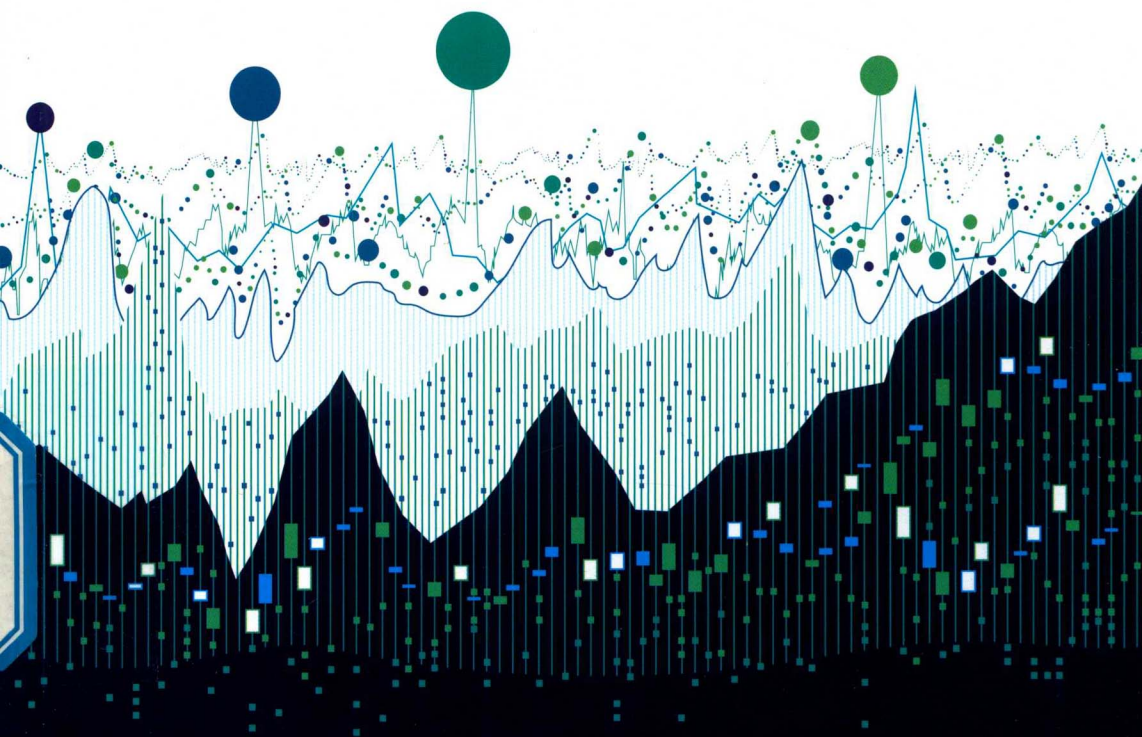


白话

冯国双 著

统计

一本能让人看明白的“白话”统计书
一本真正教你如何分析数据的统计书
一本提供数据分析思路而非公式的统计书



白话统计

冯国双 著



电子工业出版社

Publishing House of Electronics Industry

北京•BEIJING

内 容 简 介

一本书如果没有作者自己的观点,而只是知识的堆叠,那么这类书是没有太大价值的。尤其在当前网络发达的时代,几乎任何概念和知识点都可以从网络上查到。但是有一点你很难查到,那就是统计分析的思路 and 观点。比如,你可以很容易地在网上查到什么是线性回归,但你却查不到怎么“做”线性回归分析,在你遇到实际数据时仍然不知道如何分析。在本书中,你可以获得这些思路 and 观点。尽管这些观点未必是所有人都认可的,但根据笔者多年的分析经验,它们在实践中通常是奏效的。

本书凝结了作者十多年来对统计分析的理解,对各种方法的介绍采用全新的理念和思路,不再是介绍方法本身,而是试图将各种方法之间的联系阐述清楚;不再是介绍方法如何计算出结果,而是尽量说明方法背后的思想。当然,本书同时提供了如何实现结果的软件(涵盖 SAS、R、JMP、SPSS 等)操作。

