软件工程统计方法

参数假设检验

陈振宇

南京大学软件学院

Email:zychen@software.nju.edu.cn

Homepage:software.nju.edu.cn/zychen

内 容 提 纲

- 统计学导论
- 描述统计
- 概率计算基础
- 随机变量及其分布
- 统计量及其抽样分布
- 参数估计
- 参数假设检验
- 非参数假设检验
- 方差分析
- 回归分析

本 节 内 容

- 基本概念
  - 假设建立
  - 两类错误
  - 显著性水平
  - 基本步骤
- 总体均值检验
- 总体方差检验
- 置信区间与假设检验
- 样本容量与统计功效分析

1 假设检验

根据样本的信息检验关于总体的某个假设是否正确进行判断，这类问题称作假设检验问题。

罐装可乐的容量按标准应在330毫升。生产流水线上罐装可乐不断地封装，然后装箱外运。怎么知道这批罐装可乐的容量是否合格呢？

假设检验

每隔一段时间进行抽检。如每隔2小时，抽查n罐，得n个容量的值$X_1, \cdots, X_n$，根据这些值来判断生产是否正常. 如发现不正常，就应停产检查，排除故障，然后再生产；如没有问题，就继续按规定时间再抽样，以此监督生产，保证质量. 假设抽到的9罐容量如下：

$$340, 323, 339, 335, 331, 327, 329, 342, 337$$

在正常生产条件下，由于种种随机因素的影响，每罐可乐的容量应在330毫升上下波动. 更加生产经验，我们假定每罐容量服从正态分布$X \sim \mathbb{N}(\mu, \sigma^2)$.

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

假设检验

假设检验(Hypothesis Testing)是在样本的基础上对总体的某种结论作出判断的一种方法。

- 新药的疗效是否更有效?
- 新工具是否帮助我们提高了生产效率?
- 新产品的性能是否更好?
- 新生产性的稳定性是否更高?

假设检验是统计推断的重要组成部分, 分为参数假设检验和非参数假设检验:

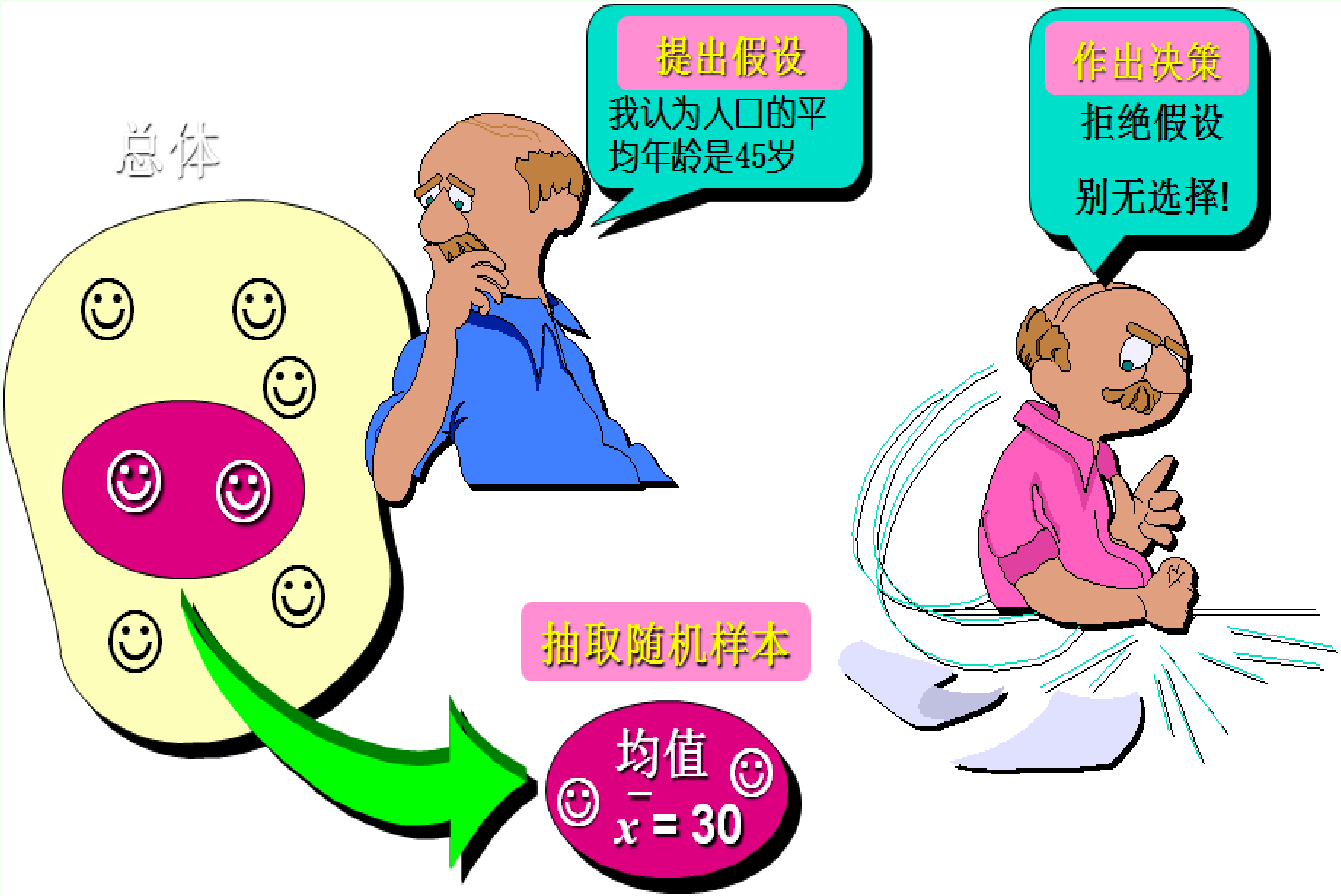
- 对总体分布中未知参数的假设检验称为参数假设检验。
- 对未知分布函数的类型或其某些特征提出的假设称为非参数假设检验。

假设检验

- 什么是假设?
  - 对总体参数的具体数值所作的陈述。
  - 总体参数包括总体均值、比例、方差等。
  - 统计分析之前必须陈述。

- 先对总体的参数(或分布形式)提出某种假设，然后利用样本信息判断假设是否成立的过程。
- 逻辑上运用反证法，统计上依据小概率原理。

假 设 检 验



思考：假如抽取的样本均值为42, 你又会下什么结论呢？

假设检验

我们先以一个例子来说明问题

Example 1 某车间用一台包装机包装葡萄糖, 包得的袋装糖重是一个随机变量, 它服从正态分布, 当机器正常时, 其均值为$\mu_0 = 0.5(kg)$, 标准差是$\sigma_0 = 0.015(kg)$. 某日开工后为检验包装机是否正常, 随机地抽取它所包装的9袋糖, 称重为

$$0.497, 0.506, 0.518, 0.524, 0.498, 0.511, 0.520, 0.515, 0.512$$

问机器是否正常?

