有3种谎言：谎言、糟糕透顶的谎言和统计资料。

对于追求效率的公民来说，统计思维有一天会和读写能力一样重要。

使我们陷入麻烦的通常并非我们不知道的事情，而是那些我们知道却不正确的事情。

绪论：

1、报纸大肆报道犯罪事件，联想起地区治安不好。

2、服用某种药物能大多数人在一个星期内治愈感冒，却忘了感冒能在一个星期内自动痊愈。

统计能夸大事实、迷惑他人的工具，使用统计一定要了解统计含义。

白衣侠客

# 1内在有偏的样本：

导致有偏原因：调查对象说真话、样本大且代表性、调查方法

1、1924级耶鲁毕业生平均年收入25111美元

（太精确？数字太大？数据真实：1、找到所有毕业生【不是随机抽样，地址不详，能够联系上（有钱），样本不具有代表性】 2、都说真话）

2、阅读杂志 琴师or真实故事

（说真话假定不可靠）

3、美国人每天刷牙1.02次

（不可能保证说真话，不太可能向陌生人承认不经常刷牙。数据没有什么价值，结果表现是美国人说自己每天刷牙多少次，不是真的做到）

4、癌症早期发现能否挽救生命

（存活率上升，可能为数据之前没有记录，只是追踪。后来才有记录）

5、心理医生说实际上所有人都是神经质的。

（样本只是他的病人，不能代表所有人）

抽样有用，一定要样本具有代表性，能排除各种误差。

无形误差和有形误差一样会破坏结果的可信度。

6、总统大选预测，兰登or罗斯福

（样本来自于有电话、订阅报纸的人（富人），不能代表所有人）

7、在街上遇到的第20个人为调查对象

（样本为在街上的人，不能代表这个国家）

8、火车站调查遇到的人类型各样

（样本唯独没有包括婴儿母亲）

随机样本（总体的每个名字或事物都有相同的几率被选进样本）

随机抽样，成本高，最经济的做法是分层抽样（准备好每层所有名单，然后随机抽中样本，代价大）。

街头调查会错过在家的，上门调查错过上班族，晚上访问错过看电影和去夜总会的

67%美国人反对。其实可能是指67%的哪些美国人。

9、调查人员会影响结果（黑人、白人调查组对黑人调查）

很多民意调查的结果因为有偏，所以 没有价值。（民众与民意结果相驳、被操纵）

# 2 精心挑选的平均数

平均数（算数平均数、中位数、众数）、区间估计比某个数值精确、

1、房价变动

（平均数：均值（算术平均数）、中位数（一半家庭是超过，一半家庭少于）、众数（年收入为多少的家庭数远大于其他家庭））2、身高

（均值、中位数差不多）

正态分布（均值、中位数、众数都是同一个值），若是收入，可能出现右偏或左偏

3、公司员工收入

（区别是哪一类平均数）

4、收入攀升

（并不意味着工资率变动，有可能兼职转正式工，工资原本就会增加）

看什么是平均，包括着什么

5、普查局数据

（附上精确的说明，有19/20概率保证真实数据会落在3107美元加减59美元范围内）

6、时代杂志“编者的话”栏目

（年龄时清晰指出为中位数，但是提及收入却没有详细指出，可能利用均值得到读者高收入，吸引广告商）

# 3 没有披露的数据

样本数太小（只用比例）、结合事件概率、原本发生事情、区间比数值精

确、表达不准确、结合自身具体（事件本身目的）情况、

1、多克斯牌牙膏使得蛀牙减少23%

（不充分的样本，仅有12人组成；蛀牙增多和无变化的试验组继续试验，直到得到测试组证明蛀牙减少）

2、可尼斯博士牙粉治愈臼齿有极大功效

（样本仅6个案例）

3、抛硬币

（样本容量少，很难得到正面与反面1:1）

4、小儿麻痹症疫苗

（样本单位足够大，但是发病概率太小 2：450，可能需要当前10-20倍样本容量才能 产生结果）

5、许多不成功的治疗方法未经证实匆匆执行

（像感冒原本就会自动治愈一样）

显著性水平（5%水平，95%可能性正确）证实结论正确（假设检验）

6、美国房产大量建造适合4人居住的房子

（对实际情况知之甚少，只用平均数来衡量 ，更多人口或更少人口没有考虑在内，3、4（45%），1、2（35%），多于4（20%））

