

统计学中的似是而非

Traps in Statistics

主讲人 傅毅

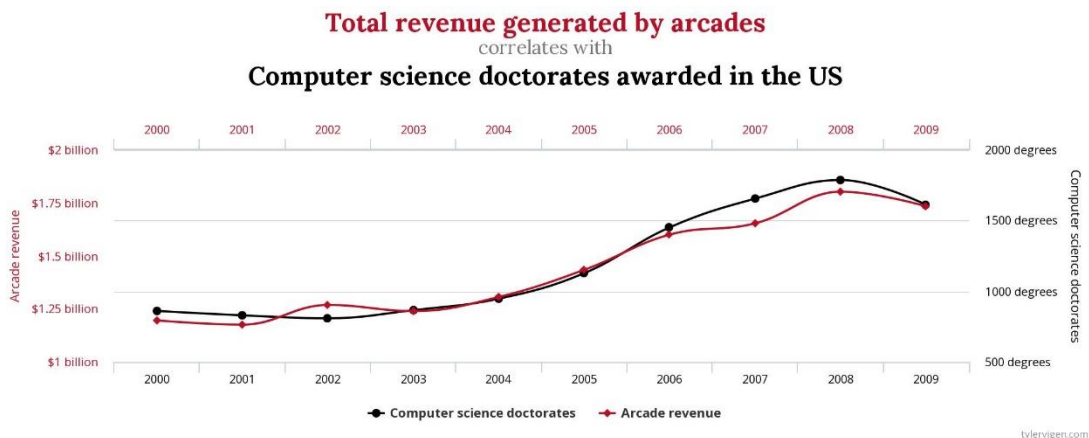
上海师范大学商学院

2022 年 10 月

目录

计算机博士增加了商场的收入? ——相关与因果	1
Computer PhDs increase arcade revenue? - Correlation and causation	3
就业率数据真的算对了吗? ——辛普森悖论.....	5
Are we right about the employment rate? - Simpson's paradox.....	7
数学和物理的成绩是负相关的? ——伯克森悖论	10
Are math and physics grades negatively correlated? - Berkson's paradox	12
数值计算是精确的, 但图表是粗糙的? ——安斯科姆四重奏	14
Calculations are accurate, but graphs are rough? - Anscombe Quartet	18
提议真的得到了大多数人的支持? ——错误概括	22
Is the proposal really supported by the majority? - Error summary.....	24
温度上升的趋势消失了? ——采摘樱桃.....	26
Has the rising temperature trend gone?-Cherry picking.....	28
小医院出生的男性婴儿更多吗? ——小样本风险	30
Are more male babies born in small hospitals?- Small sample risk.....	36
预测准确率为 99.99%的模型真的是个好模型? ——准确性悖论	39
Is a model with 99.99% prediction accuracy really a good model? - Accuracy paradox	41
人员调整提升了销售业绩? ——罗杰斯现象.....	44
Does personnel adjustment improve sales performance? - Will Rogers phenomenon	46
战机的装甲应该被安装在哪里? ——幸存者偏差	48
Where should the fighter's armor be installed?- Survivor bias	50

计算机博士增加了商场的收入？——相关与因果



图片来源: <https://tylervigen.com/spurious-correlations>

这张图展示了美国 2000 年到 2009 年之间的商场收入和计算机博士授予量数据。我们在这两张图中可以清晰的发现，商场收入曲线和计算机博士授予量曲线呈现出了高度相似的趋势。那么究竟是什么因素造成了这样的结果内？

我们为大家准备了下面的选项，你会选择哪一个呢？

- A 商场收入的提高增加了纳税，这使得高校能获得更多的资助用以招收计算机博士。
- B 计算机博士的授予量增加使得从业人员中的数字化水平大幅提高，这促进了商场的盈利能力。

C 计算机博士的授予量增加体现了整体居民受教育水平的提升，受教育水平的提升又进一步提高了的居民收入和支出，这最终提升了商场的收入。

事实上，上面的任何一个选项都缺少严谨的逻辑和有效的证据。那么为什么两根曲线的趋势会如此相似呢？目前我们能够给出的结论是：这只是一种相关，或者说是一种时间上的偶然现象。

