Weekly Review 1-2

This week I explained the idea behind the formula of the well-known statistic standard deviation so that it is clear now why it is a measure of dispersion and concluded our discussion on descriptive statistics by the Chebyshev theorem. Then we moved our focus to probability theory, in order to prepare ourselves for the study of statistical inference in the second half of this course; we illustrated some counting techniques for elementary probability calculations, introduced formally the mathematical definition of probability, and demonstrated how to derive further theoretical results from the definition by rigorous mathematics.

I explained that although the *range* (which in Statistics, unlike in everyday language, refers to the numerical value [one single value] telling us the difference between the maximum and the minimum, rather than an interval indicating from where to where) is a simple measure of dispersion, the *standard deviation*, which takes every datum into consideration, is a much more popular and useful measure of dispersion. Considering how concentrate the data are around their centre (the mean), I argued by our intuition to "establish" (step by step) the formula of the *variance*, which is the mean (to summarize) of the squared (to get rid of the sign) deviations from the mean (to see how close to the centre each datum is). Taking the square root of the variance will give us the standard deviation, which has the same unit of measurement as the original data.

In the calculation of standard deviation, we have to know whether the data themselves are already the *population* (the largest collection of entities for which we have an interest at a particular time) or only a *random sample* (a collection of entities from the population so that each entity has the same chance to be chosen). And what we are going to learn in the rest of the course (and the rest of the book) is the so-called *Statistical Inference*, where

> *Statistical inference is the procedure by which we reach a conclusion about a population on the basis of the information contained in a sample that has been drawn from that population.*

The very first thing about Statistical inference we learnt is to estimate *parameters* of a population by the corresponding *statistics* calculated from a sample. For example, if we have a very large population, whose mean and standard deviation are μ and σ, respectively, and we want to estimate these two parameters, what we have to do is to take a random sample of size n, say $\{x_1, x_2, \ldots, x_n\}$, from the population and then calculate the corresponding statistics, namely, the sample mean $\overline{x}$ and the sample standard deviation s by

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}, \qquad s = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}}.$$

The reason for using $n-1$ in the denominator of s^2 is that this will give a better estimate for σ^2 in the sense that when there are many researchers, if each researcher has his/her own random sample from the same population and obtain his/her own s^2, then the theoretical average (loosely speaking, the average of infinitely many) of all these s^2 is equal to the true σ^2. Technically speaking, we say s^2 is an *unbiased* estimator for σ^2. (However, s is not an unbiased estimator for σ. The technical details of such terms will be discussed not in this course but in a second semester course that is devoted to mathematical statistics.)

Even we now understand that the standard deviation is a measure of dispersion, what exactly do we know if say, $\mu = 50$ and $\sigma = 10$?

The beautiful Chebyshev's inequality (see p. 114) states an amazing result that

Theorem 1 (Chebyshev) *For any set of data and any constant $k > 1$, the proportion of the data that must lie within k standard deviation on either side of the mean is* <u>*at least*</u> $1 - 1/k^2$.

Therefore, if $\mu = 50$ and $\sigma = 10$, then we know e.g. that at least 75% of the data are between $\mu \pm 2\sigma = 50 \pm 20$.

Thus, only the mean and the standard deviation can already tell us a lot about the distribution of the data. Remind you that we, of course, could not know every detail of the given data from these two numbers, which serve merely as a summary. However, knowing that e.g. at least 75% of the data are not more than 2 standard deviations (and e.g. at least 89% not more than 3 standard deviations) away from the mean seems quite helpful in having an overall picture of the distribution of the data.

(Note that this course will not present any mathematical proofs. We are not learning proofs in this course; we are learning the applications of statistics, which do not require the rigorous proofs behind the theory. For students majoring in mathematics, it is a very interesting experience to learn how to accept mathematical results without seeing their proofs.)

To study statistical inference, we first have to have a proper understanding of *elementary probability theory*, which is the theme of Chapters 4–6, and that's why we shifted from statistics to probability after the discussion of descriptive statistics. As you will see very soon, the calculation of probability of an event is not always straightforward. One major difficulty is how to count the number of all possible outcomes and the number of individual outcomes comprising a particular event.

