Chapitre 10
Tests statistiques: introduction

Un test statistique est un procédé d'inférence: son but est d'énoncer des propriétés de la population en s'appuyant sur un échantillon d'observations. A l'aide d'un test on construit des intervalles de confiance qui expriment le degré d'incertitude associé à une estimation. Ce chapitre introduit les concepts nécessaires pour développer et appliquer les tests et les intervalles de confiance. Les chapitres suivantes décrivent des tests et des intervalles de confiance spécifiques pour des situations fréquemment rencontrées en pratique.

10.1 Le concept de test statistque

L'étape préliminaire à la réalisation d'un test est la formulation d'*hypothèses* concernant les caractéristiques moyennes ou "paramètres" de la population en question. Un premier type d'hypothèse est l'*hypothèse nulle*, notée H_0. En général, elle affirme que les paramètres ont des valeurs données (par exemple, suggérées par des études antérieures). L'utilisateur du test cherche à établir si son activité pourra se fonder sur cette hypothèse ou s'il vaudra mieux d'admettre que cette hypothèse est fausse. Dans ce cas, il préférera une *hypothèse alternative,* notée H_1, qui nie H_0.

Exemple

Le coût moyen d'un séjour dans un grand hôpital suisse pour "chirurgie cardiovasculaire" était de 11'000 Fr en 1999. Le directeur des finances doit établir le budget pour 2003; peut il admettre que le coût moyen sera de 11'000 Fr ? Ne vaut-il pas mieux qu'il suppose qu'il sera supérieur ? En d'autres termes, si μ indique le coût moyen pour 2003, les hypothèses sont:

$$H_0: \quad \mu = 11'000, \qquad H_1: \quad \mu > 11'000.$$

Pour choisir entre H_0 et H_1 on utilise un *test statistique*. Un test est un procédé qui permet de décider entre H_0 et H_1 sur la base d'un *échantillon d'observations*, d'une *statistique de test* et d'une *règle de décision*. La règle repose sur la statistique; elle doit permettre d'*accepter* (*confirmer*) l'hypothèse nulle ou la *rejeter* (*infirmer*). Lorsque l'hypothèse nulle est rejetée, l'utilisateur se prononce en faveur de l'hypothèse alternative.

Exemple: continuation

Le directeur obtient un échantillon aléatoire de séjours en "chirurgie cardiovasculaire" du deuxième semestre 2002 ainsi que leurs coûts. Il calcule le coût moyen $\hat{\mu}$ et la différence

$$d_0 = \hat{\mu} - 11'000.$$

Cette différence est la statistique de test et le directeur utilise la règle suivante:
"si $d_0 < 1'000$ Fr, accepter H_0; si $d_0 > 1'000$ Fr, rejeter H_0 et choisir H_1".

Deux types d'erreurs sont possibles (Figure 1):

- rejeter une hypothèse nulle vraie: *erreur de type I*,
- accepter une hypothèse nulle fausse: *erreur de type II*.

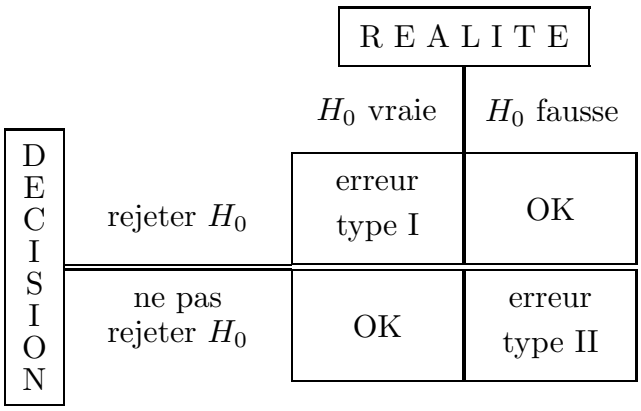

Figure 1. Les types d'erreur

La règle de décision (par exemple, la limite de 1'000 Fr) doit être telle que la probabilité de commettre une erreur de type I est plus petite d'un certain *niveau* ou *seuil* α préétabli (par exemple, $\alpha = 5\%$). Pour atteindre ce but il faudra calculer la distribution de la statistique de test en supposant que H_0 est correcte. Cette distribution s'appelle la *distribution nulle* de la statistique de test.

