第6章
数量性状的遗传统计学基础

王建康
中国农业科学院作物科学研究所
wangjiankang@caas.cn
http://www.isbreeding.net

本章的主要内容

- §6.1 数量性状的遗传学基础
- §6.2 数量性状的概率论基础
- §6.3 数量性状的数理统计基础

§6.1 数量性状的遗传学基础

- §6.1.1 质量性状和数量性状
- §6.1.2 数量性状遗传的纯系理论
- §6.1.3 数量性状遗传的多基因假说

离散变异与质量性状

- 孟德尔通过一系列精心设计的试验，证明了他的颗粒遗传理论。孟德尔的科学贡献不仅仅是建立了遗传学基本规律，他为建立遗传学基本理论而采用的科学试验方法在后来整个生命科学的研究中都发挥着至关重要的作用（Allard, 1999）。
- 为研究一个遗传分离群体中不同表型所占的比例（即分离比），要求个体在所调查的性状上有足够大的差异，以便根据性状表型对个体进行明确的分类。在遗传研究中，具有明显表型分类的变异称为不连续变异（discontinuous variation）或离散型变异（discrete variation），具有不连续变异的性状称为质量性状（qualitative trait）。

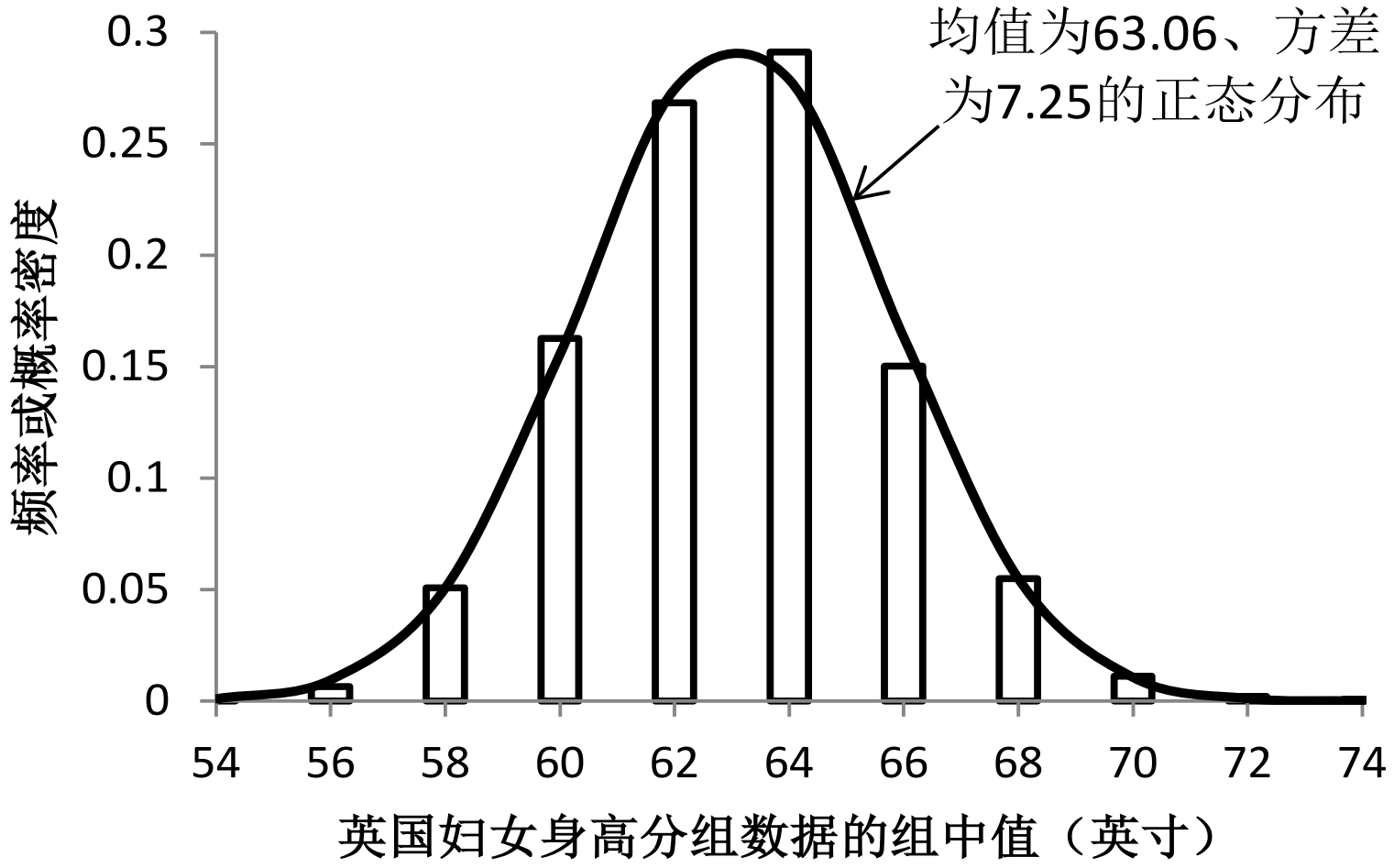
连续变异与数量性状

- 除间断性变异外，还存在大量的变异，无法或难以进行明确的表型分类。没有明显表型分类的变异称为连续变异（continuous variation），具有连续变异的性状称为数量性状（quantitative trait）。
- 数量性状的 遗传研究也有很长的历史。达尔文在他的进化论研究中就涉及了大量的数量性状；F. Galton于19世纪末在英国建立研究小组，主要研究人类群体中数量性状的遗传，在这些研究中发明了统计学中的回归和相关分析方法，并于1889年出版《Natural Inheritance》一书。

英国妇女身高（英寸）的次数分布

分组区间	组中值/in（x）	人数	频率（f）
53~55	54	5	0.0010
55~57	56	33	0.0066
57~59	58	254	0.0508
59~61	60	813	0.1628
61~63	62	1340	0.2682
63~65	64	1454	0.2911
65~67	66	750	0.1502
67~69	68	275	0.0551
69~71	70	56	0.0112
71~73	72	11	0.0022
73~75	74	4	0.0008
均值、方差、标准差	63.06、7.25、2.69		

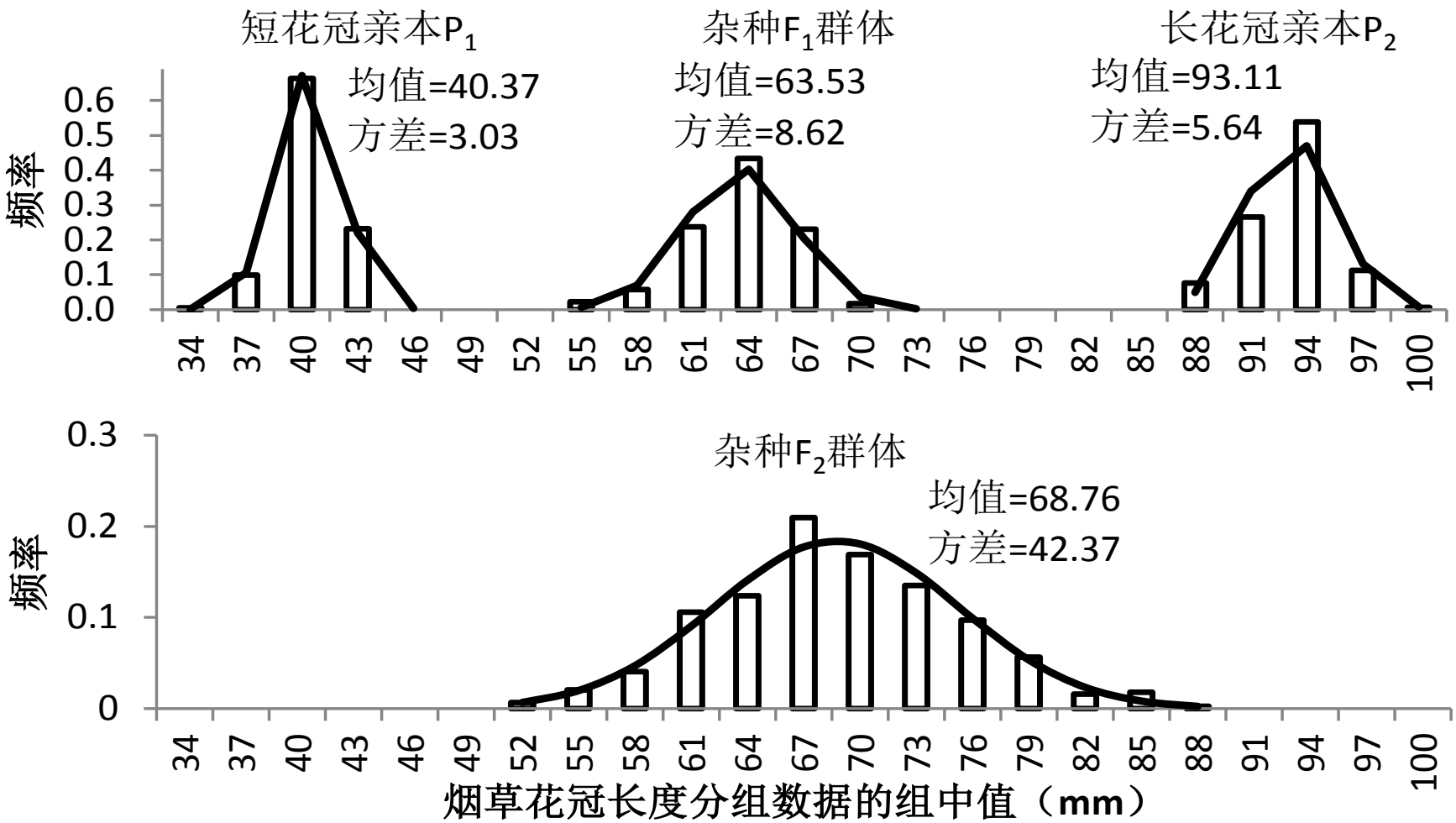
英国妇女身高的频率分布柱形图和正态拟合曲线



数量性状在纯系杂交后代中的分布

- 1900年孟德尔杂交试验被重新发现后不久，人们在其它自花授粉物种中，同时开展了大量类似豌豆的杂交试验，其中也包括异化授粉物种、经过多代自交形成的自交系之间的杂交试验，如玉米。
- 这些杂交试验进一步验证了孟德尔遗传规律（§1.1.3和§1.1.4）在质量性状中的普适性。但同时也发现，对于大量具有连续性变异的数量性状来说，难以对表型进行分组，从而难以观测到简单的孟德尔分离比。