为了简化问题, 同时根据长期实践表明标准差比较稳定, 我们设$\sigma = 0.015$. 那么我们根据样本判断$\mu = \mu_0$还是$\mu \neq \mu_0$, $\mu_0 = 0.5$. 为此我们提出两个假设:

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

然后我们根据一些合理法则去判断是接受H_0(即拒绝H_1), 还是接受H_1(即拒绝H_0).

假设检验

$\overline{X}$是μ的无偏估计. 因此如果假设H_0为真, $\overline{X}$的观察值$\overline{x}$与μ不应偏差太大. 考虑到H_0为真时, $\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$. 我们可以选定一个正数$k$, 若$\frac{|\overline{x}-\mu_0|}{\sigma/\sqrt{n}} \geq k$就拒绝假设$H_0$, 否则就接受假设$H_0$.

根据样本做决策就有可能误判, 误判的概率, 即$P\{$当H_0为真时拒绝$H_0\}$或$P_{\mu_0}\{$拒绝$H_0\}$多大?

假设检验

为了控制这类错误, 我们给定一个较小的正数α, 使得

$$P_{\mu_0}\{拒绝H_0\} \leq \alpha$$

那么如何确定α? 由$Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$, 可得$k = z_{\alpha/2}$. 则当

$$\frac{|\overline{x}-\mu_0|}{\sigma/\sqrt{n}} \geq z_{\alpha/2}$$

拒绝假设H_0; 当

$$\frac{|\overline{x}-\mu_0|}{\sigma/\sqrt{n}} < z_{\alpha/2}$$

接受假设H_0。
本例中算得$\overline{x} = 0.511$, $z_{0.025} = 1.96$, 则有

$$\frac{|0.511-0.5|}{0.015/\sqrt{9}} = 2.2 > 1.96$$

所以拒绝H_0, 认为机器今天不正常运行。

2 基本概念

H_0为**原假设(null hypothesis)**, H_1为**备择假设(alternative hypothesis)**。
例如:

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

- 关于原假设
  - 研究者想收集证据予以反对的假设。
  - 原假设又称为"0假设"。
  - 表示为H_0, 为一个算术表达式, 包含比较符$=, \geq, \leq$。
- 关于备择假设
  - 研究者想收集证据予以支持的假设。
  - 表示为H_1, 为一个算术表达式, 包含比较符$\neq, <, >$。
  - 备择假设和原假设互补。
- 原假设与备择假设的选择取决于我们对问题的态度, 一般而言我们希望从子样观测值取得某一陈述的强力支持, 这一陈述的否定作为原假设, 而陈述本身作为备择假设。
- 通常先确定备择假设再确定原假设。

原 理 与 法 则

假设检验的基本原理:

- 在"原假设"成立的条件下, 分析抽样所发生的事件是否是一个小概率事件. 若是, 根据"小概率原理"就有充分的理由拒绝原假设; 若不是, 则没有充足理由拒绝原假设, 只能不拒绝原假设(注意不称为"接受原假设")。
- 统计检验的结论应为"不拒绝" 而不为"接受"。
- 因研究目的不同, 对同一问题可能提出不同的假设(也可能得出不同的结论)。

假设检验的检验法则:
把子样空间分为两个互不相交的子集C及$\overline{C}$, 当子样$X_1, \cdots, X_n$的观测值$(x_1, \cdots, x_n) \in C$, 则拒绝原假设; 若$(x_1, \cdots, x_n) \in \overline{C}$则不拒绝原假设, 称子集$C$为拒绝域.

两类错误

- 第一类错误(Type I error): 拒真(false positive)或α错误. 当H_0为真时, 子样观测值落入拒绝域, 发生的概率记为α, 即

$$P\{拒绝H_0|H_0为真\}$$

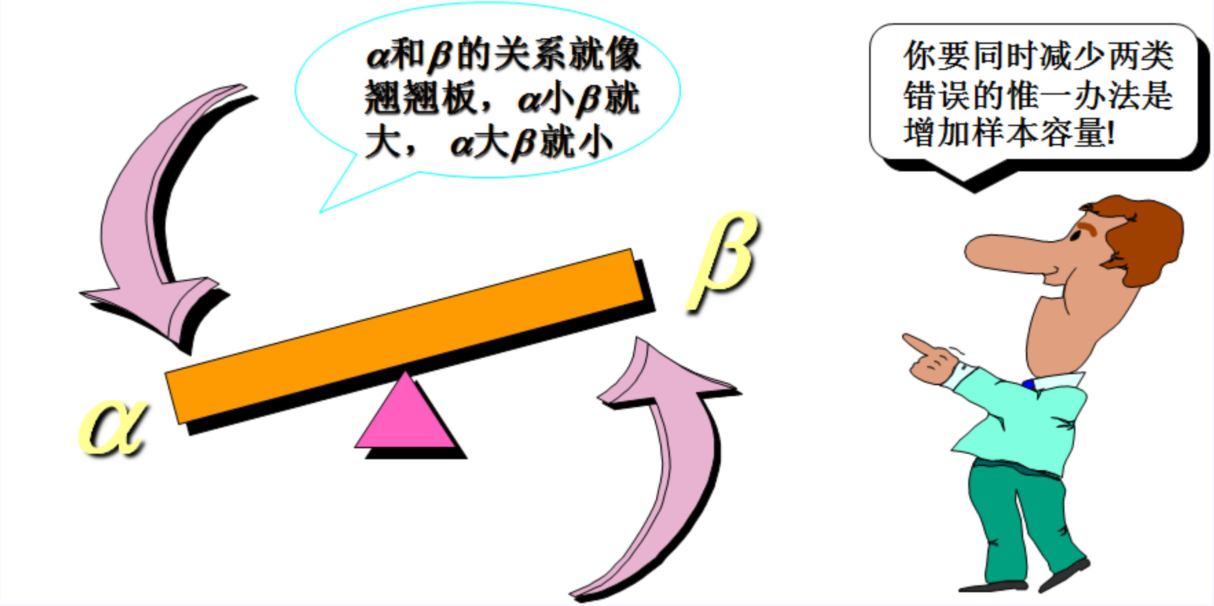
- 第二类错误(Type II error): 受伪(false negative)或β错误.
当H_0不真时, 根据检验法则我们却接受了, 其发生概率常记为β, 即

$$P\{接受H_0|H_0不真\}$$

	H_0为真	H_0不真
拒绝H_0	I类错误(α)	正确
接受H_0	正确	II类错误(β)

两 类 错 误

样本容量固定时, 若犯I类错误的概率减小, 则犯II类错误的概率将增大。在增大样本容量时可使犯两类错误的概率都减小。但样本容量太大将增加抽样成本, 甚至是不可行的。



两类错误

- 对于一个给定的样本，如果犯第I类错误的代价比犯第II类错误的代价相对较高，则将犯第I类错误的概率定得低些较为合理。反之，如果犯第I类错误的代价比犯第II类错误的代价相对较低，则将犯第I类错误的概率定得高些。
- 发生哪一类错误的后果更为严重，就应该首要控制哪类错误发生的概率。但由于犯第I类错误的概率是可以由研究者控制的，因此在假设检验中，人们往往先控制第I类错误的发生概率。
- 若只考虑对犯第I类错误的概率加以控制，而不考虑第II类错误的概率的检验，称为**显著性检验**。

