中山大學
SUN YAT-SEN UNIVERSITY

HYPERCOMP
Hyperspectral Computing Laboratory

Multivariate Statistics Analysis:

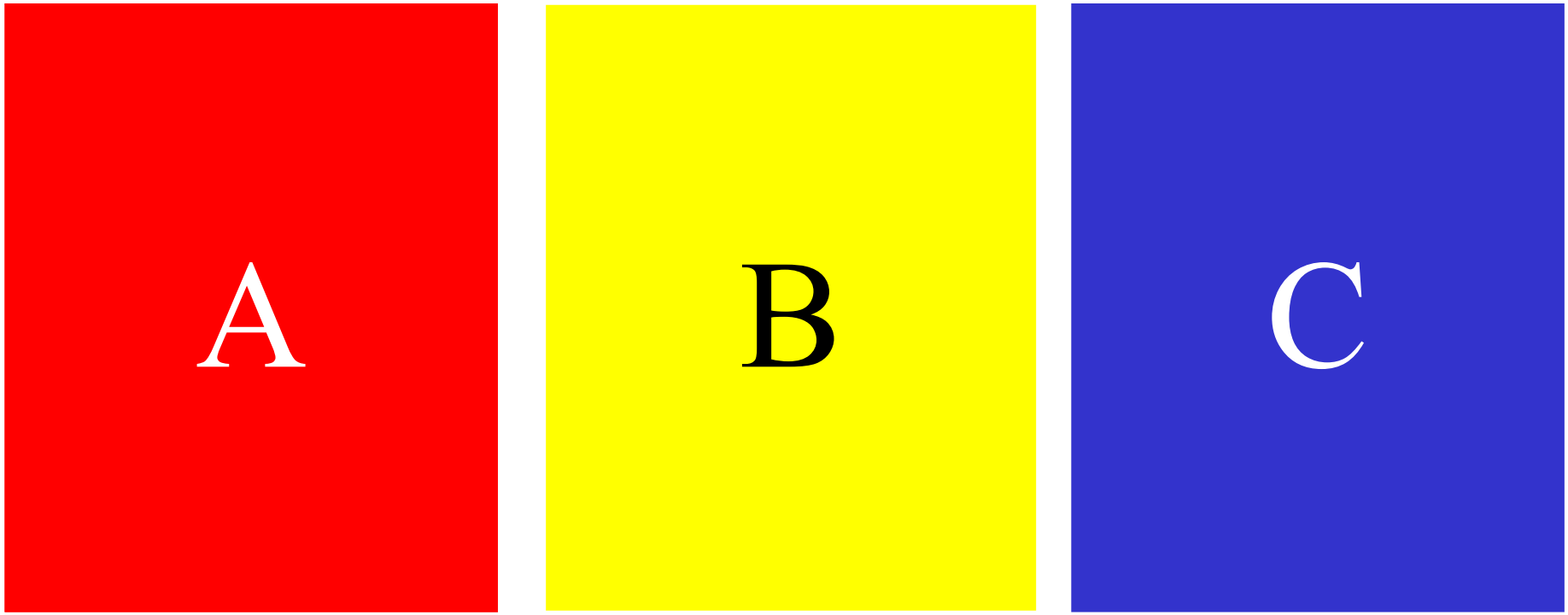
多元统计分析

Lecture 1： Basic Concept

Jun Li（李军）
School of Geography and Planning
Sun Yat-Sen University, Guangzhou, China
Mobile: 13922375250; Office: D307
E-mail: lijun48@mail.sysu.edu.cn; Webpage: http://www.lx.it.pt/~jun/

