STOCKHOLMS UNIVERSITET HT2019

Statistiska institutionen
Dan Hedlin

Sample surveys, ST306G

Examination 2019-12-03, 10.00 – 15.00

Approved aids:

1. Pocket calculator
2. Language dictionary

Separate pages with notes are not allowed.

The exam comprises 9 items, numbered 1 to 9. The maximum number of points is 50. 25 points will give you at least grade E. To obtain the maximum number of points full and clear motivations are required unless otherwise stated. You may write in English or Swedish. **There are some pages at the end of the exam with formulae that you may wish to use.**

In each of the five questions below one of the items a, b, c, d or e is incorrect. Which one? For each of the questions 1-5, answer with only one letter, a-e. Motivation is not required. Maximum 10 points.

1.

a) Suppose a population is divided into two strata. If there is a reason to believe that $S_1^2 < S_2^2$ (stratum 1 and 2; see formulas at the end of the exam), then it is often wise to take a smaller sample in stratum 1 than in stratum 2.

b) Suppose a population consists of one stratum with N_1=100 units and another stratum with N_2=50 units. A simple random sample of size 10 is taken from stratum 1 and simple random sample of size 5 is taken from stratum 2. Then the sample is proportionally allocated.

c) Suppose a population consists of one stratum with N_1=100 units and another stratum with N_2=50 units. A simple random sample of size 10 is taken from stratum 1 and a systematic sample of size 5 is taken from stratum 2. Then the sample is proportionally allocated.

d) Suppose a population consists of one stratum with 100 units and another stratum with 50 units. A simple random sample of size 10 is taken from the first stratum and another simple random sample of size 5 is taken from the other stratum. The population variance in the first stratum is $S_1^2 = 100$, and in the second stratum it is $S_2^2 = 25$. Then the sample is both proportionally and optimally allocated.

e) A survey may have different questionnaires in different subsets of the population. Then that would be a reason to stratify by those subsets.

2.

a) There may be only one domain in a survey, and that domain may be a subset of the target population.

b) In a survey on unemployment, one domain may be men, and another domain may be a subset of the first domain, for example, men 20-25 years of age.

c) Domains in a survey may be overlapping. For example, one domain may be women 20-25 years of age, another one women 22-27 years of age.

d) It is not incorrect to decide on what domains to estimate for after the data have been collected.

e) What domains you are interested in is not relevant when the optimal sample size is decided on.

3.

a) One type of nonsampling error is called 'goodness-of-fit of an estimator'.

b) One type of nonsampling error is called 'undercoverage'.

c) It is more difficult to estimate or assess the size of nonsampling errors than the size of sampling errors.

d) If sample is large, the nonsampling errors in social surveys are in general more serious than the sampling errors.

e) The nonresponse rate is in general a weak predictor of nonsampling errors.

4.

a) The sampling design in the following example is one-stage cluster sampling. The frame consists of all schools in Sweden. First, a simple random sample of schools is selected, then all teachers in the school are asked to fill out a questionnaire.

b) There is reason to believe that the sampling design described in a) is in general less efficient (that is, the variance is larger) than a simple random sample of the same number of teachers.

c) The sampling design in the following example is one-stage cluster sampling. The frame consists of all schools in Sweden. First, a simple random sample of schools is selected, then all year 9 pupils are asked to fill out a questionnaire. However, in one school the questionnaires were not handed out, because the teacher failed to read his mail.

d) The sampling design in the following example is one-stage cluster sampling where only one cluster is sampled. The seats in a large concert hall are numbered 1-900. It is sold out, every single seat is occupied. Every 15th visitor is included in a sample, starting with a random selection of the visitors in seats 1-15, then proceeding with every 15th seat until seat 900.

e) The sampling design in the following example is one-stage cluster sampling. To count blue whales a ship is equipped with a sonar that has the capacity to register a blue whale located under the ship. The sea where blue whales are to be counted is clearly marked on a map. A rectangular grid of points is laid out over the map, numbered 1-500. A simple random sample of 50 of the 500 points is selected. The ship will count the number of blue whales along a line for 10 kilometres in southern direction, starting from each selected point.

