Lecture Overview

- Estimation concepts
- Example point estimates
- Example interval estimates
- Basic sample size calculations
- Excursion: Confidence Intervals for regression coefficients

Parametric Statistical Inference: Estimation

- ***Statistical inference*** includes all procedures that ***draw conclusions*** about an unknown population based on ***one set*** of n sample observations.
- Question: Where is ***uncertainty*** in inferential statistics coming from?
  Answer: Drawing a ***random*** sample from the underlying population. Drawing different samples will lead to different sets of observed sample data and, therefore, to different estimates.

- Statistical inference deals with:

  [a] ***estimation*** of a population characteristic, which usually is a parameter of the underlying population distribution, i.e., μ or σ^2 of the normal distribution, and ***quantifying the uncertainty*** of the estimate.

  [b] ***testing hypotheses*** (initial assumptions) with regards to parameters of an underlying population distribution.

  Based on the set of sample observations a specific hypothesis can is ***rejected*** at a given ***error probability*** α under the assumption that the hypothesis is true (see BBR Chapter 8).

- In both approaches a sample $\{X_1, X_2, \ldots, X_n\}$ of given size n is drawn and a ***sampling statistics*** T is selected that is related to the unknown parameter θ of the ***population distribution***:

$$\hat{\theta} = T(x_1, x_2, \ldots, x_n).$$

  The ***hat*** on top of $\hat{\theta}$ denotes the ***estimate*** of the unknown population parameter θ.

  - For the estimation approach, $\hat{\theta}$ allows expressing that the parameter θ of the population distribution falls with a given ***level of certainty*** into an interval around θ.

  - For the hypothesis testing approach, $\hat{\theta}$ allow us to make statements about the probability that the observed ***sampling statistic differs significantly*** from an underlying

hypothetical reference population parameter θ_0 (the *naught* subscript in θ_0 refers to the ***assumed*** population characteristic).

Statistical Estimation

- Def. Point Estimation: A single number is calculated from the sample and it is used as the ***best estimate*** $\hat{\theta}$ of some unknown population parameter θ.
- Def. Interval Estimation: In interval estimation the sample is used to identify a ***range*** $[\hat{\theta}_L, \hat{\theta}_U]$ of estimation bounds for the ***unknown population parameter*** θ, within which it is believed to be embedded with a ***given probability*** $1-\alpha$.

Concept: Point Estimation

- Def. Statistical Estimator and Statistical Estimate:
  - A statistical estimator is a generic function T of the n random variables $X_1, X_2, \ldots, X_n$ from a sample. An estimator T is, therefore, also a ***random variable***.
  - Once the sample is taken, the observed values of the random variables are known. These are denoted by $x_1, x_2, \ldots, x_n$.

- The value of the estimator $\hat{\theta} = T(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$ is known as the ***statistical estimate*** of the population parameter θ.

- There are ***many possible estimation rules*** (functional specifications of the estimator) to estimate a population parameter.
  For instance, instead of the mean $\bar{X}$ we may use the mode or median to estimate the parameter μ of an underlying population.

- Two criteria to evaluate whether a selected estimation procedure is the ***best***, are
  [a] the ***bias*** and
  [b] the ***efficiency***.
  Both criteria are based on the deviation of the estimator from its true population value, i.e., $\hat{\theta} - \theta$.

- The ***mean estimation error*** (the expected variation around the true population value) is:

$$E(\hat{\theta} - \theta) = \sum_{i=1}^{p} (\hat{\theta}_i - \theta) \cdot \Pr(\hat{\theta}_i) \text{ or } E(\hat{\theta} - \theta) = \int_{-\infty}^{\infty} (\hat{\theta}_i - \theta) \cdot f(\hat{\theta}) \cdot d\hat{\theta} \ .$$

Main challenge is to develop the distribution $\Pr(\hat{\theta}_i)$ or density $f(\hat{\theta})$ of an estimation rule $\hat{\theta} = T(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$.

