Introduction à l'estimation statistique des paramètres

Sebastien Benzekry

March 4, 2018

Contents

1 Cadre et problématique

But : Ajuster un modèle dépendant de paramètres à des données expérimentales.

Les données du problème sont :

- Les **observations** : n couples de points (t_i, y_i) avec $y_i \in \mathbb{R}$ (ou $\mathbb{R}^m$). On notera parfois $y = (y_1, ..., y_n) \in \mathbb{R}^n$ et de même $t = (t_1, ..., t_n)$. Exemple : y = nombre de métastases, concentration d'un médicament dans l'organisme...
- Le **modèle structurel** : une fonction

$$f : \begin{array}{ccc} \mathbb{R} \times \mathbb{R}^p & \to & \mathbb{R} \\ (t, \theta) & \mapsto & f(t, \theta) \end{array}$$

- Les **paramètres** $\theta \in \mathbb{R}^p$.

Le problème est que l'on ne peut ni considérer le modèle comme une description exacte de la réalité (c'est d'ailleurs en substance le sens du mot modèle), ni les observations (erreurs de mesure). On se retrouve dans la situation suivante :

$$\boxed{y_i = f(t_i, \theta^*) + \varepsilon_i, \quad i = 1, ..., n} \tag{1}$$

avec θ^* la valeur du paramètre recherchée et ε_i un terme d'erreur, généralement constitué de deux termes, l'erreur de mesure et l'erreur structurelle (venant de l'imprécision intrinsèque du modèle), mais nous ne feront pas la distinction ici.

Se doter d'un **modèle statistique**, c'est donner une description des erreurs ε_i. Elles sont représentées par des variables aléatoires (v.a. en abrégé) et sont souvent supposées indépendantes (l'erreur de mesure que l'on commet à t_i n'influe pas sur l'erreur de mesure que l'on commet à t_j). Par exemple, on a souvent $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Les observations sont donc elles aussi des variables aléatoires, que l'on note Y_i. De manière générale, on essaiera de toujours noter en lettres majuscules les variables aléatoires. Il serait donc plus correct d'écrire, plutôt que (1)

$$Y_i = f(t_i, \theta^*) + \varepsilon_i,$$

les y_i étant alors des **réalisations** des v.a. Y_i. Le n-uplet $(y_1, ..., y_n)$ est appelé un **échantillon** et sa densité de probabilité, notée $p(y|\theta^*)$ car elle dépend de θ^*, est appelée la **distribution d'échantillonnage**. Si θ^* n'est plus vu comme la valeur exacte du paramètre recherché mais comme une variable notée θ, alors $p(y|\theta)$ en tant que fonction de θ et à y fixé s'appelle la **vraisemblance** et se note $L(\theta)$ (L comme *likelihood*, vraisemblance en anglais).

Le problème est de trouver une bonne valeur approchée de θ^*, appelée un **estimateur** de θ^*.

2 Méthodes d'estimation ponctuelle

On donne tout d'abord quelques définitions de statistiques.

2.1 Quelques notions de base de statistiques

Definition 1 (Estimateur)**.** *Soient $Y_1, ..., Y_n$ n variables aléatoires. On appelle estimateur toute variable aléatoire fonction de $Y_1, ..., Y_n$, c'est à dire $\hat{\theta}$ est un estimateur si il existe une fonction h telle que*

$$\hat{\theta} = h(Y_1, ..., Y_n)$$

Remark 1. *Un estimateur est donc une **variable aléatoire**, dont la loi dépend de la distribution d'échantillonnage.*

Example 1 (Sondage)**.** *Mettons qu'on veuille estimer le pourcentage de la population qui pense telle ou telle chose. On modélise la situation en disant que le résultat d'une hypothétique votation générale sur l'ensemble de la population est une variable aléatoire de Bernoulli, de paramètre θ^*, ce θ^* étant la proportion que l'on recherche. Cependant, on en va pas sonder toute la population et on dispose seulement d'un échantillon $(y_1, ..., y_n)$ de la population de petite taille. On peut définir comme estimateur de θ^* la **moyenne empirique** $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.*

D'après cette définition, $\hat{\theta}$ n'a a priori rien à voir avec le paramètre θ^* que l'on cherche à estimer. Les propriétés suivantes traduisent la manière dont l'estimateur approche le paramètre recherché.

Definition 2 (Propriétés d'un estimateur)**.** *Soit $\hat{\theta}$ un estimateur de θ^*. On dit que*

- *$\hat{\theta}$ est **sans biais** si $E[\hat{\theta}] = \theta^*$.*
- *$\hat{\theta}$ est **consistant** si $\hat{\theta} \xrightarrow[n\to\infty]{} \theta^*$, en probabilité, cad*

$$\forall \varepsilon > 0, \mathbb{P}(|\hat{\theta} - \theta^*| \geq \varepsilon) \xrightarrow[n\to\infty]{} 0$$

- $\hat{\theta}$ *est* ***efficace*** *si il est de risque quadratique minimal, cad*

$$R(\theta^*, \hat{\theta}) = min\{R(\theta^*, \hat{\theta}')\};$$

avec $R(\theta^*, \hat{\theta}) = E[(\hat{\theta} - \theta^*)^2]$

On va maintenant donner quelques exemples d'estimateurs très utilisés pour l'estimation des paramètres. Le premier, qui est à la fois le plus simple, le plus naturel et le plus utilisé est l'estimateur des moindres carrés.

2.2 Moindres carrés

Le principe des moindres carrés est de choisir le paramètre θ qui minimise la somme des carrés des **résidus** $r_i = y_i - f(t_i, \theta)$, c'est à dire :

$$\boxed{\hat{\theta}_{MC} = \underset{\theta}{\operatorname{argmin}} ||Y - f(t, \theta)||_2^2 = \underset{\theta}{\operatorname{argmin}} \left(\sum_{i=1}^{n} |Y_i - f(t_i, \theta)|^2 \right)}$$

Exemple : Modèle linéaire (en les paramètres)
On a $Y_i = a(t_i)^t \theta^* + \varepsilon_i$, avec $a(t_i) \in \mathbb{R}^p$ (par exemple $a(t_i) = (1, t_i, t_i^2)$ et $f(t_i, \theta) = \theta_1 + \theta_2 t + \theta_3 t^2$), ou plus simplement en notant $A = (a(t_i))_{1 \leq i \leq n} \in \mathbb{R}^{n \times p}$

$$Y = A\theta^* + \varepsilon.$$

On minimise la fonctionnelle $S(\theta) = ||Y - A\theta||_2^2 =< Y - A\theta, Y - A\theta >$. En remarquant que, si la matrice $A^t A$ est inversible

$$d_\theta S(h) = -2 < Ah, Y - A\theta >= -2 < h, A^t(Y - A\theta) >= 0 \Leftrightarrow \theta = (A^t A)^{-1}(A^t Y)$$

on a

$$\hat{\theta}_{MC} = (A^t A)^{-1}(A^t Y)$$

d'où on tire, si $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ que $\hat{\theta}_{MC} \sim \mathcal{N}(\theta^*, \sigma^2 (A^t A)^{-1})$.

