The Sample Mean

Let \( x_1, x_2, \ldots, x_n \) be a random sample of size \( n \) of a population measurement \( X \) with unknown mean \( \mu \) and standard deviation \( \sigma \). Let \( \bar{x} \) be the sample mean. A sample mean is only an estimate of the true population mean \( \mu \). When sampling with replacement or from a “large” population, the collection of all possible sample means \( \bar{x} \) has the following properties:

\[
\mu_{\bar{x}} = \mu \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}
\]

The average of all possible sample means is the true population average \( \mu \), but its standard deviation is a fraction of the population’s standard deviation.

Central Limit Theorem

(i) When sampling from normally distributed measurements, \( X \sim N(\mu, \sigma) \), then for all sample sizes \( n \), we have \( \bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \). That is, the distribution of all possible sample means is also normally distributed with mean \( \mu \) and standard deviation \( \sigma / \sqrt{n} \).

(ii) When sampling a non-normally distributed measurement, then for large sample sizes \( n \), we have that \( \bar{x} \) is approximately \( N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \). That is, the distribution of all possible sample means is close to a normal distribution with mean \( \mu \) and standard deviation \( \sigma / \sqrt{n} \), provided we have a large enough sample.

Normal Calculations

Using these results, we can compute probabilities such as \( P(\bar{x} < b) \) or \( P(a \leq \bar{x} \leq b) \) using the normal distribution with a mean of \( \mu \) and a standard deviation of \( \sigma / \sqrt{n} \). We also can compute inverse normal calculations with the \texttt{invNorm} command.

Example. The blood cholesterol level of men aged 20 to 34 is normally distributed with mean \( \mu = 188 \text{ mg/dL} \) and standard deviation \( \sigma = 41 \text{ mg/dL} \). You measure the cholesterol level of 100 such men chosen at random and calculate the sample mean \( \bar{x} \).

(a) What is the population \( \Omega \)? What is the measurement \( X \) and its distribution?
(b) From random samples of size 100, what is the distribution of all \( \bar{x} \) values?
(c) From random samples of size 100, how often is \( \bar{x} \) less than 180?
(d) What cholesterol level is such that 90% of all \( \bar{x} \) are below this level?
(e) Using \( z \)-scores, between what two values should 95% of sample means lie from samples of size \( n = 100 \)?
**Solution.** (a) \( \Omega = \text{Men aged 20 to 34}; X = \text{blood cholesterol level}; X \sim N(188, 41). \)

(b) For samples of size \( n = 100 \), then \( \bar{x} \sim N\left(188, \frac{41}{\sqrt{100}}\right) = N(188, 4.1). \)

(c) We simply compute \( P(\bar{x} < 180) \) for \( \bar{x} \sim N(188, 4.1) \) using the command `normalcdf(-1E99,180,188,4.1)` to obtain \( P(\bar{x} < 180) \approx 0.0255 \). So \( \bar{x} \) less than 180 about 2.55% of the time.

(d) We use the command `invNorm(.9,188,4.1)` to obtain a level of about 193.254.

(e) The 95% \( z \)-score is 1.96; thus, 95% of sample means should lie within 188 \( \pm 1.96 \times 4.1 \), or from 179.964 to 196.036.

**A Confidence Interval for \( \mu \)**

From past observations, adult heights are assumed to be normally distributed with a standard deviation of about \( \sigma = 4 \) inches. Suppose that the mean \( \mu \) is unknown. With random samples of 64 adults, where should 95% of sample means lie? Suppose \( \bar{x} = 68.2 \) from one sample of size 64, how can we estimate \( \mu \) ?

Because \( \bar{x} \sim N\left(\mu, \frac{4}{\sqrt{64}}\right) = N(\mu, 0.5) \), then 95% of all \( \bar{x} \) should be within \( \mu \pm 1.96 \times 0.5 = \mu \pm 0.98 \). But that also means that 95% of the time \( \mu \) is within \( \bar{x} \pm 0.98 \) with samples of size \( n = 64 \).

If one such sample yields \( \bar{x} = 68.2 \), then we can say \( \mu = 68.2 \pm 0.98 \), or that \( \mu \) is probably from 67.22 to 69.18, which is a 95% confidence interval for \( \mu \).

**Exercise.** A controlled test group of diabetic patients are monitored one hour after drinking a soda. After weeks of studying this procedure, their glucose levels are found to be normally distributed with \( \mu = 125 \text{ mg/dl} \) and \( \sigma = 10 \text{ mg/dl} \).

(a) What is the population \( \Omega \) under study? What is the measurement \( X \) and its distribution?
(b) If 16 such subjects are selected and tested at random, then what is the distribution of their \( \bar{x} \) values?
(c) For sample of size 16, what is the probability that \( \bar{x} \) is above 127? From 123 to 127?
(d) Find the level \( L \) such that there is only 0.05 probability that the mean glucose level of 16 test results falls above \( L \).
(e) Using \( z \)-scores, between what two values should 98% of sample means lie from samples of size 16?
Answers

(a) $\Omega =$ this control group of diabetic patients (it may not be representative of all diabetics due to the controlled situation); $X =$ glucose level one hour after drinking soda; $X \sim N(125, 10)$.

(b) For samples of size $n = 16$, then $\bar{x} \sim N\left(125, \frac{10}{\sqrt{16}}\right) = N(125, 2.5)$.

(c) $P(\bar{x} > 127) = \text{normalcdf}(127, 1\times10^9, 125, 2.5) \approx 0.2118$

$P(123 \leq \bar{x} \leq 127) = \text{normalcdf}(123, 127, 125, 2.5) \approx 0.5763$

(d) $L = \text{invNorm}(0.95, 125, 2.5) = 129.112$

(e) Within $125 \pm 2.326 \times 2.5$