NOTE 1 Census 人口普查

Target population: we wish to study **Sample**: A subset of a population. **Sampled population**: we can study

Sample unit: household. **Observation unit**: individuals living in the household.

Sampling frame: (like telephone survey) a list of all numbers. ⑥ transcription from records

Method of data collection: ① physical observations ② personal interview ③ mail. ④ web-based. ⑤ registration

Two basic sampling designs: probability sampling. nonprobability sampling

probability sampling: known probability. can judge reliability and validity.

① simple random sampling: known equal change. 所有人名单拿出来随便抽

② systematic random sampling: ordered. 每5个抽1个.

③ stratified random sampling: 之间差很多. 内部差不多 [male | female] ← strata

④ cluster sampling: 分成 cluster. 之间差不多. 里面内容不一样. (如每个N学一个cluster). 不需要所有cluster

nonprobability sampling: unknown. selection is based on some intuitive knowledge

① convenience samples: (pilot study) 先抽n个问问. Assumption: target population is homogeneous. similar to overall

② judgement sampling: 找专家 ③ Quota sampling ④ Snowball sampling: 我问你→问你朋友→...

NOTE 2. Basic concepts of statistics.

Mean: $\mu = \sum y p(y)$. Variance: $E(y-\mu)^2 = \sum (y-\mu)^2 P(y)$. S.d. $\sigma^2 = E(y-\mu)^2 = E(y^2) - \mu^2$. $\Rightarrow Ey^2 = \sigma^2 + \mu^2$.

Infinite population: Sample Mean: $\bar{y} = \frac{1}{n}\sum y_i$ Sample Variance: $S^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$

$E(\bar{y}) = \mu$. $V(\bar{y}) = \frac{\sigma^2}{n}$. $E(\bar{y}^2) = \frac{\sigma^2}{n} + \mu^2$.

Finite population: Population total $\tau = u_1 + u_2 + \cdots + u_N$ mean: $\mu = \frac{\tau}{N}$ Variance: $\sigma^2 = \frac{1}{N}\sum_1^N (u_i - \mu)^2$.

⇒ **with replacement**. Population $\{u_1, \dots, u_N\}$. Selection prob. $\{\delta_1, \dots, \delta_N\}$. Sample $\{y_1, \dots, y_n\}$. $\tau = u_1 + \cdots + u_N$

$\hat{\tau} = \frac{1}{n}\sum_1^n \frac{y_i\ (=u_i)}{\delta_i} = \frac{1}{n}\sum_1^N a_i \frac{u_i}{\delta_i}$ (a_i是u_i出现次数) $E(\hat{\tau}) = E\left(\frac{1}{n}\sum_1^N a_i \frac{u_i}{\delta_i}\right) = \frac{1}{n}\sum_1^N n\delta_i \frac{u_i}{\delta_i} = \sum_1^N u_i = \tau$. (unbiased)

$V(\hat{\tau}) = \mathrm{Var}\left(\frac{1}{n}\sum_1^n \frac{y_i}{\delta_i}\right) = \frac{1}{n^2}\sum_1^n \mathrm{Var}\left(\frac{y_i}{\delta_i}\right) = \frac{1}{n}\sum_1^N (\frac{u_i}{\delta_i} - \tau)^2 \cdot \delta_i$ $\hat{V}(\hat{\tau}) = \frac{1}{n}\cdot\frac{1}{n-1}\sum_1^n (\frac{y_i}{\delta_i} - \hat{\tau})^2$ (also unbiased)

Ex: $\{1,2,3,4\}$ $\delta: \{0.1, 0.1, 0.4, 0.4\}$ for $n=2$.

Sample $\{1,2\}$. Prob $= 0.1\times 0.1 \times 2 = 0.02$, $\hat{\tau} = \frac{1}{n}\sum_1^n \frac{y_i}{\delta_i} = \frac{1}{2}(\frac{1}{0.1} + \frac{2}{0.1}) = 15$ $\hat{V}(\hat{\tau}) = \frac{1}{n}\cdot\frac{1}{n-1}\sum_1^n (\frac{y_i}{\delta_i} - \hat{\tau})^2 = \frac{1}{2}\times\frac{1}{2}\left[(\frac{1}{0.1}-15)^2 + (\frac{2}{0.1}-15)^2\right] = 25$

$E(\hat{\tau}) = \sum \hat{\tau} P = (15\times 0.02 + \cdots) = 10$ $V(\hat{\tau}) = \sum (\hat{\tau} - E(\hat{\tau}))^2 P = (15-10)^2 \times 0.02 + \cdots = 6.250$.