Dans le cas général on a la proposition suivante sur les propriétés de l'estimateur des moindres carrés.

Proposition 1 (Propriétés de l'estimateur des moindres carrés)**.** *Supposons que les* ε_i *soient indépendantes et identiquement distribuées (iid en abrégé) avec variance* σ^2 *et qu'il existe un unique minimiseur de* $S(\theta)$*. Alors* $\hat{\theta}_{MC}$ *est un estimateur consistant de* θ^**. De plus,* $\hat{\sigma}^2_{MC} := \frac{S(\hat{\theta})}{n-p}$ *est un estimateur consistant de* σ^2*.*

Remark 2. *A mesure que le nombre de paramètres p augmente, l'estimation se fait moins précise.*

Si l'on veut pondérer les résidus pour leur donner plus ou moins d'importance, on se définit une matrice V et on minimise la fonctionnelle $(Y - f(t, \theta))^t V (Y - f(t, \theta))$.
L'estimateur des moindres carrés trouve ses limites dans le fait qu'il ne prend pas en compte la structure aléatoire de l'erreur.

2.3 Maximum de vraisemblance

Rappelons que les observations y_i sont des réalisations de v.a. dont les densités dépendent de θ^*. Ces densités sont connues une fois que l'on a défini le modèle structurel f et le modèle statistique (la loi des ε_i), et sont alors des fonctions du paramètre θ, dénotées par $p(y_i|\theta)$. On définit alors la **vraisemblance** $L(\theta)$ comme cette densité appliquée aux valeurs observées $y_1, ..., y_n$, en tant que fonction de θ et l'indépendance des Y_i donne que la loi jointe du n-uplet $(Y_1, ..., Y_n)$ est donnée par le produit des lois :

$$L(\theta) = p(y_1, ..., y_n|\theta) = \prod_{i=1}^{n} p(y_i|\theta) \tag{2}$$

C'est la probabilité d'obtenir l'observation $y_1, ..., y_n$ si le paramètre vaut θ. L'estimateur du maximum de vraisemblance consiste à choisir la valeur de θ qui maximise la probabilité d'apparition de cette *observation*.

$$\boxed{\hat{\theta}_{MV} = \underset{\theta}{\operatorname{argmax}}\, L(\theta)}.$$

On a la proposition suivante sur les propriétés de l'estimateur du maximum de vraisemblance, qui stipule notamment qu'asymptotiquement on ne peut mieux faire.

Proposition 2 (Propriétés de l'estimateur du max de vraisemblance)**.** *Sous des conditions de régularité sur $L(\theta)$, $\hat{\theta}_{MV}$ est consistant et asymptotiquement efficace.*

Remark 3. *Plus précisément, on a*

$$\sqrt{n}(\hat{\theta}_{MV} - \theta^*) \rightharpoonup \mathcal{N}(0, I_{\theta^*}^{-1})$$

où I_θ est la matrice d'information de Fisher définie par:

$$(I_\theta)_{j,k} = E\left[\left\{\frac{\partial \log(p(y_i|\theta))}{\partial \theta_j}\right\}^t_{i=1,...,n} \left\{\frac{\partial \log(p(y_i|\theta))}{\partial \theta_k}\right\}_{i=1,...,n}\right]$$

En pratique, l'estimateur du max de vraisemblance se révèle utile pour les échantillons larges. Pour de petits échantillons, il est en général biaisé et non efficace.

2.3.1 Erreur iid, de loi normale, variance constante

Si $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ on a $Y_i = f(t_i, \theta) + \varepsilon_i \sim \mathcal{N}(f(t_i, \theta), \sigma^2)$ et alors

$$p(y_i|\theta, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - f(t_i,\theta))^2}{2\sigma^2}}.$$

ce qui donne

$$L(\theta, \sigma) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}||y - f(t,\theta)||_2^2}$$

On utilise alors la log-vraisemblance (= log de la vraisemblance) car maximiser la vraisemblance est équivalent à maximiser cette dernière. On obtient

$$\left(\hat{\theta}_{MV}, \hat{\sigma}_{MV}\right) = \underset{\theta,\sigma}{\operatorname{argmin}} \left\{ n \log(\sigma\sqrt{2\pi}) + \frac{||y - f(t,\theta)||_2^2}{2\sigma^2} \right\} \tag{3}$$

Si σ ne fait pas partie des paramètres à estimer, alors on voit que l'on retombe sur l'estimateur des moindres carrés. En fait, même si on cherche à estimer la valeur de σ, il suffit de voir L comme une fonction dépendant aussi de σ, $L(\theta, \sigma)$. En annulant la dérivée dans la direction σ, on obtient [1, p.32]

$$\hat{\sigma}^2_{MV} = \frac{1}{n}\sum_{i=1}^{n} |y_i - f(t_i, \hat{\theta}_{MV})|^2 = \frac{S(\hat{\theta}_{MV})}{n}.$$

C'est là une différence avec l'estimateur des moindres carrés, qui lui utilise $\hat{\sigma}^2_{MC} := \frac{S(\hat{\theta}_{MC})}{n-p}$. Néanmoins, en remplaçant ensuite dans (3), cela suggère que $\hat{\theta}_{MV} = \hat{\theta}_{MC}$. On montre cette égalité par des calculs directs. Il peut être utile d'observer que la log-vraisemblance s'écrit alors [1, p.32]:

$$l(\theta) = \log(L(\theta)) = -\frac{n}{2}\log\left(\hat{\sigma}^2(\theta)\right) - \frac{n}{2}\log(2\pi) - \frac{n}{2} \tag{4}$$

