

数理统计的基本概念

人类社会中,有生产就有管理,有管理就有统计,统计活动应用广泛,历史悠久。通常把统计对象称为**个体**,所有个体构成的集合称为**总体**。若总体中的每个个体能表示成数值,则总体就可视为一个随机变量 X ,统计的任务就是要弄清楚总体 X 的分布。当总体规模较小的时候,“普查”这种统计方法是有效且准确的。随着总体规模越来越庞大,普查形式的统计活动,其成本和周期越来越难以承受。人们尝试着从总体 X 抽取有限容量的**样本**,运用概率论的术语、理论、方法研究样本,进而推算总体的分布。用这种方式进行的统计活动,为有别于传统的普查统计,称为**数理统计**。本章起,就来讨论数理统计中的一些经典问题及其解决方法,并介绍 Python 解法。

5.1 简单样本

定义 5-1 设总体为随机变量 X ,随机地从 X 中独立地抽取 n 个个体,观察每个个体的数据值,按抽取顺序记为 $(x_1, x_2, \dots, x_n)$,称为 X 的一个容量为 n 的**样本观测值**。第 i 个个体观测值 x_i 在不同批次的取样中呈现为一个随机变量 $X_i, i = 1, 2, \dots, n$ 。随机向量 $(X_1, X_2, \dots, X_n)$ 称为来自总体 X 的一个容量为 n 的**简单样本**,简称为**样本**。

由简单样本的抽样方式不难理解,样本中的随机变量 $X_1, X_2, \dots, X_n$ 具有

- (1) 代表性: 每个 X_i 与总体 X 具有相同的分布 ($i = 1, 2, \dots, n$)。
- (2) 独立性: $X_1, X_2, \dots, X_n$ 相互独立。

5.1.1 样本观测值的直方图

对总体 X 的一次具体抽样,得到一个容量为 n 的样本观测值 $(x_1, x_2, \dots, x_n)$ 。可以通过如下步骤,绘制直方图。

- (1) 对 $x_1, x_2, \dots, x_n$ 按升序排列。不失一般性,仍记为 $x_1, x_2, \dots, x_n$, 即

$$x_1 \leq x_2 \leq \cdots \leq x_n$$

(2) 区间 $[a, b] = [x_1, x_n]$ 包含所有的个体观测值 $x_i, i = 1, 2, \cdots, n$ 。将区间 $[a, b]$ 等分成 $m(\leq n)$ 个小区间, 约定除最后一个小区间为闭区间外, 其余均为半闭半开区间。记区间长度为 Δ 。自左向右, 统计第 k 个小区间中所含个体观测值 x_i 的个数, 即个体观测值落在该区间内的频数, 记为 $f_k, k = 1, 2, \cdots, m$ 。

(3) 在坐标平面中, 横坐标轴区间 (a, b) 上依次绘制 m 个高为 $\frac{f_k}{n} / \Delta$, 宽为 Δ 的矩形, $k = 1, 2, \cdots, m$ 。

显然, 这样绘制出来的 $(x_1, x_2, \cdots, x_n)$ 的直方图中所有小矩形面积之和为 1。这是因为将直方图中每个小矩形的高设置为 $\frac{f_k}{n} / \Delta$, 乘以宽度 Δ 后面积值为频率 $\frac{f_k}{n}$, 这样设置直方图小矩形高度称为密度化。

例 5-1 下面列出 84 个伊特拉斯坎 (Etruscan) 人男子头颅的最大宽度 (mm)。绘制这些数据的直方图。

```
141  148  132  138  154  142  150  146  155  158  150  140  147  148
144  150  149  145  149  158  143  141  144  144  126  140  144  142
141  140  145  135  147  146  141  136  140  146  142  137  148  154
137  139  143  140  131  143  141  149  148  135  148  152  143  144
141  143  147  146  150  132  142  142  143  153  149  146  149  138
142  149  142  137  134  144  146  147  140  142  140  137  152  145
```

解: (1) $n = 84$ 个数据按升序排序后的序列为:

```
126 131 132 132 134 135 135 136 137 137 137 137 138 138
139 140 140 140 140 140 140 140 141 141 141 141 141 141
142 142 142 142 142 142 142 142 143 143 143 143 143 143
144 144 144 144 144 144 145 145 145 146 146 146 146 146
146 147 147 147 147 148 148 148 148 148 149 149 149 149
149 149 150 150 150 150 152 152 153 154 154 155 158 158
```

