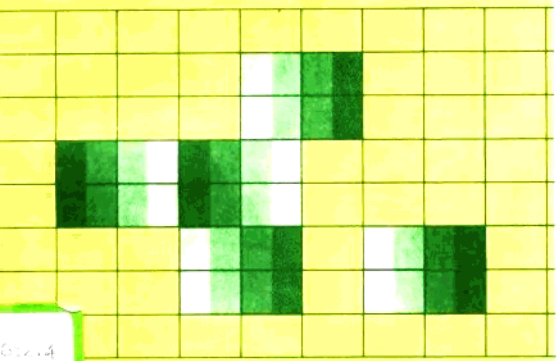
林产化工统计分析

STATISTICAL ANALYSIS ON CHEMISTRY
& INDUSTRY OF FOREST PRODUCTS

陈仁泽 编著



厦门大学出版社

内容简介

本书系数理统计在林产化学与工业的应用专著。本书所阐述的问题及数据都是来自工厂与科研的实际，对林产化学与工业的生产工艺和机理作定量与定性分析。全书共六章，包括一元统计的假设检验、方差分析、试验设计、回归分析，还有多元的回归、聚类、判别分析。本书说理浅显，思路清晰，有利于学习，尤其所述统计方法和实例都编有配套程序（磁盘）。

本书可供林产化工专业人员阅读，也可作为大专院校林化专业的教学参考书。对于从事林产工业，化工专业的人员也有一定参考价值。

前　言

有效地采集、整理和分析工业生产及科研的随机数据，运用数理统计方法对其出现的情况作出推断或预测、分析或建议，是当今为解决工业生产和科研上存在的问题，普遍采用的有效手段之一。

近十几年来，随着电子计算机的广泛应用，科学向定量化发展，林产化学与工业同数学的必然结合，促使数学林产化工这门新学科应运而生，而数理统计则成为数学林产化工的主要内容之一，其运用定性和定量相结合的方法，对林产化学与工业各分支进行研究与分析，不仅使数理统计获得新的生命力，同时也将促其本身的发展。

将数理统计应用于林产化学与工业，是作者多年的探索与研究，由于作者数理统计水平和林产化工的实践有限，书中缺点或疏漏在所难免，敬请广大林产化工工作者和热心读者不吝批评赐教。它山之石，可以攻玉。作者的探索与研究，就算是投下一块粗糙的石头，希冀吸引更多朋友加入这一研讨行列来进行切磋吧！

奉献在读者面前的《林产化工统计分析》一书，主要内容包括林产化学与工业的生产工艺和机理的定量与定性分析。全书所阐述的问题及数据，都是来自工厂与科研的实际，故此，作者力图从林产化学与工业的实际问题出发，讲清统计思路与统计分析方法；叙述各种统计方法时，力求说

理浅显，通俗易懂，有利于学习，读者可以从书中实例掌握不同统计方法，并应用到自己的实践中。

本书兼融专著与实际可用性的特点，书中所述统计方法及实例都编有配套程序（磁盘），读者在学习有关方法后均可按相应程序文件名，自行计算与检验，对尽快掌握统计分析方法大有好处。

十分感激南京林业大学博士生导师程芝教授、中国林科院南京林化所徐进研究员，当我初涉林产化工方面的应用时，他俩就给我以极大的鼓励与支持；在编著过程中，喜得洪伟教授、张宗辉、张肇文高级工程师的指导，得到了许多宝贵意见，杨学文工程师参与林产化学与工业专业内容分析的解释，陈瑞典硕士参与修改部份章节的初稿；多年来支持各种试验和提供资料的林化厂、研究所，以及许多热情帮助的同志，特别值得一提是王仰高高级经济师及潘金福同志对编著本书所给予的倾心关注和极大支持，作者谨向他们表示由衷的感谢。