⇒ **without replacement** π_i: P(the ith element in the population, u_i is selected in the sample)

$\frac{\pi_i}{n}$ = average prob. the u_i is selected across the n draws that will ouur in a sample

Ex: $\{1,2,3\}$. $\tau = 6$. $n = 2$. then $\pi_1 = \pi_2 = \pi_3 = \frac{2}{3}$. $\hat{\tau} = \frac{1}{n}\sum \frac{y_i}{\pi_i}$ ($\pi_i = \sum_{j\neq i} P_{\{i,j\}}$)

Sample (1.2). prob $= \frac{1}{3}$. $\hat{\tau} = \frac{1}{2}\cdot\frac{1}{3} + \frac{2}{3} = 4.5$ $E(\hat{\tau}) = 4.5\times\frac{1}{3} + 6\times\frac{1}{3} + 7.5\times\frac{1}{3} = 6$. $V(\hat{\tau}) = 4.5^2\times\frac{1}{3} + 6^2\times\frac{1}{3} + 7.5^2\times\frac{1}{3} - 6^2 = 1.5$ $(E(\hat{\tau}^2) - E(\hat{\tau})^2)$

NOTE 2 (2) Supp_Categorical

Test Statistic for Goodness-of-Fit Tests: H_0: the distribution is the same today as it was in 2000.

$\chi^2_{(\alpha, K-1)} = \sum \frac{(O_i - E_i)^2}{E_i}$ for O_i: observed. E_i: expected ($i = 1 \dots K$) $\frac{(252-228)^2}{228} + \cdots$

Answer: since χ^2_0 is less/larger than $\chi^2_{0.05}$. we (fail to) reject H_0.

There's (in)sufficient evidence to conclude that the distribution for ... at the $\alpha = 0.05$ level of significance

Test for Independence (chi-square test) H_0: the variables are independent for $r=2$. $c=3$ degree: $(2-1)(3-1) = 2$.

expected frequency $= \frac{(\text{row total})(\text{column total})}{\text{table total}}$ $\chi^2_0 = \frac{(82-63.5808)^2}{63.5808} + \cdots \approx 30.82$

Test for Homogeneity of Proportions H_0: $p_1 = p_2 = p_3$ $\chi^2_0 = \frac{(418-475.554)^2}{475.554} + \cdots$

Number of Years	Observed Counts	Expected Counts
<1	252	228
1-3	255	239
3-4	162	176
≥5	331	357

	Money	Health	Love	
Men	82 (63.5808)	446 (507.048)	355 (312.2208)	883
Women	46 (64.2912)	574 (512.9088)	273 (315.7728)	893
	128	1020	608	

(128×883 + 1776)

	1992	2002	2008	
Yes	418 (475.554)	479 (475.554)	525 (470.892)	1422
No	602 (544.446)	541 (544.446)	485 (539.108)	1628
	1020	1020	1010	

1020 × 1422 + 3050 = 475.554

NOTE 3. Simple Random Sampling

⇒ without replacement. possible samples: $\binom{N}{n} = C^N_n$. prob of select each is $\frac{1}{\binom{N}{n}}$

$P(y_i = u_3) = \frac{1}{N}$. $P(y_i = u_3, y_j = u_t) = \frac{1}{n(n-1)}$

Estimation of a population mean: Population $\{u_1, u_2, \dots, u_N\}$, mean $\mu = \frac{u_1 + \cdots + u_N}{N} = \frac{1}{N}\sum u_i$, variance $\sigma^2 = \frac{1}{N}\sum (u_i - \mu)^2$

Sample $\{y_1, \dots, y_n\}$. mean $\bar{y} = \frac{1}{n}\sum y_i$ $s^2 = \frac{1}{n-1}\sum (y_i - \bar{y})^2$

$E(y_i)=\sum_1^N \frac{1}{N}u_i=u$. $Var(y_i)=\sum_1^N \frac{1}{N}(u_i-u)^2=\sigma^2$. $Cov(y_i,y_j)=E(y_iy_j)-u^2=-\frac{\sigma^2}{N-1}$ ① negative. ②. N↑ then var→0.

$E(\bar{y})=u$. $Var(\bar{y})=\frac{\sigma^2}{n}\frac{N-n}{N-1}$ ①. 不放回会小些. ②. $n\to\infty$, $Var(\bar{y})\to 0$. ③ $N\to\infty$, $Var(\bar{y})\to\frac{\sigma^2}{n}$.

Ex. precision. $\frac{n}{N}\frac{N-n}{N-1}\sigma^2$, like An SRS of size 400 from population 4000. is $\frac{1}{400}\frac{4000-400}{4000-1}\sigma^2=0.00225\sigma^2$. (f called finite population correction) ($z_{\frac{\alpha}{2}}=1.96$) S找题目已知量

(Estimation of population variance). $E(s^2)=\frac{N}{N-1}\sigma^2$. unbiased estimator for $\sigma^2=\frac{N-1}{N}s^2$, for $Var(\bar{y})=(1-f)\frac{s^2}{n}$ $f=\frac{n}{N}$

if n, N, N-n suggiciently large. $\frac{(\bar{y}-u)}{\sqrt{\widehat{var}(\bar{y})}}\sim N(0,1)$ and 100(1-α)% C.I. for u is $B=z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}(\bar{y})}=z_{\frac{\alpha}{2}}\sqrt{(1-f)\frac{s^2}{n}}$

• to find minimum required sample size (n). $D=\frac{B^2}{z^2_{\frac{\alpha}{2}}}=\frac{\sigma^2}{n}\frac{N-n}{N-1}\Rightarrow n\approx\frac{N\sigma^2}{(N-1)D+\sigma^2}$ (B会给. $\sigma=\frac{range}{4}$) if pop is large $\sigma^2=E(X^2)-E(X)^2$

Estimation of a population total: Population $\{u_1,\dots,u_N\}$. total $\tau=u_1+\dots+u_N=Nu$.

