Mathematical Statistics 2, Part II: Confidence intervals and regions

Statistics Team TSE

2022–2023

Syllabus

1. Confidence intervals
   - Definition
   - Examples
2. Pivotal quantities
   - Definition
   - Examples
   - From pivotal statistics to CI
3. Large-sample CIs
   - General idea
   - Variance stabilization strategy
   - CIs for a difference of two means

Syllabus

What is a confidence interval ?

Consider a statistical model, indexed by a parameter $\theta \in \mathbb{R}^1$.
Denote the observation as $X = (X_1, \ldots, X_n)$.

Fix $\alpha \in (0, 1)$, which yields a confidence level $1 - \alpha$.

Definition : A confidence interval (CI) for θ at confidence level $1 - \alpha$ is an interval $[LB(X), UB(X)]$ such that

- $LB(X)$ and $UB(X)$ are statistics (i.e., measurable functions of X)
- $\mathbb{P}_\theta[\theta \in [LB(X), UB(X)]] \geq 1 - \alpha$ for any θ.

In general, we will denote the CI as $\mathcal{I}_{n,\alpha}$ when $X = (X_1, \ldots, X_n)$.

What is a confidence region ?

Consider a statistical model, indexed by a parameter $\theta \in \mathbb{R}^d$.
Denote the observation as $X = (X_1, \ldots, X_n)$. Fix $\alpha \in (0, 1)$, which yields a confidence level $1 - \alpha$.

For $\theta \in \mathbb{R}^d$, one may define confidence regions (ellipsoids, rectangles,...)

Definition : A confidence region (CR) for θ at confidence level $1 - \alpha$ is a set $C(X)$ such that :

- $C(X)$ is a measurable functions of X
- $\mathbb{P}_\theta[\theta \in C(X)] \geq 1 - \alpha$ for any θ.

First example : Confidence interval for a proportion

Let $(X_1, \ldots, X_n)$ be a random sample from a Bernoulli distribution $\mathcal{B}(\theta)$ with $\theta \in \Theta = [0,1]$.

- We denote by $\bar{X}_n$ the mean number of success :

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

- The Bienayme-Tchebychev inequality yields

$$\mathbb{P}_\theta\left(|\bar{X}_n - \theta| \geq \delta\right) \leq \delta^{-2} Var(\bar{X}_n) = \frac{\theta(1-\theta)}{n\delta^2} \leq \frac{1}{4n\delta^2}.$$

- For any $\alpha \in (0,1)$ and $\theta \in \Theta$, we obtain

$$\mathbb{P}_\theta(\theta \in \mathcal{I}_{n,\alpha}) \geq 1-\alpha \quad \text{with} \quad \mathcal{I}_{n,\alpha} := \left[\bar{X}_n \pm \frac{1}{2\sqrt{n\alpha}}\right]$$

Second example : Confidence interval for the mean of a Gaussian distribution

Here we change the notation θ into μ (as it refers to the mean). Let $(X_1, \ldots, X_n)$ be a random sample from the $\mathcal{N}(\mu, 1)$ distribution. Denote as $z_\beta = \Phi^{-1}(\beta)$ the β-quantile of the standard normal.

Since $\bar{X}_n \sim \mathcal{N}(\mu, \frac{1}{n})$, we then have :

$$\mathbb{P}_\mu\left[- z_{1-\alpha/2} \leq \frac{\bar{X}_n - \mu}{\frac{1}{\sqrt{n}}} \leq z_{1-\alpha/2}\right] = 1 - \alpha,$$

which rewrites with $\mathcal{I}_{n,\alpha} = \left[\bar{X}_n \pm z_{1-\alpha/2}\right]$: $\mathbb{P}_\mu\left[\mu \in \mathcal{I}_{n,\alpha}\right] = 1 - \alpha$.

This is a CI for μ, that is centered at $\bar{X}$ (the estimator we adopted here).

Second example : illustration

Take $n = 50$ and $\mu = 2$.

