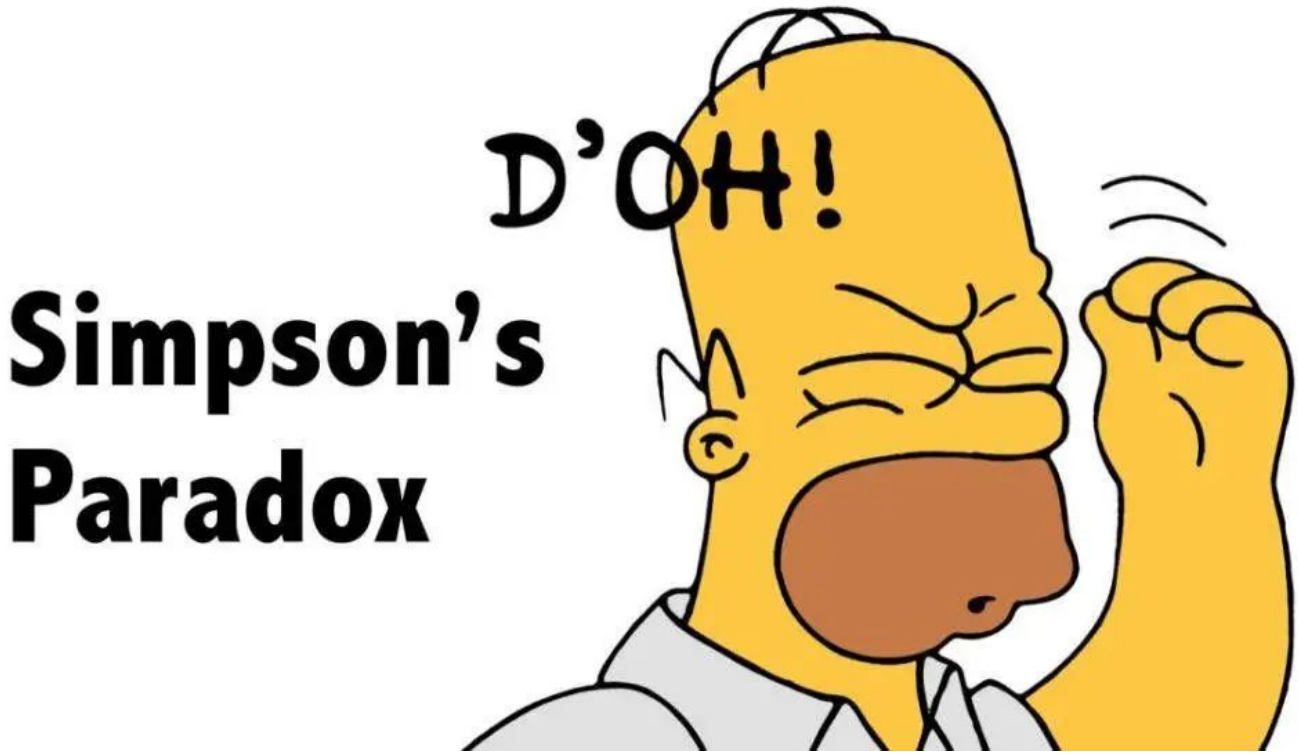


数据迷思2：辛普森悖论下的香港死亡数字

曹天元 Capo
科普作家

如果有两名篮球手A和B，本来，无论是两分球还是三分球，A都要比B投得准，但是一个赛季下来，我们在汇总数据的时候却发现：A的总体命中率居然比B要低！这可能吗？别说，还真有可能，而且在数据分析中极其常见。这就是以英国统计学家E.H.辛普森命名的所谓“辛普森悖论”。

我们可以举出具体的数字来证明这一点。如下表：

	两分球出手	两分球命中	两分球命中率	三分球出手	三分球命中	三分球命中率	总体出手	总体命中	总体命中率
A	100	60	60%	400	160	40%	500	220	44%
B	600	300	50%	200	60	30%	800	360	45%