2.3.2 Erreur iid, de loi normale, variance non-constante

Dans le cas où la variance des ε_i dépendrait de la quantité modélisée, *via* un modèle d'erreur, par exemple considéré proportionnel, i.e.

$$Y_i = f(t_i, \theta) + \sigma f(t_i, \theta)\varepsilon_i \sim \mathcal{N}(f(t_i, \theta), \sigma^2 f(t_i, \theta)^2),$$

il faut utiliser $f(t_i, \theta)$ plutôt que y_i dans l'expression de l'erreur, cf [RIG89, p.593] et [Cou, chap.2, slide 40]), on obtiendrait

$$l(\theta) = \log(L(\theta)) = -\frac{n}{2}\log\left(\hat{\sigma}^2(\theta)\right) - \sum_{i=1}^{n}\log\left(f(t_i, \theta)\right) - \frac{n}{2}\log(2\pi) - \frac{n}{2}$$

avec cette fois

$$\hat{\sigma}^2(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - f(t_i, \theta)}{f(t_i, \theta)}\right|^2$$

Dans ce cas, l'estimation diffère de celle des moindres carrés car le second terme dépend aussi de θ.

Remark 4. *Dans le cas d'une comparaison de modèles, pour éviter d'avoir un terme d'erreur dépendant du modèle structurel, il pourrait sembler pertinent, bien que nous n'ayons aucune référence à citer pour ceci, de considérer la variance proportionnelle à la donnée, c'est à dire:*

$$Y_i = f(t_i, \theta^*) + \sigma Y_i \varepsilon_i \sim \mathcal{N}(f(t_i, \theta^*), \sigma^2 Y_i^2).$$

Cela conduit à

$$l(\theta) = -\frac{n}{2}\log\left(\hat{\sigma}^2(\theta)\right) - \sum_{i=1}^{n}\log\left(y_i\right) - \frac{n}{2}\log(2\pi) - \frac{n}{2}$$

avec

$$\hat{\sigma}^2(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - f(t_i, \theta)}{y_i}\right|^2$$

et la minimization ne porte plus que sur le premier terme, ce qui rend l'estimateur du maximum de vraisemblance équivalent à celui des moindres carrés.

2.4 Estimation bayésienne. Maximum a posteriori

On est dans la situation où on a une **information a priori** sur θ. Celle-ci se traduit par une densité de probabilité sur la répartition de θ dans la population, donnée par une étude statistique préalable. Attention, ce n'est pas le même aléa pour θ et pour y.

- y : erreur de mesure $\leftrightarrow$ Précision
- θ : aléa dans la population $\leftrightarrow$ Dispersion

L'estimation bayésienne consiste à utiliser la **distribution a posteriori** $p(\theta|y)$.

$$\boxed{\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\, p(\theta|y)}.$$

Le problème est qu'en général, on ne la connaît pas explicitement. D'où le nom d'estimation bayésienne qui vient du théorème de Bayes qui dit :

$$p(\theta|y) = \frac{p(y,\theta)}{p(y)} = \frac{p(y|\theta) \times p(\theta)}{p(y)}$$

Donc, une fois que l'on a fixé y (donné par les observations), $p(y)$ est une constante dans la maximisation en θ et on a:

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \left\{\log(p(y|\theta) + \log(p(\theta))\right\}.$$

Remark 5.
- *Si $p(\theta) = cte$ (ie θ uniformément distribuée), alors $\hat{\theta}_{MAP} = \hat{\theta}_{MV}$.*
- *On pénalise la vraisemblance par la distribution a priori : si le max de vraisemblance donne une valeur mais que celle-ci a une petite probabilité d'apparition dans la population, alors elle est pénalisée.*

Proposition 3. *Si $p(\theta)$ est continue et $p(\hat{\theta}_{MV})\mathcal{N}eq0$,*

$$\hat{\theta}_{MAP} \xrightarrow[n\to\infty]{} \hat{\theta}_{MV}$$

3 Intervalles de confiance. Inférence statistique

En pratique, connaître les valeurs des paramètres qui ajustent au mieux le modèle, c'est bien mais il est encore plus intéressant d'avoir un intervalle autour de la valeur estimée dans lequel on sait que le vrai paramètre θ^* se trouve avec forte probabilité (par exemple 95%).

Definition 3 (Intervalle de confiance)**.** *Un intervalle de confiance au seuil α pour $\theta^* \in \mathbb{R}$ est un intervalle I tel que :*

$$\mathbb{P}(\theta^* \in I) = 1 - \alpha$$

3.1 La base : le théorème central limite (TCL)

Soient $Y_1, ..., Y_n$ n variables aléatoires iid (indépendantes, identiquement distribuées), d'espérance μ. On pose

$$\overline{Y_n} := \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Theorem 1 (Loi forte des grands nombres)**.** *On suppose $\mu < \infty$. Alors*

$$\overline{Y_n} \xrightarrow[n\to\infty]{} \mu, \quad ps$$

Mais on sait mieux, on sait comment est répartie la moyenne empirique $\overline{Y_n}$ autour de l'espérance μ. C'est l'objet du théorème suivant.

Theorem 2 (Théorème Central Limite)**.** *Supposons que $\sigma^2 = E[(Y_1 - E[Y_1])^2] < \infty$. Alors*

(i) $E[\overline{Y_n}] = \mu$, $V(\overline{Y_n}) = \frac{\sigma^2}{n}$.

(ii) $\sqrt{n}\overline{Y_n} \rightharpoonup \mathcal{N}(\mu, \sigma^2)$

Remark 6 (Vitesse de convergence)**.** *En pratique le théorème central limite est presque toujours employé comme si la convergence était une égalité. Cela se base sur le théorème de Berry-Essen qui établit la vitesse de convergence. Si $\mathbb{R}ho := E[Y_1^3] < \infty$, alors :*

$$||F_n - \phi||_\infty \leq \frac{C\mathbb{R}ho}{\sigma^3\sqrt{n}},$$

où F_n est la fonction de répartition de $\frac{\sqrt{n}\overline{Y_n}}{\sigma}$ et ϕ celle d'une loi normale centrée réduite. La constante C est inférieure à 0.7056 (Shevtsova 2007). Par exemple, pour Y_1 de loi uniforme sur $[-1, 1]$, on obtient

$$||F_n - \phi||_\infty \leq \frac{1}{\sqrt{n}}$$

On représente sur la figure 1 les valeurs de $\frac{1}{\sqrt{n}}$ pour n de 1 à 30. Pour $n \geq 20$, on a $||F_n - \phi||_\infty \leq 0.22$.

