抽样调查 第五周作业

20307100013 蒋翌坤

§3.8 Exercises：3

（a）总体均值$\overline{y_u} = 71.83$，总体方差$S^2 = 86.17$

（b）样本量为 4 的 SRS 的选择方式有 15 种

（c）以下表格为所有可能的样本量为 4 的 SRS 的选择方式：

		15 种选择方式														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
样本	1	66	66	66	66	66	66	66	66	66	66	59	59	59	59	70
	2	59	59	59	59	59	59	70	70	70	83	70	70	70	83	83
	3	70	70	70	83	83	82	83	83	82	82	83	83	82	82	82
	4	83	82	71	82	71	71	82	71	71	71	82	71	71	71	71
样本均值		**69.5**	**69.25**	**66.5**	**72.5**	**69.75**	**69.5**	**75.25**	**72.5**	**72.25**	**75.5**	**73.5**	**70.75**	**70.5**	**73.75**	**76.5**

$V(\bar{y}) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right) = \frac{86.17}{4} \times \left(1 - \frac{4}{6}\right) = 7.181$

（d）分层抽样的选择方式有 9 种。

（e）以下表格为所有可能的分层抽样的选择方式以及无法利用分层抽样获得的 SRS 选择方式：

		9 种选择方式									无法利用分层抽样获得的 6 种 SRS 选择方式					
		1	2	3	4	5	6	7	8	9	1	2	3	4	5	6
样本	1	66	66	66	66	66	66	59	59	59	66	66	66	66	59	70
	2	59	59	59	70	70	70	70	70	70	59	59	59	83	83	83
	3	83	83	82	83	83	82	83	83	82	70	70	70	82	82	82
	4	82	71	71	82	71	71	82	71	71	83	82	71	71	71	71

（f）以下表格为分层抽样各个选择方式的均值$\overline{y_{str}}$

	9 种选择方式								
	1	2	3	4	5	6	7	8	9
样本均值	72.5	69.75	69.5	75.25	72.5	72.25	73.5	70.75	70.5

$V(\overline{y_{str}}) = 3.14$，与$V(\bar{y}) = 7.181$相比，$V(\overline{y_{str}})$比$V(\bar{y})$小。

§3.8 Exercises：7

（a）教职员工发稿数量预测为 $\widehat{t_{str}} = \sum_{h=1}^{4} N_h \overline{y_h} = 1321$，标准误为 $SE(\widehat{t_{str}}) = \sqrt{\sum_{h=1}^{4} N_h^2 \left(1 - \frac{n_h}{N_h}\right)\left(\frac{s_h^2}{n_h}\right)} = 261.91$

（b）在第二章 SRS 练习中，我们预测教职员工发稿数量为 1430，标准误为 287.3。和分层抽样相比，分层抽样获得了更低的预测值以及更低的标准误。更低的标准误意味着预测准确性的提升。

（c）教职员工中没有发稿的比例预测为$\widehat{p_{str}} = \frac{1}{N}\sum_{h=1}^{4} n_h \overline{p_h} = 0.567$，标准误为$SE(\widehat{p_{str}}) = \sqrt{\sum_{h=1}^{4} \frac{N_h^2}{N^2}\left(1 - \frac{n_h}{N_h}\right)\left(\frac{p_h(1-p_h)}{n_h - 1}\right)} = 0.0658$

（d）分层抽样在这个例子中提升了准确性。从（b）题的结论中就可以明显看出，通过分层抽样来估计总体获得了更小的标准误，准确性有大幅提升；同时，通过分层抽样所获得的变异系数的估计（估计值/标准误）也十分小，说明预测值距离真实值的距离也会较小。

§3.8 Exercises：12

（a）各层的抽样权重为：$w_1 = \frac{280}{100} = 2.8$，$w_2 = \frac{103}{38} = 2.71$，$w_1 = \frac{280}{60} = 2.73$。分层的方式为各层随机抽样，各层的样本容量按层权比例分配。

