课前提问

1. 生物统计学的研究对象？
2. 生物统计学的任务？
3. 什么是数据？数据的类型？
4. 数据的性质？

提问：

1. 生物统计学的研究对象？

总体和样本

2. 生物统计学的目的？

由样本的信息推断总体的特征。

有几个小孩很想当天使，上帝给他们一人一个烛台，叫他们保持烛台光亮。结果几天过去了，上帝都没来，几乎所有的小孩都不再擦拭那烛台，有一天上帝突然造访，他们每个人的烛台都蒙上了厚厚的灰尘。只有一个小孩大家都叫他笨小孩，因为上帝没来，他也每天都擦拭，结果这个笨小孩成了天使。

其实当天使很简单，只要实实在在去做就行了。

第一章 统计数据的收集与整理

本章内容

1.1 总体与样本

1.2 数据

1.3 频数表与频数图的编绘

1.4 生物统计的常用术语

1.5 平均数与标准差

本章要点

- 掌握数据的基本性质
- 频数表与频数图的编绘
- $\bar{x}$ 与 S 的意义与计算（重点）

数据整理

主要任务 → 资料审核、分组、汇总、制表、制图等。

- 分组
- 频数分布
- 统计表
- 统计图

目的：1.区别不同类型，揭示内在联系

2、找出数据的重要特征：a：集中情况；b：变异情况：

3、曲线情况:变化趋势。

4、不规则情况：极端值。

1.2 数据

- 数据
- 数据的性质–变异性
- 数据的类型

1.2.1　数据

对客观事物通过调查、测量、实验所得到的数量化的观察结果，叫做数据(Data)。

1.2 .2 数据的性质

- **数据的基本性质就是变异性(Variability)。**

  变异性是统计学的出发点和根本所在。

生物统计学就是研究生物群体内个体间的**变异性规律**以及研究对生物性状观察过程中的**误差规律**。

- **变异性产生的原因**

- 变异使自然界丰富多彩，保持了生物的多样性。

科学研究的目的就在于把这些由内外因素所引起的变异的客观规律剖析清楚，以便通过它们对同类的事物或同型的现象加以估计和预测。

1.2.3. 数据的类型

两类数据：
- 离散型
- 连续型

离散型数据（Discrete Data）

用计数方式得到的数据叫离散型数据。每个数必须以整数表示。两个整数间的数值是不连续的、离散的。

- 连续型数据（Continuous Data）

重量、速度、高度、温度等可以用一定的尺度来度量。而且得到的数据可以有整数，又可有小数，中间连续不断。

1.2.4 原始数据的检查与核对（完整、真实和准确）

1、检查数据本身有无差错

记录不全（丢失、损坏、遗漏）

记载错误（笔误、虚构）

测量工具不够、测量技术不熟练

对一些特殊值（极大、极小）应反复核实，以保证计算的正确与可靠。

2、取样有无差错

取样不全或非随机取样都会影响统计的准确性。

3、数据不合理的合并

不同性别、健康状况、品种、年龄、时间与年度、试验因素不同等数据进行了不合理合并，应予以纠正。

1.3 频数表与频数图的编绘

当对某一事物做初步调查或实验时，可得到大量数据（如几百个），把这些数据按数值大小进行分组，制成频数分布表，可以看出资料的集中和变异情况，对资料得到一个初步的概念。

频数分布

- 是一种重要的分组资料，反映总体单位在各组的分布状态。

基本形式：

分组	单位数	频率
…	…	…
合计		100

频数分布

频率分布

1.3.1 离散型数据的整理

例1．调查每天出生的10个新生儿中，体重超过3公斤的人数。共调查120天。

7 7 6 6 7 5 8 8 3 9 6 7 5 8 8 8 7 8 9 8 6 7
8 7 5 7 6 8 8 7 9 5 6 8 5 9 7 7 6 8 6 8 10
8 6 8 7 9 7 7 8 7 5 8 6 7 8 7 10 8 9 7 5 8
7 6 7 8 7 5 9 8 7 7 6 7 8 5 7 9 4 7 8 75 7
7 8 9 8 7 8 6 8 4 6 8 7 5 8 6 7 9 10 86 5 7
7 8 7 6 7 8 7 7 6 6 7 8

频数：也称次数，它是落在各类别中的数据个数。

频数（率）分布:把频数（率）按组值的顺序排列起来，便得到离散型数据的频数（率）分布。

（1）分组

当所调查数据的变化较少时，以每种变数为一组。每天的10个新生儿中，体重超过3公斤的人数，可能有11种情况：0个，1个，2个，……，10个。

（2）以唱票方式记录每种情况出现的天数（频数）

组值	频数计算	频数	频率
0	0	0	0.000
1	0	0	0.000
2	0	0	0.000
3	一	1	0.008
4	丅	2	0.017
5	正正丅	12	0.100
6	正正正下	19	0.158
7	正正正正正正正下	39	0.325
8	正正正正正正下	34	0.283
9	正正	10	0.083
10	下	3	0.025
总计		120	0.999

