Session 04

Sampling Distributions

Suppose that we draw all possible samples of size *n* from a given population. Suppose further that we compute a statistic (e.g., a mean, proportion, standard deviation) for each sample. The probability distribution of this statistic is called a **sampling distribution**.

Variability of a Sampling Distribution

The variability of a sampling distribution is measured by its variance or its standard deviation. The variability of a sampling distribution depends on three factors:

- N: The number of observations in the population.
- n: The number of observations in the sample.
- The way that the random sample is chosen.

If the population size is much larger than the sample size, then the sampling distribution has roughly the same sampling error, whether we sample with or without replacement. On the other hand, if the sample represents a significant fraction (say, 1/10) of the population size, the sampling error will be noticeably smaller, when we sample without replacement.

Central Limit Theorem

The *Central Limit Theorem* states that for sufficiently large sample sizes ($n \geq 30$), regardless of the shape of the population distribution, if samples of size *n* are randomly drawn from a population that has a mean
μ and a standard deviation σ, the samples' means X are approximately normally distributed. If the populations are normally distributed, the samples' means are normally distributed regardless of the sample sizes.
The implication of this theorem is that for sufficiently large populations, the normal distribution can be used to analyze samples drawn from populations that are not normally distributed, or whose distribution characteristics are unknown.
When means are used as estimators to make inferences about a population's parameters and $n \geq 30$, the estimator will be approximately normally distributed in repeated sampling. The mean and standard deviation of that sampling distribution are given as

$$\mu_{\bar{x}} = \mu$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where μ_x is the mean of the samples and σ_x is the standard deviation of the samples. If we know the mean and the standard deviation for the population, we can easily derive the mean and the standard deviation for the sample distribution,

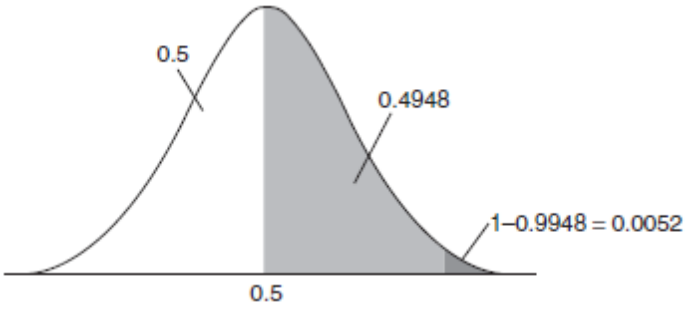
$$\mu = \mu_{\bar{X}}$$
$$\sigma = \sigma_{\bar{X}}\sqrt{n}$$

Example: Gajaga Electronics is a company that manufactures circuit boards. The average imperfection on a board is $\mu = 5$ with a standard deviation of $\sigma = 2.34$ when the production process is under statistical control. A random
sample of $n = 36$ circuit boards has been taken for inspection and a mean of $x = 6$ defects per board was found. What is the probability of getting a value of $x \leq 6$ if the process is under control?

Solution Because the sample size is greater than 30, the Central Limit Theorem can be used in this case even though the number of defects per board follows a Poisson distribution. Therefore, the distribution of the sample mean x is approximately normal with the standard deviation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.34}{\sqrt{36}} 0.39$$
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{6-5}{0.39} = \frac{1}{0.39} = 2.56$$

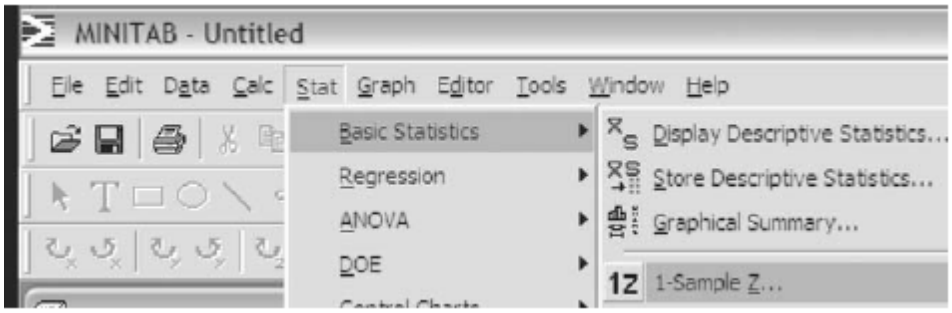
The result $Z = 2.56$ corresponds to 0.4948 on the table of normal curve areas



