



统计学原理(Statistic)

胡华平

西北农林科技大学

经济管理学院数量经济教研室

huhuaping01@hotmail.com

2021-05-08

西北农林科技大学

第1章 导论

1.1 为什么学习统计学

1.2 变量和数据

1.3 数据的计量层次

1.4 数据的时间状态

1.5 统计学的体系

1.6 统计分析的基本过程

1.1 为什么学习统计学



辛普森悖论的警示

故事是这么说的：

录用女性的六大部门

gender	申请数	录用率
男性	2590	46%
女性	1835	30%

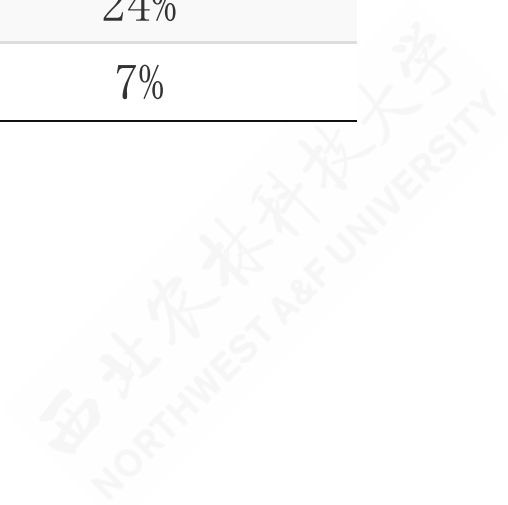


辛普森悖论的警示

但故事背后却另有蹊跷：

聚焦六大招聘部门

部门	男性申请数	男性录用率	女性申请数	女性录用率
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%



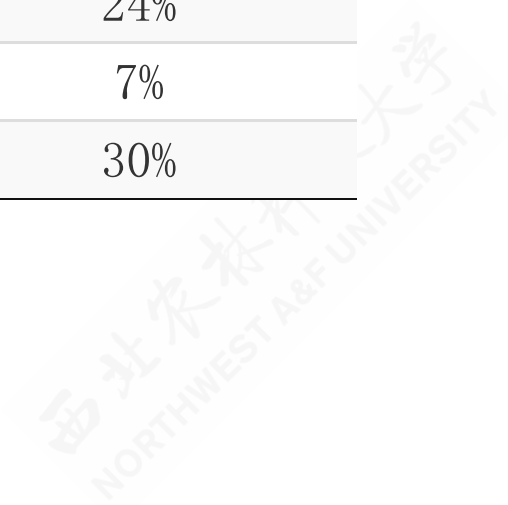


辛普森悖论的警示

事情的“真相”是：

录用女性的六大部门

部门	男性申请数	男性录用率	女性申请数	女性录用率
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%
合计	2590	46%	1835	30%





辛普森悖论的警示

对比一下

男性最喜欢的部门排序

部门	gender	录用率	申请数	录用数
A	男性	62%	825	512
B	男性	63%	560	353
D	男性	33%	417	138
C	男性	37%	325	120
E	男性	28%	191	53
F	男性	6%	272	16

女性最喜欢的部门排序

部门	gender	录用率	申请数	录用数
C	女性	34%	593	202
D	女性	35%	375	131
E	女性	24%	393	94
A	女性	82%	108	89
F	女性	7%	341	24
B	女性	68%	25	17



辛普森悖论的警示

更加细节的数据:

39个部门的录用情况

部门	女性申请数	总录用数	总申请数
1	2	82	202
2	3	76	356
3	4	63	628
4	4	64	158
5	3	59	365

Showing 1 to 5 of 39 entries

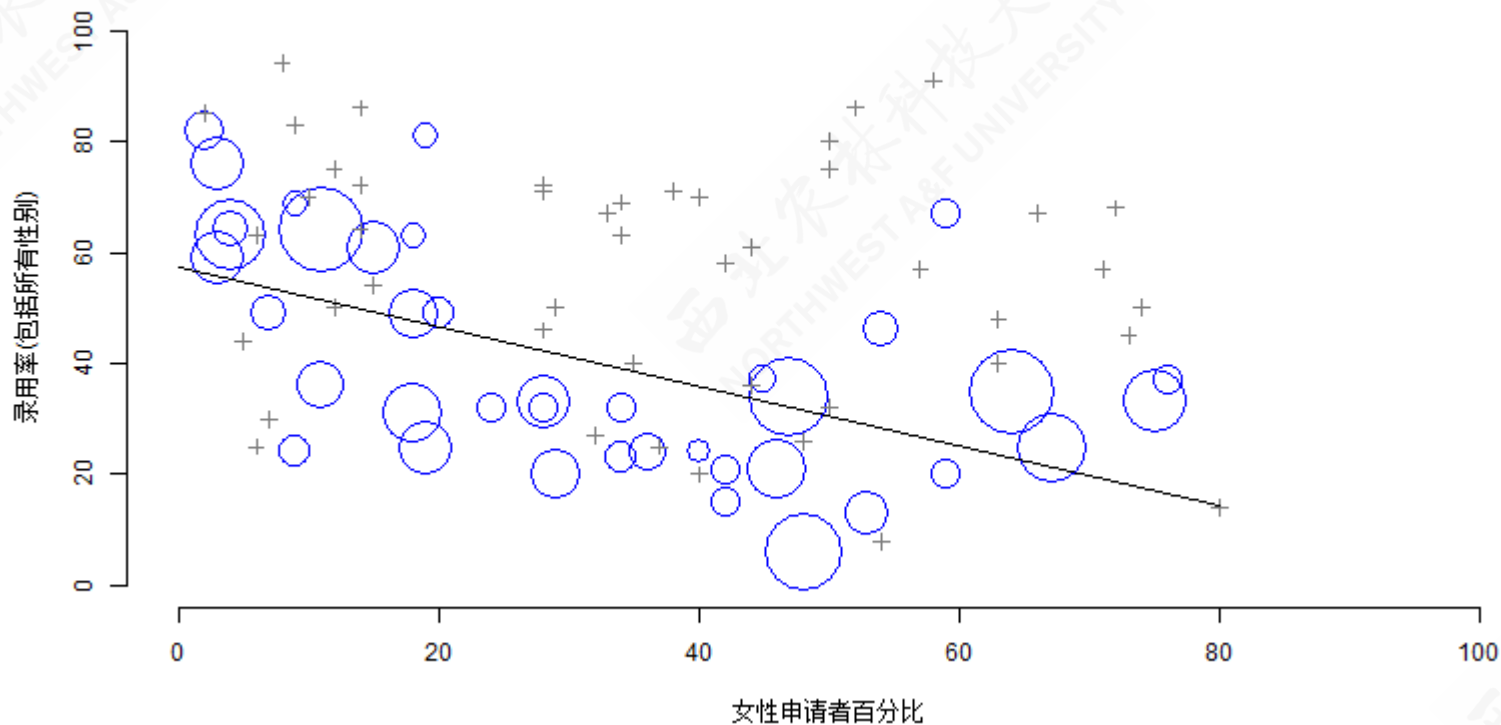
Previous 1 2 3 4 5 ... 8 Next





辛普森悖论的警示

令人吃惊的对称性分布(圆圈表示总申请数大于40人的部门, 叉叉表示总申请数小于40人的部门):





信念偏见诅咒

信念偏见效应：如果你让人们决定一个特定的论点是否在逻辑上是有效的，我们往往会受到结论可信度的影响，即使我们不应该这样做。

一个有效的论据，其结论是可信的：

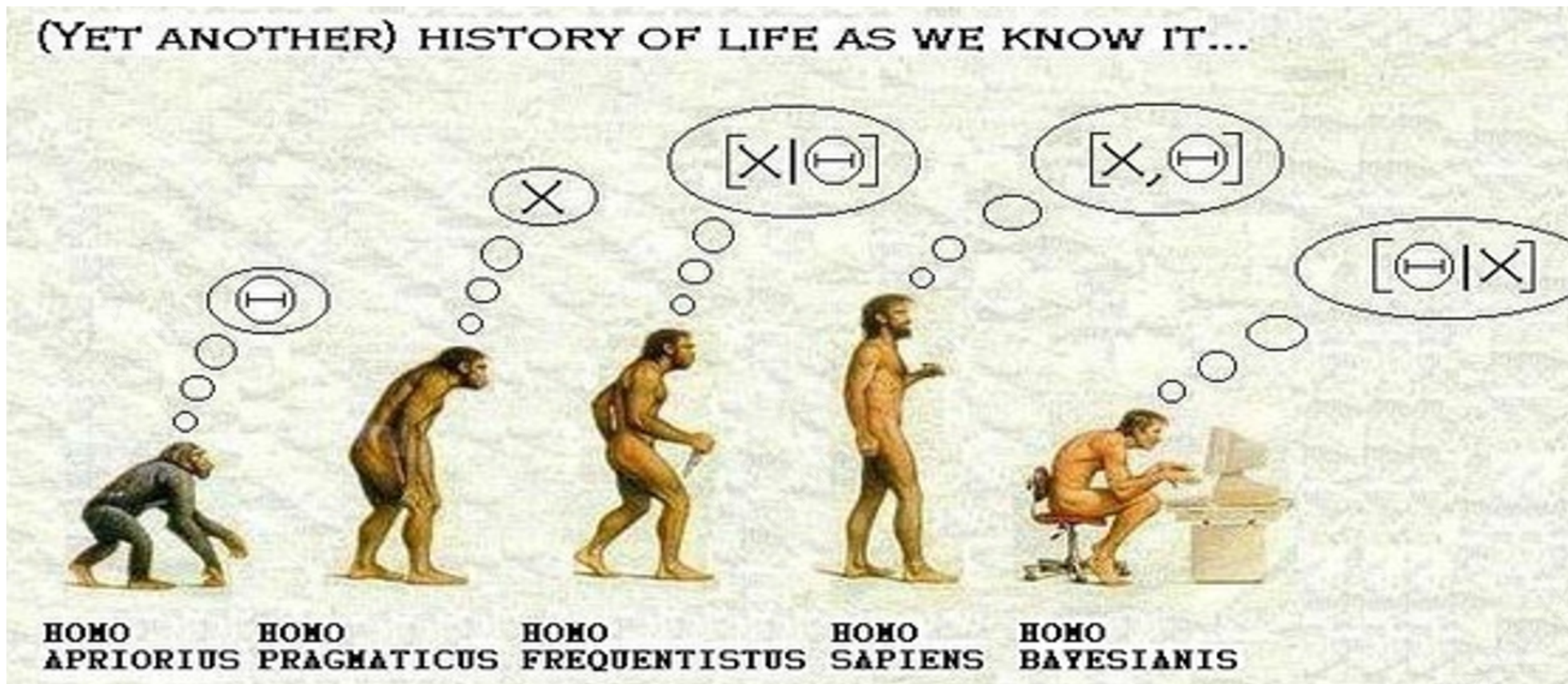
- 没有香烟很便宜（前提1）
- 有些令人上瘾的东西很便宜（前提2）
- 因此，有些令人上瘾的东西不是香烟（结论）

一个有效的论据，但其结论是不可信的：

- 没有令人上瘾的东西很便宜（前提1）
- 有些香烟很便宜（前提2）
- 因此，有些香烟不会上瘾（结论）



信念偏见诅咒



A 先验学派
(Apriorism)

B 实用学派
(Pragmatics)

C 频率学派
(Frequentists)

D 现代学派
(Sapiens)

E 贝叶斯学派
(Bayesians)





信念偏见诅咒：持枪和控枪的美国现象

素材1:

- 美国宪法规定，人民持有和携带武器的权利不受侵犯，这是宪法权利。美国历史上地广人稀，一旦发生暴力事件，警察没办法及时赶到现场，所以美国人民应该有枪支自卫。

观点:

- A. 宪法很大程度考虑到了公民个人私人持枪的选择权
- B. 菜刀在坏人手里会成为凶器，私人持枪以保护自己和身边的人不受侵害也同样重要
- C. 其他观点



信念偏见诅咒：持枪和控枪的美国现象

素材2:

- 1968年美国人口2亿，拥有枪支1.1亿；今天美国人口3.2亿，拥有枪支3.6亿！50年前是两人一支，现在是几乎一人一支。
- 2016年美国枪支产业雇佣了30多万人，对美国经济的贡献是500亿。美国枪支市场巨大，提供了几十万个工作机会，也养活了很多利益集团，他们不答应禁枪等等。

观点:

- A. 军火产业虽然提供了一些就业，但也助长了枪击事件的发生
- B. 军火市场根深蒂固，进而让私人禁枪的提案或修宪变得困难
- C. 其他观点





信念偏见诅咒：持枪和控枪的美国现象

素材3:

- 美国有三亿多支枪，每年被枪杀的人3万多。2014年美国有4万人自杀，其中超过一半选择用枪；70岁以上自杀的老人，选择用枪的比例最高，74%。
- 美国虽然总发生枪击案，但发生的概率仅有千分之0.1。过去23年美国枪杀案的比例整整下降了1倍！

观点:

- A. 枪击事件很大程度上是因为影响恶劣而被社会过度关注和解读
- B. 引发死亡率的主因很多，私人持枪造成的社会伤害并没有愈演愈烈
- C. 其他观点



信念偏见诅咒：持枪和控枪的美国现象

素材4:

- 美国枪支管理法律不是越来越严，而是越来越松。
- 正反观点针锋相对：“美国游泳淹死的人是5倍于被枪走火打死的”。“我亲戚三岁的小姑娘被他爸枪走火打死。如果这个小孩是你的，你愿意她成为你玩枪的社会成本吗？”
- 如果不发生这次拉斯维加斯枪击案，美国又会出台一个放宽枪支管制的议案，要讨论允许私人购买枪支消声器的议案，就是可以公开卖无声枪了。在美国，要自杀的人没枪也会用别的方式死，美国人对死亡的观念：毒药、安乐死与尊严

观点:

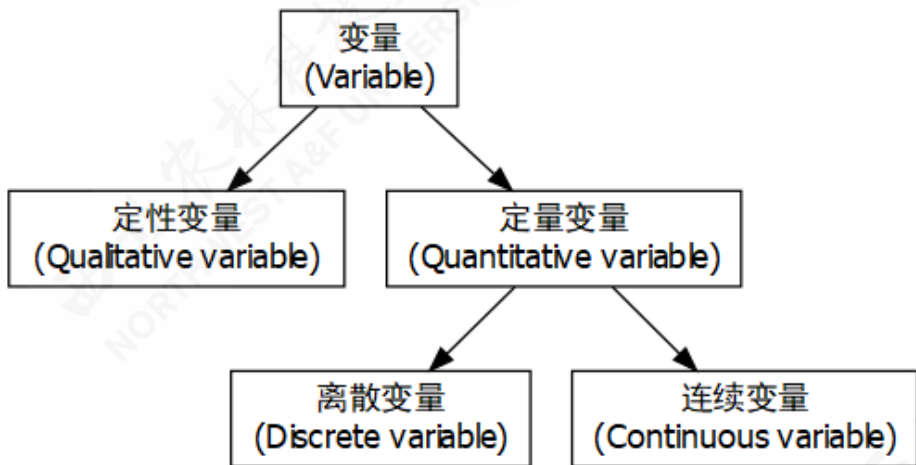
- A. 控枪政策日趋放松尽管有其可解释之处，但并不是长远明智之举
- B. 持枪派，控枪派和中立派或许都无可厚非，各方相互制衡最为重要
- C. 其他观点

1.2 变量和数据



变量及其类型A

变量：描述事物或现象的变化特征。



A. 按性质不同可分为：

- 定性变量：非数值化的变量。
- 定量变量：数值化的变量。
 - 离散变量：可能的取值比较有限、并能轻松列示的一类定量变量。
 - 连续变量：可能的取值较多、以特定微小数值间隔的一类定量变量。



随堂测验 (2min)

请分别说出如下情形分别属于那种变量类型？

- 教育程度（可能取值为：文盲、非文盲）
- 教育程度（可能取值为：小学及以下、初中、高中、大学、研究生及以上）
- 教育程度（可能取值为：1=小学及以下、2=初中、3=高中、4=大学、5=研究生及以上）
- 教育程度（单位为年，可能取值为：6、9、12、15、18、21、24、...）
- 教育程度（单位为年，可能取值为：6、7、8、9、12、14.5、15、18、20、21、...）



变量及其类型B

B. 按变量是否随机:

- 随机变量 (random variable) : 在事件集合上的一个特定函数的随机分布状态。
- 非随机变量 (non-random variable) : 也称为确定性变量 (Deterministic variable) 。

100天降雨量记录 (mm): 下面只列出了前50天

```
day1  day2  day3  day4  day5  day6  day7  day8  day9  day10  day11
28.0  11.5  77.9   3.5   6.5  85.8  23.0  63.3  34.3  22.3  61.2
day12  day13  day14  day15  day16  day17  day18  day19  day20  day21  day22
18.0  20.0   5.5  27.8  89.3  24.9  98.3  35.1  23.6  53.4  10.9
day23  day24  day25  day26  day27  day28  day29  day30  day31  day32  day33
51.3  36.4  31.3  84.3  41.9   7.7  56.9  62.7  21.3  14.8  44.8
day34  day35  day36  day37  day38  day39  day40  day41  day42  day43  day44
43.9  41.1  34.4  27.7   3.1  15.3  19.0  34.7  10.4  63.3  108.4
day45  day46  day47  day48  day49  day50
60.4  56.2  20.1  23.3  39.0   4.2
```



课堂讨论 (2min) :

非随机变量 (non-random variable) 实际存在么?

