Math Notes

My English Note

Xia Wenxuan

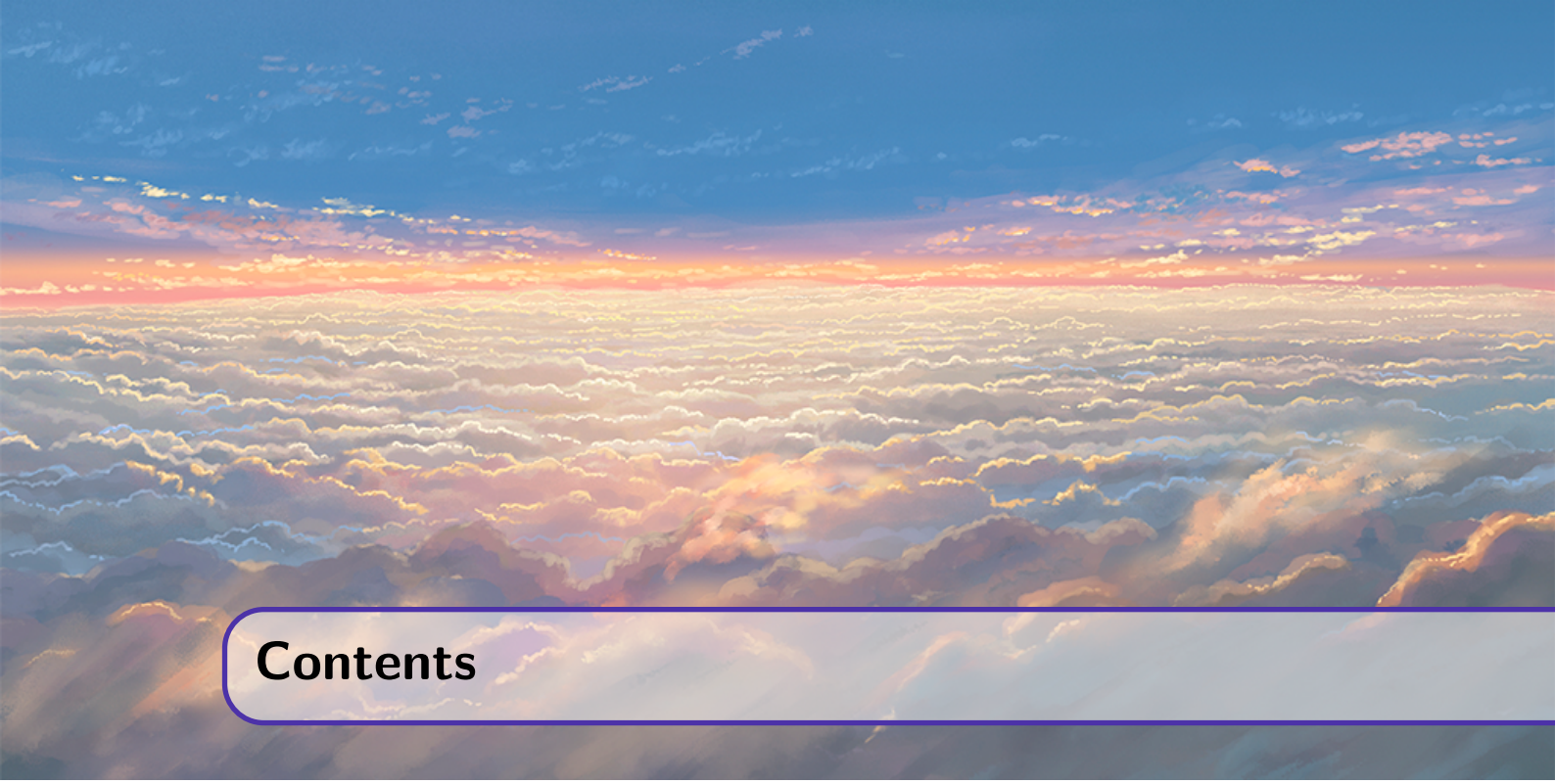
Written by Xia Wenxuan, 2021

PUBLISHED BY MYSELF

https://github.com/gegeji

All the material in this document is taken from my textbook or catechism video material, and some of it uses optical character recognition (OCR) to aid input. It may contain typographical or content inaccuracies. This document is for my personal study only. I am not responsible for the accuracy of the content of the text.

First printing, October 2, 2021



Contents

Basics of Linear Algebra

1. 对向量的介绍

1.1 Vector

Definition 1.1.1 — Vector. 一个有序的数字列表.

$$\begin{bmatrix} -1.1 \\ 0.0 \\ 3.6 \\ -7.2 \end{bmatrix} \text{ 或者 } \begin{pmatrix} -1.1 \\ 0.0 \\ 3.6 \\ -7.2 \end{pmatrix} \text{ 或者 } (-1.1, 0, 3.6, -7.2)$$

表中的数字是元素 (项、系数、分量)。元素的数量是向量的大小 (维数，长度)。大小为 n 的向量称为n 维向量。向量中的数字通常被称作标量。

用符号来表示向量，比如 α ，b，一般小写字母表示. 其它表示形式 $\boldsymbol{g}, \vec{a}$

Definition 1.1.2 — n 维向量 a 的第 i 元素. n 维向量 a 的第 i i 元素表示为 a_i.
有时 i 指的是向量列表中的第 i 个向量.

Definition 1.1.3 — $a = b$. 对于所有 i，如果有 $a_i = b_i$，则称两个相同大小的向量 a 和 b 是相等的，可写成 $a = b$

Definition 1.1.4 — stacked vector. 假设 b、c、d 是大小为 m、n、p 的向量

$$a = \begin{bmatrix} b \\ C \\ d \end{bmatrix}$$

$$a = (b_1, b_2, \ldots, b_m, c_1, c_2, \ldots, c_n, d_1, d_2, \ldots, d_p)$$

Definition 1.1.5 — 零向量. 所有项为 0 的 n 维向量表示为 0_n 或者 0

Definition 1.1.6 — 全一向量. 所有项为 1 的 n 维向量表示为 $\square_n$ 或者 1

Definition 1.1.7 — 单位向量. 当第 i 项为 1，其余项为 0 时表示为 e_i

$$e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad e_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Definition 1.1.8 — 稀疏向量. 如果一个向量的许多项都是 0，该向量为稀疏 (Sparse) 的。稀疏向量能在计算机上高效地存储和操作。

$\text{nnz}(x)$ 是指向量 x 中非零的项数 (number of non-zeros)，有时用 ℓ_0 表示。

向量 $x = (x_1, x_2)$ 可以在二维中表示一个位置或一个位移、图像、单词统计等。

1.2 Vector Space

Definition 1.2.1 — 向量空间 V. 设 V 是非空子集, P 是一数域, 向量空间 V 满足:

1. 向量加法: $V + V \to V$, 记作 $\forall x, y \in V$, 则 $x + y \in V$(加法封闭)
2. 标量乘法: $F \times V \to V$, 记作 $\forall x \in V, \lambda \in P$, 则 $\lambda x \in V$(乘法封闭)

上述两个运算满足下列八条规则 $(\forall x, y, z \in V, \lambda, \mu \in P)$

1. $x + y = y + x$ (交换律)
2. $x + (y + z) = (x + y) + z$ (结合律)
3. V 存在一个零元素, 记作 0, $x + 0 = x$
4. 存在 x 的负元素，记作 $-x$, 满足 $x + (-x) = 0$
5. $\forall x \in V$, 都有 $1x = x, 1 \in P$
6. $\lambda(\mu x) = (\lambda\mu)x$
7. $(\lambda + \mu)x = \lambda x + \mu x$
8. $\lambda(x + y) = \lambda x + \lambda y$

Corollary 1.2.1 向量空间也称为线性空间.

Corollary 1.2.2 如果 $x, y \in \mathbb{R}^2$, 则 $x + y \in \mathbb{R}^2, \lambda x \in \mathbb{R}^2 (\lambda \in \mathbb{R})$

1.3 向量运算

Definition 1.3.1 — 向量加法. n 维向量 a 和 b 可以相加，求和形式表示为 $a + b$

设向量 a, b, C 是向量空间 V 的元素，即 $a, b, c \in V$。

1. 交换律：$a + b = b + a$
2. 结合律：$(a + b) + c = a + (b + c)$ (因此可写成 $a + b + c$)
3. $a + 0 = 0 + a = a$
4. $a - a = 0$

Corollary 1.3.1 — 向量位移相加. 如果二维向量 a 和 b 都表示位移，则它们的位移之和为 $a + b$

Definition 1.3.2 — 标量与向量的乘法.

$$\beta a = \begin{bmatrix} \beta a_1 \\ \vdots \\ \beta a_n \end{bmatrix}$$

标量 β,γ 与向量 a、b
1. 结合律：$(\beta\gamma)a=\beta(\gamma a)$
2. 左分配律：$(\beta+\gamma)a=\beta a+\gamma a$
3. 右分配律：$\beta(a+b)=\beta a+\beta b$

Definition 1.3.3 — 线性组合. 对于向量 $a_1,\dots,a_m$ 和标量 $\beta_1,\dots,\beta_m$,

$$\beta_1 a_1+\cdots+\beta_m a_m$$

是向量的线性组合。$\beta_1,\dots,\beta_m$ 是该向量的系数。

■ **Example 1.1** 对于任何向量 $b\in\mathbb{R}^n$, 有如下等式

$$b=b_1e_1+\cdots+b_ne_n, b=\begin{bmatrix} b_1\\ b_2\\ \vdots\\ b_n\end{bmatrix}$$

■

1.4 内积

Definition 1.4.1 — 内积. 在数域 $\mathbb{R}$ 上的向量空间 V, 定义函数 $\langle\cdot,\cdot\rangle: V\times V\to\mathbb{R}$, 满足:
1. $\langle a,a\rangle\geq 0,\forall a\in V$, 当且仅当 $a=0$ 时 $\langle a,a\rangle=0$
2. $\langle\alpha a+\beta b,c\rangle=\alpha\langle a,c\rangle+\beta\langle b,c\rangle,\forall\alpha,\beta\in\mathbb{R}$, 且 $a,b,c\in V$
3. $\langle a,b\rangle=\langle b,a\rangle,\forall a,b\in V$

函数 $\langle\cdot,\cdot\rangle: V\times V\to\mathbb{R}$ 成为内积。

■ **Example 1.2** 在向量空间 $\mathbb{R}^n$ 上，计算两个向量对应项相乘之后求和函数

$$\langle a,b\rangle=a_1b_1+a_2b_2+\cdots+a_nb_n=a_b^T$$

$$a=\begin{bmatrix} a_1\\ a_2\\ \vdots\\ a_n\end{bmatrix}, b=\begin{bmatrix} b_1\\ b_2\\ \vdots\\ b_n\end{bmatrix}\in\mathbb{R}^n$$

■

Proof. $\langle a,a\rangle=a_1a_1+a_2a_2+\cdots+a_na_n=\sum_{i=1}^n a_i^2\geq 0, \langle a,a\rangle=0$, 则 $a=0$

$$\begin{aligned}\langle\alpha a+\beta b,c\rangle&=(\alpha a_1+\beta b_1)c_1+(\alpha a_2+\beta b_2)c_2+\cdots+(\alpha a_n+\beta b_n)c_n\\ &=\alpha\sum_{i=1}^n a_ic_i+\beta\sum_{i=1}^n b_ic_i\\ &=\alpha\langle a,c\rangle+\beta\langle b,c\rangle\end{aligned}$$

$$\langle a,b\rangle=a^Tb=b^Ta=\langle b,a\rangle$$

■

内积的性质：交换律、结合律、分配律。
交换律：$a^Tb=b^Ta$
结合律：$(\gamma a)^Tb=\gamma\left(a^Tb\right)$
分配律：$(a+b)^Tc=a^Tc+b^Tc$

1.4.1 常用的内积等式

Corollary 1.4.1 — 选出第 i 项.

$$e_i^T a = a_i$$

Corollary 1.4.2 — 向量每一项之和.

$$\mathbf{1}^T a = a_1 + \cdots + a_n$$

Corollary 1.4.3 — 向量每一项的平方和.

$$a^T a = a_1^2 + \cdots + a_n^2$$

1.5 Cauchy-Schwartz Inequality

Theorem 1.5.1 — Cauchy-Schwartz Inequality. 设 $\langle \cdot, \cdot \rangle$ 是向量空间 V 上的内积, $\forall x, y \in V$, 则有

$$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle$$

Proof. 令 $\lambda \in \mathbb{R}$, 则有 $0 \leq \langle x + \lambda y, x + \lambda y \rangle = \langle x, x \rangle + \lambda \langle y, x \rangle + \lambda \langle x, y \rangle + \lambda^2 \langle y, y \rangle = \langle x, x \rangle + 2\lambda \langle y, x \rangle + \lambda^2 \langle y, y \rangle$

则有 $\lambda^2 \langle y, y \rangle + 2\lambda \langle y, x \rangle + \langle x, x \rangle \geq 0, \forall \lambda \in \mathbb{R}$.

$\nabla = (2\langle y, x \rangle)^2 - 4\langle y, y \rangle \langle x, x \rangle \leq 0$

$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle$

当 $|\langle x, y \rangle|^2 = \langle x, x \rangle \langle y, y \rangle$ 时，有 $\langle x, x \rangle^2 + 2\lambda \langle y, x \rangle + \lambda^2 \langle y, y \rangle = 0$

也即 $\langle x + \lambda y, x + \lambda y \rangle = 0$, 因此 $x + \lambda y = 0$, 即 $x = -\lambda y$. ■

1.6 浮点运算

计算机以浮点格式存储 (实) 数值。

基本的算术运算 (加法，乘法等) 被称为浮点运算 (flop)。

算法或操作的时间复杂度：作为输入维数的函数所需要的浮点运算总数。

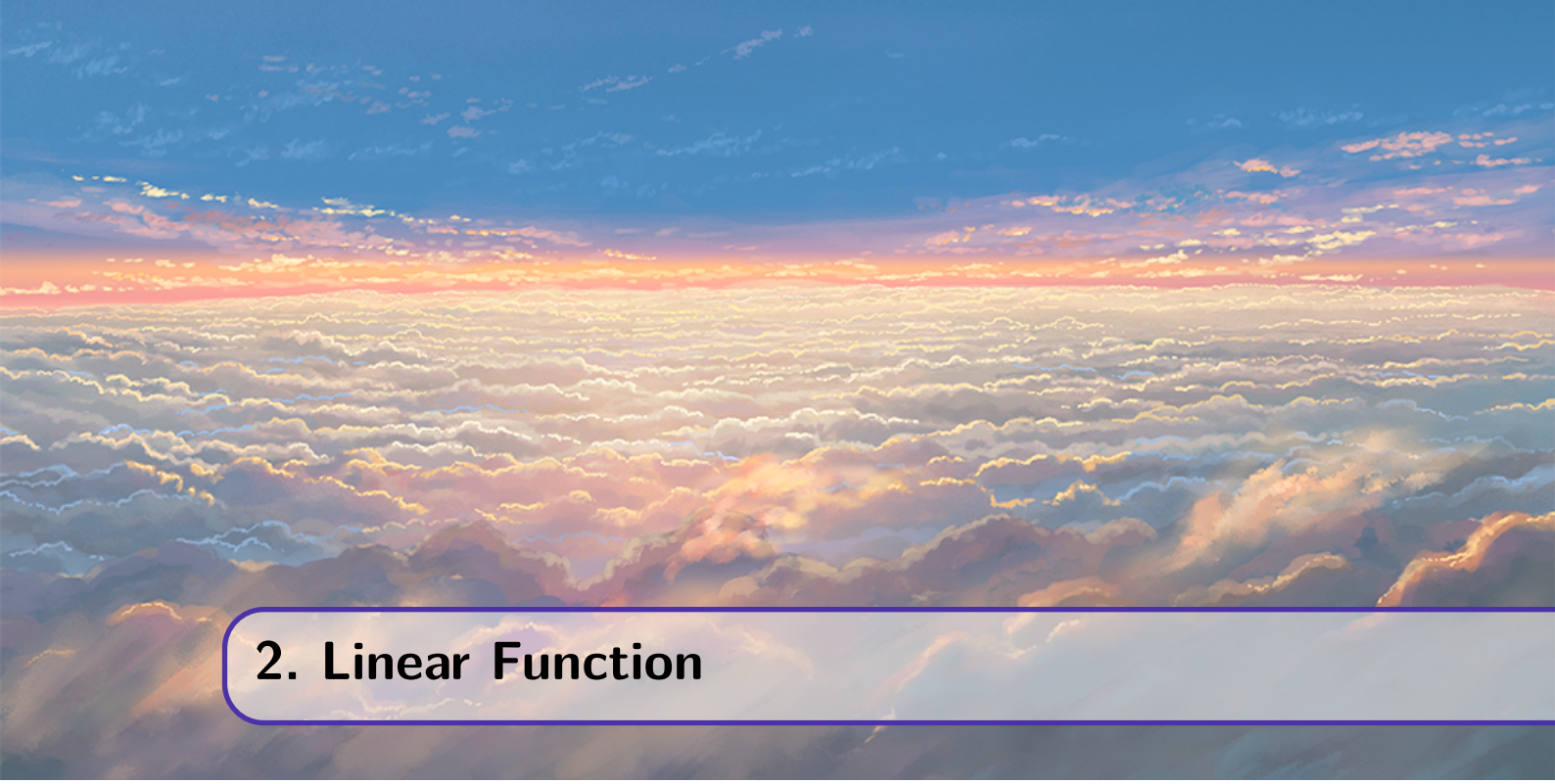
算法复杂度通常以非常粗略地近似估算。

(程序) 执行时间的粗略估计：计算机速度/flops

目前的计算机大约是 1Gflops/秒 (10^9flops/秒)

Corollary 1.6.1 假设有 n 维向量 x 和 y:

- $x + y$ 需要 n 次加法，所以时间复杂度为 (n)flops。
- $x^T y$ 需要 n 次乘法和 $n - 1$ 次加法，所以时间复杂度为 $(2n - 1)$flops。
- 对于 $x^T y$，通常将其时间复杂度简化为 $2n$，甚至为 n。
- 当 x 或 y 是稀疏的时候，算法的实际运算时间会比理论时间更少。



2. Linear Function

2.1 Linear Function

Definition 2.1.1 — Linear Function. f 是一个将 n 维向量映射成数的函数。

$$f : \mathbb{R}^n \to \mathbb{R}$$

线性函数 f 满足以下两个性质 $(k \in \mathbb{R}, x, y \in \mathbb{R}^n)$：

- 齐次性 (homogeneity): $f(kx) = kf(x)$
- 叠加性 (Additivity): $f(x + y) = f(x) + f(y)$

■ **Example 2.1** 求平均值: $f(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$ 为线性函数。 ■

■ **Example 2.2** 求最大值: $f(x) = \max\{x_1, x_2, \ldots, x_n\}$ 并不是线性函数。 ■

Proof. 令 $x = (1, -1), y = (-1, 1), \alpha = 0.5, \beta = 0.5$, 有 $f(\alpha x + \beta y) = 0 \neq \alpha f(x) + \beta f(y) = 1$

$$\begin{aligned} f(x + y) &= \max\{x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n\} \\ &\leq \max\{x_1, x_2, \ldots, x_n\} + \max\{y_1, y_2, \ldots, y_n\} \\ &\leq f(x) + f(y) \end{aligned}$$

■

Theorem 2.1.1 设 $\alpha_{1,\ldots,}\alpha_m \in \mathbb{R}, u_1, \ldots, u_m \in \mathbb{R}^n$, 则线性函数 f 满足

$$\begin{aligned} f(\alpha_1 u_1 + \alpha_2 u_2 + \ldots + \alpha_m u_m) &= f(\alpha_1 u_1) + f(\alpha_2 u_2 + \ldots + \alpha_m u_m) \\ &= \alpha_1 f(u_1) + f(\alpha_2 u_2 + \ldots + \alpha_m u_m) \\ &= \alpha_1 f(u_1) + \alpha_2 f(u_2) + \ldots + \alpha_m f(u_m) \end{aligned}$$

Definition 2.1.2 — 内积函数 (inner product function). 对于 n 维向量 a, 满足以下形式的函数被称为内积函数

$$f(x) = a^T x = a_1 x_1 + a_2 x_2 + \ldots + a_n x_n$$

上述 $f(x)$ 可以看作是每项 x_{i} 的加权之和。

Corollary 2.1.2 内积函数都是线性的.

