Estimation Theory

Confidence Interval Estimation

Benchikh Tawfik

Faculty of Medicine
1st Year Medicine

February 15, 2026

Plan de cours

INTRODUCTION

- Point estimation of a parameter θ provides a single numerical value, but gives no information about the accuracy of this estimate.
- In particular, it does not account for uncertainty due to sampling variability.
- To assess the reliability of an estimate, it is therefore necessary to associate it with an interval that contains the true value of the parameter with a specified probability. This is called *interval estimation* or *confidence interval estimation*.

Definition of a Confidence Interval

- Confidence interval estimation for a parameter θ consists in associating with a sample a random interval I, constructed in such a way that the probability that it contains the unknown true value of the parameter is equal to a predetermined level.
- Formally, we write:

$$\Pr(\theta \in I) = 1 - \alpha.$$

- The quantity $1 - \alpha$ is the probability that the interval contains the true value of the parameter. It is called the *confidence level*.

Probability Interval: Reminder

- Let X be a random variable with probability density function f.
- Given a probability level α, we choose two nonnegative numbers α_1 and α_2 such that

$$\alpha_1 + \alpha_2 = \alpha,$$

  and define two values x_1 and x_2 satisfying:

$$\Pr(X < x_1) = \alpha_1 \quad \text{and} \quad \Pr(X > x_2) = \alpha_2.$$

- The interval $I = [x_1, x_2]$ then contains an observed value of X with probability $1 - \alpha$.
- By neglecting the probability α, the distribution of X is summarized by restricting attention to values in the interval I. Such an interval is called a *probability interval* at level $1 - \alpha$, where α is known as the *critical level*.

Probability Interval: Principle

- The construction of a probability interval raises two fundamental questions:
  - What probability level α can reasonably be considered negligible?
  - For a given probability distribution and a fixed level α, infinitely many intervals $[x_1, x_2]$ satisfy the condition. How should α_1 and α_2 be chosen?
- The answers to these questions depend on the context and objectives of the problem under consideration.

Probability Interval: Example

- Suppose that a blood test measurement is a random variable X following a normal distribution $\mathcal{N}(100, 20)$.
- Values of X between two limits a and b are considered *normal* if

$$\Pr(a < X < b) = 0.95,$$

  while values outside this range are considered *pathological*.
- Knowing only the critical level $\alpha = 0.05$ is not sufficient to determine a and b, since infinitely many probability intervals satisfy this condition.
- However, regardless of the chosen interval, the probability of observing a pathological value remains equal to $\alpha = 0.05$.

Probability Interval: Example (continued)

- Assume now that the limits a and b are symmetric with respect to the mean $m = 100$.
- Introducing the standardized random variable

$$U = \frac{X - 100}{20},$$

we have:

$$\Pr(-1.96 < U < 1.96) = 0.95.$$

- Therefore,

$$a = 100 - 1.96 \times 20 = 60.8, \quad b = 100 + 1.96 \times 20 = 139.2,$$

and the corresponding probability interval is

$$\Pr(60.8 < X < 139.2) = 0.95.$$

- If low values of X are not considered pathological, only the upper bound $b = 139.2$ is retained.
- The probability of observing a pathological value then

CONSTRUCTION OF A CONFIDENCE INTERVAL

- Let X be a random variable whose probability density function $f(x;\theta)$ depends on an unknown parameter θ, and let

$$X = (X_1, \ldots, X_n)$$

be a sample of size n drawn from this distribution.

- Let $T = \varphi(X)$ be an estimator of the parameter θ, and let $g(t;\theta)$ denote the probability distribution of this estimator.
- Given a probability level α, and assuming the distribution of T is known, one can construct a probability interval for the random variable T of the form:

$$\Pr(\theta - h_1 < T < \theta + h_2) = 1 - \alpha. \tag{1.1}$$

Properties of Confidence Intervals

- A confidence interval is a *random interval*, since its bounds are random variables that depend on the observed sample.
- For a given level α, the values α_1 and α_2 must be specified, with $\alpha_1 + \alpha_2 = \alpha$.
- Their choice depends on the problem at hand and on the relative consequences of underestimating or overestimating the parameter.
- If $\alpha_1 = \alpha_2 = \alpha/2$, the resulting interval is a two-sided confidence interval with symmetric risks.
- One-sided confidence intervals can also be constructed, either with $\alpha_1 = 0$ or with $\alpha_2 = 0$.