正方: 存在!

- 有随机变量就自然有非随机变量, 它取值的产生是经过特意安排的, 也即非随机的。

反方: 不存在!

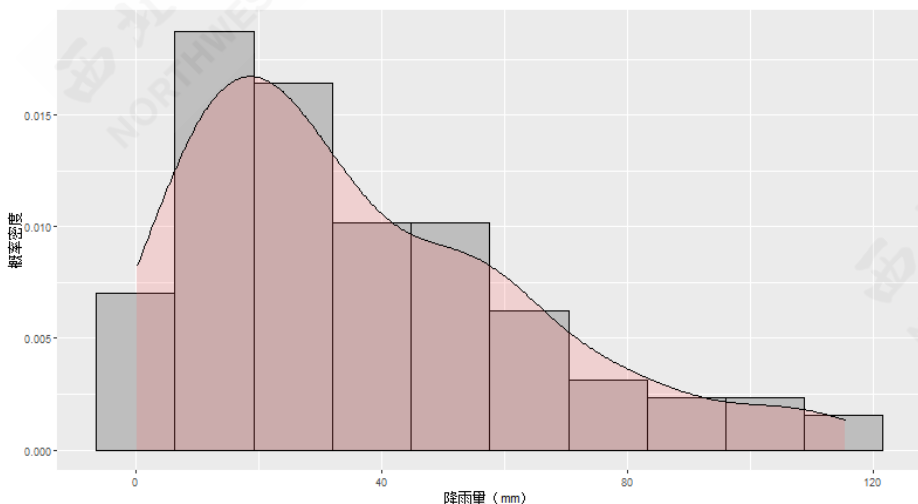
- 非随机变量是杜撰的, 如果是变量则都应该是随机的。非随机变量其实就是常数 (constant), 当然不能称为变量, 也就不能叫做非随机变量。



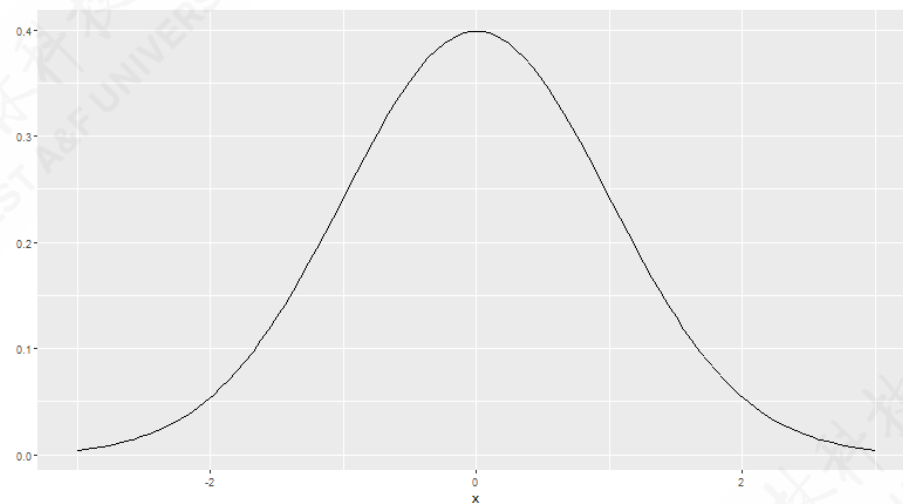
变量及其类型C

C. 按变量是否抽象化:

- 经验变量: 经验变量所描述的是我们周围可以观察到的事物。
- 理论变量: 理论变量则是由统计学家用数学方法所构造出来的一些变量, 比如z 统计量、t 统计量、 χ^2 统计量、F 统计量等。



降雨量记录数据的分布图 $n=100$

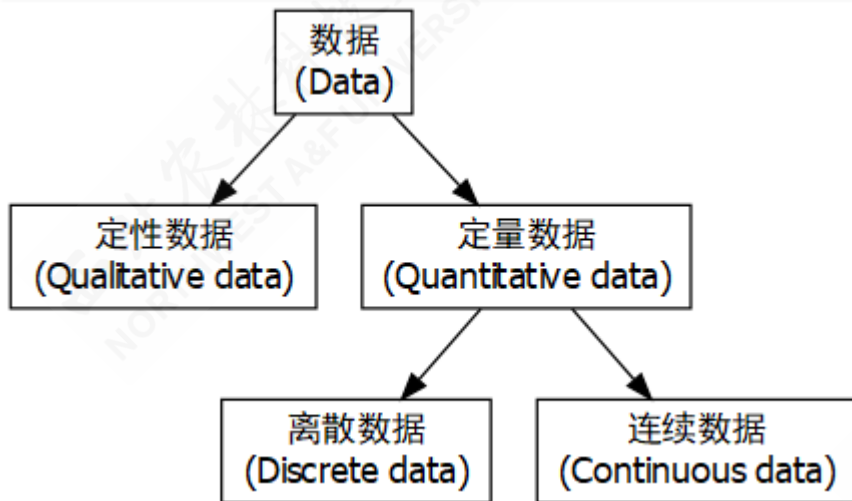


理论正态分布图



数据及其类型

数据：变量的实际具体取值组成的集合。



A. 按数据取值的性质分：

- 定性数据：定性变量的取值。
- 定量数据：定量变量的取值。
 - 离散数据：离散变量的取值。
 - 连续数据：连续变量的取值。



观察性数据和实验性数据

观察性数据 (observational data)：人们在观察性研究中被被动性记录下来的数据。

- 选择性偏差难以避免 Bias in Observational Studies - Sensitivity Analysis with R package episensr

(<https://homerhanumat.github.io/elemStats/design-of-studies.html>)

实验性数据 (experimental data)：人们通过主动控制和改变实验条件获得并记录的数据。



数据与变量的关系

- 变量是数据存身之所，而数据是变量的现实表达。
- 变量是一个抽象概念，而数据是实体概念。
- 数据和变量是相互依存的关系，如同一个硬币的两面。
- 数据和变量可以进一步共同描述或表达事物或现象的信息（information）。



SPSS统计软件下的变量和数据

下面给大家展示传统商业（收费）统计分析软件SPSS的变量和数据视窗：

西北农林科技大学
NORTHWEST A&F UNIVERSITY

西北农林科技大学
NORTHWEST A&F UNIVERSITY

西北农林科技大学
NORTHWEST A&F UNIVERSITY



SPSS软件的数据视图

可见: 55变量的 55

	问卷编码	样本编码	省份	县(乡、镇)	村庄名称	户主姓名	姓名1	与受访者关系1	性别1	年龄1	民族1	民族分组	教育水平
1	TN110056261	N1901-10	广东	惠东县多祝镇	上村村	曾新平	曾新平		男性	48	汉族	汉族	
2	TN110056262	N1901-13	广东	惠东县多祝镇	上村村	曾伟文	曾伟文		男性	53	汉族	汉族	
3	TN110056263	N1901-17	广东	惠东县多祝镇	上村村	曾秋生	曾秋生		男性	43	汉族	汉族	
4	TN110056264	N1510-1	山东	东平县接山镇	夏谢五村	徐玉平	尹燕兰		女性	54	汉族	汉族	
5	TN110056266	N1809-1	湖南	永顺县勺哈乡	基湖村	杨永久	杨永久		男性	49	汉族	汉族	
6	TN110056269	N1910-8	广东	郁南县平台镇	石台村	岑敬判	岑敬判		男性	54	汉族	汉族	
7	TN110056270	N1910-3	广东	郁南县平台镇	石台村	刘耀魁	刘耀魁		男性	43	汉族	汉族	
8	TN110056272	N1513-8	山东	沾化县大高乡	薛家村	薛常印	薛常印		男性	52	汉族	汉族	
9	TN110056273	N1513-4	山东	沾化县大高乡	薛家村	薛庆海	薛庆海		男性	73	汉族	汉族	
10	TN110056274	N1513-13	山东	沾化县大高乡	薛家村	薛呈收	薛呈收		男性	55	汉族	汉族	
11	TN110056275	N1513-11	山东	沾化县大高乡	薛家村	薛呈才	薛呈才		男性	54	汉族	汉族	
12	TN110056276	N1513-14	山东	沾化县大高乡	薛家村	薛建国	薛建国		男性	55	汉族	汉族	
13	TN110056277	N1513-16	山东	沾化县大高乡	薛家村	薛玉明	薛玉明		男性	77	汉族	汉族	
14	TN110056278	N1513-3	山东	沾化县大高乡	薛家村	薛贵臣	薛贵臣		男性	61	汉族	汉族	
15	TN110056279	N1513-7	山东	沾化县大高乡	薛家村	薛玉柱	薛玉柱		男性	75	汉族	汉族	
16	TN110056280	N1513-17	山东	沾化县大高乡	薛家村	薛玉喜	薛玉喜		男性	49	汉族	汉族	
17	TN110056281	N1513-6	山东	沾化县大高乡	薛家村	薛新贵	薛新贵		男性	44	汉族	汉族	
18	TN110056282	N1513-12	山东	沾化县大高乡	薛家村	薛建军	薛建军		男性	47	汉族	汉族	
19	TN110056283	N1513-9	山东	沾化县大高乡	薛家村	薛玉增	薛玉增		男性	61	汉族	汉族	
20	TN110056284	N1513-10	山东	沾化县大高乡	薛家村	李花叶	李花叶		女性	53	汉族	汉族	
21	TN110056285	N1513-5	山东	沾化县大高乡	薛家村	薛呈文	薛呈文		男性	64	汉族	汉族	
22	TN110056286	N1513-2	山东	沾化县大高乡	薛家村	薛荣举	薛荣举		男性	71	汉族	汉族	
23	TN110056289	N1815-1	湖南	保靖县阳朝乡	溪州村	黄立文	黄立文		男性	74	土家族	少数民族	
24	TN110056290	N1815-3	湖南	保靖县阳朝乡	溪州村	向明莲	向明莲		女性	60	土家族	少数民族	
25	TN110056300	N1727-1	湖北	大冶市陈贵镇	江添爱村	廖旺宝	廖旺宝		女性	69	汉族	汉族	
26	TN110056311	N1811-1	湖南	衡山县白果镇	绍庄村	李海贤	李思		男性	24	汉族	汉族	
27	TN110056317	N2803-1	甘肃	泾川县丰台乡	丰台村	吕爱子	吕爱子		女性	43	汉族	汉族	
28	TN110056318	N1815-4	湖南	保靖县阳朝乡	溪州村	黄召权	向云		女性	45	土家族	少数民族	
29	TN110056322	N0901-4	上海	崇明县建设镇	浜西村	胡学祥	胡学祥		男性	77	汉族	汉族	
30	TN110056323	N0901-1	上海	崇明县建设镇	浜西村	施惠福	施惠福		女性	67	汉族	汉族	



SPSS软件的变量视图

名称	类型	宽度	小数	标签	值	缺失	列	对齐	度量	角色	
1	问卷编码	字符串	33	0		无	无	12	左	名义(N)	输入
2	样本编码	字符串	27	0		无	无	12	左	名义(N)	输入
3	省份	字符串	18	0		无	无	12	左	名义(N)	输入
4	县(乡、镇)	字符串	72	0		无	无	12	左	名义(N)	输入
5	村庄名称	字符串	48	0		无	无	12	左	名义(N)	输入
6	户主姓名	字符串	66	0		无	无	12	左	名义(N)	输入
7	姓名1	字符串	72	0		无	无	12	左	名义(N)	输入
8	与受访者关...	字符串	66	0		{1, 夫妻关系...	无	12	左	名义(N)	输入
9	性别1	数值	11	0		{1, 男性}...	无	12	右	名义(N)	输入
10	年龄1	数值	11	0		无	无	12	右	度量	输入
11	民族1	字符串	24	0		无	无	12	左	度量	输入
12	民族分组	数值	11	0		{1, 汉族}...	无	12	右	有序(O)	输入
13	教育水平1	数值	11	0		无	无	12	右	名义(N)	输入
14	户口类型1	数值	11	0		{1, 农业户口...	无	12	右	名义(N)	输入
15	婚否1	数值	11	0		{1, 未婚}...	无	12	右	名义(N)	输入
16	职业1	数值	11	0		{1, 农业劳动...	无	12	右	名义(N)	输入
17	是否干部1	数值	11	0	是否干部 1	{1, 是}...	无	12	右	名义(N)	输入
18	政治面貌1	数值	11	0		{1, 党员}...	无	12	右	名义(N)	输入
19	健康状况1	数值	11	0		{1, 优}...	无	12	右	名义(N)	输入
20	宗教信仰1	数值	11	0		{1, 没有宗教...	无	12	右	名义(N)	输入
21	打工者外出...	数值	11	0		无	无	12	右	度量	输入
22	打工者打工...	数值	11	0		{1, 本村(社...	无	12	右	名义(N)	输入
23	打工者行业1	数值	11	0		{1, 制造业}...	无	12	右	名义(N)	输入
24	工种1	数值	11	0		{1, 体力劳动...	无	12	右	名义(N)	输入
25	打工截止年...	字符串	66	0		无	无	12	左	名义(N)	输入
26	打工是否寄...	字符串	6	0		{1, 是}...	无	12	左	名义(N)	输入
27	承包地总面积	字符串	66	0		无	无	12	左	名义(N)	输入
28	水旱承包地...	数值	11	0		无	无	12	右	度量	输入
29	水田面积	数值	11	0		无	无	12	右	度量	输入
30	旱地面积	数值	11	0		无	无	12	右	度量	输入
31	草地(场)...	数值	11	0		无	无	12	右	度量	输入



R统计软件下的变量和数据

下面给大家展示开源（免费）统计分析软件R的变量和数据视窗：

西北农林科技大学
NORTHWEST A&F UNIVERSITY

西北农林科技大学
NORTHWEST A&F UNIVERSITY

西北农林科技大学
NORTHWEST A&F UNIVERSITY



纽约机场数据库：变量视图

```
require("nycflights13")  
data(flights) # 导  
flights_jan <- flights %>%  
  filter(month == 1) # 只
```

```
Rows: 27,004  
Columns: 19  
$ year      <int> 2013, 2013, 2013, ...  
$ month     <int> 1, 1, 1, 1, 1, ...  
$ day       <int> 1, 1, 1, 1, 1, ...  
$ dep_time  <int> 517, 533, 542, ...  
$ sched_dep_time <int> 515, 529, 540, ...  
$ dep_delay <dbl> 2, 4, 2, -1, -6, ...  
$ arr_time  <int> 830, 850, 923, ...  
$ sched_arr_time <int> 819, 830, 850, ...  
$ arr_delay <dbl> 11, 20, 33, -18, ...  
$ carrier   <chr> "UA", "UA", "AA", ...  
$ flight    <int> 1545, 1714, 114, ...  
$ tailnum   <chr> "N14228", "N242, ...  
$ origin    <chr> "EWR", "LGA", "L, ...  
$ dest      <chr> "IAH", "IAH", "M, ...  
$ air_time  <dbl> 227, 227, 160, ...  
$ distance  <dbl> 1400, 1416, 1089, ...  
$ hour      <dbl> 5, 5, 5, 5, 6, ...  
$ minute    <dbl> 15, 29, 40, 45, ...  
$ time_hour <dtm> 2013-01-01 05:00, ...
```



