



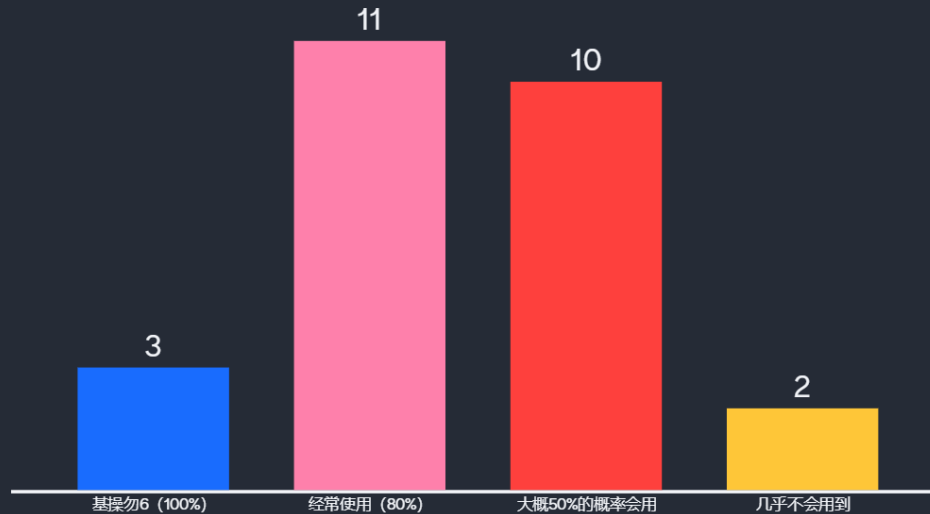
请同学们以组为单位
入座





你在将来的课题中有多大概率/频率会用到R/bioinformatic?

Mentimeter



26







应用生物信息学



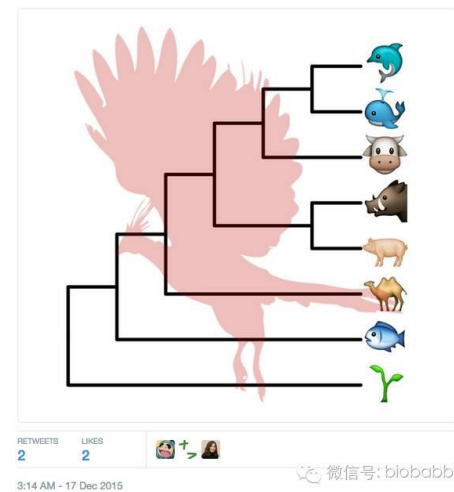
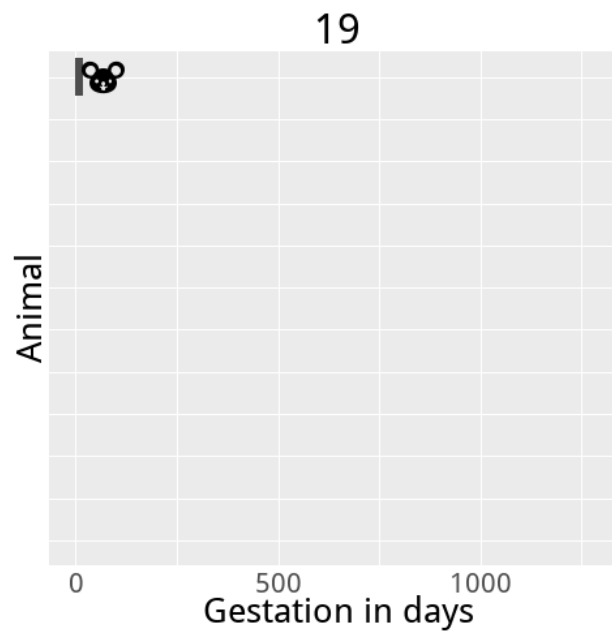
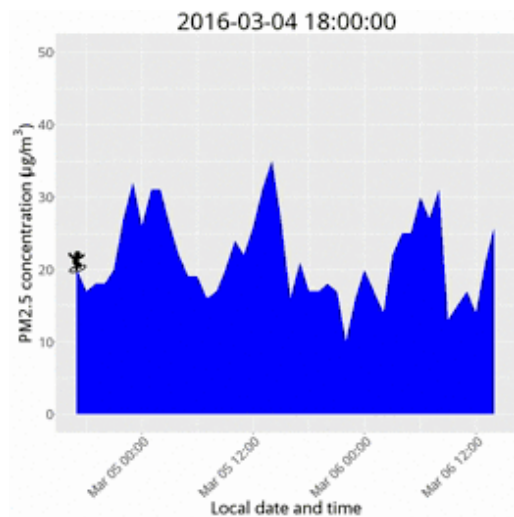
R统计与画图

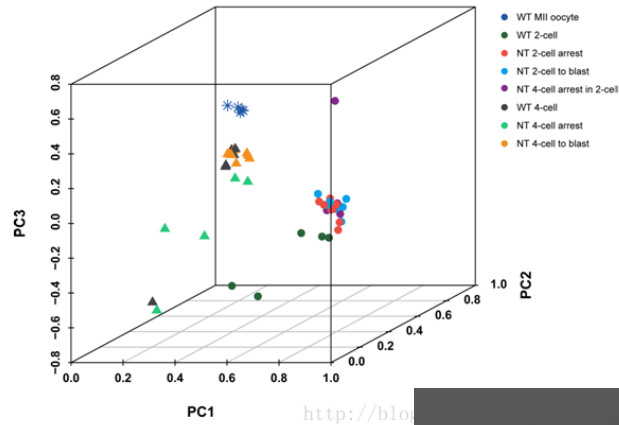
中南大学生命科学学院

刘可夫

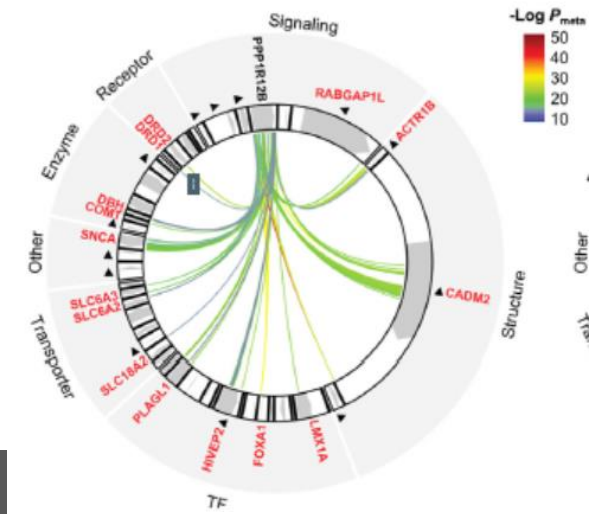
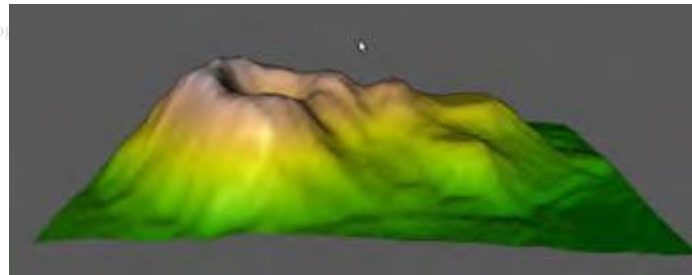
2021年



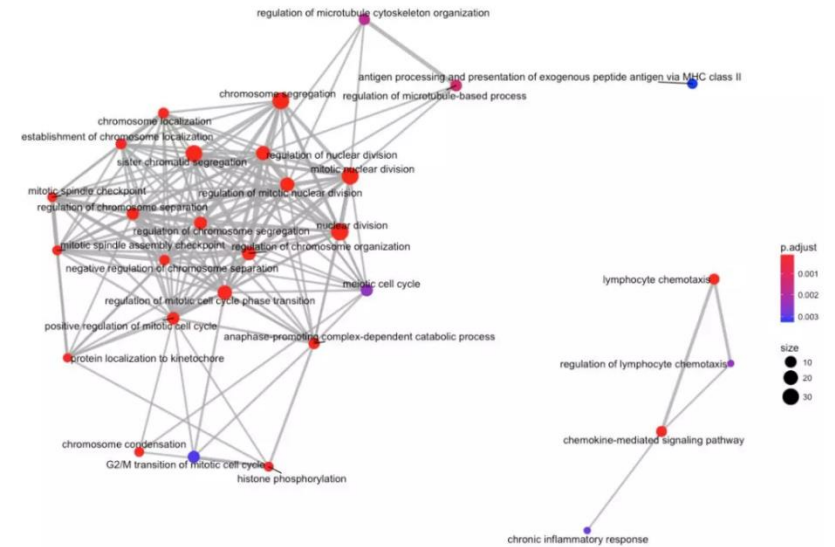
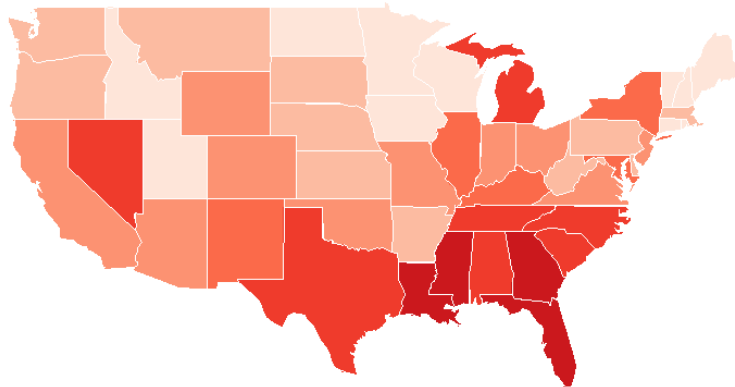




<http://blog>

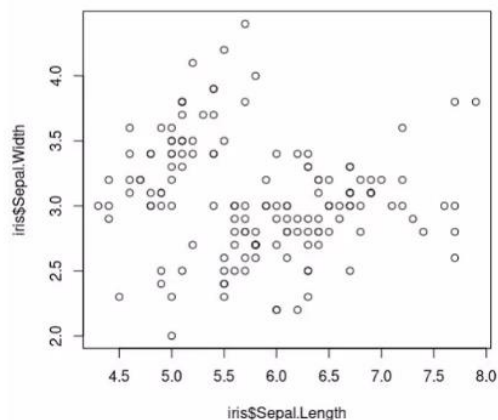


Murder Rates by US State in 1973
(arrests per 100,000 residents)



