附件：分析数据的解释与处理指导原则公示稿（第一次）

分析数据的解释与处理指导原则

药品质量的保证是通过一系列实践活动联合实现的，这些活动包括严格的工艺和处方设计、研发、验证以及执行稳健的控制策略，所有实践活动都依赖于可靠的检测数据及统计学分析方法。

可靠的检测数据依赖于规范的实验室活动，包括完整详实的记录保存、遵循现行版中国药典、药品生产质量管理规范（GMP）和实验室评定认可管理的相关法规要求的分析方法和样品性能确认等。在此基础上，恰当的将统计分析方法应用在产品全生命周期的控制策略中，可进一步确保产品质量。

本指导原则提供对药品生产工艺和产品的检测数据进行决策时的基本统计规范，包括对分析数据的描述性统计、推断性统计、基本统计原理及常用统计分析方法等，有助于科学规范的对分析数据进行处理、解释和科学呈现。

需要注意的是：（1）除本指导原则外其他合理的统计学方法也可由使用者根据自己的判断去使用；（2）本指导原则所提供数据/表格没有规定单位，其量纲仅用于说明计算方法的通用性，不直接与具体产品或检测方法相关；（3）本指导原则所提供的计算方式得到的结果不与产品的安全性和有效性直接相关，还需要结合具体产品的药理毒理及临床监测信息一同做出判定。

1. 描述性统计

描述性统计包括描述数据的统计量、统计表和统计学图形。统计量分为集中趋势和离散程度统计量。统计表包括描述连续性数据的频数统计表、描述离散性数据的列联表和汇总统计表。统计学图形包括直方图、箱线图、Q-Q 图等。下面仅对药品领域最常使用的一些统计量、统计学图形进行介绍，其他内容可参考相关统计资料和工具。

1.1 统计量

统计量用于评估样本所代表总体的集中趋势或离散程度，包括均值、标准差以及从中衍生出的其他指标，例如变异系数，也称为相对标准偏差。统计量还可

用于计算置信区间、预测区间和容忍区间等。

1.1.1 集中趋势统计量

（1）算术均值（arithmetic mean） 算术均值是对观测值分布的中心性或集中趋势的度量。它最适用于对称分布，并受分布非对称性(形状)和极值的影响。算术均值的计算是n个样本值（$x_1,x_2,...x_n$）的和除以数值个数n，公式为：

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1+x_2+\cdots+x_n}{n} \quad (1)$$

（2）几何均值（geometic mean，GM） 以对数转换方式获得的均值（如生物反应的剂量值），应使用几何均值(GM)表达，详见“4.4.3.1对数转换”。

（3）中位数（median）或第50百分位数 中位数也是中心性或集中趋势的度量，通常不受分布的极端影响。它是一个将分布划分为两个相等部分的值。对于连续分布，中位数的两侧各有50%的数据。要获得样本的第50百分位数，按递增的顺序排列样本的n个值。当n为奇数时，中位数是第$(n+1)/2$个值。当n为偶数时，中位数位于第$(n/2)$个和第$(n/2)+1$个值之间，然后取这两个值的算术平均值。

与均值相比较，中位数对观测值分布的集中趋势的度量更为稳健，特别是对于呈偏态分布的数据集。中位数这一统计量较为粗糙，使用时会使数据失去较多信息。

1.1.2 离散程度统计量

（1）极差（range） n 个观测值中最大和最小观测值之间的差值称为极差 R，它用作变异的度量。公式为：

$$R = max(x) - min(x) \quad (2)$$

极差有助于评估样本的变异，具有易于计算和理解的特性。但是，当样本大小从适中到较大时，需要谨慎使用，因为最小值和最大值均来自分布的尾部，通常较为多变。因此，样本极差直接受到极端值的影响。一般来说，样本的标准差比极差能更好的描述数据的离散程度。

极差对于小样本较为适用。例如当n = 2 ~ 12时，标准差的计算比较密集和抽象，可能会导致计算不合理，此时可使用极差。

极差在质量控制的应用中非常重要。当有多个可用的样本极差，且所有样

本极差都具有相同的样本大小n时，取平均极差除以常数d_2可作为对标准差的估计。d_2是正态分布下样本大小为n的期望极差与标准差之比，其取值详见“4.1.1 单值控制图”中的表9。

（2）方差（variance，S^2） 样本的方差是观测值（$x_1,x_2...x_n$）与其平均值（$\bar{x}$）的差方和，除以观测值的数量（n）减去 1，用以描述变异的度量。公式为：

$$S^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} \quad (3)$$

（3）标准差（standard deviation, S） 也叫标准偏差，是方差的平方根，见公式 4。对于样本量适中到较大且呈对称的钟形分布，如正态分布时，可以结合经验规则使用标准差。即大约 68%的数据将落在均值的±1 个标准差范围内；大约 95%的数据将落在±2 个标准差范围内；几乎所有（99.7%）数据将落在±3 个标准差范围内（见表 1 和图 4）。当样本量非常大或接近无限且分布为正态分布时，样本均值接近真值的程度会提高。基于它们与正态分布的相似性，该规则适用于其他对称的钟形分布。

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} \quad (4)$$

表 1 正态分布的曲线下面积

区间	曲线下面积	区间	曲线下面积
$\mu\pm1\sigma$	0.68270	$\mu\pm0.674\sigma$	0.50
$\mu\pm2\sigma$	0.95450	$\mu\pm1.645\sigma$	0.90
$\mu\pm3\sigma$	0.99730	$\mu\pm1.960\sigma$	0.95
$\mu\pm4\sigma$	0.99994	$\mu\pm2.576\sigma$	0.90

（4）标准误差（standard error，se） 也叫标准误，用于描述抽样统计量对总体统计量的偏离程度，即通过抽样计算得到统计量的准确性。当对同一总体进行多次重复抽样时，每次抽样计算的统计量不会完全一样，这是因为统计量是一个随机变量，每次在重复抽样中观测值是不同的，必然导致计算的统计量不同。而标准误是重复抽样中样本统计量（如均值）的标准差，也是样本统计量的标准不确定度。公式为：

$$se(\bar{x}) = \frac{\sigma}{\sqrt{n}} \approx \frac{S}{\sqrt{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n(n-1)}} \quad (5)$$

其中，σ 代表未知总体标准差，S 是样本标准差，n 是样本量。

（5）变异系数（coefficient of variation，CV） 变异系数也叫相对标准差（relative standard deviation, RSD），是标准差与均值的非负数率值，通常以百分位数表示。应注意，当使用所测特性的绝对量值时，*CV* 才能恰当的度量变异度，如药品含量值。针对以对数转换方式获得的值（如生物反应的剂量值），应使用几何变异系数(*GCV*)表达，详见“4.4.3.1 对数转换”。

（6）Z-分数（Z-score） 在包含 n 个不同观测值（$x_1,x_2...x_n$）的样本中，每个观测值都有一个关联的 *Z*-分数。对于观测值 x_i，相关联的 Z-分数表示为观测值 x_i 偏离样本均值$\bar{x}$的标准差 S 倍数。Z-分数的计算公式为：

$$Z = \frac{(x_i - \bar{x})}{s} \quad (6)$$

Z-分数可用于识别一组数据是否存在异常值。特别是在正态分布中，大于3倍标准差的值不常见（少于0.3%），因此利用这一性质可有效识别异常值。但使用此规则时应注意样本量适中且数据呈正态分布，对于偏态分布的数据不一定适用于该方法。

1.2 统计学图形

统计学图形是一种对数据直观描述的方式，其种类较多，此处简要介绍直方图、箱线图和残差图。

1.2.1 直方图

从一组变量数据的概率分布和描述性统计量中，可以构造出对数据集的解释有很大帮助的各种图，其中最基本的图形是频率或相对频率直方图。直方图的绘制首先要将数据集分成一系列相等的间隔，通常以间隔的中点为中心，其高度等于数据集中落在该间隔的频率(或相对频率)。

绘制直方图的一个重要目的是能可视化地评估数据集分布的一般形状，如图1。可以直观回答如数据是否来自对称分布、数据中是否存在异常值等问题。

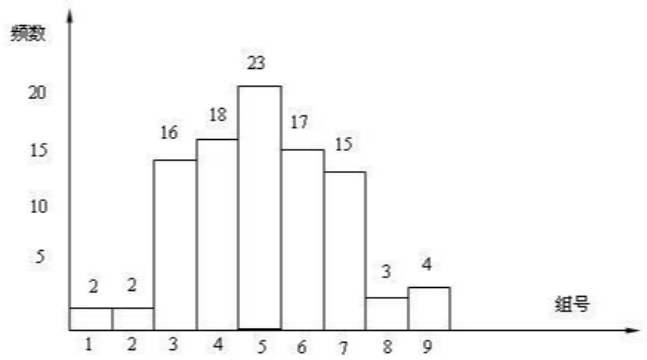

图 1 直方图

1.2.2 箱线图

箱线图又称为盒须图，也是描述分布状态的较为直观的图形。为了构建箱线图，需要从样本中获得 4 个数值：最小值、最大值、第 25 和第 75 百分位数值。其中两个百分位数值将表示为第 1 和第 3 四分位数(*Q1* 和 *Q3*)。中位数（*Q2*）和均值也可以在图上表示出来，见图 2。

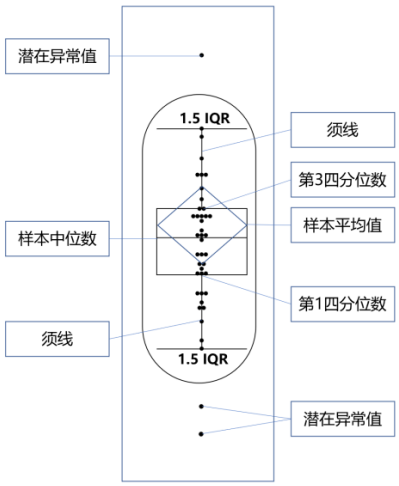

图 2 箱线图图解

箱线图可以作为识别潜在异常值的图解方法。图中超过任意一侧(从*Q1*或*Q3*) 1.5倍四分位间距（interquartile range, *IQR*）距离的数据点，表明是潜在的异常值。需要注意的是该方式是异常值识别的图示法，但不是异常值检验法，有关异常值检验可参考“4.4.5异常值”和中国药典通则1431生物检定统计法“异常值剔除和缺项补足”部分。

