第一讲 多元分析

2024.2.26

主讲：杨亚宁（ynyang@ustc.edu.cn)
助教：丁浩伦（dhldhldhl@mail.ustc.edu.cn）
吴佳铖 （wujiach@mail.ustc.edu.cn）

QQ课程群：364233462

内容

1. 课程简介
2. 数据可视化
3. 超立方体
4. 超球体

1. 课程简介

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

先修 微积分、线性代数、概率论、数理统计

向量数据 多元分析主要考虑多元（向量、多变量）数据：

$\mathbf{x}_1, \ldots, \mathbf{x}_n \in R^p$，独立，

习惯: 向量指的是列向量，以小写黑正体字母表示

数据矩阵 样本 $\mathbf{x}_1, \ldots, \mathbf{x}_n$ 按行排列组成 $n \times p$ 数据矩阵：

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

多元分析

多元分析研究向量数据的统计分析方法，主要包括但不限于

- 经典的统计推断： 参数估计、多元方差分析、多元线性回归。
- 降维/可视化： 主成分分析、因子分析、多维标度法。
- 机器学习/统计学习：无监督学习，有监督学习。

多元统计与一元统计

数理统计、线性回归模型基本属于一元统计，注意线性模型

$$y_{1\times 1} = \boldsymbol{\beta}^{\top}\mathbf{x} + \varepsilon$$

中主要的、感兴趣的响应是一元的（虽然自变量 $\mathbf{x}$ 是多元的）。

但如果同一个研究对象的响应指标有多个，则此时的线性回归模型称为多元线性回归模型。

一个例子

例1. 6件出土陶器，每个陶器测量9种化学成分含量，陶器i的成分是一个9×1向量$\mathbf{x}_i$，这些向量按行排列构成数据矩阵$X_{6\times 9}$如下

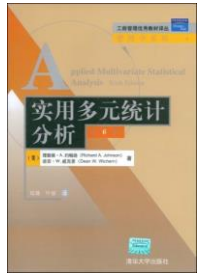
$$X_{6\times 9} = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_9^\top \end{pmatrix} =$$

ID	Al2O3	Fe2O3	MgO	CaO	Na2O	K2O	TiO2	MnO	BaO
1	18.8	9.52	2.00	0.79	0.40	3.20	1.01	0.077	0.015
2	16.9	7.33	1.65	0.84	0.40	3.05	0.99	0.067	0.018
3	13.8	7.06	5.34	0.20	0.20	4.31	0.71	0.101	0.021
4	14.6	7.09	3.88	0.13	0.20	4.36	0.81	0.124	0.019
5	15.8	2.39	0.63	0.01	0.04	1.94	1.29	0.001	0.014
6	18	1.50	0.67	0.01	0.06	2.11	0.92	0.001	0.016

- 统计推断：已知陶器1-3，4-6分别属于类型A和B，它们是否有差异？
- 分类/判别分析：如何从成分决定新挖掘陶器属于A还是B？
- 聚类分析：如果没有类别信息，从成分能否推断出有两类陶器？
- 降维/可视化：如何将变量个数(9)减少到2而不丢失过多信息？

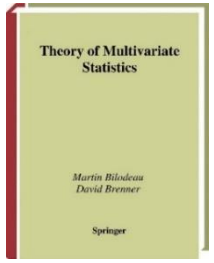
教材

1. R.A.Johnson and D.W.Wichern (2008) 实用多元统计分析, 第6版，英文版/中文版（应用，美）



全面(700+页)、理论较少、大量实际数据例子。我们将采用其结构、部分内容和实际例子。

2. M.Bilodeau, D.Brenner (1999) *Theory of Multivariate Statistics*. Springer.（理论，加）



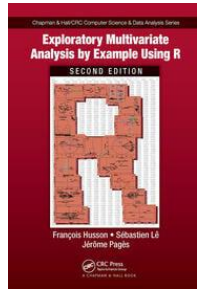
简明(287页)、理论、无实际数据例子。我们将参考其中的分布理论。

课程前几讲将主要参考Bilodeau &Brenner第2-9章部分内容，

之后内容将主要参考Johnson & Wichern的第8-12章

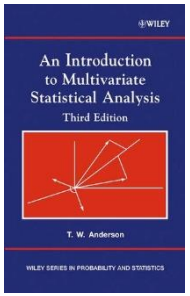
参考书

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



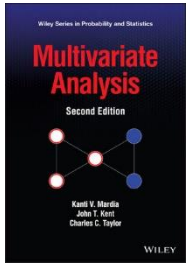
应用(250页)，仅含主成分分析，对应分析。法国学派。我们只采用其中一或两个数据例子。

