資料迷思2：辛普森悖論下的香港死亡數字

曹天元 Capo
科普作家

如果有兩名籃球手A和B，本來，無論是兩分球還是三分球，A都要比B投得準，但是一個賽季下來，我們在彙總資料的時候卻發現：A的總體命中率居然比B要低！這可能嗎？別說，還真有可能，而且在資料分析中極其常見。這就是以英國統計學家E.H.辛普森命名的所謂“辛普森悖論”。

我們可以舉出具體的數字來證明這一點。如下表：

	兩分球出手	兩分球命中	兩分球命中率	三分球出手	三分球命中	三分球命中率	總體出手	總體命中	總體命中率
A	100	60	60%	400	160	40%	500	220	44%
B	600	300	50%	200	60	30%	800	360	45%

從表中的資料，我們可以看出：A在賽季中的兩分命中率是60/100=60%，而B是300/600=50%，A高於B。另外，A的三分球命中率是160/400=40%，而B則是30/100=30%，同樣，A也高於B。

然而，如果把所有的資料“彙總”起來，計算一個整體的命中率，此時結論就會發生180度的反轉。雖然A無論兩分還是三分命中率都要高於B，但他的總體命中率卻只有44%，低於B的45%！

這是怎麼回事呢？如果仔細研究資料，我們會發現，這是因為三分球的命中率在整體上明顯要比兩分球低，所以，哪怕一個“好的三分投手”，其命中率也要低於一個“壞的兩分投手”。現在，雖然A同時是一個“好的三分投手”和一個“好的兩分投手”，而B同時是“壞的三分投手”以及“壞的兩分投手”，但如果A一直熱衷於投三分，而B則更多地投兩分，那麼，就算A在兩者的命中率上都高於B，他的整體命中率也會被更多的三分球出手而大幅拉低，最後反而落後於B。

在現實當中，我們也很容易找到類似的例子。比方說NBA著名的神射手庫裡，我們可以把他跟76人的中鋒恩比德做一下比較。庫裡職業生涯的兩分球命中率是53.3%，高於恩比德的53.2%，而三分命中率則高達42.8%，更是遠高於後者的33.8%。然而，由於庫裡出手的三分球比例要大大超過恩比德，這導致他的總體命中率只有47.3%，反而低於後者的49.0%。

然而，這說明什麼呢？說明恩比德在整體上是一個比庫裡更優秀的射手嗎？顯然，沒人會這麼認為。事實上，庫裡無論是投兩分，還是投三分，命中率都要比前者出色。只不過由於個人風格，或者戰術安排等原因，他在比賽中更多地選擇了“三分投手”的角色，而恩比德則更多地充當“內線”。這樣一來，在兩人的總出手次數當中，兩分和三分球的比例就有很大不同。所以，是這個“戰術原因”，而不是“技術原因”，才導致庫裡的整體命中率低於恩比德。但如果仔細考察分組資料，我們仍然可以得出結論：實際上庫裡才是那位更加出色的投手，無論是兩分還是三分。

庫裡和恩比德：誰是更好的射手？

所以，辛普森悖論告訴我們，光看一個合併起來的“總資料”，有時候會具有欺騙性。很有可能，當我們把這個資料細分到更具體的組別時，會得到截然相反的結論。尤其是當這些組別之間存在著很大的整體性差異，而由於某種原因，資料又恰好在這些組別之間分佈得很不均衡時，就特別容易導致辛普森悖論的出現。

現在，讓我們回到上次提起的香港疫情死亡數字。乍看上去，香港因新冠死亡的人群當中，似乎高齡老人特別多，以80歲以上為例，佔比高達71.05%。這是因為Omicron對老人特別“偏愛”嗎?

