Novogene 诺禾致源

Providing advanced genomic solutions!

R语言与数理统计、高级绘图及ggplot2

王　鹏

2018.8

目录

CONTENTS

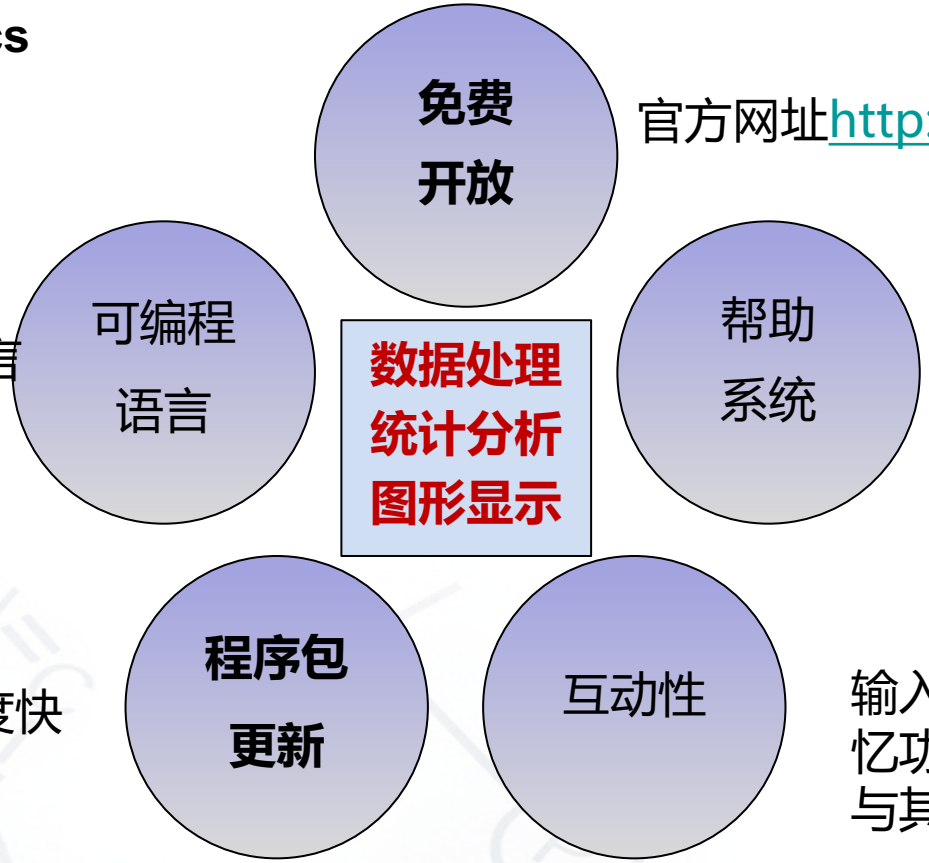
起源与特点

S语言
贝尔实验室

Ross Ihaka和Robert Gentleman开发

"R开发核心团队"

Statistics



官方网址http://www.r-project.org

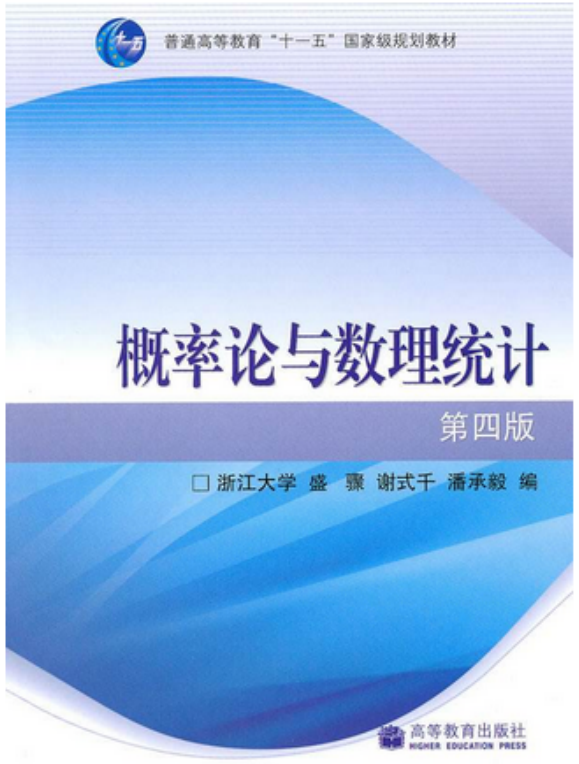
编制自己的函数、扩展现有的语言

全世界最大、最前沿的统计学家思维的聚集地

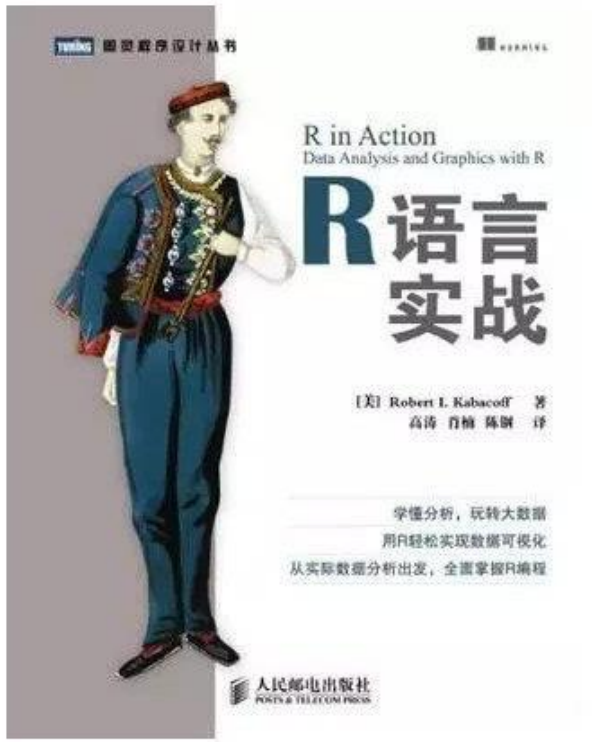
封装好、种类多、功能强、速度快

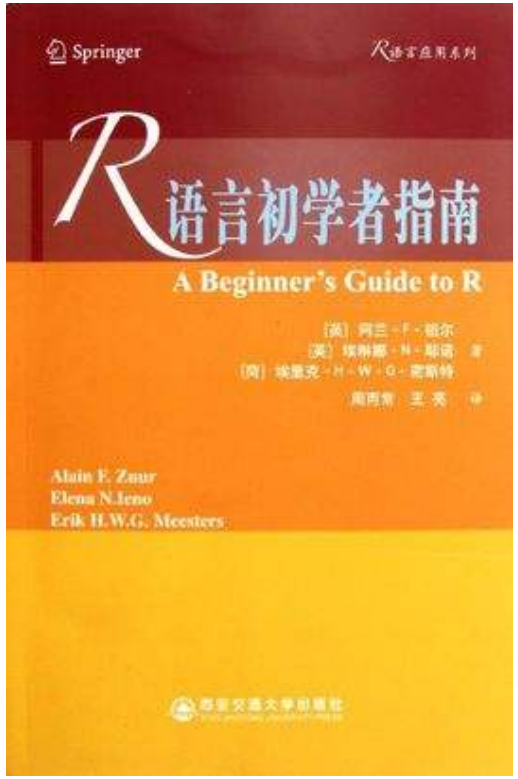
输入输出窗口+图形输出窗口、错误提示、记忆功能
与其他编程语言和数据库之间有很好的接口

参考资料

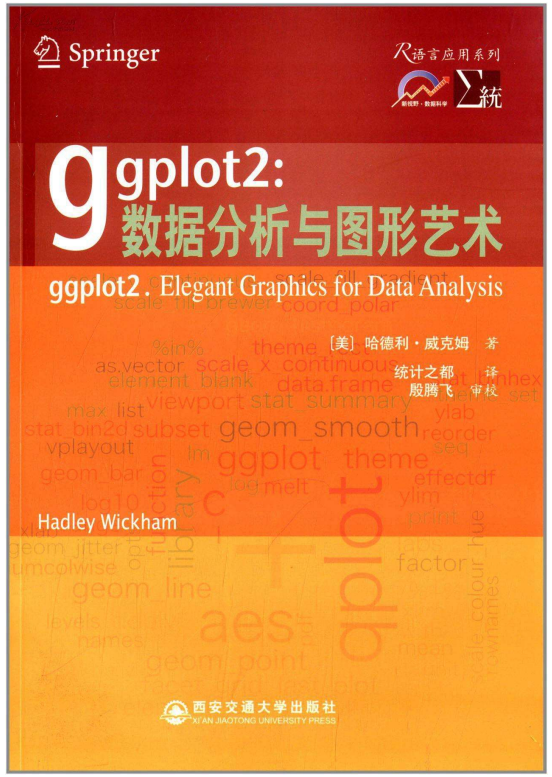


数理统计基础





R入门（数据分析和绘图）



高级绘图

数理统计

描述性统计 ⊕ 推断性统计

概率论

描述性统计：数据收集、处理、汇总、图表描述概括与分析等统计方法。

推断性统计：利用样本数据来推断总体特征的方法。

集中趋势：平均数、中位数、众数

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

离散趋势：方差、标准差

$$\operatorname{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2,$$

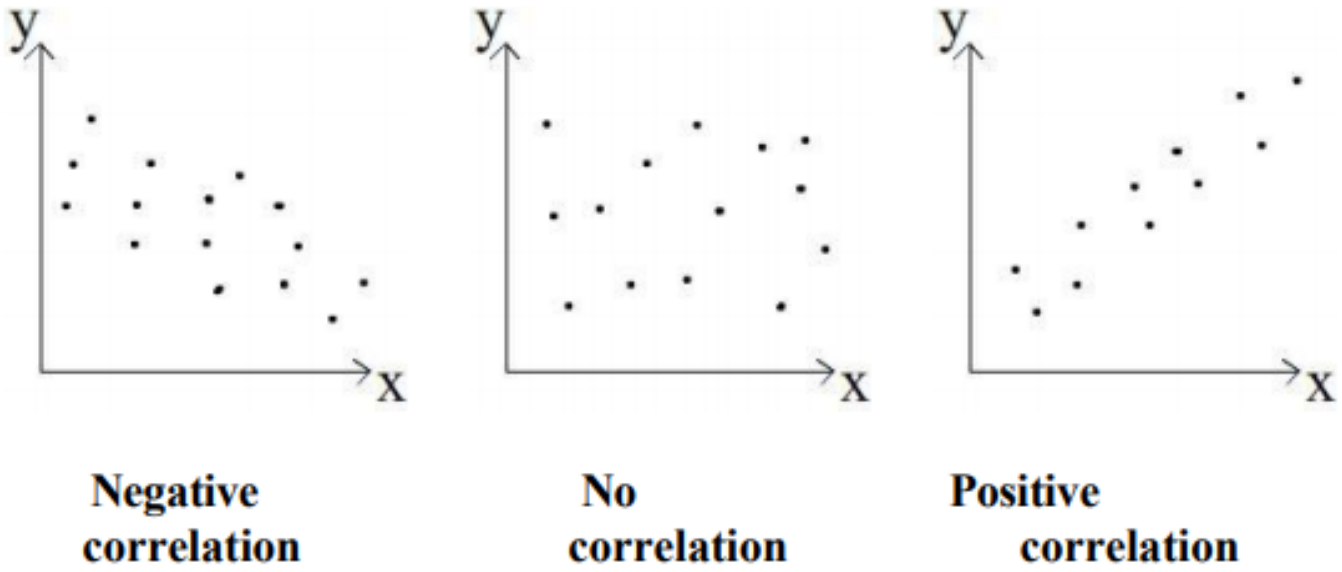
直方图、箱线图、柱形图、线图、茎叶图、饼图

常用统计函数

函数	功能
max(x)	最大值
min(x)	最小值
mean(x)	均值
median(x)	中位数
range(x)	数值的范围
which.max(x)	最大值下标
which.min(x)	最小值下标
var(x)	方差
sd(x)	标准差
sum(x)	总和

函数	功能
abs(x)	取绝对值
sqrt()	开方，平方如2^3
ceiling(x)	向上取整
floor(x)	向下取整
round(x,digits=n)	保留小位数
log(x,base=n)	以n为底，x的对数
log(x)	以e为底，x的对数
exp(x)	自然对数
cos、sin、tan(x)	余弦、正弦、正切
acos、asin、atan	反余弦、反正弦、反正切

相关系数



Pearson相关 ⊕ **Spearman相关** ⊕ **Kendall相关**

1、只表相关，而不表因果 2、相关亦有强弱之分 3、不同数据分布，适用条件不同

相关系数1

Pearson's correlation coefficient (PCC)

总体相关系数

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

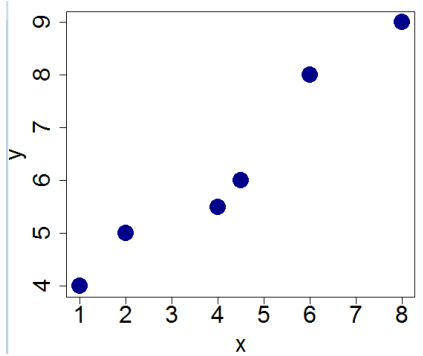
where:

- n is the sample size
- x_i, y_i are the individual sample points indexed with i
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

样本相关系数

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y},$$

```
x <- c(1,2,4,4.5,6,8)
y <- c(4,5,5.5,6,8,9)
cor(x,y,method='pearson')
plot(x,y,cex=3,col='darkblue',pch=19,cex.axis=2,cex.lab=2)
```



相关系数2

Spearman's rank correlation coefficient (SCC)

$$r_s = \rho_{\mathrm{rg}_X, \mathrm{rg}_Y} = \frac{\mathrm{cov}(\mathrm{rg}_X, \mathrm{rg}_Y)}{\sigma_{\mathrm{rg}_X} \sigma_{\mathrm{rg}_Y}}$$

where

- ρ denotes the usual Pearson correlation coefficient, but applied to the rank variables.
- $\mathrm{cov}(\mathrm{rg}_X, \mathrm{rg}_Y)$ is the covariance of the rank variables.
- σ_{rg_X} and σ_{rg_Y} are the standard deviations of the rank variables.

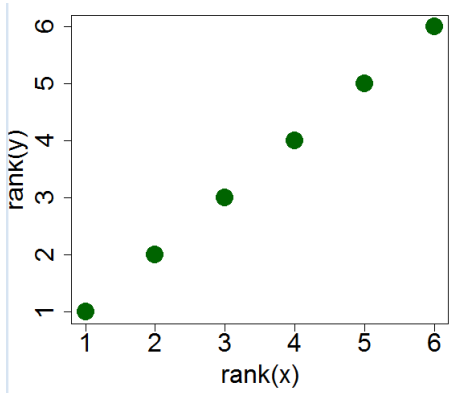
$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}.$$

where

- $d_i = \mathrm{rg}(X_i) - \mathrm{rg}(Y_i)$, is the difference between the two ranks of each observation.
- n is the number of observations

x	1	2	4	4.5	6	8
rg(x)	1	2	3	4	5	6
y	4	5	5.5	6	8	9
rg(y)	1	2	3	4	5	6
rg(x)-rg(y)	0	0	0	0	0	0

```
x <- c(1,2,4,4.5,6,8)
y <- c(4,5,5.5,6,8,9)
cor(x,y,method='spearman')
plot(rank(x),rank(y),col='darkgreen',pch=19,cex=3,cex.axis=2,cex.lab=2)
```



x <- c(1, 2, 2, 3)

rank(x)

相关系数3

Kendall rank correlation coefficient (KCC)

The Kendall τ coefficient is defined as:

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n(n-1)/2}.$$

If there are tied (same value) observations then τ_b is used:

$$\tau_b = \frac{S}{\sqrt{\left[n(n-1)/2 - \sum_{i=1}^{t} t_i(t_i-1)/2\right]\left[n(n-1)/2 - \sum_{i=1}^{u} u_i(u_i-1)/2\right]}}$$

- where t_i is the number of observations tied at a particular rank of x and u_i is the number tied at a rank of y.

x	1	3	2	4.5	6	8
rg(x)	1	3	2	4	5	6
y	4	5	5.5	6	8	9
rg(y)	1	2	3	4	5	6

一致的个数：5+3+3+2+1=14
不一致的个数：1

```
a <- c('High', 'Middle', 'Low', 'High', 'Low', 'Middle')
af <- factor(a, ordered=TRUE, levels=c('Low', 'Middle', 'High'), labels=c(1,2,3))
b <- c('High', 'Low', 'Low', 'High', 'Low', 'Middle')
bf <- factor(b, ordered=TRUE, levels=c('Low', 'Middle', 'High'), labels=c(1,2,3))
cor(as.numeric(af), as.numeric(bf), method='kendall')
```

三种相关系数比较

Pearson：连续变量、双变量正态分布、线性相关

Spearman：数据分布未知、连续或离散数据均可

Kendall：数据分布未知、适用于类别变量

cor()计算相关性值

cor.test()检验相关性的显著性

psych包中的corr.test()

```
x <- c(2,2.3,2.5,2.7,3,4)
y <- c(4,4.6,4.8,5.3,6,6.5)
cor(x,y,method = 'pearson')
cor.test(x,y,method = 'pearson')
library(psych)
dt <- data.frame(OTU_1 = c(1,2,2,3,4,6,8), OTU_2 =
c(2,4,4,5,6,8,10), OTU_3 = c(5,6,7,6,8,9,11))
rp <- corr.test(dt, method = 'pearson')
rp$r;rp$p
```