1.2.3 残差图

残差图也叫残差分布图，是在垂直轴上绘制残差，在水平轴上绘制预测响应构成的图；正常情况下，残差应处于随机分布的散落点，如图3-A。残差的任何线性或非线性趋势都表明所用模型的方程可能不正确，或模型中缺少重要的处理因素。例如，弯曲的残差图案可表明模型中需要增加二次项，如图3-B。此外，落在点集之外较远的残差可能是异常值，如图3-C。当残差图上的观测值沿水平轴的展宽增加或减少时，表明违反了恒定方差的假设，如图3-D。

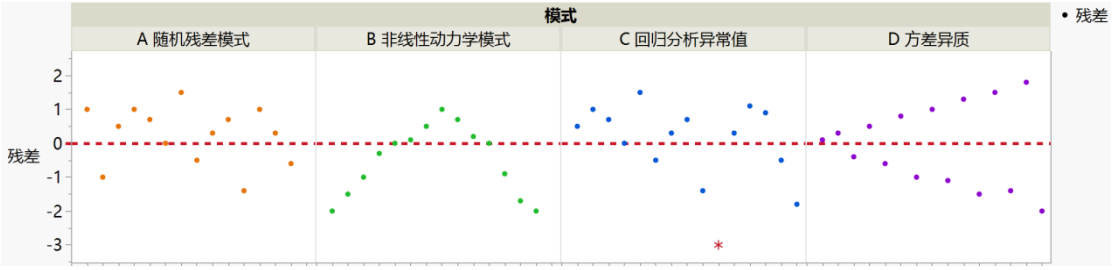

图 3 残差模式分类图

2. 推断性统计

统计推断是通过随机样本获取的信息推断总体特征的过程。统计学推断有三种方式，分别为频率推断方式、贝叶斯推断方式和可信推断方式。这里主要以频率推断方式为重点进行介绍。贝叶斯推断法在“4.5贝叶斯推断统计”中进行专门的简介，并与频率推断法进行比较。可信推断法目前已很少使用。

统计推断的基本问题可以分为两大类：一类是对总体的未知参数进行估计；另一类是关于参数的假设检验。

2.1 总体参数估计

2.1.1 值的估计

研究者期望通过检测数据估计其真值，即总体参数。估计值的类型分为数值型和非数值型。本指导原则重点介绍数值型估计值。

数值型估计值分为点估计值、区间估计值和分布估计值。点估计值是指“最佳”地表示总体参数的一个未知真值，其典型实例是从研究总体中抽取样本估计出的均值和标准差。这里所谓的“最佳”，表示估计值是相对接近未知参数值，尽

管这些点估计值与参数值之间会因样本的不同而产生差异。

由于点估计值的报告结果不能反映与真值间差异的不确定度，故使用区间估计值来表达。有关区间估计值的讨论详见 “2.1.2 区间估计”。

2.1.2 区间估计

2.1.2.1 置信区间

1、置信区间介绍

未知总体参数的置信区间是使用样本数据构建，并以概率的形式提供有关该参数估计的不确定性信息。置信区间是表达总体参数在规定置信水平下可能出现的范围，它由置信下限 (C_L) 和置信上限 (C_U)组成。其中置信水平$C=1-\alpha$反映了在相同条件下，一系列重复随机样本中置信区间$[C_L, C_U]$包含或覆盖参数真值的占比，通常以百分比表示，最高不包括100%，常用的置信水平为90%、95%和99%。

置信限是由样本得出的统计量，故每次重复抽样时会有所不同。如果置信下限 (C_L) 和置信上限 (C_U)落在真实参数值的两侧，则说明置信区间包含、覆盖或捕获了所考察的参数。

2、置信区间的计算

（1）未知参数的近似置信区间计算：

$$\hat{\theta} \pm z_{1-\frac{\alpha}{2}} \times se\left(\hat{\theta}\right) \quad (7)$$

式中，参数$\hat{\theta}$是一个统计量值，$se(\hat{\theta})$是$\hat{\theta}$的标准误差估计值；乘数$z_{1-\alpha/2}$是从标准正态分布的$(1-\alpha)$双侧置信区间选择的$1-\frac{\alpha}{2}$分位数。例如，当使用95%置信水平（$\alpha = 0.05$）时，$z_{0.975} = 1.960$；当使用99%置信水平时，$z_{0.995} = 2.576$。

（2）当参数为均值且总体分布为正态分布时的精确双侧置信区间计算：

$$\bar{x} \pm \frac{t_{1-\alpha/2,df} \times S}{\sqrt{n}} \quad (8)$$

式中，$t_{1-\alpha/2,\, df}$是当标准差S，具有df自由度，且置信水平为α时，具有df自由度的学生t分布的$1-\frac{\alpha}{2}$分位数。

（3）单侧置信区间的计算：

A. 置信区间下限：

$$C_L = \bar{x} - \frac{t_{1-\alpha,\ df} \times s}{\sqrt{n}} \quad (9)$$

B. 置信区间上限：

$$C_U = \bar{x} + \frac{t_{1-\alpha,\ df} \times s}{\sqrt{n}} \quad (10)$$

由于不同分布的相对精确置信区间的计算方式各不相同，需要根据具体分布确定其公式。可参考相关资料进行计算。当参数不符合正态分布或找不到合适的分布时，也可以参考蒙特卡洛法，该方法几乎适用于计算各种分布参数的置信区间。

2.1.2.2 预测区间

1、预测区间的介绍

预测区间是根据在相同条件下生产的早期样本的结果，预测一定数量的未来样本结果所处的范围。根据来自正态分布的 n 个观测值的样本构建一定置信水平下包含一个或多个未来观测值的区间，这样的区间称为预测区间。

2、预测区间的计算

（1）正态总体中单个未来观测值的双侧预测区间计算

正态总体中单个未来观测值 y 的预测区间是使用正态分布中 n 个观测值的样本构建的，并提供未来值在一定置信水平下预计所处的范围。假设 y 是一个未来观测值，其双侧区间预测限为 $P_L \leq y \leq P_U$。预测区间的计算公式为：

$$y = \bar{x} \pm t_{1-\frac{\alpha}{2},n-1} S\sqrt{1+1/n} \quad (11)$$

式中，$t_{1-\frac{\alpha}{2},n-1}$是在*n-1*个自由度下，来自学生t分布的第$1-\frac{\alpha}{2}$分位数；$\bar{x}$和$S$是原始样本的均值和标准差；$n$为样本数量。

（2）对于正态分布总体下单个未来结果单侧预测区间计算

正态总体单个未来观测值的单侧预测区间可使用下面的公式：

A. 预测区间下限：

$$P_L = \bar{x} - t_{1-\alpha,n-1}\ S\sqrt{1+1/n} \quad (12)$$

未来观测值*y*在置信度*100(1–α)%*下应满足$y \geq P_L$。$t_{1-\alpha,n-1}$ 是在*n-1*个自由度

下，学生*t*分布的第*1-α*分位数。

B. 预测区间上限：

$$P_U = \bar{x} + t_{1-\alpha,n-1}S\sqrt{1+1/n}\quad(13)$$

未来观测值*y*在置信度*100(1–α)%*下满足$y \le P_U$。$t_{1-\alpha,n-1}$ 是在*n-1*个自由度下，学生*t*分布的第*1-α*分位数。

（3）来自正态总体的多个未来观测值的预测区间计算

预测区间公式11~13可以修改为适用于多个未来观测值。仅将公式中使用的t值的分位数水平略有修改。当使用*C = 1–α*的置信水平计算*m*个未来观测值的预测区间，且学生*t*分布的分位数值在双侧检验时，使用$t_{1-\frac{\alpha}{2m},n-1}$；在单侧检验时，使用$t_{1-\frac{\alpha}{m},n-1}$。

2.1.2.3 容忍区间

1、容忍区间的介绍

正态分布的容忍区间是使用一组样本数据构建一个在规定置信水平下，所有未来观测值中至少有指定比例（*P*）的样本数据所处的区间。这里计算的容忍区间均是基于正态分布的稳定过程中随机选择的样本。容忍区间与预测区间的情况一样，也存在单侧和双侧两种情况且具有各种假设的其他情况。

预测区间和容忍区间的目的都是在考虑总体均值和标准差的不确定性以及总体保持稳定的情况下，从所关注总体中获取多个未来值。通常，预测区间用于预测总体或过程中一个或几个未来值的范围，而容忍区间用于预测未来整个总体或过程中一定比例（*P*）数据结果所处的范围值。基于n个观测值的当前样本，在一定置信水平下，可以计算m个未来值的至少一部分*P*的预测区间。如果保持*P*不变，并增加m至无穷大，则预测区间即变成一个容忍区间。

2、容忍区间的计算

报告容忍区间结果时，常用“99/95容忍区间”表示在置信水平为*C*=0.95 (95%)，未来观测值中占比*P*=99%比例的区间大小。不同情形下的容忍区间计算公式如下：

（1）正态分布下的双侧容忍区间计算:

$$\bar{x} \pm k_2 S\quad(14)$$

这里，$\bar{x}$是样本均值，S是样本标准差，k_2为双侧容忍区间的包含因子。样本量为n，因子k_2的确定公式是n、P和置信水平$C=1-\alpha$的函数，可通过公式15或查表5-3获得：

$$k_2 = \sqrt{\frac{Z^2_{(1+P)/2} \times (n-1)}{\chi^2_{\alpha:n-1}} \times \left(1+\frac{1}{n}\right)} \quad (15)$$