从表中的数据，我们可以看出：A在赛季中的两分命中率是60/100=60%，而B是300/600=50%，A高于B。另外，A的三分球命中率是160/400=40%，而B则是30/100=30%，同样，A也高于B。

然而，如果把所有的数据“汇总”起来，计算一个整体的命中率，此时结论就会发生180度的反转。虽然A无论两分还是三分命中率都要高于B，但他的总体命中率却只有44%，低于B的45%！

这是怎么回事呢？如果仔细研究数据，我们会发现，这是因为三分球的命中率在整体上明显要比两分球低，所以，哪怕一个“好的三分投手”，其命中率也要低于一个“坏的两分投手”。现在，虽然A同时是一个“好的三分投手”和一个“好的两分投手”，而B同时是“坏的三分投手”以及“坏的两分投手”，但如果A一直热衷于投三分，而B则更多地投两分，那么，就算A在两者的命中率上都高于B，他的整体命中率也会被更多的三分球出手而大幅拉低，最后反而落后于B。

在现实当中，我们也很容易找到类似的例子。比方说NBA著名的神射手库里，我们可以把他跟76人的中锋恩比德做一下比较。库里职业生涯的两分球命中率是53.3%，高于恩比德的53.2%，而三分命中率则高达42.8%，更是远高于后者的33.8%。然而，由于库里出手的三分球比例要大大超过恩比德，这导致他的总体命中率只有47.3%，反而低于后者的49.0%。

然而，这说明什么呢？说明恩比德在整体上是一个比库里更优秀的射手吗？显然，没人会这么认为。事实上，库里无论是投两分，还是投三分，命中率都要比前者出色。只不过由于个人风格，或者战术安排等原因，他在比赛中更多地选择了“三分投手”的角色，而恩比德则更多地充当“内线”。这样一来，在两人的总出手次数当中，两分和三分球的比例就有很大不同。所以，是这个“战术原因”，而不是“技术原因”，才导致库里的整体命中率低于恩比德。但如果仔细考察分组数据，我们仍然可以得出结论：实际上库里才是那位更加出色的投手，无论是两分还是三分。

库里和恩比德：谁是更好的射手？

所以，辛普森悖论告诉我们，光看一个合并起来的“总数据”，有时候会具有欺骗性。很有可能，当我们把这个数据细分到更具体的组别时，会得到截然相反的结论。尤其是当这些组别之间存在着很大的整体性差异，而由于某种原因，数据又恰好在这些组别之间分布得很不均衡时，就特别容易导致辛普森悖论的出现。

现在，让我们回到上次提起的香港疫情死亡数字。乍看上去，香港因新冠死亡的人群当中，似乎高龄老人特别多，以80岁以上为例，占比高达71.05%。这是因为Omicron对老人特别“偏爱”吗?

在这里，我们需要首先明白一点，就是哪怕在自然状态下，每年“本来”就应该是老年人死得多，尤其是香港这样一个高度老龄化的城市。按2021年的情况，每年死亡约5万2千人，其中80岁以上占比57.31%。

但是，有人肯定要说了，本来只占57%的死亡，现在却占了71%，这还不能说明Omicron对老年人伤害更大？哎，这就是“辛普森悖论”所带来的错觉了。上回说了，在香港的例子里，我们还需要考虑到一个“潜在”的变量，统计学上称为lurking variable，就是在不同年龄层之间，存在着差异极大的疫苗接种率。

在本轮疫情爆发之前，港府为了推行疫苗，着实下了不少力气，比方说规定如果没有“疫苗通行证”的话，就不能进入各种公众场所，包括公务员不能上班，学生不能上学，不能进入特定的商场、超市、食肆，理发店等。在二月初甚至宣布过：未来如果没有疫苗通行证，将不得到公司工作。

众所周知，香港人向来是“返工大过天”。在如此严格的举措下，但凡有上学或工作需求的香港人，基本上都接种了疫苗。尤其是20-50岁之间的青壮年，根据港府公布的数字，接种人数甚至超过了香港在这些年龄层的总人口（这是因为港府公布的数字还包括非香港居民等）。