未经许可,不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有,侵权必究。

图书在版编目(CIP)数据

白话统计 / 冯国双著. —北京:电子工业出版社, 2018.3

ISBN 978-7-121-33518-1

I. ①白… II. ①冯… III. ①统计学—通俗读物 IV. ①C8-49

中国版本图书馆 CIP 数据核字(2018)第 012651 号

策划编辑:张月萍

责任编辑:牛 勇

特约编辑:赵树刚

印 刷:天津千鹤文化传播有限公司

装 订:天津千鹤文化传播有限公司

出版发行:电子工业出版社

北京市海淀区万寿路 173 信箱 邮编:100036

开 本:720×1000 1/16 印张:18.5 字数:444 千字

版 次:2018 年 3 月第 1 版

印 次:2018 年 4 月第 3 次印刷

定 价:69.00 元

凡所购买电子工业出版社图书有缺损问题,请向购买书店调换。若书店售缺,请与本社发行部联系,联系及邮购电话:(010) 88254888, 88258888。

质量投诉请发邮件至 zlts@phei.com.cn, 盗版侵权举报请发邮件到 dbqq@phei.com.cn。

本书咨询联系方式:010-51260888-819, faq@phei.com.cn。

推荐序

宋代禅宗大师青原行思曾提出参禅的三重境界：参禅之初，看山是山，看水是水；禅有悟时，看山不是山，看水不是水；禅中彻悟，看山仍是山，看水仍是水。统计学的学习过程何尝不是这样：初学统计时，看到一种方法，觉得就是这种方法，如学到 t 检验，觉得 t 检验就是用于两组均值比较的方法；经过一段时间的学习，突然发现原来看过的方法并不是所想的那样，如在线性回归中也看到了 t 检验的身影，这时觉得 t 检验不再是自己想象中的 t 检验了；再经过一段时间的学习，就会明白方法还是原来的方法，所谓的疑惑只是自己心动而已，这时就真正理解了 t 检验的含义，而不再把它当作一种固定于特定场合的检验方法。

对于统计学初学者而言，通过课堂老师的讲授，不难达到第一种境界；然后通过接触各种资料，慢慢就会发现各种方法跟课堂上所讲可能有所不同，剪不断，理还乱，逐渐进入第二种境界；此时，如果不能厘清思路，则很容易陷入迷惑和混乱，想达到第三种境界就难上加难了。当然，你可以充分利用网络资源查找各种信息，从而为自己解惑。然而这种资源虽然多，但较为零散，而且可能会看到众说纷纭的情况，使得你更加凌乱。幸运的是，冯国双博士的这本《白话统计》以通俗、幽默的语言深入浅出地介绍了统计学中的各种概念和思想，为很多人消除了留在心中多年的存疑，可以在这一阶段帮助读者答疑解惑。其中有些内容曾经发布在微信公众号“小白学统计”中，也因此受到了很多网友的好评。

目前，欧美有不少类似白话介绍统计方法之类的书籍，这类书的特点是：不像教材那样用沉重、拘谨的语言来介绍统计学方法，而是以风趣、浅显的话语来解释说明；内容不一定多，但尽量深入。相比之下，国内的这类书籍并不多见，在生物统计学领域尤其缺乏。这就导致很多非统计学专业人士觉得统计学“面目可憎”，始终无法真正喜欢上统计学。可喜的是，目前逐渐成长起来的年轻的统计学家已经意识到这一问题，冯博士的这本《白话统计》正是让统计学尽量接地气的一种有益尝试。

