Théorie statistique de la décision
(Tests d'hypothèse)

1 La décision statistique.

Dans la pratique, on est souvent amené à **prendre des décisions** diverses au sujet d'une population, à partir de l'information que donne **l'observation d'un échantillon.** On appellera de telles décisions des décisions statistiques. On voudra, par exemple, décider à partir d'un échantillon si un nouveau sérum est effectivement efficace pour guérir une maladie, si une méthode pédagogique est meilleure qu'une autre, si une pièce de monnaie est bien équilibrée, si deux variables dépendent linéairement l'une de l'autre, si l'hypothèse Gaussienne est raisonnable, si on peut modéliser telle variable par une loi de Poisson etc...

Dans le chapitre précédent, il était question d'estimation. On cherchait à avoir une valeur (même approximative) d'un paramètre pour lequel on n'avait aucune connaissance a priori. On ne prenait pas de décision !

Afin de prendre une décision pertinente concernant un problème, il est nécessaire de disposer d'une certaine connaissane de la distribution d'un (ou de plusieurs) caractère(s) sur une population. En plus de la complexité de certains cas, le fait de ne disposer que de données observées sur des échantillons de la population (mais pas sur la population toute entière) empêche généralement de connaître de manière explicite la nature du problème que l'on souhaite étudier. A partir des connaissances a priori et des observations partielles dont on dispose, on fait souvent des hypothèses (notées H_0 et H_1 dans la suite) sur la population correspondante (sa loi, la valeur d'un paramètre, une relation entre 2 variables, ...). L'objectif des méthodes de décision statistique est de permettre de prendre une décision concenant la validité de ces hypothèses à partir de nos observations en contrôlant le risque de faire erreur.

Exemple introductif :
Un sondage est effectué peu de temps avant une élection importante. L'objectif est de dire lequel des deux candidats a le plus convaincu les Français. On note H_0, l'hypothèse que le candidat n°1 reçoit plus de 50% d'intentions de vote, et H_1 l'hypothèse complémentaire que le candidat n°2 a convaincu plus de 50% des Français. On suppose que l'on observe sur le sondage que le candidat n°2 reçoit 53% d'intentions de vote. On peut donc penser qu'il a plus convaincu les Français. Cependant, il ne s'agit ici que d'une estimation de la proportion de Français qui pensent voter pour ce candidat. On ne peut pas quantifier le risque de se tromper en faisant l'hypothèse que le candidat n°2 a plus d'intentions de vote. On cherche à construire un outil statistique qui permette de conclure à partir des données du sondage en contrôlant le risque de se tromper.

Définition 1 *Un test d'hypothèse est une règle de décision qui, à partir d'un échantillon aléatoire, permet de choisir entre deux hypothèses : H_0 appelée hypothèse nulle et H_1 appelée hypothèse alternative.*

On peut considérer différents types de tests :

– Tests de conformité : consistent à vérifier si un échantillon peut être considéré comme issu d'une population dont certains paramètres (espérance, variance, fréquence,...) sont connus.
Ex. : le taux de glucose moyen mesuré dans un échantillon d'individus traités est-il conforme au taux de glucose moyen connu dans la population ?
– Tests d'homogénéité : consistent à comparer plusieurs populations dont les paramètres sont généralement inconnus.
Ex. : Y a-t-il une différence entre le taux de glucose moyen mesuré pour deux échantillons d'individus ayant reçu des traitements différents ?
– Tests d'ajustement : consistent à vérifier si la distribution observée sur un échantillon peut être considérée comme « proche » d'une distribution théorique fixée.
Ex : la distribution du nombre de personnes dans une file d'attente observé différents jours est-elle conforme à celle d'une loi exponentielle ?
– Tests d'indépendance : consistent à tester l'indépendance entre deux caractères discrets (ou discrétisés).
Ex : La distribution des guérisons de personnes malades est-elle indépendante du fait qu'elles ont reçu ou non le sérum ? De la quantité de sérum reçue ?

On construit le test de manière à **maitriser les risques** qu'on prend en prenant une décision. La démarche est shématisée par le tableau suivant :

		Décision	
		ne pas rejeter H_0	rejeter H_0
Réalité	H_0	Correct	erreur de première espèce ($P = \alpha$)
	H_1	erreur de deuxième espèce	Correct

Un test est bon si les probabilités des erreurs sont faibles. On note α le risque de première espèce et et β le risque de seconde espèce. C'est à dire $\alpha := P(\text{Rejeter } H_0 | H_0 \text{ est vraie})$, et $\beta := P(\text{Ne pas rejeter } H_0 | H_0 \text{ est vraie})$. En pratique, il est difficile de rendre ces deux quantités petites simultanément.

En général, on privilégie l'hypothèse H_0, qui désigne souvent les situations d'absence de changement par rapport à un statu quo, ou encore l'absence de différence entre des paramètres. Ainsi, on construit le test en fixant α **le risque de première espèce** (par défaut, $\alpha = 0.05$), et on limite donc le risque de rejeter l'hypothèse H_0 alors qu'elle est vraie.

Quand on fait un test, c'est le rejet de H_0 qui est significatif. **Le risque de seconde espèce est souvent noté β et $1 - \beta$ est appelée la puissance du test.**

Dans la plupart des cas, on formule donc l' hypothèse H_0 dans l'espoir de la rejeter et d'avoir ainsi un résultat significatif.

Par exemple, si on veut montrer qu'une pièce est déséquilibrée, on prendra l'hypothèse $H_0 : p = 0.5$.

Si un commerçant affirme que son chiffre d'affaire moyen est de 5000 frs et que l'inspecteur des impots pense que c'est plus, il fera un test en prenant l'hypotèse $H_0 : m = 5000$.

De même, si on veut tester si un procédé est meilleur qu'un autre, on prendra l'hypothèse H_0 : "il n'y a aucune différences entre les procédés."

Un premier exemple : Pour une population E, le poids des enfants a une distribution normale $\mathcal{N}(m, \sigma)$ avec $m = 3300$ et $\sigma = 500$. Un médecin s'intéresse au poids des enfants dont la mère fut atteinte d'une maladie M durant a grossesse (sous population F de E). Diverses observations

antérieures suggèrent que dans cette population F, les poids de naissance seraient distribués suivant une loi Normale $\mathcal{N}(m, \sigma)$ avec $m = 3200$ et $\sigma = 500$. Pour vérifier cette dernière hypothèse, il décide de réaliser une étude personnelle de la manière suivante :

- Déterminer le poids de naissance moyen $\overline{X}_n$ sur un échantillon de 100 enfants dont la mère fut atteinte par la maladie M pendant la grossesse.
- Choisir entre les deux hypothèses :
  H_0 : le poids de naissance suit une loi Normale $\mathcal{N}(3300, 500)$ dans F
  H_1 : le poids de naissance suit une loi Normale $\mathcal{N}(3200, 500)$ dans F
  en adoptant la règle suivante : rejet de H_0 si $\overline{x}_n < 3250$ et non rejet sinon.

1. Représenter graphiquement la distribution de $\overline{X}_n$ sous H_0 puis sous H_1.
2. Déterminer les risques de première et de seconde espèce de ce test.
3. En deça de quelle valeur seuil faut-il décider de rejeter H_0 pour limiter à 0.05 le risque de première espèce. Que vaut alors le risque de seconde espèce ?

Un peu plus de précisions concernant le vocabulaire utilisé et les différents types d'hypothèses :

- Un test est dit paramétrique si les hypothèses portent sur les paramètres d'une loi et non paramétrique dans le cas contraire.
- Dans le contexte paramétrique on suppose que la distribution d'un caractère dans la population totale se modélise par un type de loi particulier mais que l'on ne connaît pas certain(s) paramètre(s) θ de celle-ci (espérance, variance, proportion, ...).
  On prend souvent l'hypothèse nulle suivante $\mathcal{H}_0 : \theta = \theta_0$, où la valeur de θ_0 correspond à la norme. On fait alors un test de conformité sur la moyenne. On peut également considérer des hypothèses nulles différentes (voir l'exemple du comptable).
  La qualification d'un test dépend de l'hypothèse alternative :

Test bilatéral	Test unilatéral gauche	Test unilatéral droit
$H_1 : \theta \neq \theta_0$	$H_1 : \theta < \theta_0$	$H_1 : \theta > \theta_0$

  Ces différentes appellations sont également valables pour les tests d'homogénéité.

Voici maintenant la méthode générale pour effectuer un test : il y a 4 étapes, les étapes 1,2,3 s'effectuent indépendamment des résultats de l'expérience aléatoire tandis que l'étape 4 s'effectue aprés observation de ceux-ci.

1. Enoncer les hypothèses à tester H_0 et H_1 et annoncer le niveau α du test.
2. Fabriquer une statistique de test (appelée aussi variable de décision), $T = f(X_1, \cdots X_n)$, fonction des observations dont la valeur permettra de prendre une décision. Il est important que la loi de cette variable soit connue (ou corresponde approximativement à une loi connue) quand H_0 est vraie.
3. Déterminer la zone de rejet R_α, c'est à dire l'ensemble des valeurs de la statistique de test pour lesquelles on sera amené à rejeter l'hypothèse H_0. Ceci se fait en deux temps :
   - on détermine (selon le problème) la "forme" de R en fonction de la statistique T et d'une ou plusieurs valeurs critiques à ne pas dépasser.
   -on détermine la ou les valeurs critiques en fonction du niveau de test et de la loi de T sous l'hypothèse nulle en écrivant $\alpha = P_{H_0}(R)$.
   La région de rejet R_α d'un test dépend donc :
   - du risque de première espèce α,
   - de la loi (sous H_0) de la variable de décision,
   - de la qualification du test (unilatéral droit, gauche ou bilatéral).

4. Prendre la décision. On relève les observations $x_1, \ldots, x_n$ réalisées sur un échantillon de la population puis on calcule la valeur de $t = f(x_1, \ldots, x_n)$ obtenue à partir de ces données.
   - Si t appartient à la région de rejet R_α, on prend la décision de rejeter l'hypothèse H_0, avec un risque d'erreur α.
   - Si t n'appartient pas à la région de rejet R_α, on prend la décision de ne pas rejeter l'hypothèse H_0 au risque α (ce qui n'est pas équivalent à affirmer que l'hypothèse H_0 est vraie avec un risque d'erreur α !!!)

