CONTRIBUTEURS

MAS-I : Modern Actuarial Statistics I (ACT-2000, ACT-2003, ACT-2005)

aut., cre. Alec James van Rassel

Référence (manuels, YouTube, notes de cours)

src. Tse, Y., Nonlife Actuarial Models, Theory Methods and Evaluation, Cambridge University Press, 2009.

src. Hogg, R.V. ; McKean, J.W. ; and Craig, A.T., Introduction to Mathematical Statistics, 7th Edition, Prentice Hall, 2013.

src. Weishaus, A., CAS Exam MAS-I, Study Manual, 1st Edition, Actuarial Study Materials, 2018.

src. Starmer, J. (2015). StatQuest. Retrieved from https ://statquest.org/.

src. Luong, A., ACT-2000 : Analyse statistique des risques actuariels, Université Laval, Québec (QC).

src. Côté, M.-P., ACT-2000 : Analyse statistique des risques actuariels, Université Laval, Québec (QC).

Contributeurs

pfr. Sharon van Rassel

pfr. Louis-Philippe Vignault

pfr. Philippe Morin

Motivation

Inspiré par la chaîne de vidéos YouTube StatQuest et mon étude pour MAS-I, je crée cette feuille dans le but de simplifier tous les obstacles que j'ai encourus dans mon apprentissage des statistiques et ainsi simplifier la vie des étudiants en actuariat.
L'objectif est d'expliquer les concepts statistiques de façon claire et concise. Je vous prie de me faire part de tous commentaires et de me signaler toute erreur que vous trouvez !

Première partie

Analyse statistique des risques actuariels

Échantillonnage et statistiques

Notation

X Variable aléatoire d'intérêt X avec fonction de densité $f(x;\theta)$;

Θ Ensemble des valeurs possible pour le paramètre θ tel que $\theta \in \Theta$;

› Par exemple, pour une loi normale $\Theta = \{(\mu, \sigma^2) : \sigma^2 > 0, -\infty < \mu < \infty\}$.

$\{X_1, \ldots, X_n\}$ Échantillon de taille n.

› On pose que les observations ont la même distribution que X;

› On pose habituellement l'indépendance entre les observations;

› L'indépendance et la distribution identique rend l'échantillon un **échantillon aléatoire**;

› Lorsque nous avons des observations, ont dénote l'échantillon par $\{x_1, \ldots, x_n\}$ pour représenter des *réalisations* de l'échantillon.

Vraisemblance

Notation

$\mathcal{L}(\theta;\boldsymbol{x})$ Fonction de vraisemblance de θ en fonction des observations $\boldsymbol{x}$;

$$\mathcal{L}(\theta;\boldsymbol{x}) = \prod_{i=1}^{n} f_X(x_i;\theta)$$

où $\boldsymbol{x}^\top = (x_1, \ldots, x_n)$.

$\{X_1, \ldots, X_n\}$ Échantillon de n observations.

› Si les n observations sont indépendantes entres-elles et proviennent de la même distribution paramétrique (identiquement distribué) c'est un **échantillon aléatoire (iid)**;

› On peut le dénoter comme $\{X_n\}$.

Pour bien saisir ce que représente la fonction de vraisemblance $\mathcal{L}(\theta;\boldsymbol{x})$, il faut songer à ce que représente $f(x;\theta)$.
La fonction de vraisemblance $\mathcal{L}(\theta;\boldsymbol{x})$ se résume à une différente façon de voir la fonction de densité $f(x;\theta)$.

› Au lieu de faire varier x pour un (ou des) paramètre θ fixe, on fait varier θ pour un échantillon d'observations $\boldsymbol{x}$ fixé;

Qualité de l'estimateur

La première section traite de «**estimateurs ponctuels** ». C'est-à-dire, on produit une seule valeur comme notre meilleur essai pour déterminer la valeur de la population inconnue. Intrinsèquement, on ne s'attend pas à ce que cette valeur (même si c'en est une bonne) soit la vraie valeur exacte.

Une hypothèse plus utile à des fins d'interprétation est plutôt un **estimateur par intervalle** ; au lieu d'une seule valeur, il retourne un intervalle de valeurs plausibles qui peuvent toutes être la vraie valeur. Le type principal d'*estimateur par intervalle* est *l'intervalle de confiance* traité dans la deuxième sous-section.

En bref :

Estimateur ponctuel L'estimateur $\hat{\theta}_n$ assigne une valeur précise à θ selon l'échantillon.

Estimateur par intervalle Un *intervalle aléatoire*, construit avec l'échantillon aléatoire, ayant une certaine probabilité de contenir la vraie valeur θ.

Estimation ponctuelle

Biais

Notation

θ Paramètre inconnu à estimer ;

Θ Ensemble des valeurs possibles pour θ ;

› Dans le cas multivarié, on a un vecteur $\boldsymbol{\theta}$ et on défini un ensemble des valeurs possibles $\boldsymbol{\Theta}$;

› Par exemple, une loi Gamma a $\boldsymbol{\theta} = \{\alpha, \beta\}$ et, puisque ces paramètres sont strictement positif, $\boldsymbol{\Theta} = \{\mathbb{R}^+, \mathbb{R}^+\}$.

$\hat{\theta}_n$ Estimateur de θ basé sur n observations ;

› Souvent, on écrit $\hat{\theta}$ pour simplifier la notation.

$\mathrm{B}(\hat{\theta}_n)$ Biais de l'estimateur $\hat{\theta}_n$.

Lorsque nous avons un estimateur $\hat{\theta}_n$ pour un paramètre inconnu θ on espère que, **en moyenne**, ses erreurs de prévision seront nulles. Le **biais** $\mathrm{B}(\hat{\theta}_n)$ d'un estimateur quantifie les erreurs de l'estimateur dans ses prévisions de la vraie valeur du paramètre θ.

Biais d'un estimateur

$$\mathrm{B}(\hat{\theta}_n) = \mathrm{E}[\hat{\theta}_n|\theta] - \theta$$

où :

$\mathrm{E}[\hat{\theta}_n|\theta]$ l'espérance de l'estimateur $\hat{\theta}_n$ sachant que la vraie valeur du paramètre est θ.

Estimateur sans biais lorsque le biais d'un estimateur est nul :

$$\mathrm{B}(\hat{\theta}_n) = 0$$

Estimateur asymptotiquement sans biais lorsque le biais d'un estimateur tends vers 0 alors que le nombre d'observations sur lequel il est basé tends vers l'infini :

$$\lim_{n\to\infty} \mathrm{B}(\hat{\theta}_n) = 0$$

Bien que le biais quantifie les erreurs de prévisions de l'estimateur $\hat{\theta}_n$, il n'indique pas la variabilité de ses prévisions. Imagine une personne ayant ses pieds dans de l'eau bouillante et sa tête dans un congélateur. **En moyenne**, sa température corporelle est tiède. *En réalité*, sa température corporelle est à la fois extrêmement élevée et faible.

Les prévisions des estimateurs non biaisés seront toujours proches de la vraie valeur θ. Cependant, être bon ***en moyenne*** n'est pas suffisant et on souhaite évaluer la variabilité des prévisions d'un estimateur $\hat{\theta}_n$ avec sa variance $\mathrm{Var}(\hat{\theta}_n)$.

Borne Cramér-Rao

Notation

$S(\theta)$ Fonction de Score, $S(\theta) = \frac{\partial \ln f(\theta;x)}{\partial \theta}$;

$\boldsymbol{I}_n(\theta)$ Matrice d'information de Fisher d'un échantillon aléatoire $\{X_n\}$;

› La matrice d'information Fisher pour une seule observation sera donc dénotée $\boldsymbol{I}(\theta)$;

› On obtient une "matrice" lorsque nous estimons plusieurs paramètres et donc $\boldsymbol{\theta}$ n'est pas juste un scalaire θ.

$\hat{\theta}^{EMV}$ Estimateur du maximum de vraisemblance de θ.

Lorsque l'on analyse la variance d'un estimateur sans biais, on débute par définir la **borne inférieure de Cramér-Rao** de sa variance $\mathrm{Var}(\hat{\theta}_n)$. Cette borne utilise **la matrice d'information de Fisher** $\boldsymbol{I}_n(\theta)$:

Borne inférieure Cramér-Rao

Sous certaines conditions de régularité,

$$\text{Var}(\hat{\theta}_n) \geq \frac{1}{\boldsymbol{I}_n(\theta)}$$

où

$$\boldsymbol{I}(\theta) = \text{E}\left[\left(\frac{\partial}{\partial\theta}\ln f(\theta;x)\right)^2\right] \overset{\text{iid}}{=} \text{E}\left[-\frac{\partial^2 \ln f(\theta;x)}{\partial\theta^2}\right]$$

$$\boldsymbol{I}_n(\theta) = \text{E}\left[\left(\frac{\partial}{\partial\theta}\ln \mathcal{L}(\theta;\boldsymbol{x})\right)^2\right] \overset{\text{iid}}{=} \text{E}\left[-\frac{\partial^2 \ln \mathcal{L}(\theta;\boldsymbol{x})}{\partial\theta^2}\right]$$

Note Dans le cas d'un échantillon aléatoire (alias, les données sont iid) on obtient la deuxième équation et $\boldsymbol{I}_n(\theta) = n\boldsymbol{I}(\theta)$.

Détails sur la borne Cramér-Rao

La borne de Cramér-Rao est un concept qui échappe souvent aux étudiants. Sur la base de ce vidéo et de ce vidéo, je vais tenter d'expliquer l'intuition sous-jacente au concept. Ce concept va réapparaître plus tard dans le bac et donc, s'il n'est pas clair d'ici la fin de la section, je vous conseille d'aller visionner les vidéos.

Premièrement, on définit l'utilité des deux premières dérivées :

$\frac{\partial}{\partial\theta}\mathcal{L}(\theta)$: Représente le « *rate of change* » de la fonction ;

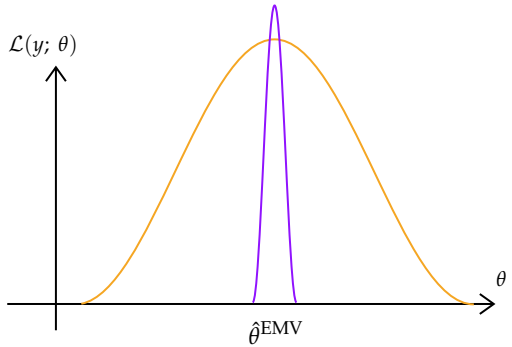
$\frac{\partial^2}{\partial\theta^2}\mathcal{L}(\theta)$: Représente la concavité de la fonction ; on peut y penser comme sa forme.

L'estimateur du maximum de vraisemblance (EMV) $\hat{\theta}^{\text{EMV}}$ du paramètre θ d'une distribution maximise la fonction de vraisemblance en fonction d'un échantillon aléatoire. En posant la première dérivée de la fonction de vraisemblance comme étant égale à 0, on trouve le "point" auquel l'EMV est égale à θ—$\theta^{\hat{\text{EMV}}} = \theta$.

Note : L'EMV devient un "point" lorsqu'on le calcule pour un échantillon aléatoire d'observations.

La fonction de vraisemblance **est concave** et, puisque sa première dérivée est nulle à $\hat{\theta}_n^{\text{EMV}}$, elle va augmenter avant puis diminuer après. La première dérivée permet donc de trouver une fonction qui est maximisée à $\hat{\theta}_n^{\text{EMV}}$. Cependant, ceci ne permet pas d'identifier une fonction unique—plusieurs fonctions peuvent être maximisées au même **point** tout en ayant des formes différentes.

Par exemple, on trace ci-dessous la fonction de vraisemblance et une autre fonction maximisée à $\hat{\theta}_n^{\text{EMV}}$:



On peut donc voir que la forme de la fonction de vraisemblance est plus comprimée, alias que la concavité est plus forte, que l'autre fonction qui se maximise au même point. C'est-à-dire. la fonction de vraisemblance correspond à la fonction avec *la plus forte concavité* dont le maximum est à $\hat{\theta}_n^{\text{EMV}}$.

On peut observer que plus la concavité augmente, plus la variabilité de la fonction de vraisemblance se rapetisse. En effet, une faible concavité implique que la fonction de vraisemblance a un grand étendue de valeurs possibles et moins de points près de $\hat{\theta}^{\text{EMV}}$. En bref, la deuxième dérivée assure que, parmi les fonctions se maximisant à $\hat{\theta}_n^{\text{EMV}}$, la fonction de vraisemblance est celle dont la variabilité des prévisions est minimisée.

L'information de Fisher permet de quantifier cette fonction de la deuxième dérivée. Puis, la borne de Cramér-Rao se défini comme son réciproque $1/\boldsymbol{I}(\theta)$. L'intuition est que plus la concavité est faible, plus l'étendue est grand. Prendre le réciproque de l'information de Fisher permet donc de quantifier l'agrandissement de l'étendu.