4. T.W.Anderson (2003) *An Introduction to Multivariate Statistical Analysis*, Wiley, 3rd ed（理论，美）



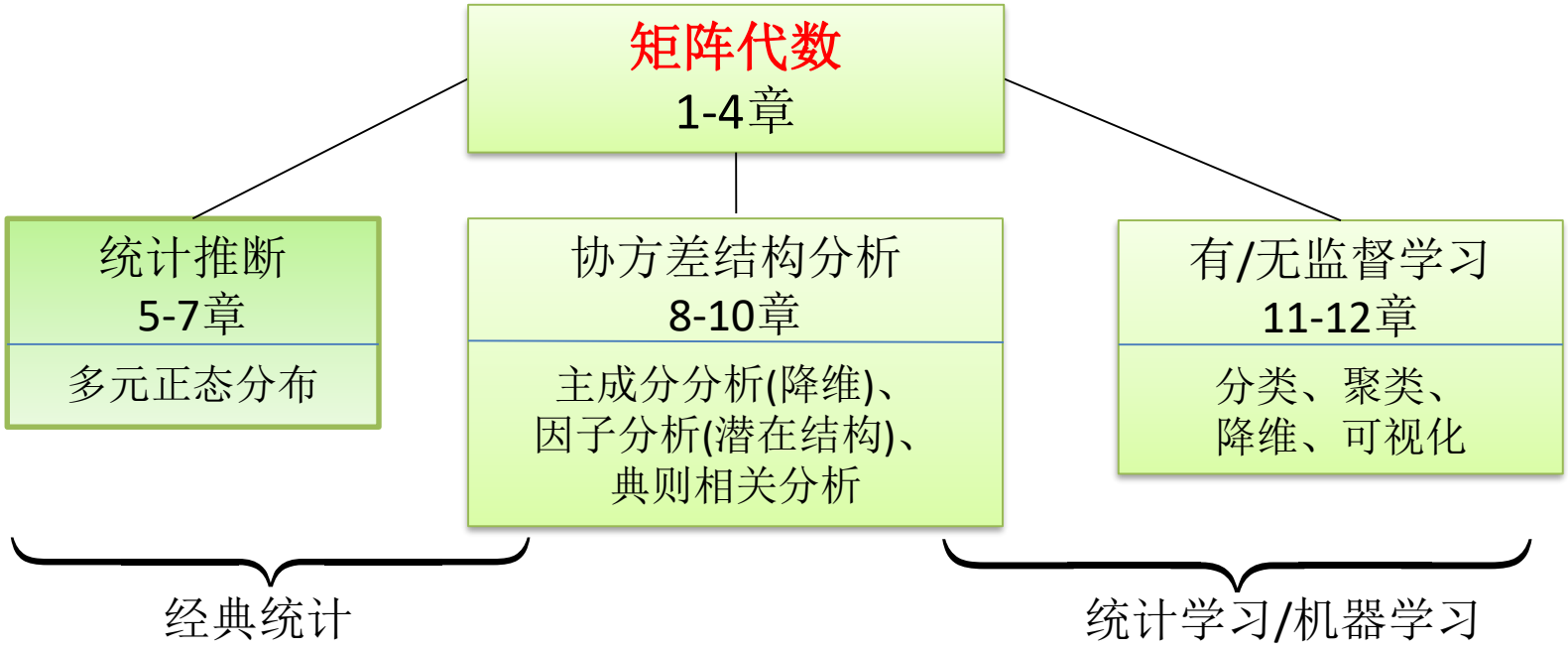
经典全面(700+页)、理论、无实际例子，供查阅(不建议阅读)。许宝騄的学生。

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



经典(400页)、理论，今年四月将出第二版，敬请期待。

Johnson&Wichern教材目录



- 多元正态分布及其它分布，基于向量样本的统计推断。

- 主成分分析: 随机向量的少数分量或者组合能否近似地代表原始的随机向量？
- 因子分析：是否可用少数几个潜变量描述可见的随机向量内部的相关性？
- 典则相关分析：如何度量随机向量之间的相关性？

- 分类判别：预测类别。
- 聚类分析：在类别未知的数据中发现分类聚簇特征。
- 降维、可视化：高维数据的低维表示和展示。

2. 数据可视化

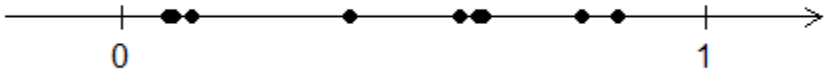
一元数据

散点图(plot), 5-number summary (summary),直方图 (hist), 枝叶图 (stem), 盒型图 (boxplot)…

例2. 随机产生10个$U(0,1)$随机数:

0.389, 0.583, 0.095, 0.853, 0.787, 0.119, 0.606, 0.081, 0.391, 0.619

散点图



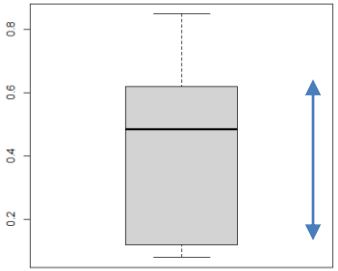
数轴上的散点图显示，均匀分布的数据点个数较少时，它们并不像我们相像的那样均匀地分布在[0,1]区间：容易出现聚簇，不等间隔。

数据汇总: Boxplot，5-number summary

```
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0800  0.1875  0.4850  0.4520  0.6175  0.8500
```

25%分位数

> boxplot(x) :



75%分位数

IQR

25%分位数

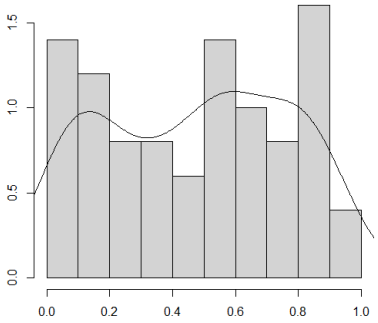
Interquantile range (度量分散程度):

IQR=3rd quantile−1st quantile

对于正态分布$IQR = 1.35\sigma$ (标准差)

分布

> 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

- 对每个边际应用一元数据的图示方法
- 二元散点图: plot,
- 二元分布： image, persp, contour

例3. 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：Strength_MD, Strength_CD, Density（密度）。 这里我们以两个强度变量为例，简单回顾各种基本的图表示方法。

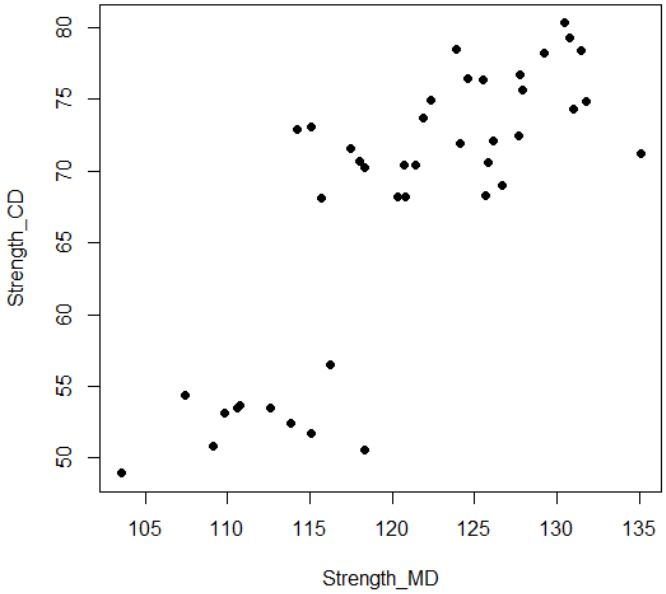
Table 1.2 Paper-Quality Measurements

Specimen	Density	Strength	
		Machine direction	Cross direction
1	.801	121.41	70.42
2	.824	127.70	72.47
3	.841	129.20	78.20
4	.816	131.80	74.89
5	.840	135.10	71.21
6	.842	131.50	78.39
7	.820	126.70	69.02
8	.802	115.10	73.10
9	.828	130.80	79.28
10	.819	124.60	76.48
11	.826	118.31	70.25
12	.802	114.20	72.88
13	.810	120.30	68.23
14	.802	115.70	68.12
15	.832	117.51	71.62
16	.796	109.81	53.10
17	.759	109.10	50.85
18	.770	115.10	51.68
19	.759	118.31	50.60
20	.772	112.60	53.51
21	.806	116.20	56.53
22	.803	118.00	70.70.
23	.845	131.00	74.35
24	.822	125.70	68.29
25	.971	126.10	72.10
26	.816	125.80	70.64
27	.836	125.50	76.33
28	.815	127.80	76.75
29	.822	130.50	80.33
30	.822	127.90	75.68
31	.843	123.90	78.54
32	.824	124.10	71.91
33	.788	120.80	68.22
34	.782	107.40	54.42
35	.795	120.70	70.41
36	.805	121.91	73.68
37	.836	122.31	74.93
38	.788	110.60	53.52
39	.772	103.51	48.93
40	.776	110.71	53.67
41	.758	113.80	52.42

Source: Data courtesy of SONOCO Products Company.

