

A/B tests allow you to determine scientifically how to optimize a website or a mobile app by trying out possible changes and seeing what performs better with your users.

course structure:

- ① overview:
  - example : choose a metric
  - review statistics
  - design
  - analyze
- ② how to protect the participants of your experiment, and what questions you should be asking yourself regarding the ethicality of experiments
- ③ choosing and characterizing metrics to use when evaluating your experiment.
- ④ designing the experiment, which includes choosing which users go

Definition:

A/B testing is a general methodology used online when you want to test out a new product or a feature.

2 sets of users {  
    1 set (control set) : existing product / feature  
    experiment

⇒ how did these users respond differently, in order to determine which version of your feature is better?

It is really useful for helping you climb to the peak of your current mountain. But if you want figure out whether you want to be on this mountain or another mountain, it isn't so useful.

着重于优化而非选择

e.g. ranking changes -

linkedin: whether they should show a news article or an encouragement to basically add new contacts.

google: search list and ads.

not useful testing out new experiences.

{ change aversion  
novelty effect -

2 issues:

1. what's your baseline for comparison?
2. how much time you need in order to actually have your users adapt to the new experience

log → analyze them retrospectively or observationally to see if a hypothesis can be developed about what's causing changes in their behavior — design, randomize and experiment, do a perspective analysis

complementary

A/B testing 用 broad quantitative data. 宽泛的定量数据.

其他用 qualitative data 定性数据. 作为 A/B test 的补充

当 online testing, goal 是 用户是否会 喜欢 这个新产品或新功能

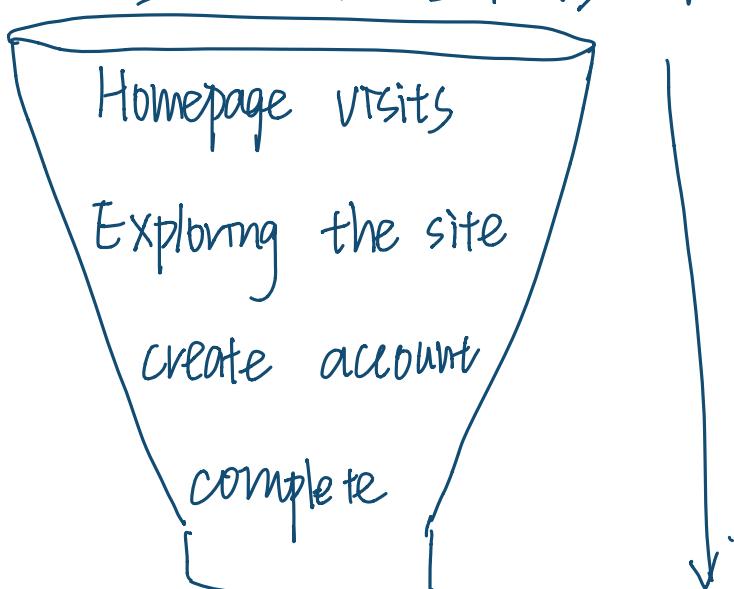
robust and repeatable

目标: 合理且能够检测到你可复验的结果.

example: Audacity example — creates online finance courses

Customer funnel

但  
回访者会跳过中间步骤聚



Experiment

Hypothesis: changing the "Start

"Now" button from orange to pink will increase how many students explore Audacity courses.

which metric to use?

Total number of courses completed  $\times$  费用

Number of clicks  $\times$

Number of clicks  $\times$  click-through

Number of page views - rate

Group 1      Group 2

• • •  
• • •

0 5  
0 1  
rate = 2.5  
prob = 0.15

unique visitors who click  $\checkmark$  click-through  
unique visitors to page - probability

updated hypothesis: changing the "Start Now" button from orange to pink will increase the click-through-probability of the button.

概率会告诉你他们能够经常找到这个按钮。

概率：用户会多经常地进入网站的二级页面。

into your control group and your experiment group.

⑤ what your results could look like,

how to analyze those results

how to draw valid conclusions.

## Binomial Distribution $(0, 1)$

$$\text{mean} = P$$

$$\text{std dev} = \sqrt{\frac{P(1-P)}{N}} \quad \text{e.g. } N=20, x=16$$

$$\text{estimated probability } \hat{P} = \frac{16}{20} = \frac{4}{5}$$

- 适用条件 {
1. 2 types of outcomes: - success and failure
  2. Independent events
  3. Identical distribution -  $P$  same for all

e.g. clicks on a search result pages - click or no click  $\times$

因为人会因为第一次 search 没有找到自己想要的结果，更换 keyword 进行第二次 search  $\Rightarrow$  事件关联。

怎样确定是否显著 — confidence interval.

Calculating a confidence interval

$$\hat{p} = \frac{x}{N} \quad \begin{matrix} \# \text{ user who clicked} \\ \# \text{ user} \end{matrix}$$

$$\hat{p} = \frac{100}{1000} = 0.1$$

$m$  = margin of error



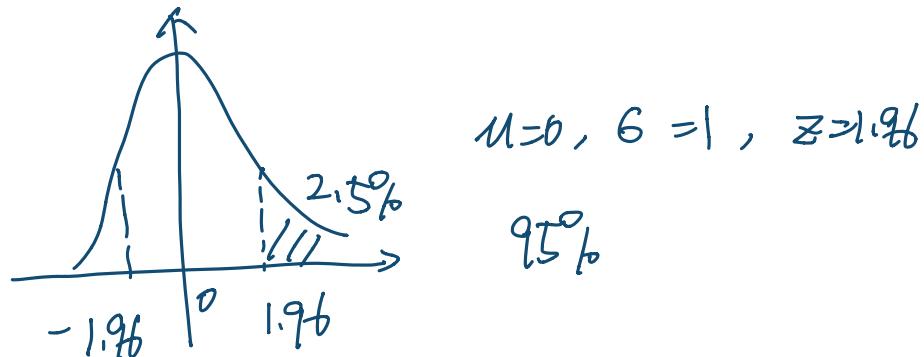
To use normal: check  $N \cdot \hat{p} > 5$  and  $N \cdot (1 - \hat{p}) > 5$

$$m = Z \times \text{SE} \quad \begin{matrix} Z \text{ score of confidence error} \\ (= \text{置信度}) \end{matrix}$$

$$m = Z \times \sqrt{\frac{\hat{p}(1-\hat{p})}{N}} \quad (= \text{置信度})$$

$$m = 0.019 \quad ?$$

Z distribution



$$p = \frac{360}{2000} = \frac{3}{20} = 0.15 \quad \frac{1 - 99\%}{2} = 0.5\% = 0.005 \Rightarrow 0.995$$

$$Z = 2.57 \quad \Rightarrow 0.995$$

$$m = 2.57 \times \sqrt{\frac{0.15 \times 0.85}{2000}} = 0.021$$

$$(0.15 - 0.021, 0.15 + 0.021) = (0.129, 0.171)$$

→ 确定实验中观察到的差异是否具有统计显著性.

How do we analyze the result?

⇒ Hypothesis testing: is a quantitative way to establish how likely it is that your results occurred by chance.

Null hypothesis (baseline): there is no difference in click-through probability between our control and our environment.

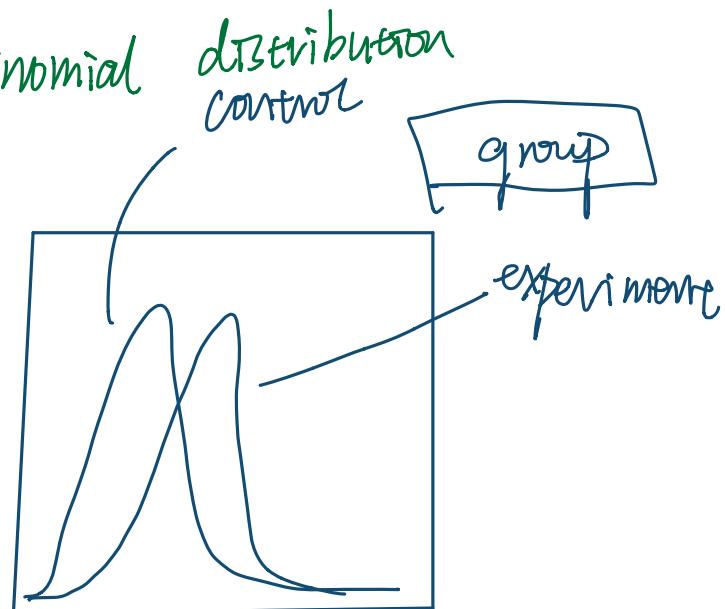
assume: each group follows a binomial distribution

Hypothesis Testing.

P(results due to chance)

$P_{\text{cont}}$

$P_{\text{exp}}$



Null hypothesis:  $P_{\text{cont}} = P_{\text{exp}}$

or  $H_0$

Measure  $\hat{P}_{\text{cont}}$  and  $\hat{P}_{\text{exp}}$

$$P_{\text{exp}} - P_{\text{cont}} = 0$$

Calculate  $P(|\hat{P}_{\text{exp}} - \hat{P}_{\text{cont}}| > H_0)$

Alternative hypothesis:  $P_{\text{exp}} - P_{\text{cont}} \neq 0$  Reject null.

or  $H_A$

Comparing two samples: ~~哪个~~? 而不是 difference

我们应该 choose a standard error that gives us a good

comparison of both  $\Rightarrow$  calculate pool standard error.

对照组转化率的随机变量

实验组转化率的随机变量

$$\hat{P}_{\text{pool}} = \frac{\hat{X}_{\text{cont}} + \hat{X}_{\text{exp}}}{N_{\text{cont}} + N_{\text{exp}}} \quad \text{total Number of 1 clicks}$$

$\hat{P}_{\text{pool}}$  estimated probability  
 $\hat{X}_{\text{cont}}, \hat{X}_{\text{exp}}$  total number of 1

$\hookrightarrow$  total probability of a click across groups

$$SE_{\text{pool}} = \sqrt{\hat{P}_{\text{pool}} * (1 - \hat{P}_{\text{pool}}) * \left( \frac{1}{N_{\text{cont}}} + \frac{1}{N_{\text{exp}}} \right)}$$

pooled standard error

$$\hat{d} = \hat{P}_{\text{exp}} - \hat{P}_{\text{cont}}$$

$\downarrow$  怎么算出来的这2个数?

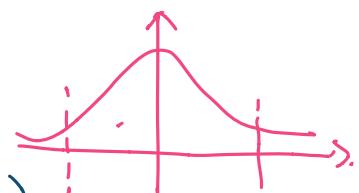
$$P = \frac{X}{N}$$

estimate the difference

表示 probability 没有区别。

$$H_0: d = 0$$

$$\hat{d} \sim N(0, SE_{\text{pool}})$$



$\downarrow$  true difference

$\hookrightarrow$  distributed normally

$\mapsto z \text{ score}$

If  $\hat{d} > 1.96 * SE_{\text{pool}}$  or  $\hat{d} < -1.96 * SE_{\text{pool}}$ , reject null

$\Rightarrow$  our difference represents a statistically significant difference.