而与之形成鲜明对比的是，香港老年人的接种率却一直上不去。因为一方面，很多老年人并没有出行的刚需，又担心身体虚弱，经受不起疫苗的副作用。加上香港部分媒体长期炒作“打疫苗死了很多人”，在老年人当中造成了很大的恐慌。直到二月份疫情爆发时，香港仍有大量老年人连一针也未接种。在80岁以上的超高龄人群当中，未接种比例甚至接近一半。

所以事情很明显，香港老人在阳性人群中超高的死亡比例，很可能是因为更多老人没有去打疫苗而造成的一种假象。如果我们想要认真地探寻一下Omicron是不是对老年人危害更大，那么，首先需要严格地控制“是否打了疫苗”这个变量才行。

现在，为了简单起见，让我们把全体香港人分成两大组：接种0针和1针的归类为“未完成全程疫苗”，而接种2针或以上的则归类为“全程接种疫苗”。在某种程度上，你可以想象，现在香港被“分割”成了两座不同的城市，一座叫“无疫苗香港”，其居民没有任何人完成全程接种。而另一座叫“疫苗香港”，其居民全部完成了疫苗接种。

根据官方统计，这两座“城市”的人口数量和相应的年龄分布如下（注1）：

年龄组别	“无疫苗香港”总人数	“疫苗香港” 总人数
<3	123600	0
3-11	373376	129224
12-19	110111	337189

20-29	82159	689741
30-39	118508	975992
40-49	99101	1061199
50-59	141384	1049916
60-69	246062	876038
70-79	190155	401145
80+	217286	180914
总数	1701743	5701357

可以明显看出，由于青壮年基本都去打了疫苗，而大量老人则未接种，导致这两座“城市”的人口年龄分布出现了巨大的差异。相比之下，“无疫苗香港”的老龄化程度要比“疫苗香港”严重得多。

好，现在让我们来看看，Omicron对这两座“城市”分别造成了怎样的冲击。根据港府的报告，从今年初至5月11日为止，香港新冠死亡共9142人，其中有2人年龄“待定”，无法纳入统计，暂且排除。在剩下的9140人中，有8026人死在了“无疫苗香港”，而仅有1114人死在“疫苗香港”。考虑到前者的“总人口”仅有后者的1/3不到，其中死亡率差距之大，实在令人瞠目结舌。

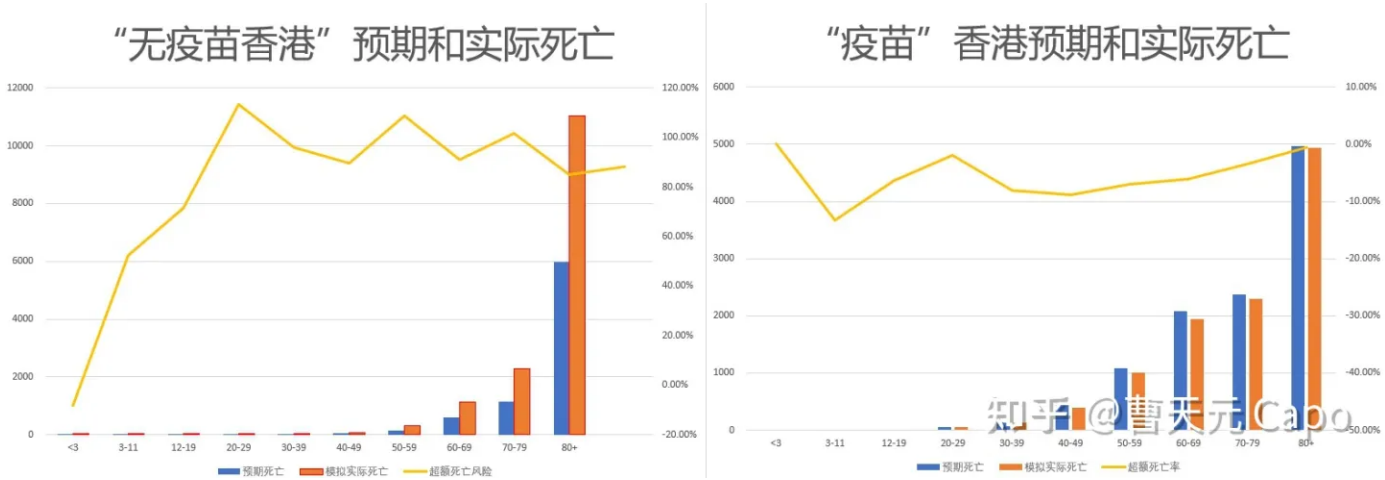
但是，死亡率高是一回事，这个高死亡率带来的额外风险是否有特别针对某个年龄段呢？为了研究这个事情，首先我们需要求出在“自然”状态下，“无疫苗香港”这座“城市”每年的预期死亡分布，然后再把它跟实际数字进行对比。这很容易，因为按照香港的“人口生命表”，我们可以获得每个年龄段每年的自然死亡率，再乘以无疫苗人口相应的年龄分布，就能得到最后的答案，如下表：

