统计数据的迷惑性

摘要：伴随着日常生活中“统计数据”出现的日益频繁，数据因其给人的严谨科学的第一印象而备受推崇。这种推崇被一些人为了实现某一目的利用，通过提供一些所谓的统计数据来欺骗人们。本文的想法来自于 Peter Donnelly 的 TED 演讲《揭示统计数据是如何迷惑陪审团》，旨在通过举出一些例子并作出相应分析，来解释这些统计数据的迷惑性，端正人们对于“统计数据”的认识。

关键词：统计数据　迷惑性　平均值　百分比

一、总论

统计数据是一种以数字形式表现出来的证据。这样的证据可以给人留下深刻的印象，因为数字使证据看起来非常科学、精确，让人感觉似乎这就代表着“事实”。然而，统计数据可以并且经常欺骗大家！表面上它们很有说服力，事实上却不一定。作为一个批判性思考者，你必须力求查明误用统计数据的推理。

这其中包括很多种方法，产生具有迷惑性或欺骗性的方法，下面将对于这些方法举出例子并做出分析，并提供出合理可行的应对这类统计数据的思路。

二、统计数据的产生

1、统计数据的来源

通常呈现在我们面前的统计数据都是经过统计学处理的，但是最原始的数据是一切分析的基础，然而原始数据的得到是和统计的方法有关的，我们来看一下下面的这个例子。

1995 年，一名报刊专栏作家对一些女性读者进行了访问，询问她们：“你情愿被丈夫紧紧抱住并温柔体贴地对待，而忘掉‘行动’吗？”这名作家报告说，接受访问的女性中有 72％的人对这个问题回答了“是”。所以她得出这样的结论：“这次调查表明，相当多的妇女对性生活不感兴趣。”

你发现这名作者是如何在提供一个事实时得出另一个结论的冯？你是否认为，如果这名专栏作家的问题是：“你喜欢过性生活吗？”所得的结果将会与这次调查的结果不同。

在面对这种情况的时候，我们只需要了解统计的方法就能对于统计数据的可信程度做一个界定了。

2、统计数据的处理

在保证数据来源科学可靠时，我们并不能够保证统计数据的正确性，前文已经提到，统计学数据的得到是需要对原始数据进行统计学分析处理的。这里就存在另一个问题，统计学分析这个过程是否科学严谨。

在做统计学分析时，我们需要有一个非常扎实的概率论和统计学基础，最好这个工作由统计学家来做，因为他们通过研究，有一个非常完善的统计学知识体系，对于不同的情况，能做出相应的分析，使用相应的统计学模型进行处理，但我们却并不能保证做这个工作的一定是统计学家，如果是其他人（没有形成完善的统计学知识体系）来做，常常被自己的主观想法所迷惑，而做出错误的分析，下面举几个例子。

1、连续抛掷硬币，直到出现某一种特定的序列（比如正反正）。考虑两种序列，正反正和正反反，下列那种说法是正确的（）

A.从开始抛掷到出现正反正时的抛掷次数大于从开始抛掷到出现正反反时的抛掷次数

B.从开始抛掷到出现正反正时的抛掷次数等于从开始抛掷到出现正反反时的抛掷次数

C.从开始抛掷到出现正反正时的抛掷次数小于从开始抛掷到出现正反反时的抛掷次数

对于这两种序列，正反正和正反反，由于在连续抛掷三次硬币，出现正反正和正反反的概率是相同的，都是 1/8，于是我们最直观的想法是，从开始抛掷硬币到出现正反正时的抛掷次数等于从开始抛掷到出现正反反的抛掷次数。

但是，换另一种思路来看如果我们想要得到正反反序列，当出现到正反时，若再出现反，显然就得到了，若出现正，则可认为完成了下一轮的 1/3。而如果我们想要的到正反正序列，当出现正反时，若再出现正，显然就得到了，若出现反，则前面出现的正反都白费了，而且这次出现的反也是对于想要的到的正反正没有贡献的，我们需等到下一次出现正面朝上，才能开始期待正反正的出现。由以上，我们可以知道，出现正反反序列比出现正反正序列要“容易”很多，于是，“从开始抛掷到出现正反正时的抛掷次数大于从开始抛掷到出现正反反时的抛掷次数”。

显然是后一种思路更严谨，更有说服力，却并不是我们最直接的想法。

2、假设我们有一种疾病的测试方法，测试结果正确的概率是 99%，现任意抽取一个人，对其进行测试，测试的结果呈阳性，问这个人患病的概率是多少？

A.等于 99%　　B.大于 99%　　C.小于 99%　　D.不能确定

乍一看之下，对于这个问题，我们会想，该人被测试呈阳性，而测试正确的概率为 99%，所以这个人患病为 99%。

但是这个问题我们应该使用概率论中的贝叶斯公式。

A 表示“测试呈阳性”,B 表示“患有该疾病”，则 P（A|B）= 99%， P（~A|~B）= 99%，设随机抽取一人，该人患该疾病的概率为 x=P(B)；则由贝叶斯公式，

$$P(B|A)=\frac{P(B)P(A|B)}{P(B)P(A|B)+P(\sim B)P(A|\sim B)}=\frac{99x}{99x+(1-x)}$$

