

《统计与数据分析》

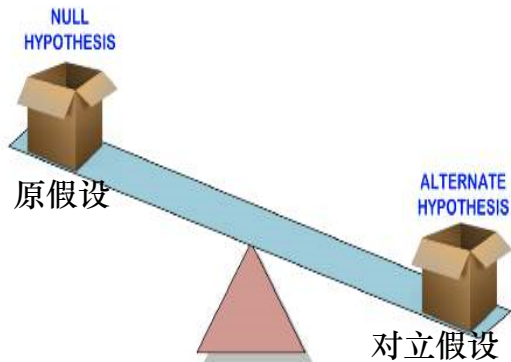
Statistics & Data Analysis



Zhu Huaqiu
@Peking University

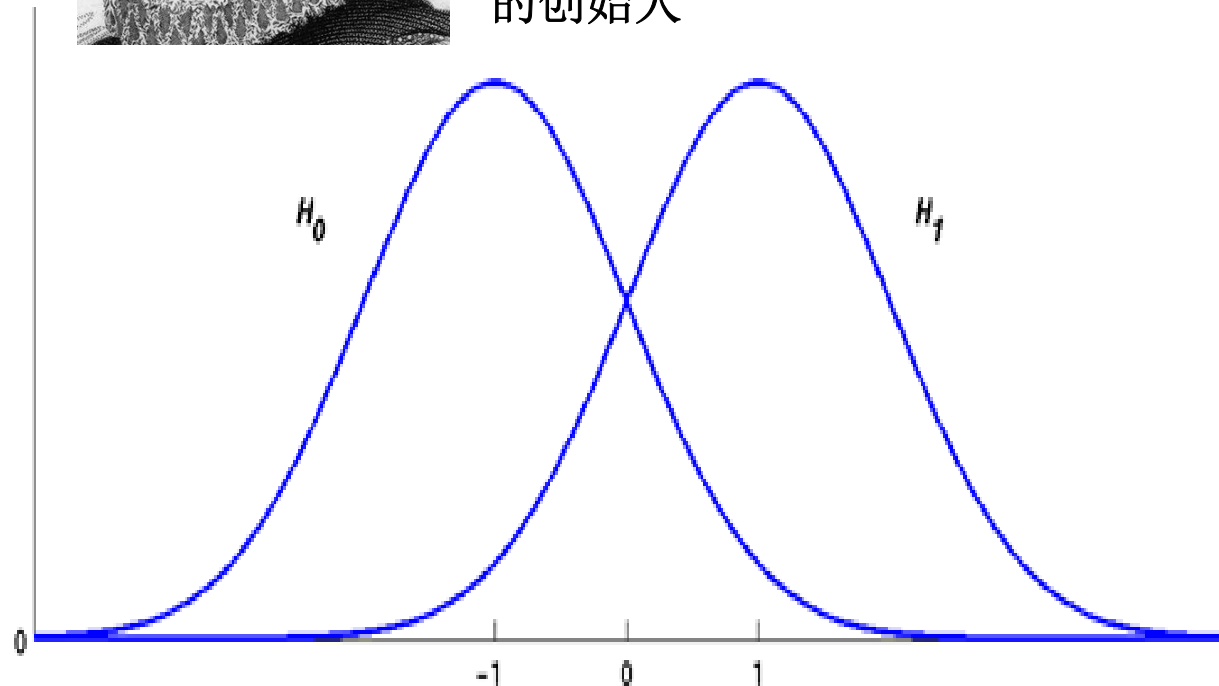
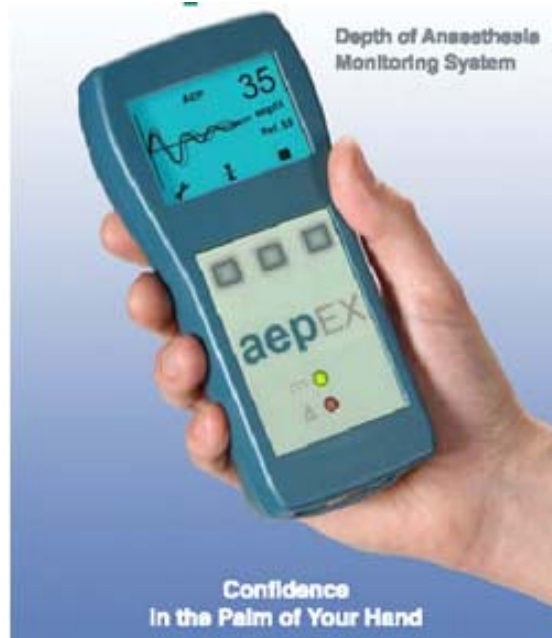
The strongest arguments prove nothing so long as the conclusions are not verified by experience. Experimental science is the queen of sciences and the goal of all *speculation*.

(Francis Bacon (1561—1626))



“英国唯物主义和整个现代实验科学的真正始祖”

唯物主义哲学家，实验科学的创始人，近代归纳法的创始人



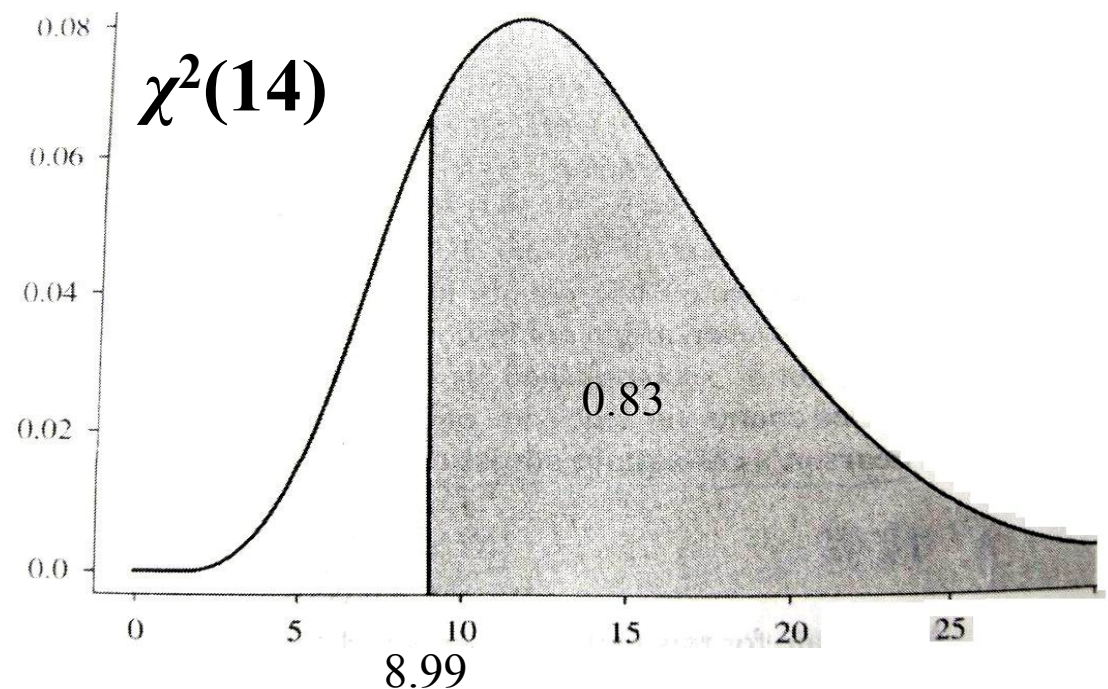
α粒子数的Poisson分布研究

n	0-2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17+
Observed	18	28	56	105	126	146	164	161	123	101	74	53	23	15	9	5
Expected	12.2	27.0	56.5	94.9	132.7	159.1	166.9	155.6	130.6	99.7	69.7	45.0	27.0	15.1	7.9	7.1
χ^2	2.76	0.04	0.01	1.07	0.34	1.08	0.05	0.19	0.44	0.02	0.27	1.42	0.59	0.00	0.57	0.57

问题 估计参数：根据观察的数据来估计Poisson分布的参数 λ

讨论：

1. 估计值的概率分布性质
2. 参数估计的数学方法
3. 总体分布的假设检验
4. 总体分布的拟合检验



假设检验 (hypothesis test) :

根据 (简单随机抽样) 抽取的 (简单随机) 样本信息来判别总体是否具有假设所定义的某种概率分布性质

—— 总体 X

—— (简单随机抽样) 样本 $X_1, X_2, \dots, X_n$

—— **原假设 (null hypothesis)** : 总体 X 的某一种概率分布性质 (参数、分布类型)

—— **对立假设 (alternative hypothesis)** : 总体 X 的另一种概率分布性质 (参数、分布类型)

§ 6.1 假设检验概述

§ 6.1.1 问题的提出

【Example 6.1】某药品生产车间用粉剂定量自动包装机包装粉剂药品，每袋标准质量为50mg。长期实践表明该设备包装的这一药品质量服从正态分布，且标准差为1.5mg。现从某天的包装产品中随机抽取9袋，精确秤得它们的质量分别为

49.5, 50.6, 51.8, 52.1, 49.3, 51.1, 52.0, 51.5, 50.0

它们的平均值为50.9mg。问当日该包装机工作是否正常？

问题：假定方差 $\sigma^2=1.5^2$ 不变，根据样本数据的平均值50.9来判断正态分布 $N(50, 1.5^2)$ 是否成立？

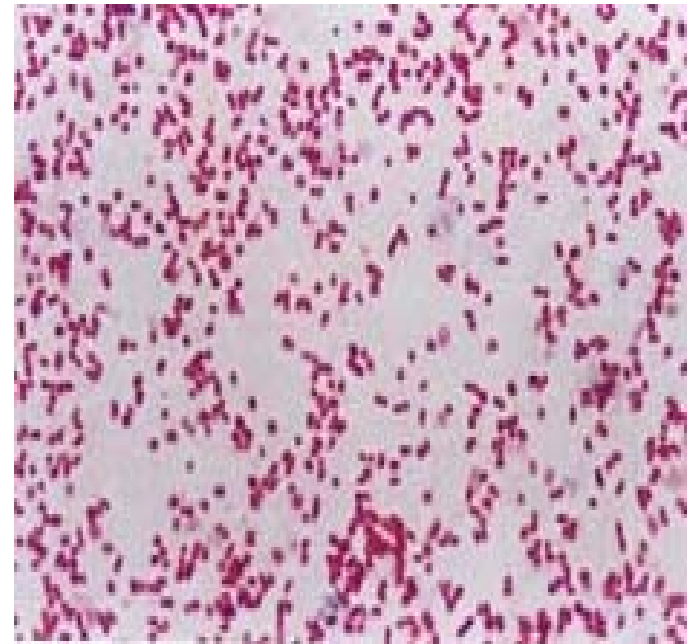
【Example 6.2】 将 0.1ml 受细菌污染的牛奶均匀涂在 1cm^2 的切片上，用显微镜观察切片每个小网格内的细菌菌落数目。根据400 (20×20) 个小网格的计数结果，统计出如下表格：

菌落数：	0	1	2	3	4	5	6	7	8	9	10	19
频 数：	56	104	80	62	42	27	9	9	5	3	2	1

试问单位面积内的菌落数是否服从Poisson分布？

问题：

根据样本信息，判别随机变量 X 是否服从Poisson分布。



(Bliss C. and Fisher R. A., (1953). Fitting the negative binomial distribution to biological data. *Biometric*, 9: 174-200)

参数假设检验问题（【例6.1】）

- 已知总体的分布函数类型，对分布函数中的未知参数提出某种假设
- 要求利用样本的信息对所提出的假设进行检验，根据检验结果作出接受或者拒绝所提假设的判断

非参数假设检验问题（【例6.2】）

- 总体的分布类型未知
- 要求根据样本的信息来对分布类型的假设进行检验，对总体的分布类型作出判断

§ 6.1.2 假设检验的基本思想与 Neyman-Pearson理论

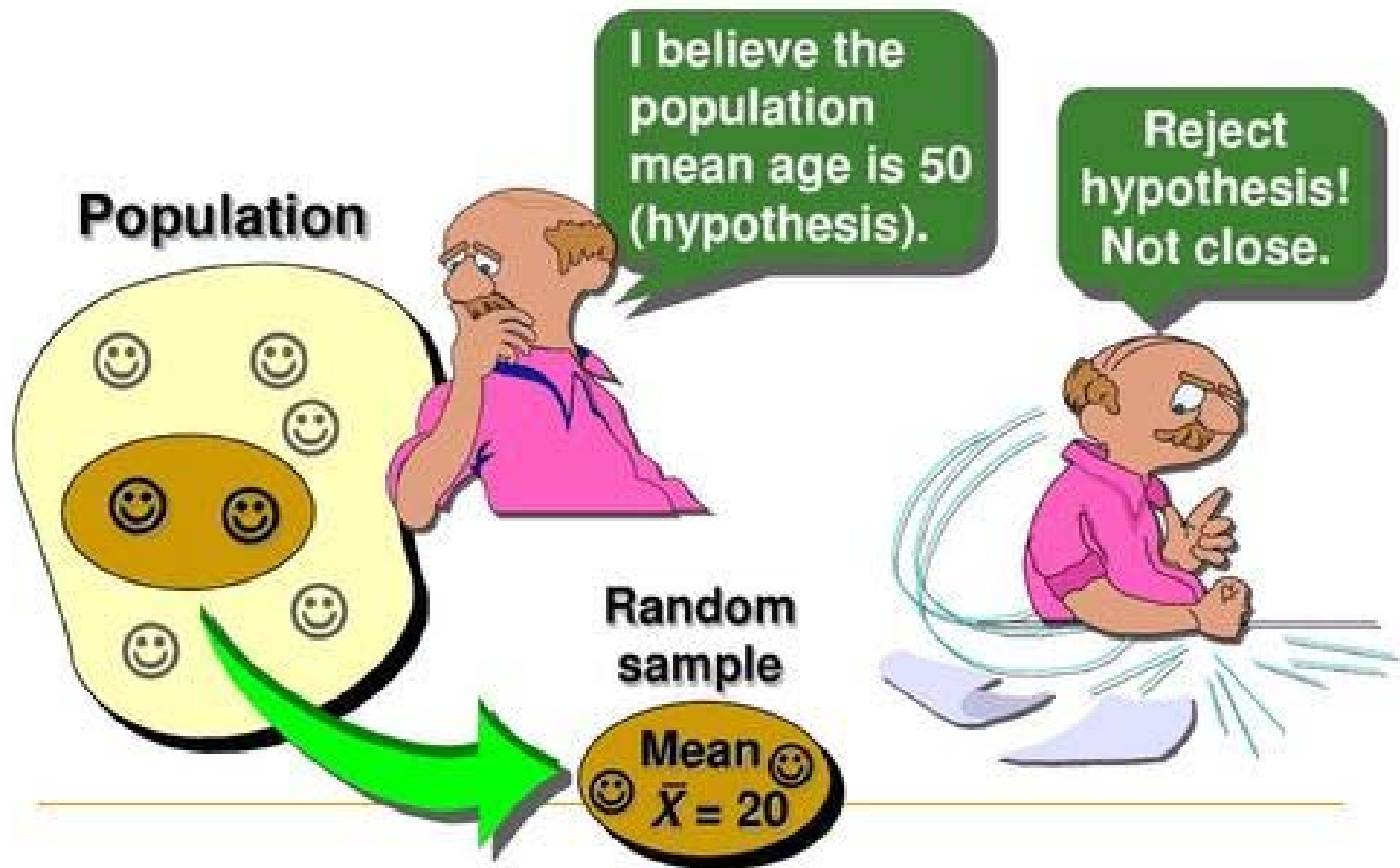
【Example 6.1】某药品生产车间用粉剂定量自动包装机包装粉剂药品，每袋标准质量为50mg。长期实践表明该设备包装的这一药品质量服从正态分布，且标准差为1.5mg。现从某天的包装产品中随机抽取9袋，精确秤得它们的质量分别为

49.5, 50.6, 51.8, 52.1, 49.3, 51.1, 52.0, 51.5, 50.0

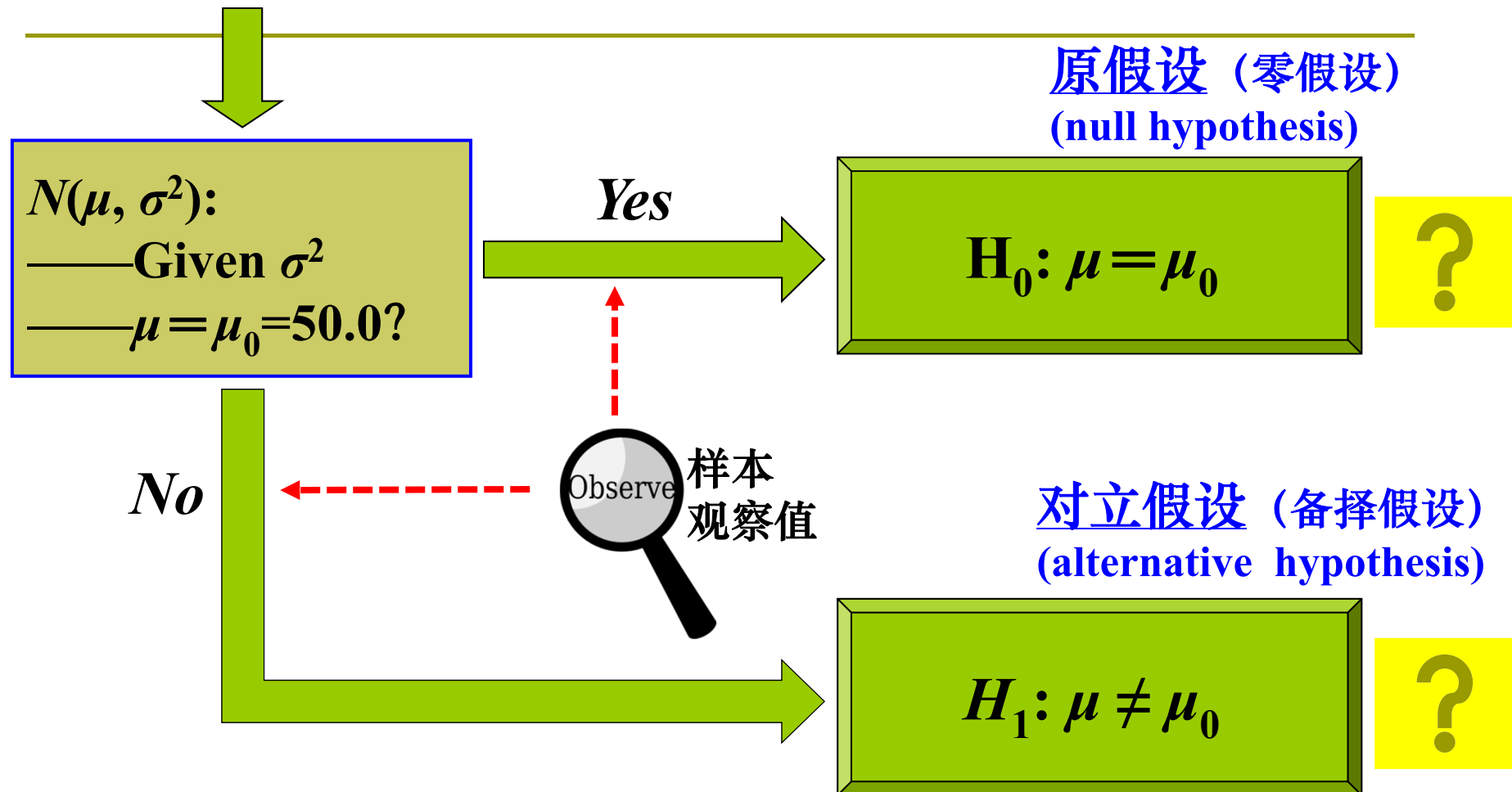
它们的平均值为50.9mg。问当日该包装机工作是否正常？

Hypothesis Testing

作判断并非总能靠显而易见……



假设方差 σ^2 不变，根据样本的信息，正态分布 $N(\mu, \sigma^2)$ 的 $\mu = \mu_0 = 50.0$ 是否成立？



H_0 and H_1 : which can we accept based on the sampling evidence?

【Example 6.1】某药品生产车间用粉剂定量自动包装机包装粉剂药品，每袋标准质量为50mg。长期实践表明该设备包装的这一药品质量服从正态分布，且标准差为1.5mg。现从某天的包装产品中随机抽取9袋，精确秤得它们的质量分别为

49.5, 50.6, 51.8, 52.1, 49.3, 51.1, 52.0, 51.5, 50.0

它们的平均值为50.9mg。问当日该包装机工作是否正常？

问题：给定 $\sigma^2 = 1.5^2$, $H_0: \mu = \mu_0 = 50.0$; $H_1: \mu \neq \mu_0$

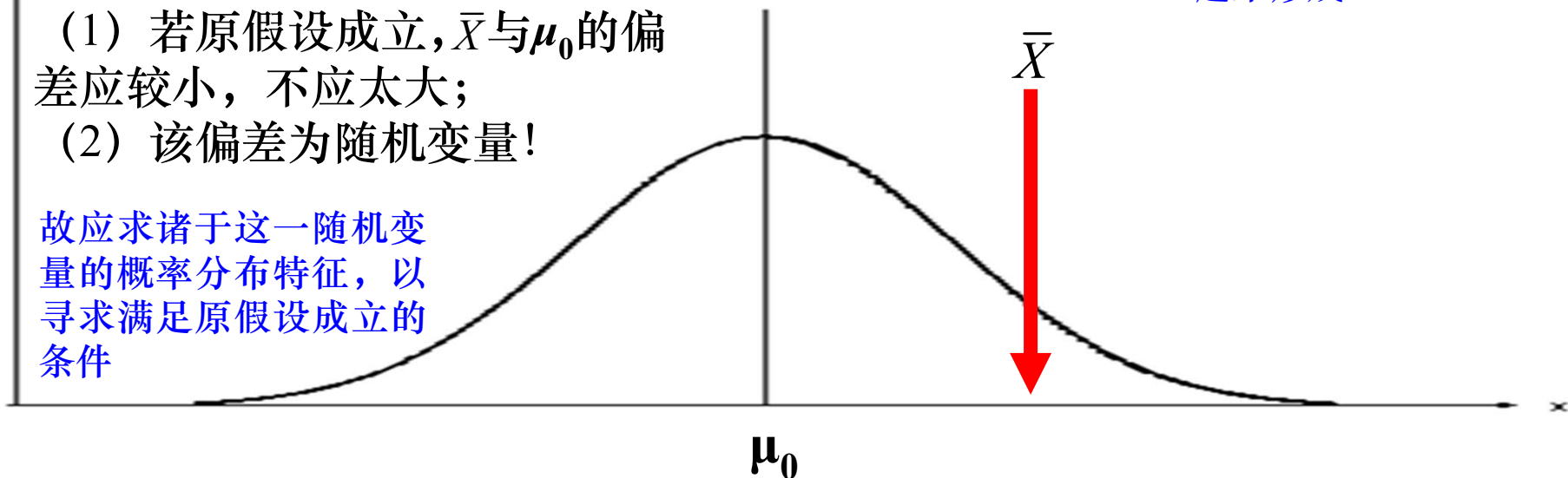
$$H_0 \Rightarrow E(\bar{X}) = \mu_0$$

原假设、对立假设的内容要根据人们关心的问题来形成！

基本思想：

- (1) 若原假设成立, $\bar{X}$ 与 μ_0 的偏差应较小, 不应太大;
- (2) 该偏差为随机变量!