在這裡，我們需要首先明白一點，就是哪怕在自然狀態下，每年“本來”就應該是老年人死得多，尤其是香港這樣一個高度老齡化的城市。按2021年的情況，每年死亡約5萬2千人，其中80歲以上佔比57.31%。

但是，有人肯定要說了，本來只佔57%的死亡，現在卻佔了71%，這還不能說明Omicron對老年人傷害更大? 哎，這就是“辛普森悖論”所帶來的錯覺了。上回說了，在香港的例子裡，我們還需要考慮到一個“潛在”的變數，統計學上稱為lurking variable，就是在不同年齡層之間，存在著差異極大的疫苗接種率。

在本輪疫情爆發之前，港府為了推行疫苗，著實下了不少力氣，比方說規定如果沒有“疫苗通行證”的話，就不能進入各種公眾場所，包括公務員不能上班，學生不能上學，不能進入特定的商場、超市、食肆，理髮店等。在二月初甚至宣佈過：未來如果沒有疫苗通行證，將不得到公司工作。

眾所周知，香港人向來是“返工大過天”。在如此嚴格的舉措下，但凡有上學或工作需求的香港人，基本上都接種了疫苗。尤其是20-50歲之間的青壯年，根據港府公佈的數字，接種人數甚至超過了香港在這些年齡層的總人口（這是因為港府公佈的數字還包括非香港居民等）。

而與之形成鮮明對比的是，香港老年人的接種率卻一直上不去。因為一方面，很多老年人並沒有出行的剛需，又擔心身體虛弱，經受不起疫苗的副作用。加上香港部分媒體長期炒作“打疫苗死了很多人”，在老年人當中造成了很大的恐慌。直到二月份疫情爆發時，香港仍有大量老年人連一針也未接種。在80歲以上的超高齡人群當中，未接種比例甚至接近一半。

所以事情很明顯，香港老人在陽性人群中超高的死亡比例，很可能是因為更多老人沒有去打疫苗而造成的一種假象。如果我們想要認真地探尋一下Omicron是不是對老年人危害更大，那麼，首先需要嚴格地控制“是否打了疫苗”這個變數才行。

現在，為了簡單起見，讓我們把全體香港人分成兩大組：接種0針和1針的歸類為“未完成全程疫苗”，而接種2針或以上的則歸類為“全程接種疫苗”。在某種程度上，你可以想像，現在香港被“分割”成了兩座不同的城市，一座叫“無疫苗香港”，其居民沒有任何人完成全程接種。而另一座叫“疫苗香港”，其居民全部完成了疫苗接種。

根據官方統計，這兩座“城市”的人口數量和相應的年齡分佈如下（注1）：

年齡組別	“無疫苗香港”總人數	“疫苗香港” 總人數
<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
總數	1701743	5701357

可以明顯看出，由於青壯年基本都去打了疫苗，而大量老人則未接種，導致這兩座“城市”的人口年齡分佈出現了巨大的差異。相比之下，“無疫苗香港”的老齡化程度要比“疫苗香港”嚴重得多。

好，現在讓我們來看看，Omicron對這兩座“城市”分別造成了怎樣的衝擊。根據港府的報告，從今年初至5月11日為止，香港新冠死亡共9142人，其中有2人年齡“待定”，無法納入統計，暫且排除。在剩下的9140人中，有8026人死在了“無疫苗香港”，而僅有1114人死在“疫苗香港”。考慮到前者的“總人口”僅有後者的1/3不到，其中死亡率差距之大，實在令人瞠目結舌。

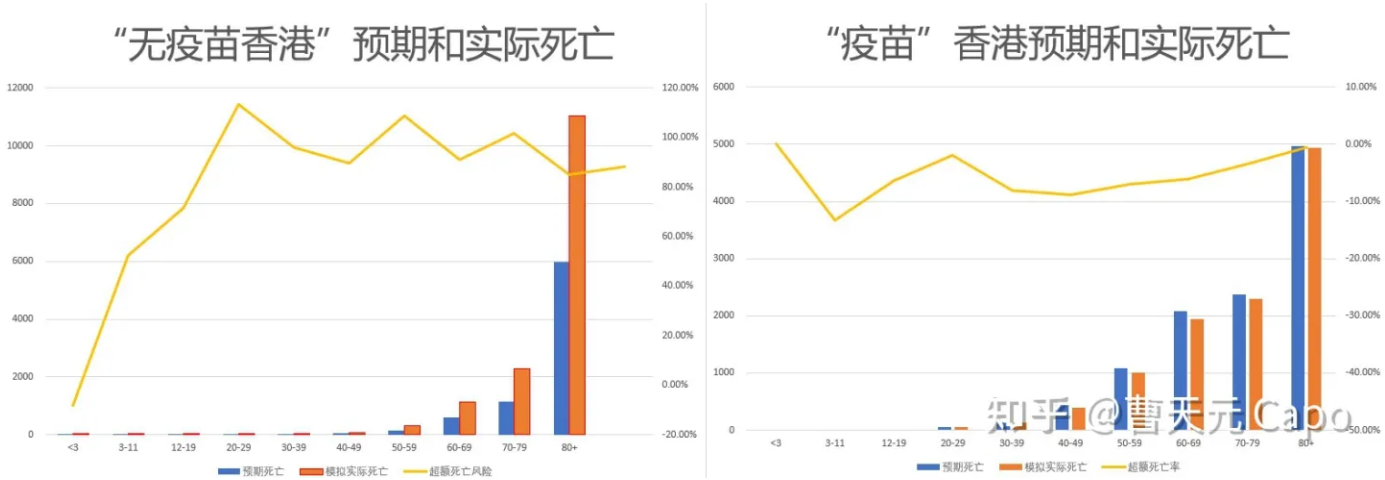
但是，死亡率高是一回事，這個高死亡率帶來的額外風險是否有特別針對某個年齡段呢？為了研究這個事情，首先我們需要求出在“自然”狀態下，“無疫苗香港”這座“城市”每年的預期死亡分佈，然後再把它跟實際數字進行對比。這很容易，因為按照香港的“人口生命表”，我們可以獲得每個年齡段每年的自然死亡率，再乘以無疫苗人口相應的年齡分佈，就能得到最後的答案，如下表：