East（1916）烟草杂交试验中花冠长度（mm）的分布



数量性状遗传的纯系理论

- 在East（1916）的杂交试验中，如果对三个调查年份分别进行计算，得到亲本P_1的平均花冠长度为40.46、40.61和39.76mm，亲本P_2的平均花冠长度分别为93.22、93.37和92.13mm。每个亲本的平均花冠长度在年份之间几乎没有差异。
- 短花冠亲本自交后代的平均花冠长度与短花冠亲本相似，长花冠亲本自交后代的平均花冠长度与长花冠亲本相似，花冠的长短在两个亲本的自交后代中都能稳定地遗传下去。
- 数量性状在纯合基因型的自交后代中能够稳定遗传，这一现象称为纯系理论（pure line theory）。纯系理论最早由W. L. Johannsen（1903）在菜豆（*Phaseolus vulgaris*）粒重性状的遗传研究中提出来。

Johannsen的纯系理论试验

- Johannsen的系列试验从1900年种植一袋8kg菜豆品种‘Princess’开始，这份从市场上得到的种子在大小和粒重上有很大差异。
- 1901年，Johannsen根据种子大小和粒重收获了287个单株，单株粒重呈连续性变异，没有明显的分组趋势。在以后几年的试验中，Johannsen详细记录了亲代和子代的粒重。结果发现粒重高的亲本，它们后代的平均粒重也较高；粒重轻的亲本，它们后代的平均粒重也较轻。
- 这其实与Galton在人类身高和体重等数量性状的遗传研究中，观测到的亲子之间的相关结果是一致的。这些结果都充分表明，数量性状是可以遗传的，对数量性状的选择也是有效的。

纯系间差异不是单基因控制的

- 根据亲子之间的粒重数据，Johannsen随后建立了19个不同籽粒大小的纯系。以厘克（cg）为单位，第1个纯系最重，平均粒重达64.2cg；第19个纯系最轻，平均粒重仅为35.1cg。这19个纯系的平均粒重从低到高呈连续性变化，没有明显的分组界限，说明粒重不可能是单基因控制的质量性状。

纯系内的差异不能遗传到下一代

- 除研究粒重在纯系之间的遗传外，Johannsen还研究了纯系内不同粒重个体产生后代的平均粒重。
- 他一开始选择了平均粒重为45.5cg的第13个纯系，从中获得4组粒重分别为20、30、40、50cg的种子，种植后发现它们后代的平均粒重分别为47.5、45.0、45.1、45.8cg，均接近于第13个纯系的平均粒重45.5cg。
- 随后他对粒重最重的第1个纯系，连续6个世代选择粒重最重和最轻的种子。两个方向选择得到的最终平均粒重分别为69cg和68cg，表明纯系内的选择是无效的。

粒重在纯系间和纯系内的亲子相关

- Johannsen还利用相关系数来估计表型变异中可遗传到后代的比例。在这19个纯系中，得到亲子之间的相关系数为0.336±0.08，说明粒重在家系间的表型变异中，大约有1/3是可遗传的。
- 从§6.3可以看出，纯系群体中后代表型均值和亲代表型之间的回归系数，等于性状的遗传力，它们之间的相关系数等于遗传力的平方根。单个后代表型和亲代表型之间的回归系数和相关系数均等于性状的遗传力。因此，也可以认为这19个纯系构成的群体中，粒重的个体水平遗传力等于0.336。
- 利用第13个纯系得到亲子之间的相关系数只有0.018±0.038，遗传力几乎为0，说明粒重在纯系内不具有遗传性。

纯系理论在数量遗传中的作用

- 纯系理论：数量性状的表型是基因和环境共同作用的结果，基因型的作用是可以遗传的，环境的作用是不能遗传的。
- Johannsen的纯系理论，将数量性状变异区分为可遗传的变异与非遗传的变异，并首次明确了基因型（genotype）和表现型（phenotype）的概念，这为理解连续性变异的遗传规律提供了依据，也为随后连续性变异多基因假说的形成起到了促进作用。

纯系理论的线性模型表示

- 对于n个纯合基因型构成的一组纯系来说，下面的等式给出基因型效应和随机环境效应对表型的线性模型。
- 其中，μ表示这些纯系的平均表现，G_i表示纯系i的基因型效应，ε_i代表环境效应及其他不可预测的随机误差效应。

$$P_i = \mu + G_i + \varepsilon_i, \quad i=1, 2, \ldots, n$$

纯系理论的线性模型表示

$$P_i = \mu + G_i + \varepsilon_i，\ i=1, 2, \ldots, n$$

- 模型中，μ+G_i可以看作每个纯系的表型均值或表型平均数。
- 纯系的自交后代与亲代有相同的基因型，基因型值G_i能够完整地遗传给后代。除了环境和随机误差外，后代表型与亲代是完全一致的。因此，后代的表型与亲代的表型具有一定的相关性。
- 就粒重这一数量性状来说，环境效应中既有年份间的差异，也有籽粒在豆荚中生长部位、豆荚在植株上生长位置、植株生长的土壤和水分等方面的差异，这些效应都不能遗传给下一代。

纯系理论的拓广

- 对于无性繁殖物种来说，任何一个基因型的无性系后代都与亲代有着相同的基因型，也因此与亲代有着相同的基因型效应。对于一个杂交品种来说，生产上每年种植的杂交种，都是同样的两个纯系的杂交后代，不同年份、不同地块种植的杂交种具有相同的基因型，也因此具有相同的基因型效应。所以，广义地讲，纯系理论对于无性繁殖物种和杂种品种这些杂合基因型来说也是适用的。
- 换句话说，无性系品种之间或杂种品种之间的差异，是由基因型差异引起的，是可以遗传的，选择是有效的；一个无性系品种或杂种品种内的个体差异，是随机环境误差造成的，是不能遗传的，选择是无效的。
- 在随机交配群体中，亲本个体的基因型是杂合的；每个后代个体具有雌雄两个亲本，它们对后代的贡献各占一半；后代的基因型与亲本不再保持一致。因此，随机交配后代的基因型值不再简单地等于亲本的基因型值。

数量性状是否服从孟德尔遗传规律之争

- 以K. Pearson为首的生物统计学派认为，连续变异是进化的重要原因；由于在连续变异中无法观察到简单的孟德尔分离比，认为孟德尔规律不适用于连续变异。
- 以W. Bateson为首的孟德尔学派认为，不连续变异是进化的重要因素；由于在连续变异中无法观察到简单的孟德尔分离比，认为连续变异不服从孟德尔规律，也因此是不能遗传的。
- 两派都有大量证据来支持各自的观点，但又难以说服对方。例如，Johannsen的菜豆粒重试验，就充分说明了数量性状的可遗传性。因此，笼统地说数量性状不能遗传是站不住脚的。

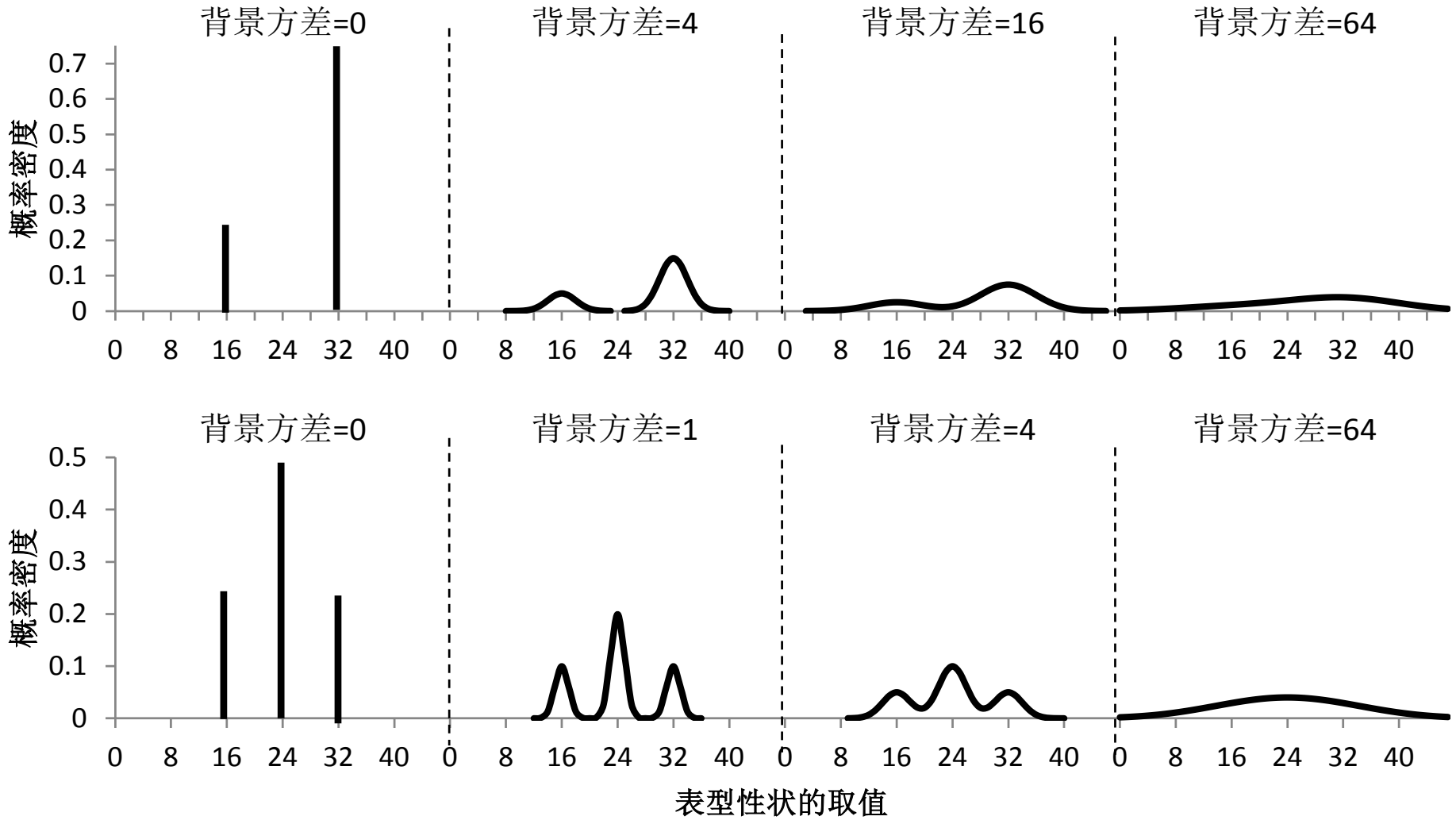
争论的解决方案

- Yule（1906）就明确指出，若是连续变异是由很多效应较小且相似的基因控制的话，孟德尔的颗粒遗传与连续变异性状的遗传之间就不会存在任何矛盾。
- 在这场争论的过程中，人们对大量数量性状开展了类似孟德尔的杂交试验，试图通过试验来验证数量性状的遗传是由多个基因共同控制的。Nilsson-Ehle（1909）的小麦粒色杂交试验表明，一个性状可以受多个孟德尔遗传因子共同控制，同时提出了数量性状的多因子假说。这一假说随后由East在烟草花冠长度和玉米穗长等杂交遗传试验中得到证实。

