抽样调查 第七周作业

20307100013 蒋翌坤

第四章 补充题：

题 1

（1）引入示性函数$Z_i=\begin{cases}1, if\ i\in S\\0, if\ i\notin S\end{cases}$, 则$E(Z_i)=\frac{n}{N}$, $E\left(Z_i^2\right)=\frac{n}{N}$, $E\left(Z_iZ_j\right)=\frac{n(n-1)}{N(N-1)}$

$s_{yx}=\frac{1}{n-1}\sum_{i\in S}(y_ix_i-x_i\bar{y}-y_i\bar{x}+\bar{y}\bar{x})=\frac{1}{n-1}(\sum_{i\in S}(y_ix_i)-n\bar{x}\bar{y}-n\bar{x}\bar{y}+n\bar{y}\bar{x})$

$=\frac{1}{n-1}\left(\sum_{i\in S}(y_ix_i)-\frac{1}{n}(\sum_{i\in S}(y_i)\sum_{i\in S}(x_i))\right)=\frac{1}{n-1}\left(\frac{n-1}{n}\sum_{i\in S}(y_ix_i)-\frac{1}{n}\sum_{i\neq j,i,j\in S}\left(y_ix_j\right)\right)$

$=\frac{1}{n-1}\left(\frac{n-1}{n}\sum_{i=1}^{N}\left(y_ix_iZ_i^2\right)-\frac{1}{n}\sum_{i\neq j}^{N}\left(y_ix_jZ_iZ_j\right)\right)$

$E\left(s_{yx}\right)=\frac{1}{n-1}\left(\frac{n-1}{N}\sum_{i=1}^{N}(y_ix_i)-\frac{(n-1)}{N(N-1)}\sum_{i\neq j}^{N}\left(y_ix_j\right)\right)$

$=\frac{1}{n-1}\left(\frac{n-1}{N}\sum_{i=1}^{N}(y_ix_i)-\frac{(n-1)}{N(N-1)}\left(\sum_{i=1}^{N}(y_i)\sum_{i=1}^{N}(x_i)-\sum_{i=1}^{N}(y_ix_i)\right)\right)$

$=\frac{1}{n-1}\left(\frac{(n-1)(N-1+1)}{N(N-1)}\sum_{i=1}^{N}(y_ix_i)-\frac{(n-1)}{N(N-1)}(N\overline{y_u}N\overline{x_u})\right)$

$=\frac{1}{N-1}\left(\sum_{i=1}^{N}(y_ix_i)-N(\overline{y_u}\times\overline{x_u})\right)=\frac{1}{N-1}\left(\sum_{i=1}^{N}(y_i-\overline{y_u})\,(x_i-\overline{x_u})\right)=S_{yx}$

（2）可知$V(Z_i)=\frac{n}{N}\left(1-\frac{n}{N}\right)$, 当$i\neq j$时，$cov\left(Z_i,Z_j\right)=-\frac{1}{N-1}\left(1-\frac{n}{N}\right)\left(\frac{n}{N}\right)$,

$cov(\bar{y},\bar{x})=\frac{1}{n^2}\sum_{i,j\in S}cov\left(y_i,x_j\right)=\frac{1}{n^2}\sum_{i,j=1}^{N}cov\left(Z_iy_i,Z_jx_j\right)=\frac{1}{n^2}\sum_{i,j=1}^{N}cov\left(Z_iy_i,Z_jx_j\right)$

$=\frac{1}{n^2}\sum_{i,j=1}^{N}y_ix_jcov\left(Z_i,Z_j\right)=\frac{1}{n^2}\left(\sum_{i=1}^{N}y_ix_i\frac{n}{N}\left(1-\frac{n}{N}\right)-\sum_{i\neq j}^{N}y_ix_j\frac{1}{N-1}\left(1-\frac{n}{N}\right)\left(\frac{n}{N}\right)\right)$

$=\frac{1-f}{n}\left(\frac{1}{N}\sum_{i=1}^{N}y_ix_i-\frac{1}{N(N-1)}\sum_{i\neq j}^{N}y_ix_j\right)=\frac{1-f}{n}S_{yx}$

