Variance Stabilizing Transformations

Main Idea:

Suppose we wish to construct a CI for an unknown population parameter θ on the basis of a random sample $(X_1, \ldots, X_n)$, and $\widehat{\theta}_n = \widehat{\theta}_n(X_1, \ldots, X_n)$ is an estimator for θ. If $Var\left(\widehat{\theta}_n\right)$ is a function of the unknown parameter θ, the goal is to find a transformation g, such that $Var\left(g\left(\widehat{\theta}_n\right)\right)$ does not depend on θ. Then one can often construct a CI for $g(\theta)$, and then convert it into a CI for θ itself.

Motivation:

Ex: Let $X_1, X_2, \ldots, X_n$ be iid $Bernoulli(\theta)$.
The goal is to construct a CI with the confidence level $1-\alpha$ for an unknown parameter θ. The population mean is equal to θ, the population variance $\sigma^2 = \theta(1-\theta)$. A natural choice of $\widehat{\theta}_n$ is $\widehat{\theta}_n = \overline{X}_n$. Then

$$E(\widehat{\theta}_n) = E(\overline{X}_n) = \theta, \quad Var\left(\widehat{\theta}_n\right) = Var(\overline{X}_n) = \frac{\sigma^2}{n} = \frac{\theta(1-\theta)}{n}\,.$$

By the CLT,

$$\sqrt{n}(\overline{X}_n - \theta) \xrightarrow{d} N(0, \theta(1-\theta)), \quad \text{as} \quad n \to \infty,$$

or, equivalently,

$$\frac{\sqrt{n}(\overline{X}_n - \theta)}{\sqrt{\theta(1-\theta)}} \xrightarrow{d} N(0,1), \quad \text{as} \quad n \to \infty.$$

Thus, for a large sample size n,

$$P\left(-z_{\alpha/2} \le \frac{\sqrt{n}(\overline{X}_n - \theta)}{\sqrt{\theta(1-\theta)}} \le z_{\alpha/2}\right) \approx 1-\alpha,$$

where

$$P(Z > z_{\alpha/2}) = \frac{\alpha}{2} \quad \text{for} \quad Z \sim N(0,1).$$

In other words,

$$P\left(\theta \in \left[\overline{X}_n - z_{\alpha/2}\frac{\sqrt{\theta(1-\theta)}}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2}\frac{\sqrt{\theta(1-\theta)}}{\sqrt{n}}\right]\right) \approx 1-\alpha,$$

but the above interval is not good for estimation of θ since the interval itself depends on the unknown parameter θ.

Now suppose we can find a transformation $g(\overline{X}_n)$ such that g is an invertible and at least twice continuously differentiable function, satisfying

$$\sqrt{n}\left(g(\overline{X}_n) - g(\theta)\right) \xrightarrow{d} N(0, c^2), \tag{1}$$

where c is some known constant.
Then g is called a "variance-stabilizing" transformation and the $(1-\alpha)$ level CI for $g(\theta)$ is given by:

$$\left[g(\overline{X}_n) - z_{\alpha/2}\frac{c}{\sqrt{n}},\;\; g(\overline{X}_n) + z_{\alpha/2}\frac{c}{\sqrt{n}}\right].$$

And, therefore, whenever g is a monotone nondecreasing function, the $(1-\alpha)$ level CI for θ is given by:

$$\left[\; g^{-1}\left(g(\overline{X}_n) - z_{\alpha/2}\frac{c}{\sqrt{n}}\right),\;\; g^{-1}\left(g(\overline{X}_n) + z_{\alpha/2}\frac{c}{\sqrt{n}}\right)\;\right],$$

and whenever g is a monotone nonincreasing function, the $(1-\alpha)$ level CI for θ is given by:

$$\left[\; g^{-1}\left(g(\overline{X}_n) + z_{\alpha/2}\frac{c}{\sqrt{n}}\right),\;\; g^{-1}\left(g(\overline{X}_n) - z_{\alpha/2}\frac{c}{\sqrt{n}}\right)\;\right].$$

Our next step is to identify the variance-stabilizing transformation g.
In general, note that

$$\sqrt{n}\left(g(\widehat{\theta}_n) - g(\theta)\right) = \sqrt{n}\left((\widehat{\theta}_n - \theta)g'(\theta) + \frac{1}{2}g''(\theta^*)(\widehat{\theta}_n - \theta)^2\right),$$

where θ^* is a certain point between $\widehat{\theta}_n$ and θ.
Then

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \xrightarrow{d} N(0, \sigma^2(\theta)) \tag{2}$$

implies that for $g'(\theta) > 0$,

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) g'(\theta) \xrightarrow{d} N(0, \sigma^2(\theta)\left[g'(\theta)\right]^2), \;\text{ as }\; n \to \infty.$$