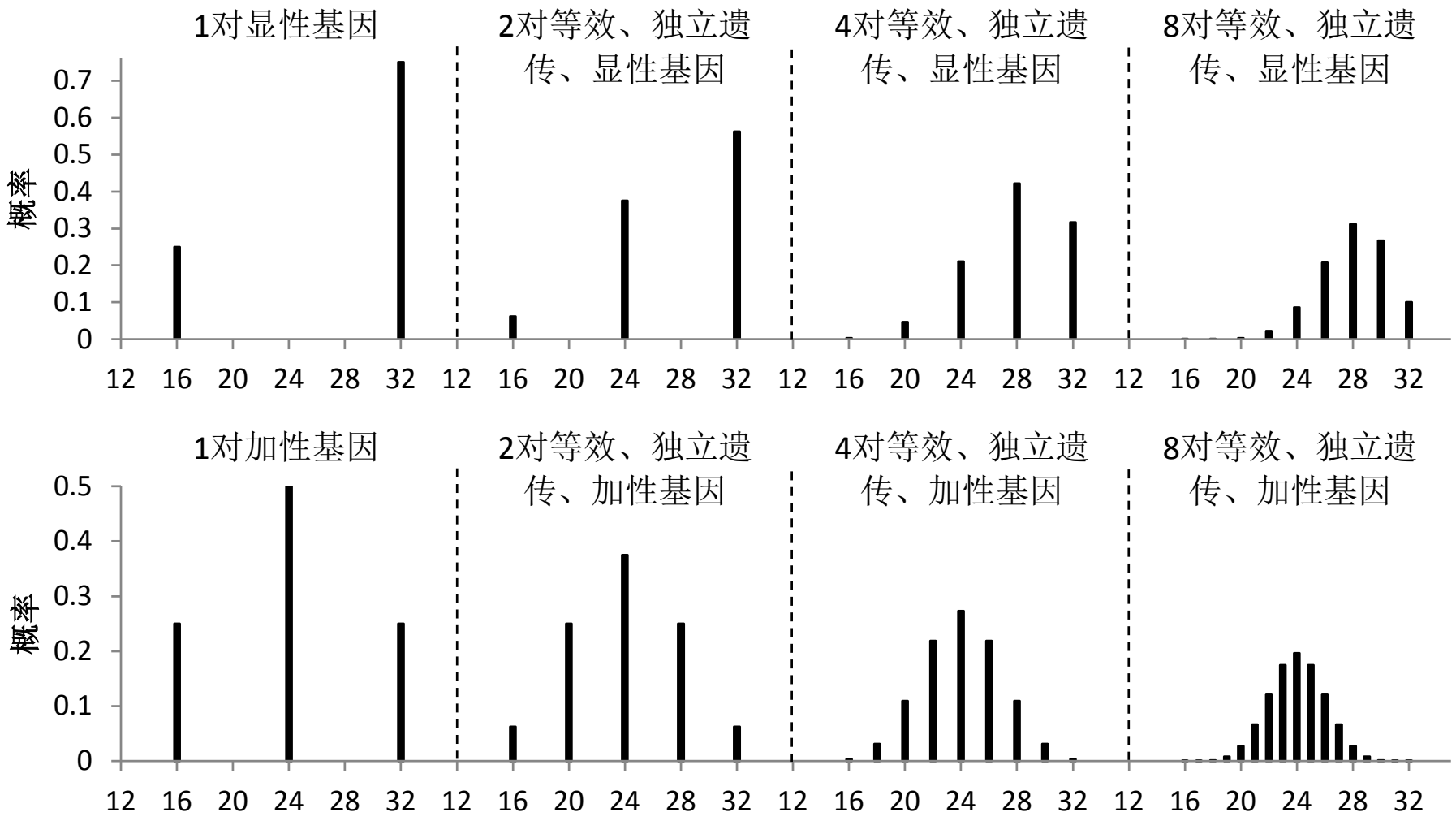
争论的终结

- R. A. Fisher（1918）“孟德尔遗传条件下的亲属间相关（*The correlation between relatives on the supposition of Mendelian inheritance*）”一文的发表，这场争论被宣告结束。
- Fisher在这篇文章中，进一步明确了数量性状多因子假说（Multi-factorial hypothesis）这一理论，将数量性状的遗传纳入孟德尔遗传的轨道，从而使两个学派的观点得到统一。
- 不仅如此，Fisher还在承认孟德尔遗传规律的前提下，提出了数量性状遗传分析的一般方法，即方差分析。这篇文章中关于遗传方差分解的基本原理和思想，构成了现代统计学中方差分析的理论基础。

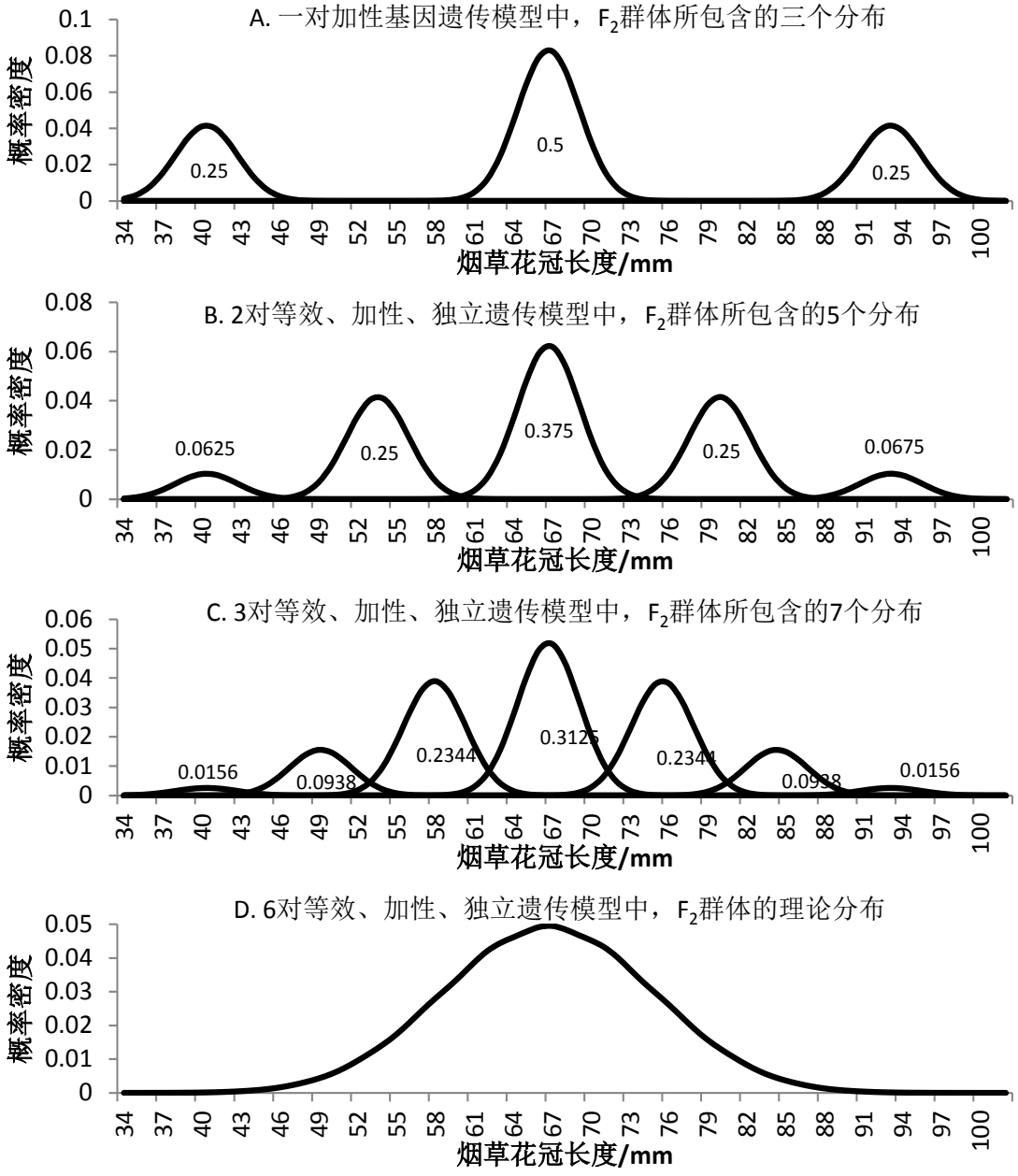
一个基因座位在不同的背景方差下，F_2群体的理论表型分布



一对和多对等效独立遗传模型中，F_2群体包含各种表型的理论频率



多基因和环境共同作用下，表型正态分布的必然性！



数量性状的取值类型

- 与简单孟德尔遗传的质量性状相比较，多基因控制性状的遗传更为复杂，有时甚至还可能存在基因间的连锁和上位型互作。为强调多基因控制性状在遗传上的复杂性，有时也把多基因控制的性状称为复杂性状（complex trait）。
- 有些数量性状在一定范围内可以取任意实型数值，如人类身高和体重、动物的产奶量、作物的产量和品质等。有些数量性状在一定范围内只能取整型数值，例如动物每胎产仔数、作物中的分蘖数、小穗数和穗粒数等。
- 还有一些人类和动物的疾病性状，表型只有有病和无病两种状态。是否发病是由感病风险（或易感性）决定的，当易感性超过一定的阈值后，就表现为有病；低于一定的阈值时，就表现为无病。这类性状又称阈值性状（threshold trait）。

数量遗传的多基因假说

- 数量遗传学（quantitative genetics）又称为遗传统计学（genetic statistics）或统计遗传学（statistical genetics），是研究数量性状遗传变异规律的一门学科。经典数量遗传学建立在多基因假说基础之上。
- 多基因假说认为，控制数量性状的基因座位较多、基因效应微小、基因间是独立遗传的；与控制质量性状的单基因一样，多基因也同样存在于染色体上，也同样服从孟德尔的分离和自由组合定律、以及连锁和交换定理；同时，数量性状还容易受环境的影响。