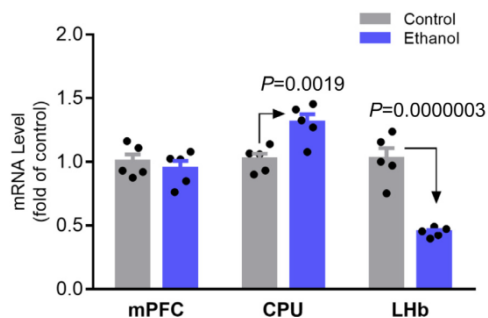


R base plot



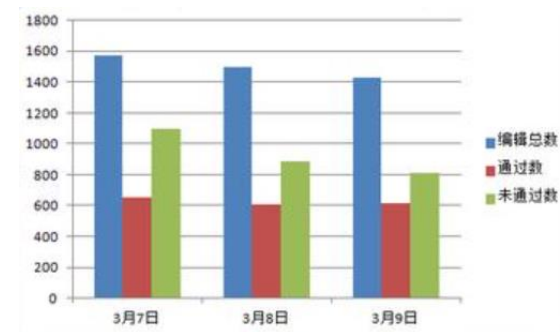
复古风, 代码保藏值高, 灵活度差

Graphpad prism



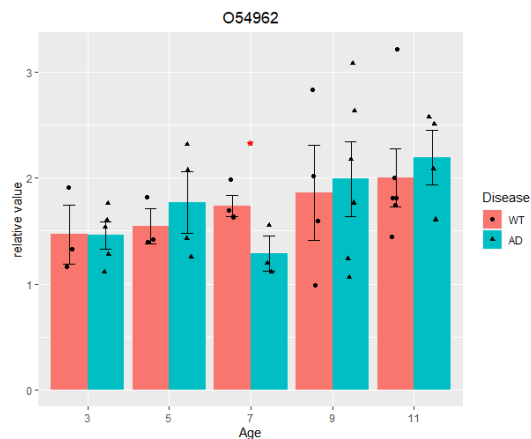
`github("csdaw/ggprism")`

excel

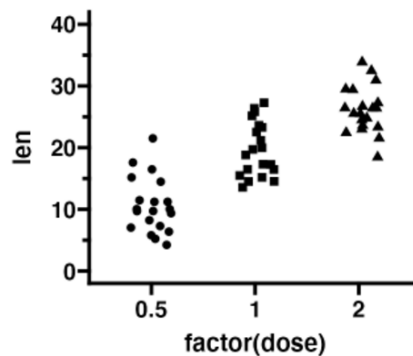


`theme_excel()`
`scale_fill_excel()`

ggplot2



美观, 代码保藏度低, 灵活度高

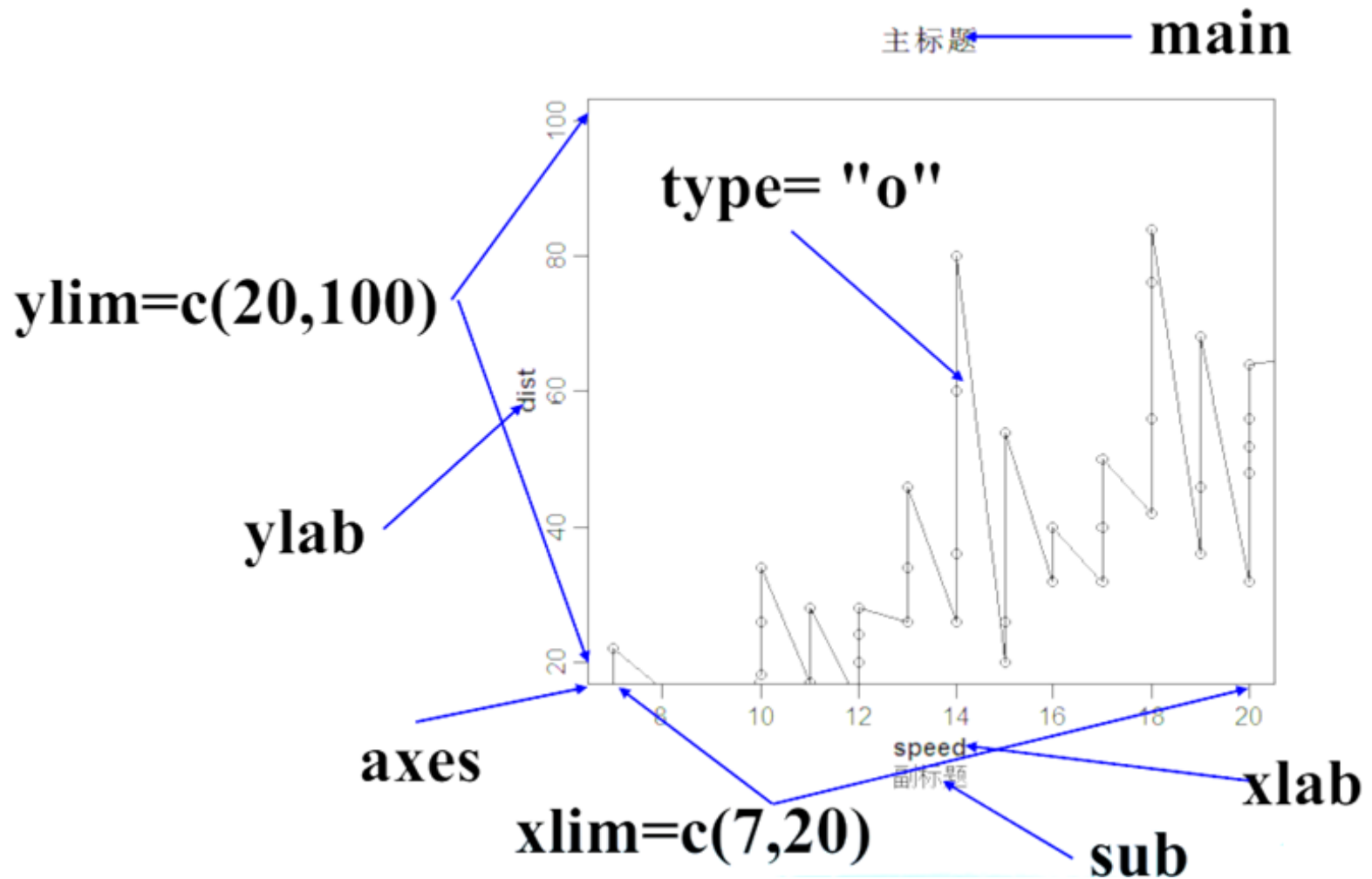




R base plot

R base plot

```
plot(cars,font.lab = 2,main = "主标题",sub = "副标题", type  
= "o", xlim = c(7,20), ylim = c(20, 100))
```





绘图步骤

1. 打开绘图窗口，不绘制任何对象

```
plot(x, y, type="n", xlab="", ylab="", axes=F)
```

2. 添加坐标点 **points(x,y)**

3. 添加坐标轴

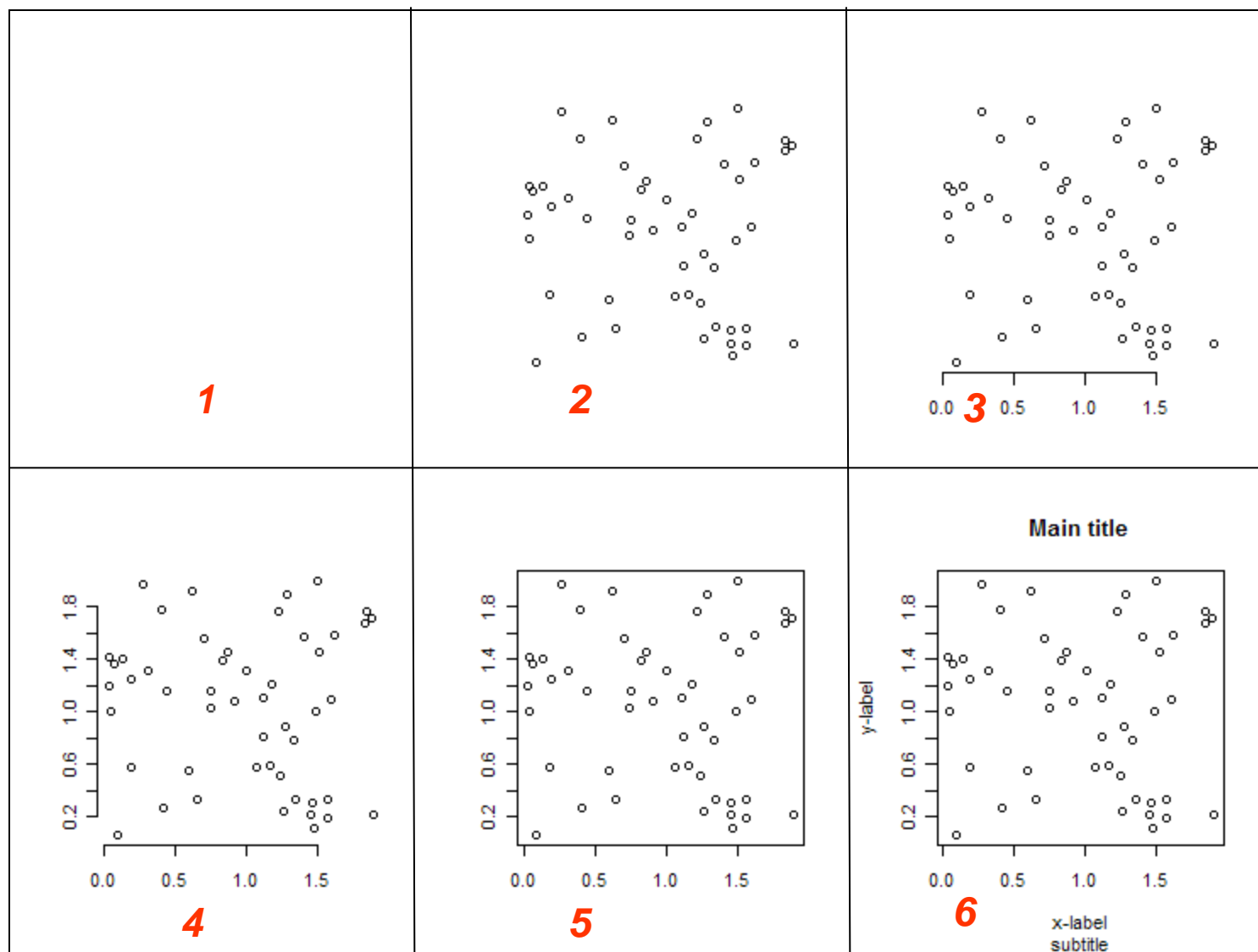
```
axis(1); axis(at=seq(0.2,1.8,0.2), side=2)
```

4. 补齐散点图的边框 **box()**

5. 添加标题、副标题、横轴说明、纵轴说明

```
title(main="Main title", sub="subtitle", xlab="x-label", ylab="y-label")
```