Sample: $\{y_1,\dots,y_n\}$. Estimator of τ: $\hat{\tau}=N\bar{y}$. $E(\hat{\tau})=E(N\bar{y})=NE(y)=Nu=\tau$ $Var(\hat{\tau})=Var(N\bar{y})=N^2\frac{\sigma^2}{n}\frac{N-n}{N-1}$

$\widehat{Var}(\hat{\tau})=N^2(1-f)\frac{s^2}{n}$ 100(1-α)% C.I. for τ is $\hat{\tau}\pm z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{\tau})}=N(\bar{y}\pm z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\sqrt{1-f})$ (if σ known. $B=N\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$)

• minimum required sample size (n). $D=\frac{B^2}{N^2z^2_{\frac{\alpha}{2}}}=\frac{\sigma^2}{n}(\frac{N-n}{N-1})$ also $n\approx\frac{N\sigma^2}{(N-1)D+\sigma^2}$ (n等向上取整) compared to mean.

Estimation of a population proportion:

for population $\{u_1,\dots,u_N\}$. $u_i=1$ if ith element has a specified characteristic 0 if not. ($u_i^2=u_i$)

let p be the population proportion: $p=\frac{\sum u_i}{N}$, population mean $u=\frac{1}{N}\sum_1^N u_i=p$. $\sigma^2=\frac{1}{N}\sum_1^N(u_i-u)^2=pq$

Sample $\{y_1,\dots,y_n\}$. proportion: $\hat{p}=\frac{\sum y_i}{n}=\bar{y}$ Estimator of $p=\hat{p}$, $E(\hat{p})=p$. $S^2=\frac{1}{n-1}\sum(y_i-\bar{y})^2=\frac{n}{n-1}\hat{p}\hat{q}$

$E(s^2)=\frac{N}{N-1}pq$. $Var(\hat{p})=\frac{pq}{n}\frac{N-n}{N-1}$, $\widehat{Var}(\hat{p})=(1-f)\frac{\hat{p}\hat{q}}{n-1}$, 100(1-α)% C.I. for p is $\hat{p}\pm z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{p})}=\hat{p}\pm z_{\frac{\alpha}{2}}\frac{\hat{p}\hat{q}}{n-1}\sqrt{1-f}$

Minimum required sample size n. $D=\frac{B^2}{z^2_{\frac{\alpha}{2}}}$, $n\approx\frac{Npq}{(N-1)D+pq}$ (p=0.5 取 0.5. or 求 max)

Comparing estimates: $E(\bar{x}-\bar{y})=u_x-u_y$, $Var(\bar{x}-\bar{y})=Var(\bar{x})+Var(\bar{y})-2Cov(\bar{x},\bar{y})$ (if samples are indep. $Cov(\bar{x},\bar{y})=0$)

Ex. difference in mean in 95% CI is $(\bar{y_1}-\bar{y_2})\pm z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$ ($S^2=\frac{N}{N-1}\sigma^2=\frac{N}{N-1}\cdot\frac{1}{N}\sum(u_i-\bar{u})^2=\frac{1}{N-1}\sum(u_i-\bar{u})^2$) to see if CI contains 0.

• 100(1-α)% C.I for p_1-p_2 is $(\hat{p_1}-\hat{p_2})\pm z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{p_1}-\hat{p_2})}=(\hat{p_1}-\hat{p_2})\pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p_1}\hat{q_1}}{n}+\frac{\hat{p_2}\hat{q_2}}{n}+\frac{2\hat{p_1}\hat{p_2}}{n}}$ (if ignore, then underestimate)

Ex. multinomial sampling. X = num of responses in the first category (p_1). then $E(x)=np_1$, $E(y)=np_2$, $E(z)=np_3$

$Var(x)=np_1q_1$. $Var(y)=np_2q_2$. $E(xy)=n(n-1)p_1p_2$. $Cov(x,y)=E(xy)-n^2p_1p_2=-np_1p_2$.