本书得于顺利出版，大大得益于下列单位的热忱支持与赞助，它们是：武平林化厂，永安造纸胶料厂，三明市林产工业公司，龙岩地区林产工业公司，德化县林业委员会，德化胶合板厂，上杭、连城、长汀、永定、龙岩林化厂，作者对他们表示诚挚的谢意！厦门教育学院的领导和同仁，对本书的编著给予极大的关注，在出版过程中，厦门大学出版社对此书予以大力支持和帮助，在此一并表示深切的谢意！

作者

1993 年 3 月于厦门教育学院

目　录

第一章　产品指标差异性的检验

对任何事物作出推断，都必须有一定的假设前提．例如用间歇法生产松香，每锅取一个样品进行化验，看软化点、色泽等指标是否合乎标准规定．若所取的样品化验合格，就说这锅松香合格，在这里，我们实际上作了“取样能代表这锅松香”的假设．通常是通过随机样本来了解总体情形．又比如，已知松脂蒸馏在正常生产的一段时期内软化点的平均数为 μ_0，标准差为 σ_0，当工艺改革为连续蒸馏之后，我们抽取 n 桶作为样本，得其软化点的平均数 $\bar{X}$，怎样推断改革后的工艺与原有工艺的软化点有否显著差异．类似的问题就是本章所要研究的对象．假设是推断的首要步骤，只有先对研究对象总体作出一定的统计假设后，通过对随机样本数据的计算，以一定的可靠性检验其假设成立与否，才能作出合理的统计推断来．本章介绍林产化工常见的几种检验方法．

§1.1　统计假设检验概述

为了正确地理解假设检验的基本思想和步骤，我们先考察下面三个例子．

例 1.1.1 设某林化厂正常生产时松香的软化点服从正态分布，均值 $\mu = 76.6℃$，标准差 $\sigma = 7.4℃$. 今抽取 9 锅化验结果的软化点平均数为 $\bar{X} = 77.8℃$，问当天生产情况是否正常？

现在的问题是要推断这个样本所来自的总体均值 μ 是否等于指定的总体均值 $\mu_0 = 76.6$？更确切地说，就是要检验这个样本是否来自正态总体 $N(76.6,\ 7.4^2)$？

类似这样的问题，我们可对子样所来自的总体作出假设. 然后再对这个假设作出推断. 今后我们把任意一个有关未知分布的假设称为**统计假设**或简称为假设，常记为 H_0. 譬如本例，可作统计假设为

$$H_0:\ \mu = 76.6$$

我们的目的就是要推断 H_0 是否成立？而建立推断假设 H_0 是否成立的方法，称为**假设检验**. 下面通过讨论例 1.1.1 来阐述这种统计检验方法的基本思想和步骤.

在例 1.1.1 中，待检验的假设为 $H_0:\ \mu = 76.6$，在 H_0 成立的条件下，这天生产的软化点分布应该服从 $N(76.6,\ 7.4^2)$，于是样本均值 $\bar{X}$ 的分布应近似地遵从 $N(76.6,\ \frac{7.4^2}{9})$，由抽样分布理论知，统计量

$$Z = \frac{\bar{X} - 76.6}{\frac{7.4}{\sqrt{9}}} \sim N(0,\ 1).$$

如果我们希望有 95% 的把握来推断 H_0 是否成立，可查正态分布表（附表 1）得 $P(|X| \leqslant 1.96) = 0.95$，即

$$P(76.6-1.96\times\frac{7.4}{\sqrt{9}}\leqslant\bar{X}\leqslant 76.6+1.96\times\frac{7.4}{\sqrt{9}})=0.95,$$