纽约机场数据库：数据视图

只看2013年1月数据集的前500条数据。

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched
2013	1	1	517	515	2	830	
2013	1	1	533	529	4	850	
2013	1	1	542	540	2	923	
2013	1	1	544	545	-1	1004	

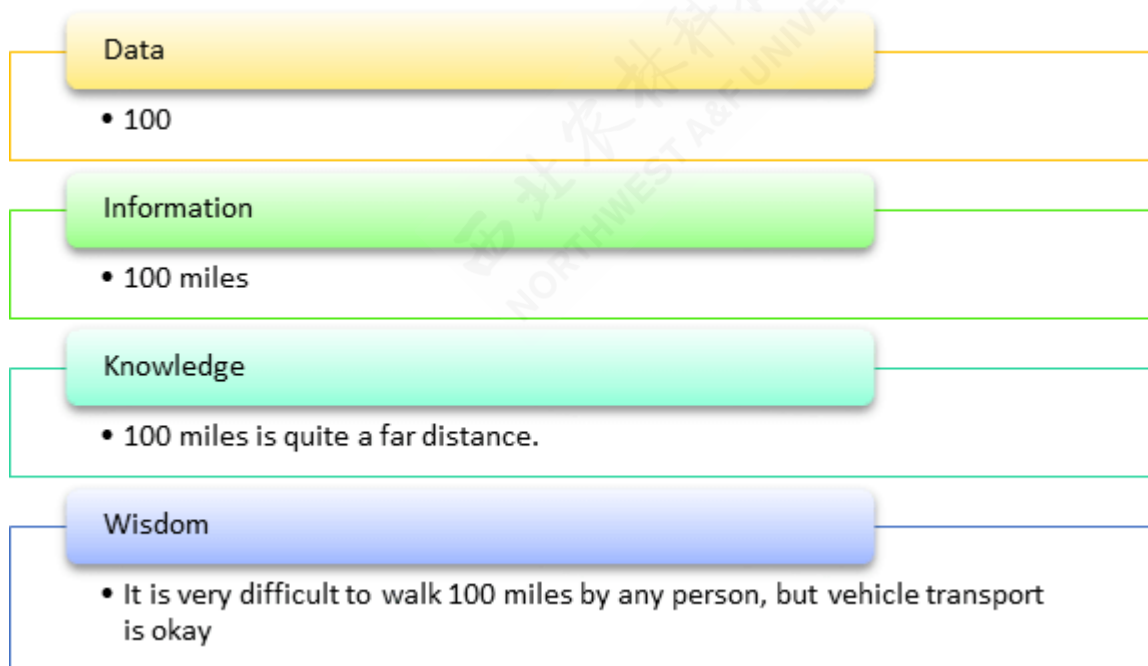




数据和信息的关系

数据和信息有何联系与不同？

“The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning.” —Statistician Nate Silver in the book *The Signal and the Noise*





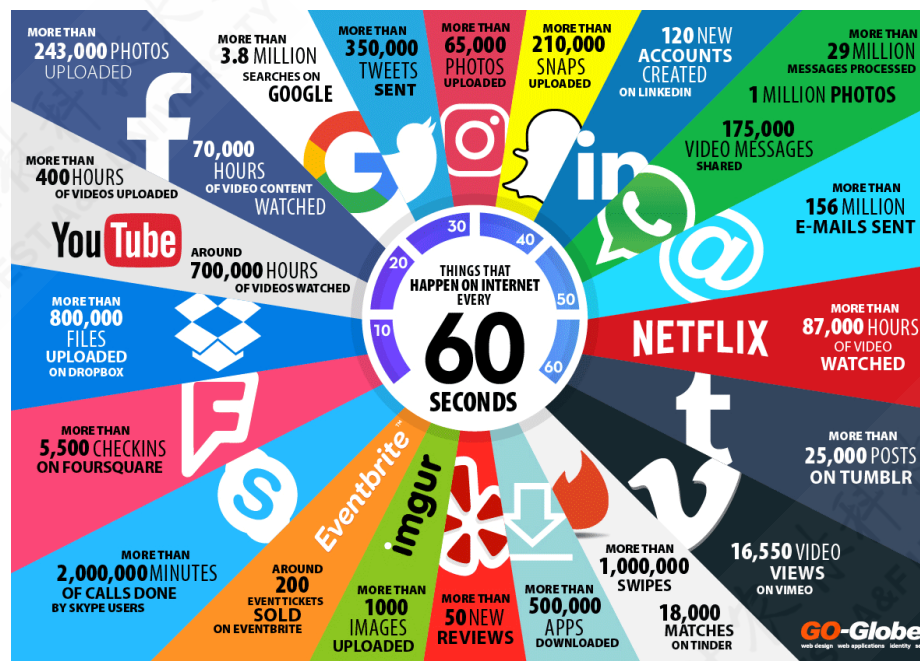
数据与信息的关系

大数据 (big data) 是不是一定好?

- 优点: 可能包含更多信息;
- 缺点: 信息量太大以至于无法处理。

大数据的四大特征 (4V) :

- Volume (数据量大)
- Variety (变异性大)
- Veracity (精确记录)
- Velocity (迅疾变化)





数据与信息的关系

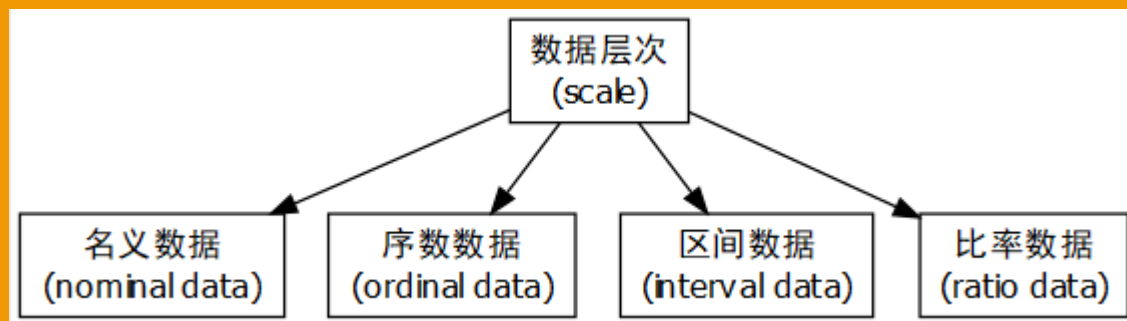
- 天体运行规律：第谷的海量天文数据VS开普勒筛选分析。



- 总统选举预测：《文学文摘》240万人大调查VS盖洛普5000人调查。



1.3 数据的计量层次





名义数据 (nominal data)

- 取值只用于区分所属类别的定性数据。
- 对事物进行分类的结果，数据表现为类别，一般用文字来表述。
 - 性别(男、女)
 - 婚姻状况(已婚、未婚、离婚、分居)



(示例) 名义数据举例

变量及可能取值VS数据集

What is your gender?

- M - Male
- F - Female

What is your hair color?

- 1 - Brown
- 2 - Black
- 3 - Blonde
- 4 - Gray
- 5 - Other

Where do you live?

- A - North of the equator
- B - South of the equator
- C - Neither: In the international space station

name	gender	color	area
tom	M	brown	North
jenny	F	blonde	south
lilei	M	black	North
hanmeimei	F	black	Neither



顺序数据(ordinal data)

- 变量的不同取值有顺序差异，即存在自然顺序
- 变量的不同取值的差值，没有实际意义
- 常用的五分量表

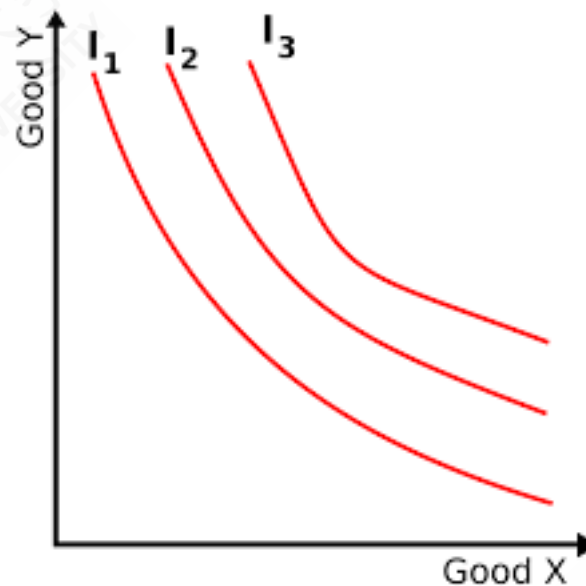
How do you feel today?

- 1 - Very Unhappy
- 2 - Unhappy
- 3 - OK
- 4 - Happy
- 5 - Very Happy

How satisfied are you with our service?

- 1 - Very Unsatisfied
- 2 - Somewhat Unsatisfied
- 3 - Neutral
- 4 - Somewhat Satisfied
- 5 - Very Satisfied

- 经济学中的无差异曲线





(示例) 顺序数据举例

原始数据 (按工作日)

weekday	temp	feel
Mon	High	VU
Tue	Low	VH
Wen	High	UH
Thu	Low	OK
Fri	Medium	H

排序后数据 (按feel)

weekday	temp	feel
Tue	Low	VH
Fri	Medium	H
Thu	Low	OK
Wen	High	UH
Mon	High	VU

课堂思考：序数数据如何排序？如果数据量超级大，怎样快速排序（数据概览如下）：

```

Rows: 5
Columns: 3
$ weekday <chr> "Mon", "Tue", "Wen", "Thu", "Fri"
$ temp      <fct> High, Low, High, Low, Medium
$ feel      <fct> VU, VH, UH, OK, H

```

说明：H表示Happy；V表示very；U表示unhappy。三个字母可以组合表达不同含义。



区间数据 (interval data) :

区间数据 (interval data) 的数值存在自然顺序、可以比较大小 (加减)、但乘除比率没有意义。

- 变量的不同取值有顺序差异, 即存在自然顺序
- 变量的不同取值的差值, 也具有实际意义
- 但不同取值的比率是没有实际意义的
 - 两个时期之内的距离 (如2000 - 1995) 是有意义的, 但两个时期的比率 (2000/1995) 就没有什么意义。
 - 2013年8月11日上午11点天气预报说杨凌的温度是华氏60度, 而长沙达到华氏90度。长沙比杨凌温度高30华氏度 (90-60), 是可以的。但说长沙比杨凌暖和1.5倍 (90/60), 是没有意义。



(示例) 区间数据举例：起飞延误时长与星期几有关系么？

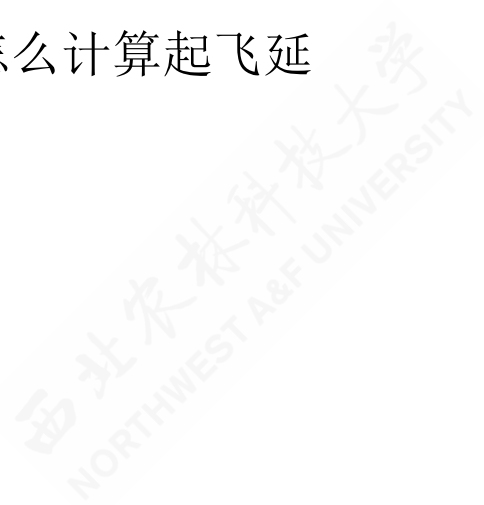
2013年1月JKF机场起飞情况 (原始)

day	wd	dep_time	sched_dep_time
1	周二	542	540
1	周二	544	545
1	周一	557	600
1	周二	558	600
1	周二	558	600
1	周二	558	600
1	周二	559	559

Showing 1 to 7 of 9,161 entries

纽约JKF机场在2013年1月起飞航班数据集：

- 计划起飞时间 (`sched_dep_time`) 和实际起飞时间 (`dep_time`) 能直接相减么？
- 如果延误超过一天，该怎么计算起飞延误时长？
- 如果取消航班了，该怎么计算起飞延误时长？





(示例) 区间数据举例：起飞延误时长与星期几有关系么？

2013年1月1日机场起飞情况 (计算并分类)

day	wd	dep_time	sched_dep_time	dep_delay	delay_cat
1	周二	542	540	2	延误0-1小时内
1	周二	544	545	-1	提前起飞
1	周二	557	600	-3	提前起飞
1	周二	558	600	-2	提前起飞
1	周二	558	600	-2	提前起飞
1	周二	558	600	-2	提前起飞
1	周二	559	559	0	提前起飞

Showing 1 to 7 of 9,161 entries

Previous 1 2 3 4 5 ... 1309 Next



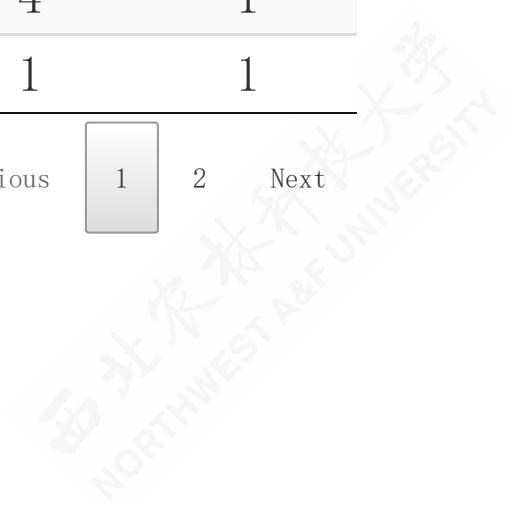


(示例) 区间数据举例：起飞延误时长与星期几有关系么？

delay_cat	周日	周一	周二	周三	周四	周五	周六
合计	1182	1213	1427	1460	1529	1227	1123
提前起飞	772	800	1043	909	972	768	703
延误0-1小时内	320	333	333	397	453	371	364
延误1-2小时内	50	51	34	57	65	46	37
延误2-3小时内	17	10	10	32	20	23	9
延误3-4小时内	8	5	1	16	4	2	1
延误4-5小时内		4	1	1	5	4	1
延误5-6小时内	1	1		2		1	1

Showing 1 to 8 of 11 entries

Previous 1 2 Next





比率数据(ratio data)

比率数据(ratio data)的取值存在自然顺序、可以比较大小（加减）、乘除比率有实际意义。

- 变量存在真实“零点”
- 变量的不同取值存在自然顺序（ $X_2 \leq X_1$ 或 $X_2 \geq X_1$ ）
- 变量的不同取值之差是有实际意义的（ $X_2 - X_1$ ）
- 变量的不同取值的乘除都是有意义的（ X_1/X_2 ）
 - 如：GDP(亿元)、个人收入(元)等



(总结) 数据层次与数据运算可能性

属性特征	名义变量	序数变量	区间变量	比率变量
可排序		✓	✓	✓
可计数	✓	✓	✓	✓
有众数	✓	✓	✓	✓
有中位数		✓	✓	✓
有均值			✓	✓
差值有意义			✓	✓
可加减			✓	✓
可乘除				✓
有真实“零点”				✓



1.4 数据的时间状态



Typel : 时间序列数据(time series data) :

时间序列数据：对一个变量在不同时间取值的一组观测结果。按取值间隔可分为高频数据和低频数据。

- 实时牌价：如股票价格
- 每日(daily)：如天气预报
- 每周(weekly)：如货币供给数字
- 每月(monthly)：如失业率和消费者价格指数
- 每季度(quarterly)：如GDP
- 每年(annually)：如政府预算
- 每5年(quinquennially)：如制造业普查资料
- 每10年(decennially)：如人口普查资料



Typel : 时间序列数据(time series data) :

平稳性(stationary): 如果一个时间序列的均值和方差不随时间而系统地变化, 那它就是平稳的(stationary)。



1951年1月-1999年9月美国的M1货币供给



Type2 : 截面数据 (cross-section data) :

横截面数据：对一个或多个变量在同一时间点上收集的数据

异质性 (heterogeneity) : 当我们的统计分析包含有异质的单位时, 我们必须考虑尺度 (size) 或规模效应 (scale effect) 以避免造成混乱。



