# 动物育种统計原理和方法

（全国畜牧科研单位数量遗传学学习班教材）

动物数量遗传理论及其应用科研协作组

一九七九年六月

# 目　　录

# 一、均数、标准差和方差

生物统计是经常与生物现象中各种数量打交道的。数量有两种：一种叫常量，或称常数，这种数量是在特定条件下保持不变的。例如家畜在正常状态下腿数都是4。另一种叫做变量。它表现为一系列变化不同的数（变数）。例如猪的离奶体重，有9公斤、10公斤、11公斤……20公斤等，每头猪有一个数（当然也有相同的）。这种数量在生物界是最常见的，也是生物统计研究的主要对象。变量是由一系列变数组成的，所以也可叫做变数列。

## (一)均数（即平均数）

一般指算术平均数而言。算术平均数就是组成一个变量的全部变数的总和被变数的个数除而得的商。用公式表示，即：$\overline{X}=\frac{\sum X}{N}$。这里$\overline{X}$代表均数，$\sum X$代表全部变数的总和，N代表变数的个数。

均数是同质的变数列（变量）的代表值。不同质的变数不能加在一起求均数。例如一头猪体重80公斤，另一头猪体重100公斤，两者的平均体重为90公斤，如果一个是体重80公斤，另一个是体长100厘米，两者不可能有均数。

均数具有以下几个特性：

1．各变数与均数的离差（即离均差）的总和等于0。

$$\sum (X-\overline{X})=0$$

例如：2、4、6、8四个变数的均数为5。

$$\sum (X-\overline{X}) = (2-5)+(4-5)+(6-5)+(8-5)$$

$$= (-3)+(-1)+1+3=0$$

各变数与任何其他数的离差的总和都不等于0，所以从绝对值来看，离均差的总和是最小值。因此均数是与各变数最接近的一数值，所以最能代表这个变量。

2．不但离均差总和的绝对值是最小的，而且离均差平方和也是最小的离差平方和。这就是所谓“最小二乘”。

例如还是这四个变数，其各离差平方和与离均差平方和的比较如下表。

| 变数 | 与2的 | | 与4的 | | 与6的 | | 与8的 | | 与均数5的 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 离差 | 离差平方 | 离差 | 离差平方 | 离差 | 离差平方 | 离差 | 离差平方 | 离差 | 离差平方 |
| 2 | 0 | 0 | −2 | 4 | −4 | 16 | −6 | 36 | −3 | 9 |
| 4 | 2 | 4 | 0 | 0 | −2 | 4 | −4 | 16 | −1 | 1 |
| 6 | 4 | 16 | 2 | 4 | 0 | 0 | −2 | 4 | 1 | 1 |
| 8 | 6 | 36 | 4 | 16 | 2 | 4 | 0 | 0 | 3 | 9 |
| 离差平方和 | | 56 | | 24 | | 24 | | 56 | | 20 |

由表可见，离均差20<24<56，是最小的离差平方和。

3．各变数都加或减一个常数，其均数等于原均数加或减该常数。例如2、4、6、8各加3，其均数就等于：

$$\frac{(2+3)+(4+3)+(6+3)+(8+3)}{4}=\frac{5+7+9+11}{4}=\frac{32}{4}=8=5+3$$

因此，在计算变数的均数时有时可以简化，例如求102、104、106、108四个变数的均数，就可先从各变数中减去100，求出均数，然后再加上100。

$$X=\frac{2+4+6+8}{4}+100=105$$

根据这个特性，还可以利用“假定均数”来计算均数，这种方法叫做“假定均数法”。

例如：2、4、6、8四个变数，我们可以任意假定哪个数为假定均数，计算各变数与此假定均数的平均离差，然后加到假定均数上，即得真均数。如我们假定4为假定均数，各变数与它的平均离差等于：

$$\frac{(2-4)+(4-4)+(6-4)+(8-4)}{4}=\frac{(-2)+0+2+4}{4}=\frac{4}{4}=1$$

均数＝假定均数＋平均离差＝4＋1＝5

这种方法在计算大量变数的均数时，可大大简化计算过程，具体方法以后再介绍。

4、各变数乘或除以一常数，其均数等于原均数乘或除以该常数。例如2、4、6、8四个变数各除以2，其均数等于：

$$\frac{(2\div2)+(4\div2)+(6\div2)+(8\div2)}{4}=\frac{1+2+3+4}{4}=\frac{10}{4}=2.5=5\div2$$

根据这一特性，在计算小数的均数时，可先化成整数，求出均数后再化回小数，例如求0.2、0.4、0.6、0.8的均数，可先各乘以10，得2、4、6、8均数为5，然后再除以10，得0.5，此即这四个小数的均数。

