Contents

Chapter 5

ELEMENTARY STATISTICS, EMPIRICAL PROBABILITY DISTRIBUTIONS, AND MORE ON SIMULATION

5.1 Connecting Probability with Observations of Data

Scope of the chapter

In chapter2 to chapter4: axioms of *probability*, various *probability* relationships

$\Longrightarrow$ *world of mathematical models*

$\Longrightarrow$ we now deal with *problems of real world*

$\Longrightarrow$ $\ni$: sample mean, variance, standard deviation, empirical distribution of random data etc..

$\Longrightarrow$ realm of *statistics*

Remark:

1. We will also consider *estimation theory*, and *decision making* based on probabilistic concepts in later chapters.
2. Computer simulation of random phenomena will be revisited as well.

5.2 Sample Mean and Sample Variance

Suppose we take a sample of manufactured items $\{x_1, x_2, \ldots, x_n\}$, e.g. resistors:

1. *population*: totality of items $\{x_1, x_2, \ldots, x_n\}$
2. *sample*: individual x_i's

Simple Statistics (of interest)

Definition 5.1 Sample mean*:*
The sample mean of a population is simply the arithmetic average of all the sample values, i.e.

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

*Remarks***:**

1. Notice that the sample mean $\overline{x}$ is a r.v., dependeing on the sample values selected from the population.
2. Hope that it is close to the *actual mean*: to be discussed later.....

Definition 5.2 Sample variance*:*
The sample variance of a population is defined in a similar manner as the variance of a random variable discussed in chapter 2, i.e. [1]

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

where $\overline{x}$ *is the sample mean defined above.*

Remarks:

1. Notice that s_x^2 is a *random variable* as sample mean $\overline{x}$ is a r.v.
2. It represents the measure of the spread of population around the sample mean
3. The square root of sample variance is called the *standard deviation*:

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Example 5.1

Find the sample mean of the following resistance values in Ω; 900, 1013, 939, 1062, 1017, 996, 970, 1079, 1065, and 1049.

Solution:

Applying the definition of the sample mean above, we easily get:

$$\overline{x} = 1009\,\Omega$$

Example 5.2

Obtain the sample standard deviation of the resistance samples given in the above example.

Solution:

Using thye definition od sample standard deviation, we obtain:

$$s_x = 58.7\Omega$$

[1]The reason for division by $n-1$ rather than n will bcome apparent in subsequent chapter when we consider estimation. Note, for a moment, that as n becomes large, the difference is small.

Note: As the # n of population gets large, computaions involved in the sample variance becomes tedious, due to the substraction procedure... [2]

Computationally efficient way of obtaining s_x^2:

Expanding the RHS of the sample variamce, we get

$$\begin{aligned} s_x^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2 - 2\overline{x}x_i + \overline{x}^2 \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2\overline{x}\sum_{i=1}^{n} x_i + n\overline{x}^2\right) \end{aligned}$$

applying the definition of the sample mean above:

$$\begin{aligned} s_x^2 &= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{2}{n}\left(\sum_{i=1}^{n} x_i\right)^2 + \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right] \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right] \\ &= \frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)} \end{aligned}$$

Remark: We only need the sums of the sample mean and the squares of the sample mean, without the *tedious* step of substraction...

Example 5.3

Redo the example 5-2 following the procedure mentioned above.

Solution:

We get: $\sum_{i=1}^{n} x_i^2 = 10,211,806$, $\sum_{i=1}^{n} x_i = 10,090$, and thus obtain $s_x^2 = 3,444$ and $s_x = 58.7\Omega$, which is the same result obtained previously.

[2] First, you calculate the sample mean, and then substract it from each sample before squaring and summing again.

Example 5.4

The *residuals* are defined as:

$$d_i = x_i - \overline{x}$$

Show that the sum of all rsiduals is equal to zero.

Solution:

From the definition of the sample mean, we have:

$$\sum_{i=1}^{n} x_i - n\overline{x} = 0$$

or

$$\sum_{i=1}^{n} (x_i - \overline{x}) = 0$$

which means

$$\sum_{i=1}^{n} d_i = 0$$

5.3 Regression Techniques

Suppose we are given measurements of a pair of data

$\Longrightarrow$ $\exists$ some relationship b/w the data [3]

$\Longrightarrow$ matching the data relation to a simple *straight line*...

