课程主页：http://staff.ustc.edu.cn/~ynyang/vector

QQ群: 373409155, 密码: 5502

第一讲 多元统计分析简介

2025.2.24

主讲：杨亚宁（ynyang@ustc.edu.cn)

助教：张明杰（mrqianmo@mail.ustc.edu.cn）

赖志鑫 （ lzx13421565815@mail.ustc.edu.cn ）

内容

1. 课程简介
2. 数据可视化
3. 超立方体
4. 超球体的体积、表面积
5. 超球体上的均匀分布（以概率研究几何）

符号约定：
向量/随机向量：黑正体小写字母，$\mathbf{x}, \boldsymbol{\theta}$
变量/随机变量：斜体小写字母，x, θ
矩阵：大写字母，X

注意我们不以大小写区分随机变量和变量（即不以大写X表示随机变量，不以小写x表示其实现）。

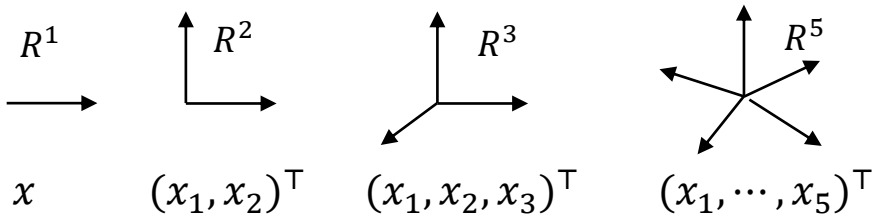
课程简介

先修

多变量微积分、线性代数、概率论、数理统计

多元/向量

多元统计分析（或多元分析）的主要研究对象是向量$\mathbf{x} \in R^p$
（向量：多元、多维、多变量）



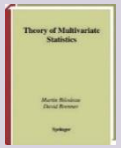
数据矩阵

向量数据 $\mathbf{x}_1, \ldots, \mathbf{x}_n \in R^p$ 按行排列组成 $n \times p$ 数据矩阵：

$$X = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \in R^{n \times p}$$

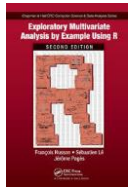
注意这里的X不是回归分析中的设计阵。一元线性回归模型一般认为不属于多元分析，这是因为主要研究对象响应变量是一元变量；如果响应是多元的（即多元线性回归）则属于多元分析。

课程内容

	前半学期：Normal	后半学期：Singular
主要内容	多元正态(normal distribution)，理解高维	奇异值分解(singular value decomposition), 统计学习
主要工具	多变量微积分	线性代数
参考书	M.Bilodeau, D.Brenner (1999) Theory of Multivariate Statistics. Springer （2-9章） 	R.Johnson and D.Wichern (2008) 实用多元统计分析, 第6版，英文版/中文版 （8-12章） 
大纲	•球对称分布 •多元正态 •卡方 •高斯图模型 •马氏随机场 •Wishart分布 •Hotelling's T^2 检验 •多元方差分析MANOVA •多元线性模型	•奇异值分解•主成分分析、双标图 •因子分析 •结构方程模型 •对应分析 •典则相关分析 •距离与相似系数 •配列 •多维标度法 •聚类 •分类

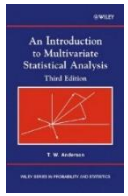
其它参考书

F.Husson, S.Le, J.Pages(2017) *Exploratory Multivariate Analysis by Example Using R*. CRC（法）



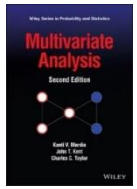
应用(250页)，仅含主成分分析，对应分析。法国学派。我们只采用其中一或两个数据例子。

T.W.Anderson (2003) *An Introduction to Multivariate Statistical Analysis*, Wiley, 3rd ed（美,许宝騄的学生）



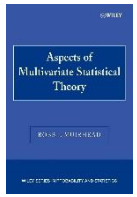
理论、经典全面(700+页)、无实际例子，供查阅。

K.V.Mardia, J.T.Kent, J.M.Bibby (1979, 2024) *Multivariate Analysis*, Academic Press （英）



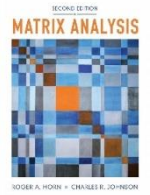
理论、经典(400页)（无电子版)。

Robb J. Muirhead (2005) Aspects of Multivariate Statistical Theory, 2nd ed., Wiley （美）



理论，Jacobian，外微分。

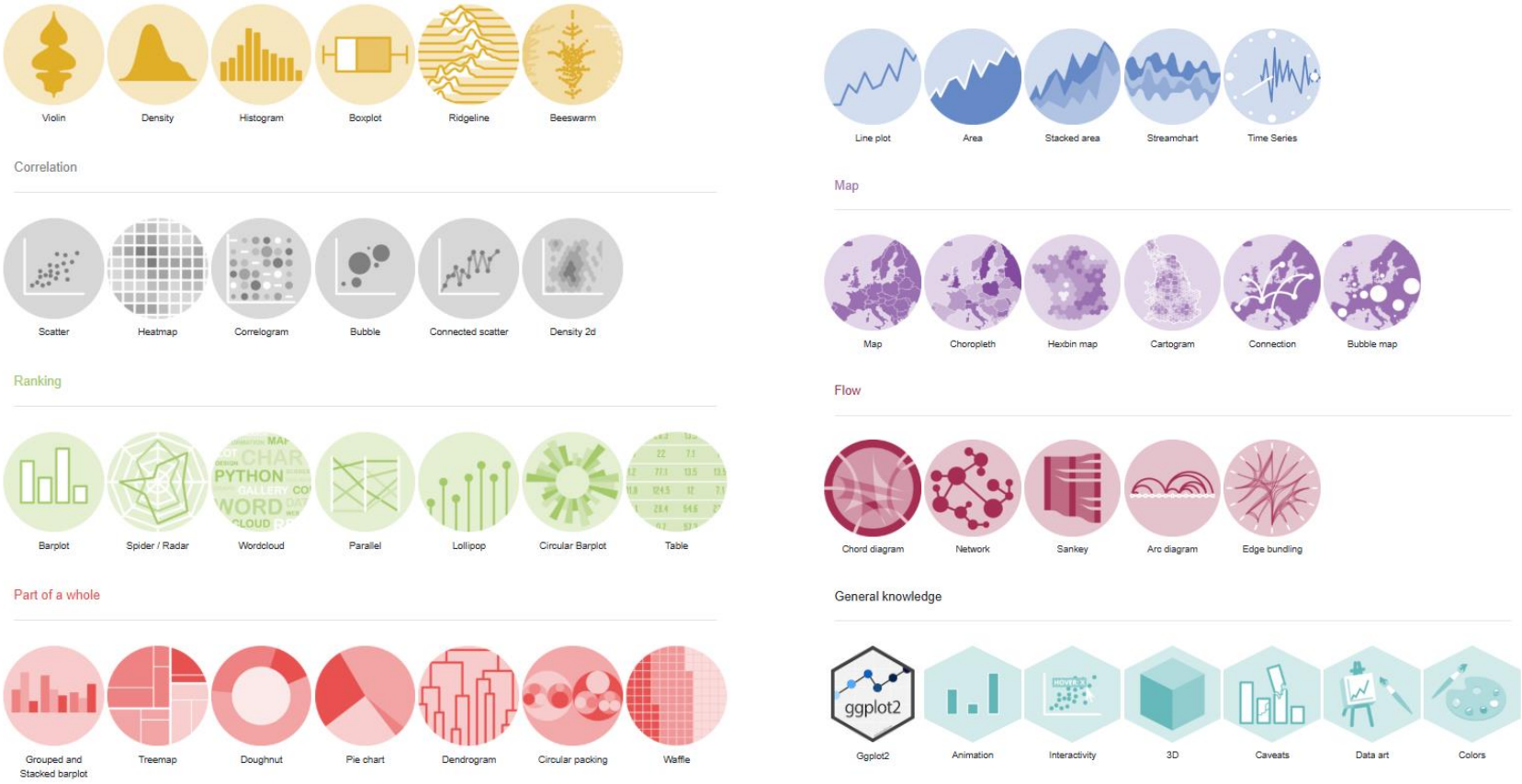
R. Horn, C. Johnson (2013) *Matrix Analysis*, 2nd ed. Cambridge University Press.