相关关系并不等于因果关系！

我们还可以给出个更加直接的例子。假如你在儿时在院子里种下一棵树苗，随着时间的流逝，你和树苗都不断长高。如果我们把你和树苗的高度数据放在一起，你认为两组数据相关吗？我们推断两组数据应该高度相关，但显然两者并非因果关系。

Computer PhDs increase arcade revenue? - Correlation and causation

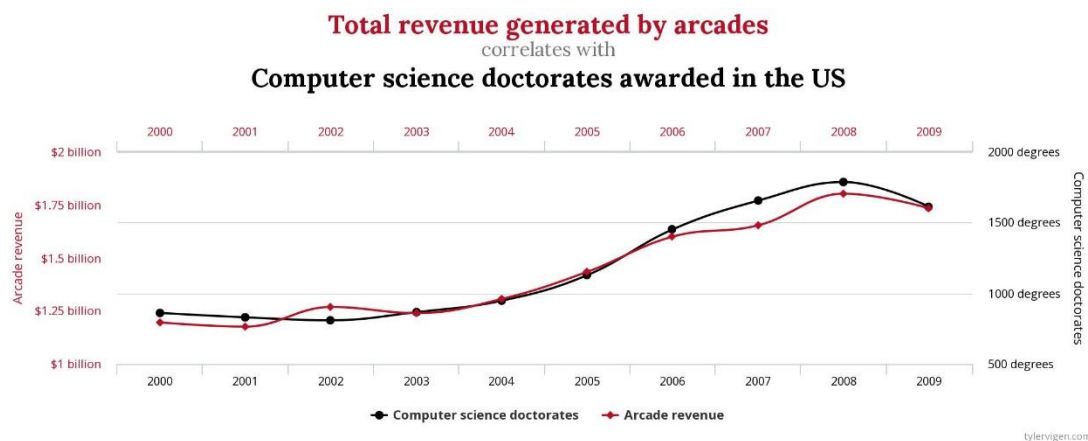


Image Source: <https://tylervigen.com/spurious-correlations>

This graph shows U.S. arcade revenue and computer doctorate grants between 2000 and 2009. We can clearly find in this figure that the arcade revenue curve and the computer doctoral degree curve show a highly similar trend. So what exactly caused this result?

We have prepared the following options for you, which one would you choose?

- A. A Higher income of the arcade increases taxes, which allows colleges and universities to get more funding to train PhDs in computer science.
- B. The increase in the awarding of computer PhDs has led to a substantial increase in the level of digitization among practitioners, which has

boosted the profitability of the arcade.

- C. The increase in the number of computer doctorates awarded reflects the improvement of the overall residents' educational level, which further increases residents' income and expenditure, which ultimately increases the income of arcades.

In fact, any of the above options lack rigorous logic and valid evidence. So why are the trends of the two curves so similar? The conclusion we can give so far is that this is just a correlation, or a temporal contingency.

Correlation does not equal causation!

We can also give a more direct example. If you planted a sapling in your yard as a child, both you and the sapling will grow taller over time. If we put you and sapling height data together, do you think the two sets of data are related? We infer that the two sets of data should be highly correlated, but clearly the two are not causal.

就业率数据真的算对了吗？——辛普森悖论

	系科 1			系科 2		
	已就业	总人数	就业率	已就业	总人数	就业率
大学 A	42	48	87.5%	530	560	94.6%
大学 B	210	236	89.0%	137	142	96.5%

现在看到的是一张被简单化处理过的就业率统计表。表格中假设大学 A 和大学 B 只有两个相同的系科，即系科 1 和系科 2。不难发现，大学 B 中系科 1 的就业率为 89.0%，系科 2 的就业率为 96.5%。这两个就业率都要高于大学 A 中对应系科的就业率。由此我们不难得到一个结论，那就是在该年度大学 B 的整体就业率要高于大学 A 的。

