The sample size n being 'sufficiently large', the 95% confidence interval is estimated in hectares as follows:

$$\left[\hat{\bar{Y}} \pm 1.96\sqrt{\frac{N-n}{N}\frac{s_y^2}{n}}\right] = \left[29.07 \pm 1.96\sqrt{\frac{2010-100}{2010} \times \frac{707.45}{100}}\right]$$
$$= [23.99; 34.15].$$

Exercise 2.2 *Occupational sickness*

We are interested in estimating the proportion of men P affected by an occupational sickness in a business of 1500 workers. In addition, we know that three out of 10 workers are usually affected by this sickness in businesses of the same type. We propose to select a sample by means of a simple random sample.

1. What sample size must be selected so that the total length of a confidence interval with a 0.95 confidence level is less than 0.02 for simple designs with replacement and without replacement ?
2. What should we do if we do not know the proportion of men usually affected by the sickness (for the case of a design without replacement) ?

To avoid confusions in notation, we will use the subscript WR for estimators with replacement, and the subscript WOR for estimators without replacement.

Solution

1. a) Design with replacement.
   If the design is of size m, the length of the (estimated) confidence interval at a level $(1-\alpha)$ for a mean is given by

$$\text{CI}(1-\alpha) = \left[\hat{\bar{Y}} - z_{1-\alpha/2}\sqrt{\frac{\hat{S}_y^2}{m}}, \hat{\bar{Y}} + z_{1-\alpha/2}\sqrt{\frac{\hat{S}_y^2}{m}}\right],$$

   where $z_{1-\alpha/2}$ is the quantile of order $1-\alpha/2$ of a random normal standardised variate. If we denote $\hat{P}_{WR}$ as the estimator of the proportion for the design with replacement, we can write

$$\text{CI}(1-\alpha) = \left[\hat{P}_{WR} - z_{1-\alpha/2}\sqrt{\frac{\hat{P}_{WR}(1-\hat{P}_{WR})}{m-1}}, \right.$$
$$\left. \hat{P}_{WR} + z_{1-\alpha/2}\sqrt{\frac{\hat{P}_{WR}(1-\hat{P}_{WR})}{m-1}}\right].$$

Indeed, in this case,

$$\widehat{\text{var}}(\hat{P}_{WR}) = \frac{\hat{P}_{WR}(1-\hat{P}_{WR})}{(n-1)}.$$

So that the total length of the confidence interval does not exceed 0.02, it is necessary and sufficient that

$$2z_{1-\alpha/2}\sqrt{\frac{\hat{P}_{WR}(1-\hat{P}_{WR})}{n-1}} \leq 0.02.$$

By dividing by two and squaring, we get

$$z^2_{1-\alpha/2}\frac{\hat{P}_{WR}(1-\hat{P}_{WR})}{n-1} \leq 0.0001,$$

which gives

$$n-1 \geq z^2_{1-\alpha/2}\frac{\hat{P}_{WR}(1-\hat{P}_{WR})}{0.0001}.$$

For a 95% confidence interval, and with an estimator of P of 0.3 coming from a source external to the survey, we have $z_{1-\alpha/2} = 1.96$, and

$$n = 1 + 1.96^2 \times \frac{0.3 \times 0.7}{0.0001} = 8068.36.$$

The sample size (n=8069) is therefore larger than the population size, which is possible (but not prudent) since the sampling is with replacement.

b) Design without replacement.
If the design is of size n, the length of the (estimated) confidence interval at a level $1-\alpha$ for a mean is given by

$$\text{CI}(1-\alpha) = \left[\hat{\bar{Y}} - z_{1-\alpha/2}\sqrt{\frac{N-n}{N}\frac{s^2_y}{n}}, \hat{\bar{Y}} + z_{1-\alpha/2}\sqrt{\frac{N-n}{N}\frac{s^2_y}{n}}\right].$$

For a proportion P and denoting $\hat{P}_{WOR}$ as the estimator of the proportion for the design without replacement, we therefore have

$$\text{CI}(1-\alpha) = \left[\hat{P}_{WOR} - z_{1-\alpha/2}\sqrt{\frac{N-n}{N}\frac{\hat{P}_{WOR}(1-\hat{P}_{WOR})}{n-1}}, \right.$$
$$\left. \hat{P}_{WOR} + z_{1-\alpha/2}\sqrt{\frac{N-n}{N}\frac{\hat{P}_{WOR}(1-\hat{P}_{WOR})}{n-1}}\right].$$

So the total length of the confidence interval does not surpass 0.02, it is necessary and sufficient that

$$2z_{1-\alpha/2}\sqrt{\frac{N-n}{N}\frac{\hat{P}_{WOR}(1-\hat{P}_{WOR})}{n-1}} \leq 0.02.$$

By dividing by two and by squaring, we get

$$z^2_{1-\alpha/2}\frac{N-n}{N}\frac{\hat{P}_{WOR}(1-\hat{P}_{WOR})}{n-1} \leq 0.0001,$$

which gives

$$(n-1)\times 0.0001 - z^2_{1-\alpha/2}\frac{N-n}{N}\hat{P}_{WOR}(1-\hat{P}_{WOR}) \geq 0,$$

or again

$$n\left\{0.0001 + z^2_{1-\alpha/2}\frac{1}{N}\hat{P}_{WOR}(1-\hat{P}_{WOR})\right\} \geq 0.0001 + z^2_{1-\alpha/2}\hat{P}_{WOR}(1-\hat{P}_{WOR}),$$

or

$$n \geq \frac{0.0001 + z^2_{1-\alpha/2}\hat{P}_{WOR}(1-\hat{P}_{WOR})}{\left\{0.0001 + z^2_{1-\alpha/2}\frac{1}{N}\hat{P}_{WOR}(1-\hat{P}_{WOR})\right\}}.$$

For a 95% confidence interval, and with an *a priori* estimator of P of 0.3 coming from a source external to the survey, we have

$$n \geq \frac{0.0001 + 1.96^2\times 0.30\times 0.70}{\left\{0.0001 + 1.96^2\times\frac{1}{1800}\times 0.30\times 0.70\right\}} = 1264.98.$$

Here, a sample size of 1265 is sufficient. The obtained approximation justifies the hypothesis of a normal distribution for $\hat{P}_{WOR}$. The impact of the finite population correction $(1-n/N)$ can therefore be decisive when the population size is small and the desired accuracy is relatively high.

2. If the proportion of affected workers is not estimated *a priori*, we are placed in the most unfavourable situation, that is, one where the variance is greatest: this leads to a likely excessive size n, but ensures that the length of the confidence interval is not longer than the fixed threshold of 0.02. For the design without replacement, this returns to taking a proportion of 50%. In this case, by adapting the calculations from 1-(b), we find $n \geq 1298$. We thus note that a significant variation in the proportion (from 30% to 50%) involves only a minimal variation in the sample size (from 1265 to 1298).