(2) 将区间 $[126, 158]$ 等分成 $m = 8$ 个小区间, 统计各小区间内所含个体观测值个数, 形成表 5-1。

(3) 根据表 5-1 在坐标平面上逐一画出高为 $\frac{f_k}{n} / \Delta$, 宽为 Δ 的矩形, $k = 1, 2, \cdots, m$, 如图 5-1 所示。

沿密度化直方图的外轮廓描绘曲线, 得到总体 X 的密度函数的近似曲线 (见图 5-1)。

表 5-1 小区间内所含个体观测值统计

小 区 间	频数 f_k	频率 f_k/n	累 积 频 率
[126, 130)	1	0.0119	0.0119
[130, 134)	3	0.0357	0.0476
[134, 138)	8	0.0952	0.1429
[138, 142)	16	0.1905	0.3333
[142, 146)	23	0.2738	0.6071
[146, 150)	21	0.2500	0.8571
[150, 154)	7	0.0833	0.9405
[154, 158]	5	0.0595	1.0

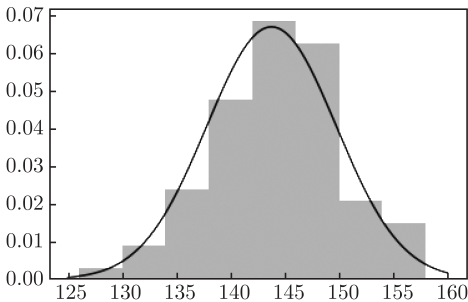


图 5-1 样本观测值 $(x_1, x_2, \dots, x_n)$ 的直方图

练习 5-1 下面列出了 30 个美国 NBA 球员的体重数据（单位：磅），这些数据是从 NBA 球队（1990—1991）赛季的花名册中抽样得到的。

225 232 232 245 235 245 270 225 240 240
217 195 225 185 200 220 200 210 271 240
220 230 215 252 225 220 206 185 227 236

绘制这些数据分为 6 组的直方图。

参考答案：见图 5-2。

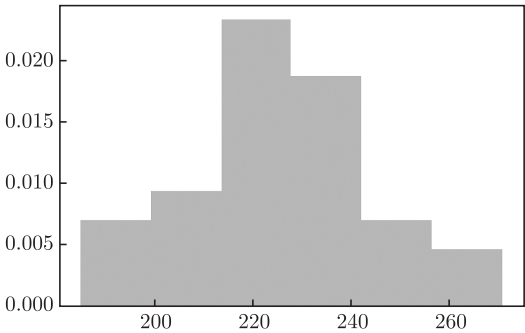


图 5-2 练习 5-1 的参考答案

5.1.2 经验分布函数

设 $(x_1, x_2, \dots, x_n)$ 是总体 X 的一个样本观测值。与绘制直方图相仿，通过下列步骤即可得到 $(x_1, x_2, \dots, x_n)$ 的经验分布函数 $F_n(x)$ 。

(1) 对 $x_1, x_2, \dots, x_n$ 按升序排列，记 $a = x_1, b = x_n$ 。

(2) 将区间 $[a, b]$ 等分成 $m (\leq n)$ 个小区间，约定除最后一个小区间为闭区间外，其余均为半闭半开区间 $\Delta_k, k = 1, 2, \dots, m$ 。统计第 k 个小区间中所含个体观测值 x_i 的个数，即个体观测值落在该区间内的频数，记为 $f_k, k = 1, 2, \dots, m$ 。

(3) 定义函数

$$F_n(x) = \begin{cases} 0, & x < a \\ \frac{f_1}{n}, & x \in \Delta_1 \\ \frac{f_1 + f_2}{n}, & x \in \Delta_2 \\ \vdots & \vdots \\ \frac{f_1 + \dots + f_{m-1}}{n}, & x \in \Delta_m \\ 1, & x \geq b \end{cases}$$