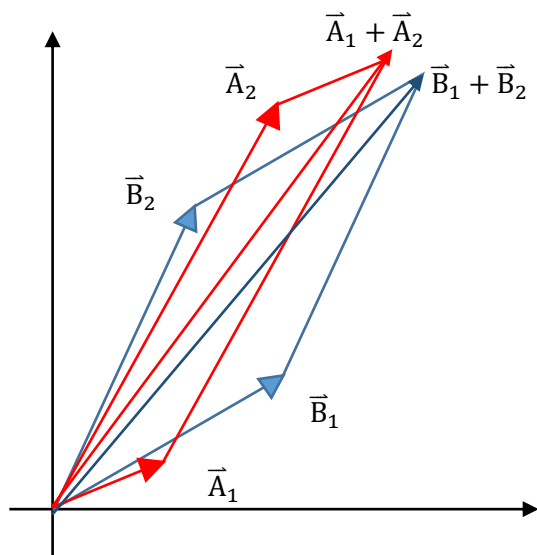
那么这个结论真的是正确的吗？

我们不妨手动计算一下两个大学的就业率吧。大学 A 的就业率等于 $(42+530)/(48+560)=94.1\%$ 。同理，大学 B 的就业率等于 91.8%，具体可以用下面这张表来进行展示。这就很奇怪了，为什么这个结果与我们的直观感受是相反的呢？

	系科 1			系科 2			大学就 业率
	已就业	总人数	就业率	已就业	总人数	就业率	

大学 A	42	48	87.5%	530	560	94.6%	94.1%
大学 B	210	236	89.0%	137	142	96.5%	91.8%

这个例子所阐述的问题通常被成为辛普森悖论。辛普森悖论是统计中的一种现象。当若干子数据被组合起来，原先子数据中的趋势有时会出现消失，甚至出现反转。这个现象在社会科学和医学科学统计中经常遇到。



事实上，用向量的方式来描述上方的例子会更加直观。 $\vec{A}_1 = (42, 48)$, $\vec{A}_2 = (530, 560)$ ，那么系科 1 的就业率就是 $\vec{A}_1$ 的斜率，系科 2 的就业率就是 $\vec{A}_2$ 的斜率，大学 A 的就业率就是 $\vec{A}_1 + \vec{A}_2$ 对应的斜率。同理可以得到 $\vec{B}_1 = (210, 236)$, $\vec{B}_2 = (137, 142)$ ，那么大学 B 的就业率就是 $\vec{B}_1 + \vec{B}_2$ 对应的斜率。我们直观就可以看出，大学 A 的就业率大于大学 B 的就业率。

Are we right about the employment rate? - Simpson's paradox

	Department 1			Department 2		
	employed	total	employment rate	employed	total	employment rate
University A	42	48	87.5%	530	560	94.6%
University B	210	236	89.0%	137	142	96.5%

What we now see is a simplified employment rate statistics table. The table assumes that University A and University B have only two departments, namely Department 1 and Department 2. It is not difficult to find that the employment rate of Department 1 in University B is 89.0%, and the employment rate of Department 2 is 96.5%. Both of these employment rates are higher than the employment rates of the corresponding departments in University A. From this, it is not difficult to draw a conclusion that the overall employment rate of University B is higher than that of University A in that year.

So is this conclusion really correct?

Let's manually calculate the employment rate of the two universities.

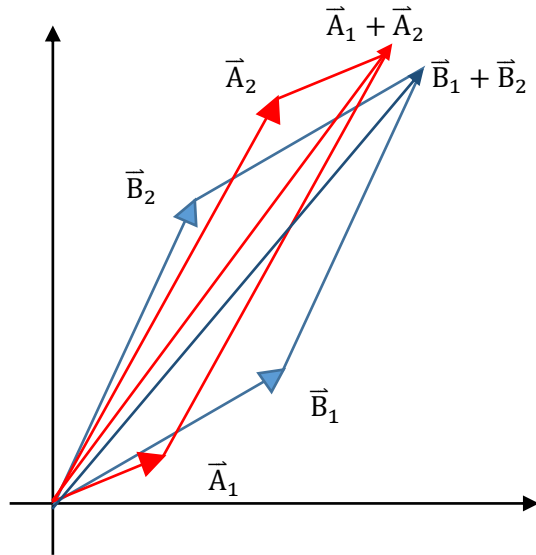
The employment rate of University A is equal to $(42+530)/(48+560)=94.1\%$. Similarly, the employment rate of University B is equal to 91.8%, which can be shown in the following table. This is very strange, why this result is contrary to our intuitive feeling?

