

Statistique de base

Ce document rassemble des résultats élémentaires de statistiques. Il reprend une grande partie des notes du cours de Bertrand Michel, et s'appuie sur le livre *Statistique en action* de Rivoirard et Stoltz (Vuibert, 2009).

1 Le Modèle statistique

Etant donnée une certaine expérience aléatoire, le statisticien construit d'abord un modèle statistique censé modéliser cette expérience. L'observation \mathbf{Y} est le résultat de cette expérience. Si \mathbf{y} est une réalisation de \mathbf{Y} , on aimerait s'aider de cette information pour en déduire la loi de \mathbf{Y} . Si nous ne faisons aucune hypothèse sur la loi de \mathbf{Y} , on dit que le modèle est non-paramétrique. Si on suppose que la loi de \mathbf{Y} est de forme connue mais dépend d'un nombre fini de paramètres réels qui sont inconnus, on dira que le modèle est paramétrique.

Soit (E, \mathcal{E}) un espace mesurable.

Définition. On appelle modèle statistique la donnée de $(E, \mathcal{E}, (P_\theta)_{\theta \in \Theta})$ où $(P_\theta)_{\theta \in \Theta}$ désigne une famille de lois de probabilités sur (E, \mathcal{E}) . On notera E_θ l'espérance associée à P_θ et V_θ la variance.

Si $\Theta \subset \mathbb{R}^d$, on dit que le modèle est *paramétrique*. Sinon le modèle est dit *non-paramétrique*.

Définition. Une observation \mathbf{Y} est une variable aléatoire à valeurs dans (E, \mathcal{E}) dont la loi appartient à la famille de lois $(P_\theta)_{\theta \in \Theta}$.

Définition. Lorsque l'observation \mathbf{Y} a la forme $\mathbf{Y} = (Y_1, \dots, Y_n)$ avec $(Y_i)_{1 \leq i \leq n}$ indépendantes et identiquement distribuées, on parlera d'échantillon. Dans ce cas, $P_\theta = p_\theta^{\otimes n}$ où p_θ est la loi de Y_1 .

Exemples.

Sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ on considère $P_\theta = \mathcal{N}(\mu, \sigma^2)$ avec $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$. L'observation \mathbf{Y} est la donnée d'une variable Gaussienne $\mathcal{N}(\mu, \sigma^2)$.

Modèle d'un lancer de pièces. On lance une pièce n fois. On a $E = \{0, 1\}^n$, \mathcal{E} la tribu triviale, p_θ est la loi de Bernoulli de paramètre $\theta \in \Theta = [0, 1]$ et $P_\theta = p_\theta^{\otimes n}$. L'observation \mathbf{Y} est un échantillon de Bernoulli.

Soit une certaine variable aléatoire sans atomes à valeurs dans $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ dont on cherche à connaître la loi. Dans ce cas, on peut prendre pour Θ l'ensemble des lois sans atomes ou de manière équivalente l'ensemble des fonctions θ continues croissantes de \mathbb{R} sur $[0, 1]$ avec $\theta(-\infty) = 0$ et $\theta(+\infty) = 1$, puis P_θ est la loi de fonction de répartition θ . Ce modèle est non-paramétrique.

2 Estimateurs

2.1 Premières propriétés

Soit g une fonction de Θ dans \mathbb{R}^p .

Définition. Un estimateur \hat{g} de $g(\theta)$ est toute application mesurable en l'observation \mathbf{Y} ne dépendant pas de θ .

On peut donc écrire $\hat{g} = h(\mathbf{Y})$ pour une certaine fonction mesurable $h : (E, \mathcal{E}) \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}^p))$.

Définition.

- La fonction de biais d'un estimateur \hat{g} est (si elle existe) la fonction $\theta \in \Theta \rightarrow b(\theta) := E_\theta[\hat{g}] - g(\theta)$.
- L'erreur quadratique d'un estimateur \hat{g} est la fonction $\theta \in \Theta \rightarrow R(\theta) := E_\theta[(\hat{g} - g(\theta))^2]$. Sa décomposition biais-variance s'écrit

$$R(\theta) = V_\theta(\hat{g}) + b(\theta)^2$$

Définition. On dit qu'un estimateur \hat{g} est **sans biais** si $E_\theta[\hat{g}] = g(\theta)$ pour tout $\theta \in \Theta$.

La définition d'un estimateur sans biais contient que \hat{g} est P_θ -intégrable pour tout $\theta \in \Theta$. Un estimateur sans biais donne donc en moyenne la bonne valeur de ce qu'il estime, ce qui est satisfaisant. Cependant, c'est un critère parfois contraignant (dans certains cas un estimateur sans biais n'existe pas, dans d'autres cas un estimateur "naturel" est avec biais). L'erreur quadratique mesure la dispersion des valeurs données par l'estimateur autour de $g(\theta)$, donc sa précision.

Soit un échantillon $\mathbf{Y} = (Y_1, \dots, Y_n)$ et $\hat{g}_n = h_n(Y_1, \dots, Y_n)$ un estimateur de $g(\theta)$. Les propriétés suivantes permettent de connaître l'erreur commise par l'estimation lorsque le nombre de répétitions devient grand.

Définition. On dit que l'estimateur \hat{g}_n est

- **asymptotiquement sans biais** si $\lim_{n \rightarrow +\infty} E_\theta[\hat{g}_n] = g(\theta)$ pour tout $\theta \in \Theta$.
- **consistant** si pour tout $\theta \in \Theta$, \hat{g}_n converge vers $g(\theta)$ en probabilité.
- **fortement consistant** si la convergence a lieu p.s.
- **asymptotiquement normal** si, pour tout $\theta \in \Theta$, $a_n(\hat{g}_n - g(\theta))$ converge en loi vers une loi Gaussienne centrée, pour une certaine suite déterministe a_n (dépendant éventuellement de θ) telle que $\lim_{n \rightarrow +\infty} a_n = +\infty$.

2.2 Estimateurs classiques

On donne ici deux méthodes pour trouver des estimateurs.

2.2.1 Méthode des moments

Soit $\mathbf{Y} = (Y_1, \dots, Y_n)$ un échantillon. Supposons que la quantité $g(\theta)$ que l'on cherche à estimer est donnée par $E_\theta[\phi(Y_1)]$ pour une certaine fonction mesurable ϕ (ne dépendant pas de θ) avec $\phi(Y_1)$ P_θ -intégrable pour tout $\theta \in \Theta$. Dans ce cas,

$$\hat{g}_n = \frac{1}{n} \sum_{i=1}^n \phi(Y_i).$$

est un estimateur fortement consistant par la loi des grands nombres. Si la variance de $\phi(Y_1)$ sous P_θ est finie pour tout $\theta \in \Theta$, alors \hat{g}_n est asymptotiquement normal par le théorème central limite, avec $a_n = \sqrt{n}$.

Exemple. On peut ainsi estimer le moment d'ordre p (s'il existe) de p_θ $\nu_p := E_\theta[Y_1^p]$ par

$$\hat{\nu}_{p,n} := \frac{1}{n} \sum_{k=1}^n Y_k^p.$$

Exemple. Un estimateur sans biais et fortement consistant de la variance est

$$\hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{k=1}^n (Y_k - \bar{Y}_n)^2.$$

2.2.2 Méthode du maximum de vraisemblance

Dans cette partie, on prend $g(\theta) = \theta$ cad que l'on souhaite estimer θ . Grossièrement, étant donnée notre observation \mathbf{y} , le principe de la méthode est de trouver la valeur de θ qui maximise la probabilité de voir l'observation \mathbf{y} .

