

Simple Inferential Statistics

To make inferences about, say, the sample mean \bar{X} , we need to 1) have an estimator for the sample mean (i.e. an algorithm for calculating it, 2) know the name of the theoretical (reference) distribution the sample mean follows, and 3) have an estimate of the standard error (standard deviation), $\sigma_{\bar{X}}$, of the sample mean.

The estimator of the sample mean is

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$$

From the Central Limit Theorem, we know that the sample mean follows the normal distribution.

The standard error of the sample mean can be estimated by

$$\begin{aligned} \sigma_{\bar{X}} &= \sqrt{\sigma_X^2 / n} \\ &= \sigma_X / \sqrt{n}, \end{aligned}$$

using the sample standard deviation, s , as an estimate of the population standard deviation σ_X .

With this information, questions of the following kind can be answered.

1) *What is the chance that the value 11.0 will be equalled or exceeded in a normal distribution with a mean of 10.0 and a standard deviation of 2.0?* (Note that question deals with a single value of a variable, and not with the sample mean, and is used to illustrate the idea of "looking up" a probability value using a cumulative density function.)

To answer this question, obtain the value of z , a variable that follows the standard normal distribution, and use the cdf of the standard normal distribution to determine probability of observing a value greater than or equal to z :

$$z = \frac{x - \mu}{\sigma} = \frac{11.0 - 10.0}{2.0} = 0.5$$

The R `pnorm(z, mean=0, sd=1)` function can be used to return this value, with `z <- 0.5`; the value returned represents the area under the cdf to the left of the value "plugged in" which in this case is 0.6915. The probability of observing a value greater than or equal to 11 is $1.0 - .6915 = .3085$.

2) *How unusual is the value of a sample mean equal to 5.15, relative to population with a true mean $\mu = 5.0$, and standard deviation, $\sigma = 1.2$, for a sample size of $n = 40$?*

To answer this question, first obtain the standard error of the sample mean, $\sigma_{\bar{X}}$

$$\begin{aligned}\sigma_{\bar{X}} &= \sigma / \sqrt{n} \\ &= 1.2 / \sqrt{40} \\ &= 1.2 / 6.3245 \\ &= .18974.\end{aligned}$$

Then, obtain the value of z :

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{5.15 - 5.0}{.18974} = 0.79056.$$

Note that now the sample mean and standard error for the sample mean appear in the equation for z . The `pnorm()` function returns the value 0.7854 for $z = 0.79056$.

3) *Suppose we know that the "true" mean of a particular process is 8.5, and that the standard error, $\sigma_{\bar{X}}$, of the means of samples of 100 observations that describe that process is 1.5. What is the range of values that will contain 95% of the sample means that might be observed in the future?*

To answer this question, we need to find the (two) values of \bar{X} that cut off areas in the tails of the pdf for \bar{X} equal to 0.025 ($= (1.0 - 0.95)/2.0$). The inverse cumulative density function is used to get this information. The R function `qnorm(0.025, mean=8.5, sd=1.5)` returns the value 5.5501 while `qnorm(0.975, mean=8.5, sd=1.5)` returns the value 11.4399.