$(x_1, x_2, \dots, x_n)$ 的经验分布函数 $F_n(x)$ 就是累积频率函数，其函数图像为一个阶梯状曲线，如图 5-3 所示折线部分即为例 5-1 中伊特拉斯坎人男子头颅的最大宽度数据的经验分布函数图像。沿 $F_n(x)$ 的边缘描绘一条光滑曲线则是总体 X 的分布函数的近似图像（见图 5-3）。

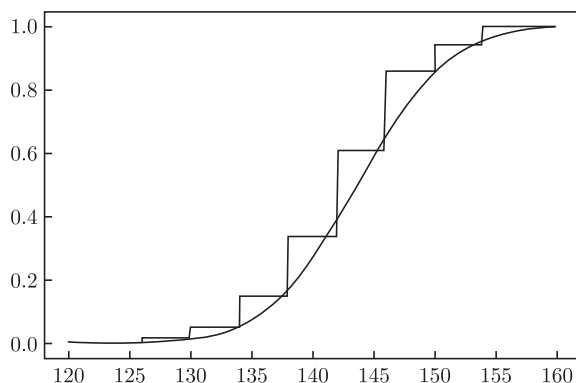


图 5-3 样本观测值 $(x_1, x_2, \dots, x_n)$ 的经验分布函数

练习 5-2 构造练习 5-1 中球员体重数据分为 6 组的经验分布函数。

$$\text{参考答案: } F_6(x) = \begin{cases} 0, & x < 185 \\ \frac{3}{30}, & 185 \leq x < 200 \\ \frac{7}{30}, & 200 \leq x < 214 \\ \frac{17}{30}, & 214 \leq x < 228 \\ \frac{25}{30}, & 228 \leq x < 242 \\ \frac{28}{30}, & 242 \leq x < 256 \\ 1, & x \geq 256 \end{cases}$$

5.1.3 Python 解法

1. 绘制数据直方图

设样本观测数据存储于数组中。matplotlib 包中的 pyplot 对象有一个专门绘制数据数组的直方图的函数 hist。该函数的调用接口为

hist(x, bins, density)

参数 x 传递数据数组。bins 传递分组个数，即小区间个数。参数 density 传递布尔值 True 或 False，默认值为 False。若为 True，意味着密度化直方图。

例 5-2 下列代码绘制例 5-1 中伊特拉斯坎人男子头颅的最大宽度数据的直方图。

```

1 import numpy as np                                #导入numpy
2 from matplotlib import pyplot as plt              #导入绘图对象pyplot
3 sample=np.array([141,148,132,138,154,142,150,146,155,158, #设置样本数据数组
4                  150,140,147,148,144,150,149,145,149,158,
5                  143,141,144,144,126,140,144,142,141,140,
6                  145,135,147,146,141,136,140,146,142,137,
7                  148,154,137,139,143,140,131,143,141,149,
8                  148,135,148,152,143,144,141,143,147,146,
9                  150,132,142,142,143,153,149,146,149,138,
10                 142,149,142,137,134,144,146,147,140,142,
11                 140,137,152,145])
12 plt.hist(sample, bins=8, density=True)           #分成8组绘制直方图
13 plt.show()
```

程序 5.1 绘制例 5-1 中样本数据直方图的 Python 程序

注意第 12 行调用 pyplot 对象的 hist 函数时所传递的各个参数：传递给接收数据数组参数 x 的是样本观测值数组 sample（在第 3~11 行设置），传递给确定涵盖所有样

本数据的区间分组数的 bins 为 8, 传递给决定小矩形高度的参数 density 为 True。运行程序 5.1, 展示的直方图如图 5-1 所示。

练习 5-3 用 Python 绘制练习 5-1 中球员体重数据的直方图。

参考答案: 见文件 chapter05.ipynb 对应代码。

2. 样本观测值的经验分布函数

除了 matplotlib 包中的 pyplot 对象提供的绘制数据数组的直方图的函数 hist, numpy 包也提供了一个 histogram 的函数, 该函数的调用接口为:

histogram(x, bins)

