

ICS 67.240
CCS X 04



中华人民共和国国家标准

GB/T 22366—2022/ISO 13301:2018

代替 GB/T 22366—2008

感官分析 方法学 采用三点强迫选择法(3-AFC)测定 嗅觉、味觉和风味觉察阈值的一般导则

Sensory analysis—Methodology—General guidance for measuring odour, flavour
and taste detection thresholds by a three-alternative forced-choice (3-AFC) procedure

(ISO 13301:2018, IDT)

2022-10-12 发布

2023-05-01 实施

国家市场监督管理总局 发布
国家标准化管理委员会

目 次

前言	I
引言	II
1 范围	1
2 规范性引用文件	1
3 术语和定义	1
4 基本原则	3
4.1 检验步骤	3
4.2 数据处理	3
5 检验程序	3
5.1 样品制备	3
5.2 刺激浓度选择	4
5.3 样品提供	4
5.4 评价员训练	5
5.5 评价员选拔	5
5.6 实验设计	6
6 数据处理	8
6.1 数学和统计模型	8
6.2 数据预检验	8
6.3 logistic 模型数据拟合的最大似然法和误差限估计	9
6.4 结果解释	9
6.5 觉察概率(p_d)不等于 0.5 时的情况	10
6.6 最优估计阈值(BET)估计	10
6.7 结果表述	11
附录 A (资料性) 给定精度下所需的评价员数量估计	12
附录 B (资料性) 示例	13
参考文献	25

前 言

本文件按照 GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

本文件代替 GB/T 22366—2008《感官分析 方法学 采用三点选配法(3-AFC)测定嗅觉、味觉和风味觉察阈值的一般导则》，与 GB/T 22366—2008 相比，除结构调整和编辑性改动外，主要技术变化如下：

- a) 将“范围”中的注更改为正文内容，将“化学刺激物”更改为“刺激”(见第1章，2008年版的第1章)；
- b) 将术语“刺激物”“待测样”“三点选配法检验”更改为“刺激”“测试样”“三点强迫选择检验”，并修改了“三点强迫选择检验”的定义(见3.1、3.4和3.5，2008年版的3.1、3.4和3.5)；
- c) 更改了“检验步骤”中的内容(见4.1，2008年版的4.1)；
- d) 删除了“数据处理”中关于评价员提交结果的要求(见2008年版的4.2)；
- e) 删除了“评价员的训练”中“若评价员对样品比较熟悉，则这种偏差可通过强化训练来降低”的内容(见2008年版的5.4)；
- f) 更改了“总则”中遵守的要求(见5.5.1，2008年版的5.5.1)；
- g) 更改了图2“群体嗅觉阈值的分布”(见图2，2008年版的图2)；
- h) 将“数学和统计模型”中的“ p_e ”更改为“ p_c ”(见6.1，2008年版的6.1)；
- i) 增加了参数 b 正、负情况的内容(见6.3.2)；
- j) 更改了“结果解释”的内容(见6.4，2008年版的6.4)；
- k) 更改了“基于最大似然法的结果和计算”中的内容(见B.2.3，2008年版的B.2.3)；
- l) 更改了“阈值估计中步长误差限初始估计”表中的内容(见表B.4，2008年版的表B.4)；
- m) 增加了“用线性回归法计算”中对回归模型使用的相关内容，将“log”更改为“ln”(见B.2.4，2008年版的B.2.4)；
- n) 增加了“结果和计算”中对B.2的说明(见B.3.3)；
- o) 将模型3的比较由“1与3”更改为“2与3”(见表B.6，2008年版的表B.6)；
- p) 将“松油烯”更改为“萜烯”(见B.4，2008年版的B.4)；
- q) 将“ α -萜烯”和“ β -萜烯”的 C 为1.000 mL/m³时的ln(C)由“−2.30”更改为“0.00”(见表B.7，2008年版的表B.7)。

本文件等同采用 ISO 13301:2018《感官分析 方法学 采用三点强迫选择法(3-AFC)测定嗅觉、味觉和风味觉察阈值的一般导则》。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由全国感官分析标准化技术委员会(SAC/TC 566)提出并归口。

本文件起草单位：中国标准化研究院、江苏大学、黑龙江飞鹤乳业有限公司、利和味道(青岛)食品产业股份有限公司、中国科学院数学与系统科学研究院、元气森林(北京)食品科技集团有限公司、中国绿色食品有限公司、浙江工商大学、北京工商大学、北京林业大学、爱普香料集团股份有限公司。

本文件主要起草人：钟葵、张迪、赵镭、刘文、汪厚银、冯士雍、宋焕禄、史波林、姚敏、李懿霖、张永昌、李林、曹珂璐、唐飞、刘野、田师一、朱保庆、邹小波、陈博、王鹤妍、张璐璐、王思思。

本文件及其所代替文件的历次版本发布情况为：

——2008年首次发布为 GB/T 22366—2008；

——本次为第一次修订。

引 言

“阈值”是感官分析中的重要术语，广泛应用于食品和饮料的感官研究中。刺激的感官阈值的数据常应用于两方面研究：测量评价员或评价小组对特定刺激的敏感性，以及检验物质引起评价员产生感官响应的能力。前者用于判别评价员的评价水平高低，后者主要作为某种物质特性的度量指标。

术语“阈值”是 19 世纪心理物理学家提出的，用来表示一个刺激的浓度，高于此浓度时刺激可被觉察，低于此浓度时则无法被觉察[见图 1a)]。

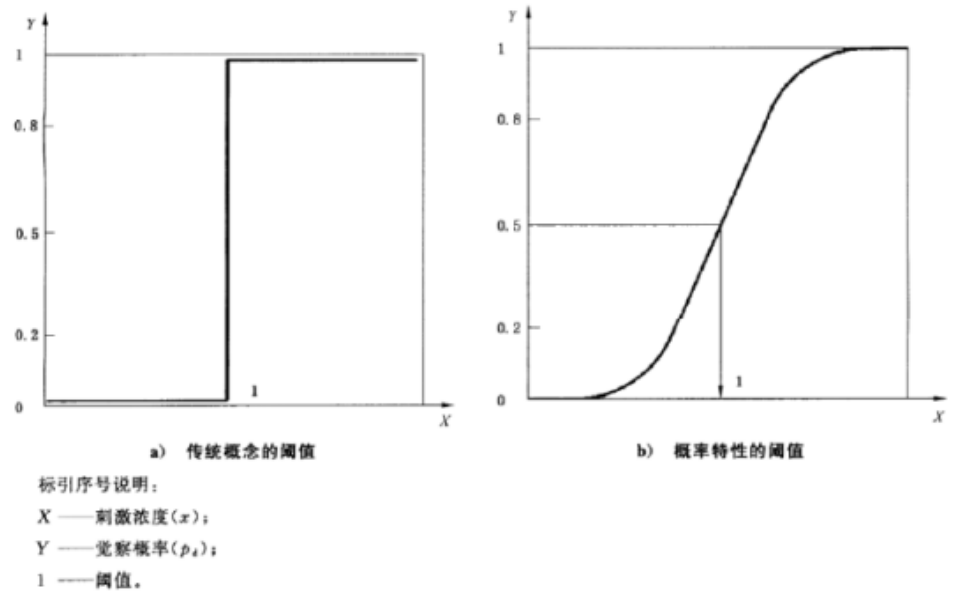


图 1 传统概念和概率特性的阈值

但是，多数情况下刺激浓度与觉察概率¹⁾之间的关系表现为 S 形曲线[图 1b)]。出于计算的目的，一种便捷方式是假定阈值是波动的，某一特定刺激浓度有时可能高于阈值，有时又可能低于阈值。因此通过估计这些瞬时值的中位数可得到阈值，即阈值是觉察概率为 0.5 时的刺激浓度。这种阈值定义类似于生物学中的药理学、毒理学等学科中有关化学物质对生物体半数有效量的定义。

多家实验室检测结果均表明水或空气中某一物质觉察阈值的变化范围通常可达 2 个~3 个甚至更多数量级，大大超出了预期的实验误差或者数据处理造成的差异(见参考文献[10]、[13]和[17])。推测原因可能与各实验室对阈值定义存在差异或采用了不同的检验程序有关。因此，Devos 等提出了一个检测空气中觉察阈值的标准化程序(见参考文献[10])。

本文件使用时需要注意，觉察阈值测定并非简单的表现描述，需要开展大量的实验工作。实验结果

1) 本文件采用 3-AFC 法对刺激进行检验，觉察概率 $p_d = 1.5 \times p_c - 0.5$ 。其中， p_c 是做出正确选择的概率。严格来说，这是评价员行为的“猜测模型”，并非评价员判断过程的心理测量学模型，例如信号检测模型同样适用，见参考文献[16]。

表明,反复多次测试后会观察到评价员的个体阈值趋于降低、个体间的差异趋于减小。评价员在阈值测定初期对测试并不熟悉,但随着对测试样和检验程序的逐渐熟悉,其灵敏度也逐步提高。3-AFC 法要求评价员能识别出刺激。训练程序尽管很繁琐,但也会提供给实验者一些所需信息,如每位评价员的觉察浓度范围。当实验者了解了评价员个体的觉察浓度范围,并为其提供适合的浓度范围时,检验结果将会改进,见 6.3。

感官分析 方法学

采用三点强迫选择法(3-AFC)测定 嗅觉、味觉和风味觉察阈值的一般导则

1 范围

本文件提供了采用三点强迫选择法(3-AFC)测定刺激引起的嗅觉、味觉和风味响应数据,并通过数据处理进行阈值估计、误差估计以及与检测相关的其他统计程序的指导。

本文件中给出的方法适用于下列两种情况中任意一种(大多数实验均能包括下列两种情况)。

- a) 测定评价员对特定刺激的敏感性;主要用于个体或特定群体中个体敏感性差异的研究,以及年龄、性别、生理状态、疾病、药物作用和环境条件对敏感性的影响研究。
- b) 检测化学物质引起感官响应的特性,主要包括下列五个方面:
 - 风味化学研究及特定化学物质对食品风味影响的研究;
 - 根据目前环境中存在的化学物质对人类的影响进行分类;
 - 化学物质的分子结构与其刺激能力之间的关联性研究;
 - 排放气体、水、食品和饮料的质量保证;
 - 嗅觉相关机制研究。

