上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

数据世界探秘

社会生活中的统计思维

授课要点

1.新闻媒体中的统计思维

2.日常生活中的统计思维

3.历史中的统计思维

4.文学中的统计思维

5.经济中的统计思维

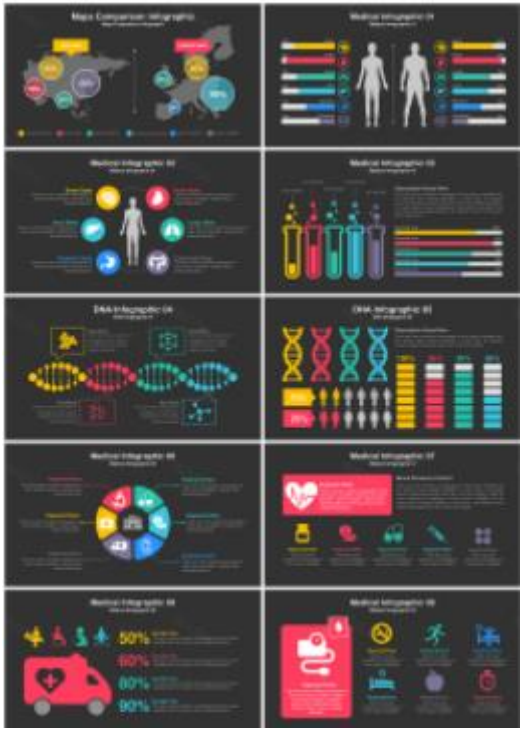
什么是统计思维

“学者不能离开统计而究学，政治家不能离开统计而施政，事业家不能离开统计而执业。”

——经济学家和教育学家马寅初

什么是统计思维

- 统计学是一种由经验到理性的认识，是一种运用偶然发现规律的科学。
- 从统计学的角度看，人们从经验或实验中所获取的知识是含有不确定性的，统计关注的是这种知识当中所含不确定性的度量问题，一旦能得到不确定性的量度，人们的知识就得到扩充，对世界的认知就朝前跨越，这个过程在人类知识积累的进程中不断重复。
- 在日益依赖数据的今天，树立正确的统计思维，才能有效地开展数据处理与分析。



➢ 什么是统计思维

Q：统计学究竟在做些什么？

从随机性中寻找规律性，是统计的基本思想，也是统计的魅力所在。

简单来说，统计学里所表达的两个核心理念就是：

- **允许误差下的概率保证；**
- **允许误差下的统计推断。**

统计学里面，则是处处存在随机性问题。它允许有误差，没有误差反令人怀疑其中有假。统计也会对一个问题拍胸脯保证，但它的保证都是基于概率形式的。而且所能保证的概率，不但不是百分之百，而且还附有误差。统计里则处处是“说不准”。例如，宣称某饮料的容量有百分之九十五的概率介于425毫升至431毫升之间，就是一典型的统计上的保证。统计代表了一种我们看待这个世界的方式。

什么是统计思维

概率和**误差**，构成了统计思维的两大支柱。并发展出统计学里几乎所着的关键要点。

统计学里的方法，和人们的思维方式有一定的对应关系。下面我们就来列举下统计学中常见的思维方式。

4.1、 新闻媒体中的统计思维

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

➢ 关于收入信心的案例

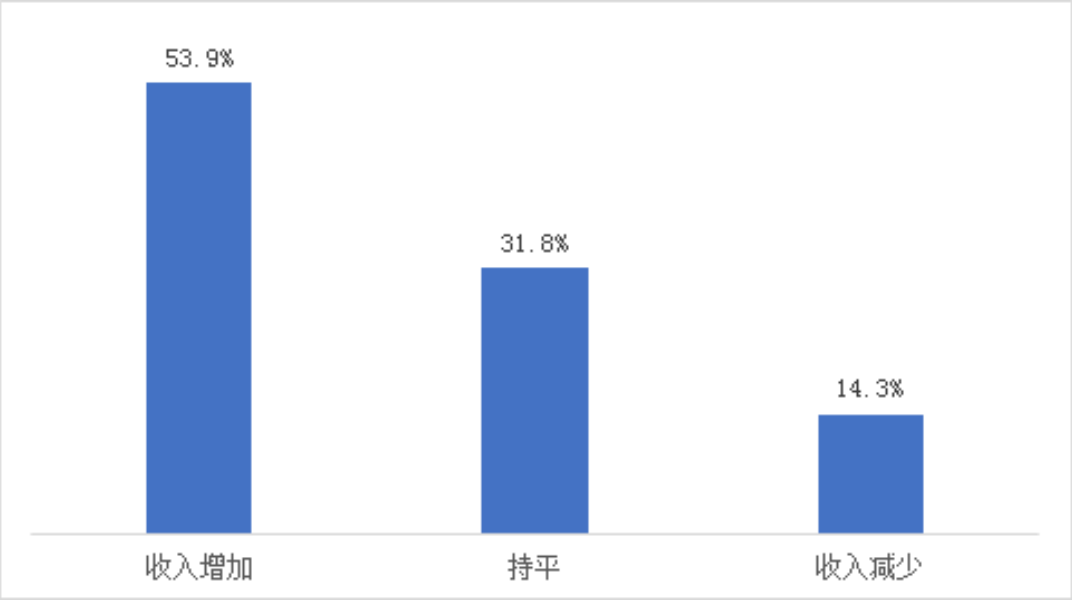
《中国经济生活大调查》是中央电视台、中国邮政集团公司、国家统计局、北京大学国家发展研究院联合发起的，全球最大规模的媒体入户问卷调查，每年通过明信片的方式，对中国10万家庭进行入户问卷调查。通过调查数据，建立善于利用数据的统计思维。

4.1、 新闻媒体中的统计思维

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

➢ 关于收入信心的案例

基于《中国经济生活大调查（2017-2018）》的当前数据来看，有31.8%的人认为2018年的收入会与2017年持平，而超过一半的人认为收入水平会增加，其中90后对收入增加信心最足。