	Department 1			Department 2			Total rate
	employed	total	employment rate	employed	total	employment rate	
University A	42	48	87.5%	530	560	94.6%	94.1%
University B	210	236	89.0%	137	142	96.5%	91.8%

The problem illustrated by this example is often referred to as Simpson's paradox. Simpson's paradox is a phenomenon in statistics. When several sub-data are combined, the trend in the original sub-data sometimes disappears or even reverses. This phenomenon is often encountered in social science and medical science statistics.

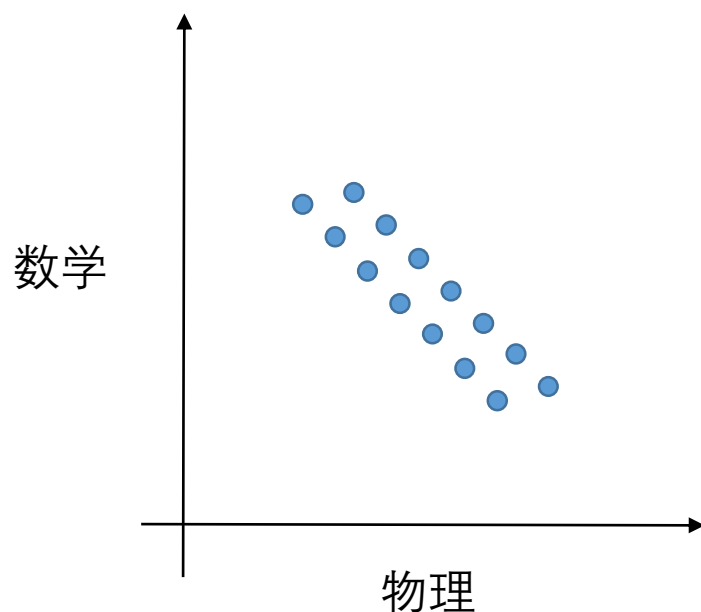
In fact, it would be more intuitive to describe the above example in terms of vectors. $\vec{A}_1 = (42,48)$, $\vec{A}_2 = (530,560)$, then the employment rate of department 1 is the slope of $\vec{A}_1$, The employment rate of department 2 is the slope of $\vec{A}_2$, and the employment rate of University A is the slope corresponding to $\vec{A}_1 + \vec{A}_2$. Similarly, $\vec{B}_1 = (210,236)$, $\vec{B}_2 = (137,142)$, then the employment rate of University B is the slope of

$\vec{B}_1 + \vec{B}_2$. We can intuitively see that the employment rate of University A is greater than that of University B.

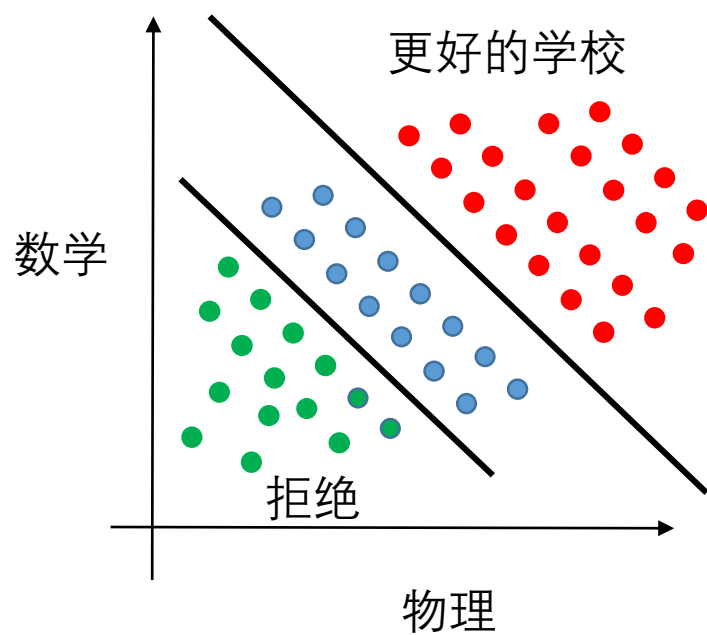


数学和物理的成绩是负相关的？——伯克森悖论

A 学校招生部门在招生的时候主要考虑学生的数学成绩和物理成绩。招生部门在完成招生后，统计了学生的数学成绩和物理成绩。他们惊讶的发现，学生的数学成绩和物理成绩居然是负相关的。具体如下图所示。