（b）当我们计算估计值的方差、标准误时，需要考虑进 fpc，因为样本容量和总体接近。

（c）估计百分比时，利用以下公式：$\widehat{p_{str}} = \frac{1}{N}\sum_{h=1}^{3} n_h \overline{p_h}$,；估计 95%置信区间时，利用以下公式：$[\widehat{p_{str}} \mp 1.96 SE(\widehat{p_{str}})]$，其中，$SE(\widehat{p_{str}}) = \sqrt{\sum_{h=1}^{3} \frac{N_h^2}{N^2}\left(1 - \frac{n_h}{N_h}\right)\left(\frac{p_h(1-p_h)}{n_h - 1}\right)}$

以下表格总结了问题中 4 种情况下的预测比例与 95%置信区间：

Estimation		CI only	HypTest Only	Both	Neither
Percentage		0.098	0.281	0.646	0.071
95% CI	Lower limit	0.059	0.223	0.578	0.039
	Upper limit	0.136	0.339	0.703	0.106

（d）每篇文章作者数的均值估计为$\overline{y_{str}} = \frac{1}{N}\sum_{h=1}^{3} N_h \overline{y_h} = 6.106$，95%置信区间为$[\overline{y_{str}} \mp 1.96 SE(\overline{y_{str}})]$，$SE(\overline{y_{str}}) = \sqrt{\frac{1}{N^2}\sum_{h=1}^{3} N_h^2 \left(1 - \frac{n_h}{N_h}\right)\left(\frac{s_h^2}{n_h}\right)}$，可得 95%置信区间为[6.053, 6.159]

（e）不用 random selection 和 random assignment 的比例估计为 0.669，95%置信区间为[0.603, 0.731]，总文章数估计为 365，95%置信区间为[330, 400]（计算方法与（b）题一样）

（f）题目中所计算的统计量不能描述所有公共健康研究文章。这些统计量只适用于总体为题目中所提到的在三本杂志中的公共健康研究文章。

§3.8 Exercises：16

（a）Bushels of clams 总数估计$\widehat{t_{str}} = \sum_{h=1}^{4} N_h \overline{y_h} = 18153$，

标准误$SE(\widehat{t_{str}}) = \sqrt{\sum_{h=1}^{4} N_h^2 \left(\frac{s_h^2}{n_h}\right)} = 2412.5$

（b）Bushels of clams 总数估计$\widehat{t_{str}} = \sum_{h=1}^{4} N_h \overline{y_h} = 7229$，

标准误$SE(\widehat{t_{str}}) = \sqrt{\sum_{h=1}^{4} N_h^2 \left(\frac{s_h^2}{n_h}\right)} = 971.5$

§3.8 Exercises：19

（b）$\overline{y_u} = \overline{y_{str}} = \frac{1}{N}\sum_{h=1}^{3} N_h \overline{y_h} = 3.939$，95%置信区间为$[\overline{y_{str}} \mp 1.96SE(\overline{y_{str}})]$，$SE(\overline{y_{str}}) = \sqrt{\frac{1}{N^2}\sum_{h=1}^{3} N_h^2 \left(1 - \frac{n_h}{N_h}\right)\left(\frac{s_h^2}{n_h}\right)}$，可得 95%置信区间为[3.854, 4.023]

（c）进行假设检验，。检验$\begin{cases} H_0: \mu_1 = \mu_2, H_1: \mu_1 < \mu_2 \\ H_0: \mu_1 = \mu_3, H_1: \mu_1 < \mu_3 \\ H_0: \mu_2 = \mu_3, H_1: \mu_2 < \mu_3 \end{cases}$，检验统计量$U_{ij} = \frac{\overline{y_i} - \overline{y_j}}{\sqrt{\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}}}$，当$U \le u_{0.05} = -1.64$时拒绝原假设。$U_{12} = -0.8, U_{13} = -1, U_{23} = -0.2$，都大于$u_{0.05}$。因此，在$\alpha = 0.05$时，没有显著的证据表明各层间价格不一样。