一般绘图步骤

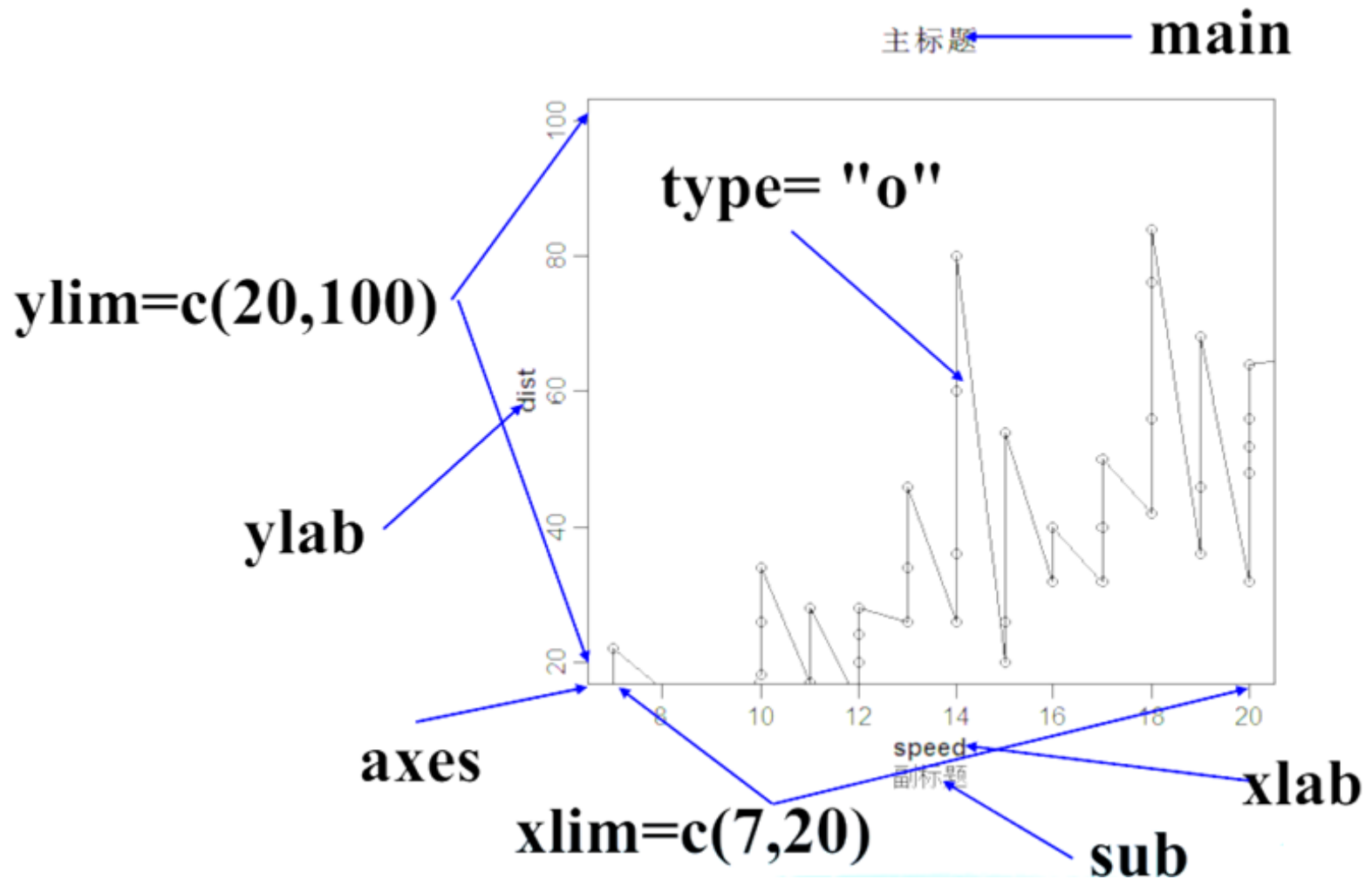




R base plot

R base plot

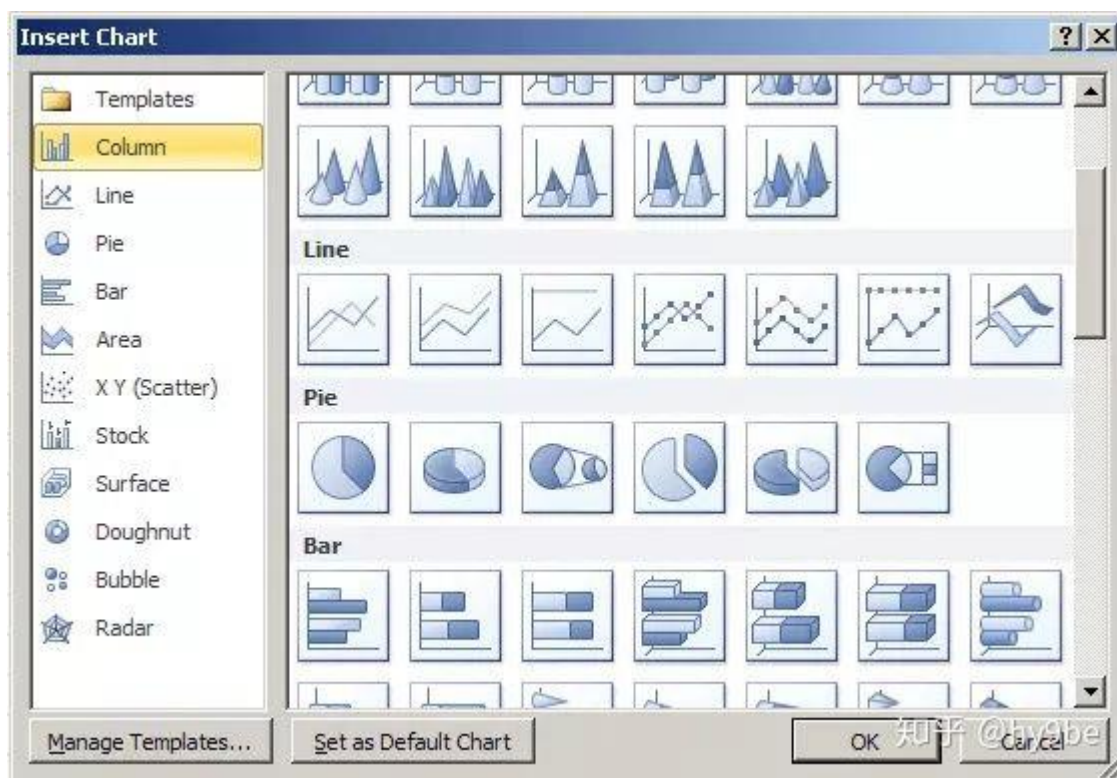
```
plot(cars,font.lab = 2,main = "主标题",sub = "副标题", type  
= "o", xlim = c(7,20), ylim = c(20, 100))
```





列举法

ink on paper



每种图形有自己单独的绘图逻辑，数据特征，格式设置

也是R base plot的绘图逻辑

所以R base plot 有高级绘图函数 每一种函数对应一种类型图 (barplot,

一旦有大的需求变化，就需要重新开发

只能在图形的最顶端进行绘画，而不能修改或删除已有的内容

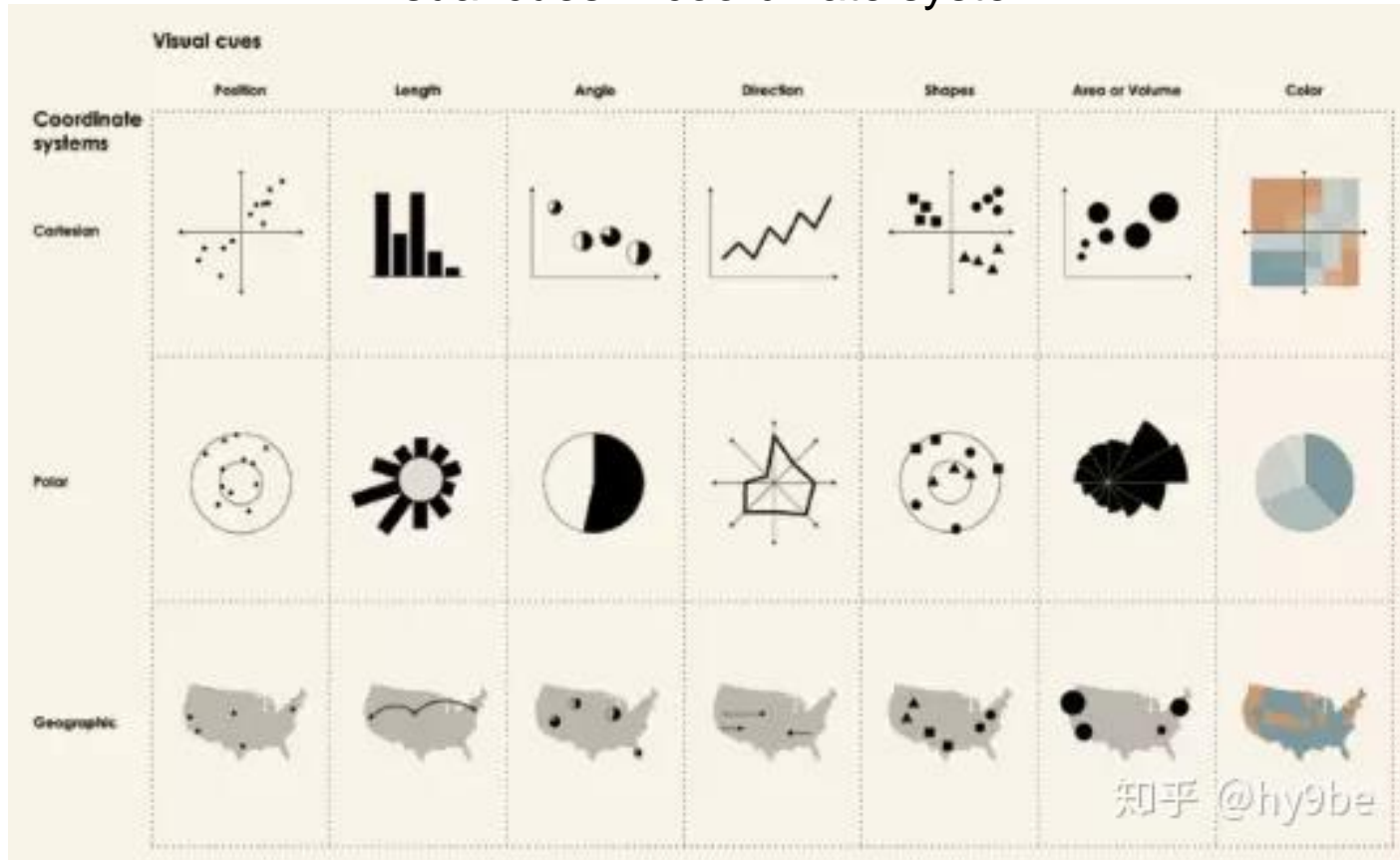




图形语法

Leland Wilkinson 《The Grammar of Graphics》

Visual cues X coordinate system



基于图形语法的可视化工具的特征是：生成每一个图形的过程就是组合不同的基础图形语法的过程。

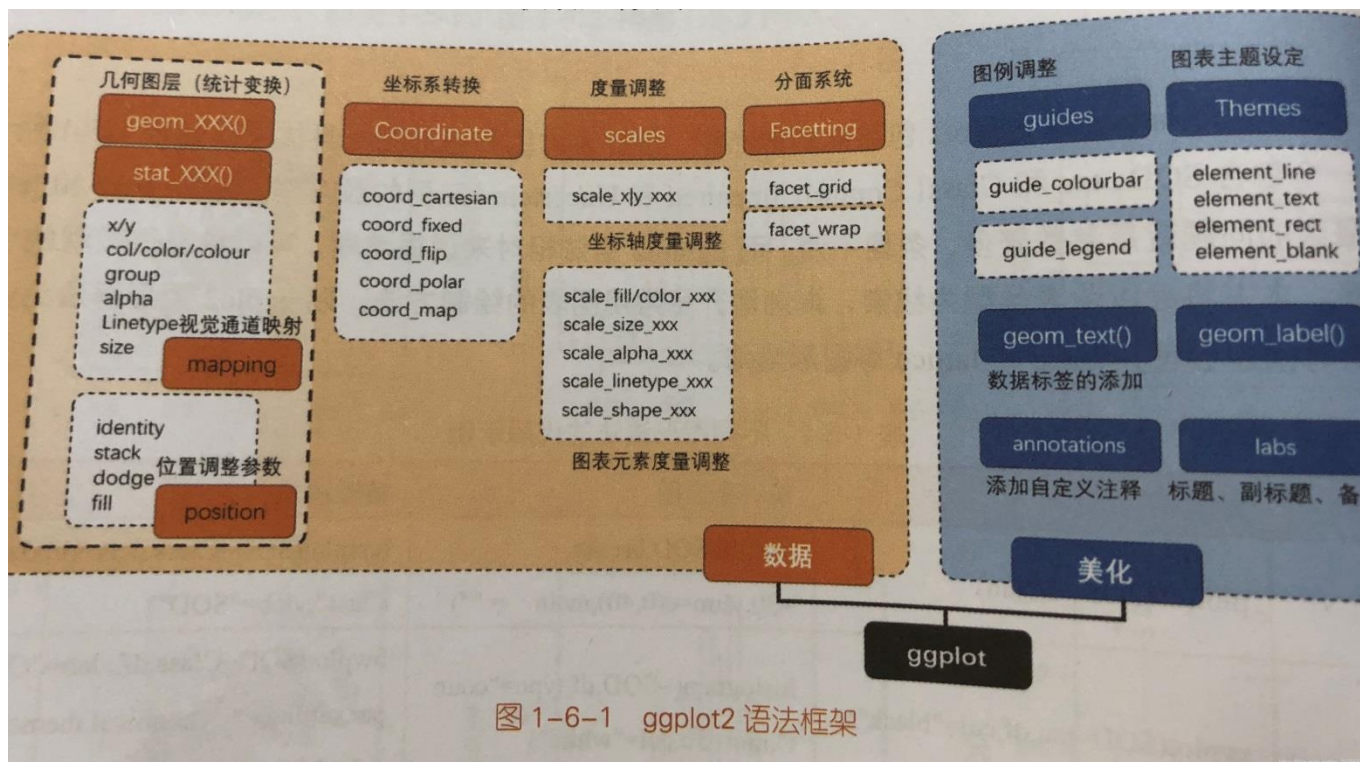


Hadley Wickham

ggplot2



1. 采用图层的设计方式，有利于结构化思维实现数据可视化。图层之间的叠加是靠“+”实现的。
2. 将表征数据和图形的细节分开，能快速将图形表现出来，使创造性的绘图更加容易实现。
3. 图形美观，扩展包丰富，有专门调整颜色（color），字体（font）和主题（theme）等辅助包。可以帮助用户快速指定个性化的图表。




```
ggplot(data = <DATA>, mapping = aes(<MAPPINGS>),  
  stat = <STAT>, position = <POSITION>) +
```

基础图层, 不出现图形元素,

```
geom_xxx() | stat_xxx() + # 几何图层或统计变换, 出现图形元素
```

```
scale_xxx() + # 标度调整, 调整具体的标度
```

```
coord_xxx() + # 坐标变换, 默认笛卡尔坐标系
```

```
facet_xxx() + # 分面系统, 将某个变量进行分面变换
```

```
guides() + # 图例调整
```

```
theme() # 主题设定
```

