BIOSTATISTICS
TOPIC 4: SAMPLING DISTRIBUTION I
DISCRETE DISTRIBUTIONS

Most of the phenomena that we observe every day can be measured. The measurements can be roughly classified into two groups: one related to observations which can be assessed quantitatively and another related to observations which can not be quantified but may be assessed qualitatively. For example, blood pressure (mmHg), bone mineral density (g/cm^2), wave velocity, etc. can be classified as quantitative measurements since they are defined by a certain physical scale. In statistics, we refer to these as **continuous** measurements. On the other hand, we used qualitative values such as male or female to describe sex of an animal, rich or poor to describe wealthy status, death or survival to indicate an ultimate event etc. These are referred to as **discrete** measurements. The set that takes on these values is called **variable**. Thus, we have continuous variable and discrete variable.

Each of these variables have their own characteristics of distribution. Characteristics here refers to the mean, variance, value ranges, shape distribution etc. which we have leaned briefly in topic 2 under the heading of descriptive statistics. We will specifically discuss the discrete distribution in this topic first. Characteristics of continuous distribution is the subject of a next topic.

I. CHARACTERISTICS OF RANDOM VARIABLES

(A) RANDOM VARIABLE

Example 1: Let us consider the possible number of heads (H) that can appear when three balanced coins are tossed once. The eight possible outcomes of this experiment, together with the number of heads associated with each outcome, are listed below:

Outcome	No. of heads
TTT	0
HTT	1
THT	1

TTH	1
THH	2
HTH	2
HHT	2
HHH	3

We have assigned a value (0, 1, 2 or 3) to each of the eight possible outcomes or sample points. Now, let the variable *X* represents the number of heads in this experiment. Depending on the outcome of the experiment, *X* can assume any of the values 0, 1, 2 or 3. When the numerical value of a variable is determined by the outcome of an experiment, the variable is called a **random variable**.

DEFINITION: *A random variable is a variable whose numerical value is determined by the outcome of a random trial.*

It is conventional in statistics to distinguish between a random variable and the possible values it can assume. We shall use an upper case letter, such as *X*, to denote the random variable and the corresponding lower case letter, *x* in this case, to denote a particular value assumed by *X*.

A random variable that takes on distinct values only is called a discrete variable.

Some discrete variables are treated as if they can assume one of an infinity of possible distinct values. For instance, one may treat the number of traffic violations in a city as having possible values 0, 1, 2, ..., ad infinitum. While realistically there is upper limit to the number of violations, it is often useful to model the number of possible outcomes as potentially infinite.

A random variable that may take on any value on a continuum is called a continuous variable. For example, the bone mineral content may be any value between 0.3 to 1.9. Of course, any measurement can be made only to a finite number of significant digits and so, strictly speaking, measured BMD is a discrete random variable. However, it is conceptually advantageous to view measured BMD as being inherently continuous. Additional examples of measured variables that are often treated as continuous are blood pressures, weight and height etc.

(B) PROBABILITY DISTRIBUTION:

DEFINITION: *The probability distribution of a random discrete variable X associates with each of the distinct outcomes x_i (i = 1, 2, . . ., k) a probability P(X = x_i).*

In example 1, the probability that 1 head appears is P(X = 1) = 3/8, probability that 2 heads appear is P(X = 2) = 3/8, probability that 2 heads appear is P(X = 3) = 1/8 and probability that no head appear is P(X = 0) = 1/8. The sum of these probabilities is equal to 1. Two general properties of probabilities can be deduced from these observations: (i) the probability of an event ranges between 0 and 1 and (ii) the sum of all probability values is 1. That is:

(a) $0 \le P(X = x_i) \le 1, \quad i = 1, 2, 3, \ldots, k.$

(b) $\sum_{i=1}^{k} P(X = x_i) = 1$

(C) EXPECTED VALUE OF DISCRETE RANDOM VARIABLE

Consider example 1 again. Let us assume that the three coins are to be tossed an infinite number of times. Although the number of heads that can appear in any of the trials is 0, 1, 2, or 3; in this infinite number of trials we expect to obtain an average of 1.5 heads per toss. This long run average of 1.5 heads is called the **mathematical expectation** or **expected value**. In every day language, expected value is actually the mean.