1.1 Tests paramétriques.

Il s'agit des tests portant sur la valeur d'un paramètre θ. On utilise le test plutôt qu'une estimation de ce paramètre quand on a une idée à priori de la valeur du paramètre et que l'on cherche à valider cette valeur.
Dans ce contexte, la statistique de test est généralement construite à partir d'un estimateur de θ.

On ne peut pas traiter tous les cas de tests paramétriques ! A chaque situation correspond un test qu'il faut construire étape par étape.
Qu'est ce qui change d'un test à l'autre ?
-La loi de la variable étudiée (loi normale,...,ou loi inconnue) ;
-le paramètre à tester (moyenne, variance,...) ;
-la taille de l'échantillon (grand i.e. $n > 30$ ou non) ;
(ces informations dictent l'estimateur à utiliser dans l'étape 2. On pourra se reporter aux tableaux donnés en annexe "Quel estimateur utiliser ?")
-la forme des hypothèses (simple ou composées, unilatérales ou bilatérales) ;
(ces informations dictent la forme de la zone de rejet en faisant appel au bon sens du praticien).

1.1.1 Tests sur les espérances

Partie A : Tests de conformité On s'intéresse à une population sur laquelle on étudie la variable X d'espérance m et de variance σ^2.
On dispose d'un échantillon aléatoire et simple prélevé sur cette population et on cherche à tester si la moyenne m de la population est conforme à une norme m_0. On peut également prendre d'autres hypothèses nulles comme par exemple $m \leq m_0$ ou $m \geq m_0$.
En général les hypothèses à tester sont :

$$H_0 : \text{" } m = m_0 \text{ " contre } H_1 : \text{"} m \neq m_0\text{"}, \text{"} m < m_0\text{" ou "} m > m_0\text{"}.$$

Soit $\overline{X}$ l' estimateur de m pour l'échantillon $(X_1, \ldots, X_n) \overset{i.i.d.}{\sim} X$.

1. On veut tester l'hypothèse H_0 dite de conformité, selon laquelle il n'y a pas de différence entre m et la norme m_0. L'hypothèse alternative étant unilatérale ou bilatérale selon le contexte. On se fixe un risque d'erreur de première espèce α (en général $\alpha = 0.05$).

2. Pour construire notre test on remarqe que la variable $\overline{X} - m_0$ prend de petites valeurs sous l'hypothèse nulle et a tendance à être grande sous l'alternative. On utilise ensuite que si X est de loi

normale ou si la taille de l'échantillon est grande la statistique de test $T = \frac{\overline{X}-m_0}{\sqrt{\frac{\sigma^2}{n}}}$ suit approximativement la loi $N(0,1)$.(Quand on ne connait pas les variances on utilise leurs estimateurs respectifs.)

3. La zone de rejet R_α sera du type $|T| > \lambda_\alpha$ ou $T > \lambda_\alpha$ ou $T < \lambda_\alpha$ selon les cas (bilatéral, unilatéral droit, unilatéral gauche). Le seuil λ_α étant calculé de façon à ce que $P_{H_0}(T \in R_\alpha) = \alpha$.

4. On conclut ensuite après observation.

Dans le cas d'un petit échantillon de loi Normale pour lesquel la variances est inconnue, on la remplace par l'estimateur de la variance S^2_{n-1} et la statistique T obtenue suit une loi de Student de paramètre $n-1$.

Remarque : Lorsque l'hypothèse nulle n'est pas simple (on dit qu'elle est composée), c'est à dire que H_0 n'est pas "$m = m_0$", il faut choisir R_α tel que pour tout m vérifiant H_0 on ait bien un risque inférieur à α (voir l'exemple qui suit).
Exemple : lors d'un contrôle fiscal, on s'intéresse au chiffre d'affaire quotidien d'un commerçant et on admet que c'est une variable aléatoire de loi Normale d'écart-type $\sigma = 500F$. (Des études statistiques ont été faites précédemment sur d'autres commerçants, qui nous permettent de faire cette hypothèse). Le commerçant affirme que son chiffre d'affaire quotidien est de 5000 Frs, tandis que le contrôleur pense plutôt à un chiffre moyen de 5200 Frs et décide donc de faire un test.

1. Il prend par exemple pour H_0 l'hypothèse $m = 5000$, et pour H_1 l'hypothèse $m > 5000$. Souvent on prend un niveau de risque de 0.05. S'il est sévère, il prendra un risque plus grand pour rejeter plus facilement.
2. Il décide de prendre pour échantillon les comptes du mois précédent et considère la variable X_n donnant le chiffre d'affaire pour le nième jour, et la statistique $\overline{X}_{25} = \dfrac{X_1 + \cdots + X_{25}}{25}$. Sous l'hypothèse H_0, la loi de $\overline{X}_{25}$ est $N(5000, \dfrac{500}{5})$.
3. On rejettera l'hypothèse H_0, si $\overline{X}_{25}$ est "trop grand ", c'est à dire :
   -$R = \left(\overline{X}_{25} > c\right)$
   -$P_{H_0}\left(\overline{X}_{25} > c\right) \leq 0.05$, c'est à dire que $P\left(U > \dfrac{c-5000}{100}\right) \leq 0.05$, où U suit la loi $N(0,1)$.
   On choisit finalement $c = 5000 + (100 * 1.64) = 5164Frs$.
4. On consulte ensuite les comptes sur l'échantillon de 25 jours et on trouve 5200 Frs. C'est plus grand que le seuil c fixé par l'inspecteur des impots, qui rejette donc l'hypothèse H_0.

Dans l'exemple que l'on vient de voir, les hypothèses portent sur la valeur d'un paramètre, on parle de test paramétrique. Par contre, on peut remarquer que pour la construction de ce test, on suppose que le chiffre d'affaire suit une loi Gaussienne. Ce genre d'hypothèse sur les lois des phénomènes peut également être validé comme on va le voir par des tests statistiques, mais il s'agit cette fois de tests non paramétriques. Enfin, l'hypothèse que cherche à rejeter le contrôleur pourrait être $\mathcal{H}_0 : m \leq 5000$ correspondant au fait que le commerçant n'a pas sous évalué son chiffre d'affaire. Dans ce cas, on voit que sous l'hypothèse nulle on a seulement $\overline{X}_{25} \sim \mathcal{N}(m, 500/5)$ avec $m \leq 5000$. Dans ce cas on a seulement $Z_{25} := (\overline{X}_{25} + (5000 - m)) \sim \mathcal{N}(5000, 500/5)$. Mais comme sous $\mathcal{H}_0$ on a $5000 - m \geq 0$, on voit que l'on a bien

$$P_{\mathcal{H}_0}(X_{25} > c) = P_{\mathcal{H}_0}(X_{25}+5000-m > c+5000-m) = P(Z_{25} > c+5000-m) = P(U > c+5000-m) \leq P(U > c) = c$$

Partie B : Tests d'homogénéité

On s'intéresse à deux populations indépendantes sur lesquelles on étudie les variables respectives X_1 et X_2 d'espérance m_1 et de variance σ_1^2 pour X_1, et d'espérance m_2 et de variance σ_2^2 pour X_2. On dispose de deux échantillons aléatoires et simples de tailles respectives n_1 et n_2 provenant de chacune des populations.

Soient $\overline{X}_1$ et $\overline{X}_2$ les estimateurs respectifs de la moyenne pour ces deux échantillons.

1. On veut tester l'hypothèse H_0 dite d'homogénéité, selon laquelle il n'y a pas de différence entre m_1 et m_2. L'hypothèse alternative étant unilatéral ou bilatéral selon le contexte.

$$H_0 : \text{" } m_1 = m_2 \text{ " contre } H_1 : \text{"}m_1 \neq m_2\text{","}m_1 < m_2\text{" ou "}m_1 > m_2\text{".}$$

2. Pour cela, on regarde l'écart $\overline{X}_1 - \overline{X}_2$ et si il s'agit de grands échantillons indépendants, on utilise la statistique $T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ qui suit approximativement la loi $\mathcal{N}(0,1)$.(Quand on ne connait pas les variances on utilise leurs estimateurs respectifs.)

3. La zone de rejet R_α sera du type $|T| > \lambda_\alpha$ ou $T < \lambda_\alpha$ ou $T > \lambda_\alpha$ selon les cas. Le seuil λ_α étant calculé de façon à ce que $P_{H_0}(T \in R_\alpha) = \alpha$.

4. On conclut ensuite après observation.

Si les écarts-types sont supposés inconnus et identiques dans les deux groupes, on peut estimer cette valeur commune à partir des deux échantillons (voir annexe). Dans le cas de petits échantillons de loi Normale et que les variances sont inconnues, on remplace les variances par leurs estimateurs $S^2_{n_1-1}$ et $S^2_{n_2-1}$. La statistique T obtenue suit alors une loi de Student de paramètre $n_1 + n_2 - 2$.

Exemple 1 : lors d'un examen d'orthographe dans la classe élémentaire, la note moyenne de 40 garçons est 74 avec un écart-type de 8 et celles de 50 filles est 78 avec un écart-type de 7. Tester au niveau 0.05, puis 0.01 l'hypothèse que les filles sont meilleures en orthographe que les garçons.

1. On prend pour H_0 "$m_1 = m_2$", $\alpha = 0.05$. .
2. La statistique $T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{8^2}{40} + \frac{7^2}{50}}}$ prend la valeur -2.49.
3. La zone de rejet est unilatère $T < \lambda_\alpha$.
4. On lit dans la table ou avec R $\lambda_{0.05} = -2.3263$ et $\lambda_{0.01} = -16449$.
5. On rejette donc H_0 avec un probabilité de 0.05 de se tromper. Le résultat est significatif au risque 0.05. Ce n'est plus le cas pour un risque de 0.01.

Remarque : on pourrait vouloir prendre $\mathcal{H}_\prime : m_1 \geq m_2$. En fait en faisant le même raisonnement que pour l'exemple du commerçant, on peut montrer que la valeur seuil reste la même.