5.

a) One disadvantage with register-based statistics (relative to a survey) is that the register may not contain exactly the variables you need.
b) Nonsampling errors in register-based statistics include under- and overcoverage.
c) In business surveys it is common to draw a random sample (e.g. stratified simple random sample) among medium-sized and small businesses and make a census among large businesses.
d) Register-based statistics may suffer from nonresponse, although we usually say 'missing values' rather than 'nonresponse' when we talk about registers.
e) Cluster sampling and two-stage sampling are often used in register-based statistics.

6.

The purpose of survey is to estimate the maintenance costs in an archive. There is a cost associated with each unit in the archive, which consists of a physical archive (one room) and a digital one. A unit in the digital archive may be, for example, a Word document. One common maintenance activity is to write or rewrite the description of the file in the digital archive. The frame consists of one list of all units in the digital archive and another one for the physical archive. There are 1000 records in each list.

A random sample is selected in the following way. First, a simple random sample of one record among the first ten records in the list of the physical archive. Then every tenth record is selected to be included in the sample until the end of the list. The sampling process is repeated in the same way for the list of the digital archive, including a simple random sample among the first ten units. Every unit in the sample is inspected manually to assess maintenance costs. There was no cost for 50% of physical units and no cost for 60% of the digital units. The costs for the sampled physical and digital units summed to 20 000 kronor and 500 kronor, respectively.

a) What is this sampling design called? Be as specific as possible.
b) Estimate the total costs.
c) Unbiased variance estimation is impossible in this case, because there is an issue associated with the second order inclusion probabilities. What issue is this?
d) The costs of another archive is going to be estimated. This archive has also 1000 physical and 1000 digital units, and the same sampling design with the same total sample size is going to be employed. However, this time the costs found in the previous survey is going to be used to improve on precision. Suggest a way to make use of the numbers given in the text above to (hopefully) increase precision in the coming survey. A good motivation is required for full points.

Maximum 8 points.

7.

In a survey on how many times a year people travel by air, a simple random sample of 1000 individuals was selected from the population register of a Swedish town. A paper questionnaire was mailed out. 600 responded. The size of the population eligible for the survey was 60 000. The target population could be identified in the frame.

Number of times an individual has travelled by airplane	Number of respondents	Number of women	Number of men
0	500	280	220
1	75	30	45
2+	25	10	15
Sum	600	320	280

a) What does MCAR, missing completely at random, and MAR, missing completely at random, mean? Be as specific as possible.

Assume MCAR in b)-e).

b) Using the data above, estimate the proportion of women who have travelled by airplane and give the standard error for the estimated proportion.
c) Is it fair to say that the data in the table above suggest that women travel by airplane more often than men do?
d) Estimate the total number of flights in the population. No variance estimation is required.
e) Use the fact that the population contains 60% women and 40% men and re-estimate the total number of flights in the population. No variance estimation is required.

Maximum 12 points.

8.

The aim of a business survey is to estimate the mean value of new orders. The team conducting the survey needs a sample size large enough for the confidence interval be $\pm 20\%$ of the population mean. That is, half the total width of the confidence interval should be 20% of the mean.

To assess the required sample size, calculations will be based on the frame variable number of employees, as is common in business surveys. Hence, the population parameters in each stratum given in the table below are based on number of employees in the businesses. N_h, t_h and S_h^2 denote the number of businesses, the total number of employees and the variance in stratum h. You can ignore the finite sampling correction. The population mean of number of employees is 1.4, $S^2 = 9.2$ and $\sum_{h=1}^{3} N_h S_h = 950$.

Note that you may be able to do c) even if you have not done a) or b).

Stratum	N_h	t_h	S_h^2
1	500	150	1
2	125	300	4
3	50	500	16
Sum	675	950	

a) The team is not aware of the method of stratification. They intend to draw a simple random sample from the population, and the estimator will be $\hat{t}_y = \frac{N}{n}\sum_{i\in s} y_i$. What sample size do they need to meet their requirement?
If you do not know to solve this, assume that their requirement is a confidence interval width of 1.4, that is, total length 1.4. (It will not give you the full points though, but at least something.)

b) You tell them that for this skewed population they should stratify by size. They opt for the three strata in the table above. What sample size is needed to make $1.96\sqrt{V(\hat{t}_{str}\ /\ N)}/\bar{y}_U = 0.2$, where $\bar{y}_U$ is the population mean of number of employees and $\hat{t}_{\mathrm{str}} = \sum_{h=1}^{H}\sum_{i\in S_h}\frac{N_h y_i}{n_h}$. Again, to make it convenient, ignore the finite population correction. Hint: substitute something in the formula for $V(\hat{t}_{str})$ with something from optimal allocation.

c) What are the best stratum sample sizes n_1, n_2 and n_3, given the total sample size from b)? If you have not been able to solve b), assume that n = 50.

