

生物统计学： 生物信息中的概率统计模型

2022年秋



有关信息

- 授课教师: 宁康
 - Email: ningkang@hust.edu.cn
 - Office: 华中科技大学东十一楼504室
 - Phone: 87793041, 18627968927
- 课程网页
 - <http://www.microbioinformatics.org/Biostatistics.html>
 - QQ群: 648612956



2022生物统计学

群号: 648612956



扫一扫二维码, 加入群聊。



课程安排

- 生物背景和课程简介
- 传统生物统计学及其应用
- 生物统计学和生物大数据挖掘
 - Hidden Markov Model (HMM)及其应用
 - Markov Chain
 - HMM理论
 - HMM和基因识别 (Topic I)
 - HMM和序列比对 (Topic II)
 - 进化树的概率模型 (Topic III)
 - Motif finding中的概率模型 (Topic IV)
 - EM algorithm
 - Markov Chain Monte Carlo (MCMC)
 - 基因表达数据分析 (Topic V)
 - 聚类分析-Mixture model
 - Classification-Lasso Based variable selection
 - 基因网络推断 (Topic VI)
 - Bayesian网络
 - Gaussian Graphical Model
 - 基因网络分析 (Topic VII)
 - Network clustering
 - Network Motif
 - Markov random field (MRF)
 - Dimension reduction及其应用 (Topic VIII)
- 面向生物大数据挖掘的深度学习

研究对象：
生物序列，
进化树，
生物网络，
基因表达
...

方法：
生物计算与生物统计

第2章：传统生物统计学及其应用

- 生物统计学基本名词
- 生物统计学理论基础
- 生物统计学基本应用

Part I

生物统计学基本名词

Logistic Regression

Uniform Distribution

Percentage

Average

BIOSTATISTICS

p-values

Population Genetics

Probability

Statistics

Correlation

Epidemiology

定义

统计学（Statistics）是把数学的语言引入具体的科学领域，把具体科学领域中要待研究的问题抽象为数学问题的过程，它是收集、分析、列示和解释数据的一门艺术和科学。

统计是以数据为食物的动物

统计的本业是消化数据，并产生有营养的结果。它的本质，和母牛差不多。

Grass——Cow——Milk

Data—— Statistics ——Information

统计学：无处不在！

习近平说：中国是世界第二大经济体，有13亿多人口的大市场，有960多万平方米的国土，中国经济是一片大海，而不是一个小池塘。大海有风平浪静之时，也有风狂雨骤之时。没有风狂雨骤，那就不是大海了。狂风骤雨可以掀翻小池塘，但不能掀翻大海。经历了无数次狂风骤雨，大海依旧在那儿！经历了5000多年的艰难困苦，中国依旧在这儿！面向未来，中国将永远在这儿！

概率：

$$P(\text{掀翻} \cap \text{狂风骤雨} | \text{小池塘}) = \text{high}$$

$$P(\text{掀翻} \cap \text{小雨} | \text{小池塘}) = \text{low}$$

$$P(\text{掀翻} \cap \text{狂风骤雨} | \text{大海}) = \text{low}$$

$$P(\text{掀翻} \cap \text{小雨} | \text{大海}) = \text{low}$$

$$P(\text{掀翻} | \text{大海}) = P(\text{掀翻} \cap \text{狂风骤雨} | \text{大海}) + P(\text{掀翻} \cap \text{小雨} | \text{大海})$$

贝叶斯推断：

$$P(\text{掀翻} | \text{大海}) = P(\text{大海} | \text{掀翻}) * P(\text{掀翻}) / P(\text{大海}) = \text{low}$$

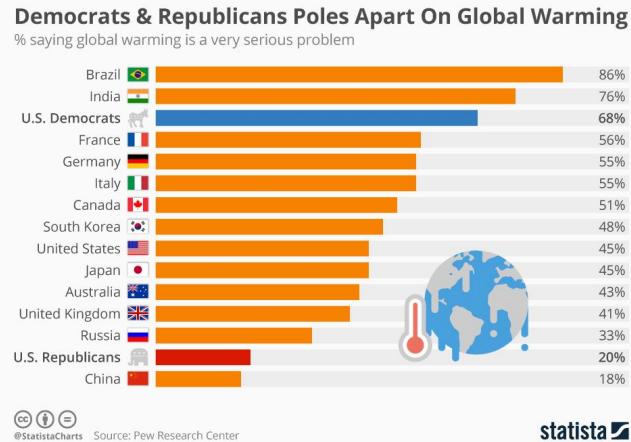
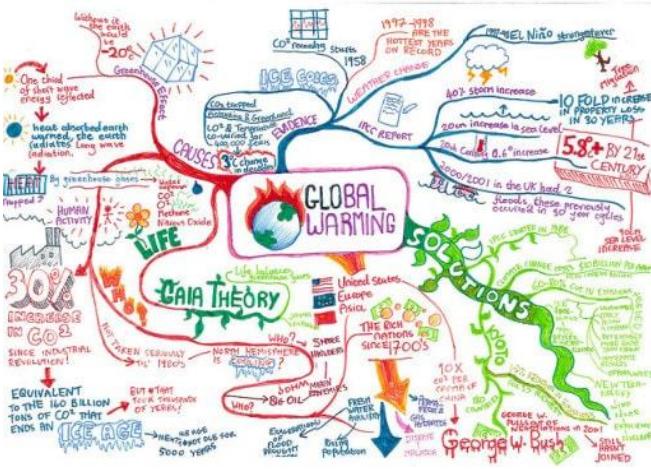
vs.

$$P(\text{掀翻} | \text{小池塘}) = P(\text{小池塘} | \text{掀翻}) * P(\text{掀翻}) / P(\text{小池塘}) = \text{high}$$



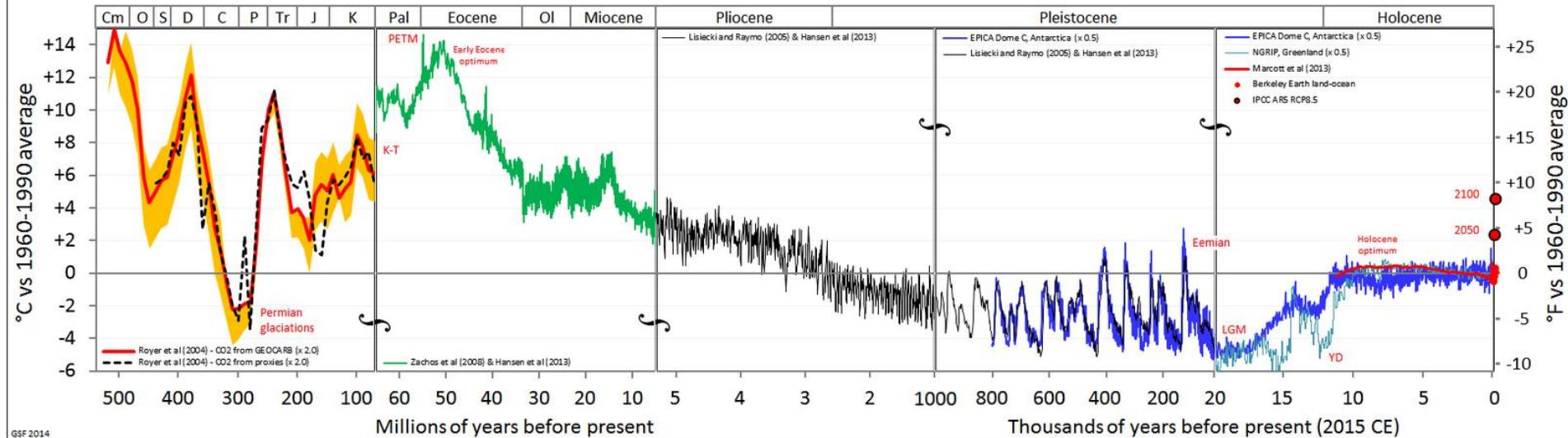
统计学：无处不在！

从以往气温情况，预测未来气温变化趋势？



statista

Temperature of Planet Earth



统计学：无处不在！

大学排名？
眼花缭乱。。。
有规则可循！

指标类别	指标名称	指标内涵	权重
人才培养 (45%)	生源质量 (新生高考成绩)	录取新生的高考成绩	30%
	培养结果 (毕业生就业率)	本科毕业生的就业率	10%
	社会声誉 (社会捐赠收入)	学校基金会年度社会捐赠收入	5%
科学研究 (40%)	科研规模 (论文数量)	Scopus数据库收录的论文数	10%
	科研质量 (论文质量)	学科标准化后的论文影响力	10%
	顶尖成果 (高被引论文)	被引用次数位居各个学科世界前1%的论文数	10%
服务社会 (10%)	顶尖人才 (高被引学者)	各个学科被引用次数最高的中国学者数	10%
	科技服务 (企业科研经费)	企事业单位委托的科技经费数	5%
国际化 (5%)	成果转化 (技术转让收入)	大学技术转让当年实际收入	5%
	学生国际化 (留学生比例)	学历留学生在校生总数的比例	5%

2019软科中国最好大学排名					人才培养	科学研究院
排名	学校名称	省市	总分		新生高考成绩得分	91.1
1	清华大学	北京	94.6		毕业生就业率	97.92%
					社会捐赠收入(千元)	275,235
					服务社会	高被引论文(篇)
					企业科研经费(千元)	875,715
					技术转让收入(千元)	12,574
					人才培养	国际化学术影响
					新生高考成绩得分	6.05%
5	复旦大学	上海	65.6		毕业生就业率	论文数量(篇)
					社会捐赠收入(千元)	论文质量(FWCI)
					服务社会	高被引论文(篇)
					企业科研经费(千元)	高被引学者(人)
					技术转让收入(千元)	留学生比例
2	北京大学	北京	76.5		人才培养	科学研究院
					新生高考成绩得分	91.6
					毕业生就业率	96.09%
					社会捐赠收入(千元)	251,272
					服务社会	高被引论文(篇)
					企业科研经费(千元)	828
					技术转让收入(千元)	高被引学者(人)
					人才培养	国际化学术影响
6	中国科学技术大学	安徽	60.9		新生高考成绩得分	6.77%
					毕业生就业率	论文数量(篇)
					社会捐赠收入(千元)	35028
					服务社会	论文质量(FWCI)
					企业科研经费(千元)	1.384
					技术转让收入(千元)	81
3	浙江大学	浙江	72.9		人才培养	科学研究院
					新生高考成绩得分	91.1
					毕业生就业率	93.40%
					社会捐赠收入(千元)	71,038
					服务社会	高被引论文(篇)
					企业科研经费(千元)	992
					技术转让收入(千元)	高被引学者(人)
					人才培养	国际化学术影响
7	华中科技大学	湖北	58.9		新生高考成绩得分	2.32%
					毕业生就业率	论文数量(篇)
					社会捐赠收入(千元)	25036
					服务社会	论文质量(FWCI)
					企业科研经费(千元)	1.56
					技术转让收入(千元)	43
7	南京大学	江苏	58.9		人才培养	科学研究院
					新生高考成绩得分	80.1
					毕业生就业率	95.70%
					社会捐赠收入(千元)	33,858
					服务社会	高被引论文(篇)
					企业科研经费(千元)	924
					技术转让收入(千元)	33
					人才培养	国际化学术影响
9	中山大学	广东	58.2		新生高考成绩得分	5.02%
					毕业生就业率	论文数量(篇)
					社会捐赠收入(千元)	32298
					服务社会	论文质量(FWCI)
					企业科研经费(千元)	1.327
					技术转让收入(千元)	80.1
					人才培养	国际化学术影响
10	哈尔滨工业大学	黑龙江	56.7		新生高考成绩得分	3.25%
					毕业生就业率	论文数量(篇)
					社会捐赠收入(千元)	28082
					服务社会	论文质量(FWCI)
					企业科研经费(千元)	1.474
					技术转让收入(千元)	797
					人才培养	国际化学术影响
7	华中科技大学	湖北	58.9		新生高考成绩得分	3.49%
					毕业生就业率	论文数量(篇)
					社会捐赠收入(千元)	114,607
					服务社会	论文质量(FWCI)
					企业科研经费(千元)	33
					技术转让收入(千元)	33,349
					人才培养	国际化学术影响

统计学：无处不在！

大学排名？
眼花缭乱。。。
有规则可循！

US.News国内部分大学排名变化								
排名	高校名称	2016	2017	2018	2019	2020	平均	变化
1	清华大学	59	57	64	50	36	53.2	+23
2	北京大学	41	53	65	68	59	57.2	-18
3	中国科学技术大学	131	136	145	138	128	135.6	+3
4	上海交通大学	136	138	156	145	136	142.2	0
5	浙江大学	106	138	159	165	157	145	-51
6	南京大学	180	187	190	179	168	180.8	+12
7	复旦大学	96	121	148	159	171	139	-75
8	中山大学	198	225	237	224	208	218.4	-10
9	华中科技大学	265	295	282	260	245	269.4	+20
10	哈尔滨工业大学	319	303	304	280	249	291	-70
11	同济大学	335	337	327	302	279	316	-56
12	武汉大学	251	324	321	299	285	296	-34
13	北京师范大学	296	286	324	330	332	313.6	-36
14	厦门大学	275	306	344	336	332	318.6	-57
15	东南大学	359	364	382	341	311	351.4	+48

★ CNUR中国大学排行榜 China University Rankings ★ CNR中国大学综合评价

排名	学校名称	省市	类型	总分
1	清华大学	北京	综合	852.5
2	北京大学	北京	综合	746.7
3	浙江大学	浙江	综合	649.2
4	上海交通大学	上海	综合	625.9
5	南京大学	江苏	综合	561.1
6	复旦大学	上海	综合	556.7
7	中国科学技术大学	安徽	理工	526.4
8	华中科技大学	湖北	综合	497.7
9	武汉大学	湖北	综合	488.0

模块 得分 模块 得分 模块 得分
 办学层次 38.2 人才培养 256.8 重大项目与成果 131.0
 学科水平 72.4 科学研究 69.1 国际竞争力 79.9
 办学资源 39.6 服务社会 40.6
 师资规模与结构 48.4 高端人才 76.5

模块 得分 模块 得分 模块 得分
 办学层次 36.1 人才培养 237.6 重大项目与成果 105.8
 学科水平 73.1 科学研究 71.0 国际竞争力 61.2
 办学资源 24.6 服务社会 16.2
 师资规模与结构 49.2 高端人才 71.9

模块 得分 模块 得分 模块 得分
 办学层次 33.9 人才培养 215.3 重大项目与成果 81.7
 学科水平 65.3 科学研究 68.6 国际竞争力 43.0
 办学资源 20.1 服务社会 23.9
 师资规模与结构 48.3 高端人才 49.1

模块 得分 模块 得分 模块 得分
 办学层次 35.4 人才培养 192.8 重大项目与成果 93.0
 学科水平 53.6 科学研究 81.2 国际竞争力 40.1
 办学资源 22.1 服务社会 18.1
 师资规模与结构 43.8 高端人才 45.8

模块 得分 模块 得分 模块 得分
 办学层次 35.1 人才培养 218.6 重大项目与成果 71.2
 学科水平 47.8 科学研究 59.6 国际竞争力 29.0
 办学资源 10.3 服务社会 5.3
 师资规模与结构 47.4 高端人才 42.0

模块 得分 模块 得分 模块 得分
 办学层次 36.6 人才培养 198.5 重大项目与成果 62.0
 学科水平 48.4 科学研究 65.7 国际竞争力 34.8
 办学资源 14.9 服务社会 6.5
 师资规模与结构 46.3 高端人才 42.9

模块 得分 模块 得分 模块 得分
 办学层次 40.0 人才培养 191.5 重大项目与成果 49.2
 学科水平 39.1 科学研究 52.6 国际竞争力 42.2
 办学资源 10.6 服务社会 0.2
 师资规模与结构 45.9 高端人才 55.1

模块 得分 模块 得分 模块 得分
 办学层次 31.9 人才培养 182.8 重大项目与成果 44.9
 学科水平 45.2 科学研究 58.3 国际竞争力 31.8
 办学资源 11.3 服务社会 22.0
 师资规模与结构 44.2 高端人才 25.5

模块 得分 模块 得分 模块 得分
 办学层次 31.7 人才培养 198.8 重大项目与成果 44.2
 学科水平 48.4 科学研究 51.3 国际竞争力 25.2
 办学资源 9.9 服务社会 11.8
 师资规模与结构 45.3 高端人才 21.4

中国最好学科排名 2020

指标体系

人才培养
立德树人典型
模范先进教师
模范先进学生
精品课程教材
国家级精品资源共享课
国家精品在线开放课程
国家级精品视频公开课
中宣部编马工程教材
教育部组编马工程教材
教学成果奖励
国家级教学成果奖
研究生教育成果奖
造就学术人才
科学院院士 工程院院士 长江特聘 国家杰青 科技领军 哲社领军 工程领军

科研项目
重大重点项目
重点研发计划
自然科学中心
重大仪器研制
自科重大计划
自科重大项目
自科重点项目
社科重大项目
哲社攻关项目
社科重点项目
面上青年项目
自科面上项目
自科青年项目
社科一般项目
社科青年项目

成果获奖
国家科技奖励
国家自然科学奖
国家技术发明奖
国家科技进步奖
(仅用于理、工、农、医门类学科权重加倍，文科学科权重减半)
教育部奖励
教育部人文社科奖
教育部科学技术奖
(文科学科、艺术门类学科、交叉学科)

学术论文
国际重要期刊论文
(理、工、农、医门类学科权重加倍，文科学科权重减半)
中文期刊论文
(仅用于文科学科、艺术门类学科、交叉学科)
国际顶尖期刊论文
(仅用于理学、医学门类学科和交叉学科)
中文顶尖期刊论文
(仅用于文科学科和交叉学科)

高端人才
资深学术权威
科学院院士(80岁以下)
工程院院士(80岁以下)
(仅用于理、工、农、医门类学科)
中年领军专家
长江特聘 国家杰青 科技领军 哲社领军 工程领军
青年拔尖英才
国家优青 海外青年 青年拔尖 青年长江
国际知名学者
科睿唯安高被引 爱思唯尔高被引 (文科学科、艺术门类学科权重减半)

统计学的基本特点

- 概率性
 - 以概率论为理论基础
 - 其结果均伴随着某种概率
- 二元性
 - 理论与实际数据
- 归纳性
 - 由现实数据资料中归纳出一般的原理，由特殊推导一般（由样本推导总体）

二、发展概况

生物统计学

古典记录统计学

17世纪中叶—
19世纪中叶：
大数定律，正态分布等。

19世纪中叶—
20世纪上半叶：
生物统计学的开始，卡方检验等。

近代描述统计学

20世纪初至今：
F检验、差分析，
方差分析等，各统计学软件。

现代推断统计学

方兴未艾：
艾尔夫马科链，
信息，机器学习中的统计推断。

迅速发展
形成分支

二、发展概况



it's a long long story

人口普查

F检验

大数定理

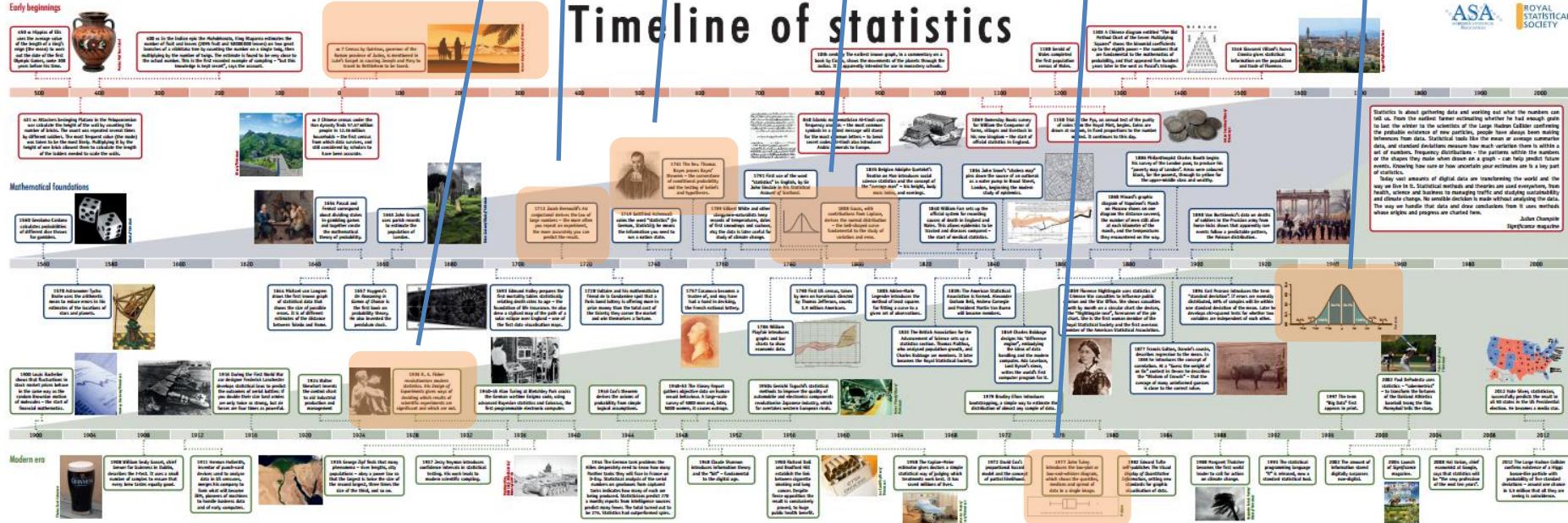
贝叶斯推断

方差

正态分布

箱式图

Timeline of statistics



统计学发展史中的重大事件与重要代表人物



J.Bernoulli (伯努利, 瑞士, 1654~1705)

系统论证了“大数定律”，即样本容量越大，样本统计数与总体参数之差越小。



P.S. Laplace (拉普拉斯, 法国, 1749~1827)

最早系统的把概率论方法运用到统计学研究中去，建立了严密的概率数学理论，并应用到人口统计、天文学等方面的研究上。

伯努利大数定律

设 m 是 n 次独立试验中事件A出现的次数，而 p 是事件A在每次试验中出现的概率，则对于任意小的正数 ε ，有如下关系：

$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{m}{n} - p \right| < \varepsilon \right\} = 1$$