拒绝原假设  $\Leftrightarrow$  example 在 tail 上面。

## Practical, or Substantive, Significance

### Hypothesis testing.

Then next, decide from a business perspective, what change in the click-through probability ability is practically significant.

- = What size change matters to us?
- = What's substantive in addition to being statistically significant

为什么不能做所有变更:

1% ~ 2% change is large enough

1. 变更本身需要投资.

2. 等下一次变更

Statistical significance

统计显著性讨论的是 repeatability. 可重复性

change

从 business point 来看, 当你看到有关取巧的 ✓ 它就是 practical

significant, 同时也是 statistical significant

⇒ Size your experiment appropriately, such that the statistical

significance bar is actually lower than the practical significance bar.

## Size vs. Power Trade off      "Design"

main question: decide how many page views we need in order to get a statistically significant result. = "Statistical power"

The power has an inverse trade-off with size:

The smaller the change that you want to detect, or the  
结果置信度↑  
increased confidence that you basically want to have in the  
result, means that you have to run a larger experiment. N↑

为什么?

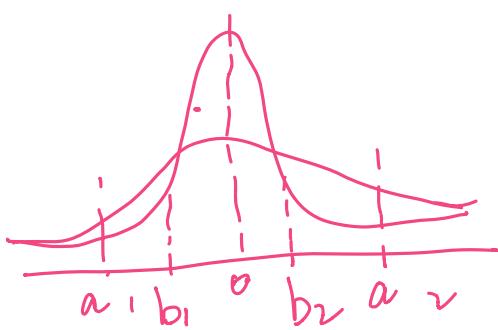


解释.

how the distribution changes when you change your sample size.

①  $n=1000$  时

(因为2条线重合在一起)



两组之间没有明显差异  $\Rightarrow \text{mean} \approx 0$

如果你测量发现值比这个低或这个高 ( $a_1, a_2$ )

reject  $H_0$ ,  $\Rightarrow$  两者有差异.  $\alpha = 0.05$ .

$\hookrightarrow$  错误的  $\alpha \Rightarrow \alpha = P(\text{reject null} | \text{null true})$

② 当 n↑  $\Rightarrow n=5000$  时, 标准误差  $\downarrow \Rightarrow$  结果分布变窄.

如果保持相同的  $\alpha$ , 又变成  $b_1, b_2$ .

拒绝  $H_0$  的临界值 (cut-off) close to 0.



此时有真正的差异:

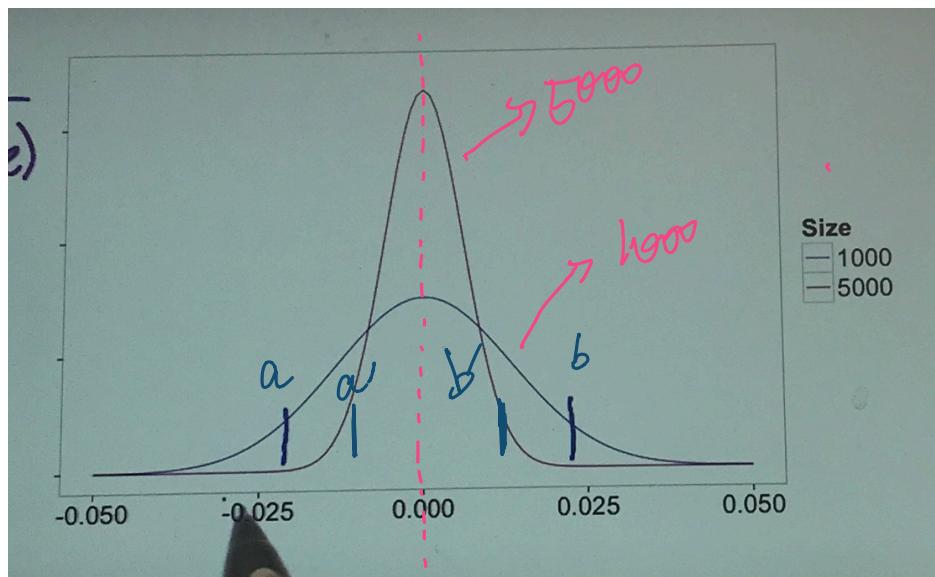
$\Rightarrow$  差异 = 实际显著性.  $= 0.02$ .

所以不能拒绝  $H_0$  得出没有显著性的差异

此时  $\beta = P(\text{fail to reject} | \text{null false})$

# How many page views N

为什么?



当取  $n = 1000$  时, 2组之

间没有真正的差异.

所以  $\text{mean} = 0$

① 当你测量的值  $< \alpha$  或  $> \beta$ , 错误得出两者有差异的机率率.

$$\alpha = P(\text{reject null} \mid \text{null is true}) = 0.05$$

当 sample  $\uparrow$ , standard error  $\downarrow$ , 当保持同样子值,

临界值

Cut-offs for rejecting the  $H_0$  nulls close to 0.

取决于你的效应有多大

~~difference = practical significance~~

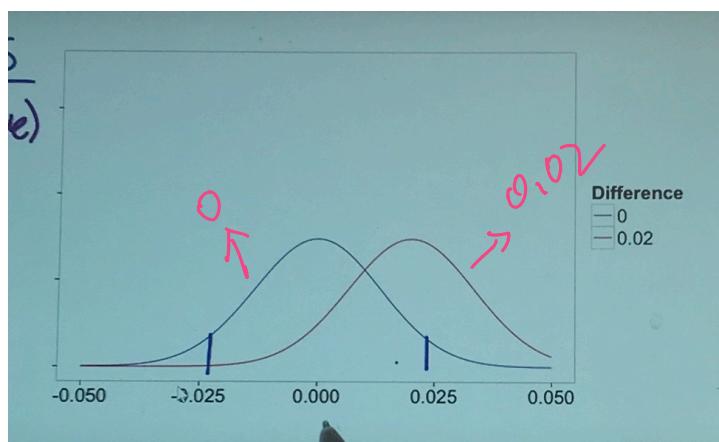
所以 failed to reject the nulls

$\Rightarrow$  没有显著性差异.

$$\beta = P(\text{fail to reject} \mid \text{null false})$$

不太能进行较差的实验

Small sample,  $\downarrow$  low,  $\beta$  is high 很大机会无法在实验中得到差异结果



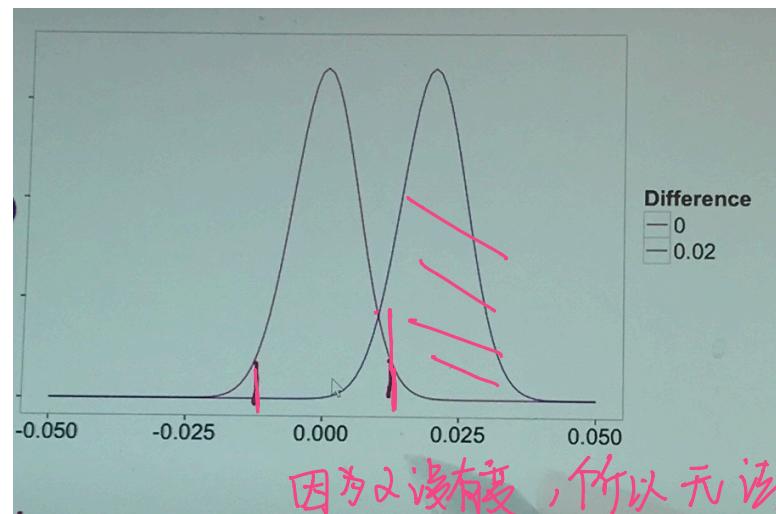
Large change of  $\mu$

lower chance of error

对于良态的分布 (正态分布), 当变化↑  $\rightarrow \beta \downarrow \rightarrow$  误差机率↓  
e.g.

$\beta$  是在实际显著性边界内. 因为你不在意更小的变化.

$1 - \beta = \text{sensitivity}$  (often 80%)  $\rightarrow$  practical significant boundary  
一般喜欢较高 sensitivity -



large sample:

& same,  $\beta$  lower

因为又没幅度, 所以无法拒绝 Ho 的机率大大降低

Calculating number of page views:

- Built-in library
- Look up answer in a table
- Use online calculator

△ Baseline conversion rate: the estimated CTR before making the change  
Min. Detectable effect: practical significance level { absolute ✓  
relative lesson 3