这些书目一般不需要翻看。当需要阅读参考书的某些章节时，我们会在课程主页“阅读材料”中指定。

数据可视化

The R Graph Gallery （https://r-graph-gallery.com/）列举了常见的数据可视化工具：



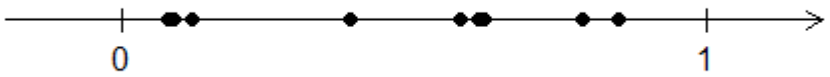
一元数据

- 散点图：plot
- 分布：直方图，盒型图，枝叶图

散点图/数轴点

实轴描点有助于了解一维数据的大小次序、间隔甚至分布。比如，“随机取10个点”通常指的是从均匀分布中产生10个随机数，其均匀性如何表现？下面产生10个[0,1]区间上的均匀随机数:

x=(0.389, 0.583, 0.095, 0.853, 0.787, 0.119, 0.606, 0.081, 0.391, 0.619)



可以看到，均匀随机数并不是我们想象的那么”均匀”，数值之间的间隔(spacing) 差别较大，容易出现聚簇（样本量较大时，每个局部都是如此）。

简单情形：

- 一个$U(0,1)$随机数将[0,1]区间划分成2段，较小一段的期望为1/4;
- 两个$U(0,1)$随机数将[0,1]区间划分成3段，最小段的期望等于1/9，最小段长度小于0.1的概率大约为0.5

Spacing/ 间隔

一般结果: 假设 $x_1, \dots, x_n \ iid \sim U(0,1)$,从小到大排列记为次序统计量 $x_{(1)} \le \cdots \le x_{(n)}$, 间隔spacing定义为:

$$d_i = x_{(i)} - x_{(i-1)}, i = 1, \dots, n+1, \text{其中} x_{(0)} = 0, x_{(n+1)} = 1,$$

已知事实:

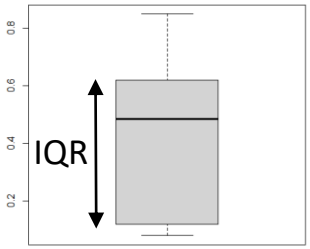
- 间隔期望相同: $E(d_i) = \frac{1}{n+1}$
- 次序统计量服从$\{(t_1, \dots, t_n): \ 0 \le t_1 \le \cdots \le t_n \le 1\}$上的均匀分布。
- $d_1, \dots, d_{n+1}$ 服从均匀分布$U(\Delta), \Delta = \{(d_1, \dots, d_{n+1}): d_i \ge 0, d_1 + \cdots + d_{n+1} = 1\}$。

问题:

- $\mathrm{P}(\min(d_i) < t) = ?$
- $E(\min(d_i)) = 1/(n+1)^2$?

分布

> boxplot(x) :

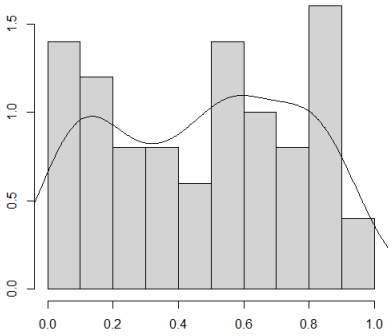


Interquantile range (度量分散程度):

IQR=75%分位数−25%分位数

=0.6175 − 0.1875

> hist(x， prob=T) :



> stem(x)

The decimal point is 1 digit(s) to the left of the |

0 | 892
2 | 99
4 | 8
6 | 129
8 | 5

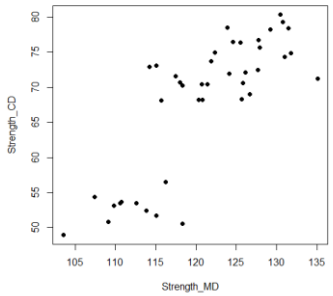
二元数据

- 二元散点图: plot
- 二元分布： image, persp, contour

散点图

散点图(scatter plot)是最基本、也是最重要的数据展示方法。

例1. 纸张的强度在机器制造方向（MD：machine direction）和与之垂直的方向（CD：cross direction）有所不同，课本Table1.2（数据集：paper ）提供了41张纸张的三项指标：x=Strength_MD, y=Strength_CD, z=Density （密度）。

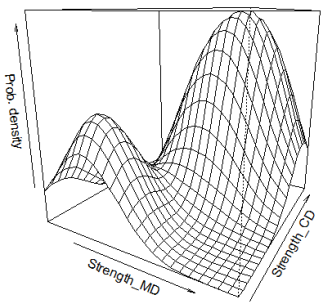


两个strength正相关数据聚簇为两类

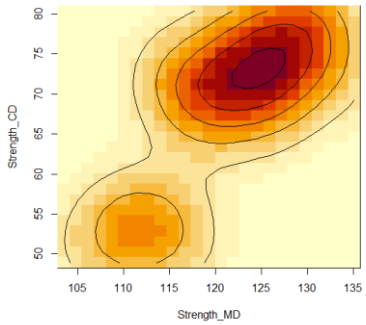
二元分布

透视图(perspective)、热图(heat map)、等高线图(contour)刻画二元数据(x,y)的分布：

persp()



image(), contour()



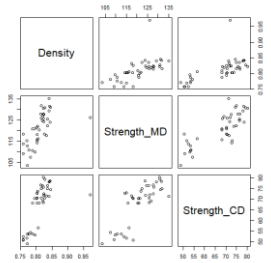

```
## kde2d估计概率密度 library(MASS)
k <- kde2d(paper[,2],paper[,3], n=25) #n: x,y轴划分区间的个数
#二维变量的密度函数（左）和概率密度的热图、等高线图：
persp(k, xlab="x", ylab="y",zlab="Prob. density",theta=30 )
image(k, xlab="Strength_MD", ylab="Strength_CD" )
contour(k, add = TRUE, drawlabels = FALSE,nlevels=6)
```

三元数据

- 对边际应用一、二元数据的图示方法。
- 三维散点图(3dplot)

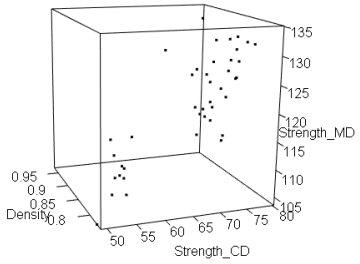
两两二维散点图

例1（续）
>plot (paper)
>pairs(paper)



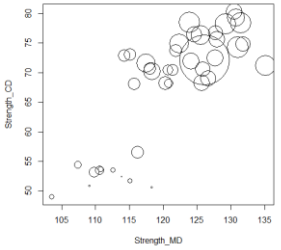
透视散点图

>library(rgl)
>plot3d(paper) #手工拖动旋转



泡泡图
bubble plot

在x-y散点图上，可利用每个数据点的大小、颜色或形状表示第三个变量z。
plot(paper[,2:3],type="n")
symbols(paper[,2:3], circles = paper[,1]-0.75,add=T,inches=0.5)