必须

可选

数据 (Data) 和映射 (Mapping)

将数据中的变量映射到图形属性。映射控制了二者之间的关系。

dat

length	width	depth	trt
2	3	4	a
1	2	1	a
4	5	15	b
9	10	80	b

mapping



x	y	colour
2	3	a
1	2	a
4	5	b
9	10	b

```
ggplot(data = dat,  
       mapping = aes(x = length, y = width, colour = trt))
```




标度 (Scale)

标度负责控制映射后图形属性的显示方式。具体形式上来看是图例和坐标刻度。Scale和Mapping是紧密相关的概念。

```
scale_colour_manual(values=c("green", "blue"))
```

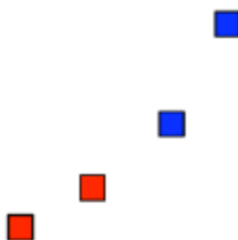
```
, scale_x_log10()
```





几何对象 (Geometric)

几何对象代表我们在图中实际看到的图形元素，如点、线、多边形等。



Geoms

```
geom_point()
```





统计变换 (**statistics**)

对原始数据进行某种计算，例如对二元散点图加上一条回归线。



Geoms

Stat

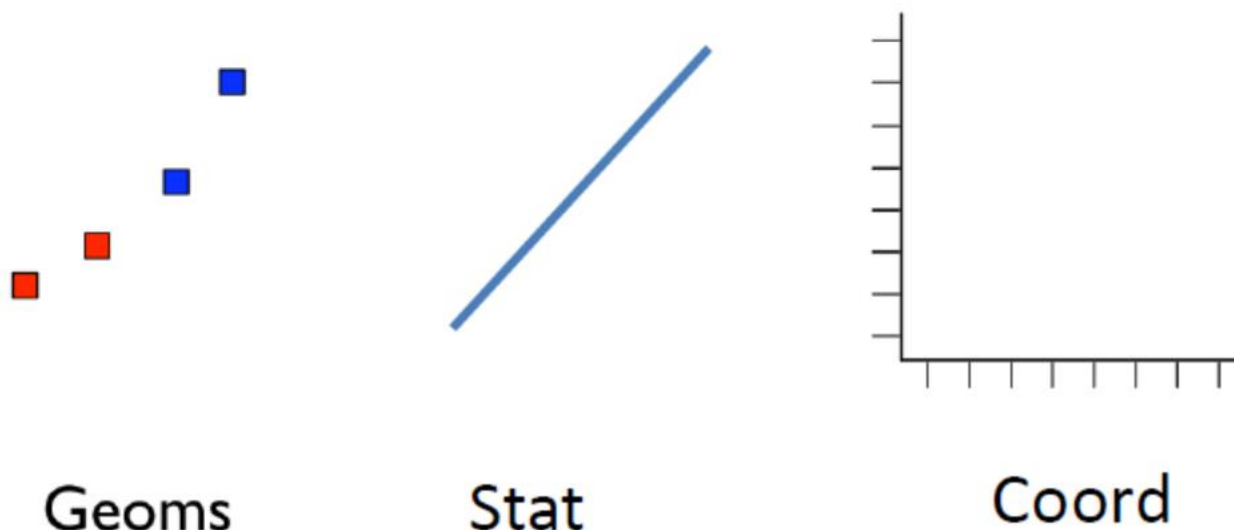
`stat_smooth()`





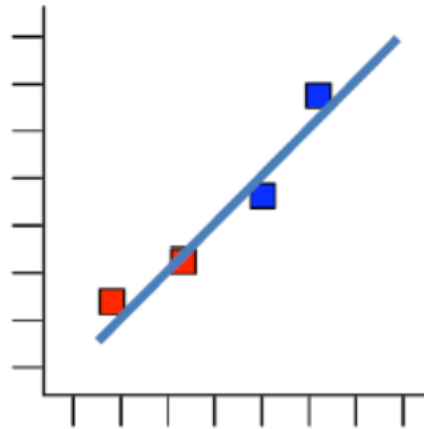
坐标系 (Coordinate)

坐标系统控制坐标轴并影响所有图形元素，坐标轴可以进行变换以满足不同的需要。



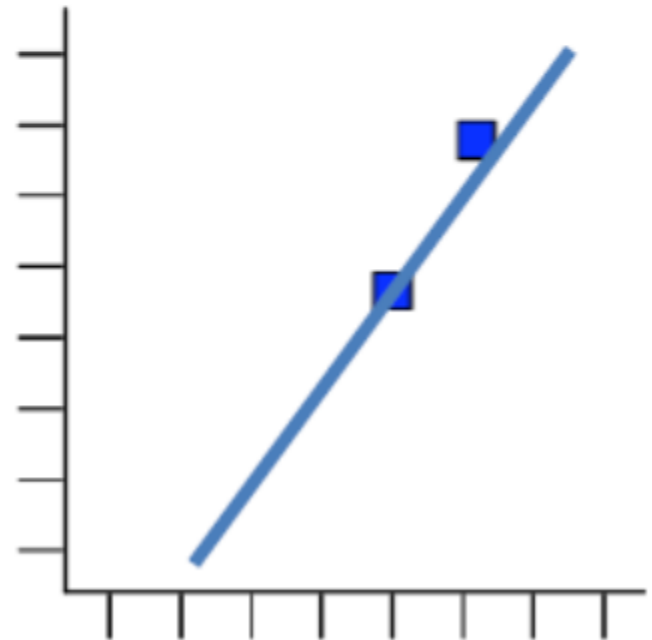
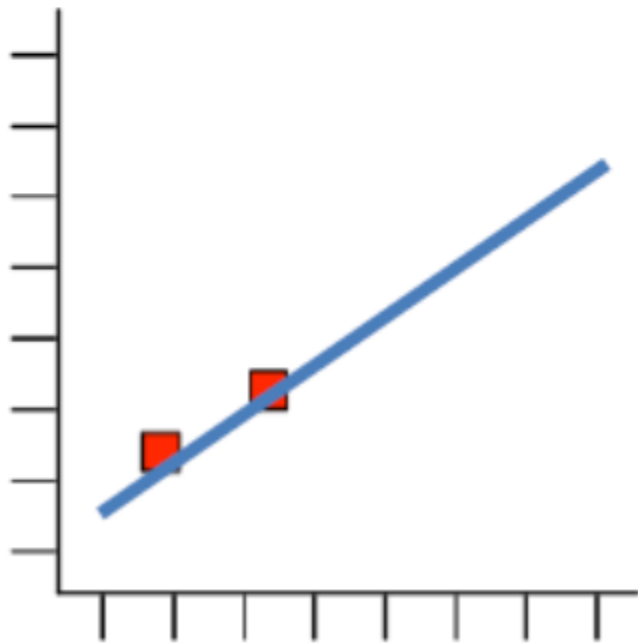
图层 (Layer)

数据、映射、几何对象、统计变换等构成一个图层。
图层可以允许用户一步步的构建图形，方便单独对图层进行修改。



分面 (Facet)

条件绘图，将数据按某种方式分组，然后分别绘图。
分面就是控制分组绘图的方法和排列形式。





ggplot2的基本概念

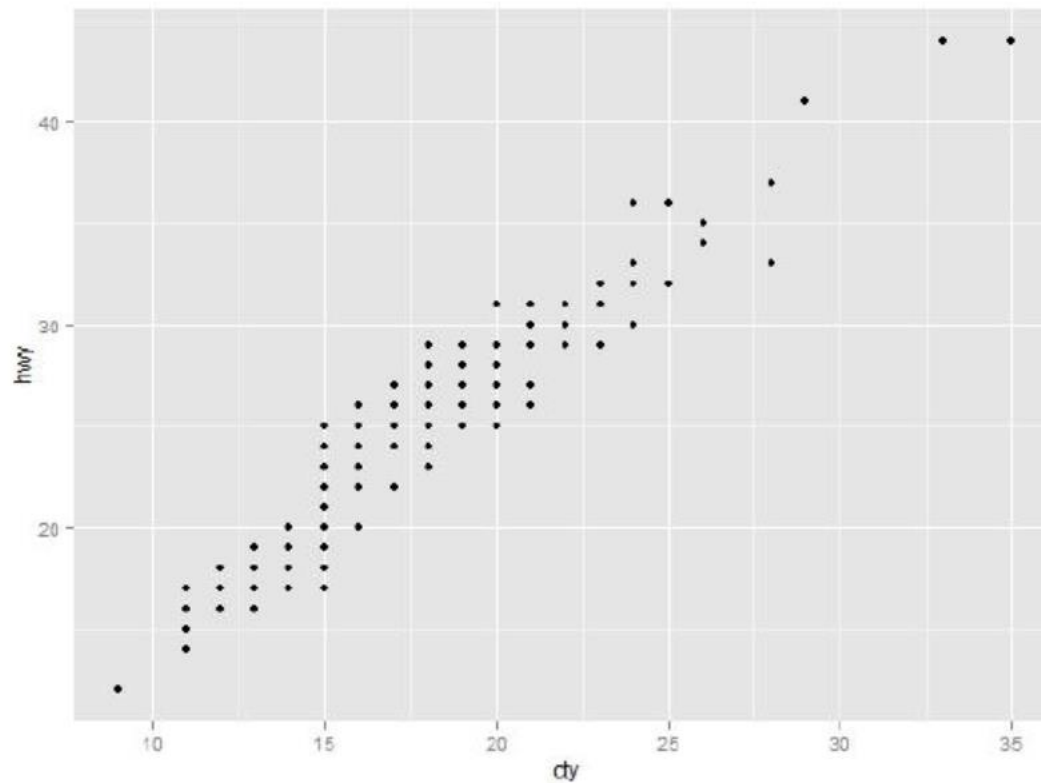
- 数据 (Data) 和映射 (Mapping)
- 标度 (Scale)
- 几何对象 (Geometric)
- 统计变换 (Statistics)
- 坐标系统 (Coordinate)
- 图层 (Layer)
- 分面 (Facet)





```
> library(ggplot2)  
> p <- ggplot(data=mpg, mapping=aes(x=cty, y=hwy))  
> p + geom_point()
```

aesthetics

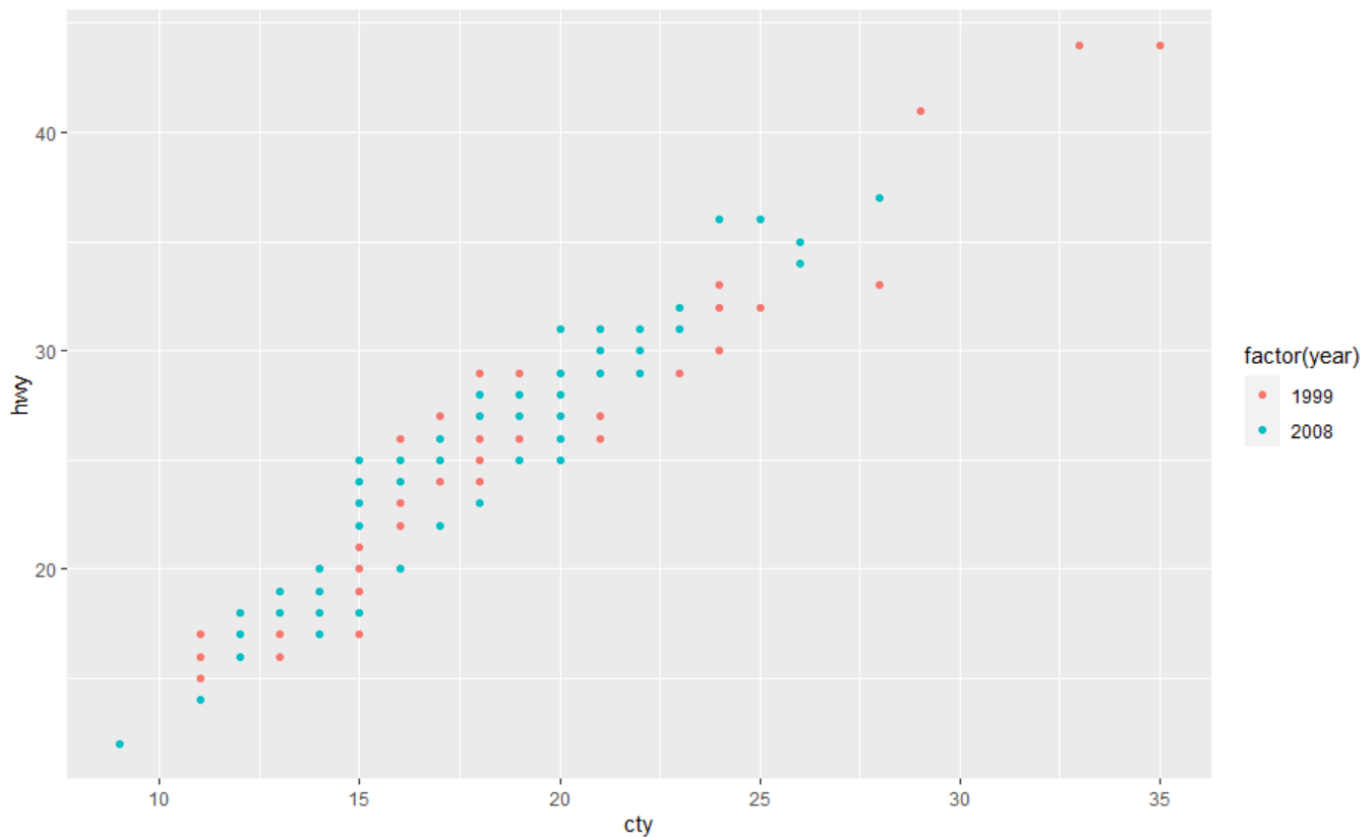




依据年份分组

ggplot2

```
ggplot(data = mpg, mapping = aes(x = cty, y = hwy, colour = year)) +  
  geom_point()
```