故应求诸于这一随机变量的概率分布特征, 以寻求满足原假设成立的条件



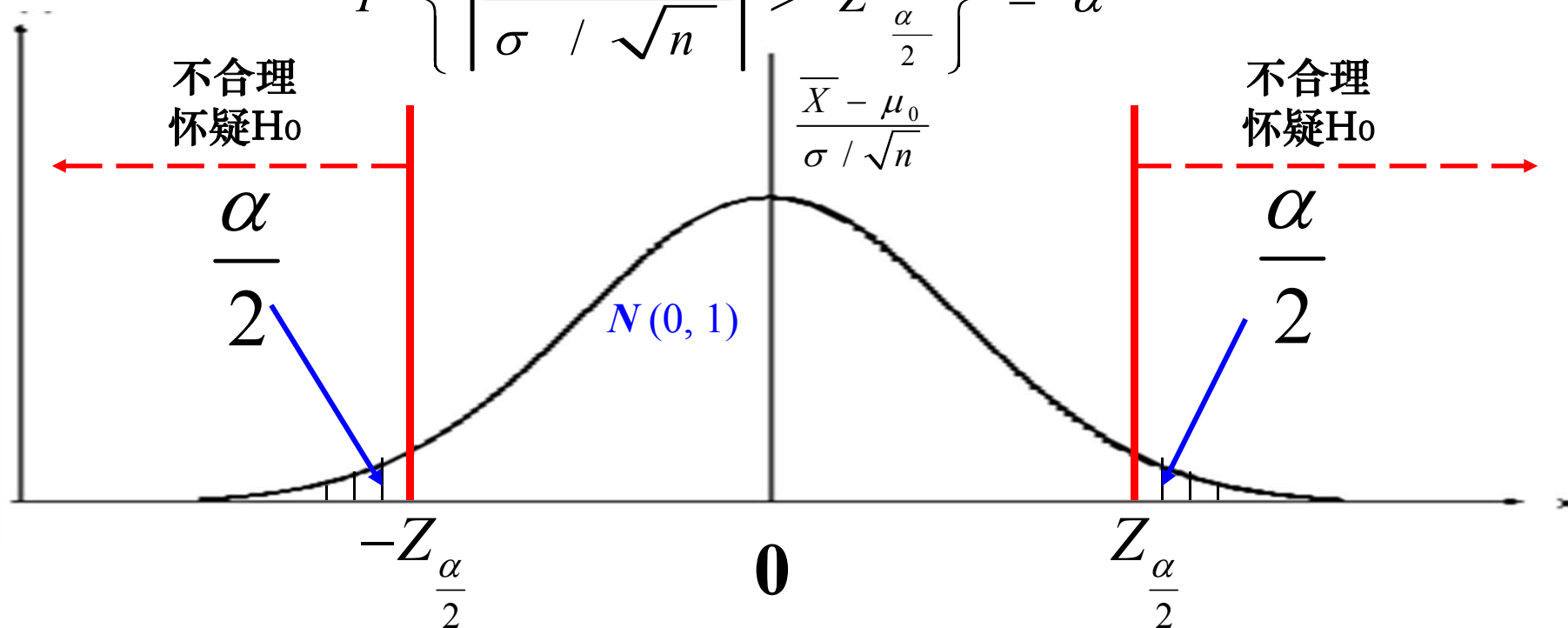
若 H_0 成立，则有：

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

中心极限定理

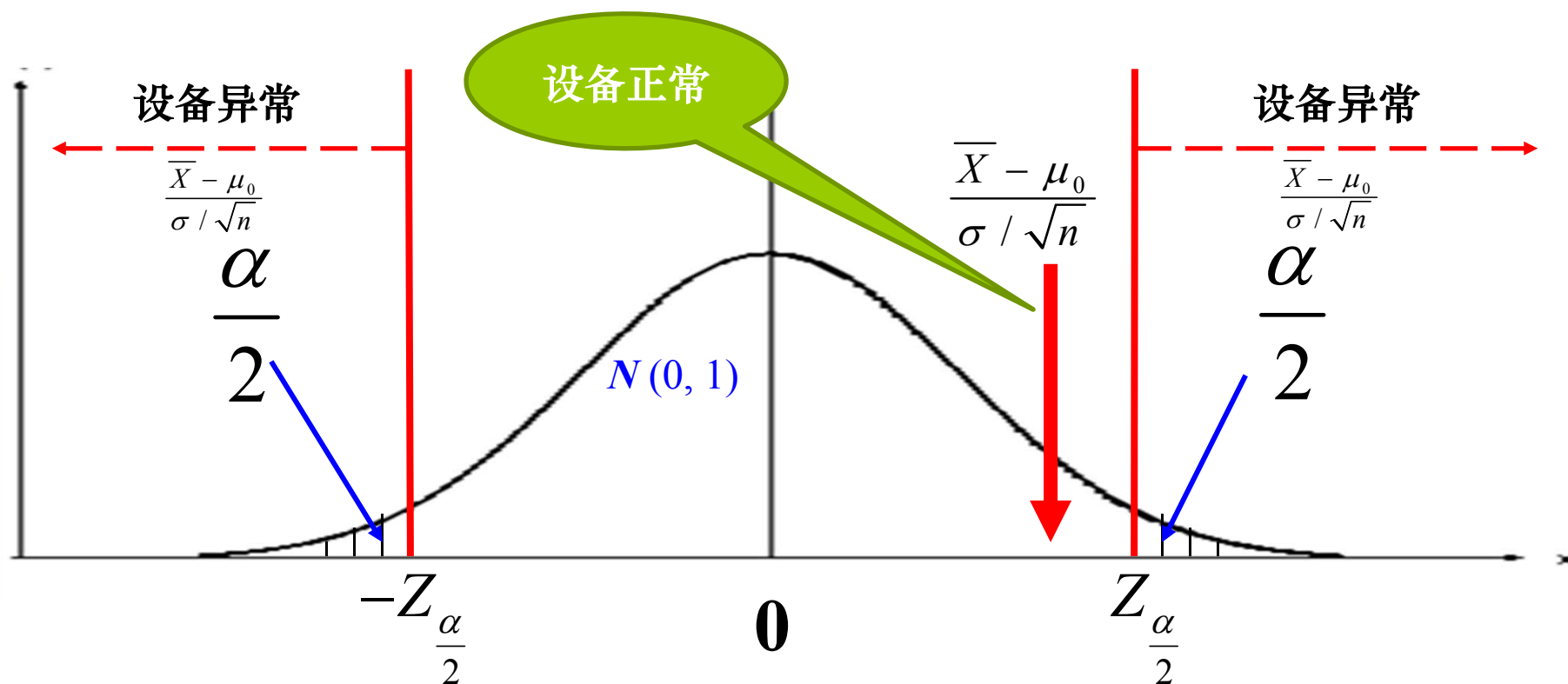
由于随机性，偏差总是存在，但偏差越大的可能性越小。因此，偏差太大的事件，其概率也很小。这种小概率事件一旦发生，对 H_0 而言是不合理的（亦即 H_0 值得怀疑）。由此，可以构造概率为 α 的小概率事件为不合理的后果（如 $\alpha=0.001, 0.01, 0.05, 0.10$ ）：

$$P \left\{ \left| \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \right| > Z_{\frac{\alpha}{2}} \right\} = \alpha$$



令 $\alpha = 0.05$ 有 $Z_{\frac{\alpha}{2}} = 1.96$
 $\bar{X} = 50.9, \quad \mu_0 = 50.0$

$$\left| \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \right| = \left| \frac{50.9 - 50.0}{1.5 / \sqrt{9}} \right| = 1.8 < Z_{\frac{\alpha}{2}} = 1.96$$



若 H_0 成立：

U检验or
Z检验

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

检验统计量
test statistic

构造不合理的小概率事件（如小概率 $\alpha=0.001, 0.01, 0.05, 0.1$ ）：

$$P \left\{ \left| \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \right| > Z_{\frac{\alpha}{2}} \right\} = \alpha$$

显著性水平
significance level

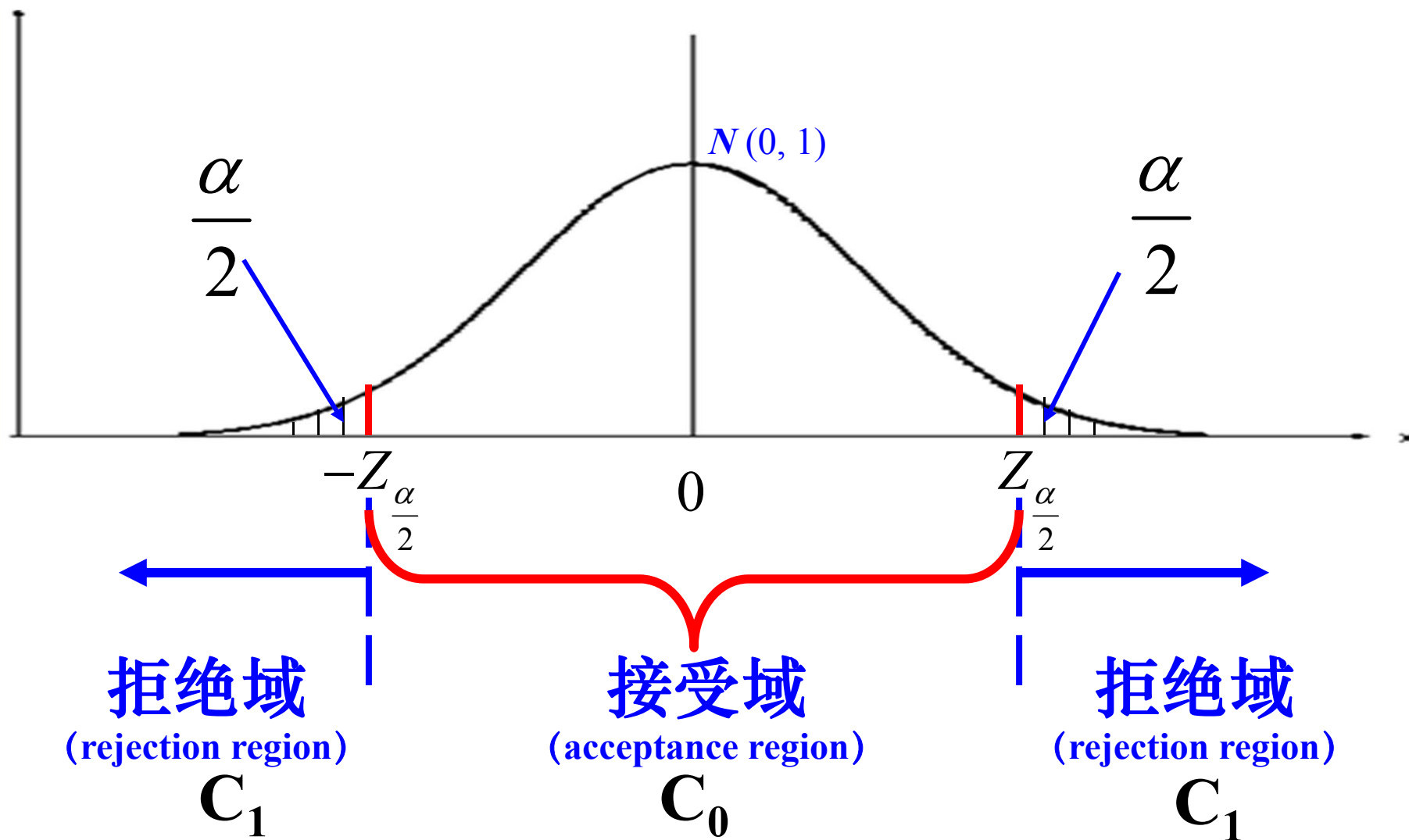
检验统计量 (test statistic)

为检验原假设和对立假设，在原假设成立的条件下，根据样本（随机变量）、样本容量和总体参数构造的随机变量，具有较明确的、或渐进的概率分布性质。

显著性水平 (significance level)

根据检验统计量的（渐进）概率分布性质，定义不合理的小概率事件来推断拒绝原假设成立的条件。

$$\frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$



Jury system

Hypothesis Testing

Hypothesis

H_0 : The defendant is not guilty

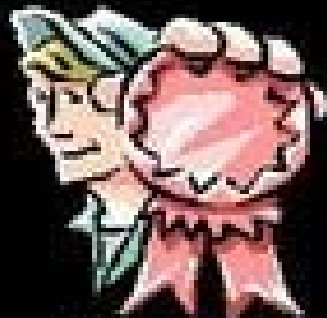
H_1 : Otherwise



Judge

Hypothesis : Assume
the accuse is not
guilty

形成原假设的原则：
保护被告原则
——无罪推断，疑罪从无



Plaintiff

Keep supplying evident
to proof the defendant is
guilty

Evident not strong
enough to prove me
wrong



Defendant

Keep rejecting the
evident supply by the
plaintiff

假设检验的检验错误

检验错误：抽样的随机性、实际推断中小概率事件仍然有可能发生

第一类错误 (Type I error) (弃真错误)：

原假设 H_0 为真，但经过检验统计量观察值的判断， H_0 被拒绝。

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

第二类错误 (Type II error) (纳伪错误)：

原假设 H_0 非真，但经过检验统计量观察值的判断， H_0 被接受。

$$\begin{aligned} &P\{\text{接受}H_0|H_0\text{非真}\}=\beta \\ \text{或} \quad &P\{\text{拒绝}H_0|H_0\text{非真}\}=1-\beta \end{aligned}$$

假设检验的两类错误

α : 显著性水平

$(1 - \beta)$: 检验功效 (Power of test)

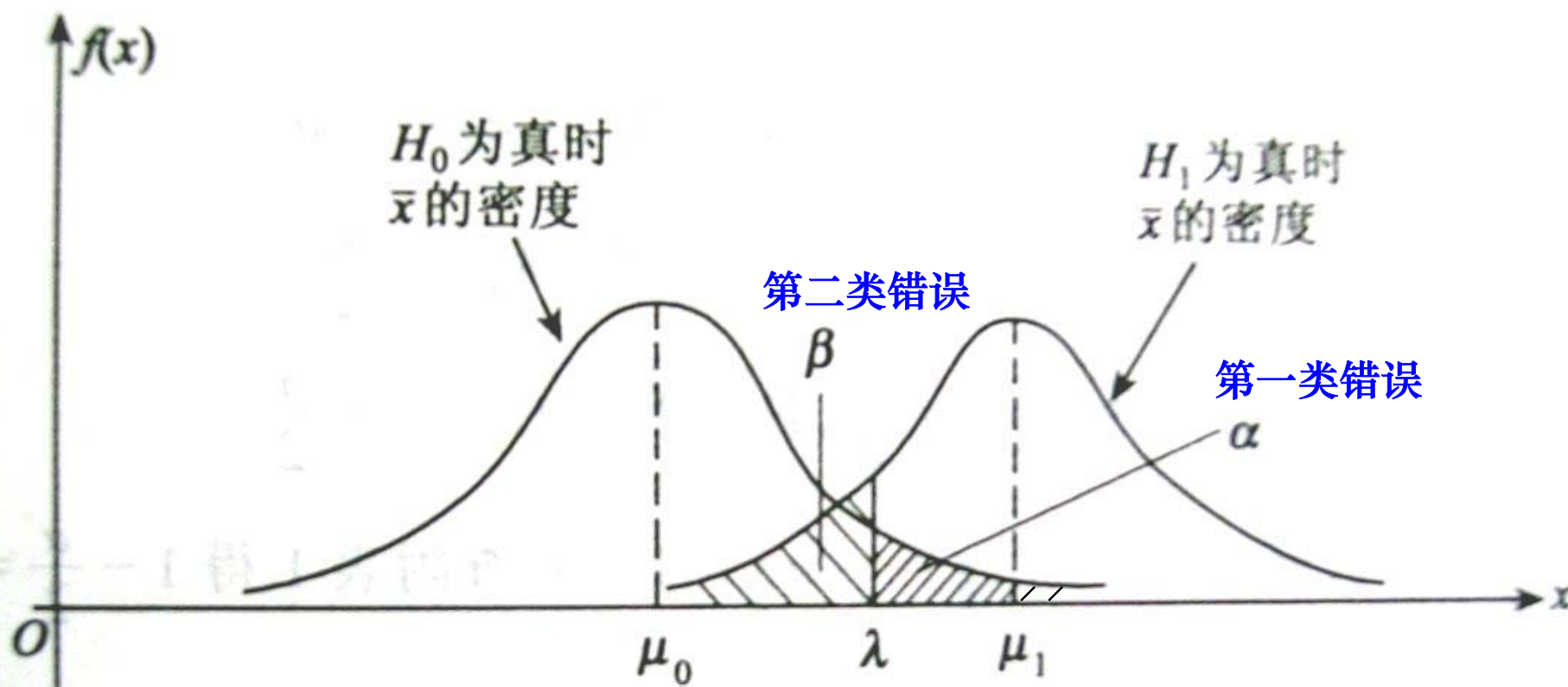
$(1 - \alpha)$: 置信水平 (Confidence Level)

第一类错误 α 与第二类错误 β 的关系

【Example 6.3】设总体 $X \sim N(\mu, \sigma^2)$ ， σ^2 已知，而 μ 只能取两个值 μ_0 、 μ_1 ，且 $\mu_0 < \mu_1$ ，现有总体 X 中抽取的容量为 n 的样本 $X_1, X_2, \dots, X_n$ ，在显著性水平 α 下检验假设：

$$H_0: \mu = \mu_0; H_1: \mu = \mu_1 (> \mu_0)。$$

请讨论第一类错误与第二类错误的关系。



解：当 H_0 成立时，构造 U 统计量：
$$U = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0, 1)。$$

在显著性水平 α 下，得到： $P\{\text{拒绝 } H_0 | H_0 \text{ 为真}\} = \alpha$ ，即：

$$\begin{aligned} P\left\{\frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \geq Z_{1-\alpha} \middle| H_0 \text{ 为真}\right\} &= P\left\{\bar{X} \geq \left(\mu_0 + \sigma Z_{1-\alpha} / \sqrt{n}\right) \middle| H_0 \text{ 为真}\right\}, \\ &= P\left\{\bar{X} \geq \lambda \middle| H_0 \text{ 为真}\right\} = \alpha \end{aligned}$$

此为第一类错误的概率，如图所示的面积，见图。

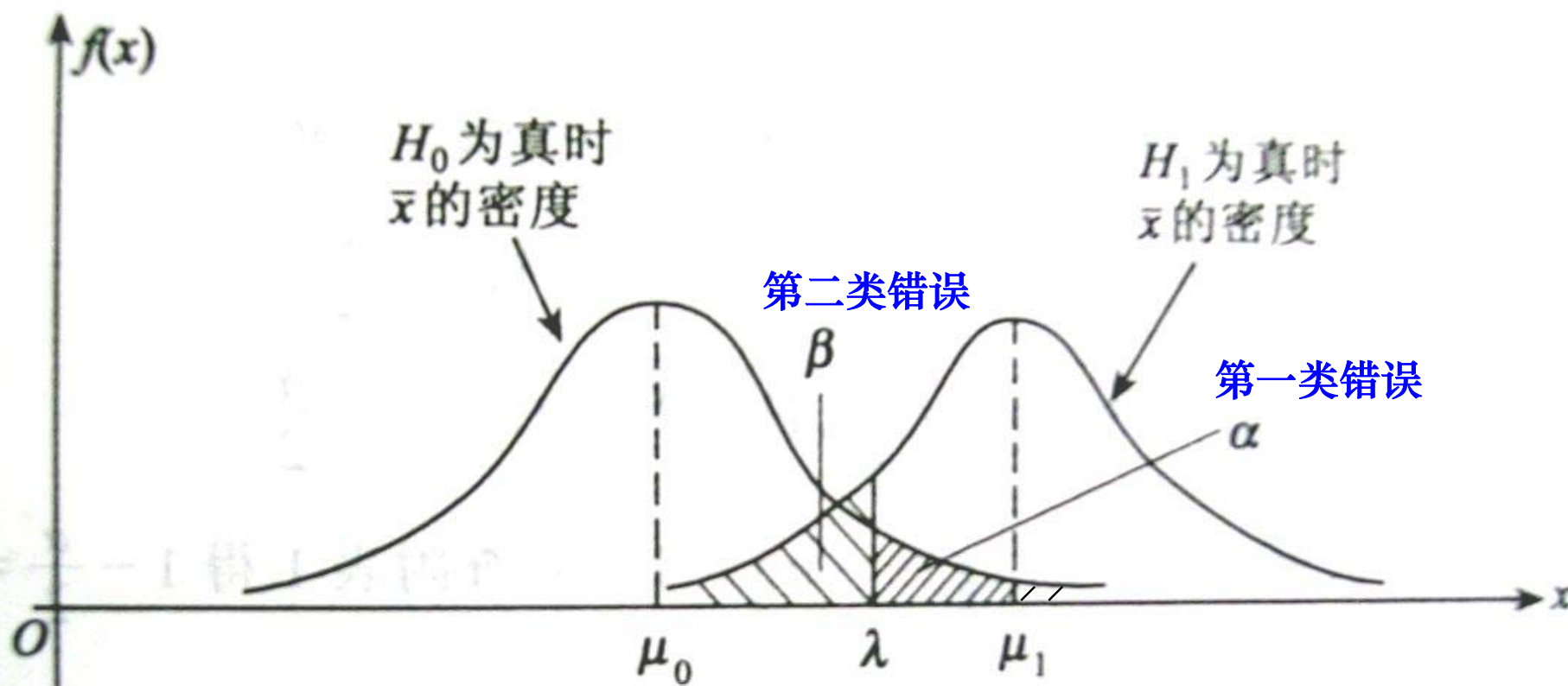
接着看第二类错误：由定义知， $\beta = P\{\text{接受 } H_0 | H_0 \text{ 非真}\}$ ，即

$$\begin{aligned} \beta &= P\left\{\frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} < Z_{1-\alpha} \middle| H_0 \text{ 非真}\right\} \\ &= P\left\{\bar{X} < \lambda \middle| H_0 \text{ 非真}\right\} = P\left\{\bar{X} < \lambda \middle| H_1 \text{ 为真}\right\} \end{aligned}$$

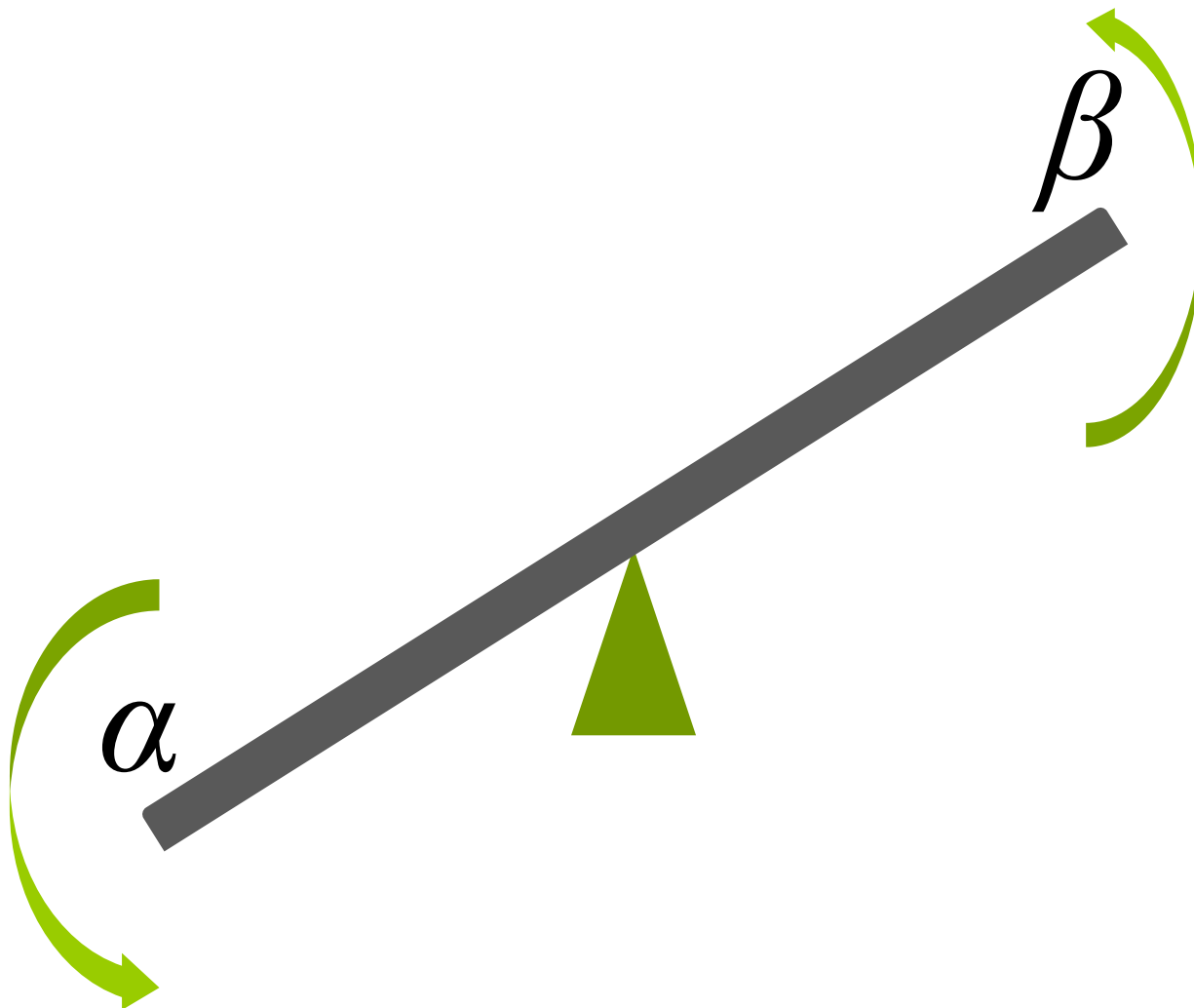
如图所示的面积，见图。

讨论:

