Annexe technique : Démonstration de la variance de l'estimateur de Kaplan–Meier

Cette annexe présente la démonstration classique de la variance de l'estimateur de Kaplan–Meier, connue sous le nom de *formule de Greenwood*. L'objectif est d'obtenir une approximation de

$$\mathrm{Var}(\hat{S}(t)) \quad \text{où} \quad \hat{S}(t) = \prod_{t_{(j)} \leq t} \left(1 - \frac{d_j}{n_j}\right).$$

1. Préliminaires

On note :

$$\hat{p}_j = \frac{d_j}{n_j}, \qquad X_j = 1 - \hat{p}_j,$$

où

— $t_{(j)}$ est le j-ème instant où survient au moins une défaillance,
— d_j est le nombre de défaillances à $t_{(j)}$,
— n_j est le nombre d'individus à risque juste avant $t_{(j)}$.

L'estimateur Kaplan–Meier s'écrit :

$$\hat{S}(t) = \prod_{t_{(j)} \leq t} X_j.$$

Sous la modélisation standard,

$$d_j \sim \mathrm{Binom}(n_j, p_j), \qquad \hat{p}_j = d_j/n_j.$$

2. Passage au logarithme

On préfère étudier la variance de $\log \hat{S}(t)$ car :

$$\log \hat{S}(t) = \sum_{t_{(j)} \leq t} \log X_j.$$

On utilise l'approximation de Taylor :

$$\log(1 - x) \approx -x, \qquad \text{pour } x \text{ petit.}$$

Ainsi :

$$\log X_j = \log(1 - \hat{p}_j) \approx -\hat{p}_j,$$

et donc :

$$\log \hat{S}(t) \approx -\sum_{t_{(j)} \leq t} \hat{p}_j.$$

3. Variance de la somme

Comme

$$\hat{p}_j = \frac{d_j}{n_j} \quad \text{avec} \quad d_j \sim \text{Binom}(n_j, p_j),$$

on a :

$$\text{Var}(\hat{p}_j) = \frac{1}{n_j^2}\,\text{Var}(d_j) = \frac{p_j(1-p_j)}{n_j}.$$

Sous l'approximation usuelle $p_j \approx \hat{p}_j = d_j/n_j$, on obtient :

$$\text{Var}(\hat{p}_j) \approx \frac{d_j(n_j - d_j)}{n_j^3}.$$

L'approximation de la variance du log-KM est donc :

$$\text{Var}\big(\log \hat{S}(t)\big) \approx \sum_{t_{(j)} \le t} \frac{d_j}{n_j(n_j - d_j)}.$$

4. Retour à l'échelle de $\hat{S}(t)$ (Delta method)

On utilise que :

$$\log \hat{S}(t) \approx g(\hat{S}(t)), \qquad g(s) = \log s, \qquad g'(s) = \frac{1}{s}.$$

Par la méthode Delta :

$$\text{Var}(\hat{S}(t)) \approx \Big(g'(\hat{S}(t))^{-2}\Big)\,\text{Var}\big(\log \hat{S}(t)\big) = \hat{S}^2(t)\,\text{Var}\big(\log \hat{S}(t)\big).$$

En substituant l'expression précédente :

$$\text{Var}(\hat{S}(t)) \approx \hat{S}^2(t) \sum_{t_{(j)} \le t} \frac{d_j}{n_j(n_j - d_j)}.$$

5. Formule finale (Greenwood)

On obtient ainsi la formule de Greenwood :

$$\boxed{\widehat{\text{Var}}(\hat{S}(t)) = \hat{S}^2(t) \sum_{t_{(j)} \le t} \frac{d_j}{n_j(n_j - d_j)}}$$

Cette expression est utilisée pour :
— construire des intervalles de confiance autour de $\hat{S}(t)$;
— tracer des bandes de confiance sur les courbes de Kaplan–Meier ;
— diagnostiquer les zones de forte incertitude (fin de suivi).

Résumé. La démonstration repose sur trois idées :

1. Passer au log pour transformer un produit en somme.
2. Approximations locales : $\log(1-x) \approx -x$ et variance binomiale.
3. Retour à l'échelle de $S(t)$ via le Delta method.