上述情况中,由于评价员做出正确选择的概率随刺激强度变化而改变,所以对于这种变化的分析,即对剂量-响应(效应)关系曲线斜率的分析也是除阈值测量外的一个重要研究内容。本文件中给出的数据处理程序中提供了这方面的相关信息。

本文件的核心是数据获取和计算程序,对于数据的有效性,本文件仅给出了一般规则和预防措施。本文件中并未严格区分“觉察阈值”和“差别阈值”。实际上,本文件给出的方法是测定差别阈值,即通过测试样和参比样比较进行阈值测量。本文件不排除在实验设计时参比样中含有刺激,或者并不确定参比样中是否含有刺激。本文件给出的方法不适用于 ISO 5492 中定义的“识别阈”的检验,也不涉及 EN 13725 中空气质量测定的标准化方法。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中,注日期的引用文件,仅该日期对应的版本适用于本文件;不注日期的引用文件,其最新版本(包括所有的修改单)适用于本文件。

ISO 5492 感官分析 术语(Sensory analysis—Vocabulary)

注: GB/T 10221—2021 感官分析 术语(ISO 5492:2008, IDT)

3 术语和定义

ISO 5492 界定的以及下列术语和定义适用于本文件。

ISO 和 IEC 维护的用于标准化的术语数据库网址如下:

——IEC 电子百科: <http://www.electropedia.org/>

——ISO 在线浏览平台: <https://www.iso.org/obp>

3.1

刺激 stimulus

能激发感受器的因素。能否被一种或多种感受器感知,主要取决于该刺激物存在的量。

3.2

介质 medium

阈值测定时,用于溶解、稀释、分散或吸附刺激(3.1)的物质。

3.3

参比样 reference sample

不添加待测刺激(3.1)的介质(3.2)。

3.4

测试样 test sample

含有已知浓度刺激(3.1)的介质(3.2)。

3.5

3-AFC 检验 3-AFC test

三点强迫选择检验 three-alternative forced-choice test

提供给评价员三个样品,其中一个是测试样(3.4),含有评价员熟悉的某种已知刺激(3.1),另外两个是参比样(3.3),要求评价员从中识别出测试样。

注:标准 3-AFC 检验属于一种特定差别检验方法,要求评价员根据某种指定特性进行判断。阈值测定时,评价员对所关注特性的识别阈值通常会偏低。因此,阈值测试时通常要求评价员采用非特定 3-AFC 检验,选择与另外两个不同的样品进行比较,而非依据某种特定的特性。

3.6

提供样品 presentation

一套用于三点强迫选择检验(3.5)的三个样品。

3.7

阈值模型 threshold model

特定检验中,检验刺激(3.1)能(结果是做出正确选择)或未能被觉察到(结果是做出随机选择)的感官检测模型。

3.8

信号检测模型 signal-detection model

特定检验中,有证据表明刺激(3.1)存在的感官检测模型。

注:这些证据有助于评价员判断刺激是否存在。

3.9

觉察阈值 detection threshold

引起感觉所需要的感官刺激(3.1)的最小值,即在测试条件下,从阈值模型(3.7)计算出的觉察概率为 0.5 时引起感官刺激所需要的刺激的最低浓度。

3.10

个体阈值 individual threshold

单个评价员的觉察阈值(3.9)。

3.11

平均阈值 average threshold

个体阈值(3.10)的平均值(明确均值的类型,如算术平均值、几何平均值或中位数等)。

3.12

合并数据的组阈值 group threshold from pooled data

将一组特定的评价员对每个浓度刺激(3.1)的评价结果输入到统计模型中,通过模型拟合得到的估计值。

4 基本原则

4.1 检验步骤

用介质配制一定浓度的刺激样品,连同两个参比样一起提供给评价员,要求评价员从中选出含有刺激或者刺激浓度更大的样品。评价员宜做出选择。通常,刺激溶解于空气中或水中。检测时,若提供给评价员的是气态介质,使用空气而非其他气体作为介质;若提供的是液体或固体介质,使用液态或固态食品代替水作为介质去稀释或吸附刺激。无论使用何种介质,介质均均匀一致,确保所提供的两个参比样及所有测试样的一致性和均一性。

配制一系列不同浓度的刺激样品,对每个浓度点进行重复检验。确保检验次数足够多,以获得期望的阈值精度和数学模型的参数。实验设计时,重复实验中包括评价员个体、评价员之间以及兼顾两者的重复检验。

4.2 数据处理

每次检验时,评价员提交的结果都是一个二元数据,即评价员指出的样品不是测试样(正确选择)就是参比样(错误选择)。统计模型的判断依据是某一浓度下做出正确选择的次数是服从二项分布的。

3-AFC 检验中,阈值是正确选择次数比例在 $2/3$ 时的刺激浓度,此时对应的觉察概率是 0.5 [见式(1)]。正确选择次数的比例能简单地通过计算和内插得到。但对阈值和阈值范围更准确的估计是通过拟合模型获得的。许多准则中常采用 logistic 模型,该模型是通过最大似然法或最小二乘法进行拟合。拟合过程主要是估计模型中的两个参数——位置参数和形状参数。前者确定拟合曲线的位置,后者决定曲线的陡峭性。通过拟合曲线,可推导出除 50% 以外的其他觉察比例的估计值。

最简单的拟合模型是正确选择次数的比例近似服从一个简单的正态分布。这种情况通常发生在数据来自单个评价员重复评价的情形。一般来说,一个简单 logistic 函数在确定其中一对参数值后,就具有较好的拟合性。但在评价员对化学物质敏感性的评价中,更常见的情形是数据既不服从正态分布,也并非对称分布。对于某些刺激,其分布呈明显的双峰分布,但其与正态分布的偏离很难被表达,除非评价员数量很大,如在 100 名以上。简单的 logistic 函数不能用来拟合显著偏离正态分布的数据,此情形下建立其他数学模型。

5 检验程序

5.1 样品制备

5.1.1 一般原则

依据 ISO 6658 中的规定,选取的刺激和介质在整个检验过程中保持稳定,且无毒性、无致敏性。确保选取的刺激和介质能实现检验目的,例如排出的气体可随着处理过程进行而发生变化,所以可先对化学物质进行纯化处理,去除异常气味或刺激后再用于检验。准备足量、均匀的刺激和介质,确保整个检验过程中无论刺激的浓度和位置如何设置,呈现给评价员时的样品都是均匀一致的。样品制备设施符合 ISO 8589 中的规定。使用的容器不会吸附测试样,也不产生任何气味和味道。除化学检验外,确保评价员不能以视觉或其他任何方式觉察出介质中是否含有刺激。非检验期间,样品避光、避热存放。

5.1.2 气体样品

将刺激和介质收集在带有聚四氟乙烯(PTFE)涂层的瓶状或球状容器中。若刺激是一种含异味的无嗅气体时,用新鲜样品连续吹扫冲洗容器、阀门和管道数次,使气体样品在整个容器内达到饱和。同

样理由,为避免气体体积变化,在一个近似恒定的温度下向评价员提供气体样品。使用具有聚四氟乙烯涂层的光滑管道和阀门,避免压力的突然改变。

5.1.3 液体样品

用液态介质配制刺激时,确保样品能在整个检验过程中始终保持完全溶解的状态。对于不能完全溶于水的物质,先用经活性炭处理除味过的乙醇或乙二醇溶解和稀释。由于蒸馏水和无水乙醇常带有一定气味,必要时用食用级产品代替或使用前先用活性炭纯化。对于完全不溶于水的疏水性刺激,可使用无嗅疏水溶剂配制,如液体石蜡、邻苯二甲酸二壬酯等,避免接触塑料容器,防止容器中的某些物质溶解析出。在逐级稀释时,随稀释比例增高,刺激因器壁吸附造成的损失比例也相应增大。因此,尽可能采用微量注射器进行稀释,且采用直接原液稀释而非逐级稀释。

5.1.4 固体样品

通常的典型介质有奶酪、鱼和肉等食品。如果无法实现将固体样品分散后再重新组合,将刺激精细切分或粉碎后再加入适当的介质中,与介质充分混合,确保刺激在介质中充分扩散一段时间后,再提供给评价员。

5.2 刺激浓度选择

根据 3-AFC 检验的需要设置一个样品系列,样品浓度逐渐递增,后一个样品浓度近似为前一个样品浓度乘以一个因子 X 。 X 的值依据可接受的阈值估计误差限而定。对于精度要求不高的研究,一般选取 $X \approx 3 \sim 5$;而对于精度要求较高的研究,一般选取 $X \approx 2$ 。根据每位评价员的个体特性选择一个合适的检验方案,确保检验结果能够对应 logistic 模型 S 形曲线的某个范围,且分布在评价员的偏好性觉察阈值范围内。最有效的数值点是对应于正确选择概率范围为 45%~90%的点,即 $p_d = 0.18 \sim 0.85$ 。

为节约样品和评价员时间,通常先设置一个较大的 X 值,确定每位评价员能感知的刺激浓度范围,阐明检验机理,并教会评价员如何在刺激浓度超过其偏好性觉察阈值范围时识别出刺激。

然后,再选取一个较小的 X 值为每位评价员设置一套适合的 3-AFC 检验浓度梯度。检验完成后,若检验结果无法确定评价员觉察阈值时,负责人通过增加刺激浓度水平的设置继续检验,直到满足要求为止。每隔几个样品就要求评价员描述其对刺激的感觉,防止对刺激的记忆消失。或者采用询问方式,测定评价员在没有觉察刺激时随机猜对的次数,例如评价员连续三次都猜对的概率是 1/27。

