Sampling Distributions

Dr. N. N. Mahto, Visthapit College Balidih (Bokaro)

If we take a sample from a population over and over again. We will see that the means of the samples are normally distributed, regardless of the distribution of the original population. This is called the Central Limit Theorem and is the backbone of most of the statistical analysis we will perform in the future. Parameters and Statistics .

A parameter is a number that describes the population. In statistical practice, the value of a parameter is not known because we cannot examine the entire population.

A statistic is a number that can be computed from the sample data without making use of any unknown parameters.

In practice, we often use a statistic to estimate an unknown parameter.. The text emphasizes this with the comment: “Statistics come from Samples, and Parameters come from Populations.”

For example, the population mean (a parameter) is denoted μ and a sample mean (a statistic) is denoted x. . Statistics and Stooge Parameters.

A sample of size 30 is taken from the population of 190 Three Stooges films. In the sample, 13 of the films have Curly as the third stooge, 13 of the films have Shemp as the third stooge, and 4 of the films have Joe as the third stooge. For this sample, the percentage of Curly films is 100%× 13/30 = 43.33%. (1) Is this a parameter or a statistic? (2) What are the three statistics this sample yields and what are the three corresponding

parameters of the population? Statistical Estimation and the Law of Large Numbers Theorem. Draw observations at random from any population with finite mean μ. As the number of observations drawn increases, the mean x of the observed values is expected to get closer and closer to the mean μ of the population. This example illustrates the Law of Large Numbers. It describes a situation where there is a population with mean μ = 25. Samples of size 1 are taken and a running average from the samples is computed. Notice that the average of the samples gets close to the population mean of 25 as the number of samples n gets large: The BPS Applet Law of Large Numbers is a simulation of an experiment similar to that in the previous example. It is based on rolling a die a repeated number of times. Input 100 rolls and watch the running average approach . Now for a real subtlety. We are ready to consider two populations. One is a population from which we will sample and then use the statistics from these samples to estimate parameters of this population. The second population is the population of samples from the original population. We will see that the population of samples is normally distributed, regardless of the distribution of the original population.

sampling distribution.

Definition.

The sampling distribution of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population..

The Sampling Distribution of x Theorem.

Suppose that x is the mean of a simple random sample (SRS) of size n drawn from a large population with mean μ and standard deviation σ. Then the sampling distribution of x has mean μ and standard deviation σ/√ n.

An unbiased estimator of a population parameter is a statistic which is "correct on average" in many samples.

As illustrated above, x is an unbiased estimator of μ. . The fact that the sampling distribution of x has standard deviation σ/√ n implies that large samples will give better estimates of population parameters than small samples (since the sampling distributions are less spread out when n is large). In other words, if the population has a normal distribution N(μ, σ), then the sampling distribution will have the normal distribution N(μ, σ/√ n). The next section will liberate us from the assumption of a normal original population. The Central Limit Theorem. If the original population is not normally distributed, then it turns out that the sampling distribution will still be normally distributed. This is why the normal distribution is so important! The most important result for our use of statistics is the following theorem. Theorem. Central Limit Theorem. Draw an SRS of size n from any population with mean μ and finite standard deviation σ. When n is large, the sampling distribution of the sample mean x is approximately normal with distribution N(μ, σ/√ n). . Again, this is extremely important It justifies the use of the normal distributions when dealing with sample data. Informally, the "when n is large" comment means that for non-normal populations, we need large samples for the Central Limit Theorem to apply.

Example. There is a BPS Central Limit Theorem Applet. Access it and play with different sample sizes n and probabilities/proportions p to see how it affects the distribution.

Suppose the average number of slaps per film in the Three Stooges' films is $\mu = 12.95$ with a standard deviation of $\sigma = 4.50$. You want to estimate μ by taking a sample of Stooges' films. You only have time to watch 10 films. (1) What are the mean and standard deviation of the average number of slaps per film x in a sampling distribution for samples of size 10 films? (2) Use the Central Limit Theorem to find the probability that the average number of slaps per film in the sample of size 10 is less than 11 slaps.

Sample mean and variance:

both are important statistics that we can use to make predictions about a population. In this lesson, learn how to calculate these important values.

Tony owns a plant nursery, and one of his biggest sellers is blueberry bushes. He sells the bushes to his customers when they are at least 18 inches tall. Tony wants to know how long it will take each of his blueberry bushes to grow tall enough to sell. To get an estimate of this time, he selects ten plants at random and records the number of days each one takes to grow from a seed into an 18 inch tall plant.

Plant #	Days
1	92
2	103
3	99
4	108
5	86
6	94
7	90
8	102
9	97
10	96

Time each plant required to grow to 18 inches tall

Sample Mean:

A **sample** is a set of measurements taken from a larger population. In this case, the population would be all of Tony's blueberry bushes, and the sample would just include the specific

ten bushes he selected to observe. Tony's measurements represent a **random sample** because they were selected at random from the population. Each seed had an equal chance of being chosen for the sample. In order for a sample to give a good approximation of the population, it must be randomly selected.

The **sample mean** is simply the average of all the measurements in the sample. If the sample is random, then the sample mean can be used to estimate the population mean.

Sample mean equation

$$\underline{x = x1 + x2 + x3 + x4 + X5 + x6 + x7 + x8 + x9 + x10}$$

N

For Tony's data, the sample mean is:

$$\bar{x} = \frac{92 + 103 + 99 + 108 + 86 + 94 + 90 + 102 + 97 + 96}{10}$$

$$\boxed{\bar{x} = \frac{967}{10} = 96.7}$$

<u>Sample Variance:</u>

Another important statistic that can be calculated for a sample is the sample variance. **Variance** measures how spread out the data in a sample is. Two samples can have the same mean, but be distributed very differently. Variance is one way to quantify these differences. The variance of a sample is also closely related to the **standard deviation**, which is simply the square root of the variance. The symbol typically used to represent standard deviation is *s*, so the symbol for variance is s^2.

To find the sample variance, follow these steps:

- First, calculate the sample mean.
- Next, subtract the mean value from the value of each meas

  Square the resulting values.

- Add the results together to get the sum of squared deviations from the mean.
- Finally, divide this by the number of degrees of freedom, which is equal to the total number of measurements minus one (*n* -1)

In equation form, this looks like:

$$s^2 = \left(\frac{1}{n-1}\right)\sum_{i=1}^{n}(x_i - \bar{x})^2$$

The easiest way to do this is to make a table like this:

sample	mean	$(x - \bar{x})$	$(x - \bar{x})^2$
92	96.7	4.7	**22.09**
103	96.7	-6.3	**39.69**
99	96.7	-2.3	**5.29**
108	96.7	-11.3	**127.69**
86	96.7	10.7	**114.49**
94	96.7	2.7	**7.29**
90	96.7	6.7	**44.89**
102	96.7	-5.3	**28.09**
97	96.7	-0.3	**0.09**
96	96.7	0.7	**0.49**
		sum =	**390.1**
		n-1 =	**9**
		s^2=	**43.344**

For Tony's data, the sample variance is equal to 43.344.

Standard deviation often gives you more useful information than variance. About 70% of the values in the population are expected

to fall within one standard deviation on each side of the mean. Tofind the standard deviation from the variance, simply take the square root.

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

$$s = \sqrt{43.344} = 6.58$$

Since the mean number of days in Tony's sample was 96.7, he can expect about 70% of his trees to reach 18 inches tall between 90 days and 103 days.