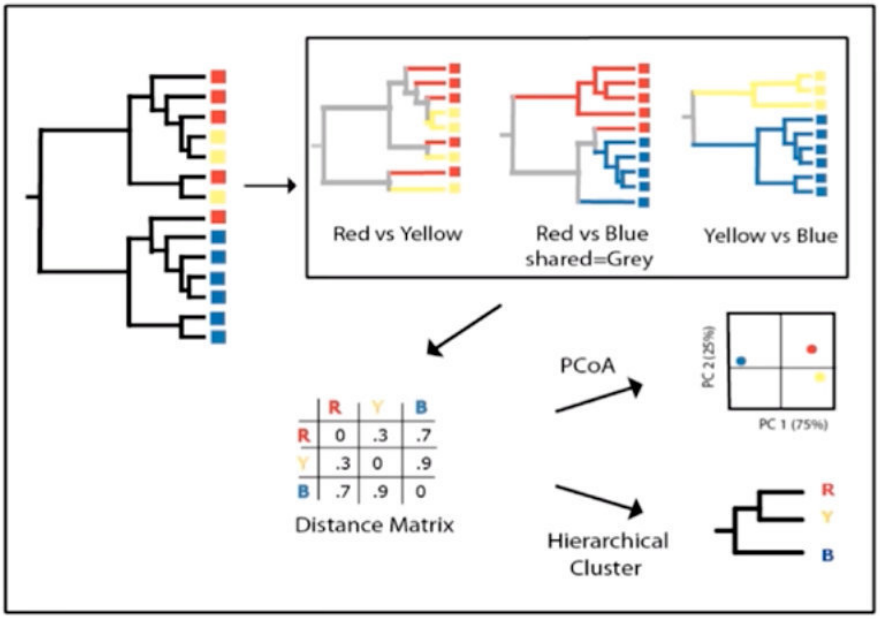
Phylogenetic-based alpha diversity: sum of branch lengths covered by a sample(**一个样本中所有分支的长度之和**)



Phylogenetic-based alpha diversity: sum of branch lengths covered by a sample

Unifrac: Phylogenetic-based beta diversity Percent of observed branch length unique to one or the either sample according to its **species**(**一个sample/community观察到的唯一的分支长度的比率**; 下图中的two communities: 左图, red/blue samples都拥有一样的物种, each sample/community has bugs in the exact same places in the tree of life; 中图, half of the phylogenetic tree is unique to the red sample or the blue sample; 右图, red sample/community are completely unrelated from all the bugs in the blue sample/community, the whole evolutionary tree is unique to one smaple of the other; 意味着上下各是两个不同的大类(branch), 虽然这里仅看的是最底端的物种(tips, species))

Beta diversity using UniFrac: Using unifrac distance between every pair of samples in data set that gives you a distance matrix(分别计算两两之间距离, 获得距离矩阵, 针对该矩阵分析 PCoA/Hierachical Cluster...)



Wighted Unifrac: take into account relative abundances

Weighted UniFrac weights the branch lengths by the abundances of bugs, emphasizes the dominant bugs; Unweighted UniFrac only uses presence/absence, emphasizes the minor bugs(in general)(**W-UniFrac强调了主要bugs/species, 而UW-UniFrac强调了次级的bugs/species**)

Summary:

Beta diversity measures diversity $between$ communities

UniFrac(phylogenetic beta diversity) is very useful

Most people use UniFrac and Bray Curtis

Chi-square is often best for gradients(but not phylogenetically informed...reaserch project?)

Statistical testing part 1

Enterotypes: how did it happen?

