LECTURE 3: BASIC STATISTICAL FACTS FOR INFERENCE

1 Simple random sampling

1. Unbiased estimator of the population mean, μ, is the sample mean

$$\overline{Y} = \frac{1}{n} \sum X_i$$

where

$$E\overline{Y} = \mu$$

The variance of the sample mean is
a. With replacement

$$var(\overline{Y}) = \frac{\sigma^2}{n}$$

b. Without replacement

$$var(\overline{Y}) = \frac{\sigma^2}{n} \frac{N-n}{N-1}$$

2. Un unbiased estimator for π is the proportion of elements that belong to that category in the sample

$$\widehat{p} = \frac{\#belong}{n}$$

$$E\widehat{p} = \pi$$

The variance of the proportion estimation is
a. With replacement

$$var(\widehat{p}) = \frac{\pi(1-\pi)}{n}$$

b. Without replacement

$$var(\widehat{p}) = \frac{\pi(1-\pi)}{n}\frac{N-n}{N-1}$$

Notice that:

- The distribution of both estimators tend to be more concentrated when n increases.
- In sample with replacement population remains unchanged throughout the selection of the sample. Then if the sample is large, N≈N-1, and with replacement then

$$\frac{N-n}{N-1} \approx \frac{N-n}{N} = 1 - \frac{n}{N} \approx 1$$

- Both estimators are unbiased no matter if the sample is with or without replacement but the variance of the estimator without replacement is lower

$$var(\overline{Y})_{noreplac} = \frac{\sigma^2}{n}\frac{N-n}{N-1} < \frac{\sigma^2}{n} = var(\overline{Y})_{replac}$$

Group exercise

We have information on the 6 firms of a little town. The characteristics are the following

Firm	Number of employees	Intent to hire?
A	9	1
B	8	1
C	6	0
D	2	0
E	1	0
F	5	0

where 1 in the third columns is they intent to hire and 0 is not. With this information we can calculate the basic descriptive statistics.

Number of firms=N=6

$\mu = 31/6$

$\pi = 2/6 = 0.333$

Variance number of employees=σ^2=8.472

a. If we take samples of two firms, obtain the probablity tree assuming random sampling.

b.Assuming no replacement obtain the distribution of the sample mean and the sample proportion.

2 How large the sample should be?

Usually you try to set a tolerable error

$$P(|\overline{Y} - \mu| \leq e) = 1 - \alpha$$

where e is the margin of error and α is the confidence level.

What happen if we increase the sample by a factor of k?

$$var(\overline{Y}) = \frac{\sigma^2}{kn}\frac{N - kn}{N - 1}$$

The improvement in accuracy is

$$improvement = \frac{\sigma^2}{kn}\frac{N - kn}{N - 1} - \frac{\sigma^2}{n}\frac{N - n}{N - 1} = \frac{N\sigma^2}{kn}\frac{(k - 1)}{N - 1}$$

In relative terms the porcentual improvement in variance when going from n to kn observations is

$$(1 - \frac{1}{k})(\frac{N}{N - n})$$

Group exercise

Imagine that we have a population of N=500 and we double the sample size from 100 to 200 observations.What is the relative improvement in accuracy?

Answer=62.5%.

Group exercise

Assume that you do not have extra information and you want to estimate a proportion using the worst case scenario (p=0.5). If you want to get a (plus-minus) 5% error with a level of significance (probability) of 95% and , what should the size of the sample be? Assume replacement (or not correction for small sample)

Answer: Notice that the interval size in this simple case is

$$\begin{aligned} e &= Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \\ 0.05 &= 1.96 * \sqrt{\frac{0.5 * 0.5}{n}} \\ n &\approx 384 \end{aligned}$$

To determine the sample size we need to estimate a population characteristic with given accuracy, say $var(\overline{Y}) = c$. From the definition of the variance we can obtain the following result

$$n = \frac{N\sigma^2}{(N-1)c + \sigma^2}$$

For the proportions the size should be calculated as

$$n = \frac{N\pi(1-\pi)}{(N-1)c + \pi(1-\pi)}$$

Problema:
a. what is σ^2? Use an earlier similar sample or a pilot.
b. Very difficult to defend a particular value for c.

