下述模型基于这样一个民政局的统计数据分析（见最后附件）：其中用到的数据为：400个监护缺失儿童。想通过400来分析总的感染人数。这里主要想指出网络常见的一种分析方法的错误之处，以及基于此新闻的正确估计方法。本文纯数学分析，并无对去世者的不敬。尽量使用恭敬语言，以免引起不适@Gaoda

先上结论，不想看过程的看完结论可以直接关闭文件。想了解为什么常见分析是错误的，可以看看(a)，（b），（c）。希望讨论的最好看完全文。

结论1：病故率A=1%，家庭患者率20%，95%的信心，则X=15，400个病故家庭对应总患者数：600,000=60万。如果400其中只有200个是这个模型对应的，则总患者数30万。以下是个对应表格：

结论2：400个病故家庭对应总患者数上限：364万。注意达到这个上限的假设很苛刻，概率非常低。

|  |  |  |
| --- | --- | --- |
| 已知病故家庭数 | 估计出的总患者数 | 估计出的总患者数上限 |
| 400 | 60万 | 364万 |
| 200 | 30万 | 182万 |
| 100 | 15万 | 91万 |

结论3：@DonaldQ猜的家庭病故率0.5%，很准，但这几乎是一个下限，真实值有很大的概率比这个要大。

--------------------------结论结束，请自行关闭文件----------------------------

本文程序下载地址：

https://github.com/miscvariables/covid-19-models/tree/master/model1

(0)本分析的所有事件假设条件独立，并无条件概率，即不考虑a患者病故和b患者病故的关联，所有病故都是独立事件，即使是一个家庭内部。注意如果考虑事件的非独立性，则结论会比现在估计的人数更少。

(1)以下先描述网络常见的一个错误模型：1%的病故率，则一个家庭内部二人a和b同时患病的情况下，同时病故的概率为1%\*1%=1/10000，则总的患者人数为：400\*10000=4,000,000。注意该模型认为，4,000,000是一个下限，因为只考虑了一个家庭a和b同时患病的情况1，而没有计入家庭只有1个人患病的情况2。总人数应该包含：情况1+情况2，所以4,000,000只是一个下限。

(a)上述模型的错误在哪里？正如@DonaldQ早期一针见血地指出那样，这里有个概率抽样空间的问题，1%这个病故概率是不能直接用的。我们先看1%怎么来的（定义为**病故率A**），1%是以所有确诊的人为抽样数量，病故率A=病故人数A/总确诊人数=1%。但在上述模型中，我们需要的病故率是以家庭为单位的抽样，即首先抽出所有家庭2人同时感染的样本，然后计算：病故率B=病故人数B/(确诊家庭\*2)，2是指家庭含2人。基于**病故率B**，再用上面的模型计算才是正确的。

(b)病故率B从公开的数据无法获得，如@Jiong所说，我们需要更详细的统计数据才能做这个计算。那是否我们只能就此停住无法进行估计了呢？答案是否定的，只需简单地引入另外一个很容易估算的参数，就可以把病故率B的范围准确地估计出来。

(c)一个马上的疑问是，病故率B比A大还是小？因为上面两个概率计算分子分母同时改变，所以可能大也可能小。如果病故率B<病故率A，那上述模型所谓的估计下限依然成立，实际人数比400万还要大得多！幸运的是，在样本完全随机的情况下，从A是可以估算B的。先把最后结论写在这里：我们有很强的信心（概率大于99%）认为，B远远大于A。接下来进入正题

(d)如前分析，要计算病故率B，我们需要对数据重新抽样，抽出家庭患者，这里有一个简单的办法，引入家庭患者比例。基于众所周知的原因，家庭成员同时感染的概率比较高，这是有共识的。我们引入：家庭患者率=(确诊家庭\*2)/总确诊人数。这个概率每个人很容易想象其范围，我觉得一个合理的范围是5%-30%。我们选择5%这个下限作为参数。注意如果选择大于5%的数值，只会让总的估计人数更少，所以这里做出的计算可以说是最保守的计算，估计的是总人数的上限。

(e)总结一下，现在模型变成了总确诊人数N（比如10000），病故率A=1%，家庭患者率5%，求病故率B。

(f)直接求B是做不到的，但我们可以以很高的置信度估计B的范围。把它转换成这样一个问题：总确诊人数N（比如10000），病故率A=1%，家庭患者率5%，问我们需要在多少个病故患者中（定义为数量X），才能找到一个家庭2人同时病故的情况。得到X以后，如何计算病故率B后面会分析。为什么要转换成这个问题，因为这是一个稍作变化的生日悖论问题，是一个常见概率问题，计算很容易。生日悖论问题见后面参考链接，这个问题和生日悖论问题的区别在于多了一个变量：家庭患者5%，但推导很类似。具体推导略去，可以参考本文代码和生日悖论解释自行推导。