Lorsque l'information de Fisher tend vers l'infini (alias la force de la concavité croît infiniment), on dit que la distribution de l'estimateur est "asymptotiquement normale" tel que $\hat{\theta}^{\text{EMV}} \overset{a.s.}{\to} \mathcal{N}\left(\mu = \theta, \sigma^2 = \frac{1}{\boldsymbol{I}(\theta)}\right)$ où a.s. veut dire asymptotiquement.

Efficacité

Notation

$\text{eff}(\hat{\theta}_n)$ Efficacité d'un estimateur $\hat{\theta}_n$;
$\text{eff}(\hat{\theta}_n, \tilde{\theta}_n)$ Efficacité de l'estimateur $\hat{\theta}_n$ relatif à l'estimateur $\tilde{\theta}_n$.

Avec le concept de l'information de Fisher, on défini **l'efficacité d'un estimateur** comme le ratio de la borne Cramér-Rao sur la variance de l'estimateur :

Efficacité d'un estimateur

$$\text{eff}(\hat{\theta}_n) = \frac{\text{Var}(\hat{\theta}_n)^{\text{Rao}}}{\text{Var}(\hat{\theta})} = \frac{1}{\boldsymbol{I}(\theta)\text{Var}(\hat{\theta})}$$

Estimateur « *efficient* » Lorsque la variance de l'estimateur $\text{Var}(\hat{\theta}_n)$ est égale à la borne de Cramér-Rao.

$$\text{eff}(\hat{\theta}_n) = 1$$

› Étant égale à la borne, il *doit* être l'estimateur avec la plus petite de tous les estimateurs sans biais.
On dit qu'il est le « ***Minimum Variance Unbiased Estimator (MVUE)*** ».

De plus, on peut généraliser cette formulation pour obtenir l'efficacité relative d'un estimateur à un autre :

Efficacité relative

$$\text{eff}(\hat{\theta}_n, \tilde{\theta}_n) = \frac{\text{Var}(\hat{\theta}_n)}{\text{Var}(\tilde{\theta}_n)}$$

où les estimateurs $\hat{\theta}_n$ et $\tilde{\theta}_n$ sont sans biais.

Lorsque :
$\text{eff}(\hat{\theta}_n, \tilde{\theta}_n) < 1$: L'estimateur $\hat{\theta}_n$ est plus efficace que l'estimateur $\tilde{\theta}_n$, et vice-versa si $\text{eff}(\hat{\theta}_n, \tilde{\theta}_n) > 1$.

Convergence

Nous pouvons également évaluer si un estimateur converge avec des très grands échantillons ; ceci évalue si un estimateur est cohérent. Un estimateur $\hat{\theta}_n$ est dit d'être « ***consistent*** » si la probabilité que sa prévision $\hat{\theta}$ du paramètre θ diffère de la vraie valeur par une erreur, près de 0, ϵ tend vers 0 alors que la taille de l'échantillon n tend vers l'infini :

Convergence (consistency) d'un estimateur

$$\lim_{n\to\infty} \Pr(\left|\hat{\theta}_n - \theta\right| > \epsilon) = 0, \quad \epsilon > 0$$

Ce critère pour qu'un estimateur $\hat{\theta}_n$ soit « *consistent* » peut être satisfait lorsque :

1. l'estimateur est **asymptotiquement sans biais** ;
$$\lim_{n\to\infty} \text{B}(\hat{\theta}_n) = 0$$
2. la **variance de l'estimateur tend vers 0**.
$$\lim_{n\to\infty} \text{Var}(\hat{\theta}_n) = 0$$

D'ailleurs, nous avons déjà raisonné ceci avec la borne inférieure Cramér-Rao. Cependant, l'inverse n'est pas vrai—qu'un estimateur soit « *consistent* » n'implique pas que sa variance ni que son biais tendent vers 0.

Malgré la nature plaisante de la convergence d'un estimateur, beaucoup d'estimateurs ont cette propriété. Nous voulons alors une mesure qui n'indique pas seulement qu'un estimateur arrive près de la bonne valeur souvent *(alias, une très petite variance)*, mais qu'il est mieux que d'autres estimateurs. De plus, dût à la sélection arbitraire de l'erreur ϵ pour la *consistency* d'un estimateur, il est possible de la choisir malicieusement afin de faire parler les données comme on le souhaite.

Détails sur la convergence

On reprend les résultats de la section précédente en expliquant plus en détails la mathématique sous-jacente.

Convergence en probabilité

Notation

$\{Y_n\}$ Séquence de variables aléatoires ;
Y Variable aléatoire comprise dans $\{Y_n\}$.

On dit que Y_n converge en probabilité à Y si $\forall \varepsilon > 0$,

$$\lim_{n\to\infty} \Pr\left[|Y_n - Y| \geq \varepsilon\right] = 0$$

ou de façon équivalente,

$$\lim_{n\to\infty} \Pr\left[|Y_n - Y| < \varepsilon\right] = 1$$

On dénote la convergence en probabilité par : $Y_n \xrightarrow{P} Y$.

Note : La convergence en probabilité est d'ailleurs le théorème sous-jacent à la loi faible des grands nombres vue en prob.

Loi faible des grands nombres

Notation

$\{Y_n\}$ Séquence de variables aléatoires iid avec moyenne μ et variance σ^2 où $\sigma^2 < \infty$;

$\overline{X}_n$ Moyenne empirique.

On pose que $\overline{X}_n \xrightarrow{P} \mu$.

Théorèmes résultant de la convergence en probabilité

Soit $X_n \xrightarrow{P} X$ et $Y_n \xrightarrow{P} Y$. Alors $X_n + Y_n \xrightarrow{P} X + Y$.

Soit $X_n \xrightarrow{P} X$ et une constante a. Alors $aX_n \xrightarrow{P} aX$.

Soit $X_n \xrightarrow{P} a$ et la fonction $g(\cdot)$ continue à a. Alors $g(X_n) \xrightarrow{P} g(a)$.

Soit $X_n \xrightarrow{P} X$ et la fonction continue $g(\cdot)$. Alors $g(X_n) \xrightarrow{P} g(X)$.

Soit $X_n \xrightarrow{P} X$ et $Y_n \xrightarrow{P} Y$. Alors $X_n Y_n \xrightarrow{P} XY$.

« *Consistency* »

Notation

Y Variable aléatoire avec une distribution paramétrique de paramètre θ ;

$\{Y_1, Y_2, \ldots, Y_n\}$ Échantillon de la distribution de Y ;

$\hat{\theta}_n$ Estimateur de θ.

On dit que $\hat{\theta}_n$ est un estimateur « *consistent* » si $\hat{\theta}_n \xrightarrow{P} \theta$.

Erreur quadratique moyenne

Notation

$\mathbf{MSE}_{\hat{\theta}_n}(\theta)$ Erreur quadratique moyenne d'un estimateur $\hat{\theta}_n$

On défini alors l'**Erreur Quadratique Moyenne** (EQM), ou **Mean Squared Error (MSE)**, permettant de comparer les différents estimateurs ayant tous une bonne *consistency* en assurant une cohérence d'interprétation. Cette mesure permet de quantifier l'écart entre un estimateur $\hat{\theta}_n$ et le vrai paramètre θ.

Erreur Quadratique Moyenne (Mean Squared Error)

$$\mathrm{MSE}_{\hat{\theta}}(\theta) = \mathrm{E}[(\hat{\theta}_n - \theta)^2] \Leftrightarrow \mathrm{Var}(\hat{\theta}_n) + \left[\mathrm{B}(\hat{\theta}_n)\right]^2$$

En combinant tous ces critères, le meilleur estimateur est alors l'estimateur **sans biais** ayant la **plus petite variance** possible parmi tous les estimateurs *sans biais*. C'est-à-dire, le **Uniformly Minimum Variance Unbiased Estimator *(UMVUE)***.

Estimation par intervalles

Notation

$\hat{\theta}_L$ **et** $\hat{\theta}_U$ Fonctions de l'échantillon aléatoire $\{X_1, \ldots, X_n\}$ où $\hat{\theta}_L < \hat{\theta}_U$;

$(\hat{\theta}_L, \hat{\theta}_U)$ Intervalle de confiance de $100(1-\alpha)\%$ de θ si $\Pr(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1-\alpha$;

› Avec les réalisations, on a un intervalle de nombres réels $(\hat{\theta}_l, \hat{\theta}_u)$.

$(1-\alpha)$ Niveau de confiance de l'intervalle où $\alpha \in (0,1)$.

Le type principal d'estimateur par intervalle est l'**intervalle de confiance** :

Intervalle de confiance

Nous sommes confiants à un niveau de $100(1-\alpha)\%$ que le paramètre inconnu θ est entre $(\hat{\theta}_L, \hat{\theta}_U)$.
De façon équivalente, nous sommes confiant à un seuil de $\alpha\%$ que θ est entre $(\hat{\theta}_L, \hat{\theta}_U)$.

Donc, $\theta \in (\hat{\theta}_L, \hat{\theta}_U)$ et nous pouvons dire que $\Pr(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) \geq (1-\alpha)$ pour tout θ.

Ce qu'il faut bien saisir avec les intervalles de confiance, c'est que soit θ est contenu dans l'intervalle $(\hat{\theta}_l, \hat{\theta}_u)$ ou il ne l'est pas.
On peut conceptualiser les intervalles comme une distribution binomiale avec probabilité de succès de $(1-\alpha)$. Si l'on effectue M essais indépendants, on s'attend à ce que $(1-\alpha)M$ intervalles de confiance contiennent θ. Donc on se sent confiant à $(1-\alpha)\%$ que la vraie valeur de θ est contenue dans l'intervalle observé $(\hat{\theta}_l, \hat{\theta}_u)$.

Efficacité des intervalles de confiance Typiquement, la largeur de l'intervalle $(\hat{\theta}_L, \hat{\theta}_U)$ augmente si on augmente le niveau de confiance $(1-\alpha)$. Par exemple, pour être certain à 100% que l'intervalle va contenir la valeur, on a qu'à faire un intervalle $(-\infty, \infty)$.

Donc, un intervalle plus petit nous donne plus d'information si le niveau est adéquat. On dit que pour un même niveau $(1-\alpha)$, l'intervalle avec la plus petite largeur est *plus efficace* que l'autre.

Statistiques

Rappel : Loi du khi-carré

Soit un échantillon aléatoire $(X_1, X_2, \ldots, X_n)$ de variables aléatoires normales de moyenne μ et variance σ^2.

Soit $Q = \sum_{i=1}^{n} (X_i - \mu)^2$.

Alors, $Q/\sigma^2 \sim \chi^2_{(n)}$.

Rappel : Loi de Student

Soit les variables aléatoires indépendantes :

› $Z \sim \mathcal{N}(0,1)$.

› $W \sim \chi^2_{(n)}$.

Alors, $T = \frac{Z}{\sqrt{W/n}} \sim t_{(n)}$.

La loi de Student tend vers la normale lorsque n est très grand.

Rappel : Loi de Fisher-Snedecor (F)

Soit les variables aléatoires indépendantes :

› $W_1 \sim \chi^2_{(\nu_1)}$.

› $W_2 \sim \chi^2_{(\nu_2)}$.

Alors, $F = \frac{W_1/\nu_1}{W_2/\nu_2} \sim \mathcal{F}_{(\nu_1,\nu_2)}$.

On peut relier la loi de Student et la loi F : $T^2 = \frac{Z^2}{W/n} \sim \mathcal{F}_{(1,n)}$ puisque $Z^2 \sim \chi^2_{(1)}$ où $Z \sim \mathcal{N}(0,1)$.

Statistique de test T_n

T_n est une statistique de test basée sur un échantillon aléatoire de n observations.

› C'est donc une **fonction** d'un échantillon aléatoire ;

› Sa distribution est la **distribution d'échantillonnage** qui dépend de :

1. La statistique.
2. La taille de l'échantillon.
3. La distribution sous-jacente des données.

Moyenne échantillonale $\bar{X}$

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

› Estime sans biais la moyenne μ ;

› Si on pose que l'échantillon aléatoire est normalement distribué, $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$;

› On centre et réduit pour trouver que $T_n = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$;

› Si σ^2 est inconnue, on l'estime avec s_n^2 pour obtenir une distribution student— $T_n = \frac{\bar{X}-\mu}{S_n/\sqrt{n}} = \frac{Z}{\sqrt{W/(n-1)}} \sim t_{(n-1)}$ où $W \sim \chi^2_{(n-1)}$.

Variance échantillonale S_n^2

$$S_n^2 = \frac{\sum(X_i-\bar{X})^2}{n-1}.$$

› Estime sans biais la vraie variance σ^2 ;

› S_n^2 n'est pas normalement distribuée, cependant la statistique $T_n = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{(n-1)}$.

Variance empirique $\hat{\sigma}^2$

$$\hat{\sigma}^2 = \frac{\sum(X_i-\bar{X})^2}{n}.$$

› Estime avec biais la vraie variance σ^2.