散点图

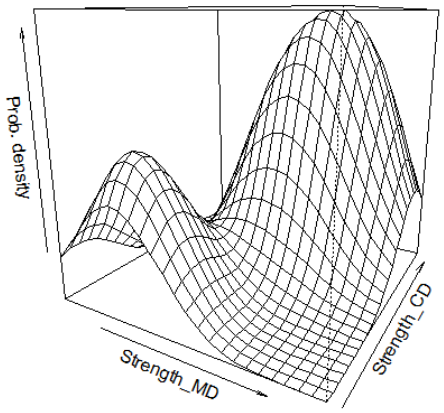
散点图是最基本，也是最重要的数据展示方法。



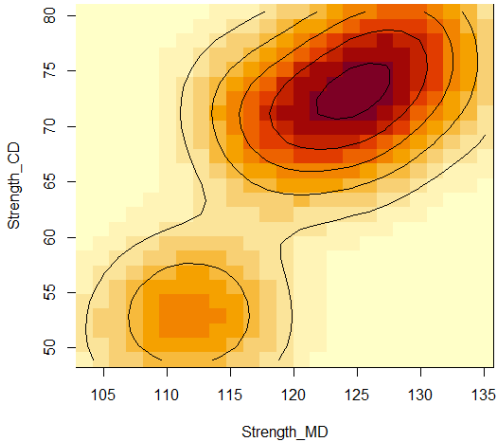
两个strength正相关
数据聚簇为两类

分布: 密度函数和密度等高线图

二元分布、密度图刻画数据在平面上的分布情况，与散点图同样重要和基本。



persp()



image(), contour()

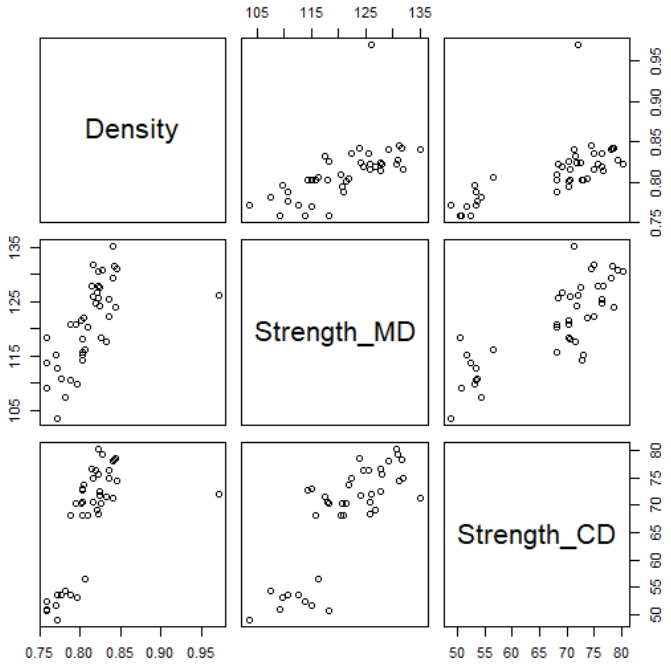
```
## kde2d估计概率密度
library(MASS)
k <- kde2d(paper[,2],paper[,3], n=25) #n: x,y轴划分区间的个数

#二维变量的密度函数（左）和概率密度的热图、等高线图：
image(k, xlab="Strength_MD", ylab="Strength_CD" )
contour(k, add = TRUE, drawlabels = FALSE,nlevels=6)
persp(k, xlab="x", ylab="y",zlab="Prob. density",theta=30 )
```

三维数据

- 对边际应用一、二元数据的图示方法。比如两两变量之间的散点图、泡泡图、条件散点图等。
- 三维散点图(3dplot)。
- 三个变量的联合分布：难以图示。

例3(续）. 两两之间的散点图表明，三个变量之间正相关

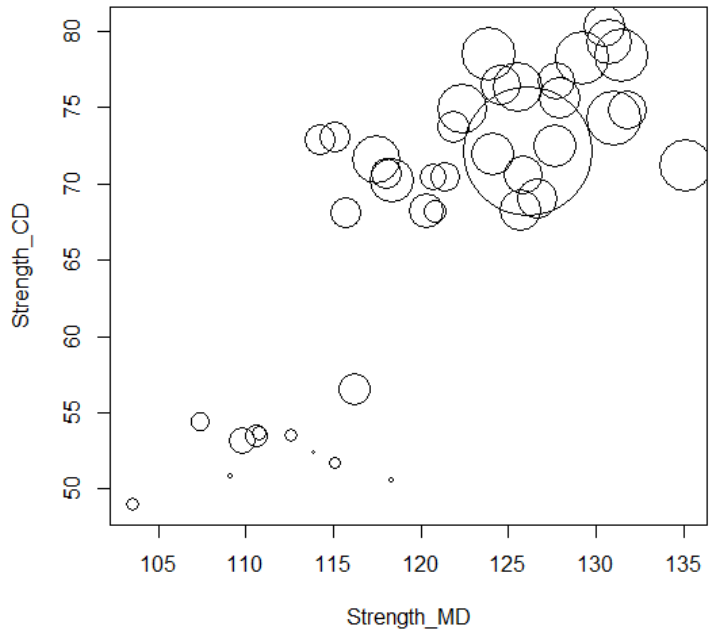
散点图



>plot (paper)