若试验条件不变，重复次数 n 接近无限大时，频率与理论概率的差值必定要小于一个任意小的正数 ε ，即这两者可以基本相等，这几乎是一个必然要发生的事情。

统计学：概率不等于事实！

盖洛普民意测验与美国总统大选关联度一览表（1936—2000）

年代	候选人	盖洛普最后 民意测验结果 (%)	总统选举真 实结果 (%)	盖洛普 误差 (%)
2000	布什	48.0	47.9	+0.1
1996	克林顿	52.0	49.2	+2.8
1992	克林顿	49.0	43.3	+5.7
1988	老布什	56.0	53.9	+2.1
1984	里根	59.0	59.2	-0.2
1980	里根	47.0	50.8	-3.8
1976	卡特	48.0	50.1	-2.1
1972	尼克松	62.0	61.8	+0.2
1968	尼克松	43.0	43.5	-0.5
1964	约翰逊	64.0	61.3	+2.7
1960	肯尼迪	51.0	50.1	+0.9
1956	艾森豪威尔	59.5	57.8	+1.7
1952	艾森豪威尔	51.0	55.4	-4.4
1948	杜鲁门	44.5	49.5	-5.0
1944	罗斯福	51.5	53.8	-2.3
1940	罗斯福	52.0	55.0	-3.0
1936	罗斯福	55.7	62.5	-6.8



盖洛普民意测验创始人
乔治·盖洛普



K. Pearson (卡.皮尔逊, 英国, 1857~1936)

首创频数分布表与频数分布图；观察到许多生物的度量并不呈现正态分布，利用相对斜率得到矩形分布、J型分布、U型分布或铃型分布等；发现了 χ^2 分布，提出了有名的卡方检验法；



W.S.Gosset (歌赛特, 英国, 1777~1855)

创立了小样本检验代替大样本检验的理论，即t分布和t检验法，也称为学生式分布。

$$x \rightarrow N(\mu, \sigma^2)$$

t分布

当 σ^2 已知

$$u = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \rightarrow N(0,1)$$

当 σ^2 未知,
且 $n > 30$

$$u = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{\bar{x} - \mu}{s / \sqrt{n}} \rightarrow N(0,1)$$

当 σ^2 未知,
且 $n < 30$

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{\bar{x} - \mu}{s / \sqrt{n}} \neq N(\mu, \sigma^2)$$

t分布是英国统计学家Gosset 1908年以笔名“student”所发表的论文提出的，因此又称为学生氏t分布。

t分布概率密度函数

$$f(t) = \frac{\Gamma(\frac{df+1}{2})}{\sqrt{\pi df} \Gamma(\frac{df}{2})} \left(1 + \frac{t^2}{df}\right)^{-\frac{df+1}{2}}$$

$$t = \frac{\bar{x} - \mu}{S_{\bar{x}}} = \frac{\bar{x} - \mu}{S / \sqrt{n}} \neq N(\mu, \sigma^2)$$



R.A.Fisher (费歇尔, 英国, 1890~1962)

发展了显著性检验及估计理论，提出了F分布和F检验，首创“方差”和“方差分析”两个概念，1925年提出随机区组和正交拉丁方试验设计，在试验设计中提出“随机化”原则，1938年和Yates合编了Fisher Yates随机数字表。

设从一正态总体 $N(\mu, \sigma^2)$ 中随机抽取样本容量为 n_1 、 n_2 的两个独立样本，其样本方差为 s_1^2 、 s_2^2 ，则定义其比值：

$$F = \frac{s_1^2}{s_2^2}$$

此 F 值具有 s_1^2 的自由度 $df_1=n_1-1$ 和 s_2^2 的自由度 $df_2=n_2-1$ 。

如果对一正态总体在特定的 df_1 和 df_2 进行一系列随机独立抽样，则所有可能的 F 值就构成一个 F 分布。

$$f(F) = \frac{\Gamma(\frac{df_1 + df_2}{2})}{\Gamma(\frac{df_1}{2})\Gamma(\frac{df_2}{2})} df_1^{\frac{df_1}{2}} df_2^{\frac{df_2}{2}} \frac{F^{\frac{df_1}{2}-1}}{(df_1 F + df_2)^{\frac{df_1+df_2}{2}}}$$

F 分布是随自由度 df_1 和 df_2 进行变化的一组曲线。

F 分布的概率累积函数

$$F(F) = \int_0^F f(F) dF$$

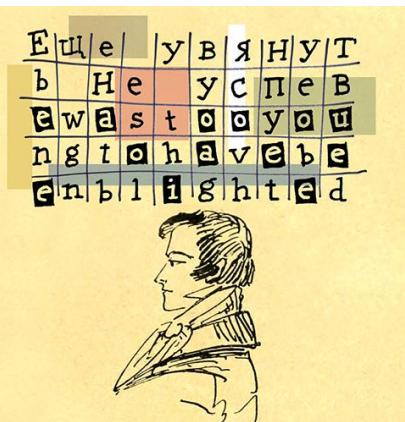
马尔科夫链 (Markov Chain)

- 马尔可夫链模型（马氏链）：每一步活动只与当前处在什么“状态”有关，与过去的“状态”没有关系。
- 把传统的基于离散状态的统计分析，转变为基于前后相关状态的统计分析。
- 能够解决生物信息、语音和图像处理等广泛的统计学相关问题。

Markov Chain

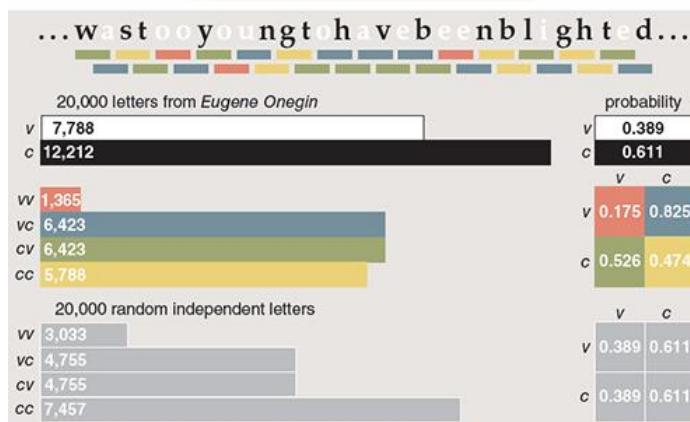
马尔科夫设计马氏链的最初应用，本质上就是机器学习。。。.

普希金诗作
《叶甫盖尼·奥涅金》



基于统计，确定转移概率参数

*He was too young to have been blighted
by the cold world's corrupt finesse;
his soul still blossomed out, and lighted
at a friend's word, a girl's caress.
In heart's affairs, a sweet beginner,
he fed on hope's deceptive dinner;
the world's éclat, its thunder-roll,
still captivated his young soul.
He sweetened up with fancy's icing
the uncertainties within his heart;
for him, the objective on life's chart
was still mysterious and enticing—
something to rack his brains about,
suspecting wonders would come out.*



马氏链作诗（不动点问题）

First order
Theg sheso pa lyiklg ut. cout Scrpauscricre cobaiives wingervet Ners, whe ilened te o
wn taulie wom uld atimorerteansourocono weveiknt hef ia ngry'sif farl t mmat and,
tr iscond frnid riliof th Gureckpeag

Third order
At oness, and no fall makestic to us, infessed Russia-nbently our then a man thouz al-
ways, and toops in he roquestill shoed to displic! Is Olga's up. Italked fore declainsel
the Juan's conven night toget nothem,

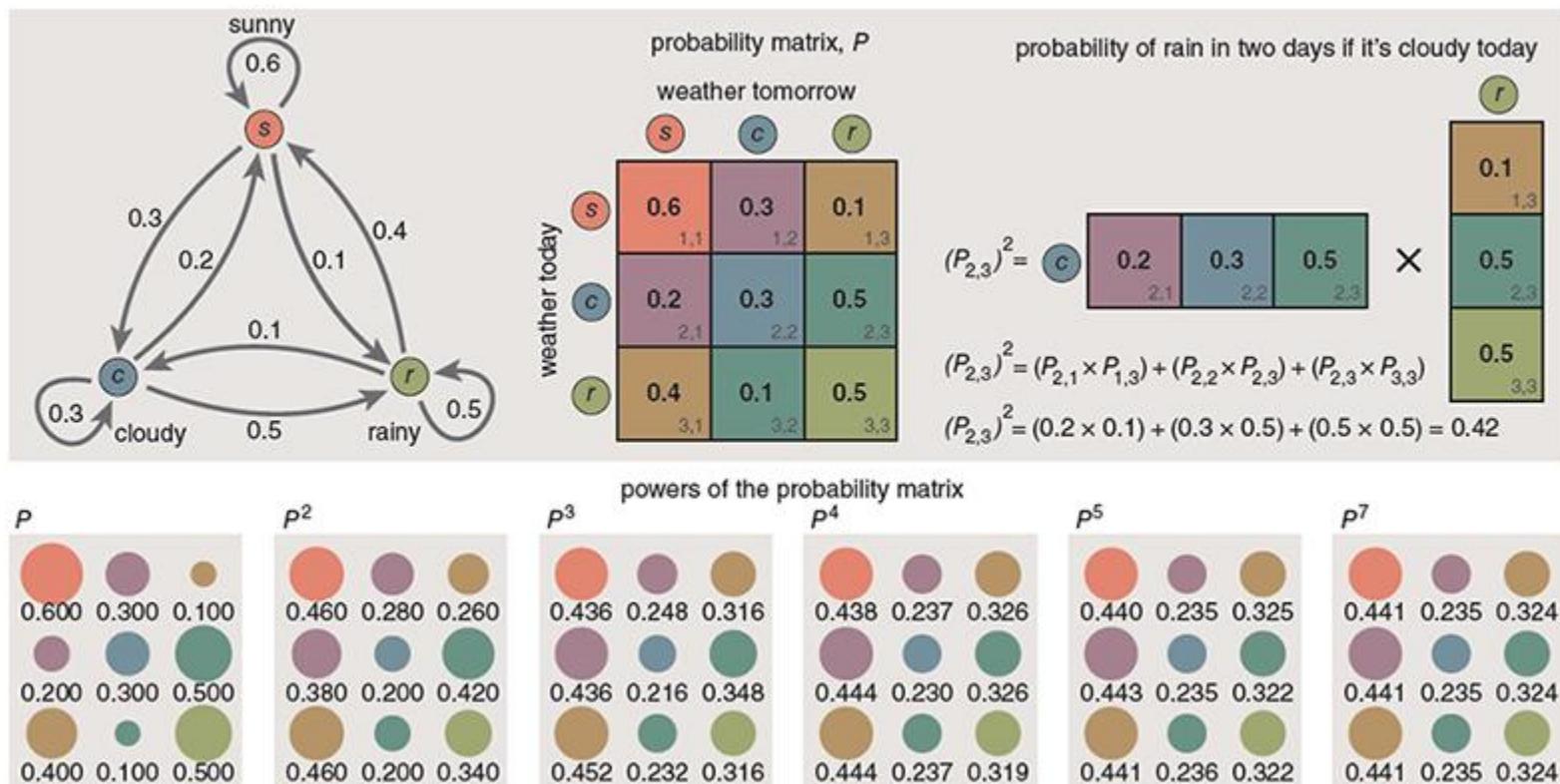
Fifth order
Meanwhile with jealousy bench, and so it was his time. But she trick. Let message
we visits at dared here bored my sweet, who sets no inclination, and Homer, so prose,
weight, my goods and envy and kin.

Seventh order
My sorrow her breast, over the dumb torment of her veil, with our poor head is stoop-
ing. But now Aurora's crimson finger, your christening glow. Farewell. Evgeny loved
one, honoured fate by calmly, not yet seeking?

Markov Chain

马氏链的广泛应用： Past, Present and Future

基于马氏链的天气预报系统



参考文献：<https://www.americanscientist.org/article/first-links-in-the-markov-chain>

Mutual information

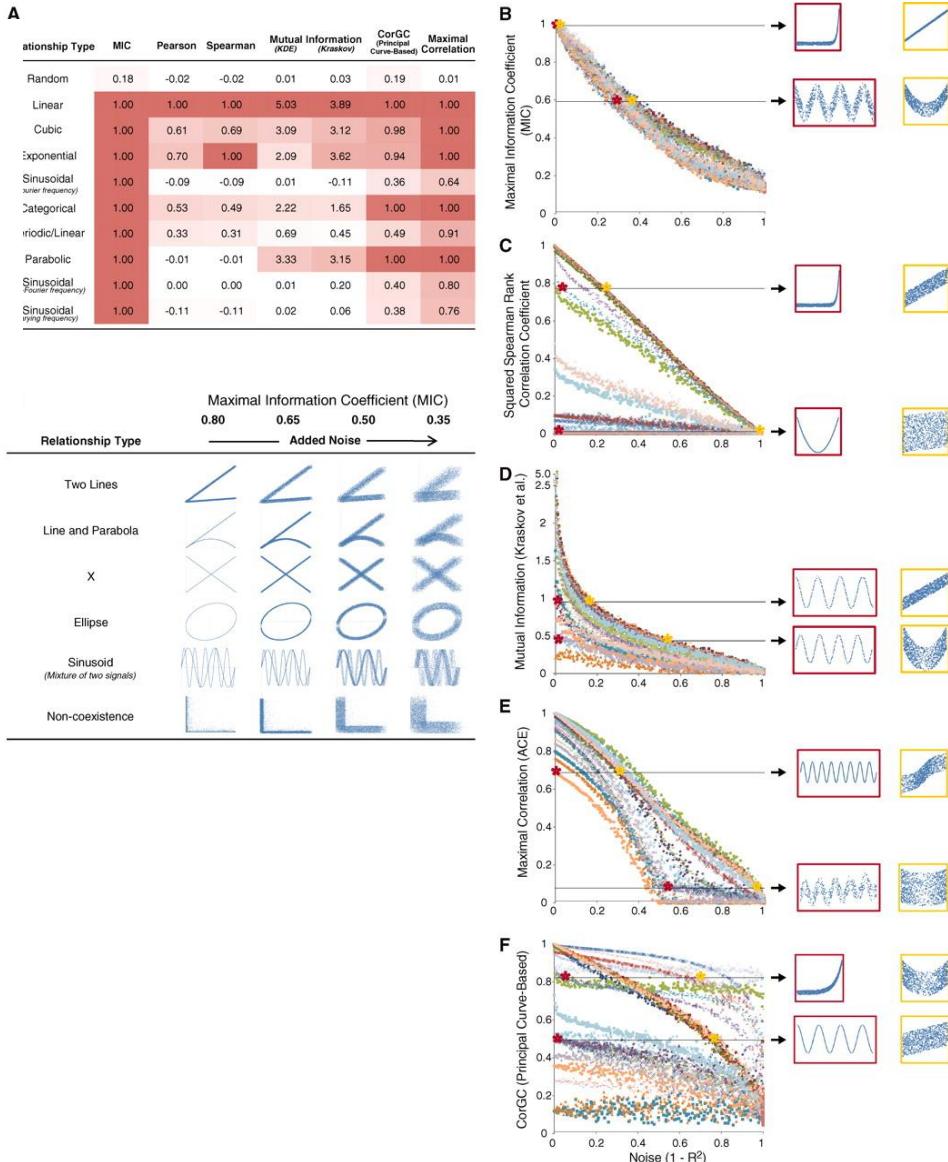
PERSPECTIVE | MATHEMATICS

A Correlation for the 21st Century

Terry Speed

* See all authors and affiliations

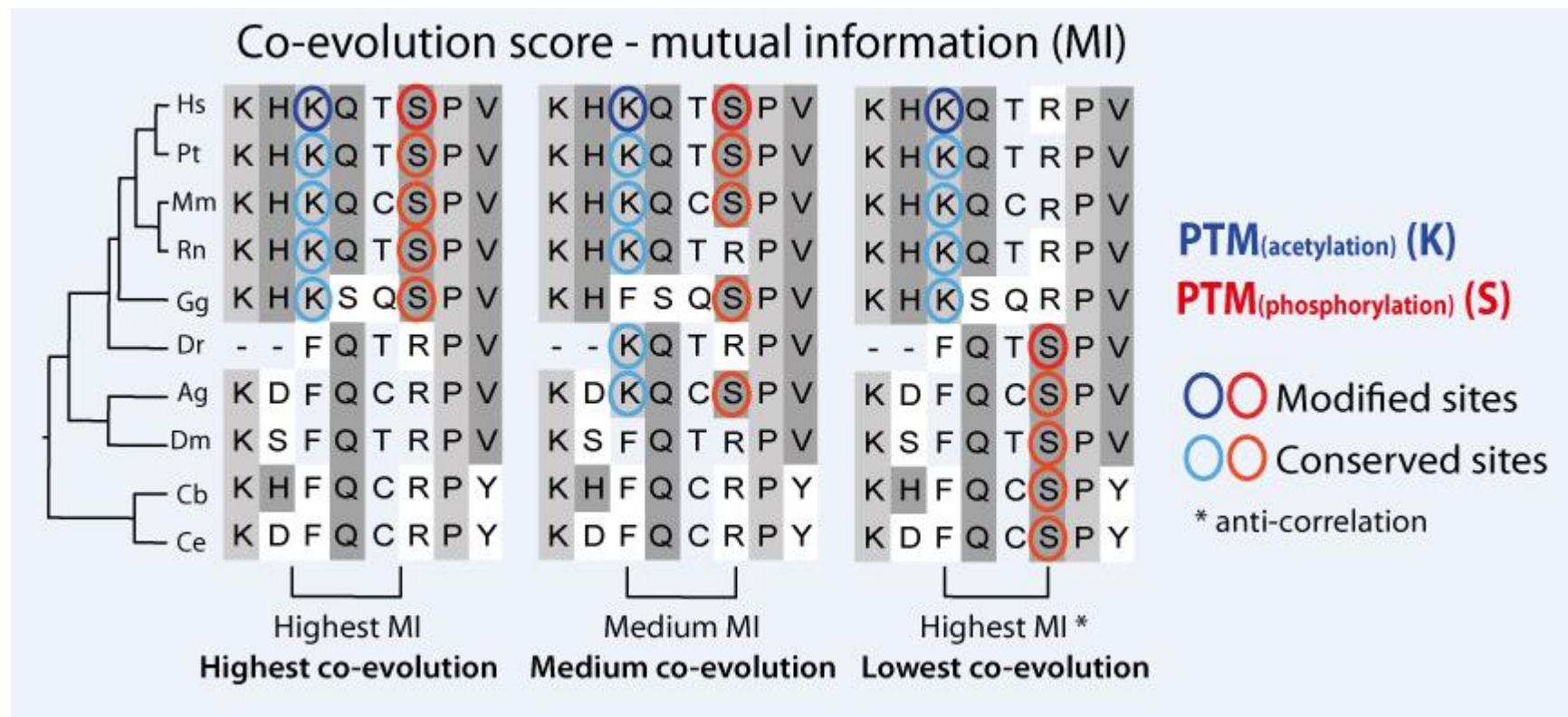
Science 16 Dec 2011:
Vol. 334, Issue 6062, pp. 1502-1503
DOI: 10.1126/science.1215894



Mutual information:

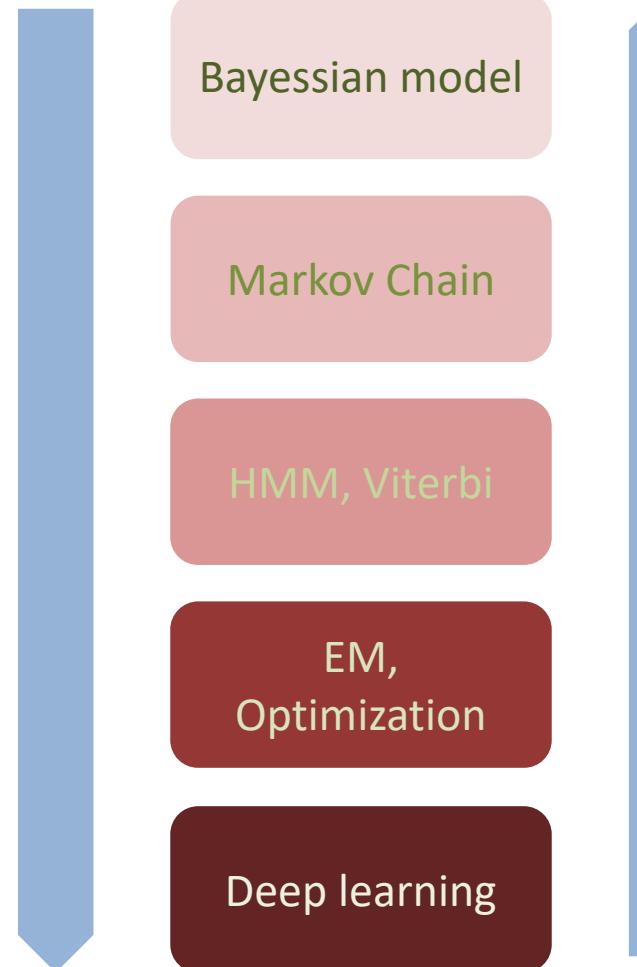
Mutual information is how much information about X that can be obtained by observing Y