《白话统计》凝结了作者十多年来对统计学的理解，采用别具一格的编写理念和思路，对各种方法的介绍不再是介绍方法本身，而是试图将各种方法之间的联系阐述清

楚；不再是介绍方法如何计算出结果，而是尽量说明方法背后的思想。本书对基本概念和方法的介绍没有采用教材中非常严谨的语言，而是尽可能用白话来解释说明。例如，在介绍累积分布和概率密度时，用到了《神雕侠侣》中的情节；在介绍假设检验思想时，用到了“女士品茶”的故事；在介绍中心极限定理时，采用了里约奥运会期间的一则虚拟故事。这不仅使得抽象理论形象化、复杂问题简单化，也增加了本书的趣味性和可读性，体现了作者深厚的理论功底和丰富的实践经验。

在此，我把《白话统计》这本著作推荐给各位读者，无论是初学者、统计学专业人士，还是统计学应用工作者，相信都能从中获益。

陈 峰

南京医科大学教授

前言

在一本书中如果没有自己的观点，而只是各种已知知识的堆叠，那这本书还有什么意义呢？

——作者

时光飞逝，从我的第一本书出版到现在，转眼已经过去了 7 年。期间，我的统计分析经验在不断积累，对统计学方法也有了更深一步的认识。但有一点始终未变，那就是我对统计学的热爱以及推广统计学的理念。从我的第一本书开始，我的理念就是，尽量写一本让非统计学专业人员也能看懂的统计书。直到本书的面世，尽管每本书的风格不一，但我始终遵循这一理念。

推广和普及统计学并不是一件容易的事情。由于各种原因，不少人在初学统计学时有点抵触，所以我尽量通过各种方式努力使统计学“平易近人”，写书便是其中之一。多年前我曾在百度空间创建了自己的“卫生统计空间”，阅读量达到近百万次，后来由于百度空间关闭而结束。后来我又在微信公众平台上创建了“小白学统计”公众号。在本书出版之际，该公众号恰好满两岁，粉丝也刚好满 2 万人。不少人在公众号上留言，建议我将内容整理成册并出版。因此，在后期我就一边写新的文章，一边将其整理修订。

虽然本书脱胎于“小白学统计”公众号，但内容其实和公众号里的文章并不相同。事实上，我几乎重写了所有内容，因为正式出版的书籍不能像公众号里的文章那样随意。所以，尽管你会看到书中的内容有与公众号相似之处，但本书更为系统和详细。

关于本书的名字，我曾在公众号上征集各位读者的意见，很多人建议直接用《小白学统计》这一书名。但经过仔细考虑，以及几位试读朋友的建议，最终还是命名为《白话统计》。因为本书并没有像教材一样系统地介绍各种统计方法，而更像随笔或补充读物。如果用武侠小说来类比，则更像梁羽生笔下的《玄功要诀》或金庸小说中的《易筋经》，虽然并不侧重教你具体招式，但却可以让你在学其他招数的时候事半功倍。

这是一本怎样的书

第一，你可以把本书看作一本“翻译”书。我在尽力把统计学中的公式“翻译”成白话文。比如，你在很多统计学书中看到正态分布的公式，而在本书中，你看到的是正态分布公式的解释及其实际含义；再如，你在统计学书中看到最大似然估计的公式时可能会有点困惑，本书则通过例子通俗地解释最大似然估计的计算思想。

第二，你还可以把本书看作统计学教材的补充内容。本书并不是简单地介绍各种统计方法，而是尽量把各种方法串联起来，从思路上理解方法本身。例如，教材中都会介绍如何用 t 检验进行两组比较，本书则重在阐述 t 检验的思想本身，这样你在任何场合（如线性回归、相关分析）看到 t 检验都不会觉得陌生；而且本书还专门用了一章的篇幅对常见的各种统计学方法进行了串讲，从一般线性模型到广义可加模型，尽量让你明白它们之间的关系。

第三，如果你愿意，则也可以把本书看作一本打发时间的消遣读物。本书对所有概念和方法都尽量以通俗的语言而非官方语言来阐释。例如，在介绍分布时，通过大家熟知的郭靖如何消耗内力来进行说明，从而避免了不少人对“分布”这一概念完全摸不着头脑的尴尬；在介绍假设检验思想时，用了“女士品茶”这一浅显的例子来说明其思路。