Objective:

Given the measurements of data pairs, $\{(x_i, y_i), i = 1, 2, \ldots, n\}$, find a straight line which is *best* fit to the data:

$$y = \alpha x + \beta$$

Criterion:

Choose the constants α and β $\ni$: the straight line fit to the data is taken in the sense of *minimum squared error*:

$$(\alpha_o, \beta_o) = \underset{(\alpha,\beta)}{argmin}\, \epsilon$$

where subscript "o" stands for *optimum*, and the ϵ is the average squared error defined as follows:

$$\epsilon = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \alpha x_i - \beta)^2$$

[3] Precise determination of the relationship is not possible due to measurement errors...

Procedure:*(regression technique)*

To find (α_o, β_o) making ϵ as small as possible, we differentite ϵ w.r.t. α and β, and put to zero:

$$\frac{\partial \epsilon}{\partial \alpha} = -\frac{2}{n-1}\sum_{i=1}^{n}(y_i - \alpha_o x_i - \beta_o)x_i = 0$$

$$\frac{\partial \epsilon}{\partial \beta} = -\frac{2}{n-1}\sum_{i=1}^{n}(y_i - \alpha_o x_i - \beta_o) = 0$$

We can re-arrange the above equations into the followings:

$$\left(\sum_{i=1}^{n} x_i^2\right)\alpha_o + \left(\sum_{i=1}^{n} x_i\right)\beta_o = \sum_{i=1}^{n} x_i y_i$$

$$\left(\sum_{i=1}^{n} x_i\right)\alpha_o + n\beta_o = \sum_{i=1}^{n} y_i$$

$\Longrightarrow$ Solving for (α_o, β_o) yields:

$$\alpha_o = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n(n-1)s_x^2}$$

$$\beta_o = \overline{y} - \alpha_o \overline{x}$$

where s_x^2 represents the sample variance of the x values.

For further simplification, define the *sample covariance* c_{xy} and the *sample correlation coefficient* r_{xy} respectively as:

$$c_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \qquad \text{and} \qquad r_{xy} = \frac{c_{xy}}{s_x s_y}$$

where s_x and s_y represent the sample standard deviation of the data sets $\{x_i\}$ and $\{y_i\}$, respectively.

By expanding the product in c_{xy}, we have:

$$c_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n(n-1)}$$

And the constant α_o can simply be expressed as:

$$\alpha_o = \frac{c_{xy}}{s_x^2}$$

$\Longrightarrow$ The **regression line** which is best fit to the data becomes:

$$y = \alpha_o x + \beta_o = \frac{c_{xy}}{s_x^2}x + \overline{y} - \frac{c_{xy}}{s_x^2}\overline{x}$$

$\Longrightarrow$ The regression line can be arranged into:

$$y - \overline{y} = \frac{c_{xy}}{s_x^2}(x - \overline{x})$$

$\Longrightarrow$ Or, the regression line can be expressed as follows: [4]

$$\frac{y - \overline{y}}{s_y} = r_{xy}\frac{x - \overline{x}}{s_x}$$

[4]Note that this is a somewaht easier expression to remember...

Remarks:

1. If $r_{xy} = 0$ the regression line vanishes, and the data set are called *uncorrelated*.
2. If $r_{xy} = \pm 1$, the data are *linearly* related as follows:
3. The *goodness* of the fit is determined by the squared error ϵ: even though it has been minimized, the regression line might still not well fit to data, especially when the correlation b/w/ data is small.

$$y_i = mx_i + b$$

Assignment: Prove the followings:

(a) If the data sets are linealy related, $r_{xy} = \pm 1$.

(b) The magnitude of r_{xy} is no greater than 1, i.e., $-1 \leq r_{xy} \leq 1$.

Example 5.5

Find the correaltion coefficient and the regression line for the following data set:

x_i	0.68	0.72	1.27	2.01	2.63	3,06	3.15	4.00	4.03	4.50
y_i	12.45	9.93	6.64	10.14	8.93	13.34	11.56	16.72	19.62	15.03

Solution:

From the data set, we can get: $r_{xy} = 0.71$, $\overline{x} = 2.6$, $\overline{y} = 12.44$, and $s_x = 1.39$, $s_y = 3.88$. Thus, the regression line becomes:

$$\frac{y - 12.44}{3.88} = 0.71\frac{x - 2.6}{1.39}$$

which is in Figure5-1.

Figure 5.1: Data pairs and regression line for the random data set of Example5-5.

