2020 年《抽样调查》期末考试

对如下抽样调查，描述其目标总体，抽样框，抽样单元和观测单元。讨论可能的选择偏差或响应不精确的来源。

1. 1994 年 6 月期的《个人电脑世界》（PC World）有一篇关于个人电脑可靠性与维修保障的报道，其中一个结论是“25%的新个人电脑存在问题”，这是 1994 年 5 月 23 日《美国近日报》（USA Today）的头版头条从 1993 年 10 月起，《个人电脑世界》每期都会有一篇有关用户硬件问题的调查报告。每月调查的回答者都有机会赢得一台新电脑，有 45000 以上的人接受了调查。

答：① 目标总体：所有使用个人电脑的用户.

② 抽样框：《个人电脑世界》的订阅者名录.

③ 抽样单元：个人.

④ 观测单元：每个用户电脑是否存在问题.

⑤ 可能存在的选择偏差或响应不精确的来源：

可能存在选择偏差，因为抽样框为《个人电脑世界》的订阅者名录，不包括所有的拥有个人电脑的用户，同时，也有可能部分未拥有个人电脑却参与调查的用户，他们的目的是获得奖品，这会影响调查结果. #

2. 1996 年 8 月的《消费者报道》中提供了由杂志读者对其接受服务的各种健康维护组织（HMOs）的满意评分。编辑们对调查作出如下说明：“评分时根据 1995 年度回收的 20000 铎份关于从 1994 年 5 月到 1995 年 4 月在 HMOs 的经历的问卷得到的。这些结果反映了《消费者报道》订阅者们的经历，他们是美国人口中更富有且更有教育的群体”（40 页）。试确定该调查的目标总体，抽样框与抽样单元。你认为该调查对于选择健康计划提供了有用的消息吗？如果你为自己选择一个 HMO，你更愿意参考哪类信息：是该项调查的结果还是由 HMOs 自行实施的顾客满意度调查的结果？

答：① 目标总体：《消费者报道》的读者.

② 抽样框：《消费者报道》读者的名录.

③ 抽样单元：个人.

④ 观测单元：对 HMOs 的满意评分.

⑤ 我认为该调查为选择健康计划提供了有用的消息，因为这是根据回收的20000份问卷得到的.

⑥. 如果我选择一个 HMO，我就会选择该项调查结果，因为该结果有 HMOs 的对比，可以让我对各个 HMO 进行比较，选择适合自己的健康计划.

#

3. 马瑞特（1994）从去过他们诊所小儿门诊就诊的 2 到 6 岁的儿童里，简单随机抽取了 240 个。他们发现这些孩子当中开始走路的频数分布为：

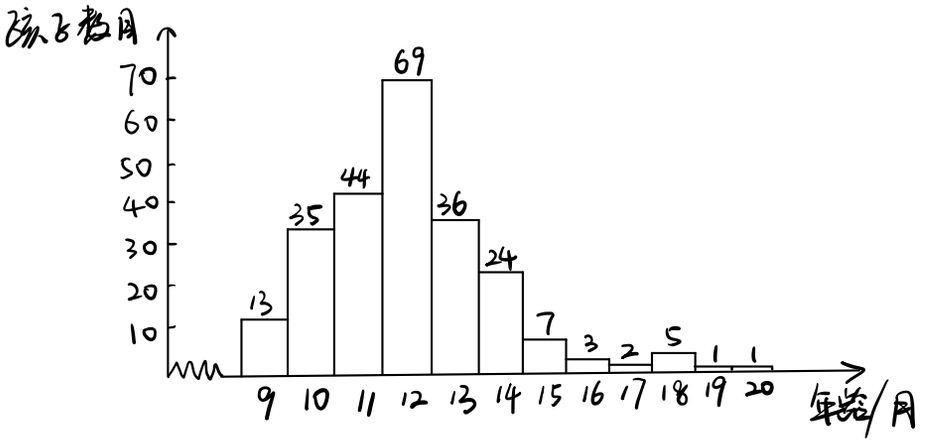
年龄（月）	9	10	11	12	13	14	15	16	17	18	19	20
孩子的数目	13	35	44	69	36	24	7	3	2	5	1	1