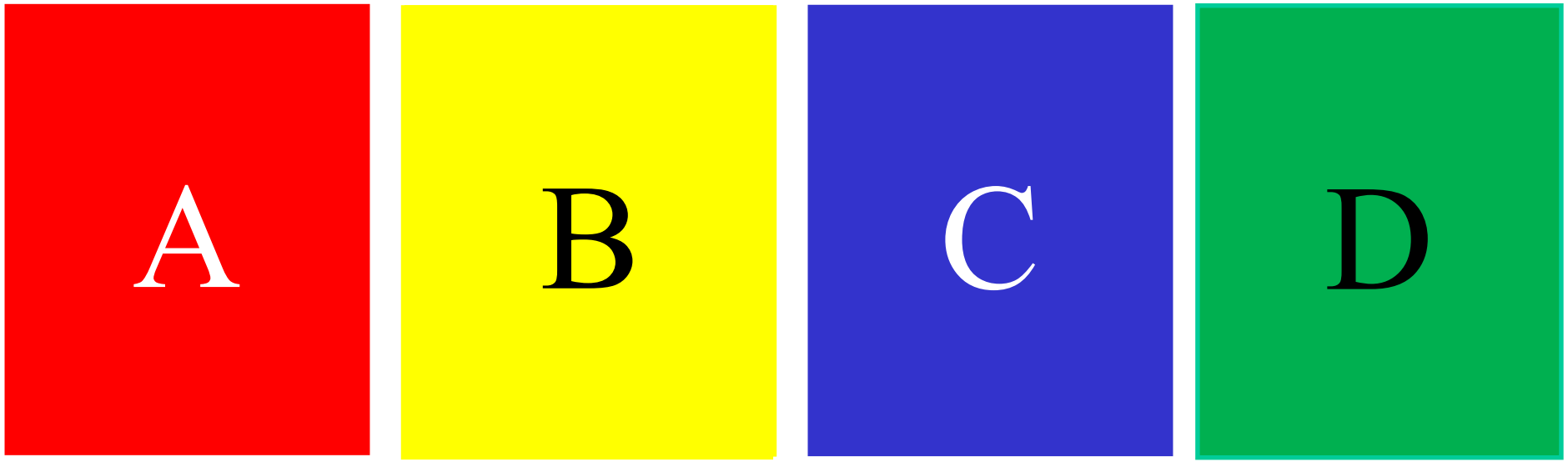
Discussion

其中一张后面有‘Good Luck’，请随机选一张，去掉一张空白的，请问是否更换选择？



Discussion

其中一张后面有‘Good Luck’，请随机选一张，去掉两张空白的，请问是否更换选择？



Let's Make a Deal:

Monty Knows



LOSER!!

RECAP: You originally picked door 2 and then switched to door 1.

Here is a summary of how previous contestants have fared.

	# of Players	Winners	Percent Winners
Switched	802	540	67.3
Didn't Switch	649	254	39.1

Play Again Play the Monty Does Not Know Version An Explanation

Back | Home | Programs | Documentation | Internet | People

[Back|Home|Programs|Documentation|Internet|People]

WIKIPEDIA
The Free Encyclopedia

Article Talk | Read Edit View history | Search

Monty Hall problem

From Wikipedia, the free encyclopedia

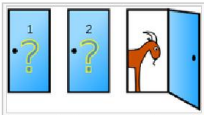
The **Monty Hall problem** is a brain teaser, in the form of a probability puzzle (Gruber, Krauss and others), loosely based on the American television game show *Let's Make a Deal* and named after its original host, Monty Hall. The problem was originally posed in a letter by Steve Selvin to the *American Statistician* in 1975 (Selvin 1975a), (Selvin 1975b). It became famous as a question from a reader's letter quoted in Marilyn vos Savant's "Ask Marilyn" column in *Parade* magazine in 1990 (vos Savant 1990a):

> Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?

Vos Savant's response was that the contestant should switch to the other door (vos Savant 1990a). Under the standard assumptions, contestants who switch have a 2/3 chance of winning the car, while contestants who stick to their choice have only a 1/3 chance.

Many readers of vos Savant's column refused to believe switching is beneficial despite her explanation. After the problem appeared in *Parade*, approximately 10,000 readers, including nearly 1,000 with PhDs, wrote to the magazine, most of them claiming vos Savant was wrong (Tierney 1991). Even when given explanations, simulations, and formal mathematical proofs, many people still do not accept that switching is the best strategy (vos Savant 1991a). Paul Erdős, one of the most prolific mathematicians in history, remained unconvinced until he was shown a computer simulation confirming the predicted result (Vazsonyi 1999).