Properties of Confidence Intervals

- **For fixed** values of the risk level α, the tail probabilities α_1 and α_2, and the sample size n, a confidence interval can be constructed for each possible sample.
- Among all such intervals, a proportion equal to α will fail to contain the true value of the parameter.
- The quantity α therefore represents the *risk* that the confidence interval does not include the true parameter value.
- The most desirable situation corresponds to choosing a small risk level α while keeping the interval length as short as possible.

Properties of Confidence Intervals

- The risk level α can be decreased, and in the limiting case $\alpha = 0$ one would obtain absolute certainty.
- However, in this case the confidence interval would extend over the entire parameter space: for example,

  $$(-\infty, +\infty)$$

  for a mean, or

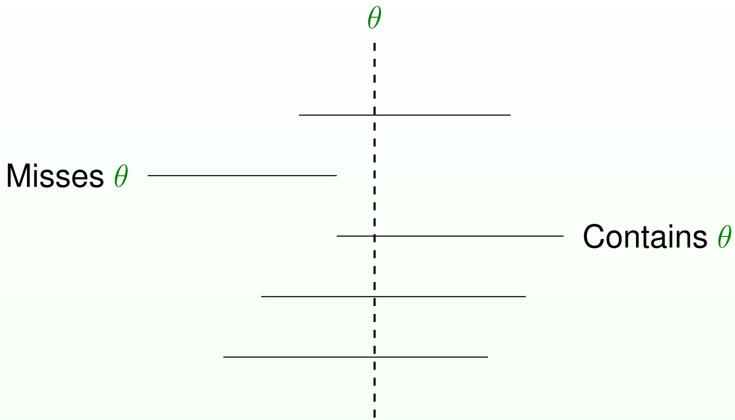
  $$[0, +\infty)$$

  for a standard deviation.
- Hence, decreasing α necessarily leads to an increase in the length of the confidence interval.
- In practice, an acceptable value of α is typically chosen (often $\alpha = 5\%$), and when possible, the sample size n is increased to improve precision.
- The probability $1 - \alpha$ is called the **confidence level** of the interval; it is associated with the interval itself and *not* with the unknown parameter value.
- Constructing a confidence interval therefore requires both a point estimator of the parameter and knowledge of its sampling distribution.

Graphical Illustration of a Confidence Interval

- Consider repeated random samples of size n drawn from the same population.
- For each sample, a confidence interval $I_n = [L_n, U_n]$ for the parameter θ is constructed.
- Graphically, this corresponds to a collection of random intervals on the real line.
- A proportion $1 - \alpha$ of these intervals contain the true parameter value θ, while a proportion α do not.
- The confidence level $1 - \alpha$ refers to this long-run frequency property, and not to the probability that a single realized interval contains θ.

The parameter θ is fixed, while the interval is random.



- Each horizontal segment represents a confidence interval from one sample.
- The vertical dashed line represents the true (fixed) parameter θ.
- Most intervals intersect θ, but some do not.

ESTIMATION AND CONFIDENCE INTERVAL FOR THE MEAN

- Let X be a random variable following a normal distribution $\mathcal{N}(m, \sigma)$. The parameters to be estimated are the mean m and the standard deviation σ.
- An unbiased estimator of the mean m is the sample mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

which follows the normal distribution

$$\overline{X} \sim \mathcal{N}\left(m, \frac{\sigma}{\sqrt{n}}\right).$$

- Two cases must be distinguished, depending on whether the standard deviation σ is known or unknown.