均数是变量的代表值。一个变量内各变数有一定的集中趋势，其中心就是均数。所

以我们往往用均数来代表一个变量。例如哈白猪的离奶体重是一个变量，表现为很多变数。如果问哈白猪离奶体重多少，该用什么数来回答呢？最恰当的就是用均数。当我们说哈白猪的离奶体重平均为12公斤，我们对这个变量就有了一个一定的概念。

## (二)标准差

用均数代表变量只说明它的集中性，而没有说明它的离中性，即其内部各变数的变动情况，因此就有其一定的片面性。均数相同的两个变量，往往可以有很大差异。例如一个由3、4、5、6、7组成，另一个由1、2、4、8、10组成，均数虽然都是5，但两者却很不相同，在育种工作中，了解一个性状的变异情况尤其重要。

表示变量内各变数离中情况的指标很多，例如全距、最大最小值、平均差、变异系数等。但最全面最基本的还是标准差，标准差即标准离均差，是各变数与均数这个中心的标准差距。计算的基本公式是：

$$S=\sqrt{\frac{\sum(X-\bar{X})^2}{N-1}}=\sqrt{\frac{SSx}{df}}$$

根号内的分子是离均差的平方和（简称平方和，以SS代表），分母是自由度，以df代表，在这里自由度是变数个数减1。

自由度在统计学中的概念就是自由变数的个数。例如a、b、c三个变数，在没有条件的情况下，这三个都是自由变数，每一个都可以是任何数值。但在一定的条件下，譬如在a+b+c=10这样一个条件下，自由变数就只有两个了。若a=4、b=5，则c就必须等于1；若a=10，b=5，则c就必须等于-5。三个变数中只有两个是自由变数。这时的自由度就是N-1=3-1=2。自由度一般比总个数小，等于总个数减去计算过程中使用的条件数。在计算标准差时，条件就是一个，即$\sum(X-\bar{X})=0$，故自由度=N-1。

由公式可见，标准差是平均离均差平方的方根。那么为什么不直接用平均离均差呢？因为离均差的总和等于0，平均离均差当然也等于0，用平均离均差就什么也说明不了。所以离均差先用平方来消除其正负，平均以后再开方。由于平方又开方，因此标准差的单位与变数的单位相同。例如变数是公斤，计算出标准差的单位也是公斤。

这个公式的主要计算部分是分子，即平方和。为了便于利用计算机或巴罗表，通过简单的演算，平方和可以化成：

$$\sum(X-\bar{X})^2=\sum(X^2-2X\bar{X}+\bar{X}^2)=\sum X^2-\sum 2X\bar{X}+\sum\bar{X}^2$$

$$=\sum X^2-\bar{X}(2\sum X-\sum\bar{X})=\sum X^2-\bar{X}\cdot\sum X$$

$$=\sum X^2-\frac{(\sum X)^2}{N}\quad(\text{注意：}\sum\bar{X}=\sum X)$$

这样，在计算时就可以不必先求出均数，然后将每个变数减去均数，再将这些离均差一一平方，最后总加起来，而可以利用计算机或巴罗表直接求出各变数平方的总和，减去变数总和的平方的平均值就行了。平方和的这个公式由两部分组成，一部分是变数平

方的总和$\sum X^2$，它与离均差的平方和之间有一定的差距，这个差距就是公式的第二部分，即变数总和平方的平均值$\frac{(\sum X)^2}{N}$。从这个意义上讲，后一部分起校正前一部分的作用，因此统计学上称这一部分为校正数；一般以C代表，即：$C=\frac{(\sum X)^2}{N}$。

例如一样本由3、4、5、6、7五个变数组成，其标准差计算方法如下：

$$S=\sqrt{\frac{SS}{df}}=\sqrt{\frac{\sum x^2-\frac{(\sum x)^2}{N}}{N-1}}=\sqrt{\frac{3^2+4^2+5^2+6^2+7^2-\frac{(3+4+5+6+7)^2}{5}}{5-1}}$$

$$=\sqrt{\frac{135-\frac{(25)^2}{5}}{4}}=\sqrt{\frac{135-125}{4}}=\sqrt{\frac{10}{4}}=\sqrt{2.5}=1.581$$

统计学中常以б代表总体的标准差，S代表样本的标准差。

标准差的特性有：

1．各变数都加或减一个常数，标准差不变。

$$б_{x+a}=\sqrt{\frac{\sum[(X+a)-(\bar{X}+a)]^2}{N-1}}=\sqrt{\frac{\sum(X-\bar{X})^2}{N-1}}=б_x$$