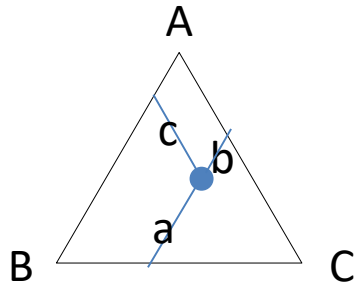


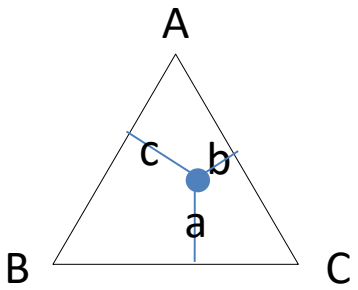
每个点的大小代表强度Density

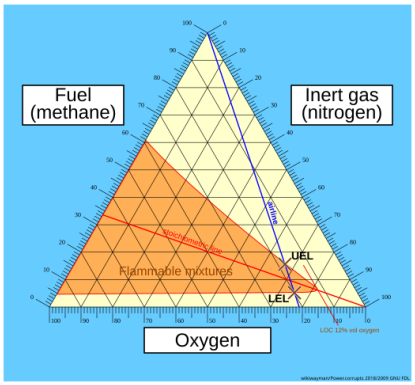
即使是三元的数据实际上也只能在2维展示。

三元图 Ternary plot

等边三角形顶点代表三个变量/元素A,B,C，任一样品三种元素的含量分别为 a, b, c, $a+b+c=1$。三原图用重心坐标(barycentric coordinate)表示含量, 有多种画法。下图（左）过样本点与三个边做平行线，右图向三个边做垂线，A对面的线段长度为A的含量a (两图中，三个线段长度都是常数)。



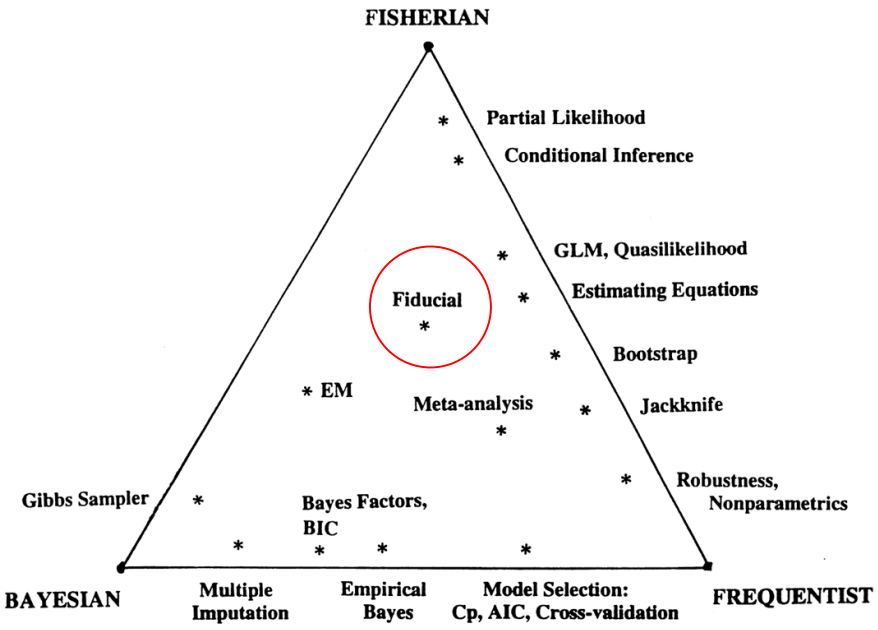




甲烷、氧气、氮气混合比例（红色区域可燃）。

土质划分(sand,clay,silt比例)

再如，Efron (1998)利用三元图描述了后Fisher时代发展起来的各种统计方法与三个统计学派(Frequency, Bayes, Fisher) 的关系：

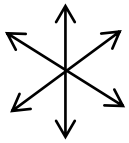


B. Efron (1998) R. A. Fisher in the 21st century. *Statistical Science* 13(2):95-122

多元数据

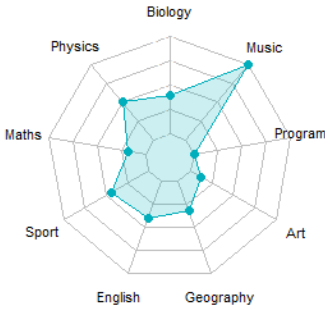
- 雷达图/蜘蛛图
- 降维、投影后可视化

雷达图

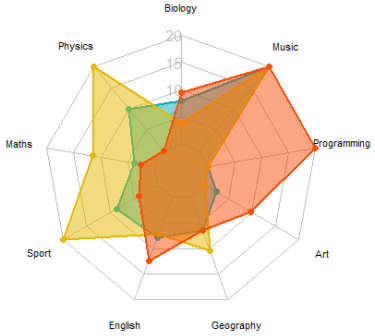


6维坐标系

雷达图/蜘蛛图在二维平面上表示多维向量的各个分量**取值大小**（而不是空间位置）。对于非负元素的n维实向量，在平面上画出n个坐标轴/箭头，代表n个维度，在每个坐标轴标出每个维度的坐标值，相邻的坐标连线，形成n边形。



9个科目成绩



3个人的成绩

马龙: 六边形战士

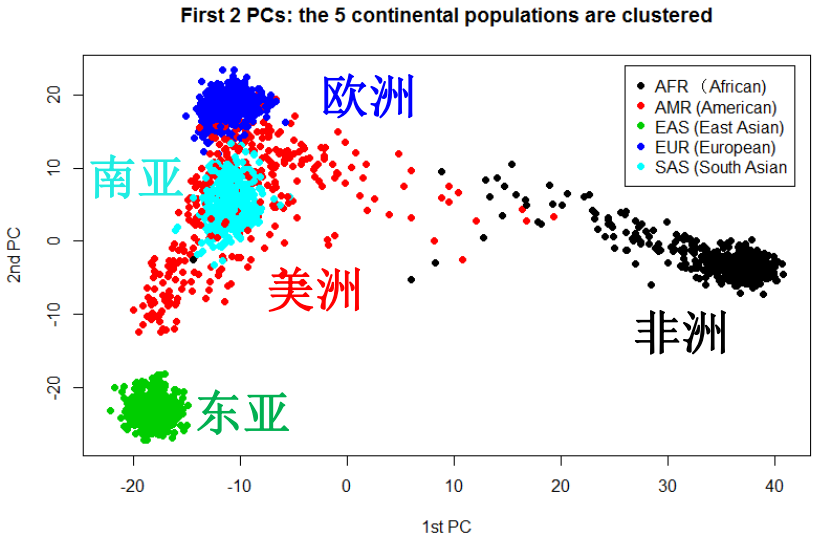
降维/投影

人类只能看见二维，通过光线、透视感知三维。多元统计提供了很多高维数据降维方法，最常用的是主成分分析方法(PCA)。

例2.（基因组数据） 2504个人，每个人在9932个位点上的基因值为0，1，2, 因此每个人的数据是9932 × 1向量。下表是部分数据（行代表人，列代表基因）。此外还有种族信息：AFR (African), AMR (American), EAS (East Asian), EUR (European), SAS (South Asian)。下图为主成分分析降低到2维的散点图。

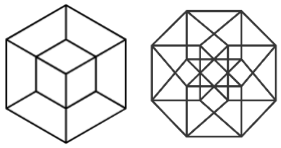
Race	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	…	V9932	V9932
AFR	2	0	0	1	2	0	2	1	1	2	0		2	2
SAS	2	0	0	1	2	0	2	0	0	1	0		2	2
…														

PCA

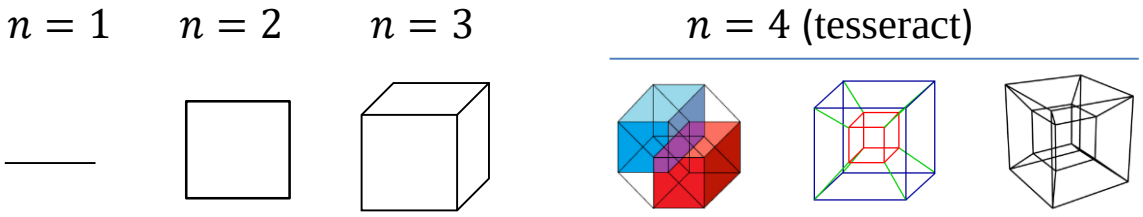


可视化高维几何体

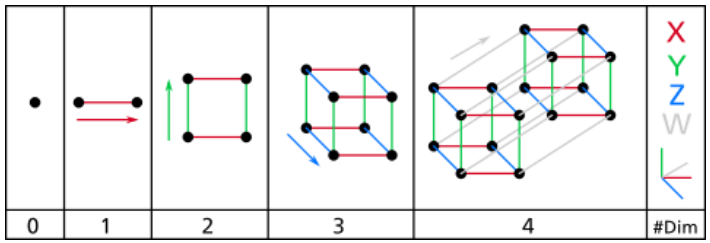
四维正方体
Tesseract

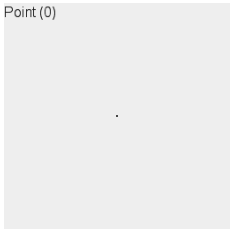


一维正方体是线段，二维是正方形，三维是立方体，都可以看作是低维立方体在新增维度上平移得到的。故四维超正方体(tesseract)可以想象为3维正方体沿第4个维度平移得到。