$E(\hat{p_i})=p_i$, $Var(\hat{p_i})=\frac{p_iq_i}{n}$, $Cov(\hat{p_1},\hat{p_2})=\frac{-p_1p_2}{n}$, $Var(\hat{p_1}-\hat{p_2})=\frac{p_1q_1}{n}+\frac{p_2q_2}{n}+\frac{2p_1p_2}{n}$, $\widehat{Var}(\hat{p_1}-\hat{p_2})=\frac{\hat{p_1}\hat{q_1}}{n}+\frac{\hat{p_2}\hat{q_2}}{n}+\frac{2\hat{p_1}\hat{p_2}}{n}$

Population (number in) Stratum (strata) — like male/female — heterogeneous between

NOTE 6. Stratified Random Sampling

separating the population element into nonoverlapping groups stratum

population total = number of sampling units × population mean ($\tau_i=N_iu_i$), $\tau=\sum_1^L N_iu_i$. $u=\frac{1}{N}\sum_1^L N_iu_i$ 代入算CI

Estimation of population mean: $\hat{u}=\bar{y}_{st}=\frac{1}{N}\sum_1^L N_i\bar{y}_i$ $E(\bar{y}_{st})=E[\frac{1}{N}\sum_1^L N_i\bar{y}_i]=\frac{1}{N}\tau=u$. $\widehat{Var}(\bar{y}_{st})=\frac{1}{N^2}\sum_1^L N_i^2\frac{s_i^2}{n_i}(1-\frac{n_i}{N_i})$

Estimation of total: $Var(N\bar{y}_{st})=Var(\hat{\tau}_{st})=\sum_1^L N_i^2\frac{S_i^2}{n_i}(1-\frac{n_i}{N_i})$ (for $S_i^2=\frac{N_i}{N_i-1}\sigma_i^2$) $\hat{\tau}_{st}\pm z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{\tau}_{st})}$ ($s_i^2=\frac{N_i}{N_i-1}\sigma_i^2$)

Sample size determination (means) $D=\frac{B^2}{z^2_{\frac{\alpha}{2}}}$ and $n=(\sum_1^L N_i^2\sigma_i^2/a_i)/(N^2D+\sum_1^L N_i\sigma_i^2)$ ($n_i=na_i$) 求出n以后再通过a_i比例分配

Sample size determination (total) $D=B^2/(N^2z^2_{\frac{\alpha}{2}})$ and $n=(\sum_1^L N_i^2\sigma_i^2/a_i)/(N^2D+\sum_1^L N_i\sigma_i^2)$ (题目会告诉你是estimate total?)

Optimal allocation of the sample: to gain the most information for the least cost. $C=c_0+\sum_1^L c_in_i$

cost = overhead cost (operating expenses, fixed) + variable cost (depending on number of observed units)

I. Min $Var(\bar{y}_{st})$, for a given total cost $C=c_0+\sum_1^L c_in_i$ (min var for fix cost)

$$n_i=\left(\frac{C-c_0}{\sum_{i=1}^L N_i\sigma_i\sqrt{c_i}}\right)\frac{N_i\sigma_i}{\sqrt{c_i}} \qquad n=\left(\frac{C-c_0}{\sum_{i=1}^L N_i\sigma_i\sqrt{c_i}}\right)\sum_{i=1}^L\frac{N_i\sigma_i}{\sqrt{c_i}}$$

$\frac{n_i}{n}=a_i=(N_i\sigma_i/\sqrt{c_i})/(\sum N_i\sigma_i/\sqrt{c_i})$ $B=z_{\frac{\alpha}{2}}\sqrt{V}$ (给B就是给V)

II. Min total cost C, given $Var(\bar{y}_{st})=V$ 就是D.

$$n_i=\left(\frac{\sum_{i=1}^L N_i\sigma_i\sqrt{c_i}}{N^2V+\sum_{i=1}^L N_i\sigma_i^2}\right)\frac{N_i\sigma_i}{\sqrt{c_i}} \qquad n=\left(\frac{\sum_{i=1}^L N_i\sigma_i\sqrt{c_i}}{N^2V+\sum_{i=1}^L N_i\sigma_i^2}\right)\sum_{i=1}^L\frac{N_i\sigma_i}{\sqrt{c_i}}$$

II. (special case) same cost: $c_1=c_2=\dots=c_L$. then it's **Neyman allocation** $a_i=\frac{n_i}{n}=\frac{N_i\sigma_i}{\sum N_i\sigma_i}$ $n_i=\frac{(N_i\sigma_i)\sum N_i\sigma_i}{N^2D+\sum N_i\sigma_i^2}$

Proportional allocation (ignore both cost and variability), Allocation scheme $\frac{n_i}{n}=\frac{N_i}{N}=a_i$

$n=(\sum N_i\sigma_i^2)/(ND+\sum N_i\sigma_i^2/N)=(\sum N_i^2\sigma_i^2/a_i)/(N^2D+\sum N_i\sigma_i^2)$, where $D=B^2/z^2_{\frac{\alpha}{2}}$, then $V(\hat{\tau}_{st})=\sum N_i^2\frac{\sigma_i^2}{n_i}\frac{N_i-n_i}{N_i-1}$