Statistique F

$$F = \frac{S_n^2/\sigma_1^2}{S_m^2/\sigma_2^2}.$$

› Si on pose que les deux échantillons aléatoires indépendants $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_m)$ sont normalement distribués, $F \sim \mathcal{F}_{(n-1,m-1)}$.

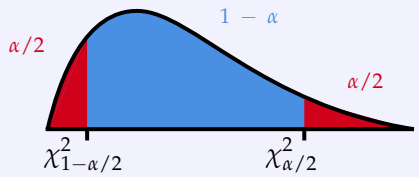
Note sur majuscule vs minuscule On écrit les statistiques avec des majuscule lorsqu'elle sont aléatoires et avec des minuscules lorsque ce sont des réalisations. Par exemple, dans une probabilité on utilise une majuscule puisque la statistique est aléatoire. Pour un seuil α **fixé** d'un intervalle de confiance, le quantile n'est pas aléatoire et jusqu'à ce que l'on calcule l'intervalle avec l'échantillon observé, les statistiques sont également aléatoires.

Intervalles de confiance

Intervalle de confiance sur la variance

Pour l'échantillon aléatoire $\{X_1, X_2, \ldots, X_n\}$ issu d'une distribution normale avec σ^2 inconnue, $\Pr\left(\chi^2_{1-\alpha/2} \leq \frac{(n-1)S_n^2}{\sigma^2} \leq \chi^2_{\alpha/2}\right) = (1-\alpha)$.

Graphiquement :



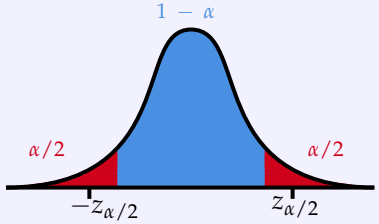
Nous sommes donc confiants à un niveau de $100(1-\alpha)\%$ que :

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

Intervalle de confiance sur la moyenne (σ^2 connue)

Pour l'échantillon aléatoire $\{X_1, X_2, \ldots, X_n\}$ issu d'une distribution normale avec μ inconnu et σ^2 connue, $\Pr\left(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = (1-\alpha)$.

Graphiquement :



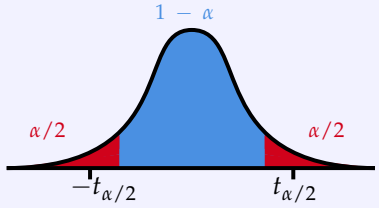
Nous sommes donc confiants à un niveau de $100(1-\alpha)\%$ que :

$$\mu \in \left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Intervalle de confiance sur la moyenne (σ^2 inconnue)

Pour l'échantillon aléatoire $\{X_1, X_2, \ldots, X_n\}$ issu d'une distribution normale avec σ^2 inconnue, $\Pr\left(-t_{\alpha/2,n-1} \leq \frac{\bar{X}-\mu}{S_n/\sqrt{n}} \leq t_{\alpha/2,n-1}\right) = (1-\alpha)$.

Graphiquement :



Nous sommes donc confiants à un niveau de $100(1-\alpha)\%$ que :

$$\mu \in \left[\bar{X} - t_{\alpha/2,n-1}\frac{S_n}{\sqrt{n}}, \bar{X} + t_{\alpha/2,n-1}\frac{S_n}{\sqrt{n}}\right].$$

Intervalle de confiance *approximatif* sur la moyenne

Pour l'échantillon aléatoire $\{X_1, X_2, \ldots, X_n\}$ issu d'une distribution avec moyenne μ et une variance inconnue.

Pour n très grand, nous sommes *approximativement* confiants à un niveau de $100(1-\alpha)\%$ que :

$$\mu \in \left[\bar{X} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right].$$

Intervalle de confiance *approximatif* sur la proportion

Pour l'échantillon aléatoire $\{X_1, X_2, \ldots, X_n\}$ issu d'une distribution Bernoulli de paramètre p.

Pour n très grand, nous sommes *approximativement* confiants à un niveau de $100(1-\alpha)\%$ que :

$$p \in \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right].$$

On défini le « *pooled estimator* » comme la moyenne pondérée des deux variances

échantillonnales $S_p^2 = \frac{(n-1)S_n^2+(m-1)S_m^2}{n+m-2}$.

Intervalle de confiance pour une différence de moyennes

Pour les échantillons aléatoires $\{X_1, X_2, \ldots, X_n\}$ et $\{Y_1, Y_2, \ldots, Y_m\}$ issus de distributions normales de moyennes μ_1 et μ_2 et variance $\sigma_1^2 = \sigma_2^2 = \sigma^2$ inconnues.
Nous sommes confiants à un niveau de $100(1-\alpha)\%$ que :

$$(\mu_1 - \mu_2) \in \left[\bar{x}_n - \bar{y}_m \pm t_{\alpha/2,n+m-2} S_p \sqrt{\frac{1}{n} + \frac{1}{m}}\right].$$

Intervalle de confiance *approximatif* pour une différence de moyennes

Pour les échantillons aléatoires $\{X_1, X_2, \ldots, X_n\}$ et $\{Y_1, Y_2, \ldots, Y_m\}$ issus de distributions normales de moyennes μ_1 et μ_2 et variances σ_1^2 et σ_2^2 inconnues.

Pour n très grand, nous sommes *approximativement* confiants à un niveau de $100(1-\alpha)\%$ que :

$$(\mu_1 - \mu_2) \in \left[\bar{X}_n - \bar{Y}_m \pm z_{\alpha/2} \sqrt{\frac{S_n^2}{n} + \frac{S_m^2}{m}}\right].$$

Intervalle de confiance *approximatif* pour une différence de proportions

Pour les échantillons aléatoires $\{X_1, X_2, \ldots, X_n\}$ et $\{Y_1, Y_2, \ldots, Y_m\}$ issus de distributions Bernoulli de paramètres p_1 et p_2.

Pour n très grand, nous sommes *approximativement* confiants à un niveau de $100(1-\alpha)\%$ que :

$$(p_1 - p_2) \in \left[\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}\right].$$

Tests d'hypothèses

Introduction

Contexte

Les statistiques classiques posent que tout phénomène observable est régi par un *"**processus**" sous-jacent*.
On ne peut jamais savoir exactement ce qu'est ce "processus" ; le mieux que l'on peut faire est d'émettre des ***hypothèses*** vraisemblables sur ce qu'il pourrait être.
Par la suite, on analyse les observations en présumant qu'elles sont régies par le processus hypothétique et détermine la ***vraisemblance des observations***. On accepte le processus hypothétique si la vraisemblance est suffisamment élevée.

Notation

Θ_0 **et** Θ_1 Sous-ensembles disjoints de Θ tel que $\Theta_0 \cup \Theta_1 = \Theta$;

$\mathbf{H}_0$ Hypothèse nulle ;

› Représente généralement le statu quo jusqu'à preuve contraire.

$\mathbf{H}_1$ Hypothèse alternative.

› Représente généralement un changement du statu quo.

Test d'hypothèse

On spécifie une hypothèse nulle et par conséquent une hypothèse alternative :

$$\mathrm{H}_0 : \theta \in \Theta_0 \qquad \text{vs} \qquad \mathrm{H}_1 : \theta \in \Theta_1$$

Puis, on spécifie une expérience et un test pour décider si l'on accepte ou rejette l'hypothèse nulle.

Terminologie

Hypothèse simple Spécifie **entièrement** une distribution de probabilité.

› Par exemple, $\mathcal{H}_0 : q = 0.50$—on connaît la valeur exacte du paramètre q pour une distribution Bernoulli.

Hypothèse composite Spécifie **partiellement** une distribution de probabilité.

› Par exemple, $\mathcal{H}_1 : q \neq 0.50$.—on ne connaît pas la valeur exacte du paramètre q, il pourrait être n'importe quel chiffre sauf 0.50.

Exemple du laissez-passer universitaire (LPU)

Par exemple, on veut savoir si le monde utilisent l'autobus (oui ou non) avant et après l'implantation du LPU.
On pose que la proportion des gens qui utilisent l'autobus est $q = 0.44$.
Il y a deux types de tests qu'on peut faire,

› Tester si l'utilisation est différente est un test "**bilatéral**", car on test si elle a augmentée *ou* diminuée ;
$\mathrm{H}_0 : q = 0.44$ $\mathrm{H}_1 : q \neq 0.44$

› Tester si l'utilisation a augmentée est un test "**unilatéral**", car on test uniquement si elle a augmentée.
$\mathrm{H}_0 : q = 0.44$ $\mathrm{H}_1 : q > 0.44$

Un test unilatéral requiert que l'on sache déjà que la proportion de gens "*doit*" être supérieure. Un test bilatéral est plus conservatif et test les deux possibilités, il devrait donc être celui qu'on applique par défaut.

L'hypothèse :

nulle dans les deux cas est que, en moyenne, l'utilisation de l'autobus n'a pas *changée*.

alternative dans le cas d'un test :

unilatéral est que, en moyenne, l'utilisation a *augmentée*.

bilatéral est que, en moyenne, l'utilisation a *changée*.

Région critique

Notation

$\mathcal{S}$ "Ensemble" de tous les résultats possible pour l'échantillon aléatoire ;

$\mathcal{C}$ **Région critique** du test qui est un sous-ensemble de $\mathcal{S}$.

On rejette H_0 si $\{X_1, \ldots, X_n\} \in \mathcal{C}$.
On conserve H_0 si $\{X_1, \ldots, X_n\} \in \mathcal{C}^c$.

› On peut aussi dire « **région de rejet** ».

Exemple du laissez-passer universitaire (LPU)

On reprend l'exemple du LPU.
L'ensemble des résultats possibles est $\mathcal{S} = [0, 1]$.

› Un test "**bilatéral**" a comme région critique $\mathcal{C} = [0, 0.44) \cup (0.44, 1]$;

› Un test "**unilatéral**" testant l'augmentation a comme région critique $\mathcal{C} = (0.44, 1]$.

On peut donc faire 2 types d'erreurs :

	Vrai état	
Décision	H_0	H_1
Rejeter H_0	Erreur de type I	Bonne décision
Accepter H_0	Bonne décision	Erreur de type II

Certitude du test

Lorsque nous voulons quantifier le degré auquel nous sommes confiants du test, nous utilisons la valeur p.
La valeur p a trois composantes :

1. La probabilité que l'événement se produise aléatoirement.
2. La probabilité qu'un événement tout aussi rare se produise.
3. La probabilité qu'un événement encore plus rare se produise.

Exemple de pile ou face

On souhaite tester si, en obtenant deux piles sur deux lancers, nous avons une pièce de monnaie truquée :

Hypothèse nulle Ma pièce de monnaie n'est pas truquée même si j'ai obtenu deux piles.

Étapes du calcul de la valeur p :

1. On calcule la probabilité d'obtenir 2 piles : $0.5 \times 0.5 = 0.25$.
2. Puis, on calcule la probabilité d'obtenir 2 faces (un événement autant rare) : $0.5 \times 0.5 = 0.25$.
3. Finalement, il n'y a pas d'autres séquences plus rares.

Donc, la valeur p du test est de 0.50.

- Ceci est plutôt élevé ;
- Souvent, on pose que la valeur p du test doit être d'au plus 0.05 ;
- Ce qui veut dire que des événements tout aussi (ou plus) rares doivent arriver moins que 5% du temps pour que l'on considère la pièce de monnaie comme étant truquée ;
- Donc, dans notre cas, on ne peut pas rejeter l'hypothèse nulle que notre pièce de monnaie n'est pas spéciale.

Dans le cas continu, on somme les probabilités d'être plus rare ou d'être moins rare. C'est la même idée que les intervalles de confiance avec la valeur p, ou *seuil de signifiance* α, représenté en rouge.

- Si la valeur p est petite, ceci indique que d'autres distributions pourraient potentiellement mieux s'ajuster aux données puisque l'événement est très rare ;
- Si la valeur p est grande, ceci indique que l'événement est très courant et que la distribution semble être bien ajustée.

Il y a plusieurs termes semblables qui peuvent devenir mélangeants.

Terminologie

p La **valeur** p du test.
- On peut la définir comme la probabilité d'un événement autant (ou plus) rare sous l'hypothèse nulle ;
- On peut la définir comme la **taille** de la région critique $\mathcal{C}$; c'est-à-dire, l'*aire* de la région de rejet de l'hypothèse nulle H_0 alors qu'elle est vraie ;
- On peut la définir comme le **seuil de signifiance** ; c'est-à-dire, la probabilité de rejeter H_0 alors qu'elle est vraie ;
- Elle correspond donc également à la **probabilité d'une erreur de type I**.

α Dénote habituellement le **seuil de signifiance** ou la **taille** du test.
- Même idée qu'avec les intervalles de confiance ;
- On peut parfois aussi utiliser α pour dénoter la valeur de p qui détermine si on rejette ou pas un test ;
- En anglais, « *threshold for significance* ».

Formellement, on défini $\alpha = \max_{\theta \in \Theta_0} \Pr\{(X_1, \dots, X_n) \in \mathcal{C}; \theta\}$.
C'est-à-dire :

- on maximise la probabilité que l'échantillon aléatoire soit contenu dans la région critique (alias rejeter H_0),
- où la distribution est tracée en fonction du paramètre θ de l'hypothèse nulle.