从这本书中你能学到的和不能学到的

本书分为两大部分：基础篇和实用篇。其中，基础篇介绍了统计学中常见的概念及初学者容易存在的疑惑。例如，很多人都头疼的分布（第3章）、初学者不易理解的假设检验（第7章）、比较重要的中心极限定理（第6章）、参数估计和置信区间（第8章和第9章）等。

实用篇则侧重介绍各种方法的思路及实现，先对各种常见方法进行了串讲（第10章），然后分别介绍了 t 检验（第12章）、方差分析（第13章）、卡方检验（第14章）、相关分析（第15章）、回归分析（第16章）。但是千万不要被我列举的表面现象所迷惑，这些方法可能你觉得都“会”，但如果你打开本书，则会发现原来这些内容并不是这么简单的。

一本书如果没有作者自己的观点，而只是知识的堆叠，那么这类书是没有太大价值的。尤其在当前网络发达的时代，几乎任何概念和知识点都可以从网络上查到。但是有一点你很难查到，那就是统计分析的思路和观点。比如，你可以很容易地在网上查到什么是线性回归，但你却查不到怎么“做”线性回归分析，在你遇到实际数据时仍然不知道如何分析。在本书中，你可以获得这些思路和观点。尽管这些观点未必是所有人都认可的，但根据我多年的分析经验，它们在实践中通常是奏效的。

如果你想从本书中学习如何一步步地进行软件操作，那你可能会失望。本书没有

教你具体的软件操作过程，因为软件实现是最简单的，而统计分析思路则是更为复杂的。学习统计，最遥远的距离是，你面对着电脑中的统计软件（中文版），菜单上的每个中文都认识，却始终不知道该点什么。这很让人受挫，不是吗？然而这不是软件操作的问题，而是统计思路的问题。本书将教你摆脱这一尴尬局面。

什么人适合阅读本书

如果你对统计学是完全的零基础，那么看前半部分内容应该问题不大，但对后半部分的统计方法可能需要花点心思，因为学习后半部分内容还是需要一定的统计学基础的。所以本书主要适合以下读者：

- 刚刚接触统计学，跟着老师听了几堂课，但是脑子里依然是一团糨糊的人。
- 学过统计学，但是对各种方法之间的关联并不清楚，想更进一步融会贯通的人。
- 在课堂上学过统计学，但遇到实际数据却不知如何下手，想了解数据分析思路的人。
- 统计学已经学得不错了，但有点眼高手低，对有些概念并未真正理解的人。
- 会一点数据分析，但是在实际数据面前容易头脑不清晰，想学习数据分析技巧的人。
- 会用简单的统计软件（如 SPSS），但点开菜单却不知道里面的选项是什么意思、不知该如何勾选、也不知如何解读结果的人。
- 喜欢本人作品、“小白学统计”公众号的粉丝。

.....

本书所用的软件

本书中结果的展示主要基于 SAS 9.4 和 JMP Pro 13。书中凡是涉及计算或结果展示的内容，大都给出了软件实现的语句或操作过程。考虑到不同读者有各自的软件使用习惯，书中分别给出了 SAS 9.4、R 3.4.3、SPSS 20 和 JMP Pro 13 的软件实现过程，部分特殊内容也用到了 Medcalc 和 Stata 12.0。

本书中的配套资源下载

本书中所有例子的数据、SAS 程序和 R 程序都可以在知了帮网站（read.zhiliaobang.com）上下载，以节省读者输入数据和程序的时间。

致谢

首先，非常感谢陈峰老师在百忙之中帮忙作序，陈老师是一位儒雅型的生物统计学教授，能够请到陈老师作序，为本书增色不少。其次，感谢成都道然科技有限责任公司在整体策划和插图上的努力，你们看到书中那些生动的漫画都出自他们的手笔。最后还必须感谢“小白学统计”公众号的粉丝，正是你们的鼓励，才让我义无反顾地将其内容整理成册并出版。