$\chi^2_{\alpha:n-1}$是卡方分布的右尾概率界值；$Z^2_{(1+P)/2}$是标准正态分布在（1+P）/2面积左侧的百分位数的平方。

表 2 双侧正态包含因子 k_2 值表

n	*C*	*P*	*k*	*n*	*C*	*P*	*k*
10	0.90	0.950	3.021	10	0.95	0.950	3.382
10	0.90	0.990	3.970	10	0.95	0.990	4.445
10	0.90	0.9973	4.623	10	0.95	0.9973	5.176
20	0.90	0.950	2.565	20	0.95	0.950	2.752
20	0.90	0.990	3.3721	20	0.95	0.990	3.617
20	0.90	0.9973	3.926	20	0.95	0.9973	4.213
30	0.90	0.950	2.413	30	0.95	0.950	2.550
30	0.90	0.990	3.171	30	0.95	0.990	3.351
30	0.90	0.9973	3.694	30	0.95	0.9973	3.904
40	0.90	0.950	2.334	40	0.95	0.950	2.445
40	0.90	0.990	3.067	40	0.95	0.990	3.213
40	0.90	0.9973	5.372	40	0.95	0.9973	3.742
50	0.90	0.950	2.284	50	0.95	0.950	2.379
50	0.90	0.990	3.001	50	0.95	0.990	3.126
50	0.90	0.9973	3.495	50	0.95	0.9973	3.641

（2）正态分布下的单侧容忍区间计算:

$$(-\infty, \bar{x}+k_1S] \text{或} [\bar{x}-k_1S, +\infty) \quad (16)$$

公式16所使用的形式取决于是期望容忍下限还是容忍上限。其中k_1为单侧容忍区间的包含因子，可通过公式17或查表3得到：

$$k_1 = \frac{t_{\alpha,n-1,\delta}}{\sqrt{n}} \text{ 其中，} \delta = z_p\sqrt{n} \quad (17)$$

z_p是标准正态分布在P面积左侧的百分位数。

表 3 单侧正态包含因子 k_1 值表

n	C	P	k_1	n	C	P	k_1
10	0.90	0.950	2.568	10	0.95	0.950	2.911
10	0.90	0.990	3.532	10	0.95	0.990	3.981
10	0.90	0.99865	4.498	10	0.95	0.99865	5.058
20	0.90	0.950	2.208	20	0.95	0.950	2.396
20	0.90	0.990	3.052	20	0.95	0.990	3.295
20	0.90	0.99865	3.895	20	0.95	0.99865	4.197
30	0.90	0.950	2.080	30	0.95	0.950	2.220
30	0.90	0.990	2.884	30	0.95	0.990	3.064
30	0.90	0.99865	3.686	30	0.95	0.99865	3.909
40	0.90	0.950	2.010	40	0.95	0.950	2.125
40	0.90	0.990	2.793	40	0.95	0.990	2.941
40	0.90	0.99865	3.574	40	0.95	0.99865	3.756
50	0.90	0.950	1.965	50	0.95	0.950	2.065
50	0.90	0.990	2.735	50	0.95	0.990	2.862
50	0.90	0.99865	3.502	50	0.95	0.99865	3.659

上限区间$(-\infty, \bar{x} + k_1S]$保证在置信水平$C$下，至少有总体中$P$比例的样本将小于$\bar{x} + k_1S$限值。对应地，容忍区间下限$[\bar{x} - k_1S, +\infty)$保证在置信水平$C$下，至少有总体中$P$比例的样本将大于$\bar{x} - k_1S$限值。表3是正态分布下，通过所选$n$、$C$和$P$值，确定单侧容忍区间$k_1$值的基本表。

2.1.3 分布估计

若总体参数被视为随机变量，可使用贝叶斯分析的计算值来定义期望值；它

利用先验信息和样本信息相结合形成的后验分布来确定未知参数落在给定范围内的概率。“4.5 贝叶斯推断统计”详细地描述了分布估计值的应用。

2.2 假设检验

假设检验用于决定是否拒绝原假设的过程和决策标准。同义词包括统计学检验、统计学假设检验和显著性检验。

2.2.1 假设检验的要素

假设检验包含三个要素，即假设、检验统计量和显著性水平。具体含义如下：

假设（包括原假设和备择假设）：原假设H_0是关于概率分布参数或概率分布类型的陈述，在使用统计假设检验被拒绝之前暂定为真。备择假设（H_a）通常是与原假设不同的概率分布或概率分布类型。备择假设代表着研究目标，是一种希望用真实数据证明比原假设更可信的陈述。

检验统计量：从感兴趣变量的样本观测值中计算获得的一个量值，其概率分布在原假设下是已知的。在选择检验统计量时，要求偏向备择假设的方向相对灵敏，常见的检验统计量见表6。

显著性水平：假设检验基于检验统计量的分布并假设原假设为真而否定原假设的概率，由假定原假设成立的数据分布计算得出。常用的显著性水平为α = 0.05、0.01和0.001。许多统计表给出了在一定显著性水平下的临界值大小。

2.2.2 两类错误

I 型错误（type I error）：当原假设实际为真时，拒绝原假设的错误。发生 I 型错误的概念由显著性水平 α 控制，故也叫 α 错误。

II 型错误（type II error）：当备择假设为真时，不拒绝原假设的错误。发生 II 型错误的概率指定为 β，故也叫 β 错误。

在假设检验中，I 型错误（α）和 II 型错误（β）的关联见表 4 和表 5。

表 4 统计学检验中的结论

	若原假设为真	若原假设为假
拒绝原假设	错误结论（Ⅰ型错误）	正确结论

无法拒绝原假设	正确结论	错误结论（II 型错误）

表 5 错误结论的概率

误判类型	发生概率
I 型错误	α（显著性水平）
II 型错误	β（$1-\beta$ 为效能）

为了同时控制 I 型错误（α）和 II 型错误（β），确定所需的样本量非常重要。许多教科书和软件包中都有支持推断研究的样本量计算公式，这些公式依赖于（α）和（β）所选的值。当研究的设计中包含区组，如分析人员或检测日期等实验因素时，这些公式将变得更加复杂。这时，最好使用计算机模拟，并请教专业统计人员。

2.2.3 差异显著性假设检验步骤

差异显著性检验步骤包括：首先基于研究目的制定一个统计假设；其次，建立所检验统计量的分布，最后，比较其与其他相同或类似假设的检验性能。当一个统计量通过假设计算出观测结果的显著性水平（即*p*值）小于规定的显著性水平（如$\alpha = 0.05$）时，拒绝原假设。

1、制定一个统计假设

差异显著性假设检验中原假设（H_0）和备择假设（H_a）均与未知总体参数相关。总体参数通常用希腊字母*theta*（θ）表示。双侧差异显著性假设检验可以写为：

$$H_0: \theta = \theta_0$$

$$H_a: \theta \neq \theta_0 \qquad (18)$$

其中θ_0表示θ的假定值。一般备择假设代表着研究目标，所以有时也称它为研究假设。假定用一个线性模型的真实斜率代表某化合物的纯度随时间的平均变化。通常，斜率这个参数用希腊字母beta（β）表示。研究人员打算确定是否有证据表明纯度的平均变化是时间的函数。也就是说，是否可以证明斜率的真实值不为零。因此，公式18写为：

$$H_0: \beta = 0$$

$$H_a: \beta \neq 0 \qquad (19)$$

由于研究目标是关注并找出差异的变化，并没有特别规定具体变化方向（可能为正向或负向），故称为双侧假设检验。

若研究的目标是单纯地证明正向变化，如确定化合物的杂质含量随放置时间推移而增加，这时，需要使用单侧差异显著性假设检验。单侧假设陈述如下：

$H_0: \beta \geq 0$

$H_a: \beta < 0$ （20）

在制定研究目标、进行设计和实施该研究之前，应先选定好是使用双侧还是单侧假设检验。选择的依据应基于研究的科学目标，而不是研究结果来决定。

如在比较新老两个方法时，原假设会假定两方法的总体均值相等，而备择假设则假定二者均值不等。表示为：

$H_0: \mu_N = \mu_O$

$Ha: \mu_N \neq \mu_O$ （21）

或等同地：

$H_0: \mu_N - \mu_O = 0$

$Ha: \mu_N - \mu_O \neq 0$ （22）

其中μ_N和μ_O分别是新老方法的均值。

2、选择检验统计量

统计假设检验的方法根据其研究目的、检验统计量等不同，而选择不同的统计检验方法。这可在教科书和文献中找到，一般统计软件也提供必要的计算，这里不再赘述。表6给出了一些常用的假设检验统计量示例。

表 6 样本比较的假设检验常见检验统计量

假设		检验统计量	分布
一个正态均值 H_0：$\mu=\mu_0$	当已知标准差σ时	$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	标准正态分布（0,1）
	当未知标	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	学生 t 分布，