Puissance d'un test

La puissance d'un test

La probabilité de *correctement* rejeter l'hypothèse nulle.

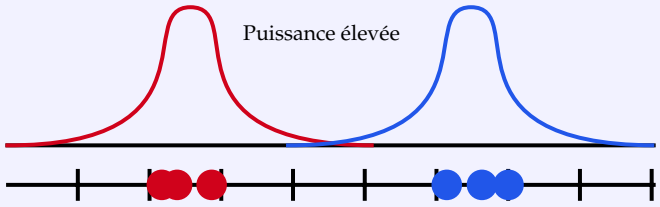
Une analyse de la puissance détermine le nombre d'observations qu'il faut afin d'avoir une probabilité élevée de correctement rejeter l'hypothèse nulle.

Plusieurs facteurs influencent la puissance d'un test. Lorsqu'on teste si deux échantillons d'observations proviennent de la même distribution,

La forme de la distribution

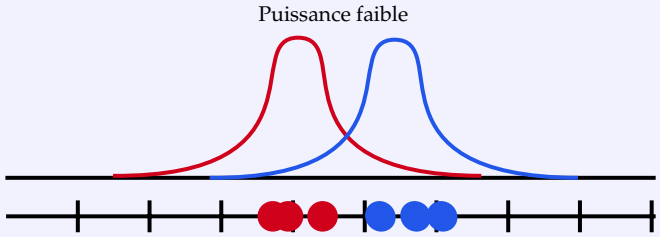
Si les deux distributions sont :

› Très **distinctes**, la puissance sera très **élevée** :



- La probabilité de **correctement** rejeter l'hypothèse nulle (que les deux échantillons proviennent d'une même distribution) est élevée ;
- On peut aussi dire qu'il y a une forte probabilité de **correctement** obtenir une faible valeur p.

› Se **chevauchent**, la puissance sera **faible** :

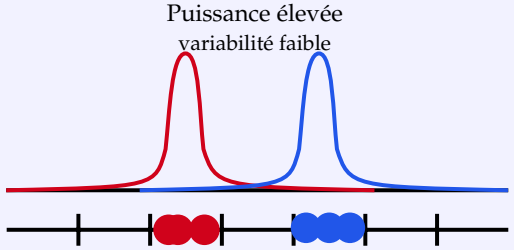


- La probabilité de **incorrectement** rejeter l'hypothèse nulle (que les deux échantillons proviennent d'une même distribution) est élevée ;
- On peut aussi dire qu'il y a une forte probabilité de **incorrectement** obtenir une faible valeur p ;
- Cependant, la puissance peut être augmentée avec plus d'observations.

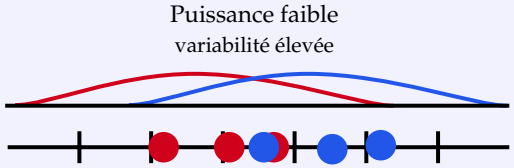
La variabilité des données

Si la variabilité de la distribution est

› **Faible**, alors la variabilité de l'échantillon sera probablement faible aussi menant à une puissance très **élevée** :



› **Élevée**, alors la variabilité de l'échantillon sera probablement élevée aussi menant à une puissance **faible** :



Il existe plusieurs mesures qui permettent de considérer la variabilité des données ainsi que la forme de la distribution. Entres autre, il y a le « ***effect size (d)*** » où $d = \frac{\bar{x}-\bar{y}}{s_p^2}$.

Le taille de l'échantillon de données

Un grand échantillon de données peut compenser pour des distributions qui se chevauchent ou une variabilité élevée. Ça permet d'augmenter notre *confiance* qu'il y a bel et bien une différence entre les échantillons.

En contraste, nous n'avons pas besoin d'un grand échantillon de données pour des distributions très distinctes ou avec une faible variabilité ; nous sommes déjà confiants que les distributions sont différentes.

Le test statistique

Certains tests ont une puissance plus élevée que les autres.
Cela dit, le test t habituel est très puissant.

La fonction de puissance

La fonction de puissance est $\gamma(\theta) = \Pr\{(X_1, \ldots, X_n) \in \mathcal{C}; \theta\}$; c'est-à-dire, la probabilité de rejeter l'hypothèse nulle H_0 si la **vraie** valeur du paramètre est $\theta \in \Theta$.

- C'est une fonction de θ ;
- Idéalement, si l'hypothèse nulle est :

  acceptée on souhaite que $\gamma(\theta) = 0$ puisque $\theta \in \Theta_0$.
  - On dénote $\gamma(\theta_0) = \Pr\{(X_1, \ldots, X_n) \in \mathcal{C}; \theta \in \Theta_0\} = 0$.

  rejetée on souhaite que $\gamma(\theta) = 1$ puisque $\theta \in \Theta_1$.
  - On dénote $\gamma(\theta_1) = \Pr\{(X_1, \ldots, X_n) \in \mathcal{C}; \theta \in \Theta_1\} = 1$.

Si, par exemple, on rejette l'hypothèse nulle on pourrait tracer la fonction de puissance pour toutes les valeurs possibles de l'ensemble Θ_1.

Tests optimaux

Notation

δ (Procédure de) test ;

$\alpha(\delta)$ Probabilité d'une erreur de type I pour un test δ ;

- $\alpha(\delta) = \Pr\{(X_1, \ldots, X_n) \in \mathcal{C}; \theta \in \Theta_0\} = \gamma(\theta_0)$.

$\beta(\delta)$ Probabilité d'une erreur de type II pour un test δ ;

- $\beta(\delta) = \Pr\left\{(X_1, \ldots, X_n) \in \mathcal{C}^{\complement}; \theta \in \Theta_1\right\} = 1 - \gamma(\theta_1)$.

Pour mettre en contexte cette notation, revoici le tableau des types d'erreur pour un test δ :

	Vrai état	
Décision	$H_0 \Rightarrow \theta \in \Theta_0$	$H_1 \Rightarrow \theta \in \Theta_1$
Rejeter H_0 $(X_1, \ldots, X_n) \in \mathcal{C}$	$\alpha(\delta)$	$1 - \beta(\delta)$
Accepter H_0 $(X_1, \ldots, X_n) \in \mathcal{C}^{\complement}$	$1 - \alpha(\delta)$	$\beta(\delta)$

- *En théorie*, on minimise la probabilité d'une erreur de type I **et** de II ;
- *En réalité*, il y a un compromis et on ne pourra pas avoir des très petites probabilités pour les deux ;
- Le contexte va déterminer ce qu'on souhaite minimiser le plus ;
  - Par exemple, soit l'hypothèse nulle que quelqu'un n'a pas le cancer ;
  - Il est plus grave de dire à quelqu'un qu'il n'a pas le cancer alors qu'il l'a (erreur de type II) que de dire qu'il a le cancer alors qu'il ne l'a pas (erreur de type I) ;
  - Dans ce contexte, on souhaiterait minimiser l'erreur de type II $\beta(\delta)$ plus que celle de type I $\alpha(\delta)$.

Puisqu'il est impossible de trouver un test δ pour lequel les probabilités d'erreurs de type I et II sont très petites, on :

1. Fixe l'erreur de type I à un seuil, alias une taille de région critique, k.
2. Trouve parmi tout les sous-ensembles de taille k celui qui minimise l'erreur de type II.

Tests optimaux

Soit un test δ^* avec les hypothèses simples :

$H_0 : \theta = \theta_0$

$H_1 : \theta = \theta_1$

› Par exemple, on pourrait avoir une distribution Bernoulli et poser H_0 : $p = 0.4$ v.s. $H_1 : p = 0.6$.

La procédure pour trouver la région critique $\mathcal{C}$ optimale de taille α du test δ^* est la suivante :

1. On trouve une région critique (alias, un sous-ensemble de $\mathcal{S}$) $\mathcal{C}$ tel que la probabilité $\alpha(\delta^*)$ d'une erreur de type I est de α.
   - › C'est-à-dire, $\alpha(\delta^*) = Pr\{(X_1, \ldots, X_n) \in \mathcal{C}; \theta = \theta_0\} = \alpha$.
   - › Cependant, ce critère n'identifie par un sous-ensemble unique ;
   - › Il y a une multitude de sous-ensemble $\mathcal{A}$ de $\mathcal{S}$ dont la probabilité que l'échantillon aléatoire y soit contenu (sous l'hypothèse nulle) est aussi α ;
   - › C'est-à-dire, $Pr\{(X_1, \ldots, X_n) \in \mathcal{A}; \theta = \theta_0\} = \alpha$.
2. On pose que la probabilité que l'échantillon aléatoire soit dans la région critique $\mathcal{C}$ (sous l'hypothèse alternative) est supérieure à la probabilité que l'échantillon aléatoire soit contenu dans tout autre sous-ensemble $\mathcal{A}$.
   - › C'est-à-dire, $Pr\{(X_1, \ldots, X_n) \in \mathcal{C}; \theta = \theta_1\} \geq Pr\{(X_1, \ldots, X_n) \in \mathcal{A}; \theta = \theta_1\}$.

Avec ces deux critères, on trouve ***la*** région critique $\mathcal{C}$ de taille α **optimale** pour tester les hypothèses simples.

En bref, on pose fixe à un seuil α la fonction de puissance posant que le vrai paramètre $\theta = \theta_0$ puis on trouve la région critique qui maximise la puissance posant que le vrai paramètre $\theta = \theta_1$.

Exemple avec une distribution Binomiale

Soit :

› La variable aléatoire $X \sim Binom(n = 3, p = \theta)$.

– Alors, $\mathcal{S} = \{x : x = 0, 1, 2, 3\}$.

› Les hypothèses :

$H_0 : \theta = 0.50$

$H_1 : \theta = 0.75$

› Le seuil de signifiance $\alpha = 0.125$.

› Les sous-ensembles $\mathcal{A}_1 = \{x : x = 0\}$ et $\mathcal{A}_2 = \{x : x = 3\}$ de $\mathcal{S}$.

Alors, $\Pr(X \in \mathcal{A}_1; \theta = 0.50) = \Pr(X \in \mathcal{A}_2; \theta = 0.50) = 0.125$ et il n'y a pas d'autres sous-ensembles de $\mathcal{S}$ avec la même taille de 0.125.
Il s'ensuit que soit $\mathcal{A}_1$ ou $\mathcal{A}_2$ est la région critique $\mathcal{C}$ optimale de taille α pour tester H_0 contre H_1.

On trouve que $\Pr(X \in \mathcal{A}_1; \theta = 0.75) = 0.015625$ alors que $\Pr(X \in \mathcal{A}_2; \theta = 0.75) = 0.421875$.

› Dans le premier cas :

$$\underbrace{\Pr(X \in \mathcal{A}_1; \theta = 0.75)}_{\text{rejeter } H_0 \text{ alors que } H_0 \text{ est faux } (\theta = 0.75)} = 0.015625 \quad < \quad \underbrace{\Pr(X \in \mathcal{A}_1; \theta = 0.50)}_{\text{rejeter } H_0 \text{ alors que } H_0 \text{ est vraie } (\theta = 0.50)} = 0.125$$

› Dans le deuxième cas :

$$\underbrace{\Pr(X \in \mathcal{A}_2; \theta = 0.75)}_{\text{rejeter } H_0 \text{ alors que } H_0 \text{ est faux } (\theta = 0.75)} = 0.421875 \quad > \quad \underbrace{\Pr(X \in \mathcal{A}_2; \theta = 0.50)}_{\text{rejeter } H_0 \text{ alors que } H_0 \text{ est vraie } (\theta = 0.50)} = 0.125$$

› Le premier sous-ensemble $\mathcal{A}_1$ n'est pas désirable car on serait plus probable de incorrectement rejeter H_0 lorsqu'elle est vraie (erreur de type I) que de correctement la rejeter lorsqu'elle est fausse !

› Alors, on choisit $\mathcal{C} = \mathcal{A}_2 = \{x : x = 3\}$.

D'ailleurs, la région est choisie en incluant dans $\mathcal{C}$ les points x pour lesquels $f(x; \theta = 0.50)$ est petite par rapport à $f(x; \theta = 0.75)$.

› On peut d'ailleurs observer que le ratio $\frac{f(x;\theta=0.50)}{f(x;\theta=0.75)}$ évalué à $x = 5$ est un minimum.

On peut utiliser ce ratio comme outil pour identifier la région critique $\mathcal{C}$ optimale pour un seuil fixe de α.

