chenqin，数据帝
阅读原文

当自然产生的数据分布在某一个位置出现一个「堆积」时，猫腻可能就隐藏其中。有些可能不是有意造假，比如人口普查数据年龄堆积在尾数为0或者5的数字中，是因为被调查人记不清自己的年龄，报告了一个模糊的数字。（https://www.zhihu.com/question/24929287/answer/29574198）有些可能也不算造假，而是数据操纵，例如上市公司的盈利总是堆积在0的右侧，那可能算是一种盈余管理，将不同年份的盈余挪腾后避免连续亏损而退市。但有些数据中出现堆积，可能就有造假的嫌疑了，最好的例子就是身高。身高是一个难以突变也无法操纵的数字，当我们对一群人的身高进行比较精密的测量时，他应该比较接近正态分布，以下数据来源于CHNS——但如果我们把人们自己报告的身高拿来做一个概率分布，他的分布就没有那么完美了，下图列出了CFPS2018数据中25-35岁男性的身高——当然，这个身高数据是自己汇报的。可以看到，这里的身高在每一个整五或者整十关口都出现了明显堆积，尤其是170处，高达19.25%的男性声称自己有170cm。以上两组数据来源于同一个年龄段的人口，出现这样大的差异显然是不可能的，一定有许多身高并非170的人将自己的身高谎报为170。那么，到底是身高比较矮的那些倾向于高报，还是身高比较高的人倾向于低报呢？我们可以将两组数据画出累计分布图。上图画出了两组数据的累计分布。可以看到，在170以上，自报身高和测量身高是基本重合的，说明身高超过170的人口没有虚报或者低报自己的身高。但在170以下差距就呈现出来了。上图画出了三个箭头，表示仪器测量身高为167cm的25-35岁男性在人群中的分布和自报身高为170cm的同年龄段男性的累积分布概率是一样的，类似的现象还在165→168以及162→165的位置出现。如果我们假设测量身高到自报身高是一个保序的映射——前者到后者不改变其排序——那么可以得出一个结论，那就是身高169、168和一部分身高为167的男性，在面对调查员时会将自己的身高报告为170。以及167、166和一部分165的男性会报告168的身高；164、163和一部分162的男性会报告165的身高。换言之，面对调查员，170cm以上的男性不太会虚报自己的身高，但是170以下的男性会虚报，且虚报不会超过3cm。上面的结论是被调查人面对调查员的反映，面对调查员的虚报可能还可以避免，但到了相亲市场上，这个虚报可能就不仅仅是可以理解，而且是完全必要的了。为了比较相亲市场上的表现，我们加入世纪佳缘的用户资料数据（ @杨阳 对网站做了数据抓取）——世纪佳缘需要每一个用户填写自己的身高。那么当我们把世纪佳缘的25-35岁男性身高累积分布放进上图的时候，奇迹出现了——世纪佳缘的身高分布和测量身高以及自报身高都出现了非常显著的差异，且在170处的「堆积」更加明显了。当然，要从上图进行推理，还需要解决几个问题。首先，虽然都是25-35岁的男性，世纪佳缘的人口分布和有群体代表性的抽样调查肯定是不一样的，比如身高很高的男性可能用不着去世纪佳缘挂牌相亲。但是这个猜测与数据并不吻合，因

为身高更高的人更不会选择挂牌相亲，会导致蓝线向上穿过灰色线，但实际并没有，蓝色线始终在灰色线的下方。另一种可能是身高比较低的男性就连在相亲网站挂牌的概率都会更低，这与数据是吻合的，但却仍然不能解释在170处的「堆积」现象——我们可以理解身高169的男性上相亲网站的比例低于身高170的男性，但却无法解释身高为170的男性要十多倍于身高为171的男性，两者的数量是连续的，这个数字差距这意味着170男性上相亲网站的比例是171男性的十多倍。因此，在「连续年龄上相亲网站的概率也连续变化」这个假设下，只有大量的身高虚报，才能解释上图在170处出现的堆积现象。虚报了多少呢？有两种假设。第一，假设世纪佳缘的人口分布和25-35岁的真实人口分布相同，那么虚报身高的为下图中点A和点C的距离，这个数字可以理解为虚报上限，为7厘米，第二，假设世纪佳缘人口和真实人口分布不同，但在170处分布连续变化，且下降速度和真实人口在该点的下降速度相同，那么虚报身高为下图中点B和点C的距离。由于在该假设下170以下人口偏少，分布下降速度会慢于真实人口分布，因此这个数字可以理解为虚报下限，为4厘米。因此，我们大概可以得出一个结论，在面对调查员时，身高170以下的25-35岁男性最多会做出3厘米左右的身高虚报，但在相亲市场上，则最多会做出4到7厘米的身高虚报。写到这里，可能有女士会开启嘲讽——相亲时声称自己为170的男性很只有166甚至163！显然，这个嘲讽是错的，因为这个数字只是我计算出的上限，他表示不会有男性虚报更多的身高，从165到169的男性都可能声称自己为170。其次，大家都能发现，上面的分析中我没有画出女性的情况，这是因为我懒。实际上女性身高的分布是这样的——可以看到，相亲市场上女性在160处的「堆积」现象，比男性还要严重，有五分之一的男性声称自己是170，同时有整整四分之一的女性声称自己是160！但这显然是不可能的。我们用类似的方法可以推算出，女性也是半斤八两，160以下的女性，在相亲市场上最多会将自己的身高虚报5-7厘米，下限比男性还要高。综上，从上文可以看到，不管是男性还是女性，在相亲市场上都会倾向于高报自己的身高，尤其是170以下的男性和160以下的女性，幅度最高可以达到7厘米之多。同时，我们并没有发现大家低报身高的证据。还有一件有趣的事——大家可以看到，虽然在相亲市场上大家的表现差不多，但女性的橙色线和灰色线之间的距离，没有男性那么远。这说明男性和女性在面对不认识的人（调查员）和潜在的相亲对象的反应差异是不一样的。从中大概可以得出另一个不太严谨的结论——在身高问题上，男性的撒谎是连续的，对不认识的人撒个中谎（最多虚报3厘米），对潜在相亲对象撒个大谎（最多虚报4-7厘米）。而女性的撒谎是跳跃的，对不认识的人会撒个小谎（最多虚报2厘米），对潜在相亲对象撒个比男性更大的谎（最多虚报5-7厘米）……

阅读原文