名词解释A

1. 非概率抽样
   抽取样本时并不依据随机原则，单元入样概率不知，无法计算抽样误差，不具备统计推断意义。

2. 概率抽样
   依据随机原则，按照某种实现设定的程序，从总体中抽取部分单元的抽样方法。用于统计推断，侧重量的分析。

3. 允许抽样误差
   $\Delta = t\sqrt{V(\bar{y})}$ 用样本对总体真值进行估计的时候，用样本的估计值减和加置信区间
   反映置信区间的下限和上限长度是两倍的允许抽样误差，在推断时可以接受的，以多大的概率把握

4. 目标总体
   所要研究对象的总体

5. 抽样总体
   产生样本的总体，即样本是从这个总体中产生的，和目标总体不是完全一致的

6. 抽样框
   样本产生的总体，抽样总体的具体表现是抽样框。是一份包含所有抽样单元的名单，给每个抽样单元边上一个号码，就可以按一定的随机化程序进行抽样。

7. 总体参数
   对总体而言，抽样调查目的是得到总体的某些特征，统计中把这些总体特征称为参数，是我们所关心的总体某个或某些方面的数量表现。例如总体总值，总体均值，总体比例，总体比率。

8. 统计量
   对样本而言，也叫估计量。从总体中按一定程序抽出的部分总体基本单元的集合称为样本，根据样本单元计算出的一个量，实现对总体参数的估计。常用统计量有样本均值，样本比例等

9. 估计量方差
   估计量分布的方差成为估计量方差，它是从平均意义上说明估计值与待估参数的差异状况。表达式为

   $$V(\hat{\theta}) = E[\hat{\theta} - E(\hat{\theta})]^2$$

   由于抽样的随机性而产生的随机性误差，没有系统性

10. 偏差
    偏差是指按某一抽样方案反复进行抽样，估计值的数学期望与待估参数之间的离差，表达式为

    $$E(\hat{\theta}) - \theta$$

    偏差是，偏于某个方向的系统性误差，并不随样本量增大而减。

11. 无回答偏差 调查过程中由于无回答或回答有误造成的误差，引起估计值与总体参数之间的差异。是一种非抽样误差

12. 抽样误差
    由于抽样随机性带来的样本统计量和总体参数之间的差异，抽样误差可以计算，可以控制。估计量方差$V(\theta)$和估计量标准差$\sqrt{V(\theta)}$都是抽样误差的表现形式。

13. 非抽样误差
    除了抽样误差以外各种原因造成的样本统计量和总体参数之间的差异。例如抽样框误差，无回答误差，计量误差。

14. 简单随机抽样
    最简单的，不加任何附加条件的随机抽样。其他抽样方法都是在其上基础的上丰富和发展。

15. 等概率抽样
    是概率样本，每个单元被选入样本的概率是相等的，由于样本单元权数相等，因此数据处理中的权数就可以忽略，从而计算更简洁方便。

名词解释B

1. 不等概率抽样
   是概率样本，每个单元被选入样本的概率是不相等的，样本单元权数也不相等，因此数据处理中的权数就可以忽略，从而计算更简洁方便。

2. 有限总体修正系数1-f
   $f = n/N$为抽样比，1-f为有限总体校正系数，反映总体未入样单元所占的比例。有限总体校正系数是不放回抽样对有放回抽样误差的修正。

3. 有放回抽样
   每次都是从N个总体单元中随机抽取1个单元，独立重复抽取n次，得到n个单元组成的样本

4. 无放回抽样
   每次都是从剩下的总体单元中随机抽取1个单元，独立重复抽取n次，得到n个单元组成的样本

5. 分层抽样
   在抽样之前，先将抽样单元按某种特征或某种规则划分为不同的层，然后从不同的层中分别独立地抽取样本，将各层的样本结合起来，对总体的目标量进行估计。这种抽样就是分层抽样。

   $$N = N_1 + N_2 + ... + N_L$$

   $$n = n_1 + n_2 + ... + n_L$$

   特点

   1. 可以提高估计精度
   2. 可以根据各层采用不同的抽样方法
   3. 可对各层进行估计

6. 分层原则

   - 层内方差尽可能小
   - 层间方差尽可能大 根据统计学方差分析原理(ANOVA)，总平方和(SST)有如下分解：
     SST = SSB + SSW
     其中，SST：Sum of Square Total，总平方和；
     SSB：Sum of Square Between；
     SSW：Sum of Square Within。
     在总体被划分之后，SSB 与SSW 是此消彼长的关系，若SSB 较小则SSW 较大，反之，所SSB 较大则SSW 较小。
     由于分层抽样的所有层都有基本单元入样，如果每层内的单元之间比较相似，每一层的方差