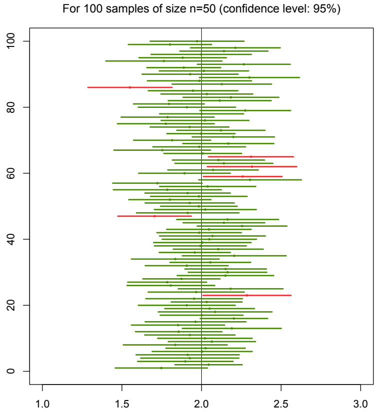
One given realization of $X_1, \ldots, X_n$ yields $\bar{X} = 2.038$ and the CIs below :

$1-\alpha$	$\alpha/2$	$z_{1-\alpha/2}$	LB	UB	UB-LB
0.90	0.05	1.64	1.80	2.27	0.465
0.95	0.025	1.95	1.76	2.31	0.554
0.99	0.005	2.57	1.67	2.40	0.730

If the confidence level increases, then the CI has a length that increases.

Second example : interpretation

Would we draw many samples $(X_1, \dots, X_{50}$ from $\mathcal{N}(\mu = 2, 1))$, then the proportion of CIs containing the true value $\mu = 2$ would be $\approx 1 - \alpha$.



Notation :

We will often write

$$\bar{X}_n \pm \frac{z_{1-\alpha/2}}{\sqrt{n}}$$

instead of

$$[\bar{X}_n - \frac{z_{1-\alpha/2}}{\sqrt{n}}, \bar{X}_n + \frac{z_{1-\alpha/2}}{\sqrt{n}}]$$

Final remarks

The two previous examples and contructions rely on an inequality “in probability”. To obtain such inequalities, several solutions

- Explicit knowledge of some distributions of some random variables : Gaussian example
- Use standard inequalities (Markov, Bienayme-Tchebychev, Chernoff, Hoeffding, ...) : Proportion example
- Use large sample properties and convergence in distributions (Central Limit Theorem) : see below

Syllabus

Pivotal quantities - 1D situation

Definition : A random variable $Q = q(X, \theta)$ is a pivotal quantity for θ if

- for all x, the function $\theta \mapsto q(x, \theta)$ is monotone ($\nearrow$ or $\searrow$)
- the distribution of Q does not depend on θ.

We could extend this definition to the multivariate case while omitting the monotonicity condition.

Big Warning !

- In the previous definition, we do not ask for a random variable Q that does not depend on θ !
- Of course Q certainly depends on θ.
- The definition is about the law of Q that regardless the value of θ, the distribution of Q under the distribution $\mathbb{P}_\theta$ is independent from θ.
- F_Q will be the cdf of Q.

Example 1 - Gaussian distribution

To make things easier to understand, let us discuss on a first example.

If $(X_1, \ldots, X_n)$ is a random sample from the $\mathcal{N}(\mu, 1)$ distribution, then

$$Q = \frac{\bar{X} - \mu}{\frac{1}{\sqrt{n}}}$$

is a pivotal quantity for μ.

- Monotonicity ?
- Distribution of Q ?

Example 2 - Uniform distribution

To make things easier to understand, let us discuss on a second example.

If $(X_1, \ldots, X_n)$ is a random sample from the uniform model $\mathcal{U}([0, \theta])$, $\theta \in \Theta = \mathbb{R}_+^*$, then :

$$Q = \frac{\max(X_1, \ldots, X_n)}{\theta}$$

is a pivotal quantity for θ.

- Monotonicity ?
- Distribution of Q ?

Construction of a CI from a pivotal quantity

(1) Since the distribution of $Q = q(X, \theta)$ does not depend on θ, we have

$$\mathbb{P}_\theta[F_Q^{-1}(\alpha/2) \leq q(X,\theta) \leq F_Q^{-1}(1-\alpha/2)] = 1-\alpha,$$

where $F_Q(t) = P_\theta[Q \leq t]$ is the cdf of Q.

(2) The monotonicity assumption then allows us to write this as

$$\mathbb{P}_\theta[LB(X) \leq \theta \leq UB(X)] = 1-\alpha.$$

This second step is called "inverting the interval".

CI for the mean of a $\mathcal{N}(\mu, \sigma^2)$ distribution, σ^2 known

Let $(X_1, \ldots, X_n)$ be a random sample from the $\mathcal{N}(\mu, \sigma^2)$ distribution, with σ^2 known. Clearly,

$$Q = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

is a pivotal quantity for μ. For any $u \in (0, \alpha)$, we have

$$\mathbb{P}_\mu\left[z_{\alpha-u} \leq \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-u}\right] = 1 - \alpha,$$

which rewrites

$$\mathbb{P}_\mu\left[\bar{X}_n - z_{1-u}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n - z_{\alpha-u}\frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha.$$

The CI length is minimized for $u = \frac{\alpha}{2}$, which yields $CI = \bar{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$.