Mutual information



Statistical modeling

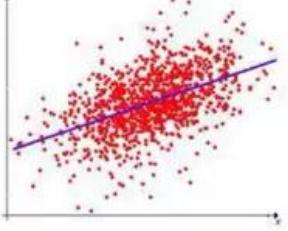
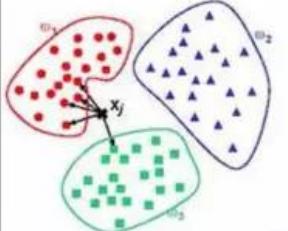
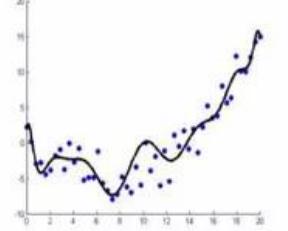
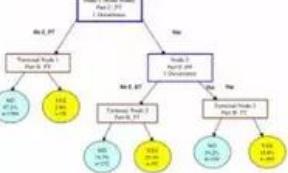
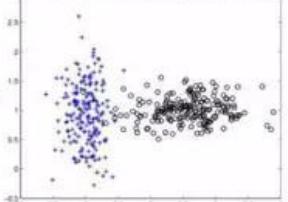
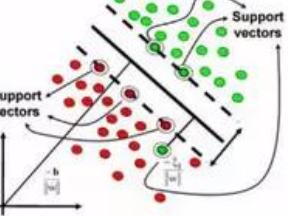
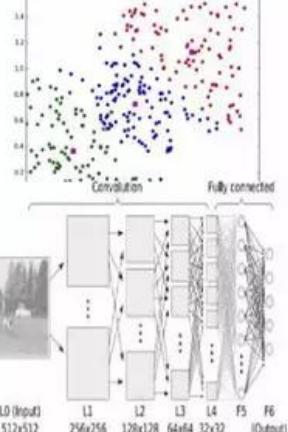
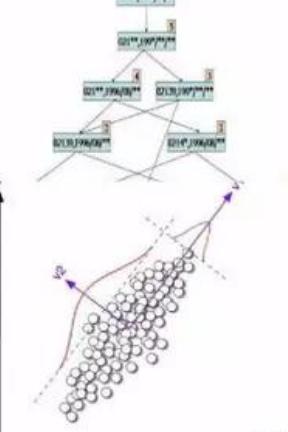
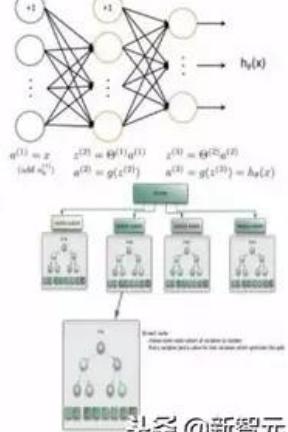
More complex model



Less data required

The main
models

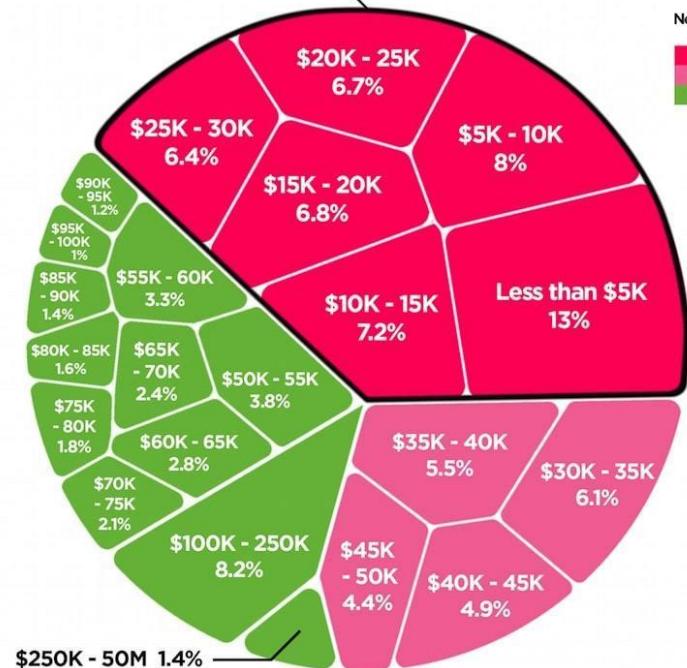
Statistical modeling

回归算法	基于实例的算法	正则化方法
		
决策树学习	贝叶斯方法	基于核的算法
		
聚类算法	关联规则学习	人工神经网络
		 <p>Diagram illustrating an Artificial Neural Network (ANN) structure:</p> <ul style="list-style-type: none">Input layer: L_0 (Input) with size 512×512.Hidden layers: L_1 (256x256), L_2 (128x128), L_3 (64x64), L_4 (32x32), L_5 (16x16), L_6 (8x8).Output layer: L_7 (Output). <p>Annotations:</p> <ul style="list-style-type: none">$\Theta^{(1)} = z^{(1)}$, $z^{(2)} = \Theta^{(1)}a^{(1)}$, $a^{(2)} = g(z^{(2)})$, $z^{(3)} = \Theta^{(2)}a^{(2)}$, $a^{(3)} = g(z^{(3)}) = h_{\theta}(x)$Cost function: $J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{cost}(h_{\theta}(x^{(i)}) - y^{(i)})$Training process: repeat $\theta := \theta - \alpha \nabla J(\theta)$ until θ converges.

统计学：无处不在！

How Much Americans Make in Wages Distribution of wage earners by level of net compensation

48% of wage earners had net compensation less than or equal to the median wage, which is estimated to be \$31,561.49 for 2017

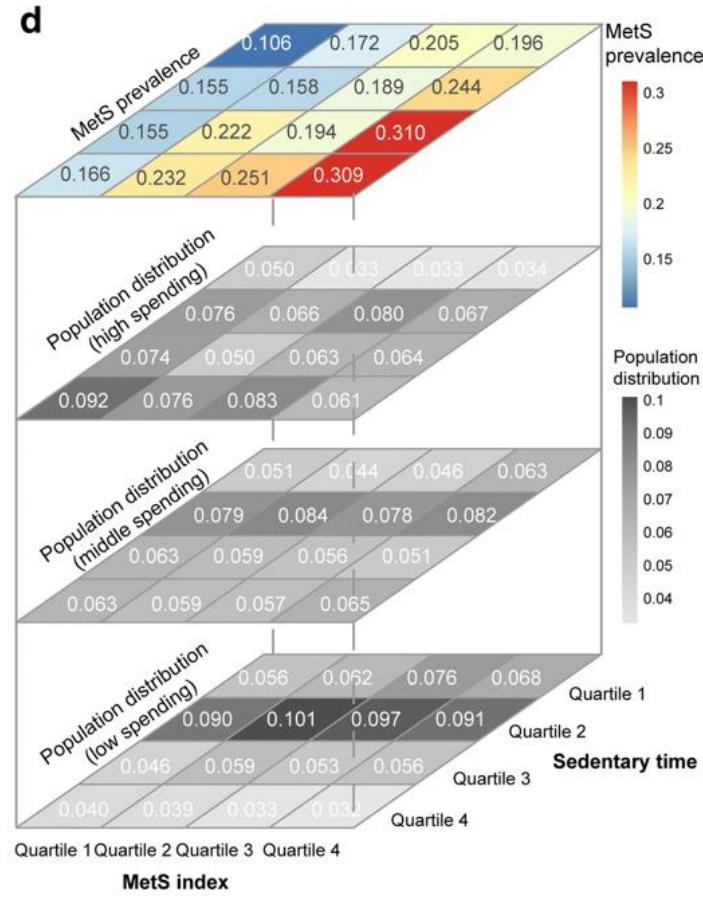
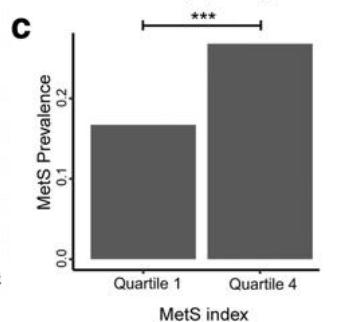
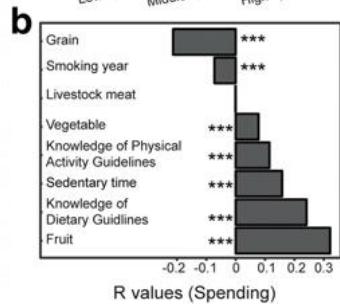
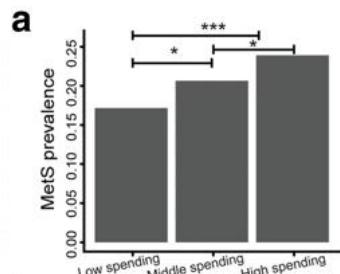


Article and Sources:

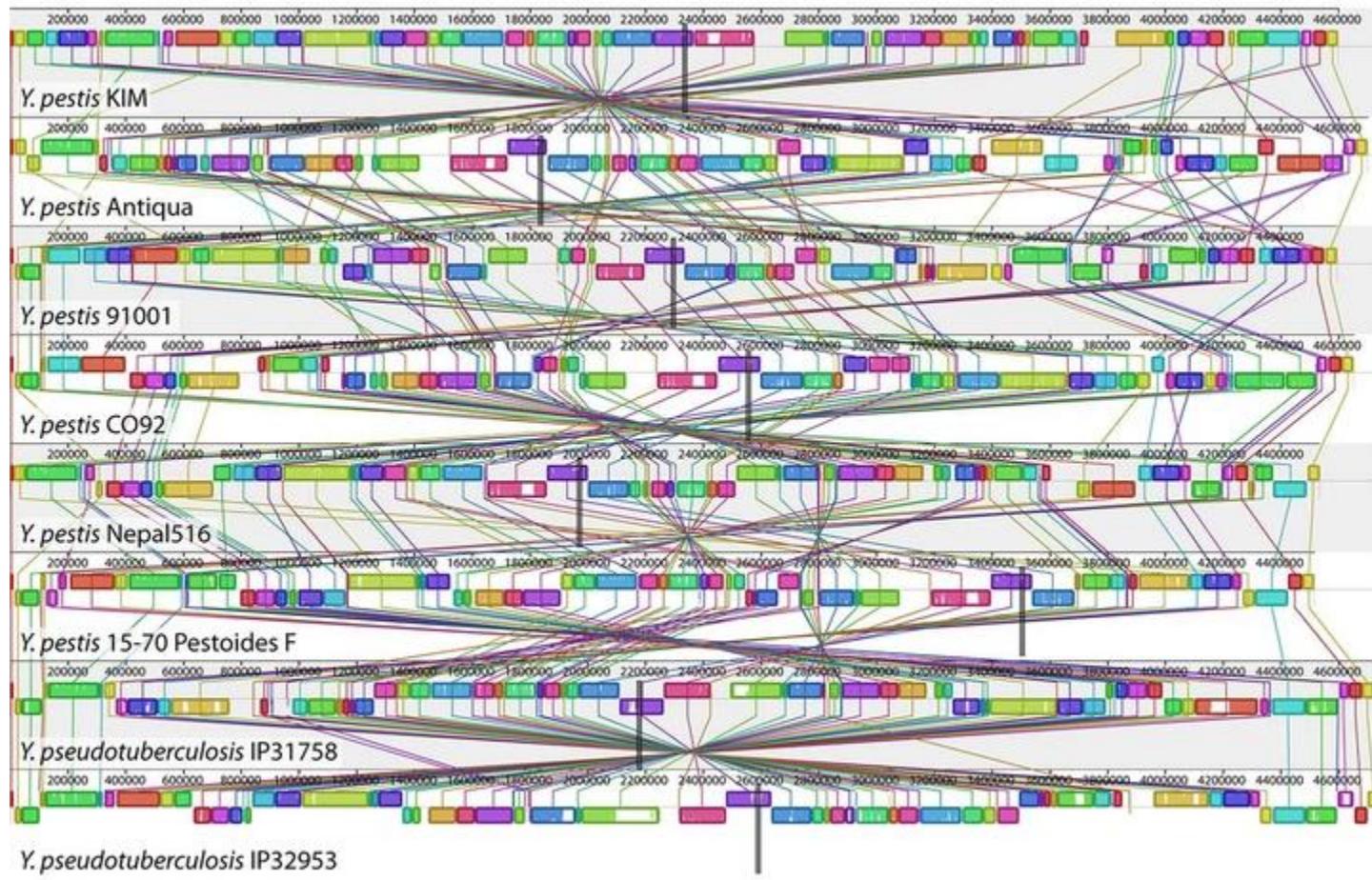
<https://howmuch.net/articles/how-much-americans-make-in-wages>

Social Security Administration - <https://www.ssa.gov>

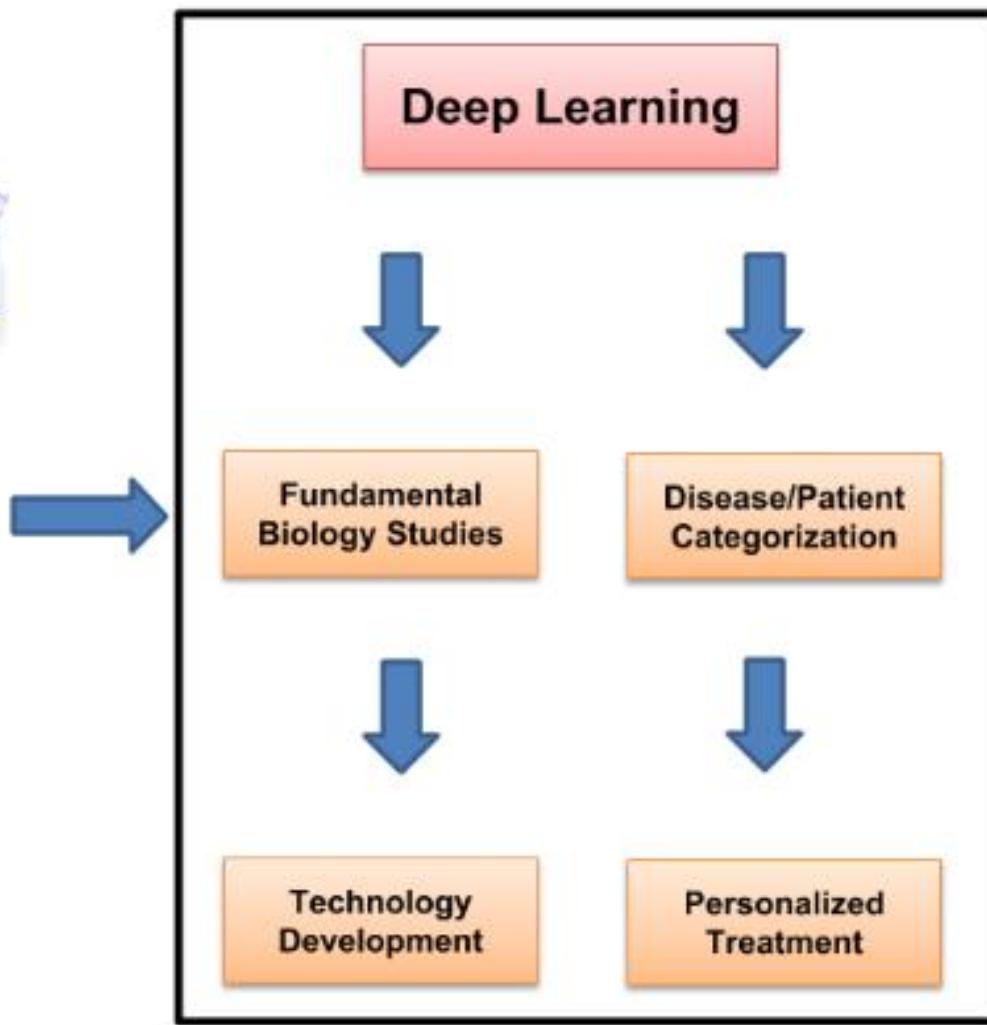
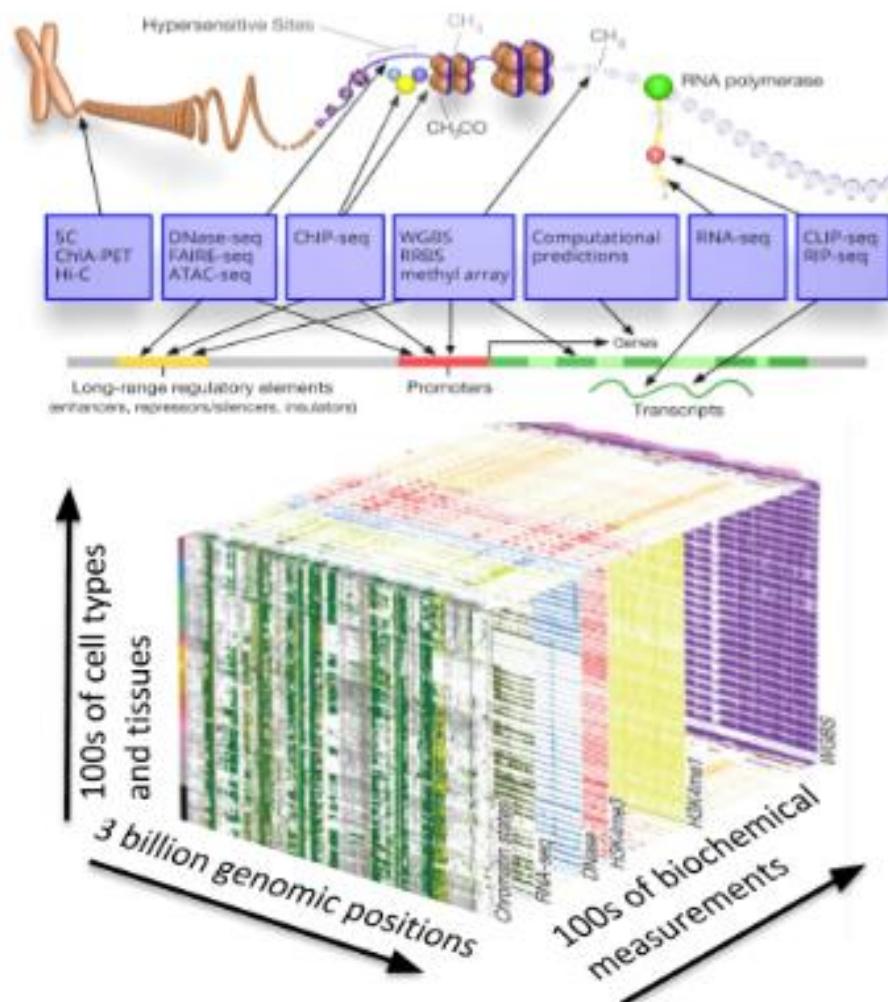
howmuch .net