这个结果显然不符合常理。我们通常认为数学成绩越高的学生，他的物理成绩也应该越高，反之亦然。那么这个现象应该如何解释呢？事实上，这是一个伯克森悖论，也称为伯克森偏差、对撞机偏差或伯克森谬误，是条件统计的结果，通常被发现是违反直觉的。其原因是由于设置了条件，部分总体缺少相应的样本。

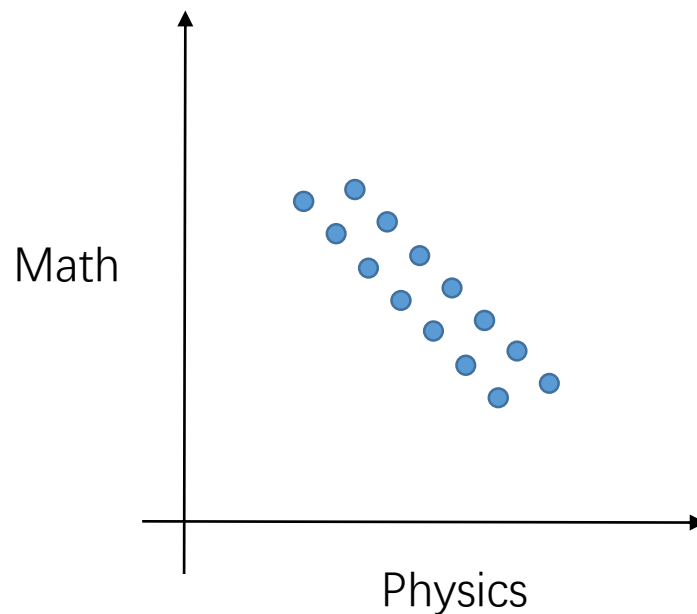


事实上, 数学和物理成绩都很高的学生会选择更好的学校(红色)。反之, 数学和物理成绩都较低的学生则会被这所学校拒绝。这两类的学生都不会被观察到。因此这所学校才会得到这个与我们直觉相反的结论。

Are math and physics grades negatively correlated?

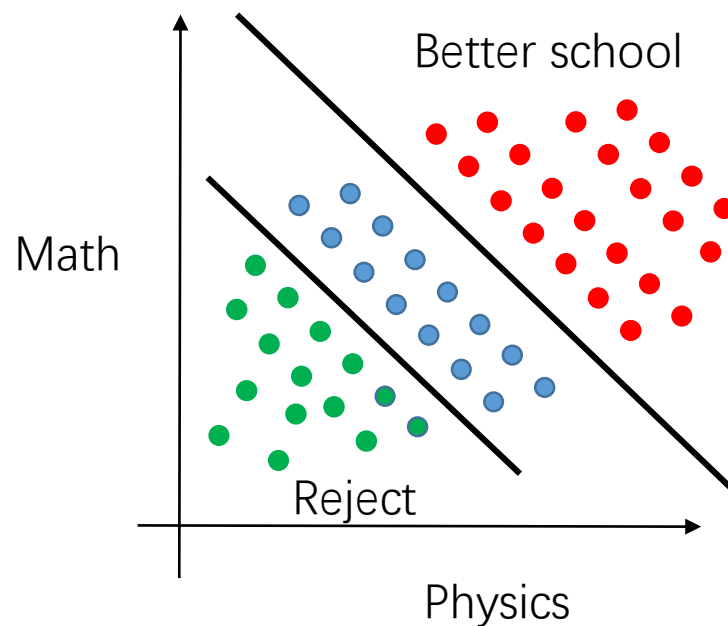
- Berkson's paradox

A school's admissions department mainly considers students' math scores and physics scores when recruiting students. After the admissions department completes the admissions, it counts the students' math scores and physics scores. They were surprised to find that students' math scores and physics scores were negatively correlated. The details are shown in the figure below.