参数 x 的意义是接收数据数组, bins 也是确定分组个数。假定传递给 bins 的置为 m , 该函数返回一个二元组 (f, b), 其中, f 为含有 m 个整数的数组, 每个元素表示对应小区间所含个体观测值个数。b 也是一个数组其中含有 $m+1$ 个元素 $\{b_0, b_1, \dots, b_m\}$, $[b_{i-1}, b_i)$ 为第 i 个小区间, $i = 1, 2, \dots, m$ 。下列代码就是利用 numpy 的这一函数, 计算样本观测值的经验分布函数 $F_n(x)$ 。

```

1 import numpy as np                #导入numpy
2 def Fn(x, sample, m):             #定义函数Fn
3     n=sample.size                  #样本容量n
4     f, b=np.histogram(sample, bins=m) #计算样本分组及频数
5     f=f/n                          #计算各组频率
6     for k in range(m-1):           #计算累积频率
7         f[k+1]+=f[k]
8     y=np.zeros(x.size)             #函数值初始化为0
9     for i in range(1, m):          #计算每个分组区间内的函数值
10        d=np.where((x>b[i-1])&(x<=b[i]))
11        y[d]=f[i-1]
12    d=np.where(x>b[m-1])            #计算最后小区间及其以后的函数值
13    y[d]=1
14    return y                        #返回y

```

程序 5.2 经验分布函数的 Python 定义

程序中第 2~14 行定义计算由参数 sample 传递的样本数据, 分成 m 组的经验分布函数, 其中, 参数 x 表示自变量, 这是一个数组类对象。第 3 行计算样本容量 n 。第 4 行调用 numpy 的 histogram 函数, 计算样本 sample 分成 m 组的频数数据 f 和分组数据 b 。第 5 行计算各组频率 (频数/样本容量)。第 6~7 行计算各分组的累积频率。第 8 行将函数值 y 初始化为 0。第 9~11 行的 for 循环用存储在 f 中的累积频率数据计算 x 中所含前 m 个小区间的函数值。第 12 行计算 x 中包含的最后一个区间及其以右部分设置为 d , 第 13 行将 d 对应的函数置 1。将程序 5.2 中定义的函数 F_n 存储于文件 utility.py, 以方便调用。

例 5-3 下列程序绘制例 5-1 中伊特拉斯坎人男子头颅的最大宽度数据分成 $m = 8$ 组的经验分布函数图形。

```

1 import numpy as np                                #导入numpy
2 from matplotlib import pyplot as plt              #导入绘图对象plt
3 from utility import Fn                            #导入Fn
4 sample=np.array([141,148,132,138,154,142,150,146,155,158, #设置样本数据数组
5                  150,140,147,148,144,150,149,145,149,158,
6                  143,141,144,144,126,140,144,142,141,140,
7                  145,135,147,146,141,136,140,146,142,137,
8                  148,154,137,139,143,140,131,143,141,149,
9                  148,135,148,152,143,144,141,143,147,146,
10                 150,132,142,142,143,153,149,146,149,138,
11                 142,149,142,137,134,144,146,147,140,142,
12                 140,137,152,145])
13 x=np.linspace(120, 160, 256)                    #设置绘图横轴
14 plt.plot(x,Fn(x, sample, 8))                    #绘制Fn(x)图形
15 plt.show()
```

程序 5.3 经验分布函数的 Python 定义

程序的第 4~12 行设置样本观测值数据 sample, 数据与程序 5.1 中的 sample 相同。第 13 行设置绘图的横坐标区间 $x = (120, 160)$, 第 14 行调用 pyplot 的 plot 函数, 利用程序 5.2 定义的函数 Fn, 绘制由数据数组 sample 算得的经验分布函数的图形, 运行此程序展示图 5-3 中的折线部分。

练习 5-4 用 Python 绘制练习 5-2 中算得的球员体重数据的经验分布函数图形。

参考答案: 见文件 chapter05.ipynb 中对应代码。

5.2 样本统计量

定义 5-2 设 $(X_1, X_2, \dots, X_n)$ 是来自总体 X 的一个简单样本, n 元连续函数 $g(X_1, X_2, \dots, X_n)$ 若不含任何未知参数, 则称其为样本的一个统计量。

例如, 设 (X_1, X_2, X_3) 为来自服从正态分布 $N(\mu, \sigma^2)$ 的总体 X 的样本, 其中, μ 为已知, σ^2 未知, 则 $X_1 + X_2 - X_3$, $3X_1 - 2X_2 + X_3$, $X_2 + 5X_3 - 4\mu$, $X_1^2 + X_2^2 + X_3^2$ 都是统计量, 但是 $X_1^2 + X_2^2 + X_3^2 + \sigma^2$ 和 $X_1 + X_2 - \sigma^2 X_3$ 都不是统计量。