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。

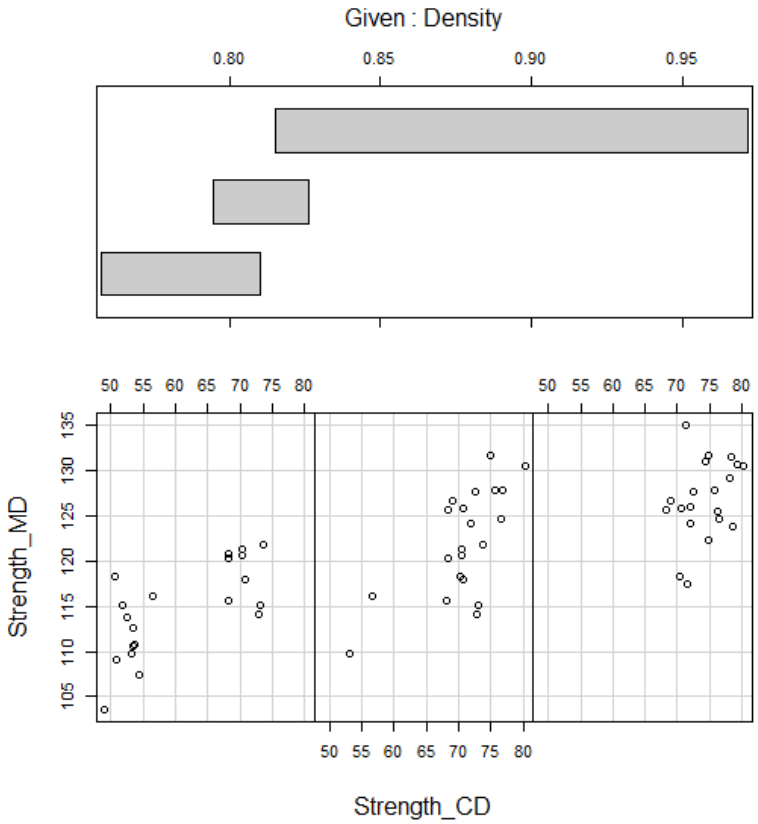


每个点的大小代表强度Density.

```
plot(paper[,2:3],type="n")
symbols(paper[,2:3], circles = paper[,1]-0.75,add=T,inches=0.5)
```

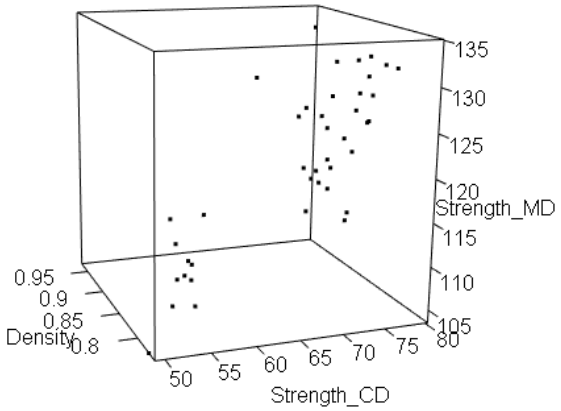
条件散点图

给定第三个变量z的条件下(x,y)的散点图，条件分布

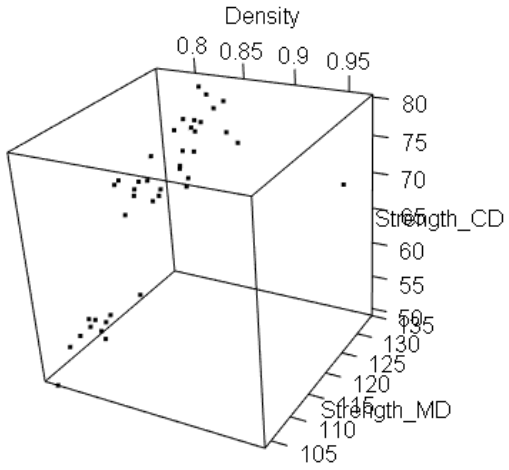



```
#R函数：coplot
coplot(Strength_MD~Strength_CD|Density,
data=table1.2,number=3,columns=3)
```

三维散点图




```
>library(rgl)
>plot3d(paper) #手工拖动旋转
```




```
##自动播放动画：
> plot3d(paper)
> M = par3d("userMatrix")
> f=par3dinterp(time = (0:2)*12,  zoom=c(1,1.1,0.9 ),
userMatrix = list(M, rotate3d(M, pi ,0, -1, 1),
rotate3d(M, pi , -1,0, -1)))
> dur=10  #播放时间
> play3d(f , duration =dur)
 #将动画存成moive.gif (dir指定存放位置)
> movie3d(f, duration = dur, dir=getwd(), clean=T)
```

高维数据

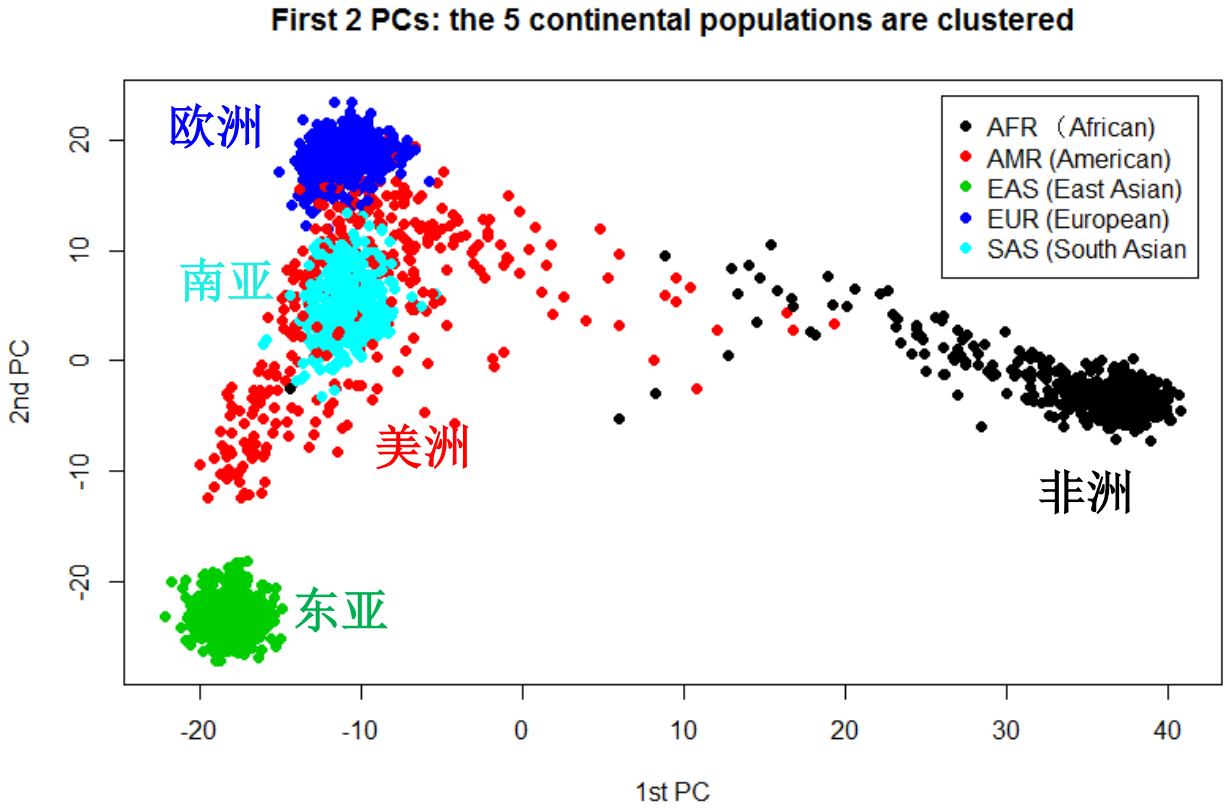
三元以上的数据只能在低维（2或3）展示。多元分析中的主成分分析方法寻找最能有效替代原始变量的线性组合，这种组合而成的“新”变量称为主成分，两个主成分（即线性组合）通常认为能够反映原始数据（多变量）的某些特征。