4.1、 新闻媒体中的统计思维

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

- 关于收入信心的案例

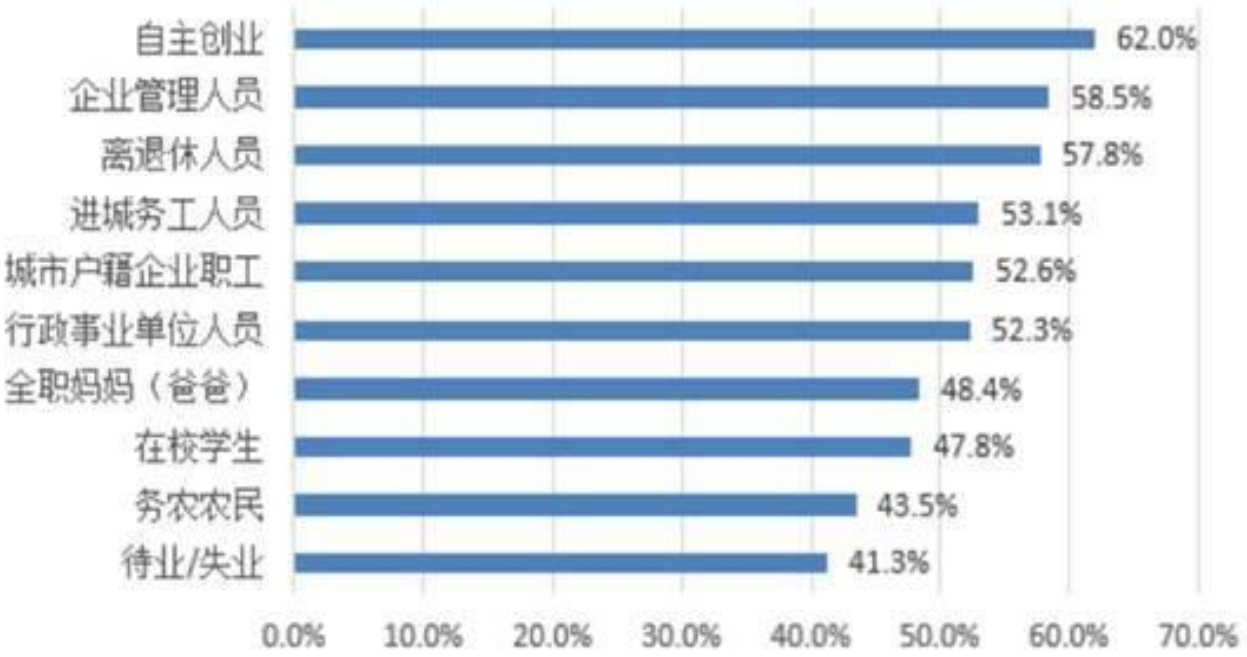


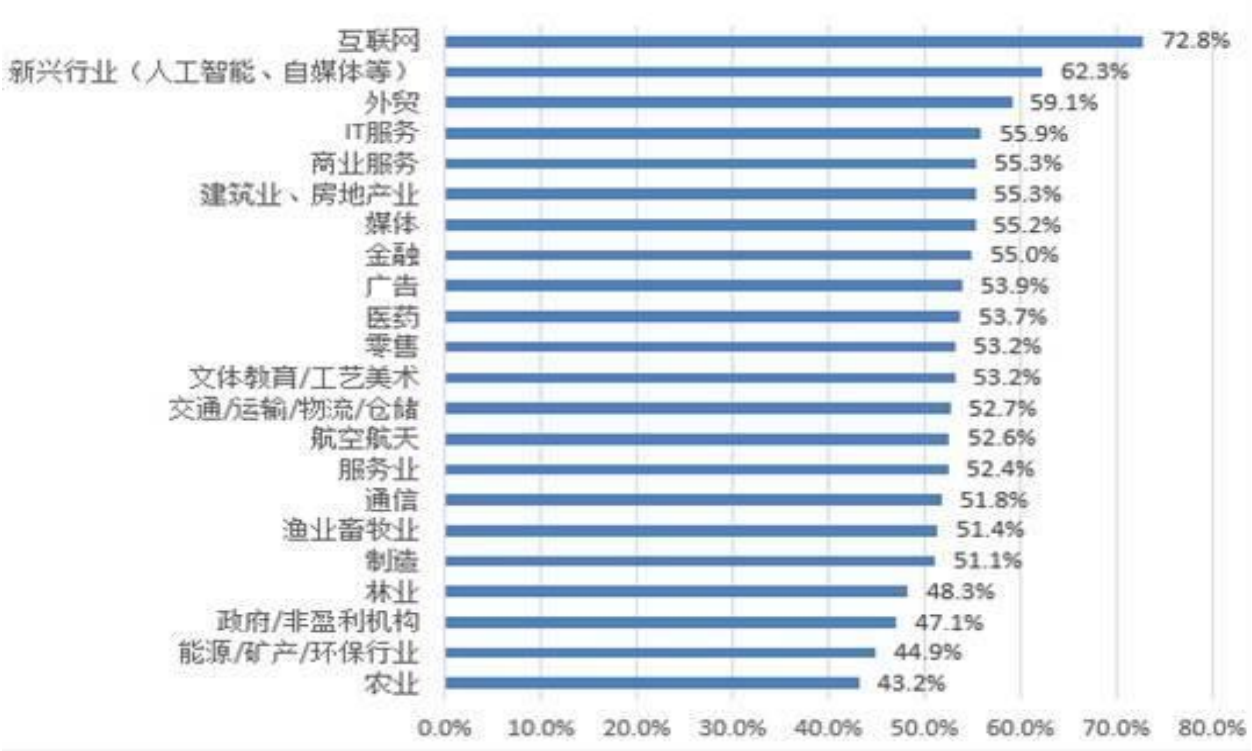
图3 各职业收入信心情况统计

自主创业人群信心最高，这个人群中有高达62%的人认为2018年收入会增加。

4.1、 新闻媒体中的统计思维

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

- 关于收入信心的案例



调查发现，各行业收入信心排在第一的是互联网行业，而紧随其后的就是新兴行业（如人工智能、自媒体等），排在第三位的是外贸行业。

4.1、 新闻媒体中的统计思维

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

➢ 关于收入信心的案例

《中国经济生活大调查（2017-2018）》
中央电视台财经频道主办

问题如：
您预计您的收入2018年会比2017年？

您认为影响幸福生活的主要因素是什么？

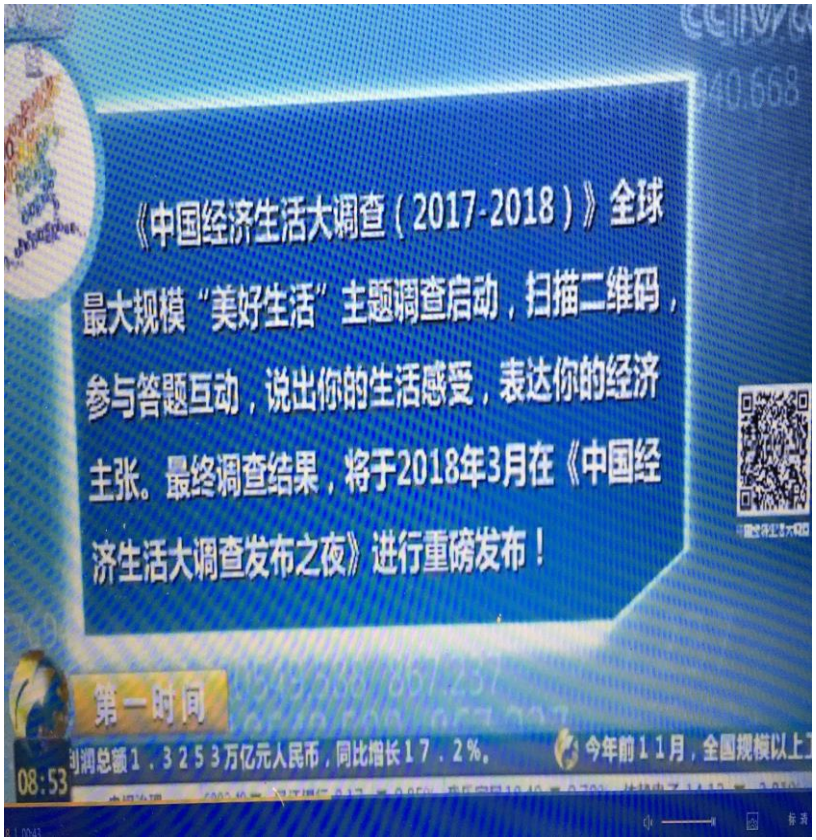
统计知识——问卷调查：

问卷设计；

调查实施方式；

调查数据整理和统计分析（包括作图）；

提出建议。

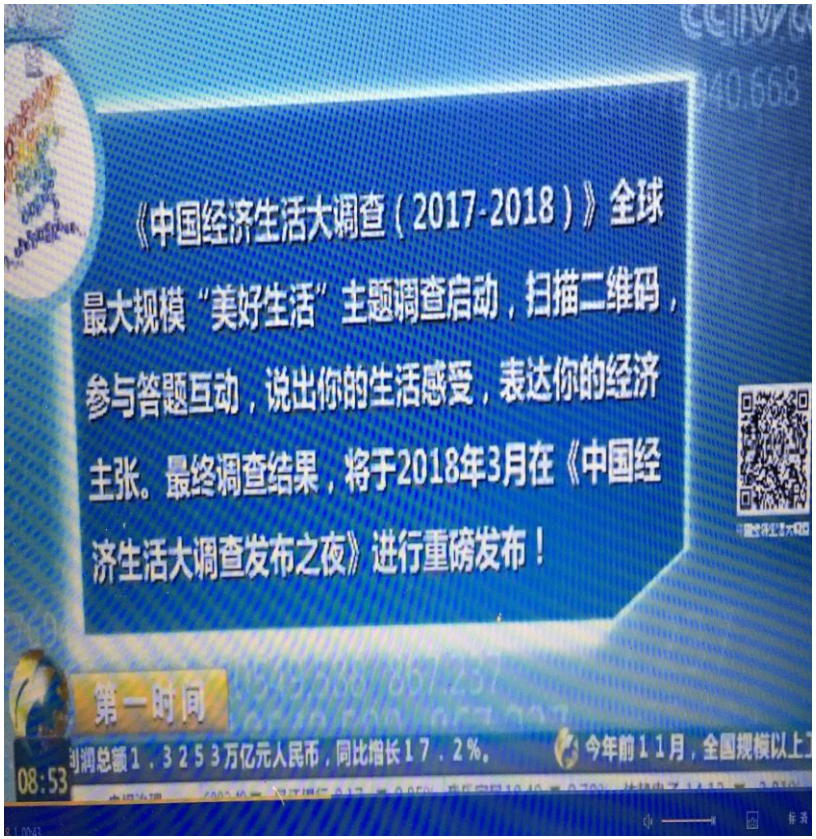


4.1、 新闻媒体中的统计思维

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

➢ 关于收入信心的案例

- 做决策要有数据，每一项数据，都可能是有用的信息。
- 统计学家的本事要能发挥，就得有善用信息的思维。



4.2、 日常生活中的统计思维

➢ 老年人身心健康自评与生活方式

案例背景*

根据2010年第六次全国人口普查详细汇总资料计算，我国人口平均预期寿命达到74.83岁。（2016年上海市人口预期寿命为83.18岁，其中男性80.83岁，女性85.61岁#）

表2.4.1 平均预期寿命变化

单位：岁

年份	合计	男	女	男女之差
1981	67.77	66.28	69.27	-2.99
1990	68.55	66.84	70.47	-3.63
2000	71.40	69.63	73.33	-3.70
2010	**74.83**	72.38	77.37	-4.99

