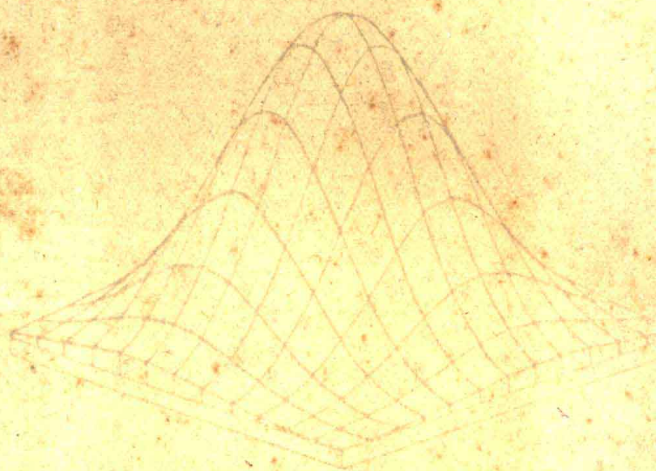
数据处理方法



中国科学技术大学数学系

数 据 处 理 方 法

*

中国科学技术大学数学系編
中国科学技术大学印刷厂印刷

*

1973 年 4 月合肥第二次印刷
印数 0001—2000

定价 0.19 元

毛主席語录

胸中有“数”。这是说，对情况和问题一定要注意到它们的数量方面，要有基本的数量的分析。任何质量都表现为一定的数量，没有数量也就没有质量。我们有许多同志至今不懂得注意事物的数量方面，不懂得注意基本的统计、主要的百分比，不懂得注意决定事物质量的数量界限，一切都是胸中无“数”，结果就不能不犯错误。

《党委会的工作方法》（一九四九年三月十三日），毛泽东选集》第四卷第一四四三頁

內 容 簡 介

这是一本为普及几种最簡易的数理統計方法用的小册子，凡有高小或初中数学程度的工人、农民或科技人員，如果想要学一点数据处理的方法，可以学习一下这本小册子的內容。在进行生产斗爭和科学实驗的工作中，如果出現一些处理数据的問題，在这本小册子里所介紹的几种方法也許可以起些帮助解决的作用。为了普及这方面的知識，所以这里凡是有关数学的略为高深一点的道理都被略去了。

中国科学技术大学数学系

1973年4月

目　　录

§1. 从破除迷信谈到事物的概率

求神問卜，拆字算命，大家都知道是迷信，是沒有根据的。这些事情，現在看来确实很可笑，但是在旧社会也确实有很多人相信。这是什么原因呢？归根到底，是因为旧社会反动統治阶級的欺騙宣传，和人們缺乏科学知識的緣故。掌握了現代科学知識，我們就可指出迷信的謬論所在，彻底揭露剝削阶級的阴謀詭計。現在，我們从数学的一門分支即概率論的角度，来看看一些迷信到底錯在哪里。

在数学中，把在自然界中无論什么事件出現的可能性称为概率，而且設法用数值去度量它。有人可能要問：一头猪有多重，可以称它的斤数；一条鋼軌有多长，可以量它的米数；你現在的体溫，可以测量其溫度度数；但是某事件出現或发生的可能性有多少，却怎样去度量呢？数学研究者們想出了好办法，使这种可能性也能够度量，并且可以进行科学分析和推导，得出正确的結論。

那么，怎样度量事件出現或发生的可能性呢？先看两种极端的情况。一种情况是某事件必然会出現或发生，譬如說“一头猪总有一天要死的”。另一种情况是某事件决不可能发生，譬如說“西天出太阳”。在数学上就把类似于“一头猪有一天会死”这类事件发生的可能性用数字一来度量，說成“一头猪有一天会死”的概率等于一。为了写起来省事，把概率記成符号 P，写成：