- <u>Def. Unbiased Estimator:</u> An estimator $\hat{\theta}$ of a population parameter is said to be unbiased if its expected value is equal to the population parameter. That is, $\hat{\theta}$ is unbiased of $E(\hat{\theta}) = \theta$.
- Example: Biased versus unbiased estimator:

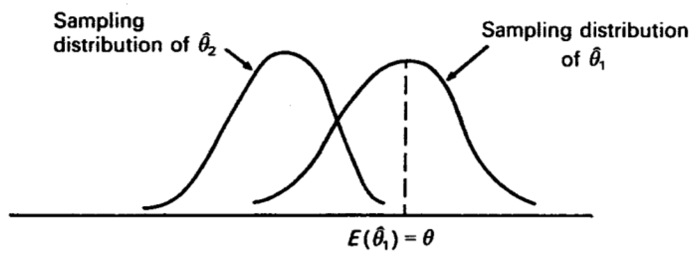


FIGURE 8-3. Sampling distributions for a biased and an unbiased estimator of θ.

- A ***biased estimator*** may become ***asymptotically*** unbiased as the sample size n increases. That is, the bias is consistently shrinking.

  An example is the biased variance estimator $\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$, because for large n we get $\lim_{n \to \infty} n \approx n - 1$.
- In how far an ***estimator varies around the population value*** <u>from sample to sample</u> is measured by the expected squared differences over all possible samples $E\left[(\hat{\theta} - \theta)^2\right]$, that is, $Var(\hat{\theta})$.

- Ultimately, we want to have an ***unbiased*** estimation rule T that, in addition, has the ***smallest possible variability for sample to sample***, i.e., the smallest variance.
- Example: Unbiased estimators with smaller versus larger variance:

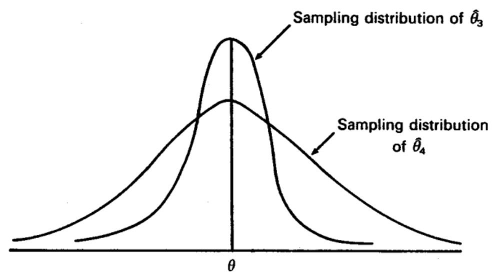


FIGURE 8-4. Sampling distributions for two unbiased estimators of θ.

- <u>Def efficient estimator:</u> An ***unbiased*** estimation rule T is called ***efficient*** or best unbiased estimator, if it has the ***smallest variance*** compared to any other possible ***unbiased*** estimation rules.
- Which estimation rule T is the most efficient may dependent on the underlying population distribution.
- <u>Def consistency:</u> An estimator is called consistent if for an increasing sample size its estimated value $\hat{\theta}$ approaches the true population value θ and its variance is shrinking.
  That is, it ***converges in probability*** to the true population value:

$$\lim_{n\to\infty} \Pr\left(\left|\hat{\theta}_n - \theta\right| < \delta\right) = 1$$

for any small, positive value of δ.

- For biased estimation rules with $E(\hat{\theta}) \neq \theta$ the concept of ***efficiency does not apply*** and their ***mean square error*** needs to be evaluated

$$MSE = E(\hat{\theta} - \theta)^2 = \underbrace{\left[E(\hat{\theta}) - \theta\right]^2}_{Bias^2\ of\ T} + \underbrace{E\left[\hat{\theta} - E(\hat{\theta})\right]^2}_{Variance\ of\ T}$$

- There may be a ***tradeoff*** between a bias and the variance of an estimator T in the MSE:

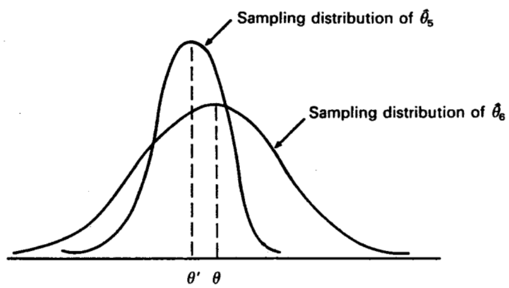


FIGURE 8-5. Difficulties in choosing a potential estimator.

Ultimately, one may ***prefer*** an estimation rule with a small bias but with a substantially smaller *MSE* compared to alternative unbiased estimation rules.

- To express the variability in original units, one can use the ***root mean square error***:

$$RMSE = \sqrt{MSE} = \sqrt{Bias^2 + Variance}.$$

Note that we ***cannot decompose*** this expression into the sum

[a] of the square-root of bias and

[b] the square-root of the variance, because both terms are jointly under the square root.

Concept: Interval Estimation

- Confidence intervals also depend on the n observed sample observations $X_1, X_2, \ldots, X_n$.
- These intervals provide ***more information*** than a simple point estimators:
    - the ***width*** defined by a lower and an upper bound of the interval $\left[\theta_L, \theta_U\right]$
    - this ***width*** is associated with the ***degree of certainty*** of the point estimator $\hat{\theta}$.
- These intervals express the degree of certainty that the true parameter θ is within the ***confidence interval***, i.e.,

$$\Pr\left(\theta \in \left[\hat{\theta}_L, \hat{\theta}_U\right]\right) = 1 - \alpha$$

where α is usually small the ***error probability*** that the ***true population parameter is outside*** the interval. ***<u>We want to keep this error probability small.</u>***

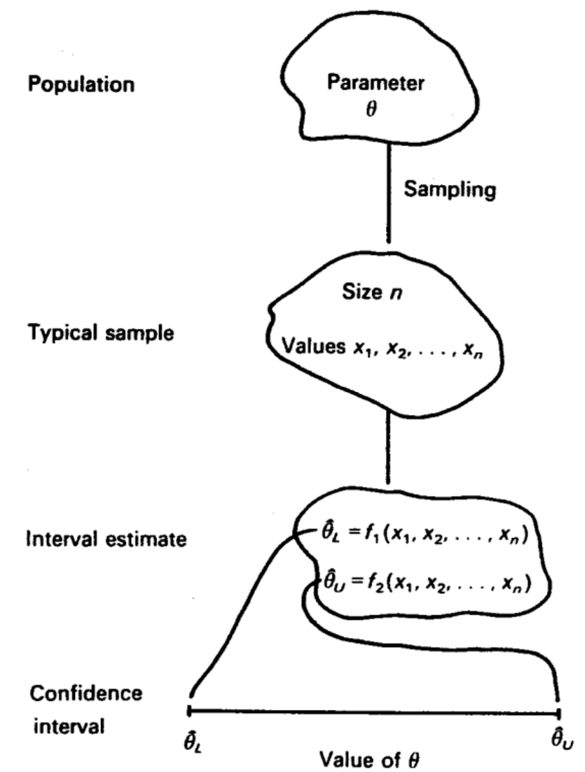


FIGURE 8-6. Interval estimation of a population parameter θ.

- The ***smaller*** the error probability α becomes the ***wider*** the confidence interval will get, because we are increasing the chance that the interval will be covering the true population parameter.
- See Figure 8-6 conceptionally shows the estimation of the confidence interval:

Examples: Point Estimators

[a] Properties of the Mean Estimation Rule $\overline{X}$

- For an underlying ***normal population distribution*** the arithmetic mean is ***unbiased*** and its variance is only 56% of that of the median's variance. It is therefore ***most efficient***.
- Since the mean minimizes the sum of the squared deviations from the central value, it is highly sensitive to extreme value (leverages).
- For ***symmetric*** distributions with ***heavy tails*** the mean may become less efficient than the median.

