

如果你觉得统计学很高深，你可以看看这个如何教小孩统计分析的故事。

从考试成绩到贸易逆差

孩子问：爸爸，你是做什么工作的？我明天要在班上介绍自己父母的工作。

我说：帮客户做统计分析。

孩子问：统计分析？

我心里想你这初中孩子，要如何给你解释统计分析呢？

我尝试用一个他可以想象的场景来介绍——假如下面是你班里同学的数学成绩，你会用什么方式把这些成绩表达出来？

孩子有点迷茫。

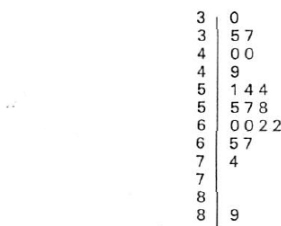
=== 20 名学生在数学考试中的分数 (满分为 100 分，按大小排序) ===

30, 35, 37, 40, 40, 49, 51, 54, 54, 55

57, 58, 60, 60, 62, 62, 65, 67, 74, 89

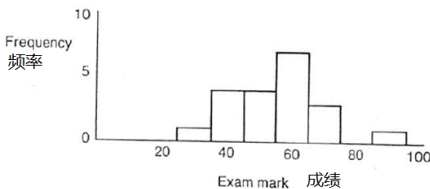
我接着说：任何样本数据，都可以用图加平均值、范围等来表达。

举例：我们可以画茎叶图 (Stem and leaf plot)[详见附件] 表达这些学生成绩：



把 20 组数，从最低排到最高。排在第十位学生和第十一位学生的平均数 56，叫中位数 (Median)。排第五名和第六名成绩的平均值 44.5，就是 Q1 四分之一。排十五名和十六名的平均值 62，就是 Q3 四分之三。这个例子就是中位数。我们也可以再加上四分位数，把排到四分之一、四分之三的两个数也标上。

看了这个茎叶图，再加上 Q1，中位数 (Q2)，Q3，[44.5, 56, 62] 便可以总括表达这二十个学生的成绩分布。也可以用下面的柱状图 / 直方图表达：



除了中位数，也常用平均值 (Mean) 来表示一堆数据的中间趋势，平均值就是把所有数据加起来，然后把总数除以数量，比如上面 20 学生数学成绩的平均值

是。。。。。

孩子充满自信，打断我说：这个平均值简单，我懂。

我说：好啊，你这么懂平均值，我问你一个小问题？

| class="wikitable" | 问题：5 10 350 355 的平均数是多少？

}

孩子拿了计算器算出：180。

我说：你知道那些数是代表什么吗？

孩子说：不知道。

如果这些数是代表角的度数，它的平均值应该是零，而不是 180。这是统计分析很重要的概念 —— 你必须知道那些数的背景，否则你的分析没意义。如果你把那四个数字当成是距离来算均值是 180, 与四个角度数求均值的结果就完全不一样了。

从以上例子，你就知道了解问题背景 (Context) 很重要。

孩子问：这些统计分析有什么用？

答：例如，有了这成绩的分析，你就可以知道自己的成绩在整个班是处于什么位置？也可以帮你比较不同科目的成绩。

比较三种教学方法

比较不同算术教学方法的实验中，45 名学生被随机分成 5 个大小相同的组。两组 (A,B) 采用目前使用的方法 (控制组)，另外三组 (C,D,E) 采用三种新方法。实验结束时，所有学生参加了一次标准测试，结果 (满分 30 分) 见表 B.1。关于教学方法的差异，可以得出什么结论？

B.1

Test results for 45 students									
Group A(control)	17	14	24	20	24	23	16	15	24
Group B(control)	21	23	13	19	13	19	20	21	16
Group C(praised)	28	30	29	24	27	30	28	28	23
Group D(reproved)	19	28	26	26	19	24	24	23	22
Group E(ignored)	21	14	13	19	15	15	10	18	20

C (praised) = 赞赏, D(reproved)= 责骂惩罚, E(ignored)= 不理睬

我说：你可以用刚学过的东西汇总每一组数据，然后看看有没有差异。这样你便可以比较五个班的成绩。

他过了十五分钟就画了出来那个五个班的箱线图：

148 Exploring data

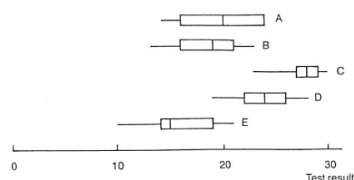


Figure B.1 Boxplots of test results for five groups of students.

