《Python统计计算》

（2021年秋季学期）

翟祥

北京林业大学

E-mail: zhaixbh@126.com

A/B测试--假设检验

A/B测试

- 在产品发布，运营等场景我们都会遇到A/B test。A/B test通常为同一个目标，设计两种方案，将两种方案随机投放市场中。A/B test让组成成分相同（相似）用户去随机体验两种方案之一，根据观测结果，判断哪个方案效果更好，结果可以通过CTR或者下单率来衡量。最终我们选择CTR或者下单率更优的版本作为线上应用的版本。

A/B测试

- 现实场景中我们避不开几个问题：
- **A/B test两组人群的转化效果是否存在差异——假设检验**
- **我们正确判断出A/B test两组人群有差异的把握有多大——统计功效**
- **在一定显著性水平和和统计功效下，我们需要选定多少样本量进行试验——反选样本量**

假设检验步骤的总结

1. 陈述原假设和备择假设
2. 从所研究的总体中抽出一个随机样本
3. 确定一个适当的检验统计量，并利用样本数据算出其具体数值
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

假设检验中的小概率原理

- **什么小概率？**
- 1. 在一次试验中，一个几乎不可能发生的事件发生的概率
- 2. 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
- 3. 小概率由研究者事先确定

什么是小概率？

小概率事件——

若$P(A)\le 0.01$，则称A为小概率事件.

小概率原理——（即实际推断原理）

一次试验中小概率事件一般是不会发生的. 若在一次试验中居然发生了，则可怀疑该事件并非小概率事件.

假设的陈述

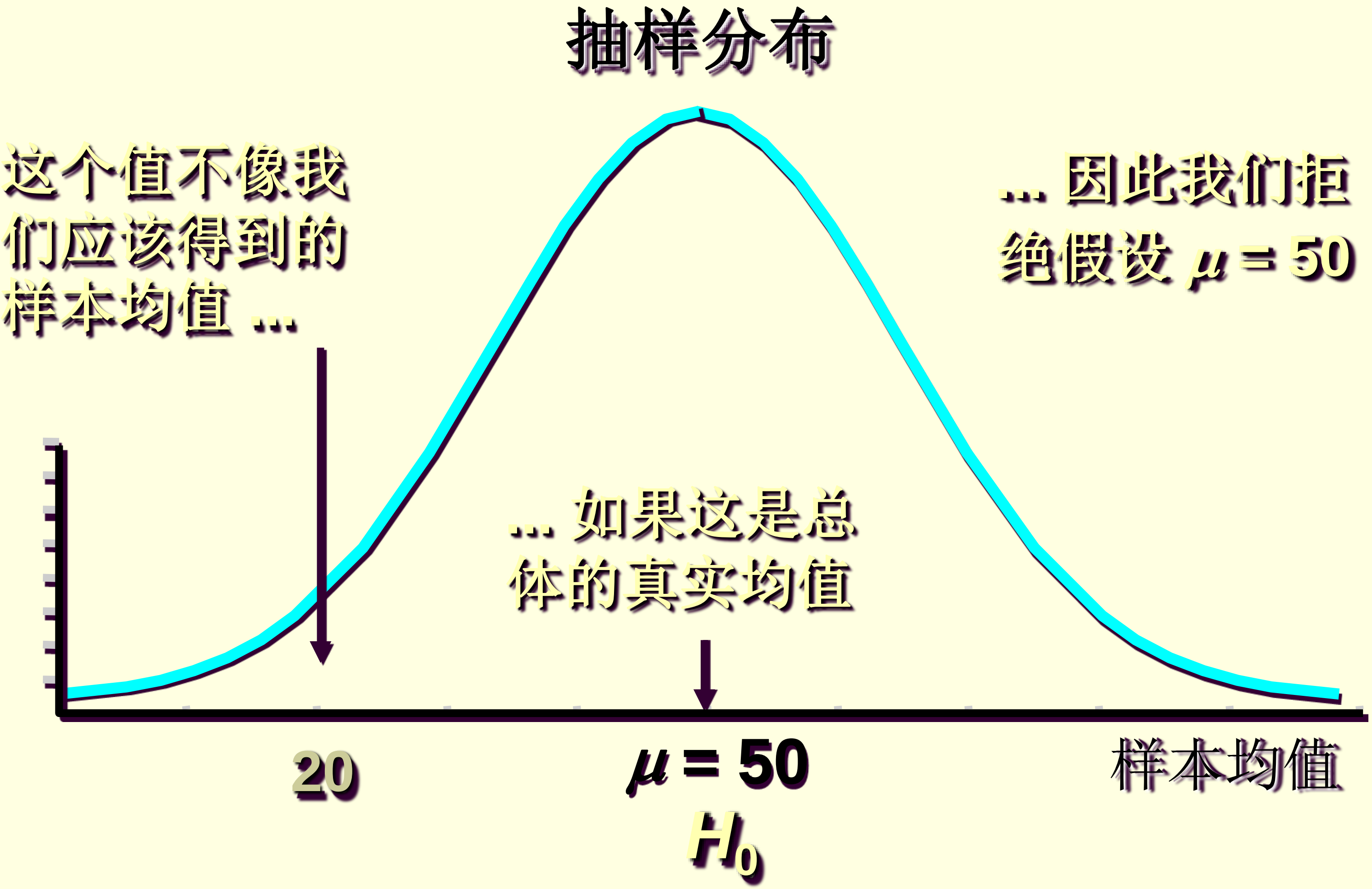
什么是假设?
(hypothesis)

- 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
  - 分析**之前**必需陈述

什么是假设检验?
(hypothesis test)

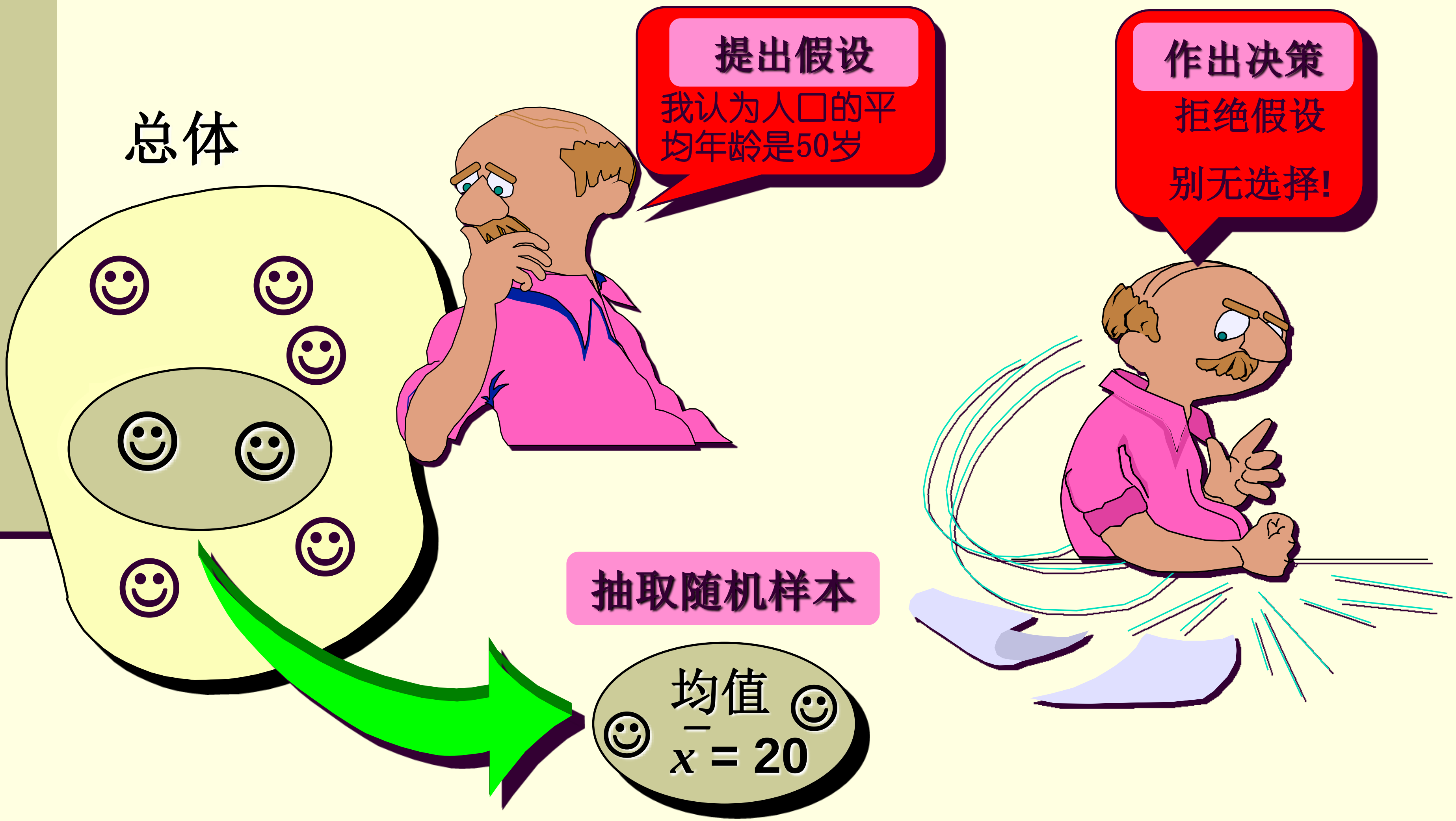
1. 先对总体的参数(或分布形式)提出某种假设，然后利用样本信息判断假设是否成立的过程
2. 有参数检验和非参数检验
3. 逻辑上运用反证法，统计上依据小概率原理

假设检验的基本思想



抽样分布
这个值不像我
们应该得到的
样本均值 ...
... 因此我们拒
绝假设 $\mu = 50$
... 如果这是总
体的真实均值
20
$\mu = 50$
H_0
样本均值

假设检验的过程



原假设与备择假设

原假设
(null hypothesis)

1. 研究者想收集证据予以反对的假设
2. 又称“0假设”
3. 总是有符号 =, ≤ 或 ≥

- 4.表示为 H_0
  - H_0： μ = 某一数值
  - 指定为符号 =，≤ 或 ≥
  - 例如, H_0： $\mu = 10\text{cm}$

为什么叫0假设?