较小而各层之间的方差较大，那么这样只需在各层中抽取少量样本单元，就能很好地代表各层的特征。所以，分层 抽样的估计量方差（精度）只是与层内方差（平方和）有关，与层间方差（平方和）无关。因此，在分层抽样时，层内方差要尽可能小，层间方差要尽可能大。如此，才能使得分层抽样的精度达到最高。

7. 比例分配
   是在分层抽样中确定不同层分配样本量的一种方式，就是使各层样本量n_n与层权W_h成正比。这种分配的估计量有比较简单的兴衰，但是没有考虑各层的方差和调查费用。

8. 内曼分配
   是分层抽样中确定样本量比较好的一种方式，使n_h与W_hS_h成正比。这种分配估计量的方差最小。但是没有考虑费用。

9. 最优分配
   在内曼分配基础上，考虑了费用的因素，使n_h与$W_hS_h/\sqrt{c_h}$成正比。这种分配方式综合考虑了精度和调查费用。

10. 事后分层
    有些情况下事先分层是无法进行的，例如：

    - 没有层的抽样框，
    - 总体特别大来不及分层
    - 几个变量都适宜于分层，而要进行事先的多重交叉分层存在一定困难
      这时如果想利用分层抽样的优点，可以采用对样本的事后分层方法。即先抽取样本量n，然后调查得到n_i和$\bar{y_i}$，又已知W_i

$$\bar{y}_{pst}=\sum_{h=1}^{L}W_h\bar{y}_h$$

11. 目录抽样
    当$n_i>N_i$时，有一些体量特别大的单位，总体比较小，把他单独分为一个层，不再做抽样。只在其他一些小的单位进行。

12. 层界
    即层的界限。
    分层抽样比简单随机抽样效率高，若分层抽样使用为了便于组织估计子总体的参数，则分层是按自然层或单元类型划分的。
    如果分层是为了提高抽样效率，按调查目标量分层最好，但如果在调查之前并不知道相应调查目标量的值，则只能通过与调查目标量高度相关的辅助指标来进行

13. 层数
    设L 个子总体所包含的单位数分别为N1,N2,...,NL，则有
    N1+N2+...+NL=N 上式中的L 即为层数，即总体被划分为“不重不漏”的子总体的个数。
    在实际工作中，要保证每个层有样本单元，层数不能超过样本量。如果要给出估计量方差的无偏估计，则每层至少需要2个样本单元，那么层数不能超过样本量的一半，即n/2

14. 设计效应
    为了比较不同抽样方法的效率，可以通过抽样方法的设计效应来进行比较。其定义为一个特定的抽样设计方案的估计量方差与相同样本量下无放回简单随机抽样的估计量方差之比

$$deff=V(\bar{y}_{st})/V(\bar{y}_{srs})$$

根据定义可以知道简单随机抽样的deff = 1。若deff > 1，表明所考虑的抽样设计效率比无放回简单随机抽样低；若deff < 1，表明该抽样的效率比无放回简单随机抽样高.

作用如下：

1. 用所采用的抽样方案的估计量的方差和简单随机抽样估计量的方差的比值，来评价抽样方案的效应。
2. 推算复杂抽样设计的样本量：
   复杂抽样设计样本量=srs样本量*设计效应

15. 比率估计
    一种估计方法，利用目标变量与辅助变量的比值关系来提高估计效果的方法。
    作用：

    - 如果我们考察的指标本身就是比率，那么对考察对象总体比率的估计要使用比率估计
    - 利用辅助变量提高估计精度。

名词解释C

1. 分别比率估计
   抽样采用分层，估计时采用比率估计，分层之后先在各层获取比率估计，然后按照层权加权平均得到总体参数估计

$$r_1 = \frac{\sum y_1}{\sum x_1}, r_2 = \frac{\sum y_2}{\sum x_2}, ..., r_L = \frac{\sum y_L}{\sum x_L}$$

$$\bar{y}_{RS} = \sum W_h \bar{y}_{Rh} = \sum W_h \frac{\bar{y_h}}{\bar{x_h}} \bar{X}_h$$