解释：3维立方体在第4个维度方向上平移：6个正方形平移后构成6个3d立方体，加上原来的立方体和平移后的立方体， 共8个3d立方体，它们构成4维立方体tesseract的8个三维“表面”



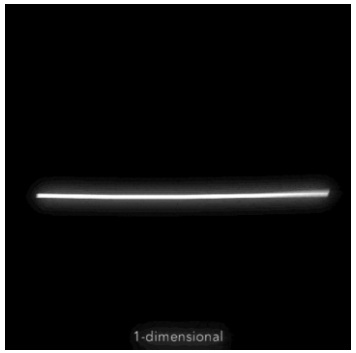


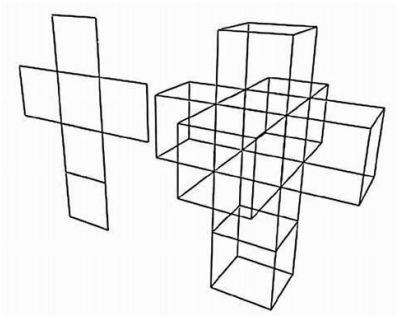
https://en.wikipedia.org/wiki/Tesseract

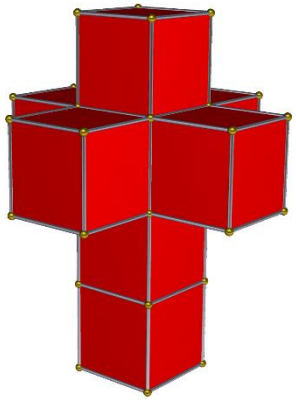
1909年《科学美国人》杂志征集对四维空间的通俗合理的解释。大多数参赛论文都提到了数学家查尔斯.霍华德.辛顿(Charles Howard Hinton， Geoffrey Hinton的曾祖父)的超立方体可视化方法。辛顿可视化方法包括解析法、投影法和切面法。

下图左是3维立方体盒子的6个表面的展开，右图类似地将四维立方体的8个3维立方体表面的展开，称为辛顿超立方体，后来甚至成了神秘主义符号。

展开立方体







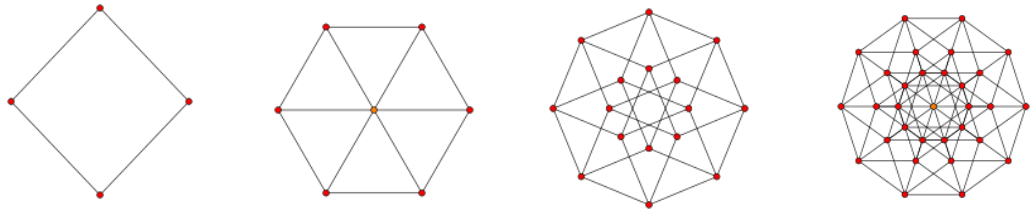
超正方体 Hypercube

较高维数欧氏空间中的超立方体

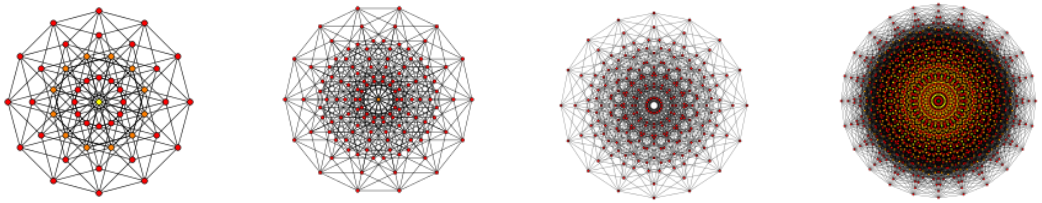
$$[0,1]^n = \{(x_1, \dots, x_n): 0 \le x_i \le 1\} \subset R^n$$

的2维投影会很杂乱，Petrie投影(Petrie polygon orthographic projections)提供了一种对称、规则的投影表示(以超立方体的$2n$个$n-1$维表面为边)：

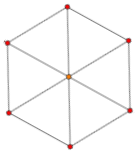
$n = 2$ $n = 3$ $n = 4$ $n = 5$



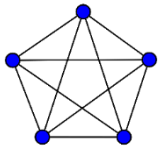
$n = 6$ $n = 7$ $n = 8$ $n = 12$



注：$n = 3$时，旋转60度才是通常的立方体画法



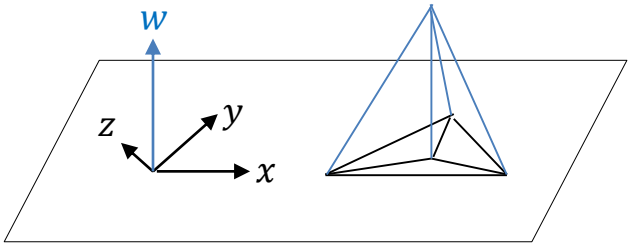
单纯形 Simplex

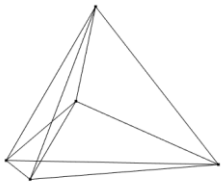


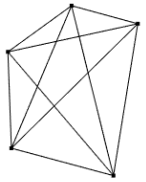
三角形的高维版本是单纯形

$$\{(x_1, \ldots, x_n): x_1 + \cdots + x_n = 1, x_i \geq 0\}$$

理解三维以上单纯形不能沿用前述低维平移方法, 我们可以将3维单纯形/四面体投影到平面上，然后再第四个维度w取一点，并与四面体的各个点连结，得到4维单纯形（五胞体）的二维投影：

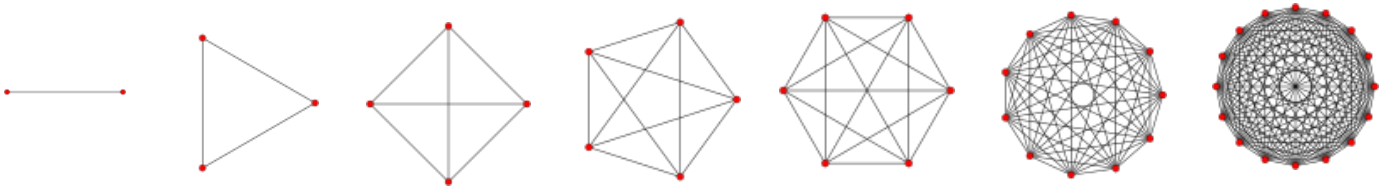




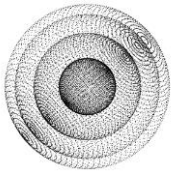


投影（Petrie polygons， $n+1$个顶点，两两连线）：

$n = 1$ $n = 2$ $n = 3$ $n = 4$ $n = 5$ $n = 10$ $n = 10$

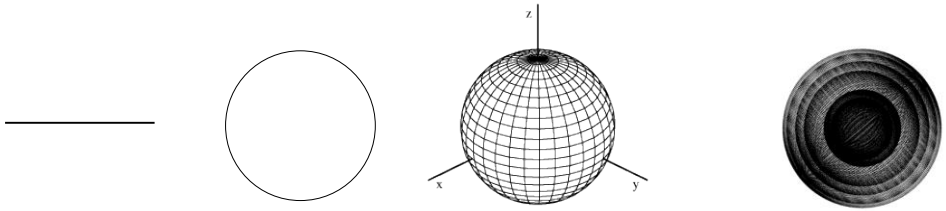


超球体 Hyperball



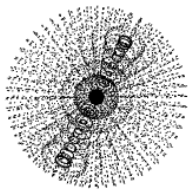
一维球是线段，二维是圆周，三维是球体，3维球可看作是由多个递增的二维圆盘沿法向平行堆积而成。以此类推，多个递增的三维球沿第四个维度平行堆积而成4维球。