	准差时		*df=n-1*
两个均值的比较 $H_0:\mu_1=\mu_2$	配对观测值$d_i=y_i-x_i$	$t=\frac{\bar{d}}{s_d/\sqrt{n}}$	学生 *t* 分布，*df=n-1*
	两独立样本	$t=\frac{\bar{y}-\bar{x}}{S_p/\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$，其中 $S_p=\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$	学生 *t* 分布，*df*=n_1+n_2-2
一个正态标准差 $H_0:\sigma=\sigma_0$		$\chi^2=\frac{(n-1)S^2}{\sigma_0^2}$	卡方分布，*df=n-1*
两独立样本的标准差 $H_0:\sigma_1=\sigma_2$		$F=\frac{S_1^2}{S_2^2}$	F 分布，$df_1=n_1-1$，$df_2=n_2-1$
一个二项分布比例 $p=p_0$		*x*=观测计数值	二项分布(P_0, *n*)
两个比值的比较 $p_1=p_2$		四格表（2×2）的卡方值 f_{11} | f_{21} | $f_{\cdot 1}$ f_{12} | f_{22} | $f_{\cdot 2}$ $f_{1\cdot}$ | $f_{2\cdot}$ | $f_{\cdot\cdot}$ 备注：$f_{1\cdot}=f_{11}+f_{12}$；$f_{2\cdot}=f_{21}+f_{22}$. $f_{\cdot 1}=f_{11}+f_{21}$；$f_{\cdot 2}=f_{12}+f_{22}$. $e_{ij}=\frac{f_{i\cdot}f_{\cdot j}}{f_{\cdot\cdot}}$ $\chi_1^2=\sum\frac{(f_{ij}-e_{ij})^2}{e_{ij}}$	卡方分布，*df*=1
一个泊松比 $\lambda=\lambda_0$		*x*=观测到的计数	泊松分布(λ_0)
两个泊松比的比较 $\lambda_1=\lambda_2$		x_1=根据比为 1 的计数，给出总计数	二项分布(*1/2*, x_1+x_2)

3、显著性水平与统计量的概率值（p）

显著性水平（α）应根据试验背景结合风险大小事先在试验方案中确定。当原假设成立的数据（类似）分布统计表所对应统计量的概率值（p）确定后，即

可将p值与α值进行比较，从而确定是否具有显著性。如，假定事先确定的显著性水平为$\alpha = 0.05$，则当统计量的p值小于0.05时，说明该数据与原假设有显著性差异。

2.2.4 差异性假设检验的局限性

使用假设检验来确定一个因素是否有影响，可通过拒绝原假设方式实现。在差异显著性检验这一经典推断方法中，如果证据足以否定原假设，那么会接受备择假设。但当无法否定原假设时，只能证明两者无显著性差异，而不能错误地认为原假设成立。实际上，未能拒绝原假设可能是由于变异性很大，或用于检验推断的样本量过小（效能过低），导致无法检测到差异。

差异显著性检验也存在一定的误判风险。其常见的局限性如下：

（1）有效的统计推断需要在收集或分析数据之前，对所关注的总体进行定义。在数据收集之后，仅选择特定子集的数据进行分析，其结论会不可靠。

（2）为了足够完善地确定一个分布，统计假设通常会涉及到不止一个特定参数。例如，对均值的 z 检验，不仅取决于 σ 的值，还取决于总体是否服从正态分布，以及观测值之间是否相互独立。如果不满足这些辅助假设，拒绝的概率就会很高，尤其是在大样本量的情况下。

（3）对一个数据集进行多重比较时，至少会有一次具有统计显著性的概率大于检验的名义显著性水平。故，在进行多重比较时，建议使用一种校正整体显著性水平的方法，如 Bonferroni 显著性水平校正法。

（4）观察研究和试验研究中的响应，可能比统计模型所能表达的更为复杂，如可能会包括霍桑效应（Hawthorne effects）和安慰剂效应（placebo effects）。在这些研究中，除了具有所应用的任何治疗效果外，还会含有影响受试者的响应效果。

基于以上差异显著性检验的局限性，除了建议关注假设检验的结论在新的独立数据样本中应具有可重复性外，更建议考虑选择使用频率推断法的等效性检验。等效性检验是一种更适合药品领域的推断方法，可以更好地控制 I 型错误和 II 型错误，但目前药品领域中应用较少，该内容详见“4.3 等效性与非劣效性检验”。

3. 基本统计原理

3.1 正态分布

正态分布（Normal distribution），又名高斯分布。正态分布的概率密度函数曲线呈钟型，两头低，中间高，左右对称；因其曲线呈钟形，因此经常称之为钟形曲线，如图 4。

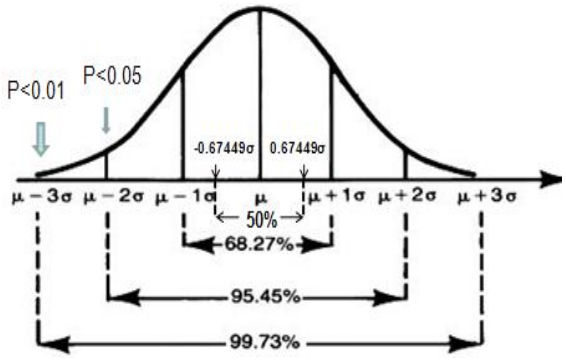

图 4 正态分布的概率密度函数曲线

图 4 中，μ 为总体均值，σ 为总体标准差。

正态分布是很多连续型数据比较分析的前提，比如 *t* 检验、方差分析、相关分析以及线性回归等，均要求数据服从正态分布或近似正态分布。

正态分布的检验方法包括图示法（详见“1.2 统计学图形”）、非参数检验方法等。针对不满足正态分布的数据集，可采用数据转换（详见“4.4.3 数据转换”）的方式转化为（近似）正态分布。

3.2 自由度

统计学上的自由度是指当以样本的统计量估计总体的参数时，样本中独立或能自由变化的变量的个数，称为该统计量的自由度。

例如，n个观测值（$x_1,x_2...x_n$）与其样本平均值的差值之和为0，即 $\sum(x-\bar{x})=0$。根据这个特性，在计算样本方差$s^2=\sum(X_i-\bar{x})^2/(n-1)$时，由于要首先计算样本均值以替代未知的总体参数$\mu$，故要损失一个自由度，即$n$个样本值中只有$n-1$个可以自由变化。一般来说，每估计一个参数，就需要消耗一个自由度。在简单线性回归中，有n对数据(x_i, y_i)，想要对n对数据拟合$y=ax+b$形式的线性模型，这里有两个需要估计的参数（a和b），此时，实际上

损失了2个自由度。该方式可进一步扩展到估计k个参数的多变量回归分析中，k个参数，意味着损失k个自由度。

3.3 抽样与样本量

3.3.1 抽样

在实践活动实施中，有效避免引入系统误差或偏倚至关重要。偏倚可能会因实验条件无意的变化、一些已知或未知因素的引入而发生。为减轻偏倚的影响，必须考虑采用有效的抽样和随机化。如何抽取样本完全取决于数据要回答的问题，在可能的情况下，使用随机过程被认为是抽样最合适的方法。

最直接的随机抽样类型称为简单随机抽样。但因这种简单的随机抽样方式不能保证各研究因素具有同等代表性，故有时该方式无法达到预期目标。生产企业在放行批次的研究设计中，可能包含所选时间、位置或平行的生产流水线（例如，多条灌装线）等影响因素；该情况下，可以使用分层抽样方式，即从每个因素中随机选择供试品。无论采用何种抽样方式，都应制定抽样计划以确保抽样能代表全部总体。

随机化不仅仅局限于抽样；在分析所研究的样本时，也应该统筹使用随机化方式，同时使用区组设计来避免研究目标与其他相关分析因素的混淆。

恰当的抽样和分析检验策略取决于生产、测量方式以及研究相关过程的知识。在对某一生产批次的检测抽样中，可能会包括一些随机选择的基本要素，这时，应收集足够的样本用于原始分析、随后的验证分析和其他支持分析。如果要对更复杂的研究进行抽样，则应通过一些设计策略来解决样本代表性问题；建议技术人员和统计人员合作，以协助选出针对特殊研究目标的最佳抽样计划和实验设计。

3.3.2 样本量

在制定研究计划时，需要考虑使用多大的样本量方可获得稳健的估计值，假定要获得总体均值，实验者从研究总体中随机抽取了 n 个样本，获得了样本均值的计算值，并使用该均值的置信区间半宽$t_{1-\frac{\alpha}{2};n-1} \times \frac{S}{\sqrt{n}}$(也称为误差范围，该处 S 为样本标准差)，表示该样本均值的（扩展）不确定度。可以定义误差范围不得大于最大允许值 H；选择置信水平$(1-\alpha)$，并对标准差(S)提供初步估计，即可使用方程求解所需的样本量大小：

$$t_{1-\frac{\alpha}{2};n-1} \times \frac{S}{\sqrt{n}} \leq H$$

$$n \geq \frac{t^2_{1-\frac{\alpha}{2};n-1} \times S^2}{H^2} \qquad (23)$$

由于 *t* 值的自由度是样本量 *n* 的函数，因此必须选择通过迭代求解公式 23。该公式也可通过其对应的 *z* 值近似替换 *t* 值来获得。对标准差 *S* 的初步估计值可使用类似研究设计，或通过相关专家的建议，在获得标准差的初步估计值或规定 *H* 值之前，应先确定数据的运算方式 (例如，转换或原始数据)。关于数据转换的内容，可参考“4.4.3 数据转换”。

3.4 不确定度

不确定度表达与一个值相关的误差大小，用于表达检测结果与真值的接近程度。该值考虑了测量或试验过程相关的系统误差和随机误差。任何研究都旨在减少不确定度，以便为药品质量做出更可靠的决策。在方法建立、确证和测量不确定度评估中，选择哪些因素为不确定度组分（即，测量不确定度的潜在来源）是至关重要的。

测量不确定度有多种来源，包括仪器、数学算法和分析人员的差异。表 7 列举了一些典型的不确定度来源。有关不确定度的更详细内容，见中国药典不确定度评定指导原则

表 7 分析实验室操作中不确定性成分的示例

分析方法设计引入的变异性	♦ 样品量或体积的影响 ♦ 被测值单元之间的变异 ♦ 关键标准物质的纯度 ♦ 样品储存条件的影响 ♦ 未能识别方法的耐用性因素
测量过程中引入的变异性	♦ 自动采样器的残留效应 ♦ 称量过程中的静电效应 ♦ 分析物回收不完全 ♦ 样品基质对校准斜率的影响 ♦ 样品温度对体积的影响

	♦ 空白校正的影响
分析人员引入的变异性	♦ 手动峰积分的影响
算法引入的变异性	♦ 线性校准强制过零点 ♦ 使用加权或未加权算法的线性拟合
样品引入的变异性	♦ 从实验样品取样的影响