Partie C : Exercices supplémentaires et Annales

1. Les charges de rupture de cables produits par une fabrique ont une valeur moyenne de 1800kg. On affirme que la charge de rupture peut être augmentée par une nouvelle technique du procédé de fabrication. Pour tester cette affirmation, on a testé unéchantillon de 50 cables et l'on a trouvé une charge de rupture moyenne de 1850 kg et un écart-type de 100kg. Peut-on admettre cette affirmation à un niveau de 0.01 ?
2. Une machine a produit dans le passé des rondelles d'épaisseur 0,05 cm. On suppose que l'épaisseur suit une loi Normale $\mathcal{N}(0.05, \sigma)$. Pour déterminer si la machine est encore en bon étatde marche, on prélève 10 rondelles et on trouve une moyenne empirique de 0.053 cm et un écart-type empirique de 0.003 cm. Tester l'hypothèse qui affirme que la machine est en bon état de marche au seuil 0.05.
3. Supposons que le fabricant de voitures "Crapaud&Cie" doive effectuer des tests de collision sur ses voitures afin de déterminer le coût moyen de réparation correspondant à une collisition frontale à 16.5 km/h. Etant donné l'ampleur des dégats on ne fait que cinq tests ! Par ailleurs, on connait ce genre de phénomènes et on sait que le coût est de loi gaussienne. On obtient les résultats suivants :

$$150, 400, 720, 500, 930 \ Euros.$$

   - Trouver un intervalle de confiance pour le coût moyen avec un niveau de sécurité de 0.95.
   - Finalement le constructeur annonce un coût moyen de 530 Euros. Trois nouveaux essais sont effectués pour tester c coût moyen. On trouve les valeurs

$$490, 630, 520 \ Euros.$$

     Doit-on rejeter l'hypothèse au risques 0.01 et 0.05 ?
   - Comment construire le test si il annonce un coût inférieur à 530 Euros ?
4. On note μ la durée de vie moyenne des ampoules fluorescentes produite par une usine. L'entreprise affirme que $\mu = 1600$. On veut tester si les ampoules produites sont bien conformes à ce qu'annonce l'usine, c'est à dire tester l'hypothèse ulle "$\mu = 1600$" contre l'alternative "$\mu \neq 1600$" avec un risque $\alpha = 0.01$. On met en test un échantillon de 100 ampoules et on trouve une durée de vie moyenne de 1570 heures et un écart-type empirique de 120 heures. Construire un test statistique. Quelle est la zone de rejet ? Que peut-on en conclure ?
5. On a étudié le temps moyen hebdomadaire (mesuré en heures) consacré au travail domestique par des femmes mariées "sans profession" des zones rurales et des villes. Les résultats numériques des observations sont les suivants :

	nombre	moyenne	$\sum_{i=1}^{n}(x_i - m)^2$
Classe 1 (Zone rurale)	120	44.8	34680
Classe 2 (Zone citadine)	82	41.2	20992

   Peut on conclure à une différence significative entre les comportements de deux classes ?
6. Une firme d'aliments pour animaux affirme avoir mis au points un nouvel aliment permettant un engraissage plus rapide des porcs. Un centre d'étude désirant s'assurer des performences de ce nouvel aliment, le teste sur deux échantillons de 18 porcs. On suppose que le poids des porcs est distribué seon un loi gaussienne et on admet que pour les deux lots les variances sont égales à 225. Les deux lots de porcs ont atteint au même moment les poids moyens

respectifs de 110Kg pour le premier lot nourri avec le nouvel aliment, et 100 Kg pour le second lot nourri de façon traditionnelle. Ces résultats confirment-ils l'efficacité de ce nouvel aliment ?

7. On a relevé pendant 30 jours ouvrables consécutifs les nombres quotdiens d'actes de délinquence commis dans un centre commercial. Les résultats sont reportés dans le tableau suivant (n_k désigne le nombre de jours et où k actes de délinquence ont été commis.

k	0	1	2	3	4	5
n_k	4	10	11	1	1	3

On suppose que le nombre X d'actes de délinquence commis chaque jour dans le centre commercial suit une loi de Poisson de paramètre λ inconnu. D'après les statistiques du ministère de l'intérieur il se produit en moyenne 1.5 actes de délinquance par jours dans chacun des centres commerciaux de la région parisienne. Mais ce chiffre est considéré comme sous-estimé par une association de comerçants. Nous allons utiliser la théorie des tests pour trancher cette question.

– En remarquant que pour une loi de Poisson, le paramètre λ est égal à la moyenne et à la variance, et que la variable centrée réduite

$$\frac{\overline{X} - \lambda}{\sqrt{\frac{\lambda}{n}}}$$

suit approximativeent une loi $\mathcal{N}(0, 1)$, construire le test correspondant à un risque de 0.05.

– Quelles en sont les conclusions ?

8. La durée du processus d'atterrissage d'un avion est le temps, mesuré en secondes, qui s'écoule entre la prise en charge par la tour de contrôle t l'imobilisation totale de l'appareil sur la piste.
Afin de faire face au flux croissant des avions se posant à l'aéroport de Toulouse-Blagnac, une restructuration des services de la tour de contrôle visant à diminuer la durée du processus d'atterrissage est réalisée.
Auparavant, cette durée s'élevait en moyenne à 160 secondes. A la suite de la retructuration, une enquête, effectuée sur 1000 avions a donné les résultats suivants

classe	[60 ;120]	]120 ;140]	]140 ;180]	]180 ;200]	]200 ;260]
effectifs	112	176	461	157	94

Faites un test de risque de première espèce 0.01 pour savoir si la durée d'attérissage a diminué significativement.

1.1.2 Tests sur les proportions

Partie A. Tests de conformité sur les proportions.

Test de conformité d'une proportion On s'intéresse à une population sur laquelle on étudie un caractère C (présence ou absence) ; on note p la proportion d'individus présentant ce caractère. On dispose d'un échantillon aléatoire et simple prélevé sur la population. On note enfin N le nombre d'individus de l'échantillon présentant le caractère C et $F_n = \frac{N}{n}$ l'estimateur de p.

Hypothèses à tester :

$$H_0 : p = p_0 \text{ contre } H_1 : p \neq p_0,\ p < p_0,\ \text{ ou } p > p_0.$$

On considère la statistique de test

$$T = \frac{F_n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

Pour de grands échantillons ($n > 30$ et $np_0(1-p_0) > 12$), cette variable aléatoire est approximativement de loi $\mathcal{N}(0,1)$ sous l'hypothèse nulle, c'est à dire lorsque $p = p_0$. On suit ensuite la même démarche que celle décrite plus haut pour le test de conformité des espérances en construisant la zone de rejet en fonction du type d'hypothèse alternative (bilatérale, unilatérale gauche, ou unilatérale droite) à partir des quantiles de cette loi. Avec R cela se fait avec la commande **prop.test**

Remarque : Dans le cas de petits échantillons on utilise la statistique de test N et les quantiles de la loi binomiale $Bin(n, p_0)$. Avec R cela se fait avec la commande **binom.test**

Exemple.

Considérons l'exemple déja cité concernant les intentions de votes observées pour deux candidats avant le second tour d'une élection. On note p la proportion de Français convaincus par le candidat n°2. On rappelle que lors d'un sondage récent effectué sur un échantillon représentatif de 1000 individus, 53% des personnes interrogées ont dit qu'elles pensaient voter pour le candidat n°2. On souhaite tester l'hypothèse H_0 que le candidat n°2 a plus d'intentions de vote (i.e. $p \leq 0.5$) contre l'aternative (i.e. $p > 0.5$)

$$H_0 \quad p \leq 0.5 \text{ ou } H_1 \quad p > 0.5$$

Ainsi, si on rejette H_0 on pourra significativement affirmer que le candidat n°2 a plus d'intentions de vote en se faveur.

1. On teste $H_0 \quad p \leq p_0$ contre $H_1 \quad p > p_0$, où $p_0 = 0.5$, avec un niveau de risque $\alpha = 0.05$.
2. On sait que la fréquence empirique F_n estime p et sous l'hypothèse H_0, F_n sera "petit" (proche de $p \leq 0.5$). Par ailleurs, n est assez grand pour que le théorème de la limite centrée soit vérifiée, c'est à dire que $\dfrac{F_n - p}{\sqrt{\frac{p(1-p)}{n}}}$ suit une loi normale centrée réduite quand p est la "vraie" côte de popularité.
3. On propose la région de rejet : $R = \{F_n > c\}$ où la valeur critique c est calculée de telle sorte que
$$P_{p_0}(F_n > c) = \alpha = 0.05.$$
Il faut remarquer ici que H_0 n'est pas "$p = p_0$" mais que si c vérifie la relation ci-dessus alors pour tout $p \leq p_0$, la probabilité $P_p(F_n > c)$ sera plus petite que $\alpha = 0.05$, car elle est

croissante sur $[0, p_0]$, si $p_0 \leq 0.5$ (ceci grâce à la croissance de $p(1-p)$ sur $[0, p_0]$) ce qui est le cas ici.
Il est alors simple par lecture dans une table de loi $N(0,1)$, de trouver

$$c = 0.50 + \left(1.64 * \sqrt{\frac{0.47 * 0.53}{1000}}\right) = 0.525884$$

4. Le sondage ayant donné $f = 0.53$, on est amené à rejeter H_0, c'est-à-dire à conclure que le candidat 2 a plus d'intention de vote avec un risque de 5% de se tromper.

Partie B. Test d'homogénéité sur les proportions (grands échantillons).

Test d'homogénéité de deux proportions On s'intéresse à deux populations indépendantes sur lesquelles on étudie un caractère C (présence ou absence) ; on note p_1 et p_2 les proportions respectives d'individus présentant ce caractère dans chacune des deux populations.
Hypothèses à tester :

$$H_0 : p_1 = p_2 \text{ contre } H_1 : p_1 \neq p_2, p_1 < p_2, \text{ ou } p_1 > p_2.$$

On dispose de deux échantillons aléatoires et simples de tailles respectives n_1 et n_2 provenant de chacune des populations. Soient F_1 et F_2 les estimateurs respectifs de la proportion d'individus pour lesquels le caractère C est observé obtenus à partir de deux échantillons de taille n_1 et n_2 (provenant de populations où les proportions sont p_1 et p_2 respectivement).

1. On veut tester l'hypothèse H_0 dite d'homogénéité, selon laquelle il n'y a pas de différence entre p_1 et p_2. L'hypothèse alternative étant unilatérale ou bilatéral selon le contexte.

2. Pour cela, on regarde l'écart $F_1 - F_2$ et si il s'agit de grands échantillons (voir un exemple de conditions en annexe) indépendants, on utilise la statistique $T = \frac{F_1 - F_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$ qui suit approximativement sous H_0 la loi $N(0,1)$. Ensuite on approxime $p = p_1 = p_2$ par la proportion $\frac{n_1 * f_1 + n_2 * f_2}{n_1 + n_2}$.