- (1) 在其它条件不变时, α 减小 (即 λ 右移), 则 β 增大; 反之, α 增大 (即 λ 左移), 则 β 减小。二者不能同时减小;
- (2) 对于确定的 α , μ_1 越接近 μ_0 , β 越大;
- (3) 在其它条件不变时, 增大样本容量 n 可有效地减小 β 。

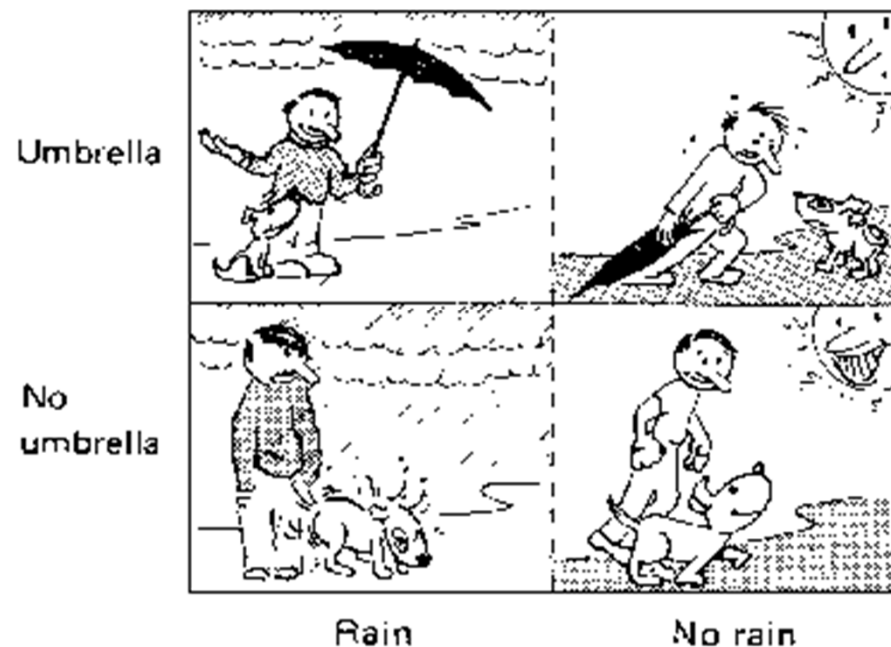


第一类错误 α 与第二类错误 β 的关系



Two types of error

It's cloudy this morning, and I wonder whether or not I should take my umbrella. There are two possible kinds of error: (1) either I take my umbrella and it doesn't rain, or (2) I don't take it and it does rain. We can't prevent these errors, but we can control them by weighing the *relative costs* of one type versus the other. Most people would say that carting around an unused umbrella is less of a cost than getting soaked in a downpour, but each individual has to decide.



	Hypothesis true	Hypothesis false
Accept. hypothesis		Type II error
Reject hypothesis	Type I error	

假设检验的Neyman-Pearson理论

在控制出现第一类错误 α 的条件下，寻求使出现第二类错误 β 尽可能小（或检验功效 $1-\beta$ 尽可能大）的检验。

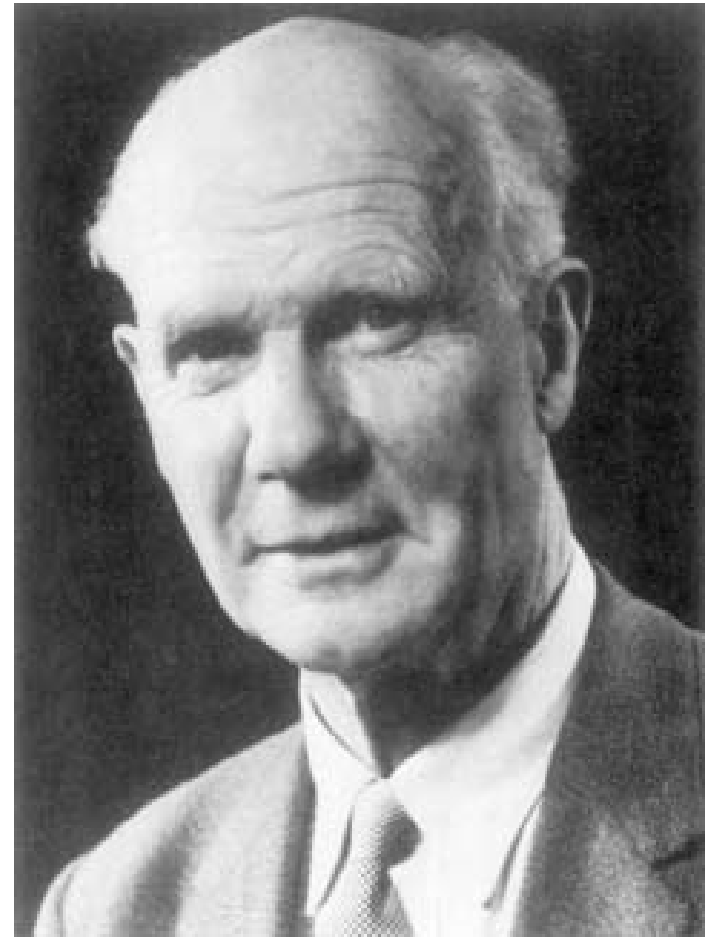
在实际操作中，当样本容量 n 给定时，优先考虑对出现第一类错误的概率并加以控制，再适当考虑出现第二类错误的概率，即显著性检验优先的基本原则。

——N-P理论强调控制第一类错误的核心地位

Jerzy Neyman
(1894-1981)
美国波兰裔统计学家



Egon Pearson
(1895-1980)
英国统计学家



1928—38年期间合作发表一系列文章，建立了假设检验的一种严格的数学理论。

判别决策的统计学基础

	H_0 is true	H_0 is false (H_1 is true)
Accept H_0	<u>Correct Decision</u>  $1 - \alpha$: Confidence Level	Type II Error  β
Reject H_0	Type I Error  α : Significance level	<u>Correct Decision</u>  $1 - \beta$: Power of a test

Sensitivity (Sn), Specificity (Sp), PPV (Positive Predictive Value), and NPV (Negative Predictive Value)

		The Truth		
		(H_0 is true) Has the disease	(H_0 is false) Does not have the disease	
Test Score:	Positive (Accept H_0)	True Positives (TP)  a	False Positives (FP)  b Type II Error	$PPV = \frac{TP}{TP + FP}$
	Negative (Reject H_0)	False Negatives (FN)  c Type I Error	True Negatives (TN)  d	$NPV = \frac{TN}{TN + FN}$

Sensitivity (Sn)

$$\frac{TP}{TP + FN}$$

Or,

$$\frac{a}{a + c}$$

Specificity (Sp)

$$\frac{TN}{TN + FP}$$

$$\frac{d}{d + b}$$

注意：
统计检验的阳性(positive)、阴性(negative)、假阳性(false positive)与假阴性(false negative)与医学诊断的概念有时不一致！

Sensitivity (Sn):

$$A/(A + C) \times 100$$

$$20/30 \times 100 = 66.7\%$$

Specificity (Sp):

$$D/(D + B) \times 100$$

$$37/70 \times 100 = 52.9\%$$

Positive Predictive Value (PPV):

$$A/(A + B) \times 100$$

$$20/53 \times 100 = 37.7\%$$

Negative Predictive Value (NPV):

$$D/(D + C) \times 100$$

$$37/47 \times 100 = 78.7\%$$

Test Result	Truth		
	Disease (number)	Non Disease (number)	Total (number)
Positive (number)	20 A (True Positive)	33 B (False Positive)	53 $T_{\text{Test Positive}}$
Negative (number)	10 C (False Negative)	37 D (True Negative)	47 $T_{\text{Test Negative}}$
	30 T_{Disease}	70 $T_{\text{Non Disease}}$	100 Total

讨论:

- (1) Sn与PPV分别对应控制Type I error与控制Type II error的水平或精度
- (2) Sp与NPV分别对应控制Type I error与控制Type II error的水平或精度

原假设的分类

简单假设 (Simple hypothesis)

- 指定了总体X的分布形式和参数具体的取值，
如： $X \sim N(\mu, \sigma^2)$ ， σ^2 已知， $H_0: \mu = \mu_0$ (点信息)
- 简单假设的参数自由度为0；

复合假设 (Composite hypothesis)

- 对于总体X的分布形式未知，或者虽然指定了分布形式但参数数值没有完全具体指定，即复合假设的参数自由度 ≥ 1 。

如：

- (1) $X \sim N(\mu_0, \sigma^2)$ ， μ_0 已知， σ^2 未指定 (区间信息)。
- (2) $X \sim$ 某种Poisson分布，其参数 λ 未指定 (区间信息) ；
- (3) $X \sim B(100, p)$ ， $p > 0.5$ (区间信息)。

对立假设的分类

单边对立假设 (one-sided alternative hypothesis)

在原假设指定或部分指定参数时，原假设参数的互补区域（即对立假设的参数区间）由一个连续区间构成

单边假设对应的假设检验问题称为单边检验。

双边对立假设 (two-sided alternative hypothesis)

在原假设指定或部分指定参数时，原假设参数的互补区域（即对立假设的参数区间）由两个不相交（不连通）的连续区间构成

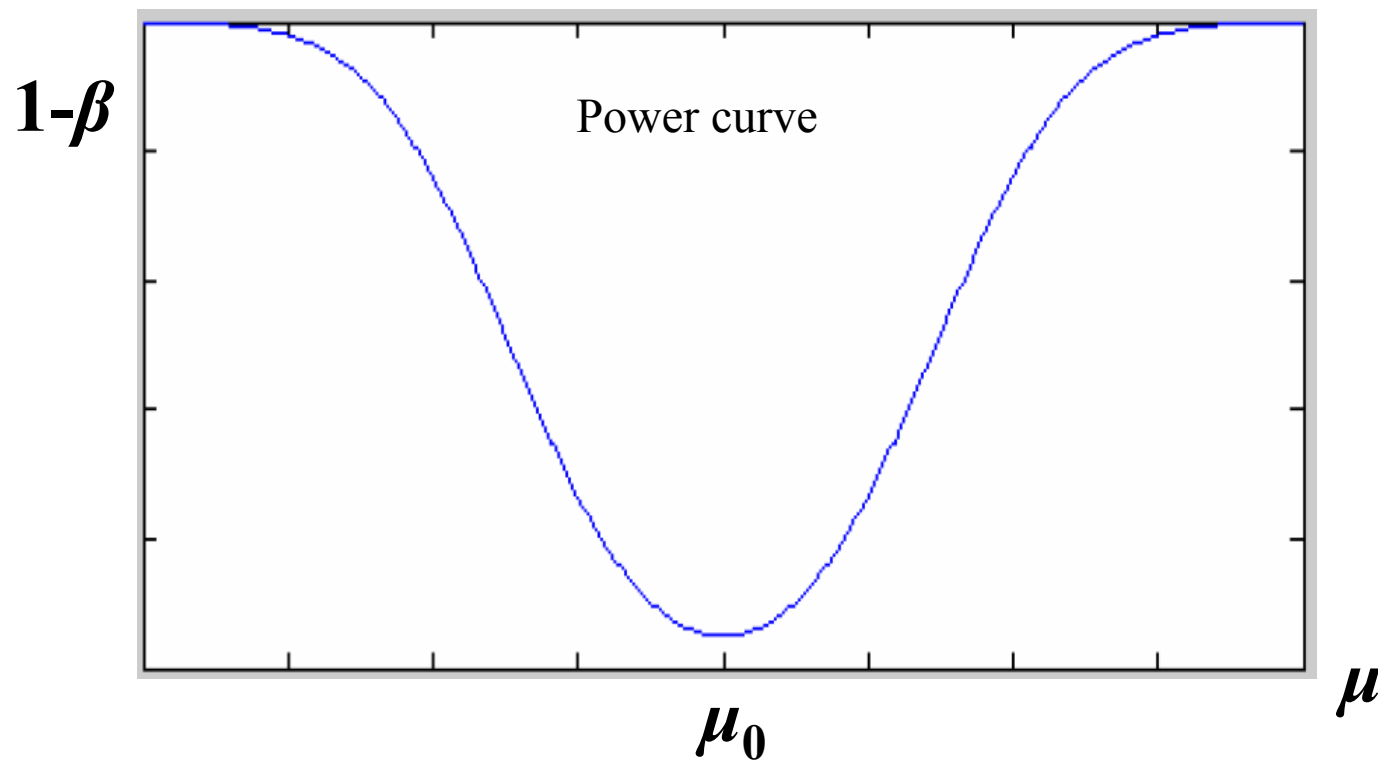
双边假设对应的假设检验问题称为双边检验。

假设检验分类通常以单边、双边分类，关心的是

- (1) 检验统计量的分布关于不合理事件的定义；
- (2) 给定显著性水平下检验功效的问题。

补充*：关于检验功效($1-\beta$)的讨论

双边检验的功效函数呈现“U”字形，“U”展得越宽，检验的功效越差。



§ 6.1.3 假设检验小结

一、假设检验过程中两个重要的经典统计学思想

(1) 反证法思想

保护原假设的思想，没有充分的理由就不能拒绝原假设。

(2) 小概率原理

小概率事件在一次试验中几乎是不会发生的，如果小概率事件发生了，就认为出现了不合理现象。

二、假设检验的一般步骤

- (1) 根据实际问题提出原假设 H_0 和对立假设 H_1 ;
- (2) 选择合理的检验统计量, 根据原假设 H_0 和对立假设 H_1 来确定拒绝域的表达式;
- (3) 选取适当的显著性水平 α , 求出使出现第一类错误的概率不超过 α 的临界值、拒绝域 C_1 (和接受域 C_0) ;
- (4) 根据样本观察值计算检验统计量的观察值, 判断属于拒绝域或接受域, 从而拒绝或接受原假设 H_0 。

§ 6.1.4 置信区间与接受域的对偶性

置信区间

以一定的置信度 $(100(1-\alpha)\%)$ 给出未知参数的所在范围。

接受域

在某一显著性水平 α 下判定未知参数是否接受原假设给定的取值。

——二者的解决途径和方法都相通

【Example 6.1】某药品生产车间用粉剂定量自动包装机包装粉剂药品，每袋标准重量为50mg。长期实践表明该设备包装的这一药品重量服从正态分布，且标准差为1.5mg。

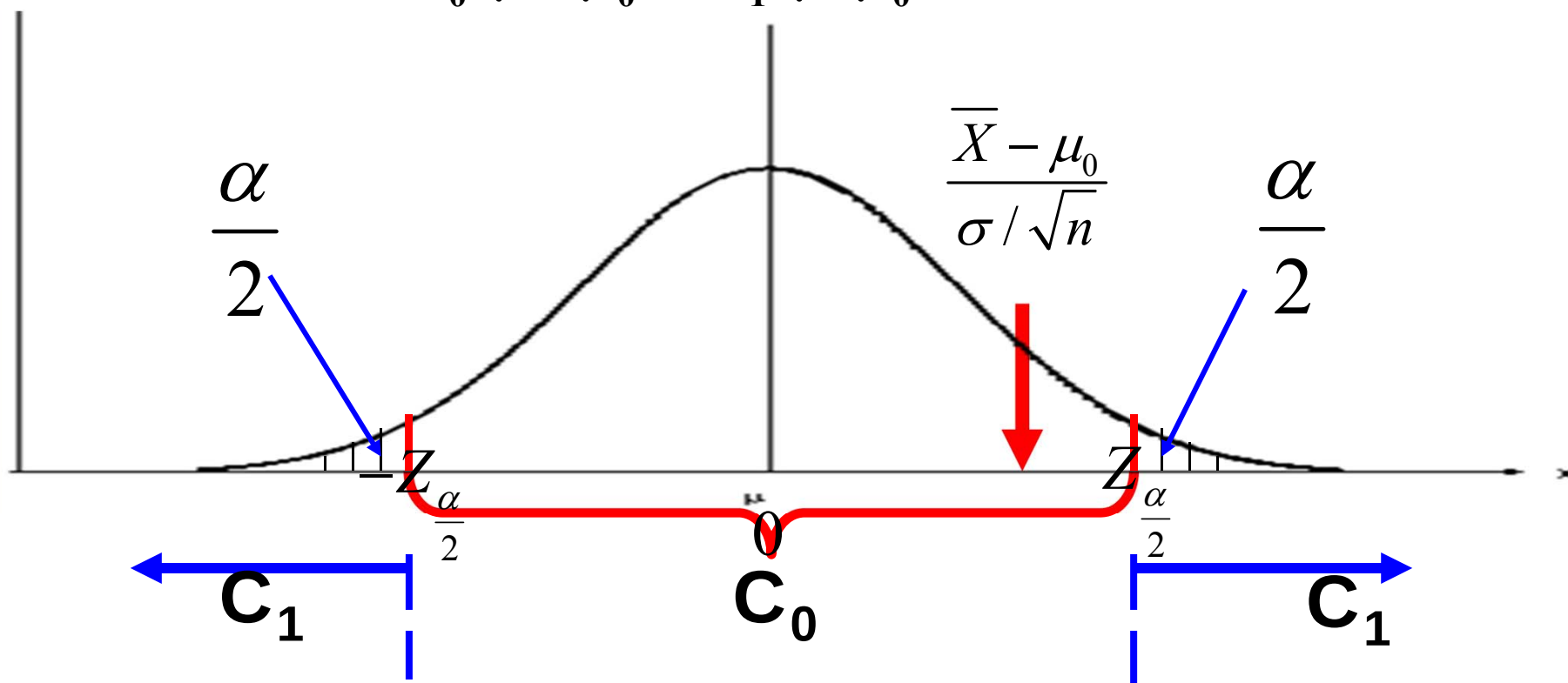
现从某天的包装产品中随机抽取9袋，精确秤得它们的重量分别为

49.5, 50.6, 51.8, 52.1, 49.3, 51.1, 52.0, 51.5, 50.0

其平均值为50.9mg。

问当日该包装机工作是否正常？

问题：给定 σ^2 ， $H_0: \mu = \mu_0$ ； $H_1: \mu \neq \mu_0$



采用检验统计量 $U = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$ ，给定显著性水平 α ，

可得到接受域形式：

$$|\bar{X} - \mu_0| \leq \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}}$$

$$\rightarrow -\frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}} \leq \bar{X} - \mu_0 \leq \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}}$$

$$\rightarrow \left(\bar{X} - \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}} \right) \leq \mu_0 \leq \left(\bar{X} + \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}} \right)$$

而上式正是关于 μ_0 的 $100(1-\alpha)\%$ 的置信区间。

可见， μ_0 的置信区间正是假设检验的接受域。

定理

设 Θ 为随机变量 X 的分布函数 $f(X|\theta)$ 的参数空间，对任一 $\theta_0 \in \Theta$ ，都存在显著性水平 α 下的对假设 $H_0: \theta = \theta_0$ 的检验，且接受域为 $A(\theta_0)$ 。则集合 $C(X) = \{X \in A(\theta)\}$ 为关于 θ 的 $100(1-\alpha)\%$ 置信区间。

定理

设 Θ 为随机变量 X 的分布函数 $f(X|\theta)$ 的参数空间，若 $C(X)$ 为关于 θ 的 $100(1-\alpha)\%$ 置信区间，即

$$P\{\theta_0 \in C(X) \mid \theta = \theta_0\} = 1 - \alpha$$

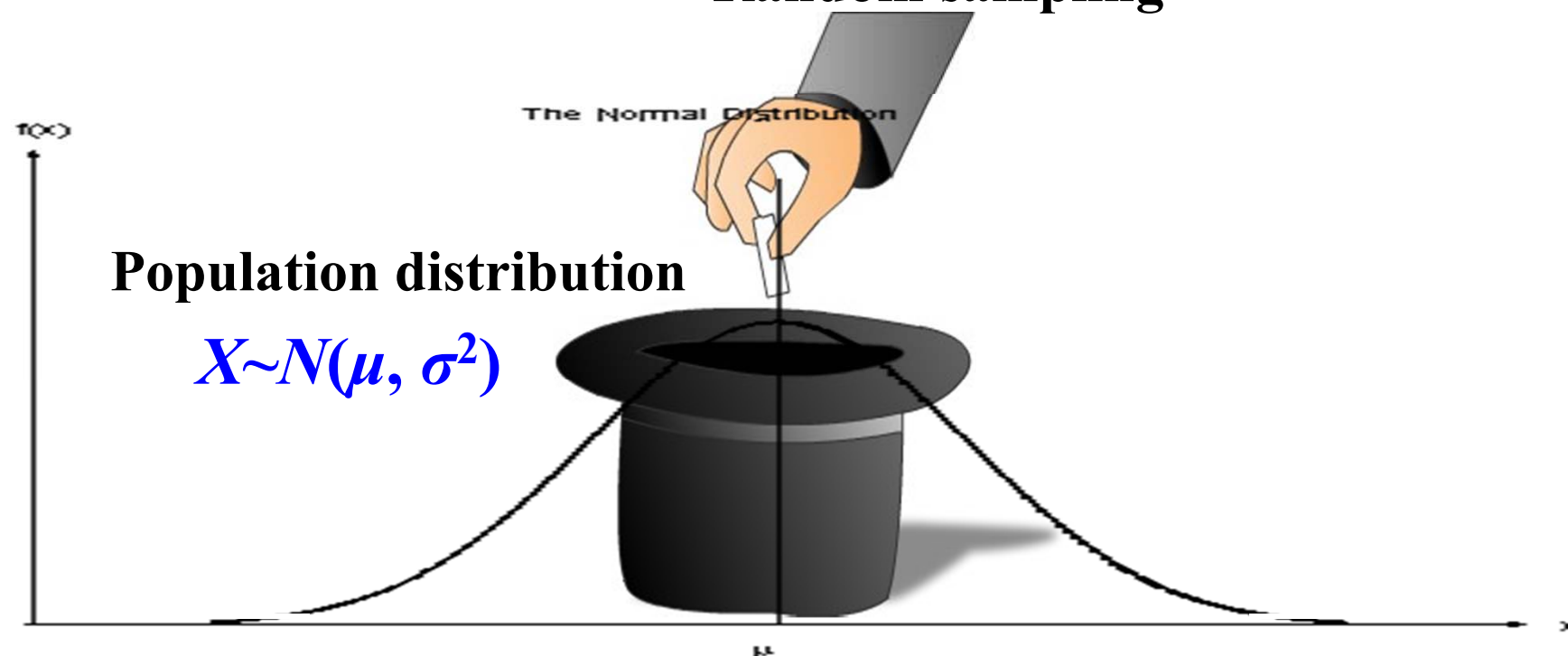
则在显著性水平 α 下的对假设 $H_0: \theta = \theta_0$ 的检验，接受域为