Revenons à notre modèle statistique $(E, \mathcal{E}, (P_\theta)_{\theta \in \Theta})$. On se place dans le cas paramétrique $\Theta \subset \mathbb{R}^d$. On suppose que les lois $(P_\theta, \theta \in \Theta)$ du modèle statistique sont dominées par une même mesure μ supposée σ -finie. Le théorème de Radon-Nikodym implique que P_θ a alors une densité que l'on note $L_\theta(\mathbf{y})$ par rapport à μ , c'est-à-dire que pour tout ensemble mesurable $A \in \mathcal{E}$, on a $P_\theta(A) = \int_{\mathbf{y} \in A} L_\theta(\mathbf{y}) d\mu(\mathbf{y})$. Cette densité $L_\theta(\mathbf{y})$ est appelée la vraisemblance du modèle.

Définition. L'estimateur du maximum de vraisemblance noté $\hat{\theta}_{\text{EMV}}$ est, s'il en existe, un θ qui maximise la vraisemblance

$$\theta \rightarrow L_\theta(\mathbf{Y}).$$

Dans le cas d'un échantillon $\mathbf{Y} = (Y_1, \dots, Y_n)$, on supposera que la loi p_θ de Y_1 (donc aussi des Y_i) a une densité $s_\theta(y_1)$ par rapport à une mesure de référence σ -finie $\mu(dy_1)$. La vraisemblance s'écrira alors

$$L_\theta(\mathbf{y}) = L_\theta(y_1, \dots, y_n) = \prod_{i=1}^n s_\theta(y_i)$$

et l'estimateur du maximum de vraisemblance sera le θ qui maximise la fonction $\theta \rightarrow L_\theta(\mathbf{Y})$.

Sous des hypothèses de régularité, on peut montrer que l'estimateur du maximum de vraisemblance est asymptotiquement normal.

2.3 Estimateur de variance minimale

Supposons que l'on veuille estimer $g(\theta)$. On aimerait comparer plusieurs estimateurs. Une idée naturelle est de comparer l'erreur commise par ces estimateurs.

Définition. *Un estimateur sans biais de variance minimale (ESBVM) est un estimateur \hat{g}_{ESBVM} sans biais qui minimise l'erreur quadratique. En d'autres termes,*

$$V_{\theta}(\hat{g}_{ESBVM}) \leq V_{\theta}(\hat{g})$$

pour tout $\theta \in \Theta$ et tout \hat{g} estimateur sans biais de $g(\theta)$. S'il existe, un tel estimateur est nécessairement unique.

Remarque. Rien ne nous dit qu'on ne peut pas trouver un estimateur biaisé avec une erreur quadratique plus petite.

La suite de cette partie montre que l'on a une borne inférieure sur la variance des estimateurs sans biais, appelée *borne de Cramer-Rao*. Cette borne inférieure dépend de la quantité d'information apportée par les observations telle que mesurée par *l'information de Fisher*.

On se place désormais dans le cas paramétrique. On a donc $\Theta \subset \mathbb{R}^p$ et supposons pour simplifier que $p = 1$.

Définition. *L'information de Fisher est définie par*

$$I(\theta) := E_{\theta} \left[(\partial_{\theta} \ln L_{\theta}(\mathbf{Y}))^2 \right]$$

lorsque cette quantité existe. Elle peut se réécrire sous certaines hypothèses de régularité

$$I(\theta) = -E_{\theta} \left[\partial_{\theta}^2 \ln L_{\theta}(\mathbf{Y}) \right].$$

Dans le cas d'un échantillon $\mathbf{Y} = (Y_1, \dots, Y_n)$, on obtient (sous hypothèses de régularité) que l'information de Fisher du modèle est donnée par

$$I_n(\theta) = nI_1(\theta)$$

avec $I_1(\theta) = E_{\theta} \left[(\partial_{\theta} \ln s_{\theta}(Y_1))^2 \right]$, qui s'écrit encore $-E_{\theta} \left[\partial_{\theta}^2 \ln s_{\theta}(Y_1) \right]$.

Intuitivement, l'information de Fisher est une mesure de l'information contenue dans l'observation \mathbf{Y} . Plus $I(\theta)$ est élevée, meilleure sera l'information et donc plus précis pourront être les estimateurs. Cette heuristique se retrouve dans le théorème suivant qui dit que l'erreur commise par un estimateur sans biais est bornée inférieurement par l'inverse de l'information de Fisher.

Théorème (Borne de Cramer-Rao). *Soit g une fonction $\Theta \rightarrow \mathbb{R}$ dérivable, et $\hat{g} = h(\mathbf{Y})$ un estimateur sans biais de $g(\theta)$. Sous certaines hypothèses de régularité, on a*

$$V_{\theta}(\hat{g}) \geq \frac{(g'(\theta))^2}{I(\theta)}.$$

La borne inférieure de cette inégalité est appelée *borne de Cramer-Rao*.

Un estimateur sans biais qui atteint la borne de Cramer-Rao est nécessairement de variance minimale. Un tel estimateur est dit *efficace*.

Dans le cas d'un échantillon, si \hat{g}_n est sans biais et vérifie quand $n \rightarrow +\infty$, $V_\theta(\hat{g}_n) \sim \frac{g'(\theta)^2}{I_n(\theta)}$, on dit que l'estimateur est *asymptotiquement efficace*. Sous des hypothèses de régularité, l'estimateur du maximum de vraisemblance est asymptotiquement efficace.

L'existence d'un estimateur efficace est en fait liée à la forme de la vraisemblance du modèle.

Théorème. (Sous certaines hypothèses de régularité). *Un estimateur efficace existe si et seulement si la vraisemblance vérifie*

$$\ln L_\theta(\mathbf{y}) = a(\mathbf{y})\alpha(\theta) + b(\mathbf{y}) + \beta(\theta).$$

Dans ce cas, $g(\theta) = -\frac{\beta'(\theta)}{\alpha'(\theta)}$ admet un estimateur efficace qui est $\hat{g} := a(\mathbf{Y})$. C'est l'unique paramètre (à une transformation linéaire près) admettant un estimateur efficace.

3 Intervalles de confiance

3.1 Définition

Soit $(E, \mathcal{E}, (P_\theta)_{\theta \in \Theta})$ un modèle statistique et \mathbf{Y} une observation. On souhaite estimer le paramètre $g(\theta)$.

Définition. Soit $\alpha \in [0, 1]$. On appelle *région de confiance au niveau $1 - \alpha$ de $g(\theta)$* un ensemble \hat{C} construit mesurablement par rapport à \mathbf{Y} , ne dépendant pas de θ , et tel que pour tout $\theta \in \Theta$,

$$P_\theta \left(g(\theta) \in \hat{C} \right) \geq 1 - \alpha.$$

Remarque. Dire que \hat{C} est construit mesurablement signifie que pour tout $\theta \in \Theta$, l'évènement $\{g(\theta) \in \hat{C}\}$ est mesurable. Si l'inégalité est en fait une égalité dans la définition précédente, on parle de niveau exact.

Remarque. Lorsque \hat{C} est un intervalle, on parlera plutôt d'intervalle de confiance.

Définition. Dans le cas d'un échantillon $\mathbf{Y} = (Y_1, \dots, Y_n)$, on appelle *région (resp. intervalle) de confiance asymptotique au niveau $1 - \alpha$ de $g(\theta)$* un ensemble (resp. un intervalle) \hat{C}_n construit mesurablement par rapport à \mathbf{Y} , ne dépendant pas de θ , et tel que pour tout $\theta \in \Theta$,

$$\liminf_{n \rightarrow +\infty} P_\theta \left(g(\theta) \in \hat{C}_n \right) \geq 1 - \alpha.$$

3.2 Méthode du pivot

La méthode du pivot consiste à trouver une fonction $f(\mathbf{y}, g(\theta))$ mesurable en $\mathbf{y} \in E$ dont la loi sous P_θ ne dépend pas de θ . On cherche ensuite a, b tels que $P_\theta(f(\mathbf{Y}, g(\theta)) \in [a, b]) \geq 1 - \alpha$. La région de confiance est alors déterminée par $\hat{\mathcal{C}} := \{g(\theta) : f(\mathbf{Y}, g(\theta)) \in [a, b]\}$.