年龄组别	“无疫苗香港”总人数	每年预期死亡占比	新冠阳性死亡	阳性死亡占比
<3	123600	0.25%	1	0.01%
3-11	373376	0.12%	6	0.07%
12-19	110111	0.06%	4	0.05%
20-29	82159	0.08%	8	0.10%
30-39	118508	0.21%	18	0.22%
40-49	99101	0.52%	42	0.52%
50-59	141384	1.85%	179	2.23%
60-69	246062	7.37%	609	7.59%
70-79	190155	14.23%	1295	16.14%
80+	217286	75.32%	5864	73.06%
总数	1701743	100.00%	8026	100.00%

我们惊讶地发现，除了10岁以下的幼儿之外，对于所有的年龄段来说，这波疫情造成的死亡比例，相比“无疫苗香港”在自然状态下的正常死亡比例，几乎都是差不多的！比方说，对于80岁以上的老人，在所有8026个死亡案例当中，他们占了5864个，占比73.06%。但是，这个比例其实一点也不“高”，因为“无疫苗香港”本身就是一座比香港更加老龄化的“虚拟城市”。从上面的数字可以看到，在170万“总人口”当中，80岁以上老人有将近22万，远超香港原先的比例。因此，换算下来，他们每年本来就应该占总死亡人数的75.32%才对。相比之下，在未接种的新冠死者当中，高龄老人的比例其实跟自然预期值相差无几，甚至还要略少。

这说明什么问题呢？显然，虽然在“无疫苗”的人群当中，绝对死亡数确实大大增加了，但是，死亡年龄的分布却仍然是“正常”的。也就是说，在没有接种疫苗的情况下，Omicron其实对所有年龄层的人都产生了同样的冲击，而并没有特别针对老年人。你可以想象，它就像是一个“死亡放大镜”，对所有年龄的死亡人数都一律“按同比例”放大。这跟我们上次得出的结论是一致的：新冠其实对所有年龄（极低龄除外）“一视同仁”，并没有对老年人造成特别大的额外伤害。