数量遗传的研究方法

- 数量性状在多基因和环境的共同作用下，没有明显的表型分组趋势，在分离群体中一般呈连续型的正态分布。尽管无法通过简单的分离比分析进行遗传研究，但是作为群体，我们可以根据统计学的原理和方法，计算群体的一些统计学参数。
- 例如，对于单个性状来说，可以计算它的均值和方差，以及亲子之间的协方差、相关系数和回归系数；对于多个性状来说，可以计算它们之间的协方差矩阵和复相关系数。然后通过不同遗传群体之间统计学参数的变化和关系，来研究数量性状的遗传规律。这其实就是20世纪80年代之前，数量性状遗传研究的通用方法。

数量性状的基因定位

- 传统数量遗传学中，由于只有表型数据可供利用，只能将控制一个数量性状的所有基因作为一个整体进行研究，而不能区分单个基因在染色体上的位置和遗传效应。
- 20世纪80年代之后出现了大量DNA水平的分子标记。遗传研究中可供利用的不仅有性状的表型数据，还有衡量亲本多态性和表征个体遗传构成的分子标记数据。利用分子标记的基因型数据，可以建立表征染色体的遗传连锁图谱。结合表型数据，就能够对单个数量性状基因在染色体上进行定位，从而把控制数量性状的多个基因分解开来，即QTL定位。

数量性状的遗传模式

- 在数量性状中，既有少数效应比较大的主基因（major gene）控制性状，也有多个效应比较小的微基因（minor gene）或多基因（polygene）控制性状，还有主基因和多基因共同控制的性状。主基因加多基因的混合遗传模型则代表了数量性状遗传的一般模式。
- 这里所说的主效基因和微效基因只是一个相对概念，它们之间并没有一个明确的界限。经典数量遗传研究中，往往把效应大到一定程度、表型上出现一定分组趋势或偏态的基因，看作是主基因。QTL定位中，一般把解释表型变异超过10%或20%的座位称为主效QTL。也有人把特定方法能够检测到的基因都称为主基因，检测不到的基因才称为微效基因或修饰基因（modifier gene）。

§6.2 数量性状的概率论基础

- §6.2.1 概率加法和乘法定理
- §6.2.2 连续型随机变量
- §6.2.3 连续随机变量的数字特征
- §6.2.4 正态分布

概率加法定律

- 若两个互斥（或互不相容）事件A与B在N次试验中各出现了n_A与n_B次，那么和事件（记为A+B）在试验中出现了（n_A+n_B）次，这两个事件之和的概率等于互斥事件A与B的概率之和，这就是概率加法定律，用下面的等式表示。

  $P(\mathrm{A}+\mathrm{B})=P(\mathrm{A})+P(\mathrm{B})$，事件A和B互不相容

- 加法定律还可被推广到n个两两互斥事件中，即n个互斥事件之和的概率等于n个互斥事件概率之和。
- 例如，个体基因型为AA和Aa是互斥事件，从F_2群体中随机抽取一个个体，表现为显性性状的概率为0.25+0.5=0.75。

独立事件和不独立事件

- 对于两个事件A和B，A发生与否不受事件B的影响，或者B发生与否不受事件A的影响，则称事件A和B相互独立；否则称事件A和B不独立。例如，田间有20株表现基本一致的小麦植株，其中18株结红粒，2株结白粒，A和B表示甲乙二人“随机抽取一株恰为白粒”的两个事件。
- 分两种情况讨论2个事件的概率。一种是甲抽取后放回，乙再抽取。显然，这时事件A和事件B发生的概率应相等，即 $P(A)=P(B)= 2/20$。和甲一样，乙仍是从同样的20株小麦中抽取，事件A的发生并不影响事件B的发生，因此A和B是相互独立的。
- 另一种情况是，甲抽到后不放回，乙后抽取。这时，如果事件A发生了，田间剩下的19个植株只有1株是白粒，这时B发生的概率等于1/19。如果事件A没有发生，田间剩下的19个植株有2株是白粒，这时B发生的概率等于1/20。显然，事件A是否发生影响了事件B发生的概率，称这两个事件不独立。

条件概率

- 事件A已经发生的条件下，事件B发生的概率称为条件概率，记为$P(\mathrm{B}|\mathrm{A})$。

$$P(\mathrm{AB}) = P(\mathrm{A}\,|\,\mathrm{B})P(\mathrm{B}) = P(\mathrm{B}\,|\,\mathrm{A})P(\mathrm{A})$$

$$P(\mathrm{B}\,|\,\mathrm{A}) = \frac{P(\mathrm{AB})}{P(\mathrm{A})} \qquad P(\mathrm{A}\,|\,\mathrm{B}) = \frac{P(\mathrm{AB})}{P(\mathrm{B})}$$

概率乘法定律

- 若事件A与B相互独立，我们有$P(B|A)=P(B)$或$P(A|B)=P(A)$，则这二事件同时发生的概率$P(AB)$等于事件A的概率$P(A)$与事件B的概率$P(B)$之积，称为概率的乘法定律。

$$P(AB)=P(A)P(B)$$

连续型随机变量

- 对于随机变量X，若存在非负可积函数$p(x)$ $(-\infty<x<+\infty)$，对于任意的a和b $(a<b)$ 都有下面的等式，则称X为连续型随机变量， $p(x)$称为随机变量X的分布密度，又称概率密度（probability density）。

$$P(a<x<b)=\int_a^b p(x)\mathrm{d}x$$

离散随机变量概率密度的性质

- 分布密度满足下面的三条基本性质。性质1说明了连续随机变量概率密度函数的非负性，以及必然事件（遍历所有可能的取值）的概率为1。性质2和3说明连续随机变量的概率是针对一个取值区间而言的，谈论单个取值点的概率是没有意义的或者说其概率为0，开区间、半开区间和闭区间上的概率是相等的。

$$p(x) \geq 0 \quad \int_{-\infty}^{+\infty} p(x)\mathrm{d}x = 1$$

$$P(x = a) = \int_{a}^{a} P(x)\mathrm{d}x = 0$$，a为任意随机变量的取值

$$P(a < x < b) = P(a \leq x < b) = P(a < x \leq b) = P(a \leq x \leq b)$$

，a和b为满足条件$a<b$的随机变量取值

累积概率分布函数

- 随机变量X在区间 上的取值概率$P(X<x)$，称为随机变量X的分布函数或概率分布，有时也称累积分布（cumulative distribution），用下面的等式表示。

$$F(x)=P(X\le x)=\int_{-\infty}^{x}p(t)\mathrm{d}t$$

- 任意连续随机变量，都可以用它的密度函数或者分布函数来进行定义。从分布函数的定义可以看出，概率密度函数等于分布函数的一阶导数或微分，即：

$$p(x)=F'(x)=\frac{\mathrm{d}F(x)}{\mathrm{d}x}$$

连续随机变量的数字特征

- 不论是离散还是连续随机变量，概率分布函数完整地描述了随机变量的取值规律。有时，概率分布只依赖于少数几个参数，这些参数或它们的函数称为随机变量的数字特征，确定概率分布有时就转变为确定数字特征的问题。
- 有些情况下，确定概率分布并不是一件容易的事。有时可能并不需要知道概率分布，而只需要知道随机变量的某些数字特征就可以了。数字特征不一定能完整地描述一个随机变量，但在理论和实际应用中都有重要意义。
- 期望（也称为均值）（expectation or mean）和方差（variance）是其中最重要的两个数字特征。对于多个随机变量来说，协方差（covariance）和相关系数（correlation coefficient）是多维随机变量最重要的数字特征。

连续随机变量的期望（或均值）

- 期望的定义 $E_X = E(X) = \int_{-\infty}^{\infty} xp(x)\mathrm{d}x$

- 期望的性质

$E(c) = c$，c是任意常数

$E(cX) = cE(X)$，c是任意常数

$E(X \pm Y) = E(X) \pm E(Y)$，$X$和$Y$是任意两个随机变量

$E(XY) = E(X)E(Y)$，X和Y是任意两个独立随机变量

连续随机变量的方差

- 方差的定义

$$V_X = V(X) = E[X - E(X)]^2 = \int_{-\infty}^{\infty} [x - E(X)]^2 \, p(x)dx$$

- 方差的性质

$$V_X = E(X^2) - E^2(X)$$

$V(c) = 0$ ， c是任意常数

$V(cX) = c^2 V(X)$ ， c是任意常数

$V(X \pm Y) = V(X) + V(Y)$， X和Y是任意两个独立随机变量

标准差和标准化变换

- 方差的平方根称为标准差（standard deviation），是实际应用中经常用到的另外一个数字特征。标准差是由方差计算而来，谈不上是一个新的数字特征。但是，它与随机变量的取值和期望有相同的量纲，在参数的区间估计和统计假设检验中经常用到。此外，在期望和方差的基础上，可以对一个随机变量作标准化变换，即：

$$X^* = \frac{X - E(X)}{\sqrt{V(X)}}$$

- 从期望的一些性质，容易证明标准化随机变量X^*的均值为0；从方差的一些性质，容易证明标准化随机变量X^*的方差为1。标准化变换在相关分析和多元统计中具有重要作用，它消除了不同随机变量在量纲上的差异，使得随机变量之间更具有可比性。

两个随机变量之间的协方差

- 协方差的定义

$$Cov_{XY} = Cov(X,Y) = E\{[(X - E(X)][Y - E(Y)]\}$$

- 协方差的性质

$$Cov_{XY} = E(XY) - E(X)E(Y)$$

$$V(X \pm Y) = V(X) + V(Y) \pm 2Cov(X,Y)$$

$Cov(aX, bY) = abCov(X,Y)$，a和b是任意常数

$$Cov(X + Y, Z) = Cov(X,Z) + Cov(Y,Z)$$

两个随机变量之间的相关系数

- 相关系数的定义 $r_{XY} = Corr(X,Y) = \dfrac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$
- 相关系数的性质
  - 与协方差一样，相关系数也没有方向性。相关系数在-1和1之间取值，描述变量X与Y之间线性关系的强弱。
  - 当r=0时，称X与Y之间无相关。这里的无相关代表的是没有线性关系，它们之间还可以具有其它函数关系。
  - 当r=1时，称X与Y完全正相关。当r=-1时，称X与Y完全负相关。当-1<r<1时，称X与Y部分相关，或有一定程度的线性关系。
  - 相关系数还可以看作是X与Y经标准化变换后，标准化变量X^*与Y^*的协方差。由于消除了方差的影响，变换后的协方差，即相关系数，比变换前的协方差能够更准确地反映随机变量X与Y之间的相关关系。