3.5 统计学分析基本假设

统计学假设应依据基础数据的生成过程加以论证，并使用适当的图形或统计工具加以验证。如果这些假设中有一个或多个假设不成立（be violated），则可能需要改用其他方式进行数据评估。本指导原则中引用的大多数统计量和假设检验均基于以下设定：测量结果所属总体符合正态分布、结果独立、并且没有异常值（详见“4.4.5 异常值”）。有关统计假设的评估详见“4.4 模型和数据考量”。

4. 常用统计方法

4.1 控制图

控制图是对总体过程进行研究的简便工具，它需要仔细地制定目标，策划设计，实施计划以合理分析等。控制图在制药行业中用于监控生产过程和分析方法的性能。

在产品或方法整个生命周期过程中可能会受到已知变异或无法预料的变异性的影响。对于产品生产过程，这种变异可能会影响产品质量或提示需要采取行动（提高警惕）措施。特别是常用于辅助决策的分析方法，这些变异可能会增加得出错误结论的风险。因此，不断地确认性能并持续保证处于控制状态是很重要的。为此，需要收集并分析来自生产过程或分析方法的相关数据。对于生产过程，这些数据可能包括原料的过程参数和检测结果。对于分析方法，这些数据可包括质控样品分析结果，分析过程中使用的标准品结果以及系统适用性数据。值得注意的是，对照样品用于监控方法的性能，而不是产品的性能或属性指标。

尽管存在各种趋势分析方法，但控制图是用于趋势分析的最简单，最有效的图形工具之一。控制图有很多类型，具体见表 8。

表 8 控制图的介绍

控制图名称	适用数据类型	应用场景
单值-移动极差图（*I-MR*）	计量类数据	用于监测一段时间内，单个值及其移动极差的趋势变化
均值-极差图（$\bar{X}-R$）	计量类数据	用于监测一段时间内，各子组样本量较少时的均值和极差随时间的趋势变化
均值-标准差图（$\bar{X}-S$）	计量类数据	用于监测一段时间内，各子组样本量不少于 3 个时的均值和标准差随时间的趋势变化
指数加权移动平均值图（*EWMA*）	计量类数据	用于监测一段时间内过程均值稍有漂移的变化
累积和图（*CUSUM*）	计量类数据	用于监测一段时间内每个样本值与目标值之间偏差

典型的控制图由中心线、控制上限和下限组成。中心线代表过程中测得的变量分布的中心。如果过程按预期执行，几乎所有结果都将落在两个确定的控制限之内。超出（控制）界限的观察值或界限内的非随机性的显示点表明存在潜在的系统性能问题。历史数据（“控制数据”）通常用于获取中心线以及控制上下限。控制图对于识别潜在性能问题的漂移、趋势和变异性提供了一种可视化手段。

4.1.1 单值控制图

单值控制图是针对单个观测结果制定控制图，通常将上下控制限设置为：

过程平均值±3×过程标准差　　（24）

上下控制限基于以下假设：过程数据遵循正态概率分布，并且均值的 3 倍标准差范围包含大约 99.7%的总体数据。给定受控过程的样本观测值为 $X_1, X_2, ..., X_n$，使用以下公式估算过程平均值：

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad (25)$$

可以通过两种方法估算标准差，但是对于 *I*-控制图，最佳计算方式是基于移动极差（*MR*）统计。该计算方式考虑到过程的“短期”变异性，并防止数据中存在意外趋势的情况下产生过宽控制限。具体而言，*MR* 代表连续观测值的平均偏差，定义为：

$$\overline{MR} = \frac{\sum_{i=2}^{n} |X_i - X_{i-1}|}{n-1} \quad (26)$$

而过程标准差的估计值是：

$$\sigma_{ST} = \frac{\overline{MR}}{d_2} \quad (27)$$

其中 d_2 是一个常数，取决于与移动极差计算相关的观察次数。极差基于两相邻的观测值。即 $n=2$ 时对应的 d_2 值为 1.1280。*I* 图的控制上限（*UCL*）和控制下限（*LCL*）为：

$$LCL = \bar{Y} - 3 \times \frac{\overline{MR}}{d_2} \quad (28)$$

$$UCL = \bar{Y} + 3 \times \frac{\overline{MR}}{d_2} \quad (29)$$

不同 n 下，d_2 值的大小见表：

表 9 不同 n 下的 d_2 值

n	2	3	4	5	6	7	8	9	10	15	20
d_2	1.128	1.693	2.059	2.326	2.534	2.704	2.847	2.970	3.078	3.472	3.735

4.1.2 检测失控的结果

创建控制图后，可以使用 Nelson 规则检测失控结果。表 A-4 中提供了详细的 Nelson 规则。这些规则的相关性取决于控制图的类型。可以将所有 8 个规则应用于 I-图，并且特定规则的选择取决于控制过程的期望灵敏度。

表 10 用于检测失控结果的 Nelson 规则

规则	描述	图表中的指示
1	1 个点超出了 LCL 或 UCL	一个点失控
2	连续 9 个点在中心线同一侧	均值发生性能变化
3	连续 6 个点持续稳定增加或减少	存在趋势
4	连续 14 点上下交替	相邻点之间存在负相关
5	3 个点中的 2 个在均值的同一侧且距离均值大于 2 倍标准差的距离	分析变异性可能增加

6	5 个点中的 4 个点在平均值的同侧，且离平均值有 1 倍标准差的距离	分析变异性可能增加
7	连续 15 个点在平均值的 1 倍标准差内	分析变异性可能降低
8	连续 8 个点在均值两侧，且没有任何点在均值的 1 倍标准差范围内	非随机样本

4.2 过程能力

4.2.1 过程能力分析

过程能力是指过程满足产品质量标准要求（规格范围等）的程度，或指工艺在一定时间里，处于控制状态下的实际加工能力。而过程能力分析是将受控过程得到输出（结果）的分布与其质量标准的限度进行比较，从而判定其输出（结果）是否能保持与质量标准相符合的一种分析方法。进行过程能力分析的前提是需要确保过程稳定，可使用控制图（详见“4.1 控制图”）来确定过程是否稳定或可控。稳定的过程是可以预测的，它有助于对未来的过程性能做出判断并改进其能力。如果过程不稳定，那么评估过程能力没有意义。当评价产品是否满足规格限（质量标准）要求或消费者期望时，常使用过程能力指数进行过程能力评估。

4.2.2 过程能力指数

过程能力指数(Process capability indices, PCI)是衡量过程能否满足规范要求的能力评价指标。最常见的指标是用规格限宽度与过程分布宽度的比值来进行评价。

过程能力指数的计算首先要满足两个基本条件：一个是过程处于受控状态；另一个是所需分析的变量数据符合正态分布。当数据符合对数正态时，应首先将数据经过对数转换，然后在对数下进行过程能力指数的计算。

过程能力指数的计算分两类情况。当仅强调使用过程内在固有的随机因素评价其能力时，所用的过程能力指数记为 C_p、C_{pk}、C_{pu}、C_{pl} 和 C_{pm}，这些指标被称为短期能力指数，这时，一般用σ_{ST}表示其标准差，其大小用控制图制作原理中的$\bar{R} / d_2$（对于子组样本数大于 1 的过程）或$\overline{MR} / d_2$（对于子组样本数等于 1 或不强调分组的过程）来估计。当从过程总波动的角度来考察过程能力时，其总波动需同时考虑随机因素和特殊因素两者的共同影响，其标准差一般记为σ_{LT}，用

样本标准差的统计学定义公式来估计。这时的过程能力指数改记为 P_p、P_{pk}、P_{pu}、P_{pl} 和 C_{pm}，也称长期能力指数。

下面首先以短期能力指数的计算为例，介绍正态分布下的过程能力指数计算过程。

（1）*Cp* 指数

Cp 指数是指过程满足技术要求的能力，用规格限宽度与过程特征分布宽度之比计算其大小。*Cp* 指数假设过程是居中的，它不包含平均值估计的过程中心相对于规格限的信息，因此 *Cp* 指数也被称之为潜在过程能力。公式如下：

$$C_p = \frac{USL-LSL}{6\sigma_{ST}} \quad (30)$$

其中，*USL* 为规格上限，*LSL* 为规格下限，σ_{ST}为受控过程的内在固有标准差（$\bar{R} / d_2$或$\overline{MR} / d_2$）。当 *Cp<1* 时，过程能力宽于规格限（质量标准），详见图 B-1-Ⅰ；当 *Cp=1* 时，过程能力等于规格限（质量标准），详见图 B-1-Ⅱ；当 *Cp>1* 时，过程能力窄于规格限（质量标准），详见图 B-1-Ⅲ。

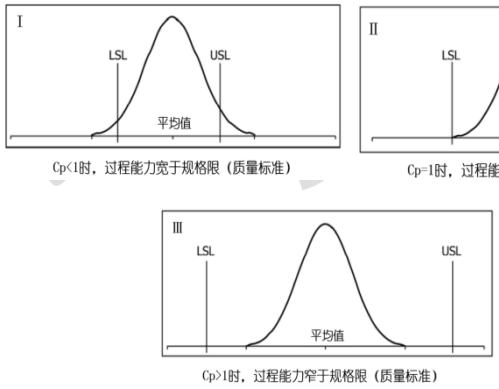

图 5 Cp 大小与规格限（质量标准）的比较

（2）*Cpk* 指数

C_{pk} 是指过程的最小能力指数，计算公式为：

$$C_{pk} = min\left[C_{pl}, C_{pu}\right] \quad (31)$$

其中：

$$C_{pl} = \frac{\bar{X}-LSL}{3\sigma_{ST}} \quad (32)$$

$$C_{pu} = \frac{USL-\bar{X}}{3\sigma_{ST}} \quad (33)$$

C_{pl} 和 C_{pu} 是单侧过程能力指数，适用于杂质检查和缺陷项比例等具有单侧限值的情形。C_{pl} 用于评估过程符合下规格限的能力，C_{pu} 用于评估过程符合上规格限的能力。