Maximum 10 points.

9.

In a project farmers in one country have been encouraged to plant trees at their farm land to stop erosion. A list of the 1132 farms that took part in the project contains addresses, names of owners and gps coordinates. The average farm land of those farms is 10.3 hectares (ha). Although that number is old, we can treat it is as a true value. To estimate the total number of trees satellite pictures are made use of. A sample of 200 farms was selected with haphazard sampling that we can view as simple random sampling.

Use three different estimators to compute three estimates of the average number of trees in the farms that took part in the project. All estimators should be either exactly unbiased or approximately unbiased. Estimate also the variances of the three estimates. In the SAS output below, farm land area is denoted by **a** and number of trees by **b**. It is part of the exam to interpret the output.

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
a	200	9.42196	2.20366	1884	4.07442	14.83234
b	200	109.44393	33.11175	21889	14.41703	202.03331
c	200	0.01385	0.52217	2.77070	-1.30727	1.49859

Pearson Correlation Coefficients, N = 200

	a	b	c
a	1.00000	0.83264	0.13388
b	0.83264	1.00000	0.11879
c	0.13388	0.11879	1.00000

Regression Analysis for Dependent Variable b

Data Summary

Number of Observations	200
Sum of Weights	1132.0
Weighted Mean of b	109.44393
Weighted Sum of b	123890.5

Fit Statistics

R-Square	0.6933
Root MSE	18.3838
Denominator DF	199

Tests of Model Effects

Effect	Num DF	F Value	Pr > F
Model	1	496.39	<.0001
Intercept	1	2.45	0.1194
a	1	496.39	<.0001

Estimated Regression Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-8.4358716	5.39423733	-1.56	0.1194
a	12.5111787	0.56154573	22.28	<.0001

Maximum 10 points.

Formulae

Population

Population of size N: $U = \{1, \ldots, i, \ldots, N\}$

Sample, size n: $s = \{1, \ldots, i, \ldots, n\}$

Population total of study variable y: $t_y = \sum_{i \in U} y_i$

Population mean of study variable y: $\bar{y}_U = \frac{1}{N} \sum_{i \in U} y_i$

Population total of auxiliary variable x: $t_x = \sum_{i \in U} x_i$

Population variance: $S_y^2 = \frac{1}{N-1} \sum_{i \in U} (y_i - \bar{y}_U)^2$ (Lohr p. 32)

A **proportion** is a special case with $y_i = \begin{cases} 1 \text{ if unit } i \text{ has the relevant characteristic} \\ 0 \text{ otherwise} \end{cases}$ (compare Lohr p. 33).

For a proportion P the population variance $S^2 \approx P(1-P)$ (Lohr p. 38)

Formulas for SRS

Expansion estimator of t_y: $\hat{t}_y = \frac{N}{n} \sum_{i \in s} y_i$

Corresponding estimator of $\bar{y}_U$: $\frac{\hat{t}_y}{N} = \bar{y}_s$

$V(\hat{t}_y) = N^2 \left(1 - \frac{n}{N}\right) \frac{S_y^2}{n}$ (Lohr (2.16))

For an estimator of $V(\hat{t}_y)$, replace S_y^2 with the following estimator of S_y^2:

$s_y^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \bar{y}_s)^2$ (Lohr (2.10) and (2.17))

Ratio estimator of t_y : $\hat{t}_{rat} = t_x \frac{\hat{t}_y}{\hat{t}_x} = t_x \hat{B}$ (Lohr (4.2))

$\hat{V}(\hat{t}_{rat}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_e^2}{n}$, where $s_e^2 = \frac{1}{n-1} \sum_{i \in s} \left(y_i - \hat{B} x_i\right)^2 = s_y^2 + \hat{B}^2 s_x^2 - 2\hat{B} s_{xy}$,

$s_{xy} = \frac{1}{n-1}\sum_{i\in s}(y_i - \bar{y}_s)(x_i - \bar{x}_s)$ (Lohr (4.8) and (4.11))

