THÉORIE DES TESTS STATISTIQUES

1) Notions de base :

a) Qu'est-ce qu'une hypothèse :

Une part importante des statistiques est simplement destinée à nous renseigner sur l'état des choses : Recensement, collationnement de données, sondages, etc. Mais on utilise aussi les statistiques pour comprendre la façon dont les choses se passent : Par exemple on a pensé que le nombre d'élèves au cours préparatoire était déterminant dans la réussite de l'apprentissage de la lecture. Le Ministère de l'Éducation Nationale a donc mis en place une expérience en 2002 pour essayer de vérifier cette affirmation. Les résultats de l'expérimentation dans les classes sont soumis à une grande variabilité, donc il faut les traiter de façon statistique, et l'affirmation « le nombre d'élèves au cours préparatoire est déterminant dans la réussite de l'apprentissage de la lecture » devient une *hypothèse statistique.*

De même, le médecin qui propose un nouveau traitement va devoir évaluer s'il est meilleur que le traitement habituel. Là encore, les effets d'un traitement varient d'un patient à l'autre, et il est nécessaire de vérifier l'hypothèse statistique que le traitement est meilleur. Enfin nous avons vu précédemment que l'analyse des séries doubles peut nous amener à tester l'hypothèse d'indépendance des deux variables.

b) Méthodologie des tests d'hypothèse :

Il nous faut, au départ, une hypothèse à tester. Nous la mettrons toujours sous une forme précise. Par exemple : «Les deux caractères sont indépendants ». Pour le médecin : « Le nouveau traitement fait le même effet que l'ancien », et non pas « est meilleur » de façon à pouvoir calculer des probabilités précises dépendant de notre hypothèse.

On appelle classiquement H_0 cette hypothèse testable. On lui associe généralement l'hypothèse alternative, notée H_1. On a donc :

H_0 : «Les deux caractères sont indépendants ».

Contre H_1 : «Les deux caractères sont dépendants ». (C'est à dire que les résultats de l'un donnent des indications sur les résultats possibles de l'autre).

Et pour le médecin :

H_0 : «Le nouveau traitement fait le même effet que l'ancien».

Contre H_1 : «Le nouveau traitement est meilleur que l'ancien». Notez que dans ce cas, on préférerait que le test statistique échoue, car c'est l'hypothèse H_1 qui nous plaît.

Il nous faut aussi savoir quelle confiance accorder au test. On donne en général un *risque*, plus précisément le risque de rejeter l'hypothèse, alors même qu'elle est vraie. Notons le r par la suite.

On détermine ensuite une procédure et un nombre (ou une série de nombres), à calculer au cours du test (à partir des résultats du test). Notons x ce nombre. Pour le test d'indépendance, on connaît ce nombre, c'est le χ^2 dont nous avons déjà parlé. Pour le médecin, il va falloir définir un

protocole de test, par exemple un test en double aveugle[1] et un critère numérique d'efficacité des traitement (taux de survie au bout de 2 mois, par exemple, ou durée moyenne d'alitement des malades) qui nous donne le nombre x cherché. Le choix de ce nombre x revient soit aux statisticiens (qui proposent des procédures toutes faites aux utilisateurs dans les situations simples), soit aux expérimentateurs, comme le médecin, seuls capables de définir ce qui est pertinent dans leur domaine.

A l'aide de ces éléments, nous déterminerons, en appliquant les techniques de calcul de probabilités, une zone de valeurs pour le nombre x, zone dans laquelle il se trouve avec la probabilité 1-r si l'hypothèse H_0 est vraie (donc H_0 étant vraie, le risque que x ne se trouve pas dans la zone est r).Cette zone est la zone d'acceptation du test. C'est la théorie statistique qui permet de la déterminer.

Le test théorique peut alors être énoncé (x est dans la zone d'acceptation : On accepte H_0 ; x n'est pas dans la zone : On rejette H_0).

