统计学原理(Statistic)

胡华平

西北农林科技大学

经济管理学院数量经济教研室

huhuaping01@hotmail.com

2022-03-26

第二章 数据收集、整理和清洗

2.1 数据目标

2.2 数据收集

2.3 资科整理和数据清洗

2.4 数据的数据库化

.2.5 数据质量

2.6 抽样设计

white[2.7 抽样分布和抽样误差]

2.8 问卷设计技术

2.7 抽样分布和抽样误差

离散和连续随机变量

总体和样本特征

抽样误差计算

离散随机变量：离散事件

六点骰子的样本空间（sample space）为：$\{1, 2, 3, 4, 5, 6\}$，随机摇一次骰子结果可能是：

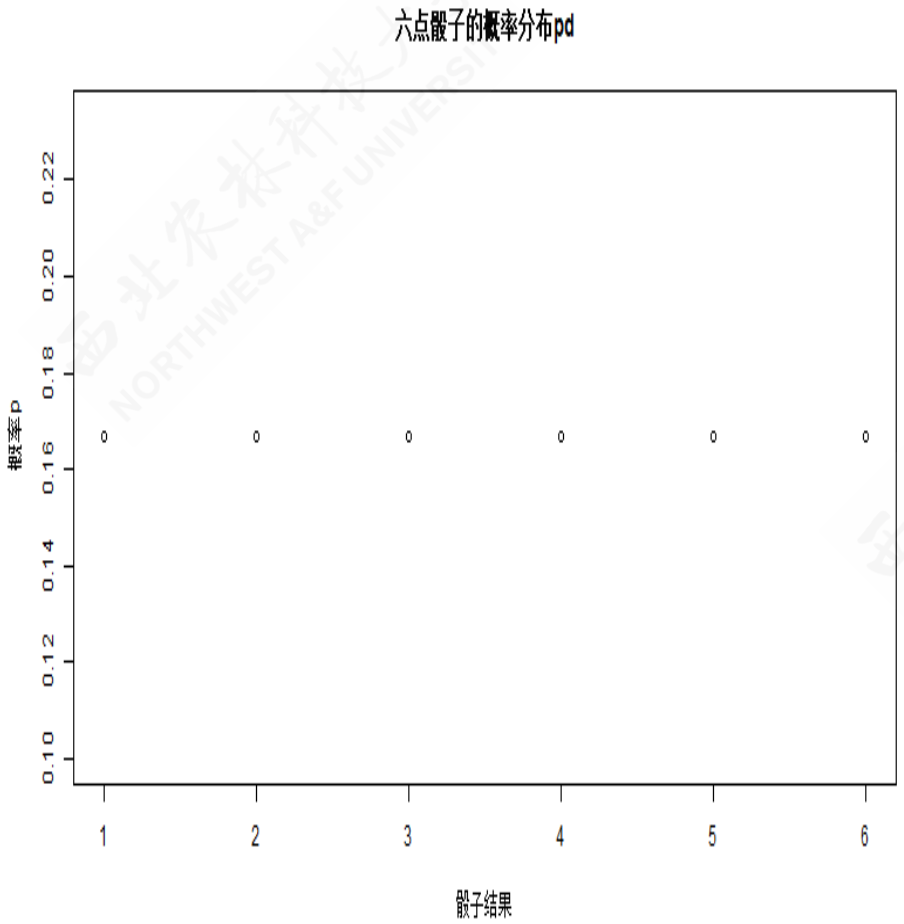
```
sample(1:6, 1)
```

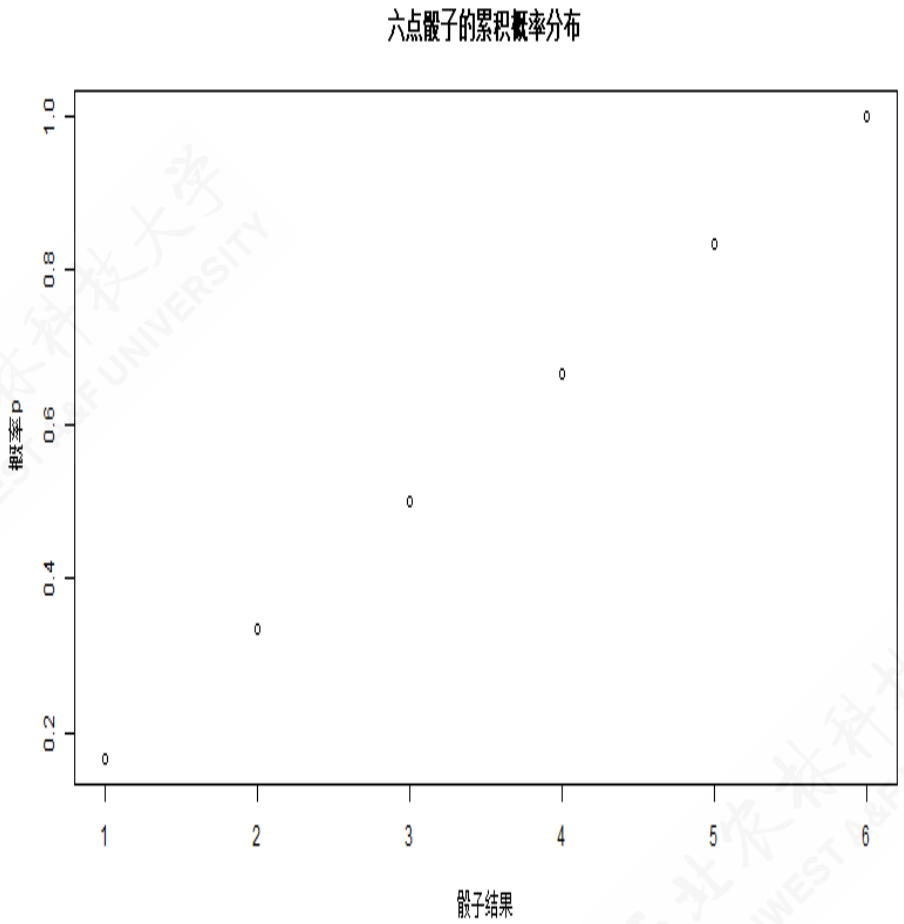
```
[1] 6
```

打六点骰子的PDF和CDF

骰子结果	1	2	3	4	5	6
概率	1/6	1/6	1/6	1/6	1/6	1/6
累积概率	1/6	2/6	3/6	4/6	5/6	1

离散随机变量：概率分布


六点骰子的概率分布pd
概率p
0.10
0.12
0.14
0.16
0.18
0.20
0.22
1
2
3
4
5
6
骰子结果


六点骰子的累积概率分布
概率p
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
骰子结果

离散随机变量：伯努利事件（概率分布）

抛硬币事件k有两种可能结果：H（头像）和T（花案）。我们随机抛一次硬币的结果可能是：

```
sample(c("H", "T"), 1)
```

```
[1] "H"
```

对于连续n次抛硬币，事件k服从伯努利$k \sim B(n,p)$分布，其概率为：

$$f(k) = P(k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} = \frac{n!}{k!(n-k)!} \cdot p^k \cdot (1-p)^{n-k}$$

离散随机变量：伯努利事件（概率分布）

例如，连续抛10次硬币且其中5次为头像朝上的伯努利概率记为$P(k=5|n=10,p=0.5)$，具体计算R函数及其结果为：

```
p_k5 <- dbinom(x = 5,size = 10, prob =
p_k5
```

```
[1] 0.25
```

连续抛10次硬币（$n=10$）且其中5次为头像朝上（$k=5$）的伯努利事件（$p=0.5$）出现的概率为24.61%。

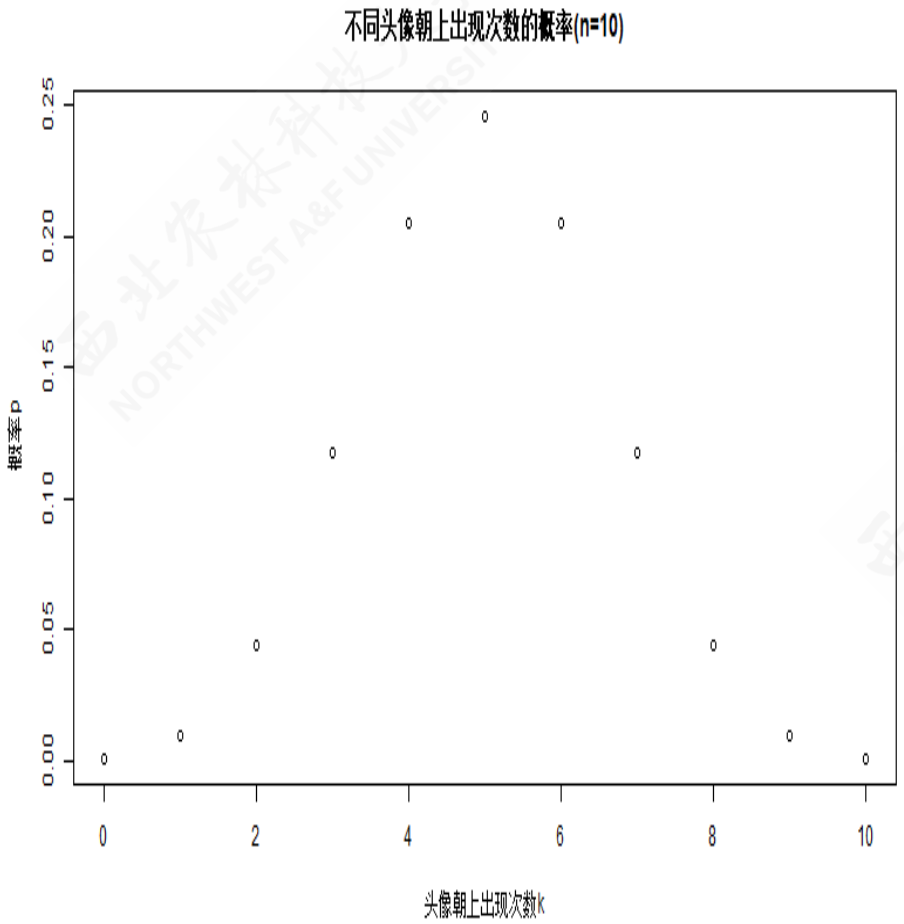
例如，连续抛10次硬币且其中头像朝上次数在$4\sim7$次之间的伯努利概率记为$P(4\leq k\leq 7)=P(k\leq 7)-P(k\leq 3)$，具体计算R函数及其结果为：

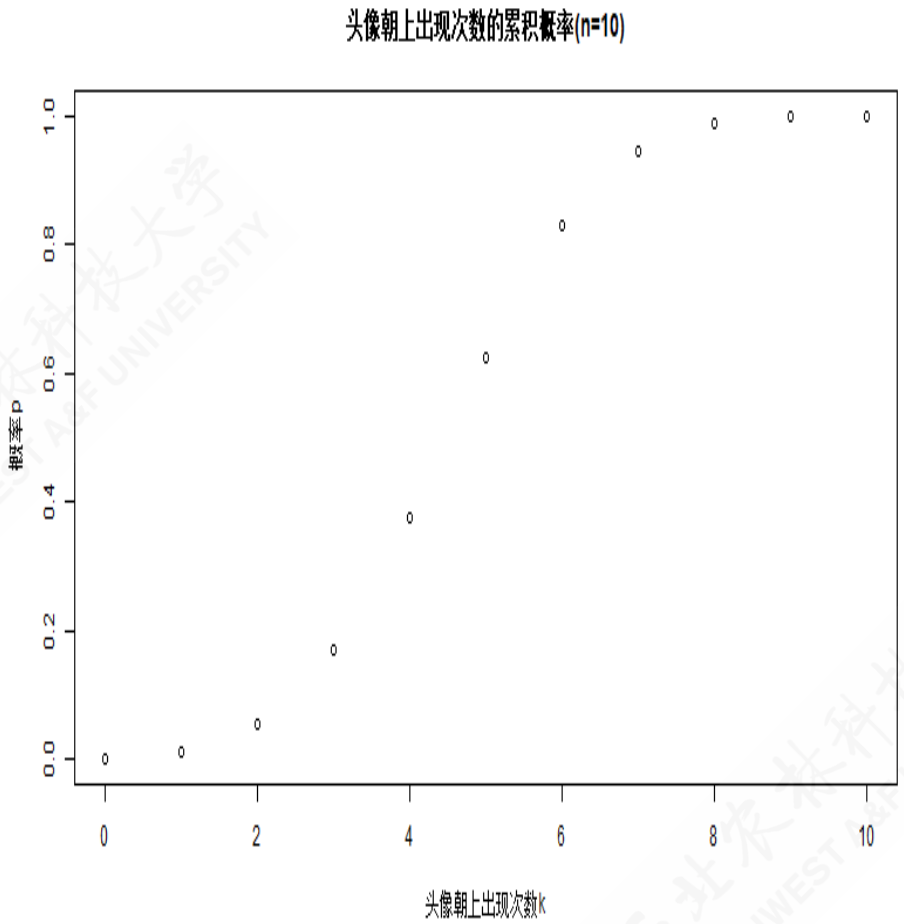
```
p_k47 <- pbinom(size = 10, prob = 0.5
p_k47
```

```
[1] 0.77
```

连续抛10次硬币（$n=10$）且其中头像朝上次数（$4\leq k\leq 7$）的伯努利事件（$p=0.5$）出现的概率为77.34%。

离散随机变量：伯努利事件（概率分布）





连续随机变量：概率、期望和方差

对于一个连续分布事件 X，给定其概率密度函数（PDF）为 $f_X(x)$，那么：

- 其累积概率密度函数（CDF）： $P(a \leq X \leq b) = \int_a^b f_X(x)\mathrm{d}x$
- 而且有完全概率密度为： $P(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = 1$。
- 进一步，其期望为： $E(X) = \mu_X = \int X f_X(x)\mathrm{d}x$。
- 其方差为： $\mathrm{Var}(X) = \sigma_X^2 = \int (X - \mu_X)^2 f_X(x)\mathrm{d}x$

连续随机变量：正态分布（PDF、CDF）

一个变量 X若服从正态分布，则由两个参数来确定，一个是期望 μ，另一个是方差 σ^2，并记为：$Y \sim \mathcal{N}(\mu, \sigma^2)$。正态分布的分布密度函数（PDF）的理论表达式为：