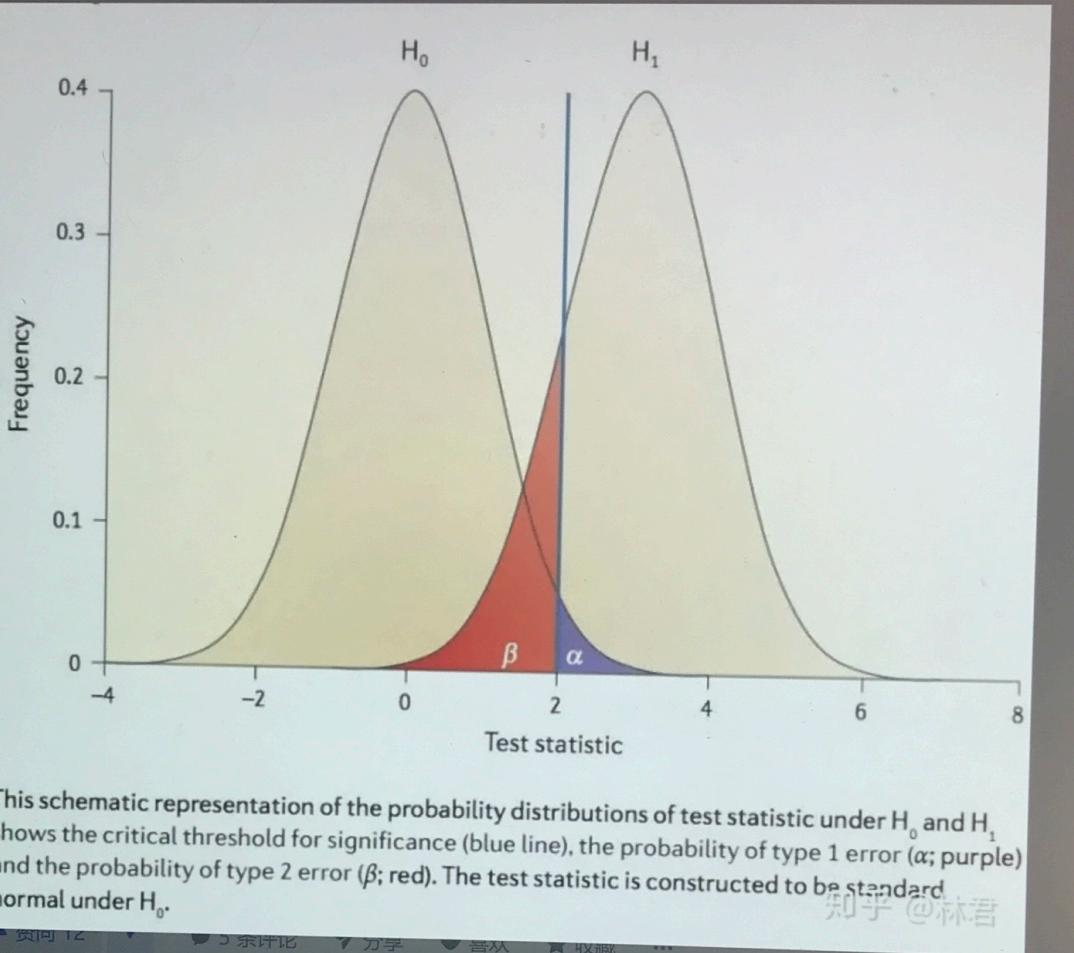
# How number of page views varies

Change	↑ page view	↓ page view
Higher click-through-probability in control (still < 0.5)	✓	
$SE = \sqrt{\frac{P(1-P)}{N}}$	$\sqrt{0.15 \times 0.85} = 0.3$ 当 $P \uparrow \rightarrow SE \downarrow \rightarrow N$ 流量使 $\sqrt{0.1 \times 0.9} = 0.3$ $SE$ 回归原水平	
Increased practical significance level ( $d_{min}$ ) $d_{min} \uparrow \Rightarrow$ 更大的变更更容易被检测到 $\Rightarrow$ 流量 $N$ 没有必要的时候	✓	
Increased confidence level ( $1-\alpha$ )	✓ 可以通过 reject $H_0$ 达到且 $\downarrow$	
Higher sensitivity ( $1-\beta$ )	✓ 保持同样 <del>power</del> $N \uparrow$	

## Minimum Detectable Effect

顾名思义，这个参数衡量了我们对实验的判断精确度的最低要求。

- 参数越大（比如10%），说明我们期望实验能够检测出10%的差别即可。检测这么大的差别当然比较容易（power变大），所以保持power不变的情况下，所需要的样本量会变小。
- 参数越小（比如1%），说明我们希望实验可以有能力检测出1%的细微差别。检测细微的差别当然更加困难（power变小），所以如果要保持power不变的话，需要的样本量会增加。



$$H_0: P \leq 4\%$$

$$H_1: P > 4\%$$

$$P_{10}(4) = 0.0042$$

# 第一节 假设检验的原理

- 假设检验的原理：小概率原理。假设检验的基本思想是概率性质的反证法。（不同于纯数学中的反证法）
- 什么是小概率？
- 概率是 $0 \sim 1$ 之间的一个数，因此小概率就是接近 $0$ 的一个数
- 著名的英国统计家Ronald Fisher 把 $20$ 分之 $1$ 作为标准，也就是 $0.05$ ，从此 $0.05$ 或比 $0.05$ 小的概率都被认为是小概率
- Fisher没有任何深奥的理由解释他为什么选择 $0.05$ ，只是说他忽然想起来的

# 什么是小概率原理？

- 小概率原理——发生概率很小的随机事件（小概率事件）在一次实验中几乎是不可能发生的。
- 根据这一原理，可以先假设总体参数的某项取值为真，也就是假设其发生的可能性很大，然后抽取一个样本进行观察，如果样本信息显示出现了与事先假设相反的结果且与原假设差别很大，则说明原来假定的小概率事件在一次实验中发生了，这是一个违背小概率原理的不合理现象，因此有理由怀疑和拒绝原假设；否则不能拒绝原假设。
- 检验中使用的小概率是检验前人为指定的。

## 假设检验的两个特点：

第一，假设检验采用逻辑上的反证法，即为了检验一个假设是否成立，首先假设它是真的，然后对样本进行观察，如果发现出现了不合理现象，则可以认为假设是不合理的，拒绝假设。否则可以认为假设是合理的，接受假设。

第二，假设检验采用的反证法带有概率性质。所谓假设的不合理不是绝对的，而是基于实践中广泛采用的小概率事件几乎不可能发生的原则。至于事件的概率小到什么程度才算是小概率事件，并没有统一的界定标准，而是必须根据具体问题而定。如果一旦判断失误，错误地拒绝原假设会造成巨大损失，那么拒绝原假设的概率就应定的小一些；如果一旦判断失误，错误地接受原假设会造成巨大损失，那么拒绝原假设的概率就应定的大一些。

■ 小概率通常用 $\alpha$ 表示，又称为检验的显著性水平。通常取 $\alpha=0.05$ 或 $\alpha=0.01$ ，即把概率不超过0.05或0.01的事件当作小概率事件。

# 原假设和备择假设

- 假设检验中，我们称作为检验对象的待检验假设为原假设或零假设，用 $H_0$ 表示。原假设的对立假设称为备择假设或备选假设，用 $H_1$ 表示。
- 例如，设 $\bar{X}_0$ 为总体均值 $\bar{X}$ 的某一确定值。  
(1) 对于总体均值是否等于某一确定值的原假设可以表示为：

$$H_0: X = \bar{X}_0 \quad (\text{如 } H_0: \bar{X} = 3190 \text{ 克})$$

其对应的备择假设则表示为：

$$H_1: X \neq \bar{X}_0 \quad (\text{如 } H_1: \bar{X} \neq 3190 \text{ 克})$$

(2) 对于总体均值  $\bar{X}$  是否大于某一确定值  $\bar{X}_0$  的原假设可以表示为：

$$H_0: \bar{X} \geq \bar{X}_0 \quad (\text{如 } H_0: \bar{X} \geq 2000 \text{ 克})$$

其对应的备择假设则表示为：

$$H_1: \bar{X} < \bar{X}_0 \quad (\text{如 } H_1: \bar{X} < 2000 \text{ 克})$$

(3) 对于总体均值  $\bar{X}$  是否小于某一确定值  $\bar{X}_0$  的原假设可以表示为：

$$H_0: \bar{X} \leq \bar{X}_0 \quad (\text{如 } H_0: \bar{X} \leq 5\%)$$

其对应的备择假设则表示为：

$$H_1: \bar{X} > \bar{X}_0 \quad (\text{如 } H_1: \bar{X} > 5\%)$$

注意：原假设总是有等号： $=$  或  $\leq$  或  $\geq$ 。

# 双侧检验和单侧检验

- 根据假设的形式不同，假设检验可以分为双侧假设检验和单侧假设检验。
- 若原假设是总体参数等于某一数值，如 $H_0: \bar{X} = \bar{X}_0$ ，即备择假设 $H_1: \bar{X} \neq \bar{X}_0$ ，那么只要 $\bar{X} < \bar{X}_0$ 和 $\bar{X} > \bar{X}_0$ 二者中有一个成立，就可以否定原假设。这种假设检验称为双侧检验。
- 若原假设是总体参数大于等于或小于等于某一数值，如 $H_0: \bar{X} \geq \bar{X}_0$ （即 $H_1: \bar{X} < \bar{X}_0$ ）；或 $H_0: \bar{X} \leq \bar{X}_0$ （即 $H_1: \bar{X} > \bar{X}_0$ ），那么对于前者当 $\bar{X} < \bar{X}_0$ 时，对于后者当 $\bar{X} > \bar{X}_0$ 时，可以否定原假设。这种假设检验称为单侧检验。可以分为左侧检验和右侧检验。

# 双侧检验与单侧检验(假设的形式)

假设	研究的问题 (总体均值检验)		
	双侧检验	左侧检验	右侧检验
$H_0$	$\bar{X} = \bar{X}_0$	$\bar{X} \geq \bar{X}_0$	$\bar{X} \leq \bar{X}_0$
$H_1$	$\bar{X} \neq \bar{X}_0$	$\bar{X} < \bar{X}_0$	$\bar{X} > \bar{X}_0$

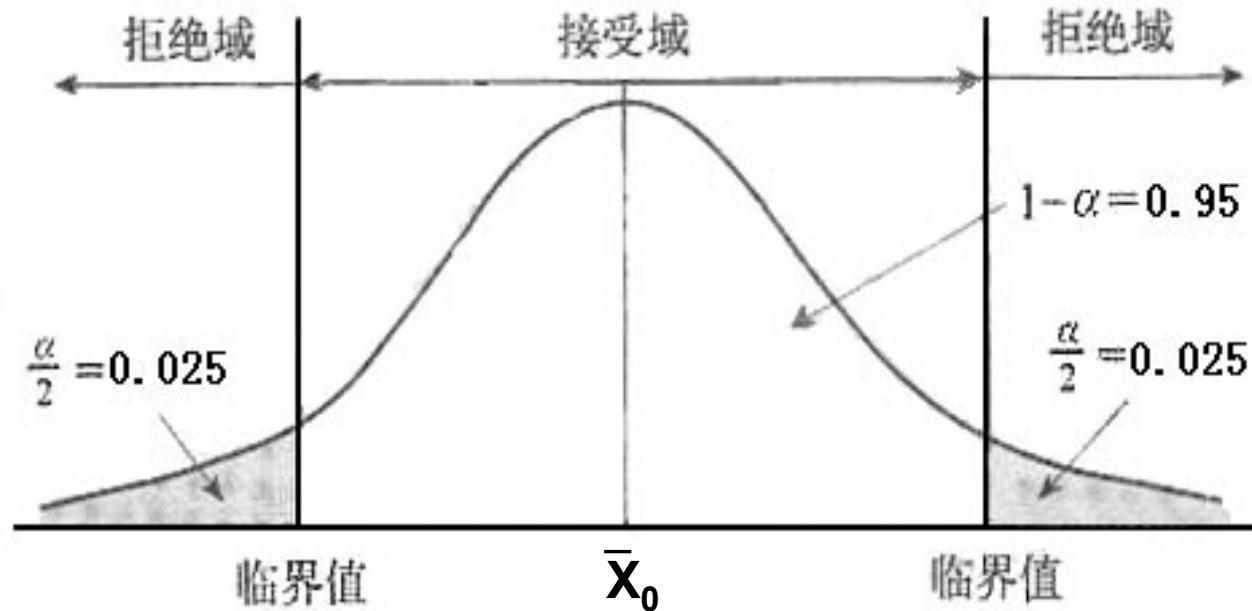
# 假设检验中的拒绝域和接受域

- 在规定了检验的显著性水平 $\alpha$ 后，根据容量为n的样本，按照统计量的理论概率分布规律，可以确定据以判断拒绝和接受原假设的检验统计量的临界值。
- 临界值将统计量的所有可能取值区间分为两个互不相交的部分，即原假设的拒绝域和接受域。
- 对于正态总体，总体均值的假设检验可有如下图示：