$$A(\theta_0) = \{X \mid \theta_0 \in C(X)\}$$

§ 6.2 正态总体的参数假设检验

Sample: $X_1, X_2, \dots, X_n$

Random sampling



§ 6.2.1 单个总体均值 μ 的检验

一、均值 μ 的双边检验

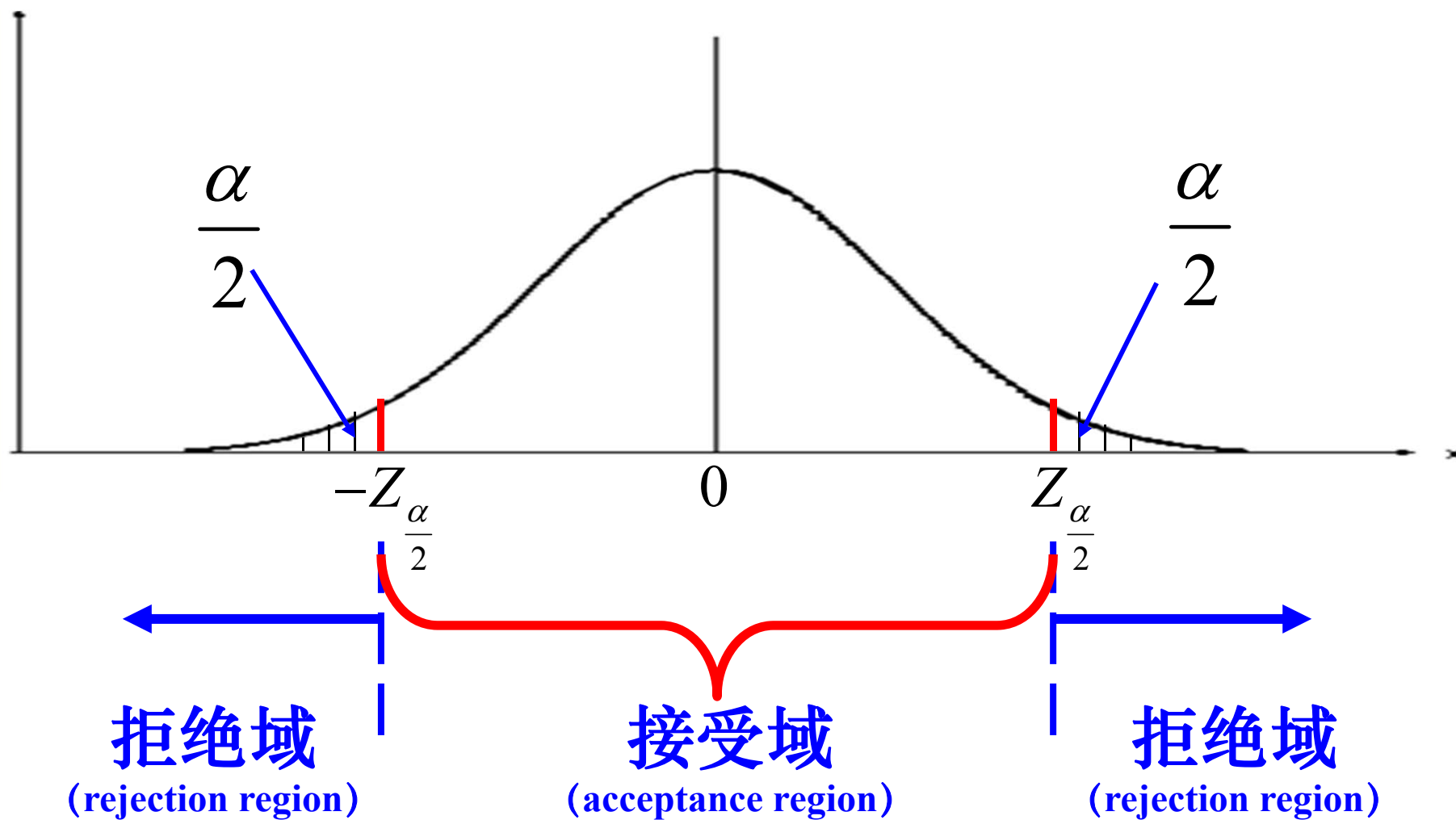
1. 已知方差 σ^2 ，检验 $H_0: \mu = \mu_0$, $H_1: \mu \neq \mu_0$

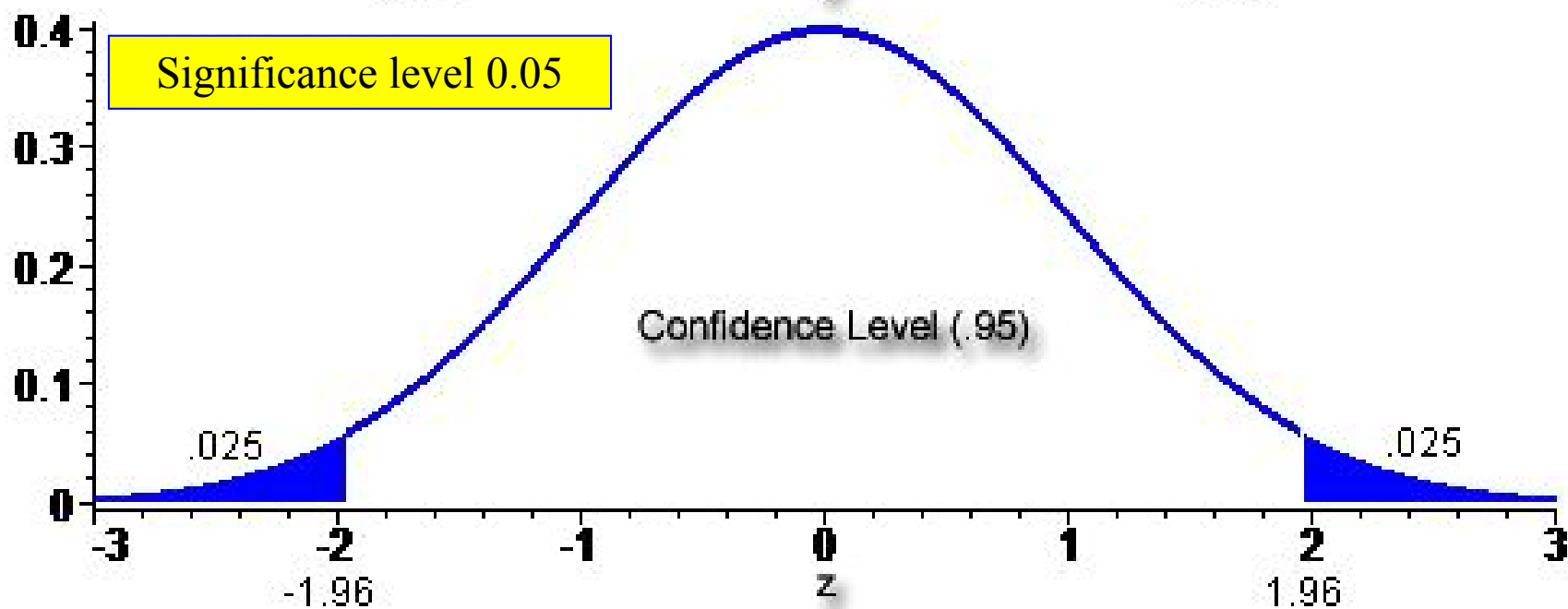
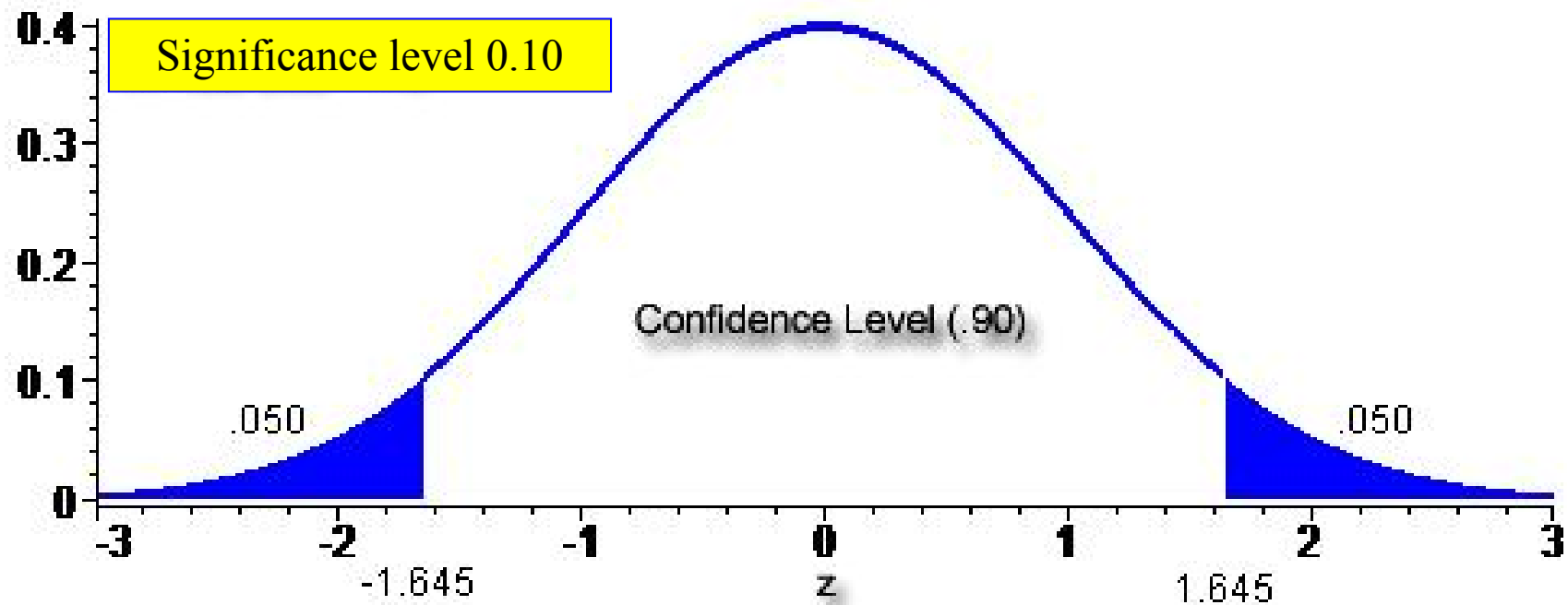
设总体 $X \sim N(\mu, \sigma^2)$ ， σ^2 已知， $X_1, X_2, \dots, X_n$ 是来自总体 X 的（简单随机抽样）样本。

采用 U 检验，检验统计量为：

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

$$\frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$





2. 未知方差 σ^2 ，检验 $H_0: \mu = \mu_0$, $H_1: \mu \neq \mu_0$

设总体 $X \sim N(\mu, \sigma^2)$, σ^2 未知, $X_1, X_2, \dots, X_n$ 是来自总体 X 的样本。

选取的检验统计量为（即**T检验**）：

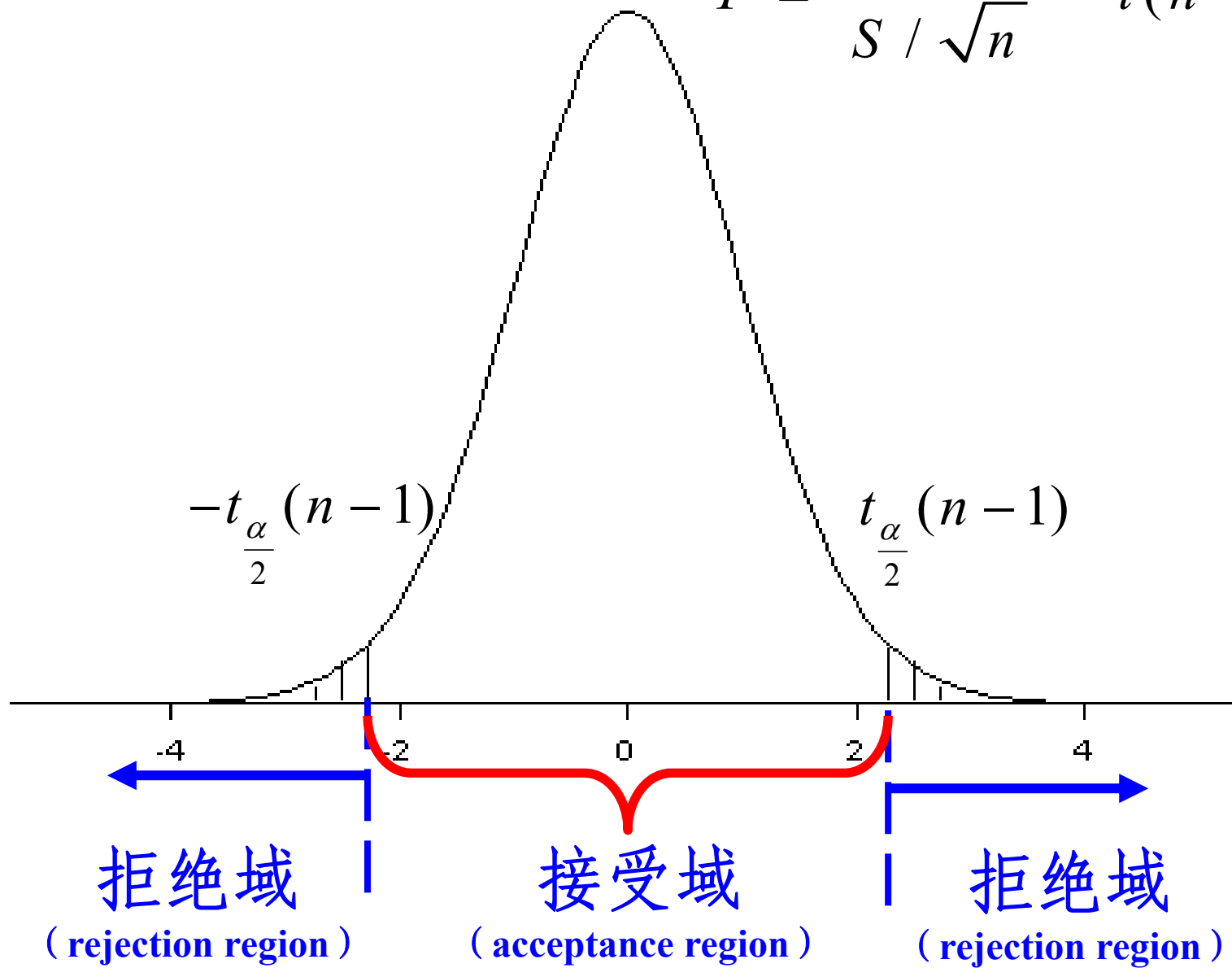
$$T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}} \sim t(n - 1)$$

显著性水平 α 下的拒绝域形式和接受域形式：

$$\left| \frac{\bar{X} - \mu_0}{S / \sqrt{n}} \right| \geq t_{\frac{\alpha}{2}}(n - 1) \quad \left| \frac{\bar{X} - \mu_0}{S / \sqrt{n}} \right| < t_{\frac{\alpha}{2}}(n - 1)$$

$t(25)$

$$T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}} \sim t(n - 1)$$



二、均值 μ 的单边检验

1. 已知方差 σ^2 ，检验 $H_0: \mu = \mu_0$ ， $H_1: \mu < \mu_0$

设总体 $X \sim N(\mu, \sigma^2)$ ， σ^2 已知， $X_1, X_2, \dots, X_n$ 是来自总体 X 的样本。

采用U检验，检验统计量为：

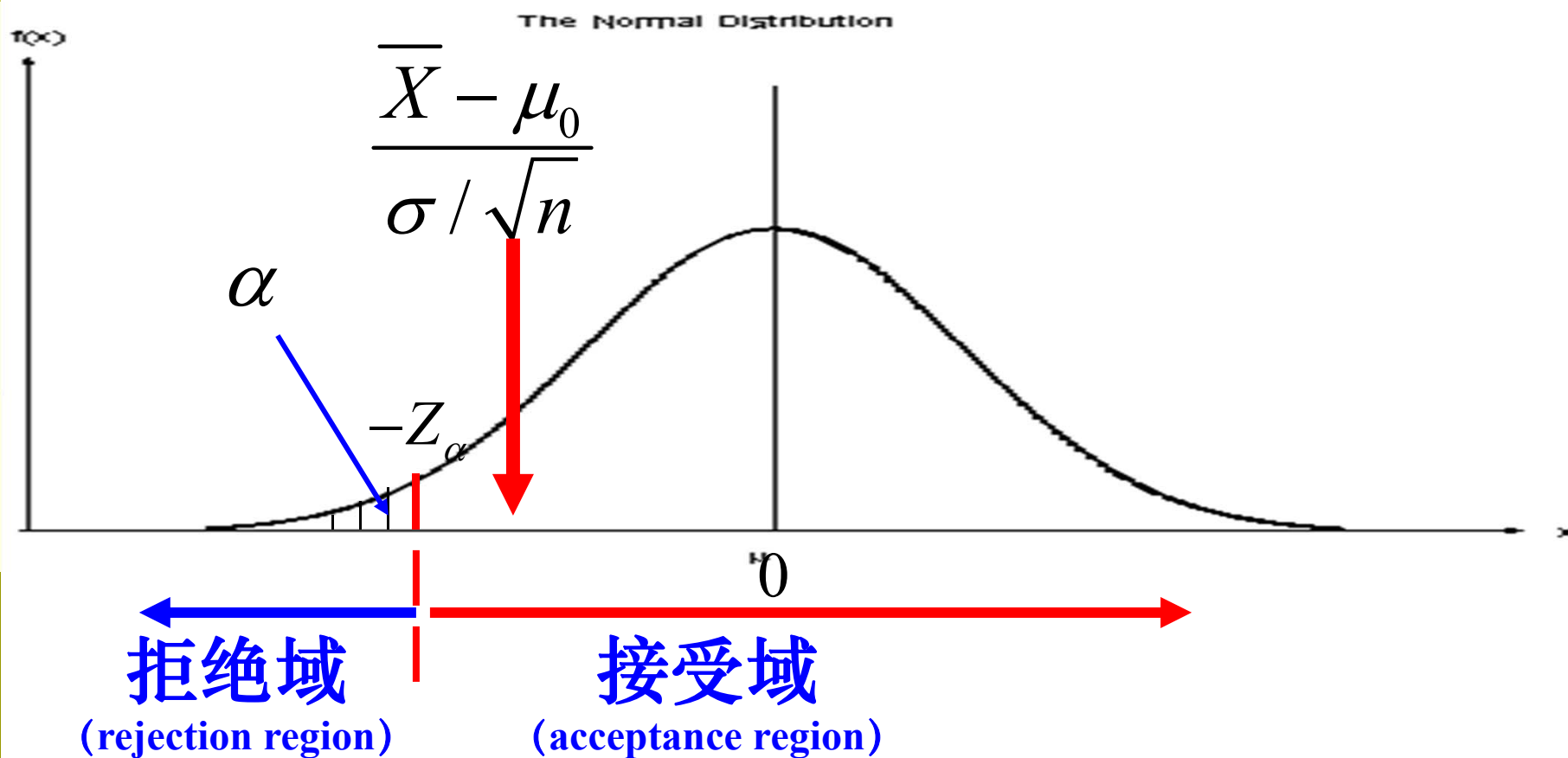
$$U = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0, 1)$$

若拒绝原假设 $H_0: \mu = \mu_0$ ，意味着接受对立假设 $H_1: \mu < \mu_0$ ，只有当样本均值的观察值比 μ_0 小很多时，才有理由拒绝原假设。