5.3 样品提供

5.3.1 准备工作

将制备好的测试样提供给检验区(检验区规定见 ISO 8589)的评价员,评价员遵守 ISO 6658 描述的感官分析良好操作规程进行样品评价。测试样采用三位随机编码,或按一定模式摆放,如将第一个样品放在评价员前方最左边,然后从左到右依次摆放样品,回答表上也采取同样的排列方式记录。采取 AAB、ABA 和 BAA 方式,平衡三个样品在评价员面前的提供次序,避免位置偏差。对高于觉察阈值浓度的样品,评价员只摄入最低摄入量。此外,每两个样品之间留有足够长的时间间隔,确保评价员的感官能力得以恢复,降低评价员的感官疲劳。

5.3.2 气体样品

气体样品的提供见参考文献[12]和[15],借助嗅觉测量装备进行样品评价。

5.3.3 液体样品

提供非挥发性化学物质液体样品时,先将非挥发性化学物质溶解于纯水或无味溶液中,再用不吸附

该物质的容器存放,如使用 100 mL 的玻璃烧杯,且溶液不超过容器的 1/4 体积。对于挥发性化学物质液体样品,应使用具塞且适宜嗅闻和品尝的广口容器,如使用 250 mL 挤压瓶,将标准测量体积的顶空气体或液体送入鼻腔或口腔中,见参考文献[8]、[12]和[14]。介质为饮料时,使用符合日常习惯的感官评价容器。

5.3.4 固体样品

当使用食品作为介质时,样品提供形态与通常感官评价时该产品的形态保持一致。

5.4 评价员训练

多数情况下,检验中所关注的是有经验观察者的觉察阈值。只有多次参与过检测物的相关训练,能够在任何时候只要刺激(例如空气或水中的某种污染物,或者是食品或饮料风味中的某种成分或杂质)存在就能觉察出的评价员才能称为有经验的评价员。因此,对检测物的熟知也是 3-AFC 检验中的一项要求。若评价员缺乏足量的训练,检测阈值范围的上限将会人为地增加 1 个~2 个数量级。若是检测非专业检验者的觉察阈值时,如家用煤气泄漏的警告剂量,则可采用未经训练的评价员,且允许存在一定的干扰物(如噪声),同时采用三点检验或者成对比较检验代替 3-AFC 检验。

训练程序首先选取一个较高的刺激浓度,再选择两种或更多的刺激浓度,要求评价员对其排序。确定评价员的偏好性觉察阈值范围后,再采用 3-AFC 法进行检验。经过该训练程序后,评价员的阈值水平会降低,并在 3 次~5 次检验后逐步趋于稳定。不同评价员对同一物质的敏感性不同,评价员之间敏感性的差异可达到 2 个~3 个甚至更高的数量级。

5.5 评价员选拔

5.5.1 总则

依据研究目的进行评价员选拔,指导原则见 ISO 8586。

5.5.2 个体阈值

可采用下列方法进行检验,如将个体的觉察阈值同已有文献中的阈值、不同环境下的测定值以及评价员对其他物质的觉察阈值进行比较。该检验可用于嗅觉缺失或嗅觉过敏症,以及味觉缺失或味觉过敏症的诊断。

5.5.3 阈值分布

实验者有时希望得到某个群体中的阈值分布状况,该群体可能是更大群体中的一个抽取样本,又或者是被选人群的全体成员,如评价小组的所有成员。本文件中的内容并不涉及群体样本选择,但实验者宜仔细研究群体及其样本。结果表述见 6.7。

5.5.4 刺激阈值测定

刺激的检验数据或平均阈值仅对检验中所使用的评价小组有效。对此评价小组以外的结果,实验者宜谨慎进行推断。实验者宜根据实验的目标和目的来选择评价小组。例如,对一组化学物质进行感官特性研究时,可选择一个人数较少的评价小组;若是对食品中潜在风味化合物特性进行研究时,可选择一个人数较多的评价小组,因其代表了一个特定的群体。

要想获取更加准确的估计值,评价员数量和样品检验数量是需要同时考虑的问题。评价员人数较少时,有必要增加样品检验次数,以获得足够的数据量。而对于人数较多的评价小组,每位评价员只需对每个样品的一个浓度或部分浓度进行检验。

5.6 实验设计

5.6.1 个体阈值

评价员做出正确选择的比例为 45%~90%时所对应的刺激浓度,可有效用于 logistic 参数估计。此浓度范围内,如果刺激浓度主要分布在阈值周围,参数的精度主要取决于样品量。表 1 列出了样品量与阈值估计值的误差限之间的关系,单位是原始浓度,见附录 A。

表 1 阈值估计值达到期望精度时的所需样品量

样品量	40	60	80	100	120	160	200
阈值估计值的误差限	2.5	2.2	2.0	1.8	1.7	1.6	1.5

阈值估计值的变化范围是通过阈值除以和乘以表 1 中的误差限来获取。例如,样品量为 80、觉察阈值为 2.4 mL/m^3 时,阈值的范围为 $1.2 \text{ mL/m}^3 \sim 4.8 \text{ mL/m}^3$ 。当样品量达到 200 以上时,精度提升变得缓慢。因此,通过增加样品量进一步提高精度的方法并不可行。此时采用下列方法可有效提高精度。先对每个浓度开展少量的重复检验,再利用 logistic 模型拟合计算阈值和误差限。通过 logistic 模型获得有效的浓度范围后,进一步对每个有效浓度点进行多次重复检验,重复上述步骤,直至达到所需的精度。

5.6.2 阈值分布

对选择的评价员按照 5.6.1 中的规定开展检验,结果以直方图或累积频率图呈现。平均阈值结果以算术平均值、几何平均值和中位数等形式给出。若阈值呈现双峰或多峰分布时,采用适合的对应处理方法,数据处理见第 6 章。

5.6.3 评价小组组阈值的测定

5.6.3.1 总则

实验设计时,评价员间的敏感度差异可能会比评价员个体敏感度差异大上数倍。因此,采用增加评价员数量,而非增加每位评价员所评价样品数的方式来设计重复检验,这种检验结果更实用且更具有代表性。

5.6.3.2 合并数据的组阈值

不同于个体阈值中将每位评价员数据单独进行 logistic 模型拟合,组阈值是采用合并数据进行拟合,即将样品每个浓度下所有评价员结果一起输入模型中,进行 logistic 模型拟合。数据量越多、分布越集中时,组阈值的拟合性就越好。3.9 中定义的觉察阈值属于组阈值,具体见附录 B 中 B.2 和 B.4。该方法适用于个体差异不作为主要考虑因素的实验设计,例如依据化学物质的污染或感官污染的严重程度进行分级的研究。

5.6.3.3 平均阈值

对选择的评价员依据 5.6.1 中的规定重复测定,结果以直方图(如图 2)或累积频率图呈现。若个体差异也是研究目的一部分时,如风味化合物、污染物或感官污染对特定群体的影响研究,采用此方法。

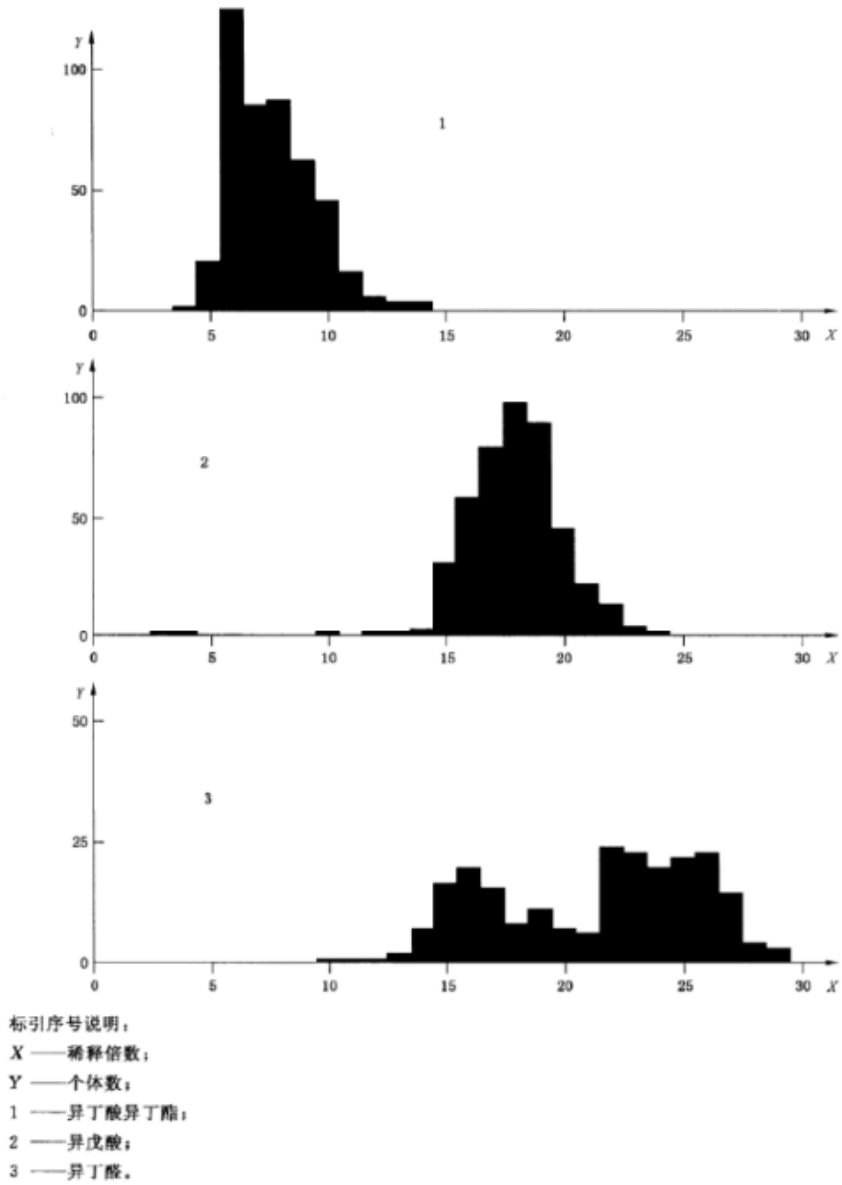


图 2 群体嗅觉阈值的分布

图 2 中 1 和 2 有 443 位评价员参与阈值测定,图 2 中 3 有 222 位评价员参与测定。为方便三个图直观比较,可将图 2 中 3 的纵坐标加倍。稀释步骤中,“0”代表每种气味的饱和浓度,为最高阈值(见参考文献[7])。