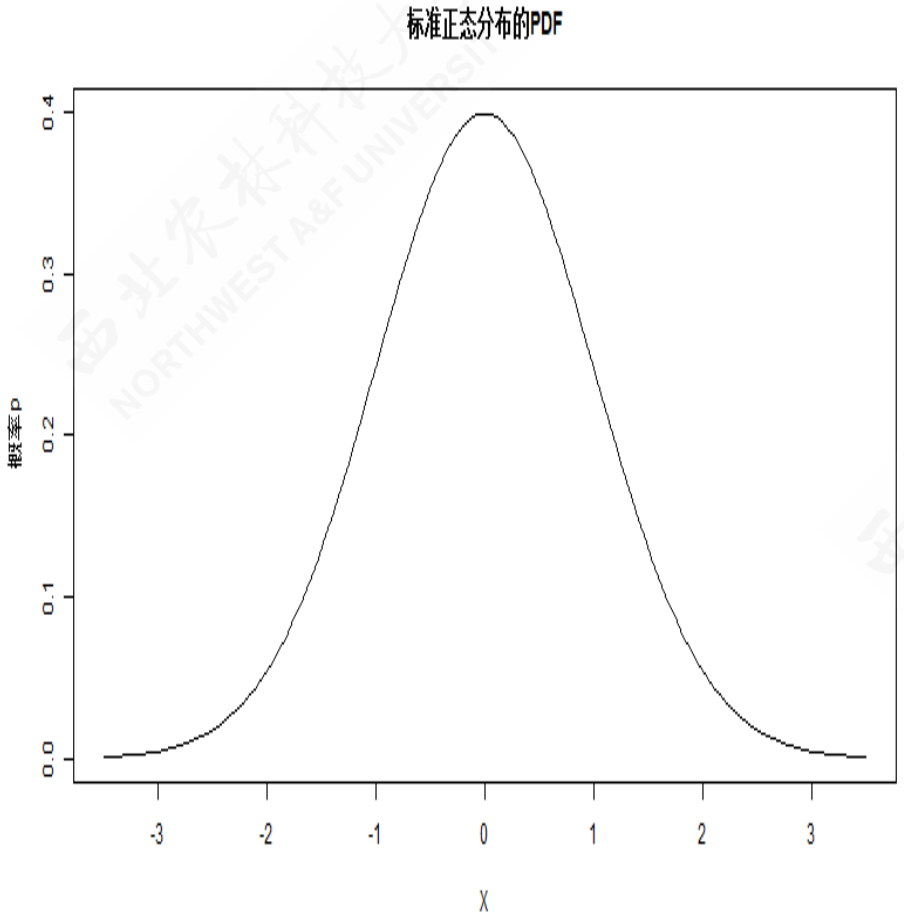
$$f(X) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-(X-\mu)^2/(2\sigma^2)\right].$$

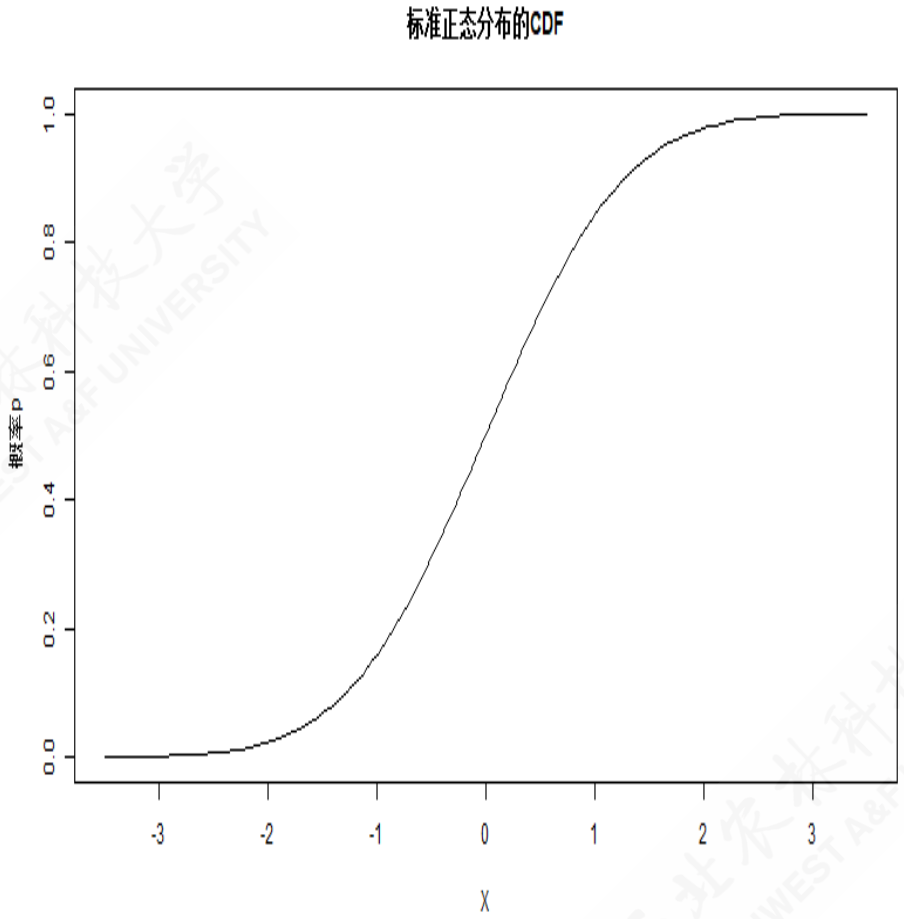
其中，标准正态分布属于一种特殊形式的正态分布，其期望为0，方差为1，一般记为：$X \sim Z(0,1)$，其概率密度函数（PDF）一般记为 ϕ，其累积概率密度函数（CDF）一般记为 Φ，也即有：

$$\phi(c) = \Phi'(c) \ , \ \Phi(c) = P(Z \leq c) \ , \ Z \sim \mathcal{N}(0,1).$$

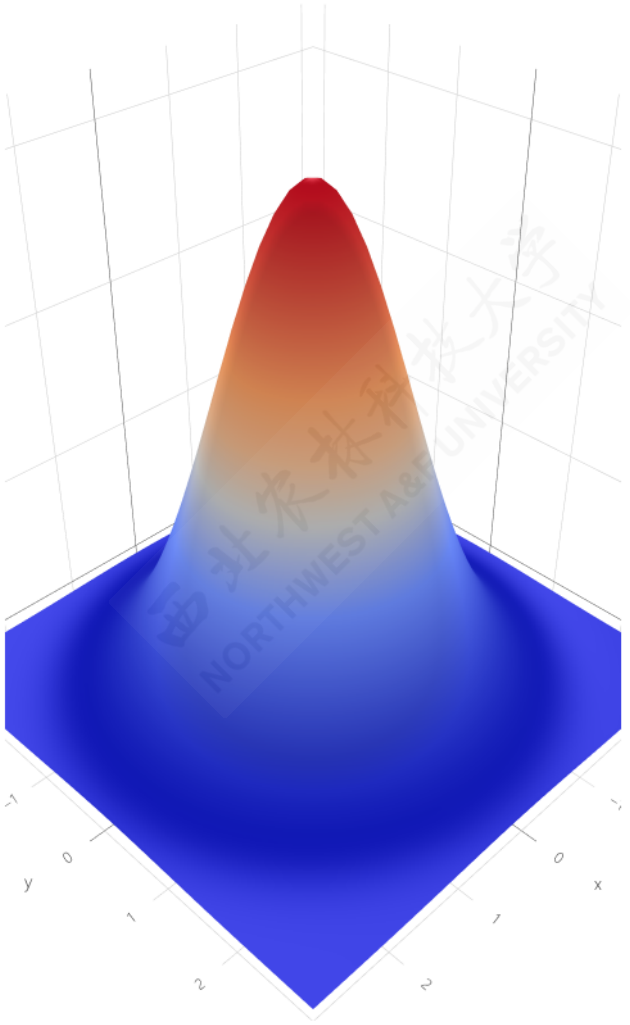
而且还有：若 $Y \sim \mathcal{N}(\mu, \sigma^2)$，则 $Y^* = \frac{Y-\mu}{\sigma} \sim Z(0,1)$。

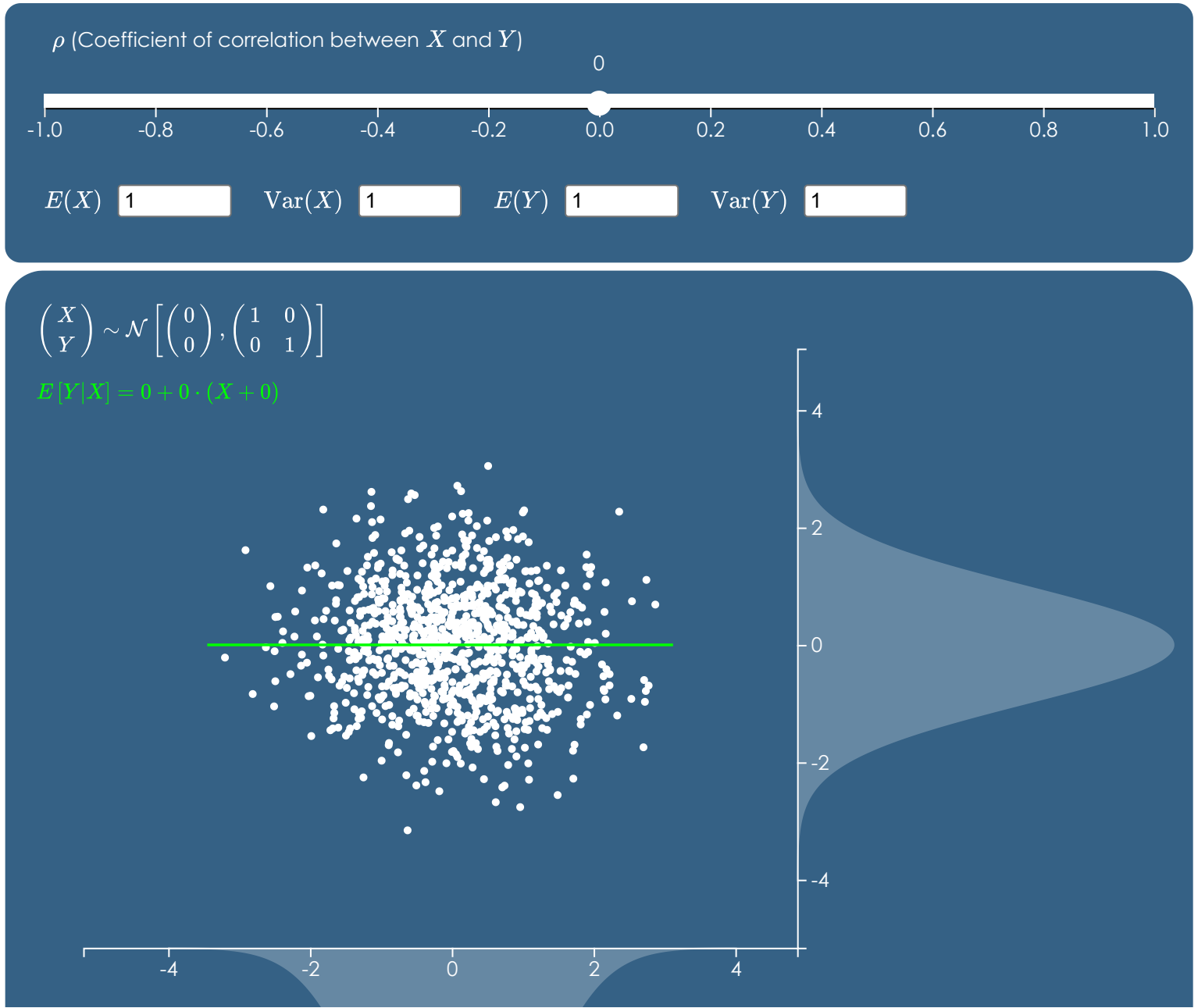
连续随机变量：正态分布（PDF、CDF）





连续随机变量：正态分布（二元联合正态）



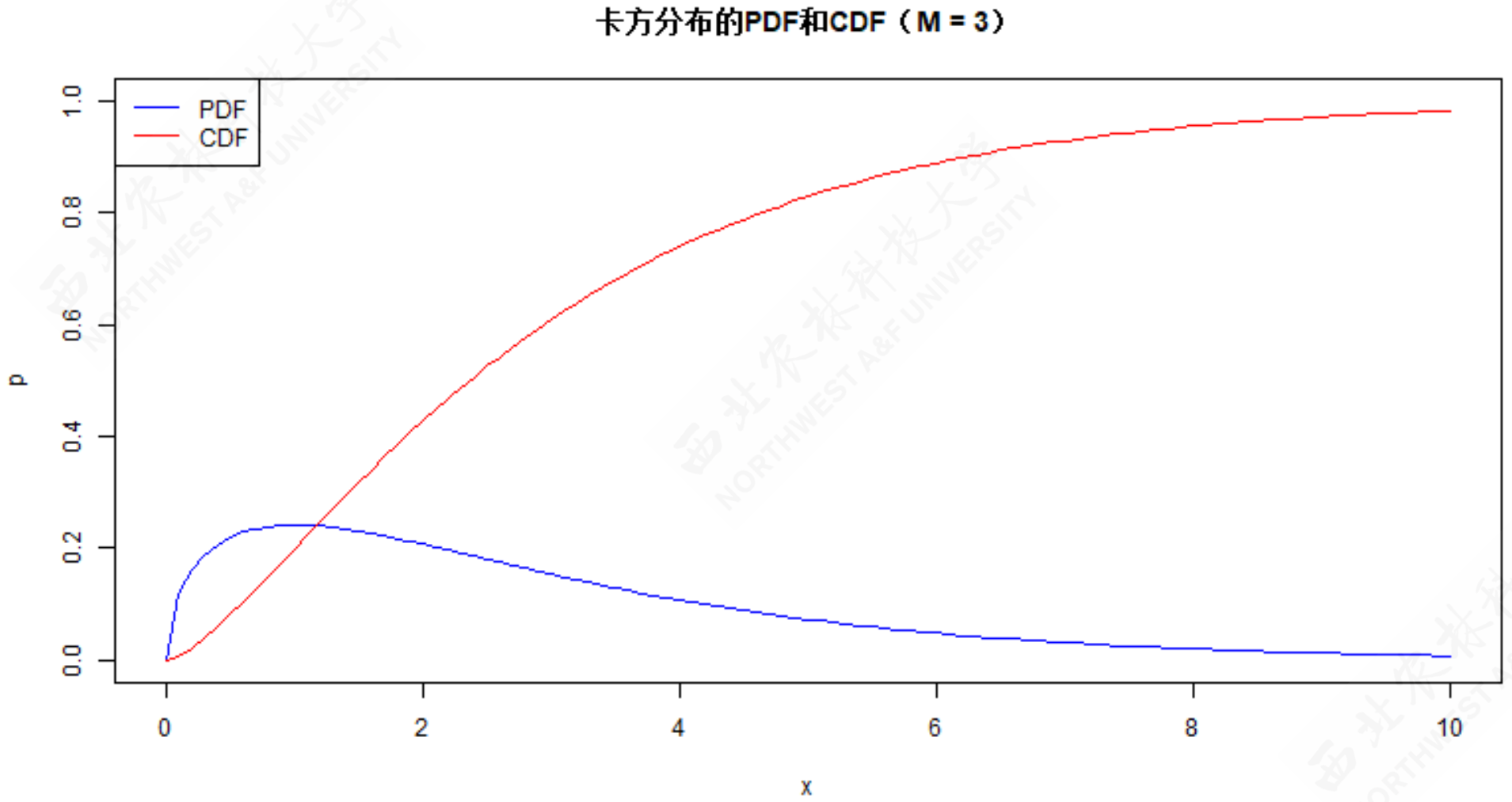

ρ (Coefficient of correlation between X and Y)
0
-1.0
-0.8
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
E(X) 1
Var(X) 1
E(Y) 1
Var(Y) 1
(X Y) ~ N[(0 0), (1 0 0 1)]
E[Y|X] = 0 + 0 · (X + 0)
4
2
0
-2
-4
-4
-2
0
2
4