$$P(\text{一头猪有一天会死})=1;$$

把类似于“西天出太阳”一类事件发生的可能性用零来度量，写

成：

$$P(\text{西天出太阳})=0.$$

但是在我們实际所遇到的事件中，属于这两类的是比較少的，

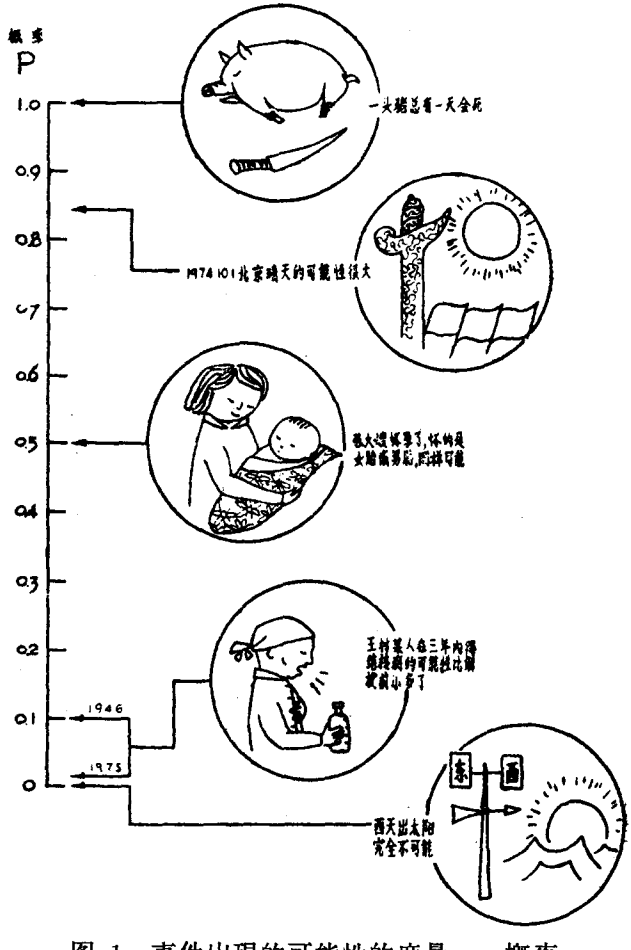


图 1　事件出现的可能性的度量——概率

大多数是在这絕对可能和絕对不可能两类之間，即它們的概率介乎一与零之間。事件E发生的可能性越大的，其概率越靠近一；可能性越小的，越靠近零；但是概率不会超过零和一的区間之外。

$$0 \leqslant P(E) \leqslant 1.$$

怎样才能度量出某一事件发生的概率呢？我們可以用图1中的几个例子来說明。图中以零到一的概率数值为縱座标，在座标对应的位置处。排列着这些事件。先看图上在$P=0.5$处的怀孕事件。怀了个男胎的概率是0.5，0.5这个数字是可以从科学理論上来推算，而且可以近似地用事实来驗証的。我們知道，一对夫妇生男育女，究竟生男还是生女在怀孕之前是不知道的。但是生男或生女的概率，也就是生男或者生女的可能性是可以度量的。讓我們看一看从生物学的实驗所得出的生男育女的規律。一个人的性别取决于她（或他）身上每个体細胞內細胞核中的一对性染色体。如果是女的，她的一对性染色体大小形状相同，我們用符号記为xx；如果是男的，他的一对性染色体是xy，其中只有一个是x，另一个是y，比x小得多。但是在生殖細胞中，染色体的数目要減半。这使女性排出的各个卵子內的性染色体只含一个x；使男性所排出的各个精子內的性染色体也只含一个，或者是x，或者是y．因为在精細胞中x和y的数目是相等的，所以在男性排出的精子中，含x的精子和含y的精子在数目上大約是相等的，应該大約各占一半。假設每个精子都有机会使一个卵子受精，从而产生一个胎儿的話，那么男胎（xy）和女胎（xx）出現的百分率应該大約各有50%。这可以用图解表示（見图2）。

虽然一对夫妇所生育的子女不多，不能驗証上述的理論比例，但是如果把許多对夫妇的子女合起来計算，就可以发現子女的数目相差不多，大約各占50%，跟理論上的說法基本上是吻合的。因此，在人数比較众多的情况下，預期男胎出現的百分率近