统计学：无处不在！



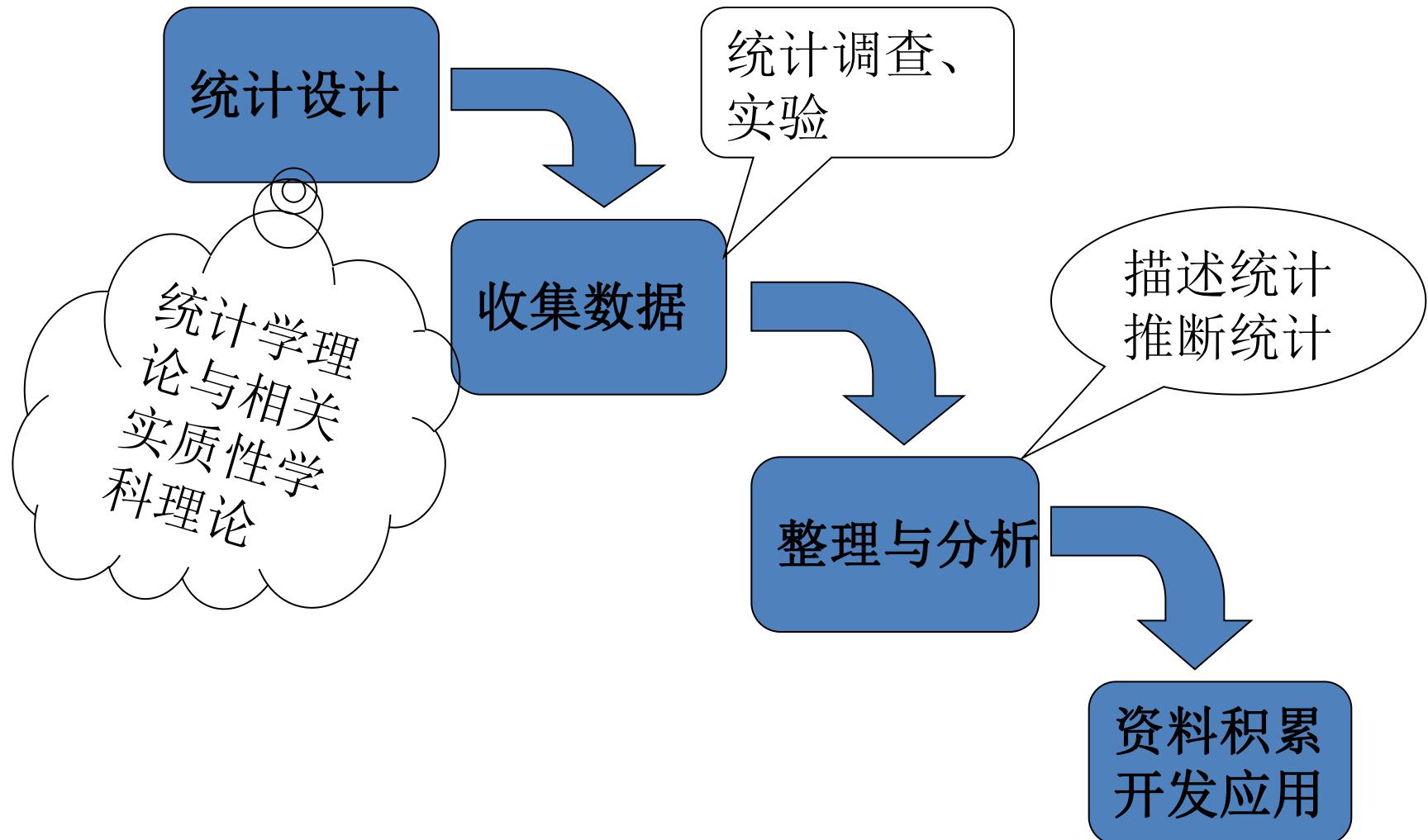
统计学：无处不在！



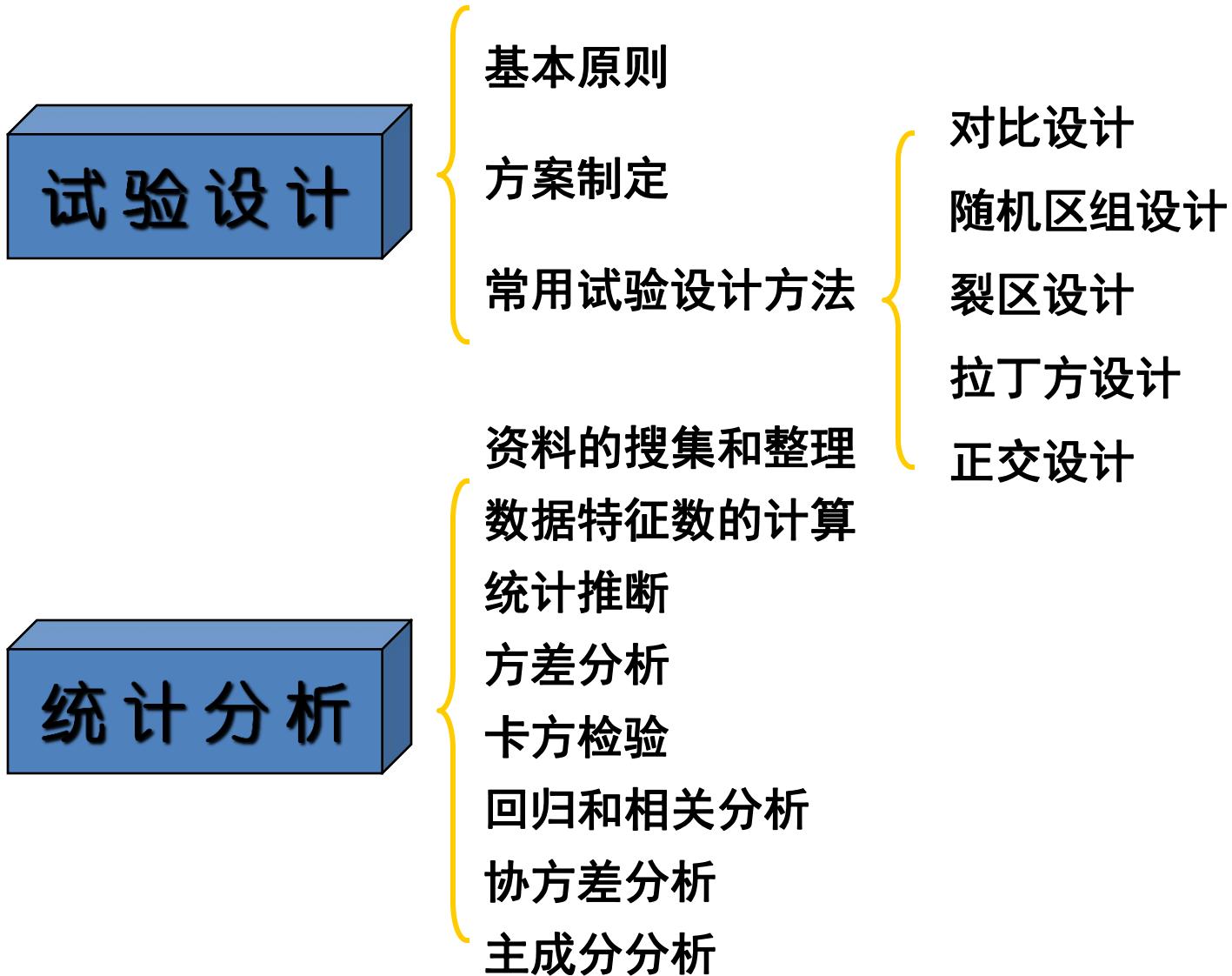
定义

生物统计学（**Biostatistics**）是数理统计在生物学研究中的应用，它是应用数理统计的原理，运用统计方法来认识、分析、推断和解释生命过程中的各种现象和试验调查资料的科学。属于生物数学的范畴。

生物统计分析的一般过程



生物统计学的基本内容



一、总体与样本

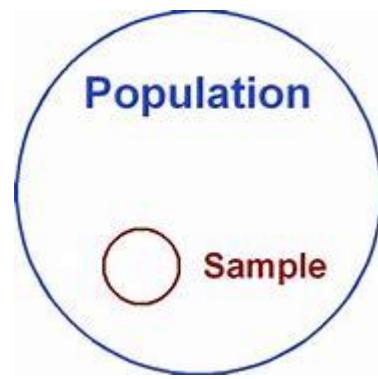
总体：具有相同性质或属性的个体所组成的集合

有限总体：含有有限个个体的总体

无限总体：包含有极多或无限多个体的总体

个体：组成总体的基本单元

样本：从总体中抽出若干个体所构成的集合



样本单位：构成样本的每个个体

样本容量（样本大小）：样本中所包含的个体数目，常记为 n 。

一般在生物学研究中，通常把 $n < 30$ 的样本叫**小样本**， $n \geq 30$ 的样本叫**大样本**。对于小样本和大样本，在一些统计数的计算和分析检验上是不一样的。

研究的目的是要了解总体，然而能观测到的却是样本，**通过样本来推断总体**是统计分析的基本特点。

二、变量与常量

变量: 或变数，指相同性质的事物间表现差异性或差异特征的数据。 (x_i)

常数: 表示能代表事物特征和性质的数值，通常由变量计算而来，在一定过程中是不变的。 (μ)



三、参数与统计数

参数：描述总体特征的数，通常未知

总体平均数(μ)，总体方差(σ^2)，…

统计数：描述样本特征的数，是样本观测值
的已知函数

样本平均数(\bar{x})，样本方差(s^2)，…

对总体的推断是通过统计数进行的

四、效应与互作

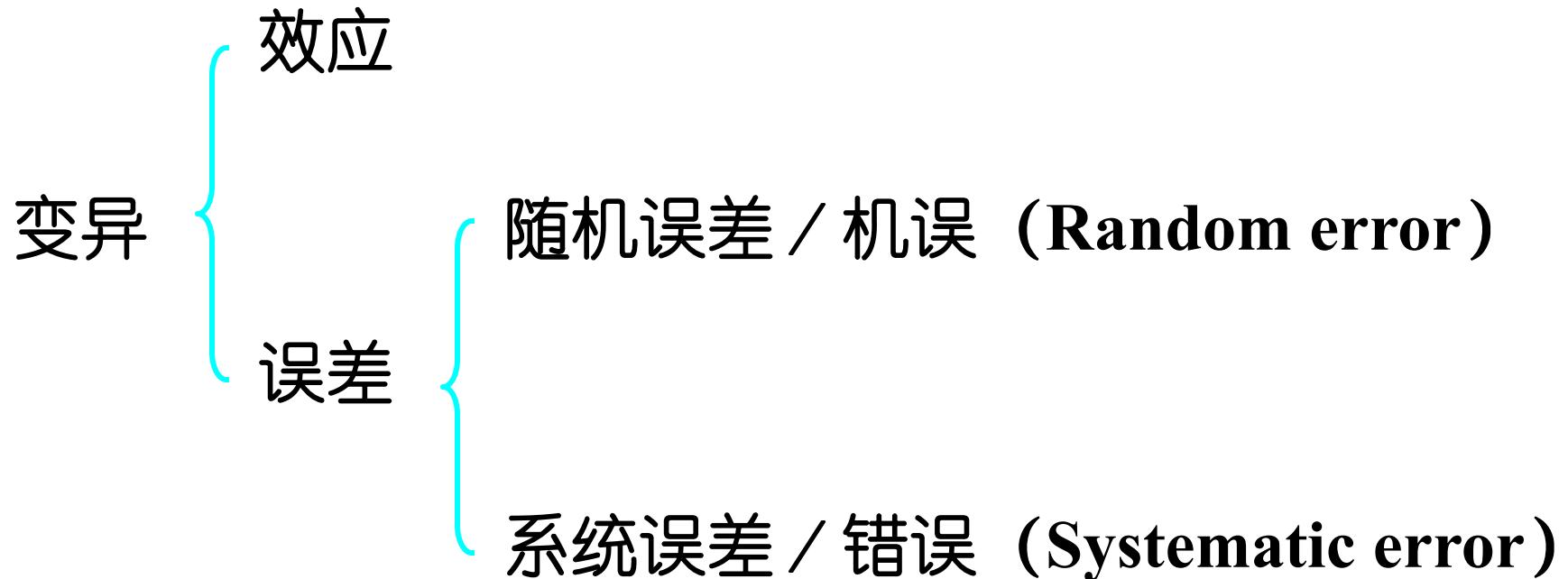
效应：通过施加试验处理，引起试验差异的作用。

效应是一个相对量，而非绝对量，表现为施加处理前后的差异。效应有**正效应**与**负效应**之分。

互作（连应）：是指两个或两个以上处理因素间相互作用产生的效应。

互作也有**正效应（协同作用）**与**负效应（拮抗作用）**之分。

五、机误与错误



随机误差，也叫**抽样误差(sampling error)**。这是由于试验中无法控制的内在和外在的偶然因素所造成。统计上的试验误差一般都指随机误差。随机误差越小，试验精确性越高。

系统误差，也叫**片面误差 (lopsided error)**。这是由于试验处理以外的其他条件明显不一致产生的。测量仪器不准、各批次药品间的差异、不同操作者操作习惯的差异等。系统误差影响试验的准确性，但是可以控制和避免的。

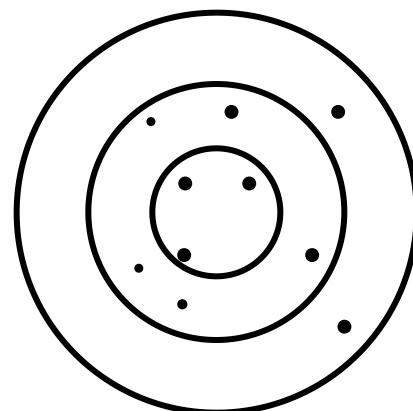
■ 错误，又称为 **过失性误差(gross error)**。在试验过程中，人为因素引起的差错。

仪器校正不准、药品配制比例不当、称量不准
确、计算出错等。**这类错误是不允许出现的。**

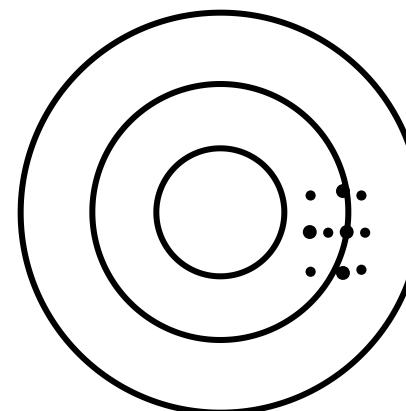
六、准确性与精确性

准确性(accuracy), 也叫**准确度**, 指在调查或试验中某一试验指标或性状的观测值与其真值接近的程度。

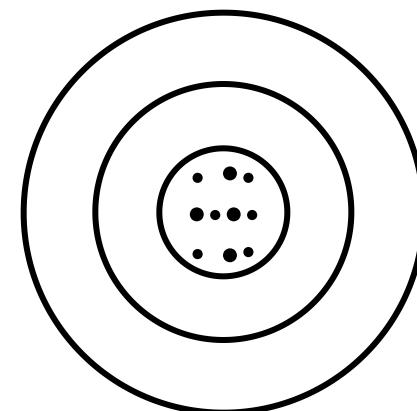
精确性(precision), 也叫**精确度**, 指调查或试验中同一试验指标或性状的重复观测值彼此接近的程度。



低准确性
低精确性



低准确性
高精确性

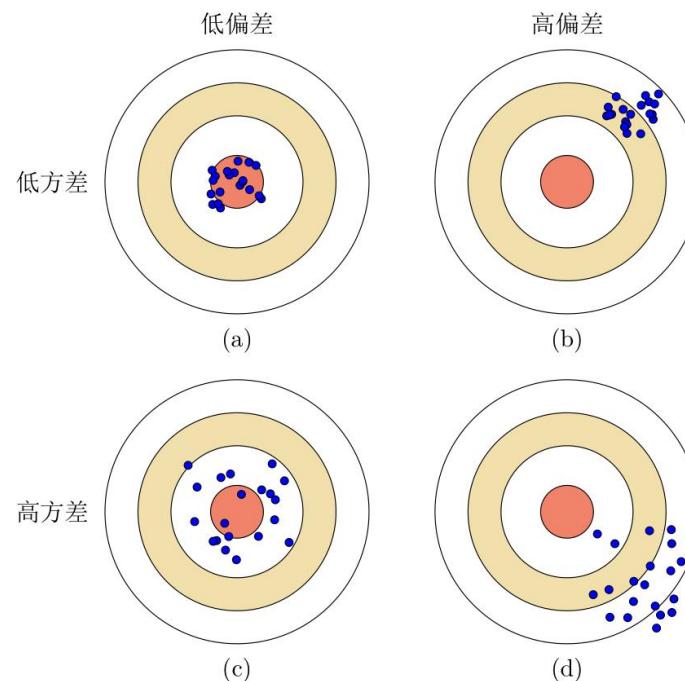


高准确性
高精确性

六、准确性与精确性

准确性(accuracy), 也叫**准确度**, 指在调查或试验中某一试验指标或性状的观测值与其真值接近的程度。

精确性(precision), 也叫**精确度**, 指调查或试验中同一试验指标或性状的重复观测值彼此接近的程度。



Part II

生物统计学理论基础



Bridging the sciences

BIOSTATISTICS

Clinical Study
Planning

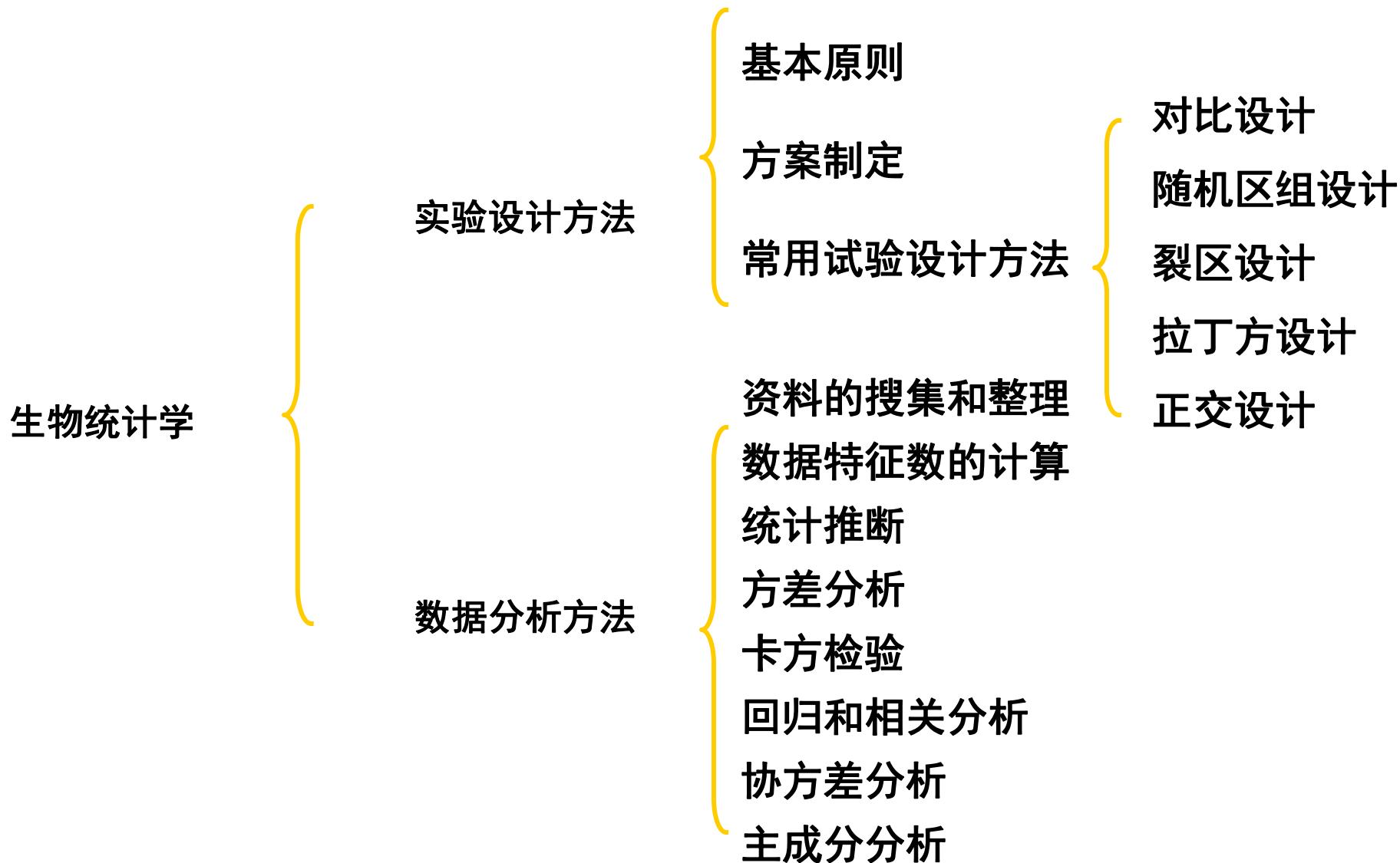
Development of
Protocol / SAP

"Operational"
statistics: tables,
listings, figures

Output: clinical
study reports,
scientific
communications

Interpret:
internal /
external consult,
regulatory,
reimbursement,
customers

一、基本框架



二、基本原则

重复、随机、局部控制三个基本原则：

是试验设计中必须遵循的原则，再采用相应的统计分析方法，就能最大程度地降低试验误差，无偏估计处理效应，从而对于各处理间的比较作出可靠的结论。

抽样方法的正确与否，直接关系到样本的代表性，影响由样本所得估计值的准确性。

随机抽样

典型抽样

顺序抽样

随机抽样

简单随机抽样

分层随机抽样

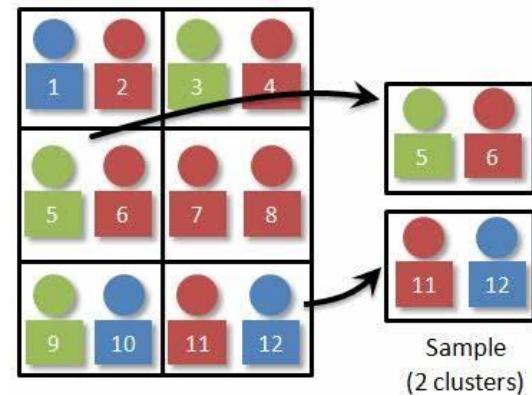
整体随机抽样

双重随机抽样

1 简单随机抽样

它是最简单、最常用的一种抽样方法，要求被抽总体内每一个体，被抽取的机会完全相等。

简单随机抽样就是采用随机的方法直接从总体中抽选若干个抽样个体组成样本的抽样方法。



- (1) 抽签法
 - (2) 随机数字表、随机数字生成器
- ◆ 处理在9个以内
- 例如：有8个处理，分别用1、2、3、4、5、6、7、8代表。用随机数字生成器得到一行随机数字为：

52648623399718302620

去掉序列中的0、9和重复数字，得到：

52648371

这就是8个处理在区组内的排列顺序，即第一小区安排5号处理，第二小区安排2号处理，第三号小区安排6号处理，余类推。

- 处理在9个以上

例如：有12个处理，随机取得97、39、24、89、90、89、86、49、15、18、25、43、80、74、30、41、67、36、43、58、42、07、04、25、17、54、60、88、49、34、42等随机数，去掉97，大于12的数用12除后取余数，将重复数字划去，所得随机排列为：