备择假设
(alternative hypothesis)

1. 研究者想收集证据予以支持的假设
2. 也称“研究假设”
3. 总是有符号 $\neq$, $<$ 或 $>$
4. 表示为 H_1
   - H_1： μ <某一数值，或μ >某一数值
   - 例如, H_1： μ < 10cm，或μ >10cm

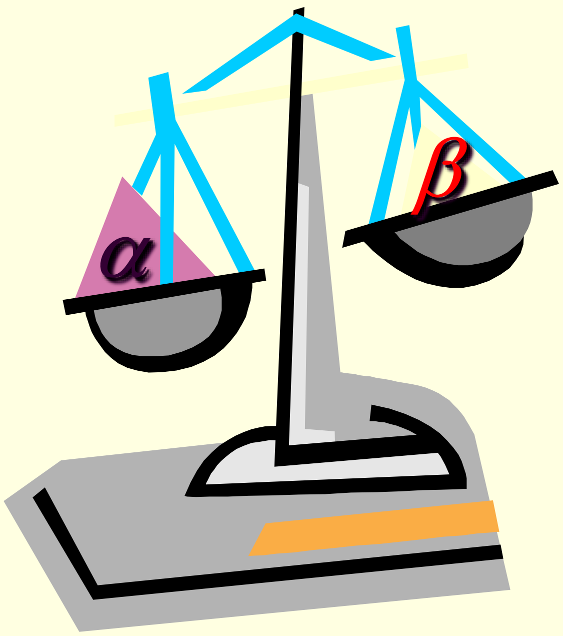
提出假设
(结论与建议)

1. 原假设和备择假设是一个完备事件组，而且相互对立
   - 在一项假设检验中，原假设和备择假设必有一个成立，而且只有一个成立
2. 先确定备择假设，再确定原假设
3. 等号“=”总是放在原假设上
4. 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)

两类错误与显著性水平

假设检验中的两类错误

- **1. 第Ⅰ类错误(弃真错误)**
  - 原假设为真时拒绝原假设
  - 第Ⅰ类错误的概率记为α
    - 被称为显著性水平
- **2. 第Ⅱ类错误(取伪错误)**
  - 原假设为假时未拒绝原假设
  - 第Ⅱ类错误的概率记为β(Beta)



假设检验中的两类错误
(决策结果)

H_0: 无罪

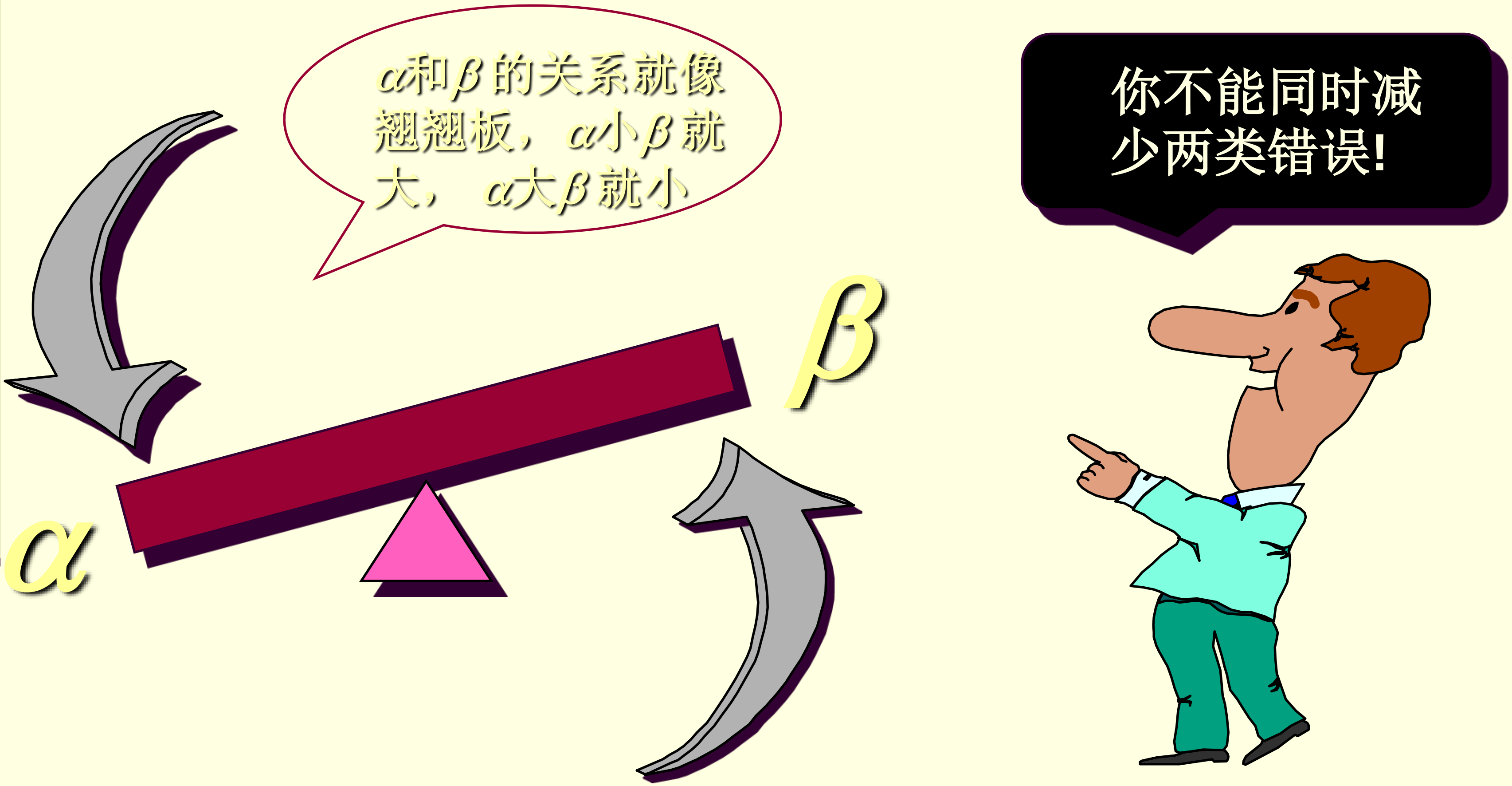
假设检验就好像一场审判过程

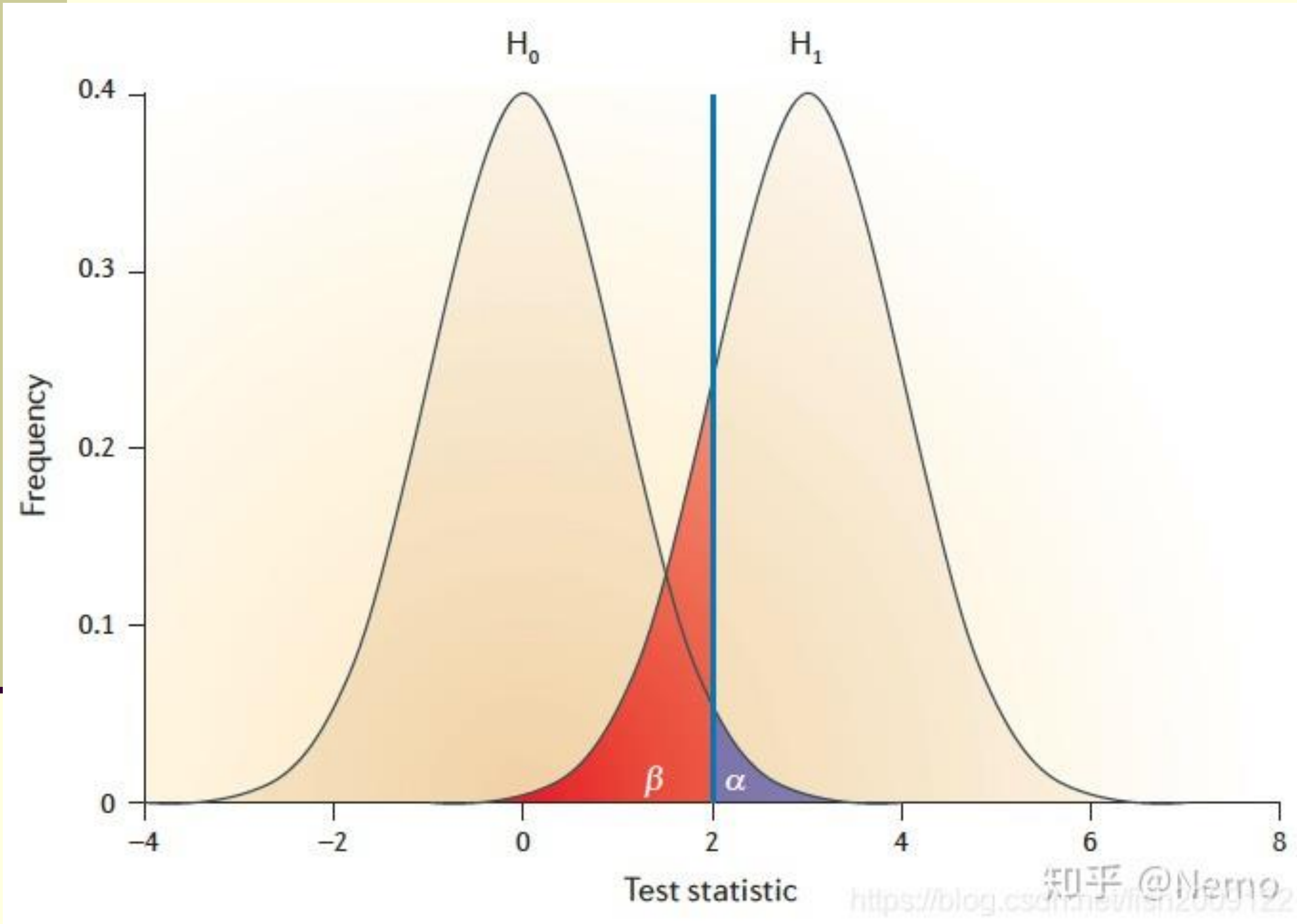
统计检验过程

陪审团审判		
裁决	实际情况	
	无罪	有罪
无罪	正确	错误
有罪	错误	正确

H_0 检验		
决策	实际情况	
	H_0为真	H_0为假
未拒绝H_0	正确决策 $(1-\alpha)$	第II类错误(β)
拒绝H_0	第I类错误(α)	正确决策 $(1-\beta)$

α 错误和 β 错误的关系

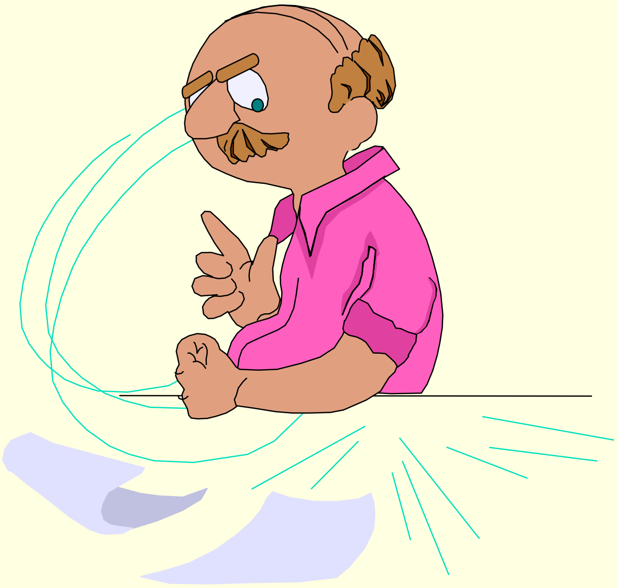




H_0
H_1
Frequency
0.4
0.3
0.2
0.1
0
β
α
−4
−2
0
2
4
6
8
Test statistic
知乎 @Nemo

影响 β 错误的因素

- 1. 总体参数的真值
  - 随着假设的总体参数的减少而增大
- 2. 显著性水平 α
  - 当 α 减少时增大
- 3. 总体标准差 σ
  - 当 σ 增大时增大
- 4. 样本容量 n
  - 当 n 减少时增大



