《抽样调查》期末答题纸

1907402030 熊雄

一.

(1) 答: ①. 目标总体: 该城市所有建筑公司之建筑师

②. 抽样框: 该城市14位建筑师在电话簿中之名录

③. 抽样单元: 最先同意受访的8位建筑师之每一位.

④. 观测单元: 最先同意受访之8位建筑师之每一位.

选择偏差/不精确之来源:

①. 每位建筑师的入样概率不等, 不是随机抽样

②. 未能得到所选定样本之回答, 从而产生无回答偏差.

③. 样本总量偏少, 不精确. ④. 建筑师可能只给出他们想听到之回答. ⑤. 访问员的措辞与语义可能产生影响.

(2). 答: A ①. 目标总体: 北京和西欧的 BRCA1 突变基因携带者.

② 抽样框: 北京和西欧33个家庭之名录.

③ 抽样单元: 被抽中之家庭的每一位家庭成员.

B. 不能精确提供, 原因如下:

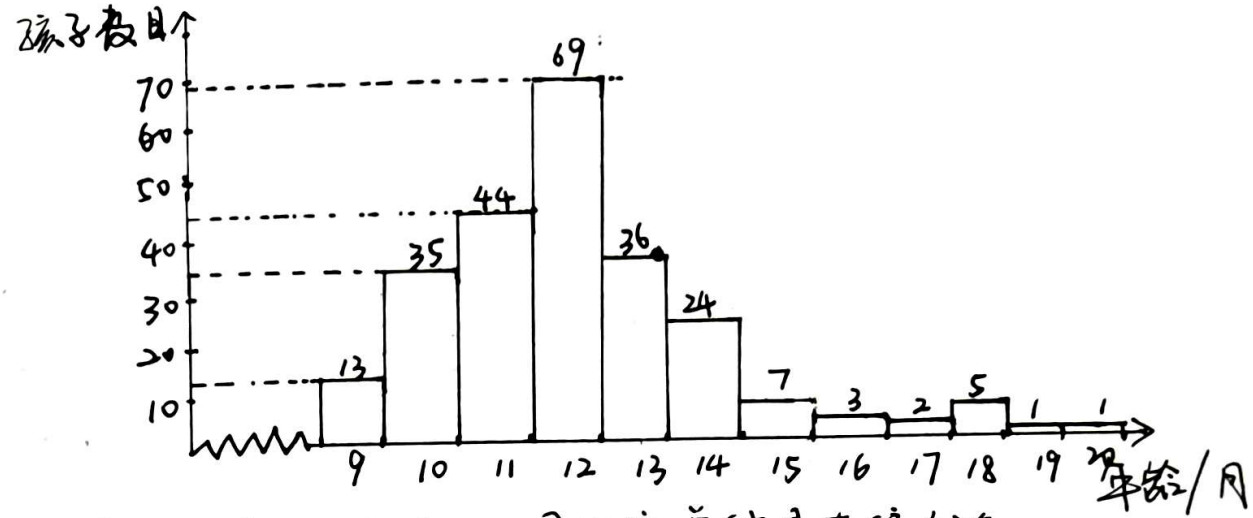
①. 调查对象中的每个家庭至少有4个人在60岁之前已被诊断有乳腺癌或卵巢癌, 因此这样选取样本是有偏之, 并不是随机抽样.

② 抽样单元之数目过少, 不能作出准确之估计.

#.

二.

答:(1) 开始走路的年龄直方图如下:



形状不是正态分布, 因为该总体是右偏分布,

是卡尔分布.

样本均值的抽样分布为正态分布, 由中心极限定理可知

均值的分布服从 $N(\mu, \frac{\sigma^2}{n})$

(2). 样本均值 $\bar{y} = \frac{1}{240}\sum_{i=1}^{240} y_i \approx 12.0792$

$$S^2 = \frac{1}{240-1}\cdot\sum_{i=1}^{240}(y_i-\bar{y})^2 \approx 3.6848$$

$$\Rightarrow \quad S = \sqrt{3.6848} \approx 1.9206$$

$$S(\bar{y}) = \sqrt{V(\bar{y})} = \frac{S}{\sqrt{n}} \approx 0.124.$$

取 $\alpha = 0.05$, $u_\alpha = 1.96$

则 独立走路的平均年龄的置信区间为 $[\bar{y} - u_\alpha\cdot\sqrt{V(\bar{y})}, \bar{y} + u_\alpha\cdot\sqrt{V(\bar{y})}]$

代入计算得 95% 的置信区间为 $[11.0862, 12.3222]$.

(3). 边际误差 $d = 0.5$, $\alpha = 0.05$, $u_\alpha = 1.96$, $S = 1.9206$

则 $n_0 = \left(\frac{u_\alpha \cdot S}{d}\right)^2 = \left(\frac{1.96 \times 1.9206}{0.5}\right)^2 = 56.68 \approx 57$

$$\Rightarrow n = \frac{n_0}{1+\frac{n_0}{N}} = \frac{57}{1+\frac{57}{240}} \approx 47$$

因此需要的样本量为 $n = 47$

#

三.（分层抽样）

解：(a). $N = 90000$, $N_1 = 35000$, $N_2 = 45000$, $N_3 = 10000$

$n = 900$. $S_1 = 2S_2 = 2S_3$. 假设每层单位抽样费用相等，采用 Neyman 分配n方法：$n_h = n \cdot \frac{N_h \cdot S_h}{\sum_{h=1}^{3} N_h S_h}$

则 $n_1 = 900 \times \frac{35000 \times S_1}{125000 \times S_2} = 504$（户）

$n_2 = 900 \times \frac{45000 \times S_2}{125000 \times S_2} = 324$（户）

$n_3 = 900 \times \frac{10000 \times S_3}{125000 \times S_2} = 72$（户）

即：应抽取独立建筑 504 户，楼房单元 324 户，公管房 72 户.

