生物统计基础知识(1)

刘来福

(北京师范大学数学系)

编者按：为了适应现代生物学研究中试验设计和数据处理的需要，我们组织了《生物统计基础知识》专题讲座，拟分四期刊载：一.平均数和标准差、二.两个样本平均数的比较、三.回归分析、四.方差分析。读后有何意见和要求可写信给我刊。

生物统计是从统计规律上来分析和解释生物界各种现象的一门科学。它根据统计学的原理在一定程度上消除原始数据中由随机因素所产生的误差，以进一步揭示有关的各因素间内在的统计规律性。因此在探索生物界各种规律的研究工作中，生物统计也应该是必须要掌握的工具之一。本文简单介绍生物统计中一些常用的方法，以利于初学者初步掌握这个工具。

一、平均数和标准误差

（一）平均数和标准差

生物学的观测资料都是来自所研究总体中的一个样本。然而，研究的目的却在于通过样本而分析总体的情况。总体中所包含的每一个个体的有关属性都是以随机的形式表现出来的。因此一般并不要求掌握其中每一个个体的有关属性，只需要掌握描述总体的某些重要特征的几个参数就够了。最常用的参数有两类：一类是总体中心值的代表数，它反映了整个随机变量分布的中心位置，集中了所有个体所提供的信息。可以做为这个总体的代表参加分析。借此可与另一个总体进行比较。例如：猪的“平均体重”。对一个有限总体来说，常用的集中性参数就是算术平均数（用μ来表示）。一般用样本的算术平均数$\bar{x}$作为μ的估计值。

所谓算术平均数是指观测数据的总和被数据的个数来除所得的商。n个数据$x_1, x_2, \cdots, x_n$的算术平均数可用$\bar{x}=\frac{1}{n}(x_1+x_2+\cdots+x_n)$式来计算，这个式子还可以简写为

$$\bar{x}=\frac{1}{n}X_{\cdot\cdot}=\frac{1}{n}\sum_{i=1}^{n}x_i。$$

式中$X_{\cdot\cdot}=\sum_{i=1}^{n}x_i$表示$x_1, \cdots, x_n$这$n$个数的总和，$\Sigma$是表示总和的符号，$\sum_{i=1}^{n}$表示观测资料$x_i$（足标从1逐个增加到$n$）这$n$个数据的总和。在不会引起误解的情形下$\sum_{i=1}^{n}x_i$也可以简记作$\Sigma x_i$。

另一类参数将反映总体分布的离散程度，或者说是整个数据变异程度的指标。例如，有三组观测数据，分别是：

8，9，10，10，10，11，12；

6，7，9，10，11，13，14；

1，4，7，10，13，16，19。

三组数的平均数都等于10。但这三组数据的离散情况却不一样。因此平均数只是一个集中数，它反映不出数据的离散状态。另外从这三组数据还可看出，虽然这三组的平均数相同，但这个平均数对这三组数据的代表程度却不一样。第一组所有的数都比较接近于平均数，平均数的代表性就较强，而第三组数据非常分散，平均数的代表性也就较差。由此可见，数据越离散，平均数的代表性就越差。因此在描述一个总体时，除了有代表总体的集中性参数（μ）之外，还需要有代表总体分布离散程度的参数。

离散程度用方差σ^2或其平方根——标准离差（简称标准差）σ来表示。

对一个有限总体来说，所谓方差，指的是各数据与其平均数离差平方和的平均值。如果总体有n个变量，则方差为：

$$\sigma^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\cdots\cdots\cdots(1)$$