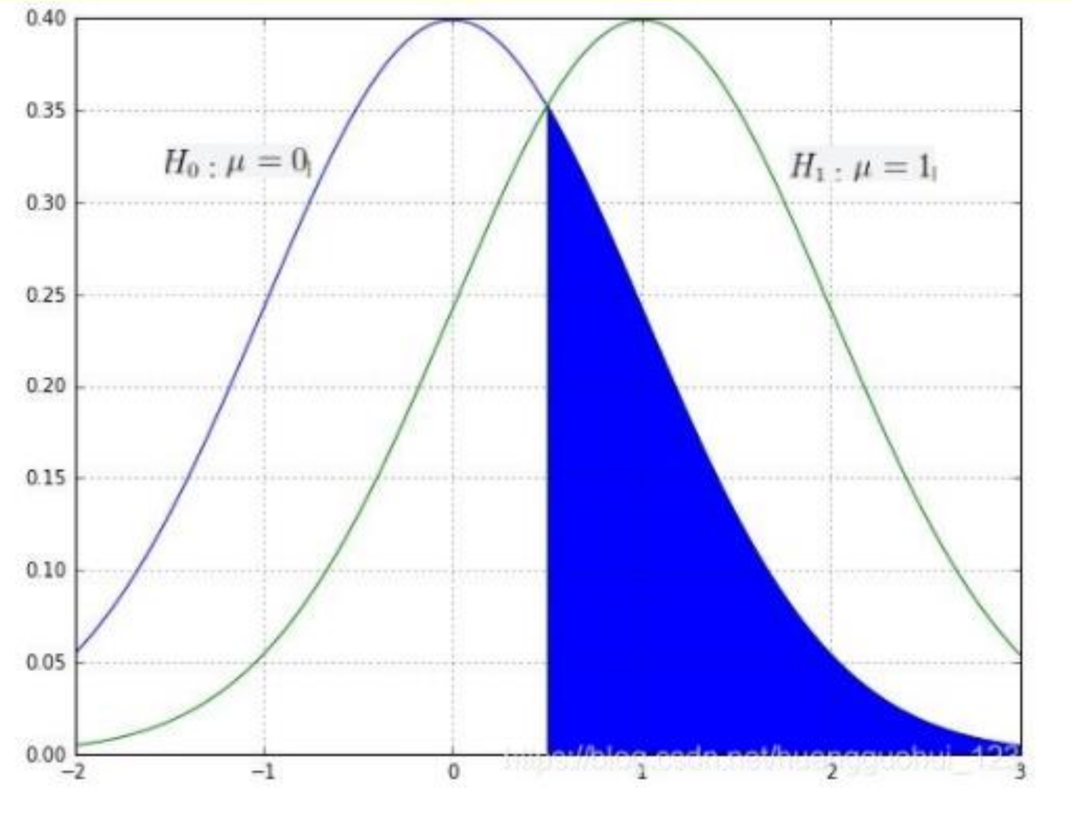
显著性水平α
(significant level)

- 1. 是一个概率值
- 2. 原假设为真时，拒绝原假设的概率
  - 被称为抽样分布的拒绝域
- 3. 表示为 α (alpha)
  - 常用的 α 值有0.01, 0.05, 0.10
- 4. 由研究者事先确定

两类错误为什么不能同时变小

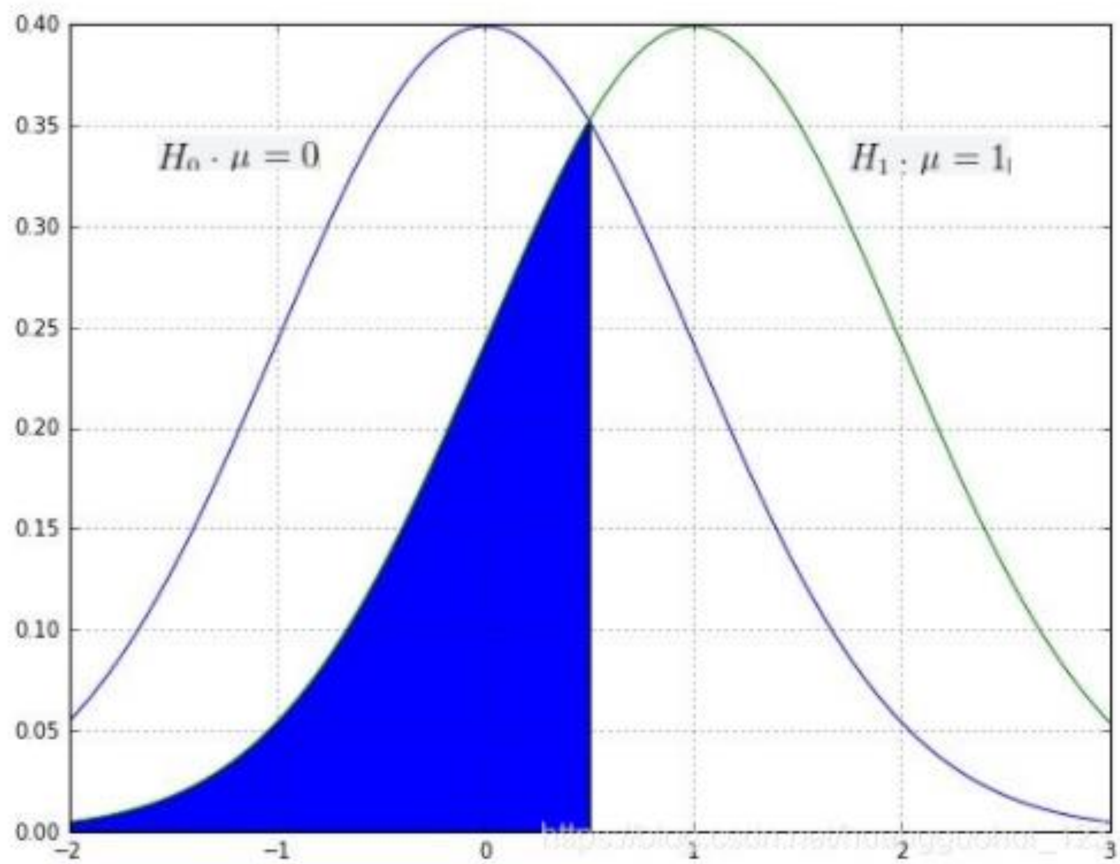
假如理论告诉我们参数μ只可能在0和1之间取值，检验统计量服从$N(\mu,1^2)$分布。检验统计量可取值于整个R，我们想要划分R为R_0，R_1，使得落在R_0时接受$H_0:\mu=0$的假设，而落在R_1时接受$H_1:\mu=1$的假设。

现在问题是如何划分R_0，R_1。事实上不同的原则可以导出不同的划分方法，其中每一个都是不同的检验。其中每种检验都对应相应的的第一类第二类错误概率值，所谓第一类第二类错误值不能同时减小，指的就是在这一阶段不可能通过单纯改变R_0，R_1的划分方式来同时降低两类错误概率。



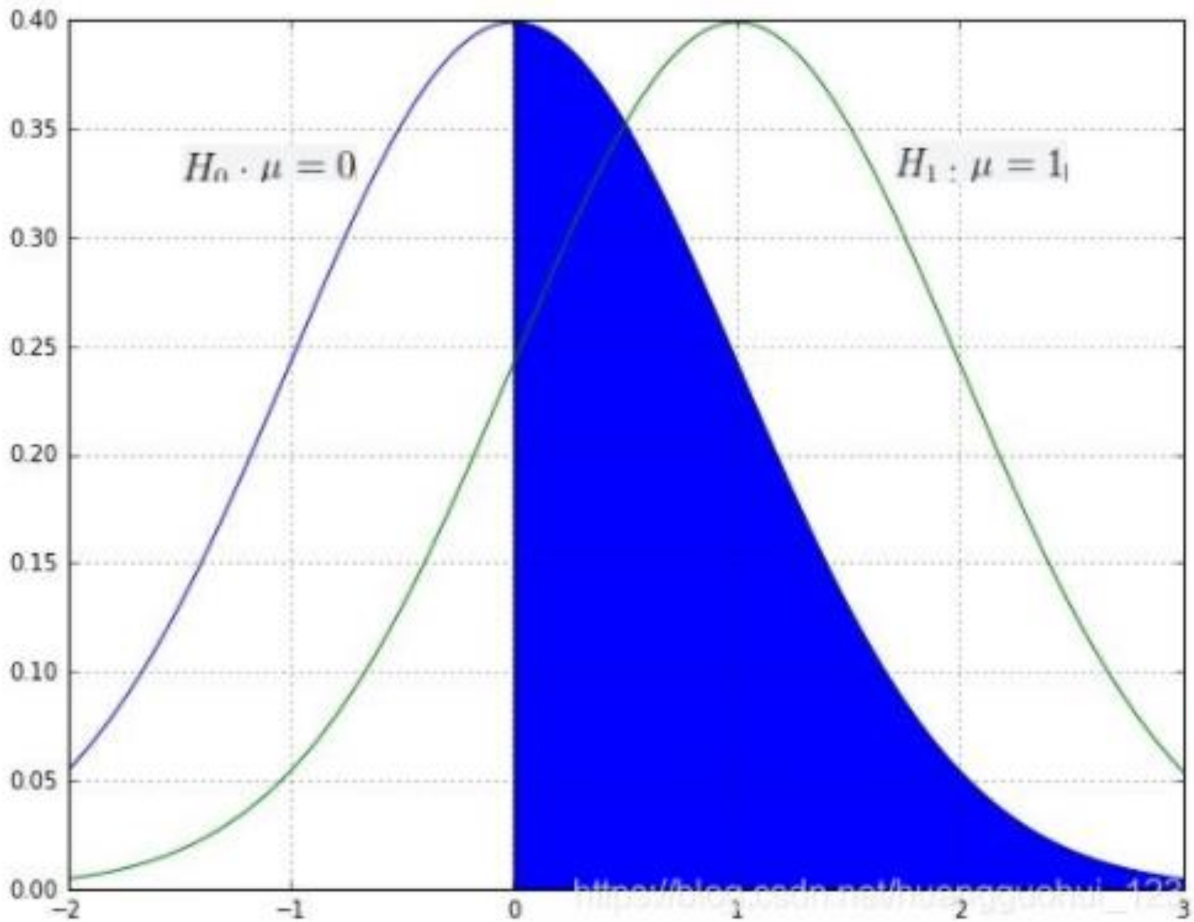
两类错误为什么不能同时变小

当μ的确为1时，错误概率（第二类错误发生的概率β）为落在1/2左边,如下图所示:



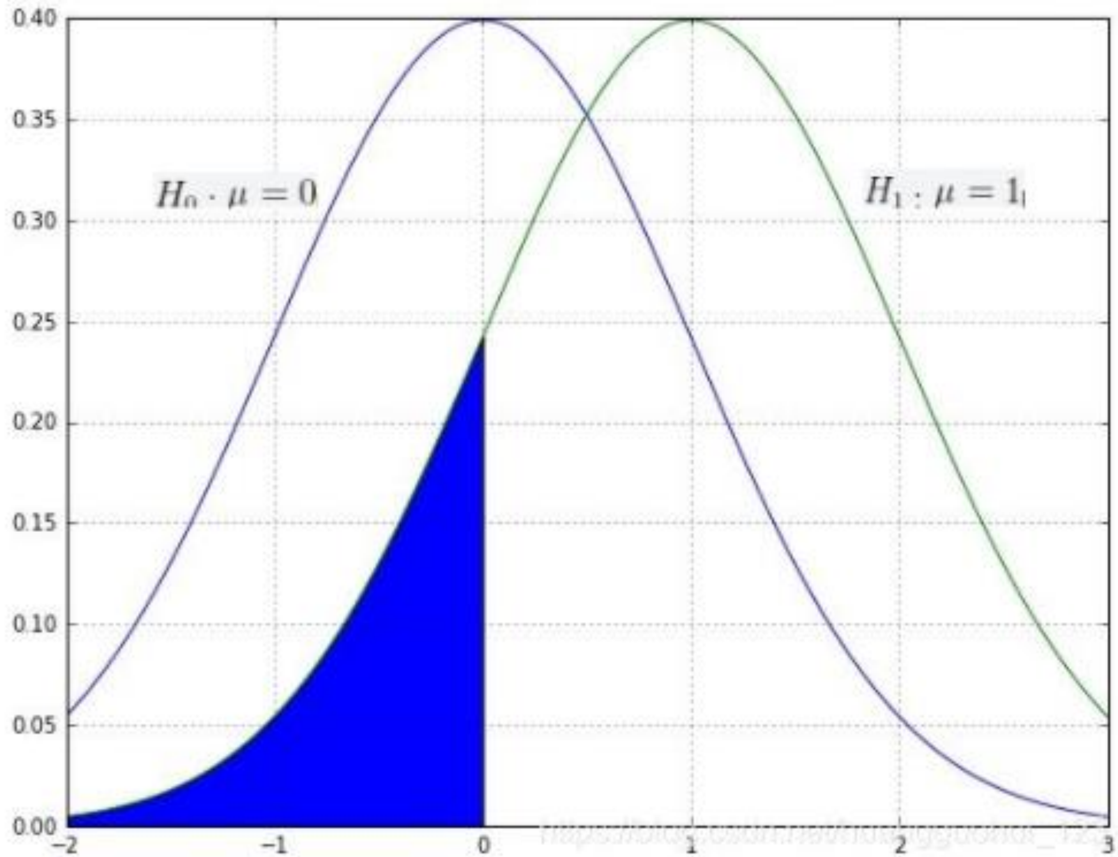
两类错误为什么不能同时变小

将分割线移到0时，$H_0 : \mu = 0$错误率升为



两类错误为什么不能同时变小

此时，$H_1:\mu=1$时错误率降为



由以上分析可获结论：对于某一具体的检验来说，当样本量n一定时，α越小β越大，α越大β越小。

势函数（Power）

- 定义：样本观察值落入拒绝域的概率为检验的势函数。

$$g(\theta)=P_{\theta}\{x\in W\}\qquad \theta\in\Theta$$

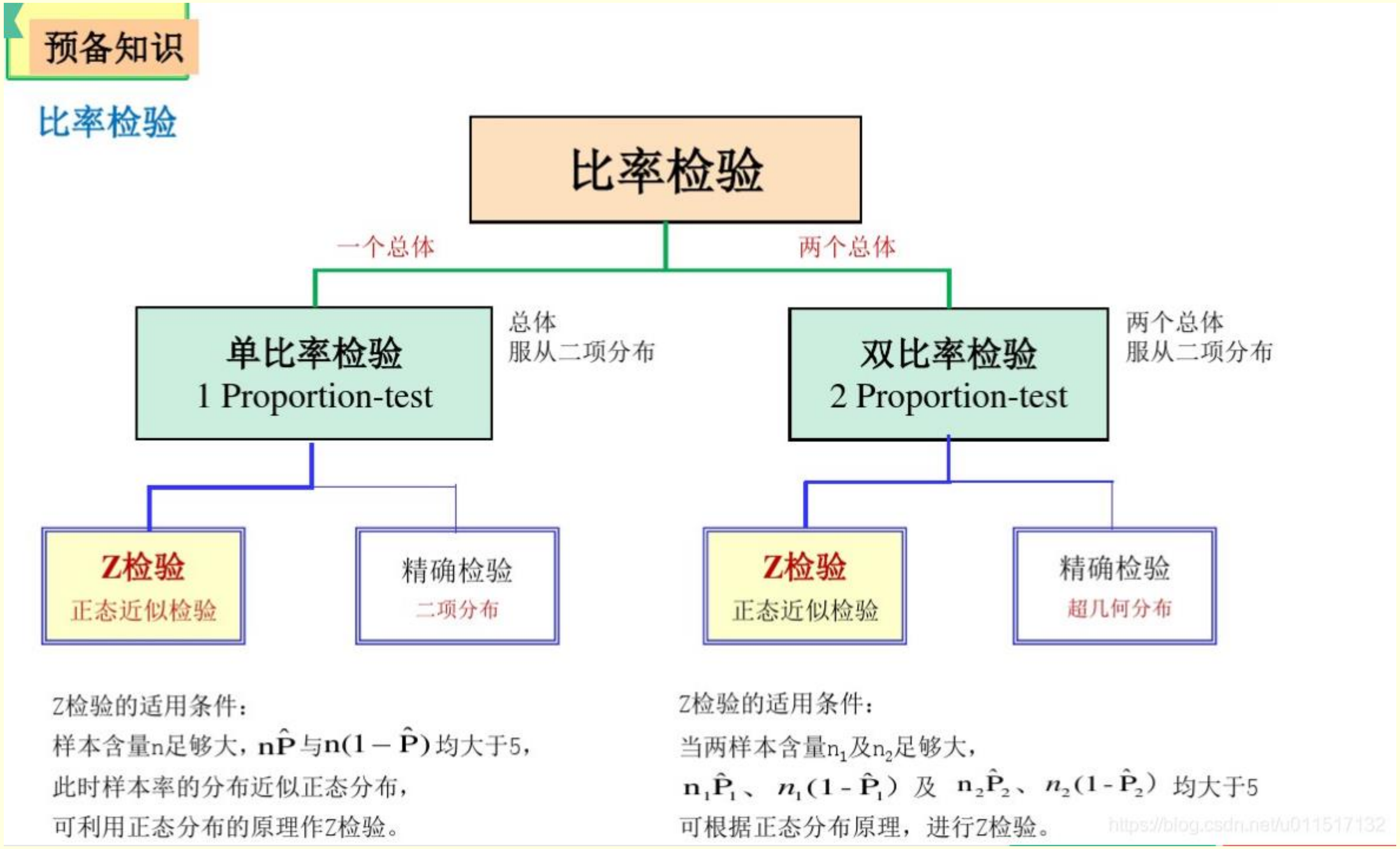
- 势函数与检验原则的关系

$$g(\theta)=\begin{cases}P_{\theta_0}\{x\in W\}=\alpha & \theta\in\Theta_0 \text{原假设为真}\\ P_{\theta_1}\{x\in W\}=1-\beta & \theta\in\Theta_1 \text{原假设为假}\end{cases}$$

- 用势函数判断检验统计量优良性的标准

单比率检验

- 对于A/B test中两组人群的对比中，我们需要对比的是ctr，转化率等指标。而ctr，用户转化率等指标，都是0-1分布，即二项分布。因此可以使用比率检验的方法进行假设检验。如果A/B test中包含多组人群，可以两两进行比较，也可以直接利用方差分析组间差异的判断。



单比率检验

现在有这样一种情景，我们新发布了一个版本或新上了一个活动，并选了一批人进行试验，我们想要知道发布了这个版本或新上活动后新的样本是否和原来有明显差异。我们可以使用单比率检验。（与平常假设检验无差别。构造统计量，看统计量是否在拒绝域内。正常是T统计量，这里由于是二项分布，n*p>5时可以认为是正态分布，即Z统计量）

1. 假定条件
   - 总体服从二项分布
   - 可用正态分布来近似(大样本)
2. 检验的 *z* 统计量

$$z=\frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}\sim N(0,1)$$

单比率检验

文档

单比率检验

单比率检验用于根据样本数据对总体比率进行推断

单比率检验
1 Proportion-test

Z检验
正态近似检验

样本含量n足够大
$n\hat{P} \geq 5$
$n(1-\hat{P}) \geq 5$

	双侧检验	左侧检验	右侧检验
检验假设	$H_0: p = p_0$ $H_1: p \neq p_0$	$H_0: p \geq p_0$ $H_1: p < p_0$	$H_0: p \leq p_0$ $H_1: p > p_0$
拒绝域	$\lvert Z \rvert \geq Z_{1-a/2}$	$Z \leq Z_a$	$Z \geq Z_{1-a}$
$P_{值}$ 决策	$P_{值} < \alpha$ 拒绝H_0		

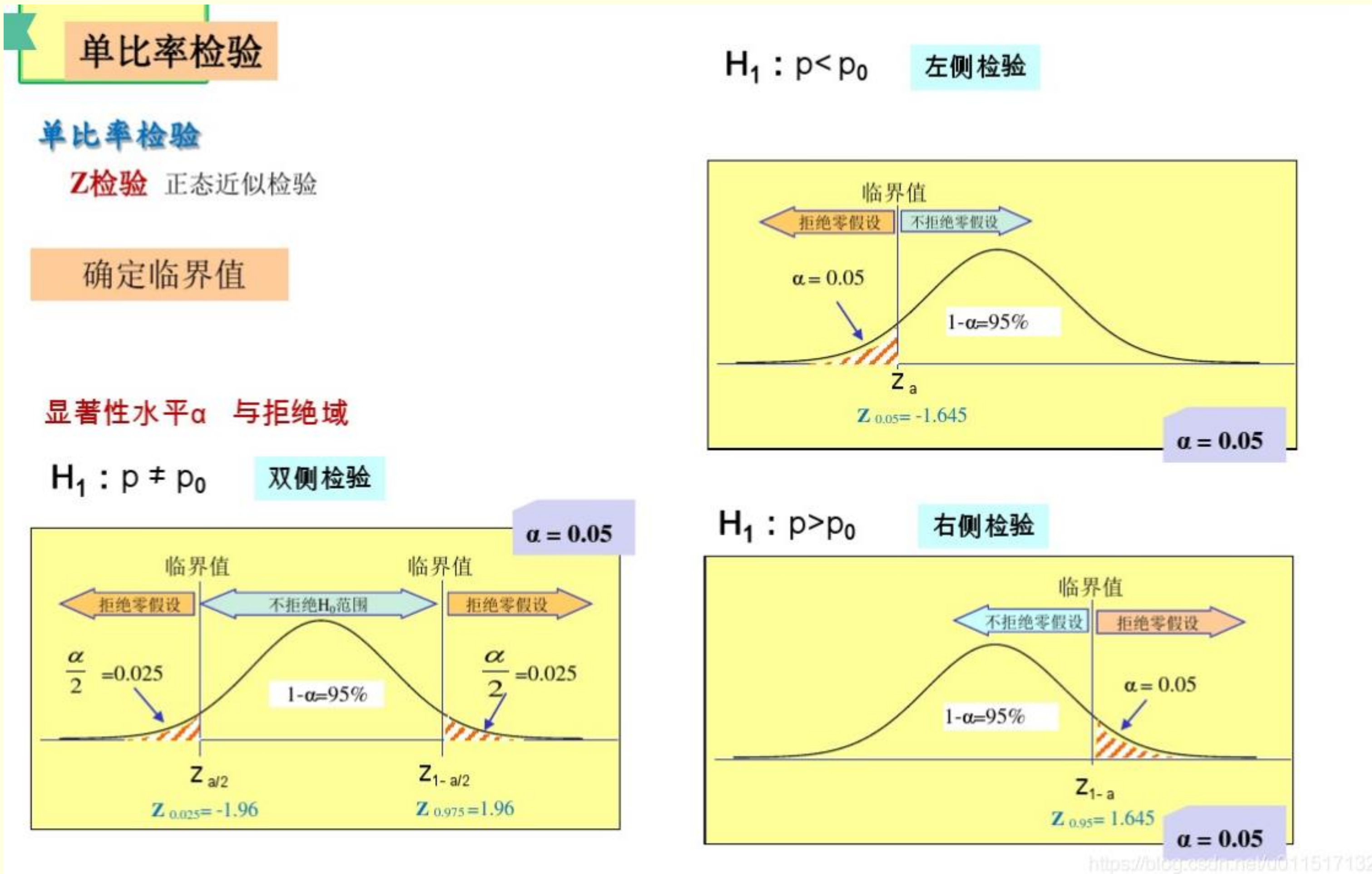
统计量

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

式中：
n ：样本数
$\hat{P}$ ：样本的比率
p_0 ：比率参考值

样本比率 $\hat{P} = x \div n$
其中x是观察到的"成功"数

单比率检验



单比率检验
单比率检验
Z检验 正态近似检验
确定临界值
显著性水平α 与拒绝域
H1：p ≠ p0
双侧检验
α = 0.05
临界值
临界值
拒绝零假设
不拒绝H0范围
拒绝零假设
α/2 =0.025
α/2 =0.025
1-α=95%
Z a/2
Z1- a/2
Z 0.025= -1.96
Z 0.975 =1.96
H1：p< p0
左侧检验
临界值
拒绝零假设
不拒绝零假设
α = 0.05
1-α=95%
Z a
Z 0.05= -1.645
α = 0.05
H1：p>p0
右侧检验
临界值
不拒绝零假设
拒绝零假设
α = 0.05
1-α=95%
Z1- a
Z 0.95= 1.645
α = 0.05