6 数据处理

6.1 数学和统计模型

3-AFC 检验中,评价员对三个样品进行随机选择,因为三个选项中只有一个是正确的,因此选择正确的概率是 $1/3$,近似为 0.33 。根据阈值模型,选择正确的概率 p_c 与觉察概率 p_d 之间的关系见式(1):

$$p_c = p_d + \frac{1}{3}(1 - p_d) = \frac{2}{3}p_d + \frac{1}{3} \quad \text{.....(1)}$$

检验中,评价员能正确觉察出刺激的概率为 p_d , $(1 - p_d)$ 表示评价员不能觉察出刺激的概率。此时,评价员一定要做出随机猜测,猜测正确的概率仅为 $1/3$ 。由于阈值是 $p_d = 0.5$ 时的刺激浓度,按式(1),即 $p_c = 0.67$ 时的数值。

p_c 是检测得到的数据, p_d 是阈值模型中的推算数据。由此反向进行推导,见式(2):

$$p_d = 1.5 p_c - 0.5 \quad \text{.....(2)}$$

通常设置一系列不同浓度的刺激,并对每个浓度进行重复检验,用评价员做出正确选择的概率 p_c 和刺激浓度作图,图形大致为 S 形曲线。根据式(2),将 p_c 转换成 p_d 后可得到另一条 S 形曲线[图 1b)]。当浓度足够低或足够高时,曲线的渐近线分别为 0 和 1。若评价员对刺激比较敏感的情况下,刺激的浓度通常以浓度对数值或稀释比表示。

p_c 与刺激浓度的 S 形曲线能用累积正态分布拟合,或更方便地用累积 logistic 分布拟合,见式(3):

$$p_c = \frac{2/3}{1 + \exp[b(t - x)]} + 1/3 \quad \text{.....(3)}$$

式中, x 代表刺激浓度,系数 t 和 b 由数据拟合求得。当 $x = t$ 时, $p_d = 0.5$,此时 t 为刺激浓度的阈值。参数 b 是 x 每变化一个单位时 p_c 的变化,决定曲线分布的陡峭度。

6.2 数据预检验

6.2.1 准备工作

采用直接观察法或将 p_c 和浓度对数值作图的方法对数据进行初步检验。需注意,无论结果是否呈现 S 形曲线,以及待测浓度是高于或低于估计阈值浓度,用于阈值估计的数据都宜准确。否则要获取更多的检验数据。先直观预估阈值,判断准确度是否能满足需要。若否,采用 6.2.2 中的 Logit 变换,或者 6.3 中的最大似然法,将数据处理后再进行模型拟合。见 B.2~B.4。

6.2.2 采用 Logit 变换预估阈值和斜率

Logit 变换等价于熟知的概率(probit)变换,只是概率变换是基于累积正态分布的,而 Logit 变换是通过式(4)将 p_d 转变成对数单位,即下列 logit 形式:

$$L_d = \ln\left(\frac{p_d}{1 - p_d}\right) \quad \text{.....(4)}$$

结合式(4)和式(2),转换成式(5):

$$L_d = \ln\left(\frac{p_d - 1/3}{1 - p_c}\right) \quad \text{.....(5)}$$

将式(3)代入式(5),得到式(6):

$$L_d = b(t - x) \quad \text{.....(6)}$$

当 p_c 服从 logistic 模型分布时, L_d 随着刺激浓度增大呈线性增加。此时决定是采用浓度的转换数据还是采用原始数据来进行计算。采用原始数据绘制的 p_c 与 $\ln x$ 的图形中,图中曲线的中间范围近似为线性时,阈值是 p_c 为 $2/3$ 时所对应的浓度。但是,无论是根据直接观测的 S 形曲线或是 p_c 中间范围

数据拟合为线性,采用变换数据并未提供任何便捷。

对于变换数据,阈值是 $L_d=0$ 时的浓度自然对数值,通过目视大致为一条直线时估计,或通过线性回归拟合估计。注意,渐近线附近的变换值可能并不稳定,该区域范围内发生很小的改变均可能会对结果带来较大影响。因此,在使用数据点进行曲线拟合时,忽略 $p_c < 0.43$ 或 $p_c > 0.9$ 的数据(即转换值 $L_d < -1.75$ 或 $L_d > 1.75$)。除非检验的关注点就在该区域,见 6.5。还要注意变换数据作图可以直接解释式(3)中参数, t 对应于 $L_d=0$ 时的刺激强度(浓度自然对数值), b 是直线的斜率。

6.3 logistic 模型数据拟合的最大似然法和误差限估计

6.3.1 总则

采用最大似然法(Maximum likelihood, ML)拟合 logistic 模型[式(3)]数据的原理是获得使似然函数结果最大的参数 t 和 b 的值,即在此参数值时,获得实际检验数据比得到其他值的可能性更大。选用合适的专用软件程序²⁾进行 ML 估计,或者更简单采用 B.2~B.4 中描述的电子表格程序。例如,ML 结果表明估计值大于 t 值上限的概率仅为 0.05。通常情况下,一种 ML 的常用替代方法是使用 logit 转换值进行线性回归拟合,如 B.2.4 所示。

6.3.2 参数 b

参数 b 是拟合方程的斜率,决定了拟合曲线的陡峭程度。 b 为正值时表明觉察概率会随着刺激浓度增加而增高,而 b 为负值时则表明觉察概率和刺激强度之间存在反比例关系。例如,用稀释倍数衡量刺激强度时, b 通常为负值(稀释度越高=强度越低=觉察概率越低)。对于评价员个体而言, b 值代表刺激强度变化时,评价员识别能力的高低,与韦伯常数或史蒂文斯幂律指数等指标相关。评价员个体不仅阈值不同,其 b 值同样存在差异, b 值的大小及其在群体中的分布情况都是阈值测定所关注的内容。 b 值高的评价员(陡峭的斜率)对刺激强度的微小变化会更敏感,在与质量控制和监测相关的任务中则更加有效。在评价员筛选时,对 b 值的认知与对阈值的认知同样重要。

6.3.3 参数估计值的置信区间

参数估计值的置信区间是包含参数真值的一个取值范围。置信区间越窄,估计值的可信度越高。估计值的精度能通过增加数据总量的方式来提高,或通过在 0.25 倍~4 倍阈值浓度范围内设置等间距浓度梯度的方式来实现。

6.4 结果解释

阈值测定结果可用于多种目的,本文件并未给出基于特定目的进行阈值检验的实验设计指导。但进行阈值结果解释和比较时,考虑数据采集和分析的方式,以及统计显著性水平的设置方式。

单个评价员的数据结果最容易解释和比较。logistic 拟合模型是评价员的心理物理学函数,由此统计分析得出的结果能用于评价员之间或者同一评价员对不同测试样之间的比较。

通过简单 logistic 扩展模型,能对来自不同评价员或不同测试样的数据进行比较。其他设计还包括选择不同评价小组代表不同群体,并在小组间重复提供样品。对于不同物质间的比较数据或不同评价小组对于某一给定物质的比较数据,通常采用标准方差分析方法来完成(假设所有估计值是以相同的方法获得,并且提供了相同的样品量,则评价员个体的 t 和 b 的估计值作为分析用的输入数据)。如果数据是 5.6.3.2 中所述的合并数据时,则 t 和 b 的估计值能用于描述物质间或评价小组间的差异。

评价员间阈值的分布可能会偏离正态分布,该情况下首先采用直方图(见图 2)、正态概率图或

2) 如 GLIM 和 SAS 软件是目前比较通用的商业化产品。该信息只是给本文件的使用者提供一些便利,并不涉及推荐认可产品。

logistic 概率图进行检验。如果偏态明显或呈明显的双峰分布,则选取合适的平均阈值,如偏态数据的中位数、双峰或多峰分布的多重平均值等。

合并数据的组阈值及其误差限估计的实验设计中,每位评价员对每个浓度仅检验一次。该方法获取的数据具有较高可变性,通常不推荐使用。多数情况下宜采用另外一种实验设计,即每位评价员在每个刺激浓度下至少进行一次检验,见示例 B.1。实验者计算平均阈值和合并数据的组阈值,并检验阈值的分布。请注意,采用这种实验设计时,估计的个体阈值精度较差,而合并数据的组阈值具有较高精度。例如,每组 50 位评价员对 5 个刺激强度进行检验,每个强度检验两次的实验设计中,共有 500 个数据可用于组阈值估计,但只有 10 个数据能用于评价员个体阈值的估计。

尽管使用每位评价员在每个刺激浓度下仅做一次检验的实验设计,即可进行合并数据的组阈值及误差限的估计。但更常用的实验设计是,要求评价员在每个浓度下至少进行一次检验。群体对刺激通常会表现出较宽范围变化的个体阈值。对于单个评价员,一个 100 倍的浓度范围即可覆盖 p_d 值从 0.05~0.95 的变化范围;但对于一组评价员,个体阈值的变化范围可达 10 000 倍。若仅向评价小组提供一个 100 倍浓度范围的刺激测试,这意味着对于部分评价员(通常是相当数量的评价员),所测试的浓度范围可能都是接近其个体心理物理学函数曲线渐进线的浓度范围。渐进线附近的结果对于 t 值和 b 值的估计几乎没有贡献,这表明敏感性较高或较低的评价员对合并数据的参数估计意义不大。若非必要,不宜将这些数据合并进行组阈值估计,或阈值估计时宜对不同敏感性的评价员赋予不同的权重。

进行数据拟合处理的前提是假定阈值分布服从 logistic 模型,与 logistic 分布的任何偏离均表明数据拟合不足。采用统计中的拟合优度检验程序对拟合不足进行检验,但对简单 logistic 模型的偏离通常不太容易被检验出,除非实验设计中考虑了 10 种以上、大于 500 倍的浓度范围以及此范围内总共数百次的检验。如果拟合优度检验结果是拟合度显著不足,则考虑采用简单 logistic 函数之外的其他模型来拟合。最简单的方法是考虑对部分评价员采用二次 logistic 函数建模,函数带有不同 t 值,也可能包括不同的 b 值,该情形适用于偏态分布和双峰分布的情况。