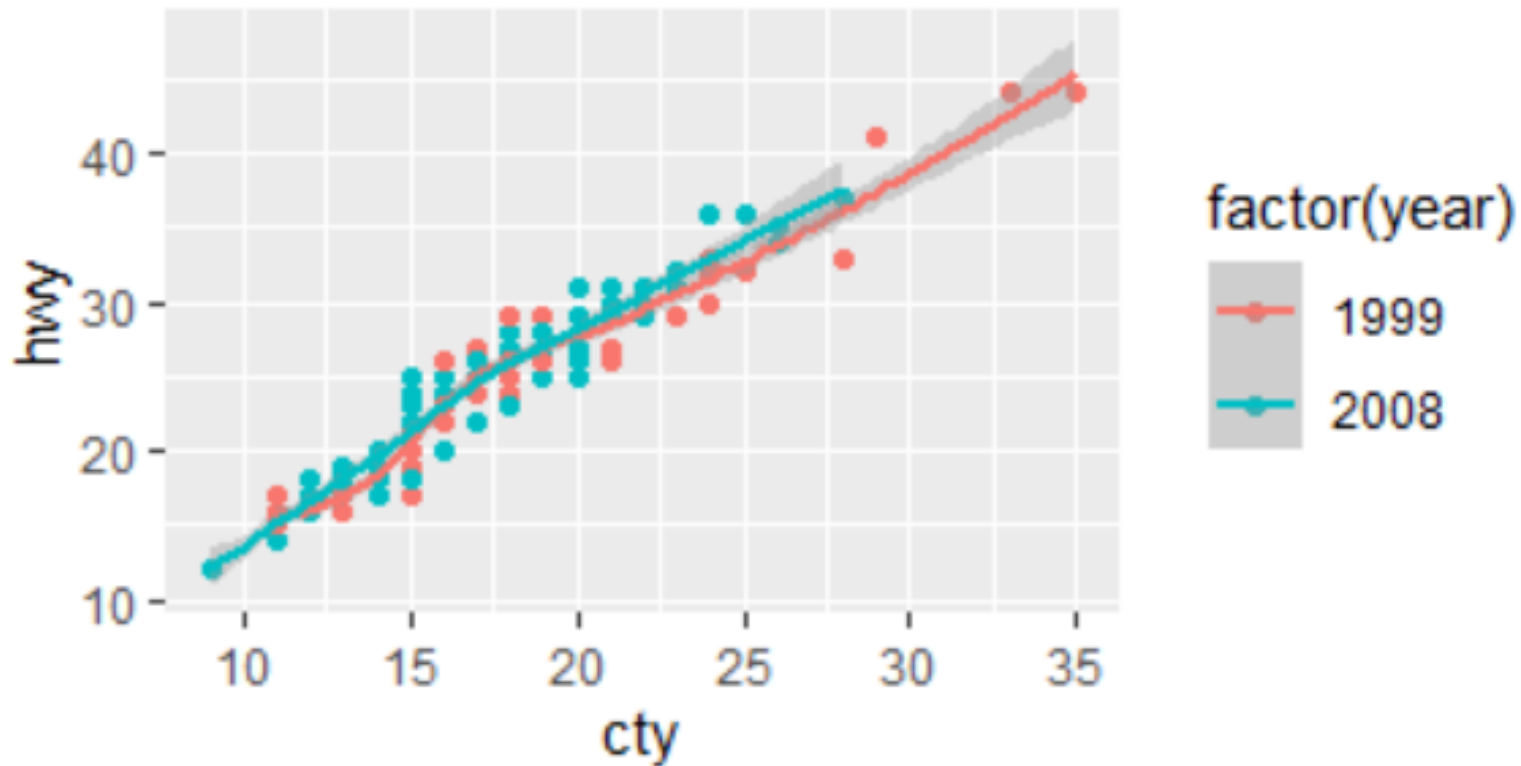




添加平滑曲线

ggplot2

```
ggplot(data = mpg, mapping = aes(x = cty, y = hwy, colour = factor(year))) +  
  geom_point() +  
  stat_smooth()
```

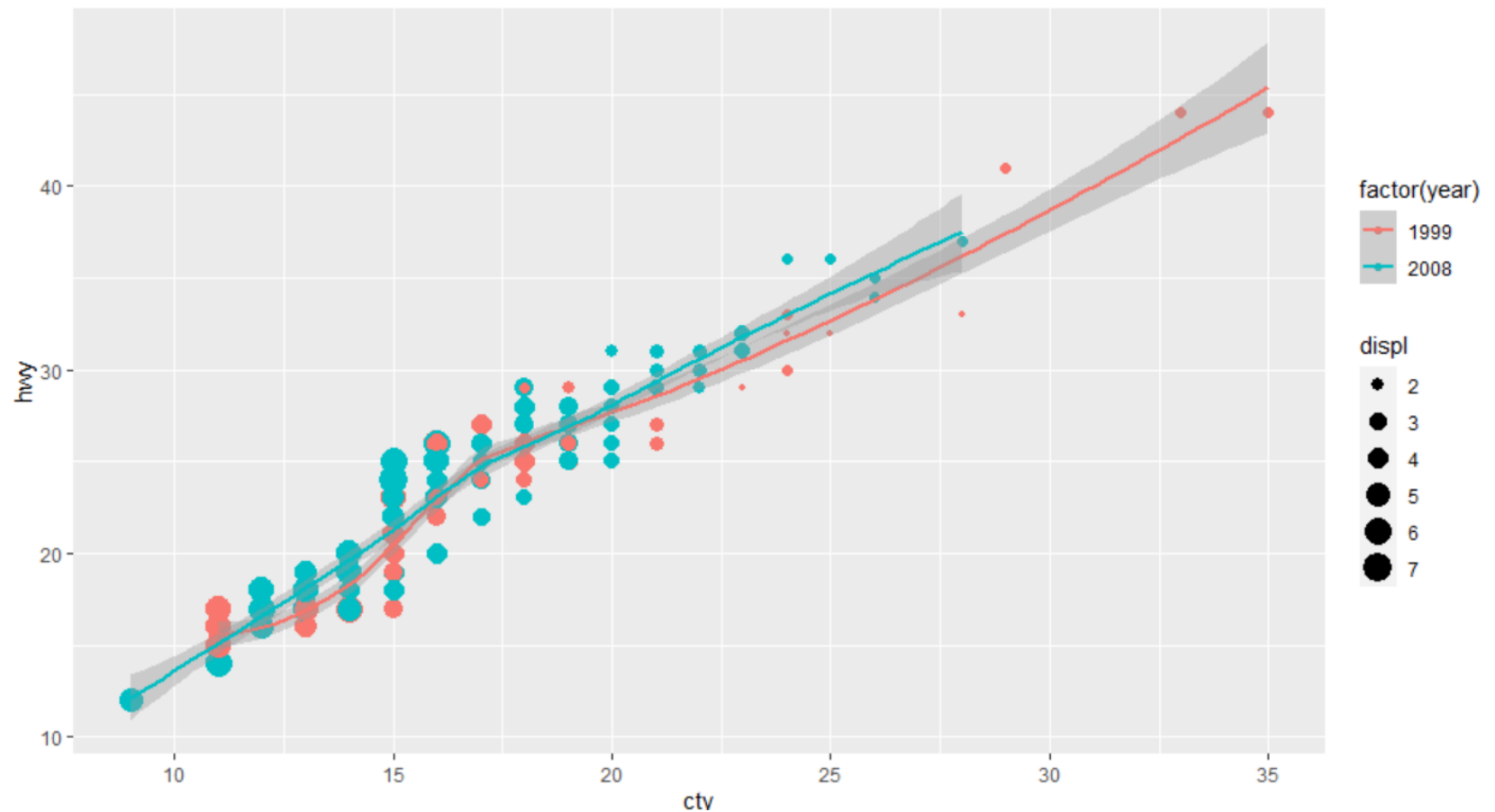




将disp 排量大小映射到点大小

ggplot2

```
ggplot(data = mpg, mapping = aes(x = cty, y = hwy, colour = factor(year))) +  
  geom_point(aes(size = displ)) +  
  stat_smooth()
```