案例：鸡蛋价格与鸡蛋产量

美国50个州的蛋类生产和价格数据

STATE	Y1	X1
AL	2206	92.7
AK	0.7	151
AZ	73	61
AR	3620	86.3
CA	7472	63.4

Showing 1 to 5 of 50 entries

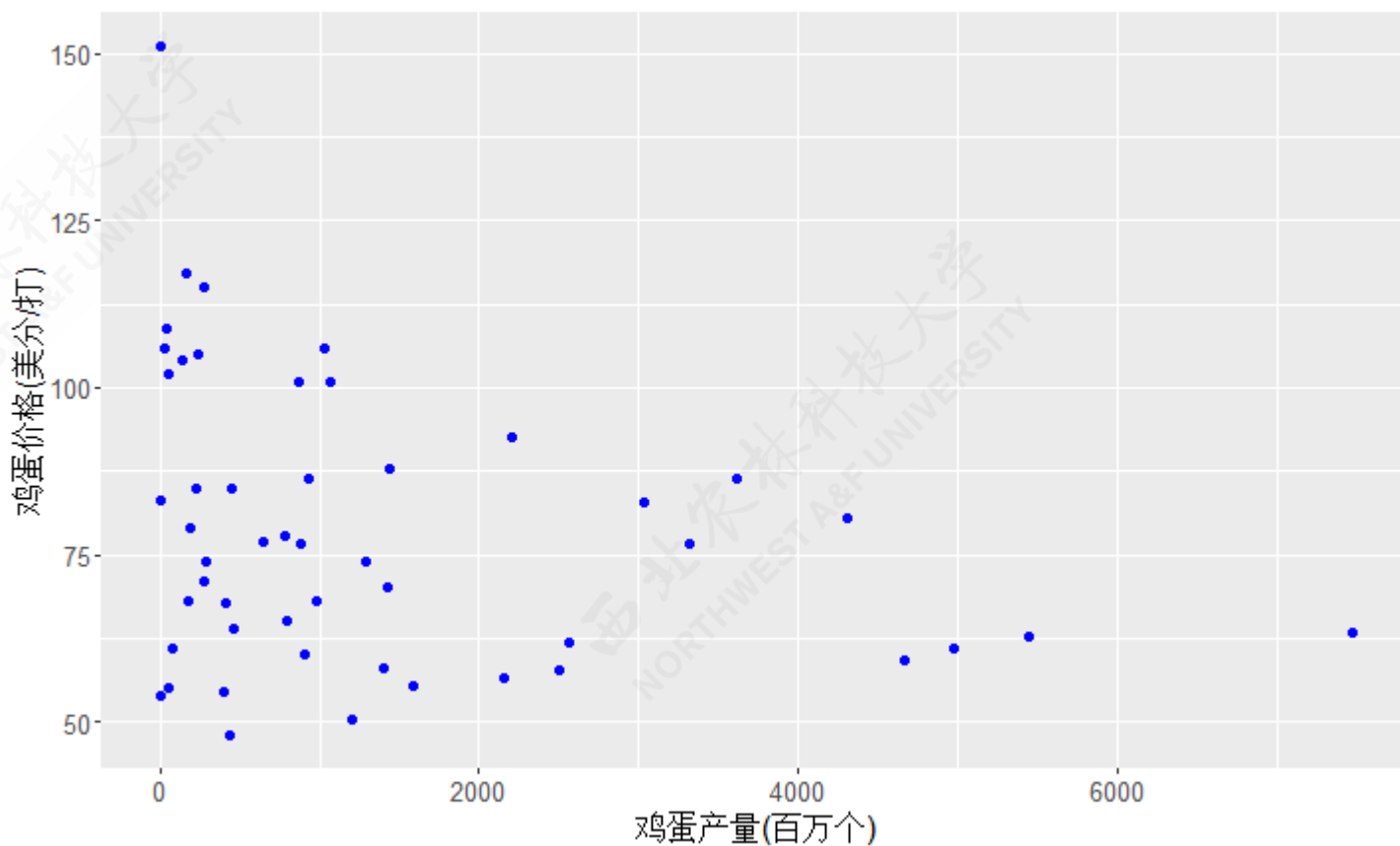
Previous 2 3 4 5 ... 10 Next

其中， Y_1 代表1990年鸡蛋产量(百万个)； X_1 代表1990年每打鸡蛋的价格(美分/打)。





案例：鸡蛋价格与鸡蛋产量



思考提问：图中特征符合经济学理论么？为什么？图中反映了数据可能存在哪些潜在问题？



Type3 : 面板数据 (Panel Data)

面板数据：是兼有时间序列和横截面数据两种成份，指对相同的横截面单元在时间轴上进行跟踪调查的数据。

- 平衡面板 (balanced panel)：所有截面单元都具有相同的观测次数
- 非平衡面板 (unbalanced panel)：并非所有截面单元都具有相同的观测次数

数据点（观测数） n ：

- 数据点（观测数）=截面单元数*时期数，也即： $n = q * t$ 。

可能存在的问题：

- “平稳性”问题：
- “异方差”问题：



案例：钢铁公司

两家钢铁公式的数据案例：

- 公司：GE=通用公司；US=美国钢铁
- I=真实总投资（百万美元）
- F=前一年的企业真实价值（百万美元）
- C=前一年的真实资本存量（百万美元）

```
Warning: `as.tibble()` was deprecated in tibble 2.0.0.  
Please use `as_tibble()` instead.  
The signature and semantics have changed, see `?as_tibble`.
```



案例：钢铁公司

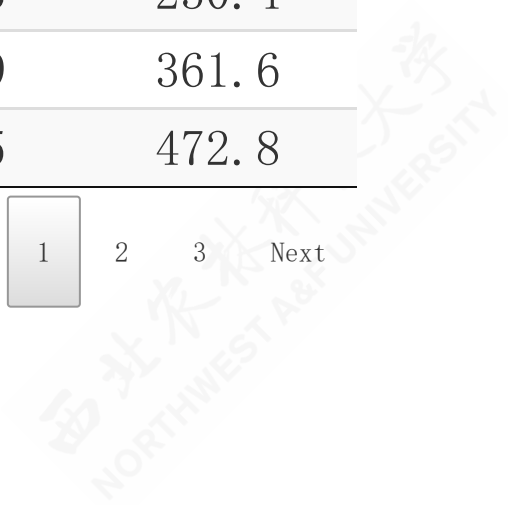
扁数据形式：

1935-1954年间美国两大钢铁公司的数据(扁数据)

year	GE. C	GE. F	GE. I	US. C	US. F	US. I
1935	97.8	1170.6	33.1	53.8	1362.4	209.9
1936	104.4	2015.8	45	50.5	1807.1	355.3
1937	118	2803.3	77.2	118.1	2673.3	469.9
1938	156.2	2039.7	44.6	260.2	1801.9	262.3
1939	172.6	2256.2	48.1	312.7	1957.3	230.4
1940	186.6	2132.2	74.4	254.2	2202.9	361.6
1941	220.9	1834.1	113	261.4	2380.5	472.8

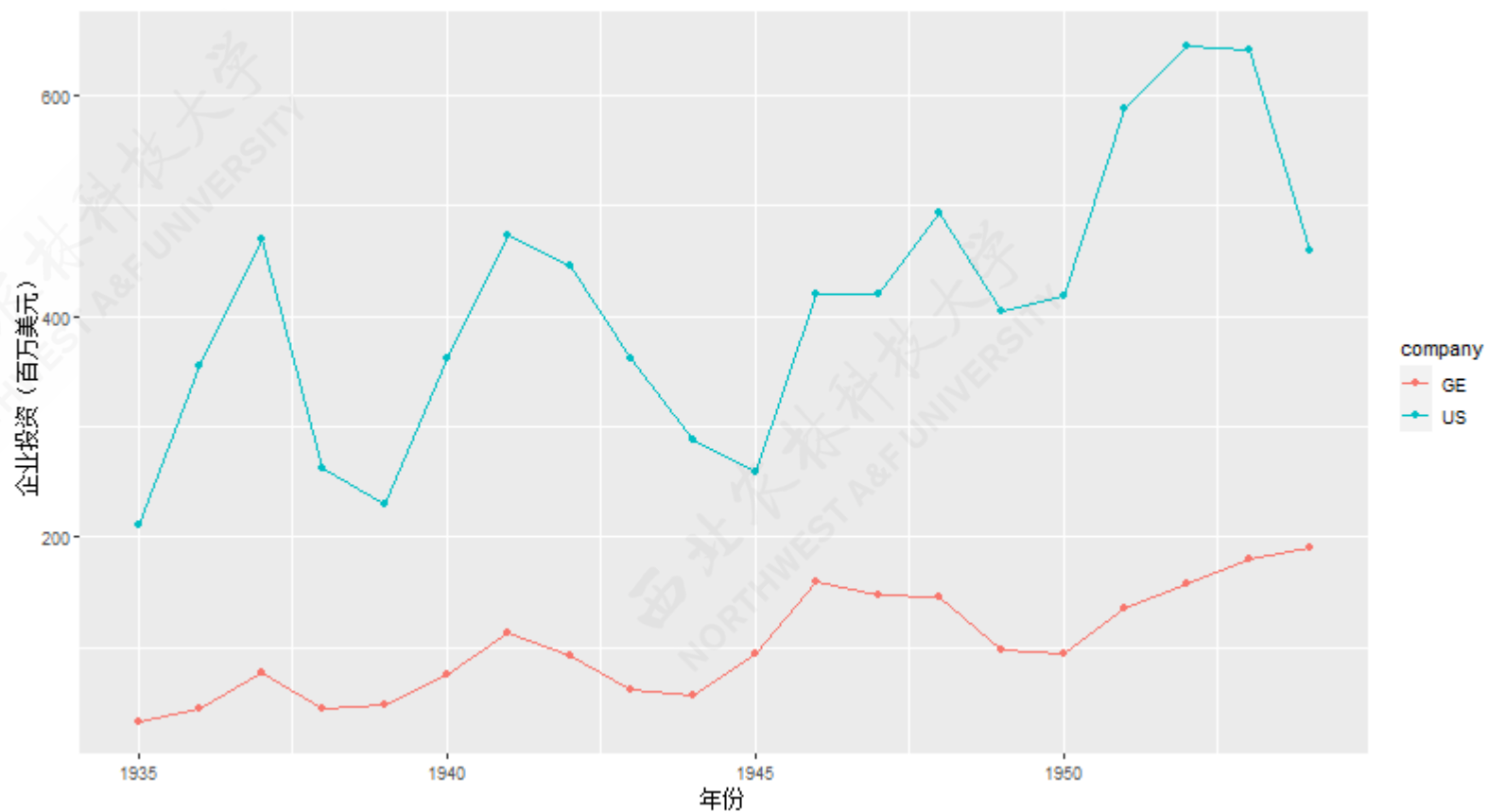
Showing 1 to 7 of 20 entries

Previous **1** 2 3 Next





案例：钢铁公司



两家公司的企业投资情况



案例：钢铁公司

长数据形式：

1935-1954年间美国两大钢铁公司的数据(长数据)

year	company	I	F	C
1935	GE	33.1	1170.6	97.8
1936	GE	45	2015.8	104.4
1937	GE	77.2	2803.3	118
1938	GE	44.6	2039.7	156.2
1939	GE	48.1	2256.2	172.6
1940	GE	74.4	2132.2	186.6
1941	GE	113	1834.1	220.9

Showing 1 to 7 of 40 entries

Previous 1 2 3 4 5 6 Next



案例：钢铁公司

1935-1954年间美国两大钢铁公司的数据(缺失部分数据)

year	GE. C	GE. F	GE. I	US. C	US. F	US. I
1935	97.8	1170.6	33.1	53.8	1362.4	209.9
1936	104.4	2015.8	45	50.5	1807.1	355.3
1937	118	2803.3	77.2	118.1	2673.3	469.9
1938	156.2	2039.7	44.6	260.2	1801.9	262.3
1939	172.6	2256.2	48.1	312.7	1957.3	230.4
1940	186.6	2132.2	74.4			
1941	220.9	1834.1	113	261.4	2380.5	472.8

Showing 1 to 7 of 20 entries

Previous

1

2

3

Next

课堂测试： 问1：平衡面板还是非平衡面板？ 问2：多少数据点？ 问3：两个公司投资函数是否相同？

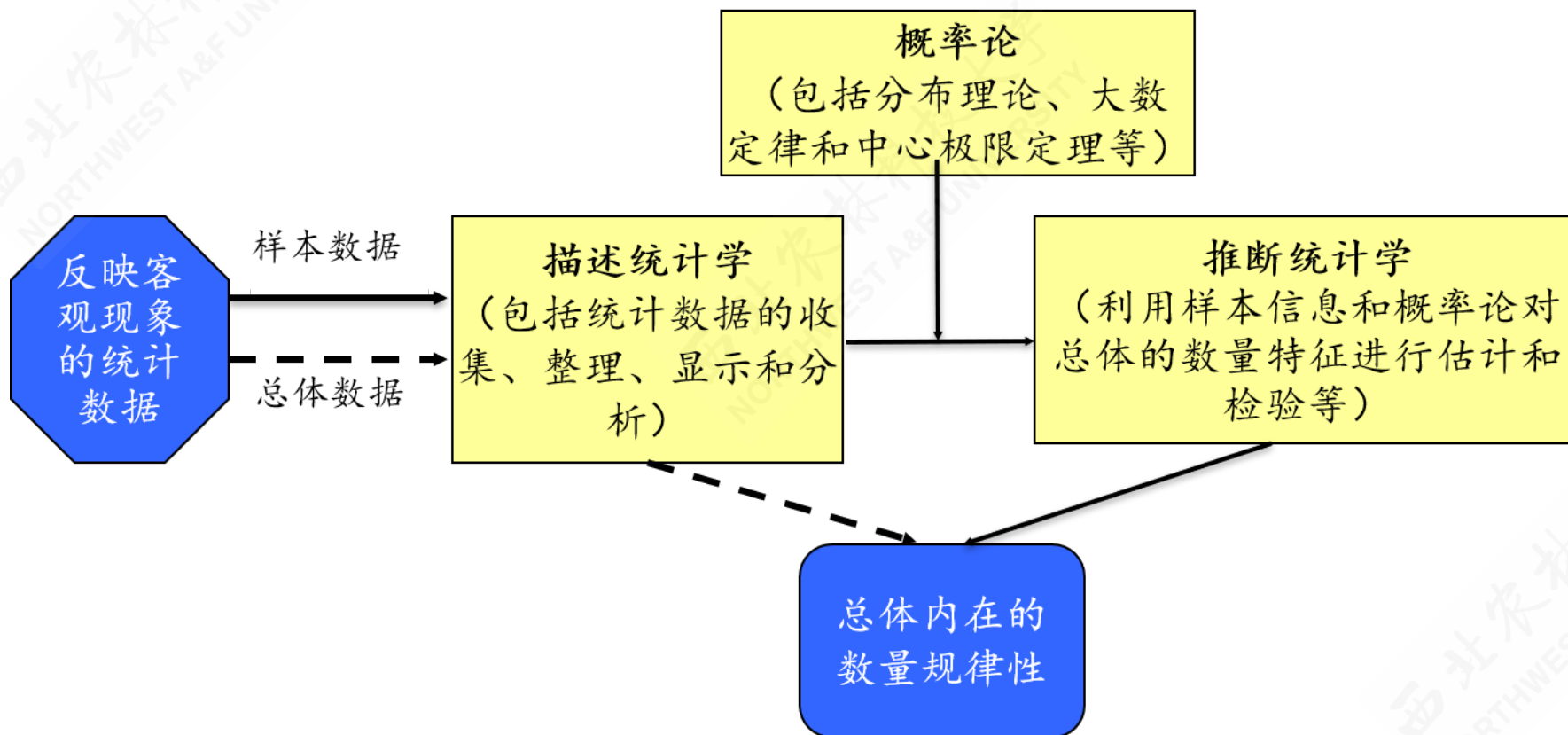
1.5 统计学的体系



统计学的方法论

方法论：用样本**描述**或**推断**总体。

主要途径：概率论在其中发挥着重要作用，也是方法论的分水岭。





描述性统计 (descriptive statistics)