Cas d'hypothèses simples

Théorème de Neymann-Pearson

Soit un test δ^* avec les hypothèses simple :

$H_0 : \theta = \theta_0$

$H_1 : \theta = \theta_1$

Soit une constante $k > 0$ et le sous-ensemble $\mathcal{C} \in \mathcal{S}$ tel que :

1. $\frac{\mathcal{L}(\theta_0;\boldsymbol{x})}{\mathcal{L}(\theta_1;\boldsymbol{x})} \leq k$ pour tout $x \in \mathcal{C}$.
2. $\frac{\mathcal{L}(\theta_0;\boldsymbol{x})}{\mathcal{L}(\theta_1;\boldsymbol{x})} \geq k$ pour tout $x \in \mathcal{C}^{\complement}$.
   - En réécrivant les équations comme $\mathcal{L}(\theta_1;\boldsymbol{x}) \leq (\geq)\ k\mathcal{L}(\theta_0;\boldsymbol{x})$ on peut l'interpréter comme qu'il doit être plus vraisemblable que $\theta = \theta_0(\theta_1)$ que $\theta_1(\theta_0)$ lorsque $x \in \mathcal{C}^{\complement}$ et que l'on rejette (accepte) H_0.
3. $\alpha = \Pr\{(X_1, \ldots, X_n) \in \mathcal{C}; \theta_1\} = \alpha(\delta^*)$.

Alors $\mathcal{C}$ est ***la*** région critique **optimale** de taille α.

Test non biaisé

Soit les mêmes hypothèses que dans la définition du théorème de Neymann-Pearson.

Un test δ est non biaisé si sa puissance est toujours d'au moins α ; c'est-à-dire, $\Pr\{(X_1, \ldots, X_n) \in \mathcal{C}; \theta\} \geq \alpha$.

Le meilleur test obtenue par le théorème de Neymann-Pearson est non biaisé.

Exemple avec une distribution Normale

Soit :

- L'échantillon aléatoire $\boldsymbol{X} = (X_1, \ldots, X_n)$ d'une distribution normale $\mathcal{N}(\mu = \theta, \sigma^2 = 0)$.
  - Alors, $\mathcal{S} = \{x : x \in \mathbb{R}\}$.

- Les hypothèses :

  $H_0 : \theta = 0$

  $H_1 : \theta = 1$

On a :

$$\frac{\mathcal{L}(\theta_0;\boldsymbol{x})}{\mathcal{L}(\theta_1;\boldsymbol{x})} = \frac{\exp\left\{-\sum_{i=1}^n x_i^2/2\right\}\frac{1}{(\sqrt{2\pi})^n}}{\exp\left\{-\sum_{i=1}^n (x_i-1)^2/2\right\}\frac{1}{(\sqrt{2\pi})^n}} = \exp\left\{-\sum_{i=1}^n x_i + \frac{n}{2}\right\}$$

Alors, la région critique $\mathcal{C}$ optimale est composée des points $(x_1, x_2, \ldots, x_n)$ tel que :

$$e^{-\sum_{i=1}^n x_i + \frac{n}{2}} \leq k \quad \Rightarrow \quad -\sum_{i=1}^n x_i + \frac{n}{2} \leq \ln(k) \quad \Rightarrow \quad \sum_{i=1}^n x_i \geq \frac{n}{2} - \ln(k)$$

$$\therefore \frac{\sum_{i=1}^n x_i}{n} \geq \underbrace{\frac{1}{2} - \frac{\ln(k)}{n}}_{c}$$

Alors, la région critique optimale $\mathcal{C} = \left\{(x_1, x_2, \ldots, x_n) : \frac{1}{n}\sum_{i=1}^n x_i \geq c\right\}$ où c est une constante choisie tel que la taille de $\mathcal{C}$ est α.

Par exemple, puisque $\bar{X} \overset{H_0}{\sim} \mathcal{N}(0, 1/n)$ on peut trouver c avec $\Pr\{\bar{X} \geq c; \theta = \theta_0\} = \alpha$.
Puis, on peut trouver la puissance du test quand H_1 est vraie avec $\Pr\{\bar{X} \geq c; \theta = \theta_1\}$.

Note sur les hypothèses Les hypothèses doivent entièrement spécifier la distribution. Si les hypothèses sont sur les paramètres, elles doivent être des hypothèses simples, mais elles peuvent être sur autre chose.
Par exemple, si on teste $H_0 : f_X(x) = g(x)$ v.s. $H_1 : f_X(x) = h(x)$ alors la vraisemblance sera un ratio de deux distributions différentes.

Cas d'hypothèses composées

Cette section n'est pas suffisamment bien expliquée pour que je la considère complète.

Exemple avec une distribution normale

Soit :

- Un échantillon aléatoire $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ tiré d'une distribution normale $\mathcal{N}(0, \theta)$;
- Les hypothèses :
  $H_0 : \theta = 1$
  $H_1 : \theta > 1$

Alors, on trouve le ratio :

$$\frac{\mathcal{L}(\theta_0 = 1; \boldsymbol{x})}{\mathcal{L}(\theta_1; \boldsymbol{x})} = \frac{\frac{1}{(1)^n(\sqrt{2\pi})^n} e^{-\frac{\sum_{i=1}^n x_i^2}{2(1)^2}}}{\frac{1}{\theta_1^n(\sqrt{2\pi})^n} e^{-\frac{\sum_{i=1}^n (x_i)^2}{2\theta_1^2}}} = \theta_1^n e^{-\frac{\sum_{i=1}^n x_i^2}{2}\left(1-\frac{1}{\theta_1^2}\right)}$$

On voit que le ratio décroît alors que $\sum x_i^2$ augmente. Par conséquent, un test uniformément le plus puissance aura une région critique définie par $\sum x_i^2 > k$ avec un k choisi selon le seuil de signifiance.

L'idée est donc de poser un θ_1 fixe pour évaluer la forme du ratio de la vraisemblance. Selon la croissance ou décroissance de la fonction, ainsi que l'hypothèse, on peut établir une région pour laquelle une augmentation du θ_1 maintient la relation.

La région uniformément la plus puissante n'existe pas toujours, mais dans le cas qu'elle existe le théorème de Neymann-Pearson permet de la trouver.

Test du khi carré

Test d'adéquation (« *goodness-of-fit test* »)

Soit n répétitions (indépendantes) d'une expérience aléatoire.
On pose :

- L'espace d'échantillon des expériences $\mathcal{A}$ qui représente l'union de k différents ensembles (disjoints) $\mathcal{A} = \{A_1 \cup A_2 \cup \cdots \cup A_k\}$;
- On pose que pour $i = 1, 2, \ldots, k$, $\Pr(A_i) = p_i$ où $p_k = 1 - p_1 - \ldots - p_{k-1}$ et $O_k = n - O_1 - \ldots - O_{k-1}$;
  - p_i représente donc la *probabilité* que le résultat de l'expérience aléatoire fait partie de l'ensemble A_i ;
  - O_i représente le *nombre d'observations* (la *fréquence*) pour lesquelles le résultat de l'expérience aléatoire fait partie de l'ensemble A_i.
- On pose que la distribution conjointe de $O_1, O_2, \ldots, O_{k-1} \sim MultiNom(n, p_1, \ldots, p_{k-1})$.

Soit le test d'hypothèse avec les nombres spécifiés $p_{1,0}, p_{2,0}, p_{k-1,0}$:
$H_0 : p_1 = p_{1,0}, p_2 = p_{2,0}, \ldots, p_{k-1} = p_{k-1,0}$
où $p_k = p_{k,0} = 1 - p_{1,0} - \ldots - p_{k-1,0}$.

Alors, sous l'hypothèse nulle : $Q = \sum_{i=1}^{k} \frac{(O_i - np_{i,0})^2}{np_{i,0}} \approx \chi^2_{(k-1)}$.

- Il y a seulement $k-1$ degrés de liberté car on estime seulement $k-1$ paramètres.
  - Le nombre n total d'observations est fixe et on déduit n_k par la somme ;
  - Si on avait à

Tableau de contingence

Dans le cas de données à deux dimensions, alias un tableau de contingence, on défini :

E_{ij} L'espérance du nombre d'observations dans la cellule i, j ;

O_{ij} Le nombre observé d'observations dans la cellule i, j.

- On pose qu'il y a c colonnes au tableau pour r rangées (les rangées sont les différents ensembles) ;
- On peut donc tester si la distribution de la fréquence est identique pour les c colonnes avec $Q = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \approx \chi^2_{(r-1)\cdot(c-1)}$;
- Cette formule est beaucoup plus intuitive visuellement que par formule ; pour la comprendre, faites un exemple.

Test du rapport de vraisemblance

Cette section n'est pas suffisamment bien expliquée pour que je la considère complète.

En fonction des observations, calculer :

1. Le maximum de vraisemblance sous l'hypothèse nulle ;
2. Le maximum de vraisemblance sous l'hypothèse alternative

La région critique correspond à la région pour laquelle le ratio des vraisemblances est en dessous d'une constante k.

› Si les deux hypothèses sont simple, ceci équivaut à utiliser le théorème de Neymann-Pearson.

Cependant, il peut s'avérer difficile d'isoler une distribution dans le ratio. Pour des grands échantillons, on peut plutôt utiliser la distribution asymptotique.
Soit une hypothèse nulle qui spécifie k paramètre et une hypothèse alternative qui en spécifie seulement l ($l < k$). Alors, la statistique du rapport de vraisemblance $Q = -2\left(\ln(\theta_0) - \ln(\theta_1)\right) \sim \chi^2_{k-l}$.

Cette statistique est vue dans les modèles linéaires généralisés aussi.

Statistiques exhaustives

Cette section n'est pas suffisamment bien expliquée pour que je la considère complète.

Statistique exhaustive

Soit l'échantillon aléatoire $(X_1, \ldots, X_n)$ d'une distribution avec paramètre θ inconnu.
Alors, la statistique T_n est "**exhaustive pour** θ" si la distribution conditionnelle $(X_1, \ldots, X_n | T_n)$ ne **dépend pas** de θ.

Exemple Bernoulli

Soit l'échantillon aléatoire d'une distribution Bernouilli de paramètre p.
Alors $T_n = \sum_{i=1}^n X_i$ est exhaustive pour p car :

$$\begin{aligned}\Pr(X_1 = x_1, \ldots, X_n = x_n | T_n = x_1 + \cdots + x_n) &= p^{x_1+\cdots+x_n}(1-p)^{n-(x_1+\cdots+x_n)} \\ &= p^t(1-p)^{n-t}\end{aligned}$$

Dépend seulement de l'échantillon par la valeur t de la statistique T_n.

Théorème de factorisation de Fisher-Neymann

Soit l'échantillon aléatoire $(X_1, \ldots, X_n)$ d'une distribution avec paramètre θ inconnu.
Alors, la statistique T_n est "**exhaustive pour** θ" si pour tout $x_i \in \mathbb{R}, i = 1, 2, \ldots, n$,

$$f(x_1;\theta) \times \ldots \times f(x_n;\theta) = g(t;\theta) \times h(x_1, \ldots, x_n)$$

où :

› $g(t;\theta)$ dépend de $(x_1, \ldots, x_n)$ seulement par T_n ;

› $h(x_1, \ldots, x_n)$ ne **dépend pas** de θ.

Pour **plusieurs paramètres**, on généralise avec le vecteur de paramètres inconnus $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$.
Alors, les statistiques $T_n^1, \ldots, T_n^k$ sont **conjointement exhaustives pour** $\boldsymbol{\theta}$ si pour tout $x_i \in \mathbb{R}, i = 1, 2, \ldots, n$,

$$f(x_1;\theta) \times \ldots \times f(x_n;\theta) = g(t_1, \ldots, t_k;\theta) \times h(x_1, \ldots, x_n)$$

où :

- $g(t^1, \ldots, t^k; \theta)$ dépend de $(x_1, \ldots, x_n)$ seulement par $T_n^1, \ldots, T_n^k$;
- $h(x_1, \ldots, x_n)$ ne **dépend pas** de θ.

Le théorème de factorisation permet d'identifier des statistiques exhaustives. Cependant, il peut y avoir plusieurs statistiques exhaustives! Certaines offrent une plus grand réduction des données; par exemple, $\bar{X}_n$ réduit les données plus que $(X_{(1)}, \ldots, X_{(n)})$.
On cherche donc la statistique exhaustive qui offre la **réduction maximale** tout en retenant toute l'information sur le paramètre visé.

Statistique exhaustive minimale

Une statistique exhaustive $T_n = T(X_1, \ldots, X_n)$ est "**minimale**" si pour tout autre statistique exhaustive $U_n = U(X_1, \ldots, X_n)$, il existe une fonction g tel que $T = g\{U(X_1, \ldots, X_n)\}$.

Critère de Lehmann-Scheffé

Soit l'échantillon aléatoire $(X_1, \ldots, X_n)$ d'une distribution avec paramètre θ inconnu.
Alors, la statistique T_n est "**exhaustive *minimale* pour** θ" si pour tout $x_i, y_i \in \mathbb{R}, i = 1, 2, \ldots, n$,

$$\frac{f(x_1;\theta) \times \ldots \times f(x_n;\theta)}{f(y_1;\theta) \times \ldots \times f(y_n;\theta)}$$

ne **dépend pas** de θ ssi $T(x_1, \ldots, x_n) = T(y_1, \ldots, y_n)$

Exemple Bernoulli

Soit l'échantillon aléatoire d'une distribution Bernouilli de paramètre p.

$$\frac{f(x_1;\theta) \times \ldots \times f(x_n;\theta)}{f(y_1;\theta) \times \ldots \times f(y_n;\theta)} = \left(\frac{p}{1-p}\right)^{(x_1+\cdots+x_n)-(y_1+\cdots+y_n)}$$

Le ratio est seulement indépendant de p si $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$ et donc $T_n = \sum_{i=1}^n X_i$ est **exhaustive minimale** pour p.