标准离差为$\sigma=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2}$。

我们通常得到的是样本资料，需要用样本平均数$\bar{x}$作为μ的估计值。代入（1）式可以得到$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$。由于

$$\frac{1}{n}\Sigma(x_i-\bar{x})^2=\frac{1}{n}\Sigma[(x_i-\mu)-(\bar{x}-\mu)]^2$$

$$=\frac{1}{n}\Sigma(x_i-\mu)^2-(\bar{x}-\mu)^2$$

$$\leqslant\frac{1}{n}\Sigma(x_i-\mu)^2=\sigma^2。$$

因此用$\frac{1}{n}\Sigma(x_i-\bar{x})^2$作为$\sigma^2$的估计值时总要偏低。对于样本观测资料，将用：

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\cdots\cdots\cdots(2)$$

作为σ^2的无偏估值，称为样本方差，其平方根$s=\sqrt{s^2}$称为样本标准差。式（2）的分子是离均差的平方总和，简称平方和，分母称为样本资料的自由度。

例如前三组数据做为样本，计算其样本标准差，第一组为：

$$s_1^2=\frac{1}{9-1}[(8-10)^2+(9-10)^2+\cdots$$

$$+(11-10)^2+(12-10)^2]$$

$$= \frac{1}{8}[4+1+1+4] = 1.25,$$

$$s_1 = \sqrt{1.25} = 1.12。$$

第二组为 $s_2 = 2.74$，第三组为 $s_3 = 6.48$。

为了计算方便和减少计算误差，采用下列公式来计算平方和。

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2$$

$$= \sum_{i=1}^{n}x_i^2 - \frac{1}{n}X^2_{\cdot}。$$

结果与前式一致[1]。例如对第一组数据有

$$\sum_{i=1}^{n}x_i^2 - \frac{1}{n}X^2_{\cdot} = (8^2+9^2+\cdots+11^2+12^2)$$

$$-\frac{1}{9}\times 90^2 = 910 - \frac{1}{9}\times 8100 = 10。$$

标准差是随机总体变异程度大小的一个指标。它的大小反映出了平均数代表整个总体的可靠程度。倘若没有变异，则 $\sigma = 0$，平均数就完全可以做为总体有关属性的代表。随着 σ 的增加，平均数对总体的代表性也就逐渐减少。

不同属性之间的标准差是不能用来对比的。例如，我们测定了某种鼠的体重，得出平均数为 $\bar{x}_1 = 53.42$ 克，标准差是 $s_1 = 13.59$ 克。又测定了它们的耗氧量，得出平均数为 $\bar{x}_2 = 4.84$ 毫升/小时/克，标准差为 $s_2 = 0.88$ 毫升/小时/克。这时我们无法判断这些鼠的体重与耗氧量哪一个属性的离散程度更大。至少体重与耗氧量的单位都不同，因此这两个数字不能放在一起比较。为了比较单位不同或者平均数大小相差悬殊的资料的变异程度，把公共的标准取为以平均数为单位的标准差，也就是取标准差与平均数的比值（百分数）当作标准（它没有单位了），称为变异系数，简称 $C.V.$，公式为：

$$C.V. = \frac{s}{\bar{x}} \times 100。$$

可以算出上面鼠的体重与耗氧量的变异系数分别为：

$$C.V._1 = \frac{s_1}{\bar{x}} \times 100 = \frac{13.59}{53.42} \times 100$$

$$= 25.44(\%),$$

$$C.V._2 = \frac{s_2}{\bar{x}} \times 100 = \frac{0.88}{4.84} \times 100$$

$$= 18.18(\%)。$$

可见体重的相对变异要大于耗氧量的相对变异。

（二）正态分布

生物的属性往往是作为随机现象出现的。如在相同的管理条件下，同一品种的猪 20 天体重增加的斤数。虽然某一头猪的增重事先我们不能估计，但是在同一随机现象大量出现时，它们也是有一定的统计规律可循。

观测了 100 头猪的 20 天增重量。其增重量在组距为 6 斤的数值范围内出现的频数列成如下的频数和频率（频数/总数）分布表（表 1）。

表 1　100 头猪 20 天增重（斤）频数、频率分布表

组（斤）	3—8.99	9—14.99	15—20.99	21—26.99	27—32.99	33—38.99	39—44.99	45—50.99	51—56.99
频数	4	5	16	23	25	18	6	2	1
频率	0.04	0.05	0.16	0.23	0.25	0.18	0.06	0.02	0.01
频率/组距	0.0067	0.0087	0.0267	0.0383	0.0417	0.0300	0.0100	0.0033	0.0017