6.5 觉察概率(p_d)不等于 0.5 时的情况

监管者通常希望空气中的恶臭物质在 5% 概率时就能被觉察,风味学家则希望能在 95% 的概率时判断出食物中是否添加了某种浓度的风味物质。此效应水平能通过 logistic 曲线计算得出,即分别是 L_d 值为 -2.94 和 2.94 时所对应的刺激强度, L_d 值通过式(4)或式(5)得到。 p_d 值过高或过低时,实验者宜确保该区域(高值区或低值区)内具有足够多数量的数据,以获得此范围相应的刺激强度。超出研究范围外的任何推断均不具有可信度。

6.6 最优估计阈值(BET)估计

最优估计阈值(BET)的快捷程序(见参考文献[4]和 B.1)能获得组阈值的粗略估计,但精度较低,且存在一定的风险性。此方法基于阈值模型,见 3.7, 3-AFC 检验中,若样品浓度低于阈值水平时,评价员做出正确选择的概率为 1/3;若样品浓度高于阈值水平时,做出正确选择的概率为 1.0。该方法耗时短,每位评价员对每个浓度水平下的样品仅评价一次,但要求有较多的评价员数量。

为方便清楚地观察到随着刺激浓度增大时做出正确选择的分布情况,按照样品浓度递增的顺序(按稀释度递减的顺序,见 B.1)将数据制成表格。BET 值计算如下:即做出错误选择所对应的最高浓度和其相邻做出正确选择的更高浓度的几何平均值。如 B.1 中所示,评价员 1 的 BET 为 $\sqrt{135 \times 45} = 78$ 。

如果评价员对一组浓度全部做出了正确选择,或者在最高浓度时就做出了错误选择,则不能使用该方法,如 B.1 中的评价员 6 和评价员 4。此时,推荐适当扩展浓度梯度后继续检验,或采用下列常用方法:若评价员在最高浓度时就做出错误选择,则假设最高浓度的相邻较高浓度会选择正确,并由此计算 BET 值。如果是一组全部选择正确的数据,则假定最低浓度的相邻较低浓度会选择错误,再进行 BET 值计算。

计算评价员个体 BET 的几何平均值即可获得评价小组的组 BET。一种简单度量评价员间变异的方法是计算 BET 对数值的标准差,如示例 B.1 所示。由于一次猜测正确的概率为 $1/3$,连续 2 次或者 3 次猜对的概率分别为 $1/9$ 和 $1/27$,因此组 BET 的结果可能存在偏差。该方法风险较大,因为每位评价员只检验一次样品,由于样品混淆或对刺激及测试方法不熟悉等原因,可能会造成对高于阈值的样品做出错误选择。如果 BET 恰恰落在浓度系列的极值范围内,或者测试的浓度扩展系列过少,则 BET 对数值的标准差可能会被低估。

6.7 结果表述

检验报告中宜包括下列内容。

- a) 所有的检验条件:包括样品性质及来源、取样方式、介质(稀释液)选取,以及提供样品时所使用的系列容器和检测设备。
- b) 样品浓度或所用的流速、温度及其他条件。
- c) 给评价员的检验步骤说明书和评分表。
- d) 每一步的稀释因子。
- e) 每位评价员检验样品的重复次数。
- f) 评价小组成员的相关信息,如年龄、性别、评价经验以及其他一些重要信息(如评价员对刺激的熟悉度、健康状况、吸烟史、是否使用假牙以及最后进食时间等)。应避免靠姓名来区分评价员,防止某位熟悉评价小组的人员对评价结果给出个人判断。
- g) 数据量不大时,结果以表格形式给出,如表 B.1。表中列出检验的浓度系列和该浓度下评价员做出正确选择的次数。报告中给出阈值估计值,如个体阈值、组阈值和误差限。
- h) 若已知评价员的阈值分布情况,报告中给出评价员的个体阈值以及相关描述统计值,如均值、方差和数据不服从正态分布时采用的检验方法。
- i) 采用合并数据时,报告中给出合并数据的组阈值和误差限,以及斜率 b 和其误差限。请注意,评价员内部差异(较小的)和评价员之间的差异(较大的)可能会影响误差限的计算。

附录 A
(资料性)

给定精度下所需的评价员数量估计

由于不同个体对于同一个刺激的响应存在差异,因此依据本文件中程序进行阈值估计时具有一定的不确定性,表 A.1 给出了典型情况下达到预期精度(±标准偏差)的示例。

表 A.1 评价小组样本量和提供样品量下的获得的检验精度

序号	评价小组	目的	3-AFC 检验中的样品数	检测精度
1	1 人	内科医生希望知道某人对物质 X 是否具有嗅觉缺失症	8 个浓度,每个浓度检验 1 次 $n=8$	$\pm 200\% \sim \pm 500\%$
2	1 人	希望知道他/她对 1 瓶调配水中物质 X 的敏感度	6 个浓度,每个浓度检验 6 次 $n=36$	$\pm 50\% \sim \pm 100\%$
			8 个浓度,每个浓度检验 10 次 $n=80$	$\pm 20\% \sim \pm 50\%$
3	6 人小组	实验者希望比较两个小组的敏感性	6 个浓度,每个浓度检验 1 次 $n=48$	$\pm 100\% \sim \pm 300\%$
			8 个浓度,每个浓度检验 6 次 $n=384$	$\pm 20\% \sim \pm 50\%$
4	4 个小组,每个小组由挑选出的 8 名具有代表性的人员组成	市政工程师希望知道污染物质 X 的阈值水平	6 个浓度,每个浓度检验 1 次 $n=192$	$\pm 50\% \sim \pm 200\%$
			6 个浓度,每个浓度检验 6 次 $n=1\,152$	$\pm 20\% \sim \pm 50\%$
5		工程师希望确定总体 95% 概率下不能检测出物质 X 的浓度水平	上述检验,加上 4×8 人中最敏感的 25% 人的重复检验 $n \approx 1\,800$	$\pm 40\% \sim \pm 100\%$

3-AFC 检验精度与样品量间的关系和所测阈值类型无关。例如一个橘子汽水的生产商希望知道某种特定成分是否能以 95% 的概率检出,这种情况下所需待测样品数量约为 1 800。评价员需求数量很大程度上取决于评价小组的总体差异。若评价小组是由差异较大的群体构成(例如许多年轻人和老人),达到既定检验精度通常需要更多的评价员数量。

附录 B

(资料性)

示 例

B.1 采用 BET 法快速测定烟囱气体样品的嗅觉阈值

B.1.1 总则

本示例给出了采用 BET 法估计阈值近似值的快捷程序。最简单操作程序如下：9 名评价员组成的评价小组对 6 种给定浓度样品仅进行一次检验，评价员 4 和评价员 6 不继续扩展浓度进行检验。更准确的方法应该是继续检验，例如将样品稀释 3 倍、5 倍、27 倍及更高倍数，让评价员 6 继续检验，直到出现错误选择。检测精度过低时，可以重复整个检验程序。

B.1.2 检验

以无嗅空气为介质制备 6 种不同浓度的烟囱气体样品，每种浓度样品均与二个无嗅空气样品组成一组。浓度梯度按照 3 倍的比例递增。从参与检验的人群中随机选择 9 位评价员，每位评价员均按照从低浓度到高浓度的顺序依次对样品进行检验，从每种浓度组的三个样品中选出一个不同于其他两个的样品，检验结果如表 B.1 所示。

表 B.1 烟囱气体的嗅觉阈值

评价员	稀释比 (浓度递增→)						最优估计阈值(BET)	
	3 645	1 215	405	135	45	15	阈值	lg 阈值
1	0	+	+	0	+	+	78	1.89
2	+	0	+	+	+	+	701	2.85
3	0	+	0	0	+	+	78	1.89
4	0	0	0	0	+	0	9	0.95
5	+	0	0	+	+	+	234	2.37
6	+	+	+	+	+	+	6313	3.80
7	0	+	+	0	+	+	78	1.89
8	+	0	0	+	+	+	234	2.37
9	+	0	+	+	+	+	701	2.85
阈值对数值(lg 阈值)总和($\sum \lg$)							20.85	
BET 阈值平均值=几何平均值							209	
阈值对数值(lg 阈值)的标准偏差							0.81	
注：“0”表示评价员在三个样品中选择错误；“+”表示评价员在三个样品中选择正确。								

B.1.3 计算

每位评价员的 BET 值由评价错误的最高浓度值和其紧邻的判断正确的更高浓度值的几何平均值确定，见参考文献[4]。计算时，假定评价员 4 在接下来稀释比为 5 时做出正确选择，评价员 6 在稀释比为 10 935 时的选择错误。BET 结果的对数值总和为 20.85，平均值约为 2.32，这与平均阈值(稀释

比 209) 的反对数值接近。用阈值对数值的标准偏差来表征阈值的离散程度, 根据表 B.1 中最右列计算得出。数值 0.81 表明评价员间的评价结果存在较大差异, 但实际值可能比 0.81 更大。由于评价员 4 和评价员 6 并未继续扩展浓度系列进行检验, 因此浓度范围被缩小。

B.2 合并数据的组阈值计算——以鲑鱼体内的柴油污染阈值测定为例, 比较最大似然拟合法与最小二乘法拟合法

B.2.1 总则

本示例给出了最大似然法的工作原理, 以及采用电子表格软件确定阈值参数和界限的方法。将该方法应用于 18 个评价员采用 3-AFC 法对 9 个非等间距的浓度系列进行检验的评价结果数据集, 见 B.2.3。为便于比较, 同时采用最小二乘法对相同数据进行拟合, 见 B.2.4。

B.2.2 检验

测试样品为柴油, 检验暴露含有柴油水样中的鲑鱼是否存在潜在柴油污染, 见参考文献[9]。鲑鱼暴露在不同浓度柴油的水样中, 24 h 后收集这些鲑鱼, 18 个有经验的评价员组成评价小组进行评价, 结果与未经暴露的对照组进行比较。表 B.2 总结了电子表格软件中的公式, 表 B.3 列出了实验数据和用于 logistic 模型拟合的工作表。