§3.8 Exercises：24

（a）由于是电话调查，可能的非抽样误差有：无应答样本、样本全由自愿回答者构成、被调查者随意回答等。

（b）将变量 *countyname* 作为各层的标志，样本的抽样权重即为各层的 *popsize* / *sampsize*

（c）在抽样过程中，有的层中样本量仅为 1，无法计算其s^2与SE，为此，在计算置信区间时，我们直接忽略这些分层样本

平均 radon 值：$\overline{y_{str}} = \frac{1}{N}\sum_{h=1}^{H} N_h \overline{y_h} = 0.426$，95%置信区间$[\overline{y_{str}} \mp 1.96SE(\overline{y_{str}})]$，$SE(\overline{y_{str}}) = \sqrt{\frac{1}{N^2}\sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right)\left(\frac{s_h^2}{n_h}\right)}$，可得 95%置信区间为[0.388, 0.464]

平均 log(radon)值$\overline{y_{log,str}} = \frac{1}{N}\sum_{h=1}^{H} N_h \overline{y_{log,h}} = 0.107$，95%置信区间$\overline{y_{log,str}} \mp 1.96SE\left(\overline{y_{log,str}}\right)$，$SE\left(\overline{y_{log,str}}\right) = \sqrt{\frac{1}{N^2}\sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right)\left(\frac{s_{log,h}^2}{n_h}\right)}$，可得 95%置信区间为[0.100, 0.113]

（d）radon level 超过 4 的比例估计：$\widehat{p_{str}} = \frac{1}{N}\sum_{h=1}^{H} n_h \overline{p_h} = 0.0375$，95%置信区间：$[\widehat{p_{str}} \mp 1.96SE(\widehat{p_{str}})]$，其中，$SE(\widehat{p_{str}}) = \sqrt{\sum_{h=1}^{3} \frac{N_h^2}{N^2}\left(1-\frac{n_h}{N_h}\right)\left(\frac{p_h(1-p_h)}{n_h-1}\right)}$，可得 95%置信区间为 [0.03377, 0.04121]

§3.8 Exercises：49

（a）全美卡车总量估计$N = \sum_{h=1}^{H} w_h n_h = 85174776$，估计量的标准差几乎为 0 是因为$w_h$中已经包含总体信息，估计值只会存在四舍五入的偏差。

（b）总卡车里程估计$\widehat{t_{str}} = \sum_{h=1}^{H} N_h \overline{y_h} = 1.115 \times 10^{12}$，95%置信区间$[\overline{t_{str}} \mp 1.96SE(\overline{t_{str}})]$，$SE(\overline{t_{str}}) = \sqrt{\sum_{h=1}^{H} N_h^2\left(1-\frac{n_h}{N_h}\right)\left(\frac{s_h^2}{n_h}\right)}$，可得 95%置信区间为$[1.102\times 10^{12}, 1.127\times 10^{12}]$

（c）可以将每一个卡车类别对应的样本拿出来构成一个新的样本总体，样本总体有 51 层，分别估计这些样本总体的总卡车里程及 95%置信区间。

以下表格总结了对 5 类卡车的总卡车里程估计及 95%置信区间：

		1	2	3	4	5
总卡车里程（$\times 10^{10}$）		42.83	54.11	4.128	3.175	7.230
95%置信区间	Lower Limit	41.91	53.25	4.050	3.107	7.128
	Upper Limit	43.75	54.97	4.206	3.244	7.332

（d）根据相关文档（https://search.r-project.org/CRAN/refmans/SDAResources/html/vius.html），当 *vius_gvw* 大于 3 时，卡车总重超过 10000 磅

卡车总重超过 10000 磅的卡车总量估计：$\widehat{t_{str}} = \sum_{h=1}^{H} N_h \widehat{p_h} = 5415209$，95%置信区间$[\overline{t_{str}} \mp 1.96SE(\overline{t_{str}})]$，$SE(\overline{t_{str}}) = \sqrt{\sum_{h=1}^{H} N_h^2\left(1-\frac{n_h}{N_h}\right)\left(\frac{p_h(1-p_h)}{n_h-1}\right)}$，可得 95%置信区间为 [5287155, 5543262]

附录：

解答题目所使用的 R 代码及输出请见 https://thisiskunmeng.github.io/sampling/hw5.html