Case 1: Known Standard Deviation σ

- For a given significance level α, we construct a probability interval for the sample mean $\overline{X}$:

$$\Pr\left(m - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{X} < m + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha,$$

  where $u_{1-\alpha/2}$ is the $(1 - \alpha/2)$ quantile of the standard normal distribution:

$$\Pr(Z \leq u_{1-\alpha/2}) = 1 - \frac{\alpha}{2}, \quad Z \sim \mathcal{N}(0, 1).$$

- By inversion, we obtain the confidence interval for the mean m:

$$\Pr\left(\bar{x} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < m < \bar{x} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

- Hence,

$$IC_\alpha(m) = \left[\bar{x} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\ ;\ \bar{x} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Confidence Interval for the Mean, σ Known: Example

- Based on previous studies, the burst resistance of a certain type of tank is assumed to follow a normal distribution with unknown mean m and known standard deviation $\sigma = 4$ kg/cm^2.
- Tests on a sample of $n = 9$ tanks yield a sample mean resistance of
$$\bar{x} = 215 \text{ kg/cm}^2.$$
- The confidence level is chosen as $1 - \alpha = 0.95$.

Confidence Interval for the Mean, σ Known: Example

- Since
$$\Pr(-1.96 < Z < 1.96) = 0.95,$$
we obtain:
$$\Pr\left(215 - 1.96 \times \frac{4}{3} < m < 215 + 1.96 \times \frac{4}{3}\right) = 0.95.$$

- Numerically,
$$IC_{0.05}(m) = [212.39\,;\,217.61].$$

- This interval has a probability of 0.95 of containing the true mean burst resistance.

Remark

- For fixed α and σ, increasing the sample size n reduces the length of the confidence interval.
- Conversely, decreasing α (increasing the confidence level) increases the interval length.

Case 2: Unknown Standard Deviation σ

- An unbiased estimator of the variance is

$$S^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n}{n-1}S^2.$$

- The statistic

$$T_{n-1} = \frac{\overline{X} - m}{S^*/\sqrt{n}}$$

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

- Let $t_{(n-1;1-\alpha/2)}$ be such that

$$\Pr(T_{n-1} \leq t_{(n-1;1-\alpha/2)}) = 1 - \frac{\alpha}{2}.$$

Confidence Interval for the Mean, σ Unknown

- The $(1-\alpha)$ confidence interval for the mean m is:

$$\begin{aligned} IC_\alpha(m) &= \left[\overline{x} - t_{(n-1;1-\alpha/2)}\frac{s^*}{\sqrt{n}}\ ;\ \overline{x} + t_{(n-1;1-\alpha/2)}\frac{s^*}{\sqrt{n}}\right] \\ &= \left[\overline{x} - t_{(n-1;1-\alpha/2)}\frac{s}{\sqrt{n-1}}\ ;\ \overline{x} + t_{(n-1;1-\alpha/2)}\frac{s}{\sqrt{n-1}}\right] \end{aligned}$$

- When n is large ($n \geq 30$), the Student distribution can be approximated by the standard normal distribution.

Confidence Interval for the Mean, σ Estimated: Example

- In order to study the hourly daily wage (in DA) of workers in a given sector of activity, a non exhaustive random sample of size $n = 16$ is drawn. The following observations are obtained:

41	40	45	50	41	41	49	43
45	52	40	48	50	49	47	46

- It is assumed that the random variable "daily wage" follows a normal distribution with unknown mean m and unknown standard deviation σ.
- The sample mean is $\bar{x} = 45.44$, and the unbiased standard deviation is $s^* = \sqrt{\frac{n}{n-1}s^2} = \sqrt{\frac{16}{15}15.25} = 4.03$.

CONFIDENCE INTERVAL FOR THE MEAN, σ ESTIMATED: EXAMPLE

- For $\alpha = 0.05$, we use $t_{(15;0.975)} = 2.131$, i.e.

$$\Pr(T_{(15)} < 2.131) = 1 - \frac{0.05}{2} = 0.975 \qquad (i.e. t_{(15;0.975)} = 2.131)$$

- The confidence interval is:

$$IC_{0.05}(m) = \left[\bar{x} - \frac{s^*}{\sqrt{n}} t_{(n-1;1-\alpha/2)}; \bar{x} + \frac{s^*}{\sqrt{n}} t_{(n-1;1-\alpha/2)}\right] =$$

$$\left[45.438 - 2.131 \times 3.907/\sqrt{15}\,;\, 45.438 + 2.131 \times 3.907/\sqrt{15})\right]$$

$$= [43.29; 47.59]\ .$$

- $IC_{0.05}(m) = [43.29;\ 47.59]$ has a probability of 0.95 of containing the true mean daily wage of workers in this sector of activity.