*来自国家统计局网页

http://www.stats.gov.cn/tjsj/tjgb/rkpcgb/qgrkpcgb/201209/t20120921_30330.html

#来自上海市老龄科学研究中心网页http://www.shrca.org.cn/76

4.2、 日常生活中的统计思维

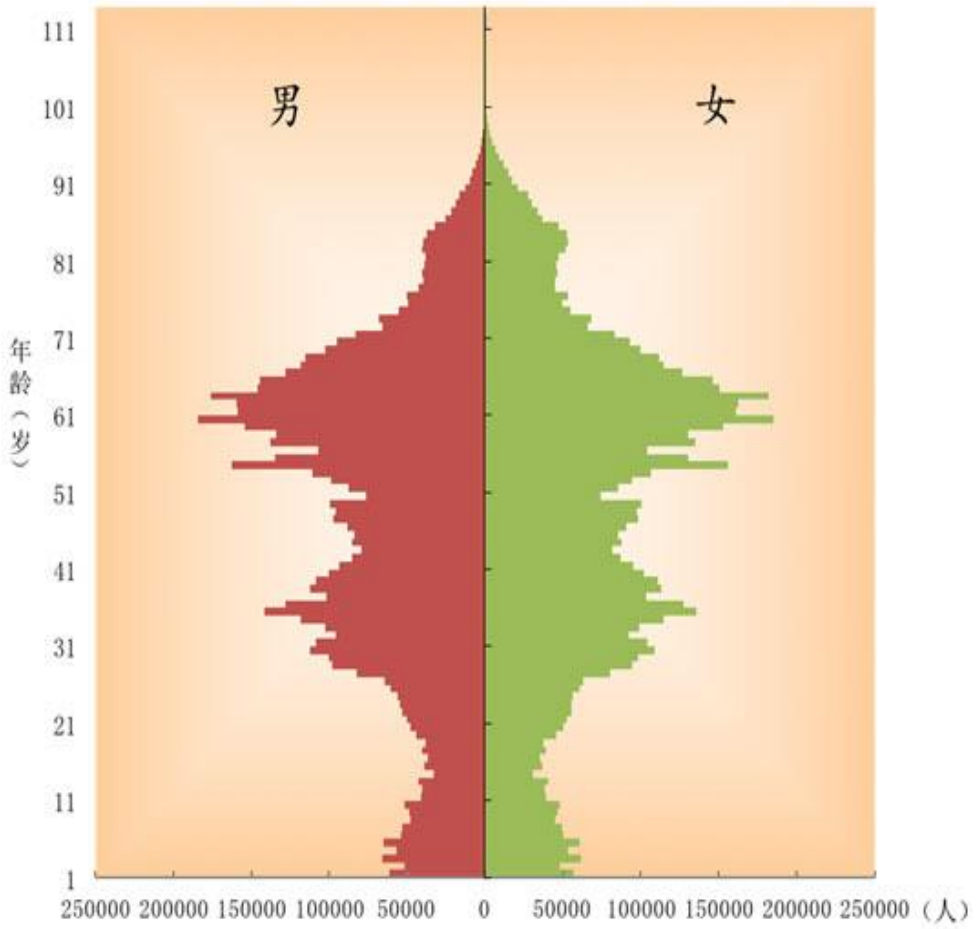
➢ 老年人身心健康自评与生活方式

老年人口基本信息

老年人口总量截至2016年12月31日：

◆全市户籍人口1449.98万人。

◆60岁及以上老年人口457.79万人，占总人口的31.6%。



4.2、 日常生活中的统计思维

➢ 老年人身心健康自评与生活方式

案例问题

关心老年人的身心健康与生活方式的关系。

问题1. 老年人身心健康如何评定、如何量化；

问题2. 老年人生活方式如何量化。

4.2、 日常生活中的统计思维

➢ 老年人身心健康自评与生活方式

案例分析

问题1分析：

考虑身心健康自评。采用问卷调查

采用入户与社区现场调查相结合的方法，对上海市18个社区60岁以上266位老年人，进行书面问卷调查，对象多为中老龄老人。

因变量为身心健康自评，对14项影响身心健康自评的变量经综合得到身心健康自评总分，分值越高，身心健康自评水平就越高。

4.2、 日常生活中的统计思维

➢ 老年人身心健康自评与生活方式

问题2分析：

生活方式的量化，考虑通过3个变量来反映。

“日常生活习惯”	“饮食规律”、“起居规律”、“吸烟”和“喝酒”
“日常体力活动”	“周锻炼次数”、“锻炼强度”、“出行购物”“室内家务”
“日常休闲活动”	“书报电视”、“社区活动参与”、“棋牌麻将”、“宗教佛事”

4.2、 日常生活中的统计思维

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

➢ 老年人身心健康自评与生活方式

案例分析

回归分析——建立多元回归模型：

因变量为身心健康自评

自变量为生活方式变量与社会人口学特征、个体客观身体健康等控制变量在内的五个变量集。

“日常生活习惯”
“日常体力活动”
“日常休闲活动”
“社会人口学变量”
“个体身体健康”

“年龄”、“性别”、“居住方式”、“教育程度”

“慢性病”、“日常疲劳感”、“感冒腹泻”、“自理能力”

4.2、 日常生活中的统计思维

➢ 老年人身心健康自评与生活方式

使用回归分析的思维解决日常生活中关心的问题：通过探索分析上海地区老龄人群日常生活方式的各种因素变化于身心健康自评程度的相联性，可为建设和谐社区的政策制定及引导老年人实现个体的成功老龄化提供理论依据，也为步入老龄化进程的其它地区提供借鉴。

4.3、 历史中的统计思维

➢ 中国皇帝寿命规律影响因素探析

近年来历史题材的电视剧深得百姓喜爱，皇帝寿命之谜备受关注，民间流传的“皇帝短命之谜”是真是假？为什么清高宗乾隆能活到89岁，南朝梁武帝萧衍能活到86岁，唐朝女皇武则天能活82岁，但是汉殇帝刘隆只有一岁，汉冲帝刘炳只活到3岁？