单比率检验

单比率检验

假设检验的例子（16） ——双侧检验

我们长园集团有个公司的一台注塑机加工某种电缆附件产品，长期以来生产过程的不合格品率 $p_0 = 2\%$，估计当前生产过程的不合格品率仍为2%。

随机抽取500个产品，测量得到不合格品数为9。

n=500
x =9

本例

样本比率提供了总体比率的估计值

样本比率 $\hat{P}$ =x÷n =9÷500 = 0.018

比率参考值 p_0 = 0.02（2%）

1- p =0.982，

① 建立检验假设

$H_0: p = 0.02$

$H_1: p \neq 0.02$

② 给定显著水平

$\alpha = 0.05$

③ 计算统计量

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.018 - 0.02}{\sqrt{\frac{0.02(1-0.02)}{500}}}$$

$= -0.3194$

检验的统计功效

- 在上面我们阐述了在显著性水平α 下一组样本的统计指标是否与原来存在差异的方法。在概率统计中，我们知道，假设检验中有两类错误，第一类错误是“弃真”，即当零假设正确时，我们拒绝的概率，记为α ；第二类错误是“收伪”，即零假设错误时，我们却没有拒绝的概率记为β 。由定义可知上面的α 和β 都是关于零假设的条件概率，实际上我们所说的显著性水平对应的就是第一类错误概率α 。现在假设我们再比较一组样本与另一组是否存在差异时，我们拒绝了零假设，即认为两组有差异，我们需要进一步知道我们正确拒绝了零假设的概率power，我们把这个概率叫做统计功效。实际上统计功效（势）power就是零假设错误时，我们却没有拒绝的概率（第二类错误概率）。即power=1−β 。

检验的统计功效

文档

单比率检验 ——双侧检验

检验功效和样本数量分析

假设检验的例子16, Z的绝对值=0.319小于临界值$Z_{0.025}$=1.96 ($P_{值}$= P = 0.7494 > α = 0.05) 出了不拒绝零假设的统计结论。

评价检验功效 ——当H0 为假时正确否定它的概率 ($p = 1 - \beta$)

双侧检验

$$\text{Power} = 1 - \Phi\left(\frac{p_0 - \hat{p} + Z_{\alpha/2}\sigma_P}{S_P}\right) + \Phi\left(\frac{p_0 - \hat{p} - Z_{\alpha/2}\sigma_P}{S_P}\right)$$

Φ : 标准正态分布的累积分布函数

将 P_0 =0.02 、$\hat{P}$ =0.018 、$Z_{\alpha/2}$ = 1.96（右尾概率分位数、当 α = 0.05 ）及 σ_p= 0.0063、S_p= 0.0059 代入上式 。

$z_{\alpha/2}$= upper α/2 point of the standard normal distribution

$$\begin{aligned}\text{Power} &= 1 - \Phi[(0.02-0.018+1.96 \times 0.0063)/0.0059] \\ &\quad + \Phi[(0.02-0.018-1.96 \times 0.0063)/0.0059] \\ &= 1 - \Phi(2.4319) + \Phi(-1.7539) \\ &= 0.0075 + 0.03972 = 0.0472\end{aligned}$$

检验 功效 Power = 0.0472

参考比率标准误

$$\sigma_p = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.02(1-0.02)}{500}} = 0.0063$$

样本比率标准误

$$S_p = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.018(1-0.018)}{500}} = 0.0059$$

计算所需样本量

- 我们总会预先设定一个显著性水平α 和统计功效β 。根据上面统计功效的公式，实际上我们可以反推出我们需要的样本量

检验功效和样本数量分析

假设检验的例子16中，如果总体比率实际为0.02 但在样本比率=0.018时，则检测到差异的可能性为4 .72%

如果我们仍然规定可以检测到的最小差值 δ =0.002，并希望功效 Power =0.9需要抽取产品样本多少个?

样本数量

双侧检验

$$n = \frac{(Z_{\alpha/2} + Z_{\beta})^2}{(2\sin^{-1}\sqrt{P_0} - 2\sin^{-1}\sqrt{\hat{P}})^2}$$

公式中

$\sin^{-1}\theta$：反正弦

三角函数采用弧度计算

将 $Z_{\alpha/2} = Z_{0.05/2} = 1.96$ 、 $Z_{\beta} = Z_{0.1} = 1.28$ 及 $P_0 = 0.02$ 、 $\hat{P} = 0.018$ 代入上式

$$n = \frac{(1.96 + 1.28)^2}{(2\sin^{-1}\sqrt{0.02} - 2\sin^{-1}\sqrt{0.018})^2} = \frac{10.4976}{(2\times 0.141897 - 2\times 0.13457)^2}$$

$= 48882.71733$

需要抽取产品样本**48883**个

容许差值越小，需要样本量越大。（为使差值符合选择，δ有时需主观规定）

特征函数的定义

在某些实际问题中, 我们除了希望控制犯第I类错误的概率外, 往往还希望控制犯第 II类错误的概率.

以上在进行假设检验时, 总是根据问题的需要, 预先给出显著性水平以控制犯第I类错误的概率, 而犯第 II类错误的概率则依赖于样本容量的选择.

在本节中, 我们将阐明如何选取样本的容量使得犯第 II类错误的概率控制在预先给定的限度内, 为此, 引入**施行特征函数(β)**.

施行特征函数的定义:

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的施行特征函数或OC函数, 其图形称为OC曲线.

施行特征函数的作用:

适当地选取样本的容量, 使得犯第 II类错误的概率控制在预先给定的限度内.

二、Z 检验法的OC 函数

1. 右边检验问题

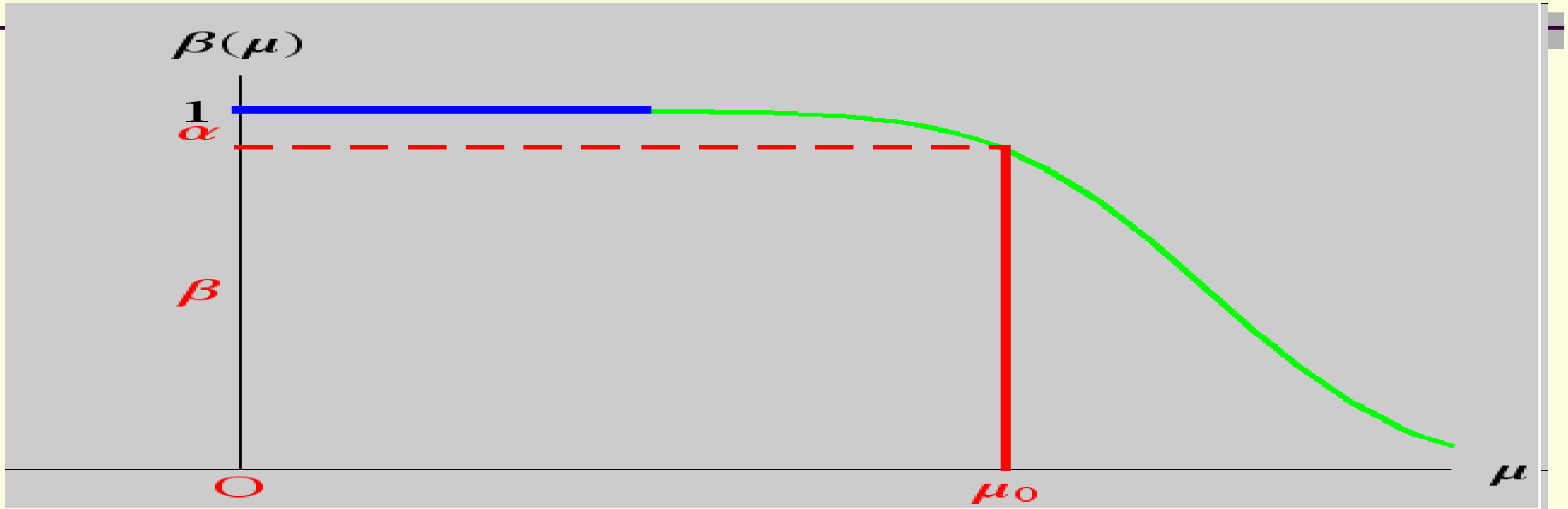
假设检验 $H_0: \mu \le \mu_0, H_1: \mu > \mu_0$的$OC$ 函数是

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

$$= P_\mu\left\{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_\alpha - \frac{\mu-\mu_0}{\sigma/\sqrt{n}}\right\}$$

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

此OC 函数的图形如下:



此OC 函数的性质如下:

(1) 它是 $\lambda = \dfrac{\mu - \mu_0}{\sigma / \sqrt{n}}$ 的单调递减连续函数;

(2) $\lim\limits_{\mu \to \mu_0^+} \beta(\mu) = 1 - \alpha,\quad \lim\limits_{\mu \to \infty} \beta(\mu) = 0.$

根据 OC 函数 $\beta(\mu)$ 可以确定样本容量 n,

使当真值 $\mu \ge \mu_0 + \delta$ ($\delta > 0$ 为取定的值)时,

犯第 II类错误的概率不超过给定的 β.

因为 $\beta(\mu)$ 是 μ 的递减函数,

故当 $\mu \ge \mu_0 + \delta$ 时, $\beta(\mu_0 + \delta) \ge \beta(\mu)$,

于是只要 $\beta(\mu_0 + \delta) = \Phi\left(z_\alpha - \frac{\sqrt{n}\delta}{\sigma}\right) \le \beta$,

即 n 满足 $z_\alpha - \frac{\sqrt{n}\delta}{\sigma} \le -z_\beta$, 只要 $\sqrt{n} \ge \frac{(z_\alpha + z_\beta)\sigma}{\delta}$,

就能使犯第 II类错误的概率不超过给定的 β.

2. 左边检验问题

假设检验$H_0: \mu \geq \mu_0, H_1: \mu < \mu_0$的$OC$函数是

$$\beta(\mu) = P_\mu(\text{接受}H_0) = \Phi(z_\alpha + \lambda), \quad \lambda = \frac{\mu - \mu_0}{\sigma / \sqrt{n}}.$$

当真值$\mu \geq \mu_0$时$\beta(\mu)$为作出正确判断的概率;

当真值$\mu < \mu_0$时$\beta(\mu)$给出犯第II类错误的概率.

只要样本容量n满足$\sqrt{n} \geq \dfrac{(z_\alpha + z_\beta)\sigma}{\delta}$

就能使犯第 II类错误的概率不超过给定的β.