Application aux statistiques.
Soit Y une va dont on cherche à estimer l'espérance μ sachant qu'on dispose d'un échantillon $(y_1, ..., y_n)$ de n réalisations indépendantes de Y, c'est à dire de n va $Y_1, ..., Y_n$ iid de même loi que Y. La moyenne empirique $\overline{Y_n}$ est un estimateur de μ, consistant d'après la loi forte des grands nombres. De plus le TCL nous dit que, pour n assez grand

$$\sqrt{n}\left(\frac{\overline{Y_n} - \mu}{\sigma}\right) \sim \mathcal{N}(0, 1),$$

d'où on tire, avec ε_α tel que $\mathbb{P}(|N| \leq \varepsilon_\alpha) = 1 - \alpha$ (valeur donnée par des tables), voir la figure 2.

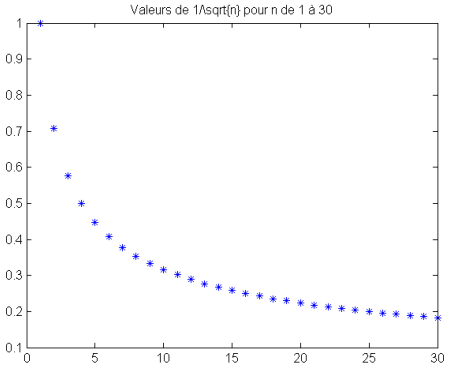

Figure 1 – Valeurs de $\frac{1}{\sqrt{n}}$ pour n de 1 à 30

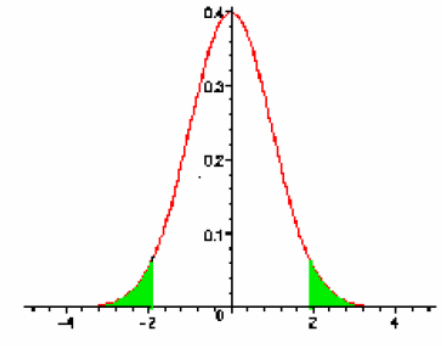

Figure 2 – Illustration de la valeur ε_α

$$\mathbb{P}\left(\left|\sqrt{n}\frac{\overline{Y_n}-\mu}{\sigma}\right| \leq \varepsilon_\alpha\right) = 1-\alpha$$

L'intervalle au seuil de confiance α de μ est donc

$$\left[\overline{Y_n} - \varepsilon_\alpha \frac{\sigma}{\sqrt{n}}, \overline{Y_n} + \varepsilon_\alpha \frac{\sigma}{\sqrt{n}}\right]$$

Quelques valeurs de ε_α pour différentes valeurs de α sont données dans le tableau suivant.

α	ε_α
0,01	2,576
0,05	1,96
0,1	1,645

Example 2. *Une balance mesure des masses avec une variance $\sigma^2 = 0,015$. On pèse trois fois la masse d'un même corps (modélisée par une variable aléatoire Y d'espérance la vraie masse du corps et de variance σ^2). On obtient $y_1 = 64,32g$, $y_2 = 64,27g$, $y_3 = 64,39g$. On calcule*

alors $\overline{Y_3} = 64,33g$ *et en prenant un seuil de confiance de* 99%*, on calcule* $\varepsilon_\alpha \frac{\sigma}{\sqrt{3}} = 0,058$*, d'où on tire :*

$$\mu = 64,33g \pm 0,058, \quad I_{0,01} = [64,27; 64,39]$$

3.2 Variance inconnue. Loi du chi-deux. Loi de Student.

Le problème avec cette méthode est qu'elle requiert la connaissance de la variance σ^2 des observations, ce qui est rarement le cas en pratique. Par contre, les données permettent le calcul de la variance empirique qui est un estimateur non-biaisé de la variance et est défini par :

$$\hat{\sigma}^2 := \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y_n})^2$$

On a alors besoin d'introduire une nouvelle loi pour décrire la loi de $\hat{\sigma}^2$.

Definition 4 (Loi du chi-deux)**.** *Soient* $\varepsilon_1, ..., \varepsilon_k$ k *va iid de loi normale centreé réduite* $\mathcal{N}(0,1)$*. Alors la loi de la va* $S = \sum_{i=1}^{k} \varepsilon_i^2$ *est appelée loi du chi-deux à* k *degrés de liberté et est notée* $\chi^2(k)$*.*

Proposition 4 (Loi de $\hat{\sigma}^2$)**.** *On a* $E[\hat{\sigma}^2] = \sigma^2$ *et la loi de* $\frac{\hat{\sigma}^2(n-1)}{\sigma^2}$ *est la loi du chi-deux à* $(n-1)$ *degrés de liberté :*

$$\frac{\hat{\sigma}^2(n-1)}{\sigma^2} \sim \chi^2(n-1)$$

On ne peut plus appliquer le théorème central limite mais les calculs suivants motivent l'introduction de la loi de Student :

$$\begin{aligned}\sqrt{n}\frac{\overline{Y_n} - \mu}{\hat{\sigma}} &= \underbrace{\sqrt{n}\frac{\overline{Y_n} - \mu}{\sigma}}_{:=N} \times \frac{\sigma}{\hat{\sigma}} \\ &= \frac{N}{\underbrace{\frac{\hat{\sigma}}{\sigma}\sqrt{n-1}}_{:=\sqrt{U}} \times \frac{1}{\sqrt{n-1}}} = \frac{N}{\sqrt{\frac{U}{n-1}}},\end{aligned}$$

avec $N \sim \mathcal{N}(0,1)$ et $U = \frac{\hat{\sigma}^2(n-1)}{\sigma^2} \sim \chi^2(n-1)$.