正态分布的普遍性

- 德国数学和物理学家高斯（Gauss，1777~1855）在研究误差理论时，首先用到了正态分布来描述误差效应的分布规律。因此，正态分布（normal distribution）又称高斯分布（Gaussian distribution）。
- 概率统计的中心极限定理（central limit theorem）表明：（1）一个随机变量如果是由大量微小、独立随机因素叠加的，这个变量一般都服从正态分布；（2）当n比较大时，二项分布$B(n, p)$近似服从正态分布，正态分布的均值和方差就等于二项分布的均值np和方差npq（其中$q=1-p$）。

正态分布的概率密度函数

- 正态分布$N(\mu, \sigma^2)$的密度函数

$$p(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}},(-\infty<x<+\infty)$$

- 正态分布$N(\mu, \sigma^2)$的分布函数

$$F(x)=P(X\le x)=\int_{-\infty}^{x}p(t)dt$$

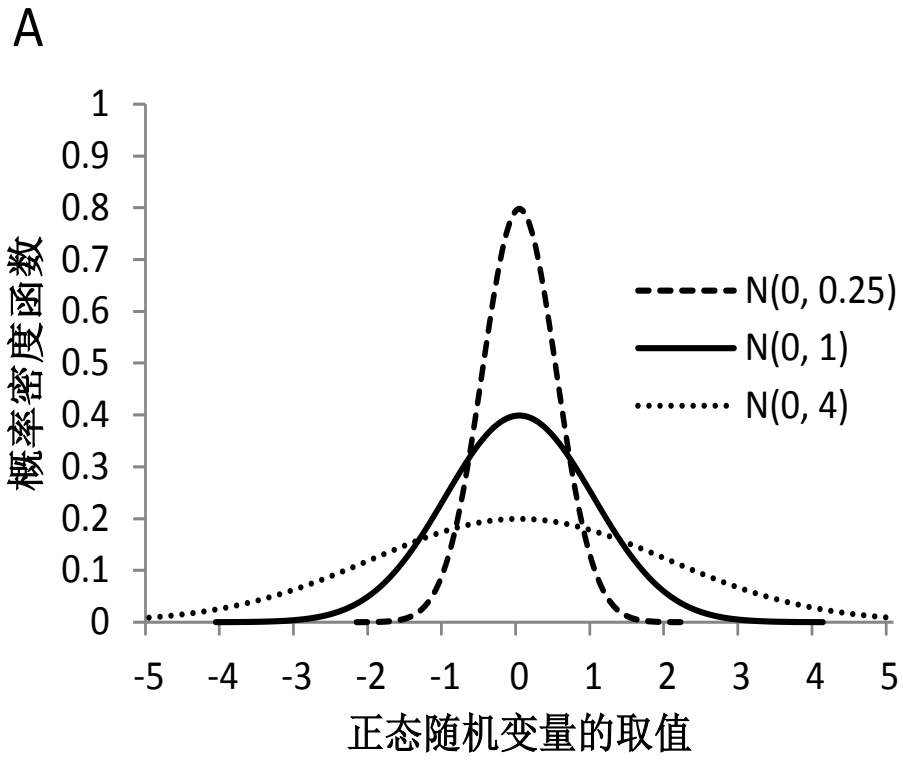
- 标准正态分布$N(0, 1)$的密度函数

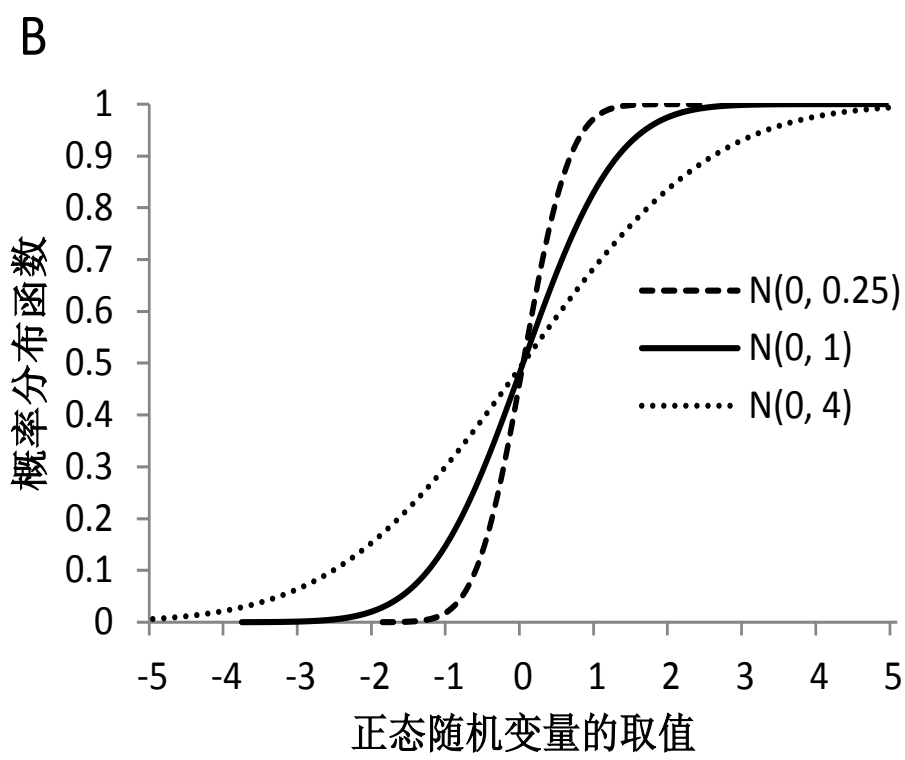
$$p(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}},(-\infty<x<+\infty)$$

正态分布的数字特征

- 从正态分布的密度函数来看，参数μ和σ^2是正态分布中仅有的两个数字特征。只要知道了这两个参数，正态随机变量的概率密度和分布函数就完全确定了。利用微积分的知识可以证明，参数μ正好等于正态随机变量X的期望，参数σ^2正好等于X的方差。因此，也将μ称为正态分布的均值，σ^2称为正态分布的方差。
- 服从正态分布的随机变量非常多，如测量误差、植物的高度、动物的体重、人的身高、健康人的红血球数目、年降水量、月平均温度、海洋的波浪高度等。在概率论和数理统计的理论研究和实际应用中，正态随机变量都起着特别重要的作用。

正态分布的概率密度（A）和概率分布（B）函数曲线





正态分布的特征

- 密度函数曲线呈钟形，以X轴为渐近线；
- 密度函数曲线关于直线$x=\mu$对称；
- $x=\mu$时曲线达到最高点，$x=\mu\pm\sigma$处有拐点；
- 正态分布的密度曲线与X轴之间的总面积等于 1，而且曲线下方介于$x=x_1$到 $x=x_2$（$x_1<x_2$）之间的面积等于随机变量落入区间（x_1，x_2）的概率；
- 任意带有参数μ和σ^2的正态随机变量X，都可以通过标准化转换，变为标准正态分布的随机变量进行研究。

正态随机变量在以均值为中心的k个标准差范围内的取值概率

$$P(|\frac{X-\mu}{\sigma}|<k) \quad 或 \quad P(\mu-k\sigma<X<\mu+k\sigma)$$

标准差的倍数k	1	2	3	1.6449	2.3263	3.0902
取值概率P	0.6827	0.9545	0.9973	0.95	0.99	0.999

- 正态随机变量在三个标准差之间的取值概率超过99.7%。在一些质量控制问题中，将这一概率称为“3σ原则”，在3σ区间之外的样品均可以被看作是次品或非正常品；3σ区间之外的数据也可被看作是奇异值。

§6.3 数量性状的数理统计基础

- §6.3.1 样本统计量
- §6.3.2 抽样分布
- §6.3.3 总体参数的估计
- §6.3.4 一元回归与相关分析
- §6.3.5 多元回归及其假设检验

总体

- 利用数理统计的方法解决实际问题时，往往把研究对象的全体称为总体（population），构成总体的每个成员称为个体。
- 在§6.1.1的妇女身高问题中，调查开始前，所有的英国成年妇女就构成了一个总体，每个妇女是这个总体中的一个个体；在花冠长度问题中，需要研究一个数量性状在两个纯合基因型亲本及它们杂交后代中的分布，双亲、F_1和F_2群体是4个不同的总体，每个群体中的单株是总体中的一个个体。
- 统计学中往往把总体看作一个特定随机变量X服从的分布，与随机变量X等同，也称为总体X或分布X。

总体的样本

- 构成总体的个体一般都有很多，甚至是无限的。在实际问题中，往往不可能对总体的所有个体进行研究。能够研究的只是总体中的一小部分个体，然后希望从这一小部分个体来推断总体的分布规律。被研究的这部分个体称为总体的一组样本（sample）或一个样本群体，样本群体中个体的数量称为样本量（sample size）。
- 为了能够利用样本对总体做出比较可靠的推断，当然就要求样本能够很好地代表总体；失去了代表性，样本中观察到的结果也就难以反映出总体的分布特征。

对总体样本的基本要求

- 通常情况下，随机性和独立性是对样本的两个最基本要求。随机性要求总体中的每个个体具有同等机会进入样本，独立性则要求每个样本个体的取值不受其他样本个体的影响。
- 随机性保证了每个样本个体X_i与总体X有相同的分布；而独立性则保证了样本个体之间是相互独立的，样本的联合概率密度就等于各个样本个体概率密度的乘积。满足随机性和独立性的样本称为简单随机样本，一般情况下均简称为样本。

样本统计量

- 样本来自总体，每个样本个体自然都包含着总体分布的信息。为了把分散在样本中的信息集中起来以反映总体的分布特征，就需要对样本进行各种加工。例如，对原始数据进行分组，统计落入每个组内的个体数或频率，然后从次数分布初步了解总体的分布特征。
- 统计上，把只包含样本的函数称为样本统计量（sample statistic），一般用于总体分布参数的点估计。样本均值（sample mean）、样本方差（sample variance）、样本协方差和样本相关系数是统计学上最重要，同时也是实际应用中最常见的统计量。
- 除此之外，还有样本矩、样本峰度、样本偏度、顺序统计量和分位数等等。为避免混淆，有时也把上一节定义的均值和方差称为总体均值和总体方差。