It is also ok (even rather better) to use $\hat{V}(\hat{t}_{rat}) = N^2\left(1-\frac{n}{N}\right)\frac{s_e^2}{n}\left(\frac{\bar{x}_U}{\bar{x}_s}\right)^2$. (Lohr (4.10) and (4.11))

Regression estimator of t_y: $\hat{t}_{reg} = N\left(\bar{y}_s + \hat{B}_1(\bar{x}_U - \bar{x}_s)\right)$,

where $\hat{B}_1 = \frac{\sum_{i\in s}(x_i-\bar{x}_s)(y_i-\bar{y}_s)}{\sum_{i\in s}(x_i-\bar{x}_s)^2}$ (Lohr (4.15))

$V\left(\hat{t}_{reg}\right) \approx N^2\left(1-\frac{n}{N}\right)\frac{S_y^2}{n}(1-R^2)$,

where $R = \frac{S_{xy}}{S_x S_y}$ is the finite population correlation coefficient. (Lohr (4.18))

A variance estimator is obtained by replacing the population quantities S_y^2 and R with sample quantities. (Lohr (4.20))

Alternative, equivalent, variance estimator: $\hat{V}\left(\hat{t}_{reg}\right) = N^2\left(1-\frac{n}{N}\right)\frac{s_e^2}{n}$,

where $s_e^2 = \frac{1}{n-1}\sum_{i\in s}\left(y_i - \hat{B}_0 - \hat{B}_1 x_i\right)^2$, $\hat{B}_0 = \bar{y}_s - \hat{B}_1\bar{x}_s$ (Lohr p. 138-139)

Domain estimation in SRS

Let $u_i = y_i x_i$ with $x_i = \begin{cases}1 \text{ if unit } i \text{ belongs to the domain}\\ 0 \text{ otherwise}\end{cases}$ (Lohr p. 134)

The part of the sample that falls in domain *d* is denoted by s_d and the number of units in s_d is denoted by n_d.

Estimation of the **mean** of study variable in domain *d*: $\bar{y}_d = \frac{\bar{u}_s}{\bar{x}_s}$

$\hat{V}(\bar{y}_d) = \left(1-\frac{n}{N}\right)\frac{s_{yd}^2}{n_d}$, where $s_{yd}^2 = \frac{\sum_{i\in s_d}(y_i-\bar{y}_d)^2}{n_d-1}$ (compare Lohr (4.13))

Estimation of the **total** of study variable in domain *d*, t_d, two cases:

1. If the population size of the domain, N_d, is known: $\hat{t}_d = N_d\bar{y}_d$ (Lohr p. 135)
2. N_d is unknown: $\hat{t}_d = N\bar{u}_s$, where $\bar{u}_s = \frac{1}{n}\sum_{i\in s}u_i$. $\hat{V}(\hat{t}_d) = N^2\left(1-\frac{n}{N}\right)\frac{s_u^2}{n}$, where $s_u^2 = \frac{1}{n-1}\sum_{i\in s}(u_i - \bar{u}_s)^2$

Sample size estimation, SRS

We want this precision: $P(|\bar{y}_s - \bar{y}_U| \le e) = 0.95$. Then, with the approximation $fpc = 1, n = \frac{1.96^2 S_y^2}{e^2}$. (compare Lohr (2.25))

Stratification and poststratification

The population is divided into nonoverlapping groups that will exhaust the population fully. I prefer subscript *g* as a generic notation of the number of one poststratum and subscript *h* for a generic notation of the number of one stratum. For example, the sample and total in stratum *h* is denoted by s_h and t_h, respectively. Lohr uses subscript *h* for both kinds of population subsets.

For **stratified simple random sampling** the population mean $\bar{y}_U$ is estimated as

$\bar{y}_{str} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i\in s_h}\frac{N_h y_i}{n_h} = \frac{1}{N}\sum_{h=1}^{H}\hat{t}_h$ (Lohr (3.1) and (3.2))

and the variance as

$\hat{V}(\bar{y}_{str}) = \sum_{h=1}^{H}\frac{N_h^2}{N^2}\left(1-\frac{n_h}{N_h}\right)\frac{s_h^2}{n_h}$ (Lohr (3.5))