即为该人患病的概率，显然这个概率是和这种疾病出现的概率有关的，因为题中

说的是测试结果正确的概率是 99%，这包括两个方面，一个是测试呈阳性该人患病和测试呈阴性该人不患病，而最直观的想法中只考虑了测试呈阳性该人患病这一种情况。

在对于这种情况，我们只有学习概率论和统计学知识，形成一个自己的知识体系，在面对这种问题时才能有一个正确的思路。

二、统计数据的呈现

在保证统计数据来源和处理时，统计数据的呈现也是具有迷惑性的，而且个人认为它的迷惑性常常远大于前两者带来的。

1 平均数

“平均数”对于我们一般人来看，最直观的认识是一种能代表一个群体平均水平的统计数据，日常生活中，我们也对于这一种数据表现出极大的兴趣，但是在统计学数据中，它常常是最具有迷惑性的。

阅读以下这些句子，看看有什么问题：

（1）当前美国人的收入比以往任何时候都高；美国工人的平均收入是 3.5 万美元。

（2）目前，工厂造成的空气污染的平均值低于危险水平。

这两个例子都使用了“平均”这个词。但是，在统计学中定义一个平均数有三种不同的方法，而且在大多数情况下不同的定义会得到不同的平均数值。是哪三种方法呢？

第一种方法是将所有的数据相加，再用所得的和除以数据的个数，得到的结果就叫做算术平均数。

第二种方法是按从大到小的顺序列出所有数据，找出位于中间的那个数。这个数叫做中数。一组数值中有一半数据大于中数，一半数据小于中数。

第三种方法是列出所有数据，然后将不同的数值排列归类。在一组数据中出现次数最多的那个数值叫做众数。

作者讨论的是算术平均数、中数还是众数，会产生很大的差异。再来分析一下美国人的收入分布状况。有的人收人极高，如年薪 200 万美元。这样高的收入将会大大地提高算术平均数。然而，这些个别的高收入对于中数或众数的影响都很小。因此，如果某人希望使平均收入看起来高一些，算术平均数可能是最能达到目的的平均数。现在你明白，当人们谈论收入时，明确他们采用的是何种平均数有多么重要了吧。

让我们来仔细看看第二个例子。如果作者所给出的是众数或者是中数，都可能使我们得出错误的判断，认为空气污染的程度还没有超过安全范围。例如，即使产生严重污染的工厂只是少数，但这些工厂排放的污染物的总和远远超过危险水平－就算把这些污染物分散到整个大气层里也是相当危险的。在这种情形下，用众数或中数来表示污染值都会非常低，但是算术平均数却会非常高。当你看见表示“平均”的数值时，都应该想想：“采用算术平均数、中数或众数是否有差别？”为了回答这个问题，请你思考使用不同平均数的含义会如何改变已知信息

的意义。

通常，不只是决定采用哪一种平均数才重要，决定最小值和最大值之间的间距（即数据的范围）、每个数据出现的频率（即数据的分布）也同样重要。例如，假设你需要一些信息来帮助你决定吃或不吃从邻近海洋里捕捉到的鱼。如果只告诉你那些鱼的平均汞含量，你会满意吗？显然，这些信息是不够的。

我们还想知道汞含量值的范围，也就是说，汞含量可能达到的最高值和最低值以及不同含量值出现的频率。因为有可能所算出的平均数是在“安全”标准内，但是如果有 10％的鱼汞含量高于“安全”标准，我想你宁愿不选择这些鱼作为晚餐。让我们再来分析另一个事例。在这个事例中，掌握数据的范围和分布是至关重要的。

美国不是一个过度拥挤的国家。就全国范围而言，每平方英里’还不到 60 人，低于大多数国家的人口密度。

首先，我们怀疑算术平均数不能代表人口密度。虽然这里用算术平均数取得的人口密度可能非常低，但是，众所周知，美国的一些地区，如东北部人口密度非常高。因此，虽然美国的平均人口密度并不高，但事实上美国的一些地区是过度拥挤的。可见，当你看到平均数时，问问自己：“我是否需要了解数据的范围和分布情况？"

2 遗漏的信息

有时候，不提供信息也是一种信息，但是这种信息常常被人们忽略。最常见的是百分比。百分比作为一个比值，必然是从在两个数中来的，对于这三个数据，若全部给出，当然就不会有问题了，但如果只给出其中的一个，其实根本说明不了任何问题。让我们通过下面这两个例子来说明问这个问题的作用。

1. 一股犯罪浪潮袭击了我市。去年杀人犯的比例增加了 67％。

2．与其他近距离接触的运动相比，拳击运动的危险性更小。纽约一项历时 30 年的关于运动引起死亡的调查显示，在这期间，棒球运动中死亡了 43 人，在死亡率方面超过了足球（22 人）和拳击（21 人）。

一开始，67％这个数字会给你留下相当深刻的印象。但是请注意，这里有信息被忽略了，即计算出这个百分比的基础－绝对数值。同样是增加 67%，从 300 个增加到 500 个与从 3 个增加到 5 个，哪个更令我们警觉呢？在第二个例子中，我们知道绝对数值，但不知道百分比。难道我们不需要了解这些绝对数值转化成百分比对运动员意味着什么吗？毕竟，从事棒球运动的人要比从事拳击运动的人多得多。

当你遇到令人印象深刻或震撼人心的数字或百分比时，千万要小心。你可能需要获得一些其他信息来判断这些数字何以能令人印象深刻。

四、总结

除以上举出的例子，还有很多的方法产生具有迷惑性的统计数据。随着社会的发展，人的平均素质在提高，“数据”的说服力被摆到了一个至高无上的位置，于是就有人利用大多数人的这个认识，利用“数据”来骗人。对此，我们需要对

于“统计数据”有一个客观的认识，最好能够掌握一定的统计学知识技能，便能看破迷雾，看到问题的本质。