The problem is a paradox of the *veridical* type, because the correct result (you should switch doors) is so counterintuitive it can seem absurd, but is nevertheless demonstrably true. The Monty Hall problem is mathematically closely related to the earlier Three Prisoners problem and to the much older Bertrand's box paradox.



In search of a new car, the player picks a door, say 1. The game host then opens one of the other doors, say 3, to reveal a goat and offers to let the player pick door 2 instead of door 1.

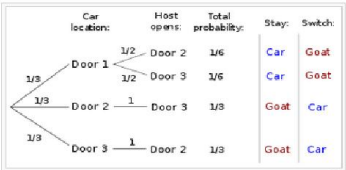
Contents [hide]

1 The paradox
2 Standard assumptions
3 Simple solutions
4 Vos Savant and the media furor

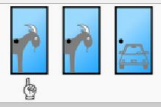
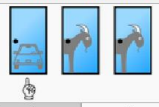
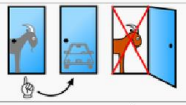
Conditional probability by direct calculation [edit]

By definition, the conditional probability of winning by switching given the contestant initially picks door 1 and the host opens door 3 is the probability for the event "car is behind door 2 and host opens door 3" divided by the probability for "host opens door 3". These probabilities can be determined referring to the conditional probability table below, or to an equivalent decision tree as shown to the right (Chun 1991; Carlton 2005; Grinstead and Snell 2006:137–138). The conditional probability of winning by switching is (1/3)/(1/3 + 1/6), which is 2/3 (Selvin 1975b).

The conditional probability table below shows how 300 cases, in all of which the player initially chooses door 1, would be split up, on average, according to the location of the car and the choice of door to open by the host.



Tree showing the probability of every possible outcome if the player initially picks Door 1

Car hidden behind Door 3 (on average, 100 cases out of 300)	Car hidden behind Door 1 (on average, 100 cases out of 300)		Car hidden behind Door 2 (on average, 100 cases out of 300)
Player initially picks Door 1, 300 repetitions			
			
Host must open Door 2 (100 cases)	Host randomly opens Door 2 (on average, 50 cases)	Host randomly opens Door 3 (on average, 50 cases)	Host must open Door 3 (100 cases)
			
Probability 1/3 (100 out of 300)	Probability 1/6 (50 out of 300)	Probability 1/6 (50 out of 300)	Probability 1/3 (100 out of 300)
Switching wins	Switching loses	Switching loses	Switching wins
On those occasions when the host opens Door 2, switching wins twice as often as staying (100 cases versus 50)		On those occasions when the host opens Door 3, switching wins twice as often as staying (100 cases versus 50)	

Discussion: More cards?

OPTIMAL STRATEGIES FOR THE PROGRESSIVE MONTY HALL PROBLEM

STEPHEN K. LUCAS, JASON ROSENHOUSE

1. Introduction

In the classical Monty Hall problem you are a contestant on a game show confronted with three identical doors. One of them conceals a car while the other two conceal goats. You choose a door, but do not open it. The host, Monty Hall, now opens one of the other two doors,

2. Switching Every Time

It has been our experience in presenting this problem to students that the strategy of switching doors at every opportunity is invariably popular. This discussion typically comes after a long struggle to persuade them of the benefits of switching in the classical version. The take home message seems to be that switching is a very good thing indeed, which might explain the popularity of this approach. In this section we will prove that your probability of success with this strategy approaches $1 - 1/e$ as $n \to \infty$.

Discusssion !