3、12、5、6、2、1、7、8、10、4、9、11

2 分层随机抽样

分层随机抽样是一种混合抽样。其特点是将总体按变异原因或程度划分成若干区层，然后再用简单随机抽样方法，从各区层按一定的抽样分数抽选抽样单位。

抽样分数：一个样本所包括抽样单位数与其总体所包括的抽样单位数的比值。

(1) 将总体变异原因与程度划分成若干区层，使得区层内变异尽可能小或变异原因相同，而区层间变异比较大或变异原因不明。

(2) 在每一个区层按一定的抽样分数独立随机抽样。



(1) 若总体内各抽样单位间的差异比较明显，那么就可以把总体分为几个比较同质的区层，从而提高抽样的准确度；



(2) 分层随机抽样既运用了随机原理，也运用了局部控制原理，这样不仅可以降低抽样误差，也可以运用统计方法来估算抽样误差；

3 整体随机抽样

整体随机抽样是把总体分成若干群，以群为单位，进行随机抽样，对抽到的样本进行全面调查。

如果总体内主要变异来源明显来自不同区层间，且每一区层均较大，则应采用分层抽样；若主要变异来源明显来自区层内各单位间，且每一区层所占面积较小，则宜用整体随机抽样。

顺序抽样

顺序抽样（系统抽样、机械抽样）

它是按某种既定顺序从总体（有限总体）中抽取一定数量的个体构成样本。

这种抽样方法可避免人们主观偏见的影响，且使用简便

如果总体内存在周期性变异，则可能会得到一个偏差很大的样本，这种现象在统计上称为系统误差。

由顺序抽样得到的样本不能计算抽样误差、估计总体值。
◦

典型抽样

根据初步资料或经验判断，有意识、有目的的选取一个典型群体作为代表（样本）进行调查，以估计整个总体，这种抽样方法就称为典型抽样。

典型样本代表着总体的绝大多数，如果选择合适，可得到可靠的结果，尤其从容量很大的总体中选取较小组数量的抽样单位时，往往采用这种抽样方法。

这种抽样多用于大规模社会经济调查，而在总体相对较小或要求估算抽样误差时，一般不采用这种方法。

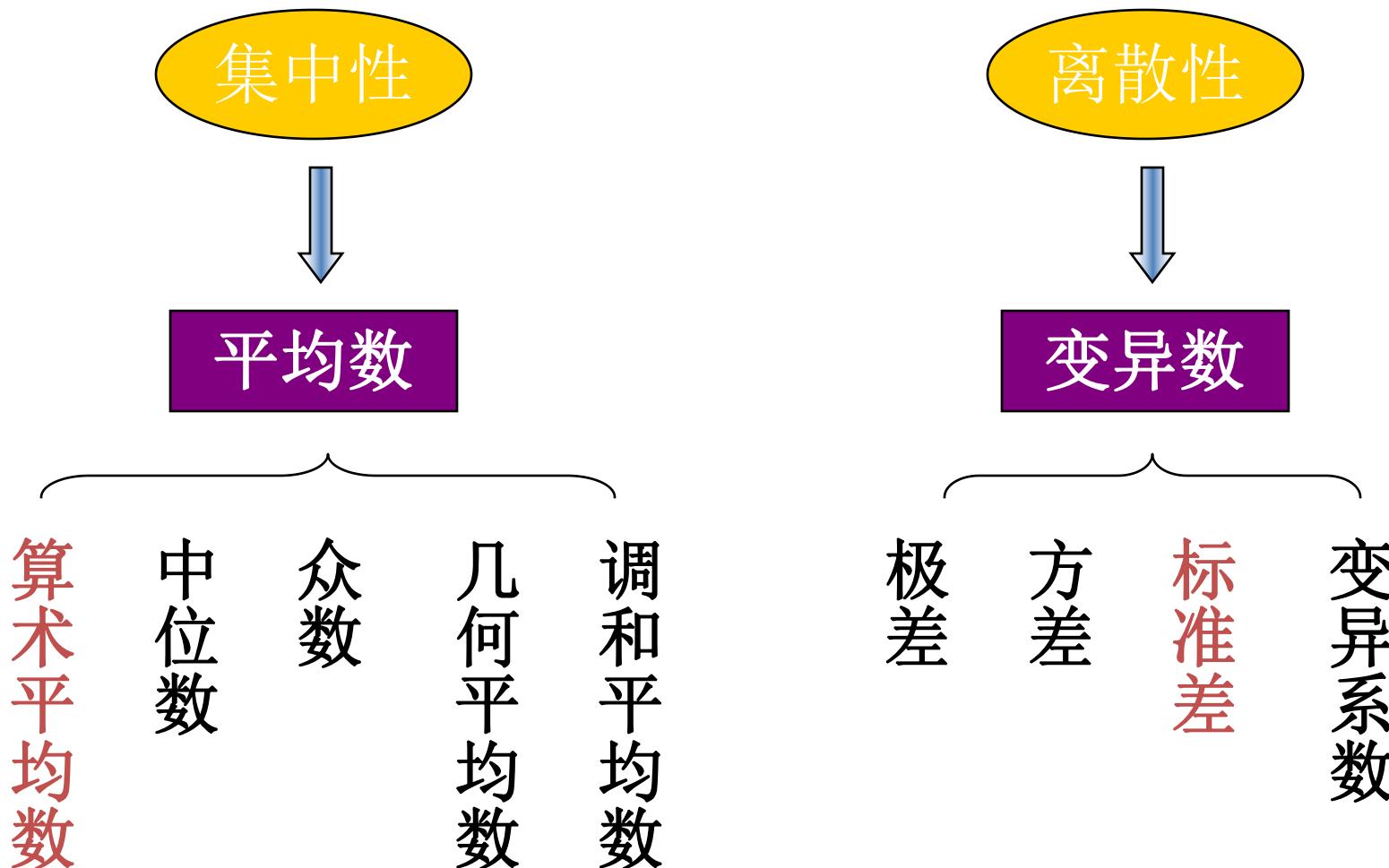
三、实验分组和统计归纳

变量的分布具有两种明显的基本特征：集中性和离散性。

集中性 变量在趋势上有着向某一中心聚集，或者说以某一数值为中心而分布的性质。

离散性 变量有着离中分散变异的性质。

三、实验分组和统计归纳



三、实验分组和统计归纳

平均数 平均数是统计学中最常用的统计量，是计量资料的代表值，表示资料中观测数的中心位置，并且可作为资料的代表与另一组相比较，以确定二者的差异情况。



三、实验分组和统计归纳

(一) 平均数的种类

算术平均数

中位数

众数

几何平均数

调和平均数



三、实验分组和统计归纳

1. 算术平均数 (arithmetic mean)

定义：总体或样本资料中所有观测数的总和除以观测数的个数所得的商，简称平均数、均数或均值。

总体： $\mu = \frac{x_1 + x_2 + x_3 + \dots + x_N}{N} = \frac{1}{N} \sum_{i=1}^N xi$

样本： $\bar{X} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n xi$

三、实验分组和统计归纳

2. 中位数(median)

$$\longrightarrow M_d$$

资料中所有观测数依大小顺序排列，居于中间位置的观测数称为中位数或中数。

三、实验分组和统计归纳

1) 当观测值个数 n 为奇数时, $(n+1)/2$ 位置的观测值, 即 $x_{(n+1)/2}$ 为中位数:

$$M_d = x_{(n+1)/2}$$

2) 当观测值个数为偶 数 时, $n/2$ 和 $(n/2+1)$ 位置的两个观测值之和的 $1/2$ 为中位数, 即:

$$M_d = \frac{x_{n/2} + x_{(n/2+1)}}{2}$$

三、实验分组和统计归纳

3. 众数(mode)

$$\longrightarrow M_0$$

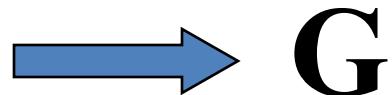
资料中出现次数最多的那个观测值或次数最多一组的组中值或中点值。

注意：

- (1) 对于某些数据而言，如均匀分布，并不存在众数；
- (2) 对于某些数据存在两个或两个以上的众数；
- (3) 主要用来描述频率分布。

三、实验分组和统计归纳

4. 几何平均数 (geometric mean)



资料中有n个观测数，其乘积开n次方所得数值。

$$G = \sqrt[n]{x_1 * x_2 * x_3 \dots * x_n}$$

适用范围：几何均数适用于变量X为对数正态分布，经对数转换后呈正态分布的资料。

三、实验分组和统计归纳

5. 调和平均数 (harmonic mean)

$$\longrightarrow H$$

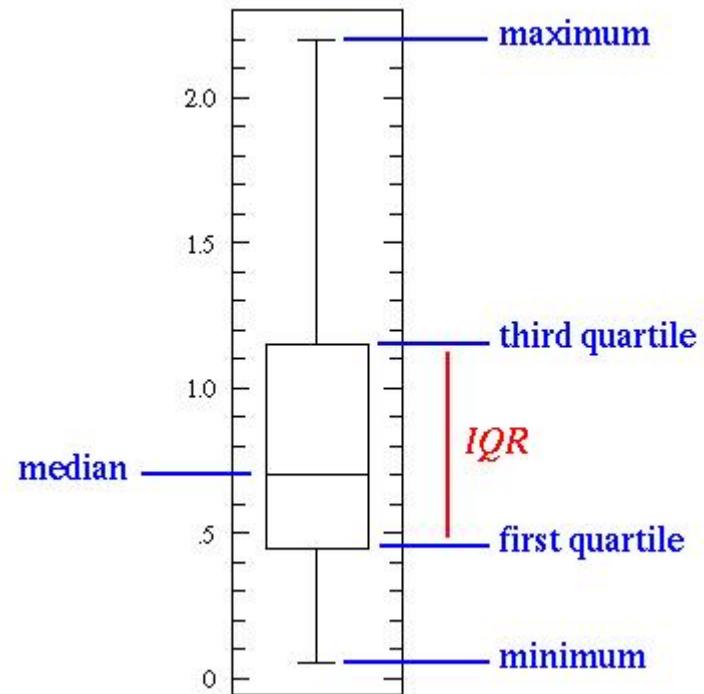
资料中各观测值倒数的算术平均数的倒数。

$$H = \frac{1}{\frac{1}{n} \sum \frac{1}{x}}$$

适用范围：主要用于反映生物不同阶段的平均增长率或不同规模的平均规模。

三、实验分组和统计归纳

箱式图 (Box plot)



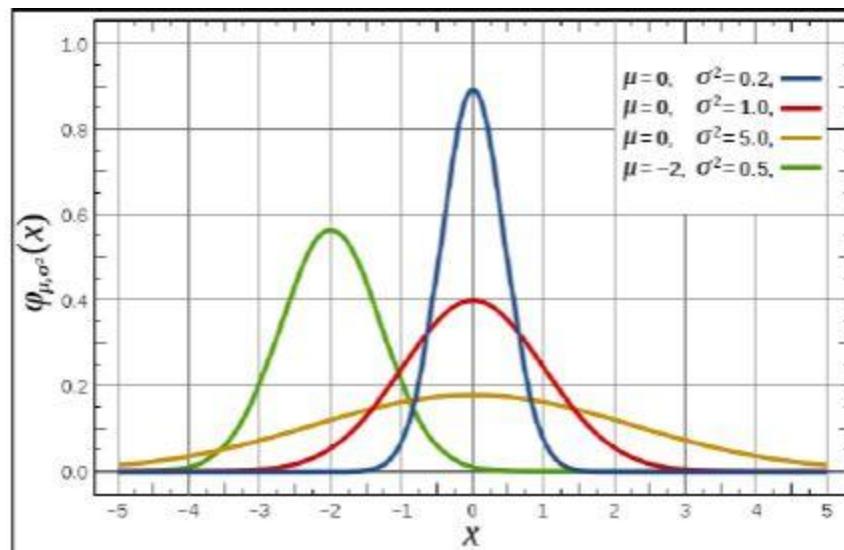
四、概率分布

统计定义：设在相同的条件下，进行大量重复试验，若事件A的频率稳定地在某一确定值p的附近摆动，则称p为事件A出现的概率。

$$P(A) = p = \lim \frac{m}{n} \approx \frac{m}{n}$$

在一般情况下，随机事件的概率P是不可能准确得到的。通常以试验次数n充分大时，随机事件A的频率作为该随机事件概率的近似值。

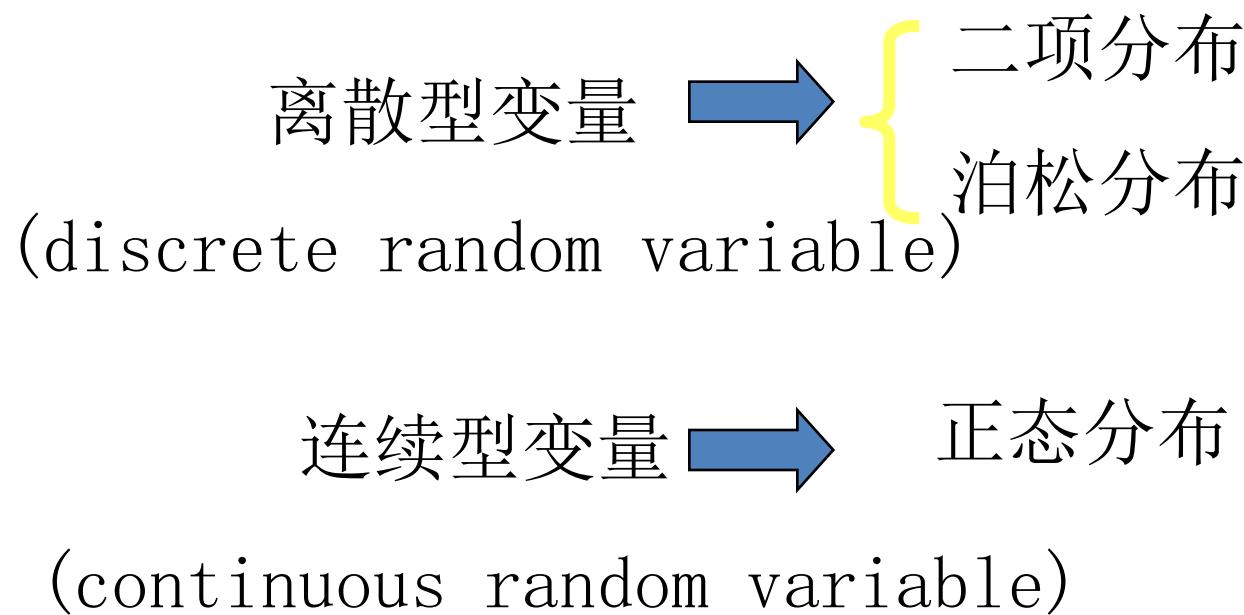
四、概率分布



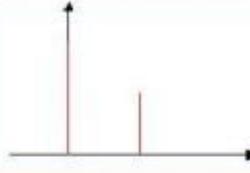
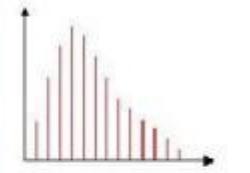
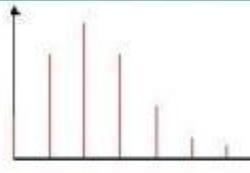
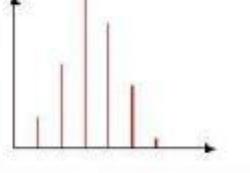
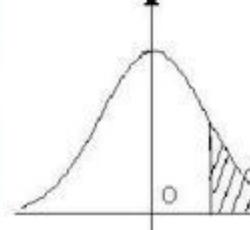
随机变量可能取得的每一个实数值或某一范围的实数值是有一个相应概率于其对应的，这就是所要研究和掌握的规律，这个规律称为随机变量的概率分布。

四、概率分布

随机变量的概率分布 (probability distribution)



四、概率分布

分布名称	概率与密度函数 $p(x)$	数学期望	方差	图形
贝努里分布	$p_k = \begin{cases} q, & k=0 \\ p, & k=1 \end{cases}$	p	pq	
两点分布	$0 < p < 1, q = 1 - p$			
二项分布 $b(k, n, p)$	$b(k; n, p) = \binom{n}{k} p^k q^{n-k}$ $k = 0, 1, \dots, n$ $0 < p < 1, q = 1 - p$	np	npq	
泊松分布 $p(k, \lambda)$	$p(k, \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}, \lambda > 0$ $k = 0, 1, 2, \dots, n$	λ	λ	
几何分布 $g(k, p)$	$g(k, p) = q^{k-1} p$ $k = 1, 2, \dots, n$ $0 < p < 1, q = 1 - p$	$\frac{1}{p}$	$\frac{q}{p^2}$	
正态分布 高斯分布 $N(\mu, \sigma^2)$	$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$ $-\infty < x < \infty, \mu, \sigma > 0, \text{常数}$	μ	σ^2	

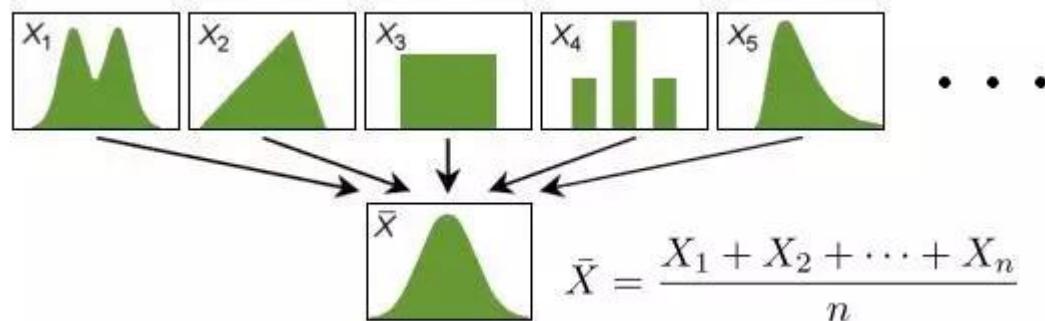
四、概率分布

中心极限定理

[Lindeberg-Levy中心极限定理] 设 X_1, \dots, X_n 独立同分布，且具有有限的均值 μ 和方差 σ^2 ，则在 $n \rightarrow \infty$ 时,有

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \rightarrow N(0, 1).$$

随意的一个概率分布中生成的随机变量，在序列和(或者等价的求算术平均)的操作之下，表现出如此一致的行为，统一的规约到正态分布。



五、假设推断

假设检验（hypothesis test）又称**显著性检验**（significance test）：

根据总体的理论分布和小概率原理，对未知或不完全知道的总体提出两种彼此对立的假设，然后由样本的实际结果，经过一定的计算，作出在一定概率意义上应该接受的那种假设的推断。

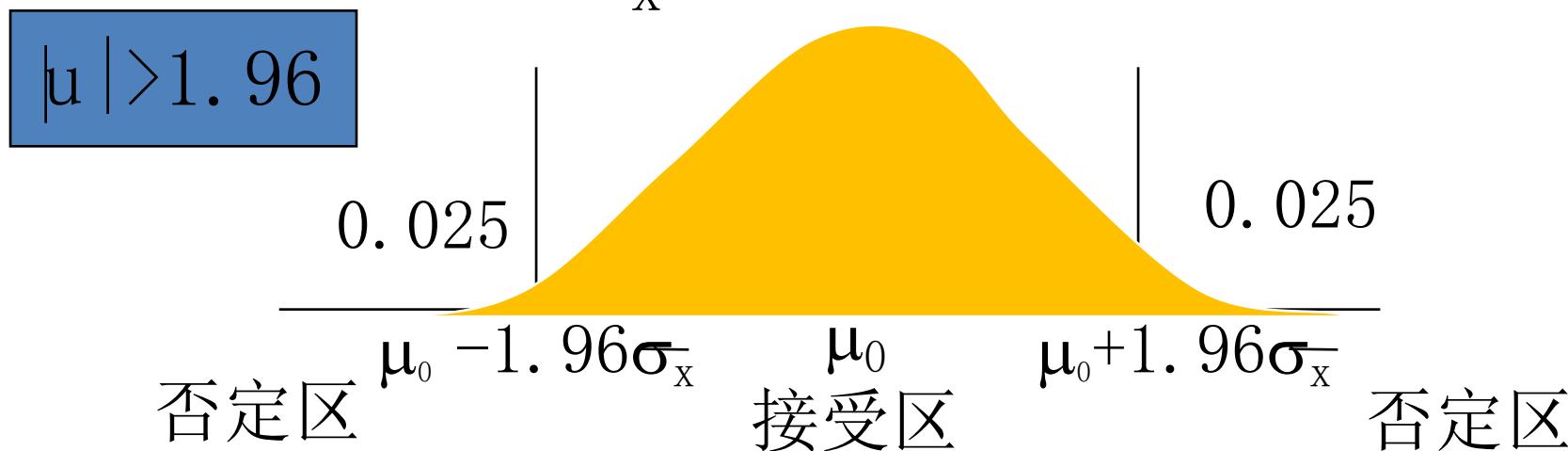
五、假设推断

根据研究设计的类型和统计推断的目的选择使用不同的检验方法。

例: $\mu_{\bar{x}} = \mu_0 = 126$

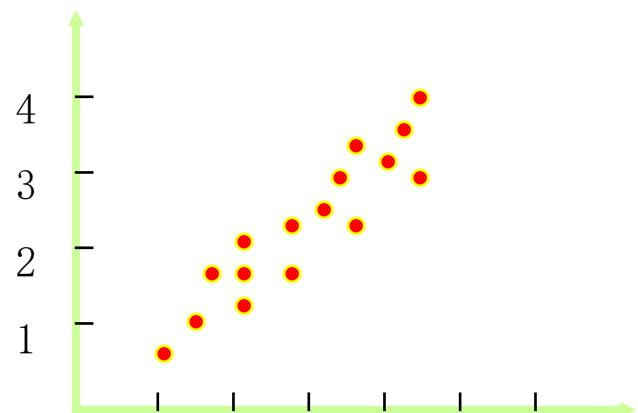
$$\sigma_{\frac{x}{x}}^2 = \frac{\sigma^2}{n} = \frac{240}{6} = 40$$

$$u = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{136 - 126}{\sqrt{40}} = 1.581$$