Although we derive the mathematical expectation in an intuitive fashion, there is a systematic procedure for determining it, and this expected value is given by:

$$\boxed{E(X) = \sum_{i=1}^{k} x_i P(x_i)} \qquad [1]$$

In the example: $E(X) = 0(1/8) + 1(3/8) + 2(3/8) + 3(1/8)$

$= 1.5.$ //

(D) VARIANCE OF DISCRETE RANDOM VARIABLE

Since the outcomes of a random variable are probabilistic, it is useful to have a measure of the dispersion or variability of the outcomes. A key measure is the variance of a random variable which is defined as:

$$\boxed{\mathrm{var}(X) = \sum_{i=1}^{k} [x_i - E(X)]^2 P(x_i)} \qquad [2]$$

It can be equivalently (but somewhat simpler) written as:

$$\boxed{\mathrm{var}(X) = E\left[(X - E(X))^2\right] = E(X^2) - [E(X)]^2}$$

In our example 1, the variance of *X*, the number of heads from the experiment, is calculated as follows:

$$\begin{aligned}\mathrm{var}(X) &= (0-1.5)^2\left(\frac{1}{8}\right) + (1-1.5)^2\left(\frac{3}{8}\right) + (2-1.5)^2\left(\frac{3}{8}\right) + (3-1.5)^2\left(\frac{1}{8}\right) \\ &= 0.75\end{aligned}$$

and $\mathrm{SD}(X) = \sqrt{0.75} = 0.86$ //

(SD = standard deviation)

(E) Expected Value and Variance of Two Discrete Random Variables

If we have two random variables, *X* and *Y*, and assume that these two variable are independent, then the following relations are true:

$$\boxed{E(X+Y) = E(X) + E(Y)}$$

$$\boxed{E(X-Y) = E(X) - E(Y)}$$

and

$$\boxed{\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y)}$$

$$\boxed{\mathrm{var}(X-Y) = \mathrm{var}(X) + \mathrm{var}(Y)}$$

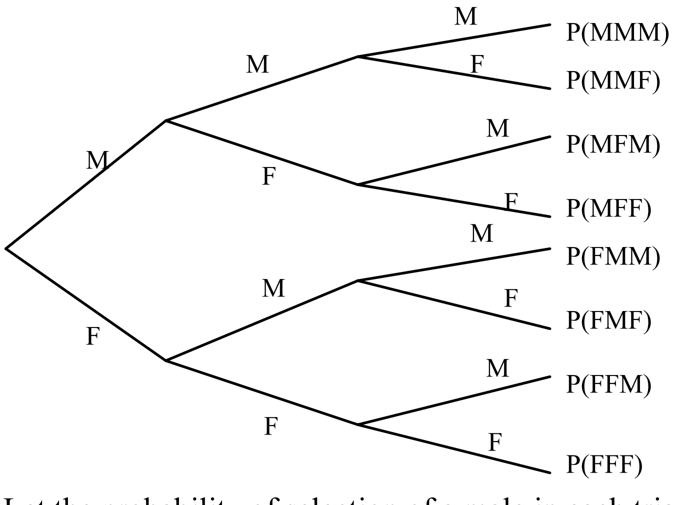
That is, the expected value of the sum of two **independent** random variables is simply the sum of the expected values of each of the two variables, and similarly for the variance. However, the expected value of the difference of two independent random variables is simply the difference of the expected values of the two random variables, but the variance of the difference is the sum (not difference) of variances of each of the two variables

II. THE BINOMIAL DISTRIBUTION

In the above section, we examined the distribution of a die experiment. There are many similar distributions in real life. We will survey some of the important ones. One of these important distributions is called the Binomial and Poisson distributions.

Example 2: Consider a selection of three consecutive persons. Each person can be classified as male (M) or female (F). Find the probability of 0, 1, 2, 3 males in these selections.

Let us represent the outcomes of these selections by a tree diagram:



Let the probability of selection of a male in each trial be p and the probability of selection of a female be $q = (1 - p)$. Then, by the above diagram, we have:

$$P(MMM) = p^3$$
$$P(MMF) = p^2q$$
$$P(MFM) = p^2q$$

P(MFF) = pq^2
P(FMM) = p^2q
P(FMF) = pq^2
P(FFM) = pq^2
P(FFF) = q^3

If X represents the number of males in the selections, we can form table as follows:

X	P(X)
0	q^3
1	$3pq^2$
2	$3p^2q$
3	p^3

In fact, this distribution can be written in a binomial formula as follows:

$$(p+q)^3 = p^3 + 3p^2q + 3pq^2 + q^3$$

The terms in this expansion give the respective probabilities of exactly 0, 1, 2, and 3 males in the 3 selections.

THEOREM: *If an experiment consists of n independent binomial trials, each with probability p of success and probability q = (1 - p) of failure, then the probability that the experiment results in exactly x successes and n-x failures is:*
$\boxed{B(x; n, p) = C_x^n p^x q^{n-x}}$ where $x = 0, 1, 2, \ldots, n$

Thus, for a given probability *p*, *x* and *n*, the binomial probability distribution is completely determined.