This result is obviously inconsistent with common sense. We generally think that students with higher math scores should also have higher physics scores, and vice versa. So how should this phenomenon be explained?

In fact, this is a Berkson paradox, also known as Berkson bias, collider bias, or Berkson fallacy, a consequence of conditional statistics that are often found to be counterintuitive. The reason is that due to the set conditions, some populations lack corresponding samples.



Students with high scores in math and physics choose better schools (red). Conversely, students with lower math and physics scores were rejected by the school (green). Neither category of students will be observed. So the school came to this counterintuitive conclusion.

数值计算是精确的，但图表是粗糙的？——安斯科姆四重奏

有人说“计算是精确的，但图表是粗糙的”，你认同这个观点吗？
下面我们来看一个例子。

现在我们有 4 组数据，每组数据都有 x 和 y 两个变量，每个变量有 11 个观察值。现在通过计算可以得到每组数据的特征都满足如下相同的结果。例如每组的 x 均值都等于 9，每组 x 和 y 的相关系数都等于 0.816，等等。那么我们是不是就能够认为，这四组数据都是相同或者相似的呢？

项目	数值	备注
x 的平均值	9	精确的
x 的样本方差	11	精确的
y 的平均值	7.50	到小数点后 2 位
y 的样本方差	4.125	± 0.003
x 和 y 的相关系数	0.816	到小数点后 3 位

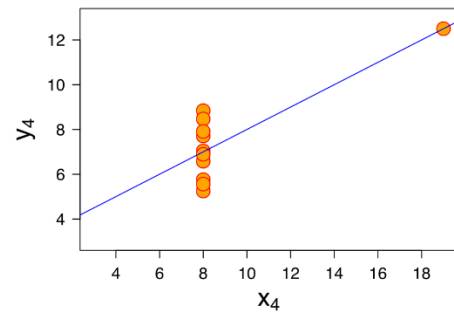
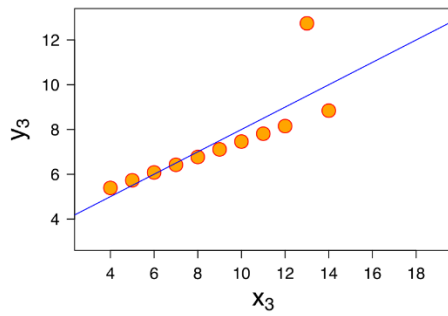
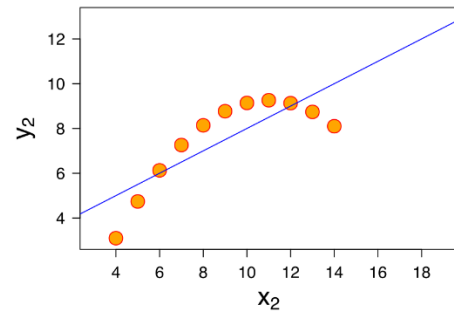
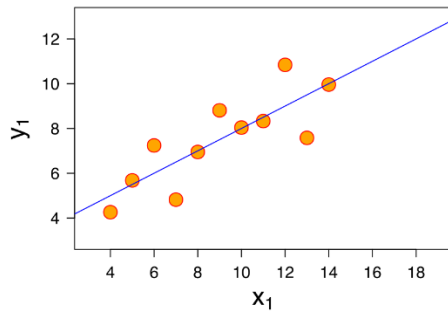
线性回归线	$y = 3.00 + 0.500 x$	分别保留 2 位和 3 位小数
线性回归的判定系数	0.67	到小数点后 2 位

我们想，也许有的读者依据表格中给出的计算结果，已经认为这四组数据是相同或者相似的了，但事实真的是这样的吗？

这四个数据集被称为安斯科姆四重奏。安斯科姆四重奏的四个数据集具有几乎相同的简单描述性统计结果，但分布却非常不同，并且在绘制图表时显得非常不同。

1		2		3		4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47

14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



这个例子说明了在分析数据时数据可视化的重要性，以及异常值对于统计结果的影响。

Calculations are accurate, but graphs are rough? -

Anscombe Quartet

Some people say "Calculations are accurate, but graphs are rough", do you agree with this point of view? Let's look at an example.