连续随机变量：卡方分布（PDF、CDF）

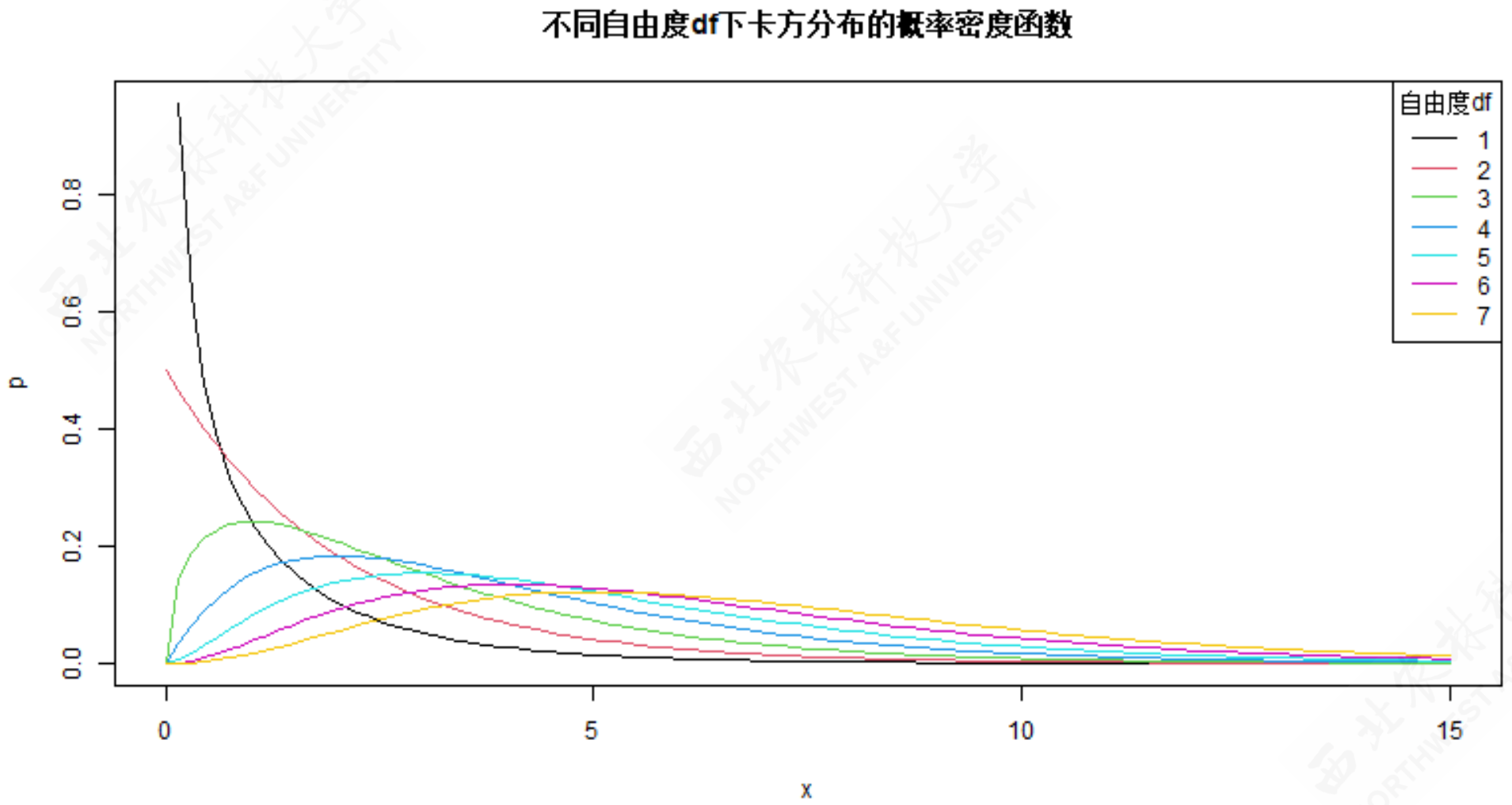
$$if,\ \ Z_m \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$$

$$then,\ \ Z_1^2 + \cdots + Z_M^2 = \sum_{m=1}^{M} Z_m^2 \sim \chi^2(M)$$

连续随机变量：卡方分布（PDF、CDF）



连续随机变量：卡方分布（PDF、CDF）

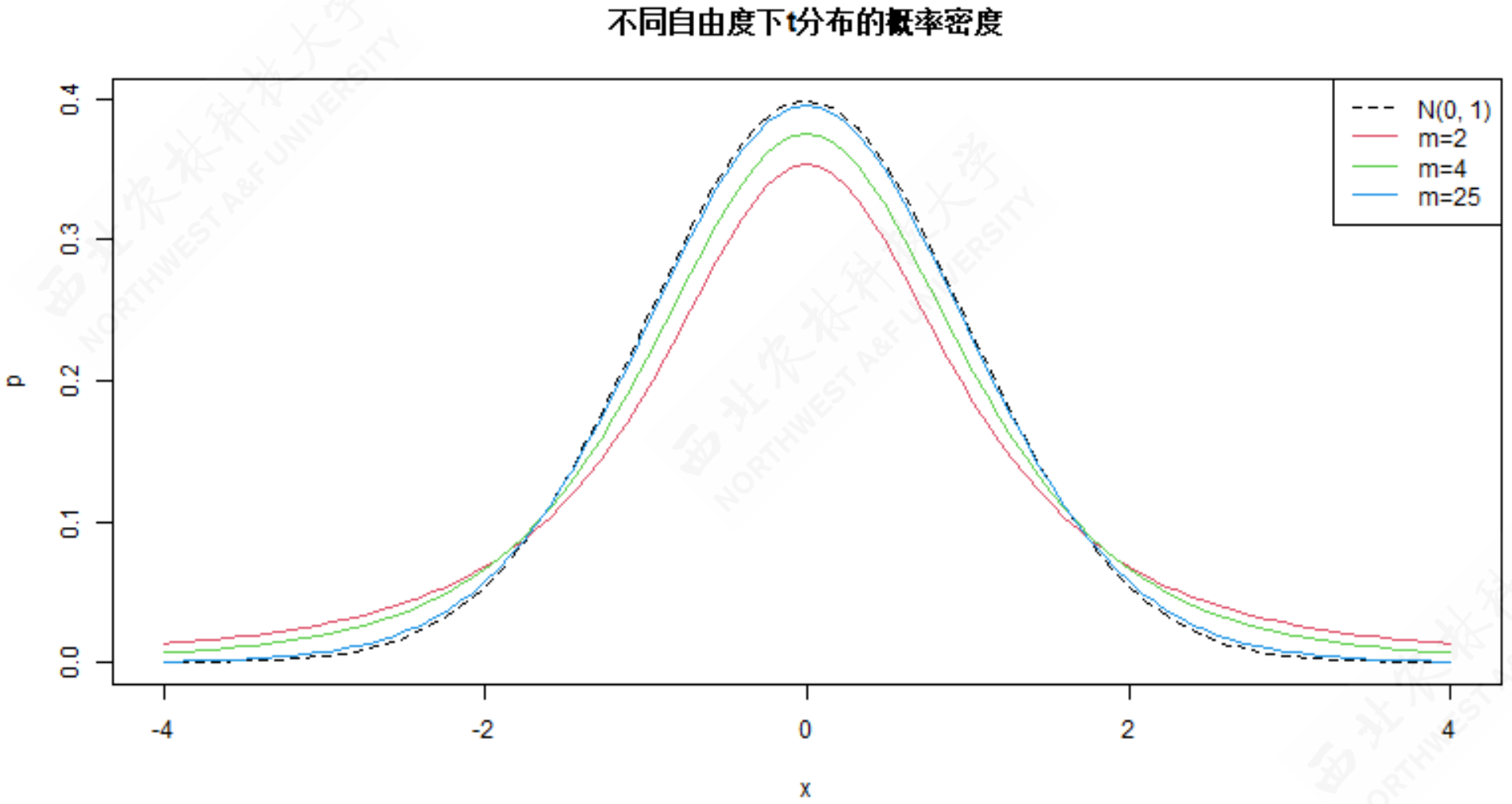


连续随机变量：t分布（PDF、CDF）

假定随机变量 $Z \sim \mathcal{N}(0,1)$服从标准正态分布，随机变量 $W \sim \chi^2(m)$卡方分布，而且二者互相独立，那么可以构造出一个如下的新随机变量 T，它将服从t分布：

$$T = \frac{Z}{\sqrt{W/m}} \sim t(m)$$

连续随机变量：t分布（PDF、CDF）



总体的特征：总体期望和总体方差

随机变量 Y有6种可能取值 $\{1,2,3,4,5,6\}$，那么每种可能取值的概率分别为：

事件Y_i	1	2	3	4	5	6
概率p	1/6	1/6	1/6	1/6	1/6	1/6

总体期望 μ_Y和总体方差 σ_Y^2：

$$E(Y) \equiv \mu_Y = \sum_1^6 (Y_i \cdot p(Y_i)) = \frac{1}{6}\sum_1^6 Y_i = \frac{1}{6} \times 21 = 3.50$$

$$\begin{aligned} Var(Y) \equiv \sigma_Y^2 &= E(Y_i - E(Y))^2 = E(Y_i - \mu)^2 = \sum_1^6 ((Y_i - \mu)p(Y_i)) \\ &= \frac{1}{6}\left[(1-3.5)^2 + (2-3.5)^2 + \cdots + (6-3.5)^2\right] \\ &= 3.89 \end{aligned}$$

样本的特征：样本均值和样本方差

从上述总体中**有放回**地随机抽选8次，得到1份样本容量 $n=8$ 的如下样本数据：

```
[1] 5 1 5 6 3 5 5 3
```

样本均值 $\bar{y}$ 和样本方差 s_y^2 分别表达并计算为：

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i=\frac{1}{8}[5+1+\cdots+3]=4.125$$

$$\begin{aligned}s_y^2&=\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}\\&=\frac{1}{8-1}\times\left[(5-4.125)^2+(1-4.125)^2+\cdots+(3-4.125)^2\right]\\&=2.6964\end{aligned}$$

样本的特征：样本均值和样本方差

因此，我们可以不断从前述总体 $Y \in \{1,2,3,4,5,6\}$ 中进行有放回的随机抽样。下表展示了10份随机样本 y，每份样本的容量都相同 $n=8$。每份样本的均值 $\bar{y}$ 见列`bar_y`，样本方差 s_y^2 见列`s2_y`。

sample	y	bar_y	s2_y
1	c(6, 6, 6, 2, 1, 3, 6, 1)	3.88	5.5536
2	c(6, 3, 1, 5, 1, 3, 3, 6)	3.50	4.0000
3	c(1, 4, 5, 1, 4, 4, 3, 4)	3.25	2.2143
4	c(4, 3, 2, 4, 5, 3, 1, 2)	3.00	1.7143
5	c(2, 2, 5, 3, 5, 3, 5, 4)	3.63	1.6964
6	c(3, 5, 5, 4, 6, 1, 1, 1)	3.25	4.2143

Showing 1 to 6 of 11 entries

Previous 1 2 Next

总体和样本特征的关系

根据**中心极限定理**和**大数定理**，我们可以推导得到总体与样本特征的如下关系：

$$E(\bar{y}) = \mu_Y \tag{eq.1}$$

$$Var(\bar{y}) = \frac{\sigma_Y^2}{n} \tag{eq.2}$$

$$E\left(Var(\bar{y})\right) = \widehat{Var}(\bar{y}) \equiv \frac{s^2}{n} \tag{eq.3}$$

其中，$s^2 = \frac{\sum_1^n (y_i - \bar{y})^2}{n-1}$ 表示随机样本的样本标准差。以上方程蕴含着如下结论：

- 方程1表明：随机变量 $\bar{y}$ 的**期望**是随机变量 Y 的**期望**的**无偏估计量**（unbiased estimator）。
- 方程2表明：随机变量 $\bar{y}$ 的**方差**与随机变量 Y 的**方差**存在以上关系。
- 方程3表明：随机变量 $\bar{y}$ 的**方差**的无偏估计量可以通过样本数据计算得到，其结果为 $\widehat{Var}(\bar{y}) \equiv \frac{s^2}{n}$。

抽样分布：骰子游戏

随机样本（random sampling）是从总体中随机抽取个体的集合。

六点骰子的可能结果为 $\{1, 2, 3, 4, 5, 6\}$，如果随机投掷2次，可以得到2次结果的数值加总：

```
set.seed(520)
toll_2 <- sample(1:6, 2, replace = T)
toll_2
```

```
[1] 3 6
```

```
sum(toll_2)
```

```
[1] 9
```

抽样分布：骰子游戏

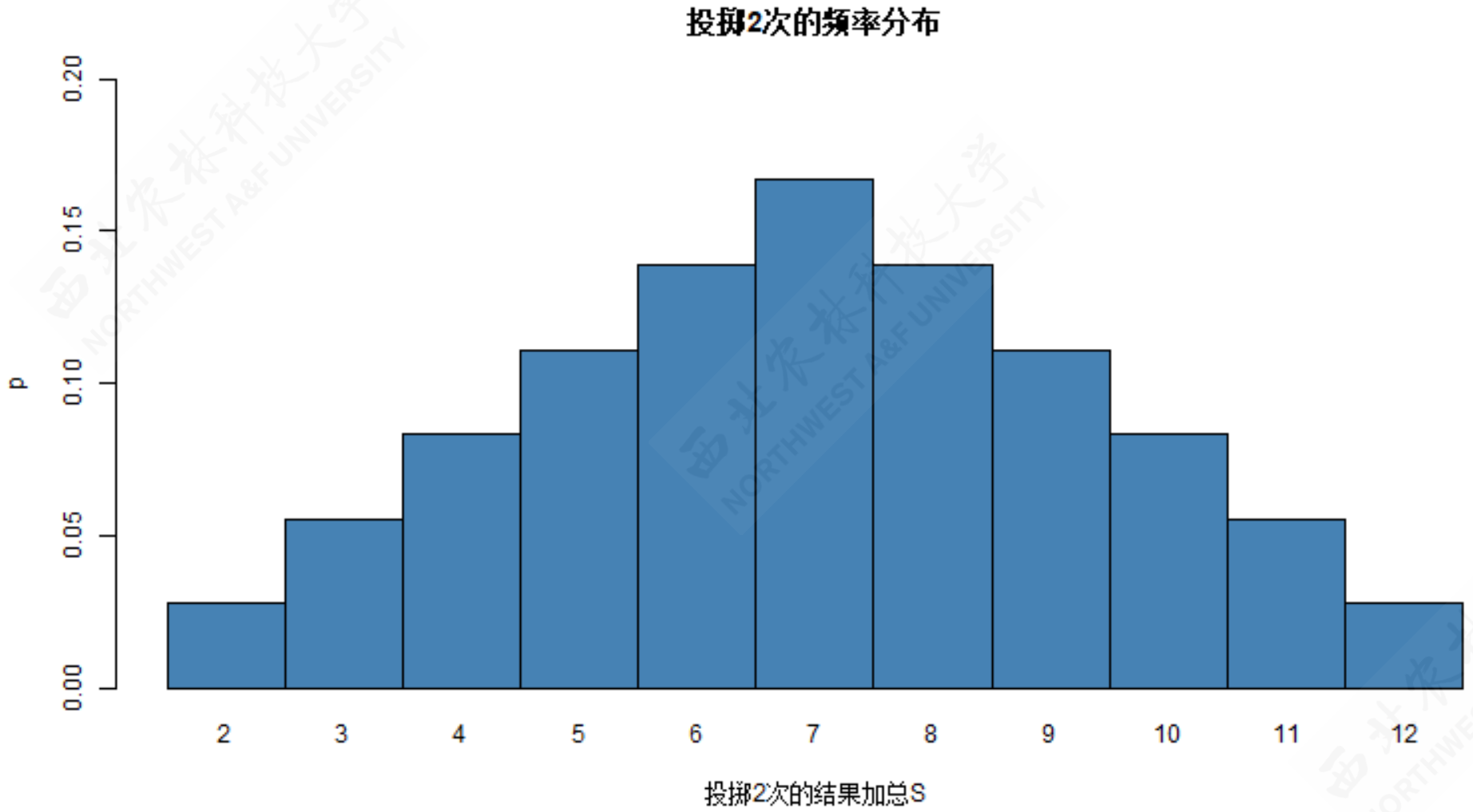
随机投掷2次的所有可能结果共有$6^2 = 36$种可能：

$$
\begin{array}{l}
(1,1)(1,2)(1,3)(1,4)(1,5)(1,6) \\
(2,1)(2,2)(2,3)(2,4)(2,5)(2,6) \\
(3,1)(3,2)(3,3)(3,4)(3,5)(3,6) \\
(4,1)(4,2)(4,3)(4,4)(4,5)(4,6) \\
(5,1)(5,2)(5,3)(5,4)(5,5)(5,6) \\
(6,1)(6,2)(6,3)(6,4)(6,5)(6,6)
\end{array}
$$