年齡組別	“無疫苗香港”總人數	每年預期死亡佔比	新冠陽性死亡	陽性死亡佔比
<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%
總數	1701743	100.00%	8026	100.00%

我們驚訝地發現，除了10歲以下的幼兒之外，對於所有的年齡段來說，這波疫情造成的死亡比例，相比“無疫苗香港”在自然狀態下的正常死亡比例，幾乎都是差不多的！比方說，對於80歲以上的老人，在所有8026個死亡案例當中，他們佔了5864個，佔比73.06%。但是，這個比例其實一點也不“高”，因為“無疫苗香港”本身就是一座比香港更加老齡化的“虛擬城市”。從上面的數字可以看到，在170萬“總人口”當中，80歲以上老人有將近22萬，遠超香港原先的比例。因此，換算下來，他們每年本來就應該佔總死亡人數的75.32%才對。相比之下，在未接種的新冠死者當中，高齡老人的比例其實跟自然預期值相差無幾，甚至還要略少。

這說明什麼問題呢？顯然，雖然在“無疫苗”的人群當中，絕對死亡數確實大大增加了，但是，死亡年齡的分佈卻仍然是“正常”的。也就是說，在沒有接種疫苗的情況下，Omicron其實對所有年齡層的人都產生了同樣的衝擊，而並沒有特別針對老年人。你可以想像，它就像是一個“死亡放大鏡”，對所有年齡的死亡人數都一律“按同比例”放大。這跟我們上次得出的結論是一致的：新冠其實對所有年齡（極低齡除外）“一視同仁”，並沒有對老年人造成特別大的額外傷害。