我就问他：能看到显著区分吗？

孩子答：看得出是 C 班最好，D 班也比较好，E 班就比较差。

我说：是的，但如果我们只是比较 5 个班成绩的平均值，没有考虑每一班成绩范围，便不能全面比较 -- 例如：上面数学考试成绩的 Q1 Q2 Q3 是（44.5，56，62）如果 C 班的数学成绩是（32，57.5，63）你会认为 C 班比你们好吗？虽然中位数比你们班高，但 Q1 比你们低很多，所以不能单看中位数（或平均值）比较。

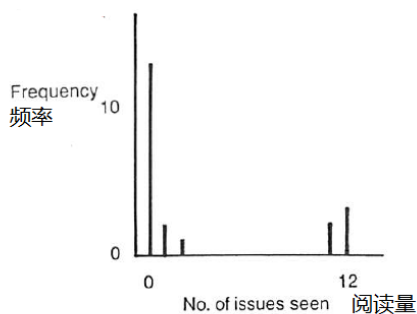
看到他对这些挺有兴趣，我便问他如何表达以下数据：

20 人一年内阅读月刊的数量

0, 1, 11, 0, 0, 0, 2, 12, 0, 0

12, 1, 0, 0, 0, 0, 12, 0, 11, 0

我看他开始使用同样方法，我便说这些统计数据跟前面学生成绩的分布不一样，它一头一尾最高，这表示大部分人要么就是不看杂志，要么就是每个月都看，这种分布就不合适用刚才那些方法去表达，只能说它有两个高峰，或者叫众数 (mode)，再配上一个柱状图。如果我们用中位数或者这些数的平均值，或三分位数来表达的话，反而是误导读者，不能正确的表达那些数据的分布。



当总体 (population) 很大时，就只能抽样，用抽样来估计 population 分布。但是要注意，如果样本不是随机抽样 (random sample)，可能会导致出来的参数偏离、有误差。

我看孩子一头雾水，但我记得他很喜欢研究二战的战斗机，我就用下面这个例子说明，这样抽样出来的结论是没有意义。

样本偏差 Survivor's Bias

错误的抽样会导致错误结论。

二战时会对那些没有被击落的战斗机，研究哪个部分被德军的攻击最多，针对这些部位去加强防护，希望增高飞机的存活率。你同意吗？我们只是抽样了未被击落的飞机，被击落的都未被抽样。一个比较正确的抽样是所有两类都抽样才比较合理。

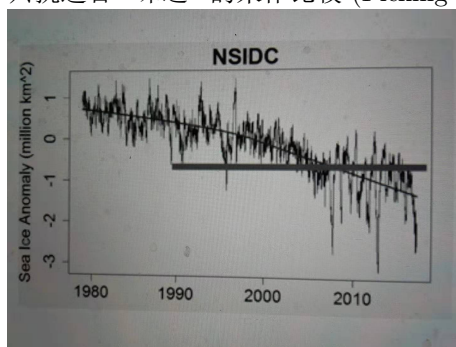
其实被击落的飞机哪个部分被击中更重要，但是我们无法得到那些抽样，只能从没有被击落的飞机去看，这是抽样的错误。所以分析得出的结论没有意义。

A WW II Fighter Plane



我说：你看看下面例子，便会更了解随机抽样的重要性：

只挑选合”味道”的来作比较 (Picking cherries)



* 从上面八零年代到现在全球海洋冰量的统计很明显看到地球在不断地暖软化。但你还可以抽两个时间，例如：2012 年四月份的冰量比 1988 年六月份多，来说明“全球变冷”！

孩子对统计分析好像越来越感兴趣。

我说：现在你开始知道什么叫统计了吗？明天可以跟老师讲故事了吗？

孩子追着问：可不可以再给我一个练习题？我觉得这统计学也没什么困难，我还可以把你给我的那些题目拿给我同桌小李试试，看他懂不懂。(男孩总是要当

英雄，出风头)

我说：好啊。下面的身高统计数据，请你用刚才的方式来汇总一下，看看你学会没有？

20 名妇女参加某种疾病研究，她们的身高 (以米计)