Proof.

$$\begin{aligned} f(\alpha x + \beta y) &= a^T(\alpha x + \beta y) \\ &= a^T(\alpha x) + a^T(\beta y) \\ &= \alpha\left(a^T x\right) + \beta\left(a^T y\right) \\ &= \alpha f(x) + \beta f(y) \end{aligned}$$

■

Definition 2.1.3 — 仿射函数 (affine function). 其一般形式为 $f(x) = a^T x + \mathrm{b}$, 其中 $a \in \mathbb{R}^n, \quad b \in \mathbb{R}$ 为标量。

Theorem 2.1.3 函数 $f : \mathbb{R}^n \to \mathbb{R}$ 为仿射函数需要满足
$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y), \alpha + \beta = 1, \alpha, \beta \in \mathbb{R}, x, y \in \mathbb{R}^n$

2.2 泰勒展开

Definition 2.2.1 — 函数 f 第 i 个分量的一阶偏导数.

$$\begin{aligned} \frac{\partial f}{\partial z_i}(z) &= \lim_{t \to 0} \frac{f\left(z_1, \cdots, z_{i-1}, z_i + t, z_{i+1}, \cdots, z_n\right) - f(z)}{t} \\ &= \lim_{t \to 0} \frac{f\left(z + t e_i\right) - f(z)}{t} \end{aligned}$$

Definition 2.2.2 — f 在点 z 的梯度.

$$\nabla f(z) = \begin{bmatrix} \frac{\partial f}{\partial z_1}(z) \\ \vdots \\ \frac{\partial f}{\partial z_n}(z) \end{bmatrix}$$

Definition 2.2.3 — Taylor's Approximation.

$$\begin{aligned} f(x) = f(z) &+ \frac{\partial f}{\partial x_1}(z)\left(x_1 - z_1\right) + \frac{\partial f}{\partial x_2}(z)\left(x_2 - z_2\right) + \cdots + \frac{\partial f}{\partial x_n}(z)\left(x_n - z_n\right) \\ &+ \frac{1}{2!} \sum_{j=1}^{n} \sum_{i=1}^{n} \frac{\partial f^2}{\partial x_i \partial x_j}(z)\left(x_i - z_i\right)\left(x_{\mathrm{j}} - z_j\right) + \cdots \end{aligned}$$

■ **Example 2.3** 泰勒公式利用多项式在一点附近逼近函数

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots + (-1)^{k-1} \frac{x^{2k-1}}{(2k-1)!} + \frac{\sin\left[\xi + (2k+1)\frac{\pi}{2}\right]}{(2k+1)!} x^{2k+1}$$

一次逼近: $\sin x \approx x$
三次逼近: $\sin x \approx x - \frac{x^3}{3!}$ ■

Proof.

$$f(x) = P_n(x) + R_n(x)$$

$P_n(x) = a_0 + a_1 (x - x_0) + a_2 (x - x_0)^2 + \cdots + a_n (x - x_0)^n$
$R_n(x) = o(x - x_0)^n$
$f(x) \approx P_n(x)$
$\therefore P_n(x_0) = f(x_0), P_n'(x_0) = f'(x_0), P_n''(x_0) = f''(x_0), \cdots, P_n^{(n)}(x_0) = f^{(n)}(x_0)$
要求 $P_n(x_0) = f(x_0) \Rightarrow a_0 = f(x_0)$

$$P_n'(x) = a_1 + 2a_2 (x - x_0) + \cdots + na_n (x - x_0)^{n-1} \Rightarrow a_1 = f'(x_0)$$

依此类推. $a_n = \frac{f^{(n)}(x_0)}{n!}$ ■

Corollary 2.2.1 — n 阶泰勒多项式.

$$P_n(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

where $a_n = \frac{f^{(n)}(x_0)}{n!}$

Corollary 2.2.2 — 对于高阶余项的公式. 带拉格朗日余项的泰勒公式

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)^{n+1}$$

$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)^{n+1}$ (ξ 在 x_0 与 x 之间)

Corollary 2.2.3 — 麦克劳林 (Maclaurin) 公式. 在零点展开麦克劳林 (Maclaurin) 公式

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \cdots + \frac{f^{(n)}(0)}{n!}x^n + \frac{f^{(n+1)}(\theta x)}{(n+1)!}x^{n+1} \qquad (0 < \theta < 1)$$

Definition 2.2.4 — 一阶泰勒公式. 假设 $f : \mathbb{R}^n \to \mathbb{R}$, 函数 f 在 z 点可导

$$\hat{f}(x) = f(z) + \frac{\partial f}{\partial x_1}(z)(x_1 - z_1) + \ldots + \frac{\partial f}{\partial x_1}(z)(x_n - z_n)$$

当 x 非常接近 z 时，$\hat{f}(x)$ 也非常接近 $f(z)$。$\hat{f}(x)$ 是关于 x 的一个仿射函数。

Corollary 2.2.4 — 一阶泰勒公式的内积形式.

$$\hat{f}(x) = f(z) + \nabla f(z)^T (x - z) \quad \nabla f(z) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(z) \\ \vdots \\ \frac{\partial f}{\partial x_n}(z) \end{bmatrix}$$

一维时，$\hat{f}(x) = f(z) + f'(z)(x - z)$

■ **Example 2.4**

$$f(x) = x_1 - 3x_2 + e^{2x_1 + x_2 - 1}$$

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \end{bmatrix} = \begin{bmatrix} 1 + 2e^{2x_1 + x_2 - 1} \\ -3 + e^{2x_1 + x_2 - 1} \end{bmatrix}$$

函数 f 在 0 点的一阶泰勒公式为:

$$\hat{f}(x) = f(0) + \nabla f(0)^T (x - 0) = e^{-1} + \left(1 + 2e^{-1}\right) x_1 + \left(-3 + e^{-1}\right) x_2$$

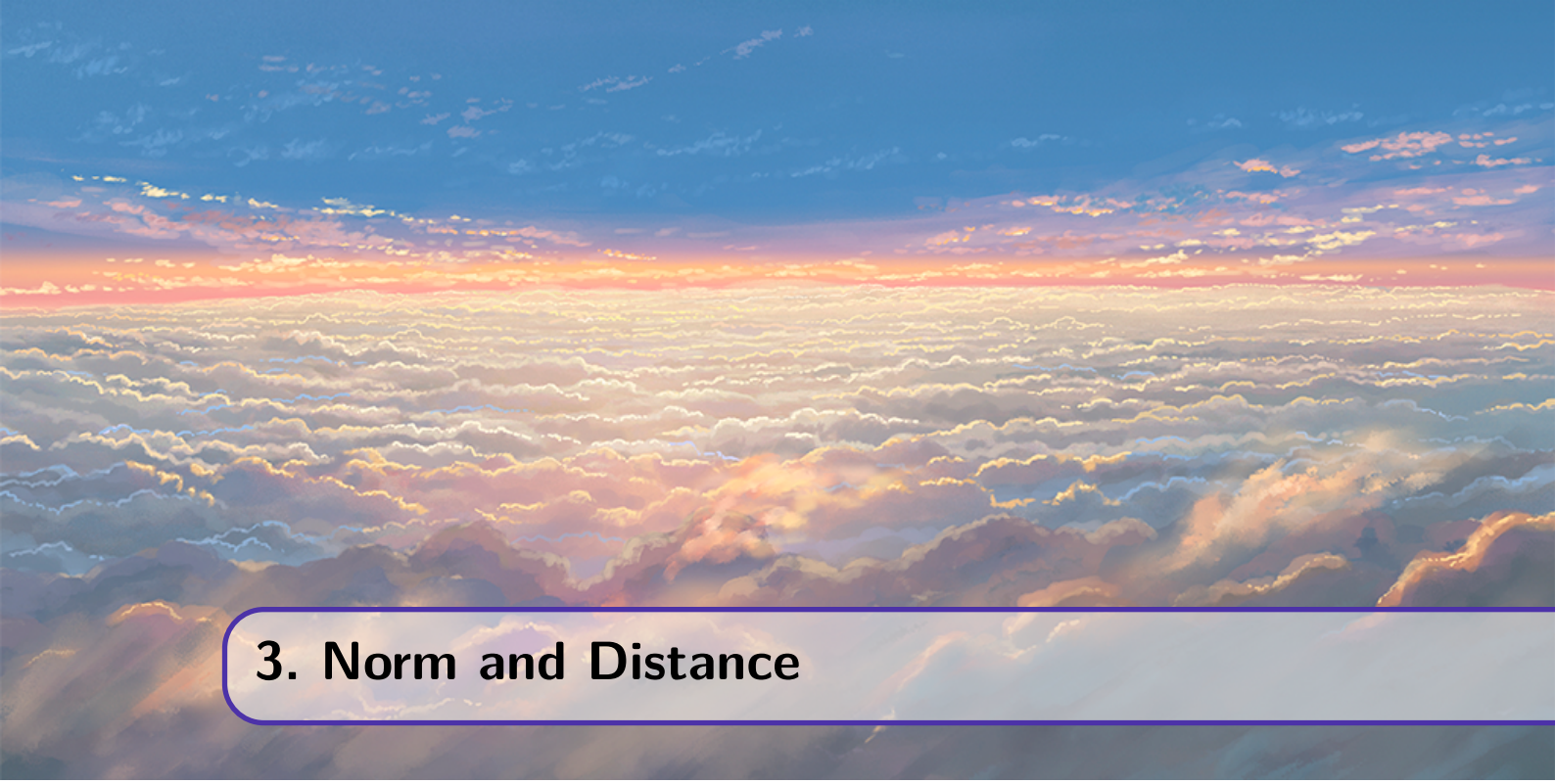
■

2.3 Regression Model

Definition 2.3.1 — Regression Model. 回归模型 (regression model) 为关于 x 的仿射函数

$$\hat{y} = x^T \beta + v$$

x 是特征向量 *(feature vector)*，它的元素 x_i 称为回归元 *(regressors)*。n 维向量 β 是权重向量 *(weight vector)*。标量 v 是偏移量 *(offset)*。标量 $\hat{y}$ 是预测值 *(prediction)*。表示某个实际结果或因变量，用 y 表示。



3. Norm and Distance

3.1 Norm

Definition 3.1.1 — Vector Norm. 在向量空间中存在一个函数 $\|\cdot\|:\mathbb{R}^n\to\mathbb{R}$, 且满足以下条件

- 齐次性: $\|\alpha x\|=|\alpha|\|x\|,\quad \alpha\in\mathbb{R}$ 且 $x\in\mathbb{R}^n$;
- 三角不等式: $\|x+y\|\le\|x\|+\|y\|,\quad x,y\in\mathbb{R}^n$;
- 非负性: $\|x\|\ge 0, \boldsymbol{x}\in\mathbb{R}^n$ 且 $\|\boldsymbol{x}\|=0\Leftrightarrow\boldsymbol{x}=0$;

则称 $\|\cdot\|$ 为向量范数。

■ **Example 3.1 — ℓ_1-范数（曼哈顿范数，Manhattan norm）.**

$$\|x\|_1=|x_1|+|x_2|+\ldots+|x_n|\quad x,y\in\mathbb{R}^n,\alpha\in\mathbb{R}$$

■

Proof.

$$\|\alpha x\|_1=|\alpha x_1|+|\alpha x_2|+\cdots+|\alpha x_n|=|\alpha|\|x\|_1\ge 0$$

$$\|x+y\|_1=|x_1+y_1|+\cdots+|x_n+y_n|\le|x_1|+|y_1|+\cdots+|x_n|+|y_n|=\|x\|_1+\|y\|_1$$

■

■ **Example 3.2 — ℓ_2-范数（欧几里得范数，Euclidean norm）.**

$$\|x\|_2=\sqrt{\left(x_1^2+x_2^2+\cdots+x_n^2\right)}=\sqrt{x^Tx}=(\langle x,x\rangle)^{\frac{1}{2}}$$

■

Proof.

$$\|\alpha x\|_2=(\langle\alpha x,\alpha x\rangle)^{\frac{1}{2}}=|\alpha|(\langle x,x\rangle)^{\frac{1}{2}}=|\alpha|\|x\|_2$$

$$\begin{aligned}\|x+y\|_2^2 &= \langle x+y, x+y\rangle = \langle x,x\rangle + \langle x,y\rangle + \langle y,x\rangle + \langle y,y\rangle \\ &= \|x\|_2^2 + 2\langle x,y\rangle + \|y\|_2^2 \le \|x\|_2^2 + 2\|x\|_2\|y\|_2 + \|y\|_2^2 \\ &= \left(\|x\|_2 + \|y\|_2\right)^2\end{aligned}$$

$$\|x+y\|_2 \le \|x\|_2 + \|y\|_2$$

■

Corollary 3.1.1 — 柯西—施瓦茨不等式.

$$|\langle x,y\rangle|^2 \le \langle x,x\rangle\langle y,y\rangle = \|x\|_2^2\|y\|_2^2$$

Definition 3.1.2 — ℓ_∞-范数.

$$\|x\|_\infty = \max_{1\le i\le n} |x_i|, x \in \mathbb{R}^n$$

Proof.

$$\begin{aligned}\max_{1\le i\le n} |x_i| &\le \left(|x_1|^p + \cdots + |x_i|^p + \cdots + |x_n|^p\right)^{1/p} \\ &\le \left(n \max_{1\le i\le n} |x_i|^p\right)^{1/p} \\ &= n^{1/p} \max_{1\le i\le n} |x_i| \\ &\to \max_{1\le i\le n} |x_i| \quad (p\to\infty)\end{aligned}$$

■

Definition 3.1.3 — ℓ_p-范数.

$$\|x\|_p = \left(x_1^p + x_2^p + \cdots + x_n^p\right)^{\frac{1}{p}}, \quad x \in \mathbb{R}^n, p \ge 1$$

ℓ_1 范数 $\|x\|_1$,ℓ_2-范数 $\|x\|_2$, ℓ_∞-范数是 ℓ_p-范数的特例。

证明可以使用以下两条不等式

Theorem 3.1.2 — Minkowshi Inequality.

$$\left(\sum_{i=1}^n |x_i+y_i|^p\right)^{\frac{1}{p}} \le \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}} + \left(\sum_{i=1}^n |y_i|^p\right)^{\frac{1}{p}}, p \ge 1, x, y \in \mathbb{R}^n$$

Theorem 3.1.3 — Hölder Inequality.

$$\sum_{i=1}^n |x_i y_i| \le \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} \left(\sum_{i=1}^n |y_i|^q\right)^{1/q}, \frac{1}{p} + \frac{1}{q} = 1, 1 < p, q < \infty$$

3.2 Root Mean Square Value (RMS)

Definition 3.2.1 — 向量 x 的均方值 (mean-square value). 向量 $x \in \mathbb{R}^n$ 的均方值 (mean-square value)

$$\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n} = \frac{\|x\|_2^2}{n}$$

Definition 3.2.2 — n 维向量 x 的均方根 (root-mean-square value, RMS).

$$\mathrm{rms}(x) = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}} = \frac{\|x\|_2}{\sqrt{n}}$$

$\mathrm{rms}(x)$ 给出了 $|x_i|$ 的 “典型” (typical) 值。例如, $\mathrm{rms}(\mathbf{1}) = 1$ (与 n 无关)。均方根 (RMS) 值对于比较不同长度的向量大小是比较有用的。

3.3 Chebyshev's Inequality

Theorem 3.3.1 — Chebyshev's Inequality.

$$P(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

$$P(|X - \mu| < \varepsilon) \geq 1 - \frac{\sigma^2}{\varepsilon^2}$$

Theorem 3.3.2 — Chebyshev's Inequality. 假设 k 为向量 x 分量满足条件 $|x_i| \geq a$ 的个数，即 $x_i^2 \geq a^2$ 的个数。

因此: $\|x\|_2^2 = x_1^2 + x_2^2 + \cdots + x_n^2 \geq ka^2$

将 a^2 移项，可得到 $k \leq \frac{\|x\|_2^2}{a^2}$

满足 $|x_i| \geq a$ 的 x_i 数量不会超过 $\frac{\|\|\|_2^2}{a^2}$

Corollary 3.3.3 — Chebyshev's Inequality Using RMS.

$$\mathrm{rms}(x) = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}} = \frac{\|x\|_2}{\sqrt{n}}$$

$|x_i| \geq a$ 的项数占整体的比例不会超过 $\left(\frac{\mathrm{rms}(x)}{a}\right)^2$, 即 $\frac{k}{n} \leq \left(\frac{\mathrm{rms}(x)}{a}\right)^2$

3.4 Distance

Definition 3.4.1 — Euclidean distance. n 维向量 a 和 b 之间的欧氏距离

$$\mathrm{dist}(a, b) = \|a - b\|_2$$

Definition 3.4.2 — RMS deviation. $\mathrm{rms}(a - b)$ 是 a 和 b 之间的均方根偏差.

Theorem 3.4.1 — Trianglar Inequality.

$$\|a-c\|_2 = \|(a-b)+(b-c)\|_2 \le \|a-b\|_2 + \|b-c\|_2$$

3.4.1 Feature Distance and Nearest Neighbor

Definition 3.4.3 — Feature Distance. 如果 x 和 y 分别为两个实体的特征向量, 那么它们的特征距离 (feature distance) 为 $\|x-y\|_2$

Definition 3.4.4 给定向量 x，一个组向量 $Z_1, \ldots, Z_m$, 当 $\hat{q}_j$ 满足:

$$\|x - z_j\|_2 \le \|x - z_i\|_2, \quad i = 1, \ldots, m$$

则称 z_j 是 x 的最近邻 (nearest neighbor)

3.5 Standard Derivation

Definition 3.5.1 — 算术平均值. 对于 n 维向量 x

$$\text{avg}(x) = \frac{\mathbf{1}^T x}{n}$$

Definition 3.5.2 — De-meaned Vector.

$$\tilde{x} = x - \text{avg}(x)\mathbf{1}$$

因此 $\text{avg}\, \boldsymbol{g}(\tilde{x}) = 0$

Definition 3.5.3 — x 的标准差.

$$\text{std}(x) = \text{rms}(\tilde{x}) = \frac{\left\|x - \left(1^T x/n\right) 1\right\|_2}{\sqrt{n}}$$

$\text{std}(x)$ 表示数据元素的变化程度。对于常数 α，当且仅当 $x = \alpha\mathbf{1}$ 时，$\text{std}(x) = 0$.