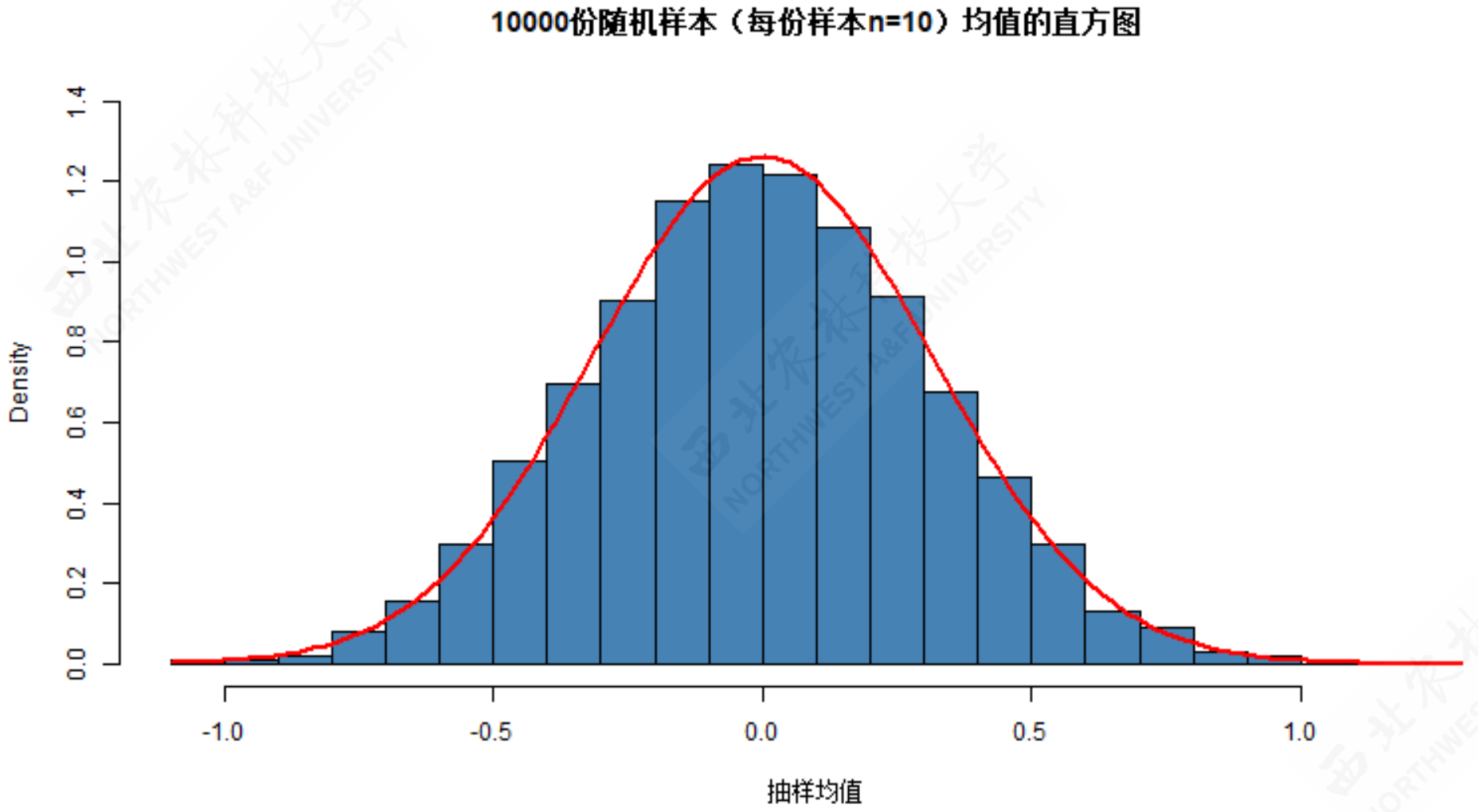
以上的全部组合，共有11种加总结果$S = \{2,3,4,5,6,7,8,9,10,11,12\}$，每种加总结果的概率分别是：

$$
P(S) = \begin{cases}
1/36, & S = 2 \\
2/36, & S = 3 \\
3/36, & S = 4 \\
4/36, & S = 5 \\
5/36, & S = 6 \\
6/36, & S = 7 \\
5/36, & S = 8 \\
4/36, & S = 9 \\
3/36, & S = 10 \\
2/36, & S = 11 \\
1/36, & S = 12
\end{cases}
$$

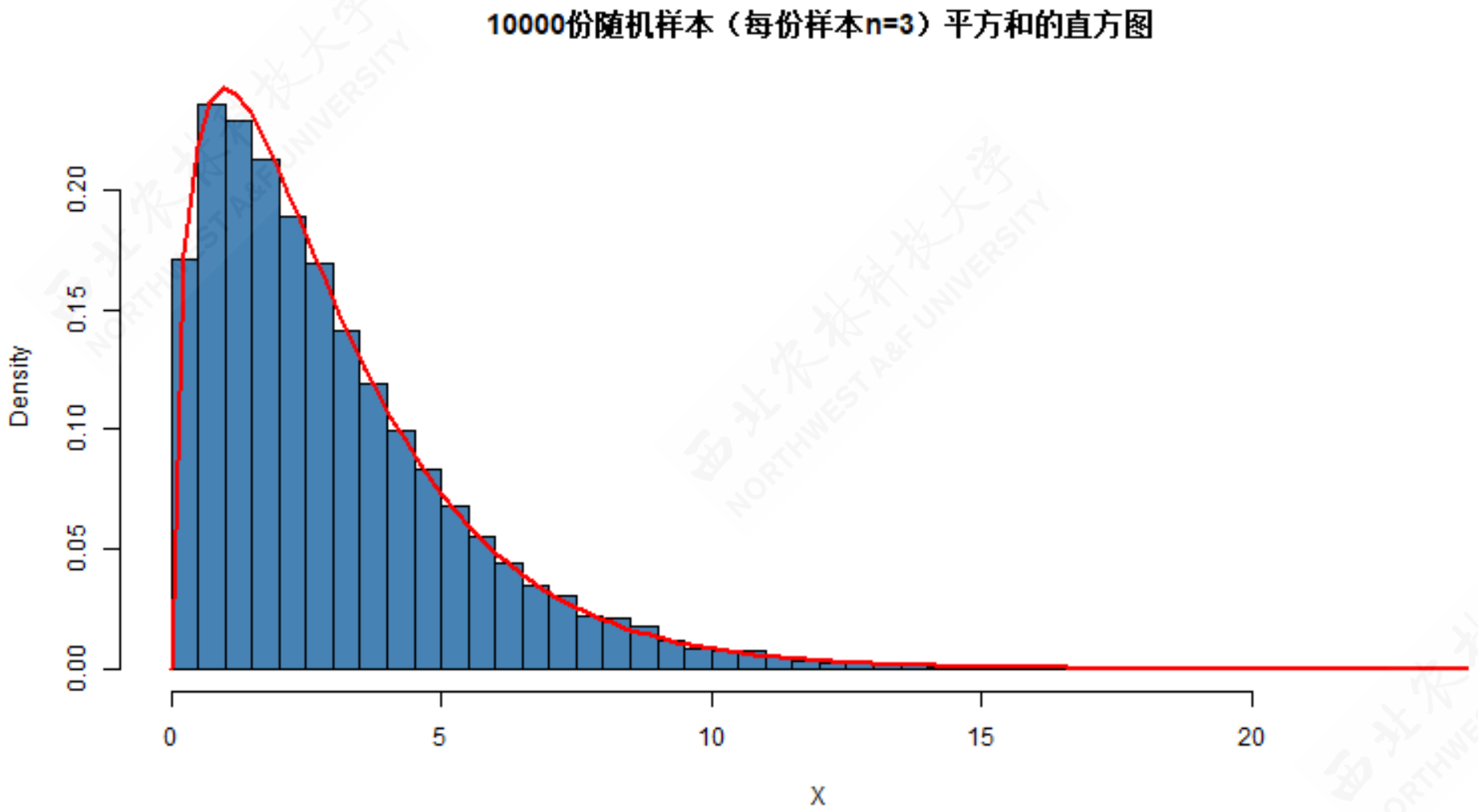
抽样分布：骰子游戏



抽样分布：骰子游戏


10000份随机样本（每份样本n=10）均值的直方图
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
-1.0
-0.5
0.0
0.5
1.0
抽样均值

抽样分布：骰子游戏


10000份随机样本（每份样本n=3）平方和的直方图
Density
0.00
0.05
0.10
0.15
0.20
0
5
10
15
20
X

抽样误差：均值和方差

假定随机样本 $y_1, \ldots, y_n$ 是独立随机抽取自正态分布总体 $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$，那么前述随机样本数据的均值 $\bar{y}$ 将服从如下正态分布：

$$\bar{y} \sim \mathcal{N}(\mu_Y, \sigma_Y^2/n) \tag{2.4}$$

其中：

$$E(\bar{y}) \equiv \mu_{\bar{y}} = E\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(y_i\right) = \frac{1}{n} \cdot n \cdot \mu_Y = \mu_Y$$

$$\begin{aligned}\mathrm{Var}(\bar{y}) \equiv \sigma_{\bar{y}}^2 &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(y_i) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n}\mathrm{cov}(y_i, y_j) = \frac{\sigma_Y^2}{n}\end{aligned}$$

抽样误差：中心极限定理

> 然而，实际中我们往往并不知道总体方差 σ_Y^2。此时，上述方差公式是不能够计算的。

有限总体中心极限定理（The finite population Central Limit Theorem）对于随机变量 $\bar{y}$的意义在于：我们可以用样本方差 s_y^2来近似替代总体方差 σ_Y^2。也即：

$$\mathrm{Var}(\bar{y}) \equiv \sigma_{\bar{y}}^2 = \frac{\sigma_Y^2}{n}$$

$$\widehat{Var}(\bar{y}) \equiv \hat{\sigma}_{\bar{y}} = \frac{s_y^2}{n}$$

抽样误差：中心极限定理

如果样本容量 n很小，随机变量 $\bar{y}$的可能分布会是多种多样的。**有限总体中心极限定理**表明，随着样本容量 n的不断增大，随机变量 $\bar{y}$的分布会越来越稳定，并趋向于**正态分布**（normal distribution），从而有：

$$\bar{y} \sim \mathcal{N}(\mu_{\bar{y}}, \sigma^2_{\bar{y}})$$

$$\frac{\bar{y} - \mu_{\bar{y}}}{\sigma_{\bar{y}}} = \frac{\bar{y} - \mu_{\bar{y}}}{\sqrt{Var(\bar{y})}} \sim \mathcal{Z}(0, 1)$$

抽样误差：置信区间

如果随机变量 $\bar{y}$的总体方差 $Var(\bar{y})$未知，则无法使用上述正态分布 $\mathcal{N}$或者标准正态 $\mathcal{Z}$分布，进行有关置信区间的样本推断。幸运的是，我们可以构造出如下服从t分布的随机变量：

$$\frac{\bar{y}-\mu_{\bar{y}}}{\hat{\sigma}_{\bar{y}}}=\frac{\bar{y}-\mu_{\bar{y}}}{\sqrt{\widehat{Var}(\bar{y})}}=\frac{\bar{y}-\mu_{\bar{y}}}{s_y/\sqrt{n}}\sim t(n-1)$$

因此可以进一步得到参数 $\mu_{\bar{y}}$的 $1-\alpha$置信区间：

$$\bar{y}-t_{1-\alpha/2}\sqrt{\widehat{Var}(\bar{y})}\leq\mu_{\bar{y}}\leq\bar{y}+t_{1-\alpha/2}\sqrt{\widehat{Var}(\bar{y})}$$

对于有放回的简单随机抽样，参数 $\mu_{\bar{y}}$的 $1-\alpha$置信区间具体为：

$$\bar{y}-t_{\alpha/2}\sqrt{\left(\frac{N-n}{N}\right)\left(\frac{s^2}{n}\right)}\leq\mu_{\bar{y}}\leq\bar{y}+t_{\alpha/2}\sqrt{\left(\frac{N-n}{N}\right)\left(\frac{s^2}{n}\right)}$$

抽样误差：简单随机抽样

对于**无放回**的简单随机抽样方案，采用**无偏估计法**（unbiased estimator）下的均值 $\bar{y}_{\text{st}}$ 和方差 $\widehat{\text{var}}\left(\bar{y}_{\text{st}}\right)$ 分别为：

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

$$\widehat{\text{Var}}\left(\hat{\mu}\right) = \frac{N-n}{N} \cdot \frac{s_y^2}{n}$$

上述方差公式中，$s_y^2 = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}{n-1}$。而 $\frac{N-n}{N}$ 又被称为**有限总体校正比值**（finite population correction）：

- 如果采用**有放回**的简单随机抽样，则上述方差公式需要去掉**有限总体校正比值**。
- 如果采用**无放回**的简单随机抽样，但是n相对于N非常小，则上述方差公式中**有限总体校正比值**会接近于1，因此也可以忽略。

（示例）田野甲虫数量案例

案例说明：为估计出一片农地中甲虫的总数。研究人员将农地细分为100个大小相等的区块。

研究者决定采用**简单随机抽样**方案，随机抽选了其中的8个区块（编号见列`field`），并分别统计出其中的甲虫数量（见列`beetles`）。最终抽样统计表见右：

简单随机抽样结果

field	beetles
41	234
42	256
18	128
13	245
80	211
68	240
25	202
100	267

（示例）简单随机抽样下估计期望和方差：计算结果

根据案例，容易计算得到：全部区块数量 $N=100$；抽选区块数量 $n=8$。抽选区块下甲虫数量的样本方差为 $s_y^2=\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}=1932.70$。

因此根据简单随机抽样**无偏估计法**下的计算公式，分别可以计算得到估计的均值 $\hat{\mu}$和方差 $\widehat{var}(\hat{\mu})$分别为：

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}y_i=\bar{y}=222.88$$

$$\begin{aligned}\widehat{\mathrm{Var}}\left(\hat{\mu}\right)&=\frac{N-n}{N}\cdot\frac{s_y^2}{n}\\&=\frac{100-8}{100}\cdot\frac{1932.70}{8}=222.2601\end{aligned}$$

