11. - 12. Uncertainty and statistical models

Contents

Introduction

These notes consider Chapters 11 (statistical models and dealing with uncertainty) and 12 (statistical tests and hypothesis testing) of the handbook. These chapters are more theoretical with little code examples so I decided to combine them. There are, however, some formulae and concepts worth writing down.

Sample mean as a simple model

See page 227.

Suppose we have n measurements $x_1, x_2, \ldots, x_n$ with a sample mean of $\hat{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. The simplest estimate for the error is the Sum of Squares (SoS) defined as

$$SoS = \sum_{i=1}^{n}(x_i - \hat{x})^2.$$

Squaring makes sure that each term of the sum stays non-negative and, therefore, describes the squared "distance" of each measurement to the sample mean. The SoS is a very rough estimate and not that usable as the value tends to increase while n increases. A better estimate is the Mean Squared Error (MSE), which is the SoS divided by n. In other words, MSE is the mean of the squares of the errors defined as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x})^2.$$

This estimate is already better but in the wrong dimension (squared) compared to the measurements. More meaningful estimate of the error can be achieved by taking the square root of the MSE i.e. $\sqrt{MSE}$.

Looking back to Chapter 6 (pages 116-117) we can see that the MSE is the population variance (σ^2) and $\sqrt{MSE}$ is the standard deviation (σ) **if** the measurements cover the whole population.

Law of large numbers and central limit theorem

See page 230. Formal definitions mentioned here are partially supplemented from Wikipedia.

Suppose we have an infinite sequence of independent and identically distributed random variables $X_1, X_2, \ldots, X_n$ with expected values $E(X_1) = E(X_2) = \cdots = \mu < \infty$. The law of large numbers (LLS) states that the sample mean

$$\hat{X_n} = \frac{1}{n}(X_1 + \cdots + X_n)$$

converges to the expected value of the population i.e.

$$\hat{X_n} \to \mu \text{ as } n \to \infty.$$

In other words, the expected value of the population can be estimated using the sample mean if the amount of samples is large enough. This, however, does not tell us anything about the precision of the estimate. In practice, it is impossible to draw infinite amount samples meaning some uncertainty always remains.

In addition, the central limit theorem (CLT) states that, as $n \to \infty$

$$\hat{X_n} \sim N(\mu, \frac{\sigma^2}{n}),$$

where σ^2 is the variance of the population.

In summary, as $n \to \infty$, the LLS states that the sample average converges to the expected value of the population and the CLT states that the distribution of $\hat{X_n}$ gets arbitrarily close to the normal distribution *regardless*[1] of the original distribution of X_i. Additionally, we now have an error estimate for the precision of the mean for the finite sample size, namely the standard error of the (sample) mean (SEM)

$$\sigma_{\hat{x}} = \frac{\sigma}{\sqrt{n}}$$

which is the standard deviation of the distribution of $\hat{X_n}$. The SEM is the estimate for how close the sample mean is to the expected value of the population. We can immediately see that the precision increases when the sample size increases or the variance of the population decreases. Note, though, that the the increase in precision by drawing more samples is not linear but slower i.e. a square root of n.

Confidence intervals

See page 234.

By definition of the normal distribution, about 68.27 % of the values are within one standard deviation from the expected value ($\mu \pm \sigma$). This is illustrated in the figure below. A confidence interval (CI) is the interval where a parameter being estimated can typically be found (with a certain confidence level). For instance, in case of our expected value, 68.27 % of the possible values for the sample mean are within one SEM from the expected value of the population. This **does not** mean that the probability to find the expected value of the population within $\pm\sigma$ interval is 68.27 %! It merely states that, if an infinitely many samples were drawn, a sample mean would be within the CI 68.27 % of the cases. Of course, more sensible confidence levels are usually used. For instance, a 95 % confidence level corresponds to roughly $\pm 2\sigma$ interval (1.96 to be more precise). Additionally, when the standard deviation of the population is not known (required to calculate the SEM), the Student's t-distribution can be used to model the distribution of the sample means.

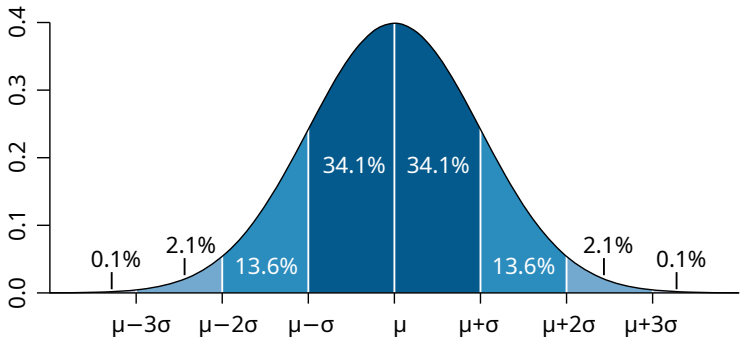

Figure 1: For the normal distribution, the values less than one standard deviation away from the mean account for 68.27 % of the set; while two standard deviations from the mean account for 95.45 %; and three standard deviations account for 99.73 %. (Wikipedia/Ainali)

[1] See Figure 11.6. in page 231.