Estimation of population proportion: $S_i^2=\frac{n_i}{n_i-1}\hat{p}_i\hat{q}_i$, unbiased estimator of p is $\hat{p}_{st}=\frac{1}{N}\sum N_i\hat{p}_i$

$Var(\hat{p}_{st})=\frac{1}{N^2}\sum N_i^2\frac{\hat{p}_i\hat{q}_i}{n_i}(\frac{N_i-n_i}{N_i-1})$, $\widehat{Var}(\hat{p}_{st})=\frac{1}{N^2}\sum N_i^2\frac{\hat{p}_i\hat{q}_i}{n_i-1}(1-f_i)$ where $f_i=\frac{n_i}{N_i}$, 100(1-α)% C.I of p is $\hat{p}_{st}\pm z_{\frac{\alpha}{2}}\sqrt{\hat{V}(\hat{p}_{st})}$

→ sample size: $n=(\sum N_ip_iq_i/a_i)/(N^2D+\sum N_ip_iq_i)$ where $D=\frac{B^2}{z^2_{\frac{\alpha}{2}}}$ (estimate proportion with error B and level 1-α)

Optimal allocation: $n_i=n\left(\frac{N_i\sqrt{p_iq_i}/\sqrt{c_i}}{\sum_{i=1}^L N_i\sqrt{p_iq_i}/\sqrt{c_i}}\right)$ Neyman allocation: $n_i=n\left(\frac{N_i\sqrt{p_iq_i}}{\sum_{i=1}^L N_i\sqrt{p_iq_i}}\right)$ Proportional allocation: $n_i=n\left(\frac{N_i}{N}\right)$ $\sqrt{p_iq_i}=\sigma_i$.

(要先找 Allocation (a_i) 才能回去算 n.)

• if we want to find CI for y_1-y_2: $\bar{y}_1-\bar{y}_2\pm z_{\frac{\alpha}{2}}\sqrt{\frac{N_1-n_1}{N_1}\frac{s_1^2}{n_1}+\frac{N_2-n_2}{N_2}\frac{s_2^2}{n_2}}$

① $Var_{Neyman}(\bar{y}_{st})\le Var_{opt}(\bar{y}_{st})$ ② $Var_{Neyman}(\bar{y}_{st})\le Var_{prop}(\bar{y}_{st})$

③ if N_i, $i=1\dots L$ are large, then $Var_{Neyman}(\bar{y}_{st})\le Var_{prop}(\bar{y}_{st})\le Var_{SRS}(\bar{y}_{st})$

④ $\frac{Var_{SRS}(\bar{y})}{Var_{prop}(\bar{y}_{st})}=\frac{\sum_i N_i\sigma_i^2+\sum_i N_i(u_i-u)^2}{\sum_i N_i\sigma_i^2}=1+\frac{\sum N_i(u_i-u)^2}{\sum N_i\sigma_i^2}$

NOTE 7. Ratio, Regression and Difference Estimation

Population N pairs $(u_1, v_1) \dots (u_N, v_N)$ when $\bar{u}_x = \frac{1}{N}\sum_1^N u_i$. $\tau_x = \sum_1^N u_i$. $\sigma_x^2 = \frac{1}{N}\sum_1^N (u_i - \bar{u}_x)^2$. 把 u_i 换成 v_i 就是 $\bar{u}_y$ τ_y σ_y^2

Sample n pairs $(x_1, y_1) \dots (x_n, y_n)$, x_i is called auxiliary variable / subsidiary variable. (to be used for forcasting)

Estimation of population ratio: $R = \frac{\sum_1^N v_i}{\sum_1^N u_i} = \frac{\tau_y}{\tau_x} = \frac{\bar{u}_y}{\bar{u}_x}$ $\quad r = \frac{\bar{y}}{\bar{x}}$. $E(r) - R \approx \frac{N-n}{N-1}\left(\frac{1}{n\bar{u}_x^2}\right)(R\sigma_x^2 - \rho\sigma_x\sigma_y)$ (E(r)=R 如果不明说)

$\mathrm{Var}(r) \approx \frac{1}{\bar{u}_x^2}\frac{\sigma_{y-Rx}^2}{n}\left(\frac{N-n}{N-1}\right)$ where $\sigma_{y-Rx}^2 = \frac{1}{N}\sum_1^N (v_i - Ru_i)^2 = \sigma_y^2 + R^2\sigma_x^2 - 2R\rho\sigma_x\sigma_y$

$\widehat{\mathrm{Var}}(r) = \frac{1}{\bar{u}_x^2}\frac{s_{y-Rx}^2}{n}(1-f)$ if $\bar{u}_x$ is known, $\widehat{\mathrm{Var}}(r) = \frac{1}{\bar{x}^2}\frac{s_{y-rx}^2}{n}(1-f)$ if $\bar{u}_x$ is unknown

then $100(1-\alpha)\%$ C.I. of R is $r \pm z_{\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}(r)}$, $\mathrm{Var}(r) = \frac{1}{\bar{u}_x^2}\frac{\sigma_{y-Rx}^2}{n}\left(\frac{N-n}{N-1}\right)$, where $s_{y-Rx}^2 = s_y^2 + r^2 s_x^2 - 2r\hat{\rho}s_x s_y = \frac{1}{n-1}\sum_1^n (y_i - rx_i)^2$