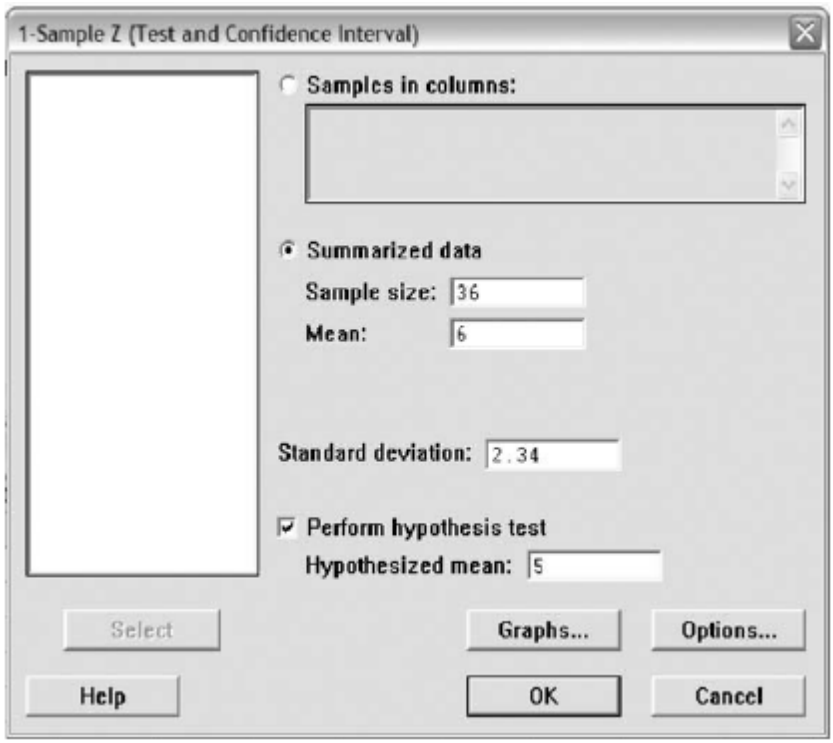
Remember from our normal distribution discussion that the total area under the curve is equal to one and half of that area is equal to 0.5. The area on the left side of any point on the horizontal line represents the probability of an event being “less than” that point of estimate, the area on the right represents the probability of an event being “more than” that point of estimate, and that the point itself represents the probability of an event being “equal to” that point of estimate. Therefore, the probability of getting a value of $x \leq 6$ is 0.5 + 0.4948 = 0.9948.

$$P(\bar{x} \leq 6) = 0.9948$$

We can use Minitab to come to the same result. From the Stat menu, select “Basic Statistics” and then select “1-Sample Z. . . ”



In the “1-Sample Z” dialog box, fill in the fields as indicated in the following Figure , then select “OK.”



One-Sample Z

Test of mu = 5 vs < 5
The assumed standard deviation = 2.34

N	Mean	SE Mean	95% Upper Bound	Z	P
36	6.00000	0.39000	6.64149	2.56	0.995

Example The average number of parts that reach the end of a production line defect-free at any given hour of the first shift is 372 parts with a standard deviation of 7. What is the probability that a random sample of 34 different productions' first-shift hours would yield a sample mean between 369 and 371 parts that reach the end of the line defect-free?

