MVE155 - Statistial Inference
Assignment 1 - Survey Sampling

Adam Wirehed
wirehed@student.chalmers.se

February 7, 2020

1 Exercise A

The following results were computed from a sample (n=600) from the data "famalies.txt". The standard deviation on the sample is calculated using the correct formula (divided by: (n - 1)) from the python library *pandas*. The sample function from the same library does the sampling without replacement by default. The standard error is calculated using the equation (1), since we sample without replacement. The equation for standard deviation on a mean measurement can be seen in 2.

$$S_{\bar{x}} = \frac{s}{\sqrt{n}}\sqrt{1 - \frac{n}{N}} \tag{1}$$

- Proportion of husband-wife family: $\hat{p}_{hw} = 0.75500$
  - Standard error: $S_{\hat{p}_{hw}} = 0.01745$
  - Confidence interval (95%): $I_p = [0.72079, 0.78921]$
- Average number of children per family: $\bar{x}_{child} = 0.97167$
  - Standard error: $S_{\bar{x}_{child}} = 0.04548$
  - Confidence interval (95%): $I_\mu = [0.88252, 1.06081]$
- Average number of persons per family: $\bar{x}_{person} = 3.11333$
  - Standard error: $S_{\bar{x}_{person}} = 0.04885$
  - Confidence interval (95%): $I_\mu = [3.01757, 3.20909]$

2 Exercise B

100 samples (n=400) was made from the data "famalies.txt" without replacement. For the mean and standard deviation of each of the samples, add a print statement for the vectors *xBarVec* and *stdVec* and execute the python code.

The formula used for the standard error is equation (1). Since we have sampling with a finite population (no replacement).

Where n is sampling size, and N is population size and s is the sample standard deviation.

- Mean income of the 100 samples (n=400): $\bar{x}_{income} = 41026.74$
- Standard deviation in income of the 100 samples: $s_{\bar{x}} = 1526.88$
- Number of confidence intervals (95%) containing the population mean ($\mu_{income} = 41335.51$) = 95 (as it should be)

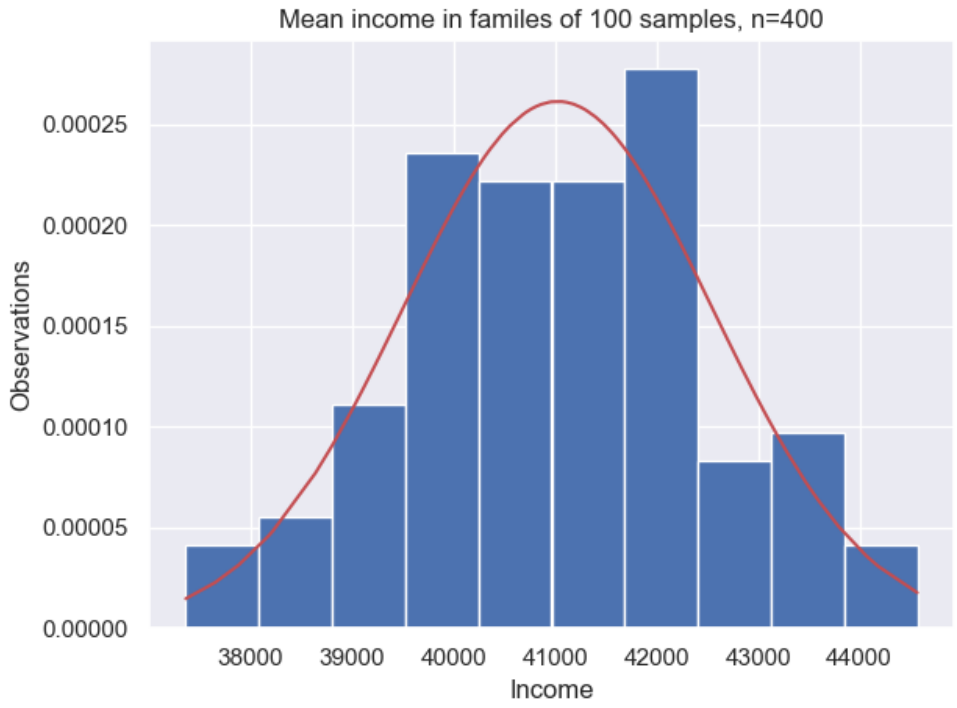

Figure 1: Histogram of mean income from 100 samples (n=400) fitted with a normal distribution, N($\bar{x}_{income}, s_{\bar{x}}$)

100 samples (n=100) was made from the same data to compare the results from the prevoius samples of size (n=400):