TABLE 8-1
Point Estimators of μ, π, and σ^2

Population parameters	Point estimator	Formula for point estimate
μ	$\bar{X}$	Mean of sample values: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
	Median	Middle value in sample (50%-ile)
	25% Trimmed Mean	Mean of middle 50% of values in sample
	10% Trimmed Mean	Mean of middle 80% of values in sample
π	P	$p = x/n$ where x = number of successes in n trials
σ^2	S^2	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

- For highly ***asymmetric distributions*** the mean is even less efficient.
- ***Trimmed means***, which disregard the most extreme observations in both tails, are more ***robust*** because are not affected by these high leverage values.
- The ***advantage*** of working with the mean is that we know from the central limit theorem its underlying sampling distribution (the normal distribution) as the sample size increases.

[b] Population Proportion

- Assuming the random variables is coded as $X_i = \begin{cases} 1 & \text{for success} \\ 0 & \text{for failure} \end{cases}$.
- Then the population proportion estimation rule is ***structurally equivalent*** to the arithmetic mean

$$\hat{\pi} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{\#\ of\ successes}{\#\ of\ trials}.$$

- Therefore, the properties of the mean apply to the proportion estimator $\hat{\pi}$ for large sample sizes.

[c] Population variance

- The estimator for the sample variance is $S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \bar{X})^2$.

- It is an unbiased estimator (over all potential samples of size n the estimator S^2 will average to the population variance σ^2).
- For a normal distributed underlying population S^2 is the most efficient estimator for σ^2.
- Division by n would lead to a biased estimator, which systematically would ***underestimate*** the variance. However, for increasing sample sizes it a consistent estimator.
- Review of the reasons:
  1. Use of $\bar{X}$ minimizes the sum of the squared deviations.
  2. We lose one degree for freedom because, once the mean is known, only $n-1$ observations need to be available (recall zero sum property $\sum_{i=1}^{n} x_i - n \cdot \bar{x} = 0$).

Example: Interval estimation for the population mean

- Review: Standard normal distribution
  1. For a ***standard normal*** distributed variable $(1-\alpha)\times 100\%$ of the observations are within the interval $\left[z_{\alpha/2}, z_{1-\alpha/2}\right]$.

     That is, $\Pr\left(z \in \left[z_{\alpha/2}, z_{1-\alpha/2}\right]\right) = 1-\alpha$

     The quantiles:

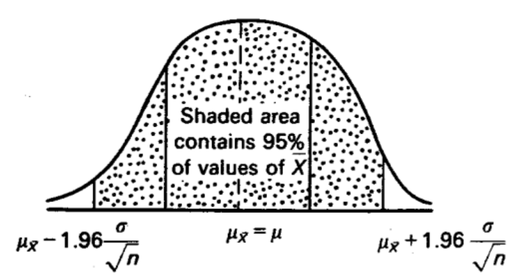


FIGURE 8-7. Sampling distribution of $\bar{X}$.

$z_{\alpha/2}$ is in the ***left tail*** of the standard normal distribution and therefore is ***negative***.

$z_{1-\alpha/2}$ is in the ***right tail*** of the standard normal distribution and therefore is ***positive***.

- The estimator for the arithmetic mean is distributed as $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$ (recall the central limit theorem).
- Therefore, the confidence interval around the ***population expectation*** μ is for a ***given mean estimation rule***:

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

$$\Pr\left(\sigma/\sqrt{n} \cdot z_{\alpha/2} \le \bar{X}-\mu \le \sigma/\sqrt{n} \cdot z_{1-\alpha/2} \right) = 1-\alpha$$

$$\Pr\left(-\bar{X} + \sigma/\sqrt{n} \cdot \underbrace{z_{\alpha/2}}_{negative} \le -\mu \le -\bar{X} + \sigma/\sqrt{n} \cdot \underbrace{z_{1-\alpha/2}}_{positive} \right) = 1-\alpha \quad \text{Multiplying by -1 re-orients the inequality}$$

$$\Pr\left(\bar{X} - \sigma/\sqrt{n} \cdot z_{\alpha/2} \ge \mu \ge \bar{X} - \sigma/\sqrt{n} \cdot z_{1-\alpha/2} \right) = 1-\alpha \quad \text{making use of } z_{\alpha/2} = -z_{1-\alpha/2} \text{ and } -z_{\alpha/2} = z_{1-\alpha/2}$$

$$\Pr\left(\bar{X} + \sigma/\sqrt{n} \cdot z_{1-\alpha/2} \ge \mu \ge \bar{X} + \sigma/\sqrt{n} \cdot z_{\alpha/2} \right) = 1-\alpha \quad \text{and rearranging the order}$$

$$\Pr\left(\bar{X} + \sigma/\sqrt{n} \cdot z_{\alpha/2} \le \mu \le \bar{X} + \sigma/\sqrt{n} \cdot z_{1-\alpha/2} \right) = 1-\alpha$$

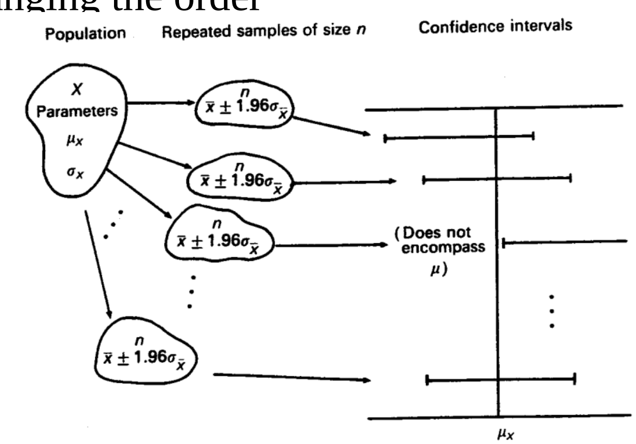


FIGURE 8-8. Interval estimates constructed from repeated samples of size n.

- Since $z_{0.025} = -1.96$ and $z_{0.975} = 1.96$ at $\alpha = 0.05$ the confidence interval can be visualized by:
- The probability $1-\alpha$ can be interpreted as:
  $(1-\alpha)\times 100\%$ of the possible samples lead to confidence intervals that ***will cover the true but unknown population expectation*** μ:

General Rules

1. The ***smaller the error probability*** α (i.e., the confidence level $1-\alpha$ increases) is the ***wider the confidence interval*** becomes and *vice versa*. Example:

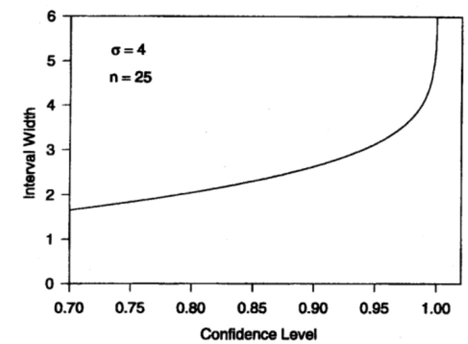


FIGURE 8-10. Effect of confidence level on interval width.

2. Furthermore, beside the error probability the ***width*** of the confidence interval also depends on the ***sample size*** through the standard error $\sigma/\sqrt{n}$ of the mean.
   As n increases, the confidence interval will shrink. Example:

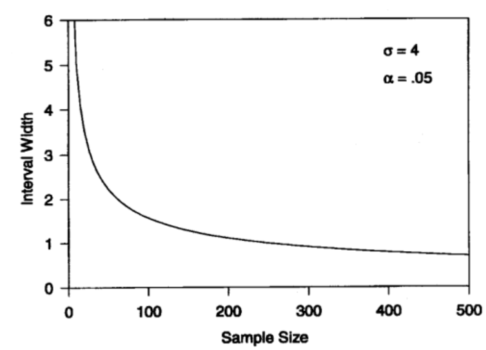


FIGURE 8-11. Effect of sample size on confidence interval width.