影响β错误的因素：

- 显著性水平α：当α减少时增大。
- 总体标准差σ：当σ增大时增大。
- 样本容量n：当n减少时增大。

3 显著性水平

当样本容量n固定时, 选定α后, k就可以确定, 也就是判定假设的门槛确定. α称为**显著性水平**(significance level)。

- 显著性水平α(significant level)。
  - 研究者事先确定：我们可以在事先确定用于拒绝原假设H_0的证据必须强到何种程度。
  - α是一个概率值：表示为α 常用的α值有0.01, 0.05, 0.10。
  - 原假设为真时, 拒绝原假设的概率。抽样分布的拒绝域。
- Significant一词的意义在这里并不是"重要的", 而是指"非偶然的"。
- 在假设检验中, 如果样本提供的证据拒绝原假设, 我们说检验的结果是显著的, 如果不拒绝原假设, 我们则说结果是不显著的。
- 一个检验在统计上是"显著的", 意思是指：这样的(样本)结果不是偶然得到的, 或者说, 不是靠机遇能够得到的拒绝原假设, 表示这样的样本结果并不是偶然得到的。不拒绝原假设(拒绝原假设的证据不充分), 则表示这样的样本结果只是偶然得到的。

p值

什么是p值(p-value)：

- 如果原假设为真, 所得到的样本结果会像实际观测结果那么极端或更极端的概率。
- p值告诉我们：如果原假设是正确的话, 我们得到得到目前这个样本数据的可能性有多大, 如果这个可能性很小, 就应该拒绝原假设。
- p值被称为观察到的(或实测的)显著性水平。
- 决策规则：若p-value$< \alpha$, 拒绝H_0。

p值描述样本所提供不利于原假设的证据有多强。但是要证明原假设不正确, p值要多小, 才能令人信服呢？这要根据两种情况来确定：

- 原假设的可信度有多高？如果H_0所代表的假设是人们多年来一直相信的, 就需要很强的证据(小的p值)才能说服他们。
- 拒绝的结论是什么？如果拒绝H_0而肯定H_1, 就需要有很强的证据显示要支持H_1。比如, H_1代表要花很多钱把产品包装改换成另一种包装, 你就要有很强的证据显示新包装一定会增加销售量(因为拒绝H_0要花很高的成本)。

α与p-value

α小概率事件的界限：

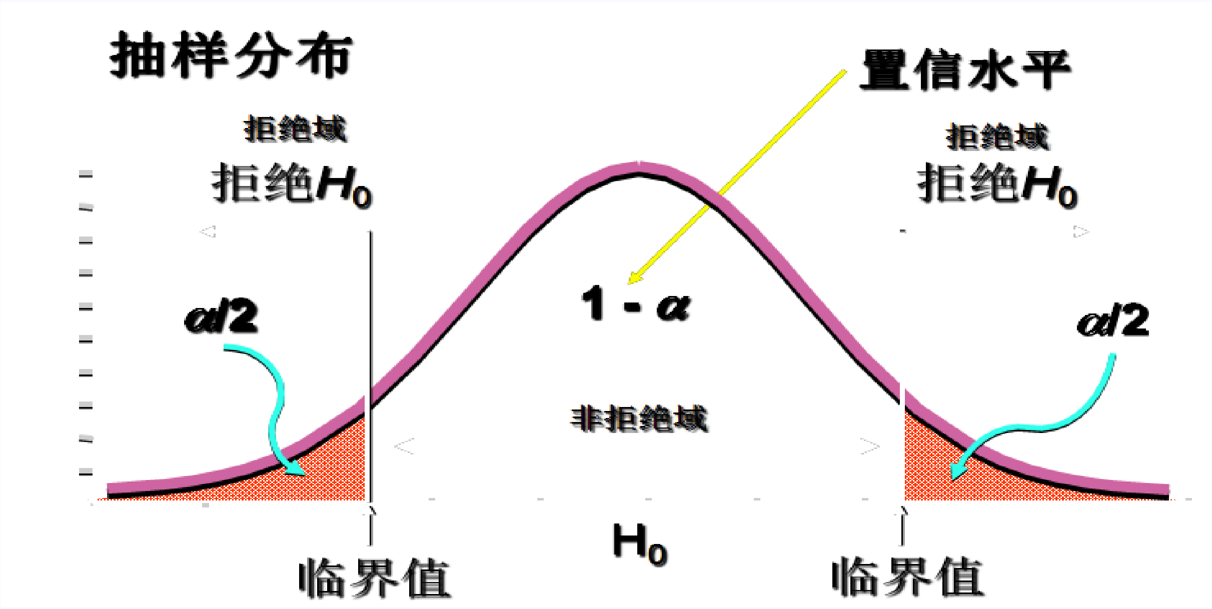
- 在一次试验中，一个几乎不可能发生的事件发生的概率。
- 在一次试验中小概率事件一旦发生，我们就有理由拒绝原假设。
- 小概率界限α由研究者事先确定。

- 我们要求多小的p值。而这个p值就叫显著性水平，用α表示：
  - 显著性水平表示总体中某一类数据出现的经常程度。
  - 假如我们选择$\alpha = 0.05$，样本数据能拒绝原假设的证据要强到：当H_0正确时，这种样本结果发生的频率不超过5%；如果我们选择$\alpha = 0.01$，就是要求拒绝H_0的证据要更强，这种样本结果发生的频率只有1%。
- 如果p值小于或等于α，我们称该组数据不利于原假设的证据有α的显著性水平。

检验统计量

检验统计量(test statistic)：根据样本观测结果计算得到的, 并据以对原假设和备择假设作出决策的某个样本统计量。例如前面例子中$Z=\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$为检验统计量。

统计量是对样本估计量的概率标准化结果



基本概念比较

- 有了p值，我们并不需要用5%或1%这类传统的显著性水平。p值提供了更多的信息，它让我们可以选择任意水平来评估结果是否具有统计上的显著性，从而可根据我们的需要来决定是否要拒绝原假设。只要你认为这么大的p值就算是显著了，你就可以在这样的p值水平上拒绝原假设。
- 传统的显著性水平，如1%、5%、10%等等，已经被人们普遍接受为"拒绝原假设足够证据"的标准，我们大概可以说：10%代表有"一些证据"不利于原假设；5%代表有"适度证据"不利于原假设；1%代表有"很强证据"不利于原假设。
- 用p值进行检验比根据统计量检验提供更多的信息。

基本概念比较

统计量检验是我们事先给出的一个显著性水平，以此为标准进行决策，无法知道实际的显著性水平究竟是多少。

- 根据统计量进行检验时，只要统计量的值落在拒绝域，我们拒绝原假设得出的结论都是一样的，即结果显著。但实际上，统计量落在拒绝域不同的地方，实际的显著性是不同的。
- 统计量落在临界值附近与落在远离临界值的地方，实际的显著性就有较大差异。而p值给出的是实际算出的显著水平，它告诉我们实际的显著性水平是多少。

