Tutorial 5:

Introduction to Data Science

Precision of SRS Estimator

Ernst C. Wit

October 14, 2024

Exercises

1. **Mark and Recapture Sampling (Equal Probabilities).**
   Context: In ecology, the *mark and recapture* method is used to estimate the population size of animals.

   Procedure:

   - A sample of animals is captured, marked, and then released back into the wild.
   - After some time, another sample is captured, and the number of marked animals in the recaptured group is counted.

   Example: Consider a population of $N = 500$ animals in a forest. A researcher captures and marks $M = 100$ animals and then releases them. After a period, the researcher recaptures a random sample of $n = 50$ animals.

   Assumptions:

   - All animals have an equal probability of being recaptured.
   - The population remains closed (no animals are added or removed during the period).

   Let X be the number of marked animals recaptured. This can be modeled as a **binomial distribution**:

   $$X \sim \text{Binomial}(n = 50, p = \frac{M}{N})$$

   where $p = \frac{M}{N}$ is the probability of capturing a marked animal in the recapture phase.

   Last week we saw that we can estimate the number of animals in the forest N via

   $$\hat{N} = \frac{50 \times 100}{X}.$$

   This week we focus on the estimate of the inverse of N, i.e.,

   $$\widehat{N^{-1}} = \frac{X}{50 \times 100}$$

   Tasks:

   (a) What is the expected value of the estimator $\widehat{N^{-1}}$ as a function of N?

   (b) What is the variance and standard deviation of the estimator $\widehat{N^{-1}}$ as a function of N?

(c) What is the approximate distribution of $\widehat{N^{-1}}$ according to the Central Limit Theorem as a function of N?

(d) We detect $X = 12$ animals. Use this value to calculate a replacement for the unknown standard deviation for N in the standard deviation in question (c).

(e) Use the distribution in (d) to calculate a 95% confidence interval for N^{-1}.
HINT: use also the figure below.

(f) By inverting the interval, find a 95% confidence interval for N. Does this interval include 500?

(g) Use Python or R to create the population with 500 individuals of which 100 hundred are marked. Then

- sample 1000 times 50 animals at random;
- for each sample ($i = 1, \ldots, 1000$) use the number of marked animals x_i to calculate the 95% confidence interval for N according to the method in (f).
- for each sample record whether 500 was inside the confidence interval.

(h) For your simulation in (g) with what relative frequence did your confidence interval include 500? Did you expect to find this relative frequency?

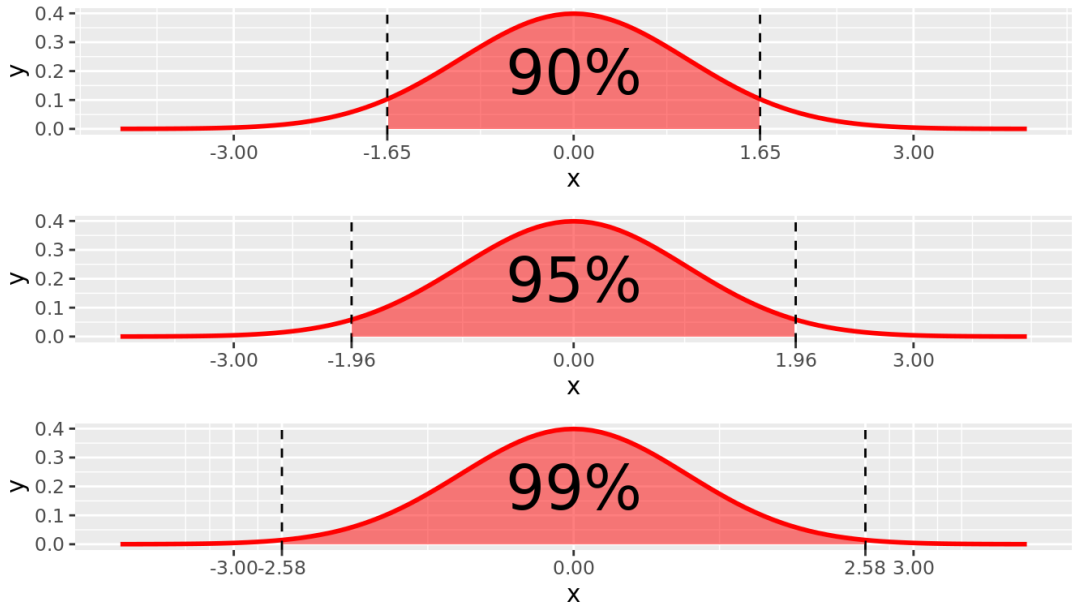

Figure 1: Quantiles for a standard normal distribution $Z \sim N(0, 1)$