例4.（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，数据矩阵是2504x9932矩阵。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)。数据有什么特点？基因能否区分种族？

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
AMR	2	0	0	0	2	0	2	0	0	0	0	1	1
AFR	2	1	0	1	2	0	2	1	0	1	0	2	2
SAS	2	0	0	0	2	0	1	0	1	1	0	0	2
EAS	2	0	0	1	2	1	2	0	1	2	0	1	2
EAS	1	0	0	0	2	0	1	0	0	0	0	2	2
EUR	2	0	0	0	2	0	2	1	1	1	0	2	2
EUR	2	0	0	1	2	2	2	0	0	1	0	2	2
AMR	2	0	0	0	2	0	2	1	0	1	0	2	2
SAS	2	0	0	1	2	0	2	0	0	1	0	2	2


北美洲
夏威夷
欧洲
亚洲
SAS
非洲
南美洲
大洋洲
南极洲

将原始数据从9932个变量降到两个主成分，在二维平面上几个种族分别聚簇成类，其中欧洲、东亚、非洲人能较好地分离，南亚和美洲人介于三者之间。

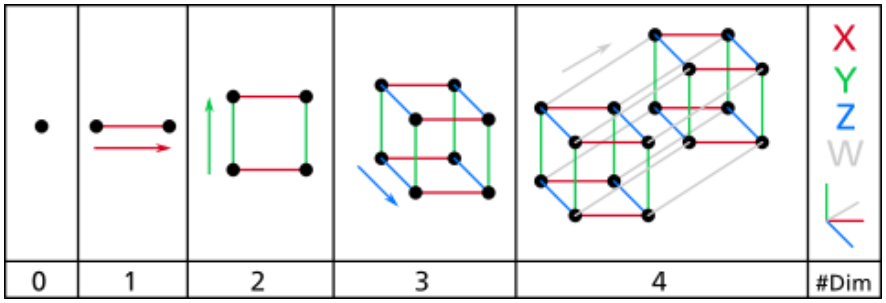


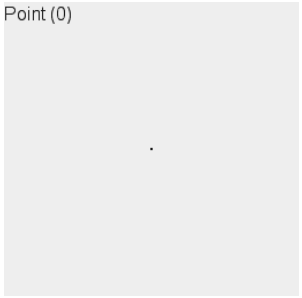
下面以n维欧氏空间中的超立方体和超球体为例，试图理解4维或4维以上欧氏空间。

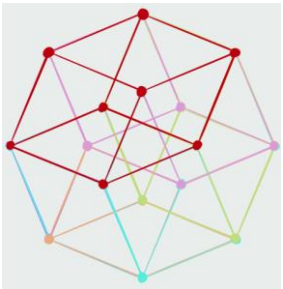
3. 超立方体

三维以上的几何形状是人类无法想象的，最简单的或许是超立方体（hypercube）。维基百科介绍了从低维到高维推广的超立方体的理解方式：$n+1$维立方体可看作是由n维立方体在新增加的维度上平移得到的。

以4维超立方体（tesseract）为例，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”。



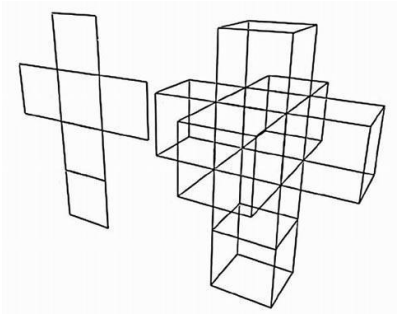


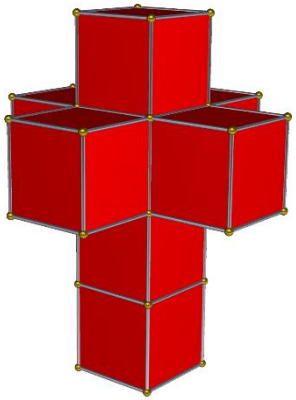


https://en.wikipedia.org/wiki/Tesseract

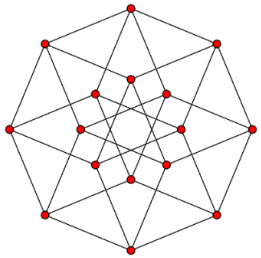
1909年《科学美国人》杂志征集对四维空间的通俗合理的解释。大多数参赛论文都提到了数学家查尔斯.霍华德.辛顿(Charles Howard Hinton， Geoffrey Hinton的曾祖父)的超立方体可视化方法。辛顿可视化方法包括解析法、投影法和切面法。

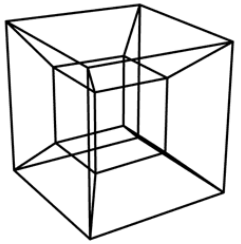
下图左是3维立方体盒子的6个表面的展开，右图类似地将四维立方体的8个3维立方体表面的展开，称为辛顿超立方体，后来甚至成了神秘主义符号。



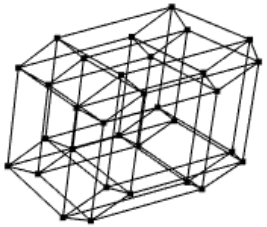


下面是四维立方体（超立方体）的几种常见的演示方式：





5维立方体的三维投影：



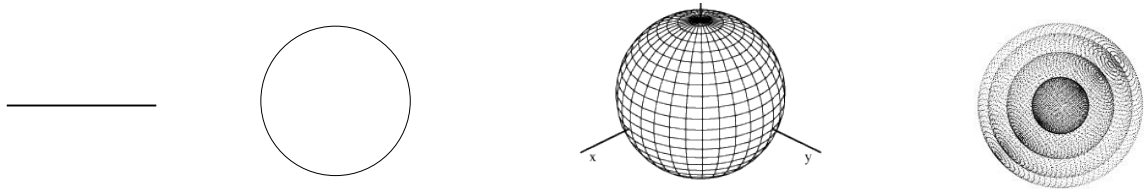
4. 超球体和超球面(hypersphere)

R^n中原点为中心、半径为R的超球体(hyperball)和超球面(hypersphere):

