Titre

Tester sans se tromper*

Antoine ROLLAND, pour Tangente 196

%%

Les tests statistiques sont bien utiles pour valider une hypothèse sur une population à partir d'une observation sur un échantillon aléatoire. Mais ils sont aussi la source de nombreuses erreurs. Nous présentons ici les erreurs de première et deuxième espèce dans les tests d'hypothèse suivant le formalise de Neyman-Pearson.

%%

Dans le domaine de la statistique inférentielle, un test d'hypothèse vise à prendre une décision sur une statistique (inconnue) d'une population en fonction d'une estimation (connue) de cette statistique sur un échantillon aléatoire de cette population (voir encadré 1 un exemple classique de test).

Effectuer un test d'hypothèse statistique consiste d'abord à poser une hypothèse de base (conservatrice) nommée H0 sur la population, et une hypothèse H1 alternative. On tire ensuite un échantillon au hasard dans la population, et on calcule la statistique recherchée dans cet échantillon. Une règle de décision nous dit alors si la statistique observée dans l'échantillon est compatible ou non avec l'hypothèse H0 posée sur la population, ou plutôt *dans quelle mesure* la statistique observée est compatible avec H0. C'est par un calcul de probabilités a priori que l'on décide si le résultat observé est probable ou peu probable compte tenu de l'hypothèse H0.

Les erreurs possibles

Il existe quatre cas possibles lors d'un test d'hypothèse, suivant la situation réelle et ce que dit le test :

	Réalité	
Résultat du test	Hypothèse H0 vraie	Hypothèse H0 fausse
Le test ne refuse pas H0	Résultat correct	Erreur de deuxième espèce β
Le test refuse H0	Erreur de première espèce α	Résultat correct

Il y a donc deux situations d'erreur : l'erreur de première espèce quand le test refuse l'hypothèse H0 alors qu'elle est vraie (erreur de première espèce, ou « faux négatif ») ; l'erreur de deuxième espèce quand le test ne refuse pas l'hypothèse H0 alors qu'elle est fausse. On peut faire une analogie avec un tribunal qui doit décider si l'accusé est innocent (ce qu'il est par défaut, c'est l'hypothèse H0 conservatrice) ou s'il y a suffisamment de raisons de penser qu'il est coupable. L'erreur de première espèce consiste à déclarer coupable un innocent. L'erreur de deuxième espèce consiste à innocenter un coupable.

Il est généralement facile de calculer l'erreur de première espèce α. Supposons que le test porte sur une statistique S. Sous l'hypothèse H0, on peut calculer la distribution de S, et donc calculer une valeur seuil pour S correspondant à une probabilité α d'être au-delà du seuil. On rejette l'hypothèse au risque α si la valeur observée S_obs est au-delà du seuil : cela signifie que si H0 est vrai, on a un risque α de rejeter à tort cette hypothèse. Typiquement, pour α=5%, cela signifie que si H0 est vrai, on a une chance sur 20 de rejeter l'hypothèse H0. La valeur 1- α est appelée degré de significativité du test.

Il est généralement plus difficile de calculer l'erreur de deuxième espèce : en effet, cette erreur β dépend entre autre du degré de fausseté de l'hypothèse alternative H1, qui n'est pas connue a priori. La valeur 1- β est appelée la puissance du test : c'est la capacité du test à détecter l'hypothèse H1.

Les valeurs de α et β sont liées : on ne peut pas avoir un test à la fois très significatif et très puissant. On s'en rend compte en prenant les situations extrêmes. En acceptant toujours H0, on accepte H0 à chaque fois qu'elle est vrai, donc α=0 … mais aussi à chaque fois qu'elle est fausse et β=1 ! De même en rejetant systématiquement H0 on a β=0 mais α=1.

En se focalisant uniquement sur l'erreur de première espèce α, on néglige la capacité du test à bien détecter ce que l'on cherche. Lors de la conception du test, il est donc impératif également de s'interroger sur la puissance du test : quelle valeur alternative est-on capable de détecter avec une puissance acceptable (voir encadré 2).

Et d'autres erreurs encore.