7、《星期天》描述孩子将在几月份坐直

（遗漏了数据、应该给定一个正常范围，而不只是一个精确值（期望），实际上很难正好符合精确值，将正常与期望混为一谈）

8、钢铁硬度提高的新闻

（记者并没有理解新闻的意思，就直接从政府新闻中挑出，放到栏目中，遗漏的原因是没有思考）

9、美国农场接上电

（表达不精确，可理解为1/4农场未接上电，接上电可理解为电线从农场铺过，或者在农场之外，并不等于用上电）

10、预测孩子未来长多高

（制作表格，只能知道某个年龄段的平均身高，若是个别例子，每个人成长速度不同，不如从父母或家族身高来分析更为准确）

11、饼干包装纸

（另一张图纵坐标没有数据，比较没有意义）

12、广告公司发展趋势

（没有纵坐标，发展趋势代表什么意思）

13、根据温度选择野营目的地

（只看平均温度，不关注波动）

# 4 毫无意义的工作

样本是否能代表总体、样本结果没有意义（不相匹配的资料）

1、智力测试

（智力测试的内容不能完全反应智商，可能只能反应阅读能力；即使是斯坦福——比奈试验，测试本身只是抽样，存在误差。问题：样本能以多大精度代表总体：

可能误差、标准误差，因为存在可能误差，智商测试结果应该有范围。（类似：步子丈量土地大小））

2、用户喜爱的文章

（45%大于35%，但是样本容量不大没有意义，也许35%只是几个人）

3、老黄金牌香烟

（虽然尼古丁及有害物质含量相对其他的少，但是对人体的危害是相同的。但是他们利用了排名滞后，误导消费者）

# 5 惊人的统计图形

绘图比文字、数字更为直观。注意图表纵坐标起始刻度、刻度单位。对图表直观效果产生影响。

1、折线图 国民收入增长

（抹去横坐标的0，从很大的值开始，上涨趋势的直观感受很不同）

（更改纵坐标单位，100%换成10%，视觉冲击）

2、股票创新高、生活成本减少

3、政府支出急剧上升or保持稳定

4、 柱状图省略中部

# 6 平面图形

为了造成视觉冲击，平面图形有时会误导观众。

1、美国工匠是罗坦提亚木匠工资2倍

（普通柱状图，或者2个钱袋表示 。一个钱袋表示，不仅高两倍，且宽两倍，其实是1:4，若考虑厚度，其实是1:8）

2、钢铁行业的10年前后熔炉比较

（宽度2倍，高度1.5倍，铁是2.5倍，总体是150%增长，视觉却是1500%）

3、美国人长寿了

（人宽度2倍，高度2倍，还是立体的，厚度2倍，总体1:8）

4、美国奶牛增多

（表示产量增长，一头是另一头3倍高，看起来奶牛变大了好多，容易引发误解，奶牛变大了）

5、“减肥”的犀牛

（表示数量减少，一头是另一头几分之几高，看起来犀牛变小了好多，容易引发误解，犀牛变小了 ）

# 7 不相匹配的资料

样本与总体结果无关，原本无关，或者因为早前因为认识浅或者条件有限导致记录少，后来记录增多，时间线比较数据不能直接说明激增。（百分数可以），记录不全，隐藏数据、总体不同的比率比较

1、杀死大量细菌的药品与感冒无关

2、调查白人与黑人是否有相同工作机会

（直接看结果，大多数人回答的是，但是应该观察他人对问题的态度，同情黑人的白人会回答不是，但是歧视黑人的人会回答是。如果种族歧视越严重，反而得到更多人认为有平等工作机会的结果）

3、内科医生香烟品牌调查

（27%内科医生选择喉宝，并不意味者该香烟的危害相对较小，虽然27%可能有含义，却和我们想要的结果无关）

4、榨汁机功能提高

（经过实验室验证，该榨汁机功能提高26%，并得到了某机构推荐。并不意味着它是所有榨汁机中功能最强的，可能它只是个老式榨汁机，现在只是比以前提高了一部分）

5、在高速公路上行驶

（以70英里每小时行驶，早上七点生还机会是晚上4倍，这个说法不足以说明早上更安全，原因只是晚上车子较多，所以车祸才发生多。因为能得出晴天比雾天驾车更危险，晴天车子多。但是常识告诉我们，雾天行车却更危险）