（示例）简单随机抽样下估计期望和方差：计算结果

根据上述计算，给定 $\alpha = 0.05$下，平均每个区块甲虫数 $\mu_{\bar{y}}$的置信区间计算结果为：

$$\hat{\mu} - t_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{\mu})} \leq \mu_{\bar{y}} \leq \hat{\mu} + t_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{\mu})}$$

$$222.88 - 2.36 \times \sqrt{222.2601} \leq \mu_{\bar{y}} \leq 222.88 + 2.36 \times \sqrt{222.2601}$$

$$187.62 \leq \mu_{\bar{y}} \leq 258.13$$

> **思考提问**：全部地块的甲虫数量和置信区间是多少？

说明：此时，t查表值为 $t_{1-\alpha/2}(n-1) = t_{0.975}(8-1) = 2.36$。

必要样本数

> 不管采用哪种抽样方法，在哪一层抽样，在哪个阶段抽样， 到底要抽多少样本合适啊？

假定 $\hat{\sigma}$是参数 σ的无偏、正态估计量。则有

$$\frac{\hat{\theta}-\theta}{\sqrt{\mathrm{Var}(\hat{\theta})}} \sim \mathcal{Z}(0,1)$$

$$P\left(\frac{|\hat{\theta}-\theta|}{\sqrt{\mathrm{Var}(\hat{\theta})}} > \mathcal{Z}_{1-\alpha/2}\right) = \alpha$$

$$P\left(|\hat{\theta}-\theta| > \mathcal{Z}_{1-\alpha/2} \cdot \sqrt{\mathrm{Var}(\hat{\theta})}\right) = \alpha$$

必要样本数

令 $d = |\bar{y} - \mu|$ 为**抽样极限误差**，则**简单随机抽样**（不放回）方案下**必要样本数**的计算公式为：

$$P\left(|\bar{y} - \mu_{\bar{y}}| > \mathcal{Z}_{1-\alpha/2} \cdot \sqrt{\frac{N-n}{N} \cdot \frac{\sigma^2}{n}}\right) = \alpha$$

$$\mathcal{Z}_{1-\alpha/2}\sqrt{\frac{N-n}{N} \cdot \frac{\sigma^2}{n}} \equiv d$$

$$n = \frac{1}{\frac{d^2}{\mathcal{Z}^2_{1-\alpha/2} \cdot \sigma^2} + \frac{1}{N}}$$

（示例）简单随机抽样方案下必要样本数的计算

在前述甲虫数量案例中，给定 $\alpha = 0.05$下，且抽样极限误差不超过1000只，请计算简单随机抽样方案下的必要样本数？

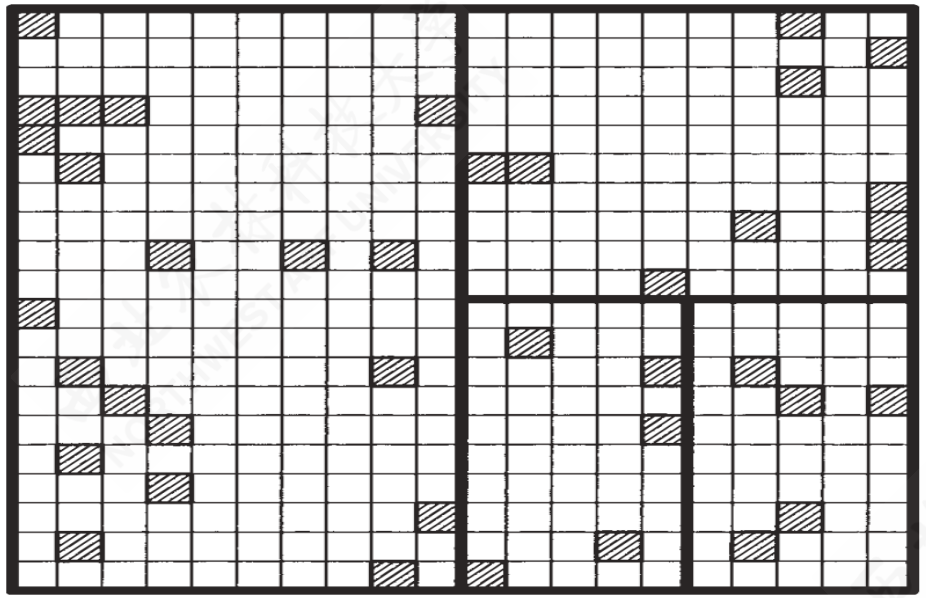
解答：根据案例已知 $N = 100$，抽样极限误差为 $d = 1000$，给定 $\alpha = 0.05$下 $\mathcal{Z}_{1-\alpha/2} = \mathcal{Z}_{0.975} = 1.96$。

因为我们不知道总体方差 $\sigma_{\bar{y}}^2$，但是可以使用样本方差 $s_{\bar{y}}^2 = 1932.70$进行替代

$$n = \frac{1}{\frac{d^2}{N^2 \cdot \mathcal{Z}_{1-\alpha/2}^2 \cdot \sigma^2} + \frac{1}{N}} = \frac{1}{\frac{d^2}{N^2 \cdot \mathcal{Z}_{1-\alpha/2}^2 \cdot s_{\bar{y}}^2} + \frac{1}{N}}$$

$$= \frac{1}{\frac{(1000)^2}{(100)^2 \cdot (1.96)^2 \cdot 1932.70} + \frac{1}{100}} = 42.61 \doteq 43$$

抽样误差：分层抽样



- 分层数量： $L=4$；各个分层的单位数：

  $N_1=200, N_2=100, N_3=N_4=50$；

  全体单位总数

  $N=\sum_{h=1}^{L} N_h=200+100+50+50=400$

  。
- 各个分层的抽样单位数：

  $n_1=20, n_2=10, n_3=n_4=5$；全部抽样总数

  $n=\sum_{h=1}^{h} n_L=20+10+5+5=40$

  。

抽样误差：分层抽样

分层抽样的均值 $\hat{\mu}_{\text{st}}$ 和方差 $\widehat{\text{Var}}\left(\hat{\mu}_{\text{st}}\right)$ 分别为：

$$\hat{\mu}_{\text{st}} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{y}_h$$

$$\widehat{\text{var}}\left(\hat{\mu}_{\text{st}}\right) = \sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2\left(\frac{N_h - n_h}{N_h}\right)\frac{s_h^2}{n_h}$$

其中：

- L 分层数量； N_h 表示第 h 个分层的所有单位数，其中 $h \in \{1, 2, \cdots, L\}$；$N = N_1 + N_2 + \cdots + N_L$ 为所有单位数。n_h 表示第 h 个分层的抽样数；$n = n_1 + n_2 + \cdots + n_L$ 为所有抽样单位总数。
- 各个分层的样本方差为：$s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}\left(y_{hi} - \bar{y}_h\right)^2$； $\bar{y}_h$ 表示各个分层的样本均值。

（示例）家庭观看电视时长案例

案例说明：一家广告公司为了有针对性地在某个县投放电视广告，公司决定进行抽样调查，以估计该县家庭每周观看电视的平均小时数。该县有两个镇区A和镇区B，还有农村区域C。A区建在一家工厂周围，大多数家庭都有带学龄儿童的工厂工人。B区主要是退休人员，C区主要是农民。A区有155户，B区有62户，C区有93户。公司决定从A区抽选20户，B区抽选8户，C区抽选12户。具体抽样结果如下：

Stratification	sampling	n_h	N_h
Town A	35, 43, 36, 39, 28, 28, 29, 25, 38, 27, 26, 32, 29, 40, 35, 41, 37, 31, 45, 34	20	155
Town B	27, 15, 4, 41, 49, 25, 10, 30	8	62
Rural Area C	8, 14, 12, 15, 30, 32, 21, 20, 34, 7, 11, 24	12	93

（示例）分层抽样下的抽样误差：计算结果

各分层的计算表如下：

Stratification	n_h	N_h	mean_h	sd_h	var_p1	var_p2	var_p3	var_all
Town A	20	155	33.90	5.95	0.25	0.87	1.77	0.38
Town B	8	62	25.13	15.25	0.04	0.87	29.05	1.01
Rural Area C	12	93	19.00	9.36	0.09	0.87	7.30	0.57

根据该分层抽样，估计的总体均值（该县住户平均收看电视时间）结果为：

$$
\begin{aligned}
\hat{\mu}_{st} &= \frac{1}{N}(N_1\bar{y}_1 + N_2\bar{y}_2 + N_3\bar{y}_3) \\
&= \frac{1}{155+62+93}[(155\times 33.9) + (62\times 25.12) + (93\times 19.0)] \\
&= 27.7
\end{aligned}
$$

（示例）分层抽样下的抽样误差：计算结果

各分层的计算表如下：

Stratification	n_h	N_h	mean_h	sd_h	var_p1	var_p2	var_p3	var_all
Town A	20	155	33.90	5.95	0.25	0.87	1.77	0.38
Town B	8	62	25.13	15.25	0.04	0.87	29.05	1.01
Rural Area C	12	93	19.00	9.36	0.09	0.87	7.30	0.57

根据该分层抽样，上述关于的总体均值估计（住户平均收看电视时间）的方差为：

$$\widehat{Var}\left(\hat{\mu}_{st}\right)=\sum_{h=1}^{3}\left(\frac{N_h}{N}\right)^2\left(\frac{N_h-n_h}{N_h}\right)\frac{s_h^2}{n_h}=\frac{1}{(310)^2}\left[\left((155)^2\cdot\frac{(155-20)}{155}\cdot\frac{(5.95)^2}{20}\right)\right.$$
$$\left.+\left((62)^2\cdot\frac{(62-8)}{62}\cdot\frac{(15.25)^2}{8}\right)+\left((93)^2\cdot\frac{(93-12)}{93}\cdot\frac{(9.36)^2}{12}\right)\right]=1.97$$

（示例）分层抽样下的抽样误差：计算结果

各分层的计算表如下：

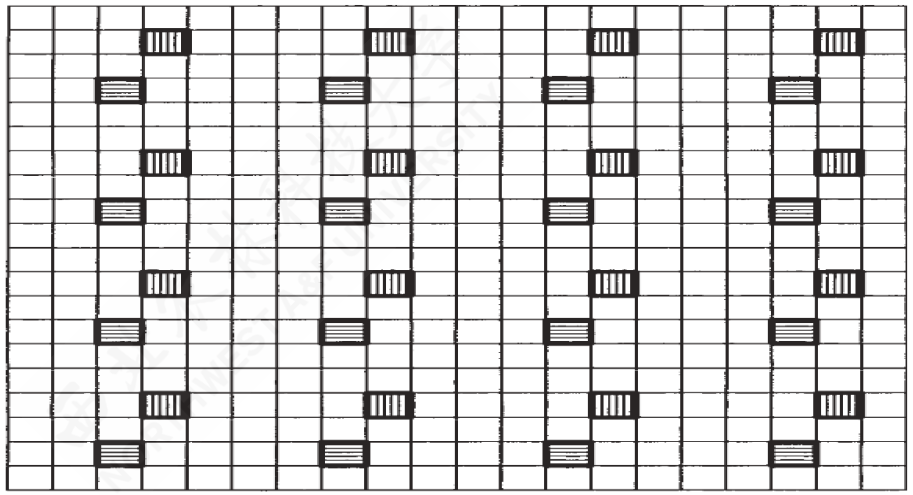
Stratification	n_h	N_h	mean_h	sd_h	var_p1	var_p2	var_p3	var_all
Town A	20	155	33.90	5.95	0.25	0.87	1.77	0.38
Town B	8	62	25.13	15.25	0.04	0.87	29.05	1.01
Rural Area C	12	93	19.00	9.36	0.09	0.87	7.30	0.57

根据该分层抽样，上述关于的总体均值估计（住户平均收看电视时间）的95%置信区间为（t查表值为 $t_{1-0.05/2}(39)=2.02$）*：

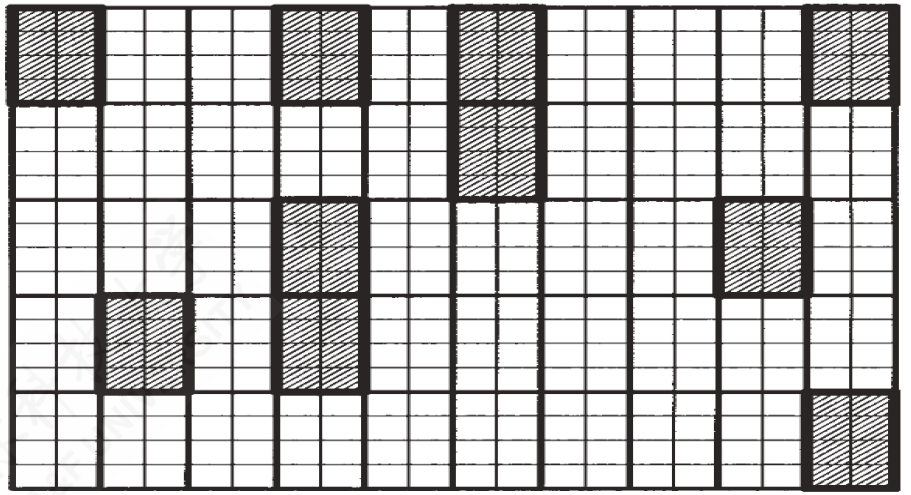
$$\hat{\mu}_{st} \pm t_{1-\alpha/2}(df) \cdot \sqrt{\widehat{Var}\left(\hat{\mu}_{st}\right)}$$
$$=27.7 \pm 2.02 \times \sqrt{1.97} = 27.7 \pm 2.84$$

说明：* 如果不是等比例分层抽取，那么自由度的确定需要用到一个计算公式

抽样误差：系统抽样和整群抽样的关系



系统抽样示例

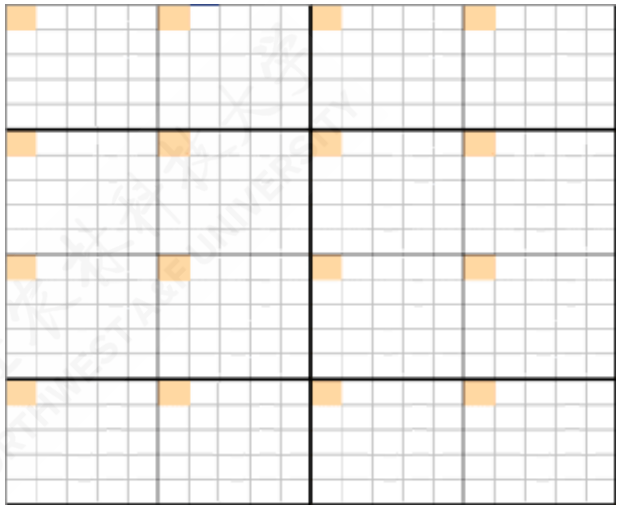


整群抽样示例

从表面上看，**系统抽样（systematic sampling）**和**整群抽样（cluster sampling）**非常不同。实际上，这两种方式具有相同的抽样结构：

