Statistics & Probability Notes

Harsh Agrawal

October 20, 2024

Abstract

This notebook compiles content from Statistics and Probability module delivered by Prof. Joseph Van Baterburg-Sherwood (2nd Year Imperial College London MBE). A total of 6 lectures are covered in these notes. Please suggest any relevant changes/improvements at ha1822@ic.ac.uk.

Contents

1 Introduction

'There are lies, damned lies, and statistics. Mark Twain'

If we begin a mission to create a drug for cancer, we need a way to test if it works. The ideal way would be to test it on every patient with that specific cancer — i.e the entire 'cancer population'. However, this is, for obvious reasons, not possible.

What we can do is test it on a subset of the population — i.e a 'sample' of the 'population'. We can then use the outcome of this test to **infer** the properties of the entire population. This is what **statistics** and **probability** allow us to do.

We start by looking at different types of statistics. On the basis of purpose and scope, we categorise them into **descriptive statistics** and **inferential statistics**.

- **Descriptive statistics** are used to describe the properties of the sample and population. It presents the underlying data in a meaningful way. Correct choice of desciptors is essential.Eg: Is the drug being trialled increasing the WBC count?
- **Inferential statistics** are used to make predictions about the population based on the sample. Eg: Is the drug being trialled effective in increasing the WBC count?

On the basis of parametrization, we categorise them into **parametric** and **non-parametric** statistics.

2 Descriptive Statistics

2.1 Sample Mean

A sample-mean (arithmetic) for n observations is given by:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

The sum of the deviations from the mean is always zero.

2.2 Sample Median

The sample median is the middle value of the data when ordered from smallest to largest. Its mathematical expression is given by:

$$\tilde{x} = \begin{cases} x\left(\frac{n+1}{2}\right) & \text{if } n \text{ is odd} \\ \frac{x\left(\frac{n}{2}\right)+x\left(\frac{n}{2}+1\right)}{2} & \text{if } n \text{ is even} \end{cases} \tag{2}$$

A sample median is a non-parametric measure of location — i.e, it doesn't assume the shape of the data distribution. Moreover, its less sensitive (more robust) to outliers in data than the sample mean.

2.3 Sample Variance and Standard Deviation

The sample variance defines, on average, how far the data points are from the mean. It is given by:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{3}$$

For convience of units, we take the square root of the variance to get the standard deviation. It is a better measure of uncertainty as it has the same units as the underlying data. It's given by:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{4}$$

Often to specify uncertainty, we specify mean $\pm$ (multiple) standard deviations. Eg: *The mean WBC count is* $\bar{x} \pm 2s$.

2.4 Sample Median Absolute Deviation

Just like the median, the median absolute deviation (MAD) is a non-parametric measure of spread. It is an alternative to standard deviation. Its calculated by obtaining the absolute deviations from the median, and then taking the median of those values. Mathematically, it is given by:

$$\text{MAD} = k \cdot \widetilde{|x_i - \tilde{x}|} \tag{5}$$

A scaling factor k is applied to make the MAD comparable to the standard deviation. For a normal distribution, $k = 1.4826$.

2.5 Sample Percentiles — Discrete

Useful for describing proportions of a dataset relative to a given number. Eg. — If you got higher marks than 90% of the class, you are in the 75^{th} percentile.

- Sample Median: 50^{th} percentile.
- Lower Quartile (LQ): Below 25^{th} percentile.
- Upper Quartile (UQ): Above 75^{th} percentile.
- Interquartile Range (IQR): Between 25^{th} and 75^{th} percentiles. $IQR = UQ - LQ$.

Let's look at an example to find the 60^{th} percentile η of:

$$X = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]$$