14、1952年小儿麻痹症年

（只是因为人们对症状认识加深，更多病人来医院诊断、治疗，医院记录增多，而死亡人数却没有大的变化）

死亡率、死亡人数比发病人数更匹配，因为可能因为民众认知，诊断人数增多，导致记录的发病人数受影响

6、去年飞机失事造成的死亡多于1910年

（并不能说明坐飞机比过去更危险，只是因为现在乘飞机的人比过去多得多，死亡人数才会增多，我认为比率比数量衡量精确）

7、死于火车事故的人数

（即使4712人死于火车事故，但是仅有132人是乘客，应该结合客户旅程总数，有可能132的死亡仅是一起火车事故造成的。）

因此关心旅行安全，不是查询事故数多少，而是关心100万客人中遇难比率，若比率很低，其实出事可能性很小。

8、职工对工会怨言

（在公司与工会发生摩擦，调查人员对工会意见，在真实询问和记录下，绝大多数都会有怨言，以相同方式，工会也能证明员工反对公司经营方式。这样达不到我们要的的目的。）

9、公司财报有隐藏利润

（每销售1美元只能盈利1.5美分，若是灯泡更换30美分，对应20美元销售额。因此隐藏了利润未报）

10、A&P商店销售净利润

（只有1.1%，比不上银行利率，为什么公司还继续经营，不直接将钱存入银行，因此有利润未报）

11、流感、肺炎数据

（并不是说疾病只出现在南方3个州，而是因为只有他们记录保留着，其他都删除了，因此数据不全，因此无意义，不能说明疾病主要发生在哪个州）

12、疟疾

（并不是说疟疾减少，而是以前人以为感冒、着凉都是疟疾，现在是确诊为疟疾才有记录，所以记录数减少）

13、海军死亡率

（交战期间，海军为9%，同时期纽约市民死亡率16%，难道参军更安全，其实是参军人多为体格强壮的年轻人，而市民包括老人、婴儿、病人，因此两组对象不可比）

15、教师最低收入提高

（前者为1942年乡村地区教师最低收入，后者为1947年纽约市区教师最低收入，可比性低，只能部分归功杜威政府）

# 8 相关关系与因果关系

a钟整点，b钟回敲响，是a引起了b吗？

相关关系无法确定何为因果。因果相反、或有第三、多个因素、相关关系有范围

1、吸烟者成绩低

（没有影响存在显著相关：”抽烟与低分有着很高的相关性，但是不能说明抽烟导致低分，有可能低分才抽烟呢。第三因素导致，性格外向与抽烟的关系，性格外向与成绩的关系）

2、牙膏防蛀牙

（将对自己不利的结果放到一边，只公开结论。）

利用小样本，可以使得任意两个事物显著相关

3、股票与钱关系

（有影响存在显著相关：股票多，钱多；钱多，股票多）

4、牧师收入与朗姆酒价格

（没有影响存在显著相关 ，可能存在第三因素，比如世界物价）

5、六月自杀率最高

（存在多个原因猜测）

6、雨和谷物

（雨下得多，谷物长得高，收成好。但是超过范围，可能破坏甚至毁灭庄稼）

7、教育价值量

（高中毕业生比中途辍学者收入高，不能说明在学校呆的时间越长，收入越高，因为博士生当老师者居多，收入不是最高的）

8、高个子与矮个子

（正相关关系：虽然平均而言，高个子比矮个子男孩更重，但是也能找到矮个子男孩比高个子男孩重的例子）

9、距离与光线

（负相关关系：距离越大，光线密度减少，越看不清书）

相关关系是一种趋势（8、9、10），不是一对一的关系，即使存在清晰的因果关系，个体不能根据相关关系作出决策。

10、大学生与收入

（总体趋势，大学生比不上大学的学生收入高，不能推到某个个体，若是大学生，那么将赚很多钱，该结论未经证实。1、聪明有钱 2、富人有钱 并不一定因为大学）

11、大学与单身关系

（真实的相关关系不能支持未经证实的因果关系，文中认为上大学使得女生单身增多，有可能那些女孩即使不上大学也是单身，事实上想单身才使得她们上大学）

12、牛奶与癌症发病率

（发病率高的州，虽然居民喝牛奶多，但是更重要的是居民寿命长，癌症多发于中年人和老年人。英国妇女癌症率虽然高于日本妇女，不是因为牛奶喝得多，而是因为研究期间的英国妇女平均寿命长于日本妇女）

