Review of probability distributions, sample statistics (infinite population)

Example: Hypothetical data on number of television sets owned per household

Number	0	1	2	3	4
Probability	.2	.4	.2	.1	.1

Here the number of sets owned is the **random variable**, and the set of probabilities above are its **probability distribution**. Some important characteristics of a probability distribution are its **expected value** (mean value), **variance**, and **standard deviation**.

Expected value of y: $\mu = E(y) = \Sigma\ y\ p(y) = 0\ p(0) + 1\ p(1) + 2\ p(2) + 3\ p(3) + 4\ p(4) = 0\ (.2) + 1\ (.4) + 2\ (.2) + 3\ (.1) + 4\ (.1) = 1.5$

Variance of y: $\sigma^2 = V(y) = \Sigma\ (\ y - E(y)\ )^2\ p(y) = (0 - 1.5)^2\ (.2) + (1 - 1.5)^2\ (.4) + (2 - 1.5)^2\ (.2) + (3 - 1.5)^2\ (.1) + (4 - 1.5)^2\ (.1) = 1.45$

The standard deviation of y is the square root of the variance, here SD(y) = sqrt(1.45) = 1.20 .

In **statistical studies**, we collect data from which we make **inferences** about unknown **population parameters**, such as the population mean and variance. For example, we use **sample statistics** such as the **mean**, **variance**, and **standard deviation**, to estimate the corresponding **population parameters**.

Example: A sample of four households have the following numbers of TVs: 2, 0, 1, 3 . The sample statistic values are:

For random samples from infinite populations, the expected value of the sample mean is the (true) population mean, and the variance of the sample mean equals the population variance divided by the sample size. Also, an unbiased estimate of the variance of the sample mean is the sample variance divided by the sample size.

Probability Sampling (finite population)

Suppose we visit a small town with four houses (denoted houses I, II, III, and IV) , and the number of TV's in the houses are: 1, 3, 4, and 4, respectively. This is a simple example of a **finite population** (the four houses), with a single measurement (the number of TV's). Suppose we consider all possible samples of size n=2 from this population of size N=4: (I, II), (I, III), (I, IV), (II, III), (II, IV), and (III, IV). In probability sampling, we assign a probability of drawing each possible sample. If we assign a probability of 1/6 of drawing each of the six samples above, then this is an example of a **simple random sample without replacement**. Many other types of sampling designs exist, and occasionally people draw samples with replacement, to mimic the process of sampling from an infinite population.

Given a sampling design such as that above, we can draw a sample, and calculate the sample mean as an **estimator** of the (unknown) population mean.

Since we know the true population in this example, we can compute the **sampling distribution** of the sample mean. The sampling distribution of a statistic is the distribution of different values that it can assume under some sampling plan. From this sampling distribution we can also learn about characteristics of the statistic such as **bias** and **mean squared error** (MSE).

Sampling distribution of the mean number of TVs in the small town:

Sample	y_i	Sample mean (ybar)	prob(ybar)
I, II	1, 3	2	1/6
I, III	1, 4	2.5	1/6
I, IV	1, 4	2.5	1/6
II, III	3, 4	3.5	1/6
II, IV	3, 4	3.5	1/6
III, IV	4, 4	4	1/6

The population mean of y is $\mu = \Sigma$ y p(y) = (1 +3 +4 +4)/4 = 3, and the population variance of y is $\sigma^2 = \Sigma$ (y - E(y))^2 p(y) = 3/2

Also, the mean of the sampling distribution of the sample mean is E(ybar) = Σ ybar p(ybar) = 3, so that the sample mean is unbiased as an estimator of the population mean.

Now, what is the variance of the sampling distribution of the mean, and how does it compare to the variance of the sample mean under infinite sampling?