如果在直角坐标系上，以随机变量可能取的数值作为横坐标，以频率/组距作为纵坐标，就可以将表 1 画成宽为 6 的一组矩形图（见图 1）。

图中每个矩形的面积恰好等于数据落在该矩形所对应的组内的频率。因此所有矩形面积的总和应该等于 1。另外还可以看出这个图的形状大致是中间高，两边低，左右对称的形状。

如果观测猪的头数不断增多，组距越来越小，分组也越来越多时，频率直方图就会越来越

1）近年来随着计算器的普及，目前市场上出售的多数计算器都带有简单的统计计算键。使用时只要将数据逐个送入机内，再按有关的键，就可以立刻得出 $\bar{x}$，$\sum x$，$\sum x^2$，σ，s 的数值，非常方便。

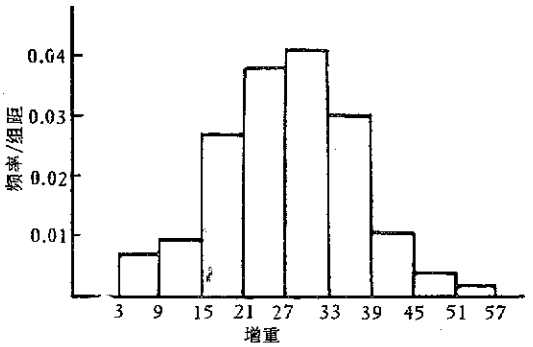

图 1　频率直方图形

对称，最后将呈图 2 那样的图形。这样的曲线称为频率分布曲线。在某一区间上频率分布曲线下的面积就表示了随机变量出现在这个范围内的概率。

随机变量如图 2 所示的分布规律称之为正

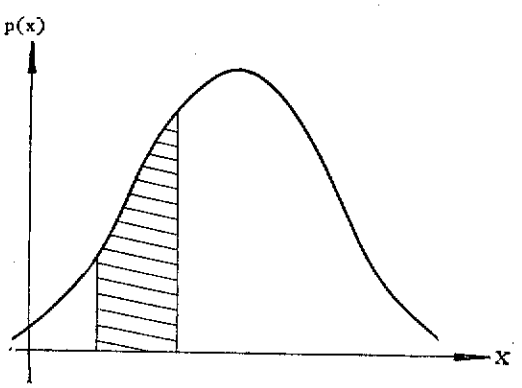

图 2　频率分布曲线

态分布。许多生物学的现象所产生的数据都是符合正态或者近似正态分布的形式。正态分布在统计学的理论及应用中占有着重要的地位，它构成了许多统计理论的基础。本文所介绍的生物统计的方法大都是对正态分布的总体来讨论的。

理论上的分析表明，正态分布曲线仅依赖于两个参数 μ, σ。随着 μ 和 σ 的不同，曲线在坐标系中的位置以及形状（高、矮、胖、瘦）均不同（如图 3 所示）。但是无论 μ, σ 取什么值正态分布曲线都有下述共同的性质：

1. 正态随机变量在整个数轴上取值。其频率分布曲线无论向左、向右延伸，都愈来愈接近横轴，并以横轴为渐近线。它与横轴所围成的图形的总面积为 1。

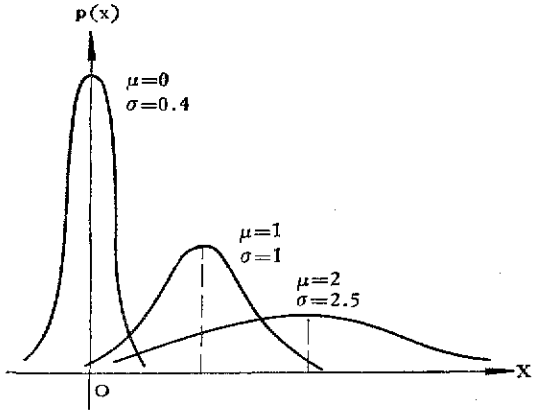

图 3　不同 μ, σ 的正态曲线

2. 当随机变量 x 取平均值 μ 时，曲线处于最高点。当 x 由此向左、向右远离时，曲线不断地降低，呈现中间高两边低的形状。并且曲线关于轴 $x=\mu$ 是对称的。