- 定义：研究数据收集、处理、汇总、图表描述、概括与分析等统计方法
- 目的：描述数据特征；找出数据的基本规律
- 内容：
 - 搜集数据
 - 整理数据
 - 展示数据
 - 描述性分析

西北农林科技大学
NORTHWEST A&F UNIVERSITY

西北农林科技大学
NORTHWEST A&F UNIVERSITY

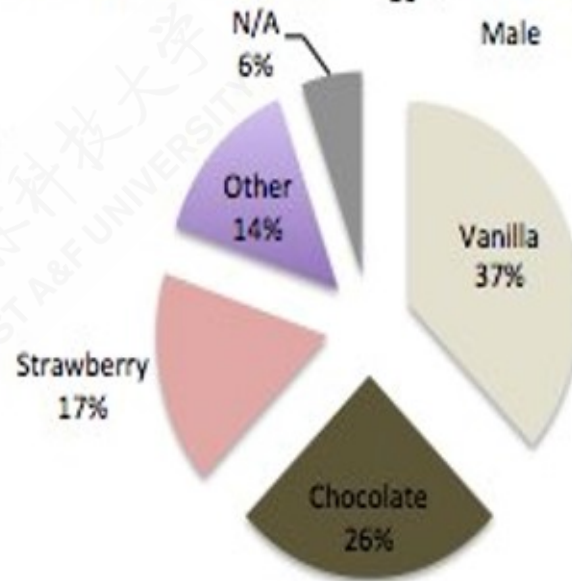


描述性统计 (descriptive statistics)

	A	B	C	D
1	Respondent #	Age	Gender	Favorite Ice Cream Flavor
2	1	36	m	Vanilla
3	2	22	f	Chocolate
4	3	61	m	Strawberry
5	4	88	m	Other
6	5	31	m	N/A
7	6	53	m	N/A
8	7	30	f	Chocolate
9	8	64	f	Chocolate
10	9	18	m	Vanilla
11	10	16	f	Vanilla
12	11	83	m	Strawberry
13	12	16	f	Strawberry
14	13	94	m	Strawberry
15	14	55	m	Vanilla
16	15	42	f	Chocolate
17	16	18	f	Vanilla
18	17	61	f	Vanilla

Raw Data

Age	
Mean	42.6
Standard Dev.	21.9



Descriptive Statistics



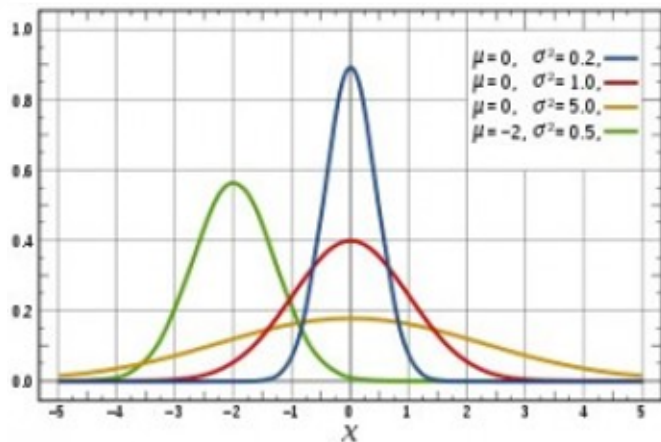
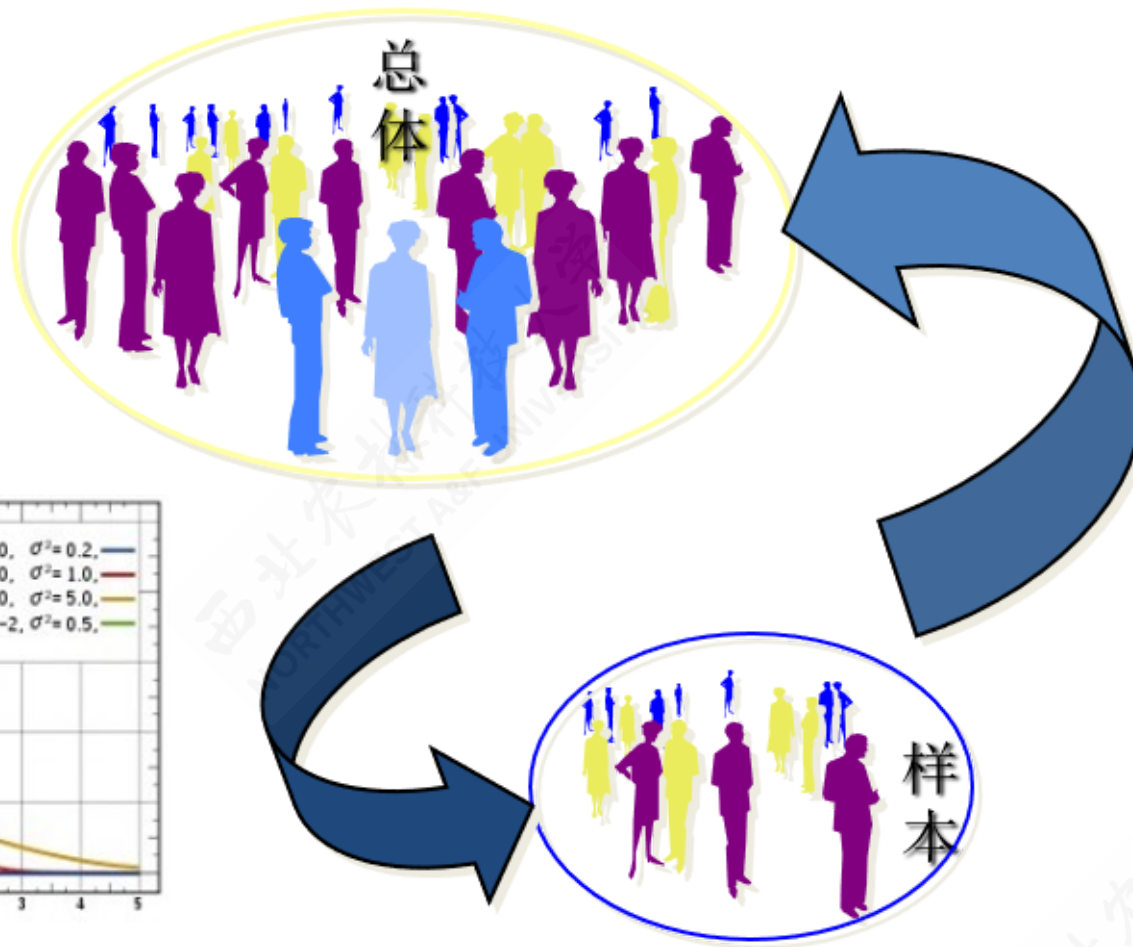
推断性统计 (inferential statistics)

- 定义：研究如何利用样本数据来推断总体特征的统计方法
- 目的：对总体特征作出推断
- 内容：
 - 参数估计
 - 假设检验



推断性统计 (inferential statistics)

西北农林科技大学
NORTHWEST A&F UNIVERSITY



西北农林科技大学
NORTHWEST A&F UNIVERSITY



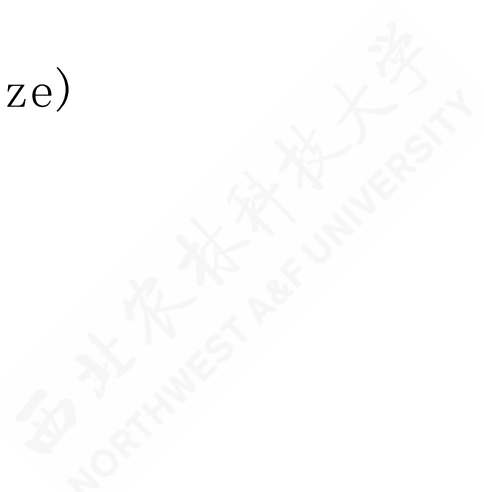
总体和样本

总体 (population):

- 定义: 所研究的全部个体(数据) 的集合, 其中的每一个个体也称为元素
- 分类: 分为有限总体和无限总体
 - 有限总体的范围能够明确确定, 且元素的数目是有限的
 - 无限总体所包括的元素是无限的, 不可数的

样本 (sample)

- 定义: 从总体中抽取的一部分元素的集合
- 样本容量: 构成样本的元素的数目称为样本容量或样本量 (sample size)





参数和统计量

参数(parameter):

- 定义: 描述总体特征的概括性数字度量, 是研究者想要了解的总体的某种特征值
- 重要统计量: 所关心的参数主要有总体均值(μ)、方差(σ^2) 等
- 记号: 总体参数通常用希腊字母表示 $\mu, \sigma^2, \Phi, \gamma, \dots$

统计量(statistic):

- 定义: 用来描述样本特征的概括性数字度量, 它是根据样本数据计算出来的一些量, 是样本的函数
- 重要统计量: 所关心的样本统计量有样本均值(\bar{X})、样本方差(S^2) 等
- 记号: 样本统计量通常用英文字母来表示 $\bar{X}, s^2, w, v, \dots$



参数和统计量

参数	par	统计量	stat
数据元	X	数据元	X
总体期望	μ	样本均值	\bar{X}
总体比率	P	样本比率	p
总体方差	σ^2	样本方差	S^2
总体标准差	σ	样本标准差	S
总体容量	N	样本观测数	n
总体相关系数	ρ	样本相关系数	r

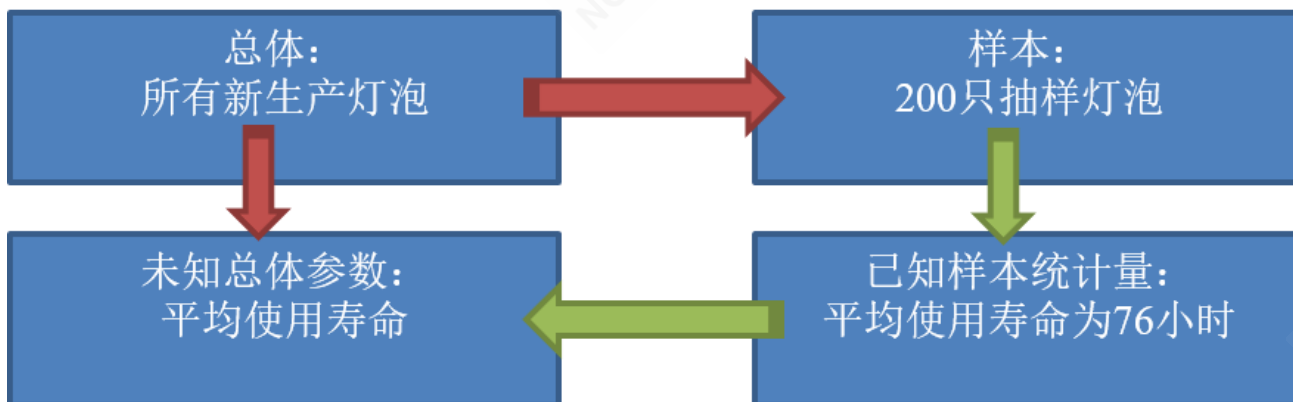




参数和统计量

数据表：NORRIS电器公司200只灯泡使用寿命（时长）

107	73	68	97	76	79	94	59	98	57
54	65	71	70	84	88	62	61	79	98
66	62	79	86	68	74	61	82	65	98
62	116	65	88	64	79	78	79	77	86
74	85	73	80	68	78	89	72	58	69
92	78	88	77	103	88	63	68	88	81



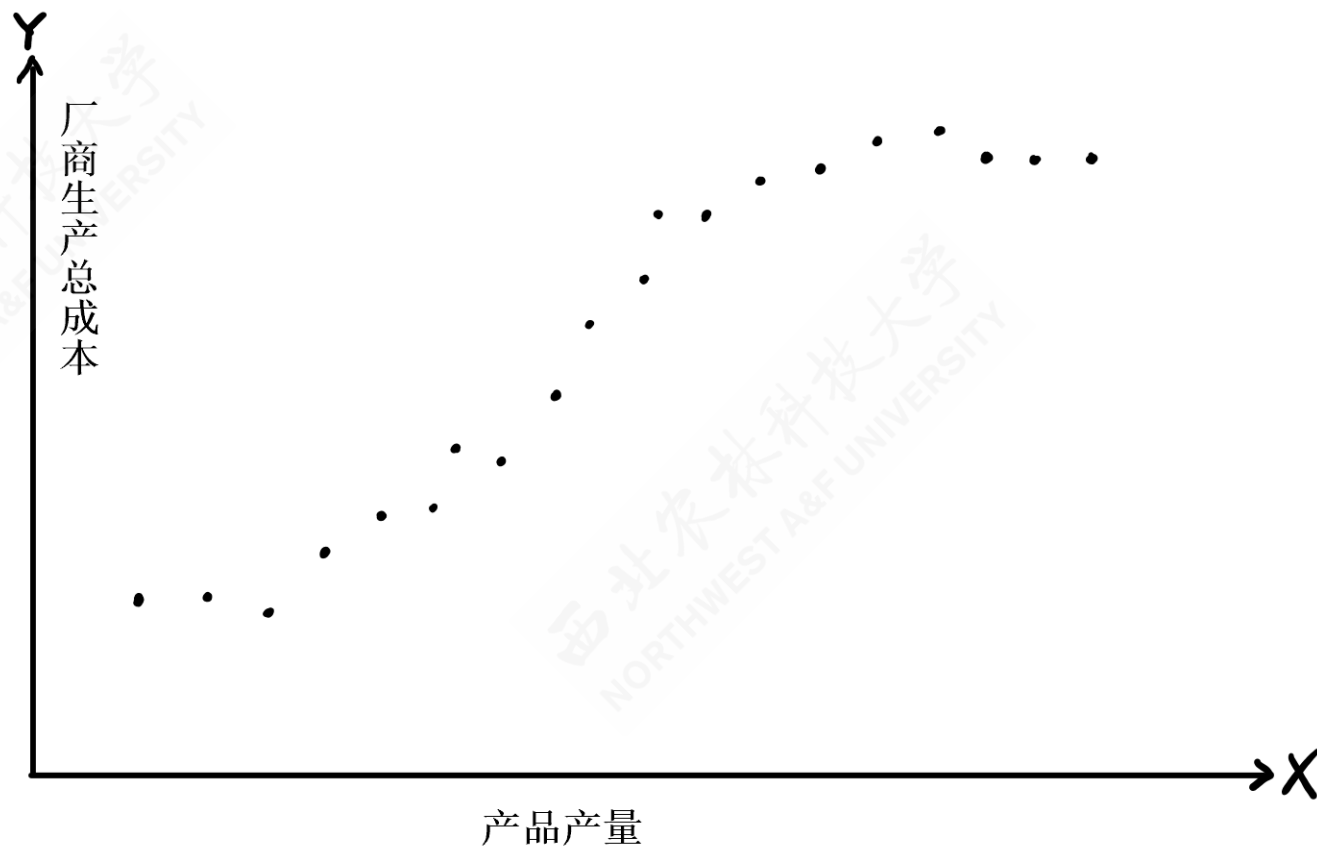


(演示) 样本VS总体视角

- 总体 ($[Y_i, X_i], N$)
- 总体参数???
- 变量的总体关系???
- 样本 ($[y_i, x_i], n$)
- 样本统计量???
- 怎么透过样本数据来窥探总体关系的秘密?