With stratified simple random sampling with proportional allocation, that is, the sample size in each stratum *h* is $n_h = n\frac{N_h}{N}$, the variance of the estimate $\hat{V}(\bar{y}_{str}) = \frac{1}{Nn}\left(1-\frac{n}{N}\right)\sum_{h=1}^{H}N_h s_h^2$ (Lohr p. 86)

Optimal allocation, equal costs: $n_h = n\frac{N_h S_h}{\sum_{h=1}^{H}N_h S_h}$ (Lohr (3.14))

For **simple random sampling followed by poststratification**, if the sample sizes in poststrata are $n_g = n\frac{N_g}{N}$, the variance estimator is the same:

$\hat{V}(\bar{y}_{post}) = \frac{1}{Nn}\left(1-\frac{n}{N}\right)\sum_{g=1}^{G}N_g s_g^2$ (Lohr (4.22))

For general poststratum sample sizes a variance estimator corresponding to the formula above marked as Lohr (3.5) can be used:

$\hat{V}(\bar{y}_{post}) = \sum_{g=1}^{G}\frac{N_g^2}{N^2}\left(1-\frac{n_g}{N_g}\right)\frac{s_g^2}{n_g}$

Poststratification estimator of the mean, SRS, general poststratum sample sizes:

$\bar{y}_{post} = \frac{1}{N}\sum_{g=1}^{G}\sum_{i\in s_g}\frac{N_g y_i}{n_g} = \frac{1}{N}\sum_{g=1}^{G}\hat{t}_g$

One-stage cluster sampling, unequal cluster sizes

N and n: number of clusters in the population and in the sample, respectively.

M_i and M_0: number of units in cluster *i* and in the population, respectively.

$t_i = \sum_{j=1}^{M_i} y_{ij}$ is the total of y_{ij} in cluster *i* (y_{ij} is the value of the study variable for unit *j* in cluster *i*). $\hat{t}_i = t_i$ because in one-stage cluster sampling, all units in the clustered are sampled.

Unbiased estimator of t_y: $\hat{t}_{unb} = \frac{N}{n}\sum_{i\in\mathcal{S}} t_i = \frac{N}{n}\sum_{i\in\mathcal{S}}\sum_{j=1}^{M_i} y_{ij}$ (Lohr p. 169)

Corresponding estimator of $\bar{y}_U$: $\hat{\bar{y}} = \frac{\hat{t}_{unb}}{M_0}$ (M_0 must be known)

$\hat{V}(\hat{\bar{y}}) = \frac{N^2}{M_0^2}\left(1-\frac{n}{N}\right)\frac{s_t^2}{n}$, where $s_t^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}}\left(\hat{t}_i - \frac{\hat{t}_{unb}}{N}\right)^2$ (Lohr (5.13) and p. 170)

Ratio estimator of $\bar{y}_U$: $\hat{\bar{y}}_{rat} = \frac{\hat{t}_{unb}}{\hat{M}_0} = \frac{\sum_{i\in\mathcal{S}}\hat{t}_i}{\sum_{i\in\mathcal{S}}M_i}$

$\hat{V}(\hat{\bar{y}}_{rat}) = \left(1 - \frac{n}{N}\right)\frac{1}{n\bar{M}^2}\frac{\sum_{i\in s}\left(t_i - \hat{\bar{y}}_{rat}M_i\right)^2}{n-1}$, where $\bar{M} = \frac{1}{n}\sum_{i\in s} M_i$

Horvitz-Thompson estimator

General sampling design, inclusion probability π_i

Unbiased estimator of t_y: $\hat{t}_{HT} = \sum_{i\in s}\frac{y_i}{\pi_i}$ (Lohr (6.19))

Response rate

Response rate computed as $\frac{(6)}{(4)+(3A)}$, where (6) is number of sample units that responds, (4) is number of sample units that are established to be in scope (i.e. belong to target population) and (3A) is the number of unresolved sample units that are believed to be in scope.