题 2

（1）$E(\overline{y_s})=\frac{\binom{N-2}{n-1}}{\binom{N}{n}}E(\bar{y}+c)+\frac{\binom{N-2}{n-1}}{\binom{N}{n}}E(\bar{y}-c)+\left(1-\frac{2\binom{N-2}{n-1}}{\binom{N}{n}}\right)E(\bar{y})=E(\bar{y})=\overline{y_u}$

所以$\overline{y_s}$是$\overline{y_u}$的无偏估计

（2）引入与题 1 相同的示性函数$Z_i=\begin{cases}1, if\ i\in S\\0, if\ i\notin S\end{cases}$

不失一般性，令$y_1=y_{(1)},y_N=y_{(N)}$

$V(\overline{y_s})=V\left(\frac{1}{n}\left(\sum_{i=1}^{N}Z_iy_i\right)+c(Z_1-Z_N)\right)=\frac{1}{n^2}V\left(\sum_{i=1}^{N}Z_iy_i+nc(Z_1-Z_N)\right)$

$=\frac{1}{n^2}Cov\left(\sum_{i=1}^{N}Z_iy_i+nc(Z_1-Z_N),\sum_{j=1}^{N}Z_jy_j+nc(Z_1-Z_N)\right)$

$= \frac{1}{n^2}\sum_{i,j=1}^{N} y_i y_j Cov(Z_i, Z_j) + y_i nc\big(Cov(Z_i, Z_1) - Cov(Z_i, Z_N)\big) + y_j nc\left(Cov(Z_j, Z_1) - Cov(Z_j, Z_N)\right) + n^2c^2\big(V(Z_1) + V(Z_N) - 2Cov(Z_1, Z_N)\big)$

$= \frac{1}{n^2}\left(\sum_{i=1}^{N} y_i^2 V(Z_i) + \sum_{i\neq j}^{N} y_i y_j Cov(Z_i, Z_j) + 2(y_1 - y_N)nc\left(\frac{n}{N}\left(1-\frac{n}{N}\right) + \frac{1}{N-1}\left(1-\frac{n}{N}\right)\left(\frac{n}{N}\right)\right) + 2n^2c^2\frac{N}{N-1}\left(\frac{n}{N}\left(1-\frac{n}{N}\right)\right)\right)$

$= (1-f)\frac{S^2}{n} + (1-f)\frac{2c}{N-1}(y_1 - y_N) + (1-f)2c^2\frac{N}{N-1}\frac{n}{N}$

$= (1-f)\left(\frac{S^2}{n} - \frac{2c}{N-1}\left(y_{(N)} - y_{(1)} - nc\right)\right)$

（3）当$V(\overline{y_s}) < V(\bar{y})$，即$\frac{2c}{N-1}\left(y_{(N)} - y_{(1)} - nc\right) > 0$，即$0 < c < \frac{y_{(N)} - y_{(1)}}{n}$时，$\overline{y_s}$优于$\bar{y}$

当$\frac{2c}{N-1}\left(y_{(N)} - y_{(1)} - nc\right)$取最大，即$c = \frac{y_{(N)} - y_{(1)}}{2n}$时，$\overline{y_s}$最优

题 3

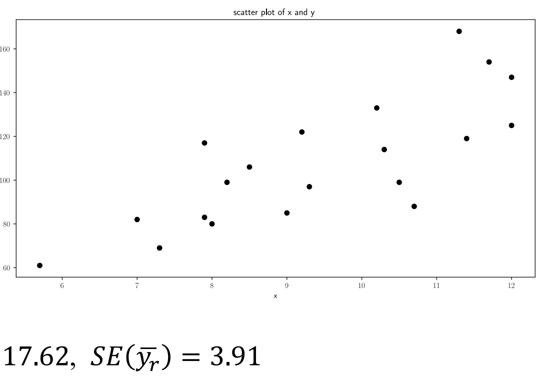
（1）计算可得$\hat{B} = 0.987$, $SE(\hat{B}) = 5.75 \times 10^{-3}$, 95%置信区间为[0.975,0.998]