1.52, 1.60, 1.57, 1.52, 1.60, 1.75, 1.73, 1.63, 1.55, 1.63

1.65, 1.55, 1.65, 1.60, 1.68, 2.50, 1.52, 1.65, 1.60, 1.65

他便拿着纸，用刚才学过的方式计算中位数、三分系数等。蛮有自信地画出图，交卷。

我就问他：你认识二点五米高的女性吗？

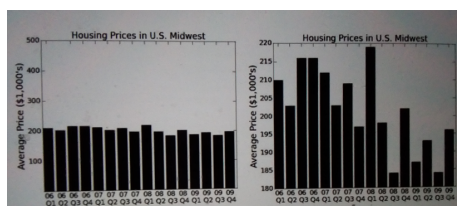
他说：真的好像没有。

所以里面那个 2.5 肯定是不对的。你为什么没有疑问，直接去做？统计分析必须要判断数据是否正确、靠谱。

统计学常常有一个说法叫“垃圾进，垃圾出 Garbage in, garbage out”，如果数据本身不可靠，那么怎么分析都没用。

我接着说：也有很多人利用图表统计数据，误导读者。所以我们看一些统计专家的统计数据分析时也要小心。

美国中西部房价趋势图



上图是 06 年到 09 年每个季度，美国中西部的房价。如果我们看右面那个柱状图，你会认为房价波动很大。但是如果同样这组数据看左面柱状图的话，你就会觉得没有太多变化。所以我们要注意人家用柱状图来表示的时候，要看左面坐标有没有标明数字。不应该仅仅看图。

虽然数据都一样，但如图形不一样，传达的信息便完全不同。

孩子说：你说这个我都懂啊，很简单嘛。

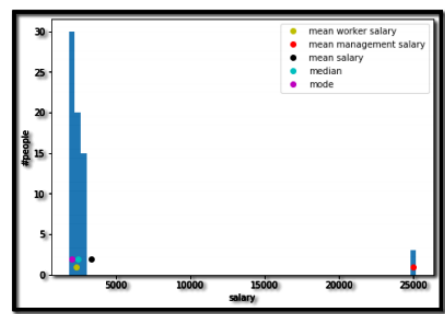
我说：用统计分析骗人的故事可多。你不是说要今年暑假找暑期工赚钱，下面是一个去工厂找工作的故事：

{| class="wikitable" | - | 小李刚中学毕业，希望去工厂找工作。

问：工厂员工薪水是多少？

人事部经理面带笑容地跟他说：我们工厂员工平均月收入大概 3300。

小李进了工厂后，才发现绝大部分的工人月薪是 2000；有多年经验的工人可以拿到 2400。但他们有少数的管理层，他们的平均月薪是 25,000。



所以人事部经理没有“骗”小李，工厂员工平均收入是每月 3309。

}

其实我们天天在媒体/广告都看到利用统计分析来误导我们：

某杂志出版社广告：。。。我们读者年龄的中位数是 34 岁，平均的家庭年收入是 7270 美元。。。

分析例子：连续与分组数据

第二天，孩子乖乖地来找我，他说：爸爸，今天我跟老师说了你的工作是做统计分析，老师听到就发我一些高年级的生物实验数据，希望帮他做些分析，看两类实验有没有差异。你可以帮忙吗？

蚕豆植物 - 双样本 T 检验？

以下数据比较 10 个剪枝和 10 个生根样本中某一化学物质的浓度：

剪枝:53 58 48 18 55 42 50 47 51 45

生根:36 33 40 43 25 38 41 46 34 29

我说: 你可以同样用昨天学过的方法（如，柱状图）比较两组数。
孩子就按这个汇总了两个实验数据的分布。

	Concentration									
	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	
Cut shoots	1					1	3	3	2	
Rooted plants			2	2	2	3	1			

我解释说：你应该也看得懂：上面那个实验有一个很明显的异常点——18。你应该要问老师，这个 18 是否错误还是一个值得相信的数？如果 18 是不对，你就很明显看到上面的平均浓度比下面高。当然我们很多数据，没有这么简单，我们就可以用一个叫“双样本 T”的检验方式来比较两种数据是否有显著差异。但

是如果有刚刚那种异常点，如直接用双样本 T 分析，因为第一组数据被那个 18 拉低了，结果很可能是两组没有显著差异。所以还是直接看数据的分布更实在。

我就给他总结一下：我们在统计学可以使用假设检验（注），其中包括双样本 T 检验比较，方差分析 (ANOVA) 等，就是针对这类统计数据分析。但是在你这个实验数据，从统计图一眼都能看出有显著区分，就不需要再采用这些假设检验方法，因采用那些技巧不会得出额外的信息。