CI for the length of the support of $\mathcal{U}([0,\theta])$

If $(X_1,\ldots,X_n)$ is a random sample from the uniform model $\mathcal{U}([0,\theta])$, $\theta \in \Theta = \mathbb{R}_+^*$, then :

$$Q = \frac{\max(X_1,\ldots,X_n)}{\theta}$$

is a pivotal quantity for θ.
An easy computation shows that : $\forall t \in (0,1),\ \mathbb{P}_\theta[Q \le t] = F_Q(t) = t^n$.
For any $\alpha \in (0,1)$, we define $t_{\alpha,n}$ s.t. : $t_{\alpha,n}^n = \alpha$, *i.e.* $t_{\alpha,n} = \alpha^{1/n}$.
The CI is obtained with $\mathbb{P}_\theta[1 \ge Q \ge t_{\alpha,n}] = 1-\alpha$:

$$\theta \in [\max(X_1,\ldots,X_n), \max(X_1,\ldots,X_n)\alpha^{-1/n}].$$

CI for the mean of a $\mathcal{N}(\mu, \sigma^2)$ distribution, σ^2 unknown

If σ is unknown, then this does not provide a valid CI.

But, if $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$ is the usual unbiased estimator of σ^2, then $Q = \frac{\bar{X}_n - \mu}{\frac{s}{\sqrt{n}}} (\sim t_{n-1})$ is a pivotal quantity for μ. Thus, for any $u \in (0, \alpha)$,

$$\mathbb{P}_{\mu,\sigma^2}\left[t_{n-1,\alpha-u} \leq \frac{\bar{X}_n - \mu}{\frac{s}{\sqrt{n}}} \leq t_{n-1,1-u}\right] = 1 - \alpha,$$

which rewrites

$$\mathbb{P}_{\mu,\sigma^2}\left[\bar{X}_n - t_{n-1,1-u}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X}_n - t_{n-1,\alpha-u}\frac{s}{\sqrt{n}}\right] = 1 - \alpha.$$

Again, the CI obtained for $u = \frac{\alpha}{2}$ has minimal length.

Distribution of Q

Reminder : $T \sim t_k$ (or $Stu(k)$, Student with k degrees of freedom) iff T has the same distribution as $Z/\sqrt{k^{-1}W}$, where

- $Z \sim \mathcal{N}(0,1)$,
- $W \sim \chi^2_k$, and
- Z and W are independent.

In the previous slide, $Q = \frac{\bar{X}_n - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$ since

- $Z = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$,
- $W = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$, and
- Z and W are independent.

CI for the variance of a Gaussian distribution

Let $(X_1, \ldots, X_n)$ be a random sample from the $\mathcal{N}(\mu, \sigma^2)$ distribution, with both μ and σ^2 unknown.

Exercise :

(i) Show that $\frac{(n-1)s^2}{\sigma^2}$ is a pivotal quantity for σ^2 (recall the previous slide !)

(ii) Check that a resulting CI for σ^2 at confidence level $1 - \alpha$ is

$$\left[\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}\right].$$

Here, working with symmetric tail probabilities does not minimize length (see JASA 1959, vol 54, page 674 for a minimal-length CI).

Syllabus

Large-sample CIs

Finding a pivotal quantity (with a known distribution) is often difficult !

Assume that $(\hat{\theta}_n)$ is such that $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2(\theta))$. Then,

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma(\theta)}$$

possibly qualifies as an "asymptotic pivotal quantity", leading to

$$\mathbb{P}_\theta\left[-z_{1-\alpha/2} \leq \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma(\theta)} \leq z_{1-\alpha/2}\right] \to 1 - \alpha.$$

If inversion is possible, then this yields an asymptotic CI for θ at confidence level $1 - \alpha$.

Large-sample CIs

If inversion is not possible, then, under minimal assumptions on $\sigma(\theta)$,

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma(\hat{\theta}_n)} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

which leads to

$$\mathbb{P}_\theta\left[-z_{1-\alpha/2} \leq \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma(\hat{\theta}_n)} \leq z_{1-\alpha/2}\right] \to 1-\alpha.$$

This can always be inverted into

$$\mathbb{P}_\theta\left[\hat{\theta}_n - z_{1-\alpha/2}\frac{\sigma(\hat{\theta}_n)}{\sqrt{n}} \leq \theta \leq \hat{\theta}_n + z_{1-\alpha/2}\frac{\sigma(\hat{\theta}_n)}{\sqrt{n}}\right] \to 1-\alpha.$$

Large-sample CIs : Example 1

Let $X = (X_1, \ldots, X_n)$ be a random sample from the density

$$f_\theta(x) = \theta \exp(-\theta x) \mathbb{1}_{[0,\infty)}(x),$$

with $\theta > 0$ (exponential with mean $1/\theta$).