3. La zone de rejet R_α sera du type $|T| > \lambda_\alpha$ ou $T < \lambda_\alpha$ ou $T > \lambda_\alpha$ selon les cas. Le seuil λ_α étant calculé de façon à ce que $P(R_\alpha) = \alpha$.

4. On conclut ensuite après observation.

Avec R, il suffit d'utiliser la commande **prop.test**. Cependant il faut mettre en argument une matrice contenant la table de contingence. Pour créer cette matrice on peut utliser la commande :

$$A < -matrix(c(p_1, p_2, n_1 - p1, n_2 - p_2), byrow = T, ncol = 2)$$

Remarque : Dans le cas d'échantillons de petite taille on utilise le test de Fisher **fisher.test**

Exemple 2 : deux groupe A et B se composent chacun de 100 personnes atteintes d'une même maladie. On administre du sérum au groupe A mais pas au groupe B (que l'on appelle groupe

de contrôle) ; ensuite les deux groupes sont traités de la même façon. Finalement, 75 malades du groupe A et 65 du groupe B ont guéri. Tester au niveau $0.01, 0.05, 0.1$ l'hypothèse que le sérum est une aide efficace dans la guérison de la maladie.

On prend la statistique $T = \frac{F_1 - F_2}{\sqrt{p(1-p)(\frac{1}{100}+\frac{1}{100})}}$ où $p = \frac{75+65}{200} = 0.3$.

On rejette H_0 si $T > \lambda$ et les valeurs critiques aux risques respectifs sont $2.33, 1.645, 1.28$. Donc c'est seulement en prenant le risque suffisamment grand de 0.1 qu'on peut rejeter H_0 et conclure que le sérum est efficace.

On peut remarquer que la conclusion dépend du risque d'erreur que l'on accepte de prendre. Si les résultats sont dus au hasard et si on a conclu qu'ils étaient dus au sérum (erreur de première espèce), on peut administrer le sérum à des groupes plus importants de malades pour s'apercevoir ensuite que le sérum n'a en réalité aucun effet bénéfique. C'est un risque que l'on n'a pas toujours envie de prendre (effets secondaires, coût du produit....)

D'un autre coté, on aurait pu conclure que le sérum n'apporte aucune aide alors qu'il est essentiel pour la guérison (erreur de 2ème espèce). Une telle conclusion peut-être dangereuse surtout si des vies humaines sont en jeu.

Remarque : ce problème peut se traiter aussi par un test d'indépendance ou l'hypothèse $\mathcal{H}_0$ est que le sérum et la guérison sont indépendants contre $\mathcal{H}_1$ qu'ils ne sont pas indépendants. Attention, le fait de rejeter $\mathcal{H}_0$ avec ce test ne dit pas si l'effet du sérum est positif ou non.

Exercices supplémentaires :

1. Lors d'une élection nationale, un candidat a recueilli 8.82% des voix. Le même candidat se présente aux élections suivantes. Une semaine avantces élections on procède à un sondage sur un échantillon de 400 électeurs jugés représentatifs, ayant déclarer vouloir voter. Le pourcentag d'intentions de vote en faveur de ce candidat sur cet échantillon est alors de 13%.
   - Au vu du résultat obtenu sur cet échantillon, peut-on rejeter avec un risque bilatéral total de 1% de se tromper, l'hypothèse que sur le plan national, le pourcentage d'électeurs favorables à ce candidat est resté de 8.82% ?
   - En vous basant sur le résultat obtenu sur cet échantillon, donner l'intervalle de confiance à 98% puis à 99% dans lequel on peut penser que se trouve le nouveau pourcentage d'électeurs voulant voter pour ce candidat.
2. Dans une expérience de perception extra sensorielle, on demande à un sujet isolé dans une pièce de dire la couleur (rouge ou bleue) d'une carte choisie parmi une pile de 50 cartes bien battues par un expérimentateur placé dans une autre pièce. Le sujet ne connait pas le nombre de cartes bleues ou rougesde la pile. En supposant que le sujet identifie correctement 32 cartes, détermiiner si les résultats sont significatis au seuil de 0.05 et 0.01%.
3. Le fabricant d'un médicament breveté affirmait qu'il était efficace à 90% pour guérir une allergie en 8 heures. Dans un échantillon de 200 personnes atteintes par cette allergie, on en a guéri 160 par le médicament. Déteriner si l'affirmation du fabricant est légitime.
4. On reprend le problème d'élection, mais on veut faire une analyse plus fine concernant les intentions de vote dans deux départements phares. Pour ceci, on interroge un échantillon dans chaque département, l'un de 300 électeurs et l'autre de 200 électeurs. Sur le premier échantillon on trouve 18% d'électeurs se déclarant prêts à voter pour ce candidat et 12% sur le second.Au vu des résultats, peut-on rejeter avec un risque bilatéral total de 5% de se tromper, l'hypothèse que les deux populations d'électeurs sont hmogènes du point de vue de leur intention de vote en faveur de ce candidat ?

5. Sur un lot de 700 boulons soumis à un test de rupture, 300 d'entre eux ont résisté à ce test. Sur un second lot de 225 boulons soumis au même test, 125 ont résitsé. Peut-on admettre au niveau de risque de0.05 puis 0.02 que les deux lots proviennent de a même fabrication ?

1.1.3 Tests sur la variance

Partie A : Test de conformité d'une variance

On s'intéresse à une population sur laquelle on étudie la variable X d'espérance m et de variance σ^2. On dispose d'un échantillon aléatoire et simple prélevé sur cette population. Hypothèses à tester :

$$H_0 : \sigma^2 = \sigma_0^2 \text{ contre } H_1 : \sigma^2 \neq \sigma_0^2, \sigma^2 < \sigma_0^2, \text{ ou } \sigma^2 > \sigma_0^2.$$

On a besoin de l'hypothèse que X est de loi normale. On utilise en général la statistique de test

$$\frac{(n-1)S_c^2}{\sigma_0^2}$$

qui suit une loi du $\chi^2(n-1)$ sous l'hypothèse nulle si la moyenne m est inconnue et $S_c = S_{n-1}$ ou une loi du $\chi^2(n)$ si m est connue et $S_c = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$.

Partie B : Test d'homogénéité de deux variances

On s'intéresse à deux populations indépendantes sur lesquelles on étudie les variables respectives X1 et X2 d'espérance m1 et de variance σ_1 pour X1, et d'espérance m2 et de variance σ_2 pour X2. On dispose de deux échantillons aléatoires et simples de tailles respectives n1 et n2 provenant de chacune des populations. Hypothèses à tester :

$$H_0 : \text{“}\sigma_1 = \sigma_2'' \text{ contre } H_1 : \text{“}\sigma_1 \neq \sigma_2'', \sigma_1 < \sigma_2'', \text{ ou “}\sigma_1 > \sigma_2''.$$

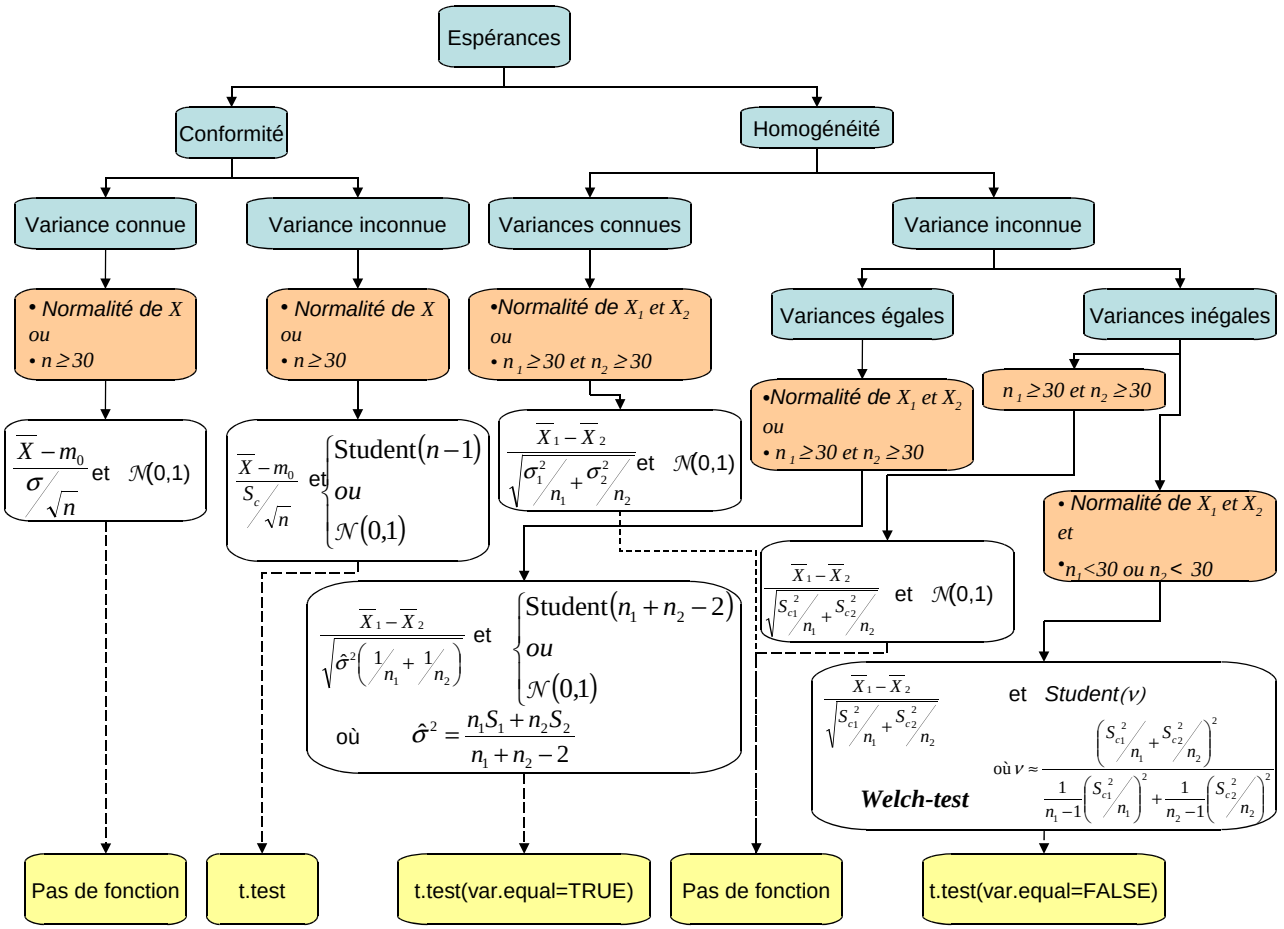
On a besoin de l'hypothèse que X est de loi normale. On utilise en général la statistique de test