Il faut se garder d'une grave confusion : la valeur 1-α indique la probabilité que le test conduise à accepter H0 si l'hypothèse est vraie. Mais elle n'indique pas du tout la probabilité de H0 d'être vraie si le test accepte l'hypothèse ! C'est un contresens courant, et dangereux, de croire qu'accepter une hypothèse au risque de 5% veut dire qu'il y a 95% de chance que cette hypothèse soit vraie. Au contraire, cette valeur peut être bien plus faible, ou bien plus importante. Regardons le cas d'un test biologique pour vérifier qu'un patient est porteur d'un certain virus[1]. On suppose qu'un tel test existe avec un degré de significativité de 99% (soit α=1% : on détecte le virus s'il est présent dans 99% des cas) et une puissance de 95% (soit β=5% : 5% des personnes saines sont détectées malgré tyout comme porteuses du virus). Regardons les deux cas suivants :

Dans le cas 1, nous avons 99000 malades et 1000 personnes saines.

	H0 vraie (malade)	H0 fausse (personne saine)
Test positif	98000	50
Test négatif	1000	950

Si le test est négatif, c'est-à-dire si le test rejette H0, alors on a encore plus d'une chance sur 2 d'être porteur du virus : non pas 1%, mais plus de 50% d'erreur d'interprétation sur le test négatif!

Dans le cas 2, nous avons 1000 malades et 99000 personnes saines.

	H0 vraie (malade)	H0 fausse (personne saine)

[1] Cette situation est plus qu'une simple analogie : pour déterminer la présence ou l'absence d'un virus, on fait une hypothèse sur une quantité de marqueurs biologiques dans le corps, que l'on analyse via la prise d'un échantillon. Si la valeur dans l'échantillon est en dessous d'un certain seuil, alors on rejette l'hypothèse H0 de présence du virus.

Test positif	10	990
Test négatif	990	98110

Dans cette situation, si le test est positif, c'est-à-dire si le test ne rejette pas H0, alors on a 99% de chance de ne pas être porteur du virus ! On voit donc qu'on ne peut pas s'arrêter à l'étude de α pour déterminer la capacité d'un test, mais qu'il faut également regarder la probabilité pour H0 d'être vraie si le test dit qu'elle l'est.

La p-valeur

De plus en plus fréquemment dans un certain nombre d'études est donnée la p-valeur plutôt que le seuil de significativité choisi. La p-valeur est calculée a posteriori une fois observée la statistique S dans l'échantillon. Elle correspond à la probabilité d'obtenir une valeur au moins aussi extrême que la valeur observée dans l'échantillon si l'hypothèse H0 est vraie. Plus la p-valeur est petite, plus on se dit que l'hypothèse H0 risque d'être fausse. Mais c'est encore un contresens qui confond la probabilité d'observer S si H0 est vraie et la probabilité que H0 soit vraie si on observe S. La mauvaise compréhension de ce qu'est la p-valeur est un problème majeur aujourd'hui dans la recherche en psychologie et en médecine, et de plus en plus les statisticiens déconseillent l'usage de la p-valeur comme critère de qualité d'une hypothèse.

Encadré 1 : un exemple de test d'hypothèse

Supposons que l'on souhaite comparer un nouveau traitement médical A à un traitement de référence B. La statistique étudiée ici est la proportion de patients dont l'état de santé est amélioré par le traitement considéré. On sait que le traitement B permet une amélioration de l'état de santé de 60% des patients. On note p_a la proportion (inconnue) de patients dont l'état de santé est amélioré par le traitement A. On ne peut évidemment pas traiter la population malade entière, on va donc traiter un échantillon pour voir le traitement A est différent du traitement B. Le test d'hypothèse sera donc

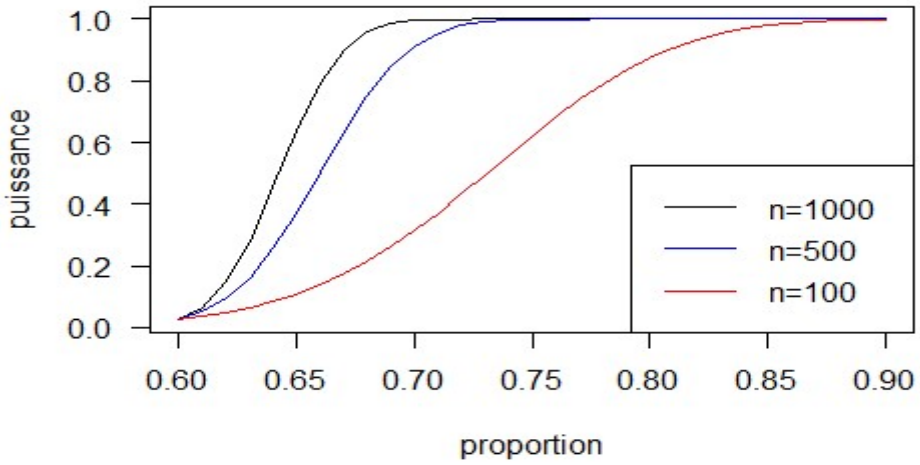
H0 : p_a =0,6 : hypothèse conservatrice : le traitement est équivalent à B