Exemple. On veut estimer la moyenne μ d'une loi Gaussienne $\mathcal{N}(\mu, \sigma^2)$ où σ^2 est connue. Pour cela on a accès à un échantillon $\mathbf{Y} = (Y_1, \dots, Y_n)$. Un estimateur est donné par

$$\hat{\mu}_n = \bar{Y}_n := \frac{1}{n} (Y_1 + \dots + Y_n).$$

On sait que $\frac{\hat{\mu}_n - \mu}{\sigma/\sqrt{n}}$ suit une loi Gaussienne $\mathcal{N}(0, 1)$. En notant z_β le quantile d'ordre β de la loi Gaussienne centrée réduite, on obtient $\hat{\mathcal{C}}_n := [\hat{\mu}_n \pm \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}]$ pour un intervalle de confiance bilatère de niveau $1 - \alpha$. Un intervalle de confiance unilatère serait $\hat{\mathcal{C}}_n :=]-\infty; \hat{\mu}_n + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}]$ (à gauche) ou $[\hat{\mu}_n - \frac{\sigma}{\sqrt{n}} z_{1-\alpha}; +\infty[$ (à droite).

Exemple. Soit un échantillon $\mathbf{Y} = (Y_1, \dots, Y_n)$ de v.a. moyenne μ et variance σ^2 toutes deux inconnues. On ne suppose plus que les v.a. suivent une loi Gaussienne. Le théorème central limite dit que $\frac{\hat{\mu}_n - \mu}{\sigma/\sqrt{n}}$ converge en loi vers $\mathcal{N}(0, 1)$. On estime σ par son estimateur $\hat{\sigma}_n$. Le lemme de Slutsky entraîne que $\frac{\hat{\mu}_n - \mu}{\hat{\sigma}_n/\sqrt{n}}$ converge en loi vers $\mathcal{N}(0, 1)$. Cela permet d'obtenir l'intervalle de confiance asymptotique de niveau $1 - \alpha$ pour μ :

$$\left[\hat{\mu}_n - \frac{\hat{\sigma}_n}{\sqrt{n}} z_{1-\alpha/2}; \hat{\mu}_n + \frac{\hat{\sigma}_n}{\sqrt{n}} z_{1-\alpha/2} \right].$$

3.3 Utilisation d'une inégalité de probabilité

Supposons que l'on veuille estimer la moyenne $\mu(\theta)$ d'une loi de probabilité dont on sait que la variance est bornée par une constante connue M^2 . Pour cela on a accès à un échantillon $\mathbf{Y} = (Y_1, \dots, Y_n)$. Un estimateur sans biais est donné par

$$\hat{\mu}_n = \bar{Y}_n = \frac{1}{n} (Y_1 + \dots + Y_n).$$

L'inégalité de Bienaymé-Tchebychev implique que

$$P_\theta (|\hat{\mu}_n - \mu| > t) \leq \frac{V_\theta(\hat{\mu}_n)}{t^2} \leq \frac{M^2}{nt^2}.$$

En choisissant t tel que $\frac{M^2}{nt^2} = \alpha$, on obtient l'intervalle de confiance de niveau $1 - \alpha$ de μ suivant

$$\left[\hat{\mu}_n - \frac{M}{\sqrt{\alpha n}}; \hat{\mu}_n + \frac{M}{\sqrt{\alpha n}} \right].$$

Exemple. Intervalle de confiance pour le paramètre d'un échantillon de Bernoulli.

3.4 Intervalle de confiance et réalisation d'un intervalle de confiance

Il est important de garder à l'esprit qu'un intervalle de confiance est une quantité aléatoire. Dans la pratique, c'est-à-dire pour les données étudiées, les bornes de l'intervalle de confiance construit sont la réalisation de l'intervalle de confiance pour les observations. Par exemple, dans le cas de l'échantillon Gaussien i.i.d., on a vu que

$$\left[\hat{\mu}_n \pm \frac{\hat{\sigma}_n t_{0,975}}{\sqrt{n}} \right]$$

est un intervalle de confiance de la moyenne au niveau 95 %. Si $n = 100$, $\hat{\mu}_n = 7,45$ et $\hat{\sigma}_n = 1,21$, la réalisation de cet intervalle vaut $[6,76; 7,24]$. Par abus, on dira parfois que $[6,76; 7,24]$ est l'intervalle de confiance de la moyenne, mais en toute rigueur cela n'a pas de sens : soit la moyenne est dans $[6,76; 7,24]$, soit elle ne l'est pas, la probabilité que la moyenne y soit est donc 0 ou 1!

3.5 Intervalles de confiance simultanés

Supposons que l'on souhaite construire une région de confiance pour un vecteur

$$g(\theta) = (g_1(\theta), g_2(\theta), \dots, g_k(\theta)).$$

Supposons que l'on dispose pour chaque $g_i(\theta)$ pris séparément d'un intervalle de confiance \hat{I}_j de niveau de confiance $1 - \frac{\alpha}{k}$. Alors $\hat{I}_1 \times \hat{I}_2 \times \dots \times \hat{I}_k$ est une région de confiance pour le vecteur $g(\theta)$ de niveau $1 - \alpha$ (méthode de Bonferroni). En effet, pour tout j ,

$$P_\theta \left(g_j(\theta) \notin \hat{I}_j \right) \leq \frac{\alpha}{k}.$$

D'où $P_\theta \left(\bigcup_{j=1}^k \{g_j(\theta) \notin \hat{I}_j\} \right) \leq k \frac{\alpha}{k}$ et donc

$$P_\theta \left(\bigcap_{j=1}^k \{g_j(\theta) \in \hat{I}_j\} \right) \geq 1 - \alpha.$$

Notons que cette méthode n'est intéressante que pour de petites valeurs de k . Pour de grandes valeurs de k , la région de confiance $\hat{I}_{1,1-\frac{\alpha}{k}} \times \hat{I}_{2,1-\frac{\alpha}{k}} \times \dots \times \hat{I}_{k,1-\frac{\alpha}{k}}$ a une probabilité beaucoup plus grande que $1 - \alpha$: elle encadre $g(\theta)$ beaucoup trop largement.

4 Tests d'hypothèses

4.1 Introduction

Exemple Une variété de souris présente des cancers spontanés avec un taux de 20%. On aimerait connaître l'efficacité d'un traitement. On fait donc l'expérience sur un échantillon de 200 souris, et on aimerait décider si le traitement a un effet ou non.

On formule donc deux hypothèses:

- \mathcal{H}_0 : le traitement est sans effet.
- \mathcal{H}_1 : le traitement a un effet.

On veut choisir entre l'hypothèse \mathcal{H}_0 (*hypothèse nulle*) et \mathcal{H}_1 (*hypothèse alternative*). Pour cela, on veut effectuer un test, c'est-à-dire que l'on veut se donner une règle pour trancher entre les deux hypothèses en fonction des résultats de l'expérience.