用 ratio 的条件：linear and pass zero. the required sample size is $n = \frac{N\sigma_{y-Rx}^2}{ND + \sigma_{y-Rx}^2}$ where $D = \frac{B^2\bar{u}_x^2}{z_{\frac{\alpha}{2}}^2}$

Estimation of population total: $\tau_y = R\tau_x$, $\hat{\tau}_y = r\tau_x$. $\mathrm{Var}(\hat{\tau}_y) = (\tau_x)^2\mathrm{Var}(r)$, $\widehat{\mathrm{Var}}(\hat{\tau}_y) = (\tau_x)^2\widehat{\mathrm{Var}}(r) = \frac{(\tau_x)^2}{N^2\bar{u}_x^2}\frac{s_{y-rx}^2}{n}(1-f)$

$100(1-\alpha)\%$ C.I. for τ_y is $\hat{\tau}_y \pm z_{\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}(\hat{\tau}_y)}$. sample size: $n = \frac{N\sigma_{y-Rx}^2}{ND + \sigma_{y-Rx}^2}$ where $D = \frac{B^2}{N^2 z_{\frac{\alpha}{2}}^2} = \left(\frac{N-n}{N-1}\right)\frac{1}{n}(\sigma_y^2 - R^2\sigma_x^2 - 2R\sigma_x\sigma_y)$

Estimation of population mean: $\bar{u}_y = R\bar{u}_x$, $\hat{\bar{u}}_y = r\bar{u}_x$, $\mathrm{Var}(\hat{\bar{u}}_y) = (\bar{u}_x)^2\mathrm{Var}(r) = \left(\frac{N-n}{N-1}\right)\frac{\sigma_{y-Rx}^2}{n}$, $\widehat{\mathrm{Var}}(\hat{\bar{u}}_y) = (\bar{u}_x)^2\widehat{\mathrm{Var}}(r) = \frac{s_{y-Rx}^2}{n}(1-f)$

$100(1-\alpha)\%$ C.I. for $\bar{u}_y$ is $\hat{\bar{u}}_y \pm z_{\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}(\hat{\bar{u}}_y)}$, sample size $n = \frac{N\sigma_{y-Rx}^2}{ND + \sigma_{y-Rx}^2}$ where $D = \frac{B^2}{z_{\frac{\alpha}{2}}^2}$

Regression Estimation (x_i, y_i) be $Y = \alpha + \beta x$. the least squares for β: $b = \frac{s_{xy}}{s_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$, α: $a = \bar{y} - b\bar{x}$. best fitted line $\hat{y} = a + bx = \bar{y} + b(x - \bar{x})$

Regression estimator of population mean $\bar{u}_y$ $\hat{\bar{u}}_{yL} = \bar{y} + b(\bar{u}_x - \bar{x})$, $\widehat{\mathrm{Var}}(\hat{\bar{u}}_{yL}) = \frac{(1-f)MSE}{n}$, $MSE = \frac{\sum(u_i - \hat{y})^2}{n-2}$, $\mathrm{Var}(\hat{\bar{u}}_{yL}) = \left(\frac{N-n}{N-1}\right)\frac{1}{n}\sigma_E^2$

$\bar{u}_y = \bar{y} + \beta(\bar{u}_x - \bar{x})$

$\sigma_E^2 = s_y^2 - b^2 s_x^2$

Difference Estimation $\bar{u}_y = \bar{u}_x + (\bar{u}_y - \bar{u}_x)$ estimate difference the difference $D = u_y - u_x$

$\hat{\bar{u}}_{yD} = \bar{u}_x + (\bar{y} - \bar{x}) = \bar{u}_x + \bar{d}$, $\bar{d} = \bar{y} - \bar{x}$, $d_i = y_i - x_i$, $\mathrm{Var}(\hat{\bar{u}}_{yD}) = \frac{(1-f)s_d^2}{n} = \left(\frac{N-n}{N-1}\right)\frac{\sigma_d^2}{n}$

and $s_d^2 = \frac{1}{n-1}\sum_1^n (d_i - \bar{d})^2 = \frac{1}{n-1}\sum_1^n [(y_i - \bar{y}) - (x_i - \bar{x})]^2 = s_y^2 + s_x^2 - 2\hat{\rho}s_x s_y$

这不就是不明说的

NOTE 8. Systematic Sampling

1-in-k systematic sampling (抽第 1, 4, 7, 10...个)，只抽一个 group. 要保证 Var between group small