13、年龄与脚尖角度

（年级较长妇女，两脚角度比较大。不是因为脚尖向外角度大，导致人变老，而是老了才导致脚尖向外角度变大）

14、跳蚤与健康

（跳蚤多更健康？事实上，每个居民身上都有跳蚤，若发烧引起体温升高，跳蚤才会死亡。因此健康人跳蚤死亡数少，并不是跳蚤多才健康）

# 9 如何进行统计操纵

平面图形、平均数不同、精确的数值、样本量过小（只显示比例）、基数

不同、统计量相加、图表

1、地图例子

（在地图上表示国民收入多少被政府征用，用阴影表明政府开支与这些州总收入相互持平，但是因为多选用地广人稀收入少的西部州，反应在图上阴影部分更多，冲击力更大，其实扭曲了关系，隐藏了事实，若用纽约州等东部区域，阴影部分会少很多。）

2、美国国民平均收入

（均值不同，前者均值，后者中位数，若按人口算平均收入，会有可笑的事情发生，四口之家收入会是两口之家2倍，事实不是如此）

3、睡眠时间

（平均每晚7.831会比7.8更精确）

4、兼职家庭帮工周收入

（4.9%周收入18美元，但是4.9%仅仅指2个人，小样本容量上的百分数可能产生误导，最好直接具体数值表示）

5、圣诞节省100%

（基数不同：实际上半折出售，50%折扣，但是用减少量比上折扣后价格，确实减少100%。鲜花便宜100%，原油价格下降从14%至220%都是相同道理）

6、商品利润率

（成本1.75美元销售价格40美元，3800%利润率，以成本为基数，结果2185%，以销售价格为基数，结果95.6%）

7、工资变动

（去年下降20%，今年5%工资提升。去年以原有工资为基数，今年以减少后工资为基数。以1美元为例子，50%减少量，以100%增加量为补偿）

8、打折扣

（50%折扣20%折扣，并不等于70%折扣。其实60%折扣，因为20%折扣以五折后价格为基数）

13、物价上涨

选择去年为基数，今年比去年涨25%

选择今年为基数，去年比今年涨25%

9、计算员工罢工损失

（损失相加：罢工一天几百万损失，1、工人生产的汽车总价值 2、供应商损失 3、零售商销售损失 4、街头停车费等一切可以加的费用）

10、百分数相加

（总成本与所有成本，所有成本上升10% 总上升10%，平均工资真的大于正常工资（加权时间计算才行））

11、百分比与百分点

（3% 6% 增长3%百分点==增长100%）

12、增加率和增长数画在一张图表

统计资料与结论像文章中刻意隐藏或者放大的词汇，需要仔细思考甄别。

# 10 对统计资料提出的五个问题

1、谁说的

有意识的偏差（错误的陈述、含糊之词、刻意挑选不合适的数据而放不合适的数据一边、测量标准的改动（基数的变动）、不正确的测量方法（平均数：均值、中位数、众数））

看清楚谁说的，权威人士掩盖了资料来源，权威人士多个定义。有可能例子，数据康奈尔大学，但是结论是作者的，形成错误印象“由康奈尔大学得出的结论”

2、他是如何知道的

样本有偏（选择不当、刻意挑选的有利样本），样本是否够大且具有代表性

3、遗漏了什么

不对未经过可信度检验的相关认识、平均数类型（均值与中位数存在较大差距）、缺乏比较的数据毫无意义、只给出百分数或描述缺少原始数据、基数可能扭曲事实、更应该关心比率（死亡率），而不是数值（死亡人数 ） 因为现在人数远超过从前了，特别是几十年前，经过几十年后这种描述

4、是否有人偷换了概念

从原始资料到形成结论间是否被偷换了概念（发病案例增多与发病率、民意测验与竞选、读者偏爱与杂志销量、犯罪人数大幅报道与犯罪率、洗澡平均次数、35年前后农场数目（定义）、35岁远大于34、36人数（年龄不确定倾向35）、人口登记目的、囚犯生活费与酒店住宿费、相关关系与因果关系、无领域的第一、直接借款与分期付款6%不同意思、文字游戏（词语引起歧义））

5、这个资料有意义吗

文章可读程度、描述不符合实际（周期性偏头痛、退休年龄、患病数与不匹配的死亡数、出生率与家庭规模、过于精确数据（平均每年汽车纳税51.13美元）、不加控制的外推法，1947年-1952年，家庭拥有的电视数量增加近10000%，若是以该趋势预测下一个五年的电视数量，你将发现电视机数量总数超过10亿台，即每个家庭拥有40台电视。