The statistical support for clusters was based on simulated data

Microbiome data are extremely hard to simulate accurately

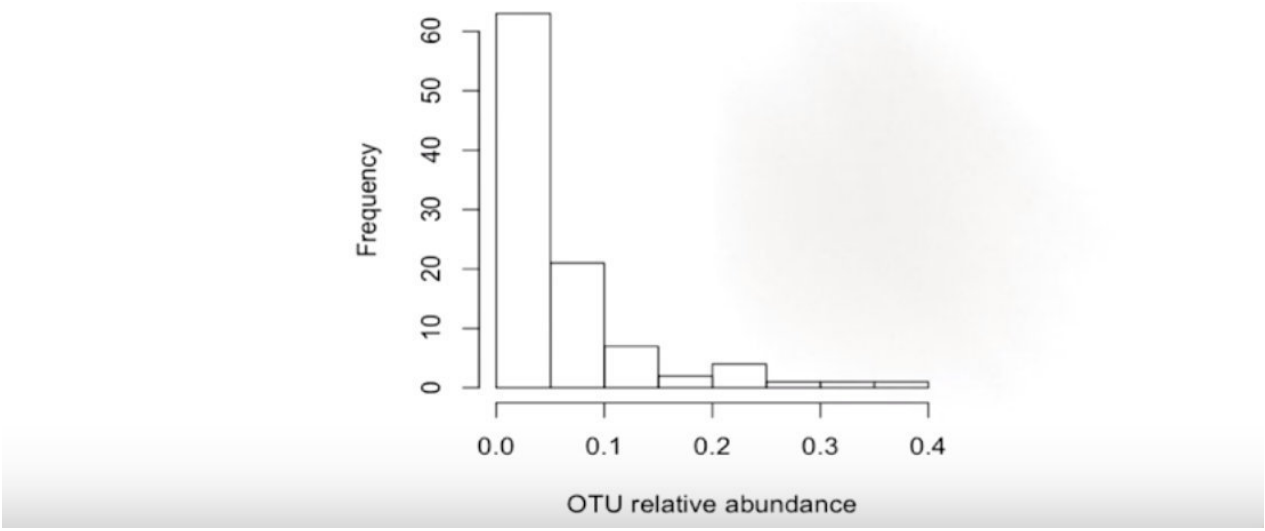
Many reviewers may not recognize this as a problem

Species not normally distributed

Often zero-inflated

Often like a negative binomial

Not like a normal distribution

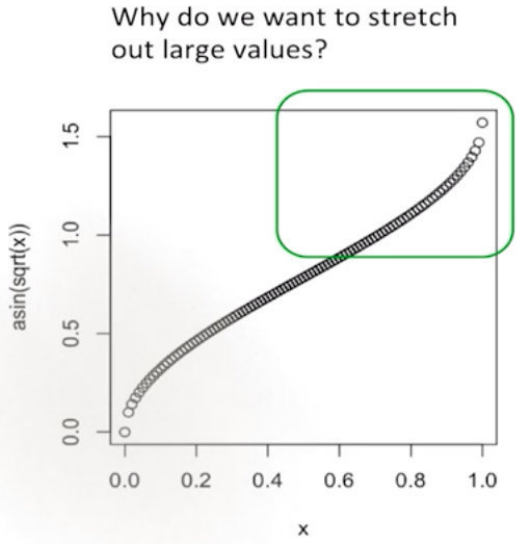


Data transform:

Arcsin-square root transform(相比压缩中间的数值, 延伸大和小的数值).

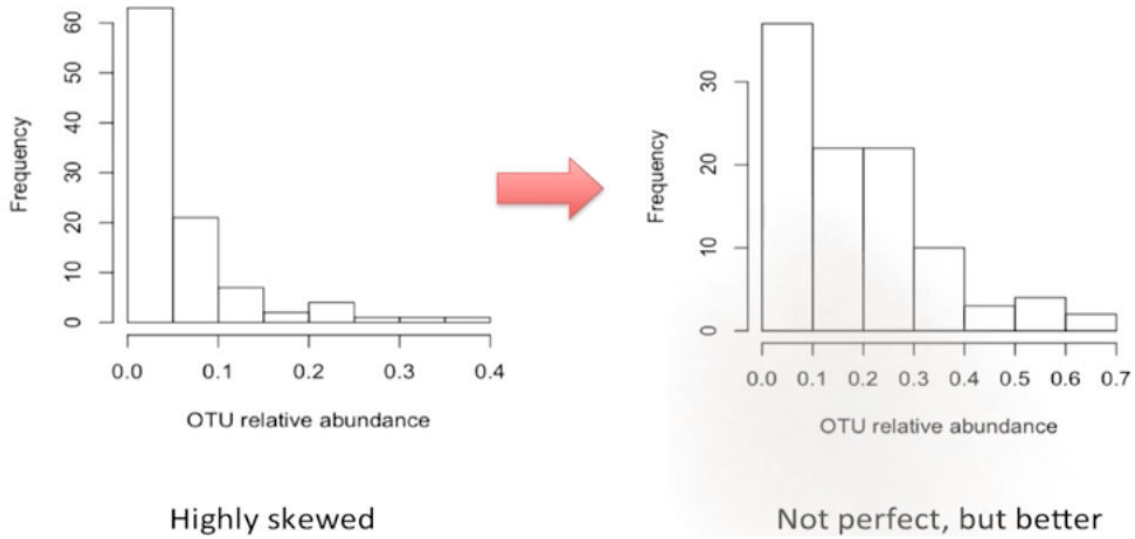
- Some prior use in ecology
- Spreads out the small numbers more than the big numbers (like log transform)
- Can handle 0 (unlike log transform)

