

Homework assignment 1

1. Investigating sources for discrepancy in an approximation.

In class we used R to compute the probability of the event $Z > 1.645$ for a standardized binomial test statistic. We obtained the value 0.044226, which is different than the target value of 0.05. Identify the possible factors that could have resulted in such a discrepancy. Which of them is the actual cause?

2. A case study: Testing genetic association in case-control trials.

An alternative approach for associating genes with diseases involves the collection of samples of unrelated individuals. In the case-control design two samples are collected: A sample of affected patients and a control sample of healthy individuals. The distribution of the alleles of a given gene in both groups is recorded and compared. For simplicity, we will assume that the gene under investigation has a *wild type* allele, which is not causing an increase in the likelihood of expressing the medical conditions and a *mutated* allele, which potentially does. A-priori, however, it is not known which of the two alleles is the mutated one and which is the wild-type.

Let n_1 be the size of the sample of cases and let n_2 be the size of the sample of controls. Mark one of the alleles by A and the other by a . Let X_1 be the number of A alleles among cases and let X_2 be the number of A alleles among controls. Under suitable assumptions it may be shown that $X_i \sim B(n_i, p_i)$, for $i = 1, 2$, and are independent. Denote $p = (p_1 + p_2)/2$ and $\delta = (p_1 - p_2)/2$.

1. Formulate the hypotheses for testing association between the gene and the medical condition.
2. Propose a test statistic and a rejection criteria. Consider separately the case when p is known and the case when it is not.
3. Investigate the distribution of your testing statistic under the null distribution. What is the effect of the different parameters on this distribution? Use R in your investigation.
4. Investigate the power function of your statistic. What is the effect of δ , p , n_1 and n_2 on the power? Use R in your investigation.