L'étape du choix des hypothèses n'est pas anodine. Les hypothèses nulle et alternative ne sont pas symétriques. On privilégie l'hypothèse \mathcal{H}_0 : je vais rejeter \mathcal{H}_0 si je suis vraiment sûr que \mathcal{H}_0 est fautive. Les deux hypothèses ne jouant pas le même rôle, dans la pratique on choisira pour H_0 l'hypothèse "la plus raisonnable", la plus communément admise, l'hypothèse privilégiée par un parti pris subjectif ou encore l'hypothèse la plus simple. Ainsi, dans un même contexte où l'on souhaite confronter deux hypothèses \mathcal{H}_a et \mathcal{H}_b , deux tests statistiques définis par $\mathcal{H}_0 = \mathcal{H}_a$, $\mathcal{H}_1 = \mathcal{H}_b$ pour le premier et par $\mathcal{H}_0 = \mathcal{H}_b$, $\mathcal{H}_1 = \mathcal{H}_a$ pour le second, peuvent aboutir à des conclusions différentes. On verra une illustration de ce phénomène dans le cas du test de Student.

Une fois que je rejette \mathcal{H}_0 , j'accepte alors l'hypothèse \mathcal{H}_1 . Sinon, j'accepte \mathcal{H}_0 , ce qui ne veut pas dire que \mathcal{H}_0 est nécessairement vraie. Cela voudra juste dire que je n'ai pas assez de certitude pour préférer \mathcal{H}_1 à \mathcal{H}_0 . C'est pourquoi l'on dit qu'un test est *significatif* lorsque la conclusion du test est que l'on rejette \mathcal{H}_0 .

La démarche.

On suppose que \mathcal{H}_0 est vraie. Dans ce cas, le nombre de souris malades suit une loi binômiale $\mathcal{B}(N, p)$ avec $N = 200$ et $p = 0.2$. La proportion \hat{p} de souris malades est une *statistique de test*. C'est une variable aléatoire dont on peut identifier la distribution sous \mathcal{H}_0 . C'est elle qui va nous dire si \mathcal{H}_0 est acceptable ou non. On calcule que la probabilité que \hat{p} (dont on connaît la distribution) se trouve dans l'intervalle $[0.1984, 0.2016]$ est de 95%.

La règle de décision sera donc la suivante: si $\hat{p} \notin [0.1984, 0.2016]$, je rejette \mathcal{H}_0 et choisis donc \mathcal{H}_1 . Sinon je ne rejette pas \mathcal{H}_0 .

La probabilité de rejeter \mathcal{H}_0 alors qu'elle est vraie est, par construction, égale à $\alpha = 5\%$. On appelle α le *risque de première espèce*. C'est ce risque que l'on cherche à minimiser dans un premier temps.

La probabilité β d'accepter \mathcal{H}_0 à tort est appelée *risque de deuxième espèce*. La puissance est égale à $1 - \beta$. À α fixé, on cherchera un test le plus puissant possible.

L'hypothèse \mathcal{H}_1 influe sur la règle de décision. Supposons que l'on prenne maintenant pour \mathcal{H}_1 l'hypothèse que le traitement est bénéfique, cad que la proportion de souris malades diminue avec le traitement. Notre test d'hypothèses est donc:

- \mathcal{H}_0 le traitement est sans effet.
- \mathcal{H}_1 le traitement a un effet bénéfique.

Il va falloir choisir entre \mathcal{H}_0 et \mathcal{H}_1 (même dans le cas où ni l'une ni l'autre ne sont vraies!). On pourrait utiliser la règle de décision précédente. Cependant, on sera amené à accepter l'hypothèse \mathcal{H}_1 lorsque \hat{p} est grand, ce qui est contre-intuitif: on ne veut pas dire que le traitement est bénéfique si le nombre de souris malades est plus élevé que la normale! On va plutôt rejeter \mathcal{H}_0 et donc accepter \mathcal{H}_1 si la proportion de souris malades est faible. Cela donnera un test plus puissant. Sous l'hypothèse \mathcal{H}_0 , la probabilité que $\hat{p} \leq 0.1986$ est de 5%. La règle de décision sera donc: si $\hat{p} \leq 0.1986$ je rejette H_0 (et choisis donc \mathcal{H}_1), sinon je choisis \mathcal{H}_0 . C'est ce qu'on appelle un test *unilatéral*, comparé au précédent qui était un test *bilatéral*.

4.2 Définitions

Soit $(E, \mathcal{E}, (P_\theta)_{\theta \in \Theta})$ un modèle statistique. Soit Θ_0 et Θ_1 deux sous-ensembles disjoints de Θ . On souhaite tester $\mathcal{H}_0 : \theta \in \Theta_0$ contre $\mathcal{H}_1 : \theta \in \Theta_1$. Lorsque Θ_i est réduit à un singleton, on parle d'hypothèse simple (par opposition avec hypothèse composite). Dans l'exemple précédent, on avait P_θ loi d'un échantillon de Bernoulli de paramètre $\theta \in \Theta = [0, 1]$, $\Theta_0 = \{0.2\}$ et $\Theta_1 = \Theta \setminus \{0.2\}$ puis $\Theta_1 =]0.2, 1]$.

Définition. *Un test est une application mesurable $\Phi : E \rightarrow \{0, 1\}$ associée à la stratégie suivante: pour l'observation \mathbf{y} , \mathcal{H}_1 est acceptée si $\Phi(\mathbf{y}) = 1$ et \mathcal{H}_0 est acceptée si $\Phi(\mathbf{y}) = 0$.*

La région $\mathcal{R} = \{\mathbf{y} : \Phi(\mathbf{y}) = 1\}$ est la zone de rejet du test. On rejettera ainsi \mathcal{H}_0 si $\mathbf{Y} \in \mathcal{R}$ et on l'acceptera sinon.

Définition. *L'erreur de première espèce du test Φ est l'application*

$$\begin{aligned} \Theta_0 &\rightarrow [0, 1] \\ \theta &\rightarrow P_\theta(\Phi = 1) \end{aligned}$$

Le niveau du test Φ est défini par $\alpha = \sup_{\theta \in \Theta_0} P_\theta(\Phi = 1)$.

Avoir un niveau α signifie que la probabilité de rejeter à tort l'hypothèse \mathcal{H}_0 est inférieure à α .

Définition. *L'erreur de seconde espèce du test Φ est l'application*

$$\begin{aligned} \Theta_1 &\rightarrow [0, 1] \\ \theta &\rightarrow P_\theta(\Phi = 0) \end{aligned}$$

La puissance du test Φ est l'application

$$\begin{aligned} \Theta_1 &\rightarrow [0, 1] \\ \pi : \theta &\rightarrow P_\theta(\Phi = 1). \end{aligned}$$

La fonction puissance est la probabilité d'accepter \mathcal{H}_1 quand celle-ci est vraie. Elle mesure donc la qualité du test. En général, diminuer l'erreur de première espèce se fait au détriment de la puissance. La démarche consiste à se fixer d'abord un niveau du test puis à trouver un test maximisant la puissance.

Exemple. On considère un échantillon $\mathbf{Y} = (Y_1, \dots, Y_n)$ de v.a. indépendantes de loi $\mathcal{N}(\theta, \sigma^2)$ avec $\theta \in \mathbb{R}$ et σ^2 connu. On veut tester $\mathcal{H}_0 : \theta = \theta_0$ contre $\mathcal{H}_1 : \theta \neq \theta_0$ où θ_0 est un réel fixé. Prenons le test $\Phi = \mathbf{1}_{\{|\sqrt{n}(\bar{Y}_n - \theta_0)| > z_{1-\alpha/2}\sigma\}}$. Alors le niveau du test est

$$P_{\theta_0}(\Phi = 1) = P_{\theta_0}(|\sqrt{n}(\bar{Y}_n - \theta_0)| > z_{1-\alpha/2}\sigma) = P_{\theta_0}(|Z| > z_{1-\alpha/2}) = \alpha$$

où Z est une variable Gaussienne centré réduite et z_β le quantile β de sa loi. La fonction puissance est donnée par

$$P_\theta(\Phi = 1) = P_\theta(|\sqrt{n}(\bar{Y}_n - \theta_0)| > z_{1-\alpha/2}\sigma) = P_\theta\left(\left|Z + \frac{\theta - \theta_0}{\sigma\sqrt{n}}\right| > z_{1-\alpha/2}\right)$$

où Z est encore une variable Gaussienne centrée réduite.