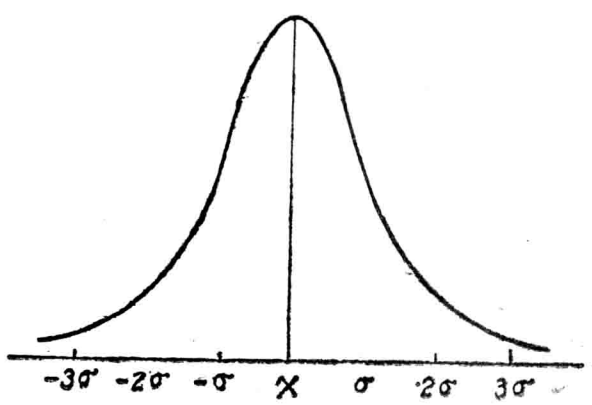
因此，在计算大变数的标准差时，有时可以简化。例如求83、84、85、86、87五个变数的标准差，可以先将各变数都减去80，得到3、4、5、6、7五个变数，然后再求这五个数的标准差：

S＝1.581（见上例），这也就是原来五个变数的标准差。

2．各变数乘或除以一个常数，求得标准差等于原标准差乘或除以该常数。

$$б_{ax}=\sqrt{\frac{\sum(aX-a\bar{X})^2}{N-1}}=\sqrt{\frac{a^2\sum(X-\bar{X})^2}{N-1}}=a\sqrt{\frac{\sum(X-\bar{X})^2}{N-1}}=aб_x$$

3．若以变数值为横坐标，以变数出现的次数为纵坐标，绘成变数分布图，则可见到一般数量性状的变数在大样本中，往往呈钟形分布，这个分布叫做正态分布：



在N相当大的情况下，标准差大致等于全距（最大与最小变数间的差距）的1/6。约有68%的变数处在均数上下各一个标准差的范围内；约95%的变数处在均数上下各两个标准差的范围内；约99%的变数处在均数上下三个标准差的范围内。

均数与标准差是统计学中最基本的两个参数，以后要介绍的一些统计学分析方法都是在其基础上推演出来的。均数表示变量的中心所在，标准差表示变量的离中情况，两者有密切关系。离中性大说明中心的代表性差；反之，标准差愈小，均数的代表性愈大。若标准差等于0，即各变数与均数都没有离差，各变数都等于均数，当然均数就能完全代表各个变数，它的代表性就最大。

标准差虽能表示变量的变异情况，但由于它是绝对值，因此，不能反映不同变数间的相对变异情况。单位不同或均数不同的变量要比较它们的变异程度时，应把标准差化成相对值。即变异系数，然后才能互比大小。

变异系数$C\cdot V=\frac{S}{\bar{X}}\times 100\%$

例如黑白花奶牛平均产奶量3800公斤，标准差800公斤；平均乳脂率3.4%，标准差0.4%。前者的变异系数$=\frac{800}{3800}\times 100\%=21.05\%$；后者的变异系数$=\frac{0.4}{3.4}\times 100\%=11.76\%$，说明产奶量的变异程度大于乳脂率的变异程度。

同一总体的各样本的均数不尽相同，这些均数的标准差称为标准误$S_{\bar{x}}$，一般是通过样本的标准差估计得来，公式为$S_{\bar{x}}=\frac{S_x}{\sqrt{N}}$。

## （三）均数差异显著性测定

从同一总体中抽取一定大小的样本，每次抽样求得的均数都不完全相同；同一现象所做的试验，每次试验所得数据的均数也不会完全相同。那末这些均数的差异能否反映它们所代表的总体或试验结论之间存在本质的差别呢？显然不能。这些差异只不过是抽样误差，是由一些我们目前还不能控制或不必要控制的偶然因素造成的。测定两个均数间的差异是实质性差异还是偶然性差异，统计学上称为均数差异显著性测定。显著的可认为是实质性差异，不显著的是偶然性差异。

均数差异显著性测定与大多数统计量的显著性测定一样，一般采用t测定。t测定的大致步骤是：首先根据统计量值与其标准误之比求出t值，然后查t表找到一定自由度下t的理论值，将求得之t值与表中查得之t值进行比较，以此确定该统计量值是否显著。