4.3、 历史中的统计思维

➢ 中国皇帝寿命规律影响因素探析

我们首先考虑可能有如下影响皇帝寿命的因素：

（1）王朝因素，即王朝持续时间和皇帝登基年龄；

（2）民族因素，即皇帝所属民族；

（3）遗传因素，即皇帝生父寿命；

（4）子嗣因素，即儿子数量；

（5）战争因素，即皇帝在位期间国家统一情况和战争次数；

（6）地理因素，即都城位置。

所以搜集了包含从秦朝到清朝28个王朝211位皇帝的相关数据并做了如下分析。

4.3、 历史中的统计思维

- 中国皇帝寿命规律影响因素探析

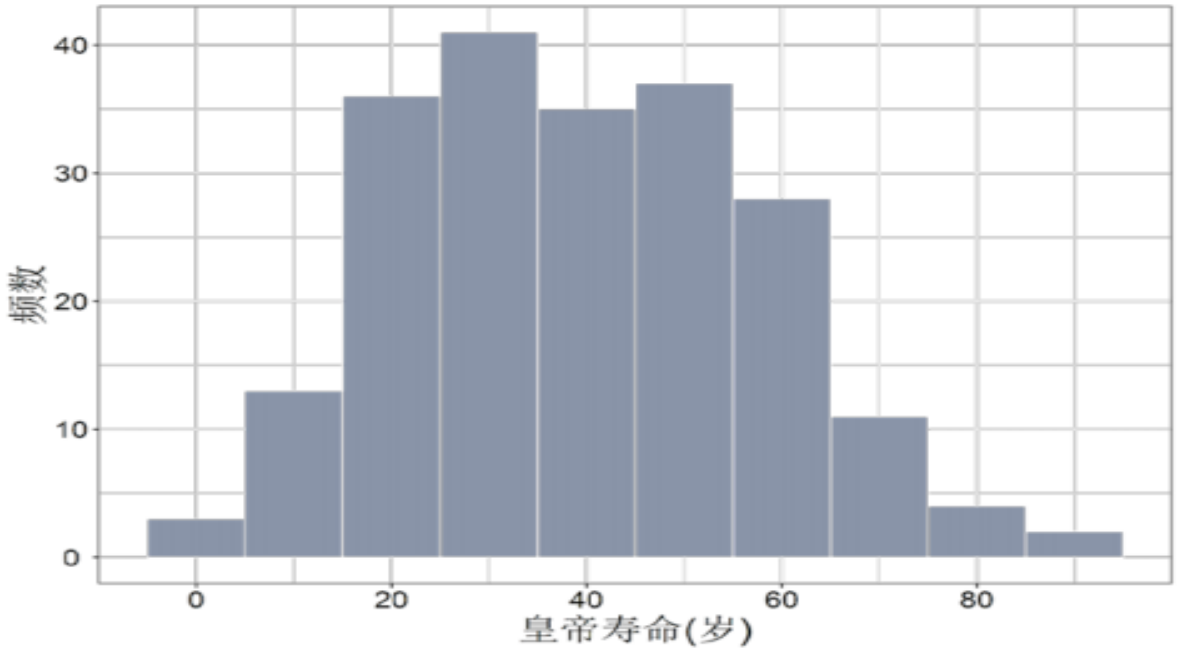


图 皇帝寿命分布图

4.3、 历史中的统计思维

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

➢ 中国皇帝寿命规律影响因素探析

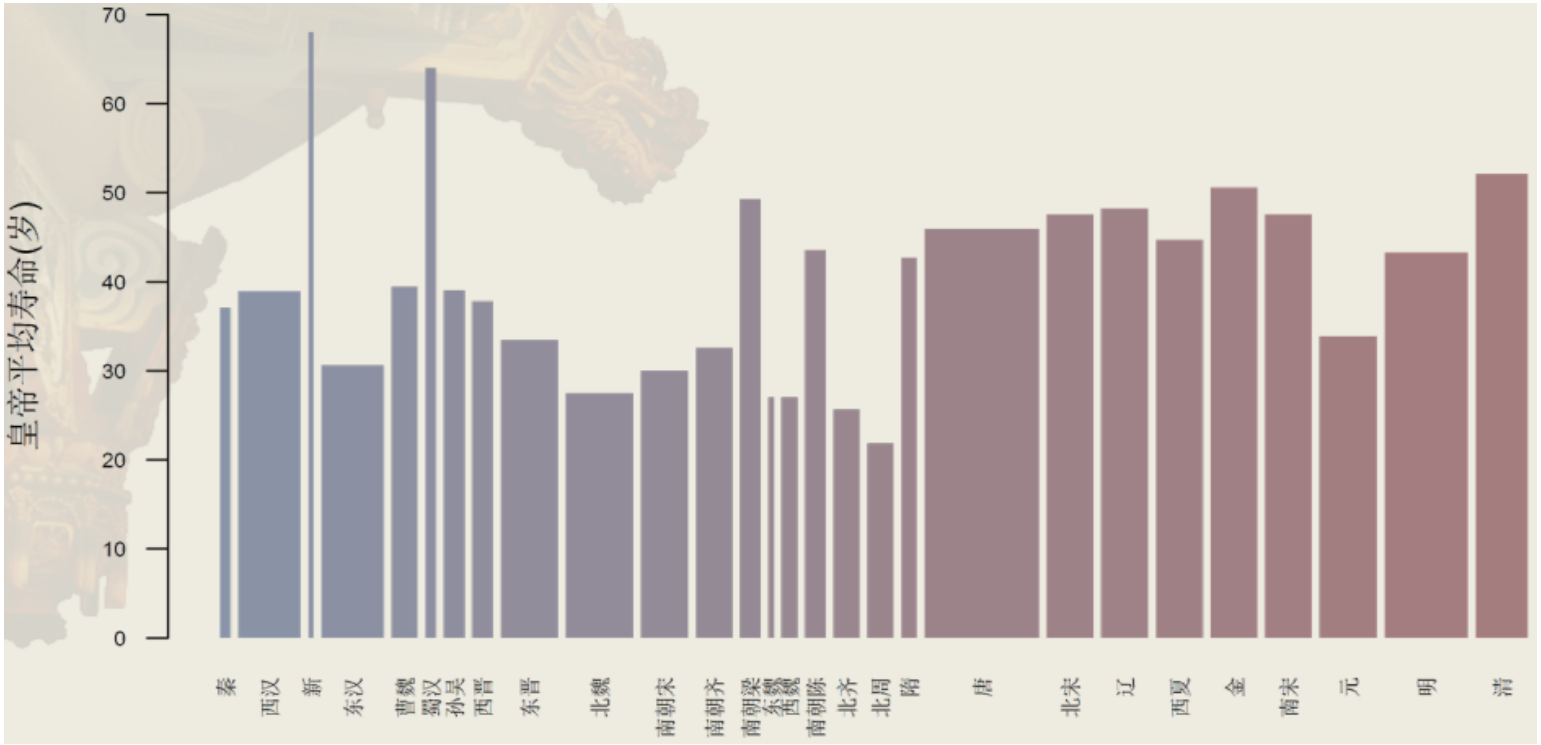


图 不同朝代皇帝的平均寿命

4.3、 历史中的统计思维

➢ 中国皇帝寿命规律影响因素探析

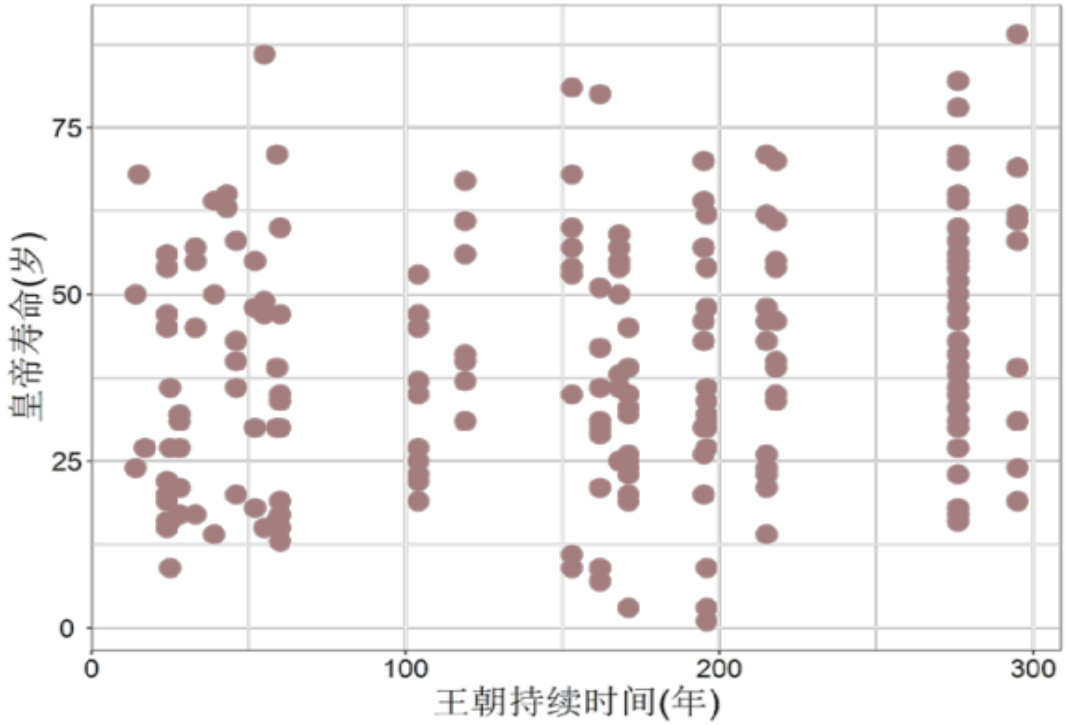


图 皇帝寿命与王朝持续时间关系图

王朝持续时间与皇帝寿命的相关系数是0.313，它们呈正相关。

4.3、 历史中的统计思维

- 中国皇帝寿命规律影响因素探析

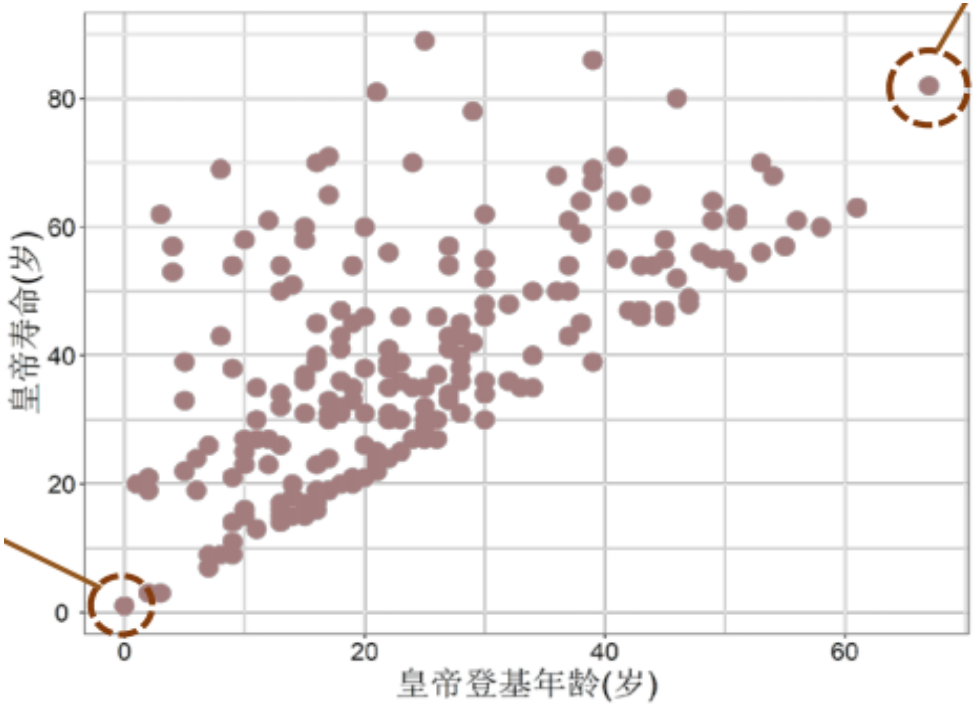


图 皇帝寿命皇帝登基年龄关系图

皇帝登基年龄与其寿命的相关系数为0.598，呈正相关。

4.3、 历史中的统计思维

➢ 中国皇帝寿命规律影响因素探析

王朝	民族	皇帝数
秦、西汉、新、东汉、蜀汉、曹魏、孙吴、西晋、东晋、南朝宋、南朝齐、南朝梁、南朝陈、北齐、隋、唐、北宋、南宋、明	汉族	141
北魏、东魏、西魏、北周	鲜卑族	22
辽	契丹族	9
金	女真族	9
西夏	党项族	9
元	蒙古族	11
清	满族	10