Theorem 3.5.1

$$\text{rms}(x)^2 = \text{avg}(x)^2 + \text{std}(x)^2$$

3.6 Angle

Definition 3.6.1 — 两个非零向量 a 和 b 之间的角 (angle).

$$\angle(a,b) = \arccos\left(\frac{a^T b}{\|a\|_2 \|b\|_2}\right)$$

$\angle(a,b)$ 的取值范围为 $[0,\pi]$, 且满足

$$a^T b = \|a\|_2 \|b\|_2 \cos(\angle(a,b))$$

在二维和三维向量之中，这里的角与普通角度 (ordinary angle) 是一致的。
- $\theta = \frac{\pi}{2} = 90°$: a 和 b 为正交，写作 $a \perp b (a^T b = 0)$。
- $\theta = 0$: a 和 b 为同向的 $(a^T b = \|a\|\|b\|)$。
- $\theta = \pi = 180°$: a 和 b 为反向的 $(a^T b = -\|a\|\|b\|)$。

- $\theta < \frac{\pi}{2} = 90°$：a 和 b 成锐角 $(a^T b > 0)$。
- $\theta > \frac{\pi}{2} = 90°$：a 和 b 成钝角 $(a^T b < 0)$。

Definition 3.6.2 — 球面的距离.

$$\mathrm{R}\angle(a, b)$$

3.6.1 相关系数

给定向量 a 和 b，其去均值向量为：

$$\tilde{a} = a - \mathrm{avg}(a)1, \tilde{b} = b - \mathrm{avg}(b)1$$

Definition 3.6.3 — a 和 b 的相关系数.

$$\rho = \frac{\tilde{a}^T \tilde{b}}{\|\tilde{a}\|_2 \|\tilde{b}\|_2} = \cos\angle(\tilde{a}, \tilde{b})$$

where $\tilde{a} \neq 0, \tilde{b} \neq 0$.

■ **Example 3.3** 高度相关的向量：
- 邻近地区的降雨时间序列。
- 类型密切相关文档的单词计数向量。
- 同行业中类似公司的日收益。

比较不相关的向量：
- 无关的向量。
- 音频信号 (比如，在多轨录音中的不同轨)。

负相关的向量：
- 深圳与墨尔本的每天气温变化

■

4. 以 k-Means 算法为例的优化问题

Problem 4.1 假设 N 个样本向量 $x_1, \ldots, x_N \in \mathbb{R}^n$，需要找到中心向量 z 满足

$$\min_{z \in \mathbf{R}^n} \sum_{i=1}^{N} \|x_i - z\|_2^2$$

Definition 4.0.1 — 高阶无穷小记号 o. 设 x, y 是同一变化过程中的无穷小，即 $x \to 0, y \to 0$, 如果它们极限

$$\lim \frac{y}{x} = 0$$

则称 y 是 x 的高阶无穷小，记作 $y = o(x)$.

Corollary 4.0.1

$$\lim \frac{y}{Cx} = \frac{1}{C} \lim \frac{y}{x} = 0$$

也即则称 y 是 Cx 的高阶无穷小，记作 $y = o(Cx)$ 。

Proposition 4.0.2 — 优化求解的必要条件. 假设函数 f 在 $\hat{x}$ 可微，则有

$$\hat{x} = \arg\min_{x \in \mathbb{R}^n} f(x) \Rightarrow \nabla f(\hat{x}) = 0$$

Proof. 假设函数 f 在 $\hat{x}$ 一阶泰勒展开, 有

$$f(x) = f(\hat{x}) + \langle \nabla f(\hat{x}), x - \hat{x} \rangle + o\left(\|x - \hat{x}\|_2\right)$$

假设 $\delta f(\hat{x}) \neq 0$, 则令 $\tilde{x} = \hat{x} - t \nabla f(\hat{x}), t > 0$, 可得

$$f(\tilde{x}) = f(\hat{x}) - t\|\nabla f(\hat{x})\|_2^2 + o\left(t\|\nabla f(\hat{x})\|_2\right)$$

当 $t \to 0$ 则 $t\|\nabla f(\hat{x})\|_2 \to 0$，高阶无穷小 $o'\left(t\|\nabla f(\hat{x})\|_2\right) \to 0$
当 t 足够小时，存在 $t\|\nabla f(\hat{x})\|_2 \geq o\left(t\|\nabla f(\hat{x})\|_2\right)$，即

$$-t\|\nabla f(\hat{x})\|_2^2 + o\left(t\|\nabla f(\hat{x})\|_2\right) \leq 0$$

$$f(\tilde{x}) = f(\hat{x}) - t\|\nabla f(\hat{x})\|_2^2 + o\left(t\|\nabla f(\hat{x})\|_2\right) \leq f(\hat{x})$$

与 $\hat{x} = \arg\min_{\mathbf{R}^n} f(x)$ 矛盾。
$\nabla f(\hat{x}) = 0$, 是最优问题解的必要条件。通常 $\nabla f(\hat{x}) = 0 \not\Rightarrow \hat{x} = \arg\min_{\mathbf{R}^n} f(x)$。 ■

■ **Example 4.1**

$$f(x) = -x^2, \quad x \in \mathbb{R}, \hat{x} = \operatorname{argmin}_{\mathbb{R}} f(x)$$

$\nabla f(\hat{x}) = 0$, 则有 $-2\hat{x} = 0$, 即 $\hat{x} = 0$

$$f(\hat{x}) = 0 \geq f(x), \quad x \in \mathbf{R}$$

(最大值!) ■

4.1 Convex Set

Definition 4.1.1 — 凸集. $\forall x, y \in \Omega, \alpha \in \mathbb{R}, 0 \leq \alpha \leq 1$ 有

$$\alpha x + (1-\alpha)y \in \Omega$$

则定义域 $\Omega \in \mathbb{R}^n$ 称为凸的 (Convex) 集合
(域内两点连线之间都属于这个域)

Definition 4.1.2 — 凸函数. 设函数 $f(x)$ 定义于称为凸的定义域 $\Omega \in \mathbb{R}^n$ 满足

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y), \forall x, y \in \Omega, \alpha \in \mathbf{R}, 0 \leq \alpha \leq 1$$

称其为凸函数。

■ **Example 4.2**

$$f(x) = x^2, x \in \mathbf{R}$$

$$\begin{aligned} f(\alpha x + (1-\alpha)y) &= (\alpha x + (1-\alpha)y)^2 \\ &= \alpha^2 x^2 + 2\alpha(1-\alpha)xy + (1-\alpha)^2 y^2 \\ &= \alpha x^2 + (1-\alpha)y^2 + \left(\alpha^2 - \alpha\right)x^2 + \left(\alpha^2 - \alpha\right)y^2 + 2\alpha(1-\alpha)xy \\ &= \alpha x^2 + (1-\alpha)y^2 - \alpha(1-\alpha)(x-y)^2 \\ &\leq \alpha x^2 + (1-\alpha)y^2 = \alpha f(x) + (1-\alpha)f(y) \end{aligned}$$

■

■ **Example 4.3** $f(x) = \|x\|$, 其中 $\|\bullet\|$ 表示 $\mathbb{R}^n$ 上的向量范数, $x \in \mathbb{R}^n$. ■

Proof.

$$\|\alpha x + (1-\alpha)y\| \leq \|\alpha x\| + \|(1-\alpha)y\| = |\alpha|\|x\| + |1-\alpha|\|y\|$$

■

■ **Example 4.4**

$$f(x) = \|x\|_2^2, x \in \mathbb{R}^n$$

■

Theorem 4.1.1 — 可微函数 f 是凸函数的充要条件.

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall x, y$$

Proof. 首先，证明一维情况 $f : \mathbb{R} \to \mathbb{R}, \alpha \in [0, 1]$.

$\Rightarrow$ 充分条件: $f(\alpha x + (1-\alpha)y) = f(x + (1-\alpha)(y-x)) \leq \alpha f(x) + (1-\alpha)f(y)$, 有

$$f(y) \geq f(x) + \frac{f(x + (1-\alpha)(y-x)) - f(x)}{(1-\alpha)(y-x)}(y-x)$$

令 $\alpha \to 1^-$, 则有 $f(y) \geq f(x) + f'(x)(y-x)$.

$\Leftarrow$ 必要条件：令 $y \neq x, z = \alpha x + (1-\alpha)y$ 则有

$$f(x) \geq f(z) + f'(z)(x-z), f(y) \geq f(z) + f'(z)(y-z)$$

可得

$$\begin{aligned} \alpha f(x) + (1-\alpha)f(y) &\geq f(z) + \alpha f'(z)(x-z) + (1-\alpha)f'(z)(y-z) \\ &= f(z) + f'(z)(\alpha x + (1-\alpha)y - z) \\ &= f(z) \end{aligned}$$

证明 n 维情况 $f : \mathbb{R}^n \to \mathbb{R}$.

$\Rightarrow$ 充分条件：令 $g(t) = f(tx + (1-t)y), t \in \mathbb{R}$, 则 $g'(t) = \langle \nabla f(tx + (1-t)y), x - y \rangle$ 由于 f 是凸函数, 证明 $g(t)$ 也是凸函数；并可得 $g(0) \geq g(1) + g'(1)(-1)$, 得证.

$\Leftarrow$ 必要条件：与一维类似。 ■

Theorem 4.1.2 如果可微函数 f 是凸函数，则有

$$\hat{x} = \arg\min_{x \in \mathbb{R}^n} f(x) \Leftrightarrow \nabla f(\hat{x}) = 0$$

Proof. 已证 $\hat{x} = \arg\min_{x \in \mathbb{R}^n} f(x) \Rightarrow$ 可得 $\nabla f(\hat{x}) = 0$

只需证 $\nabla f(\hat{x}) = 0 \Rightarrow \hat{x} = \arg\min_{x \in \mathbb{R}^n} f(x)$.

由于函数 f 是可微凸的, 则有 $\forall x \in \mathbb{R}^n$,

$$\begin{aligned} f(x) &\geq f(\hat{x}) + \langle \nabla f(\hat{x}), x - \hat{x} \rangle \\ &\geq f(\hat{x}) + \langle 0, x - \hat{x} \rangle \geq f(\hat{x}) \end{aligned}$$

可得 $f(x) \geq f(\hat{x}), \hat{x} = \arg\min_{x \in \mathbb{R}^n} f(x)$. ■

4.2 向量偏导

Definition 4.2.1 — 向量对向量的导数.

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, z = \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix}$$

$$\nabla f(z) = \begin{bmatrix} \frac{\partial f(z)}{\partial z_1} \\ \vdots \\ \frac{\partial f(z)}{\partial z_n} \end{bmatrix}$$

■ **Example 4.5**

$$f(z) = x^T z + z^T z = \sum_{i=1}^{n} \{x_i z_i + z_i^2\}$$

$$\nabla f(z) = \begin{bmatrix} \frac{\partial f(z)}{\partial z_1} \\ \vdots \\ \frac{\partial f(z)}{\partial z_n} \end{bmatrix} = \begin{bmatrix} x_1 + 2z_1 \\ \vdots \\ x_n + 2z_n \end{bmatrix} = x + 2z$$

■

问题4.1中已知目标函数是凸函数。(见4.2, 4.3, 4.1)
则可以求解

$$f(z) = \sum_{i=1}^{N} \|x_i - z\|_2^2 = \sum_{i=1}^{N} \langle x_i - z, x_i - z \rangle = \sum_{i=1}^{N} \{x_i^T x_i - 2x_i^T z + z_i^T z\}$$

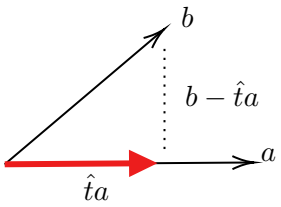
利用等价条件4.1.1

$$\nabla f(z) = \sum_{i=1}^{N} \{-2x_i + 2z\} = 0$$

(求导 4.2.1)

$$z = \frac{1}{N} \sum_{i=1}^{N} x_i$$

4.3 标量优化问题的例子：投影问题



Problem 4.2 假设 $a, b \in \mathbb{R}^n, a \neq 0, t \in \mathbb{R}$，当 t 多大时，ta 到 b 之间的距离最小

$$\hat{t} = \min_t \|ta - b\|_2^2$$

$$f(t) = \|ta - b\|_2^2 = \langle ta - b, ta - b \rangle = t^2 a^T a - 2ta^T b + b^T b$$

$$\nabla f(t) = 2ta^T a - 2a^T b = 0$$

$$\hat{t} = \frac{a^T b}{a^T a} = \frac{a^T b}{\|a\|_2^2}$$

4.4 Clustering

将物理或抽象对象的集合分成由类似特征组成的多个类的过程称为聚类 *(clustering)*.

目标：分成 k 个集合，尽量使得同一个集合中的向量彼此接近。

Notation 4.1. 给定 N 个 n 维向量 $x_1, \ldots, x_N \in \mathbb{R}^n$

- 标签 $c_i \in \{1, 2, \cdots, k\}$ 表示向量 x_i 所属类别, 例如 $c_i = 2$ 表示 x_i 属于第 2 类。
- 对于 $j = 1, \ldots, k, \quad G_j = \{i : c_i = j\}$ 表示属于第 j 类的向量 x_i 的下标集合。
- 向量 $z_j, j = 1, \ldots, k$, 表示同属于 j 类的向量 $x_i, i \in G_j$ 的聚类中心。

聚类目标是找到向量 x_i 的“标签 c_i”和“聚类中心 z_j”

Problem 4.3

$$\min_{z_j} \sum_{i \in G_j} \|x_i - z_j\|_2^2, j = 1, \cdots, k$$

$$c_i = \underset{j=\{1,\cdots,k\}}{\operatorname{argmin}} \|x_i - z_j\|_2^2, i = 1, 2, \ldots, N$$

k-means 算法是将 N 向量 $x_i \in \mathbb{R}^n$ 划分成 k 类的迭代聚类算法。

Algorithm 1: k-means Algorithm

1. 在 N 个点中随机选取 k 个点，分别作为聚类中心 z_j;
2. 更新聚类标签 c_i：计算每个点 x_i 到 k 个聚类中心 z_j 的距离, 并将其分配到最近的聚类中心 z_j 所在的聚类中 $c_i = j$;
3. 更新聚类中心 z_j ： 重新计算每个聚类现在的质心，并以其作作为新的聚类中心, 根据更新标签 c_i, 更新属于第 j 类下标集合 $G_j = \{i : c_i = j\}$, 重新计算 c_i 类的聚类中心 z_j;
4. 重复步骤 2、3，直到所有聚类中心不再变化

Proof. 更新聚类标签 c_i:

$$\|x_i - z_j\|_2^2 = \operatorname{argmin} \left\{ \|x_i - z_1\|_2^2, \|x_i - z_2\|_2^2, \cdots, \|x_i - z_k\|_2^2 \right\}$$

更新聚类中心 z_j:

$$\nabla f_j(z_j) = \sum_{i \in G_j} 2(x_i - z_j) = 0$$

$$z_j = \frac{1}{|G_j|} \sum_{i \in G_j} x_i$$

$|G_j|$ 表示集合 G_j 中元素的数目。 ■

在每一次迭代中目标函数 J 都会下降, 直到聚类中心 $z_1, \cdots, z_k$ 和划分聚类标签集合 $G_1, \cdots, G_k$ 不再变化。

但是 k-means 算法依赖于初始随机生成的聚类中心，只可得到目标函数 J 的局部局部最优。

解决方案：使用不同的 (随机的) 初始聚类中心运行 k-means 算法若干次，取目标函数 J 值最小的一次作为最终的聚类结果。

5. Linear Independence

5.1 线性相关、线性无关

Definition 5.1.1 — 线性相关 (linearly dependent). 定义：对于向量 $a_1, \dots a_m \in \mathbb{R}^n$, 如果存在不全为零的数 $\beta_1, \dots \beta_m \in \mathbb{R}$, 使得

$$\beta_1 a_1 + \cdots + \beta_m a_m = 0$$

则称向量 $a_1, \dots a_m$ 是线性相关 (linearly dependent)。

线性相关等价于至少有一个向量 a_i 是其它向量的线性组合。

Corollary 5.1.1 向量集 $\{a_1\}$ 是线性相关的，当且仅当 $a_1 = 0$ 。
向量集 $\{a_1, a_2\}$ 是线性相关的，当且仅当其中一个 $a_1 = \beta a_2, \beta \neq 0$ 。

Definition 5.1.2 — 线性独立 (linearly dependent). 如果 n 维向量集 $\{a_1, \dots, a_m\}$ 不是线性相关的，即线性独立 (linearly dependent), 也称线性无关，即：

$$\beta_1 a_1 + \cdots + \beta_m a_m = 0$$

当且仅当 $\beta_1 = \cdots = \beta_m = 0$ ，上述等式成立。

线性无关等价于不存在一个向量 a_i 是其它向量的线性组合。

Corollary 5.1.2 注：一个 n 维向量集最多有 n 个线性无关的向量, 也就是说如果 n 维向量集有 $n + 1$ 个向量，那它们必线性相关

■ **Example 5.1** n 维单位向量 $e_1, \dots, e_n$ 是线性独立的。 ■

■ **Example 5.2**

$$a_1 = \begin{bmatrix} 1 \\ -2 \\ 0 \end{bmatrix}, \quad a_2 = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}, \quad a_3 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

$$\beta_1 a_1 + \beta_2 a_2 + \beta_3 a_3 = \begin{bmatrix} \beta_1 - \beta_2 \\ -2\beta_1 + \beta_3 \\ \beta_2 + \beta_3 \end{bmatrix} = 0$$

$$\beta_1 = \beta_2 = \beta_3 = 0$$

■

Theorem 5.1.3 假设 x 是线性无关向量 $a_1, \ldots, a_k$ 的线性组合：

$$x = \beta_1 a_1 + \cdots \beta_k a_k$$

则其系数 $\beta_1, \ldots \beta_k$ 是唯一的，即如果有：

$$x = \gamma_1 a_1 + \cdots \gamma_k a_k$$

则对于 $i = 1, \ldots k$, 有 $\beta_i = \gamma_i$ 。

Proof. 系数是唯一的原因：

$$(\beta_1 - \gamma_1)\, a_1 + \cdots (\beta_k - \gamma_k)\, a_k = x - x = 0$$

由于向量 $a_1, \ldots, a_k$ 线性无关，有 $\beta_1 - \gamma_1 = \beta_k - \gamma_k = 0$ 。 ■

5.2 basis

Definition 5.2.1 — 基 (basis). n 个线性独立的 n 维向量 $a_1, \ldots, a_n$ 的集合

Definition 5.2.2 — 向量 b 在基底 $a_1, \ldots, a_n$ 下的分解. 任何一个 n 维向量 b 都可以用它们的线性组合来表示

$$b = \beta_1 a_1 + \cdots + \beta_n a_n$$

Proof. 同一向量的系数是唯一的。 ■

■ **Example 5.3** $e_1, \ldots, e_n$ 是一组基，那么 b 在此基底下的分解为

$$b = b_1 e_1 + \cdots + b_n e_n, b = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} \in \mathbb{R}^n$$

■

5.3 标准正交向量

Definition 5.3.1 — Orthogonal Vectors. 在 n 维向量集 $a_1, \ldots, a_k$ 中，如果对于 $i \neq j$, 都有 $a_i \perp a_j$ ，则称它们相互正交 (orthogonal)。

Definition 5.3.2 — Orthonormal Vectors. 如果 n 维向量集 $a_1, \ldots, a_k$ 相互正交，且每个向量的模长都为单位长度 1 ，即对于 $i = 1, \ldots k$, 有 $\|a_i\|_2^2 = 1$, 则称它们是标准正交 (orthonormal) 的。

$$a_i^T a_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

Corollary 5.3.1 标准正交的向量集是线性无关的。

Corollary 5.3.2 根据线性无关的性质，必有向量集向量个数 $k \leq n$

Definition 5.3.3 — n 维向量的一个标准正交基. 当 $k = n$ 时，$a_1, \ldots, a_n$ 是 n 维向量的一个标准正交基。

Definition 5.3.4 — x 在标准正交基下的标准正交分解. 如果 $a_1, \ldots, a_n$ 是一个标准正交基，对于任意维向量 x;

$$x = \left(a_1^T x\right) a_1 + \cdots + \left(a_n^T x\right) a_n$$

则称其为 x 在标准正交基下的标准正交分解。

这个分解可以用于计算不同标准正交基下的系数。

Proof.