均数差异的t测定也是这样，这时统计量就是两均数的差。

$$t=\frac{\bar{X}_1-\bar{X}_2}{S_{\bar{d}}}=\frac{\bar{d}}{S_{\bar{d}}}.$$

$\bar{d}$——两均数之差；　　$S_{\bar{d}}$——均数差的标准误。

均数差的标准误在不同的资料计算方法不同。大致有两种方法；在成对资料时用每对数据之差代入标准误的公式计算：在不成对资料时则利用两样本各自的标准误计算合并标准误。

兹将两种不同情况的计算步骤分别举例说明如下：

1．成对资料：

根据黑龙江省银浪羊场绵羊药物采毛试验，16只细毛母羊1977年用药物采毛法获得的产毛量与其本身1976年手工剪毛的产毛量对比资料如下：

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1977年毛量$x_1$ | 5.9 | 6.0 | 6.9 | 5.8 | 5.7 | 6.8 | 5.6 | 6.1 | 9.6 | 6.7 | 6.4 | 4.5 | 6.0 | 4.9 | 5.2 | 4.2 |
| 1976年毛量$x_2$ | 6.3 | 5.4 | 5.1 | 6.1 | 5.8 | 5.7 | 4.5 | 5.3 | 7.3 | 4.3 | 5.8 | 5.6 | 5.5 | 4.1 | 4.1 | 6.6 |
| 差 d | −0.4 | 0.6 | 1.8 | −0.3 | −0.1 | 1.1 | 1.1 | 0.8 | 2.3 | 2.4 | 0.6 | −1.1 | 0.5 | 0.8 | 1.1 | −2.4 |

$N=16$，　$\bar{X}_1=6.02$，　$\bar{X}_2=5.47$，　$\bar{d}=0.55$

(1) 计算两均数差的标准误 $S_{\bar{d}}$

$$S_d=\sqrt{\frac{\sum d^2-\frac{(\sum d)^2}{N}}{N-1}}=\sqrt{\frac{27.4-\frac{(8.8)^2}{16}}{16-1}}=\sqrt{\frac{27.4-4.84}{15}}=\sqrt{1.504}=1.23$$

$$S_{\bar{d}}=\sqrt{\frac{S_d^2}{N}}=\sqrt{\frac{1.504}{16}}=\sqrt{0.094}=0.31$$

(2) 计算t值：

$$t=\frac{\bar{d}}{S_{\bar{d}}}=\frac{0.55}{0.31}=1.77$$

(3) 查t表：自由度=15时，$P=0.05$，$t=2.131$；$P=0.01$，$t=2.947$。

计算所得的t值1.77<2.131，更<2.947，∴P>0.05，说明两均数的差异不显著。也就是说，使用药物采毛法，绵羊的剪毛量未受显著影响（若两年度的饲养管理条件基本相同）。

2．不成对资料：

据乌鲁木齐地区奶牛育种协作组1975年调查结果。五一农场90头青年母牛平均体高131.27厘米，标准差5.28；104团场80头同龄母牛平均体高133.4厘米，标准差4.14。

(1) 计算两均数差的标准误，即两均数标准误的合并标准误。

$$S_{\bar{d}}=\sqrt{\frac{SS_1+SS_2}{df_1+df_2}\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}$$

$$SS_1 = S^2{}_1 df_1 = 5.28^2(90-1) = 27.88 \times 89 = 2481.32$$

$$SS_2 = S^2{}_2 df_2 = 4.14^2(80-1) = 17.14 \times 79 = 1354.06$$

$$S_{\bar{d}} = \sqrt{\frac{2481.32+1354.06}{89+79}\left(\frac{1}{90}+\frac{1}{80}\right)}$$

$$= \sqrt{3835.38 \times \frac{1}{168}(0.0111+0.0125)}$$

$$= \sqrt{3835.38 \times 0.0059 \times 0.0236}$$

$$= \sqrt{0.5340} = 0.731$$

(2) 计算t值：

$$t = \frac{\bar{d}}{S_{\bar{d}}} = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{d}}} = \frac{131.27-133.48}{0.731} = \frac{-2.21}{0.731} = -3.023$$