4.3、 历史中的统计思维

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

➢ 中国皇帝寿命规律影响因素探析

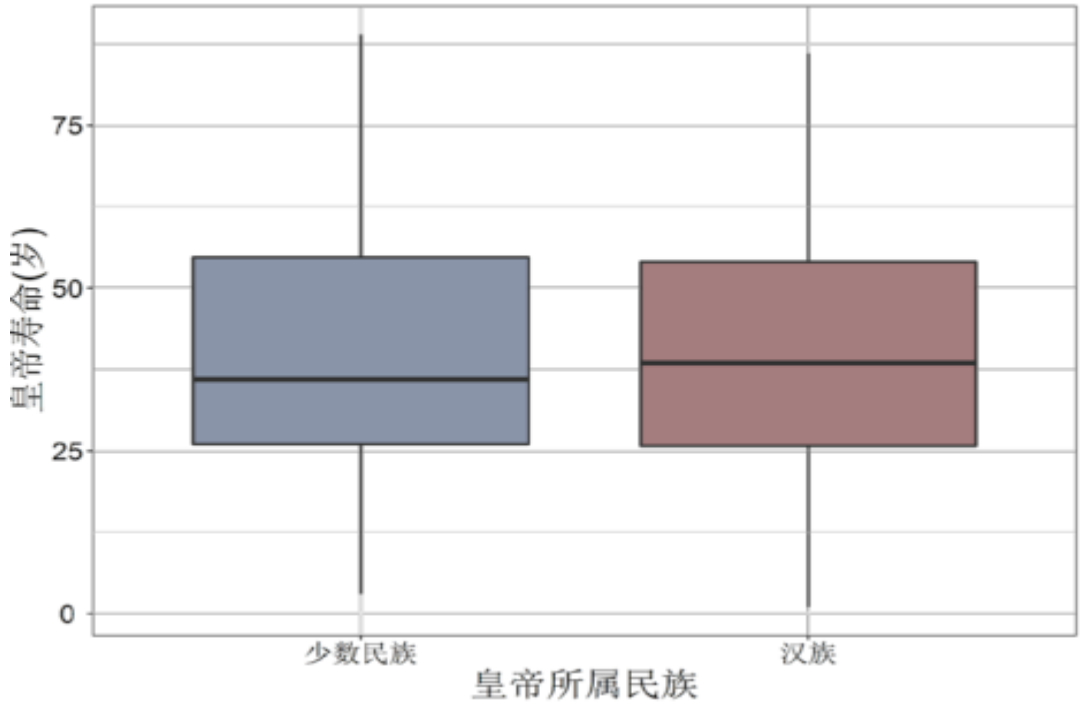


图　皇帝寿命与皇帝所属民族关系

可以看出，平均而言汉族皇帝的寿命略高于少数民族皇帝。

4.3、 历史中的统计思维

➢ 中国皇帝寿命规律影响因素探析

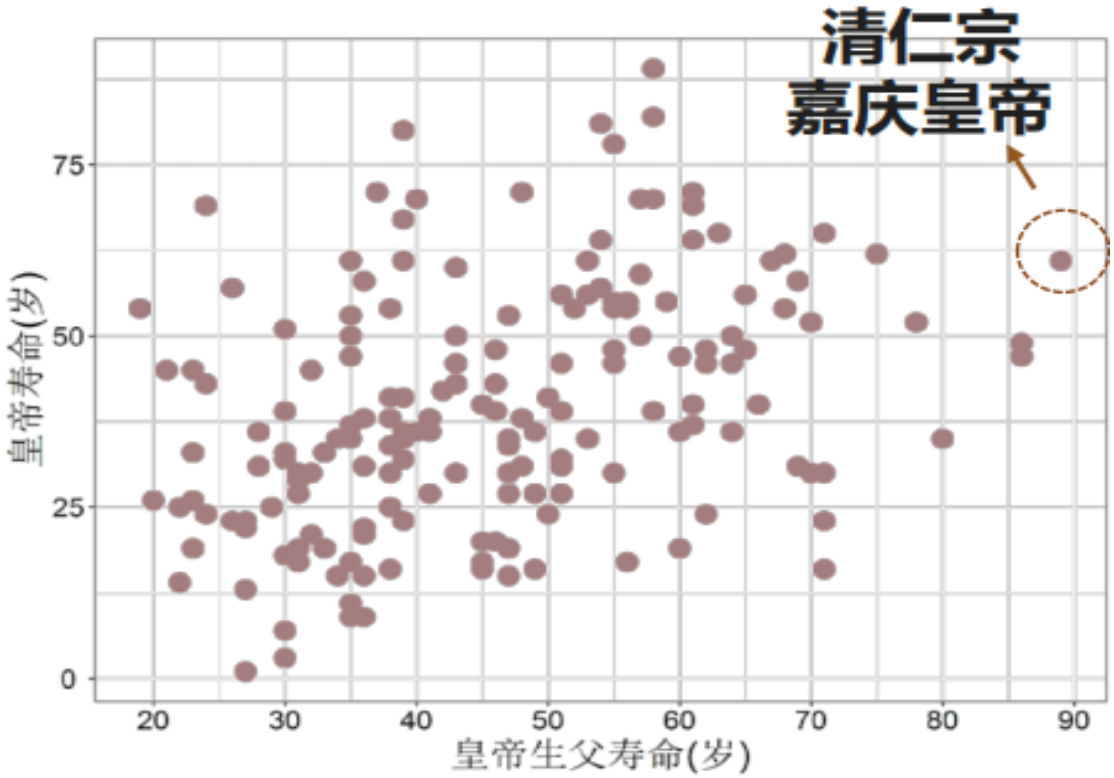


图 皇帝寿命与皇帝生父寿命关系图

皇帝生父寿命与皇帝寿命的相关系数为0.407，呈正相关。

4.3、 历史中的统计思维

➢ 中国皇帝寿命规律影响因素探析

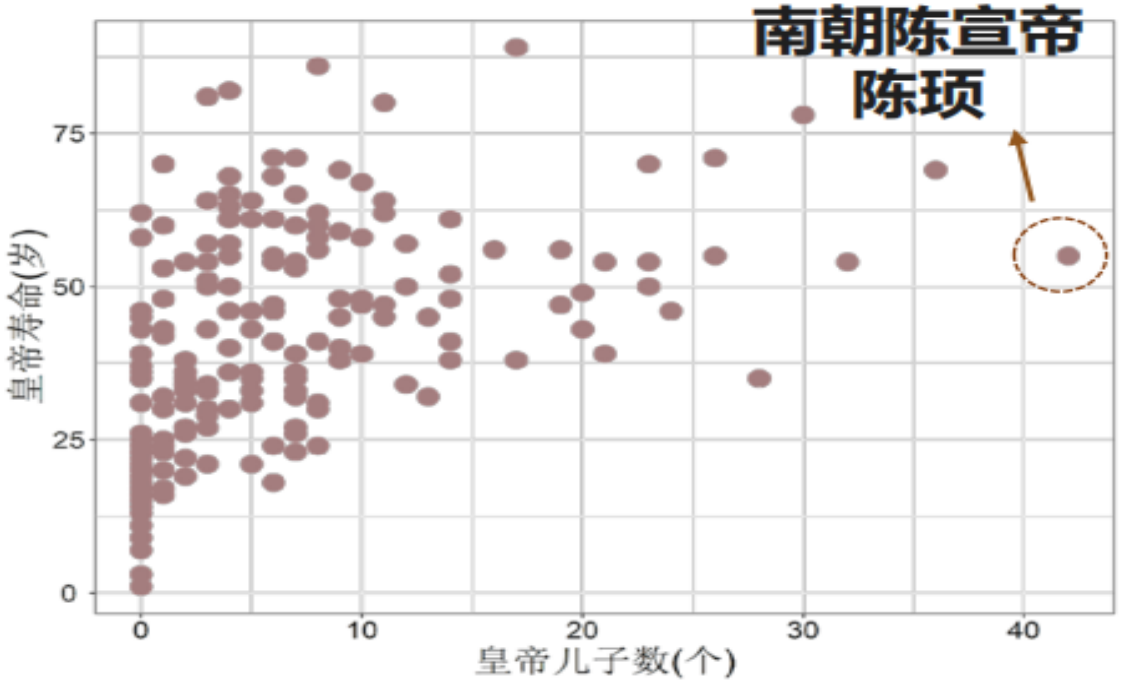


图 皇帝寿命与皇帝儿子数的关系图

皇帝寿命与其儿子数量的相关系数为0.507，呈正相关关系。

4.3、 历史中的统计思维

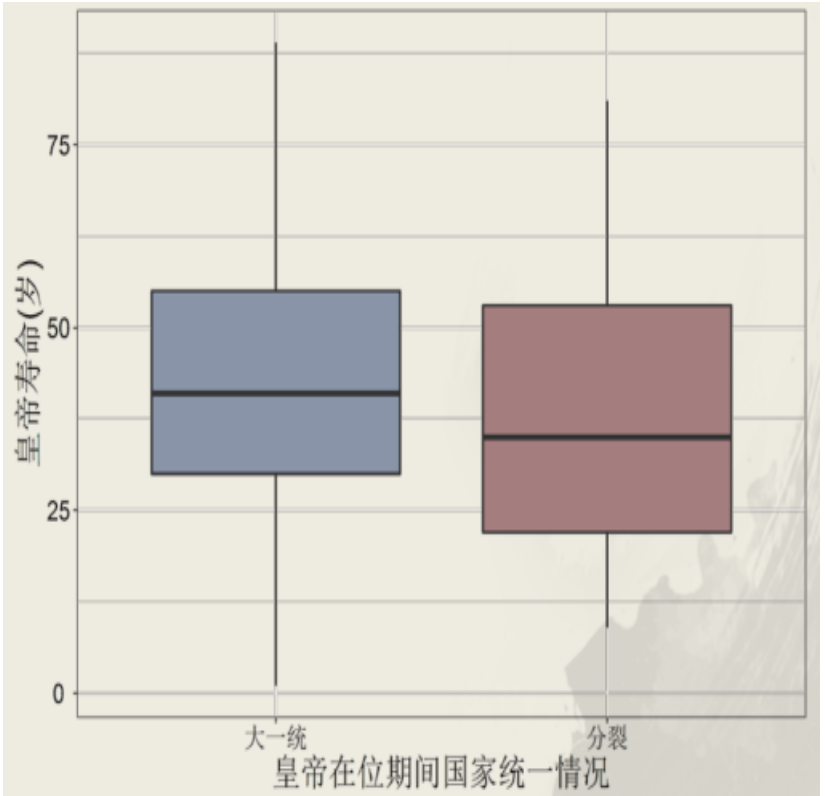
上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

➢ 中国皇帝寿命规律影响因素探析

表 历史上大一统和分裂的王朝

大一统（10朝共93帝）	分裂（18朝共118帝）
秦、西汉、新、东汉、西晋、隋、唐、元、明、清	曹魏、蜀汉、孙吴、东晋、南朝宋、南朝齐、南朝梁、南朝陈、北魏、东魏、西魏、北齐、北周、北宋、南宋、辽、西夏、金

图 皇帝寿命与国家统一情况关系图



平均而言，大一统时期皇帝寿命略高于分裂时期。

4.3、 历史中的统计思维

- 中国皇帝寿命规律影响因素探析

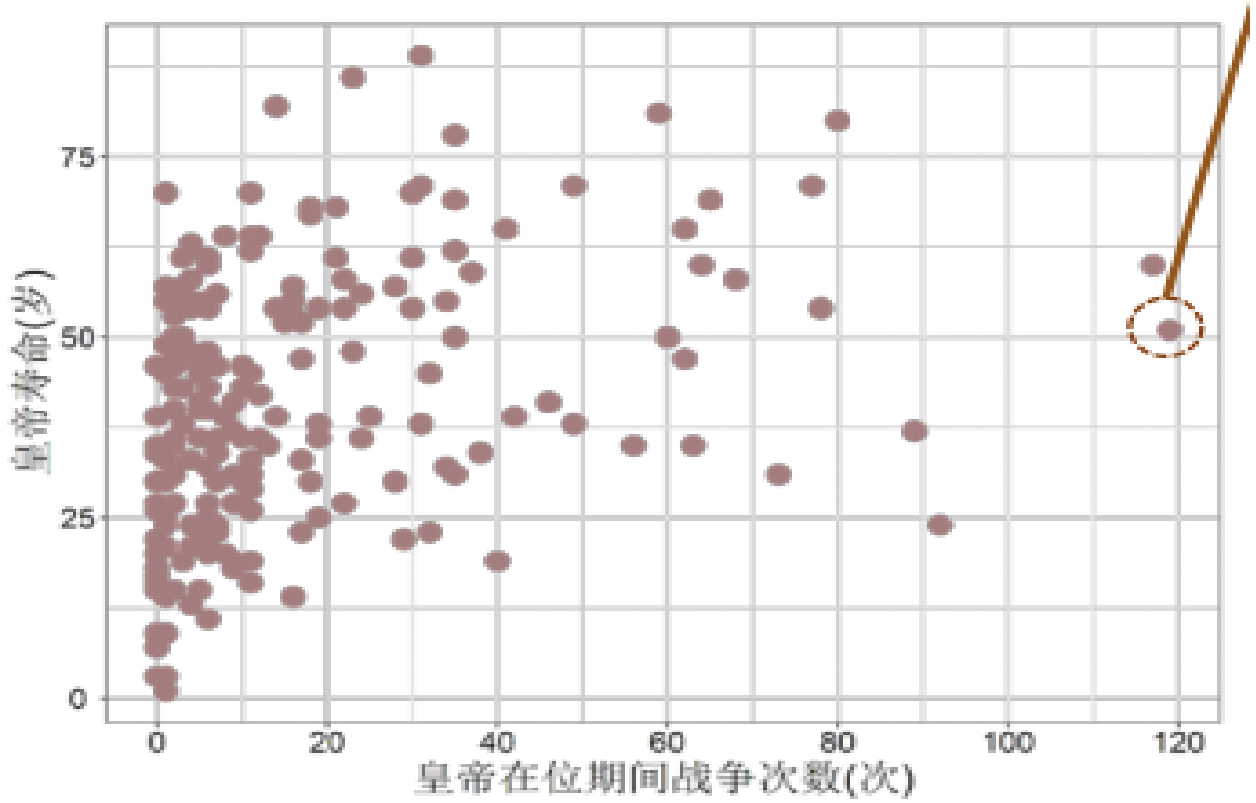


图 皇帝寿命与皇帝在位期间战争次数关系图

皇帝在位期间战争次数与其寿命的相关系数是0.373，有正相关关系。

4.3、 历史中的统计思维

➢ 中国皇帝寿命规律影响因素探析

今地名	古地名	建都王朝	地区	皇帝数
江苏南京	建业、建康、应天	孙吴、东晋、南朝宋、南朝齐、南朝梁、南朝陈	低纬度	41
四川成都	成都	蜀汉		2
浙江杭州	临安	南宋		9
陕西西安	长安、大兴	秦、西汉、新、西魏、北周、隋、唐	中纬度	48
河南洛阳	洛阳	东汉、曹魏、西晋、北魏		29
山西大同	平城	北魏		5
河北临漳	邺	东魏、北齐		6
河南开封	开封	北宋、金		11
北京	中都、大都、北京	金、元、明、清		39
宁夏银川	兴庆府	西夏		9
内蒙巴林左旗	上京	辽	高纬度	9
黑龙江阿城	会宁	金		3