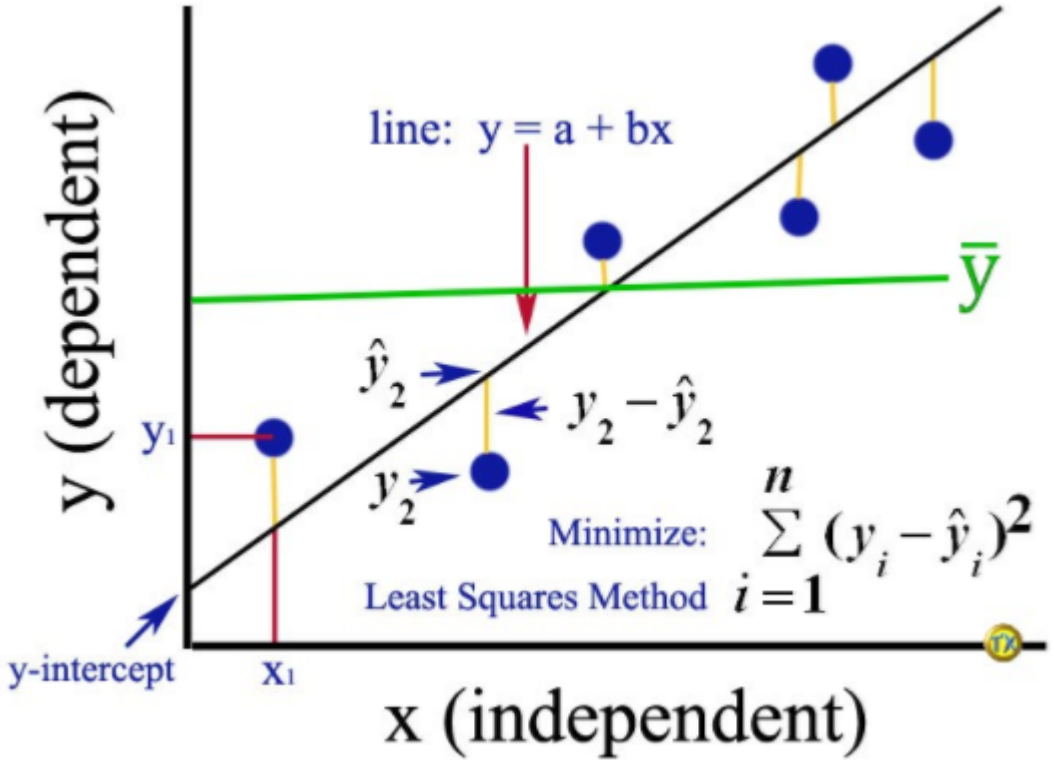
$$t_{corr} = \sqrt{\frac{r^2}{(1-r^2)/(n-2)}}$$

```
STATISTIC <- c(t = sqrt(df) * r/sqrt(1 - r^2))
```

```
        Pearson's product-moment correlation

data:  x and y
t = 6.1542, df = 4, p-value = 0.003537
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6112816 0.9947929
sample estimates:
      cor
0.9510394
```

一元线性回归

回归分析（Regression Analysis)是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法



Step1：计算中心点（x和y的平均值）

Step2：线性拟合（最小二乘）

Step3：计算斜率b、截距a和决定系数R^2

y_i $\bar{y}$ $\hat{y}$

$$b = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad a = \bar{y} - b\bar{x}$$

$SSE = \Sigma(y_i - \hat{y}_i)^2$ SSE = sum of squares due to error

$SST = \Sigma(y_i - \bar{y})^2$ SST = total sum of squares

$SSR = \Sigma(\hat{y}_i - \bar{y})^2$ SSR = sum of squares due to regression

$$R^2 = \frac{SSR}{SST}$$

$$R^2\_adjusted = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$$

$$\text{F-statistics} = \frac{SSR/(p-1)}{SSE/(n-p)}$$

Novogene 诺禾致源

Providing advanced genomic solutions!

数理统计基础

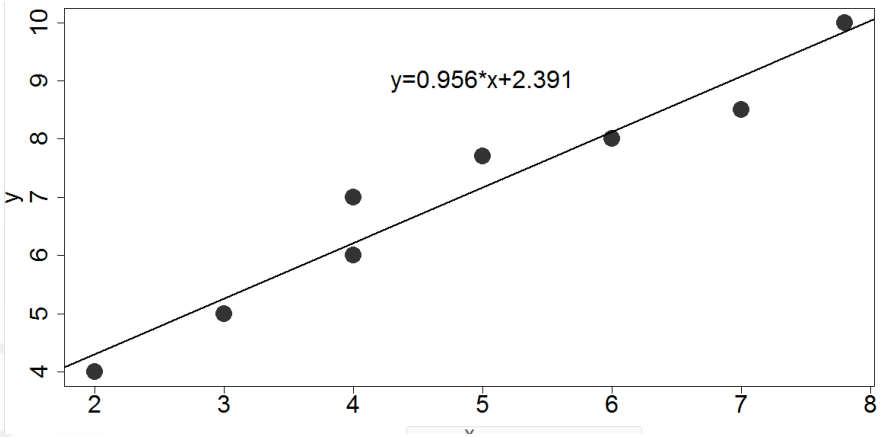
一元线性回归

$$R^2 = \frac{SSR}{SST} \qquad R^2\_adjusted = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$$

```
x <- c(2,3,4,4,5,6,7,7.8)
y <- c(4,5,6,7,7.7,8,8.5,10)
lmres <- lm(y ~ x)
summary(lmres)
plot(x,y,cex=3,pch=19,col='gray20',
cex.axis=2,cex.lab=2)
abline(lmres,lwd=2)
coff <- coefficients(lmres)
ex <- as.expression(paste('y=',
sprintf("%.3f", coff[2]), "*x+",
sprintf("%.3f", coff[1]), sep=''))
text(5,9, as.expression(ex))
```

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5794 -0.2684 -0.1683  0.2500  0.7872

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.39067    0.49056   4.873  0.00278 **
x            0.95553    0.09444  10.117 5.42e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4967 on 6 degrees of freedom
Multiple R-squared:  0.9446,    Adjusted R-squared:  0.9354
F-statistic: 102.4 on 1 and 6 DF,  p-value: 5.419e-05
```

$$\text{F-statistics} = \frac{SSR/(p-1)}{SSE/(n-p)}$$



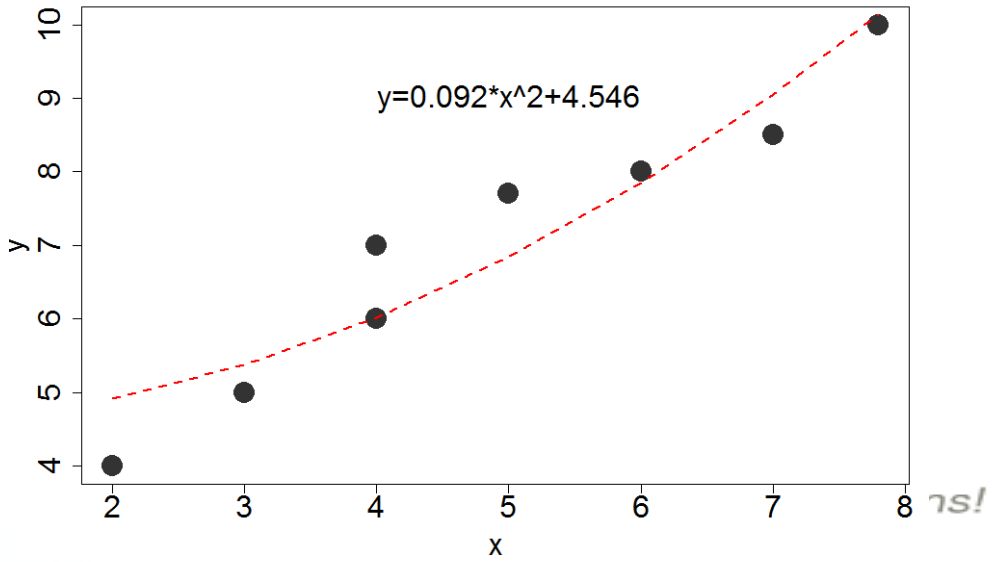
数理统计基础

多项式回归

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \varepsilon.$$

```
x <- c(2,3,4,4,5,6,7,7.8)
y <- c(4,5,6,7,7.7,8,8.5,10)
lmres2 <- lm(y ~ I(x^2))
summary(lmres2)
coff2 <- coefficients(lmres2)
plot(x,y,cex=3,pch=19,col='gray20',
cex.axis=2,cex.lab=2)
lines(x, fitted(lmres2),lwd=2,lty=2,col='red')
ex <- as.expression(paste('y=', sprintf("%.3f",
coff2[2]), "*x^2+", sprintf("%.3f", coff2[1]),
sep=''))
text(5,9, ex, cex=2)
```

```
Call:
lm(formula = y ~ I(x^2))

Residuals:
     Min       1Q   Median       3Q      Max
-0.91343 -0.41674 -0.07619  0.32387  0.98392

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.54589    0.43947  10.344 4.77e-05 ***
I(x^2)       0.09189    0.01338   6.868 0.000469 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7091 on 6 degrees of freedom
Multiple R-squared:  0.8872,    Adjusted R-squared:  0.8684
F-statistic: 47.17 on 1 and 6 DF,  p-value: 0.0004694
```



Novogene 诺禾致源

数理统计基础

多元线性回归

$$Y' = a + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

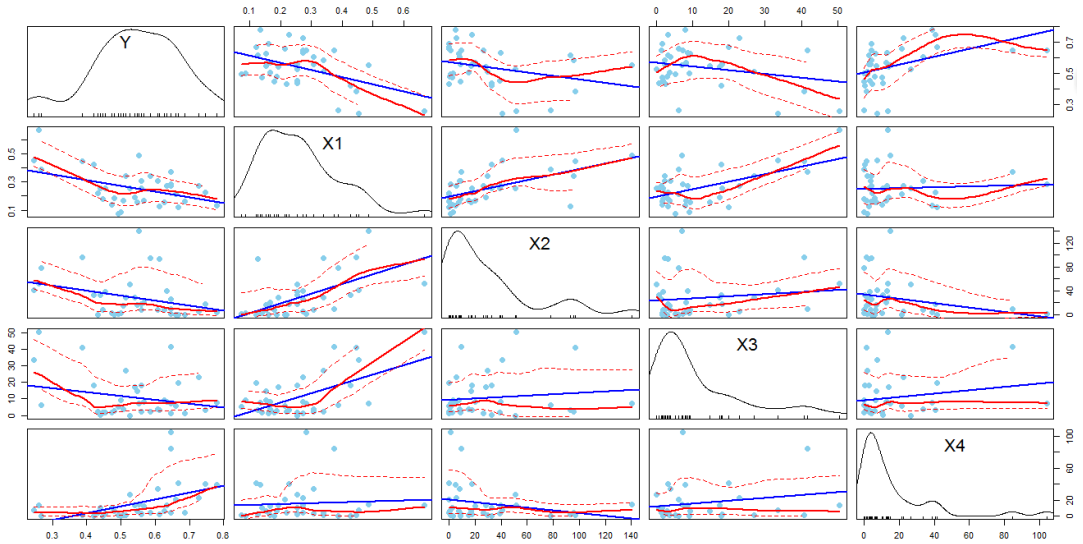
```
setwd("H:/2017~/培训/20180820_基因学院培训")
library(car)
dt <-  read.table("data/dt.list",
stringsAsFactors = F,head=T,
row.names=1,sep="\t")
scatterplotMatrix(dt, col=c('blue', 'red',
'skyblue'), cex=1.5, pch=19, lwd=2)
mullres <- lm(Y~., data=dt)
summary(mullres)
fitted(mullres)
vif(mullres)#方差膨胀因子检测多重共线性
```

vif>10，存在一定的多重共线性

$$VIF_{Weight} = \frac{1}{1 - R^2_{Weight}}$$

复相关系数

```
Call:
lm(formula = Y ~ ., data = dt)

Residuals:
      Min        1Q    Median        3Q       Max
-0.188000 -0.071922 -0.006386  0.080099  0.193943

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.6241657  0.0288888  21.606  < 2e-16 ***
X1          -0.5073685  0.1473153  -3.444  0.00101 **
X2           0.0004156  0.0004474   0.929  0.35647
X3          -0.0009471  0.0011937  -0.793  0.43043
X4           0.0029541  0.0005395   5.475 7.56e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09582 on 65 degrees of freedom
Multiple R-squared:  0.4471,    Adjusted R-squared:  0.4131
F-statistic: 13.14 on 4 and 65 DF,  p-value: 6.705e-08
```



Novogene 诺禾致源

数理统计基础

距离

相关系数：变量间的相似性(correlation)　VS　**距离**：样本间的**相似(异)性**（dissimilarity）

1、Euclidean distance

$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\cdots+(q_n-p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

`library(vegan)`　**提供17种距离计算的方法**

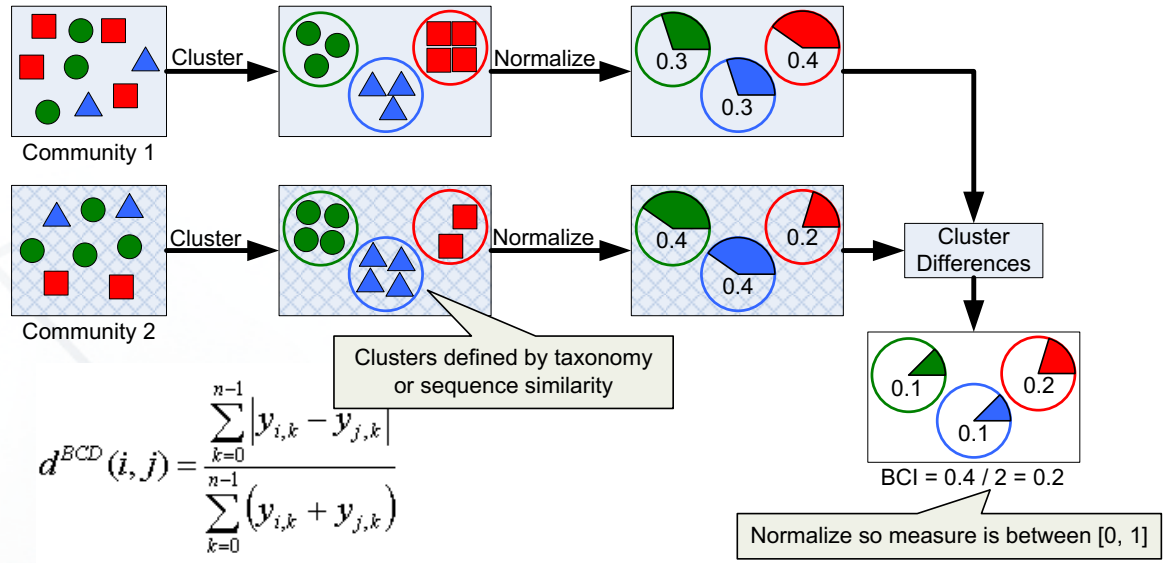
```
df <- t(data.frame(A=c(1.5,2,2.3,3.5), B=c(3.3,3.6,3.9,4.1)))
```

A	1.5	2.0	2.3	3.5
B	3.3	3.6	3.9	4.1

2、Bray-Curtis distance

An ordination of the upland forest communities of southern Wisconsin
JR Bray, JT Curtis - Ecological monographs, 1957 - Wiley Online Library
INTRODUCTION A renewed interest in objective and quantitative approaches to the classification of plant communities has led, within the past decade, to an extensive examination of systematic theory and technique. This examination, including the work of
被引次数 7689 相关文章 所有 8 个版本 引用 保存

