Chap V - Introduction aux test

Tests : démarche

Considérons $X_1, \cdot, X_n$ indépendantes et identiquement distribuées de loi P_{θ^*} avec $P_{\theta^*} \in \left\{P_\theta, \theta \in \Theta \subset \mathbb{R}^d\right\}$. Le paramétre θ^* est inconnu et on souhaite faire un test statistique.

Un exemple introductif

Considérons n lancers indépendants d'un jeton, n grand (>30). Ces n lancers sont supposés réalisés dans des conditions strictement identiques. On note x_i le résultat du $i^{\text{ème}}$ lancer, $x_i = 1$ si le résultat est "pile" et $x_i = 0$ si le résultat obtenu est "face". On compte le nombre de "piles" obtenus en n lancers et la proportion observée de "piles" en n lancers est $\overline{x} = \sum_{i=1}^n x_i/n$. On considère que les x_i sont les réalisations de variables aléatoires $X_1, \cdots, X_n$ n i.i.d. de loi $\mathcal{B}(p^*)$ avec p^* inconnue. Nous avons vu qu'une estimation naturelle est $\overline{x}$, la proportion observée de "piles" parmi les $x_1, \cdots, x_n$. Nous avons également vu que $\overline{X}$ est un bon estimateur de p^* puisqu'il est sans biais ($\mathbb{E}(\overline{X}) = p^*$) et qu'il converge en moyenne quadratique ($\text{EQM}(\overline{X}) = \text{Var}(\overline{X}) = p^*(1-p^*)/n \underset{n\to\infty}{\longrightarrow} 0$.). Nous avons donc ainsi une bonne estimation ponctuelle de p. Nous allons maintenant tenter de répondre à la question "le jeton est-il biaisé", soit aussi est-ce que $p^* = 1/2$?. Pour cela nous allons faire un test statistique.

Notons (H_0) l'hypothèse nulle, (H_0) : $p^* = 1/2$ et (H_1) : $p^* \neq 1/2$ est l'hypothèse alternative. Nous allons tenter de décider laquelle des deux hypothèses est la plus vraisemblable vues les observations. Ce faisant, nous risquons de faire deux types d'erreurs. La première, appelé erreur de première espèce est l'erreur qui consiste à rejetter (H_0) alors qu'elle est vraie, autrement dit à rejetter (H_0) à tort. La deuxième erreur que l'on peut commettre, appelée erreur de seconde espèce est celle qui consiste à garder (H_0) alors qu'elle est fausse, c'est-à-dire à garder (H_0) alors que c'est (H_1) qui est vraie. Ces deux erreurs se produisent avec une certaine probabilité. On appelle alors risque de première espèce la quantité $\mathbb{P}$(rejetter (H_0) alors qu'elle est vraie) et risque de seconde espèce la quantité $\mathbb{P}$(garder (H_0) alors que (H_1) est vraie). Pour un test de niveau α, on impose en priorité le contrôle $\alpha \geq \mathbb{P}$(rejetter (H_0) alors qu'elle est vraie), soit le risque de première espèce inférieur à α, et, à niveau α fixé, le risque de seconde espèce est $\beta_\alpha = \mathbb{P}$(garder (H_0) alors que (H_1) est vraie). Ces deux risques satisfont

$$\alpha \geq \mathbb{P}_{H_0}(\text{ rejetter } (H_0)) \text{ et } \beta_\alpha = \mathbb{P}_{H_1}(\text{ garder } (H_0)).$$

La construction d'un test est en fait la construction d'une règle de décision via la construction d'une zone de rejet. Dans notre exemple, on rejette (H_0) quand $\overline{X}$ est "loin" de $1/2$ c'est-à-dire que $|\overline{X} - 1/2| \geq t$ où t va être tel que

$$\mathbb{P}_{H_0}(|\overline{X} - 1/2| \geq t) = \alpha.$$

Pour trouver la valeur de t, on va utiliser les propriétés de $\overline{X}$. En particulier, nous allons utiliser le résultat suivant

$$\frac{\sqrt{n}(\overline{X} - p^*)}{\sqrt{p^*(1-p^*)}} \underset{n\to\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0,1).$$