2. 联合比率估计
   抽样采用分层，估计时采用比率估计，先分别对目标变量、辅助变量做分层估计，然后再采用比率估计方法,均值的联合比率估计为：

$$y_{\bar{R}c} = \frac{\bar{y_{st}}}{\bar{x_{st}}} \bar{X} = \hat{R}_c \bar{X}$$

3. πps抽样
   和pps含义一样，是与规模成比例的概率抽样。pps针对有放回的抽样，πps针对无放回抽样。特点是抽样误差计算复杂。

4. PPS抽样
   probability proportional to size，与规模成比例的概率抽样。

5. 整群抽样
   把总体划分为若干个群，以群为抽样单元，对群中的所有单位进行调查。

6. 分群原则
   群的划分有两种：

   1. 根据行政或地域形成的群体
   2. 调查人员人为确定的 分群原则: 群内差异尽可能大，群间差异尽可能小

   群内单元可以看做全面调查。

7. 群规模相等抽样
   在整群抽样中，如果总体中各群的基本单元数目相等，称为群规模相等抽样。实际中，只要群的大小接近，都可以视为群规模相等。

8. 群规模不等抽样
   在整群抽样中，如果总体中各群的基本单元数目不等，称为群规模相等抽样。

9. 系统抽样
   系统抽样又称等距抽样，它将总体N个单元按一定的顺序排列，随机抽取一个单元作为起始单元，然后按照某种确定的规则抽取其他样本单元。最简单的方式是等间隔抽取，即在随机抽取起始单元后，按照等间隔的距离抽取随后的样本单元。比简单随机抽样易于操作，在一些情况下可以提高估计效率，但方差计算较复杂。
   有三种情况

   - 按无关标志排列，单元排列顺序和要研究的变量无关
   - 按有关标志排列，单元排列顺序和要研究的变量有关。样本能够更好反应总体分布，在进行推断的时候精度更高
   - 按自然顺序排列，排列顺序自然形成的

   特点：

   - 便于操作
   - 便于审核
   - 在一些情况下可以提高估计效率
   - 估计量方差计算复杂

10. 直线等距抽样
    总体单元数为N，样本容量为n，当N是n的整数倍时，可以采用直线等距抽样，以$k = N/n$为抽样间距，把总体分为n段，从1至k之间随机抽取一个整数r，每隔k个单元抽出一个样本，知道抽满n个单元

11. 循环等距抽样
    当$N \neq n * k$时，即N不是n的整数倍时，如果仍然按直线等距抽样，则得到的结果是有偏的。为了得到无偏估计，将总体单元排列成一个首尾相接的圆，从1-N之间随机抽取一个整数r作为起始单元，每隔k抽取一个单元，直到抽满n个单元。

12. 对称等距抽样
    针对直线等距抽样的缺陷提出来的，按有关标排列的线性趋势总体，起始单元是两个，使低标志值的单元与高标志值的单元在样本中对等出现。

13. 修正直线等距抽样
    当$N \neq n * k$时，即N不是n的整数倍时，为了得到无偏估计，将k取最接近N/n的整数，在1至N之间随机抽取一个整数r，将r除以k的余数作为起始点，再以间隔k等距抽样。若余数为0，则以k为起始点。

14. 系统样本内相关系数
    反映了系统抽样的样本各单元相似性或相关性大小。

$$\rho_{wsy} = \frac{E(y_{rj} - \bar{Y})(y_{ru} - \bar{Y})}{E(y_{rj} - \bar{Y})^2}$$

若样本单元分布均匀，$\rho < 0$
若样本单元分布随机，$\rho = 0$，趋近于srs
若样本单元分布聚集，$\rho > 0$

15. 交叉子样本
    处理有周期性波动的总体的一种系统抽样。把一个大的样本，分成若干个小的样本。多个起点。
    交叉样本的设计方法：
    原系统样本k=N/n, 将n分层m份，新抽样间隔k'=mk, 每个子样本容量为 n'=n/m, 第a个子样本的均值

$$\bar{y}_\alpha = \frac{1}{n'}\sum_{j=1}^{n'} y_{ij}$$

    再将m个子样本的均值平均，得到总体均值的估计量为

$$\hat{\bar{Y}} = \bar{y}_{st} = \frac{1}{m}\sum_{\alpha=1}^{m} \bar{y}_\alpha$$

    则此时$V(\bar{y}_{st})$的估计量为

$$v(\bar{y}_{st}) = \frac{1}{m(m-1)}\sum_{\alpha=1}^{m}(\bar{y}_\alpha - \hat{\bar{Y}})^2$$

名词解释D

1. 多阶段抽样
   假设总体由N 个初级单元组成，每个初级单元又由若干个二级（次级）单元组成，先在总体中按一定方法抽取n 个初级单元，对每个抽中的初级单元再抽取若干二级单元进行调查，这种抽样方法称为二阶段抽样(two-stage sampling)（也称二阶抽样、二级抽样），如果每个二级单元又由更小的三级单元组成，那么在第二阶段抽样后，若在每个被抽中的二级单元中再进行三级单元的抽样，则是三阶段抽样（三阶抽样）。同样的道理，还可以定义更高阶段抽样。对于二阶段以上的抽样，称为多阶段抽样。 特点：
   - 构造抽样框相对容易，使得大范围抽样变为可能
   - 样本单元分布相对集中，具有较强的可操作性
   - 不同阶段中抽样方法可以灵活，有利于抽样效率的提高
   - 可用于“散料”的抽样