- 正态总体，总体均值假设检验图示：

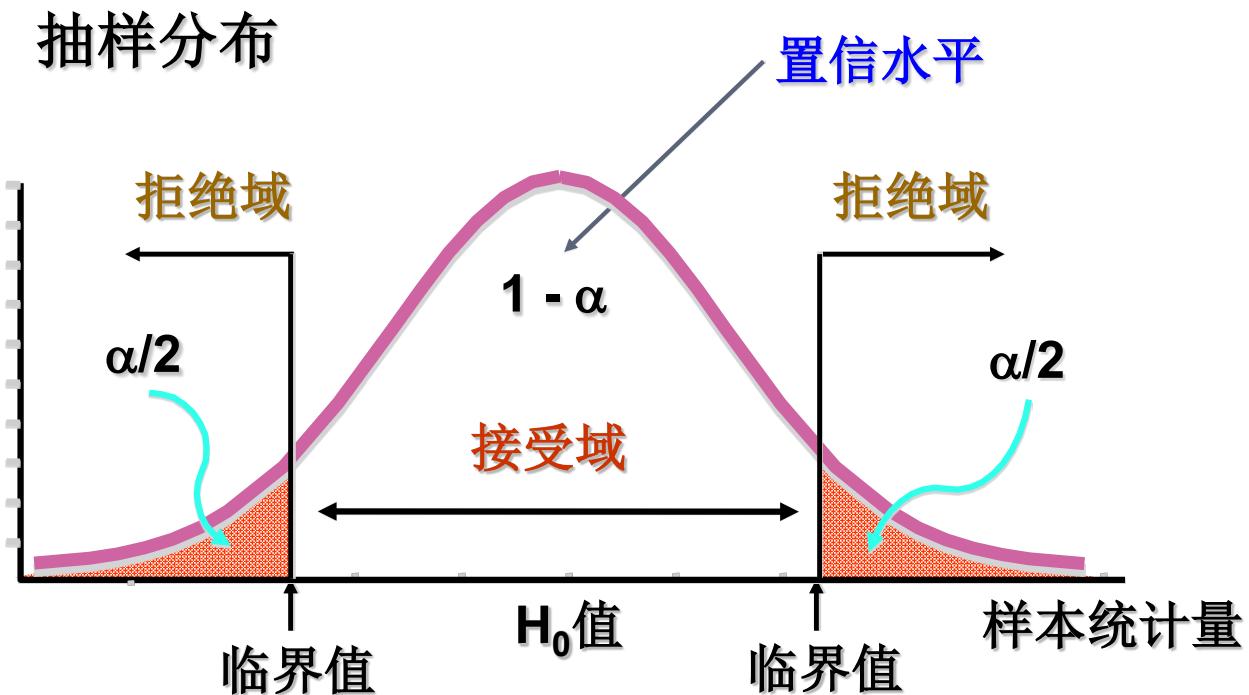
### (1) 双侧检验

设  $H_0: \bar{X} = \bar{X}_0$ ,  $H_1: \bar{X} \neq \bar{X}_0$ , 有两个临界值, 两个拒绝域, 每个拒绝域的面积为  $\alpha/2$ 。也称双尾检验。