(b). ①. 简单随机抽样

$n_{SRS} = 45\% \times N_1 + 25\% \times N_2 + 3\% \times N_3 = 27300$

$p = \frac{n_{SRS}}{N} \approx 30.33\%$ $\quad q = 1 - p = 69.67\%$

则 $v(p) = \frac{N - n}{N \cdot (n-1)} \cdot pq \approx 0.0002327$

②. 按比例配置n分层抽样.

则方差为：

$$V(p_{prop}) = \frac{1-f}{nN} \cdot \sum_{h=1}^{3} \frac{N_h^2 \cdot P_h \cdot Q_h}{N_h - 1}$$

$$= \frac{1 - \frac{900}{90000}}{900 \times 90000} \times \left(\frac{35000^2 \times 45\% \times 55\%}{35000 - 1} + \frac{45000^2 \times 25\% \times 75\%}{45000 - 1} + \frac{10000^2 \times 3\% \times 97\%}{10000 - 1} \right)$$

$$\approx 0.0002126 < 0.0002327$$

综上，按比例配置n效果优于简单随机抽样，且

$deff = \frac{0.0002126}{0.0002327} = 91.36\%$，分层抽样效率相对较高. #

四.

解: $n=12$. 计算得.

$\sum_{i=1}^{n} x_i = 16352$, $\sum_{i=1}^{n} x_i^2 = 2804330$, $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 1362.67$

$\sum_{i=1}^{n} y_i = 18459$, $\sum_{i=1}^{n} y_i^2 = 30405031$, $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = 1538.25$

$\sum_{i=1}^{n} x_i y_i = 27016552$

则 $\hat{Y}_R = \hat{R} \cdot x = \frac{18459}{16352} \times 86436 = 97573.52$ (万元)

$V(\hat{Y}_R) \approx MSE(\hat{Y}_R)$

$\approx \frac{N^2(1-f)}{n} \cdot \frac{\sum_{i=1}^{n} (y_i - \hat{R}x_i)^2}{n-1}$

$= \frac{N^2 \cdot (1-f)}{n \cdot (n-1)} \cdot \left(\sum_{i=1}^{n} y_i^2 + \hat{R}^2 \cdot \sum_{i=1}^{n} x_i^2 - 2\hat{R} \cdot \sum_{i=1}^{n} x_i y_i \right)$

$= 909860.5376$

$\Rightarrow se(\hat{Y}_R) = \sqrt{V(\hat{Y}_R)} = 953.87$

取 $\alpha = 0.05$. $u_\alpha = 1.96$

则今年总产值在 95% 置信区间为 $(97573.52 \pm 1.96 \times 953.87)$ 万元,

计算后, 即: $[97703.9348, 99443.1652]$ 万元. #

五.

解：(a). 因为该调查有名在一周时间内在特定地区进行简单随机抽样选取部分诊所，然后调查所有因健康原因寻求患病儿童去过该诊所的所有家长，因此这是一个整群样本.

① 初级单元：Chicago地区被选定的诊所

② 次级单元：一周时间内因为健康原因寻求患病儿童而去过选定诊所的所有家长.

该整群抽样为二阶整群抽样

我们令 $Y_{ij} = \begin{cases} 1, & \text{若该家庭有枪支} \\ 0, & \text{若该家庭无枪支} \end{cases}$

我们记 Chicago地区的诊所数目为 N

抽取的诊所数目为 n. $(M = \sum_{i=1}^{N} M_i)$

第 i 个 $(i = 1, 2, \cdots, N)$ 诊所抽取的家庭个数为 M_i

样本中第 i 个 $(i = 1, 2, \cdots, n)$ 诊所抽取的家庭为 m_i

第 i 个诊所中有枪支的家庭个数记为 a_i

则 $\hat{P} = \frac{N}{n \cdot M} \cdot \sum_{i=1}^{n} \frac{M_i \cdot a_i}{m_i}$ 为比例的估计量

(b). 抽样总体为 最近一周时间内因为健康原因寻求患病儿童而去过Chicago地区被选定诊所的所有家庭.

不能得到有儿童家庭的代表性样本，原因如下：

① 抽样总体位于 Chicago，该地区犯罪率高，拥有枪支的家庭较多.

② 只能选择一周内在 Chicago 地区指定诊所就诊的家庭，具有选择偏差.

#

六、

解：该抽样为初级单元大小相等时的二阶抽样

$n=4$，$m=10$。

总体均值的估计为：

$$\bar{y} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}$$

$$= \frac{1}{n}\cdot\sum_{i=1}^{n}\bar{y}_i$$

$$= \frac{1}{4}(2+2.2+1.7+2.3)$$

$$= 2.05 \approx 2 \text{(份)}.$$

因此该小镇平均每户订阅杂志约为2份。 #