Maintenant, il est possible de réaliser le test statistique. Et de conclure sur la situation (Par exemple les variables sont-elles indépendantes ? Ou bien le nouveau traitement est-il meilleur ?).

c) Un exemple :

On aimerait savoir si la pièce qu'on utilise pour jouer à pile ou face n'a pas une « préférence » pour l'un des deux côtés. Donc si c'est une pièce équilibrée, ou non. On prend :

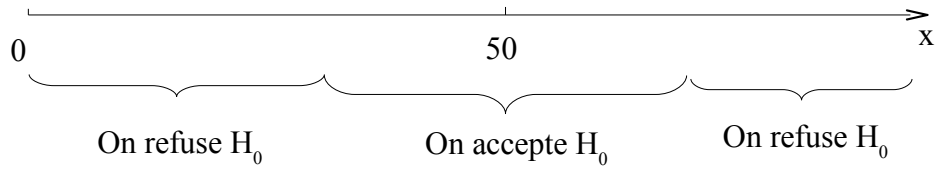
H_0 : «La pièce est équilibrée».

Contre H_1 : «La pièce sort plus souvent Pile ou sort plus souvent Face ».

Risque 5% (C'est un risque classique, assez faible pour être significatif, assez fort pour être utile).

Méthode de testage : On va lancer la pièce 100 fois et on comptera le nombre x de Piles.

Heuristique[2] : On s'attend à avoir à peu près le même nombre de Piles et de Faces si la pièce est équilibrée (exactement le même nombre est peu probable). Donc si x est proche de 50, on acceptera l'hypothèse H_0, et si x est loin de 50, on mettra en doute H_0 :



Il reste à trouver les bornes de la zone d'acceptation (La zone d'acceptation est un intervalle [a , b] avec a < 50 < b ; Comme on ne privilégie aucun des côtés de 50, on prendra 50 au milieu de l'intervalle [a , b]). Pour cela, on suppose vérifiée l'hypothèse H_0, donc que la pièce est équilibrée. Le nombre de Piles lors de 100 lancés d'une pièce équilibrée suit la loi Binomiale B(100; 0,5). On obtient les probabilités suivantes :

nombre	37	38	39	40	41	42	43	44	45	46	47	48	49
Probabilité (%)	0,27	0,45	0,71	1,08	1,59	2,23	3,01	3,9	4,85	5,8	6,66	7,35	7,8

50	51	52	53	54	55	56	57	58	59	60	61	62	63
7,96	7,8	7,35	6,66	5,8	4,85	3,9	3,01	2,23	1,59	1,08	0,71	0,45	0,27

Les autres probabilités sont nettement plus faibles, elles ne nous serviront pas.

1 Ni le malade, ni même le médecin ne savent si on donne l'ancien ou le nouveau traitement.

2 L'adjectif « euristique » ou « heuristique » (du grec heuriskêin, trouver, cf eureka), signifie « qui facilite la découverte, qui a une utilité dans la recherche (scientifique ou autre) ».

On remarque que la probabilité cumulée que le nombre de Piles soit entre 40 et 60 (compris) est de 96,5%. Donc il y a moins de 5% de risque que, si H_0 est vraie, le nombre de piles en 100 lancés soit inférieur à 40 ou supérieur à 60. On prendra comme zone d'acceptation [40 ; 60]. On aurait pu prendre [41 ; 59], avec un risque légèrement supérieur à 5 % (exactement 5,66 %).

Le test de la pièce est maintenant clair :

- Si $40 \leq x \leq 60$, on accepte l'hypothèse;
- Si $x < 40$ ou $x > 60$, on rejette l'hypothèse.

2) Problématique de la mise en oeuvre d'un test :

a) Peut-on avoir une certitude?

.

L'exemple du test de la pièce nous montre clairement qu'en général nous n'aurons pas de certitude : Si en lançant 100 fois la pièce on obtient 100 fois Pile, ce peut être :

- Soit parce que la pièce a deux faces Pile (un meilleur test aurait été de la regarder !);
- Soit qu'elle est trafiquée pour tomber très souvent sur Pile et qu'un heureux hasard a fait qu'on en a obtenu 100;
- Soit que la pièce est un peu favorable à Pile et qu'un très grand hasard a fait que pile est sorti systématiquement;
- Soit que la pièce est parfaitement équilibrée, mais qu'un hasard extraordinaire (1 chance sur 1 267 650 600 228 229 401 496 703 205 376) l'a fait tomber sur Pile 100 fois;
- Soit ...