表 B.2 电子表格中用于估计 logistic 模型参数和阈值估计误差限的公式

列	公式
B	@LN(A3)
E	+ D3/C3
F	(2/3)/(@EXP(G\$18*(F\$18-\$B3))+1)+1/3
G	+\$D3*@LN(F3)+(\$C3-\$D3)*@LN(1-F3)
H	(2/3)/(@EXP(I\$18*(H\$18-\$B3))+1)+1/3
I	+\$D3*@LN(H3)+(\$C3-\$D3)*@LN(1-H3)
J	(2/3)/(@EXP(K\$18*(J\$18-\$B3))+1)+1/3
K	+\$D3*@LN(J3)+(\$C3-\$D3)*@LN(1-J3)

表 B.3 采用电子表格软件对数据进行 logistic 模型拟合以及估计阈值误差限的示例

	A	B	C	D	E	F	G	H	I	J	K
1						最优模型		下限		上限	
2	C mL/m ³	ln(C)	n	r	P _{obs}	P _{est}	似然	P _{est}	似然	P _{est}	似然
3	0.010	-4.61	18	5	0.278	0.366	-10.951	0.414	-11.361	0.349	-10.844
4	0.028	-3.56	18	9	0.500	0.446	-12.583	0.538	-12.528	0.394	-12.892
5	0.032	-3.44	18	4	0.222	0.462	-11.762	0.558	-13.759	0.404	-10.865
6	0.060	-2.81	18	11	0.611	0.568	-12.097	0.671	-12.170	0.478	-12.672
7	0.095	-2.35	18	14	0.778	0.667	-10.072	0.755	-9.561	0.561	-11.382
8	0.285	-1.26	18	16	0.889	0.872	-6.304	0.903	-6.298	0.798	-6.812
9	0.324	-1.13	18	17	0.944	0.888	-4.206	0.914	-3.982	0.821	-5.067
10	0.673	-0.396	18	16	0.889	0.952	-6.865	0.959	-7.060	0.920	-6.389

表 B.3 采用电子表格软件对数据进行 logistic 模型拟合以及估计阈值误差限的示例 (续)

	A	B	C	D	E	F	G	H	I	J	K
11	0.992	-0.008 03	18	17	0.944	0.970	-4.031	0.973	-4.075	0.951	-3.869
12							-78.871		-80.794		-80.792
13											
14								偏差	3.841		3.841
15											
16											
17						$\ln(r)$	b	x	b	x	b
18						-2.348 9	1.311	-2.835 4	1.117	-1.867 3	1.357
19					P						
20				浓度单位		0.095 44		0.058 70		0.154 5	
21											
22					修正的阈值估计误差限			-2.835 4		-1.867 3	

B.2.3 基于最大似然法的结果和计算

最大似然法进行模型参数估计的原理是获得能使似然函数[式(B.1)]达到最大值时的参数值。电子表格软件含有能计算函数最大值或最小值的功能函数,能方便运用 3-ALPC 法检验数据进行 logistic 模型拟合。电子表格程序被广泛安装在实验室的个人电脑中,且使用时比统计包具备更多的优点。操作者能很容易理解和熟悉电子表格软件的操作过程,且无须掌握编程语言相关的知识或掌握复杂统计包的工作原理。用合理数量的观测数据(约 50 个以上,分布在阈值周围)即可顺利地获得参数和阈值范围。与统计包或其他专用统计程序相比,电子表格软件的优势在于拟合程序原理更容易被操作者理解,尤其当拟合数据未能达到理想的数据量时,优势更加明显。

不同电子表格程序中,求解函数最大值或最小值命令的命名并不相同。本示例中使用的是 QuattroPro³⁾包中的 Optimizer 函数,等同于 Excel¹⁾和 Lotus1-2-3²⁾中的 Solver 函数。

表 B.3 中,A 列~D 列是检验数据,A 列是水中的柴油浓度(C),单位为毫升每立方米(mL/m³);B 列是浓度对数值,表中采用自然对数[ln(C)](也能使用其他值为基数的对数)。刺激浓度无需为等距分布的几何序列或其他特殊间隔的浓度序列。为方便起见,通常按照浓度高低进行排序。

C 列和 D 列分别是每种浓度样品的检测数量(n)和该浓度下做出正确选择的数量(r)。每一种浓度检验时评价员数量无需一致。E 列是测试中做出正确选择次数所占的比例(P_{est}),该值在后面的计算中并未出现。

F 列和 G 列是式(3)中 logistic 函数拟合的计算数据。logistic 模型由参数 b 和 t 确定,通过给定的参数值,即能计算出正确选择的期望比例。假定一个给定参数值的模型,似然值等于获得的观测比例的乘积。所有数据的联合概率等于每一种浓度概率的乘积。当联合概率为对数似然值之和时,使用对数似然值更方便。概率对数值的计算按式(B.1):

$$\sum_{i=1}^{i=N} r_i \ln(P_{est})_i + (n_i - r_i) \ln(1 - P_{est})_i \dots\dots\dots (B.1)$$

3) 文中所提及的电子制表软件是目前比较适用的商业化产品。该信息只是给本文件的使用者提供一些便利,并不涉及推荐认可产品。

式中：

- N ——数据集(或浓度)的数量；
 n ——每种浓度重复检验的数量；
 r ——正确选择的次数；
 P_{est} ——估计的比例。

似然值由 G 列计算获得。表 B.3 中第 12 行是似然值的总和，即联合概率的对数似然值。按照与 F 列和 G 列类似的公式，在 H 列和 I 列、J 列和 K 列中分别得到 t 和 b 值的下限值和上限值。

本示例中，把数据输入到第 3 行~第 11 行中。引用表 B.2 中 F 列公式，在单元格 F18 和 G18 中分别输入阈值和斜率参数的初始估计值。程序对于初步估计是非常可靠的，初始值可以远离近似分布在阈值附近数据集的最优值，也不需要特别接近最终值。阈值的初始估计值可选取浓度范围的中间值。B 列取浓度的自然对数时，斜率的初值为 1；若取以 10 为底的对数时，斜率初值为 0.5。优化函数或者具有相同功能的函数通常来自工具箱或菜单。函数需要给定变量单元的地址(这些单元格包含的数据随着似然函数最大值的变化而改变，如本例的 F18 和 G18)和结果单元的地址(该单元为最大似然值，如 G12)。

函数有许多选项，其中很多选项设置为缺省值。检查函数是否使结果单元格中数值最大化。在本例中没有设置任何限制条件。函数停止迭代的标准是达到最大迭代次数和变量的变化量达到最小，允许操作者根据缺省值重新设置迭代次数和最小变化量。对于提供多种求解方法的电子表格，操作员宜选择用于非线性优化的广义简约梯度(Generalized reduced gradient, GRG)法或其他用于非线性问题光滑化的类似等效方法。

开始迭代后，通过个人计算机可在几秒钟内得到结果，但也有可能因为计算结果收敛太慢而停止迭代。单元格 F18 和单元格 G18 中是所需的 logistic 模型的参数估计值。

此时， $\ln(t) = -2.35$ ， $t = 0.095 \text{ mL/m}^3$ ， $b = 1.31$ 。

由这两个参数确定的 logistic 模型称为最大似然 logistic 模型，即通常所述的最优模型。其他数值参数确定的模型给出的似然值较小，模型之间的似然差异定义为偏差：

$$\text{偏差} = -2(l_1 - l_2)$$

式中， l_1 和 l_2 分别是模型 1 和模型 2 的对数似然。本例中， l_1 和 l_2 分别代表最优模型(模型 1)和另外一个模型(模型 2)的确定值。模型之间的显著性差异由卡方检验进行衡量，自由度取适当的一个值。该关系用于确定估计阈的误差限，目的是为了找到这样的一个模型，即其 t 值计算的偏差等于在确定的显著性水平和自由度为 1 时的卡方值。当自由度为 1，显著性水平 p 为 0.05 时，偏差为 3.841。

如果完全指定了适当的列，单元格 F3 和单元格 G3 中使用的公式可相应的复制到 H 列~K 列中。单元格 I14 的计算式为 $+2 \times (\$G12 - I12)$ ，表示拟合模型的下限和最优模型之间的偏差。把这个公式对应复制到单元格 K14 内。

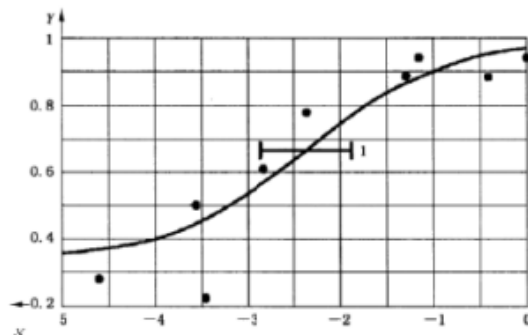
通过一次迭代过程得到误差限，在开始的时候有必要给定一个初始的误差限估计值。误差限的大小由数据集的一些特性所决定，其中最重要的一个特性就是检测总数。表 B.4 列举了阈值估计中步长误差限初始估计值与检测总数间的关系。本例中，检测次数为 162，对应的初始估计值的下限和上限分别为 $-2.35 + (-0.49) = -2.84$ 和 $-2.35 + 0.56 = -1.79$ 。这些值被输入到单元格 H18 和单元格 J18。然后，再次运行优化函数，但只使用包含 b 值的单元格，单元格 I18 或单元格 K18 作为变量。

表 B.4 阈值估计中步长误差限初始估计

检测总数	40	60	80	100	120	160	200	>200
步长下限	-1.02	-0.81	-0.76	-0.61	-0.58	-0.49	-0.45	-0.40
步长上限	1.39	1.05	0.84	0.78	0.69	0.56	0.49	0.45