- Left-sided interval:

$$IC_\alpha(m) = \left]-\infty \; ; \; \bar{x} + t_{(n-1;1-\alpha)} \frac{s^*}{\sqrt{n}}\right].$$

- Right-sided interval:

$$IC_\alpha(m) = \left[\bar{x} - t_{(n-1;1-\alpha)} \frac{s^*}{\sqrt{n}} \; ; \; +\infty\right[.$$

Case 1: Known Mean m

- As before, two cases must be distinguished, depending on whether the mean m is known or estimated.
- When the mean is known, the best estimator of the variance is the statistic

$$T = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2.$$

- The random variable $\frac{nT}{\sigma^2}$ follows a chi-square distribution with n degrees of freedom.
- A probability interval for the chi-square random variable $\chi^2(n)$ is given by (the bounds are read from the chi-square table):

$$\Pr\left(\chi^2_{\alpha/2}(n) < \chi^2(n) < \chi^2_{1-\alpha/2}(n)\right) = 1 - \alpha.$$

CASE 1: KNOWN MEAN m

- We deduce a two-sided confidence interval with symmetric risks for σ^2 (where t is the observed value of the statistic T):

$$\Pr\left(\frac{nt}{\chi^2_{1-\alpha/2}(n)} < \sigma^2 < \frac{nt}{\chi^2_{\alpha/2}(n)}\right) = 1-\alpha.$$

  where the value of $\chi^2_{1-\alpha/2}(n)$ is the upper critical value of the chi-square distribution with n degrees of freedom, such that

$$\Pr\left(\chi^2(n) \leq \chi^2_{1-\alpha/2}(n)\right) = 1-\frac{\alpha}{2}.$$

- That is,

$$IC_\alpha(\sigma^2) = \left[\frac{nt}{\chi^2_{1-\alpha/2}(n)}\,;\;\frac{nt}{\chi^2_{\alpha/2}(n)}\right].$$

Confidence Interval for the Variance, Known Mean: Example

- Let X be a random variable following the normal distribution $\mathcal{N}(40, \sigma)$.
- A sample of size $n = 25$ is drawn, and the value of the statistic T is computed.
- The observed value is $t = 12$.
- A two-sided confidence interval with confidence level $1 - \alpha = 0.95$ is required.

Confidence Interval for the Variance, Known Mean: Example

- Since $\frac{nT}{\sigma^2} \sim \chi^2(25)$, we have:

$$\Pr\left(\chi^2_{0.025}(25) < \chi^2(25) < \chi^2_{0.975}(25)\right) = \Pr(13.120 < \chi^2(25) < 40.644) = 0.95.$$

- Therefore,

$$IC_{0.05}(\sigma^2) = \left[\frac{25 \times 12}{40.644}\,;\,\frac{25 \times 12}{13.120}\right] = [7.381\,;\,22.866].$$

- The interval $[7.381; 22.866]$ contains the true variance with probability 0.95.
- Consequently, the interval $[2.716; 4.782]$ contains the true standard deviation with the same probability.

Case 2: Unknown Mean m

- The statistic

$$\frac{nS^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

  follows a chi-square distribution with $(n-1)$ degrees of freedom.
- The procedure is analogous to Case 1:

$$\Pr\left(\chi^2_{\alpha/2}(n-1) < \chi^2(n-1) < \chi^2_{1-\alpha/2}(n-1)\right) = 1 - \alpha.$$

- The bounds are obtained from the chi-square distribution table.