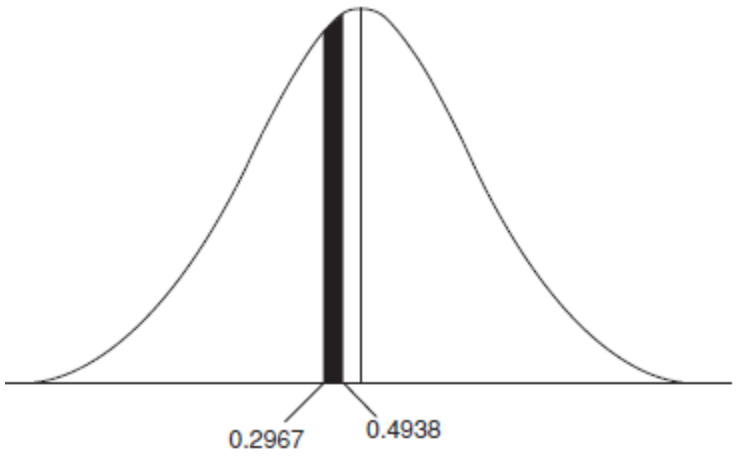
Solution In this case, $\mu = 372$, $\sigma = 7$, and $n = 34$. We must determine the probability of having the mean between 369 and 371. We will first find the probability that the mean would be equal to 369 and then for it to be equal to 371.

$$z = \frac{369 - 372}{\frac{7}{\sqrt{34}}} = \frac{-3}{1.2} = -2.5$$

In the *Z* score table, a value of 2.5 corresponds to 0.4938.

$$z = \frac{371 - 372}{\frac{7}{\sqrt{34}}} = \frac{-1}{1.2} = -0.8333$$

In the *Z* score table, a value of 0.833 corresponds to 0.2967.



The probability for the mean to be within the interval [369, 371] will be the difference between 0.4938 and 0.2967, which is equal to 0.1971.

Sampling Distribution of the Mean

If the means of all possible samples are obtained and organized, we could derive the *sampling distribution of the means*.
Consider the following example. We have five items labeled 5, 6, 7, 8 and 9 and we want to create a sampling distribution of the means for all the items. The size of the samples is two, so the number of samples will be

$$ {}_5C_2 = \frac{5!}{2!(5-2)!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{(2 \times 1)(3 \times 2 \times 1)} = 10 $$

Because the number of samples is 10, the number of means will also be 10. The samples and their means will be distributed as shown in the following Table 1

Table 1:

Combinations	Means
(6, 5)	5.5
(6, 7)	6.5
(6, 8)	7.0
(6, 9)	7.5
(5, 7)	6.0
(5, 8)	6.5
(5, 9)	7.0
(7, 8)	7.5
(7, 9)	8.0
(8, 9)	6.5

.

Exercise. How many samples of five items can we obtain from a population of 30?
Exercise. Based on the data in Table 2, build a distribution of the means for samples of two items.

Table 2:

9	12	14	13	12	16	15

Suppose we draw all possible samples of size *n* from a population of size *N*. Suppose further that we compute a mean score for each sample. In this way, we create a sampling distribution of the mean.

We know the following. The mean of the population (μ) is equal to the mean of the sampling distribution (μ_x). And the standard error of the sampling distribution (σ_x) is determined by the standard deviation of the population (σ), the population size, and the sample size. These relationships are shown in the equations below:

$$\mu_x = \mu \quad \text{and} \quad \sigma_x = \sigma * \text{sqrt}(1/n - 1/N)$$

Therefore, we can specify the sampling distribution of the mean whenever two conditions are met:

- The population is normally distributed, or the sample size is sufficiently large.
- The population standard deviation σ is known.

Note: When the population size is very large, the factor 1/N is approximately equal to zero; and the standard deviation formula reduces to: $\sigma_x = \sigma / \text{sqrt}(n)$. You often see this formula in introductory statistics texts.