Weighting class estimator

$\hat{t}_{WC} = N\sum_{c=1}^{C}\frac{n_c}{n}\bar{y}_{cR}$, where C is number of classes, n_C is sample size in class c, $\bar{y}_{cR} = \frac{\sum_{i\in s_{cR}} y_i}{n_{cR}}$ is the mean of the respondents in class c. (Lohr page 341)

Poststratified estimator to adjust for nonresponse

$\hat{t}_{post} = \sum_{g=1}^{G} N_g\bar{y}_{gR}$, where G is number of poststrata, N_g is the population size in poststratum g, $\bar{y}_{gR} = \frac{\sum_{i\in s_{gR}} y_i}{n_{gR}}$ is the mean of the respondents in poststratum g. (Lohr page 342)

Stockholms universitet

Department of Statistics

Correction sheet

Date: 191203

Room: Brunnsvikssalen

Exam: Sample Survey

Course: Sample Survey

Anonymous code: 0008-G·CS

☒ I authorise the anonymous posting of my exam, in whole or in part, on the department homepage as a sample student answer.

NOTE! ALSO WRITE ON THE BACK OF THE ANSWER SHEET

Mark answered questions

1	2	3	4	5	6	7	8	9	Total number of pages
X	X	X	X	X	X	X	X		6 7
Teacher's notes				4	7	10	4		

Points	Grade	Teacher's sign.
25	E	

SU, DEPARTMENT OF STATISTICS

Room: Brunnsviks Anonymous code: 0008-GCS Sheet number: 1

1. ~~e~~ D

2. ~~a~~ E

3. a R

4. ~~b~~ E

5 ~~e~~ R

sorry E is correct

4

SU, DEPARTMENT OF STATISTICS

Room: Brunnsviks Anonymous code: 0008-GCS Sheet number: 2

6. $N_1 = 1000$ $N_2 = 1000$ $n_1 = 100$ $n_2 = 100$

a) It's a case of a systematic sampling (systematiskt urval) stratified systematic s. ~~4~~ 2

b) We treat the two archives as two strata.

$$\hat{t} = \frac{1000}{100} \cdot 500 + \frac{1000}{100} \cdot 20.000 =$$

$$= 10 \cdot 500 + 10 \cdot 20.000 = \boxed{205000 \text{ kronor}}$$ R 1

c) It is a problem associated with systematic sampling: once we have decided on the starting point (here → some point between 1 and 10, we will be choosing every 10th unit, leaving the units in-between with 0 probability of being included. OK 2

d) One way could be to use the percentages of units without cost in every stratum, the percentage of units requiring maintenance appears to be positively correlated with the total maintenance cost.

please turn →

→

We would use a ratio estimator:

$$\hat{t}_{ret} = t_x \frac{\hat{t}_y}{\hat{t}_x}$$

In our case it would be:

$$\hat{t}_{ret} = \boxed{\text{total percentage requiring maintenance previous survey}} \; \frac{\boxed{\text{total cost in the sample}}}{\boxed{\text{percentage requiring maintenance in the sample}}}$$

will induce bias because it is not $\bar{y}_u$

2 End Q 6

7

SU, DEPARTMENT OF STATISTICS

Room: Brunnsvik Anonymous code: 0008-GCS Sheet number: 3

7. What we know:

$N = 60\,000$ $n = 1000$

$n_R = 600$

a) MCAR means that the unit non-response is not associated (correlated) with my target variable nor with my indicator variable. auxiliary variable?

MAR means that unit non-response may be associated with my indicator variable, but not my target variable 4

'Indicator variable' is not clear (CLEAR) in this context but you are on the right track

b) sum women who travelled by plane: 40

Proportion: $p_s = \frac{40}{320} = 0{,}125$. R 1

$$V_{p_s} = \left(1 - \frac{n}{N}\right)\frac{p(1-p)}{n_1} \qquad N_{women} = 30.000$$

$$\frac{n}{N} = \frac{600}{60\,000}$$

$$V_{p\,women} = \left(1 - \frac{320}{30000}\right)\frac{0{,}125 \cdot 0{,}875}{320} =$$

$$= 0{,}98933 \cdot \frac{0{,}109375}{320} = 0{,}000381499$$

st error, $S = 0{,}0183888$

c) Proportion of men flying:

1

$$\frac{45+15}{280} = \boxed{0,2142857} = p_{men}$$

Räcker med detta konstaterande. Eftersom data säger att ~~#~~ män flyger oftare än kvinnor, är data oförenliga med det motsatta påståendet

According to our data, the proportion of men flying is higher than the proportion of women who fly.

