Data Science Career Track

The Art of Statistics, Chapter 3: Why are we looking at data? Take-Away Notes

The chapter introduces basic statistical concepts by applying them in a basic **inductive inference** process. **Inductive inference** = the process of learning about general principles from specific examples, and involves four stages and three transitions (signaled by the '→' below):

1. Data → 2. Sample → 3. Study Population → 4. Target Population

Our **Data** is our raw data: whichever dataset we have to work with.

Our **Sample** is the set of things our Data gives us information about. The transition from our Data to our Sample requires that our Data give accurate information about our Sample.

Our **Study Population** is the group we are actually studying, that is, the set of things that could have been in our Sample. Perhaps our Sample is just the Study Population: this occurs when we have all the data. But most of the time, things are in our Study Population that aren't in our Sample; most of the time, there are things that could have been in our Sample, that aren't. The transition from Sample to Study Population in the inductive inference process requires that our Sample is *representative* of the Study Population.

Our **Target Population** is the set of things we want to draw conclusions on (or discover things about). The transition from the Study Population to the Target Population again requires that the Study Population is representative of the Target Population.

A **population** can be thought of as a physical group of individuals, but also as providing the *probability distribution* for a random observation: the probability of a random observation having a certain value for some attribute of interest.

The **probability distribution** is the pattern in the whole group of interest.

The **normal distribution** is correctly visualized with a **bell-shaped curve**, and can be mathematically characterized in terms of its mean and standard deviation: roughly 95% of a normally distributed population will be contained in the interval given by the mean +- two standard deviations, and 99.8% of the data in the central +- three standard deviations.