4.3、 历史中的统计思维

➢ 中国皇帝寿命规律影响因素探析

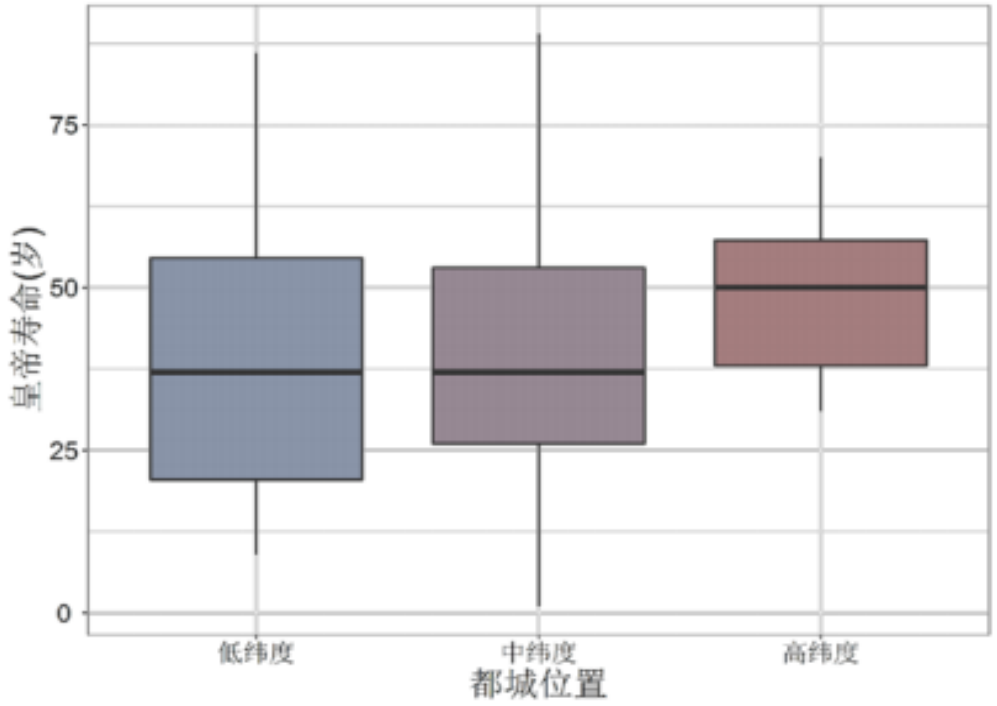


图 皇帝寿命与都城位置关系图

平均而言都城建在高纬度的皇帝的寿命高于都城建在中纬度和低纬度。

4.3、 历史中的统计思维

➢ 中国皇帝寿命规律影响因素探析

表皇帝寿命(取对数)与相关变量线性回归结果

自变量	回归系数	P值	exp（回归系数）	备注
常数项	2.334	<0.001	10.319	
王朝持续时间	0.001	0.003	1.001	
皇帝登基年龄	0.022	<0.001	1.022	
皇帝所属民族是否为汉族	0.000	0.996	1.000	基准：少数民族
皇帝生父寿命	0.004	0.096	1.004	
皇帝儿子数	0.016	0.001	1.016	
皇帝在位期间国家是否统一	0.303	0.002	1.354	基准：国家统一
皇帝在位期间战争次数	0.009	<0.001	1.009	
皇帝在位时期都城位置-高纬度	0.295	0.094	1.343	基准:低纬度
皇帝在位时期都城位置-中纬度	0.195	0.052	1.215	
F值=26.19，P值<0.001				

4.3、 历史中的统计思维

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

➢ 中国皇帝寿命规律影响因素探析

结论：

（1）王朝因素：王朝多持续1年，皇帝平均寿命增长到原来的1.001倍；皇帝登基年龄每增加1岁，皇帝平均寿命增加到原来的1.022倍；

（2）民族因素：皇帝所属民族不是皇帝寿命的重要解释因素；

（3）遗传因素：皇帝生父寿命增加1岁，皇帝的平均寿命增加到原来的1.004倍；

（4）子嗣因素：皇子数增加1，皇帝的平均寿命增加到原来的1.016倍；

（5）战争因素：国家分裂时期皇帝的平均寿命是国家统一时期皇帝平均寿命的1.354倍；皇帝在位期间战争次数增加1次，皇帝的平均寿命增加到原来的1.009倍；

（6）地理因素：都城处在高纬度的皇帝的平均寿命是都城处在低纬度皇帝平均寿命的1.343倍；都城处在中纬度的皇帝的平均寿命是都城处在低纬度皇帝平均寿命的1.215倍。

4.3、 历史中的统计思维

➢ 中国皇帝寿命规律影响因素探析

人们对历史上发生的事情，总想探究其发生的原因，以及探究其影响因素，但又不能完全掌握。在随机世界里，必然性使人们愿意事先好好准备，而不确定性则使人们对未来，充满着盼望或者恐惧。因此要有善于捕捉不确定性的思维，做到以史为鉴。

4.4、 文学中的统计思维

➢ 《红楼梦》作者考证案例

《红楼梦》是我国古典小说的巅峰之作，在世界的影响也是极大的，吸引了不少国内外“红学”研究者。红楼梦对当时的社会风气、礼仪、建筑结构、历史政治、文化、家具服饰、医药等都有相近的描写和影射，不同的人从不同的角度去研究《红楼梦》，都有一定的价值。自其问世以来，研究者甚众，研究领域甚广，从索引、考证到人物形象、艺术成就等诸多方面成果丰硕，但是对于《红楼梦》的作者到底是谁的问题，至今没有达到一致的意见。

4.4、 文学中的统计思维

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

➢ 《红楼梦》作者考证案例

对于《红楼梦》的作者问题，研究者的普遍观点有两个：
1.《红楼梦》前80回与后40回均出自曹雪芹之手；
2.《红楼梦》的前80回是出自曹雪芹，后40回是高鹗续写。

如何把文学作品的特征用量化指标概括？如何从数理统计的角度来研究《红楼梦》著作权问题？

4.4、 文学中的统计思维

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

➢ 《红楼梦》作者考证案例

以卫晶淼“统计分析《红楼梦》代词使用特色及作者辨析”一文为例。

代词系统可以说是比较能反映一种语言某个时期语法特点的一个方面，而且代词在文章中可替代性强，所以能够很好地反映出作者的语言习惯和叙事风格。

通过字数统计，知《红楼梦》前八十回的总字数为950725字，后四十回为275019字。

4.4、 文学中的统计思维

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

➢ 《红楼梦》作者考证案例

各词项在前八十回和后四十回出现的总次数如下：

表 5-1 前八十回词项出现次数统计值

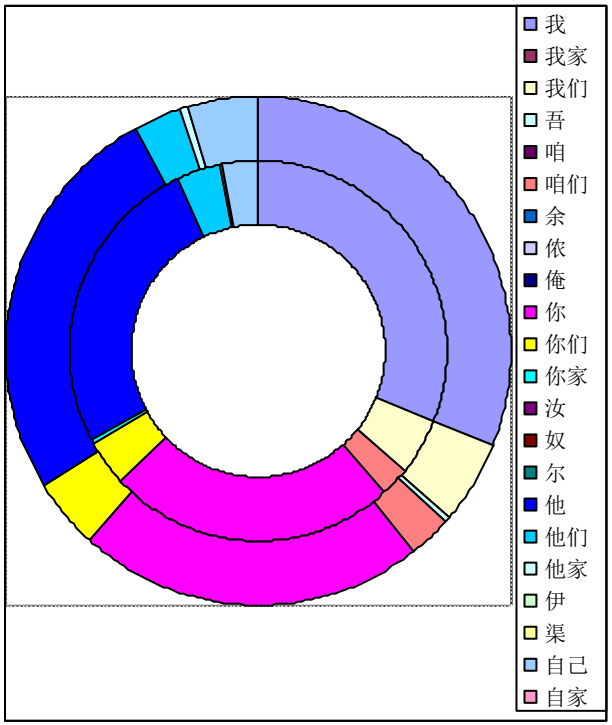
词项	我	我家	我们	吾	咱	咱们	余
出现次数	5403	21	807	16	1	404	7
词项	侬	俺	你	你们	你家	汝	奴
出现次数	10	4	4300	687	15	8	2
词项	尔	他	他们	他家	伊	渠	自己
出现次数	13	4506	615	53	2	0	503
词项	自家	这	此	那	彼	每	各
出现次数	12	5204	1034	2929	78	252	397
词项	谁	孰	甚	什么	何	如何	何妨
出现次数	659	2	30	1166	564	214	17
词项	多少	早晚	怎么	怎样	怎		
出现次数	105	34	643	13	19		

表 5-2 后四十回词项出现次数统计值

词项	我	我家	我们	吾	咱	咱们	余
出现次数	2380	15	402	3	0	216	1
词项	侬	俺	你	你们	你家	汝	奴
出现次数	0	1	1707	336	6	3	1
词项	尔	他	他们	他家	伊	渠	自己
出现次数	36	1980	236	27	7	0	331
词项	自家	这	此	那	彼	每	各
出现次数	5	2422	334	1985	13	29	185
词项	谁	孰	甚	什么	何	如何	何妨
出现次数	220	2	14	731	191	47	2
词项	多少	早晚	怎么	怎样	怎		
出现次数	49	14	487	28	24		

