24 二阶风险：以已知推断未知


30天认知训练营 · 2020
今天
进入课程 ›

24 二阶风险：以已知推断未...
12:18　11.27 MB

王烁亲述

你好，我是王烁。

上一讲，我们讲了一阶风险，也就是可能的损失。这一讲，我们讲二阶风险。它与一阶风险的差别是这样的：一阶风险是已知风险的预期值，二阶风险则是对预期的偏离，所谓预期差。用普通话来讲，一阶风险是意料，二阶风险是意外。

人是预期的动物，从过去搜集数据，在现在思考，向着未来行动。但凡风险在意料之中，就能采取行动，消化风险。但意料之外怎么办？

上一讲说到，美国平均每年有十万分之六的人死于摔倒，但每年死于摔倒的比例肯定都不一样，有的年份会高，有的年份会低。保险公司如果只按十万分之六来给意外险定价的话，遇到事故发生特别多的年份，就亏死了。长期中这个比例会回到均值，但要是短期中保险公司绷得太紧没留出余地，在短期中就挺不住倒掉了，那长期对它没有用处。对冲了意料之中，还要给意料之外留出余地。

正态分布

怎样衡量这个余地？

偏离预期的程度小风险就小，偏离预期的程度大风险就大。统计学提供了一把能够测量偏离程度的尺子，帮你度量风险，它叫标准差。

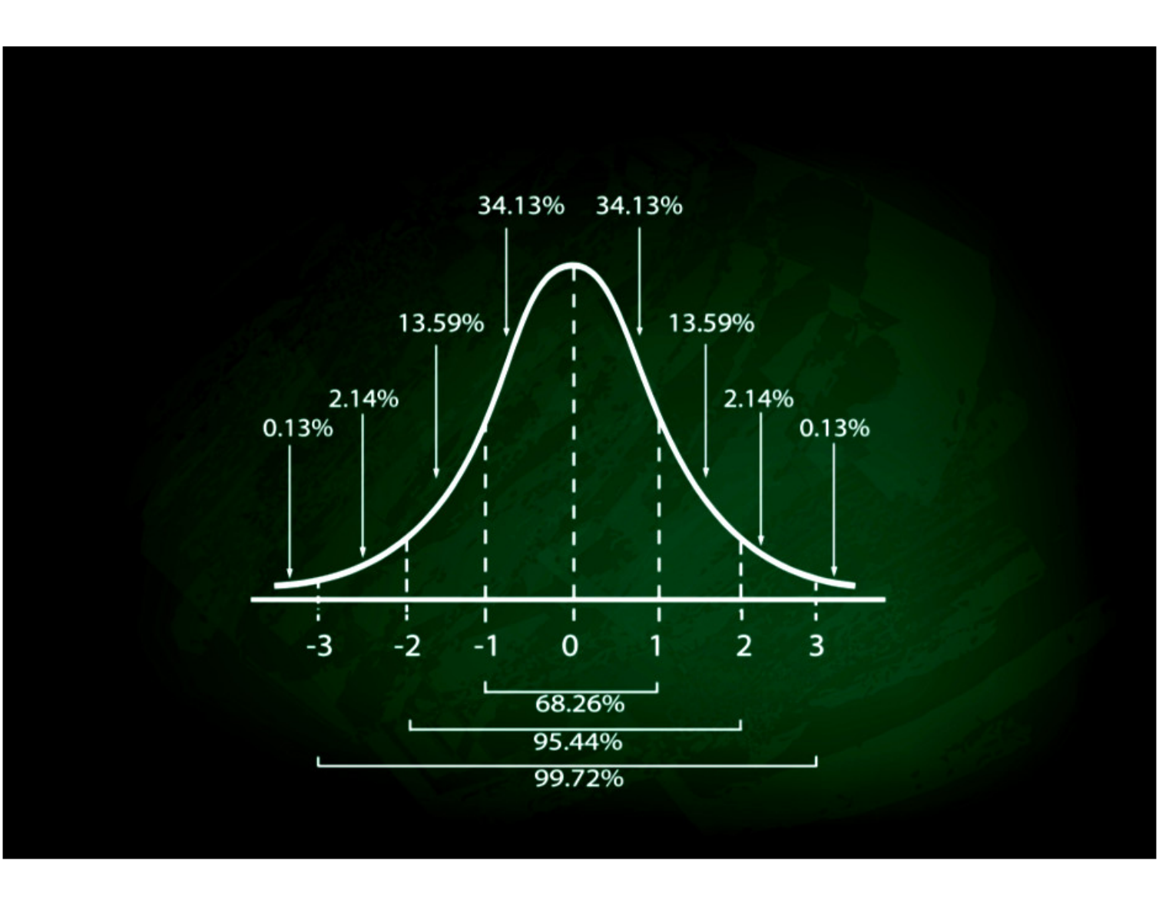
标准差度量数据偏离平均值的程度。

以上面讲的摔倒为例，它是这样得出来的：首先把每年美国人实际死于摔倒的比例，依次减去平均值；值可能为正可能为负，所以取平方再相加，再除以年数，就得到了方差，再开平方根得到标准差。标准差越大，这组数据偏离其均值的程度就越高，风险就越大，反之则越小。