```
vegdist(df, method = 'bray')
```



$$d^{BCD}(i,j) = \frac{\sum_{k=0}^{n-1}\left|y_{i,k}-y_{j,k}\right|}{\sum_{k=0}^{n-1}\left(y_{i,k}+y_{j,k}\right)}$$

3 Unifrac

Novogene 诺禾致源

Providing advanced genomic solutions!

参数估计

点估计：已知**分布**的情况下，借助**总体X中的样本数据**，估计未知**参数(θ)**

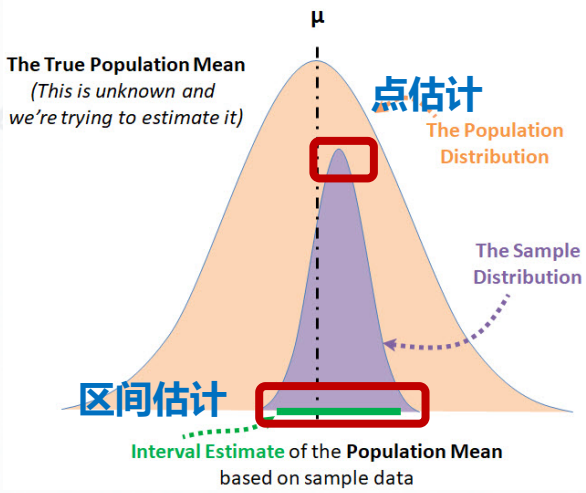
总体　样本　参数　分布

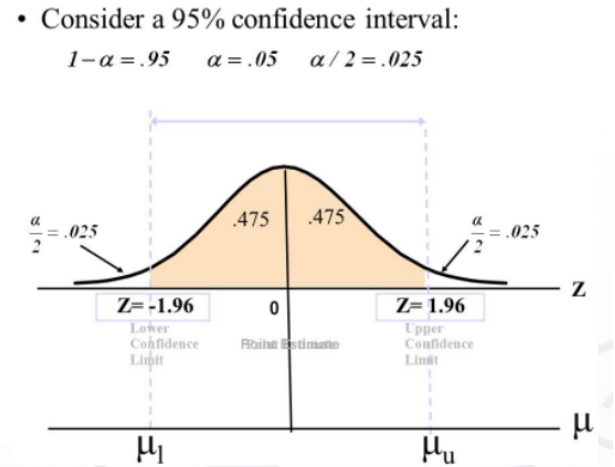
是否准确

区间估计：利用样本数据，估计未知参数在**置信度**下的可能存在的区间

置信度　置信区间

100次抽样→100个置信区间，95个区间包含真实参数





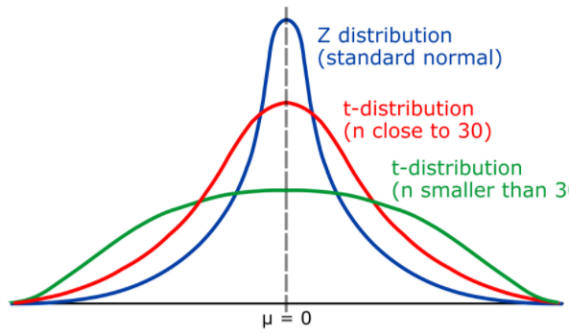
z- and t-distribution

$$\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

For a 90% confidence interval: $z_{\alpha/2} = 1.65$

For a 95% confidence interval: $z_{\alpha/2} = 1.96$

For a 99% confidence interval: $z_{\alpha/2} = 2.58$



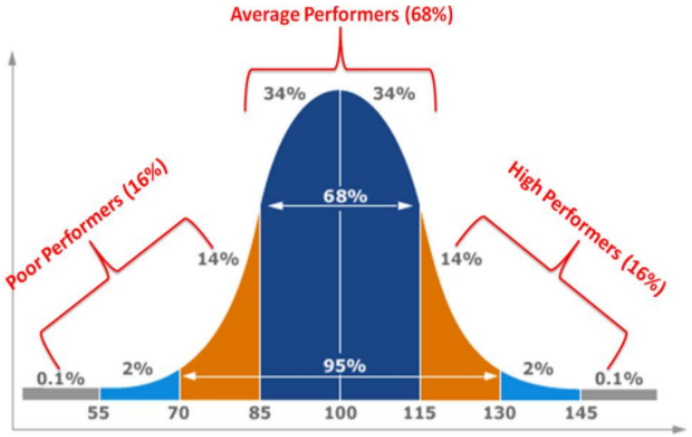
Novogene 诺禾致源

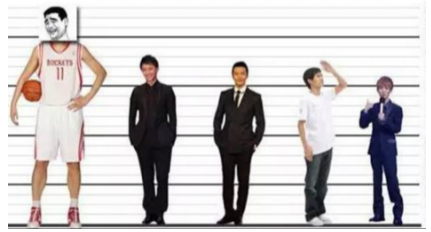
Providing advanced genomic solutions!

数理统计基础

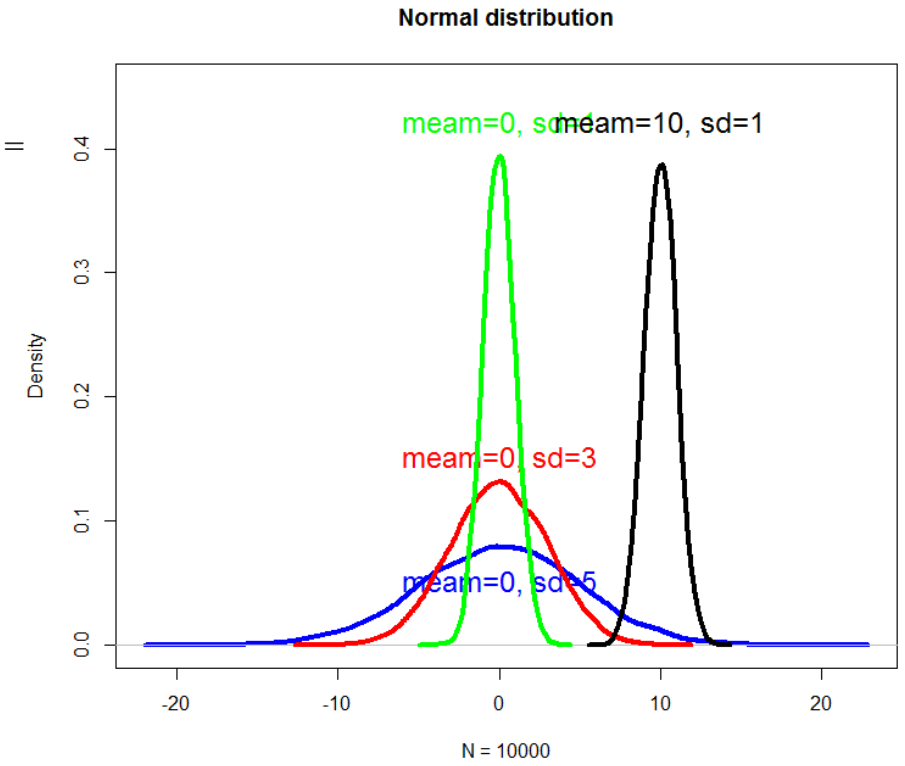
正态分布

中间高、两头低的一种概率分布





```
rn1 <- rnorm(10000, 0, 5)
rn2 <- rnorm(10000, 0, 3)
rn3 <- rnorm(10000, 0, 1)
rn4 <- rnorm(10000, 10, 1)
plot(density(rn1), lwd=4, col='blue', ylim=c(0, 0.45), xlab = 'N =
10000', main = 'Normal distribution')
text(0,0.05,labels = "meam=0, sd=5", col='blue', cex=1.5)
lines(density(rn2), lwd=4, col='red')
text(0,0.15,labels = "meam=0, sd=3", col='red', cex=1.5)
lines(density(rn3), lwd=4, col='green')
text(0,0.42,labels = "meam=0, sd=1",  col='green', cex=1.5)
lines(density(rn4), lwd=4, col='black')
text(11,0.38,labels = "meam=10, sd=1",  col='black', cex=1.5)
```



如何转化为标准正态分布

scale()

Novogene 诺禾致源

Providing advanced genomic solutions!

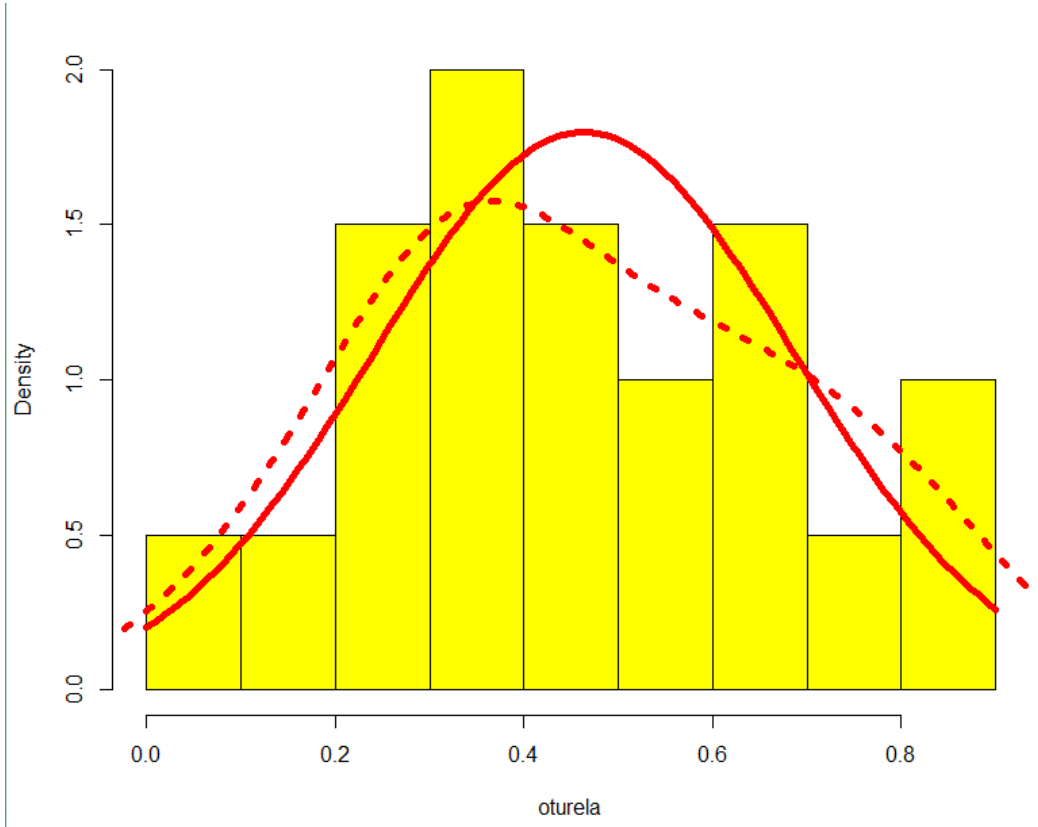
正态分布

已知OTU1的丰度符合正态分布，随机取样测得丰度，计算95%置信水平下的置信区间

```
set.seed(100)
oturela <- runif(20, 0.01, 0.99)#随机生成otu丰度数据
nx <- length(oturela)
mx <- mean(oturela)
mx - (sd(oturela)/sqrt(nx))*abs(qt(0.025, nx-1))
mx + (sd(oturela)/sqrt(nx))*abs(qt(0.025, nx-1))
t.test(oturela)
hist(oturela, col='yellow', ylim=c(0,2.0), main='data
distribution', probability = T)
lines(density(oturela), col='red', lwd=5, lty=3)
curve((dnorm(x, mx, sd(oturela))), col='red',lwd=5,
add = TRUE)
```

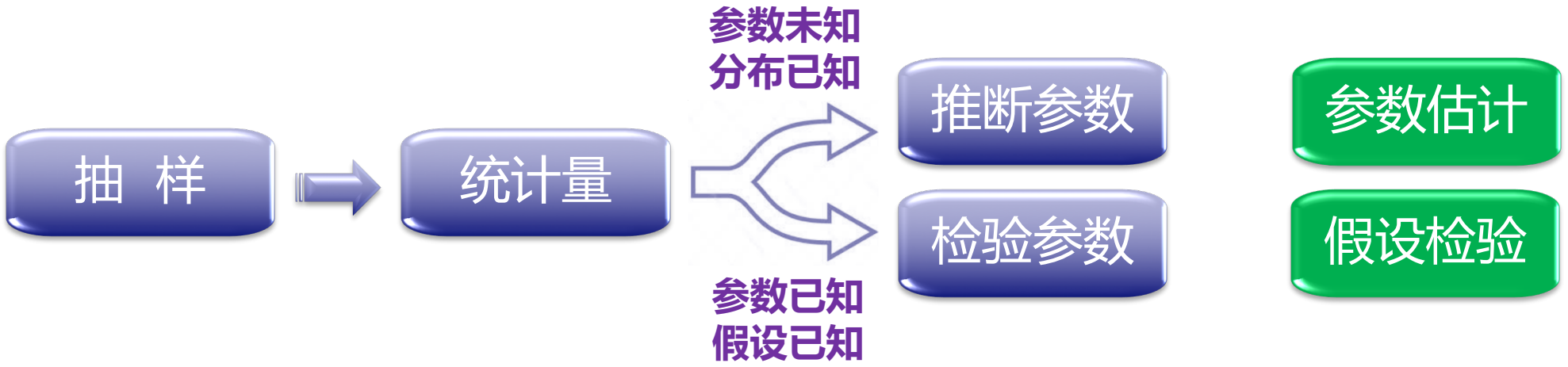
$$\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Density = Prob/Spacing



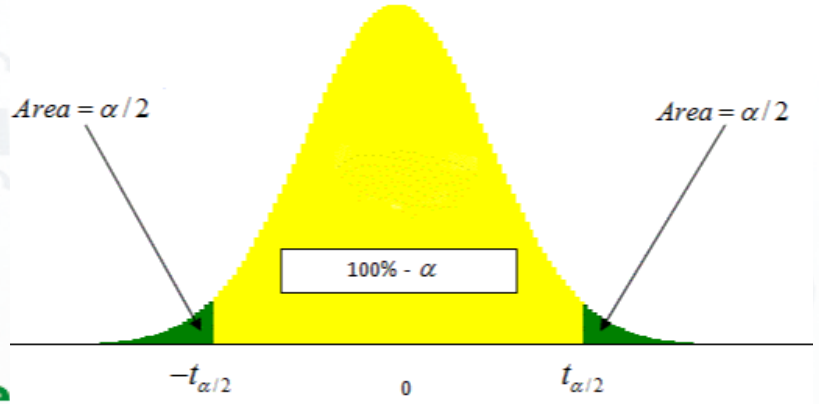
数理统计基础

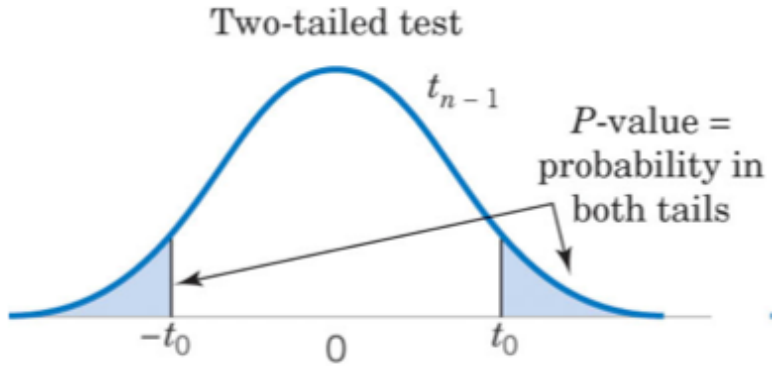
假设检验基础



显著性水平α：发生第I类错误的概率（经验值一般为0.01和0.05）

P-value（Probability）：在原假设成立的情况下，出现极端情况的概率





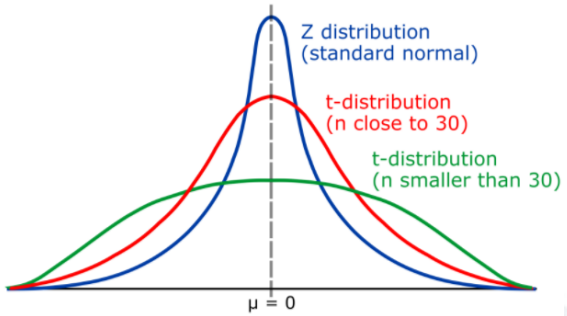
Novogene 诺禾致源

Providing advanced genomic solutions!

T-test

T-test：examine whether the two samples are drawn from the same population

t-distribution



- One sample T-test
- Two Independent Samples T-Test
- Paired T-test

```
sci <- c(8.3, 8.9, 8.8, 8.1, 7.3, 7.5, 5.8, 6.9)
t.test(sci, mu=5, alternative = 'two.sided', conf.level = 0.95)
#----------------------------------------
before <- c(5, 4, 5, 4.5, 6, 6.5, 6.3)
after <- c(10, 11, 11.1, 10.9, 9.8, 9.4, 11.3)
t.test(before, after, alternative = 'two.sided', paired=FALSE, var.equal
= FALSE, conf.level = 0.95)
#----------------------------------------
t.test(before, after, alternative = 'two.sided', paired=TRUE, var.equal
= FALSE, conf.level = 0.95)
```

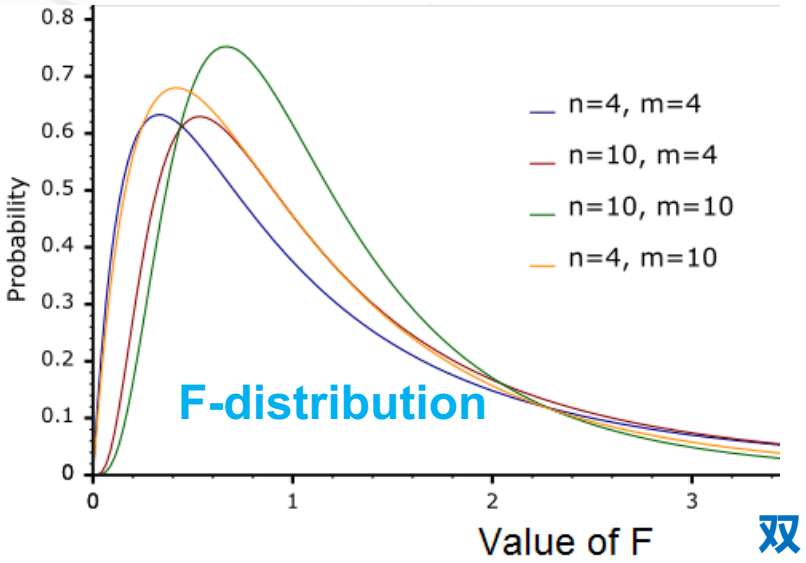
```
        Paired t-test

data:  before and after
t = -9.3624, df = 6, p-value = 8.422e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.523006 -3.819851
sample estimates:
mean of the differences
              -5.171429
```

ANOVA

Analysis of Variance：

examine the differences among group means in samples

Normally distribution　Variance is equal




```
sci <- c(3,4,5,4,2,3,2,2.5,6,7,6,7.7,9.5,9.4,8.6,8.8)
f1 <- c(rep(1,4), rep(2,4), rep(3,4), rep(4,4))
f2 <- c(rep(11, 8), rep(12,8))
df <- data.frame(sci,f1,f2)
with(df, tapply(sci,f1,mean))
ano_res <- aov(sci~f1)
summary(ano_res)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
f1           3 104.69   34.90   78.38 3.76e-08 ***
Residuals   12   5.34    0.45
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

双（多）因素ANOVA

aov(sci~f1+f2)

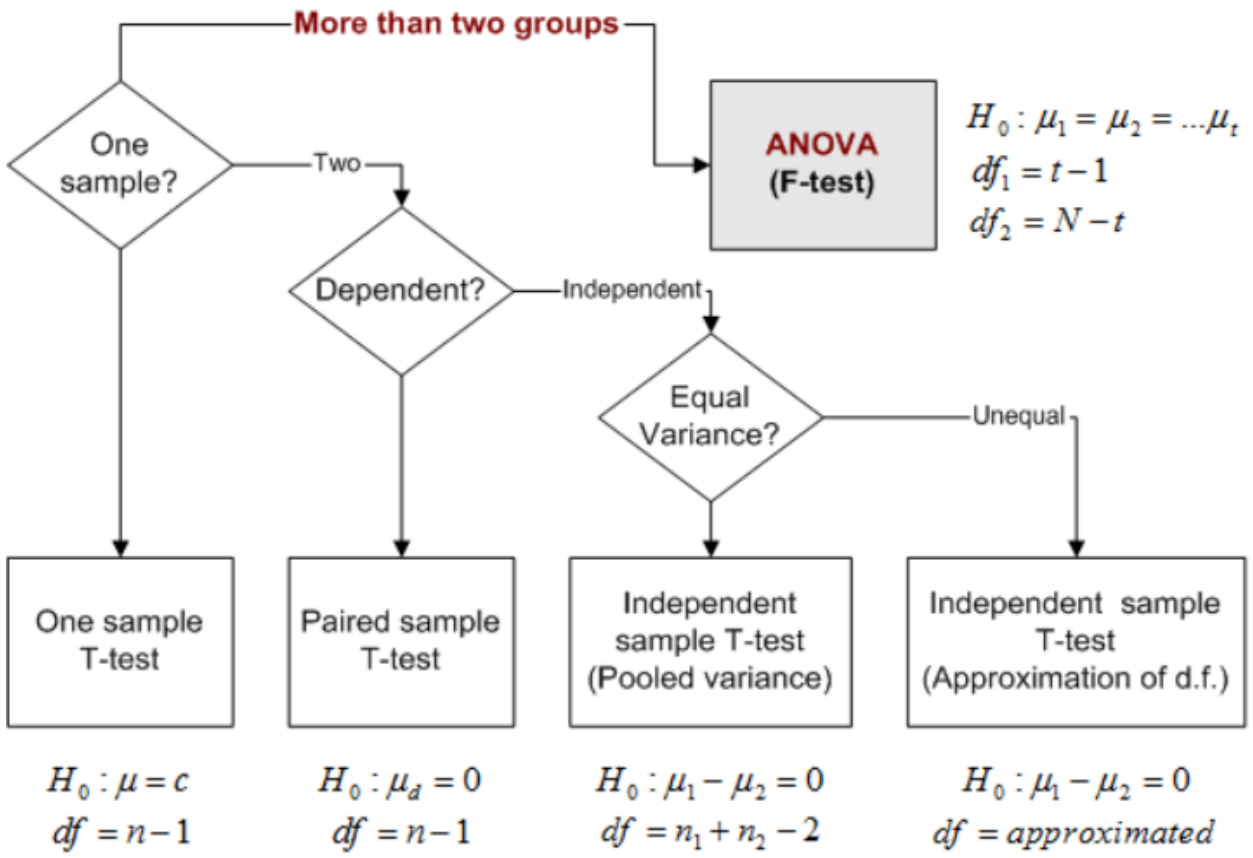
aov(sci~f1*f2)

Novogene 诺禾致源

Providing advanced genomic solutions!