$$\sin^{-1}\sqrt{y_i}$$



Square root is better(延伸小的数值, 压缩大的数值; 因为大的数值更倾向于改变(异质性), 同时有助于放大次级的数值)

Square root improves distributions(log transform works too, but does not handle zeros)



Highly skewed

Not perfect, but better

Common parametric tests:

t-test(t检验): compare 2 groups

ANOVA(方差检验): compare three or more groups

Correlation: Compare to a continuous variable(e.g. Age)

Linear Regression: Similar to correlation, but you can regress on multiple variables at the same time; and add some comfounding parts

Note: all of these assume normal distributions!

Statistical testing part2

Linear models are not always appropriate

Generalized linear models(better underlying distributions)

Non-parametric tests(no distribution assumptions)

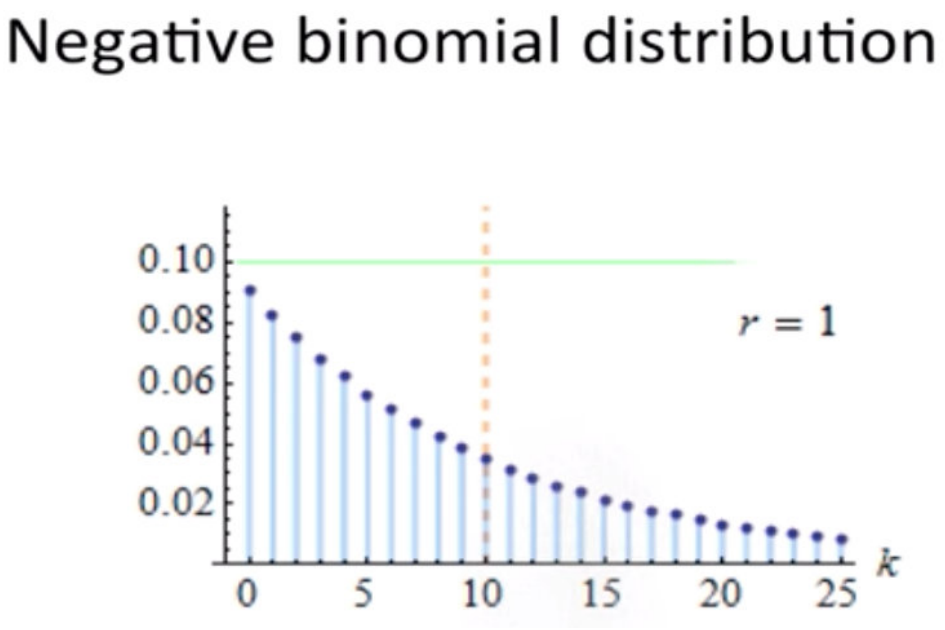
Controlling for multiple tests

- For one test, use alpha = 0.05
- For many tests:
  - For every tests, 5 will appear significant by chance
  - Bonferroni correction is most strict
    - Divide alpha by # tests
    - Controls the probability of having one or more false positive
  - False Discovery Rate(FDR) more lenient, common
    - Slightly more complex formula
    - Guarantees expected rate of false positive

Negative binomial distribution

"number of successes in a sequence of independent and identically distributed bernoulli trials before a specificed(non-random) number of failures(denoted r) occurs" 在获得5次非1前得到一个1, 需要掷骰子几次

例如, 大多数人拥有很少的bug数目, 一些人拥有中等数目bug, 很少的人拥有很多的bug



改变参数R(为成功次数), 接近正态分布: