Statistics

Haydn Cheng

November 29, 2024

Contents

1 Statistics

1.1 Standard Deviation

For a set of data which contains of N distinct values $j_1, j_2 \cdots j_i \cdots j_N$ and each with frequency $f_1, f_2 \cdots f_i \cdots f_N$, we can define $P_i = \frac{f_i}{\sum_{i=1}^{N} f_i}$ as the probability of selecting the data j_i, then the average value of any function $g(j)$ can be written as

$$\langle g(j) \rangle = \frac{f_1 g(j_1) + f_2 g(j_2) + \cdots + f_N g(j_N)}{f_1 + f_2 + \cdots + f_N} = \sum_{i=0}^{N} P_j g(j_i). \tag{1.1}$$

To measure the dispersion of the set of data, the most intuitive way is to calculate the average of difference between each data and the mean

$$\sigma' = \frac{f_1(j_1 - \langle j \rangle) + f_2(j_2 - \langle j \rangle) + \cdots + f_N(j_N - \langle j \rangle)}{f_1 + f_2 + \cdots + f_N}. \tag{1.2}$$

However, σ' always equals to zero since $\sum_{i=1}^{N} f_i j_i = \langle j \rangle \sum_{i=1}^{N} f_i$. So, either we take the absolute value of each term or we square each term in eq. (1.2) such that the result is non trivial. We adopt the latter choice since the former is tedious. We introduce the quantity standard deviation σ defined by

$$\sigma^2 = \frac{f_1(j_1 - \langle j \rangle)^2 + f_2(j_2 - \langle j \rangle)^2 + \cdots + f_N(j_N - \langle j \rangle)^2}{f_1 + f_2 + \cdots + f_N} = \sum_{i=0}^{N} P_j (j_i - \langle j \rangle)^2. \tag{1.3}$$

Note that σ in eq. (1.3) is squared so that the standard deviation has the same dimension as j.

By expanding the bracket in eq. (1.3) and applying eq. (1.1), we have

$$
\begin{aligned}
\sigma^2 &= \sum_{i=1}^{N} P_i(j_i^2 - 2j_i\langle j\rangle + \langle j\rangle^2) \\
&= \sum_{i=1}^{N}(j_i)^2 P_i - 2\langle j\rangle \sum_{i=1}^{N}(j_i)P_i + \langle j\rangle^2 \sum_{i=1}^{N} P_i \\
&= \langle j^2\rangle - 2\langle j\rangle\langle j\rangle + \langle j\rangle^2 = \langle j^2\rangle - \langle j\rangle^2.
\end{aligned}
\tag{1.4}
$$

Therefore, we have the useful identity

$$
\sigma = \sqrt{\langle j^2\rangle - \langle j\rangle^2}. \tag{1.5}
$$

1.2 Probability Density Function

When we consider continuous variable, the probability of obtaining a certain value becomes meaningless now as there are infinite choices now so we define the probability density function $\rho(x)$ such that the probability of obtaining a value between x and $x + dx$ equals to $\rho(x)dx$. Therefore, just like the previous section, we have the relations

$$
\int_{-\infty}^{+\infty} \rho(x)dx = 1 \tag{1.6}
$$

and

$$
\langle f(x)\rangle = \int_{-\infty}^{+\infty} f(x)\rho(x)dx \tag{1.7}
$$

Example: Griffith(3rd ed.) Example 1.2

Question: Consider an object being released at height h. Find the average distance $\langle x\rangle$ from the point of release if a random instant is chosen.

Solution: Let $\rho(x)dx$ be the probability of a random instant being located between x and $x + dx$ which is equals to $\frac{dt}{T}$. Since $v = \frac{\mathrm{d}x}{\mathrm{d}t} = gt$ and $T = \sqrt{\frac{2h}{g}}$, so we have $\rho(x) = \frac{1}{2\sqrt{hx}}$.$\langle x\rangle$ can then be obtained from straightforward integration

$$
\langle x\rangle = \int_0^h \frac{x}{2\sqrt{hx}}dx = \frac{h}{3} \tag{1.8}
$$