因此，若 H_0 是正确的，那么 $\bar{X}$ 的观察值 $\bar{x}$ 落在上述区间外的概率只有 0.05，即在 20 次抽样中，平均大约有一次 $\bar{x}$ 值落到这个区间外．如果在一次抽样中，$\bar{x}$ 值落在此区间外，那么根据“概率很小的事件在一次试验中不可能发生”这一原理，应该否定假设 H_0．在假设检验中，正是依据〃小概率事件的实际不可能性〃原理来作出否定或接受假设的判断的．根据上述原则，由于 $\bar{x}=77.8$，所以 $|Z|$ 的观察值 $|Z|=0.486<1.96$，即在一次抽样中，小概率事件 $\{|Z|>1.96\}$ 没有发生，所以没有理由拒绝假设 H_0：$\mu=76.6$，即应该接受假设 H_0，即认为按此工艺生产是正常的．在此例中，拒绝接受假设 H_0 的区域 $\dfrac{|\bar{X}-76.6|}{7.4/\sqrt{9}}>1.96$ 称为检验的**拒绝域**，而接受假设 H_0 的区域称为**接受域**，其分界点称为**临界点**（即分位点），$\alpha=0.05$ 称为检验的**显著性水平**（图 1.1.1）.

图 1.1.1

例 1.1.2　某林化厂按原工艺生产的松香软化

点 $X\sim N(\mu_0,\ \sigma_0^2)$，$\mu_0=79℃$，$\sigma_0=11℃$，今更换设备后，从某天生产中抽取 30 桶，经计算得样本的软化点平均值$\overline{X}=84℃$. 设方差不变，问$\overline{X}$与 μ_0 的差异是偶然的随机波动造成的，还是由于改变新设备所引起的？前面的差异称为**随机误差**，这是受偶然因素的影响，是客观存在的、不可避免的误差；后者称为**条件误差**，它是由于工艺条件或技术条件的改变而造成的，这种误差是可以控制的. 因此，上面所说的**显著性差异就是指差异的产生不是随机的，而是由系统产生的**，也称**系统误差**. 换一种说法，就是要检验对总体 $X\sim N(\mu_0,\ \sigma_0^2)$ 的假设 $H_0:\mu=\mu_0=79℃$ 是否成立？

按处理例 1.1.1 的方法，先设 H_0：$\mu=\mu_0=79$ 成立，则在此条件下，容量 n 为 30 的样本均值$\overline{X}\sim N(79,\ \frac{11^2}{30})$，

$$Z=\frac{\overline{X}-79}{\frac{11}{\sqrt{30}}}\sim N(0,\ 1).$$

如果给定的显著性水平 $\alpha=0.05$，那么在 H_0 为真时，

$$P(|Z|>Z_{\frac{\alpha}{2}})=\alpha\ ,$$

其中$|Z|=\dfrac{|\overline{X}-\mu_0|}{\frac{\sigma_0}{\sqrt{n}}}$，$Z_{\frac{\alpha}{2}}$ 为双侧分位点. 由此可得到检验的拒绝域为

$$\frac{|\overline{x}-\mu_0|}{\frac{\sigma_0}{\sqrt{n}}}>Z_{\frac{\alpha}{2}}$$

或

$$|\bar{x}-\mu_0|>\frac{\sigma_0}{\sqrt{n}}Z_{\frac{\alpha}{2}}$$

今对 $\alpha=0.05$，查正态分布表得 $Z_{0.05/2}=Z_{0.025}=1.96$，又 $n=30$，$\sigma_0=11$，故

$$\frac{\sigma_0}{\sqrt{n}}Z_{\frac{\alpha}{2}}=3.9363$$

而 $|\bar{x}-\mu_0|=|79-84|=5>3.9363$（也可由 $\mu_0-1.96\frac{\sigma_0}{\sqrt{n}}=75.08$，$\mu_0+1.96\frac{\sigma_0}{\sqrt{n}}=82.92$，接受域为 [75.08，82.92]）. 这表明 $\bar{x}$ 落入拒绝域，也就是说小概率事件 $\{|\bar{x}-\mu_0|>3.9363\}$ 竟然在一次试验中发生了，从而根据小概率事件的推断原理得出原假设 H_0 不成立. 故认为 $\bar{x}$ 与 μ_0 是有显著差异的，这种差异是由于改变工艺而产生的，属于条件误差. 留心的读者会发现，当 $\alpha=0.01$ 时，H_0 的接受域为 [73.84，84.16]，$\bar{X}=84$，正好位于其中，因此接受 H_0，认为改换设备前后所生产的松香软化点无极显著差异. 可见，显著性水平 α 的选定，对于指定 H_0 的拒绝域有直接的关系.