设 $g(X_1, X_2, \dots, X_n)$ 为样本 $(X_1, X_2, \dots, X_n)$ 的一个统计量。 $(x_1, x_2, \dots, x_n)$ 是 $(X_1, X_2, \dots, X_n)$ 的一个观测值, 则 $g(x_1, x_2, \dots, x_n)$ 称为 $g(X_1, X_2, \dots, X_n)$ 的一个观测值。例如, $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$ 是统计量 $\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$ 的一个观测值。

5.2.1 常用统计量

设 $(X_1, X_2, \dots, X_n)$ 为来自总体 X 的一个样本, 数理统计中常用的统计量有以下几种。

$$(1) \text{ 样本均值: } \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i。$$

$$(2) \text{ 样本方差: } S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right)。$$

$$(3) \text{ 样本均方差: } S = \sqrt{S^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}。$$

$$(4) \text{ 样本 } k \text{ 阶原点矩: } A_k = \frac{1}{n} \sum_{i=1}^n X_i^k, \quad k = 1, 2, \dots。$$

$$(5) \text{ 样本 } k \text{ 阶中心矩: } B_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k, \quad k = 1, 2, \dots。$$

希望提起重视的是, 虽然样本均值恰为样本 1 阶原点矩, 但是样本方差不是 2 阶样本中心矩, 因为两者的除数分别为 $n-1$ 和 n 。

设 $(x_1, x_2, \dots, x_n)$ 为 $(X_1, X_2, \dots, X_n)$ 的观测值, 则相应的统计量的观测值为

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

$$(2) s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right)。$$

$$(3) s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}。$$

$$(4) a_k = \frac{1}{n} \sum_{i=1}^n x_i^k, \quad k = 1, 2, \dots。$$

$$(5) b_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k, \quad k = 1, 2, \dots。$$

例 5-4 设总体 X 的样本观测值为 1502, 1453, 1376, 1650, 求样本均值, 样本方差和样本均方差。

解: 样本均值为

$$\bar{x} = \frac{1}{4}(1502 + 1453 + 1376 + 1650) = 1493$$

样本方差为

$$s^2 = \frac{1}{3}[(1502 - 1493)^2 + (1453 - 1493)^2 + (1376 - 1493)^2 + (1650 - 1493)^2] = 14068.667$$

样本均方差为

$$s = \sqrt{s^2} = \sqrt{14068.667} = 118.611$$

练习 5-5 设总体的一组样本观测值为 (54, 67, 68, 78, 70, 66, 67, 70, 65, 69), 计算样本均值和样本方差。

参考答案: 67.4, 35.16。

5.2.2 正态总体的样本统计量分布

定理 5-1 设总体 $X \sim N(\mu, \sigma^2)$, $(X_1, X_2, \dots, X_n)$ 是来自 X 的一个样本。记 $\bar{X}$ 、 S^2 和 S 分别为样本均值、样本方差和样本均方差, 则 $\bar{X}$ 、 S^2 和 S 有如下性质。

- (1) $\bar{X} \sim N(\mu, \sigma^2/n)$ 。
- (2) $\frac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1)$ 。
- (3) $\frac{\bar{X} - \mu}{S} \sqrt{n} \sim t(n-1)$ 。

证明见本章附录 A1。

5.2.3 两个正态总体的样本统计量分布

定理 5-2 设样本 $(X_1, X_2, \dots, X_{n_1})$ 来自总体 $X \sim N(\mu_1, \sigma_1^2)$, $(Y_1, Y_2, \dots, Y_{n_2})$ 来自总体 $Y \sim N(\mu_2, \sigma_2^2)$ 。 X 与 Y 相互独立。 $(X_1, X_2, \dots, X_{n_1})$ 的样本均值和样本方差分别为 $\bar{X}$ 和 S_1^2 , $(Y_1, Y_2, \dots, Y_{n_2})$ 的样本均值和样本方差分别为 $\bar{Y}$ 和 S_2^2 。则