Now we have 4 sets of data, each set has two variables x and y , and each variable has 11 observations. Now through the calculation, it can be obtained that the characteristics of each set of data satisfy the same result as follows. For example, the mean x of each group is equal to 9, the correlation coefficient of x and y in each group is equal to 0.816, and so on. So can we think that these four sets of data are the same or similar?

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y	4.125	± 0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively

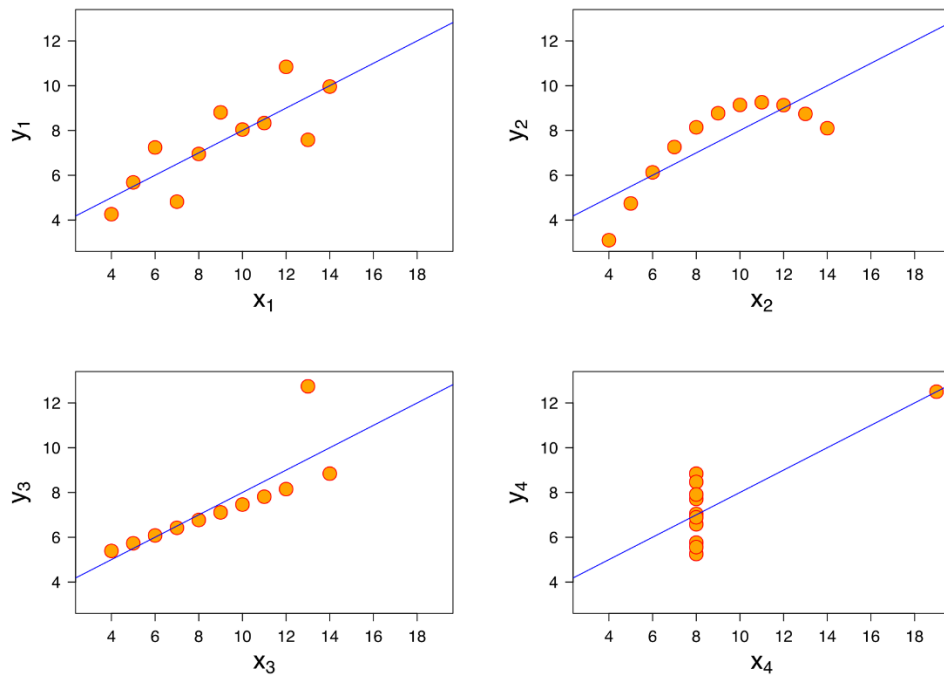
Coefficient of determination of the linear regression R^2	0.67	to 2 decimal places
---	------	---------------------

We guess that some readers may have thought that these four sets of data are the same or similar based on the calculation results given in the table, but is this true?

These four datasets are called the Anscombe Quartet. The four datasets of the Anscombe quartet have nearly identical simple descriptive statistics, but very different distributions and appear very different when plotted.

1		2		3		4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84

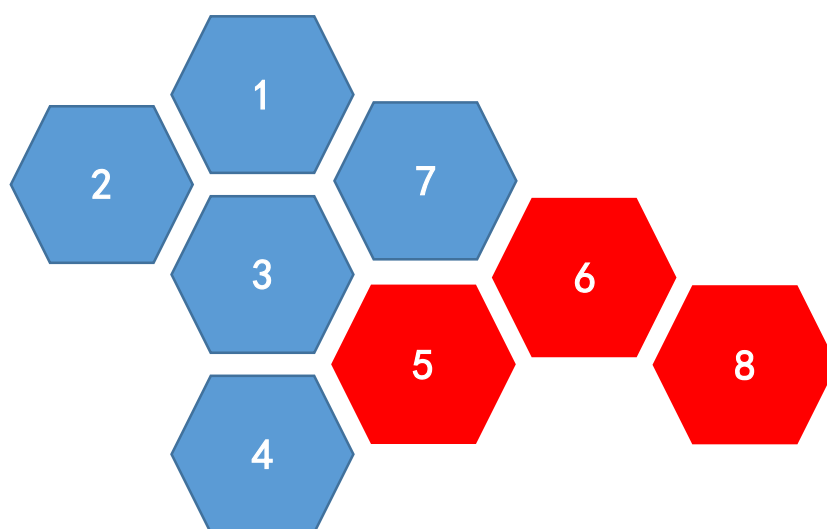
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



This example illustrates the importance of data visualization when analyzing data, and the impact of outliers on statistical results.

