1: CONFIDENCE INTERVALS FOR THE MEAN; KNOWN VARIANCE

Suppose $X_1, \ldots, X_n$ are independent and identically distributed (*iid*) random variables, and we want to make inferences about the mean, μ, of the population. That is, $\mu = E[X_i]$. Since μ determines the population distribution (at least in part), it is called a **parameter.**

A **point estimator,** such as the sample mean $\bar{X}$, provides a single guess for the true value of the parameter μ.

Used by itself, $\bar{X}$ is of limited usefulness because it provides no information about its own reliability. Furthermore, the reporting of $\bar{X}$ alone may leave the false impression that $\bar{X}$ estimates μ with complete accuracy. This is not the case. (Why?)

An **interval estimator** consists of a range of values designed to contain μ with a prespecified probability.

The interval estimator automatically provides a margin of error to account for the sampling variability of $\bar{X}$.

Eg: The Federal Trade Commission wants to estimate the average amount of Pepsi that is placed in

2-liter bottles at the Knoxville, Tennessee bottling plant. A random sample of 100 2-liter bottles from this bottling plant yielded a sample average of 1.985 liters. Can we conclude that the average for all bottles does not meet the 2-liter specification?

- **Confidence Interval:** An interval with random endpoints which contains the parameter of interest (in this case, μ) with a prespecified probability, denoted by $1-\alpha$ (the **confidence level).**

Most practitioners use either α=.05 or α=.01, but these choices are completely arbitrary.

Define $\sigma^2 = var[X_i]$. Here, we assume that σ^2 is known (and finite). In practical situations, we will rarely know the value of σ^2, but this assumption is convenient for now.

Eg: In the Pepsi example, the bottling plant has informed the FTC that the standard deviation for the amount of Pepsi placed in 2-liter bottles is .05 liters. Since $n = 100$, we can assume (using the Central Limit Theorem) that $\bar{X}$ is normally distributed with mean $\mu_{\bar{X}} = \mu$ (unknown) and standard error $\sigma_{\bar{X}} = \sigma/\sqrt{n} = .005$.

Therefore, the probability is about 95% that $\bar{X}$ will be within two standard errors of its mean.

So the probability is about .95 that μ will be within .01 liters of $\bar{X}$.

Thus, the interval $\bar{X} \pm .01$ will contain μ with probability about .95.

In general, the interval $\bar{X} \pm 2\sigma/\sqrt{n}$ will contain μ with probability about .95.

Next, we develop the general formula for a level $1-\alpha$ CI. We have the following Theorem.

- Let $z_{\alpha/2}$ denote the z value such that the area to its right under the standard normal curve is $\alpha/2$.

- Then if the population distribution is normal, the interval $\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ is a CI for μ with confidence level $1-\alpha$.

• Even if the population distribution is not normal, $Pr(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}) \to 1 - \alpha$ as the sample size tends to infinity. (We say that the CI has an *asymptotic* level of $1 - \alpha$.)

Proof: First, suppose the population is normal. We need to show that Prob(μ is in the CI) $= 1-\alpha$.

Since $\bar{X}$ is normally distributed with mean μ and SE $\sigma/\sqrt{n}$, it follows that $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ has a standard normal distribution. So does $\frac{\mu-\bar{X}}{\sigma/\sqrt{n}}$. Therefore,

$$1-\alpha = Pr\,(-z_{\alpha/2} < \frac{\mu-\bar{X}}{\sigma/\sqrt{n}} < z_{\alpha/2})$$

$$= Pr\,(-z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}} < \mu-\bar{X} < z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}})$$

$$= Pr\,(\bar{X} - z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}+z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}})$$

$$= Pr\,(\text{CI Contains } \mu)\ ,$$

and the proof is finished for normal populations. If the population is not normal, then we can use the CLT, which says that the distribution of $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ converges to the standard normal distribution as $n \to \infty$. Using the argument above and the Central Limit Theorem, we have

$$Pr\,(\text{CI Contains } \mu) = Pr\,(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) \to 1-\alpha\ .$$

Interpretation of Confidence Intervals

As stated earlier, the confidence interval $\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ will cover μ with probability $1-\alpha$.

There is a subtle difficulty in the practical interpretation of the results, however, as demonstrated in the following example.

Eg: In the Pepsi example, we had $n = 100$, $\sigma = .05$ and $\bar{x} = 1.985$. The 95% confidence interval estimator for μ is $\bar{X} \pm 1.96\,\sigma/\sqrt{n}$, that is, $\bar{X} \pm .01$.

Plugging in $\bar{x} = 1.985$, we obtain the interval (1.975,1.995).

Which of the following statements is true?

a) There is a 95% chance that μ is between 1.975 and 1.995.

b) μ will be between 1.975 and 1.995 95% of the time.

c) In 95% of all future samples, $\bar{X}$ will be between 1.975 and 1.995.

d) μ is between 1.975 and 1.995.

e) None of the above.

- Warning: The practical interpretation of confidence intervals is extremely tricky. The difficulty has to do with the distinction between an estimator and an estimate.

An **estimator** is a random variable whose value depends on a sample not yet taken (eg: $\bar{X}$).

An **estimate** is the value actually taken by the estimator for a given sample (eg: $\bar{x}$=1.985).

- The CI is an interval estimator. It has random endpoints. After the sample is taken, its endpoints take on specific values, yielding an interval estimate.

- Since μ is an (unknown) *constant*, and since the endpoints of the CI estimate are fixed numbers (eg: 1.975, 1.995), it makes *no sense* to talk about the probability that the CI estimate contains μ. Either it does or it doesn't, and we may never find out which of these events has occurred.

- Instead, it is the CI *estimator* which contains μ with probability $1-\alpha$.

The estimator has random endpoints, $\bar{X}-z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$,

$\bar{X}+z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$.

The word "probability" refers to the long-run proportion of the time that these random endpoints will contain the true mean μ, assuming a large number of repetitions of the experiment of collecting a random sample and constructing the CI.

Thus the confidence level $1-\alpha$ refers to the process of constructing confidence intervals, not to the particular CI estimate obtained from the given sample.

• Unfortunately, in practice we have only one sample. So what good is a probability statement referring to all samples which *might* have been taken?

1) You can think of $1-\alpha$ as an overall success rate. If you compute many 95% confidence intervals over your lifetime, and if the required assumptions are satisfied for each one, then approximately 95% of

these confidence intervals will contain their respective population means. Unfortunately, you may *never know* which ones were wrong.

2) Even though we can't talk about the *probability* that the given CI estimate contains μ, if we had to *bet* on it, we would say that the odds are 19 to 1 that our given 95% CI estimate fails to cover μ.