距离样本均值的均值，也就是北京市平均收入几个标准差，就看你的要求了。如果你要求估计得宽一点，标准差个数就多一点；估计得窄一点，标准差个数就少一点。

假如抽样400人得到的均值是年收入8万元，标准差是2万元，那么，推算样本均值的标准差是2万/400开平方根等于1000。那么，我就有95%的把握，北京人的平均收入在7.8万到8.2万之间，它对应着两个标准差；有99.7%的把握，北京人的平均收入在7.7万到8.3万之间，它对应着三个标准差。估计得越宽，我的把握就越大，反之亦然。

上面这个操作，统计学叫作置信区间。你日常看见的许多统计数据都是区间估计，人均收入是，美国总统大选预测也是。它表面上是一个数据，实际上总是伴随着一个区间，还有一个与区间对应的置信水平，就是有把握程度。

到这里，入门统计学的核心就都有了。在已知与未知之间，统计学就这样给我们搭了一座天梯，剩下的就是把正态分布、钟形曲线掰开揉碎的各种应用：

——如果已知均值和标准差，你就知道任何一个数据在分布中的位置。这操作叫点估计。

——如果已知样本均值和标准差，去估计未知的总体均值，这操作叫置信区间。

——如果已知总体均值，想检验一个新假设，于是用抽样获得样本均值，再考察样本均值出现的概率是否落在给定显著性水平之内，由此决定是否接受这个假设，这操作叫作假设检验。

如果已知你面对的是个正态分布，事情很简单；如果你不知道面对的是什么分布，那么，在未知的总体与你已知的抽样之间，统计学用大数定律和中心极限定理戏法，构造出一个正态分布的中间层，再根据你预定的精度要求，用抽样数据透过中间层去推断未知。

点估计、置信区间、假设检验，这些令普通人头昏脑涨的操作，本质都是同一套操作，差别只在于总体均值、样本均值、精度要求当中，你给定哪一个，求解哪一个。

本讲小结

我最后再讲几个提醒。

第一，统计学的核心不是梳理已知，绝大多数时候不是列举全部数据，那是大数据的活。统计学是小数据，核心是用已知推断未知。

第二，统计学依靠数学但不是数学，它尽可能严格但没法严格到底，在没理想工具的时候，它是有什么用什么，在我们预定的精度水平上，去推断未知。看跟统计有关的任何结论，一定不要只看单一结论，还要看区间多大，预定的精度是多少。

第三，未知地带总是有意外等着我们。以股价为例，现代金融学早期将股价波动近似地当作正态分布来处理，但正态分布过低估计了小概率重大事件出现的频率。

1987年华尔街黑色星期一，道指下跌27%，按正态分布算出来的概率小到整个宇宙诞生以来的时间都不够这事发生。股票指数的涨跌不是正态分布现在已成常识，那它到底是什么分布？用什么工具去处理它？金融界今天还在打补丁。

所有模型都是错的，好在有些还有用，能用将就用，不能用的将就修修接着用。

思考题

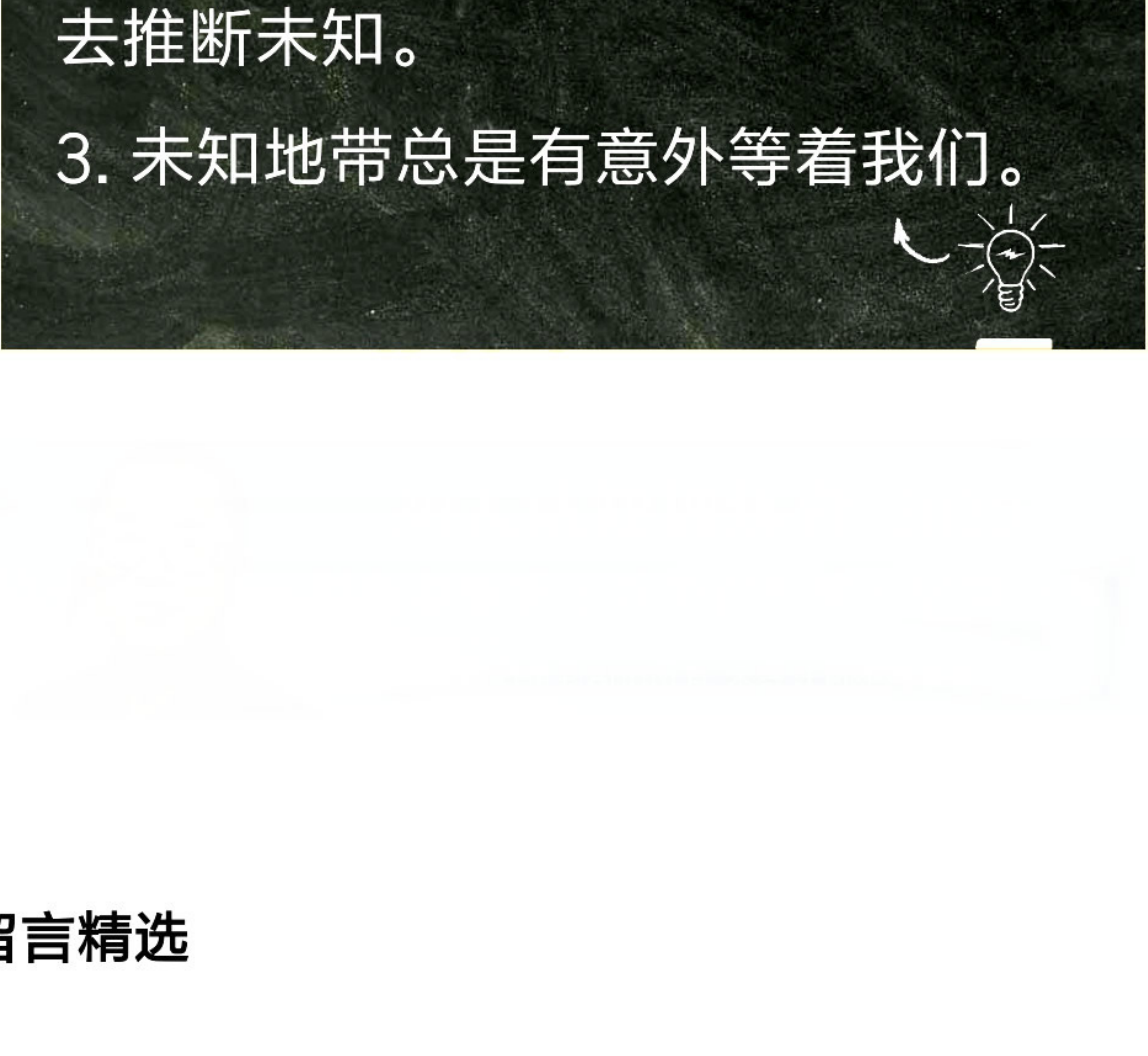
通用电气公司（GE）曾经实行过著名的6个西格玛管理，西格玛就是标准差，6个西格玛指正确率要达到6个标准差的水平。请问具体是个什么水平？我提示一下，是个有点变态的水平。

这两讲我们讲了可以计算的风险怎么逼近，下一讲，我们讲没法计算的风险怎么面对。

我是王烁，我们下一讲见。

划重点

添加到笔记



1. 统计学是小数据，核心是用已知推断未知。先已知全部数据再去分析是大数据的活。
2. 看统计结论一定要看数据的精度，因为它不完全严谨，它是用工具去推断未知。
3. 未知地带总是有意外等着我们。

留言精选

默认 最新 只看作者回复

正在加载...

写留言，与作者互动 12 44