C_{pk} 较 C_p 而言，增加对数据集中趋势这一因素影响的考虑。从公式上看是过程的均值，从实际意义上看，相当于是过程目标，C_{pk} 较 C_p 之间的关联特性如下表所示：

表 11 *Cp* 和 *Cpk* 之间的关联

关联	意义
Cp=Cpk	说明过程以目标为中心（图Ⅰ）
Cp>Cpk	说明过程不以目标为中心（图Ⅱ）
Cp>1 且 *Cpk>1*	说明该过程是有能力的并在规格限（质量标准）范围内执行
Cp＞1 且 *Cpk＜1*	说明过程是有能力的，但不是以规格限（质量标准）为中心，也没有在质量标准内执行（图Ⅲ和Ⅳ）
Cp<1 且 *Cpk<1*	说明该过程是没有能力并超出质量范围内执行

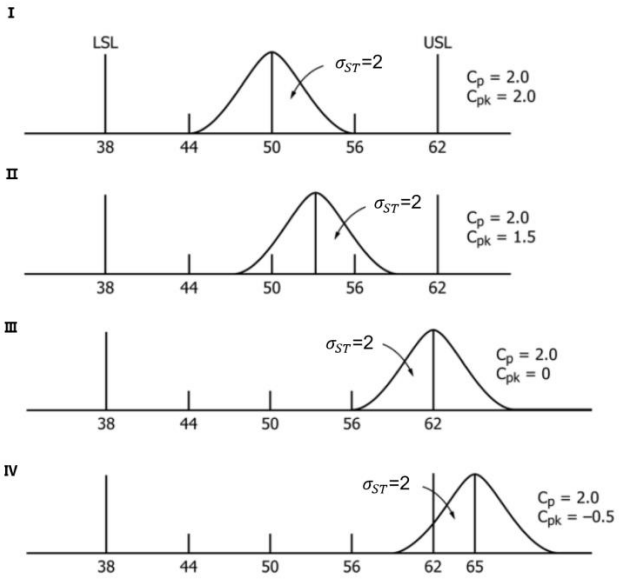

图 6 Cp 与 Cpk 大小的比较

一般当 C_{pk}=1 时，认为过程能力一般，C_{pk}≥1.33 认为过程能力达到要求，这

时，有 99.994%的概率大小满足规格要求。另外，从经济和质量两方面的要求来看，过程能力指数值并非越大越好，而应在一个适当的范围内取值。

（3）C_{pm} 指数

当强调质量特性中心偏离目标值（T）所导致的质量损失（相对偏倚）时，可以使用C_{pm}指数对过程能力进行度量。C_{pm}指数计算如下：

$$C_{pm}=\frac{\mathrm{USL}-\mathrm{LSL}}{6\sqrt{\sigma_{ST}^2+(\bar{X}-T)^2}} \quad (34)$$

这里，σ_{ST}为受控过程的内在标准差；$\bar{X}$是过程均值；T是目标值。

若过程均值等于目标值，那么 C_{pm}=C_p。随着均值偏离目标值和/或标准差的增加，C_{pm} 值将下降。

C_{pm} 在实际应用中的价值更大，特别是在药品领域中有时会出现目标值与质量标准上下限不完全等距的情况，即质量标准的目标值（T）与过程输出数据的均值（$\bar{X}$）存在偏倚时（$T\neq\bar{X}$），会导致对过程质量损失的忽视。而 C_{pm} 可有效避免此类情况的发生，但计算 C_{pm} 时，要求准确给出目标值 T。

上述是对短期过程能力指数的计算。在实际中，有时更强调从过程总波动的角度来考察过程能力，其过程标准差用σ_{LT}表示。其计算公式为：$\sigma_{LT}=\sqrt{\frac{\sum_1^n(x_i-\bar{x})^2}{n-1}}$，这时的过程能力指数记为 P_p、P_{pk}、P_{pu}、P_{pl} 和 C_{pm}，也称长期能力指数。

4.3 等效性与非劣效检验

4.3.1 等效性检验

在经典统计的假设检验中，有两个假设，即原假设和备择假设。当欲证明两个方法的均值等效时，就有必要把等效性的主张放在备择假设中。这种备择假设设定方式的统计学检验被称为等效性检验。需要理解“等效”并不意味着“相等”。等效应理解为所用新方法具有“充分相似性”。所谓“充分相似”是根据科学考量而先验决定的，并成为备择假设的基础。

假设先验知识认为两种方法的均值相差不超过某个正值 d 时，认定为等效（该值通常称为等效区间）。那么，等效性检验的假设为：

H_0: $|\mu_N-\mu_O|\geq d$

H_a: $|\mu_N - \mu_O| < d$ （35）

请注意，备择假设实际上是两个单独的单侧假设：

H_{a1}: $\mu_N - \mu_O < d$ 和

H_{a2}: $\mu_N - \mu_O > -d$ （36）

此时，该检验方法被称为双单侧检验（*TOST*）。作为单侧检验，每侧具有 α 值（通常但不一定是 0.05）的 I 型错误率。如果 100（1−2α）%的双侧置信区间（通常但不一定是 90%）完全包含在范围（$-d$，$+d$）内，则 TOST 拒绝原假设而支持备择假设。当原假设被拒绝时，得出两个方法在均值上是等效的结论。

实验室有时会想让新方法的变异度等效或优于原方法，这需要使用单侧检验。如果新方法的变异度较小，是可以接受的。需要确保的是新方法的变异度不显著大于原方法的变异度。此时，方法变异度的比较采用单侧非劣效性检验。

类似于均值的等效性检验，方法变异度的非劣效性检验也将所期望结果放在备选假设中（即将新方法的变异度非劣效于原方法作为备择假设）。由于标准差的统计特性，对其相比较的恰当参数是比率值，这里，用 σ_N 和 σ_O 分别代表新老方法的标准差。

4.3.2 非劣性检验

假定由先验知识确定出，新方法的标准差与原方法的标准差的比率值不超过系数 k（$k \geq 1$）时，方法即可使用。这时，系数 k 称为非劣效性余量。与非劣效性检验相关的假设是：

H_0: $\frac{\sigma_N}{\sigma_O} \geq k$

H_a: $\frac{\sigma_N}{\sigma_O} < k$ （37）

与均值的等效性检验不同，非劣效性假设是一个可以用显著性水平（通常但不一定是 0.05）处理的单侧假设。在实施该检验过程中，如果最终获得$\frac{\sigma_N}{\sigma_O}$的 100（1−α）%置信上限小于 k，则拒绝原假设，而接受备择假设。否定了原假设，即可得出新方法的变异度不劣于原方法变异度的结论。

建立方法可比性的统计学方法，也可用于涉及两组比较的其他情况，例如，方法转移或替代比较中。把所期望得到支持的主张放到备择假设中，会得到一个

恰当的统计结论。

等效性检验的益处显而易见，但在某些情况下，人们可能无法收集足够的样本量来提供建立等效性检验所需的效能。在这种情况下，使用差异性检验可能是唯一的选择；然而，需要注意的是，未能拒绝原假设（两方法相等）并不是证明两方法均值相同的证据。这时，仍应该报告置信区间，以表达两种方法均值间的差异。

4.4 模型和数据考量

统计分析包括模型及与数据拟合模型的可靠性相关的假设。模型可以是简单的，如仅与报告值相关的均值模型；也可以是一般的线性回归模型，如日常所用试剂盒中的标准曲线，常见理化分析的直线回归等；也可以是复杂的，如在复杂的制药环境中常见的溶出实验和生物实验中的非线性混合效应模型。一般用模型拟合残差监控模型的假设是否成立，该类监控包括对残差的正态性，方差恒定性和独立性的分析。

4.4.1 模型

统计学中的模型是指可代表总体中某（些）属性的函数化描述。模型中的参数也称总体参数，是指该属性的真实但未知的值，这也是统计学所寻求的宗旨。

均值模型表征了单变量总体的中心，可以写成：

$$Y_i = \mu + E_i \qquad (38)$$

其中，Y_i 是总体中大小为 n 的样本中的第 i 个测量值，μ 是代表总体平均值的模型参数，E_i 是误差。E_i 代表所有因素的影响，这些因素解释了为什么测量值并不总是等于 μ 的原因。这些因素通常包括产品批次之间的差异或分析方法的变异。均值模型是与总体均值相关的，通常以样本均值估计：

$$\bar{Y} = \frac{\sum_{i=1}^{n} Yi}{n} \qquad (39)$$

误差可由残差 $R_i = Y_i - \bar{Y}$ 计算。

另一熟悉的模型是简单的线性回归模型。该模型描述了带有一些协变量 X_i（例如时间或剂量）的总体均值的线性趋势，可以写成：

$$Y_i = aX_i + b + E_i \qquad (40)$$

其中（X_i，Y_i）是来自双变量总体中样本大小为 n 的第 i 个观察值，参数 a 和 b 分别是斜率和截距，它们定义了函数关系，E_i 是误差。注意，公式 38 中 μ 已被替换为公式 40 中的 aX_i+b，即，将总体的平均值转换为 X_i 的函数。公式 40 中，参数 a 和 b 是由样本数据估计而来。

一些更复杂的模型如非线性模型，可以包括定性因素（例如，验证实验中的实验者）或随机协变量（例如，连同 X_i 的另一测量 Z_i），而不是固定值。

4.4.2 有效数字

许多机构将数据存储在实验室信息管理系统（LIMS）中。这些数据虽然会达到检测方法所要求报告值的有效数字（小数）位数，但不适合需后续分析的数据。

一般要求用于计算的数字位数和报告值中出现的数字位数应分开考虑。计算过程中记录和携带的位数一定要多于报告值的位数，最好使用尽可能多的数字来执行所有统计计算。“在可行的情况下，可使用原始数据在计算设备（软件）或表格中进行计算，并且只对最终结果进行舍入”。对中间计算结果的修约处理会导致最终结果产生总体误差。