<u>Example 3</u>:

(i) The probability of a disease being cured is 0.6. If a doctor treats 10 patients, what is the probability that exactly (a) 8 patients, (b) 0, (c) 10 patients are cured.

In this case p = 0.6, according to the theorem, the probability that exactly 8 patients are cured is given by:

$$\text{P(8; 10, 0.6)} = C_6^{10}(0.6)^6(1-0.6)^4 =$$

similarly: $\quad$ P(0; 10, 0.6) = $C_0^{10}(0.6)^0(1-0.6)^{10}$ = 0.0001048

and $\quad$ P(10; 10, 0.6) = $C_{10}^{10}(0.6)^{10}(1-0.6)^0$ = 0.006046.

(ii) In a family of 4 children, what is the probability that there will be exactly two boys ?

In this case, there are two possible outcomes - boy or girl; each with equal probability of 0.5. So, by using the Binomial theorem, the required probability is
$C_2^4(0.5)^2(1-0.5)^{4-2} = \frac{3}{8}$. //

Expected Value (Mean) and Variance of a Binomial Variable

By using [1] and [2], we can determine the expected value (mean) and variance of a binomial distribution:

$$E(X) = np \qquad [3]$$
$$\text{var}(X) = np(1 - p). \qquad [4]$$

Applications of The Binomial Distribution

Suppose that we survey *n* persons in a certain district, and found that *x* of them have fractures. The questions are: (i) for any individual person we will study in another similar community, what is the average chance (probability) that he/she will have fracture, and (ii) what is the variability of this probability ?

Obviously, the estimated average probability of fracture (let us denote by *p*) is $p = \frac{x}{n}$.

So:

(a) for any ith individual, the expected probability of fracture and its variance are:

$$p_i = \frac{x}{n} = p \qquad [5]$$
$$\text{var}(p_i) = p_i(1-p_i) = p(1-p) \qquad [6]$$

(b) for *n* individuals, the expected probability of fracture and its variance are:

$$\boxed{\bar{p} = \frac{1}{n}\sum_{i=1}^{n} p_i = \frac{x}{n}} \qquad [7]$$

$$\boxed{\operatorname{var}(\bar{p}) = \frac{\bar{p}(1-\bar{p})}{n}} \qquad [8]$$

e.g. $SD(\bar{p}) = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$

Results in (b) are very important since we can use this to work out the confidence limit of any proportion.

Example 4: In a random survey, 1200 people were interviewed and assessed. Of this number, 360 subjects were found to be overweight. What is the likely true proportion of overweighted people in this community ?

It is easy to see that the proportion of overweighted subjects is 360/1200 = 0.3 (or 30%). By [7], the variance is $\frac{0.3(1-0.3)}{1200} = 0.000175$. The standard deviation is then: $\sqrt{0.000175} = 0.013$. The 95% confidence interval of the proportion of overweighted subjects is probably between $0.3 - 2(0.013) = 0.27$ to $0.3 + 2(0.013) = 0.326$ (between 27% to 32.6%).

III. The Poisson Distribution

The Poisson probability distribution (named after a great French mathematician, Simon Poisson) is usually applied in many problems involving rare events and occurred in a period of time. For example, the incidence of fractures in a certain period of times, the distribution of bacteria in a culture plate, radioactive counts per unit of time, the number of persons arriving at a hospital's emergency department, the number of typographical errors on a page etc.

Let us suppose that we know that there are, on average, λ fractures per year. This means that in a time interval of length t years, there will be λt fractures. But the actual number observed in one particular time interval may be any non-negative integers since fracture is a random event (let us accept this proposition for the time being!). We can not predict exactly how many fractures will occur in a time interval. What we can do is to predict the pattern of arrivals in a large number of such time intervals. It could be shown that the number of fractures in a time interval of length t is a Poisson random variable with parameter $\mu = \lambda t$.

Definition: *If x is the number of successes in a given space of time interval, the Poisson probability function is* $\boxed{p(x;\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}}$, *where λ is the mean of the distribution (average number of successes over a time interval), e = 2.7182818... (=* $\frac{1}{0!}+\frac{1}{1!}+\frac{1}{2!}+\frac{1}{3!}+$*......) and x = 0, 1, 2, 3, . . .* $\propto$.

Thus, it could be argued from this definition that the Poisson distribution is a special case of the Binomial distribution. In fact, we will show later that the two are related.

Expected Value (Mean) and Variance of a Binomial Variable

By using [1] and [2], we can determine the expected value (mean) and variance of a Poisson distribution as follows:

$$E(X) = \lambda \qquad [9]$$

$$\text{var}(X) = \lambda. \qquad [10]$$

Thus, for a Poisson distribution, the mean is equal to the variance.