表 1－1　每10名新生儿中体重超过3Kg的人数的频数（率）分布表

（3）为了更直观地描述数据变化规律，还可以绘成频数图。

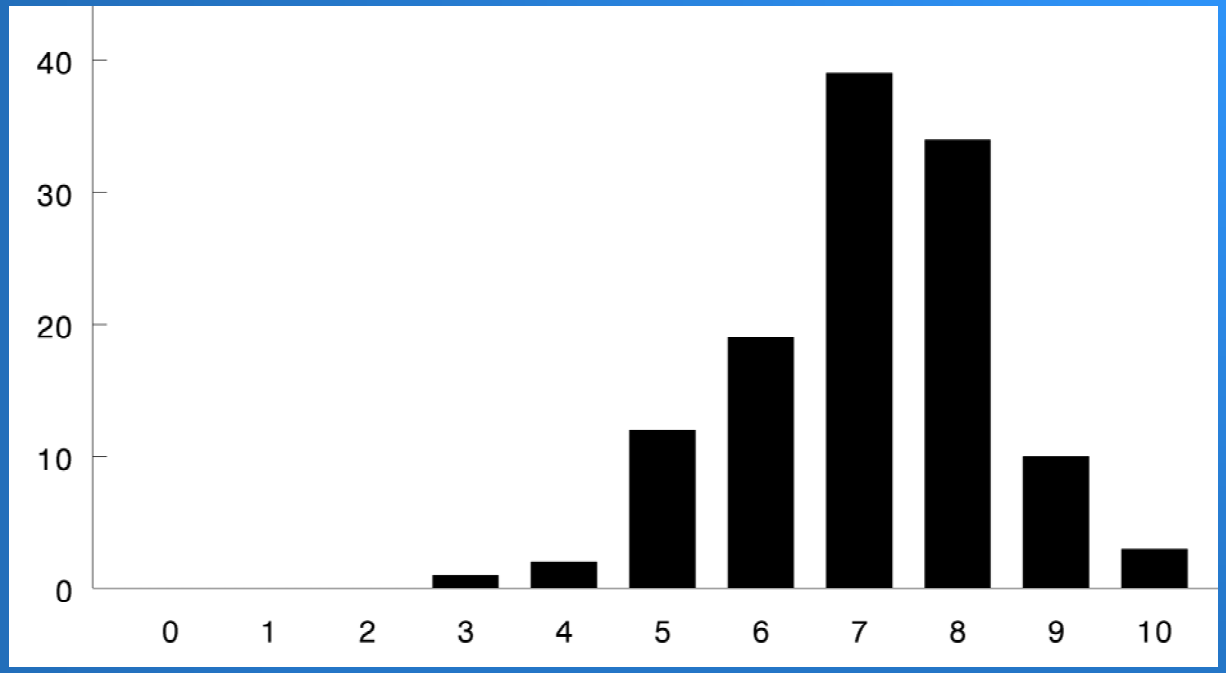


图1－2 每10名新生儿中体重超过3Kg的人数的频数分布图

1.3.2 连续型数据的整理

- **频数（率）分布：**把频数（率）按组界的顺序排列起来，便得到了连续型数据的频数（率）分布。
- 确定**组数、组距、组限，**然后按观察值的大小归组。

例：某农场做高粱“三尺三”提纯时，调查100个株高。

表1—2 “三尺三”株高测量结果

155	153	159	155	150	159	157	159	151	152	159	158	153	153	144
156	150	157	160	150	150	150	160	156	160	155	160	151	157	155
159	161	156	141	156	145	156	153	158	161	157	149	153	153	155
162	154	152	162	155	161	159	161	156	162	151	152	154	157	162
158	155	153	151	157	156	153	147	158	155	148	163	156	163	154
158	152	163	158	154	164	155	156	158	164	148	164	154	157	
165	158	166	154	154	157	167	157	159	170	158				

1、求极差（Range，R）

R=max X-min X

所有观察值中最大值与最小值的差。表示整个样本的变异幅度。

最大值**170cm**，最小值**141cm**。

R=170-141=29cm

2、确定组数和组距

组数：分组的个数

组距：每组的距离，组距= 极差 / 组数

怎样确定组数？

- 确定组数应考虑下列几个方面的要求：

（1）观察值个数的多少。一般个数多，组数也多，但还要看极差大小。

（2）极差的大小。极差大，组数多，反之则少。

（3）便于计算。

（4）能反映资料的真实面貌。既简化资料，从中得到规律性的认识，又保持资料的真实面貌。

表1-4 样本大小与组数多少的关系

样本内观察值个数	分组时组数
50	5-10
100	8-16（8-10）
200	10-20
300	12-24
500	15-30

有100个观察值，查表可分为8-16组，根据极差R=29，分为10组较合适。这样可使组距近于整数。

组距=极差/组数=29/10 =2.9≈3（cm）

3、选定组界与组中值

组界：为了使各个观察值划入一定的组内，每组应有明确的界限，称为组界。组界比观察值多一位小数。

如： 141-144　　144？　140.5-143.5

144-147　　147？　143.5-146.5

组中值/中值：位于组界中间的数值。是一组的代表值。

组限是按实验记录数据划分的每一组的上下限。

先确定第一组的中值：

通常第一组的中值以接近或等于最小观察值为好。

选第一组的中值为142。再确定第一组下界：

下界=中值-1/2组距

第一组下界=142- 3 × 1/2=140.5

当第一组的中值及下界确定后，其它各组以此为基数分别加上组距，第二组的下界为第一组的上界。

确定第一组的组值时需要考虑两个方面：

（1）第一组内次数不能过多，不能多于第二组、第三组。

（2）各组的频数以某一组最多而向两边的各组递降，因此若极端值附近的值频数很低时可适当提高第一组组值。

4. 将原始数据以唱票方式填入频数表中。

5. 计算出各组的频数和频率。

组限	组界	中值	频数计算	频数	频率
141-143	140.5-143.5	142	一	1	0.01
144-146	143.5-146.5	145	丅	2	0.02
147-149	146.5-149.5	148	正	4	0.04
150-152	149.5-152.5	151	正正丅	13	0.13
153-155	152.5-155.5	154	正正正正丅	23	0.23
156-158	155.5-158.5	157	正正正正正丅	28	0.28
159-161	158.5-161.5	160	正正正	15	0.15
162-164	161.5-164.5	163	正正	10	0.10
165-167	164.5-167.5	166	正	3	0.03
168-170	167.5-170.5	169	一	1	0.01
总计				100	1.00

表1－3　“三尺三”株高频数（率）分布表

从频数分布表中可见到的规律性：

1、植株矮的频数低，植株高的频数也低，植株中等高度的频数最高。

2、频数分布基本是两侧对称的。

3、植株平均高度在156-158厘米范围内。

组限	组界	中值	频数计算	频数	频率
141-143	140.5-143.5	142	一	1	0.01
144-146	143.5-146.5	145	丅	2	0.02
147-149	146.5-149.5	148	正	4	0.04
150-152	149.5-152.5	151	正正丅	13	0.13
153-155	152.5-155.5	154	正正正正丅	23	0.23
156-158	155.5-158.5	157	正正正正正丅	28	0.28
159-161	158.5-161.5	160	正正正	15	0.15
162-164	161.5-164.5	163	正正	10	0.10
165-167	164.5-167.5	166	正	3	0.03
168-170	167.5-170.5	169	一	1	0.01
总计				100	1.00

编制连续型数据频数（率）分布表的要点：

1、求出极差R，根据极差决定划分的组数.