Also if

$$\widehat{\theta}_n \xrightarrow{P} \theta, \tag{3}$$

then

$$\frac{1}{2}g''(\theta^*)\left[\sqrt{n}(\widehat{\theta}_n - \theta)\right]\left(\widehat{\theta}_n - \theta\right) \xrightarrow{d} 0.$$

Thus, under the conditions (2), (3),

$$\sqrt{n}\left(g(\widehat{\theta}_n) - g(\theta)\right) \xrightarrow{d} N(0, \sigma^2(\theta)\left[g'(\theta)\right]^2),$$

and g is a variance-stabilizing transformation (and (1) is satisfied) if and only if

$$\sigma^2(\theta)\left[g'(\theta)\right]^2 = c^2.$$

Without loss of generality, we can let $c = 1$, i.e.

$$g'(\theta) = \frac{1}{\sigma(\theta)},$$

or,

$$g(\theta) = \int_{t \in \Theta:\, t \le \theta} \frac{1}{\sigma(t)} dt \,. \tag{4}$$

In our <u>Bernoulli example</u> (with $\widehat{\theta}_n = \overline{X}_n$), by LLN and CLT, conditions (3) and (2) are satisfied. Now let g be a twice continuously differentiable invertible function satisfying condition (4), namely, (for $\theta \in (0,1)$

$$g(\theta) = \int_0^{\theta} \frac{1}{\sqrt{t(1-t)}} dt = \int_0^{\sin^{-1}(\sqrt{\theta})} \frac{2\sin(x)\cos(x)}{\sqrt{\sin^2(x)\cos^2(x)}} dx = 2\sin^{-1}(\sqrt{\theta}),$$

where we put

$$t = \sin^2(x), \quad \text{i.e. } x = \arcsin(\sqrt{t}) \equiv \sin^{-1}(\sqrt{t}).$$

Then $g(t) = 2\sin^{-1}(\sqrt{t})$ is a variance-stabilizing transformation and condition (1) holds with $c = 1$, namely,

$$\sqrt{n}\left(2\sin^{-1}\left(\sqrt{\overline{X}_n}\right) - 2\sin^{-1}\left(\sqrt{\theta}\right)\right) \xrightarrow{d} N(0,1) \,.$$

Thus $(1-\alpha)$ level CI for $g(\theta) = 2\sin^{-1}(\sqrt{\theta})$ is given by:

$$\left[2\sin^{-1}(\sqrt{\overline{X}_n}) - z_{\alpha/2}\frac{1}{\sqrt{n}},\ \ 2\sin^{-1}(\sqrt{\overline{X}_n}) + z_{\alpha/2}\frac{1}{\sqrt{n}}\right] \,.$$

Note that $g^{-1}(x) = \sin^2\left(\frac{x}{2}\right)$, thus a level $(1-\alpha)$ CI for θ is given by:

$$\left[\ \sin^2\left(\sin^{-1}(\sqrt{\overline{X}_n}) - z_{\alpha/2}\frac{1}{2\sqrt{n}}\right),\ \ \sin^2\left(\sin^{-1}(\sqrt{\overline{X}_n}) + z_{\alpha/2}\frac{1}{2\sqrt{n}}\right)\right] \,.$$

<u>Another example</u>: $X_1, \ldots, X_n$ iid $Poisson(\theta)$, where parameter $\theta > 0$ is unknown. Then the population mean equals to θ and the population variance $\sigma^2 = \theta$. By (4), let

$$g(\theta) = \int_0^{\theta} \frac{1}{\sqrt{t}} dt = 2\sqrt{\theta},$$

then g is a variance stabilizing transformation with $c = 1$ and by (1),

$$\sqrt{n}\left(2\sqrt{\overline{X}_n} - 2\sqrt{\theta}\right) \xrightarrow{d} N(0,1)$$

Thus, $(1-\alpha)$ level CI for $2\sqrt{\theta}$ is:

$$\left[2\sqrt{\overline{X}_n} - z_{\alpha/2}\frac{1}{\sqrt{n}},\ 2\sqrt{\overline{X}_n} + z_{\alpha/2}\frac{1}{\sqrt{n}}\right] \,.$$

And the $(1-\alpha)$ level CI for θ is given by:

$$\left[\left(\sqrt{\overline{X}_n} - z_{\alpha/2}\frac{1}{2\sqrt{n}}\right)^2,\ \left(\sqrt{\overline{X}_n} + z_{\alpha/2}\frac{1}{2\sqrt{n}}\right)^2\right] \,.$$