T-test & ANOVA

BASIS FOR COMPARISON	T-TEST	ANOVA
Meaning	T-test is a hypothesis test that is used to compare the means of two populations.	ANOVA is a statistical technique that is used to compare the means of more than two populations.
Test statistic	$(\bar{x}-\mu)/(s/\sqrt{n})$	Between Sample Variance/Within Sample Variance

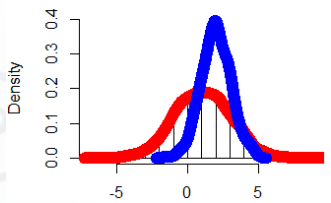


➢ 正态分布的检验

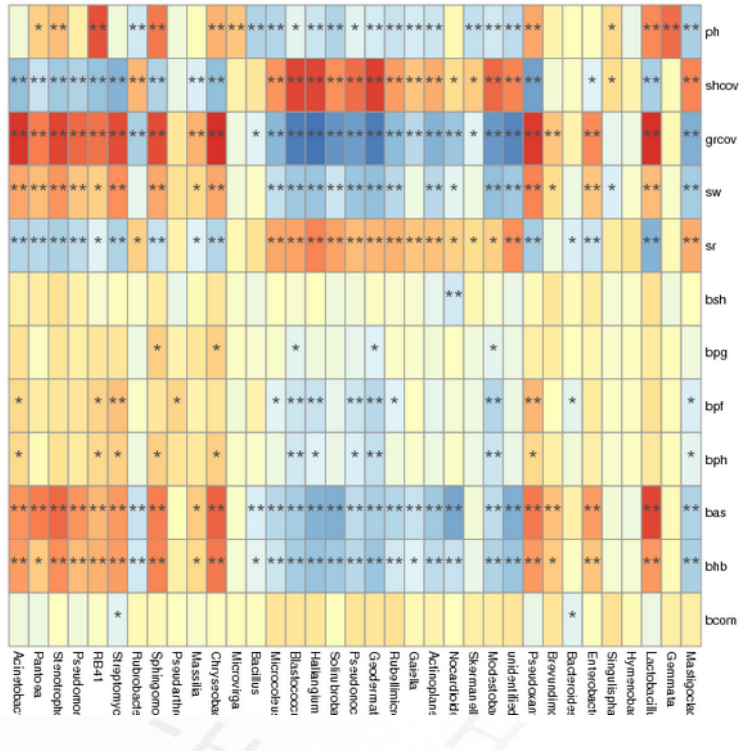
```
set.seed(50)
x <- rnorm(1000, 1, 2)
y <- rnorm(500, 2, 1)
qqnorm(x);qqline(x)
hist(x,freq=F)
lines(density(x),col="red")
shapiro.test(x)
ks.test(x,y)
```

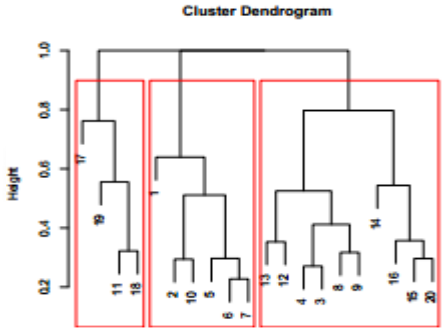
➢ 方差齐次性检验

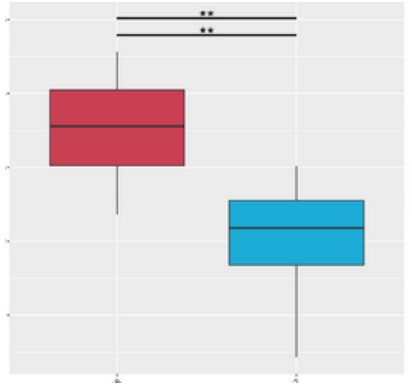
```
var(x); var(y)
hist(x, freq = F, ylim =
c(0,0.4))
lines(density(x), lwd=2,
col='red')
lines(density(y), lwd=2,
col='blue')
```

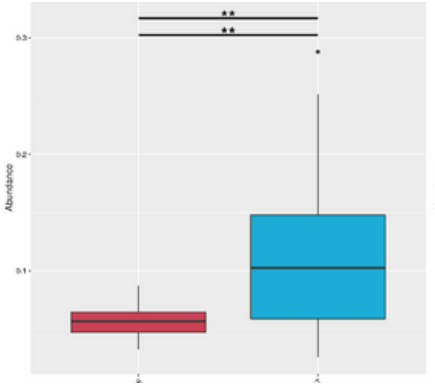


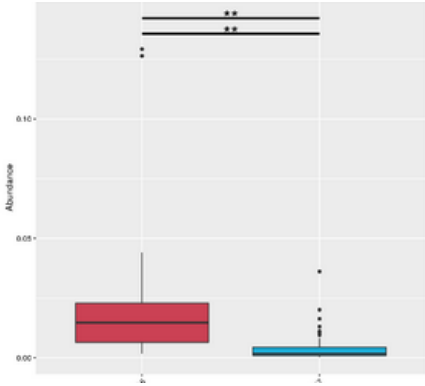
$$df_{adj} = \frac{(S_x/n_x + S_y/n_y)^2}{\frac{(S_x/n_x)^2}{nx-1} + \frac{(S_y/n_y)^2}{ny-1}}$$

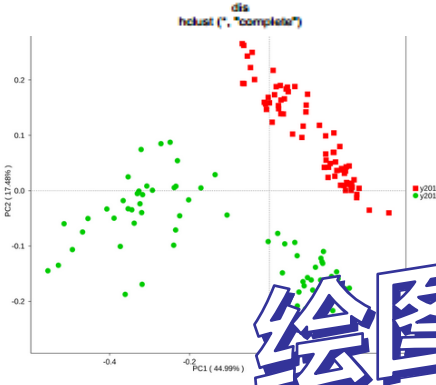


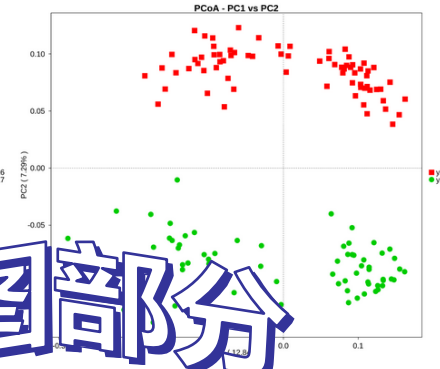


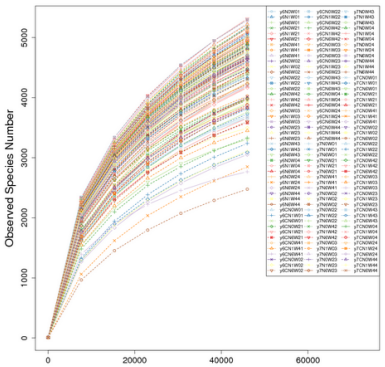


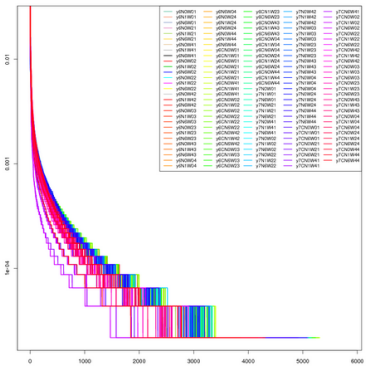




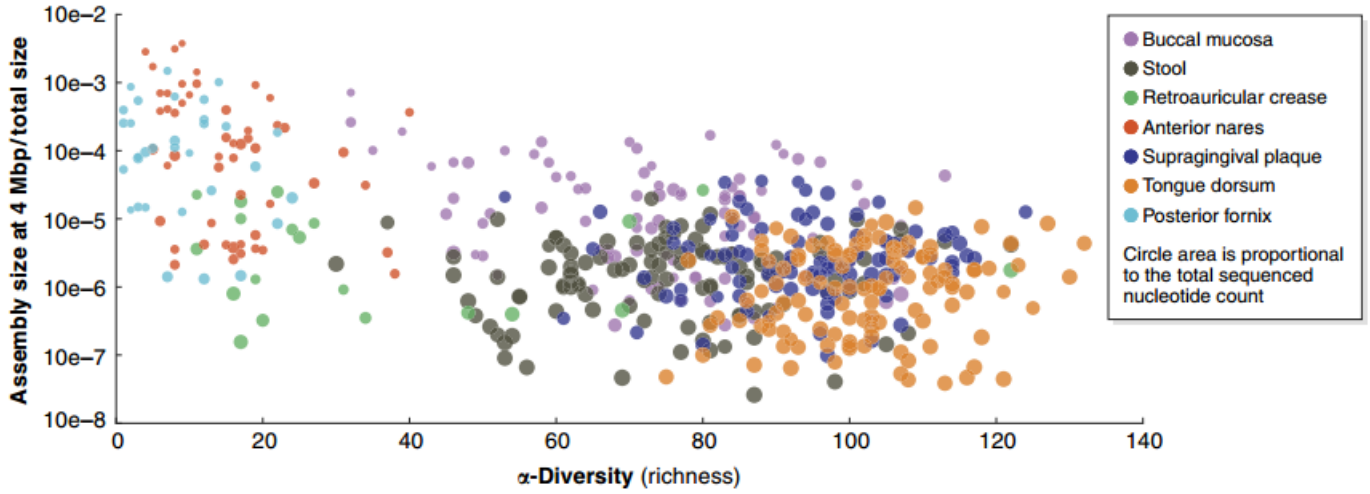


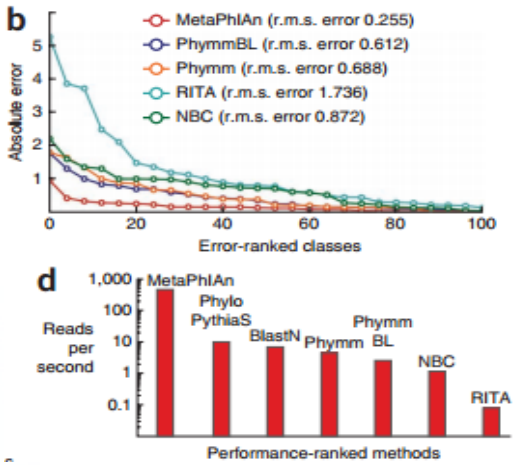


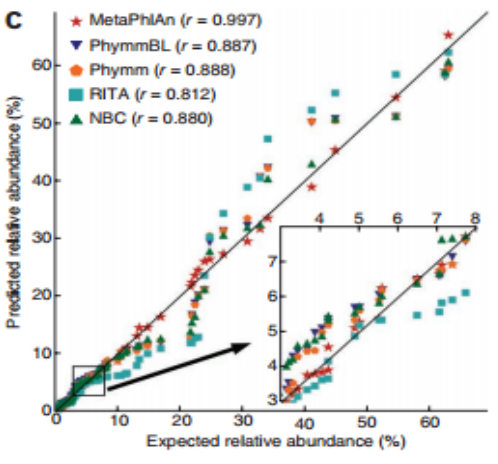




绘图部分







Novogene 诺禾致源

Providing advanced genomic solutions!

R包

R包(package)：R函数、数据、预编译代码以一种定义完善的格式组成的集合。

开箱即用

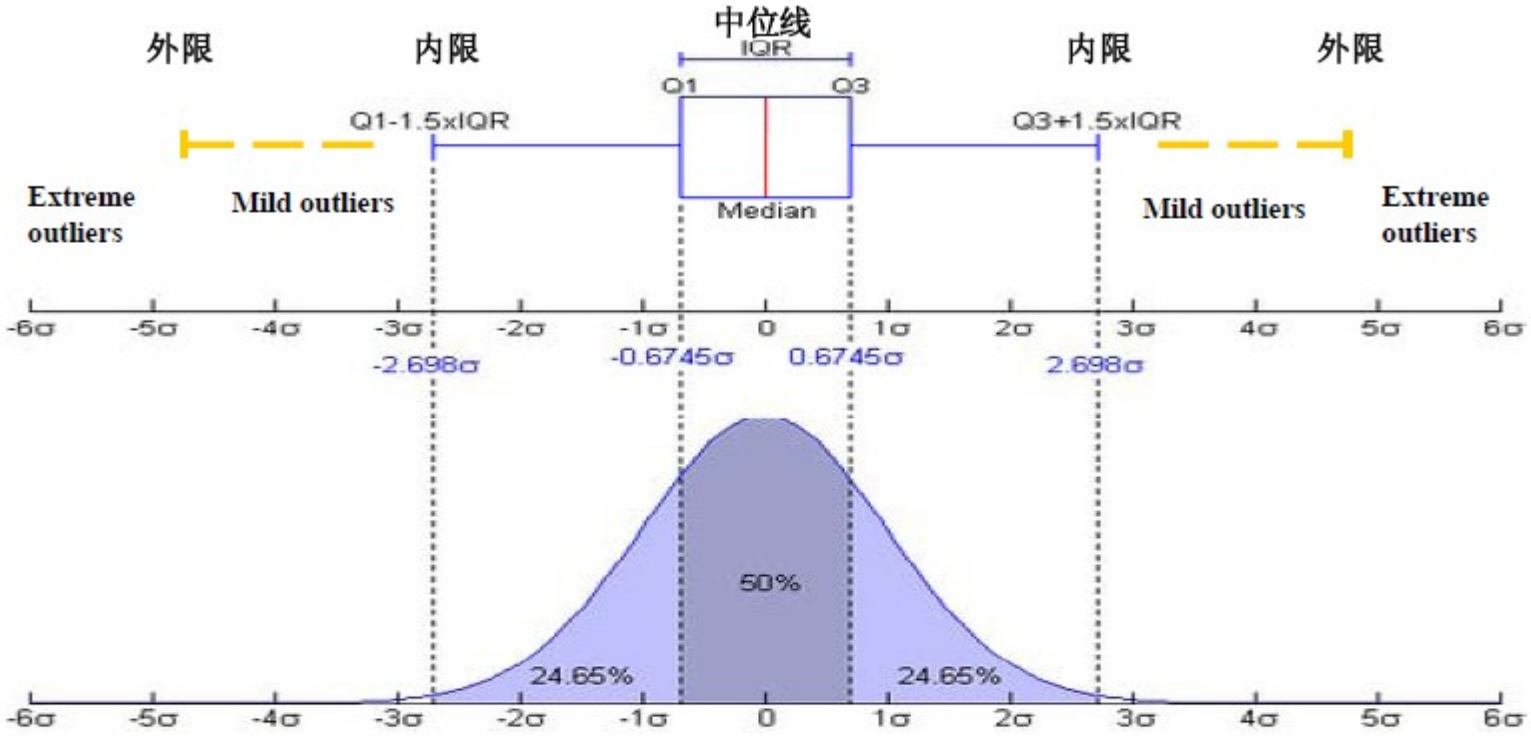
R包的获取

```
.libPaths("E:/Rstudio/R_packages") #指定安装包的路径
install.packages("vegan") #安装包
library(vegan) #加载包，也可用require()

update.packages("vegan") #包的更新
installed.packages() #查看已安装的包

help(vegan) #查看包的帮助文档
help(cca) #查看包中函数的说明文档
example(vegan) #查看包中的示例文件
```

R实战

- 注意使用"Tab"键的补齐功能；文件读写
- 可以使用小键盘"向上箭头"和"向下箭头"翻出历史记录
- 脚本中的"+" 号，代表换行而不终止句子
- **中文符号**：代码中，如果有中文的标点符号，则是无效的
- **固定简写**：如c(1,2,3)中的c，还有某些函数的参数（例如，read.table的参数sep）
- 学会查看**帮助文档**（帮助文档中有**函数**的使用示例和**参数**详解）
  - help()/example()/?/??

