数据使用中的问题：以抽样数据为例

中国人民大学　金勇进

邮箱：jinyongj@public.bta.net.cn

为什么提出这样一个题目

人们对数据的重视

抽样调查得到广泛使用

使用中存在一些“误用”和错误解读

不仅关心如何获得数据，也要注意怎样使用数据。

表现一　这样的数据是否能做统计推断

用于推断的数据应该来自于概率抽样

概率抽样特征：

随机化原则

总体单元都有机会入样

单元入样概率已知

现时中非概率抽样得到广泛应用

非概率抽样：概率抽样以外的其他抽样方法

最常见的是方便选样

有熟人，善良受访者，利用学生假期回乡做调查

非概率选样有很多优势

方便，成果率高，成本低

可以发现问题，描述情况，或做探索性研究

用非概率选样进行统计推断缺乏理论支持

但这种情况大量出现，甚至频繁出现在学术期刊的论文中，对读者起到误导作用。

表现二 进行推估没有权数

用样本数据推估总体参数，核心数据是两个，测量值和样本单元权数

权数的作用：将样本测量结果放大，还原到总体规模，也可以把权数视为“扩张系数”

权数有不同类型，如设计权数，调整权数，以及特定目标的权数修正。

在数据预处理和数据分析中，数据预处理（包括计算权数，对缺失数据的处理）占整个工作量的大部分，甚至是绝大部分比重。

而数据预处理环节，在我们许多调查数据的工作中见不到，或者不规范。

缺乏权数是突出表现

对缺失数据的处理做的就更差。

表现三. 软件陷阱

盲目使用统计软件的陷阱，抽样方法与分析方法不对应。

例：欲分析家庭有线电视与购买个人电脑是否相互独立，随机抽取了500个家庭，调查结果如下表所示：

		个人电脑		
		有	无	合计
有线电视	有	119	188	307
	无	88	105	193
	合计	207	293	500

运用传统的卡方检验，设原假设

H_0：拥有个人电脑与拥有有线电视相互独立

经计算 $P\left(\chi^2 \geq 2.28\right) = 0.1310$ 。因此没有充分理由拒绝，可以认为家庭拥有个人电脑与拥有有线电视相互独立。

但如果是采用整群抽样，向500个家庭的夫妇同时进行调查，得到的结果如下：

		个人电脑		
		有	无	合计
有线电视	有	238	376	614
	无	176	210	386
	合计	414	586	1000

这时可得检验统计量 $P\left(\chi^2 \geq 4.56\right) = 0.0327$

因为p值较小，所以拒绝原假设。说明个人电脑与有线电视有联系。

对同一群体进行调查，却得到不同的结论，为什么？

因为没有考虑群内单元之间的相关性。

盲目使用软件计算容易落入“软件陷阱”。

表现四　　平均数中的“悖论”

调查数据中经常使用平均数描述现象的一般水平。

有下面房价变动的例子。

报告期和基期平均房价比较

	总销售价格（亿元）	总销售面积（万平方米）	平均房价（万元/平方米）
基期	220	100	2.20
报告期	215	100	2.15

数据表明，与基期相比，报告期平均房价下降。

将表中数据分解，形成下表：

	总销售价格（亿元）	总销售面积（万平方米）	平均房价（万元/平方米）
基期	220	100	2.2
城区	180	60	3
郊区	40	40	1
报告期	215	100	2.15
城区	160	50	3.2
郊区	55	50	1.1

无论是城区还是郊区，报告期与基期相比，房价都在上升，总平均却是下降，主要是结构在“作怪”。

简单使用平均数，有时就会被误导。

上面所举的各种例子，错误并不复杂；
但却频频出现在我们的数据使用中，
这提示我们，在大数据时代，我们需要更高质量的原始数据，也不要忽视对数据的正确使用。

谢谢大家！