报告值的有效数字位数与方法的精密度相关。一般生物活性方法的几何变异系数（GCV）在 2%~20%之间时，报告值的有效数字位数保留 2 位。

4.4.3 数据转换

数据转换是对测量（值）的函数化重新表达，其目的是为更好地表达已知的科学关系或满足统计学模型的假设需求。数据是否应该转换或应选择哪种转换方式，可根据一组数据使用残差图凭经验找出。对数转换方式是数据分析时较为常用的转换方式。

4.4.3.1 对数转换

许多生物反应系统经常利用测量系统科学知识进行数据转换，尤其是当均值模型预测的响应值变异与响应值成比例时。对于这些测量系统，对原始响应采用对数转换非常有用，可将该响应在整个响应范围内转换为具有几乎恒定的方差。

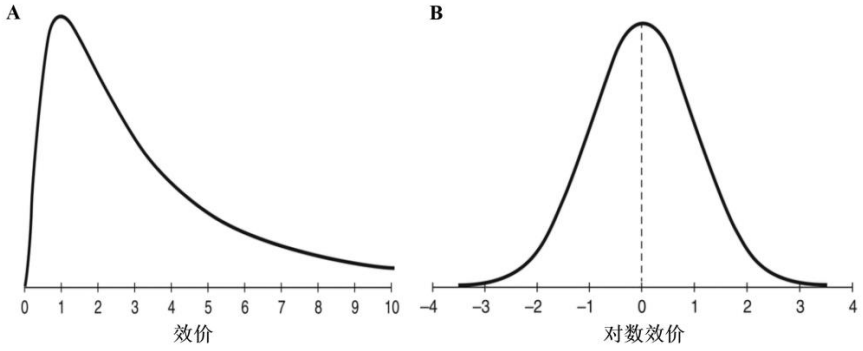

图 7 效价的偏态分布（左图 A）和效价对数转换后的正态分布（右图 B）

使用对数转换的另一个原因是，它可以将原本符合非线性函数形式的数据，转变为更简单容易使用的简便模型。如，可以使用对数变换将非线性一级动力学模型转换为线性模型。

在“1.1 统计量”中，描述了样本数据的集中和离散相关的统计量，包括样本均值，样本标准差等。这些参数只有当样本数据符合近似正态分布且没有异常值时，才有意义。如图 7 左侧中的分布呈右偏态，右侧尾部的较大值会影响平均值的估计。图 7 右图中的分布显示了响应值经对数转换后的分布。分布又称为对数正态分布，由于正态曲线的对称性，这时，所得样本均值和样本标准差作为转换后分布的集中和离散度才有代表意义。

对响应进行对数转换后，样本均值可以转换回原尺度。这种反转换产生了原尺度的几何均值（GM）。*Yi* 表示原尺度，使用 *Ti* 作为对 *Yi* 的对数转换值。则：

$$Ti=ln\ (Yi) \quad (41)$$

$$\bar{T} = \frac{\sum_{i=1}^{n} Ti}{n} \quad (42)$$

$$GM = exp(\bar{T}) = (\prod_{i=1}^{n} Y_i)^{1/n} \quad (43)$$

响应的对数转换标准差（S_T）同样可以反转换为 exp（S_T）。此术语称为几何标准差（*GSD*）。即，

$$GSD = exp\ (S_T) \quad (44)$$

由于 S_T 为非负值，因此 GSD≥1，表示原响应尺度的倍数变化。在表达方面，使用响应值的算术运算中，响应值的汇总计算表示为 $T \pm S_T$，但对于几何运算下的响应值计算，可以将其汇总计算表示为 GM×/÷GSD，或 GM / GSD 至 GM×GSD。例如，GSD＝1.25， GM＝1.0，则变异范围可以汇总为 1.0 / 1.25＝0.80 到 1.0×1.25

= 1.25。代表了 1 个标准差的波动范围。在对数转换状态下，还可以计算出一个更合适的范围。

几何变异系数表达为：

$$\%GCV = 100 \times (GSD - 1)\% \quad (45)$$

从对数正态分布状态下，通过计算该状态下的变异，可得到另一种计算原始数据状态下的%CV 方法：

$$\%CV = \sqrt{exp(S_T^2) - 1} \times 100\% \quad (46)$$

在数值上，当对数正态分布的%GCV 和%CV 均小于 20%时，两者非常接近。在报告对数正态的数据变异度或区间量值时，应明确指定将它们与所对应的 GSD 一起使用。

根据“2.1.2.1 置信区间”中的式（8），在对数状态下，100 (1-α)%均值双侧置信区间转化为：

$$C_L(T) = \bar{T} - t_{1-\alpha/2;n-1}\frac{S_T}{\sqrt{n}}$$

$$C_U(T) = \bar{T} + t_{1-\alpha/2;n-1}\frac{S_T}{\sqrt{n}} \quad (47)$$

可获得原始数据状态下几何均值的置信区间为：

$$C_L(Y) = exp(C_L(T))$$

$$C_U(Y) = exp(C_U(T)) \quad (48)$$

4.4.3.2 其他转换

对于其他类型的数据，可以考虑除对数以外的其他转换。例如，当所用响应值在 0 到 1 之间（或百分比在 0%~100%之间）时，可选用反正弦（arcsine）或 logit 转换。反正弦变换的公式为：

$$T = 2 \times sin^{-1}(\sqrt{Y}) \quad (49)$$

logit 转换的公式为：

$$T = ln\left(\frac{Y}{1-Y}\right) \quad (50)$$

当大多数数据位于上限 1.0 或下限 0.0 时，反正弦（arcsine）或 logit 转换特别有用。对于计数数据，转换方式可以选用平方根或计数数据的对数转换方式。

669 幂指数变换（最常见的是 Box-Cox 变换）也是一种常用的数据转换方式。该
670 类转换的形式为：

671 $$T=\begin{cases}\frac{Y^{\lambda}-1}{\lambda} & \lambda\neq 0\\ ln(Y) & \lambda=0\end{cases} \qquad (51)$$

672 其中 λ 是将数据集转换为正态分布的最佳转换值。

673 无论进行何种转换，在转换状态下计算的汇总统计量和区间都可以反转换为
674 原始状态下的量值。在任何情况下，都应检查数据以确定转换后的测量值是否显
675 示出几乎均匀的变异性并呈近似正态分布。

676 **4.4.4 评估模型充分性**

677 所有模型都涉及数据生成和数据自身两方面的过程假设。除假定的函数形式
678 外，公式 38 和公式 40 中残差项的分布也是非常重要的。典型的假设是残差项独
679 立、呈正态分布、且在响应范围内具有恒定的方差。当这些假设满足（合理性）
680 时，统计学模型通常易于解释且效能强大。尽管任何模型都具有其特性，但必须
681 检查模型所依赖的假设情况是否符合要求。评估模型充分性就是验证假设的过程。

682 模型充分性的评估有图形方法和定量方法。在许多数据分析项目中，研究人
683 员和专业统计人员在选定最终模型之前会进行多次讨论。讨论考虑的主题包括数
684 据的适当转换，所关注的处理和设计因素，潜在的候选模型以及模型拟合的评估。

685 用于评估模型拟合的工具包括带有原始残差和学生化残差的残差图，基于模
686 型的异常值检测方法以及回归杠杆和影响度量指标。残差图可以几种生成方式，
687 最常见的形式是在垂直轴上绘制残差，在水平轴上绘制预测响应。当残差图上的
688 观测值沿水平轴的展宽增加或减少时，表明违反了恒定方差的假设。残差的任何
689 线性或非线性趋势都表明模型选用的不恰当，或模型中缺少关键的影响因素。例
690 如，弯曲的残差图可能说明模型中需要增加二次项。另外，远离点集之外的残差
691 可能表示异常值。可以通过合适的数据转换来解决问题。

692 如果将模型用于预测未来行为，则误差项的正态性是特别重要的假设。可用
693 于监控此假设的图形方法包括箱线图、Q-Q 图等。许多常见的统计软件中均能实
694 现这些图形工具的使用。

695 当数据被特定“分组”时，通常会出现缺乏独立性的情况。例如，实验中从相
696 同的培养板上获得的测量值比在其他板上记录的测量值更相似。这种所谓的组内

相关性可以通过在模型中增加一个“分组”因子来说明相关性，从而更好地建模。

在评估模型假设时应注意统计检验易受到样本大小的影响。在样本量较小的情况下，统计检验可能对检测偏离模型假设的偏差不敏感。相反，对于大样本，即使目测评估表明假设是合理的，他们也可能检验到假设不成立。评价模型是否适用需要：（1）对所得数据的测量过程有科学理解，（2）图形分析，（3）统计检验三者结合在一起的方式。

4.4.5 异常值

有时，观测到的分析结果与预期的分析结果有很大差异，这样的结果被称为离群结果。对离群结果应进行记录、解释和处置。应注意，有些异常结果虽然与预期结果有很大差异，但的确是对被测特性的合理测量。而有些结果虽然处于所测属性的期望值范围内，却仍可能是分析系统中的错误所致。防止存在离群分析结果的首要方法（防线）是采取合适的系统适用性和控制规则（见“4.1 控制图”）。

当存在离群结果时，需进行系统的实验室和生产过程调查，以确定是否存在明确的异常原因。调查异常值时要考虑的因素包括人为错误、仪器错误、计算错误、以及产品或组件缺陷。对其彻底调查时，应考虑包括：方法的准确度和精密度、内控参考品和对照品、过程和分析趋势、以及质量限度标准。如果确定是由分析方法所产生的异常值，则可以对同一样本（如果适用）或其他类似样本进行重新检测，此时离群数据可依据调查结果宣布无效，并在后续计算中剔除。