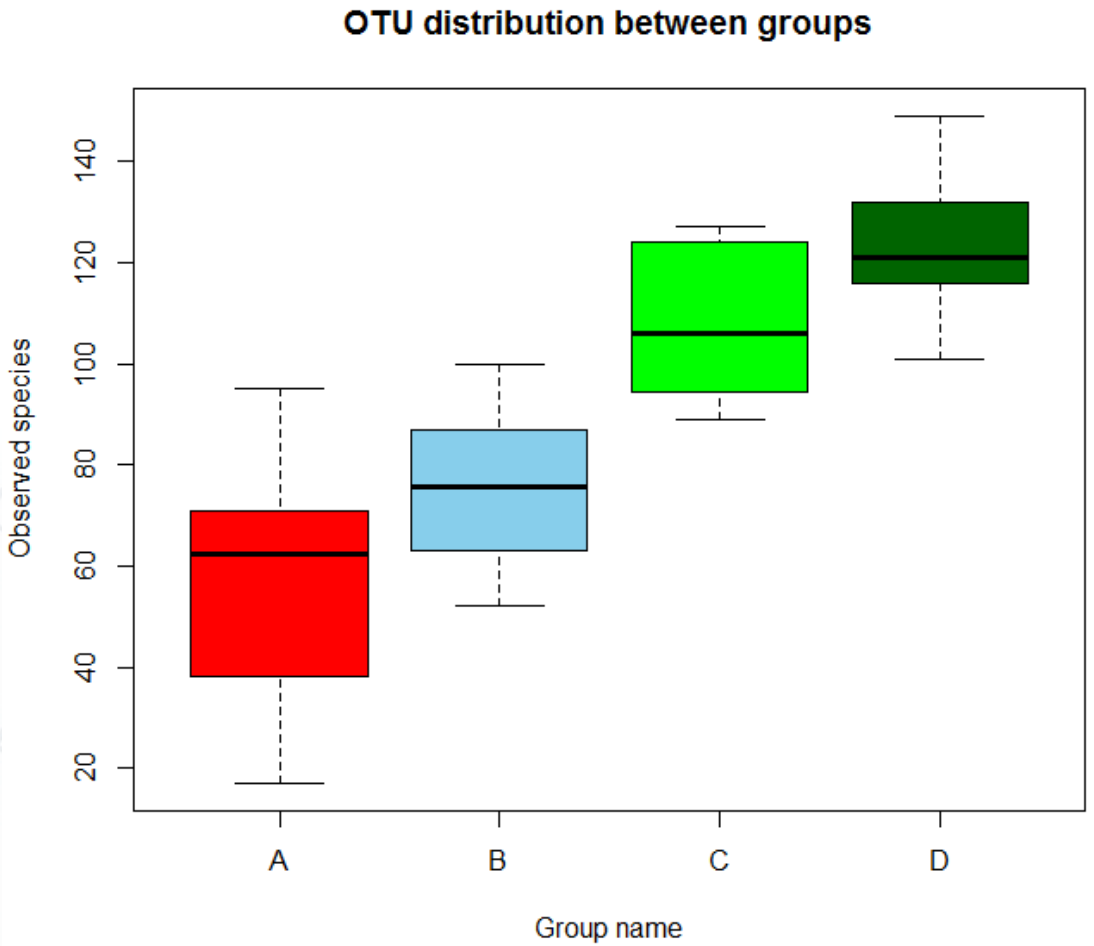
箱线图解读



Novogene 诺禾致源

Providing advanced genomic solutions!

箱线图1--boxplot(x)




```
setwd("H:/2017~/培训/20180820_基因学院培训")
otu <- read.table("data/otu_num.txt", header=T,
sep="\t", stringsAsFactors=TRUE)
str(otu)
# otu$Group <- factor(otu$Group)
# otu$Level <- factor(otu$Level)
boxplot(OTU_num~Group,
        data=otu, col=c("red","skyblue",
"green","darkgreen"),
        varwidth=F,main="OTU distribution
ween groups", xlab="Group
e",ylab="Observed species")
```

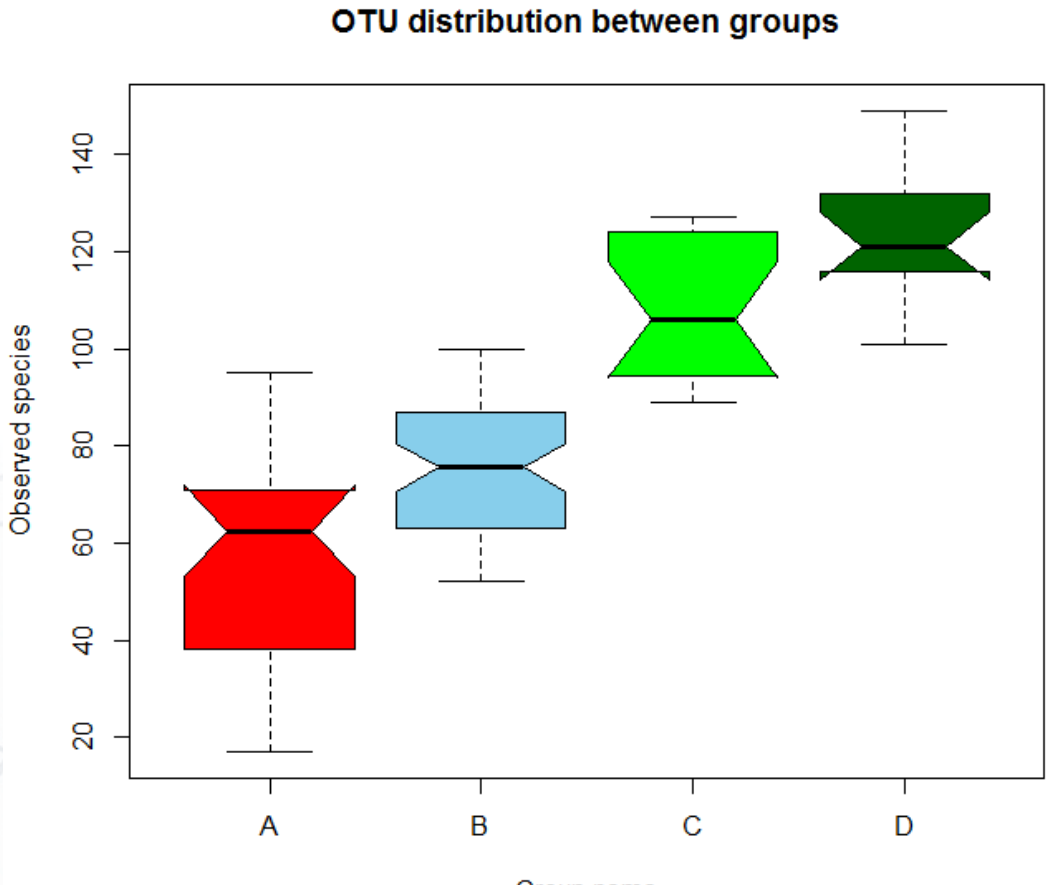
OTU_num	Group	Level
44	A	low
37	A	low
34	A	low
54	A	low
83	A	low
48	A	low
39	A	low

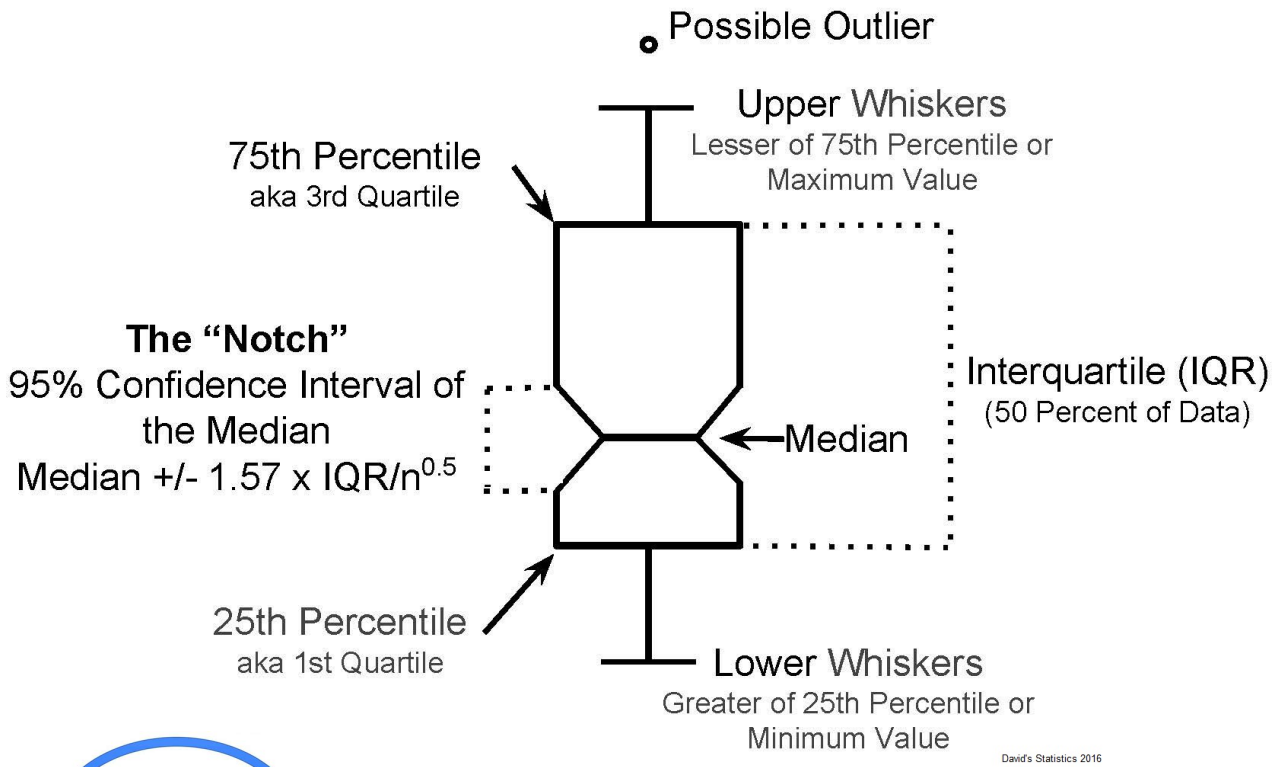
如何让盒子的宽度与样本量成正比（ varwidth=T ）

如何让A、B和C、D之间空白大一些(at=c(1,2,4,5))

试试OTU_num~Group*Level

箱线图2--boxplot(x)





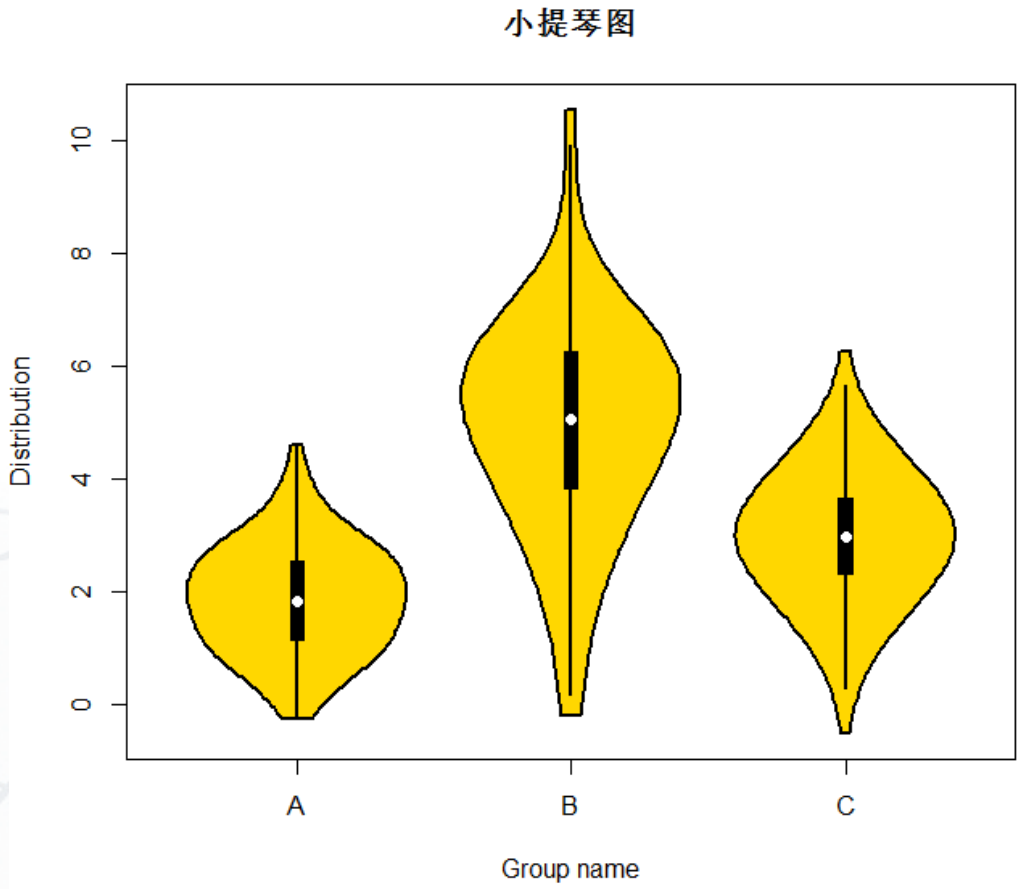
notch=TRUE

凹槽箱线图

Novogene 诺禾致源

Providing advanced genomic solutions!

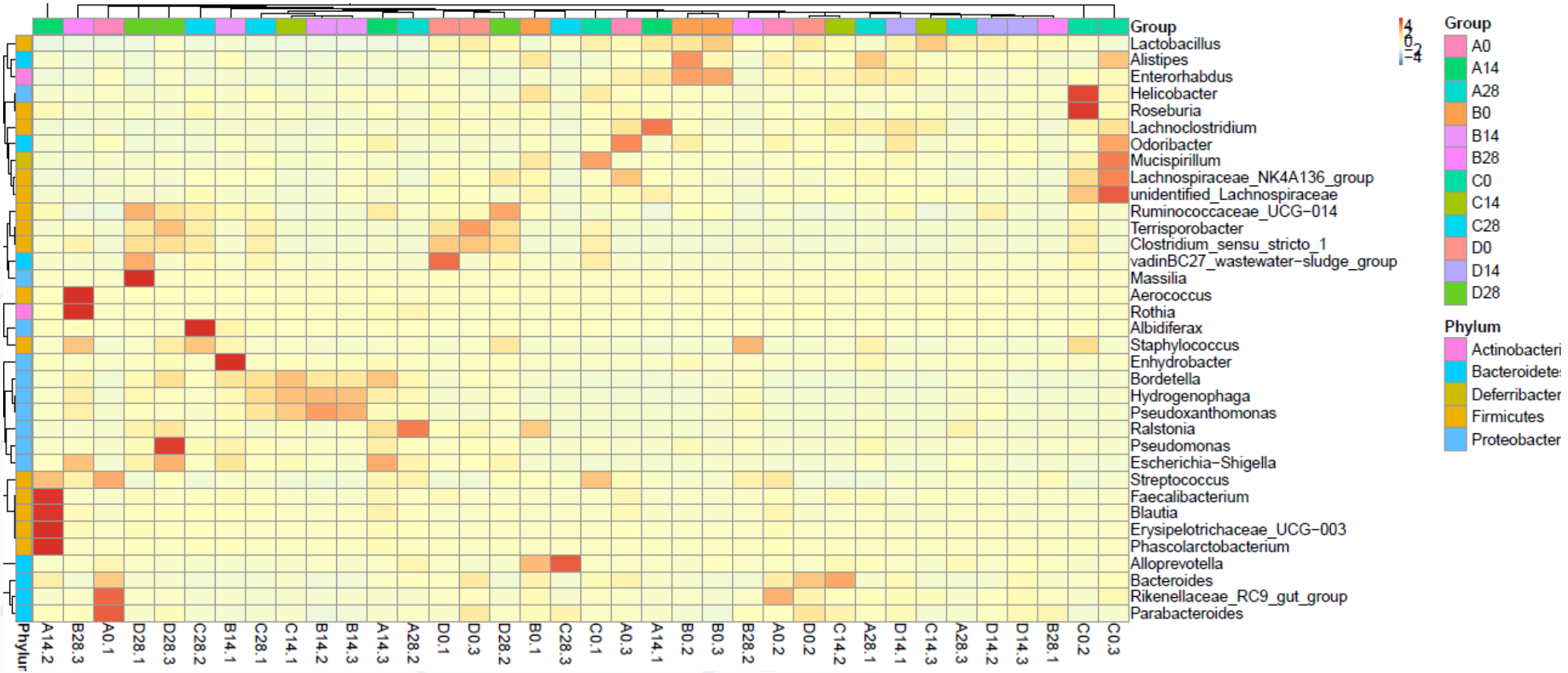
箱线图3--boxplot(x)




```
library(vioplot)
x1 <- rnorm(100, 2, 1)
x2 <- rnorm(50, 5, 2)
x3 <- rnorm(1000, 3, 1)
vioplot(x1, x2, x3, lty=1, lwd=2,
        names=c("A", "B", "C"),
col="gold")
title(main="小提琴图", xlab="Group
name", ylab="Distribution")
```

白点：中位数；黑色盒子：四分位点；外部形状：核密度图

热图--heatmap



Novogene 诺禾致源

Providing advanced genomic solutions!