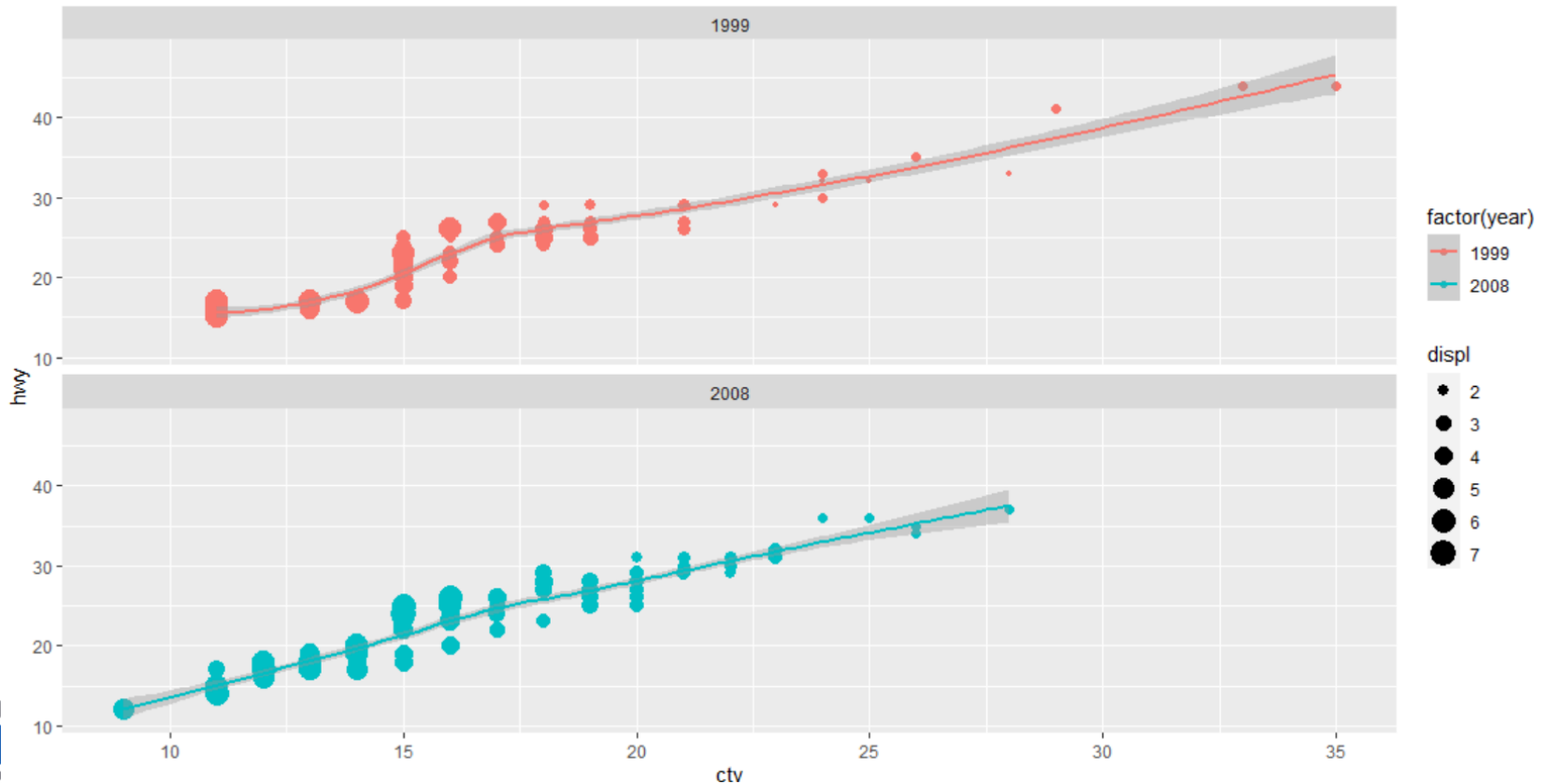




分面

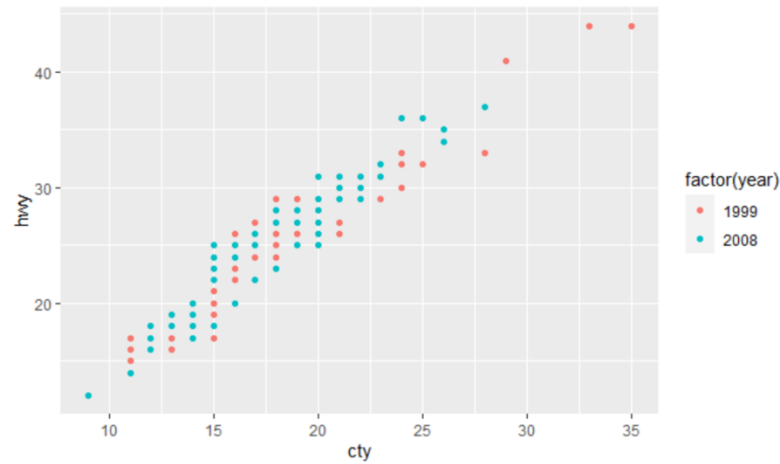
ggplot2

```
ggplot(data = mpg, mapping = aes(x = cty, y = hwy, colour = factor(year))) +  
  geom_point(aes(size = displ)) +  
  stat_smooth() +  
  facet_wrap(~year, ncol = 1)
```



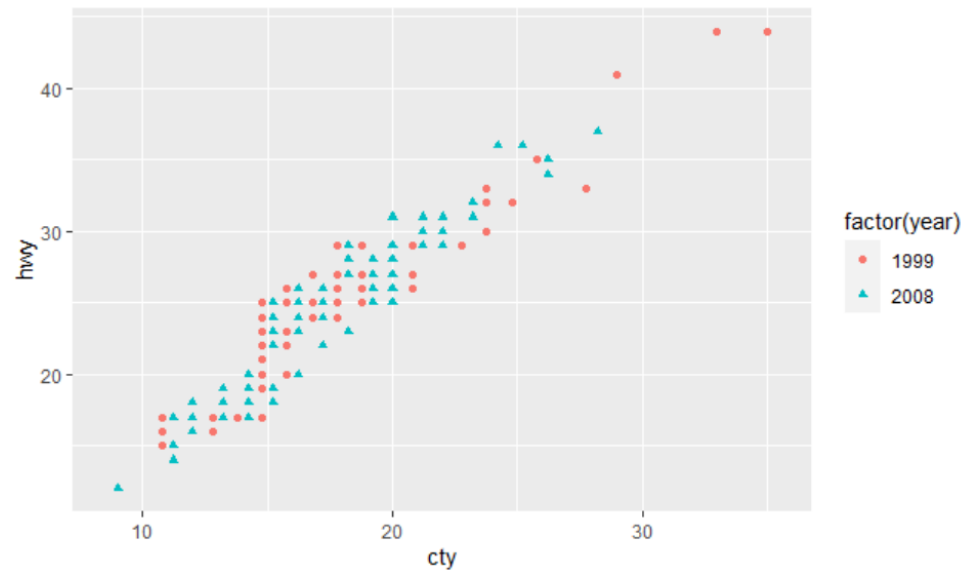


ggplot2



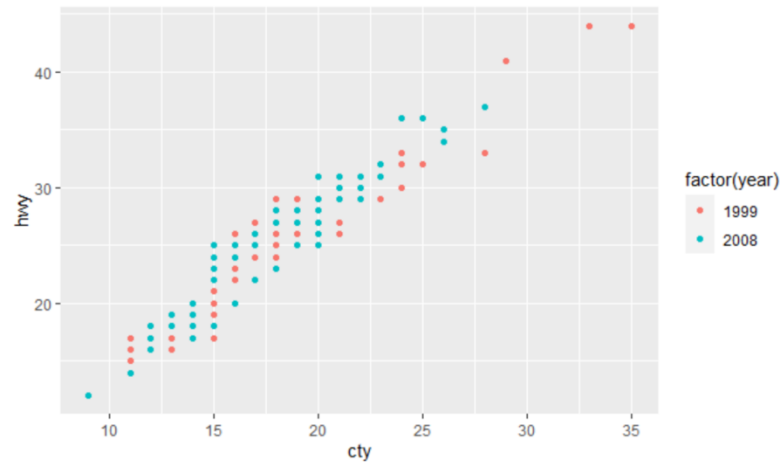
躲避
分组数据

```
ggplot(data = mpg, mapping = aes(x = cty, y = hwy, colour = factor(year))) +  
  geom_point(aes(shape = factor(year)), size = 1.5,  
    position = position_dodge(width = 0.9))
```



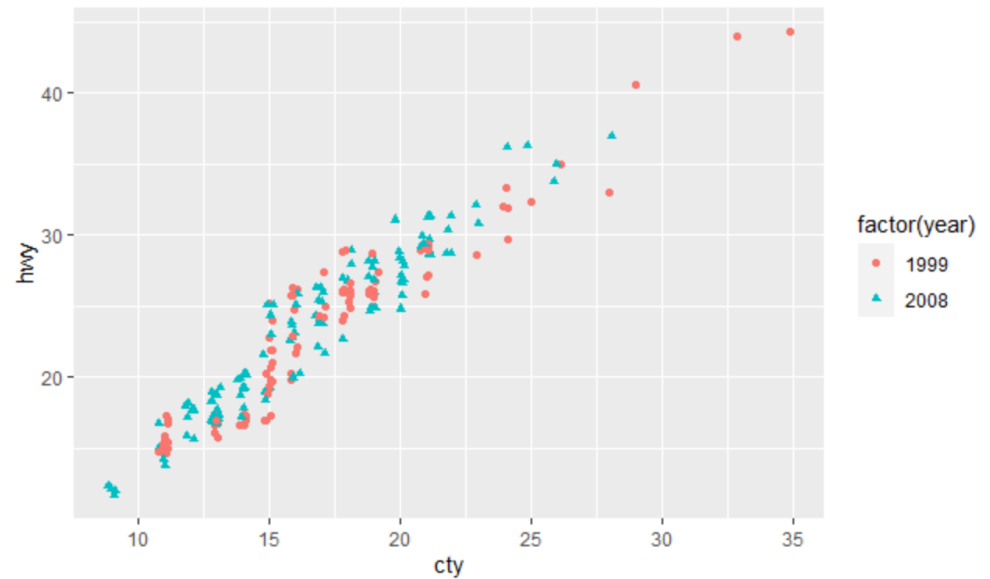


ggplot2



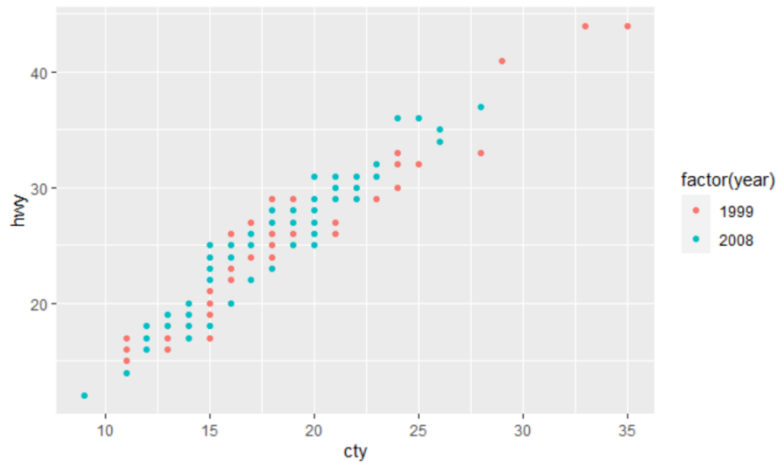
抖动
数据随机抖动

```
ggplot(data = mpg, mapping = aes(x = cty, y = hwy, colour = factor(year))) +  
  geom_point(aes(shape = factor(year)), size = 1.5,  
    position = position_jitter(width = 0.2))
```



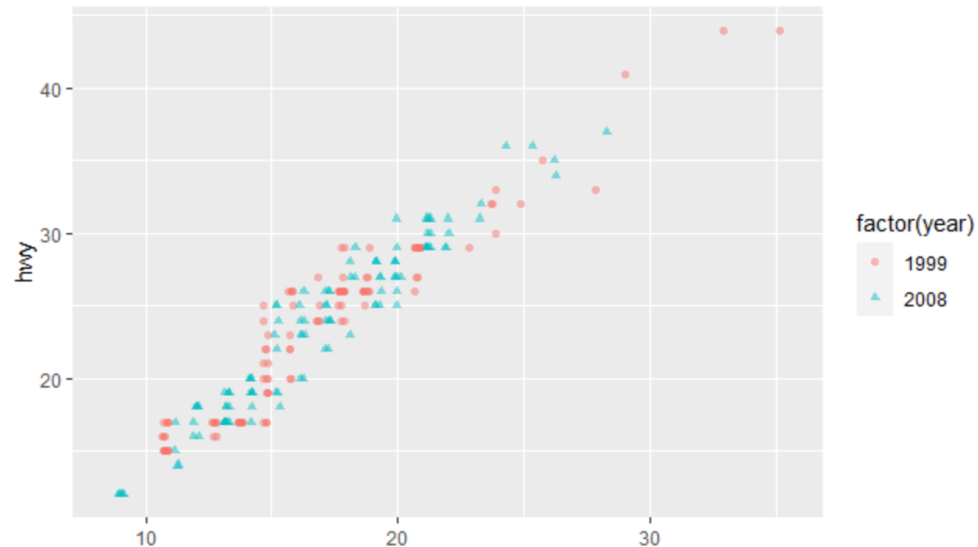


ggplot2



抖动+躲避

```
ggplot(data = mpg, mapping = aes(x = cty, y = hwy, colour = factor(year))) +  
  geom_point(aes(shape = factor(year)), size = 1.5, alpha = 0.5,  
    position = position_jitterdodge(jitter.width = 0.5,  
      dodge.width = 0.9))
```





图形保存

输出到文件

pdf , postscript , xfig, bitmap, pictex, cairo_pdf, svg, png, jpeg, bmp, tiff

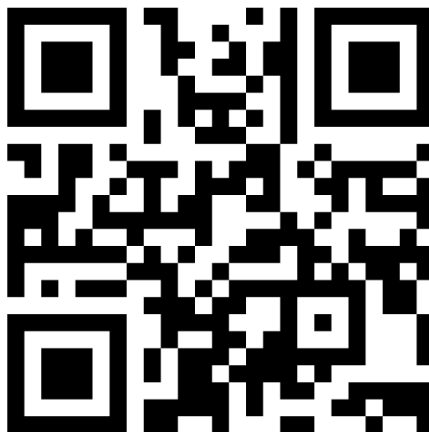
```
> pdf("plot.pdf",width=4,height=4)
```

```
> png("plot.png",width=400,height=600)
```

```
> dev.off() #绘制完图形后关闭图形设备
```