给定显著性水平 α 的拒绝域和接受域形式：

$$\frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} < -Z_\alpha$$

$$\frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \geq -Z_\alpha$$



2. 未知方差 σ^2 ，检验 $H_0: \mu = \mu_0$, $H_1: \mu < \mu_0$

设总体 $X \sim N(\mu, \sigma^2)$ ， σ^2 未知， $X_1, X_2, \dots, X_n$ 是来自总体 X 的样本。

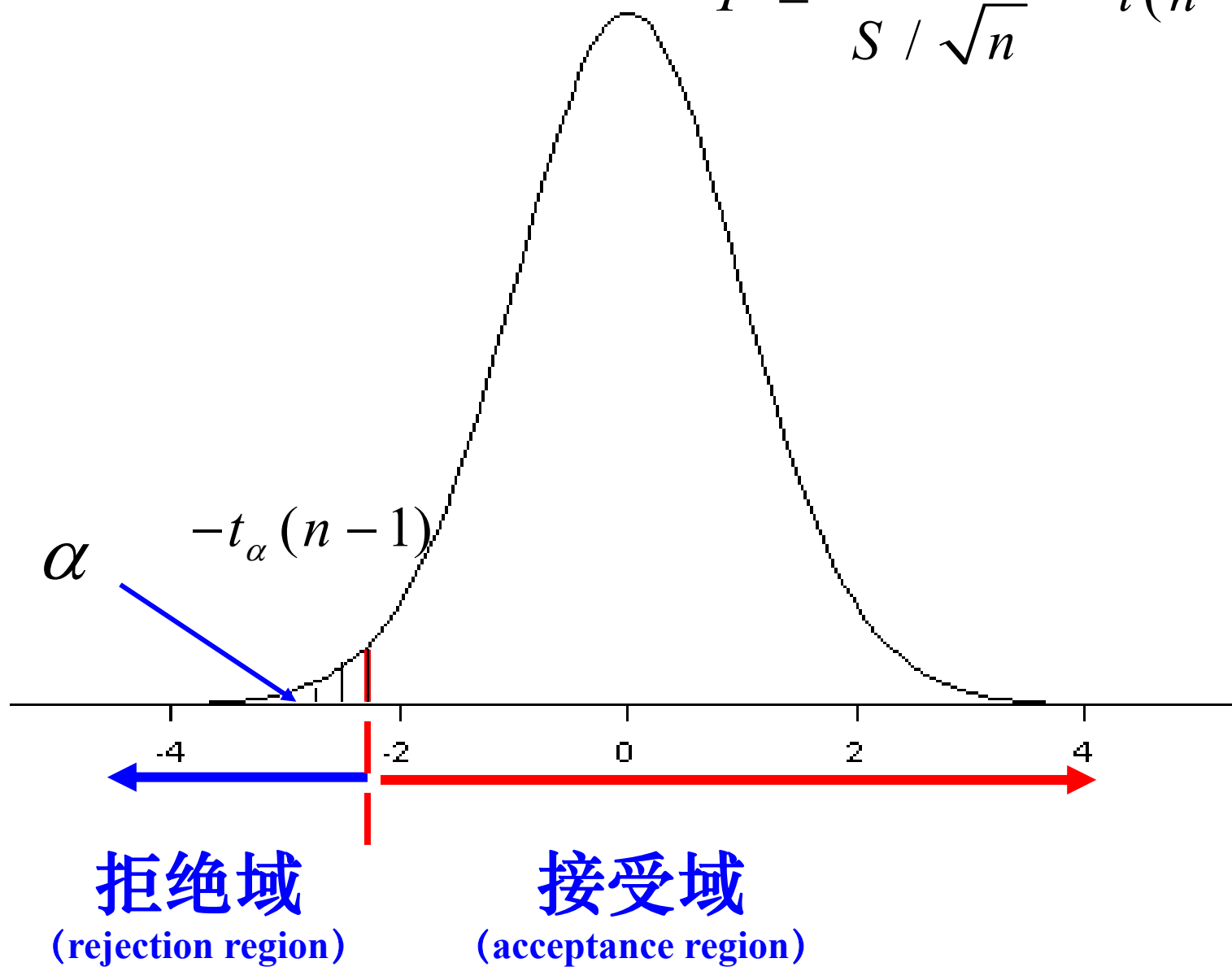
采用T检验，给定显著性水平 α 的拒绝域和接受域形式：

$$\frac{\bar{X} - \mu_0}{S / \sqrt{n}} < -t_{\alpha}(n - 1)$$

$$\frac{\bar{X} - \mu_0}{S / \sqrt{n}} \geq -t_{\alpha}(n - 1)$$

$t(25)$

$$T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}} \sim t(n - 1)$$



3. 未知方差 σ^2 ，检验 $H_0: \mu \leq \mu_0$, $H_1: \mu > \mu_0$

设总体 $X \sim N(\mu, \sigma^2)$ ， σ^2 未知， $X_1, X_2, \dots, X_n$ 是来自总体 X 的样本。

采用T检验，将原假设的复合假设分解为两种更简单的情况：

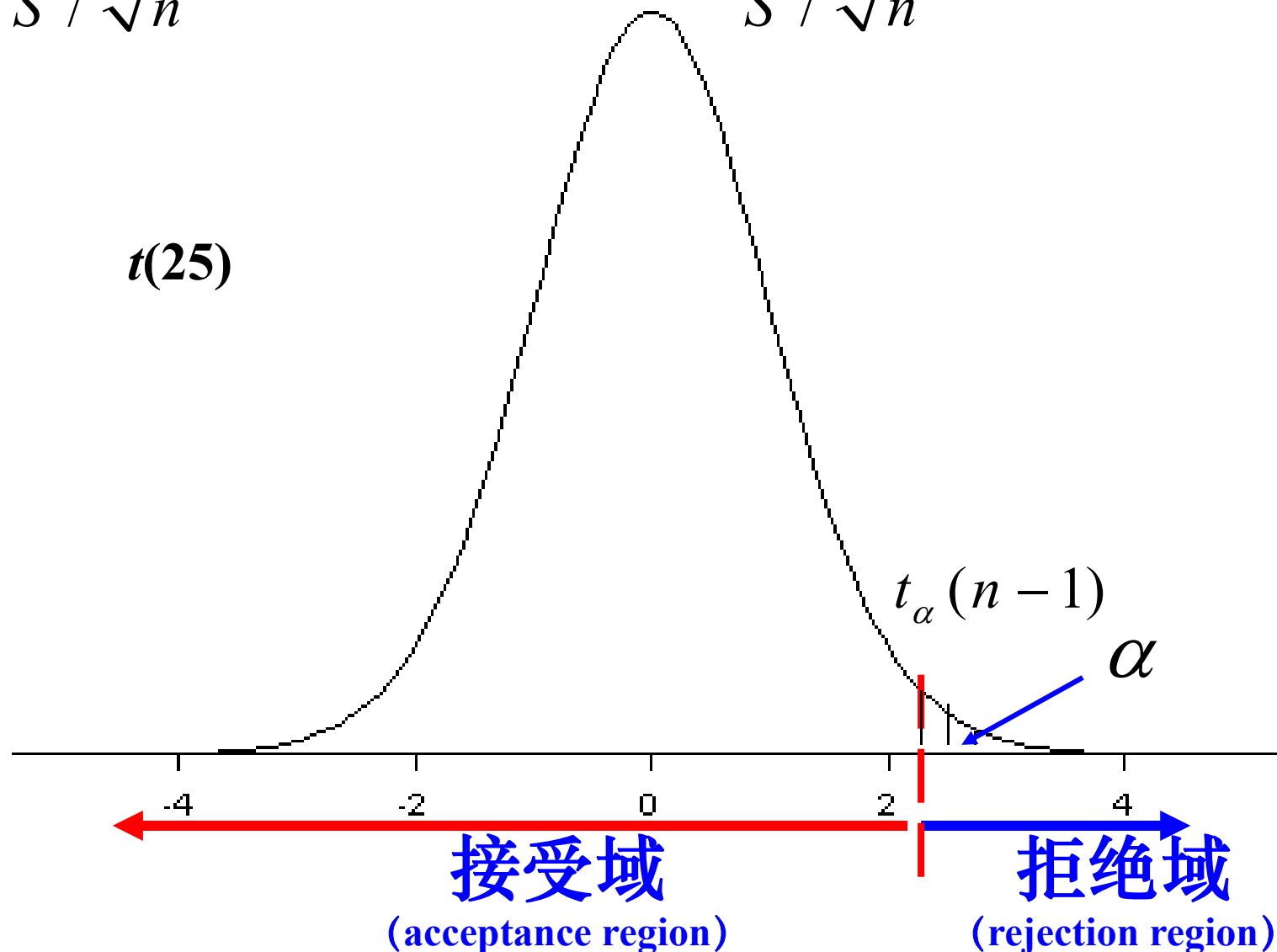
(1) 若 $\mu = \mu_0$ 成立

(2) 若 $\mu < \mu_0$ 成立

给定显著性水平 α 的拒绝域和接受域形式:

$$\frac{\bar{X} - \mu_0}{S / \sqrt{n}} > t_{\alpha}(n-1)$$

$$\frac{\bar{X} - \mu_0}{S / \sqrt{n}} \leq t_{\alpha}(n-1)$$



讨论：

- (1) 假设检验的关键：
 - 对立假设是否在小概率条件下成立
 - 对立假设是否显著地满足
- (2) 显著地满足对立假设的区域 $\Leftrightarrow$ 拒绝域
显著的程度：显著性水平 α
- (3) 对立假设 $\Leftrightarrow$ 拒绝域 $\Leftrightarrow$ 接受域 $\Leftrightarrow$ 原假设

§ 6.2.2 两正态总体均值差的检验

设总体 $X \sim N(\mu_1, \sigma_1^2)$ 和 $Y \sim N(\mu_2, \sigma_2^2)$ 相互独立, $X_1, X_2, \dots, X_{n_1}$ 是来自总体 X 的样本, $Y_1, Y_2, \dots, Y_{n_2}$ 是来自总体 Y 的样本。

1. 已知方差 σ_1^2 和 σ_2^2 , 检验

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

$$E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$$

$$(\bar{X} - \bar{Y}) \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$



$$U = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

采用U检验，选取 U 为检验统计量，即：

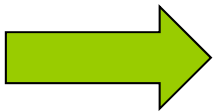
$$U = \frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

给定显著性水平 α 的拒绝域和接受域形式为：

$$\frac{|\bar{X} - \bar{Y}|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > Z_{\frac{\alpha}{2}} \quad \frac{|\bar{X} - \bar{Y}|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq Z_{\frac{\alpha}{2}}$$

2. 未知方差 σ_1^2 和 σ_2^2 ，但已知 $\sigma_1^2 = \sigma_2^2$ ，
检验 $H_0: \mu_1 = \mu_2$ ， $H_1: \mu_1 \neq \mu_2$

$$E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$$


$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

$$S_w^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

其中：

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2,$$

$$S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2$$

采用***T*检验**，选取*T*为检验统计量，即：

$$T = \frac{(\bar{X} - \bar{Y})}{S_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

给定显著性水平 α 的拒绝域和接受域形式为：

$$\frac{|\bar{X} - \bar{Y}|}{S_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{\frac{\alpha}{2}}(n_1 + n_2 - 2)$$

$$\frac{|\bar{X} - \bar{Y}|}{S_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{\frac{\alpha}{2}}(n_1 + n_2 - 2)$$

3. 已知方差 σ_1^2 和 σ_2^2 ，检验 $H_0: \mu_1 = \mu_2$ ， $H_1: \mu_1 > \mu_2$

采用**U检验**，选取U为检验统计量，即：

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

给定显著性水平 α 的拒绝域和接受域形式为：

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > Z_\alpha \qquad \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq Z_\alpha$$

§ 6.2.3 单个正态总体方差 σ^2 的检验

一、方差 σ^2 的双边检验

设总体 $X \sim N(\mu, \sigma^2)$, $X_1, X_2, \dots, X_n$ 是来自总体 X 的样本。

1. 未知均值 μ , 检验 $H_0: \sigma^2 = \sigma_0^2$, $H_1: \sigma^2 \neq \sigma_0^2$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$\frac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1)$$

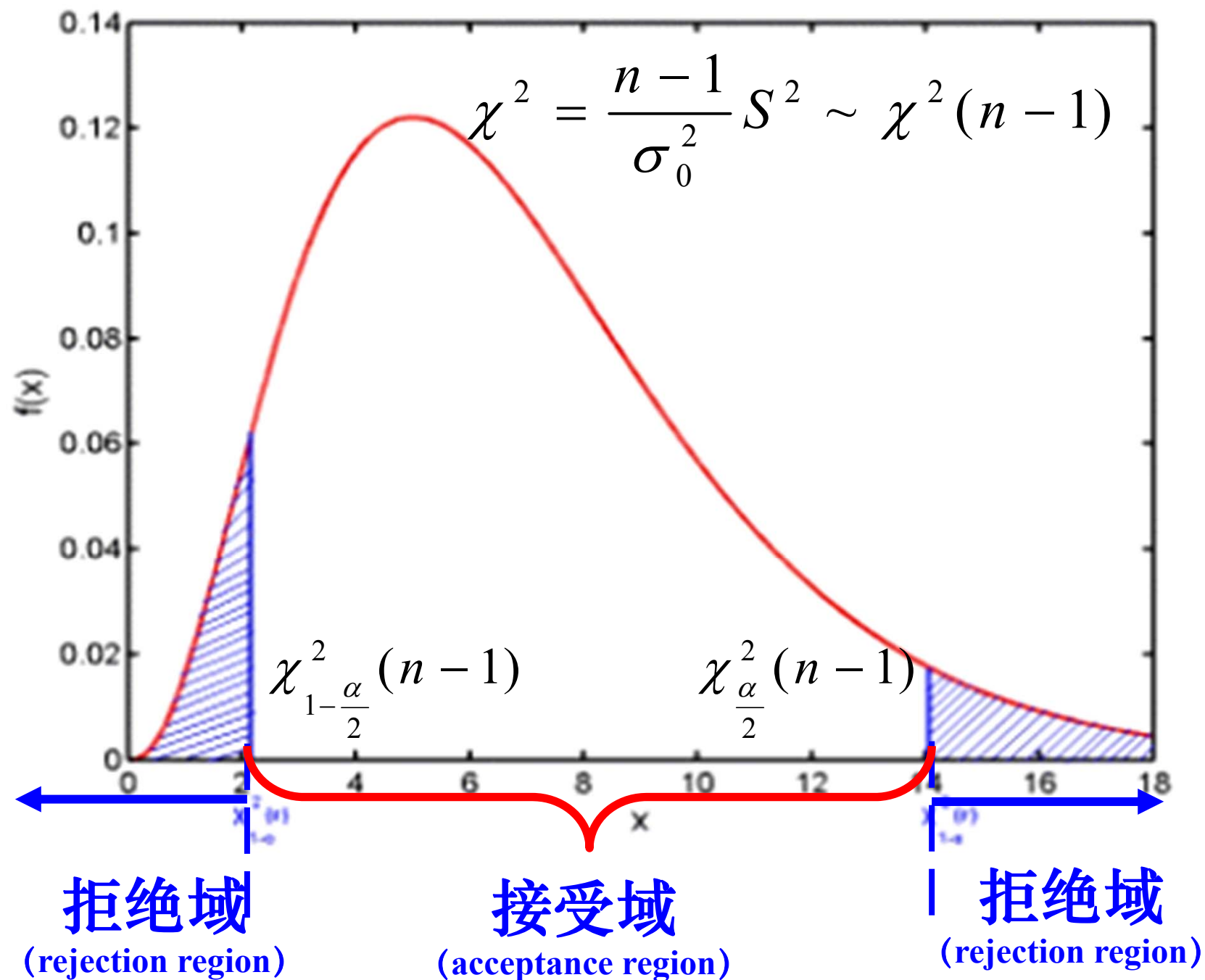
选取检验统计量（称为 χ^2 检验）：

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

给定显著性水平 α 的拒绝域和接受域形式为：

$$\left(\frac{n-1}{\sigma_0^2} S^2 < \chi_{1-\frac{\alpha}{2}}^2(n-1) \right) \cup \left(\frac{n-1}{\sigma_0^2} S^2 > \chi_{\frac{\alpha}{2}}^2(n-1) \right)$$

$$\chi_{1-\frac{\alpha}{2}}^2(n-1) \leq \frac{n-1}{\sigma_0^2} S^2 \leq \chi_{\frac{\alpha}{2}}^2(n-1)$$



2. 已知均值 μ ，检验 $H_0: \sigma^2 = \sigma_0^2$ ， $H_1: \sigma^2 \neq \sigma_0^2$

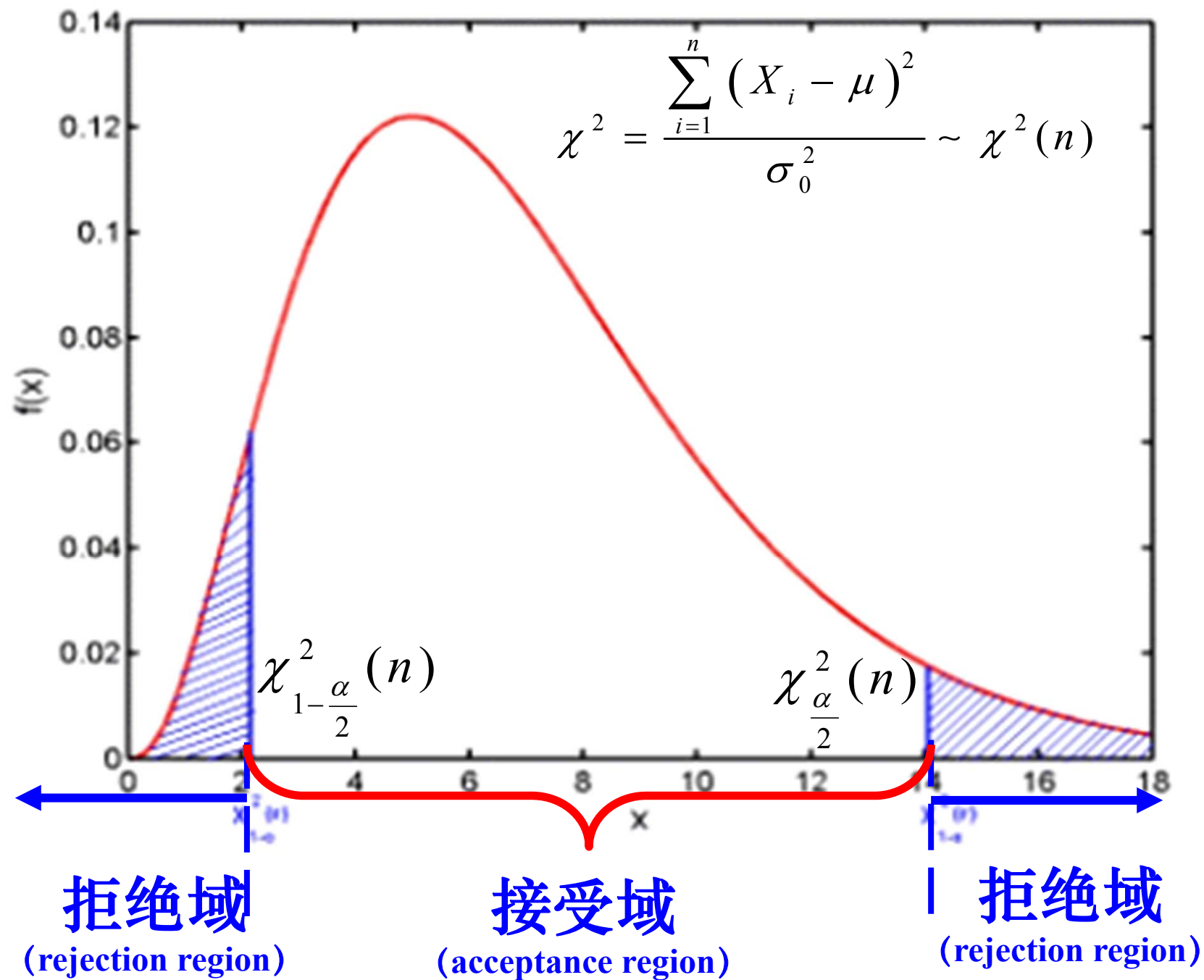
采用 χ^2 检验，选取检验统计量为：

$$\chi^2 = \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma_0^2} \sim \chi^2(n)$$

给定显著性水平 α 的拒绝域和接受域形式为：

$$\left(\frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma_0^2} < \chi_{1-\frac{\alpha}{2}}^2(n) \right) \cup \left(\frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma_0^2} > \chi_{\frac{\alpha}{2}}^2(n) \right)$$

$$\chi_{1-\frac{\alpha}{2}}^2(n) \leq \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma_0^2} \leq \chi_{\frac{\alpha}{2}}^2(n)$$



二、方差 σ^2 的的单边检验

设总体 $X \sim N(\mu, \sigma^2)$, $X_1, X_2, \dots, X_n$ 是来自总体 X 的样本。

1. 未知均值 μ , 检验 $H_0: \sigma^2 \leq \sigma_0^2$, $H_1: \sigma^2 > \sigma_0^2$

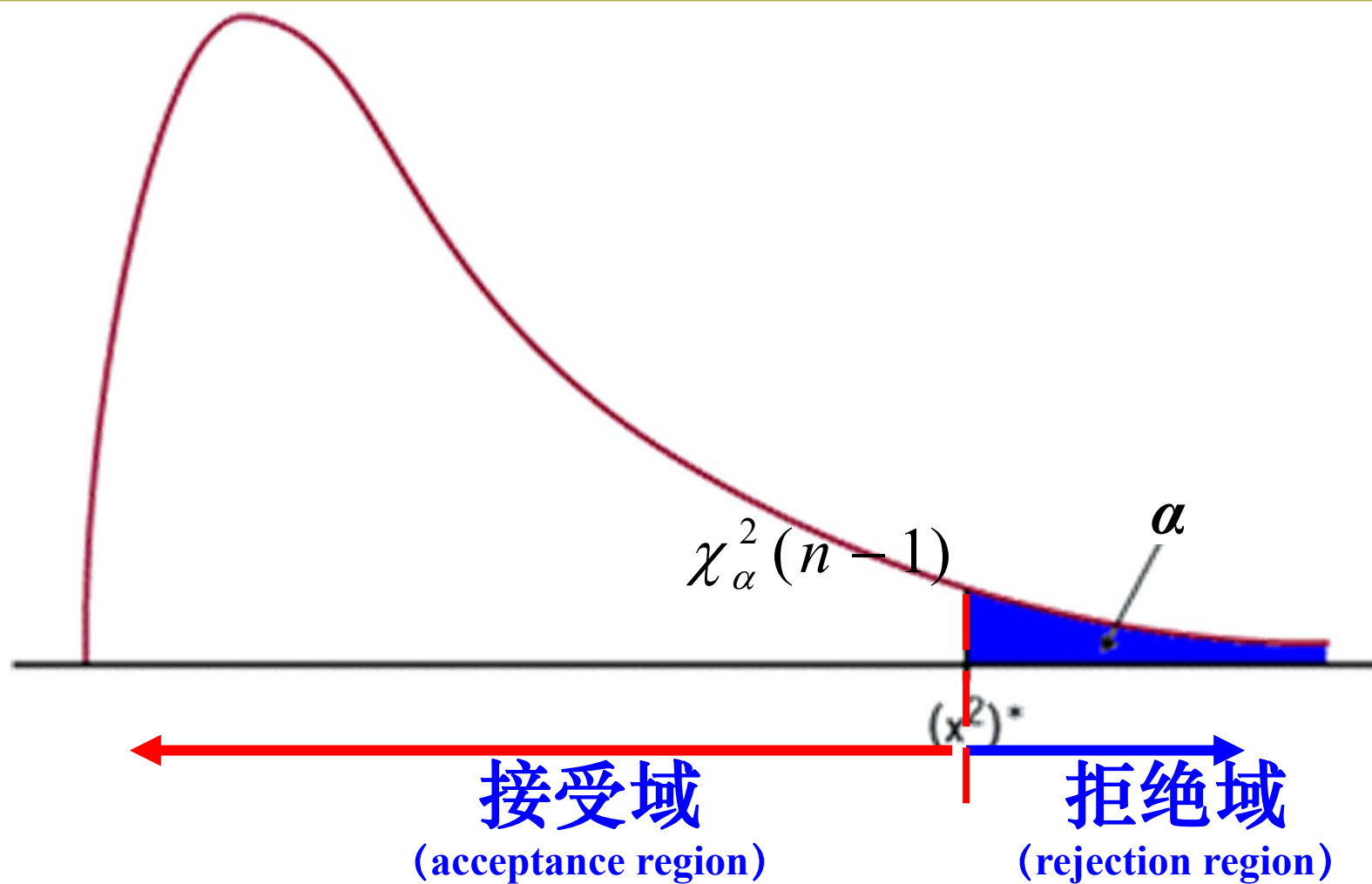
将原假设分解为更为简单的两种情况。采用 χ^2 检验, 选取检验统计量为:

$$\frac{n-1}{\sigma_0^2} S^2 \sim \chi^2(n-1)$$

给定显著性水平 α 的拒绝域和接受域形式为:

$$\frac{n-1}{\sigma_0^2} S^2 > \chi_\alpha^2(n-1) \qquad \frac{n-1}{\sigma_0^2} S^2 \leq \chi_\alpha^2(n-1)$$

$$\frac{n-1}{\sigma_0^2} S^2 \sim \chi^2(n-1)$$



2. 未知均值 μ ，检验 $H_0: \sigma^2 \geq \sigma_0^2$ ， $H_1: \sigma^2 < \sigma_0^2$

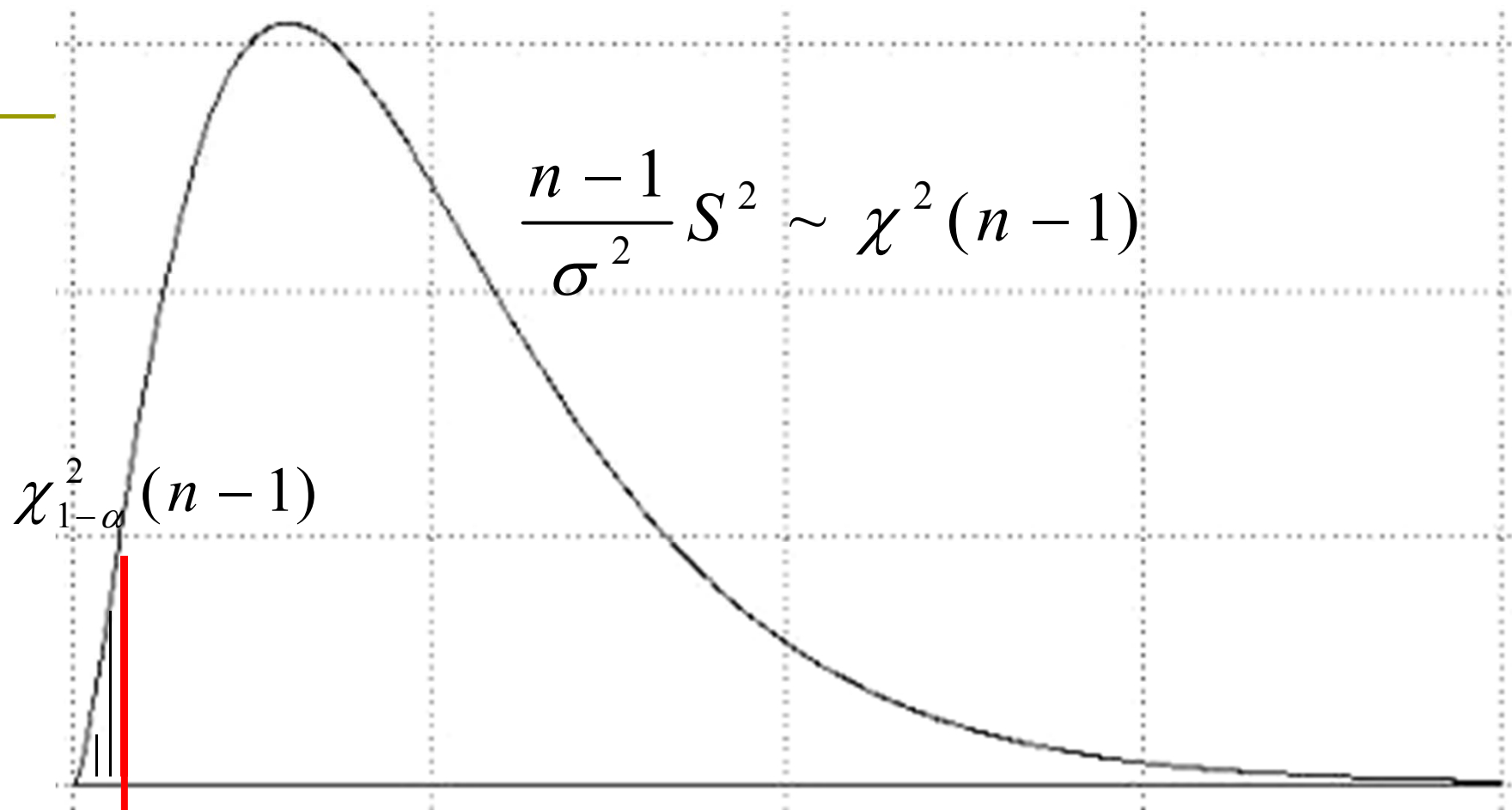
采用 χ^2 检验，选取检验统计量为：

$$\frac{n-1}{\sigma_0^2} S^2 \sim \chi^2(n-1)$$

给定显著性水平 α 的拒绝域和接受域形式为：

$$\frac{n-1}{\sigma_0^2} S^2 < \chi_{1-\alpha}^2(n-1)$$

$$\frac{n-1}{\sigma_0^2} S^2 \geq \chi_{1-\alpha}^2(n-1)$$



拒绝域

接受域
(acceptance region)