提议真的得到了大多数人的支持？——错误概括

某企业员工提议在年底的时候组织一场大型企业活动。现在公司将该提议告知了所有的子公司（8 个），并征求意见。如果员工支持率超过 50%，那么企业将举办这一活动。通过调查，8 个子公司中，有 5 个子公司的员工支持率超过 50%（1,2,3,4,7，蓝色），另有 3 个子公司的员工支持率没有超过 50%，（5,6,8 红色）。显然，现在的总体支持率情况是 $5/8=62.5\%$ ，提议得到了大多数人的支持。真的是这样的吗？



我们一起来看看具体的数据吧。

子公司	支持人数	总人数	支持率
1	25	40	62.5%
2	39	66	59.1%
3	22	25	88.0%
4	51	84	60.7%

5	29	154	18.8%
6	13	85	15.3%
7	24	43	55.8%
8	62	259	23.9%
总数	265	756	35.1%

通过计算我们可以得到整个公司的支持率只有 $265/756=35.1\%$ 。也就是说，提议并没有获得大多数人的支持。

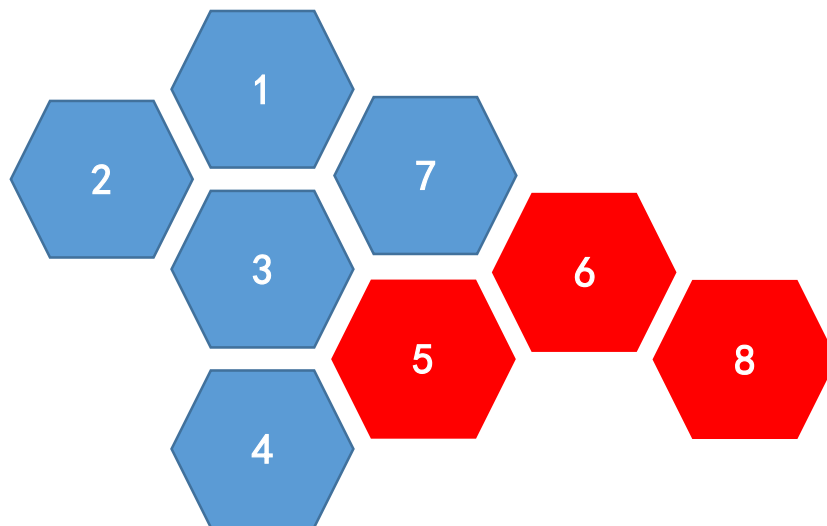
那么为什么我们会得到两个相反的结果呢？事实上，公司在第一次计算支持率的时候，错误地进行了概括。我们需要计算的是全公司员工支持率，而之前实际计算的是支持率超过 50% 的子公司数量占比。这是两个完全不同的概念，不能被相互替换。

因此我们在进行统计分析的时候，首先要明确分析的问题是什么。

Is the proposal really supported by the majority?- Error summary

A company employee proposes to organize a large corporate event at the end of the year. The company now informs all subsidiaries (8) of the proposal and collects wishes. If the employee approval rate exceeds 50%, then the business will host the event. Through the survey, among the 8 subsidiaries, the employee support rate of 5 subsidiaries exceeds 50% (1, 2, 3, 4, 7, blue), and the employee support rate of 3 subsidiaries does not exceed 50%, (5, 6, 8 red). Obviously, the overall approval rating is now $5/8 = 62.5\%$, and the proposal is supported by the majority.

Is it really like this?



Let's take a look at the specific data.

Sub-firm	support	total	rate
1	25	40	62.5%

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/296120053204011042>