4.4、 文学中的统计思维

➢ 《红楼梦》作者考证案例



按照人称代词、指示代词、疑问代词分类后，画出各个词项在其所属的类别中的比例的圆环图，可以比较直观的看到多数词在前八十回和后四十回中的比例是有很明显差异的，但是也有少数无明显差异的词汇。删除出现次数较少的词项，如：尔、吾、伊等等。计算剩下的词项所占比例的均值和方差。

4.4、 文学中的统计思维

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

- 《红楼梦》作者考证案例

表 5-3　前八十回中词项比例的统计量

词项	我	我们	咱们	你	你们	他	他们
均值	0.009727	0.001414	0.000716	0.007818	0.001193	0.008219	0.001064
方差	0.003997	0.000899	0.000545	0.003541	0.000861	0.002964	0.000848
词项	自己	这	此	那	彼	每	各
均值	0.000915	0.009391	0.001952	0.001210	0.000461	0.000724	0.001195
方差	0.000523	0.002375	0.001410	0.000460	0.000594	0.000647	0.000647
词项	谁	什么	何	如何	多少	怎么	怎
均值	0.002096	0.001044	0.000389	0.000187	0.001151	0.005355	0.000148
方差	0.001101	0.000812	0.000331	0.000194	0.000722	0.001610	0.000174

表 5-4　后四十回中词项比例的统计量

词项	我	我们	咱们	你	你们	他	他们
均值	0.008690	0.001460	0.000771	0.006200	0.001200	0.007120	0.000850
方差	0.002284	0.000815	0.000569	0.001982	0.000562	0.002767	0.000511
词项	自己	这	此	那	彼	每	各
均值	0.001210	0.008720	0.001220	0.007120	0.000047	0.000115	0.000672
方差	0.000612	0.001767	0.000783	0.001815	0.000088	0.000171	0.000532
词项	谁	什么	何	如何	多少	怎么	怎
均值	0.000781	0.002648	0.000709	0.000178	0.000188	0.001761	0.001950
方差	0.000495	0.001078	0.000619	0.000201	0.000177	0.000718	0.000740

4.4、 文学中的统计思维

➢ 《红楼梦》作者考证案例

案例解析

点估计：用样本统计量来估计总体参数，因为样本统计量为数轴上某一点值，估计的结果也以一个点的数值表示，所以称为点估计。点估计的精确程度用置信区间表示。

区间估计：从点估计值出发，按给定的概率值建立包含待估计参数的区间。其中这个给定的概率值称为置信度或置信水平，这个建立起区间估计来的包含待估计参数的区间称为置信区间，指总体参数值落在样本统计值某一区内的概率；而置信区间是指在某一置信水平下，样本统计值与总体参数值间误差范围。划定置信区间的两个数值分别称为置信下限和置信上限。

4.4、 文学中的统计思维

➢ 《红楼梦》作者考证案例

以前八十回为原总体的样本，估计出的置信区间，和以后四十回为新总体的样本，估计出的均值比对，发现：

1. 只有“我们、咱们、你们、多少”五个词项的新总体样本的点估计值落在了原总体样本以99%的置信度估计出的置信区间内，可以认为两个样本的均值没有明显差别，不能排除前八十回和后四十回来自同一总体的可能性；

2. 其他词项的新总体样本的点估计值均不能落在原总体样本以95%的置信度估计出的置信区间内，所以可以认为两个样本的均值有明显差别，排除前八十回和后四十回来自同一总体的可能性。

4.4、 文学中的统计思维

- 《红楼梦》作者考证案例

表 5-5 估计均值落在置信区间内的统计量

词项	我们	咱们	你们	多少
置信下限	0.00120	0.00059	0.00103	0.00013
估计均值	0.00141	0.00072	0.00119	0.00019
置信上限	0.00172	0.00095	0.00138	0.00024

表 5-6 估计均值小于置信下限的统计量

词项	自己	那	什么	怎么	怎
估计均值	0.00092	0.00536	0.00210	0.00115	0.00121
置信下限	0.00101	0.00655	0.00231	0.00153	0.00172
置信上限	0.00140	0.00770	0.00299	0.00199	0.00218

4.4、 文学中的统计思维

- 《红楼梦》作者考证案例

表 5-7　估计均值大于置信上限的统计量

词项	我	你	他	他们	这	此
置信下限	0.00797	0.00557	0.00624	0.00069	0.00816	0.00097
置信上限	0.00941	0.00682	0.00799	0.00101	0.00928	0.00146
估计均值	0.00973	0.00782	0.00822	0.00106	0.00939	0.00195
词项	彼	每	谁	何	如何	
置信下限	0.00002	0.00006	0.00063	0.00051	0.00011	
置信上限	0.00007	0.00017	0.00094	0.00090	0.00024	
估计均值	0.00015	0.00046	0.00120	0.00104	0.00039	

4.4、 文学中的统计思维

➢ 《红楼梦》作者考证案例

也可以直接比较每个代词的使用均值在前八十回与后四十回是否有显著不同。

1. 在显著性水平为0.05时，“我、你、他、他们、自己、这、此、那、彼、每、谁、什么、何、如何、怎么、怎”这十六个代词的使用频率的均值在前八十回的样本与后四十回的样本中差异显著，能够排除其来自同一样本的可能性；

2. “我们”“咱们”“你们”“多少”这四个代词，其使用频率的均值在前八十回与后四十回的样本中差异不显著，不能排除来自同一样本的可能性。

4.4、 文学中的统计思维

➢ 《红楼梦》作者考证案例

从代词的可替代性分析：

由于“我们”“咱们”“你们”“多少”这四个代词可替代性较弱，其使用频率均值在前八十回与后四十回的样本中差异不显著是合理的，根据其他代词有显著性差异可以得出 结论：

《红楼梦》前八十回与后四十回的作者确非一人。

4.4、 文学中的统计思维

➢ 《红楼梦》作者考证案例

统计思维体现在文学中，最直观的即为遣词造句，将词句作为数据的一种，运用统计的知识：频率，均值，置信区间等进行数据分析。

4.5、 经济中的统计思维

➢ 富人是否变得更富了

在“互联网泡沫”的最高峰时期，科技股的价格飙升到了一个前所未有的水平，微软的创始人比尔·盖茨以超过1000亿美元的身家成为世界首富。除了最富裕的18个国家，比尔·盖茨超过了全世界余下国家的国民生产总值之和，不久比尔·盖茨的身家就缩水了，一方面是由于微软股票价格的下跌，另一方面由于他把他所有股份捐给了慈善机构。然而，媒体仍然在持续不断地曝光那些巨富的故事，似乎富人变得更富有了，而我们远远地落在后面。事实真的如此吗？

4.5、 经济中的统计思维

➢ 富人是否变得更富了

表： 按收入高低划分的5组人群（包括最高收入5%的人群）分别占家庭总收入百分比

%	1/5低收入家庭	第二个1/5家庭	第三个1/5家庭	第四个1/5家庭	1/5高收入家庭	5%最高收入家庭
1967	4	10.8	17.3	24.2	43.6	17.2
1977	4.2	10.2	16.9	24.7	44	16.8
1987	3.8	9.6	16.1	24.3	46.2	18.2
1997	3.6	8.9	15	23.2	49.4	21.7
2005	3.4	8.6	14.6	23	50.4	22.2