5.4 Empirical Distribution Functions

Consider a r.v. X with distribution function $F_X(x)$, which is *unknown*. [5] We define the *empirical cumulative distribution function* [6] as follows:

Definition 5.3 Empirical Distribution Function*:*
We have a number of independent samples of a r.v. X, denoted $\{x_i, i = 1, 2, \cdots, n\}$. Then the empirical distribution function of X is defined as:

$$\tilde{F}_X(x|x_1, x_2, \cdots, x_n) = \frac{\text{number of samples } x_1, x_2, \cdots, x_n \text{ no greater than } x}{n}$$

Example 5.6

Obtain the empirical distribution of the resistance samples given in Example5-1.

Solution:

Arranging the samples in ascending order, which are 900, 939, 970, 996, 1013, 1017, 1049, 1062, 1065 and 1079, the empirical distribution can easily be obtained via the above definition. The result is plotted in Figure5-1.

Figure 5.2: The empirical distribution function for resistance samples in Example5-1.

[5] In chapter 3 when we theoretically discussed r.v.'s, we assumed a certain type of distribution functions such as Gaussian, exponential etc., even though are are not absolutely sure of it...

[6] Or simply *empirical distribution*.

Note:

1. The empirical distribution function has the same properties of a cdf. (*check!*)
2. The empirical probability mass function, which vorresponds to the pdf, can easily be obtaind from the distribution funcrion. (*how?*)

Simplified way of computing empirical distribution:

When the # of datum is large, we follow the following steps:

(1) We divide the data range into a convenient # of intervals of equal length.

(2) WE plot a *histogram* counting the number of data within each cell(or interval).

(3) The # of data within each cell is *cumulatively* summed to get empirical distribution.

Example 5.7

The intervals b/w telephone calls arriving at a certain switching office are recorded in minutes, which are: 0.026, 0.977, 0.05, 0.183, 0.597, 0.426, 1.327, 0.017, 0.191, 0.938, 0.065, 0.098, 0.271, 0.827, 0.863, 0.101, 0.372, 0.93, 0.343, 0.156, 0.451, 0.637, 0.282, 0.191, 0.14, 0.163, 0.372, 1.048, 0.5, 0.09, 1.675, 0.33, 0.206, 0.426, 1.128, 0.026, 0.041, 0.299, 0.531, 0.376, 0.49, 0.083, 0.575, 0.393, 0.651, 0.009, 0.606, 0.151, 0.283 and 0.815. Find the histogram and the empirical cdf of these time intervals.

Solution:

Following the steps described above, the result are plotted in Figure5-3. Note that the histogram appears to be roughly exponentially decreasing...

Figure 5.3: Empirical distributions for Example5-7: (a) histogram; (b) empirical cdf.

Remarks:

1. Recall the definition of the pdf in chapter 3:

$$\begin{aligned} f_X(x)\Delta x &= P(x < X \leq x + \Delta x) \\ &\simeq \frac{\text{number of data values in } (x, x+\Delta x)}{\text{total number of data values}}, \quad \Delta x \ll 1 \end{aligned}$$

   Based on this, we can figure out that the histogram obtained in above example does not match the pdf for two reasons:

   (a) The histogram is not normalized by the total # of data.

   (b) As is obvious in the above equation,, i.e. to get an approximation to the pdf, we must divide through by Δx

   With these two corrections, replotted hitogram is in Figure5-4, along with the plot of the following function:

$$\tilde{f}_X(x) = 2e^{-2x}u(x)$$

   (cf) The agreement is surprisingly good [7]...

[7] The data values in Example5-7 have been generated via a random number generator producing exponentilly distributed random variables.

Figure 5.4: Normalized histogram to approximate the theoretical pdf.

2. Notice that there $\exists$ a trade-off b/w # of bins versus the apprearance of the histogram:

   (a) Too many bins, due to too few samples per bin, causes statistical irregularity, and thus make the histogram ragged-looking.

   (b) Too few bins causes a low resolution on the abscissa of the histogram.

   (cf) Usually rely on trial and error process to select the suitable # of bins[8]

[8]Especially in the case of relatively few data samples, as in the case of these examples.....

5.5 More on Monte Carlo Simulation

Monte Carlo Simulation

Complex computer simulations of systems undergoing *random purturbations*, using random number generator...

Example 5.8

Consider an RC circuit with random resistance and capacitance [9], of which the 3-dB cutoff frequency of an RC filter is defined as:

$$f_3 = \frac{1}{2\pi RC}$$

Analyze the possible values of f_3.