在生产中，有时我们只是关心总体的均值 μ 是（能）否低于或高于某个规定的数值 μ_0. 如松香的特级香率当然希望它越大越好，这时，我们要检验的问题成为：总体中特级品个数的均值（即特级率）μ 是等于 95% 还是大于 95%，即确定：是接受 H_0：$\mu=95\%$，还是接受另一个

假设 $\mu > 95\%$. 这种检验一般的叙述方式为：在显著水平 α 下，检验假设

$$H_0:\mu = \mu_0;\ H_1:\mu > \mu_0$$

由 $P(Z \geqslant Z_{上\alpha}) = \alpha$，得拒绝域为 $(Z_{上\alpha},\ +\infty)$，其中 $z_{上\alpha}$ 为标准正态分布的上侧分位数，相当于双侧分位表上的 $Z_{2\alpha}$，如 $Z_{上0.05} = 1.645$，而双侧分位数表上 $z_{0.10} = 1.645$. 若统计假设为

$$H_0:\mu = \mu_0,\ H_1:\mu < \mu_0$$

则由 $P(Z \leqslant -Z_{上\alpha}) = \alpha$，得拒绝域 $(-\infty, -Z_{上\alpha})$.

例 1.1.3 对某林化厂（间歇式生产）不同时期随机选取该厂三个不同班次的松香各 12 锅，计 36 锅，检验该批产品软化点平均值是否低于 76°（设总体服从正态分布）.

表 1.1.1

班次 序号	1	2	3	4	5	6	7	8	9	10	11	12
Ⅰ	76.8	75.0	75.0	76.0	75.4	75.0	75.6	75.2	75.4	75.8	75.0	74.5
Ⅱ	75.0	75.4	75.2	76.0	75.5	75.4	75.8	75.2	75.5	75.5	74.0	74.2
Ⅲ	76.2	76.0	76.2	76.4	75.8	76.2	75.4	74.6	75.0	75.2	74.4	75.0

解 这是一个单边检验问题，检验假设

$$H_0:\mu = 76°,\ H_1:\mu < 76°$$

由题设，构造统计量

$$Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{75.375 - 76}{0.624/\sqrt{36}} = -6.01$$

对给定的显著性水平 $\alpha = 0.01$，由正态表查得 $Z_{\alpha} =$

-2.33, 而

$$z_0 = -6.01 < -2.33 = Z_{0.01}$$

这表明在 0.01 显著性水平上拒绝原假设 H_0，而接受备择假设 H_1，其结论为：这批松香的软化点平均值低于 76℃.

这种检验的一般叙述方式为：在显著性水平 α 下，检验假设

$$H_0: \mu = \mu_0; \quad H_1: \mu < \mu_0.$$

这里的 H_1 称为备择假设，而 H_0 为原假设或零假设. 一般地，把这种检验称为单边检验，而对假设 $H_0: \mu = \mu_0$ 的检验称为双边检验（实际上，对这类检验，备择假设 $H_1: \mu \neq \mu_0$ 常常省略）.

从以上三个例子的具体讨论，可以总结出假设检验的具体步骤如下：

1. 建立待检假设 H_0 和备择假设 H_1.

2. 选择一个合适的统计量，并知在 H_0 成立下，此统计量所遵从的抽样分布形式. 常用的抽样分布有标准正态分布、t 分布和 F 分布，对应的检验分别称为 z 检验，t 检验和 F 检验.