A very important technique in elementary probability is how to count the number of possible ways to choose r balls from n distinct balls (e.g. balls labelled by distinct numbers). If the order of the chosen balls is important, e.g. $\{1, 2, 3\}$, $\{2, 1, 3\}$, $\{3, 1, 2\}$, etc., are different

outcomes, the number of possible *permutations* (outcomes that the order matters) is

$$ {}_nP_r = \frac{n!}{(n-r)!}. $$

If the order of them is irrelevant, e.g. $\{1,2,3\}$ is the same as $\{2,1,3\}$, $\{3,1,2\}$, etc., then the total number of possible *combinations* (the order does not matter) is given by

$$ {}_nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}. $$

Using the first, the second and the third prize of our local lottery *Mark Six*, I illustrated how to count possible combinations systematically. For example, if we have to pick up r balls from n distinct balls in which a of them are red in colour (representing say the $a = 6$ drawn numbers in Mark Six) and the rest of them are green (representing the remaining 43 balls), the number of ways we can get x red balls, $0 \le x \le r$, when the order does not matter, is

$$ \binom{a}{x}\binom{n-a}{r-x}. $$

This idea can be easily extended to more than two colours. For example, if we have a red balls (again, representing say the $a = 6$ drawn numbers in Mark Six), b blue balls (representing the $b = 1$ extra drawn number) and $n-a-b$ green balls (representing the $c = 42$ untouched balls left in the original big pool), then the number of ways to get x red balls and y blue balls by chosen $r(\ge x+y)$ balls is

$$ \binom{a}{x}\binom{b}{y}\binom{n-a-b}{r-x-y}. $$

In some applications perhaps x or y is zero, or $x + y = r$; in these cases I recommend that you still write all terms down, including e.g. $\binom{a}{0}$ when $x = 0$, in the above expression so that you will not miss any colour by mistake. (And note the convention that $0! = 1$.)

In this lottery story we say we sample *without replacement*. We can make a totally different story if we consider sampling *with replacement*, meaning that the picked ball would be replaced by a new but identical ball (or we simply put the picked one back to the pool) so that the same number (the same ball) could appear more than once. Problems related to sampling with replacement will appear in Chapter 5.

We want to count the number of possible combinations because we want to calculate probability. But, wait, what actually is probability? The interpretation we may use comes from the *law of large numbers* (see p. 155), which says that

Theorem 3 (Law of Large Numbers) *If an experiment is repeated infinitely many times, the probability of an event is equal to the relative frequency of the occurrence of the event.*

An essential point in any probabilistic statement is that the experiment can be repeated. Thus, the statement "the probability that it will rain tomorrow is 0.3" is no longer a statement that we can understand, because "tomorrow" will appear only once and cannot repeat or be repeated. Such a *subjective probability*, however, is very commonly used in everyday life. We will give it a rigourous meaning when we move to Chapter 5.

There is no short-cut to learn how to count. In fact, there does not exist a general procedure for counting the number of individual outcomes comprising a particular event. As a result, counting is always challenging. The only way to master the technique is to have more practice. Questions in Assignment 1 and the next example class will give you more chance to practice the art of counting. The most important ability that you have to develop through these questions is to know how to count in a *systematic way*.

However, probability is not only about counting the number of possible combinations, because each outcome is not necessarily equally likely. For example, if we roll a die, a fair six-faced die will lead to a probability function that assigns 1/6 to each possible outcome. But as I showed you, dice are not necessarily fair. If a die is not a cube but a general cuboid (or even a U-shape object), then each outcome is not equally likely. More information on such dice can be found in

`http://www.riemer-koeln.de/joomla/index.php?option=com_wrapper&view=wrapper&Itemid=56`

Sprechen Sie Deutsch? (= Do you speak German?) If you don't, then before you learn it, you may read the following article written in English:

`http://www.riemer-koeln.de/mathematik/quader/cuboid.metrika.pdf`

which reports a scientific study of such dice.

Now, consider a question from the reverse direction: instead of calculating the probabilities of the outcomes of a unfair die, we consider the feasibility of making a die that meets our pre-specified probabilities of the outcomes. For example, can we make a loaded die such that each even number has a chance of 1/4 and each odd number has 1/12? More generally, for some arbitrarily assigned probability values, can we manufacture a die that has the assigned probabilities (as the proportions of outcomes in the long run)?

To answer this question, we have to introduce some more mathematical terms. Denote by Ω (or, as used in elementary textbooks, S) the *sample space* of an *experiment*, meaning that Ω is the collection of all possible *outcomes* of the experiment. An *event* A is a *subset* of Ω, denoted by $A \subset \Omega$, i.e. A is a collection of some possible outcomes in Ω. A function $\Pr(\cdot)$ defined for all *events* is called a *probability function*, or simply *probability*, if it satisfies

1. $\Pr(A) \geq 0$ for all events A;
2. $\Pr(\Omega) = 1$;

3. for mutually exclusive events A_1, A_2, ...,

$$\Pr(A_1 \cup A_2 \cup \cdots) = \Pr(A_1) + \Pr(A_2) + \cdots ,$$

where A_i and A_j are *mutually exclusive* if they have nothing in common, i.e. $A_i \cap A_j = \emptyset$, in which the symbol $\emptyset$ denotes the empty set, i.e. a set containing nothing. (The symbols $\cup$ and $\cap$ represent *or* and *and*, respectively, also known as *union* and *intersection*.)