t 值 表

| 自由度 (df) | 机率 (P) | | | | |
|---|---|---|---|---|---|
| | 0.20 | 0.10 | 0.05 | 0.01 | 0.001 |
| 1 | 3.078 | 6.314 | 12.706 | 63.657 | 636.619 |
| 2 | 1.886 | 2.920 | 4.303 | 9.925 | 31.598 |
| 3 | 1.638 | 2.353 | 3.182 | 5.841 | 12.924 |
| 4 | 1.533 | 2.132 | 2.776 | 4.604 | 8.610 |
| 5 | 1.476 | 2.015 | 2.571 | 4.032 | 6.859 |
| 6 | 1.440 | 1.943 | 2.447 | 3.707 | 5.959 |
| 7 | 1.415 | 1.895 | 2.365 | 3.499 | 5.405 |
| 8 | 1.397 | 1.860 | 2.306 | 3.355 | 5.041 |
| 9 | 1.383 | 1.833 | 2.262 | 3.250 | 4.781 |
| 10 | 1.372 | 1.812 | 2.228 | 3.169 | 4.587 |
| 11 | 1.363 | 1.796 | 2.201 | 3.106 | 4.437 |
| 12 | 1.356 | 1.782 | 2.179 | 3.055 | 4.318 |
| 13 | 1.350 | 1.771 | 2.160 | 3.012 | 4.221 |
| 14 | 1.345 | 1.761 | 2.145 | 2.977 | 4.140 |
| 15 | 1.341 | 1.753 | 2.131 | 2.947 | 4.073 |
| 16 | 1.337 | 1.746 | 2.120 | 2.921 | 4.015 |
| 17 | 1.333 | 1.740 | 2.110 | 2.898 | 3.965 |
| 18 | 1.330 | 1.734 | 2.101 | 2.878 | 3.922 |
| 19 | 1.328 | 1.729 | 2.093 | 2.861 | 3.883 |
| 20 | 1.325 | 1.725 | 2.086 | 2.845 | 3.850 |
| 21 | 1.323 | 1.721 | 2.080 | 2.831 | 3.819 |
| 22 | 1.321 | 1.717 | 2.074 | 2.819 | 3.792 |
| 23 | 1.319 | 1.714 | 2.069 | 2.807 | 3.767 |
| 24 | 1.318 | 1.711 | 2.064 | 2.797 | 3.745 |
| 25 | 1.316 | 1.708 | 2.060 | 2.787 | 3.725 |
| 26 | 1.315 | 1.706 | 2.056 | 2.779 | 3.707 |
| 27 | 1.314 | 1.703 | 2.052 | 2.771 | 3.690 |
| 28 | 1.313 | 1.701 | 2.048 | 2.763 | 3.674 |
| 29 | 1.311 | 1.699 | 2.045 | 2.756 | 3.659 |
| 30 | 1.310 | 1.697 | 2.042 | 2.750 | 3.646 |
| 50 | 1.299 | 1.676 | 2.008 | 2.678 | 3.496 |
| 100 | 1.290 | 1.661 | 1.982 | 2.625 | 3.390 |
| ∞ | 1.2816 | 1.6448 | 1.9600 | 2.5758 | 3.2905 |

(3) 查t表：

自由度$df_1 + df_2 = 89 + 79 = 168$

当自由度=120时，P=0.05，t=1.98，P=0.01，t=2.617计算所得t值 3.023，>2.617，更>1.98，∴P<0.01，说明两均数的差异非常显著，也就是说，两场青年母牛的体高有显著差别。

3. 当$N_1 = N_2$时，$df_1 = df_2$

$$S_{\bar{d}} = \sqrt{\frac{SS_1 + SS_2}{df_1 + df_2}\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} = \sqrt{\frac{SS_1 + SS_2}{2df} \cdot \frac{2}{N}}$$

$$= \sqrt{\left(\frac{SS_1}{df} + \frac{SS_2}{df}\right)\frac{1}{N}} = \sqrt{(S_1^2 + S_2^2)\frac{1}{N}} = \sqrt{S_{\bar{x}_1}^2 + S_{\bar{x}_2}^2}$$

两个大样本，如果N相差不大，可按N相同计算，如上例，也可计算如下：

$$S_{\bar{d}} = \sqrt{S_{\bar{x}_1}^2 + S_{\bar{x}_2}^2} = \sqrt{\frac{S_2^2}{N_1} + \frac{S_2^2}{N_2}} = \sqrt{\frac{5.28^2}{90} + \frac{4.14^2}{80}}$$

$$= \sqrt{\frac{27.88}{90} + \frac{17.14}{80}} = \sqrt{0.3098 + 0.2142}$$

$$= \sqrt{0.5240} = 0.7239$$

$$t = \frac{-2.21}{0.7239} = -3.053 > 2.617$$

∴ P<0.01，非常显著。

## (四)用分组法计算均数和标准差

当分析的资料包含大量变数时，可利用分组和级差法大大简化计算过程。

以吉林省双辽种羊场一九七六年126头基础母羊剪毛前体重资料为例：

(单位：公斤)

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 53.0 | 50.0 | 51.0 | 57.0 | 56.0 | 51.0 | 48.0 | 46.0 | 62.0 | 51.0 | 61.0 |
| 56.0 | 62.0 | 58.0 | 46.5 | 48.0 | 46.0 | 50.0 | 54.5 | 56.0 | 40.0 | 53.0 |
| 51.0 | 57.0 | 54.0 | 59.0 | 52.0 | 47.0 | 57.0 | 59.0 | 54.0 | 50.0 | 52.0 |
| 54.0 | 62.0 | 50.0 | 50.0 | 53.0 | 51.0 | 54.0 | 56.0 | 50.0 | 52.0 | 50.0 |
| 52.0 | 43.0 | 53.0 | 48.0 | 50.0 | 60.0 | 58.0 | 52.0 | 54.5 | 50.0 | 47.0 |
| 37.0 | 52.0 | 46.0 | 45.0 | 48.0 | 53.0 | 58.0 | 47.0 | 64.0 | 50.0 | 45.0 |
| 55.0 | 62.0 | 51.0 | 50.0 | 43.0 | 53.0 | 42.0 | 56.0 | 50.0 | 45.0 | 56.0 |
| 54.0 | 65.0 | 61.0 | 47.0 | 52.0 | 49.0 | 49.0 | 51.0 | 45.0 | 52.0 | 54.0 |
| 48.0 | 57.0 | 45.0 | 53.0 | 54.0 | 57.0 | 54.0 | 54.0 | 45.0 | 44.0 | 52.0 |
| 50.0 | 52.0 | 52.0 | 55.0 | 50.0 | 54.0 | 43.0 | 57.0 | 56.0 | 54.0 | 49.0 |
| 55.0 | 50.0 | 48.0 | 46.0 | 56.0 | 45.0 | 45.0 | 51.0 | 46.0 | 49.0 | 48.5 |
| 49.0 | 55.0 | 52.0 | 58.0 | 54.5 | | | | | | |