3. 无论 μ 和 σ 取什么值，也就是说无论曲线是什么形状（高、矮、胖、瘦），当以标准差 σ 做为度量单位时，以 μ 为起点的相同区间上曲线所围成的面积都是相等的。例如曲线在 $\mu\pm\sigma$ 这个区间上所围成的面积总是 0.6827。也就是说，随机变量出现在范围 $|x-\mu|<\sigma$

$$\left(\text{或者 } \frac{|x-\mu|}{\sigma}<1\right)$$

内的概率是 0.6827。同样在 $\frac{|x-\mu|}{\sigma}<2$ 内的概率是 0.9545，而在 $\frac{|x-\mu|}{\sigma}<3$ 内的概率是 0.9973。这样，利用这个性质就造出了正态分布曲线下一定区间的面积表（表 2）。表中 W 栏表示 $W=\frac{x-\mu}{\sigma}$ 的值，它是以 μ 为起点，以 σ 为单位所度量的区间长度。表中的数字表示 x 出现在区间 $[\mu, \mu+W\sigma]$ 内

$$\left(\text{即使得 } 0<\frac{x-\mu}{\sigma}<W\right)$$

的概率的大小。由曲线的对称性可知，它刚好是在区间 $(\mu-W\sigma, \mu+W\sigma)$ 上正态曲线所围成的面积的一半。

利用正态分布的知识可以对一些已知正态总体中的随机变量进行估计。

表 2　正态曲线下一定区间的面积（节录）

W	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
⋮										
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4430	0.4441
1.6	0.4452	0.4463	0.4474	0.4485	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4700	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4762	0.4767
2.0	0.4773	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
⋮										

例如，已知某种鼠的体重呈正态分布，其平均体重 $\mu = 60$ 克，标准差 $\sigma = 9.5$ 克，若了解

（1）这种鼠的体重在 60±15 克之间的概率有多大？

（2）这种鼠的体重 95% 在什么范围之内？

则：（1）我们需要求的是鼠的体重 x 在区间 $60-15<x<60+15$ 即 $|x-60|<15$ 的概率，也就是说，要求 $\frac{|x-60|}{9.5}<\frac{15}{9.5}=1.58$ 的概率。查正态曲线面积表，当 $W=1.58$ 时得 $P_1=0.4430$，这仅仅是半个区间 $60<x<60+15$ 上正态曲线的面积。利用对称性可知 x 在区间 $(60-15, 60+15)$ 内的概率应该是 $P=2P_1=2\times 0.4430=0.8860$。因此这种鼠的体重在 45—75 克之间的概率为 88.6%。

（2）首先需要找出这样的 W 值，使得 x 在范围 $\frac{|x-60|}{9.5}<W$ 的概率是 0.95。表 2 给的是半个正态曲线下的面积，因此应找 W 值，使得 x 在范围 $0<\frac{x-60}{9.5}<W$ 的概率为

$$P=\frac{1}{2}\times 0.95=0.475。$$

利用表 2 可以反查出 $P=0.4750$ 所对应的 $W=1.96$，于是 $W\sigma=1.96\times 9.5=18.62$（克）。因此体重 x 有 95% 的可能出现在范围 $|x-60|<18.62$（克）或者在区间 $[60-18.62, 60+18.62]$ 既 $[41.38, 78.62]$ 之内。

（三）标准误差和区间估计

通过样本的统计量论及总体的某些参数时，必须注意样本统计量的代表程度有多大。如果总体的属性没有变异（$\sigma=0$），那么样本的统计量与总体的参数完全一致。但是，由于种种随机因素的干扰，就使得样本统计量本身也是个随机变量。对同一个总体进行多次抽样，每次所得到的统计量也不会相同。