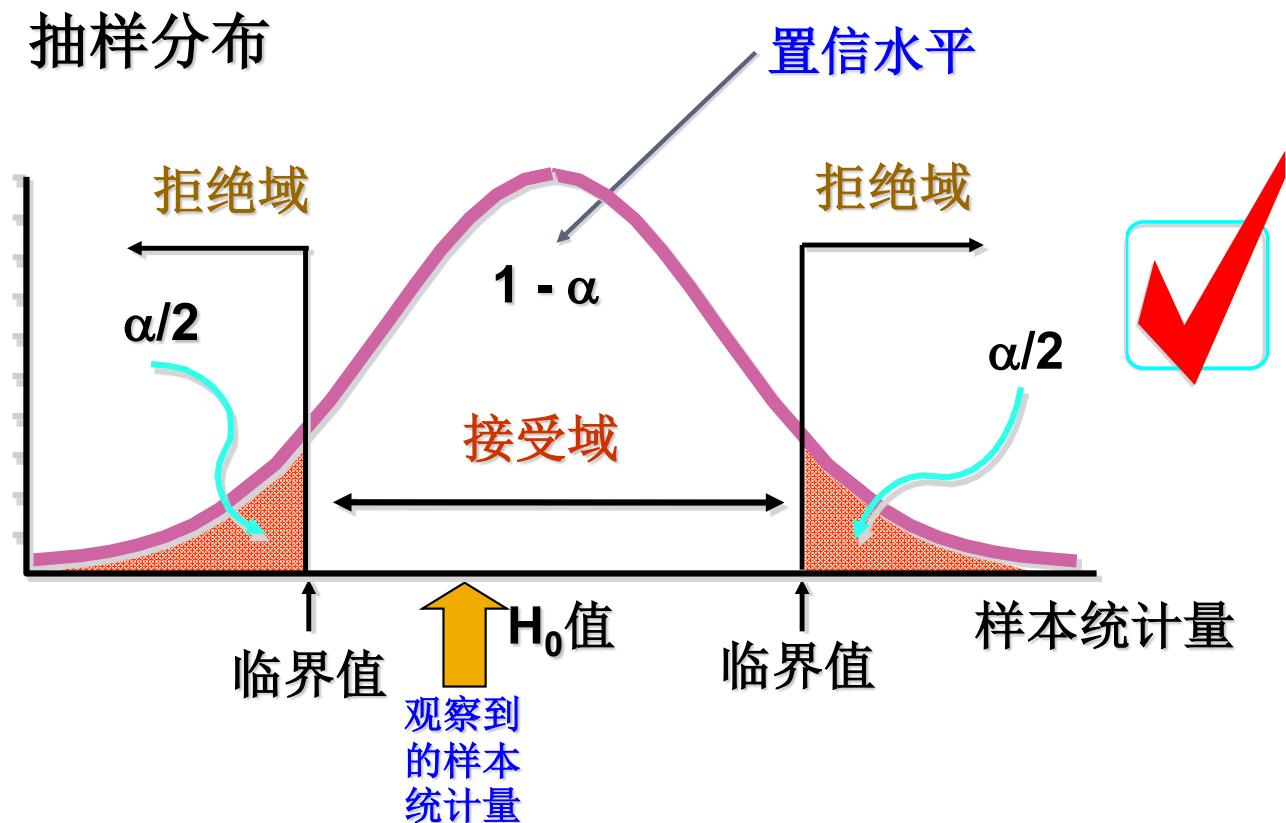


双侧检验示意图

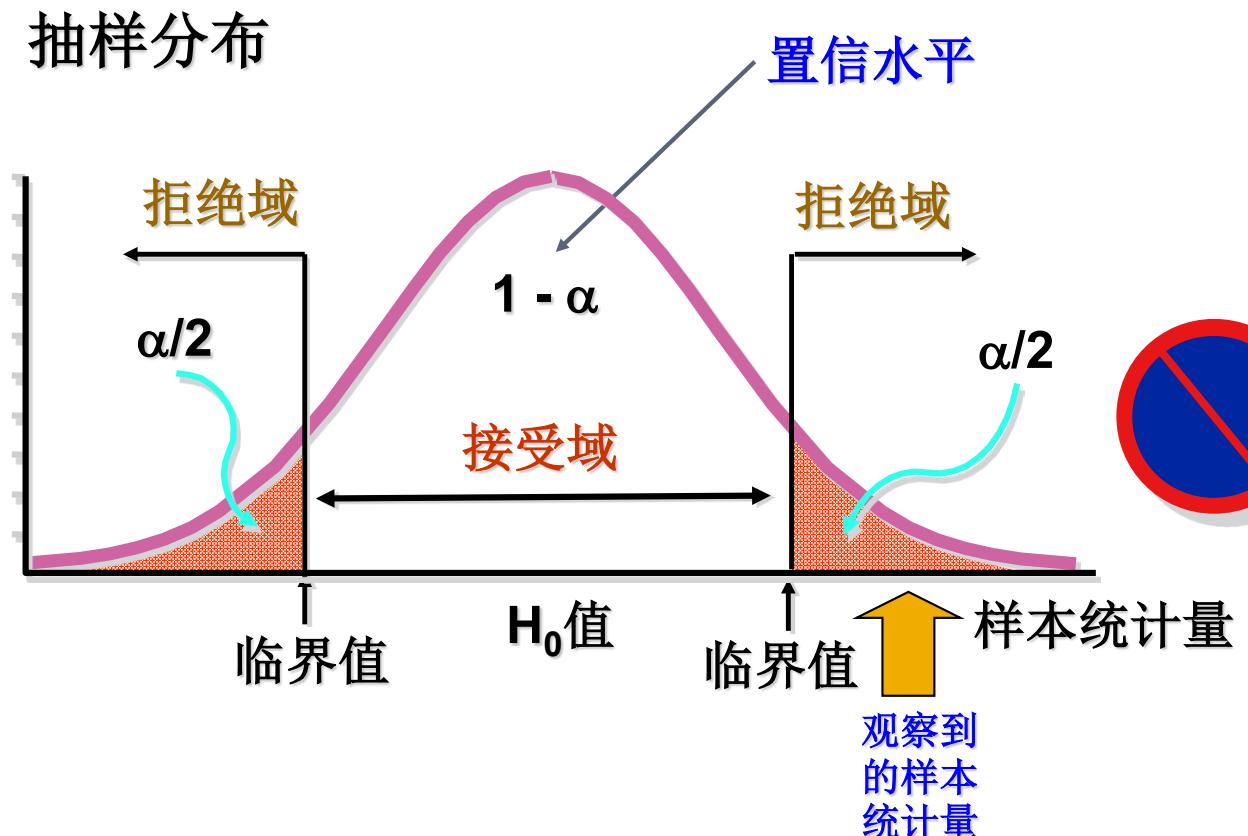
# 双侧检验示意图（显著性水平与拒绝域）



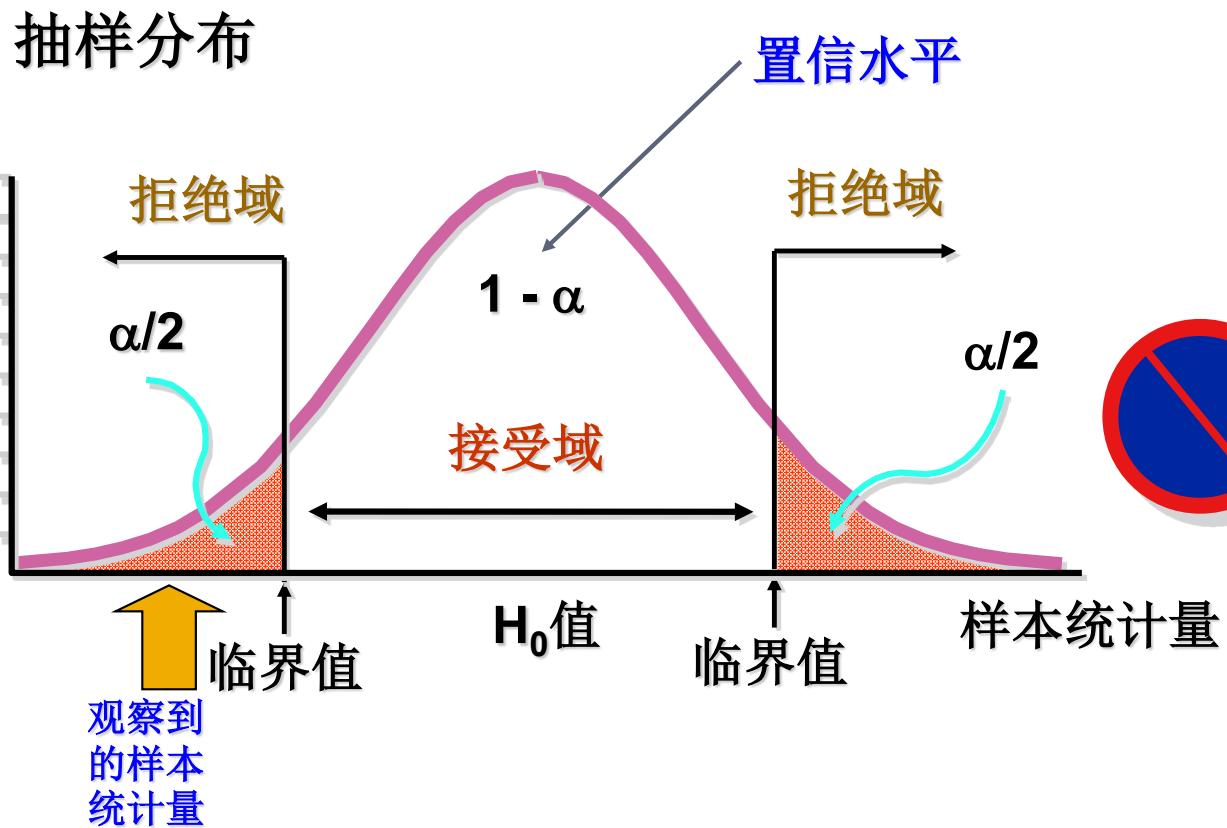
# 双侧检验示意图（显著性水平与拒绝域）



# 双侧检验示意图（显著性水平与拒绝域）



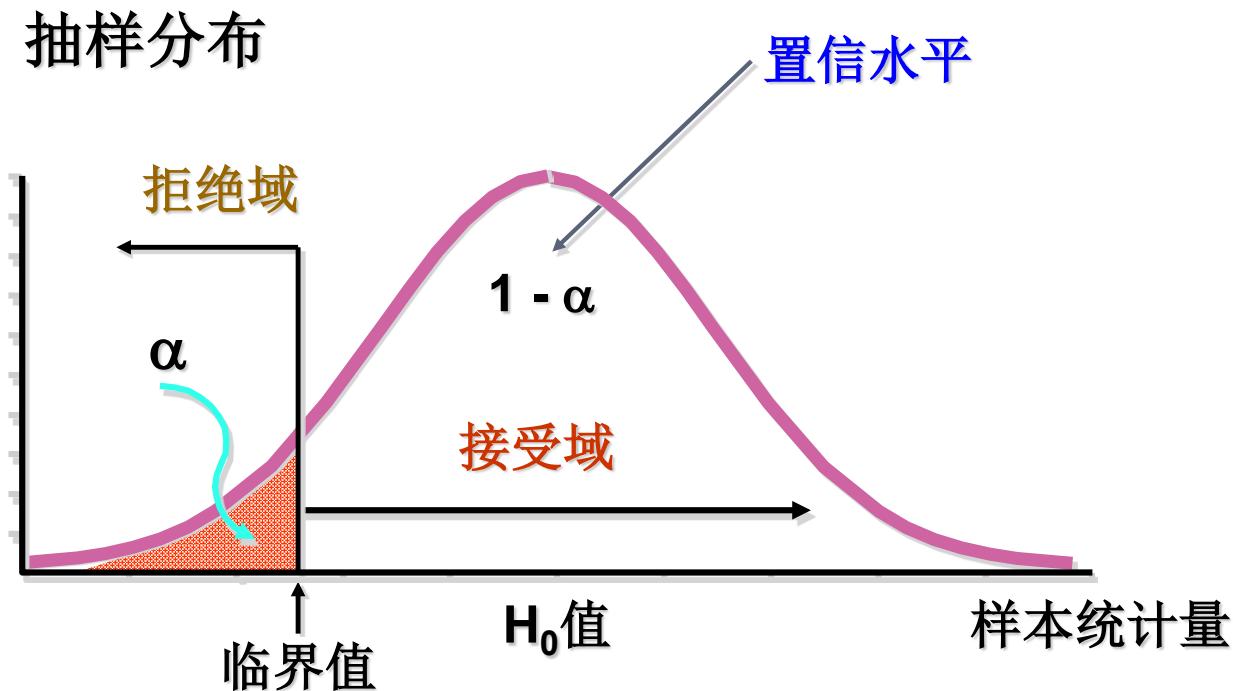
# 双侧检验示意图 (显著性水平与拒绝域)



## (2) 单侧检验

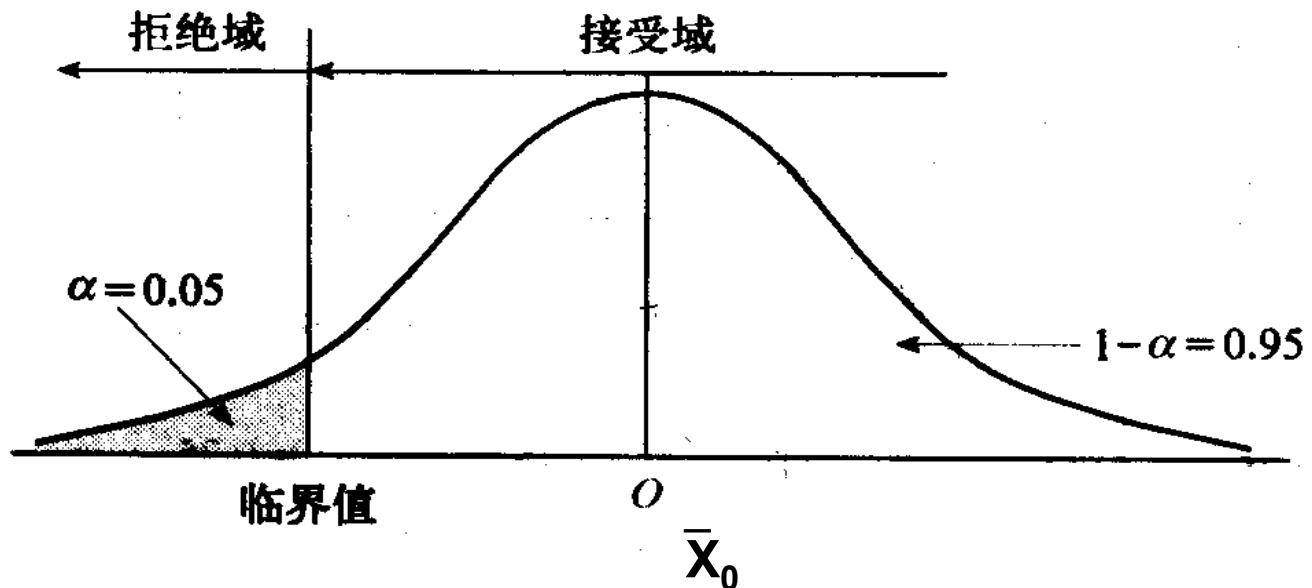
有一个临界值，一个拒绝域，拒绝域的面积为 $\alpha$ 。分为左侧检验和右侧检验两种情况。

单侧检验示意图（显著性水平与拒绝域）



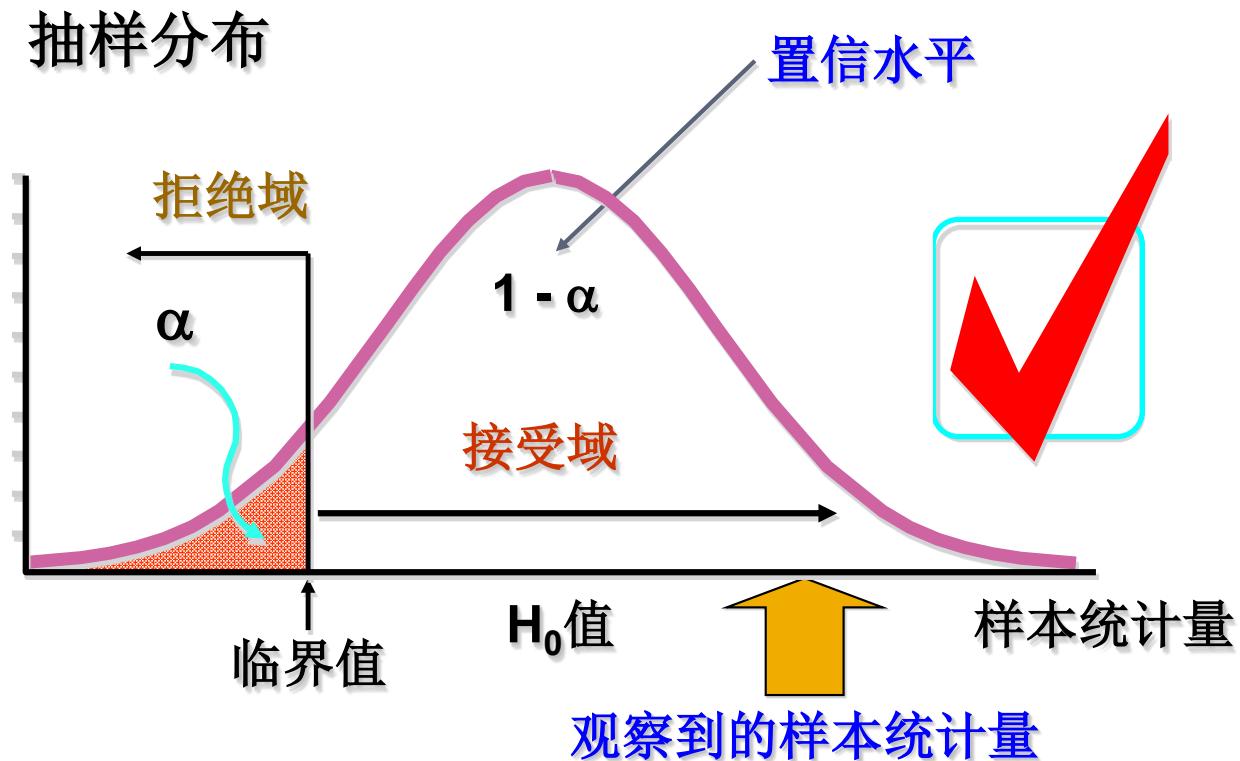
## 左侧检验

设  $H_0: \bar{X} \geq \bar{X}_0$ ,  $H_1: \bar{X} < \bar{X}_0$ ; 临界值和拒绝域均在左侧。也称下限检验。