Confidence Interval for the Variance and Standard Deviation

- Two-sided confidence interval with symmetric risks:
  - For the variance:

$$IC_\alpha(\sigma^2) = \left[\frac{ns^2}{\chi^2_{1-\alpha/2}(n-1)}\ ;\ \frac{ns^2}{\chi^2_{\alpha/2}(n-1)}\right].$$

  - For the standard deviation:

$$IC_\alpha(\sigma) = \left[\sqrt{\frac{ns^2}{\chi^2_{1-\alpha/2}(n-1)}}\ ;\ \sqrt{\frac{ns^2}{\chi^2_{\alpha/2}(n-1)}}\right].$$

CONFIDENCE INTERVAL FOR THE VARIANCE, ESTIMATED MEAN: EXAMPLE "HOURLY DAILY WAGE"

- The statistic $\frac{nS^2}{\sigma^2}$ follows a chi-square distribution with $16 - 1 = 15$ degrees of freedom.
- We obtain: for $n = 16$ and $\alpha = 0.05$ (95% confidence level),

$$\chi^2_{1-\alpha/2}(n-1) = \chi^2_{0.975}(15) \approx 27.488.$$

and

$$\chi^2_{\alpha/2}(n-1) = \chi^2_{0.025}(15) \approx 6.262.$$

$$\Pr(6.262 < \chi^2(15) < 27.488) = 0.95.$$

- Hence, small

$$IC_{0.05}(\sigma^2) = \left[\frac{16 \times 15.246}{27.488} ; \frac{16 \times 15.246}{6.262}\right] = [8.874 ; 39.955].$$

- The interval $[8.874; 39.955]$ contains the true variance with probability 0.95, and $[2.98; 6.24]$ has the same property for

One-Sided Confidence Intervals for the Variance

- Right-sided intervals:

$$\Pr\left(0 < \sigma^2 < \frac{ns^2}{\chi^2_\alpha(n-1)}\right) = 1 - \alpha.$$

- Left-sided intervals:

$$\Pr\left(\frac{ns^2}{\chi^2_{1-\alpha}(n-1)} < \sigma^2\right) = 1 - \alpha.$$

REMARK

- If the sample size satisfies $n > 30$, the random variable

$$\sqrt{2\chi^2(n)} - \sqrt{2n-1}$$

  is approximately standard normal.
- The normal distribution table may then be used to approximate confidence bounds.

One-sided Confidence Interval for the Variance: Example

- Continuing with the previous example (daily wages), we determine a value a such that the 95% confidence level is satisfied:

$$\Pr(0 < \sigma < a) = 1 - \alpha\,.$$

$$\begin{aligned}
\Pr(\chi^2_{0.05}(15) < \chi^2(15)) &= \Pr(7.26 < \chi^2(15)) \\
&= \Pr\left(7.26 < \tfrac{16\times 15.264}{\sigma^2}\right) = 0.95
\end{aligned}$$

$$\Pr(\sigma^2 < 33.64) = \Pr(\sigma < 5.80) = 0.95$$

Sample Size Determination

- Assume that the lifetime of electric light bulbs follows a normal distribution with standard deviation $\sigma = 100$ hours.
- What is the minimum sample size required so that the 95% confidence interval for the mean lifetime has length less than 20 hours?
- The length of the two-sided confidence interval is:

$$2 \times 1.96 \times \frac{\sigma}{\sqrt{n}}.$$

- Solving

$$2 \times 1.96 \times \frac{100}{\sqrt{n}} = 20 \quad \Rightarrow \quad n = 385.$$

CONFIDENCE INTERVAL FOR THE MEAN: GENERAL CASE

- Regardless of the sample size, we use the unbiased estimators for the mean and variance:

$$\text{mean: } \overline{X}, \quad \text{variance: } S^{*2}.$$

- For large samples (practically, $n > 30$), the Central Limit Theorem allows us to use the same formulas as for the normal case to compute a confidence interval for the mean.
- In particular, if we want to estimate the expectation when the population variance is known, the confidence interval is the same as the one determined assuming the sample follows $\mathcal{N}\left(m, \frac{\sigma^2}{n}\right)$.
- For small samples, it is necessary to take into account the actual distribution of the studied variable.