H1 : p_a≠0,6 : hypothèse alternative qui montre que le traitement A est différent du traitement B.

On prend un échantillon de taille n. Sous l'hypothèse H0, et pour un risque α donné, on peut calculer un intervalle de fluctuation à 1-α de la valeur p_n de la proportion de malades dans l'échantillon dont l'état de santé s'est amélioré : c'est l'intervalle $]p_a - z_{1-\alpha}\sqrt{\frac{p_a(1-p_a)}{n}}\,;\, p_a + z_{1-\alpha}\sqrt{\frac{p_a(1-p_a)}{n}}[$ où $z_{1-\alpha}$ est le 1-α quantile de la loi normale. Typiquement pour α=5%, p_a=0,6 et n=92, l'intervalle de fluctuation est]0,5 ;0,7[. La règle de décision est donc : si p_n est compris entre 0,5 et 0,7, alors on ne rejettera pas H0. Si p_n n'est pas compris dans cet intervalle, alors on rejettera H0 et on pourra dire que le traitement A est significativement meilleur (resp. moins bon) que le traitement B au risque de 5% si p_n est plus grand que 0,7 (resp. plus petit que 0,5).

Encadré 2 : une courbe de puissance

La valeur alternative étant souvent inconnue, on ne peut pas calculer directement la puissance d'un test, mais on peut calculer une courbe de puissance. Pour reprendre l'exemple de l'encadré 1, on peut calculer la puissance du test de proportion pour une taille d'échantillon n fixée, en faisant varier la « vraie » proportion alternative. On voit dans le graphe que pour un échantillon de taille 100, le test rejette l'hypothèse H0 à raison à tous les coups si la vraie valeur de p_a est de 0,85. Si la valeur de p_a est de 0,75, le test ne rejette l'hypothèse H0 que dans un cas sur deux. On voit sur le graphe que l'on peut nettement améliorer la puissance du test en augmentant la taille de l'échantillon. Il est donc indispensable, avant de mettre en œuvre un test d'hypothèse, de se demander quel est la taille nécessaire de l'échantillon pour rejeter H0 en fonction de l'effet que l'on imagine observer. Par exemple ici un échantillon de taille 100 suffit si l'on pense que la valeur de p_a est de 0,85, mais on préfèrera un échantillon de taille 500 si on pense que p_a est plus proche de 0,7.



Encadré 3 : le contrôle qualité

Le contrôle qualité dans les entreprises industrielles est un autre exemple de test d'hypothèse. Une entreprise fabriquant un produit en très grande quantité par lots veut savoir si son lot respecte le cahier des charges établi par le client. Comme on ne peut pas vérifier un par un tous les produits, le contrôle qualité en prélève un petit échantillon et prend la décision de commercialiser ou non le lot en fonction du pourcentage de produits défectueux observé dans l'échantillon. C'est ce qu'on appelle le contrôle par échantillonnage.

On définit une « qualité acceptable » q par le client (par exemple 1% de produits défectueux). Ici, H0 est que la proportion p de défectueux dans un lot est inférieure à q. Le risque de première espèce α est alors appelé « risque fournisseur » : c'est le risque de refuser un lot alors qu'il est de qualité suffisante ; le fournisseur détruit un lot pourtant acceptable. Le risque de deuxième espèce β est appelé « risque client » : c'est le risque d'accepter un lot alors qu'il est de qualité insuffisante ; le client achète un lot qui ne lui convient pas. L'élaboration d'un *plan d'échantillonnage* permet de calculer la taille de l'échantillon et la règle de décision en fonction des q, α et β souhaités.