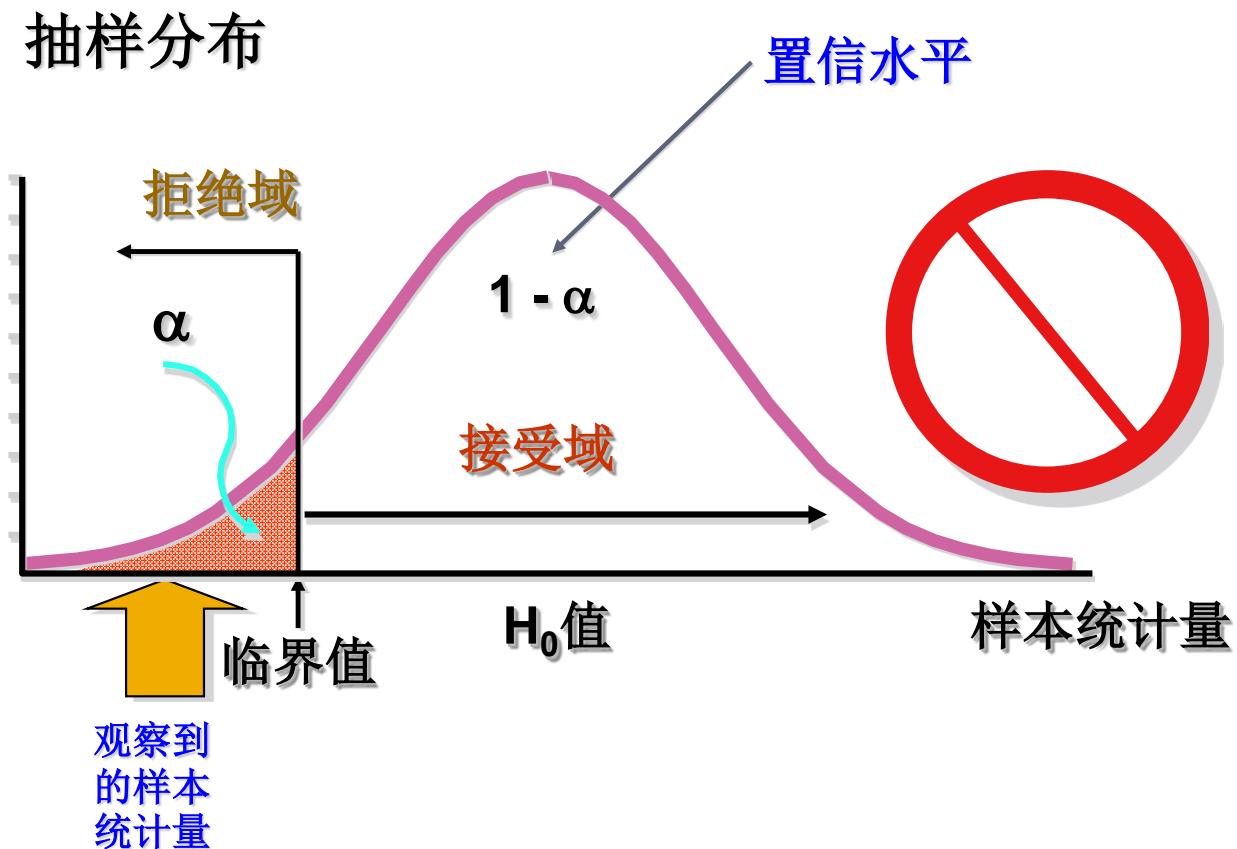


左单侧检验示意图

# 左侧检验示意图（显著性水平与拒绝域）

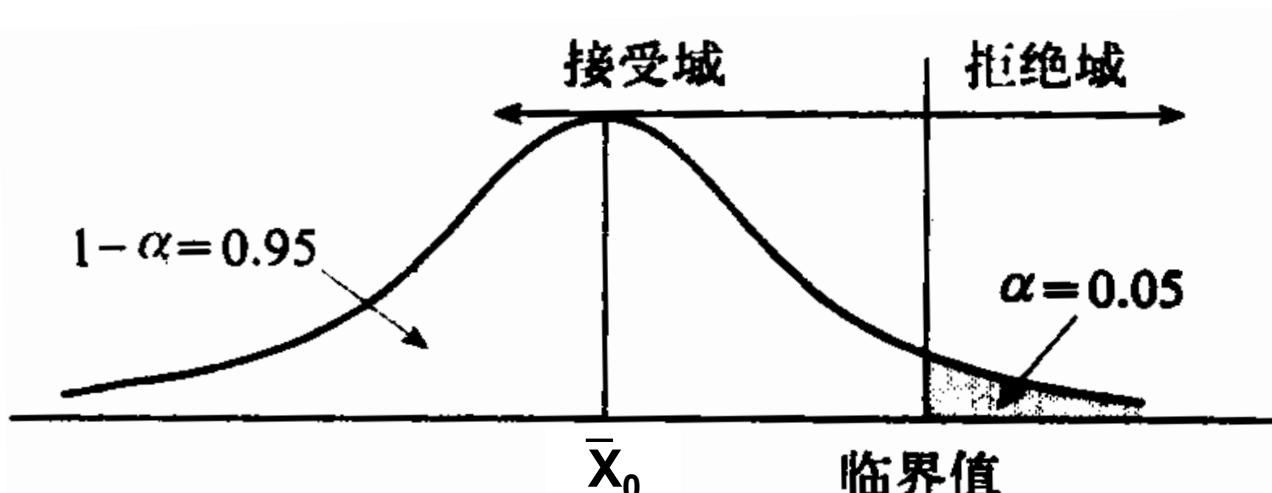


# 左侧检验示意图（显著性水平与拒绝域）



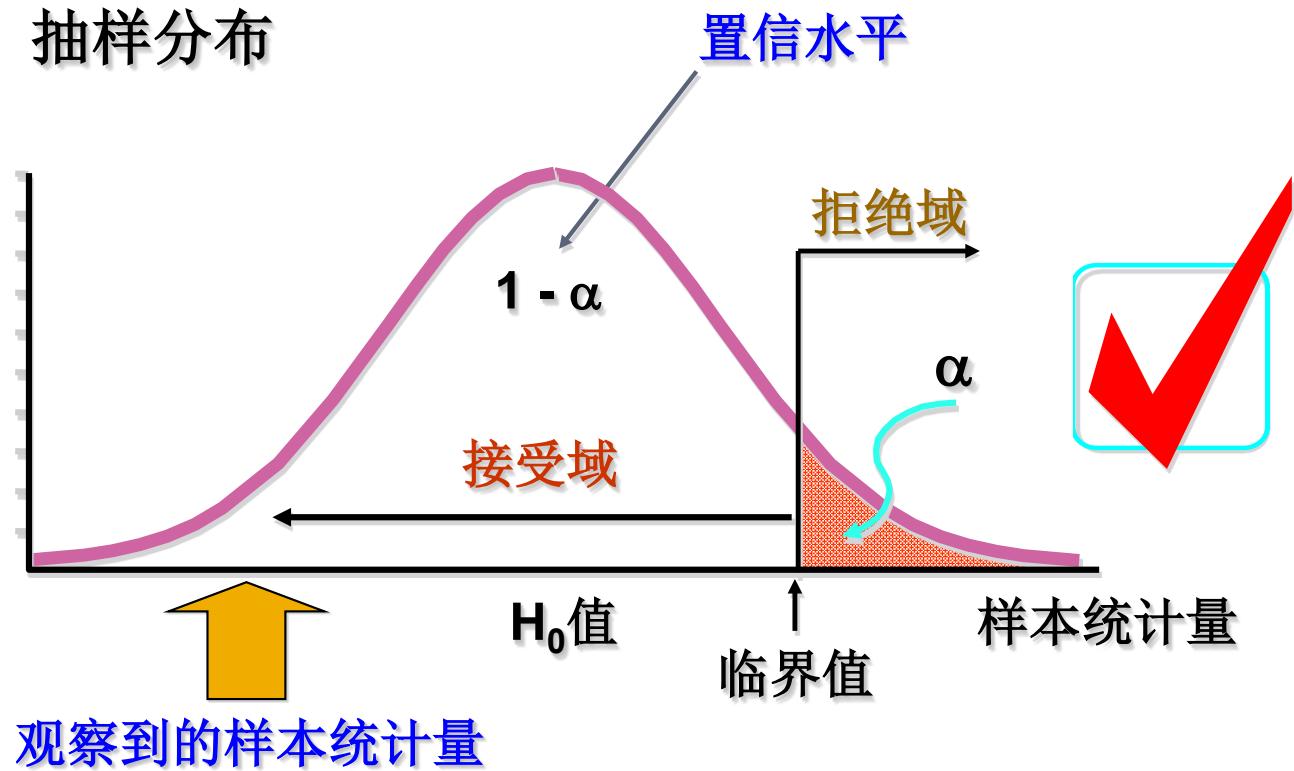
## 右侧检验

设  $H_0: \bar{X} \leq \bar{X}_0$ ,  $H_1: \bar{X} > \bar{X}_0$ ; 临界值和拒绝域均在右侧。也称上限检验。

