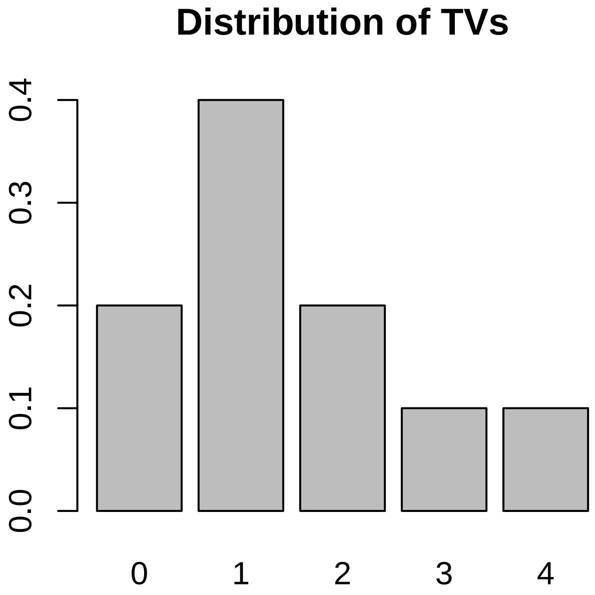
Probability review

Probability distributions, sample statistics with infinite and finite populations

Module 3

Infinite populations

Example of hypothetical data on number if television sets owned per household



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.*

tvs	**0**	**1**	**2**	**3**	**4**
$P(tvs)$	0.2	0.4	0.2	0.1	0.1

Expected value

Is a mean in which values of y do not all occur with equal probability; a weighted average. But still is a mean value, one measure of location.

Expected value of y (or x...whatever):

$$\mu = E(y) = \sum_{y} yp(y) = y_1P(y_1) + y_2P(y_2) + \ldots + y_nP(y_n)$$

For this example:

$$\mu = E(y) = 0(0.2) + 1(0.4) + 2(0.2) + 3(0.1) + 4(0.1) = 1.5$$

Variance

The variance is a measure of variation; it is the average *squared* distance each data point is from its mean. Its units of measurement are *squared* units.

Variance of y:

$$\sigma^2 = V(y) = \sum_{y} (y - \mu)^2 p(y) = (y_1 - \mu)^2 P(y_1) + (y_2 - \mu)^2 P(y_2) + \ldots + (y_n - \mu)^2 P(y_n)$$

For this example:

$$\sigma^2 = V(y) = (0 - 1.5)^2(0.2) + (1 - 1.5)^2(0.4) + (2 - 1.5)^2(0.2) + (3 - 1.5)^2(0.1) + (4 - 1.5)^2(0.1) = 1.45$$

Standard deviation

The standard deviation is a measure of variation; it is the average distance each data point is from its mean. It is just the square-root of the variance so that the units of measurement are the same as the mean.

$$\sigma = SD(y) = \sqrt{V(y)}$$

For this example:

$$\sigma = SD(y) = \sqrt{1.45} = 1.2$$

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:

sample mean

$$n = 4,\ N = 4,\ \{2, 0, 1, 3\}$$

$$\hat{\mu} = \overline{y} = \frac{\sum y_i}{n}$$

$$= \frac{2+0+1+3}{4} = 1.5$$

This sample mean was equal to the population mean but not all samples would yield the same mean.

sample variance

Estimator of true variance

$$\hat{\sigma}^2 = s^2 = \frac{\sum (y_i - \overline{y})^2}{n-1}$$

$$= \frac{(2-1.5)^2 + (0-1.5)^2 + (1-1.5)^2 + (3-1.5)^2}{4-1} = \frac{5}{3} \approx 1.67$$

sample standard deviation

Again, just take the square root of the variance.

$$s = \sqrt{s^2} = \sqrt{1.67} = 1.29$$

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).

Sample	y_i	$\overline{y}$	$P(\overline{y})$
I,II	1,3	2	$\frac{1}{6}$
I,III	1,4	2.5	$\frac{1}{6}$
I,IV	1,4	2.5	$\frac{1}{6}$
II,III	3,4	3.5	$\frac{1}{6}$
II,IV	3,4	3.5	$\frac{1}{6}$
III,IV	4,4	4	$\frac{1}{6}$

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