1．编制次数分配和计算表：

在核实原始资料的基础上，将全部变数，按其数值大小进行分组。组数的多少，根据变数的总个数（即总次数N）的多少和要求的精确度来决定，一般N小于30时不分组，50时分5组，100时分8组，200时分10组，300时分12组，500时分15组，1000时分20组为宜。但这并不是死规定，仅供参考。分组愈多愈精确，但计算就较繁。分组时每组的差距，即组距i应相等，组距应大致等于$\frac{全距}{组数}$，以整数为方便。然后从最小变数或略小于最小变数开始，按组距划分各组的组限。将全部变数分配入相应各组，分计各组的次数。

如上列资料最小变数为37，最大变数为65，全距为65－37＝28。N＝126，以分8组为宜，$i \approx \frac{28}{8} = 3.5$，为方便起见，i可定为4。将全部资料整理成下表：

| 组　限 | 中值(x) | 划　线　记　录 | 次数(f) | 级差(d) | fd | $fd^2$ |
|---|---|---|---|---|---|---|
| 35— | 37 | 一 | 1 | —4 | —4 | 16 |
| 39— | 41 | 丅 | 2 | —3 | —6 | 18 |
| 43— | 45 | 正正正丅 | 18 | —2 | —36 | 72 |
| 47— | 49 | 正正正正正正一 | 31 | —1 | —31 | 31 |
| 51— | 53 | 正正正正正正正正一 | 41 | 0 | 0 | 0 |
| 55— | 57 | 正正正正丅 | 22 | 1 | 22 | 22 |
| 59— | 61 | 正丌 | 9 | 2 | 18 | 36 |
| 63— | 65 | 丅 | 2 | 3 | 6 | 18 |
| | | | N＝126 | | $\sum fd = -31$ | $\sum fd^2 = 213$ |

2．决定假定均数和各组的级差：

分组后，各变数就丧失其原来的数值，而以其所在组的中值为代表，中值等于该组的下限加$\frac{i}{2}$，例如第1组下限为35，$\frac{i}{2} = \frac{4}{2} = 2$，中值＝35＋2＝37。中值与该组各变数的均数之间可能有出入，但各组的中值与均数的离差有正有负，在样本较大的情况下，各组的这种离差几乎可以完全抵销，所以总的误差不会太大。

为了进一步简化，可利用假定均数来计算。假定均数（$\bar{X}'$）可任意假定，计算结果都相同。但为了计算方便，一般以次数最多的组中值为假定均数，也可以$\frac{N}{2}$次数的所在组的中值为假定均数，例如$\frac{N}{2} = \frac{126}{2} = 63$，从第一组开始，顺序将各组的次数加起来，加到第五组次数总和为95、63正处于这一组内，故以这一组的中值53为假定均数。这样确定的假定均数最接近真均数，计算也就最方便。

级差的决定很简单，就是以假定均数所在组为0，向上各组顺序为－1，－2，－3，

－4，向下各组顺序为1、2、3。要注意的是即使没有次数分布的组也得顺序写上级差，不能跳过去，否则就要出错，这样定出的 $d=\frac{X-\overline{X}'}{i}$。

3．计算出fd、$\sum fd$、$fd^2$、$\sum fd^2$。

fd就是各组的次数（f）乘本组的级差（d）。$\sum fd$就是各组fd的总和。

$fd^2$就是各组fd再乘本组的d。$\sum fd^2$就是各组$fd^2$的总和。

4．代入公式：

$$\overline{X}=\frac{\sum fd}{N}\cdot i+\overline{X}'=\frac{-31}{126}\times 4+53=-0.246\times 4+53$$

$$=-0.984+53=52.016\text{公斤}$$

$$S=\sqrt{\frac{\sum fd^2-\frac{(\sum fd)^2}{N}}{N-1}}\cdot i$$

$$=\sqrt{\frac{213-\frac{(-31)^2}{125}}{125}}\times 4=\sqrt{\frac{213-\frac{961}{126}}{125}}\times 4$$