单元格 I12 和单元格 K12 是目标单元格,为似然值的总和。如果偏差不充分接近 3.84 这一目标值,修正所估计的边界值。在数据集中 50 个以上的观测结果分布在阈值周围,这个偏差大致与距离阈值的步长平方根呈线性相关。修正的估计值通过简单的比值就可计算获得。单元格 H22 为计算边界的修正估计值式: $\text{@SQRT}(3.84/\text{I14}) * (\text{H18} - \$\text{F18}) + \$\text{F18}$,该公式被复制到单元格 J22。把修正值结果输入单元格 H18 和单元格 J18(或在此编辑相应的值),再一次运行优化函数。当获得的偏差接近期望值时,迭代结束。在表 B.3 的示例中,三次迭代即足够。

操作员使用电子表格软件的绘图工具绘制测试的数据和 logistic 拟合曲线,如图 B.1 所示。使用表 B.3 中第 F 列的数值绘制 logistic 曲线。充分利用电子表格软件中的函数构造一组等间隔分布的浓度对数数据,并通过表 B.3 中第 G 列的公式计算概率和参数估计值即可得到一条更光滑的曲线。



标引序号说明:

X —— 浓度的自然对数;

Y —— 正确选择概率;

1 —— 误差限;

● —— 检测值;

—— 拟合曲线。

图 B.1 合并数据的组阈值——柴油污染的鲑鱼

参数拟合,尤其是寻找误差限的困难在于对非正常数据集的处理经验。非正常数据集包括小规模数据集、数据偏在阈值一侧的数据集以及检测数据比值随着浓度增加明显偏离平滑的单调递增曲线。该情况时优化函数通常会无法收敛,给出错误提示后退出。通常, b 值过高、过低甚至呈负值时程序会退出(这里的检验变量值通常随着浓度升高而增大,例如模型要求的觉察概率随着刺激浓度增加而增大, b 值呈正值。当测试水及气体中不良气味或异味时,通常是稀释样品使其检验浓度与稀释程度相关联。此时检验变量值随着刺激浓度递减反而增大, b 值呈现负值)。当拟合 logistic 或确定边界范围遇到困难时,对实验者而言最好的解决办法是获取更多的数据,而操作员可重复拟合程序,包括调整优化函数的选项来限制 b 值。对 b 值的约束可以是取比 0 大点或者小点的一个数值。若如表 B.3 中第 B 列那样取自然对数时, b 值可以小于 2;若取 10 为底的对数时, b 值可以在 1 左右。如果刺激浓度超过 20 倍,觉察概率范围在 0.05~0.95,对应的 b 值上限受到约束。若未能找到一个适合的解决办法时,操作员可指定一个 b 值进行阈值拟合。在确定 b 值范围和极大似然函数后,该拟合程序可反复进行。检验极大似然值和 b 值的变化能够确定是否存在一个最优的目标值。

用单元格 H22 定义的公式计算非正常数据集时,边界误差估计值有可能收敛很慢或者无法收敛。可在进行迭代时,在电子数据表格的相应列中记录 t 值和相关的偏差。在偏差为 3.84 时,经过三四次迭代,偏差与阈值估计值之间的距离可绘制成图,从图上直观地用外推或内插得到步长,即偏差减去

3.84 或者找到对应偏差为 0 时的步长。电子表格软件的回归函数也能给出偏差减少值与二次多项式步长的回归关系,求解该公式可以在偏差减少值为 0 时得到 t 值。修订的边界估计值在下次迭代计算时使用。一个简单的策略是用新的修订值 x 和对应的偏差去替换存在的偏差和阈值向量,并重新计算表达式,提供一个修正值。优化函数中的 b 值需要严格限制,以防超出正常范围。重复这一过程,如果存在一个合理的边界值时,计算很快收敛。检查偏差与阈值之间的距离变化可以提示是否能够快速找到一个边界值。

B.2.4 用线性回归法计算

该方法的原理是先采用 6.2.2 中描述的 Logit 变换,再使用最小二乘法的计算机统计软件(如 SAS⁴⁾等统计软件)确定与数据最匹配的直线。数据命令见表 B.3,结果输出如图 B.2。回归模型仅适用于数据对数值在 $-1.75 \sim 1.75$ 范围内的数据点。超出此范围的值太过极端,无法进行模型拟合和曲线确定,如 6.2.2 所述。

4) 该信息只是给本文件的使用者提供一些便利,并不涉及推荐认可产品。除 SAS 外,SPSS、S-PLUS 和 SYSTAT 软件也同样适用。

```

TITLE "Logistic Regression of Threshold Data";
DATA INPUT;
    INPUT CONC R N;          /* Input Data */
    P=MAX(R/N, 1/3);          /* Compute P(C) */
    LOGIT=LOG((P-1/3)/(1-P)); /* Perform Logit Transformation */
    LOG_C=LOG(CONC);           /* Convert to LOG Concentration */
CARDS;
.010 5 18
.028 9 18
.032 4 18
.060 11 18
.095 14 18
.285 16 18
.324 17 18
.673 16 18
.992 17 18
;
RUN;
DATA TRIMMED;                /* Trim data of extreme values */
    SET INPUT;                /* as per Section 6.2.1 */
    IF LOGIT GE -1,75 AND LOGIT LE 1,75;
RUN;
PROC REG DATA=TRIMMED OUTEST=EST; /* Fit the regression model */
    MODEL LOGIT=LOG_C;
RUN;
DATA EST;
    SET EST;
    LOG_T = -INTERCEP/LOG_C;    /* Compute Threshold value */
    THRESHLD = EXP(LOG_T);

```

图 B.2 Logit 转化数据进行线性回归分析估计阈值示例

B.3 比较两位评价员的个体阈值——以水中柴油为例

B.3.1 目的

目的是检验两位评价员在对某种物质的感知上是否存在差异。本示例给出了最大似然法分层统计模型的应用。

B.3.2 检验

将一种化学物质溶于水，目的是测定和比较两位评价员对某种气味的觉察阈值。

B.3.3 结果和计算

该方法建立了较复杂的分层统计模型，检验数据符合模型的似然。当越多的参数增加到模型中时，似然也增大。随着参数增加而引起的似然函数变化，以偏差表示。偏差用于检验这种新模型是否能引起似然的显著增加。当 3-AFC 数据符合 logistic 模型时，模型中增加的参数就是 t 值和 b 值。

表 B.5 列出了检验数据及计算结果，表 B.6 中列出了三种统计模型的拟合结果。实验中浓度梯度以及每个浓度检验数量的设置均采用相同策略，即先在一个广泛浓度范围内初步小试，然后再在接近个体阈值的浓度范围内进行多次重复。

如 B.2 中的示例，重要的是确保在提供了多种求解方法的电子表格中，操作员宜选择用于非线性优化的广义简约梯度 (GRG) 法或其他用于非线性问题光滑化的类似等效方法。

表 B.5 比较两位评价员响应的统计模型计算

C mL/m ³	ln(C)	n	r	P _{obs}	模型 1		模型 2		模型 3	
					P _{est}	对数似然	P _{est}	对数似然	P _{est}	对数似然
评价员 1										
0.001 5	-6.50	3	0	0.000	0.337	-1.23	0.334	-1.22	0.334	-1.22
0.004 0	-5.52	3	0	0.000	0.347	-1.28	0.335	-1.23	0.335	-1.23
0.004 6	-5.38	3	0	0.000	0.350	-1.29	0.336	-1.23	0.336	-1.23
0.013 9	-4.28	3	0	0.000	0.404	-1.55	0.348	-1.29	0.349	-1.29
0.020 0	-3.91	5	1	0.200	0.442	-3.15	0.360	-2.81	0.360	-2.81
0.025 0	-3.69	7	1	0.143	0.473	-4.59	0.371	-3.78	0.372	-3.78
0.041 7	-3.18	3	1	0.333	0.565	-2.24	0.415	-1.95	0.415	-1.95
0.100 0	-2.30	13	8	0.615	0.759	-9.33	0.578	-8.70	0.578	-8.70
0.125 0	-2.08	3	3	1.000	0.804	-0.65	0.636	-1.36	0.636	-1.36
0.500 0	-0.693	5	4	0.800	0.961	-3.41	0.926	-2.91	0.925	-2.90
2.500 0	0.916	2	2	1.000	0.995	-0.01	0.994	-0.01	0.994	-0.01
12.500 0	2.53	2	2	1.000	1.000	0.00	1.000	0.00	1.000	0.00
评价员 2										
0.001 5	-6.50	3	0	0.000	0.337	-1.23	0.344	-1.27	0.344	-1.27
0.004 0	-5.52	3	2	0.667	0.347	-2.54	0.384	-2.40	0.383	-2.40
0.004 6	-5.38	3	1	0.333	0.350	-1.91	0.395	-1.93	0.394	-1.93
0.012 5	-4.38	8	4	0.500	0.395	-5.73	0.560	-5.60	0.559	-5.60
0.013 9	-4.28	3	2	0.667	0.404	-2.33	0.586	-1.95	0.585	-1.95
0.020 0	-3.91	5	4	0.800	0.442	-3.85	0.682	-2.68	0.682	-2.68

表 B.5 比较两位评价员响应的统计模型计算（续）

C mL/m ³	ln(C)	n	r	P _{obs}	模型 1		模型 2		模型 3	
					P _{est}	对数似然	P _{est}	对数似然	P _{est}	对数似然
0.041 7	-3.18	3	2	0.667	0.565	-1.97	0.855	-2.25	0.856	-2.25
0.050 0	-3.00	7	6	0.857	0.604	-3.95	0.886	-2.90	0.886	-2.90
0.100 0	-2.30	5	5	1.000	0.759	-1.38	0.958	-0.22	0.958	-0.21
0.125 0	-2.08	3	3	1.000	0.804	-0.65	0.970	-0.09	0.970	-0.09
0.500 0	-0.693	5	5	1.000	0.961	-0.20	0.997	-0.02	0.997	-0.02
2.500 0	0.916	2	2	1.000	0.995	-0.01	1.000	0.00	1.000	0.00
12.500 0	2.53	2	2	1.000	1.000	0.00	1.000	0.00	1.000	0.00
总和						-54.48		-47.76		-47.76