$$a_i^T x = \left(a_1^T x\right) a_i^T a_1 + \cdots + \left(a_i^T x\right) a_i^T a_i + \cdots + \left(a_n^T x\right) a_i^T a_n = a_i^T x$$

■

5.4 Gram-Schmidt Algorithm

Algorithm 2: Gram-Schmidt Algorithm

Input: n 维向量 $a_1, \ldots, a_k$
Output: 若这些向量线性无关时 $q_1, \ldots, q_k$（标准正交基）；若线性相关时判断 a_j 是 $a_1, \ldots, a_{j-1}$ 的线性组合

1 $q_1 = a_1 / \|a_1\|_2$;
2 **while** $i = 2, \cdots, k$ **do**
3 正交化: $\widetilde{q}_i = a_i - \left(q_1^T a_i\right) q_1 - \cdots - \left(q_{i-1}^T a_i\right) q_{i-1}$;
4 检验线性相关：如果 $\widetilde{q}_i = 0$, 提前退出迭代;
5 单位化: $q_i = \widetilde{q}_i / \|\widetilde{q}_i\|_2$;
6 **end**

如果步骤 2 中未提前结束迭代，那么 $a_1, \ldots, a_k$ 是线性独立的，而且 $q_1, \ldots, q_k$ 是标准正交基。

如果在第 j 次迭代中提前结束，说明 a_j 是 $a_1, \ldots, a_{j-1}$ 的线性组合, 因此 $a_1, \ldots, a_k$ 是线性相关的。

Theorem 5.4.1 $q_1, \ldots, q_{i-1}, q_i$ 是标准正交的。

Proof. 假设第 $i-1$ 次迭代成立，即: $q_r \perp q_s, \forall r, s < i$.
正交化步骤保证有以下关系成立

$$\widetilde{q}_i = a_i - \left(q_1^T a_i\right) q_1 - \cdots - \left(q_{i-1}^T a_i\right) q_{i-1}$$

等式两边同时乘以 $q_j^T, j = 1, \ldots, i-1$

$$\begin{aligned} q_j^T \tilde{q}_i &= q_j^T a_i - \left(q_1^T a_i\right)\left(q_j^T q_1\right) - \cdots - \left(q_{i-1}^T a_i\right)\left(q_j^T q_{i-1}\right) \\ &= q_j^T a_i - q_j^T a_i = 0 \end{aligned}$$

$\because q_j^T q_r = 0, j \neq r, q_j^T q_j = 1$
$\therefore \widetilde{q}_i \perp q_1, \ldots, \widetilde{q}_i \perp q_{i-1}$.
单位化步骤保证了 $q_i = \widetilde{q}_i / \left\|\widetilde{q}_i\right\|_2$, 即 $q_1, \ldots, q_i$ 是标准正交。 ■

Algorithm 3: Gram-Schmidt Algorithm (Another Algorithm)

Input: Three independent vectors $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$
Output: Three orthonormal vectors $\boldsymbol{q}_1 = \boldsymbol{A}/\|\boldsymbol{A}\|, \boldsymbol{q}_2 = \boldsymbol{B}/\|\boldsymbol{B}\|, \boldsymbol{q}_3 = \boldsymbol{C}/\|\boldsymbol{C}\|$.

1 Choose $\boldsymbol{A} = \boldsymbol{a}$;

2 $$\boldsymbol{B} = \boldsymbol{b} - \frac{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{b}}{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}}\boldsymbol{A}$$;

3 $$\boldsymbol{C} = \boldsymbol{c} - \frac{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{c}}{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}}\boldsymbol{A} - \frac{\boldsymbol{B}^{\mathrm{T}}\boldsymbol{c}}{\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}}\boldsymbol{B}$$;

4 单位化;

II Statistics

6. Random Variables

6.1 随机变量的引入

在实际问题中，随机试验的结果可以用数量来表示，由此就产生了随机变量的概念.

有些试验结果本身与数值有关（本身就是一个数）；在有些试验中，试验结果看来与数值无关，但我们可以引进一个变量来表示它的各种结果. 也就是说，把试验结果数值化. 这种对应关系在数学上理解为定义了一种实值单值函数

6.2 随机变量的特点

它随试验结果的不同而取不同的值，因而在试验之前只知道它可能取值的范围，而不能预先肯定它将取哪个值.

由于试验结果的出现具有一定的概率，于是这种实值函数取每个值和每个确定范围内的值也有一定的概率, 所以称这种定义在样本空间 S 上的实值单值函数 $X = X(e)$ 为随机变量

6.3 优势

1. 引入随机变量后，对随机现象统计规律的研究，就由对事件及事件概率的研究扩大为对随机变量及其取值规律的研究
2. 易于表示
3. 和原来集合的表示等价

6.4 常见随机变量

6.4.1 离散型随机变量

公式法

$$\boldsymbol{P}\{\boldsymbol{X}=\boldsymbol{x}_k\}=\boldsymbol{p}_k, \boldsymbol{k}=1,2,\cdots$$

列表法

$$X \sim \begin{pmatrix} x_1 & x_2 & \cdots & x_n & \cdots \\ p_1 & p_2 & \cdots & p_n & \cdots \end{pmatrix}$$

$\boldsymbol{X}$	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\cdots$	$\boldsymbol{x}_n \cdots$
$\boldsymbol{p}_k$	$\boldsymbol{p}_1$	$\boldsymbol{p}_2$	$\cdots$	$\boldsymbol{p}_n \cdots$

6.4.2 两点分布

特殊的二项分布 (n = 1)

X	0	1
p_k	$1-\mathrm{p}$	p

6.4.3 (n 重) 伯努利试验, 二项分布

$$P\{X=k\} = \begin{pmatrix} n \\ k \end{pmatrix} p^k (1-p)^{n-k} \quad k = 0, 1, \cdots, n$$

$\boldsymbol{X}$	$\boldsymbol{0}$	$\boldsymbol{1}$	$\cdots$	$\boldsymbol{k}$	$\cdots$	$\boldsymbol{n}$
$\boldsymbol{p_k}$	$\boldsymbol{q}^n$	$\begin{pmatrix} \boldsymbol{n} \\ \boldsymbol{1} \end{pmatrix} \boldsymbol{pq}^{n-1}$	$\cdots$	$\begin{pmatrix} \boldsymbol{n} \\ \boldsymbol{k} \end{pmatrix} \boldsymbol{p}^k \boldsymbol{q}^{n-k}$	$\cdots$	$\boldsymbol{p}^n$

二项分布性质

- 将伯努利试验 E 独立地重复地进行 n 次, 则称这一串重复的独立试验为 n 重伯努利试验. 伯努利试验对试验结果没有等可能的要求“重复”是指这 n 次试验中 $P(A) = p$ 保持不变.“独立”是指各次试验的结果互不影响.
- 每次试验条件相同
- 每次试验只考虑两个互逆结果 A 或 $\overline{A}$
- 各次试验相互独立
- 二项分布描述的是 n 重伯努利试验中事件 A 出现的次数 X 的分布律

二项分布单峰性质

若在 k_0 处，概率 PX=k 达到最大（称 k_0 为随机变量 X 的最可能值）

$$\begin{cases} \frac{P\{X=k_0\}}{P\{X=k_0+1\}} \geq 1 \\ \frac{P\{X=k_0\}}{P\{X=k_0-1\}} \geq 1 \end{cases}$$

得

$$(n+1)p - 1 \leq k\_0 \leq (n+1)p$$

$$k_0 = \begin{cases} (n+1)p \text{ 和} (n+1)p-1, & \text{当} (\boldsymbol{n}+\boldsymbol{1})\boldsymbol{p} \text{ 为整数,} \\ [(n+1)p], & \text{其它,} \end{cases}$$

6.4.4 泊松分布

$$\boldsymbol{P}\{\boldsymbol{X}=\boldsymbol{k}\} = \frac{\lambda^k \mathrm{e}^{-\lambda}}{\boldsymbol{k}!}, \quad \boldsymbol{k} = \boldsymbol{0}, \boldsymbol{1}, \boldsymbol{2}, \cdots$$

其中 $\lambda > 0$ 是常数, 则称 X 服从参数为 λ 的泊松分布, 记作

$$X \sim \pi(\lambda)$$

Theorem 6.4.1 — 泊松定理. 泊松分布是作为二项分布的近似，于 1837 年由法国数学家泊松引入的.

$$\lim_{n\to\infty} C_n^k p_n^k \left(1-p_n\right)^{n-k} = e^{-\lambda}\frac{\lambda^k}{k!} \quad (\lambda = np)$$

6.4.5 连续型随机变量

6.4.6 均匀分布

$$f(x) = \begin{cases} \frac{1}{b-a}, & a < x < b \\ 0, & \text{其它} \end{cases}$$

$$X \sim U(a,b)$$

$$F(x) = P\{X \le x\} = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \le x < b \\ 1 & x \ge b \end{cases}$$

6.4.7 指数分布

$$f(x) = \begin{cases} \frac{1}{\theta}e^{-x/\theta}, & x > 0 \\ 0, & \text{其它,} \end{cases}$$

$$F(x) = P\{X \le x\} = \begin{cases} 1-e^{-x/\theta}, & x > 0 \\ 0, & \text{其它} \end{cases}$$

6.4.8 正态分布

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

6.4.9 标准正态分布

$$\text{若} X \sim N\left(\mu,\sigma^2\right), \text{则} Z = \frac{X-\mu}{\sigma} \sim N(\mathbf{0},1)$$

6.5 随机变量的分布函数

6.6 随机变量的分布函数性质

对任意实数 $x_1 < x_2$，随机点落在区间 $(x_1, x_2]$ 内的概率为:

$$P\{x_1 < X \le x_2\} = P\{X \le x_2\} - P\{X \le x_1\} = F(x_2) - F(x_1)$$

$$F(-\infty) = \lim_{x\to-\infty} F(x) = 0$$
$$F(+\infty) = \lim_{x\to+\infty} F(x) = 1$$

6.7 连续型随机变量及其概率密度

$$f(x) \ge 0$$

$$\int\_{-\infty}^{\infty} f(x)dx = 1$$

6.8 随机变量的数字特征

6.8.1 方差

$$D(X) = E\Big\{[x - E(X)]\Big\}$$

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

方差性质

$$D(C) = 0$$

$$D(CX) = C^2 D(X)$$

$$D(X + C) = D(X)$$

X 和 Y 相互独立时,

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

D(X) = 0 等价于 $P[X = E(X)] = 1$

Theorem 6.8.1 — 切比雪夫不等式.

$$P(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

$$P(|X - \mu| < \varepsilon) \geq 1 - \frac{\sigma^2}{\varepsilon^2}$$

6.8.2 协方差

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

$$Cov(X, Y) = E(XY) - E(X)E(Y)$$

协方差性质

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

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

$$Cov(X, X) = D(X)$$

$$Cov(X, c) = 0$$

$$Cov(aX + b, Y) = aCov(X, Y)$$

$$Cov(aX, bY) = abCov(X, Y)$$

$$Cov(X_1 + X_2, Y) = Cov(X_1, Y) + Cov(X_2, Y)$$

对于随机变量序列 $X_1, \ldots, X_n$ 与 $Y_1, \ldots, Y_m$, 有

$$\operatorname{cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \operatorname{cov}\left(X_i, Y_j\right)$$

对于随机变量序列 $X_1, \ldots, X_n$, 有

$$\operatorname{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{var}\left(X_i\right) + 2\sum_{i,j:i<j} \operatorname{cov}\left(X_i, X_j\right)$$

6.8.3 相关系数

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

相关系数是刻划两个变量间线性相关程度的一个重要的数字特征. 相关系数也可以看成协方差：一种剔除了两个变量量纲影响、标准化后的特殊协方差。

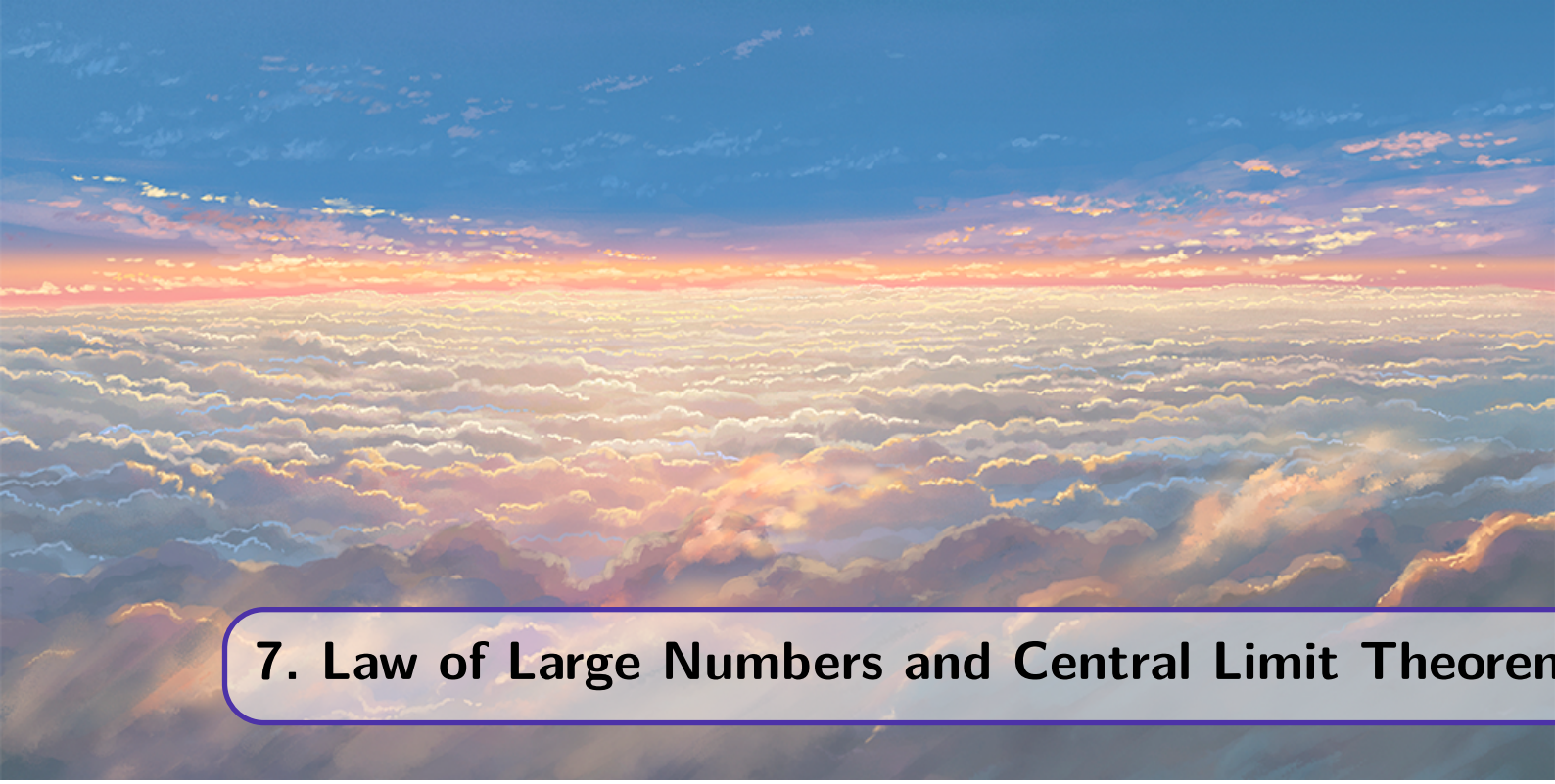
6.8.4 X, Y 不相关时候的性质

若 $\rho_{XY} = 0$, 称 X 和 Y 不 (线性) 相关。

Corollary 6.8.2 独立一定是不相关，不相关不一定独立。

Theorem 6.8.3 若随机变量 X 与 Y 的方差都存在，且均不为零; 则下列四个命题等价。

1. $\rho_{XY} = 0$
2. $\operatorname{Cov}(X, Y) = 0$
3. $E(XY) = E(X)E(Y)$
4. $D(X \pm Y) = DX + DY$。



7. Law of Large Numbers and Central Limit Theorem

7.1 辛钦大数定律

Theorem 7.1.1 — 辛钦大数定律. 设随机变量序列 $X_1, X_2, \cdots$ 相互独立，服从同一分布，具有数学期 $E(X_i) = \mu, i = 1, 2, \cdots$，则对于任意正数 ε，有

$$\lim_{n\to\infty} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| < \varepsilon\right\} = 1$$

则序列 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 依概率收敛于 μ（很可能接近于 μ）.

另一表述:

$$\overline{X} \xrightarrow{P} \mu$$

7.1.1 辛钦大数定律条件

- $X_1, X_2, \cdots$ 相互独立
- $X_1, X_2, \cdots$ 服从同一分布
- 不要求方差存在。

Corollary 7.1.2 设 $X_n \xrightarrow{P} a$, $Y_n \xrightarrow{P} b$ 则

$$g(X_n, Y_n) \xrightarrow{P} g(a, b)$$

7.2 伯努利大数定理

Theorem 7.2.1 — 伯努利大数定理. 设 f_A 是 n 次独立重复试验中事件 A 的发生次数, p 是每次试验中发生的概率, 则对于任意的正数 ε.

$$\lim_{n\to\infty}\{|f_A/n-p|<\varepsilon\}=1$$

贝努里大数定律表明，当重复试验次数 n 充分大时，事件 A 发生的频率 nA/n 与事件 A 的概率 p 有较大偏差的概率很小.