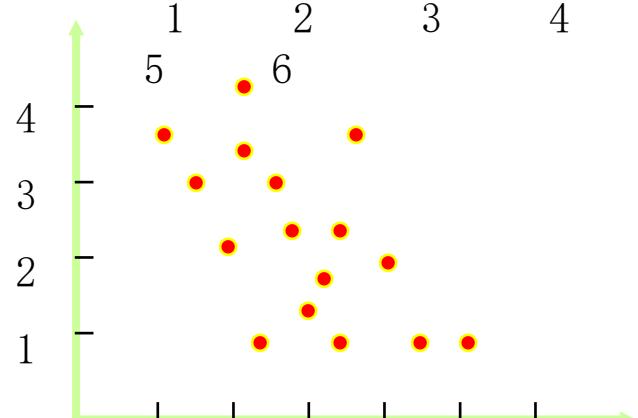


六、相关性分析

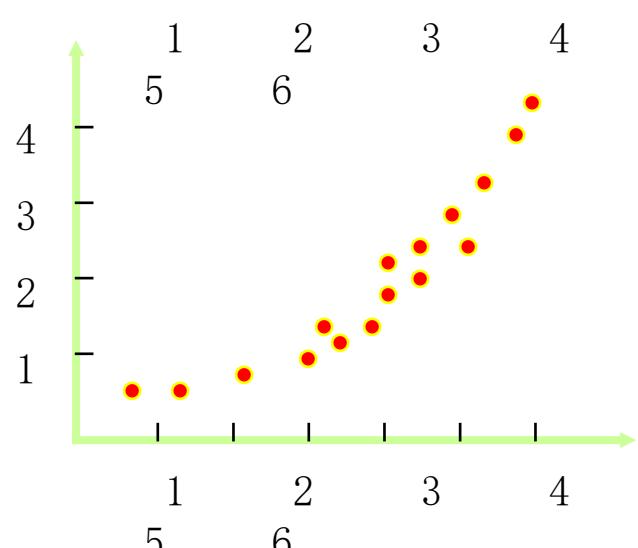
正向直线关系



负向直线关系



曲线关系



六、相关性分析

直线回归方程 (linear regression equation)

$$\hat{y} = a + bx$$

自变量

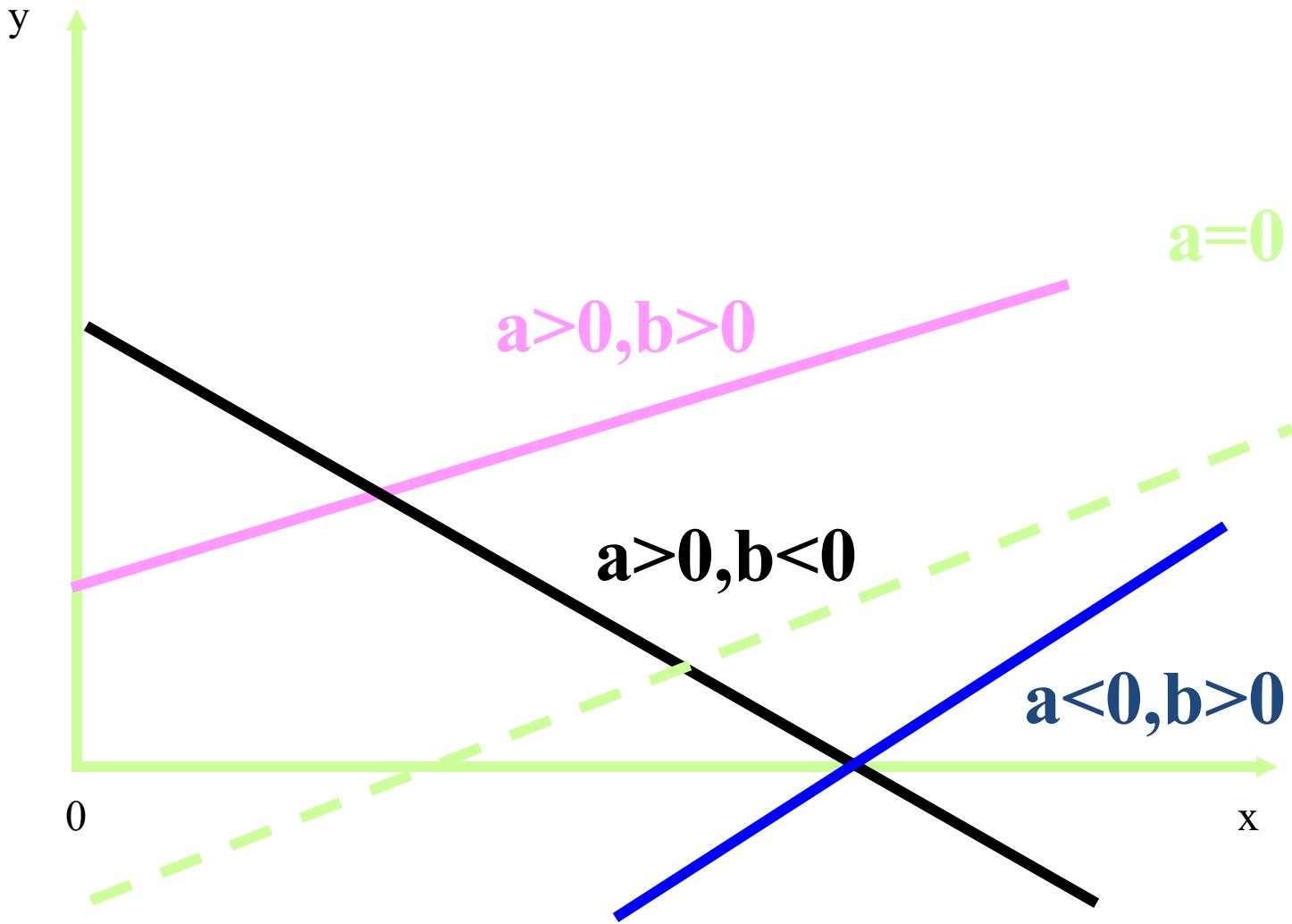
斜率(slope)
回归系数(regression coefficient)

截距(intercept)
回归截距

与x值相对应的依变量y的点估计值



直线回归方程 (linear regression equation)



六、相关性分析

直线回归方程 (linear regression equation)

最小二乘法

(method of least square)

$$\sum_1^n (y - \hat{y})^2$$

各点到回归直线的纵向距离的平方和最小

$$Q = \sum_1^n (y - \hat{y})^2 = \sum_1^n (y - a - bx)^2$$

六、相关性分析

作回归分析时要有实际意义

不能把毫无关联的两种现象勉强作回归分析，即便有回归关系也不一定是因果关系，还必须对两种现象的内在联系有所认识，即能从专业理论上作出合理解释或有所依据。

Example：统计研究发现，冰淇淋销量最高的时候，就是公共泳池的溺水事故发生得最多的时候。

六、相关性分析

进行直线回归分析之前，绘制散点图

当观察点的分布有直线趋势时，才适宜作直线回归分析。

散点图还能提示资料有无异常值，即对应于残差绝对值特别大的观测数据。异常点的存在往往对回归方程中的a和b的估计产生较大的影响。因此，需要复查此异常点的值。



六、相关性分析

直线回归的适应范围一般以自变量的取值为限

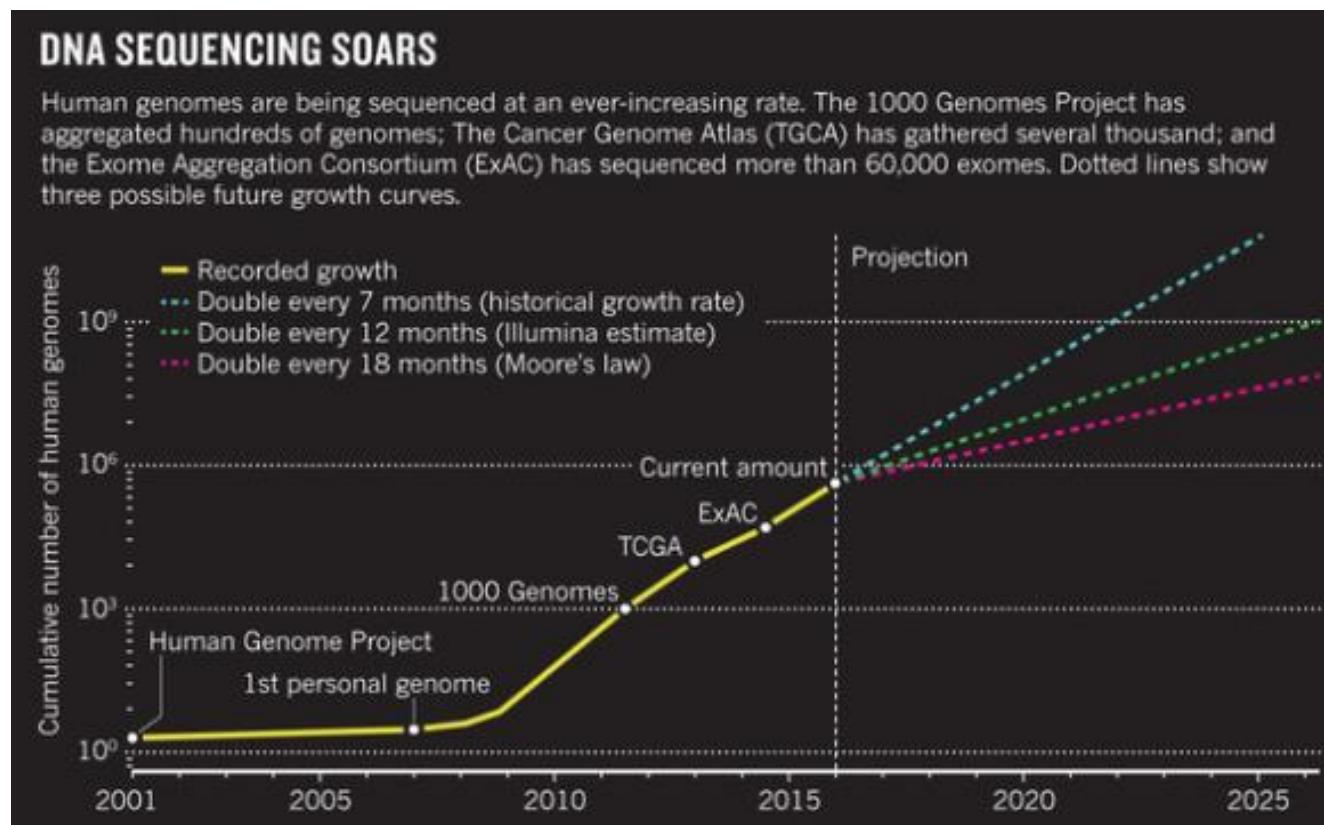
在自变量范围内求出的估计值，一般称为内插(interpolation);超过自变量取值范围所计算出的估计值，称为外延(extrapolation)。

若无充分理由证明超过自变量取值范围还是直线，应该避免外延。

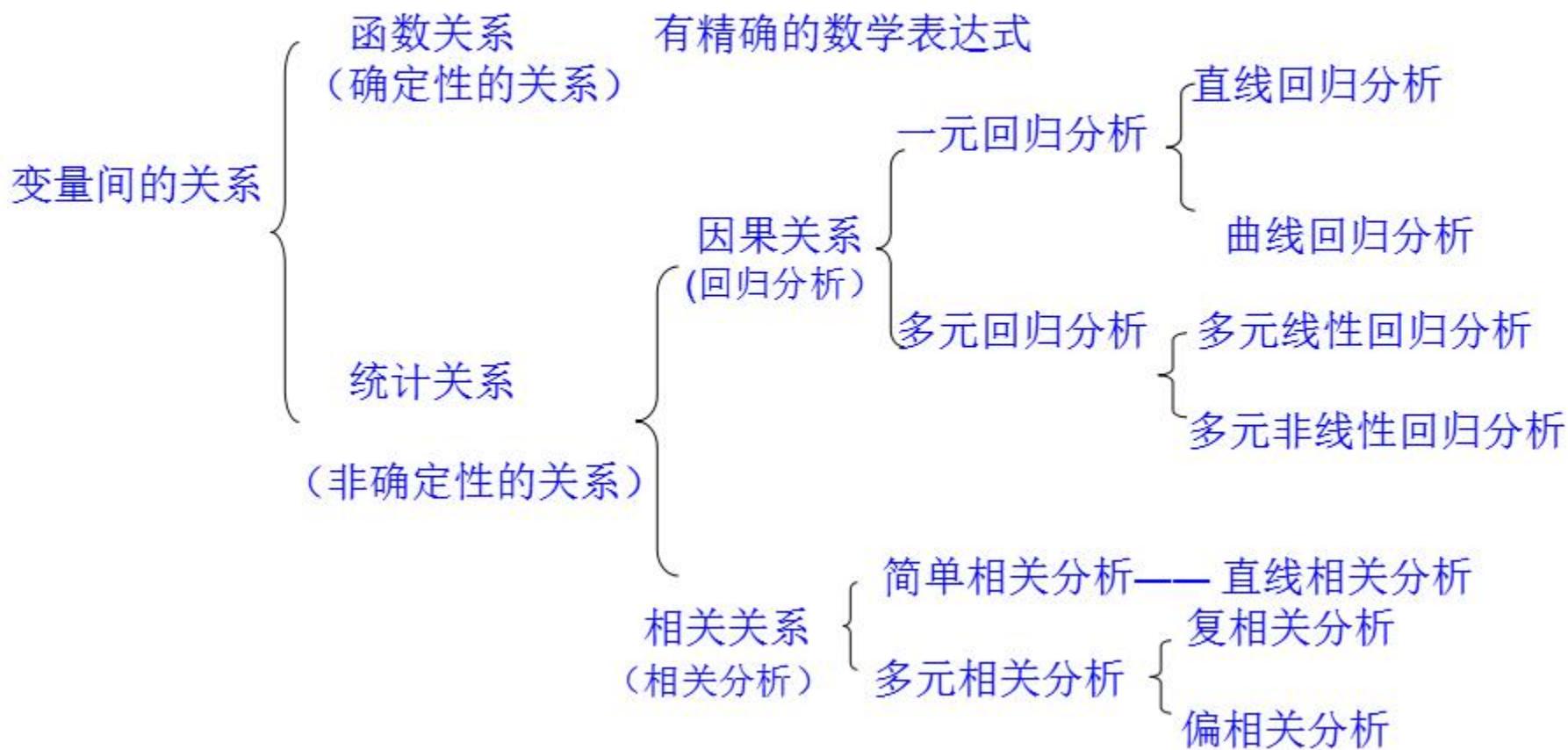


六、相关性分析

直线回归的适应范围一般以自变量的取值为限



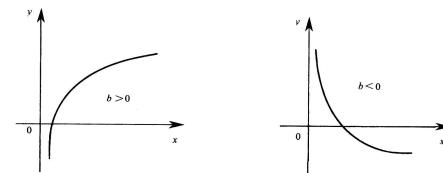
六、相关性分析



六、相关性分析

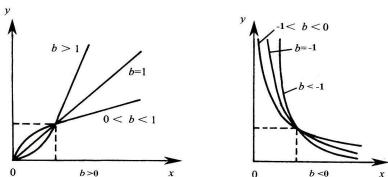
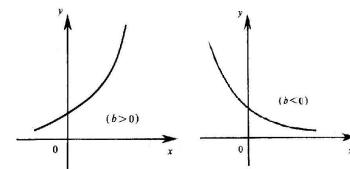
对数函数

$$\hat{y} = a + b \lg x$$



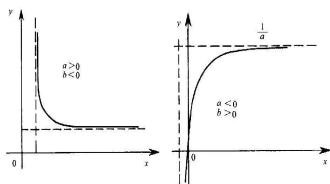
指数函数

$$\hat{y} = ae^{bx}$$



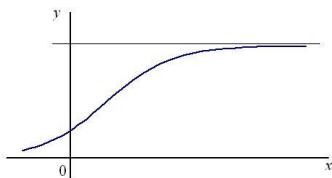
幂函数

$$\hat{y} = ax^b$$



倒数函数

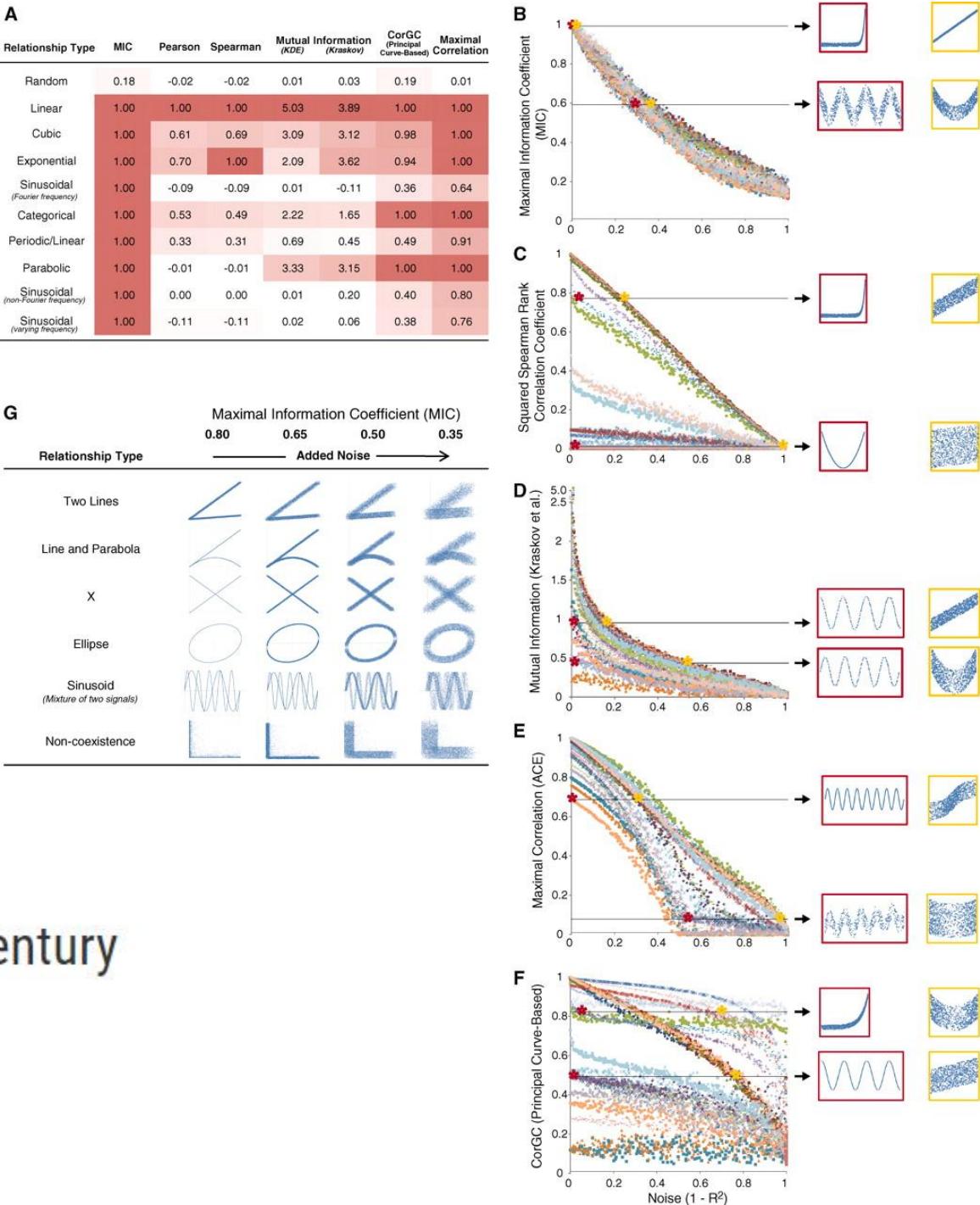
$$\frac{1}{\hat{y}} = a + \frac{b}{x}$$



S形曲线

$$\hat{y} = \frac{K}{1 + ae^{-bx}}$$

六、相关性分析



PERSPECTIVE | MATHEMATICS

A Correlation for the 21st Century

Terry Speed

* See all authors and affiliations

Science 16 Dec 2011:
Vol. 334, Issue 6062, pp. 1502-1503
DOI: 10.1126/science.1215894

七、试验设计与统计分析过程

广义

狭义

课题的名称

试验目的

研究依据、内容

研究的预期效果

试验方案

试验单位的选取

试验单位的重复数

试验单位的分组与排列方法

试验记录项目和要求

试验结果的分析方法

经济或社会效益分析

参加研究人员

已备条件

尚缺少的条件

试验时间、地点

进度安排、经费预算

成果形式

学术论文撰写

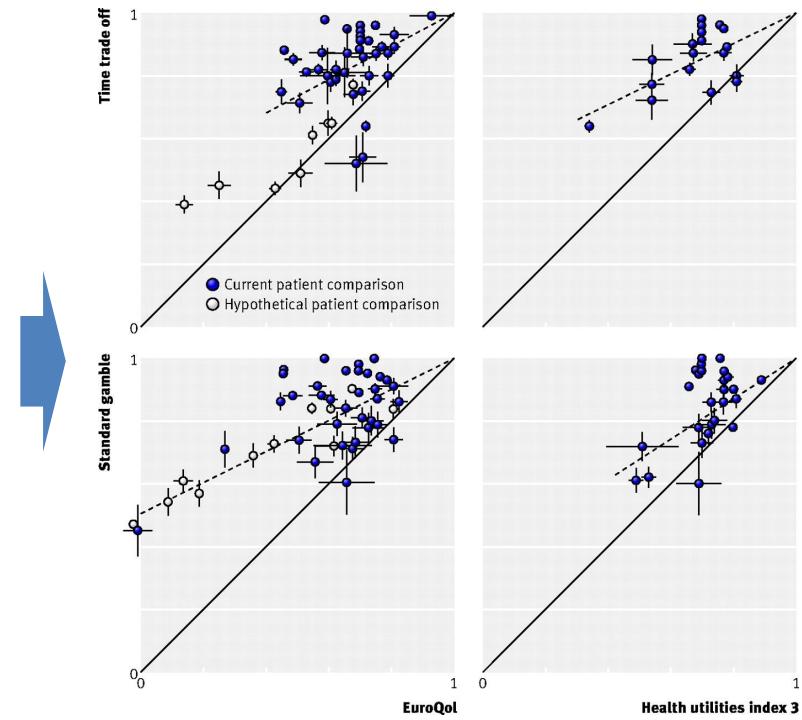
八、统计分析过程中经常会遇到的陷阱

- 相关性不等于因果性；
- 实验操作引入的组间差异；
- 降维问题；
- 深度学习得到的特征与样本无关。
- 。 。 。

八、统计分析过程中经常会遇到的陷阱

相关性不等于因果性

	A	B	C	D	E
1	Gender	Height	Weight	Height (cm)	Weight (kg)
2	Male	73.847017	241.89356	187.5714232	109.72099
3	Male	68.781904	162.31047	174.7060363	73.622732
4	Male	74.110105	212.74086	188.2396677	96.49755
5	Male	71.730978	220.04247	182.1966851	99.809504
6	Male	69.881796	206.3498	177.4997615	93.598619
7	Male	67.253016	152.21216	170.8226598	69.042216
8	Male	68.785081	183.92789	174.7141064	83.428219
9	Male	68.348516	167.97111	173.6052294	76.190352
10	Male	67.01895	175.92944	170.2281321	79.800187
11	Male	63.456494	156.39968	161.1794947	70.941642
12	Male	71.195382	186.60493	180.836271	84.642501
13	Male	71.640805	213.74117	181.967645	96.951285
14	Male	64.766329	167.12746	164.506476	75.807679
15	Male	60.98307	190.44618	175.078008	95.021979