1) 构造一个开始走路年龄分布的直方图。形状是否是正态分布？你认为样本均值的抽样分布是否是正态分布？为什么？

2) 求均值，标准差以及独立走路的平均年龄的 95%置信区间。

3) 假如研究者希望在另一个地方做另一个研究，想得到边际误差为 0.5 的开始走路年龄的 95%的置信区间。使用这些数据的估计标准差，他们需要多大的样本量？

答：(1) 开始走路年龄直方图如下：



形状不是正态分布，因为该总体是右偏分布，是长尾分布。

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)

$$\bar{y} = \frac{1}{240}\sum_{i=1}^{240} y_i \approx 12.0792$$

$$S^2 = \frac{1}{240-1}\cdot\sum_{i=1}^{240}(y_i-\bar{y})^2 \approx 3.6848 \Rightarrow S \approx 1.9206$$

$$\sqrt{V(\bar{y})} = \frac{S}{\sqrt{n}} \approx 0.124, \quad u_2 = 1.96$$

独立走路的平均年龄的置信区间为

$$\left[\bar{y} - u_\alpha \cdot \sqrt{V(\bar{y})},\ \bar{y} + u_\alpha \cdot \sqrt{V(\bar{y})}\right]$$

代入计算知独立连路的平均年龄的95%置信区间为

$[11.0862,\ 12.3222]$.

(3). 边际误差 $d = 0.5$，置信水平 $\alpha = 0.05$，$S = 1.9266$

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

$$\Rightarrow n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{57}{1 + \frac{57}{240}} = \frac{57}{1.2375} \approx 47$$

因此需要的样本量为 $n = 47$.

#

4. 1995 年 12 月发表在《获胜变化之谜》上的一封信中说道：“我注意到在你们的比赛中，前面几期中没有来自南方的获胜者。你们一直说获胜者是随机抽取的，那是否意味着南方的参与者较少？”作为回复，编辑从前面几期比赛中随机抽取了 1000 个单元作为样本，发现其中有 175 个来自南方。

1）求来自南方的参与者比例的 95%置信区间。

2）根据美国统计摘要，30.9%的美国人口生活在编辑认为是南方的州。从你的置信区间中是否有证据说明来自南方的参与者与居住在南方的人口比例之间有差异？

解：(1). $\hat{p} = \frac{175}{1000} = 0.175$, $\hat{q} = 1 - \hat{p} = 0.825$

$$\sqrt{V(\hat{p})} = \sqrt{\frac{\hat{p}\cdot\hat{q}}{n}} = \sqrt{\frac{0.175\times 0.825}{1000}} \approx 0.01202$$

$u_\alpha = 1.96$

来自南方的参与者比例的 95% 置信区间为

$$\left[\hat{p} - u_\alpha\cdot\sqrt{V(\hat{p})},\ \hat{p} + u_\alpha\cdot\sqrt{V(\hat{p})}\right].$$

代入计算知来自南方的参与者比例的 95% 置信区间为

$$[0.1514,\ 0.1986].$$

(2). 由(1)可知，30.9% 不在置信区间内，因此没有证据表明来自南方的参与者与居住在南方的人口比例之间有差异.

#

5. 在一所高校学生调查中，调查学生中对如下陈述赞同的比例：当我看到我的专业的主要期刊上的一个新问题时，我很少发现一篇是我感兴趣的文章。数据如下：

Discipline	Membership	Number Mailed	Valid Returns	Agree(%)
Literature	9100	915	636	
Classics	1950	633	451	37
Philosophy	5500	658	481	23
History	10850	855	611	23
Linguistics	2100	667	493	29
Political Science	5500	833	575	19
Sociology	9000	824	588	43
Totals	44000	5385	3835	41