Pour effectuer ce calcul il est parfois nécessaire d'admettre que les données peuvent être décrites à l'aide d'un modèle ou qu'elles satisfont à certaines conditions. On dira alors que le calcul est effectué en s'appuyant sur des *conditions d'application.*

Exemple: continuation

Conditions d'application: (1) la distribution des coûts est lognormale (c'est-à-dire, il est raisonnable de décrire cette distribution à l'aide du modèle lognormal); (2) la distribution des coûts ne change pas entre le deuxième semestre 2002 et 2003.

La *probabilité d'erreur de type I* est donc

$$P_0(\text{rejeter } H_0),$$

où le suffixe "0" indique que le calcul est effectué en supposant que H_0 est correcte. Comme on l'a déjà dit, la règle de décision est choisie de façon à maintenir cette probabilité sous un certain niveau. D'autre part, si l'hypothèse H_1 est bien spécifiée (voir remarque ci-dessous) on pourra calculer aussi

$$P_1(\text{rejeter } H_0),$$

où le suffixe "1" indique que le calcul est effectué en supposant que H_1 est correcte. Cette probabilité s'appelle la *puissance* du test. C'est la probabité de rejeter l'hypothèse nulle si l'alternative est correcte. En général, il est souhaitable que la puissance du test soit élevée (par exemple 95%) et on atteindra ce but en prenant un échantillon de "taille suffisammant élevée".

Remarques

1. Il n'est pas possible de calculer la puissance d'un test si on ne spécifie pas précisément H_1. Par exemple, on ne peut pas effectuer des calculs sous l'alternative $\mu > 11'000$. Il faut spécifier une valeur "simple" de μ, par exemple $\mu = 13'000$ Fr.

2. Lorqu'un test rejette une hypothèse, il n'est pas certain qu'elle soit fausse ! On ne peut même pas affirmer que, si un test de niveu α rejette H_0, cette hypothèse est fausse avec probabilité $1 - \alpha$. En effet, il est impossible d'établir si une hypothèse est vraie ou fausse sans examiner la population de façon exhausive. Toutefois, avant d'appliquer le test, le statisticien sait que la probabilité qu'il commettra une erreur de type I est inférieure à α. Pour interpréter cela le statisticien peut imaginer d'appliquer ce test un grand nombre de fois pendant sa vie, par exemple 1000 ! La proportion d'hypothèses H_0 qu'il aura rejeté incorrectement sera α (par exemple, environ 50 sur 1000).

Nous allons considérer deux exemples de tests de niveau α pour des situations très élémentaires. Ces exemples ne représentent pas des tests couramment utilisés en pratique. Leur but est celui d'illustrer les concepts.

Exemple: tester si une moyenne est égale à une valeur donnée

La moyenne de population $\mu(X)$ d'une certaine variable X est inconnue. Nous écrivons μ à la place de $\mu(X)$ et considérons les hypothèses

$$H_0 : \quad \mu = \mu_0 \qquad \text{et} \qquad H_1 : \quad \mu > \mu_0.$$

Par exemple, X est la taille des poissons du lac Léman. La taille moyenne μ est inconnue et un pêcheur considère les hypotèses H_0: $\mu = 5$ cm et H_1: $\mu > 5$ cm. Selon le cas, il choisit son filet.

Soit $x_1, \ldots, x_n$ un échantillon d'observations indépendants de X (par exemple, les tailles de 30 poissons pris selon échantillonnage simple). Il est raisonnable d'utiliser la statistique

$$D(X_1, \ldots, X_n) = \hat{\mu} - \mu_0$$

où $\hat{\mu}$ est la moyenne arithmétique $\hat{\mu}(X_1, \ldots, X_n) = \sum X_i/n$. L'échantillon fournit une valeur observée de D, notée d_0 (par exemple, $\hat{\mu} = 7$ cm et $d_0 = 2$ cm). Est-ce-que d_0 est suffisamment élevé pour rejeter H_0 ? Pour répondre à cette question il faut comparer d_0 à une certaine limite que nous allons déterminer. Considérons la statistique standardisée