Estimation of population mean: Sample $\{y_1, y_2, \dots, y_n\}$. $\bar{y}_{sy} = \frac{1}{n}\sum_1^n y_i$, sample variance $s_{sy}^2 = \frac{1}{n-1}\sum_1^n (y_i - \bar{y}_{sy})^2$, $\mathrm{Var}(\bar{y}_{sy}) = \frac{1}{k}\sum_1^k (\bar{u}_i - u)^2$

$\widehat{\mathrm{Var}}(\bar{y}_{sy}) = (1-f)\frac{s_{sy}^2}{n}$. $100(1-\alpha)\%$ C.I for u: $\bar{y}_{sy} \pm z_{\frac{\alpha}{2}}\sqrt{(1-f)\frac{s_{sy}^2}{n}}$

$SST = \sum_{i=1}^k\sum_{j=1}^n (u_{ij} - u)^2$, $SSW = \sum_{i=1}^k\sum_{j=1}^n (u_{ij} - \bar{u}_i)^2$ $SSB = \sum_{i=1}^k n(\bar{u}_i - u)^2$, $SST = SSW + SSB$, $MST = \frac{SST}{N-1}$, $MSW = \frac{SSW}{k(n-1)} = \frac{SSW}{N-k}$, $MSB = \frac{SSB}{k-1}$

then $\mathrm{Var}(\bar{y}_{sy}) = \frac{\sigma^2}{n}(1 + (n-1)\rho_w)$ where $\rho_w = \frac{E(y_{ij} - u)(y_{il} - u)}{E(y_{ij} - u)^2} = \frac{\sum_{i=1}^k\sum_{j=1}^n\sum_{l=1}^n (u_{ij} - u)(u_{il} - u)}{nk(n-1)\cdot\sigma^2}$

(Compare) if $\mathrm{Var}(\bar{y}_{sy}) < \mathrm{Var}(\bar{y}_{SRS}) \Leftrightarrow \sigma^2\frac{N-k}{N}MSW < \frac{\sigma^2}{n}\frac{N-n}{N-1} \Leftrightarrow MSW > \sigma^2\frac{N}{N-1}$

Estimation of population total $\hat{\tau} = N\bar{y}_{sy} = \frac{N}{n}\sum_1^n y_i$, $\widehat{\mathrm{Var}}(\hat{\tau}_{sy}) = N^2(1-f)\frac{s_{sy}^2}{n}$, $100(1-\alpha)\%$ C.I. for τ is $\hat{\tau}_{sy} \pm z_{\frac{\alpha}{2}}\sqrt{N^2(1-f)\frac{s_{sy}^2}{n}}$

Estimation of population proportion $\hat{p}_{sy} = \bar{y}_{sy} = \frac{1}{n}\sum_1^n y_i$, $\widehat{\mathrm{Var}}(\hat{p}_{sy}) = (1-f)\frac{\hat{p}_{sy}\hat{q}_{sy}}{n-1}$, $100(1-\alpha)\%$ CI for p: $\hat{p}_{sy} \pm z_{\frac{\alpha}{2}}\sqrt{(1-f)\frac{\hat{p}_{sy}\hat{q}_{sy}}{n-1}}$

NOTE 9&10 Cluster Sampling externally homogeneous but internally heterogeneous 每个cluster都大致能代表总体性质

(Clusters of equal sizes) Primary sampling units (PSU): Sampling units chosen in the first stage of selection

Secondary sampling units (SSU) Sampling units within the PSUs that are chosen in the second stage of selection.

One-stage cluster sampling: 选一个cluster 而所有都要 Two-stage cluster sampling: 在选定的cluster里SRS之类

Let $M_1 = \cdots = M_N = M$, total population size $= NM$, Total sample size $= nM$

Estimation of population mean: select n clusters in N by SRS.

$\hat{u} = \bar{y}_{cl} = \frac{1}{n}\sum_1^n \bar{y}_i$ (average of cluster mean), $\hat{u}$ unbiased of u

$Var(\bar{y}_{cl}) = \frac{N-n}{Nn} S_b^2$ where $S_b^2 = \frac{1}{N-1}\sum_1^N (u_i - u)^2$ mean sum of square between cluster means

$\widehat{Var}(\bar{y}_{cl}) = \frac{N-n}{Nn} s_b^2$ where $s_b^2 = \frac{1}{n-1}\sum_1^n (\bar{y}_i - \bar{y}_{cl})^2$

(compare to SRS) define $S^2 = \frac{1}{NM-1}\sum_{i=1}^N \sum_{j=1}^M (u_{ij} - u)^2$ and $\bar{y}$ sample mean of SRS

relative efficiency $RE = \frac{Var(\bar{y})}{Var(\bar{y}_{cl})} = \frac{1}{NM-1}\left[\frac{N(M-1)}{M}\frac{S_w^2}{S_b^2} + (N-1)\right] = \frac{S^2}{MS_b^2}$ $S_w^2 = \frac{1}{N}\sum^N S_i^2$ within clusters