Definition 5 (Loi de Student)**.** *Soit* $N \sim \mathcal{N}(0,1)$ *et* $U \sim \chi^2(k)$*. Alors la loi de la va* $\frac{N}{\sqrt{\frac{U}{k}}}$ *est appelée loi de Student à* k *degrés de liberté et est notée* $t(k)$*.*

On a donc $\sqrt{n}\frac{\overline{Y_n}-\mu}{\hat{\sigma}} \sim t(n-1)$ et en introduisant t_{n-1}^{α} tel que pour $S \sim t(n-1)$, $\mathbb{P}(|S| \leq t_\alpha^{n-1}) = 1-\alpha$ (donné par les tables), on a $\mathbb{P}\left(\left|\sqrt{n}\frac{\overline{Y_n}-\mu}{\hat{\sigma}}\right| > t_\alpha^{n-1}\right) = \alpha$ et donc un intervalle de confiance au seuil α pour μ est donné par

$$\boxed{\left[\overline{Y_n} - t_\alpha^{n-1}\frac{\hat{\sigma}}{\sqrt{n}}; \overline{Y_n} + t_\alpha^{n-1}\frac{\hat{\sigma}}{\sqrt{n}}\right]}$$

3.3 Le cas des moindres carrés

Dans le cadre de l'estimation des paramètres (qui s'appelle aussi la **régression**), on peut aussi appliquer cette théorie bien qu'a priori on ne soit pas dans le cas de va iid (les lois des Y_i ne sont pas identiques). On le fait ici pour l'estimateur des moindres carrés.

3.3.1 Cas linéaire

On reprend le modèle linéaire de la section 2.2 :

$$Y = A\theta^* + \varepsilon$$

et l'estimateur des moindres carrés défini par

$$\hat{\theta}_{MC} = \underset{\theta}{\operatorname{argmin}} \left(||Y - A\theta||_2^2 \right) = \underset{\theta}{\operatorname{argmin}} \left(S(\theta) \right) = (A^t A)^{-1} A^t Y$$

On a besoin de préciser la proposition 1.

Proposition 5 (Propriétés de l'estimateur des moindres carrés dans le cas linéaire)**.** *On suppose le modèle structurel linéaire en les paramètres et que l'erreur est normale (ie que les ε_i sont iid de loi normale centrée $\mathcal{N}(0, \sigma^2)$. Alors*

(i) $\hat{\theta}_{MC} \sim \mathcal{N}_p(\theta^*, \sigma^2 (A^t A)^{-1})$

(ii) $\hat{\theta}_{MC}$ *est indépendant de* $\hat{\sigma}^2 = \frac{S(\hat{\theta})}{n-p}$

(iii) $(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p)$

Grâce à cette proposition, il est légal de faire comme précédemment pour obtenir un intervalle de confiance pour la r-ième composante de θ^*, donné par

$$\boxed{\left[\hat{\theta}_{MC} - t_\alpha^{n-1} \frac{\hat{\sigma}}{\sqrt{n}} c_{rr} ; \hat{\theta}_{MC} + t_\alpha^{n-1} \frac{\hat{\sigma}}{\sqrt{n}} c_{rr} \right]},$$

où on a noté c_{rs} les coefficients de la matrice $C = (A^t A)^{-1}$.

3.3.2 Cas non-linéaire

Dans le cas d'un modèle non-linéaire en les paramètres .....

$$Y = f(t, \theta^*) + \varepsilon$$

..... on linéarise! (autour de θ^*). Posons $F = D_\theta f(t, \theta^*)$ et écrivons $f(t, \theta) = f(t, \theta^*) + F \cdot (\underbrace{\theta - \theta^*}_{:=\beta})$

$$S(\theta) = ||Y - f(t, \theta)||^2 \simeq ||\underbrace{Y - f(t, \theta^*)}_{=\varepsilon} - F \cdot \beta||^2 = ||\varepsilon - F \cdot \beta||^2$$

On obtient alors un estimateur pour β donné par $\hat{\beta}_{MC} = (F^t F)^{-1} F^t \varepsilon$, d'où on tire

$$\hat{\theta}_{MC} = \theta^* + \hat{\beta}_{MC} = \theta^* + (F^t F)^{-1} F^t \varepsilon.$$

Proposition 6. *Supposons que* $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 Id)$. *Alors, approximativement*

(i) $\hat{\theta}_{MC} \sim \mathcal{N}_p(\theta^*, \sigma^2 (F^t F)^{-1})$. *On pose* $C = F^t F$

(ii) $\hat{\theta}_{MC}$ *est indépendant de* $\hat{\sigma}^2 = \frac{S(\hat{\theta})}{n-p}$

(iii) $(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p)$

On obtient alors un intervalle de confiance au seuil α pour la r-ième composante de θ^* en approximant $F = D_\theta f(t, \theta^*)$ par $\hat{F} = D_\theta f(t, \hat{\theta}_{MC})$:

$$\boxed{\left[\hat{\theta}_{MC} - t_\alpha^{n-1} \frac{\hat{\sigma}}{\sqrt{n}} \hat{c}_{rr}; \hat{\theta}_{MC} + t_\alpha^{n-1} \frac{\hat{\sigma}}{\sqrt{n}} \hat{c}_{rr}\right]},$$

4 Hypothesis testing, p-values, statistical significance

When confronted to analysis of data sets, one often faces the question of the *statistical significance* of the results, for example to test if the difference in the mean of two group samples is significant so that we can conclude that the groups are "truly" different. The scientific definition of "truly" is provided by the quantitative assessment of the probability that the observed differences would occur by chance only. More precisely, the concept of hypothesis testing relies on assuming a null hypothesis H_0 to hold and then compute the probability of obtaining the observed sample data under this hypothesis. This probability is the *p-value* associated to the statistical test. When this value is below some arbitrary significant level α (often taken to 0.05 or 0.01), one rejects the hypothesis H_0 because the probability that our data would have been generated under this hypothesis is too low.
example

4.1 Quantitative variables. Test differences for the mean or distribution

4.1.1 Z-test

Test difference between two samples when variance is known. Assumes normal distribution.

4.1.2 Student's t-test

Test difference between two samples when variance is unknown. Assumes normal distribution.

4.1.3 Nonparametric tests

Wilcoxon rank-sum test (also called Wilcoxon-Mann-Whitney or Mann-Whitney U test): Non-parametric test of the null hypothesis that two *indenpendent* samples are issued from the same distribution. Similar to t-test but does not require assumption of normally distributed samples (thus can be used for small sample size).

Wilcoxon signed-rank test: Nonparametric test of the null hypothesis that two *paired* samples are issued from the same distribution. Similar to paired t-test but does not require assumption of normally distributed samples (thus can be used for small sample size).