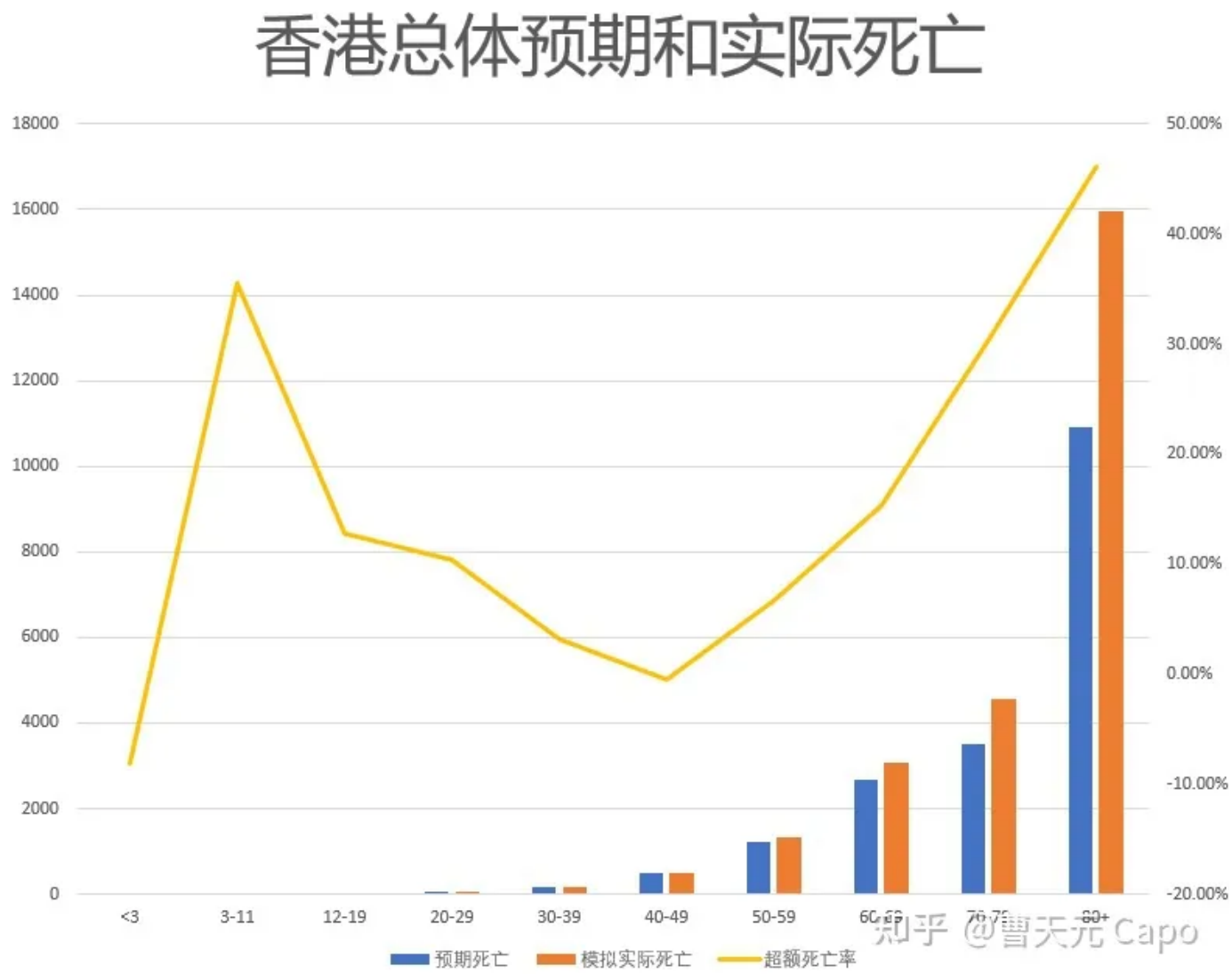
为了更加直观起见，我们还可以通过生命表估算出从年初至今，“疫苗香港”和“无疫苗香港”本来应该产生多少死亡，然后再通过模型，模拟出两者“实际上”到底各自死了多少人（注2），并与前者进行对比。结果如下图：



可以看出，一方面，在“无疫苗香港”，情况比较悲惨。这座“城市”以170万的总人口，年初至今预期死亡7918人，而模拟实际死亡为14894人，“多死”了6976人，期间总体超额死亡比例高达88%。不过，正如之前说的，除了10岁以下的幼儿之外，这个超额风险是各个年龄层“均匀承担”的，大致都在85%-110%之间，变化不大。