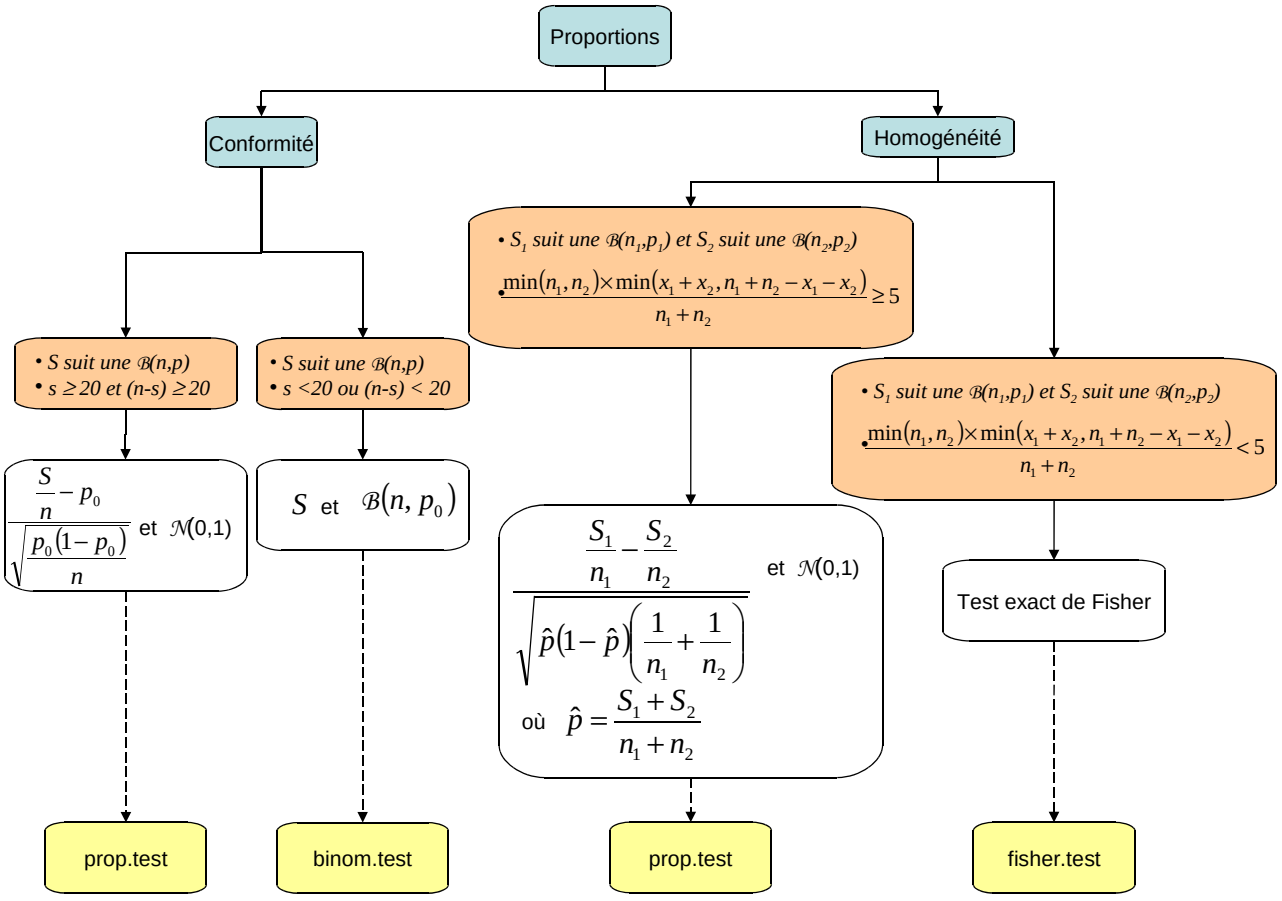
$$\frac{S_{c1}^2}{S_{c2}^2}$$

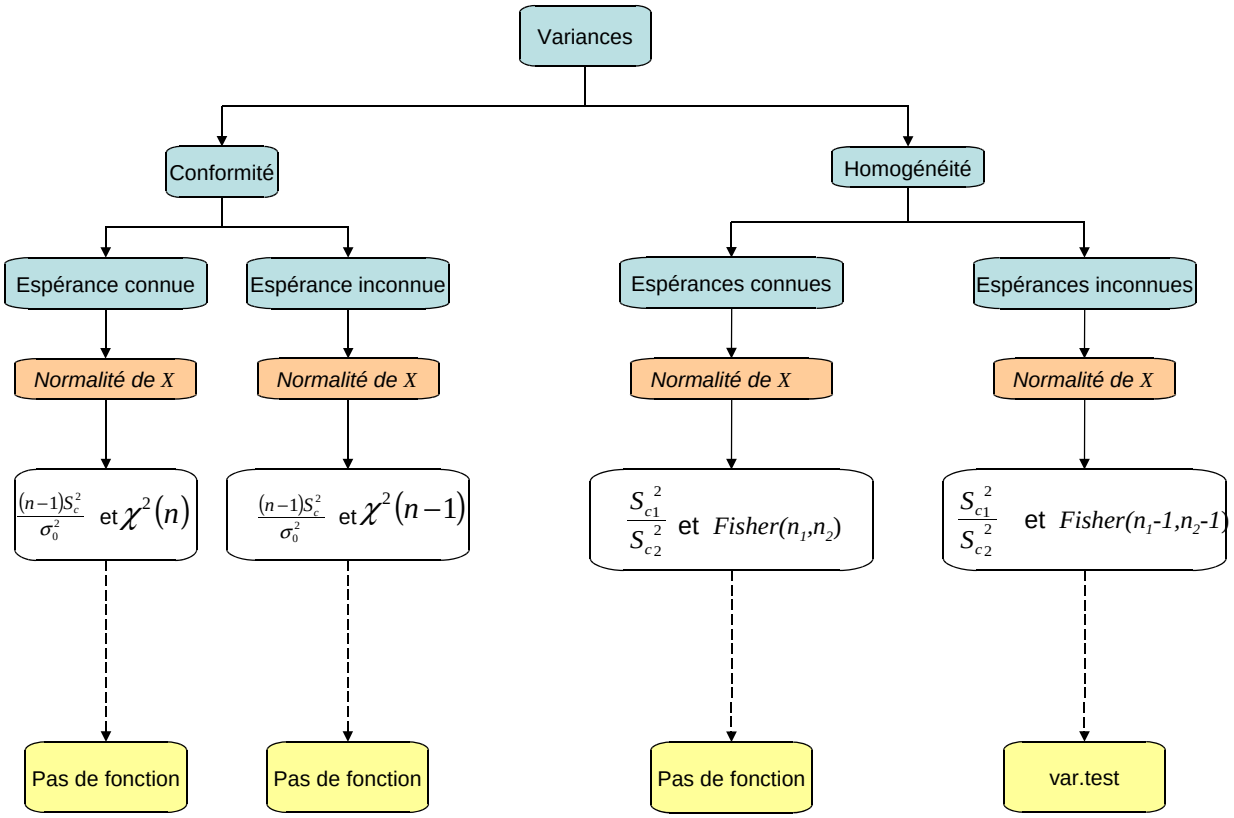
qui suit une loi de $Fisher(n_1 - 1, n_2 - 1)$ sous l'hypothèse nulle si les moyennes m1 et m2 sont inconnues et $S_{ci} = S_{n_i - 1}$ ou une loi de $Fisher(n_1, n_2)$ si m_1 et m_2 sont connues et $S_{ci} = \frac{1}{n_i}\sum_{j=1}^{n}(X_j - m_i)^2$.

Exercice : Créer deux échantillons de tailles respectives 100 et 500 de variables de lois respectives $\mathcal{N}(0,1)$ et $\mathcal{N}(1, 1.6)$. Tester successivement au risque de 5%

- l'hypothèse que l'écart-type du premier échantillon est 1 (en sachant que m=0)
- l'hypothèse que l'écart-type du premier échantillon est 1.6 (en ignorant que m=1)
- l'hypothèse que les deux échantillons ont même écart-type en sachant puis en ignorant la valeur des moyennes.


Espérances
Conformité
Homogénéité
Variance connue
Variance inconnue
Variances connues
Variance inconnue
• Normalité de X ou • n ≥ 30
• Normalité de X ou • n ≥ 30
•Normalité de X1 et X2 ou • n1 ≥ 30 et n2 ≥ 30
Variances égales
Variances inégales
•Normalité de X1 et X2 ou • n1 ≥ 30 et n2 ≥ 30
n1 ≥ 30 et n2 ≥ 30
• Normalité de X1 et X2 et •n1<30 ou n2< 30
et N(0,1)
et Student(n−1) ou N(0,1)
et N(0,1)
et N(0,1)
et Student(n1+n2−2) ou N(0,1)
où σ̂² = (n1S1 + n2S2)/(n1+n2−2)
et Student(ν)
où ν ≈
Welch-test
Pas de fonction
t.test
t.test(var.equal=TRUE)
Pas de fonction
t.test(var.equal=FALSE)


Proportions
Conformité
Homogénéité
• S suit une ℬ(n,p)
• s ≥ 20 et (n-s) ≥ 20
• S suit une ℬ(n,p)
• s <20 ou (n-s) < 20
• S1 suit une ℬ(n1,p1) et S2 suit une ℬ(n2,p2)
• min(n1,n2)×min(x1+x2,n1+n2−x1−x2) / (n1+n2) ≥ 5
• S1 suit une ℬ(n1,p1) et S2 suit une ℬ(n2,p2)
• min(n1,n2)×min(x1+x2,n1+n2−x1−x2) / (n1+n2) < 5
(S/n − p0) / √(p0(1−p0)/n) et 𝒩(0,1)
S et ℬ(n, p0)
(S1/n1 − S2/n2) / √(p̂(1−p̂)(1/n1 + 1/n2)) et 𝒩(0,1)
où p̂ = (S1+S2)/(n1+n2)
Test exact de Fisher
prop.test
binom.test
prop.test
fisher.test



tests très sensibles à la non normalité

1.1.4 Test d'ajustement.

Le test d'ajustement permet de décider si une variable suit une loi P_0 discrète ou discrétisée. On dispose toujours d'un n-échantillon de la variable X auquel on fait correspondre l'effectif observé n_i de la classe a_i pour $i = 1 \cdots r$ (r étant le nombre de classes obtenues par coupure) ; c'est ce qu'on appelle le tri à plat de la variable observée.