Confidence Interval for the Mean, General Case: Example

- Consider a sample of 40 biscuit packages taken from a production batch of 2,000 units. The sample mean weight is $\bar{x} = 336$ g, and the sample standard deviation is $s = 0.86$ g.
- What is the 98% confidence interval for the mean weight of all packages in this production?
- The underlying distribution of package weights and the population variance are unknown. However, since the sample size is large ($n = 40 > 30$), the confidence interval can be approximated as:

$$\bar{x} - t\frac{s}{\sqrt{n-1}} < m < \bar{x} + t\frac{s}{\sqrt{n-1}},$$

with $\bar{x} = 336$, $s = 0.86$, and $n = 40$.

Confidence Interval for the Mean, General Case: Example

- The critical value t is obtained from the Student's t-distribution table with $n - 1 = 39$ degrees of freedom and depends on the chosen confidence level.
- Therefore, the 98% confidence interval for the true mean weight of the biscuit production is:

$$[335.666, 336.334].$$

INTRODUCTION

- Estimating a proportion arises in many real-world situations, for example:
  - the proportion of defective items in a production batch,
  - the proportion of voters who will support a given candidate,
  - the proportion of households that will purchase a new brand of detergent.

Confidence Interval for a Proportion

- Let p denote the proportion of individuals in a population exhibiting a certain characteristic C, which is unknown.
- The natural estimator of p is the sample proportion F, defined as:

$$F = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

where X_i is a Bernoulli random variable with parameter p:

$$X_i = \begin{cases} 1 & \text{if individual } i \text{ has characteristic } C, \\ 0 & \text{otherwise.} \end{cases}$$

Distribution and Confidence Interval

- Since $X_i \sim \text{Bernoulli}(p)$, the total number of successes $nF = \sum_{i=1}^{n} X_i$ follows a Binomial distribution $\mathcal{B}(n, p)$.
- Depending on n and p, the Binomial distribution can be approximated by different limiting distributions, which are used to construct a confidence interval. In practice, one can:
  - use statistical tables to find lower and upper limits of a confidence interval for given n and k (number of successes),
  - use the normal approximation with justification.

Confidence Interval via Normal Approximation

- For small n, one should use the exact Binomial tables (or charts).
- For sufficiently large n, such that $np > 5$ and $n(1-p) > 5$, by the Central Limit Theorem,

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(np, \sqrt{np(1-p)}\right),$$

so that the sample proportion F approximately follows

$$F \sim \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right),$$

and therefore

$$T = \frac{F - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}(0,1).$$

Confidence Interval via Normal Approximation

- Using the quantiles of the standard normal distribution:

$$\Pr\left(-U_{\frac{\alpha}{2}} < T < U_{\frac{\alpha}{2}}\right) = 1 - \alpha,$$

we obtain the confidence interval for p:

$$IC_{1-\alpha}(p) = \left[F - U_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}},\ F + U_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right].$$

- This interval has probability $1-\alpha$ of containing the true p, but it depends on p itself, which is unknown.
- In practice, we replace p with its estimate F, yielding:

$$IC_{1-\alpha}(p) = \left[f - U_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}, f + U_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}\right],$$

where f is the observed sample proportion.

Confidence Interval for a Proportion: Example

- In a random sample of 100 drivers, 25 of them own a car with an engine capacity exceeding 1600 cc.
- What is the confidence interval for the proportion of drivers owning a car with engine capacity greater than 1600 cc (two-sided interval, symmetric risk, confidence level 95%)?

Example: Solution

- Point estimate of p: Let K be the number of drivers owning a car with engine capacity > 1600 cc in a sample of size $n = 100$. Then $K \sim \text{Binomial}(n, p)$.
- An unbiased estimator for the proportion p is the sample proportion:

$$\hat{p} = f_n = \frac{K}{n}.$$

- From the data, the point estimate is:

$$\hat{p} = f_n = \frac{25}{100} = 0.25.$$

Example: Solution

- Using the normal approximation, the 95% confidence interval for p is:

$$\Pr\left(f_n - u_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} < p < f_n + u_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right) = 0.95$$

$$\Pr\left(f_n - 1.96\sqrt{\frac{p(1-p)}{n}} < p < f_n + 1.96\sqrt{\frac{p(1-p)}{n}}\right) = 0.95$$

- Replacing p by its estimate $f_n = 0.25$, we obtain:

$$\Pr(0.25-0.085 < p < 0.25+0.085) = \Pr(0.165 < p < 0.335) = 0.9$$

FINITE POPULATION CASE

- Most of the previous examples (estimating a proportion, mean, or variance) implicitly assumed an infinite population.
- Some results change when the population is finite, especially when the sample is a significant fraction of the population.