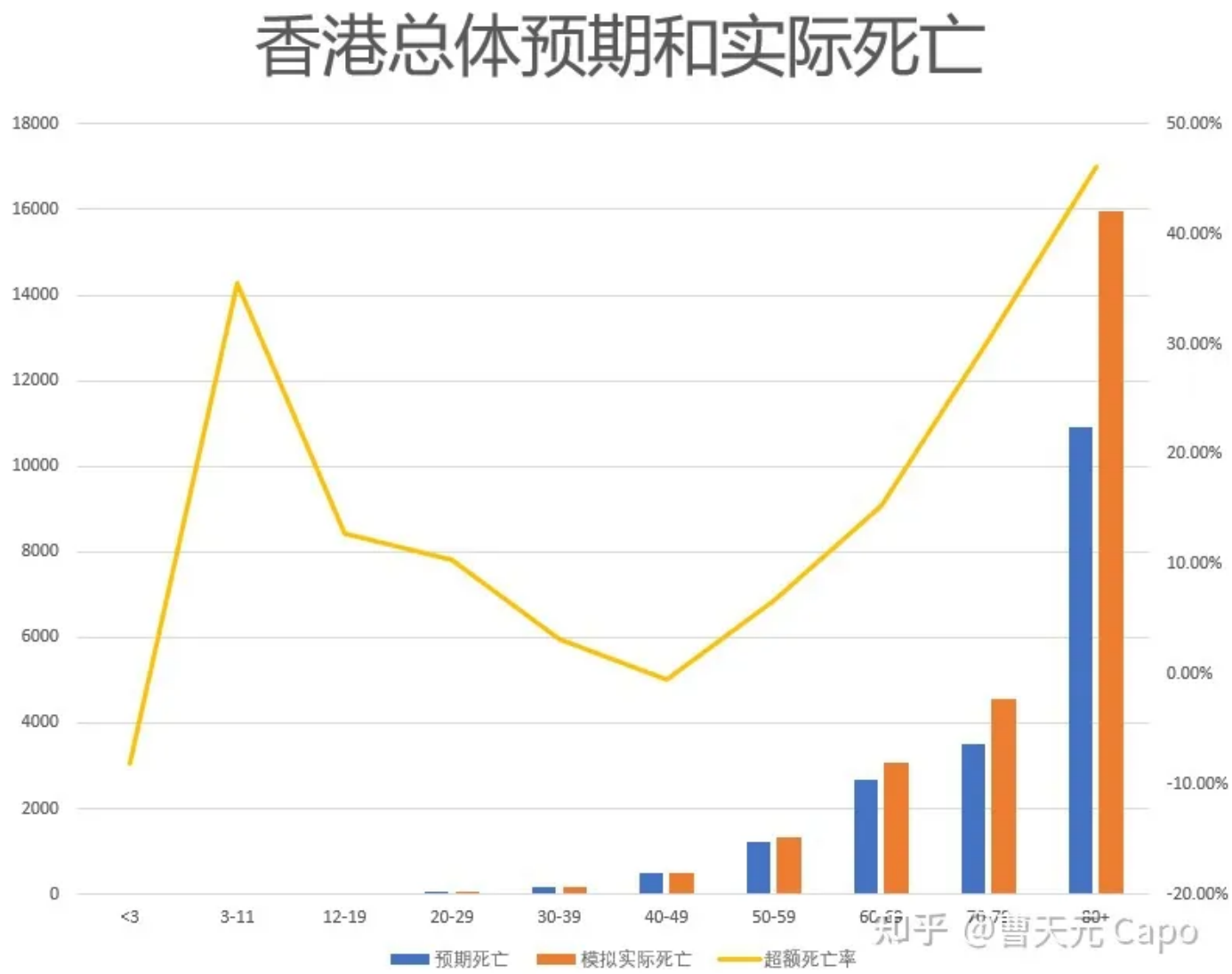
為了更加直觀起見，我們還可以通過生命表估算出從年初至今，“疫苗香港”和“無疫苗香港”本來應該產生多少死亡，然後再通過模型，模擬出兩者“實際上”到底各自死了多少人（注2），並與前者進行對比。結果如下圖：



可以看出，一方面，在“無疫苗香港”，情況比較悲慘。這座“城市”以170萬的總人口，年初至今預期死亡7918人，而模擬實際死亡為14894人，“多死”了6976人，期間總體超額死亡比例高達88%。不過，正如之前說的，除了10歲以下的幼兒之外，這個超額風險是各個年齡層“均勻承擔”的，大致都在85%-110%之間，變化不大。

另一方面，在“疫苗香港”，則幾乎沒有超額死亡。事實上，模型給出的超額死亡率是-3.27%。在這座人口為570萬的“城市”當中，本來年初至今，預計死亡11154人，而模擬實際死亡為10790人，甚至“少死”了364人。值得一提的是，這些少死的人，也基本符合該城市的年齡“自然分佈”，換句話說，雖然超額死亡風險是負數，但也基本上由各個年齡層“均勻承擔”，基本上都在-10%-0%之間輕微變動。從中，我們可以得出另一個結論，就是疫苗的保護作用也並沒有明顯的年齡偏好，它帶來的“福利”，基本上也仍然是按比例“平均分配”給各個年齡層的。

然而，如果我們把兩座“城”放在一起，把它們的數字彙總起來，“神奇”的現象就出現了。本來，在每一座“分城”當中，新冠帶來的超額死亡風險都並不隨年齡劇烈波動，但一旦把它們合起來，事情就發生了變化，超額死亡率曲線開始劇烈地上下起伏，而且看上去，似乎老年人的“風險”變得更大了。