Théorème de Rao-Blackwell

Soit l'estimateur $\hat{\theta}_n$ sans biais pour θ avec $\text{Var}(\hat{\theta}_n) < \infty$.
Si la statistique T_n est exhaustive pour θ, la statistique $\theta_n^* = \text{E}[\hat{\theta}_n | T_n]$ est un **estimateur sans biais** de θ avec $\text{Var}(\theta_n^*) \leq \text{Var}(\hat{\theta}_n)$.

Famille exponentielle

Une loi de probabilité fait partie de la famille exponentielle linéaire si :

1. Sa fonction de densité *(ou de masse)* de probabilité peut être exprimée comme :

Densité de la famille exponentielle

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y;\phi)\right)$$

où

θ paramètre canonique

ϕ paramètre de dispersion

2. La fonction c ne dépend pas du paramètre θ.
3. Le support de Y ne dépend pas de θ puisqu'il ne peut pas varier.

Statistiques d'ordre

Soit un échantillon aléatoire de taille n. Nous définissons la k^{e} **statistique d'ordre** $X_{(k)}$ comme étant la k^{e} plus petite valeur d'un échantillon.
Les crochets sont utilisés pour différencier la k^{e} statistique d'ordre $X_{(k)}$ de la k^{e} observation X_k.

Nous sommes habituellement intéressés au minimum $X_{(1)}$ et le maximum $X_{(n)}$.

Minimum

$$X_{(1)} = \min(X_1, \dots, X_n)$$
$$f_{X_{(1)}}(x) = n f_X(x) \big(S_X(x)\big)^{n-1}$$
$$S_{X_{(1)}}(x) = \prod_{i=1}^{n} \Pr(X_i > x)$$

Maximum

$$X_{(n)} = \max(X_1, \dots, X_n)$$
$$f_{X_{(n)}}(x) = n f_X(x) \big(F_X(x)\big)^{n-1}$$
$$F_{X_{(n)}}(x) = \prod_{i=1}^{n} \Pr(X_i \leq x)$$

De façon plus générale, on défini :

k^{e} statistique d'ordre

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!1!(n-k)!} \underbrace{[F_X(x)]^{k-1}}_{\text{observations} < k} \overbrace{f_X(x)}^{\text{observation} = k} \underbrace{[S_X(x)]^{n-k}}_{\text{observations} > k}$$

$$F_{X_{(k)}}(x) = \underbrace{\sum_{i=r}^{n} \binom{n}{i} [F_X(x)]^j [1 - F_X(x)]^{n-j}}_{\text{Probabilité qu'au moins } k \text{ des } n \text{ observations } X_k \text{ sont} \leq x}$$

On peut observer que $X_{(k)} \sim Beta(\alpha = k, \beta = n - k + 1)$

Nous pouvons également définir quelques autres statistiques d'intérêt :

$R = X_{(n)} - X_{(1)}$: **L'étendue** (range) est la différence entre le minimum et le maximum d'un échantillon.

- L'utilité de l'étendue est limitée puisqu'elle est très sensible aux données extrêmes.
- Par exemple, supposons qu'on observe des données historiques de température pour le 1er septembre.
  En moyenne, la température est de $16°C$, mais nous avons un cas extrême de $-60°C$ en 1745.
  L'étendue sera de $86°C$ ce qui n'est très représentatif des données.
  Donc, dans ce contexte, la mesure n'est pas d'une très grande utilité.

$M = \frac{X_{(n)} + X_{(1)}}{2}$: **mi-étendue** (Midrange), est la moyenne entre le minimum et le maximum d'un échantillon.

- Pour comprendre ce que représente la mi-étendue, on la compare à la moyenne arithmétique.
- La moyenne arithmétique considère les données observées et calcule leur moyenne.
  Il s'ensuit qu'elle ne considère pas les chiffres qui ne sont pas observés.
- La mi-étendue considère **tous** les chiffres, observés ou non, entre la plus grande et la plus petite valeur d'un échantillon et en prend la moyenne.

Exemple sur les statistiques d'ordre

Soit un échantillon de données météorologiques $\{-30°, -24°, -7°, -23°, +5°\}$ (celsius).
Je suppose que ce sont des températures du 4 février observées lors des dernières années.

- La moyenne arithmétique ($-22.25°C$) m'intéresse, car je peux savoir, en moyenne, ce qu'est la température le 4 février.
- La mi-étendue ($-12.5°C$), tout comme l'étendue ($-35°C$), ne m'intéresse pas puisqu'elle ne prend pas en considération la vraisemblance des différentes températures.

Maintenant, je suppose que ces données sont des températures observées tout au long de l'hiver passé.

- La moyenne arithmétique ne m'intéresse pas puisqu'elle est beaucoup trop biaisée par les températures de cette même journée.
- Cependant, la mi-étendue et l'étendue me donnent maintenant une meilleure idée de la température de l'hiver.

L'important à retenir est que l'utilité des mesures dépend de la situation. Également, ceci est un exemple **très** simpliste et dans tous les cas on ne peut pas tirer de conclusions sur les températures de l'hiver à partir d'une seule journée.

Nous pouvons définir la **médiane** en termes de statistiques d'ordre :

$$\text{Med} = \begin{cases} X_{((n+1)/2)}, & \text{si n est impair} \\ \frac{X_{(n/2)} + X_{(n/2+1)}}{2}, & \text{si n est pair} \end{cases}$$

Finalement, on définit la distribution conjointe du minimum et du maximum $\forall x < y$:

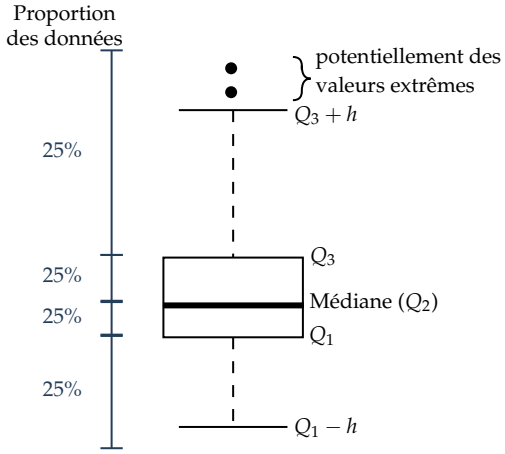
$$f_{X_{(1)}, X_{(n)}}(x, y) = n(n-1)[F_X(y) - F_X(x)]^{n-2} f_X(x) f_X(y)$$

Q-Q plots

Pour n données observées, la k^{e} statistique d'ordre $Y_{(k)}$ est le $\frac{k}{(n+1)}^{\text{e}}$ quantile.
Un sommaire à cinq chiffres consiste de :

1. Le minimum.
2. Le premier quartile Q_1.
3. La médiane (deuxième quartile) Q_2.
4. Le troisième quartile Q_3.
5. Le maximum.

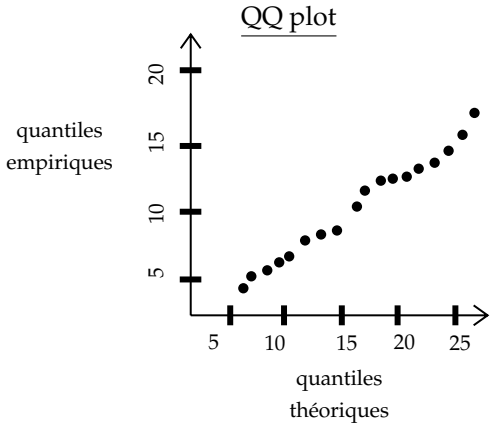
Ces chiffres forment d'ailleurs le diagramme de boîte à moustaches (« *boxplot* ») avec $h = 1.5 \cdot (Q_3 - Q_1)$:



En pratique, on pose souvent que les données suivent une distribution. Un q-q plot permet de comparer les quantiles théoriques aux quantiles empiriques observés.

› Il s'ensuit que l'on connait la distribution, mais pas les paramètres ;
› Dans le cas d'une loi normale, on peut centrer et réduire pour obtenir Z ce qui correspond donc à une q-q plot **normale** ;
› Autrement, on estime les paramètres avec l'échantillon (vérifier).

Par exemple :



Construction d'estimateurs

Précédemment, nous avons décrit les méthodes utilisées pour évaluer la **qualité** de l'estimateur. Cependant, comment obtenons-nous des estimateurs à évaluer ?
Plusieurs méthodes existent pour établir des estimateurs, de plus plusieurs méthodes existent pour estimer des paramètres. La méthode vu dans le cadre du cours de statistique est la **méthode fréquentiste**, le cours de mathématiques IARD 1 (ACT-2005) présente **l'estimation bayésienne**.
Avant de le faire, nous présentons quelques concepts :

Terminologie

$\mu'_k(\hat{\theta})$ k^{e} moment centré à 0, $\mu'_k = \mathrm{E}[X^k]$;

$\pi_g(\theta)$ $100g^{\text{e}}$ pourcentile, $\pi_g(\theta) = F_\theta^{-1}(g)$, $g \in [0,1]$;

$F_e(x)$ Fonction de répartition **e**mpirique ;

$\hat{=}$ Notation pour poser une égalité.

Les deux premiers estimateurs ci-dessous sont les plus faciles à obtenir, mais sont aussi les moins performants puisqu'ils n'utilisent que quelques traits des données au lieu de l'entièreté des données comme la troisième méthode.
Cette distinction devient particulièrement importante dans le cas d'une distribution avec une queue lourde à la droite (Pareto, Weibull, etc.) où il devient plus essentiel de connaître les valeurs extrêmes pour bien estimer le paramètre de forme (α pour une Pareto).
Un autre désavantage est que les deux premières méthodes nécessitent que les données proviennent toutes de la même distribution. Sinon, les moments et quantiles ne seraient pas clairs.
Finalement, sous les deux premières méthodes la décision de quels moments et percentiles à utiliser est arbitraire.

Méthode des moments (MoM)

Estimation de θ par la méthode des moments

Pour ajuster une distribution de p paramètres, on pose égale les p premiers moments empiriques $\hat{\mu}'_k$ au p premiers moments de la distribution μ'_k.
L'estimation de θ est alors toute solution des p équations :

$$\hat{\mu}'_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \hat{=} \mathrm{E}\left[X^k\right] = \mu'_k(\theta), \quad k = 1, 2, \ldots, p$$

La raison pour cet estimateur est que la distribution empirique aura les mêmes p premiers moments centrés à 0 que la distribution paramétrique.

Méthode du «Percentile Matching »

Estimation de θ par la méthode du « *Percentile Matching* »

Pour ajuster une distribution de p paramètres, on pose égale p pourcentiles $\hat{\pi}_g(\hat{\theta})$ de l'échantillon à ceux de la distribution $\pi_g(\theta)$.
L'estimation de θ est alors toute solution des p équations :

$$F_e(\hat{\pi}_{g_k}|\theta) = g_k, \quad k = 1, 2, \ldots, p$$

La raison pour cet estimateur est que le modèle produit aura p percentiles qui vont « *matcher* » les données.
Il peut arriver que les percentiles de distributions ne soient pas uniques, par exemple dans le cas de données discrètes lorsque le quantile recherché peut tomber entre 2 *marches* de la fonction empirique, ou mal-définis. Il est alors utile de définir une méthode d'interpolation des quantiles (bien qu'il n'en existe pas une d'officielle).
Soit le « *smoothed empirical estimate* » d'un pourcentile :

Smoothed empirical estimate

On utilise les statistiques d'ordre de l'échantillon $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ pour l'**interpolation** suivant :

$$\hat{\pi}_g = (1-h)x_{(j)} + hx_{(j+1)}, \quad \text{où}$$
$$j = \lfloor (n+1)g \rfloor \quad \text{et} \quad h = (n+1)g - j$$

Méthode du maximum de vraisemblance

Nous cherchons à maximiser la probabilité d'observer les données. Ceci est fait par la vraisemblance $\mathcal{L}(\theta; \boldsymbol{x})$ ou, puisque le logarithme ne change pas le maximum, la log-vraisemblance $\ell(\theta; x)$ où :

Maximum de vraisemblance

$$\mathcal{L}(\theta; \boldsymbol{x}) = \prod_{i=1}^{n} f(x_i; \theta) \qquad \text{et} \qquad \ell(\theta; \boldsymbol{x}) = \sum_{i=1}^{n} \ln f(x_i; \theta)$$

et l'**estimateur du maximum de vraisemblance** de $\boldsymbol{\theta}$ est celui qui maximise la fonction de vraisemblance.