我接着跟他说：有些时候统计数据不是连续数据，像学生成绩、浓度、高度等，而是分组数据。例如下面这些国外人种眼睛颜色与头发颜色的数据，这种数据我们就需要用列联表 (Contingency Table) 来看这两个变量两分组变量之间有没有关系。

头发和眼睛颜色的不同组合

592 人的头发和眼睛颜色的不同组合：

B4表	Hair colour头发颜色			
Eye colour眼睛颜色	Black黑	Brunette深褐	Red红	Blond金
Brown棕	68	119	26	7
Blue蓝	20	84	17	94
Hazel淡绿褐	15	54	14	10
Green绿	5	29	14	16

如果我们利用 B4 表的数，把它那些每一行、每一列的进行求和，我们可以看见每列或者每行的分析结果。例如，棕色 (brown) 眼睛的占大多，但金发 (blond) 例外，金发的大部分是蓝 (blue) 眼睛。其实也看到金头发的那列跟其他的一些分布很不一样。比如我们看每行的话，也可以看到棕色眼睛跟蓝眼睛的数量差不多。但在金头发 (blond) 和黑头发 (black) 那列的分布就不一样了。统计分析的卡方检验 [详见附件 A3] 也可以帮我们分析分组数据间有没有相关。但刚才这类简单 4 X 4 数据表，计算出 B7 列联表便足够，卡方检验没有带来更多有用信息。

实际 (Obs.) 数量表：

B5表	Hair colour头发颜色				
Eye colour眼睛颜色	Black黑	Brunette深褐	Red红	Blond金	Total
Brown棕	68	119	26	7	220
Blue蓝	20	84	17	94	215
Hazel淡绿褐	15	54	14	10	93
Green绿	5	29	14	16	64
Total	108	286	71	127	592

实际 (Obs.) vs 预计 (Exp.) 数量表 Contingency Table:

B7表	Hair colour头发颜色								
	Black黑		Brunette深褐		Red红		Blond金		
Eye colour眼睛颜色	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Total
Brown棕	68	40	119	106	26	26	7	47	220
Blue蓝	20	39	84	104	17	26	94	46	215
Hazel淡绿褐	15	17	54	45	14	11	10	20	93
Green绿	5	12	29	31	14	8	16	14	64
Total	108		286		71		127		592

(注：双样本 T 检验，方差分析 (ANOVA)，卡方检验等都属于假设检验，下一章会详细介绍各种假设检验方法)

总结

统计分析主要是希望利用数据分析、统计方式帮我们解决问题 (Problem Solving)，所以度量本身的重点是如何分析，帮助解决问题。
按目标策划收集那些数据？如何收集？如何分析？

数据分析师也应按以上思路，协助项目经理，更好地利用数据解决问题。

步骤	上面实例
了解背景,提问,明确目标	比较5个班的成绩是否有显著差异
	这些数字代表什么?什么单位? (角度、还是距离)
策划收集数据	怎样取样? 避免取样偏差
确保数据质量	2.5米离女性? 异常数据? (18)
初步数据分析 (描述性)	平均值, 三分数(Q1 Q2 Q3), 柱状图
数据分析	如果能简单看出来有显著差异, 便不需要用假设检验, 或其他统计分析方法
解读(Interpret), 沟通结果	当我们读统计分析报告要小心, 注意是否被骗

附件

如何画茎叶图 (Stem and leaf plot)

以文中 20 名学生在数学考试分数为例：

1. 把分数从小到大排序
2. 从上而下画一直线
3. 第一个数 30，左边写 3，右面写 0（叶）
4. 排第二是 35，因 5 大于 0 - 4, 新一行，左边写 3，右面写 5（叶）
5. 37, 因 7 属于 5 - 9，所以同一行，在 5 右面写 7（叶）
6. 最终便可以简单利用数目字形成一个横放的直方图

References

1. GUTTAG, John V.: "Introduction to computation and programming using Python" MIT Press 2021
2. CHATFIELD, Chris : "Problem Solving: a statistician's guide 2/e" Chapman & Hall 1995
3. BLUMAN: Elementary Statistics 10/e

====<>>====