Exemple. Si on teste plutôt $\mathcal{H}_0 : \theta = \theta_0$ contre $\mathcal{H}_1 : \theta > \theta_0$, on considèrera le test $\Phi = \mathbf{1}_{\{\sqrt{n}(\bar{Y}_n - \theta_0) > z_{1-\alpha}\sigma\}}$ qui est encore de niveau α . Sa fonction puissance est donnée par

$$P_\theta(\Phi = 1) = P_\theta(\sqrt{n}(\bar{Y}_n - \theta_0) > z_{1-\alpha}\sigma) = P_\theta\left(Z + \frac{\theta - \theta_0}{\sigma\sqrt{n}} > z_{1-\alpha}\right).$$

On peut montrer que ce test (unilatéral) est en effet plus puissant que le test précédent (bilatéral). On verra dans la partie 4.5 qu'il est en fait plus puissant que n'importe quel test de niveau α . De plus, si on teste $\mathcal{H}_0 : \theta \leq \theta_0$ contre $\mathcal{H}_1 : \theta > \theta_0$, le même test $\Phi = \mathbf{1}_{\{\sqrt{n}(\bar{Y}_n - \theta_0) > z_{1-\alpha}\sigma\}}$ reste de niveau α et le plus puissant.

Remarque. Il y a un lien étroit entre intervalle de confiance et tests statistiques. Supposons que l'on veuille tester $\mathcal{H}_0 : \theta = \theta_0$ contre $\mathcal{H}_1 : \theta \neq \theta_0$. Alors il suffit de choisir un intervalle de confiance I pour θ de niveau de confiance $1 - \alpha$. On prendra alors $\Phi = \mathbf{1}_{\{\theta_0 \notin I\}}$ comme test de niveau (inférieur à) α .

4.3 Tests asymptotiques

On considère le cas d'un échantillon $\mathbf{Y} = (Y_1, \dots, Y_n)$. On teste encore $\theta \in \Theta_0$ contre $\theta \in \Theta_1$.

Définition. Un test Φ_n est dit asymptotiquement de niveau α si $\sup_{\theta \in \Theta_0} \limsup_{n \rightarrow +\infty} P_\theta(\Phi_n = 1) = \alpha$.

Définition. Un test Φ_n est dit convergent si $\lim_{n \rightarrow +\infty} P_\theta(\Phi_n = 1) = 1$ pour tout $\theta \in \Theta_1$.

Cela veut dire que l'on pourra trancher entre les deux hypothèses avec de plus en plus de certitude si le nombre d'observations est suffisant.

Soit $\mathbf{Y} = (Y_1, \dots, Y_n)$ un échantillon de moyenne μ et variance σ^2 . On aimerait tester l'égalité de la moyenne à un certain réel μ_0 .

$$\mathcal{H}_0 : \mu = \mu_0 \quad \text{contre} \quad \mathcal{H}_1 : \mu \neq \mu_0$$

On utilise la statistique de test (avec $\hat{\sigma}_n$ estimateur de l'écart-type),

$$Z_n = \frac{\bar{Y}_n - \mu_0}{\hat{\sigma}_n/\sqrt{n}}.$$

La variable Z_n converge en loi vers $\mathcal{N}(0, 1)$ (par le théorème central limite et le lemme de Slutsky). Donc le test

$$\Phi = \mathbf{1}_{\{|Z_n| > z_{1-\alpha/2}\}}$$

est asymptotiquement de niveau α .

Exercice. Faire un test asymptotique $\mathcal{H}_0 : \theta = 1/2$ contre $\mathcal{H}_1 : \theta \neq 1/2$ dans le cas d'un échantillon de Bernoulli de paramètre $\theta \in [0, 1]$ (*Test de proportion*).

4.4 Pratique et interprétation des tests

Plutôt que de donner une réponse définitive telle que “j’accepte \mathcal{H}_0 ” ou telle que “ je rejette \mathcal{H}_0 ”, il est préférable de “mesurer” la facilité avec laquelle \mathcal{H}_0 peut être acceptée. Cette mesure est donnée par la *probabilité critique* (aussi appelée *niveau de significativité* ou encore *p-value*) :

$$\begin{aligned} p &:= \max\{\alpha \in [0, 1] \mid \text{le test accepte } \mathcal{H}_0 \text{ au niveau } \alpha\} \\ &= \max\{\alpha \in [0, 1] \mid X \notin \mathcal{R}_\alpha\} \\ &= \min\{\alpha \in [0, 1] \mid \text{le test rejette } \mathcal{H}_0 \text{ au niveau } \alpha\} \\ &= \min\{\alpha \in [0, 1] \mid X \in \mathcal{R}_\alpha\} \end{aligned}$$

Il s’agit donc du niveau limite, pour lequel la réponse du test change. Notons que la définition donnée ci-dessus sous-entend que la région de rejet \mathcal{R}_α croît avec α , ce qui est naturel. Les logiciels statistiques fournissent comme réponse à un test la *p-value* correspondante, il appartient à l’utilisateur d’interpréter cette valeur :

- Supposons que la *p-value* soit de 0.65, on accepte alors \mathcal{H}_0 sans problème pour des niveaux standards (par exemple de l’ordre de 0.05).
- Supposons que la *p-value* d’un test soit de 10^{-2} , on rejette alors \mathcal{H}_0 pour des niveaux standards.

De façon générale, accepter \mathcal{H}_0 ne constitue pas une validation radicale de l’hypothèse nulle car les tests sont généralement trop conservatifs (ils privilégient \mathcal{H}_0). Accepter \mathcal{H}_0 signifie uniquement que rien d’anormal n’a été détecté dans les données qui ne contredise l’hypothèse nulle. À l’inverse, on rejette \mathcal{H}_0 lorsque les observations (ou une statistique) prennent une valeur extrême sous l’hypothèse de loi imposée par \mathcal{H}_0 . En fin de compte, la réponse à un test n’est donc claire et franche que lorsque la *p-value* est très faible ; on dit alors que le résultat du test est *significatif*.

L’étude complète d’un test d’hypothèse se déroule de la façon suivante :

1. Choix des hypothèses \mathcal{H}_0 et \mathcal{H}_1 ,
2. Détermination d’une statistique adaptée pour le problème posé, choix d’une région de rejet,
3. Étude de la fonction puissance,
4. Calcul de la *p-value* pour les données observées,
5. Décision finale, interprétation.

4.5 Test du rapport de vraisemblance

Soit $(E, \mathcal{E}, (P_\theta)_{\theta \in \Theta})$ le modèle statistique, avec \mathbf{Y} l’observation. On rappelle que $L_\theta(\mathbf{y})$ désigne la vraisemblance du modèle. Le test du rapport de vraisemblance s’appuie sur la statistique de test

$$h(\mathbf{Y}) = \frac{\sup_{\theta \in \Theta_1} L_\theta(\mathbf{Y})}{\sup_{\theta \in \Theta_0} L_\theta(\mathbf{Y})}.$$

Sous l’hypothèse \mathcal{H}_1 , le numérateur prendra une valeur importante alors que le dénominateur sera faible: au final la statistique $h(\mathbf{Y})$ sera grande. À l’inverse, sous l’hypothèse \mathcal{H}_0 , la statistique $h(\mathbf{Y})$ sera proche de 0. Le test du rapport de vraisemblance s’écrit

$$\Phi := \mathbf{1}_{\{h(\mathbf{Y}) > k_\alpha\}}$$

où k_α est une constante à fixer selon le seuil α . Le lemme suivant montre que ce test est optimal lorsque les hypothèses sont des hypothèses simples: $\mathcal{H}_0 : \theta = \theta_0$ contre $\mathcal{H}_1 : \theta = \theta_1$ où θ_0 et θ_1 sont 2 éléments de Θ .