(a) 调查中抽样总体是什么？

(b) 给出一个陈述赞同者的比例的估计量，并计算估计量的标准差。

解：(a) 抽样总体：该高校中上述专业的全体学生

(b). $w_i = \frac{N_i}{N}$. $i = 1, 2, \cdots, 7$

$$\hat{p} = \sum_{i=1}^{7} w_i p_i \approx 0.3337$$

$$V(\hat{p}) = \sum_{h=1}^{7} w_h^2 \cdot (1 - f_h) \cdot \frac{\hat{p}_h(1-\hat{p}_h)}{n_h}$$

$$\sqrt{V(\hat{p})} \approx 0.00636$$

6. 某系统有 56 个公司，去年全系统总产值 86436 万元。为估计今年的总产值，年底在所辖全部公司中随机抽取 12 个公司进行调查，所得如下表，其中 x_i 和 y_i 分别为去年和今年的产值，试用比估计方法估计今年总产值的估计值。

公司	1	2	3	4	5	6	7	8	9	10	11	12
x_i	764	1642	957	1324	2131	1176	1618	1532	834	1432	1728	1214
y_i	853	1835	1028	1512	2286	1354	1826	1721	958	1648	1904	1534

解：计算可知：($n=12$)

$\sum_{i=1}^{n} x_i = 16352$，$\sum_{i=1}^{n} x_i^2 = 28043730$，$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 1362.67$

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

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

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

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

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

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

$= 909860.5376$

所以 $Se(\hat{Y}_R) = \sqrt{V(\hat{Y}_R)} = 953.87$，$u_\alpha = 1.96$

因此今年总产值的估计值的 95% 的置信区间为

$(97573.52 \pm 1.96 \times 953.87)$ 万元，即

$[97703.9348, 99443.1652]$ 万元。

#

7. 一位研究者为研究一个地区高中女生吸烟的流行情况，他从这一地区 29 所高中学校中抽取了 4 所高中。数据如下：

样本序号	学生总数	女生数	接受采访女生数	吸烟者数
1	1471	792	25	10
2	890	447	15	3
3	1021	511	20	6
4	1587	800	40	27

（a） 估计这一地区高中女生的吸烟比例，并给出 95%的置信区间。

（b） 估计这一地区高中女生吸烟者的总数，并给出 95%的置信区间。

解：

(a). $N=29$, $n=4$, $u_\alpha=1.96$

$$\sum_{i=1}^{n} M_i p_i = 792\times\frac{10}{25} + 447\times\frac{3}{15} + 511\times\frac{6}{20} + 800\times\frac{27}{40}$$

$$= 1099.5$$

$$M_0 = \sum_{i=1}^{n} M_i = 2550$$

$$\Rightarrow \hat{p} = \frac{\sum_{i=1}^{n} M_i p_i}{M_0} = \frac{1099.5}{2550} \approx 0.4311765$$

$$V(\hat{p}) = \frac{N^2\cdot(1-f_1)}{M_0^2\cdot n}\cdot\frac{\sum_{i=1}^{n} M_i(p_i-\hat{p})^2}{n-1} + \frac{N}{M_0^2\cdot n}\cdot\sum_{i=1}^{n}\frac{M_i^2\cdot(1-f_{2i})}{m_i}S_{2i}^2$$

$$\approx 0.018408$$

$$\Rightarrow \sqrt{V(\hat{p})} \approx 0.135676$$

用此该地区高中女生吸烟比例的 95% 置信区间为 (0.4311765 ± 0.135676). 即：$[16.5251\%, 69.7102\%]$.

(b). $\hat{Y} = \frac{M_0}{n} \cdot N = 18487.5 \Rightarrow \hat{\hat{Y}} = \hat{Y} \cdot \hat{p} = 7971.376$

$\sqrt{V(\hat{\hat{Y}})} = \hat{Y} \cdot V(\hat{p}) = 1081.524$, $u_2 = 1.96$

则此，抽烟女生人数估计的 95% 置信区间为

$(18487.5 \pm 1.96 \times 1081.524)$，即

$[16367.71, 20607.29]$.

#