Annexe technique : Estimateur de Breslow et sa variance

Cette annexe présente la construction et la variance de l'estimateur de Breslow pour la fonction de risque cumulée :

$$\Lambda(t) = \int_0^t \lambda(s)\, ds.$$

Contrairement à Kaplan–Meier, Breslow repose sur une **estimation additive** : il accumule les contributions de risque à chaque instant de défaillance. Cette structure rend l'analyse de sa variance beaucoup plus simple.

1. Définition : hazard cumulé discretisé

Aux temps de défaillance observés $t_{(1)}, t_{(2)}, \ldots$, on note :

$$n_j = \text{nb. à risque juste avant } t_{(j)}, \qquad d_j = \text{nb. de défaillances à } t_{(j)}.$$

L'estimateur de Breslow s'écrit :

$$\widehat{\Lambda}(t) = \sum_{t_{(j)} \le t} \frac{d_j}{n_j}.$$

Interprétation. Chaque fraction

$$\frac{d_j}{n_j}$$

est une estimation empirique du *hazard instantané* à $t_{(j)}$, si l'on approxime :

$$\lambda(t) \approx \Pr(T = t \mid T \ge t) = \frac{d_j}{n_j}.$$

2. Structure additive : un point crucial

L'estimateur de Breslow est une **somme** :

$$\widehat{\Lambda}(t) = \sum_j X_j, \qquad X_j = \frac{d_j}{n_j}.$$

Ce point simplifie fortement la variance :

$$\operatorname{Var}\big(\widehat{\Lambda}(t)\big) = \sum_{t_{(j)} \le t} \operatorname{Var}(X_j) \quad \text{(indépendance conditionnelle entre instants)}.$$

3. Variance de chaque contribution

Pour un temps $t_{(j)}$, on suppose classiquement :

$$d_j \sim \text{Binom}(n_j,\, p_j), \qquad p_j = \Pr(T = t_{(j)} \mid T \geq t_{(j)}).$$

Alors :

$$\widehat{p}_j = \frac{d_j}{n_j} = X_j, \qquad \text{Var}(X_j) = \frac{1}{n_j^2}\text{Var}(d_j) = \frac{p_j(1-p_j)}{n_j}.$$

En remplaçant p_j par $\widehat{p}_j = d_j/n_j$:

$$\text{Var}(X_j) \approx \frac{d_j(n_j - d_j)}{n_j^3}.$$

4. Formule simplifiée de la variance

On utilise l'approximation standard :

$$n_j - d_j \approx n_j \qquad \text{(valable lorsque } d_j \text{ est petit, cas courant).}$$

Ainsi :

$$\text{Var}(X_j) \approx \frac{d_j}{n_j^2}.$$

Finalement :

$$\boxed{\widehat{\text{Var}}\Big(\widehat{\Lambda}(t)\Big) = \sum_{t_{(j)} \leq t} \frac{d_j}{n_j^2}}$$

Ce résultat repose uniquement sur la structure additive de Breslow et la variance binomiale de d_j.

5. Lien avec Kaplan–Meier

Kaplan–Meier s'écrit :

$$\hat{S}(t) = \prod_j (1 - d_j/n_j),$$

ce qui implique une variance plus complexe (formule de Greenwood).

En revanche :

$$\widehat{\Lambda}(t) = \sum_j \frac{d_j}{n_j} \qquad \Longrightarrow \qquad \text{variance = somme des variances.}$$

Approximation du lien.

$$\hat{S}(t) = \exp\big(-\widehat{\Lambda}(t) + o(1)\big) \quad \text{car} \quad \log(1-x) \approx -x.$$

6. Résumé des différences clés

— Kaplan–Meier : estimateur de survie, structure en **produit**, variance compliquée (Greenwood).
— Breslow : estimateur du hazard cumulé, structure en **somme**, variance simple.
— Les deux sont cohérents et asymptotiquement équivalents.

À retenir : L'estimateur de Breslow est mathématiquement plus simple que Kaplan–Meier, et c'est pour cela qu'il est utilisé dans le modèle de Cox comme estimateur de $\Lambda_0(t)$.