7.3 中心极限定理

Theorem 7.3.1 — Lyapunov 中心极限定理. 设 $X_1, X_2, \ldots$ 相互独立, 他们拥有数学期望和方差

$$E(X_k)=\mu_k$$

$$D(X_k)=\sigma_k^2$$

则有

$$\begin{aligned}\lim_{n\to\infty}F_n(x)&=\lim_{n\to\infty}P\left\{\frac{\sum_{k=1}^n X_k-\sum_{k=1}^n\mu_k}{B_n}\le x\right\}\\&=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\mathrm{t}^2/2}\,\mathrm{dt}\\&=\Phi(x)\end{aligned}$$

其中 $B_n=\sqrt{D\left(\sum_{k=1}^n X_k\right)}$

Theorem 7.3.2 设 $X_1, X_2, \ldots$ 相互独立, 他们服从同一分布, 拥有数学期望和方差

$$E(X_k)=\mu$$

$$D(X_k)=\sigma^2$$

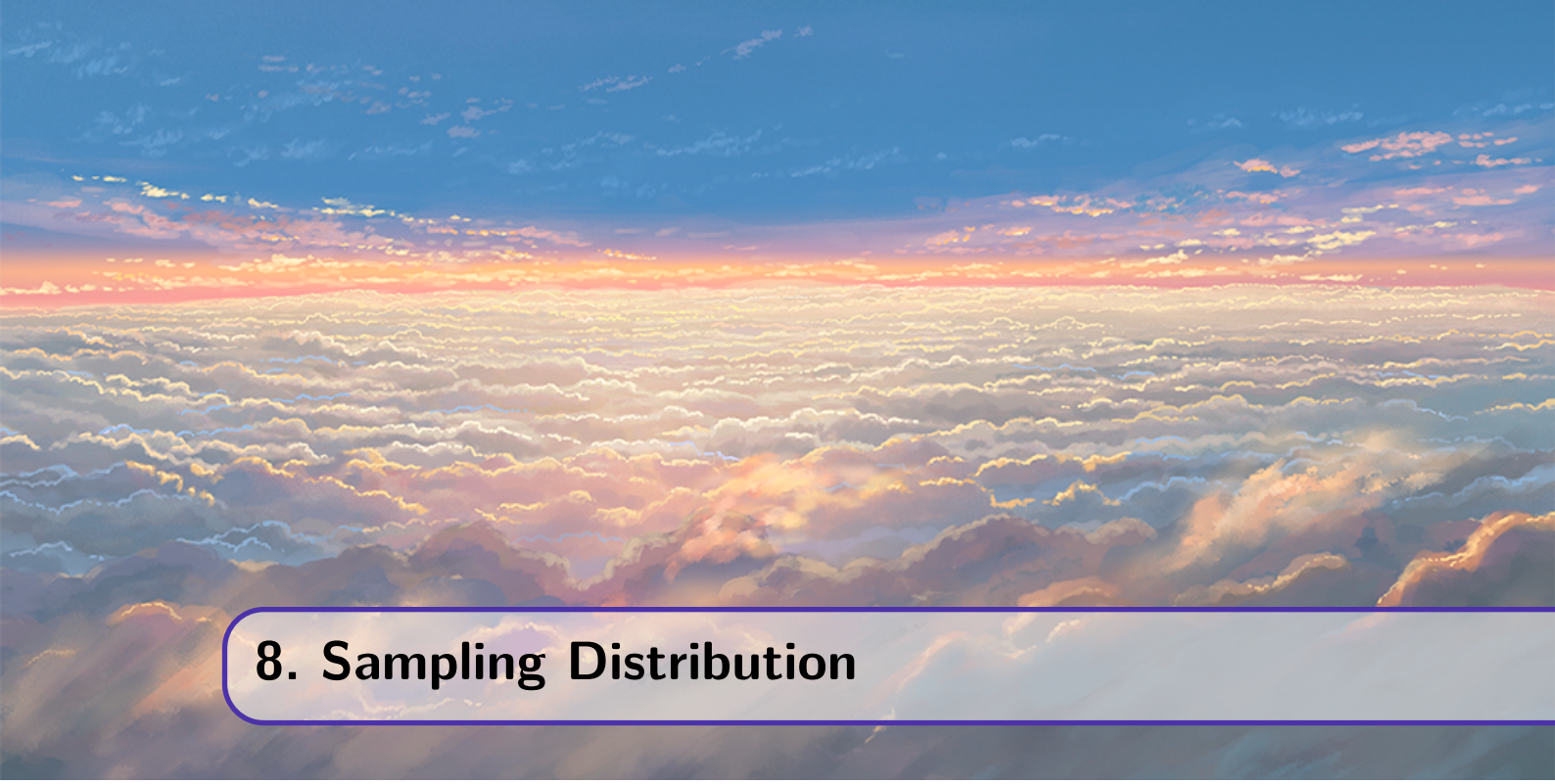
则有

$$\begin{aligned}\lim_{n\to\infty}F_n(x)&=\lim_{n\to\infty}P\left\{\frac{\sum_{i=1}^n X_i-n\mu}{\sigma\sqrt{n}}\le x\right\}\\&=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\mathrm{t}^2/2}\,\mathrm{dt}\\&=\Phi(x)\end{aligned}$$

Theorem 7.3.3 — De Moivre-Laplace 定理. 设 η_n 服从参数为 n,p 的二项分布, 则对于任意 x, 有

$$\lim_{n\to\infty}P\left\{\frac{\eta_n-np}{\sqrt{np(1-p)}}\le x\right\}=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt=\Phi(x)$$

Proof. 将 η_n 拆分成 n 个相互独立, 服从同一分布的随机变量 $X_1, X_2, \ldots$ 之和 ■



8. Sampling Distribution

8.1 统计量

不含任何未知参数的样本的函数称为统计量. 它是完全由样本决定的量.

8.2 常见统计量

Definition 8.2.1 — 样本平均值.

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

Definition 8.2.2 — 样本方差.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

Definition 8.2.3 — 样本标准差.

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}$$

Definition 8.2.4 — 样本 k 阶原点矩.

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

Definition 8.2.5 — 样本 k 阶中心矩.

$$\boldsymbol{B_k} = \frac{1}{\boldsymbol{n}} \sum_{i=1}^{n} \left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)^k$$

Definition 8.2.6 — X 和 Y 的 $k+p$ 阶混合原点矩. 若

$$E\left\{\left(X^k\right)\left(Y^p\right)\right\}, k, p = 1, 2, \ldots$$

存在, 则称它为 X 和 Y 的 $k+p$ 阶混合原点矩

Definition 8.2.7 — X 和 Y 的 $k+l$ 阶混合中心矩. 若

$$E\left\{[X - E(X)]^k [Y - E(Y)]^l\right\}, k, l = 1, \quad 2, \ldots$$

存在, 则称它为 X 和 Y 的 $k+l$ 阶混合中心矩

8.3 卡方分布

Definition 8.3.1 — 卡方分布. 设 $X_1, X_2, ...X_n$ 相互独立, 都服从正态分布 $N(0,1)$ (都是来自总体 $N(0,1)$ 的样本), 则称随机变量:

$$\chi^2 = X_1^2 + X_2^2 + ... + X_n^2$$

所服从的分布为自由度为 n 的 χ^2 分布. 记作 $\chi^2 \sim \chi^2(n)$

8.3.1 卡方分布性质

Corollary 8.3.1 $\chi^2(1) \sim X^2(1)$

Corollary 8.3.2 — χ^2 分布的可加性. 若 $X_1 \sim \chi^2(n_1), X_2 \sim \chi^2(n_2)$, 且 X_1 与 X_2 相互独立，则

$$X_1 + X_2 \sim \chi^2(n_1 + n_2)$$

Proof. 事实上, 卡方分布是 Gamma 分布的特殊情况。自由度为 n 的卡方分布 χ_n^2 其实就是 $\Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$.

根据 Gamma 分布的矩生成函数（Moment Generating Function, MGF), 若 $X_1 \sim \chi_{n_1}^2$, $X_2 \sim \chi_{n_2}^2$,

那么 X_1, X_2 对应的 MGF 分别为

$$M_{X_1}(t) = (1-2t)^{-n_1/2}$$

$$M_{X_2}(t) = (1-2t)^{-n_2/2}$$

因为 X_1, X_2 相互独立, 那么 $Y = X_1 + X_2$ 的 MGF 为

$$M_Y(t) = M_{X_1}(t) M_{X_2}(t) = (1-2t)^{-(n_1+n_2)/2}$$

由 MGF 的唯一性,

$$Y \sim \Gamma\left(\frac{n_1+n_2}{2}, \frac{1}{2}\right) = \chi_{n_1+n_2}^2$$

■

Corollary 8.3.3 — χ^2 分布的数学期望和方差.

$$E(\chi^2) = n$$

$$D(\chi^2) = 2n$$

Proof.

$$E(X) = 0 D(X) = 1 E(X_i^2) = 1 D(X_i^2) = 2$$

■

Corollary 8.3.4 — 卡方分布中心极限定理. 设 $X_1, X_2, ...X_n$ 相互独立, 都服从正态分布 N(0,1), $E(X) = \mu$, $D(X) = \sigma^2$, 则有

$$P(\frac{\chi^2(n) - n}{\sqrt{2n}} \leq x) \sim \Phi(x)$$

Proof.

$$E(X) = 0 D(X) = 1 E(X_i^2) = 1 D(X_i^2) = 2$$

■

Corollary 8.3.5

$$\frac{\chi^2(n)}{n} \leftarrow \frac{1}{n} \sum_{i=1}^{n} X_i^2 = 1$$

Proof.

$$E(X_i^2) = 1$$

■

Definition 8.3.2 — 卡方分布的上分位点.

$$p(\chi^2 > \chi_\alpha^2(n)) = \alpha$$

8.4 t 分布

Definition 8.4.1 — 自由度为 t 的 t 分布. 设 $X \sim N(0,1), Y \sim \chi^2(n)$, 且 X,Y 相互独立, 则称随机变量

$$t = \frac{X}{\sqrt{Y/n}}$$

为自由度为 n 的 t 分布, 记作 $t \sim t(n)$

8.4.1 t 分布性质

Corollary 8.4.1 — t 分布数学期望和方差.

$$E(t) = 0$$

$$D\big(t(n)\big) = \frac{n}{n-2}$$

Corollary 8.4.2 — t 分布的概率密度函数. $n = 1$ 时

$$f(t) = \frac{1}{\pi(1+t^2)} \text{(柯西密度)}$$

数学期望不存在. $n > 1$ 时, $h(t)$ 的图形关于 $t = 0$ 对称.

Corollary 8.4.3 n 足够大时 t 分布近似于 $N(0,1)$ 分布. 由于卡方分布

$$n \to \infty \text{时}, \frac{\chi^2(n)}{n} \to 1$$

所以

$$t(n) = \frac{\mathrm{N(0,1)}}{\sqrt{\frac{\chi^2(n)}{n}}} \xrightarrow{n\to\infty} \mathrm{N(0,1)}$$

Corollary 8.4.4 — 与 F 分布的关系.

$$\begin{aligned} t^2(n) &= \frac{N(0,1)^2}{\chi^2(n)/n} \\ &= \frac{\chi^2(1)/1}{\chi^2(n)/n} \sim F(1,n) \end{aligned}$$

$$\frac{1}{t^2(n)} \sim F(n,1)$$

Definition 8.4.2 — t 分布的上分位点.

$$t_{1-\alpha}(n) = -t_\alpha(n)$$

8.5 F 分布

Definition 8.5.1 — F 分布. 设 $U \sim \chi^2(n_1), V \sim \chi^2(n_2)$, 且 U, V 相互独立, 则称随机变量

$$F = \frac{U/n_1}{V/n_2}$$

服从自由度为 n_1, n_2 的 F 分布. 记作 $F \sim F(n_1, n_2)$

8.5.1 F 分布性质

Corollary 8.5.1 — F 分布数学期望.

$$E(F) = \frac{n_2}{n_2 - 2}$$

即与 n_1 无关.

Corollary 8.5.2 — $F(n_1,n_2),F(n_2,n_1)$ 的关系.

$$\frac{1}{F(n_1,n_2)} \sim F(n_2,n_1)$$

Corollary 8.5.3 — 与 t 分布的关系.

$$\begin{aligned} t^2(n) &= \frac{N(0,1)^2}{\chi^2(n)/n} \\ &= \frac{\chi^2(1)/1}{\chi^2(n)/n} \sim F(1,n) \end{aligned}$$

$$\frac{1}{t^2(n)} \sim F(n,1)$$

Corollary 8.5.4 — F 分布上　分位点的性质.

$$F_{1-\alpha}(n_1,n_2) = \frac{1}{F_\alpha(n_2,n_1)}$$

8.6 正态总体的样本均值和样本方差的分布

Definition 8.6.1 — 正态总体的样本均值的数学期望、方差和样本方差的数学期望. 设总体 $X \sim N(\mu,\sigma^2)$ 的均值为 μ, 方差为 ^{2}.

$X_1,X_2,...X_n$ 是来自 X 的一个样本, 样本均值是 $\overline{X}$, 样本方差是 S^2, 则有

$$E(\overline{X}) = \mu$$

$$D(\overline{X}) = \frac{\sigma^2}{n}$$

$$E(S^2) = \sigma^2$$

Corollary 8.6.1 — 矩估计法原理.

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

$$\frac{1}{n}\sum X_1^2 - \overline{X}^2 \to D(X) = E(X^2) - E(X)^2$$

Theorem 8.6.2 — 样本均值的分布. 设总体 X N(μ,σ^2) 的均值为 μ, 方差为 ^{2}, $X_1,X_2,...X_n$ 是来自 X 的一个样本, 样本均值是 $\overline{X}$, 则有

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

Proof.

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + ... \sim N(n\mu, n\sigma^2)$$

■

Corollary 8.6.3

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

Theorem 8.6.4 — 样本方差的分布. 设总体 $X \sim N(\mu, \sigma^2)$ 的均值为 μ, 方差为 ^{2}, $X_1, X_2, ...X_n$ 是来自 X 的一个样本, 样本均值是 $\overline{X}$, 样本方差是 S^2, 则有

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

而且 $\overline{X}$ 与 S^2 相互独立.

Theorem 8.6.5 — 样本均值和样本方差的关系. 设 $X_1, X_2, ...X_n$ 是总体 X N(μ, σ^2) 的样本, 样本均值是 $\overline{X}$, 样本方差是 S^2, 则有

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{\sqrt{n}(\overline{X} - \mu)}{S} \sim t(n-1)$$

对比

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

但是 n 很大时

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

$$S^2 \to E(X^2) - E(X)^2 = \sigma^2$$

$$\therefore S \to \sigma$$

Theorem 8.6.6 — 两总体样本均值差、样本方差比的分布. 设 $X \sim N\left(\mu_1, \sigma_1^2\right)$, $Y \sim N\left(\mu_2, \sigma_2^2\right)$, 且 X 与 Y 独立,

$X_1, X_2, \ldots, X_n$ 是来自 X 的样本, $Y_1, Y_2, \ldots, Y_{n_2}$ 是取自 Y 的样本,

$\overline{\boldsymbol{X}}$ 和 $\overline{\boldsymbol{Y}}$ 分别是这两个样本的样本均值, $\boldsymbol{S}_1^2$ 和 $\boldsymbol{S}_2^2$ 分别是这两个样本的样本方差, 则有

1. $\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2}$ 的分布

$$\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$

2.

$$\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{\sigma^2(n_1+n_2-2)}}}\sim t(n_1+n_2-2)$$

Proof. 1. $\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2}$ 的分布

$$\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2}\sim F(n_1-1,n_2-1)$$

当 $\sigma_1^2=\sigma_2^2=\sigma^2$ 时

$$\overline{X}-\overline{Y}=N(\mu_1-\mu_2,\sigma^2/n_1+\sigma^2/n_2)$$

$$\therefore U=\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim N(0,1)$$

又因为

$$\frac{n_1-1}{\sigma^2}S_1^2\sim\chi^2(n_1-1)$$

$$\frac{n_2-1}{\sigma^2}S_2^2\sim\chi^2(n_2-1)$$

$$\therefore V=\frac{n_1-1}{\sigma^2}S_1^2+\frac{n_2-1}{\sigma^2}S_2^2\sim\chi^2(n_1+n_2-2)$$

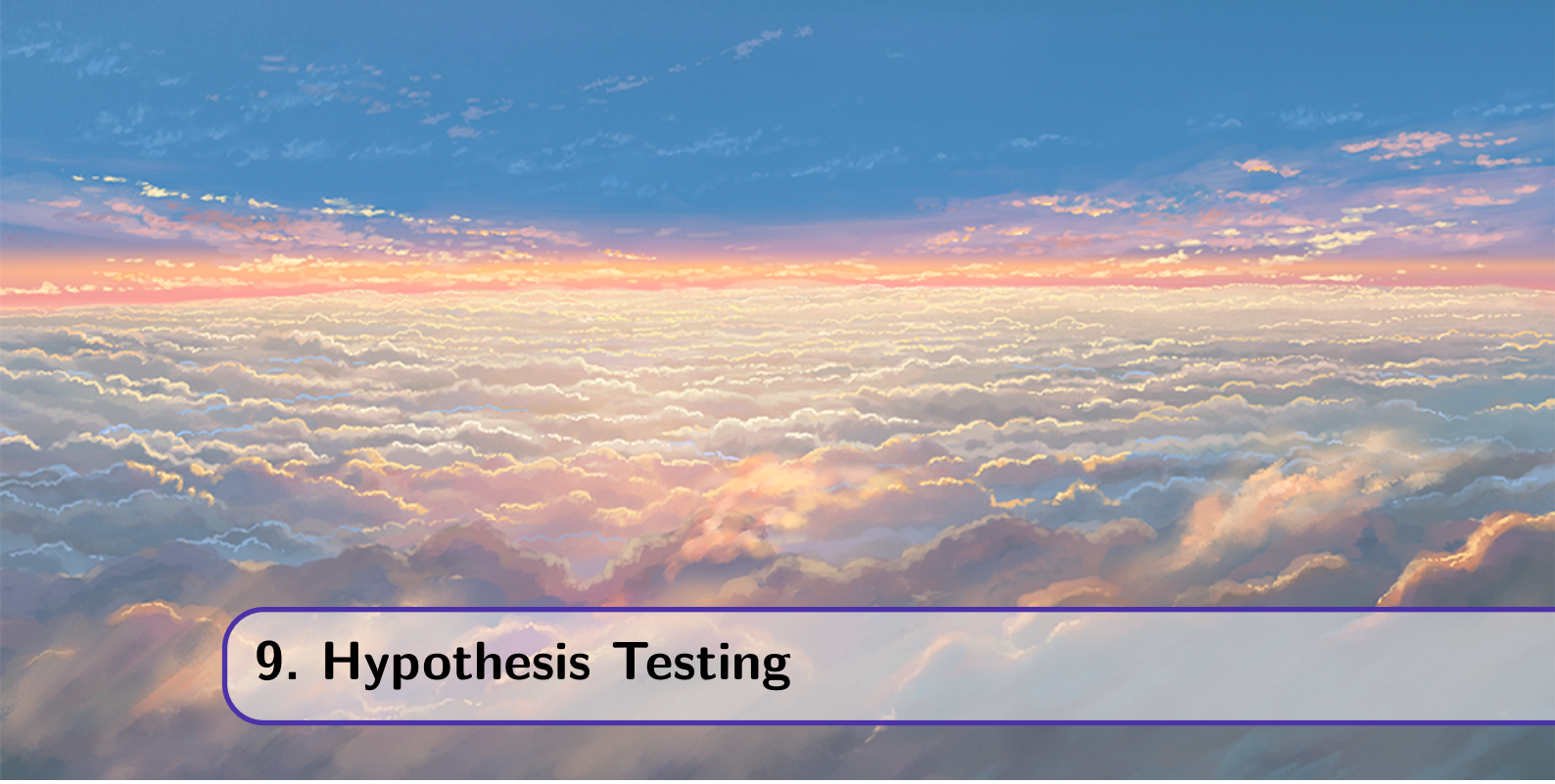
2 和 3 相互独立, 根据 t 分布定义, 因此有

$$\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t(n_1+n_2-2)$$

代入 $S_w^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1+n_2-2)}$

$$\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{\sigma^2(n_1+n_2-2)}}}\sim t(n_1+n_2-2)$$

■



9. Hypothesis Testing

9.1 正态总体均值方差的检验法

假设显著性水平为 α。参阅9.1.