The MLE of θ, namely $\hat{\theta}_n = 1/\bar{X}_n$, satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2(\theta) = \theta^2),$$

which yields

$$\mathbb{P}_\theta\left[- z_{1-\alpha/2} \leq \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\theta} \leq z_{1-\alpha/2} \right] \to 1 - \alpha.$$

Large-sample CIs : Example 1

Here, inversion is possible :

$$\mathbb{P}_\theta\left[-z_{1-\alpha/2} \leq \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\theta} \leq z_{1-\alpha/2}\right] \to 1-\alpha$$

is inverted into the asymptotic CI

$$\mathbb{P}_\theta\left[\frac{\hat{\theta}_n}{1+\frac{z_{1-\alpha/2}}{\sqrt{n}}} \leq \theta \leq \frac{\hat{\theta}_n}{1-\frac{z_{1-\alpha/2}}{\sqrt{n}}}\right] \to 1-\alpha.$$

Large-sample CIs : Example 2

Let $X = (X_1, \dots, X_n)$ be a random sample from the Bernoulli distribution with mean θ. The MLE of θ, namely $\hat{\theta}_n = \bar{X}$, satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2(\theta) = \theta(1-\theta)),$$

which yields

$$\mathbb{P}_\theta\left[-z_{1-\alpha/2} \leq \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sqrt{\theta(1-\theta)}} \leq z_{1-\alpha/2}\right] \to 1-\alpha.$$

Wilson method. Inversion is possible and yields the asymptotic CI

$$\mathrm{CI} = \left(\hat{\theta}_n + \frac{z^2_{1-\alpha/2}}{2n} \pm \frac{z_{1-\alpha/2}}{\sqrt{n}}\sqrt{\hat{\theta}_n(1-\hat{\theta}_n) + \frac{z^2_{1-\alpha/2}}{4n}}\right) \Big/ \left(1 + \frac{z^2_{1-\alpha/2}}{n}\right).$$

Large-sample CIs : Example 2

However this formula is complex, which motivates the second method.

Wald method. Inverting instead

$$\mathbb{P}_\theta\left[-z_{1-\alpha/2} \leq \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}} \leq z_{1-\alpha/2}\right] \to 1-\alpha$$

yields the simpler asymptotic CI

$$\mathrm{CI} = \bar{X} \pm z_{1-\alpha/2}\frac{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}}{\sqrt{n}}.$$

If the lower (upper) bound is < 0 (> 1), then we replace it by 0 (1).
In the `binom` package, the method is called "asymptotic".

Large-sample CIs : variance stabilization

Main idea : Use the CLT and the Delta method jointly !

- Assume that we know :

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2(\theta)),$$

- Consider ϕ a smooth function of θ, then the Delta method yields :

$$\sqrt{n}(\phi(\hat{\theta}_n) - \phi(\theta)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, (\phi'(\theta))^2\sigma^2(\theta)),$$

- **Leading nice idea :** choose ϕ such that the limiting variance factor is independent from θ, *i.e.* for example, choose ϕ :

$$(\phi'(\theta))^2\sigma^2(\theta) = 1$$

- The constant 1 above may be replaced by any constant number.
- Finally, use a CI of the Gaussian and then "inverse" the ϕ application.

Large-sample CIs : variance stabilization - Example 1

Consider a Bernoulli model $X = (X_1, \ldots, X_n)$ based on $\mathcal{B}(\theta)$.

- We compute the limiting variance factor :

$$\sigma^2(\theta) = \theta(1-\theta)$$

- We solve the differential equation :

$$(\phi'(\theta))^2\theta(1-\theta) = 1 \Longleftarrow \phi(\theta) = 2\arcsin(\sqrt{\theta})$$

- We obtain the CI : $\sqrt{n}(\phi(\hat{\theta}_n) - \phi(\theta)) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$.

