Tutorial 9a: frequentist inference for proportions

Required files: `Tutorial9a.R`, `Tutorial9aFunctions.R`, `isatest.rdata`

1 Estimating the proportion of infected fish in a bay

In this practical we will make inferences about the proportion of fish infected with the salmon anemia virus in the Bay of Fundi (Canada). The data are from a study on the quality of diagnostic tests described in McClure *et al.* (2005). The file `isatest.rdata` contains information on 1,071 fish collected in the Bay of Fundi.

Use the script `Tutorial9a.R` to run the code required to answer the questions of this session

Sampling from the population

We will assume that the `isatest` data set contains the whole population of fish in the bay, and then we will take samples of different sizes from this full data set.

Question 1: Load the `isatest` data into `R` and to take samples of sizes 10, 30 and 100. What's the true value of the parameter we need to estimate? (*Use step 1 of the `R` script)*

Model and likelihood

As described in the lecture, the binomial distribution is a suitable model for the number of infected fish in a random sample. Based on the binomial model, the likelihood function for the proportion of infected fish is

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x},$$

where n is the sample size, x in the number of infected fish in the sample and p is the proportion of fish infected with the salmon anemia virus in the population.

Using this model, the likelihood for each sample size can be calculated for the range of possible values that the proportion parameter p can take. Run the 4 lines of step 2 of the `R` script to calculate the likelihood for the different samples of fish.

Question 2: Plot and compare the likelihood for sample sizes 10, 30 and 100 (*Use step 2 of the* `R` *script)*

Frequentist inference

As shown in the lecture, the maximum likelihood estimate of the proportion based on a binomial model is

$$\hat{p} = \frac{x}{n}$$

Using the Central limit theorem, a 90% confidence interval for the population proportion is

$$\left[\hat{p} - 1.64\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + 1.64\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Question 3: Compare the point estimates with the population proportion (the true value of the parameter). Compare also the confidence intervals obtained with each sample size. What is the effect of the sample size on the interval estimates? (*use step 3 of the* `R` *script)*