本书是作者多年经验的累积，而且查阅了大量国内外文献，但仍不敢说百分之百正确。如果书中有任何观点上的错误，那说明本人水平仍有不足，所有错误均由本人承担责任，还请读者不吝指正，可在“小白学统计”公众号里留言。

冯国双

2017年12月于北京

目 录

第 1 篇 基础篇

第 1 章 为什么要学统计	2
1.1 统计学有什么用	3
1.2 生活世事皆统计	4
1.3 如何学统计	4
第 2 章 变异——统计学存在的基础	6
2.1 随机与变异	6
2.2 特朗普与罗斯福的胜出——抽样调查到底可不可靠	8
2.3 什么是抽样误差	9
第 3 章 郭靖的内力能支撑多久——谈概率分布	11
3.1 累积分布与概率密度的通俗理解	12
3.2 是生存还是死亡？这是一个问题——用 Weibull 分布寻找生存规律	16
3.3 2003 年的那场 SARS——用 Logistic 分布探索疾病流行规律	20
3.4 “普通”的正态分布	23
3.5 几个常用分布—— t 分布、 χ^2 分布、 F 分布	28
第 4 章 关于统计资料类型的思考	35
4.1 计数资料等于分类资料吗	36
4.2 计数资料可否采用连续资料的方法进行分析	37

4.3	分类资料中的无序和有序是如何确定的.....	38
4.4	连续资料什么时候需要转换为分类资料.....	39
4.5	连续资料如何分组——寻找 cut-off 值的多种方法.....	41
4.6	什么是虚拟变量/哑变量	47
第 5 章	如何正确展示你的数据	52
5.1	均数和中位数——你被平均了吗	53
5.2	方差与标准差——变异的度量	54
5.3	自由度——你有多少自由活动的范围	56
5.4	百分位数——利用百分数度量相对位置	57
5.5	如何比较苹果和橘子——利用 Z 值度量相对位置	59
5.6	某百岁老人调查报告说：少运动才能活得久——谈一下比例和率.....	61
5.7	在文章中如何正确展示百分比.....	63
第 6 章	寻找失踪的运动员——中心极限定理.....	64
6.1	中心极限定理针对的是样本统计量而非原始数据.....	65
6.2	样本量大于 30 就可以认为是正态分布了吗.....	67
第 7 章	从“女士品茶”中领会假设检验的思想.....	70
7.1	女士品茶的故事	70
7.2	零假设和备择假设	72
7.3	假设检验中的两类错误.....	73
7.4	P 值的含义.....	76
7.5	为什么 P 值小于 0.05（而不是 0.02）才算有统计学意义.....	78
7.6	为什么零假设要设定两组相等而不是两组不等.....	79
第 8 章	参数估计——一叶落而知秋	81
8.1	点估计	81
8.2	最小二乘估计	82
8.3	最大似然估计	84
8.4	贝叶斯估计	86

第 9 章 置信区间估计——给估计留点余地	88
9.1 置信区间的理论与实际含义	88
9.2 置信区间与 P 值的关系	90
9.3 利用标准误计算置信区间	91
9.4 利用 Bootstrap 法估计置信区间	92

第 2 篇 实用篇

第 10 章 常用统计方法大串讲	98
10.1 一般线性模型——方差分析与线性回归的统一	99
10.2 广义线性模型——线性回归与 Logistic 回归的统一	103
10.3 广义可加模型——脱离“线性”束缚	107
10.4 多水平模型——打破“独立”条件	112
10.5 结构方程模型——从单因单果到多因多果	119
第 11 章 正态性与方差齐性	127
11.1 用统计检验方法判断正态性	127
11.2 用描述的方法判断正态性	130
11.3 方差分析中的方差齐性判断	133
11.4 理解线性回归中的方差齐性	135
第 12 章 t 检验——不仅是两组比较	138
12.1 从另一个角度来理解 t 检验	138
12.2 如何正确应用 t 检验	140
12.3 t 检验用于回归系数的检验	141
12.4 t 检验的替代——Wilcoxon 秩和检验	142
第 13 章 方差分析与变异分解	145
13.1 方差分析中变异分解的思想	145
13.2 为什么回归分析中也有方差分析	147