$n = 1$ $n = 2$ $n = 3$ $n = 4?$

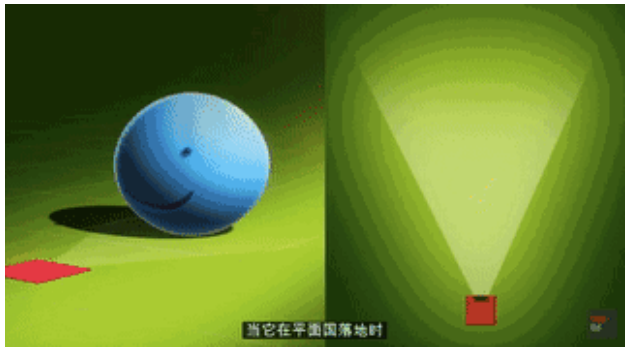


$$x_1^2 + x_2^2 + x_3^2 + x_4^2 \leq 1$$

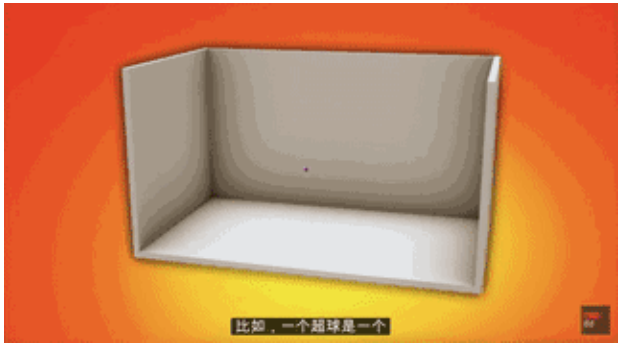
四维及以上的超球体(hyperball、hypersphere因为没有有限个顶点、边或面，超球体的2维投影都是圆盘，且无法展示其上点的紧邻关系或连续性质。



三维球在平面的投影/切片是圆盘，四维球在三维空间的投影是三维球。 当四维球穿越三维空间时，我们会看到凭空出现一个三维球越来越大，而后越来越小直至凭空消失（右图）：



穿越2维平面的三维球



穿越3维空间的四维球

但这些图示不足以（甚至极可能不正确）提供一个完整的4维或更高维的高维球的理解，我们只能通过数学语言研究高维球的性质。

超球体的体积和表面积

超球体和超球面

R^n中原点为中心、半径为R的超球体(hyperball, n-ball)和超球面(hypersphere, $(n-1)$-sphere):

$$B^n(R) = \{\mathbf{x} \in R^n: \|\mathbf{x}\| \le R\},$$

$$S^{n-1}(R) = \{\mathbf{x} \in R^n: \|\mathbf{x}\| = R\},$$

半径$R = 1$时分别记作B^n，S^{n-1}。

注: $S^{n-1}(R) \subset R^n$，维数$n-1$.

表面积：指的是$n-1$维边界的体积。

体积和面积公式

定理1. n维欧氏空间中半径为R的球体的体积和表面积：

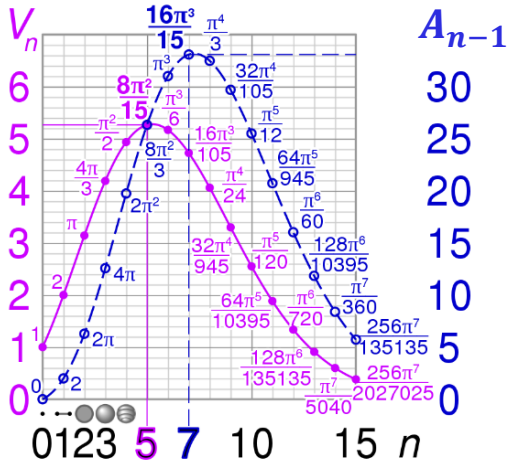
$$V_n(R) = |B^n(R)| = \frac{\pi^{n/2}}{\Gamma(n/2+1)} R^n = \frac{R}{n} \times A_{n-1}(R)$$

$$A_{n-1}(R) = |S^{n-1}(R)| = \frac{n\pi^{n/2}}{\Gamma(n/2+1)} R^{n-1}$$

- $\frac{dV_n(R)}{dR} = A_{n-1}(R)$,
- $V_n(R) = \int_0^R A_{n-1}(r)dr$。

性质1：单位球的体积和表面积分别在$n=5$、7时最大，之后单调下降趋于0. 半径不为1的的球体体积和表面积同样也在某个n处达到极大（未必是5，7），之后同样趋于0。

https://en.wikipedia.org/wiki/N-sphere



性质2：球的大部分体积集中在球表面附近。 从球体积公式容易知道，单位球表面厚度为ϵ的球壳体积

$$\Delta_n = V_n(1) - V_n(1-\epsilon) = \frac{\pi^{n/2}}{\Gamma(n/2+1)}1^n - \frac{\pi^{n/2}}{\Gamma(n/2+1)}(1-\epsilon)^n$$

它占总体积的比例即概率$\frac{\Delta_n}{V_n(1)} = 1-(1-\epsilon)^n \to 1, n \to \infty$。例如

$$1-(1-0.01)^{100} = 63.4\%,\quad 1-(1-0.01)^{500} = 99.34\%,$$

我们将看到，这一现象等同于球内均匀分布的径向分布集中于1附近。

球坐标变换

球坐标变换$\mathbf{x} \in R^n \to (r, \boldsymbol{\theta}) = (r, \theta_1, \ldots, \theta_{n-1}) \in R^n$,

$$\begin{cases} x_1 = r\cos(\theta_1) \\ x_2 = r\sin(\theta_1)\cos(\theta_2) \\ \vdots \\ x_{n-1} = r\sin(\theta_1)\cdots\sin(\theta_{n-2})\cos(\theta_{n-1}) \\ x_n = r\sin(\theta_1)\cdots\sin(\theta_{n-2})\sin(\theta_{n-1}), \end{cases}$$

$$r = \|\mathbf{x}\| \geq 0,\ 0 \leq \theta_1, \ldots, \theta_{n-2} \leq \pi,\ 0 \leq \theta_{n-1} \leq 2\pi$$

引理1：球坐标变换的Jacobian：

$$J = J(r, \boldsymbol{\theta}) = J(\mathbf{x} \to r, \boldsymbol{\theta}) = r^{n-1}\sin^{n-2}(\theta_1)\,\sin^{n-3}(\theta_2)\cdots\sin(\theta_{n-2})$$

单位球面的面积元

$$d\sigma = \sin^{n-2}(\theta_1)\,\sin^{n-3}(\theta_2)\cdots\sin(\theta_{n-2})\,d\theta_1\cdots d\theta_{n-1}$$

球坐标系下的体积元

$$dv = J(r, \boldsymbol{\theta})drd\boldsymbol{\theta} = r^{n-1}drd\sigma$$

https://en.wikipedia.org/wiki/N-sphere

定理1的证明

我们利用球坐标变换求解n维球的体积和表面积。记

$$D_n = \{(r, \theta_1, \cdots, \theta_{n-1}): 0 < r \le R, 0 < \theta_1, \ldots, \theta_{n-2} \le \pi, 0 < \theta_{n-1} < 2\pi\}$$

由引理1，$B^n(R)$的体积:

$$|B^n(R)| = \int_{\|\mathbf{x}\| \le R} d\mathbf{x} = \int_{D_n} r^{n-1} dr d\sigma = \int_0^R r^{n-1} dr \times \int_{S^{n-1}} dr d\sigma$$

$$= \frac{R^n}{n} \times \int_{S^{n-1}} d\sigma = \frac{R^n}{n} \times |S^{n-1}|$$