另一方面，在“疫苗香港”，则几乎没有超额死亡。事实上，模型给出的超额死亡率是-3.27%。在这座人口为570万的“城市”当中，本来年初至今，预计死亡11154人，而模拟实际死亡为10790人，甚至“少死”了364人。值得一提的是，这些少死的人，也基本符合该城市的年龄“自然分布”，换句话说，虽然超额死亡风险是负数，但也基本上由各个年龄层“均匀承担”，基本上都在-10%-0%之间轻微变动。从中，我们可以得出另一个结论，就是疫苗的保护作用也并没有明显的年龄偏好，它带来的“福利”，基本上也仍然是按比例“平均分配”给各个年龄层的。

然而，如果我们把两座“城”放在一起，把它们的数字汇总起来，“神奇”的现象就出现了。本来，在每一座“分城”当中，新冠带来的超额死亡风险都并不随年龄剧烈波动，但一旦把它们合起来，事情就发生了变化，超额死亡率曲线开始剧烈地上下起伏，而且看上去，似乎老年人的“风险”变得更大了。

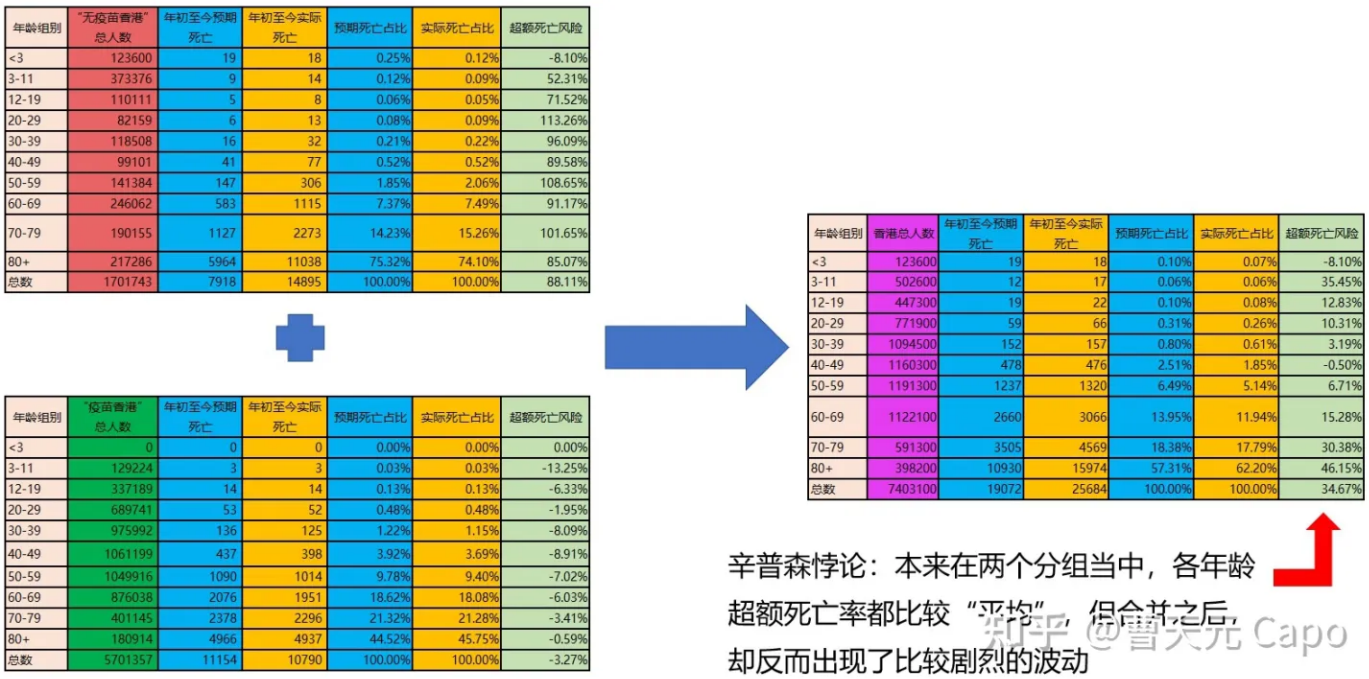


比方说，如果我们抽取两个年龄组做比较，一个是20-29岁，一个是70-79岁。本来，在“无疫苗香港”组，前者的额外死亡率是113%，后者是102%，明明是前者略高于后者。而在“疫苗香港”组，前者的额外死亡率是-1.95%，后者是-3.41%。因为是负数，所以仍然是前者略高于后者。