Jun Li (李军)
School of Geography and Planning
Sun Yat-Sen University, Guangzhou, China
Mobile: 13922375250; Office: D307
E-mail: lijun48@mail.sysu.edu.cn; Webpage: http://www.lx.it.pt/~jun/

中山大學
SUN YAT-SEN UNIVERSITY

HYPERCOMP
Hyperspectral Computing Laboratory

Multivariate Statistics Analysis:

多元统计分析

Lecture 1： Basic Concept

Jun Li（李军）
School of Geography and Planning
Sun Yat-Sen University, Guangzhou, China
Mobile: 13922375250; Office: D307
E-mail: lijun48@mail.sysu.edu.cn; Webpage: http://www.lx.it.pt/~jun/

Basics

1)总体（母体,matrix）:研究的全部元素的集合

2)个体（样本点,sample）:组成总体的每个元素

3)样本集（子样,set,vector）:从总体中取出的一部分个体的集合

Example 1

例如研究某花岗岩体中钾的含量（通常研究某一指标，即某一变量），若从该岩体中合理选取n个样品（n=3000），分析其中钾的含量为Ki（i=1，2，…，n），则

1)K1,K2,...或Kn等称为个体；

2)n个元素（个体）组成的集合（K1，K2,...，Kn）称为样本（子样）；

3)样本中包含的个体数目（n）称为样本的容量。一般样本容量n≥30称为大样本，n＜30称为小样本；

4)所有可能的个体的集合称为总体，通常地质体皆可无限取样，这时总体包含无限多个体。这样的总体称为无限总体。

Basics

1)多元总体：多个指标（变量,variable）

2)代表性(representative)：要求使总体的每一个个体都有相同的抽取机会。

3)独立性(independent)：要求每个观测结果既不影响其它观察结果。也不受其它观察结果的影响，也就是说抽样是独立的随机抽样。

Example 2

抛硬币，正、反面

Basics

1) 数字特征（特征数）：反映数据分布的集中位置，从而可以代表数据整体的特征数（表征数），称为整个代表性特征数（又叫集中性参数）；另一类是反映数据分布离散程度的参数，称为离散性特征数。

(1) 样本算术平均数

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

Basics

(1) 加权平均数

$$\bar{x}=\frac{c_1f_1+c_2f_2+\ldots+c_mf_m}{f_1+f_2+\ldots+f_m}=\frac{\sum_{j=1}^{m} f_jc_j}{\sum_{j=1}^{m} f_j}=\frac{1}{n}\sum_{j=1}^{m} f_jc_j$$

Example 3

测绘有10名学生，地信有12名学生，在一次测验中的成绩分别为：

测绘 = 70,72,74,76,78,82,84,86,88,90

地信 = 81,83,85,87,89,90,90,91,93,95,97,99

Basics

(1)几何平均数（Geometric mean），是求一组数值的平均数的方法中的一种。适用于对比率数据的平均，并主要用于计算数据平均增长（变化）率。

$$\bar{x} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

$$\ell g\bar{x} = \frac{1}{n}(\ell g x_1 + \ell g x_2 + \ldots + \ell g x_n) = \frac{1}{n}\sum_{i=1}^{n} \ell g x_i$$

气象.水.熵.复杂性 分享
http://blog.sciencenet.cn/u/zhangxw
张学文的文章，涉及气象、水分、熵、统计、复杂性、一般科学等

博客首页 动态 微博 博文 相册 主题 分享 好友 留言板

博文

问题：在什么场合应当用几何平均值，而不是用算术平均值？！

已有 14173 次阅读 2010-7-5 16:43 | 个人分类:一般科技.2. | 系统分类:观点评述 | 关键词:平均值，几何平均值，统计

问题：在什么场合应当用几何平均值，而不是用算术平均值？！

（7.6日注，也许题目改为在什么场合使用代数（几何，...）平均值合适？更妥当）
人们在很多场合（学生成绩统计、社会经济统计和科学实验等...）进行平均值的统计计算（先求N个样本的合计值再以样本数N除之）。这种平均值称为代数（或者算术）平均值。其实数学还推荐几何平均值（它是样本值的连乘积再开N次方），以至调和平均值（样本值的倒数的代数平均）。
在没有计算器或者计算机的时代，自然是代数平均值最容易计算，人们就习以为常地使用代数平均值了。但是现在计算机如此普及，普遍使用代数平均值究竟是一种习惯，还是满足某种理论要求？
在什么场合应当统计几何平均值，而不是用算术平均值？
本文引用地址： http://www.sciencenet.cn/m/user_content.aspx?id=341374