13.3	铁打的方差分析,流水的实验设计.....	148
13.4	方差分析后为什么要进行两两比较.....	152
13.5	多重比较方法的选择建议.....	154
13.6	所有的多组都需要做两两比较吗——兼谈固定效应和随机效应.....	164
13.7	重复测量方差分析详解.....	166
13.8	方差分析的替代——Kruskal-Wallis 秩和检验	176
13.9	多组秩和检验后的两两比较方法.....	178
第 14 章	卡方检验——有“卡”未必走遍天下.....	181
14.1	卡方检验用于分类资料组间比较的思想.....	181
14.2	卡方用于拟合优度评价——从 Hardy-Weinberg 定律谈起.....	184
14.3	似然比 χ^2 、M-H χ^2 、校正 χ^2 与 Fisher 精确检验	186
14.4	等级资料到底可不可以用卡方检验.....	191
14.5	卡方检验的两两比较.....	193
14.6	Cochran-Armitage 趋势检验.....	194
14.7	分类变量的赋值是如何影响分析结果的.....	196
第 15 章	相关分析与一致性检验.....	200
15.1	从协方差到线性相关系数.....	200
15.2	线性相关系数及其置信区间.....	203
15.3	如何比较两个线性相关系数有无差异.....	206
15.4	分类资料的相关系数.....	207
15.5	基于秩次的相关系数.....	210
15.6	相关分析中的几个陷阱.....	213
15.7	用 ICC 和 CCC 指标判断一致性.....	215
15.8	用 Bland-Altman 图判断一致性.....	218
15.9	Kappa 检验在一致性分析中的应用	219

第 16 章 线性回归及其分析思路	222
16.1 残差——识别回归模型好坏的关键	223
16.2 回归系数的正确理解	226
16.3 回归系数检验 VS 模型检验	227
16.4 均值的置信区间 VS 个体的预测区间	228
16.5 逐步回归筛选变量到底可不可靠——谈变量筛选策略	230
16.6 如何评价模型是好还是坏——交叉验证思路	237
16.7 线性回归的应用条件——你的数据能用线性回归吗	240
16.8 如何处理非正态——Box-Cox 变换	247
16.9 如何处理非线性——Box-Tidwell 变换	248
16.10 方差不齐怎么办——加权最小二乘法	250
16.11 当共线性导致结果异常时怎么办——岭回归、Lasso 回归	254
16.12 发现异常值应该删除吗——谈几种处理异常值的方法	260
16.13 如何处理缺失值——是删除还是填补	268
16.14 一个非教材的非典型案例——线性回归的综合分析	276

第1篇

基础篇

- 第1章 为什么要学统计
- 第2章 变异——统计学存在的基础
- 第3章 郭靖的内力能支撑多久——谈概率分布
- 第4章 关于统计资料类型的思考
- 第5章 如何正确展示你的数据
- 第6章 寻找失踪的运动员——中心极限定理
- 第7章 从“女士品茶”中领会假设检验的思想
- 第8章 参数估计——一叶落而知秋
- 第9章 置信区间估计——给估计留点余地

为什么要学统计

希特维尔：索拉的算法是一个程序，用来选择和洞悉目标。

美国队长：什么目标？

希特维尔：你、开罗的一个主持人、国防部副部长、爱荷华市的一个高中毕业生、布鲁斯班纳、奇异博士。所有能威胁到九头蛇的人，不管是现在还是未来。

美国队长：未来？怎么可能知道？

希特维尔：怎么会不知道？21 世纪是一本数码书，索拉教会九头蛇如何读这本书。你们的银行记录、病历、投票模式、电子邮件、通话信息，还有大学考试成绩。索拉的算法会评估人们的过去，预测他们的未来。