显著性意义

- 当原假设被拒绝时, 我们称样本结果在统计上是显著的(statistically significant), 当不拒绝原假设时, 我们称样本结果在统计上是不显著的。
  - p值越小, 表明结果越显著。但检验结果究竟是“显著的”、“中度显著的”还是“高度显著的”, 需要由研究者自己根据p值大小和实际问题来决定。
- 在“显著”和“不显著”之间没有清楚的界限, 只是在p值越来越小时, 我们就有越来越强的证据, 检验的结果也就越来越显著。
- 在实际检验中, 不要把统计上的显著性与实际上的显著性混同起来。一个在统计上显著的结论在实际中却不见得很重要, 也不意味着就有实际意义。因为p值不仅和样本的大小密切相关, 也和总体参数的真值有关。
  - 无论总体的状况如何, 观测值多一点, 就可以让p值准些。
  - 在假设检验时, 不仅要报告p值, 而且也要报告样本大小。

4 双边检验和单边检验

- 双边检验
  备择假设H_1没有特定的方向性, 表示为$\neq$, 即可能大于也可能小于。此类假设检验称为双边(双侧、双尾)检验(two-tailed test)。例如：

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

- 单边检验
  备择假设具有特定的方向性, 含有符号"$>$"或"$<$"的假设检验, 此类假设检验称为称为单边(单侧、单尾)检验(one-tailed test)。分为左边检验和右边检验。

  - 例如：右边检验为

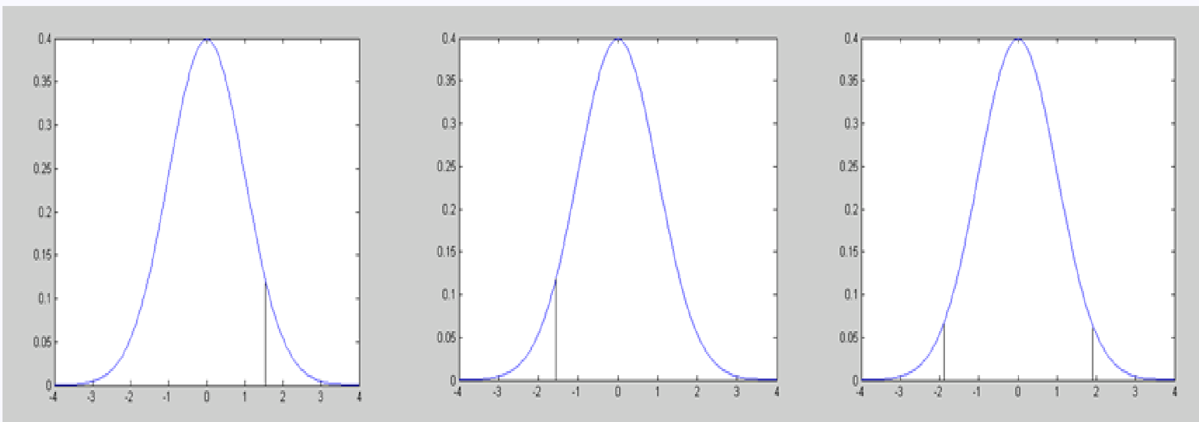
$$H_0 : \mu \leq \mu_0, H_1 : \mu > \mu_0$$

  - 例如：左边检验为

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

双边检验和单边检验

$$(a)W = \{\mu > c\}, (b)W = \{\mu < c\}, (c)W = \{\mu > c_1, or, \mu < c_2\}$$



$(a)\ H_1 : \mu > \mu_0$ $(b)\ H_1 : \mu < \mu_0$ $(c)\ H_1 : \mu \neq \mu_0$

单 边 检 验

计算单边检验的拒绝域, 右边检验的拒绝域是$\bar{x} \geq k$, 则

$$P_{\mu \leq \mu_0}\{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \geq \frac{k-\mu_0}{\sigma/\sqrt{n}}\} = \alpha$$

$$\frac{k-\mu_0}{\sigma/\sqrt{n}} = z_\alpha$$

$$k = \mu_0 + \frac{\sigma}{\sqrt{n}} z_\alpha$$

即得检验问题的拒绝域为

$$\bar{x} \geq \mu_0 + \frac{\sigma}{\sqrt{n}} z_\alpha, z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \geq z_\alpha$$

对应的左边检验拒绝域为

$$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \leq -z_\alpha$$

例 子

Example 2 某厂的固体燃料推进器的燃烧率服从正态分布$N(\mu,\sigma^2)$, $\mu = 40cm/s$, $\sigma = 2cm/s$. 现用一种新方法生产了一批推进器. 从中随机取$n = 25$只, 测得燃烧率的样本均值为$\overline{x} = 41.25cm/s$. 设均差仍为$2cm/s$, 请问新方法是否对燃烧率有显著提高? ($\alpha = 0.05$)

解: 按题意检验假设如下:

$$H_0:\mu \le \mu_0,\text{没有提高},\ H_1:\mu > \mu_0,\text{有提高}$$

这是右边检验问题, 其拒绝域为

$$z = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} = \frac{41.25-40}{2/\sqrt{25}} = 3.125 > z_{0.05} = 1.645$$

所以在显著性水平$\alpha = 0.05$下拒绝H_0, 所以认为新方法提高了燃烧率。

5 检验步骤

综上所述, 假设检验步骤为:

1. 根据问题要求建立备选假设H_1和原假设H_0。
2. 从研究的总体中抽出一个随机样本。
3. 选取合适的统计量, 其抽样分布不含任何未知参数。利用样本数据计算具体数据。以便计算其分位点.
4. 确定一个适当的显著性水平，并计算出其临界值，指定拒绝域。
5. 将统计量的值与临界值进行比较, 作出决策:
   - 统计量的值落在拒绝域, 拒绝H_0, 否则不拒绝H_0。
   - 也可以直接利用p值作出决策。

6 总体均值检验

- 检验问题：单个总体均值μ的检验。
- 检验条件：
  - 总体近似服从正态分布。
  - 总体不是正态分布但是大样本$(n \geq 30)$。

关于μ的检验(Z检验):
正态分布且σ^2已知：

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

大样本且σ^2未知：

$$Z = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim N(0,1)$$

所以利用$Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$或$Z = \frac{\overline{X}-\mu_0}{s/\sqrt{n}}$来确定拒绝域。

原假设H_0	备择假设H_1	拒绝域
$\mu \leq \mu_0$	$\mu > \mu_0$	$z \geq z_\alpha$
$\mu \geq \mu_0$	$\mu < \mu_0$	$z \leq -z_\alpha$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert z\rvert \geq z_{\alpha/2}$

例子

Example 3 一种机床加工的零件尺寸绝对平均误差为*1.35mm*。生产厂家现采用一种新的机床进行加工以期进一步降低误差。为检验新机床加工的零件平均误差与旧机床相比是否有显著降低, 从某天生产的零件中随机抽取*50*个进行检验。利用这些样本数据, 检验新机床加工的零件尺寸的平均误差与旧机床相比是否有显著降低? $(\alpha = 0.01)$