§ 6.2.4 两正态总体方差比较的检验

设总体 $X \sim N(\mu_1, \sigma_1^2)$ 和 $Y \sim N(\mu_2, \sigma_2^2)$ 相互独立, $X_1, X_2, \dots, X_{n_1}$ 是来自总体 X 的样本, $Y_1, Y_2, \dots, Y_{n_2}$ 是来自总体 Y 的样本。

1. 未知均值 μ_1 、 μ_2 , 检验

$$H_0: \sigma_1^2 = \sigma_2^2, \quad H_1: \sigma_1^2 \neq \sigma_2^2$$

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

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2, \quad S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2$$

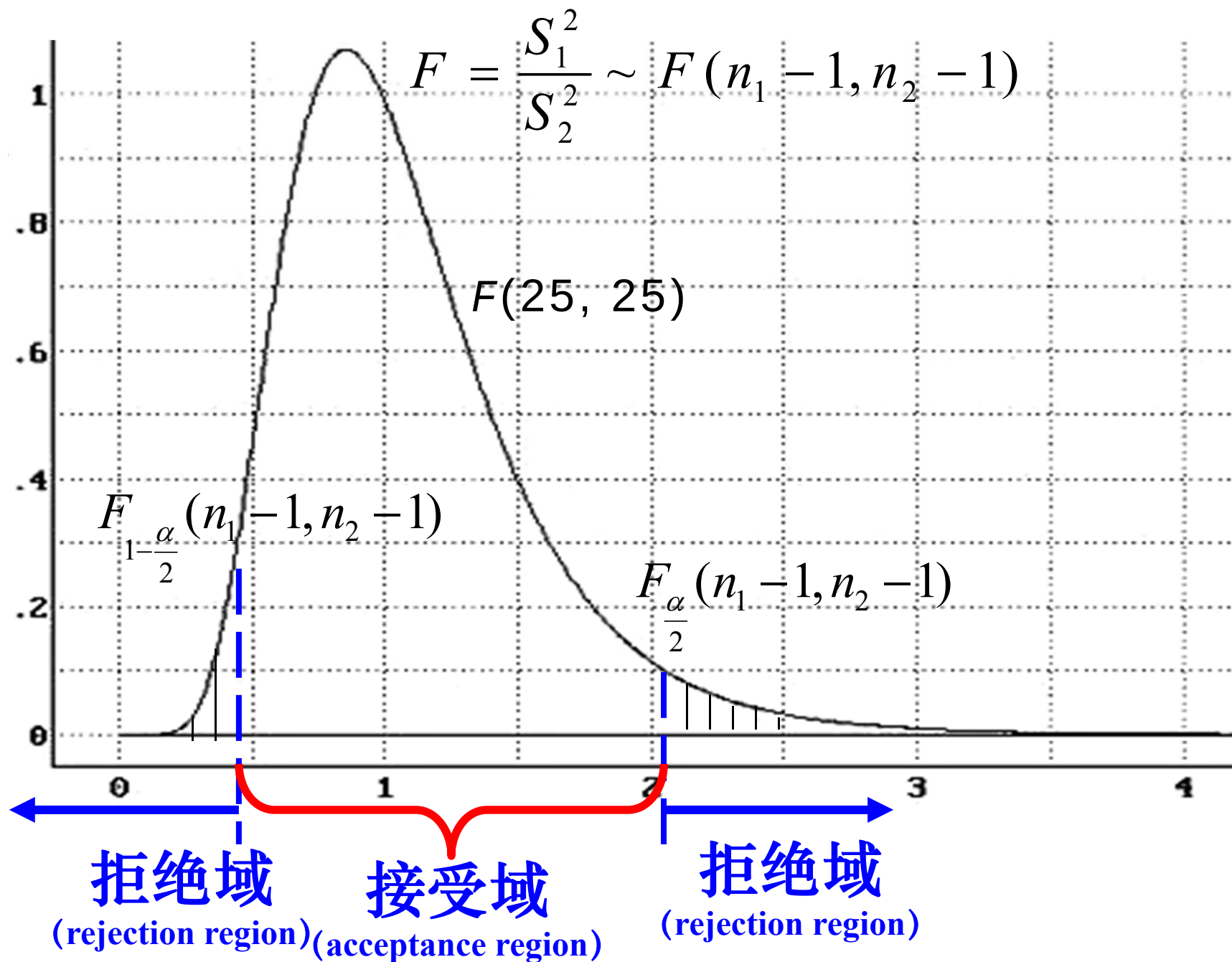
采用**F检验**，选取检验统计量为：

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

给定显著性水平 α 的拒绝域和接受域形式为：

$$\left\{ \frac{S_1^2}{S_2^2} < F_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1) \right\} \cup \left\{ \frac{S_1^2}{S_2^2} > F_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1) \right\}$$

$$F_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1) \leq \frac{S_1^2}{S_2^2} \leq F_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)$$



2 未知均值 μ_1 、 μ_2 ，检验

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

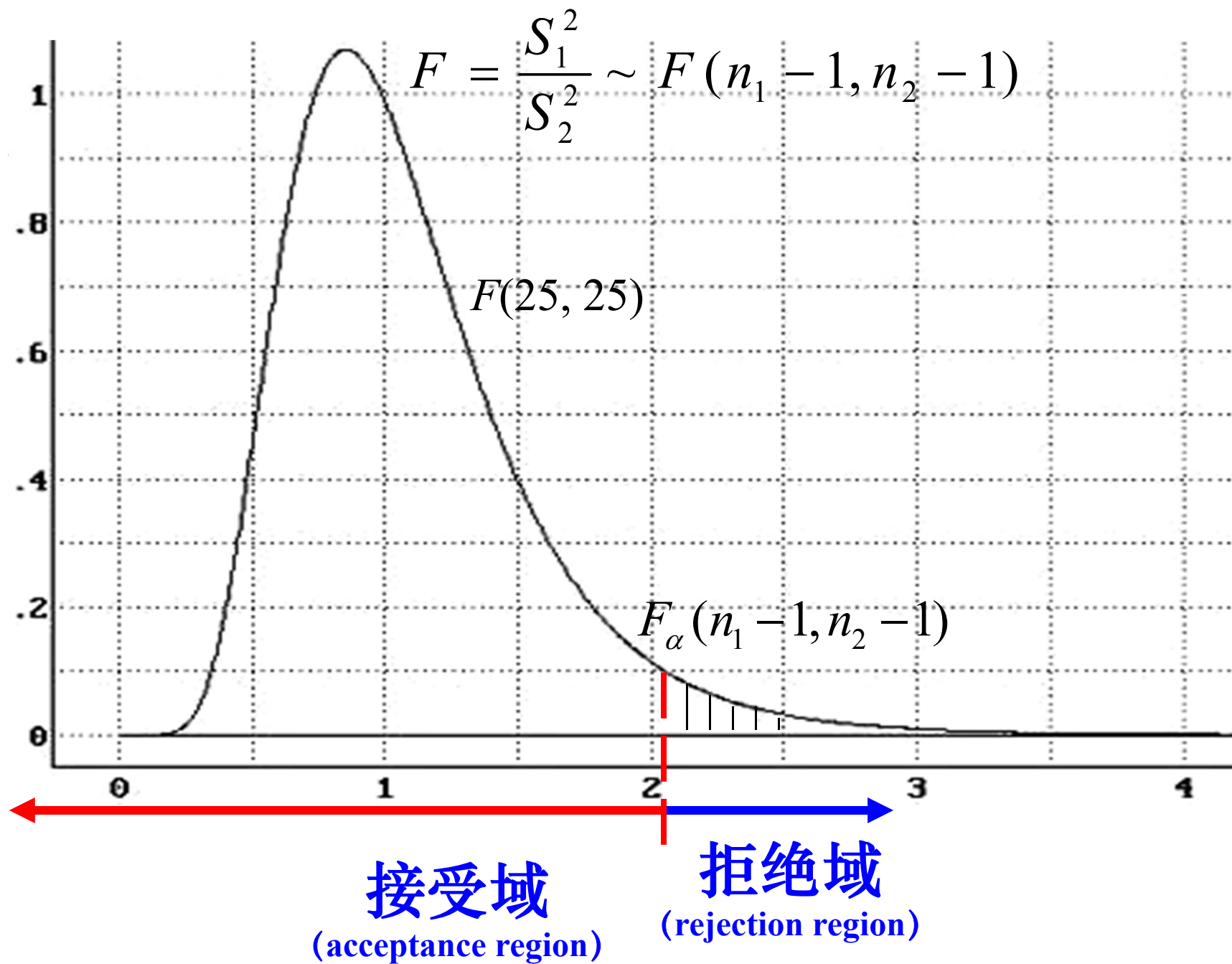
采用F检验，选取检验统计量为：

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

给定显著性水平 α 的拒绝域和接受域形式为：

$$\frac{S_1^2}{S_2^2} > F_\alpha(n_1 - 1, n_2 - 1)$$

$$\frac{S_1^2}{S_2^2} \leq F_\alpha(n_1 - 1, n_2 - 1)$$



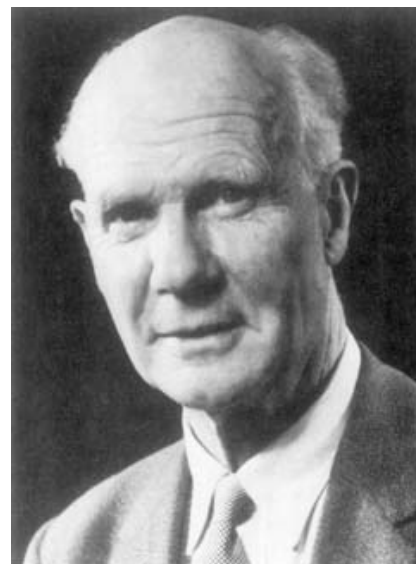
§ 6.3 似然比假设检验

Neyman-Pearson (1928)

- 更一般条件下的、统一的似然比 (Likelihood ratio) 检验方法
- 检验统计量在一般的条件下有统一的渐近分布，并在控制显著性水平优先的条件下考虑检验功效的水平
- 包括非正态的总体分布



Jerzy Neyman
(1894-1981)
美国波兰裔统计学家



Egon Pearson
(1895-1980)
英国统计学家

§ 6.3.1 N-P引理：似然比检验方法

Neyman-Pearson引理 (Neyman-Pearson Lemma)

设原假设 H_0 和对立假设 H_1 均为简单假设，它们分别给出随机变量总体 X 的概率分布密度函数（或分布律） $f_0(x|\theta_0)$ 和 $f_1(x|\theta_1)$ 。对于来自总体 X 的（简单随机抽样）样本 $X_1, X_2, \dots, X_n$ ，定义**似然比**（ LR , Likelihood Ratio）为检验统计量：

$$LR = \frac{lik(f_0)}{lik(f_1)} = \frac{\prod_{i=1}^n f_0(X_i | \theta_0)}{\prod_{i=1}^n f_1(X_i | \theta_1)}$$





(1) 似然比检验 (LR test)：在给定显著性水平 α 条件下，当 $LR < c$ 满足（ c 为临界值），则可拒绝原假设 H_0 。

(2) 对于上述给定显著性水平 α 的似然比检验，任给显著性水平 $\alpha^* \leq \alpha$ 的假设检验，均有：

$$(1 - \beta^*) \leq (1 - \beta)$$

其中， β^* 和 β 均为第二类错误的概率。

N-P引理 (LR检验) 讨论:

- (1) 基于样本 $X_1, X_2, \dots, X_n$ 的统计观察（实践），当似然比 LR 取值越大，表明随机变量的分布越趋近原假设 H_0 ，反之，越小越趋近对立假设 H_1 。
- (2) 在给定显著性水平 α 的条件下，似然比检验产生的第二类错误概率 β 最小，即**似然比检验方法是最大功效的检验方法**。
- (3) 适用条件：原假设 H_0 和对立假设 H_1 均为简单假设。

似然比 (LR) 检验问题

概率密度 f 的参数空间 Θ

θ_0

统计观察 (实践) :
 $X_1, X_2, \dots, X_n$

θ_1

H_0 : 定义 θ_0
(θ_0 为参数空间 Θ 的一点)

H_1 : 定义 θ_1
(θ_1 为参数空间 Θ 的另一点)

【Example 6.4】 设总体 $X \sim N(\mu, \sigma^2)$, σ^2 已知, 而 μ 只能取两个值 μ_0 、 μ_1 , 且 $\mu_0 < \mu_1$, $X_1, X_2, \dots, X_n$ 为来自总体 X 的容量为 n 的简单随机抽样样本, 试在显著性水平 α 下运用似然比检验假设: $H_0: \mu = \mu_0$; $H_1: \mu = \mu_1$ 。

基本步骤:

——写出 LR 的表达式

——根据 LR 取值趋小的形式找到拒绝原假设的统计量形式

——根据抽样分布求出临界点的值

§ 6.3.2 广义似然比检验方法

——**广义似然比检验** (GLR test, Generalized Likelihood Ratio test)

已知 $X_1, X_2, \dots, X_n$ 是来自总体 X 的（简单随机抽样）样本，根据总体分布函数可得到它们的联合概率密度形式为：

$$\prod_{i=1}^n f(X_i | \theta), \quad \theta \in \Theta$$

Θ 为参数空间，其维数可能大于1。

设原假设 H_0 定义参数范围为： $\theta \in \omega_0 \subset \Theta$ ，对立假设 H_1 定义参数范围为： $\theta \in \omega_1 \subset \Theta$ ，且 $\omega_0 \cap \omega_1 = \emptyset$ ，记 $\Theta' = \omega_0 \cup \omega_1$ ，显然有 $\Theta' \subseteq \Theta$ 。

定义**广义似然比** (GLR , Generalized Likelihood Ratio) 为：

$$GLR^* = \frac{\max_{\theta \in \omega_0} [lik(\theta)]}{\max_{\theta \in \omega_1} [lik(\theta)]}, \quad lik(\theta) = \prod_{i=1}^n f(X_i | \theta)$$

以GLR*为检验统计量，显然有GLR*越小，原假设越有可能被拒绝。

通常更多地采用下列广义似然比的定义：

$$GLR = \frac{\max_{\theta \in \omega_0} [lik(\theta)]}{\max_{\theta \in \Theta'} [lik(\theta)]}$$

讨论：

(1) 当原假设 H_0 为简单假设时，有：
$$GLR = \frac{lik(\theta_0)}{\max_{\theta \in \Theta'}[lik(\theta)]}$$

(2) $GLR = \min(GLR^*, 1)$

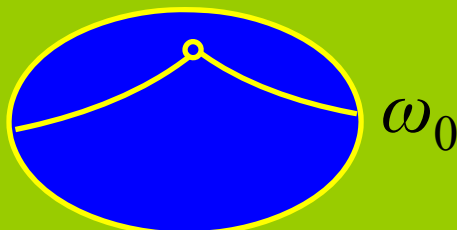
(3) 与似然比检验相比，广义似然比检验并不能保证是最优检验，尤其是当检验的问题本身并不存在最优解时。但具有统一的检验统计量和更一般性的方法。

广义似然比 (GLR) 检验问题

GLR检验是LR检验的推广 (由点推广到空间)

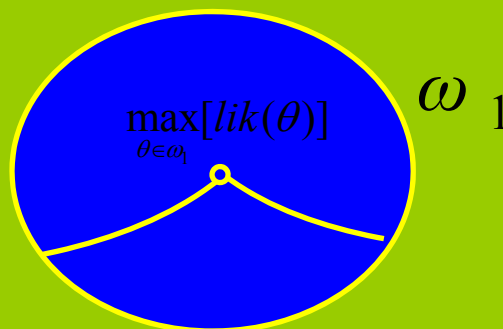
概率密度 f 的参数空间 Θ

$$\max_{\theta \in \omega_0} [\text{lik}(\theta)]$$



统计观察 (实践) :

$$X_1, X_2, \dots, X_n$$



H_0 : 定义 ω_0
(ω_0 为参数空间 Θ 的一个子空间)

H_1 : 定义 ω_1
(ω_1 为参数空间 Θ 的另一个子空间, 包括 Θ 本身)

【Example 6.5】 设总体 $X \sim N(\mu, \sigma^2)$, σ^2 已知, $X_1, X_2, \dots, X_n$ 为来自总体 X 的容量为 n 的简单随机抽样样本, 试在显著性水平 α 下运用广义似然比检验假设: $H_0: \mu = \mu_0$; $H_1: \mu \neq \mu_0$ 。

$$GLR = e^{-\frac{1}{2\sigma^2} \left[\sum_{i=1}^n (X_i - \mu_0)^2 - \sum_{i=1}^n (X_i - \bar{X})^2 \right]}$$

$$GLR' = -2 \log GLR = \frac{n(\bar{X} - \mu_0)^2}{\sigma^2} \sim \chi^2(1)$$

解：这里， $\omega_0 = \{\mu_0\}$ 是确定的取值， $\omega_1 = \{\mu \mid \mu \neq \mu_0\}$ ， $\Theta' = \{\mu \mid -\infty < \mu < +\infty\}$ 。

原假设 H_0 为简单假设（点假设），不需考虑求解其最大似然值的点，故 GLR 的分子只需计算 $\mu=\mu_0$ 时的似然值：

$$\max_{\theta \in \omega_0} [lik(\theta_0)] = lik(\theta_0) = \frac{1}{(\sigma \sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu_0)^2\right)。$$

根据对立假设 H_1 的内容，GLR 的分母是在 $\Theta' = \omega_0 \cup \omega_1$ 的空间上寻求最大似然值的解。根据上一讲《参数估计》

的内容，可知最大似然估计 $\mu = \bar{X}$ 就是其最大似然解，因此有：

$$\max_{\theta \in \Theta'} [lik(\theta)] = \frac{1}{(\sigma \sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2\right)。$$

于是检验统计量 GLR 为：

$$\begin{aligned} GLR &= \frac{\max_{\theta \in \omega_0} [lik(\theta)]}{\max_{\theta \in \Theta'} [lik(\theta)]} = \frac{lik(\theta_0)}{\max_{\theta \in \Theta'} [lik(\theta)]} \\ &= \frac{1}{(\sigma \sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu_0)^2\right) \bigg/ \frac{1}{(\sigma \sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2\right)， \\ &= \exp\left(-\frac{1}{2\sigma^2} \left[\sum_{i=1}^n (X_i - \mu_0)^2 - \sum_{i=1}^n (X_i - \bar{X})^2 \right]\right) \end{aligned}$$

采用对数形式有：↵

$$GLR' = -2 \log GLR = \frac{1}{\sigma^2} \left[\sum_{i=1}^n (X_i - \mu_0)^2 - \sum_{i=1}^n (X_i - \bar{X})^2 \right], \quad \leftarrow$$

注意到 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ ，并有：↵

$$\begin{aligned} & \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2 \\ &= \left(\sum_{i=1}^n (X_i^2) - 2\bar{X} \sum_{i=1}^n X_i + n\bar{X}^2 \right) + (n\bar{X}^2 - 2n\bar{X}\mu_0 + n\mu_0^2) \quad \leftarrow \\ &= \sum_{i=1}^n (X_i - \mu_0)^2 \end{aligned}$$

因此代入 $\sum_{i=1}^n (X_i - \mu_0)^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2$ ，↵

$$\text{最后得到: } GLR' = -2 \log GLR = \frac{n(\bar{X} - \mu_0)^2}{\sigma^2}. \quad \leftarrow$$

因此，本问题转换成检验统计量(-2logGLR)的问题。↵

同时, 因为 $\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$, 可推出: ↵

$$\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \sim N(0, 1), \quad \text{↵}$$

由随机变量函数可推出, 该统计量服从 $\chi^2(1)$ 分布。故最后有: ↵

$$GLR' = -2 \log GLR = \frac{n(\bar{X} - \mu_0)^2}{\sigma^2} \sim \chi^2(1) \quad \text{↵}$$

注意到上式的负号, 最后可写出显著性水平 α 下原假设 H_0 的拒绝域为: ↵

$$\frac{n(\bar{X} - \mu_0)^2}{\sigma^2} > \chi_{\alpha}^2(1) \circ \rightarrow \left| \bar{X} - \mu_0 \right| \geq \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}} \circ \quad \text{↵}$$

定理（一般总体的渐进分布性质）

在概率密度函数或分布律函数充分光滑的条件下，当样本容量 n 趋近无穷大时，推广到一般分布的总体 X (μ_0, σ^2 分别为 X 的均值和方差) 的检验统计量

$$GLR' = -2 \log GLR = \frac{n \left(\bar{X} - \mu_0 \right)^2}{\sigma^2}$$

的抽样分布趋近 $\chi^2(k)$ 分布，其自由度 k 为：

$$k = \dim(\Theta') - \dim(\omega_0)$$

似然比 (LR) 检验问题

概率密度 f 的参数空间 Θ

$$\frac{lik(\theta_0)}{lik(\theta_1)}$$

θ_0

H_0 : 定义 θ_0
(θ_0 为参数空间 Θ 的一点)

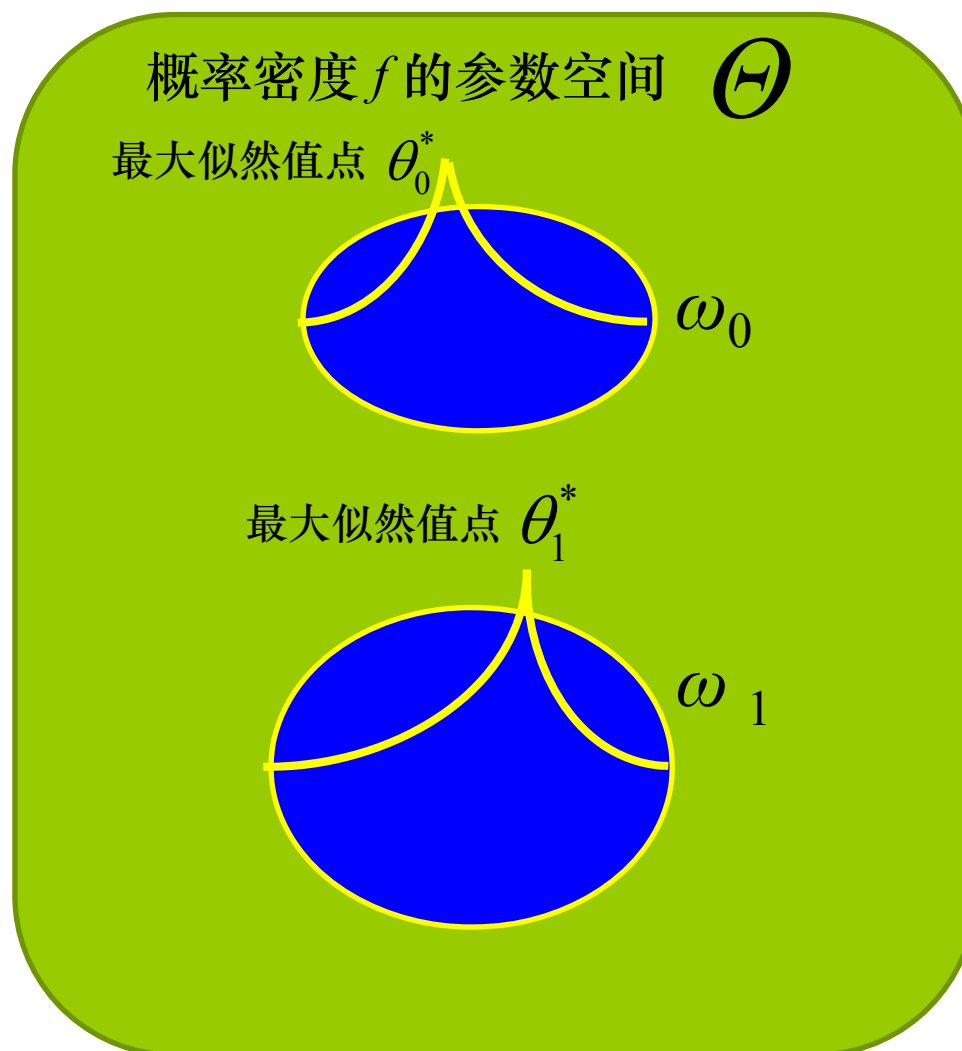
θ_1

H_1 : 定义 θ_1
(θ_1 为参数空间 Θ 的另一点)