一般，我们用样本平均数做为总体平均数的估值，而样本平均数本身就是个随机变量，它要随着样本的不同而异。因此，就需要对此样本平均数对总体平均数进行估值时所产生的误差做出判断。对来自同一个总体的所有样本，如果把各样本的平均数集中起来构成一个新的总体，这个新总体的平均值一定等于原来总体的平均值，而这些样本平均数，距离总体平均值的变异大小，就可以作为判断用样本平均数估计总体平均值时误差的一个标准。因此仍然可以用标准差来度量其变异的大小。这个标准差可称为平均数标准差或称为标准误差，以 $s_{\bar{x}}$ 表示。

不能用多次抽样估算平均数标准差的办法来判断估值的误差，而要求能从一组样本本身所提供的信息，就能找出标准误差的估值。为此，需要对标准误差进一步加以分析。

由于样本平均数之间的差异主要来自两个方面：总体本身的差异和样本含量的多少（假定取样是完全随机的）。对于同一个总体取两个样本，一个含量是 10，另一个含量是 100，可以预期，较小样本的平均数之间的变异，肯定会大于较大样本的平均数之间的变异。从不同的总体抽得含量相同的两个样本，变异大的总体，其样本平均数的变异也大。上述关系可用公式 $s_{\bar{x}}=s/\sqrt{n}$ 表示。其中 s 是样本标准差，为总体标准差 σ 的一个估计值，n 为计算 $\bar{x}$ 及 s

的样本的含量，$s_{\bar{x}}$ 就是（含量为 n 的）样本平均数的标准差即标准误差。

例如：24 只某种鼠的体重平均数为 $\bar{x}_1 = 53.42$ 克，样本标准差 $s_1 = 13.59$ 克。平均数 $\bar{x}_1$ 的标准差(标准误差)则为 $s_{\bar{x}_1} = s_1/\sqrt{n} = 13.59/\sqrt{24}$ 克 $= 2.77$ 克。

耗氧量的平均数 $\bar{x}_2 = 4.84$ 毫升/小时/克，样本标准差 $s_2 = 0.88$ 毫升/小时/克。其平均数的标准差为：

$$s_{\bar{x}_2} = s_2/\sqrt{n} = 0.88/\sqrt{24}\text{毫升/小时/克} = 0.18\text{毫升/小时/克}。$$

样本平均数 $\bar{x}$ 也是个随机变量，因此用 $\bar{x}$ 估计 μ 时 μ 不能刚好就等于 $\bar{x}$。应给出一个以 $\bar{x}$ 为中心的区间，使 μ 将在这个区间之内。

如果平均数 $\bar{x}$ 来自一个正态分布的总体，利用标准误差就能得出 μ 的一个区间估计。

类似正态总体中的 W，我们构成一个新的样本统计量 t，$t = \dfrac{\bar{x} - \mu}{s_{\bar{x}}}$。与 W 不同之处在于用样本平均数 $\bar{x}$，代替了 W 中的个体值 x，用样本的标准误差 $s_{\bar{x}}$ 代替了总体标准差 σ。这个统计量对对应的频率分布曲线与正态分布很相似，我们称为 t 分布。不同之处在于它随样本含量的不同而异，而含量无限增大时，它将趋近于正态分布（图 4）。在一定的自由度（样本含量－1）下，t 分布曲线在一定区上的面积就表示统计量 $t = \dfrac{\bar{x} - \mu}{s_{\bar{x}}}$ 出现在这个区间的概率。

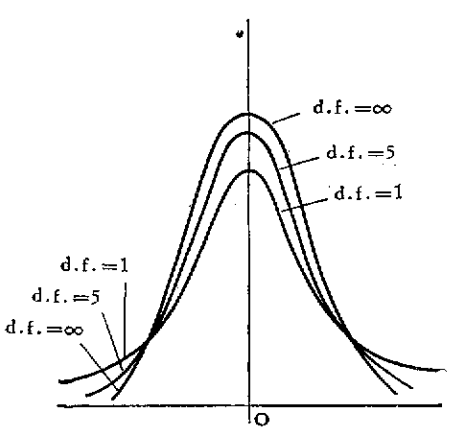

图 4　不同自由度的 t 值分布

如，当自由度 $d.f. = 24 - 1 = 23$ 时，t 分布曲线在区间 $[-1.71, 1.71]$ 上的面积是 0.9，而由 24 个数据算出的样本平均数为 $\bar{x}_1 = 53.42$，标准误差为 $s_{\bar{x}_1} = 2.77$，则其总体平均值 μ 在范围 $\left|\dfrac{53.42 - \mu}{2.77}\right| < 1.71$ 的概率为 90%，即有 90% 的把握断言：总体平均值在区间 $53.42 \pm 1.71 \times 2.77$ 即区间 $[48.68, 58.16]$ 内。