Part III

生物统计学基本应用

Biometry

Nutrition

Treatment

Public Health

Survey Research

Environmental Health

Statistics

Medical Informatics

Genetics

Bioinformatics

Health Care Management

Pharmacology

Physiology

Health Care Policy

Epidemiology

Medicine

Health Care

Maternal Child Health

Biomathematics

Drug & Program Evaluations

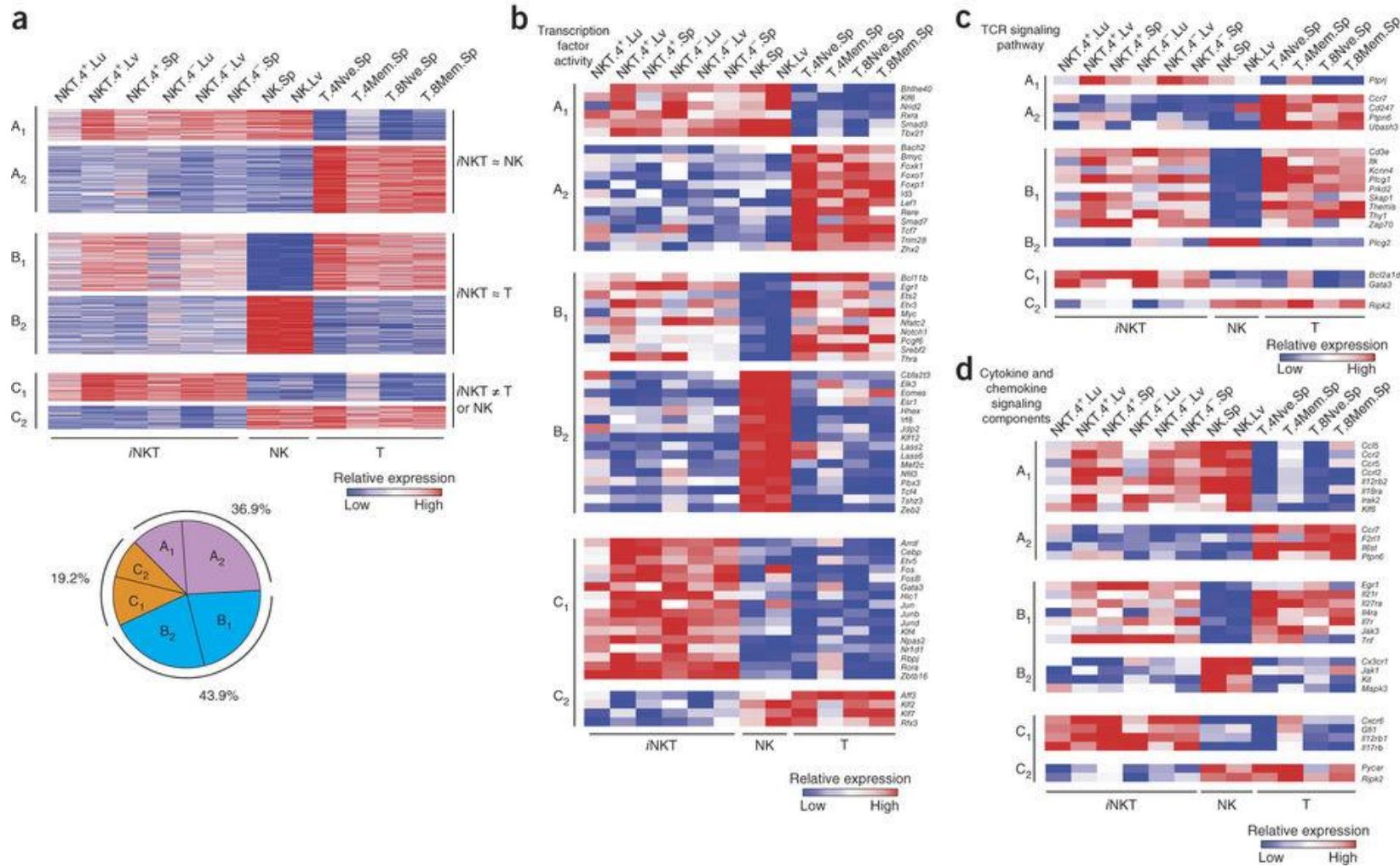
Population Health

Biology

Health & Social Behavior

Biotechnology

一、基因差异表达

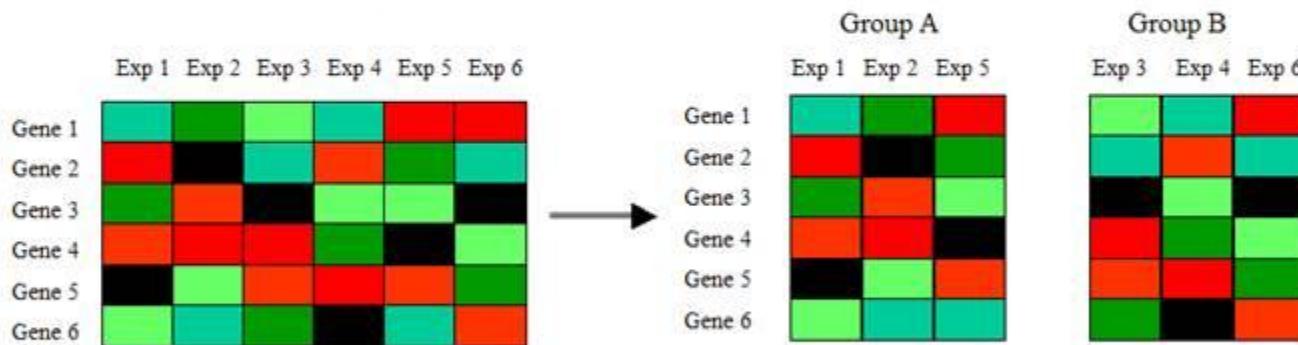


Adapted from "Shared and distinct transcriptional programs underlie the hybrid nature of iNKT cells. Nature Immunology, 2013"

一、基因差异表达

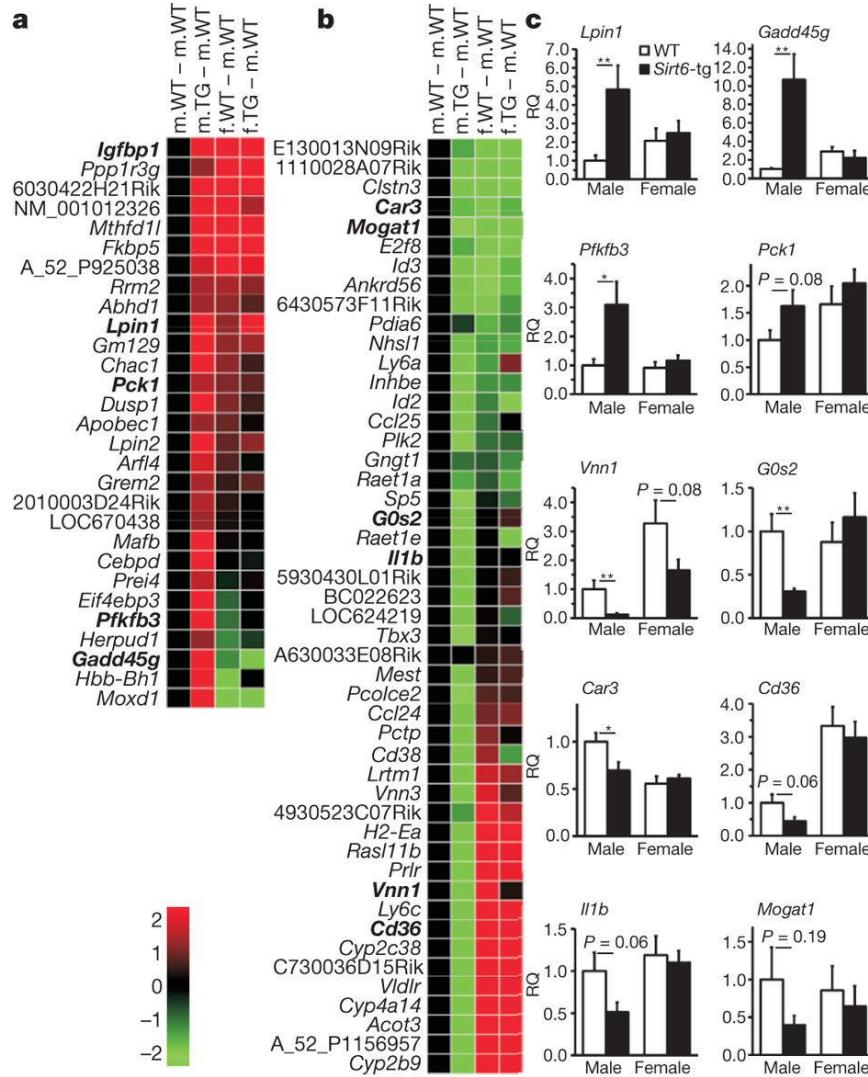
T-Tests (TTEST)

1. Assign experiments to two groups, e.g., in the expression matrix below, assign Experiments 1, 2 and 5 to group A, and experiments 3, 4 and 6 to group B.



2. Question: Is mean expression level of a gene in group A significantly different from mean expression level in group B?

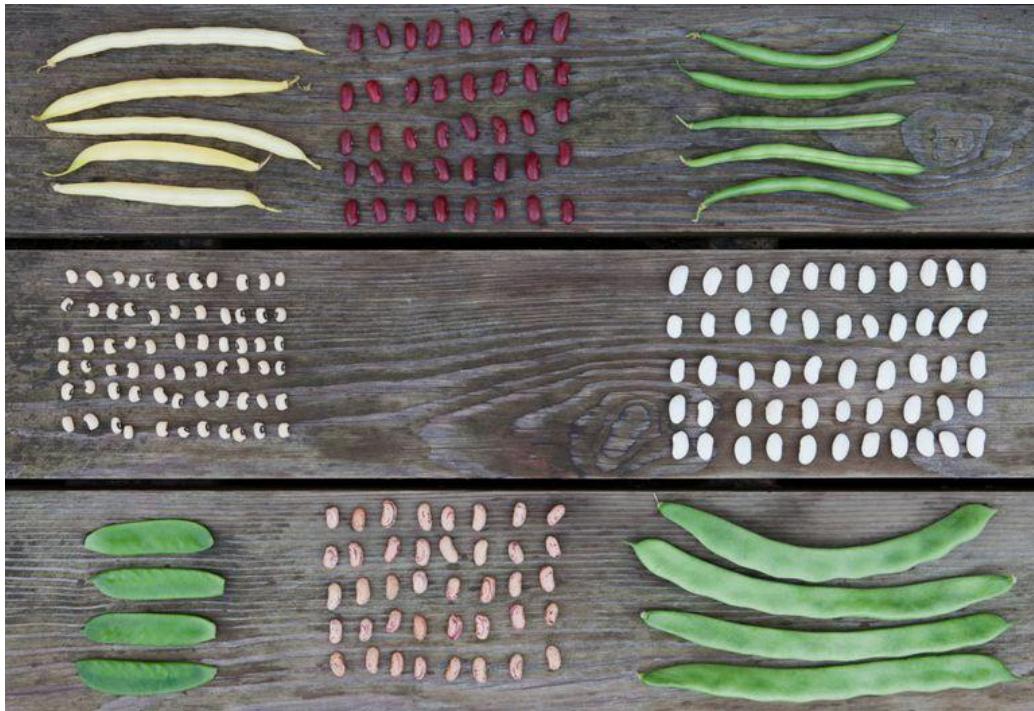
一、基因差异表达



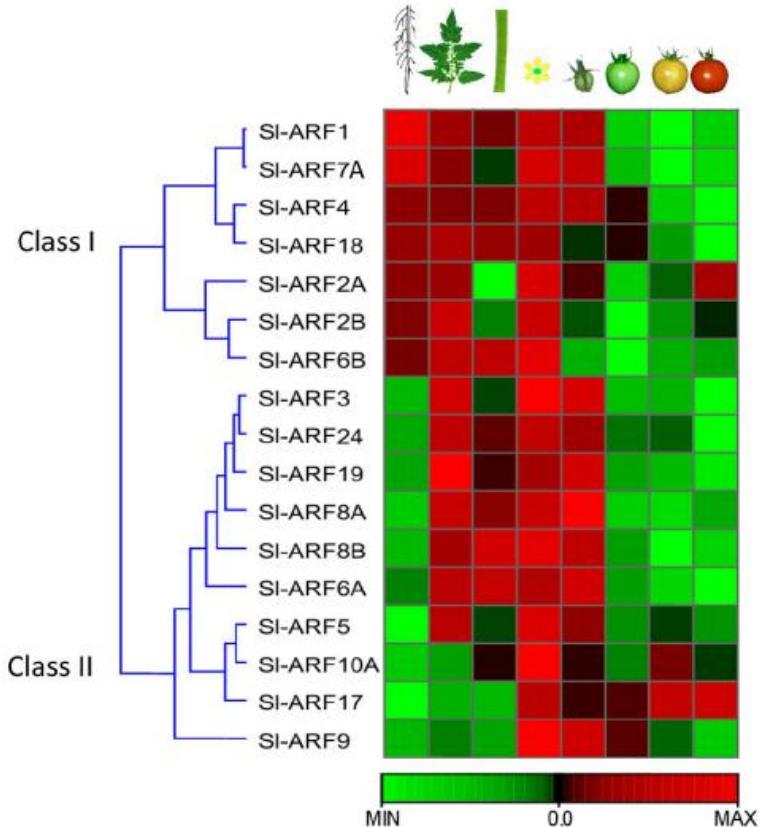
Adapted from "The sirtuin SIRT6 regulates lifespan in male mice. Nature 2012"

二、基因表达和性状关联性分析

Phenotype



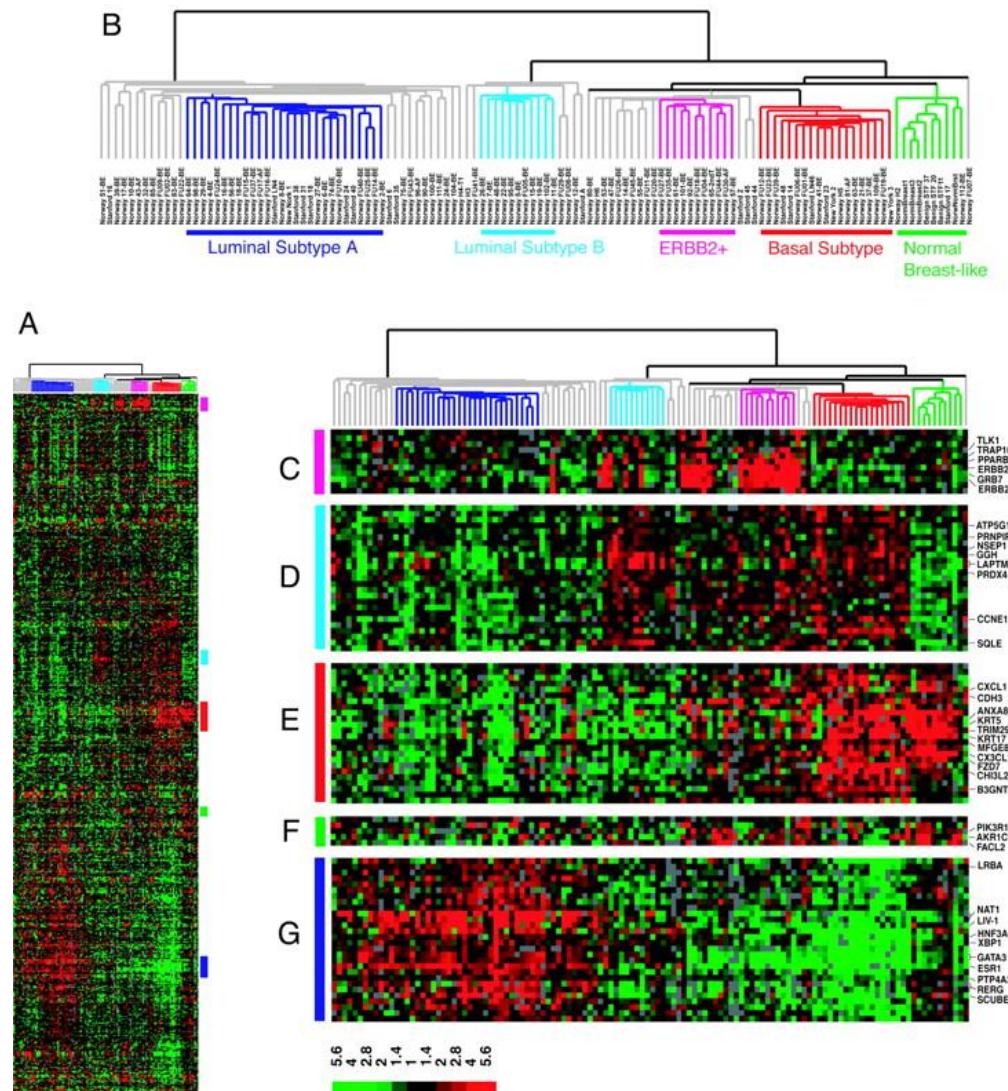
Phenotype& Genotype



Adapted from: <https://www.thoughtco.com/phenotype-373475>

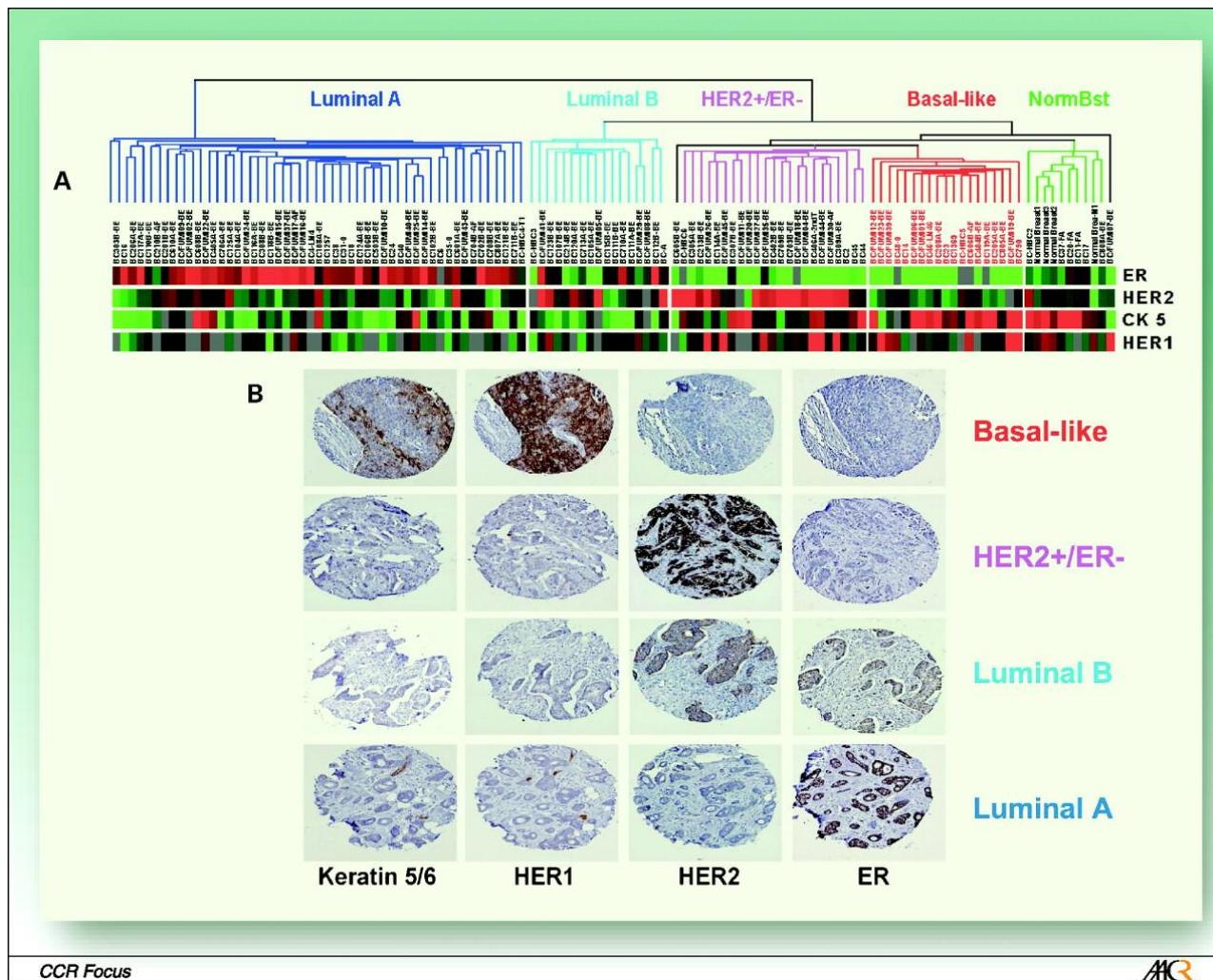
Adapted from "Characterization of the Tomato ARF Gene Family Uncovers a Multi-Levels Post-Transcriptional Regulation Including Alternative Splicing. PLoS ONE, 2014"