From mathematical point of view, whenever you give me a probability function that satisfies the above three conditions, it is possible to have a physical die whose outcomes follow your probability function. (How to make this die is an engineering problem and hence is none of our business.) If you give me a function that does not satisfy the three conditions, no one can make a die that have such probabilities for the outcomes.

Using simple mathematical argument, we derived the following four properties of a probability function:

(i) $\Pr(A) \leq 1$ for all events A;

(ii) $\Pr(A^{\mathrm{c}}) = 1 - \Pr(A)$, where A^{c} denotes the complement of A (note that the event A does not happen if and only if A^{c} happens, and so $\Pr(A^{\mathrm{c}})$ is the probability that A does not happen);

(iii) $\Pr(\emptyset) = 0$;

(iv) the *General Addition Rule* given on p. 181,

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B).$$

The General Addition Rule is quite easy to believe if we draw a *Venn diagram*. However, Venn diagrams cannot replace rigorous mathematical proofs for probability statements. We proved it mathematically by a simple technique, namely, split an event into two (or more) mutually exclusive (i.e. disjoint) parts and then apply the third condition of a probability. The General Addition Rule of course can be generalised to the case for the union of k events. The statement can be easily written down by using Venn diagrams and then proved by mathematical induction (if you know what mathematical induction is, then good; but if you do not, it does not matter for our course).

The reason for considering the 'weird' event $\emptyset$ in (iii) above is that when we have two mutually exclusive events, say A and B, the probability that both of them happen is of course zero (by the definition of mutually exclusive events), i.e. $\Pr(A \cap B) = 0$; mathematically, the fact that A and B are mutually exclusive is expressed as $A \cap B = \emptyset$, and so it is intuitively obvious that a sensible definition of probability must result in $\Pr(\emptyset) = 0$.

Please be reminded that the purpose of the above discussion is to understand probability by mathematically rigourous arguments instead of by intuitive arguments (nevertheless, the mathematical results should agree with our intuition). Probability theory, as a branch of mathematics, starts with definitions (sample space and events) and axioms (the three conditions). Then theorems (such as the four properties presented in this and the last review) are derived from definitions and axioms by mathematical arguments. However, this course is not supposed to be too mathematical and so we only illustrate the ideas but we will not really do the theorem–proof stuffs.

Then, we introduced the notion of *conditional probability*, which is as follows. Suppose B has happened or definitely will happen, and we are interested in the occurrence of event A, then *the probability of A, given B* is

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}, \qquad \text{if } \Pr(B) \neq 0.$$

However, sometimes it may be more helpful to express the relationship as

$$\Pr(A \cap B) = \Pr(A|B) \Pr(B),$$

which can be understood as follows: the chance that both A and B will happen can be split into two probabilities; one is the chance of A if B happens and the other is the chance that B really will happen. This relationship is still true even if $\Pr(B) = 0$.

For some experiments, event B has nothing to do with event A and vice versa, and so $\Pr(A|B) = \Pr(A)$ and, equivalently, $\Pr(B|A) = \Pr(B)$. For such a pair of events, we say they are *independent*. Note that for events A and B with non-zero probabilities, i.e. $\Pr(A) > 0$ and $\Pr(B) > 0$, if they are independent, they are not mutually exclusive and if they are mutually exclusive, they are not independent. That is to say, for two events with non-zero probabilities, the event that "they are independent" and the event that "they are mutually exclusive" are mutually exclusive! (But be careful: there are events A and B that are neither independent nor mutually exclusive.)

When will two events be independent? If we are talking about the outcomes of two experiments, e.g. rolling a die twice (or rolling two dice), then independence is often an *assumption*. If we are talking about two events of one experiment, then we can prove or disprove the independence. For example, suppose we roll a fair die and consider A is the event that the outcome is 1, 2, 4 or 6, and B is the event that the outcome is 1, 3 or 4. Then because $\Pr(A) = 2/3$, $\Pr(B) = 1/2$ and $\Pr(A \cap B) = 1/3$, we have $\Pr(A \cap B) = \Pr(A) \Pr(B)$, and so A and B are by definition independent. No intuitive explanation is available: they are just happened to be independent!

Next week we derive some further properties of a probability function, introduce and discuss some interesting applications of the notion of *conditional probabilities*, and then do some gambling (without money...). Afterwards, we will move to Chapter 5 for more

advanced topics in probability, which can be viewed as abstractions and generalisations of some concepts and notions in gambling.

Enjoy your Assignment 1.

Cheers,
Heng