广义似然比 (GLR) 检验问题

GLR检验是LR检验的推广（由点推广到空间，最终落实到点）

$$\frac{lik(\theta_0^*)}{lik(\theta_1^*)}$$



H_0 : 定义 ω_0
(ω_0 为参数空间 Θ 的一个子空间)

H_1 : 定义 ω_1
(ω_1 为参数空间 Θ 的另一个子空间，包括 Θ 本身)

§ 6.3.3 应用：多项分布的广义似然比检验

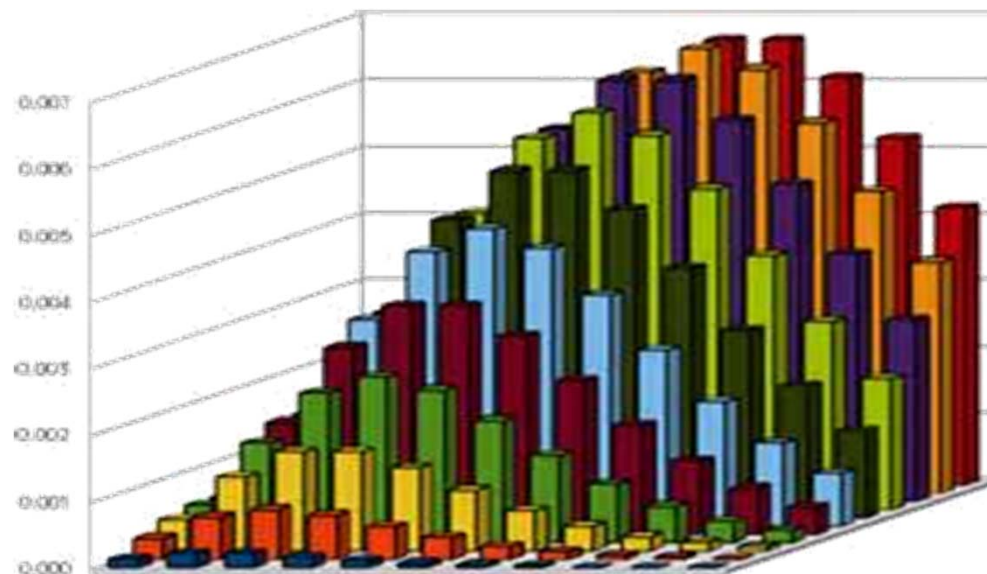
多项分布概率 (Multinomial cell probability)问题

设 $X_1, X_2, \dots, X_m$ 为分布在 m 个格子（事件）中、满足多项分布的采样计数：

$$X_1 + X_2 + \dots + X_m = n$$

令落入每个格子的概率分别为 $p_1, p_2, \dots, p_m$ ：

$$p_1 + p_2 + \dots + p_m = 1$$



【Example 5.1】Berkson对美国国家标准局的一组数据进行分析。测量数据的放射源为镅（americium）241，每10s记录一次测得的 α 粒子数 k ，共计1027次，平均每次的 α 粒子数为8.392。

k	0-2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17+
O bserved	18	28	56	105	126	146	164	161	123	101	74	53	23	15	9	5
E xpected	12.2	27.0	56.5	94.9	132.7	159.1	166.9	155.6	130.6	99.7	69.7	45.0	27.0	15.1	7.9	7.1

放射性衰变的 α 粒子数的Poisson分布

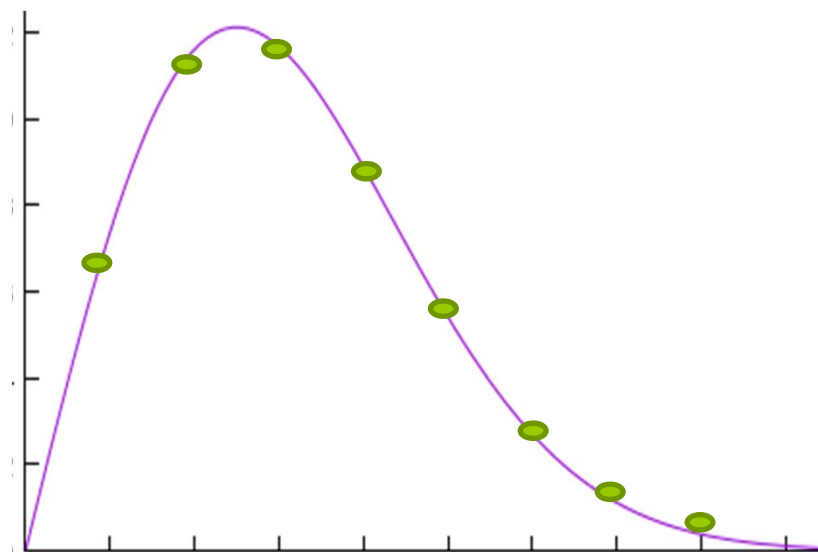
$$\pi_k = P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

$$\lambda > 0; \quad k = 0, 1, 2, \dots$$

根据上表观察的数据，可得到估计值：

$$\hat{\lambda} = 8.392$$

问题：
检验该Poisson分布是否成立



(Berkson J. (1966). Examination of randomness of alpha particle emission. In *Research Papers in Statistics*, New York: Wiley)

多项分布拟合检验问题：设 $X_1, X_2, \dots, X_m$ 为分布在 m 个格子中、满足多项分布的采样计数（即样本）且 $X_1 + X_2 + \dots + X_m = n$ 。令落入每个格子的概率分别为 $p_1, p_2, \dots, p_m$ ，且 $p_1 + p_2 + \dots + p_m = 1$ 。

原假设 H_0 ：给出多项分布概率密度矢量 $p = p(\theta) \in \omega_0$ ，其中 θ 为本问题的参数或待定未知参数。

对立假设 H_1 ：所有满足 $p_1 + p_2 + \dots + p_m = 1$ 的概率密度矢量 $p \in \Theta'$ 。

- ◆ 原假设 H_0 给定了 $p_1, p_2, \dots, p_m$ 的某一具体的分布（可能未知，由待定参数 θ 确定）。
- ◆ 对立假设 H_1 是定义在参数空间 Θ' 的所有可能的分布。

$$G L R = \frac{\max_{\theta \in \omega_0} [lik(\theta)]}{\max_{\theta \in \Theta'} [lik(\theta)]}$$

分子项（未知参数为 θ ）：

$$\max_{p \in \omega_0} \left\{ \frac{n!}{\prod_{i=1}^m X_i!} \prod_{i=1}^m p_i(\theta)^{X_i} \right\}$$

记上式的最大似然解为 $\hat{\theta}$ ，则分子项为：

$$\longrightarrow \frac{n!}{\prod_{i=1}^m X_i!} \prod_{i=1}^m p_i(\hat{\theta})^{X_i}$$

$$G L R = \frac{\max_{\theta \in \omega_0} [lik(\theta)]}{\max_{\theta \in \Theta'} [lik(\theta)]}$$

分母项（未知参数为 $p_1, p_2, \dots, p_m$ ）：

$$\max_{\theta \in \Theta'} [lik(\theta)] = \frac{n!}{\prod_{i=1}^m X_i!} \prod_{i=1}^m \hat{p}_i^{X_i},$$

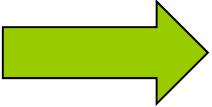
其中 $\hat{p}_i$ 为最大似然解。上一讲已经求得：

$$\hat{p}_i = \frac{X_i}{n}$$

因此，有：

$$X_i = n\hat{p}_i$$

$$G L R = \frac{\max_{\theta \in \omega_0} [lik(\theta)]}{\max_{\theta \in \Theta'} [lik(\theta)]}$$



$$\begin{aligned}
 G L R &= \frac{\max_{\theta \in \omega_0} [lik(\theta)]}{\max_{\theta \in \Theta'} [lik(\theta)]} \\
 &= \frac{\frac{n!}{\prod_{i=1}^m X_i!} \prod_{i=1}^m p_i(\hat{\theta})^{X_i}}{\frac{n!}{\prod_{i=1}^m X_i!} \prod_{i=1}^m \hat{p}_i^{X_i}} \\
 &= \prod_{i=1}^m \left(\frac{p_i(\hat{\theta})}{\hat{p}_i} \right)^{X_i}
 \end{aligned}$$

注意到：

$$X_i = n \hat{p}_i$$

$$GLR = \frac{\max_{\theta \in \omega_0} [lik(\theta)]}{\max_{\theta \in \Theta'} [lik(\theta)]}$$

记 $O_i = X_i = n\hat{p}_i$, $E_i = np_i(\hat{\theta})$, 有

$$-2 \log GLR = -2n \sum_{i=1}^m \hat{p}_i \log \left(\frac{p_i(\hat{\theta})}{\hat{p}_i} \right)$$

$$= 2 \sum_{i=1}^m n\hat{p}_i \log \left(\frac{n\hat{p}_i}{np_i(\hat{\theta})} \right)$$

$$= 2 \sum_{i=1}^m O_i \log \left(\frac{O_i}{E_i} \right)$$



$$GLR' = -2 \log GLR = 2 \sum_{i=1}^m O_i \log \left(\frac{O_i}{E_i} \right)$$

$$GLR = \frac{\max_{\theta \in \omega_0} [lik(\theta)]}{\max_{\theta \in \Theta'} [lik(\theta)]} \quad O_i = X_i = n\hat{p}_i, \quad E_i = np_i(\hat{\theta})$$

定理 (一般总体的渐进分布性质)

在概率密度函数或分布律函数充分光滑的条件下, 当样本容量 n 趋近无穷大时, 推广到一般分布的总体 X (μ_0, σ^2 分别为总体 X 的均值和方差) 的检验统计量

$$GLR' = -2 \log GLR = \frac{n(\bar{X} - \mu_0)^2}{\sigma^2}$$

的抽样分布趋近 $\chi^2(k)$ 分布, 其自由度 k 为:

$$k = \dim(\Theta') - \dim(\omega_0)$$

因此, GLR' 服从 χ^2 分布:

$$GLR' = -2 \log GLR = 2 \sum_{i=1}^m O_i \log \left(\frac{O_i}{E_i} \right) \sim \chi^2(m - k - 1)$$

$$\dim(\Theta') - \dim(\omega_0) = (m - 1) - k \quad (k \text{ 为参数 } \theta \text{ 的维度})$$

定理 (多项分布拟合检验的渐进分布性质)

对于同一多项分布的拟合假设检验问题，在原假设 H_0 成立的条件下，当样本容量 n 趋近无穷大时，检验统计量 GLR

$$GLR' = -2 \log GLR = -2n \sum_{i=1}^m \hat{p}_i \log \left(\frac{p_i(\hat{\theta})}{\hat{p}_i} \right) = 2 \sum_{i=1}^m O_i \log \left(\frac{O_i}{E_i} \right)$$

与Pearson χ^2 检验统计量

$$X^2 = \sum_{i=1}^m \frac{[X_i - np_i(\hat{\theta})]^2}{np_i(\hat{\theta})} = \sum_{i=1}^m \frac{[O_i - E_i]^2}{E_i}$$

近似等价，且二者的抽样分布都趋近自由度为 $\dim(\Theta) - \dim(\omega_0)$ 的 χ^2 分布。

这里， $O_i = X_i = n\hat{p}_i$ ， $E_i = np_i(\hat{\theta})$ 。

§ 6.4 非参数假设检验简介

参数假设检验问题 (【例6.1】)

——已知总体的分布函数类型，对分布函数中的未知参数提出某种假设

——要求利用样本的信息对所提出的假设进行检验，根据检验结果作出接受或者拒绝所提假设的判断

非参数假设检验问题 (【例6.2】)

——总体的分布类型未知

——要求根据样本的信息来对分布类型的假设进行检验，对总体的分布类型作出判断

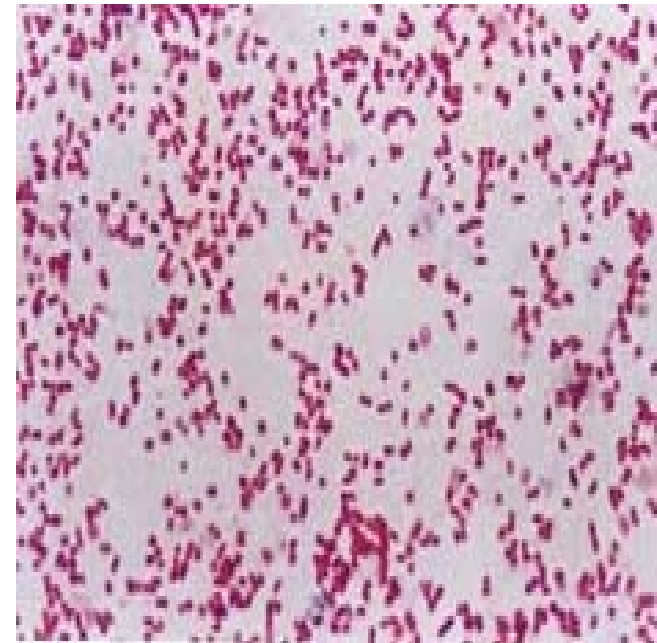
【Example 6.2】 将0.1ml受细菌污染的牛奶均匀涂在1cm²的切片上，用显微镜观察切片每个小网格内的细菌菌落数目。根据400（20×20）个小网格的计数结果，统计出如下表格：

菌落数：	0	1	2	3	4	5	6	7	8	9	10	19
频 数：	56	104	80	62	42	27	9	9	5	3	2	1

试问单位面积内的菌落数是否服从Poisson分布？

问题：

根据样本信息，判别随机变量 X 是否服从Poisson分布。



(Bliss C. and Fisher R. A., (1953). Fitting the negative binomial distribution to biological data. *Biometric*, 9: 174-200)

§ 6.4.1 K-S检验 (Kolmogorov–Smirnov test)

K-S检验:

——拟合优度检验 (Goodness-of-fit test)

——通过对经验分布、理论分布两种分布差异的分析确定是否有理由认为样本的观察结果来自所假定的理论分布总体，研究样本观察值的分布和设定的理论分布是否吻合

定义 (经验分布函数、理论分布函数)

设 $X_1, X_2, \dots, X_n$ 为来自总体 X 的样本， $x_1, x_2, \dots, x_n$ 为其观察值，将它们按照观察值的大小递增顺序排列生成顺序统计量

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)},$$

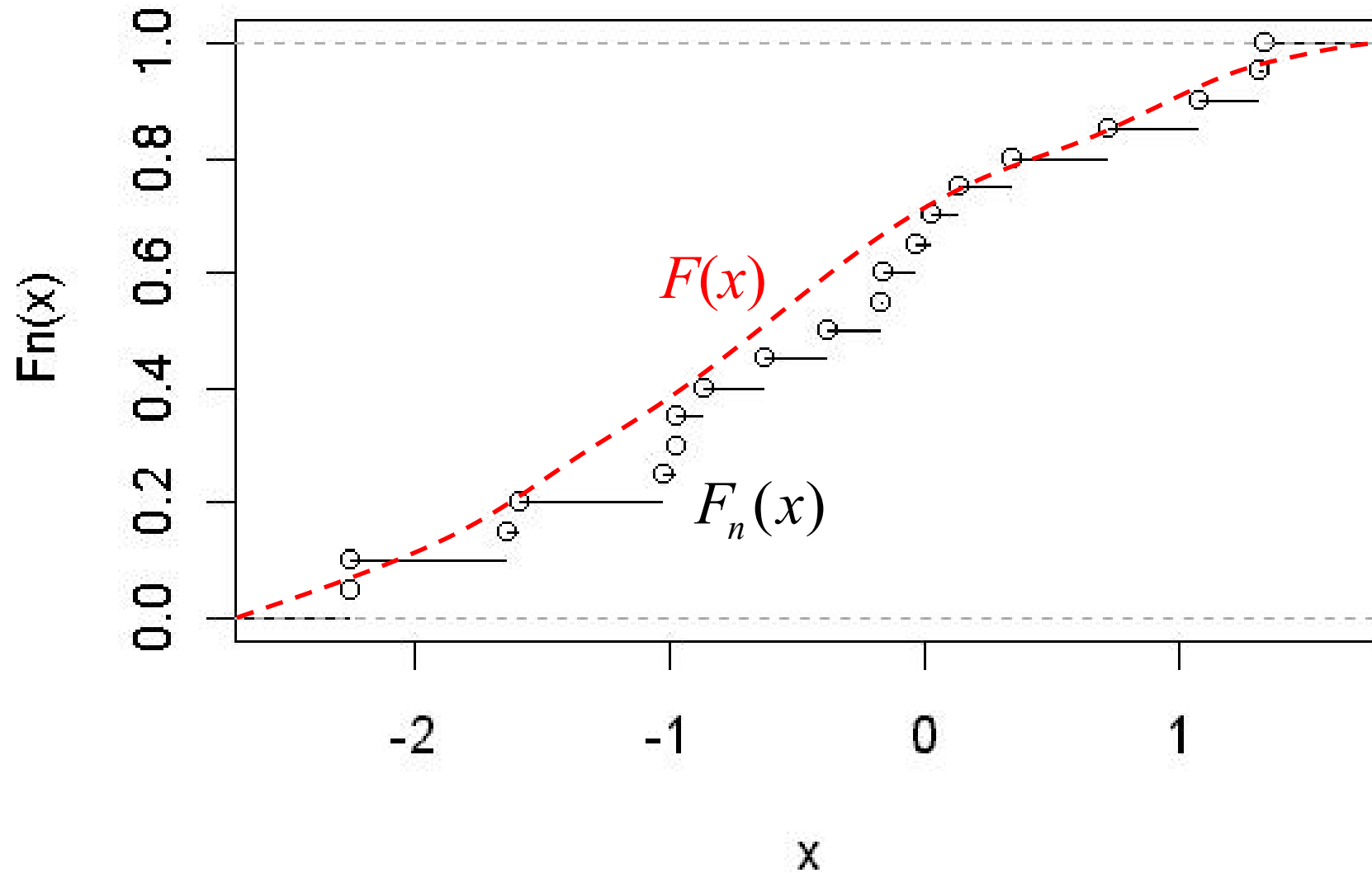
并作函数：

$$F_n(x) = \begin{cases} 0, & x < x_{(1)} \\ \frac{k}{n}, & x_{(k)} \leq x < x_{(k+1)}, (k = 1, 2, \dots, n-1) \\ 1, & x \geq x_{(n)} \end{cases}$$

称 $F_n(x)$ 为总体 X 的经验分布函数 (empirical distribution function)。对应地，总体 X 的分布函数 $F(x)$ 称为理论分布函数 (theoretical distribution function)。

Empirical distribution function

Theoretical distribution function



讨论:

(1) 对于给定的样本值 $x_1, x_2, \dots, x_n$, 经验分布函数 $F_n(x)$ 符合分布函数的三个条件:

单调非减、右连续的阶梯函数

$$0 \leq F_n(x) \leq 1$$

$$F_n(-\infty)=0, F_n(+\infty)=1$$

(2) 由于样本的随机性, 经验分布函数 $F_n(x)$ 也是随机变量。
且 $n F_n(x)$ 服从二项分布 $B(n, F(x))$

定理

设总体 X 的理论分布函数为 $F(x)$ ，经验分布函数为 $F_n(x)$ ，则当 $n \rightarrow \infty$ 时，有：

$$P\left\{\lim_{n \rightarrow \infty} \max_{-\infty < x < +\infty} |F_n(x) - F(x)| = 0\right\} = 1$$

——当样本容量 n 很大时，可以用经验分布函数去估计总体 X 的理论分布函数。

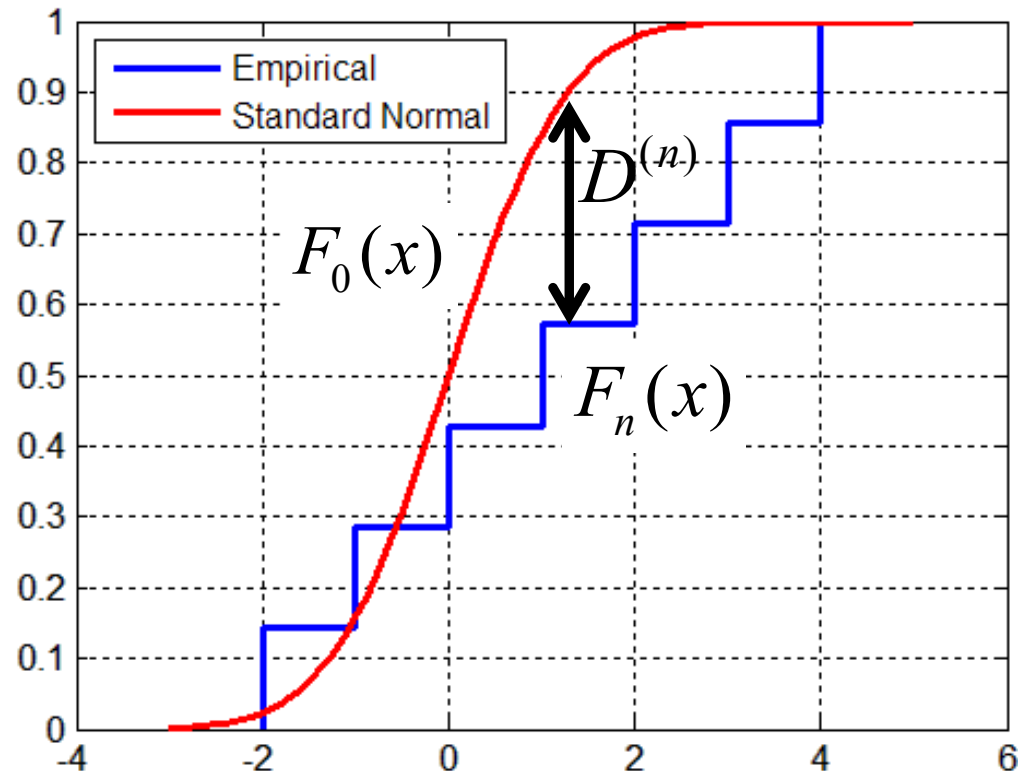
K-S检验 (K检验)

基本思想：设总体X的分布函数为 $F(x)$ ，对一个给定的分布 $F_0(x)$ ，考虑原假设 $H_0: F(x) = F_0(x)$ 。当样本容量 n 充分大时，样本的经验分布函数 $F_n(x)$ 是总体X分布 $F(x)$ 的很好的近似，因此当原假设 H_0 为真时， $F_n(x)$ 与 $F_0(x)$ 之间的差应该是一个小量。

Kolmogorov统计量：

$$D^{(n)} = \max_{-\infty < x < +\infty} |F_n(x) - F_0(x)|$$

- 反映分布函数的差异
- 是随机变量



定理 (Kolmogorov统计量的渐进分布性质)