Lemme de Neyman-Pearson. *Pour $\alpha \in]0; 1[$, s'il existe k_α tel que le test*

$$\Phi = \mathbf{1}_{\{L_{\theta_1}(\mathbf{Y}) > k_\alpha L_{\theta_0}(\mathbf{Y})\}}$$

soit de niveau α , alors ce test est le plus puissant des tests de niveau inférieur à α , cad que pour tout test Φ' de niveau inférieur à α , on a $P_{\theta_1}(\Phi' = 1) \leq P_{\theta_1}(\Phi = 1)$.

Remarque. Le test de vraisemblance ne s'exprime pas forcément en fonction de la vraisemblance. Il est possible que le test puisse s'écrire en fonction d'une statistique de test S plus simple.

Exemple. Soit $\mathbf{Y} = (Y_1, \dots, Y_n)$ un échantillon i.i.d. de loi $\mathcal{N}(\theta, \sigma^2)$ avec θ inconnu et σ^2 connue. On teste $\theta = \theta_0$ contre $\theta = \theta_1$ avec $\theta_0 < \theta_1$. On a $\ln L_\theta(\mathbf{y}) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{k=1}^n (y_k - \theta)^2$. Or, pour toute constante c ,

$$\begin{aligned} L_{\theta_1}(\mathbf{y}) > c L_{\theta_0}(\mathbf{y}) &\Leftrightarrow \ln L_{\theta_1}(\mathbf{y}) > c_1 + \ln L_{\theta_0}(\mathbf{y}) \\ &\Leftrightarrow \sum_{k=1}^n (y_k - \theta_1)^2 > c_2 + \sum_{k=1}^n (y_k - \theta_0)^2 \\ &\Leftrightarrow \sum_{k=1}^n y_k > c_3 \Leftrightarrow \frac{\bar{y}_n - \theta_0}{\sigma/\sqrt{n}} > c_4 \end{aligned}$$

où c_1, c_2, c_3, c_4 sont des constantes ne dépendant pas de \mathbf{y} (mais pouvant dépendre de $\theta_0, \theta_1, n, \sigma^2$). On voit donc que le test du rapport de vraisemblance se réécrit

$$\Phi = \mathbf{1} \left\{ \frac{\bar{Y}_n - \theta_0}{\sigma/\sqrt{n}} > z_{1-\alpha} \right\}$$

où la constante $c_4 = z_{1-\alpha}$ a été choisie de sorte que Φ soit de niveau α .

Remarque. Le lemme de Neyman-Pearson permet d'avoir un test dans le cas d'hypothèses simples. Dans certains cas, il est possible d'en déduire un test pour des hypothèses composites.

Proposition. *Soit θ_0 fixé et Θ_0 contenant θ_0 . Si le test fourni par le lemme de Neyman-Pearson ne dépend pas de $\theta_1 \in \Theta_1$ et vérifie $\sup_{\theta \in \Theta_0} P_\theta(\Phi = 1) \leq \alpha$, alors c'est aussi un test de $\theta \in \Theta_0$ contre $\theta \in \Theta_1$ et il est uniformément plus puissant parmi les tests de niveau inférieur à α ($UPP(\alpha)$).*

Dire que le test de vraisemblance ne dépend pas de θ_1 signifie que le test de vraisemblance de $\theta = \theta_0$ contre $\theta = \theta_1$ est le même quel que soit $\theta_1 \in \Theta_1$. Dire que Φ est uniformément plus puissant parmi les tests de niveau inférieur à α signifie que pour tout test Φ' de niveau inférieur à α de $\theta \in \Theta_0$ contre $\theta \in \Theta_1$, on a $P_{\theta_1}(\Phi' = 1) \leq P_{\theta_1}(\Phi = 1)$ pour tout $\theta_1 \in \Theta_1$. Remarquons qu'il n'y a pas de raison qu'un tel test existe forcément.

Exemple. Reprenons le cas de l'échantillon i.i.d. $\mathbf{Y} = (Y_1, \dots, Y_n)$ de loi $\mathcal{N}(\theta, \sigma^2)$ avec θ inconnu et σ^2 connue. Soit θ_0 fixé, $\Theta_0 = \{\theta \leq \theta_0\}$ et $\Theta_1 = \{\theta > \theta_0\}$. D'après l'exemple précédent le test de vraisemblance s'écrit

$$\Phi = \mathbf{1} \left\{ \frac{\bar{Y}_n - \theta_0}{\sigma/\sqrt{n}} > z_{1-\alpha} \right\}.$$

On peut remarquer qu'il ne dépend pas de $\theta_1 > \theta_0$ et on a $P_\theta(\Phi = 1) \leq \alpha$ pour tout $\theta \leq \theta_0$. On déduit que Φ est $UPP(\alpha)$.

5 Echantillons Gaussiens

On présente dans cette partie les tests statistiques classiques associés à des échantillons Gaussiens.

5.1 Vecteurs Gaussiens

5.1.1 Définition et premières propriétés

Définition. On appelle vecteur Gaussien de \mathbb{R}^d un vecteur aléatoire $\mathbf{Y} = (Y_1, \dots, Y_d) \in \mathbb{R}^d$ tel que toute combinaison linéaire des composantes suit une loi Gaussienne. On notera \mathbf{m} son vecteur moyenne et Σ sa matrice de covariance.

On a donc $\mathbf{m} = (E[Y_1], \dots, E[Y_d])'$, $\Sigma = E[(\mathbf{Y} - \mathbf{m})(\mathbf{Y} - \mathbf{m})'] = (Cov(Y_i, Y_j))_{1 \leq i, j \leq d}$.

Exemple. Le vecteur $\mathbf{Y} = (Y_1, \dots, Y_d)$ composé de d variables Gaussiennes i.i.d. est un vecteur Gaussien.

Proposition. Si A est une matrice $p \times d$ et b un vecteur dans \mathbb{R}^p , alors $A\mathbf{Y} + b$ est un vecteur Gaussien de \mathbb{R}^p de moyenne $A\mathbf{m} + b$ et de matrice de covariance $A\Sigma A'$.

Proposition. La fonction caractéristique d'un vecteur Gaussien de \mathbb{R}^d s'écrit

$$E[e^{i\langle t, \mathbf{Y} \rangle}] = e^{i\langle t, \mathbf{m} \rangle} e^{-\frac{1}{2}t' \Sigma t}$$

où $t \in \mathbb{R}^d$. En particulier la loi d'un vecteur Gaussien est caractérisée par sa moyenne et sa variance. On notera $\mathcal{N}(\mathbf{m}, \Sigma)$ cette loi.

Exemple. La loi $\mathcal{N}(\mathbf{m}, \Sigma)$ est la loi du vecteur $\mathbf{m} + \mathbf{Y}$ où \mathbf{Y} suit la loi $\mathcal{N}(0, \Sigma)$.

Exemple. Notons I_d la matrice identité en dimension d . La loi $\mathcal{N}(0, I_d)$ est celle de (Y_1, \dots, Y_d) où les $(Y_i)_{1 \leq i \leq d}$ sont des variables Gaussiennes i.i.d. centrées réduites.