$\hat{RE} = \frac{M-1}{M}\frac{s_w^2}{Ms_b^2} + \frac{1}{M}$

cluster sampling is good when S_w is large and S_b^2 is small

Estimation of population proportion p: $\hat{p}_{cl} = \frac{1}{n}\sum_1^n p_i$ unbiased of p

$Var(\hat{p}_{cl}) = \frac{N-n}{Nn}\frac{1}{N-1}(NPQ - \sum_1^N p_i q_i)$ $\widehat{Var}(\hat{p}_{cl}) = \frac{N-n}{Nn}\frac{1}{n-1}\sum_1^n (p_i - \hat{p}_{cl})^2$

Notation

Population quantities (refer to population data on p.8)

- N = number of clusters (psus) in the population
- M_i = number of elements (ssus) in cluster i
- $M_0 = \sum_{i=1}^N M_i$ = total number of elements (ssus) in the population — population size
- $\bar{M} = \frac{M_0}{N}$ = Average cluster size for the population
- $\tau_i = \sum_{j=1}^{M_i} u_{ij}$ = total in cluster i
- $\tau = \sum_{i=1}^N \tau_i = \sum_{i=1}^N \sum_{j=1}^{M_i} u_{ij}$ = population total
- $\mu = \frac{\tau}{M_0}$ = population mean
- $\mu_i = \frac{\tau_i}{M_i}$ = population mean in cluster i
- $\sigma^2 = \frac{1}{M_0 - 1}\sum_{i=1}^N \sum_{j=1}^{M_i}(u_{ij} - \mu)^2$ = population variance — consistant
- $\sigma_i^2 = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}(u_{ij} - \mu_i)^2$ = population variance in cluster i

Sample quantities

- n = number of clusters (psus) in the sample
- m_i = number of elements (ssus) in cluster i in the sample
- $\sum_{i=1}^n m_i$ = total number of elements (ssus) in the sample
- $\bar{m} = \frac{1}{n}\sum_{i=1}^n m_i$ = Average cluster size for the sample
- y_i = total of all observations in cluster i
- $\bar{y} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n m_i}$ = sample mean

(clusters of unequal sizes) 好处是：M_0不知道时，抽n个cluster去估效果好好.

Estimation of population mean:

(unbiased) $\hat{u} = \bar{y}_u = \frac{1}{n\bar{M}}\sum_1^n M_i \bar{y}_i$ 但$\bar{M}$不知道就无法估效.

$E(\bar{y}_u) = u$, $Var(\bar{y}_u) = \frac{N-n}{Nn} S_b^{2*}$ where $S_b^{2*} = \frac{1}{n-1}\sum_1^n \left(\frac{M_i}{\bar{M}}\bar{y}_i - \bar{y}_u\right)^2$

(biased) $\hat{u} = \bar{y} = \frac{\sum y_i}{\sum m_i}$, $E(\bar{y}) \approx u$. $\widehat{Var}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{s_r^2}{n\bar{M}^2}$ where $S_r^2 = \frac{\sum_1^n (y_i - \bar{y}m_i)^2}{n-1}$ ratio observation

Estimation of population total estimator #2 (when M_0 known) $\hat{\tau} = M_0\bar{y}$. $\widehat{Var}(\hat{\tau}) = N^2\left(1 - \frac{n}{N}\right)\frac{s_r^2}{n}$

#1 (M_0 unknown) $\hat{\tau} = N\bar{y}_\tau = N\left(\frac{1}{n}\sum_1^n y_i\right) = N\bar{M}\bar{y}_u$, $\widehat{Var}(\hat{\tau}) = \widehat{Var}(N\bar{M}\bar{y}_u) = N^2\bar{M}^2\frac{N-n}{Nn}\frac{1}{n-1}\sum_1^n\left(\frac{M_i}{\bar{M}}\bar{y}_i - \bar{y}_u\right)^2 = N^2\frac{N-n}{Nn}s_\tau^2$ where $s_\tau^2 = \frac{1}{n-1}\sum_1^n (y_i - \bar{y}_\tau)^2$

$\bar{y}_\tau = \bar{M}\bar{y}_u = \frac{1}{n}\sum_1^n y_i$, $\widehat{Var}(\bar{y}_u) = \frac{N-n}{Nn}\frac{1}{n-1}\sum_1^n\left(\frac{M_i}{\bar{M}}\bar{y}_i - \bar{y}_u\right)^2$, $B = 1.96\sqrt{\hat{V}(\hat{\tau})} = 1.96\sqrt{\frac{V(N-n)}{n}}s_\tau$

Estimation of population proportion p $\hat{p} = \frac{\sum a_i}{\sum M_i}$ (unbiased of p), $\widehat{Var}(\hat{p}) = \left(1 - \frac{n}{N}\right)\frac{s_p^2}{n\bar{M}^2}$ where $S_r^2 = \frac{\sum_1^n (a_i - \hat{p}m_i)^2}{n-1}$