2、根据极差和组数求出组距。

3、确定组界，组界是每一组实际值的上下界。

4、计算中值，中值是每一组组限的平均值。

5、以唱票的方式把原始数据添入相应的组界内，统计出每组的频数,并计算出相应的频率。

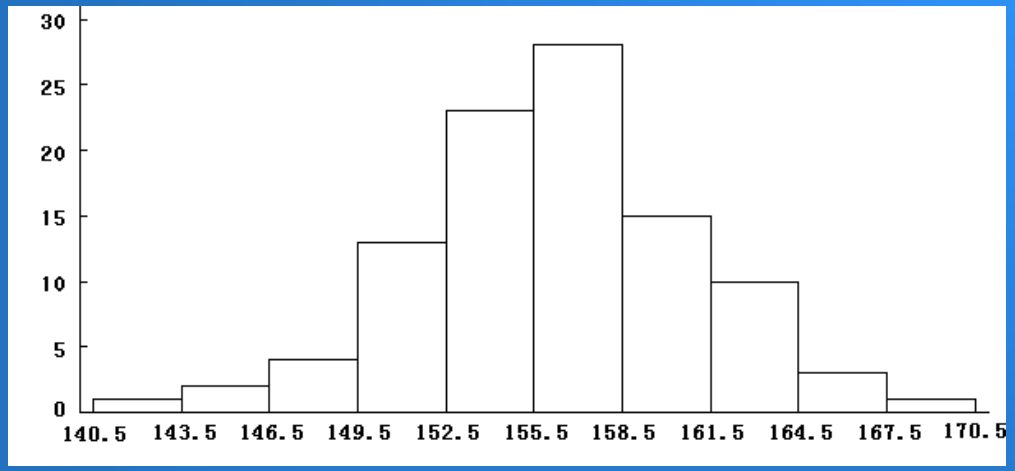
6 频数图

频数图：将频数表的结果以图形方式做更直观的描述。

三种常用频数图；

（1）直方图(Histgram)

频数



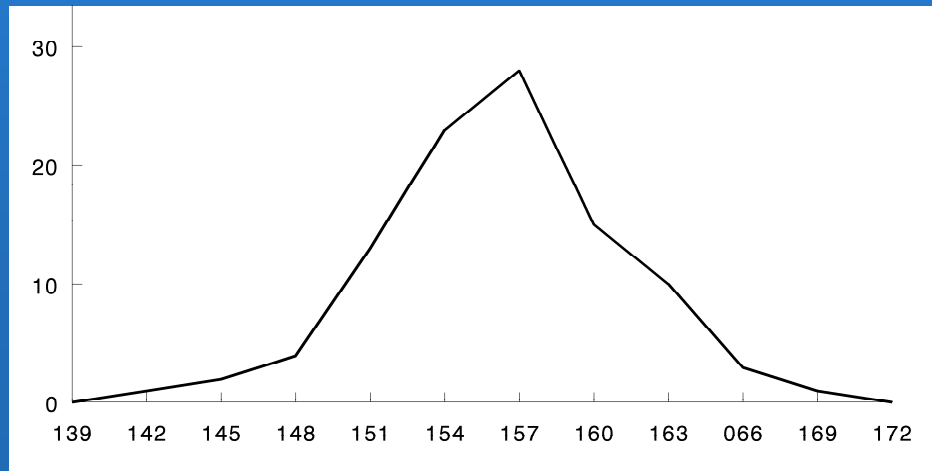
株高

图1－6 “三尺三”株高直方图

（2）多边形图(Polygon)

横轴为中值，纵轴为频数（率），

标上各点，连接各点构成多边形图。



（3）条形图

适用于离散型数据和属性资料。在横轴上标出间断的中值或分类性状，纵轴标出频。

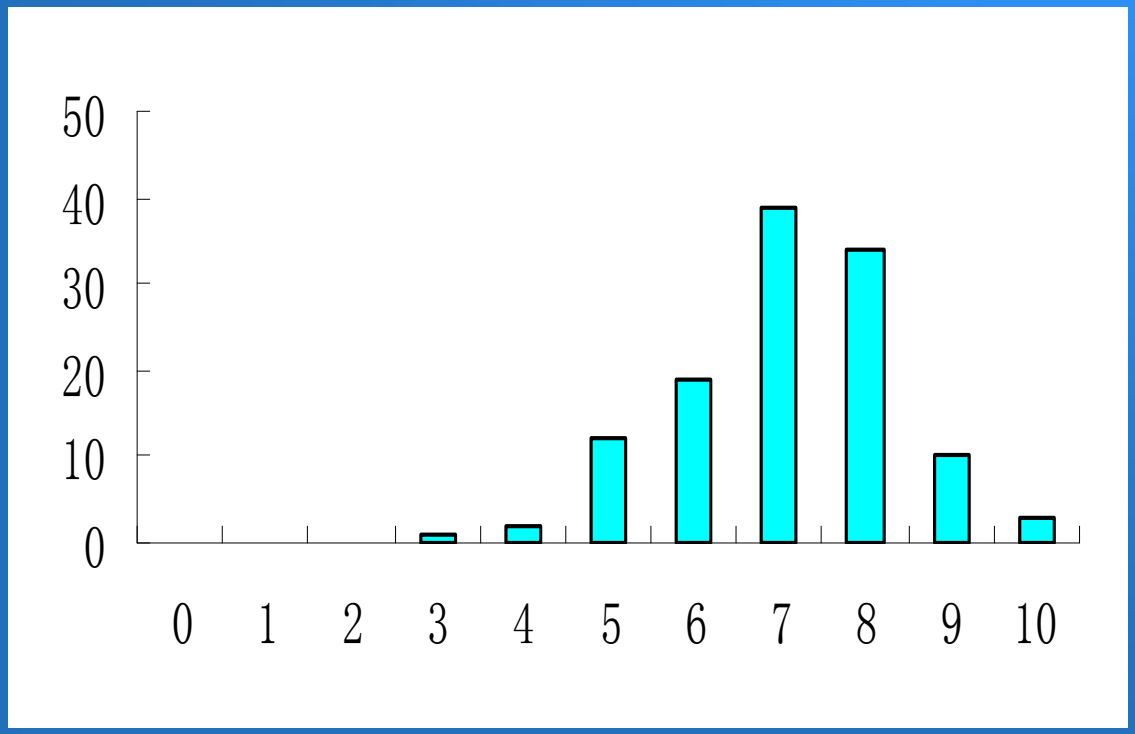


图1－8　10个新生儿中体重超过3公斤人数的频数图

（4）累积频数图

- 累积频数：将每一组的频数与前几组的频数相加。
- 累积频数图：表示某一组值以下有多少株，以及一定数量的植株在哪一高度之一。

“三尺三”株高的累积频数表

中值	频数	累积频数
142	1	1
145	2	3
148	4	7
151	13	20
154	23	43
157	28	71
160	15	86
163	10	96
166	3	99
169	1	100