$$=\sqrt{\frac{213-7.63}{125}}=\sqrt{\frac{205.37}{125}}=\sqrt{1.64}\times 4$$

$$=1.28\times 4=5.12\text{公斤}$$

## （五）方差（也称变量或变异量）

因易与变数列组成的变量相混淆，故以称方差为宜。方差的意思就是平方标准差（$S^2$）。和标准差一样，也是说明一个变量内部的变异情况的，但其计算过程比标准差省去开方这一步，因而更加方便。

$$S^2=\frac{\sum(X-\overline{X})^2}{N-1}$$

方差具有下列特性：

1．$S^2(x+a)=S^2$

2．$S^2(x\cdot a)=a^2S^2{}_x$

3．$S^2(x+y)=S^2{}_x+S^2{}_y+2COV_{xy}$

$COV_{xy}$ ——x和y的协方差，即$\frac{\sum(X-\overline{X})(Y-\overline{Y})}{N-1}$.

4．具有可分析性：

后两个特性以后还要详细解释。

## (六)习　题

计算下列资料的均数和标准差：

52头新淮猪后备母猪的胸围（厘米）

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 70 | 69 | 69 | 66 | 64 | 66 | 72 | 66 | 72 | 69 |
| 74 | 75 | 72 | 75 | 70 | 64 | 77 | 73 | 70 | 69 |
| 73 | 70 | 73 | 76 | 68 | 71 | 62 | 65 | 68 | 71 |
| 68 | 77 | 65 | 72 | 75 | 78 | 76 | 73 | 72 | 76 |
| 73 | 76 | 72 | 78 | 63 | 66 | 66 | 66 | 77 | 72 |
| 67 | 67 | | | | | | | | |

## (七)附计算实例

1973年春产260头苏白仔猪的离奶种重（公斤）

（黑龙江省生产建设兵团某部18团种畜连）

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 18 | 19.75 | 16.5 | 19.5 | 18.25 | 22 | 19 | 20 | 18.25 | 14.5 | 22.5 | 21 | 16.5 | 17.5 |
| 19.5 | 20 | 19 | 17.5 | 18.5 | 14.5 | 20.75 | 13.5 | 20 | 15 | 20.5 | 20 | 17 | 19 |
| 16 | 22.5 | 21 | 23.5 | 19 | 25 | 20 | 24.5 | 21 | 23 | 24 | 23 | I4 | 11 |
| 15.75 | 16.5 | 18 | 7 | 6 | 16 | 15 | 18.75 | 15 | 15.5 | 11.5 | 11 | 8 | 11.5 |
| 13.5 | 13 | 9 | 15.5 | 9.25 | 11 | 10 | 17 | 19 | 14.5 | 14.25 | 16 | 13.75 | 12.75 |
| 12.25 | 13.25 | 15 | 15.25 | 9 | 16 | 11.5 | 15.5 | 26 | 25 | 19 | 11.25 | 16.5 | 20.5 |
| 13.25 | 19.5 | 11.5 | 20 | 20 | 17 | 18.5 | 16.5 | 15 | 14.75 | 13 | 16.75 | 10.5 | 15 |
| 23 | 24 | 17 | 17 | 16.5 | 15.75 | 23.5 | 18 | 16.5 | 18.5 | 20.5 | 18.5 | 22 | 18.5 |
| 20.5 | 18.25 | 20.5 | 26 | 18.5 | 18.5 | 21 | 15 | 18.25 | 24 | 26.5 | 24 | 20 | 26 |
| 23.25 | 22 | 24.5 | 20.75 | 18 | 22.5 | 17 | 13.75 | 14.5 | 16.5 | 17 | 17 | 19 | 18.5 |
| 17.5 | 15 | 19 | 15.2 | 15.5 | 19 | 19.5 | 20 | 23.5 | 13.5 | 19 | 19.25 | 21.25 | 21 |
| 11.5 | 9.75 | 16 | 16 | 15.5 | 20.5 | 14 | 17.75 | 14.75 | 15.75 | 19 | 22 | 18.75 | 16.75 |
| 19 | 17 | 21.5 | 17 | 20 | 16 | 22.25 | 23.5 | 17 | 20 | 14.5 | 26 | 22.5 | 24 |
| 15 | 13.5 | 22.5 | 15.5 | 15 | 16.75 | 17 | 17.75 | 13 | 9.5 | 16.25 | 15 | 13 | 13.75 |
| 12.5 | 16.5 | 13.75 | 16 | 21.25 | 18.5 | 19.75 | 21 | 20 | 18.5 | 17 | 15.5 | 19.25 | 20.5 |
| 20.5 | 20 | 12 | 18.5 | 13.5 | 20 | 16 | 18.5 | 17.5 | 21.5 | 25.25 | 25 | 26 | 19 |
| 21 | 21.5 | 19 | 15.75 | 21 | 21.25 | 21.5 | 21.75 | 21.5 | 19 | 14.5 | 20.25 | 23.5 | 22 |
| 20.25 | 22 | 23 | 17 | 22.5 | 21 | 19.75 | 16.75 | 16.5 | 21 | 17 | 20 | 22.25 | 17.5 |
| 19 | 17.5 | 17.25 | 13.5 | 18.5 | 19 | 16 | 13.5 | | | | | | |