9.2 经验分布函数

设 $X_1, X_2, \cdots, X_n$ 是总体 $\boldsymbol{F}$ 的一个样本, 用 $S(\boldsymbol{x})$ $(-\infty < x < \infty)$ 表示 $X_1, X_2, \cdots, X_n$ 中不大于 x 的随机变量

$$F_n(x) = \frac{1}{n} S(x), -\infty < x < \infty.$$

容易得到的 ($\boldsymbol{F}_n(\boldsymbol{x})$ 的观察值仍以 $\boldsymbol{F}_n(\boldsymbol{x})$ 表示)。一般地, 设 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$ 是总体 $\boldsymbol{F}$ 的一个容量为 $\boldsymbol{n}$ 的样本值先将 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$ 按自小到大的次序排列，并重新编号。

设为

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

则经验分布函数 $F_n(x)$ 的观察值为

$$F_n(x) = \begin{cases} 0, & \text{若} x < x_{(1)} \\ \frac{k}{n}, & \text{若} x_{(k)} \leq x < x_{(k+1)} \\ 1, & \text{若} x \geq x_{(n)} \end{cases}$$

经验分布函数的任一个观察值 $F_n(x)$ 与总体分布函数 $F(x)$ 只有微小的差别，从而在实际上可当作 $F(x)$ 来使用。

9.3 Q-Q 图 (Quantile-quantile Plot)

Q-Q 图是 Quantile-Quantile Plot 的简称，是检验拟合优度的好方法，目前在国外被广泛使用，它的图示方法简单直观，易于使用。

现在我们希望知道观测数据与分布模型的拟合效果如何。如果拟合效果好，观测数据的经验分布就应当非常接近分布模型的理论分布，而经验分布函数的分位数自然也应当与分布模型的理论分位数近似相等。

Table 9.1: 正态总体均值方差的检验法

	原假设 H_0	检验统计量
1	$\mu \le \mu_0, \mu \ge \mu_0, \mu = \mu_0$ $(\sigma^2$ 已知 $)$	$Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$
2	$\mu \le \mu_0, \mu \ge \mu_0, \mu = \mu_0$ $(\sigma^2$ 果知$)$	$t = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$
3	$\mu_1 - \mu_2 \le \delta, \mu_1 - \mu_2 \ge \delta, \mu_1 - \mu_2 = \delta$ $(\sigma_1^2, \sigma_2^2$ 已知 $)$	$Z = \frac{\bar{X}-\bar{Y}-\delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$
4	$\mu_1 - \mu_2 \le \delta, \mu_1 - \mu_2 \ge \delta, \mu_1 - \mu_2 = \delta$ $(\sigma_1^2 = \sigma_2^2 = \sigma^2$ 未知 $)$	$t = \frac{\bar{X}-\bar{Y}-\delta}{S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}, S_w^2 = \frac{(n_1-1)S_1^2+(n_2-2)S_2^2}{n_1+n_2-2}$
5	$\sigma^2 \le \sigma_0^2, \sigma^2 \ge \sigma_0^2, \sigma^2 = \sigma_0^2$ $(\mu$ 未知 $)$	$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$
6	$\sigma_1^2 \le \sigma_2^2, \sigma_1^2 \ge \sigma_2^2, \sigma_1^2 = \sigma_2^2$ $(\mu_1, \mu_2$ 未知 $)$	$F = \frac{S_1^2}{S_2^2}$
7	$\mu_D \le 0, \mu_D \ge 0, \mu_D = 0$ (成对数据)	$t = \frac{\bar{D}-0}{S_D/\sqrt{n}}$

Algorithm 4: 作 Q-Q 图

Input: 观测数据 $x_1, x_2, \cdots, x_n$

1 将 $x_1, x_2, \cdots, x_n$ 依大小顺序排列成：$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$;

2 取 $y_i = F^{-1}((i-1/2)/n)$，$i = 1, 2, \cdots, n$;

3 将 $(y_i, x_{(i)})$，$i = 1, 2, \cdots, n$，这 n 个点画在直角坐标图上;

4 如果这 n 个点看起来呈一条 45° 角的直线，从 $(0, 0)$ 到 $(1, 1)$ 分布，我们就相信 $x_1, x_2, \cdots, x_n$ 拟合分布 $F(x)$ 的效果很好。

9.4 χ^2 拟合优度检验

可按照下面的五个步骤进行检验：

Algorithm 5: χ^2 拟合优度检验

1 建立待检假设 H_0：总体 X 的分布函数为 $F(x)$;
2 在数轴上选取 $k-1$ 个分点 $t_1, t_2, \cdots, t_{k-1}$，将数轴分成 k 个区间：$(-\infty, t_1)$，$[t_1, t_2)$，$\cdots$，$[t_{k-2}, t_{k-1})$，$[t_{k-1}, +\infty)$，令 p_i 为分布函数 $F(x)$ 的总体 X 在第 i 个区间内取值的概率，设 m_i 为 n 个样本观察值中落入第 i 个区间上的个数，也称为组频数;
3 选取统计量 $\chi^2 = \sum_{i=1}^{k} \frac{(m_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{m_i^2}{np_i} - n$，如果 H_0 为真，则 $\chi^2 \sim \chi^2(k-1-r)$，其中 r 为分布函数 $F(x)$ 中未知参数的个数;
4 对于给定的显著性水平 α, 确定 χ^2_α, 使其满足 $P\{\chi^2(k-1-r) > \chi^2_\alpha\} = \alpha$。;
5 依据样本计算统计量 χ^2 的观察值, 作出判为总体 X 的分布函数为 $F(x)$;

9.5 柯尔莫哥洛夫 (Kolmogorov-Smirnov) 检验

χ^2 拟合优度检验实际上是检验 $p_i = F_0(a_i) - F_0(a_{i-1}) = p_{i0}$（$i = 1, 2, \cdots, k$）的正确性，并未直接检验原假设的分布函数 $F_0(x)$ 的正确性，柯尔莫哥洛夫检验直接针对原假设 $H_0 : F(x) = F_0(x)$，这里分布函数 $F_0(x)$ 必须是连续型分布。柯尔莫哥洛夫检验基于经验分布函数 (或称样本分布函数) 作为检验统计量，检验理论分布函数与样本分布函数的拟合优度。

设总体 X 服从连续分布，$X_1, X_2, \cdots, X_n$ 是来自总体 X 的简单随机样本，F_n 为经验分布函数。当 H_0 为真时，根据大数定律，当 n 趋于无穷大时，经验分布函数 $F_n(x)$ 依概率收敛总体分布函数 $F_0(x)$。定义 $F_n(x)$ 到 $F_0(x)$ 的距离为

$$D_n = sup_{-\infty<x<+\infty} |F_n(x) - F_0(x)|$$

，当 n 趋于无穷大时，D_n 依概率收敛到 0。

Theorem 9.5.1 — Kolmogorov 定理. 在 $F_0(x)$ 为连续分布的假定下，当原假设为真时，$\sqrt{n}D_n$ 的极限分布为

$$\lim_{n\to\infty} P\{\sqrt{n}D_n \le t\} = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2t^2}, t > 0$$

在显著性水平 α 下，一个合理的检验是：如果 $\sqrt{n}D_n > k$，则拒绝原假设，其中 k 是合适的常数。

注：对于固定的 α 值，我们需要知道该 α 值下检验的临界值。常用的是在统计量为 D_n 时，各个 α 值所对应的临界值如下：在 $\alpha = 0.1$ 的显著性水平下，检验的临界值是 $1.22/\sqrt{n}$；在 $\alpha = 0.05$ 的显著性水平下，检验的临界值是 $1.36/\sqrt{n}$；在 $\alpha = 0.01$ 的显著性水平下，检验的临界值是 $1.63/\sqrt{n}$。这里 n 为样本的个数。当由样本计算出来的 D_n 值小于临界值时，说明不能拒绝原假设，所假设的分布是可以接受的；当由样本计算出来的 D_n 值大于临界值时，拒绝原假设，即所假设的分布是不能接受的。

9.6 秩和检验

秩和检验的依据是，如果两总体分布无显著差异，那么 T 不应太大或太小，以 T_1 和 T_2 为上、下界的话，则 T 应在这两者之间，如果 T 太大或太小，则认为两总体的分布有显著

Algorithm 6: 柯尔莫哥洛夫检验

（1）原假设和备择假设

$$H_0 : F(x) = F_0(x)，\ H_1 : F(x) \neq F_0(x)$$

.

（2）选取检验统计量

$$D_n = \sup_{-\infty<x<+\infty} |F_n(x) - F_0(x)|$$

，当 H_0 为真时，D_n 有偏小趋势，则拟合得越好；当 H_0 不真时，D_n 有偏大趋势，则拟合得越差。

（3）确定拒绝域给定显著性水平 α，查 D_n 极限分布表，求出 t_α 满足 $P\sqrt{n}D_n \geq t_\alpha = \alpha$，作为临界值，即拒绝域为 $[t_\alpha, +\infty)$。

（4）作判断计算统计量的观察值，如果检验统计量 $\sqrt{n}D_n$ 的观察值落在拒绝域中，则拒绝原假设，否则不拒绝原假设。

Algorithm 7: 秩和检验

Input: 设分别从 X、Y 两总体中独立抽取大小为 n_1 和 n_2 的样本, 设 $\boldsymbol{n_1} \leq \boldsymbol{n_2}$

将两个样本混合起来，按照数值大小统一编序由小到大，每个数据对应的序数称为秩。;

计算取自总体 X 的样本所对应的秩之和，用 $\boldsymbol{T}$ 表示;

根据 $\boldsymbol{n}_1, \boldsymbol{n}_2$ 与水平 α, 查秩和检验表, 得秩和下限 $\boldsymbol{T}_1$ 与上限 $\boldsymbol{T_2}$;

两总体分布有显著差异。否则认为 X、Y 两总体分布在水平 α 下无显著差异;

Table 9.2: 所有试验数据

A_1	X_{11}	X_{12}	$\dots$	X_{1n_1}
A_2	X_{21}	X_{22}	$\dots$	X_{2n_2}
$\cdots$	$\dots$	$\cdots$		$\cdots$
$\ddot{A}_r$	X_{r1}	X_{r2}	$\dots$	X_{rn_r}

差异。

9.7 方差分析 (Analysis of Variance, ANOVA)

在现实问题中，经常会遇到类似考察两台机床生产的零件尺寸是否相等，病人和正常人的某个生理指标是否一样，采用两种不同的治疗方案对同一类病人的治疗效果比较等问题。这类问题通常会归纳为检验两个不同总体的均值是否相等，对这类问题的解决可以采用两个总体的均值检验方法。但若检验总体多于两个，仍采用多总体均值检验方法会遇到困难。

9.7.1 单因素方差分析

只考虑一个因素 A 所关心的指标的影响，A 取几个水平在每个水平上作若干个试验，假定试验过程中除因素自身外其他影响指标的因素都保持不变（只有随机因素存在）。我们的任务是从试验结果推断，因素 A 对指标有无显著影响，即当 A 取不同水平时指标有无显著差异 A 取某个水平下的指标视为随机变量，判断 A 取不同水平时指标有无显著差别，相当于检验若干总体的均值是否相等。

不妨设 A 取 r 个水平, 分别记为 $A_1, A_2, \cdots, A_r$。若在水平 A_i 下总体 $X_i \sim N\left(\mu_i, \sigma^2\right), i = 1, 2, \cdots, r$, 这里 μ_i, σ^2 未知, μ_i 可以互不相同，但假定 X_i 有相同的方差。设在水平 A_i 下作了 n n_i 次独立试验，即从总体 X_i 中抽取样本容量为 n_i 的样本, 记作

$$X_{ij}, j = 1, 2, \cdots, n_i$$

其中, $X_{ij} \sim N\left(\mu_i, \sigma^2\right), i = 1, 2, \cdots, r, j = 1, 2, \cdots, n_i$, 且相互独立。

将所有试验数据列成表格

表 table 9.2中对应 A_i 行的数据称为第 i 组数据。判断 $\boldsymbol{A}$ 的 $\boldsymbol{r}$ 个水平对指标有无显著影响，相当于作以下的假设检验:

原假设 $H_0 : \mu_1 = \mu_2 = \cdots = \mu_r$; 备择假设 $\boldsymbol{H}_1 : \mu_1, \mu_2, \cdots, \mu_r$ 不全相等。

由于 X_{ij} 的取值既受不同水平 A_i 的影响, 又受 A_i 固定下随机因素的影响，所以将它分解为

$$X_{ij} = \mu_i + \varepsilon_{ij}, i = 1, 2, \cdots, r, j = 1, 2, \cdots, n_i$$

其中 $\varepsilon_{ij} \sim N\left(\mathbf{0}, \sigma^2\right)$, 且相互独立。引入记号

$$\mu = \frac{1}{n}\sum_{i=1}^{r} n_i\mu_i, \quad n = \sum_{i=1}^{r} n_i, \quad \alpha_i = \mu_i - \mu, \quad i = 1, \cdots, r$$

称 μ 为总均值, α_i 是水平 A_i 下总体的平均值 μ_i 与总评均值 μ 的差异，习惯上称为指标 $\boldsymbol{A}_i$ 的效应。

Definition 9.7.1 原假设 $H_0 : \mu_1 = \mu_2 = \cdots = \mu_r$; 备择假设 $\boldsymbol{H}_1 : \mu_1, \mu_2, \cdots, \mu_r$ 不全相等。

为检验 $\boldsymbol{H}_0$, 给定显著性水平 α, 记

$$\boldsymbol{F}=\frac{S_A/(r-1)}{S_E/(n-r)}\sim F(r-1,n-r)$$

分布的上 α 分位数为 $\boldsymbol{F}_\alpha(\boldsymbol{r-1},\boldsymbol{n-r})$, 检验规则为
$\boldsymbol{F}<\boldsymbol{F}_\alpha(\boldsymbol{r-1},\boldsymbol{n-r})$ 时接受 $\boldsymbol{H_0}$, 否则拒绝。

Proof. 由 section 9.7.1式, 模型可表为

$$\begin{cases}\boldsymbol{X}_{ij}=\mu+\alpha_i+\varepsilon_{ij}\\ \sum_{i=1}^r \boldsymbol{n}_i\alpha_i=\boldsymbol{0}\\ \varepsilon_{ij}\sim N\left(\boldsymbol{0},\sigma^2\right),i=1,\cdots,r,j=1,\cdots,n_i\end{cases}$$

原假设是

$$H_0:\alpha_1=\alpha_2=\cdots=\alpha_r=0$$

记

$$\bar{X}_{i\cdot}=\frac{1}{n_i}\sum_{j=1}^{n_i}X_{ij},\bar{X}=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i}X_{ij}$$

$\overline{\boldsymbol{X}}_{\mathrm{i}\cdot}$ 是第 $\boldsymbol{i}$ 组数据的组平均值, $\overline{\boldsymbol{X}}$ 是全体数据的总平均值。考察全体数据对 $\bar{X}$ 的偏差平方和

$$S_T=\sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}\right)^2$$

$$S_T=\sum_{i=1}^{r}n_i\left(\bar{X}_{i\cdot}-\bar{X}\right)^2+\sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_{i\cdot}\right)^2.$$

$$\text{记}S_A=\sum_{i=1}^{r}n_i\left(\bar{X}_{i\cdot}-\bar{X}\right)^2,$$

$$S_E=\sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_{i\cdot}\right)^2,$$

$$\text{则}S_T=S_A+S_E,$$

$\boldsymbol{S_A}$ 是各组均值对总平均值的偏差平方和，反映 $\boldsymbol{A}$ 不同水平间的差异, 称为组间平方和; $\boldsymbol{S}_E$ 是各组内的数据对样本均值偏差平方和的总和，反映了样本观测值与样本均值的差异，称为组内平方和，而这种差异认为是由随机误差引起的，因此也称为误差平方和。

注意到 $\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_{i0}\right)^2$ 是总体 $N\left(\mu_i,\sigma^2\right)$ 的样本方差的 $\boldsymbol{n_i-1}$ 倍, 于是有

$$\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_{i\cdot}\right)^2/\sigma^2\sim\chi^2\left(n_i-1\right)$$

由 χ^2 分布的可加性知

$$S_E/\sigma^2\sim\chi^2\left(\sum_{i=1}^{r}\left(n_i-1\right)\right),$$

即

$$\boldsymbol{S}_E/\sigma^2\sim\chi^2(\boldsymbol{n-r}),$$

且有

$$ES_E=(n-r)\sigma^2$$

方差来源	离差平方和	自由度	均方	$\boldsymbol{F}$ 值	概率
因素 $\boldsymbol{A}$ (组间)	$\boldsymbol{S}_A$	$\boldsymbol{r-1}$	$\boldsymbol{S}_A/(\boldsymbol{r-1})$	$\boldsymbol{F}=\frac{\boldsymbol{S}_A/(\boldsymbol{r-1})}{\boldsymbol{S}_E/(\boldsymbol{n-r})}\boldsymbol{p}$	
误差 (组内)	$\boldsymbol{S}_E$	$\boldsymbol{n-r}$	$\boldsymbol{S}_E/(\boldsymbol{n-r})$		
总和	$\boldsymbol{S}_T$	$\boldsymbol{n-1}$			

对 $\boldsymbol{S}_A$ 作进一步分析可得

$$ES_A=(r-1)\sigma^2+\sum_{i=1}^{r}n_i\alpha_i^2.(7.10)$$

$$ES_A=(r-1)\sigma^2$$

可知若 $\boldsymbol{H}_0$ 成立, $\boldsymbol{S}_A$ 只反映随机波动, 而若 $\boldsymbol{H}_0$ 不成立, 那它就还反映了 $\boldsymbol{A}$ 的不同水平的效应 α_i。单从数值上

$$\frac{S_A/(r-1)}{S_E/(n-r)}\approx 1$$

该比值服从自由度 $\boldsymbol{n}_1=\boldsymbol{r}-\boldsymbol{1},\boldsymbol{n}_2=(\boldsymbol{n}-\boldsymbol{r})$ 的 $\boldsymbol{F}$ 分布，即

$$F=\frac{S_A/(r-1)}{S_E/(n-r)}\sim F(r-1,n-r)$$

以上对

$$\boldsymbol{S}_A,\boldsymbol{S}_E$$

的分析相当于对组间、组内方差的分析。 ■

若白实验数据算得结果有 $\boldsymbol{F}>\boldsymbol{F}_{\boldsymbol{\alpha}}(\boldsymbol{r}-\boldsymbol{1},\boldsymbol{n}-\boldsymbol{r})$, 则拒绝 $\boldsymbol{H}_0$, 即认为因素 $\boldsymbol{A}$ 对试验结果有显著影响; 若 $\boldsymbol{F}<\boldsymbol{F}_{\alpha}(\boldsymbol{r}-\boldsymbol{1},\boldsymbol{n}-\boldsymbol{r})$, 则接受 $\boldsymbol{H}_{\boldsymbol{0}}$, 即认为因素 $\boldsymbol{A}$ 对试验结果没有显著影响。

$F>F_{0.01}(r-1,n-r)$, 则称因素 A 的影响高度显著。

如果取$\alpha=\boldsymbol{0.01}$ 时拒绝$\boldsymbol{H}_{\boldsymbol{0}}$ 如果取$\alpha=0.05$ 时拒绝$\boldsymbol{H}_0$, 但取$\alpha=0.01$ 时不拒绝 $\boldsymbol{H}_{\boldsymbol{0}}$, 即

$$F_{0.01}(r-1,n-r)\geq F>F_{0.05}(r-1,n-r)$$

则称因素 $\boldsymbol{A}$ 的影响显著。

9.7.2 双因素方差分析方法

如果要考虑两个因素对指标的影响，就要采用双因素方差分析。它的基本思想是：对每个因素各取几个水平然后对各因素不同水平的每个组合作一次或若干次试验对所得数据进行方差分析。对双因素方差分析可分为无重复和等重复试验两种情况，无重复试验只需检验两因素是否分别对指标有显著影响；而对等重复试验还要进步检验两因素是否对指标有显著的交互影响。

设 $\boldsymbol{A}$ 取 $\boldsymbol{s}$ 个水平 $\boldsymbol{A_1},\boldsymbol{A}_2,\cdots,\boldsymbol{A}_s$, $\boldsymbol{B}$ 取 $\boldsymbol{r}$ 个水平 $\boldsymbol{B}_1,\boldsymbol{B}_2,\cdots,\boldsymbol{B}_r$, 在水平组合 $(\boldsymbol{B}_i,\boldsymbol{A}_j)$ 下总体 $\boldsymbol{X}_{ij}$ 服从正态分布 $N\left(\mu_{ij},\sigma^2\right),i=1,\cdots,r,j=1,\cdots,s_{\circ}$ 又设在水平组合 $(\boldsymbol{B}_i,\boldsymbol{A}_j)$ 下作了 $\boldsymbol{t}$ 个试验, 所得结果记作 $\boldsymbol{X}_{ijk},\boldsymbol{X}_{ijk}$ 服从 N $\left(\mu_{ij},\sigma^2\right),i=1,\cdots,r,j=1,\cdots,s,k=1,\cdots,t$, 且相互独立

将 $\boldsymbol{X}_{ijk}$ 分解为

$$X_{ijk}=\mu_{ij}+\varepsilon_{ijk},\quad i=1,\cdots,r,\quad j=1,\cdots,s,\quad k=1,\cdots,t,$$