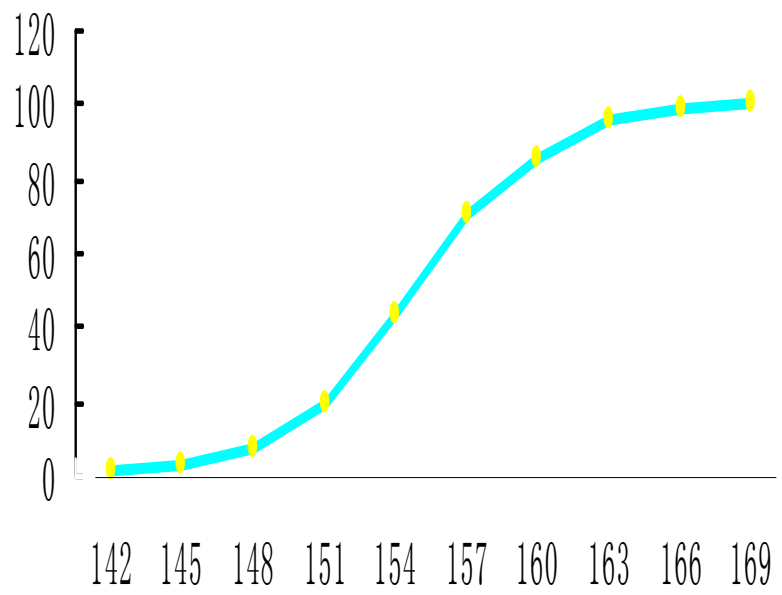


图1－9 “三尺三”株高累计频数分布图

1.3.3　研究频数（率）分布的意义

1、可以描述数据的集中点，以平均值表示。

2、可以描述数据变异的情况。

3、可以描述数据分布的形状。

4、可以显示数据中的不规则的情况。

1.3.4 数据整理的目的

1、区别不同类型，剖析结构和特点，揭示内在联系。

2、找出数据的重要特征：

a：集中情况：　b：变异情况：

3、曲线情况:变化趋势。

4、不规则情况：极端值。

1.4 生物统计的常用术语

一、总体与样本

总体(population) :根据研究目的确定的研究对象全体称为总体；

个体 (individual)：总体中的一个研究单位称为个体；

样本(sample)：总体的一部分称为样本；

样本容量(sample size)：样本中所包含的个体数目叫样本容量或大小，样本容量常记为n。通常把$n \leqslant 30$的样本叫小样本，$n > 30$的样本叫大样本。

假想总体

例如进行几种饲料的饲养试验，实际上并不存在用这几种饲料进行饲养的总体，只是假设有这样的总体存在，把所进行的试验看成是假想总体的一个样本；

Question: 研究的目的是要了解总体，然而能观测到的却是样本?

- 通过样本来推断总体是统计分析的基本特点。

Question: 样本能否推断总体？

- 有很大的可靠性但有一定的错误率这是统计分析的又一特点。

- 抽样：从总体中抽取样本这一过程称为抽样。

随机抽样：抽取个体时，使总体内所有个体有同等被抽取的机会。随机抽样也叫做概率抽样。

放回式抽样和非放回式抽样

放回式抽样：从总体中抽出一个个体，记下特征后放回总体再做第二次抽样，这种方式抽取的样本为随机样本。抽样的概率不变。

非放回式抽样：从总体中抽出个体后，不再放回，抽取的样本为非随机样本。每次抽样的概率发生变化。

随机抽样的方法

（1）抽签、拈阄(Draw lots)、扔钱币（Tossing）

（2）使用随机数字表

随机数字表是采用完全随机化方法制成，在随机抽样、随机化实验设计中应用很广。(课后阅读)

影响抽样误差的因素

样本容量

抽样方式

总体内部差异

二、参数与统计量

- **参数(parameter)：**由总体计算的特征数叫**参数；**

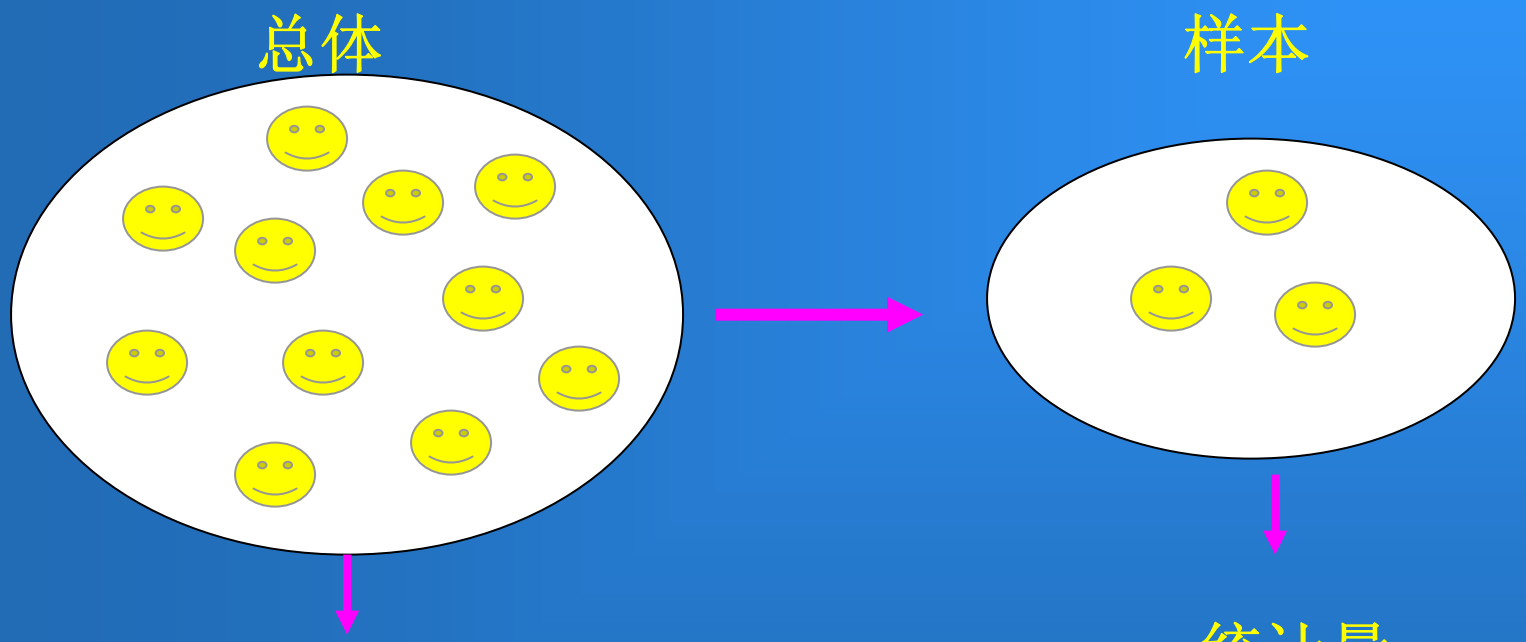
  希腊字母表示参数，例如用 μ 表示总体平均数，用 σ 表示总体标准差，总体方差 σ^2 ；