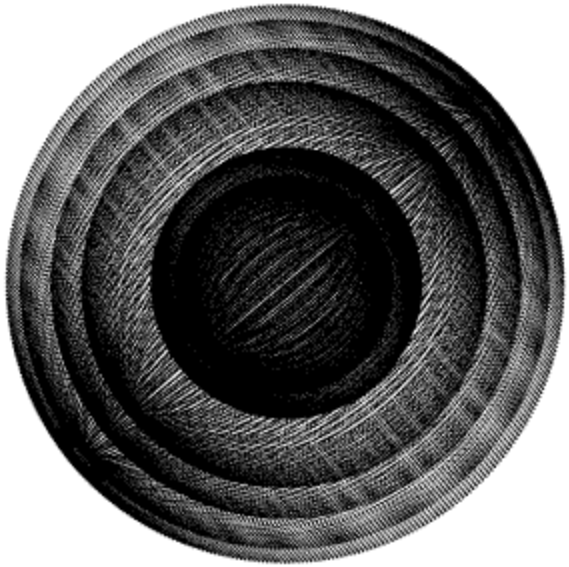
$$B^n(R) = \{\mathbf{x} \in R^n: \|\mathbf{x}\| \leq R\}, \quad S^{n-1}(R) = \{\mathbf{x} \in R^n: \|\mathbf{x}\| = R\}.$$

分别称为 n-ball和$(n-1)$-sphere.

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



4维以上超球体非常难以理解和可视化。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， 但这不足以提供一个正确的4维球的理解方式。我们只能通过数学方式得到4维球的部分性质。



球坐标变换

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

$$\begin{cases} x_1 = r\cos(\theta_1) \\ x_2 = r\sin(\theta_1)\cos(\theta_2) \\ \vdots \\ x_{n-1} = r\sin(\theta_1)\cdots\sin(\theta_{n-2})\cos(\theta_{n-1}) \\ x_n = r\sin(\theta_1)\cdots\sin(\theta_{n-2})\sin(\theta_{n-1}) \\ r = \|\mathbf{x}\| \geq 0,\ 0 \leq \theta_1, ..., \theta_{n-2} \leq \pi,\ 0 \leq \theta_{n-1} \leq 2\pi, \end{cases}$$

引理1: 球坐标变换的Jacobi:

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

球坐标表示下的体积元

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

半径为R的球面$S^{n-1}(R)$上的面积元

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

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

利用引理1求解超球体积和表面积如下：

$B^n(R)$的体积:

$$|B^n(R)| = \int_{\|\mathbf{x}\|\le R} d\mathbf{x} = \int_{C_n} d^nV$$

$$= \int_0^{2\pi}\cdots\int_0^{\pi}\int_0^{\pi}\int_0^{R} r^{n-1}\sin^{n-2}(\theta_1)\sin^{n-3}(\theta_2)\cdots\sin(\theta_{n-2})\ drd\theta_1\cdots d\theta_{n-1}$$

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

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

$S^{n-1}(R)$的面积:

$$|S^{n-1}(R)| = \int_{\partial C_n} d_{S^{n-1}}A$$

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

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

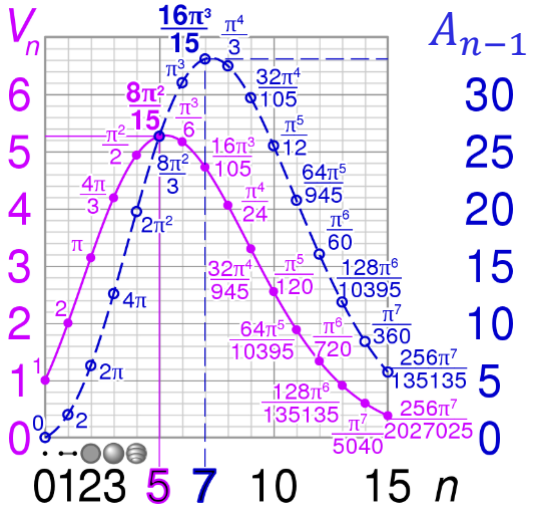
$\partial C_n = \{(\theta_1,\cdots,\theta_{n-1}):$
$0\le\theta_1,\ldots,\theta_{n-2}<\pi, 0\le\theta_{n-1}<2\pi\}$

体积和面积公式

n维球体的体积和表面积：

$$V_n = \frac{\pi^{n/2}}{\Gamma(n/2+1)} R^n \quad , \quad A_{n-1} = \frac{n\pi^{n/2}}{\Gamma(n/2+1)} R^{n-1} = \frac{nV_n}{R}$$

单位球的体积和表面积分别在$n = 5$、~~8~~ 7时最大，之后单调下降趋于0.



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

更多超球性质可以通过微积分计算（较复杂），更简单的方式是引入球内均匀分布和球面均匀分布，通过概率了解几何性质。

黎曼积分~均匀分布积分

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

特别地，若$g(\mathbf{x}) = 1_{(\mathbf{x}\in A\subset B^n)}$

$$\int_{A\subset B^n} d\mathbf{x} = |B^n|P(\mathbf{x}\in A), \mathbf{x}\sim U(B^n)$$

进一步有importance sampling.

$$\int \quad g(\mathbf{x})d\mathbf{x} = \int \quad \left[\frac{g(\mathbf{x})}{f(\mathbf{x})}\right] f(\mathbf{x})d\mathbf{x} = \boldsymbol{E}\left[\frac{g(\mathbf{x})}{f(\mathbf{x})}\right],\ \mathbf{x}\sim f,$$

f是一个容易的密度，比如上面的球均匀分布。

球均匀分布

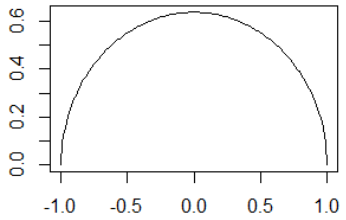
圆盘均匀分布

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

假设 $(x_1, x_2) \sim U(B^2)$, 有概率密度: $p(x_1, x_2) = \frac{1}{\pi}$, $x_1^2 + x_2^2 \leq 1$,

则x_1服从Wigner半圆律分布（ semicircle law ）

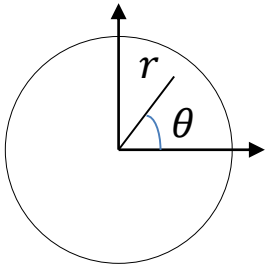
$$p(x_1) = \frac{2}{\pi}\sqrt{1 - x_1^2}, \quad |x_1| \leq 1$$



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

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

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。

球均匀分布

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

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

在标准坐标系下直接计算$\mathbf{x}\sim U(B^n)$的性质比较困难，通常转化到球坐标系

命题1. 假设$\mathbf{x}\sim U(B^n)$，考虑球坐标变换

$$\mathbf{x}=(x_1,\ldots,x_n)^\top\in R^n\to(r,\boldsymbol{\theta}),$$

则$(r,\theta_1,\ldots,\theta_{n-2},\theta_{n-1})$的联合概率密度

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

$0<r<1, 0\le\theta_1,\ldots,\theta_{n-2}\le\pi, 0\le\theta_{n-1}<2\pi$。特别地

a) 所有参数$r,\theta_1,\ldots,\theta_{n-1}$相互独立。

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

$\Leftrightarrow r\overset{d}{=}\max(u_1,\ldots,u_n), u_1,\ldots,u_n$ iid $\sim U(0,1)$。

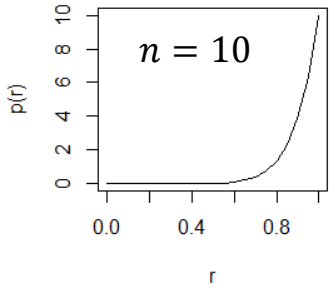
a) $\theta_k\sim p_k(\theta_k)\propto\sin^{n-k-1}(\theta_k)$，特别地，$\theta_{n-1}\sim U(0,2\pi)$.