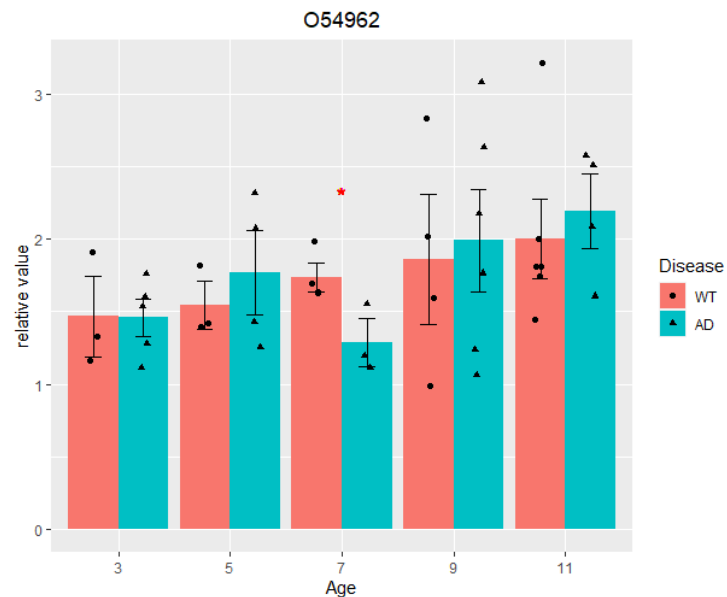
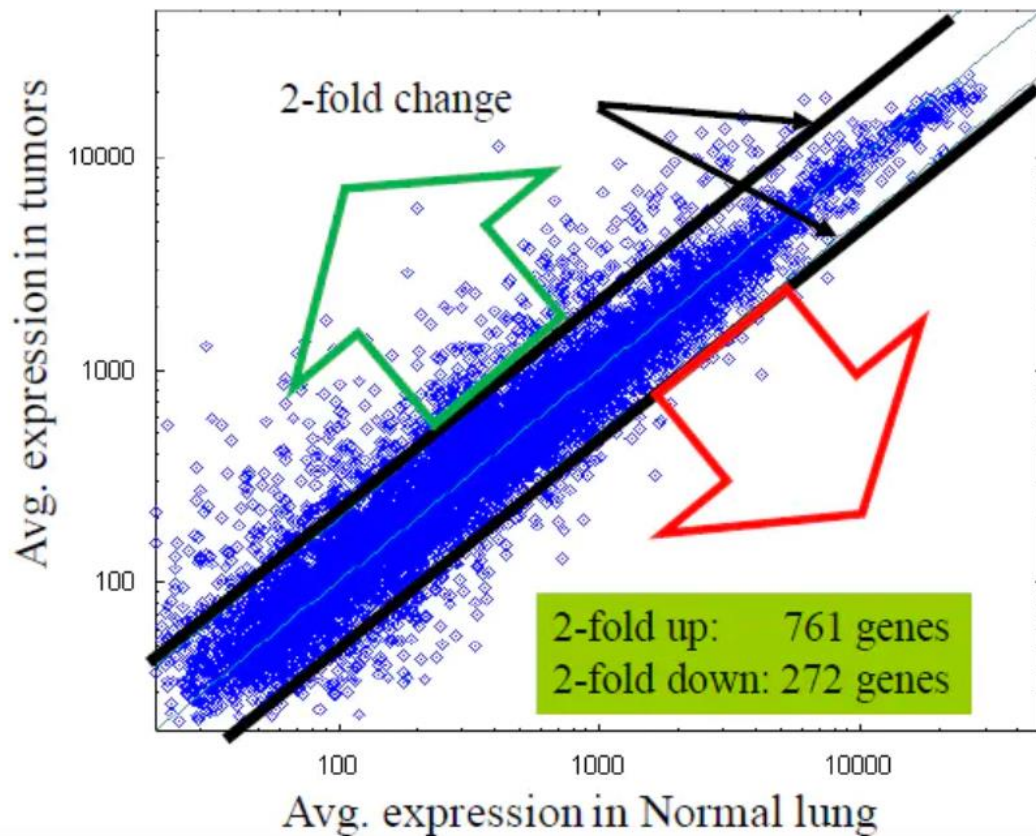
通过菜单命令保存图形





统计分析

倍数变化



没有反应数据变异特征，没有统计学上的有效分析



差异分析

显著性检验

T检验

方差分析 ANOVA

线性回归

Limma

(线性回归模型的Empirical Bayes参数估计)

基于负二项分布 edgeR, DESeq2等

P值: 也就是概率

零假设: 基因/蛋白在不同处理下的表达量相同。





t.test 的R格式

```
t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"),  
mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95,...)
```

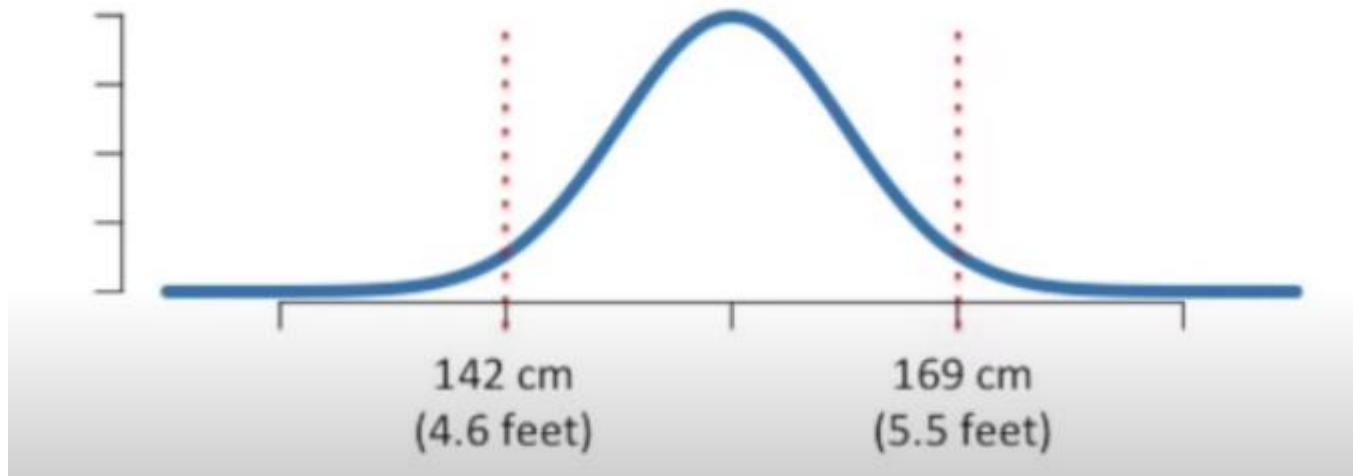
- 若仅出现数据x, 则进行单样本t检验; 若出现数据x和y, 则进行二样本的t检验
- alternative=c("two.sided", "less", "greater")用于指定所求置信区间的类型; alternative="two.sided"是缺省值, 表示求置信区间
alternative="less"表示求置信上限; alternative="greater"表示求置信下限.
- mu表示均值, 它仅在假设检验中起作用, 默认值为零。





单侧/双侧

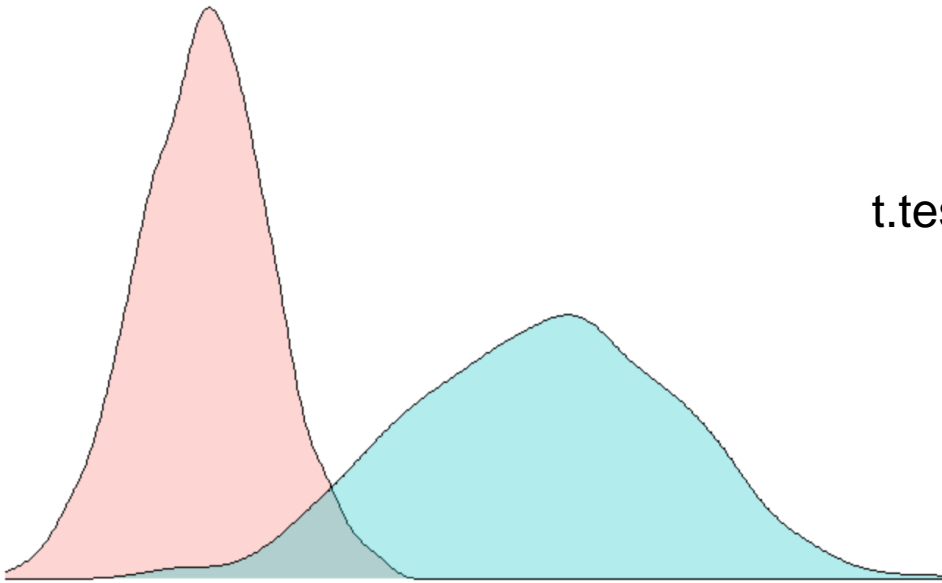
To calculate p-values, you add up the percentages of areas under the curve.





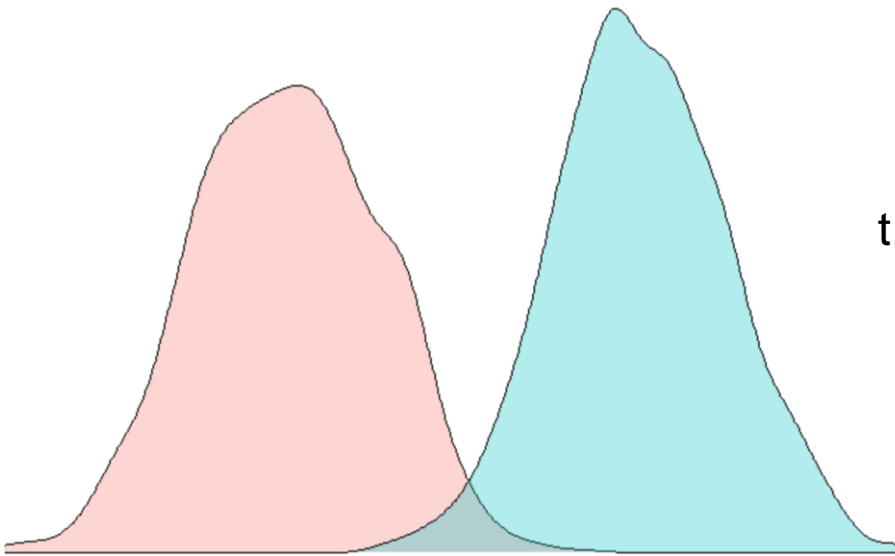
Welch t test

`t.test(x, y = NULL, var.equal = FALSE)`



Student's t-test

`t.test(x, y = NULL, var.equal = TRUE)`





在diabets.Rdata数据中，采用Welch T检验方法，看GLU数据中在Male和Female中是否有差异

t.test(x, y)

```
setwd("E:\\Rclass")
load("diabets.RData")
#
Male <- diabets$GLU[diabets$Sex == "Male"]
Female <- diabets$GLU[diabets$Sex == "Female"]
t.test(Male, Female)

total <- c(Male, Female);
group = c(rep(1:2, c(length(Male), length(Female))));
t.test(total~group)

#
dat <- data.frame(total, group)
t.test(total~group, dat)
"
```

```
t.test(GLU~Sex, diabets, subset=Sex %in% c("Male", "Female"))
```

关于T检验的Graphpad prism 和 excel 实现





t test另一种写法

Welch t test

default `t.test(x, y = NULL, var.equal = FALSE)`

`pairwise.t.test(x, y, pool.sd=FALSE, var.equal = FALSE)`

Student's t-test

`t.test(x, y = NULL, var.equal = TRUE)`

`pairwise.t.test(pool.sd=FALSE, var.equal = TRUE)`

pooled t-test

The `pool.sd` switch calculates a common SD for all groups and uses that for all comparisons (this can be useful if some groups are small). This method does not actually call `t.test`

```
> pairwise.t.test(diabets$GLU, diabets$Sex, p.adjust.method = "none")
```

```
Pairwise comparisons using t tests with pooled SD
```

```
data: diabets$GLU and diabets$Sex
```

```
      Male  Female  
Female 0.075  -  
unknown 0.112 0.787
```

```
P value adjustment method: none
```





正态分布检验

```
shapiro.test(Male)
```

方差齐性检验

```
var.test(total~group)
```





ANOVA 方差分析

方差齐性 `x <- aov(GLU~Sex, diabets)`
`summary(x)`

`oneway.test(GLU~Sex, diabets, var.equal = TRUE)`

单因素2样品比较 = Student t test

```
> summary(aov(total~group))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	1	13.20	13.204	3.603	0.0771
Residuals	15	54.96	3.664		

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
> t.test(total~group, var.equal = TRUE)
```

