Statistics and Their Distributions

Definition: A statistics is any quantity whose value can be calculated from sample data. Prior to obtaining data, there is uncertainty as to what value of any particular statistic will result. Therefore, a statistic is a random variable and will be denoted by an uppercase letter; a lower case letter is used to represent the calculated or observed value of the statistic.

Remarks: 1°/ Because of this uncertainty, before the data becomes available we view each observation as a random variable and denote the sample by $X_1, X_2, \dots, X_n$. (uppercase letters for random variables). The obtained values of the samples are denoted by $x_1, x_2, \dots, x_n$.

2°/ The <u>sample mean</u>, regarded as a statistic (before a sample has been selected or an experiment has been carried out), is denoted by $\overline{X}$; the calculated value of this statistic is $\overline{x}$.

3°/ Similarly, the <u>sample standard deviation</u>, thought of as a statistic, is denoted by S, and its computed value is s.

4°/ Any statistic, being a random variable, has a probability distribution. The <u>probability distribution of a statistic</u> is referred to as its <u>sampling distribution</u> to emphasize that it describes how the statistic varies in value accross all samples that might be selected.

5°/ The probability distribution of any particular statistics depends not only on the population distribution (normal, uniform, etc) and the sample size n but also on the method of sampling.

Random Samples

→ To describe a sampling method often encountered (at least approximately) in practice.

Definition: The rv's $X_1, X_2, \ldots, X_n$ are said to form a random sample of size n if

1°/ The X_i's are independent rv's.

2°/ Every X_i has the same probability distribution

<u>Remarks</u> 1°/ Conditions 1°/ and 2°/ say that the X_i's are <u>independent</u> and <u>identically distributed</u> (iid).

2°/ If sampling is either <u>with replacement</u> or from an infinite (conceptual) population, Conditions 1°/ and 2°/ are satisfied exactly.

3°/ If sampling is <u>without replacement</u>, yet the <u>sample size n</u> is much smaller than the <u>population size N</u> ($n/N \leq 5\%$ in practice), Conditions 1°/ and 2°/ are approximately satisfied.

4°/ The virtue of this sampling method is that the probability distribution of any statistic can be more easily obtained than for any other sampling methods.

Deriving the Sampling Distribution of a Statistic

Example 6.2. (There are relatively few different X values in the population).

- Population distribution

Table 6.1

x	40	45	50
$p(x)$	0.2	0.3	0.5

with $\mu = 46.5$

$\sigma^2 = 15.25$

- Let X_1 and X_2 constitute a random sample from the population distribution. (i.e, X_1 and X_2 are independent, each with the probability distribution shown in Table 6.1.

- Table 6.2: Outcomes, probabilities, and $\bar{X}$ and S^2,

x_1	x_2	$p(x_1, x_2) = p(x_1)\,p(x_2)$	$\bar{x}$	s^2
40	40	$0.04 = 0.2 \times 0.2$	40	0
40	45	$0.06 = 0.2 \times 0.3$	42.5	12.5
⋮	⋮	⋮	⋮	⋮

To obtain the probability distribution of $\bar{X}$, we must consider each possible value $\bar{x}$ and compute its probability.

- $\bar{x} = 40$ occurs only 1 time in the table with probability 0.04:

$$p_{\bar{X}}(40) = P(\bar{X} = 40) = 0.04$$

- $\bar{X} = 42.5$ occurs twice in the table with probability 0.06 and 0.06:

$$p_{\bar{X}}(42.5) = P(\bar{X} = 42.5) = 0.06 + 0.06 = 0.12$$

- $\bar{X} = 45$ occurs three times with probabilities 0.1, 0.09 and 0.1:

$$p_{\bar{X}}(45) = P(\bar{X} = 45) = 0.1 + 0.09 + 0.1 = 0.29$$

- etc. . .

The complete sampling distributions of $\bar{X}$ and S^2

$\bar{x}$	40	42.5	45	47.5	50
$p_{\bar{X}}(\bar{x})$	0.04	0.12	0.29	0.3	0.25

s^2	0	12.5	50
$p_{S^2}(s^2)$	.38 = 0.04 + 0.09 + 0.25	.42 = 0.06 + 0.06 + 0.15 + 0.15	.20 = 0.1 + 0.1

Then

$$\mu_{\bar{X}} = E(\bar{X}) = \sum \bar{x}\, p_{\bar{X}}(\bar{x}) = 40(0.04) + 42.5(0.12)$$
$$+ 45(0.29) + 47.5(0.3) + 50(0.25) = 46.5 = \mu.$$

$$\sigma^2_{\bar{X}} = V(\bar{X}) = \sum \bar{x}^2 \cdot p_{\bar{X}}(\bar{x}) - \mu^2_{\bar{X}}$$
$$= (40)^2(0.04) + (42.5)^2(0.12) + (45)^2(0.29) + (47.5)^2(0.3)$$
$$+ (50)^2(0.25) - (46.5)^2 = 7.625 = \frac{15.25}{2} = \frac{\sigma^2}{2}$$

$$\mu_{S^2} = E(S^2) = \sum s^2\, p_{S^2}(s^2) = (0)(.38) + (12.5)(0.42)$$
$$+ (50)(0.2) = 15.25 = \sigma^2$$

That is, the $\overline{X}$ sampling distribution is centered at the population mean μ, and the S^2 sampling distribution is centered at the population variance σ^2.