- Mean income of the 100 samples (n=100): $\bar{x}_{income} = 41004.59$

- Standard deviation in income of the 100 samples: $s_{\bar{x}} = 3398.84$

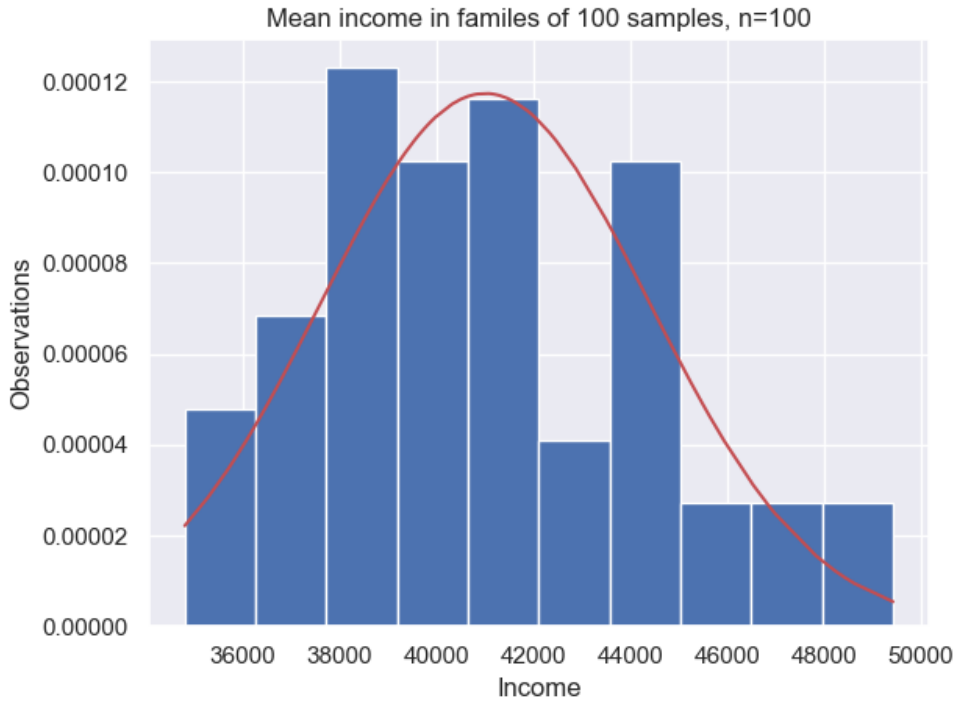

Figure 2: Histogram of mean income from 100 samples (n=100) fitted with a normal distribution, $\mathrm{N}(\bar{x}_{income}, s_{\bar{x}})$

In both the figures 1 and 2 the data somewhat follows the normal distribution. Each distribution have one or two bins that does not follow the normal distribution (eg. bin above 46000 in Figure 2). However the samples with n=400 does indeed follow the normal distribution a bit better than the samples with n=100. By the Central Limit Theorem, the sample mean distribution is approximately normal. So our observations are in order.

The value that differentiate the most between the two "surveys" are the standard deviation. For the survey with (n=100) results in a standard deviation that is more than double the size compared to the survey with (n=400). If we look at the equation that computes the sample standard deviation (2), we see that it is divided by the amount of data in the sample. Where x_1 is the mean value of sample number one and $\bar{x}$ is the mean value of all the mean values from the sampling. Since the size difference of the samples are of times four. We can expect a difference in the standard deviation by a factor of $\sqrt{4} = 2$, which we almost have. Hence the more data we have in our sample the lower standard deviation **of the mean** we can expect when we sample from the same population. In "normal" cases increasing the sample size would not necessarily decrease the standard deviation, but push it towards the population standard deviation.

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n - 1}} \tag{2}$$

If we look at the histograms we can see that the plot with the samples (n=100) is a bit more skewed than

the plot with (n=400). Since we have less data it is more vulnerable to outliers which can skew or "spread" the data. This is also represented in the difference between the standard deviation values.

3 Exercise C

For this exercise we stratify the family data by regions North, East, South and West.

The sampling fractions, optimal and proportional, are computed using equation (3) and (4) respectively. Where the standard deviation is computed on the income for the families. Since we calculate the standard deviation for the whole population (the new stratified datasets based on the region are whole populations) we don't divide by (n-1) but with n instead. This is done in *Pandas* by the input argument *std(ddof=0)*.

$$p_j = \frac{w_j \sigma_j}{\bar{\sigma}} \tag{3}$$

$$p_j = w_j \tag{4}$$

$$\bar{\sigma} = \omega_1 \sigma_1 + ... + \omega_n \sigma_n \tag{5}$$

The computed values for the different regions can be seen in Table 1.

Table 1: Computed values for optimal and proportional allocation

Region	Optimal alloc.	Proportional alloc.
North	0.25292	0.23126
East	0.22199	0.23675
South	0.28916	0.30664
West	0.23591	0.22536

Since we use simple random sampling (with replacement) we use the formula in equation (6) when computing the standard error. Both of the samples (stratified and simple) will be sampling with replacement, which is done in *Pandas* with the input argument *df.sample(replace=True)*.

$$S_{\bar{x}} = \frac{s}{\sqrt{n}} \tag{6}$$

The stratified sample mean and sample standard deviation is then computed using equation (7) and (8) respectively. Where n=4 (each of the regions) and n_n is the sample size from the stratified sample n.

$$\bar{x} = \omega_1 \bar{x}_1 + ... + \omega_n \bar{x}_n \tag{7}$$

$$s_{\bar{x}} = \sqrt{\omega_1^2 \frac{s_1^2}{n_1} + ... + \omega_n^2 \frac{s_n^2}{n_n}} \tag{8}$$

- Average income: $\bar{x}_s = 40769.93$
- Standard error: $S_{\bar{x}_s} = 1416.43$
- Confidence interval (95%): $I_\mu = [37993.71, 43546.14]$

A non-stratified sample (simple random sample with replacement) of the same sample size (n=500) was made on the same population for comparison.

- Average income: $\bar{x}_s = 42093.64$
- Standard error: $S_{\bar{x}_s} = 1572.56$
- Confidence interval (95%): $I_\mu = [39011.42, 45175.87]$

We can observe that the standard error is a bit lower for the stratified sample. Which is good since the variances for different sampling allocations are ordered in the following way:

$$Var(\bar{X}_{so}) \leq Var(\bar{X}_{sp}) \leq Var(\bar{X})$$

The mean income for the simple random sample (with replacement) is higher then for the stratified sample. This may be due to that when we sample with proportional allocation based on region, we sample more families from areas with a lot of people (smaller living area per family). And less from regions with people with lager living areas. This could reduce the amount of outliers (families with really high income) and push our estimated mean towards the median. But without proper testing, this is nothing more than speculation.