Si (H_0) est vraie, $p^* = 1/2$ et donc

$$T_n = \frac{\sqrt{n}(\overline{X} - 1/2)}{\sqrt{(1/2)(1-(1/2))}} \underset{n\to\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0,1) \quad \textbf{sous } H_0.$$

Soit u_α tel que $\mathbb{P}(|\mathcal{N}(0,1)| \geq u_\alpha) = \alpha$. Pour ce u_α, comme n est grand,

$$\mathbb{P}_{H_0}(|T_n| \geq u_\alpha) = \mathbb{P}(|\overline{X} - 1/2| \geq u_\alpha\sqrt{n}/\sqrt{(1/2)(1-(1/2))} \text{ est proche de } \alpha.$$

On choisit donc $t = u_\alpha\sqrt{n}/\sqrt{(1/2)(1-(1/2))}$. On en déduit donc la zone de rejet de (H_0) au niveau α

$$ZR_\alpha = \{|\overline{X} - 1/2| \geq u_\alpha\sqrt{n}/\sqrt{(1/2)(1-(1/2))}\} = \{|T_n| \geq u_\alpha\}.$$

La règle de décision est donc :
Si $|\overline{X} - 1/2| \geq u_\alpha\sqrt{n}/\sqrt{(1/2)(1-(1/2))}$, on rejette (H_0) au niveau α. Sinon, on ne rejette pas (H_0).
Cette règle de décision est s'écrit aussi :
Si $|T_n| \geq u_\alpha$, on rejette (H_0) au niveau α. Sinon, on ne rejette pas (H_0).

La quantité T_n est appelée la statistique de test. Elle ne dépend pas de p^* !

Cadre général

On dispose de n données $x_1, \cdots, x_n$ réalisations de n variables aléatoires $X_1, \cdots, X_n$. On va chercher à tester des hypothèses portant sur la loi de probabilité du n-uplet $(X_1, \cdots, X_n)$ modélisant les observations. On se placera essentiellement dans le cadre où les X_i sont indépendantes et identiquement distribuées (iid). On effectue un test de H_0 contre H_1, H_0 et H_1 étant deux hypothèses portant sur la loi de $X_1, \cdots X_n$.

La construction d'un test va consister à établir une règle de décision permettant de faire un choix entre les deux hypothèses H_0 et H_1 au vu d'un échantillon de même loi que X. En faisant ce test nous allons faire deux types d'erreurs. La première erreur consiste à rejetter l'hypothèse H_0 alors qu'elle est vraie. La deuxième erreur consiste à garder l'hypothèse H_0 alors qu'elle est fausse.

Choix des hypothèses : H_0 est l'hypothèse privilégiée

L'hypothèse H_0 appelée *hypothèse nulle* est celle que l'on garde si le résultat de l'expérience n'est pas clair. On conserve H_0 sauf si les données conduisent à la rejeter. Quand on ne rejette pas H_0, on ne prouve pas qu'elle est vraie ; on accepte de conserver H_0 car on n'a pas pu accumuler suffisamment d'éléments matériels contre elle ; les données ne sont pas incompatibles avec H_0, et l'on n'a pas de raison suffisante de lui préférer H_1 compte-tenu des résultats de l'échantillon. Ne pas rejetter H_0, "c'est acquitter faute de preuve". On n'abandonne pas H_0 sans de solides raisons. L'hypothèse H_1 contre laquelle on *teste* H_0 est appelée *contre hypothèse ou hypothèse alternative*. Ceci veut dire en pratique qu'on va imposer pour le test que la probabilité de rejetter H_0 à tort (alors qu'elle est vraie) soit petite, inférieure à α, que l'on appelera le niveau du test.