3. 确定检验形式，按实际问题的性质，决定是选择单侧检验还是选择双侧检验形式.

4. 选定显著性水平 α（一般常取 $\alpha = 0.05$ 或 0.01），查相应的抽样分布的统计表（z 检验查正态分布表，t 检验查 t 分布表等），确定临界值，从而确定 H_0 的拒绝域或接受域.

5. 作出对 H_0 的判断和解释. 把所求的实测统计量值与

临界值比较，若实测值落在 H_0 拒绝域中，则拒绝 H_0；若实测值落在 H_0 接受域中，则接受 H_0.

下面我们简略介绍有关假设检验可能犯的两类错误.

假设检验是以样本提供的信息依据实际推断原理而对总体所作的假设得出接受或拒绝结论的. 由于实际推断原理中小概率事件仍可能发生，所以我们接受或拒绝假设都不是绝对无误的，这就导致了假设检验中可能出现的两类错误：一种是假设 H_0 为真时，作出拒绝 H_0 的错误推断. 这个错误的概率很小，不超过指定的显著性水平 α. 即

$$P(\text{拒绝}H_0|H_0\text{为真})=\alpha,$$

我们称这类错误为**第一类错误**. 另一种是假设 H_0 为假时，作出接受 H_0 的错误推断，并称它为犯**第二类错误**，常记为 β，即：

$$P(\text{接受}H_0|H_0\text{不真})=\beta$$

因此，有时也形象地称这两类错误分别为“以真为假”和“以假为真”的错误.

图1.1.2是双边检验中可能犯的两类错误的示意图.

表 1.1.2

真实情况	判断结果	
	接受 H_0	拒绝 H_0
H_0 为真	正确 概率 $=1-\alpha$	以真为假 (第一类错误)概率 $=\alpha$
H_0 为假	以假为真 (第二类错误)概率 $=\beta$	正确 概率 $=1-\beta$

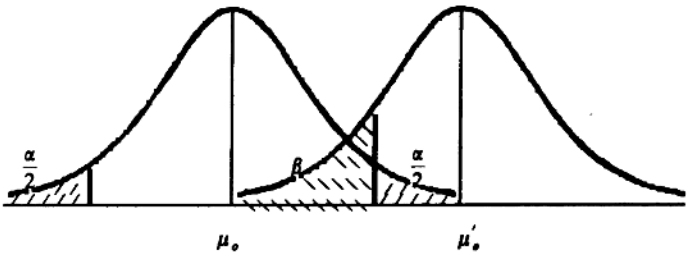


图 1.1.2 两类错误的示意图

在假设检验中，人们总是希望犯这两类错误的概率愈小愈好．但是，当样本容量 n 一定时，要同时减少犯两类错误的概率是不可能的．从图 1.1.2 中可以看到：若 α 减少，则 $\frac{\sigma_0}{\sqrt{n}} Z_{\alpha/2}$ 值必然变大，于是 β 就变大；反之，若 α 变大，则 $\frac{\sigma_0}{\sqrt{n}} Z_{\alpha/2}$ 值变小，从而 β 就变小．因此，在假设检验中，一般只规定显著水平 α，即控制犯第一类错误的概率，而使 β 尽可能地小．若要使 α 和 β 都达到预先指定的值，就必须加大样本容量（事实上，在抽样理论中，正是根据指定的 α 和 β 来确定抽样数量的）．在假设检验中，为了防止 β 过大，一般样本容量 n 不得小于 5.

最后要着重指出的是，经检验没有被拒绝的假设不一定是绝对正确的假设，“拒绝”和“接受”只是我们依据实际推断原理从其中选择一个较为合理的推断而作出的决定而已．

譬如本例，若 Z 落于区间 $[-Z_\alpha, Z_\alpha]$ 之内时，就不否定假设 H_0，这时是否就认为 H_0 成立，即 $\mu=\mu_0$ 呢？