Math 122
Introduction to Statistics
Central Limit Theorem

SAMPLING DISTRIBUTIONS

Means: Suppose that a value of a random variable x is associated with every member of a population. We can create a new random variable by collecting samples of size n from the population and averaging the associated values of x. This sample average $\bar{x}$ has a probability distribution called the **sampling distribution of the mean**.

Other Statistics: We can calculate other statistics from samples such as a proportion, variance, standard deviation, median, range, etc.

Notation: We will use the following notation:

Size	Population N	Sample n
Mean	μ	$\bar{x}$
Proportion	p	$\hat{p}$
Standard Deviation	σ	s
Variance	σ^2	s^2

Estimators: We frequently use sample statistics to estimate population parameters. Some statistics are good estimators of the corresponding population parameters. What this means is that the average of all values of the sample statistic is equal to the population parameter and that values of the sample statistic do not systematically over-estimate or under-estimate the population parameter. Such estimators are called **unbiased estimators**.

Unbiased Estimators: These are unbiased estimators: $\bar{x}$, s^2, and $\hat{p}$. The averages of these statistics are equal to the corresponding population parameters, and the statistics do not systematically over- or under-estimate the parameters.

Biased Estimators: These are biased estimators: median, range, s. Either the averages of these statistics do not equal the population parameters, or they tend to systematically over- or under-estimate the parameter.

Note about s: Even though s tends to be a biased estimator of σ, the bias is slight, and s is frequently used to estimate σ.

CENTRAL LIMIT THEOREM

Version 1: Suppose x is a random variable with mean μ and standard deviation σ and that $\bar{x}$ is the mean of simple random samples of size n of values of x. If x is normal, or if $n > 30$, then $\bar{x}$ is approximately normal with mean $\mu_{\bar{x}} = \mu$ and standard deviation

$\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

Version 2: Suppose x is a random variable with mean μ and standard deviation σ and that $\bar{x}$ is the mean of simple random samples of size n of values of x. If x is normal, or if $n > 30$, then

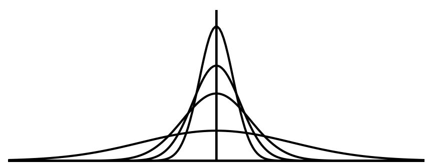
$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{\bar{x}-\mu}{\sigma}\sqrt{n}$$

is approximately standard normal.

Note: If x is normal, then $\bar{x}$ is normal for all sample sizes.

Note: The larger n is, the closer $\bar{x}$ is to being normal.

Comparison: For increasing values of n, the values of $\bar{x}$ are concentrated more closely around μ. As n gets larger, it is easier to find values of $\bar{x}$ close to μ and harder to find values any distance away from μ.



Pictured are the density curves for distributions of $\bar{x}$ with sample sizes of 1, 5, 10, and 20 (with the highest peaks corresponding to larger sample sizes).

Examples: Heights of college age females are normally distributed with a mean of 63.8 inches and a standard deviation of 2.7 inches.

1. What is the probability that a randomly selected college age female is between 63 inches and 64.6 inches tall?

2. What is the probability that a randomly selected college age female is between 65 inches and 67 inches tall?

3. What is the probability that a simple random sample of 5 college age females has an average height between 63 inches and 64.6 inches?

4. What is the probability that a simple random sample of 5 college age females has an average height between 65 inches and 64.7 inches?

5. What is the probability that a simple random sample of 50 college age females has an average height between 63 inches and 64.6 inches?

6. What is the probability that a simple random sample of 50 college age females has an average height between 65 inches and 64.7 inches?

Solution:

1. Let x be the height of a randomly selected college age female. Since this is a question about a single person, and since heights are normally distributed, we simple convert to z scores.

$$\begin{aligned} P(63 < x < 64.6) &= P\left(\frac{63 - 63.8}{2.7} < \frac{x - 63.8}{2.7} < \frac{64.6 - 63.8}{2.7}\right) \\ &= P\left(-0.2963 < z < 0.2963\right) \\ &= \textit{normalcdf}\,(-0.2963, 0.2963) \\ &= 0.2330 \end{aligned}$$

2. Let x be the height of a randomly selected college age female. Since this is a question about a single person, and since heights are normally distributed, we simple convert to z scores.

$$\begin{aligned} P(65 < x < 67) &= P\left(\frac{65 - 63.8}{2.7} < \frac{x - 63.8}{2.7} < \frac{67 - 63.8}{2.7}\right) \\ &= P\left(0.4444 < z < 1.1852\right) \\ &= \textit{normalcdf}\,(0.4444, 1.1852) \\ &= 0.2104 \end{aligned}$$