但是，把数据合并之后，我们就会惊讶地发现：20多岁年轻人的“总体”超额死亡风险为10.31%，而70多岁老年人的“总体”超额风险则高达30.38%！突然之间，后者远远超过了前者。

为什么在每一个分组当中，都是前者比后者高，而合起来之后，却反而变成后者比前者高？哎，这就是我们一开头提到的，因为“辛普森悖论”而带来的错觉了。简单来说，因为不接种疫苗组，其整体超额风险远高于接种疫苗组，而年轻人不接种疫苗的少，接种疫苗的多，老年人则正好相反。因此合并数据之后，经过加权，前者的数据就会更多受到“接种疫苗”带来的影响，后者的数据则更多受到“不接种疫苗”带来的影响。最后，就出现了系统性的差别。在这里，疫苗接种率被称为一个

“对撞因子”（Collider），它和“年龄”还有“超额死亡率”两个变量同时相关。因此，如果不仔细控制疫苗接种率这个变量，我们就很可能得出一个整体上似是而非的错误结论。



年龄组别	“无疫苗香港”总人数	年初至今预期死亡	年初至今实际死亡	预期死亡占比	实际死亡占比	超额死亡风险
<3	123600	19	18	0.25%	0.12%	-8.10%
3-11	373376	9	14	0.12%	0.09%	52.31%
12-19	110111	5	8	0.06%	0.05%	71.52%
20-29	82159	6	13	0.08%	0.09%	113.26%
30-39	118508	16	32	0.21%	0.22%	96.09%
40-49	99101	41	77	0.52%	0.52%	89.58%
50-59	141384	147	306	1.85%	2.06%	108.65%
60-69	246062	583	1115	7.37%	7.49%	91.17%
70-79	190155	1127	2273	14.23%	15.26%	101.65%
80+	217286	5964	11038	75.32%	74.10%	85.07%
总数	1701743	7918	14895	100.00%	100.00%	88.11%

年龄组别	“疫苗香港”总人数	年初至今预期死亡	年初至今实际死亡	预期死亡占比	实际死亡占比	超额死亡风险
<3	0	0	0	0.00%	0.00%	0.00%
3-11	129224	3	3	0.03%	0.03%	-13.25%
12-19	337189	14	14	0.13%	0.13%	-6.33%
20-29	689741	53	52	0.48%	0.48%	-1.95%
30-39	975992	136	125	1.22%	1.15%	-8.09%
40-49	1061199	437	398	3.92%	3.69%	-8.91%
50-59	1049916	1090	1014	9.78%	9.40%	-7.02%
60-69	876038	2076	1951	18.62%	18.08%	-6.03%
70-79	401145	2378	2296	21.32%	21.28%	-3.41%
80+	180914	4966	4937	44.52%	45.75%	-0.59%
总数	5701357	11154	10790	100.00%	100.00%	-3.27%

年龄组别	香港总人数	年初至今预期死亡	年初至今实际死亡	预期死亡占比	实际死亡占比	超额死亡风险
<3	123600	19	18	0.10%	0.07%	-8.10%
3-11	502600	12	17	0.06%	0.06%	35.45%
12-19	447300	19	22	0.10%	0.08%	12.83%
20-29	771900	59	66	0.31%	0.26%	10.31%
30-39	1094500	152	157	0.80%	0.61%	3.19%
40-49	1160300	478	476	2.51%	1.85%	-0.50%
50-59	1191300	1237	1320	6.49%	5.14%	6.71%
60-69	1122100	2660	3066	13.95%	11.94%	15.28%
70-79	591300	3505	4569	18.38%	17.79%	30.38%
80+	398200	10930	15974	57.31%	62.20%	46.15%
总数	7403100	19072	25684	100.00%	100.00%	34.67%