4.2 Normality tests χ^2

Kolmogorov-Smirnov (n¿50).

Shapiro-Wilk (n¡50).

4.3 Qualitative variables

4.3.1 Goodness-of-fit for frequencies. χ^2 test

The framework of this test is when one dispose of data on empirical frequencies of a phenomenon stratified in a number of categories. The classical example is the case of a die that can take values from one to six which would have been thrown a large number of times n and for which one would have reported the empirical frequencies of each of the six faces. The example that we have in mind is the proportion of metastatic relapses in breast cancer stratified by the primary tumor size at diagnosis (see Table 1).

Diameter of primary tumor at diagnosis (cm)	**Proportion of patients developing metastasis (%)**	**No. patients**
$1 \leq D \leq 2.5$	27.1	317
$2.5 < D \leq 3.5$	42.0	496
$3.5 < D \leq 4.5$	56.7	544
$4.5 < D \leq 5.5$	66.5	422
$5.5 < D \leq 6.5$	72.8	329
$6.5 < D \leq 7.5$	83.8	192
$7.5 < D \leq 8.5$	81.3	136

Table 1 – Probability of metastatic relapse as a function of primary tumor size at diagnosis in breast cancer, from [KTL+84]

Suppose that we have a model with output the probability of falling into each category. The question is then to determine, for a given set of parameter of the models, whether the model probabilities (the *expected* frequencies) could have generated the *observed* frequencies. The following proposition enlightens the usefulness of the χ^2 statistics to this regard.

Proposition 7 (Pearson). *Let $X_1, \cdots, X_N$ be N independent and identically distributed random variables with values in a set S. Let $B_1, \cdots, B_K$ be K subsets of S such that $\bigcup_{k=1}^{K} B_k = S$. For any $1 \leq k \leq K$, let p_k such that $\mathbb{P}(X_1 \in B_k) = p_k$, $E_k := Np_k$ and $O_k := \sum_{n=1}^{N} \mathbb{1}_{\{X_n \in B_k\}}$. Then*

$$Z^2 = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k} \underset{N \to +\infty}{\rightharpoonup} \chi^2(K-1)$$

Find a proper reference

A heuristic explanation of the reason why the number of degrees of freedom appearing in the χ^2 distribution is $K-1$ (and not K) is that the probabilities are linked by $\sum_{k=1}^{K} p_k = 1$ (from the assumption that all the possible values taken by the X_k fall within at least one B_l, i.e. $\bigcup_{k=1}^{K} B_k = S$), thus reducing the number of degrees of freedom of one.

In our example the X_n's would be the primary tumor sizes (with N individuals in total), the B_k's the ranges used for classifying these sizes, the O_k's the observed frequencies (of metastasis) and the E_k's the expected frequencies (given by the underlying probabilities p_k, $1 \leq k \leq K$).

This proposition allows us to detect whether the data could come from a theoretical distribution (defined by means of the p_k's). The associated p-value to the statistics Z^2 is $p = \mathbb{P}(U \geq Z^2)$ with $U \sim \chi^2(K-1)$. The theoretical distribution is rejected if p is lower than a given threshold α, usually taken to 0.05.

When the probabilities p_k come from a parametric model with P parameters, the proposition has to be modified and now, asymptotically

$$Z^2 \sim \chi^2(K-P-1)$$

4.3.2 Association. χ^2 test

Another use of the χ^2 statistic is in a test for association between categorical variables (such as gender and detection of brain metastasis, for instance). Data are often represented as in Table 2.

	Metastasis	No metastasis	Total
Male	55 (34.8)	103 (65.2)	158
Female	40 (37.7)	66 (62.3)	106
Total	95	169	

Table 2 – Proportion of males and females patients with non-small cell lung carcinoma developing cerebral metastasis [MAM+07]

The null hypothesis is that the event (the occurrence of brain metastasis in our example) is independent of the variable (gender in our case). Under this hypothesis, the expected frequencies for each of the four cells in the Table are computed in the following way (for instance for the cell "male with metastasis"). The proportion of men is $\frac{158}{158+106}$ and the (total) number of patients with metastasis is 95, thus:

$$E_{1,1} = \frac{95 \times 158}{158+106} = 56.9$$

The mathematical result, due to Pearson [Pea00], states that under the null hypothesis

$$\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}} \sim \chi^2((I-1)(J-1))$$

where I and J are respectively the number of rows and columns, and $O_{i,j}$ are the observed proportions.

4.4 F-test

The name F-test comes from the statistician R. Fisher (1890-1962).

4.4.1 Analysis Of VAriance (ANOVA)

4.4.2 Goodness-of-fit for regression

While usually a F test is used to compare two models, a global significance F test is performed to compare the given model to a model with only an intercept, i.e. a constant model of the average of the data. More precisely, the null hypothesis is that both fits are equal and the alternative hypothesis is that the fit of the constant model is significantly lower than your model. The F-statistic is the ratio of the two variances and is given by (assuming n observations and a model with P parameters)

$$F = \frac{\frac{SST-SSE}{P-1}}{\frac{SSE}{n-P}}$$

where SST, SSE and SSM respectively stand for the total sum of squares and error sum of squares and are given by

$$SST = \sum_i (y_i - \overline{y_n})^2, \quad SSE = \sum_i \left(y_i - M(t_i, \hat{\theta})\right)^2.$$

If we introduce the model sum of squares

$$SSM = \sum_i \left(M(t_i, \hat{\theta}) - \overline{y_n}\right)^2$$

the underlying reason for this expression of F is that in the linear case

$$SST = SSM + SSE$$

and thus

$$F = \frac{\frac{SSM}{P-1}}{\frac{SSE}{n-P}} = \frac{\text{explained variance}}{\text{unexplained variance}}.$$

Under the null hypothesis, F follows then a F-distribution which can be used to determine a p-value. We refer to [Ros15] for more details.