Lorsque la loi de la variable que l'on étudie est discrète on prend comme classes les différentes valeurs (modalités) de la variable et on dénombre combien de fois chaque valeur a été obtenue dans les données. Lorsque la variable a une loi de type continu, on fixe des classes et on dénombre le nombre de données qui se trouvent dans chacune de ces classes.

Si la loi de X est effectivement P_0, on devrait avoir approximativement égalité entre "l'effectif théorique" (en moyenne) $np_i = n * P_0(a_i)$ et l'effectif observé n_i. S'il est nécessaire pour pouvoir appliquer notre test d'avoir un effectif théorique au moins égal à 5 dans chacune des classes, il est souvent pertinent de prendre des classes d'amplitude assez réduite afin de comparer de manière assez fine les deux distributions (théorique supposée P_0 et empirique : "observations").

Construisons notre test en utilisant les étapes décrites dans l'introduction.

1. $H_0 =$ "les effectifs théoriques pour les r classes sont respectivement" $P_0(X \in a_i), \forall i = 1 \cdots r$.

   $$H_1 = \text{"X ne suit pas la loi } P_0\text{."}$$

   On se donne un niveau de test α
2. On introduit la statistique

   $$D^2 = \sum \frac{(N_i - nP_0(X \in a_i))^2}{nP_0(X \in a_i)}$$

   où N_i est la variable aléatoire dont l'observation est n_i. Cette statistique mesure la distance entre la loi observée et la loi théorique. On l'a choisie car d'après un théorème de probabilité (basé entre autre sur le théorème de la limite centrée), on a : pour n assez grand, si la loi de X est P_0, D^2 suit approximativement une loi de khi-deux à $r-1$ degés de liberté.
3. On rejettera l'hypothèse si l'écart est "trop grand", donc si $D^2 > \lambda$, où λ est choisi de façon à ce que $P(D^2 > \lambda | H_0) \leq \alpha$. On lit λ dans la table du khi-deux.
4. On calcule d^2 obtenu sur l'échantillon et on rejette l'hypothèse si $d^2 > \lambda$. Sinon, on n'est pas parvenu à mettre en évidence une non adéquation significative par rapport à la modélisation par la loi P_0.

Remarque : dans le cas où la loi P_0 dépend d'un (ou plusieurs) paramètre(s) inconnu(s) a priori θ, c'est-à -dire $P_{0,\theta}$, il faut estimer ces paramètres. Cependant, D^2 suit alors un khi-deux de degré de liberté $r-1-q$, où q est le nombre de paramètres estimés. On verra que la commande que l'on utilise avec R n'en tient pas compte. Il vaut donc mieux dans ce cas calculer à la main la valeur du quantile de la loi du χ^2 avec la commande que l'on a déjà vue : **qchisq(1-α,r-1-q)**.

Exemple : l'examen de 320 familles ayant 5 enfants s'est traduit par la distribution du tableau suivant :

Nbre de garçons et filles	$5g$ et $0f$	$4g$ et $1f$	$3g$ et $2f$	$2g$ et $3f$	$1g$ et $4f$	$0g$ et $5f$
Nbre de familles	18	56	110	88	40	8

Ce résultat est-il compatible avec l'hypothèse que les naissances respectives de garçons et filles sont indépendantes et équiprobables ? Pour répondre à cette question, on fait un test d'ajustement. Sous H_0, le nombre de garçons suit une loi binomiale $B(5, \frac{1}{2})$ et on peut ainsi calculer les probas théoriques de chaque classe, puis

$$d^2 = \frac{(18-10)^2}{10} + \frac{(56-50)^2}{50} + \frac{(110-100)^2}{100} + \frac{(88-100)^2}{100}\frac{(40-50)^2}{50} + \frac{(8-10)^2}{10} = 11.96.$$

On lit le seuil λ dans une table de khi-deux au paramètre $\nu = 6-1$ et on trouve 11.1 pour un risque de 0.05 et 15.1 pour un risque de 0.01. Donc en conclusion on rejette l'hypothèse d'équiprobabilité au risque 0.05 et on ne rejette pas au risque 0.01.

Remarque : l'option p.rescale sert à demander de renormaliser le vecteur des probabilités prob de telle maniée que la somme de ses composantes fasse 1. Elle est donc inutile si cela a été fait avant.

Remarque : On peut tester l'adéquation à une répartition arbitraire $(p_1, \ldots, p_r)$ on prend alors $prob = (p_1, \ldots, p_r)$ comme loi théorique (la définition de prob se fait donc à la main).

1.1.5 Le test d'indépendance.

On dispose d'un tableau de contingence de deux variables discrètes (ou discrétisées, c'est à dire triées à plat) X et Y dont on veut tester l'indépendance. On note p le nombre de modalités de X et r est le nombre de modalités de Y. On suppose que l'on dispose d'un échantillon de n individus et on note N_{ij} la variable aléatoire représentant le nombre d'individus de l'échantillon pour lesquels X est dans la classe a_i et Y dans la classe b_j. On définit également les variables $N_{i.} = N_{i1} + \cdots + N_{ip}$ (resp. $N_{.j} = N_{1j} + \cdots + N_{rj}$) qui représente le nombre d'individus de l'échantillons pour lesquels X est dans la classe a_i (resp. Y est dans la classe b_j) indépendamment de Y (resp. de X).

L'idée est assez similaire à celle utilisée pour construire le test d'ajustement. On va comparer les effectis observés pour chaque combinaisons d'une modalité (ou classe) de la variable X et d'une modalité (ou classe) de la variable Y avec ceux que l'on attendrait dans le cas de l'indépendance des variables. Sous l'hypothèse d'indépendance, pour tout $(i,j) \in \{1,\ldots,p\} \times \{1,\ldots,r\}$, la probabilité que X soit dans la classe a_i et Y dans la classe b_j est le produit des probabilités de chacune de ces classes :

$$\mathbb{P}(\{X \in a_i\} \cap \{Y \in b_j\}) = \mathbb{P}(\{X \in a_i\})\mathbb{P}(\{Y \in b_j\}).$$

Bien entendu, les probabilités d'observation des différentes classes ne sont pas connues. On les remplace dons par les proportions sur l'échantillon $p_X(i) = N_{i.}/n$ et $p_Y(j) = N_{.j}/n$. Par conséquent, en supposant l'indépendance des variables, on peut s'attendre à observer des effectifs théoriques

$$NT_{ij} = np_X(i)p_Y(j) = \frac{N_{i.}N_{.j}}{n}.$$

Une autre manière de le comprendre est de dire que parmi les $N_{i.}$ individus pour lesquels X est dans a_i les proportions d'observation des différentes valeurs (ou classes) de Y est la même que dans l'échantillon total, ce qui est logique dans le cas où les variables sont indépendante.

La méthode est la suivante :

1. Choix des hypothèses

$$H_0 : \text{ X et Y sont indépendantes}$$

H_1 : X et Y ne sont pas indépendantes

On se fixe $\alpha = 0.01$ par exemple.

2. On définit la statistique D^2 qui mesure la distance entre le tableau des observations et le tableau théorique que l'on s'attendrait à obtenir si les variables X et Y étaient indépendantes.

$$D^2 = \sum_{i,j} \frac{(N_{i,j} - \frac{N_{i.} * N_{.j}}{n})^2}{\frac{N_{i.} * N_{.j}}{n}},$$

où les $N_{ij}, N_{i.}, N_{.j}$ sont les variables dont les valeurs $n_{ij}, n_{i.}, n_{.j}$ sont les effectifs des "cases" du tableau, selon les notations déjà utilisées.
Le théorème de probabilité ci-dessous donne la loi de la variable aléatoire D^2 sous l'hypothèse H_0 et justifie en partie son choix.

Théorème 1 *Si X et Y sont des variables indépendantes (c'est-à-dire) si H_0 est vraie, et si n est grand > 30 (et que les effectifs théoriques $NT_{ij} > 5$) alors la statistique D^2 suit approximativement une loi de khi-deux à $(p-1)(r-1)$ degrés de liberté, où p est le nombre de modalités de X et r est le nombre de modalités de Y.*

3. On rejettera H_0 si D^2 est "grand" et donc la zone de rejet est de la forme $D^2 > \lambda$ où la valeur critique λ vérifie :

$$P(D^2 > \lambda | H_0) = P(\chi^2 > \lambda) = \alpha.$$

On lit λ dans la table du khi-deux.

4. On calcule d^2 pour l'échantillon observé et on conclut.

Remarque : Ce test d'indépendance peut être utilisé pour tester l'homogénéité d'un caractère. On étudie en effet l'indépendance entre le caractère et l'appartenance aux différentes classes. Attention, si on rejette l'hypothèse d'indépendance, le résultat du test ne nous permet pas de conclure si la proportion d'un caractère donné est plus fort dans une classe que dans l'autre. Un test d'homogénéité sur les proportions doit alors être réalisé.

Exemple 1 : on veut tester l'efficacité d'un sérum. Les résultats sont rassemblés dans le tableau ci-dessous.

	guérison	non guérison	total
avec sérum	75	25	100
sans sérum	65	35	100
total	140	60	200

$$d^2 = \frac{(75-70)^2}{70} + \frac{(65-70)^2}{70} + \frac{(25-30)^2}{30} + \frac{(35-30)^2}{30} = 2.38$$

Le khi-deux a 1 degré de liberté nous donne 3.84 au risque 0.05 et donc on ne peut pas rejeter l'indépendance. Ainsi soit on fait d'autres tests, soit on conclut que le sérum n'est pas significativement efficace.

Exemple 2 : parmi un groupe de malades qui se plaignent de ne pas bien dormir, certains ont absorbés un sommnifère sous forme de cachets, d'autres ont absorbé des placébo (cachets en sucre). Après quoi pour tester l'efficacité réelle des cachets on leur a demandé s'ils avaient bien dormi.

	bien dormi	pas bien dormi
ont eu un vrai somnifère	44	10
ont eu du sucre	81	35

On calcule d^2 et on cherche le seuil sur le khi-deux de paramètre 1 et on voit qu'on ne peut pas rejeter l'indépendance, donc on a des doutes sur le produit !