设总体 X 的分布函数 $F(x)$ 连续, $X_1, X_2, \dots, X_n$ 为来自总体 X 的样本, 则当原假设 $H_0: F(x) = F_0(x)$ 成立时, 有:

$$\lim_{n \rightarrow \infty} P \left(D^{(n)} \leq \frac{\lambda}{\sqrt{n}} \right) = K(\lambda) \quad (D^{(n)} = \max_{-\infty < x < +\infty} |F_n(x) - F_0(x)|)$$

其中,

$$K(\lambda) = \begin{cases} 0, & \lambda \leq 0 \\ \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2 \lambda^2}, & \lambda > 0 \end{cases}$$

满足这一分布函数的分布称为K分布。

Kolmogorov统计量的良好性质: “分布无关性 (distribution-free)” (对任意连续的 $F_0(x)$ 都成立)

K-S检验法 (Smirnov, 1939)

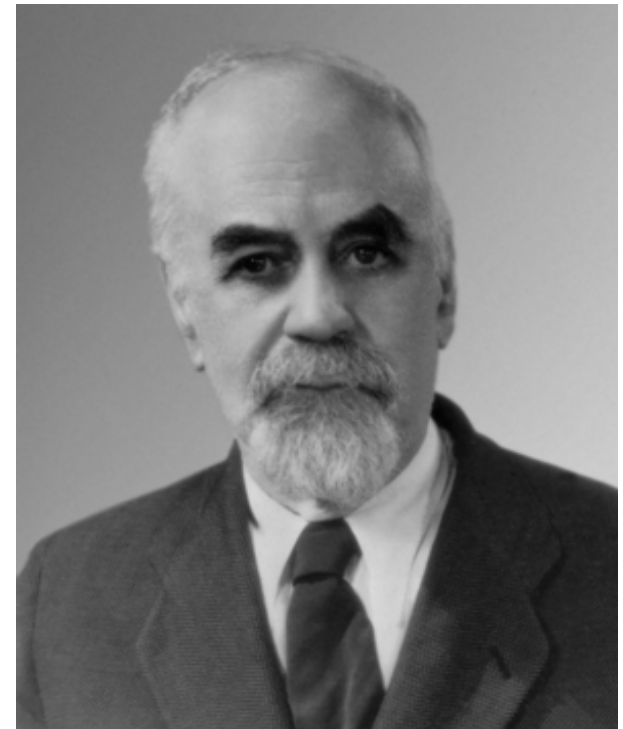
设总体 X 的分布为 $F(x)$, 考虑原假设 $H_0: F(x) = F_0(x)$, 在显著性水平 α 下的拒绝域形式为:

$$P\left(D^{(n)} \geq D_{1-\alpha}^{(n)}\right) = \alpha$$

$$D_{0.90}^{(n)} \approx \frac{1.23}{\sqrt{n}}$$

$$D_{0.95}^{(n)} \approx \frac{1.36}{\sqrt{n}}$$

$$D_{0.99}^{(n)} \approx \frac{1.63}{\sqrt{n}}$$



Vladimir Smirnov
(1887-1974)

K-S检验的基本步骤

(1) 将样本观察值 $x_1, x_2, \dots, x_n$ 生成顺序统计量 $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$, 并求出经验分布函数:

$$F_n(x) = \begin{cases} 0, & x < x_{(1)} \\ \frac{k}{n}, & x_{(k)} \leq x < x_{(k+1)}, (k = 1, 2, \dots, n-1) \\ 1, & x \geq x_{(n)} \end{cases}$$

(2) 计算Kolmogorov统计量 $D^{(n)}$:

$$D^{(n)} = \max_{-\infty < x < +\infty} |F_n(x) - F_0(x)| = \max_{1 \leq i \leq n} \left\| \frac{i}{n} - F_0(x_{(i)}) \right\|, \left\| \frac{i-1}{n} - F_0(x_{(i)}) \right\|$$

(3) 若 $D^{(n)} \geq D_{1-\alpha}^{(n)}$ 则拒绝原假设 H_0 。

$$D_{0.90}^{(n)} \approx \frac{1.23}{\sqrt{n}} \quad D_{0.95}^{(n)} \approx \frac{1.36}{\sqrt{n}} \quad D_{0.99}^{(n)} \approx \frac{1.63}{\sqrt{n}}$$

讨论

- (1) K-S检验适合对单个总体的连续型分布进行假设检验（常被称为单样本K-S检验）；
- (2) 具有较高的灵敏度；
- (3) 缺点：要依赖理论分布函数 $F_0(x)$ ，无法检验离散型总体的分布。

§ 6.4.2 χ^2 检验 (Chi-square test)

——K. Pearson (1900) ;

设总体 X 分布 $F(x)$ 未知, 已知 $X_1, X_2, \dots, X_n$ 为来自总体 X 的样本。

$$H_0: F(x) = F_0(x); \quad H_1: F(x) \neq F_0(x)$$

将总体 X 的样本空间 Ω 划分为 m 个互不相容的子空间 $A_1, A_2, \dots, A_m$, 使

$$\bigcup_{i=1}^m A_i = \Omega \quad A_i \cap A_j = \emptyset, i \neq j$$

于是当原假设 H_0 成立时, 可计算离散型分布律:

$$p_i = P(X \in A_i), i = 1, 2, \dots, m$$

在 n 次试验中实际频数为 n_i , 则 n_i/n 为事件 $\{X \in A_i\}$ 的频率。



构造Pearson χ^2 检验统计量

$$X^2 = \sum_{i=1}^m \frac{[n_i - np_i]^2}{np_i} = \sum_{i=1}^m \frac{[O_i - E_i]^2}{E_i}$$

Pearson χ^2 检验统计量的渐近分布为 $\chi^2(m-r-1)$ ，其中 r 是分布函数参数的维数。

定理 (多项分布拟合检验的渐进分布性质)

对于同一多项分布的拟合假设检验问题，在原假设 H_0 成立的条件下，当样本容量 n 趋近无穷大时，GLR检验统计量

$$GLR' = -2 \log GLR = -2n \sum_{i=1}^m \hat{p}_i \log \left(\frac{p_i(\hat{\theta})}{\hat{p}_i} \right) = 2 \sum_{i=1}^m O_i \log \left(\frac{O_i}{E_i} \right)$$

与Pearson χ^2 检验统计量

$$X^2 = \sum_{i=1}^m \frac{[X_i - np_i(\hat{\theta})]^2}{np_i(\hat{\theta})} = \sum_{i=1}^m \frac{[O_i - E_i]^2}{E_i}$$

近似等价，且二者的抽样分布都趋近自由度为 $\dim(\Theta') - \dim(\omega_0)$ 的 χ^2 分布。

这里， $O_i = X_i = n\hat{p}_i$ ， $E_i = np_i(\hat{\theta})$ 。

讨论

- (1) χ^2 检验统计量的定义与样本空间 Ω 的划分有关。只有合适的划分才能使得构造的离散型分布律能合理地近似总体分布；
- (2) 样本容量 n 应该足够大，一般 $n \geq 50$ ；
- (3) 一般要求 $np_i > 5$ ，否则应适当合并 A_i 以满足要求。
- (4) 当总体理论分布 $F_0(x)$ 是连续型时，K-S检验要比Pearson χ^2 检验效率高。

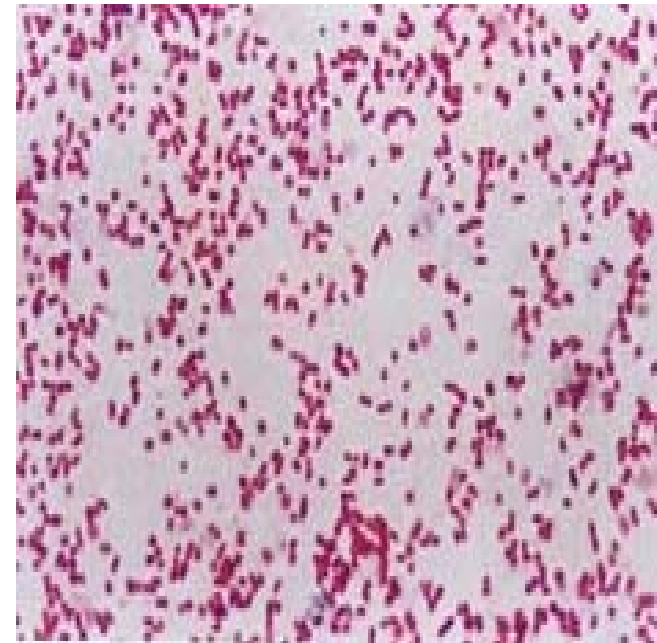
【Example 6.2】 将0.1ml受细菌污染的牛奶均匀涂在1cm²的切片上，用显微镜观察切片每个小网格内的细菌菌落数目。根据400（20×20）个小网格的计数结果，统计出如下表格：

菌落数：	0	1	2	3	4	5	6	7	8	9	10	19
频 数：	56	104	80	62	42	27	9	9	5	3	2	1

试问单位面积内的菌落数是否服从Poisson分布？

问题：

根据样本信息，判别随机变量 X 是否服从Poisson分布。



(Bliss C. and Fisher R. A., (1953). Fitting the negative binomial distribution to biological data. *Biometric*, 9: 174-200)

【Example 6.2】 将0.1ml受细菌污染的牛奶均匀涂在1cm²的切片上，用显微镜观察切片每个小网格内的细菌菌落数目。根据400（20×20）个小网格的计数结果，统计出如下表格：

菌落数：	0	1	2	3	4	5	6	7	8	9	10	19
频 数：	56	104	80	62	42	27	9	9	5	3	2	1

试问单位面积内的菌落数是否服从Poisson分布？

$$\hat{\lambda} = \bar{X} = \frac{0 \times 56 + 1 \times 104 + \dots + 19 \times 1}{400} = 2.44$$

菌落数：	0	1	2	3	4	5	6	≥7
实际频数：	56	104	80	62	42	27	9	20
期望频数：	34.9	85.1	103.8	84.4	51.5	25.1	10.2	5.0

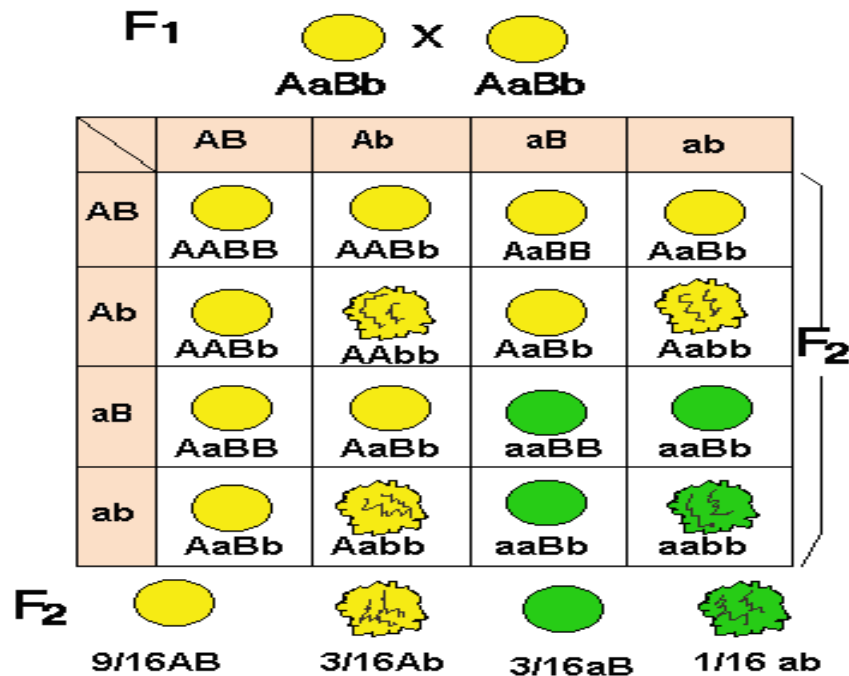
$\frac{[O_i - E_i]^2}{E_i}$:	12.8	4.2	5.5	6.0	1.8	0.14	0.14	45.0
-------------------------------	------	-----	-----	-----	-----	------	------	------

计算得到 $\chi^2=75.6$ ，服从 $\chi^2((8-1)-1) = \chi^2(6)$ 分布，查表 $\chi_{0.05}^2(6) = 18.55$ ，故 p 值小于 0.05。

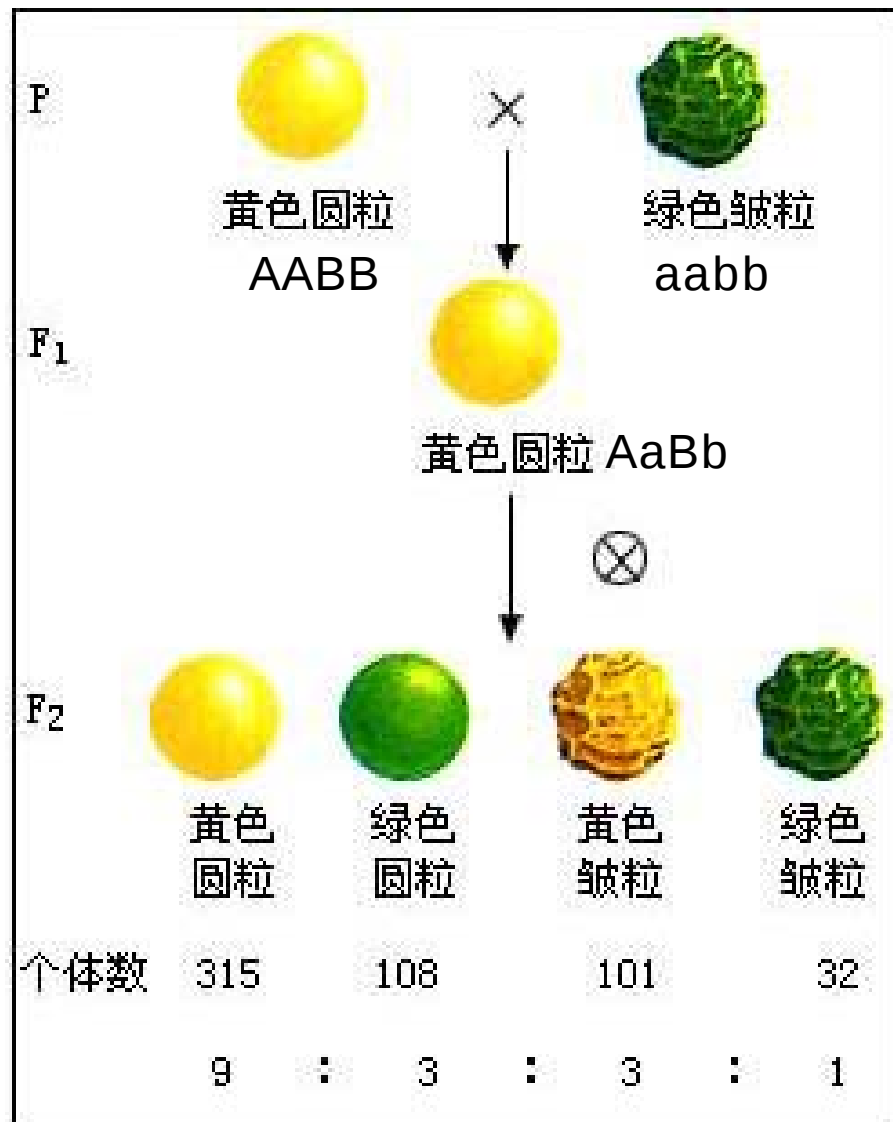
原假设被拒绝。

原因：注意第 1 栏和最后 1 栏，偏差太大！

【Example 6.6】 请对Mendel遗传试验发表的数据进行假设检验
 （注：Fisher于1936年对Mendel的实验数据进行统计分析后，认为Mendel的数据过于接近理想数据。）



孟德尔曾做过杂交试验的园地



解：分别考虑两个检验统计量 ↵

$$\begin{aligned} (1) \quad -2 \log GLR &= 2 \sum_{i=1}^m O_i \log \left(\frac{O_i}{E_i} \right) \quad \text{↵} \\ &= 0.618, GLR = e^{-0.618/2} = 0.73 \end{aligned}$$

$$(2) \quad X^2 = \sum_{i=1}^m \frac{[O_i - E_i]^2}{E_i} = 0.604 \quad \text{↵}$$

二者服从 $\chi^2((4-1)-0) = \chi^2(3)$ 分布 ($\dim(\omega_0)=0$), p-value 接近 0.9。

假设检验通过。 ↵

【Example 6.7】 调查339名50岁以上有吸烟习惯者与慢性气管炎病的关系。试问吸烟习惯与慢性气管炎是否有关系？

即检验假设 H_0 ：吸烟习惯与慢性气管炎相互独立（给定显著性水平 $\alpha=0.01$ ）。

	患病人数	未患病人数	合计	患病率
吸烟人数	43	162	205	21.0%
不吸烟人数	13	121	134	9.7%
合计	56	283	339	16.5%

§ 6.4.3 应用举例：两总体之间的独立性检验

设有二元总体 (X, Y) ， x, y 的取值范围分别划分成 m 个和 k 个互不相交的区间 $A_1, A_2, \dots, A_m$ 和 $B_1, B_2, \dots, B_k$ 。从总体中抽取容量为 n 的简单随机样本，样本值记为 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ ，记 n_{ij} 为样本值的 x 落入 A_i ， y 落入 B_j 的个数（ $i=1, 2, \dots, m; j=1, 2, \dots, k$ ），又记

$$n_{i\cdot} = \sum_{j=1}^k n_{ij}, \quad n_{\cdot j} = \sum_{i=1}^m n_{ij}$$

——列联表（contingency table）

需要检验假设

$$H_0: X \text{ 与 } Y \text{ 相互独立}$$

这种假设检验问题称为独立性检验。

	患病人数	未患病人数	合计
吸烟人数	43	162	205
不吸烟人数	13	121	134
合计	56	283	339

列联表 ($m \times k$ contingency table)

		y				$n_{i\bullet} = \sum_{j=1}^k n_{ij}$
		B ₁	B ₂	...	B _k	
x	A ₁	n_{11}	n_{12}	...	n_{1k}	$n_{1\bullet}$
	A ₂	n_{21}	n_{22}	...	n_{2k}	$n_{2\bullet}$
	...	...	...	...	...	...
	A _m	n_{m1}	n_{m2}	...	n_{mk}	$n_{m\bullet}$
$n_{\bullet j} = \sum_{i=1}^m n_{ij}$		$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet k}$	

$m \times k$ 列联表是对二元总体(X, Y)随机变量的分类表述或者连续性变量的粗粒化 (coarse-grained) 表述。

$$\left\{ \begin{array}{l} p_{ij} = P(X \in A_i, Y \in B_j) \\ p_{i\cdot} = P(X \in A_i) \\ p_{\cdot j} = P(Y \in B_j) \end{array} \right.$$

$$\left\{ \begin{array}{l} p_{i\cdot} = \sum_{j=1}^k p_{ij}, \\ p_{\cdot j} = \sum_{i=1}^m p_{ij}, \\ \sum_{i=1}^m p_{i\cdot} = \sum_{j=1}^k p_{\cdot j} = \sum_{i=1}^m \sum_{j=1}^k p_{ij} = 1 \end{array} \right.$$

回顾：随机变量的相互独立性

定理 对于二维连续型随机变量(X, Y), **X与Y相互独立的充要条件是**

$$f(x, y) = f_X(x) \cdot f_Y(y)$$

在 $f(x, y)$, $f_X(x)$, $f_Y(y)$ 的所有公共连续点上都成立。

在原假设 H_0 成立的条件下，根据二维随机变量相互独立性，有：

$$p_{ij} = p_{i\cdot} \cdot p_{\cdot j}, \quad i = 1, 2, \dots, m; j = 1, 2, \dots, k$$

假设检验就是要检验这一关系。

(1) 当 p_{ij} 已知时

		y				$n_{i\bullet} = \sum_{j=1}^k n_{ij}$
		B ₁	B ₂	...	B _k	
x	A ₁	n_{11}	n_{12}	...	n_{1k}	$n_{1\bullet}$
	A ₂	n_{21}	n_{22}	...	n_{2k}	$n_{2\bullet}$
	...	...	...	...	...	...
	A _m	n_{m1}	n_{m2}	...	n_{mk}	$n_{m\bullet}$
$n_{\bullet j} = \sum_{i=1}^m n_{ij}$		$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet k}$	n

构造检验统计量：

$$\chi^2 = \sum_{i=1}^m \sum_{j=1}^k \frac{(n_{ij} - np_{ij})^2}{np_{ij}} = \sum_{i=1}^m \sum_{j=1}^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

其渐近分布为 $\chi^2(mk - 1)$ 。

给定显著性水平 α 的拒绝域形式：

$$\chi^2 = \sum_{i=1}^m \sum_{j=1}^k \frac{(n_{ij} - np_{ij})^2}{np_{ij}} \geq \chi_{1-\alpha}^2(mk - 1)$$

(2) 当 p_{ij} 未知时

用 p_{ij} 的最大似然估计来代替

$$\hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n}, \quad \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}$$

$$\longrightarrow \hat{p}_{ij} = \hat{p}_{i\cdot} \cdot \hat{p}_{\cdot j} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n^2} = \frac{\hat{n}_{ij}}{n}$$

构造检验统计量：

$$\chi^2 = \sum_{i=1}^m \sum_{j=1}^k \frac{(n_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}} = \sum_{i=1}^m \sum_{j=1}^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

		y				$n_{i\cdot} = \sum_{j=1}^k n_{ij}$
		B ₁	B ₂	...	B _k	
x	A ₁	n_{11}	n_{12}	...	n_{1k}	$n_{1\cdot}$
	A ₂	n_{21}	n_{22}	...	n_{2k}	$n_{2\cdot}$
	...	...	...	...	...	...
	A _m	n_{m1}	n_{m2}	...	n_{mk}	$n_{m\cdot}$
$n_{\cdot j} = \sum_{i=1}^m n_{ij}$		$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot k}$	n

(2) 当 p_{ij} 未知时

		y				$n_{i\bullet} = \sum_{j=1}^k n_{ij}$
		B ₁	B ₂	...	B _k	
x	A ₁	n_{11}	n_{12}	...	n_{1k}	$n_{1\bullet}$
	A ₂	n_{21}	n_{22}	...	n_{2k}	$n_{2\bullet}$
	...	...	...	...	...	...
	A _m	n_{m1}	n_{m2}	...	n_{mk}	$n_{m\bullet}$
$n_{\bullet j} = \sum_{i=1}^m n_{ij}$		$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet k}$	n

服从 χ^2 分布，其自由度为

$$\dim(\Theta') - \dim(\omega_0) = (m-1)(k-1)$$

给定显著性水平 α 的拒绝域形式为：

$$\chi^2 = \sum_{i=1}^m \sum_{j=1}^k \frac{(n_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}} \geq \chi_{1-\alpha}^2((m-1)(k-1))$$

【Example 6.7】 调查339名50岁以上有吸烟习惯者与慢性气管炎病的关系。试问吸烟习惯与慢性气管炎是否有关系？

即检验假设 H_0 ：吸烟习惯与慢性气管炎相互独立（给定显著性水平 $\alpha=0.01$ ）。

	患病人数	未患病人数	合计	患病率
吸烟人数	43	162	205	21.0%
不吸烟人数	13	121	134	9.7%
合计	56	283	339	16.5%

解：给定 $\alpha = 0.01$ ，已知 $n=339$ ， $m=k=2$ ，

查表得： $\chi^2_{0.99}((2-1)(2-1)) = 6.635$ 。

由表计算得到： $\hat{p}_{1\cdot} = \frac{205}{339}$ ， $\hat{p}_{2\cdot} = \frac{134}{339}$ ， $\hat{p}_{\cdot 1} = \frac{56}{339}$ ， $\hat{p}_{\cdot 2} = \frac{283}{339}$

$\rightarrow \chi^2 = 7.48 > \chi^2_{0.99}(1) = 6.635$ ，

原假设被拒绝，即吸烟习惯与慢性气管炎存在关系。