1.26	1.13	0.98	1.12	1.23	0.99	1.98	1.11	1.70	1.17
1.19	0.96	1.10	1.12	0.74	1.45	1.97	1.54	2.37	1.12
1.31	1.06	1.12	0.95	1.50	1.24	0.91	1.08	1.38	1.23
0.97	1.00	1.03	1.02	0.50	1.01	1.22	1.10	1.60	0.82
1.81	0.94	1.16	1.13	0.59	2.03	1.06	1.64	1.26	0.86

解：建立假设：

$$H_0 : \mu \geq 1.35, H_1 : \mu < 1.35$$

$\overline{x} = 1.3152, s = 0.365749, n = 50,$ 计算检验统计量:

$$z = \frac{1.3152 - 1.35}{0.365749/\sqrt{50}} = -2.6061 < -z_{0.01} = -2.33$$

结论：拒绝H_0。新机床加工的零件尺寸的平均误差与旧机床相比有显著降低。

计算p值$= \Phi(-2.6061) = 1 - \Phi(2.6061) = 0.004579$。

总体均值检验(小样本)

- 检验问题：单个总体均值μ的检验。
- 检验条件：总体近似服从正态分布但是小样本($n < 30$), 且σ^2未知。

关于μ的检验(t检验): 因为

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim t(n-1)$$

所以利用t来确定拒绝域。

原假设H_0	备择假设H_1	拒绝域
$\mu \le \mu_0$	$\mu > \mu_0$	$t \ge t_\alpha(n-1)$
$\mu \ge \mu_0$	$\mu < \mu_0$	$t \le -t_\alpha(n-1)$
$\mu = \mu_0$	$\mu \ne \mu_0$	$\|t\| \ge t_{\alpha/2}(n-1)$

例 子

Example 4 某种元件的寿命X(以小时计)服从正态分布$N(\mu,\sigma^2)$, μ,σ^2均未知, 现测得*16*只元件寿命如下:

$$159, 280, 101, 212, 224, 379, 179, 264$$

$$222, 362, 168, 250, 149, 260, 485, 170$$

问是否有理由认为元件的平均寿命大于*225*小时?

解: 按题意需检验

$$H_0:\mu\le\mu_0, H_1:\mu>\mu_0$$

取$\alpha=0.05$. 有题中的数据可知$n=16$, $\overline{x}=241.5$, $s=98.7259$, $t_{0.05}(15)=1.7531$

$$t=\frac{\overline{x}-\mu_0}{s/\sqrt{n}}=0.6685<1.7531$$

t没有落入拒绝域, 故不拒绝H_0。即统计意义上元件寿命大于225小时不显著。

7 总体均值差检验

两个总体$N(\mu_1,\sigma_1^2)$和$N(\mu_1,\sigma_2^2)$的均值检验，其中σ_1^2,σ_2^2已知。
设随机变量分别为X,Y，样本均值为$\overline{X},\overline{Y}$。
现在来检验问题：

$$H_0:\mu_1-\mu_2=\delta, H_1:\mu_1-\mu_2\neq\delta$$

δ为已知常数. 取显著性水平α, 因为

$$Z=\frac{(\overline{X}-\overline{Y})-\delta}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}\sim N(0,1)$$

则拒绝域为

$$|\frac{(\overline{X}-\overline{Y})-\delta}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}|\geq k$$

$$|\frac{(\overline{x}-\overline{y})-\delta}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}|\geq z_{\alpha/2}$$

原假设H_0	备择假设H_1	拒绝域
$\mu_1-\mu_2\leq\delta$	$\mu_1-\mu_2>\delta$	$z\geq z_\alpha$
$\mu_1-\mu_2\geq\delta$	$\mu_1-\mu_2<\delta$	$z\leq -z_\alpha$
$\mu_1-\mu_2=\delta$	$\mu_1-\mu_2\neq\delta$	$\lvert z\rvert\geq z_{\alpha/2}$

总体均值差检验

两个总体$N(\mu_1,\sigma_1^2)$和$N(\mu_1,\sigma_2^2)$的均值检验, 其中$\sigma_1^2=\sigma_2^2$, 但未知. 设随机变量分别为X,Y, 样本均值为$\overline{X},\overline{Y}$. 现在来检验问题

$$H_0:\mu_1-\mu_2=\delta, H_1:\mu_1-\mu_2\neq\delta$$

δ为已知常数. 取显著性水平α, 因为

$$t=\frac{(\overline{X}-\overline{Y})-\delta}{s_w\sqrt{1/n_1+1/n_2}}\sim t(n_1+n_2-2)$$

其中$s_w^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$. 则拒绝域为

$$|\frac{(\overline{X}-\overline{Y})-\delta}{s_w\sqrt{1/n_1+1/n_2}}|\geq k$$

$$|\frac{(\overline{X}-\overline{Y})-\delta}{s_w\sqrt{1/n_1+1/n_2}}|\geq t_{\alpha/2}(n_1+n_2-2)$$

原假设H_0	备择假设H_1	拒绝域
$\mu_1-\mu_2\leq\delta$	$\mu_1-\mu_2>\delta$	$t\geq t_\alpha(n_1+n_2-2)$
$\mu_1-\mu_2\geq\delta$	$\mu_1-\mu_2<\delta$	$t\leq -t_\alpha(n_1+n_2-2)$
$\mu_1-\mu_2=\delta$	$\mu_1-\mu_2\neq\delta$	$\|t\|\geq t_{\alpha/2}(n_1+n_2-2)$

例 子

Example 5 在平炉上改进方法增加钢产量. 在两座炉上分别用两种方法交替炼钢, 各*10*炉, 钢得率如下:
老方法: $78.1, 72.4, 76.2, 74.3, 77.4, 78.4, 76.0, 75.5, 76.7, 77.3$
新方法: $79.1, 81.0, 77.3, 79.1, 80.0, 79.1, 79.1, 77.3, 80.2, 82.1$
设两个样本相互独立, 分别来自正态分布$N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$, $\mu_1, \mu_2, \sigma_1, \sigma^2$均未知, 问是否有理由认为新方法提高了钢产率($\alpha = 0.05$)?

解: 按题意需检验

$$H_0 : \mu_1 - \mu_2 \geq 0, H_1 : \mu_1 - \mu_2 < 0$$

根据题意分别求得以下样本值:

$$n_1 = 10, \overline{x} = 76.23, s_1^2 = 3.325$$

$$n_2 = 10, \overline{x} = 79.43, s_2^2 = 2.225$$

计算$s_w^2 = 2.775$. 取$t_{0.05}(18) = 1.7341$, 计算可得

$$t = -4.295 < -1.7341$$

t落入拒绝域, 故拒绝H_0, 即新方法改进了得钢率.

基于配对数据的检验

为了比较两种产品或两种方法的差异, 我们常在相同条件下做对比实验, 根据一批配对的观察值, 然后分析观察值, 分析数据作出推断, 称为逐对比较法.