3. 双边检验问题

假设检验 $H_0: \mu = \mu_0, H_1: \mu \neq \mu_0$ 的 OC 函数是

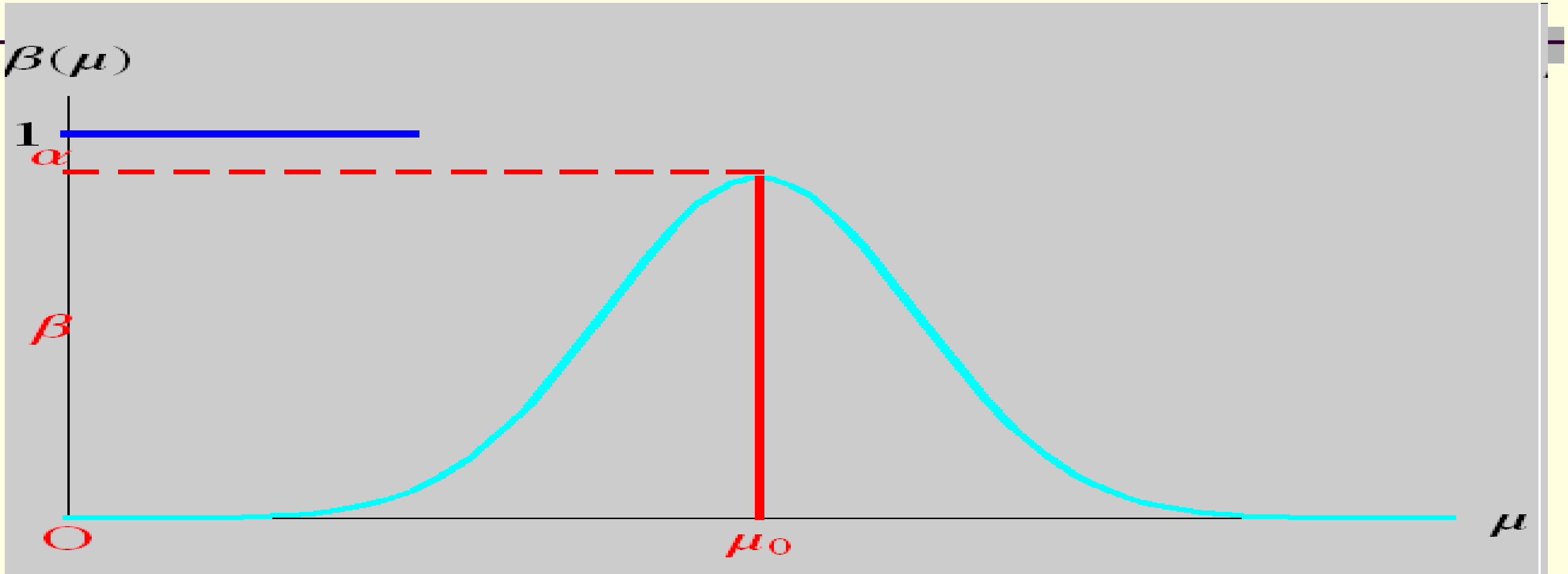
$$\beta(\mu) = P_\mu(\text{接受}H_0) = P_\mu\left\{-z_{\alpha/2} < \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2}\right\}$$

$$= P_\mu\left\{-\lambda - z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < -\lambda + z_{\alpha/2}\right\}$$

$$= \Phi(z_{\alpha/2} - \lambda) - \Phi(-z_{\alpha/2} - \lambda)$$

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

此OC 函数的图形如下:



只要样本容量n满足 $\sqrt{n} \geq \dfrac{(z_{\alpha/2}+z_{\beta})\sigma}{\delta}$,

就能使犯第 II类错误的概率不超过给定的β.

例1 (工业产品质量抽验方案) 设有一大批产品，产品质量指标 $x \sim N(\mu,\sigma^2)$. 以 μ 小者为佳，厂方要求所确定的验收方案对高质量的产品 $(\mu \le \mu_0)$ 能以高概率 $1-\alpha$ 为买方所接受. 买方则要求低质产品 $(\mu \ge \mu_0 + \delta)$ 能以高概率 $1-\beta$ 被拒绝. α, β 由买方和厂方协商给出. 并采取一次抽样确定该批产品是否为买方所接受. 问应怎样安排抽验方案. 已知 $\mu_0 = 120,\ \delta = 20,\ \sigma^2 = 900,\ \alpha, \beta$ 均取 0.05.

解　检验问题 $H_0: \mu \le \mu_0, H_1: \mu \ge \mu_0 + \delta,$

由 Z 检验,拒绝域仍为 $\dfrac{\overline{x} - \mu_0}{\sigma / \sqrt{n}} \ge z_\alpha.$

OC函数 $\beta(\mu) = P_\mu\left\{\dfrac{\overline{X} - \mu_0}{\sigma / \sqrt{n}} < z_\alpha\right\}$

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

现在要求当 $\mu \ge \mu_0 + \delta$ 时, $\beta(\mu) \le \beta.$

因为 $\beta(\mu)$ 是 μ 的递减函数, 只需 $\beta(\mu_0 + \delta) = \beta.$

由于 $\sqrt{n} \geq \dfrac{(z_\alpha + z_\beta)\sigma}{\delta}$,

根据给定的数据知 $n \geq 24.35$, 故取 $n = 25$.

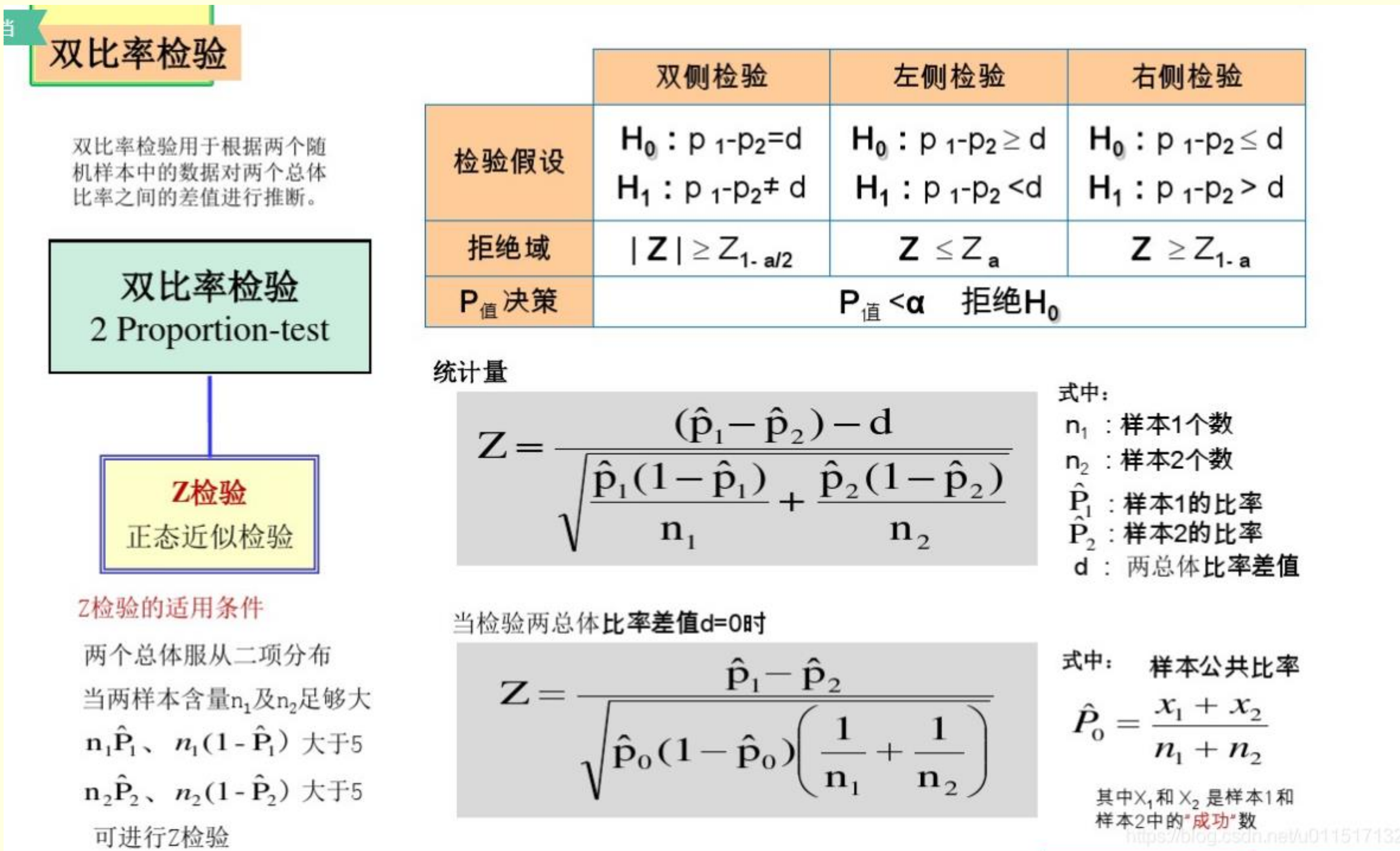
且当$\bar{x}$的观察值满足$\dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_\alpha = z_{0.05} = 1.645$时,

即$\bar{x} \geq 129.87$时,买方就拒绝这批产品,

而当$\bar{x} < 129.87$时,买方就接受这批产品.

双比率检验

- 在**A/B test**是，我们同上的做法是在同一层试验中，选用两个或多个版本（活动）进行同时实现，此时我们需要比较两组样本的差异性。于是我们就需要用到两样本双比率检验。



	双侧检验	左侧检验	右侧检验
检验假设	$H_0: p_1-p_2=d$ $H_1: p_1-p_2\neq d$	$H_0: p_1-p_2\geq d$ $H_1: p_1-p_2<d$	$H_0: p_1-p_2\leq d$ $H_1: p_1-p_2>d$
拒绝域	$\lvert Z\rvert \geq Z_{1-a/2}$	$Z \leq Z_a$	$Z \geq Z_{1-a}$
$P_{值}$决策	$P_{值}<\alpha$ 拒绝H_0		

统计量

$$Z=\frac{(\hat{p}_1-\hat{p}_2)-d}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

式中：
n_1 ：样本1个数
n_2 ：样本2个数
$\hat{P}_1$ ：样本1的比率
$\hat{P}_2$ ：样本2的比率
d ：两总体比率差值

当检验两总体比率差值d=0时

$$Z=\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

式中： 样本公共比率

$$\hat{P}_0=\frac{x_1+x_2}{n_1+n_2}$$

其中X_1和X_2是样本1和样本2中的"成功"数

双比率检验

双比率检验

——双侧检验

假设检验的例子（19）

有两台波峰焊设备，1号设备生产的板件中随机抽取1600个产品，得到优等品的件数为320；2号设备生产的板件中随机抽取2000个产品，得到优等品的件数为360。

为确定两台波峰焊设备产出优等品率是否一致，采用双比率检验。

用显著性水平α=0.05 进行检验

样本1的比率

$\hat{P}_1 = x_1 \div n_1 = 320 \div 1600 = 0.2$

样本2的比率

$\hat{P}_2 = x_2 \div n_2 = 360 \div 2000 = 0.18$

$n_1\hat{P}_1$、$n_1(1-\hat{P}_1)$ 大于5
$n_2\hat{P}_2$、$n_2(1-\hat{P}_2)$ 大于5

可用正态近似检验

检验两总体**比率差值** $d = p_1 - p_2 = 0$

两总体"成功"比率合并估计**值**:

$$\hat{P}_0 = \frac{x_1 + x_2}{n_1 + n_2} = \frac{320 + 360}{1600 + 2000} = 0.189$$