(g)重复一下最新的带入具体数字的模型问题（带入数字只是为了解释方便，计算时都可以是变量）：总确诊人数10000，病故率A=1%，家庭患者率5%，问我们需要在多少个病故患者中（定义为数量X），才能找到一个家庭2人同时病故的情况。注意你如果要100%找到一个家庭2人同时病故的情况，这是个问题是无解的。因为根据鸽笼原理，只有在501个病故者中（10000\*5%=500，500+1=501），才能以100%的概率找到找到一个家庭2人同时病故的情况，而实际病故者只有100（10000\*1%=100）。但是如果我们以99%的概率找到一个家庭2人同时病故，需要多少个人病故者呢（也就是X需要等于多少）？直觉是X可能很接近501（毕竟99%很接近100%），但生日悖论告诉我们，这个直觉是错的，并且误差大的离谱！计算表明（见程序），X=91时，就能以99.03%的概率找到一个家庭2人同时病故的情况。以下是一个直观的表格，反映生日悖论反直觉的地方。生日悖论问题计算用到阶乘，是一个强非线性问题，这就是其反直觉的根源所在。

|  |  |
| --- | --- |
| 病故者人数达到的数量 | 找到一个家庭2人同时病故的概率 |
| X=20 | 62.30% |
| X=30 | 77.46% |
| X=50 | 91.95% |
| X=90 | 98.98% |
| X=91 | 99.03% |
| X=100 | 99.39% |

(h)现在，我们得到了第一个结论：总确诊人数N，病故率A=1%，家庭患者率5%，那么在91个病故者中，我们就能找到一个家庭2人同时病故的情况，得到此结论有99.03%的信心（概率）。注意这个结论和总确诊人数N有关吗？直觉告诉我们，达到99%的X应该和N无关。但@profChen的直觉也是有道理的，我们用数值模拟，并没有证明X和N无关。我们需要确认N和X的关系，以下是模拟结果，支持X=91和N无关的结论。

|  |  |  |
| --- | --- | --- |
| 总确诊人数N | 病故者人数达到的数量X | 找到一个家庭2人同时病故的概率 |
| N=10,000 | X=91 | 99.03% |
| N=100,000 | X=91 | 99.01% |
| N=1,000,000 | X=91 | 99.01% |
| N=10,000,000 | X=91 | 99.01% |
| N=100,000,000 | X=91 | 99.01% |

(i)有了第一个结论，我们可以直接通过上述新闻的400，估计总病故者数了。每91个病故者找到一个家庭2人同时病故，则400对应400\*91=36,400个病故者，按照1%的病故率A，总患病人数3,640,000。这个结论看起来和最初的模型很接近，但需要强调的是，3,640,000现在变成了上限而不是下限。为什么是上限，这里有两个原因：一是5%的家庭患者率是很低的一个估计，如果换成10%，则X=45，总患病人数降为1,800,000。下表为家庭患者率和总患病人数的关系：

|  |  |  |
| --- | --- | --- |
| 家庭患者率 | 病故者人数达到的数量X | 总病患数 |
| 5% | X=91 | 3,640,000 |
| 10% | X=45 | 1,800,000 |
| 20% | X=22 | 880,000 |
| 30% | X=14 | 560,000 |

(j)为什么是上限，第二个原因是病故者X=91时，可以找到一个家庭2人病故者，几乎是一个上限。99%是一个很高的信心，一般的估计并不需要选择99%这么高。另外肯定找到一个家庭2人病故者是指至少找到1个家庭，也可能包括2个3个4个更多的家庭的情况。所以X=91，这是一个上限。如果我们选择95%的信心，5%的家庭患者率，则X=60，总患者数为：2,400,000。

(k)用一个更接近大家共识的真实情况的数据带入计算来结束这段讨论：病故率A=1%，家庭患者率20%，95%的信心，则X=15，400个病故家庭对应总患者数：600,000=60万。如果400其中只有200个是这个模型对应的，则总患者数30万。

(l)最后还有一个问题，病故率B在哪里？其实一旦我们转换成生日悖论问题，就不再需要病故率B这个参数了，因为我们可以直接用更准确的方式估计总患者数。不过既然开头提到这个参数，那也可以顺便算一下。

(m)回到最初的定义：病故率B=病故人数B/(确诊家庭\*2)，2是指家庭含2人。我们回到最初那个数据：总确诊人数10000，病故率A=1%，家庭患者率5%，99%的信心，则确诊家庭人数=500（家庭数250），家庭2人病故人数B的下限是：100/91\*2=2.2，则病故率B=2.2/500=0.44%。注意这个是下限，也就是病故率B不会低于0.44%，离1%\*1%=1/10000差了44倍。**@DonaldQ猜的0.5%，很准啊！**

参考链接：

生日悖论

https://zh.wikipedia.org/wiki/%E7%94%9F%E6%97%A5%E5%95%8F%E9%A1%8C

附件新闻：