设n对独立的观察结果: $D_i = X_i - Y_i$, 由于D_i是同一因素引出, 认为D_i服从同一分布. 假设$D_i \sim N(\mu_d, \sigma_d^2)$, 其中$\mu_d, \sigma_d^2$未知. 现在来检验问题

$$H_0 : \mu_d = 0, H_1 : \mu_d \neq 0$$

取显著性水平α, 则拒绝域为

$$|t| = |\frac{\bar{d}}{s_d/\sqrt{n}}| \geq t_{\alpha/2}(n-1)$$

原假设H_0	备择假设H_1	拒绝域
$\mu_d \leq 0$	$\mu_d > 0$	$t \geq t_\alpha(n-1)$
$\mu_d \geq 0$	$\mu_d < 0$	$t \leq -t_\alpha(n-1)$
$\mu_d = 0$	$\mu_d \neq 0$	$\|t\| \geq t_{\alpha/2}(n-1)$

例 子

Example 6 在平炉上改进方法增加钢产量. 在同一座炉上分别用两种方法交替炼钢, 各*10*炉, 钢得率如下*:*
老方法*:* $78.1, 72.4, 76.2, 74.3, 77.4, 78.4, 76.0, 75.5, 76.7, 77.3$
新方法*:* $79.1, 81.0, 77.3, 79.1, 80.0, 79.1, 79.1, 77.3, 80.2, 82.1$
设$\mu_1, \mu_2, \sigma_1, \sigma^2$均未知, 问是否有理由认为新方法提高了钢产率*(*$\alpha = 0.05$*)?*

解: 两种方法(新-老)的差为

$$1, 8.6, 1.1, 4.8, 2.6, 0.7, 2.9, 1.8, 3.5, 4.8$$

按题意需检验

$$H_0 : \mu_d \leq 0, H_1 : \mu_d > 0$$

根据题意分别求得以下样本值: $\overline{x}_d = 3.18, s_d = 2.41$, 计算

$$t = \frac{\overline{d}}{s_d/\sqrt{n}} = \frac{3.18}{2.41/\sqrt{10}} = 4.17$$

取$t_{0.05}(9) = 1.833$, 则$t = 4.17 > 1.833$ t落入拒绝域, 故拒绝H_0, 即新方法改进了得钢率.

例子

Example 7 某健美俱乐部宣称通过一期训练班可以有效减肥（*8.5kg*），他们随机调查了*10*名健身者训练前后的体重如下：
训练前：$94.5, 101, 110, 103.5, 97, 88.5, 96.5, 101, 104, 116.5$
训练后：$85, 89.5, 101.5, 96, 86, 80.5, 87, 93.5, 93, 102$
*(1)*问是否有理由认为减肥广告有效$(\alpha = 0.05)$?
*(2)*假如上述数据是从不同人随机调查的，是否有理由认为减肥广告有效$(\alpha = 0.05)$?

如何决定数据是否来自配对样本？

总体方差检验

单个总体$N(\mu,\sigma^2)$方差σ^2的检验. μ,σ^2已知, 关于σ的检验($\mathcal{X}^2$检验):

$$H_0:\sigma^2=\sigma_0^2, H_1:\sigma^2\neq\sigma_0^2$$

其中σ_0为常数. 因为

$$\mathcal{X}^2=\frac{(n-1)S^2}{\sigma_0^2}\sim\mathcal{X}^2(n-1)$$

所以利用$\mathcal{X}^2$来确定拒绝域.

原假设H_0	备择假设H_1	拒绝域
$\sigma^2\le\sigma_0^2$	$\sigma^2>\sigma_0^2$	$\mathcal{X}^2\ge\mathcal{X}_\alpha^2(n-1)$
$\sigma^2\ge\sigma_0^2$	$\sigma^2<\sigma_0^2$	$\mathcal{X}^2\le\mathcal{X}_{1-\alpha}^2(n-1)$
$\sigma^2=\sigma_0^2$	$\sigma^2\neq\sigma_0^2$	$\mathcal{X}^2\ge\mathcal{X}_{\alpha/2}^2(n-1)$ 或$\mathcal{X}^2\le\mathcal{X}_{1-\alpha/2}^2(n-1)$

例 子

Example 8 生产的某型号电池, 其寿命服从方差$\sigma^2 = 5000$的正态分布. 随机取*26*个电池, 测出样本方差为$s^2 = 9200$, 问能否推断波动较以往显著变化*(*$\alpha = 0.02$*)?*

解: 按题意需检验

$$H_0 : \sigma^2 = 5000, H_1 : \sigma^2 \neq 5000$$

根据题意$n = 26$, $\mathcal{X}^2_{0.01}(25) = 44.314$, $\mathcal{X}^2_{0.99}(25) = 11.524$, 拒绝域为

$$\mathcal{X}^2 \geq 44.314 \text{ 或者} \mathcal{X}^2 \leq 11.524$$

$\mathcal{X}^2 = \frac{(n-1)s^2}{\sigma_0^2} = 46 > 44.314$. 落入拒绝域, 故拒绝$H_0$, 即波动显著变化.

总体方差检验

两个总体$N(\mu_1,\sigma_1^2)$和$N(\mu_2,\sigma_2^2)$, 其中$\mu_1,\sigma_1^2,\mu_2,\sigma_2^2$均未知. 现需要检验:

$$H_0:\sigma_1^2\le\sigma_2^2, H_1:\sigma_1^2>\sigma_2^2$$

其中σ_0为常数. 因为

$$F=\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2}\sim F(n_1-1,n_2-1)$$

所以利用F来确定拒绝域.

原假设H_0	备择假设H_1	拒绝域
$\sigma_1^2\le\sigma_2^2$	$\sigma_1^2>\sigma_2^2$	$F\ge F_\alpha(n_1-1,n_2-1)$
$\sigma_1^2\ge\sigma_2^2$	$\sigma_1^2<\sigma_2^2$	$F\le F_{1-\alpha}(n_1-1,n_2-1)$
$\sigma_1^2=\sigma_2^2$	$\sigma_1^2\ne\sigma_2^2$	$F\ge F_{\alpha/2}(n_1-1,n_2-1)$ 或$F\le F_{1-\alpha/2}(n_1-1,n_2-1)$

8 置信区间与假设检验

我们在点估计中求得θ的估计值$\hat{\theta}(x_1, \ldots, x_n)$是其一种近似值, 但是点估计本身既没有反映这种近似值的精确度, 也不知道它的误差范围. 因此有必要进一步探讨用估计量来估计θ的误差范围, 因而引入区间估计。

一般而论, 我们就是要通过样本构造一个随机区间$(\underline{\theta}, \overline{\theta})$并指出该区间以多大的可信度包含未知参数$\theta$, 这类问题称作区间估计问题, 相应的区间称作置信区间.

置信区间与假设检验

Definition 1 设总体$\zeta \sim f(x,\theta)$, θ未知. 若对于事先给定的α, 存在两个统计量$\underline{\theta}(\zeta), \overline{\theta}(\zeta)$, 使得

$$P\{\underline{\theta} < \theta < \overline{\theta}\} = 1-\alpha$$

则称$(\underline{\theta}, \overline{\theta})$是参数$\theta$的置信度为$1-\alpha$的置信区间. $\underline{\theta}, \overline{\theta}$分别称为置信下限和置信上限.