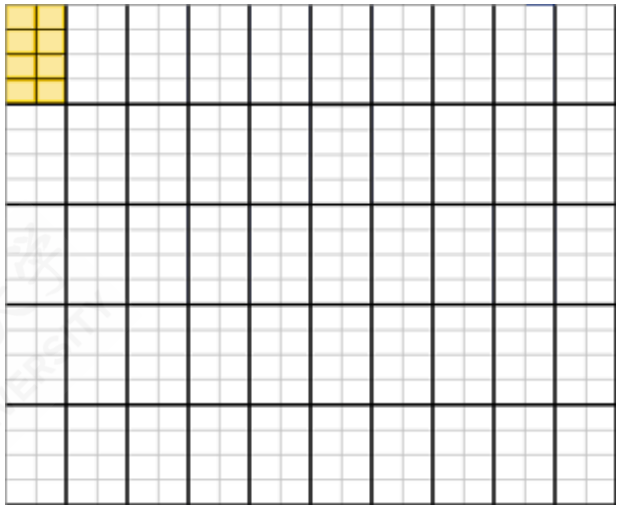
- 利用**主要抽样单位（PSU）**划分群组，而每个主要抽样单位又是由**次要抽样单位（SSU）**组成。
- 如果主要抽样单位（PSU）被随机抽中，则其所有次要单位（SSU）的 y值将都会被抽中。

抽样误差：系统抽样和整群抽样的关系



系统抽样示例

- 该案例共有25个主要抽样单位（PSU）：每个5*5中型方框都有25个小格。
- 着色色区块是一次典型的随机抽取的**系统抽样**结果，共抽取样本数(n=16)。

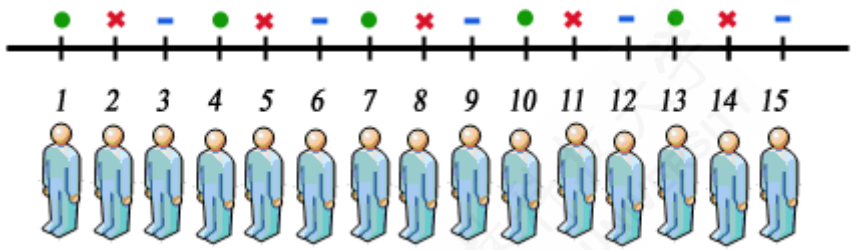


整群抽样示例

- 该案例共有50个主要抽样单位（PSU）：共有50个2*4型方框。
- 着色色区块是一次典型的随机抽取的**整群抽样**结果，共抽取样本数(n=8)。

（示例）系统抽样和整群抽样的关系

案例说明：下面展示的是“三采一”的系统采样：我们从前三个主要抽样单元（PSU）中随机选择一个，然后再连续地每隔三个选择一个。



从主要抽样单位PSU {1,2,3}中随机选择一个值。例如，如果选择2，那么我们将选择上图中所有的红叉个体×的{2，5，8，8，11 14}。

- 抽样得到的样本数据{2、5、8、11、14}，只是我们随机选定了1个主要抽样单位（PSU）红叉×，因而所有具有该主要抽样单位的全部个体都被抽中（全部红叉）。
- 实际上，只抽选1个主要抽样单位（PSU）的情况并不少见，例如以上“三采一”的系统样本。我们只采样3个主要抽样单位（分别是绿点●、红叉×、蓝短

抽样误差：系统抽样和整群抽样（记号表达）

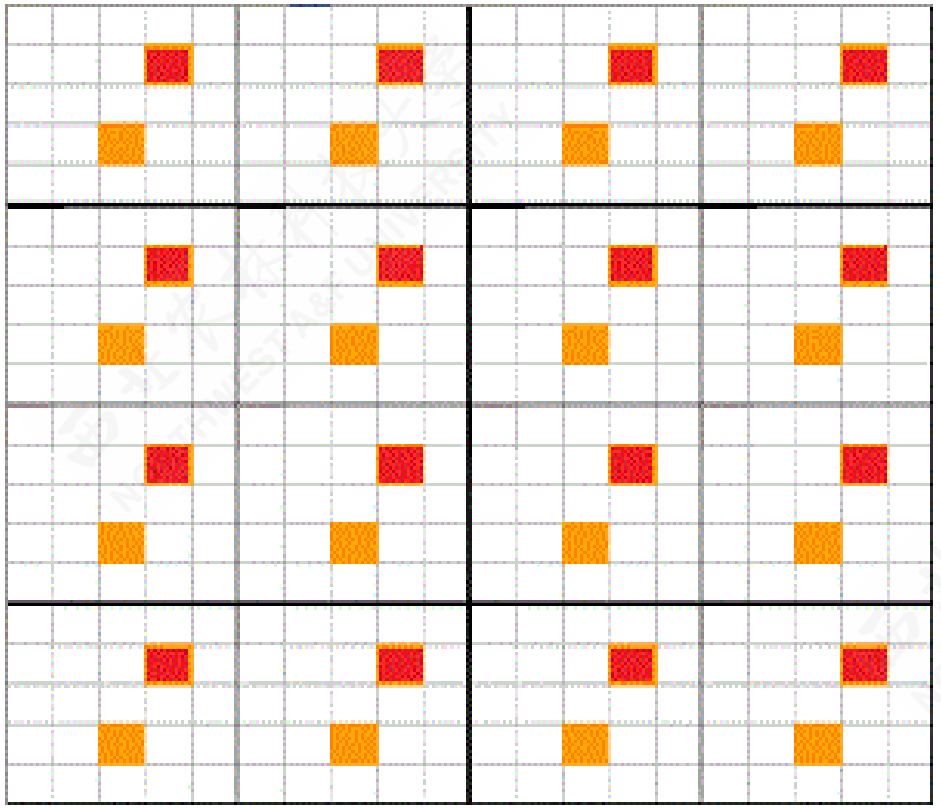
我们约定系统抽样和整群抽样的记号表达体系如下：

- N表示总体中的主要抽样单元（PSU）的数量；n表示样本中的主要抽样单元（PSU）的数量；M_i表示第 i个主要抽样单元（PSU）中次要抽样单元（SSU）的数量；$M=\sum_{i=1}^{N} M_i$表示总体中的所有次要抽样单元（SSU）的数量；
- y_{ij}表示第 i个主要抽样单元（PSU）中第 j次要抽样单元（SSU）的个体的变量值。$y_i=\sum_{j=1}^{M_i} y_{ij}$表示第 i个主要抽样单元（PSU）下所有个体的变量值之和。
- 主要抽样单元（PSU）的均值记为 μ_1，次要抽样单元（SSU）的均值记为 μ，二者的计算公式分别为：

$$\mu_1=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij}=\frac{1}{N}\sum_{i=1}^{N} y_i$$

$$\mu=\frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij}=\frac{1}{M}\sum_{i=1}^{N} y_i$$

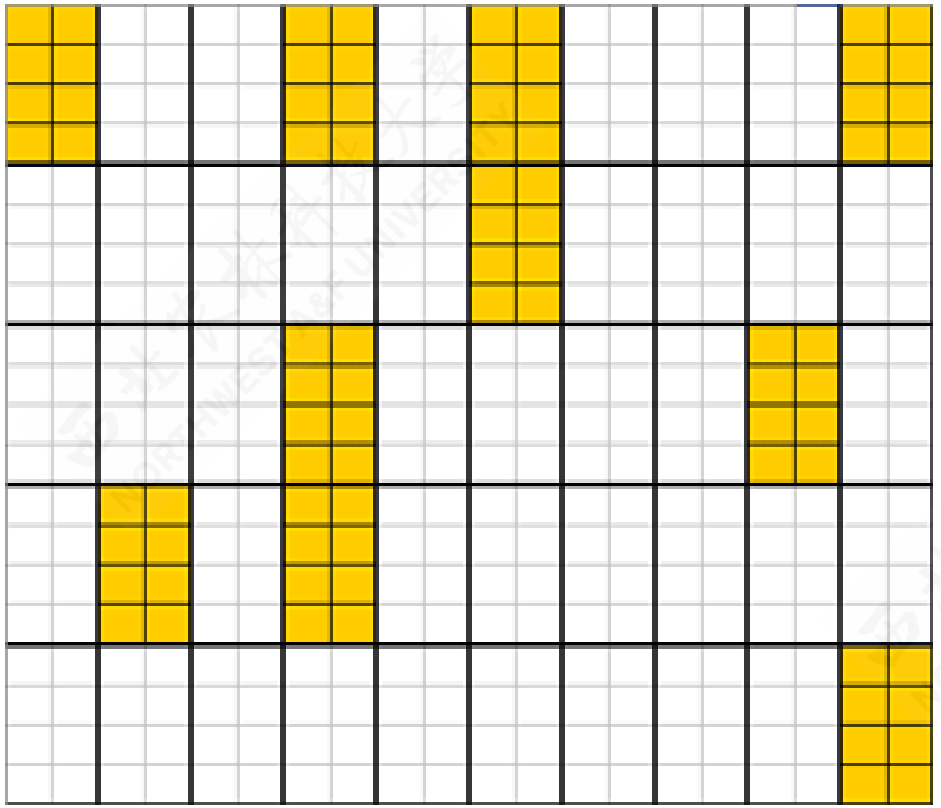
（示例）系统抽样的记号表达



系统抽样示例

- 总体的主要抽样单元（PSU）的数量 $N = 25$：每个5*5中型方框的全部小格，共25个。
- 样本中的主要抽样单元（PSU）的数量 $n = 2$：每个5*5中型方框的都抽中了2个着色小格。
- 每个主要抽样单元（PSU）中次要抽样单元（SSU）的数量 $M_i = 16$：所有5*5中型方框，共有16个。

（示例）整群抽样的记号表达



整群抽样示例

- 总体的主要抽样单元（PSU）的数量 $N = 50$：每个2*4中型方框， 共50个。
- 样本中的主要抽样单元（PSU）的数量 $n = 10$：随机抽中的2*4中型方框，共抽中10个2*4中型着色方框。
- 每个主要抽样单元（PSU）中次要抽样单元（SSU）的数量 $M_i = 8$：每个2*4中型方框中的8个小格。

抽样误差：系统抽样的抽样误差计算

在“1-in-n”系统抽样方案下，估计**次要抽样单元（SSU）**的均值$\hat{\mu}_{sy}$和方差$\widehat{Var}\left(\hat{\mu}_{sy}\right)$分别为：

$$\hat{\mu}_{sy} = \bar{y}_{sy} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i$$

$$\begin{aligned}\widehat{Var}(\hat{\mu}_{sy}) &= \frac{M - n\cdot\bar{M}}{M\cdot n}\cdot\frac{1}{(n-1)}\cdot\sum_{i=1}^{n}(\bar{y}_i - \hat{\mu})^2 \\ &= \frac{M - n\cdot\bar{M}}{M\cdot n}\cdot s^2_{\bar{y}_i}\end{aligned}$$

其中：

- $\bar{y}_i = \frac{y_i}{M_i} = \frac{\sum_{j=1}^{M_i} y_{ij}}{M_i}; i \in 1, 2, \ldots, n$
- $\bar{M} = M_1 = M_2 = \ldots = M_n$

（示例）渡船汽车载客量案例

案例说明：载有汽车横渡海湾的渡轮是按载客量而不是按人收取费用的。轮渡公司希望估算8月份每辆车的平均载人数。该公司知道去年有400辆车乘坐轮渡（见右表）。

```
  1   2   3   4   5   6   7   8  10
 12  22  59 102 108  66  24   6   1
```

公司想对其中80辆车进行采样，为了便于估计系统样本的方差，研究人员决定选择使用**系统抽样**方法，反复抽取10份样本，每份样本都包含8量汽车的记录数据。

id	persons
1	5
2	6
3	4
4	8
5	5
6	6
7	5
8	2

Showing 1 to 8 of 400 entries

Previous 1 2 3 4 5 … 50 Next

（示例）系统抽样下估计期望和方差：抽样结果

公司决定采用$1-in-50(400/8)$的系统抽样方案，也即：

- 从1到50的序号中，不重复随机选择10个序号：8、16、40、6、2、26、37、14、47、46；
- 然后分别以这10个序号作为起始点，每隔50个抽取1个单位，每份样本都会抽取得到8个单位；
- 最终共获得10份系统抽样样本（每份样本含8个个体）。抽样结果如下（括号内为车内人数）：

select	out	mean
sample_1	8(2),58(3),108(2),158(3),208(3),258(6),308(4),358(1)	3.00
sample_2	16(4),66(5),116(5),166(6),216(2),266(4),316(4),366(6)	4.50
sample_3	40(5),90(5),140(7),190(7),240(5),290(5),340(4),390(6)	5.50
sample_4	6(6),56(3),106(7),156(4),206(6),256(6),306(3),356(3)	4.75
sample_5	2(6),52(6),102(6),152(4),202(4),252(5),302(4),352(3)	4.75

（示例）系统抽样下估计期望和方差：计算结果

根据案例，容易计算得到：主要抽样单位（PSU）数量 $N=50$；样本中的主要抽样单位（PSU）数量 $n=10$；第 i个主要抽样单位下的次要抽样单位的数量 $M_i=8(i\in 1,2,\cdots,50)$；总体的全部次要抽样单位（SSU）数量 $M=\sum_{i=1}^{50}M_i=8\times 50=400$

$$\hat{\mu}_{sy}=\frac{1}{n}\sum_{i=1}^{n}\bar{y}_i=4.62$$

$$\widehat{Var}(\hat{\mu}_{sy})=\frac{M-n\cdot\bar{M}}{M\cdot n}\cdot\frac{1}{(n-1)}\cdot\sum_{i=1}^{n}(\bar{y}_i-\hat{\mu})^2=\frac{M-n\cdot\bar{M}}{M\cdot n}\cdot s^2_{\bar{y}_i}$$

$$=\frac{400-10\times 8}{400\times 10}\cdot 0.4931=0.0394$$

- 其中：$\bar{y}_i=\frac{y_i}{M_i}=\frac{\sum_{j=1}^{M_i}y_{ij}}{M_i};i\in 1,2,\ldots,n;$ 以及 $\bar{M}=M_1=M_2=\ldots=M_n$

抽样误差：整群抽样的抽样误差计算方法I

为了次要抽样单元（SSU）均值和方差，我们可以采用**无偏估计法**（unbiased estimator）：

$$\hat{\mu} = \frac{N}{M} \cdot \frac{\sum_{i=1}^{n} y_i}{n}$$

$$\widehat{Var}(\hat{\mu}) = \frac{N(N-n)}{M^2} \cdot \frac{s_u^2}{n}$$

- $s_u^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$
- N表示总体中的主要抽样单元（PSU）的数量； n表示样本中的主要抽样单元（PSU）的数量； M_i表示第 i个主要抽样单元（PSU）中次要抽样单元（SSU）的数量； M表示总体中的所有次要抽样单元（SSU）的数量；
- y_{ij}表示第 i个主要抽样单元（PSU）中第 j次要抽样单元（SSU）的个体的变量值。 $y_i = \sum_{j=1}^{M_i} y_{ij}$表示第 i个主要抽样单元（PSU）下所有个体的变量值之和。

抽样误差：整群抽样的抽样误差计算方法2

此外，当群组变量值总和与群组单位数呈正相关关系时，使用**比率估计法**（ratio estimator）比使用无偏估计更好。此时，估计的**次要抽样单元（SSU）**均值 $\hat{\mu}_r$ 和方差 $\widehat{Var}(\hat{\mu}_r)$ 分别为：

$$\hat{\mu}_r = r = \frac{\sum_{i=1}^{n} y_i}{M} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i}$$

$$\widehat{Var}(\hat{\mu}_r) = \frac{N(N-n)}{n(n-1)} \cdot \frac{1}{M^2} \sum_{i=1}^{n} (y_i - rM_i)^2$$