Le hasard intervient presque systématiquement dans les tests statistiques, et c'est le rôle du risque d'en tenir compte.

b) Risques de première et de seconde espèce :

Lorsqu'on réalise le test effectif (Par exemple on lance 100 fois la pièce), quatre situations peuvent arriver, qu'on peut résumer dans le tableau suivant :

Réalité / Résultat du test	H_0 vraie	H_0 fausse
Réussi	OK	Erreur de deuxième espèce
Échoué	Erreur de première espèce	OK

Le risque de se tromper suite au test est de deux formes, deux espèces :

- Le test peut échouer, alors même que l'hypothèse est vraie. C'est le risque de première espèce, mesuré par le nombre r que l'on a choisi au départ. On connaît précisément ce risque. On peut le diminuer, mais en général, plus le risque r est faible, plus la zone d'acceptation est grande, et moins le test sera discriminant (voir l'exemple de la pièce, avec les risques 3,5 % et 5,66 %; Pour avoir un risque de moins de 1% il aurait fallu prendre [37 ; 53]).
- Le test peut réussir, alors même que l'hypothèse est fausse. Ce cas est bien plus flou, en général : Que veut dire « la pièce n'est pas équilibrée » ? Qu'il y a 2 fois plus de Piles ? Ou 55 % de Piles et 45 % de face ? Ou 49 % de Piles et 51 % de Faces ? Ou même qu'il y a 49,99 % de Piles ? Dans ces derniers cas, il sera difficile de faire la différence entre une pièce équilibrée et celle-ci. Cette *erreur de deuxième espèce* est donc nettement plus difficile à cerner. On n'a pas de valeur directe du risque de cette erreur, et même on peut considérer que dans certains cas elle n'est pas vraiment une erreur (jouer avec une pièce qui produit 49,99 % de Piles est plutôt satisfaisant, et la plupart des pièces sont sans doute plus « fausses » que cela).

Dans certains tests, on peut mesurer la qualité du test en mathématisant cette erreur de deuxième espèce. C'est ce qu'on appelle la *puissance* du test. Plus le test est puissant, plus on pourra diminuer le risque de seconde espèce.

c) Comment conclure ?

Lorsque le test échoue, on peut prendre le risque d'affirmer que l'hypothèse H_0 est fausse. On sait même quel risque on prend : r (si H_0 est fausse, on ne s'est pas trompé, si elle est vraie, on n'a pas eu de chance, car la probabilité de se tromper était petite, elle valait r).

Par exemple si on lance une pièce 100 fois, et qu'elle tombe sur Pile 68 fois, on peut prendre le risque d'affirmer qu'elle n'est pas équilibrée. Car on sait que pour une pièce équilibrée, il n'y a qu'une chance sur 20 (5 %) de tomber en dehors de l'intervalle [40 ; 60]. En fait, au risque 1 % encore, on peut affirmer que la pièce est déséquilibrée (La zone d'acceptation devient [37 ; 63]).

Lorsque le test réussit, il est difficile d'affirmer que H_0 est vraie, car elle peut être légèrement fausse, ou même assez fausse, mais par hasard on est tombé dans le domaine d'acceptation. On conclura simplement qu'on n'a pas de raison de dire que H_0 est fausse,

Par exemple, si la pièce est tombée 58 fois sur Pile, on n'a pas de raison de douter qu'elle est équilibrée. Et pourtant elle est tombée seulement 42 fois sur Face (16 de moins que de Piles).

d) Statistiquement significatif :

On rencontre souvent cette expression, ou plus simplement « significatif » suite à un test, un sondage, etc. Il s'agit tout simplement de dire qu'un test a échoué (C'est seulement dans ce cas qu'on peut dire quelque chose de nouveau). Par exemple un médecin qui a testé une nouvelle méthode dira qu'elle est « significativement plus efficace », parce qu'il a testé l'hypothèse H_0 : « les deux méthodes donnent le même résultat » contre l'hypothèse alternative H_1 : « la nouvelle méthode est meilleure », et que le test a échoué. Si le seuil de risque est 5 %, il dira plus précisément « la nouvelle méthode est significativement plus efficace au seuil de 5 % ».