

统计学历史人物漫谈（转）

一、戈赛特（Gosset）——是 t 检验的创始人。

戈赛特据说同时拥有数学和化学两个学位（这在当时似乎并不是很少见，不少学者都同时拥有两个甚至以上的学位）。与许多学者一样，他当时并没有直接从事统计学的研究，毕竟，在 100 多年前，统计学甚至还算不上一门学科。他当时从事的是啤酒酿造行业，然后就在这一似乎与统计无关的行业里，他做了一项研究，想弄清楚发酵时需要加多少酵母最合适。当时戈赛特做出了结果并准备将其发表，可惜他所在的是酿酒行业，贸然发表的话会引起泄露机密之嫌。但戈赛特又确实想发表这一文章，因此采取了折中的办法：匿名发表。他采用了一个笔名，也就是现在我们仍可以在统计学课本上见到的“student”。

戈赛特最重要的一个贡献就是提出了小样本的检验思想。现在我们看起来似乎并无任何出奇。但在当时，统计学几乎就是大样本的科学，一提起统计学，就想到大样本。当时的 K 皮尔逊的几乎所有工作都是基于大样本的假设。但戈赛特根据自己的经验认为，有的情况下，大样本对于研究者来讲太过于奢侈了，必须专注于小样本。但是一旦用小样本分析，无可避免地会牵扯到误差的问题。在大样本情况下，你可以假定没有误差或者误差很小可以忽略不计，而小样本必须考虑到这一问题。那么小样本情况下，误差有多大呢？这就是戈赛特所关注的。

戈赛特通过自己的不断演算，最终于 1908 年发表了一篇极为重要的文章《the probable error of the mean》，提出了 t 分布，这也是至今我们仍在广泛应用的 t 检验（也叫 student t 检验）的基础。考虑一下当时的条件，可想而知戈赛特做出了多少次的计算才得出这一结论。他需要一次又一次地计算均数、标准误，以确定相关数据的概率分布，现在条件下通过计算机模拟可能很快出来结果，但当时显然是很复杂的。但不管如何，通过戈赛特的努力，最终发现了小样本的分析规律，并奠定了小样本分析的基础。现在的人通常称其为小样本理论的鼻祖。

当然戈赛特的贡献并不止于此，以后的几十年中他仍然发表了不少文章，而且他是一个非常谦逊的人，当时 pearson 和 fisher 之间的矛盾很深，而戈赛特作为二者之间的调和人，协助他们在各自领域里做出了不少贡献，而且作了很多穿针引线的事情，保持几位统计学大家的关系。这在以后也是很为人所称道的。

二、高尔顿（Galton）

高尔顿（Galton）这个名字也许有的统计书上有介绍，主要是关于“回归”这一名词。高尔顿通过测量父亲和儿子的身高，发现了“向均数回归”这一现象，并提出了回归的思想。

高尔顿早年也是学医学的，实际上，统计学只是高尔顿兴趣中很小的一部分，他在气象学、心理学、社会学、教育学、优生学、指纹学等都有很深的研究。高尔顿首次发现每个人的指纹是唯一的，迄今仍被称为 Galton Marks。后来高尔顿在他的表兄达尔文（就是我们熟知的达尔文）的影响下，开始对遗传学感兴趣，他在伦敦成立了一个生物统计实验室，收集了很多人的身高、体重等数据。如果从遗传角度，父亲身高较高的话，可能儿子身高也高，如果一代一代传下去，会不会出现两极分化的现象呢？高尔顿对收集的数据进行了整理，结果发现并非如此。如果父亲很高，往往儿子身高较低；如果父亲身高很低，儿子可能身高更高。由此，高尔顿提出了“向均数回归”这一说法。这也是近代回归分析的起源。尽管现在统计书中的“回归”已经远远超出了当初高尔顿所提出的“回归”的概念，但高尔顿对回归的建立确实功不可没。

高尔顿在研究父子身高关系时，提出了用“相关系数”进行度量二者的关系，并给出了明确的公式以及计算方法。这样，高尔顿首次提出了“相关”这一字眼。但高尔顿毕竟没有真正把精力放在统计学上。1889 年他出版了“自然遗传”一书，总结了他在这方面的发现，后来就脱离了这一领域，转而研究别的东西了。但他的研究并没有因为他的改行而中止，幸好有他的学生 K.皮尔逊（Karl Pearson）继续研究他的工作，而且完善了他的理论。所以现在的书中也称相关系数为 Pearson 相关，但要记住，它是高尔顿首先提出的，皮尔逊在一定程度上更规范地阐明了这一概念并进行了推广。

1901 年，高尔顿、皮尔逊、韦尔登共同创办《生物统计》杂志，这也是高尔顿的一个很大的贡献，高尔顿为该杂志提供了丰厚的资金支持，使得该杂志称为世界上第一本印有全彩图片的期刊，至今该杂志仍然是统计学领域的卓越期刊。其中戈赛特的 t 检验的提出，就是发表在这一杂志上。

三、K 皮尔逊（Karl Pearson）

pearson 的名字多数人都不是很陌生，当前的统计学教材中好几章可能都有他的名字，比如 Pearson 卡方检验，Pearson 相关系数。

K 皮尔逊应该算是早期统计学的创始人之一，在他年轻的时候，受到了高尔顿的影响，对统计学开始感兴趣。当高尔顿离开统计学领域，转而研究其它问题的时候，K 皮尔逊接替了他的工作，接管了高尔顿的生物统计实验室，而且后来成了《生物统计》期刊的唯一编辑（共同合伙人高尔顿和威尔登已死亡）。

当时的研究中，充满了不确定性，人们从多次试验中得到许多散乱的数值。而 K 皮尔逊认为，所有的这些实验数据，都是一个数据分布中的样本。而且 K 皮尔逊发现了“偏态分布”（这之前 Laplace 发明了正态分布），K 皮尔逊声称，只要有均数、标准差、峰度、对称度，就可以去完整描述一个分布。当然这些东西在以后被证实存在很多问题。

K 皮尔逊最大的成就之一就是创造出了拟合优度检验。现在我们依然在用，也就是卡方拟合优度检验。这一检验即使在现在依然能看出它的重要性，它可以让你模拟现实中不同的数学模型，然后利用拟合优度检验来确定哪一个更好。

K 皮尔逊还完善了高尔顿提出的相关系数，并证明了回归中的复相关系数，可以认为，K 皮尔逊是复相关系数的提出人。

可惜，K 皮尔逊晚年控制欲太强，以至于与另一伟大的统计学家 Fisher 有了很大的分歧，当年 Fisher 投稿《生物统计》，结果被 K 皮尔逊百般刁难，最终导致 Fisher 不再投稿这一杂志，而改投其它杂志，而且几乎以后所有文章也不再发表在《生物统计》上。二人之间的分歧一直是统计学的一大遗憾。当后来 K 皮尔逊被 Fisher 一举击败后，晚年的 K 皮尔逊，几乎无人关注。只是孤零零地在办公室里，似乎统计学界已经没有了他的存在。反而是他的儿子 E 皮尔逊似乎更加活跃于统计学界。