Example: Asymptotic Interval for the Expectation from a Non-normal Population

- In this case the underlying sampling distribution of the mean may become difficult to evaluate, however, we can make use of the central limit theorem for ***sufficiently large sample size*** n as then the mean becomes ***asymptotically*** normal distributed
- The required sample size depends on the underlying specific circumstances:
    1. for ***well-behaved population distributions*** it may be as low as $n \approx 30$ if just one parameter is estimated from the sample,
    2. for the ***binomial distribution*** it depends on the underlying probability level π. For π ***not too close*** to either end of its support $\pi \in [0,1]$, an $n \approx 100$ may be sufficient.

Example: Interval Estimation for the Expectation from a Normal Population with Unknown Standard Deviation

- The unknown population standard deviation σ must be replaced by a sample estimate of the standard deviation $S = \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \big/ (n-1)}$.

- Since $Z = (\bar{X} - \mu) \big/ (\frac{1}{\sqrt{n}} \cdot S)$ contains now ***two random variables*** (that is, $\bar{X}$ and S) the random variable Z will no longer follow a normal distribution.

  One can show that Z follows the t***-distribution with*** $n-1$ ***degrees for freedom***:

  1. Like the normal distribution also the t-distribution is ***symmetric***.
  2. For small degrees of freedom the t-distribution has substantially ***heavier tails***, that is, it has a ***positive kurtosis***.
  3. For $n > 30$, the t-distribution can be ***approximated*** by the normal distribution.

     $\Rightarrow$ The t-distribution approaches the ***standard normal distribution*** which does not use degrees of freedom:

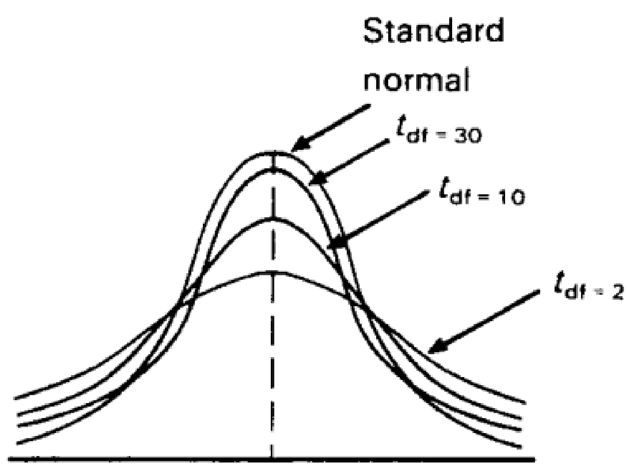


FIGURE 8-12. Normal distribution and the *t* distribution for 2, 10, and 30 df.

- The confidence interval for $n \leq 30$ becomes:

$$\Pr\left(\bar{x} + s/\sqrt{n} \cdot t_{\alpha/2,df=n-1} \le \mu \le \bar{x} + s/\sqrt{n} \cdot t_{1-\alpha/2,df=n-1}\right) = 1-\alpha\,,$$

where $t_{\alpha/2,df=n-1}$ is the lower quantile and $t_{1-\alpha/2,df=n-1}$ is the upper quantile.

- Note: For $n > 30$ the confidence interval based on the standard normal distribution with $z_{\alpha/2}$ and $z_{1-\alpha/2}$ can be used.

Example: Interval Estimation for the Success Probability of a Binomial Distributed Population

- Because the proportion estimator $\hat{\pi} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$ has the structure of an arithmetic mean, we can apply for sufficiently large n the central limit theorem.
- The proportion estimator is unbiased:

$$E(\hat{\Pi}) = \tfrac{1}{n} \cdot \sum_{i=1}^{n} E(X_i) \quad \text{with} \quad E(X_i) = 0 \cdot (1-\pi) + 1 \cdot \pi = \pi$$

$$= \tfrac{1}{n} \cdot \sum_{i=1}^{n} \pi = \tfrac{1}{n} \cdot n \cdot \pi = \pi$$