$$Z(X_1, \ldots, X_n) = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}/\sqrt{n}},$$

où $\hat{\sigma}$ est l'écart type de $x_1, \ldots, x_n$, et notons sa valeur observée par z_0 (par exemple, $\hat{\sigma} = 3.94$ et $z_0 = 2.78$). Grâce au théorème centrale limite, sous H_0 (et sous de faibles conditions d'application)

$$Z \sim \mathcal{N}(0, 1)$$

approximativement. En d'autres termes, la distribution nulle de Z est approximativement la distribution de Gauss standard. Soit donc $z_{1-\alpha}$ le quantile $1 - \alpha$ de cette distribution. Par exemple, pour $\alpha = 5\%$ on trouve $z_{0.95} = 1.645$. La règle de décision pourra enfin être formulée:

$$\text{rejeter } H_0 \text{ si } Z > z_{1-\alpha}.$$

Evidemment on applique cette règle à la valeur observée de Z. Par exemple, si $z_0 = 2.78$ il faut rejeter H_0. Pour cette règle la probabilité d'erreur de type I est

$$P_0(\text{rejeter } H_0) \approx P(Z > z_{1-\alpha}) = \alpha$$

où P est calculé à l'aide de la distribution de Gauss.

Exemple: tester si une proportion est égale à une valeur donnée

Dans une population, le taux p d'individus avec une certaine caractéristique est inconnu: par exemple le taux de fumeurs en Suisse. Nous cherchons un test pour les hypothèses

$$H_0 : p = 50\%, \qquad H_1 : p \neq 50\%.$$

Le seuil souhaité est $\alpha \approx 5\%$. Nous disposons d'un échantillon de 10 personnes prises au hasard (selon un plan d'échantillonnage simple).

Nous considérons la variable aléatoire

$$K = \text{ nombre d'individus avec la caractéristique.}$$

K sera la statistique de test. Sous l'hypothèse H_0 (et si la taille de la population est très grande – condition d'application) elle suit un modèle binomial $\mathcal{B}(n = 10, p = 0.5)$ et sa distribution est donnée dans le tableau suivant.

k:	0	1	2	3	4	5	6	7	8	9	10
$P(K = k)$:	0.0010	.0098	.0439	.1172	.2051	.2461	.2051	.1172	.0439	.0098	.0010

Nous remarquons que

$$P(K = 0 \text{ ou } K = 1 \text{ ou } K = 9 \text{ ou } K = 10) = 0.022,$$

tandis que

$$P(K = 0 \text{ ou } K = 1 \text{ ou } K = 2 \text{ ou } K = 8 \text{ ou } K = 9 \text{ ou } K = 10) = 0.109.$$

Nous définissons alors la règle de décision suivante:

$$\text{“Rejeter } H_0 \text{ si } K = 0 \text{ ou } K = 1 \text{ ou } K = 9 \text{ ou } K = 10\text{”}. \tag{10.1}$$

En effet, dans ce cas, la probabilité de rejeter H_0 si elle est vraie est $0.022 < 5\%$, tandis que si on rejetait H_0 pour $K \geq 8$ ou $K \leq 2$ cette probabilité deviendrait $0.109 > 5\%$.

Remarques

1. Le modèle binomial n'est pas applicable si la population a une taille N modérée par rapport à n (car dans ce cas, la probabilité que la deuxième, personne – ainsi que la troisième etc. – soit un fumeur n'est plus p).

2. Le même raisonnement peut servir à construire un test pour l'hypothèse

$$H_0 : \text{le taux } p \text{ est égal à une valeur donnée } p_0,$$

(où p_0 n'est pas nécessairement 50%) avec une taille d'échantillon n quelconque (mais beaucoup plus petit que la taille de la population). Evidemment, sous H_0, il faudra considérer le modèle binomial $\mathcal{B}(n, p = p_0)$.

3. Souvent, la règle de décision est du type “rejeter H_0 si $S > c_\alpha$” où S est la statistique de test et c_α une constante qui dépend du niveau α. On appelle c_α une *valeur critique*.