样本均值和样本方差

- 设$X_1, X_2, \ldots, X_n$是总体X的一组简单随机样本，样本均值和样本方差的计算方法如下：

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

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

样本方差的自由度 $S_X^2 = \frac{1}{n-1}(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2)$

- 样本方差定义为离差平方和除以样本量减1，样本量减1称为样本方差的自由度（degree of freedom）。因此，样本方差就等于离差平方和除以自由度。没有样本均值，也就无法计算样本方差。因此，自由度之所以是*n*-1，可以看作在利用*n*个独立样本个体估计样本均值时损失掉了一个自由度。
- 另外，公式中的*n*个离差项是不完全独立的，它们之间满足和为0这个约束条件。因此，自由度*n*-1中的“1”也可以看作是计算方差的所有离差项满足约束条件的个数。
- 自由度的计算在假设检验和方差分析中是必不可少的，掌握以上两点后，基本上就能计算大部分场合下样本方差的自由度了。

分组数据的样本均值和样本方差

- 用$X_1, X_2, \ldots, X_k$表示k个分组的组中值，样本落在k个组内的频率分别为$f_1, f_2, \ldots, f_k$，频率之和等于1。这时，样本均值和样本方差分别为：

$$\overline{X} = \sum_{j=1}^{k} f_j X_j$$

$$S_X^2 = \sum_{j=1}^{k} f_j (X_j - \overline{X})^2 = \sum_{j=1}^{k} f_j X_j^2 - \overline{X}^2$$

- 样本量n较小时，样本方差公式右端还应该乘以$n/(n\text{-}1)$以进行无偏性矫正；对于较大的样本量，是否矫正对样本方差的影响不大。

样本协方差和样本相关系数

- 设$X_1, X_2, \ldots, X_n$是总体X的一组简单随机样本，$Y_1, Y_2, \ldots, Y_n$是总体Y的一组简单随机样本，

$$Cov_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \frac{1}{n-1}(\sum_{i=1}^{n} X_i Y_i - n\overline{XY})$$

- 如是分组数据，样本协方差为

$$Cov_{XY} = \sum_{j=1}^{k} f_j (X_j - \overline{X})(Y_j - \overline{Y}) = \sum_{j=1}^{k} f_j X_j X_j - \overline{X}\overline{Y}$$

- 样本相关系数：

$$r_{XY} = \frac{Cov(X,Y)}{\sqrt{S_X^2 S_Y^2}}$$

样本均值的期望和方差

- 对简单随机样本X_1、X_2、…、X_n来说，它们相互独立并与总体X有相同的分布，称为独立同分布（independent and identical distribution, iid）。因此，每个样本的期望和方差等于总体的期望和方差。

  $E(X_i)=E_X$ $V(X_i)=V_X$，对任意样本个体i=1、2、…、n

- 根据期望性质公式6.17和公式6.18，容易证明样本均值的期望等于总体的期望。

$$E(\overline{X})=\frac{1}{n}\sum_{i=1}^{n}E(X_i)=E_X$$

- 根据方差性质公式6.23和公式6.24，容易证明样本方差的期望等于总体方差除以样本量n。

$$V(\overline{X})=\frac{1}{n^2}\sum_{i=1}^{n}V(X_i)=\frac{1}{n}V_X$$

正态总体样本均值的分布

- 前面的公式对总体的分布类型没有任何限制。
- 对于正态总体$N(\mu, \sigma^2)$来说，样本均值的期望当然就等于μ，样本均值的方差当然就等于σ^2/n。因此，样本均值服从均值为μ，方差为σ^2/n的正态分布，即：

$$\overline{X} \sim N(\mu, \frac{1}{n}\sigma^2)$$

非正态总体样本均值的分布

- 对于其他均值为E_X、方差为V_X的非正态总体来说，概率论的中心极限定理表明，随着样本量的增加，样本均值渐近服从均值为E_X、方差V_X/n的正态分布，

$$\overline{X} \xrightarrow{n\to\infty} N(E_X, \frac{1}{n}V_X)$$

- 大量模拟结果表明，对于大多数非正态总体来说，n=30的样本均值就十分接近正态分布，也因此把超过30的样本称为大样本（large sample）。

样本方差的均值

- 根据方差计算公式6.21可以证明样本方差的期望等于总体的方差。这一结果与样本均值是一致的。推导过程中，除要求总体均值和方差存在外，没有其它限制条件。因此，不论总体分布是什么，样本方差的期望都等于总体的方差。

$$E(X_i^2) = V(X_i) + E^2(X_i) = V_X + E_X^2$$

$$E(\overline{X}^2) = V(\overline{X}) + E^2(\overline{X}) = \frac{1}{n}V(X) + E_X^2$$

$$E(S_X^2) = \frac{1}{n-1}[\sum_{i=1}^{n} E(X_i^2) - E(\overline{X}^2)]$$

$$= \frac{1}{n-1}[n(V_X + E_X^2) - (V_X + E_X^2)] = V_X$$

样本方差的分布与三大抽样分布

- 样本方差的分布要比均值复杂得多，一般只能研究正态总体的样本（简称正态样本）中方差的分布。
- 在给出正态样本方差的分布之前，首先介绍一下从标准正态分布出发定义的χ^2分布、t分布和F分布，这里不给出它们的概率密度函数。
- 这三个分布都与样本均值和样本方差的分布有密切关系，在有关总体均值和总体方差的统计假设检验中，发挥了必不可少的作用，有时也把它们统称为三大抽样分布。

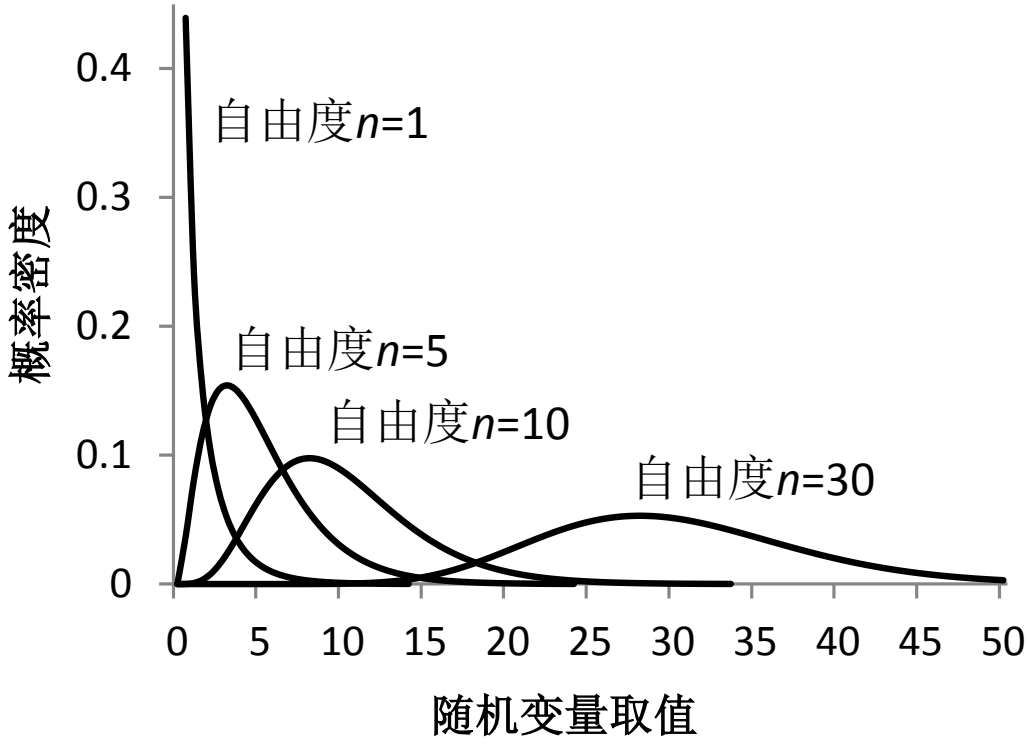
χ^2分布

- χ^2分布的定义，

  $\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2 \sim \chi^2(n)$ ，其中$X_i \sim N(0, 1)$且iid

- 从定义可以看出，χ^2分布代表的随机变量为n个独立、标准正态随机变量的平方和，n又称为χ^2分布的自由度，所以常用$\chi^2(n)$表示这个分布。
- 标准正态分布中不含任何参数，因此χ^2分布只有标准正态分布数目（或自由度）这一个参数。

χ^2分布的特点



- 从概率密度函数曲线图可以看出，χ^2分布是一个单峰、偏态分布，有一个较长的右尾巴；随着自由度的增加，分布均值和分布方差也随之增大。
- 根据分布的定义公式6.50还可以证明，χ^2分布的均值等于n，方差等于$2n$。

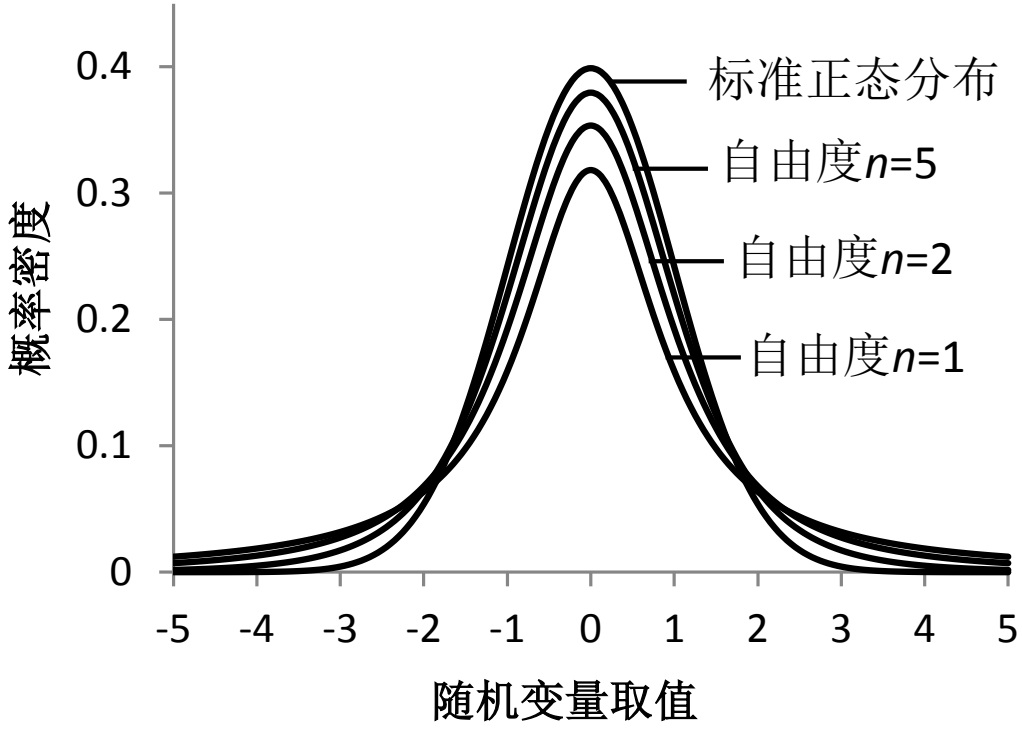
*t*分布

- *t*分布的定义，

$$t = \frac{Y}{\sqrt{\frac{1}{n}X}} \sim t(n)$$ ，其中$Y \sim N(0, 1)$，$X \sim \chi^2(n)$且相互独立