- Its variance is:

$$\begin{aligned} Var(\hat{\Pi}) &= Var\left(\tfrac{1}{n}\cdot\sum_{i=1}^{n} X_i\right) \\ &= \tfrac{1}{n^2}\cdot\sum_{i=1}^{n}\underbrace{Var(X_i)}_{=\pi\cdot(1-\pi)} \\ &= \tfrac{1}{n^2}\cdot n\cdot\pi\cdot(1-\pi) = \frac{\pi\cdot(1-\pi)}{n} \end{aligned}$$

Since the variance depends on the unknown population probability π, its ***estimated value*** $\hat{\pi}$ must ***substituted*** into the equation for the standard error: $s_{\pi} = \sqrt{\dfrac{\hat{\pi}\cdot(1-\hat{\pi})}{n}}$

- Using the normal approximation $\hat{\pi} \sim N\left(\pi, \frac{\hat{\pi}\cdot(1-\hat{\pi})}{n}\right)$ the confidence interval for population success probability π becomes

$$\Pr\left(\hat{\pi} + \sqrt{\frac{\hat{\pi}\cdot(1-\hat{\pi})}{n}}\cdot z_{\alpha/2} \le \pi \le \hat{\pi} + \sqrt{\frac{\hat{\pi}\cdot(1-\hat{\pi})}{n}}\cdot z_{1-\alpha/2}\right) = 1-\alpha$$

Summary of Confidence Intervals

TABLE 8-5
Summary of Point Estimators and Confidence Intervals for π and μ

Population parameter	Point estimator	Formula for confidence intervel	Appropriate conditions
μ	$\bar{X}$	$\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$	Exact for any sample size when population standard deviation is known and X normally distributed Approximate when X is not normally distributed but $n > 30$.
μ	$\bar{X}$	$\bar{x} \pm t_{\alpha/2,n-1f}(s/\sqrt{n})$	Exact when population standard deviation is unknown and X normally distributed Approximate when X is not normal but $n > 30$.
π	P	$p \pm z_{\alpha/2}\sqrt{p(1-p)/n}$	Approximate when $n > 100$.

Sample Size Determination

- Figure 8-11 showed that there is a relationship between the sample size n and the interval width of a confidence interval at a given confidence level $1-\alpha$.
  These components can be used to determine the required sample size for a given error probability and bounds.
- Steps:

1. Determine the estimator for population parameter for which we would like to determine the sample size and its distribution
2. Determine the ***precision*** in terms of the error E, which is half the interval width.
3. Determine the confidence level $1-\alpha$.

- Assuming a normal distribution
  1. For the population expectation from a normal distribution and a confidence level $1-\alpha$ we get:

     The confidence interval bounds are $\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$; therefore the error E becomes

     $$E = |z_{\alpha/2}| \cdot \sigma/\sqrt{n}$$

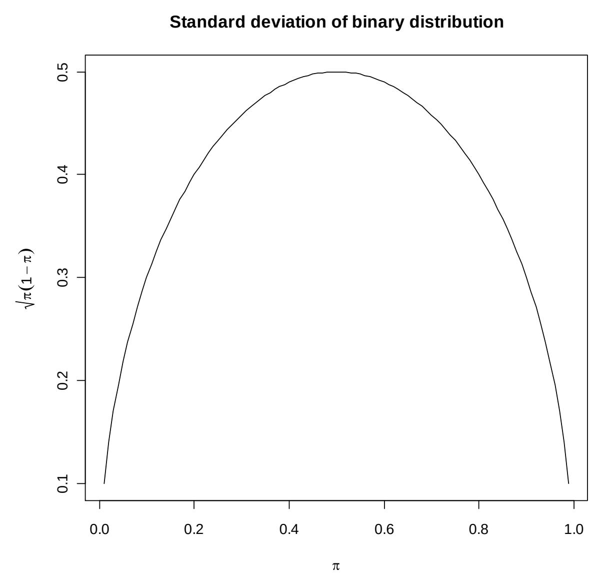
     Solving this expression for sample size leads to $n = \left(\frac{|z_{\alpha/2}| \cdot \sigma}{E} \right)^2$.