热图--heatmap

```
#install.packages("pheatmap")
library(pheatmap)
setwd("H:/2017~/培训/20180820_基因学院培训")
taxa<-read.table("data/cluster.g.txt",sep="\t",header=T,row.names=1)
group<-read.table("data/sam_group.list",sep="\t",header=F)
gp<-read.table("data/genus_phylum.list",sep="\t",header=F)
annotation_col = data.frame(Group=factor(group[[2]]))
rownames(annotation_col) = group$V1
annotation_row = data.frame(Phylum=factor(gp$V2))
rownames(annotation_row) = gp$V1
pheatmap(taxa,scale="row",fontsize=50,
         cluster_rows=T,clustering_distance_rows="correlation",
         cluster_cols=T,clustering_distance_cols="euclidean",
         clustering_method="average",
         annotation_col = annotation_col,
         annotation_row = annotation_row,
         height=30,width=70,filename="genus_heatmap.pdf")
```

数据标准化：Z-score

$$z = \frac{x - \mu}{\sigma}$$

column ? ro

变量计算系数

样本计算距离

clustering_distance_rows

clustering_distance_cols

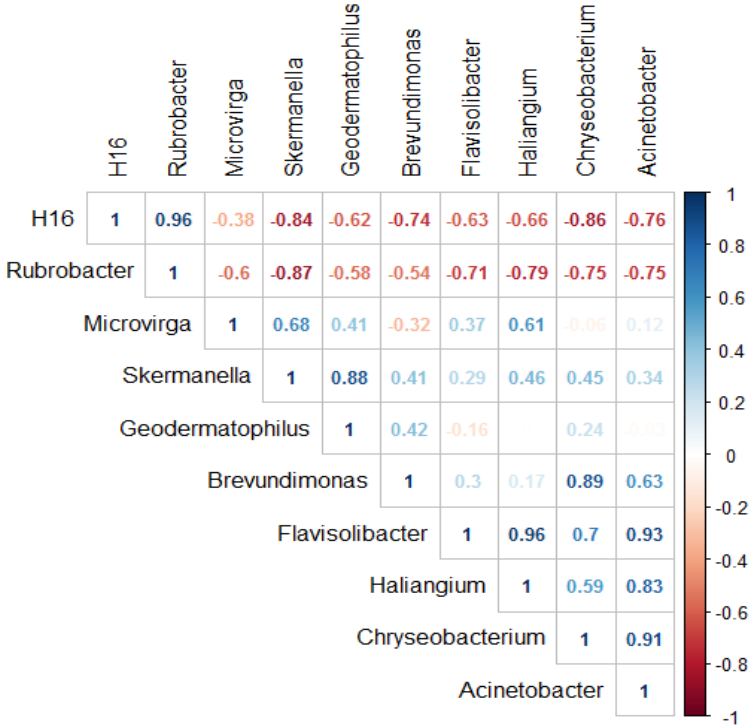
clustering_method

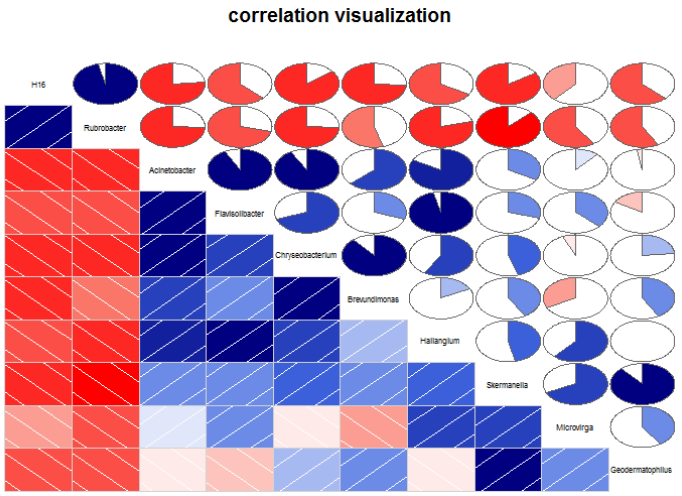
思考：方格内如何添加数字？？？

相关图

```
library(corrplot)
setwd("H:/2017~/培训/20180820_基因学院培训")
mat <- read.table("data/cluster.p.txt", header=T,
row.names=1, sep="\t")
mat1 <- t(mat)[,1:10]
a <- cor(mat1)
corrplot(a, method="number", type="upper",
         tl.col="black", number.cex=0.8, number.font=2,
         order = "hclust", hclust.method="average")
#####################################
library(corrgram)
corrgram(a, order=T,
lower.panel=panel.shade,upper.panel=panel.pie,
text.panel=panel.txt,main="correlation visualization")
```

method = c("circle", "square", "ellipse", "number", "shade", "color", "pie")





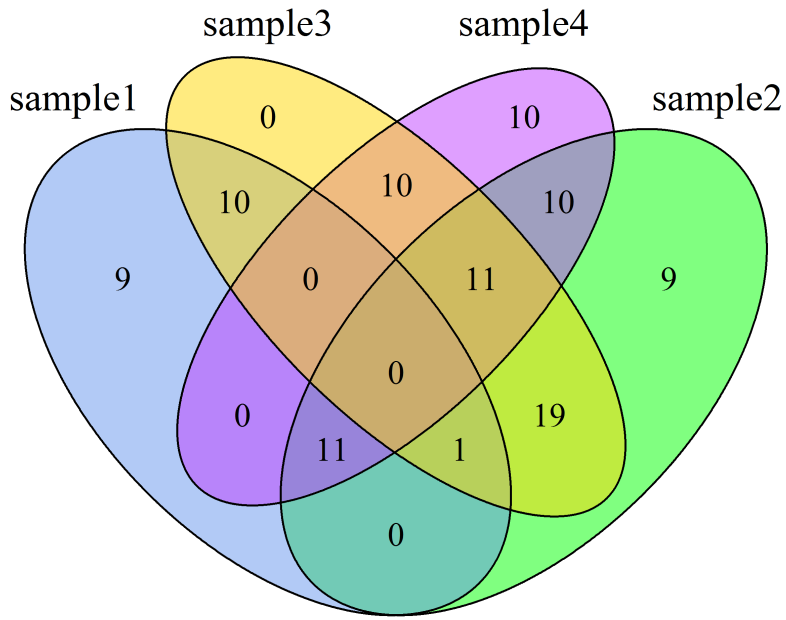
挑选感兴趣的菌属，进行相关性系数的计算和可视化展示

Novogene 诺禾致源

Providing advanced genomic solutions!

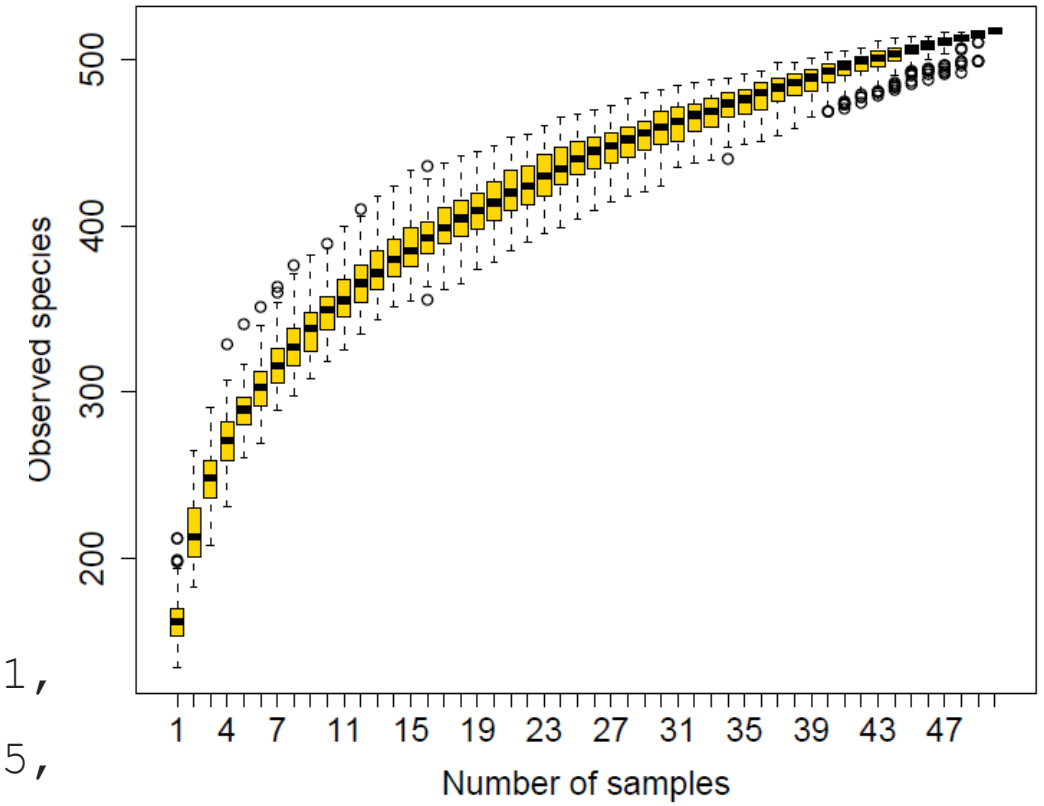
维恩图--Venndiagram

```
library(VennDiagram)
ven<-list(sample1=20:50,  sample2=c(1:30,50:80), sample3=40:90, sample4=c(10:30,70:100)
venn.diagram(ven, filename='venn.png', cex=1.2, col="black", alpha= 0.50,lwd =1.2,
cat.cex=1.4, fill=c("cornflowerblue", "green", "Gold1","darkorchid1"),  margin=0.15)
```

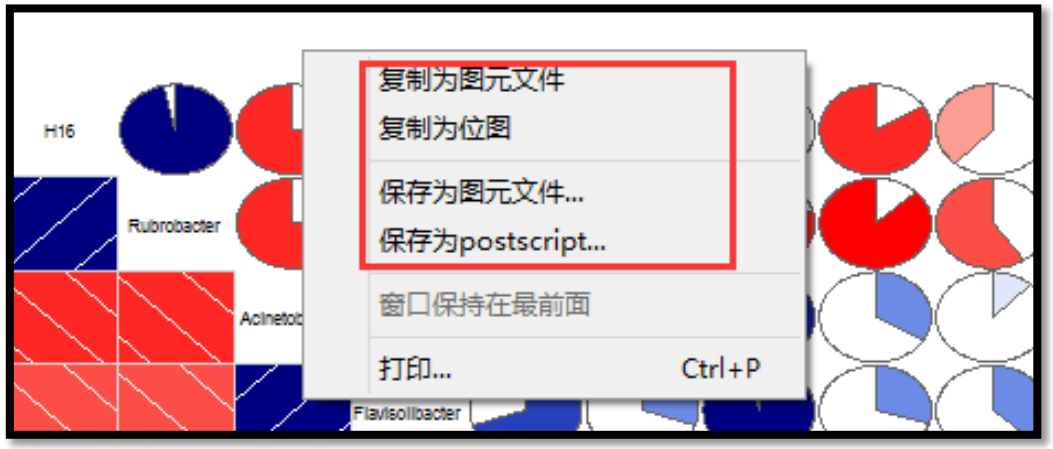


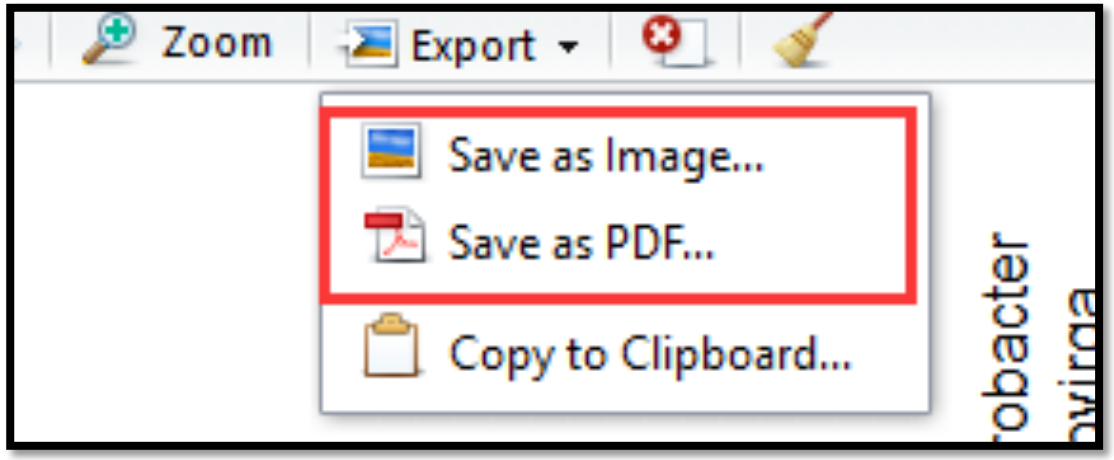
累积箱线图--Specaccum

```
library(vegan)
library(reshape2)
setwd("H:/2017~/培训/20180820_基因学院培训")
otu <-
read.table("data/otu_for_specaccum.txt",header=T)
otu <- t(otu)
sp2 <- specaccum(otu, "random")
perm <- as.data.frame(sp2$perm)
permdata <- melt(sp2$perm)
boxplot(value~Var1,data=permdata, breaks=seq(0, 51,
3), col="gold", lwd=1.5, cex.axis=1.5, cex.lab=1.5,
xlab="Number of samples",ylab="Observed species")
```



图片的输出





1、直接输出

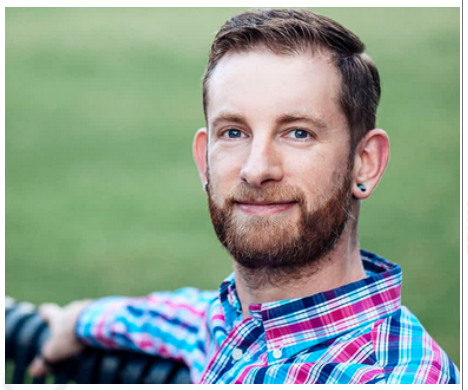
2、命令模式

```
pdf(file="file.pdf", width=7, height=10)
png(file="file.png",width=480,height=480)
jpeg(file="file.png",width=480,height=480)
tiff(file="file.png",width=480,height=480)
.....
dev.off()
```

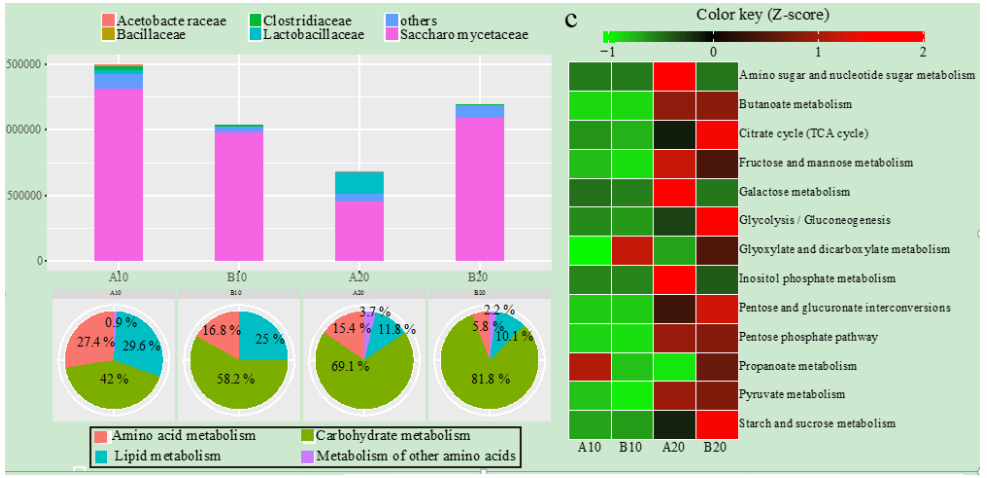
ggplot2绘图

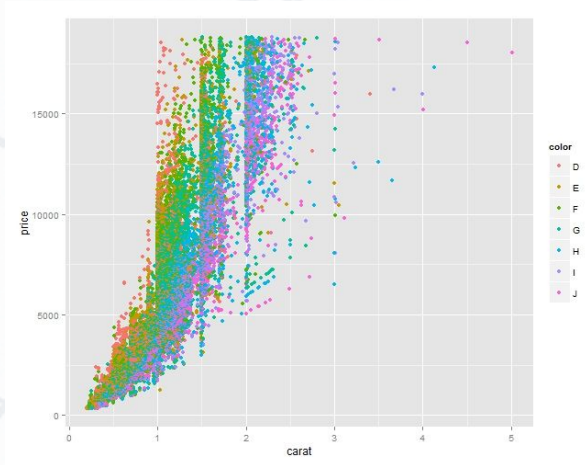
ggplot2简介

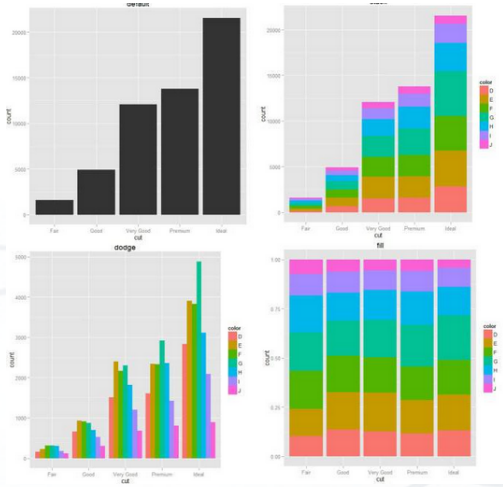
ggplot2是一个用来绘制统计图形(数据图形)的R包，背后也有一套独特的图形语法。

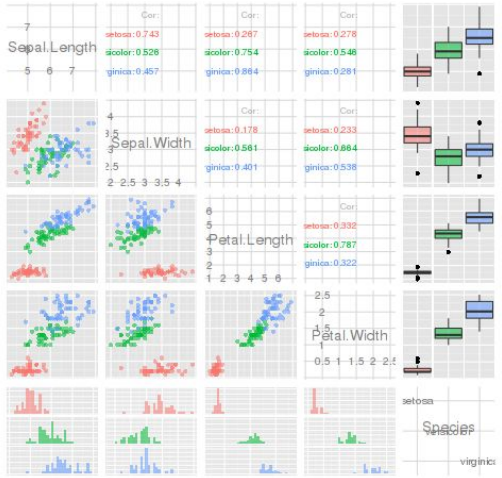


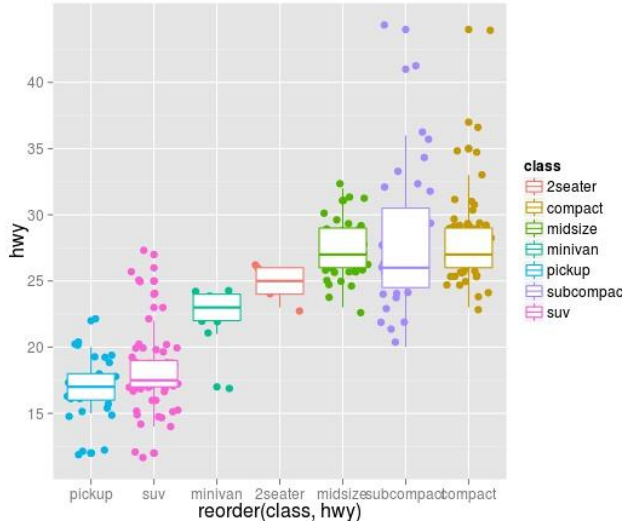
Hadley Wickham











Novogene 诺禾致源

Providing advanced genomic solutions!

ggplot2语法

可视化数据data ⊕ 图形的映射mapping

数据框格式　　aes(x, y)