Solution:

Suppose the nominal value of R is $1000(\Omega)$ w/ $\pm 10\%$ tolerance, i.e. $R \sim U[900, 1100]$, whereas the capacitance has a nominal value of $1(\mu F)$ w/ $\pm 5\%$ tolerance, i.e. $C \sim U[0.95, 1.05]$.

(1) Conventional method:

Without any computer simulations, all we can deduce from the given data are the nominal value, maximum value, and the minimum value of f_3, i.e.:

(i) nominal f_3:

$$f_{3,\,\text{nom}} = \frac{1}{2\pi \times 10^3 \times 10^{-6}} = \frac{500}{\pi} \approx 159(\text{Hz})$$

[9] This is due to the uncertain manufacturing process...

(ii) maximum f_3:

$$f_{3,\,\max} = \frac{1}{2\pi \times 900 \times 0.95 \times 10^{-6}} \approx 186(\text{Hz})$$

(i) minimum f_3:

$$f_{3,\,\min} = \frac{1}{2\pi \times 1100 \times 1.05 \times 10^{-6}} \approx 137(\text{Hz})$$

Therefore, all we can say about f_3 is that *it would have values of somewhere in between* 127(Hz) and 186(Hz), with its nominal value of 159(Hz).

(2) Monte Carlo simulation:

We generate, say, 5000 values of R and C which are uniformly distributed within their allowed range about the nominal values, and compute f_3 for each pair of R and C (usually using high level programming languages ∋: C^{++}, FORTRAN etc., or utilizing mathematics package ∋: MATLAB), and then plot the histogram of the cutoff frequency... (figure5-5)

Figure 5.5: Histograms for (a) resistance, (b) capacitance, and (c) cutoff frequency.

Table5.1 MATLAB program for generating histogram.

With this method, we can generate a histogram for f_3 to determine the influence of component variation on cutoff frequency

$\Longrightarrow$ This gives *more information* that the above conventional method [10]

$\Longrightarrow$ (e.g.) *estimate* of the most likely value of f_3 at which the histogram is its peak.

Additional use of data from MC simulations:

(a) If the RC circuit is part of a larger system, we can find the maximum or minimum values of the cutoff frequencey, and carry out a worst-case design [11].

(b) Find the approximate probability that f_3 lies in a specified region [12], for example, from figure5-5(c):

$$P\left(140 \leq f_3 \leq 150\right) \approx \frac{900}{5000} = 0.18$$

whereas

$$P\left(150 \leq f_3 \leq 160\right) \approx \frac{1600}{5000} = 0.32$$

(c) Design a more suitable system, in this case, the only apparent solution to provide a more precise circuit is to tighten the tolerances on R and C.

[10] We call this conventional method a *extreme value analysis*.

[11] This is the case with more complex functional relationships or probabilitic models for the component values... Note that, in a simple model as in the above example, we can easily figure out the extreme values w/o MC simulation.

[12] Note that, for this purpose, the empirical cdf would be more accurate...

5.6 Statistical Process Control

In a manufacturing process, there $\exists$ several steps that take place

♠ Monitor and determine *when* things have gone worng...

♠ Close down the process and look for the offending steps

: **Control Chart**

Illustration

Suppose we are manufacturing transistors, the *gain* of which is of our concern

$\Longrightarrow$ measure the gain, devide the emasurements into lots

$\Longrightarrow$ compute the sample mean of each lot: $m_i,\ i = 1, 2, \ldots, N$

$\Longrightarrow$ compute the sample mean of the sample means[13] : m

$\Longrightarrow$ set the upper and lower limits as ± 3 sample std of the m_i's from m

$\Longrightarrow$ check if any m_i falls outside the control limit.....

$\Longrightarrow$ decide whether the manufacturing process has a problem or not

[13] i.e. sample mean of entire samples!

Example 5.9

In a transistor manufacturing process, for 25 lots of size 5, the sample means of the current gains are measured and plotted in figure5-6 by the $\times$'s. The sample mean of the entire $5 \times 25 = 125$ transistor gains is 98.8, and the sample standard deviation of the lot's sample mean is 6.95. We, then, set the upper and lower control limits as:

$$\text{UCL} = 98.81 + 3 \times 6.95 = 119.65$$

$$\text{LCL} = 98.81 - 3 \times 6.95 = 77.96$$

Solution:

The corresponding control chart, i.e. the sample means of the 25 lots along w/ the UCL and LCL, is in figure 5-6.