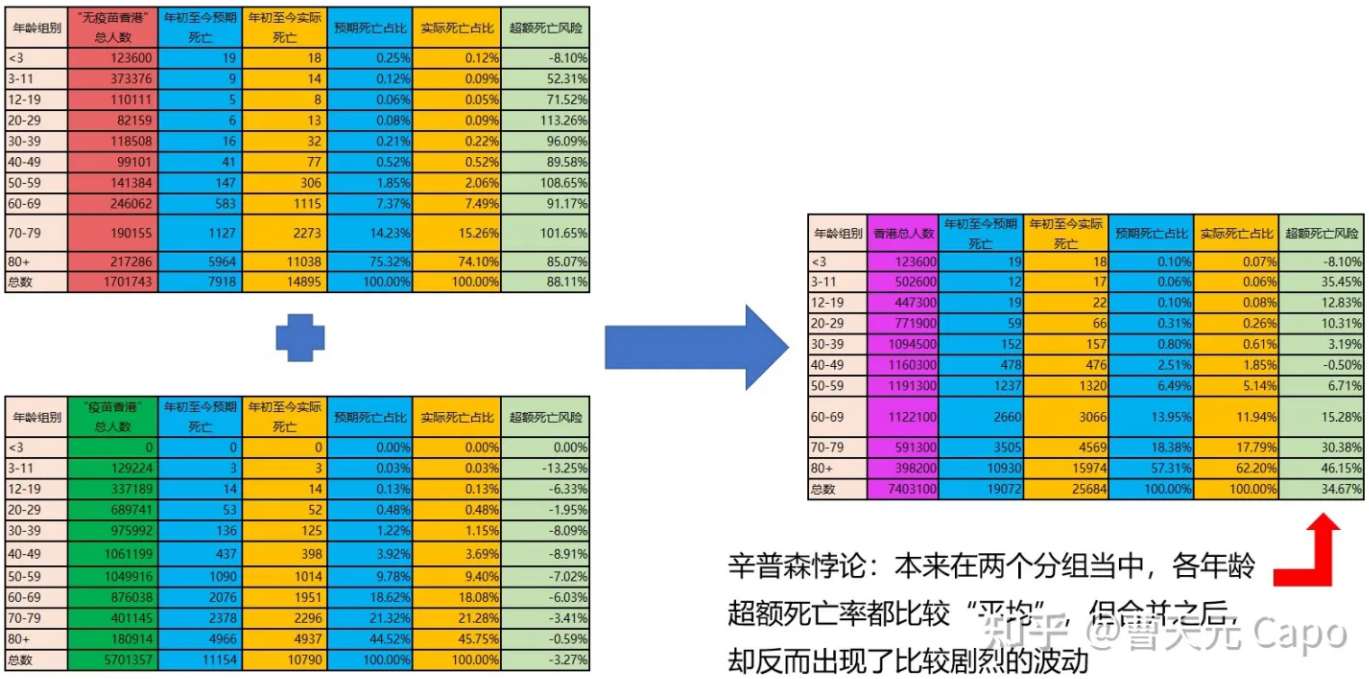


比方說，如果我們抽取兩個年齡組做比較，一個是20-29歲，一個是70-79歲。本來，在“無疫苗香港”組，前者的額外死亡率是113%，後者是102%，明明是前者略高於後者。而在“疫苗香港”組，前者的額外死亡率是-1.95%，後者是-3.41%。因為是負數，所以仍然是前者略高於後者。

但是，把資料合併之後，我們就會驚訝地發現：20多歲年輕人的“總體”超額死亡風險為10.31%，而70多歲老年人的“總體”超額風險則高達30.38%！突然之間，後者遠遠超過了前者。

為什麼在每一個分組當中，都是前者比後者高，而合起來之後，卻反而變成後者比前者高？哎，這就是我們一開頭提到的，因為“辛普森悖論”而帶來的錯覺了。簡單來說，因為不接種疫苗組，其整體超額風險遠高於接種疫苗組，而年輕人不接種疫苗的少，接種疫苗的多，老年人則正好相反。因此合併資料之後，經過加權，前者的資料就會更多受到“接種疫苗”帶來的影響，後者的資料則更多受到“不接種疫苗”帶來的影響。最後，就出現了系統性的差別。在這裡，疫苗接種率被稱為一個