- N表示总体中的主要抽样单元（PSU）的数量； n表示样本中的主要抽样单元（PSU）的数量； M_i表示第 i个主要抽样单元（PSU）中次要抽样单元（SSU）的数量； M表示总体中的所有次要抽样单元（SSU）的数量；
- y_{ij}表示第 i个主要抽样单元（PSU）中第 j次要抽样单元（SSU）的个体的变量值。 $y_i = \sum_{j=1}^{M_i} y_{ij}$表示第 i个主要抽样单元（PSU）下所有个体的变量值之和

（示例）家庭休假支出案例

案例说明：社会学家想要估计某个城市中每个家庭的平均年休假预算。据统计，这个城市有3100户。

社会学家将整个城市划分为400个街区，并将其视为400个集群。然后，他随机抽样了24个集群，采访了该集群中的每个家庭。

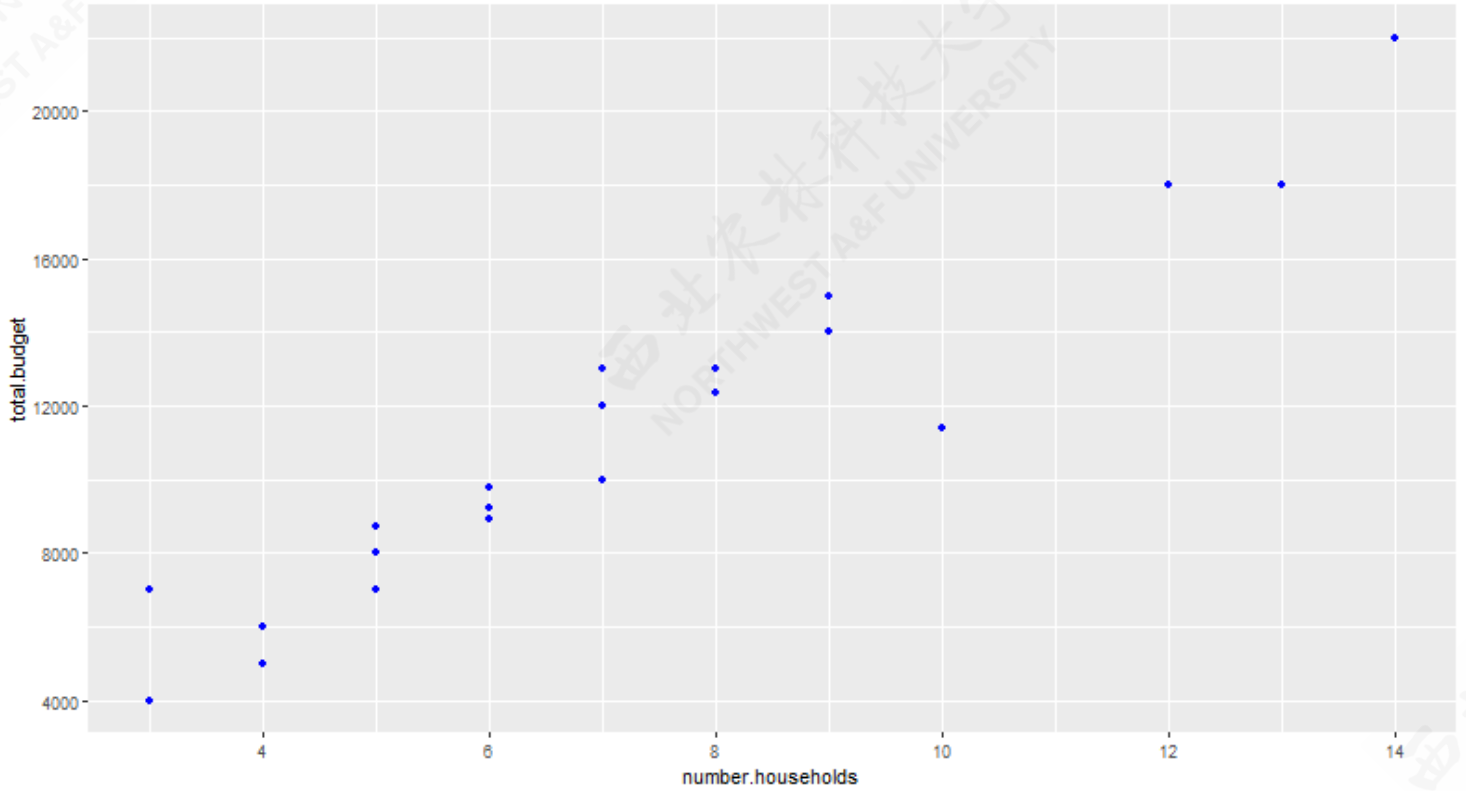
整群抽样的结果见右边数据表：

cluster	number.households	total.budget
389	7	12000
202	9	15000
39	5	8000
286	8	13000
6	12	18000
180	5	7000
143	4	6000
280	8	13000

Showing 1 to 8 of 25 entries Previous 1 2 3 4 Next

（示例）整群抽样下的抽样误差：相关性分析

初步分析抽样的各群组我们可以发现，主要抽样单元（PSU）的**变量总值** y_i（各群组内全部家庭的旅游支出总和）与主要抽样单元（PSU）的**单位规模** M_i（各群组的家庭数）存在高度正相关关系。



（示例）整群抽样下的抽样误差：回归分析

利用R软件进行回归分析，可以进一步发现二者呈现显著线性关系。

$$
\begin{array}{lccc}
total.budget = & +647.98 & +1441.94 number.households_i & +e_i \\
(s) & (705.8674) & (92.5852) & \\
(t) & (+0.92) & (+15.57) & \\
(p) & (0.3686) & (0.0000) &
\end{array}
$$

（示例）整群抽样下的抽样误差：比率估计法

根据整群抽样数据，我们可以计算得到次要抽样单元（SSU）的均值为：

$$\hat{\mu}_r = r = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i} = \frac{259240}{169} = 1533.96$$

$$\widehat{Var}\left(\hat{\mu}_r\right) = \frac{N(N-n)}{n \cdot M^2} \cdot \frac{1}{n-1} \sum_{i=1}^{n} \left(y_i - rM_i\right)^2$$

因为已知： $M = 3100; N = 400; n = 24$，次要抽样单元（SSU）的方差计算结果为：

$$\begin{aligned}
\widehat{Var}\left(\hat{\mu}_r\right) &= \frac{N(N-n)}{n \cdot M^2} \cdot \frac{1}{n-1} \sum_{i=1}^{n} \left(y_i - rM_i\right)^2 \\
&= \frac{400 \times (400-24)}{24 \times 9610000} \times \frac{1}{24-1} \times 40387519.04 \\
&= \frac{150400}{230640000} \times \frac{1}{23} \times 40387519.04 = 1145.07
\end{aligned}$$

（示例）整群抽样下的抽样误差：比率估计法

前述**比率估计法**需要用到的计算表如下所示：

cluster	number.households	total.budget	r_Mi	part_sqr
389	7	12000	10,737.75	1,593,271.33
202	9	15000	13,805.68	1,426,399.13
39	5	8000	7,669.82	109,017.19
286	8	13000	12,271.72	530,397.62
6	12	18000	18,407.57	166,116.54
180	5	7000	7,669.82	448,662.16
143	4	6000	6,135.86	18,457.39
280	8	13000	12,271.72	530,397.62
126	14	22000	21,475.50	275,097.15

Showing 1 to 9 of 25 entries

Previous 1 2 3 Next

（示例）整群抽样下的抽样误差：无偏估计法

作为对比，下面我们再采用**无偏估计法**公示进行计算。

容易计算：$M=3100; N=400; n=24$，以及 $s_u^2=\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2=20208762.32$。因此，次要抽样单元（SSU）的均值和方差计算结果分别为：

$$\hat{\mu}=\frac{N}{M}\cdot\frac{\sum_{i=1}^{n}y_i}{n}=\frac{400}{3100}\cdot\frac{259240}{24}=1393.81$$

$$\begin{aligned}\widehat{Var}(\hat{\mu})&=\frac{N(N-n)}{M^2\cdot n}\cdot\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2\\&=\frac{400(400-24)}{(3100)^2\cdot 24}\cdot s_u^2=\frac{400(400-24)}{(3100)^2\cdot 24}\times 20208762.32=13178.1\end{aligned}$$

抽样误差：整群抽样两个误差估计方法的比较

- 当群组变量值总和与群组单位数大小成正比时，使用比率估计比使用无偏估计更好。因为无偏估计法的方差会非常大，估计结果非常不满意。

> 我们可以简单使用的随机抽样的公式来计算方差吗？——抱歉不行！

- 如果采用简单随机抽样，那么就应该相对应使用简单随机抽样公式计算方差，而且必须通过简单随机抽样收集数据。注意：如果不按照抽样方案计算方差，这会是一个很大的错误！

抽样误差：整群抽样的抽样误差计算方法3

有时候，群组被抽中的概率p_i就等于群组单位数占总体单位数的比率，也即$p_i = M_i/M$。我们一般称的这种情形为**主要抽样单元**（PSU）满足**比例概率条件**（probabilities proportional to size, pps）。那么，在满足pps条件下进行的整群抽样，估计**次要抽样单元（SSU）**的均值$\hat{\mu}_p$和方差$\widehat{Var}\left(\hat{\mu}_p\right)$分别为：

$$\hat{\mu}_p = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{M_i}\right)$$

$$\widehat{Var}\left(\hat{\mu}_p\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\bar{y}_i - \hat{\mu}_p\right)^2$$

- $\bar{y}_i = \frac{y_i}{M_i}$表示第$i$个群组的抽样均值。$n$表示样本中的主要抽样单元（PSU）的数量；$M_i$表示第$i$个主要抽样单元（PSU）中次要抽样单元（SSU）的数量；M表示总体中的所有次要抽样单元（SSU）的数量。

（示例）请求计算机援助案例

案例说明：一家大型公司共有10个部门，每个部门的员工人数各不相同（见下左表）。IT部门主管计划对该公司的3个部门进行随机抽样，以估计该公司平均每个部门的计算机帮助请求数。然后，他采用**可重复抽样**的比例概率整群抽样法（pps），随机抽取了三个部门的样本数据（见下右表）：

cluster	employees	requires
1	1000	643
2	650	427
3	2100	1266
4	860	544
5	2840	1938
6	1910	1308

Showing 1 to 6 of 11 entries Previous 1 2 Next

cluster	employees	requires
2	650	427
8	3200	1933
10	1200	770

（示例）整群抽样下的抽样误差：计算结果

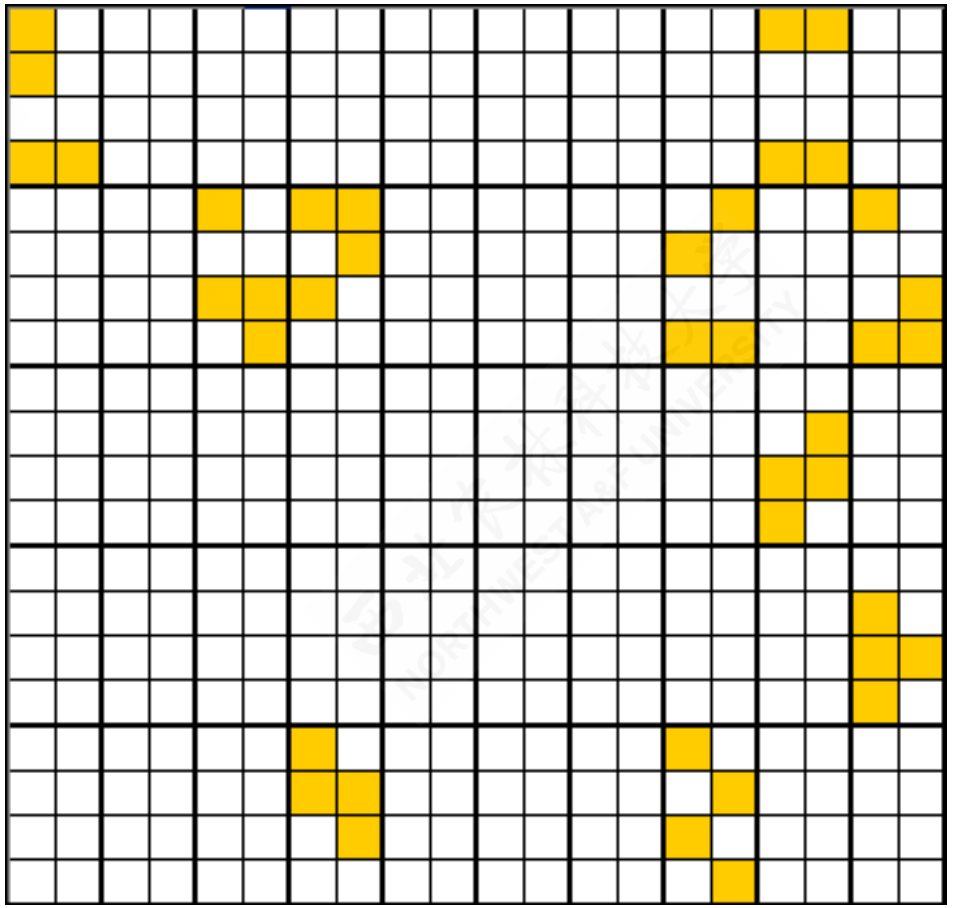
cluster	employees	requires	ratio	minus	sqr
2	650	427	0.6569	0.0227	0.000516
8	3200	1933	0.6041	-0.0302	0.000909
10	1200	770	0.6417	0.0074	0.000055
Total	5050	3130	1.9027	0.0000	0.001480

依据抽样数据，我们知道 $n=3$，容易计算得到 $\bar{y}_i=\frac{y_i}{M_i}$，并进一步计算得到估计的均值 $\hat{\mu}_p$，然后再计算出方差 $\widehat{Var}(\hat{\mu}_p)$：

$$\hat{\mu}_p=\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{M_i}=mean(\bar{y}_i)=\bar{\bar{y}}_i=0.6342$$

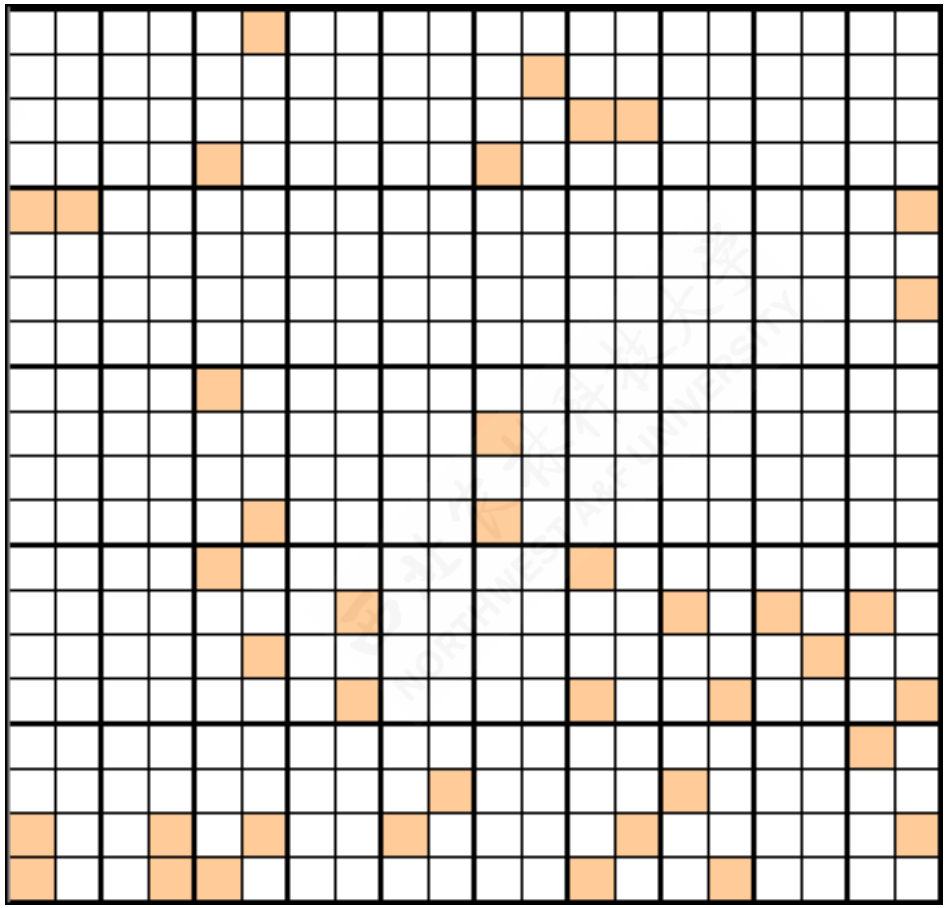
$$\widehat{Var}\left(\hat{\mu}_p\right)=\frac{1}{n(n-1)}\cdot\sum_{i=1}^{n}\left(\bar{y}_i-\hat{\mu}_p\right)^2=\frac{1}{n}\cdot s^2(\bar{y}_i)=0.000247$$