计算其平均数和标准差：

1．编制次数分配表：

先找出最小值和最大值及算出全距。

最小值为 6 公斤，最大值为26.5公斤。

全距 = 最大值—最小值 = 26.5 − 6 = 20.5公斤。

决定组数和组距，因资料总次数为260头，故可分为10—12组。考虑全距与组距，具体可分11组。

$$\text{组距 (i)} = \frac{\text{全距}}{\text{组数}} = \frac{20.5}{11} = 1.86 \approx 2 \text{ 公斤}$$

根据已定组数和组距将资料进行分组，并作各组次数的计算；

| 组　限 | 中　值 | 划　线　计　数 | 次　数 (f) |
|---|---|---|---|
| 5 — | 6 | 一 | 1 |
| 7 — | 8 | 丅 | 2 |
| 9 — | 10 | 正丅 | 7 |
| 11— | 12 | 正正下 | 13 |
| 13— | 14 | 正正正正正下 | 28 |
| 15— | 16 | 正正正正正正正正正下 | 49 |
| 17— | 18 | 正正正正正正正正正丅 | 47 |
| 19— | 20 | 正正正正正正正正正正下 | 53 |
| 21— | 22 | 正正正正正正下 | 33 |
| 23— | 24 | 正正正丅 | 17 |
| 25— | 26 | 正正 | 10 |
| | | $N = \Sigma f = 260$ | |

计算各组的中值，例如第一组的中值，因下限为 5，$\frac{i}{2} = 1$，故中值 = 下限 + $\frac{i}{2}$ = 5 + 1 = 6 。

用划线计数方法计算各组的次数。

2．计算平均数和标准差：

根据次数分配表列出平均数和标准差计算表（表见下页）。

决定假定均数（$\overline{X}'$）和各组的级差（d）。

设$\overline{X} = 16$，则假定均数所在组的级差为 0，比 $\overline{X}'$ 组小的各组的级差依次为 −1、−2、−3……，比 $\overline{X}_1$ 组大的级差依次为1、2、3……。

计算fd、$\Sigma fd$、$fd^2$及$\Sigma fd^2$值。

例如第一组的d = −5，∴ $fd = 1 \times (-5) = -5$，$fd^2 = fd \times d = (-5) \times (-5) = 25$，把各项fd值及$fd^2$值总加起来即得$\Sigma fd$值和$\Sigma fd^2$值。

| 组　限 | 中　值 | 次　数 (f) | 级　差 (d) | fd | $fd^2$ |
|---|---|---|---|---|---|
| 5— | 6 | 1 | —5 | —5 | 25 |
| 7— | 8 | 2 | —4 | —8 | 32 |
| 9— | 10 | 7 | —3 | —21 | 63 |
| 11— | 12 | 13 | —2 | —26 | 52 |
| 13— | 14 | 28 | —1 | —28 | 28 |
| 15— | 16 | 49 | 0 | 0 | 0 |
| 17— | 18 | 47 | 1 | 47 | 47 |
| 19— | 20 | 53 | 2 | 106 | 212 |
| 21— | 22 | 33 | 3 | 99 | 297 |
| 23— | 24 | 17 | 4 | 68 | 272 |
| 25— | 26 | 10 | 5 | 50 | 250 |
| Σ | | 260 | | 282 | 1278 |

代入公式计算平均数（$\overline{X}$）及标准差（S）：

$$\overline{X} = \overline{X}' + \frac{\Sigma fd}{N} i = 16 + \frac{282}{260} \times 2 = 16 + 2.17 = 18.17\text{公斤}$$

$$S = i\sqrt{\frac{\Sigma fd^2 - \frac{(\Sigma fd)^2}{N}}{N-1}} = 2\sqrt{\frac{1278 - \frac{(282)^2}{260}}{260-1}}$$

$$= 2\sqrt{\frac{1278 - \frac{79524}{260}}{259}} = 2\sqrt{\frac{1278 - 305.86}{259}}$$

$$= 2\sqrt{\frac{972.14}{259}} = 2\sqrt{3.75} = 2 \times 1.94 = 3.88\text{公斤}$$