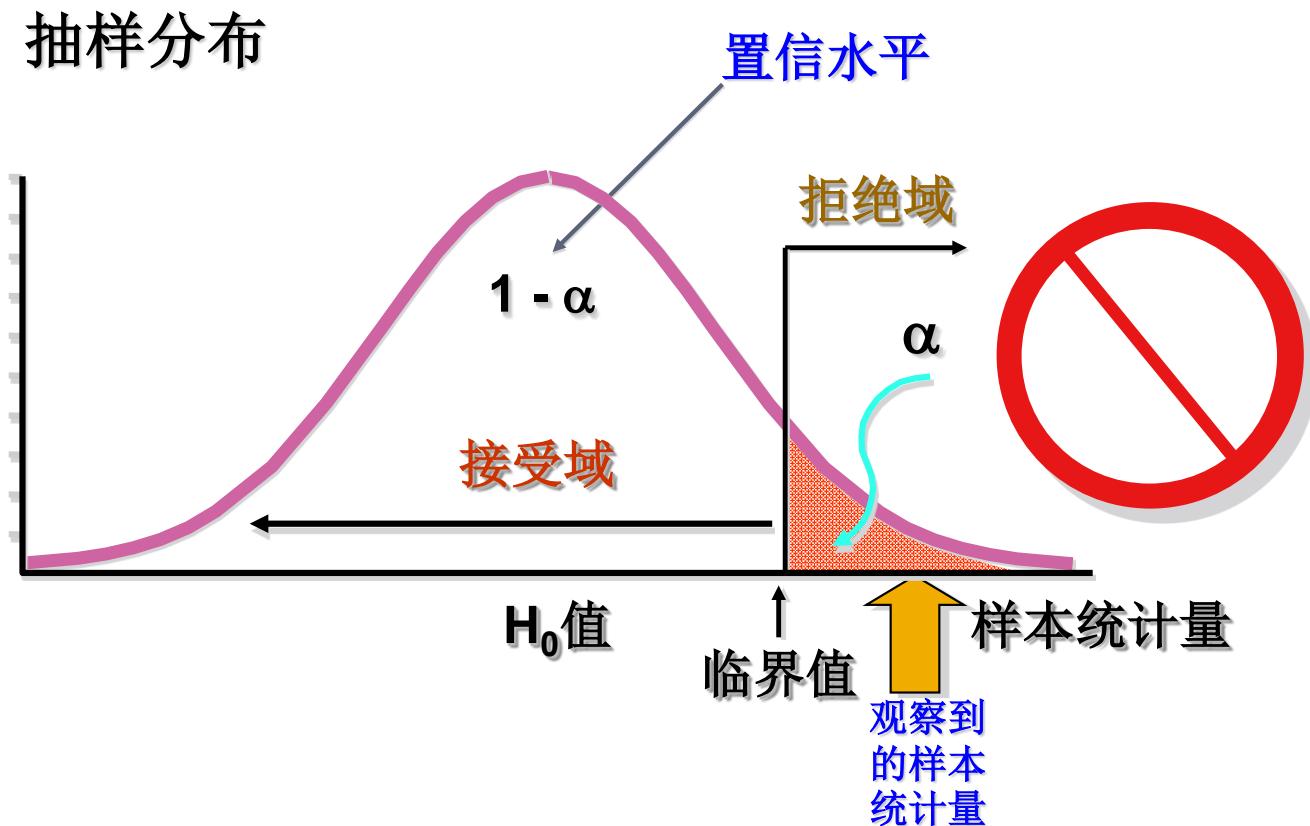


右单侧检验示意图

# 右侧检验示意图（显著性水平与拒绝域）



# 右侧检验示意图（显著性水平与拒绝域）



# 假设检验的两类错误

- 根据假设检验做出判断无非下述四种情况：
  - 1、原假设真实，并接受原假设，判断正确；
  - 2、原假设不真实，且拒绝原假设，判断正确；
  - 3、原假设真实，但拒绝原假设，判断错误；
  - 4、原假设不真实，却接受原假设，判断错误。
- 假设检验是依据样本提供的信息进行判断，有犯错误的可能。所犯错误有两种类型：
- 第一类错误是原假设 $H_0$ 为真时，检验结果把它当成不真而拒绝了。犯这种错误的概率用 $\alpha$ 表示，也称作 $\alpha$ 型错误或弃真错误。与前面一个相反
- 第二类错误是原假设 $H_0$ 不为真时，检验结果把它当成真而接受了。犯这种错误的概率用 $\beta$ 表示，也称作 $\beta$ 型错误或取伪错误。

# 假设检验的两类错误

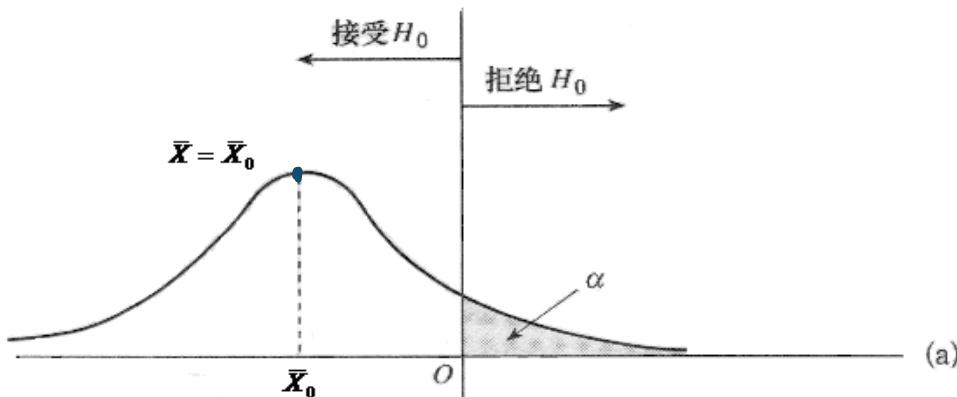
正确决策和犯错误的概率可以归纳为下表：

## 假设检验中各种可能结果的概率

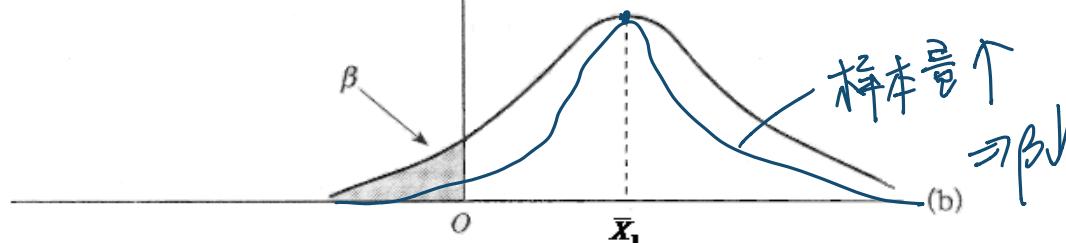
		接受 $H_0$	拒绝 $H_0$ 接受 $H_1$
审判过程： $H_0$ : 无罪			
裁决	实际情况		
无罪	无罪	$H_0$ 为 真	$1-\alpha$ (正确决策)
有罪	错误		$\alpha$ (弃真错误)
有罪	正确	$H_0$ 为 伪	$\beta$ (取伪错误)
			$1-\beta$ (正确决策)

## •假设检验两类错误关系的图示

以单侧上限检验为例，设  $H_0: \bar{X} \leq \bar{X}_0$ ,  $H_1: \bar{X} > \bar{X}_0$



图(a)  
 $\bar{X} \leq \bar{X}_0$   
 $H_0$  为真



图(b)  
 $\bar{X} = \bar{X}_1 > \bar{X}_0$   
 $H_0$  为伪

从上图可以看出，如果临界值沿水平方向右移， $\alpha$ 将变小而 $\beta$ 变大，即若减小 $\alpha$ 错误，就会增大犯 $\beta$ 错误的机会；如果临界值沿水平方向左移， $\alpha$ 将变大而 $\beta$ 变小，即若减小 $\beta$ 错误，也会增大犯 $\alpha$ 错误的机会。

## 两类错误的控制

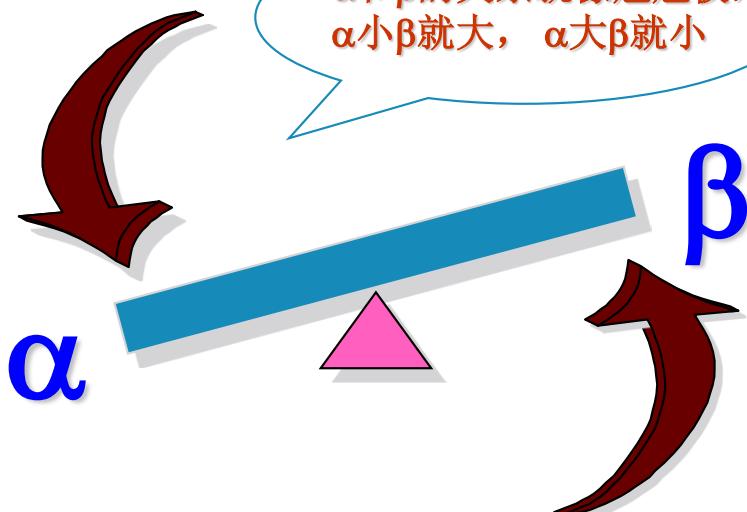
(1) 利用已知的实际总体参数值与假设参数值之间的大小关系，合理安排拒绝区域的位置

- ① 若弃真的后果较大，则将 $\alpha$ 定小一点， $\beta$ 就大。
- ② 若存伪的后果较大，则将 $\beta$ 定小一点， $\alpha$ 就大。

(2) 使样本容量增大，可以同时减少两类错误，或减少其中一种错误而不致于增大另一种错误（如在 $\alpha$ 和其他条件不变时， $\beta$ 会减小）。（因为样本容量增大，抽样误差越小，样本分布就越高狭，两侧的面积就越小。）

## $\alpha$ 错误和 $\beta$ 错误的关系

■ 在样本容量 $n$ 一定的情况下，假设检验不能同时做到犯 $\alpha$ 和 $\beta$ 两类错误的概率都很小。若减小 $\alpha$ 错误，就会增大犯 $\beta$ 错误的机会；若减小 $\beta$ 错误，也会增大犯 $\alpha$ 错误的机会。要使 $\alpha$ 和 $\beta$ 同时变小只有增大样本容量。但样本容量增加要受人力、经费、时间等很多因素的限制，无限制增加样本容量就会使抽样调查失去意义。因此假设检验需要慎重考虑对两类错误进行控制的问题。



你不能同时减少两类错误！



## ●两类错误的控制准则

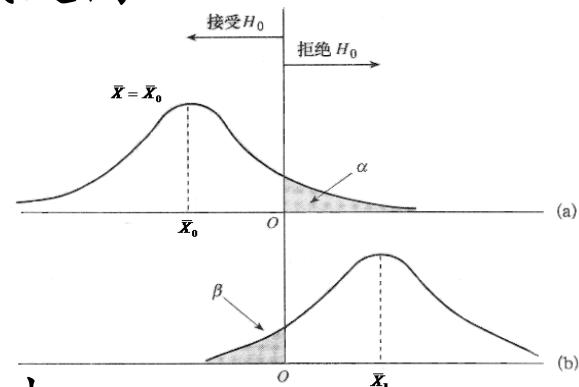
- 假设检验中人们普遍执行同一准则：首先控制弃真错误（ $\alpha$  错误）。 假设检验的基本法则以  $\alpha$  为显著性水平就体现了这一原则。
  - 两个理由：
    - 统计推断中大家都遵循统一的准则，讨论问题会比较方便。
    - 更重要的是：原假设常常是明确的，而备择假设往往是模糊的。如  $H_0: \bar{X} = \bar{X}_0$  很清楚，而  $H_1: \bar{X} \neq \bar{X}_0$  则不太清楚，是  $\bar{X} < \bar{X}_0$  还是  $\bar{X} > \bar{X}_0$ ？大多少小多少都不清楚。对含义清晰的数量标准进行检验更容易被接受。
- 因此，第一类错误成为控制两类错误的重点。