ESTIMATING THE MEAN m AND STANDARD DEVIATION σ

- Let X be a random variable defined on a finite population of size N.
- Since X takes only a finite number of values, it is effectively discrete.
- Consider a sample of size n drawn from this population.

Estimating the Mean and Variance

- **Sampling with replacement:**
  - Unbiased estimator of the mean:

$$\overline{X}, \quad \mathbb{E}(\overline{X}) = m, \quad \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$$

  - Unbiased estimator of the variance:

$$S^{*2} = \frac{n}{n-1} S^2$$

- **Sampling without replacement:**
  - Unbiased estimator of the mean:

$$\overline{X}, \quad \mathbb{E}(\overline{X}) = m, \quad \mathrm{Var}(\overline{X}) = \frac{N-n}{N-1} \frac{\sigma^2}{n}$$

  - Unbiased estimator of the variance:

$$\frac{N-1}{N} \frac{n}{n-1} S^2$$

Estimating a Proportion in a Finite Population

- **Sampling with replacement:**
  - The sample proportion f is an unbiased estimator of p:

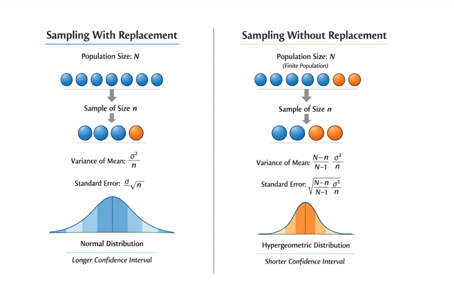
$$\mathbb{E}(f) = p, \quad \text{Var}(f) = \frac{p(1-p)}{n}$$

- **Sampling without replacement:**
  - The appropriate distribution is the hypergeometric distribution.
  - The sample proportion f remains an unbiased estimator of p:

$$\mathbb{E}(f) = p, \quad \text{Var}(f) = \frac{N-n}{N-1}\frac{p(1-p)}{n}$$

  - Confidence intervals can be determined using special tables or statistical software.

Estimating a Proportion in a Finite Population



Example: Proportion in a Finite Population

- A factory produces $N = 500$ items, and we want to estimate the proportion p of defective items.
- We randomly select $n = 50$ items **without replacement** and observe $x = 5$ defective items.
- Sample proportion:

$$f = \frac{x}{n} = \frac{5}{50} = 0.10$$

- Variance adjusted for finite population:

$$\text{Var}(f) = \frac{N-n}{N-1}\frac{p(1-p)}{n} \approx \frac{500-50}{499}\frac{0.10 \cdot 0.90}{50} \approx 0.00163$$

- Standard deviation: $\sqrt{0.00163} \approx 0.0404$
- Approximate 95% confidence interval:

$$f \pm 1.96 \cdot 0.0404 \implies [0.02, 0.18]$$

- Interpretation: With 95% confidence, the true proportion of defective items is between 2% and 18%.

EXERCISE 1

As part of a workplace health study, a random sample of 500 employees from different sectors and regions of Algeria was surveyed. Among them, 145 reported having experienced workplace bullying.

1. Identify the population, the variable of interest, its type, and its parameter(s).
2. Provide a point estimate of the proportion of employees who have experienced workplace bullying.
3. Construct a 90% confidence interval for this proportion.
4. If a 95% confidence interval were calculated using the same data, would it be larger or smaller than the 90% interval? Explain without performing calculations.