Note that for this particular process, we are well within the contol limits. [14]

The MATLAB program producing the control chart is in Table5-2.

Figure 5.6: Control chart for transistor gains: 25 lots of 5 transistor each.

Table5.1 MATLAB program for generating histogram.

[14] If the sample mean of any lot exceeds or goes below the UCL or LCL, respectively, the process is said to be *out of control*, and we seek for the cause of the excursion, and correct it!

5.7 Convergence of the Sample Mean to the Mean

Recall: The sample mean $\overline{x}$ of a population, given below, is an *estimate* of the true mean μ_X:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{(r.v.)}$$

Question: How good this *estimate* is?

Answer: We get a bound by applying the Chebyshev's inequality [15]...

$$P(|X - \mu_X| \geq k\sigma_X) \leq \frac{1}{k^2}$$

To apply the Chebyshev's inequality, we need the (1) mean and (2) variance of the sample mean $\overline{x}$:

(1) mean[16]:

$$E(\overline{x}) = \mu_X$$

[15] We replace X with $\overline{x}$.
[16] This is shown in problem 4-24a.

(2) variance:

$$\begin{aligned}
\mathrm{Var}(\overline{x}) &= E\{[\overline{x} - E(\overline{x})]^2\} \\
&= \left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n}\mu_X\right)^2\right] \\
&= \left\{\frac{1}{n^2}\left[\sum_{i=1}^{n}(X_i - \mu_X)\right]^2\right\} \\
&= \frac{1}{n^2}E\left[\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - \mu_X)(X_j - \mu_X)\right] \\
&= \frac{1}{n^2}\left\{\sum_{i=1}^{n}\sum_{j=1}^{n}E[(X_i - \mu_X)(X_j - \mu_X)]\right\} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}E[(X_i - \mu_X)^2] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{\mathrm{Var}(X_i)}{n} \\
&= \frac{\sigma_X^2}{n}
\end{aligned}$$

where σ_X^2 is the variance of each sample from the population.

Thus, the Chebyshev's inequality [17] becomes:

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} x_i - \mu_X\right| \geq \frac{k\sigma_X}{\sqrt{n}}\right) \leq \frac{1}{k^2}$$

[17] Unfortunately, it is usually true that we also do not know the variance σ_X^2 of the samples, so we cannot apply this Chebyshev's inequality in *real* estimation problems. We will discuss this subject in Chapter 6 again.

Example 5.10

Samples drawn from a population are known to have standard deviation of 2. We want the probability that the absolute value of the difference b/w their sample mean and the true mean is greater than 0.5 to be less than 1%.

How many samples should be drawn [18]?

Solution:

From the Chebyshev's inequality, we want:

$$\frac{1}{k^2} = 0.01 \quad \text{or} \quad k^2 = 100 \quad \text{or} \quad k = 10$$

and

$$\frac{k\sigma_X}{\sqrt{n}} = \frac{10 \times 2}{\sqrt{n}} = 0.5 \quad \text{or} \quad \sqrt{n} = \frac{10 \times 2}{0.5} = 40 \quad \text{or} \quad n = 1600$$

Note that this is a fairly large # of samples.

Remark: If we know something about the distribution of the sample mean $\overline{x}$, the required number of samples can be reduced.....

Recall[19] the *central limit theorem* which says that the sample mean is approximately Gaussian for a large number of samples, and then the LHS of Chebyshev's inequality becomes:

$$2\int_{k\sigma_X/\sqrt{n}}^{\infty} \frac{e^{-\nu^2/(2\sigma_X^2/n)}}{\sqrt{2\pi\sigma_X^2/n}} d\nu = 2Q(k) = 0.01$$

where the substitution $u = n^{1/2}\nu/\sigma_X$ has been used in the integrand to get the Q-function expression.

Using the table of Q-function, we find that $k = 2.57$, and since k must be an integer, we round it up to 3. Therefore, we have:

$$\frac{k\sigma_X}{\sqrt{n}} = 0.5 \quad \text{or} \quad \frac{3 \times 2}{\sqrt{n}} = 0.5 \quad \text{or} \quad \sqrt{n} = \frac{3 \times 2}{0.5} = 12 \quad \text{or} \quad n = 144$$

Notice that the # of samples is considerably lass than the previous case.....

[18] So that the sample mean $\overline{x}$, as an *estimate* of the true mean, maintains the precision given in the problem...

[19] In Chapter 4.