——电影《美国队长 2》

我个人其实并不喜欢对统计学做出定义，就像电影 *Three Idiots* 中所说的，我们没有必要纠结“书”的严谨概念，关键是要明白“书”的用处、能给我们带来什么好处。

那么，统计学能做什么呢？统计学是帮你寻找规律，然后做出决策的方法或技术。任何你收集到的一手数据都是杂乱无章的，而统计学就像魔法镜，透过镜子，你看到的不再是一堆纷繁无序的数据，而是整齐有序的规律（见图 1.1）。

21 世纪是一本数字书，而统计学恰是处理数据的工具。30 年前，如果你不识字，则会觉得寸步难行；如今，如果你不懂一点统计学知识，则同样会觉得无法理解这个数字社会。著名科幻小说作家赫伯特·乔治·威尔斯在 1903 年就预言：“在未来社会，



图 1.1 统计学作用示意图

统计学思维将像阅读能力一样成为社会人必不可少的能力。”我想，当年威尔斯说的未来社会指的就是现在。

所谓的“大数据时代”刚刚来临的时候，有的人声称，统计学已经死了，有这么多的数据，还用统计推断做什么？不过现在“大数据热”终于慢慢有点降温了，很多理智的人已经发现，原来大数据并不像我们想象的那样是一座金矿。大数据的确存在，但如果没有合理的统计学思维和技术，那么你挖出来的闪光的东西也许只是玻璃，而不是金子。

1.1 统计学有什么用

为什么要学习统计学呢？如果我跟你说，因为统计学是一门“%¥#*”的学科，那么你一定厌倦了，我们就说点实际的。

先说大家最感兴趣的。首先，掌握统计学，会让你在职场生涯中游刃有余。在当前的“大数据时代”，最缺乏的就是能够正确利用数据的专业人员。麦肯锡曾在一份报告中提到，预计到2018年，大数据或者数据工作者的岗位需求将激增，其中大数据科学家的缺口为14万~19万人，而懂得利用大数据做出决策的统计分析师和经理的岗位缺口则将达到150万人。

欧美职场曾流行一句话：“只要有SAS（统计分析软件）认证，你就永远不会失业。”最近又有人戏称应用统计学是“全宇宙就业排名第一的专业”，因为工作缺口太大，而且起薪远高于其他专业。对很多人来说，可能这一点理由就足够了。但我们不能只想着眼前的赚钱，还是要有一点“诗和远方”这种理想的，所以我们继续往下看。

其次，学习统计学，可以助力你的科研工作。如果你是科研人员，那么一定十分清楚这一点。目前国内对科研项目的申请几乎都要求有统计学专家的参与。当然，你不一定（也没有必要）达到专业统计的水准，但是如果你不懂一点统计学知识，那么如何跟统计学专家沟通呢？在我的科研经历中，有一点认识是深刻的：很多时候，统计学不仅仅是锦上添花，有时往往是雪中送炭。当你限于思路、限于方法而无法进行的时候，统计学有可能帮你开拓一片新的天地，让你耳目一新。

最后，学习统计学，能够培养一种理性看待事物的能力。统计学是一门让人理智的学科，因为它要求“用数据说话”，而不是感情用事。在有些情况下，你可能会直观地认为某些事情是没有意义的或者不可能的，只有统计学分析才能给出真正的答案。

当前是一个信息社会，各种有用无用的信息每天都在铺天盖地地进入你的视线，如何判断这些信息？当你看到一则新闻说“老父在女儿门前打地铺睡觉”时，你是不是会当即指责女儿不孝？当你看到一篇报道说“宝马女与骑电动车的人争执”时，你会觉得是宝马女不对吗？如果你有这些想法，那你不具备真正的理性思维。但如果你掌握了一定的统计学知识，那么理性思维就会成为你的下意识，当你再看到一些信