如果你知道一组数据的均值和标准差，那么，哪怕不知道每个数值是什么，你已经驯服了其中蕴含的大部分风险，如果这组数据是正态分布的话，简直就是全部风险。

所谓正态分布，呈现为钟形曲线，左右对称，最高点是这组数据的平均值，向左右两侧放下伸展出尾部，好像一口钟。

在正态分布里，数据非常集中，绝大多数数据集中在平均值周围，极少部分分布在两侧尾部。具体来说，在距均值一个标准差的距离内，有68.3%的数据，两个标准差之内有95.4%，三个标准差之内就能涵盖99.7%。得到钟形曲线，意味着二阶风险尽在掌握。



正态分布图示意

面对钟形曲线，你能精确地表达自己想要拥有多少风险。来一个标准差的？还是两个三个的？

你读论文看科普，最常见一个数字，5%，这不是偶然，它对应着约两个标准差。两个标准差以外的尾部是小概率事件，被认为不大会发生，管住95%风险足够了。

如果是智商，则两个标准差之内对应着智商70到130的区间，其两侧分别是弱智和天才的门槛。普通人的智商则在一个标准差之内，85到115之间，2/3的人都在这里。

钟形曲线刻画风险，就是这般了如指掌。

中间层工具

当然，正态分布、钟形曲线，并不是这个世界的常态。正态分布要求随机可重复事件且彼此相互独立。世界上太多事情或者不随机，或者不可重复，或者并不相互独立。

比如收入就不是正态分布，它既不随机也不独立，因此收入分布的形状不是钟形曲线，更像是金字塔结构，所谓二八法则，20%的人拥有80%的财富，层层往上。

之所以说世界是未知的，从统计思维的角度看，就是因为把世界当作总体，它的数据怎样分布对我们是未知的。黑夜里走路，手电筒照亮眼前一小段路，但灯光之外的黑暗里有多少坎多少沟多少洞，看不见。

所幸统计学有大数定律和中心极限定理这些工具，在看得见与看不见之间，给我们搭了一座天梯。不管总体是怎样分布的，只要抽样过程满足随机、独立性要求，样本量足够，那就有两个关键结果：

- 第一，根据中心极限定理，样本的均值是正态分布。
- 第二，根据大数定律，样本均值的均值约等于总体均值。

也就是说，统计学能构造出一个我们恰好会处理的正态分布，并根据它的性质来对总体作出推断。

总体均值，样本均值，样本均值的均值，三个层次的概念很绕，我来举个例子。

比如说你想知道北京市所有人的平均收入。前面说了，收入不是正态分布，另外北京有2000万常住人口，你也不可能去穷尽每个人的收入信息。那你怎么才能知道平均收入？

如果你多次随机抽样，每次相互独立，样本容量足够大——比如每次抽400人，总共抽样1000次——那么：

- 第一，我们想知道的是北京市人口的平均收入，这里叫作总体均值。
- 第二，通过1000次抽样，产生1000个样本，每个样本中的400人有个平均值，叫作样本均值，1000个样本均值之间还有个平均值，叫作样本均值的均值。
- 第三，样本均值的均值约等于总体均值，所以，只要获得这些样本均值，再取其平均值，我们就得到了北京市人口的平均收入。

抽样1000次只是我为了方便解释，现实中没有人这么麻烦，抽样调查毕竟是有成本的。现实中往往只抽一次，一次就够用。

为什么？

复习一下：中心极限定理告诉我们，样本均值是正态分布。大数定律告诉我们，样本均值的均值，约等于总体平均值。

现在，只抽一次样，只产生一个样本均值数据，要用它去推断北京市总人口平均收入，等于要回答这个问题：在那个样本均值的正态分布里，如果只知道其中一个样本的均值数据，那么，它与样本均值的均值，也就是总体均值，也就是北京市平均收入的距离是多少？

回答这个问题，只需要再知道一样东西：样本均值的标准差。

我们并不知道样本均值的标准差，惟一知道的是这一次抽样400个人收入的样本标准差。统计学的处理方法是，用样本标准差除以标本容量的平方根，来推算样本均值的标准差。这法子不完美，但没有更好的办法。统计思维有什么就用什么，并不因为不完美就停在那里不往前走。

好了，现在推断北京市平均收入的条件都已具备：

- 第一，你只要判断样本均值的均值就可以了，因为它等于总体均值。
- 第二，根据样本标准差，推算样本均值的标准差。抽样得到的单个样本均值，距样本均值的均值的距离，等于样本均值的均值距样本均值的距离——你跟我的距离，等于我跟你的距离。
- 第三，用样本均值的标准差，来衡量样本均值与样本均值的均值的距离。

在这个例子里，你抽样已经得到了样本均值，样本均值的标准差已推算，至于距离样本均值的均值，也就是北京市平

写留言，与作者互动　12　44