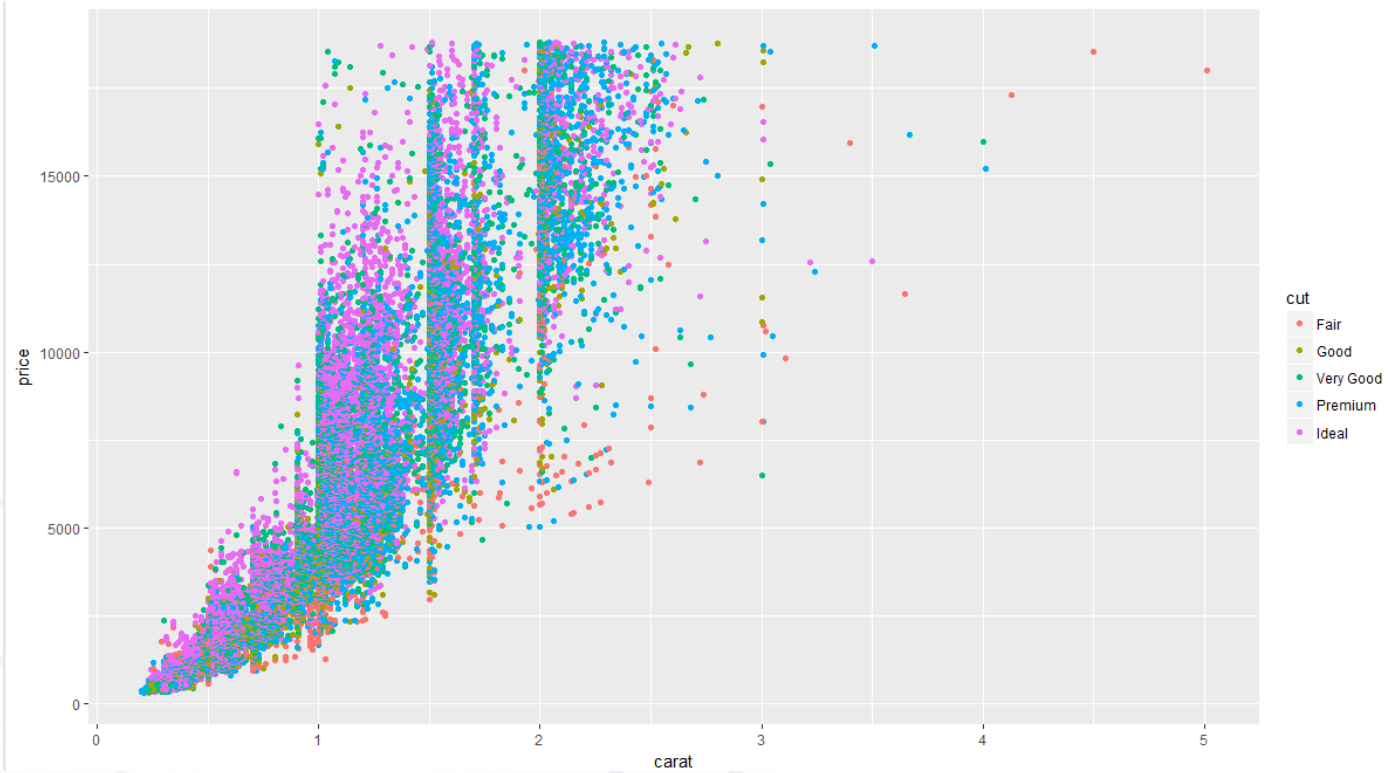
- 几何对象（geom）：实际看到的图形元素，点、线、多边形
- 统计变换（stats）：自动对数据进行某种汇总，直方图、线性模型
- 标度（scale）：将数据的取值映射到图形空间（颜色、形状或大小表示不同的数据）
- 坐标系（coord）：笛卡尔坐标系、极坐标系、对数坐标系
- 分面（facet）：将数据进行拆分并针对子数据集分别作图
- 图层(Layer)：数据+图形属性映射，一种统计变换，一种几何对象，一种位置调整方式。

绘图对象

ggplot2语法

```
ggplot(diamonds,aes(carat,price, colour=cut)) + geom_point()
```

绘图对象　　几何对象



ggplot2绘图

PCA

```
library(ade4)
library(ggplot2)
setwd("H:/2017~/培训/20180820_基因学院培训")
#读入数据
data <- read.table("data/otu_table.relative.xls", stringsAsFactors = F,
                   head=T, row.names=1,sep="\t", comment.char = "")
data <- data[, c(1:(ncol(data)-1))]
data <- t(data)
groups = read.table("data/group.list",
                    head=F,colClasses=c("character","character"))
#pca分析
pca<-dudi.pca(data[,1:ncol(data)], scannf=F, nf=5)
PC1 <-pca$li[,1]
PC2 <-pca$li[,2]
#绘图数据的整理
plotdata<-data.frame(rownames(pca$li),PC1,PC2,groups$V2)
colnames(plotdata) <-c("sample","PC1","PC2","groups")
pc1 <-round(pca$eig[1],2)
pc2 <-round(pca$eig[2],2)
```

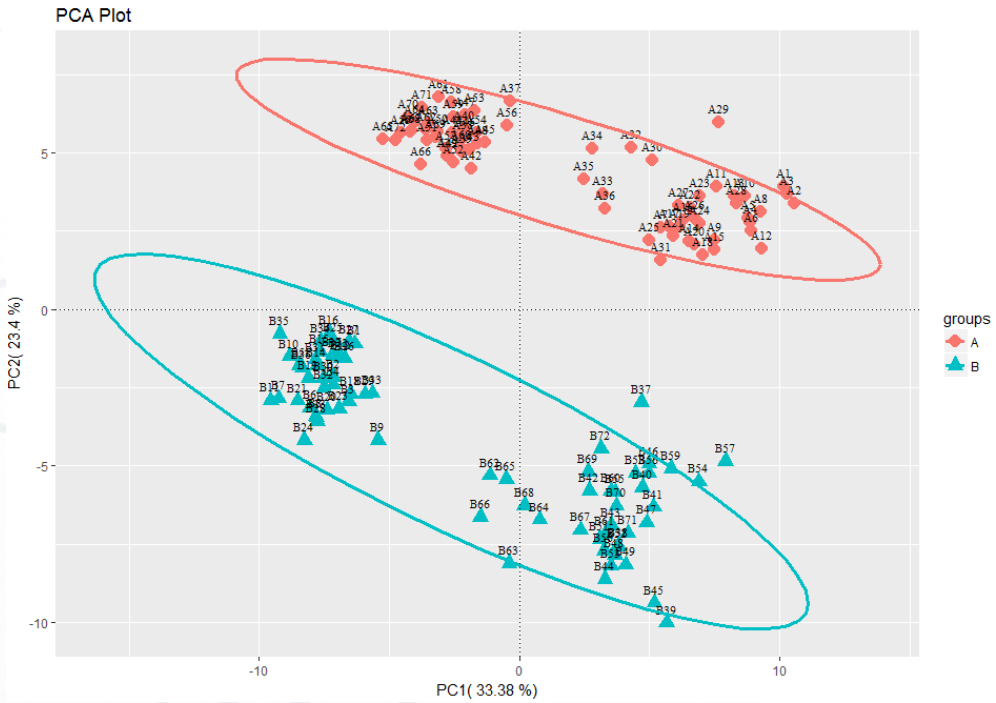
	V1	V2
1	A1	A
2	A2	A
3	A3	A
4	A4	A
5	A5	A
6	A6	A
7	A7	A
8	A8	A
9	A9	A
10	A10	A
11	A11	A

	A1	A2	A3	A4	A5	A
OTU_1	0.000000000	0.000000000	0.000021700	0.000021700	0.000021700	0
OTU_2	0.000065200	0.000087000	0.000108717	0.000152204	0.000173947	0
OTU_3	0.020395295	0.020286578	0.018264443	0.015524777	0.014894218	0
OTU_4	0.003348481	0.003935553	0.013546129	0.006392555	0.004718314	0
OTU_5	0.000239177	0.000195690	0.000173947	0.000239177	0.001021939	0
OTU_6	0.000087000	0.000021700	0.000000000	0.000087000	0.000195690	0
OTU_7	0.015503033	0.019460329	0.017481681	0.018851514	0.018047009	0
OTU_8	0.000369638	0.000521841	0.000413124	0.000630558	0.001108913	0
OTU_9	0.001130656	0.001021939	0.001804701	0.001739471	0.001108913	0
OTU_10	0.024374334	0.021091083	0.015720467	0.012611163	0.020069144	0
OTU_11	0.007958079	0.008392946	0.006305582	0.005848970	0.008349460	0

ggplot2绘图

PCA

```
#ggplot作图
P<-ggplot(plotdata, aes(PC1, PC2)) #绘图对象(data + mapping)
P<-P+geom_point(aes(colour=groups,shape=groups),size=4)
P<-P+geom_text(aes(label=sample),size=3,family="serif", hjust=0.5,vjust=-1)
P<-P+labs(title="PCA Plot",x=paste("PC1(",pc1,"%)"), y=paste("PC2(",pc2,"%)"))
P<-P+geom_vline(xintercept=0,linetype="dotted")
P<-P+geom_hline(yintercept=0,linetype="dotted")
P<-P+stat_ellipse(aes(group=groups, colour = groups), size=1.2)
```



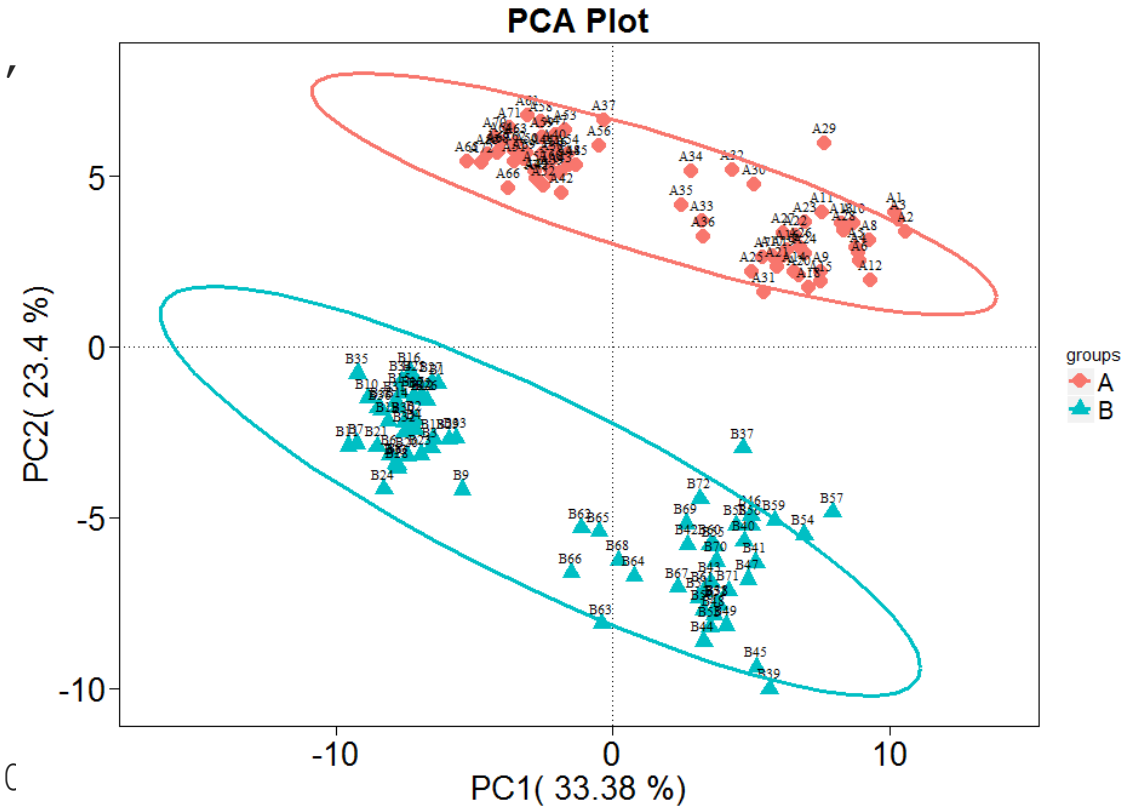
Novogene 诺禾致源

Providing advanced genomic solutions!

ggplot2绘图

PCA

```
#图形细节设置
P +  theme(panel.background = element_rect(fill='white',
colour='black'),
           panel.grid=element_blank(),
           #axis.title = element_text(color='black',
size=20),
           axis.ticks.length = unit(0.4,"lines"),
axis.ticks = element_line(color='black'),
           # axis.ticks.margin =
unit(0.6,"lines"),axis.line = element_line(colour =
"black"),
           axis.title.x=element_text(colour='black',
size=20),
           axis.title.y=element_text(colour='black',
size=20),
           axis.text=element_text(colour='black',size=20
           #legend.title=element_blank(),
           legend.text=element_text(size=15))+
      theme(plot.title = element_text(size=20,colour =
"black",face = "bold", vjust=0.5, hjust=0.5))
```



ggplot2绘图

CCA/RDA/DCA/NMDS

```
library(vegan)
library(ggplot2)
setwd("H:/2017~/培训/20180820_基因学院培训")
community = read.table("data/otu_table.relative.xls",
stringsAsFactors = F,
                  head=T, row.names=1,sep="\t",
comment.char = "")
community <- community[, c(1:(ncol(community)-1))]
community <- t(community)
envdata <- read.table("data/env.list",head=T,
row.names=1,sep="\t")
group <- read.table("data/group.list",
head=F,colClasses=c("character","character"))
```

otu_table.relative.xls

	A1	A2	A3	A4
OTU_1	0.000000000	0.000000000	0.000021700	0.000021700
OTU_2	0.000065200	0.000087000	0.000108717	0.000152204
OTU_3	0.020395295	0.020286578	0.018264443	0.015524777
OTU_4	0.003348481	0.003935553	0.013546129	0.006392555
OTU_5	0.000239177	0.000195690	0.000173947	0.000239177
OTU_6	0.000087000	0.000021700	0.000000000	0.000087000
OTU_7	0.015503033	0.019460329	0.017481681	0.018851514
OTU_8	0.000369638	0.000521841	0.000413124	0.000630558

env.list

	A	B	C	D	E	F	G
A1	8.17	0.5000000	0.08666667	1.0983551	9	142.72320	2.960
A2	8.17	0.6133333	0.12666667	1.1616870	12	229.54990	93.080
A3	8.06	0.5266667	0.25333333	1.1548166	9	206.43750	51.090
A4	8.13	0.5066667	0.16666667	1.2050198	7	120.30800	15.080
A5	8.08	0.4533333	0.25333333	1.4827793	9	122.41460	38.940
A6	8.18	0.5666667	0.16666667	1.2985625	12	172.37540	20.080
A7	8.03	0.5600000	0.30666667	1.3922382	12	153.28980	8.800

group.list

	V1	V2
1	A1	A
2	A2	A
3	A3	A
4	A4	A
5	A5	A
6	A6	A
7	A7	A

CCA/RDA/DCA/NMDS

```
cca<-cca(community, envdata, scale=T)
scorcca <- scores(cca)
sam <- data.frame(scorcca$sites, group$V2)   #提取样本得分
colnames(sam) <- c("CCA1","CCA2","group")
spec <- scorcca$species   #物种得分
spec <- as.data.frame(spec)
env <- cca$CCA$biplot[,c(1,2)] #环境因子得分
env <- as.data.frame(env)
cca1 =round(cca$CCA$eig[1]/sum(cca$CCA$eig)*100,2) #第一轴标签
cca2 =round(cca$CCA$eig[2]/sum(cca$CCA$eig)*100,2) #第二轴标签
```

	CCA1	CCA2	group
A1	-1.36210166	-0.8036976	A
A2	-1.27082473	-0.8048199	A
A3	-1.29508761	-0.7201427	A
A4	-1.30281090	-0.4206726	A
A5	-1.23087047	-0.6506768	A

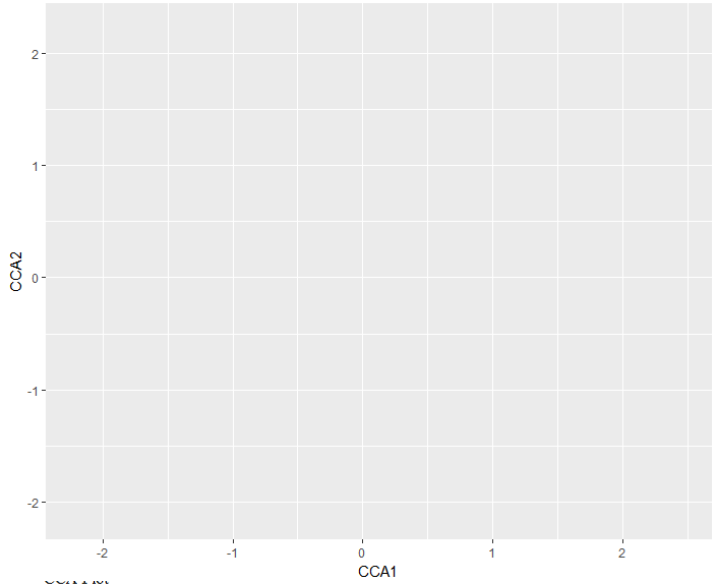
	CCA1	CCA2
OTU_1	1.1430926385	-0.035684056
OTU_2	0.9204109083	-0.238505565
OTU_3	-0.3887077214	-0.147354452
OTU_4	-0.0822042197	0.043417285
OTU_5	1.0124002464	-0.080052047

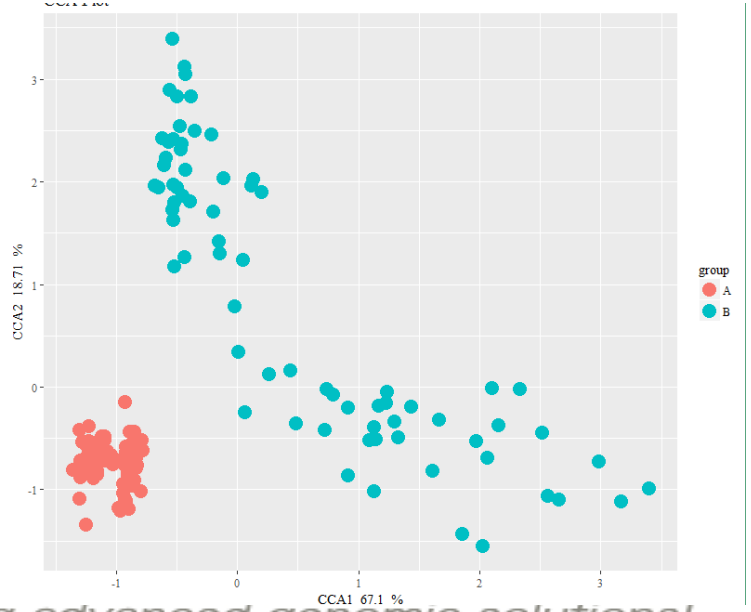
	CCA1	CCA2
A	-0.328298942	0.91550427
B	-0.365471896	-0.40131650
C	0.520723323	0.52146569
D	0.264999692	0.34250365
E	-0.442916520	-0.23640547

ggplot2绘图

CCA/RDA/DCA/NMDS

```
#绘图对象的创建
p <- ggplot(data=sam,aes(CCA1,CCA2))
#几何对象
p <- p + geom_point(aes(colour=group,shape=group),size=5) +
    #geom_text(aes(label=rownames(sam)),
     #          size=4,hjust=0.5,vjust=-0.7,position =
"jitter") +
    scale_shape_manual(values=c(19,19)) +
    labs(title="CCA Plot",x=paste("CCA1 ",cca1,"
%"),y=paste("CCA2 ",cca2," %")) +
    theme(text=element_text(family="serif"))
```





Novogene 诺禾致源

Providing advanced genomic solutions!