抽样误差：多阶段抽样



两阶段抽样示例1：10个PSU，4个SSU/PSU

抽样误差：多阶段抽样



两阶段抽样示例2：20个PSU，2个SSU/PSU

抽样误差：多阶段抽样的符号约定

多阶段抽样下，一些重要的符号约定如下：

- N表示总体中的全部群组数量； n表示随机抽样后抽选得到的群组数量； M_i表示总体中，第 i个群组中的单位数量； m_i表示随机抽中的第 i个群组中的单位数量；$M = \sum_{i=1}^{N} M_i$表示总体中的所有单位数量；
- y_{ij}表示随机抽中的第 i个群组中的第 j个单位的变量值； $\bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{jij}$表示被抽中的第 i个群组的样本均值。 $\hat{y}_i = M_i \frac{\sum_{j=1}^{m_i} y_{ij}}{m_i} = M_i \bar{y}_i$表示对总体中第 i个群组的变量总值的估计值。

抽样误差：多阶段抽样下的抽样误差计算方法I

多阶段抽样下，**次要抽样单元（SSU）**均值 $\hat{\mu}$和方差 $\widehat{Var}(\hat{\mu})$的**无偏估计法**（unbiased estimator）计算公式分别为：

$$\hat{\mu} = \frac{N}{M} \cdot \frac{\sum_{i=1}^{n} \hat{y}_i}{n} = \frac{N}{M} \cdot \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{n}$$

$$\widehat{Var}(\hat{\mu}) = \frac{N(N-n)}{M^2} \cdot \frac{s_u^2}{n} + \frac{N}{nM^2} \sum_{i=1}^{n} M_i \left(M_i - m_i\right) \frac{s_i^2}{m_i}$$

两个样本方差，其中 s_u^2表示主要抽样单位（PSU）的样本方差；而 s_i^2表示被抽中的第 i个群组的样本方差。

$$s_u^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\hat{y}_i - \frac{\sum_{i=1}^{n} \hat{y}_i}{n} \right)^2$$

$$s_i^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} \left(y_{ij} - \bar{y}_i\right)^2$$

抽样误差：多阶段抽样下的抽样误差计算方法2

> 对于两阶段抽样方案：第一阶段和第二阶段都采用简单随机抽样。
>
> - 如果总体的次要抽样单元（SSU）总数 M不可知，则不能使用前述的**无偏估计法**。
> - 此外，如果群组的变量加总值（sum value）与群组的个体数量（element size）与呈比率关系，则应该采用下述**比率估计法**。

对于这样的多阶段抽样方案，**次要抽样单元（SSU）**均值 $\hat{\mu}_r$和方差 $\widehat{Var}\left(\hat{\mu}_r\right)$的**比率估计法**（ratio estimator）计算公式分别为：

$$\hat{\mu}_r = \hat{r} = \frac{\sum_{i=1}^{n} \hat{y}_i}{\sum_{i=1}^{n} M_i} = \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{\sum_{i=1}^{n} M_i}$$

$$\widehat{Var}\left(\hat{\mu}_r\right) = \frac{N(N-n)}{nM^2} \cdot \frac{1}{n-1} \sum_{i=1}^{n} \left(\hat{y}_i - M_i \hat{r}\right)^2 + \frac{N}{nM^2} \sum_{i=1}^{n} M_i \left(M_i - m_i\right) \frac{s_i^2}{m_i}$$

（示例）连锁餐厅满意度案例

案例说明：一家餐饮连锁店想估计员工对工作的平均满意度（里克特量表1-7分制）。该连锁店共有120家餐厅，连锁店的全体员工总数为6860人。研究人员决定使用**两阶段随机抽样**方案，第一阶段采用简单随机抽样来采样10家餐厅（被抽中的序号见列`id`，餐厅总员工数见列`tot_worker`）。然后，第二阶段也使用简单随机抽样对这些餐厅中约20%的员工（被抽中的员工数量见列`sel_worker`）进行抽样和工作满意度采访（见列`satisfaction`）。最终抽样数据结果如下：

id	tot_worker	sel_worker	satisfaction
41	54	11	5,7,4,7,6,7,6,5,3,4,7
119	48	10	6,3,7,3,6,3,5,6,7,7
42	68	14	5,5,7,6,4,3,5,5,6,3,6,4,7,4
18	70	14	3,3,3,4,7,5,7,4,6,7,3,3,3,3
13	52	11	6,5,5,3,6,3,3,5,7,7,3

Showing 1 to 5 of 10 entries

Previous 1 2 Next

（示例）多阶段抽样的抽样误差：基本计算量

根据案例数据，容易得到：

- 所有餐厅数量为 $N = 120$。
- 简单随机抽样选中的餐厅数量为 $n = 10$。
- 第 i个餐厅的总员工数为 $M_i =$ `total_worker`列。
- 随机抽中的第 i个餐厅中的被抽中的员工数量为 $m_i =$ `sel_worker`列。
- 连锁店全体员工的总人数为 $M = \sum_{i=1}^{N} M_i = 6860$。
- 随机抽中的第 i个餐厅中的平均工作满意度评分为 $\bar{y}_i =$ `mean`列，工作满意度的样本方差 $s_i^2 =$ `variance`列。
- 估计得到的第 i个酒店的加总满意度评分为 $\hat{y}_i =$ `y_hat`列，10家被抽中酒店估计的平均加总满意度评分为 $\bar{\hat{y}} = \frac{\sum_{i=1}^{n} \hat{y}_i}{n} = 280.29$。

（示例）多阶段抽样的抽样误差：无偏估计法的计算量I

我们可以根据**无偏估计法**的相关理论公式，得到如下的计算表：

id	tot_worker	sel_worker	satisfaction	mean	variance	y_hat	s
41	54	11	5,7,4,7,6,7,6,5,3,4,7	5.55	2.07	299.5	
119	48	10	6,3,7,3,6,3,5,6,7,7	5.30	2.90	254.4	
42	68	14	5,5,7,6,4,3,5,5,6,3,6,4,7,4	5.00	1.69	340.0	
18	70	14	3,3,3,4,7,5,7,4,6,7,3,3,3,3	4.36	2.86	305.0	
13	52	11	6,5,5,3,6,3,3,5,7,7,3	4.82	2.56	250.5	
80	62	13	5,5,4,3,6,5,7,7,3,5,3,3,6	4.77	2.19	295.7	
68	41	9	6,4,3,6,6,3,3,5,7	4.78	2.44	195.9	
25	53	11	3,4,5,7,4,7,7,7,4,7,5	5.45	2.47	289.1	

Showing 1 to 8 of 10 entries

Previous 1 2 Next

（示例）多阶段抽样的抽样误差：无偏估计法的计算量2

从而容易计算得到到如下两个**无偏估计法**需要用到的样本方差：

$$s_u^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{y}_i - \frac{\sum_{i=1}^{n}\hat{y}_i}{n}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{y}_i - \bar{\hat{y}}\right)^2 = 1591.18$$

$$s_i^2 = \frac{1}{m_i-1}\sum_{j=1}^{m_i}\left(y_{ij} - \bar{y}_i\right)^2 = \text{variance}$$

> 上述 s_i^2计算结果见前页ppt的计算表中的variance列。

（示例）多阶段抽样的抽样误差：无偏估计法的结果

因此，采用**无偏估计法**估计得到的酒店满意度的均值和方差分别为：

$$\hat{\mu} = \frac{N}{M} \cdot \frac{\sum_{i=1}^{n} \hat{y}_i}{n} = \frac{N}{M} \cdot \bar{\hat{y}} = \frac{120}{6860} \times 280.29 = 4.90$$

$$\begin{aligned}
\widehat{Var}(\hat{\mu}) &= \frac{N(N-n)}{M^2} \cdot \frac{s_u^2}{n} + \frac{N}{nM^2} \sum_{i=1}^{n} M_i \left(M_i - m_i\right) \frac{s_i^2}{m_i} \\
&= \frac{120(120-10)}{10 \times 6860^2} \times 1591.18 + \frac{120}{10 \times 6860^2} \times 4615.55 \\
&= 0.0458
\end{aligned}$$

上述求和项内部个值计算结果见前页ppt的计算表，其中：

- $M_i \left(M_i - m_i\right) \frac{s_i^2}{m_i}$ 见计算表中的`sum_right`列；

（示例）多阶段抽样的抽样误差：比率估计法的计算量I

我们可以根据**比率估计法**的相关理论公式，得到如下的计算表：

id	tot_worker	sel_worker	satisfaction	mean	variance	y_hat	s
41	54	11	5,7,4,7,6,7,6,5,3,4,7	5.55	2.07	299.5	
119	48	10	6,3,7,3,6,3,5,6,7,7	5.30	2.90	254.4	
42	68	14	5,5,7,6,4,3,5,5,6,3,6,4,7,4	5.00	1.69	340.0	
18	70	14	3,3,3,4,7,5,7,4,6,7,3,3,3,3	4.36	2.86	305.0	
13	52	11	6,5,5,3,6,3,3,5,7,7,3	4.82	2.56	250.5	
80	62	13	5,5,4,3,6,5,7,7,3,5,3,3,6	4.77	2.19	295.7	
68	41	9	6,4,3,6,6,3,3,5,7	4.78	2.44	195.9	
25	53	11	3,4,5,7,4,7,7,7,4,7,5	5.45	2.47	289.1	

Showing 1 to 8 of 10 entries

Previous 1 2 Next

（示例）多阶段抽样的抽样误差：比率估计法的计算量2

容易计算得到到如下**比率估计法**需要用到的样本方差：

$$s_i^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} \left(y_{ij} - \bar{y}_i \right)^2 = \text{variance}$$

> 上述 s_i^2计算结果见前页ppt的计算表中的`variance`列。

（示例）多阶段抽样的抽样误差：比率估计法的结果

因此，采用**比率估计法**估计得到的酒店满意度的均值和方差分别为：

$$\hat{\mu}_r = \hat{r} = \frac{\sum_{i=1}^{n} \hat{y}_i}{\sum_{i=1}^{n} M_i} = \frac{2802.86}{555} = 5.05$$

$$\begin{aligned}
\widehat{Var}\left(\hat{\mu}_r\right) &= \frac{N(N-n)}{nM^2} \cdot \frac{1}{n-1} \sum_{i=1}^{n} \left(\hat{y}_i - M_i \hat{r}\right)^2 + \frac{N}{nM^2} \sum_{i=1}^{n} M_i \left(M_i - m_i\right) \frac{s_i^2}{m_i} \\
&= \frac{120(120-10)}{10 \times 6860^2} \times \frac{1}{10-1} \times 7120.48 + \frac{120}{10 \times 6860^2} \times 4615.55 \\
&= 0.0234
\end{aligned}$$

> 上述两个求和项内部个值计算结果见前页ppt的计算表，其中：
>
> - $\left(\hat{y}_i - M_i \hat{r}\right)^2$见计算表中的`sum1_yi_sqr`列；
> - $M_i \left(M_i - m_i\right) \frac{s_i^2}{m_i}$见计算表中的`sum2_si_sqr`列。

抽样误差：多阶段抽样下的抽样误差计算方法3

> 对于两阶段抽样方案：第一阶段采用比例概率抽样法（PPS），第二阶段采用简单随机抽样法：
>
> - 那么抽样误差计算应该使用**比例概率估计法**（pps估计法，具体为Hansen-Hurwitz estimator）。

对于这样的多阶段抽样方案，**次要抽样单元（SSU）**均值$\hat{\mu}_p$和方差$\widehat{Var}\left(\hat{\mu}_p\right)$的**比例概率估计法**（pps estimator）计算公式分别为：

$$\hat{\mu}_p = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{\hat{y}_i}{M_i} = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{\bar{y}_i * M_i}{M_i} = \frac{1}{n} \cdot \sum_{i=1}^{n} \bar{y}_i$$

$$\widehat{Var}\left(\hat{\mu}_p\right) = \frac{1}{n(n-1)} \cdot \sum_{i=1}^{n} \left(\bar{y}_i - \hat{\mu}_p\right)^2$$

（示例）学生书本支出案例

案例说明：一个学院共有36个专业（major）。研究者想估算出上学期学生在教科书上花费（expenses）的平均金额。由于每个专业的规模差异很大，因此采用的两阶段抽样方案，其中第一阶段采用的是pps抽样，第二阶段是简单随机抽样。最终抽样数据结果如下：

major	tot_students	sel_students	expenses
18	10	4	326,400,423,443
13	20	8	278,312,450,350,227,438,512,403
16	30	12	512,256,332,402,512,309,411,610,422,630,550,470
4	15	6	426,312,512,440,342,533

（示例）多阶段抽样的抽样误差：PPS估计法的计算表

我们可以根据**比例概率估计法**的相关理论公式，得到如下的计算表：

major	tot_students	sel_students	expenses
18	10	4	326,400,423,443
13	20	8	278,312,450,350,227,438,512,403
16	30	12	512,256,332,402,512,309,411,610,422,630,550,470
4	15	6	426,312,512,440,342,533

建议使用html浏览本课件，此表可以往右拉动，可以查看更多计算列。

（示例）多阶段抽样的抽样误差：PPS估计法的结果

因此，采用**比例概率估计法**估计得到学生书本支出的均值和方差分别为：

$$\hat{\mu}_p = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{\hat{y}_i}{M_i} = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{\bar{y}_i * M_i}{M_i} = \frac{1}{n} \cdot \sum_{i=1}^{n} \bar{y}_i = \bar{\bar{y}}_i = 412.02$$

$$\widehat{Var}\left(\hat{\mu}_p\right) = \frac{1}{n(n-1)} \cdot \sum_{i=1}^{n} \left(\bar{y}_i - \hat{\mu}_p\right)^2 = \frac{1}{n} \cdot s^2_{\bar{y}_i}$$

$$= \frac{1}{4} \times 1214.6406 = 303.6602$$

> 上述计算结果的中间步骤计算值见前页ppt的计算表。 其中：
>
> - $\bar{y}_i$见计算表中的`mean`列；
> - $\hat{y}_i$见计算表中的`y_hat`列。

本节结束