Remark 7. *In [SWK+12, BWS+14], the authors mention p-values for a goodness-of-fit test (not specified but coming out of the NLIN procedure from SAS). It seems that the test is actually a F test (see SAS user manual).*

4.5 Type I and type II errors. Sample size. Sensitivity/Specificity. ROC curve analysis

In terms of hypothesis testing, the result is considered as **positive** when H_0 is rejected and **negative** when H_0 is failed to be rejected. This makes sense when considering the example of

testing whether a patient is affected by the disease. In this case:

$$H_0 : \text{“the patient is healthy” versus } \overline{H_0} : \text{ “the patient is affected the disease”.}$$

An example from oncology could be:

$$H_0 : \text{“the patient has no tumor” versus } \overline{H_0} : \text{ “the patient has a tumor”.}$$

In these terms, when considering a statistical test of a hypothesis H_0, two types of errors can be made: the rejection of H_0 when it is in fact true (type I error, false positive) or the failure to reject H_0 when it is false in the reality (type II error, false negative). The probability of a type I error is denoted by α. It is the same as the significance level and $1-\alpha$ is called the **specificity** of the test. Value of α is often taken to 0.05. The probability of a type II error is denoted by β and the **power** (also called **sensitivity**) of a test is defined by $1-\beta$.

In mathematical terms, denoting P the event of a positive test result and N the event of a negative test result:

$$\alpha = \mathbb{P}(P|H_0), \quad \beta = \mathbb{P}(N|\overline{H_0}).$$

Stated differently, the power of a test is the probability to detect an effect when this effect is indeed true. It depends on the sensitivity level defined for the test and the sample size n.

In the example of a medical test determining if a patient is healthy or not, a type I error would then misjudge a healthy patient as sick (or detecting a malignant tumor when there is none) and a type II error would classify as healthy a sick patient (respectively, fail to detect the presence of a malignant tumor). See the Table 3 for a summary of these concepts.

Another common example is testing the effect of a drug when compared to a placebo in a clinical trial. Then H_0 is “the drug has no effect” and the alternative H_1 is “the drug is effective”. A type I error would be to conclude that the drug has an effect when in fact it does not (falsely reject H_0) whereas a type II error would be to fail to detect an effect of a drug when it fact it is truly active (falsely accept H_0). In other words, when we conclude that the drug has an effect, there are two possibilities for the truth. Either the drug had no effect and we observed an effect just by “chance”. This has probability α. Or the drug does have an effect and the probability that we are correct in our conclusion is $1-\beta$.

Several clinical and preclinical studies are specifically about calculating the power of the test (which depends on the size of the sample, i.e. the number of subjects) in the design of the experiment in order to have important chances to detect an effect.

4.5.1 Positive and negative value predictive

Knowing the specificity and sensitivity of a test is not sufficient to determine the real quantity of interest in the case of medical testing. Indeed, what we are interested is to determine whether a patient who has a positive result of a test indeed carries the disease or not. The previous notions dealt with the reverse conditional probability, i.e. what was known was the state of the disease and what was evaluated was the performance of the test with regards to that. Let us take the example of detection of a tumor (event T) for simplicity, where $H_0 = \overline{T}$ and $\overline{H_0} = T$. What we are interested in is $\mathbb{P}(T|P)$ (known as the positive value predictive, PV^+), which is

Test result \ Reality	H_0 $\overline{D}$ (no disease)	$\overline{H_0}$ D (disease)
Positive (reject H_0)	False Positive $P\|H_0$ $P\|\overline{D}$ α Type I	True Positive Sensitivity $P\|\overline{H_0}$ $P\|D$ $1-\beta$
Negative (don't reject H_0)	True Negative Specificity $N\|H_0$ $N\|\overline{D}$ $1-\alpha$	False negative $N\|\overline{H_0}$ $N\|D$ β Type II

Table 3 – Type I and type II errors.

the fraction of persons who do have a tumor among the persons that have a positive test result. To compute this quantity, we have to invoke Bayes theorem for conditional probabilities which states that, for two events A and B

$$\mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A).$$

Applying this to our case, we get:

$$PV^+ = \mathbb{P}\left(T|P\right) = \frac{\mathbb{P}(P|T)\mathbb{P}(T)}{\mathbb{P}(P)}$$

where we start to recognize the sensitivity at the numerator. The quantity at the denominator can also be computed in terms of sensitivity and specificity using Bayes formula one more time:

$$\begin{aligned}\mathbb{P}(P) &= \mathbb{P}(P|H_0)\mathbb{P}(H_0) + \mathbb{P}(P|T)\mathbb{P}(T) = (1 - \mathbb{P}(N|H_0))\mathbb{P}(H_0) + \mathbb{P}(T)SP \\ &= (1 - SP)(1 - \mathbb{P}(T)) + \mathbb{P}(T)SE\end{aligned}$$

where SE and SP stand for the specificity and sensitivity of the test (usually assessed during the experimental or clinical study). Therefore, we see that one last quantity remains to be determined to compute the PPV is the prevalence $\mathbb{P}(T)$, often denoted p. This quantity can be given by large epidemiological studies. In the end, we get:

$$PV^+ = \frac{SE \times p}{(1 - SP) \times (1 - p) + p \times SE}.$$

Similarly, the predictive negative value is the probability of not having the disease, conditionally to having a negative test result. Computations give:

$$PV^- = \frac{SP \times (1 - p)}{(1 - SE) \times p + (1 - p) \times SP}.$$

Example 3 (Lung cancer and smoking status)**.** *The percentage of smokers among lung cancer patients is 90%, which makes that the sensitivity of a screening test (or symptom) based on the smoking status 0.9. On the other hand, approximately 30% of the population is composed of smokers, thus the specificity of this test (true negative rate, i.e. proportion of people of don't smoke and don't have cancer) is 70%. Assuming a lifetime risk of having lung cancer of 7.19%*[1] *(= prevalence), then the predictive value of the smoking status is 18.9%.*

4.5.2 ROC curve analysis

4.5.3 Sample size

Sample size calculations require to specify: a test, desired specificity levels and power and an expected alternative.

References:
B Rosner. Fundamentals of biostatistics, 2015
Good webpage with summary for power analysis and sample size calculations: http://www.3rs-reduction.co.uk/html/6__power_and_sample_size.html

[1] https://www.cancer.org/cancer/cancer-basics/lifetime-probability-of-developing-or-dying-from-cancer.html

5 Model selection

5.1 Information-theoretic approach. Akaike Information Criterion (AIC)

This information criterion can be applied in the context of likelihood maximization (estimator $\hat{\theta}_{MV}$). As can be found in [BA03, p.61] (see also the original paper of Akaike [Aka74]), its general expression is

$$AIC = -2l(\hat{\theta}_{MV}) + 2K$$

where $l(\theta) = log(L(\theta))$ is the log-likelihood and $K = p + 1$ (the $+1$ being because in maximum likelihood estimation, the σ of the error model is an additional parameter (to the parameters of the structural model f), subject to optimization.