乎 50%，女胎的也近乎 50%。現在如果有个张大嫂怀了孕，不知

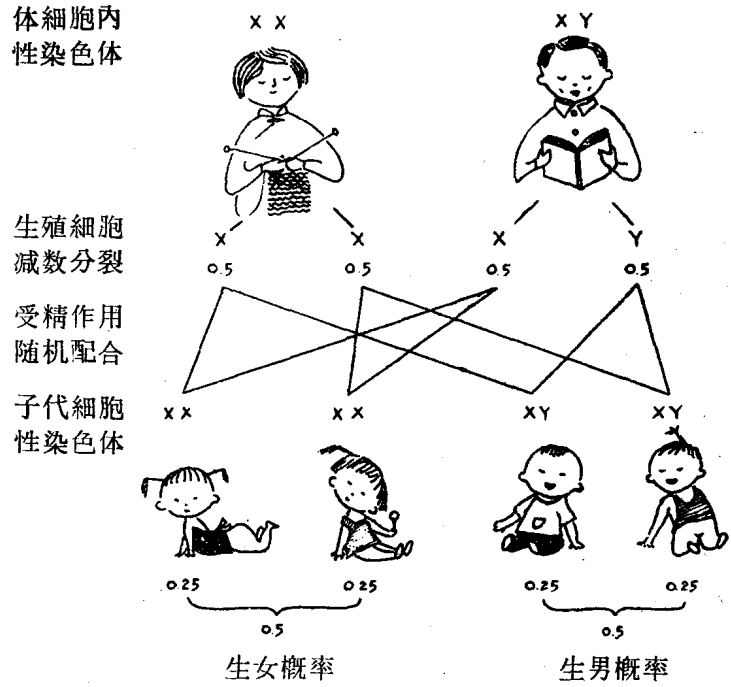


图 2　男女胎出現的概率

道是男胎还是女胎，我們就可以有科学根据地說她怀男胎的可能性有一半，怀女胎的可能性也有一半，写成概率的符号，这就是：

$$P(\text{男胎})=0.50,$$

$$P(\text{女胎})=0.50。$$

这种概率因为是从科学理論推导出来，然后經过事实或实驗証明了的，称为理論概率。

現在讓我們看一下旧社会里的情况。旧社会重男輕女，在农村里，大家尤其希望得男孩。許多人怀了孕，要到送子观音庙去

烧香叩头，祈求她送来个儿子。当然泥塑的菩薩影响不了男女胚胎的概率。只要怀了胎，总有一半的可能怀个男胎。但是在旧社会里，在迷信泥菩薩的心理支配下，如果老张果眞后来得了个男孩，他一定欢天喜地馬上传出去，說这个菩薩有求必应，眞的灵驗！如果生了个女孩，他也不怪菩薩，因为他缺乏科学知識，迷信菩薩总是对的；于是他就另找得不到男孩的原因：是不是因为上庙那天沒有戒葷呢？是不是上代沒有“积德”呢？反正牵强附会，总要找出一条可以替菩薩开脫的原因。当然，这些原因都是不相干的。

其实，决定性别的不是菩薩，而是每个人生殖細胞中的性染色体。人类群体总是按照科学的客观規律，一代代男女大約各半地传下去。

在現再来看一个生病的例子，怎样估計得病的概率。在旧社会里卫生差、医院少、葯費貴，人們得結核病的很多；一个好好的人，在三、五年內传染上这种病也是难說。譬如說1940年有人在王村地区調查了一万个人，到了1943年再去調查，发現原来健康的人在这三年中有10%传染上了結核病。換句話說，調查的結果告訴我們，王村1940－1943年間新得結核病的病例的发生頻率是10%或者0.10。（頻率与概率不同的地方在于：頻率是事件已发生以后所占总数的百分率，概率則是在事件未发生之前它会发生的可能性的度量。它們在数字上尽管可以相同，但在意义上大有区别。）有了这样一个經驗数字，我們就可以推想：如果王村每个人传染上結核病的机会都一样，那么在以后三年內，即1943—1946年的期間內，任一个王村人得結核病的概率或可能性是0.10，可写成为：

$$P(1943—1946\text{年王村某人得結核病})=0.10.$$

这种概率的度量是根据大量調查得来的頻率，也是一个从經驗得

来的值。象这种概率我們称为經驗概率。

現在大家知道結核病是一种桿状細菌侵入体內所引起的，只要注意卫生就不易得这种病。但是那时的一般农民怎能知道呢？他們缺乏关于結核病的知識，往往以为有什么痨病鬼在作祟。生了病，就請巫师来，祈求把鬼捉住；或者对鬼賄賂，請它饒命。巫师鬼混了一陣，乘机哄騙了些錢財，吃得酒醉肉飽而去。結核病呢？照旧以10%上下的概率而发生，决不因此而稍減。