Example 5: The average number of new patients entering a clinical trial is 2 per week. What is the probability that on any given week,
(a) exactly one; (b) no subject; (c) exactly 3
are entering the study

Let X be the event of subjects entering the study. By using the Poisson distribution definition we have for (a) P(X = 1) = $\frac{2^1 e^{-2}}{1!} = 0.27$; for (b) P($X$ = 0) = $\frac{2^0 e^{-2}}{0!} = 0.135$ and for (c) P(X = 3) = $\frac{2^3 e^{-2}}{3!} = 0.1804$ //

FITTING A POISSON DISTRIBUTION

Example 6: A set 100, 0.1 minute radiological counts from a single source were made. The observed frequencies are as follows:

Count	Observed Frequency
0	11
1	20
2	28
3	24
4	12
5	5
$\geq$6	0

Are these frequencies consistent with a Poisson distribution ?

If the data follow a Poisson distribution, then we would expect the observed frequencies to be similar to the expected frequencies under the Poisson assumption. Therefore, we need to determine the expected frequencies. To do this, we need to find the mean λ, and the expected frequencies can be calculated by $p(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$.

The mean counts for the data is: $\frac{(0\times 11)+(1\times 20)+(2\times 28)+(3\times 24)+...+(6\times 0)}{100}$ = 2.2 counts. We can estimated the probability for each count under the Poisson distribution assumption by the function $p(x;\lambda)=\frac{e^{-\lambda}\lambda^{x}}{x!}$. In this case, λ = 2.2 and x = 0, 1, 2, 3, . . ,6. For example, $P(x=0)=\frac{e^{-2.2}(2.2)^{0}}{0!}=0.1108$, $P(x=1)=\frac{e^{-2.2}(2.2)^{1}}{1!}=0.2437$, etc. The respective expected number of counts are then: 0.1108 × 100 = 11, 0.2437× 100 = 24.4, etc. The full tabulation of these calculation is given in the following table:

Count	$p(x;\lambda)=\frac{e^{-2.2}(2.2)^{x}}{x!}$	**Frequency**	
(x)		Expected	Observed
0	0.118	11.1	11
1	0.2438	24.4	20
2	0.2681	26.8	28
3	0.1966	19.7	24
4	0.1082	10.8	12
5	0.0476	4.8	5
$\geq$6	0.0174	1.7	0
Total	**1.0000**	**100.0**	**100**

As can be seen from the last two columns of this table, the observed frequencies agree pretty closely with the expected Poisson frequencies. It is therefore reasonable to conclude that the data are distributed according to the Poisson probability law.

IV. Relationships between The Poisson Distribution and Binomial Distribution

We remarked earlier that there is a relationship between a Binomial distribution and the Poisson distribution. In fact, the Poisson distribution can be regarded as a limits of binomial distribution, when the probability of an event of interest p is *very* small (close to 0) and as n gets very large in such a way that np remains constant. For instance, the probability *p* that a given entity enters the examined aliquot is very small, and *n* the population of entities is large and the assumption of throughout mixing without clumping or aggregation guarantees that the long term average number of entities per aliquot *np* will remain constant. These considerations suggest that *np* , the Binomial mean, should be equal to μ, the mean of the Poisson distribution. Now, the variance of the Binomial distribution is $np(1 - p) = \mu(1 - p)$; but as *p* approach 0, $(1 - p)$ approaches 1 and the variance of Poisson distribution approaches μ.

This limit was first proposed and proved by S. Poisson in the 1733..

V. EXERCISES

1. A doctor engaged in cancer research is interested in the proportion of treated patients surviving at least five years. In one Sydney hospital she found that of 150 patients treated, 45 were alive after 5 years.
(a) Estimate the proportion of treated cancer patients in the population surviving at least 5 years.
(b) Estimate the standard error of this estimate.
(c) What assumptions, if any, are needed for this estimation procedure to be valid?

2. The quoted figure for the 5-year mortality rate for a particular form of leukemia is 80%. In the hospital where you are a resident interested in malignant neoplasm research, of the last 5 cases with this form of leukemia 4 are cured and 1 died. Do you feel you should check to see if some new procedure was used on the patients or that they were special in some other way, or pass off the cures to chance.

3. Suppose that 60% of the voting population in a city, about to have a referendum on adding sodium fluoride the drinking water, favour fluoridation.
(a) A sample of 10 persons are interviewed. What is the probability that 5, 6, or 7 fluoridation?
(b) A sample of 100 persons are interviewed. What is the probability that between 55 and 65 inclusive favour fluoridation?