（2）每次 bootstrap 从 300 个 SRS 样本中有放回的抽取 100 个样本，进行 1000 次，得到 1000 个$\hat{B}$。从这 100 个$\hat{B}$中，可以得到 95%置信区间为[0.967,1.007]

§4.8 Exercises：

题 3

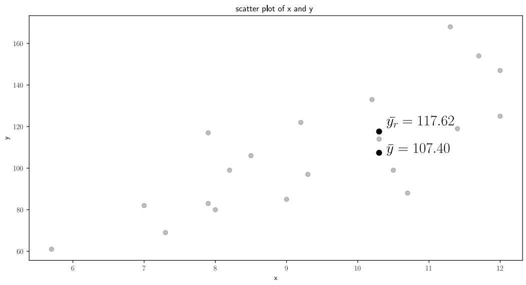
（a）x和y的散点图如下所示：



（b）计算可得：$\overline{y_r} = 117.62$, $SE(\overline{y_r}) = 3.91$

（c）计算可得：$\bar{y} = 107.4$, $SE(\bar{y}) = 6.19$

（d）标出所得到的估计如下图所示：



由于样本中相关系数$r = 0.78 > 0.5$，所以用比估计更合适。

题 4

（a）Domain 2 中有孩子的家庭比例为0.486，95%置信区间为[0.456, 0.517]

（b）Domain 2 中家庭平均孩子数为0.884，95%置信区间为[0.8200.949]

（c）Domain 2 中家庭总孩子数为242400, 95%置信区间为[224691,260109]

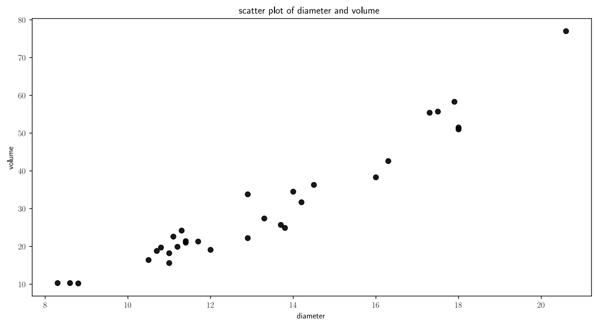
题 9

（a）在拘捕中，burglary 占比为0.059，95%置信区间为[0.052,0.066]

（b）家庭内的严重攻击行为数为76120, 95%置信区间为[55927,96312]

题 11

（a）volume 和 diameter 的散点图如下图所示：

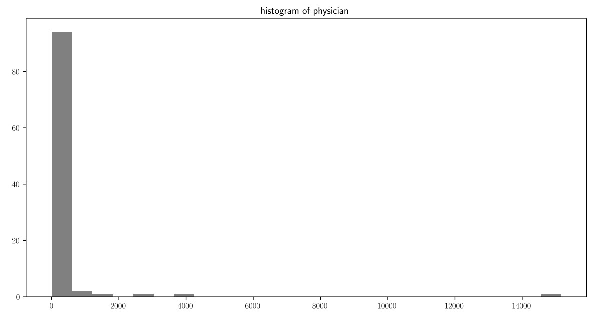


（b）利用比估计，所有树的总体积为95272, 95%置信区间为[84548,105996]

（d）利用简单估计，所有树的体积为89517, 95%置信区间为[72438,106596]。由于样本中相关系数$r = 0.97 > 0.5$，所以用比估计更合适。

题 13

（a）physicians 的直方图如下图所示：



（b）利用简单估计，全美内科医生总数为 933411，标准误为 491983

（c）利用比估计，全美内科医生总数为 639506，标准误为 87885.3

（e）比估计更接近总体中内科医生的真实人数。这是由于内科医生与县人口的相关系数高达 0.98，如果用简单估计，由于县人口差距很大，导致得出的样本均值不能很好的反应各个县的平均内科医生数量，而用比估计可以很好的避免这个问题。

附录：

解答题目所使用的代码及输出请见：
https://thisiskunmeng.github.io/sampling/hw7.html