- 置信区间是一个随机区间.
- 置信区间的端点和区间长度都是样本的函数, 即统计量.
- 置信区间的含义是: 大量重复抽样下, 将样本观察值代入$\underline{\theta}, \overline{\theta}$可求得许多确定区间, 其中约$100(1-\alpha)\%$的区间包含$\theta$在内.

置信区间与假设检验

- 作一项统计推断, 不能只去看是否有统计上的显著性, 置信区间往往会更有用。
- 置信区间的宽度会帮助我们把真正的总体参数定位得更准确。

- 双边检验问题$H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$的接受域可以定出正态均值$\mu$的$1-\alpha$置信区间.
- 右边检验问题$H_0 : \mu \leq \mu_0, H_1 : \mu > \mu_0$的接受域可以定出正态均值$\mu$的$1-\alpha$置信上限.
- 左边检验问题$H_0 : \mu \geq \mu_0, H_1 : \mu < \mu_0$的接受域可以定出正态均值$\mu$的$1-\alpha$置信下限.

例 子

Example 9 设$X \sim N(\mu, \sigma^2)$, μ未知, $\sigma^2 = 1$, $n = 16$, 且由一样本算得$\overline{x} = 5.20$, 于是得参数μ的一个置信水平为*0.95*的置信区间为

$$\overline{x} \pm \frac{1}{\sqrt{16}} z_{0.025} = (5.20 \pm 0.49) = (4.71, 5.69)$$

现考虑检验问题:

$$H_0 : \mu = 5.5, H_1 : \mu \neq 5.5$$

由于$5.5 \in (4.71, 5.69)$, 所以不能拒绝H_0.

例 子

Example 10 $X \sim N(\mu, \sigma^2)$, μ未知, $\sigma^2 = 1$, $n = 16$, 且由一样本算得$\overline{x} = 5.20$, $\alpha = 0.05$

1. 采用假设检验$H_0 > 5$是否成立。
2. 通过置信区间检验上述假设。
3. 若三条生产线的样本分别为$\overline{x}_1 = 5.10, \overline{x}_2 = 5.20, \overline{x}_3 = 5.30$, 如何进行检验更方便？

例 子

Example 11 两批电子元件样品的性能如下：
A(X): 140, 138, 143, 142, 144, 137
B(Y): 135, 140, 142, 146, 138, 140
假设$X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, 试分别采用区间估计和假设检验方法解决

1. $\sigma_1^2 = \sigma_2^2$?
2. $\mu_1 = \mu_2$?

9 样本容量问题

在前面的讨论中，我们都是假设样本容量(Sample Size)n是已知的, 但是在实际问题中, 需要自己动手设计调查方案, 这时, 如何决定样本容量大有学问.

- 如果n选得过大, 会增加费用;
- 如果n选得过小, 会使估计误差增大.

样 本 容 量 问 题

要增强检验法的可信度和说服力, 需要提交检验正确判断的概率.

- 我们可以通过提高置信水平α(比如从0.05调整为0.01)来减少犯I类错误的概率.
- 我们可以通过降低β值来减少犯II类错误的概率.
- 样本容量确定的情况下, 不可能同时提高上述两方面.

一般的做法是: 先限制犯第一类错误的概率α, 然后利用备择假设确定β的值. 如果β太大, 则增大样本容量使β减小; 如果实际问题不需要β太小, 则可考虑减小样本容量以节省人力、物力与时间.

OC 函数与统计功效

Definition 2 (OC函数) 若C是参数θ的某检验问题的一个检验法,

$$\beta(\theta) = P_\theta(\text{接受}H_0)$$

称为检验法C的OC函数.

设此检验法的显著性水平为α,

- 当$\theta \in H_0$, $\beta(\theta)$就是做出正确判断的概率, 故$\beta(\theta) \geq 1-\alpha$.
- 当$\theta \in H_1$, $\beta(\theta)$就是犯II类错误的概率, 故$1-\beta(\theta)$为正确判断的概率.

Definition 3 (功效) 函数$1-\beta(\theta)$称为检验法C的功效函数. 当$\theta' \in H_1$时, $1-\beta(\theta')$称为检验法C在θ'上的统计功效.

统计功效代表参数真值为θ'时, 检验法做出正确判断的概率.

Z检验的功效分析

首先考察右边检验问题

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

相应的OC函数为

$$\beta(\mu) = P_\mu(\text{接受}H_0) = P_\mu\{\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} < z_\alpha\}$$

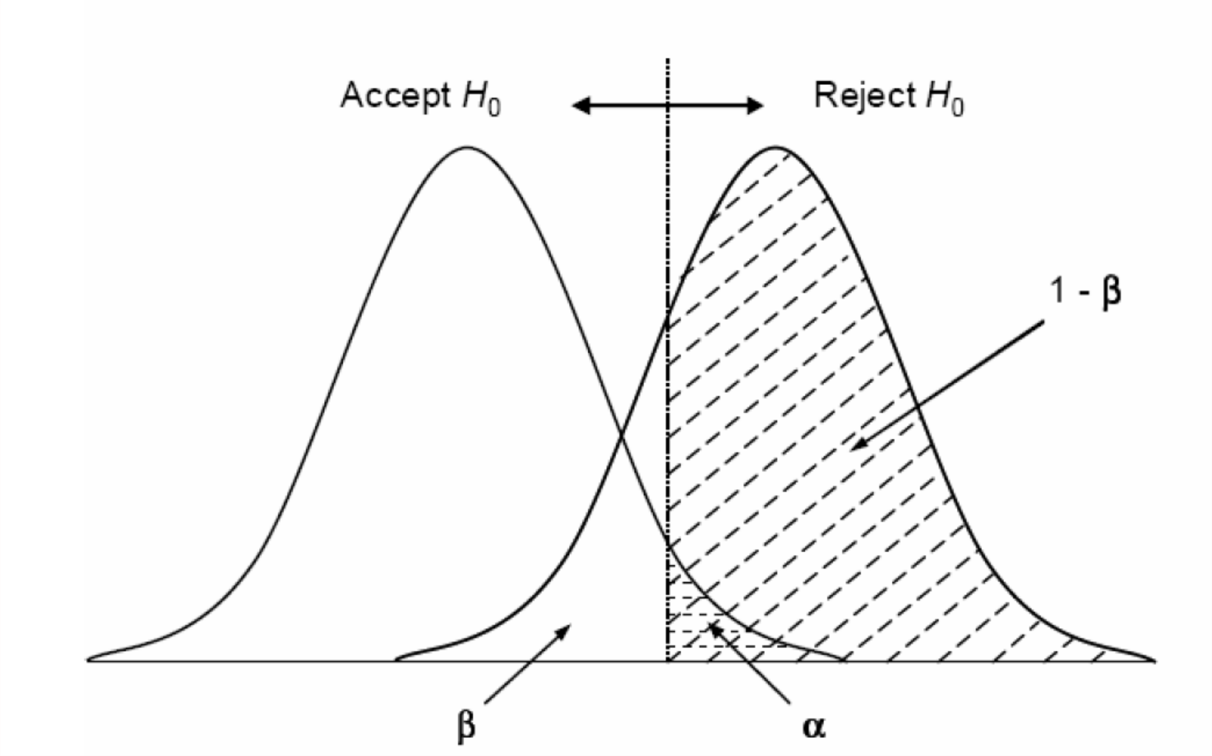
$$= P_\mu\{\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_\alpha - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\} = \Phi(z_\alpha - \lambda), \lambda = \frac{\mu - \mu_0}{\sigma/\sqrt{n}}$$

OC函数具有以下性质:

- 它是λ的单调递减连续函数;
- $\lim_{\mu\to\mu_0^+} \beta(\mu) = 1 - \alpha$;
- $\lim_{\mu\to\infty} \beta(\mu) = 0$.