Sampling Distribution of the Proportion

In a population of size *N*, suppose that the probability of the occurence of an event (dubbed a "success") is P; and the probability of the event's non-occurence (dubbed a "failure") is Q. From this population, suppose that we draw all possible samples of size *n*. And finally, within each sample, suppose that we determine the proportion of successes *p* and failures *q*. In this way, we create a sampling distribution of the proportion.

We find that the mean of the sampling distribution of the proportion (μ_p) is equal to the probability of success in the population (P). And the standard error of the sampling distribution (σ_p) is determined by the standard deviation of the population (σ), the population size, and the sample size. These relationships are shown in the equations below:

$$\mu_p = P \quad \text{and} \quad \sigma_p = \sigma * \text{sqrt}(1/n - 1/N) = \text{sqrt}[PQ/n - PQ/N]$$

where $\sigma = \text{sqrt}[PQ]$.

Note: When the population size is very large, the factor PQ/N is approximately equal to zero; and the standard deviation formula reduces to: $\sigma_p = \text{sqrt}(PQ/n)$. You often see this formula in intro statistics texts.

Example 1

Assume that a school district has 10,000 6th graders. In this district, the average weight of a 6th grader is 80 pounds, with a standard deviation of 20 pounds. Suppose you draw a random sample

of 50 students. What is the probability that the average weight of a sampled student will be less than 75 pounds?

Solution: To solve this problem, we need to define the sampling distribution of the mean. Because our sample size is greater than 40, the Central Limit Theorem tells us that the sampling distribution will be normally distributed.

To define our normal distribution, we need to know both the mean of the sampling distribution and the standard deviation. Finding the mean of the sampling distribution is easy, since it is equal to the mean of the population. Thus, the mean of the sampling distribution is equal to 80.

The standard deviation of the sampling distribution can be computed using the following formula.

$$\sigma_x = \sigma * \text{sqrt}(1/n - 1/N)$$
$$\sigma_x = 20 * \text{sqrt}(1/50 - 1/10000) = 20 * \text{sqrt}(0.0199) = 20 * 0.141 = 2.82$$

Let's review what we know and what we want to know. We know that the sampling distribution of the mean is normally distributed with a mean of 80 and a standard deviation of 2.82. We want to know the probability that a sample mean is less than or equal to 75 pounds. To solve the problem: mean = 80, standard deviation = 2.82, and value = 75. The probability that the average weight of a sampled student is less than 75 pounds is equal to 0.038.

Example 2

Find the probability that of the next 120 births, no more than 40% will be boys. Assume equal probabilities for the births of boys and girls. Assume also that the number of births in the population (N) is very large, essentially infinite.

Solution: The Central Limit Theorem tells us that the proportion of boys in 120 births will be normally distributed.

The mean of the sampling distribution will be equal to the mean of the population distribution. In the population, half of the births result in boys; and half, in girls. Therefore, the probability of boy births in the population is 0.50. Thus, the mean proportion in the sampling distribution should also be 0.50.

The standard deviation of the sampling distribution can be computed using the following formula.

$$\sigma_p = \text{sqrt}[PQ/n - PQ/N]$$
$$\sigma_p = \text{sqrt}[(0.5)(0.5)/120] = \text{sqrt}[0.25/120] = 0.04564$$

In the above calculation, the term PQ/N was equal to zero, since the population size (N) was assumed to be infinite.

Let's review what we know and what we want to know. We know that the sampling distribution of the proportion is normally distributed with a mean of 0.50 and a standard deviation of 0.04564. We want to know the probability that no more than 40% of the sampled births are boys. To solve the problem: mean = .5, standard deviation = 0.04564, and value = .4. The probability that no more than 40% of the sampled births are boys is equal to 0.014.

Note: This use of the Central Limit Theorem provides a good approximation of the true probabilities. The exact probability, computed using a binomial distribution, is 0.018 - very close to the approximation obtained with the Central Limit Theorem. The accuracy of the approximation increases as sample size increases.