## • 两类错误的控制准则

- $\alpha$  错误：控制显著性水平。
  - 实验条件控制较好： $\alpha=0.05$
  - 实验条件难于控制： $\alpha=0.01$ ，或更高

P227 统计检验力  $1-\beta$

- $\beta$  错误的影响因素与控制
  - 实际值与假设值相差越大， $\beta$  越小。
  - $\alpha$  越小， $\beta$  越大。同时控制，增加n。
  - $\alpha$ 、n 固定时，适当的检验类型可减小 $\beta$ 。



## ● 假设检验的步骤

- (一)根据研究需要提出原假设 $H_0$ 和备择假设 $H_1$
- (二)确定适当的检验统计量
- (三)确定显著性水平 $\alpha$ 和临界值及拒绝域
- (四)根据样本数据计算检验统计量的值 (或P值)
- (五)将检验统计量值与临界值比较, 作出拒绝或接受原假设的决策

## 假设检验的步骤

(一) 根据研究需要提出原假设 $H_0$ 和备择假设 $H_1$

- 应该注意：

- (1) 对任一假设检验问题，其所有可能结果均应包括在所提出的两个对立假设中，原假设与对立假设总有一个、也只能有一个成立。
- (2) 原假设一定要有等号：=或 $\leq$ 或 $\geq$ 。

- 原假设不是随意提出的，应该本着“不轻易拒绝原假设”的原则。

## • 双侧检验原假设与备择假设的确定

- 双侧检验属于决策中的假设检验。即不论是拒绝 $H_0$ 还是接受 $H_0$ ，都必需采取相应的行动措施。
- 例如，某种零件的尺寸，要求其平均长度为10厘米，大于或小于10厘米均属于不合格。待检验问题是该企业生产的零件平均长度是10厘米吗？(属于决策中的假设)则建立的原假设与备择假设应为

$$H_0: \bar{X} = 10 \quad H_1: \bar{X} \neq 10$$

## • 单侧检验原假设与备择假设的确定

- 应区别不同情况采取不同的建立假设方法。
  - 对于检验某项研究是否达到了预期效果
  - 一般是将研究的预期效果（希望、想要证明的假设）作为备择假设  $H_1$ ，将认为研究结果无效作为原假设  $H_0$ 。先确立备择假设  $H_1$ 。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，原假设不会轻易被拒绝，就使得希望得到的结论不会轻易被接受，从而减少结论错误。
- 例如，有研究预计，采用新技术生产后将会使某产品的使用寿命明显延长到1500小时以上。则建立的原假设与备择假设应为：  $H_0: \bar{X} \leq 1500$      $H_1: \bar{X} > 1500$
- 例如，有研究预计，改进生产工艺后会使某产品的废品率降低到2%以下。则建立的原假设与备择假设应为：  
 $H_0: \bar{X} \geq 2\%$      $H_1: \bar{X} < 2\%$

## 单侧检验原假设与备择假设的确定

### →对于检验某项声明的有效性

- 一般可将所作的声明作为原假设。将对该声明的质疑作为备择假设。先确立原假设 $H_0$ 。因为除非有证据表明“声明”无效，否则就应认为该“声明”是有效的。
- 例如，某灯泡制造商声称，该企业生产的灯泡平均使用寿命在1000小时以上。通常除非样本能提供证据表明使用寿命在1000小时以下，否则就应认为厂商的声称是正确的。建立的原假设与备择假设应为：

$$H_0: \bar{X} \geq 1000 \quad H_1: \bar{X} < 1000$$

- 对于上述问题还可以结合不同背景建立假设。同样的问题背景不同可以采用不同的原假设。
- 例如，一商店经常从某工厂购进某种商品，该商品质量指标为  $\bar{X}$ ， $\bar{X}$ 值愈大商品质量愈好。商店提出的进货条件是按批验收，只有通过假设“ $\bar{X} \geq \bar{X}_0$ ”检验的批次才能接受。有两种可能情况：

(1)如果根据过去较长时间购货记录，商店相信该厂产品质量好，于是同意把原假设定为  $\bar{X} \geq \bar{X}_0$ ，而且选择较低的检验显著性水平。这对工厂是有利的，使得达到质量标准的产品以很小的概率被拒收。虽然这会使商店面临接受不合标准产品的风险，但历史记录显示出现这种情况的可能性很小，而且商店也可因此获得较好的货源。

(2)如果过去一段时期的记录表明，该厂产品质量并不理想，商店则会坚持认为  $\bar{X} \leq \bar{X}_0$  为原假设，并选定较小的检验显著性水平。这对商店是有利的，不会轻易地拒绝原假设，有  $1-\alpha$  的可能把劣质产品拒之门外。

## Analyze Results

$$N_{cont} = 10,072$$

$$N_{exp} = 9886$$

$$X_{cont} = 974$$

$$X_{exp} = 1242$$

$$\hat{P}_{pool} = \frac{974 + 1242}{10072 + 9886} = 0.111$$

$$SE_{pool} = \sqrt{0.111(1-0.111)\left(\frac{1}{10072} + \frac{1}{9886}\right)} = 0.00445$$

$$\hat{d} = \hat{P}_{exp} - \hat{P}_{cont} = \frac{1242}{9886} - \frac{974}{10072} = 0.02893$$

$$m = 1.96 \times 0.00445 = 0.008722$$

statistical significance  
= (0.0202, 0.0376)

confidence level :  $(\hat{d} - m, \hat{d} + m)$

$$= 0.008722 \cancel{+} \frac{1242}{9886}, 0.008722 \cancel{+} \frac{1242}{9886}$$

launch?

$$(0.1169, 0.1344)$$

yes. 因为  $\hat{d} > d_{min}$

$$(0.08798, 0.1054)$$

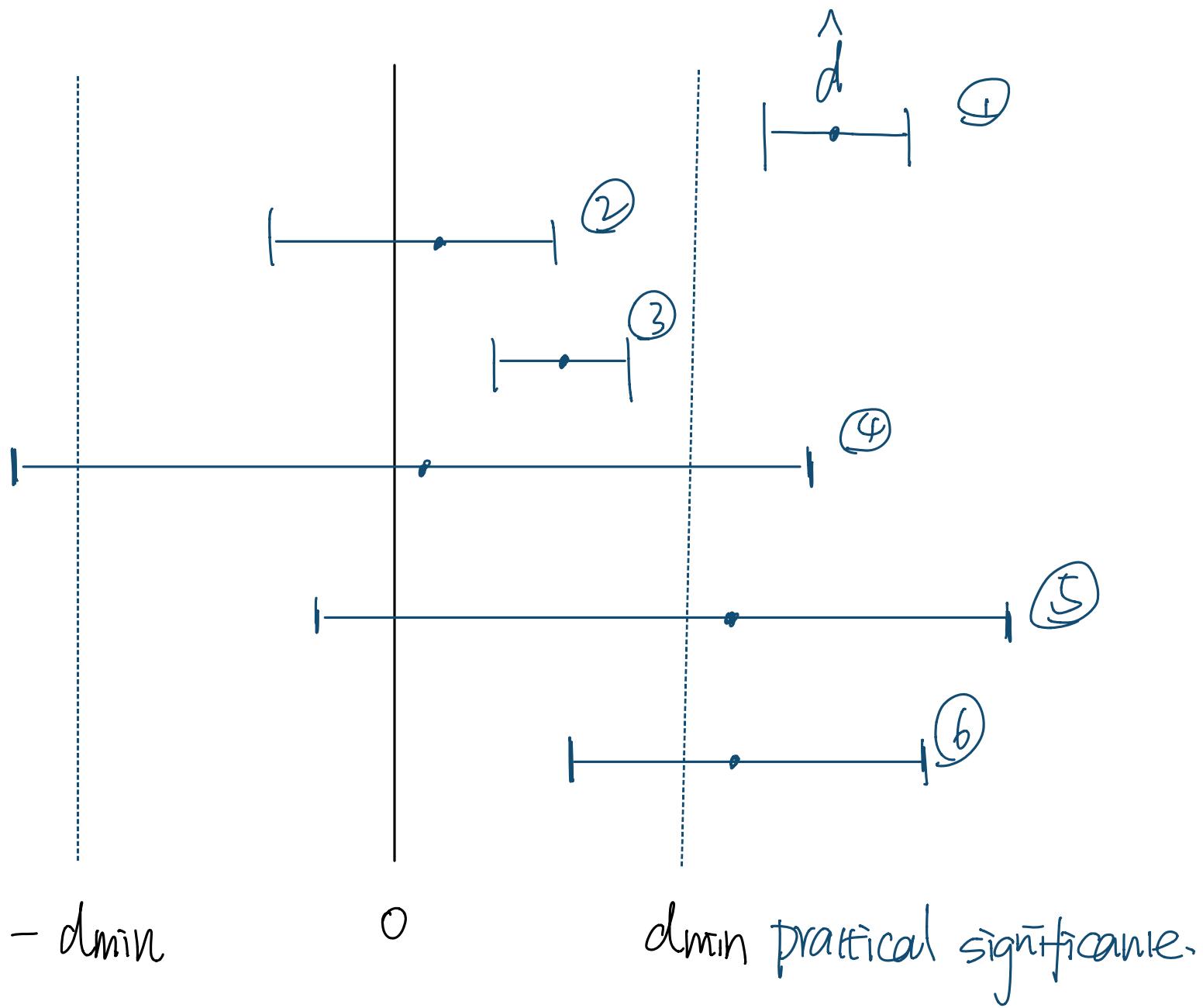
因为 confidence level 的最小值  $> 0.02 (d_{min})$

practical confidence

$$d_{min} = 0.02$$

confidence level = 95%

# Confidence Interval Cases



①  $\hat{d} > \text{practical significance} \Rightarrow$  两端都  $>$  practical  
significant, change 会顯示實際顯著性.  
 $\Rightarrow$  launch

②.本性结果.因为置信区间包括0.因此无变更会更具有统计显著性.  $\Rightarrow$  no launch.

③具有统计显著性,存在 positive change. 但不具有实际显著性.  $\Rightarrow$  no launch

④.置信区间超出了实际显著性的范围.

e.g. 会使顾客↑10% 或 ↓10%  
 $\Rightarrow$  power 不够,重新 test.

⑤ point estimate > practical significance

$\Rightarrow$ 你可以猜测这个变更是否是你关注的  
但 confidence level contain 0.

$\Rightarrow$  additional test.

⑥可以猜测存在一个 positive change, 也可能不具有

实际显著性

⇒ addition test