- **统计量(statistic)：**由样本计算的特征数叫**统计量**。

拉丁字母表示统计量，例如用 $\overline{X}$ 表 示样本平均数，用**S**表示样本标准差。

观察值（Observation）：个体的某一性状、特性的测定数 值叫做**观察值**。以 $X_1, X_2, \ldots\ldots X_i \ldots\ldots X_n$ 表示。同一总体内每一个体的观察值有变异。

总体参数由相应的统计量来估计，例如用 $\bar{x}$ 估计 μ ，用 S估计 σ 等。



μ	平均数	$\bar{x}$
σ	标准差	S

图1-10. 总体和样本、参数和统计量

三、随机误差与系统误差

- 随机误差：也叫 抽样误差，是由测量过程的随机变异性导致。 随机误差影响试验的精确性。统计上的实验误差指随机误差。这种误差愈小，试验的精确性愈高。

- 系统误差：也叫片面误差， 这是 由于试验动物的初始条件相差较大，饲料种类、品质、数量、饲养条件未控制相同，测量的仪器不准 、 标准试剂未经校正，以及观测、记载、抄录、计算中的错误所引起。系统误差影响试验的准确性。

近似的测量值=真实测量值+系统误差+随即误差

三、准确性与精确性

- 准确性(accuracy)：也叫准确度，观测值与其真值接近的程度。是由测量工具的精度和系统误差所决定的。
- 精确性(precision)：也叫精确度，指调查或试验中同一试验指标或性状的重复观测值彼此接近的程度。

在调查或试验中要注意：

- 应严格按照调查或试验计划进行，
- 准确地进行观测记载，力求避免人为差错，
- 特别要注意试验条件的一致性，即除所研究的各个处理外，供试材料的初始条件如品种、性别、年龄、健康状况、饲养条件、管理措施等应尽量控制一致，并通过合理的调查或试验设计努力提高试验的准确性和精确性。

由于真值μ常常不知道，所以准确性不易度量，但利用统计方法可度量精确性。

1.5 平均数与标准差

- 统计数据经过整理和显示后，对数据的形状和特征就可以有一个大致的了解。为了进一步掌握数据的分布特征和规律需进行更深入的分析，还需要找到反映数据分布特征的各个代表值。对一组数据分布的特征可以从第三方面进行描述：

分布的集中趋势，反应各数据向其中心值靠拢或聚集的程度。

分布的离散程度，反映各数据远离中心值的趋势。

分布的偏态和峰态，反映数据的分布形状。

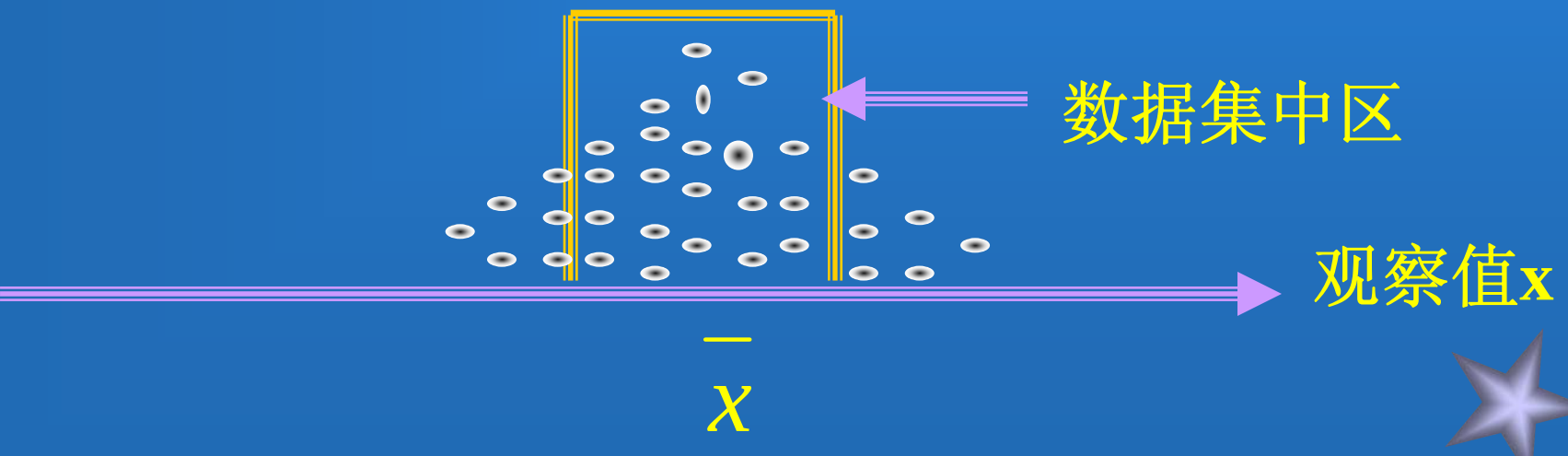
1.5.1 平均数的意义和种类

意义： 平均数是数量资料的代表值，表示资料中观察值的中心位置与集中趋势。

功用：

1. 指示资料内数据的中心位置，可以代表研究对象的一般水平

2. 可作为资料的代表而与其他资料相比较，借以明确两者间的差异。



种类：

主要有 算术平均数（Average）、

中位数（Median）、

众 数（Mode）

几何平均数（Geometric mean）.