  2. To explore the effect of the error on the required sample size let us assume that error becomes twice as large: $E^* = 2 \cdot |z_{\alpha/2}| \cdot \sigma / \sqrt{n}$ we get as required sample size

     $$n = \left(\frac{|z_{\alpha/2}| \cdot \sigma}{2 \cdot E} \right)^2 = \frac{1}{4} \cdot \left(\frac{|z_{\alpha/2}| \cdot \sigma}{E} \right)^2 .$$

Thus just a ***quarter*** of sample data is needed if we are willing to make the error ***twice*** as large.

- General rules:
  1. As the error E ***decreases*** the sample size n ***increases***.
  2. As the error probability α ***decreases***, the critical tail value $|z_{\alpha/2}|$ becomes larger and, therefore, the required sample size n ***increases***.
- Assuming an unknown population variance of normally distributed random variables:
  1. This leads to critical values of the *t*-distribution
  2. The *t*-distribution depends on the sample size n. However, for a sample size of $n > 30$ a normal approximation can be used.
- Assuming a binomial distribution
  1. For the proportion estimator $\hat{\pi}$ of a binomial distribution we get $n = \left[\frac{|z_{\alpha/2}| \cdot \sqrt{\hat{\pi} \cdot (1-\hat{\pi})}}{E} \right]^2$



2. Since $\hat{\pi}$ is ***unknown*** before the sample is drawn, it must be determined exogenously (e.g., through past experience).
3. Alternatively, the worst case scenario of $\pi = 0.5$ can be used, for which the variance $\pi \cdot (1 - \pi)$ becomes the largest. This provides a conservative upper bound for n.
4. Note, the error E must be substantially smaller than $1/2$ in order for the interval covering the support $\pi \in \left]0,1\right[$.

Excursion: Confidence Intervals for Regression Coefficients

- The regression coefficients $b_0, b_1, b_2, \ldots$ are estimated of the true relationship

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \varepsilon$$

between the dependent variable and Y and the independent variables $X_1, X_2, \ldots$

- A sampling perspective from an underlying population can be applied:
  - Thus estimated regression coefficients are estimates of the population parameters.

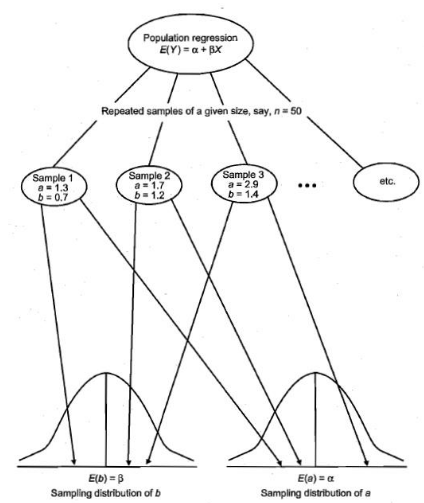


FIGURE 12-9. Sampling for a regression model.

  - Each estimated regression coefficient has a sampling distribution with a standard error.
- If the assumptions of regression analysis are satisfied then the estimated regression coefficients are unbiased: $E[b_j] = \beta_j$ for all independent variables X_j.
- The confidence intervals at a given error probability α can be calculated (technical details are covered in "Advanced Data Analysis"):

$$\Pr\left[\beta_j^{lower,\alpha} \leq \beta_j \leq \beta_j^{upper,\alpha}\right] = 1 - \alpha$$

- <u>Lay-person's interpretation (no statistical rigor):</u>
  - If the value 0 is within the confidence interval $0 \in \left[\beta_j^{lower,\alpha}, \beta_j^{upper,\alpha}\right]$, then this implies that that the true population parameter β_j is not be different from zero.
  - Consequently, the associated independent variable X_j has no influence on the variability of Y.
  - Explore the script **`StateSchoolConfint.r`**.