当然，很多人肯定还会想到，关于疫苗接种问题上，还存在另外一个“对撞因子”，就是“疫苗接种意愿”，它和“身体健康程度”以及“接种率”同时都有关系。简单地说，就是身体越差，越有基础病的人，就越是“不愿意”去接种疫苗，而这些人以老年为多。这样一来，就会造成一个“自我选择”的偏差，导致老年人更多地不去接种，最后造成疫情中的死亡率偏高。

无疑，这也是一个问题，不过，从目前的数据看来，自我选择也许会导致疫苗的效率被高估（比如说“疫苗香港”甚至出现了负数的超额死亡，这很可能是因为健康人群自我选择导致的，而并非完全是疫苗本身的作用）。但是，它似乎并没有造成总体上的年龄偏差。简单地说，如果身体虚弱的老人不愿意去接种疫苗，那么，身体虚弱的年轻人也会做出同样的选择，而他们之间的比例仍然是“自然”的。关于疫苗的问题，我们以后有机会再来谈。

总之，由于辛普森悖论的存在，我们在分析数据的时候，时刻需要留意，是否其中存在着潜在的“对撞因子”？否则，光是单看整体的数字，得出的结论很可能会南辕北辙。

注1：人口数字和疫苗接种情况分别来自香港政府网站上的人口报告和“新冠死亡个案报告初步数据分析”文档。但是，后者关于疫苗接种的详细数据最早只能追溯到4月21日，无法反应疫情初起时的状态（疫情爆发后，香港的疫苗接种也迎来了一轮高峰，所以如今的数字要远高于当初）。加上之前提到过的，由于接种数字当中还包括在香港工作的非本地居民等，导致有些年龄段的接种数甚至大于总人口。为此，我们对这些数据进行了一些处理，降低总接种人口的比例，对于某些年龄段

还要乘上0.98-0.99不等的系数，使得未接种人数不至于是负数。总而言之，这里的数据尽量试图还原二月底时的疫苗接种状态。

注2：模型采用的估算方法，跟我们在上海案例中用的办法是类似的。在估算超额死亡率时同样如此，就是根据阳性人数每日的变化，画出一条“阳性活跃曲线”，然后将这条曲线对时间做积分，求出其占“全民总时间”的比例。这样就可以知道所有的阳性人口在活跃期间“应该”正常死亡多少人，以便和实际报告死亡数对比。

当然，香港的情况稍微有些不同。第一，港府判断死亡人数的标准是“新冠检测阳性后28天”，只要在这个期间死亡的都算。所以我们应该画的是“28天内阳性活跃曲线”，而不是“每日活跃”。第二，港府至今仅报告了117万个阳性病例，但因为香港从未进行过全民核酸筛查，疫情高峰起来之后更是干脆放弃了严格的检测，所以这个数字很明显是大大低估的。事实上，早在3月22日，港大的报告就认为当时至少已经感染了400万人。

由于缺乏可靠的检测数据，我们只能根据各种其他信息，对模型进行调整和测试，以拟合实际发生的情况。就目前使用的参数来说，它显示至今为止，香港总共感染病毒人口已高达550万之多，几乎已经快要达到群体免疫阈值（这也就是为什么香港疫情如今大大放缓的原因）。根据该模型，全香港从年初1月1日至5月14日，“本该”死亡19072人，而实际死亡25684人，“多死”了6612人，期间超额了34%。

有人可能会质疑模型的准确度，但是，模型给出的数字本身有多准确，在这里并不重要，只是用来举例而已。实际上，超额死亡率肯定是一个定值，所以就算有误差，相差的无非就是一个比例。这最多影响具体的数字，而并不影响文中的结论，也就是超额死亡率的分布，在“接种疫苗”和“未接种”两个分组当中，并不和年龄分布高度相关。