三、基因分型



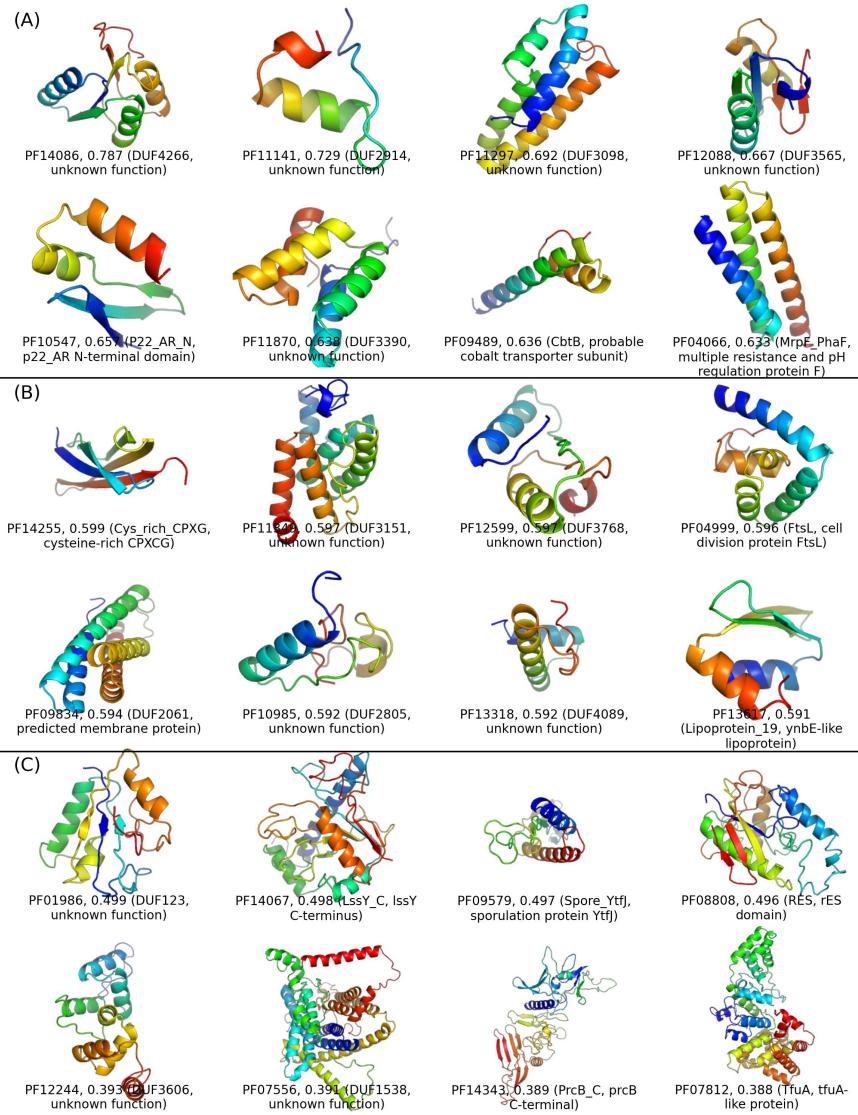
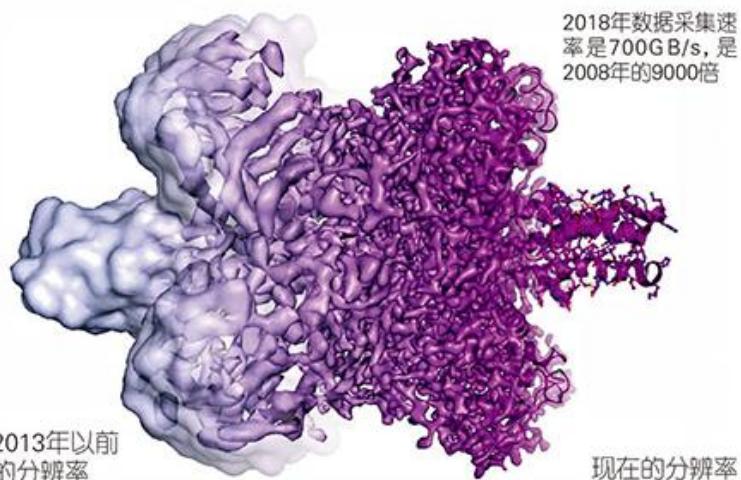
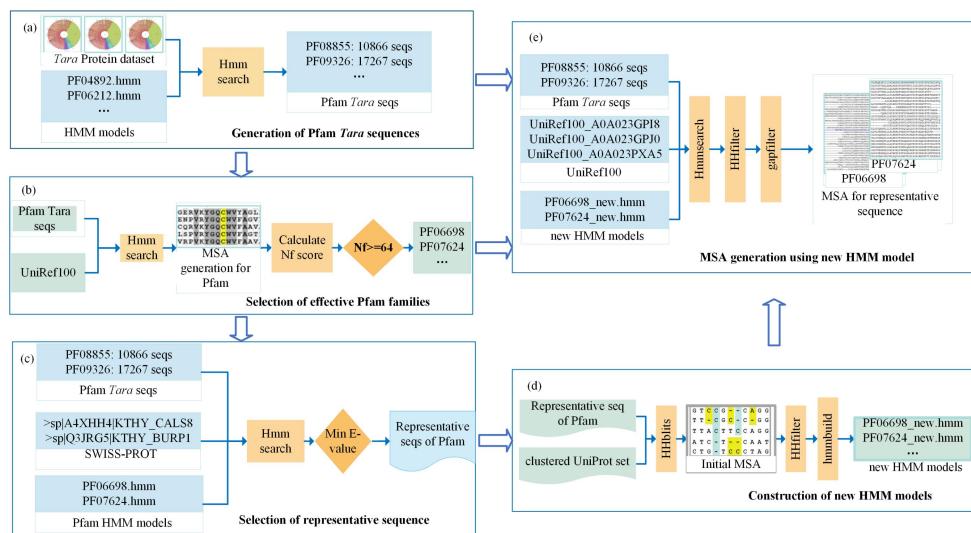
Adapted from "Repeated observation of breast tumor subtypes in independent gene expression data sets. PNAS, 2003"

三、基因分型



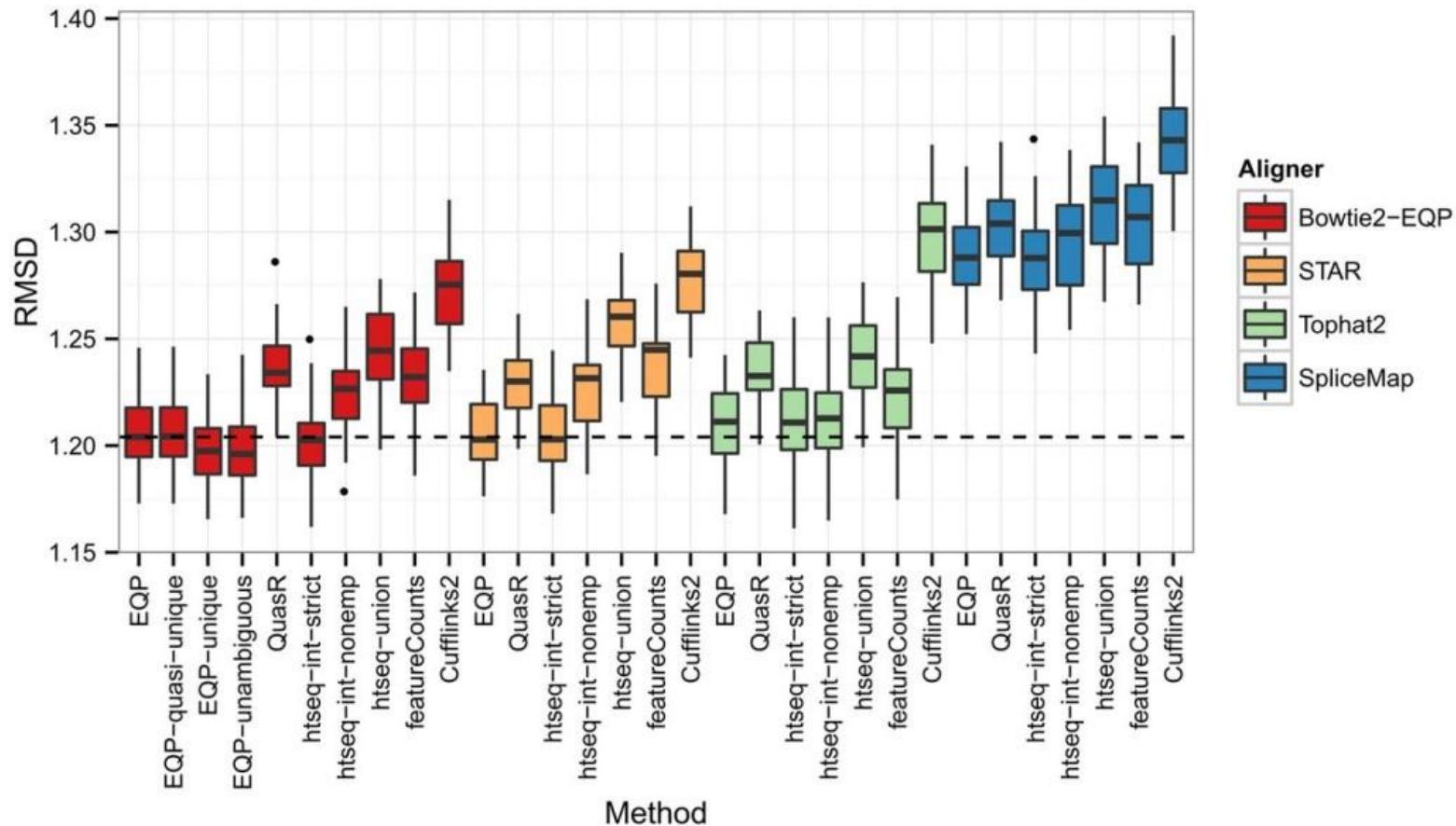
Adapted from "Triple-Negative Breast Cancer: Risk Factors to Potential Targets. Clinical Cancer Research, 2008"

四、结构预测：打分函数



Adapted from “Fueling ab initio folding with marine microbiome enables structure and function predictions of new protein families. Genome Biology, 2019”

五、方法比较

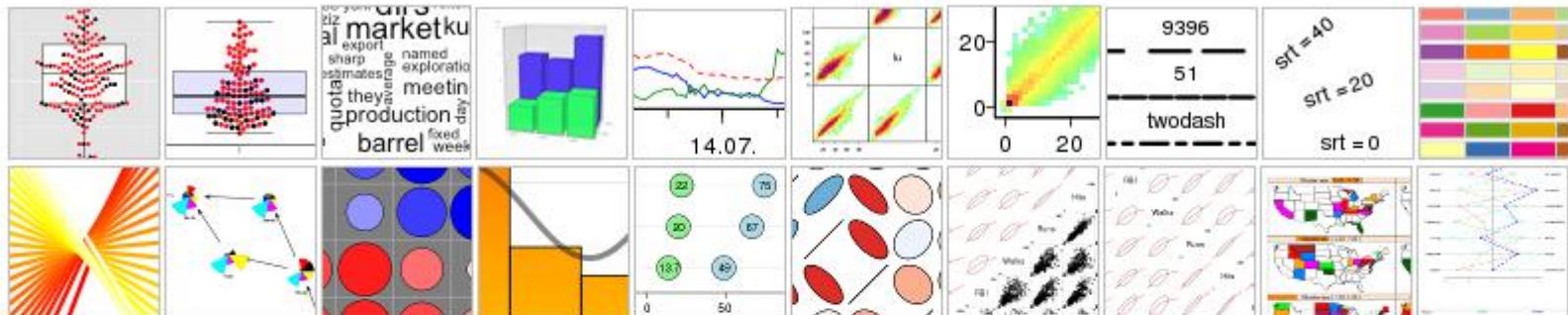


Adapted from “The exon quantification pipeline (EQP): a comprehensive approach to the quantification of gene, exon and junction expression from RNA-seq data. NAR, 2016”

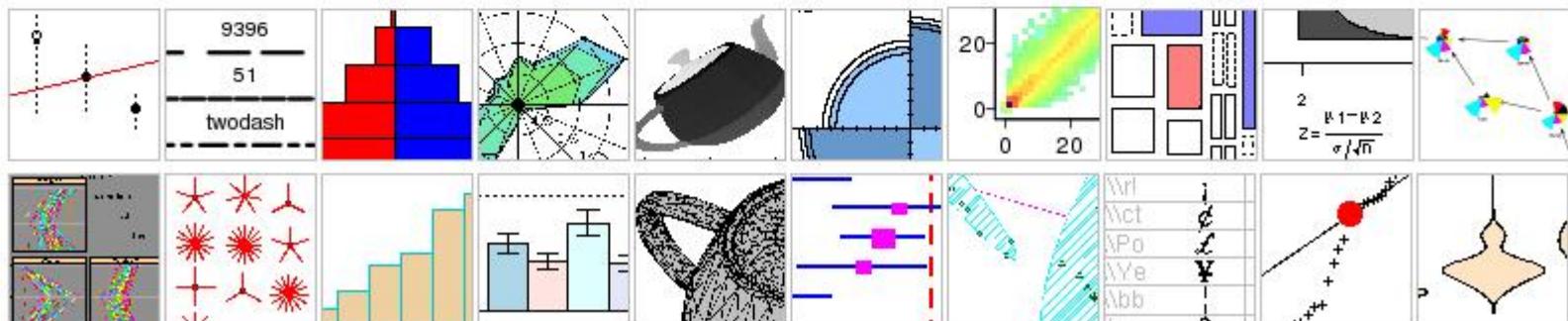
Hand-ons

R: <https://www.r-project.org>

» Last entries ...



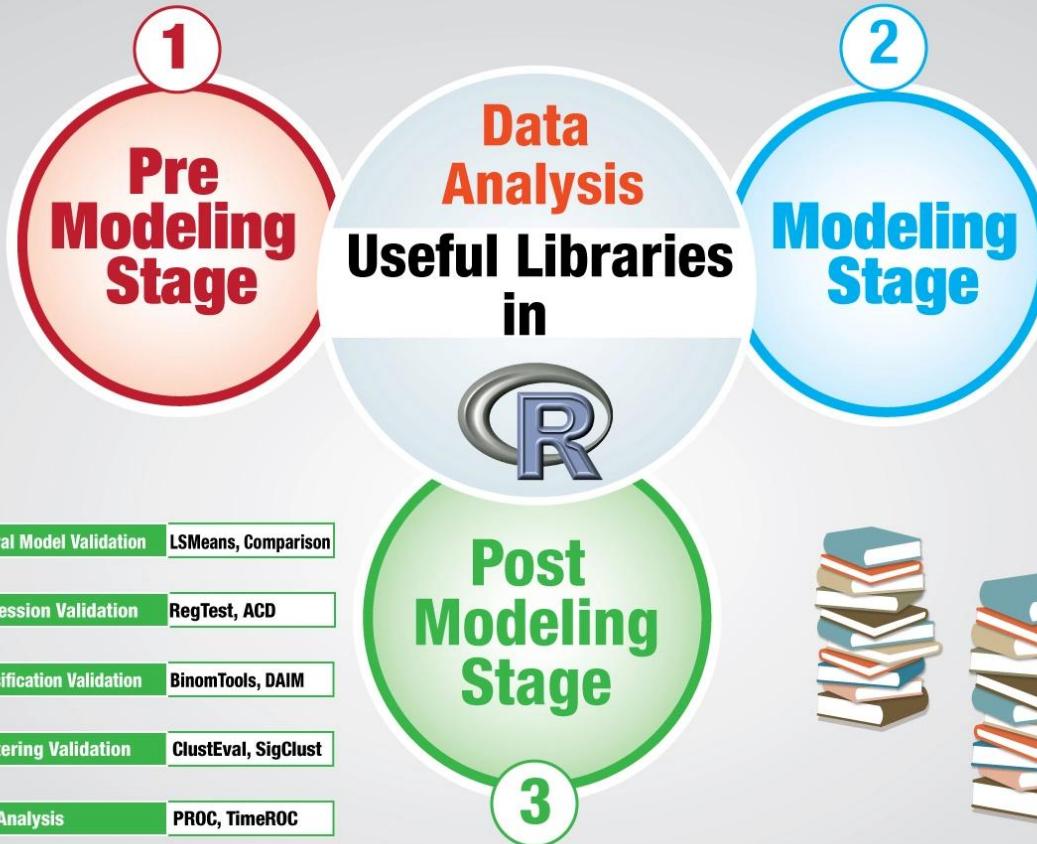
» Random entries



Hand-ons

>install.packages("package name")

- 1. Data Visualization ggplot2, googleVis
- 2. Data Transformation plyr, data.table
- 3. Missing Value Imputations MissForest, MissMDA
- 4. Outlier Detection Outliers, EVIR
- 5. Feature Selection Features, RRF
- 6. Dimension Reduction FactoMineR, CCP



Other Libraries

A. Improve performance Rcpp, parallel

B. Work with web XML, jsonlite, httr

C. Report results shiny, RMarkdown

D. Text Mining tm, twitteR

E. Database sqldf, RODBC, RMongo

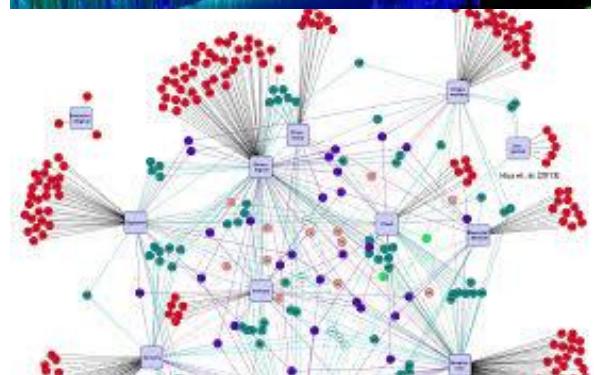
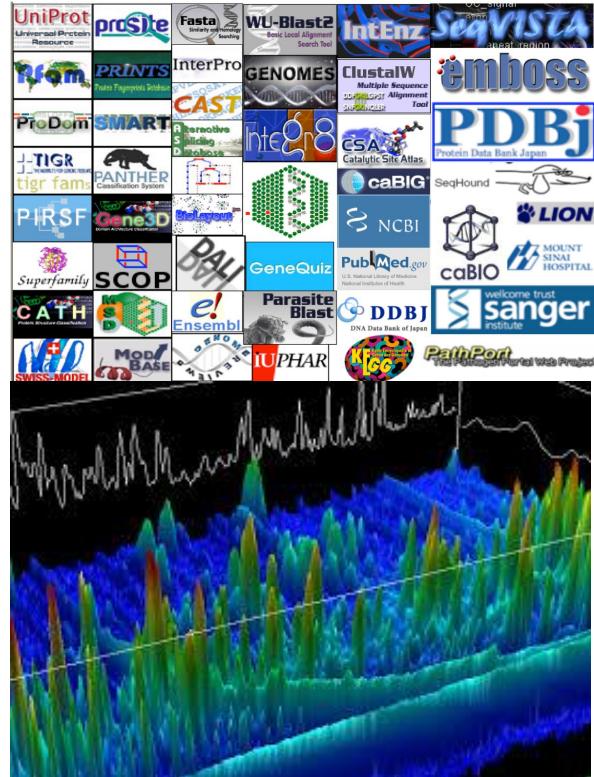
F. Miscellaneous swirl, reshape2, qcc

Hand-ons

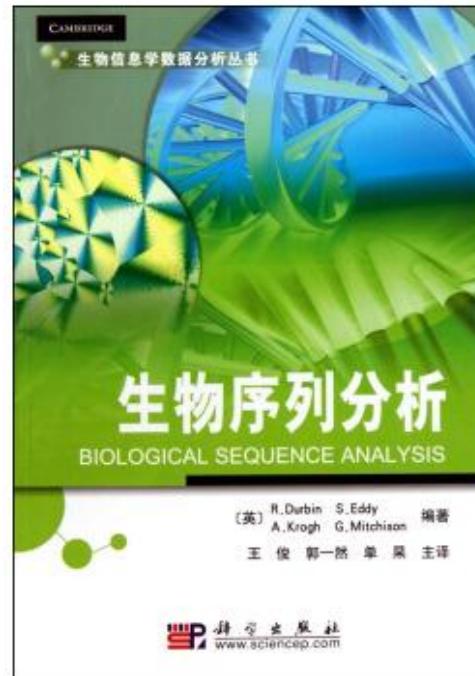
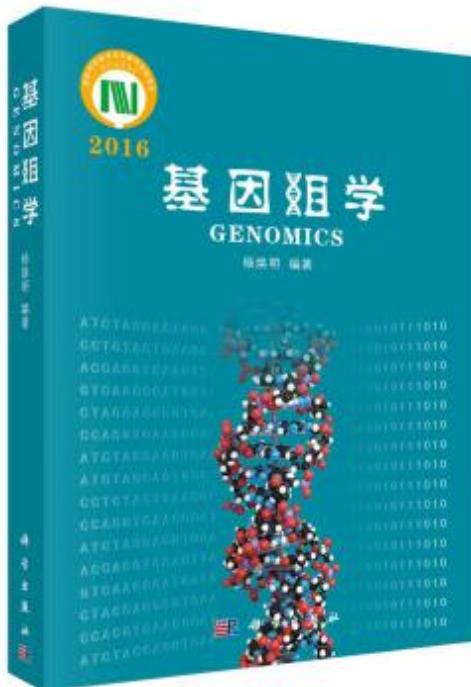
眼见为实！

经典生物信息和生物统计分析平台

- 基因组: [Genome Browser](#)
- 进化树: [iTOL](#)
- 蛋白质组: [Firmiana](#)
- 蛋白质结构: [PDB](#)
- 微生物组: [EBI-Metagenome](#), [NODE-Microbiome](#)



References



References



Slides credits

- 生物信息学研究方法概述：北京大学生物信息中心
- 生物统计学：卜东波@中国科学院计算技术研究所，邓明华@北京大学
- 神经网络与深度学习：邱锡鹏@复旦大学
- Introduction to Computational Biology and Bioinformatics: Xiaole Shirley Liu Lab@Harvard University
- Combinatorial Methods in Computation Biology: Ken Sung Lab@NUS
- Deep Learning in the Life Sciences: MIT
- Probabilistic Graphical Models: Eric Xing@CMU
- Numerous other leading researchers and leading labs.....

补充知识

- 生物统计方法的实例；
- 生物统计方法的实际操作；
- 生物大数据可视化。

1. 生物统计方法的实例

➤ Transcriptomics:

<http://gepia.cancer-pku.cn/> (GEPIA platform)

➤ Proteomics:

<http://tools.proteomecenter.org/wiki/index.php?title=Software:TPP> (TPP pipeline)

2. 生物统计方法的实际操作

- StatPage: <http://statpages.info/>
- Gapminder: <https://www.gapminder.org/>
- MORPHEUS:
<https://software.broadinstitute.org/morpheus/>
- TAIR:
<https://www.arabidopsis.org/portals/expression/microarray/microarraySoftwareV2.jsp>

3. 生物大数据可视化

- R: <https://www.r-graph-gallery.com/>
- Interactive: <https://d3js.org/>
- Illustration: <http://echarts.baidu.com/examples/>

生物统计经典软件

基因组可视化: Genome Browser, (<http://genome.ucsc.edu/>), (tracks, annotations, etc.)

序列保守性: WebLogo, (<http://weblogo.berkeley.edu/logo.cgi>),

基因预测: MEME, (<http://meme-suite.org/>).

进化树: iTOL, (<https://itol.embl.de/>),

基因调控网络: GeneNetwork, (<http://gn2.genenetwork.org/>), Cytoscape, (<https://cytoscape.org/>),

代谢通路: KEGG, (<https://www.kegg.jp/>); iPATH, (<https://pathways.embl.de/>),

蛋白结构与功能: PDB, (<http://www.rcsb.org>); pFAM, (<http://pfam.xfam.org/>),

微生物组: EBI Magnify. (<https://www.ebi.ac.uk/metagenomics/>),

蛋白和小分子互作数据: STITCH, (<http://stitch.embl.de/>); STRING, (<http://string-db.org>),

药物数据库: DrugBank, (<https://www.drugbank.ca/>),

生物数据分析平台: Galaxy, (<https://usegalaxy.org/>),

生物数据可视化: Echart, (<https://www.echartsjs.com/examples/zh/index.html>),