So, the initial answer to the question is – no, our data does not suggest that women fly more frequently than men. However we do not know yet whether the difference is significant.

$$V_{men} = \left(1 - \frac{280}{30.000}\right) \cdot \frac{0,21428 \cdot 0,78572}{280} =$$

$$= 0,9906 \cdot \frac{0,1683673469}{280} =$$

$$\boxed{= 0,000595648}$$

$$S_{men} = 0,024405$$

men ↑

Difference of proportions: $0,2142857 - 0,125$

$$= \boxed{0,0892857}$$

We assume 0 covarience

$$V_{p_{diff}} = V_{p_{men}} + V_{p_{women}} - 2\,cov(p_{men}\, p_{women})$$

forts nästa side:

7c continued.

$$V_{p\,diff} = 0,000595648 + 0,000381499$$

$$= \boxed{0,000977147}$$

$$S_{diff} = 0,031259$$

Is the difference in flying frequency between men and women significant?

H_0 – there is no difference

H_A – there is a difference.

test variable $z = \dfrac{P_{diff} - 0}{S_{diff}}$

We perform a $\underline{2\text{-tailed}}$ test, 0,95 sig. level.

Reject the null hypothesis if $|z| > 1,96$.

$$z_{obs} = \frac{0,0892857}{0,031259} = 2,856319$$

H_0 – rejected

Answer: No, our data does not suggest that women travel by plane more often than men!

d) Travel by year.

Single flight a year, women

$F_d = N_d \bar{y}_d$

$\bar{y}_d = \frac{30}{320} = 0{,}09375$

$\hat{F}_{d_W} = 0{,}09375 \cdot 30000 = \boxed{2\,812{,}5}$

Assuming $N_d = 0.5N$?

~~OK with me, but you could have done this simpler,~~

$\hat{t} = \frac{60000}{600}(75 + 2 \cdot 25) = 12\,500$

as simple as that

Single flight a year, men

$F_d = \frac{45}{280} = 0{,}160714$

$\hat{F}_{d_m} = 0{,}160714 \cdot 30000 = \boxed{4821}$

2 flights or more, women

$\bar{y}_d \quad \frac{10 \cdot 2 \text{ (or more)}}{320} = 0{,}0625$

$\hat{F}_{d_W} = 0{,}0625 \cdot 30000 = 1875$ or more

2 flights or more, men:

$y_{d_m} = \frac{15 \cdot 2 \text{ (or more)}}{280} = 0{,}10714$

→ cont next p.

SU, DEPARTMENT OF STATISTICS

Room: Brunnsvik Anonymous code: 0008-GCS Sheet number: 5

$\hat{f}_{d_{men}} = 0,10714 \cdot 30\,000 = 3214,28$

The sum: $2812,5 + 4821 + 1875 + 3214,28 = \approx \boxed{12\,723}$ flights per year, or more.

1

e) $60000 \cdot 0,6 = 36\,000$ – women

$24\,000$ – men.

Single flight a year, women:

$\hat{f}_{d\,w} = 0,09375 \cdot 36000 = \boxed{3375}$

Single flight, men:

$\hat{f}_{d_m} = 0,160714 \cdot 24\,000 = \boxed{3857,136}$

2 or more, women:

$\hat{f}_{d_m} = 0,0625 \cdot 36000 = \boxed{2250}$

→

2 or more, men:

$$\hat{f}_{\substack{d\\m}} = 0{,}10714 \cdot 24000 = \boxed{2571{,}36}$$

R $\boxed{\text{The Sum: } \approx 12\,053 \text{ flights}}$

3

A note about MCAR, MAR (from part e):

If my intention is to assess the flight attitudes of women as opposed to men, I do not want the unit non-response to be correlated with that → that is that, for instance, women do not answer the question more often than men. If, on the other hand, the question is equally sensitive to men and women, the only problem is the indicator, that is the questio itself (both men and women may be ashamed to answer) → then the situation is MAR.

10

SU, DEPARTMENT OF STATISTICS

Room: Brunnsvik Anonymous code: 0008-GCS Sheet number: 6

(8.)

$$n = \frac{1{,}96^2 \cdot s^2_y}{e^2} \qquad (fpc \to 1)$$

a)

Pop mean = 1,4

$e = 1{,}4 \cdot 0{,}2 = 0{,}28$

$s^2_y = 9{,}2$

$$n = \frac{1{,}96^2 \cdot 9{,}2}{0{,}28^2} = 450{,}8 \approx \boxed{451}$$

$\Downarrow$

required sample size.

4