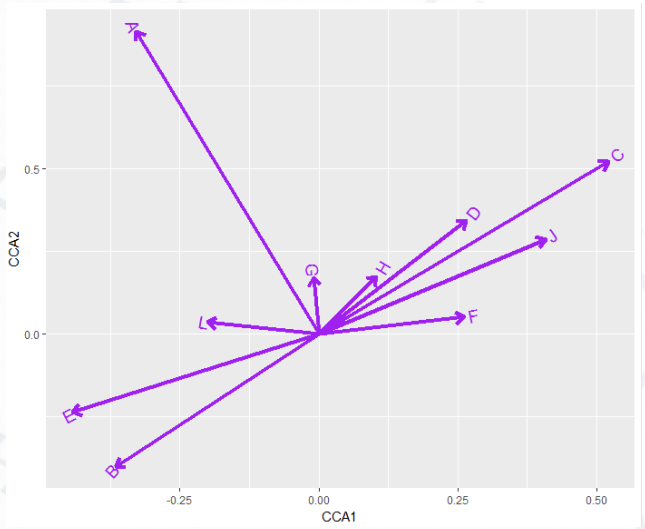
ggplot2绘图

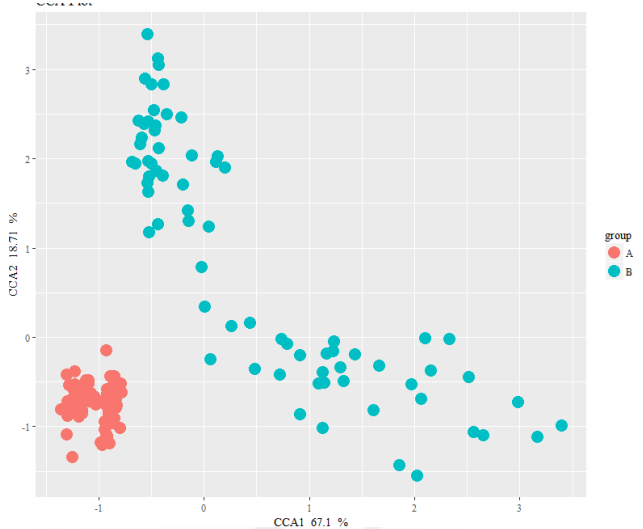
CCA/RDA/DCA/NMDS

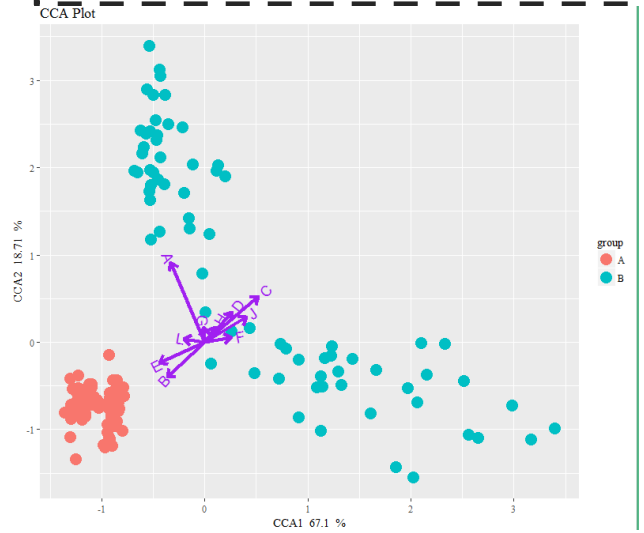
```
p + geom_segment(data = env,aes(x=0,y=0,xend = env[,1], yend = env[,2]), colour="purple",
size=1.5, arrow=arrow(angle=35, length=unit(0.3, "cm"))) +
    geom_text(data=env, aes(x=env[,1], y=env[,2], label=rownames(env)),
              size=5, colour="purple",hjust = (1 - 2 * sign(env[ ,1])) / 3,
              angle = (180/pi) * atan(env[ ,2]/env[ ,1]))
```

输出ggplot图

```
ggsave(filename="CCA.pdf",
plot=p, height=9, width=12)
```







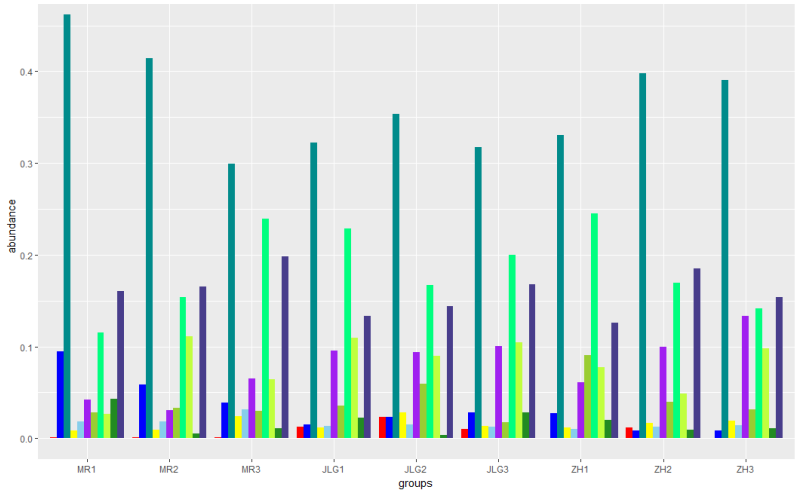
Novogene 诺禾致源

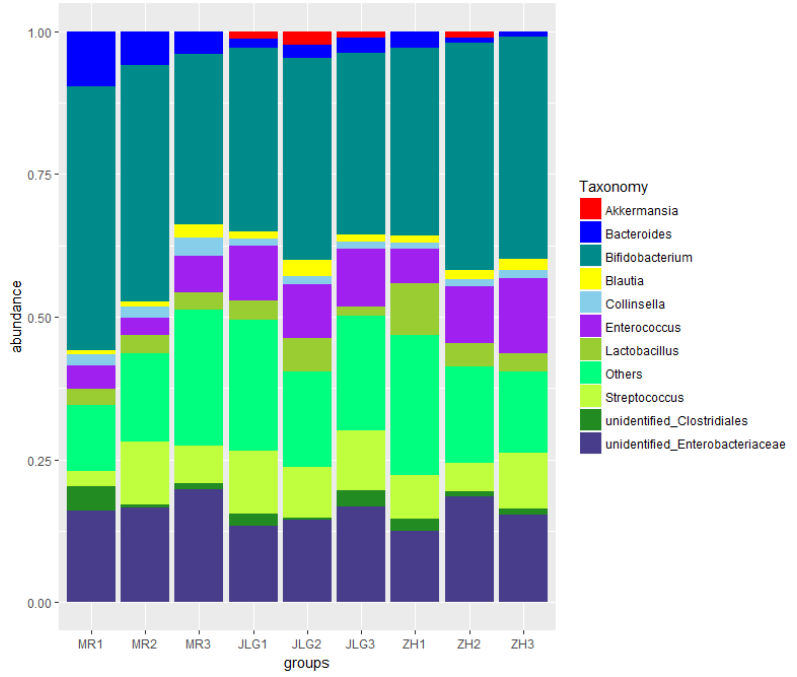
Providing advanced genomic solutions!

ggplot2绘图

柱状图

```
library(ggplot2)
library(reshape2)
setwd("H:/2017~/培训/20180820_基因学院培训")
otudata <-
read.table("data/otu_table.g10.group.relative.xls",header
=T)
meltdata <- melt(otudata, id.vars = 'Taxonomy',
value.name = "abundance", variable.name = 'groups')
col <- c('red','blue','cyan4','yellow','skyblue','purple',
'olivedrab3','springgreen','olivedrab1','forestgreen','da
rkslateblue')
ggplot(meltdata, aes(groups, abundance, group=Taxonomy,
fill=Taxonomy)) +
  geom_bar(stat = "identity", position = "dodge")+
  scale_fill_manual(values=col)
```





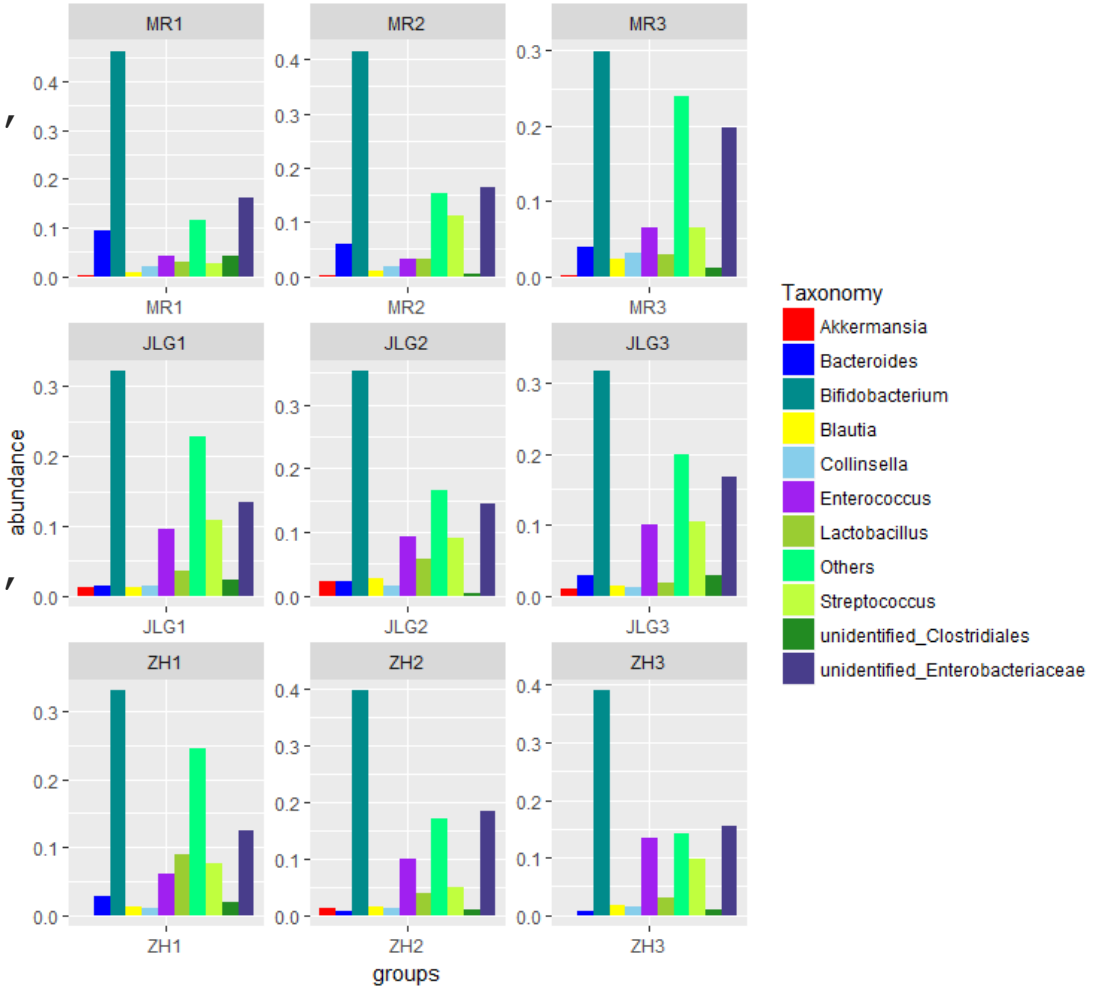
Novogene 诺禾致源

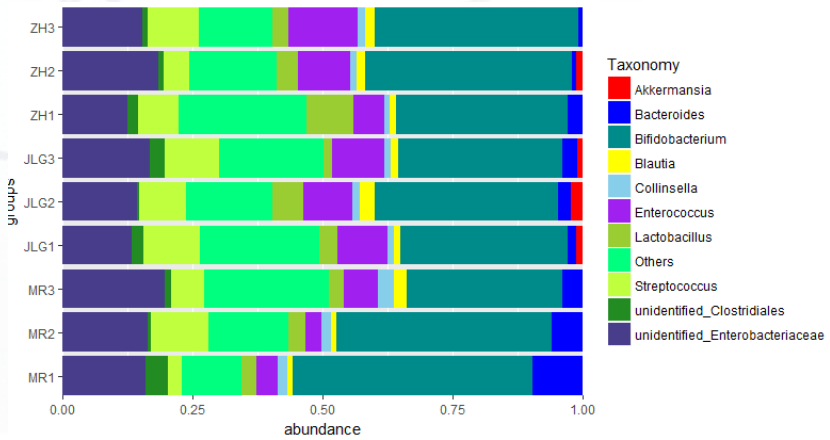
Providing advanced genomic solutions!

ggplot2绘图

柱状图+分面/坐标旋转

```
ggplot(meltdata, aes(groups, abundance, group=Taxonomy,
fill=Taxonomy)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_manual(values=col) +
  facet_wrap(~ groups, scales = 'free')
#==============================================
ggplot(meltdata, aes(groups, abundance, group=Taxonomy,
fill=Taxonomy)) +
  geom_bar(stat = "identity", position = "fill")+
  scale_fill_manual(values=col)+
  coord_flip()
```



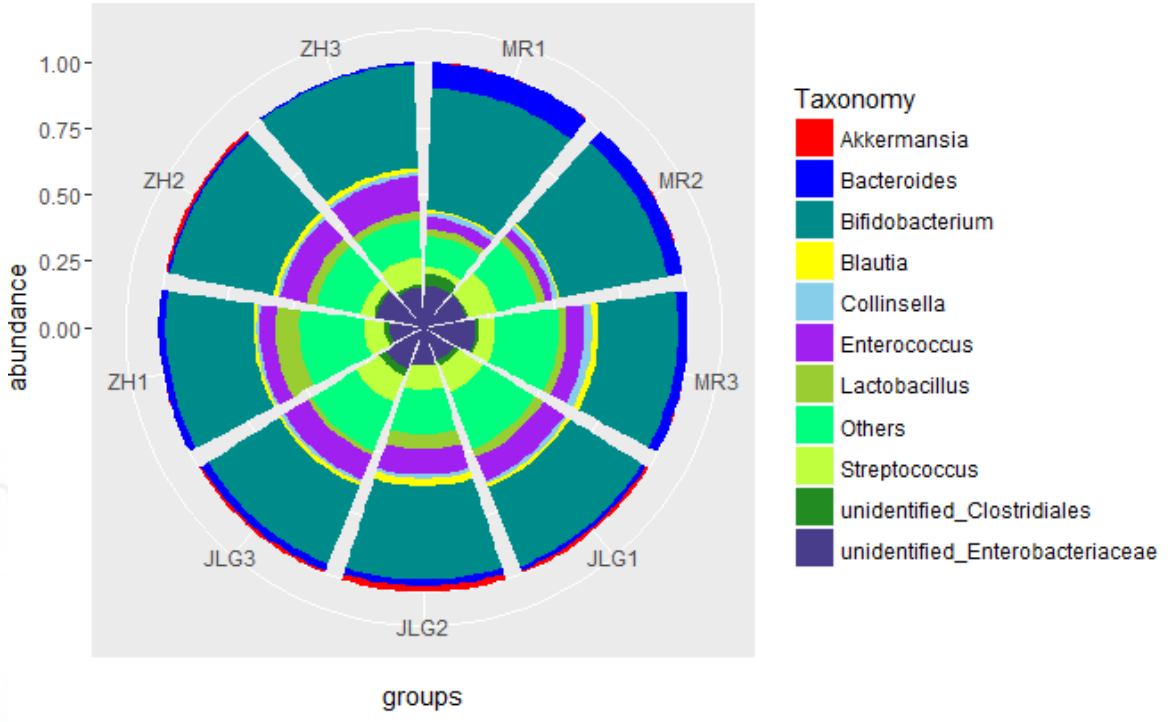


Novogene 诺禾致源

Providing advanced genomic solutions!

柱状图+坐标旋转

```
ggplot(meltdata, aes(groups, abundance, group=Taxonomy, fill=Taxonomy)) +
  geom_bar(stat = "identity", position = "stack")+
  scale_fill_manual(values=col)+
  coord_polar(theta = 'x')
```



Summary

数理统计

- 描述性统计
- 相关系数、线性回归、距离
- 区间估计、正态分布
- 假设检验

R绘图

- 描述性绘图：箱图
- 包绘图：热图、相关图、韦恩图、累积箱线图
- ggplot2绘图：PCA、CCA、柱状图等

数据处理

统计分析

图形展示

Novogene 诺禾致源

Providing advanced genomic solutions!

Thanks for your attention!

更多关注，敬请留意：www.novogene.cn