评论人1：zhaoxing [2010-7-5 16:59:41]
也被这一问题困惑很久，统计学的书看过十几二十种，但没有一本把这个问题说清楚过。
有的说几何平均适用于时序数据，有的说适用于增长或变化率计算，但都没说为什么。
从算法上，个人感觉几何平均可能更"平滑"一些，面对样本中存在极端值时的稳健性更好。
评论人2：lix [2010-7-5 18:05:19]
这个问题张老师是专家，我算回答张老师的课堂提问吧。我的理解，平均值是对一个分布的简化描述。这一简化，肯定要丢失一些信息。那么什么时候该用什么样的平均值，取决于您后继用这个平均值来干什么。比如说要用某地居民的平均收入来计算其平均幸福指数？由于岁入千万的个人，未必比岁入十万的幸福100倍（后继计算中有非线性），所以就可以考虑用几何平均数，压低岁入千万的个人对平均幸福指数的拔高效应。一般说来，可以考虑这个分布本身接近正态分布，还是这个分布的对数更接近正态分布。但是没有什么固定的原则。

张学文
加为好友 给我留言
打个招呼 发送消息

作者的精选博文 全部
- 胡焕庸线就那么重要吗?
- 漫话空中水
- 大规模调水对降水的影响的
- 感谢许局长为我的书写了序
- 乌鲁木齐地震了!
- 悼丁裕国教授!

作者的其他最新博文 全部
- 温度的年变幅与当地海拔高
- 秦大河院士与何祚休院士语
- 积雪天数与海拔高度的气候
- 天山附近气压年变幅的一个
- 图书价格排序符合幂律?
- 欢迎全国人大代表把改用智

热门博文导读 全部
- 毛左派为什么反柴静?
- 这些年我基金申请失败的...
- 对科学家来说，这是个好...
- 近十年不出人才
- 解读柴静不必动机

1. 代数平均值在很多理论分析中也常用到，它计算方便，理解容易，这没有什么不对之处。毕竟，我做统计，我做主。
2. 从数学角度看，几何平均值没有什么理由比代数平均值地位低一等。它的计算困难问题，也因为电脑的普及而消失。但是使用几何平均值时，必需注意数学上的连乘积与开方运算的脾气。需要明确，变量值（样本值）可能出现负数的情况，不能用样本的连乘积或者几何平均值，因为变量的负值会带来连乘积的值时正时负，让你不放心，开方还出现负数开方，更是不可理解。所以对于变量可能存在负值的样本（如摄氏气温）不能统计其几何平均值（连乘积）。
3. 类似地，变量可能为0的样本，会使连乘积=0，所以这类变量也不能统计几何平均值。即，变量可能=0，或者小于0的样本，不能分析其几何平均值。
4. 既然有上面这些注意点，何必还计算几何平均值？请注意有的变量本身就天然具有不可能为0，不能取负值的本性，抓住这个特征，说不定使更容易看透其规律。物体具有的动能、人的年龄、人具有的财富、百分比的值等很多变量仅能取正值。这些特点，有时需要把握和利用而不是忽略。
5. 利用信息熵最大（我称为复杂程度最大），仅配合代数平均值确定不变，可以推导出该随机变量的概率密度只能服从负指数分布。在这种知识提示下，你统计代数平均值自然是妥当的。而且你会发现，另外一批样本的代数平均值与第一批几乎相同，而其几何平均值却不同。
6. 另外，利用信息熵最大，仅配合几何平均值（不是代数平均值！）确定不变，可以推导出该随机变量的概率密度只能服从幂率分布。在这种知识提示下，你统计几何平均值自然是妥当的。而且会发现，另外一批样本的几何平均值与第1批几乎相同，而其代数平均值却不同。是的，现在幂率分布在分形研究里很时髦，为什么一些分形现象里满足幂率分布？因为该系统的变量几何平均值具有保守性（不变），并且满足熵最大（最混乱、最复杂）。
7. 前面两段说明，在系统里随机性体现信息熵最大，并且仅存在一个约束条件，如果这个条件是（而且仅是）代数平均值不变则分布为负指数分布，如果是（而且仅是）几何平均值不变，则符合幂率。