其中单位球面面积

$$|S^{n-1}| = \int_{S^{n-1}} d\sigma$$

$$= \int_0^{2\pi} \cdots \int_0^{\pi} \sin^{n-2}(\theta_1) \sin^{n-3}(\theta_2) \cdots \sin(\theta_{n-2}) \, d\theta_1 \cdots d\theta_{n-1} = \frac{n\pi^{n/2}}{\Gamma(n/2+1)}$$

$$\int_0^{\pi} \sin^{n-k-1}(\theta_k) \, d\theta_k = B\left(\frac{n-k}{2}, \frac{1}{2}\right)$$

令$t = \sin^2(\theta_k)$

$$\Rightarrow |B^n(R)| = \frac{\pi^{n/2} R^n}{\Gamma(n/2+1)}, \quad |S^{n-1}(R)| = R^{n-1}|S^{n-1}| = \frac{n\pi^{n/2}}{\Gamma(n/2+1)} R^{n-1}$$

引理2. 假设$g(r)$是一元连续函数，则

$$\int_{\|\mathbf{x}\|\le R} g(\|\mathbf{x}\|)d\mathbf{x} = A_{n-1}\int_0^R g(r)r^{n-1}dr,$$

其中$A_{n-1} = |S^{n-1}| = \frac{n\pi^{n/2}}{\Gamma(n/2+1)}$为单位超球面$S^{n-1}$的面积。

证明： 引入球坐标变换$\mathbf{x} \to (r, \boldsymbol{\theta}) = (r, \theta_1, \dots, \theta_{n-1})$，

$$\int_{\|\mathbf{x}\|\le R} g(\|\mathbf{x}\|)d\mathbf{x} = \int_{D_n} g(r)r^{n-1}drd\sigma = \int_0^R g(r)r^{n-1}dr\int_{S^{n-1}} d\sigma$$

$= A_{n-1}\int_0^R g(r)r^{n-1}dr$。

注：引理2将在后面反复用到。在已知球表面积的条件下，引理2的证明1其实不需要球坐标的具体形式，只用到体积元变换(今后将主要采用这种表达)：

$$d\mathbf{x} = r^{n-1}\, drd\sigma,$$

关于引理2可参考：

- 程艺，陈卿，李平（2019），数学分析讲义，第三册 ，p223.
- 常庚哲，史济怀（2012）数学分析教程下册，p60, 例3.
- Courant, R. (1936). *Differential and Integral Calculus II*. Wiley，p303

下面应用引理2和正态分布的已知结论给出定理1的另外一个证明（避免使用球坐标积分或gamma积分）

定理1的证明2：

易知 $A_{n-1}(r) = r^{n-1}A_{n-1}(1)$，$V_n(r) = r^n V_n(1)$.

由引理2，

$$V_n(R) = \int_0^R dV_n(r) = \int_0^R A_{n-1}(r)dr = A_{n-1}(1)\int_0^R r^{n-1}dr = \frac{A_{n-1}(1)R^n}{n}$$

只需求$A_{n-1}(1) = |S^{n-1}|$.

再由引理2，

$$\int\cdots\int e^{-x_1^2-\cdots-x_n^2}dx_1\ldots dx_n = \int_0^R e^{-r^2}dV_n(r)$$

$$= A_{n-1}(1)\int_0^R e^{-r^2}r^{n-1}dr = A_{n-1}(1)\Gamma(n/2)/2$$

但是左端$= \left(\int e^{-x_1^2}dx_1\right)^n = \pi^{n/2}$, 所以$A_{n-1}(1) = \frac{2\pi^{n/2}}{\Gamma(n/2)}$.

定理1的证明3（递归方法）：

$$V_n(1) = \int \cdots \int_{x_1^2+\cdots+x_n^2\le 1} dx_1 \ldots dx_n = \int_{-1}^{1} dx_1 \int \cdots \int_{x_2^2+\cdots+x_n^2\le 1-x_1^2} dx_2 \ldots dx_n$$

$$= \int_{-1}^{1} V_{n-1}\left(\sqrt{1-x_1^2}\right) dx_1 = V_{n-1}(1) \int_{-1}^{1} (1-x_1^2)^{(n-1)/2} dx_1$$

$$= 2^n(\Gamma((n+1)/2))^2/\Gamma(n+1) V_{n-1}(1)。$$

$$x_1 = \pm\sqrt{1-(x_2^2+\cdots+x_n^2)}$$

$$A_{n-1}(1) = 2\int \cdots \int_{x_2^2+\cdots+x_n^2\le 1} \sqrt{1+\left(\frac{\partial x_1}{\partial x_2}\right)^2+\cdots+\left(\frac{\partial x_1}{\partial x_n}\right)^2}\, dx_2 \cdots dx_n$$

$$= 2\int \cdots \int_{x_2^2+\cdots+x_n^2\le 1} \frac{1}{\sqrt{1-(x_2^2+\cdots+x_n^2)}} dx_2 \cdots dx_n = \int_0^1 \frac{1}{\sqrt{1-r^2}} dV_{n-1}(r)$$

$$= A_{n-2}(1)\int_0^1 \frac{1}{\sqrt{1-r^2}} r^{n-2} dr = 2A_{n-2}(1)\text{Beta}\left(\frac{1}{2}, \frac{n-1}{2}\right),$$

由上述两个递推式，即可求得$V_n(1)$，$A_{n-1}(1)$。

参见：

常庚哲，史济怀（2012），数学分析教程， p57 ， p398-399.

程艺，陈卿，李平（2019），数学分析讲义，第二册 ， p176.

教材中用的是$V_n(1)$和$V_{n-2}(1)$的递归

这是importance sampling的基本想法

黎曼积分~均匀分布积分：

$$\int_{B^n} g(\mathbf{x})d\mathbf{x} = \int_{B^n} g(\mathbf{x})\frac{1}{|B^n|}d\mathbf{x} = |B^n|Eg(\mathbf{x}), \mathbf{x}\sim U(B^n)$$

$$\int_{S^{n-1}} g(\mathbf{x})d\sigma = \int_{S^{n-1}} g(\mathbf{x})\frac{1}{|S^{n-1}|}d\sigma = |S^{n-1}|Eg(\mathbf{x}), \mathbf{x}\sim U(S^{n-1})$$

下面通过概率方法研究几何(理论计算或蒙特卡洛，主要是后者)，具体地，以球内均匀分布$U(B^n)$，球面均匀分布 $U(S^{n-1})$研究球和球面。

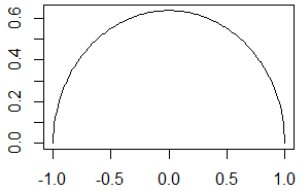
球内、球面均匀分布

圆盘均匀分布

我们首先以R^2单位圆盘为例: $B^2 = \{(x_1, x_2): x_1^2 + x_2^2 \le 1\}$,

假设(x_1, x_2)在圆盘内服从均匀分布，记作 $(x_1, x_2) \sim U(B^2)$,
其概率密度:

$$p(x_1, x_2) = \frac{1}{\pi}, \ x_1^2 + x_2^2 \le 1,$$



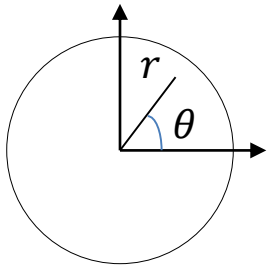
则x_1服从Wigner半圆律分布（ semicircle law ）

$$p(x_1) = \frac{2}{\pi}\sqrt{1 - x_1^2}, \ |x_1| \le 1$$

极坐标变换 $x_1 = r\cos(\theta), x_2 = r\sin(\theta)$, 则

$$\theta \sim U(0, 2\pi), \ r \sim p(r) = 2r, \ 0 < r < 1$$

角度均匀分布，模长r的概率密度单调增加，说明
球内均匀分布的点倾向于靠近圆周附近。



注：调整r的分布可产生其它分布,比如二元正态/Box-Muller方法。

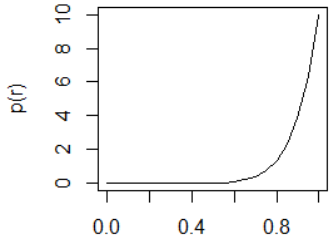
B^n球内均匀分布

随机向量$\mathbf{x}=(x_1,\ldots,x_n)^\top$服从$B^n=B^n(1)=\{\mathbf{x}:\ \|\mathbf{x}\|\le 1\}$内均匀分布，即$\mathbf{x}\sim U(B^n)$，概率密度

$$f(\mathbf{x})=\frac{1}{|B^n|}=\frac{\Gamma(1+n/2)}{\pi^{n/2}},\ \mathbf{x}\in B^n.$$