关于模长r

$n > 1$ 时, r的概率质量集中于1附近

$$P(1-\epsilon < r < 1) = \int_{1-\epsilon}^{1} nr^{n-1}dr = 1-(1-\epsilon)^n$$

另外， $E(r) = n/(n+1)$.



这说明单位球内均匀取的点较大可能在球表面附近, 即大部分体积集中在球表面附近。事实上，球表面厚度为ϵ的球壳体积

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

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

注1：通过球内一个随机点（均匀分布）或者随机抽取多个点（蒙特卡洛），我们就可研究球有关的几何性质。比如算出$P(\mathbf{x} \in A)$，就能得到A的体积：

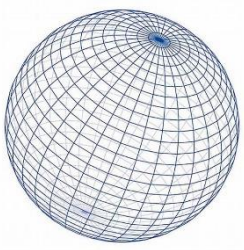
$$|A| = P(\mathbf{x} \in A)\left|B^n\right|$$

概率几何?

注2：改变 r 的概率分布可以产生其它分布（比如正态分布），用于描述球体物理性质（而不是单纯几何性质），比如地球的质量密度随半径 r 增加而增加, 既不是线性的，也不是 r^{n-1}.

可能是 r^{α}?
$1 < \alpha < n-1$

球面均匀分布



我们以球面均匀分布研究超球面，实际上也是研究超球体。球面均匀分布没有概率密度函数，数学计算更为复杂，下面主要使用蒙特卡洛模拟进行考察。所有模拟观察到的现象都归结于理论结果定理1(p41)，它提供了球面均匀分布与低维球分布之间的关系。

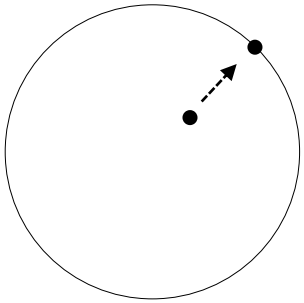
球面均匀分布

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$\mathbf{u} \sim U(S^{n-1})$。$U(S^{n-1})$关于勒贝格测度没有概率密度。

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 $\mathrm{B}^1 = [-1,1]$, 球面: $S^0 = \{-1,+1\}$, $u \sim U(S^0)$, 概率分布:

$$P(u = \pm 1) = 1/2$$

如何产生$U(S^0)$随机数?

除了用伯努利随机数直接产生，也可以通过正态随机数产生:

产生$x \sim N_1(0,1), u = x/|x| = sgn(x)$

圆周均匀分布 $U(S^1)$

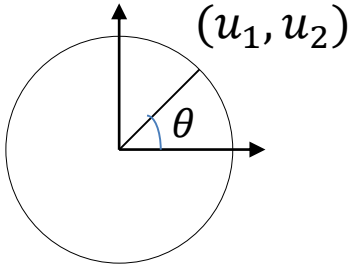
假设$\mathbf{u} = (u_1, u_2)^{\mathrm{T}}$在圆周$S^1 = \{(x_1, x_2): x_1^2 + x_2^2 = 1\}$上均匀分布, 记作 $\mathbf{u} \sim U(S^1)$, 即对任何圆弧L

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

注意$\mathbf{u}$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

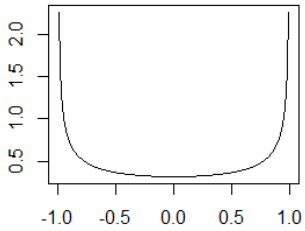
$$P(u_1 \le t) = 1 - 2P(\theta \le \arccos(t)) = 1 - 2\arccos(t)$$



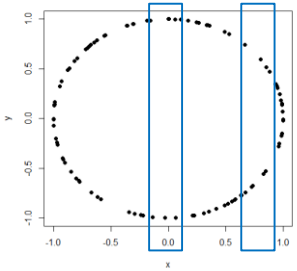
求导得u_1的概率密度

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

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



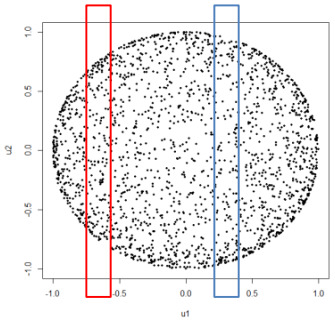
直观：如右图所示，在圆周上产生若干均匀随机数，在边界± 1附近方框截取的圆弧较长（点数最多），即u_1在± 1附近概率较大.

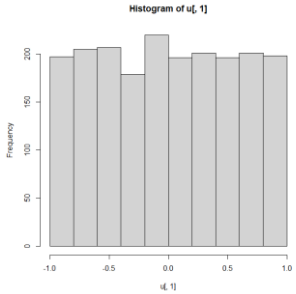


$U(S^2)$

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 $\mathbf{u} = (u_1, u_2, u_3)^\top \sim U(S^2)$

一元
边际





蒙特卡洛模拟显示u_1均匀分布 (参见定理1):

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

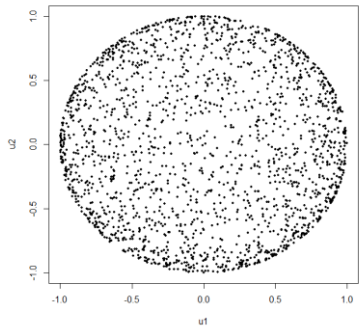
这表明等高的球台(蓝红框，spherical segment)的表面积应该相同。

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：

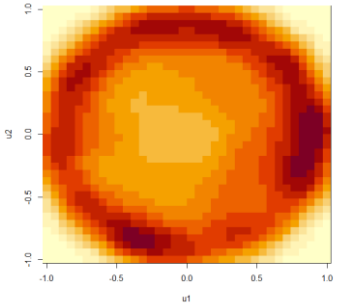
Archimedes' Hat-Box Theorem：
若球台高度为h, 球半径为R，则球台表面积为$2\pi Rh$.