解放后，全国展开了爱国卫生运动，大家講究卫生，医院成倍成倍的增加，葯費也便宜許多，在毛主席的无产阶級革命路綫的正确指引下，赤脚医生在农村更是大受欢迎，結核病的发生頻率就大大降低。譬如說从1969—1972年在王村再作調查，得到的数字說明該区在三年內新发生結核病的頻率已降低到1%或0.01。因此可以同样推測一个健康的王村人在以后的三年內得結核病的概率已下降到0.01以下；可以写成为：

$$P(1972—1975\text{年王村某人得結核病})\leqslant 0.01.$$

一般說来，經驗概率必須积累相当多的数据或資料来决定的。这些数据或資料从哪里来呢？那就是通过实践和調查。譬如說查遍了过去50年来北京地区在十月一日那天的气象記录，計算出晴天占其中的42天；我們就可以估計在未来的某一年十月一日碰上晴天的概率是：

$$P(1974.10.1,\ \text{北京天晴})=\frac{42}{50}=0.84.$$

在旧社会，許多人相信有些事件的发生是在冥冥之中注定了的，說它是“命”、“运气”，而且还認为可以請算命先生或者拆字先生去算出来。其实，算命拆字都是騙人的勾当。譬如說，解放前有些人由于失业，就去看相算命，自認为命相不好，想知道何时好轉。他們并不知道，失业是旧社会剝削制度的产物，也

就是說，在旧社会里失业概率本来是很大的，于是就受騙了。北京有个看相算命多年的牛乐天，解放后自己坦白說："看相，其实是察顏观色，揣摸来看相人的身分职业，談天說地，連詐带套地猜測对方的心思，然后故弄玄虛，說几句不着边际的話，或者給他增点恐惧，或者給他点'希望'，引誘他来占卦，卦金就稳稳收进腰包了"。我們只要能通过实际調查，根据事实估計出各种事件发生的概率，也就可以看出星相占卜的不可信了。

§2. 概率怎样出现在工农业生产中？

在工农业生产中，經常遇到一些和概率这个可能性的度量有关系的事例。下面举几个普通的例子說一說，使大家对于概率有进一步的認識。

例如某灯泡厂生产一种电灯泡。根据生产經驗，知道在所生产每一批同一型号的灯泡中每只灯泡的使用寿命都是参差不齐的。要想制成一批寿命恰恰完全相同的灯泡实际上是不可能做到的。經过一定的抽样检驗質量的方法，厂里的技术人員能对每批灯泡的寿命作出有一定的可信程度的評估。譬如說他对某批灯泡的寿命評估說："这批灯泡，总数五万只，通过抽样检驗，評估它們的使用寿命有95%可以超过1000小时，有10%可以超过5000小时。"这里他对这批灯泡的寿命作出了一个定量的估計，要驗証他这一估計的准确程度，必須等到这批灯泡全部使用时間超过5000小时后才能分晓。因此在此以前，对这批灯泡中任一只尚未使用的灯泡来說，只能根据他的估計說一下它的寿命超过1000小时和5000小时的可能性。除此而外，别无他法。那就是

說这只灯泡寿命超过 1000 小时的概率是 95%，而超过 5000 小时的概率則是 10%。写成式子即：

$$P(t\geqslant 1000\text{ 小时})=0.95,$$

$$P(t\geqslant 5000\text{ 小时})=0.10.$$

这是概率在工业生产中出現的一种方式。

可是对于顧客来說，他感兴趣的只是是他所买回家去的那只灯泡的寿命，对于其他 49999 只灯泡的寿命如何他是不管的。如果对他說："你这只灯泡的寿命可以超过 1000 小时"，幸而他所买的刚好是在 95% 以內的那些寿命超过 1000 小时的灯泡中的一只，那么这句話总算說好了；但不幸而他所买的却是那 5% 以內的那些寿命較短的灯泡中的一只，那么这句話就說錯了。因此上面这句話說对的概率虽达 95%，而說錯的概率也有 5%，这是概率在工业生产中出現的又一种重要方式。这里对某一只灯泡的寿命估計得对与不对的概率虽然在数值上和上述的寿命 t 大于某一时間的概率有对等的关系，但在意义上显然是有区别的。

