4 Hypothesis Testing

4.1 Background

The word *statistics* means different things to different folks. To a football fan, statistics are passing, and scoring; to the manager of a power station, statistics are the amount of pollution being released into the atmosphere. To a bank, statistics is the chance that a loner will repay her loan on time. To a student taking this course, statistics are the grades on your quizzes and final exam in the course.

Each of these people is using the word correctly, yet each person uses it in a different way. All of them are using statistics to help them make decisions. The word *statistik* comes from the Italian word *statista* meaing "statesman". It was first used by Gottfried Achenwall (1719-1772), a professor at Marlborough and Gottingen. However, long before this, people had been recording and using data. Statisticians commonly separate statistical techniques used into two broad categories: *descriptive statistics* and *inferential statistics*. Suppose a professor computes an average grade for one math class and use it to describe the performance of that one class then this is descriptive statistics. Graphs, tables, and charts that display data are all under descriptive statistics.

Now suppose this professor decides to use this average grade achieved in one class to estimate the average grade achieved in all other class of the same math course. This is inferential statistics. Obviously, any conclusion the professor makes about the other class is based on generalization that goes far beyond the data for the original math class; this generalization may not be valid, so the professor must state how likely it is to be true. Statistical inference involves **generalization** and the **probability** of their validity.

4.1.1 Some Concepts

1. Random Variable

2. Expected Value of a Random variable

   The expected value of a discrete random variable is nothing more than the weighted average of each possible outcome, multiplied by the probability of that outcome happening.

4.1.2 Normal Distribution

There are two basic reasons why the normal distribution occupies such a prominent place in statistics. First it has properties that make it applicable to a great many situations. Second, the normal distribution comes close to fitting the actual observed frequency distributions of many phenomena, including human characteristics (weights, heights).

Characteristics:

1. The curve has a single peak; thus, it is unimodal. It has the **bell shape**.
2. The mean of a normally distributed population lies at the center of its curve.
3. Because of the symmetry of the normal distribution, the median and the mode are also at the center. The mean, median and mode are the same value.
4. The two tails of the distribution extend indefinitely and never touch the horizontal axis.

4.1.3 Estimation

The probability theory forms the foundation for statistical inference. Statistical inference is based on estimation and hypothesis testing. In both estimation and hypothesis testing we shall be making inferences about characteristics of population from information contained in samples. To calculate the exact mean would be an impossible goal. Even so, we will be able to make an estimate, make a statement about the error that will probably accompany this estimate.

1. **Point Estimate**

   A point estimate is a single number that is used to estimate an unknown population parameter. It is often insufficient, because it is either right or wrong. And you do not know how wrong or right you it is.

2. **Interval Estimate**

   A interval estimate is a range of values used to estimate an unknown population parameter. It indicates the error in two ways: by the extent of its range and by the probability of the true population parameter lying within that range.

Example 1. Suppose the marketing research director needs an estimate of the average life in months of car batteries his company manufactures. We select a random sample of 200 batteries, record the car owners' names and interview these owners about the battery life they have experienced. Say the mean battery life is 36 months.

If we use the point estimate of the sample mean as the best estimator of the population mean, we would report that the mean life of the company's batteries is 36 months. But the director also asks for a statement about the uncertainty that will be likely to accompany this estimate. To provide such a statement, we need to find the *standard error of the mean.*

To measure the spread or dispersion, in our distribution of sample means, we can use the following formula: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, where σ is the standard deviation of the population. We can use this formula because mean of the sample mean is the population mean.

Now suppose we know the standard deviation of the population mean to be 10 months. Then $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = 0.707$ month. We could report to the director that our estimate of the life of the company's batteries **may** lie somewhere in the interval estimate of 35.293 to 36.707 months. This is helpful but insufficient information for the director. Next, we need to calculate the chance that the actual life will lie in this interval or in other intervals of different widths that we might choose, $\pm 2\sigma$, $\pm 3\sigma$, and so on.

Now, a large number of sample means from a population, the distribution of these means will approximate a normal distribution. Applying this to the battery life, our best estimate of the life of the company's batteries is 36 months, and we are 68.3 percent confident that the life lies in the interval of $\pm\sigma$. Similarly, we are 95.5 percent confident that the life falls within the interval of $\pm 2\sigma$, and 99.7 percent confident that battery life falls in the interval of $\pm 3\sigma$.