Exercise 1 - Solutions

1. Population: all employees in Algeria from the sectors studied.
   Variable: whether an employee has experienced workplace bullying (Yes/No).
   Type: categorical (binary).
   Parameter: proportion of employees who have experienced bullying, p.
2. Point estimate of the proportion:

$$\hat{p} = \frac{\text{number of employees reporting bullying}}{\text{sample size}} = \frac{145}{500} = 0.29$$

3. 90% Confidence Interval for p using normal approximation:

$$CI = \hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

   Here, $\hat{p} = 0.29$, $n = 500$, $z_{0.95} \approx 1.645$:

$$SE = \sqrt{\frac{0.29 \times 0.71}{500}} \approx 0.0203$$

Exercise 2

To design a rehabilitation program, researchers administered a cognitive neuropsychology questionnaire to a random sample of 150 dyslexic children. The questionnaire contains 20 questions. For each child, the number of correct answers x_i was recorded. The collected data satisfies:

$$\sum_i x_i = 1502, \quad \sum_i x_i^2 = 19486.$$

1. The statistical population studied is:
   (A) the researchers (B) the rehabilitation program (C) the dyslexic children (D) the questionnaire (E) the number of correct answers.
2. The statistical variable X being studied is:
   (A) the researchers (B) the rehabilitation program (C) the dyslexic children (D) the questionnaire (E) the number of correct answers.

Exercise 2 - Questions

3. The possible values of X are:
(A) only $\{20\}$ (B) $\{0, 1, \ldots, 20\}$ (C) $\{0, 1, \ldots, 50\}$
(D) {correct, incorrect} (E) A, B, C, and D are incorrect.
4. The sample size is:
(A) 20 (B) 150 (C) 1502 (D) 3000 (E) 19486.
5. The unbiased point estimate of the population mean μ is:
(A) 1502 (B) 10.8 (C) 10.01 (D) 75.01 (E) 79.05.
6. The unbiased point estimate of the population variance is:
(A) 28.3 (B) 29.17 (C) 29.7 (D) 29.9 (E) 30.59.

Exercise 2 - Confidence Interval Questions

7. The population mean has a 95% probability of lying within which interval?
(A) [9.13, 10.89] (B) [9.27, 10.75] (C) [10.01, 12.29]
(D) [3.77, 16.25] (E) [8.87, 11.15]
8. The probability that the population mean lies within the interval [9.27, 10.75] is:
(A) 0.05 (B) 0.1 (C) 0.9 (D) 0.95 (E) 0.99
9. The margin of error for estimating the population mean at a 99% confidence level is:
(A) 0.99 (B) 0.95 (C) 0.88 (D) 6.24 (E) 1.16

Exercise 2 - Solutions (Step-by-step)

Given data: $n = 150$, $\sum_i x_i = 1502$, $\sum_i x_i^2 = 19486$, X_i = number of correct answers per child.

1. Population: the 150 dyslexic children.
   Variable: number of correct answers X_i.
   Sample size: $n = 150$.
2. Sample mean (unbiased estimator of population mean):

$$\overline{X} = \frac{\sum_i x_i}{n} = \frac{1502}{150} = 10.0133 \approx 10.01$$

3. Sample variance (biased):

$$S^2 = \frac{1}{n}\sum_i (x_i - \overline{X})^2 = \frac{\sum_i x_i^2}{n} - (\overline{X})^2 = 29.641$$

   Sample variance (unbiased):

$$S^*2 = \frac{n}{n-1}S^2 \approx 29.83$$

Exercise 2 - Solutions (Step-by-step)

4. 95% confidence interval for the mean:

$$CI = \overline{X} \pm t_{1-\alpha/2,n-1} \frac{S}{\sqrt{n}}$$

Using $t_{0.975,149} \approx 1.976$, $S \approx \sqrt{29.82} \approx 5.46$, $n = 150$:

$$SE = \frac{5.46}{\sqrt{150}} \approx 0.445$$

$$CI = 10.01 \pm 1.976 \times 0.445 \approx 10.01 \pm 0.879$$

$$\boxed{CI_{95\%} \approx [9.13, 10.89]}$$

5. Margin of error for 99% confidence interval ($t_{0.995,149} \approx 2.61$):

$$ME = t \cdot SE = 2.61 \times 0.445 \approx 1.16$$

$$CI_{99\%} \approx [10.01 - 1.16, 10.01 + 1.16] \approx [8.85, 11.17]$$