Erreurs de première et de seconde espèce

Lors de la prise de la décision de rejetter ou non l'hypothèse H_0, on peut commettre deux erreurs, soit rejetter à tort l'hypothèse H_0, soit la garder à tort. Les situations possibles lors de la prise de décision sont résumées dans le tableau suivant.

Décision / Réalité	H_0	H_1
H_0	Décision correcte Probabilité : $1-\alpha$	Erreur de première espèce Probabilité α
H_1	Erreur de seconde espèce Probabilité β	Décision correcte Probabilité $1-\beta$

— L'erreur de première espèce est l'erreur que l'on commet lorsqu'on rejette H_0 à tort, ie lorsqu'on choisit H_1 alors que H_0 est vraie.
La probabilité de commettre cette erreur, que l'on appelle le risque de première espèce, est notée $P(rejeter\ H_0\ à\ tort) = \mathbb{P}_{H_0}(rejeterH_0)$. Pour un niveau de test α, on impose que

$$\mathbb{P}_{H_0}(rejeter\ H_0) \leq \alpha.$$

— L'erreur de deuxième espèce l'erreur que l'on commet lorsqu'on accepte H_0 à tort, ie lorsqu'on ne rejette pas H_0 alors qu'elle est fausse ; la probabilité de commettre cette erreur, que l'on appelle le risque de deuxième espèce, est notée β avec

$$\beta = P(accepter\ H_0\ à\ tort) = P_{H_1}(accepter\ H_0)$$

La stratégie consiste à fixer le niveau du test α et à imposer que le risque de première espèce soit inférieur à α. C'est-à-dire que la probabilité de rejeter H_0 à tort est fixée à un seuil aussi faible que l'on veut, par exemple $\alpha = 5\%$. Ce qui signifie qu'il y a 5 chances sur 100 que, si H_0 est vraie, l'échantillon ne donne pas une valeur de l'observation comprise dans la zone d'acceptation de H_0. On est donc prêt à rejeter H_0 si le résultat fait partie d'une éventualité improbable n'ayant que 5% de chances de se produire. L'hypothèse H_0 est privilégiée : on veut avant tout contrôler le risque de rejeter H_0 à tort. Une fois le risque de première espèce controlé par α, on cherchera alors à minimiser le risque de deuxième espèce β, qui sera du coup une fonction de α. Autrement dit, une fois qu'on a controlé le risque de rejetter H_0 à tort, on va chercher à minimiser le risque de la garder à tort.

Choix de H_0 Selon les cas, le nom d'hypothèse nulle est réservé
— soit à *l'hypothèse telle qu'il est le plus grave de rejeter à tort.* On a vu que la probabilité de rejeter H_0 à tort est choisie petite, inférieure à α, le niveau du test. On choisit donc H_0 et H_1 de telle sorte que le scénario catastrophique, soit d'accepter H_1 alors que H_0 est vraie : ce scénario "le pire" a ainsi une petite probabilité de se réaliser α fixée (le scénario catastrophique dépend souvent du point de vue considéré).
— soit à *l'hypothèse dont on a admis jusqu'à présent la validité*, H_1 représentant la contre hypothèse suggérée par une nouvelle théorie ou une expérience récente.
— *soit à l'hypothèse qui permet de faire le test* (seule hypothèse facile à formuler, permettant calcul de la loi d'une variable aléatoire sur laquelle on peut fonder le test).

Statistique de test- zone de rejet de H_0

Le test est basé sur l'utilisation d'une *statistique de test*, T_n, fonction des variables aléatoires $X_1, \cdots, X_n$, en général liée à un estimateur. La statistique de test T_n est une variable aléatoire dont la loi sous H_0 va déterminer la zone de rejet de H_0. On rejette H_0 quand T_n est "loin" de H_0, vers H_1. Cette zone de rejet ZR_α est déterminée de telle sorte que

$$\alpha \geq P(rejeter\ H_0\ à\ tort) = \mathbb{P}_{H_0}(rejeter\ H_0) = \mathbb{P}_{H_0}(T_n \in ZR_\alpha).$$