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），因而边缘处的点投影后拥挤在一起，概率较大。

$(u_1, u_2, u_3)^{\mathrm{T}} \sim U(S^2) \Rightarrow$

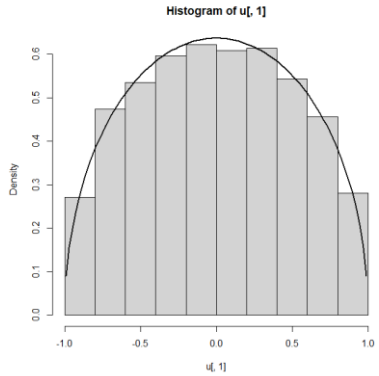
$$f(u_1,\ u_2) = \frac{1}{2\pi^{3/2}} \frac{1}{\sqrt{1-u_1^2-u_2^2}},\ u_1^2 + u_2^2 < 1$$

（参见后面定理1）

$U(S^3)$

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

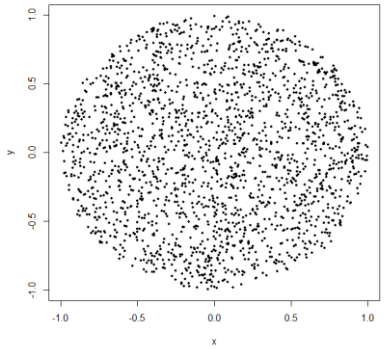


Wigner semicircle law (物理):

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

参见定理1

二元边际



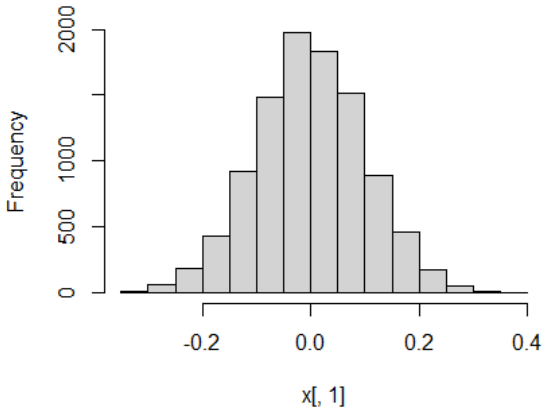
$\mathbf{u} \sim U(S^3) \Rightarrow (u_1, u_2) \sim U(B^2)$

4维空间的球表面从人类(二维)视角看起来处处平坦?

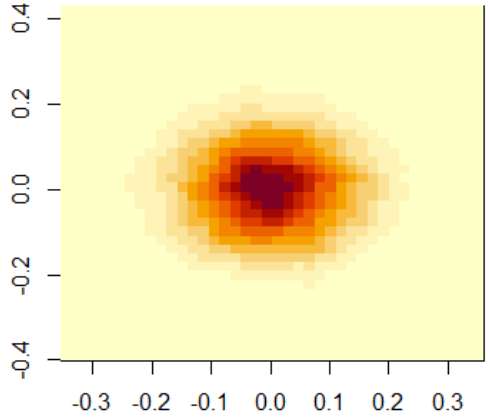
$U(S^{99})$

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



前述观察到的现象是下述定理1的特殊情况。

$U(S^{n-1})$的边际分布

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

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

该定理说明$U(S^{n-1})$的k维边际分布是k维单位球内的球对称分布（正交变换下不变，密度函数仅依赖于模长 $\|\mathbf{u}_1\|$ ）

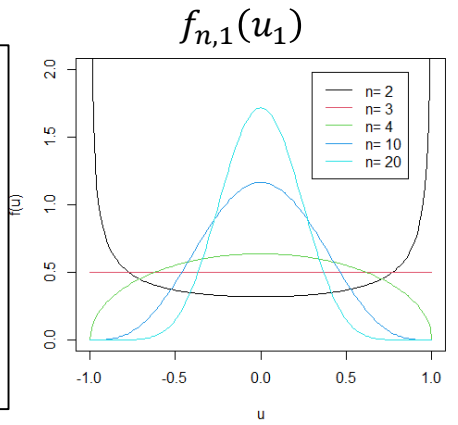
我们将在第2-3讲证明该结论（参见参考书 Bilodeau and Brenner (1999) P53-54）。 这里我们略过技术细节，只关注蒙特卡洛观察到的现象与理论结果的一致性，以及理论结果蕴含的涵义。

高维球面的面积大部分集中于赤道带

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

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

换言之，$\frac{1}{2}(u_1+1) \sim Beta\left(\frac{n-1}{2}, \frac{n-1}{2}\right)$,

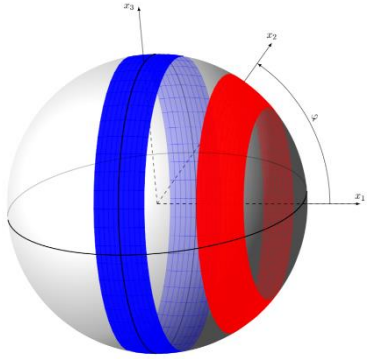


- 对任何$n > 1$, $E(u_1) = 0, \mathrm{var}(u_1) = 1/n.$
- $n > 3$时，u_1集中于0附近, 由切比雪夫不等式

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

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%。



- 从$f_{n,1}$公式可以看出：

$n = 2$时，$u_1 \sim$ arcsin law； $n = 3$时，$u_1 \sim U[-1,1]$。

当n 很大，$f_{n,1}(u_1) \approx \phi(\sqrt{n}u_1)$，$u_1$近似$\sim N(0, n^{-1})$

-

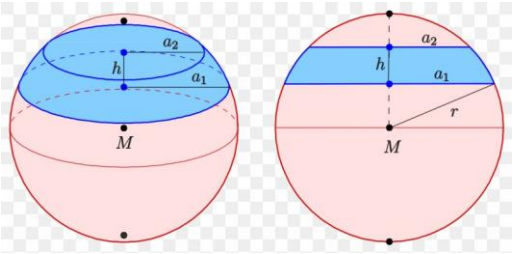
推广的阿基米德定理

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

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

这表明用低2维的平面截取球面,得到的高度固定的球带面积（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



如果地球受到来自于n维空间的降维打击，$n=?$

总结（超球部分）

概率方法提供了理解球体几何性质的简便工具，通过概率计算/蒙特卡洛（当然也可以通过微积分），我们对高维球有了至少如下认识：

- 高维球的大部分体积集中于球壳附近
- 高维球面的面积大部分集中于赤道带附近（切比雪夫concentration不等式）
- 推广的阿基米德定理：$\mathbf{u}\sim U(S^{n-1}) \Rightarrow \mathbf{u}_{[1:(n-2)]}\sim(B^{n-2})$

下一讲球对称分布是这些结果的一般化。