- 从定义可以看出， *t*分布是两个分布的商，分子为一个标准正态分布，分母为一个χ^2分布除以其自由度的平方根，同时还要求分子中的标准正态分布与分母中的χ^2分布相互独立。
- *t*分布中也只包一个参数，即分母中χ^2分布的自由度*n*，所以常用*t*(*n*)表示这个分布。

*t*分布的特点



- 从概率密度函数曲线图可以看出，$t(n)$是单峰态分布，关于$x=0$对称；自由度越小，方差就越大；随着自由度的增加，逐渐趋近于标准正态分布。
- 根据概率密度函数可以证明，当自由度$n=1$时，分布的均值和方差都不存在；当$n=2$时，均值为0，方差不存在；当$n>2$时，均值等于0，方差等于$n/(n-2)$。因此，随着样本量的增大，方差趋近于标准正态分布的方差1。

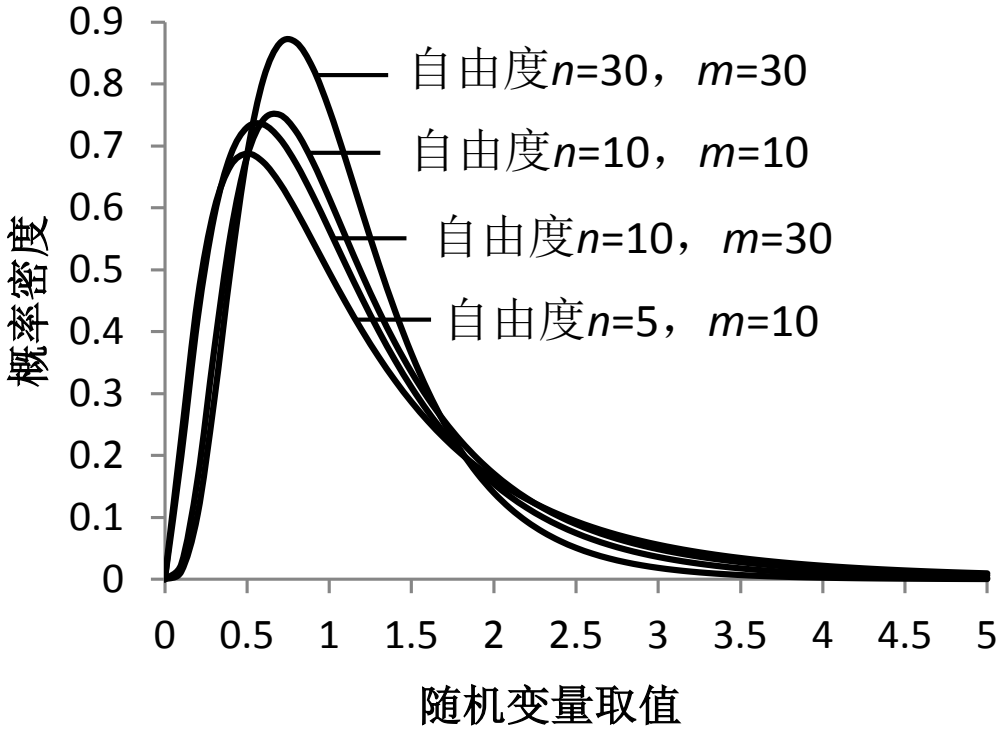
*F*分布

- *F*分布的定义，

$$F = \frac{\frac{1}{n}X}{\frac{1}{m}Y} \sim F(n,m)$$，其中$X \sim \chi^2(n)$，$Y \sim \chi^2(m)$且相互独立

- 从定义可以看出*F*分布也是两个分布的商，分子为一个χ^2分布除以其自由度，分母为另一个独立的χ^2分布除以其自由度。显然，*F*分布中包含了两个参数，一是分子χ^2分布的自由度*n*，另一个是分母χ^2分布的自由度*m*，所以常用$F(n, m)$表示这个分布。分子χ^2分布的自由度也称为第一自由度，分母χ^2分布的自由度也称为第二自由度。

*F*分布的特点



- 从概率密度函数曲线图可以看出，$F(n, m)$是一个单峰左偏态分布，有一个较长的右尾巴。
- 根据概率密度函数可以证明，$F(n, m)$分布的均值等于$m/(m\text{-}2)$，只依赖于第二自由度。$F(n, m)$分布的方差同时依赖两个自由度，表达式比较复杂。

正态总体的抽样分布

- 正态总体$N(\mu, \sigma^2)$中，样本均值和样本方差的分布如下，且相互独立

$$\overline{X} \sim N(\mu, \frac{1}{n}\sigma^2) \qquad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

- 因此，

$$\frac{\overline{X}-\mu}{\sqrt{\frac{1}{n}\sigma^2}} \sim N(0,1) \qquad \frac{\overline{X}-\mu}{\sqrt{\frac{1}{n}S^2}} \sim t(n-1)$$

抽样分布的作用

- 有了上面的分布，就可以在总体方差未知的情况下，利用样本均值和样本方差对均值这一总体参数进行统计假设检验。
- 实际应用中，除了对总体均值进行检验外，还可以对总体方差进行检验，这时就要用到χ^2分布。如要检验两个总体的方差是否相等，或是多个总体是否具有相同的均值，这时就要用到F分布。没有三大抽样分布，也就没有统计假设检验。
- 从定义来看，三大抽样分布都是在正态分布基础上衍生出来的分布，仅适合于从正态分布总体中抽取的简单随机样本。由此也可看出正态分布在数理统计中所占的地位。

总体参数的估计

- 常用的参数估计方法包括最小二乘估计、极大似然估计、矩估计等，这些估计都是样本的函数，即样本统计量。
- 假定θ^是一个样本统计量，作为总体参数θ的一个估计，给定一组样本值，就能计算出一个估计值θ^。为与区间估计（interval estimate）相区分，这样的估计又称点估计（point estimate），在不易混淆的场合下简称估计。

参数估计的无偏性

- 有时，不同的估计方法可能会给出不同的估计值。无偏性和有效性是评价估计值优劣的常用标准。如果θ^的期望等于待估计的总体参数θ，则称θ^是θ的无偏估计（unbiased estimate）；否则称有偏估计（biased estimate）。样本均值和样本方差分别是总体均值和总体方差的无偏估计。
- 有些估计虽然是有偏的，但随着样本量的增加，估计值的期望会逐渐趋近于待估参数，这样的估计称为渐近无偏估计。下面会看到，正态分布方差的极大似然估计是渐近无偏的。
- 对于无偏估计来说，线性变换不改变其无偏性，即无偏估计的线性函数仍然是参数线性函数的无偏估计。对于其他的非线性函数，这一结论不一定成立。

参数估计的有效性

- 绝大多数场合下，都要求估计值能够具有无偏性或者渐近无偏性。参数的无偏估计有时会有很多。一个直观的想法就是估计值围绕真实值的波动越小越好，也就是希望估计值的方差尽可能小，这就是无偏估计的有效性（effectiveness）。
- 如果一个无偏估计有更小的方差，就称这个无偏估计更有效。例如，对于总体X的2个简单样本X_1和X_2来说，满足$a+b=1$的样本统计量aX_1+bX_2都是总体均值的无偏估计，作为练习请读者证明：a和b相等时，无偏估计aX_1+bX_2的方差最小，因此这时的无偏估计是最有效的。

极大似然估计

- 样本似然函数

$$L(\theta; X_1, X_2, \cdots, X_n) = \prod_{i=1}^{n} p(X_i; \theta)$$

- 极大似然估计（maximum likelihood estimate, MLE）

$$L(\hat{\theta}) = \max_{\theta \in \Theta} \{L(\theta; X_1, X_2, \cdots, X_n)\}$$

正态总体参数的极大似然估计

- 样本似然函数

$$L(\theta)=\prod_{i=1}^{n} f(X_i;\theta)=\prod_{i=1}^{n}[\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}]$$

$$=\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2\}$$

- 对数似然函数

$$\ln L(\theta)=-\frac{n}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$$

似然方程及其求解

- 对数似然方程

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu) = 0$$

$$\frac{\partial \ln L}{\partial \sigma^2} = \frac{n}{2\sigma^2} - \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i - \mu)^2 = 0$$

- 均值和方差的最大似然估计

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$$

极大似然估计的统计性质

- MLE具备很多其他估计都没有的优良性质，这也正是MLE之所以得以广泛应用的原因。之前说过，无偏估计仅在线性变换下具有不变性。如果θ^是θ的MLE，那么任一函数g(θ)的MLE都是g(θ^)。这一性质称为MLE的变换不变性，不论是线性变换还是非线性变换都是成立的。有了这一性质，有时就能比较简单地计算一些复杂总体参数的MLE。
- 通常情况下，MLE都渐近服从正态分布，渐近正态分布的均值等于θ，方差等于$[nI(\theta)]^{-1}$，即

$\hat{\theta} \xrightarrow{n\to\infty} N(\theta, \frac{1}{nI(\theta)})$，其中的$I(\theta)$称为Fisher信息量

最小二乘估计

- 最小二乘估计（least square estimation, LSE）一般用于线性模型中未知参数的估计。这里以包含一个未知参数的线性模型为例，说明这一方法。对样本X_1、X_2、…、X_n来说，定义下面的线性模型，其中，μ为待估总体均值，ε_i称为每个样本与总体均值之间的离差。