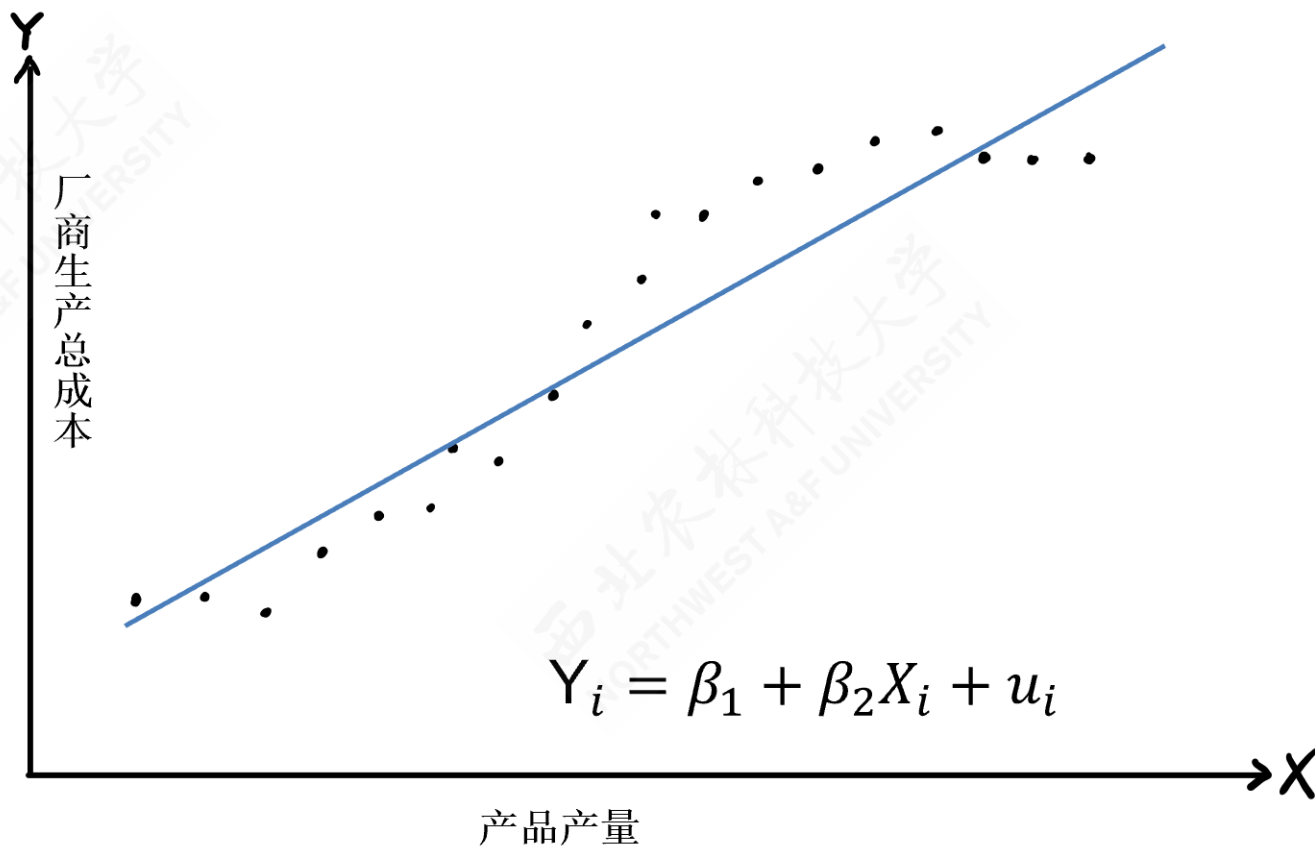
(演示) 样本VS总体视角I



一份样本数据



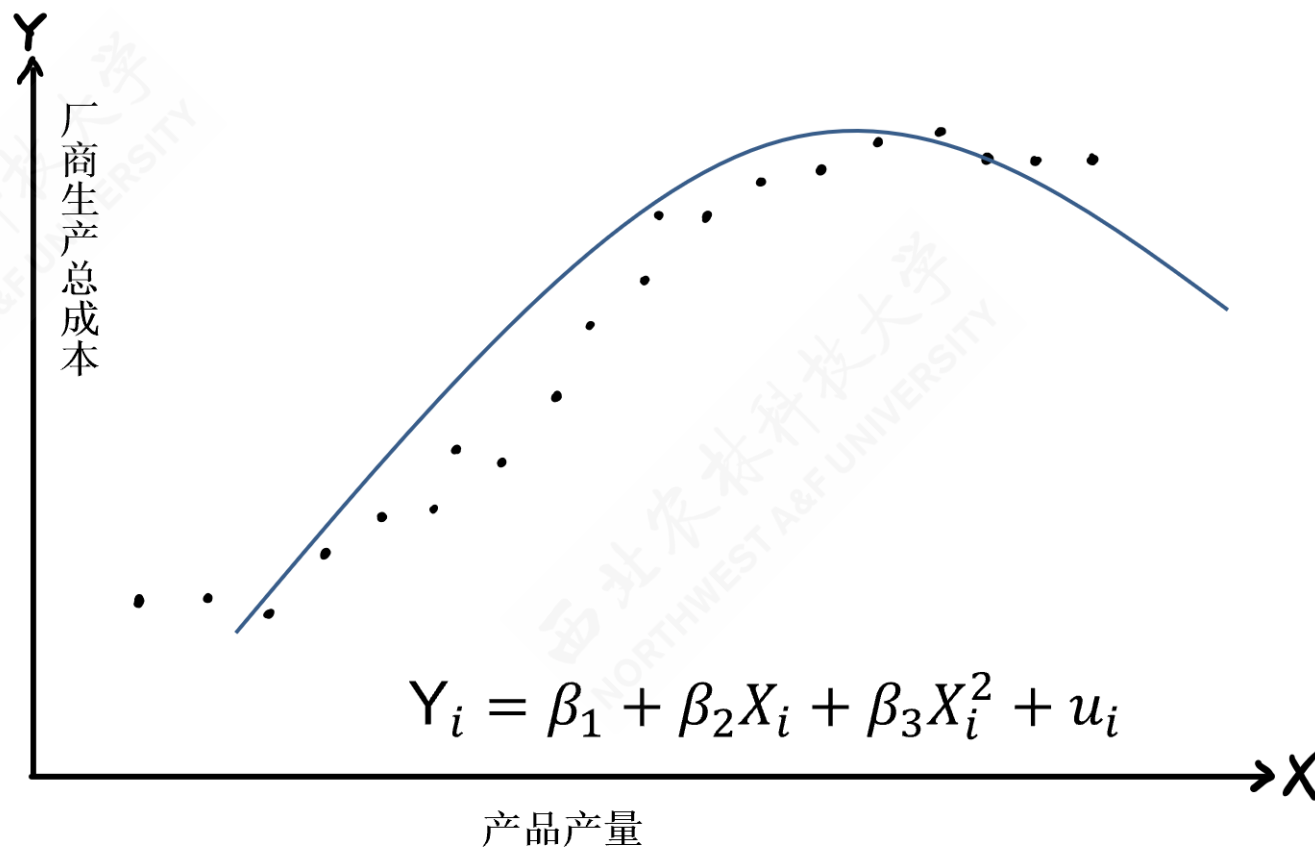
(演示) 样本VS总体视角?



同学的视界：一条直线



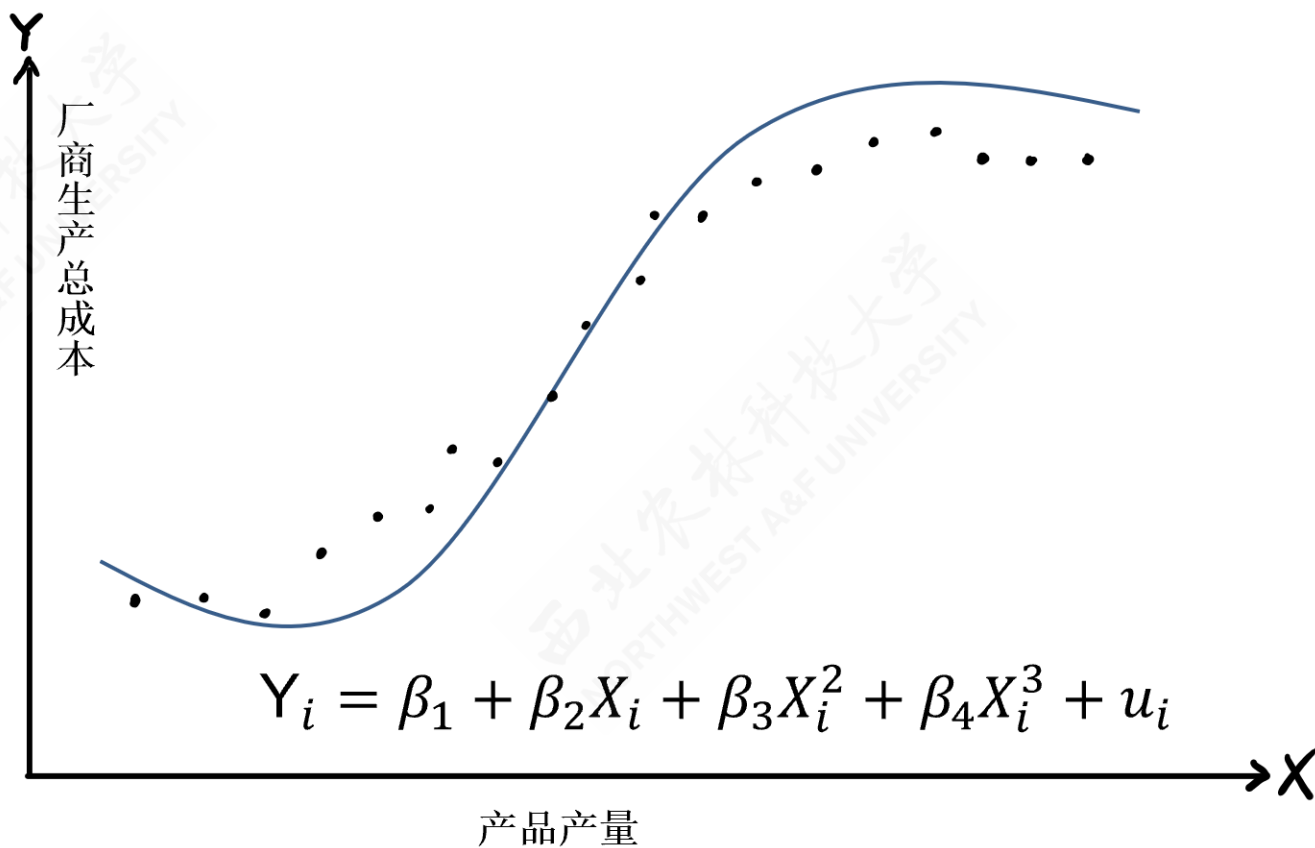
(演示) 样本VS总体视角3



B同学的视界：一条抛物线



(演示) 样本VS总体视角4



0同学的视界：一条S型曲线



(实验演示) 参数VS统计量?

总共162名同学参与随机抽样；每个同学都随机抽取 $n = 10$ 个数据点。因此共有162份样本数据（每份含10个观测数）。

西北农林科技大学
NORTHWEST A&F UNIVERSITY

西北农林科技大学
NORTHWEST A&F UNIVERSITY

西北农林科技大学
NORTHWEST A&F UNIVERSITY



(实验演示) 参数VS统计量3

总体参数有哪些?

样本统计量又有哪些?

总体:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

$$Y_i = 25 + 0.5X_i + u_i$$

样本:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$$

1.6 统计分析的基本过程



统计分析的基本过程

基本过程包括：

- 实际问题：发现问题
- 收集数据：取得数据
- 处理数据：整理与图表展示
- 分析数据：利用统计方法分析数据
- 解释数据：结果的说明
- 得到结论：从数据分析中得出客观结论





统计分析需要具备的能力

A. 基础背景教育：

- 知识：高等数学、概率论与数理统计、商业经济统计、科学哲学...
- 技能1：前沿阅读：学术期刊academic journal（[西农电子期刊数据库](#)；[开源期刊库 arXiv.org](#)）、RSS订阅（[统计学之都](#)；[R-bloggers](#)）、[google学术](#)...
- 技能2：知识管理：Onenote、印象笔记、Mindmanage（思维导图）...

B. 广泛数据获取：

- 知识：实验设计、调研经历、数据爬虫

技能：大学生创新创业训练、开源数据库（国内微观调研数据库、国外[google dataset search](#)）、问卷星、Chrome开发工具、HTML标记语言、CSS语言、Regex语言、docker操作、Selenium/RSelenium



统计分析需要具备的能力?

C. 系统数据管理:

- 知识: 数据库查询语言、连续整合或协作系统 (CI)
- 技能: SQL、DBI、[RSQLite](#); git语言及[GitHub](#)自动化托管。

D. 数据可视化:

- 知识: 现代图形理论、传统/现代编程语言
- 技能: Microsoft Excel、SPSS、

E. 数据建模分析

- 知识: 数据清洗、数据变换、非结构化数据处理、平行运算、计量建模、深度学习与人工智能...
- 技能: Hadoop、Machine Learning and AI、交互制图技术 ([R Graph Gallery](#))、[shiny APP](#)、books ([Interpretable Machine Learning](#))



统计分析需要具备的能力3

F. 报告写作交流

- 知识：学术道德规范、科技论文规则、高效写作工具、论文发表、学术研讨交流
- 技能：传统写作排版（Word、Latex语言）、现代标记写作语言（Markdown/Rmarkdown）、文献管理工具（Endnote、Zotero）、可复现图文技术、演讲汇报slide（Powerpoint、Xaringan）、个人网站（hugo语言、blogdown）

G. 技术技巧之外

- 知识：对智慧和科学的好奇心、生活或商业上的感知与睿智
- 技能：沟通技能、团队协作





统计分析过程：纽约案例

航班延误背后的故事：纽约繁忙的天空。数据来源nycflights13

This package contains information about all flights that departed from NYC (e.g. EWR, JFK and LGA) in 2013: 336,776 flights in total. To help understand what causes delays, it also includes a number of other useful datasets. This package provides the following data tables.

- flights: all flights that departed from NYC in 2013
- weather: hourly meteorological data for each airport
- planes: construction information about each plane
- airports: airport names and locations
- airlines: translation between two letter carrier codes and names





实际问题：航班延误有规律可循么？变量情况

从2013-01-01 05:00:00到2013-12-31 23:00:00期间，纽约市三个机场EWR、JFK、LGA，起落航班总数有336776架次。

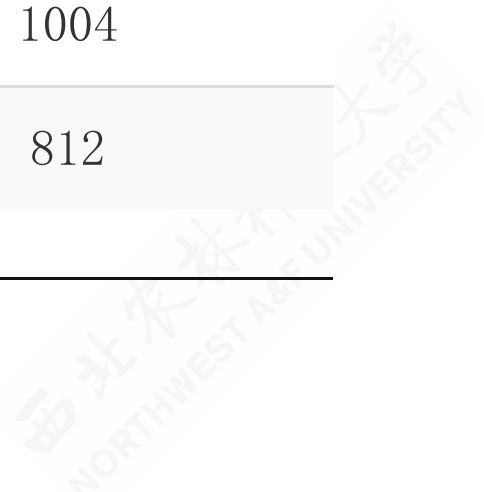
```
tibble [336,776 x 19] (S3: tbl_df/tbl/data.frame)
 $ year          : int [1:336776] 2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 2013
 $ month         : int [1:336776] 1 1 1 1 1 1 1 1 1 1 ...
 $ day           : int [1:336776] 1 1 1 1 1 1 1 1 1 1 ...
 $ dep_time      : int [1:336776] 517 533 542 544 554 554 555 557 557 558 ...
 $ sched_dep_time: int [1:336776] 515 529 540 545 600 558 600 600 600 600 ...
 $ dep_delay     : num [1:336776] 2 4 2 -1 -6 -4 -5 -3 -3 -2 ...
 $ arr_time      : int [1:336776] 830 850 923 1004 812 740 913 709 838 753 ...
 $ sched_arr_time: int [1:336776] 819 830 850 1022 837 728 854 723 846 745 ...
 $ arr_delay     : num [1:336776] 11 20 33 -18 -25 12 19 -14 -8 8 ...
 $ carrier       : chr [1:336776] "UA" "UA" "AA" "B6" ...
 [list output truncated]
```




分析数据：描述性统计1——1天的航班情况

2013年1月1日这一天的航班总数为842架次。从这样的数据表能看出什么规律？

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched
2013	1	1	517	515	2	830	
2013	1	1	533	529	4	850	
2013	1	1	542	540	2	923	
2013	1	1	544	545	-1	1004	
2013	1	1	554	600	-6	812	