- (1) $\bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$ 。
- (2) $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$ 。
- (3) 若 $\sigma_1^2 = \sigma_2^2 = \sigma^2$,

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

其中, $S_w^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}$, $S_w = \sqrt{S_w^2}$ 。证明见本章附录 A2。

在以后各章的讨论中会看到本节的关于正态总体样本统计量分布性质的定理 5-1 和定理 5-2 是数理统计中诸如正态总体参数的区间估计和假设检验等计算的基本根据。

5.2.4 Python 解法

5.2.2 节及 5.2.3 节涉及的关于正态分布簇: 正态分布、 χ^2 分布、 t 分布和 F 分布的 Python 表示及分位点计算在 3.5.6 节中已有讨论。此处仅就 5.2.1 节的样本均值和样本方差的计算介绍 Python 解法。

numpy 包中的 array 类数组对象拥有函数 mean, 直接调用该函数就可算得存储在其中的样本均值; 函数 var 计算样本方差, 函数 std 计算样本均方差。值得注意的是, var 和 std 都带有命名参数 ddof, 其意义是计算中使用的除数是 n-ddof, 其中, n 表示元素的数量。默认情况下, ddof 为零。因此, 按样本方差定义, 需要传递 ddof=1。

例 5-5 下列代码验算例 5-4 所求的结果。

```
1 import numpy as np          #导入numpy
2 x=np.array([1502, 1453, 1367, 1650]) #设置样本数据
3 mean=x.mean()              #计算样本均值
```

```

4 s2=x.var(ddof=1)           #计算样本方差
5 s=x.std(ddof=1)           #计算样本均方差
6 print('x_=%.4f'%mean)
7 print('s^2=%.4f'%s2)
8 print('s=%.4f'%s)

```

程序 5.4 计算例 5-4 中样本均值、样本方差、样本均方差的 Python 程序

运行程序，输出如下。

```

x_=1493.0000
s^2=14068.6667
s=118.6114

```

此即样本均值、样本方差和样本均方差的观测值 $\bar{x}$ 、 s^2 和 s 的精确到万分位的值。

练习 5-6 用 Python 验算练习 5-5 的计算结果。

参考答案：见文件 chapter05.ipynb 中对应代码。

5.3 本章附录

A1. 定理 5-1 的证明

证明： (1) 由于 $X_1, X_2, \dots, X_n$ 独立且均服从 $N(\mu, \sigma^2)$ ，根据定理 3-7 的 (1) 知 $\sum_{i=1}^n X_i \sim N(n\mu, n\sigma^2)$ 。再根据练习 3-22 的结论，得 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim N(\mu, \sigma^2/n)$ 。

(2) 根据定理 3-7，需要说明存在相互独立的随机变量 $Z_k \sim N(0, 1)$, $k = 1, 2, \dots, n-1$ ，使得 $\frac{n-1}{\sigma^2} S^2 = Z_1^2 + Z_2^2 + \dots + Z_{n-1}^2$ 。为此，首先注意到变量 $Y_i = \frac{X_i - \mu}{\sigma} \sim N(0, 1)$, $i = 1, 2, \dots, n$ ，且相互独立。

接下来，构造一个 n 阶正交矩阵^① $\mathbf{A} = (a_{ij})_{n \times n}$ ，其中第 n 行元素 $a_{nj} = 1/\sqrt{n}$ ，

$j = 1, 2, \dots, n$ 。记 $\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$ ，令 $\mathbf{Z} = \mathbf{A}\mathbf{Y}$ ，即

$$\begin{cases} Z_1 = a_{11}Y_1 + a_{12}Y_2 + \dots + a_{1n}Y_n \\ Z_2 = a_{21}Y_1 + a_{22}Y_2 + \dots + a_{2n}Y_n \\ \vdots \\ Z_n = a_{n1}Y_1 + a_{n2}Y_2 + \dots + a_{nn}Y_n \end{cases}$$

^① 满足 $\mathbf{A}\mathbf{A}^T = \mathbf{A}^T\mathbf{A} = \mathbf{I}$ 的矩阵 $\mathbf{A}$ ，其中， $\mathbf{A}^T$ 表示 $\mathbf{A}$ 的转置， $\mathbf{I}$ 为单位矩阵（详见参考文献 [7]9.2 节）。