“對撞因子”（Collider），它和“年齡”還有“超額死亡率”兩個變數同時相關。因此，如果不仔細控制疫苗接種率這個變數，我們就很可能得出一個整體上似是而非的錯誤結論。



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%

辛普森悖论：本来在两个分组当中，各年龄超额死亡率都比较“平均”，但合并之后，却反而出现了比较剧烈的波动

當然，很多人肯定還會想到，關於疫苗接種問題上，還存在另外一個“對撞因子”，就是“疫苗接種意願”，它和“身體健康程度”以及“接種率”同時都有關係。簡單地說，就是身體越差，越有基礎病的人，就越是“不願意”去接種疫苗，而這些人以老年為多。這樣一來，就會造成一個“自我選擇”的偏差，導致老年人更多地不去接種，最後造成疫情中的死亡率偏高。

無疑，這也是一個問題，不過，從目前的資料看來，自我選擇也許會導致疫苗的效率被高估（比如說“疫苗香港”甚至出現了負數的超額死亡，這很可能是因為健康人群自我選擇導致的，而並非完全是疫苗本身的作用）。但是，它似乎並沒有造成總體上的年齡偏差。簡單地說，如果身體虛弱的老人不願意去接種疫苗，那麼，身體虛弱的年輕人也會做出同樣的選擇，而他們之間的比例仍然是“自然”的。關於疫苗的問題，我們以後有機會再來談。

總之，由於辛普森悖論的存在，我們在分析資料的時候，時刻需要留意，是否其中存在著潛在的“對撞因子”？否則，光是單看整體的數字，得出的結論很可能會南轅北轍。

注1：人口數字和疫苗接種情況分別來自香港政府網站上的人口報告和“新冠死亡個案報告初步資料分析”文件。但是，後者關於疫苗接種的詳細資料最早只能追溯到4月21日，無法反應疫情初起時的狀態（疫情爆發後，香港的疫苗接種也迎來了一輪高峰，所以如今的數字要遠高於當初）。加上之前提到過的，由於接種數字當中還包括在香港工作的非本地居民等，導致有些年齡段的接種數甚至大於總人口。為此，我們對這些資料進行了一些處理，降低總接種人口的比例，對於某些年齡段

還要乘上0.98-0.99不等的係數，使得未接種人數不至於是負數。總而言之，這裡的資料儘量試圖還原二月底時的疫苗接種狀態。

注2：模型採用的估算方法，跟我們在上海案例中用的辦法是類似的。在估算超額死亡率時同樣如此，就是根據陽性人數每日的變化，畫出一條“陽性活躍曲線”，然後將這條曲線對時間做積分，求出其佔“全民總時間”的比例。這樣就可以知道所有的陽性人口在活躍期間“應該”正常死亡多少人，以便和實際報告死亡數對比。

當然，香港的情況稍微有些不同。第一，港府判斷死亡人數的標準是“新冠檢測陽性後28天”，只要在這個期間死亡的都算。所以我們應該畫的是“28天內陽性活躍曲線”，而不是“每日活躍”。第二，港府至今僅報告了117萬個陽性病例，但因為香港從未進行過全民核酸篩查，疫情高峰起來之後更是乾脆放棄了嚴格的檢測，所以這個數字很明顯是大大低估的。事實上，早在3月22日，港大的報告就認為當時至少已經感染了400萬人。

由於缺乏可靠的檢測資料，我們只能根據各種其他資訊，對模型進行調整和測試，以擬合實際發生的情況。就目前使用的參數來說，它顯示至今為止，香港總共感染病毒人口已高達550萬之多，幾乎已經快要達到群體免疫閾值（這也就是為什麼香港疫情如今大大放緩的原因）。根據該模型，全香港從年初1月1日至5月14日，“本該”死亡19072人，而實際死亡25684人，“多死”了6612人，期間超額了34%。

有人可能會質疑模型的精準度，但是，模型給出的數字本身有多精準，在這裡並不重要，只是用來舉例而已。實際上，超額死亡率肯定是一個定值，所以就算有誤差，相差的無非就是一個比例。這最多影響具體的數字，而並不影響文中的結論，也就是超額死亡率的分佈，在“接種疫苗”和“未接種”兩個分組當中，並不和年齡分佈高度相關。