① 建立检验假设

$H_0: p_1 - p_2 = 0$ （两个总体比率相等）

$H_1: p_1 - p_2 \neq 0$ （两个总体比率存在差异）

② 给定显著水平

$\alpha = 0.05$

③ 计算统计量

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}_0(1-\hat{P}_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{0.2 - 0.18}{\sqrt{0.189(1-0.189)\left(\frac{1}{1600} + \frac{1}{2000}\right)}}$$

$= 1.523$

双比率检验

双比率检验

接上页

——双侧检验

④ 查临界值

显著性水平　α =0 .05、 α /2=0 .025

反查正态分布表（右尾概率）

$Z_{临界值}$为：　$Z_{0.025}$=1.96

Z-Value	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.3	0.382	0.378	0.374	0.371	0.367	0.363	0.359	0.356	0.352	0.348
0.4	0.345	0.341	0.337	0.334	0.330	0.326	0.323	0.319	0.316	0.312
…	…	…	…	…	…	…	…	…	…	…
1.9	0.029	0.028	0.027	0.027	0.026	0.026	0.025	0.024	0.024	0.023
2.0	0.023	0.022	0.022	0.021	0.021	[illegible]	0.020	0.019	0.019	0.018

$Z_{0.025}$=1.96

α /2=0 .025
双侧检验
临界值=±1.96

⑤ 用算得的统计量与相应的临界值作比较

∵　| Z |=1.523< $Z_{0.025}$=1.96

作出不拒绝零假设的统计结论
两台波峰焊设产出优等品率是一致的

◆ 计算检验 P-值

查正态分布表 | Z | =1.52　（右尾概率）

Z-Value	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.3	0.382	0.378	0.374	0.371	0.367	0.363	0.359	0.356	0.352	0.348
0.4	0.345	0.341	0.337	0.334	0.330	0.326	0.323	0.319	0.316	0.312
…	…	…	…	…	…	…	…	…	…	…
1.4	0.081	0.081	0.081	0.081	0.081	0.081	0.081	0.081	0.081	0.081
1.5	0.067	0.067	0.067	0.067	0.067	0.067	0.067	0.067	0.067	0.067

$P = P_{(Z>1.52)} = 0.067$

$P = P_{(Z<-1.52 \text{ 及 } Z>1.52)}$
$= 0.067 \times 2 = 0.134$

∵ $P_{值}$= 0.134 > α = 0.05

按α =0. 05 的水平**无法拒绝零假设H0**

两台波峰焊设产出优等品率是一致的

| Z | ≤临界值与$P_{值}$>0. 05是对应的

统计功效

档

双比率检验

——右单侧检验

假设检验的例子（20）

有两台相同设备，1号设备和2号设备生产相同产品且质量无差异，工程师在2号设备换上新设计的刀具，由于新刀具使用价格较贵的合金材料；希望使用新刀具能使废品率有所下降，预计废品率至少会降低0.5%。

使用原来刀具的1号设备生产的零件中和换上新刀具的2号设备生产的零件中各分别随机抽取3000个；经测量得出：1号设备生产的零件中废品数是75个，2号设备生产的零件中废品数是21个。

检验使用新刀具使废品率至少会降低0.5%

样本 $n_1 = n_2 = n = 3000$

样本1的比率

$$\hat{P}_1 = x_1 \div n_1 = 75 \div 3000 = 0.025$$

样本2的比率

$$\hat{P}_2 = x_2 \div n_2 = 21 \div 3000 = 0.007$$

$n_1\hat{P}_1$、$n_1(1-\hat{P}_1)$ 大于5
$n_2\hat{P}_2$、$n_2(1-\hat{P}_2)$ 大于5 } 可用正态近似检验

我们将预期效果（想要证明的假设）作为备择假设H_1，因为只有当检验结果与原假设有明显差别时才能拒绝原假设而接受备择假设，拒绝是有说服力的，减少结论错误。

① 建立检验假设

$H_0: p_1 - p_2 \le 0.005$

$H_1: p_1 - p_2 > 0.005$ （想要证明）

② 给定显著水平

$\alpha = 0.05$

③ 计算统计量 （两总体比率差值d不为时）

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - d}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

$$= \frac{(0.025 - 0.007) - 0.005}{\sqrt{\frac{0.025(1-0.025)}{3000} + \frac{0.007(1-0.007)}{3000}}}$$

$Z = 4.023$

统计功效

双比率检验

——右单侧检验

样本 $n_1 = n_2 = n = 3000$

检验功效和样本数量分析

计算在假设检验的例子20中的检验功效，先计算“使用原来刀具比换上新刀具生产的零件废品率多”的检验功效

评价检验功效①

——当H0 为假时正确否定它的概率 ($p = 1 - \beta$)

——右单侧检验 $H_1 : p_1 > p_2$

$$\text{Power} = 1 - \Phi\left(\frac{\hat{p}_2 - \hat{p}_1 + Z_\alpha \sqrt{\dfrac{2\hat{p}_c(1-\hat{p}_c)}{n}}}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n}}} \right)$$

$$\hat{p}_c = \frac{\hat{p}_1 + \hat{p}_2}{2}$$

z_α = one-sided critical value

(upper point of the standard normal distribution)

将 $\hat{p}_1 = 0.025$、$\hat{p}_2 = 0.007$ 、$Z_\alpha = 1.645$ 及右面计算值代入上式

$$\hat{p}_c = \frac{\hat{p}_1 + \hat{p}_2}{2} = \frac{0.025 + 0.007}{2} = 0.016$$

$$\frac{2\hat{p}_c(1-\hat{p}_c)}{n} = \frac{0.016(1-0.016)}{3000} = 0.00229$$

$$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}} = \sqrt{\frac{0.025(1-0.025)}{3000} + \frac{0.007(1-0.007)}{3000}} = 0.00323$$

Power $= 1 - \Phi\,[(0.007 - 0.025 + 1.645 \times 0.00229) \div 0.00323]$

$= 1 - \Phi\,[-4.40649]$

$= 0.99999$

统计功效

双比率检验

检验功效和样本数量分析

评价检验功效

样本数量n
是指每个组的

左单侧检验　$H_1: p_1 < p_2$

$$\text{Power} = \Phi\left(\frac{\hat{p}_2 - \hat{p}_1 - Z_\alpha\sqrt{\frac{2\hat{p}_c(1-\hat{p}_c)}{n}}}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}}\right) \qquad \hat{p}_c = \frac{\hat{p}_1 + \hat{p}_2}{2}$$

z_α= one-sided critical value (upper point of the standard normal distribution)

双侧检验　$H_1: p_1 \neq p_2$

$$\text{Power} = 1 - \Phi\left(\frac{\hat{p}_2 - \hat{p}_1 + Z_{\alpha/2}\sqrt{\frac{2\hat{p}_c(1-\hat{p}_c)}{n}}}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}}\right) + \Phi\left(\frac{\hat{p}_2 - \hat{p}_1 - Z_{\alpha/2}\sqrt{\frac{2\hat{p}_c(1-\hat{p}_c)}{n}}}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}}\right) \qquad \hat{p}_c = \frac{\hat{p}_1 + \hat{p}_2}{2}$$

$z_{\alpha/2}$= upper $\alpha/2$ point of the standard normal distribution

计算样本量

双比率检验

检验功效和样本数量分析

计算在假设检验的例子20中的**检验是**“使用原来刀具比换上新刀具生产的零件废品率多0. 5%以上”的 **检验**

使用原来刀具生产的零件废品率 $\hat{p}_1 = 0.025$

如果使用新刀具生产的零件废品率比原来低0. 013时, 差异可以检测到,我们规定基线 比率 $\hat{p}_2 = 0.012$

若希望检测功效 Power =0.9，需要抽取产品样本多少个?

样本数量①

单侧检验

左或右侧检验

$$n = \frac{2(Z_\alpha + Z_\beta)^2}{(2\sin^{-1}\sqrt{\hat{P}_1} - 2\sin^{-1}\sqrt{\hat{P}_2})^2}$$

公式中

sin⁻¹θ: 反正弦

三角函数采用弧度计算

将 Z_α = 1.645、Z_β = $Z_{0.1}$ =1.28 及 $\hat{p}_1 = 0.025$、$\hat{p}_2 = 0.012$ 代入上式

$$n = \frac{2(1.64 + 1.28)^2}{(2\sin^{-1}\sqrt{0.025} - 2\sin^{-1}\sqrt{0.012})^2} = \frac{2\times\cancel{10.4976}}{(2\times 0.15878 - 2\times 0.109765)^2}$$

= 2184.713656

需要抽取产品样本数量 n =2185个

样本 数量 n

是指每个组的

计算样本量

双比率检验

检验功效和样本数量分析

计算在假设检验的例子20中的检验是“使用原来刀具比换上新刀具生产的零件废品率多0.5%以上”的 检验

使用原来刀具生产的零件废品率　$\hat{p}_1 = 0.025$

如果使用新刀具生产的零件废品率比原来低0.005（0.5%）时,差异可以检测到,我们规定基线 比率 $\hat{p}_2 = 0.02$

若希望检测功效 Power =0.9，需要抽取产品样本多少个？

样本数量②

单侧检验

左或右侧检验

$$n = \frac{2(Z_\alpha + Z_\beta)^2}{(2\sin^{-1}\sqrt{\hat{P}_1} - 2\sin^{-1}\sqrt{\hat{P}_2})^2}$$

公式中

$\sin^{-1}\theta$: 反正弦

三角函数采用弧度计算

将 Z_α = 1.645、Z_β = $Z_{0.1}$ =1.28 及 $\hat{p}_1 = 0.025$、$\hat{p}_2 = 0.02$ 代入上式

$$n = \frac{2(1.64 + 1.28)^2}{(2\sin^{-1}\sqrt{0.025} - 2\sin^{-1}\sqrt{0.02})^2} = \frac{\boxed{2 \times 10.4976}}{(2 \times 0.15878 - 2 \times 0.141897)^2}$$

= 18414.18704

需要抽取产品样本数量 n = **18415**个

样本 数量 n

是指每个组的

计算样本量

双比率检验

32页上计算的**样本** 数量 n = 2185

样本数量

用另一个公式计算一下

双比率检验样本数量的估计公式（2）

此公式更为常用，不要求二项分布，分母是均值差，后面的相乘项是方差

$$n=\left(\frac{Z_\alpha+Z_\beta}{\hat{p}_1-\hat{p}_2}\right)^2[\hat{p}_1(1-\hat{p}_1)+\hat{p}_2(1-\hat{p}_2)]$$