10.2 Le p-value

Très souvent, surtout lorsqu'on utilise les logiciels statistiques, la fixation du niveau α ne précède pas la détermination de la règle de décision. En effet, un cheminement inverse est utilisé: on calcule, sous l'hypothèse nulle, la probabilité p d'obtenir "une valeur plus extrême que celle qu'on a observé". Si cette probabilité – dite *p-value* – est très petite, l'événement serait surprenant et alors l'hypothèse H_0 est rejetée.

Supposons que dans l'échantillon de l'exemple on ait observé 0 fumeurs. Calculons la probabilité $p = P(K = 0 \text{ ou } K = 10) = 0.002$ d'obtenir une valeur aussi extrême que celle observée, sous l'hypothèse H_0. Cette probabilité est très petite: elle est inférieure à $\alpha = 5\%$. En d'autres termes, l'événement observé ne soutient pas l'hypothèse H_0. Nous rejetons alors H_0 en faveur de l'alternative H_1.

En général on peut directement définir la règle de décision à l'aide du p-value. Dans l'exemple du test pour la moyenne, la règle "rejeter H_0 si $Z > z_{1-\alpha}$" est équivalente à "rejeter H_0 si le p-value est inférieur à α".

La limite de ce qu'on appelle surprenant (rare) est arbitraire, mais dans beaucoup de domaines (biologie, médecine) on utilise assez systématiquement 5% (on lit "$p < 5\%$" dans les publications). La borne de 5% peut être abaissée dans le cas où une erreur de type I pourrait avoir des conséquences jugées graves. Considérons par exemple le problème de comparer la survie moyenne de patients soumis à un procédé opératoire nouveau et très coûteux à la survie obtenue par un procédé traditionnel. Supposons que l'hypothèse nulle d'égalité entre les survies moyennes soit rejetée par un test statistique en faveur du nouveau traitement. Les conséquences financières d'une introduction généralisée du nouveau procédé pourraient être très lourdes en cas d'erreur de type I. Evidemment, tout dépend du point de vue !

10.3 Test unilatéral et test bilatéral

Considérons encore les exemples de la Section 10.1. Dans le cas du taux, nous avons développé un *test bilatéral*, c'est à dire, nous avons cherché à savoir si la fréquence de fumeurs dans l'échantillon était suffisamment "petite ou élevée" pour rejeter H_0. Le rejet de H_0 était équivalent à l'acceptation de H_1, c'est à dire, à affirmer que le taux de population est "inférieur ou supérieur" à 50%. Or, le chercheur a souvent en tête une alternative H_1 unilatérale; c'est à dire qu'il cherche à savoir si par exemple le taux p est "supérieur" à 50%. Il peut alors adopter un *test unilatéral* et rejeter H_0 seulement si le nombre de fumeurs dans l'échantillon est élevé, par exemple:

"Rejeter H_0 si $K = 9$ ou $K = 10$".

Cette règle est associée à une probabilité d'erreur de type I égale à $P(K = 9 \text{ ou } K = 10) = 0.011$, qui est la moitié de la probabilité d'erreur associée à la règle bilatérale. Evidemment, le chercheur peut aussi adopter la règle "Rejeter H_0 si $K = 8$ ou $K = 9$ ou $K = 10$". avec $P(K = 8 \text{ ou } K = 9 \text{ ou } K = 10) = 0.0547 \approx 5\%$. On remarque ainsi, que tout en gardant un seuil à 5%, il devient plus probable que l'alternative soit acceptée si elle est vraie: en d'autres termes, la puissance du test augmente.

Dans le cas du test pour la moyenne nous avons considéré l'hypothèse alternative unilatérale H_1: $\mu > \mu_0$ et nous avons utilisé la règle de décision unilatérale: rejeter H_0 si $Z > z_{1-\alpha}$. Pour tester H_0 contre l'alternative bilatérale H_1: $\mu \neq \mu_0$ au niveau α il faut utiliser la règle bilatérale

"Rejeter H_0 si $Z < z_{\alpha/2}$ ou si $Z > z_{1-\alpha/2}$".