如果有人要問，这位同志是怎样来估計产品寿命的呢？上述的估計概率的数字又是怎样得来的呢？关于寿命評估方法，将来在另一本小册子里介紹給大家，这里暫时从略。

其实通常所謂"十拿九稳"呀，"有把握"呀，"冒点风险"呀，都包含有概率的含义，不过概率的数学把这些俗語更进一步地定量化和精确化罢了。在工业中，如果某厂对他們所生产的灯泡的寿命有 95% 的把握說它們超去 1000 小时的話，就意味着每只灯泡的寿命将超过 1000 小时的概率为 0.95。如果用戶买了一只这种灯泡，那么这只灯泡的寿命将不超过 1000 小时的概率是 5%，这也就是他必須冒的"风险"。

工厂的产品，不論是成品或是半成品，都必須交給其他工厂或用戶使用。这一交一收，是一道不可缺少的手續。在工厂方面，

一批产品出厂了，要进行检查，以保証質量达到一定的規格，基本上不致把一批次品当作好品送出門，使用戶吃亏。在用戶方面，一批产品进来了，需进行驗收，即确保質量滿足自己的要求，又基本上不致把一批好品当作次品退回去，使工厂遭受損失。不把次品当作好品送出門，是工厂为用戶打算，为用戶保証的。不把好品作为次品退回去，是用戶为工厂打算，为工厂保証的。如果基本上能做到这样，工厂和用戶就双方有益。如果个个工厂家家用戶都这样做，国家的建設，就更加多、快、好、省。对一大批产品，通常采用抽样检驗的方式。但由样品来估計总体，就不可避免地有可能出現誤差，从而使两方各承担一定的风险。厂方所承担的这种风险称为厂方风险，用戶所承担的則是用戶风险。如果一批产品眞正的次品率是 P，P 如低于某个規定的較低的次品率 P_0，（称为接受質量水准），应該是接受它們的，但如因抽样之故，不幸而不接受这批产品，則出現这个事件的概率至多是 α，即愿冒的风险是 α（厂方风险）。如果一批产品眞正的次品率是 P，P 如高于某个規定的較高的次品率 P_1（称为不接受質量水准），应該是不接受它們的，但如因抽样之故，不幸而接受了这批产品，則出現这个事件的概率至多是 β，即愿冒的风险是 β（用戶风险）。

为了使两方面都尽可能地免受損失，就需把这两項风险規定在一个較低的概率上，并且把双方所商定的以上两个水准 P_0 和 P_1 以及两个概率 α 和 β 写在定貨的合同上。这样双方就有了共同一致的抽样检驗的方案，可以省却許多不必要的麻煩了。至于这些水准怎么定？从它們又如何定出抽样方案？都将在有关抽样检驗方法的一本小册子中予以介紹。

工业上的产品还有一个使用时灵不灵的問題。例如一支上天的火箭，由上千个零件装配而成。这些零件在装配前虽然反复試

驗，但試驗的条件毕竟不能和它在火箭上实际使用时的条件完全一致。它在試驗时即使是灵了，却不能保証它在实际使用中百分之百地也灵，但也許可以保証百分之 99.99 或 99.999 地灵。这样火箭上各个零件的可能灵可能不灵的程度或者靠得住靠不住的程度。影响到总装成的火箭整体的灵不灵或靠得住靠不住的問題。这类的問題称为工业产品的可靠性問題，这类問題在航空工业和火箭工业中尤其显得重要。实际上工业产品的可靠性的程度就是以它們在使用时是灵的概率来度量的，研究它們的可靠性，就必須定量地研究这一概率究竟有多少。我們說一支火箭的可靠性是 99.99%，等于說它在放射时是灵的概率为 0.9999；不灵的概率（即不可靠性）因此只有 0.0001，确实很小很小，但总比零大一些。关于可靠性的評估方法这里也暫时从略。