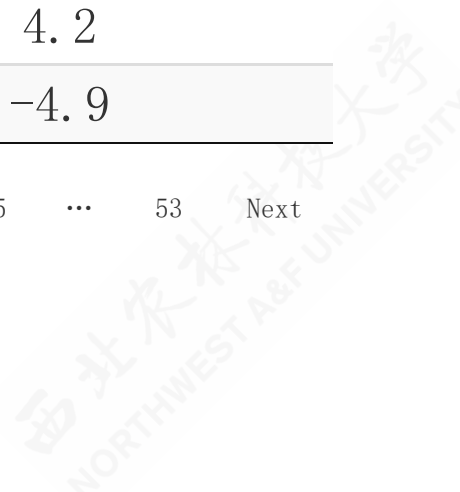
分析数据：描述性统计2——每天的航班数和平均延误时长

这样的统计分析，能不能看出任何蛛丝马迹呢？

date	n_flight	mean_dep_delay	mean_dep_arr
2013-01-01	842	11.5	12.7
2013-01-02	943	13.9	12.7
2013-01-03	914	11.0	5.7
2013-01-04	915	9.0	-1.9
2013-01-05	720	5.7	-1.5
2013-01-06	832	7.1	4.2
2013-01-07	933	5.4	-4.9

Showing 1 to 7 of 365 entries

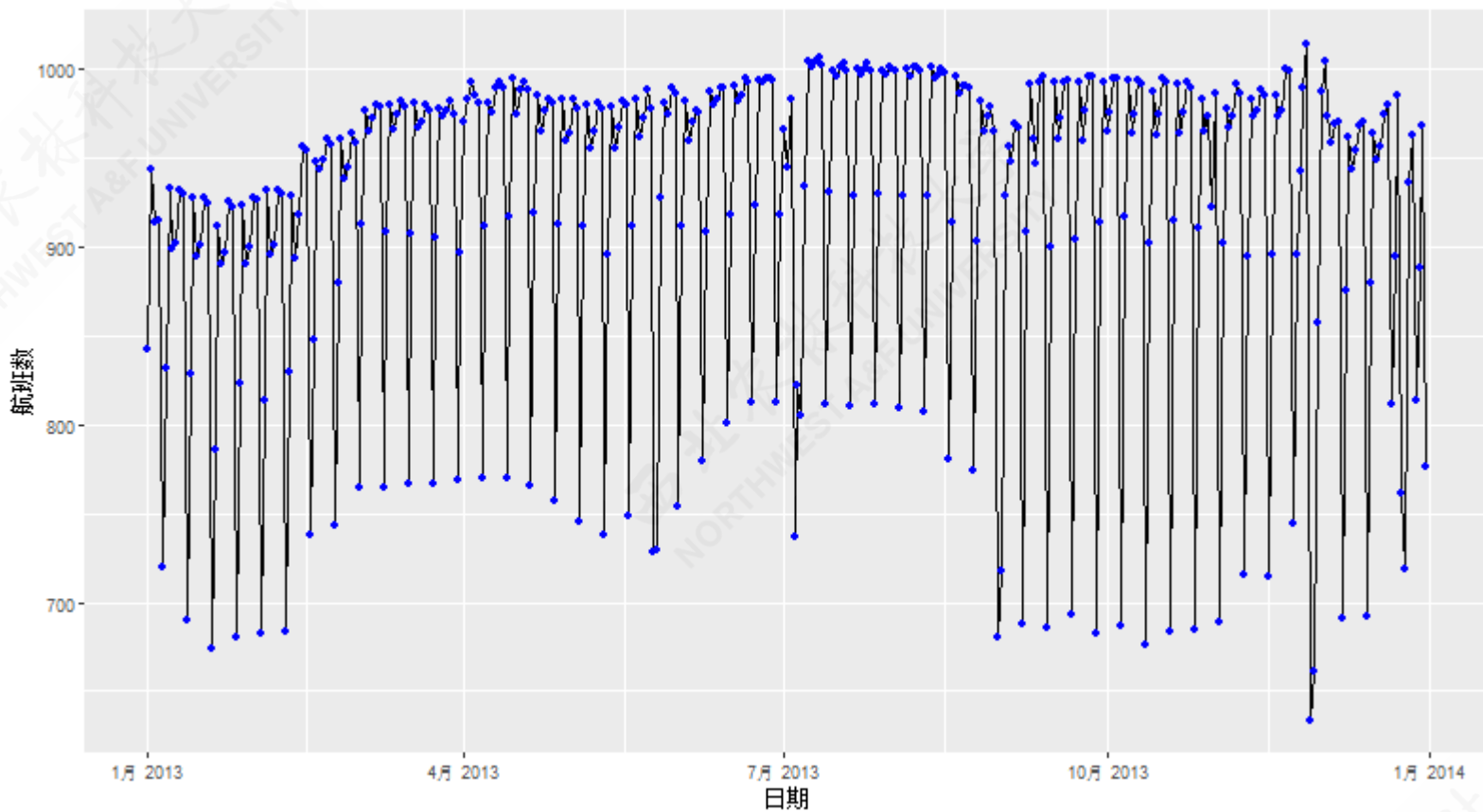
Previous **1** 2 3 4 5 ... 53 Next





分析数据：描述性统计3(每天航班数)

按时间轴查看每天航班数的全貌图，是不是会更轻松一点？





分析数据：描述性统计4(航班数和平均延误时长)

星期几与航班数和平均延误时长有关系？童鞋们，咱先上一个表：

date	wday	n_flight	mean_dep_delay	mean_dep_arr
2013-01-01	周二	842	11.5	12.7
2013-01-02	周三	943	13.9	12.7
2013-01-03	周四	914	11.0	5.7
2013-01-04	周五	915	9.0	-1.9
2013-01-05	周六	720	5.7	-1.5
2013-01-06	周日	832	7.1	4.2
2013-01-07	周一	933	5.4	-4.9

Showing 1 to 7 of 365 entries

Previous

1

2

3

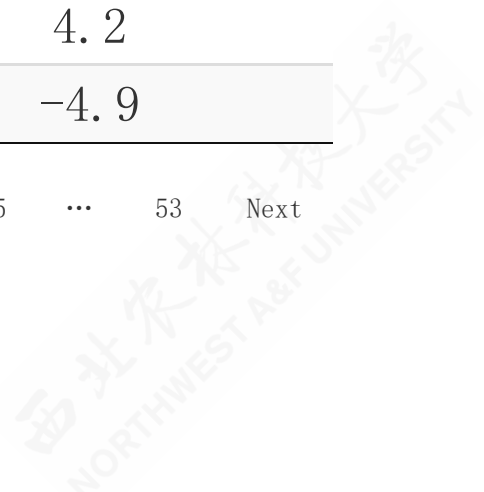
4

5

...

53

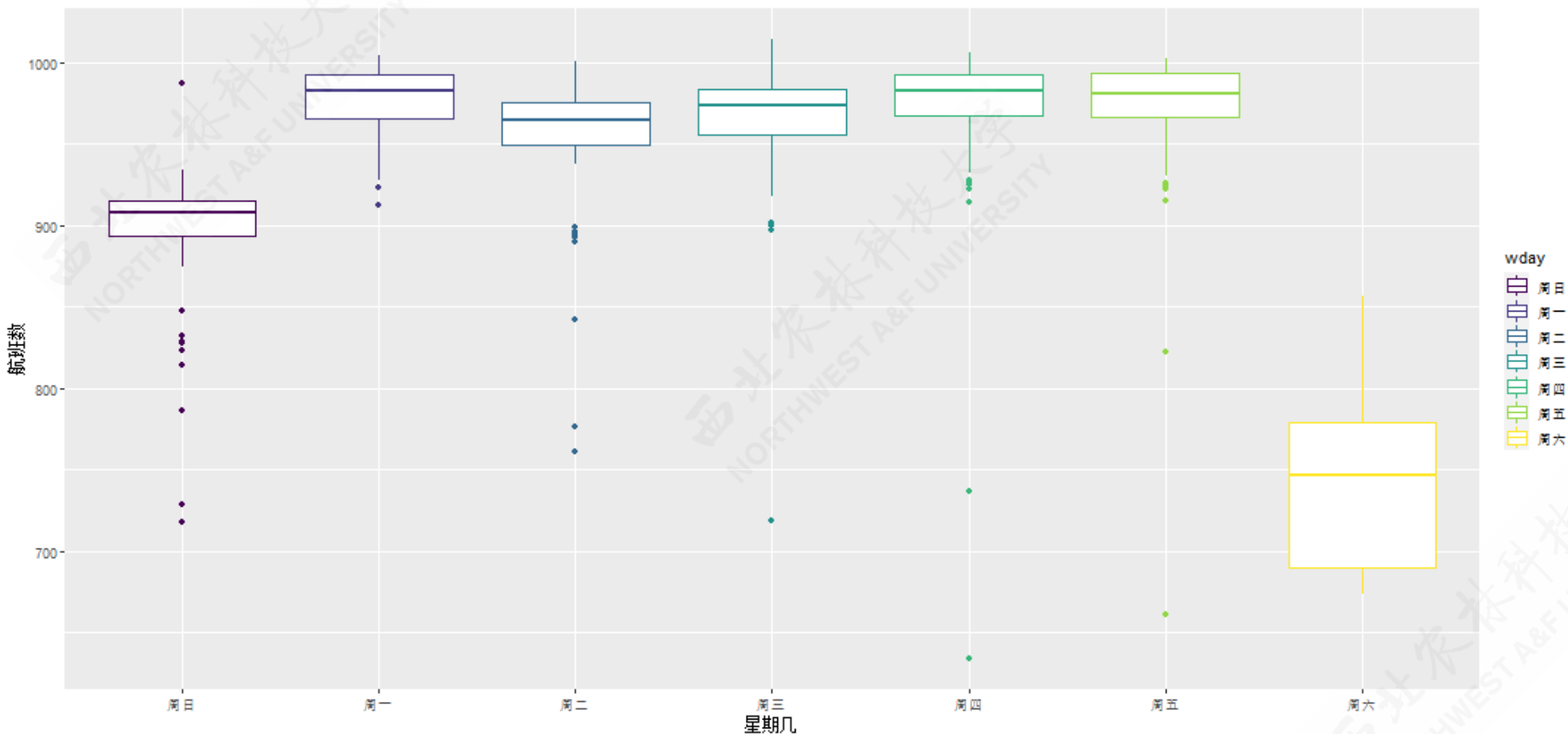
Next





分析数据：描述性统计5(星期几matter?)

小伙伴们，那咱们再上个图（多批箱线图）瞧一瞧吧





分析数据：描述性统计6(幸运星期六)

星期六会更适合坐飞机么？星期六好像航班很少耶！！扣扣鼻子问，WHY？具体是怎样呢？

每天(星期)的航班数和平均延误时长

date	wday	n_flight	mean_dep_delay	mean_dep_arr
2013-01-05	周六	720	5.7	-1.5
2013-01-12	周六	690	1.6	-13.0
2013-01-19	周六	674	3.5	-8.5
2013-01-26	周六	680	7.2	0.8
2013-02-02	周六	682	5.4	-4.8
2013-02-09	周六	684	18.5	6.6

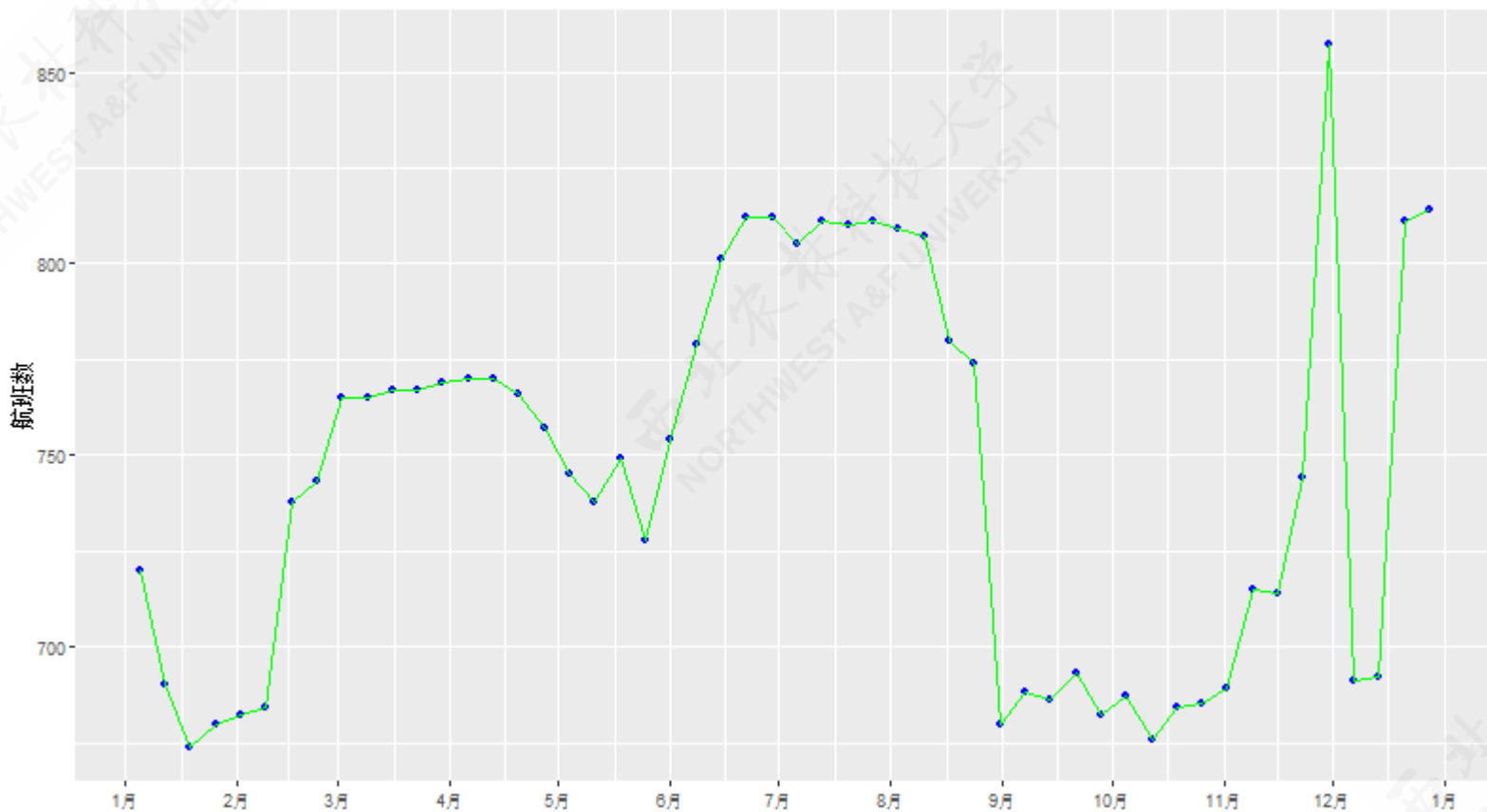
Showing 1 to 6 of 52 entries

Previous 1 2 3 4 5 ... 9 Next



分析数据：描述性统计7(全年的周六)

全年只看星期六的航班数是怎样的赶脚？惊讶的小伙伴们，看看星期六的表现吧！





分析数据：描述性统计8(机场matter?)

选择起飞机场是不是能避免延误？

三大机场每月航班数和月平均延误时长

origin	month	n_flight	mean_dep_delay	mean_dep_arr
EWR	1	9893	14.9	12.8
EWR	2	9107	13.1	8.8
EWR	3	10420	18.1	10.6
EWR	4	10531	17.4	14.1
EWR	5	10592	15.4	5.4
EWR	6	10175	22.5	16.9

Showing 1 to 6 of 36 entries

Previous

1

2

3

4

5

6

Next





分析数据：描述性统计9(机场matter?2)

选择起飞机场是不是能避免延误？

三大机场每月航班数和月平均延误时长

origin	n_flight	mean_dep_delay	mean_dep_arr
EWR	120835	15.0	9.1
JFK	111279	12.0	5.4
LGA	104662	10.4	5.9

西北农林科技大学
NORTHWEST A&F UNIVERSITY

西北农林科技大学
NORTHWEST A&F UNIVERSITY



分析数据：描述性统计9(机场matter?)