4. Under microscopic investigation, on the average 5 particular microorganisms are found on a one square centimetre untreated specimen, One such specimen was chemically treated. If it is assumed that the treatment was ineffective and if the Poisson distribution is used, what is the probability that (a) less than 3 (b) exactly 5 (c) more than 6 (d) 2 or 3 organisms.

5. A student has been told that the probability of obtaining a successful end point in a particular titration is 0.7. This student carries out 5 such titrations and obtains only one successful end point. Should he think of a career that does not involve chemistry ?

6. Let *x* be a random variable assuming the values of 1 and 0 for the presence and absence of amblyopia. Consider the results of ocular tests for 500 children:

Amblyopia present: 50

Amblyopia absence: 450.

(a) Using the grouped computing formula (Topic 2), where $x_i = 0$ or 1 and f_i represents the associated frequencies, find:

$$\mu = \frac{\sum_{i=1}^{2} f_i x_i}{\sum_{i=1}^{2} f_i} \text{ and } \sigma^2 = \frac{N \sum_{i=1}^{2} f_i x_i^2 - \left(\sum_{i=1}^{2} f_i x_i \right)^2}{N^2}.$$

(b) Let $P(x = 1) = p$, find p and compute μ and σ^2, using the binomial formulae ([7] and [8] in text). Verify that these agree with the values obtained in (a).

7. Suppose that a particular strain of staphlococcus produces a certain symptom in 2% of persons infected. At a church picnic 200 persons ate contaminated food and were infected with the organism. What is the probability that (a) 10 or fewer, (b) none, (c) more than 4, (d) exactly 4, were infected?

8. According to genetic theory, blood type MM, MN, and NN should occur in a very large populations with frequencies p^2, $2p(1 - p)$ and $(1 - p)^2$, where p is (unknown) gene frequency.

(a) Suppose that in a random sample size n from the population, there are *a, b,* and *c* of the three genotypes. Find an expression for p.

(b) The observed frequencies in a sample of size 100 were 32, 46 and 22, respectively. Find p and the expected frequencies under the model.

9. In cells changes in genetic (heredity) material occur which are called mutations. These may be spontaneous or induced by external agents. If mutations occur in the reproductive cells (gametes) then the offspring inherits the mutant genes. In human, the rate at which spontaneous mutation occur per gene is about 4 per 100,000 gametes. In the common bacterium *E. coli* a mutant variety is resistant to the drug streptomycin. In one experiment of 150 petri dishes were plated with 1,000,000 bacteria each. It was found that 98 petri dishes had no resistant colonies, 40 had one, 8 had two, 3 had three and 1 had four. The average number of mutants per million cells (bacteria) is therefore $\frac{(40 \times 1) + (8 \times 2) + (3 \times 3) + (1 \times 4)}{150} = 0.46$.

Use this average to calculate the expected number of mutants expected under the Poisson hypothesis. Is the distribution of the number of mutants consistent with a Poisson distribution ?

10. The probability that an individual with a rare disease will be cured is 1%.
(a) A random sample of 10 persons with the disease is selected; find the probability that 1 person is cured, using Binomial distribution theory.
(b) A random sample of 300 persons witch this disease is selected; find the probability that less than 4 persons are cured, using the Poisson approximation to the binomial distribution.

*11. THE NEGATIVE BINOMIAL (PASCAL) DISTRIBUTION. Suppose that an experiment consists of a number of patients, each of those can result in either success or failure. This experiment continues until *k* successes are noted. Let *p* and *q* be the probabilities of success and failure, respectively, for a given patient. The probability that exactly *k*+*x* patients are needed to observed *k* successes are given by: $C_{k-1}^{k+x-1} p^k q^x$.

Suppose that in a clinical trial of osteoporosis, we need to recruit 8 patients from a population with a osteoporotic prevalence of 30%. What is the probability that 18 patients are to be screened ?

*12. THE GEOMETRIC DISTRIBUTION. As a special case of the negative binomial distribution, suppose that an experiment continues until the first success is noted. Again, each patient can result in either success or failure with probabilities *p* and *q*. Show that the probability that *x* patients are required is $P(1st succes) = q^{x-1} p$.

*13. THE HYPERGEOMETRIC DISTRIBUTION. Suppose that a population consists of *n* patients with disease A and *m* patients with disease B. In an experiment if *k* patients are selected at random without replacement. Show that the probability of exactly *r* out of *k* patients selected are of disease A is given by: $\frac{C_r^m \times C_{k-r}^{n-m}}{C_k^n}$

*: Optional exercises.