An alternative approach: assume that the purpose of the sample is to estimate the proportion π of elements belonging to a given category. One wants the estimator of π that is the closest to the true parameter value. Suppose that is is desired that $\widehat{p}$ be in the interval from $\pi + d$ to $\pi - d$ with a probability $1 - \alpha$. The parameter d give us an indications of "closeness" and 2d is the size of the interval. Therefore, d and α determine the desired accuracy.

$$n = \frac{N\pi(1-\pi)}{(N-1)d^2 + \pi(1-\pi)}$$

where

$$d = c/Z_{\alpha/2}$$

Additional problem: π. Worst case scenario: $\pi = 0.5$ (variance is 0.25, the largest). If other information is available, for instance we know that π cannot be larger than 0.2, then we will use 0.2.

Group exercise

How large should a sample without replacement be taken of a district with N=50000 households so that the esimated population proportion of households buying a given product is in the interval $\pi - 0.01$ $\pi + 0.01$ with probability 95%? A survey taken two year ago gave an estimator of the proportion equal to 0.4.

$$n = \frac{50000 * 0.4 * (1-0.4)}{(50000-1)(0.01/1.96)^2 + 0.4 * (1-0.4)} = 7,784$$

Group exercise

A telephone company plant to ascertain the condition of telephone poles in the region it services and the cost of the repair. There are 10,000 poles, a list of which is mantained by the company. From this list a random sample of 100 poles is selected without replacement. Crews were sent to calculate

the cost of needed repairs. Results: average repair cost=$83. sample variance=121. How many additional poles must be sampled if the estimate of the total cost of repairing all telephone poles (formed by pooling observations in the pilot and the planned sample) is to be within (plus-minus) $5,000 of the true cost with probability 90%?

Solution: we are now looking for the total number in a category.
So, instead of

$$\pi - c \leq \widehat{p} \leq \pi + c$$

we have

$$\begin{aligned} N\pi - Nc &\leq N\widehat{p} \leq N\pi + Nc \\ \tau' - c' &\leq T' \leq \tau' + c' \\ c &= c'/N \end{aligned}$$

therefore the sample size required for the estimate of the population mean of the variable Y to be in the interval from $\mu - c'$ to $\mu + c'$ with probability (1-α) is

$$n = \frac{N\sigma^2}{(N-1)D^2 + \sigma^2}$$

with $D = c/Z_{\alpha/2}$
N=10,000, 1-$\alpha = 0.90$, $Z_{\alpha/2} = 1.645$, average cost per pole=5,000/10,000=0.5. Using the variance of the pilot sample as an estimate of the population variance and c=0.5 the required sample is

$$n = \frac{10000 * 121}{(10000-1) * (0.5/1.645)^2 + 121} = 1,158$$

3 Confidence interval for simple random sampling

When n and N-n are large, an approximate 100(1-α)% confidence interval for μ , the population mean of the variable Y is

$$\overline{Y} \pm Z_{\alpha/2}\sqrt{\widehat{var}(\overline{Y})}$$

while one for π, the population proportion in a given category is

$$\widehat{p} \pm Z_{\alpha/2}\sqrt{\widehat{var}(\widehat{p})}$$

An unbiased estimator for the variance of the sample mean is

$$\widehat{var}(\overline{Y}) = \frac{S^2}{n-1}\frac{N-n}{N}$$

An unbiased estimator for the variance of a proportion is

$$\widehat{var}(\widehat{p}) = \frac{\hat{p}(1-\widehat{p})}{n-1}\frac{N-n}{N}$$

Notice that S^2 is the sample variance

$$S^2 = \frac{1}{n}\sum(Y_i - \overline{Y})^2 = \frac{1}{n}\sum Y_i^2 - \overline{Y}^2$$