下面看看美国所有机场信息(airport):变量情况

faa	name	lat	lon	alt	tz	dst	tzone
04G	Lansdowne Airport	41.1	-80.6	1044	-5	A	America/New_York
06A	Moton Field Municipal Airport	32.5	-85.7	264	-6	A	America/Chicago
06C	Schaumburg Regional	42.0	-88.1	801	-6	A	America/Chicago
06N	Randall Airport	41.4	-74.4	523	-5	A	America/New_York
09J	Jekyll Island Airport	31.1	-81.4	11	-5	A	America/New_York

Showing 1 to 5 of 1,458 entries

Previous 1 2 3 4 5 ... 292 Next

美国总共有1458个机场。





分析数据：描述性统计10(机场matter?4)

概括地来看看机场数据吧：

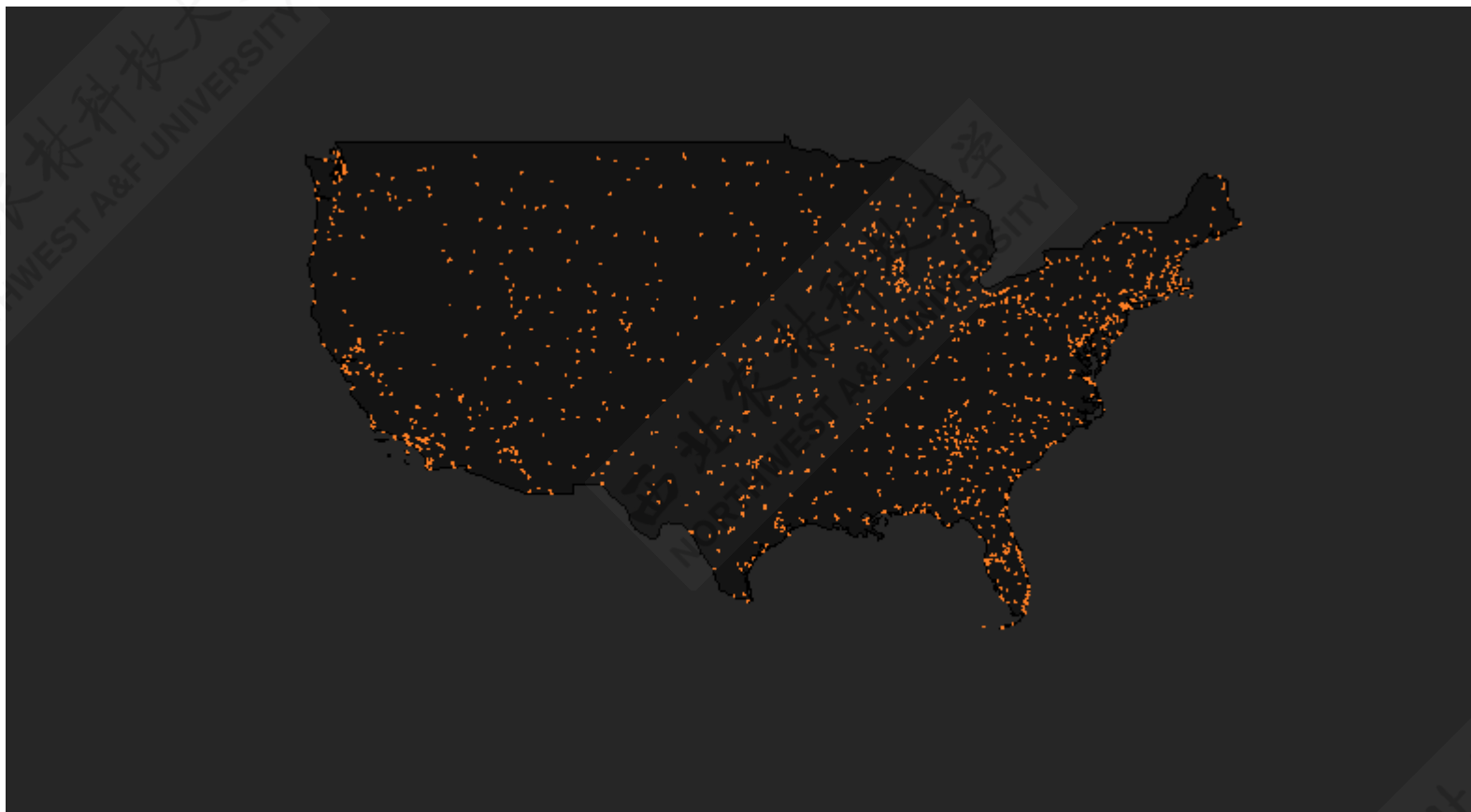
```
Rows: 1,458
Columns: 8
$ faa    <chr> "04G", "06A", "06C", "06N", "09J", "0A9", "0G6", "0...
$ name   <chr> "Lansdowne Airport", "Moton Field Municipal Airport...
$ lat    <dbl> 41, 32, 42, 41, 31, 36, 41, 43, 40, 48, 40, 42, 41,...
$ lon    <dbl> -81, -86, -88, -74, -81, -82, -85, -77, -77, -123, ...
$ alt    <dbl> 1044, 264, 801, 523, 11, 1593, 730, 492, 1000, 108,...
$ tz     <dbl> -5, -6, -6, -5, -5, -5, -5, -5, -5, -8, -5, -6, -5,...
$ dst    <chr> "A", "A", "A", "A", "A", "A", "A", "A", "U", "A", "...
$ tzone  <chr> "America/New_York", "America/Chicago", "America/Chi...
```





分析数据：描述性统计II(机场matter?5)

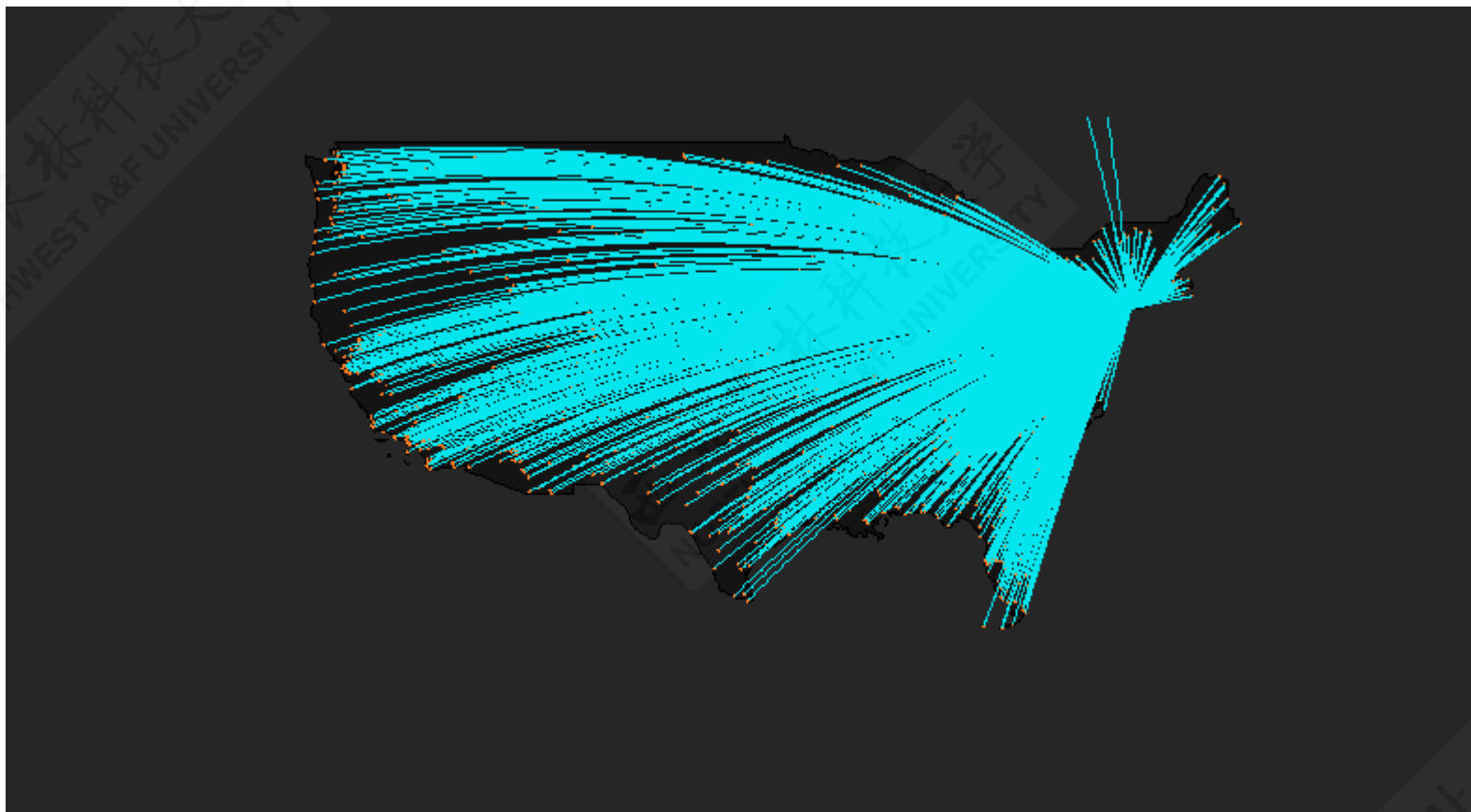
来一个炫酷一点的可视化直观地图吧





分析数据：描述性统计12(机场matter?6)

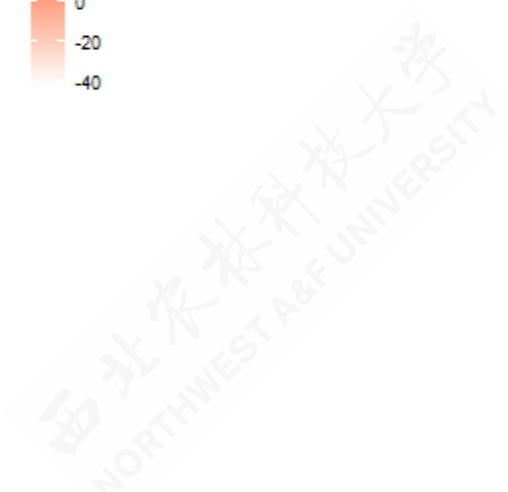
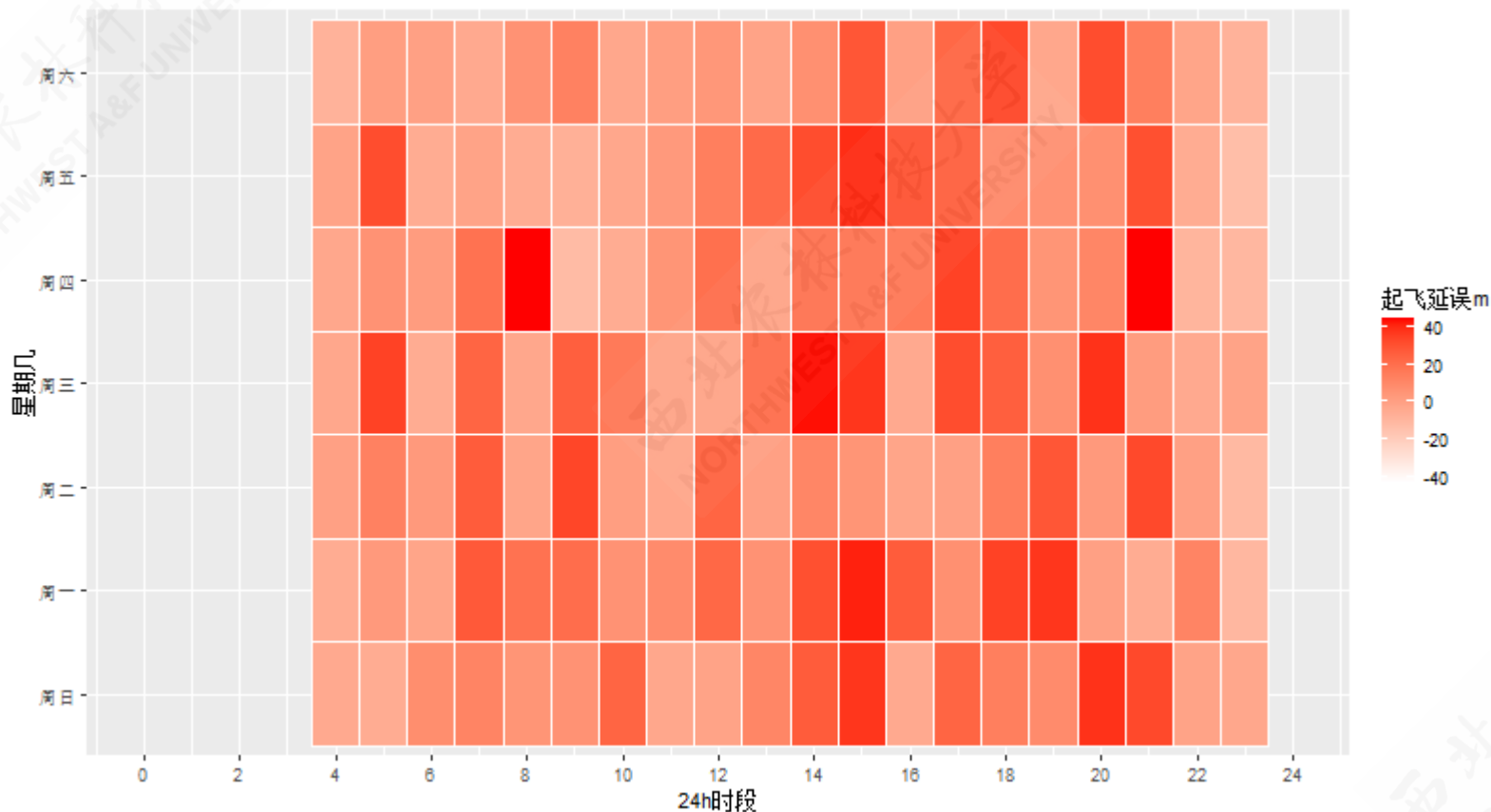
机场信息(airport):可视化





分析数据：描述性统计(JFK机场的热力图)

下面重点看看纽约JFK机场全年的起飞延误情况（热力图heat map plot），日期和星期几确实影响起飞延误么？





分析数据：推断性统计(JFK机场ANOVA分析I)

```
# 起飞延误时长VS星期几  
summary(aov(dep_delay ~ wd, data = flights_jfk))
```

```
|  
      Df    Sum Sq Mean Sq F value           Pr(>F)  
wd          6    246401   41067      27 <0.00000000000000002 ***  
Residuals 109409 166473256    1522  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
1863 observations deleted due to missingness  
|
```





分析数据：推断性统计(JFK机场ANOVA分析2)

```
# 起飞延误时长VS24h时间段  
summary(aov(dep_delay ~ hh, data = flights_jfk))
```

```
      Df    Sum Sq Mean Sq F value    Pr(>F)  
hh      1   5701754 5701754    3874 <0.00000000000000002 ***  
Residuals 109414 161017904    1472
```

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

1863 observations deleted due to missingness





分析数据：推断性统计(JYK机场回归分析)

$$dep_delay = \beta_0 + \beta_1 Mon + \beta_1 Tue + \beta_1 Wen + \beta_1 Thu + \beta_1 Fri + \beta_1 Sta + \beta_1 hh + u_i$$

Min	1Q	Median	3Q	Max
-63.7	-17.7	-9.4	-0.9	1296.5

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-7.138	0.330	-21.63	< 0.000000000000000002	***
wd.L	-0.987	0.308	-3.20	0.0014	**
wd.Q	-1.232	0.308	-3.99	0.0000650013111561	***
wd.C	-1.193	0.307	-3.89	0.0000999799605708	***
wd^4	-2.448	0.306	-8.00	0.000000000000000013	***
wd^5	2.276	0.305	7.46	0.0000000000000000843	***
wd^6	-0.573	0.306	-1.87	0.0611	.
hh	1.494	0.024	62.25	< 0.000000000000000002	***

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

Residual standard error: 38 on 109408 degrees of freedom
(1863 observations deleted due to missingness)

Multiple R-squared: 0.0356, Adjusted R-squared: 0.0356

F-statistic: 578 on 7 and 109408 DF, p-value: <0.000000000000000002





得到结论：

西北农林科技大学
NORTHWEST A&F UNIVERSITY

西北农林科技大学
NORTHWEST A&F UNIVERSITY

西北农林科技大学
NORTHWEST A&F UNIVERSITY

本章結束