命题1. 假设$\mathbf{x}\sim U(B^n)$，$r=\|\mathbf{x}\|$的概率密度

$$p(r)=nr^{n-1},0<r\le 1\Leftrightarrow r^n\sim U(0,1)$$



$r=\|\mathbf{x}\|$的概率密度

证：半径$r=\|\mathbf{x}\|$到$r+\Delta r$的球壳体积为$\Delta V=|S^{n-1}(r)|\Delta r$

$=\frac{n\pi^{n/2}r^{n-1}}{\Gamma(n/2+1)}\Delta r$，所以球壳上的概率质量为

$$f(\mathbf{x})\Delta V=f(\mathbf{x})\times\frac{n\pi^{n/2}r^{n-1}}{\Gamma(n/2+1)}\Delta r=\frac{\Gamma(1+n/2)}{\pi^{n/2}}\times\frac{n\pi^{n/2}r^{n-1}}{\Gamma(n/2+1)}\Delta r$$

$$=nr^{n-1}\Delta r\triangleq p(r)\Delta r$$

所以$p(r)=nr^{n-1},\ 0<r\le 1$.

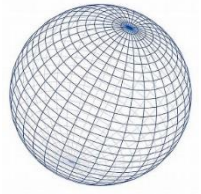
注：$n>1$时，r的概率质量集中于1附近:

$$P(1-\epsilon<r<1)=P\left((1-\epsilon)^n<r^n<1\right)=1-(1-\epsilon)^n$$

因为概率与体积成正比，所以球体积大部分集中于球壳表面附近.

更多球内均匀分布性质可通过引入球坐标进行研究（第二讲）. 下面我们以球面均匀分布研究超球面，因为其理论计算（第3讲）较为复杂，我们将主要通过蒙特卡洛方法考察球面均匀分布的性质。

球面均匀分布



定义1. R^n中原点为中心、半径为1的单位超球面,即$(n-1)$-sphere

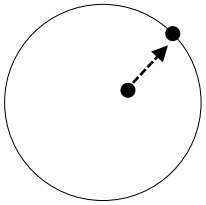
$$S^{n-1} = \{\mathbf{x} \in R^n: \|\mathbf{x}\| = 1\}.$$

若$\mathbf{u}$落在球面上面积相等的区域的概率相同，则称$\mathbf{u}$服从球面均匀分布，记作$\mathbf{u} \sim U(S^{n-1})$ 。

注：$U(S^{n-1})$关于R^n勒贝格测度μ^n没有密度, 关于μ^{n-1}的密度为

$$p(\mathbf{u}) = \frac{1}{|S^{n-1}|}, \mathbf{u} \in S^{n-1} \text{ w.r.t } \mu^{n-1}$$

定义2. 若$\mathbf{x} \sim U(B^n)$， 则 $\mathbf{u} = \frac{\mathbf{x}}{\|\mathbf{x}\|} \sim U(S^{n-1})$ （第2讲）。



$U(S^0)$

R^1中单位球 $B^1 = [-1,1]$, 球面: $S^0 = \{-1,+1\}$,

$$u \sim U(S^0) \Leftrightarrow P(u = \pm 1) = 1/2$$

如何产生$U(S^0)$随机数？

除了用伯努利随机数直接产生，也可以通过均匀分布随机数产生：

产生$x \sim U(-1,1)$, $u = \frac{x}{|x|} = sgn(x) \sim U(S^0)$

$U(S^1)$

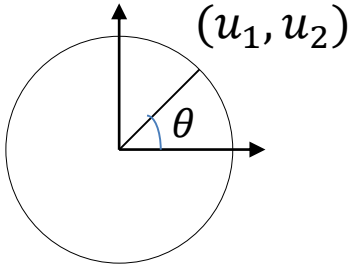
假设$\mathbf{u} = (u_1, u_2)^{\mathrm{T}}$在圆周$S^1 = \{(x_1, x_2): x_1^2 + x_2^2 = 1\}$上均匀分布, 记作 $\mathbf{u} \sim U(S^1)$, 即对任何圆弧L

$$P(\mathbf{u} \in L) = |L|/2\pi, \quad L \subset S^1$$

注意$\mathbf{u}$在R^2没有概率密度。u_1的分布？

引入极坐标：$u_1 = \cos(\theta), u_2 = \sin(\theta)$,因为圆弧长度与对应的圆心角成正比，故角度$\theta \sim U(0,2\pi)$, 所以

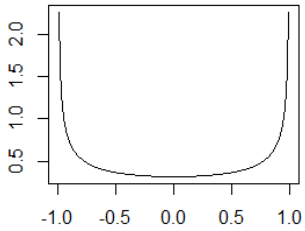
$$P(u_1 \le t) = 1 - 2P(\theta \le \arccos(t)) = 1 - 2\arccos(t)$$



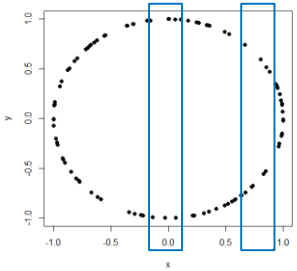
求导得u_1的概率密度

$$f(t) = \frac{1}{\pi\sqrt{1-t^2}}, \ |t| \le 1$$

该分布称为区间$[-1,1]$上的arcsin分布 (arcsin law，右图).



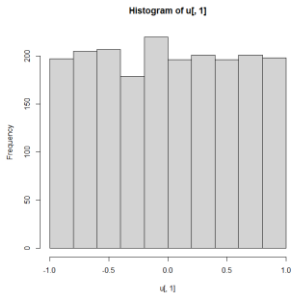
蒙特卡罗：如右图所示，在圆周上产生若干均匀随机数，在边界± 1附近方框截取的圆弧较长（点数最多），即u_1在± 1附近概率较大.

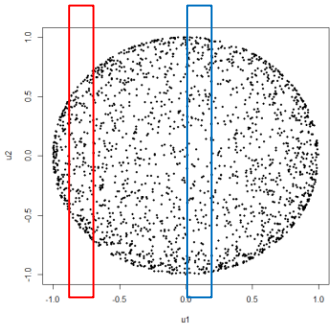


$U(S^2)$

三维空间中，单位球面$S^2 = \{(x_1, x_2, x_3): x_1^2 + x_2^2 + x_3^2 = 1\}$,
假设 $\mathbf{u} = (u_1, u_2, u_3)^\top \sim U(S^2)$。

$U(S^2)$的一元边际





两个等宽的框
内点数一样多

上图显示u_1服从均匀分布 (参见定理2):

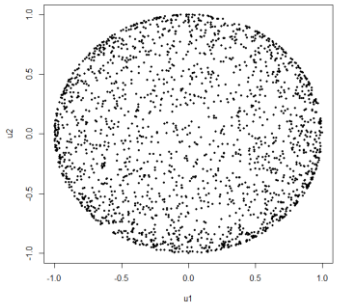
$$(u_1, u_2, u_3)^\top \sim U(S^2) \quad \Rightarrow u_1 \sim U(-1,1)$$

这表明等高的球台(蓝红框，spherical segment)的表面积应该相同。换言之，将球体苹果切成若干等厚度的切片(球台)，则每个切片的苹果表皮面积相同, 这实际上是2200多年前阿基米德发现的结果：

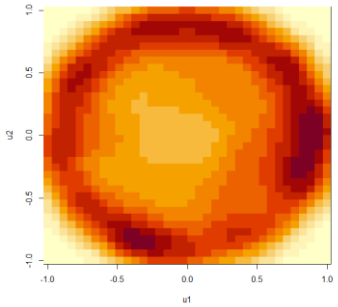
Archimedes' Hat-Box Theorem：
若球台高度为h, 球半径为R，则球台表面积为$2\pi Rh$.