19

8. 如果某变量体现的不是负指数分布，也不是幂率，而是所谓gamma分布，它对应的统计特点是说明？答案是：该系统体现信息熵最大的同时（也可以夸张地说这是热力学第二定律的体现，我使用了这个神秘的定律！）受到两个（而且仅是两个）条件的约束：变量的代数平均值以及（同时）几何平均值也是不变量（这两个平均值不需要相同，但是几何平均值必然小于代数平均值）。即此时概率分布不再是负指数或者幂率而是gamma分布了。此时，你会发现取用两批不同的大样本，双方的代数平均值应当相等，而且双方的几何平均值也相等。这里变量的代数平均值、几何平均值同等重要。而这些认识为你从理论解释该分布为什么恰好如此，提供了依据。合适的平均值的选取可能是迈向理论分析的合理跳板。
9. 以上这些认识基本来自《组成论》里对分布与复杂程度最大的系统性分析。这里就点到为止了。以上考虑与李小文老师考虑分布问题是一致的。
10. 初中3年级200个学生的百米成绩平均值是多少？这里需要先问百米成绩是以秒计算，去求平均值，还是以速度计算平均值。要知道速度是目前计算百米成绩的秒数的倒数！你是统计速度的平均值还是统计速度的倒数的平均值？这联系着统计学里的调和平均值的利用。如果你从调和平均值的角度分析（包括对于的分布）更容易获得理论说明，那么就应当统计调和平均值。这一切取决于随后的分析需要，看看哪种平均值是你需要踩的恰当的阶梯。
11. 结合目前是近千万学生高考，其学生分数如果符合正态分布，统计其代数平均值就可以了。如果符合gamma分布（偏态），我认为需要代数平均值、几何平均值都分析，并且用最大熵下加上代数平均值不变和几何平均值不变给予理论说明（这个理论结论已经准备在哪里了）。
12. 以上说明可能不尽合适，供进一步讨论。

Basics

方差

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

$D(X)=E(X^2)-[E(X)]^2$

均方差（标准差）

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

$$S = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2} = \sqrt{\overline{x^2} - \bar{x}^2}$$

Basics

协方差（Covariance）

期望值分别为$E(X) = \mu$与$E(Y) = \nu$的两个实数随机变量X 与Y 之间的协方差定义为：

$$\operatorname{cov}(X, Y) = \mathrm{E}((X - \mu)(Y - \nu)),$$

其中E是期望值。它也可以表示为：

$$\operatorname{cov}(X, Y) = \mathrm{E}(X \cdot Y) - \mu\nu,$$

（1）设c是常数，则D(c)=0。

（2）设X是随机变量，c是常数，则有$D(cX)=c^2D(X)$。

（3）设 X 与 Y 是两个随机变量，则

$D(X+Y)= D(X)+D(Y)+2Cov(X,Y)$

$D(X -Y)= D(X)+D(Y)-2Cov(X,Y)$

特别的，当X，Y是两个不相关的随机变量则

$D(X+Y)=D(X)+D(Y),D(X-Y)=D(X)+D(Y),$

此性质可以推广到有限多个两两不相关的随机变量之和的情况。

（4）D(X)=0的充分必要条件是X以概率为1取常数值c，即X=c,a.s.其中E(X)=c。

（5）$D(aX+bY)=a^2DX+b^2DY+2abCov(X,Y)$。

Basics

极差

$$R = x_{\max} - x_{\min}$$

Discusssion !

Jun Li (李军)
School of Geography and Planning
Sun Yat-Sen University, Guangzhou, China
Mobile: 13922375250; Office: D307
E-mail: lijun48@mail.sysu.edu.cn; Webpage: http://www.lx.it.pt/~jun/