5.1.1 Classical least squares. Constant error

When a constant variance error model is considered, which is often the case in classical least squares regression, the formula becomes

$$AIC = n \log\left(\frac{1}{n}\sum_{i=1}^{n} |y_i - f(t_i,\theta)|^2\right) + n\left(\log(2\pi) + 1\right) + 2(p+1)$$

The AIC being meaningful only when comparing two models and the third term being a constant depending only on n, it is often omitted, leading to (this formula can be found in [BA03, p.63], [MC04, p.143] and [**?**, p.25])

$$AIC = n \log\left(\frac{1}{n}\sum_{i=1}^{n} |y_i - f(t_i,\theta)|^2\right) + 2(p+1)$$

5.1.2 Proportional error

In view of (4) for the expression of $l(\theta)$, under a gaussian error model with proportional variance, this leads to

$$AIC = n \log\left(\frac{1}{n}\sum_{i=1}^{n} \left|\frac{y_i - f(t_i,\theta)}{f(t_i,\theta)}\right|^2\right) + 2\sum_{i=1}^{n} \log\left(f(t_i,\theta)\right) + 2(p+1)$$

Under a variance error model that is proportional to the data rather than the model (see remark 4), the second term is $2\sum_{i=1}^{n} \log(y_i)$, does not depend on the model anymore and could also be omitted, leading to the formula

$$AIC = n \log\left(\frac{1}{n}\sum_{i=1}^{n} \left|\frac{y_i - f(t_i,\theta)}{y_i}\right|^2\right) + 2(p+1)$$

5.2 Frequentist point of view

References for the formula and mention of the fact that the likelihood ratio statistic is asymptotically chi-squared distributed: [Hoe62], chap.9.1.4 and [**?**]. See also [**?**] for a simple exposition of the proof.

Another test that can be performed from the residuals is an F-test [**?**].

6 Historique

- Premiers moindres carrés : Legendre en 1805, qui montre aussi que l'estimateur des moindres carrés est le meilleur lorsque la distribution des erreurs est normale (où il est alors égal au maximum de vraisemblance).
- L'estimation statistique débute en 1900 avec Pearson.
- R. A. Fisher introduit dans les années 1920 -1930 l'estimateur du maximum de vraisemblance. "The method of maximum likelihood is usually credited to the English statistician RA Fisher, who described the method in 1922" [HC97].
- Type I and type II errors in 1928 by Neyman and Pearson [NP28, NP33]
- L'application dans le cadre de modèles vient de Koopman (1930's), en économétrie.
- La résolution numérique des moindres carrés non-linéaires date de 1958 (Booth et Peterson).

References

[1] Seber GAF, Wild CJ, Nonlinear Regression Analysis. Wiley Series in Probability and Mathematical Statistics, ed. Barnett V, Bradley RA, Hunter JS, Kendall DG, Miller RG, Smith AFM, Stigler SM, Watson GS. 1989, New York: John Wiley. 768.

[2] Bard Y, Nonlinear parameter estimation. 1974, New York: Academic Press. 341.

[3] Poly de statistiques de G. Biau et A. Tsybakov.

[4] H. Motulsky et A. Christopoulos, Fitting models to biological data using linear and nonlinear regression.

References

[Aka74] H Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, December 1974.

[BA03] K P Burnham and D R Anderson. Model selection and multimodel inference: a practical information-theoretic approach. Springer Science & Business Media, 2003.

[BWS+14] Mauricio Burotto, Julia Wilkerson, Wilfred Stein, Robert Motzer, Susan Bates, and Tito Fojo. Continuing a Cancer Treatment Despite Tumor Growth May Be Valuable: Sunitinib in Renal Cell Carcinoma as Example. *PLoS ONE*, 9(5):e96316–128, May 2014.

[Cou] *Cours Pharmacocinétique*.

[HC97] John P Huelsenbeck and Keith A Crandall. PHYLOGENY ESTIMATION AND HYPOTHESIS TESTING USING MAXIMUM LIKELIHOOD. *Annu. Rev. Ecol. Syst.*, 28(1):437–466, November 1997.

[Hoe62] P G Hoel. *Introduction to mathematical statistics*. Wiley & Sons, New York, 3rd edition edition, 1962.

[KTL+84] S Koscielny, M Tubiana, M G Le, A Valleron, H Mouriesse, G Contesso, and D Sarrazin. Breast cancer: relationship between the size of the primary tumour and the probability of metastatic dissemination. *Br J Cancer*, 49(6):709–15, June 1984.

[MAM+07] Amol Mujoomdar, John H M Austin, Rohin Malhotra, Charles A Powell, Gregory D N Pearson, Maria C Shiau, and Haralambos Raftopoulos. Clinical Predictors of Metastatic Disease to the Brain from Non–Small Cell Lung Carcinoma: Primary Tumor Size, Cell Type, and Lymph Node Metastases1. *Radiology*, 242(3):882–888, March 2007.

[MC04] H Motulsky and Arthur Christopoulos. *Fitting models to biological data using linear and nonlinear regression*. Oxford University Press, 2004.

[NP28] J Neyman and E S Pearson. On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I. *Biometrika*, 20A(1/2):175–240, 1928.

[NP33] J Neyman and E S Pearson. The testing of statistical hypotheses in relation to probabilities a priori. *Mathematical Proceedings of the Cambridge Philosophical Society*, 29(04):492–510, October 1933.

[Pea00] Karl Pearson. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 50(302):157–175, July 1900.

[RIG89] C Ressayre, A Iliadis, and J Gouvernet. Simultaneous modelling of a nonlinear multiresponse pharmacokinetic system. *Applied Mathematical Modelling*, 1989.

[Ros15] B Rosner. Fundamentals of biostatistics, 2015.

[SWK+12] Wilfred D Stein, Julia Wilkerson, Sindy T Kim, Xin Huang, Robert J Motzer, Antonio Tito Fojo, and Susan E Bates. Analyzing the pivotal trial that compared sunitinib and IFN-α in renal cell carcinoma, using a method that assesses tumor regression and growth. *Clin Cancer Res*, 18(8):2374–2381, April 2012.