  $X_i = \mu + \varepsilon_i$，对任意样本$i$=1, 2, …, n

离差平方和

- 对模型中的离差求平方和，得到离差平方和$Q(\mu)$。

$$Q(\mu)=\sum_{i=1}^{n}(X_i-\mu)^2$$

- 在给定样本值的条件下， $Q(\mu)$是未知参数μ的函数，它达到最小时 的取值，称为 的最小二乘估计。因此，求解LSE实质上也是一个极值问题。

最小二乘估计的计算

- 对函数$Q(\mu)$求导、并令导数等于0，就得到μ的LSE

$$\frac{d}{d\mu}Q(\mu)=\sum_{i=1}^{n}-2(X_i-\mu)=-2(\sum_{i=1}^{n}X_i-n\mu)=0$$

$$\hat{\mu}=\frac{1}{n}(\sum_{i=1}^{n}X_i)=\overline{X}$$

- 总体均值μ的LSE等于样本均值，也等于正态总体均值的MLE。对于任何总体的样本，都可以如上计算总体均值的LSE。
- 在正态总体的假定下，LSE有更多优良性状（即Gauss-Markov定理），也只有这时才能对参数进行区间估计和假设检验。

回归与相关

- 回归与相关是Galton等人研究人类身高这一数量性状在亲子之间遗传规律时提出的分析方法，之后很快成为数理统计中分析两个或多个变量关系的主要方法。
- 利用1078对父、子身高数据，Galton发现父亲身高（用X表示）的均值是67英寸，儿子身高（用Y表示）的均值是68英寸，Y与X的回归关系是$Y=33.73+0.516X$。
- 也就是说，高个子父亲有生高个子儿子的趋势，矮个子父亲有生矮个子儿子的趋势；父亲身高每增加1英寸，儿子身高平均增加0.516英寸，而不是1英寸，身高在父子之间是部分遗传的。

回归与相关

- Galton随后考察了高个子父亲所生儿子的平均身高和矮个子父亲所生儿子的平均身高。结果发现，平均身高为80英寸的高个子父亲所生儿子的平均身高为75.01英寸，低于父亲的身高；平均身高为60英寸的矮个子父亲所生儿子的平均身高为64.69英寸，高于父亲的身高。Galton把这种现象看作子代身高有向中心（或群体均值）的回归趋势。正是这种趋势，使得一段时间内的人类身高相对稳定，不会出现高的越高、矮的越矮的两极分化。
- Galton随后利用家庭身高数据，得到子女平均身高Y与父母亲平均身高X的回归系数是0.65。在第8章中可以看到，上面的两个回归系数就是身高性状的狭义遗传力。

回归分析与相关分析

- 如果两个变量之间有明确的因果关系，一般采用回归分析；否则采用相关分析。
- 协方差和相关系数衡量的都是变量之间的线性关系。因此，变量之间的线性回归分析与相关分析是密切关联的两种分析方法。

一元回归的线性模型

- 假定变量X与Y之间有线性关系，b称为回归系数，a称为截距。对于一组X和Y的样本来说，用下面的公式表示他们之间的线性关系，公式中ε_i称为回归离差或残差。

  $Y_i = a + bX_i + \varepsilon_i$，$i$=1、2、…、$n$表示样本

- 仅关心参数估计时，残差ε_i只需满足均值为0的iid即可，对具体的分布类型没有要求；
- 如同时关心模型和回归参数的检验，则还要求残差服从正态分布。

一元回归模型的参数估计

- LSE是残差平方和达到最小时参数a和b的取值。通过求导，并令导数等于0，可以得到a和b的LSE。

$$Q(a,b)=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}(Y_i-a-bX_i)^2$$

$$\hat{b}=\frac{\sum_{i=1}^{n}X_iY_i-\frac{1}{n}(\sum_{i=1}^{n}X_i)(\sum_{i=1}^{n}Y_i)}{\sum_{i=1}^{n}X_i^2-\frac{1}{n}(\sum_{i=1}^{n}X_i)^2}\qquad \hat{a}=\overline{Y}-\hat{b}\overline{X}$$

利用相关和方差计算回归系数

- 回归系数可以看作是变量X与Y之间协方差与自变量X方差的商。在第8章研究数量性状的亲子关系时，要经常用到这一结论。

$$\hat{b} = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{1}{n}(\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{\sum_{i=1}^{n} X_i^2 - \frac{1}{n}(\sum_{i=1}^{n} X_i)^2} = \frac{Cov(X,Y)}{V(X)}$$

回归系数与相关系数的关系

- 回归系数带有量纲，如变量Y的量纲为kg、X的量纲为cm，则回归系数的量纲是kg/cm。而相关系数是无量纲的。
- 从定义不难看出，它们之间满足下面的关系式

$$r = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}} = b\sqrt{\frac{V(X)}{V(Y)}}$$

回归系数与相关系数的检验

- 在正态总体的条件下，相关系数可用下面的t分布进行显著性检验。

$$t=\frac{r}{\sqrt{\frac{1-r^2}{n-2}}}\sim t(n-2)$$，其中n为样本量

- 同时还可以证明，一元回归方程的显著性检验、回归系数的显著性检验及相关系数的显著性检验是等价的。
- 在多元回归方程中，这一结果只对正交设计矩阵才是成立的。

单个纯系群体内的亲子相关

- 一个纯系群体的均值用G表示，个体i的表型用线性模型$X_i=G+\varepsilon_i$表示，其中ε_i为随机误差。根据纯系理论，个体i大量子代的平均表型等于均值G，根据协方差的性质，

$$Cov(X,\bar{Y})=Cov(G,G)+Cov(G,\varepsilon_i)=0 \quad V(X)=\sigma_\varepsilon^2$$

- 如果每个个体只有一个后代，个体i的后代表型用线性模型$Y_i=G+\delta_i$表示，假定上下代的随机误差效应有相同的方差，且相互独立，那么

$$Cov(X,Y)=0 \quad V(X)=V(Y)=\sigma_\varepsilon^2$$

- 因此，子代均值对亲代表型X的回归系数为0，它们之间的相关系数也是0。子代与亲代的回归系数为0，相关系数也是0。

多个纯系群体间的亲子间相关

- 在n个不同基因型构成的遗传群体中，假定繁殖后的基因型保持不变，亲代表型和子代表型用线性模型$X_i=G_i+\varepsilon$和$Y_i=G_i+\varepsilon$表示，其中ε为随机误差。
- 根据协方差的性质，

$$Cov(X,\overline{Y}) = Cov(G,G) = \sigma_G^2 \qquad V(X) = \sigma_G^2 + \sigma_\varepsilon^2$$

- 子代均值对亲代表型X的回归系数其实就是第7章要介绍的广义遗传力；它们之间的相关系数就是广义遗传力的平方根。

$$b = \frac{\sigma_G^2}{\sigma_P^2} \qquad r = \frac{\sigma_G^2}{\sqrt{\sigma_P^2\sigma_G^2}} = \sqrt{b}$$

随机交配群体的亲子间相关

- 随机交配群体中，子代的平均表现等于亲代的育种值（加性效应）

$$Cov(X,\bar{Y}) = Cov(G,A) = \sigma_A^2 \qquad V(X) = \sigma_G^2 + \sigma_\varepsilon^2$$

- 子代均值对中亲的回归系数等于狭义遗传力；后代平均表现与中亲的相关系数等于狭义遗传力的平方根。

$$b = \frac{\sigma_A^2}{\sigma_P^2} = h^2 \qquad r = \frac{\sigma_G^2}{\sqrt{\sigma_P^2 \sigma_G^2}} = \sqrt{b} = h$$

多元回归分析

- 对于一个因变量和多个自变量来说，假定m个自变量为$X_1, X_2, \ldots, X_m$，因变量为Y，它们之间的线性关系模型为

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_m X_m + \varepsilon$$

线性回归的矩阵模型

- 对于大小为n的一组样本来说，将因变量Y的n个观察值用下面的列向量$\mathbf{Y}$表示，常数项的系数1和回归系数对应的因变量观测值用如下的矩阵$\mathbf{X}$表示，所有待估参数用如下的向量$\mathbf{b}$表示，残差用如下的向量$\boldsymbol{\varepsilon}$表示，向量和矩阵后面的下标表示阶数。

$$\mathbf{y}_{n\times 1}=\begin{bmatrix}Y_1\\Y_2\\\vdots\\Y_n\end{bmatrix}\quad \mathbf{X}_{n\times(m+1)}=\begin{bmatrix}1 & X_{11} & \cdots & X_{1m}\\1 & X_{21} & \cdots & X_{2m}\\\vdots & \vdots & \ddots & \vdots\\1 & X_{n1} & \cdots & X_{nm}\end{bmatrix}\quad \mathbf{b}_{(m+1)\times 1}=\begin{bmatrix}b_0\\b_1\\\vdots\\b_m\end{bmatrix}\quad \boldsymbol{\varepsilon}_{n\times 1}=\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_n\end{bmatrix}$$

$$\mathbf{y}=\mathbf{X}\mathbf{b}+\boldsymbol{\varepsilon}$$

回归系数的最小二乘估计

- 矩阵模型中的**X**称为设计矩阵或发生矩阵（design matrix or incidence matrix），参数向量**b**的最小二乘估计满足下面的正规方程，其中$\mathbf{X}^{\mathrm{T}}$表示设计矩阵 的转置。

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- 如果$\mathbf{X}^{\mathrm{T}}\,\mathbf{X}$的逆矩阵存在，则**b**的最小二乘估计为

$$\hat{\mathbf{b}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

回归模型的预测值和平方和分解

- 获得回归参数的估计后，在给定一组自变量的取值后，就能够预测变量Y的表现，称其为预测值，

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}}$$

- 总离差平方和的分解

$$SS_T = \sum_i (Y_i - \overline{Y})^2 = \sum_i (Y_i - \hat{Y}_i + \hat{Y}_i - \overline{Y})^2 = \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \overline{Y})^2$$

$$SS_{\text{Reg}} = \sum_i (\hat{Y}_i - \overline{Y})^2 \qquad df_{\text{Reg}} = m \qquad MS_{\text{Reg}} = \frac{MS_{\text{Reg}}}{df_{\text{Reg}}}$$

$$SS_\varepsilon = \sum_i (Y_i - \hat{Y}_i)^2 \qquad df_{\text{Reg}} = n - m - 1 \qquad MS_\varepsilon = \frac{MS_\varepsilon}{df_\varepsilon}$$

回归模型的显著性检验

- 利用前面的两个均方，就可得到回归模型显著性检验的F统计量，回归项和剩余项的自由度对应于F分布的两个自由度。

$$F = \frac{MS_{\text{Reg}}}{MS_{\varepsilon}} \sim F(df_{\text{Reg}}, df_{\varepsilon})$$

- 根据上述等式计算出的F值，就可以对回归模型进行显著性检验。