Two sample t-test

```
data: total by group
t = 1.8983, df = 15, p-value = 0.07707
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2552151  4.4104734
sample estimates:
mean in group 1 mean in group 2
  7.669905      5.592276
```





ANOVA 方差分析

方差不齐

```
oneway.test(GLU~Sex, diabets)
```

```
oneway.test(GLU~Sex, diabets, var.equal = FALSE)
```

单因素2样品比较 = Welch t test

```
> oneway.test(total~group)
```

```
One-way analysis of means (not assuming equal variances)
```

```
data: total and group
```

```
F = 3.6656, num df = 1.0000, denom df = 5.0771, p-value =
```

```
0 > t.test(total~group)
```

```
welch Two Sample t-test
```

```
data: total by group
```

```
t = 1.9146, df = 5.0771, p-value = 0.1128
```

```
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
```

```
-0.6991812  4.8544395
```

```
sample estimates:
```

```
mean in group 1 mean in group 2
```

```
7.669905
```

```
5.592276
```





方差齐性检验

Bartlett.test

leveneTest

多因素方差分析

```
diabets2 <- data.frame(diabets, age = c(rep(1:2, c(7, 22))))
```

	Sex	GLU	GHb	age
1	Male	8.800000	5.500000	1
2	Male	6.500000	3.400000	1
3	Female	5.400000	5.500000	1
4	Female	5.600000	4.300000	1
5	unknown	6.700000	3.500000	1
6	unknown	7.800000	3.400000	1
7	unknown	4.500000	5.500000	1
8	Male	9.700000	7.000000	2
9	Female	5.000000	4.300000	2





#有相互作用的双因素

```
x <- aov(GLU~Sex+age+Sex:age, diabets2)
```

```
summary(x)
```

```
x <- aov(GLU~Sex*age, diabets2)
```

```
summary(x)
```

#只有相互作用

```
x <- aov(GLU~Sex:age, diabets2)
```

```
summary(x)
```

#无相互作用

```
x <- aov(GLU~Sex+age, diabets2)
```

```
summary(x)
```





DESeq2

Differential expression analysis based on the Negative Binomial (a.k.a. Gamma-Poisson) distribution


Description

This function performs a default analysis through the steps:

1. estimation of size factors: [estimateSizeFactors](#)
2. estimation of dispersion: [estimateDispersions](#)
3. Negative Binomial GLM fitting and Wald statistics: [nbinomWaldTest](#)

1. Estimation of size factors

```
locfunc=stats::median;

type <- "ratio"
sanitizeColData <- function(object) {}
object2 <- sanitizeColData(object)
sizeFactors(object2) <- estimateSizeFactorsForMatrix(counts(object2),
                                                    locfunc=locfunc)

if(1 != 1)
{
  counts <- counts(object)
  incomingGeoMeans <- FALSE
  loggeomeans <- rowMeans(log(counts))
  sf <- apply(counts, 2, function(cnts) {
    exp(locfunc((log(cnts) - loggeomeans)[is.finite(loggeomeans) & cnts > 0]))
  })
  sf
}
```





相关性分析

(1) 相关系数

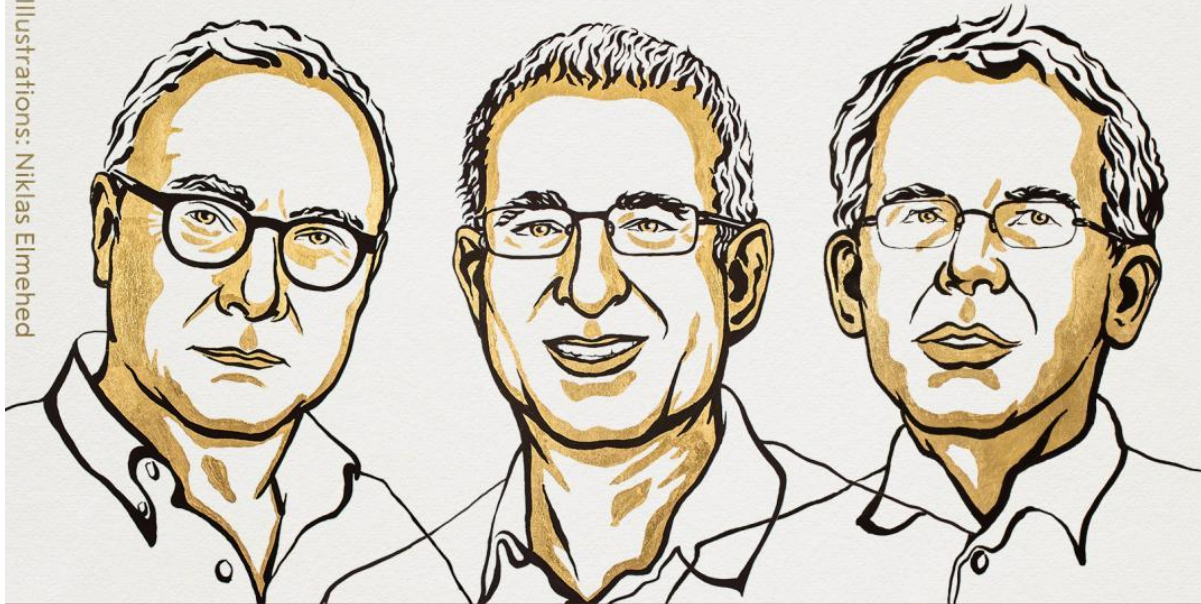
$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$





THE SVERIGES RIKSBANK PRIZE
IN ECONOMIC SCIENCES IN MEMORY
OF ALFRED NOBEL 2021

Illustrations: Niklas Elmehed



David
Card

"for his empirical
contributions to labour
economics"

Joshua
D. Angrist

"for their methodological
contributions to the analysis
of causal relationships"

Guido
W. Imbens

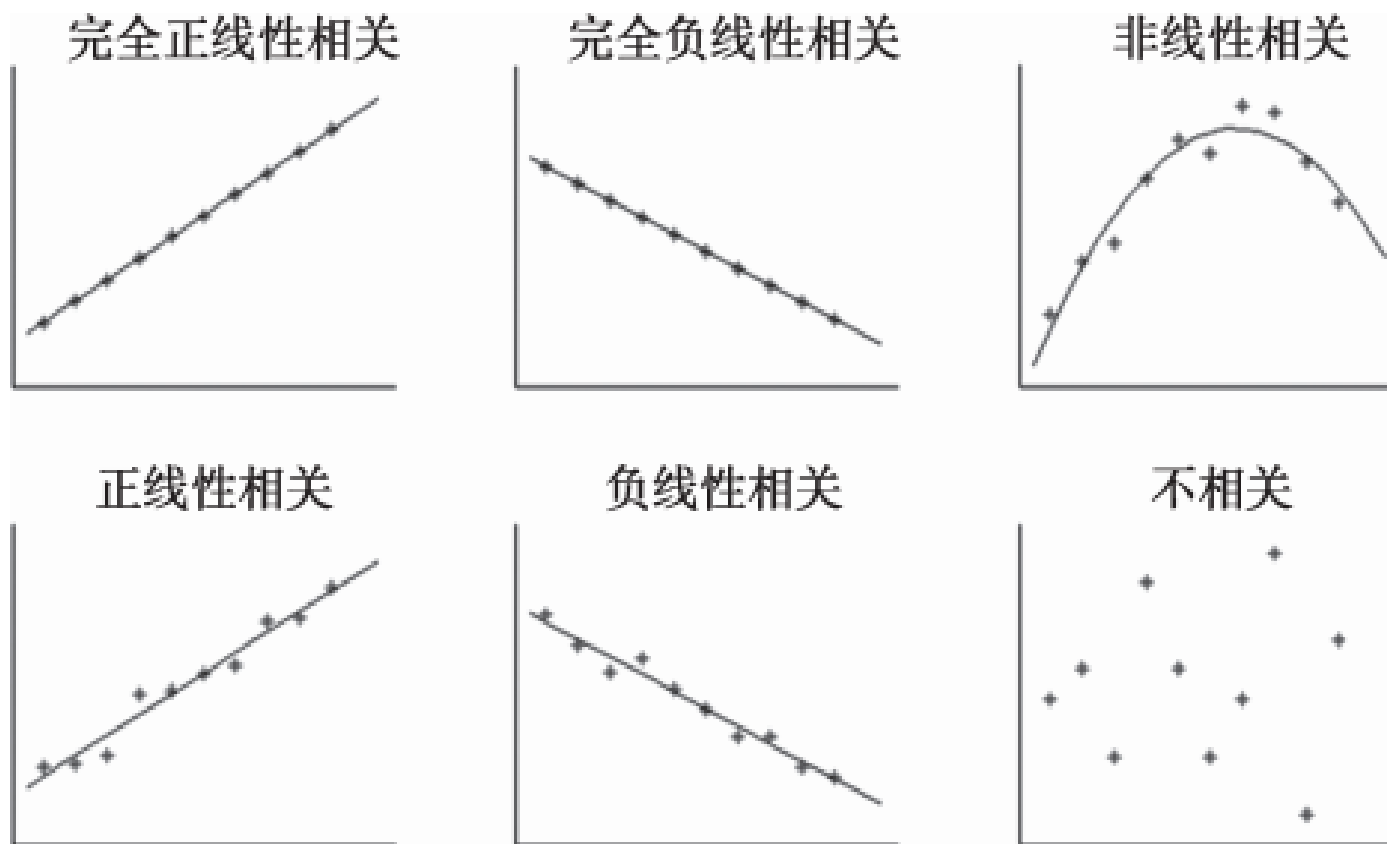
THE ROYAL SWEDISH ACADEMY OF SCIENCES



因果必相关，相关未必因果。



(2) 相关分类





Pearson 相关 Spearman 相关 Kendall 相关

Test for Association/Correlation Between Paired Samples

Description

Test for association between paired samples, using one of Pearson's product moment correlation coefficient, Kendall's *tau* or Spearman's *rho*.

Usage

```
cor.test(x, ...)  
  
## Default S3 method:  
cor.test(x, y,  
         alternative = c("two.sided", "less", "greater"),  
         method = c("pearson", "kendall", "spearman"),  
         exact = NULL, conf.level = 0.95, continuity = FALSE, ...)  
  
## S3 method for class 'formula'  
cor.test(formula, data, subset, na.action, ...)
```





富集分析

#超几何分布

`dhyper(q,m,n,k)`

#在m个白球，n个黑球中抽k个球，其中出现q个白球的概率 概率总和是1

`x = 0; for(i in 0:5)`

`x = x+dhyper(i,10,20,5);`

```
x0 <- dhyper(0,10,20,5); x1 <- dhyper(1,10,20,5);  
x2 <- dhyper(2,10,20,5); x3 <- dhyper(3,10,20,5);  
x4 <- dhyper(4,10,20,5); x5 <- dhyper(5,10,20,5);  
.
```

Values

x0	0.108795419140247
x1	0.339985684813271
x2	0.359984842743463
x3	0.159993263441539
x4	0.0294724432655467
x5	0.0017683465959328

`phyper(q,m,n,k)`

#`low.tail=TRUE`,在m个白球，n个黑球中抽k个球，其中出现q个白球或更少的概率

#`low.tail=FALSE`,在m个白球，n个黑球中抽k个球，其中出现大于q个白球的概率

`phyper(3,10,20,5,lower.tail = FALSE)`#5个球抽到4个或5个白的概率





#Fisher extract 检验

```
fisher.test(data=data.frame(c(q,k-q),c(m-q,n-k+q)))  
data<-data.frame(c(4,1),c(6,19))
```

	抽取	剩下	合计
白球	4	6	10
黑球	1	19	20
合计	5	25	30

fisher.test(data) #5个球抽到4个以及比其概率更小概率的总和

phyper(3,10,20,5,lower.tail = FALSE)#5个球抽到4个或5个白的概率





实验设计

前瞻性

回顾性

临床试验

生物实验类研究

Meta分析

真实世界研究

基于实验的生物信息学分析

基于数据库的生物信息学分析

尽可能的消除其他因素的干扰

单一变量原则

Covariates校正