De façon formelle, on dit que $\hat{\theta}^{\text{EMV}} = \max_{\theta}\{\mathcal{L}(\theta; \boldsymbol{x})\} = \max_{\theta}\{\ln \mathcal{L}(\theta; \boldsymbol{x})\}$.

Raccourcis

Si la fonction de vraisemblance est de la forme :

› $\mathcal{L}(\gamma) = \gamma^{-a} e^{-b/\gamma}$ alors $\hat{\gamma}^{\text{MLE}} = \frac{b}{a}$.

› $\mathcal{L}(\lambda) = \lambda^{a} e^{-\lambda b}$ alors $\hat{\lambda}^{\text{MLE}} = \frac{a}{b}$.

› $\mathcal{L}(\theta) = \theta^{a}(1-\theta)^{b}$ then $\hat{\theta}^{\text{MLE}} = \frac{a}{a+b}$.

Propriétés

Propriété d'invariance

Soit une fonction bijective $g(\cdot)$ et l'estimateur du maximum de vraisemblance (EMV) $\hat{\theta}^{\text{EMV}}$ de θ.

Alors, selon la propriété d'invariance $g(\hat{\theta}^{\text{EMV}})$ est l'EMV de $g(\theta)$.

L'EMV satisfait cette propriété.

Convergence en distribution de l'EMV

Théorème : $\hat{\theta}^{\text{EMV}} \approx \mathcal{N}\left(0, \frac{1}{\boldsymbol{I}_n(\theta)}\right)$.

Sous certaines conditions de régularité, la distribution de $\sqrt{n}\,(\hat{\theta} - \theta)$ converge en distribution vers une distribution normale avec une moyenne nulle et une variance égale à la borne de Cramér-Rao.

$$\sqrt{n}\,(\hat{\theta} - \theta) \xrightarrow{D} \mathcal{N}\left(0, \frac{1}{\boldsymbol{I}_n(\theta)}\right)$$

Ce qui implique :

1. $\hat{\theta}$ est asymptotiquement sans biais.
2. $\hat{\theta}$ est « *consistent* ».
3. $\hat{\theta}$ est approximativement normalement distribué avec moyenne θ et variance $1/\boldsymbol{I}_n(\theta)$ pour des grands échantillons.
4. $\hat{\theta}$ est asymptotiquement efficace puisque sa variance tend vers la borne Cramér-Rao.

Souvent les professeurs ne montrent pas ces conditions puisqu'elles sont compliquées. Alors, ne vous en faites pas si vous ne les comprenez pas complètement.

Conditions de régularité

R0 Les variables X_i sont iid avec densité $f(x_i; \theta)$ pour $i = 1, 2, \ldots$.

R1 Les fonctions de densité ont tous le même support pour tout θ.

› C'est-à-dire que le support de X_i ne dépend pas de θ ;

› C'est une condition restrictive que certains modèles ne respectent pas.

R2 La "vraie valeur" de θ est contenue dans l'ensemble des valeurs possibles Θ.

R3 Les deux premières dérivés de la fonction de densité $f(x; \theta)$ existent.

Note : La condition R3 additionnelle pour la borne de Cramér-Rao.

Cas multivarié

On généralise du cas où θ est un scalaire (un seul paramètre) au cas multivarié avec k paramètres et le vecteur $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)^{\top}$.

Notation

En notation matricielle, on multiple le vecteur $\boldsymbol{\theta}$ par la transposée $\boldsymbol{\theta}^{\top}$ au lieu de mettre θ au carré.

› La matrice d'information Fisher d'*une* observation est donc une matrice

$k \times k$:

$$\boldsymbol{I}(\boldsymbol{\theta}) = \mathrm{E}\left[\frac{\partial \ln f(X;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln f(X;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\right] \overset{iid}{=} \mathrm{E}\left[\frac{\partial^2 \ln f(X;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\boldsymbol{\theta}^\top}\right]$$

› Pour la matrice d'information Fisher d'un échantillon aléatoire de n observation, on utilise la relation $\boldsymbol{I}_n(\boldsymbol{\theta}) = n\boldsymbol{I}(\boldsymbol{\theta})$.

$\boldsymbol{I}_n^{-1}(\boldsymbol{\theta})$ Inverse de la matrice d'information Fisher $\boldsymbol{I}_n(\boldsymbol{\theta})$.

Soit $\tilde{\boldsymbol{\theta}}$ un estimateur sans bais de $\boldsymbol{\theta}$.

Notation

$\mathbf{Var}(\tilde{\boldsymbol{\theta}})$ Matrice de variance de $\tilde{\boldsymbol{\theta}}$.

› Le $(i,j)^{\mathrm{e}}$ élément est donc $\mathrm{Cov}(\tilde{\boldsymbol{\theta}}_i, \tilde{\boldsymbol{\theta}}_j)$.

La version multivariée de l'inégalité Cramér-Rao stipule que $\mathrm{Var}(\tilde{\boldsymbol{\theta}}) - \boldsymbol{I}_n^{-1}(\boldsymbol{\theta})$ est une matrice « *nonnegative definite* ».

› Puisque les éléments de la diagonale doivent être positifs, la borne inférieure de $\mathrm{Var}(\tilde{\boldsymbol{\theta}}_i)$ est le i^{e} élément de la diagonale de $\boldsymbol{I}_n^{-1}(\boldsymbol{\theta})$.

En bref, on trouve que sous certaines conditions de régularité, la distribution de $\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)$ converge en distribution vers une distribution normale multivariée (de k dimensions) avec une moyenne nulle et une variance égale à la borne de Cramér-Rao.

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{D} \mathcal{N}_k\left(0, \boldsymbol{I}_n^{-1}(\boldsymbol{\theta})\right)$$

Deuxième partie

Modèles linéaires en actuariat

Régression linéaire simple

Modèle de régression linéaire simple

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Exemple de compréhension

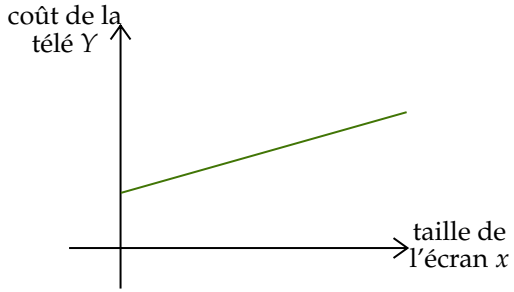
On illustre le concept et la signification des paramètres de régression avec cet exemple illustratif

Objectif On veut deviner le coût d'une télévision (télé) selon la taille de son écran.

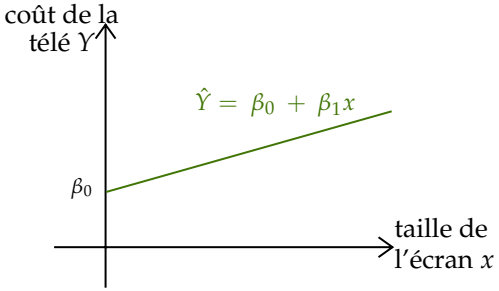
L'idée de la "régression" est de deviner, ou "prédire" du mieux qu'on peut le coût d'une télé en fonction de la taille de son écran.
Deviner le coût *exact* d'une télé *seulement* en fonction de la taille de son écran est impossible. Il y a de nombreuses raisons qui déterminent le prix d'une télé et un bon exercice est de réfléchir à ce qu'elles pourraient être. J'inclus ci-dessous une liste de quelques raisons, ou "facteurs", qui me sont survenus :

- La compagnie qui la produit (Sony vs LG, etc.).
- La résolution (4K vs 360p).
- L'année de fabrication (1990 vs 2020).
- L'endroit de l'achat (Amazon vs BestBuy, Mexique vs Canada, etc.).
- Le temps de l'année (été vs hiver, Boxing Day, etc.).

Maintenant supposons que tu joues à un jeu avec tes amis où qu'ils doivent deviner le coût d'une télé en fonction de sa taille. Ils vont probablement tous te donner des différentes réponses.
Si tu crées un modèle de prévision, il doit être systématique et toujours deviner le même prix pour la même taille d'écran—même si la prévision est erronée.
Alors, supposons que tu changes le jeu un peu et stipules que la personne qui devine le prix le plus éloigné doit prendre une gorgée de sa bière. Les réponses de tes amis vont probablement se ressembler un peu plus, mais il y a un problème qui demeure—tu veux que les prévisions soient proportionnelles à la taille de l'écran.
C'est-à-dire, si ton ami devine qu'une télé de 25" coûte 100$, tu t'attends à ce qu'il devine qu'une télé de 50" coûte 200$.
La raison est qu'une régression **linéaire** *simple* est simplement une ligne droite :

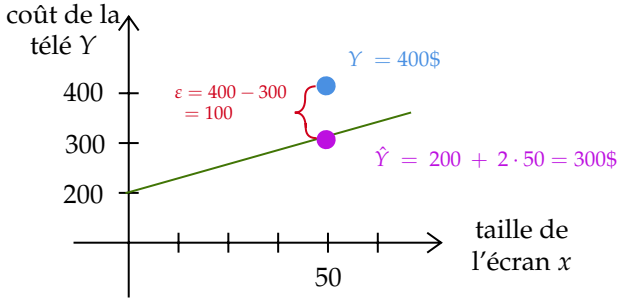


L'intuition est que ton ami se base uniquement sur la taille de l'écran comme information pour deviner le coût. Une régression **linéaire** simple applique un facteur **multiplicatif**. Il ne peut pas se dire que plus grand l'écran est grand, plus le prix va augmenter—ceci serait plutôt une régression avec un paramètre **exponentiel**.
On crée donc un facteur surnommé "paramètre". Dans le cas d'une régression linéaire simple, on a deux paramètres d'intérêts : un "niveau de base" pour le coût β_0 et un "multiplicateur" de la taille d'écran β_1 :



On suppose qu'une télé doit coûter au moins un certain prix. Ce "niveau de base" est l'intercepte sur le graphique ci-dessus surnommé l'ordonnée β_0. De ton gré, tu supposes au moins $\beta_0 = 200\$$ pour cet exemple.
Ensuite, le multiplicateur va multiplier la taille de l'écran pour obtenir un prix. Ce paramètre représente donc la pente β_1. De ton gré, tu suppose une pente de $\beta_1 = 2\$$ pour cet exemple.
Le coût (l'axe des Y) est la variable qui dépend de la taille—c'est la variable "dépendante" Y. La taille (l'axe des x) est la variable que l'on connaît indépendamment du coût—c'est la variable "indépendante" x.
Finalement la droite elle-même est le coût que le modèle devine $\hat{Y}$. Le chapeau signifie que c'est une estimation, ou "prévision".

Par exemple, le modèle devine que le prix d'une télé de 50" est de 300$; soit, $\hat{Y} = \beta_0 + \beta_1 x = 200 + (2) \cdot (50) = 300$. Selon le modèle, on estime que le coût de la télé est de 300$.
Supposons que tu connais le *vrai* coût Y, alors tu peux mesurer à quel point tu est dans le champ. Supposons que le vrai coût est de $Y = 400$\$. Alors, l'erreur dans ta prédiction est de $\varepsilon = 400 - 300 = 100$\$.
Graphiquement :

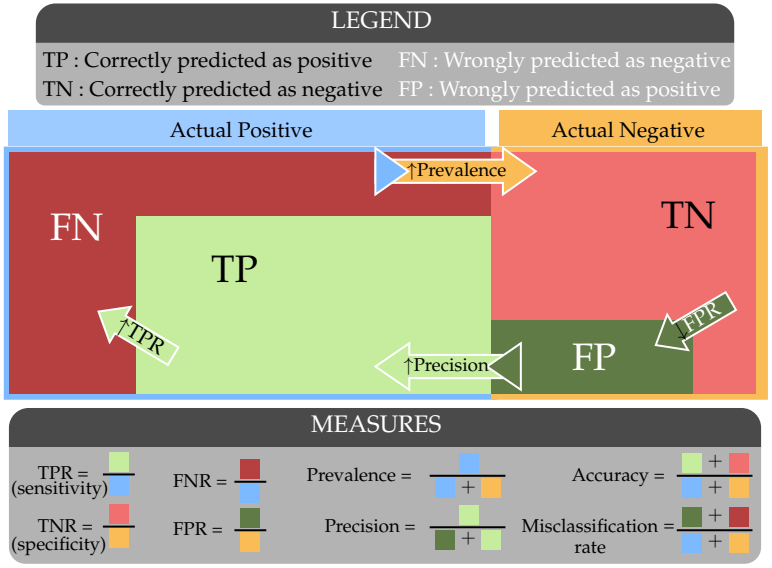


On voit donc que $Y = \beta_0 + \beta_1 x + \varepsilon$ est un "modèle" théorique pour obtenir une variable dépendante Y en fonction de :

› Une variable indépendante x multipliée par un facteur β_1.
› Un niveau de base l'intercepte β_0.
› Une erreur aléatoire ε inconnue.