Exercise : check that this leads to the asymptotic CI

$$\mathrm{CI} = \left[\sin^2\left(\arcsin\sqrt{\hat{\theta}_n} - \frac{z_{1-\alpha/2}}{2\sqrt{n}}\right), \sin^2\left(\arcsin\sqrt{\hat{\theta}_n} + \frac{z_{1-\alpha/2}}{2\sqrt{n}}\right)\right].$$

Large-sample CIs : variance stabilization - Example 2

Gaussian model $X = (X_1, \ldots, X_n)$ based on $\mathcal{N}(0, \sigma^2)$, CI on σ^2 ?

- CLT application : define $\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n X_i^2$ and observe that :

$$\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, Var(X^2)\right).$$

- We compute the limiting variance factor :

$$Var(X^2) = 2\sigma^4$$

- We solve the differential equation :

$$2(\phi'(\sigma^2))^2\sigma^4 = 2 \Longleftarrow \phi(t) = \log t$$

- We obtain the CI : $\sqrt{n}(\log(\hat{\sigma}_n^2) - \log \sigma^2) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, 2\right)$.

Exercise : What is the CI obtained in this way ?

Syllabus

CIs for the difference of two means

Several cases

- Independent Gaussian samples
  1. known variances
  2. unknown, equal, variances
- Independent samples, large sample size
  1. known variances
  2. unknown, equal, variances
- Dependent samples : matched pairs experiment

Independent Gaussian samples ; known variances

Let $X_1, \ldots, X_{n_1}$ i.i.d. $\mathcal{N}(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_{n_2}$ i.i.d. $\mathcal{N}(\mu_2, \sigma_2^2)$ be two independent samples, with σ_1^2, σ_2^2 known.

Building a CI for $\mu_1 - \mu_2$ is based on the pivotal quantity

$$Q = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

and leads to

$$\mathrm{CI} = \bar{X} - \bar{Y} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Independent Gaussian samples ; unknown, equal, variances

Consider the case where σ_1^2, σ_2^2 are unknown.

Then, under the additional assumption $\sigma_1^2 = \sigma_2^2$, the pooled estimator

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{\sum_{i=1}^{n_1}(X_i - \bar{X})^2 + \sum_{i=1}^{n_2}(Y_i - \bar{Y})^2}{n_1 + n_2 - 2}$$

is an unbiased estimator of $\sigma^2 (\stackrel{\text{def}}{=} \sigma_1^2 = \sigma_2^2)$, which follows, e.g., from

$$\begin{aligned}\frac{(n_1 + n_2 - 2)s_p^2}{\sigma^2} &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{\sigma^2} \\ &= \frac{(n_1 - 1)s_1^2}{\sigma_1^2} + \frac{(n_2 - 1)s_2^2}{\sigma_2^2} \sim \chi^2_{n_1+n_2-2}.\end{aligned}$$

Independent Gaussian samples ; unknown, equal, variances

The construction is then based on the pivotal quantity

$$Q = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} \sim t_{n_1+n_2-2}$$

and leads to

$$\mathrm{CI} = \bar{X} - \bar{Y} \pm t_{n_1+n_2-2,1-\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\,.$$

Independent samples, large sample sizes

Case 1 : known variances

The same quantity Q as for the Gaussian case (known variances) is now asymptotically pivotal.

We obtain the same expression for the CI as for the Gaussian case with known variances.

Case 2 : equal, unknown, variances

The same quantity Q as for the Gaussian case (unknown variances) is now asymptotically pivotal.

The expression for the CI is obtained from the one in the Gaussian case with unknown equal variances by replacing t-quantiles with Gaussian ones.

Dependent samples, matched pairs

When X_i and Y_i are dependent because, e.g., they are measured on the same subject, we work on the differences $D_i = X_i - Y_i$.

The variance $\sigma_D^2 = \mathrm{Var}[D_i] = \mathrm{Var}[X_i] + \mathrm{Var}[Y_i] - 2\mathrm{Cov}[X_i, Y_i]$ is usually smaller than in the independent case (differences between twins tend to be smaller than between independently selected people).

With $s_D^2 = \frac{1}{n-1}\sum_{i=1}^n (D_i - \bar{D})^2$ (an unbiased estimator of σ_D^2),

$$Q = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\frac{s_D}{\sqrt{n}}} \ (\sim t_{n-1})$$

is a pivotal quantity for $\mu_1 - \mu_2$, which leads to

$$\mathrm{CI} = \bar{X} - \bar{Y} \pm t_{n-1,1-\alpha/2}\frac{s_D}{\sqrt{n}}.$$