当参数真值μ仅仅比μ_0大很少时, $\beta(\mu)$很大, 接近$1 - \alpha$, 即统计功效很小. 也就是犯II类错误的概率很大. 此时无论样本容量多大, 控制犯II类错误的概率基本不可能.

统计功效示意图



样本容量确定

假设$\mu \geq \mu_0 + \delta(\delta > 0)$, 对于给定的$\beta$, 我们如何确定样本容量$n$?

$\beta(\mu)$是μ的递减函数, 则有

$$\beta(\mu) \leq \beta(\mu_0 + \delta) = \Phi(z_\alpha - \sqrt{n}\delta/\sigma) \leq \beta$$

有

$$z_\alpha - \sqrt{n}\delta/\sigma \leq -z_\beta$$

即样本容量n满足

$$\sqrt{n} \geq \frac{(z_\alpha + z_\beta)\sigma}{\delta}$$

就能使得$\mu \in H_1$且$\mu \geq \mu_0 + \delta$时犯II类错误的概率不大于β.

例 子

Example 12 (产品质量抽检) 某产品质量指标$X \sim N(\mu, \sigma^2)$. μ小为佳, 厂家的验收方案中要求买家对高质量产品($\mu \le \mu_0$)能以高概率$1-\alpha$接受. 买家的验收方案中要求对低劣产品($\mu \ge \mu_0 + \delta$)能以高概率$1-\beta$拒绝. 已知$\mu_0 = 120$, $\delta = 20$, 由经验知$\sigma^2 = 900$, 双方协定$\alpha = \beta = 0.05$. 试确定抽样验收方案.

解: 检验问题表达为

$$H_0 : \mu \le \mu_0, H_1 : \mu > \mu_0$$

由Z检验法可知拒绝域为

$$\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \ge z_\alpha$$

由

$$\sqrt{n} \ge \frac{(z_\alpha + z_\beta)\sigma}{\delta} = \frac{(1.645 + 1.645)30}{20} = 4.935$$

故取$n = 25$, 且当抽样均值$\overline{x}$满足

$$\frac{\overline{x} - 120}{30/5} \ge z_{0.05} = 1.645$$

即$\overline{x} \ge 129.87$时买方拒绝这批产品, 否则接受这批产品.

Z检验的功效分析

考虑左边检验问题$H_0: \mu \geq \mu_0, H_1: \mu < \mu_0$, 则OC函数

$$\beta(\mu) = \Phi(z_\alpha + \lambda), \lambda = \frac{\mu - \mu_0}{\sigma/\sqrt{n}}$$

样本容量确定公式为

$$\sqrt{n} \geq \frac{(z_\alpha + z_\beta)\sigma}{\delta}$$

考虑双边检验问题$H_0: \mu = \mu_0, H_1: \mu \neq \mu_0$, 则OC函数

$$\beta(\mu) = \Phi(z_{\alpha/2} - \lambda) + \Phi(z_{\alpha/2} + \lambda) - 1, \lambda = \frac{\mu - \mu_0}{\sigma/\sqrt{n}}$$

Z检验的功效分析

样本容量确定公式需要求解以下形式的超越方程.

$$\beta = \Phi(z_{\alpha/2} - \sqrt{n}\delta/\sigma) + \Phi(z_{\alpha/2} + \sqrt{n}\delta/\sigma) - 1$$

通常我们采取以下近似方法, n较大, 可认为$z_{\alpha/2} + \sqrt{n}\delta/\sigma > 4$, 则$\Phi(z_{\alpha/2} + \sqrt{n}\delta/\sigma) = 1$, 所以近似有

$$\beta = \Phi(z_{\alpha/2} - \sqrt{n}\delta/\sigma)$$

即样本容量确定公式近似为

$$\sqrt{n} \geq \frac{(z_{\alpha/2} + z_\beta)\sigma}{\delta}$$

t检验的功效分析

首先考察右边检验问题

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

t检验法的OC函数为

$$\beta(\mu) = P_\mu(\text{接受}H_0) = P_\mu\{\frac{\overline{X} - \mu_0}{S/\sqrt{n}} < t_\alpha(n-1)\}$$

其中变量

$$\frac{\overline{X} - \mu_0}{S/\sqrt{n}} = (\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} + \lambda)/\frac{S}{\sigma}, \lambda = \frac{\mu - \mu_0}{\sigma/\sqrt{n}}$$

我们称变量$\frac{\overline{X}-\mu_0}{S/\sqrt{n}}$服从非中心参数为$\lambda$、自由度为$n-1$的非中心t分布. $\lambda = 0$时, 退化成普通t分布.

对于给定α, β和δ, 可由附表查得所需样本容量n, 使得$\mu \in H_1$且$\frac{\mu-\mu_0}{\sigma} \ge \delta$(相应地左边检验问题为$\frac{\mu-\mu_0}{\sigma} \le -\delta$, 双边检验问题为$\frac{|\mu-\mu_0|}{\sigma} \ge \delta$)时犯II类错误的概率不超过$\beta$.

例 子

Example 13 考虑$\alpha = 0.05$下进行t检验:

$$H_0 : \mu \leq 68, H_1 \mu > 68$$

(1) 要求在H_1中$\mu \geq \mu_1 = 68 + \sigma$时犯$II$类错误的概率不超过$\beta = 0.05$. 试确定样本容量.

(2) 若样本容量为$n = 30$, 问在H_1中$\mu = \mu_1 = 68 + 0.75\sigma$时犯$II$类错误的概率为多少?

例 子

Example 14 考虑$\alpha = 0.05$下进行t检验:

$$H_0 : \mu \le 68, H_1 \mu > 68$$

(1) 要求在H_1中$\mu \ge \mu_1 = 68 + \sigma$时犯*II*类错误的概率不超过$\beta = 0.05$. 试确定样本容量.

(2) 若样本容量为$n = 30$, 问在H_1中$\mu = \mu_1 = 68 + 0.75\sigma$时犯*II*类错误的概率为多少?

解:

(1) $\delta = \frac{\mu_1 - \mu_0}{\sigma} = 1, \alpha = \beta = 0.05$, 查表得$n = 13$.

(2) $\alpha = 0.05, n = 30, \delta = 0.75$, 查表得$\beta = 0.01$.

内 容 回 顾

- 基本概念
  - 假设建立
  - 两类错误
  - 显著性水平
  - 基本步骤
- 总体均值检验
- 总体方差检验
- 置信区间与假设检验
- 样本容量与统计功效分析