Proposition. Lorsque Σ est inversible, la densité du vecteur Gaussien $\mathcal{N}(\mathbf{m}, \Sigma)$ en $\mathbf{x} \in \mathbb{R}^d$ s'écrit

$$\frac{1}{(2\pi)^{d/2}} \frac{1}{\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(\mathbf{x} - \mathbf{m})' \Sigma^{-1} (\mathbf{x} - \mathbf{m})}.$$

Dans le cas où Σ n'est pas inversible, la loi de \mathbf{Y} est portée par un sous-espace vectoriel strict de \mathbb{R}^d .

Proposition. Soient X et Y deux variables aléatoires réelles. Si (X, Y) est un vecteur Gaussien, alors X et Y sont indépendantes si et seulement si $Cov(X, Y) = 0$.

Théorème central limite. Soit $(\mathbf{X}_i)_{i \geq 1}$ une suite i.i.d. de vecteurs aléatoires de \mathbb{R}^d de moyenne nulle et de matrice de covariance finie Σ . Alors $\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{X}_i$ converge en loi vers $\mathcal{N}(0, \Sigma)$.

5.1.2 Loi du khi-deux, loi de Student, loi de Fisher

On définit ici trois lois qui vont revenir fréquemment dans le cadre des échantillons Gaussiens.

Définition. On appelle loi du χ^2 à d degrés de libertés de paramètre de décentrage $\|\mathbf{m}\|^2$, notée $\chi^2(d, \|\mathbf{m}\|^2)$ (ou simplement $\chi^2(d)$ si $\mathbf{m} = 0$) la loi de la variable $\|\mathbf{Y}\|^2$ où \mathbf{Y} est de loi $\mathcal{N}(\mathbf{m}, I_d)$.

Remarque. Dans la définition, $\|\mathbf{x}\| = \sqrt{x_1^2 + \dots + x_d^2}$ représente la norme euclidienne du vecteur $\mathbf{x} = (x_1, \dots, x_d)' \in \mathbb{R}^d$.

Définition. On appelle loi de Student à d degrés de libertés de paramètre de décentrage $\mu \in \mathbb{R}$, notée $\mathcal{T}(d, \mu)$ (et $\mathcal{T}(d)$ dans le cas $\mu = 0$) la loi de la variable aléatoire

$$\frac{Z}{\sqrt{U/d}}$$

où U, Z sont indépendantes, U suit une loi $\chi^2(d)$ et Z suit une loi $\mathcal{N}(\mu, 1)$.

Définition. Soient Z_1 et Z_2 deux variables aléatoires indépendantes de loi respectivement $\chi^2(d_1, \lambda)$ et $\chi^2(d_2)$. On appelle loi de Fisher à d_1 et d_2 degrés de liberté et de paramètre de décentrage λ la loi de la variable aléatoire

$$F = \frac{Z_1/d_1}{Z_2/d_2}.$$

On note $\mathcal{F}(d_1, d_2, \lambda)$ cette loi, et plus simplement $\mathcal{F}(d_1, d_2)$ si $\lambda = 0$ (loi de Fisher à d_1 et d_2 degrés de liberté).

5.1.3 Le théorème de Cochran

Voici un théorème fondamental pour les échantillons Gaussiens.

Théorème de Cochran. Soit \mathbf{Y} de loi $\mathcal{N}(\mathbf{m}, I_d)$. Soit $E_1 \oplus \dots \oplus E_\ell$ une décomposition de \mathbb{R}^d en ℓ espaces vectoriels orthogonaux de dimension respective d_1, \dots, d_ℓ . Alors les projections orthogonales $\Pi_{E_1}\mathbf{Y}, \dots, \Pi_{E_\ell}\mathbf{Y}$ sont indépendantes et la loi de $\|\Pi_{E_k}\mathbf{Y}\|^2$ est $\chi^2(d_k, \|\Pi_{E_k}\mathbf{m}\|^2)$.

Corollaire. Soit $\mathbf{Y} = (Y_1, \dots, Y_n)$ un échantillon de n variables Gaussiennes i.i.d. de loi $\mathcal{N}(\mu, \sigma^2)$. Alors les variables \bar{Y}_n et $\frac{1}{\sigma^2} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$ sont indépendantes, de loi respective $\mathcal{N}(\mu, \sigma^2/n)$ et $\chi^2(n-1)$. Ainsi,

$$\frac{\bar{Y}_n - \mu}{\hat{\sigma}_n/\sqrt{n}} \sim \mathcal{T}(n-1)$$

où $\hat{\sigma}_n^2 := \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$.

5.2 Intervalles de confiance

On déduit des résultats précédents des intervalles de confiance pour la moyenne et la variance d'un échantillon Gaussien. Soit donc (Y_1, \dots, Y_n) des variables aléatoires Gaussiennes $\mathcal{N}(\mu, \sigma^2)$ indépendantes.

Moyenne. On a déjà vu comment construire un intervalle de confiance lorsque σ^2 est connue. Dans le cas où σ^2 est inconnue, un intervalle de confiance (bilatéral) au niveau $1 - \alpha$ de μ est donné par

$$\left[\bar{Y}_n - \frac{\hat{\sigma}_n}{\sqrt{n}} t_{n-1, 1-\frac{\alpha}{2}}, \bar{Y}_n + \frac{\hat{\sigma}_n}{\sqrt{n}} t_{n-1, 1-\frac{\alpha}{2}} \right]$$

où $t_{d,\beta}$ est le quantile β de la loi de Student $\mathcal{T}(d)$.

Variance. Un intervalle de confiance (bilatéral) au niveau $1 - \alpha$ de σ^2 est donné par

$$\left[\frac{(n-1)\hat{\sigma}_n^2}{c_{n-1, 1-\alpha/2}}, \frac{(n-1)\hat{\sigma}_n^2}{c_{n-1, \alpha/2}} \right]$$

où $c_{d,\beta}$ est le quantile β de la loi khi-deux $\chi^2(d)$.

Exercice. Donner les intervalles de confiance unilatéraux à droite.

5.3 Test de comparaison à une référence

Soit une série de variables aléatoires Gaussiennes indépendantes (Y_1, \dots, Y_n) de loi $\mathcal{N}(\mu, \sigma^2)$.

Moyenne. On veut tester (test bilatéral)

$$\mathcal{H}_0 : \mu = \mu_0 \quad \text{contre} \quad \mathcal{H}_1 : \mu \neq \mu_0$$

pour un certain nombre μ_0 . On utilise que sous \mathcal{H}_0 , la variable

$$T = \frac{\bar{Y}_n - \mu_0}{\hat{\sigma}_n / \sqrt{n}} \sim \mathcal{T}(n-1).$$

La variable T est notre statistique de test. Pour obtenir un test de risque α , on rejettera \mathcal{H}_0 si

$$|T| > t_{n-1, 1-\alpha/2}.$$

- Dans le cas où $\mathcal{H}_1 : \mu > \mu_0$, on rejette \mathcal{H}_0 si $T > t_{n-1, 1-\alpha}$.
- Dans le cas où $\mathcal{H}_1 : \mu < \mu_0$, on rejette \mathcal{H}_0 si $T < t_{n-1, \alpha}$.

Pro-OGM et anti-OGM. Cet exemple (inventé) a pour objectif d'illustrer le caractère non équitable des tests d'hypothèses déjà souligné plus haut. La protéine P435 permet de mesurer le taux de mutation cellulaire dans le pancréas. Un échantillon de $n = 20$ souris est nourri avec du maïs OGM H76. Il est communément admis que pour un taux supérieur à 80 unités de P435, le risque de cancer est élevé. On modélise ce taux de P435 chez une souris nourrie au maïs H76 par une variable aléatoire notée Y , et on suppose que $Y \sim \mathcal{N}(\mu, \sigma^2)$ où μ et σ^2 sont inconnus. Sur l'échantillon des 20 souris, les mesures donnent $\bar{Y} = 79.5$ et $\hat{\sigma}_n^2 = 2.13$. Soit la statistique de Student $T_n := \sqrt{n} \frac{\bar{Y}_n - 80}{\hat{\sigma}_n} = 1.53$.