以上在工业中所遇到有关概率的問題在农业中也有之。例如用抽样方法对某区小麦的产量进行了数理統計的估計，估計它們的平均亩产量有 95% 的可能介乎 900 斤到 1100 斤之間，即 1000 斤 ±100 斤，这种估計也不是百分之百地保险的，也有可能估錯，虽然估錯的可能性不大，譬如說只有 5 %。因此这是一种带有估錯的概率有多少的估計，是一种比較老实的估計。話虽如此，即使由于抽样之故，眞的估計錯了的話，也不会估錯太远。如估計得更粗率一点，譬如說把所估計的亩产量区間扩大到 1000 斤 ±130 斤，就可以把估錯的概率降低到 1 %。关于总体平均数的估計方法在这本小册子里以后还会詳予介紹。

在农业中进行田間試驗是一項发展农业生产的必不可少的重要工作。在比較几种不同品种的作物的产量的高低时，在比較几种杀虫剂的杀虫功效的大小时，在比較几种田間管理方法的上下时，在比較几种施肥方案的优劣时，所有对从科学实驗的結果作出的判断和結論，无不隐藏着一个概率的尺度，用来度量这些科

学实驗的成果在推广时的成效的可能性如何。推而广之，在工业的发展道路中所不可缺少的科学实驗也是同样必須从实驗結果来作出一些具有一定的可能性的推断。有趣的事是这种推断的概率往往在实驗未作之前就已經事前决定好了。如不事前决定好这种概率，科学实驗工作者就往往只能无根据地来决定他們所要做的实驗的条件因素、实驗水准和重复实驗的数目。这样做往往造成人力物力和时間的浪費，因为实驗数目如果定得太少了，从得出的数据不足以作出有价值的估計和推断，从而使整个实驗成为徒劳而无功，白費一場苦辛，但如实驗数目定得太多，那么最后发現本来可以不必做这么多的实驗也可得出同样的估計推断，这样就做了許多虛功，枉費了許多人力物力和时間。那么如何多快好省地来做科学实驗呢？就是怎样想法設計一項科学实驗，使它的結果所得的計数或度量的数据，在进行了一定的数学分析之后，可以以預期的置信概率来作出科学的估計和推断。这个概率既不能定得太高，太高使人对实驗的結果无法分辨之，不能作出定量的估計和推断；也不能定得太低，太低使人事后会感到有許多数据其实不起什么作用，倒令人化費了不少牛劲去求它們。因此对科学实驗結果的估計和推断的置信概率的預先規定在科学实驗設計中是一樁具有头等重要的事情。根据經驗和判断，規定了一个合宜的置信概率，再从而設計实驗的規模和安排，往往可以收到事半而功倍的效果。沒有这种設計，往往会造成不必要的浪費。关于科学实驗設計的一般知識，請参考有关的一类小册子，例如科普出版社出版的《科学实驗設計一百例評注》。在这本小册子中也有联系到实驗設計的地方，但只是附带的。

概率这个重要的对可能性的数学度量在工农业的发展中几乎到处有它的踪迹，从以上所举的一些事例中我們可以見其一斑，大家对概率是怎么回事，它在哪些場合下出現，可能已有了一个

初步的認識。在下一节中我們想談一談有关概率的两个最基本的数学法則，即概率的相加和相乘的情况。

§ 3. 概率的相加与相乘

曾听到过一段有趣的的对話：

张：你老兄快做爸爸啦，恭喜恭喜。你希望得个女孩，还是得个男孩？

李：全一样。

张：我猜呀，不是得女，就是得男。

这虽是一个笑話，但是其中确实包含有数学的道理。老李的妻子生孩子是一樁将要发生的未知事件。如果只記单胎，不計双胞胎、怪胎及其他不幸事件，那么不是得女就是得男。換句話說，全部可能发生的事件就只有这两种，而且这两樁事件的发生是互不相容的，所謂“互不相容”就是不能兼而有之的意思。我們已經知道，在未分娩之前，生女生男的可能性各有一半，其概率各为0.5，用符号来表示，就可以写成：

$$P(\text{女})=P(\text{男})=0.5.$$

老李的妻子生孩子（女或男）这一事件是必然要发生的，发生的可能性是百分之百，也就是发生这样事件的概率等于1。写成公式是：

$$P(\text{女或男})=1.$$

我們可以看到，生孩子这件事与生男育女这两件事发生的概率之間存在着相加的关系，就是：

$$P(\text{女或男})=P(\text{女})+P(\text{男})=0.5+0.5=1.$$