Table 9.3: 双因素试验数据表

	A_1	A_2	$\cdots$	A_s
$\boldsymbol{B}_1$	$X_{111},\cdots,X_{11t}$	$X_{121},\cdots,X_{12t}$	$\cdots$	$X_{1s1},\cdots,X_{1st}$
B_2	$X_{211},\cdots,X_{21t}$	$X_{221},\cdots,X_{22t}$	$\cdots$	$X_{2s1},\cdots,X_{2st}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
B_r	$X_{r11},\cdots,X_{r1t}$	$X_{r21},\cdots,X_{r2t}$	$\cdots$	$X_{rs1},\cdots,X_{rst}$

其中 $\varepsilon_{ijk}\sim N\left(0,\sigma^2\right)$, 且相互独立。记

$$\begin{array}{ll}\mu=\frac{1}{rs}\sum_{i=1}^{r}\sum_{j=1}^{s}\mu_{ij}, & \\ \mu_{\cdot j}=\frac{1}{r}\sum_{i=1}^{r}\mu_{ij}, & \\ \alpha_j=\mu_{\cdot j}-\mu, & \gamma_{ij}=(\mu_{ij}-\mu)-\alpha_i-\beta_j \\ \mu_{i\cdot}=\frac{1}{s}\sum_{j=1}^{s}\mu_{ij} & \\ ,\beta_i=\mu_{i\cdot}-\mu & \end{array}$$

μ 是总均值, α_j 是水平 A_j 对指标的效应, β_i 是水平 $\boldsymbol{B}_i$ 对指标的效应, γ_{ij} 是水平 B_i 与 A_j 对指标的交互效应。

模型为

$$\begin{cases}\boldsymbol{X}_{ijk}=\mu+\boldsymbol{\alpha}_j+\beta_i+\gamma_{ij}+\varepsilon_{ijk} \\ \sum_{j=1}^{s}\alpha_j=0,\sum_{i=1}^{r}\beta_i=0,\sum_{i=1}^{r}\gamma_{ij}=\sum_{j=1}^{s}\gamma_{ij}=0 \\ \varepsilon_{ijk}\sim N\left(0,\sigma^2\right),i=1,\cdots,r,j=1,\cdots,s,k=1,\cdots,t\end{cases}$$

原假设为

$$\begin{array}{l}H_{01}:\alpha_j=0(j=1,\cdots,s) \\ H_{02}:\beta_i=0(i=1,\cdots,r) \\ H_{03}:\gamma_{ij}=0(i=1,\cdots,r;j=1,\cdots,s).\end{array}$$

9.7.3 无交互影响的双因素方差分析

没有交互影响，每组试验就不必重复，即可令 $\boldsymbol{t}=\mathbf{1}$, 过程大为简化。

$$\mu_{ij}=\mu+\alpha_j+\beta_i,\quad i=1,\cdots,r,j=1,\cdots,s,$$

$$\mu_{ij}=\mu+\alpha_j+\beta_i,\quad i=1,\cdots,r,\quad j=1,\cdots,s$$

$$\begin{cases}X_{ij}=\mu+\alpha_j+\beta_i+\varepsilon_{ij} \\ \sum_{j=1}^{s}\alpha_j=0,\sum_{i=1}^{r}\beta_i=0 \\ \varepsilon_{ij}\sim N\left(0,\sigma^2\right),i=1,\cdots,r,j=1,\cdots,s\end{cases}$$

采用与单因素方差分析模型类似的方法导出检验统计量。记

$$\bar{X}=\frac{1}{rs}\sum_{i=1}^{r}\sum_{j=1}^{s}X_{ij},\bar{X}_{i\cdot}=\frac{1}{s}\sum_{j=1}^{s}X_{ij},\bar{X}_{\cdot j}=\frac{1}{r}\sum_{i=1}^{r}X_{ij}$$
$$S_T=\sum_{i=1}^{r}\sum_{j=1}^{s}\left(X_{ij}-\bar{X}\right)^2$$

其中 S_T 为全部试验数据的总变差, 称为总平方和

方差来源	离差平方和	自由度	均方	F 值
因素 A	S_A	$s-1$	$\frac{S_A}{s-1}$	$F_A=\frac{S_A/(s-1)}{S_E/[(r-1)(s-1)]}$
因素 B	S_B	$r-1$	$\frac{S_B}{r-1}$	$F_B=\frac{S_B/(r-1)}{S_E/[(r-1)(s-1)]}$
误差	S_E	$(r-1)(s-1)$	$\frac{S_E}{(r-1)(s-1)}$	
总和	$\boldsymbol{S_T}$	$\boldsymbol{rs-1}$		

对其进行分解

$$\begin{gathered}S_T=\sum_{i=1}^r\sum_{j=1}^s\left(X_{ij}-\bar{X}\right)^2\\==\sum_{i=1}^r\sum_{j=1}^S\left(X_{ij}-\bar{X}_{i\cdot}-\right.\\\left.\bar{X}_{\cdot j}+\bar{X}\right)^2+S\sum_{i=1}^r\left(\bar{X}_{i\cdot}-\bar{X}\right)^2+\\r\sum_{j=1}^s\left(\bar{X}_{\cdot j}-\bar{X}\right)^2\\=S_E+S_A+S_B\end{gathered}$$

可以验证，在上述平方和分解中交叉项均为 0，其中

$$\begin{gathered}S_E=\sum_{i=1}^r\sum_{j=1}^s\left(X_{ij}-\bar{X}_{i\cdot}-\bar{X}_{\cdot j}+\bar{X}\right)^2\\S_A=r\sum_{j=1}^s\left(\bar{X}_{\cdot j}-\bar{X}\right)^2,\\S_B=s\sum_{i=1}^r\left(\bar{X}_{i\cdot}-\bar{X}\right)^2\end{gathered}$$

我们先来看看 $\boldsymbol{S}_A$ 的统计意义。因为 $\overline{\boldsymbol{X}}_{\cdot j}$ 是水平 $\boldsymbol{A}_j$ 下所有观测值的平均，所以 $\sum_{j=1}^s\left(\bar{X}_{\cdot j}-\bar{X}\right)^2$ 反映了 $\overline{\boldsymbol{X}}_{-\mathbf{1}},\overline{\boldsymbol{X}}_{\cdot 2},\cdots,\overline{\boldsymbol{X}}_{\cdot}$ 差异的程度 。这种差异是由于因素 $\boldsymbol{A}$ 的不同水平所引起的，因此 S_A 称为因素 A 的平方和。类似地，$\boldsymbol{S}_B$ 称为因素 $\boldsymbol{B}$ 的平方和。至于 $\boldsymbol{S}_E$ 的意义不甚明显，我们可以这样来理解：

因为 $S_E=S_T-S_A-S_B$, 在我们所考虑的两因素问题中，除了因素 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 之外，剩余的再没有其它系统性因素的影响，因此从总平方和中减去 $\boldsymbol{S}_A$ 和 $\boldsymbol{S}_B$ 之后，剩下的数据变差只能归入随机误差, 故 S_E 反映了试验的随机误差。

有了总平方和的分解式 $S_T=S_E+S_A+S_B$, 以及各个验统计量应取为 $\boldsymbol{S}_A$ 与 $\boldsymbol{S}_E$ 的比。

$$F_A=\frac{\frac{S_A}{s-1}}{\frac{S_E}{(r-1)(s-1)}}\sim F(s-1,(r-1)(s-1))$$

$$F_B=\frac{\frac{\boldsymbol{S}_B}{\boldsymbol{r-1}}}{\frac{\boldsymbol{S}_E}{(\boldsymbol{r-1})(s-\mathbf{1})}}\sim\boldsymbol{F}(\boldsymbol{r-1},(\boldsymbol{r-1})(\boldsymbol{s-1}))$$

检验规则为

$F_A<F_\alpha(s-1,(r-1)(s-1))$ 时接受 H_{01}, 否则拒绝 H_{01}；$F_B<F_\alpha(r-1,(r-1)(s-1))$ 时接受 H_{02}，否则拒绝 H_{02}；

9.7.4 关于交互效应的双因素方差分析

与前面方法类似，记

$$\begin{gathered}\bar{X}=\frac{1}{rst}\sum_{i=1}^r\sum_{j=1}^s\sum_{k=1}^t X_{ijk},\bar{X}_{ij\cdot}=\frac{1}{t}\sum_{k=1}^t X_{ijk},\\\bar{X}_{i\cdot\cdot}=\frac{1}{st}\sum_{j=1}^s\sum_{k=1}^t X_{ijk},\bar{X}_{\cdot j\cdot}=\frac{1}{rt}\sum_{i=1}^r\sum_{k=1}^t X_{ijk}\end{gathered}$$

将全体数据对 $\bar{X}$ 的偏差平方和

$$S_T=\sum_{i=1}^r\sum_{j=1}^s\sum_{k=1}^t\left(X_{ijk}-\bar{X}\right)^2$$

进行分解，可待

$$S_T = S_E + S_A + S_B + S_{AB},$$

其中

$$S_E = \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}\left(X_{ijk} - \bar{X}_{ij\cdot}\right)^2$$

$$S_A = rt\sum_{j=1}^{s}\left(\bar{X}_{\cdot j\cdot} - \bar{X}\right)^2$$

$$S_B = st\sum_{i=1}^{r}\left(\bar{X}_{i\cdot\cdot} - \bar{X}\right)^2$$

$$S_{AB} = t\sum_{i=1}^{r}\sum_{j=1}^{s}\left(\bar{X}_{ij\cdot} - \bar{X}_{i\cdot 0} - \bar{X}_{\cdot j\cdot} + \bar{X}\right)^2$$

称 S_E 为误差平方和，$\boldsymbol{S_A}$ 为因素 $\boldsymbol{A}$ 的平方和 (或列间平方和), S_B 为因素 B 的平方和 (或行间平方和), S_{AB} 为交互作用的平方和 (或格间平方和) 。

$$F_{AB} = \frac{\frac{S_{AB}}{(r-1)(s-1)}}{\frac{S_E}{rs(t-1)}} \sim F((r-1)(s-1), rs(t-1))$$

据此统计量, 可以检验 $\boldsymbol{H}_{03}$ 。

水平 α, 检验的结论为: 为交互作用显著。将试验数据按上述分析、计算的结果排成表 7.20 的形式，称为双因素方差分析表。

方差来源	离差平方和	自由度	均方	F 值
因素 A	S_A	$s-1$	$\frac{S_A}{s-1}$	$F_A = \frac{S_A/(s-1)}{S_E/[rs(t-1)]}$
因素 B	S_B	$r-1$	$\frac{S_B}{r-1}$	$F_B = \frac{S_B/(r-1)}{S_E/[rs(t-1)]}$
交互效应	S_{AB}	$\boldsymbol{(r-1)(s-1)}$	$\frac{S_{AB}}{(r-1)(s-1)}$	$F_{AB} = \frac{S_{AB}/[(r-1)(s-1)]}{S_E/[rs(t-1)]}$
误差	$\boldsymbol{S_E}$	$\boldsymbol{rs(t-1)}$		
总和	$\frac{\boldsymbol{S_E}}{\boldsymbol{rs(t-1)}}, S_T$	$\boldsymbol{rst-1}$		

9.8 多元线性回归

多元线性回归分析的模型为

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon \\ \varepsilon \sim N\left(0, \sigma^2\right) \end{cases}$$

式中 $\beta_0, \beta_1, \cdots, \beta_m, \sigma^2$ 都是与 $x_1, x_2, \cdots, x_m$ 无关的未知参数其中 $\beta_0, \beta_1, \cdots, \beta_m$ 称为回归系数。

现得到 $\boldsymbol{n}$ 个独立观测数据 $[b_i, a_{i1}, \cdots, a_{im}]$, 其中 b_i 为 y 的观察值, $\boldsymbol{a}_{i1}, \cdots, \boldsymbol{a}_{im}$ 分别为 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m$ 的观察值,

$$i = 1, \cdots, n, n > m, \quad \text{由(7.19) 得}$$

$$\begin{cases} b_i = \beta_0 + \beta_1 a_{i1} + \cdots + \beta_m a_{im} + \varepsilon_i \\ \varepsilon_i \sim N\left(0, \sigma^2\right), \quad i = 1, \cdots, n. \end{cases}$$

记 $X=\begin{bmatrix}1 & a_{11} & \cdots & a_{1m}\\ \vdots & \vdots & \dots & \vdots\\ 1 & a_{n1} & \cdots & a_{nm}\end{bmatrix}$, $Y=\begin{bmatrix}b_1\\ \vdots\\ b_n\end{bmatrix}$, **7.21** $\varepsilon=[\varepsilon_1,\cdots,\varepsilon_n]^T$, $\beta=[\beta_0,\beta_1,\cdots,\beta_m]^T$

(7.19) 表示为

$$\begin{cases}\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}\\ \boldsymbol{\varepsilon}\sim\boldsymbol{N}\left(\boldsymbol{0},\sigma^2\boldsymbol{E}_n\right)\end{cases}$$

9.8.1 回归模型的假设检验

因变量 y 与自变量 $x_1,\cdots,x_m$ 之间是否存在如模型式 (7.19) 所示的线性关系是雪　巾雨检验的, 显然, 如果所有的 $|\hat{\beta}_j|(j=1,\cdots,m)$ 都很小, y与$x_1,\cdots,x_m$ 的线性关

$$H_0:\beta_j=0,\quad j=1,\cdots,m$$

$$F=\frac{U/m}{Q/(n-m-1)}\sim F(m,n-m-1),$$

在显著性水平 α 下，对于上 α 分位数 $\boldsymbol{F}_\alpha(\boldsymbol{m},\boldsymbol{n-m-1})$, 注：接受 $\boldsymbol{H}_0$ 只说明 $\boldsymbol{y}$ 与 $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_m$ 的线性关系不明显，可能存在非线性关系，如平方关系。还有一些衡量 y 与 $x_1,\cdots,x_m$ 相关程度的指标，如用回归平方和在总平方和中的比值定义复判定系数

$$\boldsymbol{R}^2=\frac{\boldsymbol{U}}{\boldsymbol{SST}}$$

$\boldsymbol{R}=\sqrt{\boldsymbol{R}^2}$ 称为复相关系数, $\boldsymbol{R}$ 越大, $\boldsymbol{y}$ 与 $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_m$ 相关关系越密切，通常，$\boldsymbol{R}$ 大于 0.8 (或 0.9) 才认为相关关系成立。

9.8.2 回归系数的假设检验和区间估计

其中若千个等于零。所以应进一步作如下 $\boldsymbol{m+1}$ 个检验

$$H_0^{(j)}:\beta_j=0,\quad j=0,1,\cdots,m$$

由式 (7.30), $\hat{\beta}_j\sim N\left(\beta_j,\sigma^2c_{jj}\right),c_{jj}$ 是 $\left(X^TX\right)^{-1}$ 中的

$\boldsymbol{H}_0^{(j)}$ 成时

$$t_j=\frac{\hat{\beta}_j/\sqrt{c_{jj}}}{\sqrt{Q/(n-m-1)}}\sim t(n-m-1)$$

绝。式 (7.36) 也可用于对 β_j 作区间估计，在置信水平 $1-\alpha$下, β_j 的置信区间为

$$\left[\hat{\beta}_j-t_{\frac{\alpha}{2}}(n-m-1)s\sqrt{c_{ij}},\hat{\beta}_j+t_{\frac{\alpha}{2}}(n-m-1)s\sqrt{c_{ij}}\right]$$

式中, $s=\sqrt{\frac{Q}{n-m-1}}$ o

9.8.3 利用回归模型进行预测

$[x_1,\cdots,x_m]$ 的取值 $[a_{01},\cdots,a_{0m}]$ 预测 y 的取值 b_0,b_0 是随机的，显然其预测值（点估计）为

$$\hat{b}_0=\hat{\beta}_0+\hat{\beta}_1a_{01}+\cdots+\hat{\beta}_ma_{0m}.(7.38)$$

给定 α 可以算出 $\boldsymbol{b_0}$ 的预测区间（区间估计）, 结果较复简化为

$$\left[\hat{b}_0-z_{\frac{\alpha}{2}}s,\hat{b}_0+z_{\frac{\alpha}{2}}s\right]$$

式中, z_α 是标准正态分布的上恙分位数。对 b_0 的区间估计方法可用于给出已知数据残差 $e_i = b_i - \hat{b}_i (i = 1, \cdots, n)$ 的　口直信区间, e_i 服从均值为零的正态分布，所以若某个 e_i 的置信区间不包含零点, 则认为这个数据是异常的，可予以剔除。

9.9 逐步回归

实际问题中影响因变量的因素可能很多，有些可能关联性强一些，而有些可能影响弱一些。人们总希望从中挑选出对因变量影响显著的自变量来建立回归模型，逐步回归是一种从众多变量中有效地选择重要变量的方法以下只讨论多元线性回归模型的情形

简单地说，就是所有对因变量影响显著的变量都应选入模型，而影响不显著的变量都不应选入模型；从便于应用的角度，变量的选择应使模型中变量个数尽可能少

基本思想: 记 $S = \{x_1, x_2, \cdots, x_m\}$ 为候选的自变量集合 $S_1 \subset S$ 是从集合 S 中选出的一个子集。设 S_1 中有 l 个自变量 $(1 \leq l \leq m)$, 由 S_1 和因变量 y 构造的回归模型的残差平方和为 Q, 则模型的残差方差 $s^2 = Q/(n - l - 1)$ n 为数据样本容量。所选子集 S_1 应使 s^2 尽量小。通常回若模型中包含有对 y 影响很小的变量, 那么 Q 不会由于包含这些变量在内减少多少, 却因 l 的增加可能使 s^2 反而增大, 同时这些对 y 影响不显著的变量也会影响模型的稳定性，因此可将残差方差 s^2 最小作为衡量变量选择的一个数量标准。

可以从另外一个角度考虑自变量 x_j 的显著性。y 对自变量 $x_1, x_2, \cdots, x_m$ 线性回归的残差平方和为 Q, 回归平方和为 U, 在剔除掉 x_j 后，用 y 对其余的 $m - 1$ 个自变量做回归, 记所得的残差平方和为 $Q_{(j)}$, 回归平方和为 $U_{(j)}$, 则自变量 x_j 对回归的贡献为

$$\Delta U_{(j)} = U - U_{(j)}$$

称为 x_j 的偏回归平方和。由此构造偏 F 统计量

$$F_j = \frac{\Delta U_{(j)}/1}{Q/(n - m - 1)}$$

F_i 服从自由度为 $(1, n - m - 1)$ 的 F 分布，此 F 检验与式方程中剔除变元时，回归平方和减少，残差平方和增加。根据平方和分解式可知

$$\Delta U_{(j)} = \Delta Q_{(j)} = Q_{(j)} - Q$$

残差平方和减少, 两者的增减量同样相等。

当自变量的个数较多时，求出所有可能的回归方程是非常困难的。为此，人们提出了一些较为简便、实用、快速的选择自变量的方法。这些方法各有优缺点，至今还没有绝对最优的方法，目前常用的方法有前进法、后退法、逐步回归法，而逐步回归法最受推崇。

9.9.1 前进法

前进法的思想是变量由少到多, 每次增加一个, 直至没有可引入的变量为止。具体做法是首先将全部 m 个自变量分别对因变量 y 建立一元线性回归方程, 利用归系数的 F 检验值, 记为 $\{F_1^1, F_2^1, \cdots, F_m^1\}$, 选其最大者记为

$$F_j^1 = \max\{F_1^1, F_2^1, \cdots, F_m^1\}$$

给定显著性水平 α, 若 $F_j^1 \geq F_\alpha(1, n - 2)$, 则首先将 x_j 引入回归方程，为了方便，不妨设 x_j 就是 x_1 。