$U(S^2)$的二元边际

(u_1, u_2)散点图



(u_1, u_2) 分布热图

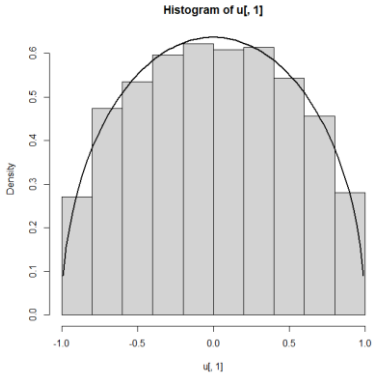


二维投影(u_1, u_2) 多数点拥挤在边界上，这符合我们对球面的直观理解：从u_3方向观察球面，中间部分比较平坦，边缘部分陡峭（曲率大），因而边缘处的点较多，概率较大。

$U(S^3)$

R^4中单位球面上的均匀分布, $\mathbf{u} = (u_1, u_2, u_3, u_4)^\top \sim U(S^3),\ \|\mathbf{u}\| = 1$

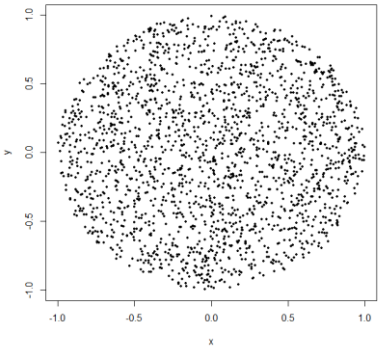
一元边际



Wigner 半圆律

$$u_1 \sim f(t) = \frac{2\sqrt{1-t^2}}{\pi},\ |t| \le 1$$

二元边际



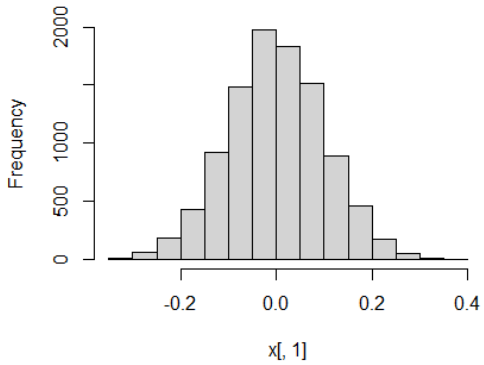
$(u_1, u_2) \sim U(B^2)$

4维空间的球表面从人类(二维)视角看起来处处平坦？

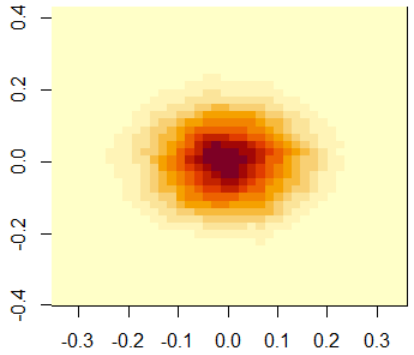
$U(S^{99})$

5维以上欧氏空间单位球面上的均匀分布的一元、二元边际聚集于0附近，近似正态分布。$n = 100$情形：

一元边际分布



二元边际分布



前面所观察到的现象是下述定理2的特殊情况，我们将在第3讲证明（参见参考书 Bilodeau and Brenner (1999) P53-54）。

$U(S^{n-1})$的边际分布

定理2. 假设 $\mathbf{u} = (u_1, \ldots, u_n)^{\mathsf{T}} = (\mathbf{u}_1^{\mathsf{T}}, \mathbf{u}_2^{\mathsf{T}})^{\mathsf{T}} \sim U(S^{n-1})$，则对任何$1 \le k \le n-1$，$\mathbf{u}_1 = (u_1, \ldots, u_k)^{\mathsf{T}}$的边际概率密度为

$$f_{n,k}(\mathbf{u}_1) = \frac{\Gamma(n/2)}{\pi^{k/2}\Gamma((n-k)/2)} (1 - \|\mathbf{u}_1\|^2)^{(n-k-2)/2}, \|\mathbf{u}_1\| \le 1.$$

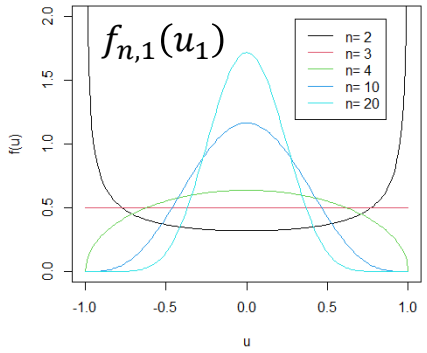
前面实验看到的各种现象都蕴含在定理2中

⇩

一元边际

推论1. $U(S^{n-1})$ 的一元边际分布称为Wigner球对称分布：

$$f_{n,1}(u_1) = \frac{\Gamma(n/2)}{\pi^{1/2}\Gamma((n-1)/2)} (1 - u_1^2)^{(n-3)/2}, |u_1| < 1.$$

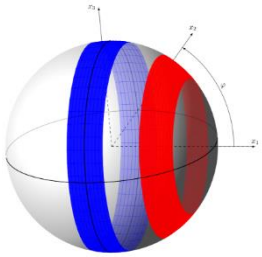


高维球面的面积大部分集中于赤道带

从推论1(一元边际密度函数)可以知道: 对任何$n > 1$, $E(u_1) = 0, \mathrm{var}(u_1) = 1/n$。 由切比雪夫不等式

$$P(|u_1| < \varepsilon) > 1 - \frac{1}{n\varepsilon^2}$$

这表明高维球面面积大部分在赤道球带(蓝色)附近



例如，当$n = 100$时，u_1以38%的概率集中于$[-0.05, 0.05]$区间，即宽为0.1的赤道带的面积占单位球面面积的38%。

推广的阿基米德定理

推论2. R^n中$U(S^{n-1})$的$k = n - 2$元边际分布为$U(B^{n-2})$，即R^{n-2}中单位球内的均匀分布

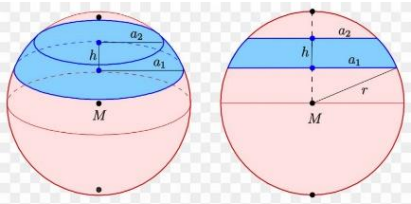
$$f(\mathbf{u_1}) = \frac{\Gamma(n/2)}{\pi^{(n-2)/2}}, \quad \mathbf{u}_1 = \mathbf{u}_{[1:(n-2)]} \in B^{n-2}$$

推论2证实了我们在蒙特卡洛观察到的如下结果：

- $U(S^2)$的一维边际在$[-1,1]$区间均匀；
- $U(S^3)$的二维边际在单位圆内均匀。

推论2证明了如下高维球的几何性质：

用低2维的平面截取球面,得到的高度固定的球台表面积与截取位置无关，故我们称之为推广的阿基米德定理。



球台是球带和截得它的两个平行截面所围成的几何体

推论2表明了球面S^{n-1}与球体B^{n-2}的关系，表明研究球面和球体性质时可以互相转换。

如果地球受到来自于n维空间的降维打击，$n=?$

$n=5$：$\mathbf{u}\sim U(S^4)\subset R^5 \Rightarrow \mathbf{u}_{[1:3]}\sim U(B^3)$
5维球面均匀分布的3维边际在3维球内均匀分布。

总之，概率方法提供了理解高维空间（球体）几何性质的简便工具