Une fois cette zone de rejet de H_0 déterminée, le risque de deuxième espèce est donné par est

$$\beta(\alpha) = P(accepter\ H_0\ à\ tort) = P_{H_1}(accepter\ H_0) = P_{H_1}(T_n \notin ZR_\alpha).$$

L'allure de la région de rejet notée ZR_α, est déterminée par H_1, mais le calcul précis de ZR_α est fonction de α et de la loi de T_n sous H_0. Le risque de deuxième espèce $\beta = \beta(\alpha)$ est alors calculé en utilisant la loi de T_n sous H_1. On va alors espérer que ce risque de espèce $\beta = \beta(\alpha)$ soit petit α fixé.

Puissance

La qualité d'un test est donnée par sa capacité à séparer les deux hypothèses H_0 et H_1. Elle est mesurée par la puissance du test qui est la probabilité d'accepter H_0 quand H_1 est vraie. Cette puissance π est donc donnée par

$$\Pi_\alpha = P(rejetter\ H_0\ qd\ elle\ est\ fausse) = 1 - \beta(\alpha) = P_{H_1}(rejeter\ H_0).$$

La puissance mesure l'aptitude du test à rejeter une hypothèse fausse. C'est un mesure de la qualité du test. On va chercher à ce que cette puissance Π_α soit grande à niveau du test α fixé.
Si α diminue (décroît vers 0), $1-\alpha$ augmente, et la règle de décision est plus stricte : on n'abandonne H_0 que dans des cas rarissimes, et on conserve H_0 bien souvent à tort. A force de ne pas vouloir abandonner H_0, "on la garde presque tout le temps" : le risque de la garder à tort augmente, autrement dit le risque de deuxiìeme espèce β augmente, et la puissance diminue.
Rien ne dit que conserver H_0 mette à l'abri de se tromper. La probabilité d'accepter H_0 à tort est β, et cette probabilité (qui ne se calcule que si l'on connaît la loi de T_n sous H_1) peut être très importante, contrairement à α qui est fixé aussi petit qu'on veut à l'avance. Les hypothèses H_0 et H_1 ne jouent pas un rôle symétrique.

Degré de signification ou $p - value$

Lorsque le test n'est pas significatif (ie on ne rejette pas H_0), on peut se demander à quel niveau H_0 serait rejettée. Ce niveau est appelé degré de signification α_s, appelé aussi probabilité critique ou $P - value$. De même lorsqu'on rejette H_0, le test est significatif. On peut alors, pour mesurer la conviction avec laquelle on rejette H_0, calculer le degré de signification α_s. Ce degré de signification mesure la probabilité d'obtenir t_n (ou une valeur encore plus éloignée de H_0) si H_0 est vraie. C'est une mesure de l'accord entre l'hypothèse testée H_0 et le résultat obtenu. Plus il est proche de 0, plus forte est la contradiction entre H_0 et le résultat de l'échantillon, et plus on rejettera H_0 avec assurance.

Le degré de signification est le plus petit niveau α_s pour lequel le test correspondant serait significatif. Ce α_s ne peut être calculé qu'une fois que les observations ont été faites : c'est le niveau du test obtenu en choisissant la zone de rejet de H_0 la plus petite possible qui contienne l'observation. Calculer le degré de signification évite le problème de se fixer un risque α à l'avance pour effectuer le test. Beaucoup de logiciels statistiques effectuent automatiquement les tests en donnant le degré de signification. En fonction du risque choisi, on décide si on accepte ou non H_0 : il suffit de comparer α_s à α : si $\alpha_s < \alpha$ on rejette H_0. Généralement on admet que :

degré de signification	significativité du test
$0,01 < \alpha_s \leq 0,05$	significatif
$0,001 < \alpha_s \leq 0,01$	très significatif
$\alpha_s \leq 0,001$	hautement significatif