接下来因变量 y 分别与 $(x_1, x_2), (x_1, x_3), \cdots, (x_1, x_m)$ 建立二元线性回归方程, 对这 $m - 1$ 个回归方程中 $x_2, x_3, \cdots, x_m$ 的回归系数进行 F 检验, 利用 (7.43) 式计算 F 值, 记为 $\{F_2^2, F_3^2, \cdots, F_m^2\}$, 选其最大者记为

$$F_j^2 = \max\{F_2^2, F_3^2, \cdots, F_m^2\}$$

若 $\boldsymbol{F}_j^2 \geq \boldsymbol{F}_\alpha(\mathbf{1}, \mathbf{n}-\mathbf{3})$, 则接着将 $\boldsymbol{x_j}$ 引入回归方程。

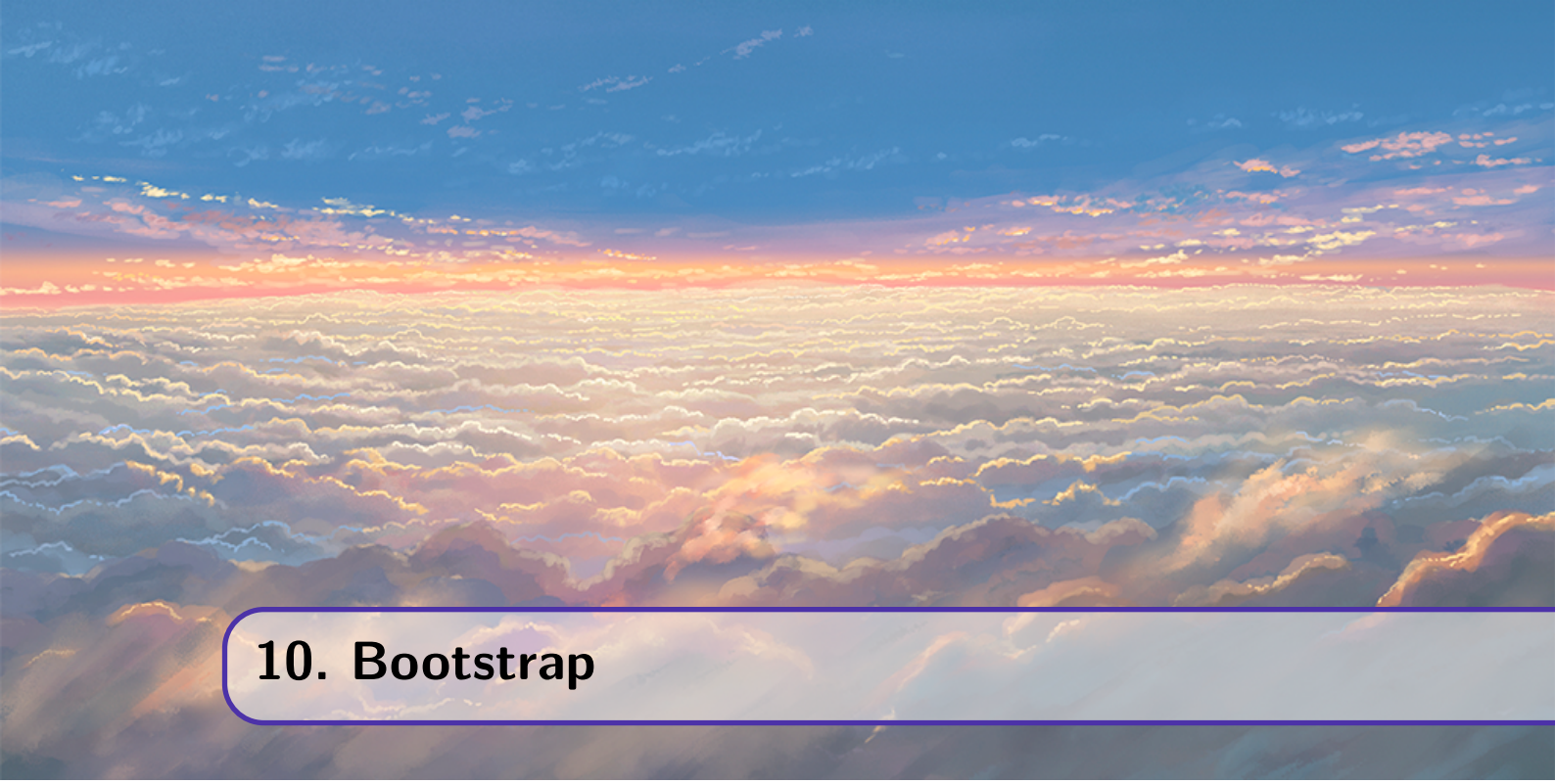
依上述方法接着做下去，直至所有未被引入方程的入变量的个数。这时, 得到的回归方程就是确定的方程。有关，在用软件计算时，我们实际是使用显著性 $\boldsymbol{P}$ 值做检验。

9.9.2 后退法

后退法易于掌握，我们使用 t 统计量做检验，与 $\boldsymbol{F}$ 统计量做检验是等价的。具体步骤如下:

Algorithm 8: 后退法

1 以全部自变量作为解释变量拟合方程;
2 每一步都在未通过 $\boldsymbol{t}$ 检验的自变量中选择一个 $\boldsymbol{t}_j$ 值最小的变量 $\boldsymbol{x}_j$, 将它从模型中删除;
3 直至所有的自变量均通过 $\boldsymbol{t}$ 检验, 则算法终止;



10. Bootstrap

设总体的分布 F 未知，但已知有一个容量为 n 的来自分布 F 的数据样本，自这一样本按放回抽样的方法抽取一个容量为 n 的样本，这种样本称为 bootstrap 样本或称为自助样本。相继地，独立地自原始样本中取很多个 Bootstrap 样本，利用这些样本对总体 F 进行统计推断，这种方法称为非参数 Bootstrap 方法，又称自助法。这一方法可以用于当人们对总体知之甚少的情况，它是近代统计中的一种用于数据处理的重要实用方法。这种方法的实现需要在计算机上作大量的计算，随着计算机威力的增长，它已成为一种流行的方法。

10.1 估计量的标准误差的 Bootstrap 估计

在估计总体未知参数 θ 时，人们不但要给出 θ 的估计 $\hat{\theta}$, 还需指出这一估计 $\hat{\theta}$ 的精度。通常我们用估计量 $\hat{\theta}$ 的标准差 $\sqrt{D(\hat{\theta})}$ 来度量估计的精度。估计量 $\hat{\theta}$ 的标准差 $\sigma_{\hat{\theta}}=\sqrt{D(\hat{\theta})}$ 也称为*估计量 $\hat{\theta}$ 的标准误差*。

设 $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, X_n$ 是来自 $\mathcal{F}(x)$ 为分布函数的总体的样本, θ 是我们感兴趣的未知参数, 用 $\hat{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}\left(\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n\right)$ 作为 θ 的估计量，在应用中 $\hat{\theta}$ 的抽样分布常是很难处理的, 这样, $\sqrt{D(\hat{\theta})}$ 常没有一个简单的表达式，不过我们可以用计算机模拟的方法来求得 $\sqrt{D(\hat{\theta})}$ 的估计。

为此, 自 $\boldsymbol{F}$ 产生很多容量为 $\boldsymbol{n}$ 的样本 (例如 $\boldsymbol{B}$个), 对于每一个样本计算 $\hat{\theta}$ 的值, 得 $\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_B$, 则 $\sqrt{D(\hat{\theta})}$ 可以用

$$\hat{\sigma}_{\hat{\theta}}=\sqrt{\frac{1}{B-1} \sum_{i=1}^{B}\left(\hat{\theta}_i-\bar{\theta}\right)^2}$$

来估计，其中 $\bar{\theta}=\frac{1}{B} \sum_{i=1}^{B} \hat{\theta}_i$ 。

然而 F 常常是未知的，这样就无法产生模拟样本，需要另外的方法。

设分布 F 未知, $x_1, x_2, \cdots, x_n$ 是来自 F 的样本值，$\boldsymbol{F}_n$ 是相应的经验分布函数。当 n 很大时, F_n 接近 F 。1 用 F_n 代替上一段中的 F, 在 F_n 中抽样。在 F_n 中抽样, 就是在原

始样本 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$ 中每次随机地取一个个体作放回抽样。如此得到一个容量为 $\boldsymbol{n}$ 的样本 $\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \cdots, \boldsymbol{x}_n^*$, 这就是第一段中所说的 Bootstrap 样本。用 Bootstrap 样本按上一段中计算估计 $\hat{\theta}(x_1, x_2, \cdots, x_n)$ 那样求出 θ 的估计 $\hat{\theta}^* = \hat{\theta}(x_1^*, x_2^*, \cdots, x_n^*)$ 估计 $\hat{\theta}^*$ 称为 θ 的 Bootstrap 估计。

相应地、独立地抽得 B 个 Bootstrap 样本, 以这些样本分别求出 θ 的相应的 Bootstrap 估计如下:

Bootstrap 样本 1 $x_1^{*1}, x_2^{*1}, \cdots, x_n^{*1}$, Bootstrap 估计 $\hat{\theta}_1^*$;
Bootstrap 样本 2 $x_1^{*2}, x_2^{*2}, \cdots, x_n^{*2}$, Bootstrap 估计 $\hat{\theta}_2^*$;
Bootstrap 样本 $\boldsymbol{B}$ $\boldsymbol{x}_1^{*B}, \boldsymbol{x}_2^{*B}, \cdots, \boldsymbol{x}_n^{*B}$, Bootstrap 估计 $\hat{\theta}_B^*$.

则 $\hat{\theta}$ 的标准误差 $\sqrt{D(\hat{\theta})}$, 就以

$$\hat{\sigma}_{\hat{\theta}} = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\theta}_i^* - \bar{\theta}^*\right)^2}$$

来估计, 其中 $\bar{\theta}^\star = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^*$, 上式就是 $\sqrt{D(\hat{\theta})}$ 的 Bootstrap 估计。

Algorithm 9: 求 $\sqrt{D(\hat{\theta})}$ 的 Bootstrap 估计

1 自原始数据样本 $x_1, x_2, \cdots, x_n$ 按放回抽样的方法, 抽得容量为 n 的样本 $x_1^*, x_2^*, \cdots, x_n^*$ (称为 *Bootstrap 样本*);
2 相继地、独立地求出 $B(B \geq 1000)$ 个容量为 n 的 Bootstrap 样本, $x_1^{*i}, x_2^{*i}, \cdots, x_n^{*i}, \boldsymbol{i} = \boldsymbol{1}, \boldsymbol{2}, \cdots, \boldsymbol{B}$ 。对于第 $\boldsymbol{i}$ 个 Bootstrap 样本, 计算 $\hat{\theta}_i^* = \hat{\theta}\left(x_1^{*i}, x_2^{*i}, \cdots, x_n^{*i}\right), \boldsymbol{i} = \boldsymbol{1}, \boldsymbol{2}, \cdots, B$ $\left(\hat{\theta}_i^*\right.$ 称为 θ 的第 $\boldsymbol{i}$ 个 *Bootstrap 估计*$\left.\right)$;
3 计算 $\hat{\sigma}_{\hat{\theta}} = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\theta}_i^* - \bar{\theta}^*\right)^2}$, 其中 $\bar{\theta}^* = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^*$

10.2 估计量的均方误差的 Bootstrap 估计

设 $X = (X_1, X_2, \cdots, X_n)$ 是来自总体 F 的样本, F 未知,$\boldsymbol{R} = \boldsymbol{R}(\boldsymbol{X})$ 是感兴趣的随机变量, 它依赖于样本 $\boldsymbol{X}$。。假设我们希望去估计 R 的分布的某些特征。例如 R 的数学期望 $E_F(\boldsymbol{R})$, 就可以按照上面所说的三个步骤 $1^\circ, 2^\circ 3^\circ$ 进行, 只是在 2° 中对于第 $\boldsymbol{i}$ 个 Bootstrap 样本 $\boldsymbol{x}_i^* = \left(x_1^{*i}, x_2^{*i}, \cdots, x_n^{*i}\right)$, 计算 $R_i^* = R_i^*(x_i^*)$ 代替计算 θ_i^*, 且在 3° 中计算感兴趣的 R 的特征。例如如果希望估计 $\boldsymbol{E}_F(\boldsymbol{R})$ 就计算

$$E_*(R^*) = \frac{1}{B}\sum_{i=1}^{B}R_i^*$$

10.3 Bootstrap 置信区间

设 $X = (X_1, X_2, \cdots, X_n)$ 是来自总体 F 容量为 n 的样本 $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)$ 是一个已知的样本值。$\boldsymbol{F}$ 中含有未知参数 $\theta, \hat{\theta} = \hat{\theta}(X_1, X_2, \cdots, X_n)$ 是 θ 的估计量。现在来求 θ 的置信水平为 $1-\alpha$ 的置信区间。

相继地, 独立地从样本 $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ 中抽出 B 个容量为 n 的 Bootstrap 样本, 对于每个 Bootstrap 样本求出 θ 的 Bootstrap 估计: $\hat{\theta}_1^*, \hat{\theta}_2^*, \cdots, \hat{\theta}_B^*$ 。将它们自小到大排序, 得

$$\hat{\theta}^*_{(1)} \leq \hat{\theta}^*_{(2)} \leq \cdots \leq \hat{\theta}^*_{(B)}$$

取 $R(X) = \hat{\theta}$, 用对应的 $R(X^*) = \hat{\theta}^*$ 的分布作为 $R(X)$ 的分布的近似, 求出 $R(X^*)$ 的分布的近似分位数 $\hat{\theta}^*_{\alpha/2}$ 和 $\hat{\theta}^*_{1-\alpha/2}$ 使

$$P\left\{\hat{\theta}^*_{\alpha/2} < \hat{\theta}^* < \hat{\theta}^*_{1-\alpha/2}\right\} = 1 - \alpha$$

于是近似地有

$$P\left\{\hat{\theta}^*_{\alpha/2} < \theta < \hat{\theta}^*_{1-\alpha/2}\right\} = 1 - \alpha.$$

记 $\boldsymbol{k}_1 = \left[\boldsymbol{B} \times \frac{\alpha}{2}\right], \boldsymbol{k}_2 = \left[\boldsymbol{B} \times \left(\mathbf{1} - \frac{\alpha}{2}\right)\right]$, 在上式中以 $\hat{\theta}^*_{(k_1)}$ 和 $\hat{\theta}^*_{(k_2)}$ 分别作为分位数 $\hat{\theta}^*_{\alpha/2}$ 和 $\hat{\theta}^*_{1-\alpha/2}$ 的估计, 得到近似等式

$$P\left\{\hat{\theta}^*_{(k_1)} < \theta < \hat{\theta}^*_{(k_2)}\right\} = 1 - \alpha$$

于是由上式就得到 θ 的置信水平为 $1-\alpha$ 的近似置信区间 $\left(\hat{\theta}^*_{(k_1)}, \hat{\theta}^*_{(k_2)}\right)$, 这一区间称为 θ 的置信水平为 $1-\alpha$ 的 Bootstrap 置信区间。这种求置信区间的方法称为分位数法。

10.4 参数 Bootstrap 方法

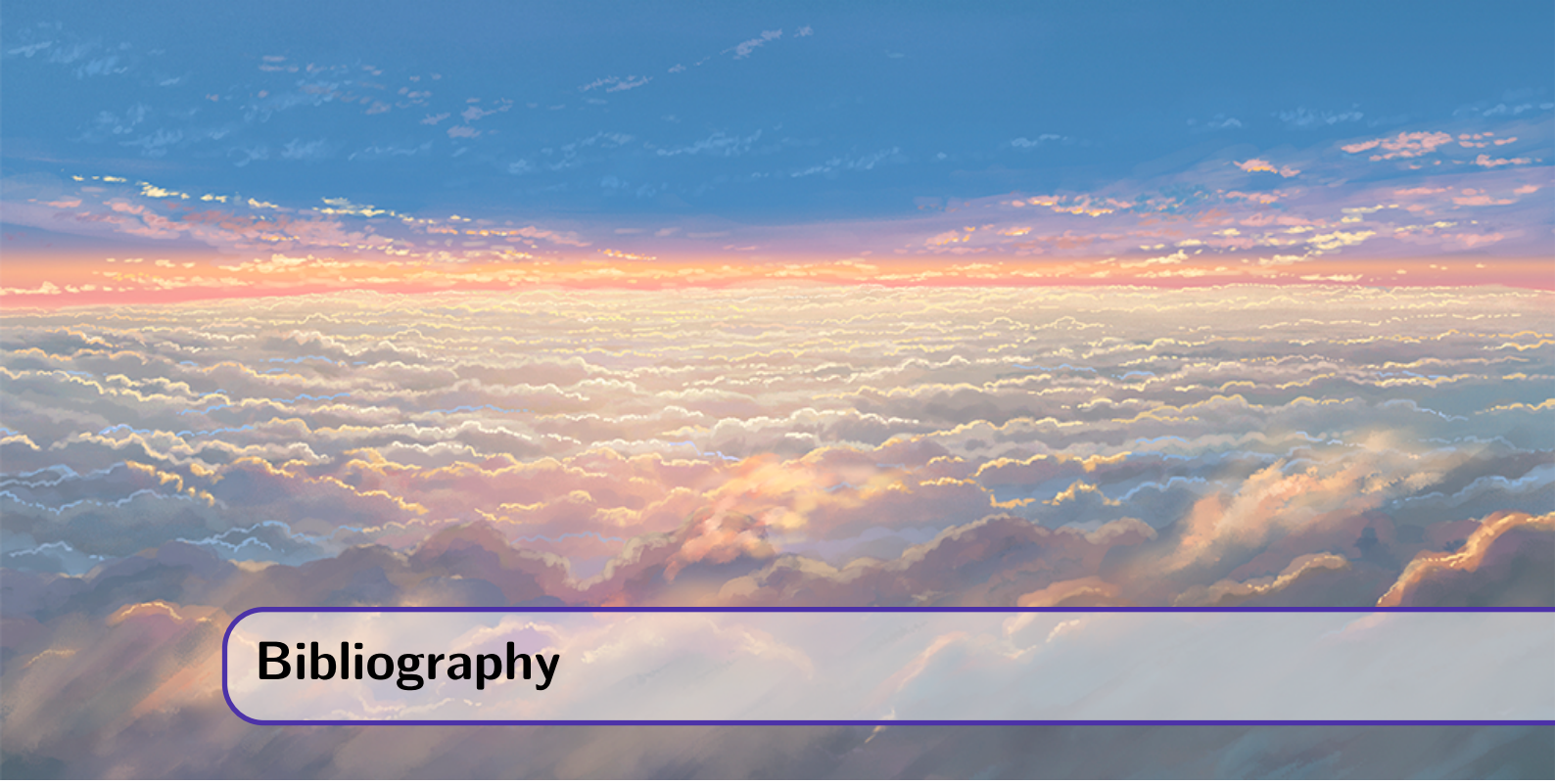
假设所研究的总体的分布函数 $F(x;\beta)$ 的形式已知, 但其中包含未知参数 β (β 可以是向量)。现在已知有一个来自 $F(x;\beta)$ 的样本

$$X_1, X_2, \cdots, X_n$$

利用这一样本求出 β 的最大似然估计 $\hat{\beta}$ 。在 $F(x;\beta)$ 中以 $\hat{\beta}$ 代替 β 得到 $F(x;\hat{\beta})$, 接着在 $F(x;\hat{\beta})$ 中产生容量为 n 的样本

$$X_1^*, X_2^*, \cdots, X_n^* \sim F(x;\hat{\beta})$$

这种样本可以产生很多个, 例如产生 $B(B \geq 1000)$ 个, 就可以利用这些样本对总体进行统计推断，其做法与非参数 Bootstrap 方法一样。这种方法称为参数 Bootstrap 方法。



Bibliography

Articles

Books