3. Let $\bar{x}$ be the average height of a simple random sample of 5 college age females. Since x is normally distributed, then so is $\bar{x}$. The mean for $\bar{x}$ is $\mu_{\bar{x}} = 63.8$ and the standard deviation is

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 2.7/\sqrt{5} = 1.2075.$$

We now proceed as above with this new standard deviation.

$$\begin{aligned} P(63 < \bar{x} < 64.6) &= P\left(\frac{63 - 63.8}{1.2075} < \frac{\bar{x} - 63.8}{1.2075} < \frac{64.6 - 63.8}{1.2075}\right) \\ &= P\left(-0.6625 < z < 0.6625\right) \\ &= \textit{normalcdf}\,(-0.6625, 0.6625) \\ &= 0.4923 \end{aligned}$$

NOTICE: In this case, the probability for the sample is HIGHER than the probability for the individual. This is because we are looking at a range of heights CONTAINING the mean.

4. Let $\bar{x}$ be the average height of a simple random sample of 5 college age females. Since x is normally distributed, then so is $\bar{x}$. The mean for $\bar{x}$ is $\mu_{\bar{x}} = 63.8$ and the standard deviation is

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 2.7/\sqrt{5} = 1.2075.$$

We now proceed as above with this new standard deviation.

$$\begin{aligned} P(65 < \bar{x} < 67) &= P\left(\frac{65-63.8}{1.2075} < \frac{\bar{x}-63.8}{1.2075} < \frac{67-63.8}{1.2075}\right) \\ &= P\left(0.9938 < z < 2.6501\right) \\ &= \mathit{normalcdf}(0.9938, 2.6501) \\ &= 0.1561 \end{aligned}$$

NOTICE: In this case, the probability for the sample is LOWER than the probability for the individual. This is because we are looking at a range of heights NOT CONTAINING the mean.

5. Let $\bar{x}$ be the average height of a simple random sample of 50 college age females. Since our sample size is larger than 30 (or since the population is normal), $\bar{x}$ is normal. The mean for $\bar{x}$ is $\mu_{\bar{x}} = 63.8$ and the standard deviation is

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 2.7/\sqrt{50} = 0.3818.$$

We now proceed as above with this new standard deviation.

$$\begin{aligned} P(63 < \bar{x} < 64.6) &= P\left(\frac{63-63.8}{0.3818} < \frac{\bar{x}-63.8}{0.3818} < \frac{64.6-63.8}{0.3818}\right) \\ &= P\left(-2.0953 < z < 2.0953\right) \\ &= \mathit{normalcdf}(-2.0953, 2.0953) \\ &= 0.9639 \end{aligned}$$

6. Let $\bar{x}$ be the average height of a simple random sample of 50 college age females. Since our sample size is larger than 30 (or since the population is normal), $\bar{x}$ is normal. The mean for $\bar{x}$ is $\mu_{\bar{x}} = 63.8$ and the standard deviation is

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 2.7/\sqrt{50} = 0.3818.$$

We now proceed as above with this new standard deviation.

$$\begin{aligned} P(65 < \bar{x} < 67) &= P\left(\frac{65-63.8}{0.3818} < \frac{\bar{x}-63.8}{0.3818} < \frac{67-63.8}{0.3818}\right) \\ &= P\left(3.1430 < z < 8.3814\right) \\ &= \mathit{normalcdf}(3.1430, 8.3814) \\ &= 0.0008 \end{aligned}$$

VARIANTS

There are theorems which similar to the central limit theorem for sample proportions and sample variances. These theorems can help to simplify calculations, and they are the basis for the techniques we will be using later for confidence intervals and hypothesis testing.

Means Without σ: Suppose x is a random variable with mean μ and that $\bar{x}$ is the mean of simple random samples of size n of values of x. If x is normal, or if $n > 30$, then the distribution of

$$\frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{\bar{x}-\mu}{s}\sqrt{n}$$

is approximately a Student t distribution with $n-1$ degrees of freedom.

Proportions 1: Suppose that x is the number of successes in simple random samples of size n selected from a population in which the proportion of success is p. If $np \geq 5$ and $n(1-p) \geq 5$, then the sample successes x are approximately normally distributed with mean $\mu = np$ and standard deviation $\sigma = \sqrt{np(1-p)}$.

Proportions 2: Suppose that simple random samples of size n are selected from a population in which the proportion of success is p. If $np \geq 5$ and $n(1-p) \geq 5$, then the sample proportion $\hat{p}$ is approximately normally distributed with mean $\mu = p$ and standard deviation $\sigma = \sqrt{\frac{p(1-p)}{n}}$.

Variance: The distribution of sample variances involves a different type of distribution called a χ^2 distribution. (χ is the Greek letter "chi.")