Intervalle de pari

On peut calculer, sous H_0, un intervalle de pari IP_α pour les réalisations de T_n au risque α. Cet intervalle de pari IP_α est défini par

$$\mathbb{P}_{H_0}(T_n \in IP_\alpha) = 1 - \alpha.$$

On alors

$$\mathbb{P}_{H_0}(T_n \notin IP_\alpha) = 1 - \mathbb{P}_{H_0}(T_n \in IP_\alpha) = \alpha$$

La règle de décision est donc

$$\begin{cases} \text{si } T_n \notin IP_\alpha & \text{alors on rejette } H_0, \\ \text{si } T_n \in IP_\alpha & \text{alors on accepte } H_0 \end{cases}$$

Méthode de construction d'un test

1. On dispose de données $x_1, \cdots, x_n$ réalisations de n variables aléatoires $X_1, \cdots, X_n$. On va chercher à tester des hypothèses de modélisation portant sur la loi de probabilité du n-uplet $(X_1, \cdots, X_n)$ modélisant les observations. On se placera essentiellement dans le cadre où les X_i sont indépendantes et identiquement distribuées (i.i.d.). Par exemple, supposons que l'on s'intéresse à la taille moyenne des hommes dans la population, dont on prétend qu'elle est égale à 175cm. Considérons un échantillon de n hommes choisis au hasard dans la population et on note X_i la taille du ième homme . On suppose alors que les $X_1, \cdots, X_n$ sont n variables aléatoires indépendantes et identiquement distribuées, de loi $\mathcal{N}(\mu, \sigma^2)$, $\sigma = 15$ cm étant connue. On note x_i la taille observée du i ème homme. Les x_i sont les réalisations des X_i. On a observé $\overline{x} = 176$ avec $\overline{x} = n^{-1} \sum_{i=1}^{n} x_i$.
2. On choisit l'hypothèse nulle H_0 et l'hypothèse alternative H_1. L'hypothèse H_0 correspond à une hypothèse sur le modèle donné. Très souvent, la loi commune des X_i dépend d'un (ou plusieurs) paramètre θ, et H_0 est une assertion concernant θ. Dans notre exemple, la loi des X_i est une loi normale dépendant de deux paramètres, espérance μ et variance σ^2, σ étant connue on pose $\theta = \mu$. Si on se demande si la taille moyenne a augmenté, on va tester l'hypothèse $H_0 : \mu = 175$ contre $H_1 : \mu > 175$.
3. On choisit une statistique de test T_n, *ie* une variable aléatoire T_n, *ie* une fonction de $X_1, \cdots, X_n$, dont on connaît la loi sous H_0 et qui a un comportement différent sous H_1, que sous H_0. Dans notre exemple, si la variance σ^2 est connue, alors la statistique de test T_n est $T_n = \sqrt{n}(\bar{X} - 175)/\sigma$. On sait que $\sqrt{n}(\bar{X} - \mu)/\sigma \sim \mathcal{N}(0,1)$ et donc sous H_0, $\mu = 175$ donc T_n suit sous H_0 une loi $\mathcal{N}(0,1)$, mais ne suit plus une loi $\mathcal{N}(0,1)$ si on n'est plus sous H_0. La décision va porter sur la valeur prise par cette variable aléatoire. En général cette statistique de test est liée à un estimateur ($\overline{X}$ par exemple) de la quantité qui nous intéresse et sur laquelle on effectue le test (μ par exemple).
4. L'étape suivante consiste à déterminer la zone de rejet de H_0, cette zone de rejet étant déterminée par l'hypothèse H_1 et par le contrôle du risque de première espèce, $\mathbb{P}_{H_0}(\textit{rejetter } H_0) \leq \alpha$, pour un test de niveau α. La zone de rejet, notée ZR_α est telle que

$$P(\text{rejeter } H_0 \text{ à tort}) = \mathbb{P}_{H_0}(\text{rejeter } H_0) = \mathbb{P}_{H_0}(T_n \in ZR_\alpha) = \alpha,$$

où α est le niveau du test fixé à l'avance. La zone de rejet ZR_α est un sous-ensemble des valeurs possibles de T_n qui sont improbables sous H_0. La région de rejet ne peut être déterminée que si l'on connaît la loi de T_n sous l'hypothèse H_0. Dans notre exemple, on rejette H_0 quand $\bar{X} > 175 + t$ où t est tel que

$$P_{H_0}(\bar{X} > 175 + t) \leq \alpha, \text{ soit aussi tel que } \mathbb{P}_{H_0}\left(T_n > t\sqrt{n}/\sigma\right) \leq \alpha.$$