常用的几种平均数

	概　念	计算公式	特　　点
1. 算术平均数 $\bar{x}$	标志总量与总体单位总数的比值	直接法： $\bar{x}=\frac{\sum x_i}{n}$ 加权法： $\bar{x}=\frac{\sum x_i f_i}{\sum f_i}$	性质 ① $\bar{x}$ 的大小受样本内每个值的影响； ② 若每个Xi都乘以相同的数K，则 $\bar{x}$ 也应乘以K ③离均差之和等于零； $\sum(x-\bar{x})=0$ ④离均差之平方和最小 $\sum(x-\bar{x})^2=$最小 缺点：①易受极值影响 ②在偏斜分布和U形分布中，不具有代表性

- 平均数的计算

直接法计算算术平均数

$$\bar{x}=\frac{x_1+x_2+\ldots\ldots+x_n}{n}=\frac{\sum_{i=1}^{n}x_i}{n}=\frac{\sum x}{n}$$

加权法计算算术平均数

对于样本含量 $n \geqslant 30$ 以上且已分组的资料，可以在次数分布表的基础上采用加权法计算平均数，计算公式为：

$$\bar{x}=\frac{f_1x_1+f_2x_2+\cdots+f_kx_k}{f_1+f_2+\cdots+f_k}=\frac{\sum_{i=1}^{k}f_ix_i}{\sum_{i=1}^{k}f_i}=\frac{\sum fx}{\sum f}$$

式中：x_i —第i组的组中值；

f_i —第i 组的次数；

k —分组数

	概 念	计算 公 式	特 点
2. 中位数（M_e）	把一组观察值按大小顺序排列，位次居中的那个数值即中位数	***n***为奇数时 $Md = x_{(n+1)/2}$ ***n***为偶 数 时 $M_d = \frac{x_{n/2} + x_{(n/2+1)}}{2}$	优点：①容易理解，②不受极值影响③中位数是一种位置上的平均数。 缺点：①灵敏度和计算功能差

	概　念	特　　点
3. 众数（M_o）	资料中最常见、频数最多的一个数，或次数最多一组的组中值。	优点： ①容易理解， ②不受极值影响 缺点： ①灵敏度和计算功能差 ②稳定性差 ③具有不唯一性

	概 念	计算公式	特 点
4. 几何平均数 ($\overline{x}_G$)	几个变量值连乘积的几次根。	$G = \sqrt[n]{x_1 x_2 \ldots\ldots x_n}$	优点: ①用于几何平均数的观察值往往彼此相差较大，成倍数关系 缺点:①有“ 0”或负值时不能计算

要点解释

权数

权数（**Weighted**），是分布数列中的频数或频率。对求平均数具有权衡轻重的作用，是影响平均数变动的两个因素之一（另一因素是变量值）。

例

(1) $\bar{x}$ =5

X	频数	频率(%)
4	10	25.0
5	20	50.0
6	10	25.0
合计	40	100.0

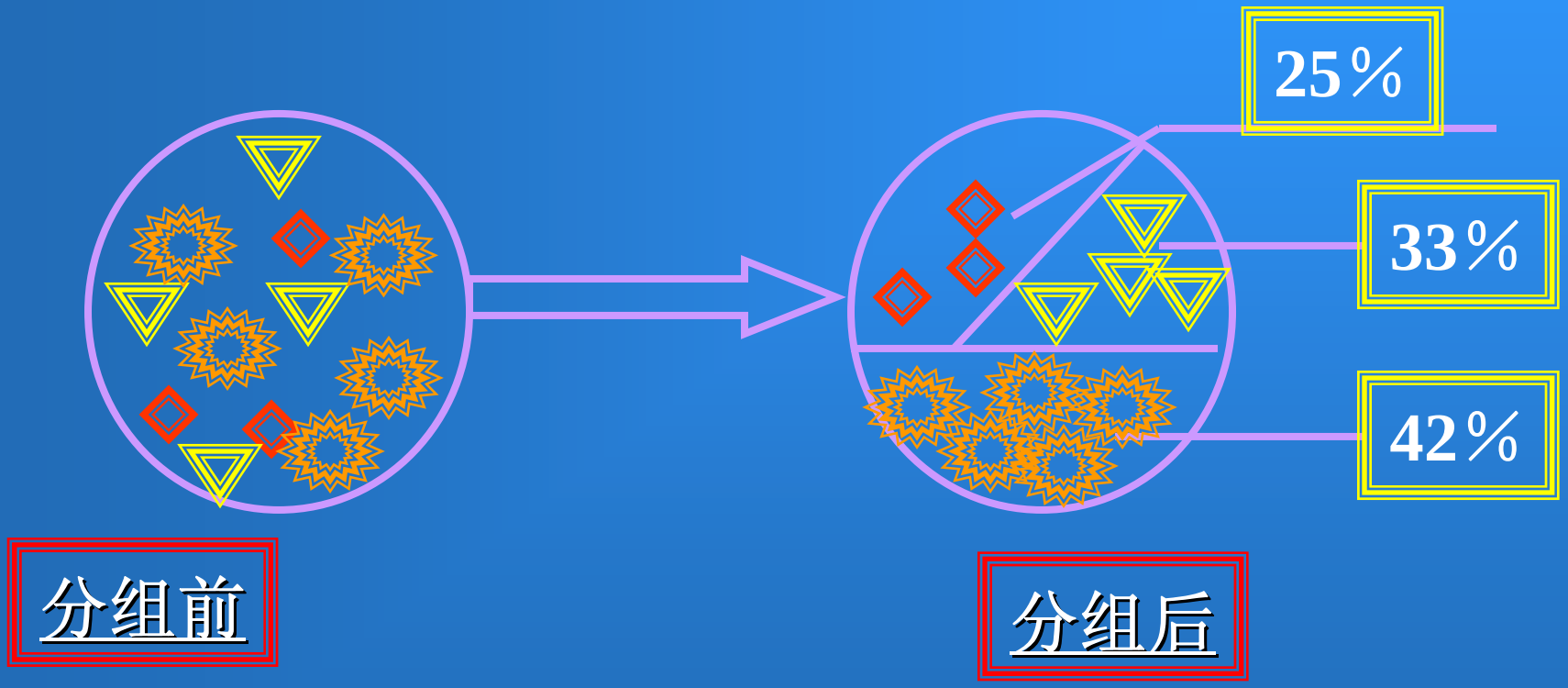
(2) $\bar{x}$ =5

X	频数	频率(%)
4	20	25.0
5	40	50.0
6	20	25.0
合计	80	100.0

(3) $\bar{x}$ =4.75

X	频数	频率(%)
4	20	50.0
5	10	25.0
6	10	25.0
合计	80	100.0

分组



课前提问

- 平均数的意义和种类
- 算术平均数的性质
- 频数资料平均数的计算
- 标准差的意义、性质
- 实验数据的科学表示法
- 变异系数的意义、应用条件

意义：

平均数是数量资料的代表值，表示资料中观察值的中心位置与集中趋势。

种类：

主要有 算术平均数（Average）、

中位数（Median）、

众 数（Mode）

几何平均数（Geometric mean）.