“异常值标注”是对离群结果的非正式识别，其中，异常值标注最常用的方式是通过图形技术进行视觉判别，例如使用残差图、标准化残差图或箱线图。“异常值辨识”是使用统计显著性检验来确证某些值与已知或假定的数据分布不一致的方法。准确选择异常值辨识技术通常取决于对测定值的数量和位置的初始识别。除了上述标注法外，还有很多异常值检验方法。其中常见的是狄克森（Dixon）检验法和格拉布斯（Grubb’s）检验法，见中国药典通则 1431 生物检定统计法中“异常值剔除和缺项补足”部分。如前所述，使用狄克森检验或格拉布斯检验也要求数据符合正态性假设，若实验样本量小无法验证则需要依靠历史测量结果来支持这一假设。

在识别统计异常值的过程中，通常需要找出导致出现该异常值的原因。在分析实验室中，所用异常值检验方法通常用于单变量。选择该类方法时，要考虑：

（1）是否可以假定分布是正态分布，还是应用非正态分布的其他特定分布类型进行检验，（2）是否怀疑有多个异常值，以及哪些观测值应该被标记。异常值检验可以分类为参数（基于模型）或非参数检验。选择参数化结构的检验通常应是正态分布，即常见的狄克森检验法和格拉布斯检验法。考虑是否存在一个或多个标记的异常值，以及其相对位置（如大于或小于大部分测量值）。如果怀疑有不止 1 个异常值，可能需要多次使用异常值检验方法。但一般不建议对同一数据集采用多次异常值检验方法，而应该查找试验原因，重新采集试验数据。

4.5 贝叶斯推断统计

利用贝叶斯方法可以得出一个区间，该区间表示总体真值在概率为 $100\times(1-\alpha)\%$的范围内。频率论是基于统计量的频率推断，而贝叶斯推断则是基于总体参数的概率表达。总体参数是未知的，它出现在统计模型中（例如，均值、方差、均值差等）；而统计量是基于（所得样本）数据的汇总统计量或估值（例如，参数估值）。基于频率论的统计推断方法认为总体的参数值不可知但确定，而贝叶斯统计推断则认为总体参数具有不确定性，但可以通过概率分布进行估计。贝叶斯统计为基于风险的决策研究提供了一种新的观点和工具。贝叶斯推断还可以将样本数据和总体参数的先验信息整合，进一步提升对总体参数的估计。当样本量较小或者研究本身不能很好量化某些影响因素时，贝叶斯统计通过整合有价值的先验知识，可更好地支持决策判断的正确性。

4.5.1 参数不确定性与抽样变异性

参数（Parameter） 是未知的假定值或总体量值，如总体的平均值、标准差，或方法间的均值差。尽管这些参数未知，但是可以估计。参数值及其固有不确定性是贝叶斯方法的基础。

统计量（Statistics） 是从所关注总体或过程获取的样本中得到的观测量值或观测量值的汇总量值。统计量有：分析的结果（测量值）、样本均值、样本标准差、方法间获得的均值差、或其值的置信区间。当对总体进行重复抽样时，由于抽样变异性，统计量值会有所不同。

基于频率论的统计方法认为总体参数是不变的固定值。概率分布模型被用来对随机样本统计量的抽样变异性进行估计和统计推断。置信区间的计算就是一种

常见方法，通过构造 95％置信区间，可确保在重复随机取样时，总体参数有 95％的次数落入该区间。需要指出的是，这里的 95%描述的是方法学的可靠性（如其覆盖范围），而不是总体参数被单个区间包含的概率。

贝叶斯统计学方法认为参数值是不确定的（不是固定值），并使用概率分布对其可能的水平进行建模。它进一步拓展经典的频率论统计学方法，用概率的思想同时对样本统计量的抽样误差和总体参数的不确定性进行研究。贝叶斯模型有时被称为“完整”的概率模型，因为它同时考虑样本统计量的抽样误差，相关先验信息和观测数据对参数不确定性的估计。在频率论统计方法中必须预先确定概率水平（例如 95％），并且所得区间是随机的；而贝叶斯方法则与频率论者的区间计算方法不同，它对参数事先确定区间，并估计参数值在该区间范围内有多大概率。该方式对于判断一个分析方法产生超出给定范围的结果的概率是非常重要的。

4.5.2 先验分布和后验分布

频率论和贝叶斯推断都使用概率分布来表征模型。两者都使用相同的似然（likelihood）模型来解释抽样变异性。贝叶斯方法在选择似然比模型时，其独特之处是根据有关统计变异性的先验知识来确定。

贝叶斯推断还要求在针对观察数据之前，先建立参数不确定性的概率模型，称为先验分布。如同似然比（模型）一样，先验分布是一项基于先验数据、可靠知识、或常识的选择。贝叶斯方法要求确保所选的先验分布科学合理，并且不会过分影响推断。使用适当合理的先验分布知识可以有效地减少决策所需的样本量。但是，当没有足够的理论，历史数据或者专业知识的情况下，先验分布采取不倾向于任何取值的均匀分布，从而对后续的贝叶斯推断的影响也最小。这样的先验分布通常被称为“无信息的”先验分布。当使用该类分布时，贝叶斯推断通常与频率推断相一致，因为这时两者均依赖于似然。

贝叶斯在方法学上将似然和先验分布模型与观测数据结合起来，以产生描述参数不确定性的更新分布模型，更新后的模型即称为后验分布。后验分布提供了总体参数值在其区间的概率；这种区间称为可信区间。当采用某些类别的非信息性先验分布时（例如，正态似然法中使用的 Jeffrey 先验），可以从后验分布中计算贝叶斯可信区间，并且有时在数值上等于传统频率论法相应的置信区间。但是，

对于两种区间的解读是不同的。与可信区间相关联的概率，是给定现有观测数据情况下对总体参数不确定度的估计；而与置信区间相关联的概率，是多个样本的样本统计量对总体参数的覆盖概率。

从贝叶斯统计角度来看，有关目标总体参数的所有知识都基于后验分布。先前研究的后验分布可以为后续研究提供先验分布。随着新数据的获取，以这种方式更新先验分布，为知识构建提供了范例，从而为在药物开发过程中应用先验知识提供了统计基础。

参数的后验分布也可以与未来观察到的数据或统计量重新组合，获得后验预测分布的似然比。与后验分布一样，贝叶斯的观点是将所有关于（参数）预测值所有的知识建立在该后验预测分布的基础之上，该后验分布可用于构造与概率分布类似的贝叶斯容忍区间和预测区间。但与概率分布的区间不同，贝叶斯区间不需要预先设定概率水平。例如，后验预测分布可用于估计未来超出规格结果的发生概率（*OOS%*）。

4.5.4 频率论和贝叶斯方法的比较

频率论和贝叶斯方法对统计推断都非常有价值。频率论分析法应用广泛，操作简便，可提供已知参数覆盖概率下的可靠性。贝叶斯方法可以量化参数的不确定度，从而支持基于定量风险的决策。虽然应用贝叶斯-蒙特卡洛模拟技术上更具挑战性，但它通常可以解决频率论分析法难以解决的问题。如果可以明确先验信息的合理性（正确性），那么与频率论方法比贝叶斯方法可使用更小的样本量来制定决策。表 12 提供了两者的特性比较。

表 12 频率推断和贝叶斯推断间的特性差异

特性	频率推断法	贝叶斯推断法
统计量	概率建模的抽样变异性	
参数	视为固定和未知的	概率模型的不确定性
覆盖（包含）率	（通常）理论上已知	（通常）需要通过计算机模拟确定
先验信息	通过抽样变异性模型引入	通过抽样变异性模型和参数的先验分布引入

估值类型	点估计值和区间估计值	后验分布（可从中得出点和区间估计值）
观测值	视为假设的一系列重复样本的实现	当作推论依据的固定值
参数推断	参数值基于重复采样覆盖概率的固定概率	参数值由后验分布概率量化
统计学设计的影响	可能影响重复抽样的覆盖概率	不重要
多重比较	可影响重复抽样的覆盖概率	不重要
风险评估	间接风险评估	适用于量化风险的估计概率
参数值的先验知识	从推断中排除	需要模型参数的先验分布
持续的知识构建	历史研究的非正式评估	历史研究的后验分布为今后研究提供了先验分布
未来观测值的预测	基于容忍或预测区间的间接推断	基于后验预测分布概率的直接推断
软件	日常中广泛可及	需要专业统计软件

起草单位：中国食品药品检定研究院。
参与单位：山东省食品药品检验研究院、浙江省食品药品检验研究院、江苏省食品药品监督检验研究院、陕西省食品药品监督检验研究院、山西省检验检测中心、沈阳药科大学等
主要起草人及联系电话：谭德讲，010-53851589；陈华，010-53851622

分析数据的解释与处理指导原则增订说明

一、目的和意义

分析数据的科学解释和/或评价直接关系到产品质量的最终决策是否科学可靠。无论是对检测产品的方法评估，还是科学地使用所建方法去评价产品，都需要对所获得数据进行解释和处理，以便做出科学的决策。因此，分析数据的解释与处理对药品研发、生产和检验检测，都是不可或缺的一环，是对药品全生命周期中质量保障的基础或支柱内容。

目前《中国药典》各论中已经引入了较多的统计术语和统计计算方法/原理，需要制定一个专门涵盖药典中所需统计分析方法的通用性指导原则以供规范应用。建立分析数据的解释与处理指导原则，可以指导药品领域选择科学可靠的统计方法，给出如何利用统计分析方法对药品生产工艺和产品进行规范地决策；也为使用数据评估药品质量提供一个科学、可接受的监管方向。

二、主要内容

本指导原则参考国内外相关标准进行构建，为药典中使用的统计提供基本原理、数据分析和处理方法。主要内容包括前言、描述性统计、推断性统计、基本统计原理、常用统计方法等。