1.1.6 Exercices supplémentaires et Annales

1. Dans dix corps d'armée prussiens, pendant une période de 20 ansallant de 1875 à 1894, le nombre de morts par corps d'armée et par an dus à la ruade d'un cheval est donnée par le tableau suivant :

X	0	1	2	3	4
Effectif	109	65	22	3	1

   – Quelle est la population étudié ? Quelle est la taille de l'échantillon considéré ?
   – Démontrer que l'estimateur du maximum de vraisemblance du paramètre λ d'une loi de poisson est égal à la moyenne arithmétique $\overline{X}$.
   – Faire un test d'ajustement des données précédens par uen loi de poisson en prenant un risque $\alpha = 0.1$. (Détailler bien les quatre étapes de la construction du test).

2. On a interrogé 200 élèves d'un lycée sur le type d'études supérieures qu'ils désiraient entreprendre. Les résultats de l'enquête figurent dans le tableau ci-dessous. Au vu de ces résultats, en prenant un risque de 0.01, semble-t-il exister uen relation entre le choix des études et le sexe ?

	Garçons	Filles	Total
Littéraires	60	60	120
Scientifiques	42	18	60
Techniques	18	2	20
Total	120	80	200

3. Le tableau suivant donne le nombre d'étudiants qui ont été brillants et médiocres devant trois examinateurs : M.X, M.Y, et M.Z.

	M.X	M.Y	M.Z	Total
Brillant	50	47	56	153
Médiocre	5	14	8	27
Total	55	61	64	180

   Peut on dire que l'on a les mêmes chances selon l'examinateur avec qui on passe ?

4. On étudie la population des ménages d'un pays lointain. On regarde le nombre d'enfants par ménage pour un échantillon de taille 1000 de cette population.

Enfants	1	2	3	4	5	6	7	8	10
Effectif	510	248	132	55	29	15	2	8	1

   – Calculer la moyenne empirique de l'échantillon ci-dessous et tracer sa boîte à moustache.
   – On fait l'hypothèse que l'on peut modéliser par une loi géométrique. Proposer un estimatimateur du paramètre de la loi et en donner une estimation.
   – Faire un test d'ajustement par rapport à cette loi géométrique avec R.

5. On a effectué dans le département du Loiret un tirage aléatoire d'un échantillon de 1000 élèves agés de 6 à 12 ans. On a fait passer à ces élèves un test psychométrique comportant deux épreuves. Le tableau de contingence ci-dessous donne la répartition par âge et nombre d'épreuves réussies.

Réussite Age	[6 ;7[	[7 ;9[	[9 ;11[	[11 ;12]
0	35	70	80	45
1	57	125	170	114
2	23	75	110	96

Peut on dire avec un risque de 0.05 puis de 0.01 que l'âge et la réussite aux épreuves ne sont pas indépendants ?

6. On a mené une étude sur un échantillon de 500 appareils d'un même type, concernant la surface d'usure S d'une pièce déterminée au bout de 5000 heures de fonctionnement. Les résultats sont regroupés dans le tableau ci-dessous et létude est menée ensuite dans R.

S.	[0.65 ;0.75[	[0.75 ;0.95[	[0.95 ;1.15[	[1.15 ;1.35[	[1.35 ;1.55[	[1.55 ;1.75[	[1.75 ;1.95[	[1.95 ;2.05[	[2.05 ;2.25[
E.	2	12	49	118	155	110	44	9	1

Calculer les effectifs théoriques puis tester au risque de 0.05 l'hypothèse que la surface d'usure peut être modélisée par une loi normale.

7. On a enregstré, pendant 50 jours, le nombre d'accidents automobiles à Pearson-Gulch. Les résultats sont les suivants :

Nombre d'accidens	0	1	2	3	4
Nombre de jours correspondant	21	18	7	3	1

On fait l'hypothèse que les données proviennent d'une loi de Poisson de paramètre a inconnu.
- Donner l'estimateur de a par la méthode du maximum de vraisemblance.
- Donner une estimation $\hat{a}$ de a à partir des données.
- Calculer les coefficiens théoriques en supposant que X suit une loi de Poisson de paramètre $\hat{a}$.
- Effectuer un test concernant la validité de cette modélisation.

1.2 Utilisation des logiciels de statistiques.

Quand on utilise un logiciel, on raisonne un peu différemment. Le logiciel calcule la valeur t de la statistique de test observée sur nos données et fournit la probabilité que la statistique de test T (variable aléatoire) dépasse cette valeur. On utilise ici le fait que sous l'hypothèse nulle on connaît, au moins approximativement, la loi de T. Le logiciel ne renvoit généralement pas la valeur seuil du test c_α mais la p-value.
La p-value est le risque de première espèce qu'on prend si on décide de rejeter l'hypothèse nulle. Si on choisit de prendre un risque de première espèce de valeur α, on rejette H_0 si la p-value est infèrieure à α. Car le fait que la p-value soit inférieure à α est équivalent au fait que t soit dans R_α.

1.3 Retour sur l'analyse de la variance

L'analyse de variance permet de mesurer les effets d'une ou plusieurs variables qualitatives appelées aussi facteurs sur une variable quantitative Y.

Ce test vise à établir la possibilité de rejeter l'hypothèse H_0 d'égalité simultanée des moyennes correspondant aux sous-populations formées par les différentes modalités des facteurs. On peut donc l'interpréter comme une génralisation des tests d'homogénéité sur les moyennes que nous avons vus précédemment qui se limitaient au cas de deux sous-populations.

Voici quelques exemples :
- On souhaite évaluer les effets de différents traitements sur le taux de virus dans le sang chez des patients atteints d'une maladie M .
- On mesure le rendement pour différentes variétés de mais soumis à 6 types de fertilisants azotés.
- On étudie des rendements laitiers sur des vaches d'une espèce donnée en fonction du régime alimentaire (paille, foin, herbe,aliments ensilés) et de la dose (faible, forte).
- On étudie les temps de germination de différentes variétés de carottes sur 4 types de sol.
- On étudie la corrosion de différents tuyaux en fonction de la nature du sol dans lesquels ils se trouvent et du type de protection (peinture) qu'ils ont reçu.
- Des études marketting examinent régulièrement l'impact de différentes campagnes publicitaires sur les ventes de différents aliments.
- On peut examiner le rendement d'un paquet d'action en fonction de la stratégie de palcement.
- On peut aussi comparer les salaires d'embauches selon les écoles d'origine.

Comme on l'a déja évoqué dans le chapitre sur les statistiques descriptives, l'usage préalable des boites à moustaches pour opérer un rapide jugement graphique est vivement conseillé.

EXEMPLE INTRODUCTIF : Un forestier s'intéresse aux hauteurs moyennes de trois forêts. Pour les estimer, il échantillonne un certain nombre d'arbres de chaque forêt et mesure leur hauteur. Les résultats se présentent sous la forme suivante :

Hauteur (m)	23,4	22,1	22,5	22,5	24	18,9	23,7	21,1	24,4	24,6	24	24,9	23,5	25	24,5	26,2
N° Forêt	1	3	3	2	2	3	2	3	1	1	2	1	3	1	3	1

La variable qualitative (numéro de forêt) a ici trois modalité, à qui on associe 3 groupes. Pour représenter graphiquement un tel tableau, on représente sur un même graphique les résultats correspondants à chacun des groupes, et on compare ainsi la dispersion des résultats obtenus, souvent à l'aide des 3 boites à moustaches placées parallèllement sur un même graphique. On voit ainsi immédiatement, les différences de médianes et de dispersions. Il semble qu'en moyenne la première forêt soit plus élevée que la seconde, elle même plus élevée que la troisième. La variance de la troisième étant plus élevée que celle des autres, il y un arbre de 18,9 mètre et un de 24,5 (on avait déja remarqué ça avec les boites). Les forêts semblent différentes, mais la variance des arbres est relativement grande et on peut se demander si les forêts sont significativement différentes et si les écarts ne sont pas dus à l'échantillonnage.

Pour aller plus loin dans l'analyse, on réalise une analyse de variance ou ANOVA dont l'objectif est de détecter des différences significatives entre les valeurs moyennes prises par Y sur les différentes sous-populations (apelées aussi classes). En d'autres termes on se pose la question suivante : **la variabilité observée dans les données est-elle uniquement due au hasard**

de l'échantillonnage *(c'est à dire le choix des arbres que l'on mesure)*, ou bien existe-t-il effectivement des différences significatives entre les classes imputables au facteur *(c'est à dire les différentes forêts)*. Pour cela, on va comparer **la variance intraclasse (ou variance résiduelle)** qui résume la variabilité à l'intérieur des classes (à l'intérieur des forêts) et la **variance interclasse (ou variance des moyennes)** qui décrit la variabilité entre les valeurs moyennes des différentes classes *(des différentes forêts)*.

Dans ce paragraphe, on suppose qu'on a un seul facteur (une seule variable qualitative explicative) à k modalités (on peut cependant étendre la méthode de l'ANOVA pour prendre en compte plusieurs facteurs). On souhaite par exemple tester les effets de k traitements sur une maladie M mesurée par le taux de virus dans le sang. On administre donc respectivement les k traitements à $n_1, \cdots, n_k$ patients et on veut tester l'égalité des moyennes.

De manière générale, on note dans la suite de ce chapitre k le nombre de modalités du facteur et $n_1, \ldots, n_k$ les effectifs des sous-populations $P_1, \ldots, P_k$ correspondant aux différentes modalités du facteur. On note $m_1, \ldots, m_k$ les moyennes de la variable qualitative Y sur ces différentes sous-populations. On note généralement $Y_{i,j}$, $1 \leq i \leq n_j$, $1 \leq j \leq k$ les variables aléatoires représentant la variable étudiée Y sur les différents individus de l'échantillon afin de faire apparaître la sous-population à laquelle il appartient au travers de l'indice j. On notera dans ce qui suit

$$\overline{Y}_j := \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{i,j} \text{ et } \overline{Y} := \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} Y_{i,j}.$$

Quelques idées clés amenant à la contsruction de l'ANOVA :

Les écarts à la moyenne se décomposent de la façon suivante :

$$Y_{i,j} - \overline{Y} = (Y_{i,j} - \overline{Y}_j) + (\overline{Y}_j - \overline{Y}).$$

Remarque : On écrit souvent le modèle sous la forme suivante

$$\forall j \in \{1, \ldots, k\}, \forall i \in \{1, \ldots, n_j\}, Y_{i,j} = \overline{Y} + a_j + \epsilon_{i,j},$$

dans lequel $a_j = \overline{y}_j - \overline{y}$ représente l'effet de la modalité j du facteur sur la valeur moyenne et les variables $\epsilon_{i,j}$ (appelées résidus) représentent l'aléa qui n'est pas expliqué par les moyennes de chaque classe.