算术平均数($\bar{\chi}$) 的性质

（1） $\bar{\chi}$ 的大小受样本内每个值的影响

（2）若每个X_i都乘以相同的数K，则 $\bar{\chi}$ 也应乘以K

（3）若每个Xi都加上相同的数A，则 $\bar{\chi}$ 也应加上A。

（4）离均差之和等于零。

（5）离均差之平方和最小。

本次内容

- 1.6 平均数与标准差
- 第二章 概率分布

例　　题

例1：调查9个某种传染病人，他们的潜伏期（天）分别为：2、5、4、3、3、6、16、9、3，求中位数。

将数据按大小顺序排列，得：2、3、3、3、4、5、6、9、16，n=9是奇数，所以第5位的4就是中位数。平均潜伏期4天。

（ $\bar{x}$ = 5.67天）

例：随机抽取某轧钢厂工人20名，测得其白细胞棘突百分比如下，求中位数。

0 0 0 0 1 2 2 4 4 5 6 6 6 7 8 9 10 11 13 14

n=20 是偶数，中位数是第10和第11位观察值的算术平均数

$$M_d = \frac{5+6}{2} = 5.5$$

$$\bar{x} = 5.4$$

例 2 鞋码： 38 39 40 40 40 41 Mo=40

38 39 39 40 41 41 Mo=39、41

38 39 40 41 42 43 无Mo

例3 有5人，其血清抗体效价分别为1:10，1:100，1:1000，1:10000，1:100000，求其效价倒数的平均水平。

如果用算术平均数计算，

$$\bar{x}=\frac{10+100+1000+10000+100000}{5}=22222$$

$$G=\sqrt[n]{x_1x_2\ldots\ldots x_n}$$

$$G=\lg^{-1}3=1000$$

例4 有 2 20 25三个数，其几何平均数为

$$\sqrt[3]{2\times20\times25}=\sqrt[3]{1000}=10$$

计算若干个来自同一总体的样本平均数的平均数时，如果样本含量不等，也应采用加权法计算。

例 5　某牛群有黑白花奶牛 1500头，其平均体重为750 *kg*，而另一牛群有黑白花奶牛1200头，平均体重为725 *kg*，如果将这两个牛群混合在一起，其混合后平均体重为多少？

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{750 \times 1500 + 725 \times 1200}{2700} = 738.89(kg)$$

对同一资料，几种主要的平均数之间的关系

算术平均数>几何平均数>位置平均数

- 请注意：重点
- ↓
- 算术平均数。

1.6.2 标准差 (Standard deviation)

在研究时只考虑数据的集中程度而不考虑变异性就不能很好的刻画数据的特征.

度量数据的离散程度：

平均离差

标准差

方差

极差

变异系数

常用的几种标志变异指标

	概　念	计　算	特　点
1．极差（R）	数列中最大值与最小值之差	R=最大值-最小值	优点：容易理解，计算方便 缺点：不能反映全部数据分布状况
2．平均离差（A、D）	各观察值与均值离差绝对值的算术平均	简单：$A\cdot D=\frac{\sum\lvert X_i-\bar{X}\rvert}{n}$ 加权：$A\cdot D=\frac{\sum\lvert X_i-\bar{X}\rvert f_i}{\sum f_i}$	优点：反映全部数据分布状况 缺点：取绝对值，数字上不尽合理

	概　念	计　算	特　点
3.方差（S^2, σ²）和 标准差（S, σ）	各标志值与均值离差平方的平均。 方差的平方根（取正根）	简单：$\sigma^2=\frac{\sum(X_i-\bar{X})^2}{n}$ 加权：$\sigma^2=\frac{\sum(X_i-\bar{X})^2 f_i}{\sum f_i}$ $\sigma=\sqrt{\sigma^2}$	优点：反映全部数据分布状况. 缺点：受计量单位和平均水平影响，不便于比较
4. 变异系数（CV）	标准差与均值之商，可以用来比较不同资料的相对变异程度	$CV=\frac{S}{\bar{X}}$	优点：适宜不同数据的比较,是无量纲的系数 缺点：对数据结构变化反应不灵敏

方差（σ²）和标准差（σ）是应用最广的标志变异指标

$$(x-\bar{x})$$

$$\Downarrow$$

$$\sum(x-\bar{x})=0 \Rightarrow$$

它不能表示整个资料中所有观察值的总偏离程度

$$\Downarrow$$

$$\sum|x-\bar{x}| \Rightarrow$$

使用不方便，在统计学中未被采用

标准差

$$\downarrow$$

$$s=\sqrt{\sum(x-\bar{x})^2/(n-1)}$$

$$\Uparrow$$

$$(x-\bar{x}) \Rightarrow (x-\bar{x})^2 \Rightarrow \sum(x-\bar{x})^2 \Rightarrow \sum(x-\bar{x})^2/(n-1)$$

$$\downarrow$$

消除离均差的负号

$$\downarrow$$

离均差的平方之和（简称**平方和**，记为*SS*）

$$\downarrow$$

称为**均方**（缩写为*MS*），又称为**样本方差**，记为S^2

标准差 (Standard deviation)

- 意义:表示每个观察值与平均值的平均变异程度，反映观察值分布的离散趋势。
- 公式

（一）直接法

$$S=\sqrt{\frac{\sum\left(x_i-\bar{x}\right)^2}{n-1}}$$

$$s=\sqrt{\frac{\sum x^2-(\sum x)^2/n}{n-1}}$$

（二）加权法

$$s=\sqrt{\frac{\sum f_i(x_i-\bar{x})^2}{\sum f_i-1}}=\sqrt{\frac{\sum f_i x_i^2-(\sum f_i x_i)^2/\sum f_i}{\sum f_i-1}}$$

标准差的性质：

（1）标准差大小受观察值影响，其单位与观察值相同。

（2）在计算时，对各观测值加上或减去一个常数，其数值不变。

（3）当每个观测值乘以或除以一个常数a，则所得的标准差是原来标准差的a倍或1/a倍。

（4）在资料服从正态分布的条件下，约有68.26%的观测值在平均数左右一倍标准差（$\bar{x}$ ±S）范围内；约有95.43%在平均数左右两倍标准差（$\bar{x}$ ±2S）范围内；约有99.73%在平均数左右三倍标准差（$\bar{x}$ ±3S） 范 围内。也就是说全距近似地等于6倍标准差，可用（全距/6）来粗略估计标准差。