Comme sous H_0, $\mu = 175$, on en déduit que sous H_0, $T_n \sim \mathcal{N}(0,1)$. Soit u_α tel que si $Z \sim \mathcal{N}(0,1)$ alors $P(Z > u_\alpha) = \alpha$. Pour ce u_α, on a

$$\mathbb{P}_{H_0}\left(T_n > u_\alpha\right) = \alpha.$$

On choisit donc t tel que $t\sqrt{n}/\sigma = u_\alpha$, soit $t = u_\alpha\sigma/\sqrt{n}$. On en déduit la zone de rejet de H_0 au niveau α,

$$ZR_\alpha = \{T_n > u_\alpha\} = \{\bar{X} > 175 + u_\alpha\sigma/\sqrt{n}\}.$$

5. On détermine ensuite la règle de décision :

$$\begin{cases} \text{si } T_n \in ZR_\alpha & \text{alors on rejette } H_0, \\ \text{si } T_n \notin ZR_\alpha & \text{alors on ne rejette pas } H_0 \end{cases}$$

On peut également formuler la règle de décision en terme de zone d'acceptation, ou d'intervalle de pari :

$$\begin{cases} \text{si } T_n \in IP_\alpha & \text{alors on ne rejette pas } H_0 \\ \text{si } T_n \notin IP_\alpha & \text{alors on rejette } H_0 \end{cases}$$

6. Mise en oeuvre du test : application numérique et conclusion
   Soit t_n la réalisation de T_n. Dans notre exemple $t_n = \sqrt{n}(176 - 175)/\sigma$. La conclusion du test vient de :
   • si $t_n \in ZR_\alpha$ alors on rejette H_0 (au risque α de se tromper) : il est très peu probable d'obtenir les résultats que l'on a trouvés si H_0 est vraie. Les données sont en contradiction avec H_0
   • si $t_n \notin ZR_\alpha$ alors on ne rejette pas H_0 : les données ne sont pas en contradiction avec H_0
7. Lorsque l'on est amené à rejeter H_0, on dit que le test est significatif. Lorsqu'on ne rejette pas H_0, au niveau α, on dit que le test n'est pas significatif au niveau α.
8. Pour évaluer de façon précise la significativité du test on va calculer le degré de signification ou p-value. On va chercher à évaluer l'incompatibilité des observations avec l'hypothèse H_0, en calculant le degré de signification. Plus il sera proche de 0, plus forte est la contradiction entre H_0 et le résultat de l'échantillon, et plus on rejettera H_0 avec assurance. Dans notre exemple, l'hypothèse H_1 est (H_1) : $\mu > 175cm$. Par conséquent, si t_n est la réalisation de la variable aléatoire T_n, statistique du test, alors le degré de signification α_s est donné par

$$\alpha_s = \mathbb{P}_{H_0}(T_n > t_n).$$

Remarques : Il est nécessaire pour construire le test, de connaître la loi de T_n sous H_0 et que T_n ait un comportement différent sous H_0 et sous H_1 ; mais dans de nombreux cas, on ne connaît pas la loi de T_n sous H_1. C'est le cas dans notre exemple ou la loi de $T_n = \sqrt{n}(\overline{X} - 175)/\sigma$ n'est pas connue sous l'hypothèse H_1 : $\mu > 175$. Dans ces cas là, on calcule la fonction puissance qui à toute valeur $\mu = \mu_1$ sous (H_1) associe

$$\mu_1 \mapsto \Pi_{\mu_1} = \mathbb{P}_{\mu=\mu_1}(ZR_\alpha).$$