模型 1：该模型中，只有一个 t 值和一个 b 值用于估计表 B.5 中 P_{est} 列的值。 P_{est} 列中每个单元格的计算公式均涉及包含 t 和 b 的两个单元格，这两个单元格构成电子表格软件中最大化程序的变量单元格。目标单元格中列出了两位评价员通过最大化程序得到对数似然值的和。该函数的运行结果是给出 t 和 b 的最大似然估计值。模型中 t 和 b 的值如表 B.6 所示。

表 B.6 模型分析表

模型	参数	评价员		对数似然	比较	偏差	D.F.	χ^2 检验的 P 值
		1	2					
1	t	-2.72		-54.48				
	b	1.37						
2	t	-1.97	-3.97	-47.76	1 与 2	13.43	1	0.001 2
	b	1.62						
3	t	-1.97	-3.97	-47.76	2 与 3	0.00	1	
	b	1.63	1.62					

模型 2：数据检查结果表明两位评价员的阈值并不相同。模型 2 利用评价员各自的 t 值，但只有一个 b 值。表 B.5 中模型 2 P_{est} 列中的值就是采用评价员各自的 t 值和共同的 b 值计算得到。包含有两个 t 值和一个 b 值的三个单元格共同构成最大化程序的变量单元格，目标单元格是该系列数据对数似然值的总和。函数运算结果给出了该模型 t 值和 b 值的最优值。此时模型数据拟合度较好，且对数似然值较小。模型 2 和模型 1 的参数值和偏差如表 B.6 所示。

模型 3：模型 3 是模型 2 的扩展，包含两位评价员各自独立的 t 值和 b 值，即两个 t 值和两个 b 值，共计四个变量单元。而且，目标单元格也包含上述数据对数似然值的总和。

B.3.4 结论和推断

表 B.6 列出了模型参数和偏差的统计结果。模型每增加一个参数，自由度就会减一个。三个模型中，两两模型之间的偏差可采用一个自由度的卡方检验。模型 1 和模型 2 之间的偏差非常显著，概率为 0.001 2。继续增加第三个参数，即第二个 b 值后，构成的模型 3 的拟合精度并未提高，但模型 3 与模型 2 之间偏差变为 0。

由此推断，两位评价员对柴油样品气味的觉察阈值存在显著差异，但对强度增加的敏感性并未表现出差异。

B.4 两种物质的阈值比较——水中的 α -萘烯和 β -萘烯

B.4.1 总则

B.3 中的应用模型同样适用于两种物质的阈值比较。多位评价员的数据合并后得到每种物质的组阈值数据。

B.4.2 检验

检验目的是测量和比较 α -萘烯和 β -萘烯的感官特性。评价小组由 24 名有经验的评价员组成,每种物质设置 10 个浓度水平,采用 3-AFC 法对每个浓度进行检验。

B.4.3 结果和计算

检验数据和电子表格软件计算如表 B.7 所示,表 B.8 中列出模型参数 t 和 b 的值及偏差的统计结果。

表 B.7 不同统计模型在计算 α -萘烯和 β -萘烯阈值上的应用

C mL/m ³	ln(C)	n	r	P _{obs}	模型 1		模型 2		模型 3	
					P _{est}	对数似然	P _{est}	对数似然	P _{est}	对数似然
α-萘烯										
0.004	-5.52	24	11	0.458	0.371	-16.93	0.387	-16.80	0.412	-16.66
0.020	-3.91	24	8	0.333	0.408	-15.56	0.444	-15.89	0.473	-16.24
0.040	-3.22	24	13	0.542	0.432	-17.13	0.480	-16.73	0.508	-16.61
0.100	-2.30	24	14	0.583	0.473	-16.89	0.540	-16.39	0.563	-16.32
0.200	-1.61	24	16	0.667	0.511	-16.46	0.593	-15.55	0.609	-15.45
0.500	-0.693	24	15	0.625	0.571	-16.02	0.669	-15.98	0.673	-16.00
1.000	0.00	24	20	0.833	0.622	-13.40	0.727	-11.57	0.722	-11.64
2.500	0.916	24	17	0.708	0.692	-14.50	0.797	-15.02	0.782	-14.85
5.000	1.61	24	19	0.792	0.743	-12.44	0.843	-12.50	0.823	-12.36
25.000	3.22	24	22	0.917	0.847	-7.41	0.920	-6.89	0.897	-6.94
β-萘烯										
0.004	-5.52	24	11	0.458	0.371	-16.93	0.349	-17.16	0.334	-17.35
0.020	-3.91	24	10	0.417	0.408	-16.30	0.368	-16.42	0.336	-16.64
0.040	-3.22	24	12	0.500	0.432	-16.86	0.381	-17.33	0.338	-17.96
0.100	-2.30	24	8	0.333	0.473	-16.24	0.407	-15.55	0.346	-15.28
0.200	-1.61	24	11	0.458	0.511	-16.69	0.433	-16.58	0.358	-17.06
0.500	-0.693	24	9	0.375	0.571	-17.74	0.479	-16.40	0.393	-15.89
1.000	0.00	24	8	0.333	0.622	-19.35	0.523	-17.02	0.445	-15.90
2.500	0.916	24	13	0.542	0.692	-17.73	0.591	-16.67	0.558	-16.57
5.000	1.61	24	18	0.750	0.743	-13.50	0.648	-14.07	0.672	-13.85
25.000	3.22	24	21	0.875	0.847	-9.12	0.779	-9.77	0.894	-9.09
总和						-307.21		-300.32		-298.63

表 B.8 不同拟合模型计算 α -萘烯和 β -萘烯阈值的统计结果分析

模型	参数	α -萘烯	β -萘烯	对数似然	模型比较	偏差	自由度	χ^2 检验 p 值
1	t	-0.591		-307.2				
	b	0.459						
2	t	-0.723	-1.831	-300.3	1 与 2	13.79	1	0.000 20
	b	0.506						
3	t	-0.787	-1.581	-298.3	1 与 3	17.16	2	0.000 19
	b	0.426	1.016		2 与 3	3.37	1	0.066

B.4.4 结论和推断

模型 2 对数据的拟合性显著优于模型 1，证实了两种异构体的阈值存在差异的假设。模型 3 有两个 t 值和两个 b 值，拟合度和对数似然值均更大。但模型 2 和模型 3 之间的偏差为 3.37， χ^2 检验结果 p 值为 0.066，与常用的 $p=0.05$ 水平相比，并不显著。但拟合曲线的斜率存在显著差异，需要进一步展开研究。对于 β -萘烯，四个低浓度数据点均靠近 logistic 曲线的渐近线，对曲线拟合并不提供作用； α -萘烯的浓度较均匀地涵盖了渐进线之间的浓度范围，对曲线斜率进行较好的估计。

此外，模型 2 中的 b 值与模型 3 中 α -萘烯的 b 值更接近，表明模型 3 中 β -萘烯的 b 值估计精度低于异构体 α -萘烯。对于 β -萘烯，建议在该检验选取的最高浓度之上再设置两个或更多个浓度点，可在 $5 \text{ mL/m}^3 - 25 \text{ mL/m}^3$ 间选取一个浓度点，进行重复评价。

参 考 文 献

- [1] ISO 6658 Sensory analysis—Methodology—General guidance
 - [2]ⁱ ISO 8586 Sensory analysis—General guidelines for the selection, training and monitoring of selected assessors and expert sensory assessors
 - [3] ISO 8589 Sensory analysis—General guidance for the design of test rooms
 - [4] ASTM E679-91(1997) Standard Practice for Determination of Odor and Taste Thresholds by a Forced-Choice Ascending Series Method of Limits.
 - [5] EN 13725 Air Quality—Determination of odour concentration by dynamic olfactometry
 - [6] Aitkin M., Anderson D., Francis B., Hinde J. Statistical Modelling in GLIM. Clarendon, Oxford, 1989.
 - [7] Amoore J.E. Specific anosmia and the concept of primary odors. Chemical Senses and Flavor, 1977, 2 pp.267-281.
 - [8] Buttery R.G. Qualitative and sensory aspects of flavor of tomato and other vegetables and fruits. In: Flavor Science; Sensible Principles and Techniques, Acree T.E. and Teranishi R., Chapter 8, pp.277-278, ACS Professional Reference Book, American Chemical Society, Washington, DC, 1993.
 - [9] Davis H.K., Geelhoed E.N., MacRae, A.W., Howgate P. Sensory analysis of trout tainted by diesel fuel in ambient water. Water Sci. Technol. 1992, 25 (2) pp.11-18.
 - [10] Devos M., Patte F., Rouault J., Laffort P., van Gemert L.J. Standardized Human Olfactory Thresholds. IRL Press, Oxford, 1990.
 - [11] Doty R.L., Gregor T., Settle R.G. Influences of intertrial interval and sniff bottle volume on the phenyl ethyl alcohol olfactory detection threshold. Chem. Senses. 1986, 11 pp.259-264.
 - [12] Doty R.L., Deems D.A., Frye R., Pelberg R., Shapiro A. Olfactory sensitivity, nasal resistance, and autonomic function in the multiple chemical sensitivities (MCS) syndrome. Arch. Otolaryngol. Head Neck Surg. 1988, 114 pp.1422-1427. [Description of the University of Pennsylvania's Dynamic Air-Dilution Olfactometer].
 - [13] Fazzalari F.A. Compilation of Odor and Taste Threshold Values Data. American Society for Testing and Materials, Philadelphia, 1978.
 - [14] Guadagni D.G., Buttery R.G. Odor threshold of 2,3,6-trichloroanisole in water. J. Food Sci. 1978, 43 pp.1346-1347.
 - [15] van Harreveld A., Ph., Heeres P., Harssema H. A review of 20 years of standardization of odor concentration measurement by dynamic olfactometry in Europe. J. Amer. Waste Management Assoc., 49 (5).
 - [16] Macmillan N.A., Creelman C.D. Detection Theory, A User's Guide. Cambridge University Press, 2005, pp.395.
 - [17] van Gemert L.J., Nettenbreijer A.H. Compilation of Odour Threshold Values in Air and Water. Central Institute for Nutrition and Food Research TNO, Zeist, Netherlands, 1977. Supplement V, 1984.
-