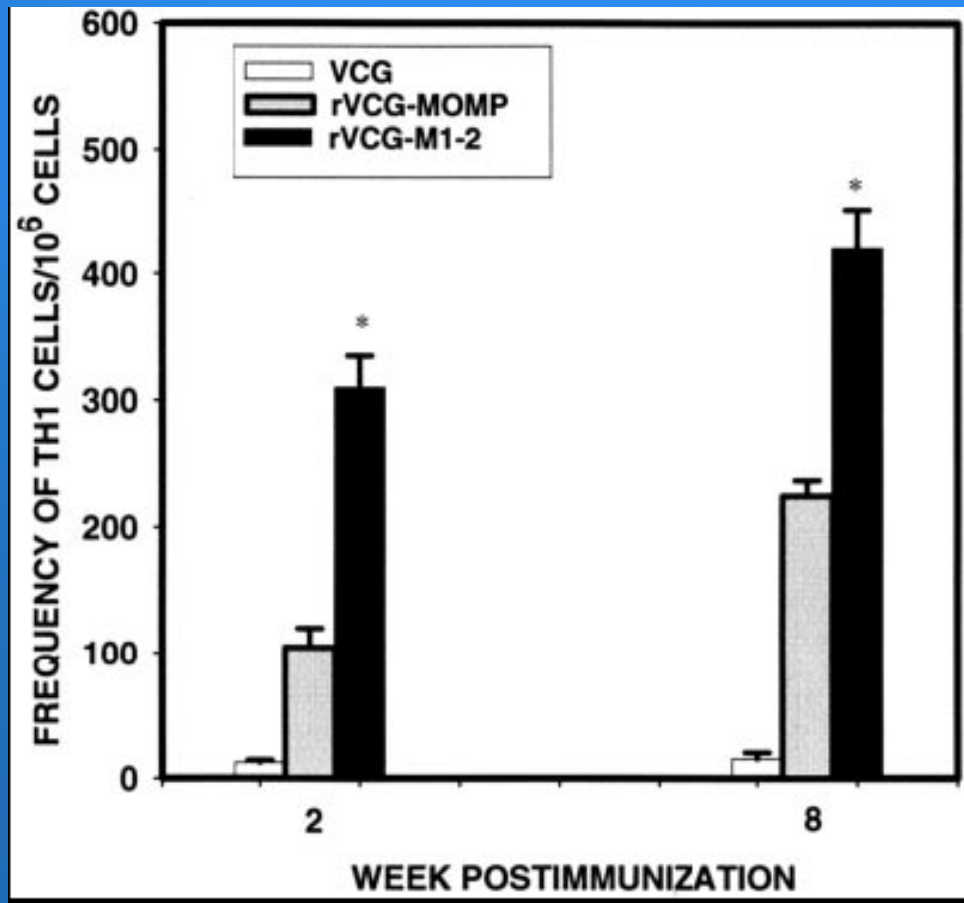
实验数据的科学表示法：

$\bar{x} \pm s$ 或 $\bar{x} \pm SD$

Table 1
Fresh ovarian weight and preovulatory follicles from infected and uninfected rabbits

Groups	Left ovary (g)	Right ovary (g)	Preovulatory follicles
Controls	0.32 ± 0.04	0.33 ± 0.04	6.5 ± 1.7
MV-ZPB Day 5	0.22 ± 0.05*	0.25 ± 0.06	5.0 ± 2.2
MV-ZPB Day 15	0.15 ± 0.05**	0.13 ± 0.06**	4.3 ± 1.9
MV-HA Day 15	0.16 ± 0.07**	0.16 ± 0.06**	4.3 ± 1.3
MV-ZPB Day 30	0.19 ± 0.05**	0.19 ± 0.07*	2.3 ± 1.5†
MV-HA Day 30	0.13 ± 0.03**	0.14 ± 0.02**	5.8 ± 1.3

Notes: the values are mean ± SD for four rabbits in each group.
* $P < 0.05$ against uninfected controls.
** $P < 0.01$ against uninfected controls.
† $P < 0.05$ against infected (MV-HA Day 30) and uninfected controls.



4．变异系数（Coefficient of variance，CV）

标准差可以反映数据的离散程度，

但对于**平均水平不同或计量单位不同的不同样本**，是不能用标准差直接比较离散程度。

$$CV = \frac{S}{\bar{X}}$$

注意，变异系数的大小，同时受平均数和标准差两个统计量的影响，因而在利用变异系数表示资料的变异程度时，最好将平均数和标准差也列出。

两组样本的 $\bar{X}$ 接近时可用S比较它们的变异性大小，极差较大时不可。例：10岁身高 A=120±5.0cm

8个月身高 B=70±4.0cm

$$cv = \frac{s}{\bar{x}} \qquad cv_1 = \frac{s}{\bar{x}} = \frac{5}{120} = 0.042 \qquad cv_2 = \frac{4}{70} = 0.057$$

当两个样本的观察值单位不同时，也不能直接比较。

即两种性状，如身高、体重

例：大学男生身高：165±12cm CV_1=0.073

体重：58.3±4.8kg CV_2=0.082

恭喜大家！

- 你们现在已基本入门。下一章的难度要大些，

一定要加油喔！

例 计算以下数据的标准差：26 25 28 24 23 25 27 27 30 21。

解 最好列成以下表格的形式计算

	x	x^2
	26	676
	25	625
	28	784
	24	576
	23	529
	25	625
	27	729
	27	729
	30	900
	21	441
$\sum x$	256	6614

$$s=\sqrt{\frac{\sum x_i^2-\frac{(\sum x_i)^2}{n}}{n-1}}$$

=2.59

某纯系蛋鸡200枚蛋重资料次数分布及标准差计算

组别	组中值（x）	次数（f）	fx	fx^2
44.15—	45.0	3	135.0	6075.0
45.85—	46.7	6	280.2	13085.34
47.55—	48.4	16	774.4	37480.96
49.25—	50.1	22	1102.2	55220.22
50.95—	51.8	30	1554.0	80497.20
52.65—	53.5	44	2354.0	125939.00
54.35—	55.2	28	1545.0	85317.12
56.05—	56.9	30	1707.0	97128.30
57.75—	58.6	12	703.2	41207.52
59.45—	60.3	5	301.5	18180.45
61.15—	62.0	4	248.0	15376.00
合计		$\Sigma f=200$		Σ

利用某纯系蛋鸡200枚蛋重资料的次数分布表（见表）计算标准差。

将表中的Σf、Σfx、$\sum fx^2$代入上式得：

$$S=\sqrt{\frac{\sum fx^2-(\sum fx)^2/\sum f}{\sum f-1}}=\sqrt{\frac{57550.711-107051^2/200}{200-1}}=3.55(g)$$

即某纯系蛋鸡200枚蛋重的标准差为3.55g。