Le point clé de la méthode est la décomposition suivante de la variance totale (provenant d'une identité de type pythagore)

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{i,j} - \overline{Y})^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{i,j} - \overline{Y}_j)^2 + \sum_{j=1}^{k} (\overline{Y}_j - \overline{Y})^2.$$

qui s'écrit aussi
SCT(somme des carrés totaux) (variance totale)= SCR(sommes des carrés résiduels) (variance intraclasse)+ SCE(somme des carrés expliqués)(variance interclasse).

Lorsque le facteur n'a pas d'effet, les moyennes sur les différentes sous-populations $m_1, \ldots, m_k$ sont les mêmes. Par conséquent, les moyennes arithmétiques des différenes sous-populations $\overline{Y}_1$, $\overline{Y}_k$ sont semblables (par la loi des grands nombres) et la variance interclasse est faible (proche de 0)

tandis que la variance intraclasse est proche de la variance totale. Lorsque le facteur à un effet assez important, les moyennes sur les différentes sous-populations $m_1, \ldots, m_k$ sont sensiblement différentes. Dans ce cas, les moyennes arithmétiques des différenes sous-populations $\overline{Y}_1$, $\overline{Y}_k$ ont tendance à être différentes (par la loi des grands nombres) et la variance interclasse prend des valeurs plus importantes. La variance interclasse quantifie l'effet du facteur (qui peut être plus ou moins forte). La variance intraclasse est plus faible que la la variance totale et correspond à la variabilité qui n'est pas expliquée par les valeurs moyennes sur chaque classe (c'est à dire la variabilité des résidus $\epsilon_{i,j}$).

On peut dans un premier temps penser à considérer le ratio R^2 de la variance interclasse par la variance totale qui représente la part de variance interclasse (celle qui est expliquée par le facteur). Cependant, on préfère considérer la statistique de test :

$$F = \frac{SCE/k-1}{SCR/n-k}$$

dont on connaît la loi de probabilité (loi de Fisher à (k-1,n-k) degrés de libeté) sous certaines hypothèses (normalité de la loi de Y, indépendance entre les sous-populations et homogénéité des variances) lorsqu'il n'y a pas d'effet du facteur. Plus le facteur a une influence sur la variable à expliquer, plus F a tendance à être grand. Par conséquent, il paraît logique de rejeter l'hypothèse qu'il n'y a pas d'effet du facteur si l'indicateur F est supérieur à un seuil. On fixe ce seuil c_α de manière à contrôler la probabilité α de se tromper en concluant que le facteur a un effet. On se sert du fait que l'on connaît la loi de F lorsque le facteur n'a pas d'effet pour choisir le seuil $c_\alpha := Q_{F(k-1,n-k)}(1-\alpha)$ tel que :

$$\mathbb{P}_{\mathcal{H}_0}(F > c_\alpha) \approx \alpha.$$

La démarche de l'ANOVA est la suivante : *Remarque importante :* pour que l'ANOVA soit valide, on doit vérifier les hypothèses suivantes :

1. les variables $Y_{i,j}$ sont de loi normale.
2. les variances des $Y_{i,j}$ ne varient pas d'une sous-population à l'autre (i.e. la variance de la variable quantitative à expliquer ne dépend pas des modalités du facteur).
3. les variables $Y_{i,j}$ sont indépendantes.

Une étude préalable peut être faite pour vérifier ces hypothèses. Pour contrôler l'homoscedasticité (c'est à dire que la variance de la variable quantitative à expliquer ne dépend pas des modalités du facteur) on compare les écarts-type intra-groupe et on fait un test de Bartlett (ou un test de Levenne). Ensuite, on fait souvent l'histogramme des résidus suivi d'un test de Shapiro-Wilks sur les résidus pour tester la normalité (on peut également réaliser un test de Kolmogorov-Smirnov). Enfin, la dernière condition d'indépendance est d'ordinaire satisfaite lorsque l'on a un échantillon aléatoire simple ou en utilisant une procédure "d'aléatorisation" (ou de randomisation) : procédure par laquelle on affecte au hasard chaque individu à un groupe expérimental.
Si certaines de ces conditions (normalité et homoscédasticité) ne sont pas vérifiées on utilisera plutôt un test non-paramétrique de Kruskal-Wallis.

Etape 1 :

$$H_0 : m_1 = m_2 = \cdots m_k \text{ et } H_1 : \exists 1 \leq j \neq l \leq k,\, m_j \neq m_l.$$

Fixer le seuil du test α (en général 0.05)

Etape 2 :

La statistique de test est

$$F = \frac{SCE/k-1}{SCR/n-k}$$

sous H_0 elle suit une loi de Fisher à $(k-1, n-k)$ degrés de liberté. Elle "n'est pas grande si H_0 est vraie" et a tendance à prendre des valeurs positives plus fortes si H_0 n'est pas vérifiée.

Etape 3 :

Il semble assez logique au vu de l'étape 2 et des commentaires qui précèdent de choisir de rejeter l'hypothèse H_0 si la statistique F est supérieure à une valeur seuil c_α. On prend donc $R_\alpha = \{F > c_\alpha\}$ en choisissant c_α le plus petit possible tel que $\mathbb{P}_{H_0}(F > c_\alpha) \leq \alpha$. Puisque F est de loi de Fisher(k-1,n-k) sous H_0 on prend $c_\alpha = Q_{F(k-1,n-k)}(1-\alpha)$.

Etape 4 :

On calcule finalement la valeur f de F obtenues à partir de nos données.

Si $t > c_\alpha$ on rejette l'hypothèse nulle avec un risque α de se tromper. Sinon, pour un risque α, l'effet du facteur n'est pas significatif.

De façon générale, après l'ANOVA, plusieurs cas se présentent :

1. on décide de ne pas tenir compte du facteur, puisqu'on n'a pas observé d'effet significatidf de celui-ci sur le phénomène étudié. Ainsi, pour modéliser la variable réponse pour un individu, on prend la moyenne de l'échantillon total.
2. au contraire, il semble que le phénomène dépende du facteur, cette fois ci, pour modéliser la variable réponse sur un individu, on évalue d'abord la modalité du facteur prise par cet individu et on prend la moyenne sur l'échantillon des individus ayant pris cette modalité.
   Cepedant, attention, il est important de comprendre que l'analyse de la variance n'est pas un test permettant de « classer » les moyennes des différentes sous-populations (par exemple de dire que la moyenne d'une sous population est supérieure à la moyenne d'une aute mais plus petite que celle d'une troisième). Comme on l'a noté précédemment, l'hypothèse nulle $\mathcal{H}_0$ revient à dire que toutes les moyennes sont égales. Le but ici est donc beaucoup plus « humble » : il s'agit de comparer des moyennes de différents groupes et de dire si, parmi l'ensemble, au moins une d'entre elles diffère des autres, mais on ne sait ni laquelle ni combien d'entre elles. Déterminer quel groupe a un effet différentiel, c'est-à-dire quel groupe présente une moyenne de la variable étudiée différente des autres, est un problème tout à fait différent. Il peut se poser après une ANOVA et les tests associés sont dits « tests de comparaisons multiples », ou MCT pour Multiple Comparison Test. Ces tests obligent en général à augmenter les risques de l'analyse (en termes de risque statistique). Dans la biologie moderne, notamment, des tests MCT permettent de prendre en compte le risque de façon correcte malgré le grand nombre de tests effectués (par exemple pour l'analyse de biopuces). On pourra se reporter notamment aux procédures de Bonferonni et de Sidak.
   Il s'agit d'une généralisation à k populations du test T de Student de comparaison de moyennes de deux échantillons. Voir également les test de Dunett et le HSD de Tuckey.

Utilisation du logiciel R

Le logiciel nous donne la probabilité qu'une variable de Fisher dépasse F et si cette p-value est inférieure au risque on rejette H_0 et on conclut que le facteur a bien une influence sur la variable à expliquer.

Tableaux récapitulatif proposé par les logiciels.

source de variation	degrés de liberté	Somme	carrés moyen	F	$p-value$
Expliqués	$p-1$	SCE	CME	CME/CMR	
Résidus	$n-p$	SCR	CMR		
Total	$n-1$	SCT			

EXERCICE :
Retour à notre exemple concernant les forêts.

1. Réaliser une analyse de variance.
2. Discuter les résultats obtenus et conclure.

EXERCICE : On étudie la résistance à la corrosion de différents tuyaux. On dispose de deux variables qualitatives pour tenter d'expliquer la variabilité des observations (le type de sol dans lequel ils se trouvent et le type de protection (peinture) qu'ils ont reçu. Les données sont disponibles dans le fichier tuyaux de la librairie lycee.

1. Charger les données avec R.
2. Tracer les boîtes à moustaches (box plot) de la corrosion pour comparer les résultats selon les différents types de sol puis selon les différentes protections.
3. Effectuer une analyse de variance pour mieux cerner les effets de chacun des facteurs. Commenter et conclure.
4. Pour aller plus loin, lorsque l'on a deux facteurs, on réalise plutôt une ANOVA à deux facteurs que deux ANOVA séparée à un facteur. On peut le faire avec R de la manière suivante : $aov(y\ x1 + x2)$ où + signifie que l'on considère que les effets des deux facteurs sont indépendants. On peut remplacer le signe + par le signe $*$ afin de considérer également les interactions entre les deux facteurs (l'effet d'un facteur peut dépendre de la valeur de l'autre facteur). Essayer ces commandes et interpréter les résultats.

EXERCICE : Revenons au jeu de données entreprise (dans la librairie lycee) que nous avions étudié il y quelques semaines. Pour essayer de prévoir la défaillance des entreprises, l'économiste W. BEAVER inroduit le ratio défini, pour chaque entreprise, par le quotient de la marge brute d'autofinancement (cash flow) par la dette totale.

1. Charger les données avec R.
2. Tracer les boîtes à moustaches (box plot) du ratio pour comparer les résultats selon les différents types d'entreprises (saines ou défaillanes).
3. Effectuer une analyse de variance pour mieux cerner les effets du facteur. Commenter et conclure.