10.4 L'intervalle de confiance

Considérons l'exemple du test concernant la moyenne $\mu(X)$ et supposons que $\hat{\mu}$ ait été observée, par exemple 7 cm. Quelle est la qualité de cette estimation ? Quelle mesure d'erreur pouvons-nous associer à cette estimation ? Une réponse nous est fournie par l'intervalle de confiance. Ce concept est fondé sur celui du test. En effet, ayant observé une certaine valeur de $\hat{\mu}$, imaginons de tester les hypothèses nulles

$$H_0 : \mu = \mu_0$$

pour toutes les valeurs de μ_0 comprises entre $-\infty$ et ∞, et ceci avec probabilité d'erreur de type I inférieure à α (par exemple, 5%). Certaines de ces hypothèses seront rejetées, d'autres acceptées. L'ensemble des valeurs μ_0 acceptées forme un intervalle appelé *intervalle de confiance* pour $\mu(X)$ avec *coefficient de couverture* ou *coefficient de confiance* $1-\alpha$ (par exemple, 95%).

Plus précisément, supposons que H_1 est bilatérale: $\mu \neq \mu_0$. L'hypotèse H_0: $\mu = \mu_0$ est donc acceptée si

$$z_{\alpha/2} < \frac{\hat{\mu} - \mu_0}{\hat{\sigma}/\sqrt{n}} < z_{1-\alpha/2}$$

c'est à dire, comme $z_{\alpha/2} = -z_{1-\alpha/2}$, si

$$\hat{\mu} - z_{1-\alpha/2}\hat{\sigma}/\sqrt{n} < \mu_0 < \hat{\mu} + z_{1-\alpha/2}\hat{\sigma}/\sqrt{n}.$$

L'ensemble des valeurs μ_0 qui satisfont ces inégalités est un intervalle de confiance pour μ avec coefficient de couverture $1-\alpha$. Dans l'exemple, avec $\hat{\mu} = 7$ cm, $\hat{\sigma} = 3.94$ cm, $n = 30$, $z_{1-\alpha/2} = z_{0.975} = 1.960$, on trouve l'intervalle (5.59,8.41).

Notons que les limites de l'intervalle sont des variables aléatoires et que, sous H_0,

$$P(\hat{\mu} - z_{1-\alpha/2}\hat{\sigma}/\sqrt{n} < \mu_0 < \hat{\mu} + z_{1-\alpha/2}\hat{\sigma}/\sqrt{n}) \approx 1 - \alpha.$$

En d'autre termes, avant de tirer l'échantillon, on sait que l'intervalle de confiance de coefficient $1-\alpha$ couvrira le paramètre $\mu(X)$ avec probabilité $1-\alpha$. Comment interpréter cette affirmation ? Si on observe un nombre très élevé d'échantillons et si chaque fois on calcule l'intervalle de confiance, une proportion $1-\alpha$ de ces intervalles couvrira la valeur inconnue du paramètre ! De façon moins précise, on interprète un intervalle de confiance comme un ensemble de valeurs plausibles du paramètre étant donné l'échantillon.

Considérons encore l'exemple du test pour le taux p de fumeurs. Supposons que le nombre de fumeurs observé dans un échantillon de 10 personnes est $K = 8$. On peut alors construire un intervalle de confiance pour le paramètre p avec coefficient de couverture $1-\alpha$ à l'aide d'un programme informatique qui calcule des tests de niveau α de H_0: $p = p_0$ contre H_1: $p \neq p_0$ pour toutes les valeurs p_0 comprises entre 0% et 100%. Certaines de ces hypothèses seront rejetées, d'autres acceptées. L'ensemble des valeurs p_0 acceptées forme un intervalle de confiance (p_g, p_d). La proportion observée (80%) est approximativement au milieu de l'intervalle. A noter que l'on obtient des intervalles différents selon les valeurs de K ! L'interprétation de l'intervalle reste la même que dans l'exemple précédent.

Remarque

Un test bilatéral fournit une borne inférieure ainsi qu'une borne supérieure d'un intervalle de confiance. Un test unilatéral fournit une seule borne, l'autre est $-\infty$ ou $+\infty$.