Autres

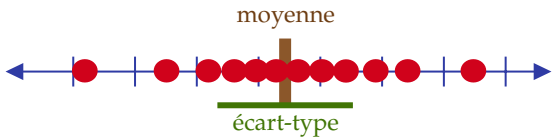
Matrice de confusion :



Erreur

Écart-type Mesure la variation entre les observations d'**un** ensemble de données.

› « *standard deviation* ».



Erreur type Mesure la variation entre les moyennes de **plusieurs** ensembles de données.

› « *standard error* ».

Troisième partie

Mathématiques IARD I

Estimations et types de données

Distributions empiriques

Notation

X Variable aléatoire de perte ;

θ Paramètre de la distribution de X ;

› Le paramètre peut être un scalaire θ ou un vecteur $\boldsymbol{\theta}$;

› Par exemple, pour une loi Gamma $\boldsymbol{\theta} = \{\alpha, \beta\}$;

› Pour simplifier la notation, on le traite comme un scalaire θ.

$F_X(x;\theta)$ Fonction de répartition de X avec paramètre θ ;

› Pour simplifier la notation, on écrit $F(x;\theta)$ sauf s'il faut être plus spécifique.

$f_X(x;\theta)$ Fonction de densité de X avec paramètre θ ;

› Pour simplifier la notation, on écrit $f(x;\theta)$ sauf s'il faut être plus spécifique.

$\{X_1, \ldots, X_n\}$ Échantillon aléatoire de n observations de X ;

$\hat{\theta}$ Estimateur de θ établit avec l'échantillon aléatoire $\{X_1, \ldots, X_n\}$;

$F(x;\hat{\theta})$ Estimation *paramétrique* de la fonction de répartition de X ;

$f(x;\hat{\theta})$ Estimation *paramétrique* de la fonction de densité de X ;

› Si θ est connu, la distribution de X est complètement spécifiée ;
En pratique, θ est inconnu et doit être estimé avec les données observées.

› On peut estimer $F_X(x)$ et $f_X(x)$ directement pour toute valeur x sans présumer une forme paramétrique ;
Par exemple, un histogramme est une estimation *non-paramétrique*.

Données complètes

Notation

X Variable d'intérêt (e.g., la durée de vie ou la perte) ;

$\{X_1, \ldots, X_n\}$ Valeurs de X pour n individus ;

$\{x_1, \ldots, x_n\}$ n valeurs observées de l'échantillon ;

› Il peut y avoir des valeurs dupliquées dans les valeurs observées.

$0 < y_1 < \ldots < y_m$ m valeurs distincts où $m \leq n$;

w_j Nombre de fois que la valeur y_j apparaît dans l'échantillon pour $j = 1, \ldots, m$;

› Il s'ensuit que $\sum_{j=1}^{m} w_j = n$;

› Pour des données de mortalité, w_j individus décèdent à l'âge y_j ;

› Si tous les individus sont observés de la naissance jusqu'à la mort c'est un « *complete individual data set* ».

r_j « *risk set* » au *temps* y_j ;

› Le nombre d'individus exposés à la possibilité de mourir au temps y_j ;

› Par exemple, $r_1 = n$ car tous les individus sont exposés à la risque de décéder juste avant le temps y_1 ;

› On déduit que $r_j = \sum_{i=j}^{m} w_i$, alias le nombre d'individus qui survivent juste avant le temps y_j.

Données incomplètes

Exemple

Soit une étude sur le nombre d'années nécessaire pour obtenir un diplôme universitaire. L'étude commence cette année et tient compte de tous les étudiants présentement inscrits, ainsi que ceux qui vont s'inscrire au courant de l'étude. Tous les étudiants sont observés jusqu'à la fin de l'étude et on note le nombre d'années nécessaire pour ceux qui complètent leurs diplômes.

Si un étudiant a commencé son cursus scolaire avant l'étude et suit présentement des cours, le chercheur a de l'information sur le nombre d'années qu'il a déjà investi. Cependant, d'autres étudiants qui se sont inscrits en même temps, mais ont cessé leurs études ne seront pas observés dans cet échantillon. Alors, l'individu est observé d'une population **tronquée à la gauche** puisque l'information sur les étudiants qui ont quitté l'université avant le début de l'étude n'est *pas disponible*.

Si un étudiant n'est pas encore diplômé lorsque l'étude prend fin, le chercheur ne peut pas savoir combien d'années supplémentaire seront nécessaires. Cet individu fait donc partie d'une population **censurée à la droite** puisque le chercheur a de l'information *partielle* (le nombre d'années minimale) sans savoir le nombre exact.

Notation

d_i État de troncature de l'individu i de l'échantillon ;

- $d_i = 0$ s'il n'y a pas de troncature ;
- Par exemple, un étudiant à commencé son programme universitaire d_i années avant le début de l'étude.

x_i Temps de "survie" de l'individu i ;

- Par exemple, le nombre d'années avant d'obtenir son diplôme ;
- Si l'étude prend fin avant que x_i soit observé, on dénote le temps de survie jusqu'à ce moment u_i ;
- Donc chaque individu a *soit* une valeur x_i ou u_i mais *pas les deux*.

Données groupées

Notation

$(c_0, c_1], (c_1, c_2], \ldots, (c_{k-1}, c_k]$ k intervalles regroupant les observations ;

$0 \leq c_0 < c_1 < \ldots < c_k$ Extrémités des k intervalles ;

n Nombre d'observations de x_i dans l'échantillon ;

n_j Nombre d'observations de x_i dans l'intervalle $(c_{j-1}, c_j]$;

- Il s'ensuit que $\sum_{j=1}^{k} n_j = n$.

r_j « *risk set* » de l'intervalle $(c_{j-1}, c_j]$ lorsque les données sont complètes ;

- Il s'ensuit que $r_j = \sum_{i=j}^{k} n_i$.

Estimation de modèles non paramétriques

Distribution empirique

Notation

g_h Somme partielle du nombre d'observations inférieur, ou égale, à y_j ;

› Il s'ensuit que $g_j = \sum_{h=1}^{j} w_h$.

Distribution empirique Distribution discrète prenant comme valeurs $y_1, \dots, y_m$ avec probabilités $\frac{w_1}{n}, \dots, \frac{w_m}{n}$;

› On peut également la définir comme la distribution discrète équiprobable des valeurs $x_1, \dots, x_n$.

$\hat{f}()$ Fonction de densité empirique ;

$$\hat{f}(y) = \begin{cases} \frac{w_j}{n}, & \text{si } y = y_j \, \forall j \\ 0, \text{sinon} \end{cases}$$

$\hat{F}()$ Fonction de répartition empirique ;

$$\hat{F}(y) = \begin{cases} 0, & y < y_1, \\ \frac{g_j}{n}, & y_j \leq y < y_{j+1},\ j = 1, \dots, m-1 \\ 1, & y_m \leq y \end{cases}$$

$\tilde{F}()$ Fonction de répartition lissée ;

› En anglais, « *smoothed empirical distribution function* » ;

› Estimation de la fonction de répartition *lissée* pour une valeur de y pas dans l'ensemble $y_1, \dots, y_m$;

› Lorsque $y_j \leq y < y_{j+1}$ et $j \in \{1, 2, \dots, m-1\}$, $\tilde{F}(y)$ est une interpolation linéaire de $\hat{F}(y_{j+1})$ et $\hat{F}(y_j)$:

$$\tilde{F}(y) = \frac{y - y_j}{y_{j+1} - y_j} \hat{F}(y_{j+1}) + \frac{y_{j+1} - y_j}{y_{j+1} - y_j} \hat{F}(y_j)$$

Estimation par noyaux

La fonction de répartition empirique résume les données d'une distribution discrète. Cependant, lorsque la variable d'intérêt X est continue on souhaite estimer une fonction de densité.
Pour une observation x_i de l'échantillon, la fonction de répartition empirique assigne une masse de probabilité de $1/n$ au point x_i. Puisque X est continue, il est normal que l'on souhaite *distribuer* cette masse *autour* de x_i.
Si l'on souhaite distribuer cette masse de façon égale, on le fait sur l'intervalle $[x_i - b, x_i + b]$ avec la fonction de x_i $f_i(x)$:

$$f_i(x) = \begin{cases} \frac{0.5}{b}, & x_i - b \leq x \leq x_i + b, \\ 0, & \text{sinon} \end{cases}$$

› Cette fonction est rectangulaire avec une base de longueur $2b$ et une hauteur de $0.5/b$ pour avoir une aire de 1.

› On peut l'interpréter comme la fonction de densité contribué par l'observation x_i ;

› On note que ceci correspond à la fonction de densité d'une distribution uniforme $U(x_i - b, x_i + b)$;

› Alors, seulement les valeurs de x contenues dans l'intervalle $(x_i - b, x_i + b)$ reçoivent une "contribution" de x_i ;

› La fonction de densité de X est donc la somme des masses de probabilité contribuées $\tilde{f}(x) = \frac{1}{n} \sum_{i=1}^{n} (x)$.

On défini $\phi_i = \frac{x - x_i}{b}$ et $K_R(\phi)$:

$$K_R(\phi) = \begin{cases} \frac{1}{2}, & -1 \leq \phi \leq 1, \\ 0, & \text{sinon} \end{cases}$$

› On trouve donc que $f_i(x) = \frac{1}{b} K_R(\phi_i)$ et $\tilde{f}(x) = \frac{1}{nb} \sum_{i=1}^{n} K_R(\phi_i)$.

Notation

b « *bandwith* » où $b > 0$;
$K_R(\phi)$ « *rectangular (box, uniform) kernel function* » ;
$\tilde{f}(x)$ Estimation de la fonction de densité selon le noyaux rectangulaire ;
$K_T(\phi)$ « *triangular kernel* » ;

$$K_R(\phi) = \begin{cases} 1 - |\phi|, & -1 \leq \phi \leq 1, \\ 0, & \text{sinon} \end{cases}$$

$K_G(\phi)$ « *Gaussian kernel* » ;
$$K_G(\phi) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{\phi^2}{2}}, -\infty < \phi < \infty$$

Estimation de modèles paramétriques

Estimation par maximum de vraisemblance pour des données incomplètes et groupées

Lorsque les données sont groupées et/ou incomplètes, les observations ne sont plus iid mais on peut quand même formuler la fonction de vraisemblance et trouver l'EMV.
La première étape est d'écrire la fonction de (log) vraisemblance adéquate pour la méthode d'échantillonnage des données.

Par exemple, soit des données groupées en k intervalles :

- On trouve avec la fonction de répartition $F(\cdot;\theta)$ que la probabilité d'être dans l'intervalle $(c_{j-1}, c_j]$ est $F(c_j;\theta) - F(c_{j-1};\theta)$;
- On pose que les observations individuelles sont iid ;
- Donc, la vraisemblance d'avoir n_j observations dans l'intervalle $(c_{j-1}, c_j]$, pour $j = 1, \ldots, k$ et $\boldsymbol{n} = (n_1, \cdots, n_k)$ est :

$$\mathcal{L}(\theta; \boldsymbol{n}) = \prod_{j=1}^{k} \left[F(c_j;\theta) - F(c_{j-1};\theta)\right]^{n_j}$$

Fonction de vraisemblance

$$\mathcal{L}(\theta; \boldsymbol{x}) = \prod_{j=1}^{k} \underbrace{f(x_j;\theta)}_{\substack{\text{probabilité de chaque} \\ \text{observation à la} \\ \text{valeur observée}}}$$

données complètes

Données groupées en k intervalles :

$$\mathcal{L}(\theta; \boldsymbol{n}) = \prod_{j=1}^{k} \underbrace{\left[F(c_j;\theta) - F(c_{j-1};\theta)\right]^{n_j}}_{\substack{\text{probabilité d'une} \\ \text{observation dans} \\ \text{l'intervalle}}}$$

données groupées

Données censurées vers la droite avec n_1 observations complètes et n_2 observations censurées à la limite de u :

$$\mathcal{L}(\theta; \boldsymbol{x}, n_2) = \underbrace{\left[\prod_{i=1}^{n_1} f(x_i;\theta)\right]}_{\substack{\text{probabilité de chaque} \\ \text{observation à la} \\ \text{valeur observée}}} \underbrace{\left[1 - F(u;\theta)\right]^{n_2}}_{\substack{\text{probabilité qu'une} \\ \text{observation soit} \\ \text{d'au moins } u}}$$

données censurées vers la droite

Données tronquées vers la gauche avec un déductible de d :

$$\mathcal{L}(\theta; \boldsymbol{x}) = \underbrace{\frac{1}{\left[1 - F(d;\theta)\right]^{n}}}_{\substack{\text{pondère par la} \\ \text{probabilité d'être} \\ \text{supérieur au déductible}}} \prod_{i=1}^{n} f(x_i;\theta)$$

données censurées vers la droite

Évaluation et sélection de modèles

Graphiquement

Densité empirique

Tests de mauvaise classification et diagnostiques

Test de Kolmogorov-Smirnov Test de Anderson-Darling Test d'adéquation du khi-carré Test du rapport de vraisemblance

Critère d'information pour la sélection de modèles

AIC BIC