将 Z_α = 1.645、Z_β = $Z_{0.1}$ =1.28 及 $\hat{p}_1=0.025$、$\hat{p}_2=0.012$ 代入上式

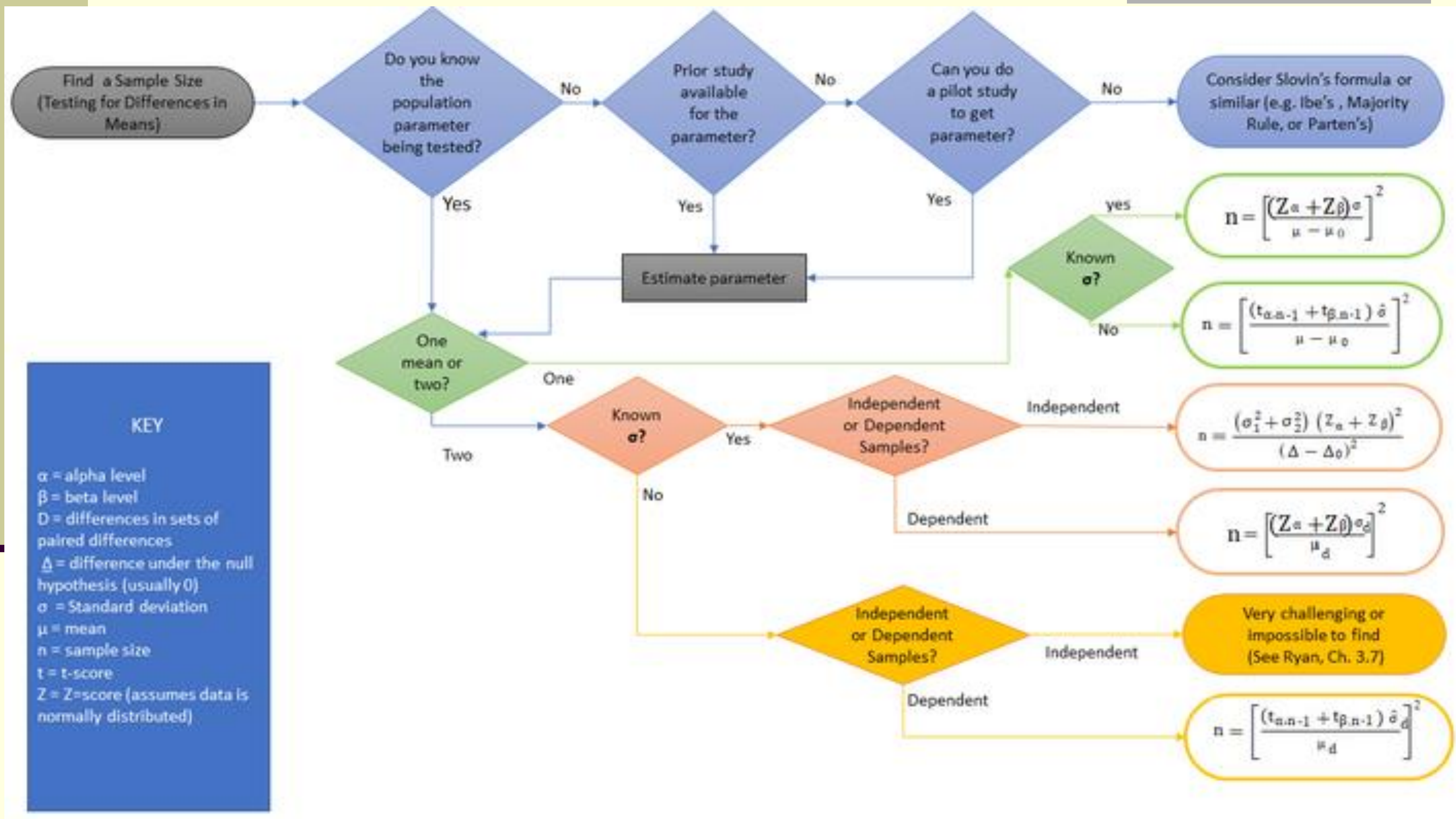
$$n=\left(\frac{1.645+1.28}{0.025-0.012}\right)^2[0.025(1-0.025)+0.012(1-0.012)]$$

= 1834.194

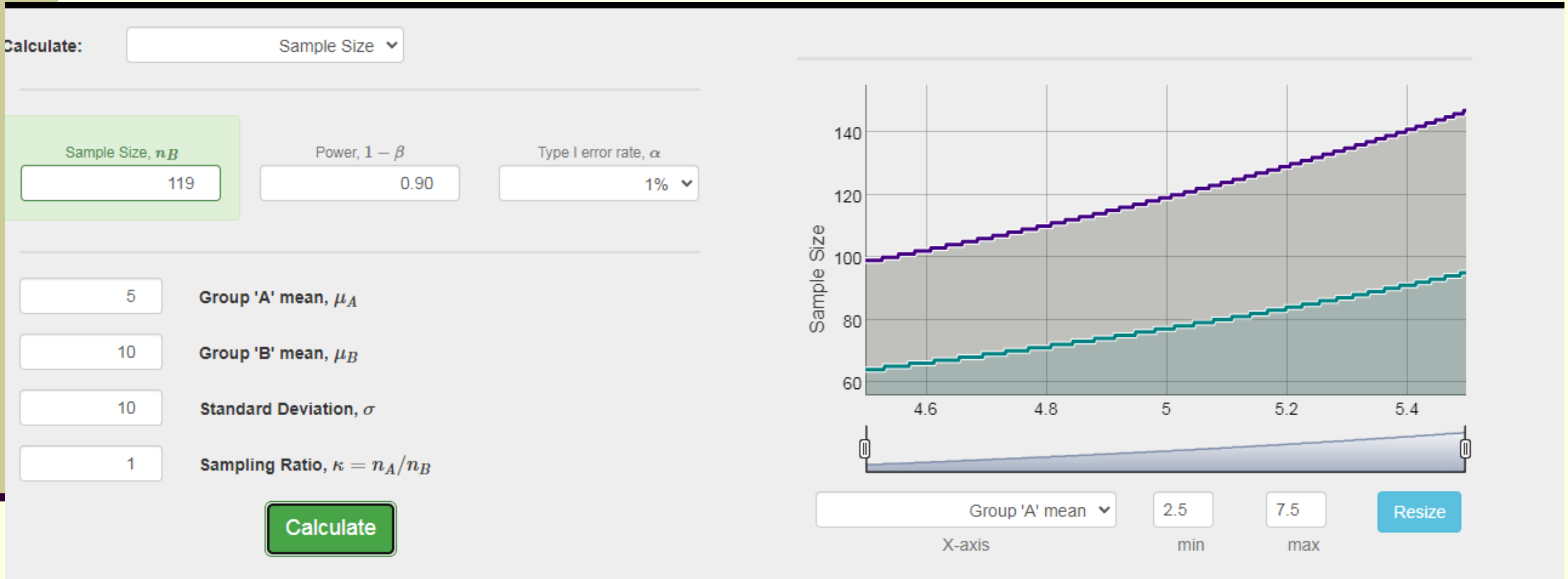
需要抽取产品样本数量 n = 1835个　　32页计算 ~~n = 2185~~

用不同的的估计公式计算的**样本** 数量会有差异

计算样本量



Find a Sample Size (Testing for Differences in Means)
Do you know the population parameter being tested?
No
Prior study available for the parameter?
No
Can you do a pilot study to get parameter?
No
Consider Slovin's formula or similar (e.g. Ibe's , Majority Rule, or Parten's)
Yes
Yes
Yes
Estimate parameter
One mean or two?
One
Known σ?
yes
$n = \left[\frac{(Z_\alpha + Z_\beta)\sigma}{\mu - \mu_0}\right]^2$
No
$n = \left[\frac{(t_{\alpha,n-1} + t_{\beta,n-1})\hat{\sigma}}{\mu - \mu_0}\right]^2$
Two
Known σ?
Yes
Independent or Dependent Samples?
Independent
$n = \frac{(\sigma_1^2 + \sigma_2^2)(Z_\alpha + Z_\beta)^2}{(\Delta - \Delta_0)^2}$
Dependent
$n = \left[\frac{(Z_\alpha + Z_\beta)\sigma_d}{\mu_d}\right]^2$
No
Independent or Dependent Samples?
Independent
Very challenging or impossible to find (See Ryan, Ch. 3.7)
Dependent
$n = \left[\frac{(t_{\alpha,n-1} + t_{\beta,n-1})\hat{\sigma}_d}{\mu_d}\right]^2$
KEY
α = alpha level
β = beta level
D = differences in sets of paired differences
Δ = difference under the null hypothesis (usually 0)
σ = Standard deviation
μ = mean
n = sample size
t = t-score
Z = Z=score (assumes data is normally distributed)


Calculate:
Sample Size
Sample Size, n_B
119
Power, $1-\beta$
0.90
Type I error rate, α
1%
5
Group 'A' mean, μ_A
10
Group 'B' mean, μ_B
10
Standard Deviation, σ
1
Sampling Ratio, $\kappa = n_A/n_B$
Calculate
Sample Size
140
120
100
80
60
4.6
4.8
5
5.2
5.4
Group 'A' mean
X-axis
2.5
min
7.5
max
Resize

Statsmodels计算power

- stats.power.FTestAnovaPower()
- stats.power.FTestPower()
- stats.power.GofChisquarePower()
- stats.power.NormalIndPower()
- stats.power.tt_ind_solve_power
- stats.power.tt_solve_power
- stats.power.TTestIndPower()
- stats.power.TTestPower()
- stats.power.zt_ind_solve_power

stats.power.FTestAnovaPower()

statsmodels.stats.power.FTestAnovaPower

`class statsmodels.stats.power.FTestAnovaPower(**kwds)` [source]

Statistical Power calculations F-test for one factor balanced ANOVA

Methods

plot_power ([dep_var, nobs, effect_size, ...])	plot power with number of observations or effect size on x-axis
power (effect_size, nobs, alpha[, k_groups])	Calculate the power of a F-test for one factor ANOVA.
solve_power ([effect_size, nobs, alpha, ...])	solve for any one parameter of the power of a F-test

Methods

plot_power ([dep_var, nobs, effect_size, ...])	plot power with number of observations or effect size on x-axis
power (effect_size, nobs, alpha[, k_groups])	Calculate the power of a F-test for one factor ANOVA.
solve_power ([effect_size, nobs, alpha, ...])	solve for any one parameter of the power of a F-test

class statsmodels.stats.power.TTestPower(***kwds*)

class statsmodels.stats.power.TTestIndPower

TTestIndPower(**kwds)	Statistical Power calculations for t-test for two independent sample
TTestPower(**kwds)	Statistical Power calculations for one sample or paired sample t-test
GofChisquarePower(**kwds)	Statistical Power calculations for one sample chisquare test
NormalIndPower([ddof])	Statistical Power calculations for z-test for two independent samples.
FTestAnovaPower(**kwds)	Statistical Power calculations F-test for one factor balanced ANOVA
FTestPower(**kwds)	Statistical Power calculations for generic F-test
normal_power_het(diff, nobs, alpha[, ...])	Calculate power of a normal distributed test statistic
normal_sample_size_one_tail(diff, power, alpha)	explicit sample size computation if only one tail is relevant
tt_solve_power([effect_size, nobs, alpha, ...])	solve for any one parameter of the power of a one sample t-test
tt_ind_solve_power([effect_size, nobs1, ...])	solve for any one parameter of the power of a two sample t-test
zt_ind_solve_power([effect_size, nobs1, ...])	solve for any one parameter of the power of a two sample z-test

随机模拟计算统计功效

Monte Carlo 方法计算检验的功效:

1. 选择一个特定的$\theta_1 \in \Theta$.
2. 对每个重复j, $j = 1, \cdots, m$.
   (a) 在对立假设空间$\theta = \theta_1$的情况下产生第j个随机样本 $x_1^{(j)}, \cdots, x_n^{(j)}$;
   (b) 基于第j个样本计算检验统计量T_j;
   (c) 记录决策结果$I_j = 1$,若H_0在显著性水平α下被拒绝; 否则 $I_j = 0$.
3. 计算显著的检验比例$\frac{1}{m}\sum_{j=1}^{m} I_j$, 此比例即为观测到的一型错误率.

作业

- 用模拟法比较三种正态性检验：Shapiro-Wilk检验、k-s检验和偏度检验的功效（power）。