2. 二阶段抽样
   假设总体由N 个初级单元组成，每个初级单元又由若干个二级（次级）单元组成，先在总体中按一定方法抽取n 个初级单元，对每个抽中的初级单元再抽取若干二级单元进行调查，这种抽样方法称为二阶段抽样(two-stage sampling)（也称二阶抽样、二级抽样）

3. 自加权样本
   在自加权抽样设计方案中，每个样本单元的设计权数都是相同的，即每个单元最终入样的概率是相等的。在各种抽样方法下，如果总体中每一个最基本单元被抽中的概率相等，这种抽样设计就称为自加权设计，这样的样本就称为自加权(self-weighting)样本

4. 入样概率

5. 包含概率

6. 汉森-赫维茨估计量
   令$Z_i = \frac{M_i}{\Sigma M_i}$为第i个单元（群）入样概率
   $\Sigma M_i = M$
   均值估计量为：

$$\hat{\bar{Y}}_{pps} = \frac{1}{n}\sum_{i=1}^{n}\bar{Y}_i$$

   均值估计量的方差为：

$$V(\hat{\bar{Y}}_{pps}) = \frac{1}{n}\frac{\sum_{i=1}^{n}(\bar{Y}_i - \hat{\bar{Y}}_{pps})^2}{n-1}$$

   总量的估计量为：

$$\hat{Y}_{pps} = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{Z_i} = \frac{\Sigma M_i}{n}\Sigma\bar{Y}_i = \frac{M}{n}\Sigma\frac{Y_i}{M_i}$$

   总量估计量的方差：

$$V(\hat{Y}_{pps}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\frac{y_i}{z_i} - \hat{Y}_{pps})^2$$

7. 二重抽样
   先从总体中随机抽取一个样本量较大的样本，对其进行简单调查以获取有关总体的结构或者辅助信息，为下一步抽样提供条件；然后再从中随机抽取一个样本量较小的样本对总体研究的目标量进行估计，称为二重抽样。 作用：
   - 进行样本筛选
   - 进行事后分层，抽选好的样本，是的样本结构与总体结构一致
   - 获得辅助信息的估计，提高估计效率
   - 用于对无回答的调整

8. 抽样中的权数

9. 样本量
   从总体中抽取若干单元的集合的数量叫样本量，用n表示

10. 随机化回答
    是针对敏感问题或者高度隐私问题的调查，采用的一种特殊方法。这种抽样调查基于概率理论，通过设计一套或若干套的特殊问答卡片，由受访者随机抽取卡片回答问题，然后再收集答卷，通过样本数据对目标总体进行推断。
    技术特征：
    - 随机化回答方式
    - 对受访者起到保密作用
    - 同时可以对目标参数进行推断

11. 逆抽样方法
    捕获再捕获法有时会遇到第二次捕获的样本中没有标记单元，即r=0的特殊情况，或者预估的第二次捕获样本量比实际需要的大很多。针对这种情况霍尔丹提出一种逆抽样方法： 逆抽样法的实施流程为：先获取数量为n1 的样本并在样本上进行标记，然后将这些样本放回总体，在已标记的样本充

分融入总体后，进行第二次抽样。在这一阶段中，事先确定一个整数r(r>1)，然后进行逐个抽样，直到抽到r 个具有标记特征的单元为止。设抽到r 个具有标记特征的单元时实 际抽取的样本量为n，根据第二次捕获时总体分布结构与第一次捕获时的结构相同的原理推算出：
N的估计为

$$\hat{N} = \frac{n_1 n}{r}$$

方差估计为

$$v(\hat{N}) = \frac{n^2 n(n-r)}{r^2(r+1)}$$

12. 沃纳模型
    沃纳模型是沃纳(Warner)于1965 年首先提出的一种随机化回答模型。该模型是为了解决社会经济现象中敏感性问题而采用的一种随机化回答技术。 模型设计的基本思想是：
    为了调查某个敏感问题或者高度隐私问题，同时列出两个存在相关关系的问题制成卡片（一个问题是具有某种特征，记为卡片A；另一个问题是不具有某种特征，记为卡片B），卡片A所占比例为P，卡片B所占比例为1-P。被调查者随机抽取卡片（调查人员不知道抽取结果）进行回答，答案只能是“是”或“不是”

    弱点：
    A,B卡片均为敏感问题
    P不能等于1/2

13. 西蒙斯模型
    针对沃纳模型的弱点提出的一种改进。
    模型设计的基本思想是：为了调查某个敏感问题或者高度隐私问题，同时列出两个无关的问题制成卡片（一个问题是具有某种特征，记为卡片A；另一个问题是无关问题，记为卡片B），卡片A所占比例为P，卡片B所占比例为1-P。

14. 多阶段抽样

15. 辅助变量