在进行区间估计时，一般总要求在一定的信度下求出所需要的区间来。因此就需在一定的概率 $1-\alpha$ 下求出 $\dfrac{|\bar{x} - \mu|}{s_{\bar{x}}}$ 的一个界限值 t_α，使得 $\dfrac{|\bar{x} - \mu|}{s_{\bar{x}}} < t_\alpha$ 的概率刚好是 $1 - \alpha$。这样就有 $1 - \alpha$ 的把握估计总体平均值 μ 将出现在区间 $[\bar{x} - t_\alpha s_{\bar{x}}, \bar{x} + t_\alpha s_{\bar{x}}]$ 之内。这个区间称之为平均值 μ 的 $1-\alpha$ 置信区间，$1-\alpha$ 称为置信度或信度。

对于不同的自由度，在一定置信度 $1-\alpha$ 下的界限 t_α 的值，已被算出，并列成 t 值表（表 3）供使用。

表 3　t 值表（节录）

自由度 \ α	0.50	0.25	0.10	0.05	0.01
⋮					
11	0.6975	1.2145	1.7959	2.2010	3.1058
13	0.6938	1.2041	1.7709	2.1604	3.0123
15	0.6912	1.1967	1.7530	2.1315	2.9467
17	0.6892	1.1910	1.7396	2.1098	2.8982
19	0.6876	1.1866	1.7291	2.0939	2.8609
21	0.6863	1.1831	1.7207	2.0796	2.8314
23	0.6853	1.1802	1.7139	2.0687	2.8073
⋮					

例如，在前述某种鼠的例子中，总体平均值 μ 的 95% 置信区间计算方法如下：

置信度 $1-\alpha = 0.95$，故 $\alpha = 0.05$。自由度 $d.f. = 24 - 1 = 23$。查 t 值表，得界限值 $t_{0.05} = 2.0687$。于是鼠的体重的 95% 置信区间为

$$[53.42 - 2.0687 \times 2.77, 53.42 + 2.0687 \times 2.77] = [47.69, 59.15]。$$

同样可以求出其耗氧量的95% 置信区间为 $[4.84-2.0687\times0.18, 4.84+2.0687\times0.18]=[4.47, 5.21]$。

利用标准误差，除了用置信区间给出总体平均数的误差范围外，还可以判断总体平均数与某些事先给定的标准之间的差异是否显著。

例如：测得 14 只 60 日龄雄鼠在X射线照射前与后之体重(克)的改变如下：

2.2，1.2，0.5，1.8，1.0，2.4，0.9，1.0，0.5，0.6，3.2，0.3，0.1，0.4，求照射前后体重的差别是否显著。

由这 14 个数据可以算出平均数 $\bar{x}=1.15$ 克，标准差 $s=0.9171$ 克，标准误差 $s_{\bar{x}}=0.2451$ 克。问题是要判断用X射线照射的小鼠这个总体中体重的平均改变量是否显著地异于零。为此先求出总体平均数的置信区间，当自由度 $d.f.=14-1=13$，置信度为 0.95 时，由 t 值表查得 $t_{0.05}=2.1604$。由此得到体重的平均改变量将在区间

$$[1.15-2.1604\times0.2451, 1.15+2.1604\times0.2451]=[0.6, 1.7],$$

即平均改变量的 95% 的置信区间为[0.6克，1.7克]。当置信度为 0.99 时，$t_{0.01}=3.0123$，还可以算出 99% 的置信区间为 [0.39 克，1.91 克]。由于 0 在 99% 的置信区间之外，所以有 99% 的把握断言：照射前后雄鼠的体重有明显的差别。