- Les militants anti-OGM effectuent le test de $\mathcal{H}_0 : \mu > 80$ contre $\mathcal{H}_1 : \mu \leq 80$ au niveau 5%. Il s'agit du test

$$\Phi_{\text{anti}}(X) := \mathbf{1}\{T_n < t_{n-1, 0.05}\}$$

où $t_{n-1, 0.05} \approx -1.73$. Les militants anti-OGM gardent donc l'hypothèse $\mu > 80$.

- Les industriels pro-OGM effectuent le test de $\mathcal{H}_0 : \mu \leq 80$ contre $\mathcal{H}_1 : \mu > 80$ au niveau 5%. Il s'agit du test

$$\Phi_{\text{pro}}(X) := \mathbf{1}\{T_n > t_{n-1,0.95}\}$$

où $t_{n-1,0.95} \approx 1.73$. Les pro-OGM gardent donc l'hypothèse $\mu \leq 80$.

La faible puissance du test de Student pour $n = 20$ explique cette situation paradoxale. On peut calculer par exemple $\pi(\mu = 0.85) = 0.12$ et $\pi(\mu = 0.9) = 0.25$; le test est très conservatif !

Variance. *On veut tester (test bilatéral)*

$$\mathcal{H}_0 : \sigma^2 = \sigma_0^2 \text{ contre } \mathcal{H}_1 : \sigma^2 \neq \sigma_0^2$$

où σ_0^2 est un certain nombre positif. On se rappelle que la variable $\frac{(n-1)\hat{\sigma}_n^2}{\sigma^2} \sim \chi^2(n-1)$. Ainsi sous \mathcal{H}_0 , on aura

$$U = \frac{(n-1)\hat{\sigma}_n^2}{\sigma_0^2} \sim \chi^2(n-1).$$

La variable U est notre statistique de test. Pour obtenir un test de risque α , on rejettera \mathcal{H}_0 si

$$U \notin [c_{n-1,\alpha/2}, c_{n-1,1-\alpha/2}].$$

- Dans le cas où $\mathcal{H}_1 : \sigma^2 > \sigma_0^2$, on rejette \mathcal{H}_0 si $U > c_{n-1}(1-\alpha)$.
- Dans le cas où $\mathcal{H}_1 : \sigma^2 < \sigma_0^2$, on rejette \mathcal{H}_0 si $U < c_{n-1}(\alpha)$.

Remarque. On aurait pu déduire ces tests des intervalles de confiance de la partie précédente.

Vocabulaire. On appelle communément le test sur la moyenne test de Student ou t-test (en anglais).

5.4 Test d'homogénéité de deux échantillons Gaussiens

On considère deux échantillons Gaussiens indépendants $\mathbf{X} = (X_1, \dots, X_n)$ et $\mathbf{Y} = (Y_1, \dots, Y_m)$. On aimerait comparer les caractéristiques des deux échantillons.

5.4.1 Comparaison des variances

On suppose que \mathbf{X} et \mathbf{Y} sont indépendants. Les variables $(X_i)_i$ sont i.i.d. de loi $\mathcal{N}(\mu_X, \sigma_X^2)$ et les variables $(Y_i)_i$ de loi $\mathcal{N}(\mu_Y, \sigma_Y^2)$.

On aimerait tester $\mathcal{H}_0 : \sigma_X = \sigma_Y$ contre $\mathcal{H}_1 : \sigma_X \neq \sigma_Y$.

Sous \mathcal{H}_0 , on remarque que

$$F = \frac{\hat{\sigma}_X^2}{\hat{\sigma}_Y^2} \sim \frac{\chi^2(n-1)/(n-1)}{\chi^2(m-1)/(m-1)} = \text{Fisher}(n-1, m-1).$$

On rejette donc \mathcal{H}_0 si $F < f_{n-1,m-1}(\alpha/2)$ ou $F > f_{n-1,m-1}(1-\alpha/2)$.

- Dans le cas où $\mathcal{H}_1 : \sigma_X > \sigma_Y$, on rejette \mathcal{H}_0 si $F > f_{n-1,m-1}(1-\alpha)$.
- Dans le cas où $\mathcal{H}_1 : \sigma_X < \sigma_Y$, on rejette \mathcal{H}_0 si $F < f_{n-1,m-1}(\alpha)$.

Remarque. Ce test est appelé test de Fisher.

5.5 Comparaison des moyennes

Echantillons indépendants. On suppose que $\mathbf{X} = (X_1, \dots, X_n)$ et $\mathbf{Y} = (Y_1, \dots, Y_m)$ sont indépendants et ont **même variance** $\sigma_X = \sigma_Y = \sigma$ (il existe encore un test dans le cas $\sigma_X \neq \sigma_Y$ appelé test de Welch). Les variables $(X_i)_i$ sont i.i.d. de loi $\mathcal{N}(\mu_X, \sigma^2)$ et les variables $(Y_i)_i$ i.i.d. de loi $\mathcal{N}(\mu_Y, \sigma^2)$.

On teste (test bilatéral)

$$\mathcal{H}_0 : \mu_X = \mu_Y \text{ contre } \mathcal{H}_1 : \mu_X \neq \mu_Y$$

Sous \mathcal{H}_0 ,

$$T = \frac{\bar{X}_n - \bar{Y}_m}{\hat{\sigma}_{intra} \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim \mathcal{T}(n + m - 2)$$

avec $\hat{\sigma}_{intra}^2 = \frac{1}{n+m-2} (\sum_{k=1}^n (X_k - \bar{X}_n)^2 + \sum_{k=1}^m (Y_k - \bar{Y}_m)^2)$.

On rejette \mathcal{H}_0 si $|T| > t_{n+m-2}(1 - \alpha/2)$.

- Dans le cas où $\mathcal{H}_1 : \mu_X > \mu_Y$, on rejette \mathcal{H}_0 si $T > t_{n+m-2}(1 - \alpha)$.
- Dans le cas où $\mathcal{H}_1 : \mu_X < \mu_Y$, on rejette \mathcal{H}_0 si $T < t_{n+m-2}(\alpha)$.

Echantillons appariés. On suppose cette fois-ci que les échantillons \mathbf{X} et \mathbf{Y} ne sont plus indépendants, mais qu'ils proviennent de la même population. On supposera donc que les couples $(X_1, Y_1), \dots, (X_n, Y_n)$ sont des vecteurs Gaussiens indépendants, chaque vecteur (X_i, Y_i) étant de moyenne \mathbf{m}_i et de variance Σ . Le vecteur moyenne $\mathbf{m}_i = (\mu_{X,i}, \mu_{Y,i})$ peut dépendre de i , mais la matrice de covariance Σ est identique pour tous les couples.

On aimerait tester $\mathcal{H}_0 : \mu_{X,i} = \mu_{Y,i}$ pour tout i contre $\mathcal{H}_1 : \text{il existe } i \text{ tel que } \mu_{X,i} \neq \mu_{Y,i}$.

Cela revient à faire un test de comparaison de la moyenne des variables $\Delta_i := X_i - Y_i$ à zéro. On appelle ce test le test de Student apparié. La statistique de test est

$$T = \frac{\bar{\Delta}_n}{\hat{\sigma}_n / \sqrt{n}}$$

avec $\hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{k=1}^n (\Delta_k - \bar{\Delta}_n)^2$. On compare ensuite T aux quantiles de la loi de Student $\mathcal{T}(n-1)$.