4.5、 经济中的统计思维

➢ 富人是否变得更富了

问题

1. 如何描述收入分配均衡状况？
2. 从收入高低不同人群占家庭总收入百分比可以得出富人更富了的结论吗？
3. 如果能得出此结论你怎么看？

4.5、 经济中的统计思维

➢ 富人是否变得更富了

问题解析

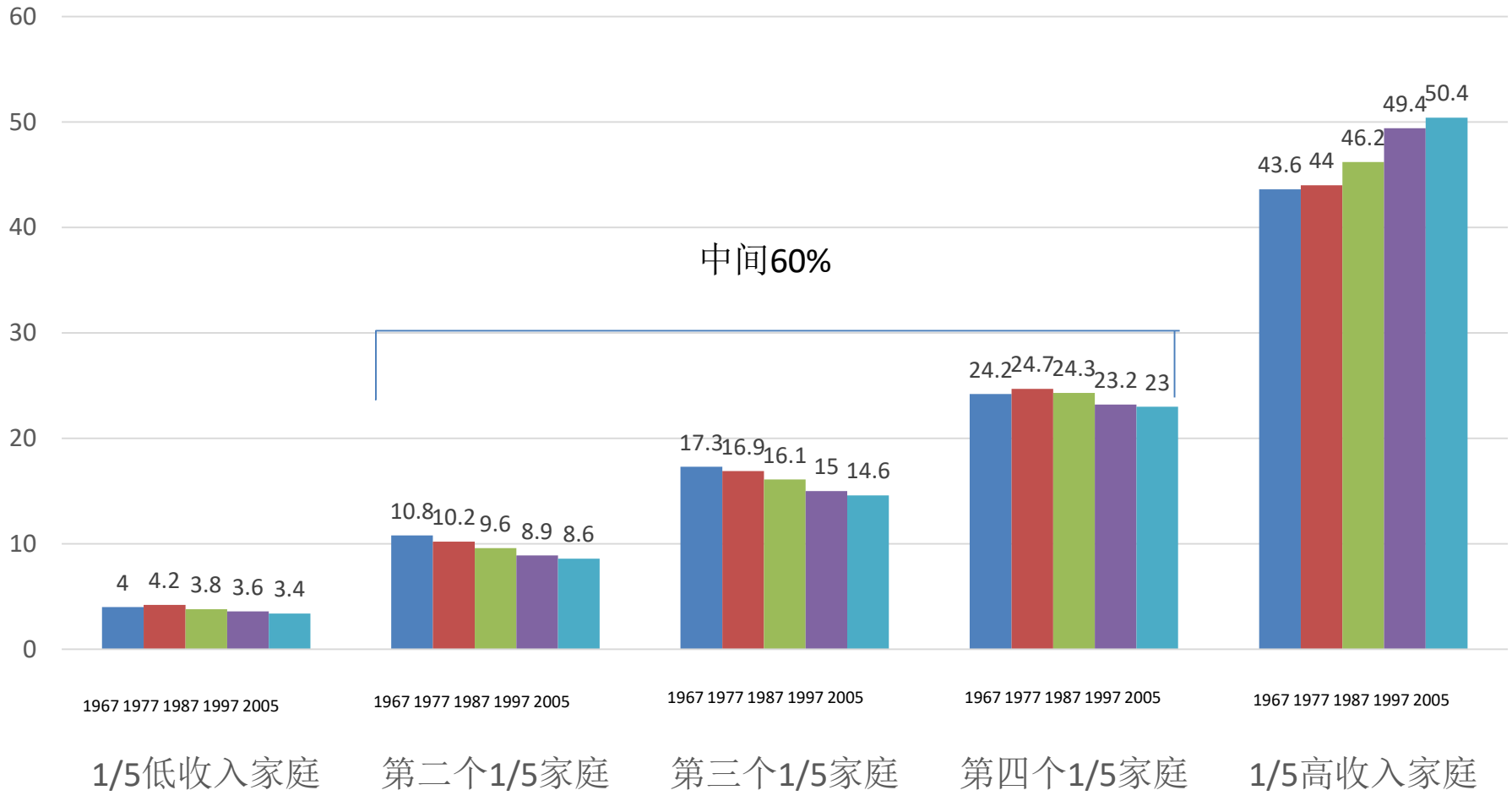
描述收入分配均衡状况的方法：

5分法，即按收入高低区分为5组人群。通常在收入最高的一组中进一步区分出收入最高的5%的群体与其他人群进行比较。

4.5、 经济中的统计思维

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

- 富人是否变得更富了



4.5、 经济中的统计思维

➢ 富人是否变得更富了

结论：
我们可以发现，除了最富裕的20%群体之外的全部群体，所享有的社会收入份额从1967年以来一直在下降。与此同时，最富有的20%的群体所占的收入份额却在稳步增长，与那5%的最富有群体一样。即确认了一个事实：与普通公众相比，富人的确是变得更富了。

4.5、 经济中的统计思维

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

➢ 富人是否变得更富了

对此现象分析：

帕累托法则——二八定律

1897年，意大利经济学者巴莱多偶然注意到19世纪英国人的财富和收益模式。在调查取样中，发现大部分的财富流向了少数人手里。同时，他还从早期的资料中发现，在其他的国家，都发现有这种微妙关系一再出现，而且在数学上呈现出一种稳定的关系。

于是，帕累托从大量具体的事实中发现：社会上20%的人占有80%的社会财富，即：财富在人口中的分配是不平衡的。

同时，人们还发现生活中存在许多不平衡的现象。因此，二八定律成了这种不平等关系的简称。习惯上，二八定律讨论的是顶端的20%，而非底部的80%。

4.5、 经济中的统计思维

➢ 富人是否变得更富了

帕累托二八效率（最优）法则：

只要在让一部分人变得更好的同时并没有让其他人变得更糟，那么这种改变就是好的。

不断增长的美国经济是符合帕累托法则的，因为尽管绝大多数人群在收入份额中所占的比例要比过去要小，但是每个人的绝对财富以及生活条件要比过去要好。

今天的富人也与过去不同。过去你想成为有钱人，只能靠出身，而现在，你可以通过接受良好的教育以及努力工作获得财富。因此，这可以促进人们更好地受教育和努力工作。

4.5、 经济中的统计思维

- 富人是否变得更

分析我国的收入分配均衡情况

6-2 全国居民按收入五等份分组的人均可支配收入

单位：元

组 别	2013	2014	2015	2016
低收入户 (20%)	4402.4	4747.3	5221.2	5528.7
中等偏下户 (20%)	9653.7	10887.4	11894.0	12898.9
中等收入户 (20%)	15698.0	17631.0	19320.1	20924.4
中等偏上户 (20%)	24361.2	26937.4	29437.6	31990.4
高收入户 (20%)	47456.6	50968.0	54543.5	59259.5

6-7 城镇居民按收入五等份分组的人均可支配收入

单位：元

组 别	2013	2014	2015	2016
低收入户 (20%)	9895.9	11219.3	12230.9	13004.1
中等偏下户 (20%)	17628.1	19650.5	21446.2	23054.9
中等收入户 (20%)	24172.9	26650.6	29105.2	31521.8
中等偏上户 (20%)	32613.8	35631.2	38572.4	41805.6
高收入户 (20%)	57762.1	61615.0	65082.2	70347.8

数据来自中国统计年鉴2017

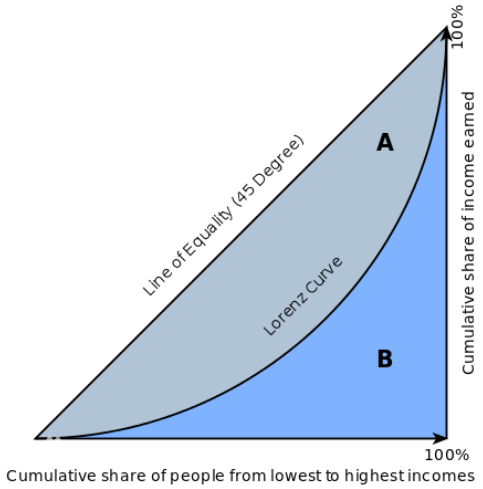
4.5、 经济中的统计思维

➢ 富人是否变得更富了

问题延伸

描述收入分配均衡性——基尼系数

洛伦兹曲线（Lorenz curve），也译为“劳伦兹曲线”。就是，在一个总体（国家、地区）内，以“最贫穷的人口计算起一直到最富有人口”的人口百分比对应各个人口百分比的收入百分比的点组成的曲线。（为了研究国民收入在国民之间的分配问题，美国统计学家（或说奥地利统计学家）M.O.洛伦兹（Max Otto Lorenz，1876- 1959）1907年（或说1905年）提出了著名的洛伦兹曲线。）



设实际收入分配曲线和收入分配绝对平等曲线之间的面积为A，实际收入分配曲线右下方的面积为B。并以A除以（A+B）的商表示不平等程度。这个数值被称为基尼系数或称洛伦茨系数。

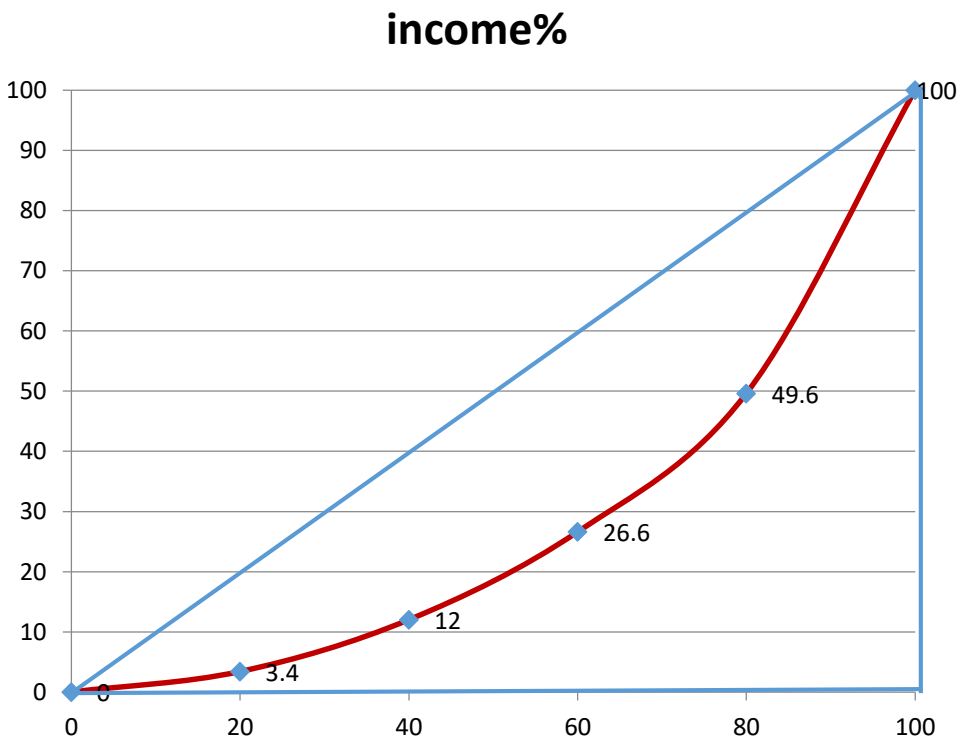
4.5、 经济中的统计思维

➢ 富人是否变得更富了

问题延伸

基尼系数，按照联合国有关组织规定：

低于0.2	收入绝对平均
0.2-0.3	收入比较平均
0.3-0.4	收入相对合理
0.4-0.5	收入差距较大
0.5以上	收入差距悬殊



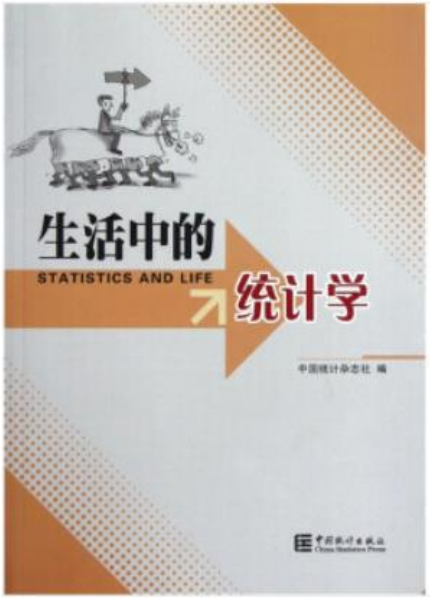
4.5、 经济中的统计思维

➢ 富人是否变得更富了

- 借助统计思维，能够快速知道一组数据呈现的形式和分布。
- 在数据分析中，统计思维就是用统计的相关思维，来解决数据分析的问题。
- 通过查看相关统计量的形式，来了解这组数据的概要，从局部到整体，以点带面地看这组数据的大小，分布以及其他特征。通常的统计量包括了平均值，最大最小值，中位数、百分位数等等。

推荐阅读

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS



1. 中国统计杂志社编，生活中的统计学，中国统计出版社，2010.
2. Jeffrey O.Bennett，William L.Briggs， Mario F.Triola，封文波译，心理统计-日常生活中的统计推理，第3版，机械工业出版社，2013.
3. 杨轶莘编著， 大数据时代下的统计学， 电子工业出版社， 2016.

谢 谢!

Thank You